

Building a Genomics Resource

Progenetix - From Experiments to APIs

Michael Baudis | UZH BIO390 HS25

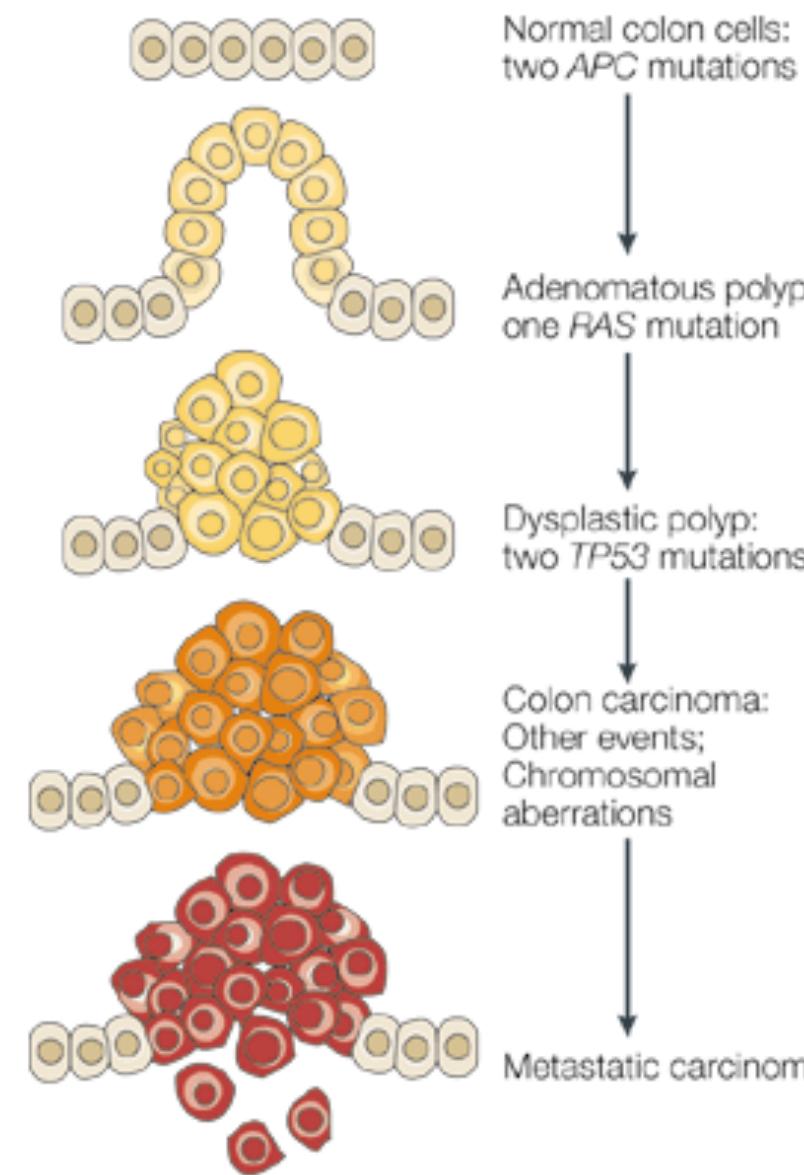
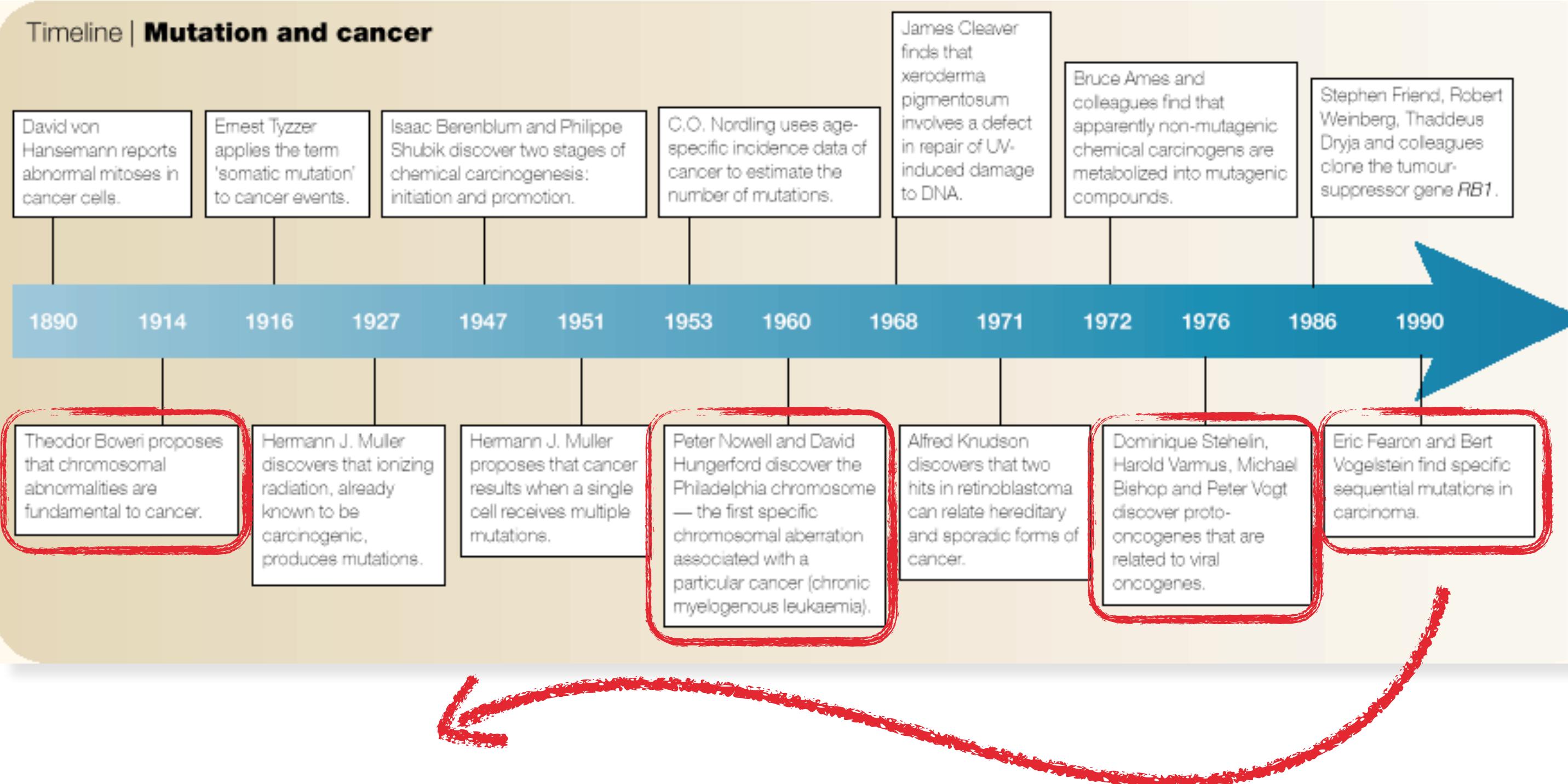


Building a Genomics Resource

A (personal) journey through time...

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

Timeline | Mutation and cancer



Cancers are based on acquired and inherited genomic mutations

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



Theodor Boveri (1914)

Observations in sea urchin eggs

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

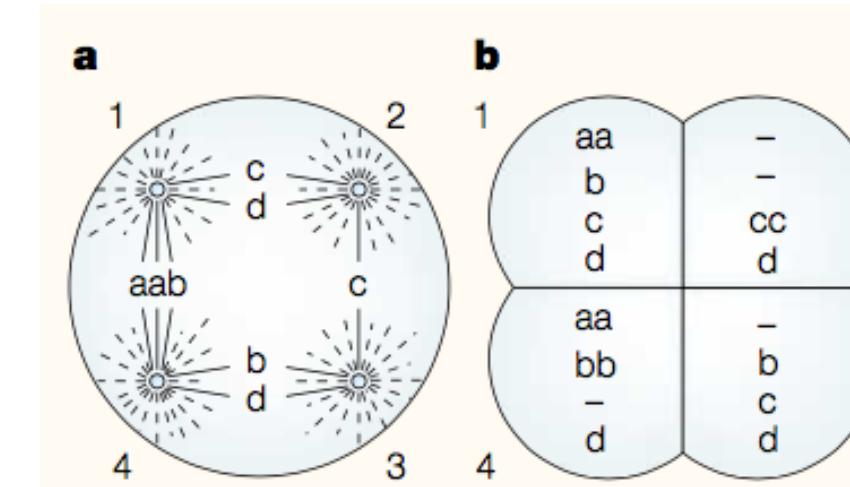
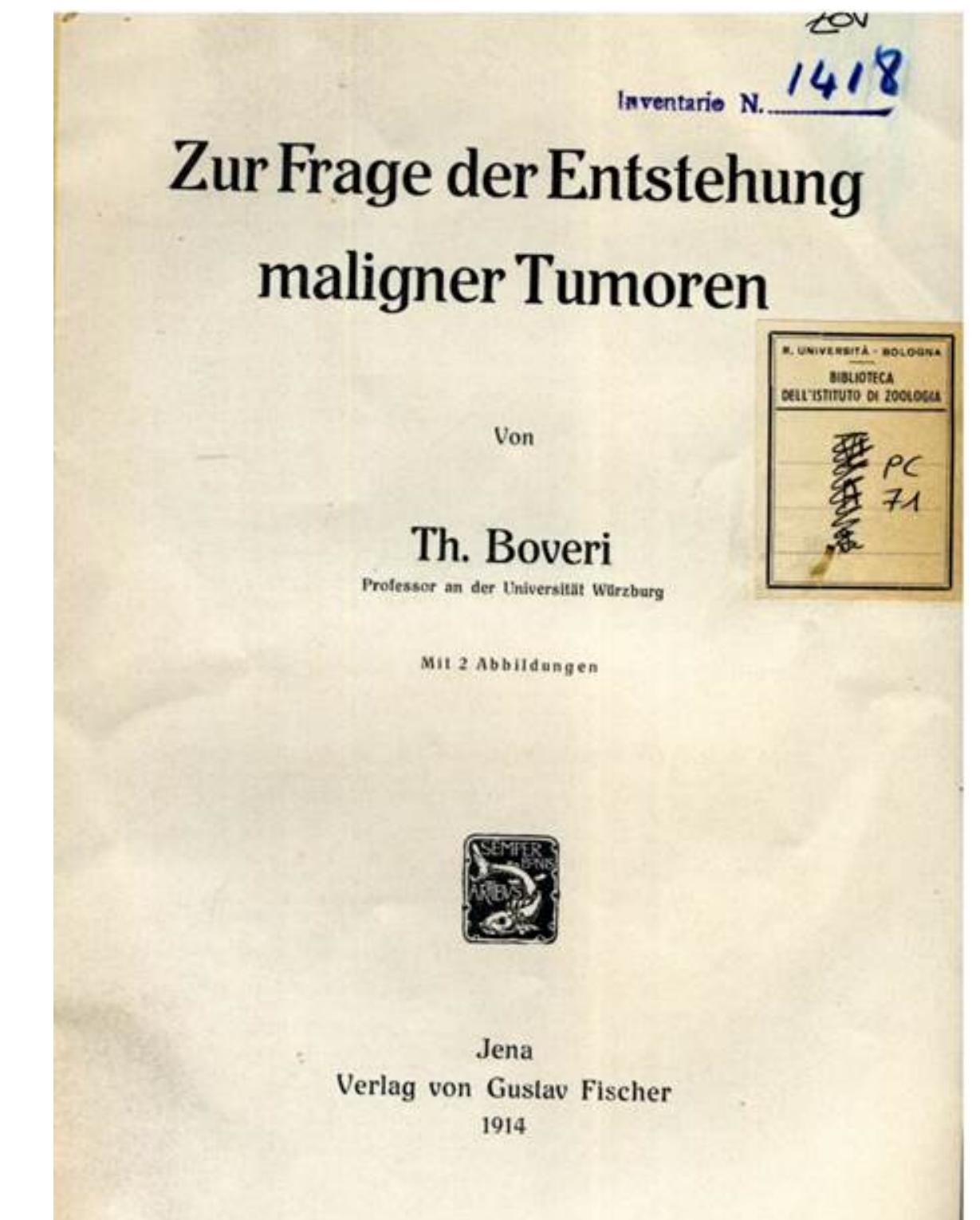
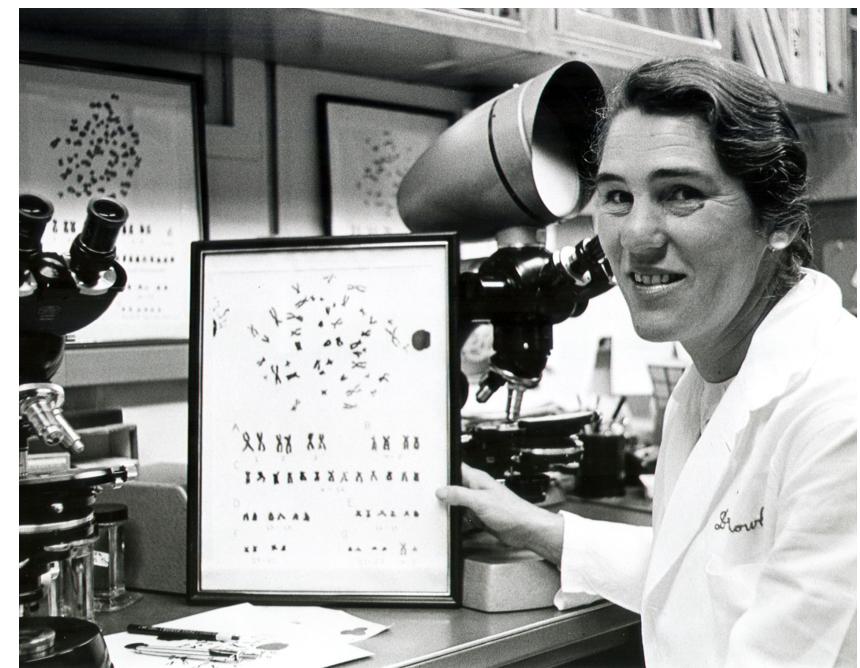


Figure 2 | Multiple cell poles cause unequal segregation of chromosomes. **a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached to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



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

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



Janet Rowley (1972/73)

Chromosomal translocations in cancer

- Recurrent chromosomal translocations in leukemias /lymphomas
- "Philadelphia chromosome" in CML (Nowell & Hungerford, 1960) represents a reciprocal translocation between chromosomes 9 and 22
- 1972: t(8;21) ALL manuscript rejected by NEJM
- 1973: t(9;22) manuscript rejected by *Nature* "with some reasonable comments and some truly wrong"
- Clinical implications: **Tyrosine Kinase inhibitors** as standard first-line therapy in CML
 - first trials in 1998 (STI-571; Imatinib/Gleevec)
 - cf. Druker BJ, Lydon NB (2000). Lessons learned from the development of an Abl tyrosine kinase inhibitor... J Clin Invest 2000;105:3-7

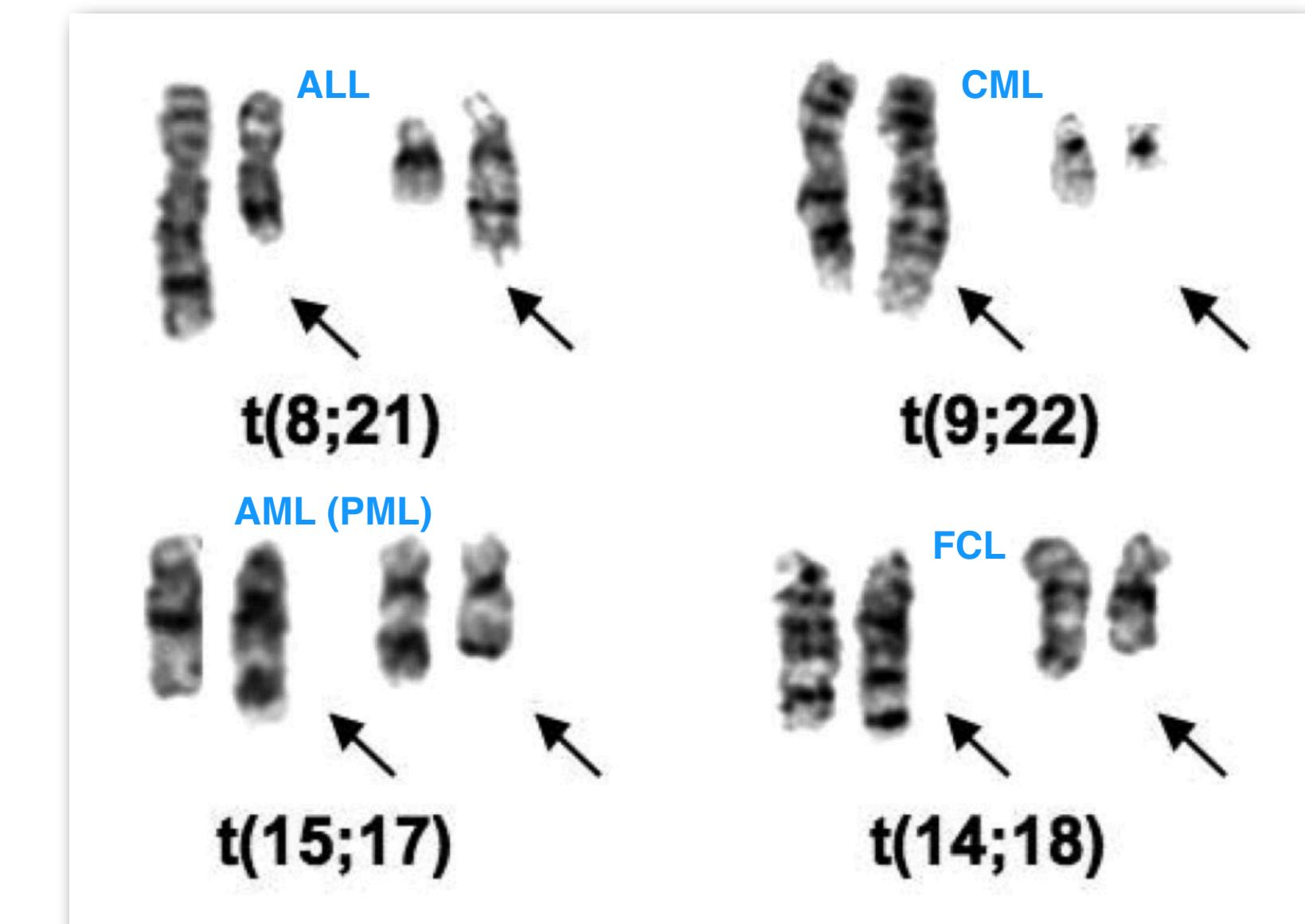
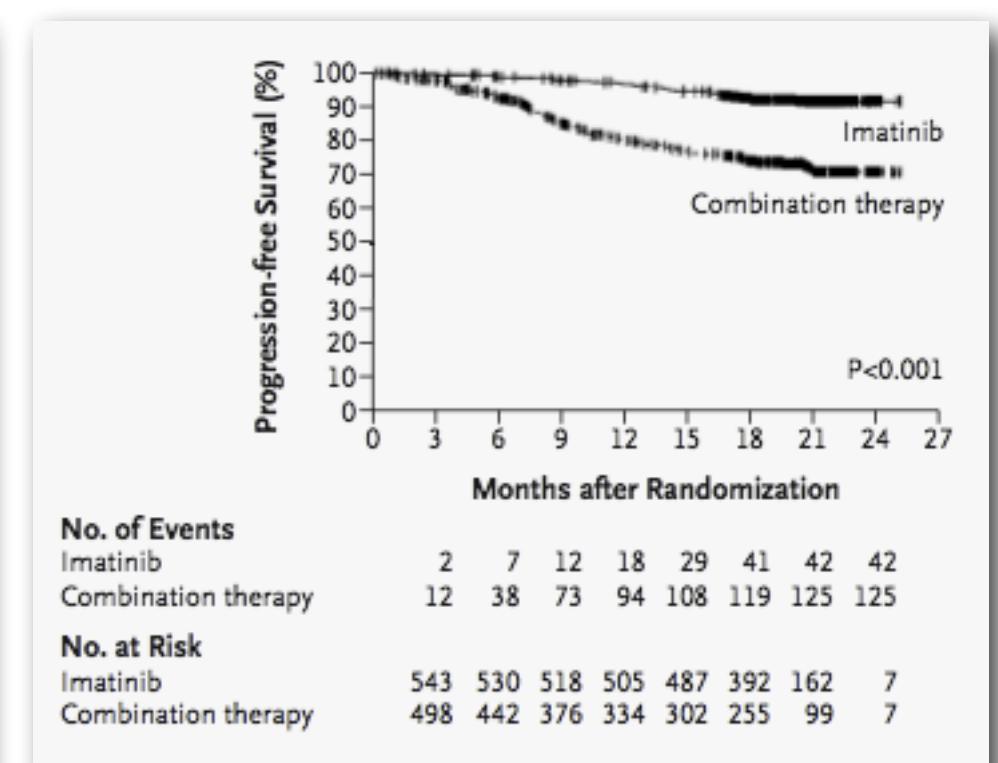
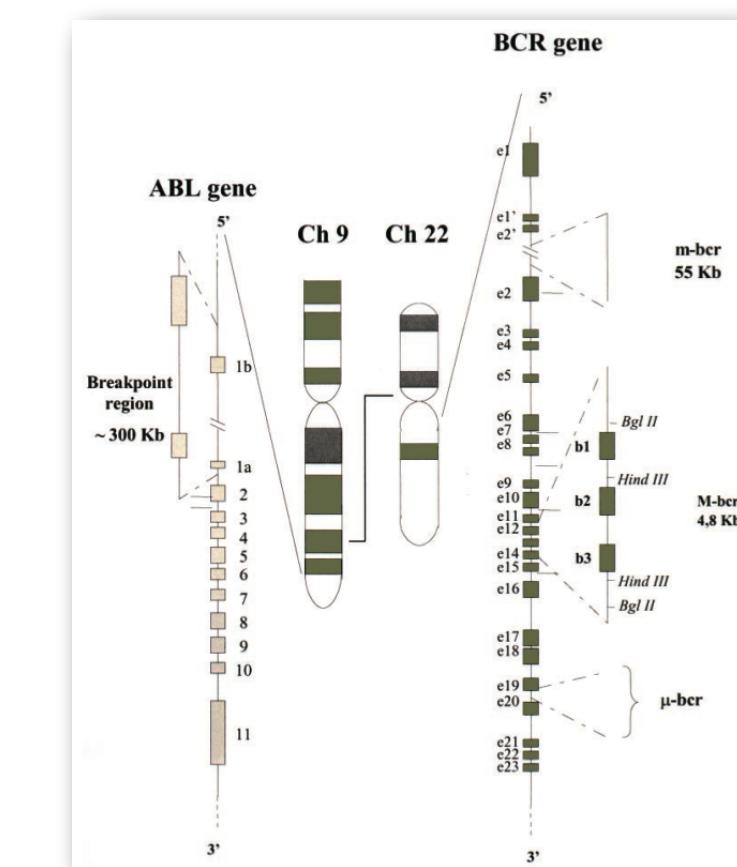


Figure 1. Partial karyotypes of common translocations discovered by Rowley.
The translocations appear in the order in which they were discovered.

Janet D Rowley. Chromosomal translocations: revisited yet again
Blood (2008), 112(6)

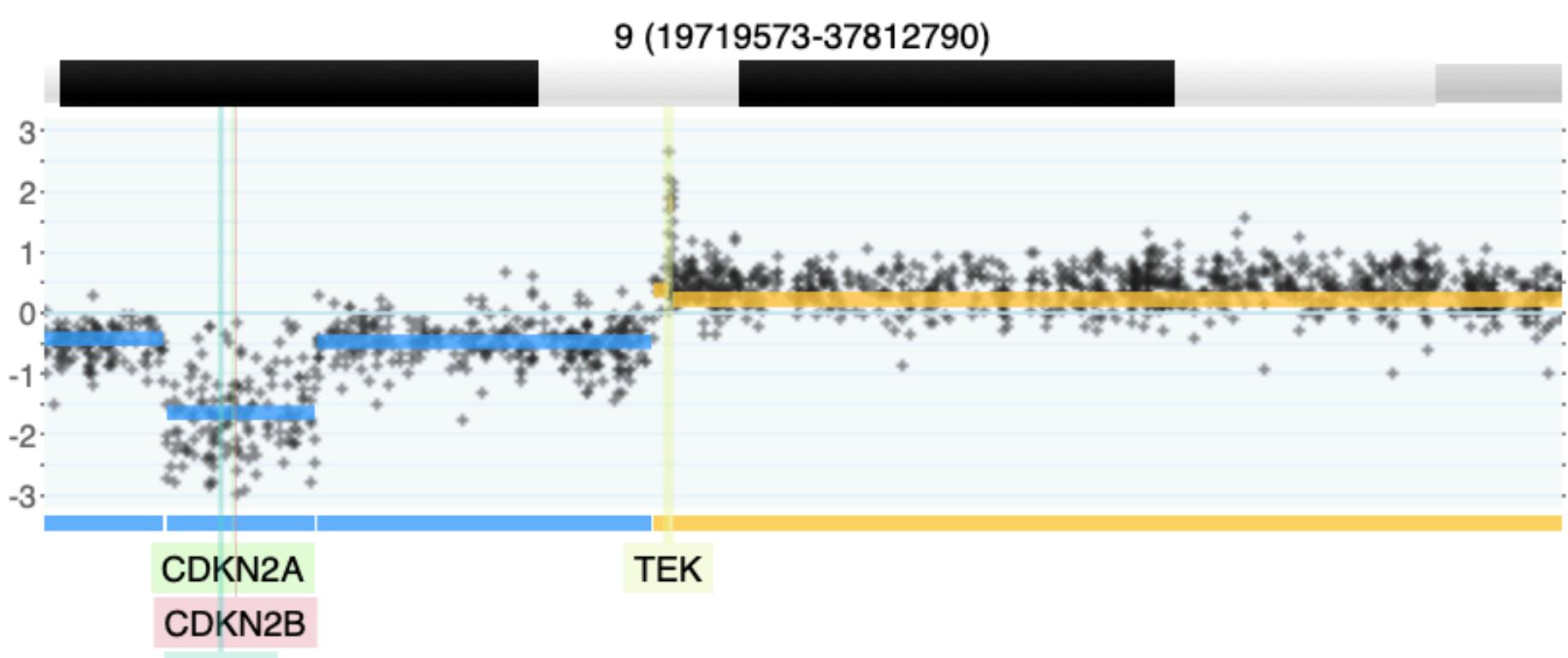
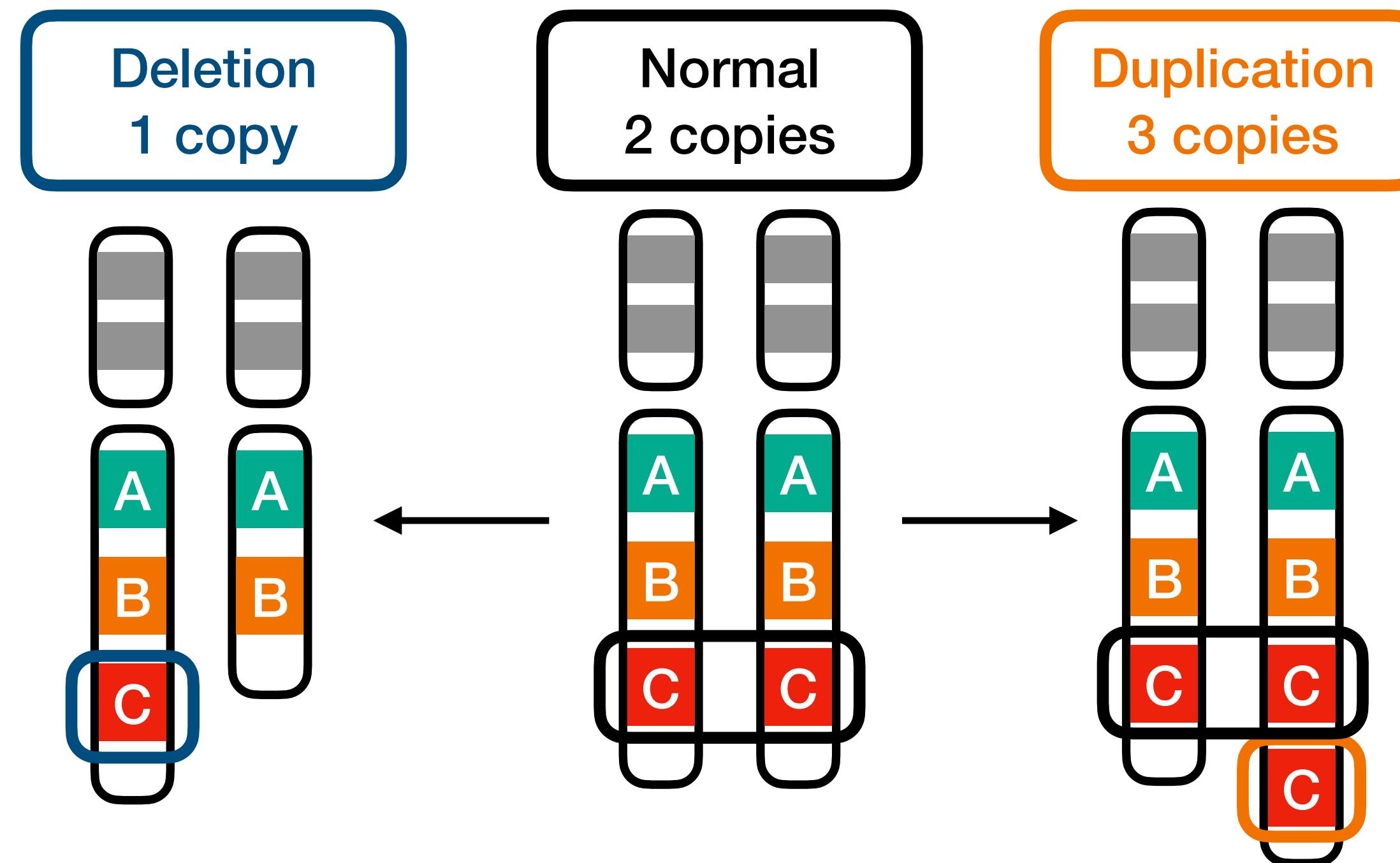


Event free Survival in first large Imatinib Trials

Pane et al. BCR/ABL genes
Oncogene (2002), 21 (56)

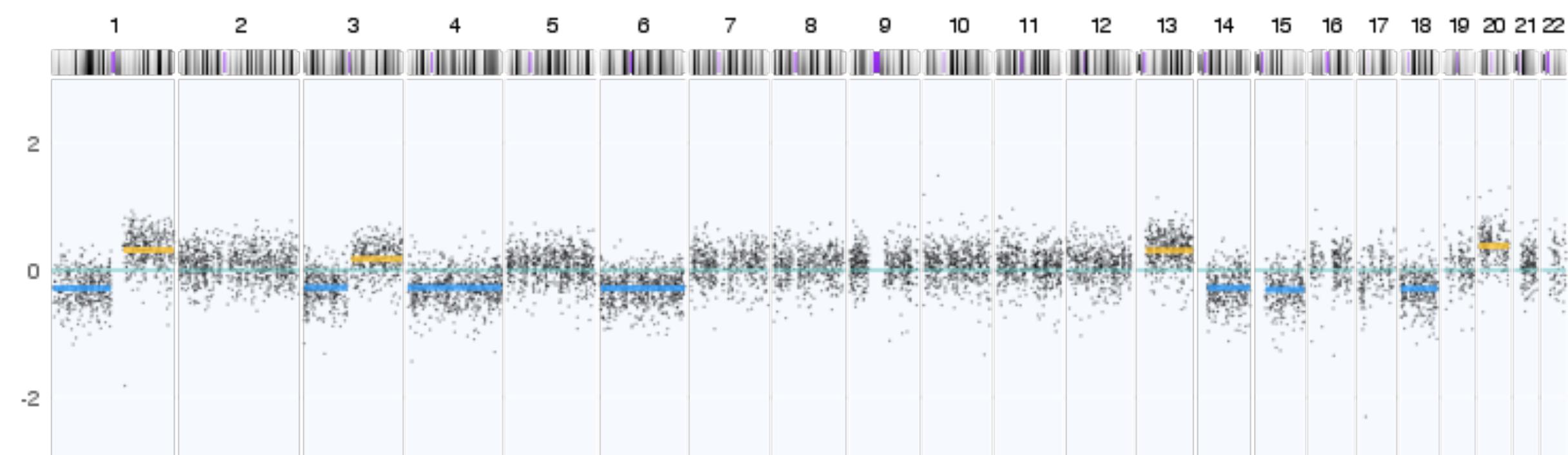
O'Brien et al. Imatinib compared with interferon and low-dose cytarabine...
NEJM (2003) vol. 348 (11)

Copy Number Variant (CNV)



2-event, homozygous deletion in a Glioblastoma

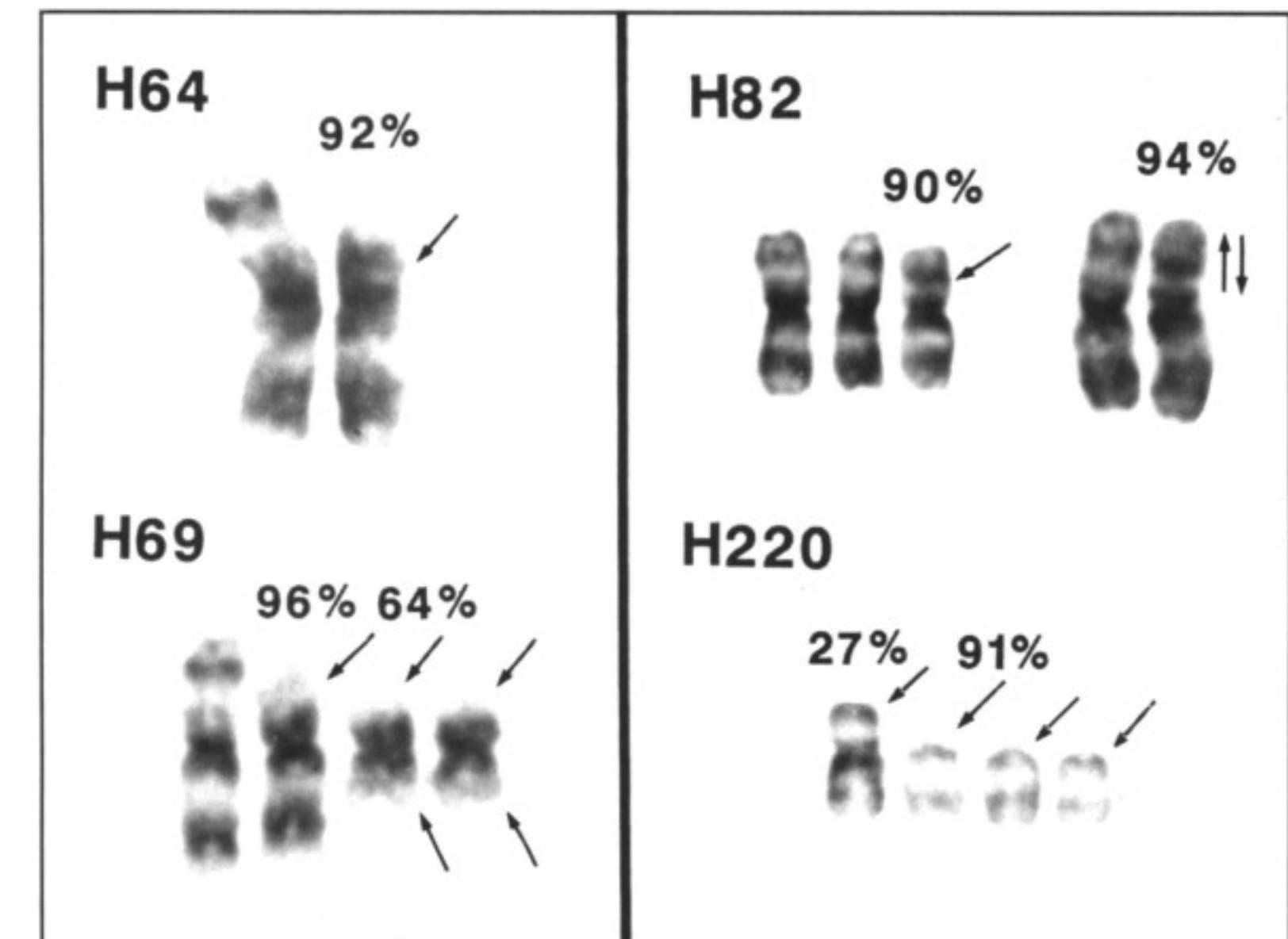
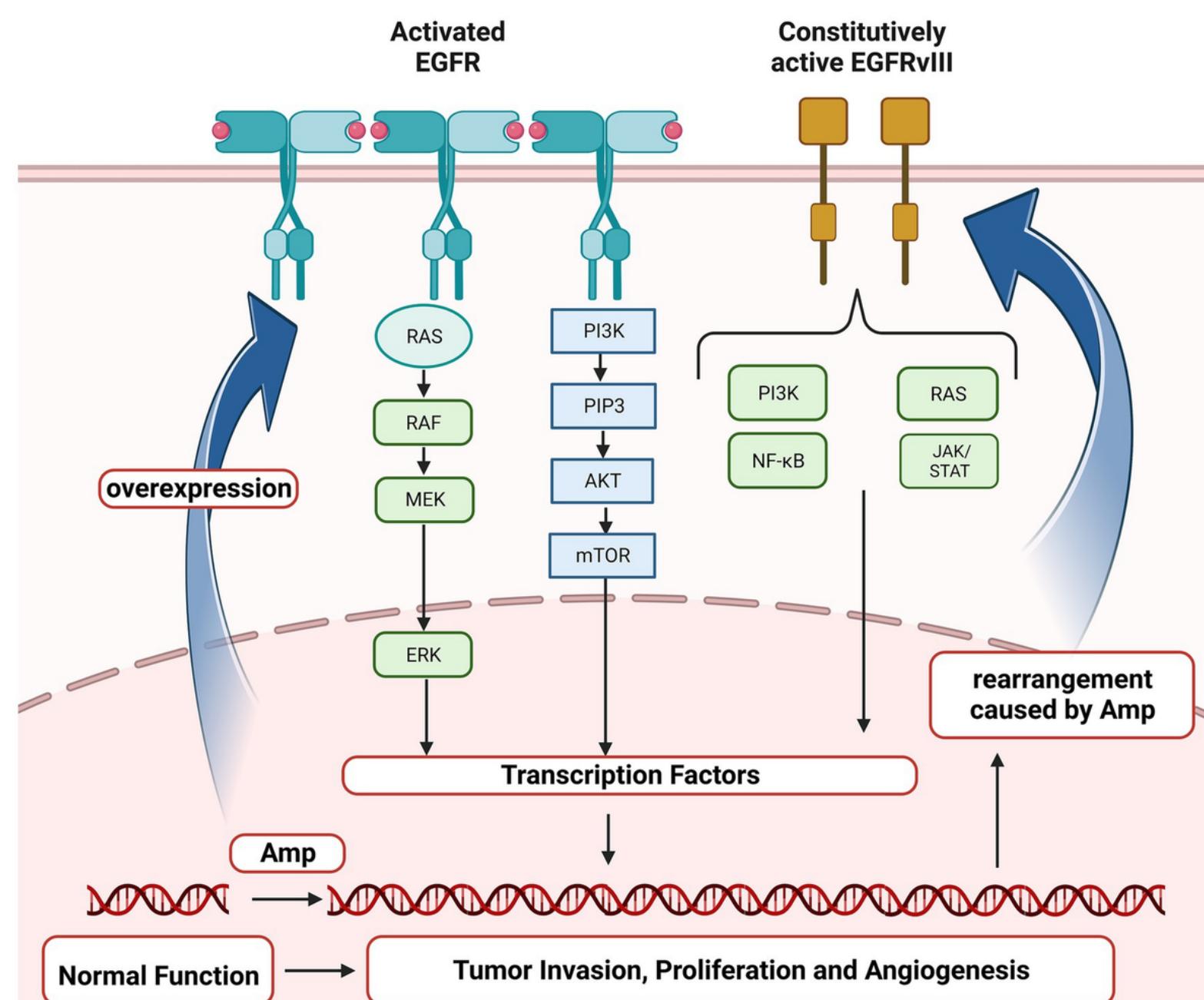
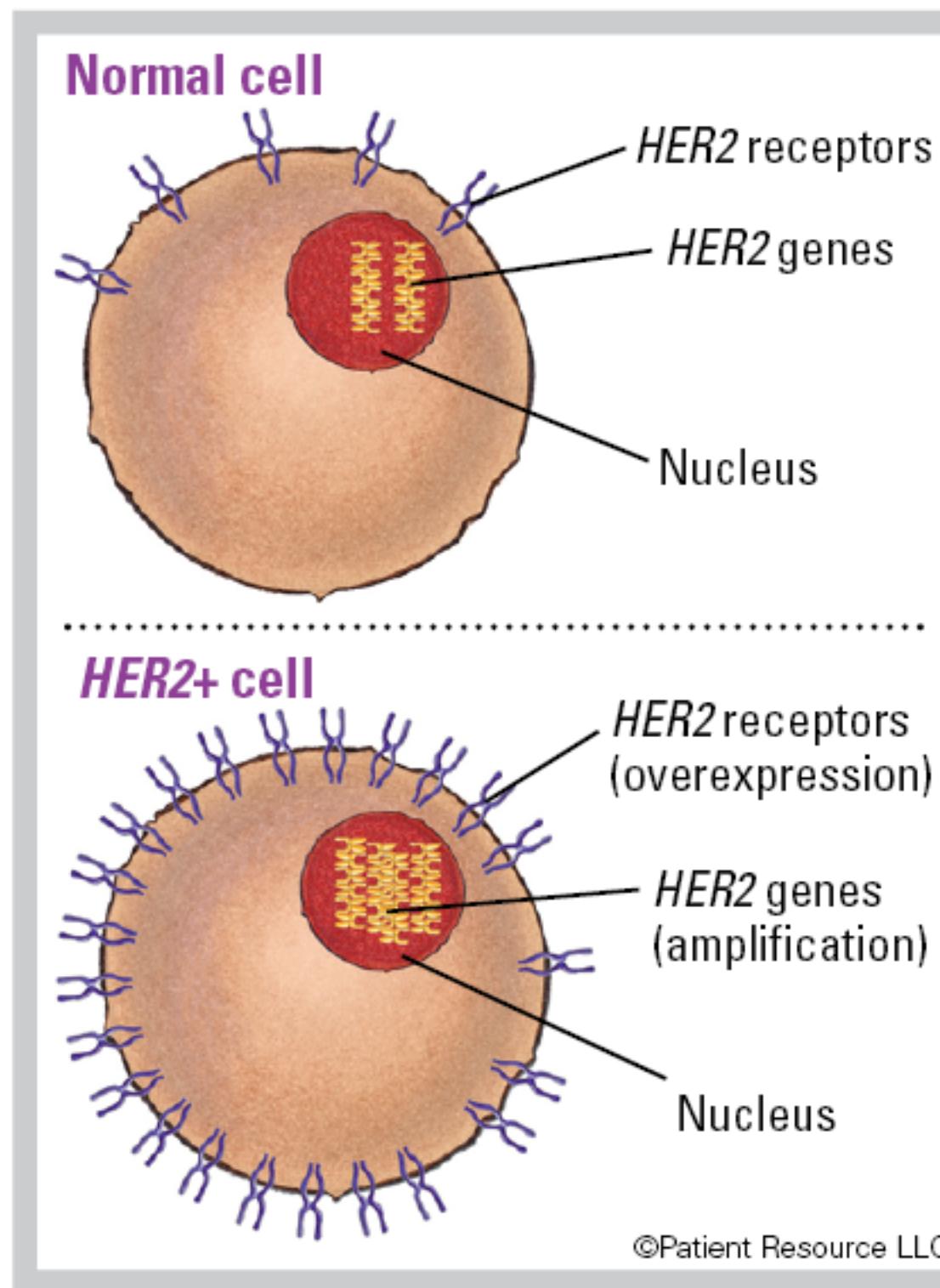
- Intermediate-scale genetic change
- Size: 1kb to multiple megabase
- Additional copies of sequence (**duplications**) and losses of genetic material (**deletions**)



Gain of chromosome arm 13q in colorectal carcinoma

Somatic CNVs related to cancer

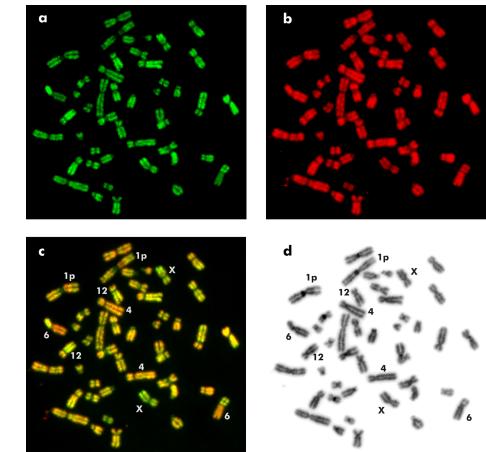
▲ FIGURE 1
BREAST CELLS



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

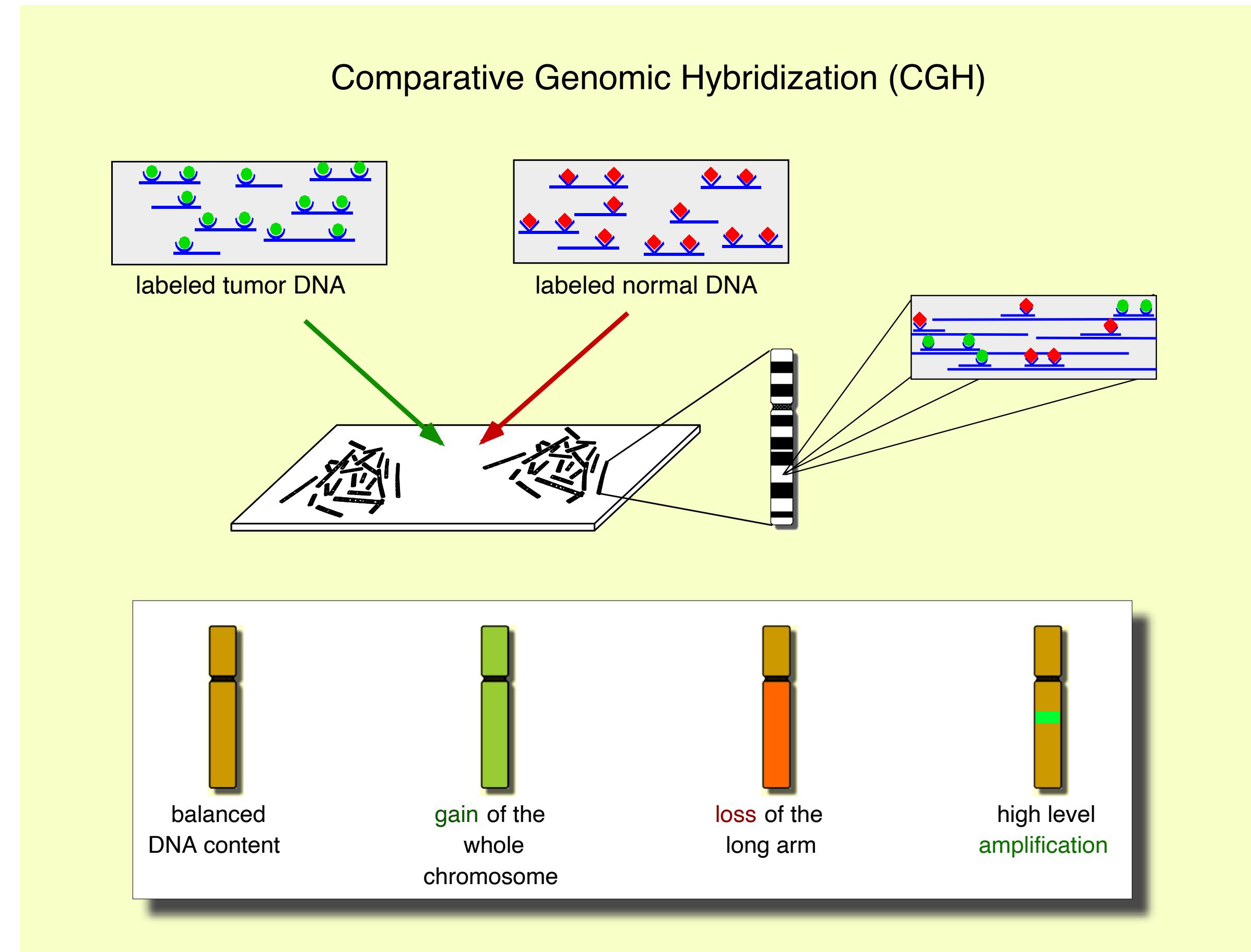
Comparative Genomic Hybridization

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;258(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

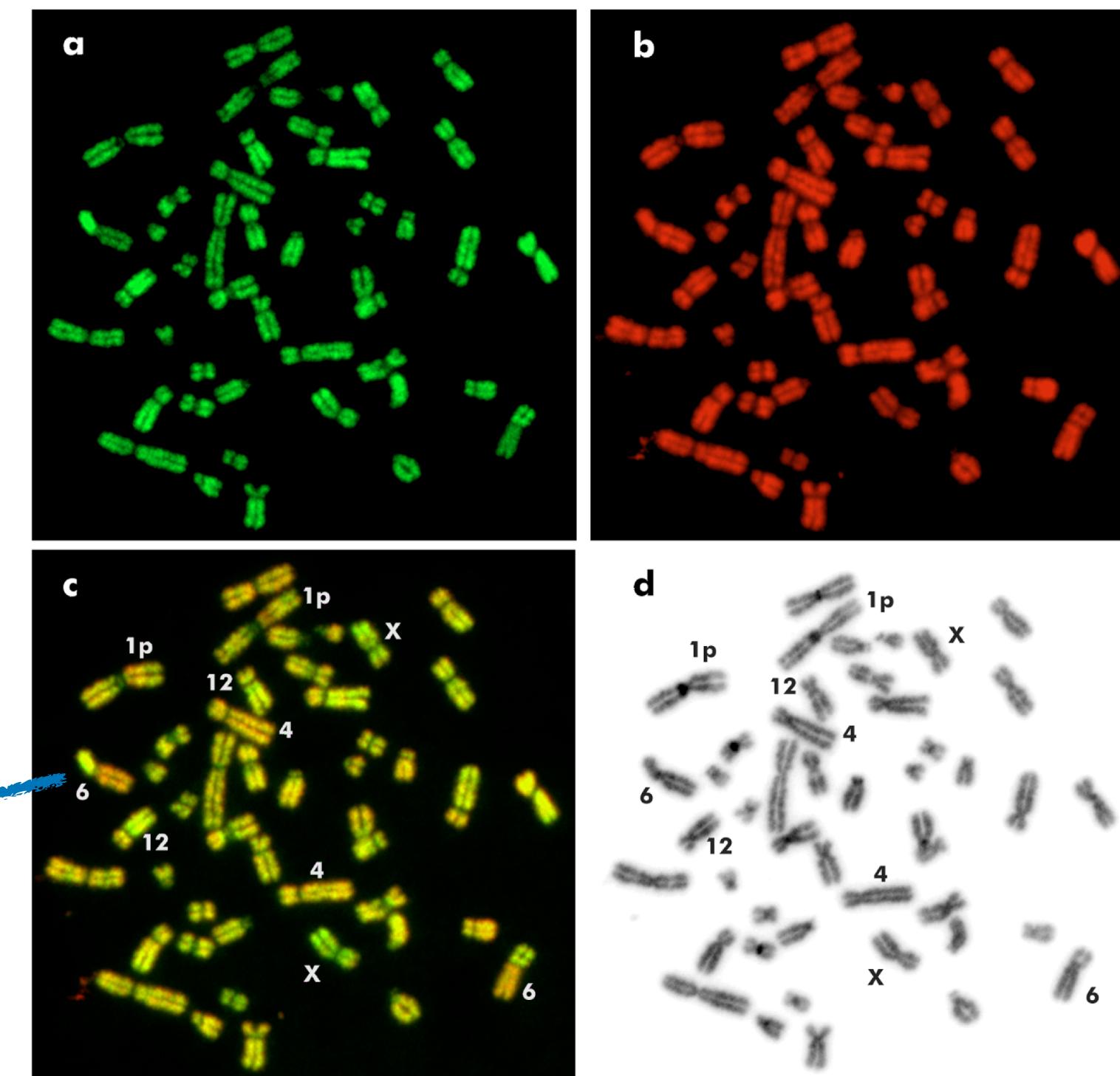
Comparative Genomic Hybridization

Molecular-Cytogenetic Technology for Genomic Imbalance Screening

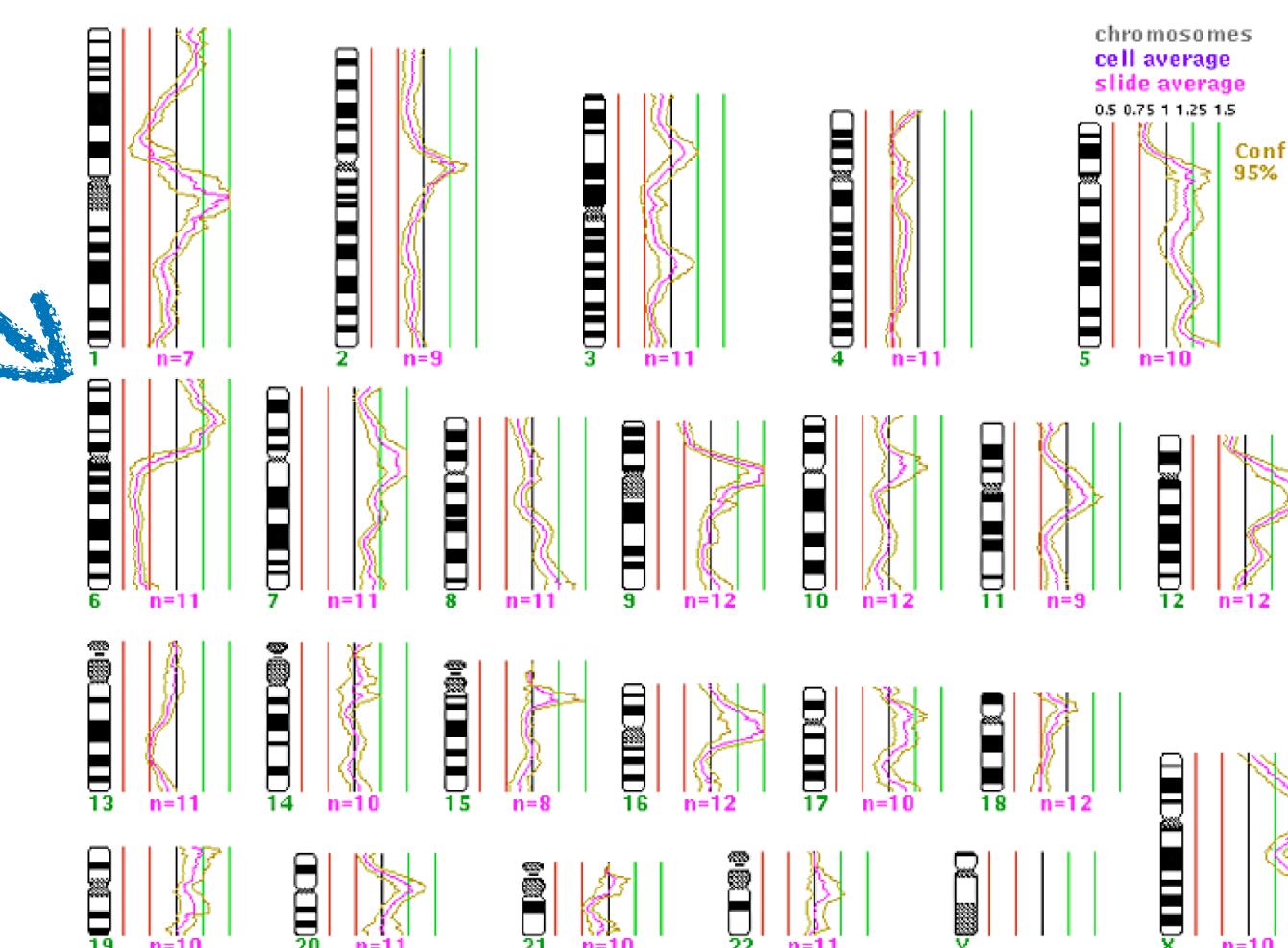
- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression hybridization of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative 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;258(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



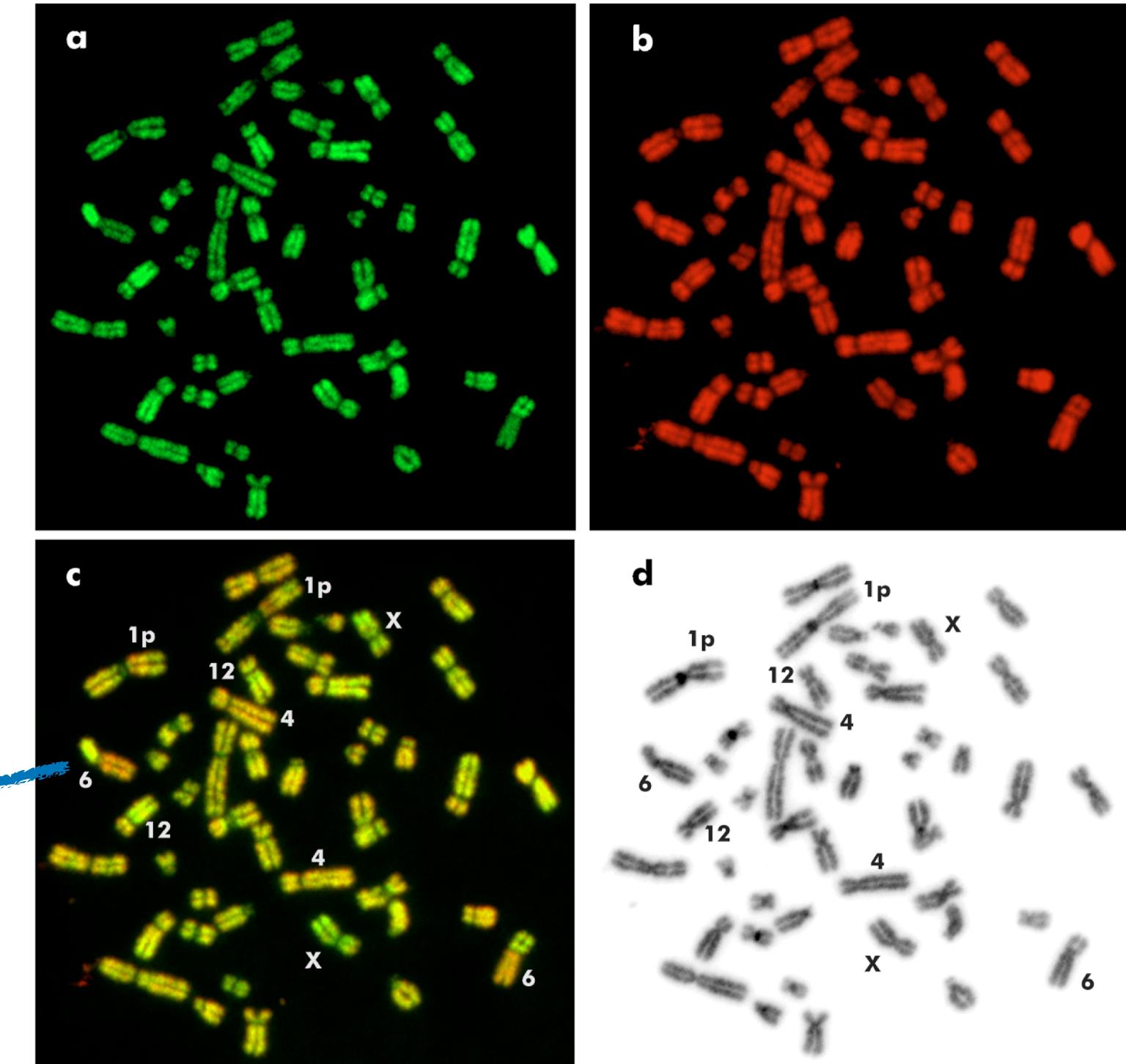
Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

Comparative Genomic Hybridization

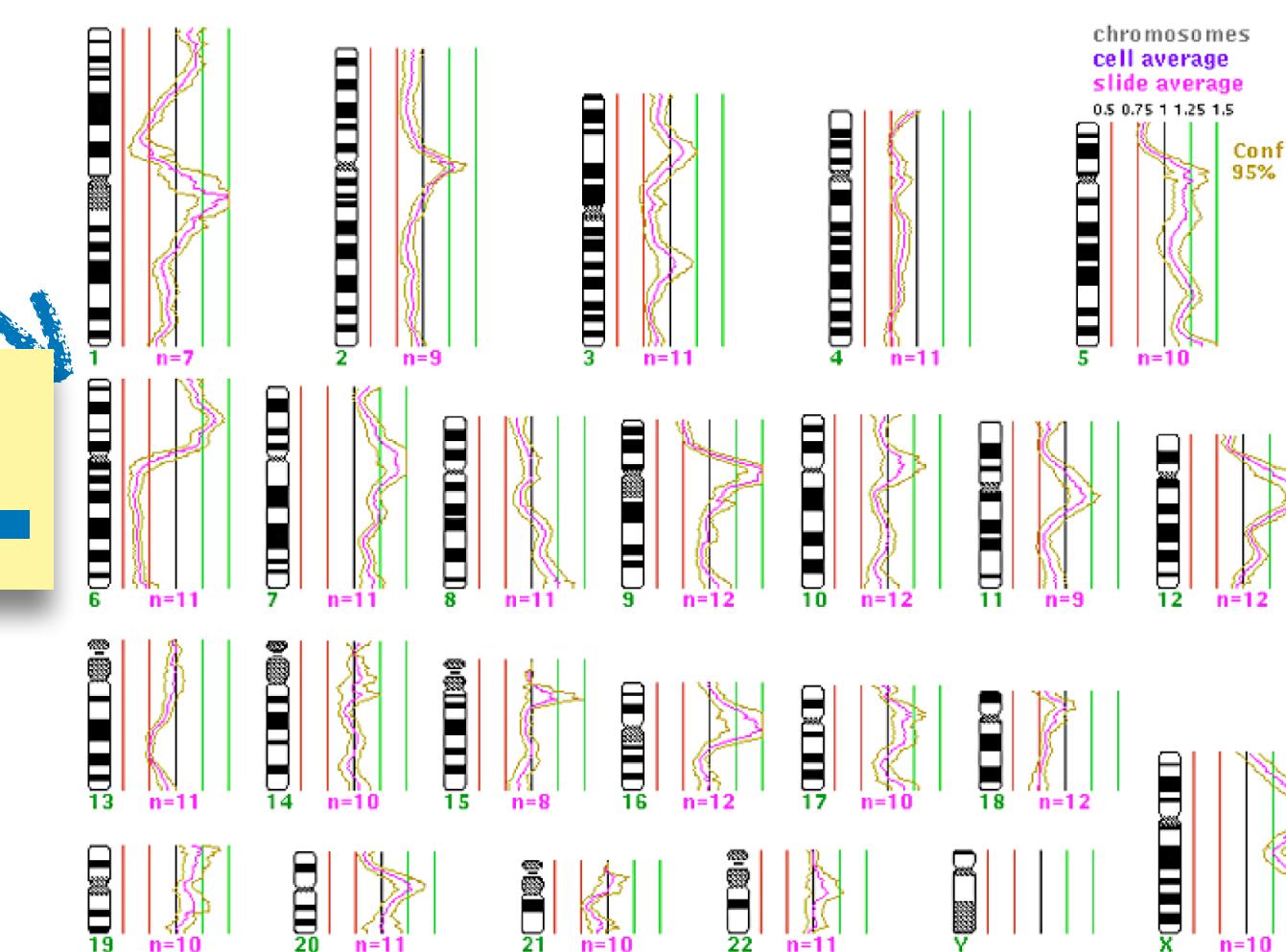
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- 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)

+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

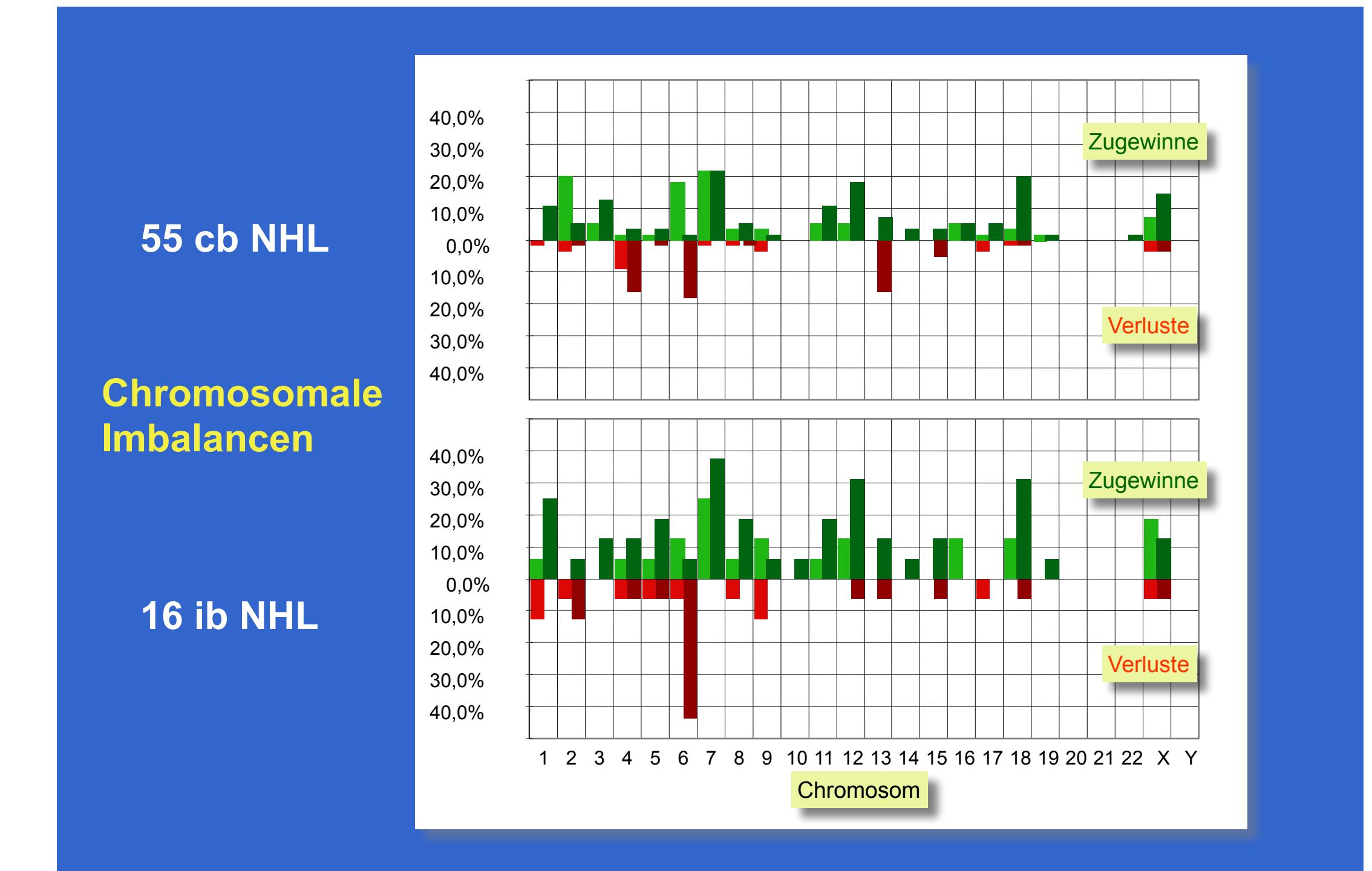
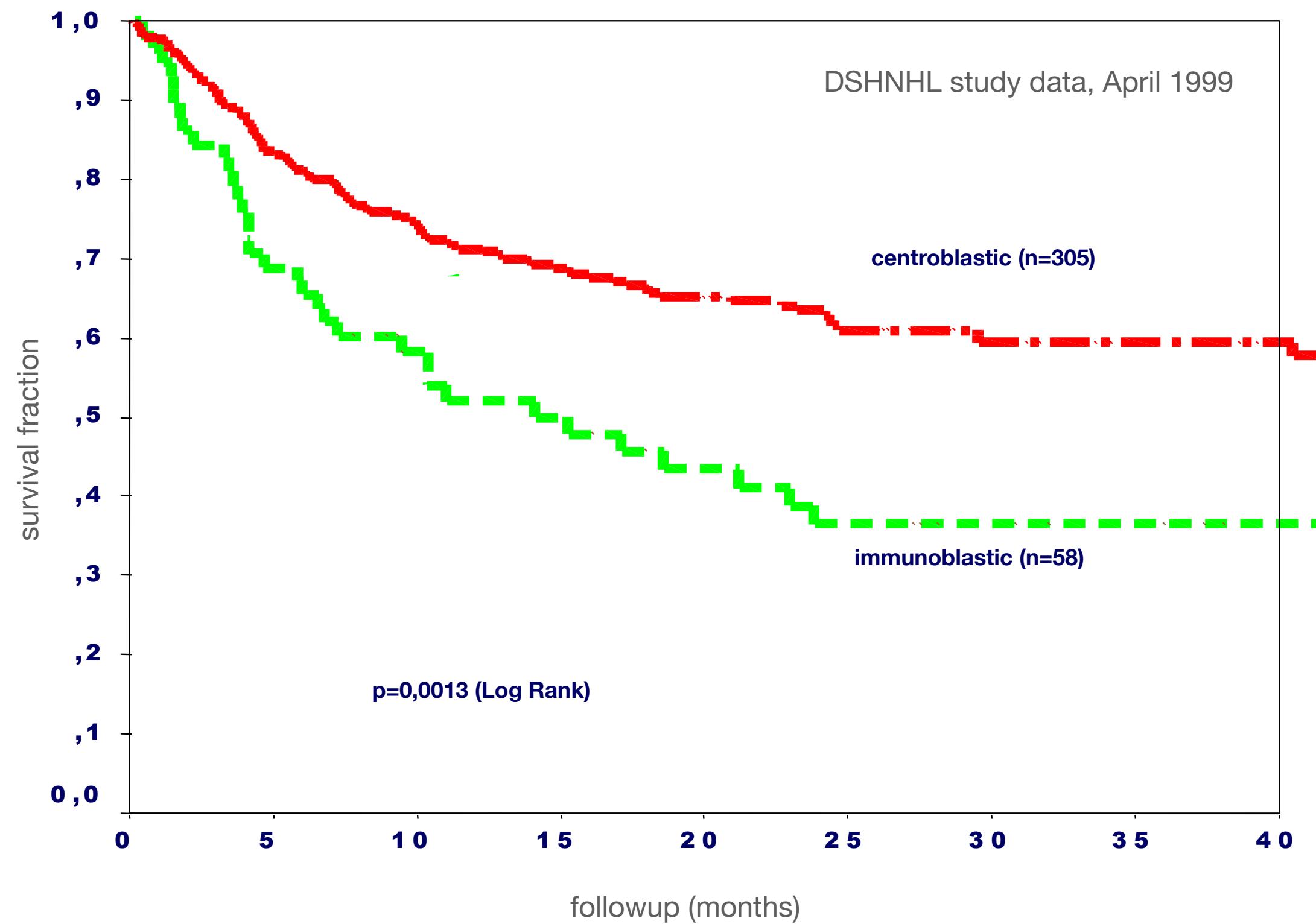


Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

- 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;258:818-821.
- 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.

Cancer CNVs | Diagnostics | Prognosis

Single-study CNV frequencies correspond to diagnostic subsets



Let's build a database!

Progenetix CGH Database and Website

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

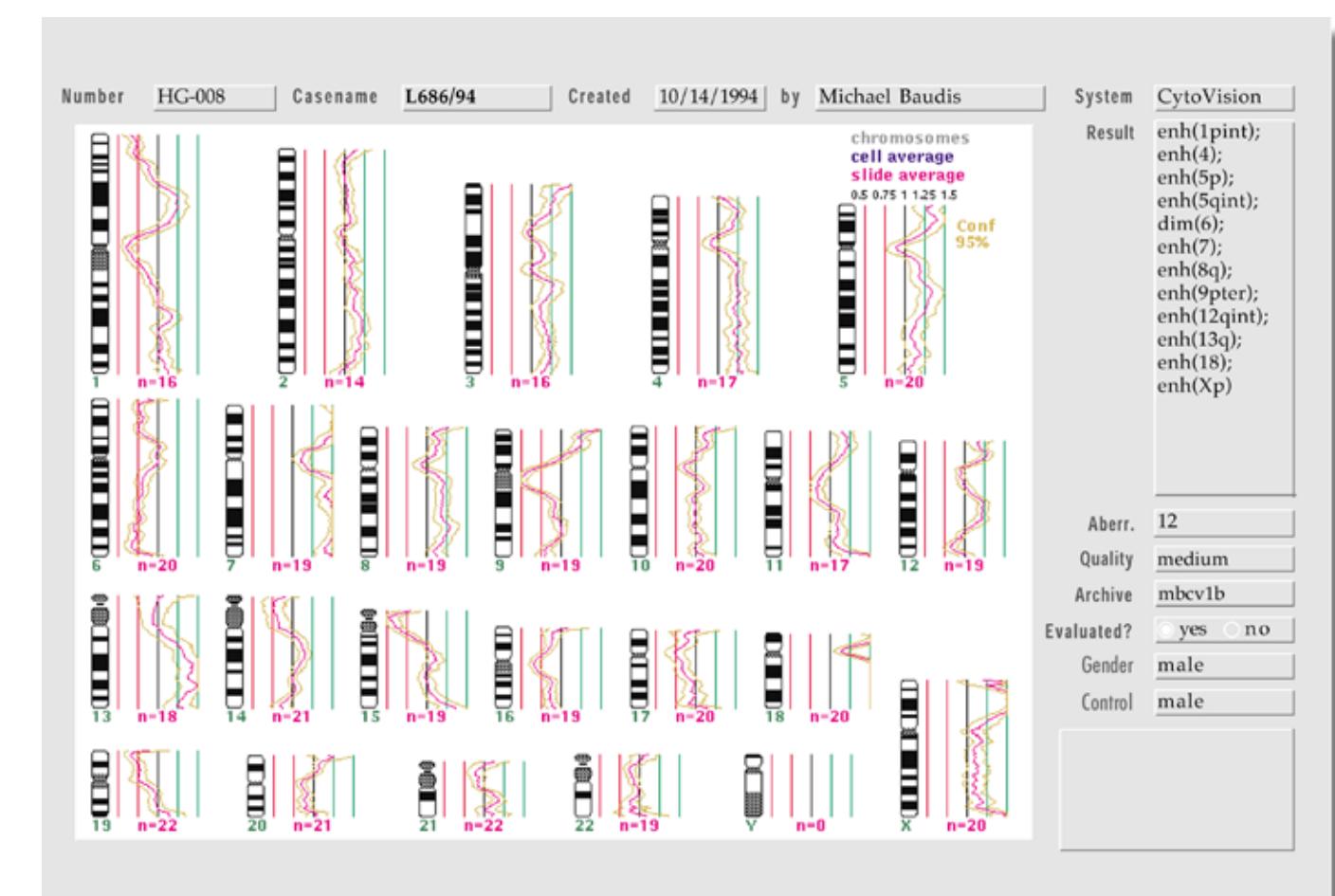
- Dec 6, 2000
 - first time online
- Nov 30, 2000
 - addition of graphical representation and gene table
- Nov 17, 2000
 - generation of website layout and database automatisation

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



Selected will be cases with gain of chromosomal material involving chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, q, included in the project: High Grade N of . Only cases with the histology shall be included. Alternatively, you may select cases which have shown to be for the - translocation.

Only evaluated cases?

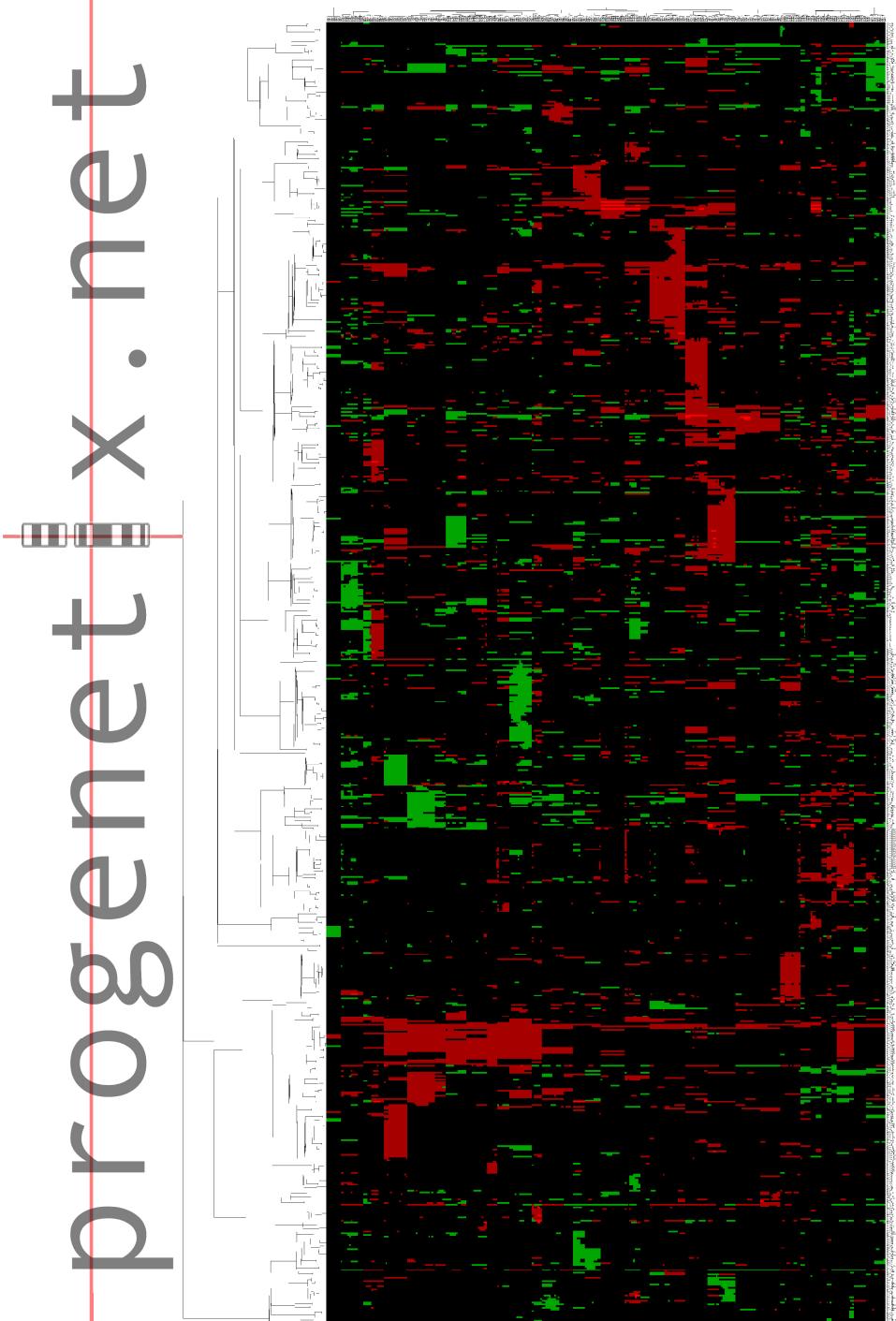


Over the last decade, techniques for the genome wide scanning for genomic imbalances in malignant neoplasia have been developed, e.g. Comparative Genomic Hybridization (CGH).

Currently, no comprehensive online source for CGH data with a standardized format suitable for data mining procedures has been made available for public access. Such a data repository could be valuable in identifying genetic aberration patterns with linkage to specific disease entities, and provide additional information for validating data from large scale expression array experiments.

A case and band specific aberration matrix was selected as most suitable format for the mining of CGH data. The [progenetix.net] data repository was developed to provide the according data to the research community for a growing number of human malignancies.

In the current implementation, two main purposes are being served. First, access to the band specific pattern of chromosomal imbalances allows the instantaneous identification of genomic "hotspots". Second, the band specific aberration matrices can be included in data mining efforts. As an example, the clustering off all informative cases from the current (September 2001) dataset is shown here (online source under www.progenetix.net/bcats/clustered.png).



Data selection

PubMed is searched for publications applying CGH to the analysis of malignant tumors. Articles are selected according to their online availability and the description of genomic imbalances on a per case basis.

Transformation of input data

Chromosomal aberration data is transformed via customized parsing commands to a common format adherent to ISCN 1995 recommendations. In some cases, aberration data was transcribed from graphical representations or provided by the authors.

Data storage

Currently, the primary data is stored in a dedicated "off-line" database. Besides case identifier and ISCN adapted chromosomal imbalance data, tumor classification and source information including the PubMed identifier is recorded. Disease entities are reclassified to ICD-O-3 codes.

Text parsing and generation of aberration matrix

For the generation of the case and band specific aberration matrix, a dedicated text pattern comparison model was developed using Perl. Briefly, for each chromosomal band, the aberration field of each case is searched for a variety of patterns containing aberration information applying to that band. A matrix with currently 324 band resolution is generated, annotating chromosomal gains with "1" and losses with "-1"; localized high-level gains are designated "2".

Website generation

For graphical representation of chromosomal imbalances, HTML pages containing different views of the underlying aberration matrices are generated using Perl. Graphics are implemented using HTML syntax. Besides band specific, whole genomic overviews, chromosome specific pages with links to all involved cases are generated for each ICD-O-3 entity as well as for each registered project. Additionally, those representations are available for several subsets combining related data (e.g. all lymphoid neoplasias, breast carcinoma cases). For each of the groups, the according aberration matrix is linked for download.

Hierarchical clustering of band specific chromosomal imbalances from 999 human neoplasias, contained in the [progenetix.net] collection. Cases without aberrations were excluded.

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

Michael Baudis^{1,2,*} and Michael L. Cleary²

¹Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany

²Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA

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



ABSTRACT

Summary: Through sequencing projects and, more recently, array-based expression analysis experiments, a wealth of genetic data has become accessible via online resources. In contrast, few of the (molecular-) cytogenetic aberration data collected in the last decades are available in a format suitable for data mining procedures. www.progenetix.net is a new online repository for previously published chromosomal aberration data, allowing the addition of band-specific information about chromosomal imbalances to oncologic data analysis efforts.

Availability: <http://www.progenetix.net>

Contact: mbaudis@stanford.edu

Neoplastic transformation and progression is the result of genetic defects arising in normal cells and giving rise to a malignant clone. During the process of oncogenesis, some of the usually multiple steps required for acquisition of the full neoplastic phenotype may represent themselves as numerical or structural abnormalities in the chromosomes of the transformed cells.

Over the last decades, the analysis of chromosomal abnormalities in malignant cells has gained importance in oncologic research as well as in clinical practice. A vast number of genetic abnormalities has been identified in the virtually complete range of human neoplasias. Several attempts have been undertaken for collection and classification of those abnormalities, the most widely recognized being the catalog by Mitelman and co-workers (Mitelman, 1994; online access through <http://cgap.nci.nih.gov/Chromosomes/Mitelman>).

In addition to metaphase analysis of short-term cultivated tumor cells or tumor cell lines, molecular cytogenetic techniques have recently been applied to the analysis of chromosomal abnormalities in primary tumor tissues. One of the more widely used screening techniques is Comparative Genomic Hybridization (CGH; Kallion-

iem et al., 1992; du Manoir et al., 1993). Briefly, this method is based on the competitive *in-situ* hybridization of differentially labeled tumor versus normal genomic DNA to normal human metaphase spreads. The calculation of the intensity ratios of the two fluorochromes gives an overview about relative gains and losses of DNA in the tumor genome with mapping to the respective chromosomal bands. The identification of frequently imbalanced regions in tumor entities may point towards tumor suppressor gene or proto-oncogenes mapping to the respective chromosomal bands. Usually, the result of those experiments is communicated either in text format according to the International System for Cytogenetic Nomenclature (Mitelman, 1995) or graphically, with aberration bars next to chromosomal ideograms for the representation of chromosomal gains and losses.

Because in each experiment CGH analysis covers the whole number of chromosomes, the comparision of data sets from related malignancies could lead to the delineation of common as well as divergent genetic pathways defining the respective malignant phenotypes. Although an extremely large number of malignant tumors has been analyzed using this technique, no comprehensive CGH database with band-specific chromosomal aberration information is publicly available[†].

A minimal requirement for such a database would be the conversion of the text or graphical information used in publications to data tables, representing the information about the aberration status of single chromosomal bands for each case. For the site discussed here, this process includes: (1) the transformation of the published results in a format adapted from the ISCN, and (2) the automatic generation of the band specific aberration table.

Due to format variations of the published data, step 1 consists of the manual conversion of the text data or evaluation and conversion of the graphical representations, respectively. Due to the (in computational terms) odd

[†]Links to a number of online CGH resources with different scopes can be found at www.progenetix.net.

*To whom correspondence should be addressed.

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

[ideogram] [casetable] [clustering] [download source]

About [progenetix]

Contents, Aims and FAQs

Publications

ICD-O Entities

Site Codes and Misc. Groups

ISCN2matrix Converter

Data Source Access

Sponsors and Contributors

News and History

Links

PLOS

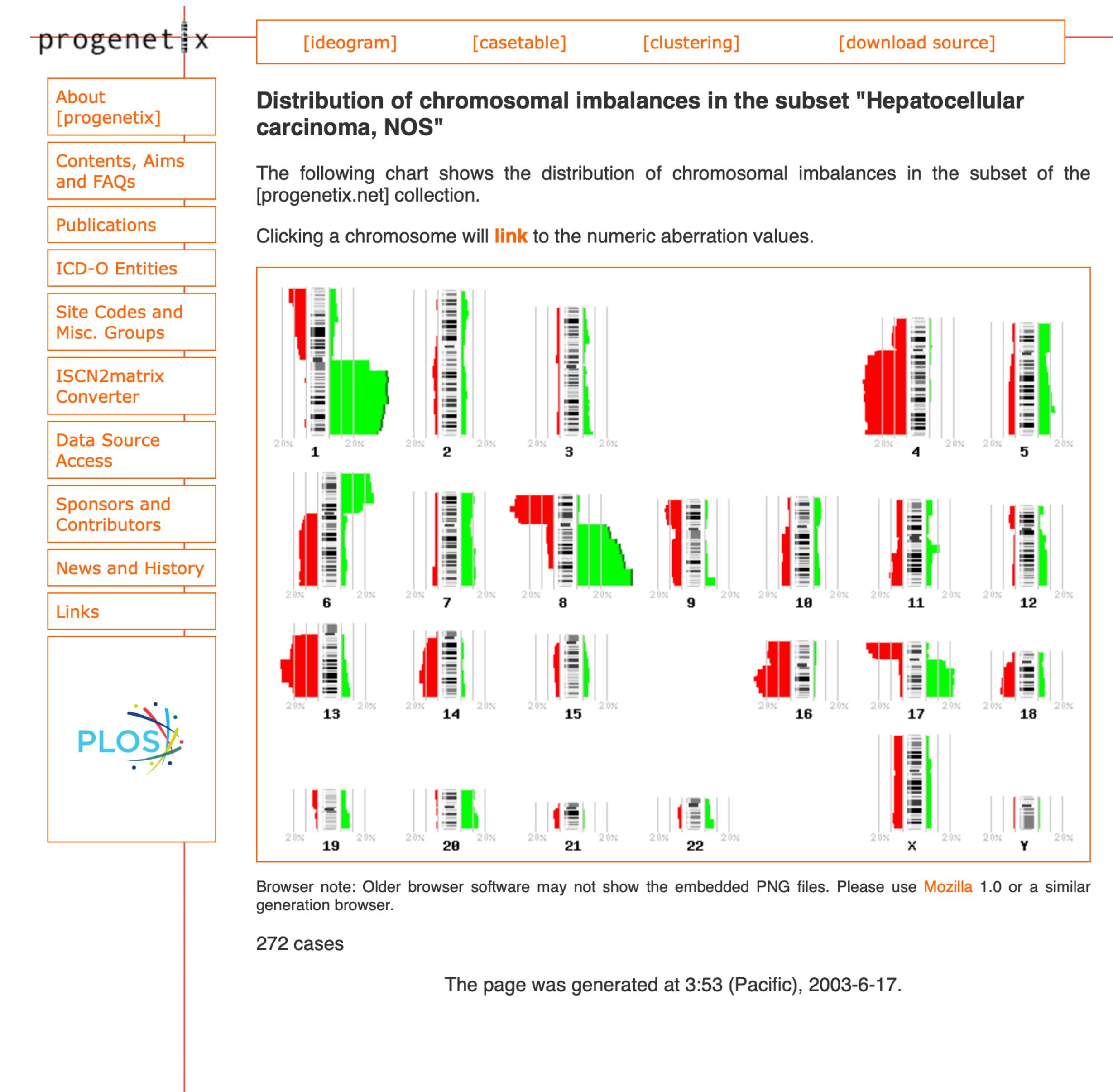
List of cases included in the subset "Hepatocellular carcinoma, NOS"

Casename	Original diagnosis	PUBMED ID	Aberrations (by CGH)
HCC-vir-dys-ca-01sat	Hepatocellular carcinoma (HBV, satellite tumor)	12666986	rev ish enh(1q21qter, 7p11.2pter, 7q11.2q31, 8q13qter, 9p22pter, 10, 11p11.2p12, 11q12qter, 15q26) dim(1p22pter, 2q32qter, 4, 5, 7q32qter, 8p12pter, 14q21qter, 15q11.2q21, 16, 17p11.2pter, 17q11.2q21, 18, 19)
HCC-vir-dys-ca-01tu	Hepatocellular carcinoma (HBV)	12666986	rev ish enh(1q21qter, 5p12pter, 8q12qter, 9p21pter, 11q12qter, 20) dim(1p31pter, 4, 7q32qter, 8p12pter, 14q21qter, 16, 17p12pter, 18, X)
HCC-vir-dys-ca-02tu	Hepatocellular carcinoma (HCV)	12666986	rev ish enh(1q21q43, 6q12q14, 7, 8p11.2, 8p21p23, 8q11.2q13, 8q23, 10p11.2p13, 10q11.2qter, 17q11.2q24, Xq13qter) dim(11, 14q31, 15q11.2q21, 16p12pter, 17p11.2pter, 19p13.1pter, 19q13.1q13.2, Xp21)
HCC-MF-01T1	Hepatocellular carcinoma	12579536	rev ish enh(16q13qter)
HCC-MF-01T2	Hepatocellular carcinoma	12579536	rev ish enh(12q22qter, 17q) dim(16q)
HCC-MF-01T3	Hepatocellular carcinoma	12579536	rev ish enh(12q21.3qter, 17q21qter) dim(16q21qter)
HCC-MF-02T1	Hepatocellular carcinoma	12579536	rev ish dim(6q13qter)
HCC-MF-02T2	Hepatocellular carcinoma	12579536	rev ish enh(1q, 17q) dim(17p)
HCC-MF-03T1	Hepatocellular carcinoma	12579536	rev ish enh(1q, 3q26.2qter, 4p, 6p21.1pter, 11p15, 19q) dim(16q10q12.2)
HCC-MF-03T2	Hepatocellular carcinoma	12579536	rev ish enh(8q, 11p15, 12pterq12) dim(3p, 4q, 5q, 8p23.1, 9q, 16q) amp(1q)
HCC-MF-04T1	Hepatocellular carcinoma	12579536	rev ish enh(1p33qter, 8q21.2qter) dim(1pterp34, 4q, 9q) amp(6p, 13q21qter)
HCC-MF-04T2	Hepatocellular carcinoma	12579536	rev ish enh(1q, 5q31.3qter, 8q) dim(6q, 16, 17pterq21) rev ish enh(6q, 8q, 10p, 12q21.1qter, 13q22qter, 17q, 18p) dim(4p15qter, 5, 7p21qter, 7q, 9p, 9q10q34.2, 11q, 16q) amp(10p)
HCC-MF-05T1	Hepatocellular carcinoma	12579536	rev ish enh(6q, 8q12qter, 12q21.1qter, 13q22qter, 17q, 18p) dim(4p15qter, 5, 7p21qter, 7q, 9q10q34.2, 11q, 16q) amp(10p)
HCC-MF-05T2	Hepatocellular carcinoma	12579536	rev ish enh(6q, 8q12qter, 12q21.1qter, 13q22qter, 17q) dim(4q, 5q, 7p, 7q, 9q10q31, 11q, 14q, 16q) amp(10p)
HCC-MF-06T1	Hepatocellular carcinoma	12579536	rev ish enh(1q, 5p23pter, 18p, 22) dim(4q, 6q, 9pterq33, 13q, 14q, 16pterq23) amp(8q)

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 revised 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-pter, 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

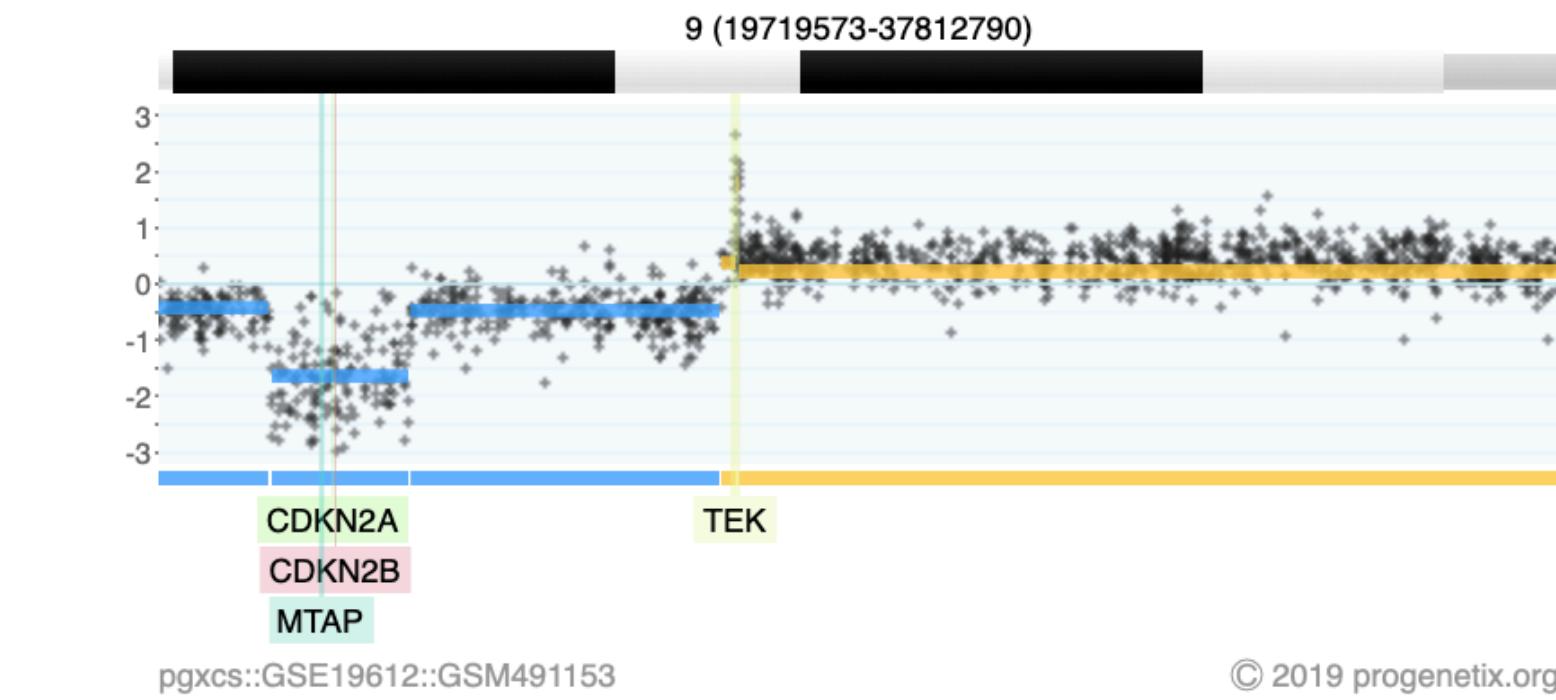
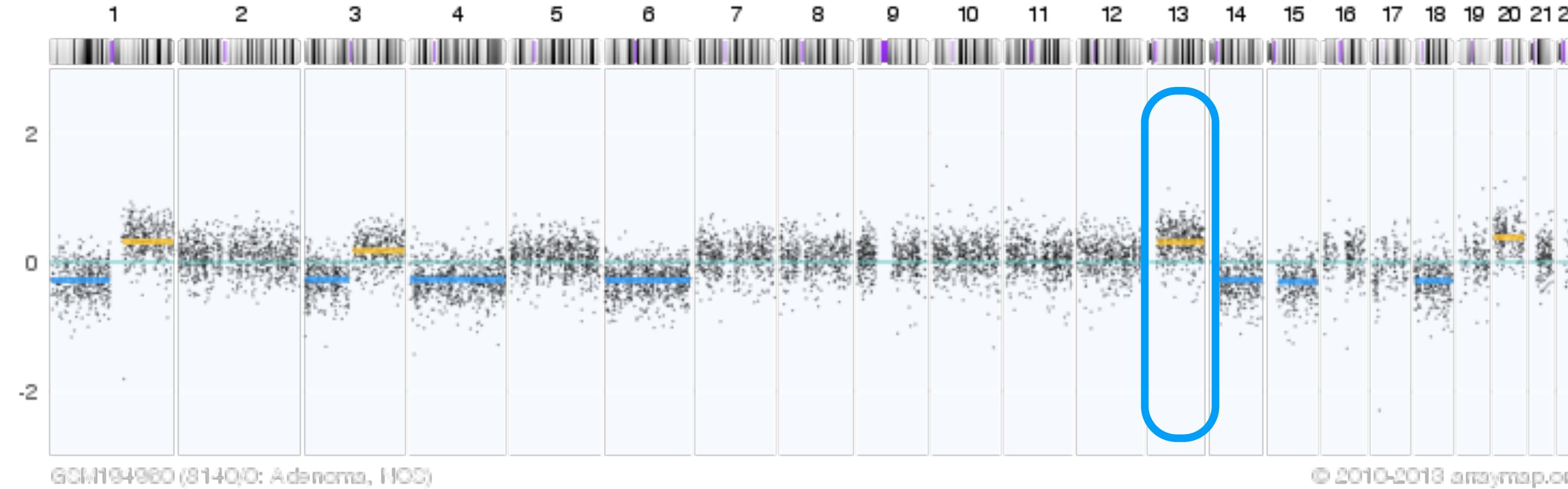
^aAJCC/UICC staging system (Hutter and Sabin, 1986).^bGrade of primary tumor: 1–3, low, moderate, high grade; 9, grading unknown.^cSynchronous, diagnosis of metastatic disease within 12 months following diagnosis of primary tumor; metachronous, diagnosis of metastatic disease after 12 months or later.

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

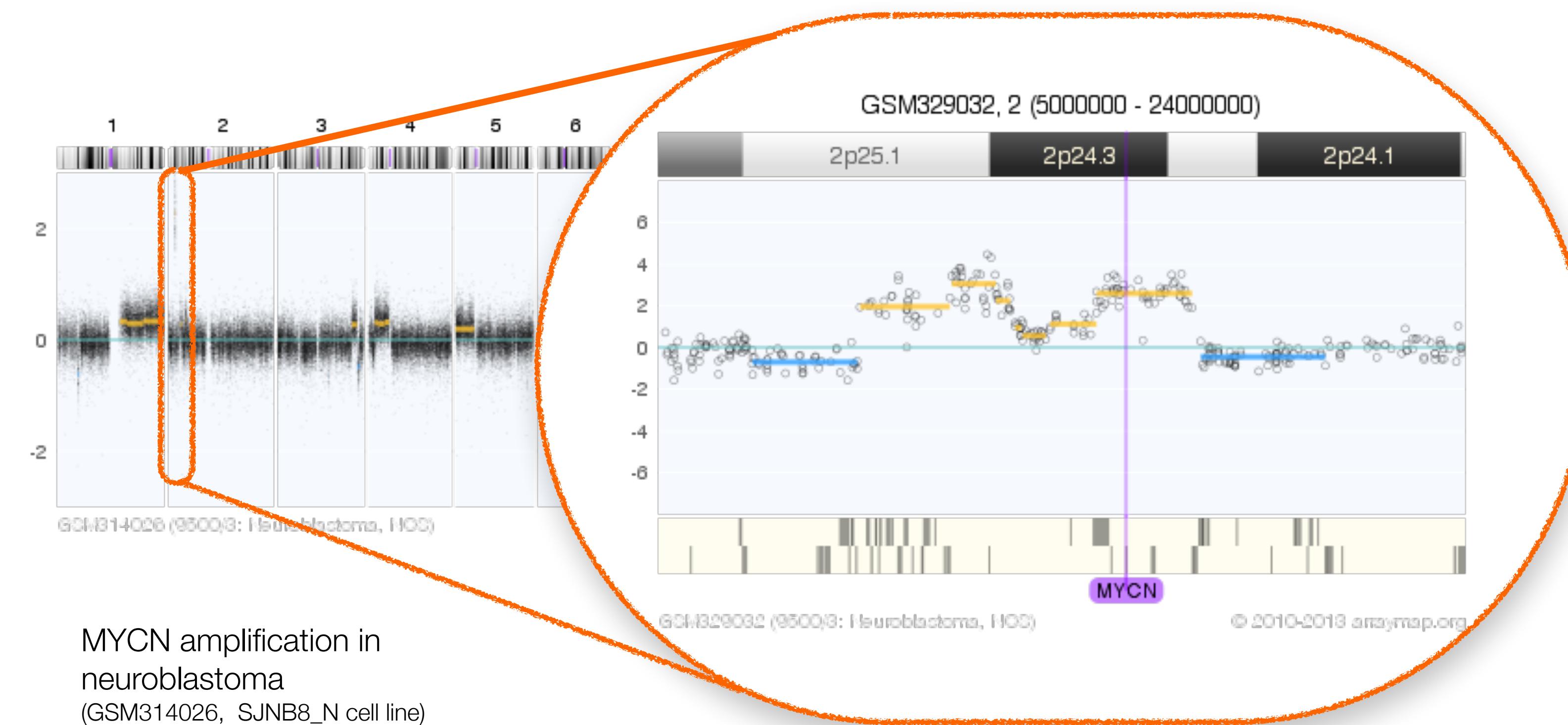
W. Michael Korn,¹ Toru Yasutake,² Wen-Lin Kuo,¹ Robert S. Warren,³ Colin Collins,¹ Masao Tomita,² Joe Gray,¹ and Frederic M. Waidman¹

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

Array-based Detection of Copy Number Variations



2-event, homozygous deletion in a Glioblastoma



low level/high level copy number alterations (CNAs)

arrayMap



arrayMap (2012 - 2020)

Probe-Level Genomic Array Data in Cancer



Search Samples

Search Publications

Progenetix



Citation & Licensing

User Guide

People

Beacon⁺

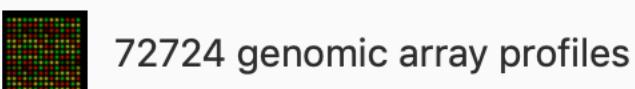


162.158.150.56

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



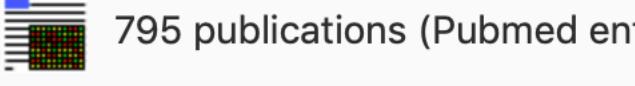
898 experimental series



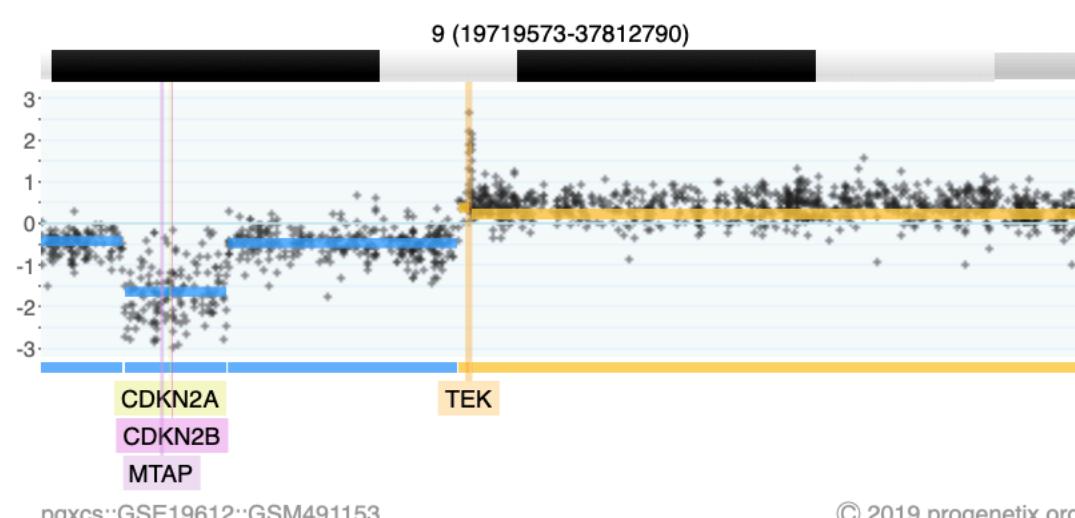
257 array platforms



341 ICD-O cancer entities



795 publications (Pubmed entries)



© 2019 progenetix.org

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.

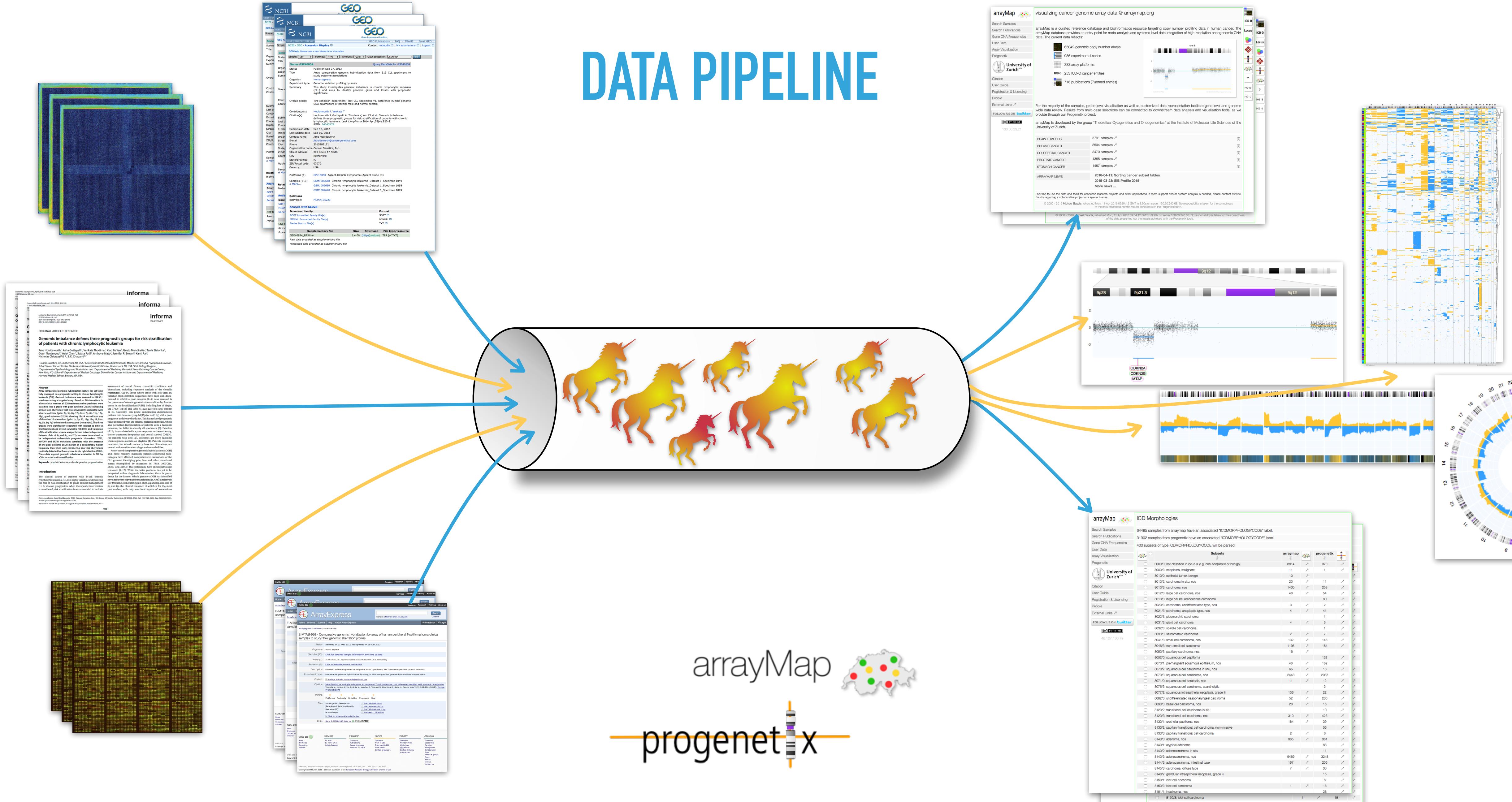
© 2000 - 2019 Michael Baudis, refreshed 2019-06-12T21:00:19Z in 6.00s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.



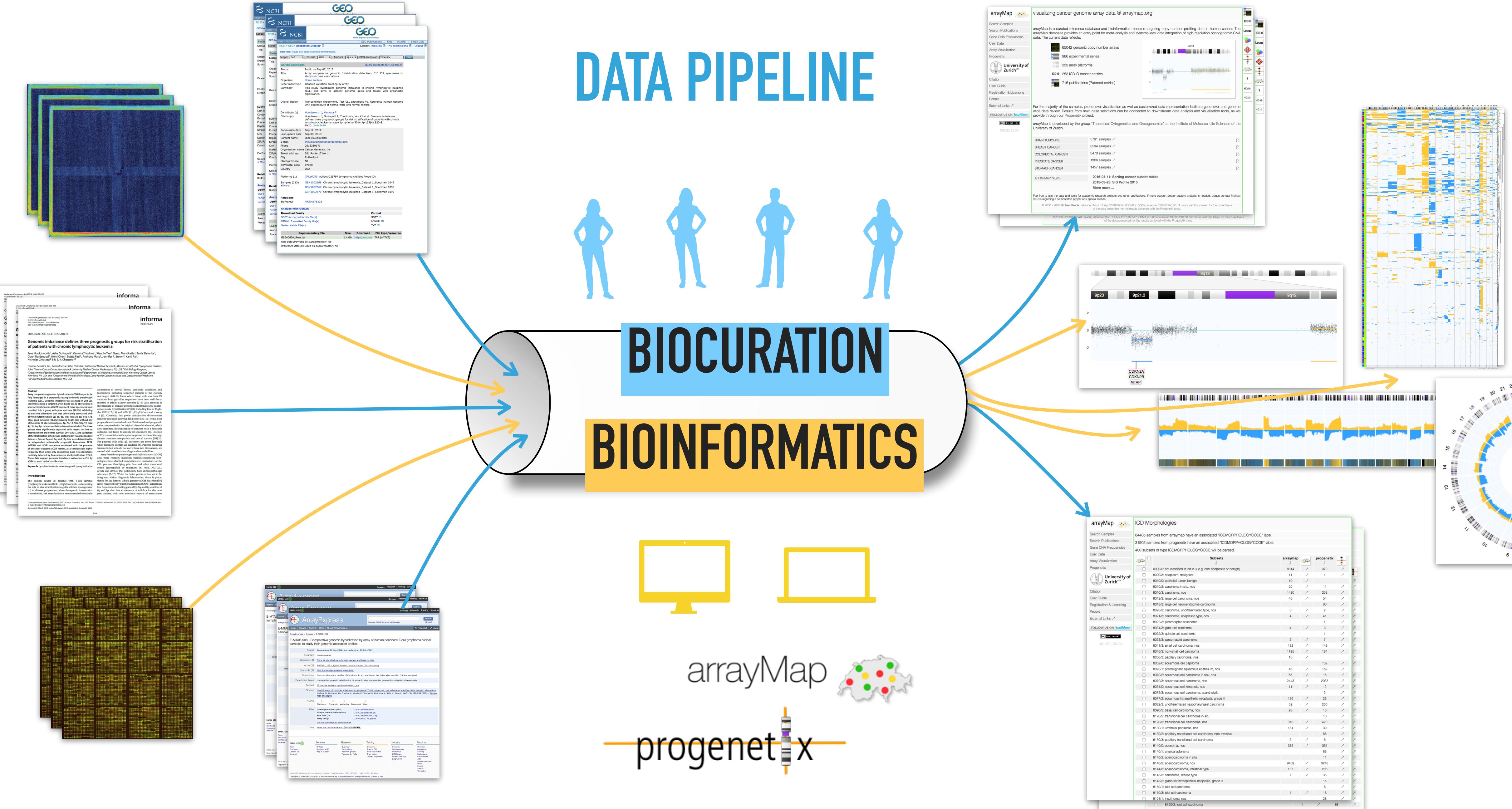
arrayMap



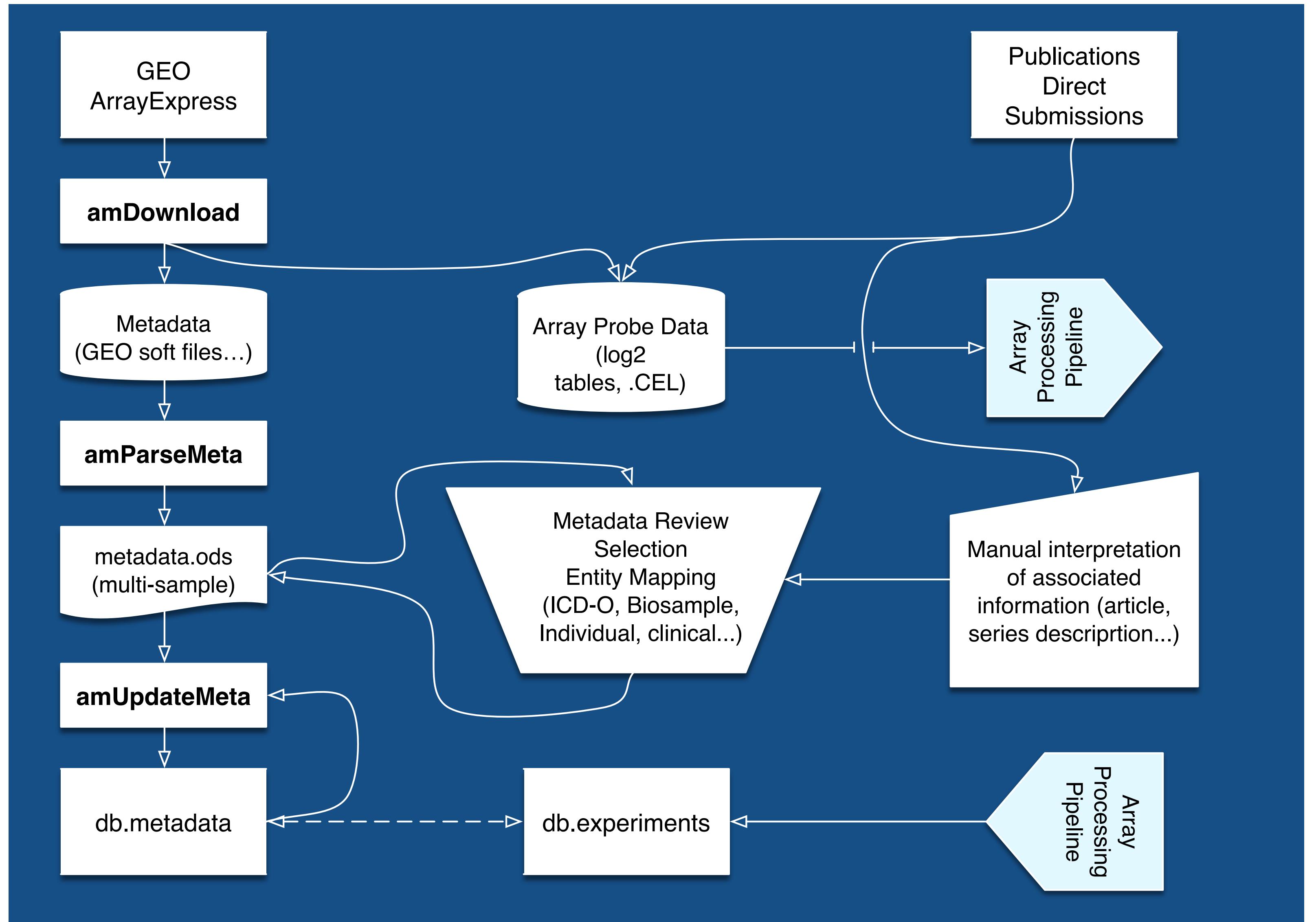
DATA PIPELINE



DATA PIPELINE



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



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
```

Regular Expressions!



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

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

survival: alive

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

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

Channel 1 is normal -> Cave value swap!

Gender or "chromosomal sex"?

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

Not yet registered way to express "alive"!

```
$mkey->{ samplekey } = 'DEATH';
$mkey->{ matches } = [
    'death',
    'dead ',
    'vital_status',
    'dead_alive',
    'alive_dead',
];
(
    $mkey->{ retv }, $mkey->{ retk } ) = _grepmeta( $mkey, $meta );
if ( $mkey->{ retv } =~ /^(.+?)$/ ) {
    $sample->{ $mkey->{ samplekey } } = _normDeath($mkey->{ retv }) }
```

Disease annotations in Progenetix

From some text, somewhere, to ontology classes

 NCIthesaurus

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

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



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

	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

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

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

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

Standardized Data

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

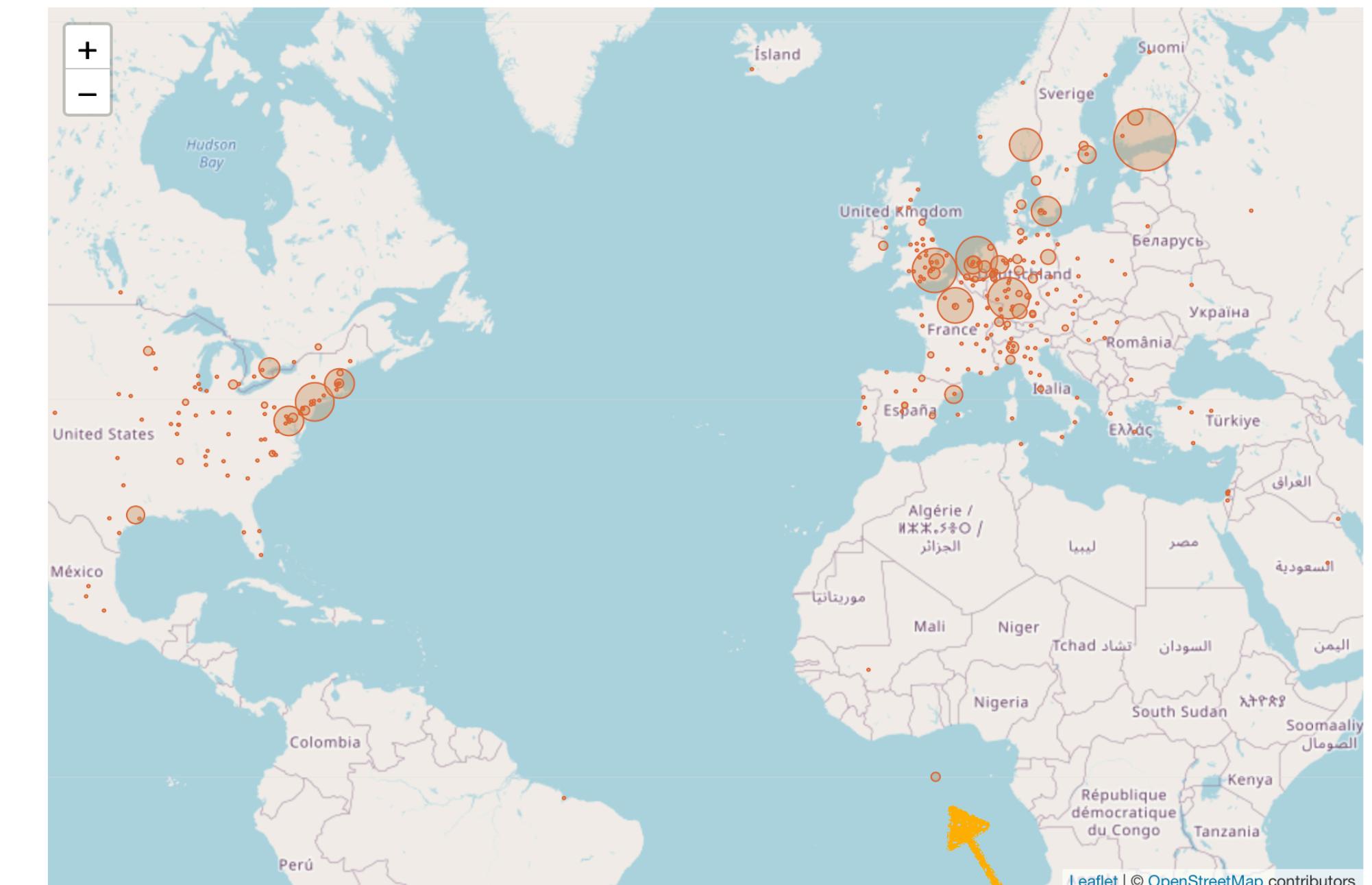
- International initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of **data annotation standards**
- established principles are the use of **hierarchical coding systems** where individual codes are represented as **CURIEs**
- other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

```
"data_use_conditions" : {  
    "label" : "no restriction",  
    "id" : "DUO:0000004"  
},  
  
"provenance" : {  
    "material" : {  
        "type" : {  
            "id" : "EFO:0009656",  
            "label" : "neoplastic sample"  
        }  
    },  
    "geo" : {  
        "label" : "Zurich, Switzerland",  
        "precision" : "city",  
        "city" : "Zurich",  
        "country" : "Switzerland",  
        "latitude" : 47.37,  
        "longitude" : 8.55,  
        "geojson" : {  
            "type" : "Point",  
            "coordinates" : [  
                8.55,  
                47.37  
            ]  
        },  
        "ISO-3166-alpha3" : "CHE"  
    },  
    {  
        "age": "P25Y3M2D"  
    }  
}
```

Data Curation

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

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



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

Data Curation - Geolocations

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

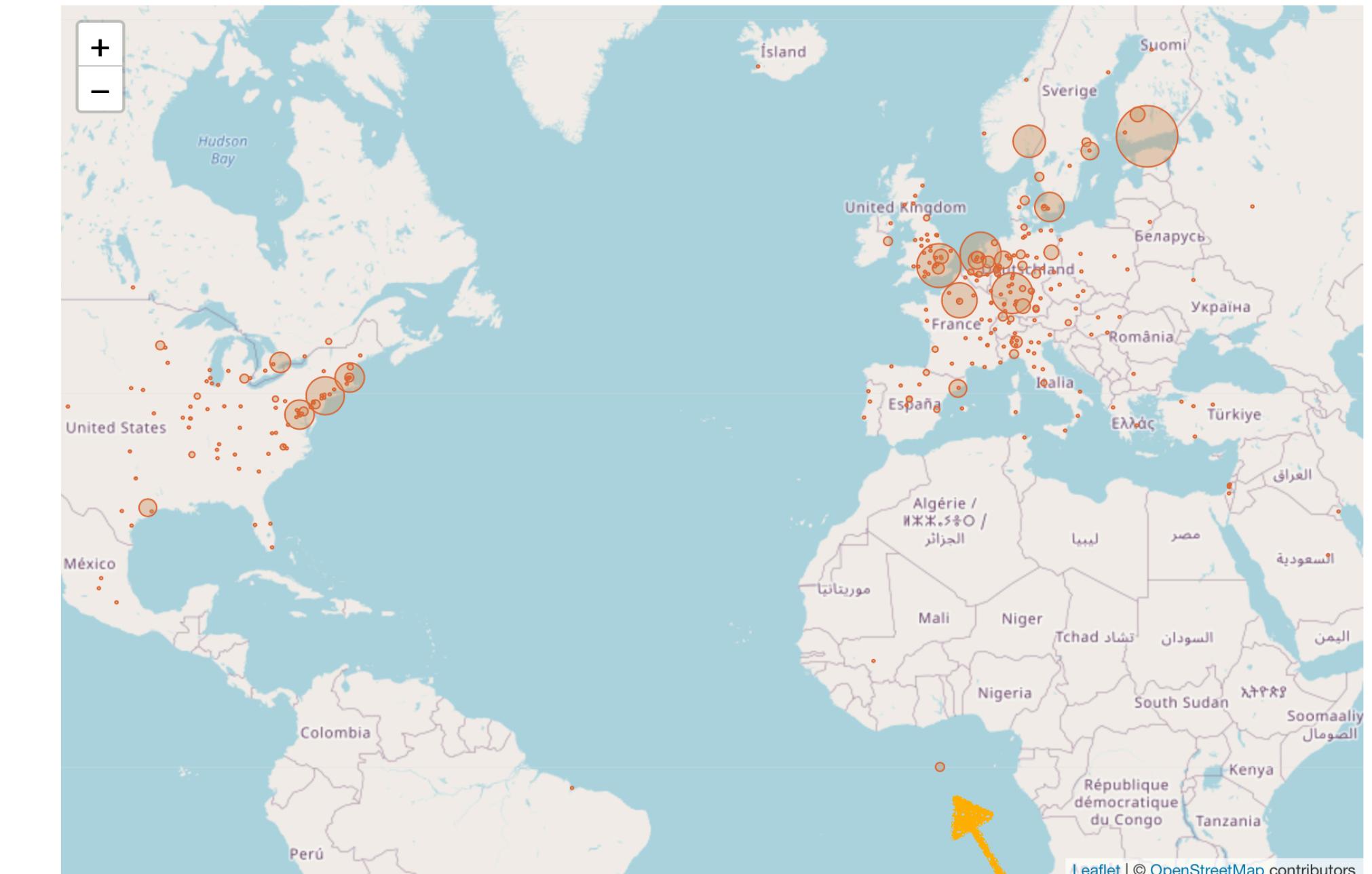
The most geo-tagged place on earth is Null Island



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

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

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



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

25 / 3306 publications

Data Curation - Geolocations

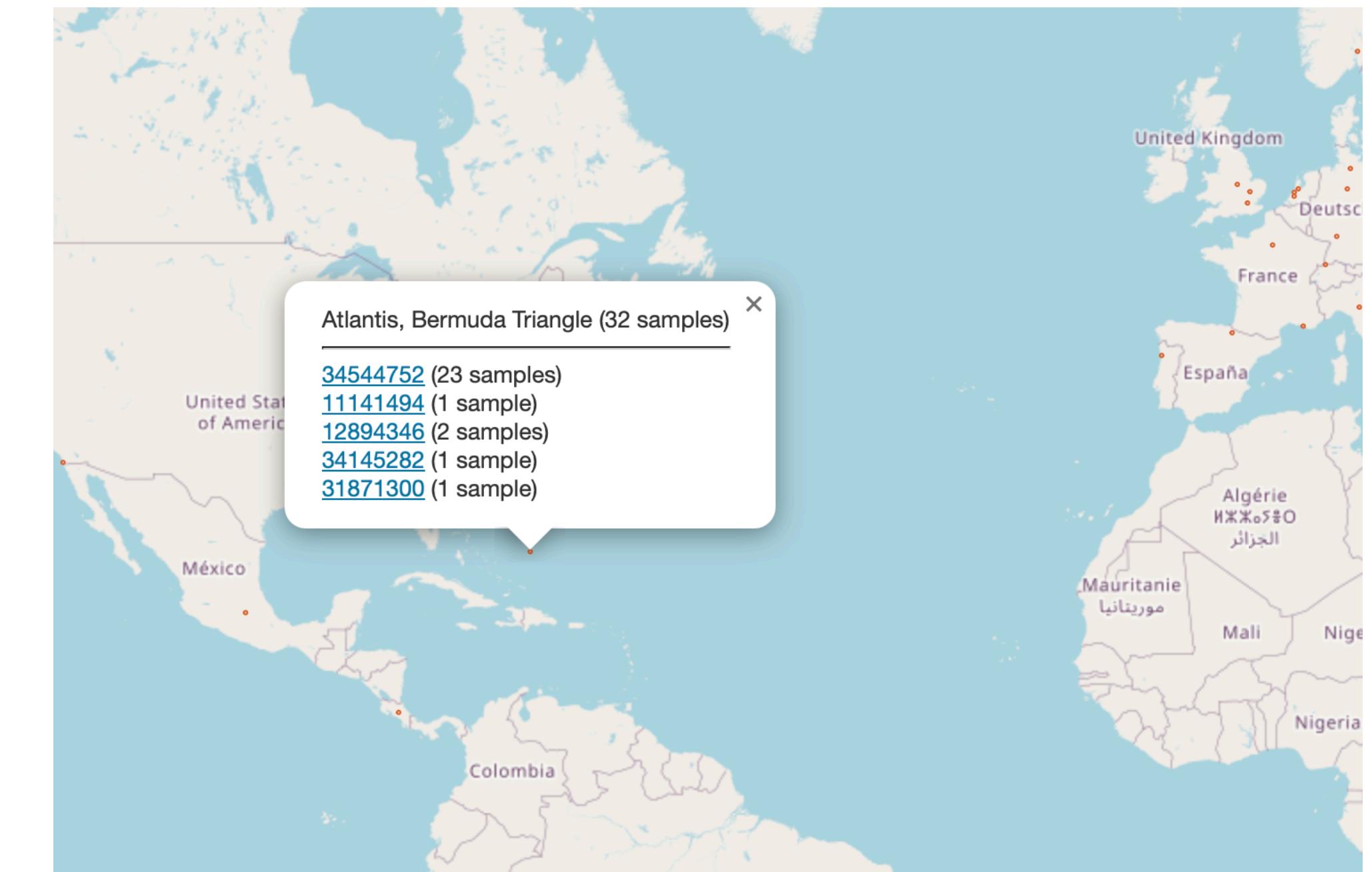
Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

The most geo-tagged place on earth is Null Island



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

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



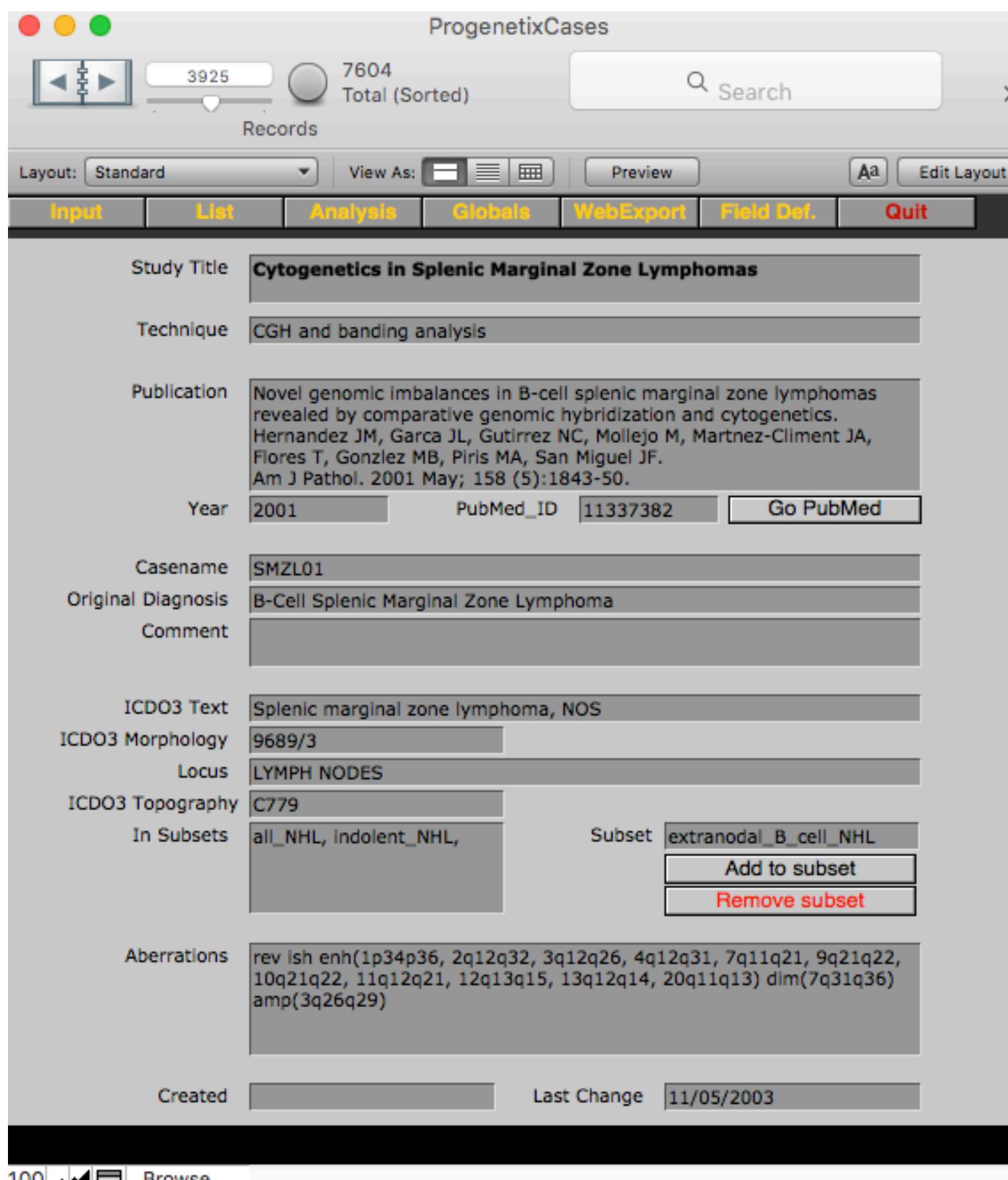
Progenetix in 2025

An oncogenomic reference resource



Database Structure

From flat database to hierarchical object storage



Archived version of 2003 "ProgenetixCases" FMP solution

2003

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

2025

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

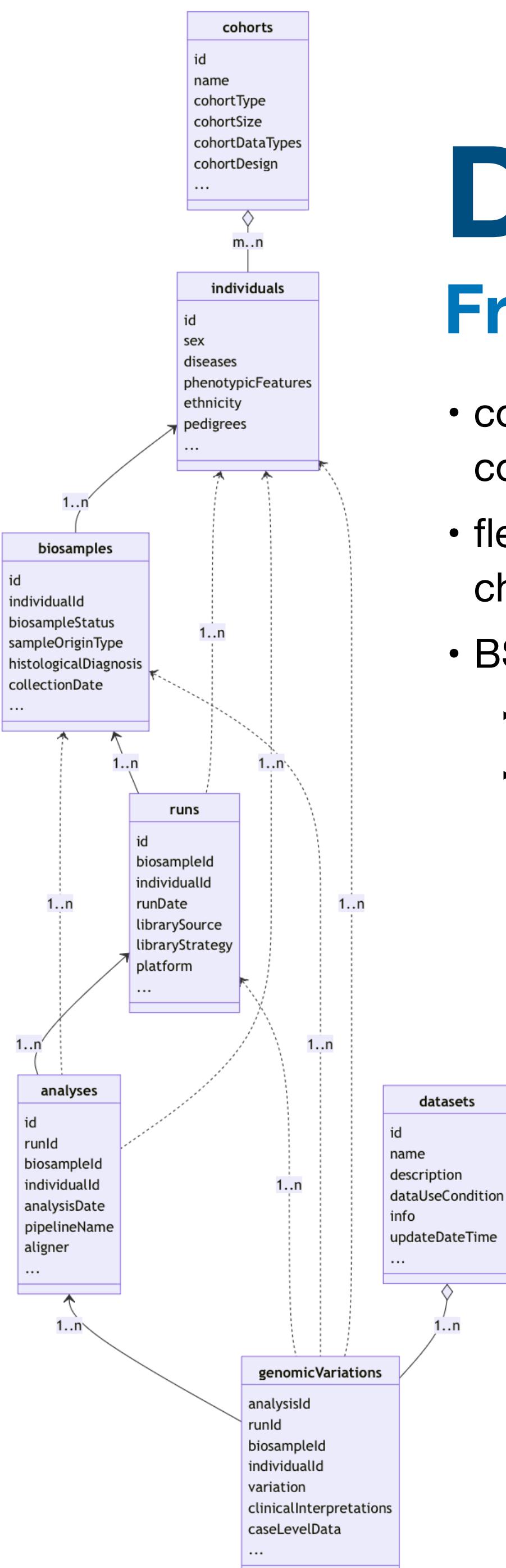
```
{
  "id" : "pgxind-kftx394x",
  "biocharacteristics" : [
    {
      "description" : "female",
      "type" : {
        "id" : "PATO:0020002",
        "label" : "female genotypic sex"
      }
    },
    {
      "description" : null,
      "type" : {
        "id" : "NCBITaxon:9606",
        "label"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

```
{
  "_id" : ObjectId("5bab583e727983b2e01255ae"),
  "callset_id" : "pgxcs-kftvv618",
  "biosample_id" : "pgxbs-kftvhcao",
  "assembly_id" : "GRCh38",
  "digest" : "7:107200000-158821424:DEL",
  "reference_name" : "7",
  "variant_type" : "DEL",
  "start" : 107200000,
  "end" : 158821424,
  "info" : {
    "cnv_value" : null,
    "cnv_length" : 51621424
  },
  "updated" : "2018-09-26 09:51:39.775397"
}
```

```
{
  "_id" : ObjectId("5bab56cd727983b2e00b0bde"),
  "id" : "pgxbs-kftvhcao",
  "description" : "Splenic Marginal Zone Lymphoma",
  "biocharacteristics" : [
    {
      "type" : {
        "id" : "UBERON:0002106",
        "label" : "spleen"
      }
    },
    {
      "type" : {
        "id" : "icdot-C42.2",
        "label" : "Spleen"
      }
    },
    {
      "type" : {
        "id" : "icdom-96893",
        "label" : "Splenic marginal zone B-cell lymphoma"
      }
    },
    {
      "type" : {
        "id" : "NCIT:C4663",
        "label" : "Splenic Marginal Zone Lymphoma"
      }
    }
  ],
  "individual_id" : "pgxind-kftx394x",
  "individual_age_at_collection" : "P67Y",
  "info" : {
    "death" : "0",
    "followup_months" : 53,
    "callset_ids" : [
      "pgxcs-kftvv618"
    ],
    "legacy_id" : "PGX_AM_BS_SMZL01"
  },
  "external_references" : [
    {
      "type" : {
        "id" : "PMID:11337382"
      }
    }
  ],
  "provenance" : {
    "material" : {
      "type" : {
        "id" : "EFO:0009656",
        "label" : "neoplastic sample"
      }
    },
    "geo" : {
      "label" : "Salamanca, Spain",
      "precision" : "city",
      "city" : "Salamanca",
      "country" : "Spain",
      "geojson" : {
        "type" : "Point",
        "coordinates" : [
          -3.68,
          40.43
        ]
      },
      "ISO-3166-alpha3" : "ESP"
    }
  },
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  }
}
```

Database Structure

From flat database to hierarchical object storage



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

→ rapid prototyping and implementation

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

2025

```

{
  "_id" : ObjectId("5bab56cd727983b2e00b0bde"),
  "id" : "pgxbs-kftvhcao",
  "description" : "Splenic Marginal Zone Lymphoma",
  "biocharacteristics" : [
    {
      "type" : {
        "id" : "UBERON:0002106",
        "label" : "spleen"
      }
    },
    {
      "type" : {
        "id" : "icdot-C42.2",
        "label" : "Spleen"
      }
    },
    {
      "type" : {
        "id" : "icdom-96893",
        "label" : "Splenic marginal zone B-cell lymphoma"
      }
    }
  ],
  "individual_id" : "pgxind-kftx394x",
  "individual_age_at_collection" : "P67Y",
  "info" : {
    "death" : "0",
    "followup_months" : 53,
    "callset_ids" : [
      "pgxcs-kftvv618"
    ],
    "legacy_id" : "PGX_AM_BS_SMZL01"
  },
  "external_references" : [
    {
      "type" : {
        "id" : "PMID:11337382"
      }
    }
  ],
  "provenance" : {
    "material" : {
      "type" : {
        "id" : "EFO:0009656",
        "label" : "neoplastic sample"
      }
    }
  },
  "geo" : {
    "label" : "Salamanca, Spain",
    "precision" : "city",
    "city" : "Salamanca",
    "country" : "Spain",
    "latitude" : 40.43,
    "longitude" : -3.68
  },
  "info" : {
    "legacy_id" : "PGX_IND_SMZL01"
  },
  "updated" : ISODate("2018-09-26T09:51:39.775Z")
}

{
  "_id" : ObjectId("5bab583e727983b2e01255ae"),
  "callset_id" : "pgxcs-kftvv618",
  "biosample_id" : "pgxbs-kftvhcao",
  "assembly_id" : "GRCh38",
  "digest" : "7:107200000-158821424:DEL",
  "reference_name" : "7",
  "variant_type" : "DEL",
  "start" : 107200000,
  "end" : 158821424,
  "info" : {
    "cnv_value" : null,
    "cnv_length" : 51621424
  },
  "updated" : "2018-09-26 09:51:39.775397"
}
  
```

```

{
  "_id" : ObjectId("5bab56cd727983b2e00b0bde"),
  "id" : "pgxbs-kftvhcao",
  "description" : "Splenic Marginal Zone Lymphoma",
  "biocharacteristics" : [
    {
      "type" : {
        "id" : "UBERON:0002106",
        "label" : "spleen"
      }
    },
    {
      "type" : {
        "id" : "icdot-C42.2",
        "label" : "Spleen"
      }
    },
    {
      "type" : {
        "id" : "icdom-96893",
        "label" : "Splenic marginal zone B-cell lymphoma"
      }
    }
  ],
  "individual_id" : "pgxind-kftx394x",
  "individual_age_at_collection" : "P67Y",
  "info" : {
    "death" : "0",
    "followup_months" : 53,
    "callset_ids" : [
      "pgxcs-kftvv618"
    ],
    "legacy_id" : "PGX_AM_BS_SMZL01"
  },
  "external_references" : [
    {
      "type" : {
        "id" : "PMID:11337382"
      }
    }
  ],
  "provenance" : {
    "material" : {
      "type" : {
        "id" : "EFO:0009656",
        "label" : "neoplastic sample"
      }
    }
  },
  "geo" : {
    "label" : "Salamanca, Spain",
    "precision" : "city",
    "city" : "Salamanca",
    "country" : "Spain",
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        -3.68,
        40.43
      ]
    },
    "ISO-3166-alpha3" : "ESP"
  },
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  }
}
  
```

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,*}

¹Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland

²Swiss 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

^aset of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Thesaurus.

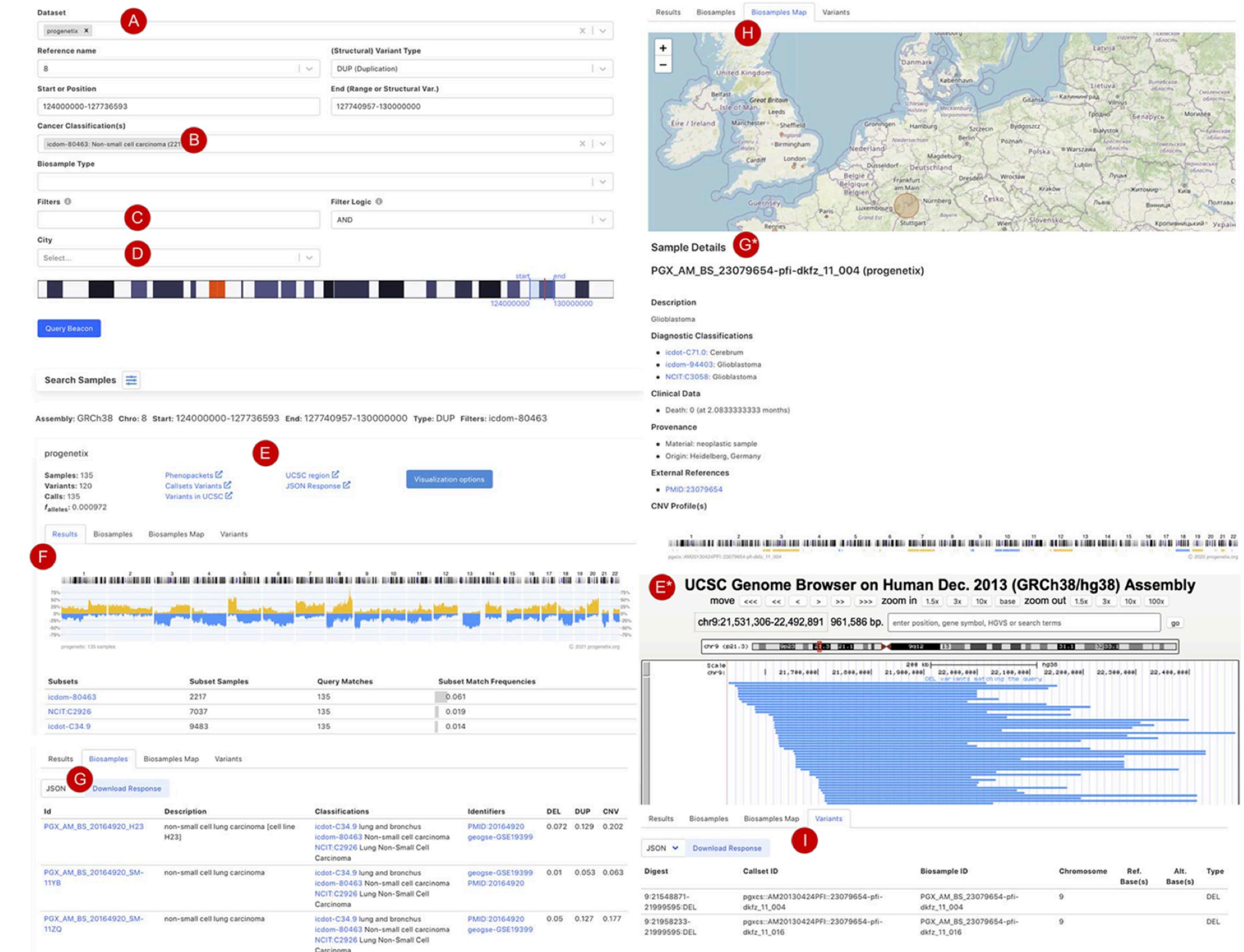
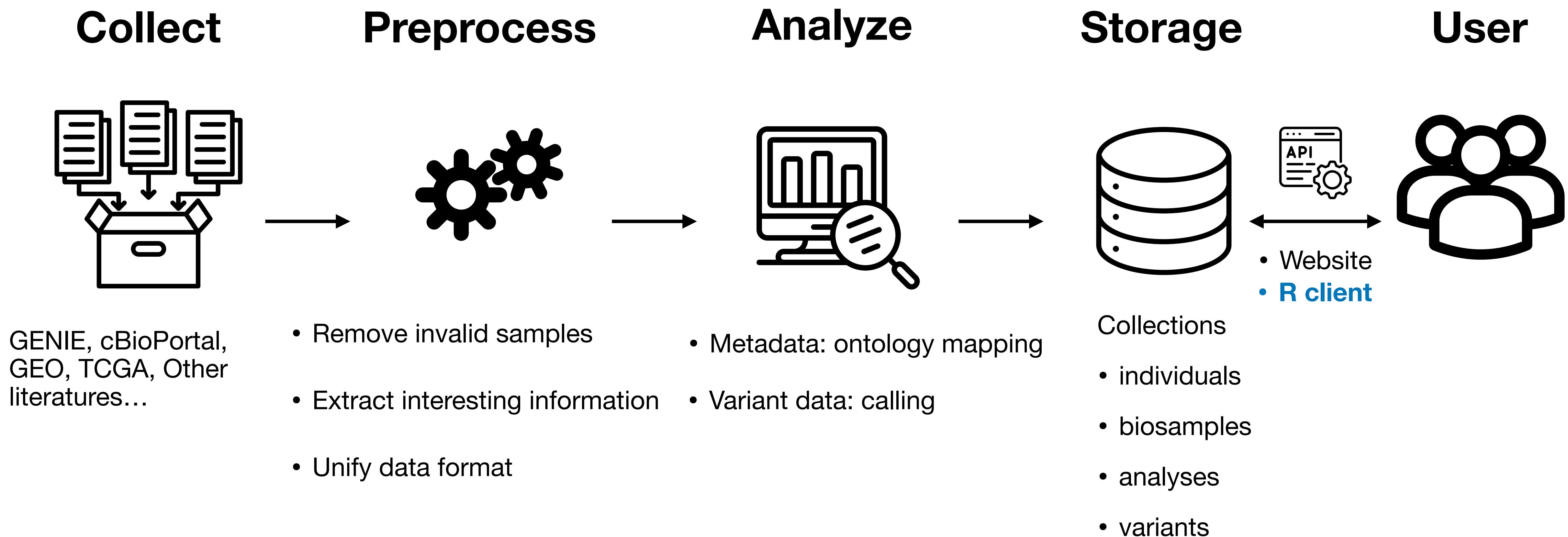


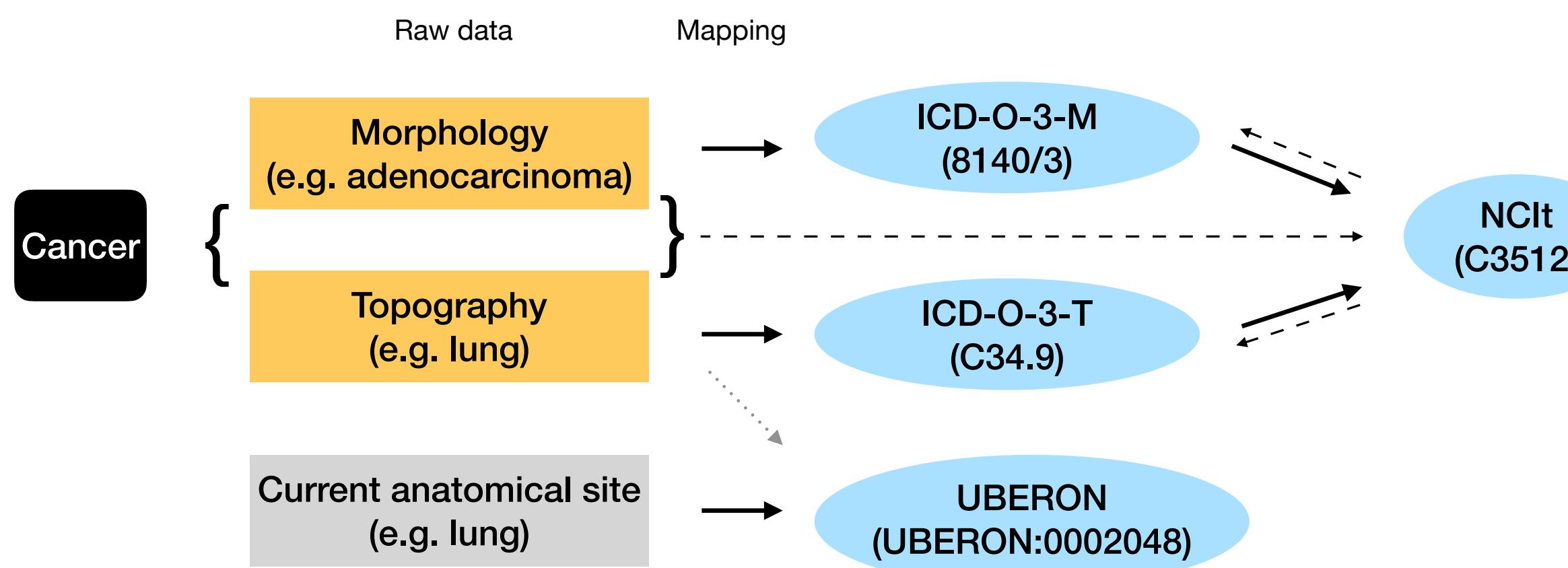
Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range. This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with ≤ 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and ICD-O-3 ontology terms available); C: additional filter, e.g. Cellosaurus; D: additional filter for geographic location; E: external link to UCSC browser to view the alignment of matched variants; F: cancer type classification sorted by frequency of the matched biosamples present in the subset; G: list of matched biosamples with description, statistics and reference. More detailed biosample information can be viewed through 'id' link to the sample detail page; H: matched variants with reference to biosamples can be downloaded in json or csv format.

End-to-End Data Pipeline

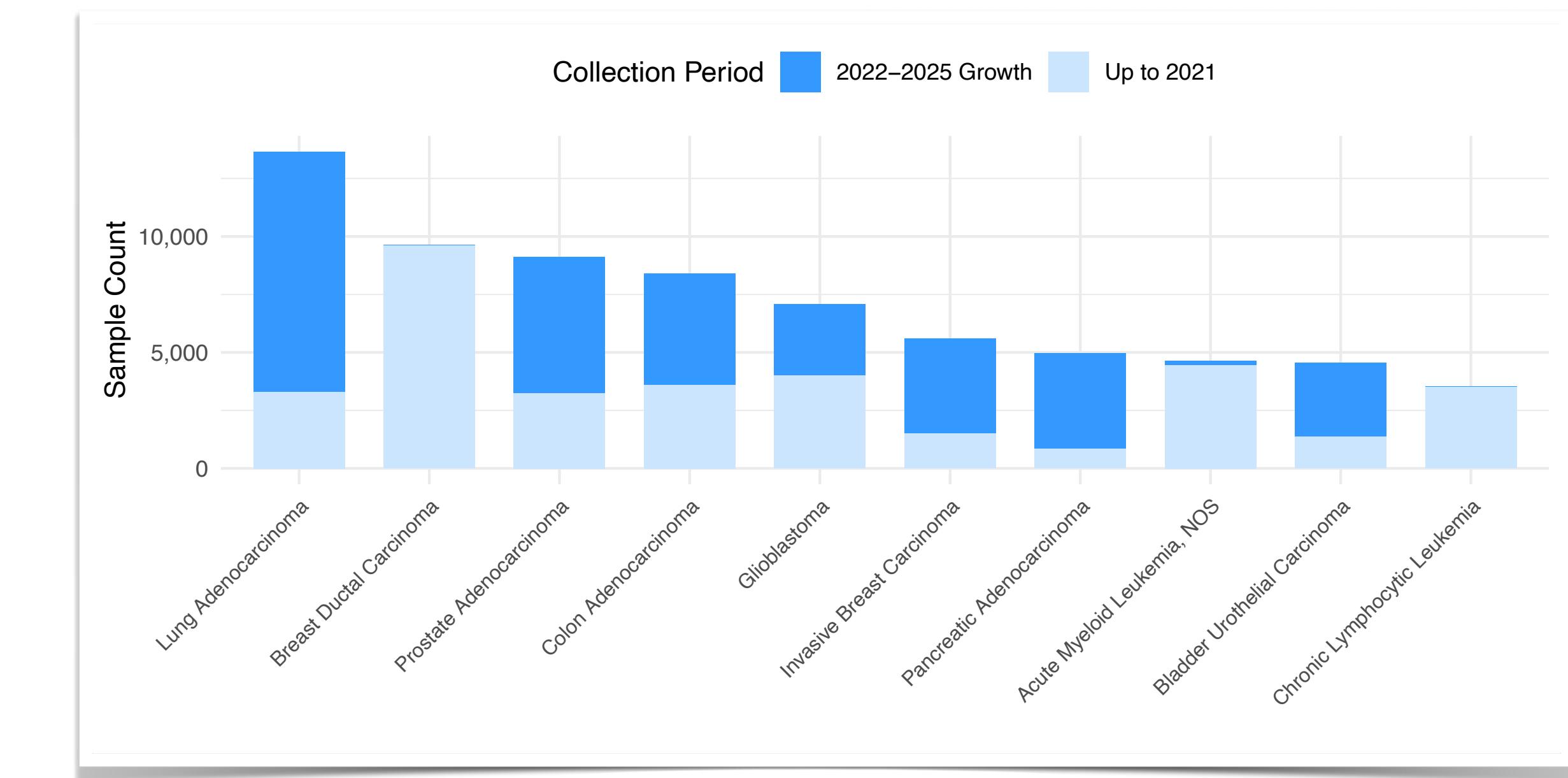


Enhance Progenetix

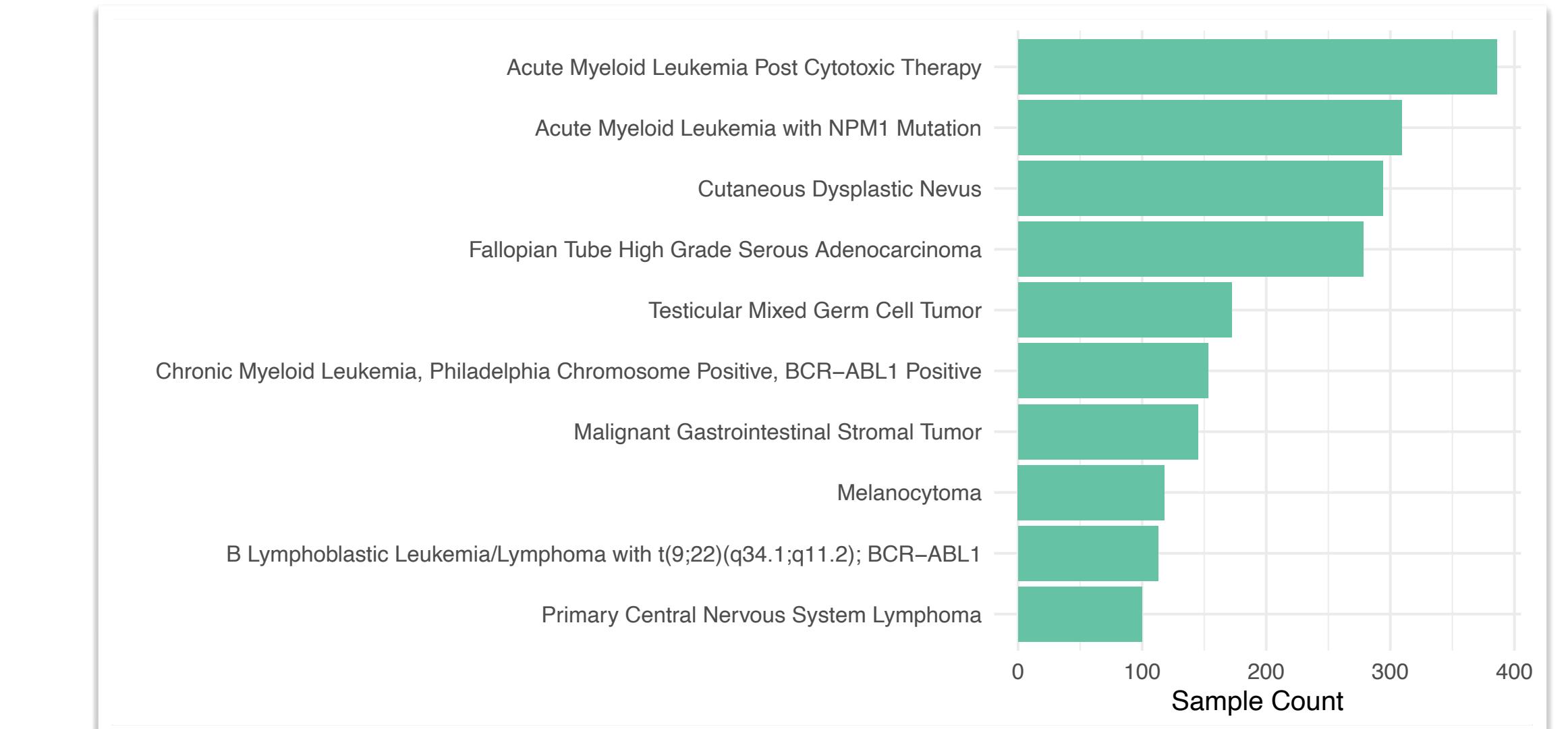
Optimization and expansion of cancer type representation



Top 10 most frequent cancer

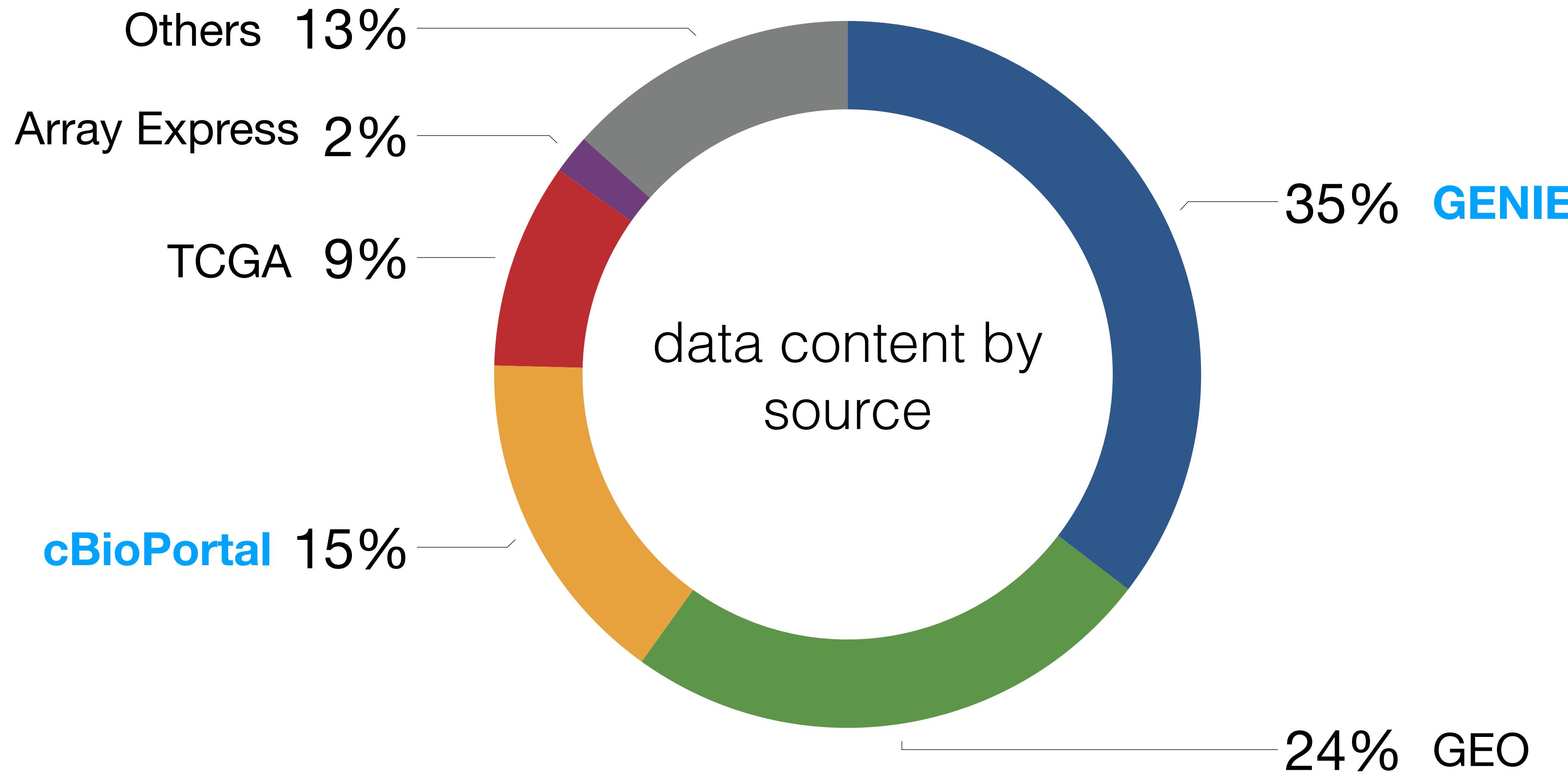


More granular terms in new NCIt cancer types



Enhance Progenetix

Import ~100k new samples



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

Cancer Genomics Reference Resource

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



CNV Profiles

- ... by NCIT
- ... by ICD-O Morphology
- ... by ICD-O Site
- ... by TNM & Grade

Search Samples

arrayMap

- TCGA Data
- cBioPortal Studies

Publication DB

Progenetix Use

NCIT - ICD-O Mappings

UBERON Mappings

Upload & Plot

OpenAPI Paths and Examples

Cancer Cell Lines

Beacon+

Documentation

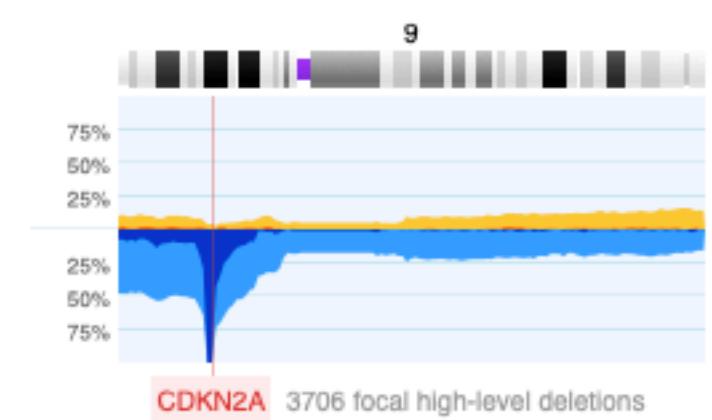
Baudisgroup @ UZH

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* of currently **240600** samples from **1126** different cancer types (NCIt neoplasm classification)

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

Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the respective Cancer Types pages (e.g. [NCIT Neoplasia Codes](#) ) and compared through the [Compare CNV Profiles](#)  option. Below is an example of aggregated CNV data in 11 samples in Oligoastrocytoma with the frequency of regional **copy number gains (high level)** and **losses (high level)** displayed for the 22 autosomes.



[Download SVG](#) | [Go to NCIT:C4050](#) | [Download CNV Frequencies](#)

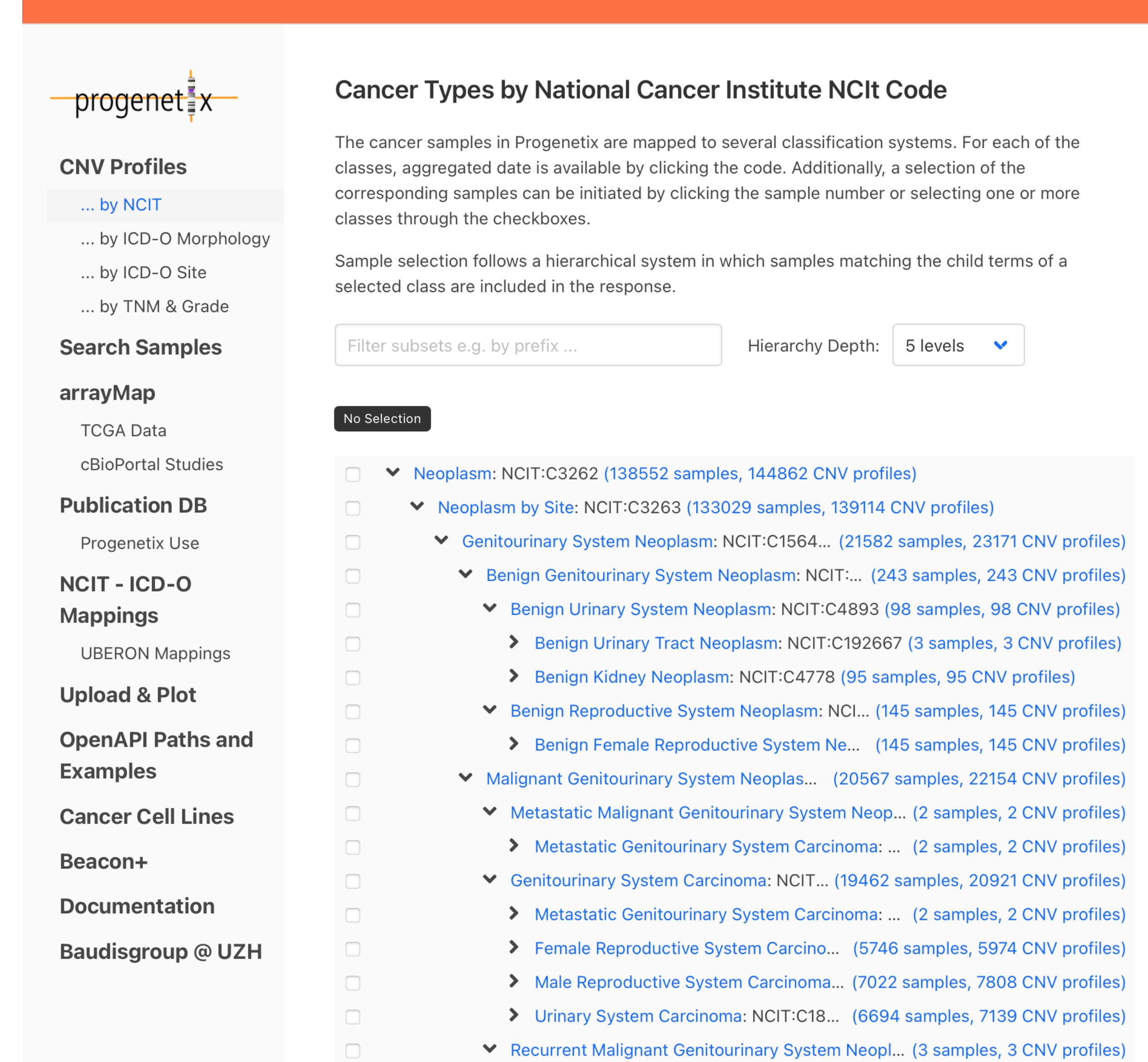
Cancer Genomics Publications

Through the [[Publications](#)] page Progenetix provides 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.

© CC-BY 2001 - 2025 progenetix.org

Cancer Genomics Reference Resource

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

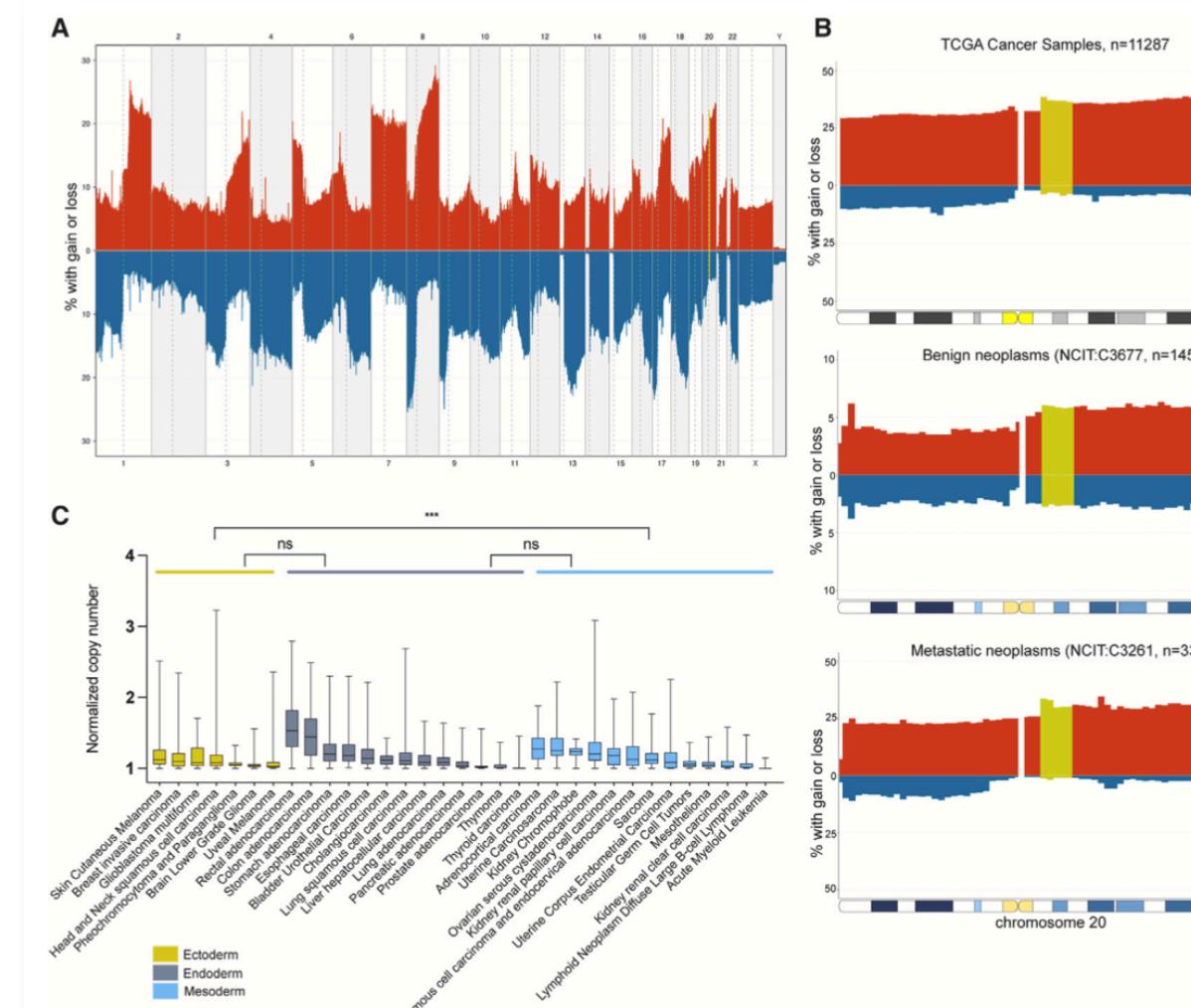


The screenshot shows the Progenetix interface with a sidebar on the left containing links like 'CNV Profiles', 'Search Samples', 'arrayMap', 'Publication DB', 'NCIT - ICD-O Mappings', 'Upload & Plot', 'OpenAPI Paths and Examples', 'Cancer Cell Lines', 'Beacon+', 'Documentation', and 'Baudisgroup @ UZH'. The main area is titled 'Cancer Types by National Cancer Institute NCI Code' and contains a detailed hierarchical tree of cancer types. A search bar at the top says 'Filter subsets e.g. by prefix ...' and a dropdown says 'Hierarchy Depth: 5 levels'. The tree starts with 'Neoplasm' (138552 samples, 144862 CNV profiles), which branches into 'Neoplasm by Site' (133029 samples, 139114 CNV profiles), then into 'Genitourinary System Neoplasm' (21582 samples, 23171 CNV profiles), 'Benign Genitourinary System Neoplasm' (243 samples, 243 CNV profiles), 'Benign Urinary System Neoplasm' (98 samples, 98 CNV profiles), 'Benign Urinary Tract Neoplasm' (3 samples, 3 CNV profiles), 'Benign Kidney Neoplasm' (95 samples, 95 CNV profiles), 'Benign Reproductive System Neoplasm' (145 samples, 145 CNV profiles), 'Benign Female Reproductive System Neoplasm' (145 samples, 145 CNV profiles), 'Malignant Genitourinary System Neoplasms' (20567 samples, 22154 CNV profiles), 'Metastatic Malignant Genitourinary System Neoplasms' (2 samples, 2 CNV profiles), 'Metastatic Genitourinary System Carcinoma' (2 samples, 2 CNV profiles), 'Genitourinary System Carcinoma' (19462 samples, 20921 CNV profiles), 'Metastatic Genitourinary System Carcinoma' (2 samples, 2 CNV profiles), 'Female Reproductive System Carcinoma' (5746 samples, 5974 CNV profiles), 'Male Reproductive System Carcinoma' (7022 samples, 7808 CNV profiles), 'Urinary System Carcinoma' (6694 samples, 7139 CNV profiles), and 'Recurrent Malignant Genitourinary System Neoplasms' (3 samples, 3 CNV profiles).

Progenetix Use

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

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



Progenetix References

arrayMap progenetix cancercellines

Articles Citing - or Using - Progenetix

This page lists articles which we found to have made use of, or referred to, the Progenetix resource ecosystem. These articles may not necessarily contain original case profiles themselves. Please contact us to alert us about additional articles you are aware of. Also, you can now directly submit suggestions for matching publications to the oncopubs repository on Github.

Filter

Publications (121)	Samples		
id	Publication	Genomes	pgx
PMID:38157850	Krivec N, Ghosh MS et al. (2024) Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research. ... Stem Cell Reports	0	0
PMID:37627037	Austin BK, Firooz A, Valafar H et al. (2023) An Updated Overview of Existing Cancer Databases and Identified Needs. Biology (Basel)	0	0
PMID:37393410	Liu SC, Wang CI, Liu TT, Tsang NM et al. (2023) A 3-gene signature comprising CDH4, STAT4 and EBV-encoded LMP1 for early diagnosis ... Discov Oncol	0	0

Stem Cell Reports Review



OPEN ACCESS

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

Nuša Krivec,^{1,2} Manjusha S. Ghosh,^{1,2} and Claudia Spits^{1,2,*}

¹Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium

²These authors contributed equally.

*Correspondence: claudia.spits@vub.be
<https://doi.org/10.1016/j.stemcr.2023.11.013>

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

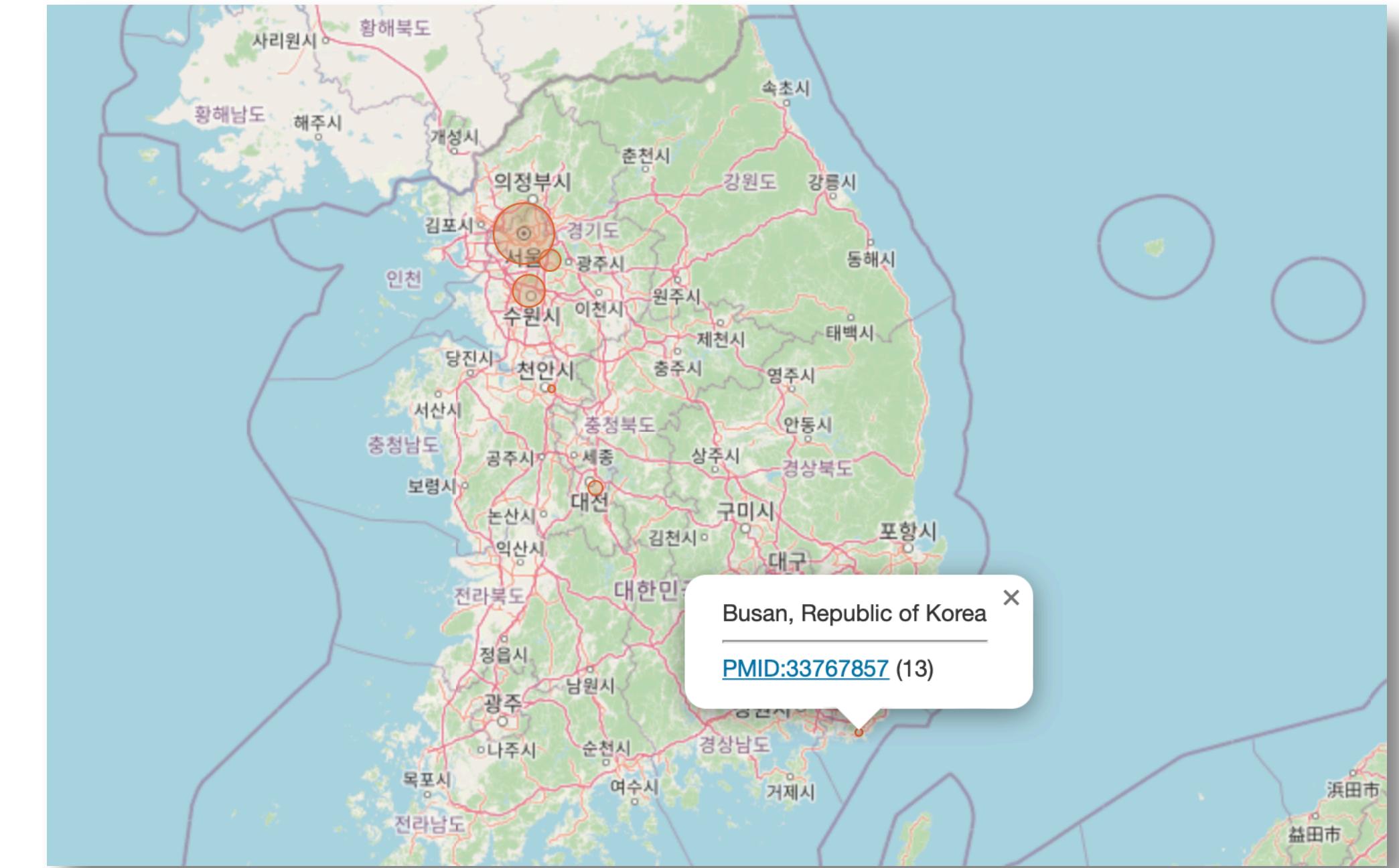
(A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal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

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

Service: Publications

Location Mapping for Statistics and Discovery...

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



[PMID:33767857](#) ↗

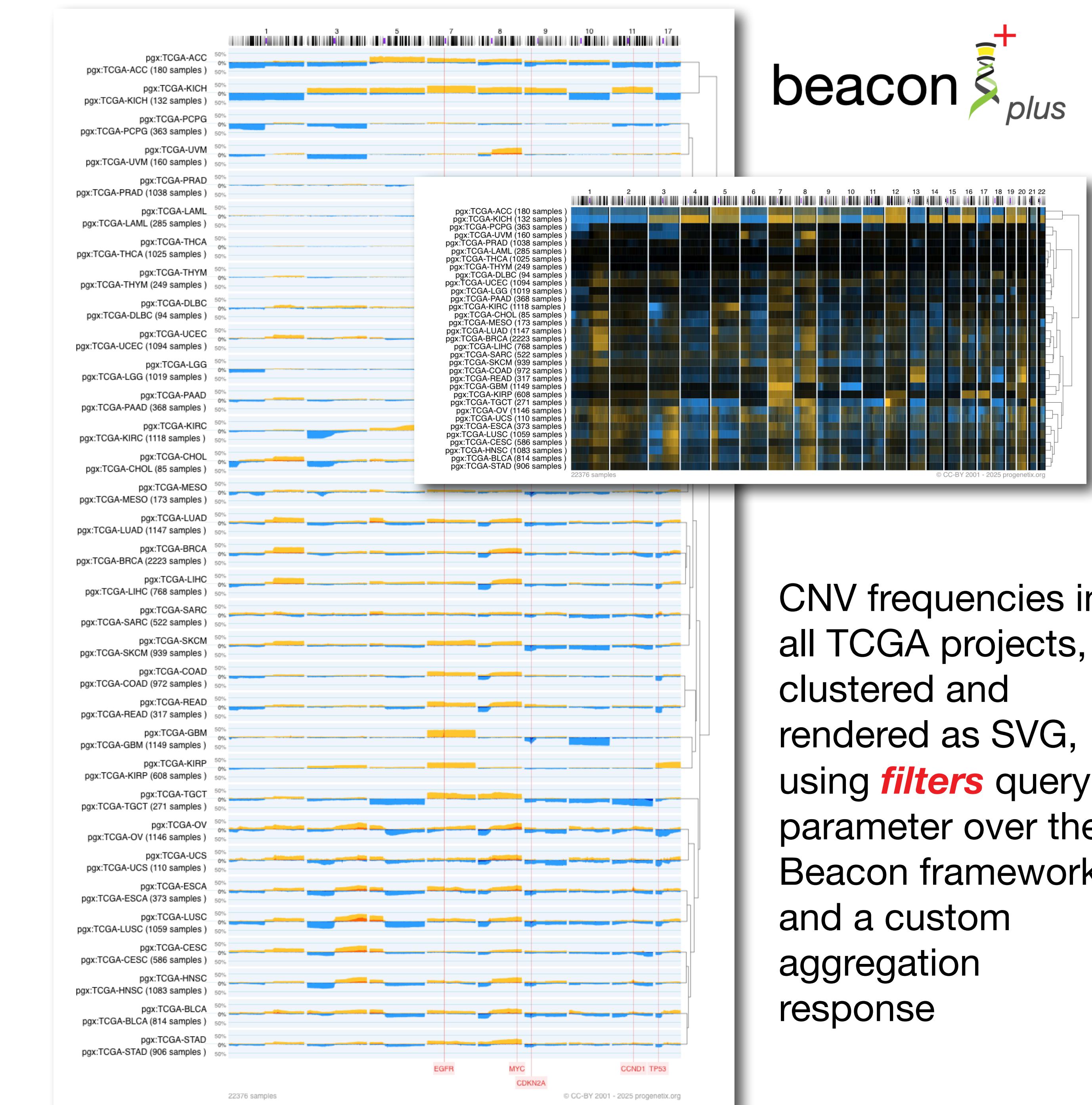
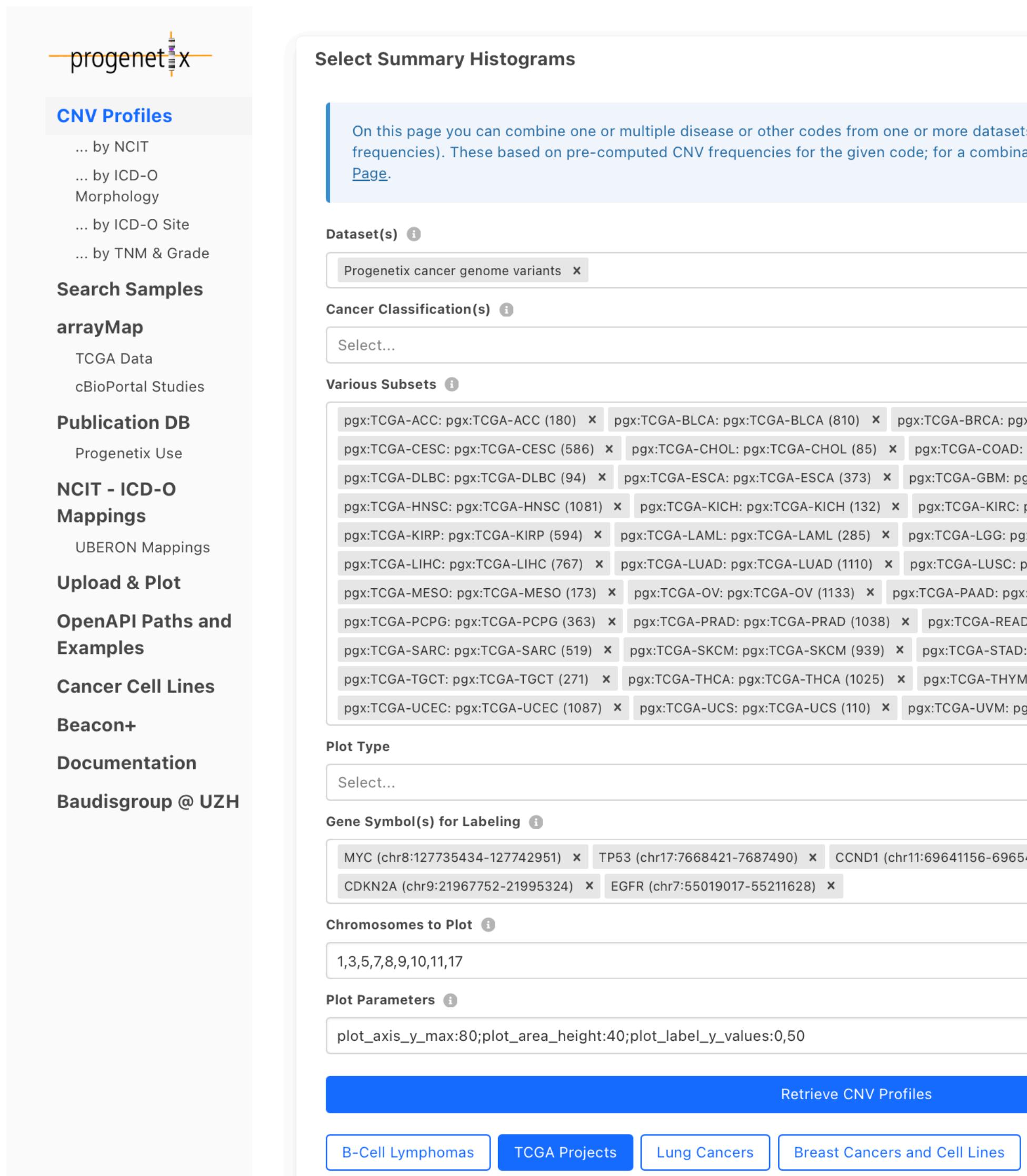
Methylation and molecular profiles of ependymoma: Influence of patient age and tumor anatomic location.

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

Mol Clin Oncol PMID:33767857 ↗

Pushing the envelope...

Custom Beacon aggregation response for displaying CNV frequencies



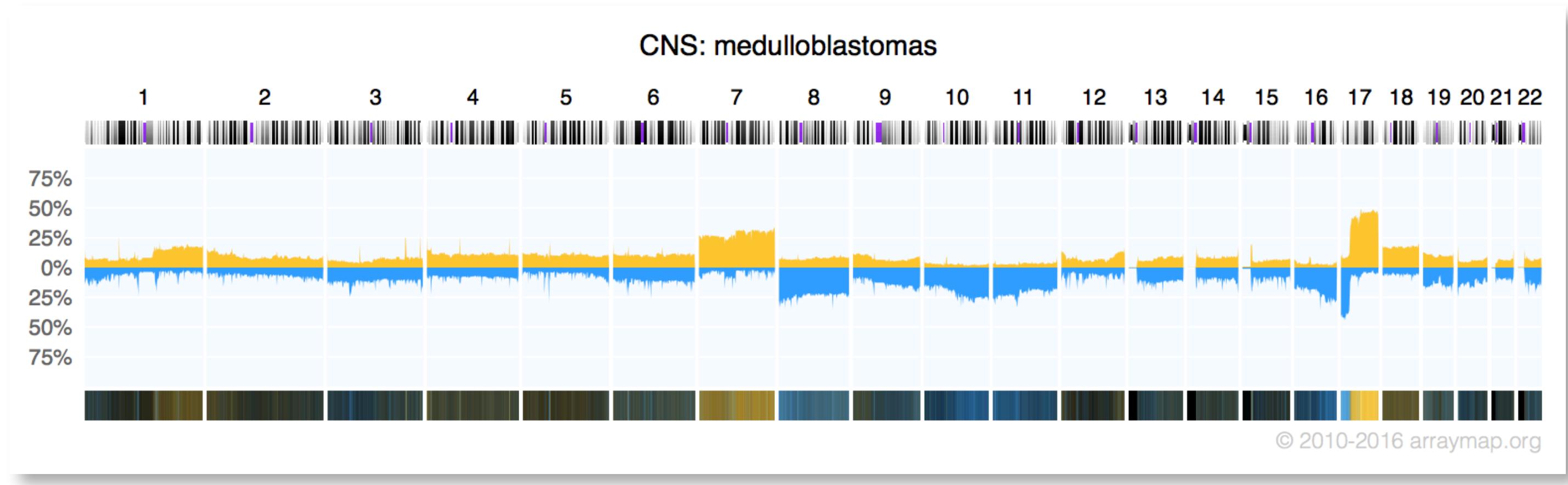
CNV frequencies in all TCGA projects, clustered and rendered as SVG, using ***filters*** query parameter over the Beacon framework and a custom aggregation response

Data Use Cases

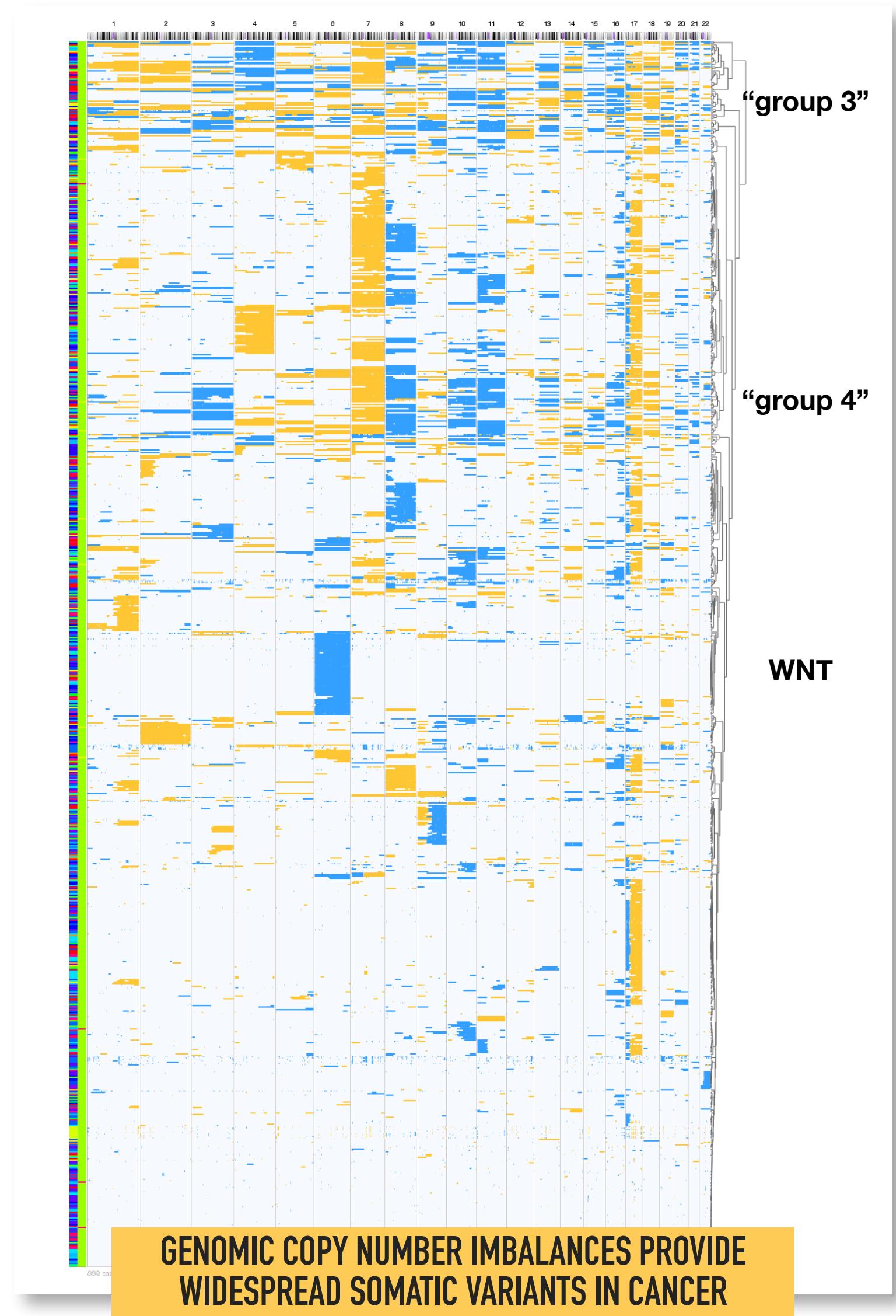
Somatic CNVs In Cancer

Recurrent mutation patterns

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



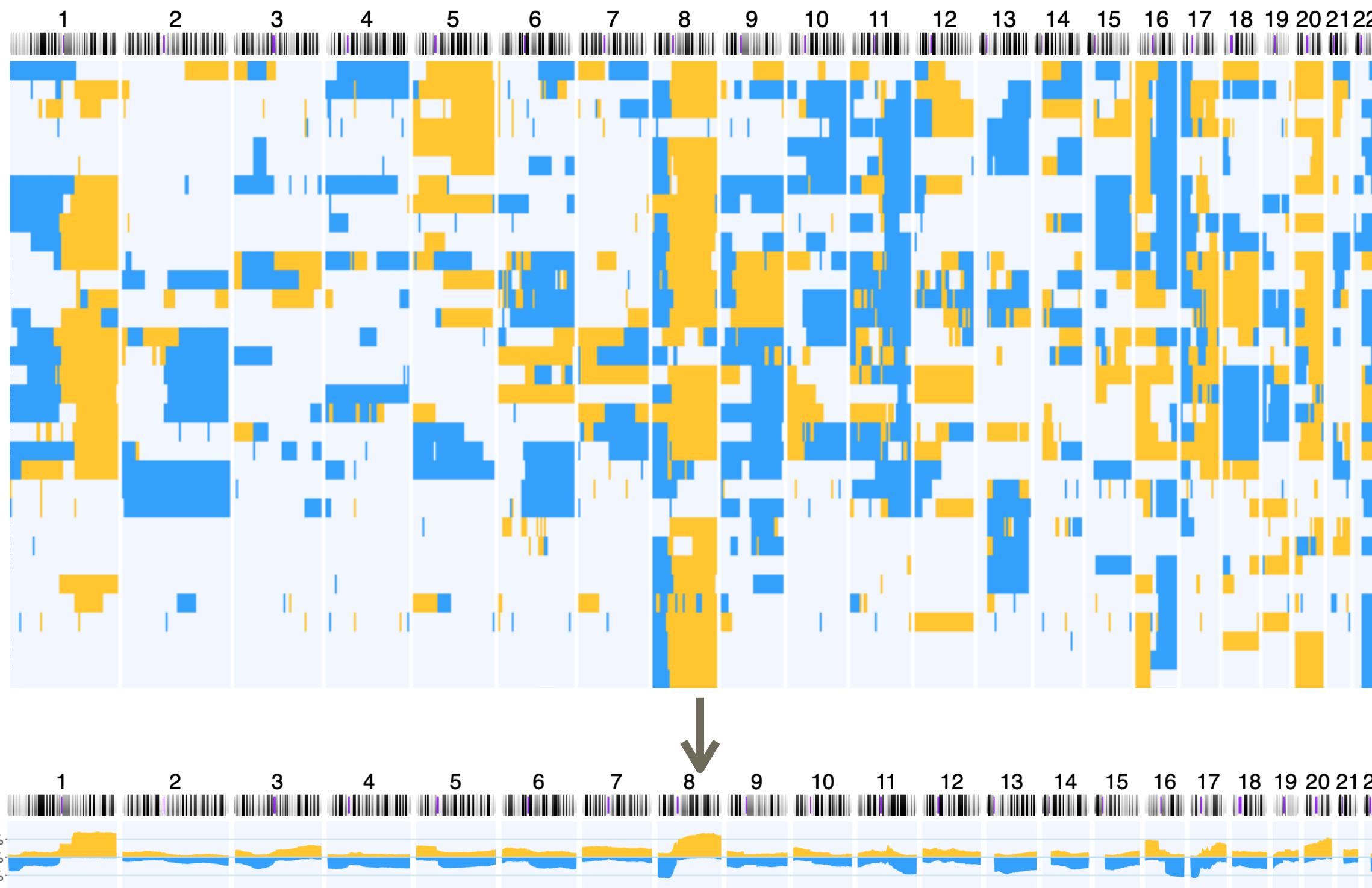
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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



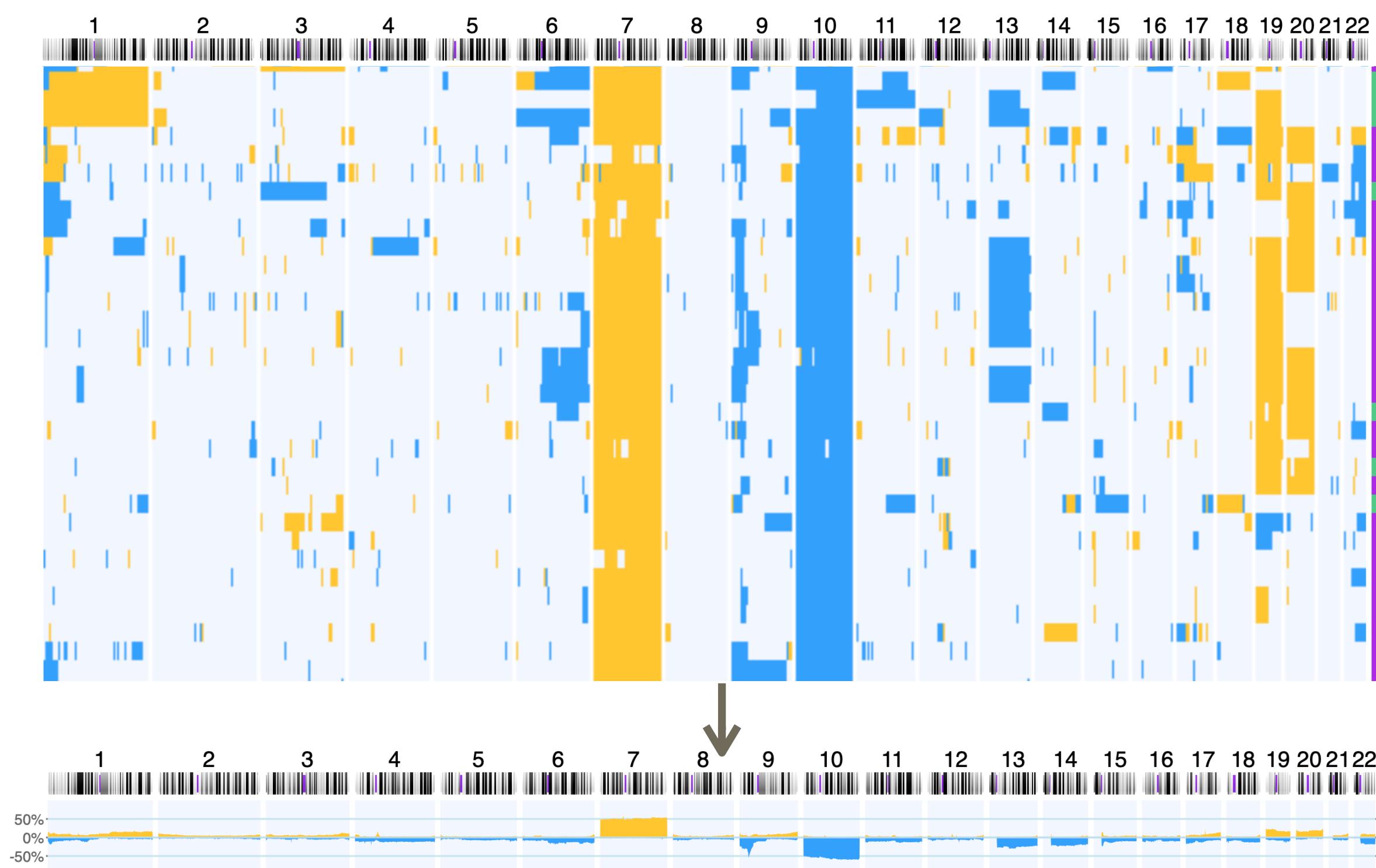
Drivers? Passengers? Markers?

Disentangling CNA Patterns

Ductal Breast Carcinoma



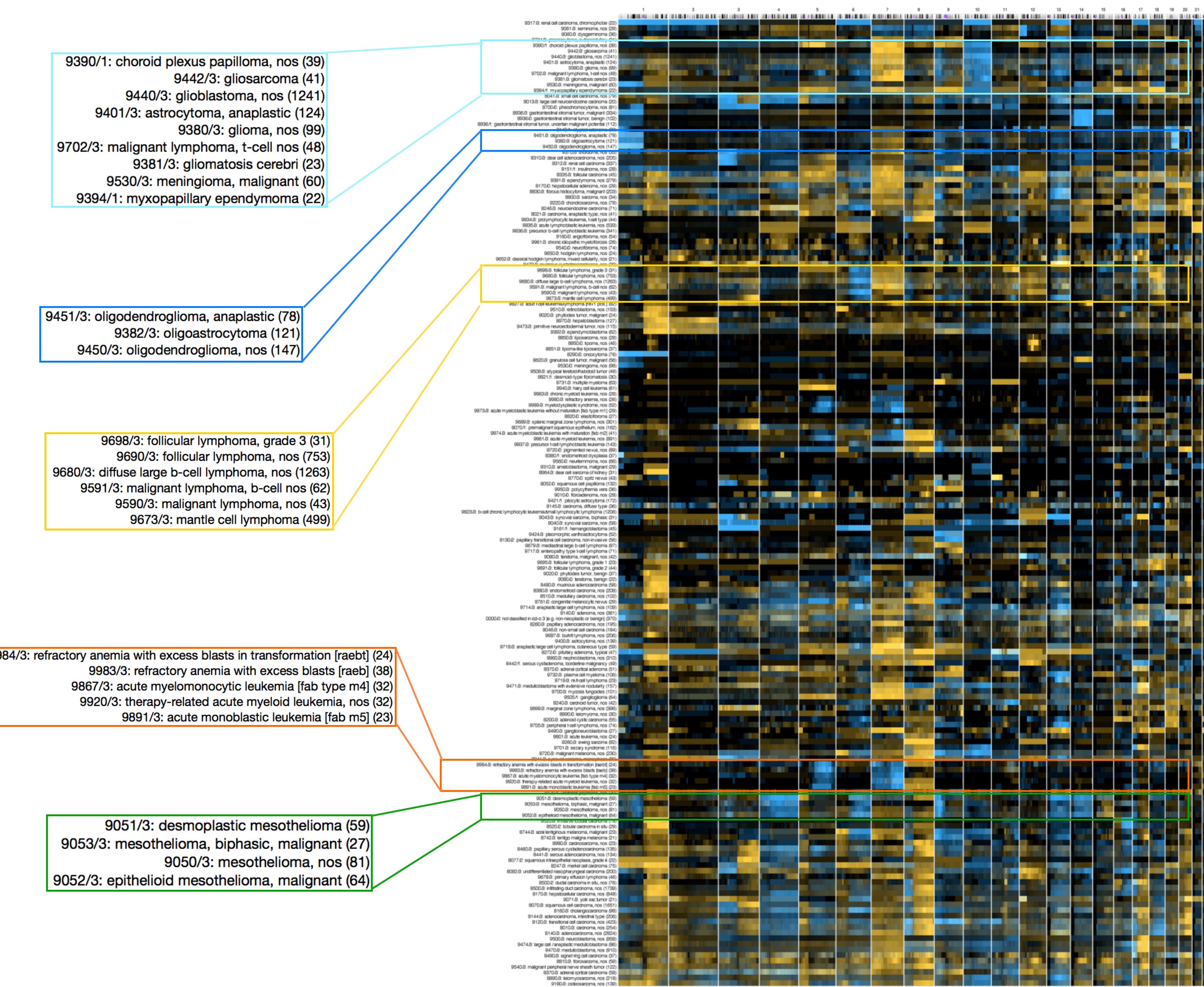
Glioblastoma

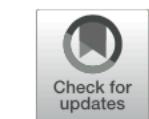


Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles





Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes

Bo Gao^{1,2} and Michael Baudis^{1,2*}

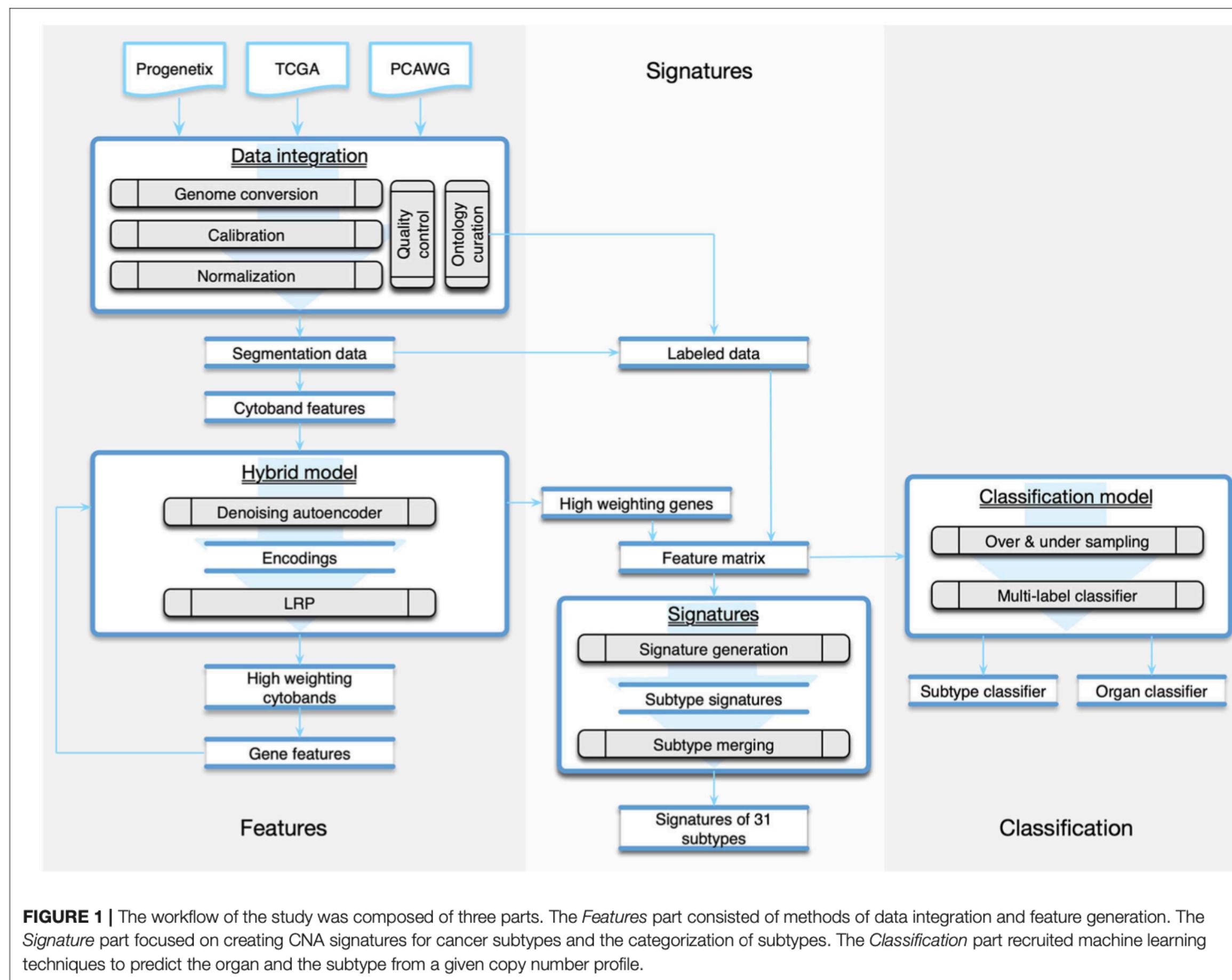


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine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

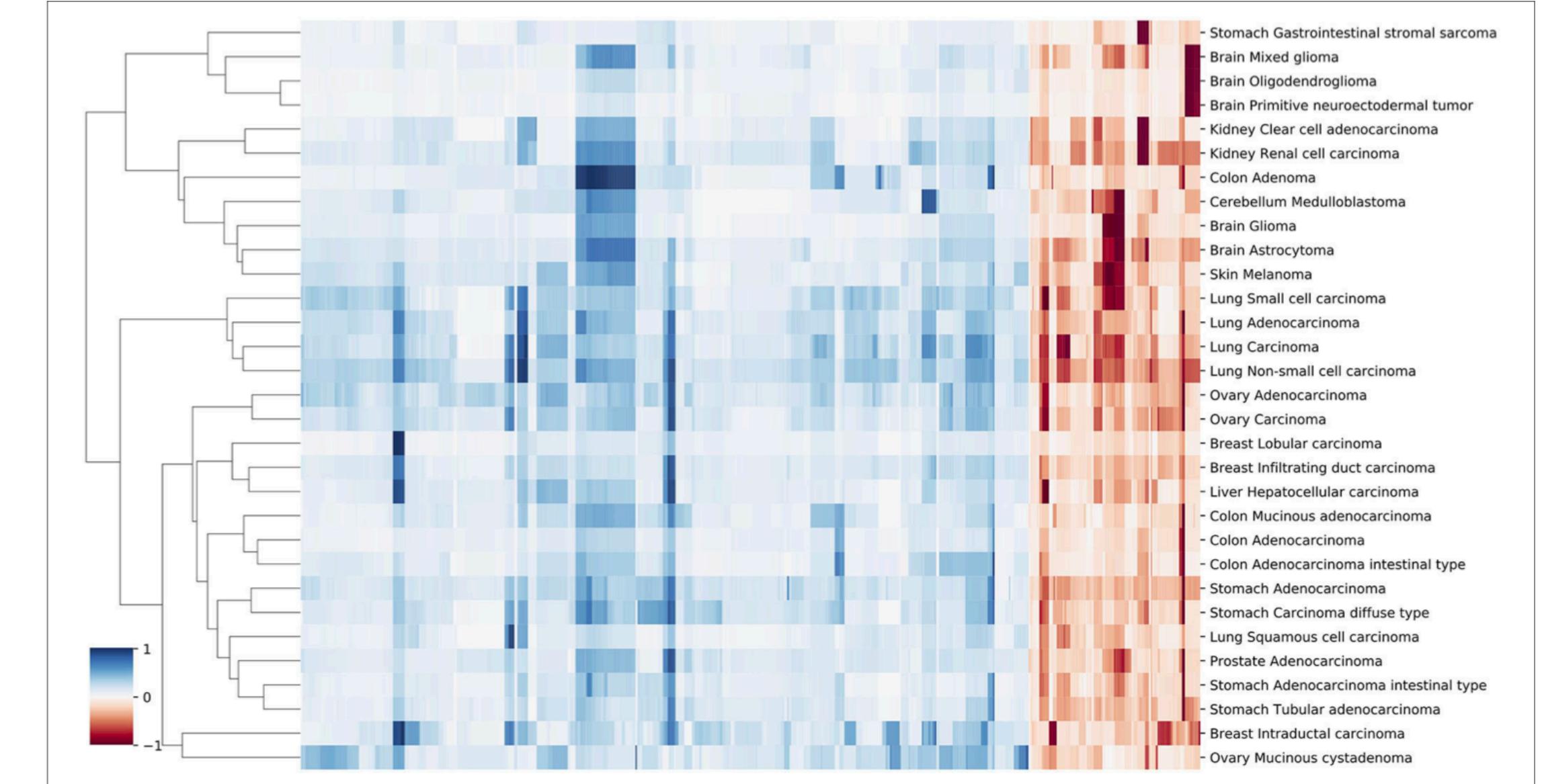


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

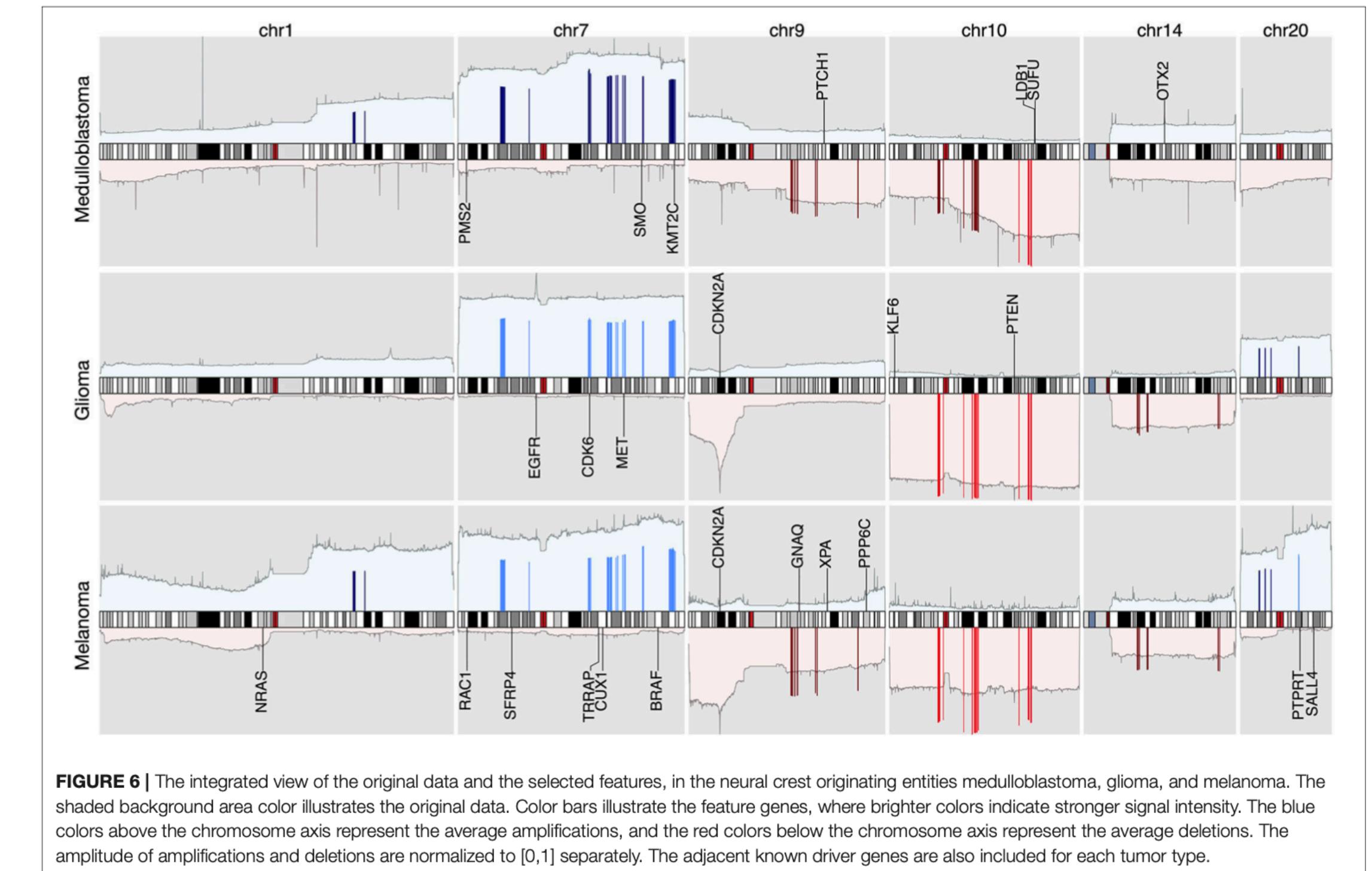
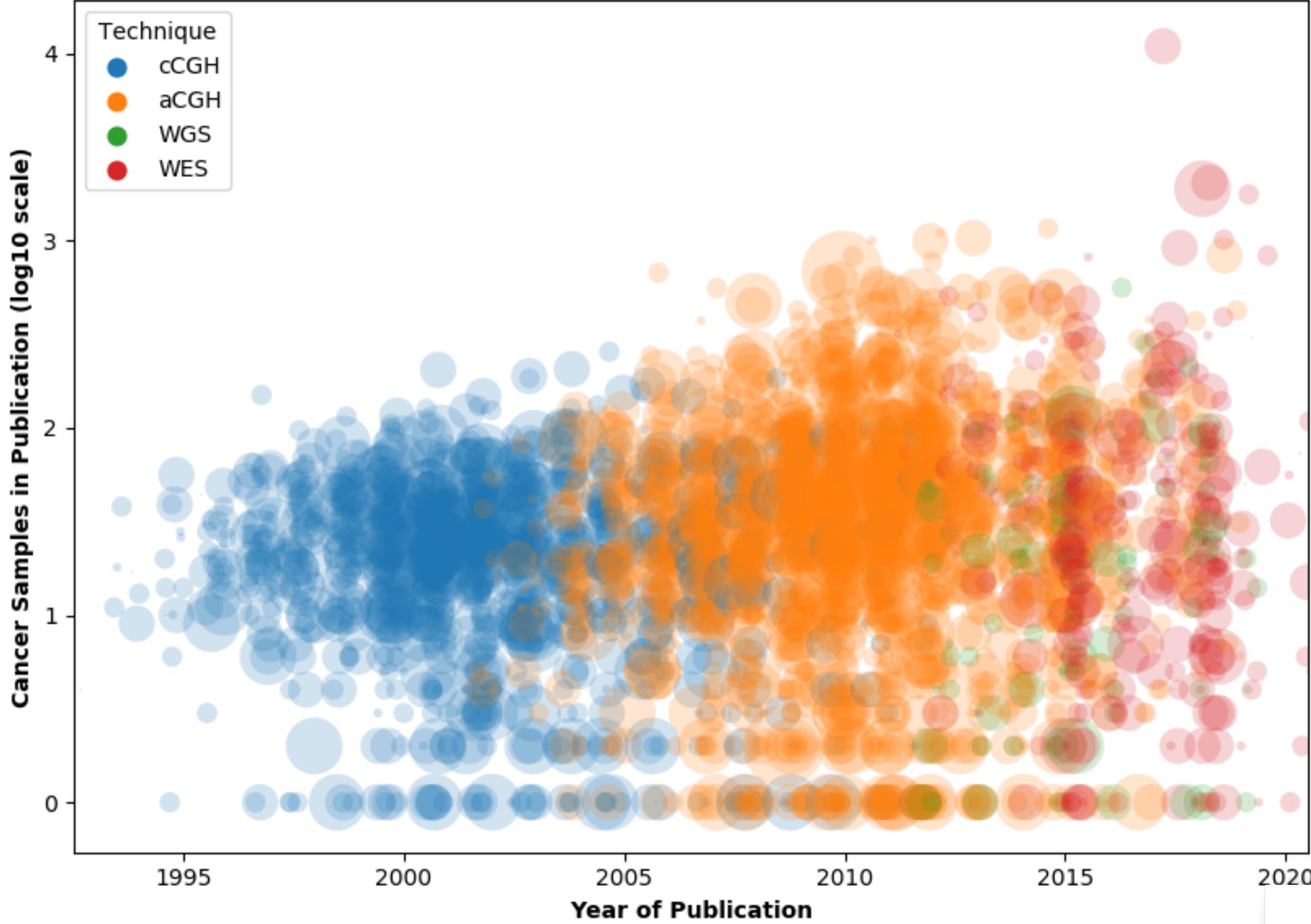


FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger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

Number of tumor samples for each publication across the years



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 Publication Collection

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

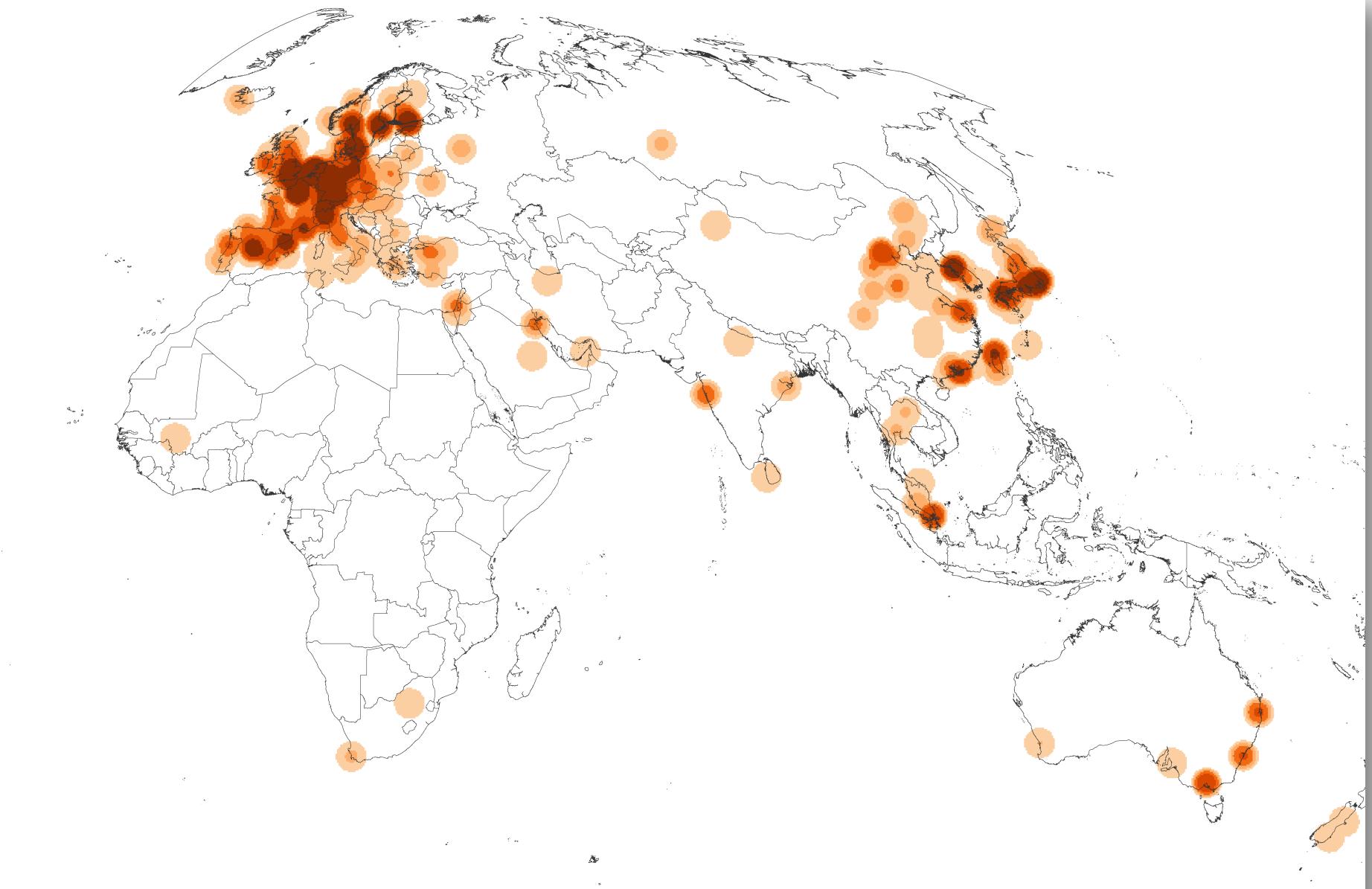
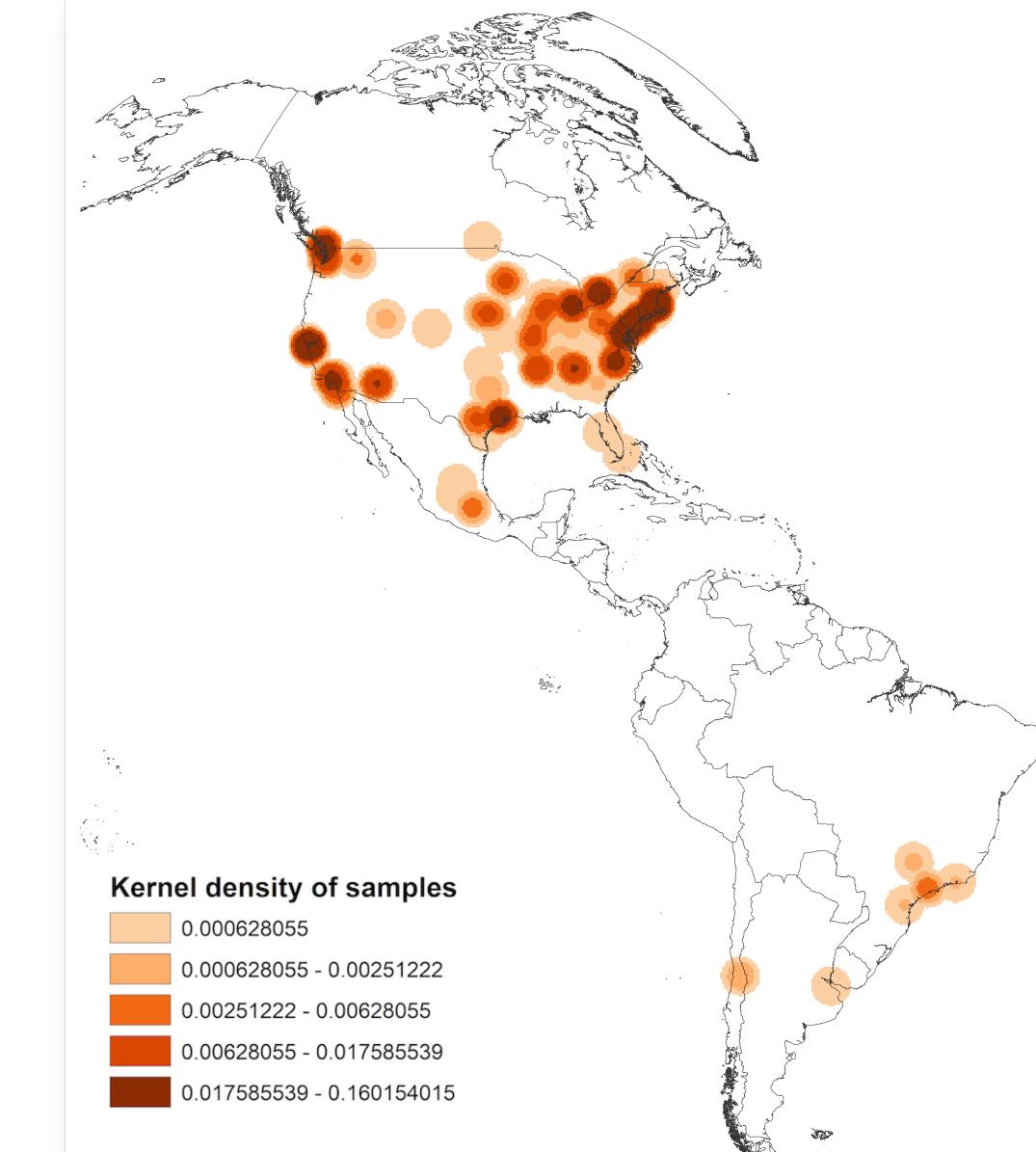
Please [contact us](#) to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation [🔗](#).

Filter [i](#) City [i](#)

 Type to search... [▼](#)

Publications (3324)

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



Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard



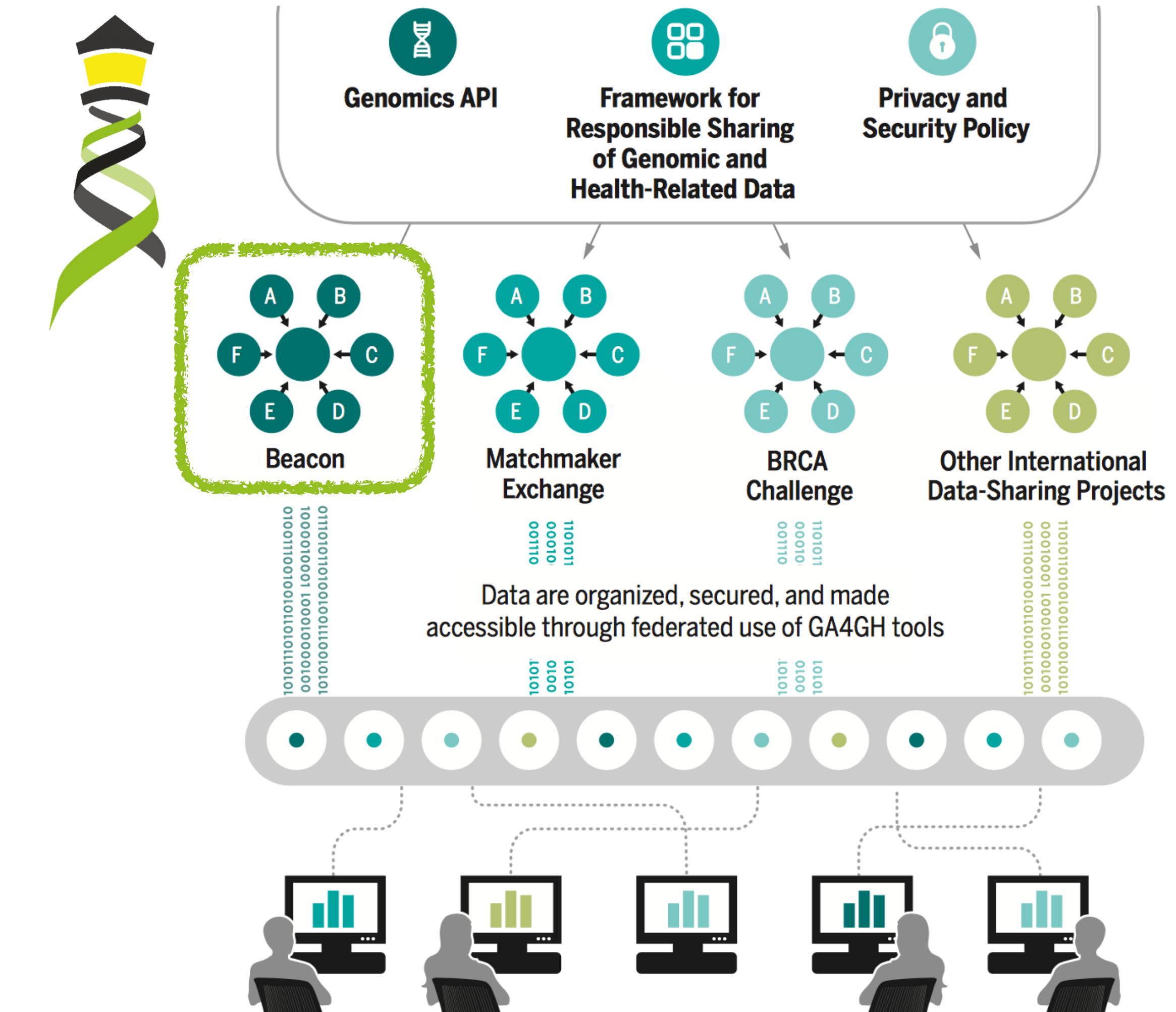


GENOMICS

A federated ecosystem for sharing genomic, clinical data

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

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



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

Beacon v2 GA4GH Approval Registry

Beacons:    

 **European Genome-Phenome Archive (EGA)**

[Visit us](#) [Beacon API](#) [Contact us](#)

GA4GH Approval Beacon Test

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	██████████
Bioinformatics analysis	██████████
Biological Sample	██████████
Cohort	██████████
Configuration	██████████
Dataset	██████████
EntryTypes	██████████
Genomic Variants	██████████
Individual	██████████
Info	██████████
Sequencing run	██████████

 **progenetix**

[Visit us](#) [Beacon UI](#) [Beacon API](#) [Contact us](#)

Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

Progenetix Cancer Genomics Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the...

BeaconMap	██████████
Bioinformatics analysis	██████████
Biological Sample	██████████
Cohort	██████████
Configuration	██████████
Dataset	██████████
EntryTypes	██████████
Genomic Variants	██████████
Individual	██████████
Info	██████████
Sequencing run	██████████

 **UNIVERSITY OF LEICESTER**

[Beacon UI](#) [Beacon API](#) [Contact us](#)

University of Leicester

Cafe Variome Beacon v2

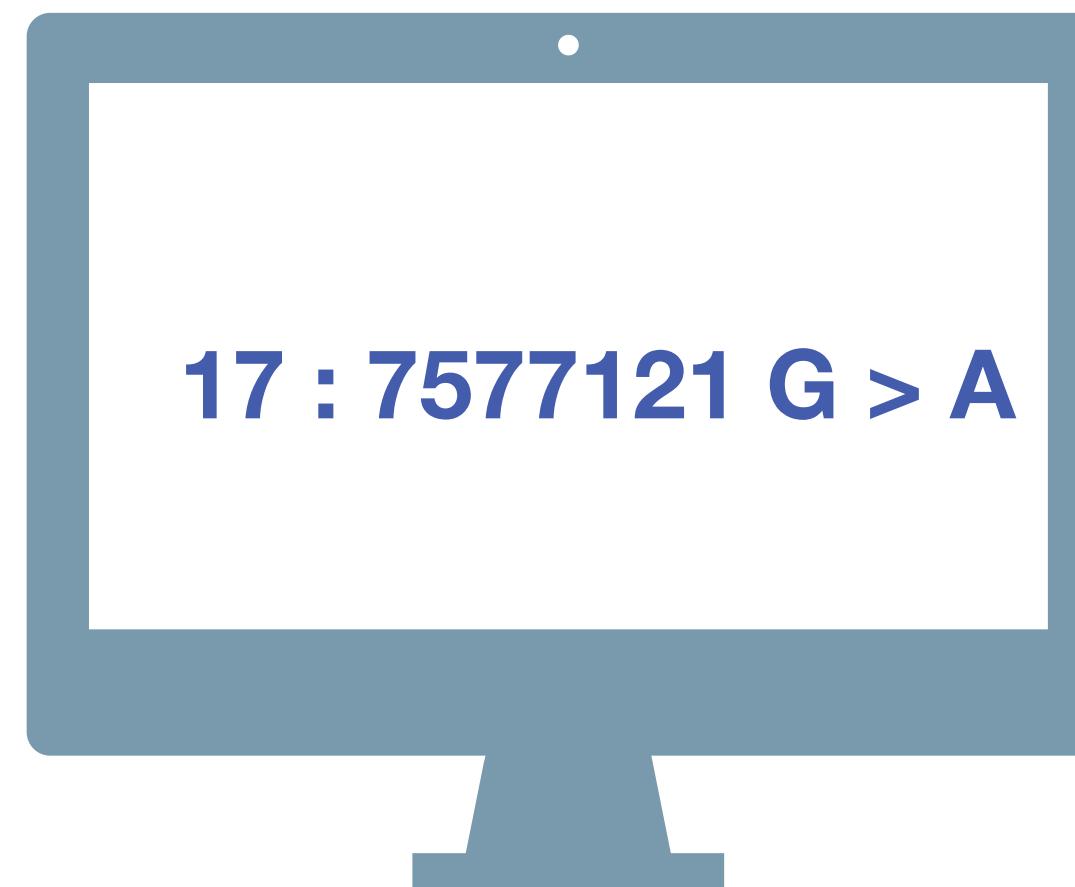
This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	██████████
Bioinformatics analysis	██████████
Biological Sample	██████████
Cohort	██████████
Configuration	██████████
Dataset	██████████
EntryTypes	██████████
Genomic Variants	██████████
Individual	██████████
Info	██████████
Sequencing run	██████████

Matches the Spec Not Match the Spec Not Implemented


Global Alliance
for Genomics & Health



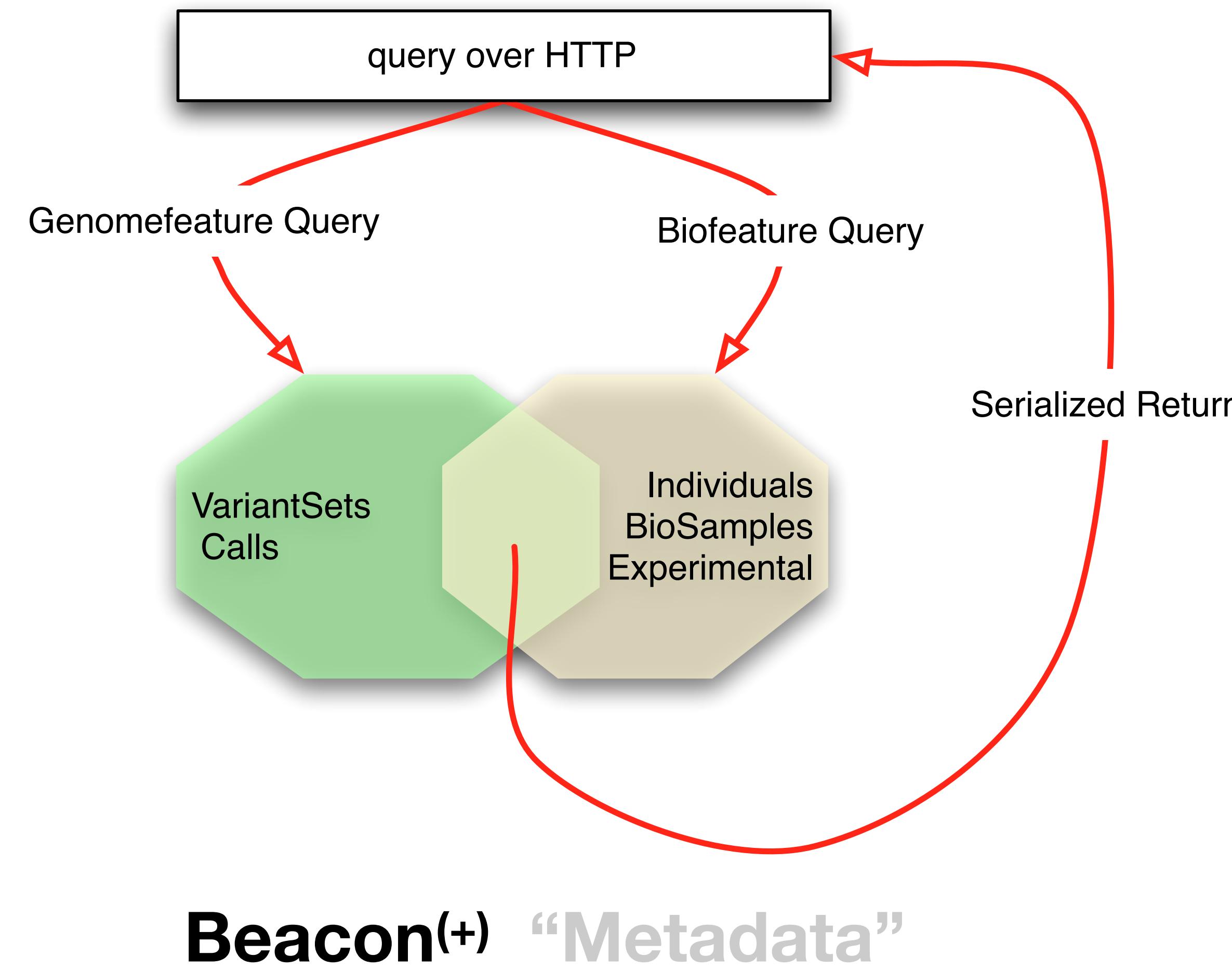


Beacon

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

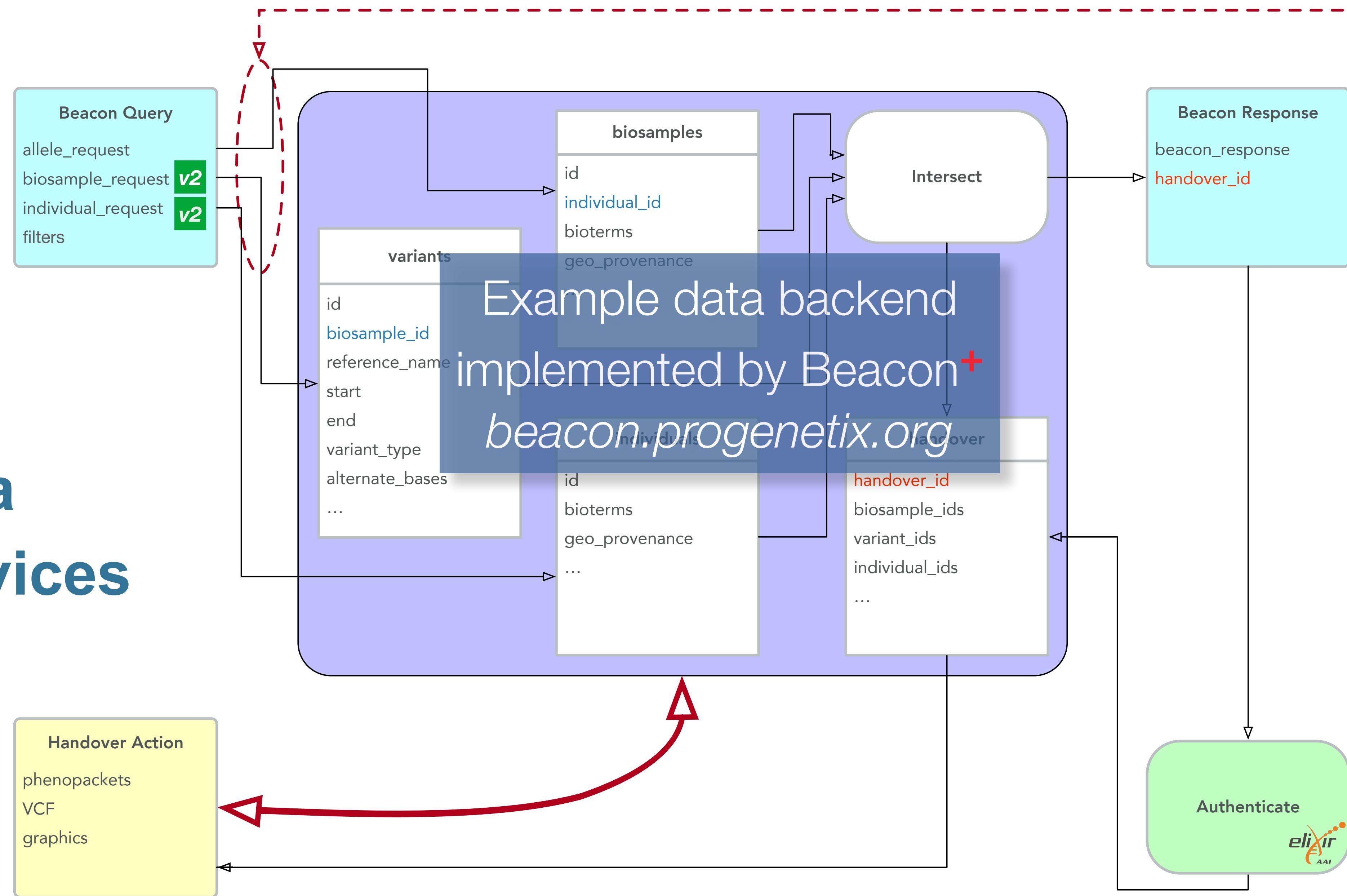
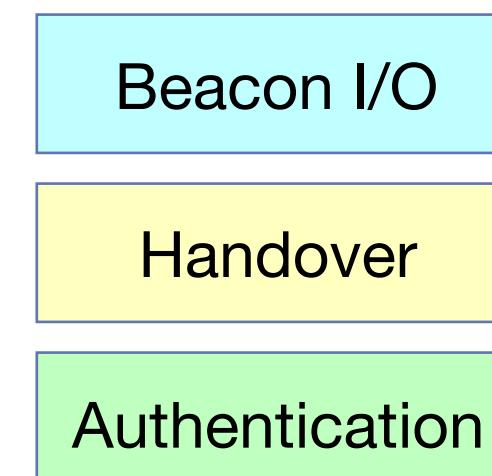
YES | NO | \0

Minimal GA4GH query API structure



Beacon & Handover

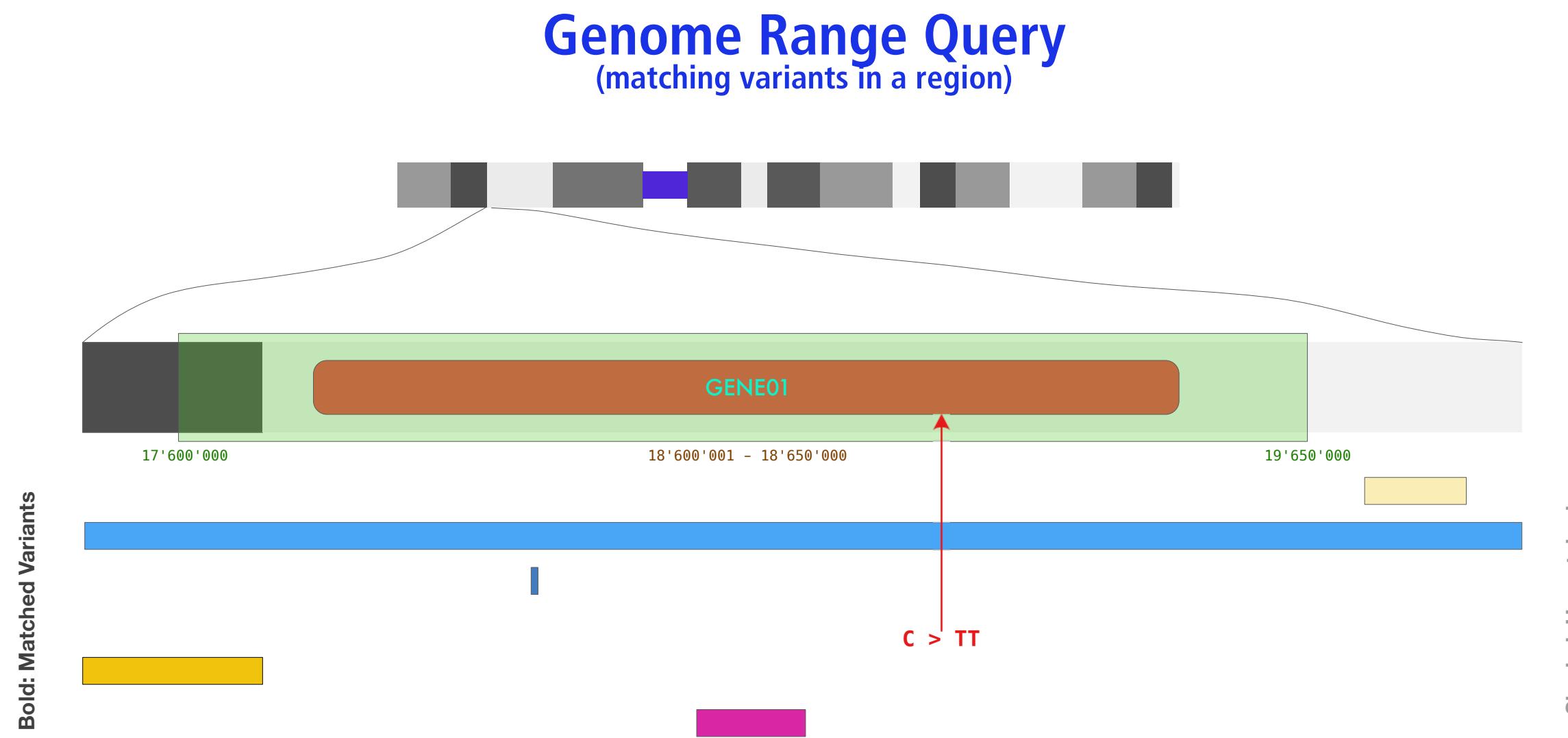
Beacons v1.1
supports data
delivery services



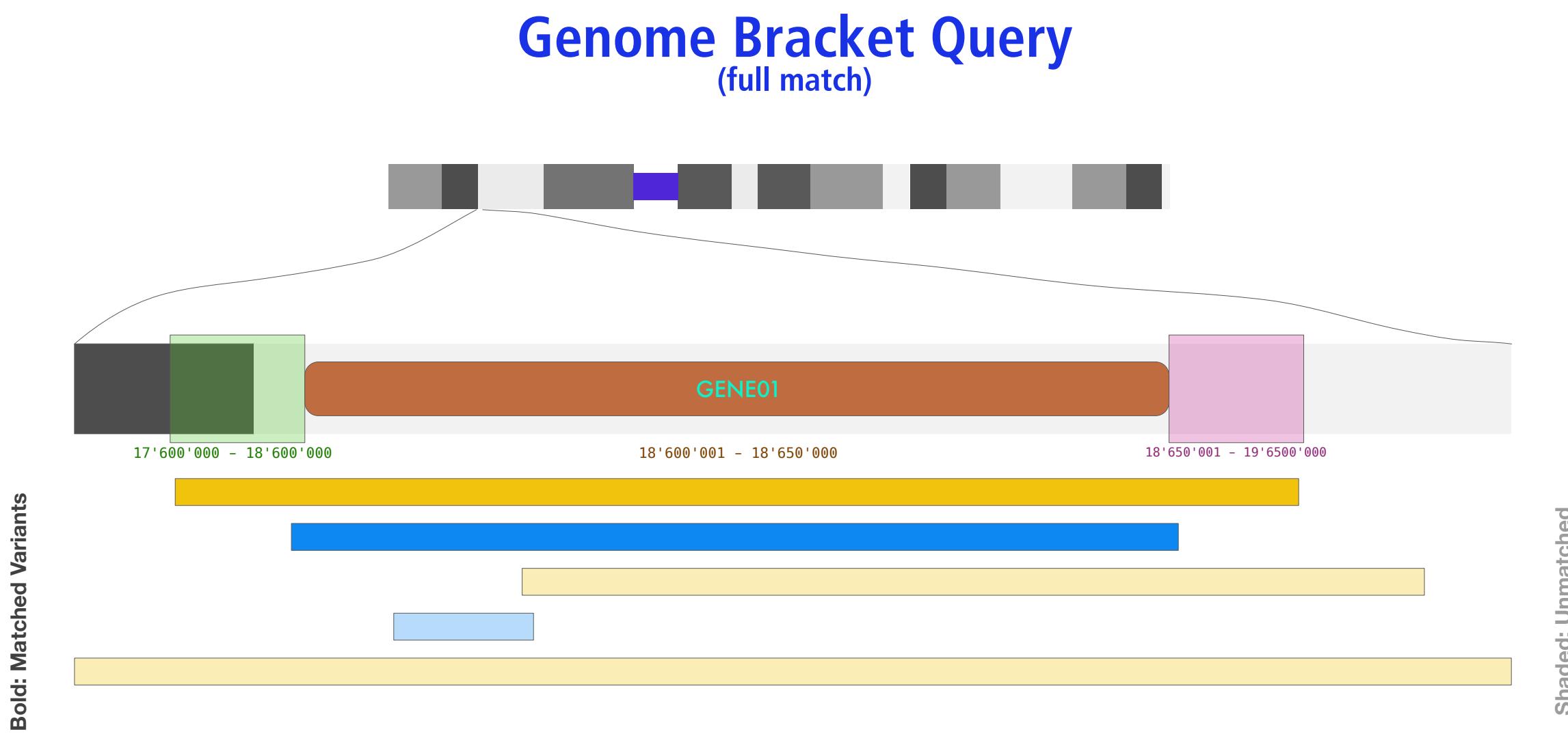
Beacon v2: Extended Variant Queries



Range and Bracket queries enable positional wildcards and fuzziness



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



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

Beacon v2 Paths

Progenetix utilizes Beacon v2 REST paths and GET requests

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



Base `/biosamples`

`/biosamples/` + **query**

- `/biosamples/?filters=cellosaurus:CVCL_0004`

◦ this example retrieves all biosamples having an annotation for the Cellosaurus CVCL_0004 identifier (K562)

`/biosamples/{id}/`

- `/biosamples/pgxbs-kftva5c9/`

◦ retrieval of a single biosample

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

- `/biosamples/pgxbs-kftva5c9/variants/`

- `/biosamples/pgxbs-kftva5c9/variants_in_sample/`

◦ retrieval of all variants from a single biosample

◦ currently - and especially since for a mostly CNV containing resource - `variants` means "variant instances" (or as in the early v2 draft `variantsInSample`)

Base `/variants`

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

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

`/variants/` + **query**

- `/variants/?`

`assemblyId=GRCh38&referenceName=17&variantType=DEL&filterLogic=AND&start=7500000&start=7676592&end=7669607&end=7800000`

◦ This is an example for a Beacon "Bracket Query" which will return focal deletions in the TP53 locus (by position).

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

- `/variants/5f5a35586b8c1d6d377b77f6/`

- `/variants_in_sample/5f5a35586b8c1d6d377b77f6/`

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

- `/variants/5f5a35586b8c1d6d377b77f6/biosamples/`

- `/variants_in_sample/5f5a35586b8c1d6d377b77f6/biosamples/`

Beacon v2 Requests

POSTing Queries

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

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



Progenetix

Genomic resource utilizing Beacon v2 calls

- Progenetix uses Beacon v2 queries to drive its UI
- all individuals, biosamples, variants, analyses matched by a given query are stored by their object ids
- handovers for variant purposes (e.g. to retrieve all matched variants) are returned in the original response and asynchronously retrieved by the front end app

The screenshot shows the Progenetix web application interface. At the top, it displays assembly information: GRCh38 Chro: 9 Start: 21500001-21975098 End: 21967753-22500000, Type: EFO:0030067, and Filters: NCIT:C3058. Below this, a summary box shows Matched Samples: 660, Retrieved Samples: 660, Variants: 279, Calls: 667, and links to UCSC region, Variants in UCSC, Dataset Response (JSON), and Visualization options.

The main content area has tabs for Results, Biosamples, Biosamples Map, Variants, and Annotated Variants. The Results tab is active, showing a bar chart for 'progenetix' with values from -75% to 1%. Below the chart is a table of matched subset codes and their frequencies:

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
UBERON:0002021	14	1	0.029
pgx:icdot-C71.0	14	1	0.029
pgx:icdot-C71.0	14	1	0.029
NCIT:C3796	84	4	0.048
UBERON:0001869	1714	14	0.008
pgx:icdot-C71.0	1714	14	0.008

Below the table are download links: Download Sample Data (JSON) and Download Sample Variants (JSON), both with a count of 1-660.

A yellow callout box highlights a network request: /beacon/biosamples/?requestedGranularity=record&limit=1000&skip=0&assemblyId=GRCh38&referenceName=9&variantType=EFO:0030067&start=21500000,21975098&end=21967753,22500000&filters=NCIT:C3058.

A cyan callout box highlights another network request: /beacon/biosamples/?skip=0&limit=1000&accessid=fbffda57-0f41-4d6a-99fc-41d4cfdea9f6&requestedSchema=biosample.

A pink callout box highlights a third network request: /beacon/genomicVariations/?accessid=e2dadd91-9326-46de-97e4-6b88413b6bfe&requestedSchema=genomicVariant.

The bottom right corner shows a browser developer tools Network tab with several requests listed, including biosamples, genomicVariations, samplePlots.cgi, and collations, with total times of 10.00s, 20.00s, and 30.00s respectively.

Website populated by asynchronous retrieval of Beacon query results using handovers

progenetix

Edit Query

CNV Profiles

- ... by NCIT
- ... by ICD-O Morphology
- ... by ICD-O Site
- ... by TNM & Grade

Search Samples

arrayMap

- TCGA Data
- cBioPortal Studies

Publication DB

- Progenetix Use

NCIT - ICD-O Mappings

- UBERON Mappings

Upload & Plot

OpenAPI Paths and Examples

Cancer Cell Lines

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

Filters: NCIT:C3058

progenetix

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

UCSC region ↗ Geographic Map ↗ Variants in UCSC ↗ Dataset Responses (JSON) ↗ Visualization options

Results Biosamples Variants

progenetix (198 samples)

© CC-BY 2001 - 2025 progenetix.org

Reload histogram in new window ↗

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

progenetix Data Downloads

Download Sample Data (TSV)

Part1 ↗ Part2 ↗ Part3 ↗ Part4 ↗ Part5 ↗

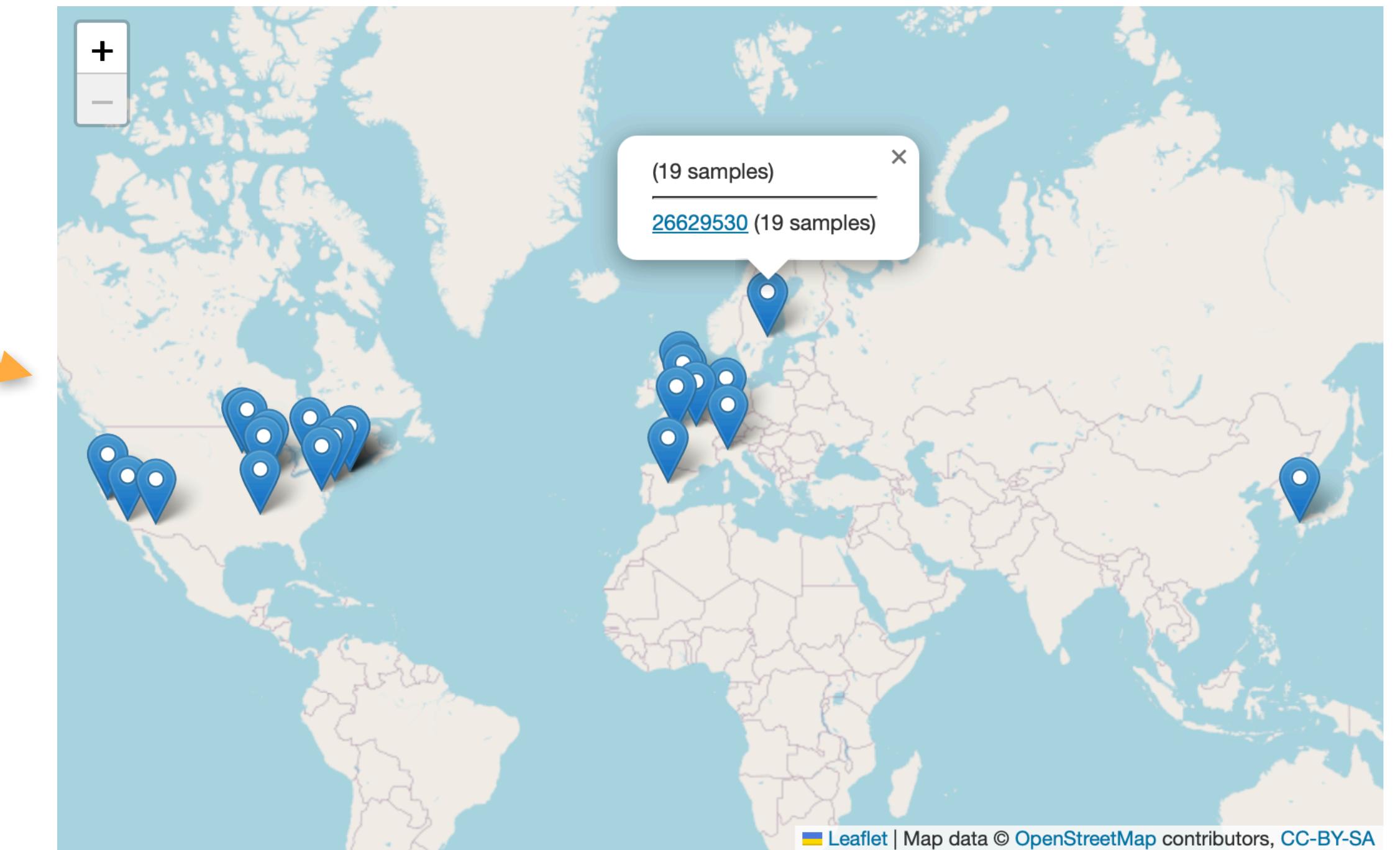
Download Sample Data (JSON)

Part1 ↗ Part2 ↗ Part3 ↗ Part4 ↗ Part5 ↗

Download Variants (Beacon VRS)

Part1 ↗ Part2 ↗ Part3 ↗ Part4 ↗ Part5 ↗

Download Variants (VCF)



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

ga4gh-beacon / **beacon-v2**

Type / to search

Code Issues Pull requests Discussions Actions Projects Security Insights Settings

beacon-v2 Public

Edit Pins Unwatch 10 Fork 22 Starred 32

add-aggregation-resp... 37 Branches 5 Tags Go to file Add file Code About

This branch is 25 commits ahead of main.

#259

mbaudis re-adding distributions 9682bed · 2 days ago 717

.github/workflows adding github actions demo file

bin fixes for aggregation PR

docs fixes for aggregation PR

framework re-adding distributions

models measures => measurements re-fix (this branch)

.gitattributes re-structuring intro pages

.gitignore Fix file naming conflict error in schemas-md on macOS A...

CHANGELOG.md Merge branch 'main' into schema-urgent-fixes

LICENSE Initial commit

README.md Merge branch 'develop' into develop_changelog

mkdocs.yaml some v2 naming/version use cleanup

requirements.txt switch to mermaid2 plugin

README License Security

Unified repository for Beacon Code & Documentation

progenetix / **bycon**

Type / to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

bycon Public

Edit Pins Unwatch 3 Fork 10 Starred 8

main 10 Branches 49 Tags Go to file Add file Code About

mbaudis Merge pull request #44 from mbaudis/main 68ee58b · last month 931 Commits

.github/workflows docs & formatting last month

beaconServer 2.4.3 "Bologna" 4 months ago

beaconplusWeb vrsifier and vrs format last month

bycon VCF sequence fix; some clean-up last month

byconServices going VRSv2 alpha last month

docs VCF sequence fix; some clean-up last month

housekeepers going VRSv2 alpha last month

importers refactor importers 2 months ago

local vrsifier and vrs format last month

rsrc going VRSv2 alpha last month

tests going VRSv2 alpha last month

.gitignore 2.1.2 9 months ago

LICENSE Create LICENSE 5 years ago

README.md export tables last month

install.py 2.4.9 2 months ago

markdowner.py v2.4.7 "Thessaloniki" 3 months ago

About

Bycon - A Python Based Beacon API (beacon-project.io) implementation leveraging the Progenetix (progenetix.org) data model

Readme

CC0-1.0 license

Activity

Custom properties

8 stars

3 watching

10 forks

Report repository

Releases 15

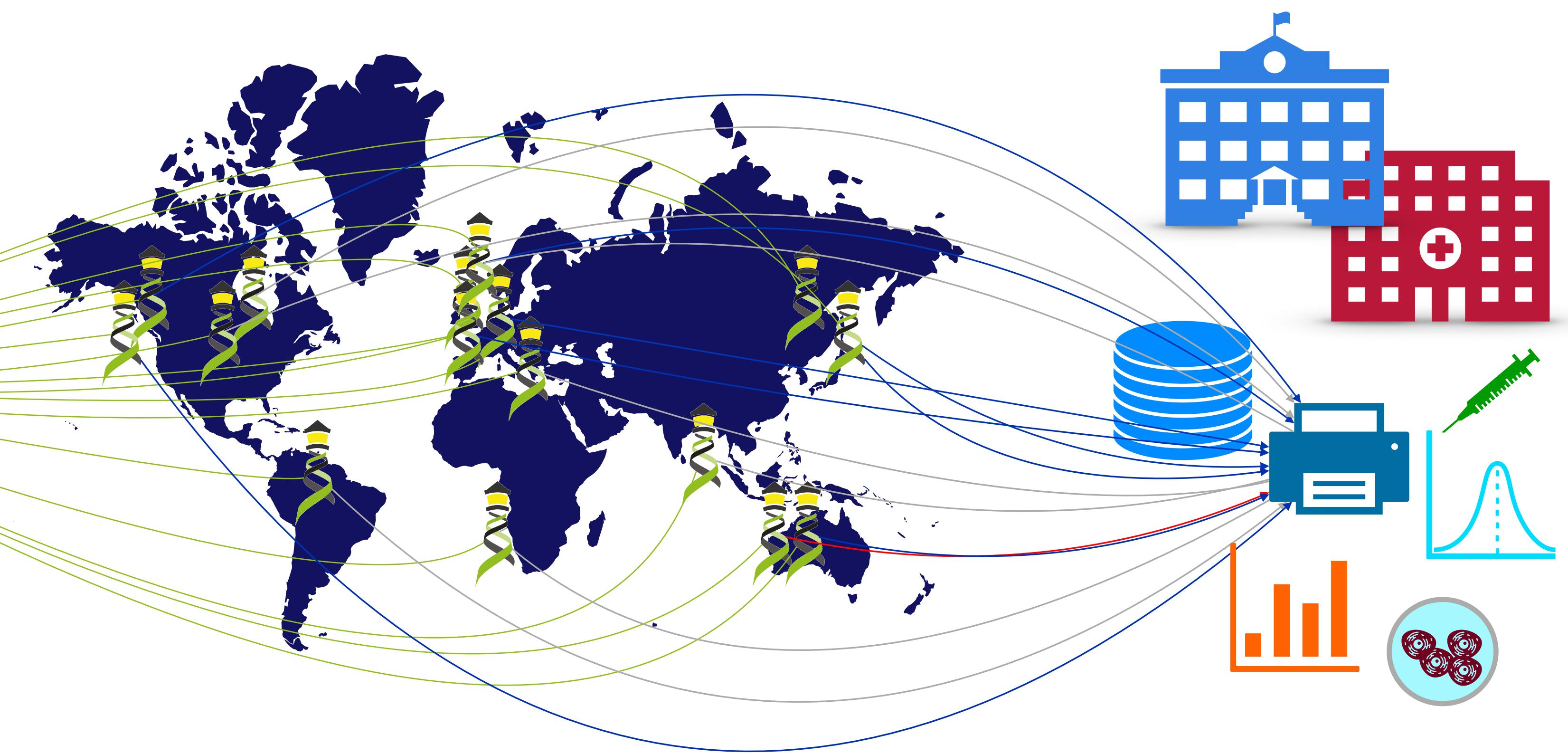
v2.5.0 "Forked" Latest on Jul 30

+ 14 releases

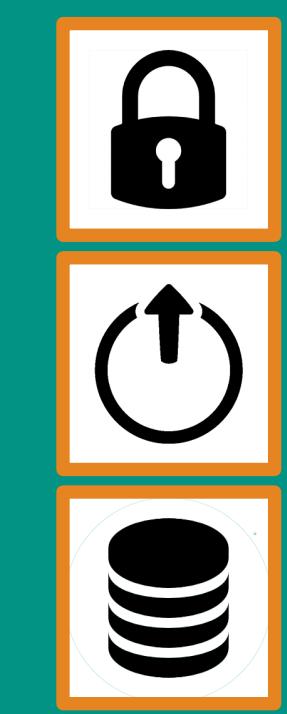
Packages

No packages published Publish your first package

Contributors 6

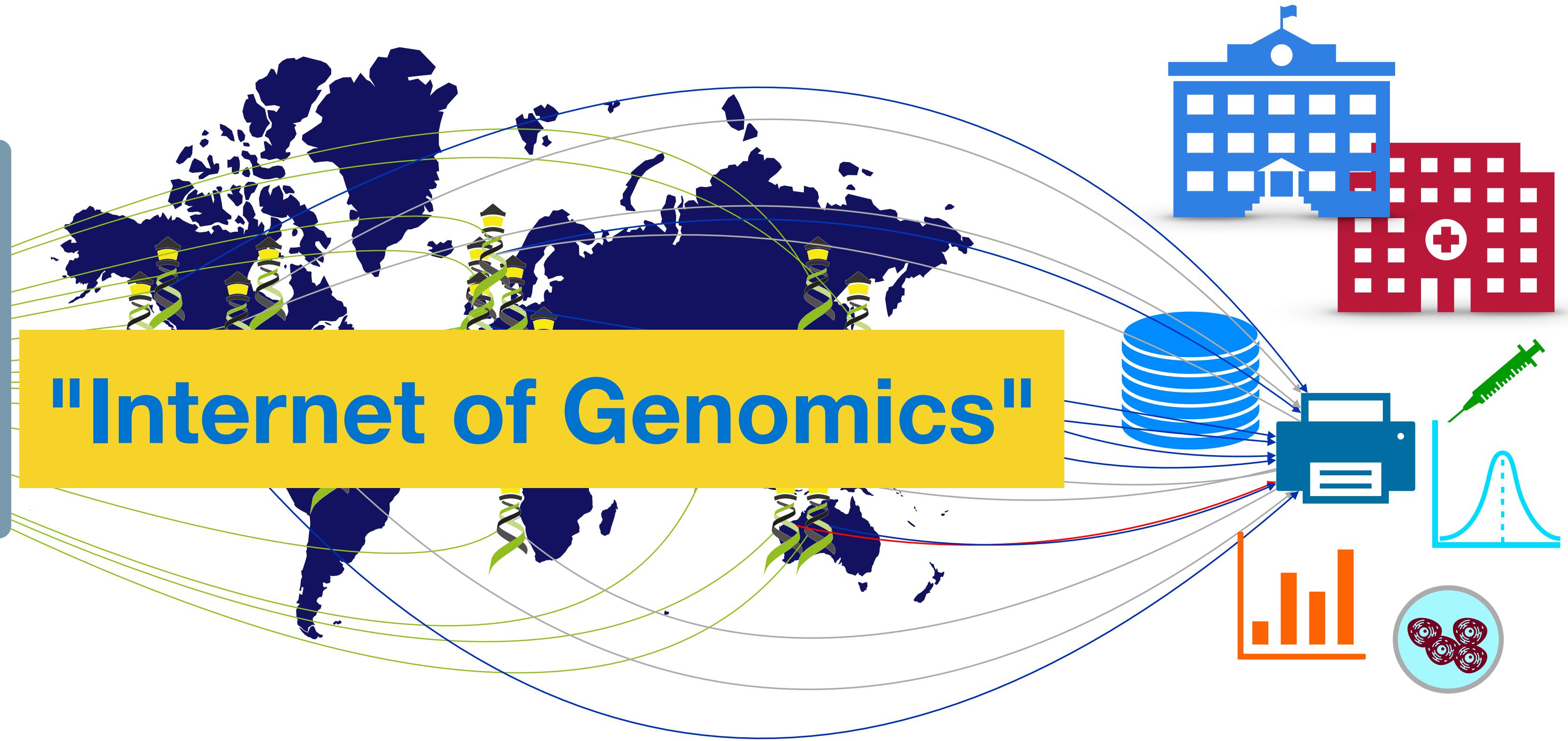


Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

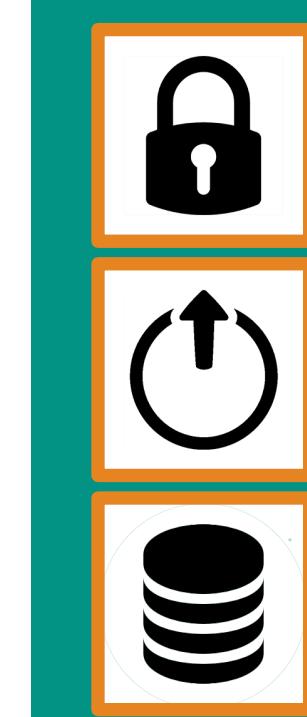


Beacon API

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



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



Beacon API

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



Making use of Progenetix' Beacon API

Data analysis through integration with R



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

All users
Interface

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

Output

```
"results": [
  {
    "caseLevelData": [
      {
        "analysisId": "pgxcs-kftvu6cg",
        "biosampleId": "pgxbs-kftvh94d",
        "id": "pgxvar-5bab5837727983b2e0121e97"
      }
    ],
    "variantInternalId": "11:0-134452384:DEL",
    "variation": {
      "copyChange": "efo:0030067",
      "identifiers": {},
      "subject": {
        "interval": {
          "end": {
            "type": "Number",
            "value": 134452384
          },
          "start": {
            "type": "Number",
            "value": 0
          }
        },
        "type": "SequenceInterval"
      },
      "sequence_id": "refseq:NC_000011.10",
      "type": "SequenceLocation"
    },
    "variantAlternativeIds": []
  },
  {
    "caseLevelData": [
      {
        "analysisId": "pgxcs-kftvu6cg",
        "biosampleId": "pgxbs-kftvh94d",
        "id": "pgxvar-5bab5837727983b2e0121e99"
      }
    ],
    "variantInternalId": "1:0-84699999:DEL",
    "variation": {
      "copyChange": "efo:0030067",
      "identifiers": {},
      "subject": {
        "interval": {
          "end": {
            "type": "Number",
            "value": 84699999
          },
          "start": {
            "type": "Number",
            "value": 0
          }
        },
        "type": "SequenceInterval"
      },
      "sequence_id": "refseq:NC_000001.11",
      "type": "SequenceLocation"
    }
  }
]
```



R users
R Studio®

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

	variant_id	biosample_id	analysis_id	reference_genome	variant
1	pgxvar-5bab5837727983b2e0121e99	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000001.11	1:0-84699999:DEL
2	pgxvar-5bab5837727983b2e0121e9a	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000001.11	1:124300000-247249719:DEL
3	pgxvar-5bab5837727983b2e0121e9c	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000002.12	2:12800000-61099999:DEL
4	pgxvar-5bab5837727983b2e0121e9d	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000002.12	2:197100000-242951149:DEL
5	pgxvar-5bab5837727983b2e0121e94	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000003.12	3:14700000-71799999:DEL
6	pgxvar-5bab5837727983b2e0121e8d	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000004.12	4:35500000-191273063:DUP
7	pgxvar-5bab5837727983b2e0121e8e	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000005.10	5:18500000-143099999:DUP
8	pgxvar-5bab5837727983b2e0121e91	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000006.12	6:0-60499999:DEL
9	pgxvar-5bab5837727983b2e0121e92	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000006.12	6:130400000-170899992:DEL

pgxRpi

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

Metadata query

[https://progenetix.org/beacon/individuals/?
filters=NCIT:C3697](https://progenetix.org/beacon/individuals/?filters=NCIT:C3697)

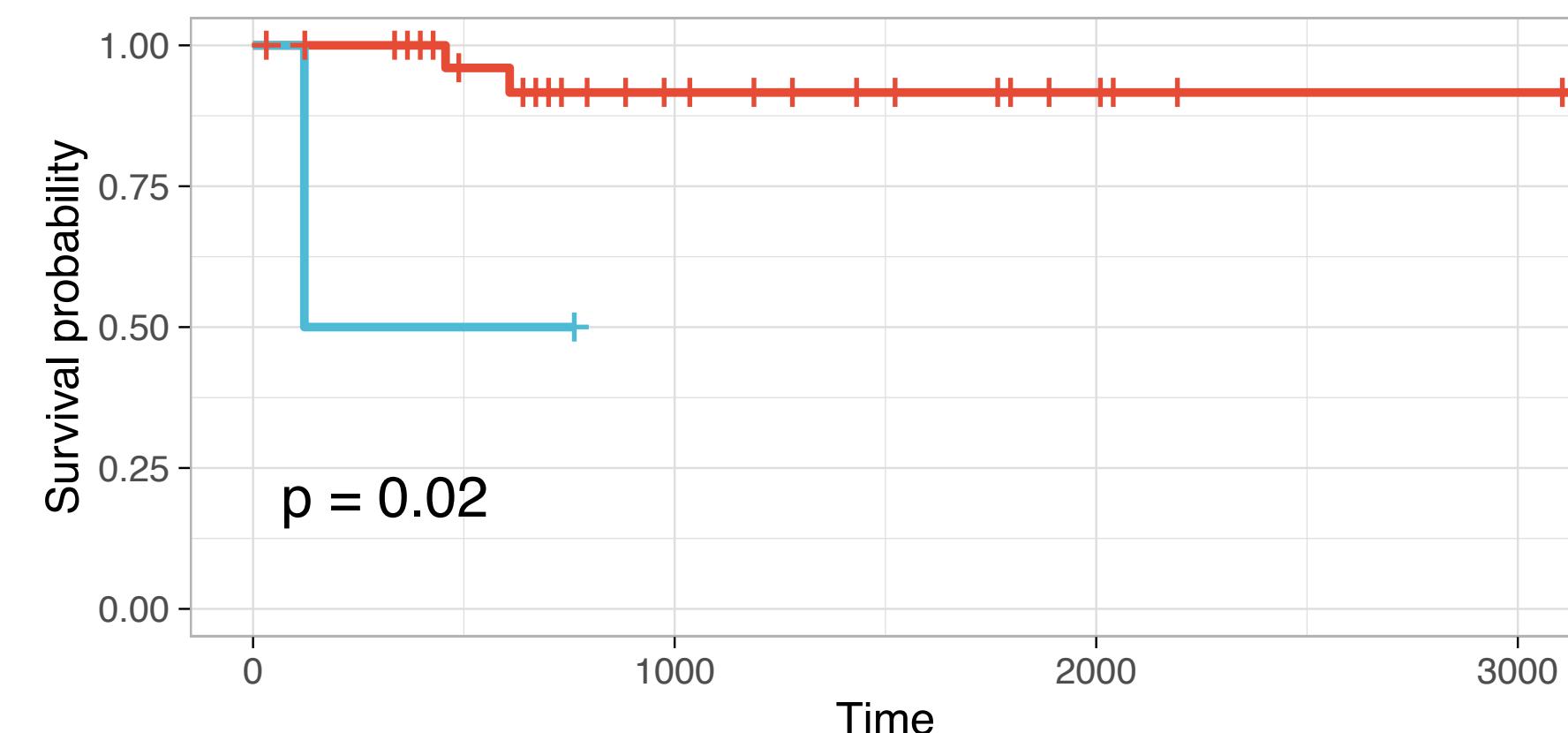
Output

```
{
  "description": null,
  "id": "pgxind-kftx359j",
  "indexDisease": {
    "clinicalTnmFinding": [],
    "diseaseCode": {
      "id": "NCIT:C3697",
      "label": "Myxopapillary Ependymoma"
    },
    "followupState": {
      "id": "EFO:0030041",
      "label": "alive (follow-up status)"
    },
    "followupTime": "P178M",
    "onset": {
      "age": "P16Y",
      "ageDays": 5843.88
    },
    "stage": {
      "id": "NCIT:C92207",
      "label": "Stage Unknown"
    }
  },
  "info": {
    "legacyIds": [
      "PGX_IND_Epend-car-01"
    ]
  },
  "provenance": {
    "geoLocation": {
      "geometry": {
        "coordinates": [
          -1.4,
          50.9
        ],
        "type": "Point"
      },
      "properties": {
        "city": "Southampton",
        "continent": null,
        "country": "United Kingdom",
        "latitude": 50.9,
        "longitude": -1.4,
        "precision": "city"
      },
      "type": "Feature"
    }
  },
  "sex": {
    "id": "PATO:0020001",
    "label": "male genotypic sex"
  },
  "updated": "2018-09-26 09:51:34.766000",
  "vitalStatus": {
    "status": "ALIVE",
    "survivalTimeInDays": 5384
  }
},
```

individuals <- pgxLoader(type='individual', filters='NCIT:C3697')

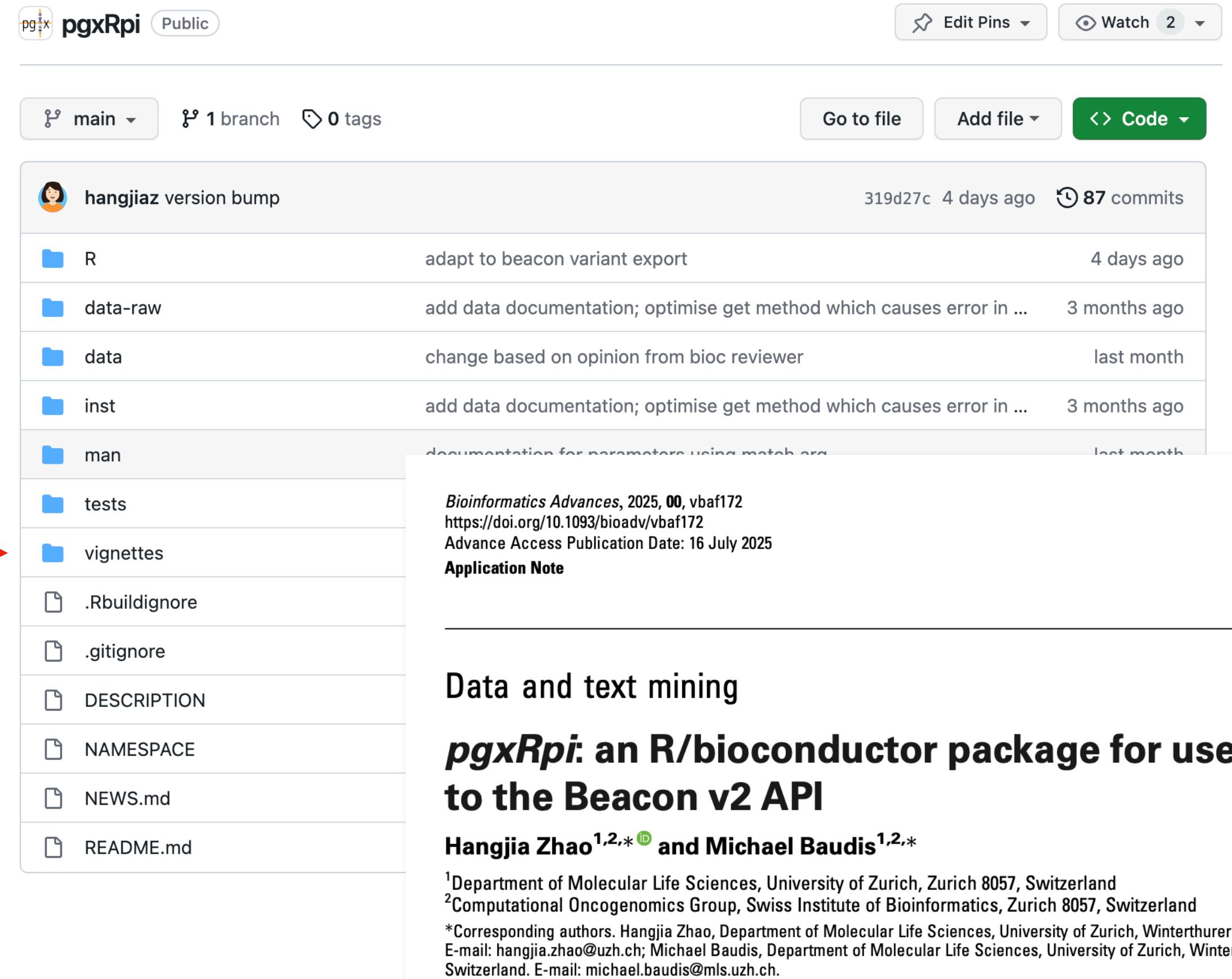
individual_id	sex_label	age_iso	histological_diagnosis_id	index_disease_followup_time	index_disease_followup_state_label
pgxind-kftx359j	male genotypic sex	P16Y	NCIT:C3697	P178M	alive (follow-up status)
pgxind-kftx35a0	male genotypic sex	P23Y	NCIT:C3697	P115M	alive (follow-up status)
pgxind-kftx35aa	male genotypic sex	P15Y	NCIT:C3697	P114M	alive (follow-up status)
pgxind-kftx35ac	male genotypic sex	P24Y	NCIT:C3697	P30M	alive (follow-up status)
pgxind-kftx35ai	female genotypic sex	P44Y	NCIT:C3697	P101M	alive (follow-up status)
pgxind-kftx35as	male genotypic sex	P50Y	NCIT:C3697	P331M	dead (follow-up status)
pgxind-kftx35bb	male genotypic sex	P28Y	NCIT:C3697	P48M	alive (follow-up status)

Strata — group_id=NCIT:C27243 + group_id=NCIT:C40359



pgxRpi

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



pgxRpi Public

Edit Pins Watch 2

main 1 branch 0 tags

Go to file Add file Code

hangjiaz version bump 319d27c 4 days ago 87 commits

R adapt to beacon variant export 4 days ago

data-raw add data documentation; optimise get method which causes error in ... 3 months ago

data change based on opinion from bioc reviewer last month

inst add data documentation; optimise get method which causes error in ... 3 months ago

man documentation for parameters using match_... last month

Bioinformatics Advances, 2025, 00, vba172
https://doi.org/10.1093/bioadv/vba172
Advance Access Publication Date: 16 July 2025

Application Note

OXFORD

Data and text mining

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

Hangjia Zhao^{1,2,*} and Michael Baudis^{1,2,*}

¹Department of Molecular Life Sciences, University of Zurich, Zurich 8057, Switzerland

²Computational Oncogenomics Group, Swiss Institute of Bioinformatics, Zurich 8057, Switzerland

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

2 Retrieve metadata of samples

2.1 Relevant parameters

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

2.2 Search by filters

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

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

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

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

```
biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3512")
# data looks like this
biosamples[c(1700:1705),]
#>   biosample_id group_id group_label individual_id callset_ids
#> 1700 pgxbs-kftvjjhx NA NA pgxind-kftx5fyd pgxcs-kftwjevi
#> 1701 pgxbs-kftvjjhz NA NA pgxind-kftx5fyf pgxcs-kftwjew0
#> 1702 pgxbs-kftviji1 NA NA pgxind-kftx5fyh pgxcs-kftwjewi
#> 1703 pgxbs-kftvjjn2 NA NA pgxind-kftx5g4r pgxcs-kftwjg5r
#> 1704 pgxbs-kftvjjn4 NA NA pgxind-kftx5g4t pgxcs-kftwjg6q
#> 1705 pgxbs-kftvjjn5 NA NA pgxind-kftx5g4v pgxcs-kftwjg78
```

Components of an Online Bioinformatics Resource

Going Full Stack?

Components of an Online Bioinformatics Resource

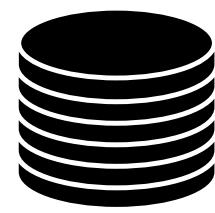
A Stack to work with/through

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

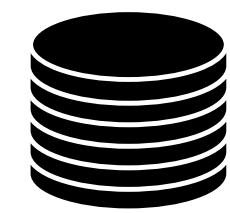
bycon based Beacon+ Stack

progenetix

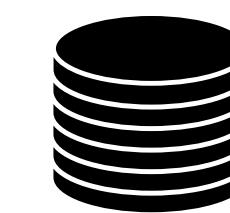
- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - ▶ [pubmed:10027410](#), [NCIT:C3222](#), [pgx:cohort-TCGA](#), [pgx:icdom-94703...](#)
 - ▶ precomputed frequencies per collection informative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



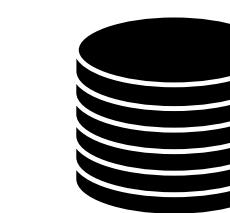
variants



analyses



biosamples



individuals



collations



geolocs



genespans

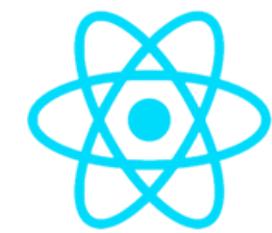


qBuffer

Entity collections

Utility collections

github.com/progenetix/bycon



React



Last but NOT Least...

Documentation is, actually, rather important

Documentation Strategies (Not so) Best Practices

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

- What is documentation? I'll remember this! ↗ ↘ ↙ ↚
 - Just email me if help is needed, unexpectedly
 - We had money for a chat bot.
 - Clean code documents itself - Just use explicit variable/function names.
 - Clean code documents itself - Never use explicit variable/function names.
 - Perl POD it is. There is a command to show the notes in your terminal...
 - I wrote a paper about the resource. In 2001.
 - Haven't you found the GoogleGroups account?
 - Documentation? StackOverflow, whelp!

`normalize_variant_values_for_export(v, by)`

BIOINFORMATICS APPLICATIONS NOTE

 Progenetix.net: an online repository of molecular cytogenetic aberrations in cancer

Michael Baudis^{1, 2,*} and Michael L. Cleary¹

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



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

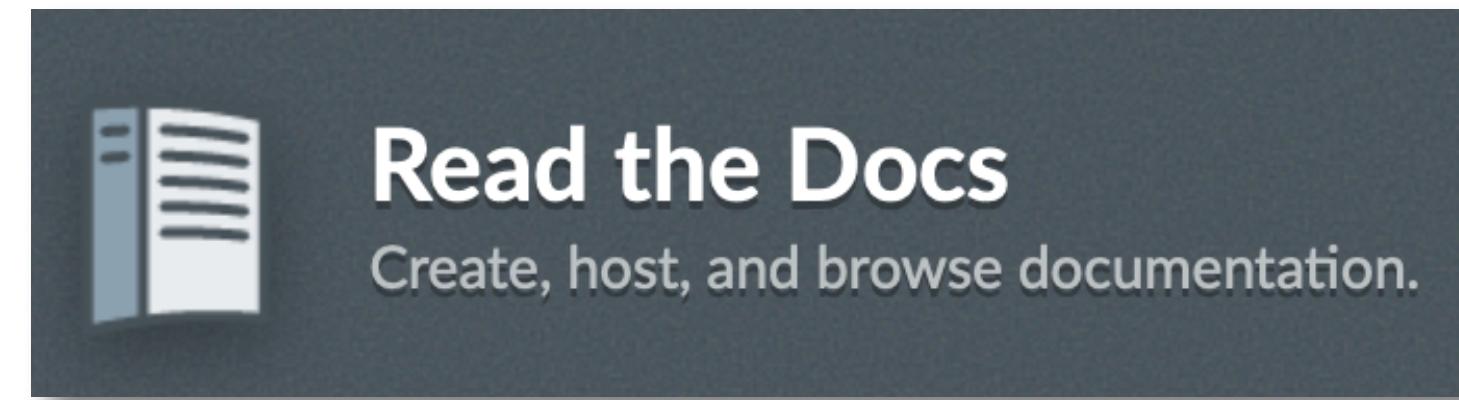
Michael Baudis^{1, 2,*} and *Michael L. Cleary*²

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

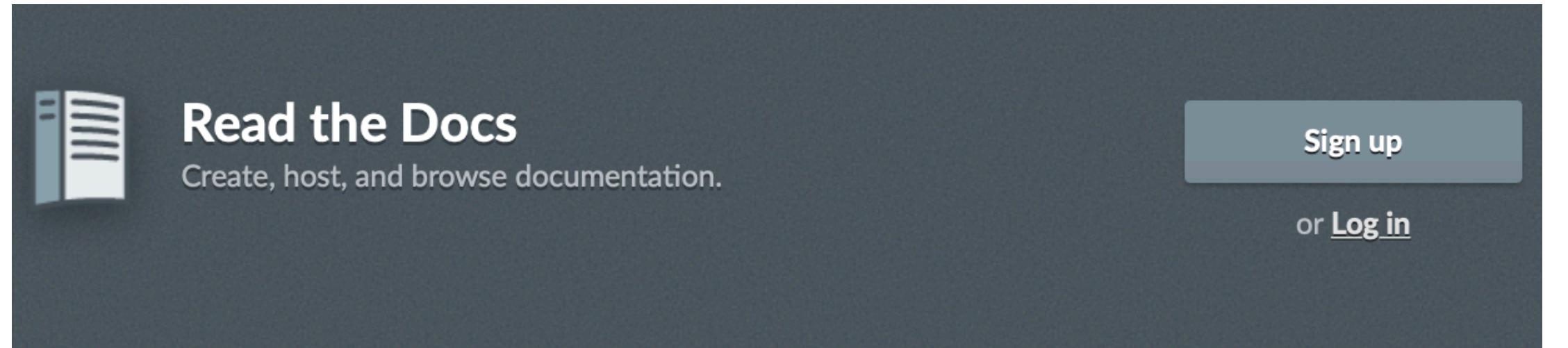
Documentation Strategies

Currently en Vogue

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



Documentation Strategies



The screenshot shows the Read the Docs homepage. It features a dark header with the "Read the Docs" logo and the tagline "Create, host, and browse documentation." Below the header is a "Sign up" button and a "Log in" link. A sidebar on the right contains links to "Introduction," "Terminology & Information Model," "Schema," "Implementation Guide," "Releases," and "Appendices." At the bottom of the page is a promotional banner for the Malala Fund.

Technical documentation lives here

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

Free docs hosting for open source

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

Always up to date

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

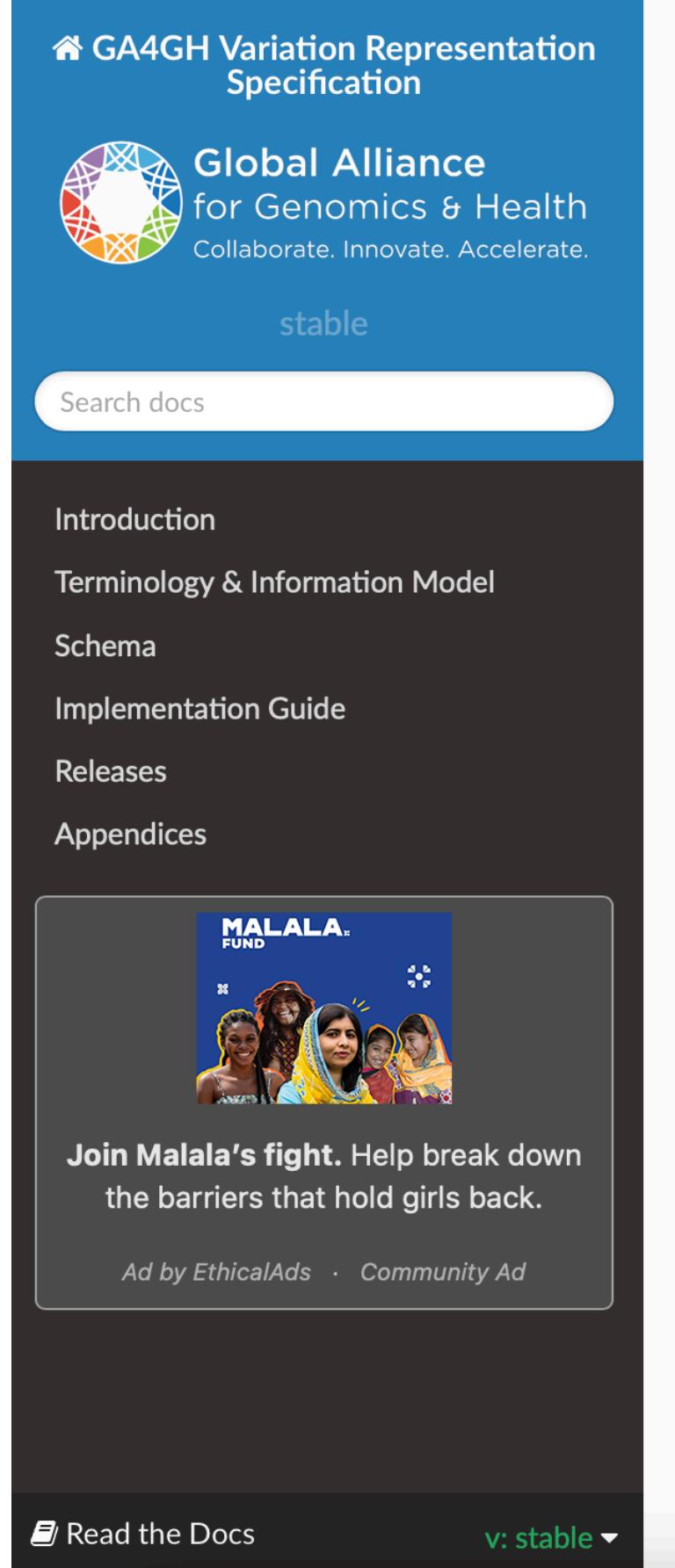
Downloadable formats

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

Multiple versions

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

Example: GA4GH Variation Representation Standard ->



The screenshot shows the GA4GH Variation Representation Specification documentation. It features a blue header with the "GA4GH Variation Representation Specification" logo and the "Global Alliance for Genomics & Health" logo. Below the header is a search bar and a sidebar with links to "Introduction," "Terminology & Information Model," "Schema," "Implementation Guide," "Releases," and "Appendices." At the bottom of the page is a footer with the "Read the Docs" logo and a "v: stable" link.

GA4GH Variation Representation Specification

The Variation Representation Specification (VRS, pronounced “verse”) is a standard developed by the Global Alliance for Genomic Health to facilitate and improve sharing of genetic information. The Specification consists of a JSON Schema for representing many classes of genetic variation, conventions to maximize the utility of the schema, and a Python implementation that promotes adoption of the standard.

Citation

The GA4GH Variation Representation Specification (VRS): a computational framework for variation representation and federated identification. Wagner AH, Babb L, Alterovitz G, Baudis M, Brush M, Cameron DL, ..., Hart RK. *Cell Genomics*. Volume 1 (2021). doi:10.1016/j.xgen.2021.100027

- [Introduction](#)
- [Terminology & Information Model](#)
 - [Information Model Principles](#)
 - [Variation](#)
 - [Locations and Intervals](#)
 - [Sequence Expression](#)
 - [Feature](#)
 - [Basic Types](#)

Output

		on Jan 29	History
 ahwagner	add docs	...	
	_static	Use shared metaschema tooling (#354)	13 months ago
	appendices	remove reference to develop branch (#344)	14 months ago
	images	Closes #324: Removed Abundance from current schema; re-implemente...	14 months ago
	impl-guide	fix link to Data Proxy class	14 months ago
	releases	Closes #320: Add note about attributes that permit identifiable and n...	17 months ago
	conf.py	Closes #345: Fix sphinx theming (#346)	14 months ago
	defs	Use shared metaschema tooling (#354)	13 months ago
	index.rst	update citation	
	introduction.rst	update doc urls to use vrs.ga4gh.org	

Source

2 years ago

FOLDERS

- progenetix-web
 - .github
 - .next
 - docs
 - css
 - img
 - javascripts
 - news
 - beaconplus.md
 - changelog.md
 - classifications-an
 - CNAME
 - index.md
 - progenetix-data-r
 - progenetix-websi
 - publication-colle
 - services.md
 - technical-notes.m
 - ui.md
 - use-cases.md
- extra
- node_modules
- out
- public
- src
 - .babelrc
 - .env.development
 - .env.production
 - .eslintrc.json
 - .gitignore
 - .prettierrc
 - .jest.config.js
 - mkdocs.yaml
 - next.config.js
 - package-lock.json
 - package.json
- README.md

MkDocs & Material for MkDocs & Github Actions

```

1 | site_name: Progenetix Documentation
2 | site_description: 'Documentation for the Progenetix oncogen
3 | site_author: Michael Baudis
4 | copyright: '&copy; Copyright 2022, Michael Baudis and proge
5 | repo_name: 'progenetix-web'
6 | repo_url: https://github.com/progenetix/progenetix-web
7 |
8 ######
9
10 nav:
11   - Documentation Home: index.md
12   - News & Changes: news
13   - Pages & Forms: ui
14
15
16
17
18   - Publication Collection: publication-collection
19   - Data Review: progenetix-data-review
20   - Technical Notes: technical-notes
21   - Progenetix Website Builds: progenetix-website-builds
22   - Progenetix Data : http://progenetix.org
23   - Baudisgroup @ UZH : http://info.baudisgroup.org
24
25 #####
26
27 markdown_extensions:
28   - toc:
29     toc_depth: 2-3
30     permalink: true
31   - admonition
32   - attr_list
33   - footnotes
34   - md_in_html
35   - pymdownx.critic
36   - pymdownx.caret
37   - pymdownx.details
38   - pymdownx.keys
39   - pymdownx.magiclink:
40     hide_protocol: true
41   - pymdownx.mark
42   - pymdownx.tilde
43   - pymdownx.saneheaders

```

```

1 | # Classifications, Ontologies and Standards
2 |
3 | The Progenetix resource utilizes standardized diagnostic coding systems, with a
4 | move towards hierarchical ontologies. As part of the coding process we have
5 | developed and provide several code mapping resources through repositories, the
6 | Progenetix website and APIs.
7 |
8 | Additionally to diagnostic and other clinical concepts, Progenetix increasingly
9 | uses hierarchical terms and concepts for the annotation and querying of technical
10 | parameters such as platform technologies. Overall, the Progenetix resource uses a
11 | query syntax based around the [Beacon v2 "filters"](https://beacon-project.io/v2/filters.html) concept with a [CURIE](https://www.w3.org/TR/2010/NOTE-curie-20101216/)
12 | based syntax
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

```

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

#####

30 | #### Private filters
31 |
32 | Since some classifications cannot directly be referenced, and in accordance with
33 | the upcoming Beacon v2 concept of "private filters", Progenetix uses
34 | additionally a set of structured non-CURIE identifiers.

Local Testing

```

FOLDERS
progenetix-web
  .github
  .next
  docs
  css
mkdocs.yaml
  1 | site_
  2 | site_
  3 | site_
  4 | copyr
  5 | repo_name: 'progenetix-web'
  6 | repo_url: https://github.com/progenetix/progenetix-web

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

```

Web Deployment (Github)

the Progenetix oncogenes and their role in cancer development
Michael Baudis and progenetix.org

```

classifications-and-ontologies.md
# Classification and Ontology
The Progenetix team is moving towards a more modular and flexible architecture. This involves
decentralizing certain components and creating a more dynamic system for managing data and
processes. One key aspect of this transition is the use of GitHub Actions to handle deployment
and automation tasks. In this section, we will discuss the workflow setup and how it enables
efficient and reliable deployment of the Progenetix website and documentation.

## Workflow Setup
The Progenetix GitHub repository contains a workflow named 'mk-progenetix-docs' defined in 'mk-progenetix-docs.yaml'. This workflow is triggered by pushes to the 'main' branch and consists of several steps:
1. **refseq_ids_in_examples_aggregator_start**: A step that starts the aggregator process for refseq IDs in examples.
2. **Update_VariantsDataTable_js**: A step that updates the VariantsDataTable.js file.
3. **Update_VariantsDataTable_js**: Another step that updates the VariantsDataTable.js file.

## Workflow Details
The 'mk-progenetix-docs' workflow is triggered by pushes to the 'main' branch. It includes management, caches, and deployments. The workflow runs 178 times. Recent runs include:
- 'refseq_ids_in_examples_aggregator_start' (3 days ago)
- 'Update_VariantsDataTable_js' (11 days ago)
- 'Update_VariantsDataTable_js' (16 days ago)

## Contributors
The workflow is maintained by mbaudis, who has cleaned up the code. There is 1 contributor: mbaudis.
```

progenetix.org/services/ids/geo:GSE13547 (GSM491153) | arrayexpress | EBI cellosaurus | Cellosaurus:CVCL_1655 UBERON | UBERON cbioportal | cBioPortal //progenetix.org/services/ids/geo:MEXP-1008 | impact_2017](http://progenetix.org/services/ids/geo:MEXP-1008|^5] |

Private filters

Since some classifications and ontologies are not yet available in the upcoming Beacon API, the team has developed a set of structured non-CURIE identifiers. These identifiers are used to map between different classification systems and ensure consistency across the platform.

**Progenetix Documentation**[Documentation Home](#)[News & Changes](#)[Pages & Forms](#)[Services API](#)[Beacon+ API & bycon](#)[Use Case Examples](#)[Classifications, Ontologies & Standards](#)[Publication Collection](#)[Data Review](#)[Technical Notes](#)[Progenetix Website Builds](#)[Progenetix Data ↗](#)[Baudisgroup @ UZH ↗](#)

Classifications, Ontologies and Standards



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

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

Table of contents

List of filters recognized by different query endpoints

[Public Ontologies with CURIE-based syntax](#)

[Private filters](#)

[Diagnoses, Phenotypes and Histologies](#)

[NCIt coding of tumor samples](#)

[ICD coding of tumor samples](#)

[UBERON codes](#)

[Genomic Variations \(CNV Ontology\)](#)

[Geolocation Data](#)

[Provenance and use of geolocation data](#)

List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

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

Documentation Strategies

Best Practices

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



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

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

Progenetix as Example Genomics Resource

Some trajectories ...

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



(Bio)informatics Skill Set

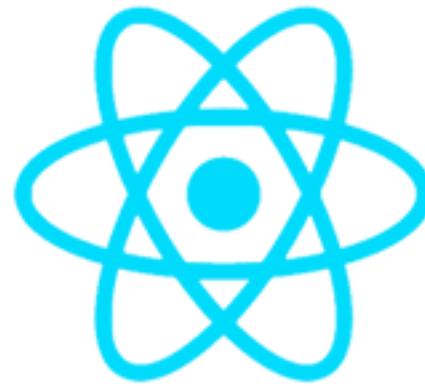
What has been needed to develop & maintain progenetix.org?

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

(Bio)informatics Skill Set

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

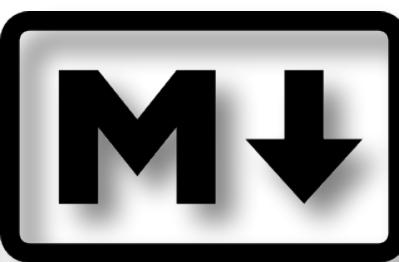
text mining



React



regular expressions
s/knowledge/mastery/



MkDocs

Project documentation with Markdown.



array & sequencing pipelines



Master Project in Data Wrangling? Ask!