

# A Reference Resource for Copy Number Variations in Cancer

Implementing GA4GH Standards to Drive an Open Oncogenomics Resource

1992



Heidelberg

Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Licher) | resident in clinical hematology/oncology | data, clinical studies & cancer systematics

2001



Stanford

Post-doc in hemato-pathology (Michael Cleary) | molecular mechanisms of leukemogenesis | transgenic models | expression arrays | systematic cancer genome data collection | *Progenetix* website

2003



Gainesville

Assistant professor in paediatric haematology | molecular mechanisms of leukemogenesis | focus on bioinformatics for cancer genome data analysis

2006



Aachen

Research group leader in genetics | genomic array analysis for germline alterations | descriptive analysis of copy number aberration patterns in cancer entities

2007



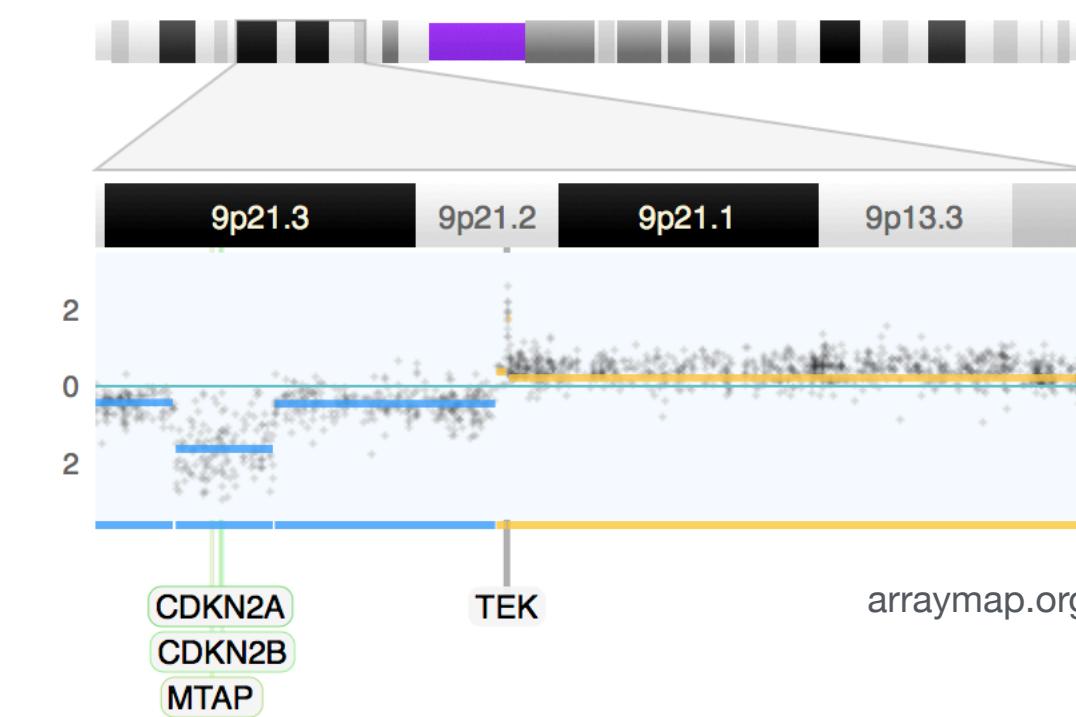
Zürich

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



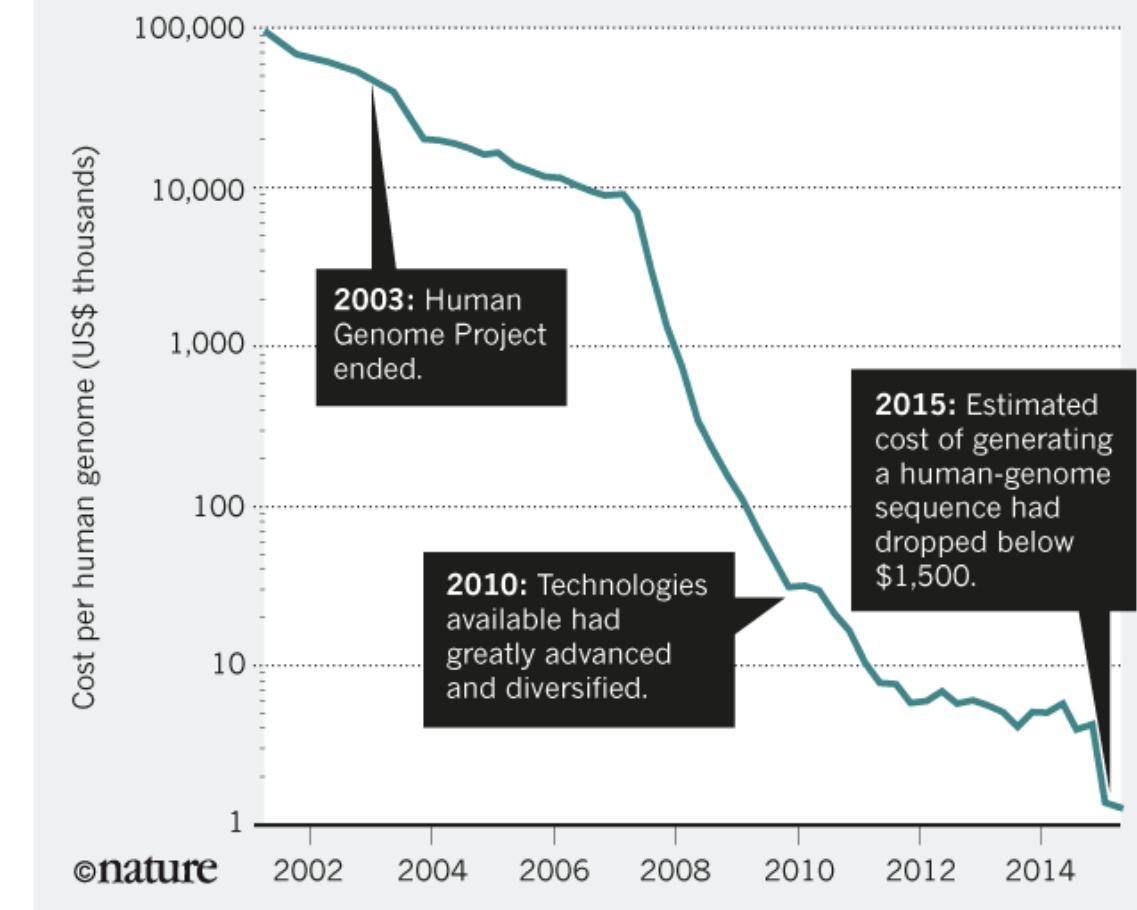
## Genome screening at the core of “Personalised Health”

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

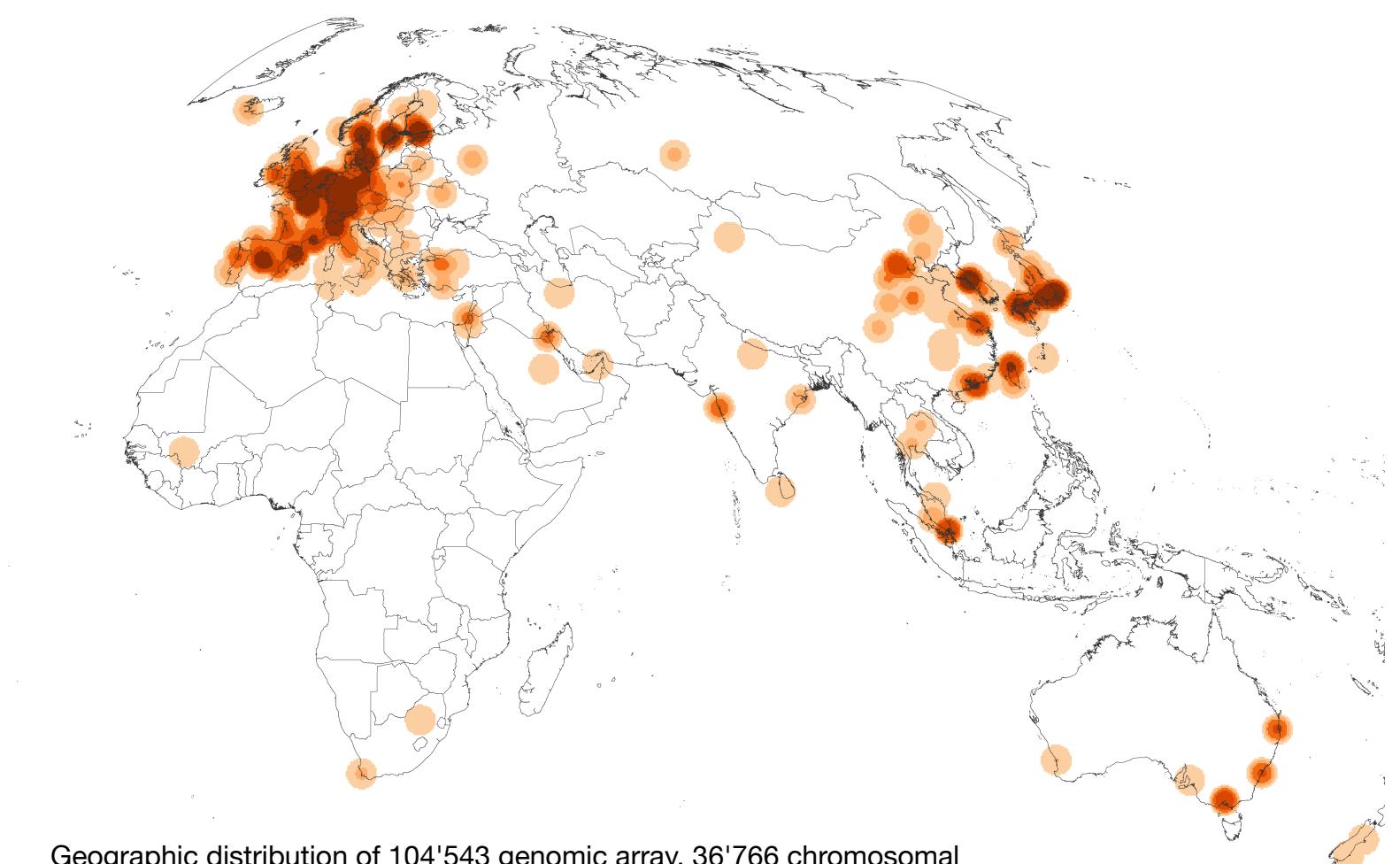
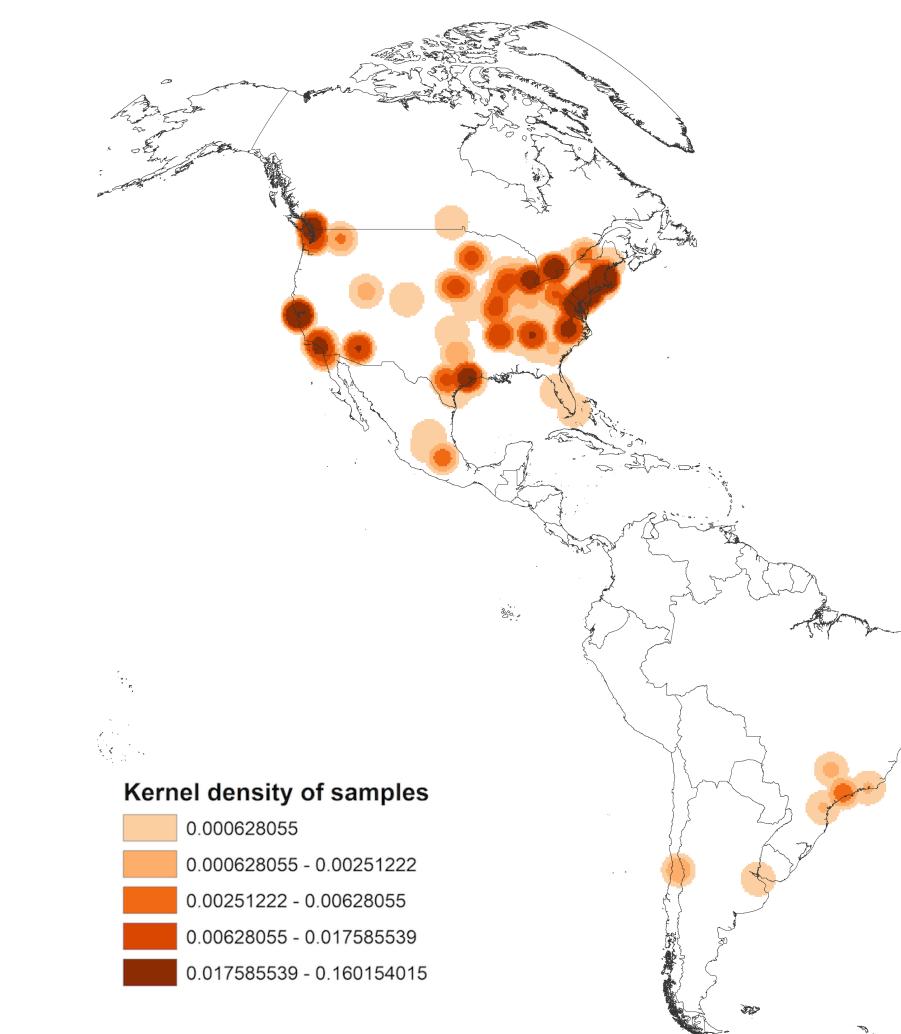


### BETTER, CHEAPER, FASTER

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)

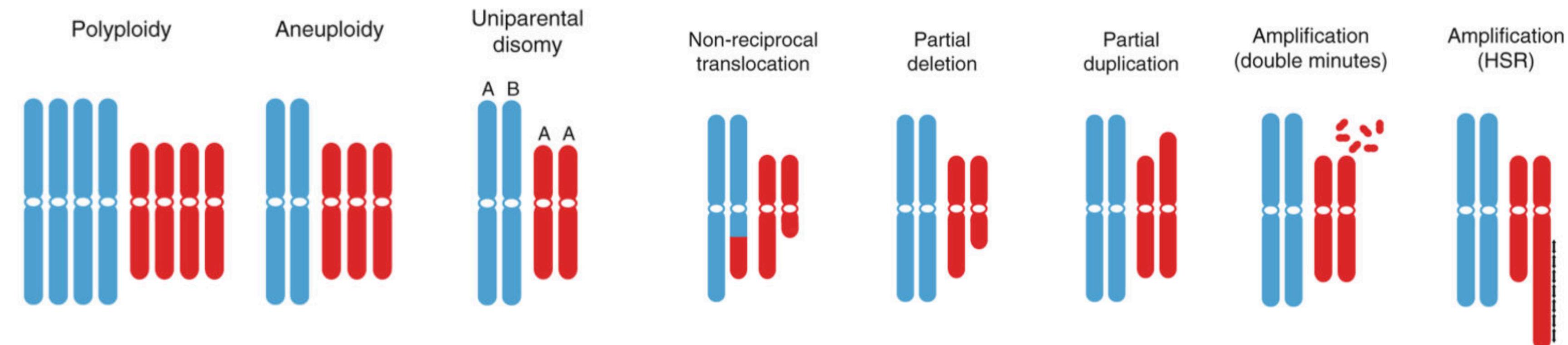


Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

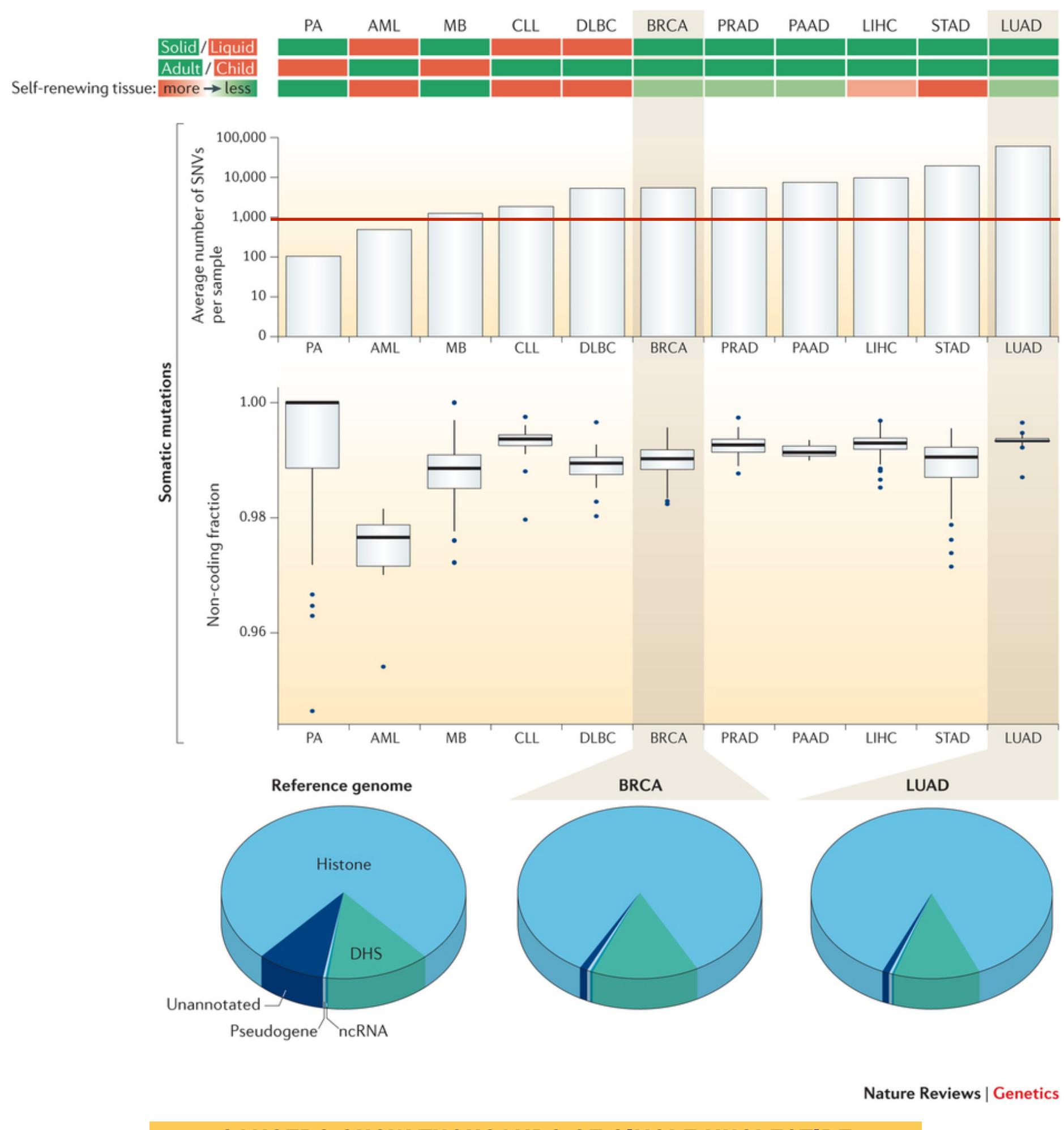
# Types of genomic alterations in Cancer

## Imbalanced Chromosomal Changes: CNV

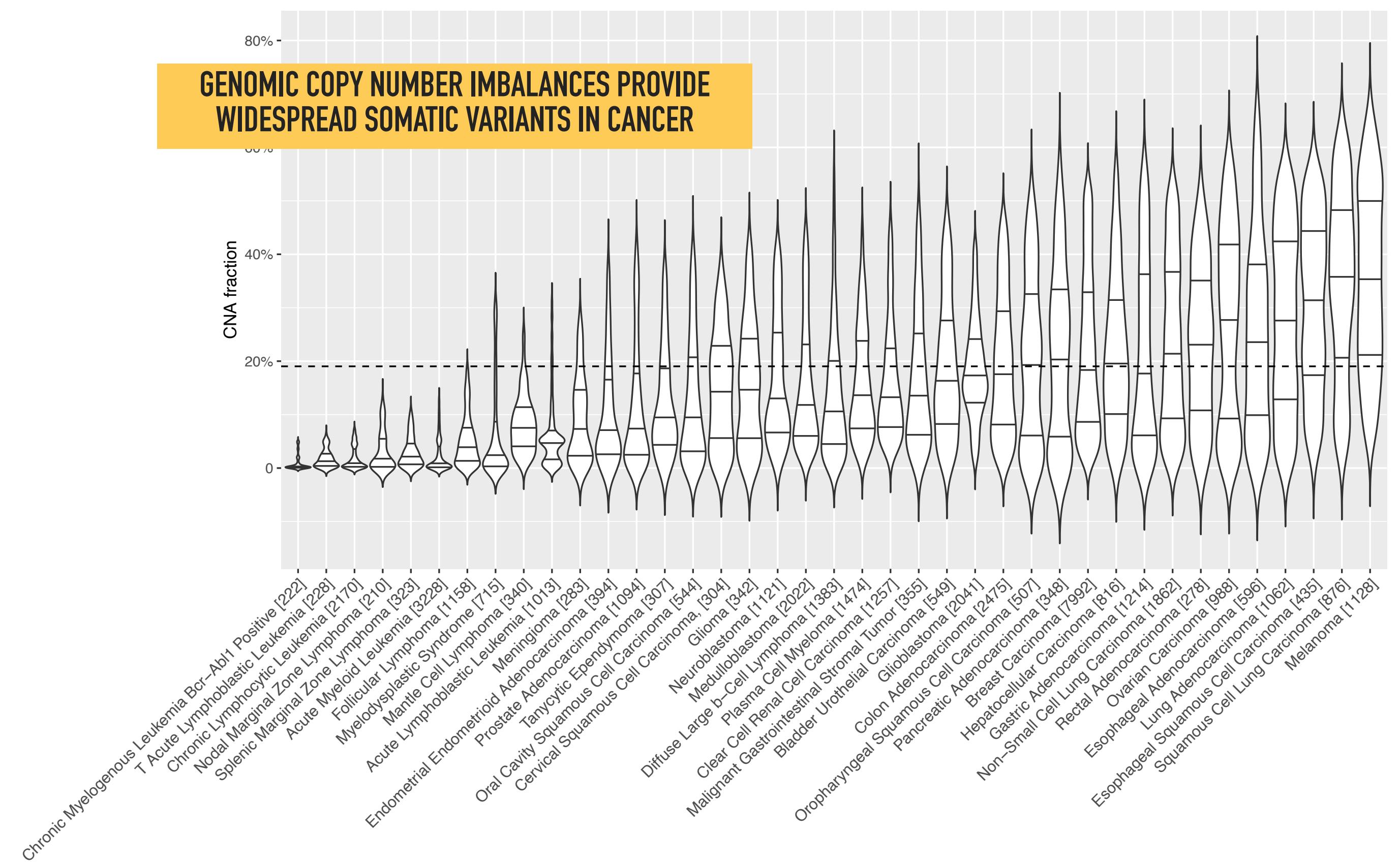
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



# 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

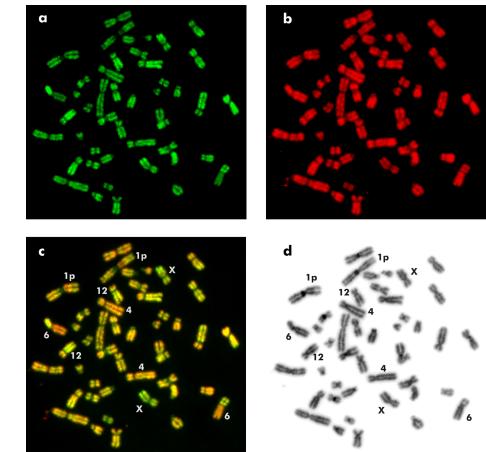
# **History & Current State...**

## **Origins & trajectory of the Progenetix Resource**



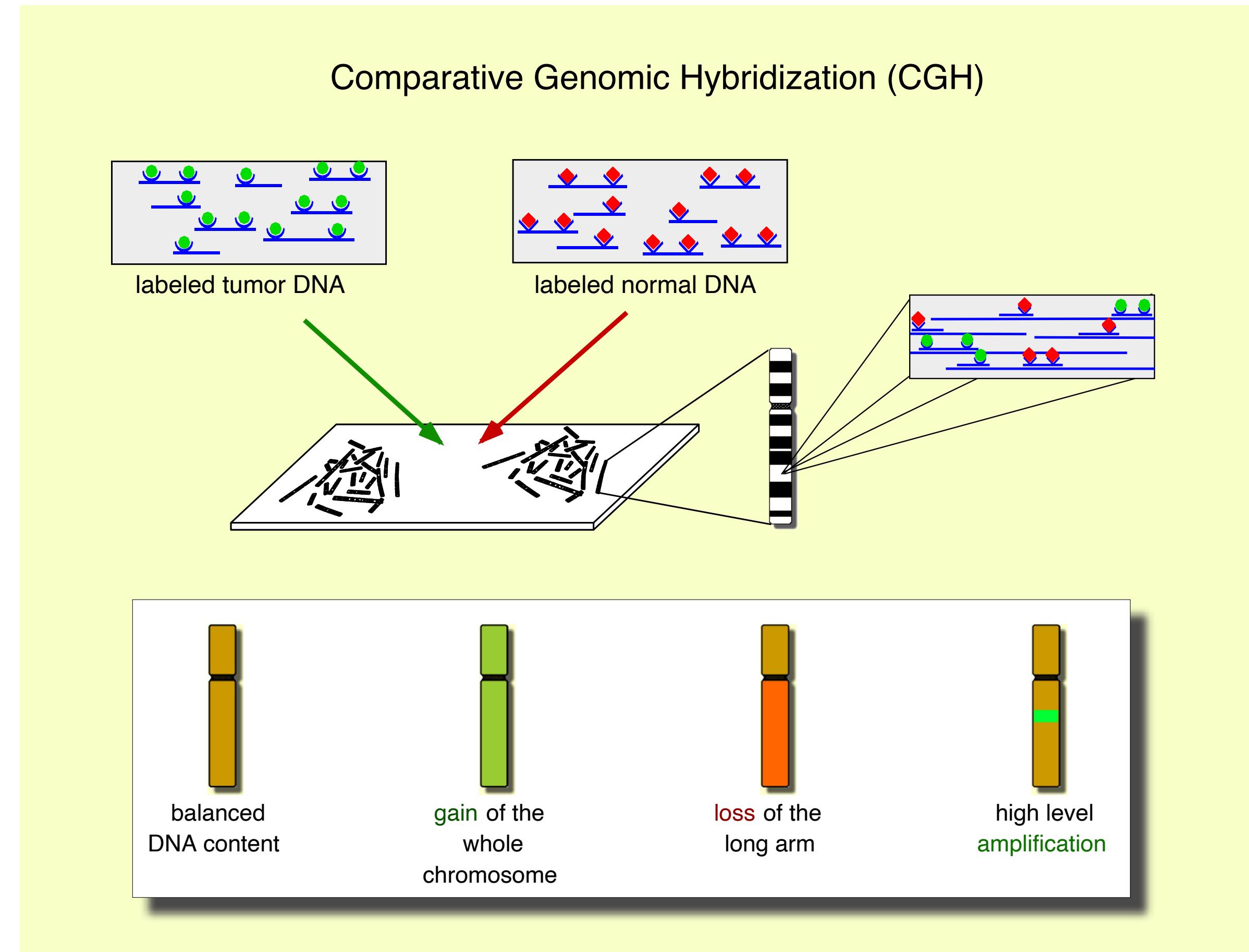
# Comparative Genomic Hybridization

## 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.

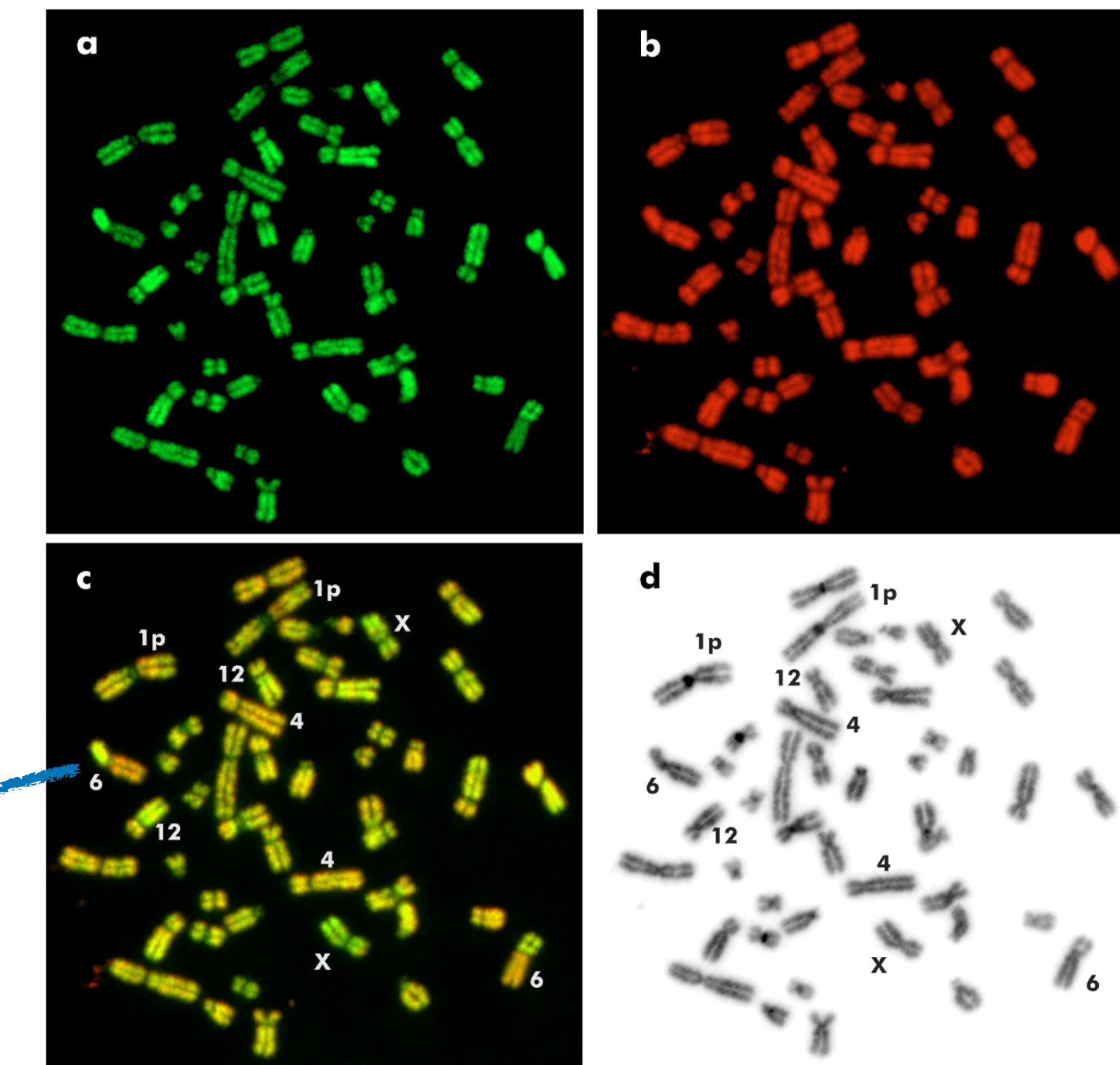


Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes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

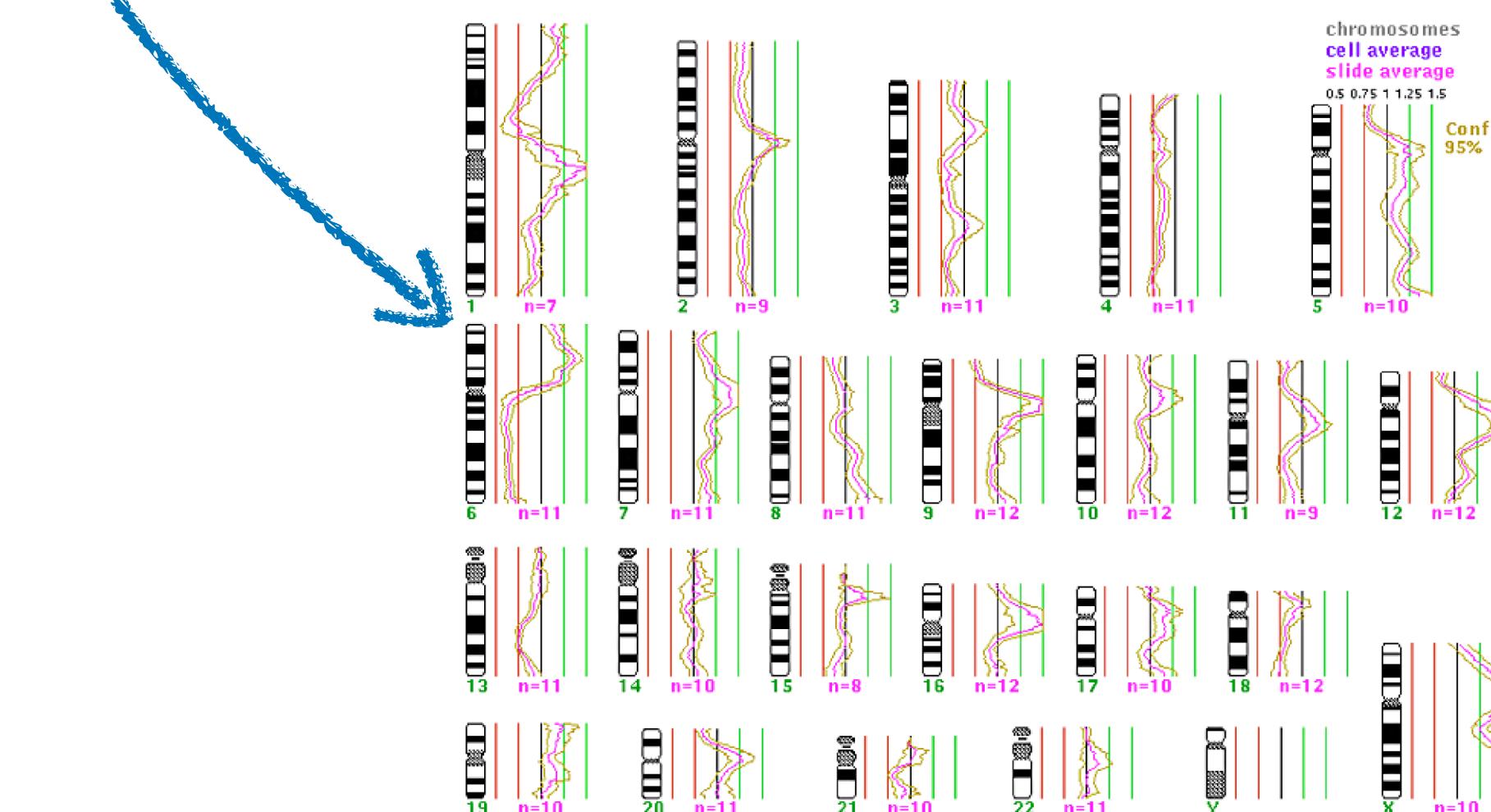
# Comparative Genomic Hybridization

## 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)



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

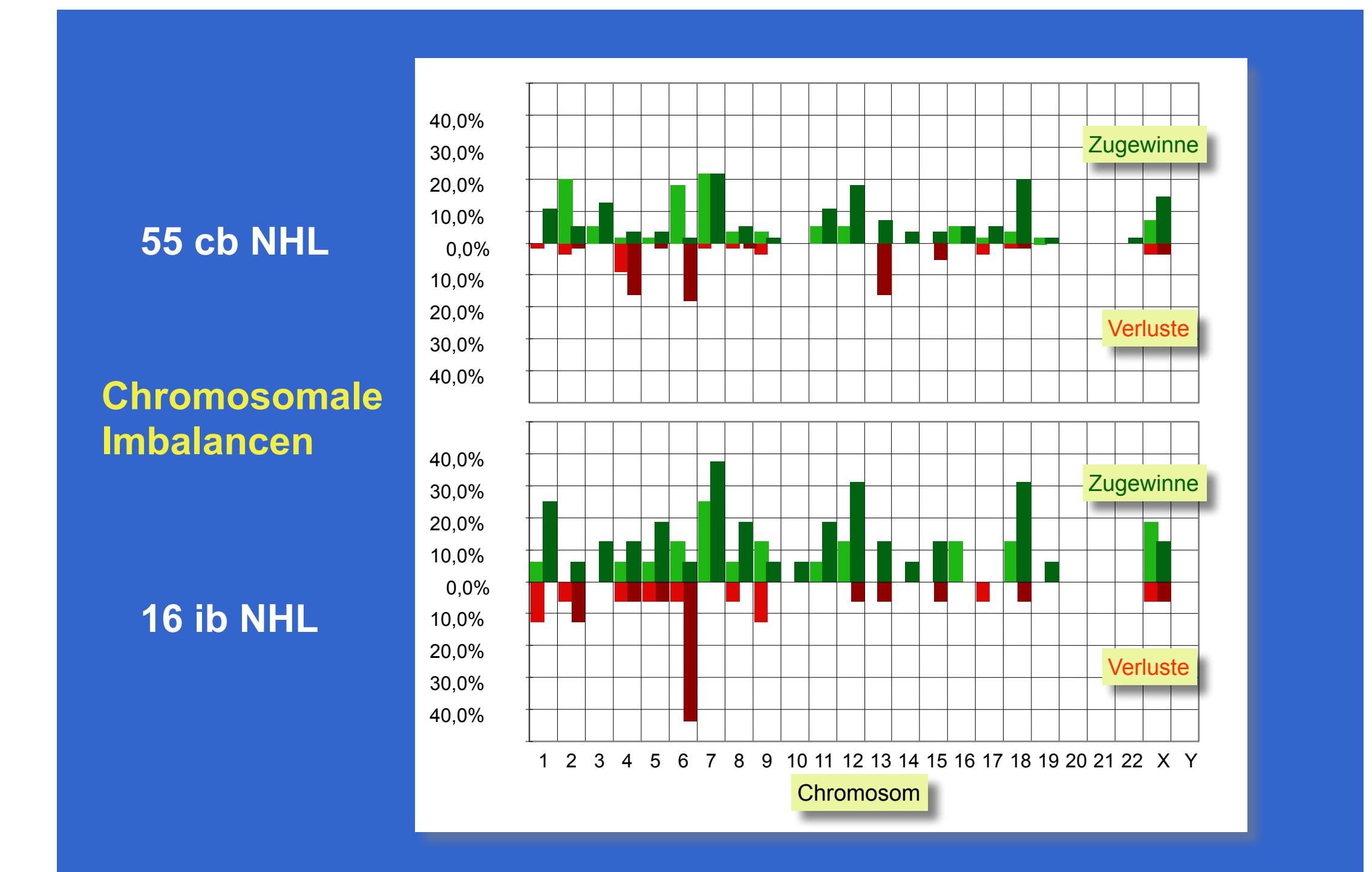
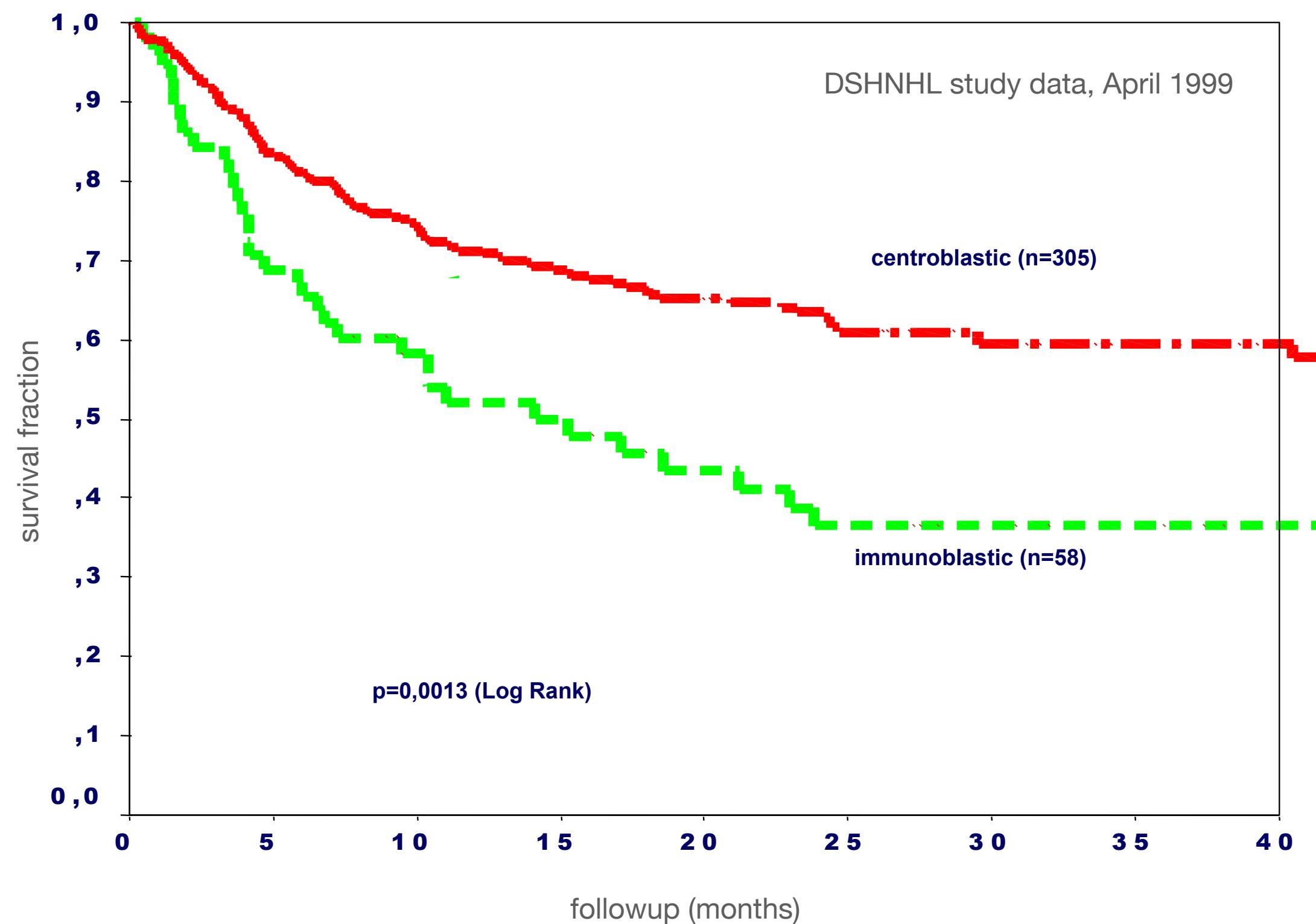


Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet. 1993;90:584-589.

# Cancer CNVs | Diagnostics | Prognosis

Single-study CNV frequencies correspond to diagnostic subsets

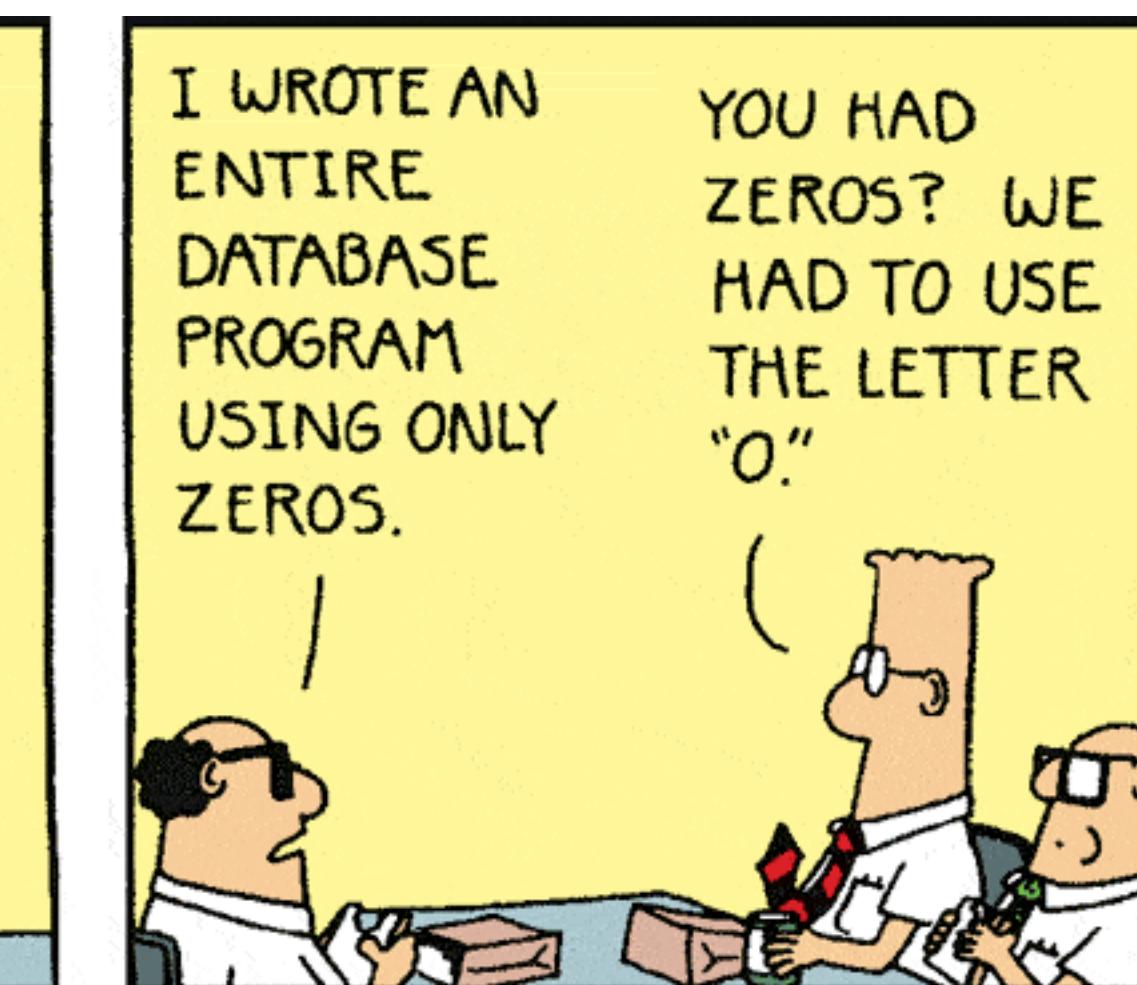


# Let's build a database!



| Tuesday February 27, 1996

## ... using archaic tools



| Tuesday September 08, 1992

# Progenetix CGH Database and Website

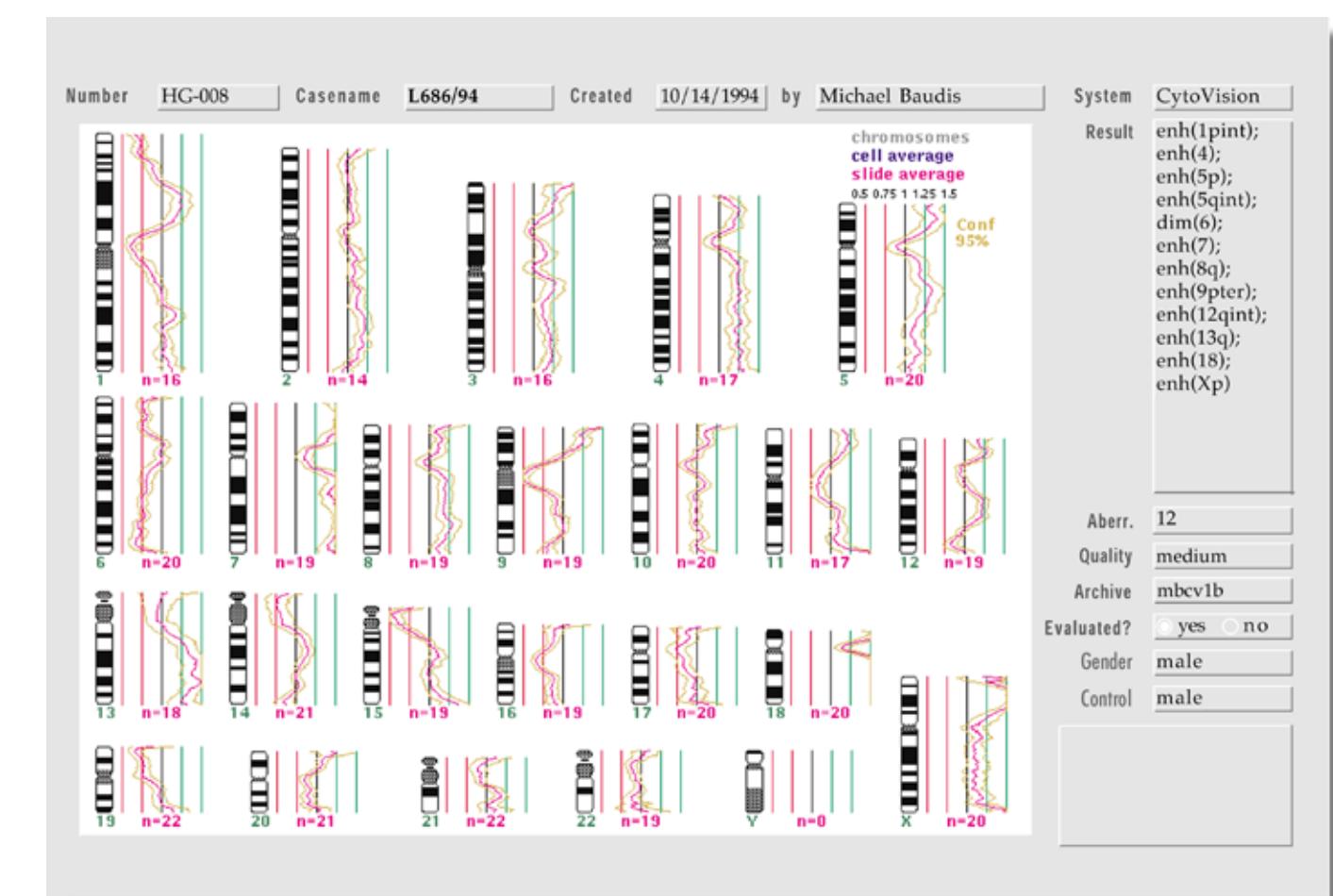
- originally an internal FileMaker Pro database, to store CGH profiles and annotations for the "Organization of Complex Genomes" group (head: Peter Lichter) at the German Cancer Research Center (DKFZ), starting in 1998
- expansion to include literature derived data, with a focus on malignant non-Hodgkin's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

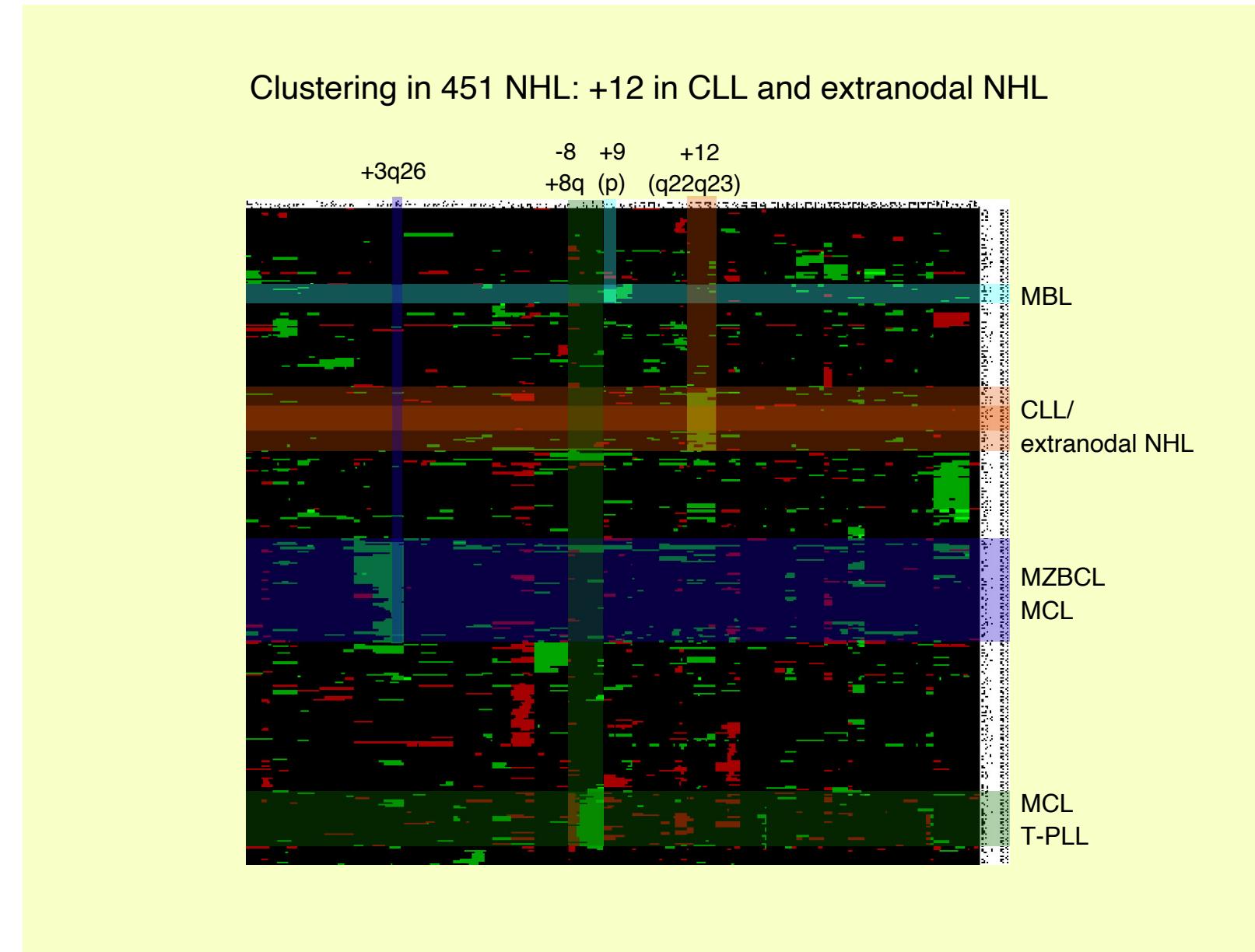
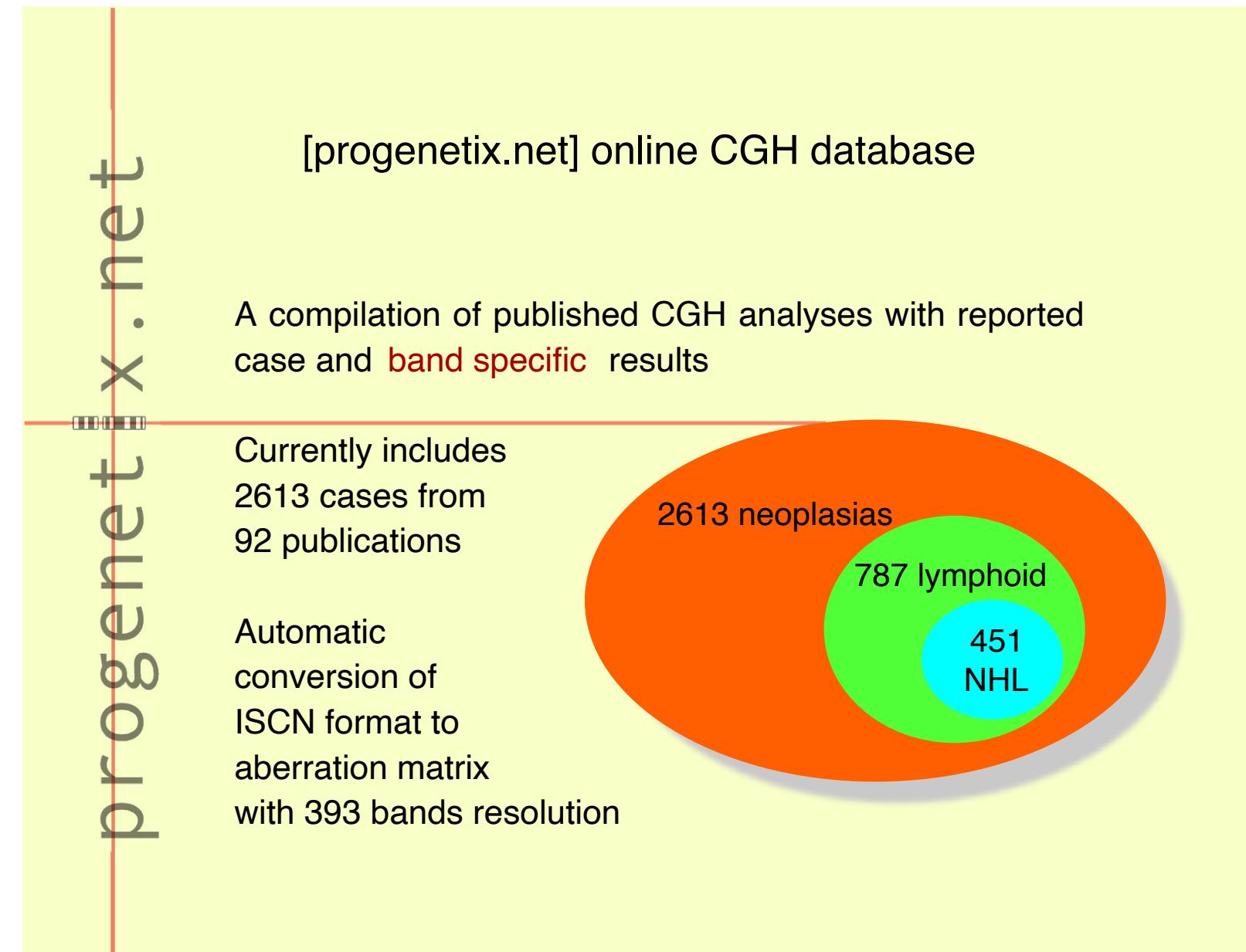
- Dec 6, 2000
  - first time online
- Nov 30, 2000
  - addition of graphical representation and gene table
- Nov 17, 2000
  - generation of website layout and database automatisation

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?



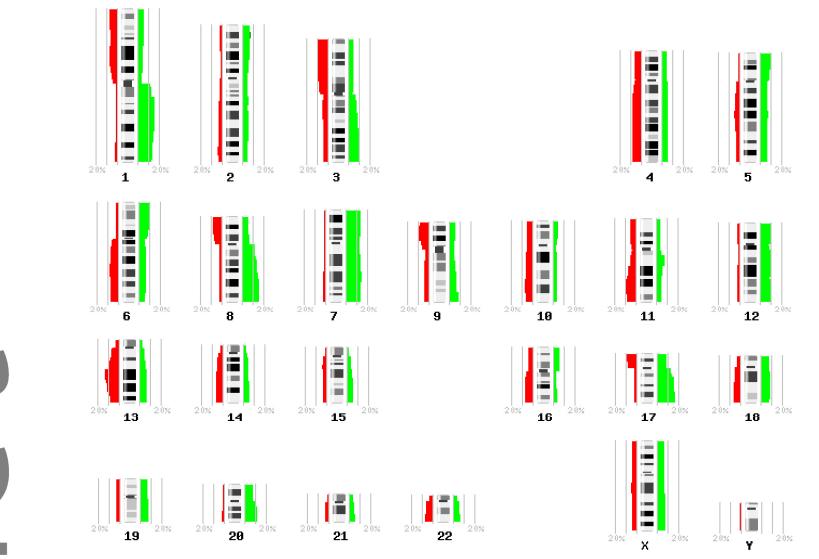


## Collection and Transformation of Chromosomal Imbalances in Human Neoplasias for Data Mining Procedures

michael baudis, dept. of pathology, stanford university

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

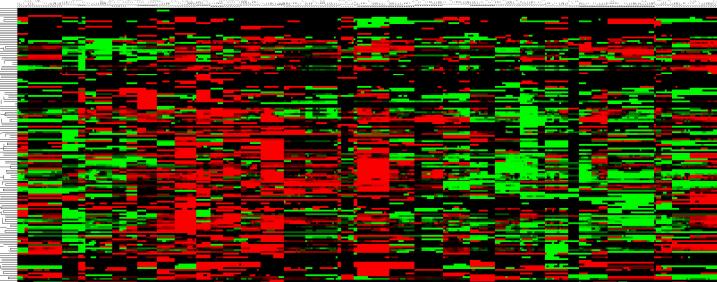
**Chromosomal imbalances in 5478 clinical cases from 196 publications**  
Although not as prominent as in specific subgroups, this large collection shows the non-random distribution of chromosomal gains (green) and losses (red).



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

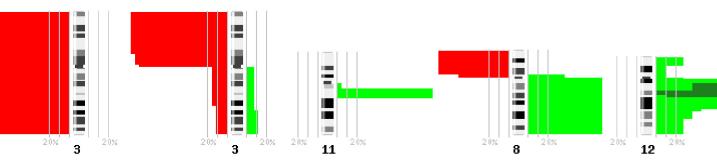


**Clustering of the band averages for the different ICD-O entities**  
Two dimensional clustering groups related disease entities and chromosomal bands with related aberrations.



**Results** Out of 4896 tumor samples, 3862 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.5% for a loss (max. 12.9% at 13q1) and 6.5% for a gain (max. 15.6% at 8q23). Differences between neoplastic entities showed in the average frequency and distribution pattern of imbalanced chromosomal regions. Tumor subsets (10 or more cases) with the strongest hot spots for losses were small cell lung carcinomas (ave. 23.3% with max. 96.2% at 3p14p26) and pheochromocytomas (ave. 10.9% with max. 92.7% at 3p); prominent gain maxima were found in pure high grade infiltrating duct carcinomas of the breast (ave. 5.9% with max. 95.7% at 11q13), T-PLL (ave. 4.7% with max. 81.8% for whole 8q) and dedifferentiated liposarcomas (ave. 10.4% with max. 81.8% at 12q13), among others. By cluster analysis, different combinations of chromosomal hot spot regions could be shown to occur in tumors subsummed in the same diagnostic entity; the example of neuroblastomas is shown.

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



**Conclusion** So far, progenetix.net project was able to:  
1. collect a large dataset of genomic aberration data generated through a molecular-cytogenetic screening technique (CGH)  
2. develop the software tools to transform those data to a meta format compatible to commonly used genomic interval descriptions  
3. produce graphical and numerical output from those data for hot spot detection and statistical analysis.

For future approaches, the data collection will be valuable for filtering data from expression array experiments for relevant genes, and possibly for the description of common and divergent genetic pathways in the oncogenetic process of different tumor entities. The transformed raw data of the progenetix.net collection is available for research purposes over the website.

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



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

# Progenetix Database in 2003

## Text conversion for CNVs

- based on listed CGH results from publications
  - ▶ literature detection using optimized PubMed queries
  - ▶ extraction (copy/paste, typing) of 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

[progenetix.net] molecular-cytogenetic data collection

Please read the [license](#), especially if you are not from an academic institution.

**Collection of published cytogenetic abnormalities in human malignancies**  
For all cases registered in [progenetix], band specific chromosomal aberration data is available to be included in data mining projects. The complete dataset can be accessed for download (see [\[here\]](#) for information).

The **ISCN2matrix converter** allows the online conversion from an aberration list in ISCN format to a band specific aberration matrix, with optional generation of a graphical representation.

**Software source**  
for storage and visualization of CGH data

7604 cases from 274 publications  
Newest resolution: 863 bands, matched to the "Golden Path" and ENSEMBL CytoView  
presented at BCATS 2001 and 2002 ([poster](#)) and the ASH 2001 meeting

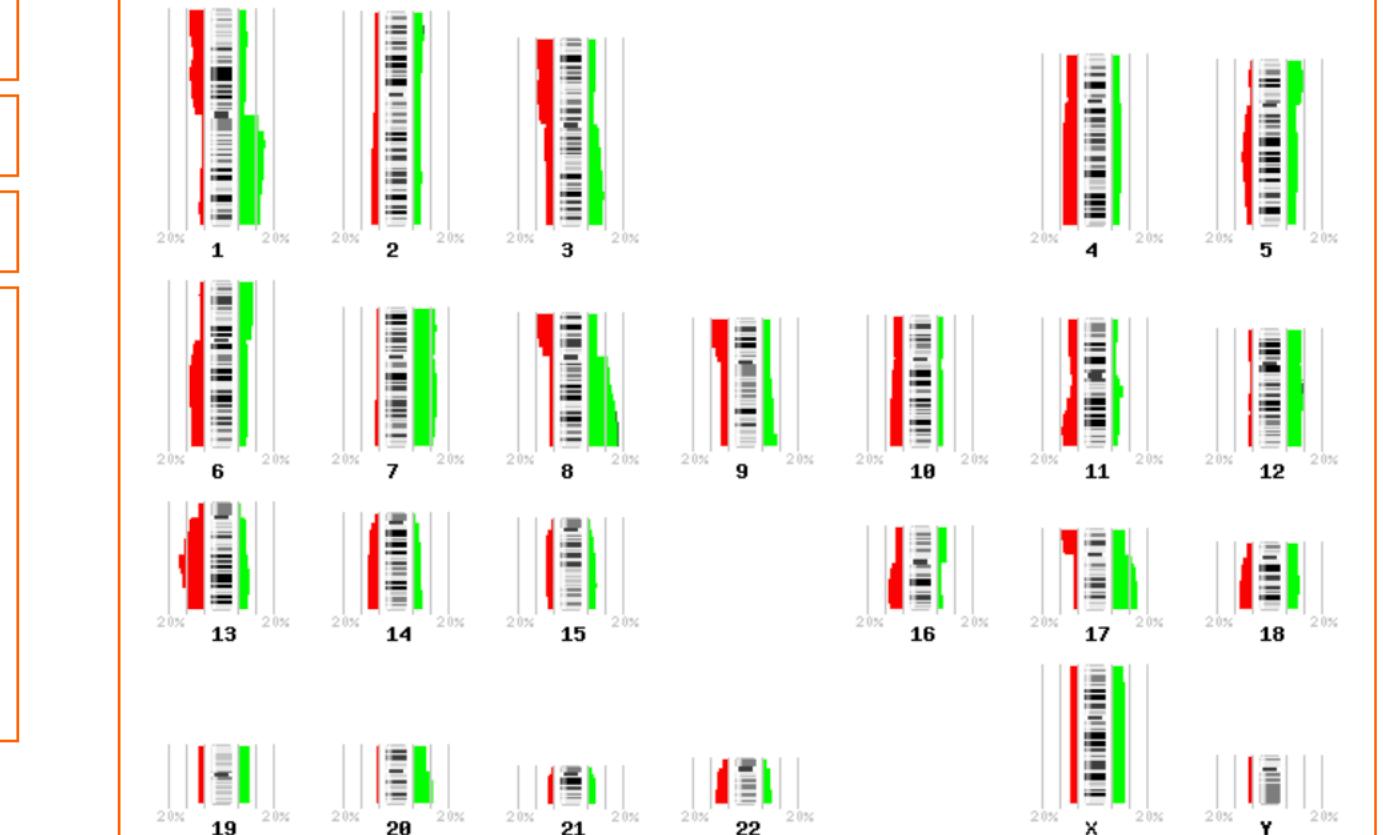
**Citation**

- Progenetix CGH online database. Baudis M. (2000-2003): [www.progenetix.net](http://www.progenetix.net).
- Progenetix.net: an online repository for molecular cytogenetic aberration data. Baudis M. and Cleary M. *Bioinformatics* 17 (12) 2001: 1228-1229.

**Submission**  
Casestables should be sent to [progenetix.net](mailto:progenetix.net).

sponsored by a gift from METASYSTEMS

**Server & Browser**  
The new version of the site is run on a commercial server, using RedHat Linux and [Apache](#) server software. It is optimized for newer generation browsers and is tested using [Camino](#) under [OS X](#).



**PLOS**

**Publications** lists the articles currently contained in the database with links to PubMed. Casestables list all cases of the according project with their chromosomal imbalances in an ISCN adapted format.

**ICD-O Entities** lists all disease entities throughout the collection according to their ICD-O (3) codes and links to the respective graphical representations

**Predefined Groups** combine data from related disease entities

# 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)	<a href="#">12666986</a>	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)	<a href="#">12666986</a>	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)	<a href="#">12666986</a>	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	<a href="#">12579536</a>	rev ish enh(16q13qter)
HCC-MF-01T2	Hepatocellular carcinoma	<a href="#">12579536</a>	rev ish enh(12q22qter, 17q) dim(16q)
HCC-MF-01T3	Hepatocellular carcinoma	<a href="#">12579536</a>	rev ish enh(12q21.3qter, 17q21qter) dim(16q21qter)
HCC-MF-02T1	Hepatocellular carcinoma	<a href="#">12579536</a>	rev ish dim(6q13qter)
HCC-MF-02T2	Hepatocellular carcinoma	<a href="#">12579536</a>	rev ish enh(1q, 17q) dim(17p)
HCC-MF-03T1	Hepatocellular carcinoma	<a href="#">12579536</a>	rev ish enh(1q, 3q26.2qter, 4p, 6p21.1pter, 11p15, 19q) dim(16q10q12.2)
HCC-MF-03T2	Hepatocellular carcinoma	<a href="#">12579536</a>	rev ish enh(8q, 11p15, 12pterq12) dim(3p, 4q, 5q, 8p23.1, 9q, 16q) amp(1q)
HCC-MF-04T1	Hepatocellular carcinoma	<a href="#">12579536</a>	rev ish enh(1p33qter, 8q21.2qter) dim(1pterp34, 4q, 9q) amp(6p, 13q21qter)
HCC-MF-04T2	Hepatocellular carcinoma	<a href="#">12579536</a>	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	<a href="#">12579536</a>	rev ish enh(6q, 8q12qter, 12q21.1qter, 13q22qter, 17q, dim(4q, 5q, 7p, 7q, 9q10q31, 11q, 14q, 16q) amp(10p)
HCC-MF-05T2	Hepatocellular carcinoma	<a href="#">12579536</a>	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	<a href="#">12579536</a>	rev ish enh(1q, 5p23pter, 18p, 22) dim(4q, 6q, 9pterq33, 13q, 14q, 16pterq23) amp(8q)

# 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 <sup>a</sup>	Grade <sup>b</sup>	Diagnosis of metastatic disease <sup>c</sup>
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

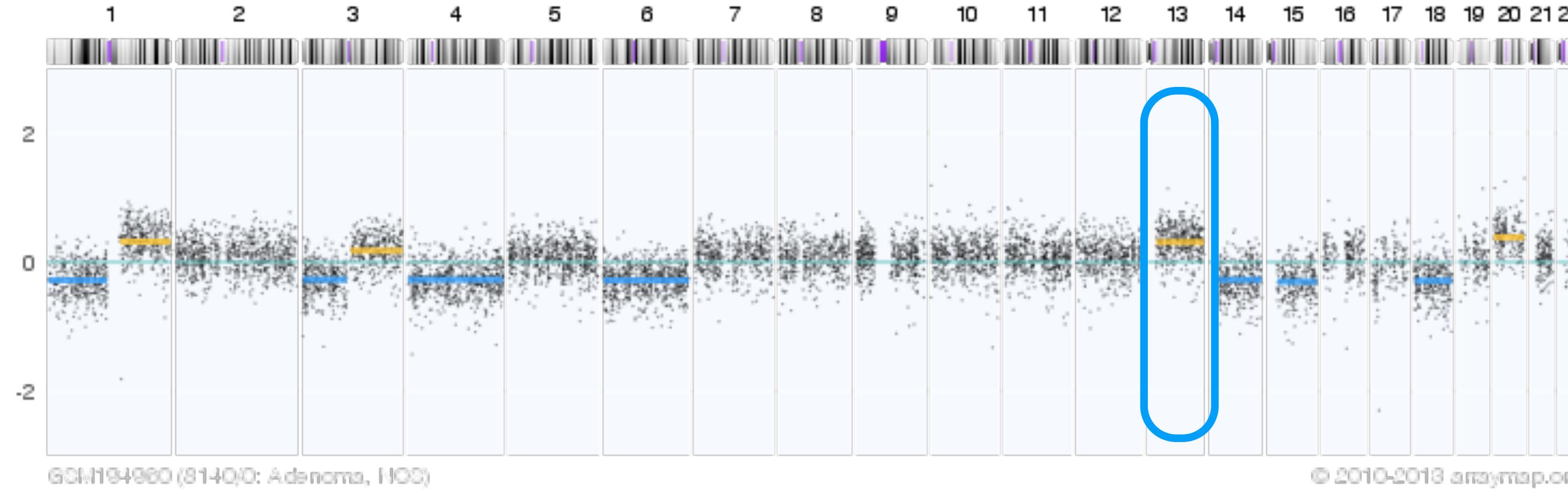
<sup>a</sup>AJCC/UICC staging system (Hutter and Sabin, 1986).<sup>b</sup>Grade of primary tumor: 1–3, low, moderate, high grade; 9, grading unknown.<sup>c</sup>Synchronous, diagnosis of metastatic disease within 12 months following diagnosis of primary tumor; metachronous, diagnosis of metastatic disease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

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

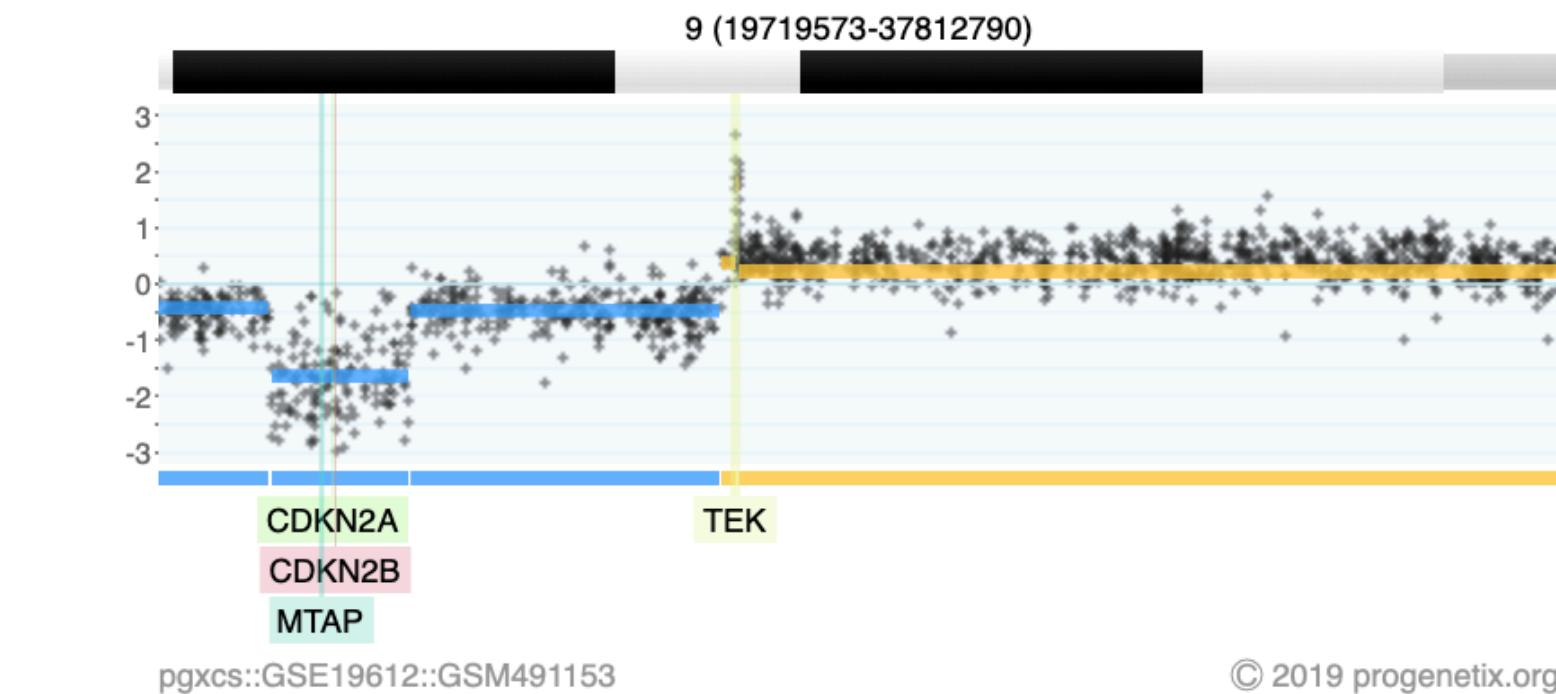
**Chromosome Arm 20q Gains and Other Genomic Alterations in Colorectal Cancer Metastatic to Liver, as Analyzed by Comparative Genomic Hybridization and Fluorescence In Situ Hybridization**

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

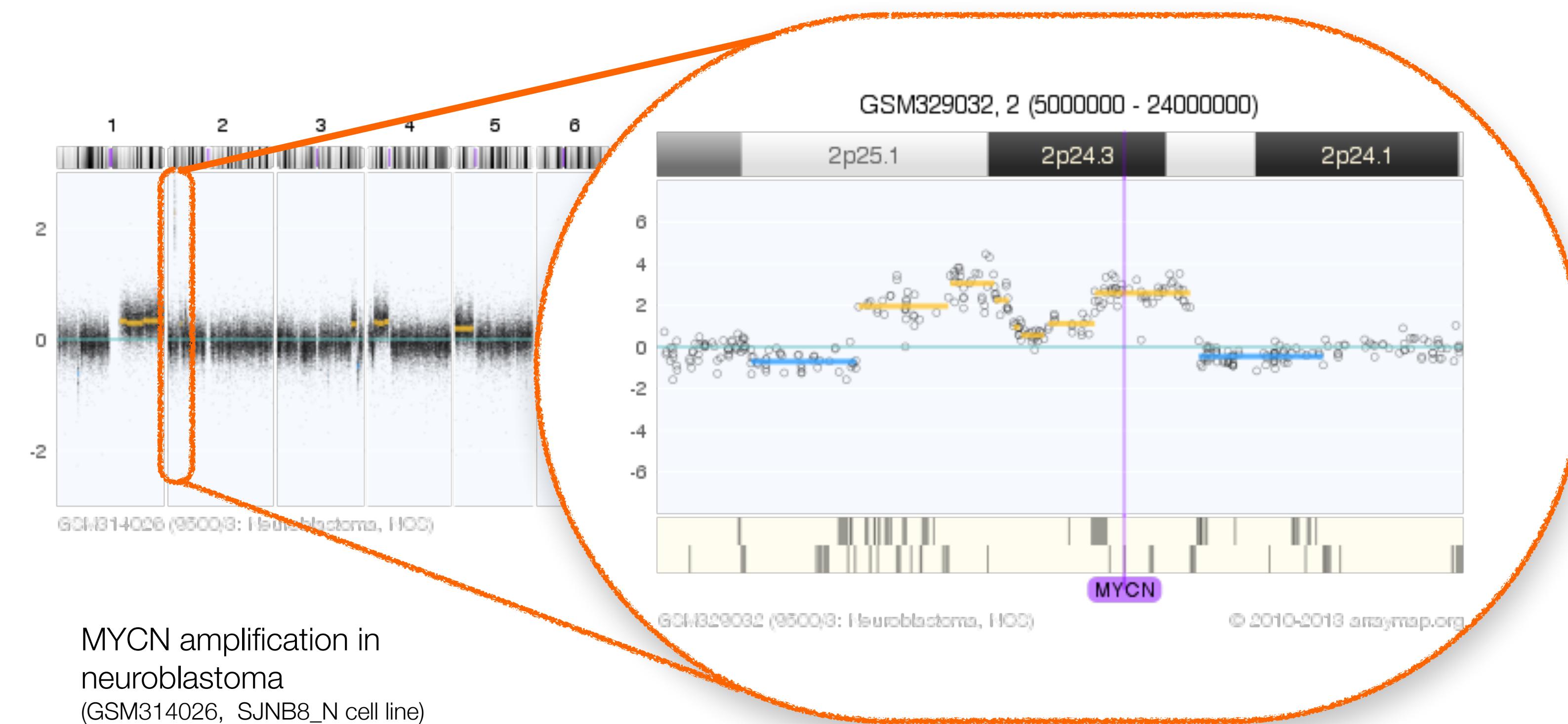
# Array-based Detection of Copy Number Variations



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<sup>+</sup>

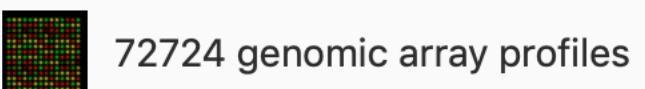


162.158.150.56

### 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.

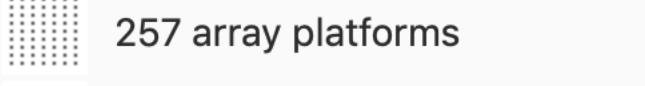
The current data reflects:



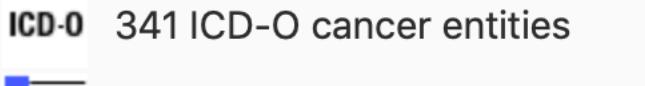
72724 genomic array profiles



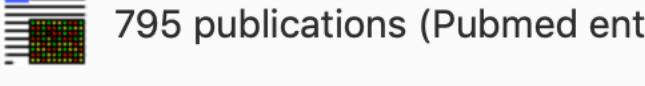
898 experimental series



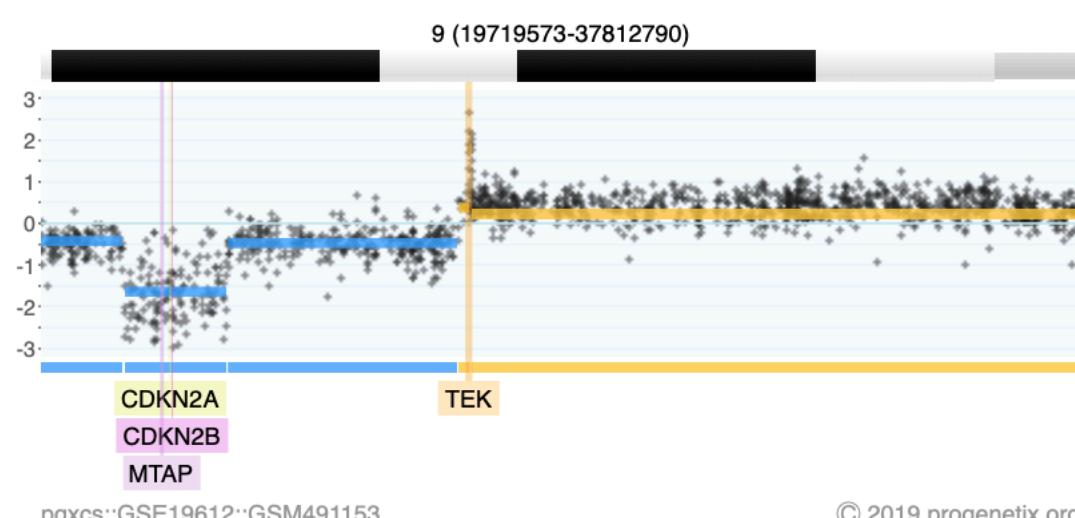
257 array platforms



341 ICD-O cancer entities



795 publications (Pubmed entries)



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

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

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

#### RELATED PUBLICATIONS

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

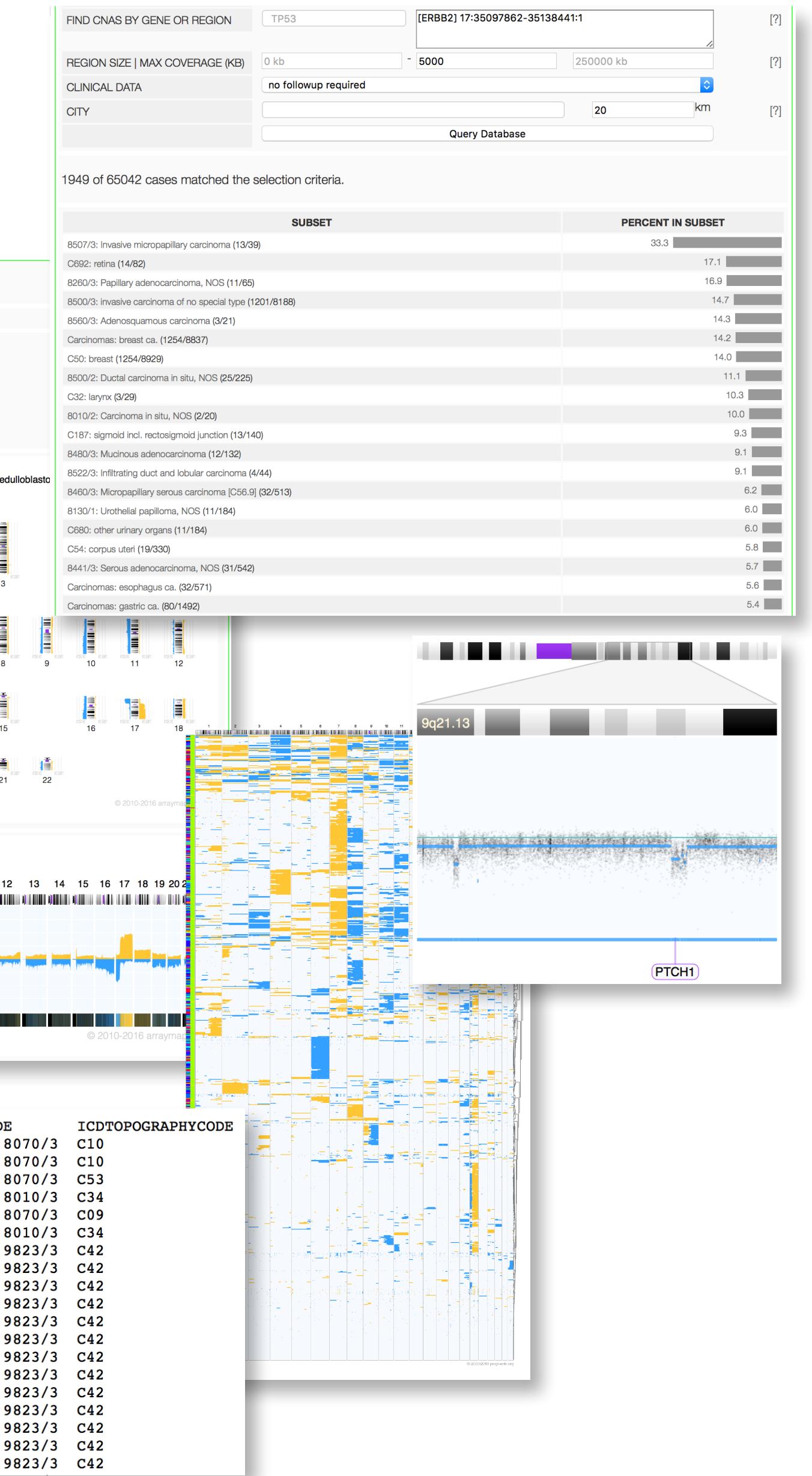
Cai, H., Kumar, N., & Baudis, M. 2012. arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies. *PLoS One* 7(5), e36944.

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

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

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

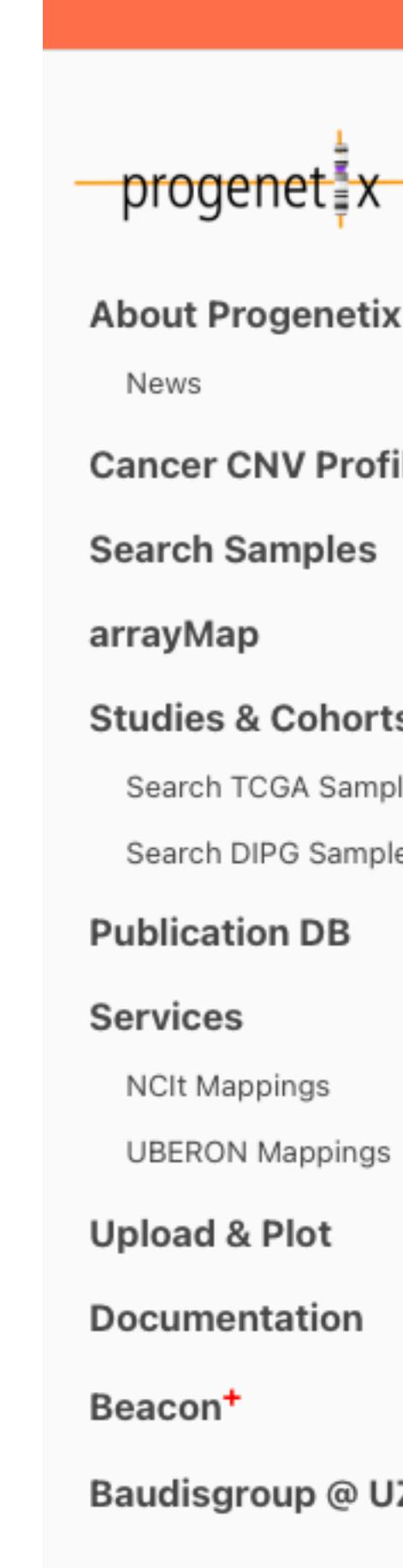
Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project.



# Progenetix in 2021

## Cross-platform Oncogenomics

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



The image shows the navigation menu of the Progenetix website. It includes links for About Progenetix, News, Cancer CNV Profiles, Search Samples, arrayMap, Studies & Cohorts, Search TCGA Samples, Search DIPG Samples, Publication DB, Services, NCIt Mappings, UBERON Mappings, Upload & Plot, Documentation, Beacon<sup>+</sup>, and Baudisgroup @ UZH.

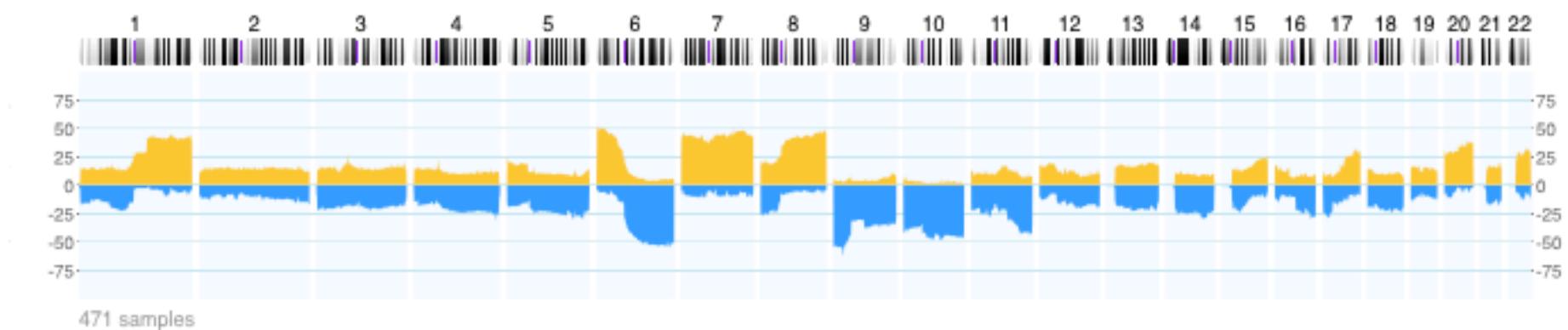
### Cancer genome data @ [progenetix.org](https://progenetix.org)

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.

For exploration of the resource it is suggested to either start with:

- [Cancer Types](#)
- [searching](#) for CNVs in genes of interest

Non-Cutaneous Melanoma (NCIT:C8711)



#### [Download SVG](#)

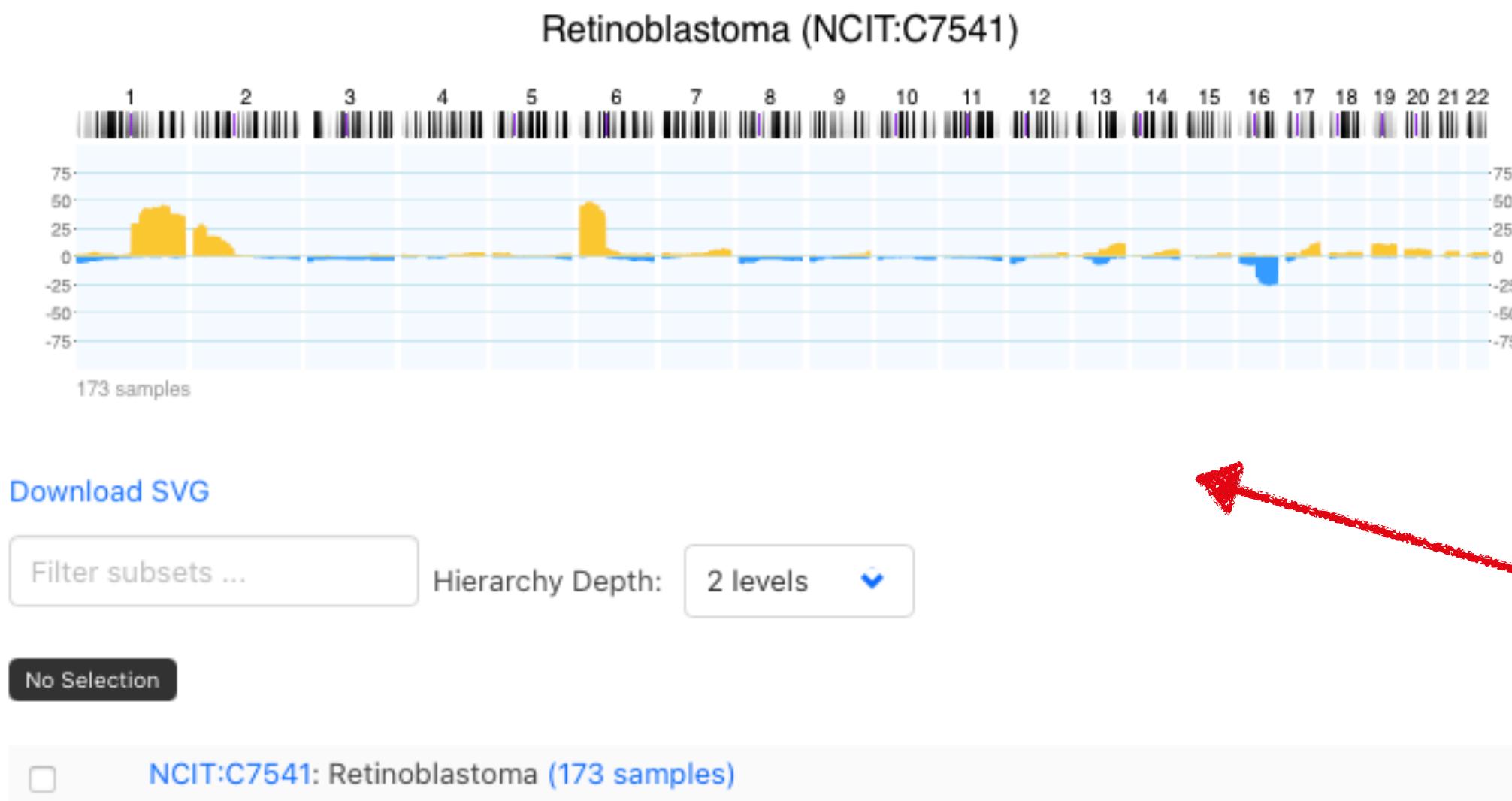
The resource currently contains genome profiles of **139448** individual samples and represents cancer types, according to the NCIt "neoplasm" classification.

Additionally to this genome profiles and associated metadata, the website present information about publications (currently **4024** articles) referring to cancer genome profiling experiments.

# Progenetix

## Cancer Type CNA Data

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



progenetix

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

Cancer Types

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

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

Cancer Classification: NCIT Cancer Core

Filter subsets ... Hierarchy Depth: 2 levels

No Selection

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

[About Progenetix](#)[Cancer CNV Profiles](#)[Search Samples](#)[Publication DB](#)[Services](#)[Upload & Plot](#)[Documentation](#)[Beacon<sup>+</sup>](#)[Baudisgroup @ UZH](#)

## Search Samples

[CDKN2A Deletion Example](#)[MYC Duplication](#)[TP53 Del. in Cell Lines](#)

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

 [Gene Spans](#) [Cytoband\(s\)](#)**Reference name** i

9

**(Structural) Variant Type** i

DEL (Deletion)

**Start or Position** i

21500001-21975098

**End (Range or Structural Var.)** i

21967753-22500000

**Cancer Classification(s)** i

NCIT:C3058: Glioblastoma (4358)

|

**Biosample Type** i**Filters** i**Filter Logic** i

AND

**City** Select... [Query Beacon](#)

[About Progenetix](#)[Cancer CNV Profiles](#)[Search Samples](#)[Publication DB](#)[Services](#)[Upload & Plot](#)[Documentation](#)[Beacon<sup>+</sup>](#)[Baudisgroup @ UZH](#)

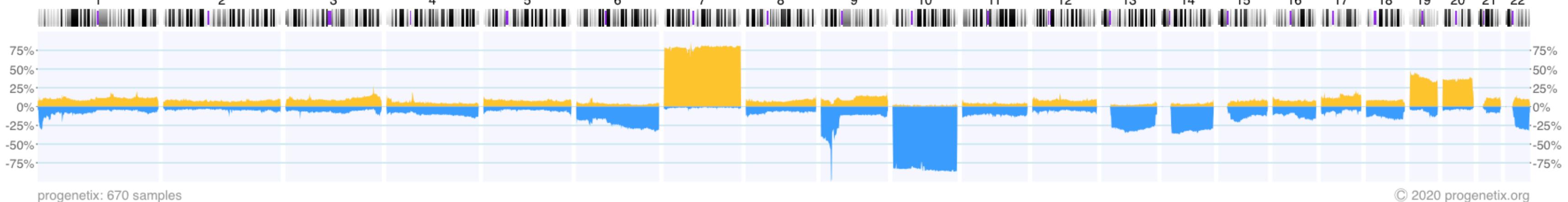
Assembly: GRCh38 Chro: 9 Start: 21500001-21975098 End: 21967753-22500000 Type: DEL Filters: NCIT:C3058

progenetix

Samples: 668

Variants: 286

Calls: 675

 $f_{alleles}$ : 0.000088[Phenopackets](#)[Callsets Variants](#)[Variants in UCSC](#)[UCSC region](#)[JSON Response](#)[Visualization options](#)[Results](#) [Biosamples](#) [Biosamples Map](#) [Variants](#)

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



Search Samples



About Progenetix

Cancer CNV Profiles

Search Samples

Publication DB

Services

Upload & Plot

Documentation

Beacon<sup>+</sup>

Baudisgroup @ UZH

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

progenetix

Samples: 668

Phenopackets

UCSC region

Variants: 286

Callsets Variants

JSON Response

Calls: 675

Variants in UCSC

f<sub>alleles</sub>: 0.000088

Visualization options

Results

Biosamples

Biosamples Map

Variants

JSON

[Download Response](#)

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



Page 1 of 135

# Progenetix

## Services, Documentation...

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

**progenetix**

**Services: Ontologymaps (NCIt)**

The **ontologymaps** service provides equivalence 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**

gio|

NCIT:C3058: Glioblastoma  
NCIT:C3288: Oligodendrogloma  
NCIT:C4326: Anaplastic Oligodendrogloma  
NCIT:C3903: Mixed Glioma  
NCIT:C3059: Glioma  
NCIT:C3796: Gliosarcoma  
NCIT:C4822: Malignant Glioma  
NCIT:C3308: Paraganglioma  
NCIT:C4831A: Renal Paraganglioma

**About Progenetix**

News  
Cancer CNV Profiles  
Search Samples  
arrayMap  
Studies & Cohorts  
Publication DB  
Services  
Upload & Plot  
Documentation  
Beacon+  
Baudisgroup @ UZH

**NCIthesaurus**

**Progenetix :: Info**

Structural Cancer Genomics Resource Documentation and Example Pages

**News**  
**About...**  
**API**  
**Documentation**  
**Publications**  
**Progenetix Home**

**Related Sites**

Progentix Data  
Baudisgroup @ UZH  
Beacon+  
SchemaBlocks {S}[B]  
Beacon Project

**Github Projects**

baudisgroup  
progenetix  
ELIXIR Beacon

**Tags**

API Beacon BeaconPlus  
BeaconV2 GA4GH Perl Python  
TCGA article bycon code  
documentation identifiers licensing  
ontologies prefixes schemas  
services statistics tools website

## Welcome to the *Progenetix* documentation pages

The **Progenetix Resource Documentation** provides information and links related to the **Progenetix** cancer genome resource and the related **Progenetix code repositories** contains projects, such as data conversion scripts, ontology mappings and code for the **Beacon+** project.

### Progenetix Website Code Repositories

- **Progenetix Source Code**
- **Related Projects**

### Latest News

#### Progenetix File Formats

##### Standard Progenetix Segment Files [.pgxseg](#)

Progenetix uses a variation of a standard tab-separated columnar text file such as produced by array or sequencing CNV software, with an optional metadata header for e.g. plot or grouping instructions.

@mbaudis 2021-02-22: [more ...](#)

#### Beacon+ and Progenetix Queries by Gene Symbol

We have introduced a simple option to search directly by Gene Symbol, which will match to *any* genomic variant with partial overlap to the specified gene. This works by expanding the Gene Symbol (e.g. *TP53*, *CDKN2A* ...) into a range query for its genomic coordinates (maximum CDR).

Such queries - which would e.g. return all whole-chromosome CNV events covering the gene of interest, too - should be narrowed by providing e.g. **Variant Type** and **Maximum Size** (e.g. 2000000) values.

@mbaudis 2021-02-22: [more ...](#)

Gene Symbol
MYC
MYCBP (1:38864669-38873304)
MYCBPAP (17:56508545-50531427)
MYCL (1:39897371-39901887)
MYCN (2:15940586-15946096)

#### The Progenetix oncogenomic resource in 2021

Qingyao Huang, Paula Carrio Cordero, Bo Gao, Rahel Paloots, Michael Baudis

bioRxiv. doi: <https://doi.org/10.1101/2021.02.15.428237>

This article provides an overview of recent changes and additions to the Progenetix database and the services provided through the resource.

2021-02-15: [more ...](#)

#### Diffuse Intrinsic Pontine Glioma (DIPG) cohort

Diffuse Intrinsic Pontine Glioma (DIPG) is a highly aggressive tumor type that originates from glial cells in the pons area of the brainstem, which controls vital functions including breathing, blood pressure and heart rate. DIPG occurs frequently in the early childhood and has a 5-year survival rate below 1 percent. Progenetix has now incorporated the DIPG cohort, consisting of 1067 individuals from 18 publications. The measured data include copy number variation as well as (in part) point mutations on relevant genes, e.g. TP53, NF1, ATRX, TERT promoter.

@qingyao 2021-02-15: [more ...](#)

#### arrayMap is Back

After some months of dormancy, the **arrayMap** resource has been relaunched through integration with the new **Progenetix** site. All of the original arrayMap data has now been integrated into Progenetix, and of today the [arraymap.org](#) domain maps to a standard Progenetix search page, where only data samples with existing source data (e.g. probe specific array files) will be presented.

@mbaudis 2021-02-06: [more ...](#)



# DATA PIPELINE

## BIOCURATION BIOINFORMATICS



NCBI GEO Accession Display

Series GSE640034 Public on Sep 07, 2013

Organism: Human

Experiment type: Genomic variation profiling by array

Summary: This study investigates genomic variation in chronic lymphocytic leukemia (CLL) specimens with prognostic significance.

Overall design: Overall design experiment, Test vs. Specimens vs. Reference human genome

Contributor(s): Houldsworth J, Venkata T, Guttagji A, Thoduri V, Yan XI et al.

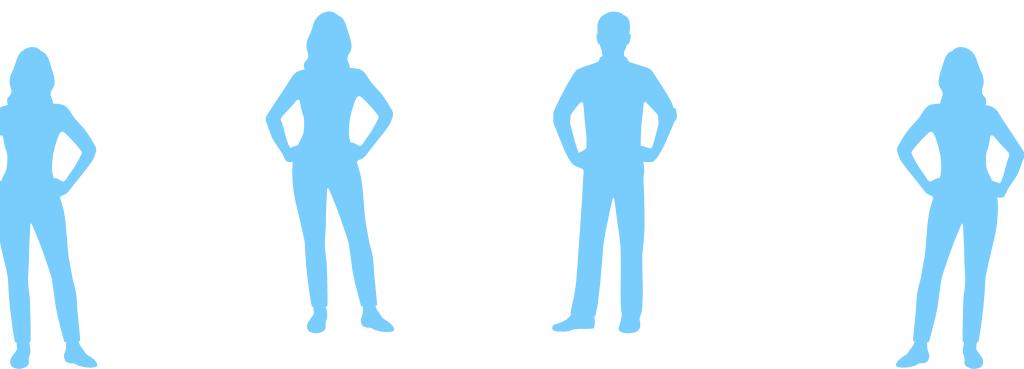
Phone: +41 61 267 32 32

Address: University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Sample ID: GSE640034

Platform: Agilent G1317P Lymphoma (Agilent Probe ID)

Supplementary file: GSE64034.RAW.tar



arrayMap

985 experimental series

333 array platforms

253 ICD-O cancer entities

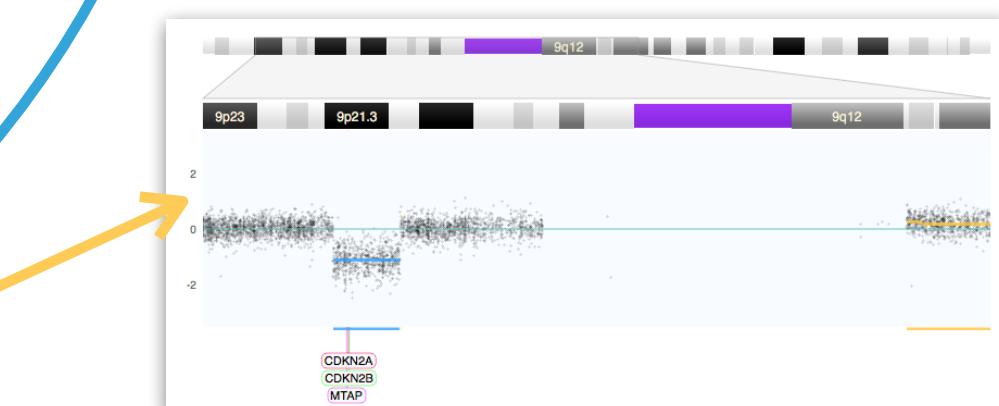
716 publications (PubMed entries)

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.

Platforms (1): GPR100, Agilent G1317P Lymphoma (Agilent Probe ID)

Supplementary file: GSE64034.RAW.tar



informa healthcare

ORIGINAL ARTICLE RESEARCH

Genomic imbalance defines three prognostic groups for risk stratification of patients with chronic lymphocytic leukemia

Jane Houldsworth<sup>1</sup>, Asha Guttapalli<sup>1</sup>, Venkata Thoduri<sup>1</sup>, Xiao Jie Yan<sup>1</sup>, Geeta Mendiratta<sup>1</sup>, Tamja Zelenka<sup>2</sup>, Gouri Nangisetty<sup>3</sup>, Wei Chen<sup>3</sup>, Supratik Pati<sup>3</sup>, Anthony Mato<sup>3</sup>, Jennifer R. Brown<sup>3</sup>, Kanti Rai<sup>4</sup>

<sup>1</sup>Cancer Genetics, Inc., Rutherford, NJ, USA; <sup>2</sup>Weinstein Institute of Medical Research, Manhattan, NY, USA; <sup>3</sup>Lymphoma Division, Department of Hematology and Oncology, Department of Medicine, Division of Hematology/Oncology, Department of Epidemiology and Biostatistics, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY, USA; <sup>4</sup>Department of Hematology and Oncology, David Hahn Cancer Institute and Department of Medicine, Harvard Medical School, Boston, MA, USA

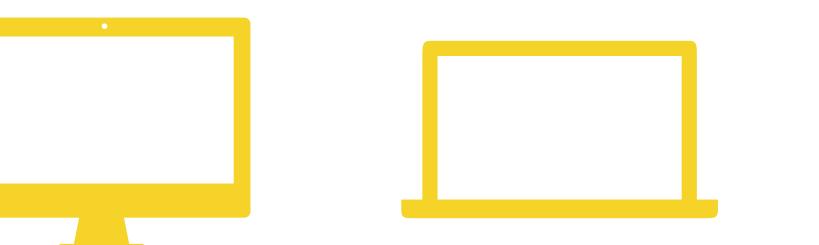
Abstract

Abstract: Several recent genomic hydridization (GCH) projects have been fully leveraged in a prognostic setting in chronic lymphocytic leukemia (CLL). In this study, we analyzed 20 CLL specimens using a targeted array. Based on 20 aberrations in each specimen, we identified a group with poor outcome (n = 12) and a group with good outcome (n = 8). We found that patients with poor outcome had a significantly higher number of gains and losses compared to the good outcome group. The presence of somatic genetic abnormalities by fluorescence in situ hybridization (FISH) was also associated with poor outcome. The presence of TP53 (17q11) and ATM (11q13) gain and loss, respectively, were associated with poor outcome. The presence of CDKN2A (1p13) and SP10 (17q12) intermediate outcomes. The presence of TP53BP1 (17q12) and ATM (11q13) loss and loss of heterozygosity (LOH) of 1q (q11-q12) and 17q loss were determined to be associated with intermediate outcome. The presence of TP53BP1 and SP10 mutations correlated with the presence of TP53 and ATM gains, respectively. These results suggest that when regions contain an allelic LOH, Patients requiring further investigation should be considered for clinical trials. These data support genomic imbalance evaluation in CLL by FISH and, more recently, massively parallel sequencing techniques. These findings may facilitate the identification of the CLL genes identifying genes, loci and other molecular mechanisms that are associated with CLL prognosis.

Keywords: chronic lymphocytic leukemia, molecular genetics, prognostication

Introduction

The clinical course of patients with B-cell chronic lymphocytic leukemia (CLL) is highly variable, undergoing disease progression, stabilization or regression. In CLL, disease progresses, when therapeutic intervention is required, and the clinical relevance of which is the most important factor in determining the best treatment.



arrayMap

progenetix

ArrayExpress

E-MTAB-998 Comparative genomic hybridization array of human peripheral T-cell lymphoma clinical samples to study their genomic aberration profiles

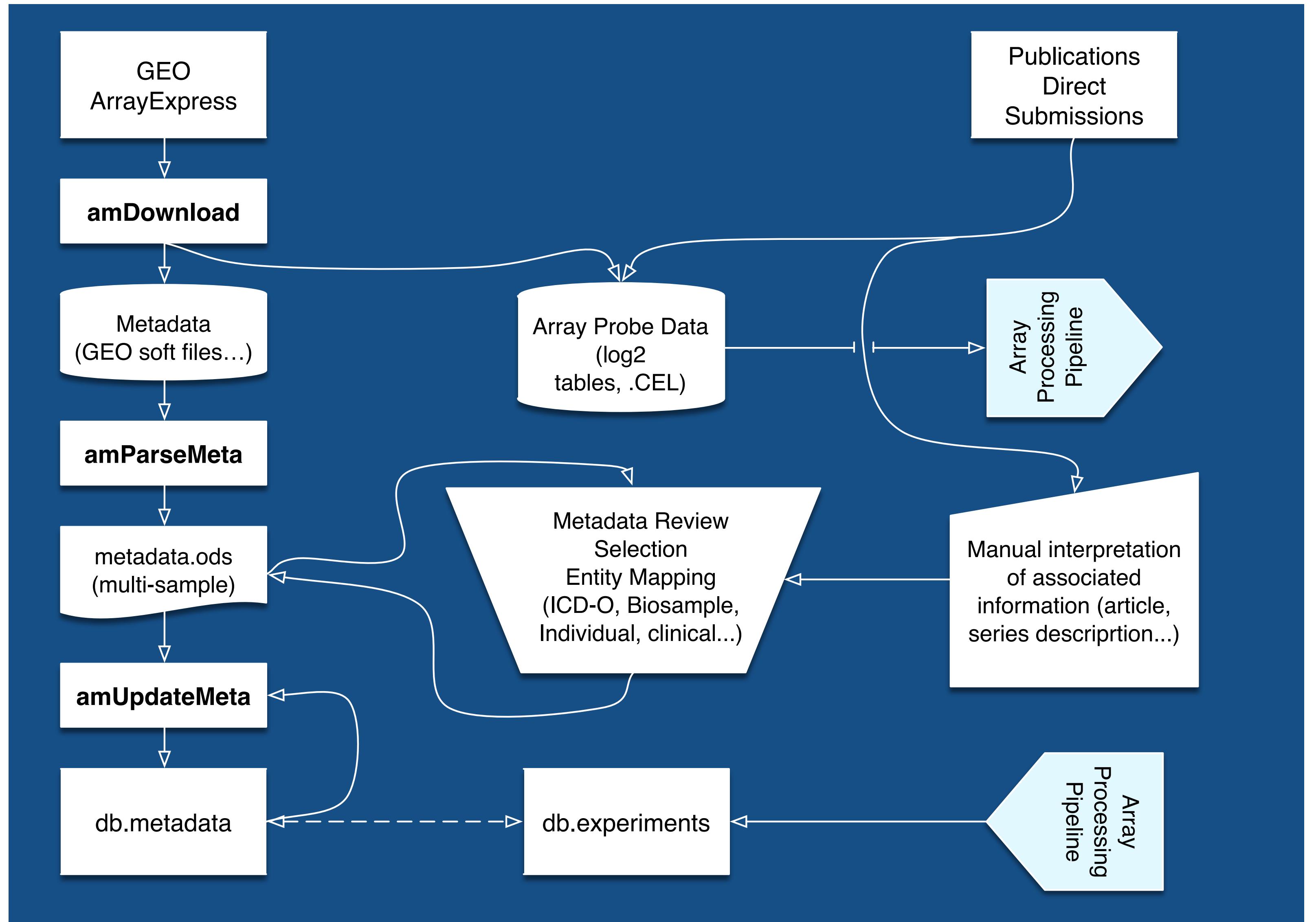
Organism: Homo sapiens

Sample: E-MTAB-998 Click for detailed sample information and links to data

Description: Genomic aberration profiles of Peripheral T-cell Lymphoma, not otherwise specified (clinical samples)

Experiment type: comparative genomic hybridization array, *a* vs *b*, *c* vs *d*, *e* vs *f*, *g* vs *h*, *i* vs *j*, *k* vs *l*, *m* vs *n*, *o* vs *p*, *q* vs *r*, *s* vs *t*, *u* vs *v*, *w* vs *x*, *y* vs *z*, *aa* vs *ab*, *ac* vs *ad*, *ae* vs *af*, *ag* vs *ah*, *ai* vs *aj*, *ak* vs *al*, *am* vs *an*, *ao* vs *ap*, *ar* vs *as*, *ar* vs *at*, *ar* vs *au*, *ar* vs *av*, *ar* vs *aw*, *ar* vs *ay*, *ar* vs *az*, *ar* vs *ba*, *ar* vs *bc*, *ar* vs *bd*, *ar* vs *be*, *ar* vs *bf*, *ar* vs *bg*, *ar* vs *bh*, *ar* vs *bi*, *ar* vs *bj*, *ar* vs *bk*, *ar* vs *bl*, *ar* vs *bm*, *ar* vs *bn*, *ar* vs *bo*, *ar* vs *bp*, *ar* vs *br*, *ar* vs *bs*, *ar* vs *bu*, *ar* vs *bv*, *ar* vs *bw*, *ar* vs *by*, *ar* vs *bz*, *ar* vs *ca*, *ar* vs *cb*, *ar* vs *cd*, *ar* vs *ce*, *ar* vs *cf*, *ar* vs *cg*, *ar* vs *ch*, *ar* vs *ci*, *ar* vs *ck*, *ar* vs *cl*, *ar* vs *cm*, *ar* vs *cn*, *ar* vs *co*, *ar* vs *cp*, *ar* vs *cr*, *ar* vs *cs*, *ar* vs *cu*, *ar* vs *cv*, *ar* vs *cw*, *ar* vs *cy*, *ar* vs *cz*, *ar* vs *da*, *ar* vs *db*, *ar* vs *dc*, *ar* vs *de*, *ar* vs *df*, *ar* vs *dg*, *ar* vs *dh*, *ar* vs *di*, *ar* vs *dk*, *ar* vs *dl*, *ar* vs *dm*, *ar* vs *dn*, *ar* vs *do*, *ar* vs *dp*, *ar* vs *dr*, *ar* vs *ds*, *ar* vs *du*, *ar* vs *dv*, *ar* vs *dw*, *ar* vs *dy*, *ar* vs *dz*, *ar* vs *ea*, *ar* vs *eb*, *ar* vs *ec*, *ar* vs *ed*, *ar* vs *ef*, *ar* vs *eg*, *ar* vs *eh*, *ar* vs *ei*, *ar* vs *ej*, *ar* vs *ek*, *ar* vs *el*, *ar* vs *em*, *ar* vs *en*, *ar* vs *eo*, *ar* vs *ep*, *ar* vs *er*, *ar* vs *es*, *ar* vs *eu*, *ar* vs *ev*, *ar* vs *ew*, *ar* vs *ey*, *ar* vs *ez*, *ar* vs *fa*, *ar* vs *fb*, *ar* vs *fc*, *ar* vs *fd*, *ar* vs *fe*, *ar* vs *fg*, *ar* vs *fh*, *ar* vs *fi*, *ar* vs *fk*, *ar* vs *fl*, *ar* vs *fm*, *ar* vs *fn*, *ar* vs *fo*, *ar* vs *fp*, *ar* vs *fr*, *ar* vs *fs*, *ar* vs *fu*, *ar* vs *fv*, *ar* vs *fw*, *ar* vs *fy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs *pp*, *ar* vs *qq*, *ar* vs *rr*, *ar* vs *ss*, *ar* vs *tt*, *ar* vs *uu*, *ar* vs *vv*, *ar* vs *ww*, *ar* vs *yy*, *ar* vs *zz*, *ar* vs *aa*, *ar* vs *bb*, *ar* vs *cc*, *ar* vs *dd*, *ar* vs *ee*, *ar* vs *ff*, *ar* vs *gg*, *ar* vs *hh*, *ar* vs *ii*, *ar* vs *jj*, *ar* vs *kk*, *ar* vs *ll*, *ar* vs *mm*, *ar* vs *nn*, *ar* vs *oo*, *ar* vs

# Bioinformatics & Data Curation - arrayMap data “Pipeline”



# Progenetix & arrayMap: Data Scopes

## Biomedical and procedural "Meta"data types

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



# Data sets in tutorials



# Data sets in the wild



# Data Curation - Happy RegExing!

## Extracting clinical and technical metadata from GEO SOFT file

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

# 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 Grooteplein 10
!Sample_contact_city = Nijmegen
!Sample_contact_zip/postal_code = 6525GA
!Sample_contact_country = Netherlands
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CEL.gz
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CHP.gz
!Sample_series_id = GSE7255
```

```
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...

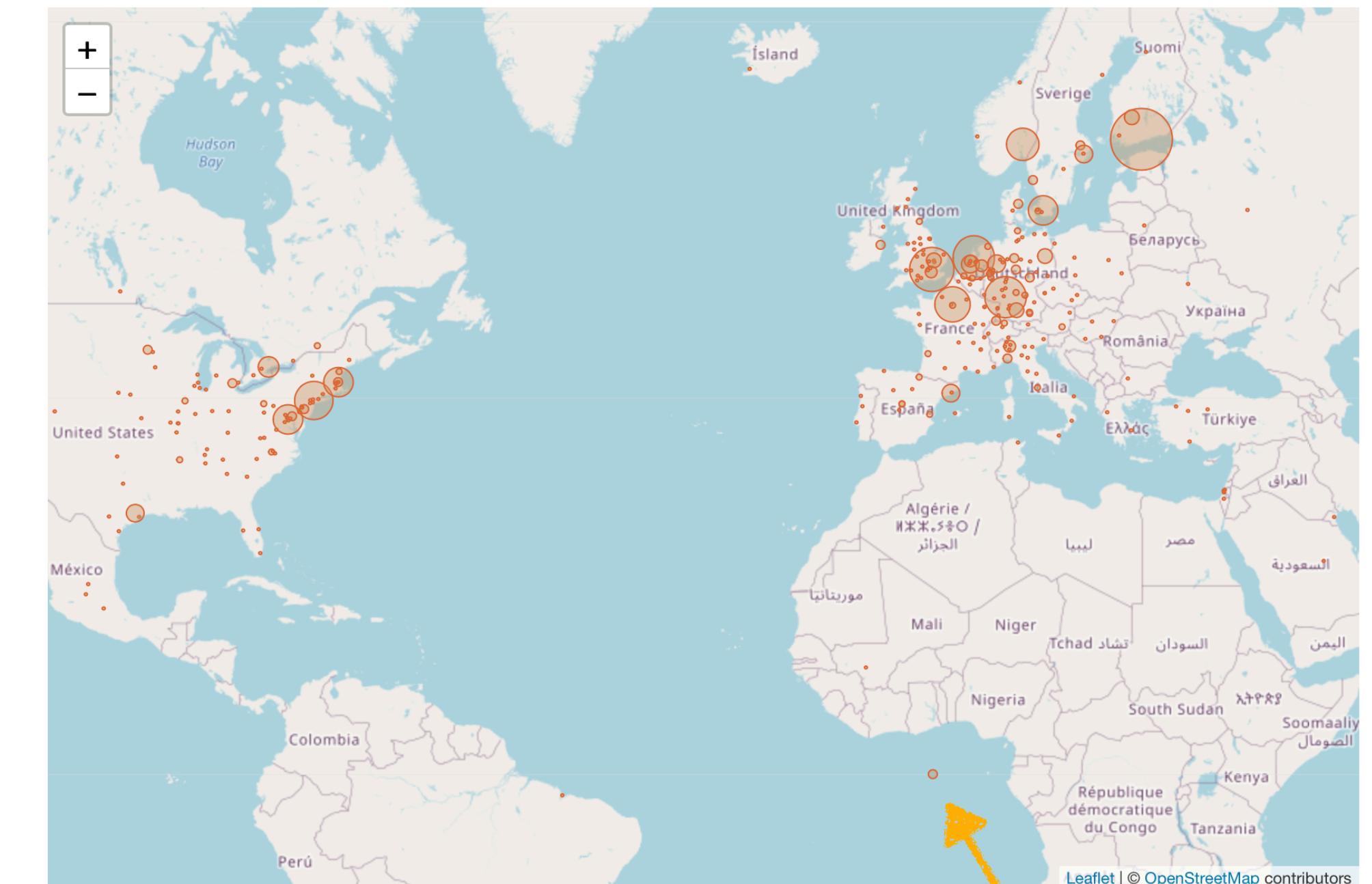
Unknown way to express "alive"!

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

# Data Curation

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

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



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

Progenetix publication collection

2020-11-28

25 / 3306  
publications

# Data Curation - Geolocations

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

The most geo-tagged place on earth is Null Island

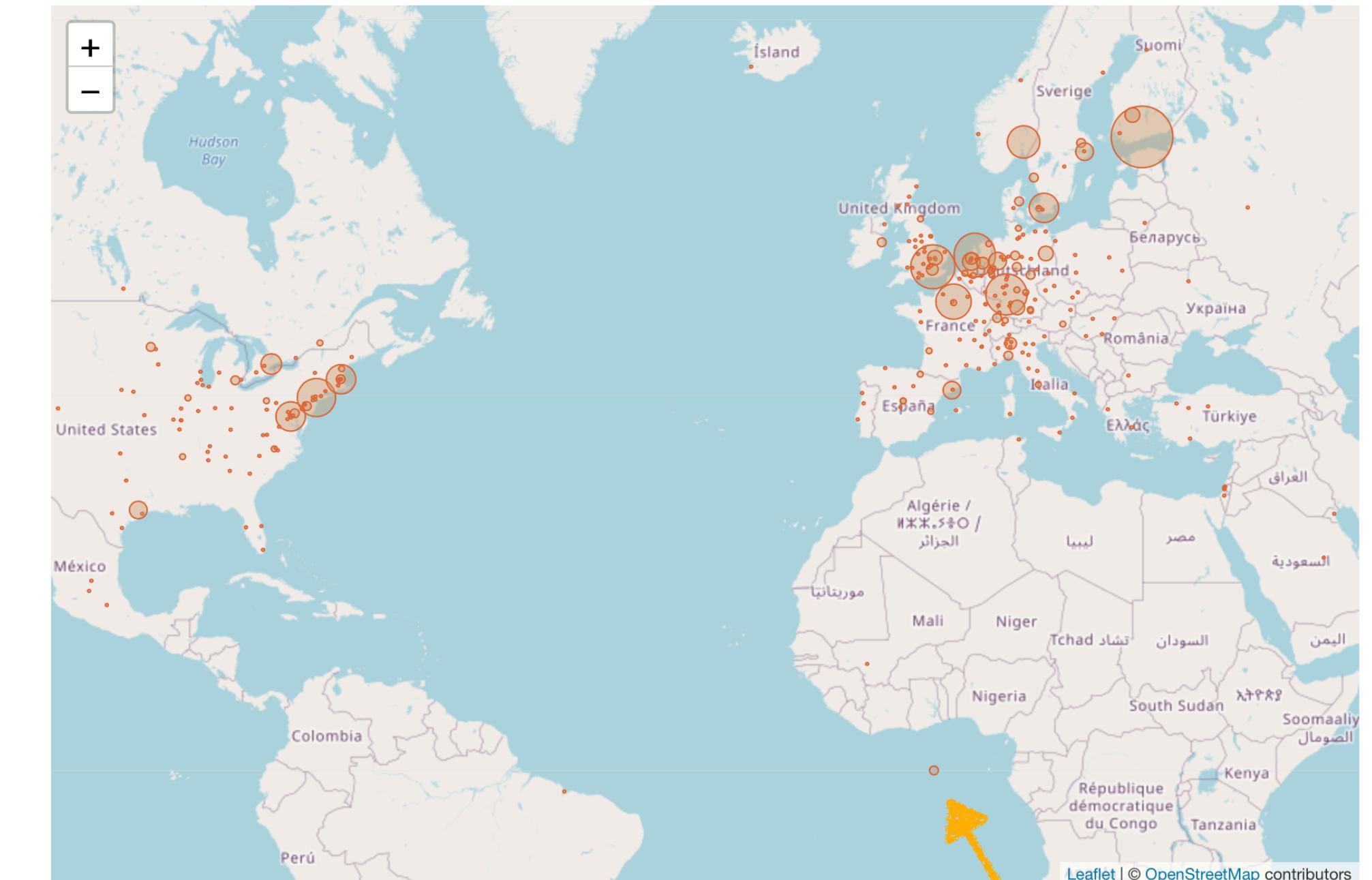


A troubleshooting country has been added with an Indeterminate sovereignty class called **Null Island** ([1](#), [2](#)). It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging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

[https://en.wikipedia.org/wiki/Null\\_Island](https://en.wikipedia.org/wiki/Null_Island)

Michael Szell: The Data Science Process 2

2020-11-25



Progenetix publication collection

2020-11-28

25 / 3306 publications

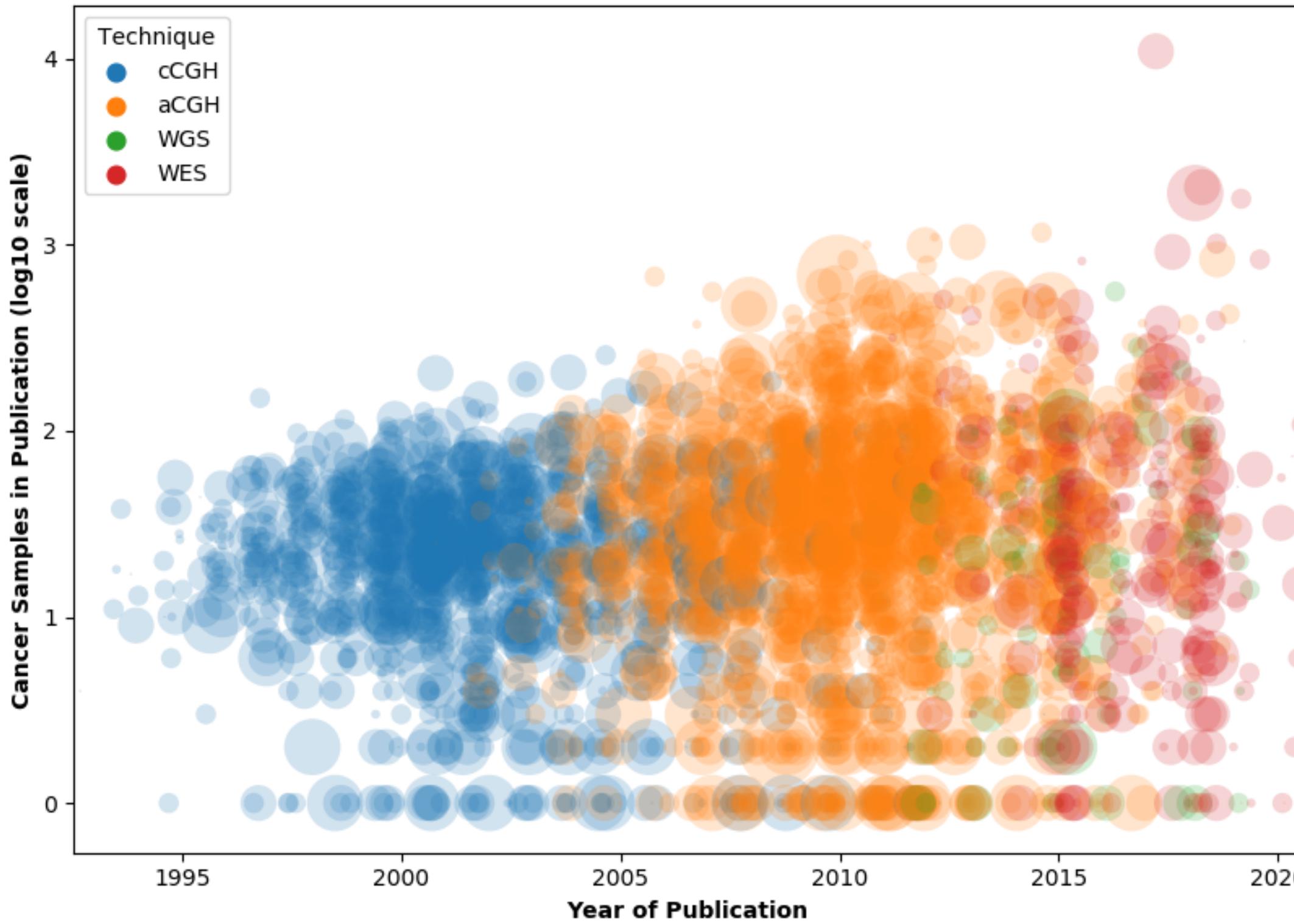
# Standardized Data

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

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

```
"data_use_conditions" : {  
    "label" : "no restriction",  
    "id" : "DUO:0000004"  
},  
  
"provenance" : {  
    "material" : {  
        "type" : {  
            "id" : "EFO:0009656",  
            "label" : "neoplastic sample"  
        }  
    },  
    "geo" : {  
        "label" : "Zurich, Switzerland",  
        "precision" : "city",  
        "city" : "Zurich",  
        "country" : "Switzerland",  
        "latitude" : 47.37,  
        "longitude" : 8.55,  
        "geojson" : {  
            "type" : "Point",  
            "coordinates" : [  
                8.55,  
                47.37  
            ]  
        },  
        "IS0-3166-alpha3" : "CHE"  
    },  
    {  
        "age": "P25Y3M2D"  
    }  
}
```

### Number of tumor samples for each publication across the years



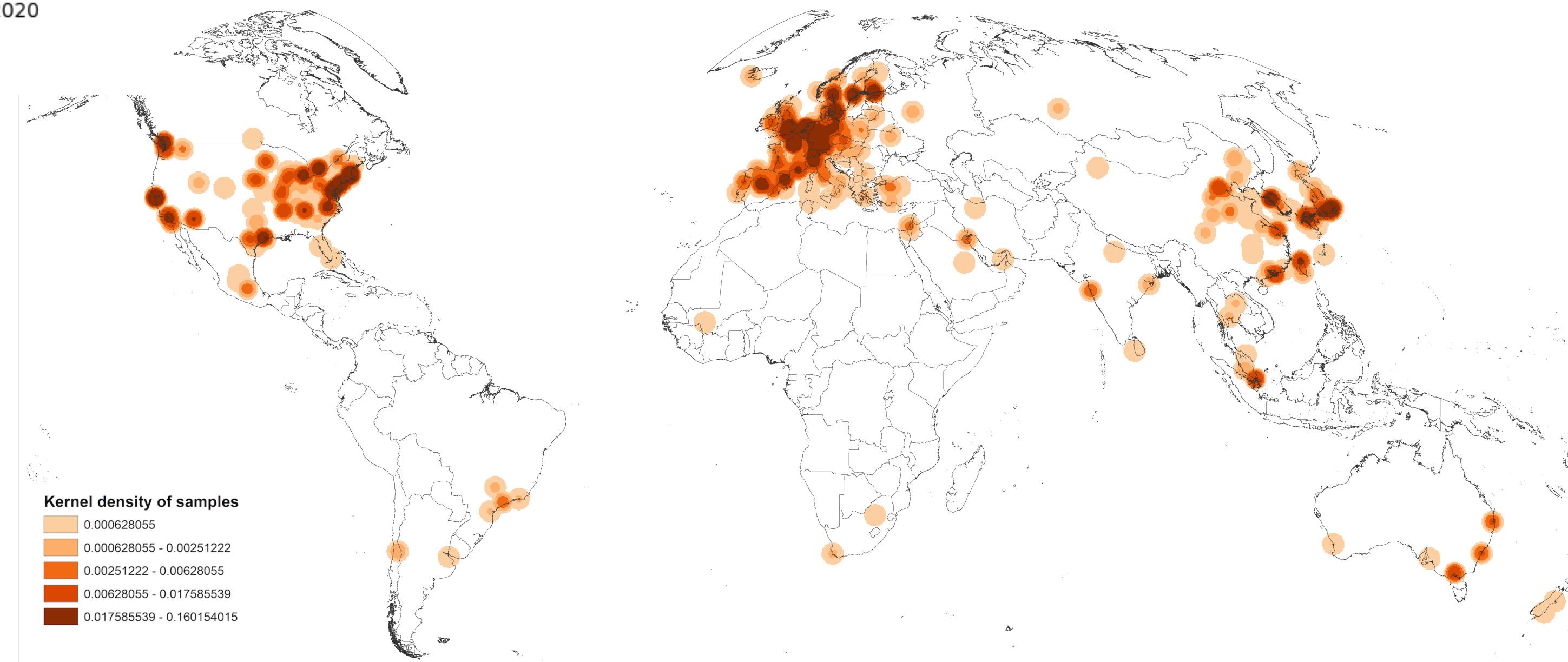
Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets.

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

## Publication Landscape of Cancer CNV Profiling

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

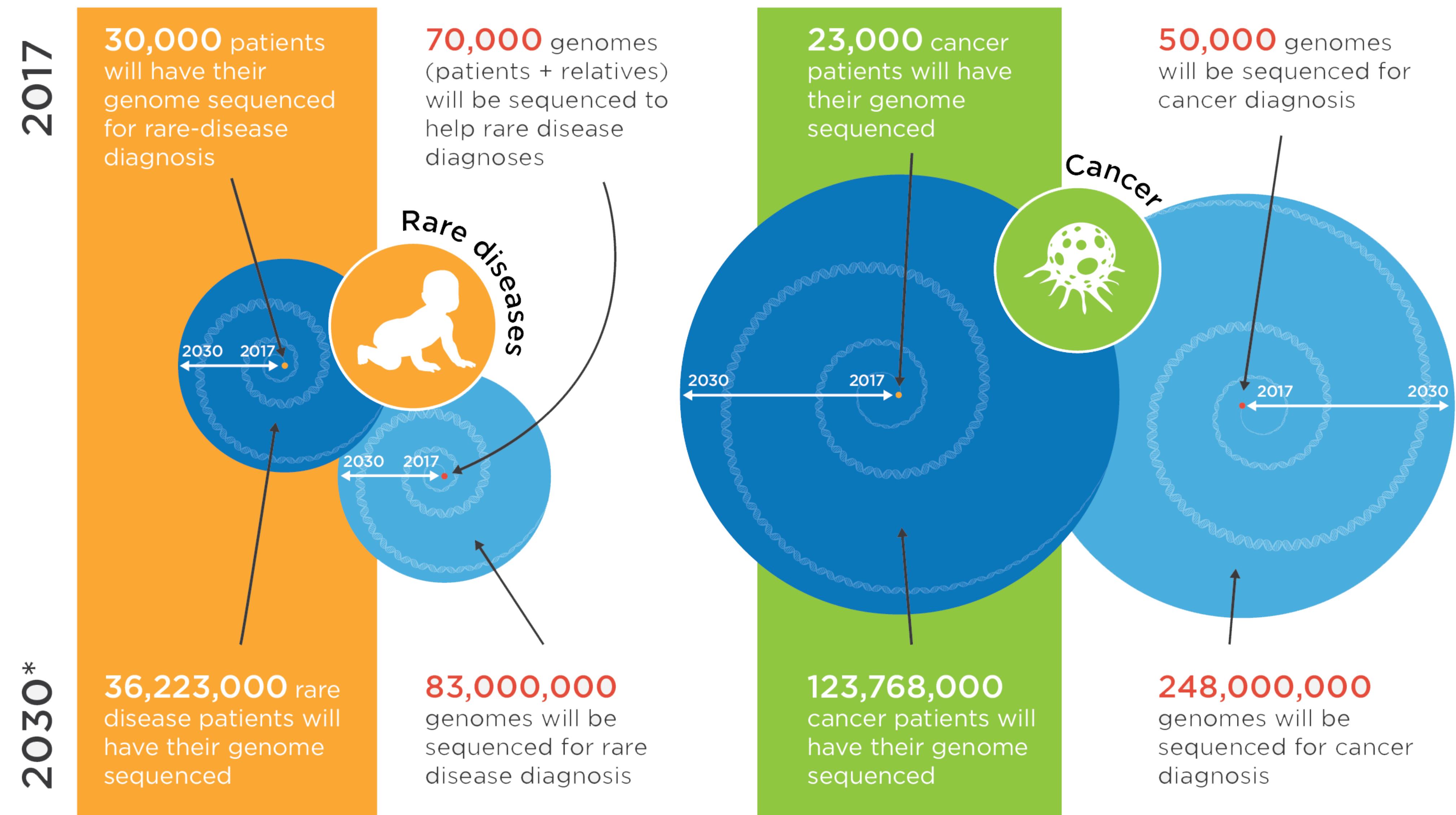
For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.





# Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.



# The vision: Federation of data



# The Global Alliance for Genomics and Health

## Making genomic data accessible for research and health

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

GENOMICS

*A federated ecosystem for sharing genomic, clinical data*

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

The Global Alliance for Genomics and Health\*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291



Global Alliance  
for Genomics & Health

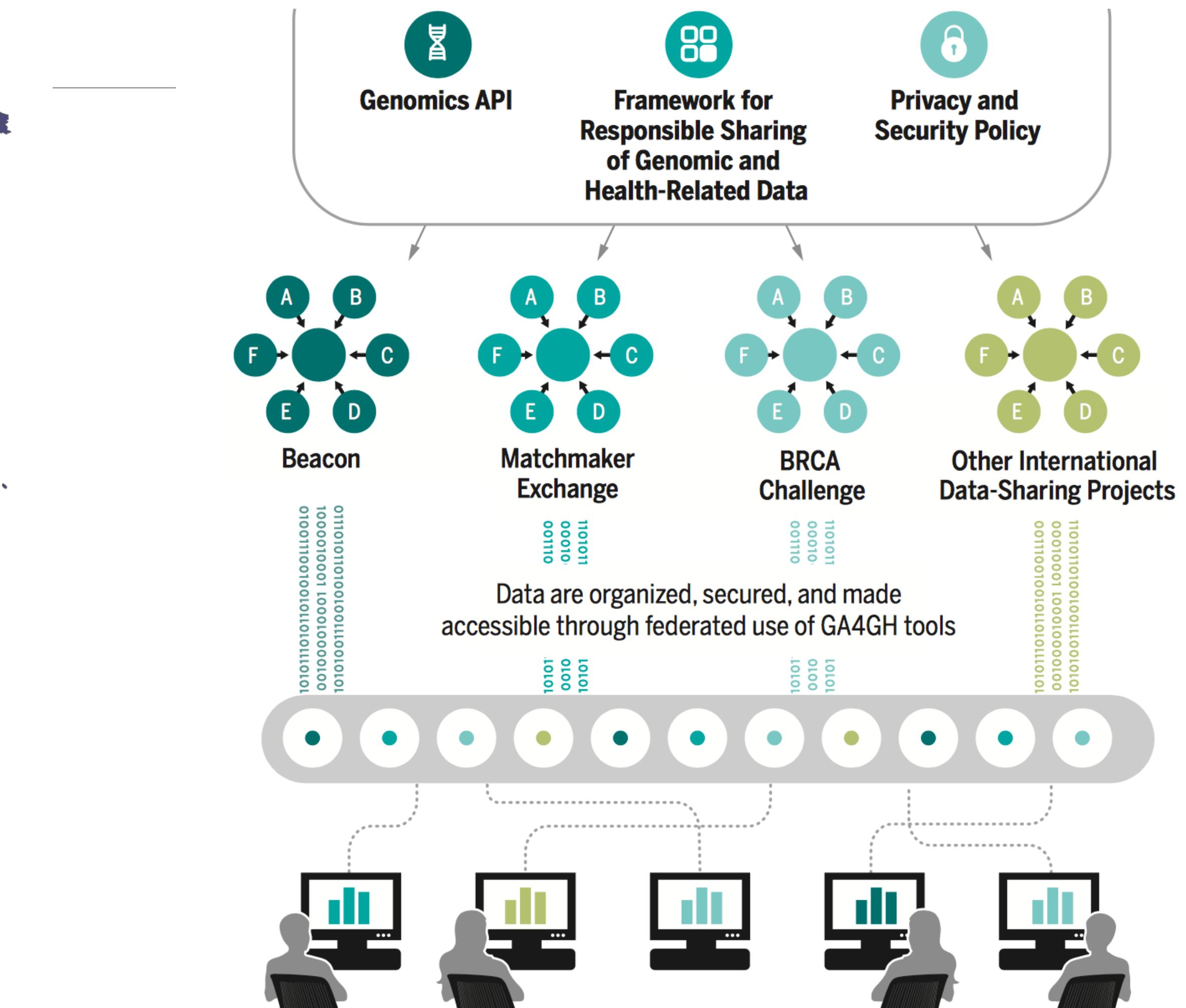


## GENOMICS

# *A federated ecosystem for sharing genomic, clinical data*

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

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





GENOMICS

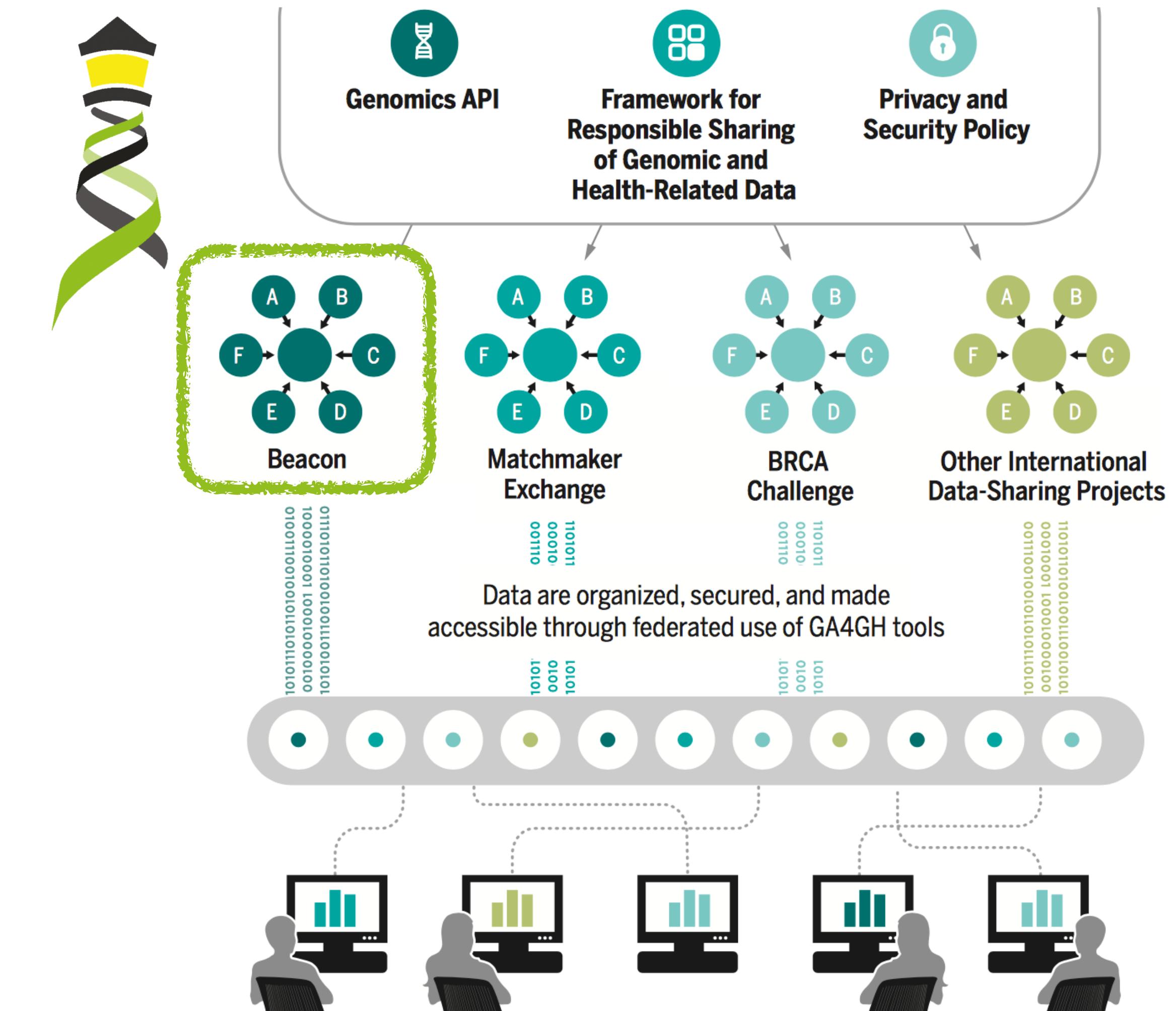
# *A federated ecosystem for sharing genomic, clinical data*

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

## **The Global Alliance for Genomics and Health\***

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 62

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



# DNASTACK



Global Alliance  
for Genomics & Health

## Introduction

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

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

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

## Utility

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

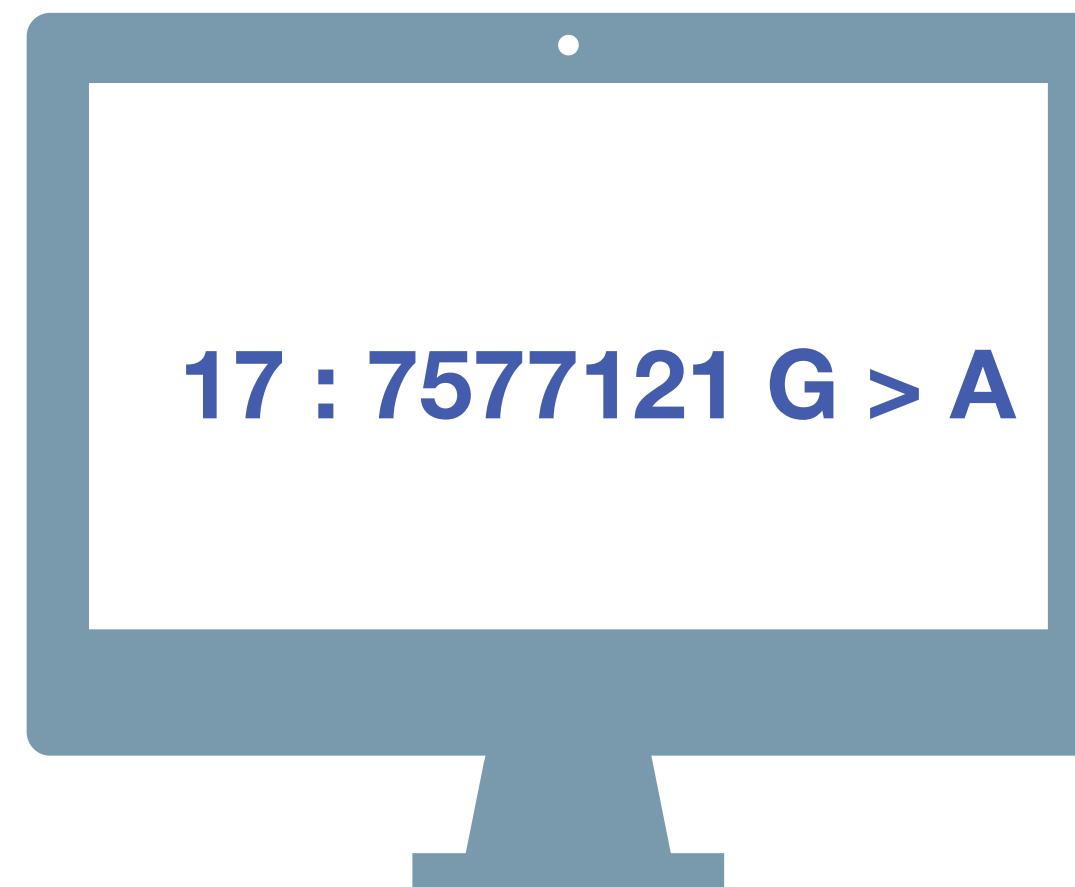
A number of more useful first versions have been suggested.

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

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

## Implementation

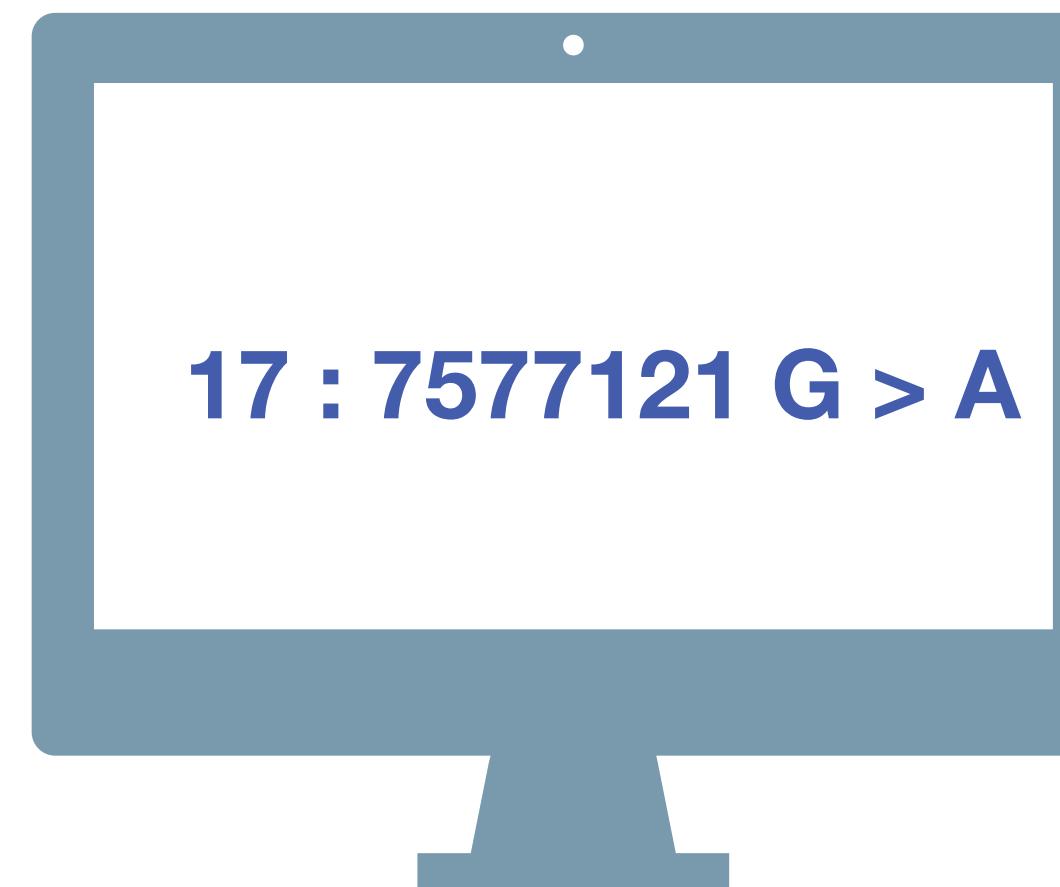
1. Specifying the chromosome ... The interface needs to specify the **accession.version** of a chromosome, or **build number**...
2. Return values ... right to **refuse** to answer without it being an error ... DOS **attack** ... or because ...especially **sensitive**...
3. Real time response ... Some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



# Beacon

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

**YES | NO | \0**



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

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

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

# Beacon Project in 2016

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



Beacon Network

Search Beacons

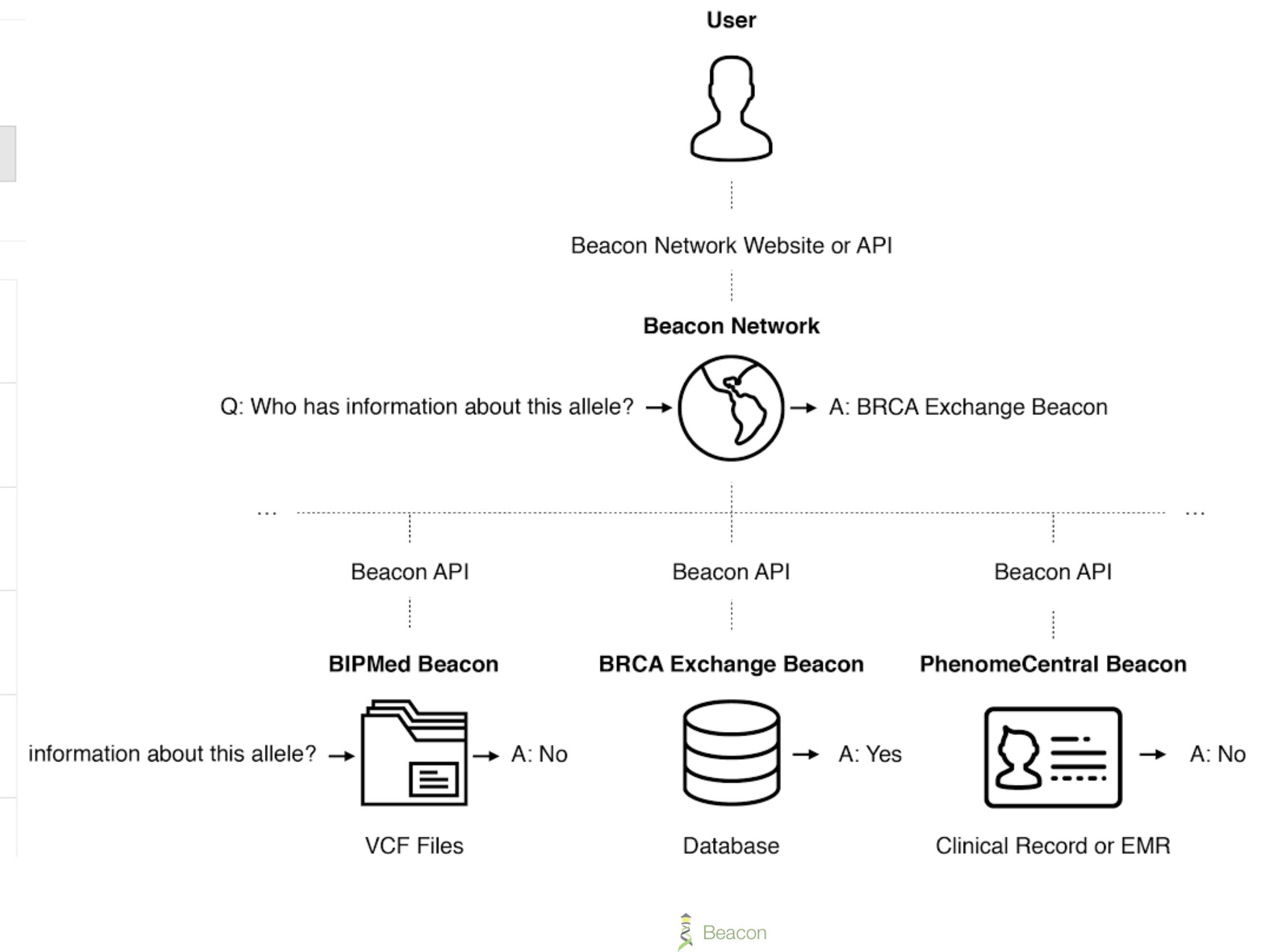
Search all beacons for allele

GRCh37 ▾ 10:118969015 C / CT Search

Response All None  
Found 16  
Not Found 27  
Not Applicable 22

Organization All None  
AMPLab, UC Berkeley  
BGI  
BioReference Laboratories  
Brazilian Initiative on ...  
BRCA Exchange  
Broad Institute  
Centre for Genomic R...  
Centro Nacional de A...  
Curoverse  
EMBL European Bio...  
Global Alliance for G...  
Google  
Institute for Systems ...  
Instituto Nacional de ...

Organization	Allele	Response
BioReference	10:118969015 C / CT	Found
Catalogue of Somatic Mutations in Cancer	10:118969015 C / CT	Found
Cell Lines	10:118969015 C / CT	Found
Conglomerate	10:118969015 C / CT	Found
COSMIC	10:118969015 C / CT	Found
dbGaP: Combined GRU Catalog and NHLBI Exome Seq...	10:118969015 C / CT	Found



35+

Organizations

90+

Beacons

200+

Datasets

100K+

In

Releases

Date	Tag	Title
2018-01-24	v0.4.0	Beacon
2016-05-31	v0.3.0	Beacon

# ELIXIR - Making Beacons Biomedical



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

# Beacon v2 - Clinical Beacon requirements

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

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

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

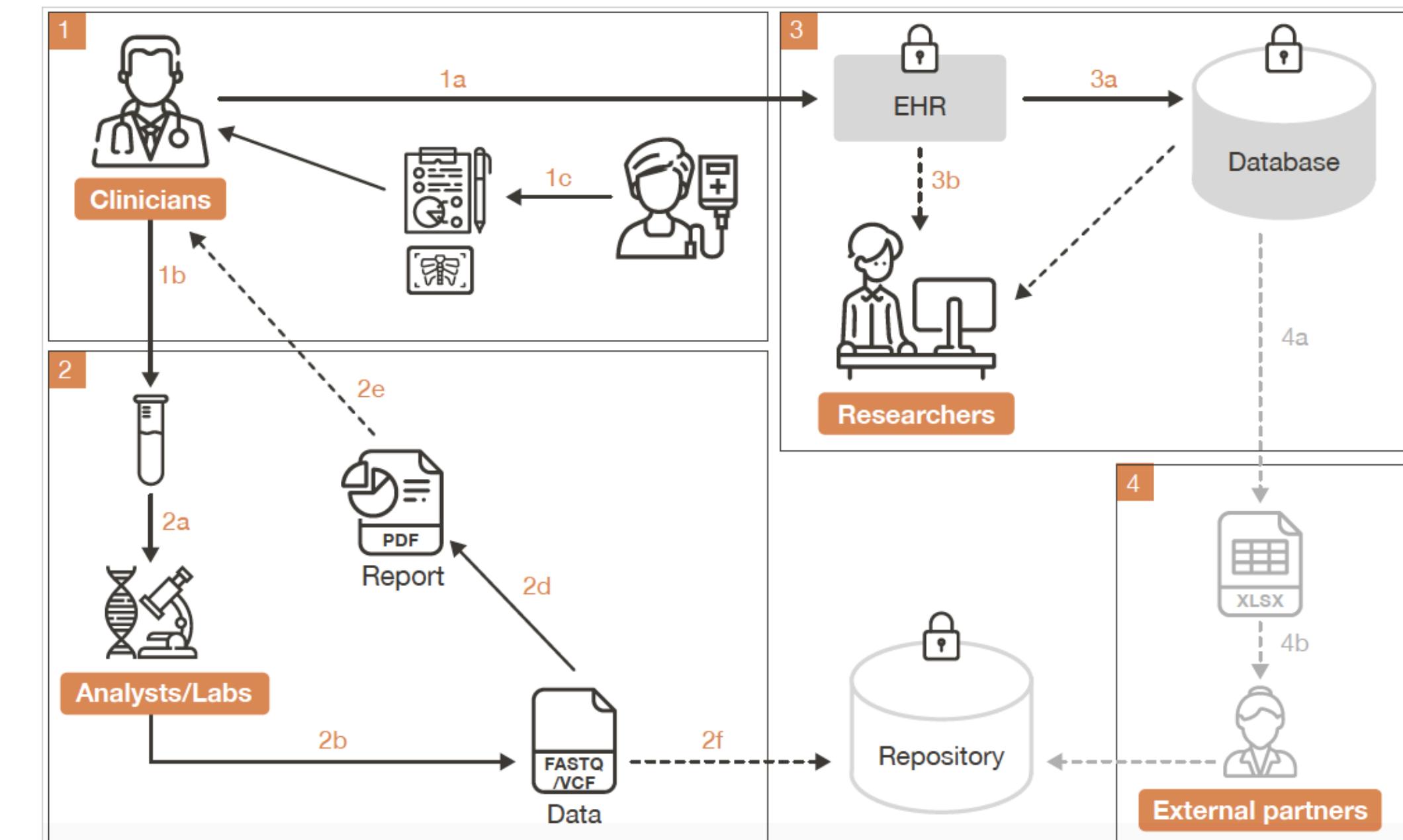
- Allowing more informative queries, like filtering by gender or age
- Allowing to trigger the next step in the data access process, e.g. who to contact or which are the data use conditions
- Jumping to another system where the data could be accessed, e.g. if the Beacon is internal to a hospital, to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

## The process

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

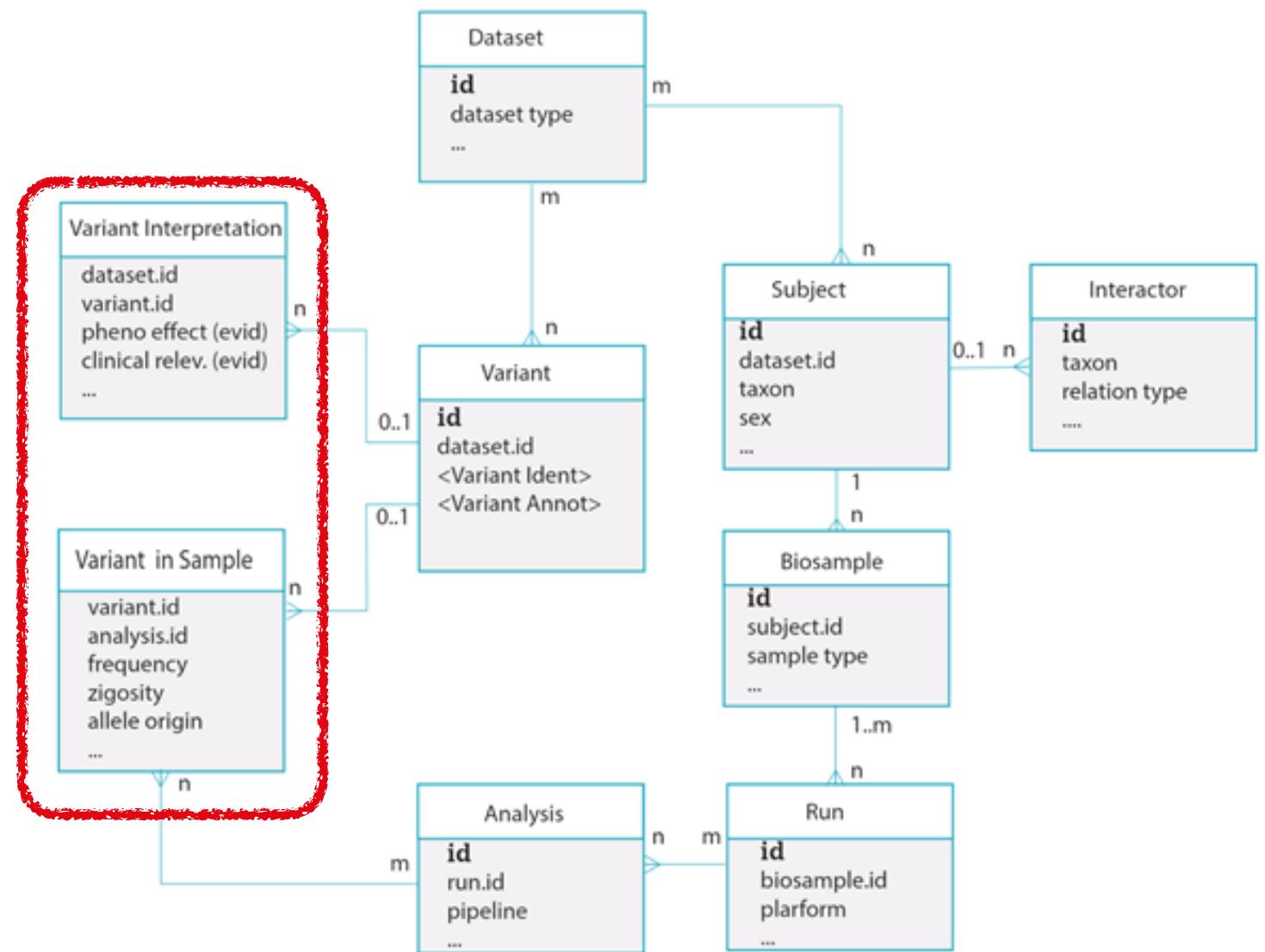
Interviews were conducted with the following GA4GH Driver Projects:

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



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

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



# ELIXIR Beacon Network



- developed under lead from ELIXIR Finland
- **authenticated access** w/ ELIXIR AAI
- **incremental extension**, starting with ELIXIR Beacon resources adhering to the **latest specification** (contrast to legacy networks)
- service details provided by individual Beacons, using **GA4GH service-info**
- **registration service**
  - integrator** throughout ELIXIR Human Data
  - starting point for "**beyond ELIXIR**" **feature rich** federated Beacon services

GRCh38 ▾ 17 : 7577121 G > A

[Example variant query](#) [Advanced Search](#)

baudisgroup at UZH and SIB  
Progenetix Cancer Genomics Beacon+

Beacon+ provides a forward looking implementation of the Beacon API, with focus on structural variants and metadata based on the cancer and reference genome profiling data represented in the Progenetix oncogenomic data resource (<https://progenetix.org>).

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

National Bioinformatics Infrastructure Sweden  
SweFreq Beacon

Beacon API Web Server based on the GA4GH Beacon API

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

LCSB at University of Luxembourg  
ELIXIR.LU Beacon

ELIXIR.LU Beacon

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

Research Programme on Biomedical Informatics  
DisGeNET Beacon

Variant-Disease associations collected from curated resources and the literature

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

European Genome-Phenome Archive (EGA)  
EGA Beacon

This [Beacon](https://beacon-project.io/) is based on the GA4GH Beacon [v1.1.0](https://github.com/ga4gh/beacon/specification/blob/develop/beacon.yaml)

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

University of Tartu Institute of Genomics, Estonia  
Beacon at the University of Tartu, Estonia

Beacon API Web Server based on the GA4GH Beacon API

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)

CSC - IT Center for Science Production Beacon

Beacon API Web Server based on the GA4GH Beacon API

[Visit Us](#) · [Beacon API](#) · [Contact Us](#)



# Beacon Project - Partner Engagement & Next Steps

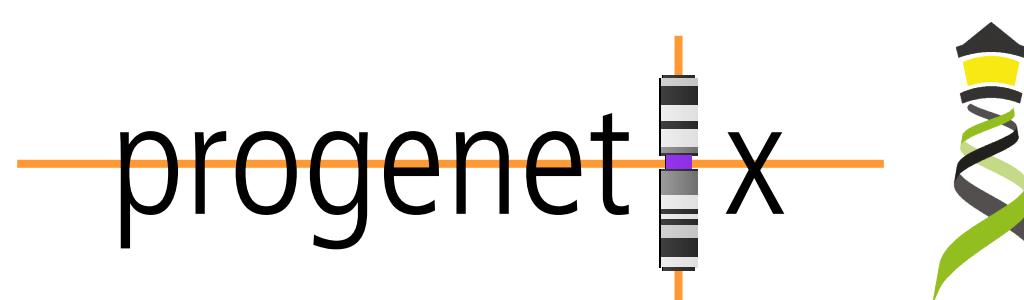
- Working with **partner communities & projects** on *deploying Beacons*
    - ELIXIR hCNV Community
    - European Joint Program on Rare Diseases
    - clinical groups & data initiatives (e.g. Andalucia, Cancer Core Europe, SPHN)
    - variant annotation resources, with optional clinical components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



# Beacon+ by Progenetix

## From Beacon Query to Explorative Analyses of CNV Patterns

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



**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**

**Genome Assembly** GRCh38 / hg38

**Gene Symbol**

**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**

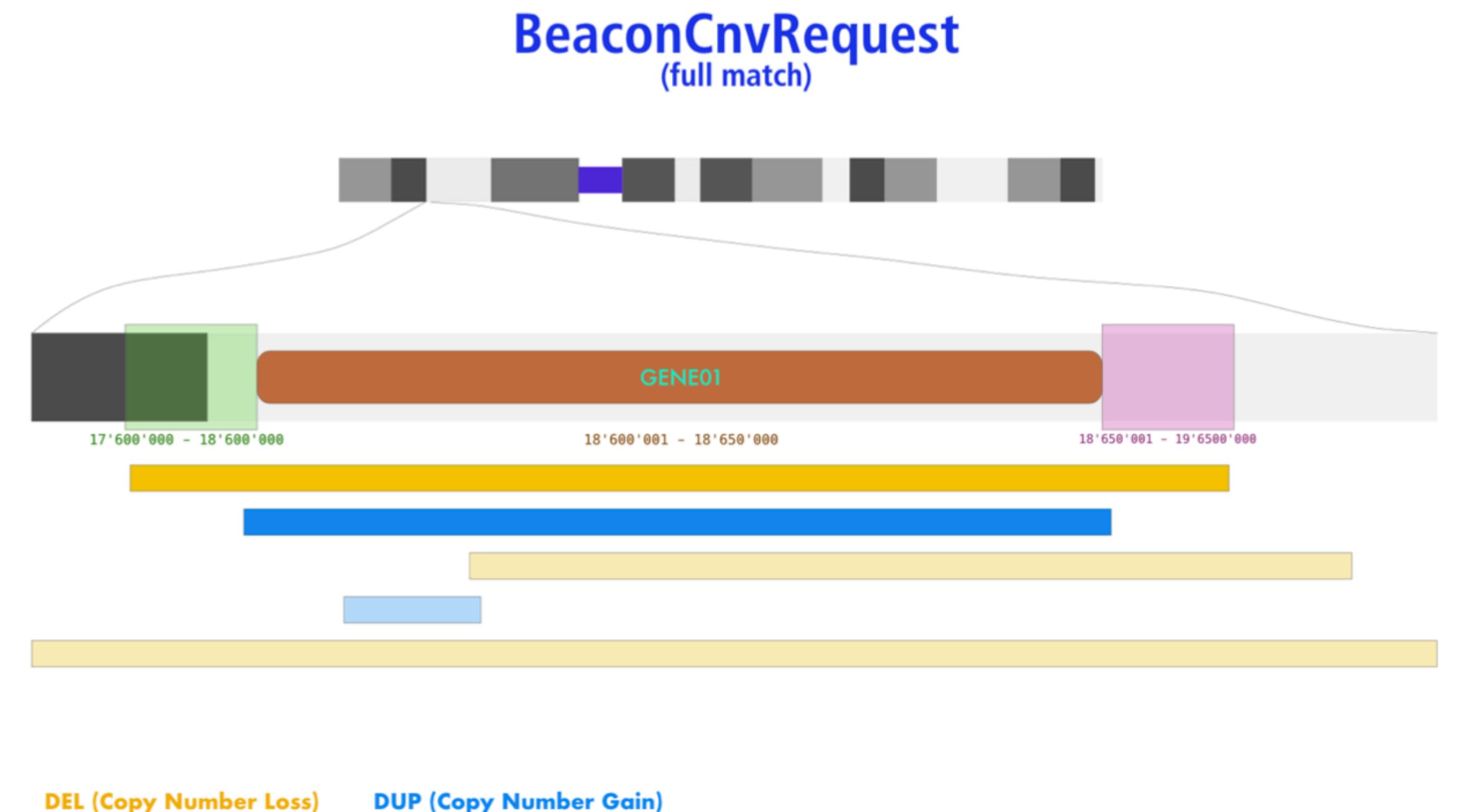
**Cancer Classification(s)**

**Filters**

**City**

**Query Database**

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

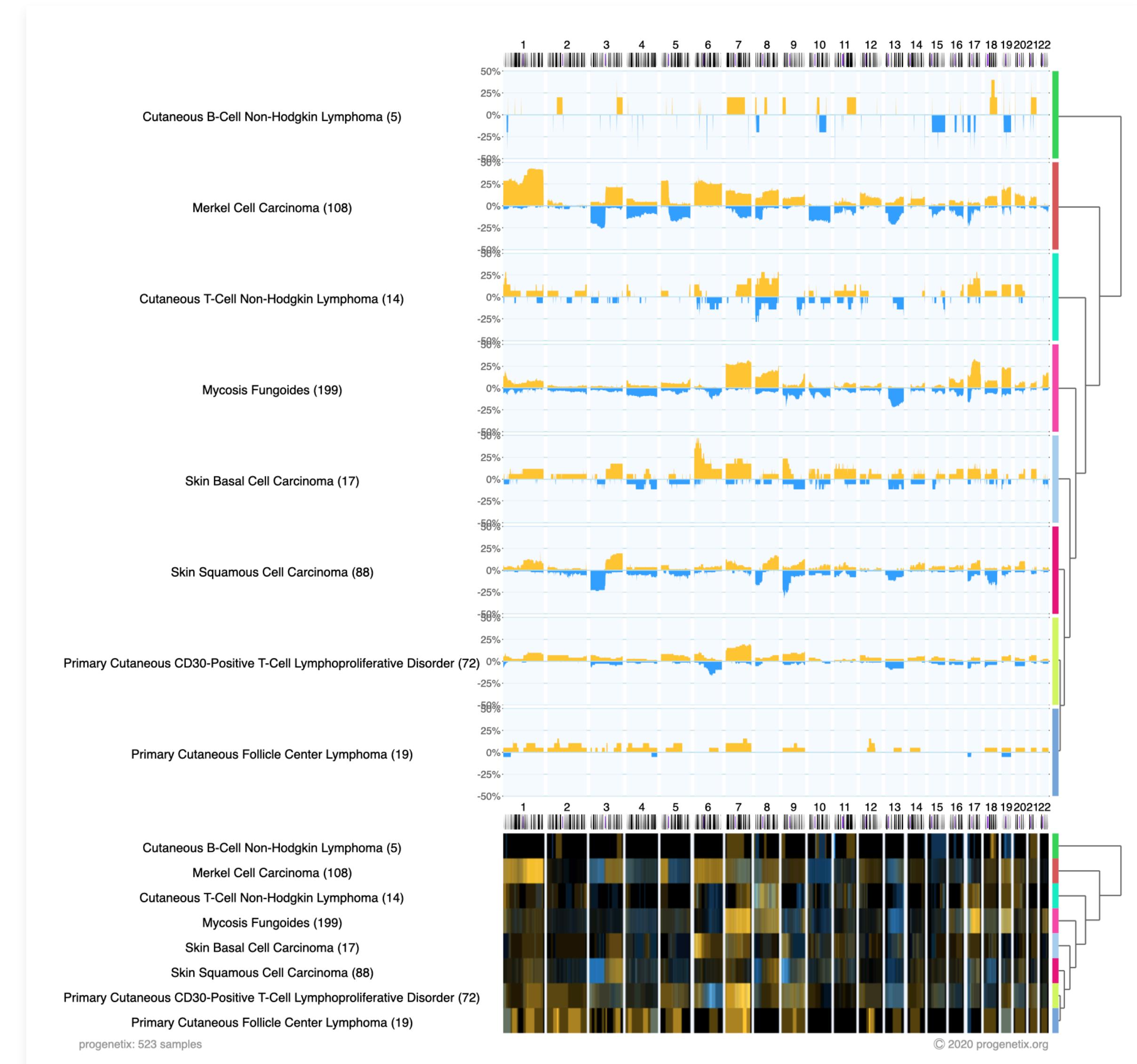
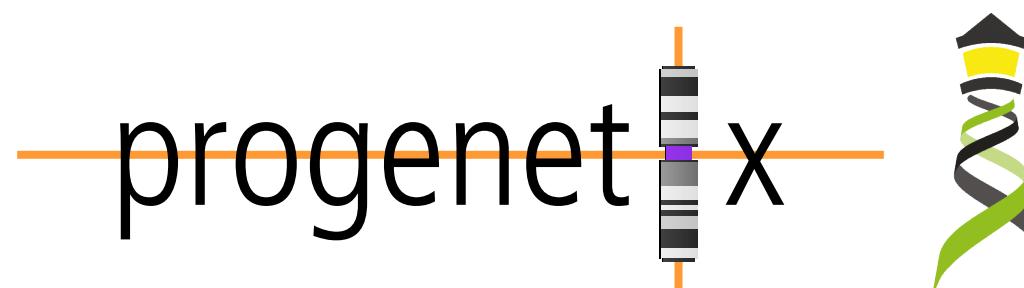


A “full match” BeaconCnvRequest is a typical scenario for e.g. matching CNVs in which the whole CDR of a gene has been duplicated. Here, both start and end search intervals lie outside of the region of interest. The maximum size of matched CNVs can be limited through the extend of the outer bounds (`start[0]`,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

# Beacon+ by Progenetix

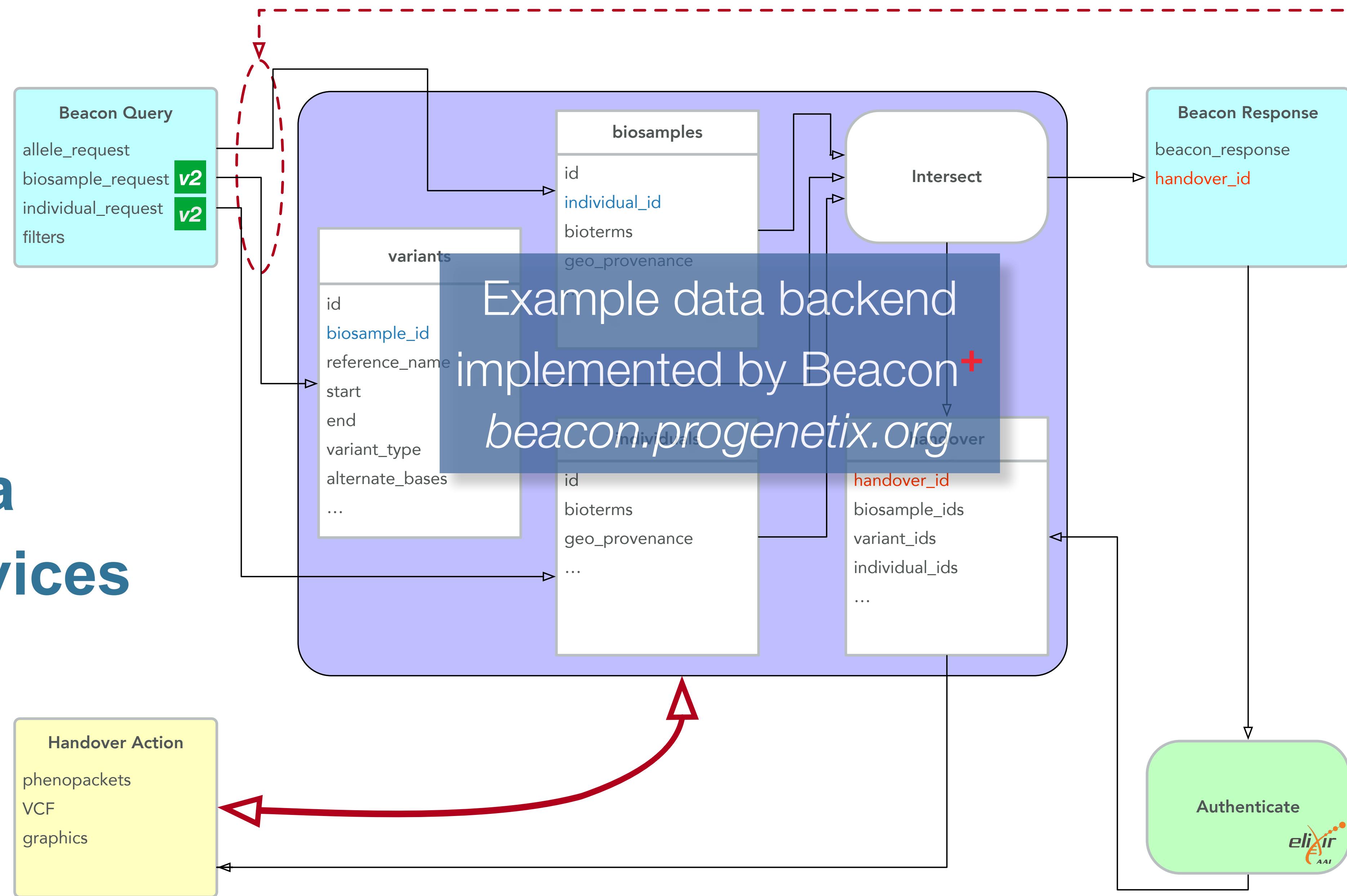
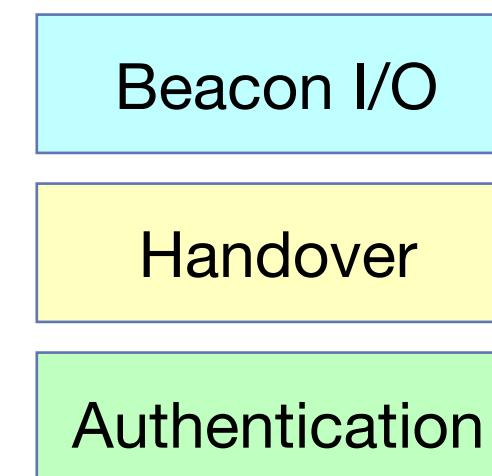
## From Beacon Query to Explorative Analyses of CNV Patterns

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



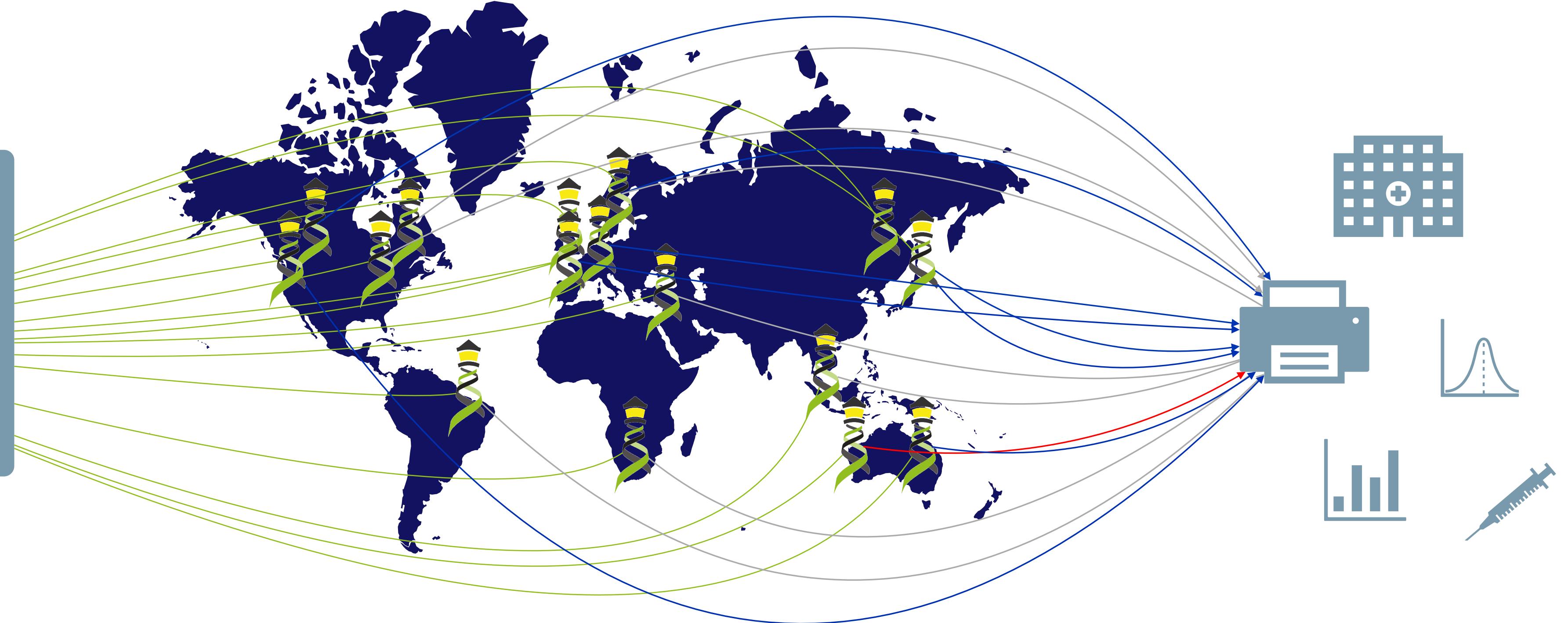
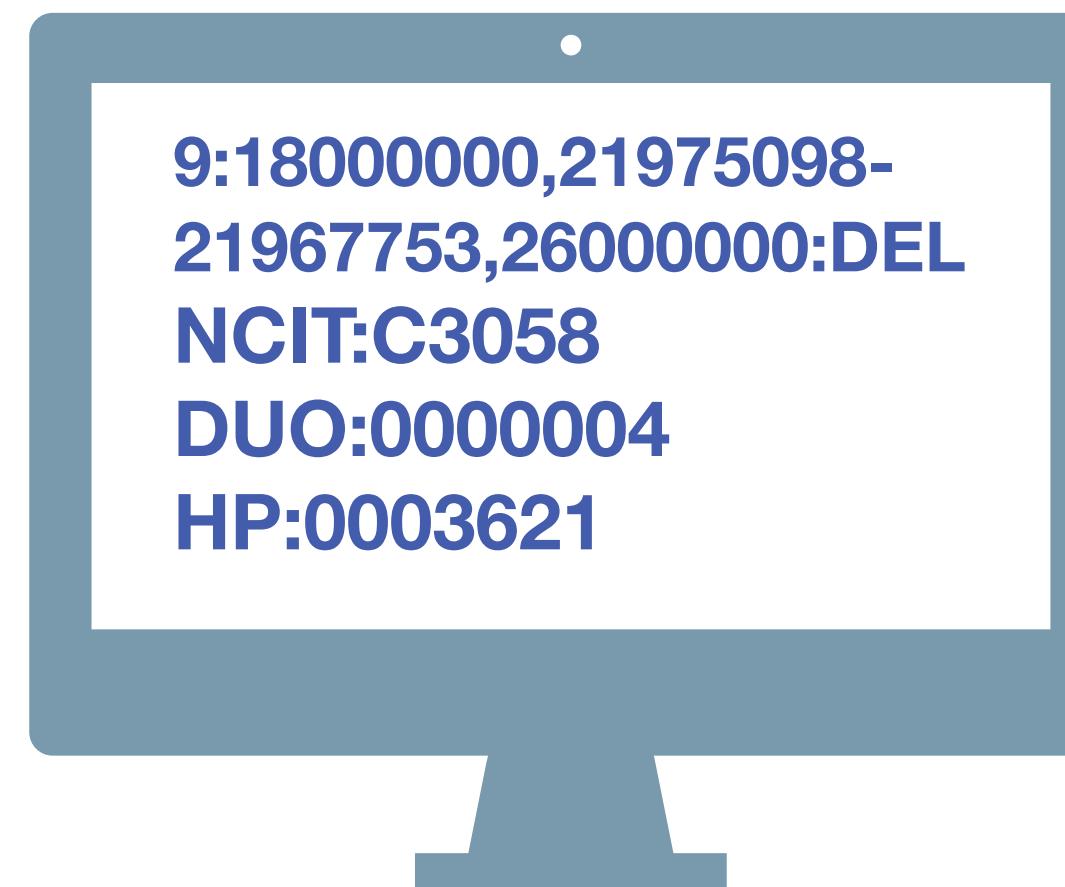
# Beacon & Handover

Beacons v1.1  
supports data  
delivery services



# Beacon & GA4GH

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

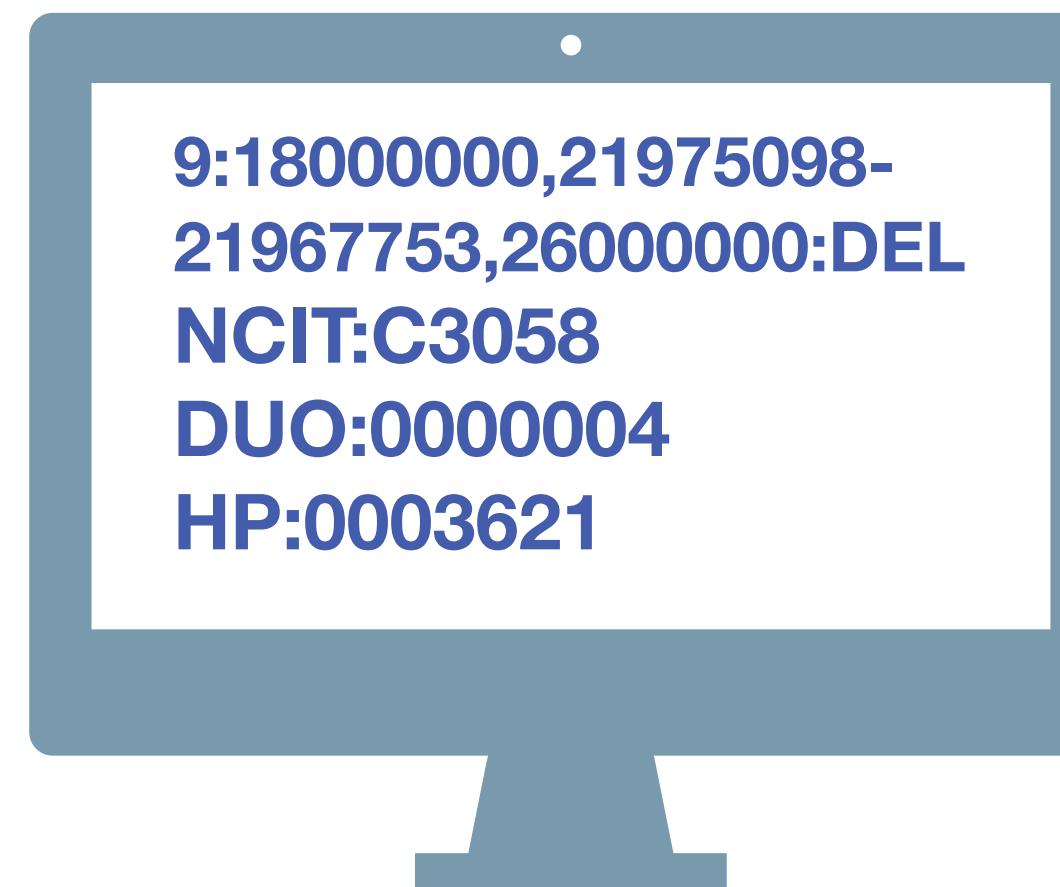


Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

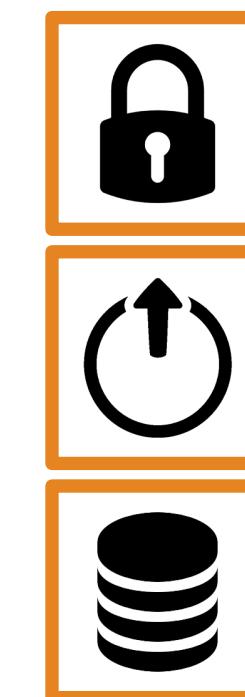


## Beacon v2 API

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



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

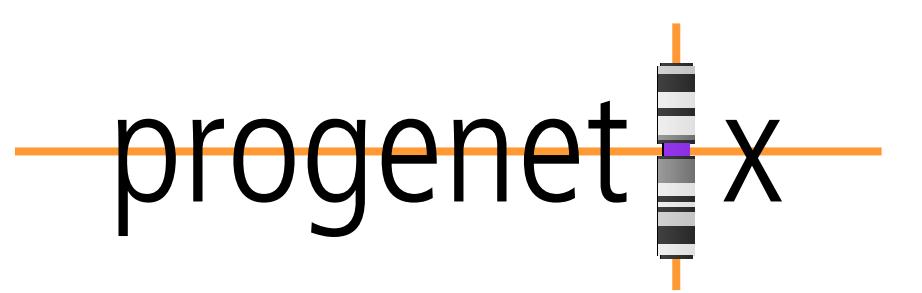


## Beacon v2 API

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

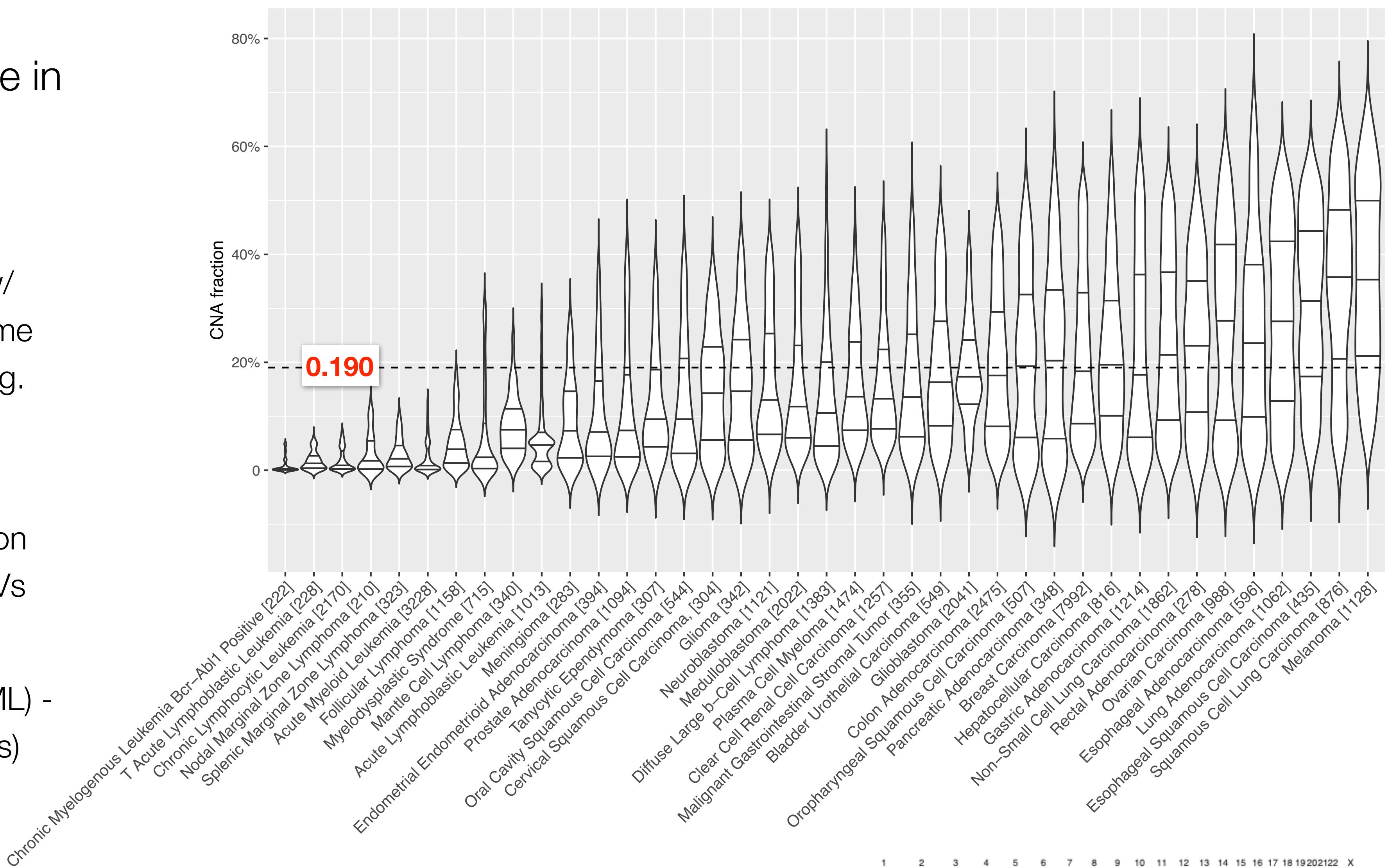


# Progenetix Data Use Cases

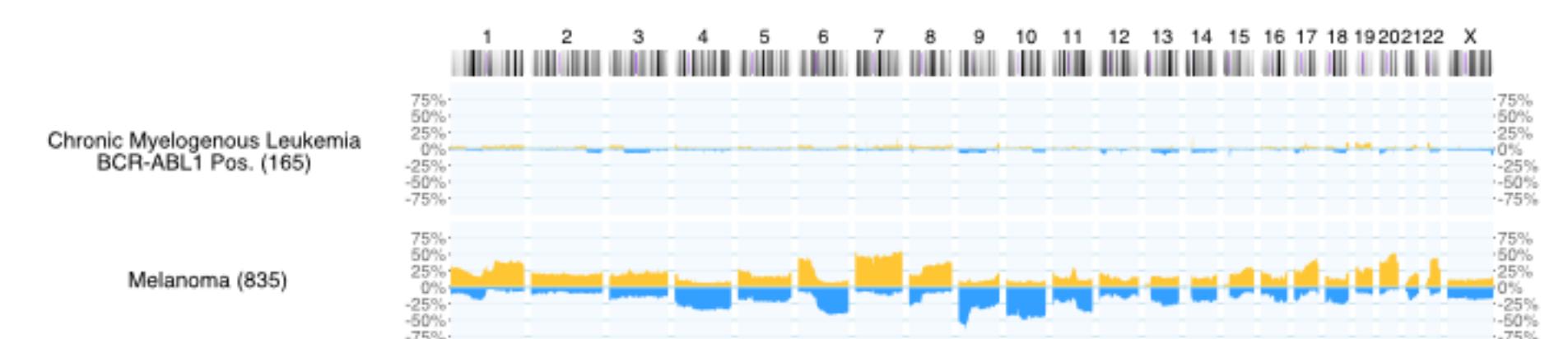


# Genome CNV coverage in Cancer Classes

- 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)



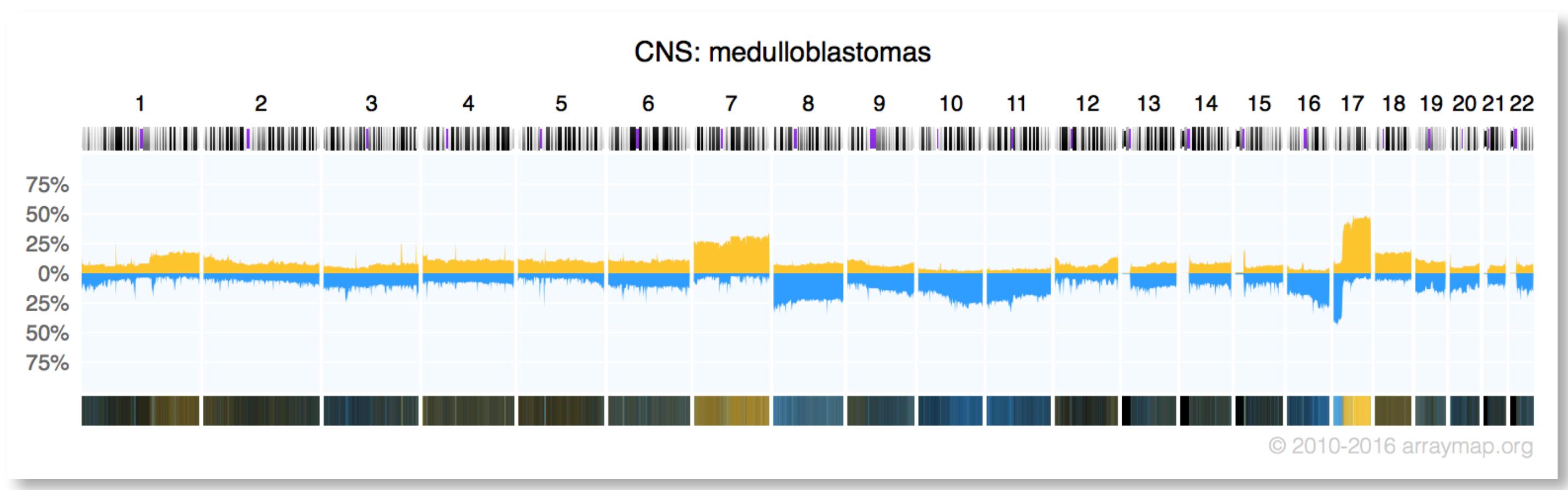
Lowest / Highest CNV fractions =>



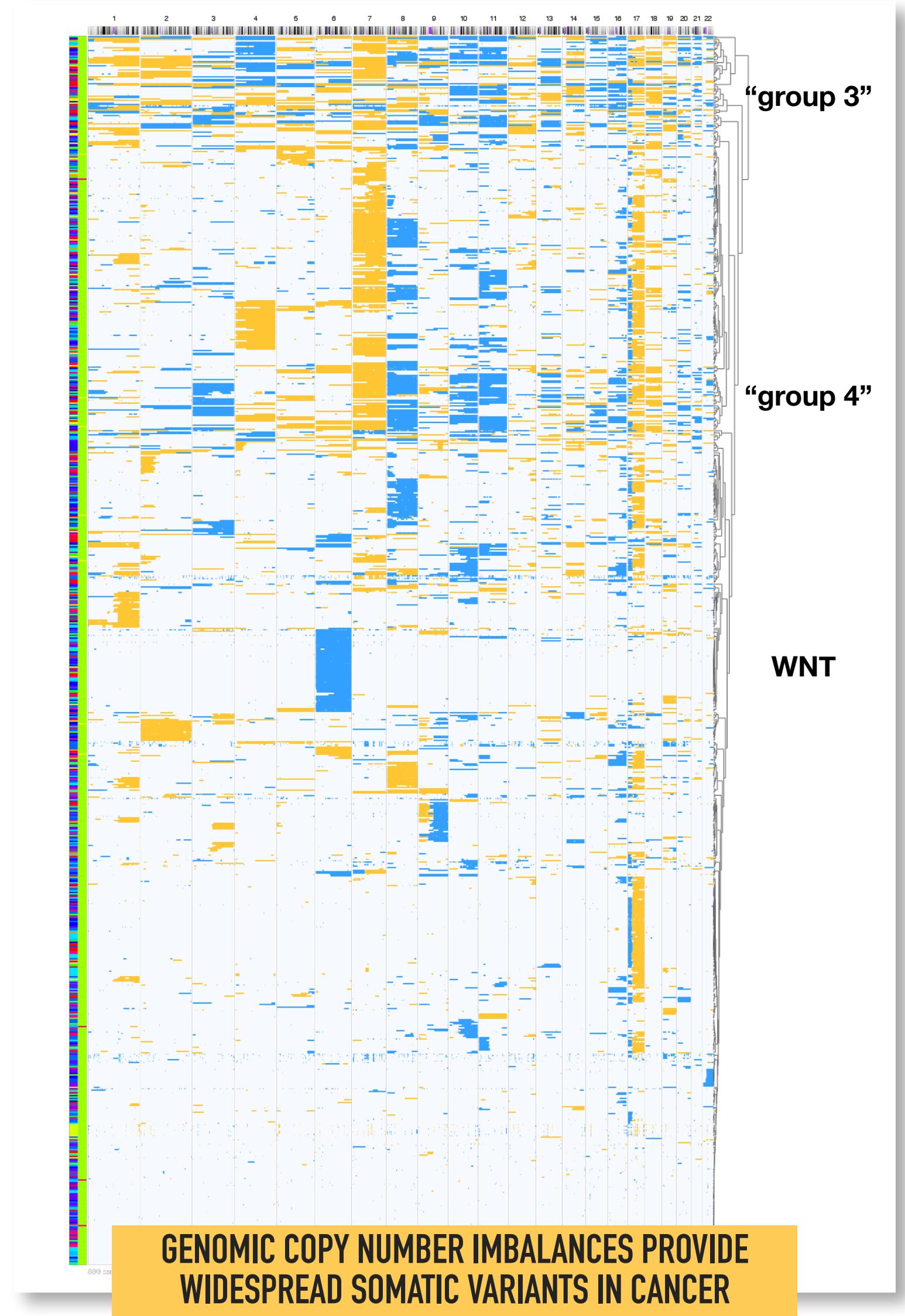
# Somatic CNVs In Cancer

## Recurrent mutation patterns

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



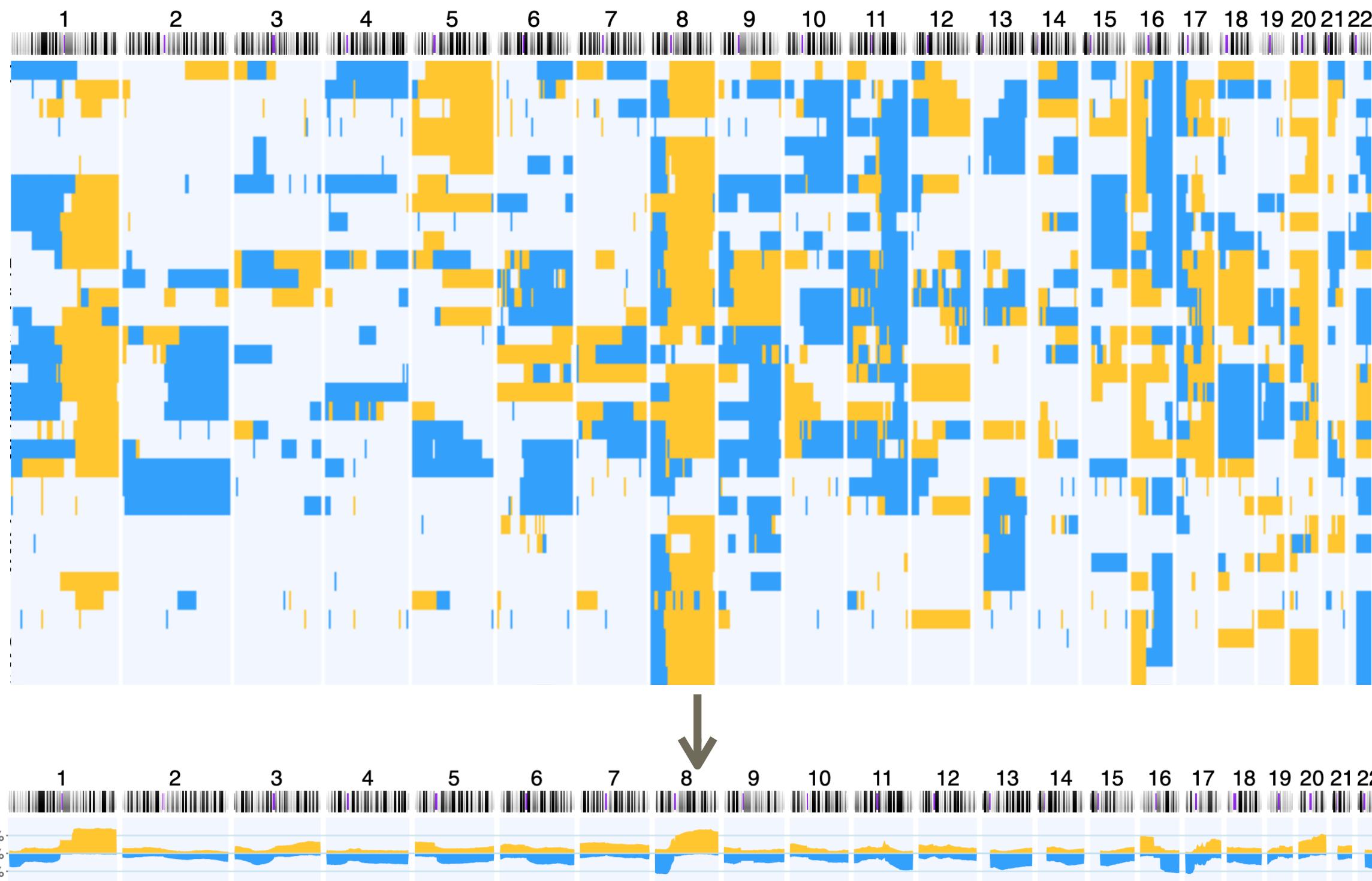
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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



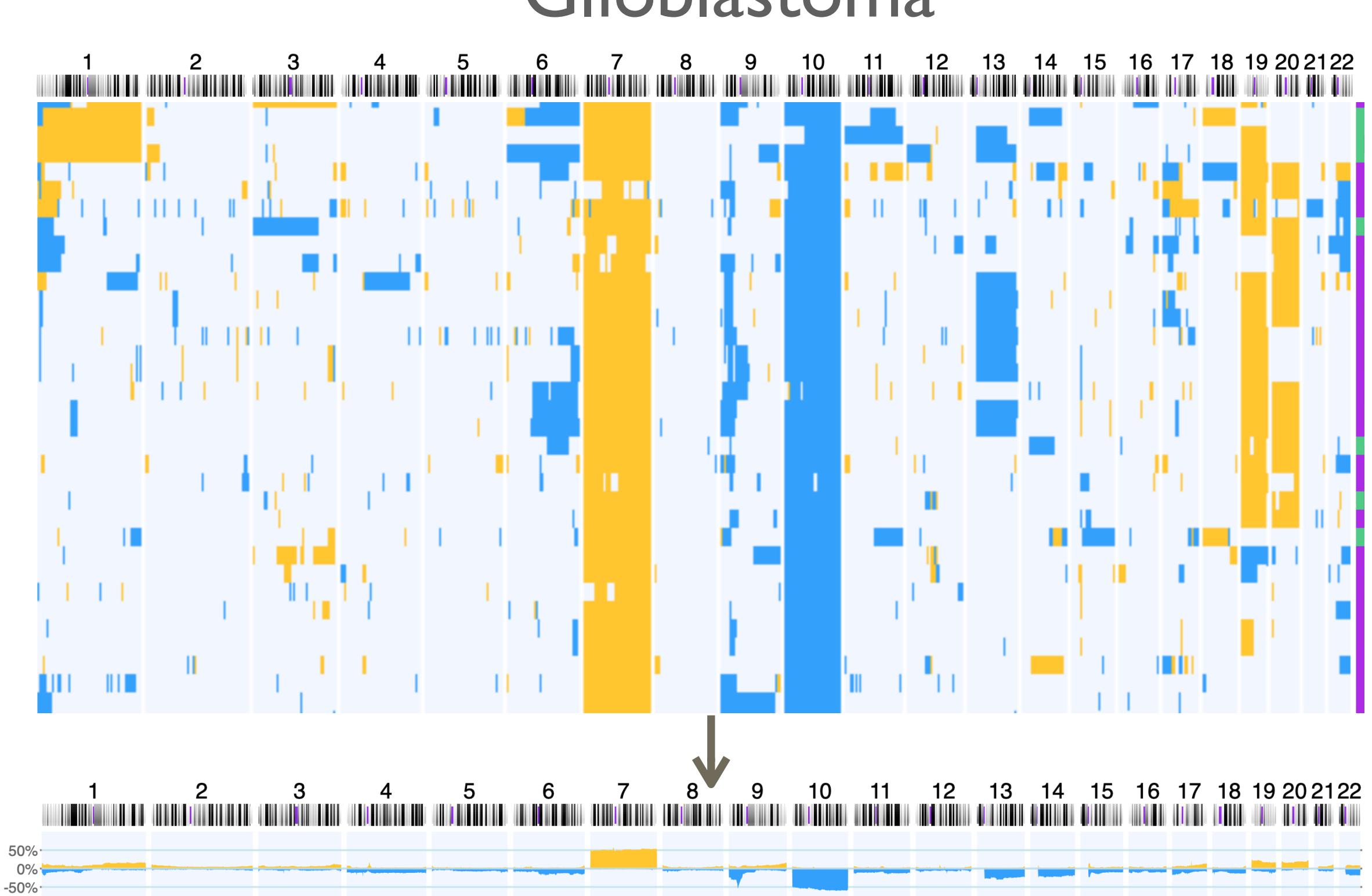
# Drivers? Passengers? Markers?

## Disentangling CNA Patterns

Ductal Breast Carcinoma



Glioblastoma

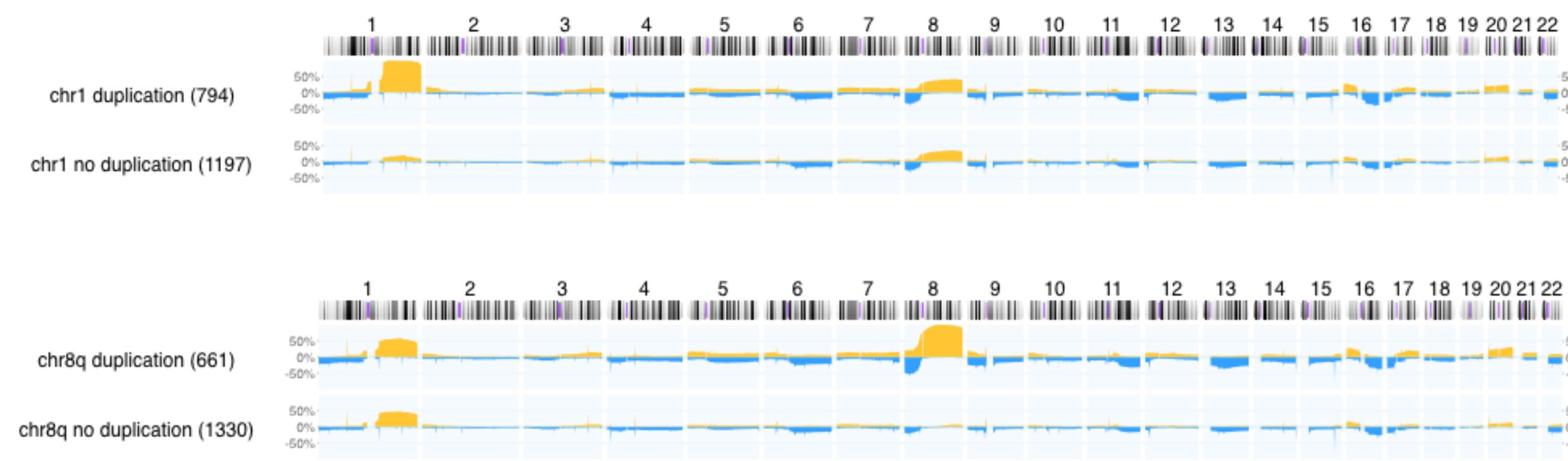


Thousands of genes involved (passengers)  
Descriptive report with arbitrary cutoff

# Intra-Disease CNA patterns are not random

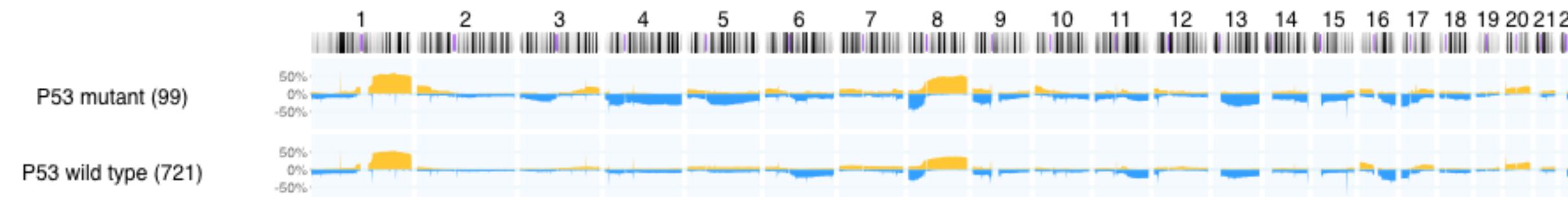
## I. Robust

Independent evolution of CNV landscape



## 2. Regulated

P53 mutant slightly higher CNV



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

## 3. Functionally relevant

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

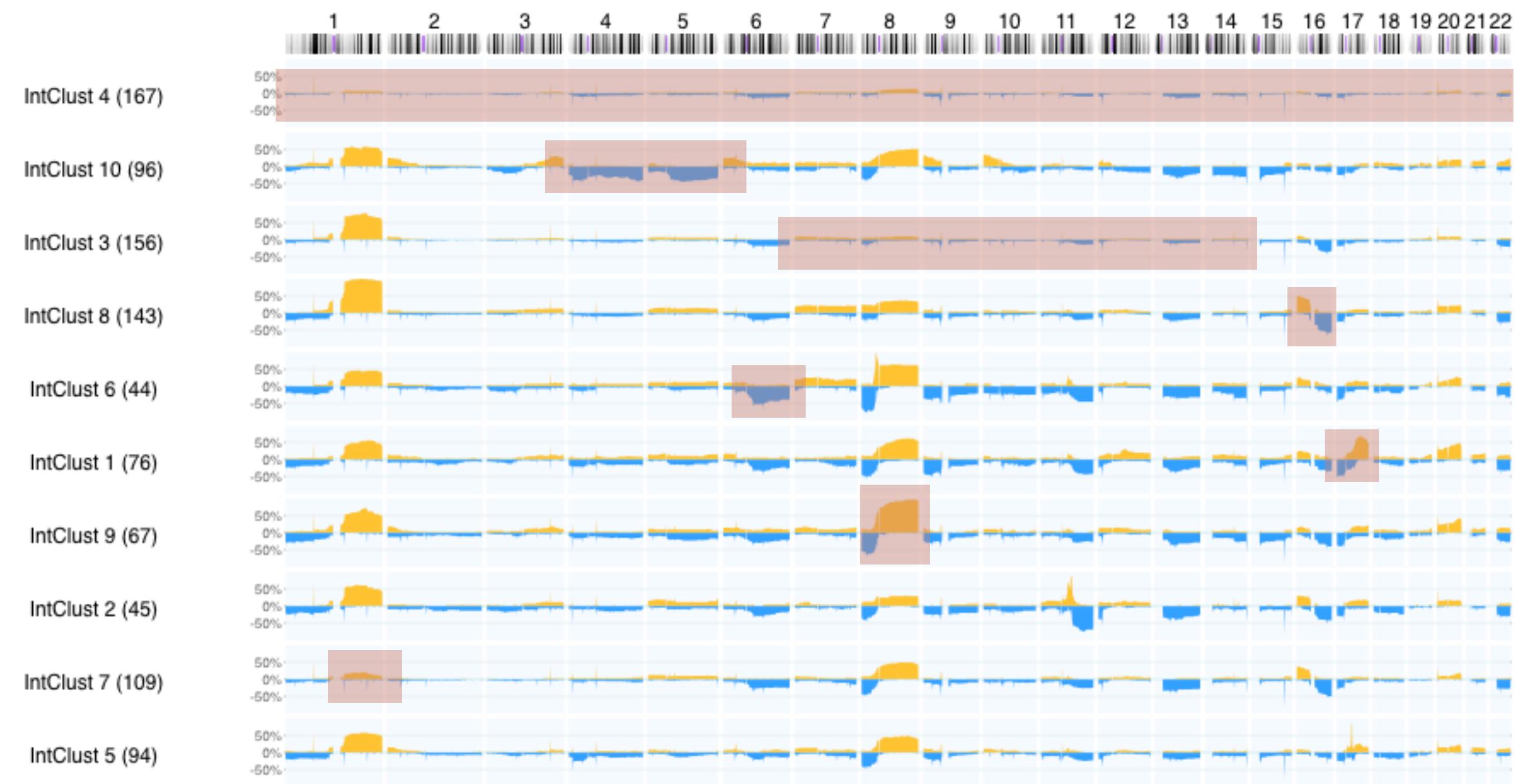
SJ Dawson, OM Rueda, S Aparicio, C Caldas - The EMBO journal, 2013 - [embopress.org](#)

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

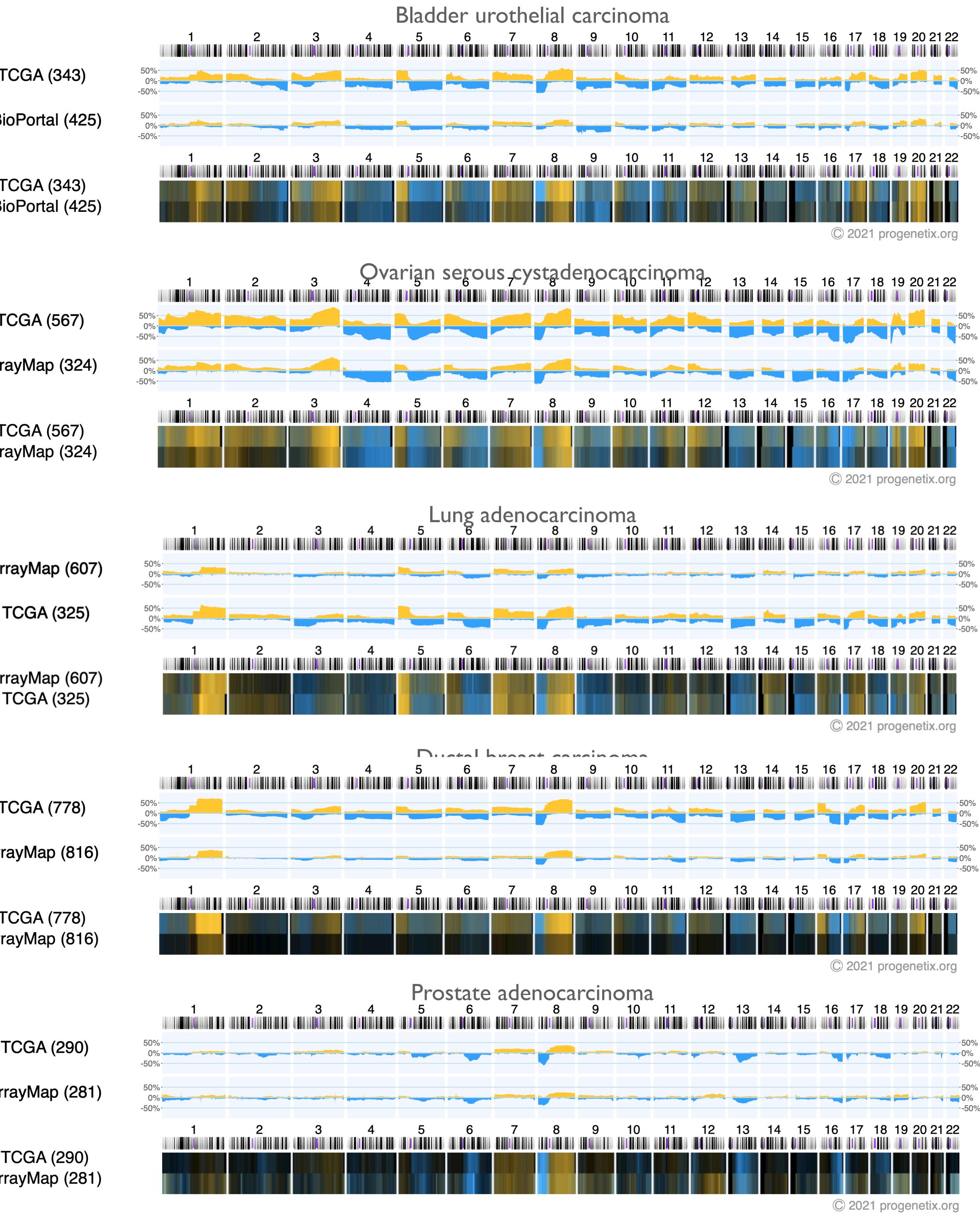
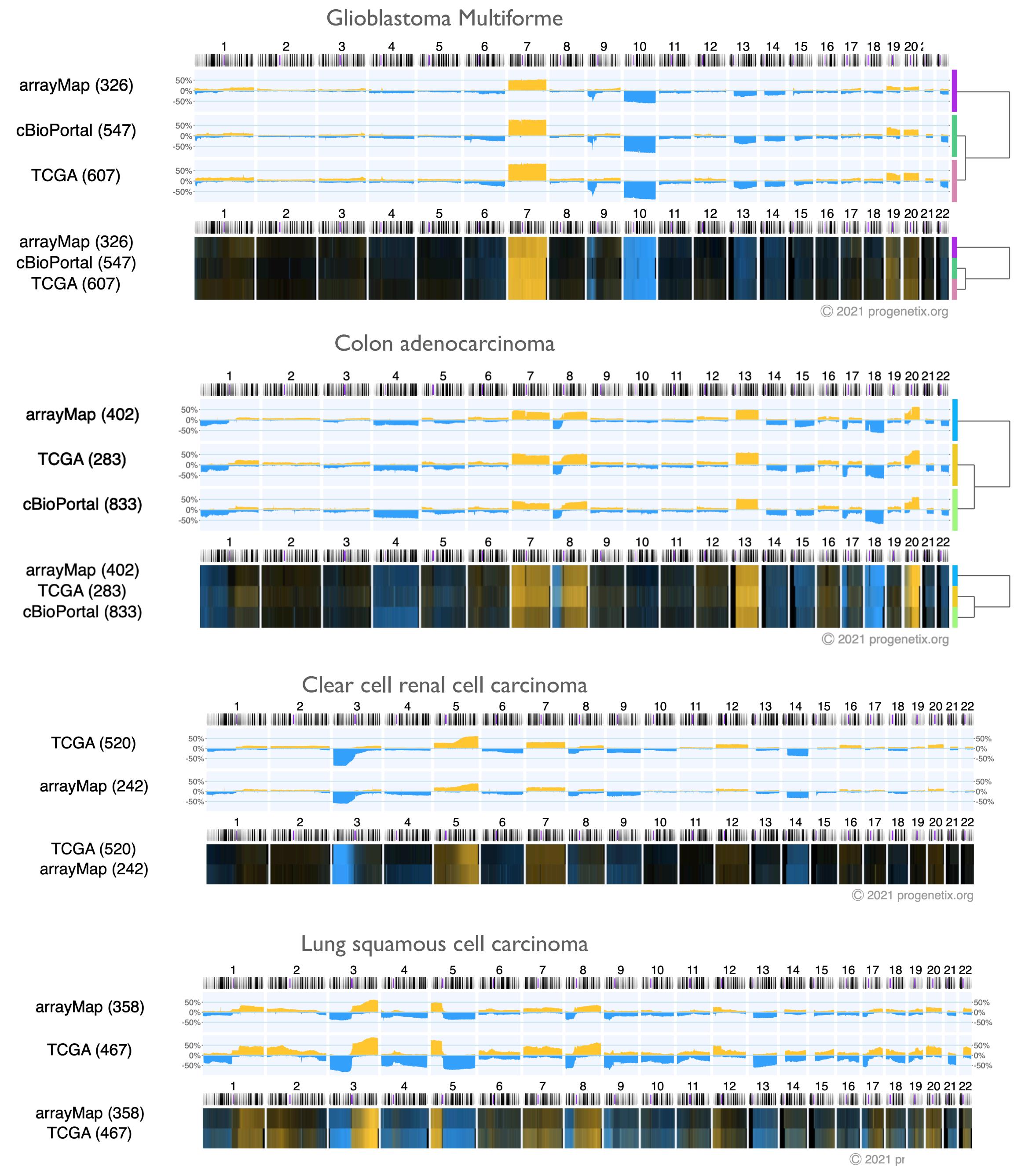
☆ 99 Cited by 278 Related articles All 8 versions

### IntClust BRCA subtypes

Distinctive CNV patterns among IntClust groups



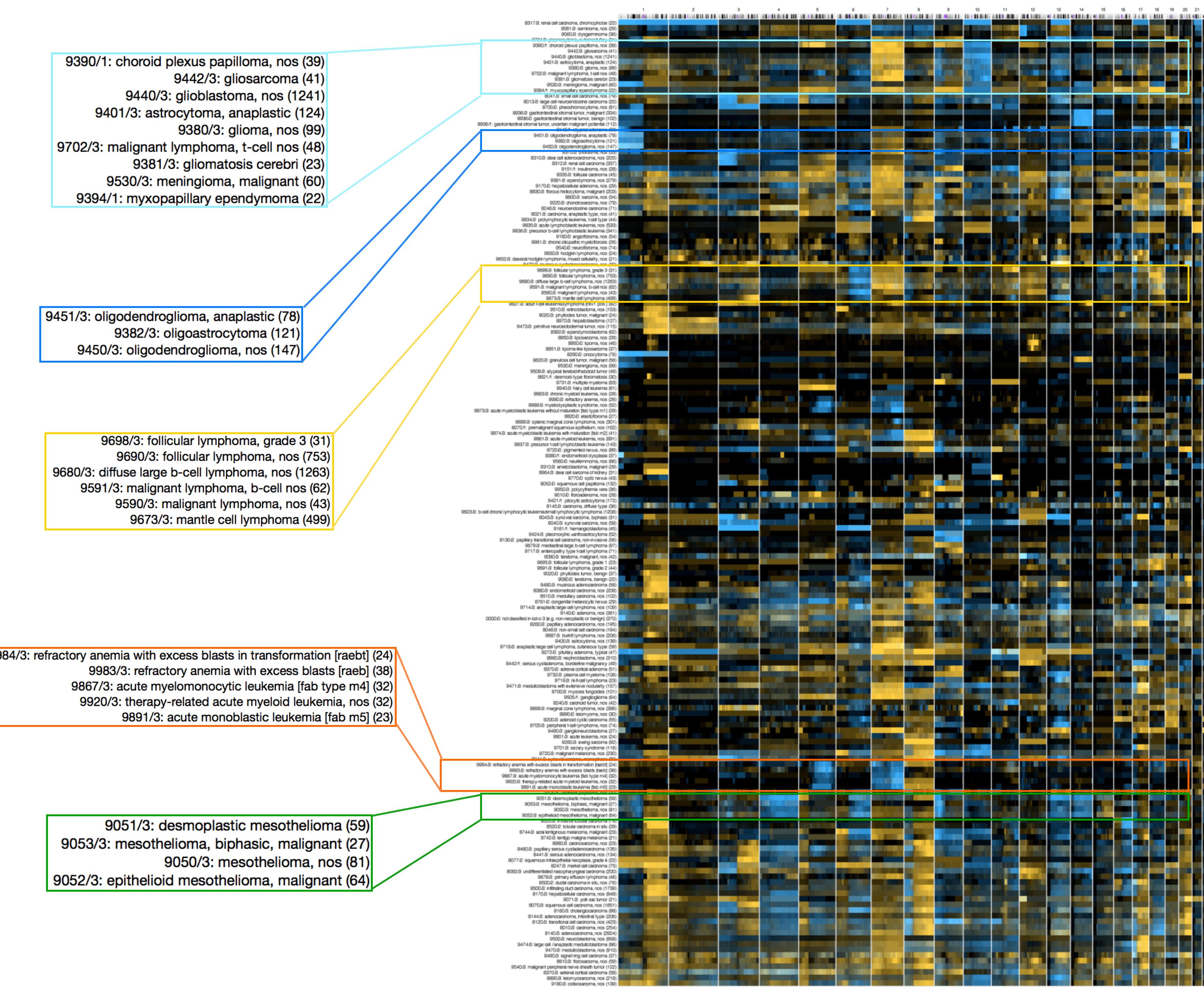
# Limited Resource-Specific Biases in CNA Data



# Somatic Mutations In Cancer: Patterns

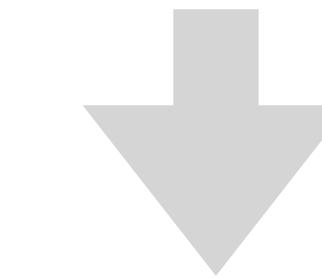
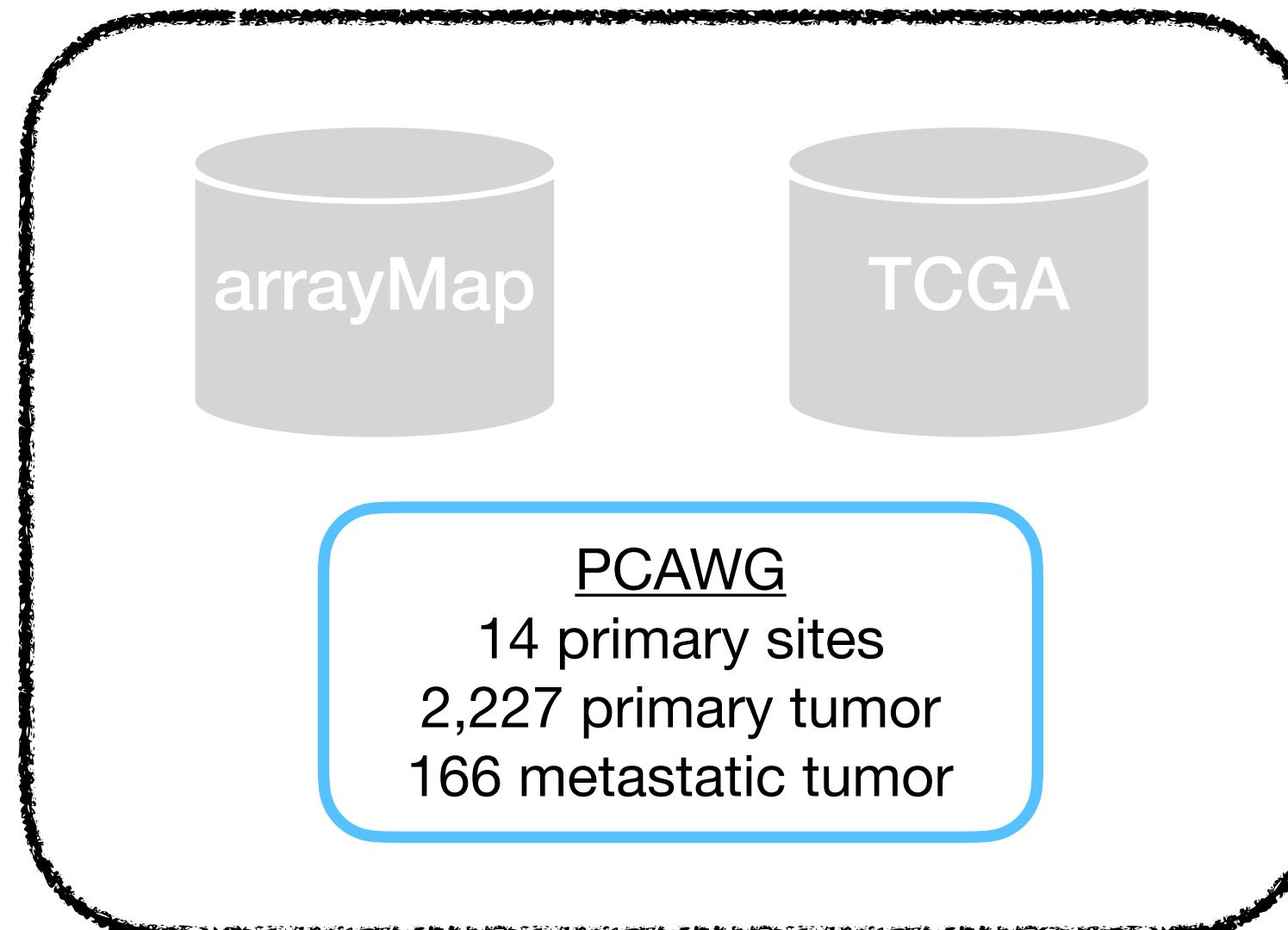
## Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



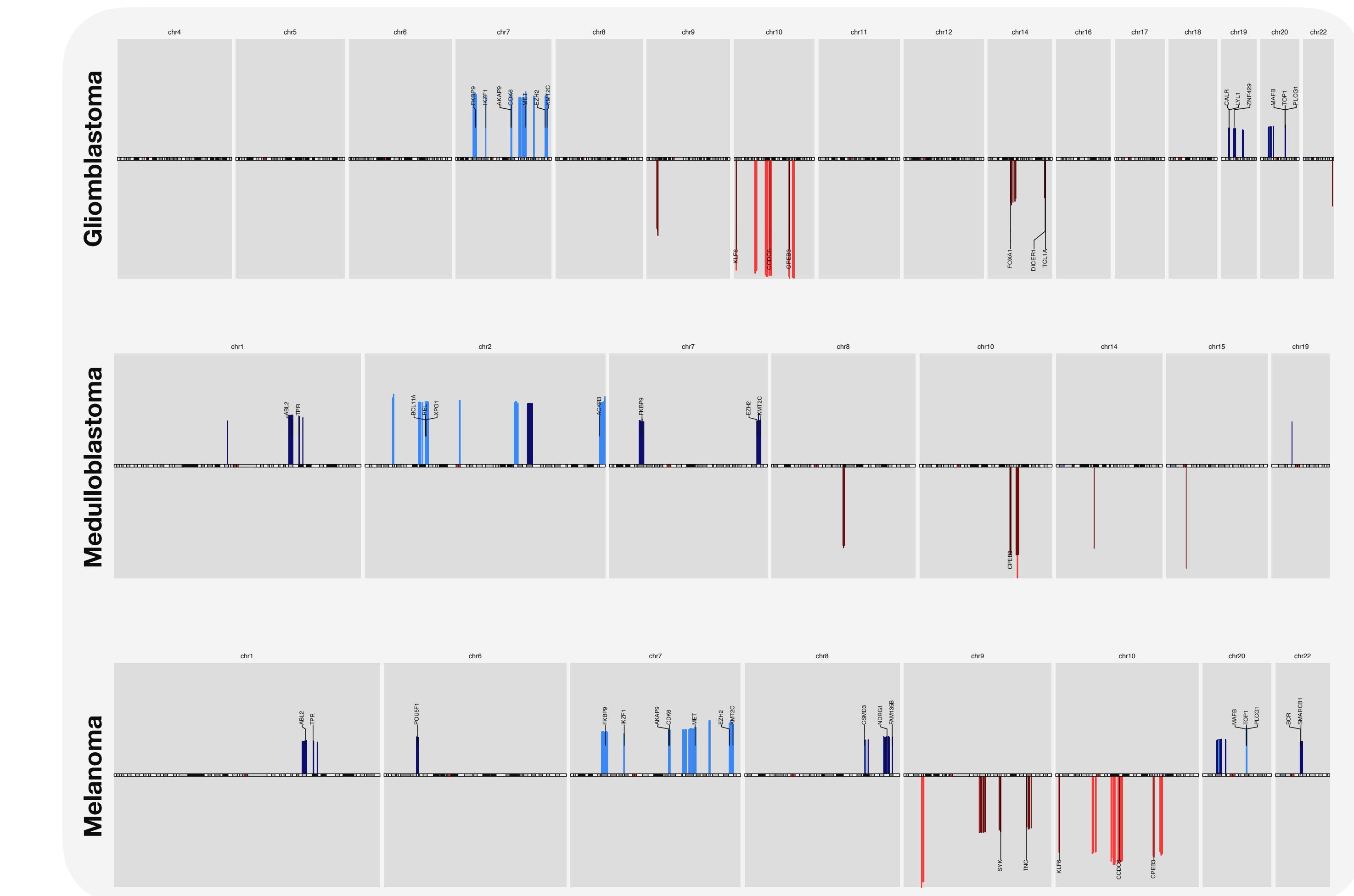
# Unique Patterns of Copy Number Mutations Across Cancer Types

An extensive collection of tumor CNV



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

Examples of unique CNV patterns



# Progenetix as Example Genomics Resource

## Some trajectories ...

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



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



**ICR** The Institute of  
 Cancer Research

UNIVERSITÄTSMEDIZIN  
 GÖTTINGEN **UMG**

FOLLOW US ON [twitter](#)

Sponsors & Support

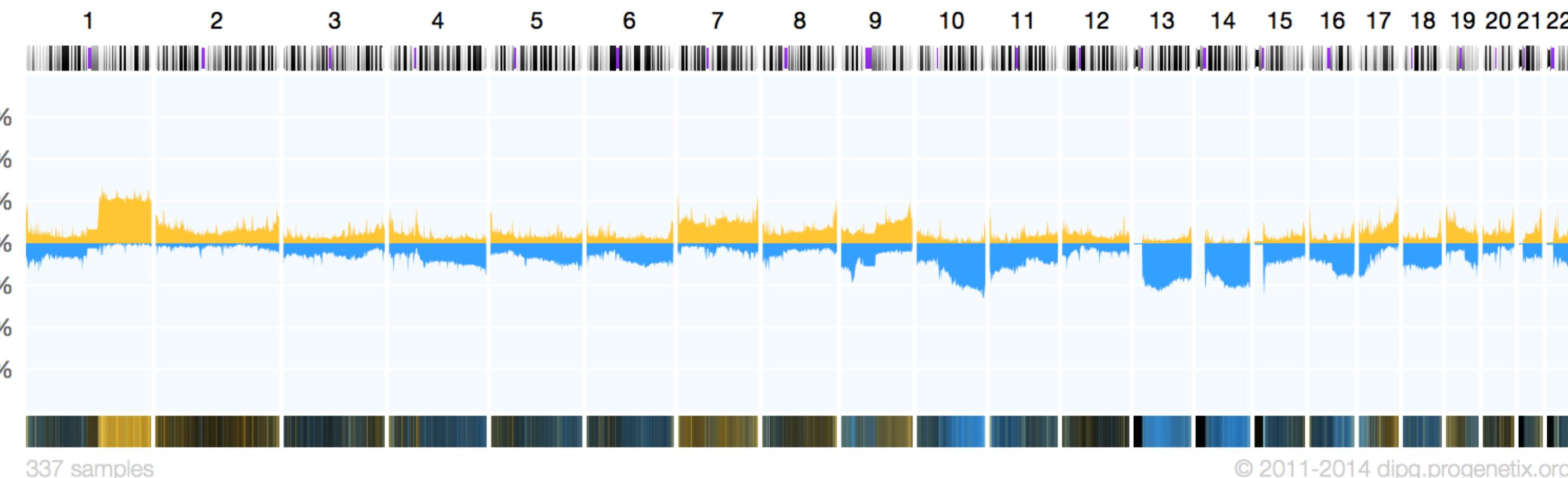


# Diffuse Intrinsic Pontine Glioma (DIPG) Genomics Repository

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

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

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



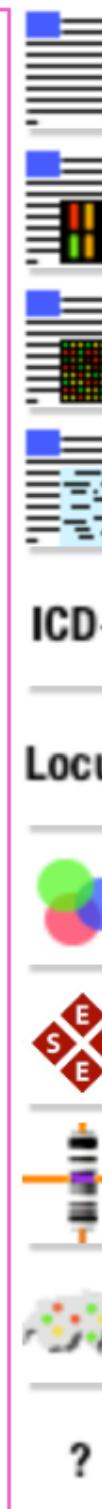
Genomic copy number aberrations in 337 DIPG and pediatric high grade gliomas

DIPG NEWS

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

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

[More news ...](#)



2014

# Restart DIPG... Topical Resource

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

The screenshot shows the Progenetix DIPG Mutation Data search interface. The left sidebar contains links for About Progenetix, News, Cancer CNV Profiles, Search Samples, arrayMap, Studies & Cohorts (with TCGA and DIPG samples), Publication DB, Services (NCIt and UBERON mappings), Upload & Plot, Documentation, Beacon+, and Baudisgroup @ UZH. The main content area has a header "DIPG Mutation Data" and a sub-header "Search Genomic Variants in DIPG and Related Pediatric Gliomas". It states that the portal explores genomic variants in aggressive childhood gliomas, enabled by data from Mackay et al. (2017). Below this is a "Search Samples" section with tabs for Allele Request, Range Query (which is selected), and CNV Requests. A "SNV Range Example" box contains text about querying mutations in the EIF4A1 gene across the DIPG dataset. Another box explains that a range query returns variants between given positions, with exact variants retrievable via a variant handover link. The bottom section contains input fields for Gene Symbol (Select...), Reference name (17), (Structural) Variant Type (Select...), Start or Position (7572826), and End (Range or Structural Var.) (7579005).

**GA4GH Genome Beacons** A Driver Project of the Global Alliance for Genomics and Health GA4GH and supported through ELIXIR

**Beacon Protocol for Genomic Data Sharing**

Beacons provide discovery services for genomic data using the Beacon API developed by the **Global Alliance for Genomics and Health (GA4GH)**. The **Beacon protocol** itself standard for genomics data discovery. It provides a framework for public web services against genomic data collections, for instance from population based or disease specific genome repositories.

**Baudisgroup @ UZH**

(Ni Ai)  
Michael Baudis  
(Haoyang Cai)  
Paula Carrio Cordo  
Bo Gao  
Qingyao Huang  
(Saumya Gupta)  
(Nitin Kumar)  
Sofia Pfund  
Rahel Paloots  
Pierre-Henri Toussaint

The original Beacon protocol has five main characteristics:

- **Simple:** focus on robustness and ease of implementation.
- **Federated:** maintained by a community of developers and can be modified e.g. through changing the codebase.
- **General-purpose:** used for copy number queries ("variantCNVrequest"), which capture a set of similar variants.
- **Aggregative:** provide a way to aggregate results from multiple sources.
- **Privacy protecting:** query results are aggregated and do not reveal individual patient information.

Sites offering *beacons* can scale their services to handle complex queries among a potentially large number of users. Since 2015 the development of the Beacon protocol has been internationalized, involving participants from around the world. Recent developments include:

- providing a framework for handling structural variants
- allowing for data delivery in various environments and allowing for more complex queries.

**Beacon v2 - Towards Flexibility**

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**

9-19000000,21975098-21967753,23000000;DEL;ncit:C3058;DUO:0000004;HP:0003621

[beacon-project.io](http://beacon-project.io)

 Swiss Personalized Health Network

**Beacon+** About Progenetix Help

**Baudisgroup @ UZH**

(Ni Ai)  
Michael Baudis  
(Haoyang Cai)  
Paula Carrio Cordo  
Bo Gao  
Qingyao Huang  
(Saumya Gupta)  
(Nitin Kumar)  
Sofia Pfund  
Rahel Paloots  
Pierre-Henri Toussaint

The original Beacon protocol has five main characteristics:

- **Simple:** focus on robustness and ease of implementation.
- **Federated:** maintained by a community of developers and can be modified e.g. through changing the codebase.
- **General-purpose:** used for copy number queries ("variantCNVrequest"), which capture a set of similar variants.
- **Aggregative:** provide a way to aggregate results from multiple sources.
- **Privacy protecting:** query results are aggregated and do not reveal individual patient information.

Sites offering *beacons* can scale their services to handle complex queries among a potentially large number of users. Since 2015 the development of the Beacon protocol has been internationalized, involving participants from around the world. Recent developments include:

- providing a framework for handling structural variants
- allowing for data delivery in various environments and allowing for more complex queries.

**Beacon v2 - Towards Flexibility**

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**

9-19000000,21975098-21967753,23000000;DEL;ncit:C3058;DUO:0000004;HP:0003621

**Cancer Classification(s)**

NCIT:C3058: Glioblastoma (2119)

**City**

Select...  
21000001 21975098  
21967753 23000000

**Query Beacon**

**{S}{B} and GA4GH**

Melanie Courtot  
Helen Parkinson  
many more ...

[beacon.progenetix.org/ui/](http://beacon.progenetix.org/ui/)

**Beacon API Leads**

Jordi Rambla  
Anthony Brooks  
Juha Törnroos

**Discovery WS**

Michael Baudis (Beacon)  
Marc Fiume (Networks)

**ELIXIR**

Gary Saunders  
David Lloyd  
Serena Scollen

The Beacon protocol defines an open standard for genomics data discovery, developed by members of the Global Alliance for Genomics & Health. It provides a framework for public web services responding to queries against genomic data collections, for instance from population based or disease specific genome repositories.

This repository contains the specification for the v2 major version upgrade of the Beacon API. It is now (2020) under active development and has *not* seen a stable code release.

For further information, please follow the work here and consult the [Beacon Project website](#).

[github.com/ga4gh-beacon/beacon-specification](https://github.com/ga4gh-beacon/beacon-specification)

**About**

GA4GH Beacon v2 specification.

ga4gh beacon openapi

Readme Apache-2.0 License

Releases No releases published Create a new release

**Contributors**

sdelatorrep sdelatorrep mbaudis mbaudis blankdots blankdots

[github.com/ga4gh-beacon/beacon-specification](https://github.com/ga4gh-beacon/beacon-specification)





University of  
Zurich<sup>UZH</sup>



Global Alliance  
for Genomics & Health



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

[progenetix.org](http://progenetix.org)  
[info.baudisgroup.org](mailto:info.baudisgroup.org)  
[sib.swiss/baudis-michael](http://sib.swiss/baudis-michael)  
[imls.uzh.ch/en/research/baudis](http://imls.uzh.ch/en/research/baudis)  
[beacon-project.io](http://beacon-project.io)  
[schemablocks.org](http://schemablocks.org)