

Building a Genomics Resource

From Experiments to APIs

Michael Baudis | UZH BIO390 HS22



Building a Genomics Resource

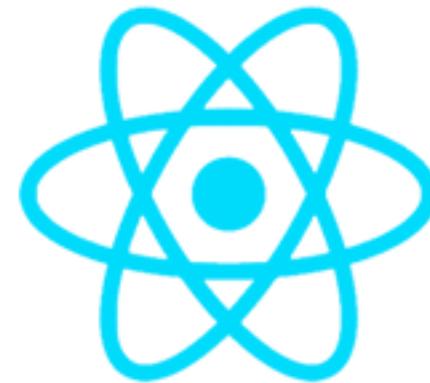
A (personal) journey through time...

- Genomic Copy Number Variations in cancer (CNA / CNV)
- Comparative Genomic Hybridization (CGH) as the original CNV screening technique
- CNVs differ between cancer (sub)types and may correlate to clinical outcome
- single studies are limited in understanding disease-specific changes - 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

(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.



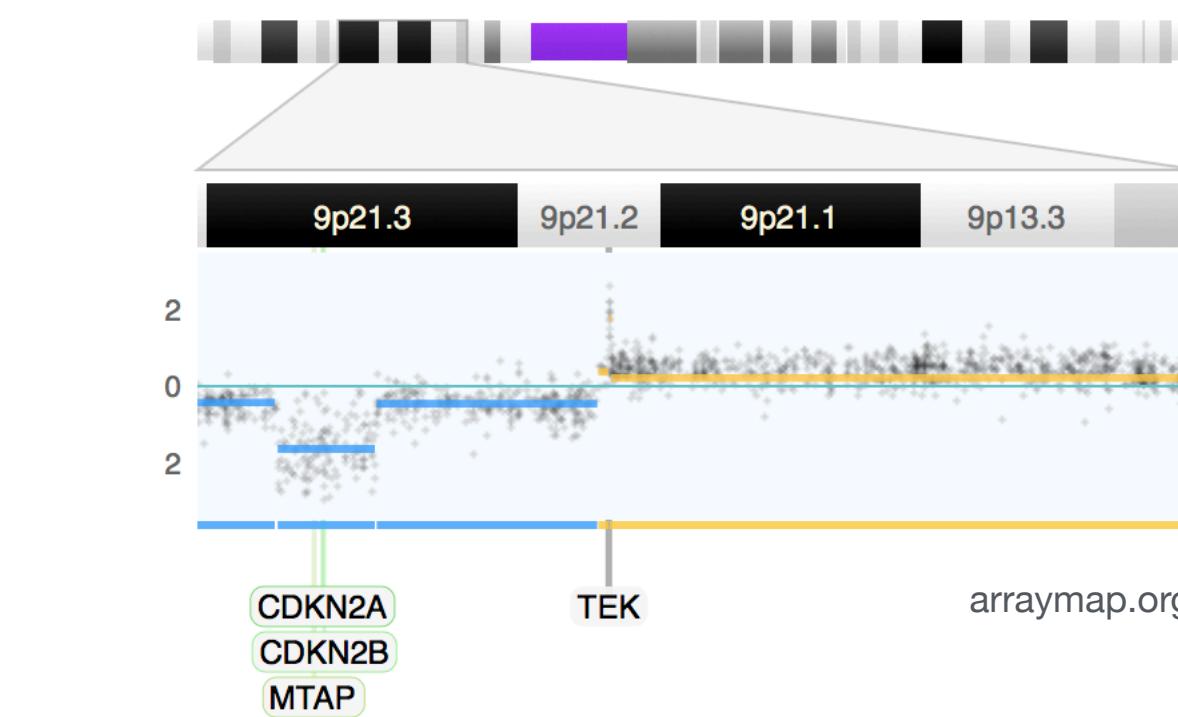
array & sequencing pipelines



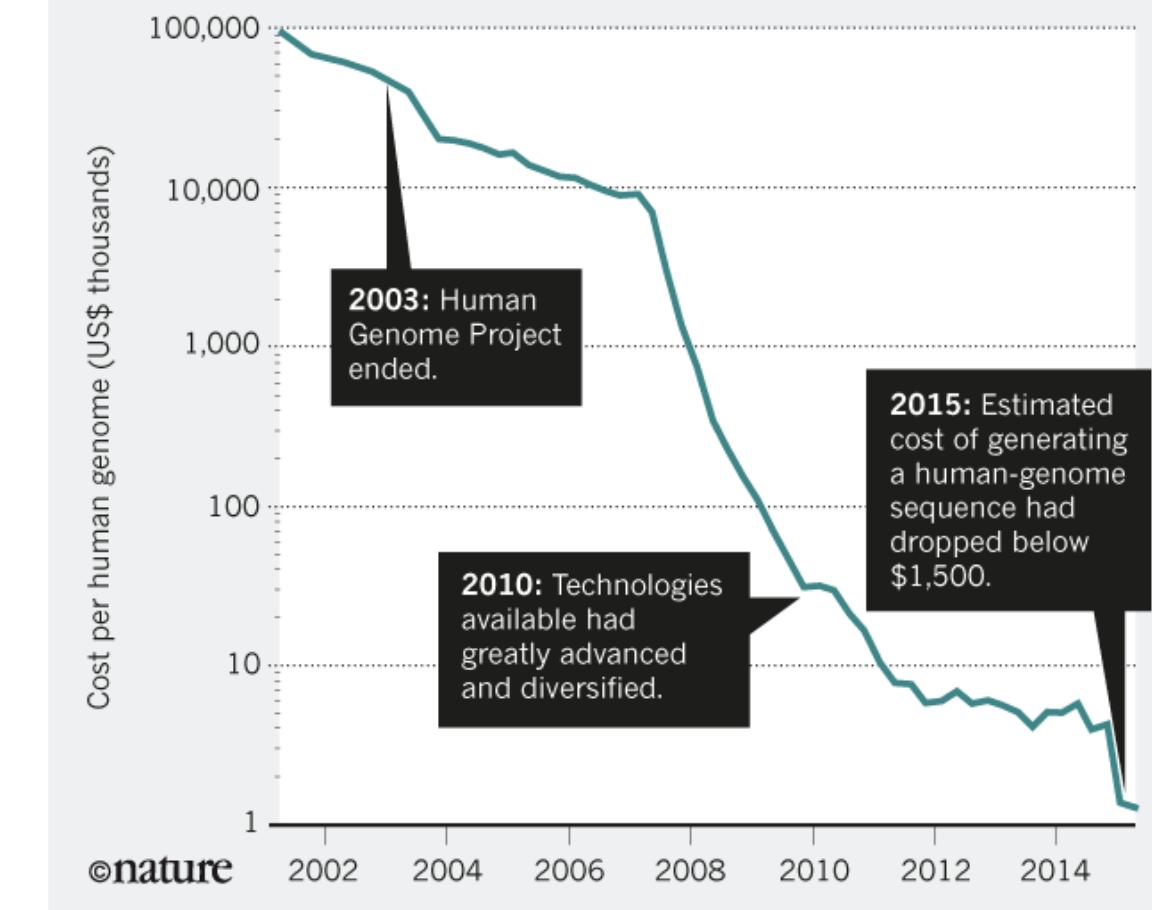


Genome screening at the core of “Personalised Health”

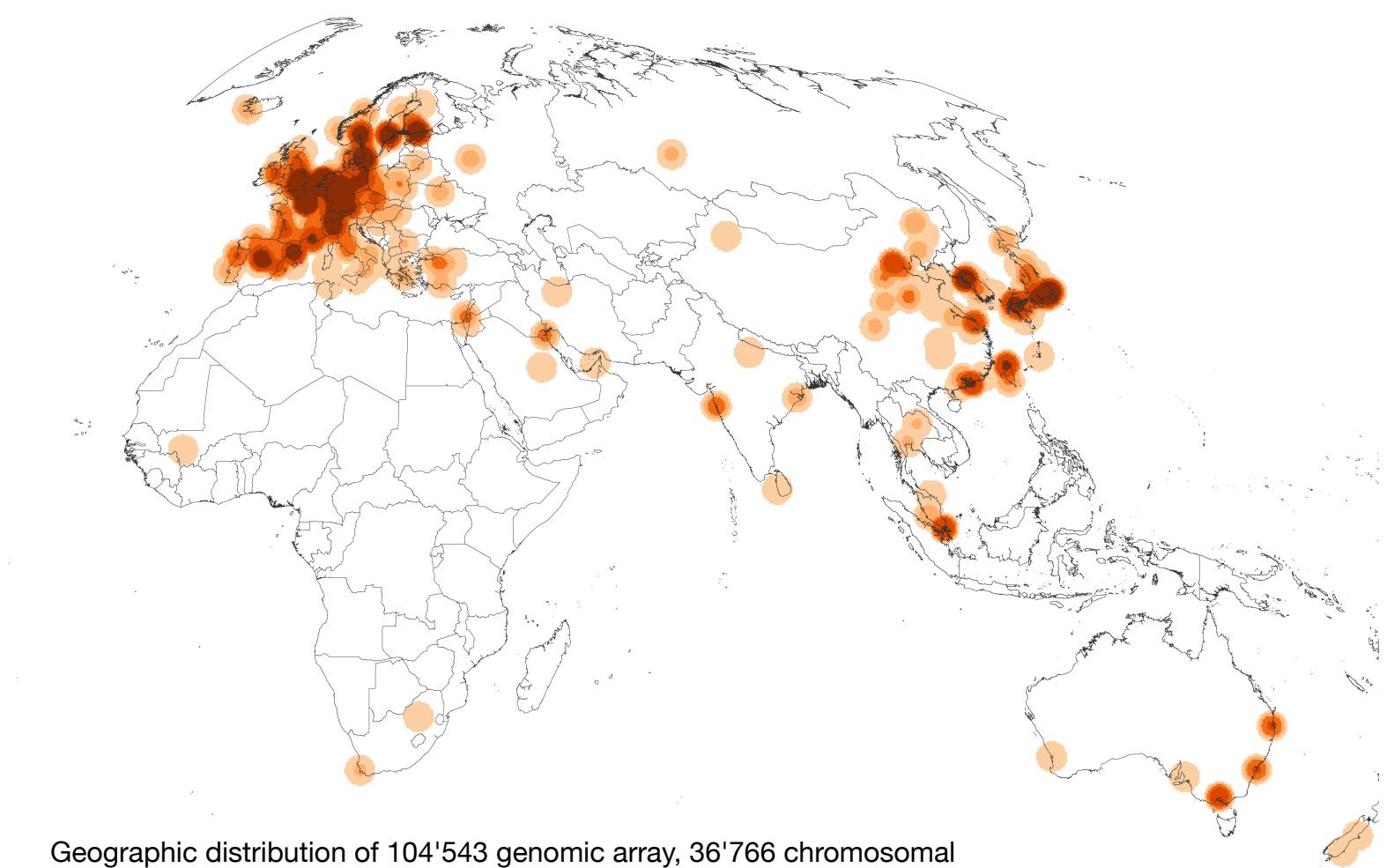
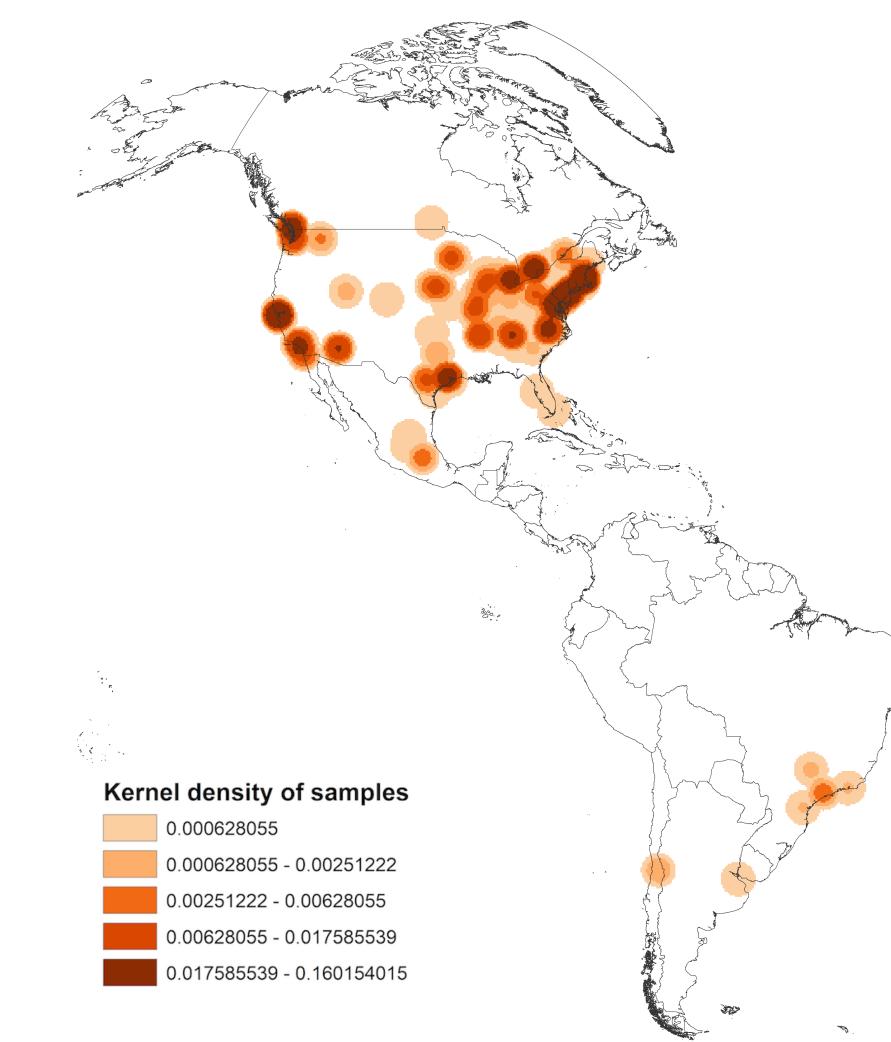
- ▶ **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- ▶ The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**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



BETTER, CHEAPER, FASTER
The cost of DNA sequencing has dropped dramatically over the past decade, enabling many more applications.

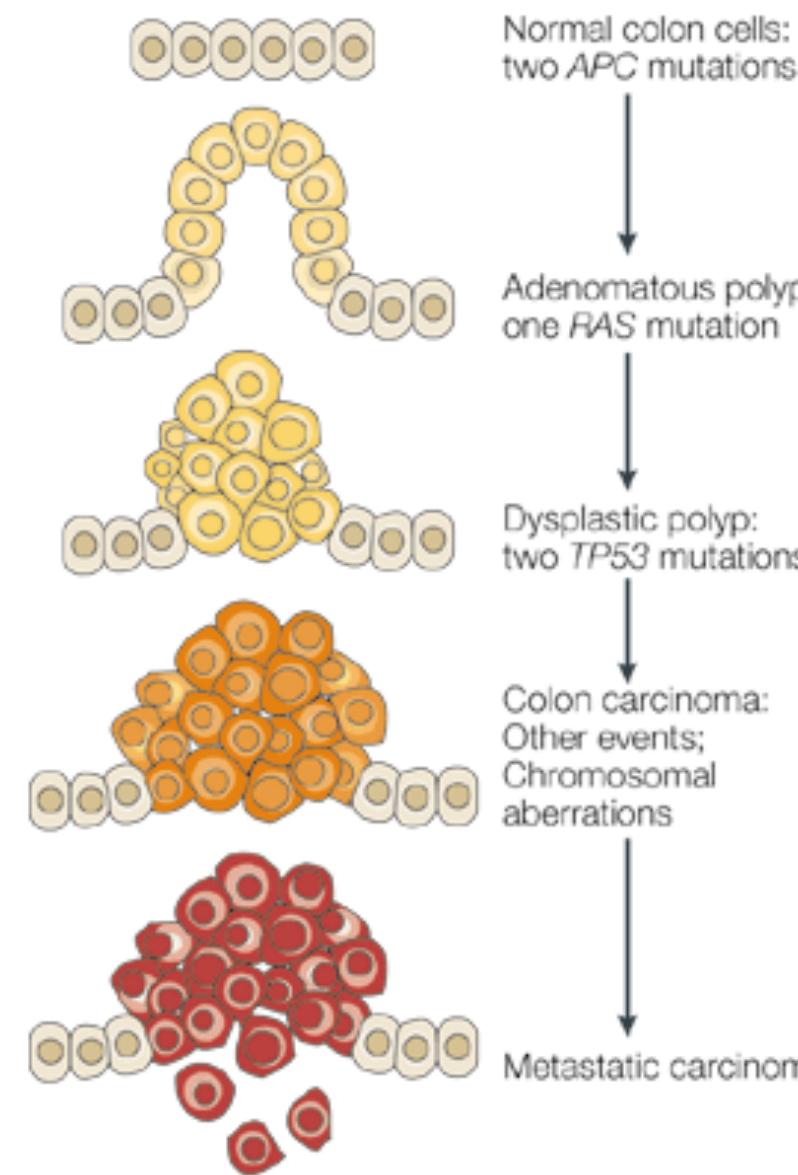
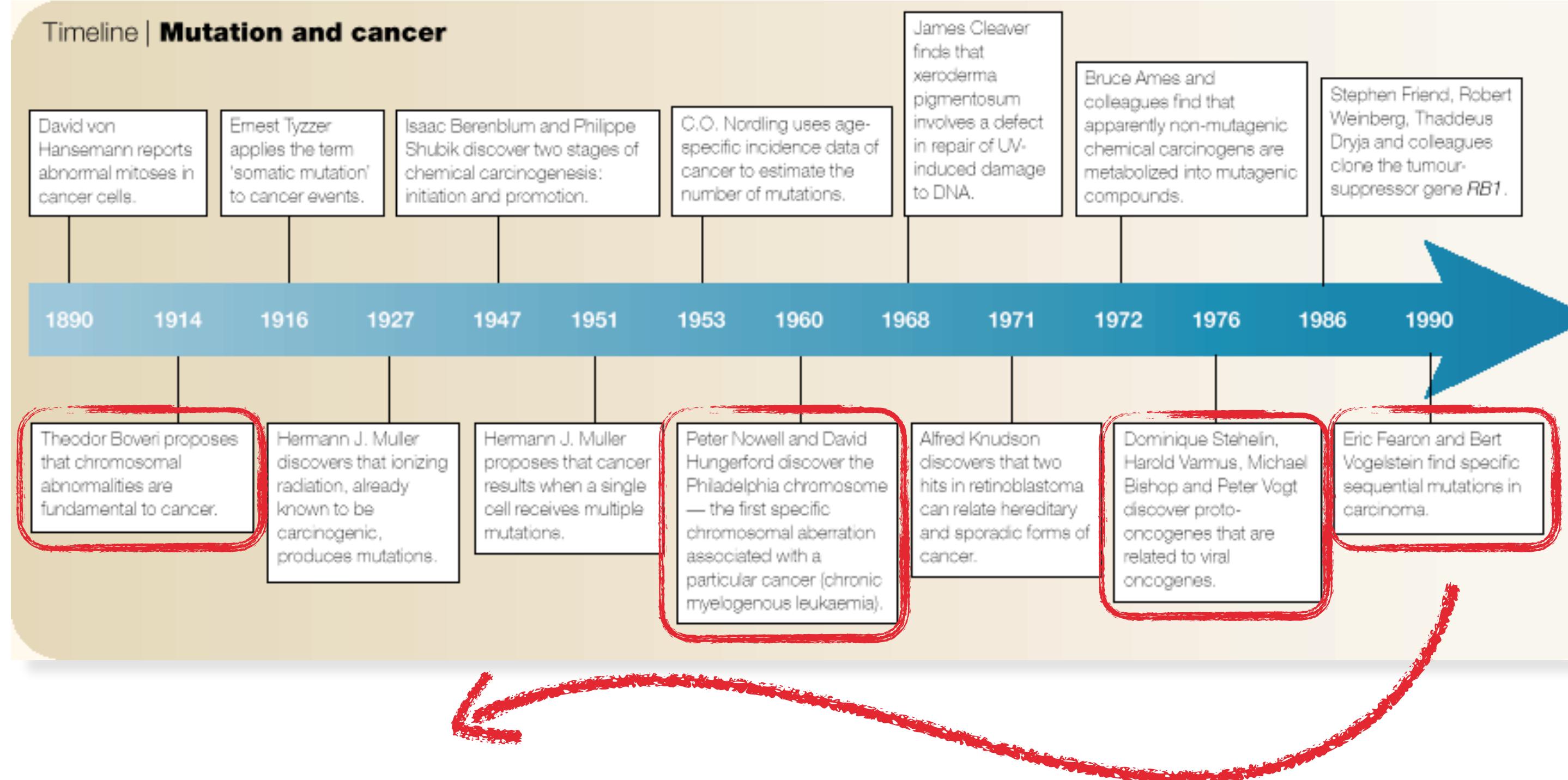


The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)



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

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.*)

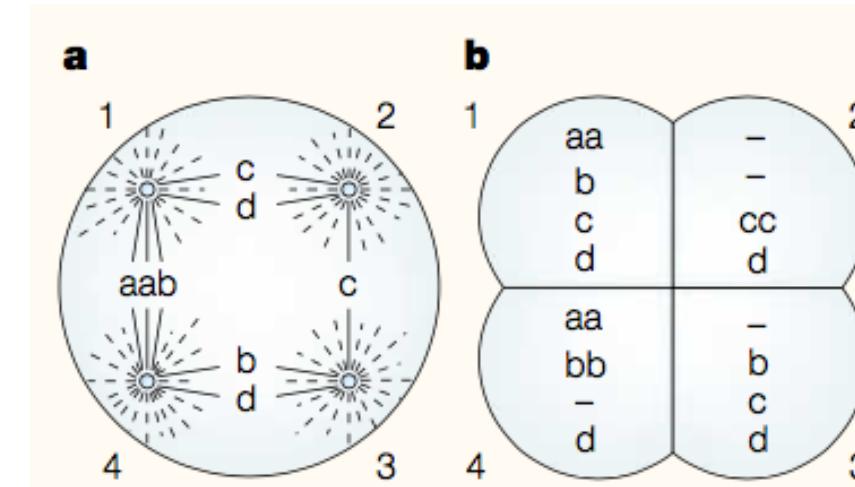
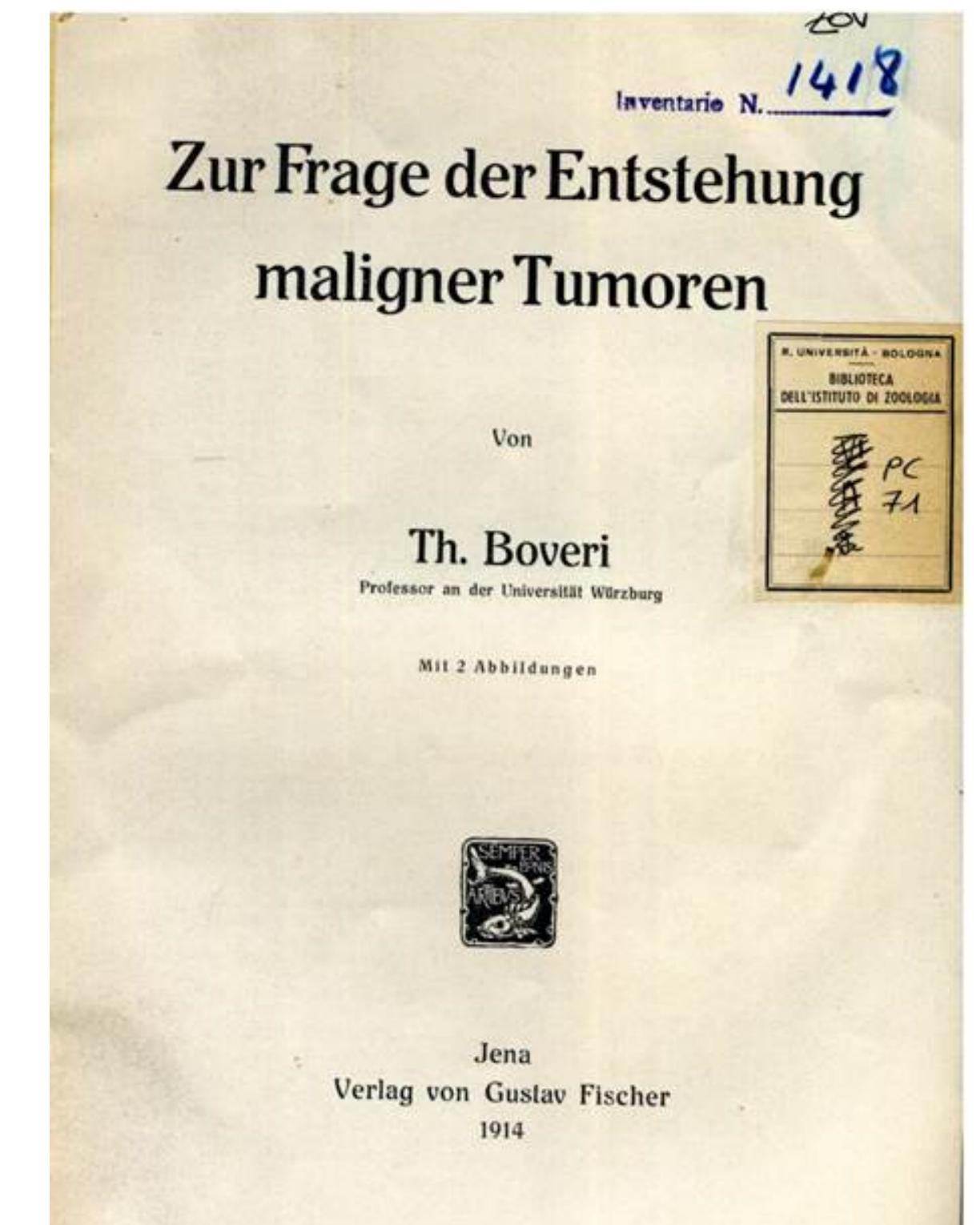


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes.** **a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached 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.

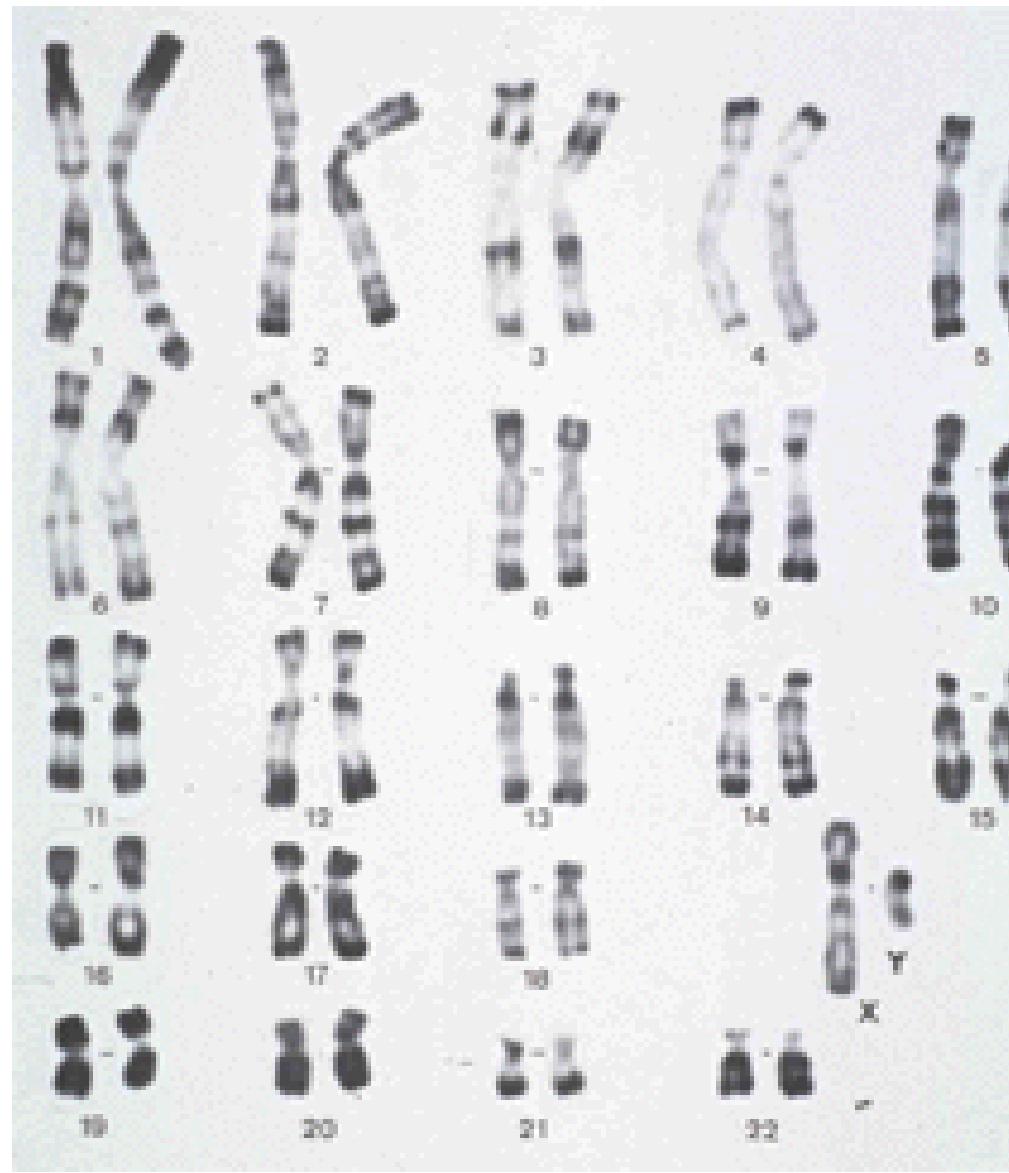


Allan Balmain
Cancer genetics: from Boveri and
Mendel to microarrays.
NatRev Cancer (2001); 1: 77-82

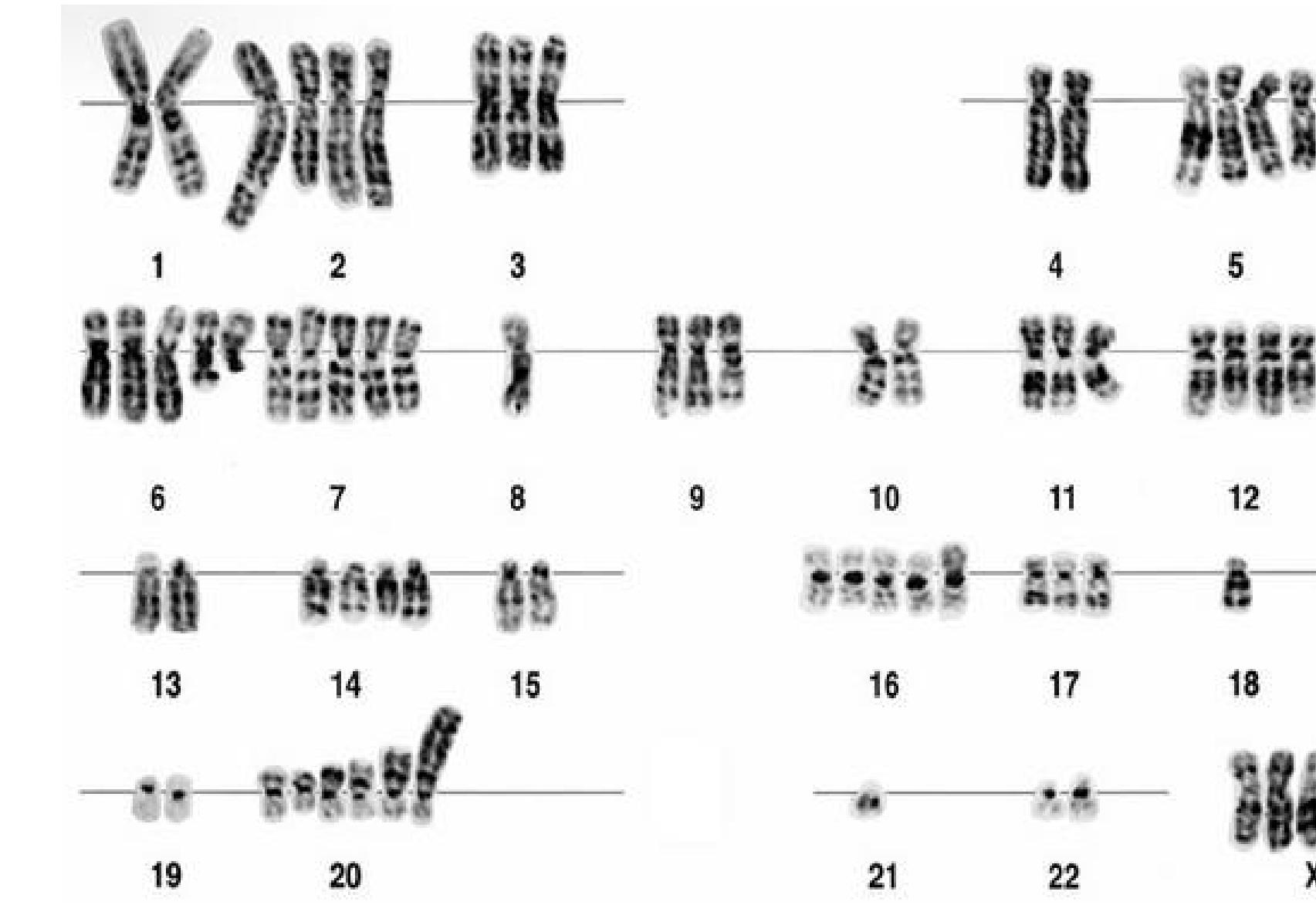
Anna Di Leonardo , Sergio Nasi , Simonetta Pulciani
Cancer: We Should Not Forget The Past
Journal of Cancer (2015), Vol. 6: 29-39
(for book cover & summary)

Genomic changes at the DNA level are hallmarks of cancer

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

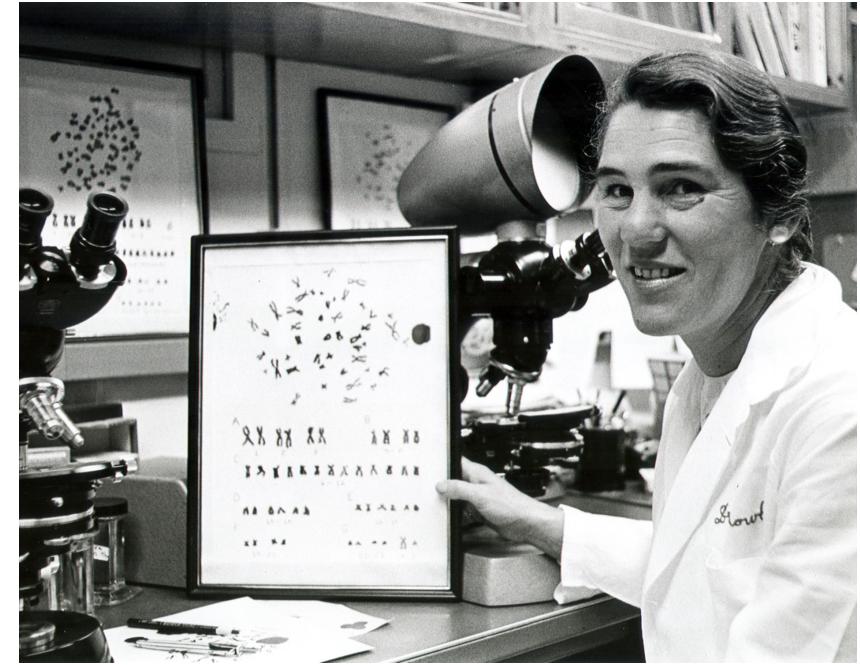


Normal karyotype



Tumor karyotype

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



Janet Rowley (1972/73)

Chromosomal translocations in cancer

- Recurrent chromosomal translocations in leukemias and 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)

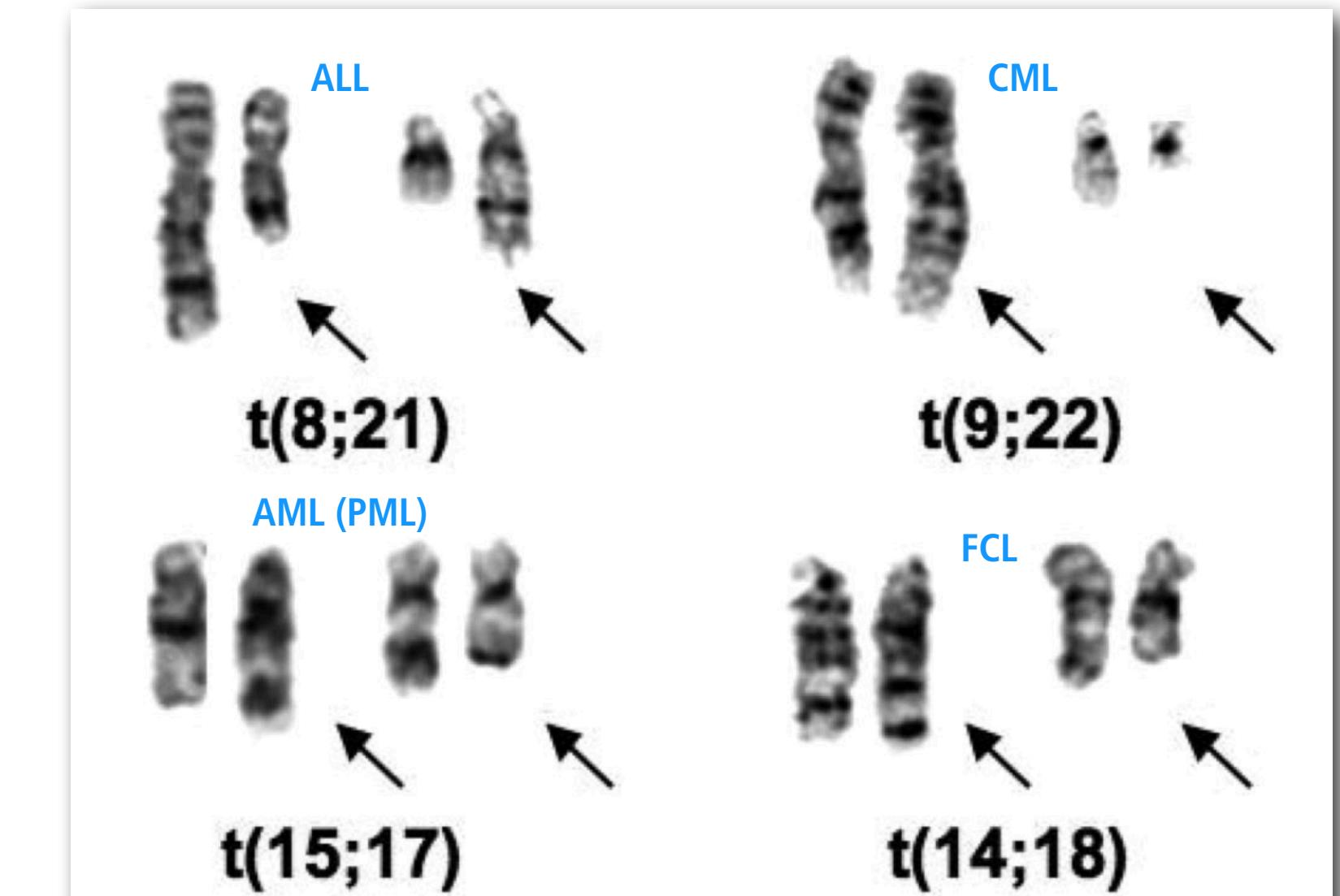
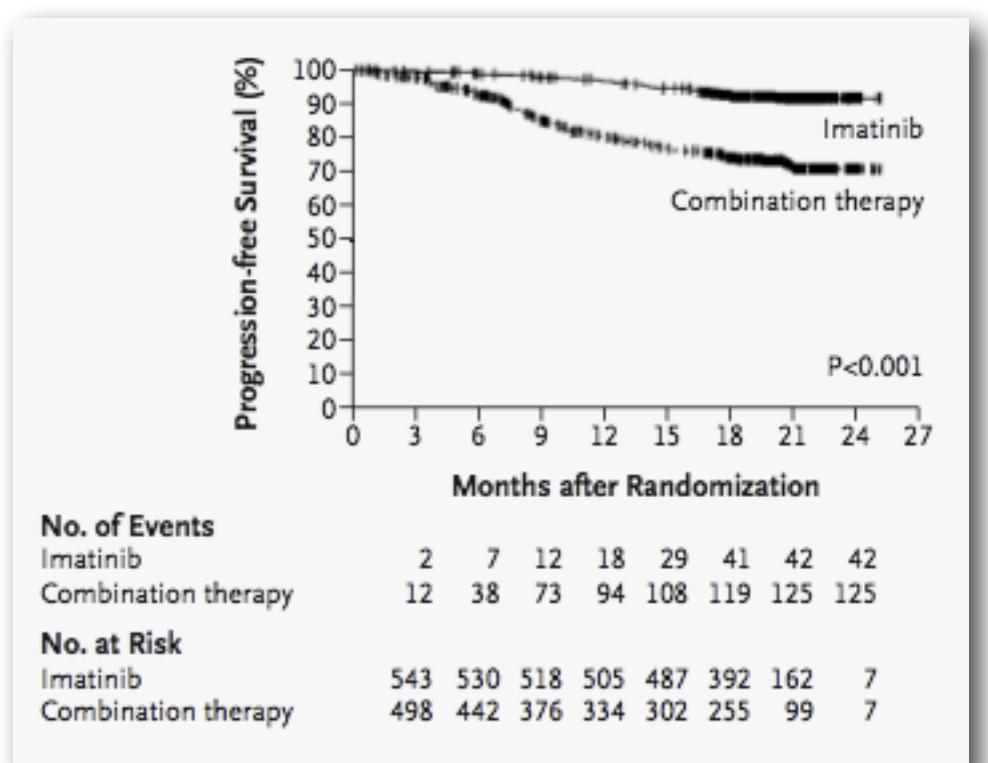
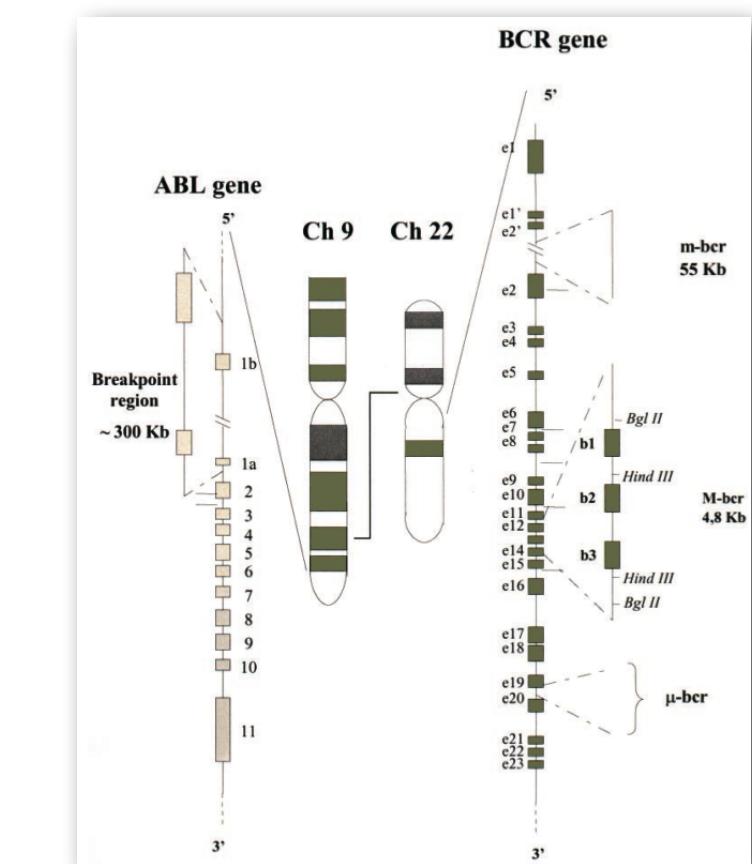


Figure 1. Partial karyotypes of common translocations discovered by Rowley.
The translocations appear in the order in which they were discovered.

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



Event free Survival in first large Imatinib Trials

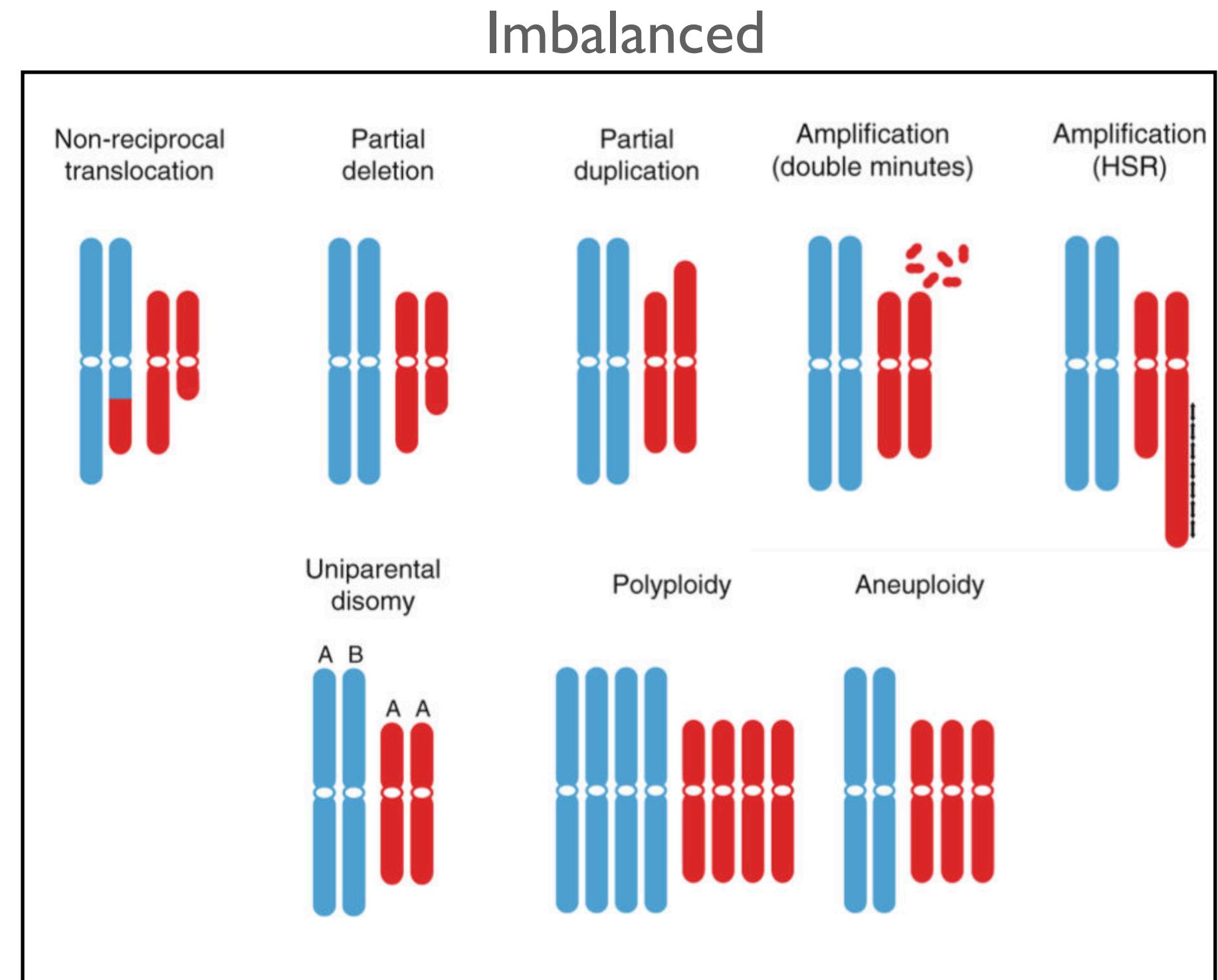
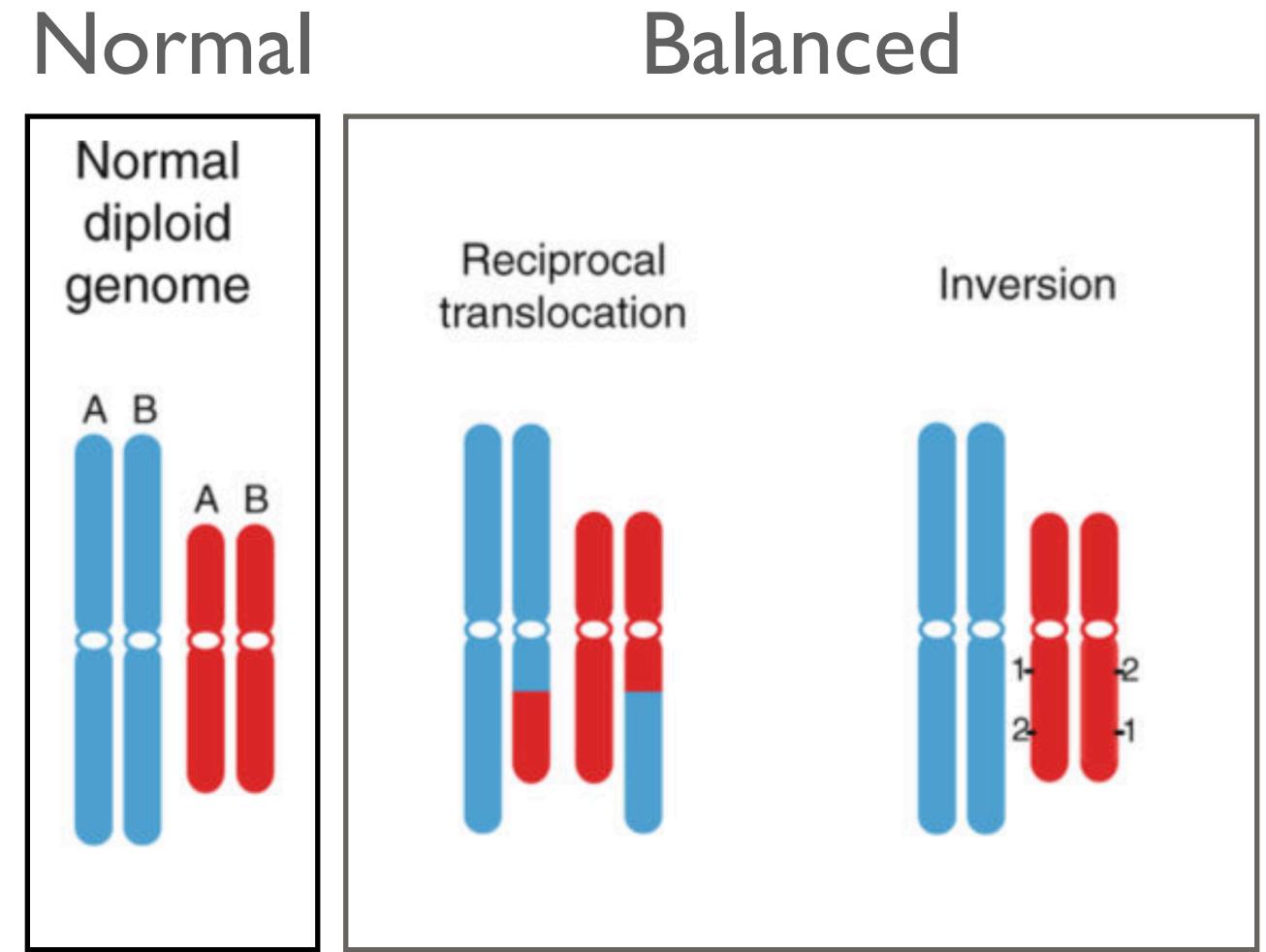
Pane et al. BCR/ABL genes
Oncogene (2002), 21 (56)

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

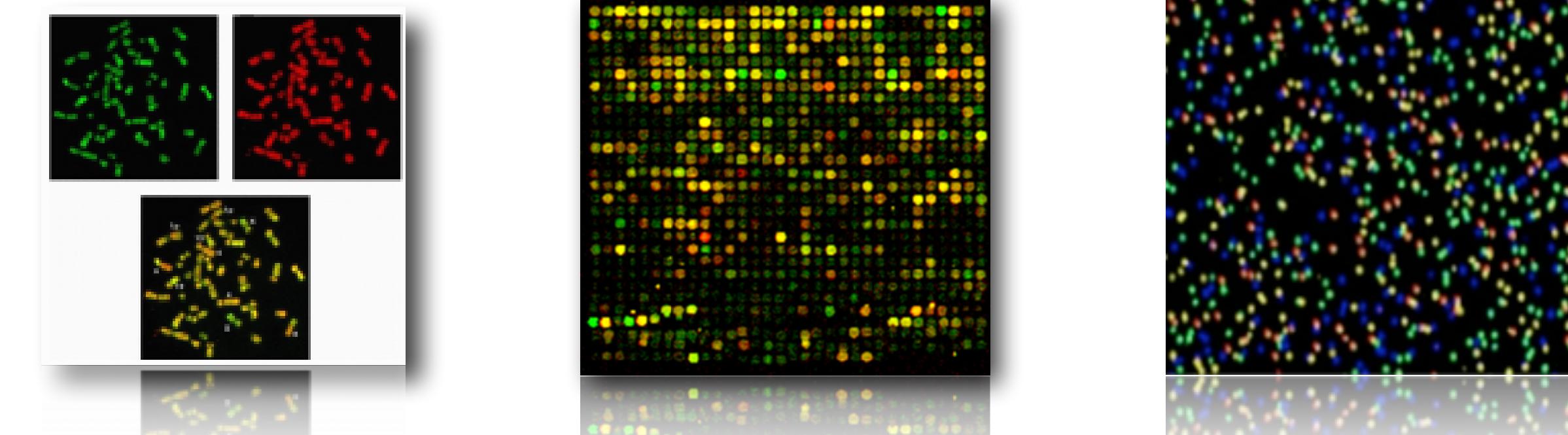
Types of genomic alterations in Cancer

Imbalanced Chromosomal Changes: CNV

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

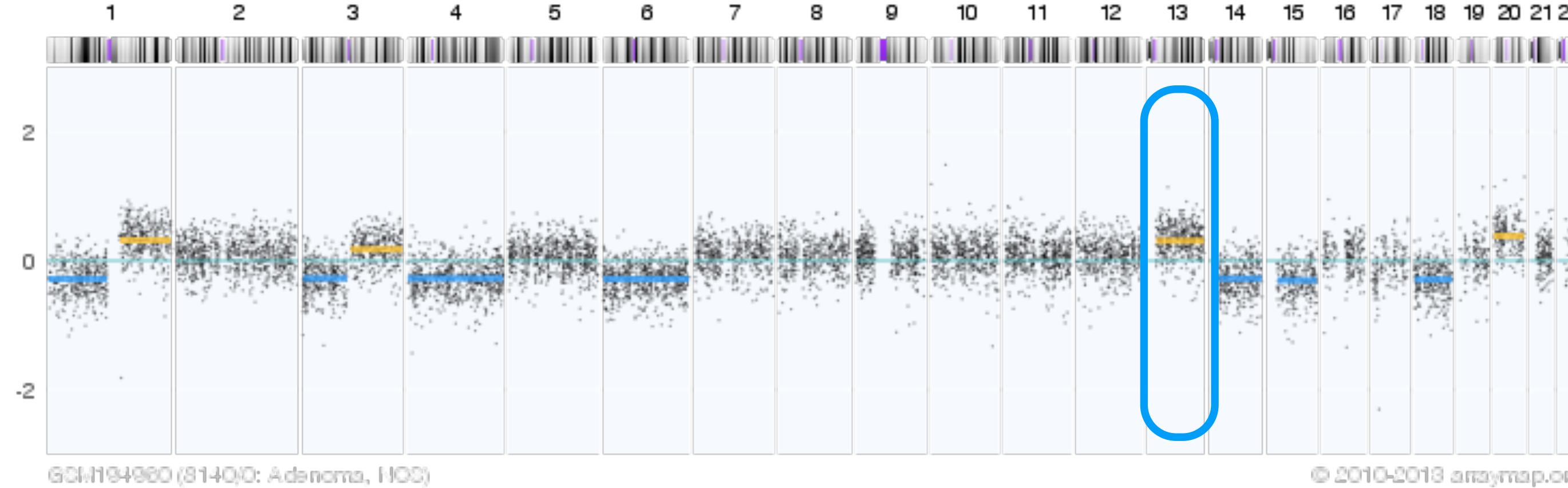


WHOLE GENOME SCREENING IN CANCER

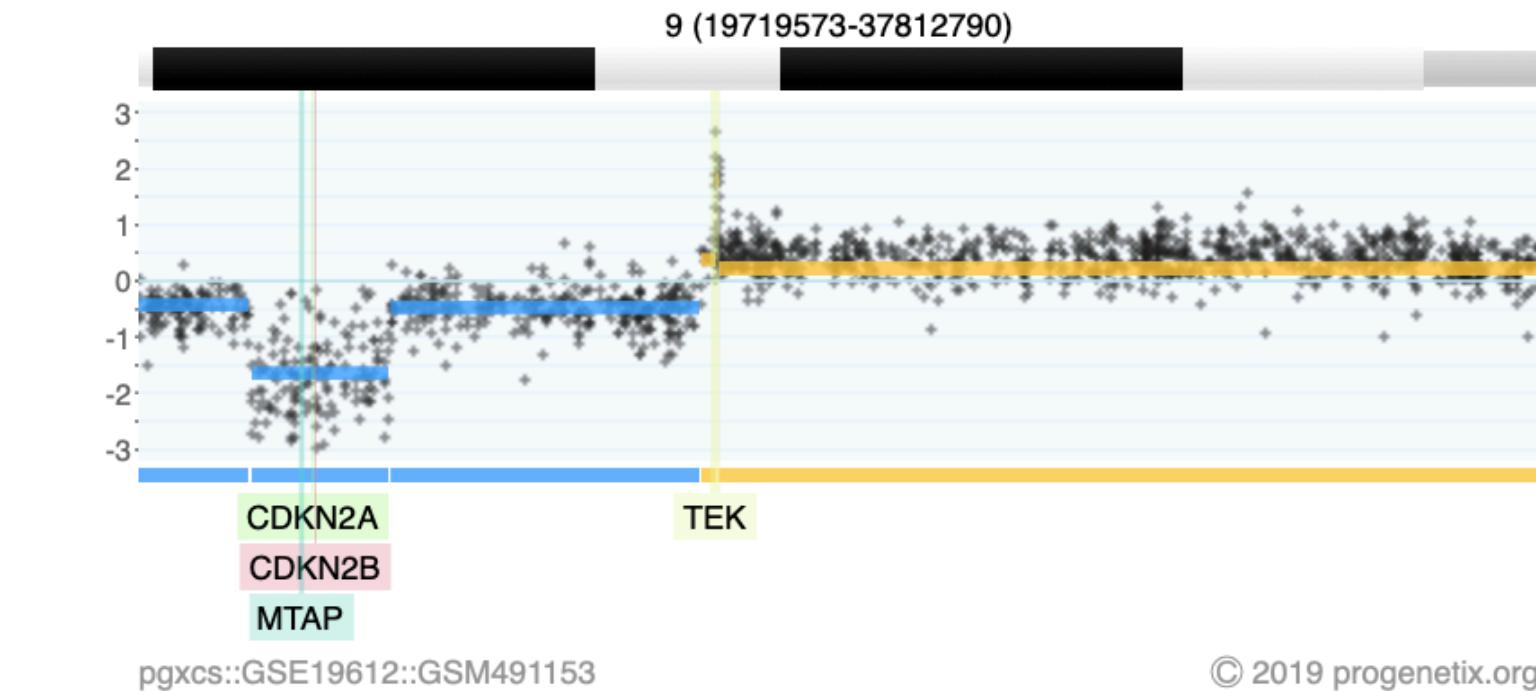


	chromosomal CGH	genomic arrays	“NGS” genome sequencing (WES, WGS)
1st application report	1992	1997	2010
source	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)
main source problems	mixed/degraded source tissue	mixed/degraded source tissue	mixed/degraded source tissue
resolution	chromosomal bands = few megabases	mostly in the 100kb range, but tiling possible	single bases
target identification	surrogate (position)	“semidirect” (segmentation spanning probes)	direct quantitative and qualitative
structural	no	depending on type	yes
available data	>24,000 cases (57%) through Progenetix	raw data repositories (GEO, EMBL, SMD), Progenetix	Limited for raw data (BAMs ...); variant call data in dbgap, clinvar; selected studies with called CNV segments
predominant data format	ISCN = static	raw => depends on bioinformatics	mostly annotated variant calls or SNVs

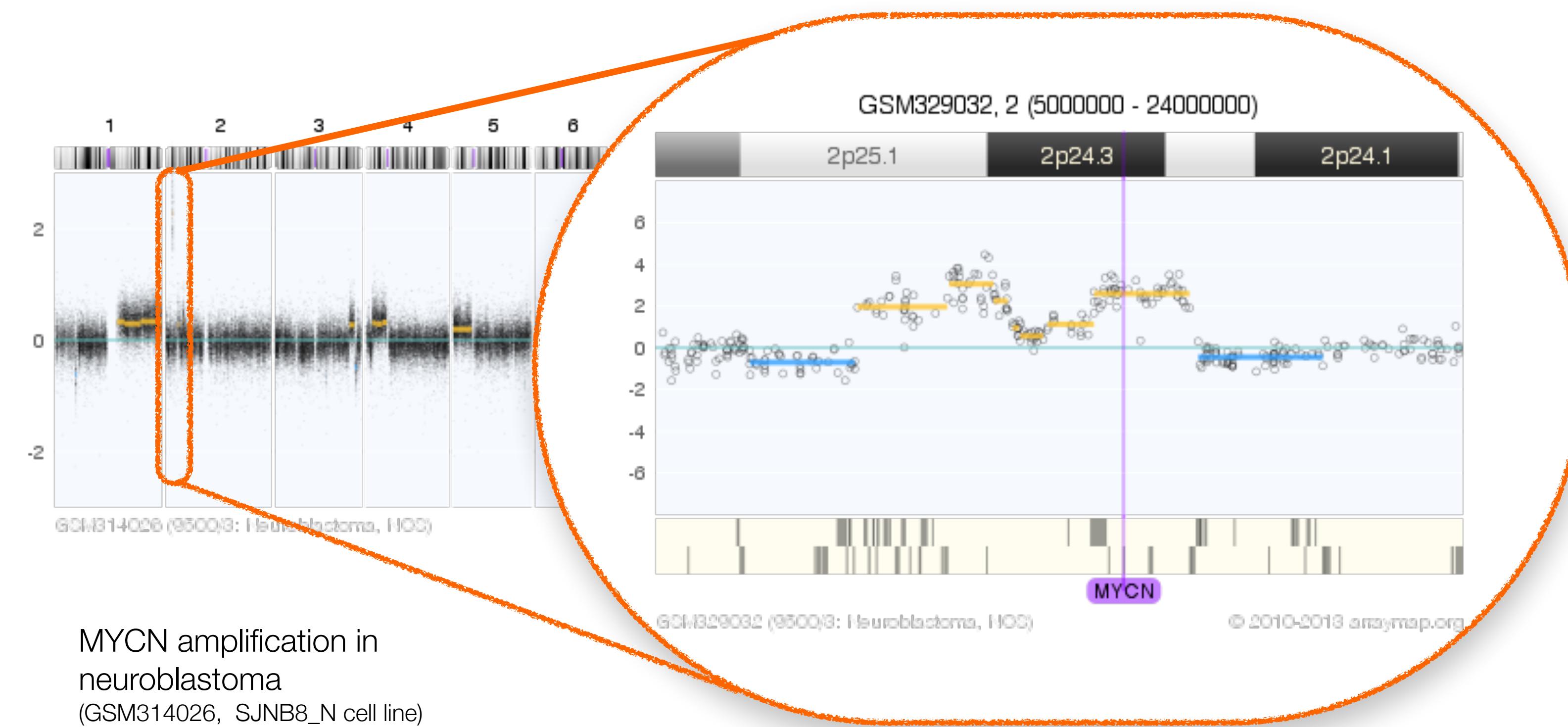
Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma



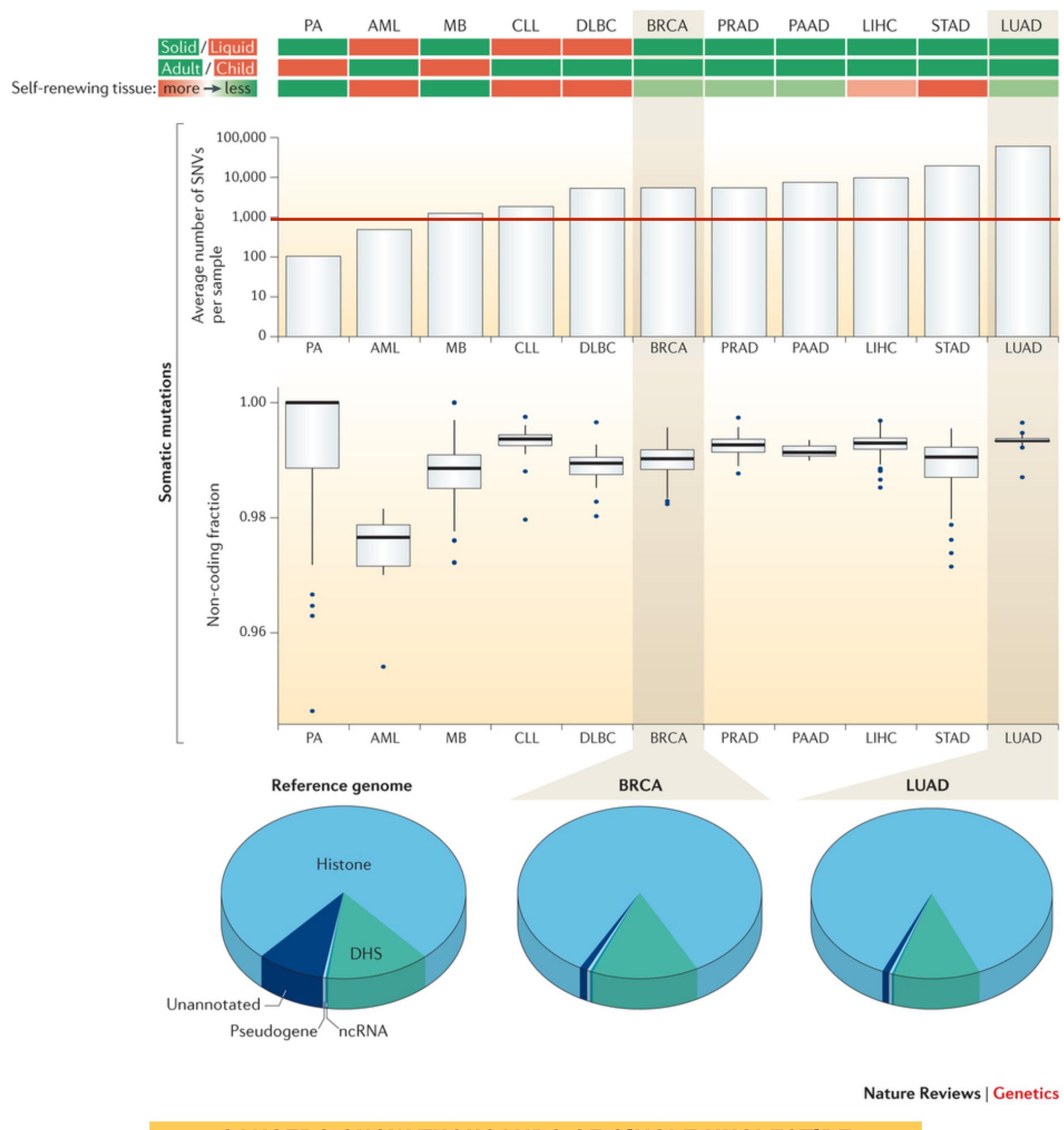
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

low level/high level copy number alterations (CNAs)

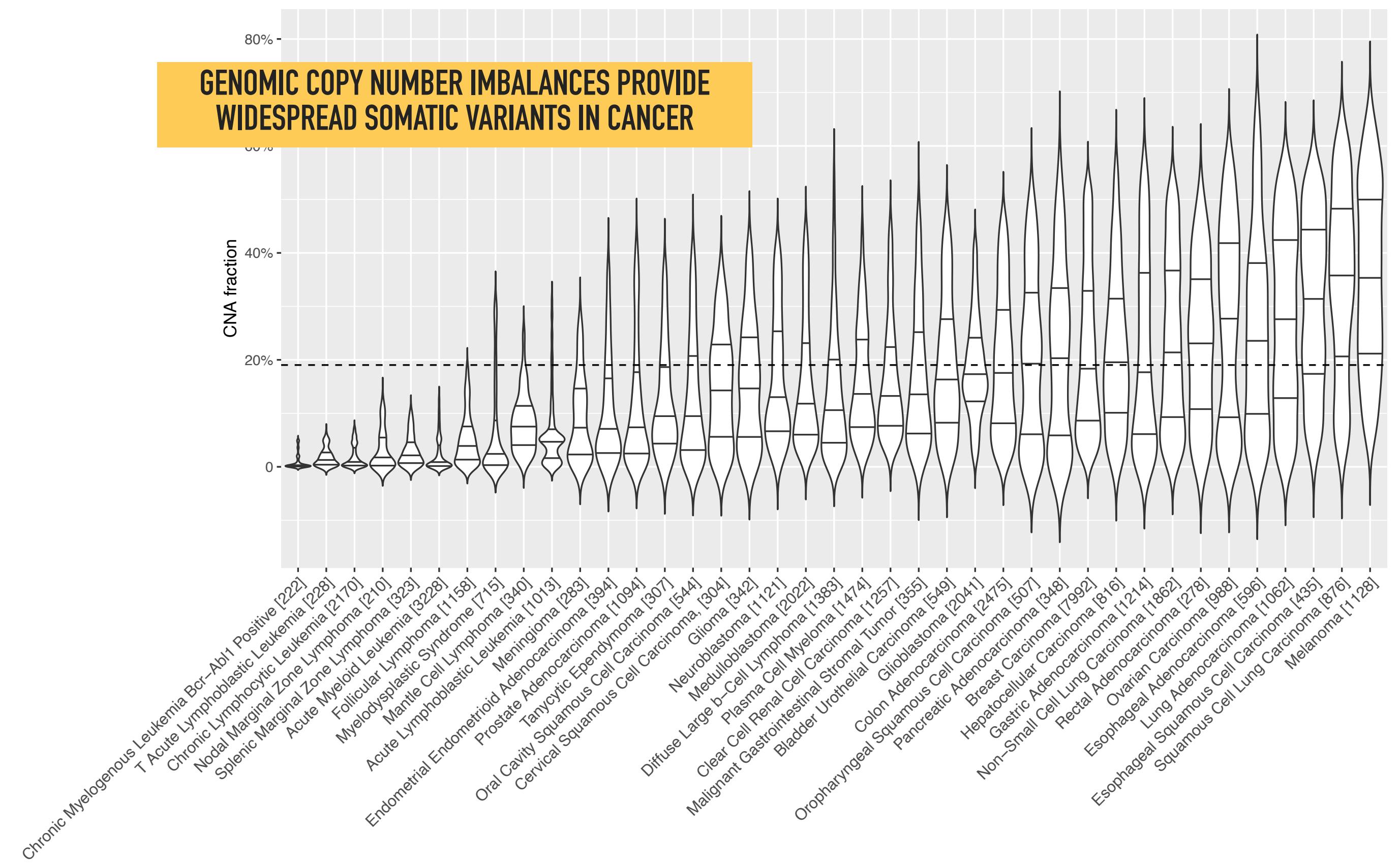
arrayMap



Quantifying Somatic Mutations In Cancer



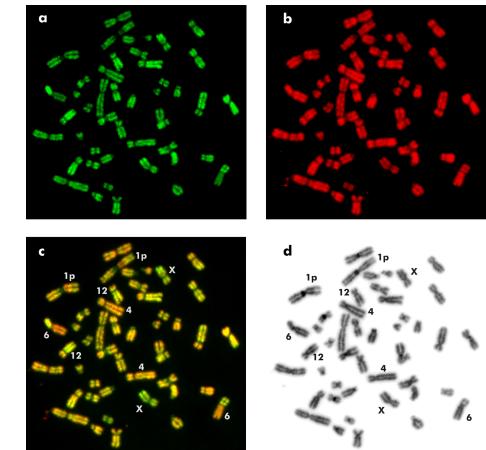
Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016))



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

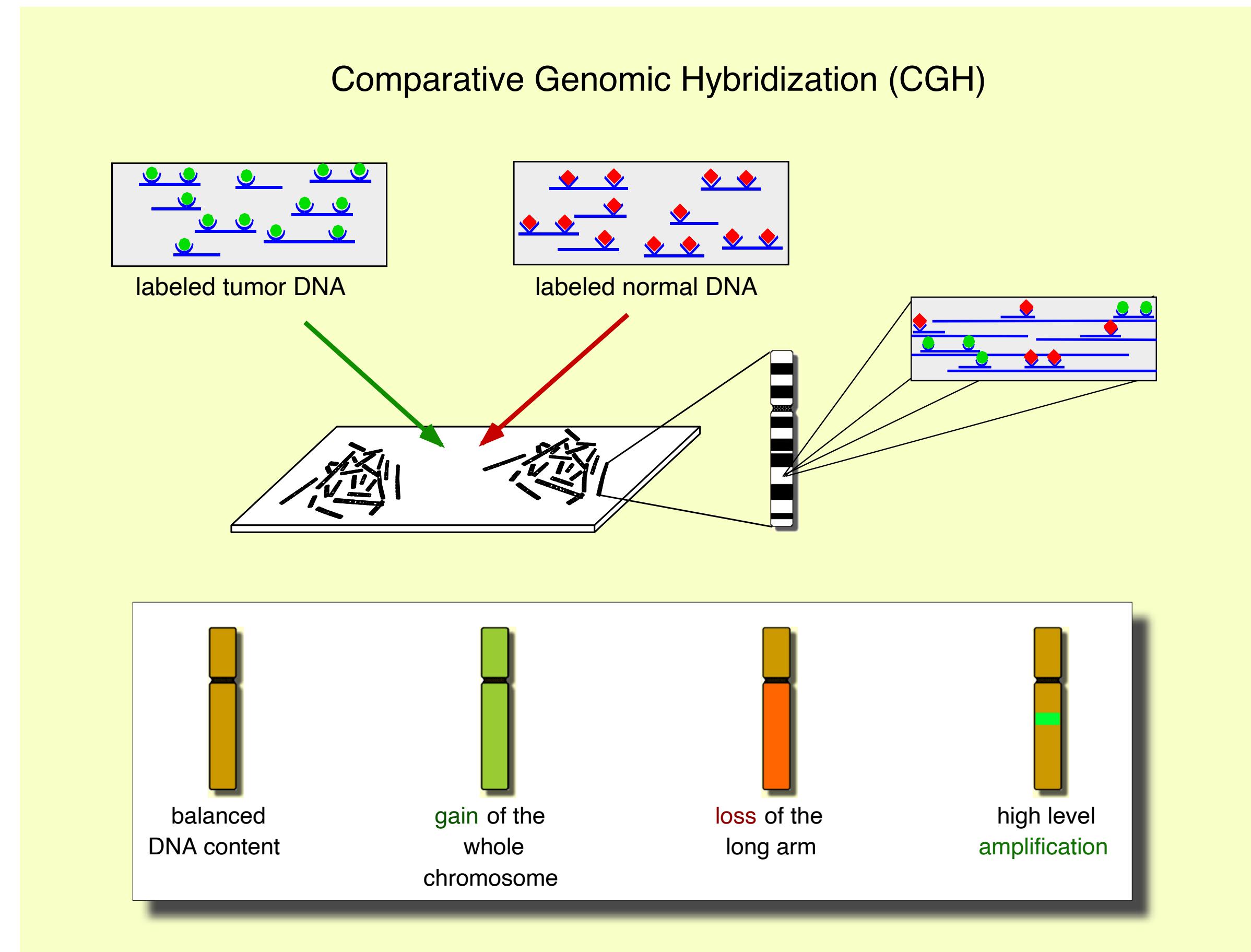
Comparative Genomic Hybridization

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

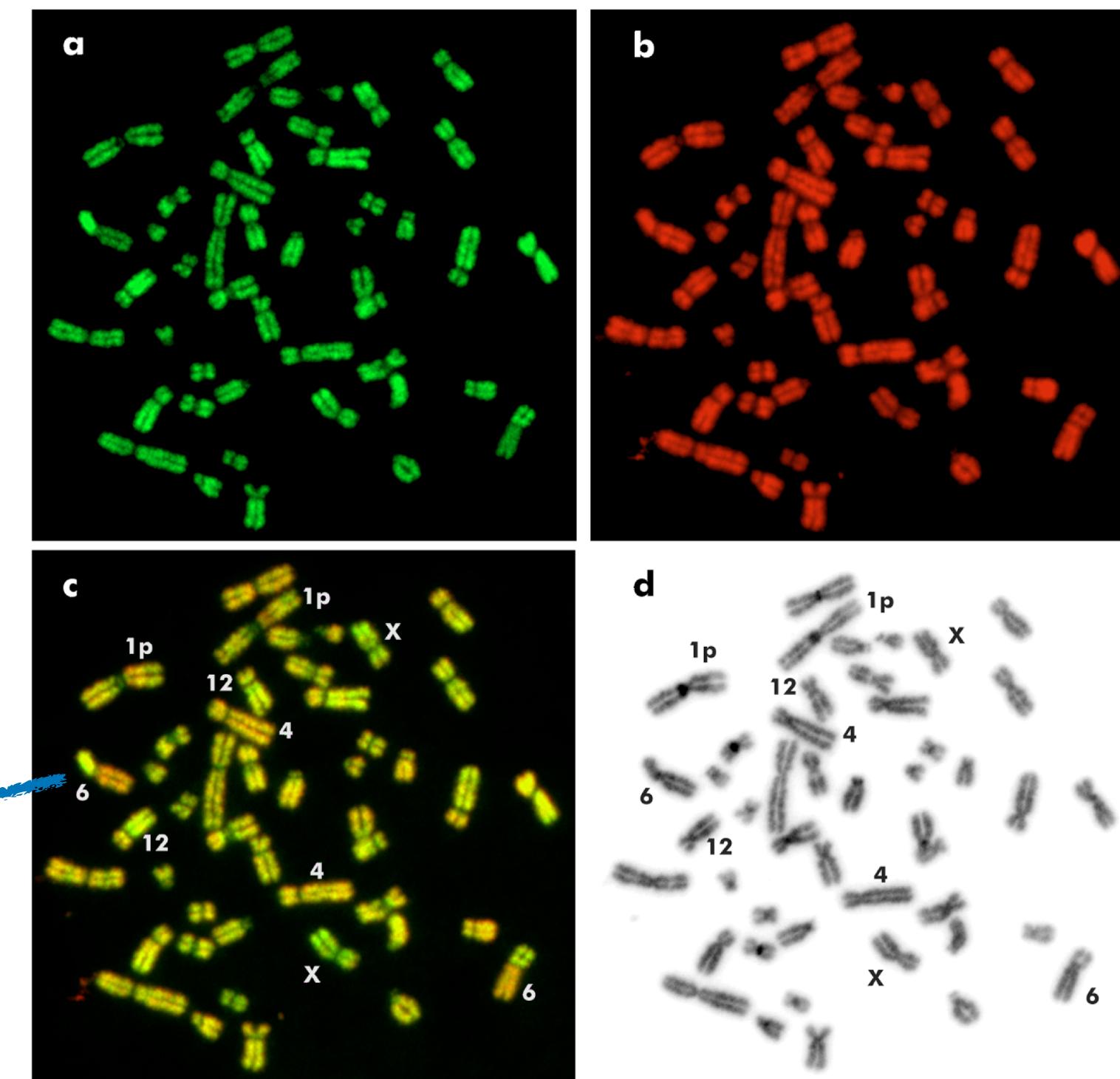
Comparative Genomic Hybridization

Molecular-Cytogenetic Technology for Genomic Imbalance Screening

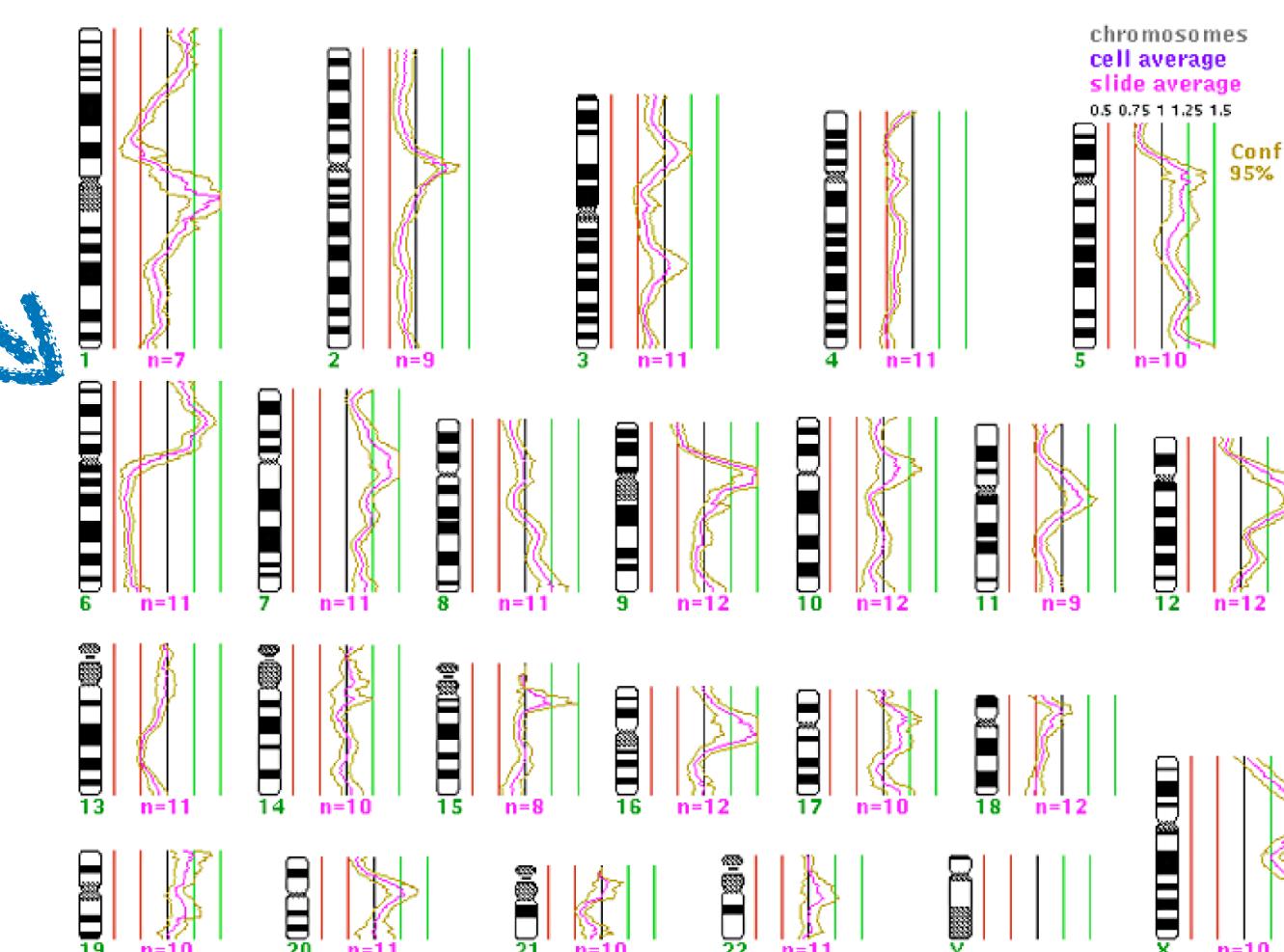
- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression hybridization of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative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



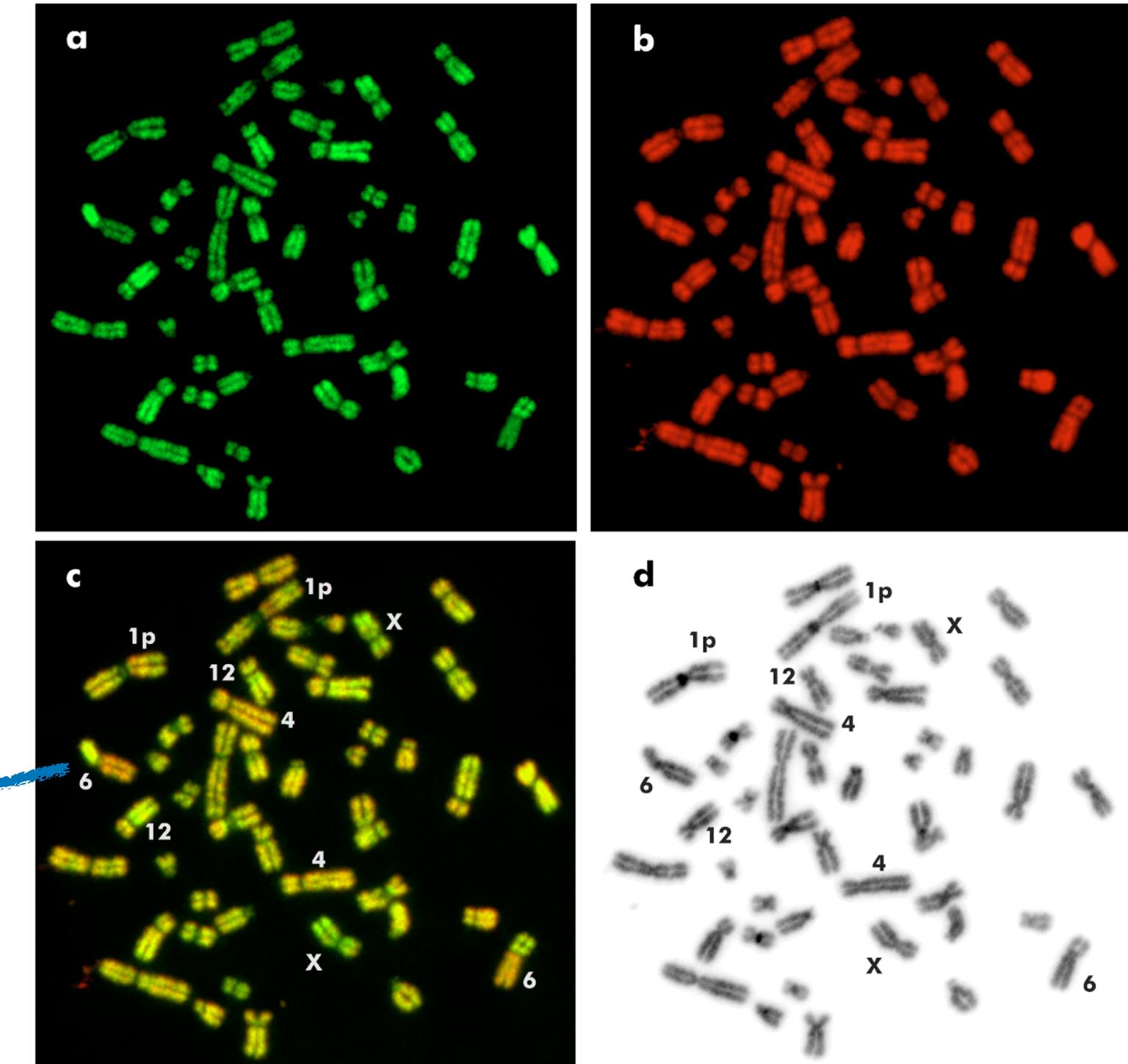
Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

Comparative Genomic Hybridization

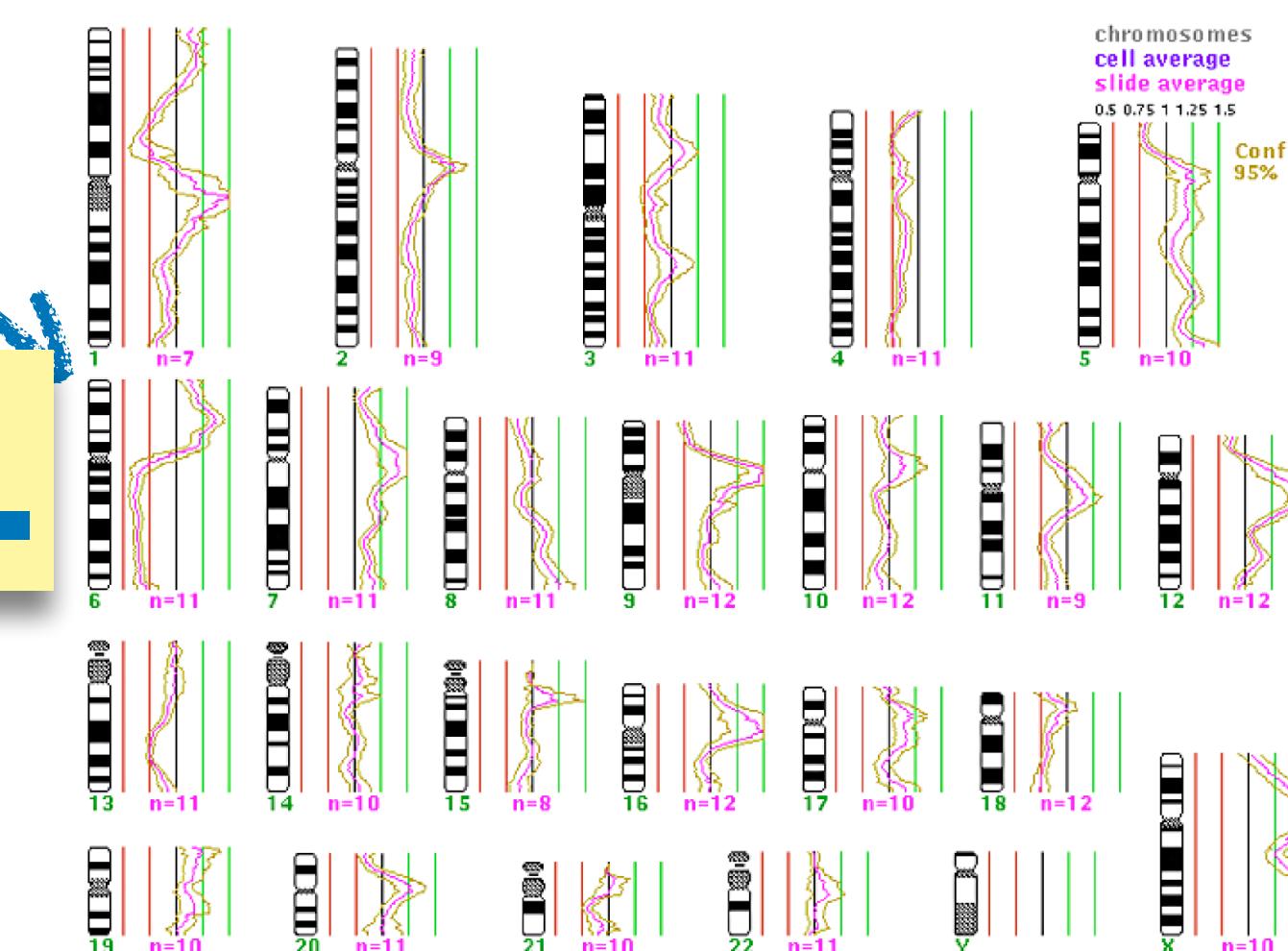
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- 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

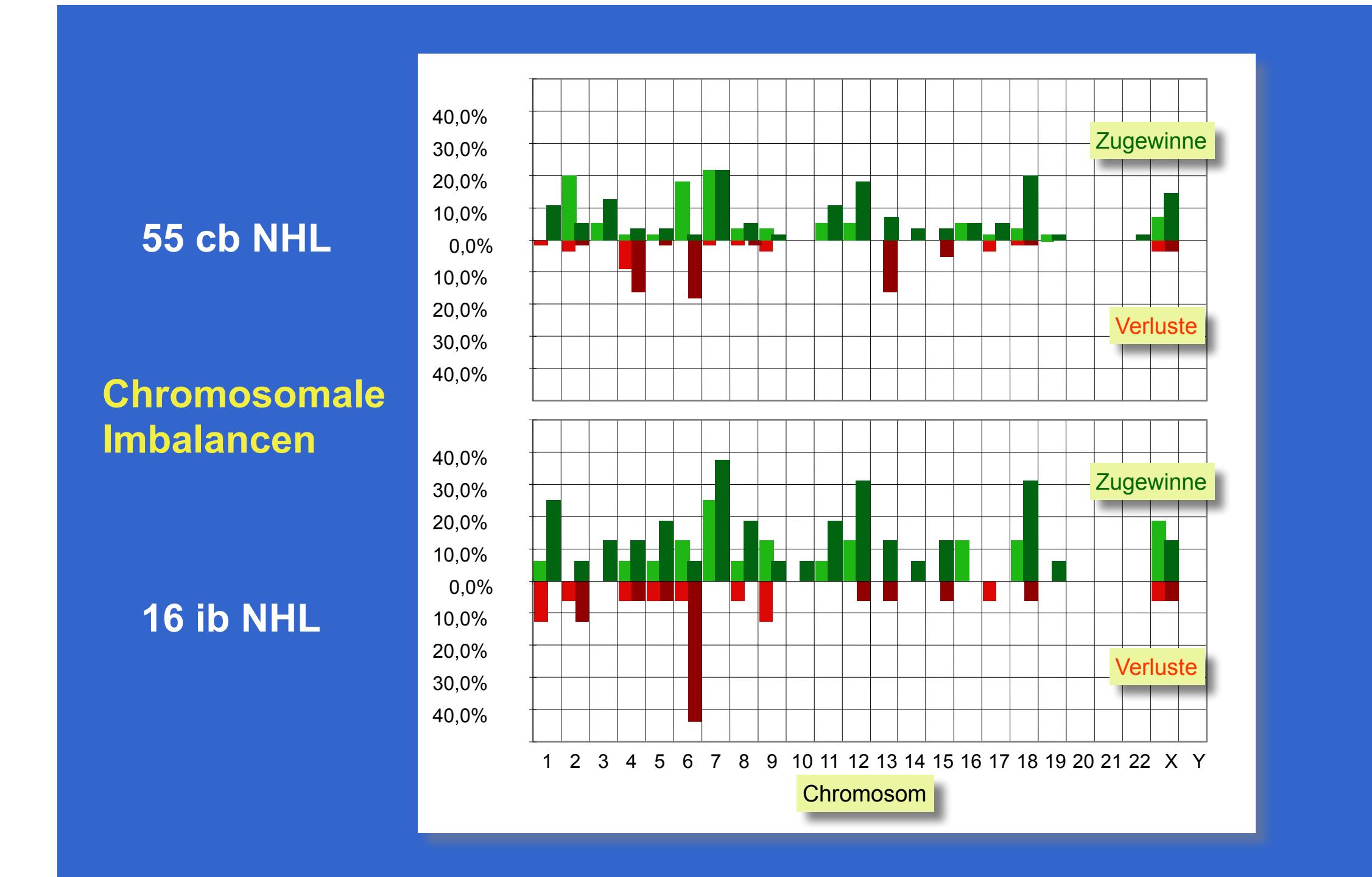
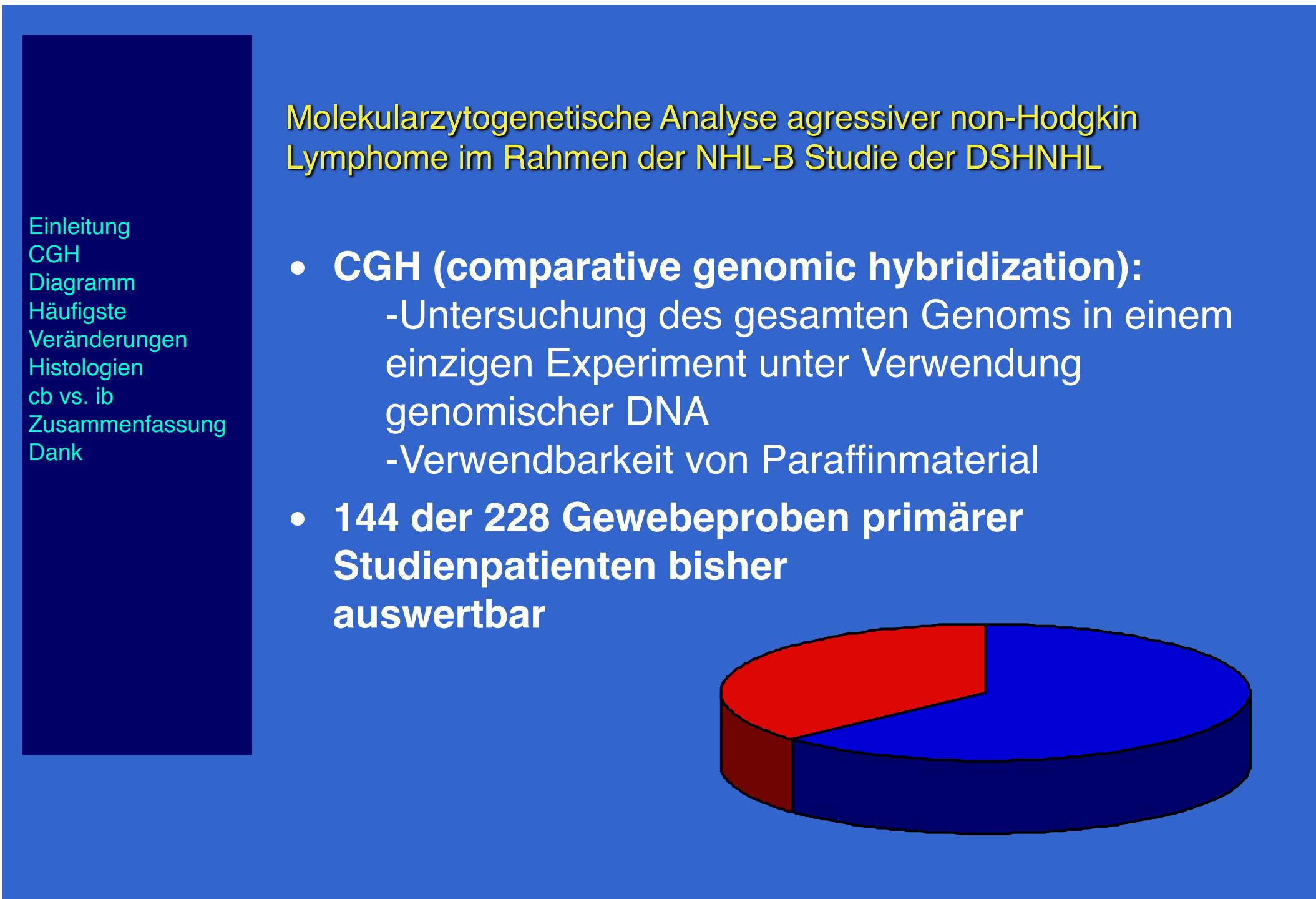


Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

- 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.

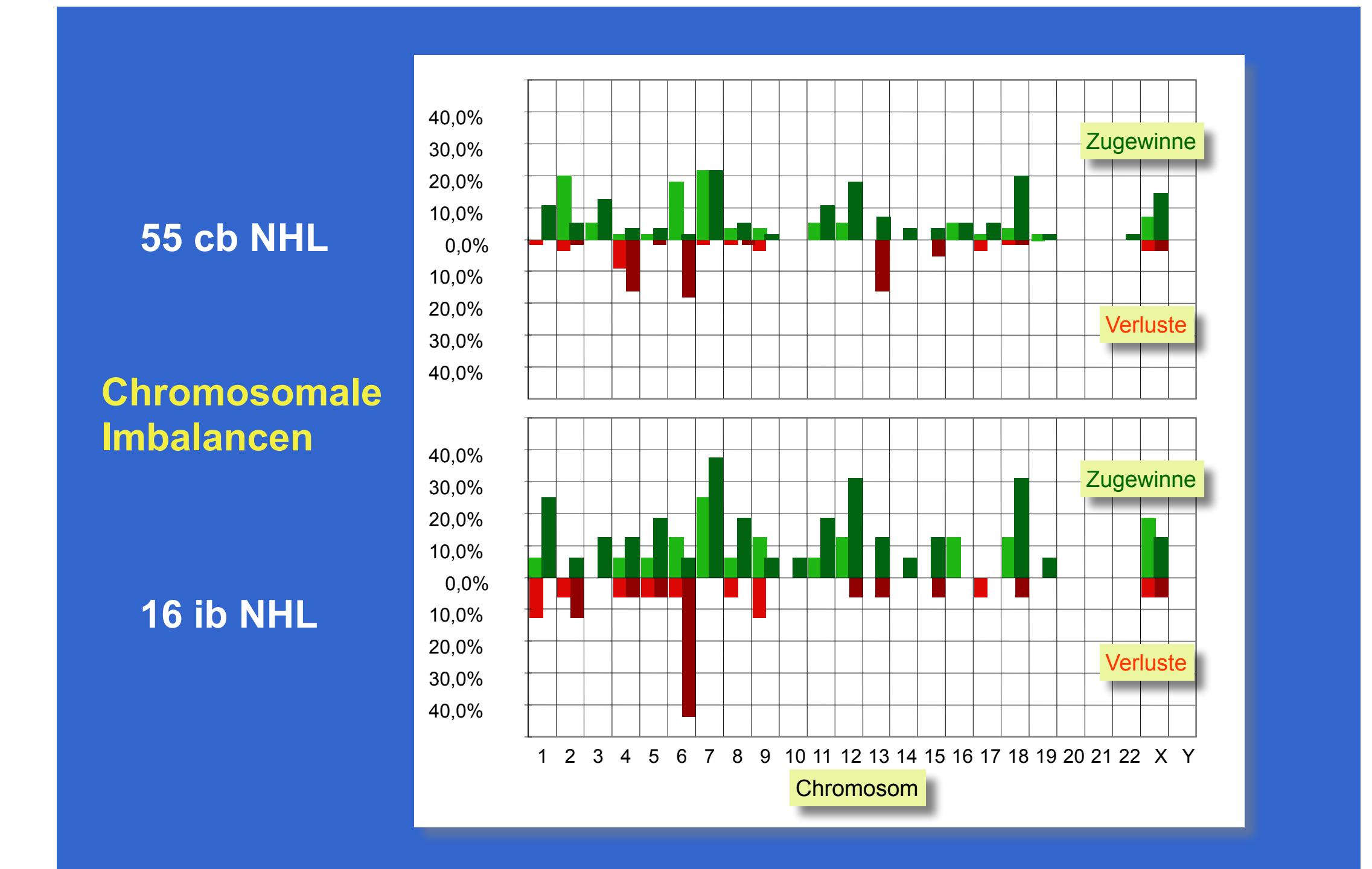
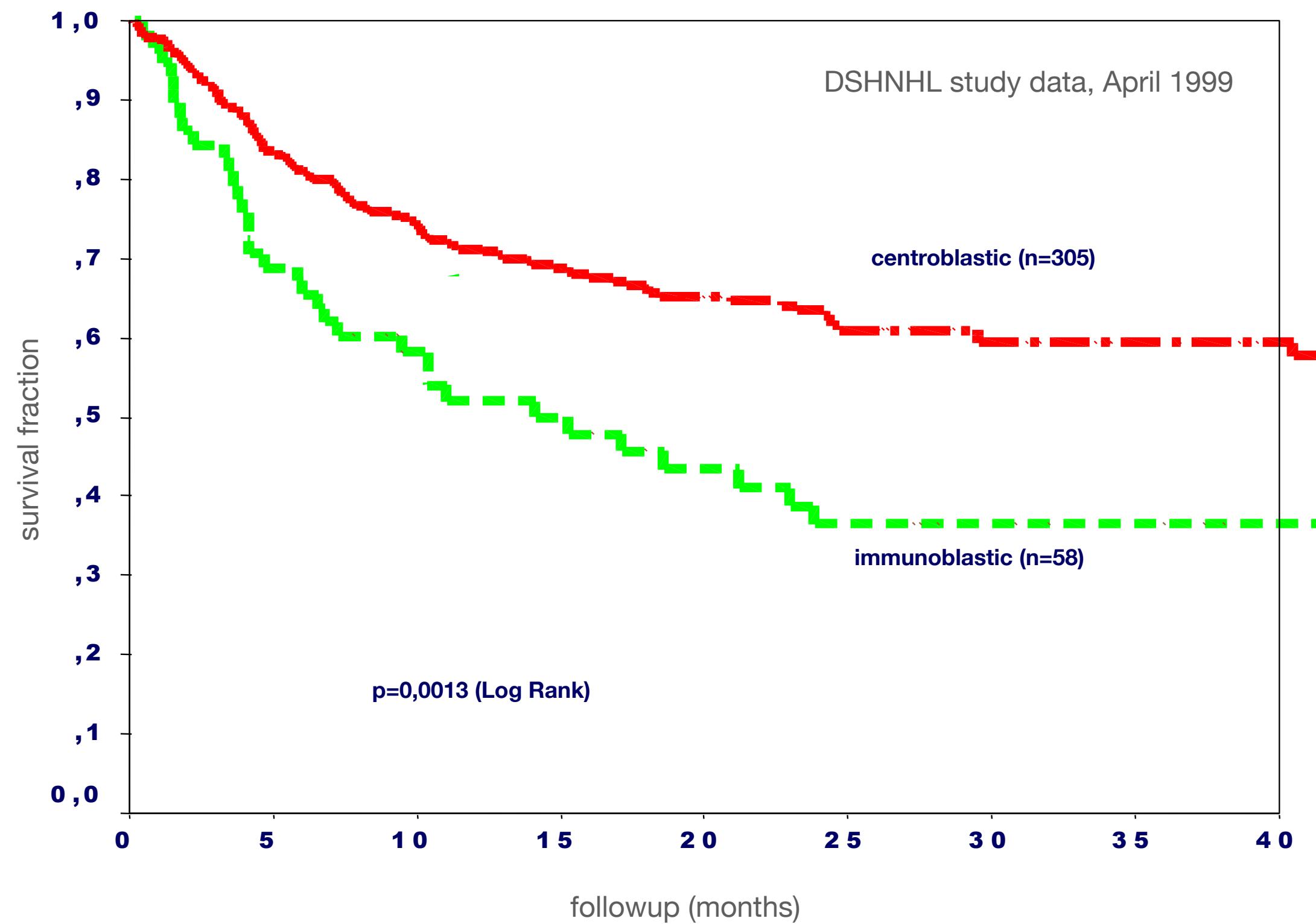
Cancer CNVs | Diagnostics | Prognosis

Single-study CNV frequencies correspond to diagnostic subsets



Cancer CNVs | Diagnostics | Prognosis

Single-study CNV frequencies correspond to diagnostic subsets



Let's build a database!



dilbert.com | Tuesday February 27, 1996

... using archaic tools



dilbert.com | Tuesday September 08, 1992

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

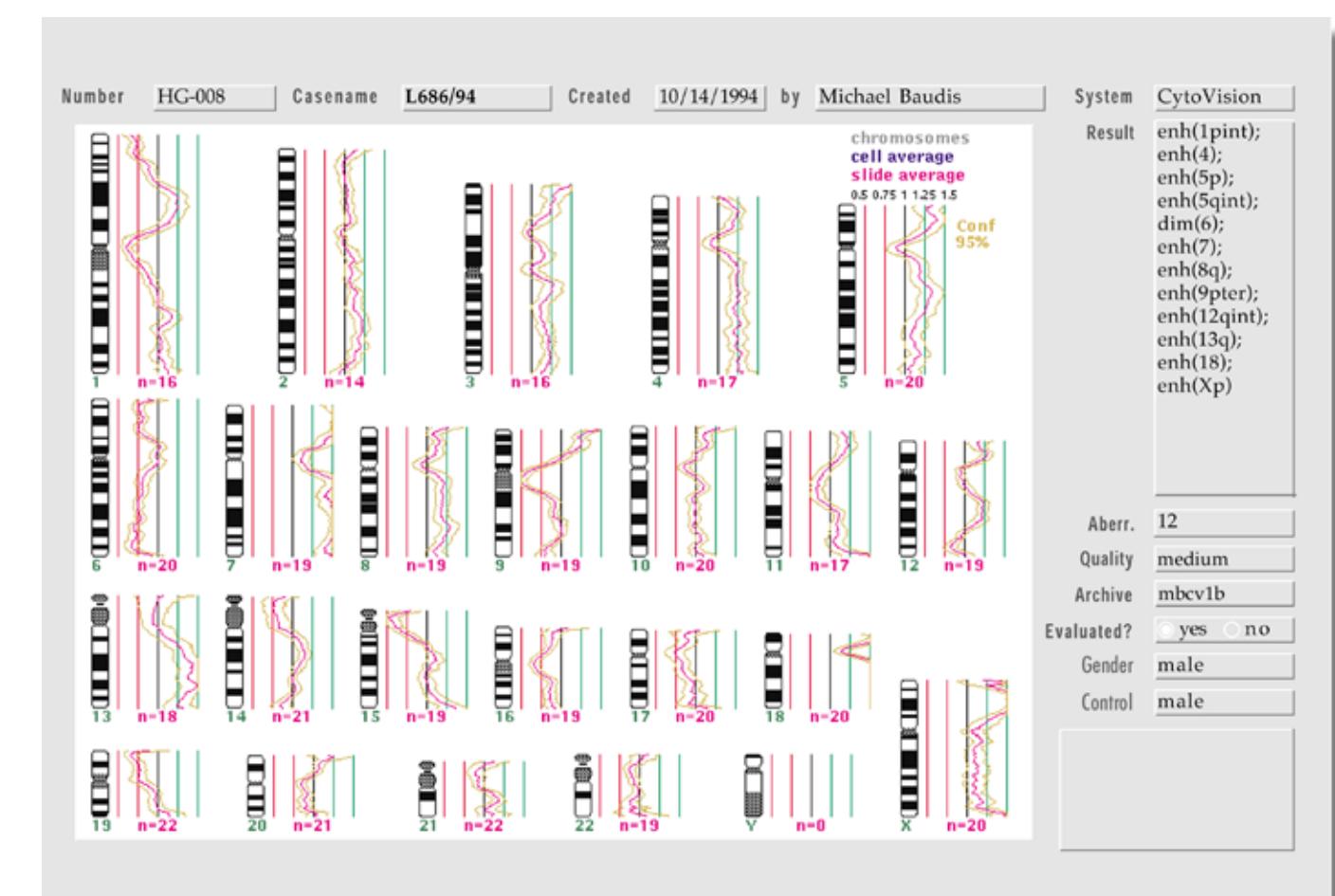
- Dec 6, 2000
 - first time online
- Nov 30, 2000
 - addition of graphical representation and gene table
- Nov 17, 2000
 - generation of website layout and database automatisation

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?

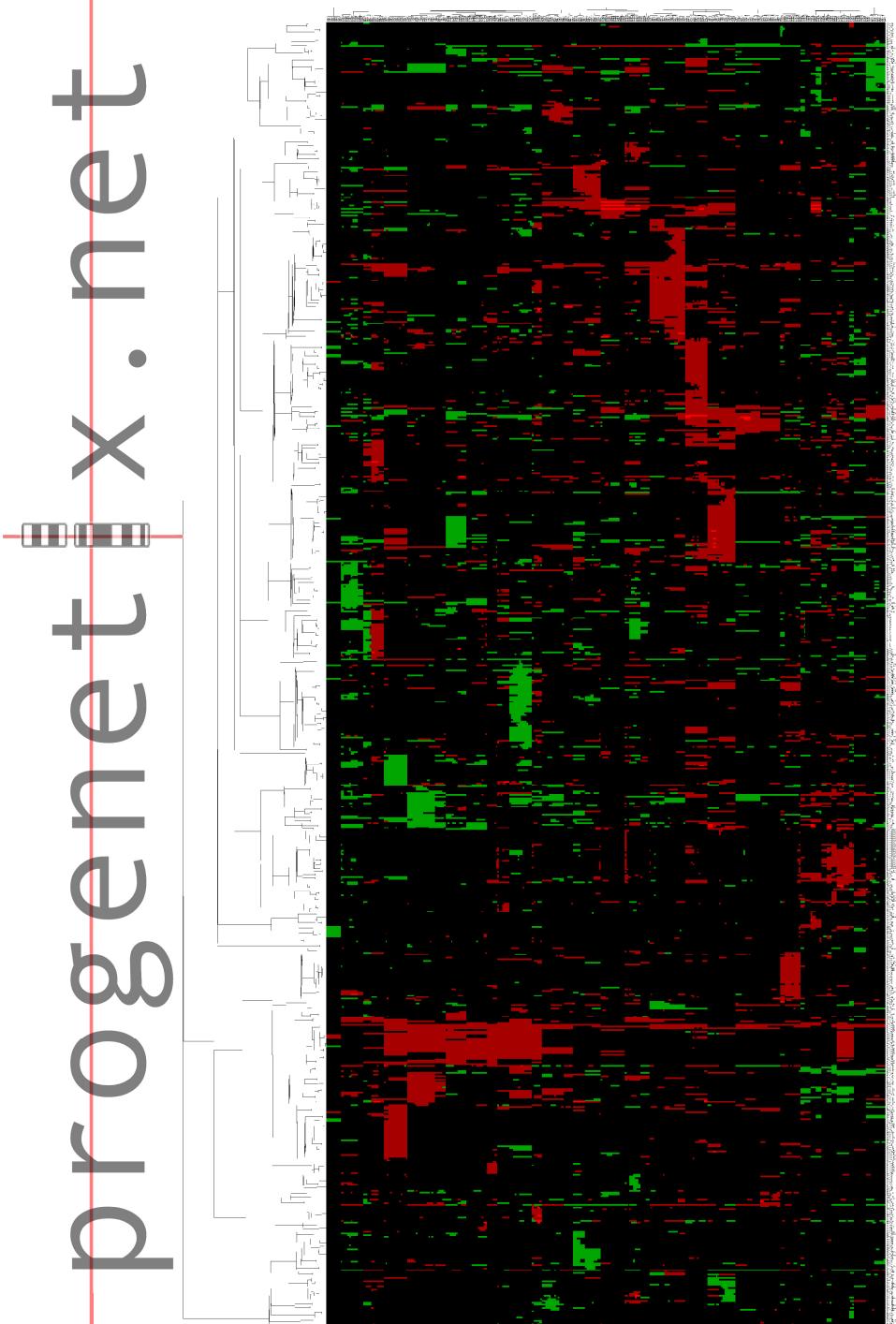


Over the last decade, techniques for the genome wide scanning for genomic imbalances in malignant neoplasia have been developed, e.g. Comparative Genomic Hybridization (CGH).

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

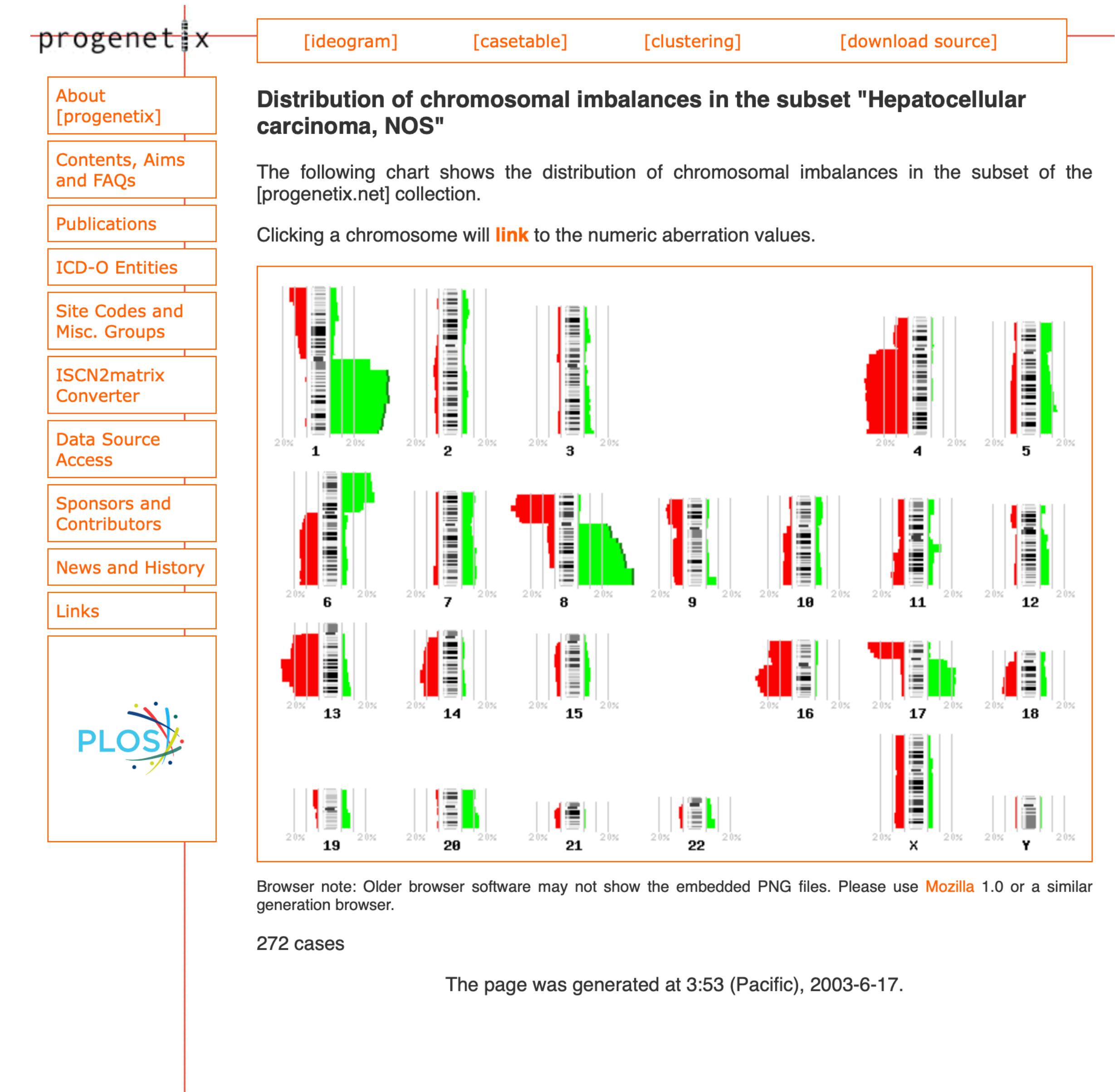
List of cases included in the subset "Hepatocellular carcinoma, NOS"

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

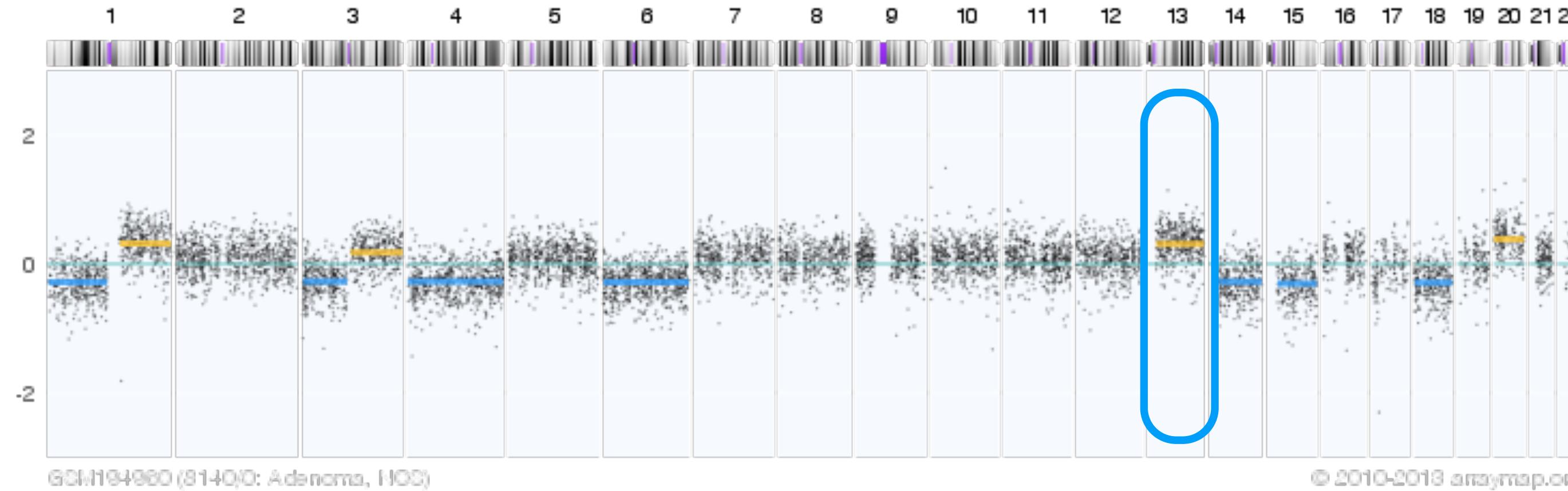
W. Michael Korn,¹ Toru Yasutake,² Wen-Lin Kuo,¹ Robert S. Warren,³ Colin Collins,¹ Masao Tomita,² Joe Gray,¹ and Frederic M. Waidman¹

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

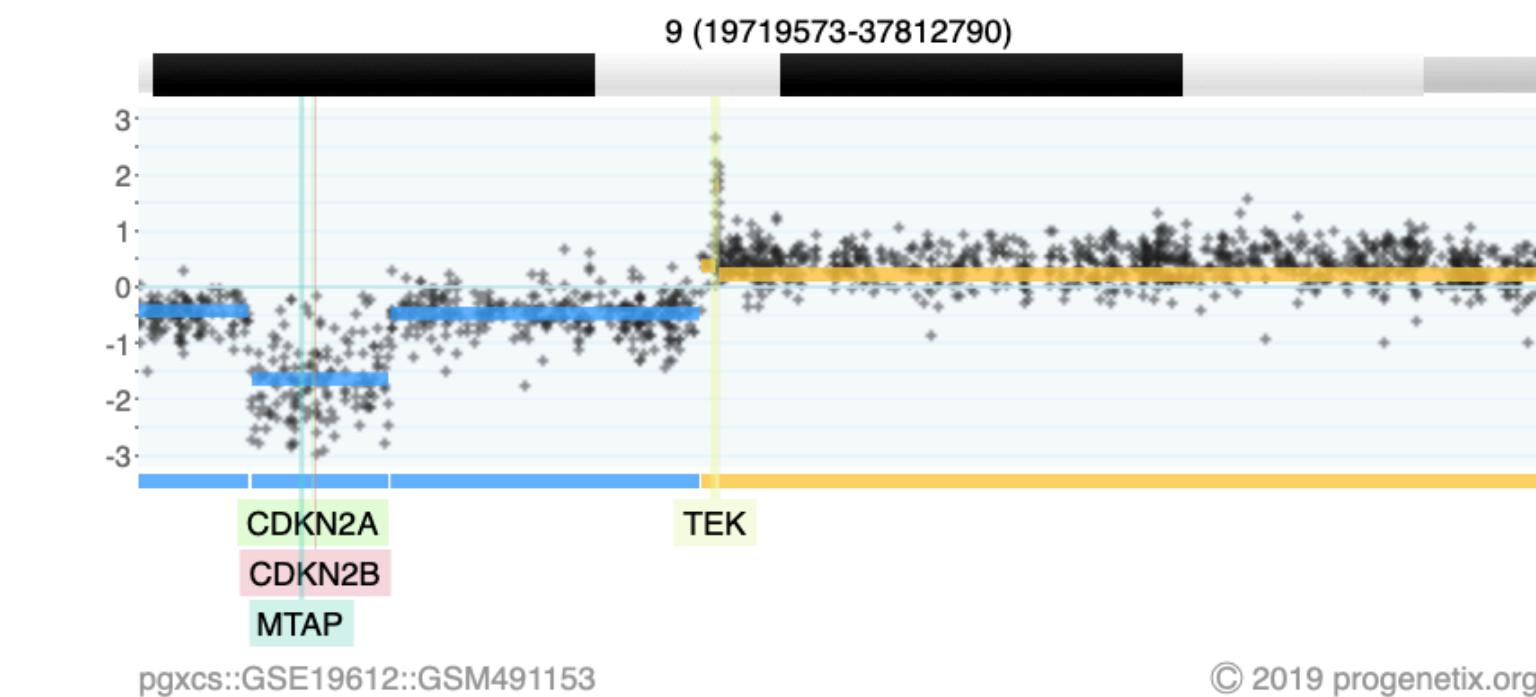
<https://progenetix.org/2003-06-17/>

progenetix

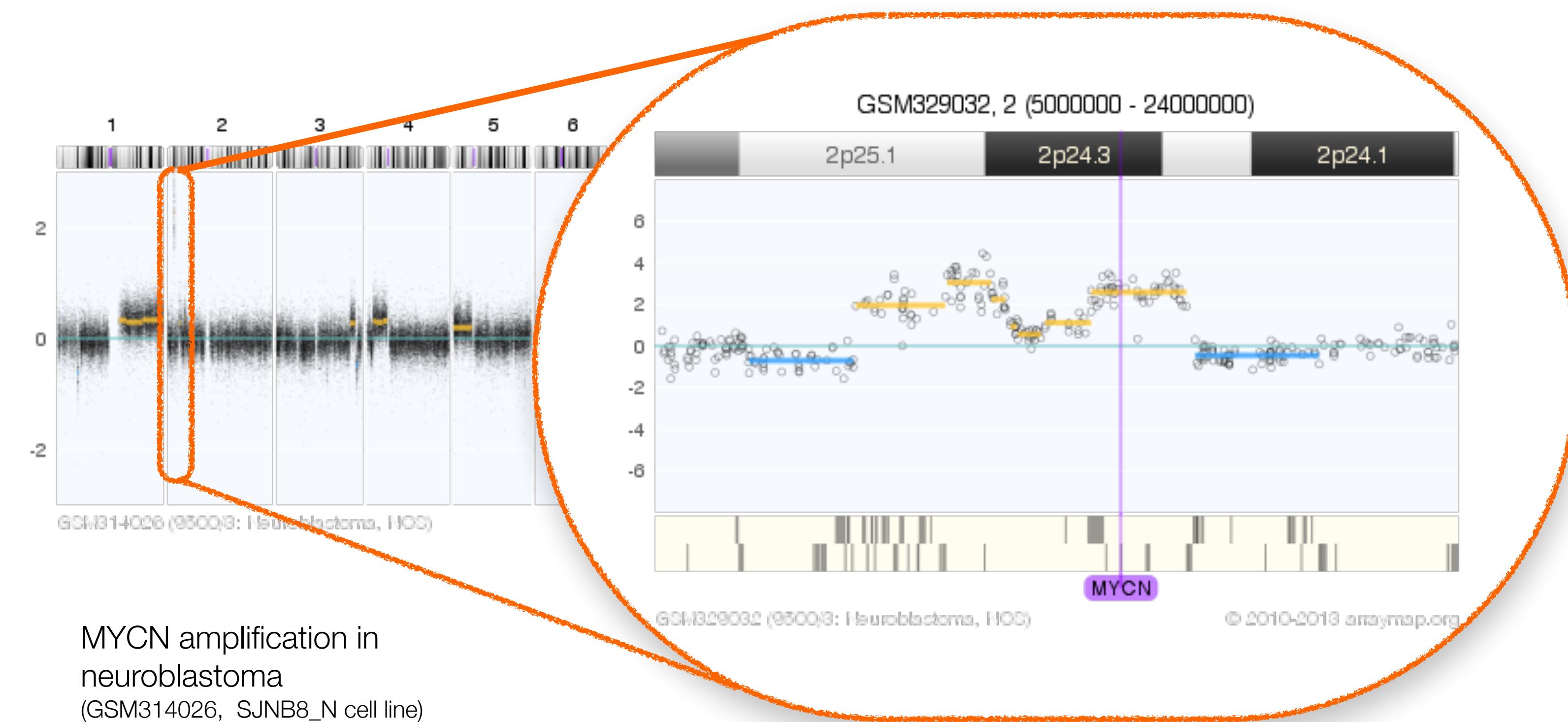
Array-based Detection of Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma



MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

low level/high level copy number alterations (CNAs)

arrayMap



arrayMap (2012 - 2020)

Probe-Level Genomic Array Data in Cancer



Search Samples

Search Publications

Progenetix



Citation & Licensing

User Guide

People

Beacon⁺

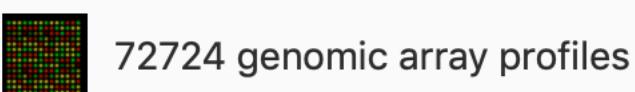


162.158.150.56

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:



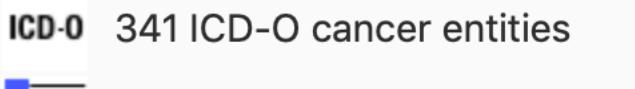
72724 genomic array profiles



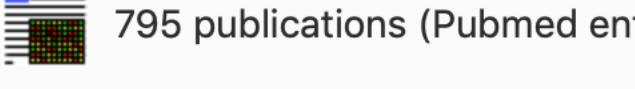
898 experimental series



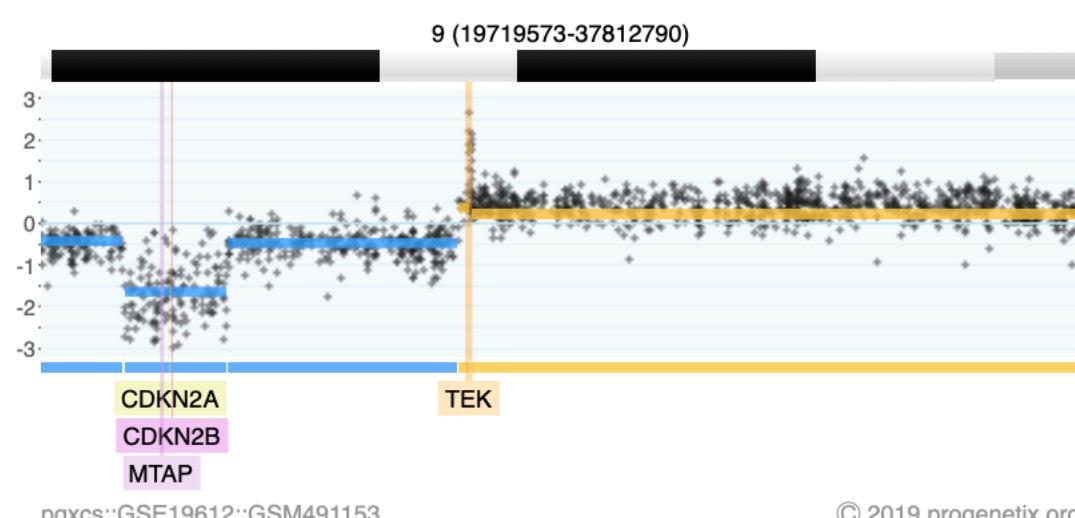
257 array platforms



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.

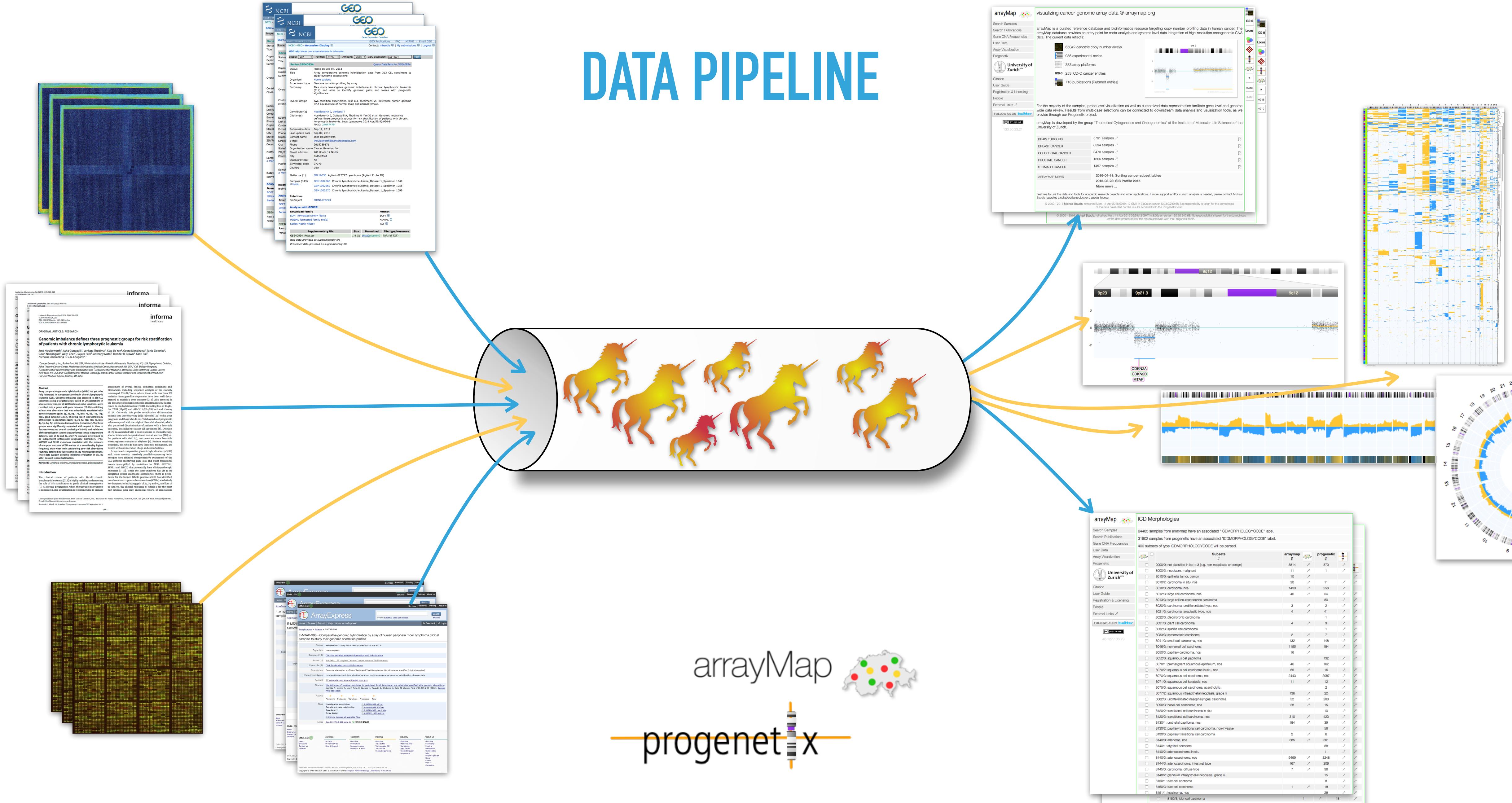
© 2000 - 2019 Michael Baudis, refreshed 2019-06-12T21:00:19Z in 6.00s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.



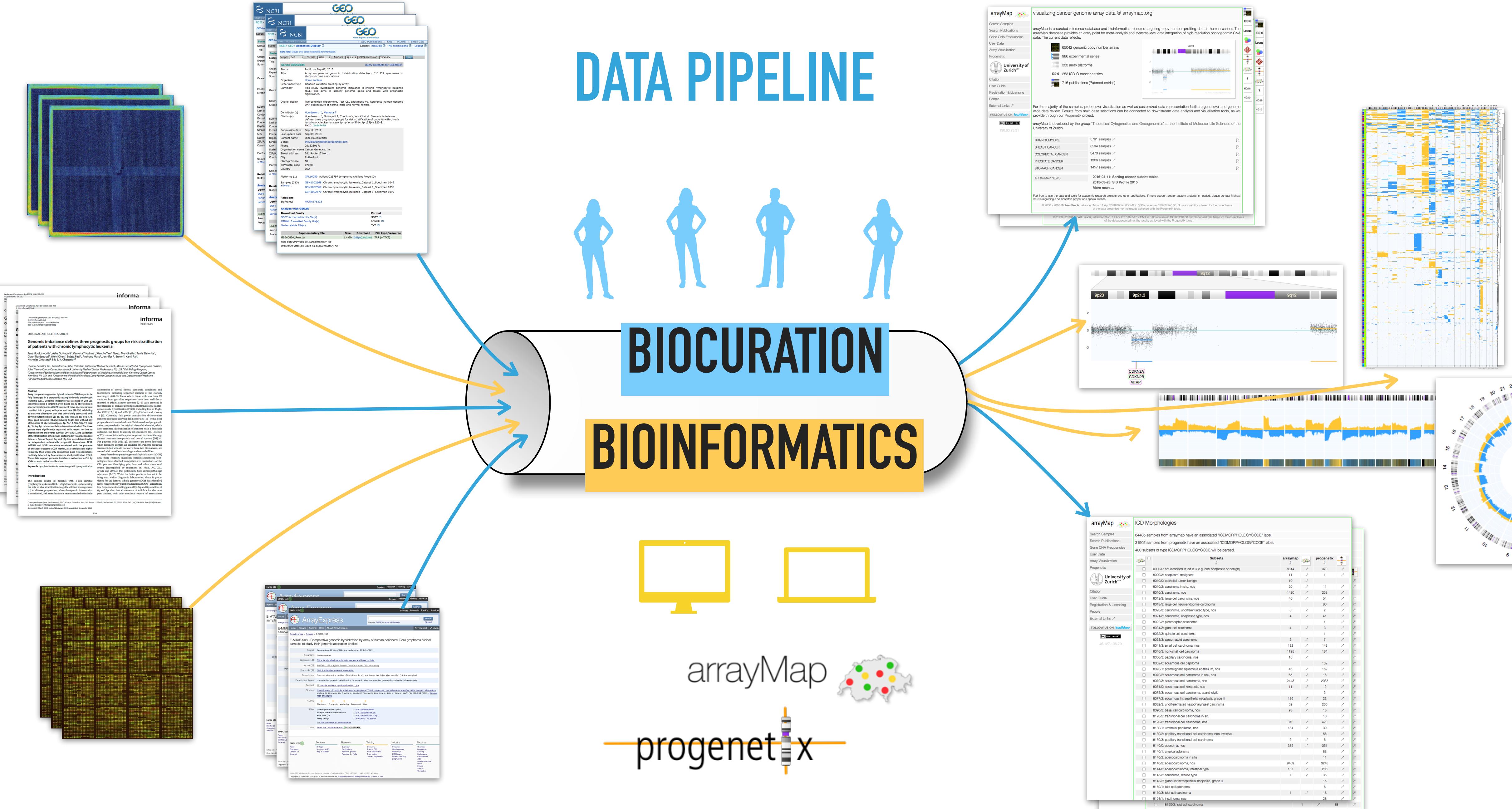
arrayMap



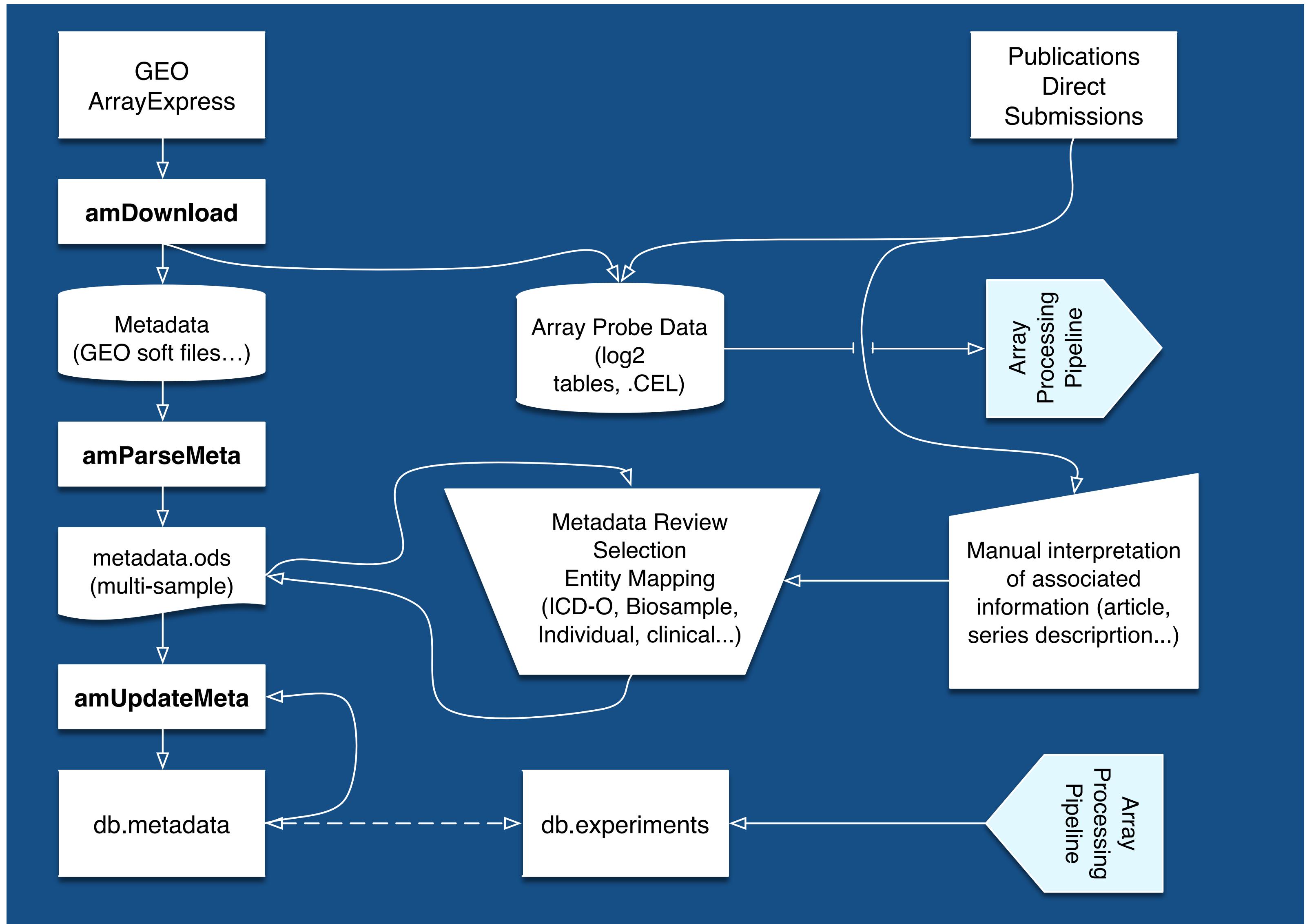
DATA PIPELINE



DATA PIPELINE



Bioinformatics & Data Curation - arrayMap data “Pipeline”



Recent Publications

CNV Data Analysis & Methods

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



Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes

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

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



ORIGINAL PAPER

CNARA: reliability assessment for genomic copy number profiles

Ni Ai^{1*}, Haoyang Cai², Caius Solovan³ and Michael Baudis^{1*}

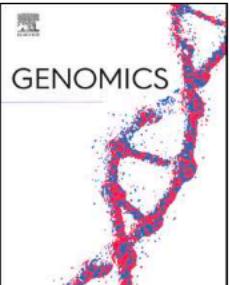
Ai et al. *BMC Genomics* (2016) 17:799
DOI 10.1186/s12864-016-3074-7



Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno



RESEARCH ARTICLE
Minimum error calibration and normalization for genomic copy number analysis

Bo Gao^{a,b}, Michael Baudis^{a,b,*}

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

Haoyang Cai^{1,2}, Nitin Kumar^{1,2}, Homayoun C Bagheri³, Christian von Mering^{1,2}, Mark D Robinson^{1,2*} and Michael Baudis^{1,2*}

SOFTWARE TOOL ARTICLE

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

Bo Gao 1,2, Qingyao Huang^{1,2}, Michael Baudis 1,2

OPEN **Enabling population assignment from cancer genomes with SNP2pop**

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



SCIENTIFIC
METHODS
research

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



Cancer Classifications need an Einstein to sort them out



BRADY'S NCI:038 NCI:BRADY'S MORPHOLOGY CODES
GSM393858 C2853 Acute Myeloid Leukemia Not Otherwise Specified 9861/3 C42
GSM302285 C2852 Adenocarcinoma 8140/3 C34
GSM918983 C3222 Medulloblastoma 9480/3 C716
GSM551398 C4017 Ductal Breast Carcinoma 8500/3 C50
GSM412384 C3163 Chronic Lymphocytic Leukemia 9823/3 C42
GSM1218286 C4017 Ductal Breast Carcinoma 8500/3 C50
GSM714412 C2852 Adenocarcinoma 8140/3 C569
GSM1109923 C9306 Soft Tissue Sarcoma 8800/3 C499
GSM711848 C2852 Adenocarcinoma 8140/3 C25
GSM746294 C89426 8022/2 C53
GSM1981528 C4017 Ductal Breast Carcinoma 8500/3 C50
GSM281399 C8949 8500/2 C50
GSM533469 C9349 Plasmacytoma 9831/3 C42



Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

!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 Grootplein 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>

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
manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard
701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix
!Sample_scan_protocol = Scanning performed according to the standard Affymetrix
or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix scanner 3000
!Sample_description = primary ALL diagnosis sample
!Sample_data_processing = copy number detection using CNAG2.0 software (http://
!Sample_platform_id = GPL3718
!Sample_contact_name = Roland,P.,Kuiper
!Sample_contact_email = r.kuiper@antrg.umcn.nl, e.verwiel@antrg.umcn.nl
!Sample_contact_phone = +31243610868
!Sample_contact_fax = +31243668752
!Sample_contact_department = Human Genetics
!Sample_contact_institute = Radboud University Nijmegen Medical Centre
!Sample_contact_address = Geert Grootplein 10
!Sample_contact_city = Nijmegen
!Sample_contact_zip/postal_code = 6525GA
!Sample_contact_country = Netherlands
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CEL.gz
```

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

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

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

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

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

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 }) }
```

Data Curation

Happy RegExing!



```
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).*?(?:(:deaf?:d|th))|alive|surviv|outcome|status'
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'
41     s: 'survival: dead'
42     - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\\w]+?NA'
43     s: ''
44     - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\\w]+?CR'
45     s: 'survival: alive'
46     - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\\w]+?RD'
47     s: '# alive but not responding to therapy so removed?'
48     - m: 'Event Free Survival[^\\w]+?no event'
49     s: 'recurrence: no'
50     - m: 'Event Free Survival.event'
51     s: 'recurrence: yes'
52     - m: 'Outcome[^\\w]+?no event'
53     s: 'survival: alive'
54     - m: 'Outcome[^\\w]+?event'
55     s: 'survival: dead'
56     - m: 'survival status[^\\w]+?0'
57     s: 'survival: dead'
58     - m: 'survival status[^\\w]+?1'
59     s: 'survival: alive'
60     - m: 'overall[^\\w]+?survival[^\\w]+?days[^\\w]+?NA'
61     s: ''
62     - m: 'survival(?: time|from diagnosis)?[^\\w]+?(days|months|years?)[^\\w]+?(\\d\\d?\\d?\\d?\\.?\\d?\\d?)'
63     s: 'survival: \\2\\1'
```

Disease annotations in Progenetix

From some text, somewhere, to ontology classes

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

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

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

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

Standardized Data

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

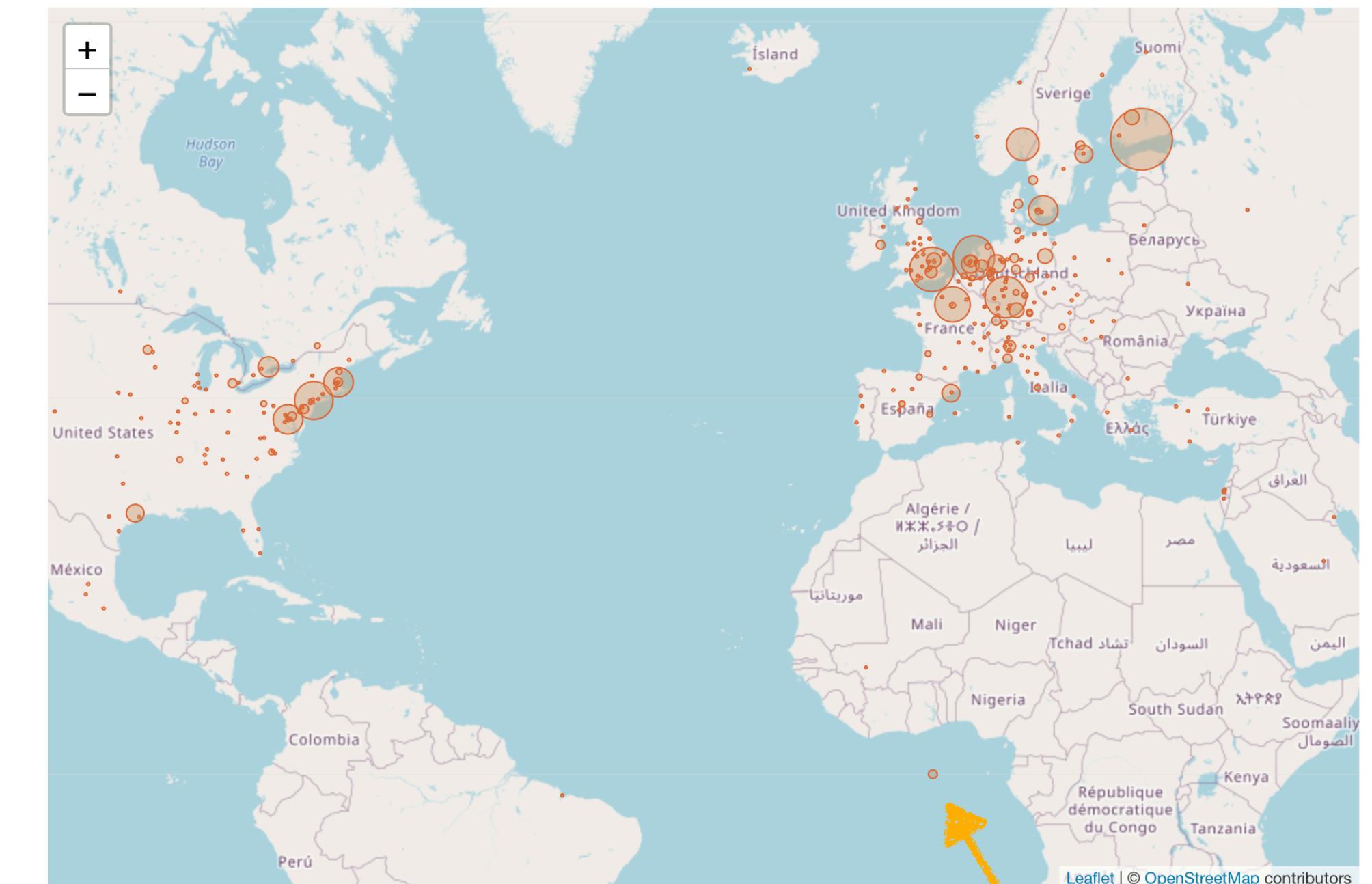
- Multiple international initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of data annotation standards
- emerging / established principles are the use of hierarchical coding systems where individual codes are represented as CURIEs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

```
"label" : "no restriction",
"id" : "DUO:0000004"
},
"provenance" : {
  "material" : {
    "type" : {
      "id" : "EF0:0009656",
      "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 ...

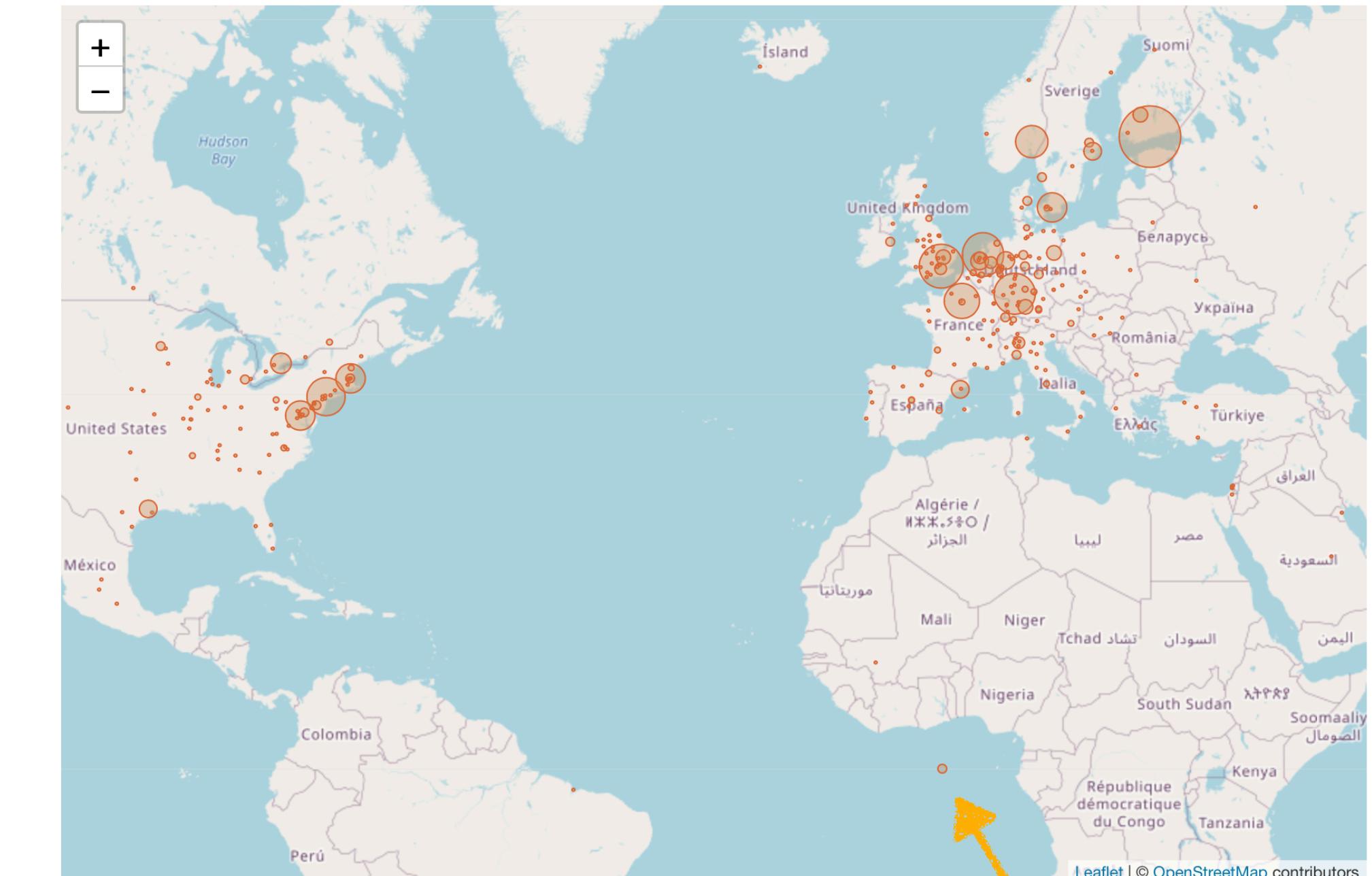
The most geo-tagged place on earth is Null Island



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

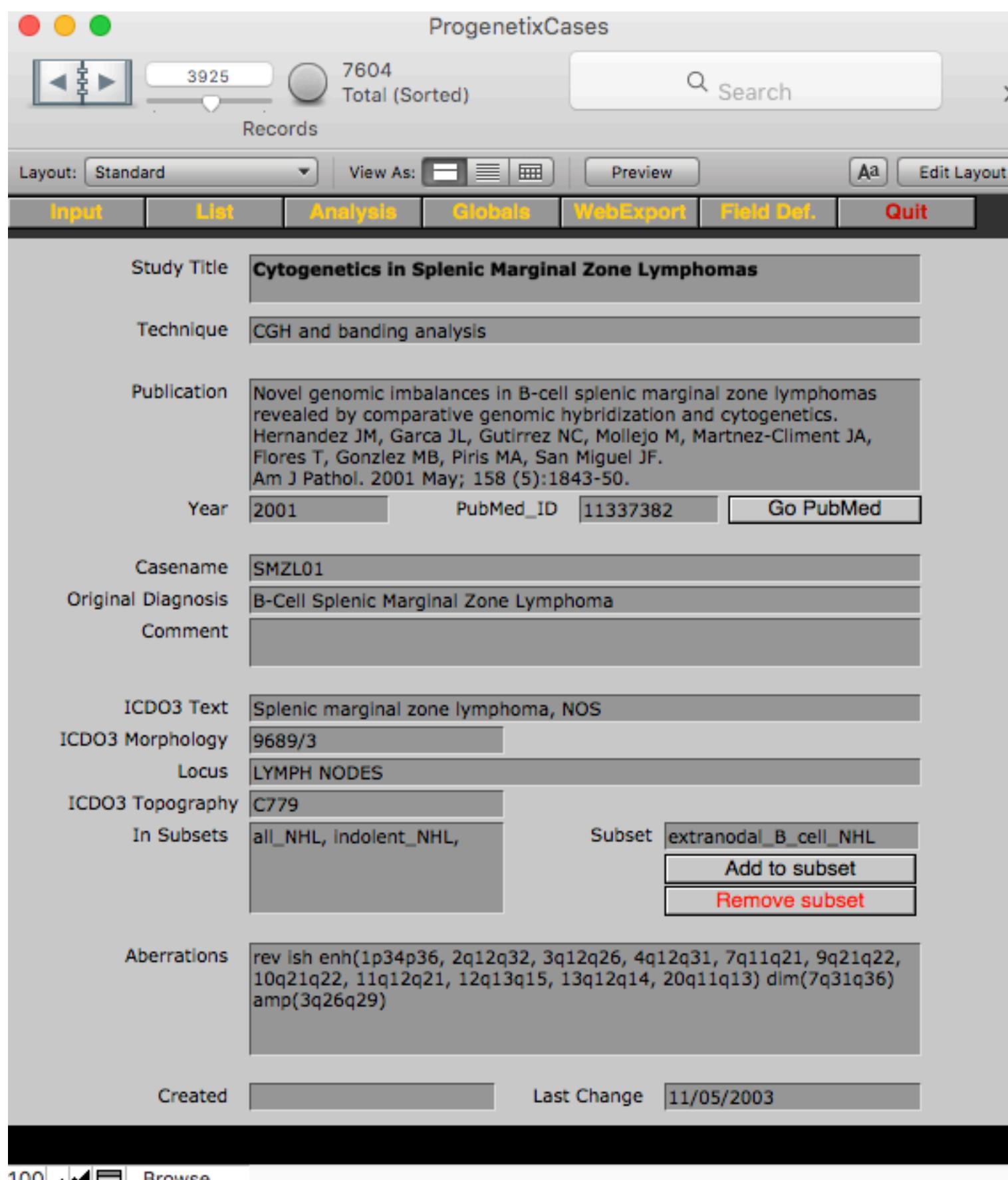


Progenetix publication collection
progenetix.org/publications/list
2020-11-28

25 / 3306 publications

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

2021

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

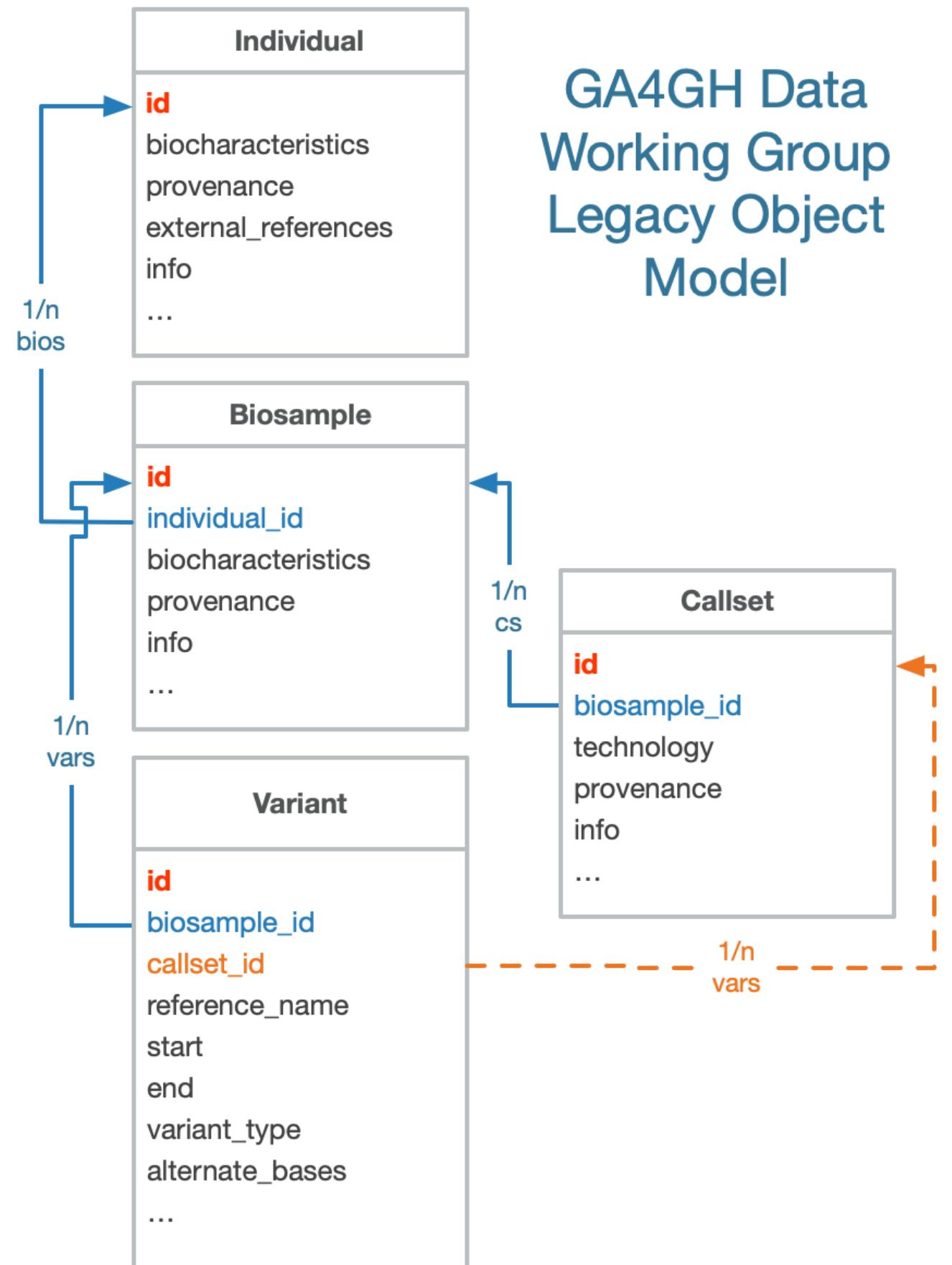
```
{
  "id" : "pgxind-kftx394x",
  "biocharacteristics" : [
    {
      "description" : "female",
      "type" : {
        "id" : "PATO:0020002",
        "label" : "female genotypic sex"
      }
    },
    {
      "description" : null,
      "type" : {
        "id" : "NCBITaxon:9606",
        "label"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

{
  "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"
}
```

```
{
  "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"
    }
  }
}
```

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

2021

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

```

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

{
    "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"
}
    
```

```

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

Progenetix in 2022

An oncogenomic reference resource



Progenetix in 2022

Cancer Genomics Reference Resource

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

Cancer CNV Profiles

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

Search Samples

arrayMap

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

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Beacon⁺

Documentation

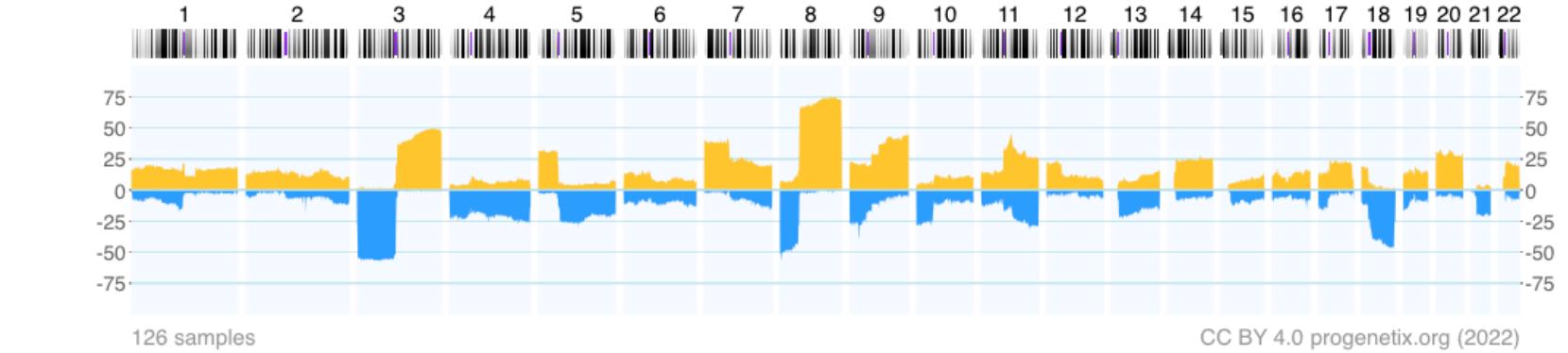
- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

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

Floor of the Mouth Neoplasm (NCIT:C4401)



[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

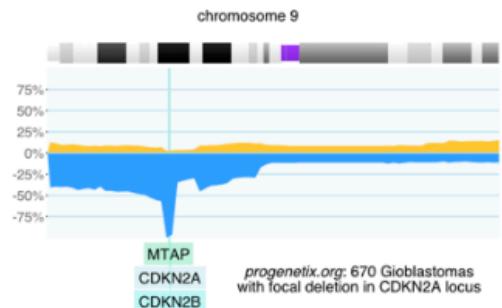
Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.

Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

Progenetix Use Cases

Local CNV Frequencies

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



progenetix.org: 670 Glioblastomas with local deletion in CDKN2A locus

Cancer CNV Profiles

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

Cancer Genomics Publications

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

Progenetix in 2022

Cancer Genomics Reference Resource

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

progenetix

Search Samples Modify Query

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

progenetix

Samples: 668 Variants: 286 Calls: 675

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

UCSC region JSON Response Visualization options

Results Biosamples Biosamples Map Variants

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

75% 50% 25% 0% -25% -50% -75%

-75% -50% -25% 0% 25% 50% 75%

progenetix: 670 samples CC BY 4.0 progenetix.org (2021)

Matched Subset Codes

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

Progenetix in 2022

Cancer Genomics Reference Resource

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

TCGA CNV Data

Search Genomic CNV Data from TCGA

This search page accesses the TCGA subset of the Progenetix collection, based on 22142 samples (tumor and references) from The Cancer Genome Atlas project. The results are based upon data generated by the [TCGA Research Network](#). Disease-specific subsets of TCGA data (aka. projects) can be accessed below.

TCGA Cancer samples (pgx:cohort-TCGAcancers)

11090 samples

CC BY 4.0 progenetix.org (2022)

[Download SVG](#) | [Go to pgx:cohort-TCGAcancers](#) | [Download CNV Frequencies](#)

Edit Query

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Beacon⁺

Documentation

- News
- Downloads & Use Cases
- Sevices & API

TCGA Cancer Studies

Filter subsets e.g. by prefix Hierarchy Depth: 2 levels

No Selection

- pgx:TCGA-ACC: TCGA ACC project (180 samples)
- pgx:TCGA-BLCA: TCGA BLCA project (810 samples)
- pgx:TCGA-BRCA: TCGA BRCA project (2219 samples)
- pgx:TCGA-CESC: TCGA CESC project (586 samples)

Progenetix in 2022

Cancer Genomics Reference Resource

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



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

Search Samples

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

Publication DB
Genome Profiling
Progenetix Use

Services
NCIt Mappings
UBERON Mappings

Upload & Plot

Beacon⁺

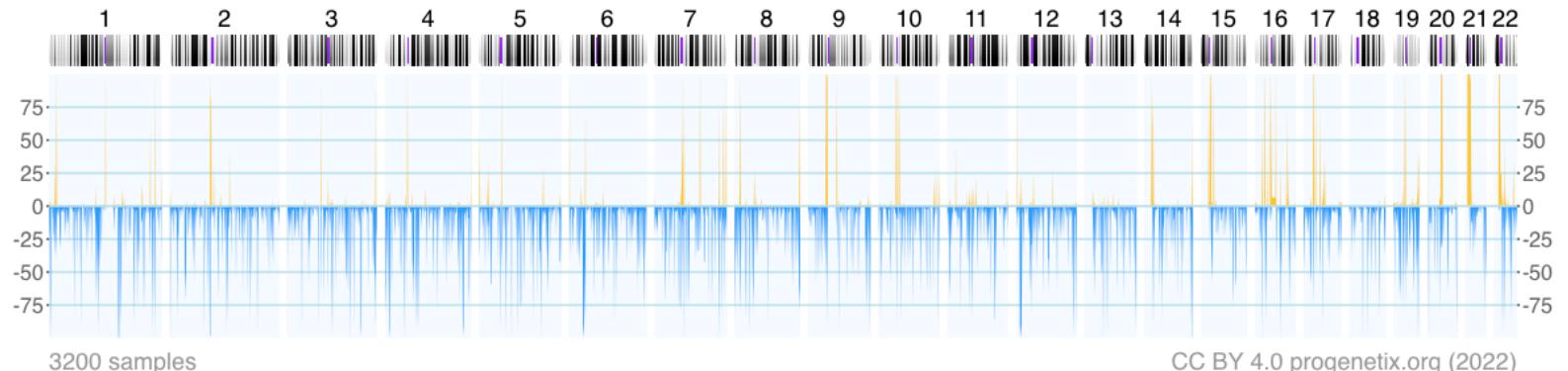
Documentation
News
Downloads & Use Cases
Sevices & API

1000 Genomes Germline CNVs

Search Genomic CNV Data from the Thousand Genomes Project

This search page accesses the reference germline CNV data of 3200 samples from the 1000 Genomes Project. The results are based on the data from the Illumina DRAGEN caller re-analysis of 3200 whole genome sequencing (WGS) samples downloaded from the AWS store of the Illumina-led reanalysis project.

1000 genomes reference samples (pgx:cohort-oneKgenomes)



Download SVG | Go to pgx:cohort-oneKgenomes | Download CNV Frequencies

Please note that the CNV spikes are based on the frequency of occurrence of any CNV in a given 1Mb interval, not on their overlap. Some genome bins may have at least one small CNV in each sample - especially in peri-centromeric regions - and therefore will display with a 100% frequency - although many of those may not overlap.

Search Samples

Range Example

Chromosome (Structural) Variant Type

Start or Position End (Range or Structural Var.)

Reference Base(s) Alternate Base(s)

Progenetix in 2022

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiling data, with focus on copy number variations (CNV)
- >116'000 cancer CNV profiles, mapped to >800 NCI^t codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCI^t, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

progenetix

Cancer CNV Profiles

Search Samples

Studies & Cohorts

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

Publication DB

Genome Profiling
Progenetix Use

Services

NCI^t Mappings
UBERON Mappings

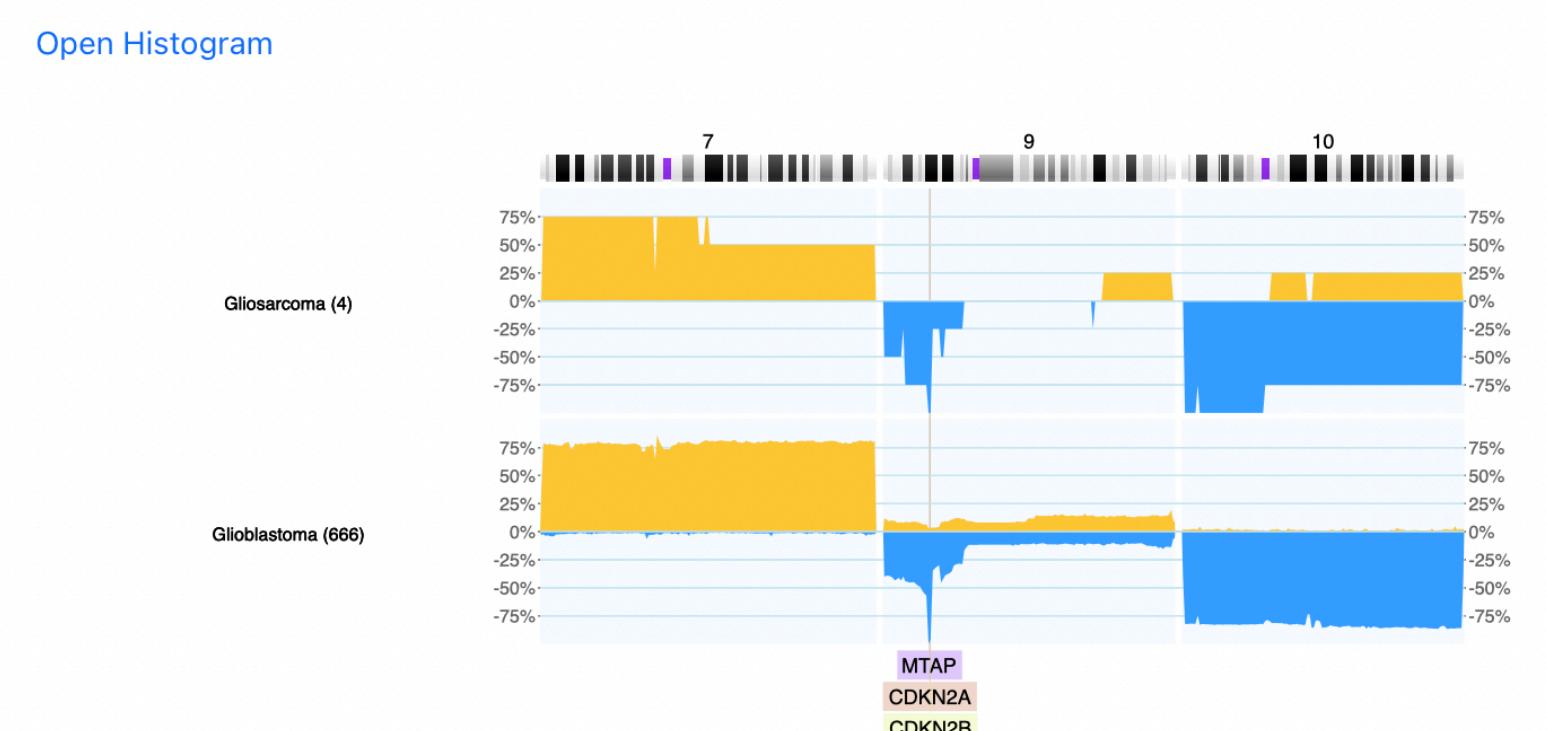
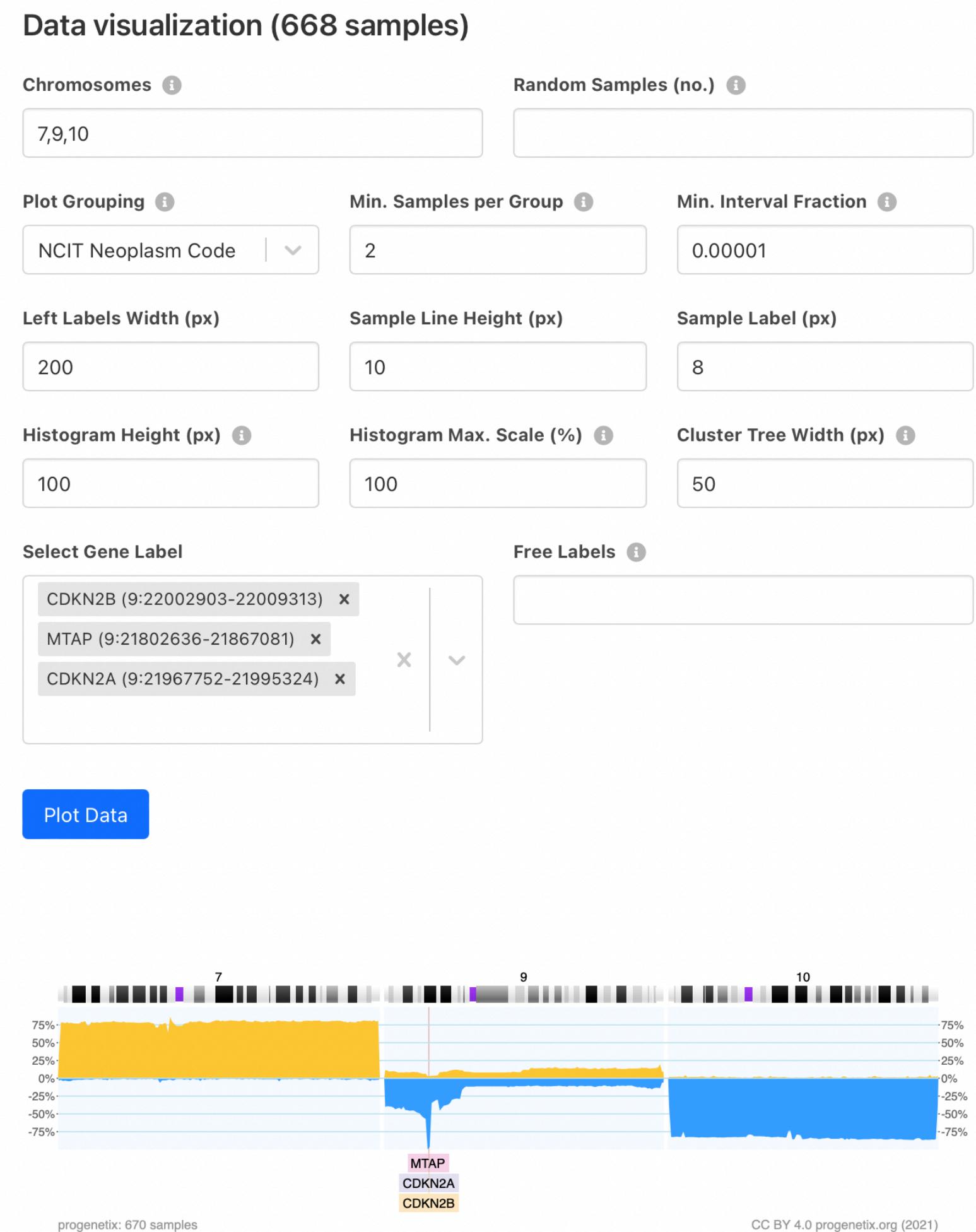
Upload & Plot

Download Data

Beacon⁺

Progenetix Info

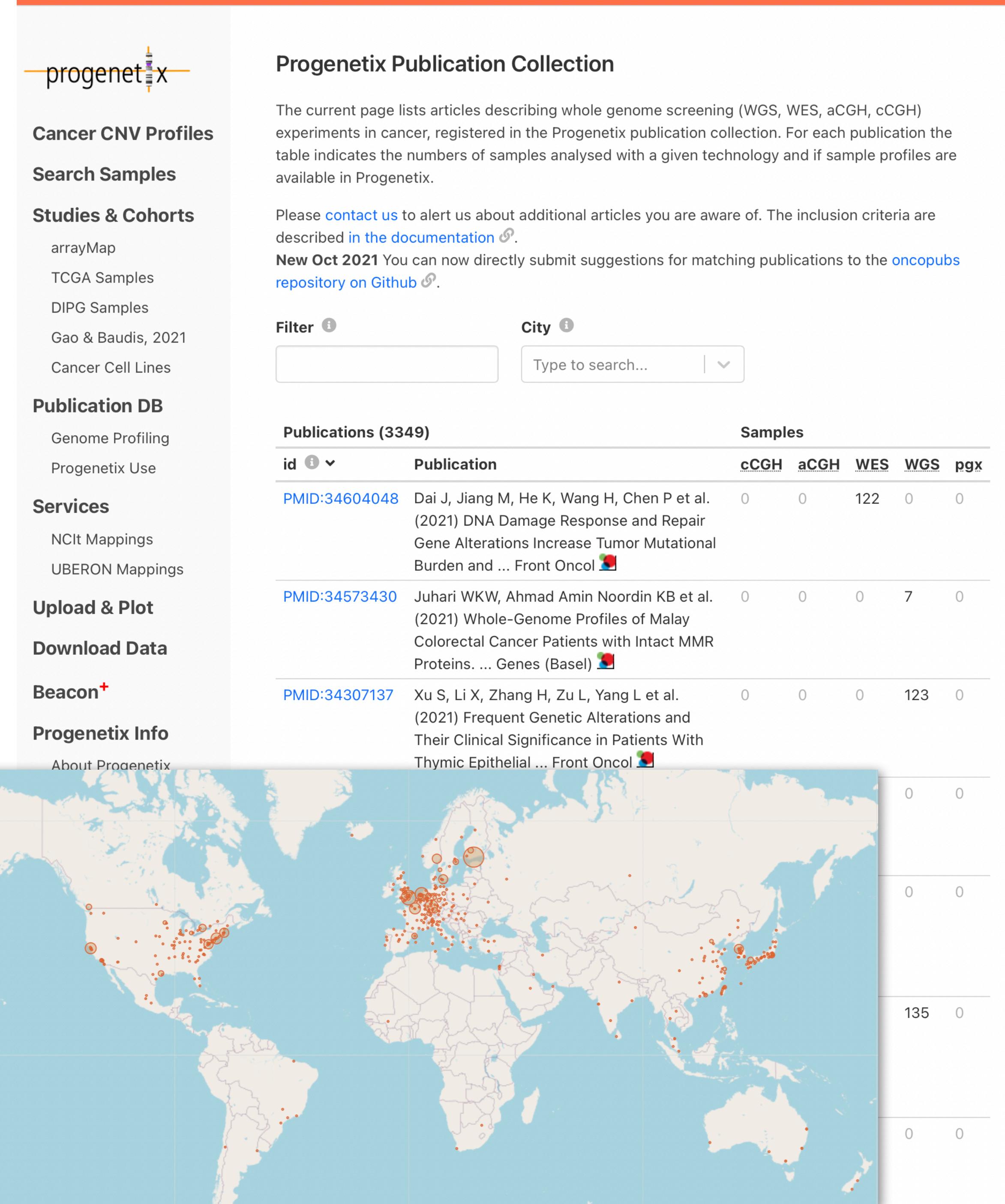
About Progenetix
Use Cases
Documentation
Baudisgroup @ UZH



Progenetix in 2022

Cancer Genomics Reference Resource

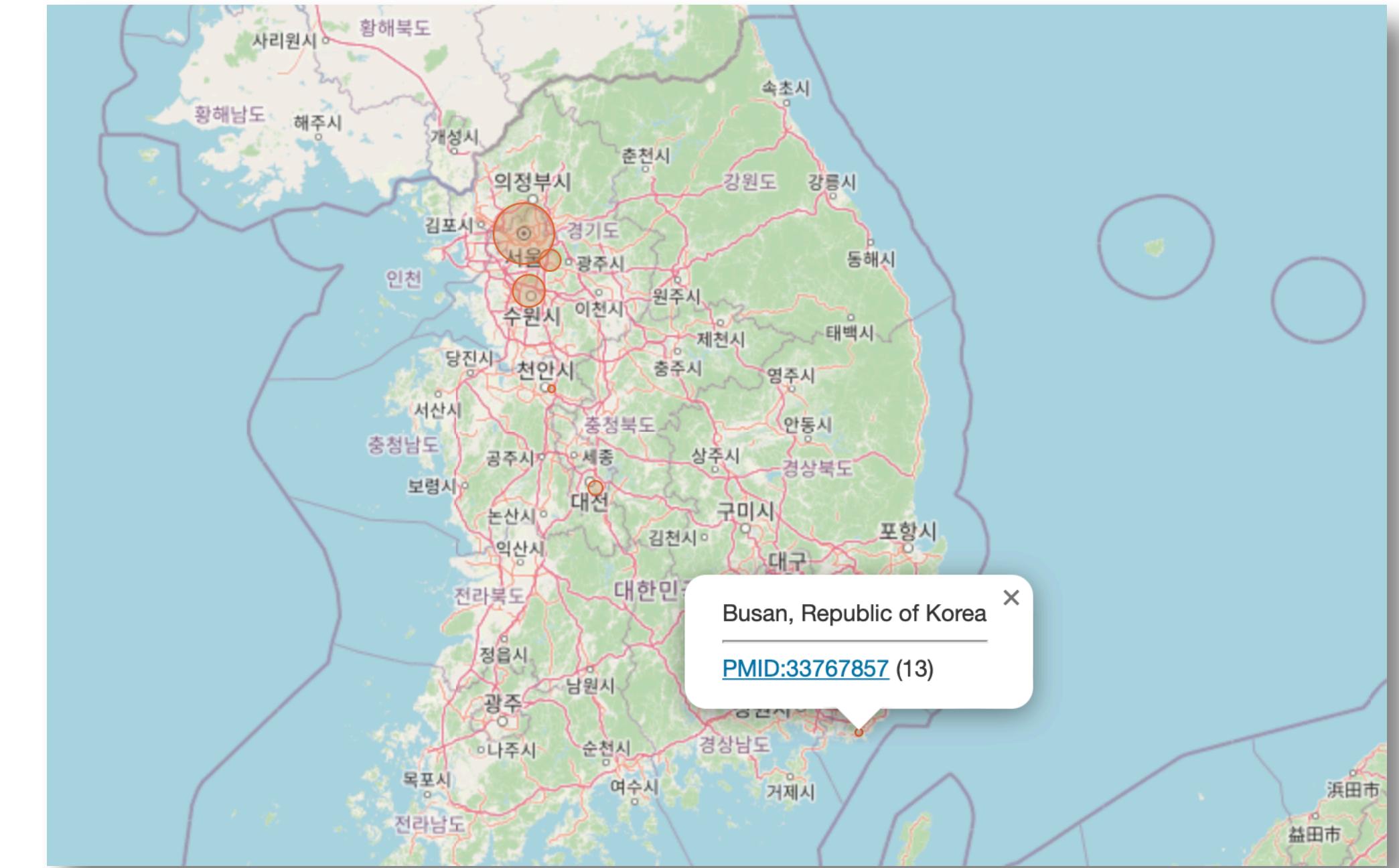
- open resource for curated oncogenomic profiles
 - >116'000 cancer CNV profiles, from >800 types
 - majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
 - standardized encodings (e.g. NCIIt, ICD-O 3)
 - identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
 - core biosample and technical metadata where accessible (TNM, sex, survival ...)
 - publication database and code mapping services



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 ↗

Progenetix in 2022

Cancer Genomics Reference Resource

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



[Cancer CNV Profiles](#)

[Search Samples](#)

[Studies & Cohorts](#)

arrayMap

TCGA Samples

DIPG Samples

Gao & Baudis, 2021

Cancer Cell Lines

[Publication DB](#)

Genome Profiling

Progenetix Use

[Services](#)

NCIt Mappings

UBERON Mappings

[Upload & Plot](#)

[Download Data](#)

[Beacon⁺](#)

[Progenetix Info](#)

About Progenetix

Use Cases

Documentation

Baudisgroup @ UZH

Services: Ontologymaps (NCIt)

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

NCIT and ICD-O 3

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

More documentation with focus on the API functionality can be found on the [documentation pages](#).

The data of all mappings can be retrieved through this API call: [{JSON}](#)

Code Selection i

NCIT:C4004: Gastric Adenocarcinoma

x | ▾

Optional: Limit with second selection

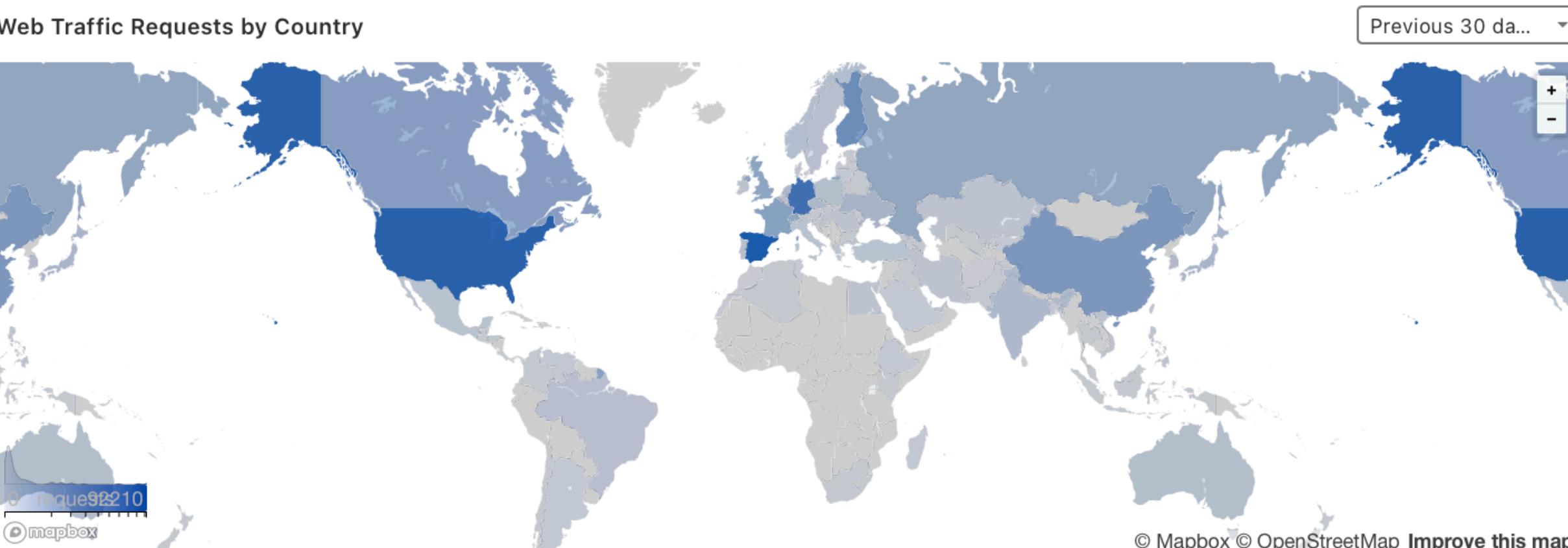
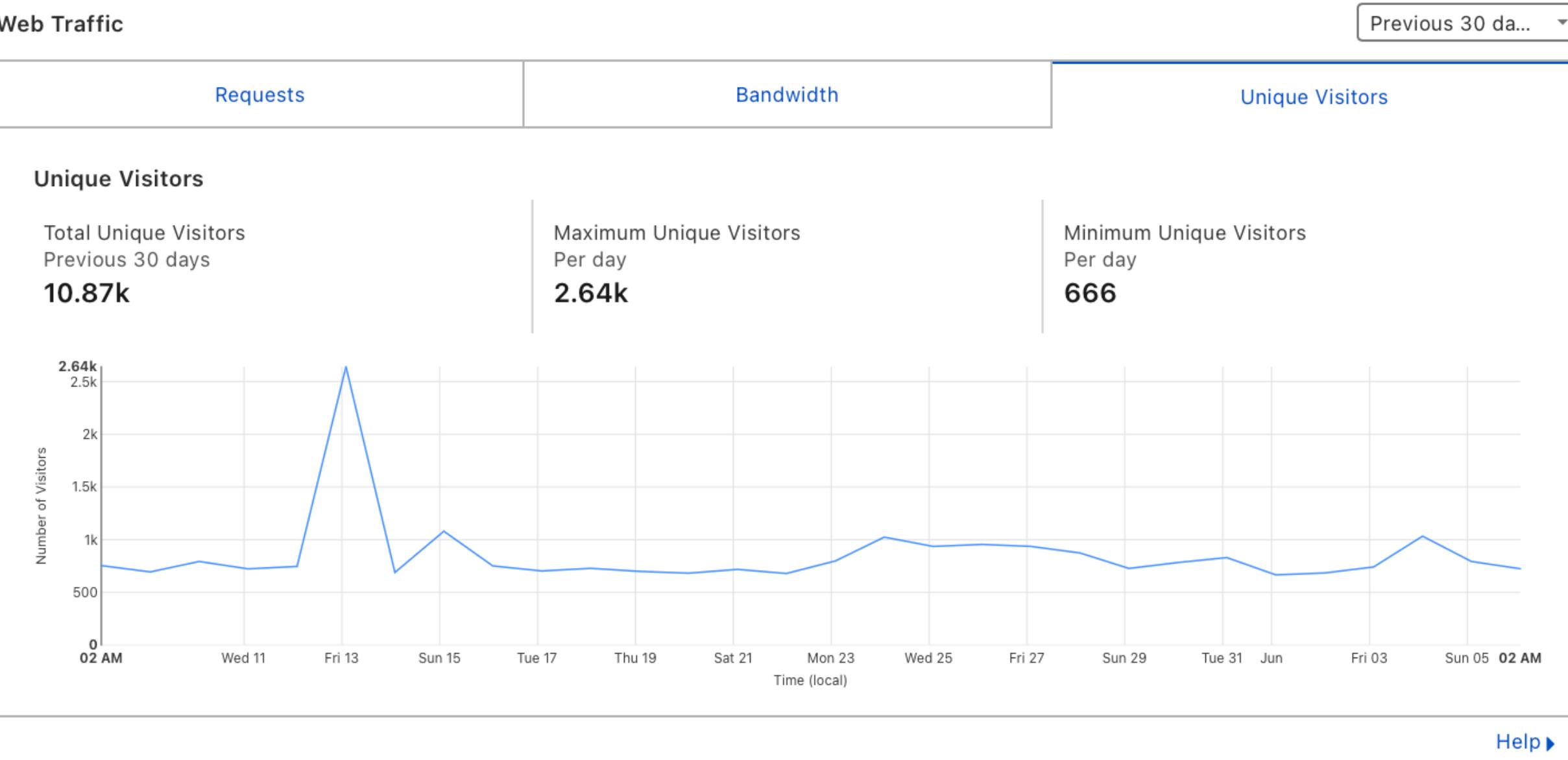
Matching Code Mappings [{JSON}](#)

NCIT:C4004: Gastric Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C16.9: stomach
NCIT:C4004: Gastric Adenocarcinoma	icdom-82603: Papillary adenocarcinoma, NOS	icdot-C16.9: stomach
NCIT:C4004: Gastric Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C16.5: Lesser curvature of stomach, NOS
NCIT:C4004: Gastric Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C16.3: Gastric antrum
NCIT:C4004: Gastric Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C16.2: Body of stomach
NCIT:C4004: Gastric Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C16.0: Cardia, NOS
NCIT:C4004: Gastric Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C16.1: Fundus of stomach
NCIT:C4004: Gastric Adenocarcinoma	icdom-82603: Papillary adenocarcinoma, NOS	icdot-C16.2: Body of stomach
NCIT:C4004: Gastric Adenocarcinoma	icdom-82603: Papillary adenocarcinoma, NOS	icdot-C16.3: Gastric antrum
NCIT:C4004: Gastric Adenocarcinoma	icdom-82553: Adenocarcinoma with mixed subtypes	icdot-C16.3: Gastric antrum

Progenetix in 2022

Cancer Genomics Reference Resource

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



Top Traffic Countries / Regions

Previous 30 days

Country / Region	Traffic
Spain	88,010
United States	59,321
Germany	28,826
Finland	21,311
Singapore	

Cloudflare | progenetix.org | June 2022

Help ▶

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.

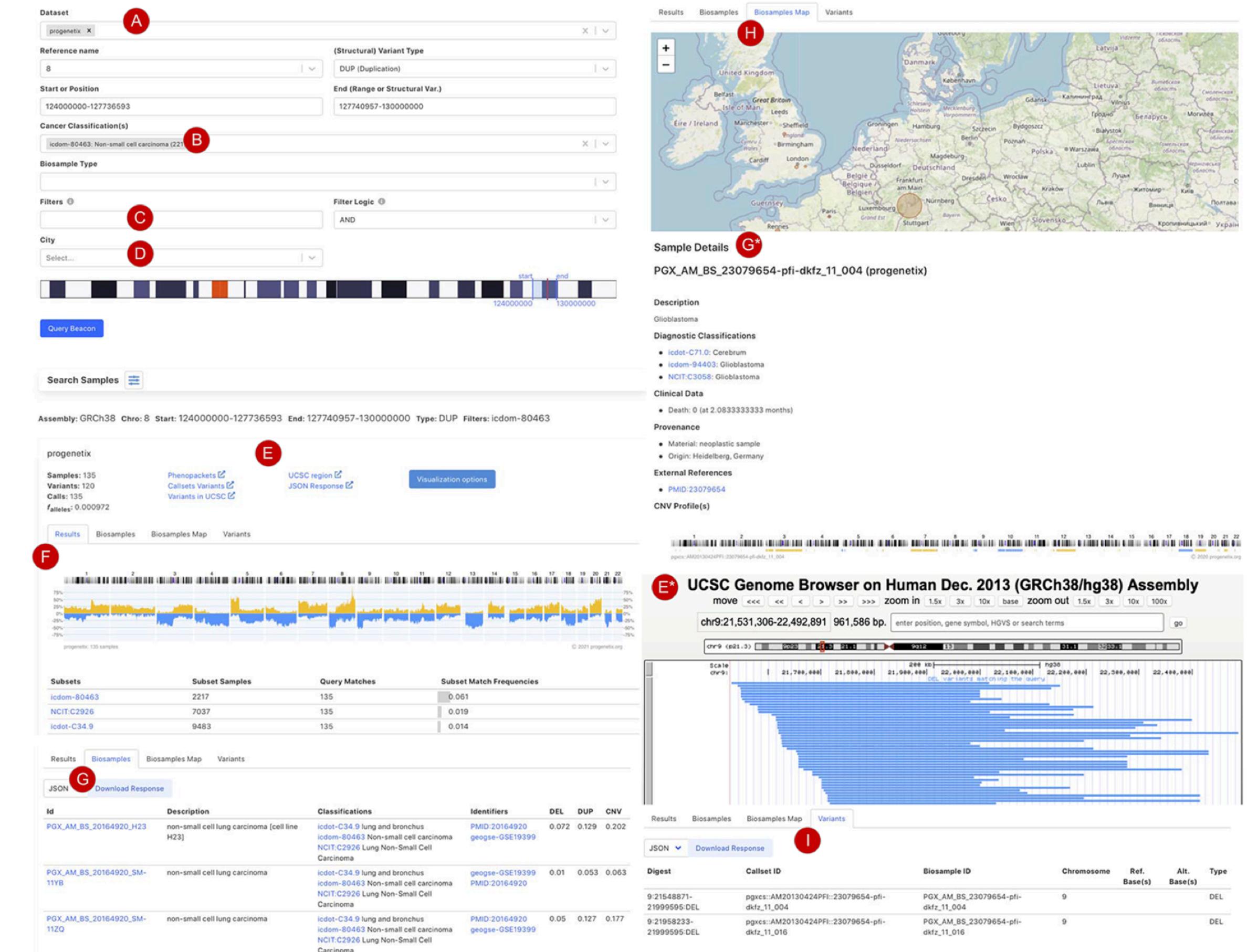
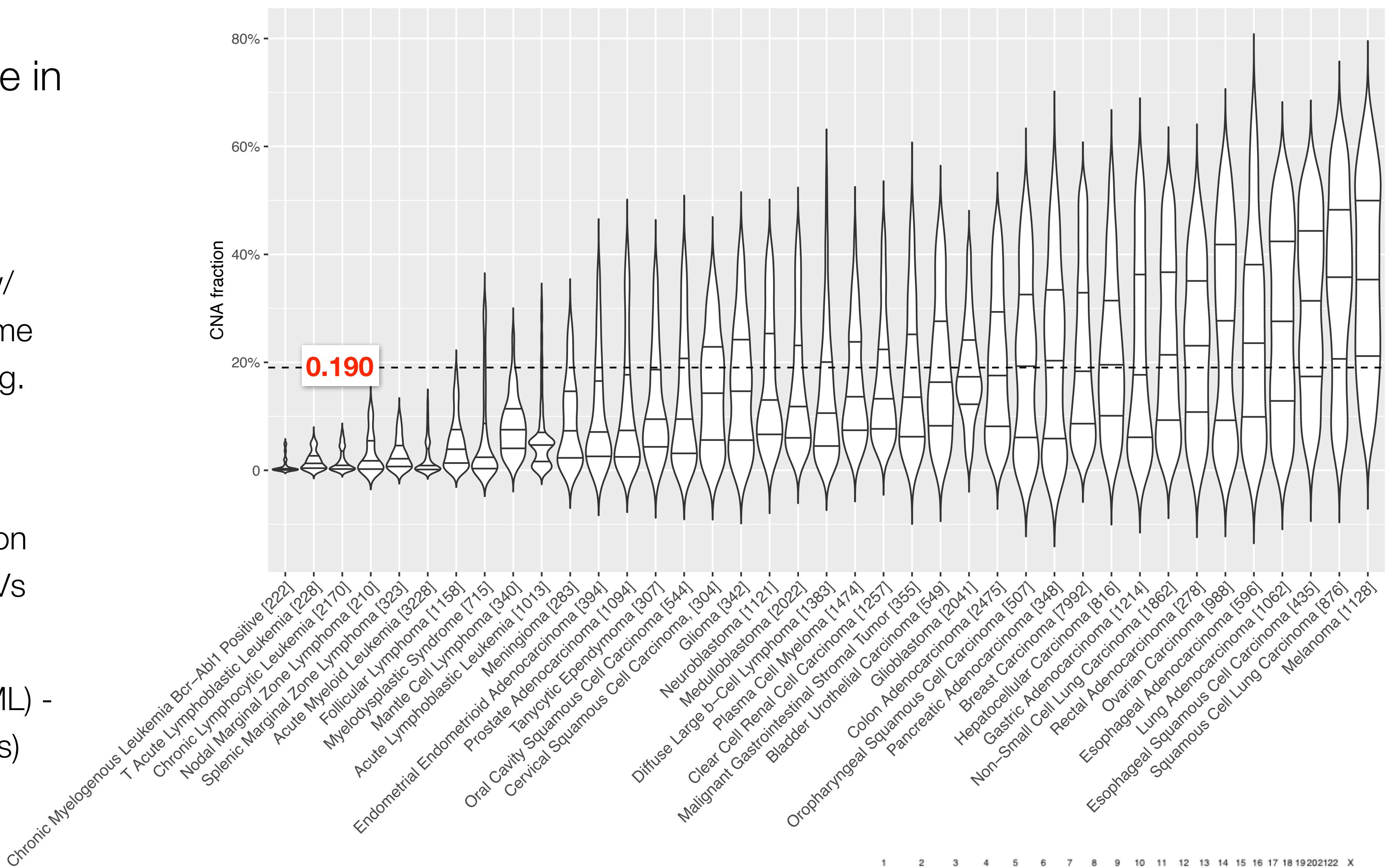


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range. This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with ≤ 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and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

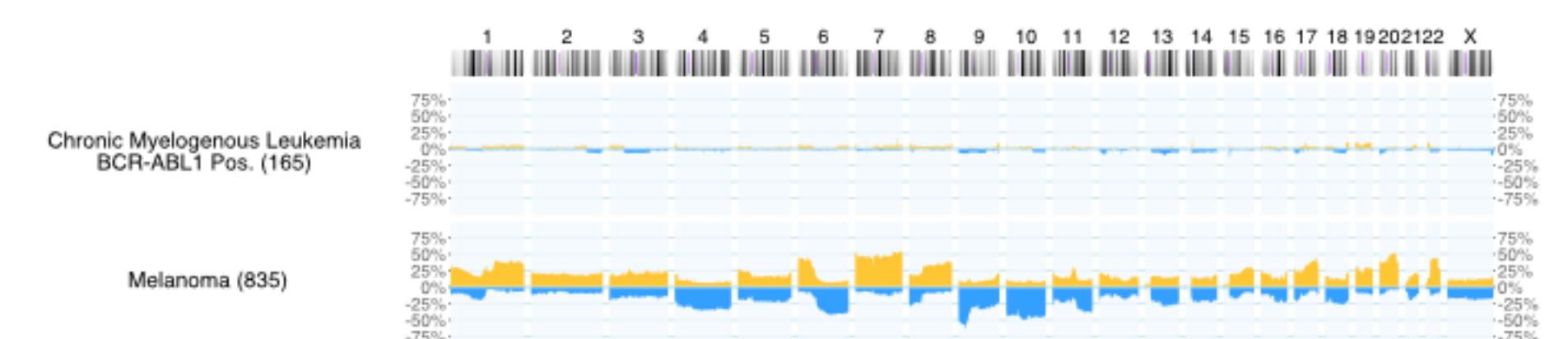
Data Use Cases

Genome CNV coverage in Cancer Classes

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



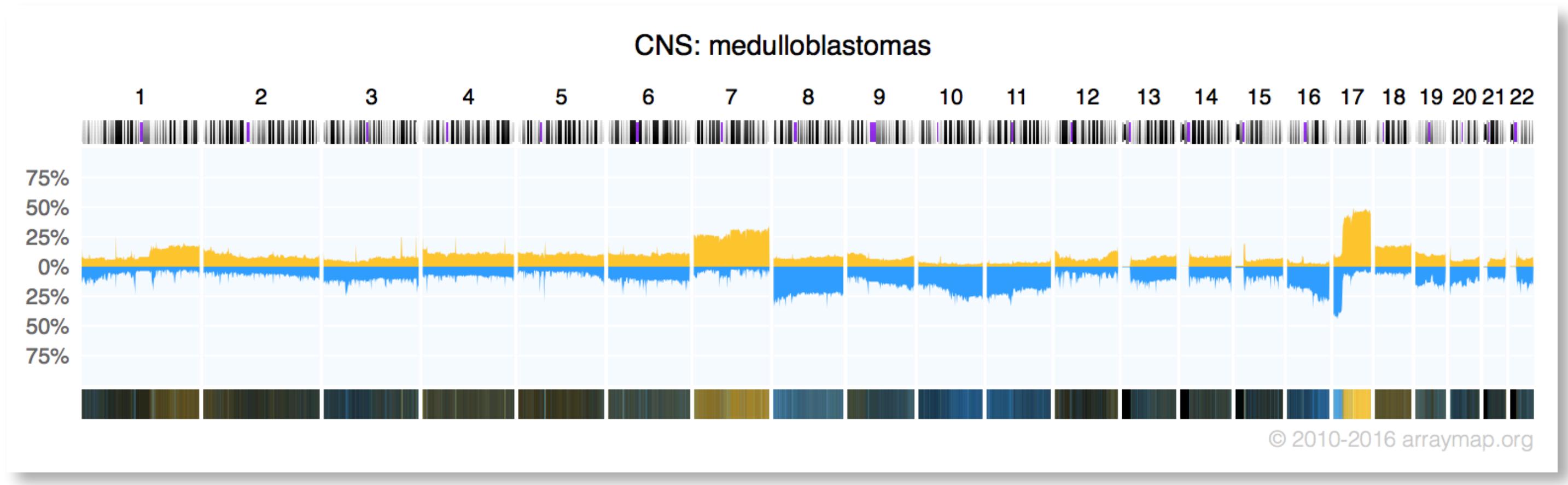
Lowest / Highest CNV fractions =>



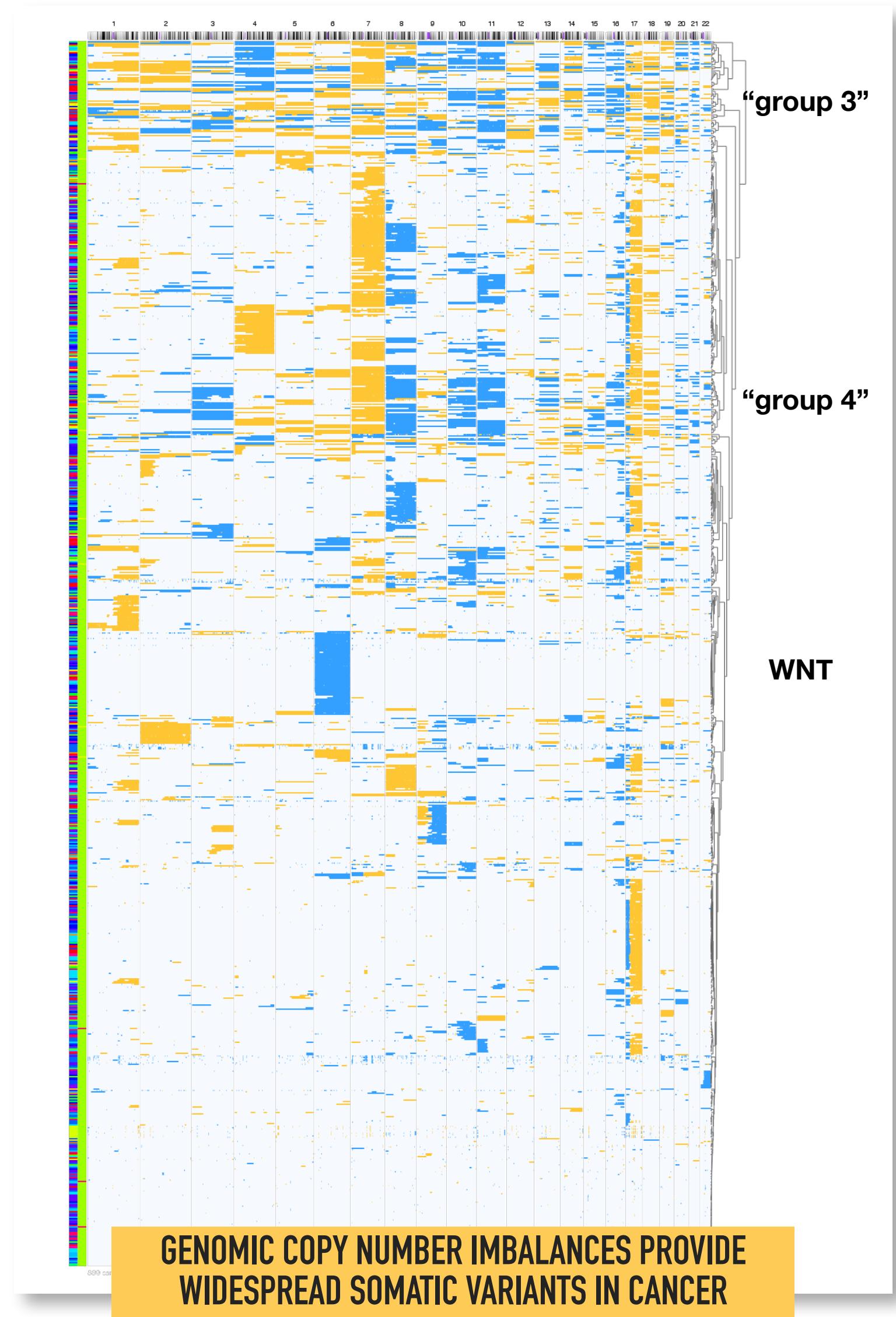
Somatic CNVs In Cancer

Recurrent mutation patterns

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



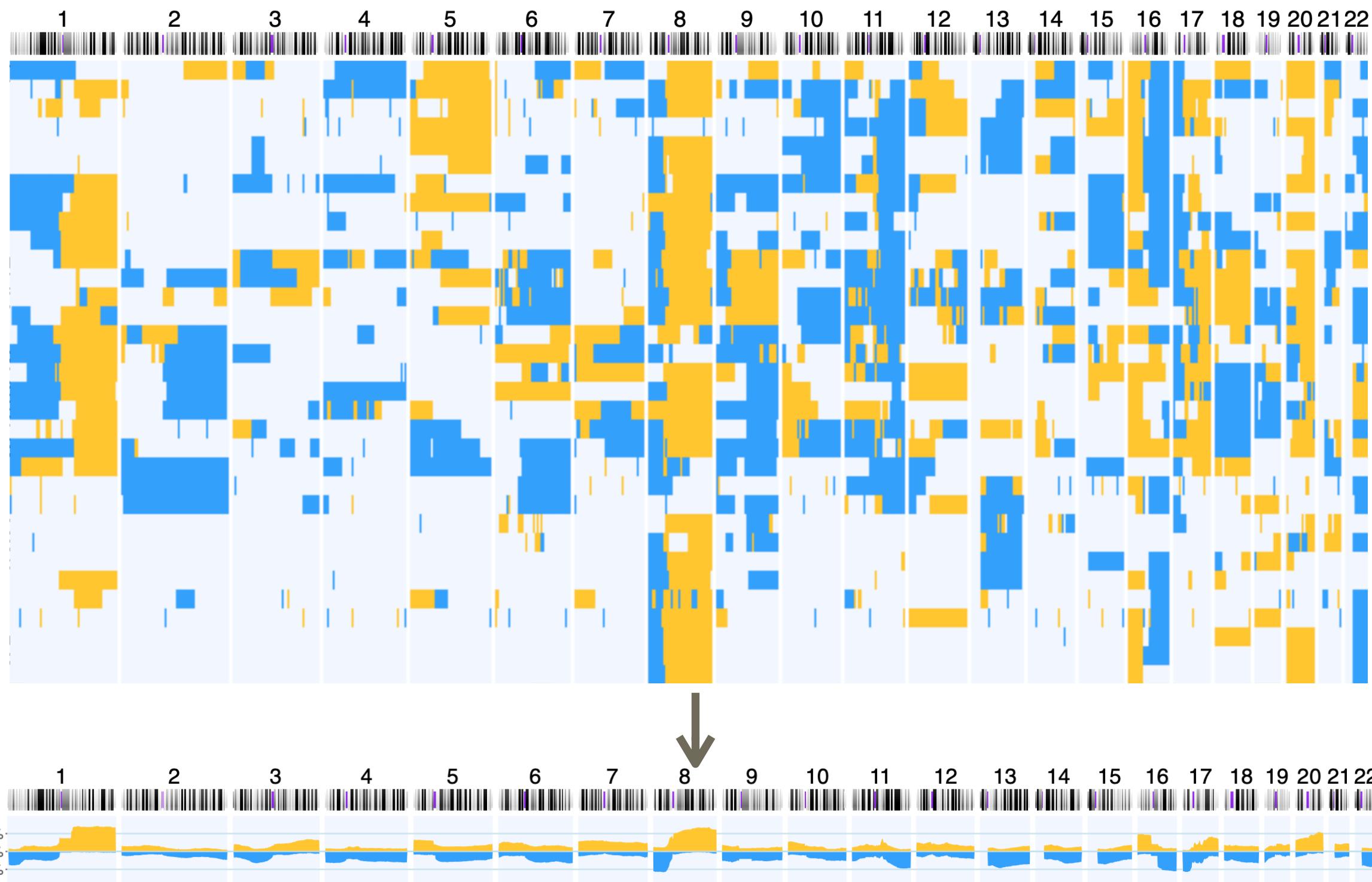
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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



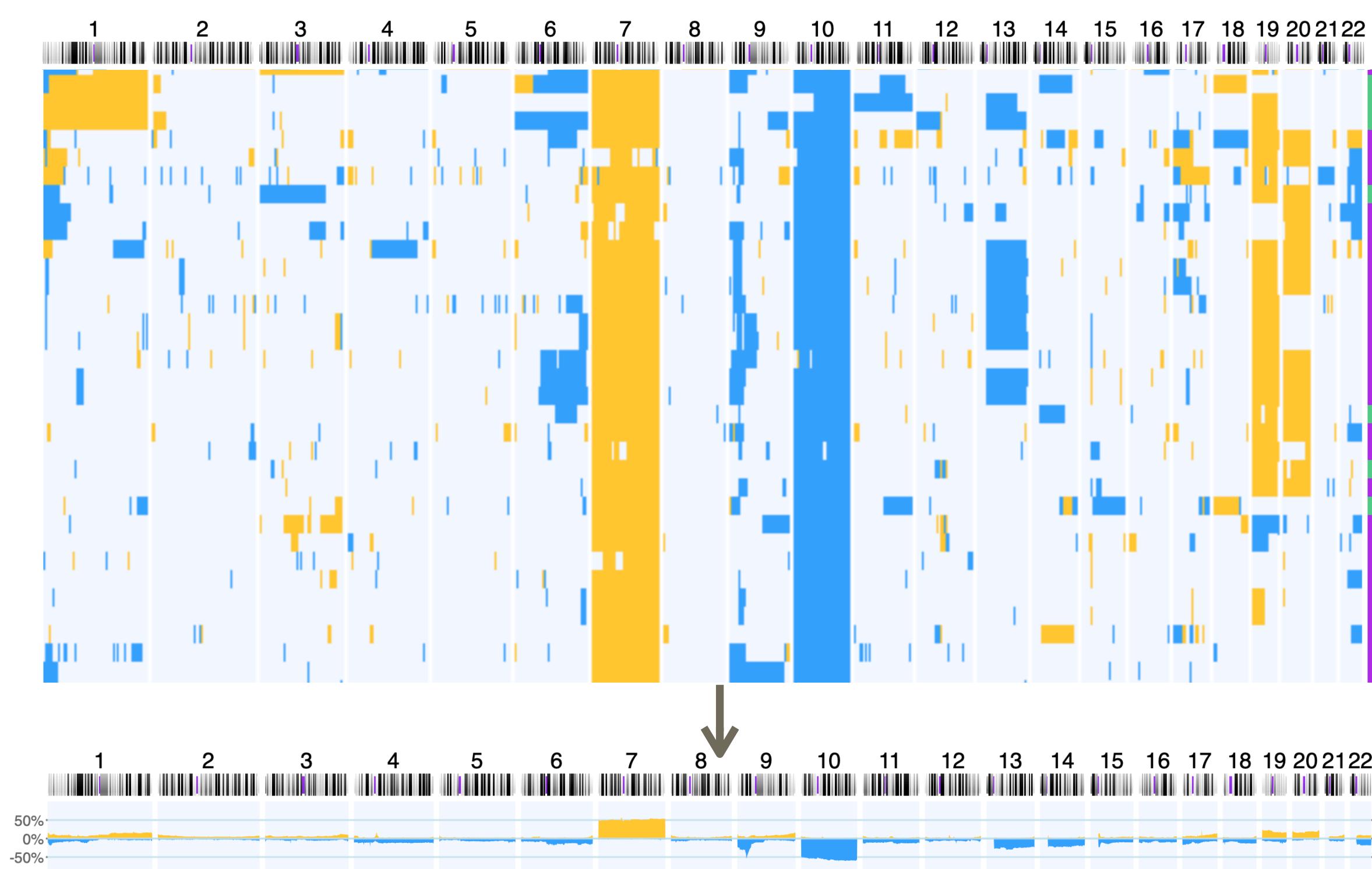
Drivers? Passengers? Markers?

Disentangling CNA Patterns

Ductal Breast Carcinoma



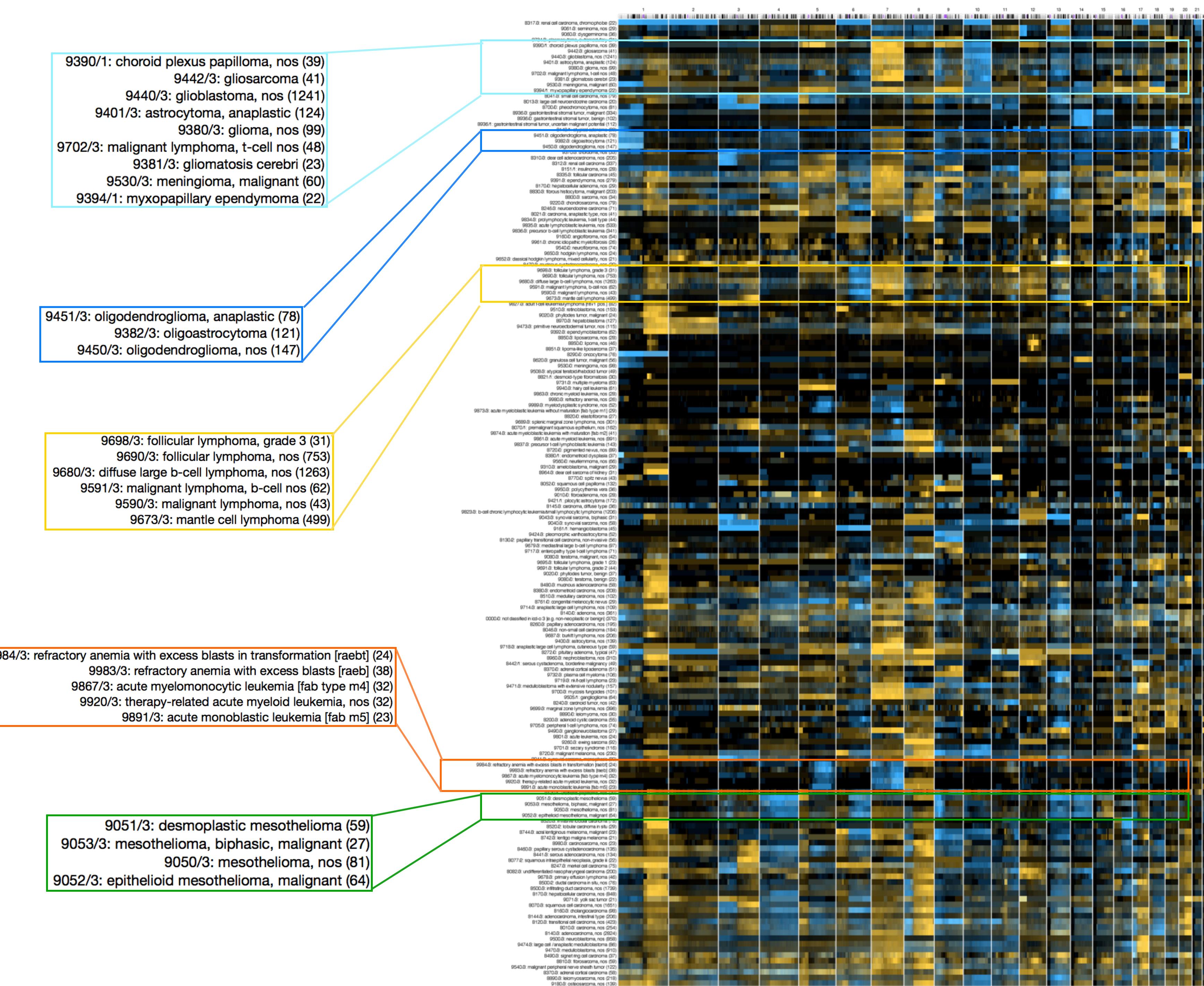
Glioblastoma



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*}

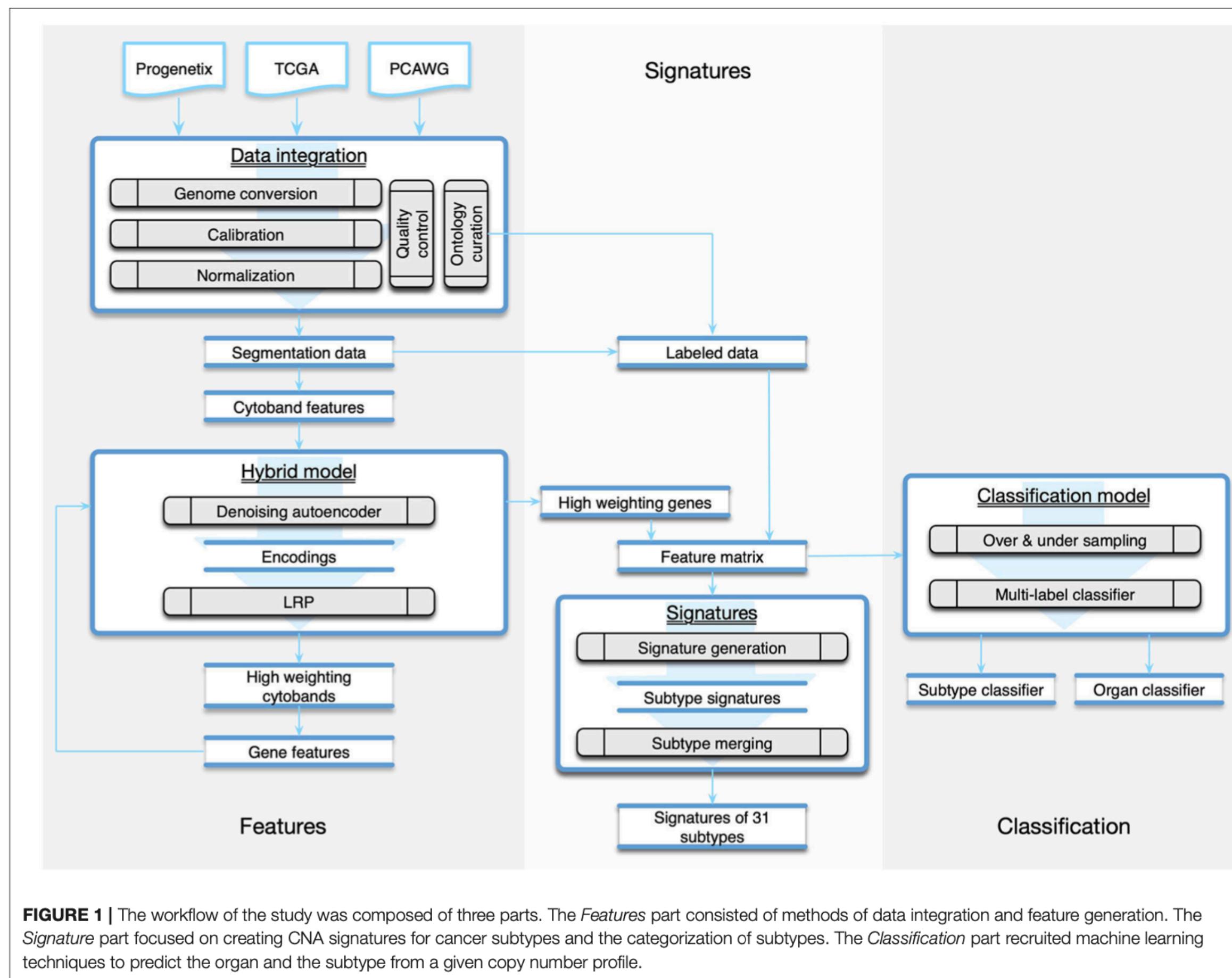


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine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

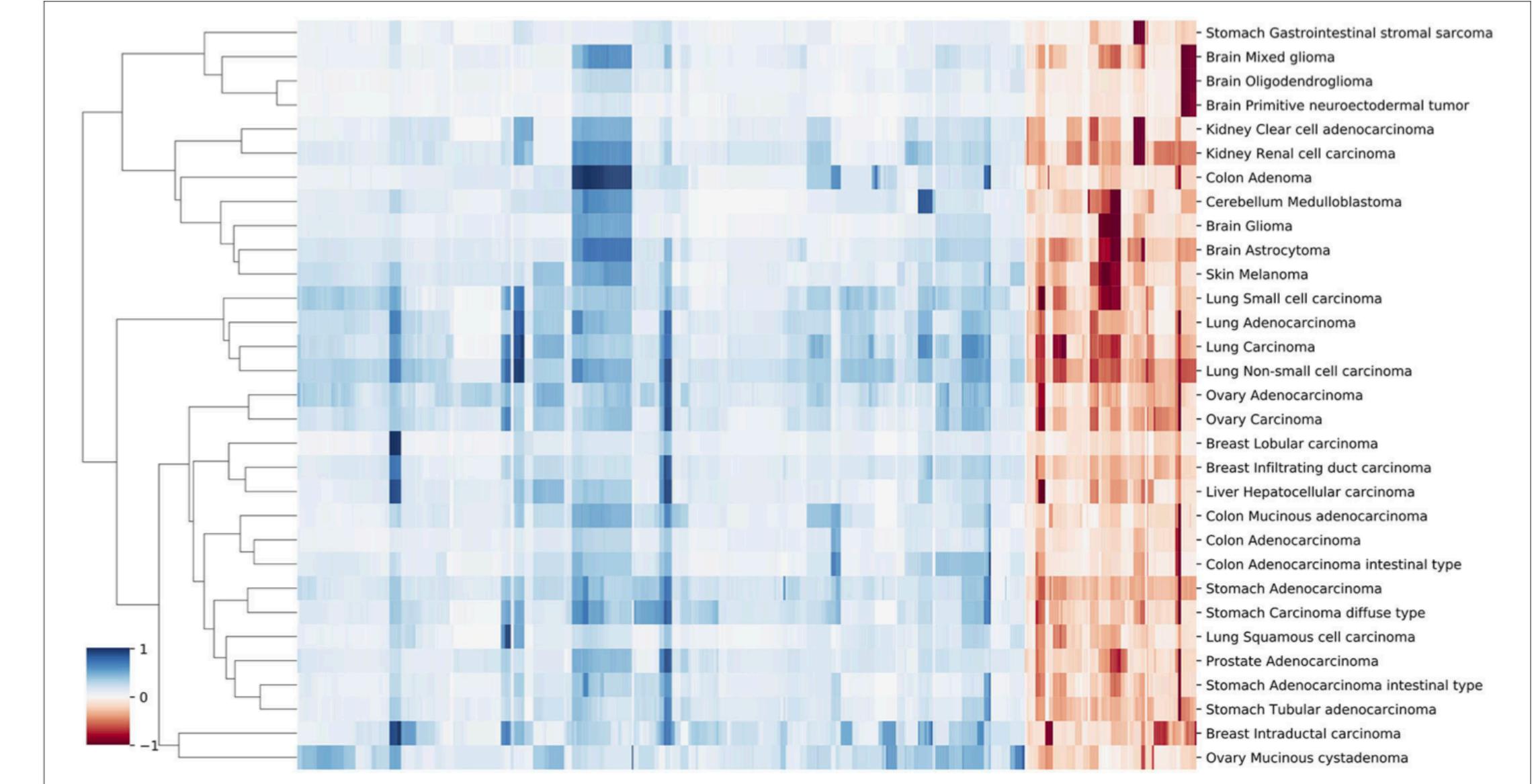


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

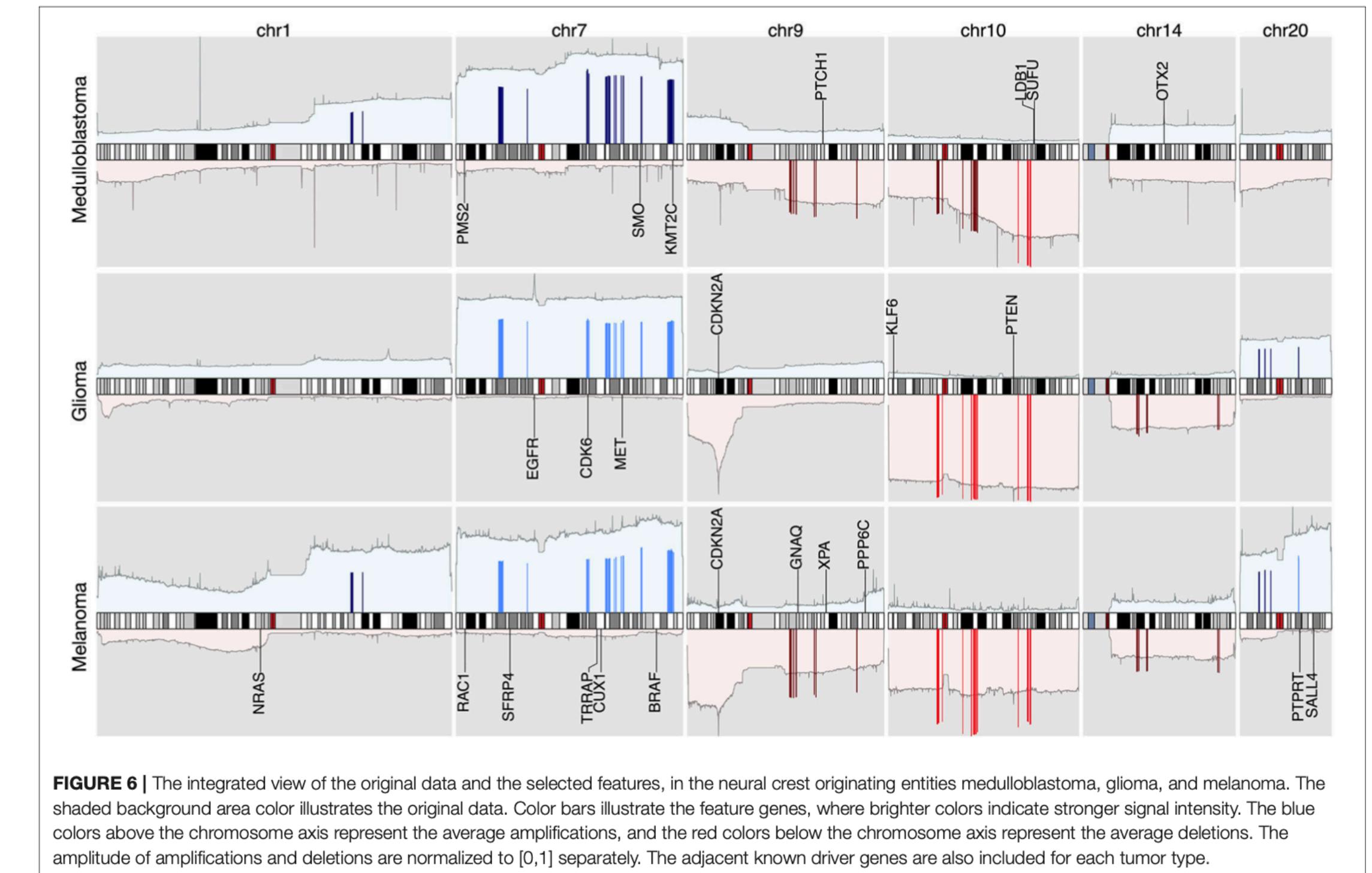
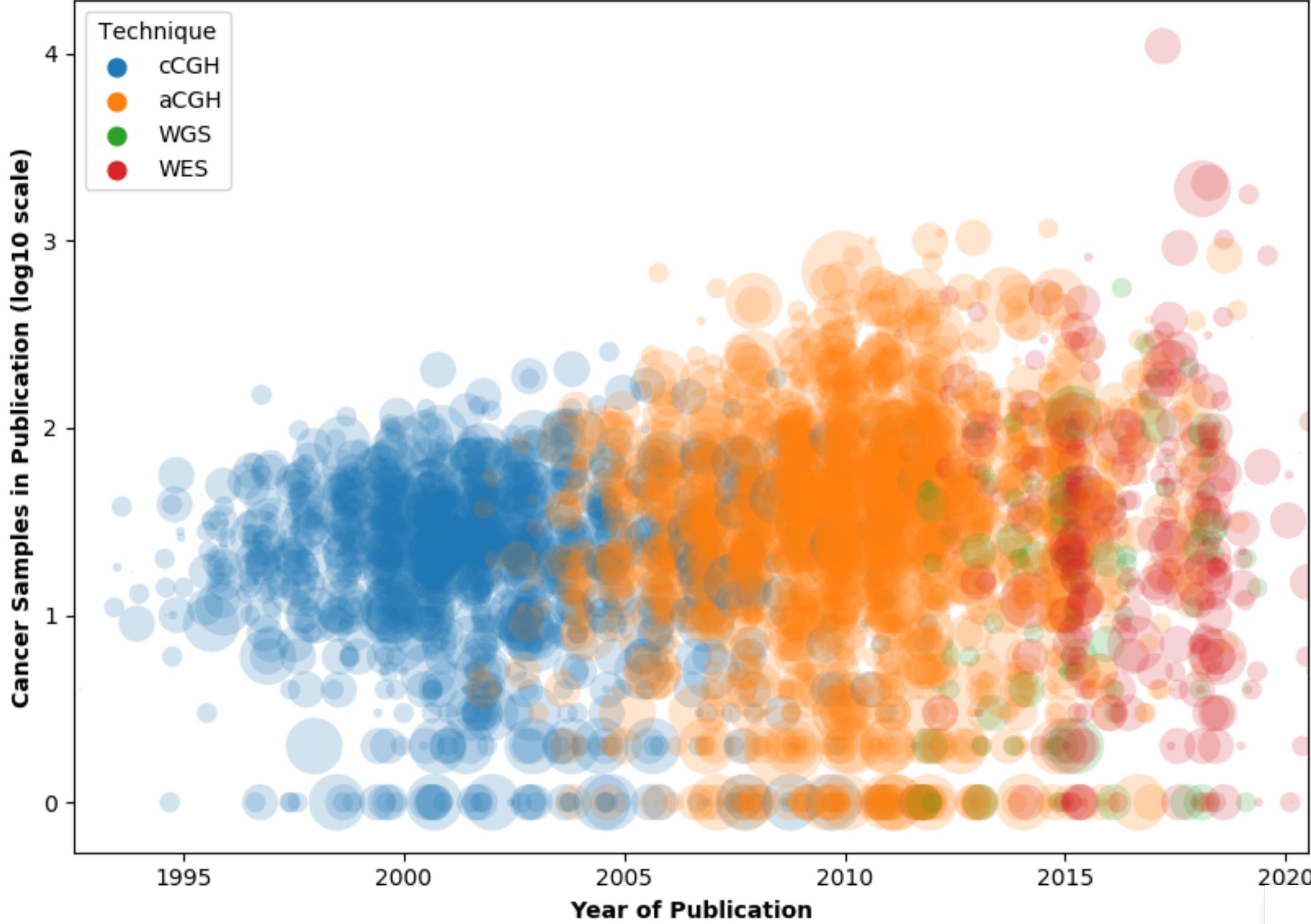


FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger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

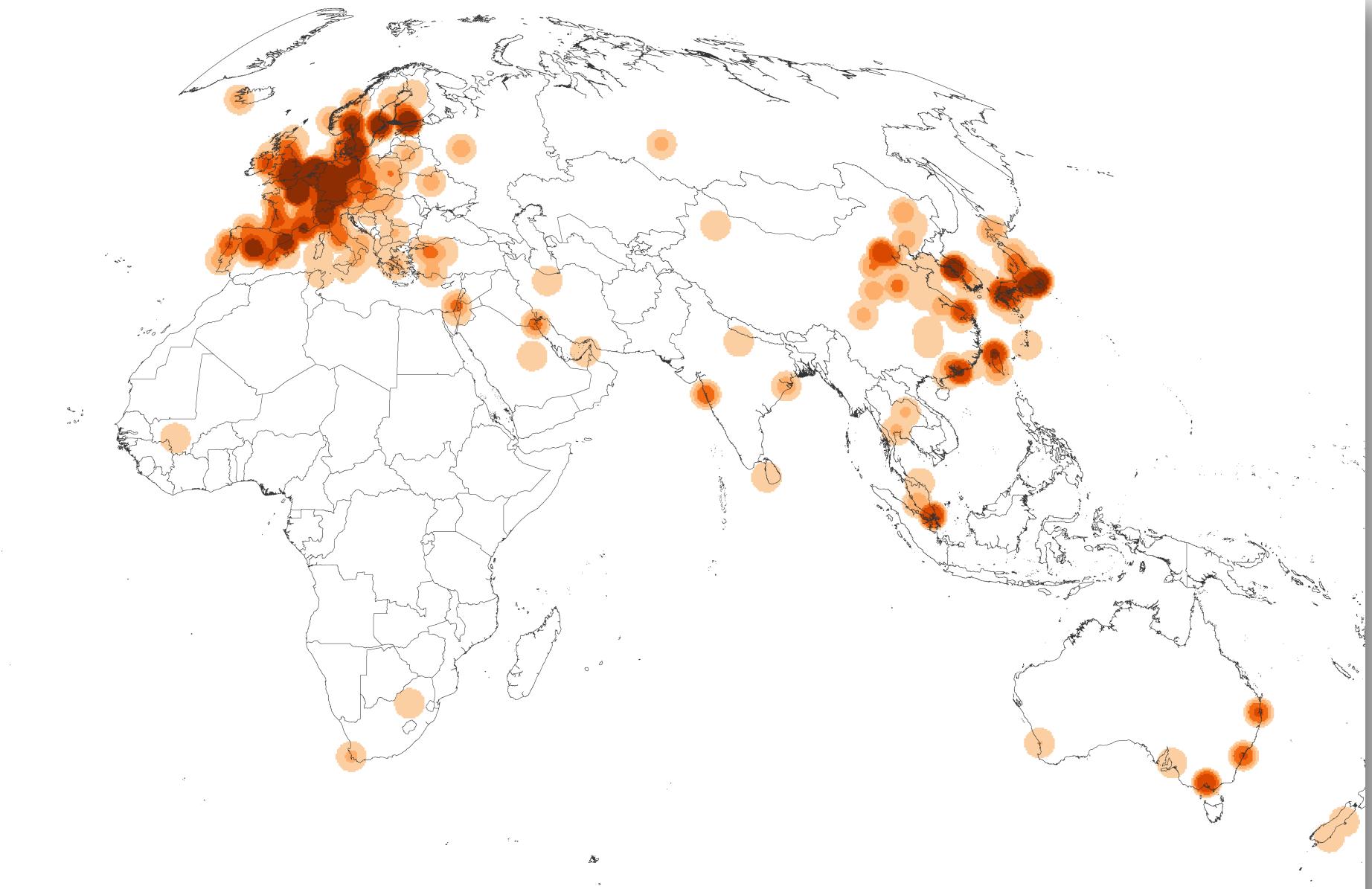
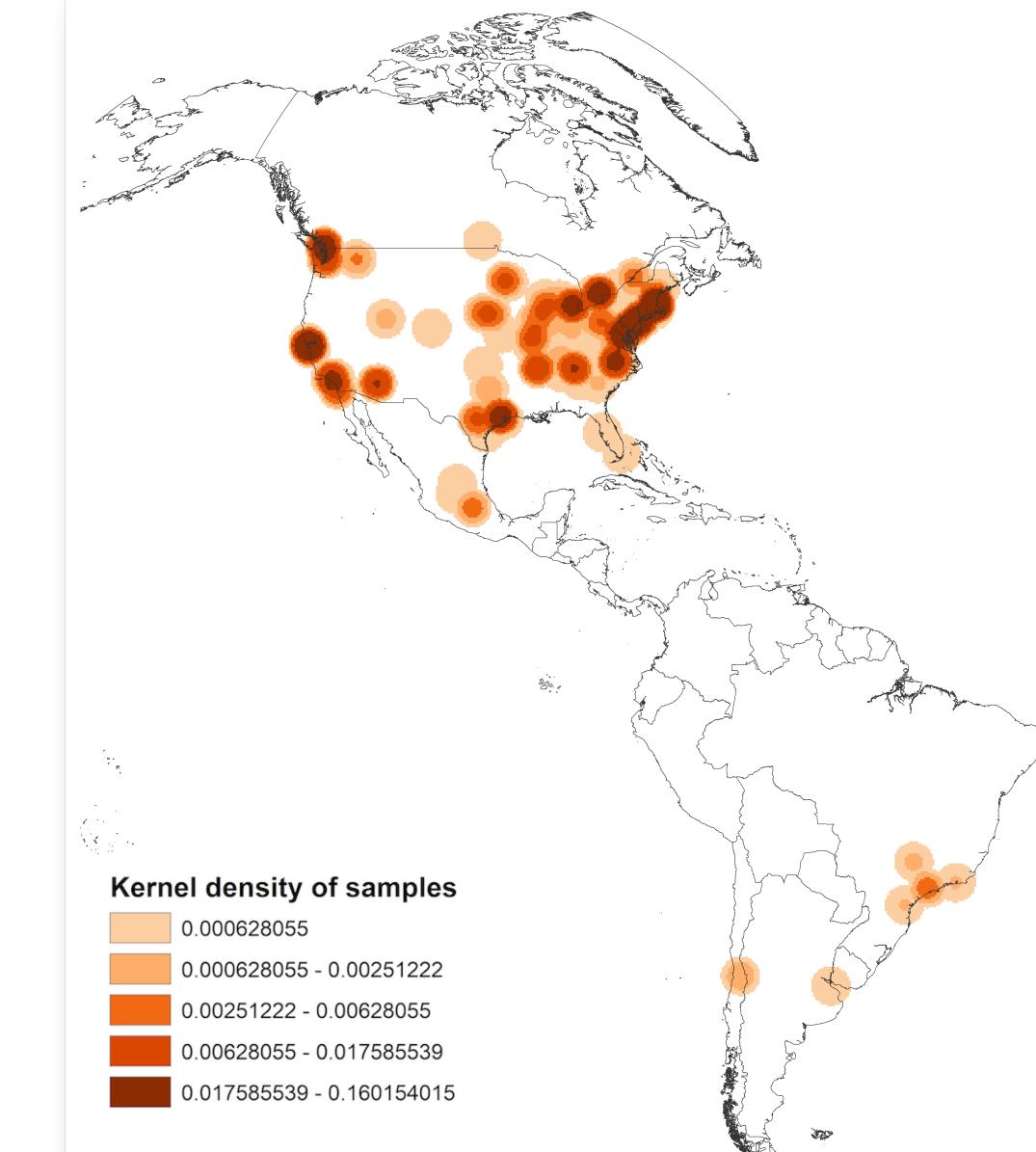
Please [contact us](#) to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation [🔗](#).

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



Population stratification in cancer samples based on SNP array data

- Despite extensive somatic mutations of cancer profiling data, consistency between germline and cancer samples reached 97% and 92% for 5 and 26 populations
- Comparison of our benchmarked results with self-reported meta-data estimated a matching rate between 88 % to 92%.
- Ethnicity labels indicated in meta-data are vague compared to the standardized output from our tool

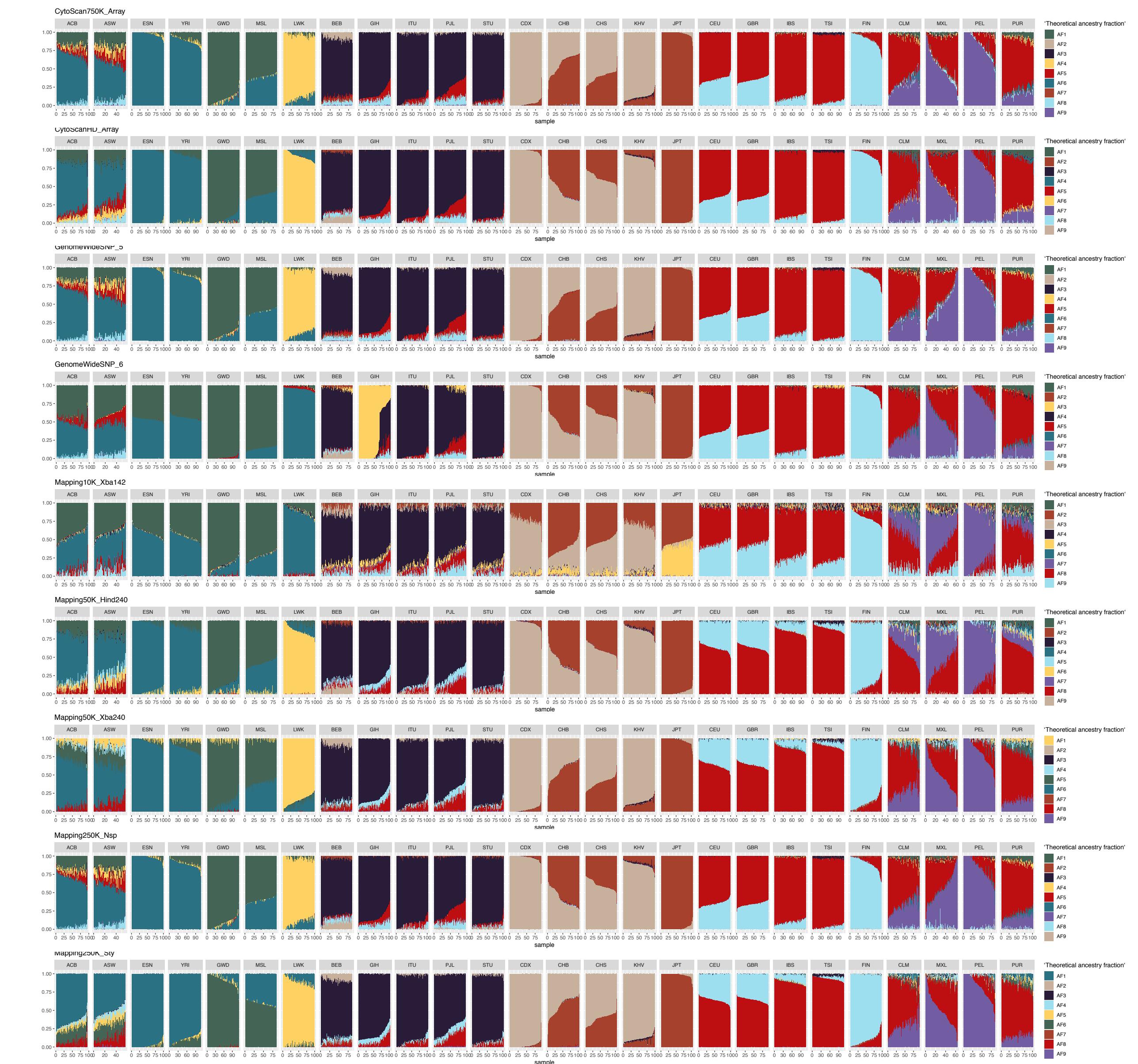


Figure S1 The fraction or contribution of theoretical ancestors ($k=9$) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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

Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard



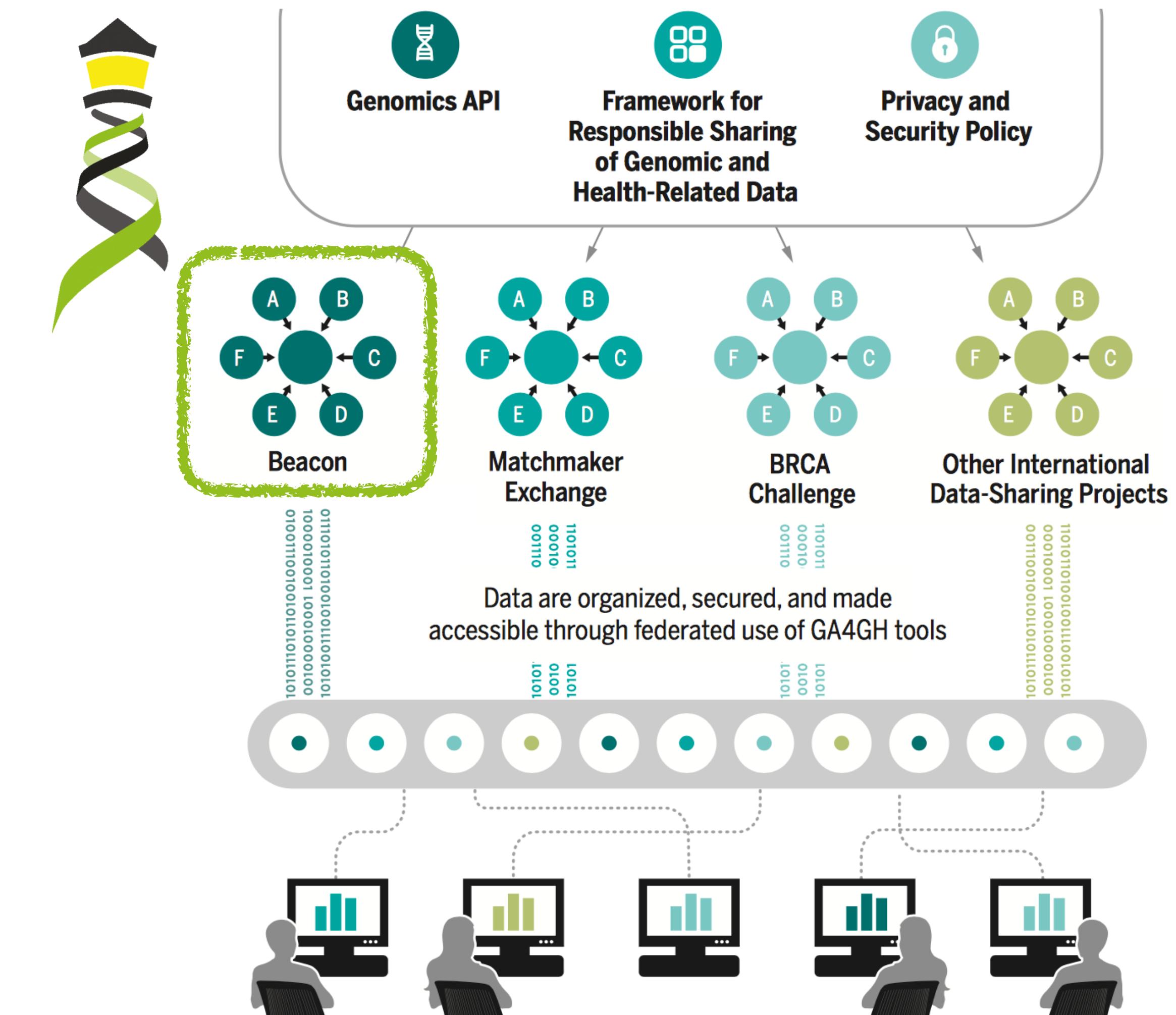


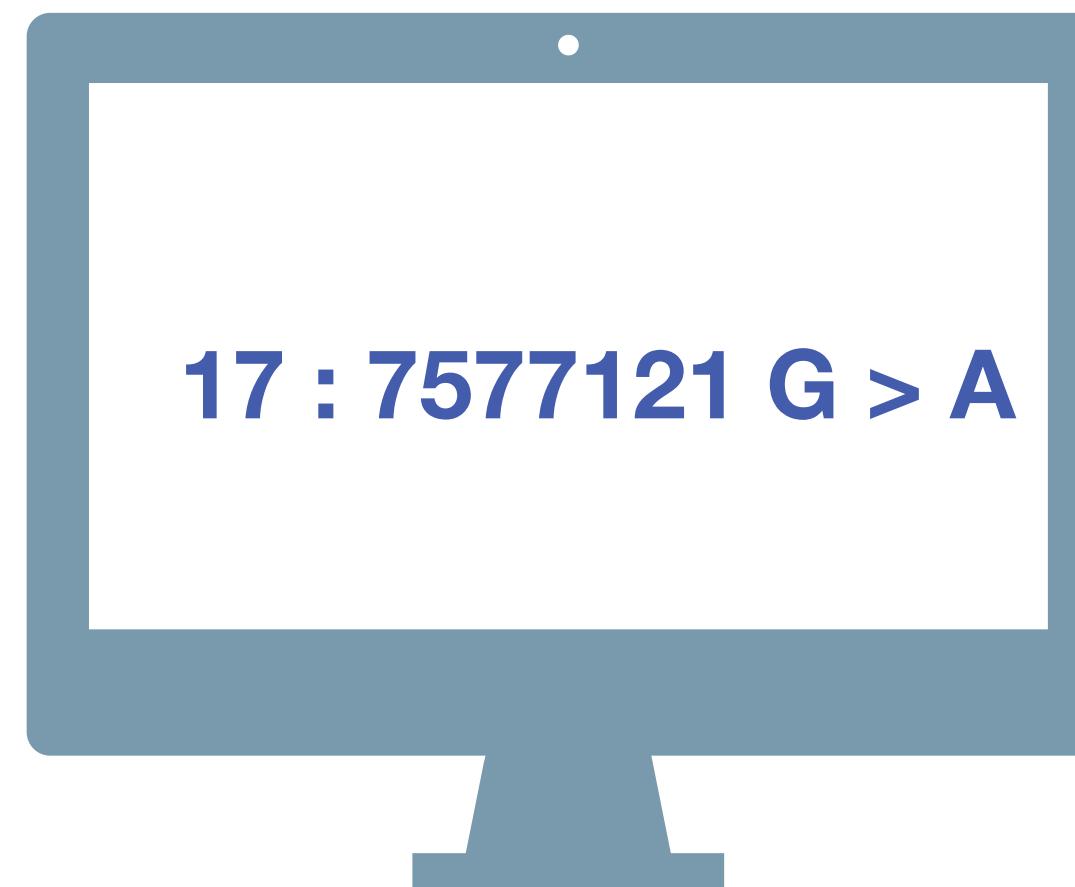
GENOMICS

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.





Beacon

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

YES | NO | \0

Introduction

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

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

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

Utility

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

A number of more useful first versions have been suggested.

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

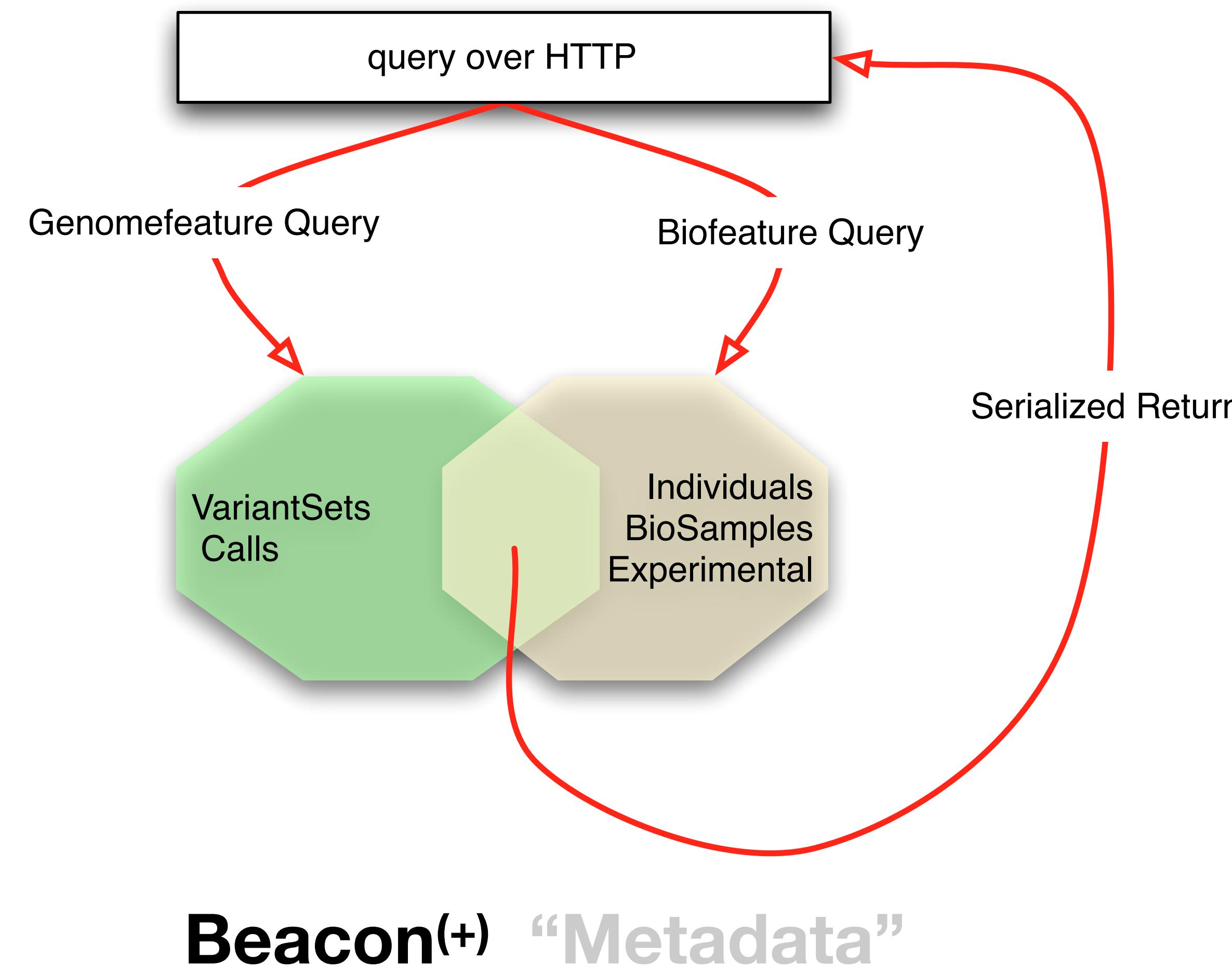


"I would personally recommend all those be held for version 2, when the beacon becomes a service."
Jim Ostell, 2014

Implementation

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

Minimal GA4GH query API structure



Beacon+ by Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

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



Search Samples

CNV Request Allele Request Range Query All Fields

CNV Example

This query type is for copy number queries ("variantCNVrequest"), e.g. using fuzzy ranges for start and end positions to capture a set of similar variants.

Dataset: progenetix

Cohorts: Select...

Genome Assembly: GRCh38 / hg38

Gene Symbol: Select...

Reference name: 9 **(Structural) Variant Type**: DEL

Start or Position: 19000001-21975098 **End (Range or Structural Var.)**: 21967753-24000000

Minimum Variant Length: **Maximal Variant Length**:

Cancer Classification(s): Select...

Filters:

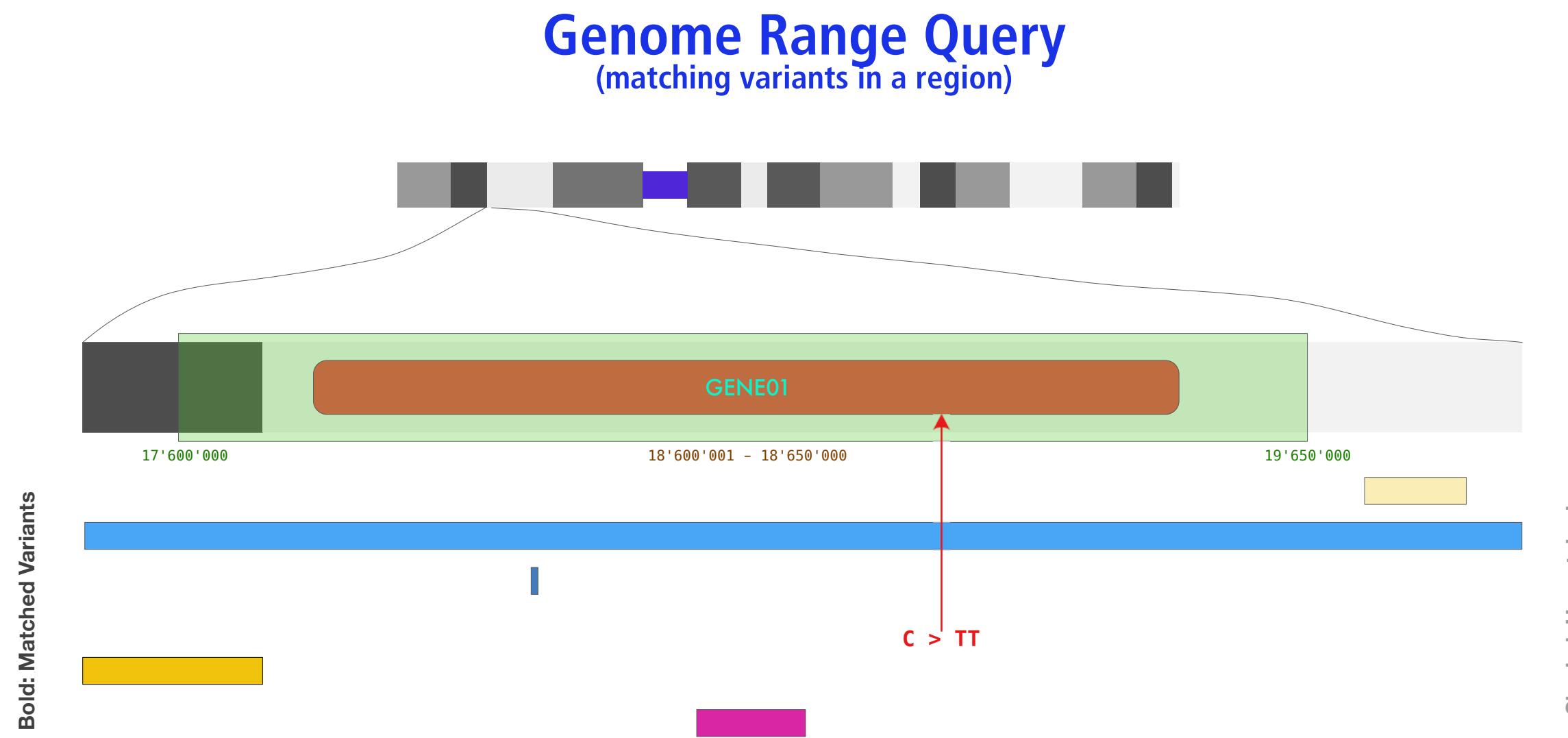
City: Select...

Query Database

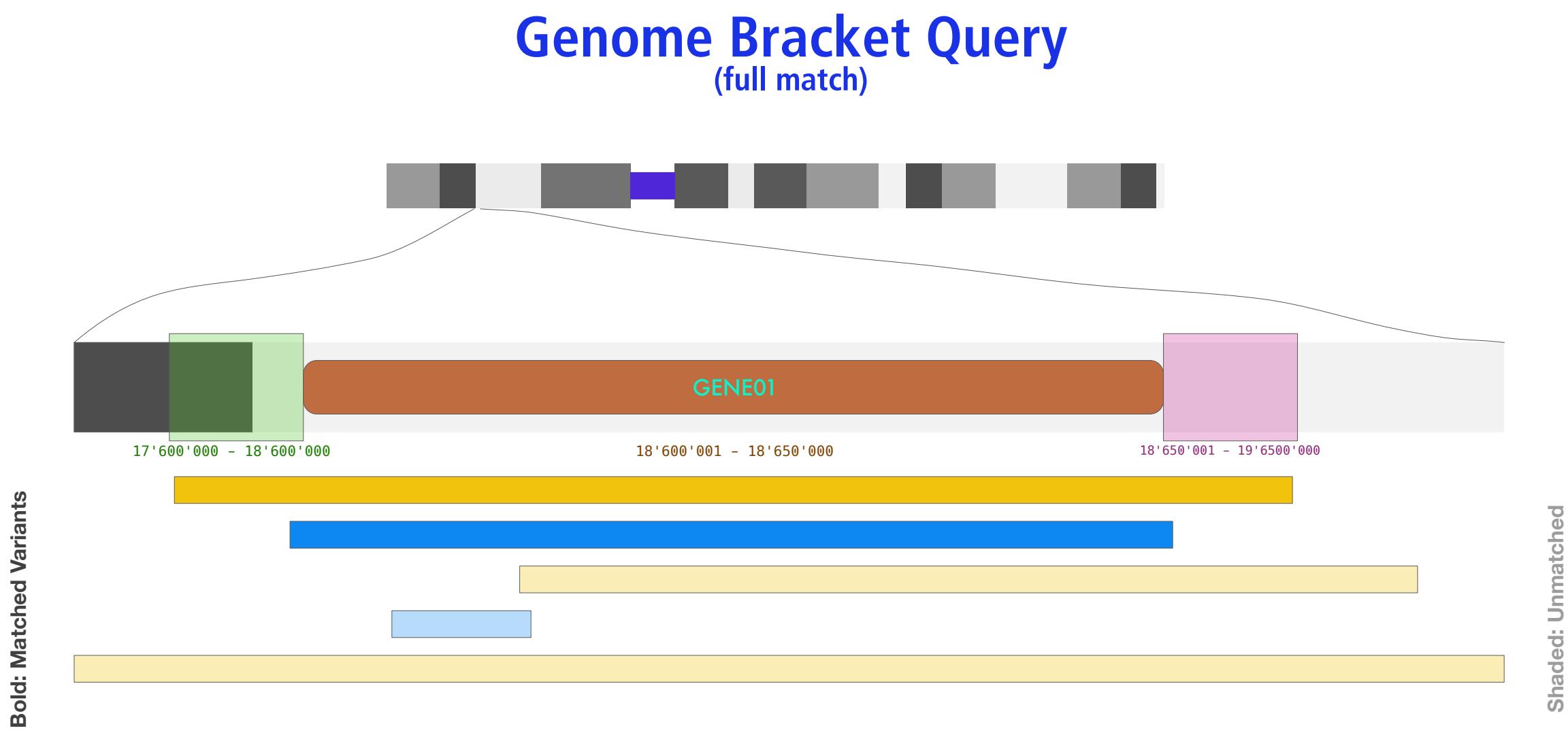
Beacon v2: Extended Variant Queries



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)

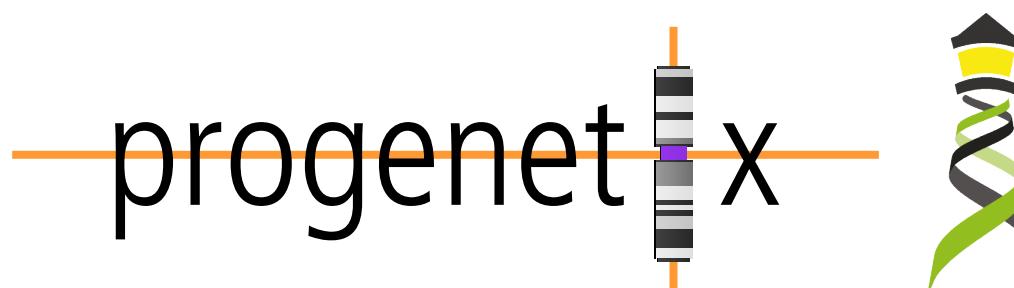


- Genome Bracket Queries allow to search for structural variants with start and end positions falling into defined sequence ranges
- allows to query any contiguous genomic variant (and in principle also can step in for range queries)
- typical use case is e.g. the query for variants such as duplications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

Beacon v2 Filters

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

- Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications
 - implicit *OR* with otherwise assumed *AND*
 - implementation of hierarchical annotations overcomes some limitations of "fuzzy" disease annotations



Beacon+ specific: Multiple term selection with OR logic

<input checked="" type="checkbox"/>	> NCIT:C4914: Skin Carcinoma	213
<input type="checkbox"/>	> NCIT:C4475: Dermal Neoplasm	109
<input checked="" type="checkbox"/>	> NCIT:C45240: Cutaneous Hematopoietic and Lymphoid Cell Neoplasm	310

Filters: NCIT:C4914, NCIT:C4819, NCIT:C9231, NCIT:C2921, NCIT:C45240, NCIT:C6858, NCIT:C3467, NCIT:C45340, NCIT:C7195, NCIT:C3246, NCIT:C7217



progenetix

Variants: 0 *f*alleles: 0 Callsets Variants ↗ UCSC region ↗ Legacy Interface ↗ [Show JSON Response](#)

Results Biosamples

Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.154	0.056	0.21
PGX_AM_BS_MCC03	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.137	0.21	0.347
PGX_AM_BS_MCC04	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.158	0.056	0.214
PGX_AM_BS_MCC05	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.107	0.327	0.434

Beacon v2 Requests

POSTing Queries

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

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



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 UI with a search bar at the top containing the query: "Assembly: GRCh38 Chro: 9 Start: 21500001-21975098 End: 21967753-22500000 Type: EFO:0030067 Filters: NCIT:C3058". Below the search bar, a summary box displays: Matched Samples: 660, Retrieved Samples: 660, Variants: 279, Calls: 667. It also includes links for UCSC region, Variants in UCSC, Dataset Response (JSON), and Visualization options.

The main content area shows a chart titled "progenetix" with a y-axis from -75% to 1%. Below the chart is a table with columns: Matched Subset Codes, Subset Samples, Matched Samples, and Subset Match Frequencies. The table lists several rows, with the first row highlighted in yellow:

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
UBERON:0002021	14	1	0.025
pgx:icdot-C71.0	1714	14	0.008
pgx:icdot-C71.0	1714	14	0.008
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 (UCSC)" and "Download Sample Variants (JSON)".

A network request table on the right side of the screen shows the following requests:

Name	Do...	T Transf...	T...	10.00s	20.00s	30.00s
biosamples	pro...	fr	5.14 KB	2...		
biosamples	lock	fr	52.60...	1...		
genomicVariations	lock	fr	25.99...	1...		
genomicVariations	lock	fr	3.98 KB	8...		
samplePlots.cgi	lock	fr	26.13 ...	2...		
collations	pro...	fr	199.4...	1...		

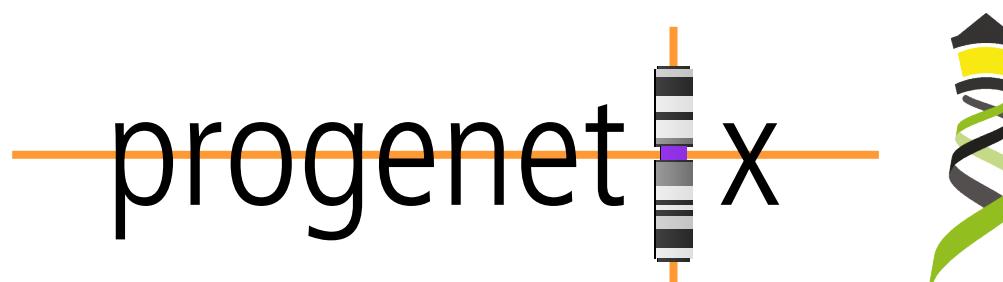
Specific requests highlighted with colored boxes are:

- A yellow box highlights the 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 box highlights the request: `/beacon/biosamples/?skip=0&limit=1000&accessid=fbffda57-0f41-4d6a-99fc-41d4cfdea9f6&requestedSchema=biosample`
- A pink box highlights the request: `/beacon/genomicVariations/?accessid=e2dadd91-9326-46de-97e4-6b88413b6bfe&requestedSchema=genomicVariant`
- A light red box highlights the request: `/cgi-bin/PGX/cgi/samplePlots.cgi?accessid=fbffda57-0f41-4d6a-99fc-41d4cfdea9f6&method=cnvhistogram&-size_plotimage_w_px=645`

Beacon v2 Paths

Progenetix utilizes Beacon v2 REST paths

- 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 (see next)
- 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/`

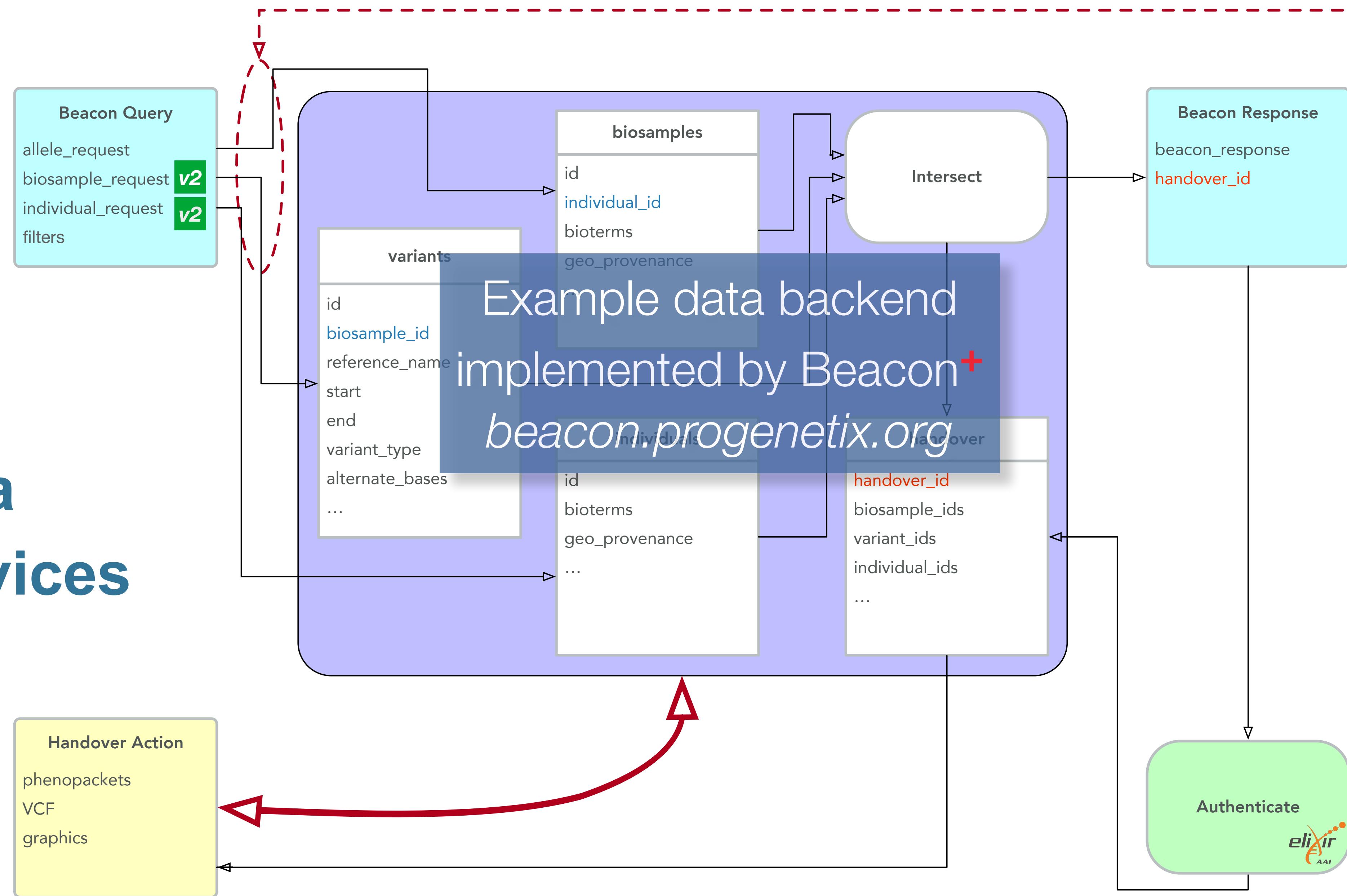
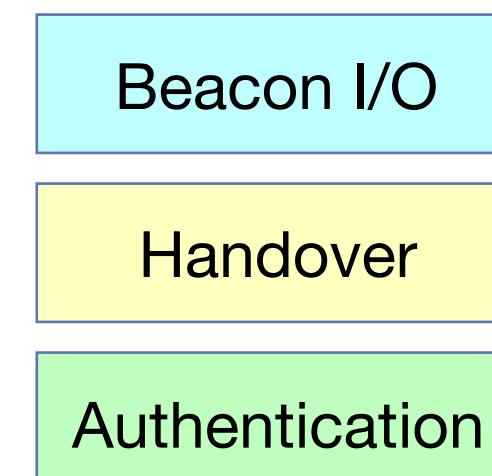
`/variants/{id}/biosamples/` & `variants_in_sample/{id}/biosamples/`

- `/variants/5f5a35586b8c1d6d377b77f6/biosamples/`

- `/variants_in_sample/5f5a35586b8c1d6d377b77f6/biosamples/`

Beacon & Handover

Beacons v1.1
supports data
delivery services



ga4gh-beacon / beacon-framework-v2 Public

Code Issues 18 Pull requests 2 Discussions Actions Wiki Security Insights Settings

main 7 branches 0 tags Go to file Add file Code About

jrambla Merge pull request #51 from ga4gh-beacon/configuration-typos-fixes ...

common	de-lining \n
configuration	speling in configuration -> filteringTermsSchema
requests	de-lining \n
responses	de-lining \n
.gitignore	Initial commit
LICENSE	Initial commit
README.md	Adding naming conventions to readme
endpoints.json	de-lining \n

README.md

beacon-framework-v2

Beacon Framework version 2

Introduction

The GA4GH Beacon specification is composed by two parts:

- the Beacon Framework (in *this* repo)
- the Beacon Model (in the [Models repo](#))

The Beacon Framework is the part that describes the overall structure of the API

progenetix / bycon Public

Code Issues Pull requests 1 Actions Projects Wiki Security Insights Settings

master 3 branches 0 tags Go to file Add file Code About

mbaudis Update README.md 5064e89 11 seconds ago 519 commits

beaconServer	datatables, genesRefresher	6 days ago
byconeer	datatables, genesRefresher	6 days ago
config	datatables, genesRefresher	6 days ago
lib	intervalFrequencies service & some library shuffling	5 months ago
schemas	datatables, genesRefresher	6 days ago
services	genespan method for gene request size reduction	2 days ago
remnants	biocharacteristics removal; shuffling of beaconsv2 references...	21 days ago
.gitignore	biocharacteristics removal; shuffling of beaconsv2 references...	21 days ago
LICENSE	Create LICENSE	12 months ago
README.md	Update README.md	11 seconds ago
__init__.py	intervalFrequencies service & some library shuffling	5 months ago
requirements.txt	add non-interactive mode	16 months ago

README.md

License CC0 1.0

Bycon - a Python-based environment for the Beacon v2 genomics API

The `bycon` project - at least at its current stage - is a mix of Progenetix (i.e. GA4GH object model derived, MongoDB implemented) - data management, and the implementation of middleware & server for the Beacon API.

More information about the current status of the package can be found in the inline documentation which is also presented in an accessible format on the [Progenetix website](#).

About

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

Readme

CC0-1.0 License

Releases

No releases published [Create a new release](#)

Packages

No packages published [Publish your first package](#)

Contributors 4

mbaudis Michael Baudis
sofiapfund Sofia
qingyao
KyleGao Bo Gao

Languages

Python 99.9% Shell 0.1%

Onboarding

Demonstrating Compliance

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

 Centre Nacional Analisis Genomica (CNAG-CRG)

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

Beacon @ RD-Connect

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	██████████

 University of Leicester

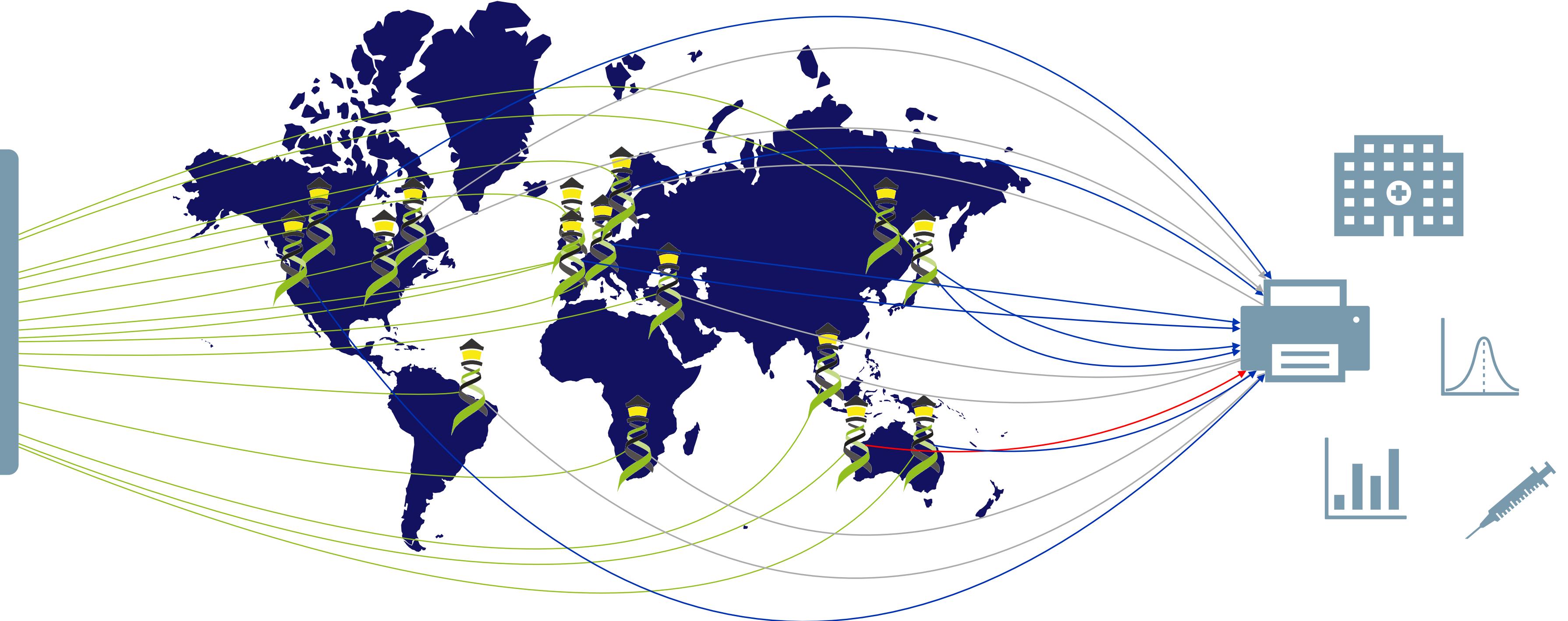
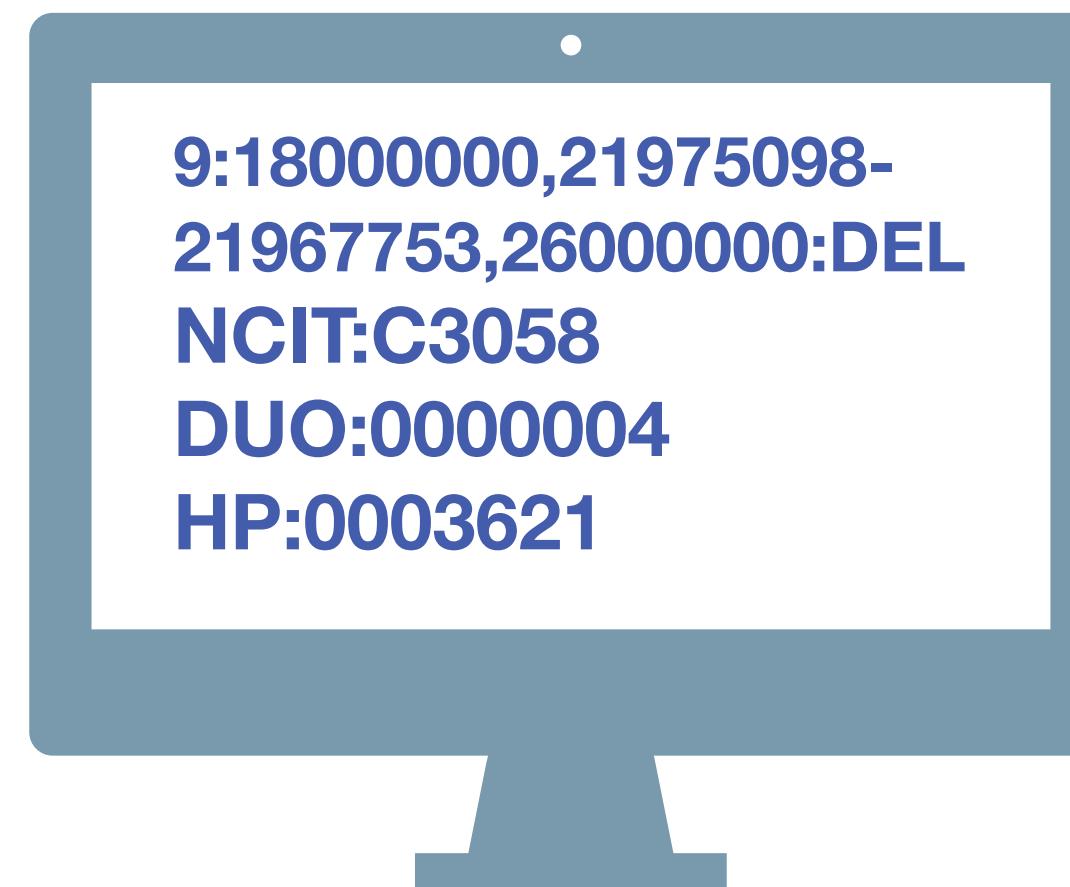
[Beacon UI](#) [Beacon API](#) [Contact us](#)

Cafe Variome Beacon v2

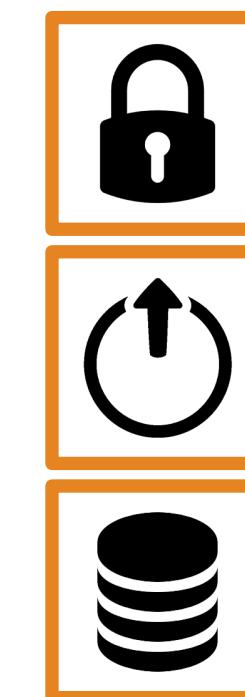
This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

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

✓ Matches the Spec ✗ Not Match the Spec ● Not implemented

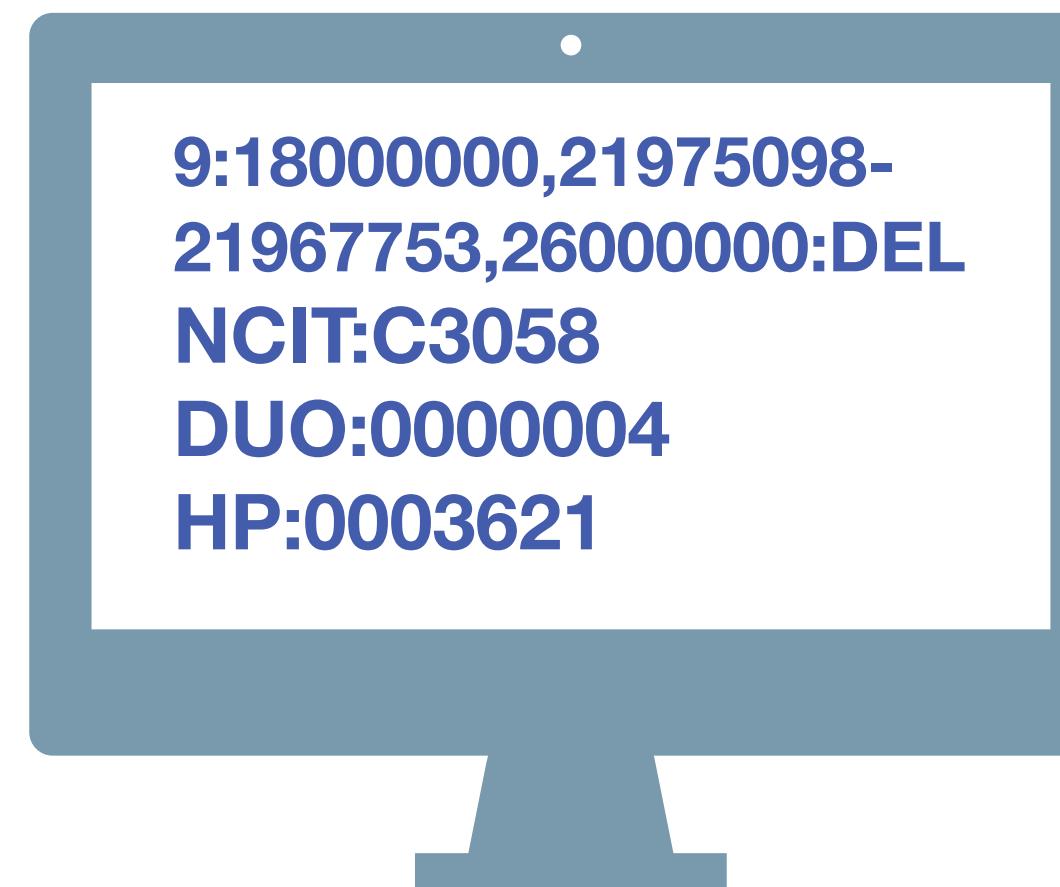


Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?



Beacon v2 API

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



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



Beacon v2 API

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

Components of an Online Bioinformatics Resource

Going Full Stack?

Components of an Online Bioinformatics Resource

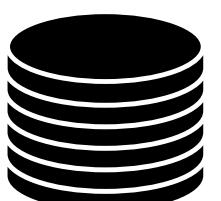
A Stack to work with/through

- 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)?

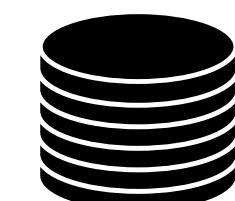
Progenetix Stack



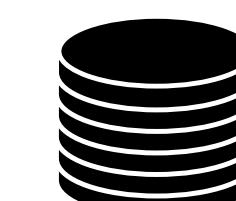
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the *bycon* package
 - schemas, query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



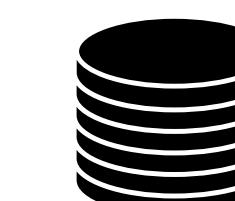
variants



analyses



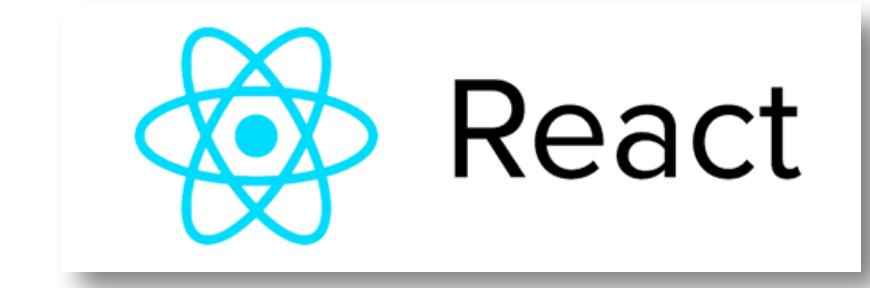
biosamples



individuals

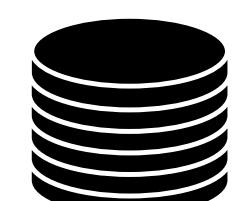


Entity collections

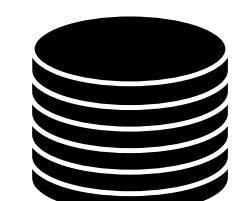


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and provides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

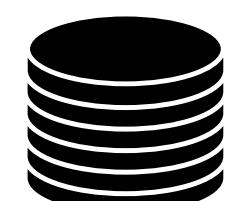
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



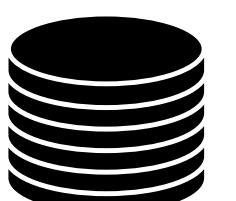
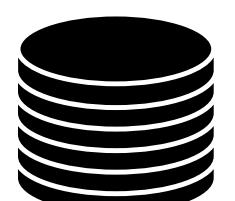
collations



geolocs



genespans publications



qBuffer

Utility collections

Last but NOT Least...

Documentation is, actually, rather important

Documentation Strategies

(Not so) Best Practices

- 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!

mbaudis@netscape.net

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

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. Cleary²

¹Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and

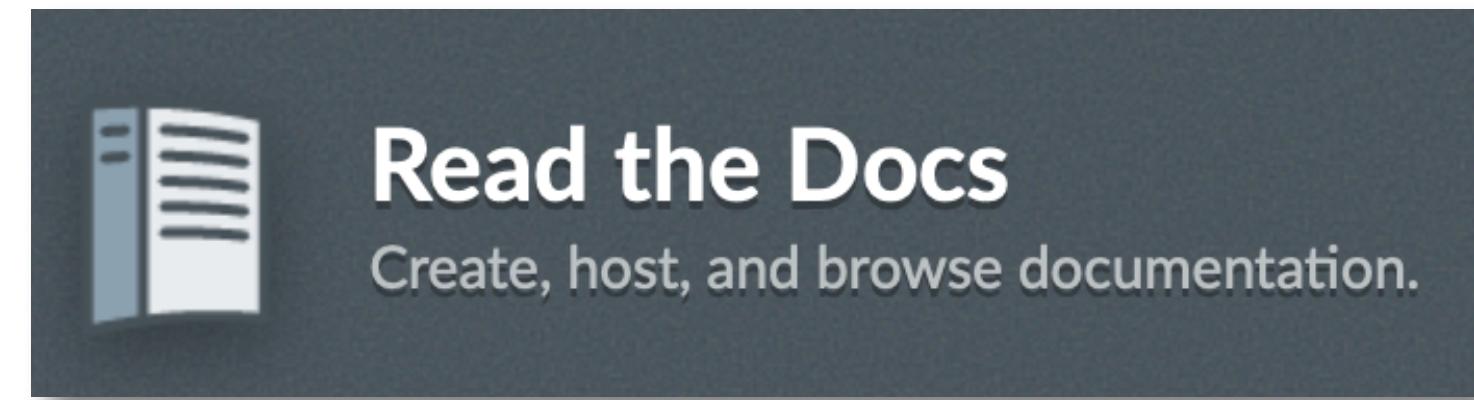
²Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA

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

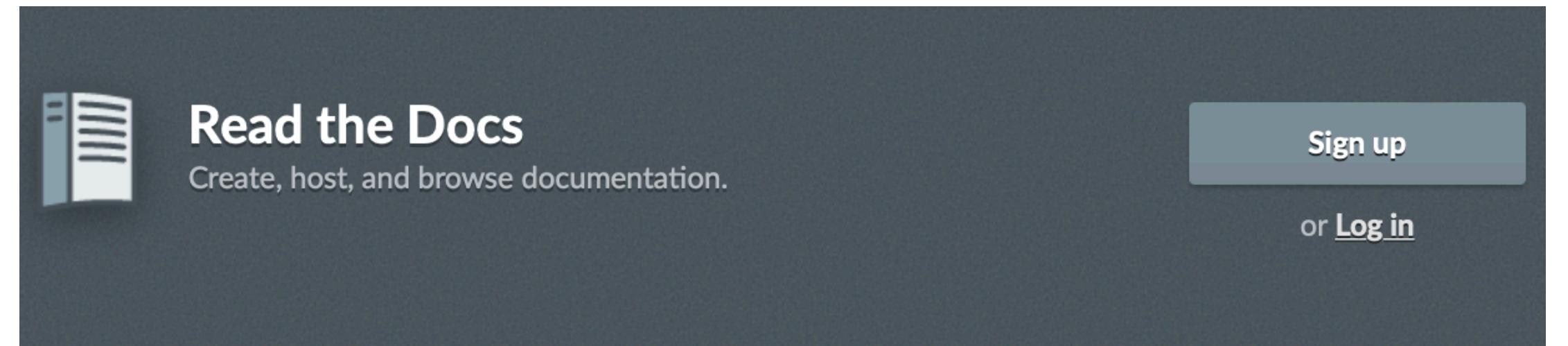
```
f_d = f_d_s[c_t]
r = {}
for k in res_schema.keys():
    if k in f_d:
        r.update({k:f_d[k]})
```

Documentation Strategies Currently en Vogue

- Cloud-based documentation systems with online compilation
- written in simplified markup languages
 - Markdown (Yeah!)
 - Restructured Text (Meeh...)
- 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." A central content area displays a "Malala Fund" advertisement with the text "Join Malala's fight. Help break down the barriers that hold girls back." and "Ad by EthicalAds · Community Ad".

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.

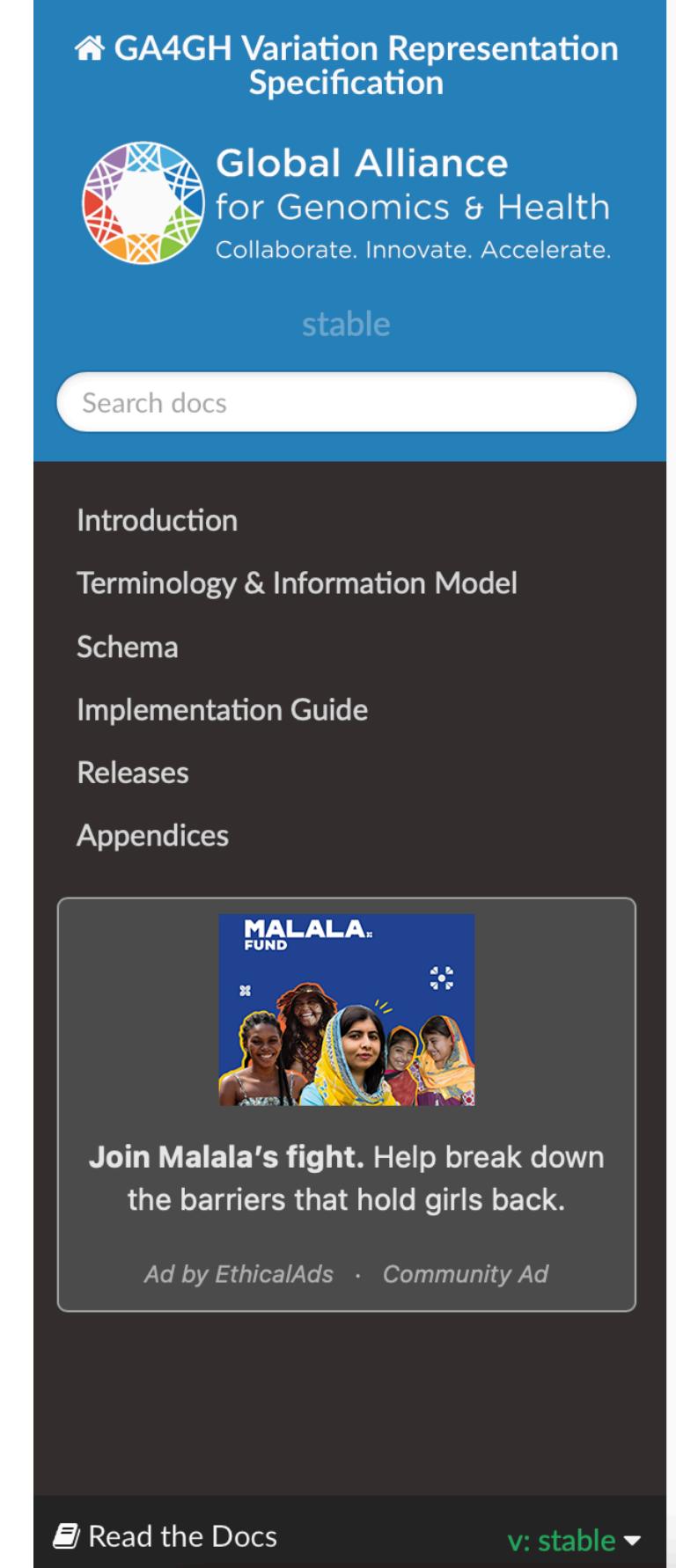
Downloadable formats

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 "stable" link and a search bar. The main content area contains links to "Introduction," "Terminology & Information Model," "Schema," "Implementation Guide," "Releases," and "Appendices." A sidebar on the left lists files such as "_static," "appendices," "images," "impl-guide," "releases," "conf.py," "defs," "index.rst," and "introduction.rst." An "Output" section on the right shows a list of recent commits from user "ahwagner" with timestamps and commit details.

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]	[cBioPortal: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 by Michael Baudis and progenetix.org

```

# 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
Progenetix.org

# Local Testing
the Progenetix oncogenes and their role in cancer development by Michael Baudis and progenetix.org

# Web Deployment (Github)
with a
live

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/

```

progenetix / progenetix-web Public

Actions

All workflows

mk-progenetix-docs

pages-build-deployment

Management

Caches

Deployments

Links

Private filters

Since some classifications and ontologies are now represented by CURIEs, the upcoming Beacon API will also support additional a set of structured non-CURIE identifiers.

Actions

New workflow

178 workflow runs

refseq ids in examples, aggregator UI start

Update VariantsDataTable.js

Update VariantsDataTable.js

Events Status Branch Actor

main 3 days ago ... 41s

main 11 days ago ... 39s

main 16 days ago ... 42s

mbaudis cleanup ✓

1 contributor

19 lines (19 sloc) | 491 Bytes

```

name: mk-progenetix-docs
on:
  push:
    branches:
      - main
  jobs:
    deploy:
      runs-on: ubuntu-latest
      steps:
        - uses: actions/checkout@v2
        - uses: actions/setup-python@v2
          with:
            python-version: 3.x
        - run: pip install mkdocs-material
        - run: pip install mkdocs-macros-plugin
        - run: pip install pymdown-extensions
        - run: pip install mkdocs-mermaid2-plugin
        - run: pip install mdx_gh_links
        - run: mkdocs gh-deploy --force

```

**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 ...

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



(Bio)informatics Skill Set

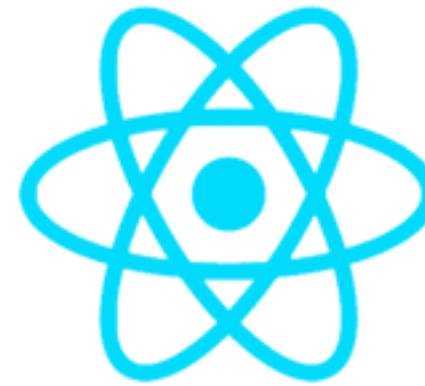
What has been needed to develop & maintain progenetix.org?

- Scripting and application development using Python, Perl and JavaScript
- Data analysis and plotting in R, Python and Perl
- Regular expressions for data entry and (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 (some OpenOffice Calc / Google Sheets spreadsheets for data cleanup)

(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!

BIO390: Course Schedule

- 2022-09-20: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2022-09-27: Christian von Mering - Sequence Bioinformatics
- 2022-10-04: Mark Robinson - Statistical Bioinformatics
- 2022-10-11: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2022-10-18: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2022-10-25: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2022-11-01: Katja Baerenfaller (SIAF) - Proteomics
- 2022-11-08: Pouria Dasmeh - Biological Networks
- 2022-11-15: Patrick Ruch - Text mining & Search Tools
- 2022-11-22: Ahmad Aghaebrahimian (ZHAW) - Semantic Web
- 2022-11-29: Michael Baudis - Building a Genomics Resource
- **2022-12-06: Valérie Barbie (SIB) - Clinical Bioinformatics**
- 2022-12-13: Michael Baudis - Genome Data & Privacy | Feedback
- 2022-12-20: Exam (Multiple Choice)

BIO390 HS22

Exam planning

- On site exam!
- 2022-12-20
- time: 08:15-09:45
- multiple (single + multiple) choice w/ one or two open questions
- no material, phones etc.
- student ID for entrance