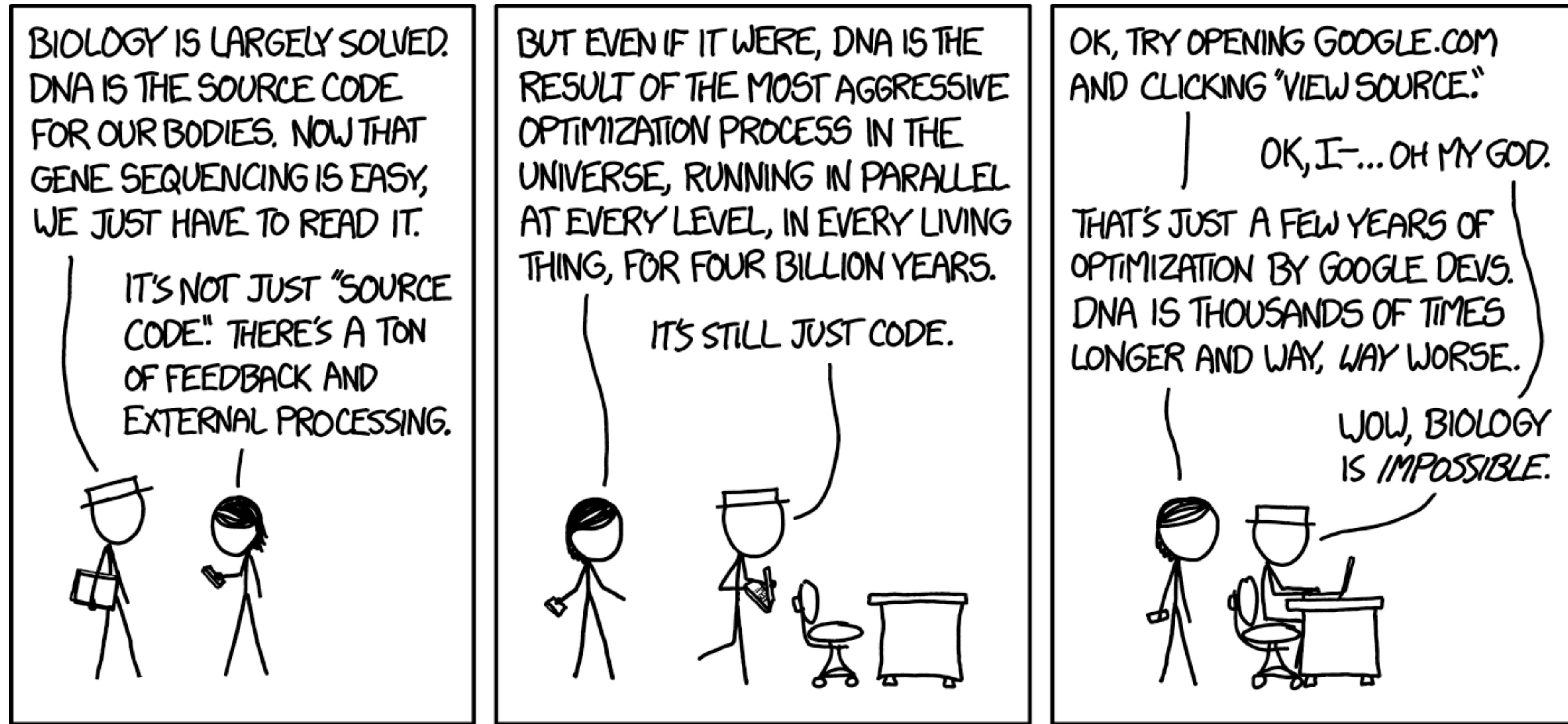


Theoretical Cytogenetics and Oncogenomics

**Cancer Genome Profiles | Oncogenomic Data Resources | Bioinformatics
Methods | Data Exchange Standards for Genomics and Personalized Health**

Biology is has complex source code - Bioinformatics might help

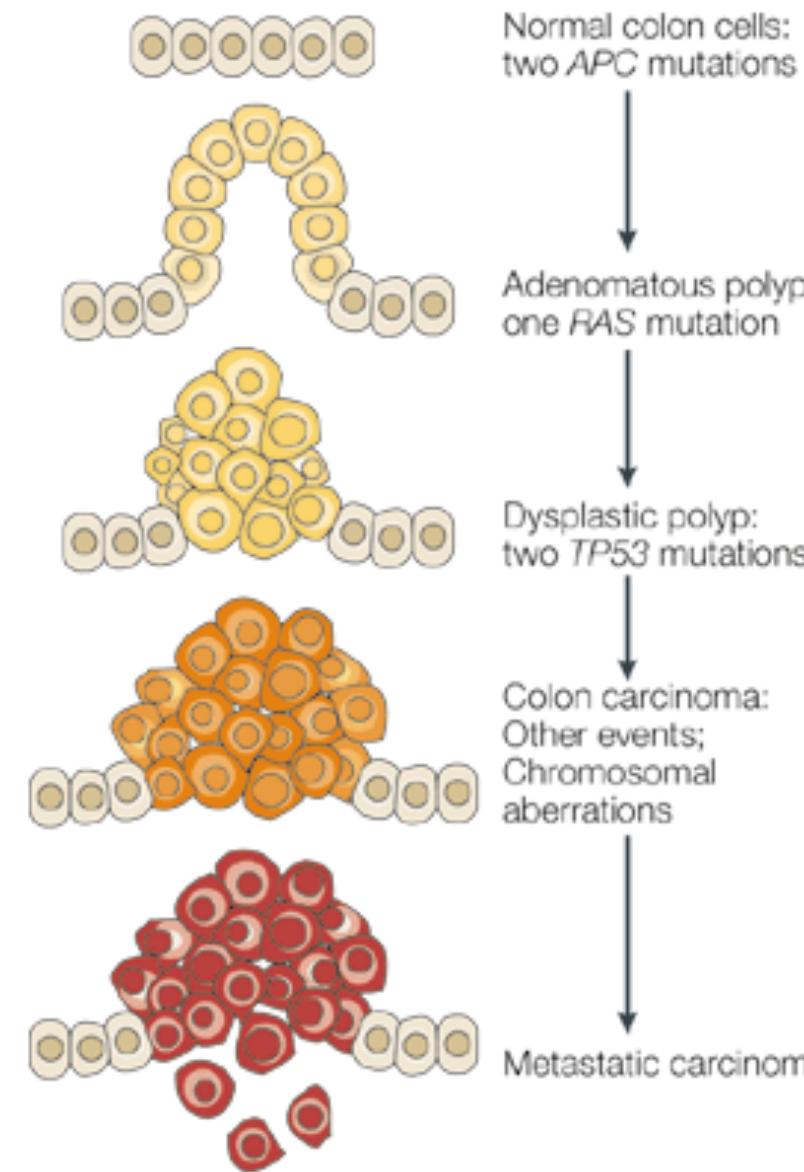
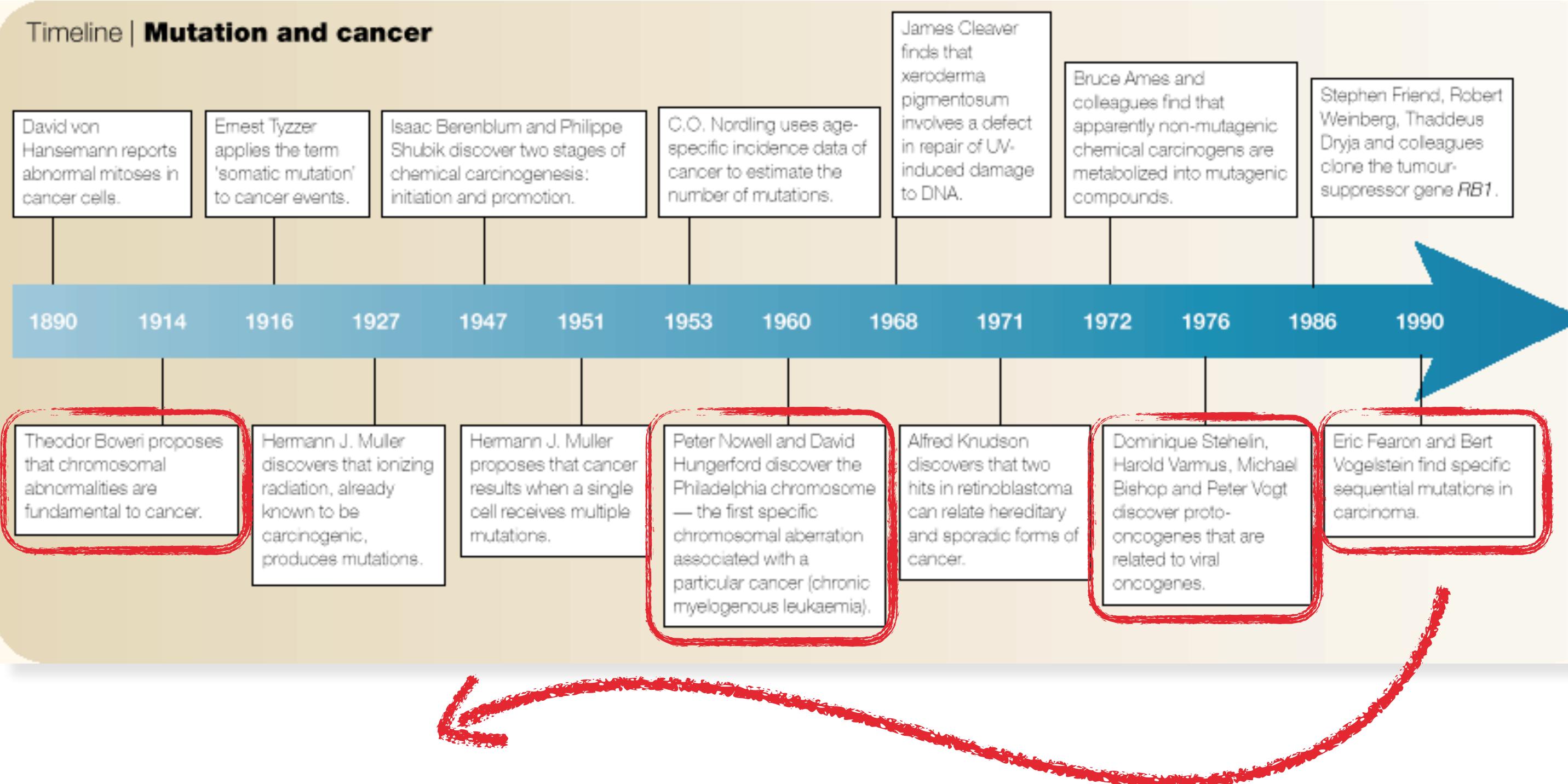


Theoretical Cytogenetics and Oncogenomics

... but what does this entail @baudisgroup?

- patterns & markers in cancer genomics, especially somatic structural genome variants
- bioinformatics support in collaborative studies
- reference resources for curated cancer genome variations
- bioinformatics tools & methods
- standards and reference implementations for data sharing in genomics and personalized health
- open research data "ambassadoring"

Timeline | Mutation and cancer



Cancers are based on acquired and inherited genomic mutations

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



Theodor Boveri (1914)

"Structural Genome Variants" in sea urchin eggs

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

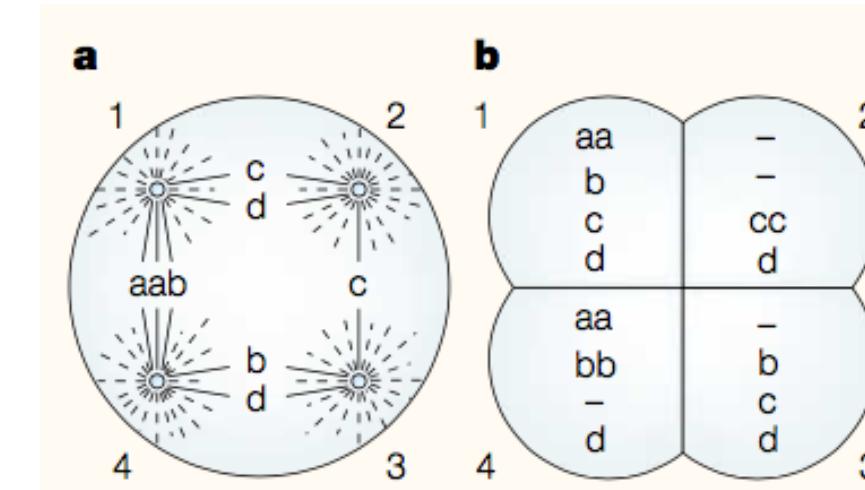
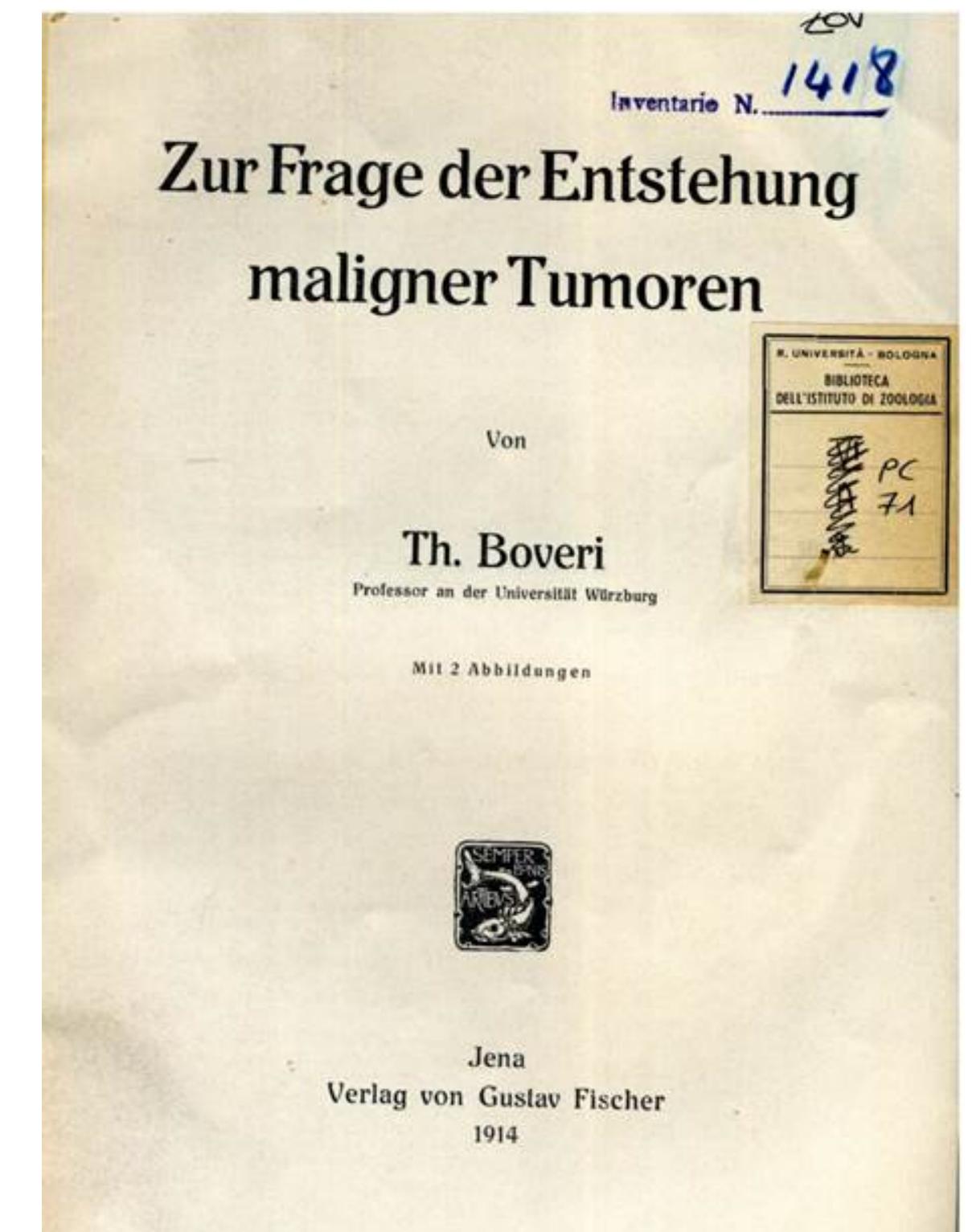
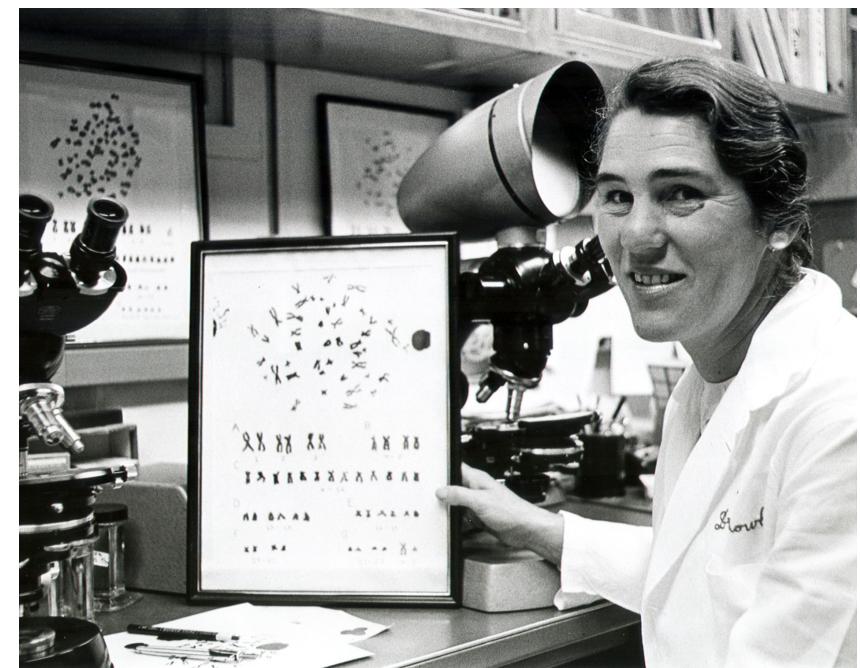


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes.** **a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached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



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

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



Janet Rowley (1972/73)

Chromosomal translocations in cancer

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

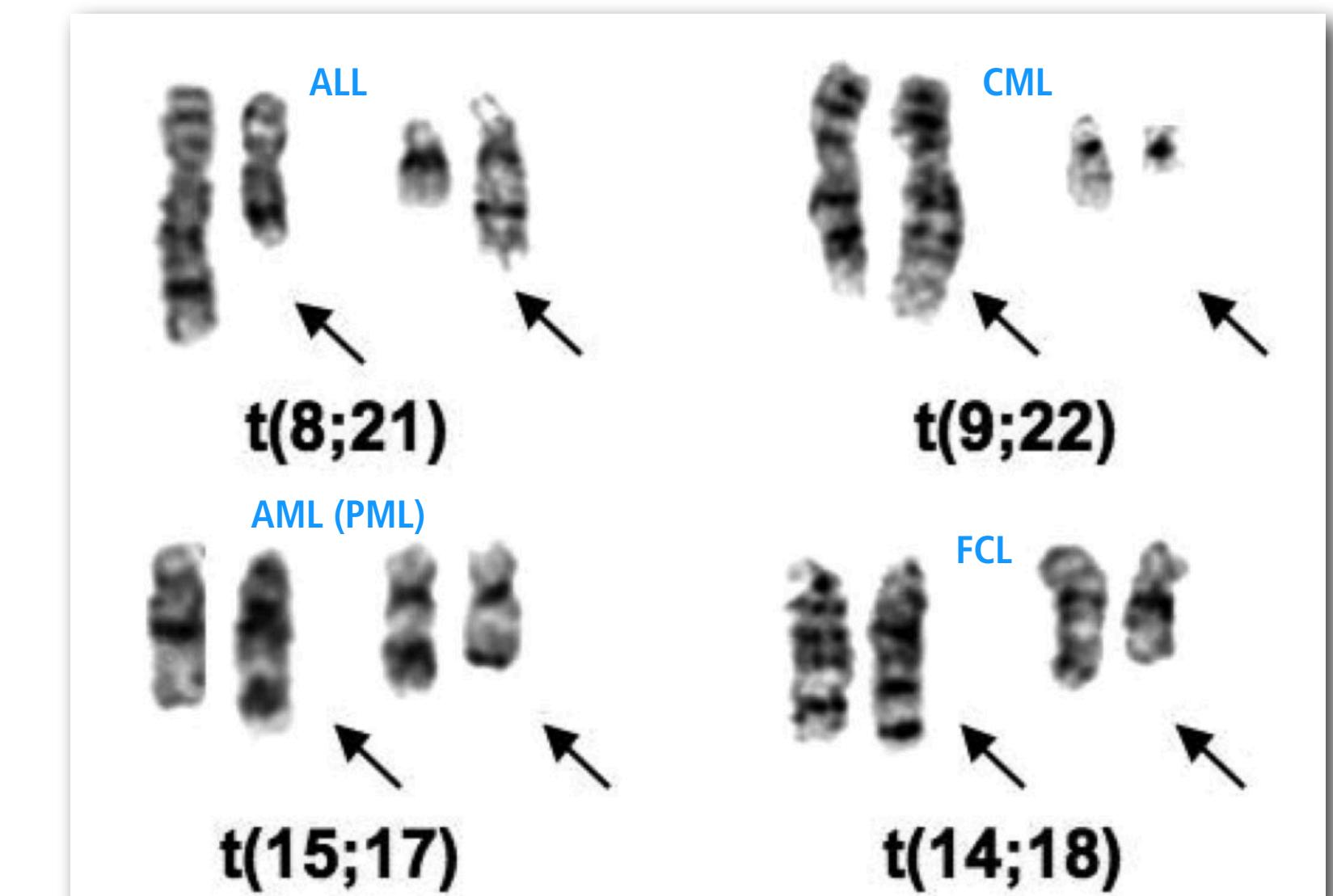
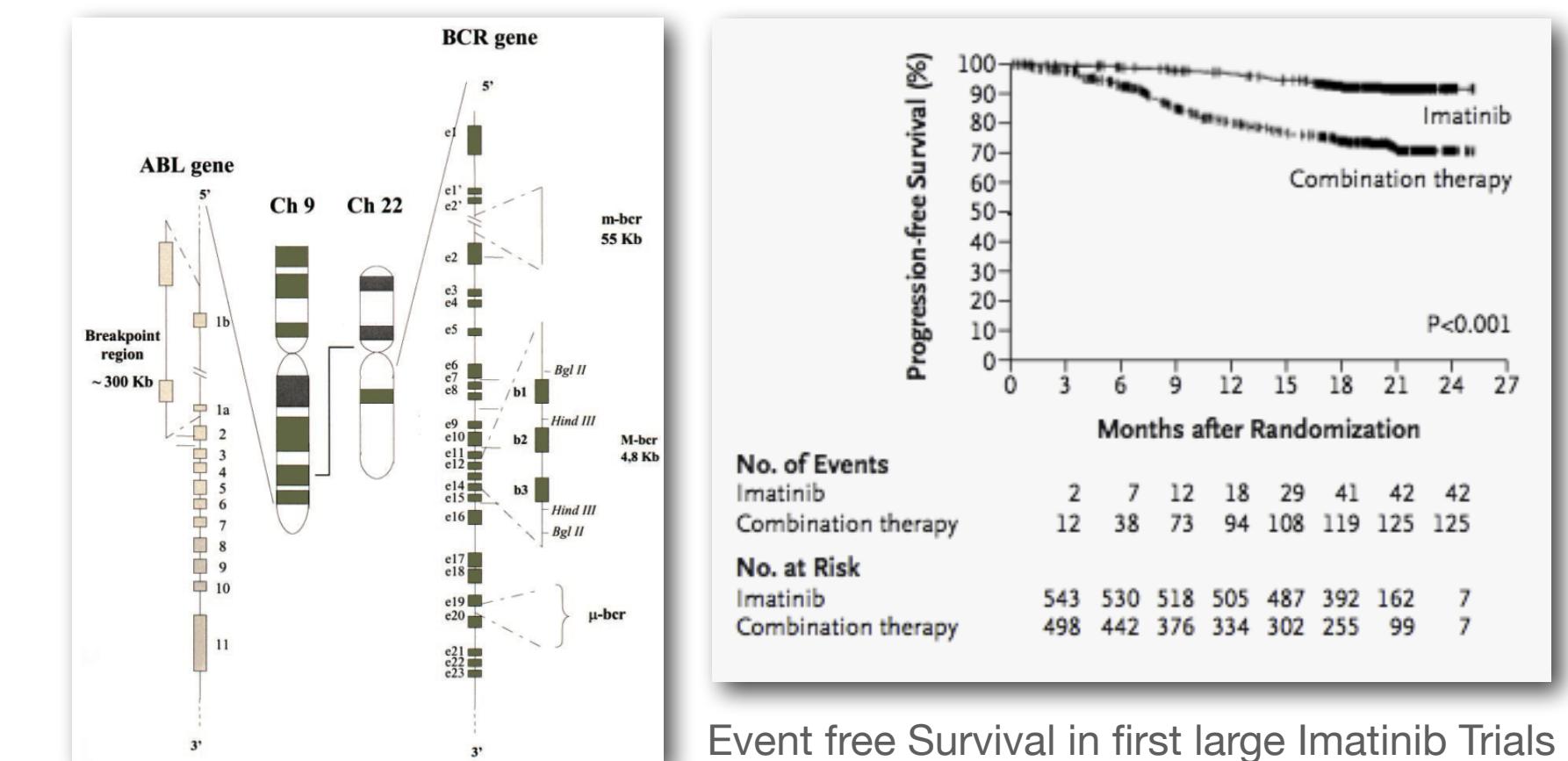


Figure 1. Partial karyotypes of common translocations discovered by Rowley.
The translocations appear in the order in which they were discovered.

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



Event free Survival in first large Imatinib Trials

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

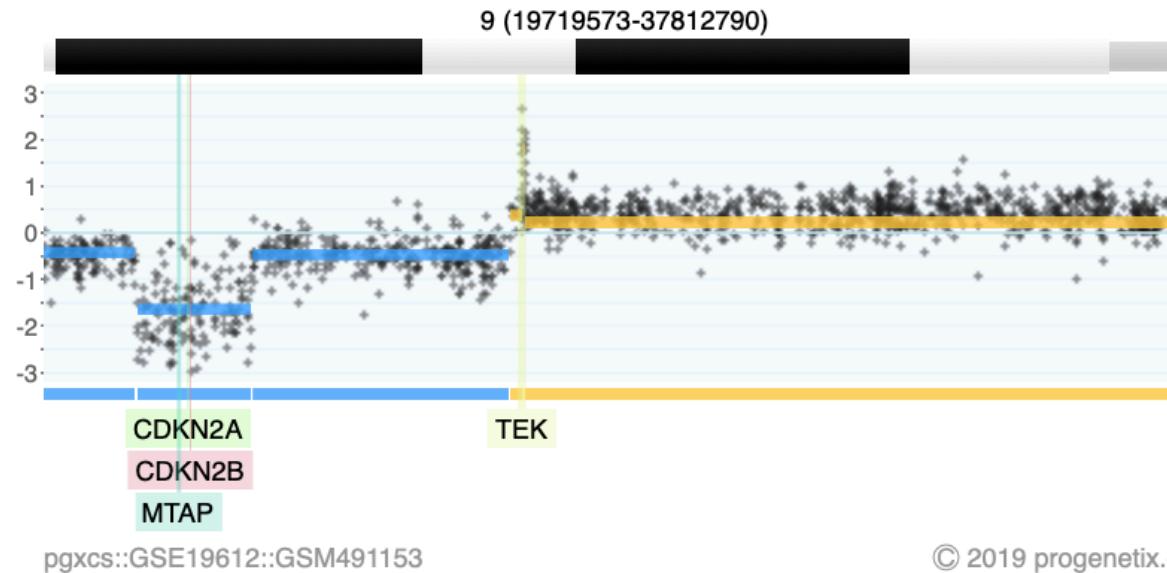
O'Brien et al. Imatinib compared with interferon and low-dose cytarabine...
NEJM (2003) vol. 348 (11)

Theoretical Cytogenetics and Oncogenomics

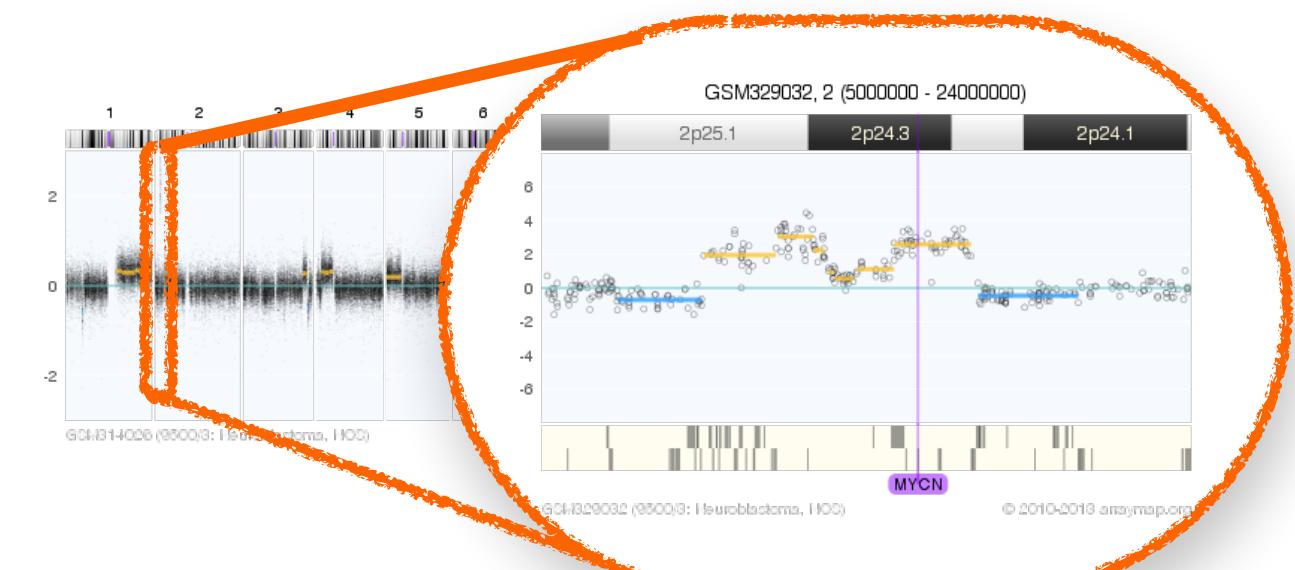
Research | Methods | Standards

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

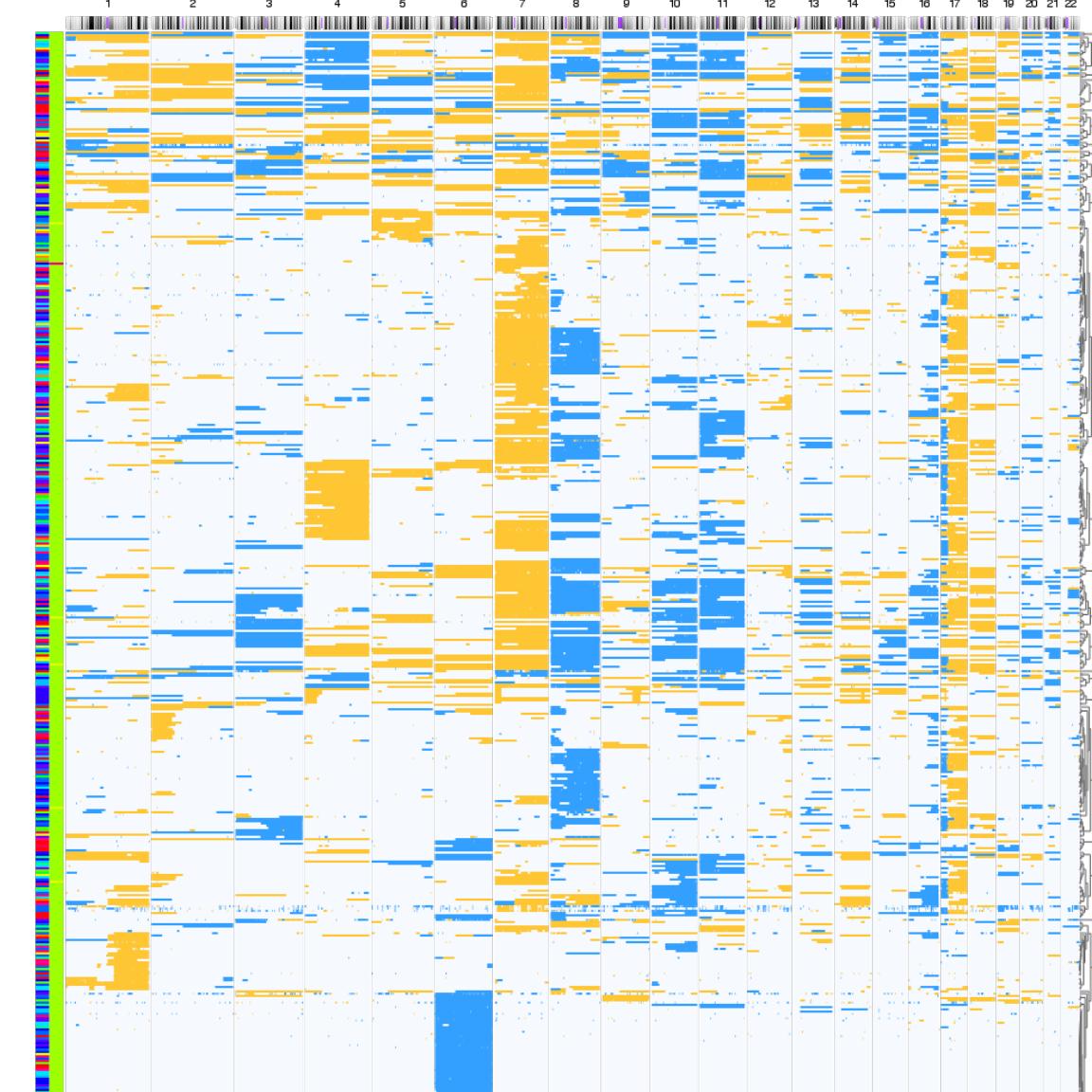
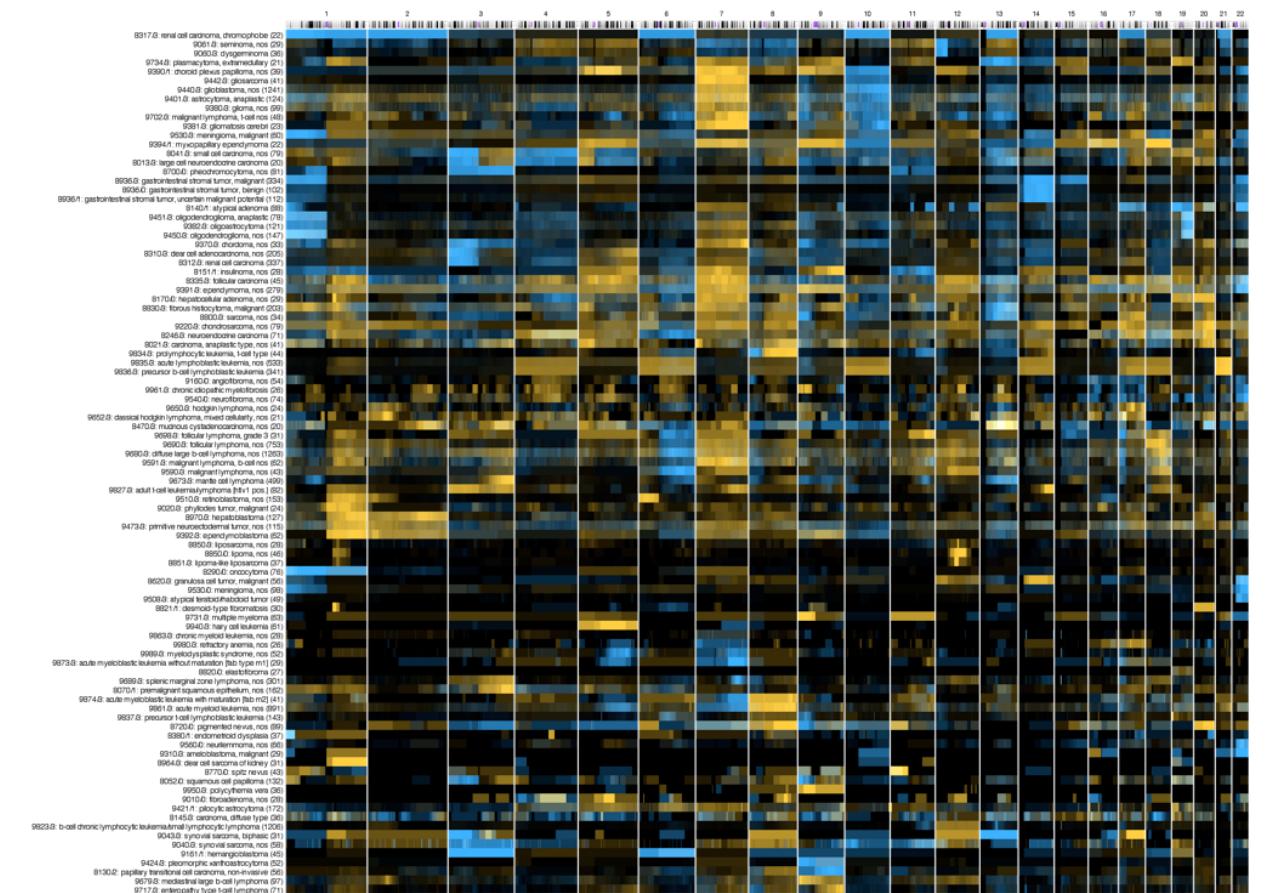
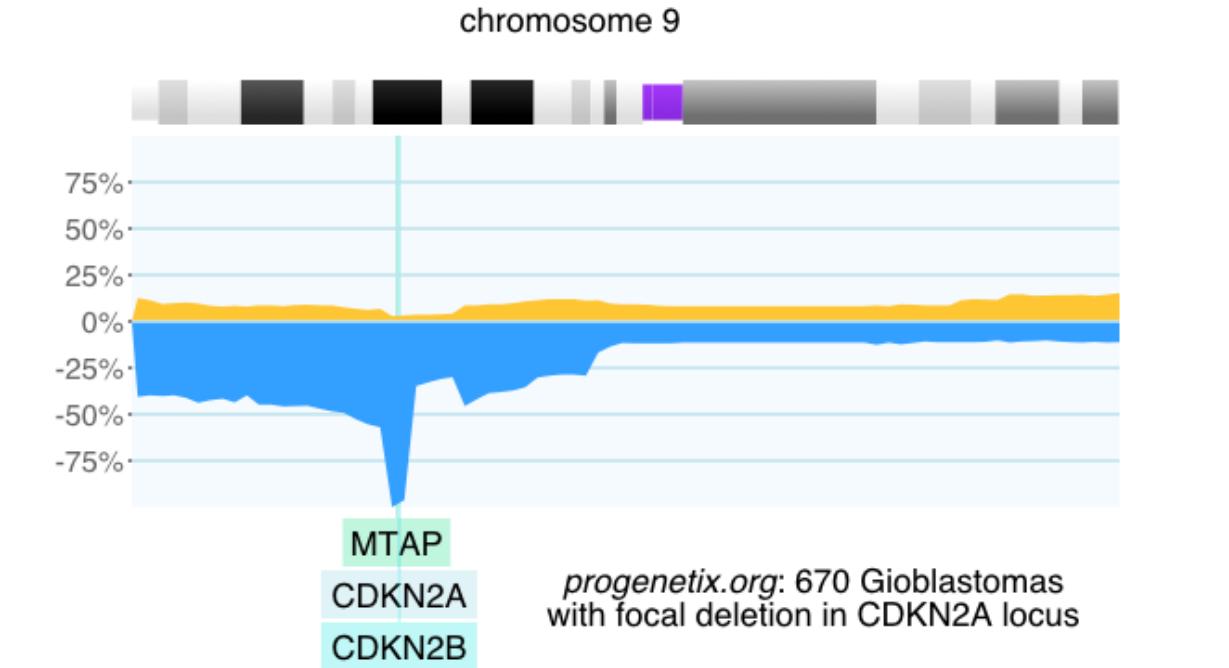
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



2-event, homozygous deletion in a Glioblastoma



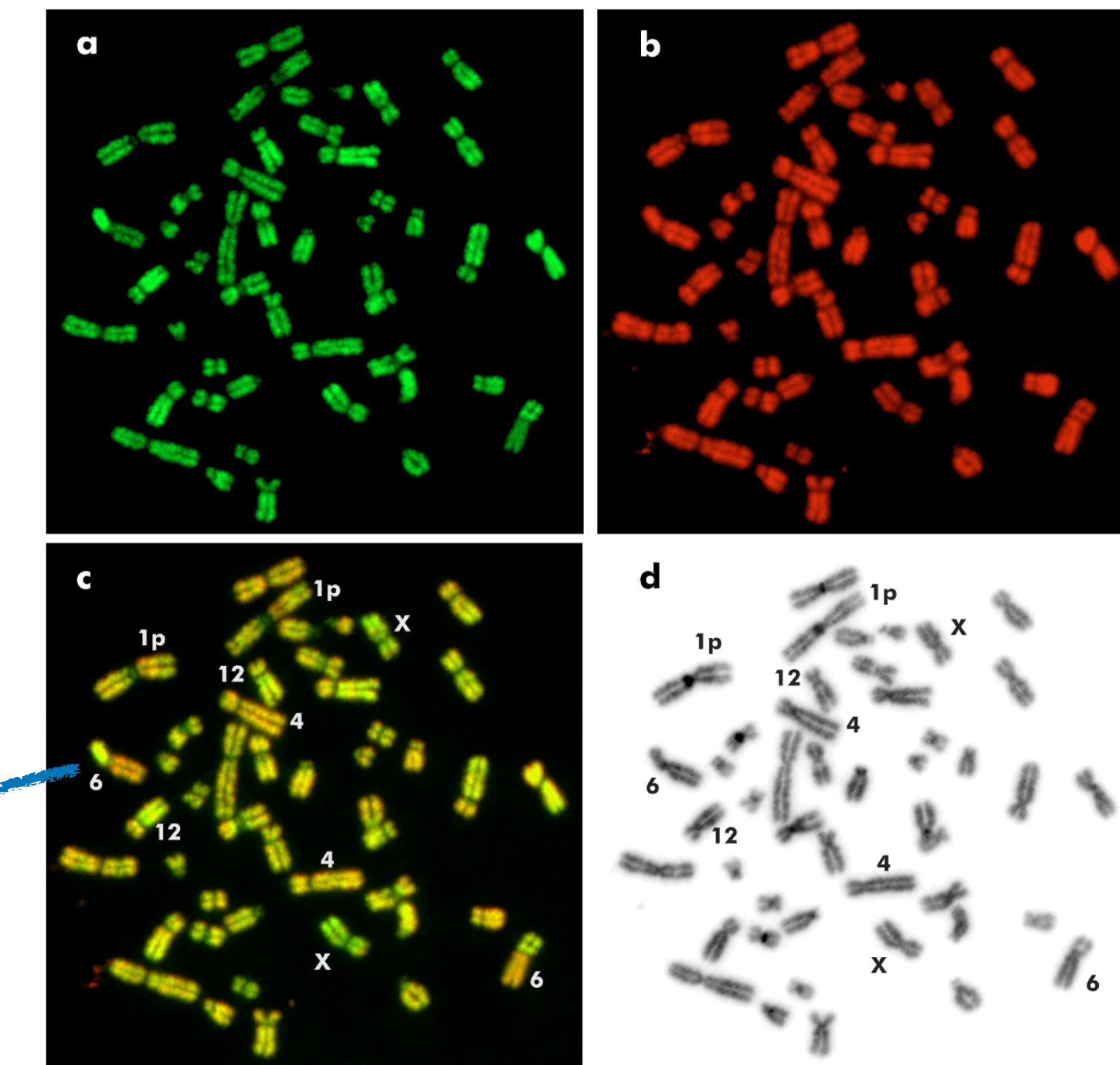
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)



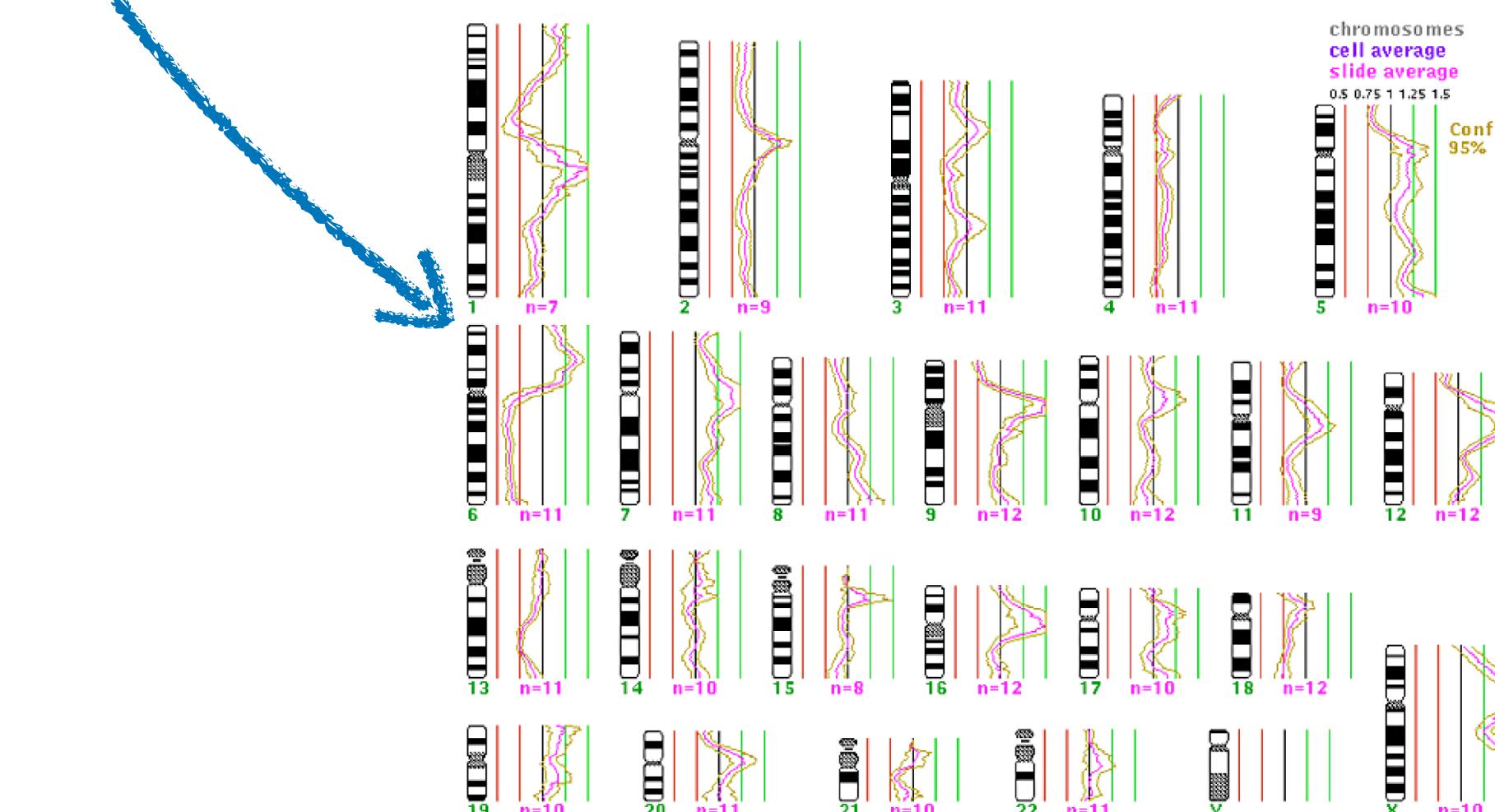
Comparative Genomic Hybridization

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)



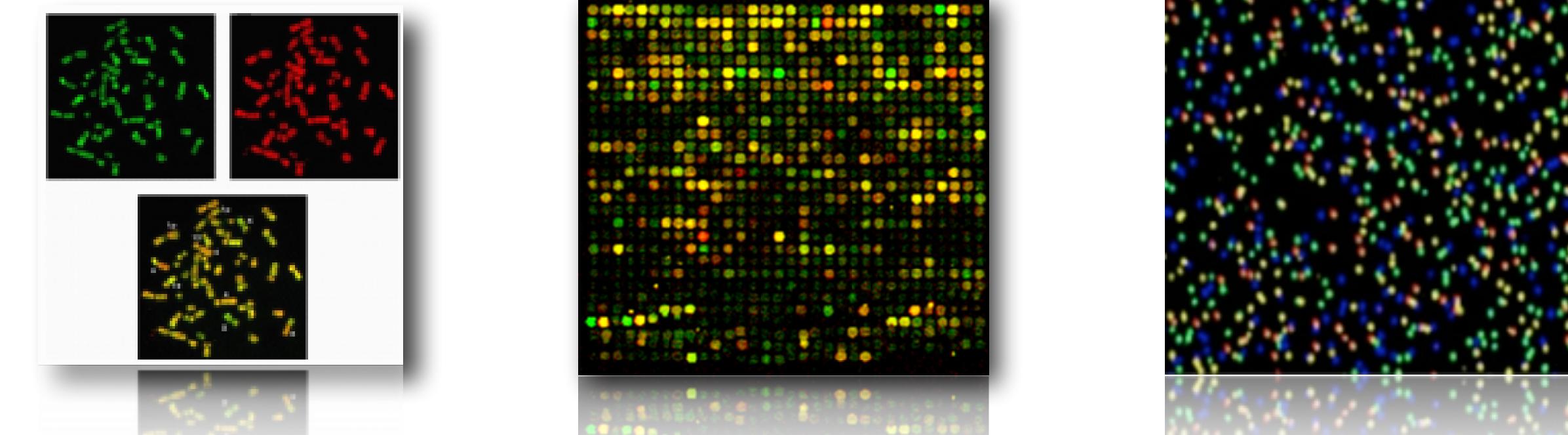
CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen



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

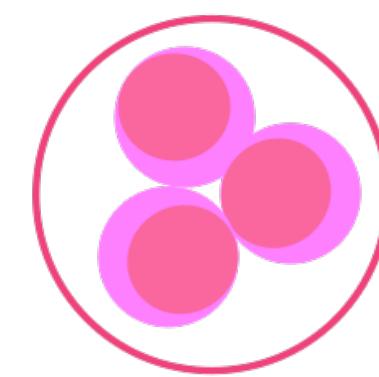
- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences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

WHOLE GENOME SCREENING IN CANCER



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

Data Resources



cancerCellLines



Progenetix in 2021

Cancer Genomics Reference Resource

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

progenetix

Search Samples Modify Query

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

progenetix

Samples: 668 Variants: 286 Calls: 675

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

UCSC region JSON Response Visualization options

Results Biosamples Biosamples Map Variants

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

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

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

progenetix: 670 samples

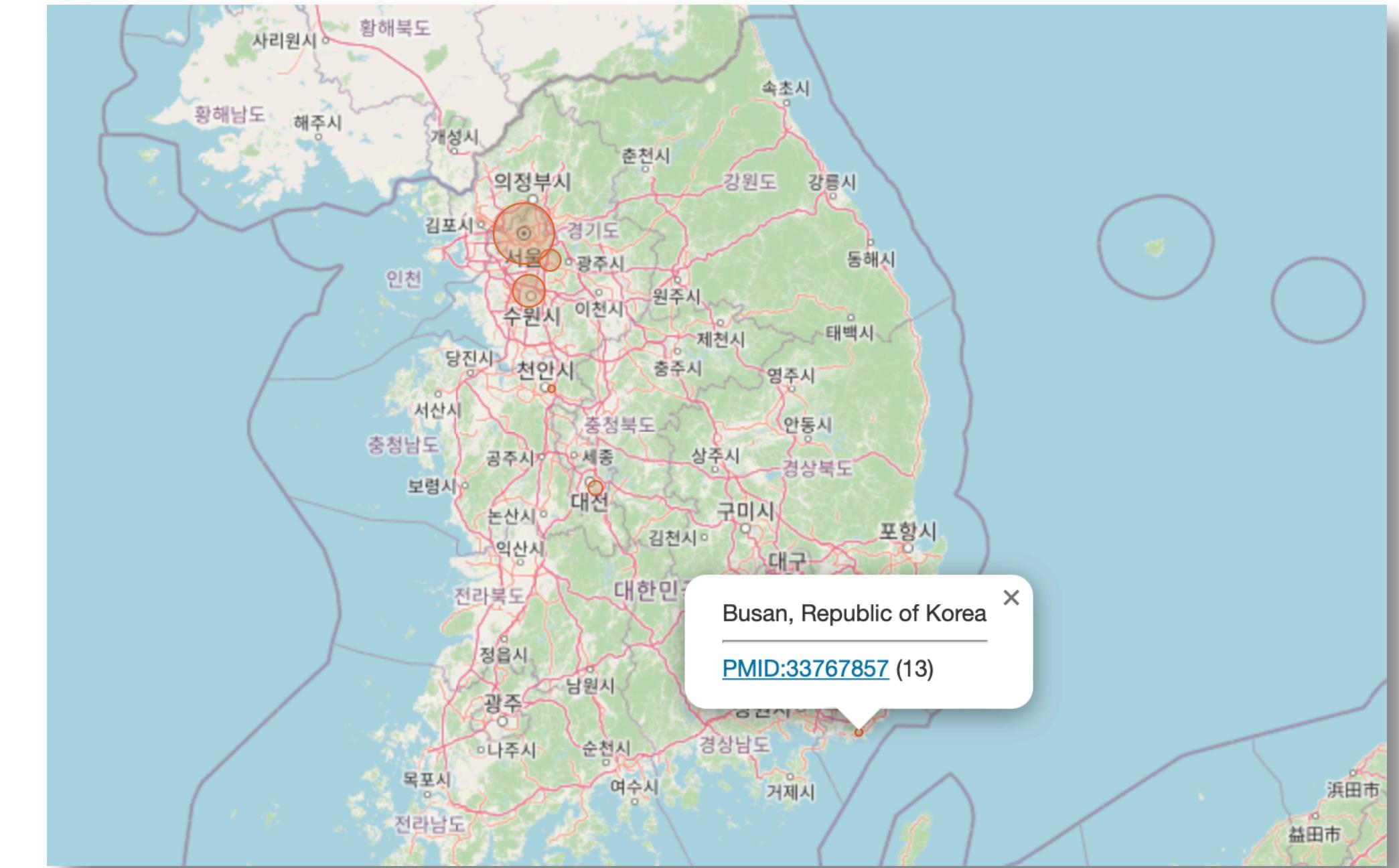
CC BY 4.0 progenetix.org (2021)

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

Service: Publications

Location Mapping for Statistics and Discovery...

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



[PMID:33767857](#) ↗

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

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

Mol Clin Oncol PMID:33767857 ↗

The Progenetix oncogenomic resource in 2021

Qingyao Huang^{1,2}, Paula Carrio-Cordo^{1,2}, Bo Gao^{1,2}, Rahel Paloots^{1,2} and Michael Baudis^{1,2,*}

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

²Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich 8057, Switzerland

*Corresponding author: Tel: +41 44 635 34 86; Email: michael.baudis@mls.uzh.ch

Citation details: Huang, Q., Carrio-Cordo, P., Gao, B. et al. The Progenetix oncogenomic resource in 2021. *Database* (2021) Vol. 2021: article ID baab043; DOI: <https://doi.org/10.1093/database/baab043>

Abstract

In cancer, copy number aberrations (CNAs) represent a type of nearly ubiquitous and frequently extensive structural genome variations. To disentangle the molecular mechanisms underlying tumorigenesis as well as identify and characterize molecular subtypes, the comparative and meta-analysis of large genomic variant collections can be of immense importance. Over the last decades, cancer genomic profiling projects have resulted in a large amount of somatic genome variation profiles, however segregated in a multitude of individual studies and datasets. The Progenetix project, initiated in 2001, curates individual cancer CNA profiles and associated metadata from published oncogenomic studies and data repositories with the aim to empower integrative analyses spanning all different cancer biologies. During the last few years, the fields of genomics and cancer research have seen significant advancement in terms of molecular genetics technology, disease concepts, data standard harmonization as well as data availability, in an increasingly structured and systematic manner. For the Progenetix resource, continuous data integration, curation and maintenance have resulted in the most comprehensive representation of cancer genome CNA profiling data with 138 663 (including 115 357 tumor) copy number variation (CNV) profiles. In this article, we report a 4.5-fold increase in sample number since 2013, improvements in data quality, ontology representation with a CNV landscape summary over 51 distinctive National Cancer Institute Thesaurus cancer terms as well as updates in database schemas, and data access including new web front-end and programmatic data access.

Database URL: progenetix.org

Table 1. Statistics of samples from various data resources

Data source	GEO	ArrayExpress	cBioPortal	TCGA	Total
No. of studies	898	51	38	33	1939
No. of samples	63 568	4351	19 712	22 142	138 663
Tumor	52 090	3887	19 712	11 090	115 357
Normal	11 478	464	0	11 052	23 306
Classifications					
ICD-O (Topography)	100	54	88	157	209
ICD-O (Morphology)	246	908	265	140	491
NCIt	346	148	422	182	788
Collections					
Individuals	63 568	4351	19 712	10 995	127 549
Biosamples	63 568	4351	19 712	22 142	138 663
Callsets ^a	63 568	4351	19 712	22 376	138 930
Variants	5 514 126	118 4170	1 778 096	2 654 065	10 716 093

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

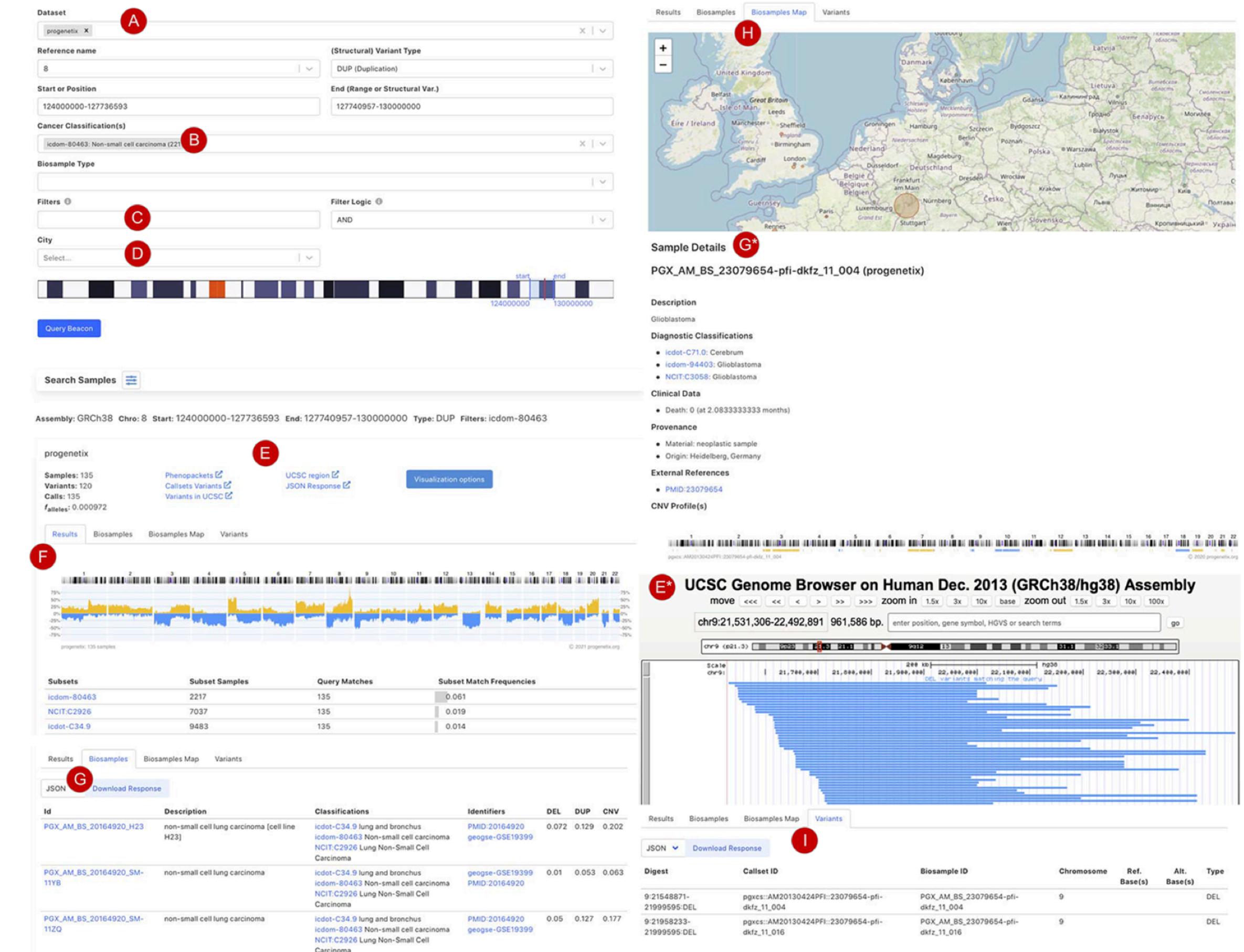


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range. This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with ≤ 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and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

Cancer Cell Lines

Cancer Genomics Reference Resource

- starting from >5000 cell line CNV profiles
 - 5754 samples | 2163 cell lines
 - 256 different NCIT codes
- genomic mapping of annotated variants and additional data from several resources (ClinVar, CCLE, Cellosaurus...)
 - 16178 cell lines
 - 400 different NCIT codes
- query and data delivery through Beacon v2 API
 - integration in data federation approaches

The screenshot shows the homepage of cancercelllines.org. At the top is a pink header bar. Below it is a navigation menu with the following items: Cancer Cell Lines (with a red circular icon), Search Cell Lines, Cell Line Listing, CNV Profiles by Cancer Type, Documentation, News, Progenetix (which is highlighted in grey), and Baudisgroup @ UZH.

cancercellines.org

Cancer Cell Lines by Cellosaurus ID

The cancer cell lines in [cancercelllines.org](#) are labeled by their parentage hierarchically: Daughter cell lines are displayed below the primary cell line. For example, HeLa is listed as a daughter cell line of **HeLa (CVCL_0030)** and so forth.

Sample selection follows a hierarchical system in which samples are retrieved based on the selected cell line. For example, selecting HOS for HeLa will also return the daughter lines by default - but one can also select the specific daughter line.

Cell Lines (with parental/derived hierarchies)

Filter subsets e.g. by prefix

Hierarchy Depth:

No Selection

- > cellosaurus:CVCL_0312: HOS (204 samples)
- > cellosaurus:CVCL_1575: NCI-H650 (6 samples)
- > cellosaurus:CVCL_1783: UM-UC-3 (9 samples)
- > cellosaurus:CVCL_0004: K-562 (28 samples)
- cellosaurus:CVCL_3827: K562/Adr (1 sample)
- > cellosaurus:CVCL_0589: Kasumi-1 (9 samples)
- > cellosaurus:CVCL_XK00: M397 (2 samples)
- > cellosaurus:CVCL_1650: Reh (11 samples)
- cellosaurus:CVCL_8857: EU-1 (1 sample)
- cellosaurus:CVCL_0011: KM-3 (1 sample)
- cellosaurus:CVCL_8462: NOI-90 (1 sample)
- cellosaurus:CVCL_ZV66: Reh/EphA2 (1 sample)
- cellosaurus:CVCL_A049: WSU-CLL (1 sample)
- > cellosaurus:CVCL_2063: HCC827 (27 samples)

Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

- Cellosaurus - a knowledge resource on cell lines [cellosaurus:CVCL_0312](#)

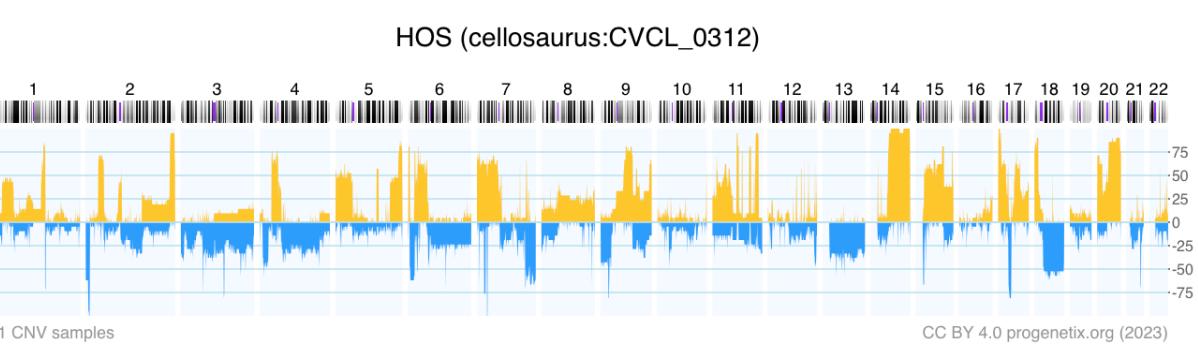
Sample Counts

- 204 samples
- 57 direct cellosaurus:CVCL_0312 code matches
- 21 CNV analyses

Search Samples

Select cellosaurus:CVCL_0312 samples in the [Search Form](#)

Raw Data (click to show/hide)



[Download SVG](#) | Go to cellosaurus:CVCL_0312 | [Download CNV Frequencies](#)

Gene Matches Cytoband Matches Variants

ALK	. ABC-14 cells harbored no ALK mutations and were sensitive to ... crizotinib while also exhibiting MNNG HOS transforming gene (MET)	Rapid Acquisition of Alectinib Resistance in ALK-Positive Lung Cancer With High Tumor Mutation Burden (31374369)	ABSTRACT
AREG	crizotinib while also exhibiting MNNG HOS	Rapid Acquisition of Alectinib Resistance	ABSTRACT

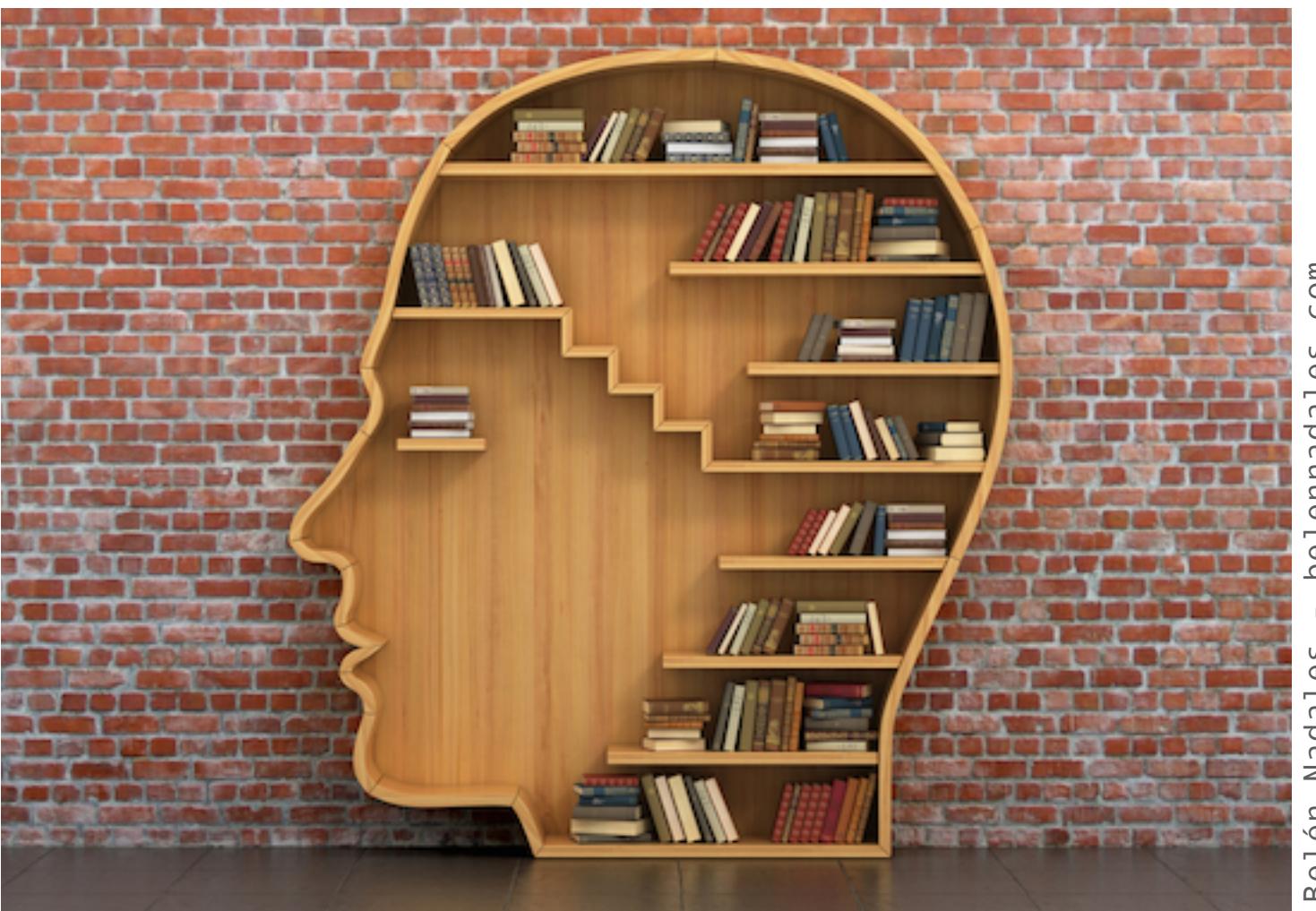
Data Parasites



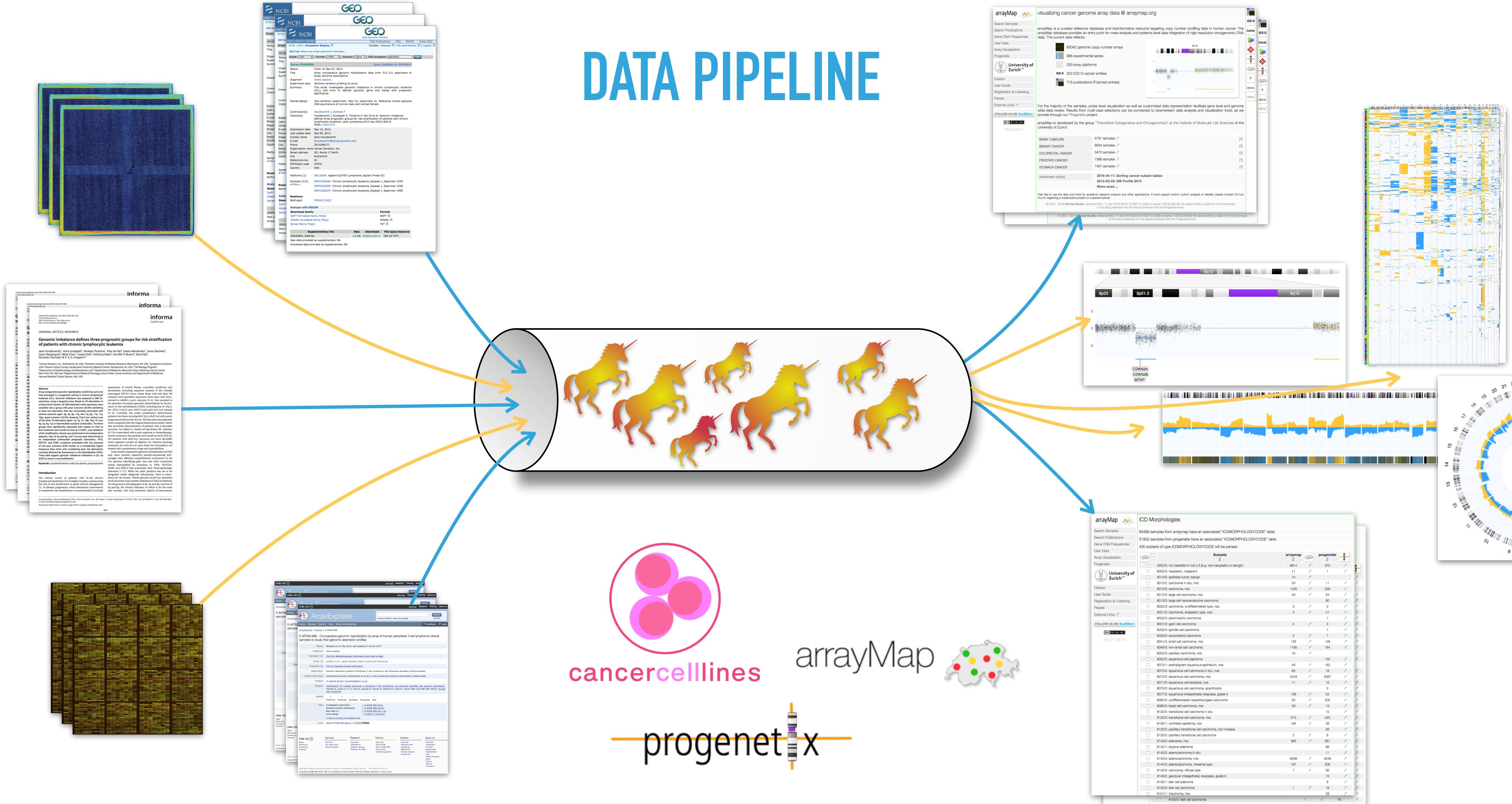
PARASITE

Data Parasites

... or let's call us CURATORS



DATA PIPELINE



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

ZIP/Postal Code: 8008 Zurich

City: Zurich

Country: Switzerland

Platform: Agilent G1317P Lymphoma (Agilent Probe ID)

Sample ID: GSE640034 Chronic lymphocytic leukemia, Dataset 1, Specimen 1949

Platform: Agilent G1317P Lymphoma (Agilent Probe ID)

Sample ID: GSE640034 Chronic lymphocytic leukemia, Dataset 1, Specimen 1949

Platform: Agilent G1317P Lymphoma (Agilent Probe ID)

Sample ID: GSE640034 Chronic lymphocytic leukemia, Dataset 1, Specimen 1999

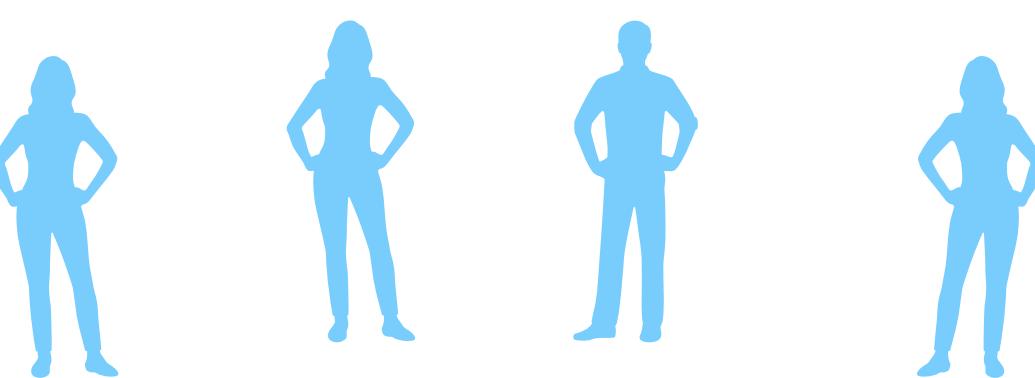
Supplementary file: GSE640034.RAW.tar Size: 1.4 Gb Download File type/resource: Tar (of TXT)

Raw data provided as supplementary file

Processed data provided as supplementary file



cancercelllines
progenetix



arrayMap

985 experimental series

333 array platforms

253 ICD-O cancer entities

716 publications (PubMed entries)

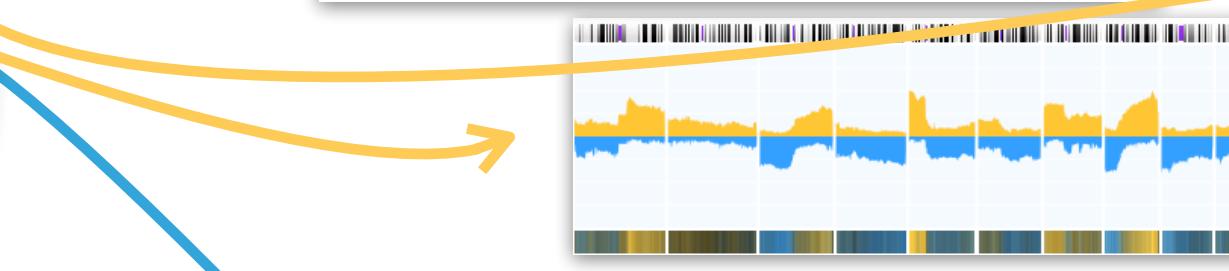
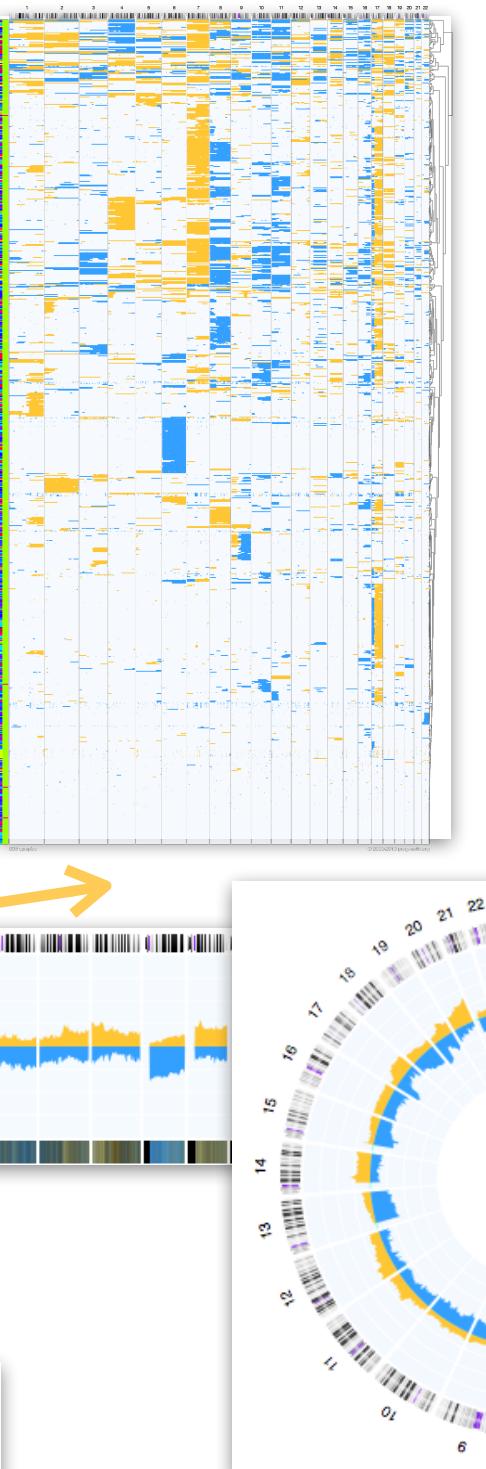
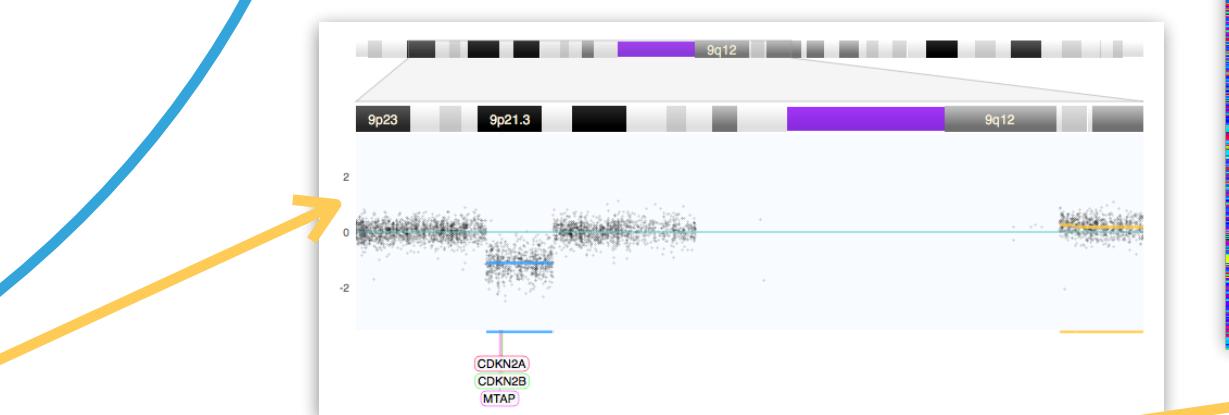
For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Prognetic project.

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

BRAIN TUMOUR: 5791 samples ✓
BREAST CANCER: 8594 samples ✓
COLORECTAL CANCER: 3470 samples ✓
PROSTATE CANCER: 1366 samples ✓
STOMACH CANCER: 1457 samples ✓

ARRAYMAP NEWS: 2016-04-11: Sorting cancer subset tables
2015-03-23: SIB Profile 2015
More news ...

© 2000 - 2016 Michael Baudis, refreshed Mon, 11 Apr 2016 09:54:12 GMT in 3.00s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Prognetic tools.



ICD Morphologies

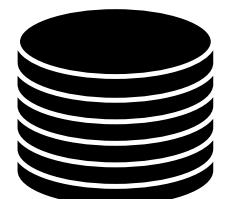
64485 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.
31922 samples from progneticx have an associated "ICDMORPHOLOGYCODE" label.
400 subsets of type ICDMORPHOLOGYCODE will be parsed.

	arraymap	progneticx
Subsets	6	6
00000: not classified in icd-3 (e.g. non-neoplastic or benign)	8614 ↗ 370	
00003: neoplasm, malignant	11 ↗ 1	
00100: epithelial tumor, benign	10 ↗ 1	
00102: carcinoma, nos	20 ↗ 11	
00120: large cell carcinoma, nos	1430 ↗ 258	
00200: squamous cell carcinoma, nos	46 ↗ 54	
00203: carcinoma, unendifferentiated type, nos	3 ↗ 2	
00210: adenocarcinoma, nos	4 ↗ 41	
00220: giant cell carcinoma	4 ↗ 3	
00300: small cell carcinoma	1 ↗ 1	
00330: sarcomatoid carcinoma	2 ↗ 7	
00410: anal cell carcinoma, nos	132 ↗ 148	
00500: mesothelioma, nos	1195 ↗ 184	
00503: pleural carcinoma, nos	16 ↗ 15	
00700: squamous cell carcinoma, nos	46 ↗ 162	
00702: squamous cell carcinoma, nos	65 ↗ 16	
00703: squamous cell carcinoma, nos	2443 ↗ 2087	
00707: squamous cell carcinoma, nos	11 ↗ 12	
00750: squamous cell carcinoma, nos	132 ↗ 22	
00800: differentiated/recapitulated carcinoma	52 ↗ 200	
00900: basal cell carcinoma, nos	28 ↗ 15	
01200: transitional cell carcinoma, nos	310 ↗ 423	
01300: papillary transitional cell carcinoma, non-invasive	184 ↗ 39	
01303: papillary transitional cell carcinoma	2 ↗ 6	
01400: squamous cell carcinoma, nos	385 ↗ 361	
01402: squamous cell carcinoma, nos	88 ↗ 89	
01403: adenocarcinoma, nos	11 ↗ 11	
01404: adenocarcinoma in situ	9469 ↗ 3248	
01443: adenocarcinoma, intestinal type	167 ↗ 206	
01450: carcinoma, diffuse type	7 ↗ 36	
01500: squamous cell carcinoma, nos	8 ↗ 15	
01501: adenocarcinoma	1 ↗ 18	
01502: squamous cell carcinoma	1 ↗ 28	
01511: insularoma, nos	1 ↗ 18	
01512: insulinoma, nos	29 ↗ 29	

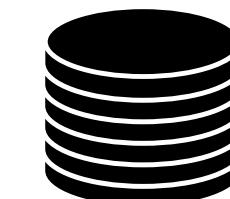
Progenetix Stack



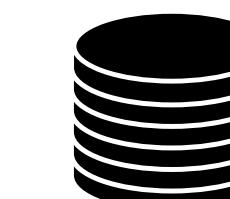
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the *bycon* package 
 - ▶ schemas,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



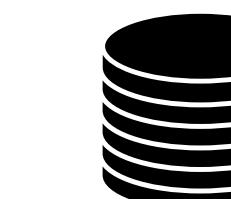
variants



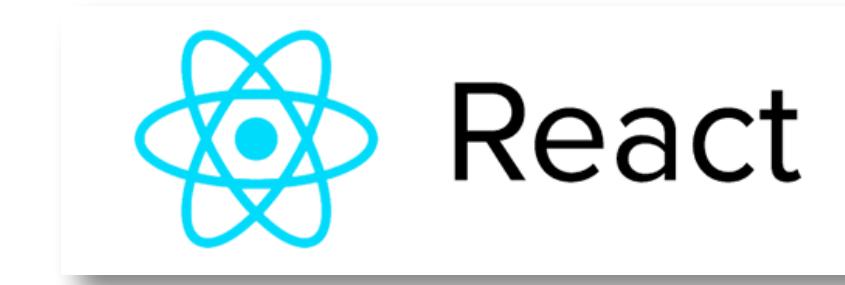
analyses



biosamples

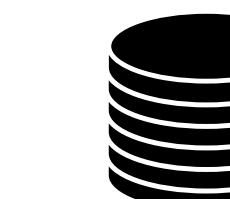


individuals

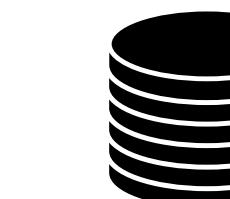


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and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

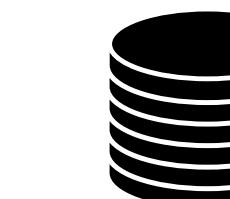
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
_id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



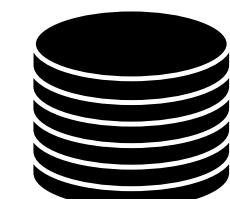
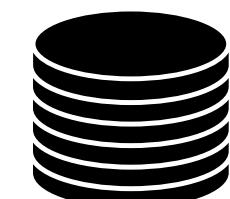
collations



geolocs



genespans publications



qBuffer

Entity collections

Utility collections



FITTING
THE MODEL

EVER
CLEANING
THE DATA

Data sets in tutorials



Data sets in the wild



Cancer Classifications need an Einstein to sort them out



BRADY'S NCI:038 NCI:BRADY'S MORPHOLOGY CODES
GSM393858 C2853 Acute Myeloid Leukemia Not Otherwise Specified 9861/3 C42
GSM302285 C2852 Adenocarcinoma 8140/3 C34
GSM18983 C3222 Medulloblastoma 9480/3 C716
GSM1551398 C4017 Ductal Breast Carcinoma 8500/3 C50
GSM1412384 C3163 Chronic Lymphocytic Leukemia 9823/3 C42
GSM1218286 C4017 Ductal Breast Carcinoma 8500/3 C50
GSM14412 C2852 Adenocarcinoma 8140/3 C569
GSM1109923 C9306 Soft Tissue Sarcoma 8800/3 C499
GSM11848 C2852 Adenocarcinoma 8140/3 C25
GSM246294 C89426 8022/2 C53
GSM1981528 C4017 Ductal Breast Carcinoma 8500/3 C50
GSM281399 C8949 8500/2 C50
GSM533469 C9349 Plasmacytoma 9831/3 C42



Ontologies and Classifications



Services: Ontologymaps (NCIt)

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

NCIT and ICD-O 3

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

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

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

Code Selection ⓘ

NCIT:C4337: Mantle Cell Lymphoma X | ▾

Optional: Limit with second selection | ▾

Matching Code Mappings [{JSON ↗}](#)

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

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

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

List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

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

Private filters

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

For terms with a `pgx` prefix, the [identifiers.org resolver](#) will

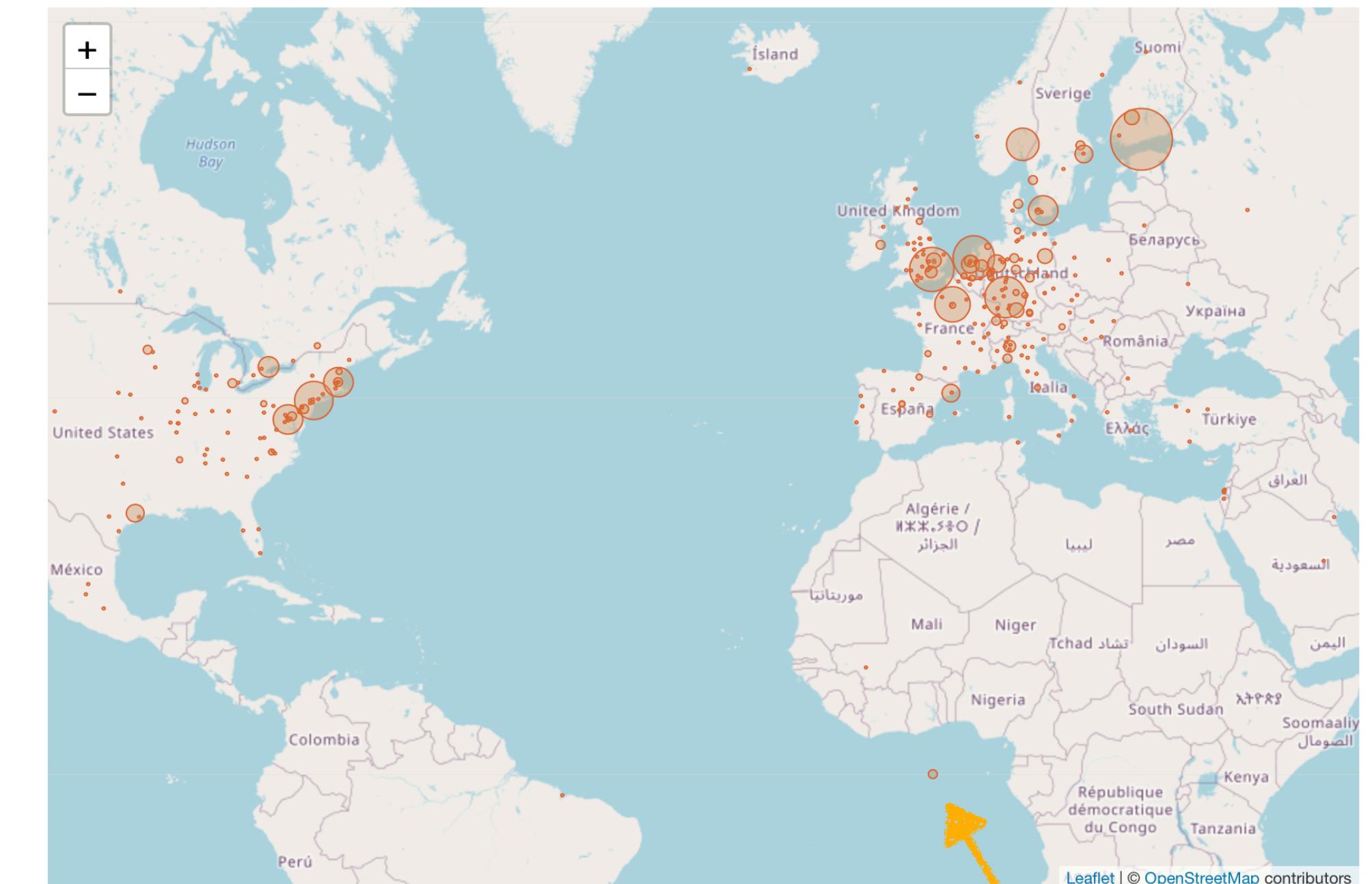
Filter prefix / local part	Code/Ontology	Example
pgx:icdom...	ICD-O 3 ⁷ Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3 ⁷ Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix) ⁸	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort...	Progenetix cohorts ¹⁰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

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
progenetix.org/publications/list
2020-11-28

25 / 3306
publications

Data Curation - Geolocations

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

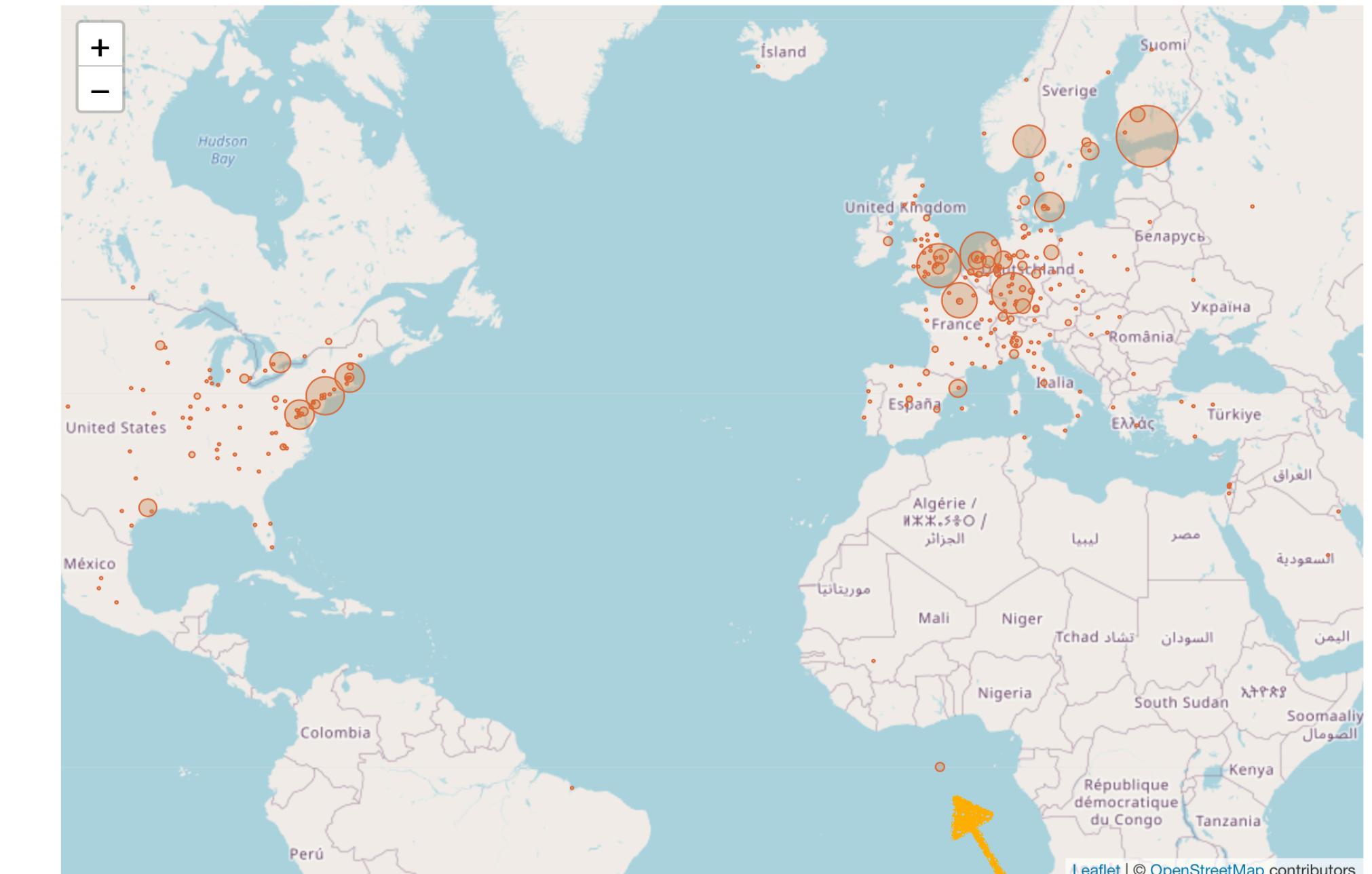
The most geo-tagged place on earth is Null Island



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

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

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



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

25 / 3306 publications

pgxRpi

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

- **Query and export variants**

https://progenetix.org/beacon/g_variants/?biosampleIds=pgxb-s-kftvh94d&output=pgxseg

```
> variants <- pgxLoader(type="variant", biosample_id="pgxb-s-kftvh94d", output = "pgxseg")
```

- **Query metadata of biosamples and individuals by filters (e.g. NCIt, PMID)**

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

```
> biosamples <- pgxLoader(type="biosample", filters="NCIT:C3697")
```

- **Query and visualize CNV frequency by filters**

<http://www.progenetix.org/services/intervalFrequencies/?filters=NCIT:C3512&output=pgxfreq>

```
> freq <- pgxLoader(type="frequency", output="pgxfreq", filters="NCIT:C3512")
> pgxFreqplot(freq)
```

- **Process local .pgxseg files**

```
> info <- pgxSegprocess(file=file, show_KM_plot = T,
  return_seg = T, return_metadata = T, return_frequency = T)
```

README.md

pgxRpi

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

You can install this package from GitHub using:

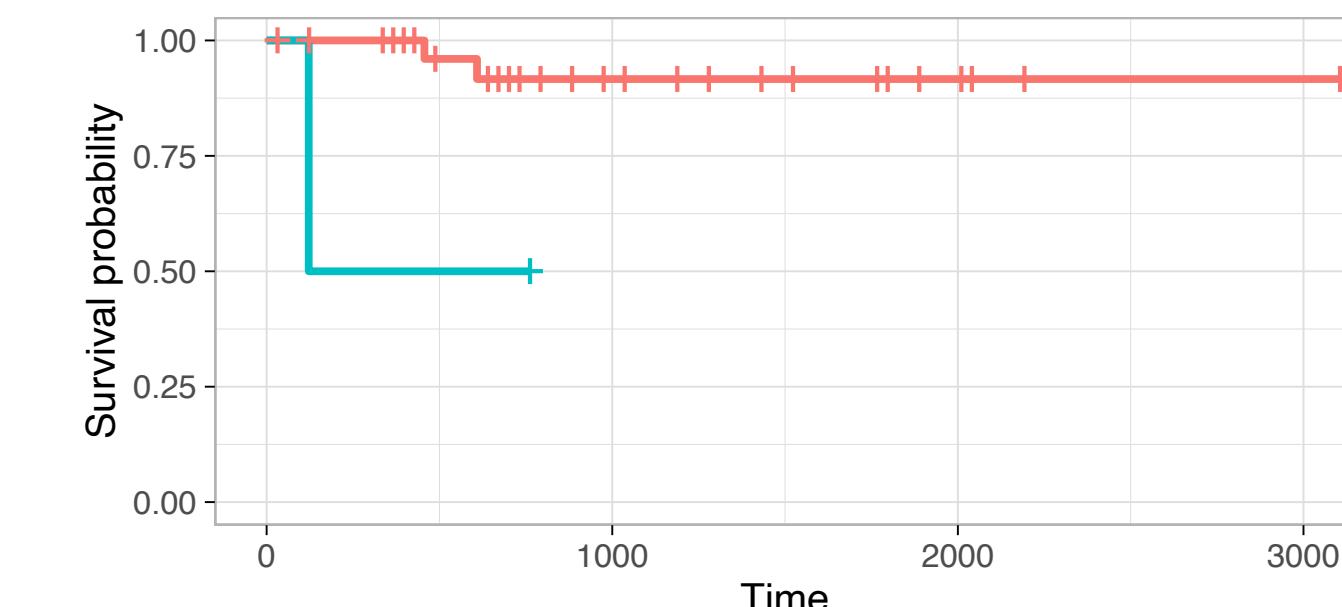
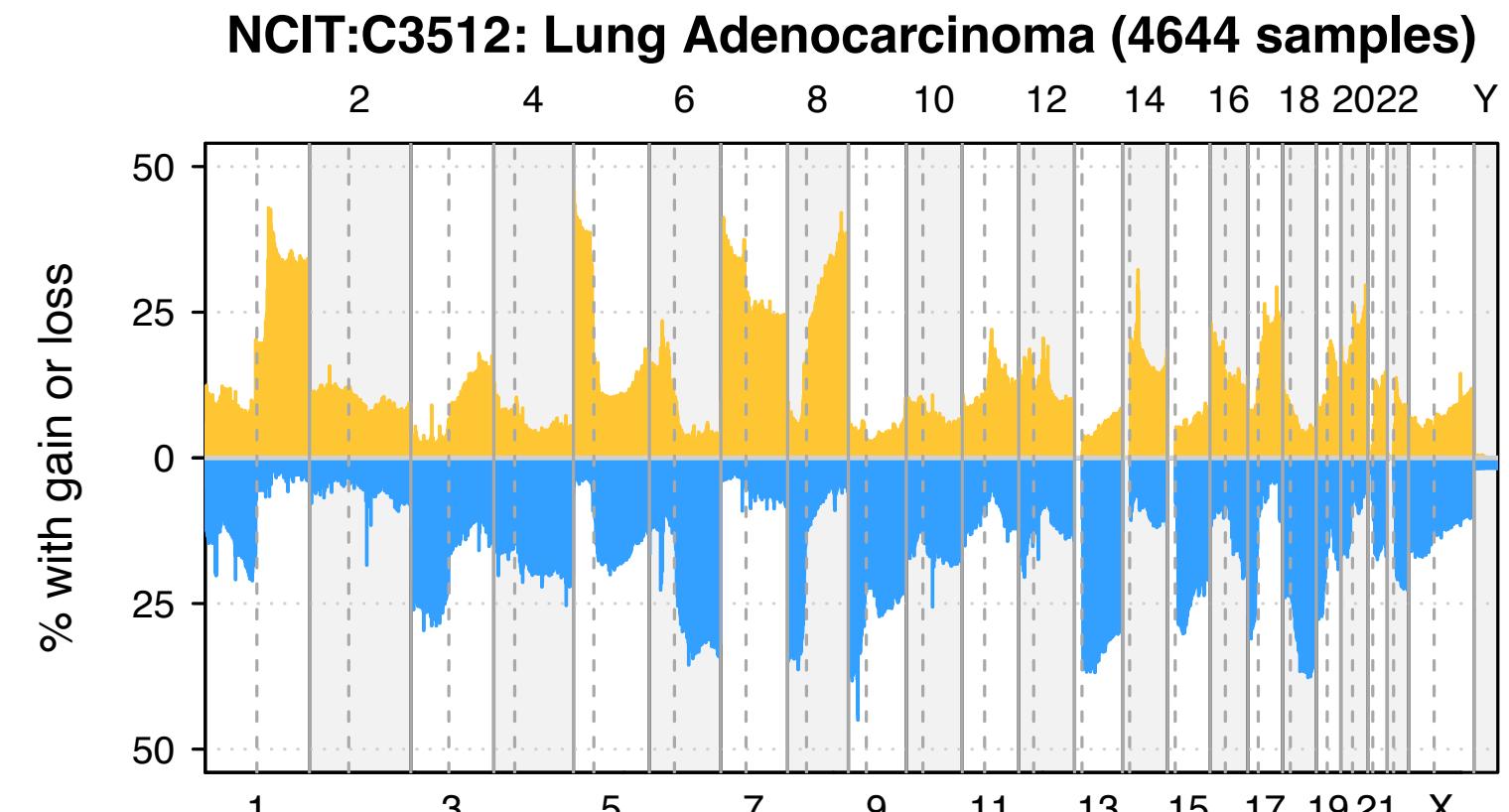
```
install.packages("devtools")
devtools::install_github("progenetix/pgxRpi")
```

If you are interested in accessing CNV variant data, get started from this [vignette](#)

If you are interested in accessing CNV frequency data, get started from this [vignette](#)

If you are interested in processing local pgxseg files, get started from this [vignette](#)

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



Standardized Data

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



xkcd

```
"label" : "no restriction",
"id" : "DUO:0000004"
},
"provenance" : {
  "material" : {
    "type" : {
      "id" : "EF0:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Zurich, Switzerland",
    "precision" : "city",
    "city" : "Zurich",
    "country" : "Switzerland",
    "latitude" : 47.37,
    "longitude" : 8.55,
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        8.55,
        47.37
      ]
    },
    "ISO-3166-alpha3" : "CHE"
  }
},
{
  "age" : "P25Y3M2D"
```



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

GA4GH Standards Development



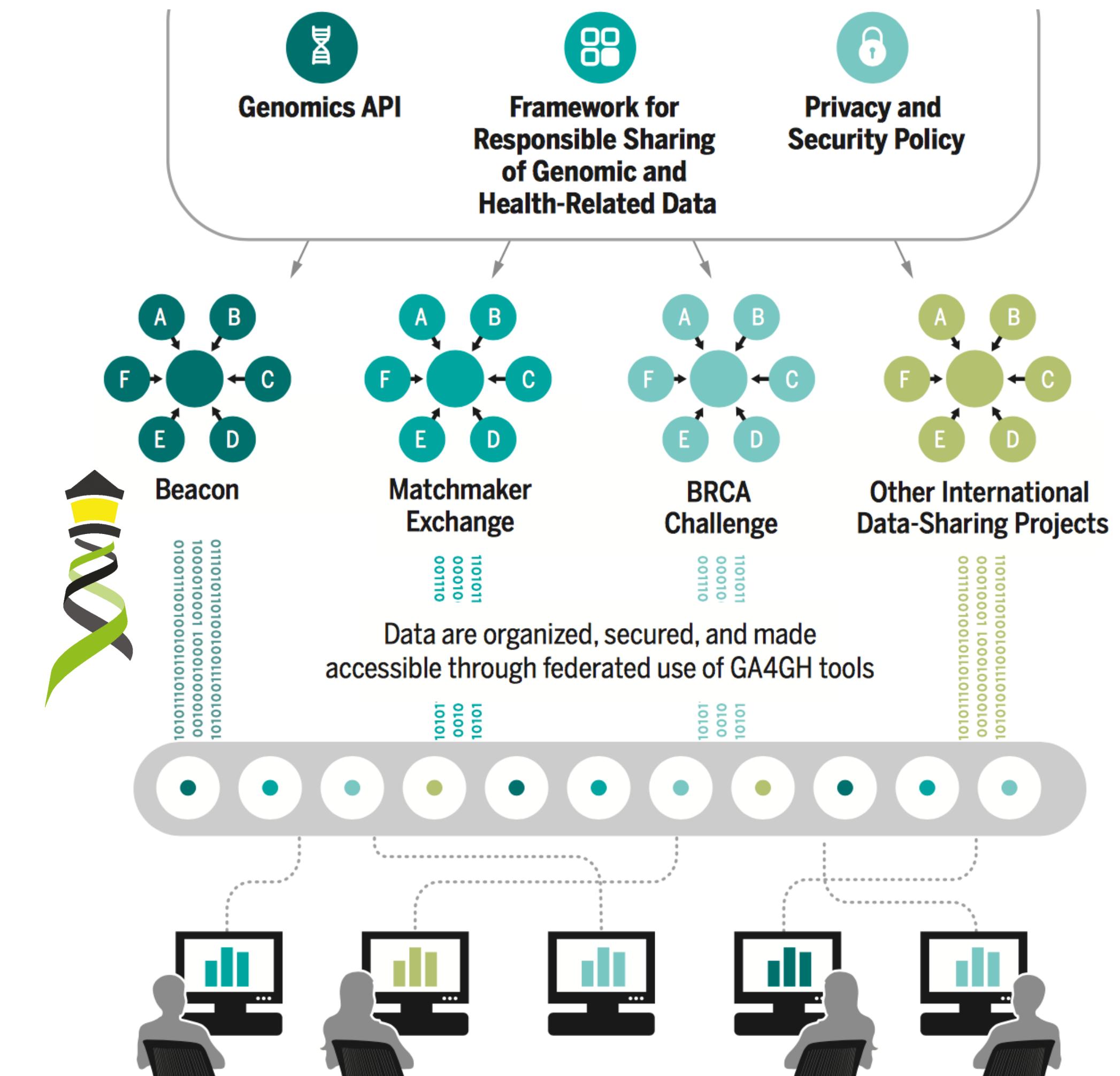


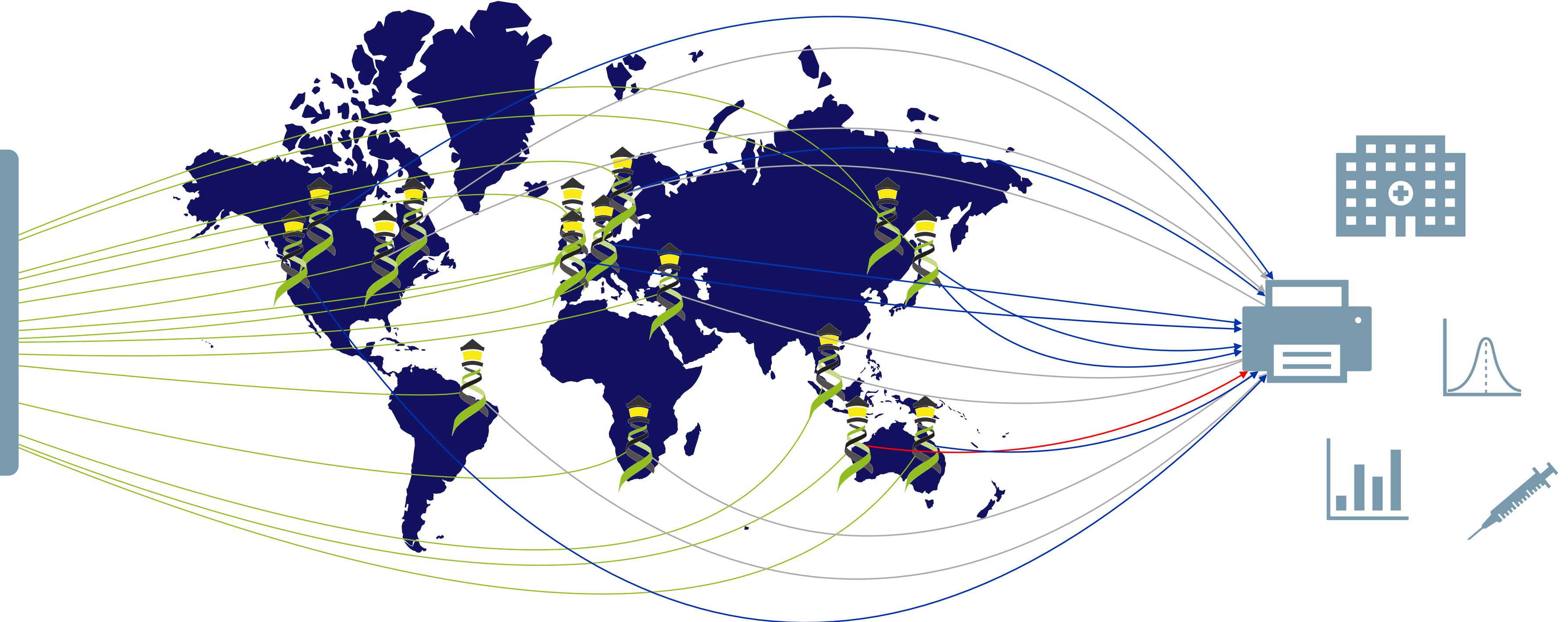
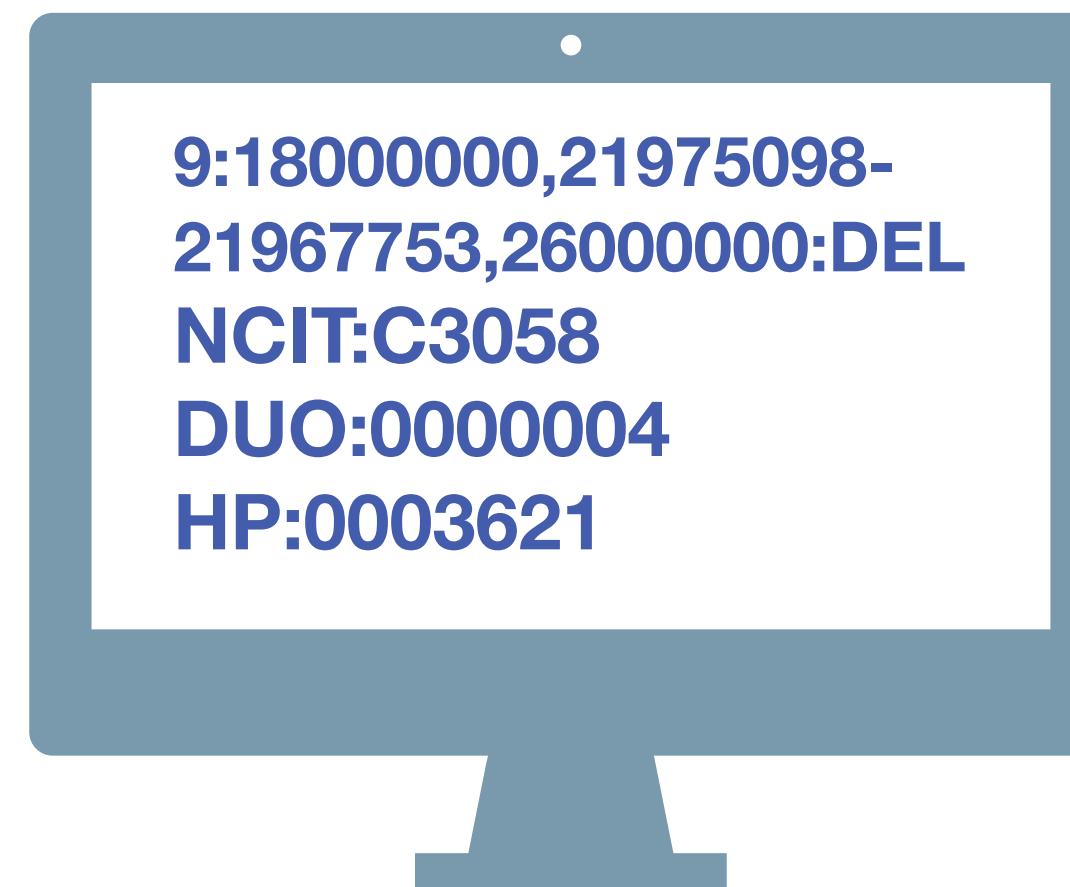
GENOMICS

A federated ecosystem for sharing genomic, clinical data

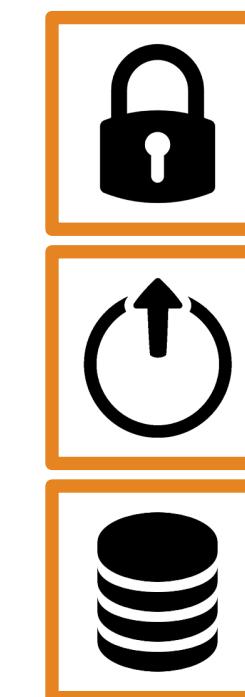
Silos of genome data collection are being transformed into seamlessly connected, independent systems

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





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".

Onboarding v2

Driving Compliance

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

Beacon v2 GA4GH Approval Registry

Beacons:    

European Genome-Phenome Archive (EGA)

GA4GH Approval Beacon Test

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

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

progenetix

Visit us 

Beacon UI 

Beacon API 

Contact us 

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

Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

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

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

cnag

Beacon @ RD-Connect

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

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

University of Leicester

Cafe Variome Beacon v2

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

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

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



The GA4GH Phenopackets v2 Standard

A Computable Representation of Clinical Data

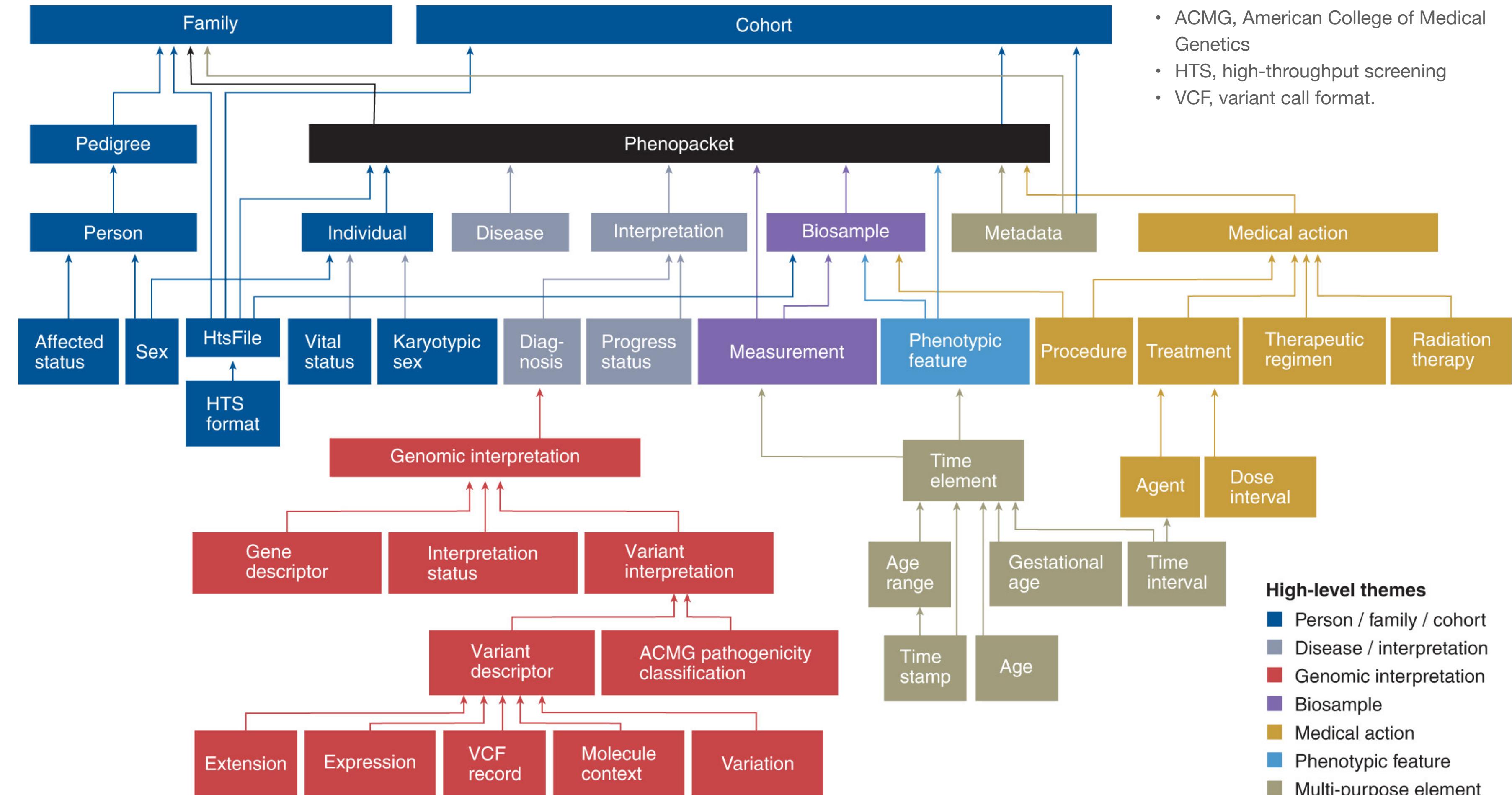


The GA4GH Phenopacket schema consists of several optional elements, each containing information about a certain topic, such as phenotype, variant or pedigree. An element can contain other elements, which allows a hierarchical representation of data.

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

Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, Callahan TJ, et al. 2022.

The GA4GH Phenopacket Schema Defines a Computable Representation of Clinical Data.
Nature Biotechnology 40 (6): 817–20.



The GA4GH Phenopackets v2 Standard

A Computable Representation of Clinical Data



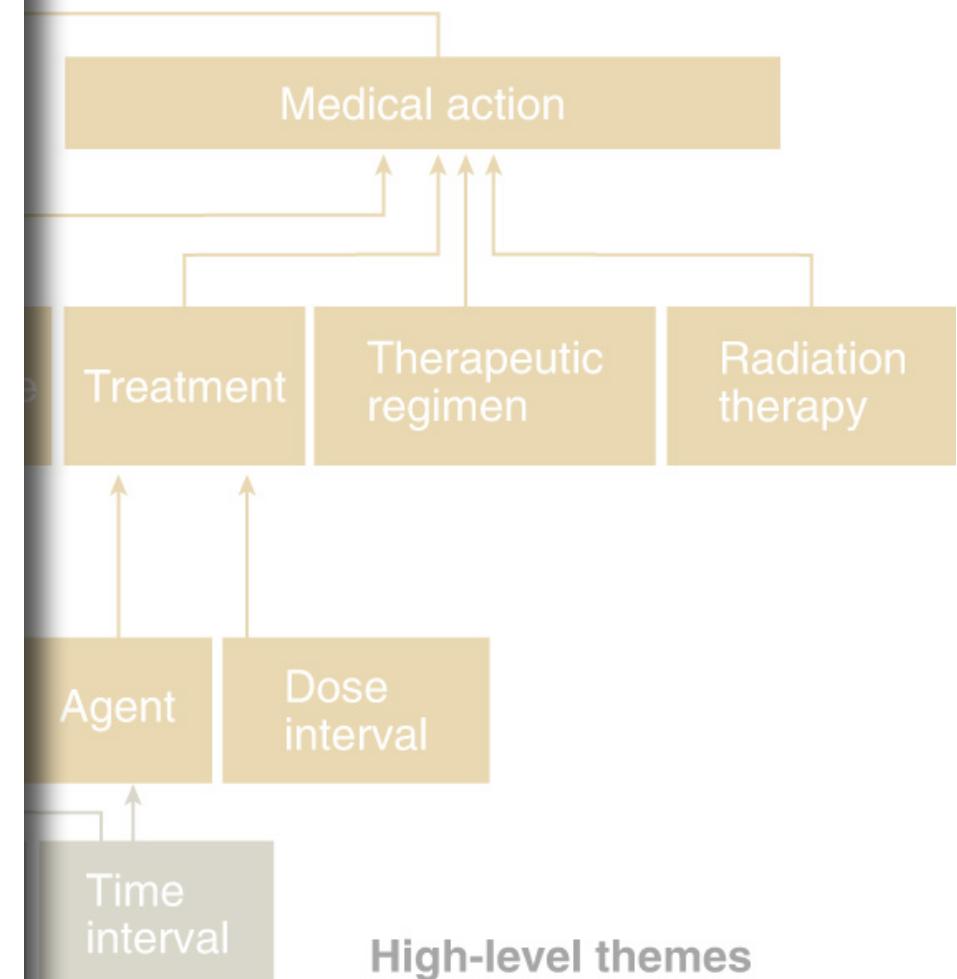
The GA4GH Phenopacket schema consists of several optional elements, each containing information about a certain aspect such as phenotype, variation, pedigree. An element can contain other elements, which allows for hierarchical representation. For instance, Phenopackets contains elements of type *Individual*, *PhenotypicFeature*, *Biosample* and so on. Individual elements can therefore be regarded as **building blocks** for larger structures.

Jacobsen JOB, Baudis M, Beckmann JS, Beltran S, Bussey TJ, et al. 2022.

The GA4GH Phenopacket Schema: A Computable Representation of Clinical Data.
Nature Biotechnology 40 (6): 817–20.

The screenshot shows a web page for ISO 4454:2022. At the top, there's a navigation bar with tabs for 'Family' (which is selected), 'Cohort', 'Standards', 'About us', 'News', 'Taking part', and 'Store'. Below the navigation is a search bar with a magnifying glass icon. The main content area has a breadcrumb trail: ← ICS ← 35 ← 35.240 ← 35.240.80. The title is 'ISO 4454:2022 Genomics informatics — Phenopackets: A format for phenotypic data exchange'. To the right of the title is a 'Buy this standard' button with a price of CHF 198. Below the title, there's an 'Abstract' section with a 'Preview' button. The abstract describes the document as a uniform, machine-readable, phenotypic description of an individual, patient or sample in the context of rare disease, common/complex disease or cancer. It also states that it is applicable to academic, clinical and commercial research, as well as clinical diagnostics. The 'General information' section includes fields for Status (Published), Publication date (2022-07), Edition (1), Number of pages (86), and Technical Committee (ISO/TC 215/SC 1 Genomics Informatics). At the bottom of the page are five red buttons labeled 'Extension', 'Expression', 'VCF record', 'Molecule context', and 'Variation'.

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



- High-level themes**
- Person / family / cohort
 - Disease / interpretation
 - Genomic interpretation
 - Biosample
 - Medical action
 - Phenotypic feature
 - Multi-purpose element

Maintaining some Standards

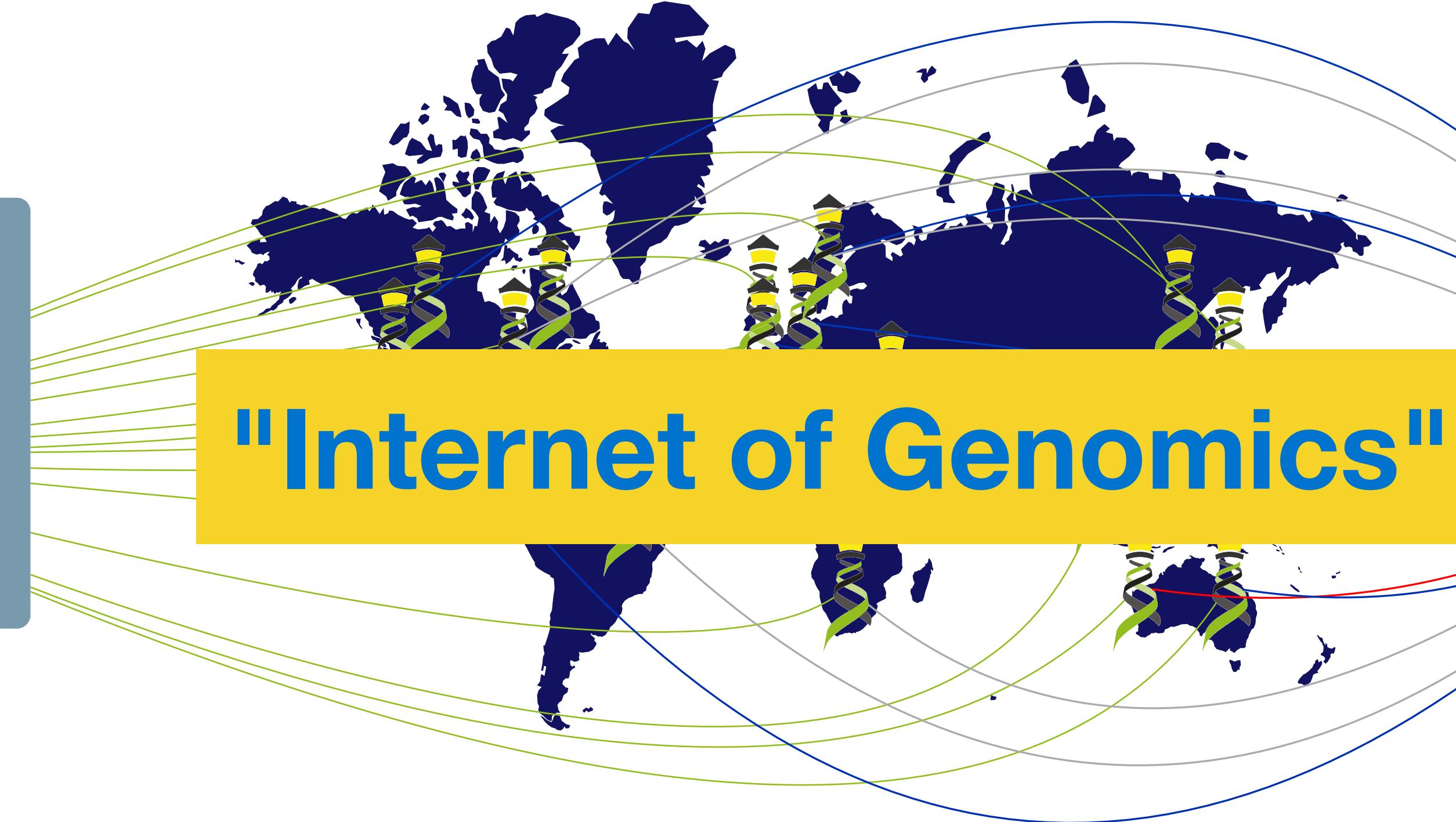
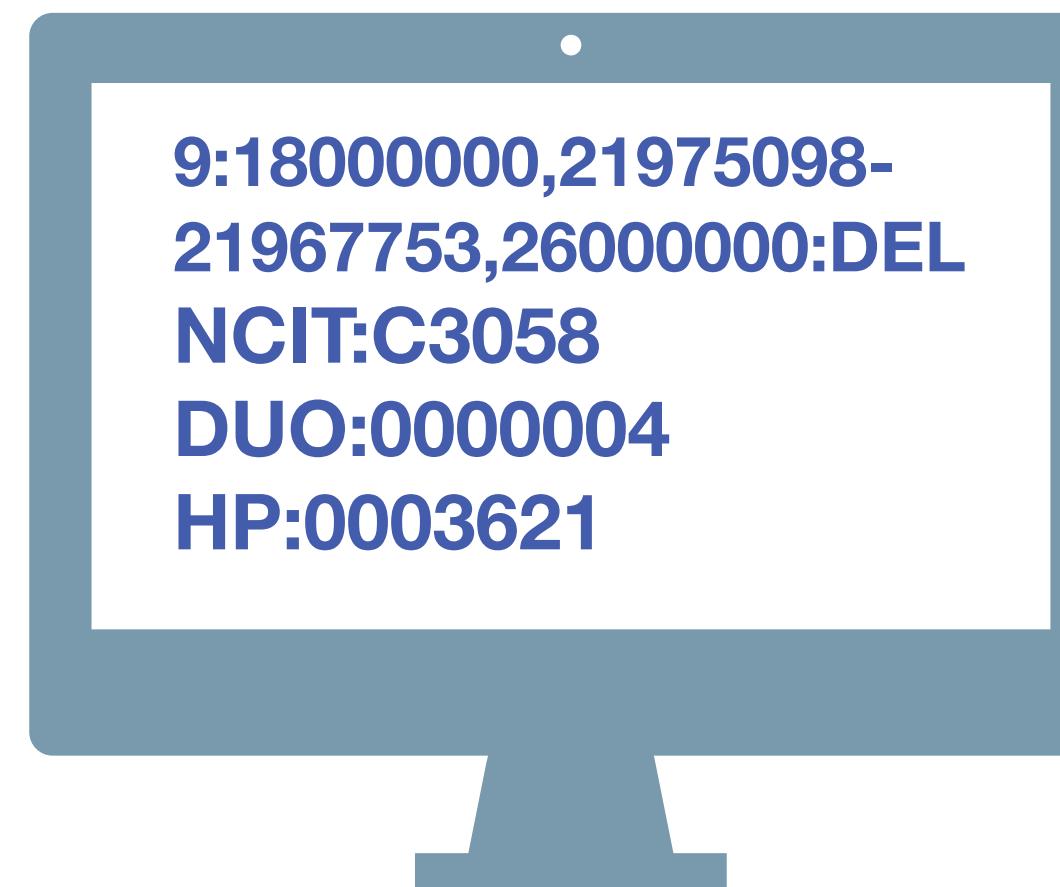
CNV Term Use in Computational (File/Schema) Formats



<https://cnvar.org/resources/CNV-annotation-standards/>

EFO	Beacon	VCF	SO	GA4GH VRS ⇒ VRS proposal ¹	Notes
EFO:0030070 copy number gain	DUP ² or EFO:0030070	DUP	SO:0001742 copy_number_gain	low-level gain (implicit) ⇒ EFO:0030070 copy number gain	a sequence alteration whereby the copy number of a given genomic region is greater than the reference sequence
EFO:0030071 low-level copy number gain	DUP ² or EFO:0030071	DUP	SO:0001742 copy_number_gain	low-level gain ⇒ EFO:0030071 low-level copy number gain	
EFO:0030072 high-level copy number gain	DUP ² or EFO:0030072	DUP	SO:0001742 copy_number_gain	high-level gain ⇒ EFO:0030072 high-level copy number gain	commonly but not consistently used for >=5 copies on a bi-allelic genome region
EFO:0030073 focal genome amplification	DUP ² or EFO:0030073	DUP	SO:0001742 copy_number_gain	high-level gain ⇒ EFO:0030073 focal genome amplification	commonly but not consistently used for >=5 copies on a bi-allelic genome region, of limited size (operationally max. 1-5Mb)
EFO:0030067 copy number loss	DEL ² or EFO:0030067	DEL	SO:0001743 copy_number_loss	partial loss (implicit) ⇒ EFO:0030067 copy number loss	a sequence alteration whereby the copy number of a given genomic region is smaller than the reference sequence
EFO:0030068 low-level copy number loss	DEL ² or EFO:0030068	DEL	SO:0001743 copy_number_loss	partial loss ⇒ EFO:0030068 low-level copy number loss	
EFO:0020073 high-level copy number loss	DEL ² or EFO:0020073	DEL	SO:0001743 copy_number_loss	partial loss ⇒ EFO:0020073 high-level copy number loss	a loss of several copies; also used in cases where a complete genomic deletion cannot be asserted
EFO:0030069 complete genomic deletion	DEL ² or EFO:0030069	DEL	SO:0001743 copy_number_loss	complete loss ⇒ EFO:0030069 complete genomic deletion	complete genomic deletion (e.g. homozygous deletion on a bi-allelic genome region)

Hangjia Zhao
Michael Baudis
(& the VRS group!)



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".

Recent Publications Standards & Resources

- GA4GH & ELIXIR
- Standards & reference implementations for data sharing and standards in genomics and personalized health

Cell Genomics

Perspective

GA4GH: International policies and standards for data sharing across genomic research and health

Heidi L. Rehm,^{1,2,47} Angela J.H. Page,^{1,3,*} Lindsay Smith,^{3,4} Jeremy B. Adams,^{3,4} Gil Alterovitz,^{5,47} Lawrence J. Maxmillian P. Barkley,⁶ Michael Baudis,^{7,8} Michael J.S. Beauvais,^{9,8} Tim Beck,¹⁰ Jacques S. Beckmann,¹¹ Sergi Beltran,^{12,13,14} David Bernick,¹ Alexander Bernier,⁹ James K. Bonfield,¹⁵ Tiffany F. Boughtwood,^{16,17} Guillaume Bourque,^{9,18} Sarion R. Bowers,¹⁵ Anthony J. Brookes,¹⁰ Michael Brudno,^{18,19,20,21,30} Matthew H. Brudno,^{18,38} Tony Burdett,²³ Orion J. Buske,²⁴ Moran N. Cabili,¹ Daniel L. Cameron,^{25,26} Robert J. Carroll,¹ Esmeralda Casas-Silva,¹²³ Debyani Chakravarty,²⁹ Bimal P. Chaudhari,^{30,31} Shu Hui Chen,³² J. Michael Cherry,¹ Justina Chung,^{3,4} Melissa Cline,³⁴ Hayley L. Clissold,¹⁵ Robert M. Cook-Deegan,³⁵ Mélanie Courtoot,²³ Fiona Cunningham,²³ Miro Cupak,⁶ Robert M. Davies,¹⁵ Danielle Denisko,¹⁹ Megan J. Doerr,³⁶ Lena I. Dolman,¹



ANALYSIS

<https://doi.org/10.1038/s41588-020-0603-8>

OPEN

A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer

Alex H. Wagner^①, Brian Walsh^②, Georgia Mayfield², David Tamborero^{3,4}, Dmitriy Sonkin^⑤, Kilannin Krysiak^①, Jordi Deu-Pons^{6,7}, Ryan P. Duren⁸, Jianjiong Gao^⑨, Julie McMurry², Sara Patterson¹⁰, Catherine del Vecchio Fitz¹¹, Beth A. Pitel¹², Ozman U. Sezerman¹³, Kyle Elliott², Jeremy L. Warner^⑭, Damian T. Rieke^⑮, Tero Aittokallio^{⑯,⑰}, Ethan Cerami¹¹, Deborah I. Ritter^{⑯,⑲}, Lynn M. Schriml²⁰, Robert R. Freimuth^⑫, Melissa Haendel^{⑬,⑳}, Gordana Raca^{22,23}, Subha Madhavan²⁴, Michael Baudis²⁵, Jacques S. Beckmann^⑳, Rodrigo Dienstmann²⁷, Debyani Chakravarty⁹, Xuan Shirley Li⁸, Susan Mockus^⑩, Olivier Elemento²⁸, Nikolaus Schultz⁹, Nuria Lopez-Bigas^{3,6,7}, Mark Lawler²⁹, Jeremy Goecks², Malachi Griffith^{①,⑳}, Obi L. Griffith^{①,⑳}, Adam A. Margolin² and Variant Interpretation for Cancer Consortium^{*}

Cell Genomics

Commentary

International federation of genomic medicine databases using GA4GH standards

Adrian Thorogood,^{1,2,*} Heidi L. Rehm,^{3,4} Peter Goodhand,^{5,6} Angela J.H. Page,^{4,5} Yann Joly,² Michael Baudis,⁷ Jordi Rambla,^{8,9} Arcadi Navarro,^{8,10,11,12} Tommi H. Nyronen,^{13,14} Mikael Linden,^{13,14} Edward S. Dove,¹⁵ Marc Fiume,¹⁶ Michael Brudno,¹⁷ Melissa S. Cline,¹⁸ and Ewan Birney¹⁹

Roadmap | Published: 27 August 2019

Leveraging European infrastructures to access 1 million human genomes by 2022

Gary Saunders, Michael Baudis, Regina Becker, Sergi Beltran, C. Brooksbank, Søren Brunak, Marc Van den Bulcke, Rachel Drysdale, Paul Flicek, Francesco Florindi, Peter Goodhand, Ivo Gut, Jaap Heringa, Juty, Thomas M. Keane, Jan O. Korbel, Ilkka Lappalainen, Brane Mayrhofer, Andres Metspalu, Arcadi Navarro, Steven Newhouse, Per Persson, Aarno Palotie, Helen Parkinson, Jordi Rambla, David Salter, Swertz, Alfonso Valencia, Susheel Varma, Niklas Blomberg & Serena Scollen

— Show fewer authors

Nature Reviews Genetics 20, 693–701 (2019) | [Cite this article](#)

ADVANCED GENETICS

Open Access

Research Article | Open Access |

GA4GH Phenopackets: A Practical Introduction

Markus S. Ladewig, Julius O. B. Jacobsen, Alex H. Wagner, Daniel Danis, Baha El Kassaby, Michael Gargano, Tudor Groza, Michael Baudis, Robin Steinhaus, Dominik Seelow ... See all authors

First published: 25 Aug

Correspondence | Published: 15 June 2022

The GA4GH Phenopacket schema defines a computable representation of clinical data

Julius O. B. Jacobsen , Michael Baudis, Gareth S. Baynam, Jacques S. Beckmann, Sergi Beltran, Orion J. Buske, Tiffany J. Callahan, Christopher G. Chute, Mélanie Courtot, Daniel Danis, Olivier Elemento, Andrea Essewanger, Robert R. Freimuth, Michael A. Gargano, Tudor Groza, Ada Hamosh, Nomi L. Harris, Rajaram Kaliyaperumal, Kevin C. Kent Lloyd, Aly Khalifa, Peter M. Krawitz, Sebastian Köhler,

Cell Genomics

CellPress
OPEN ACCESS

Technology

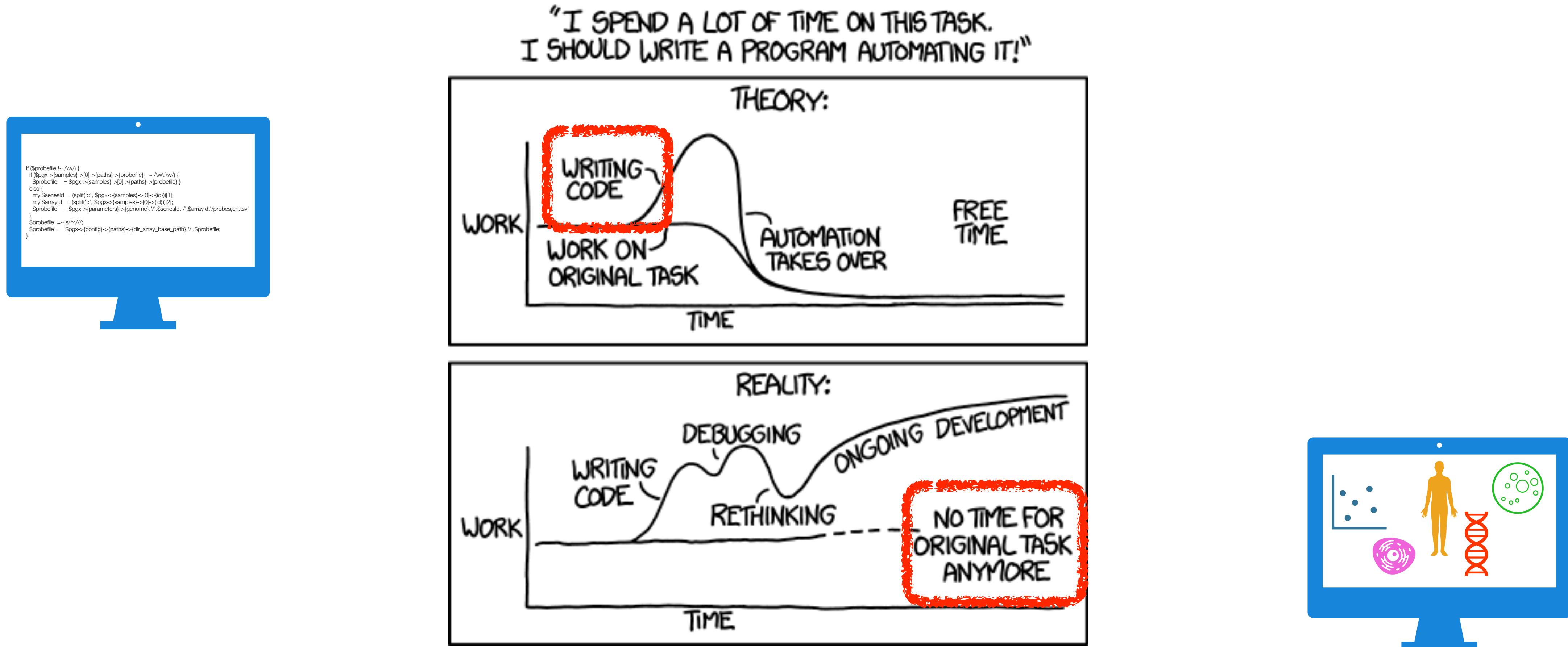
The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification

Alex H. Wagner,^{1,2,25,*} Lawrence Babb,^{3,*} Gil Alterovitz,^{4,5} Michael Baudis,⁶ Matthew Brush,⁷ Daniel L. Cameron,^{8,9} Melissa Cline,¹⁰ Malachi Griffith,¹¹ Obi L. Griffith,¹¹ Sarah E. Hunt,¹² David Kreda,¹³ Jennifer M. Lee,¹⁴ Stephanie Li,¹⁵ Javier Lopez,¹⁶ Eric Moyer,¹⁷ Tristan Nelson,¹⁸ Ronak Y. Patel,¹⁹ Kevin Riehle,¹⁹ Peter N. Robinson,²⁰ Shawn Rynearson,²¹ Helen Schuilenburg,¹² Kirill Tsukanov,¹² Brian Walsh,⁷ Melissa Konopko,¹⁵ Heidi L. Rehm,^{3,22} Andrew D. Yates,¹² Robert R. Freimuth,²³ and Reece K. Hart^{3,24,*}

{bio_informatics_science}



{bio_informatics_science}



Population stratification in cancer samples based on SNP array data

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

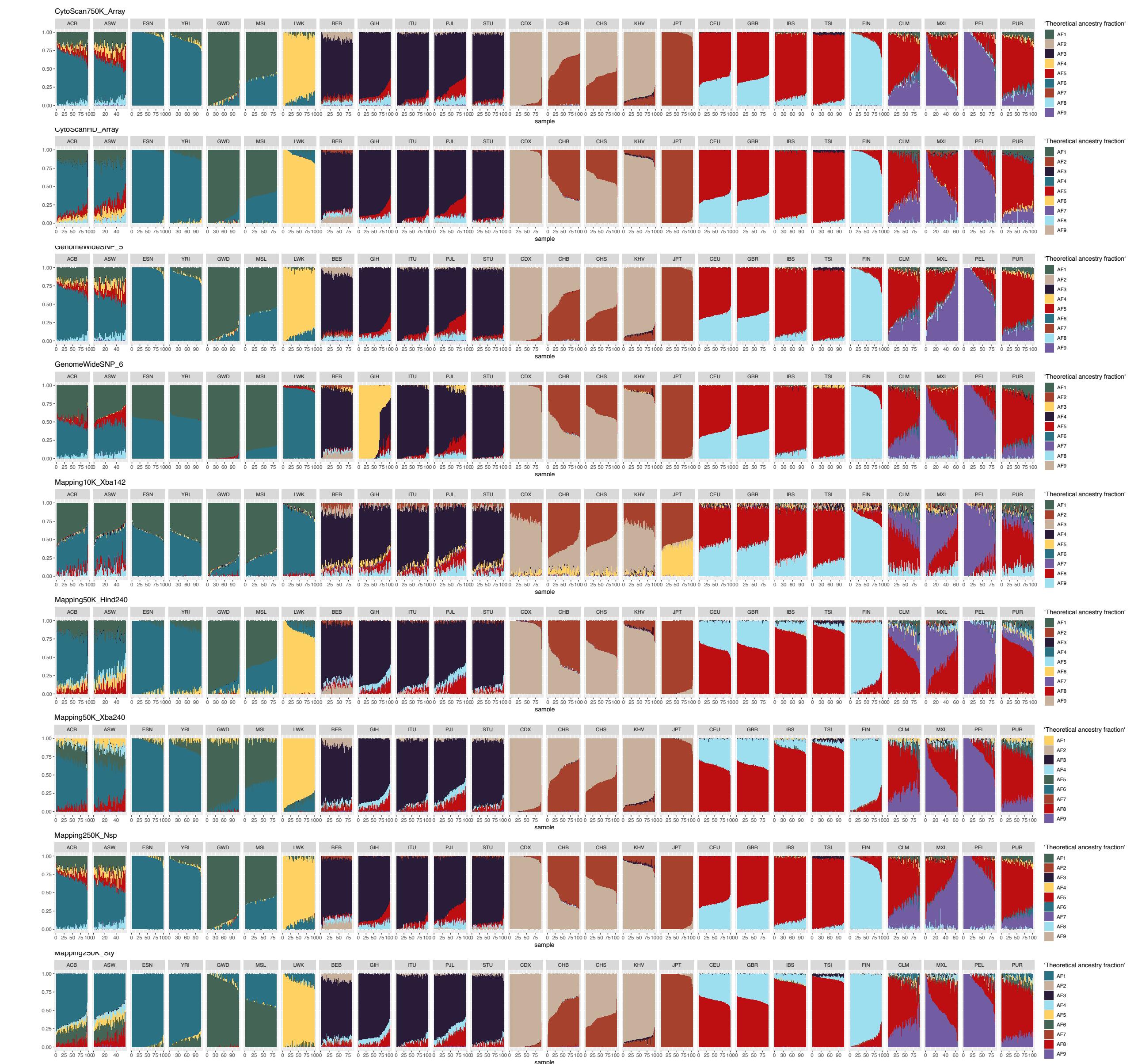
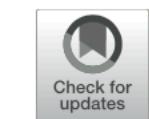


Figure S1 The fraction or contribution of theoretical ancestors ($k=9$) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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



Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes

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

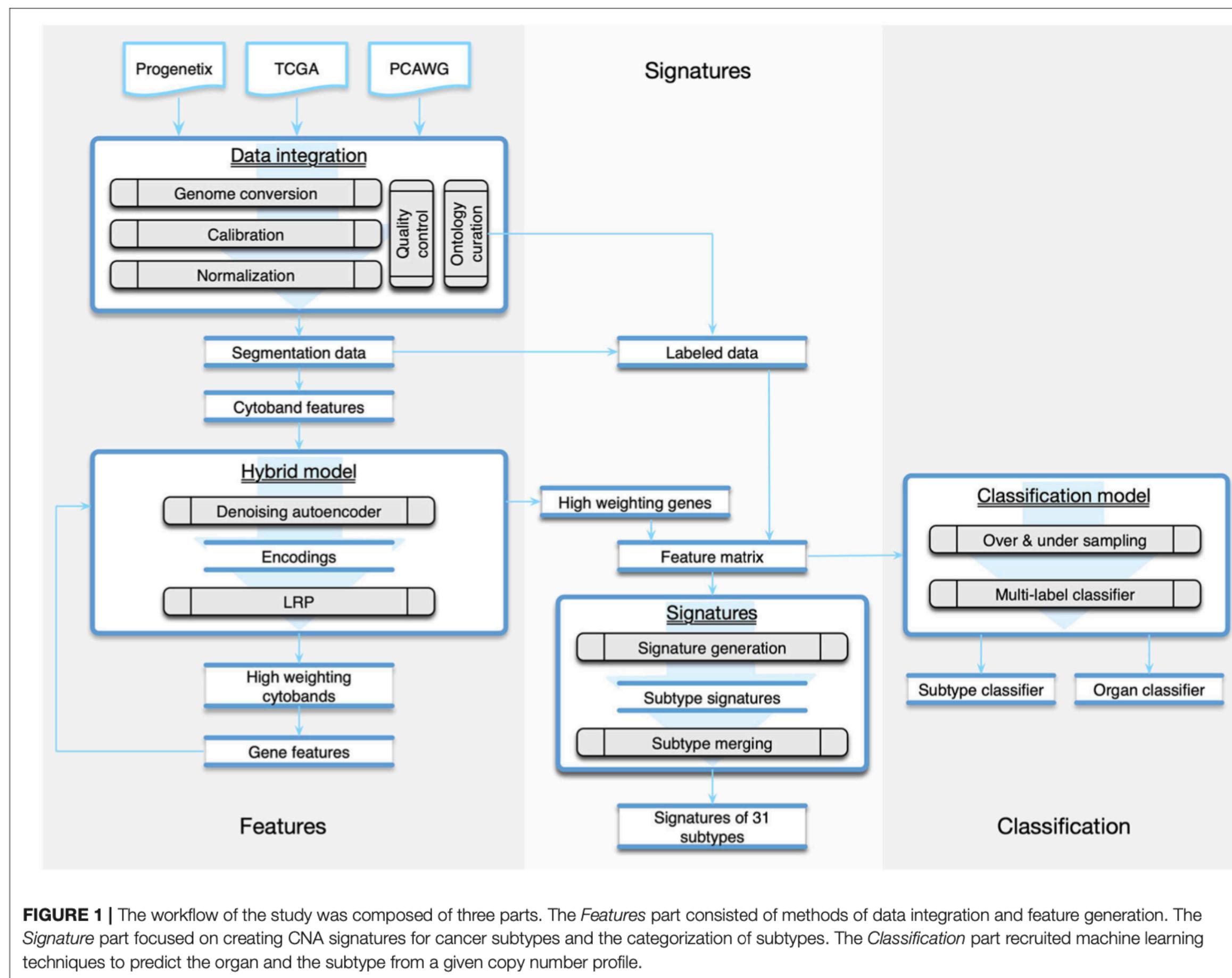


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine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

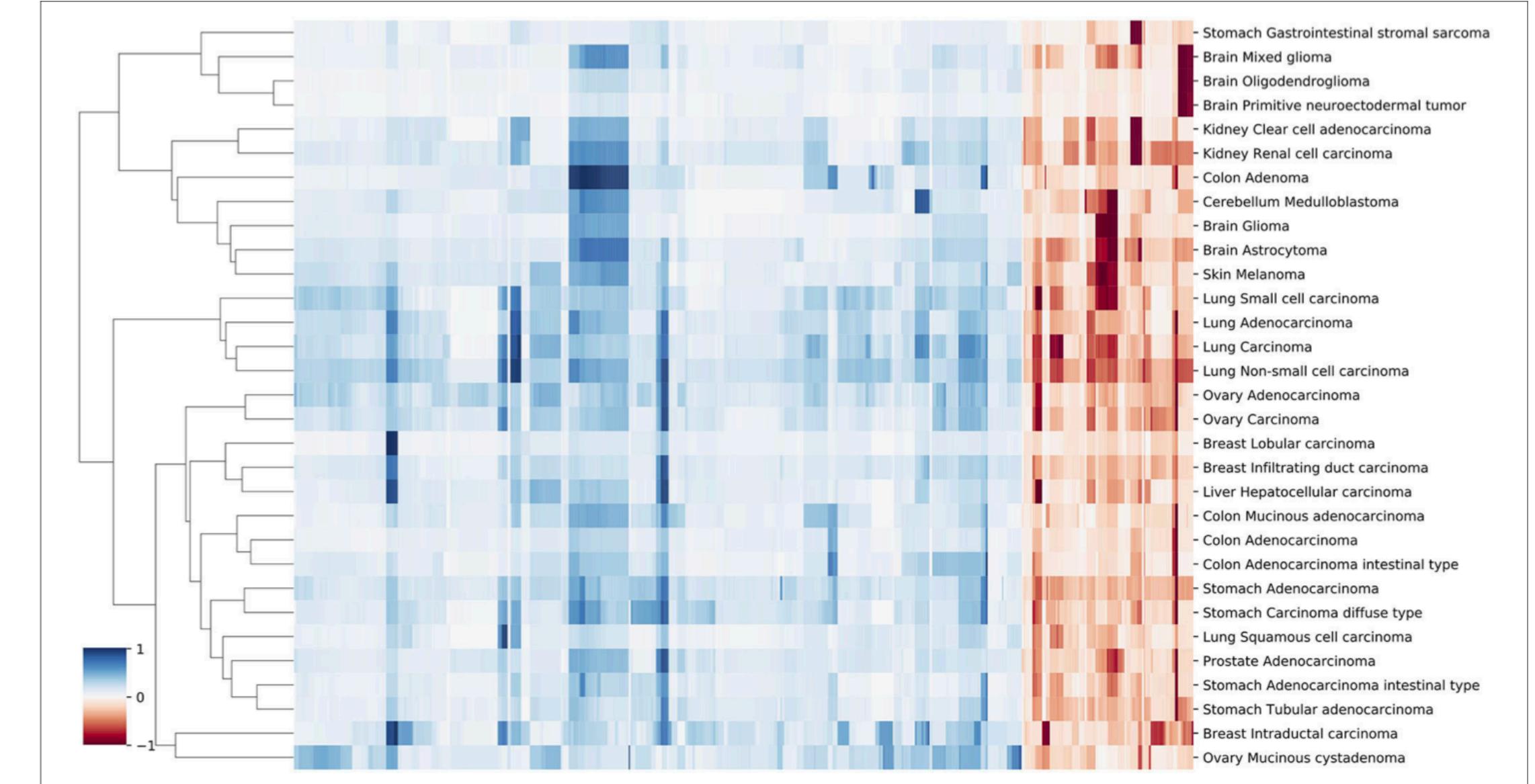


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

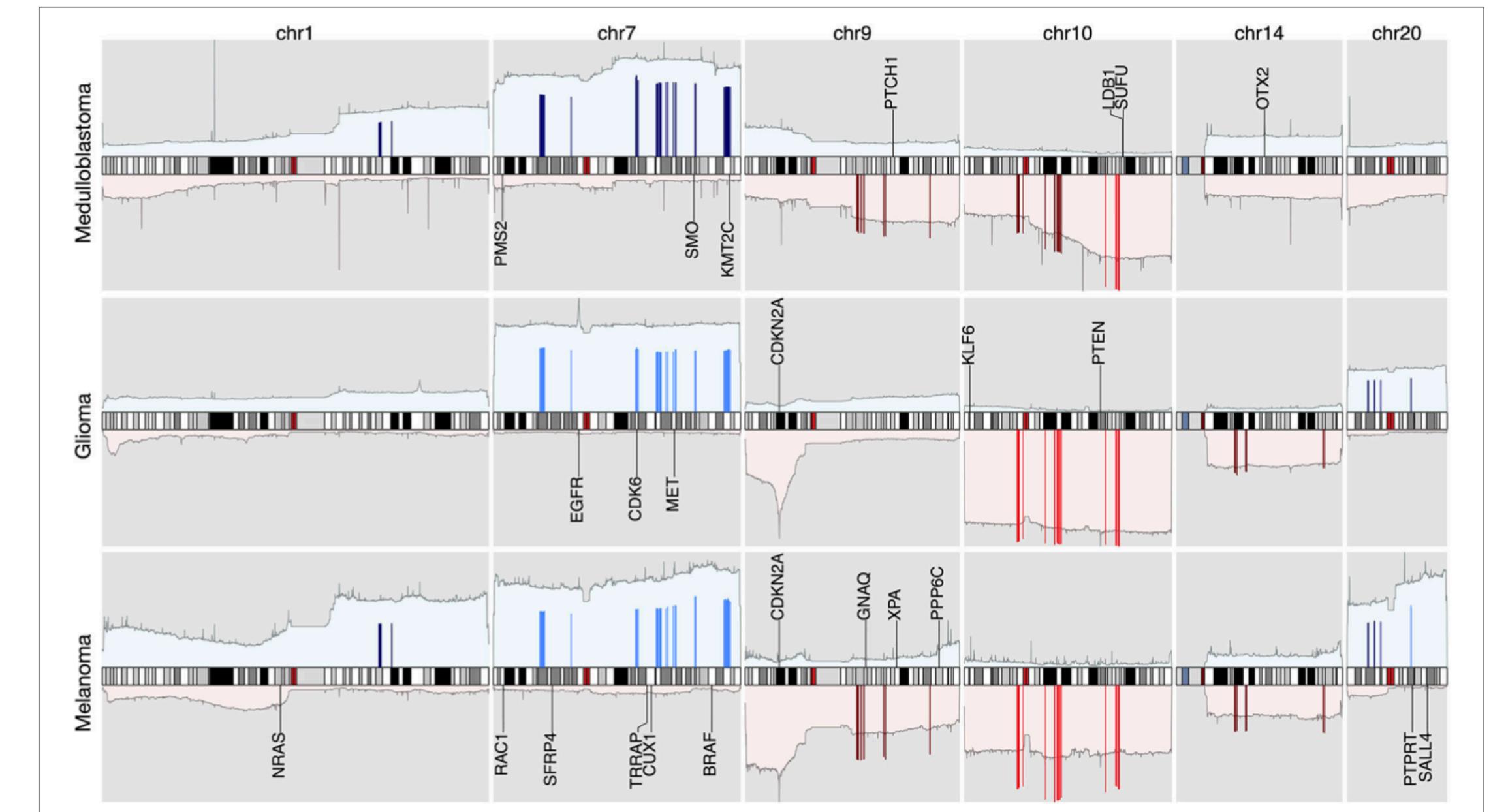
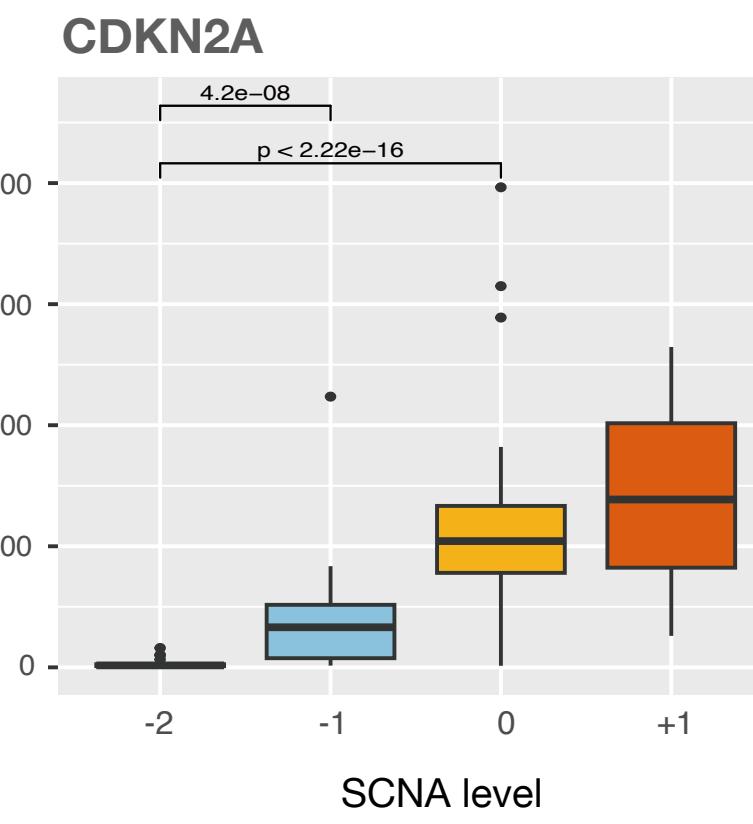
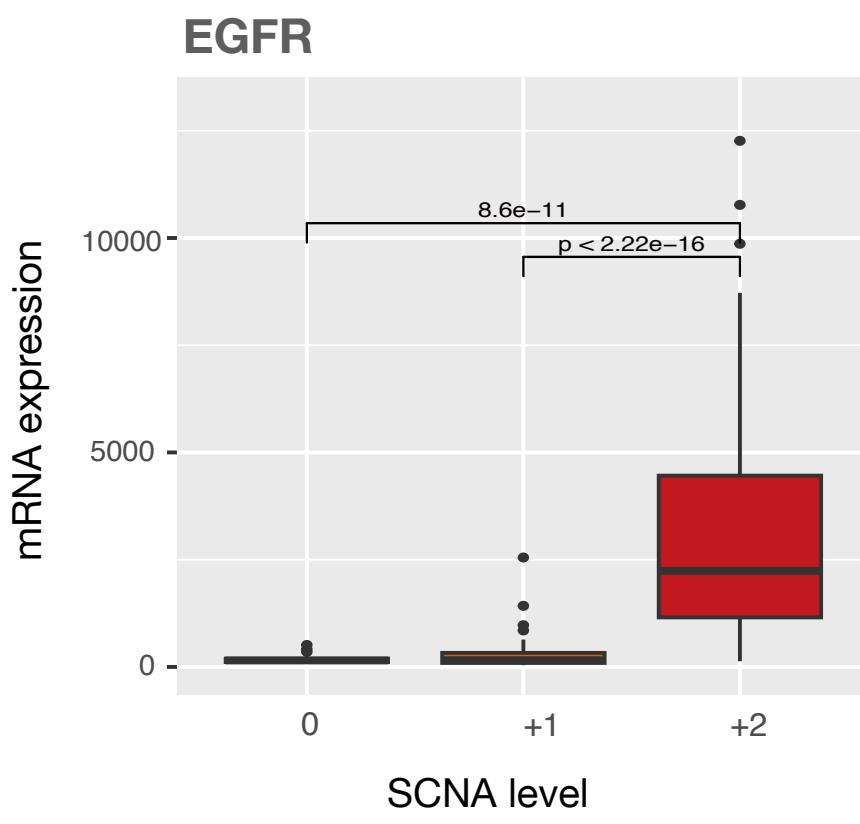


FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger signal intensity. The blue colors above the chromosome axis represent the average amplifications, and the red colors below the chromosome axis represent the average deletions. The amplitude of amplifications and deletions are normalized to [0,1] separately. The adjacent known driver genes are also included for each tumor type.

- Stomach Gastrointestinal stromal sarcoma
- Brain Mixed glioma
- Brain Oligodendrogloma
- Brain Primitive neuroectodermal tumor
- Kidney Clear cell adenocarcinoma
- Kidney Renal cell carcinoma
- Colon Adenoma
- Cerebellum Medulloblastoma
- Brain Gioma
- Brain Astrocytoma
- Skin Melanoma
- Lung Small cell carcinoma
- Lung Adenocarcinoma
- Lung Carcinoma
- Lung Non-small cell carcinoma
- Ovary Adenocarcinoma
- Ovary Carcinoma
- Breast Lobular carcinoma
- Breast Infiltrating duct carcinoma
- Liver Hepatocellular carcinoma
- Colon Mucinous adenocarcinoma
- Colon Adenocarcinoma
- Colon Adenocarcinoma intestinal type
- Stomach Adenocarcinoma
- Stomach Carcinoma diffuse type
- Lung Squamous cell carcinoma
- Prostate Adenocarcinoma
- Stomach Adenocarcinoma intestinal type
- Stomach Tubular adenocarcinoma
- Breast Intraductal carcinoma
- Ovary Mucinous cystadenoma

LabelSeg

A tool for fast and accurate profiling of CNA segment profiles in cancer



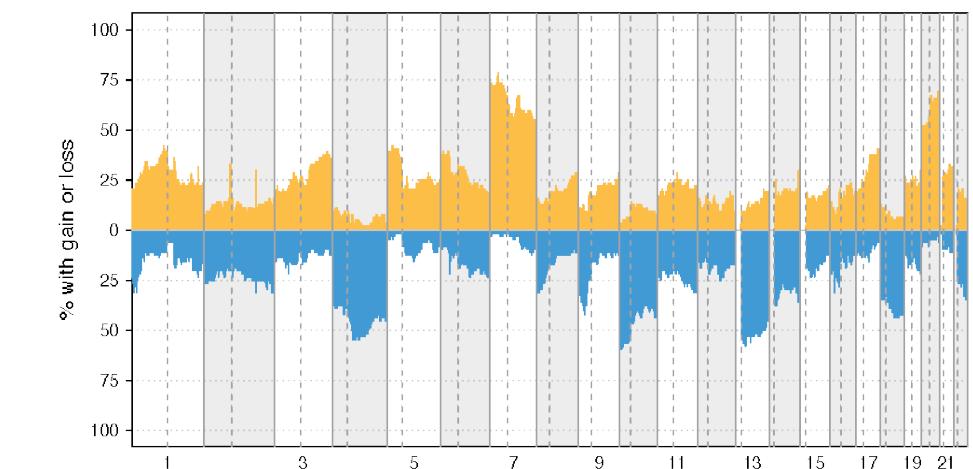
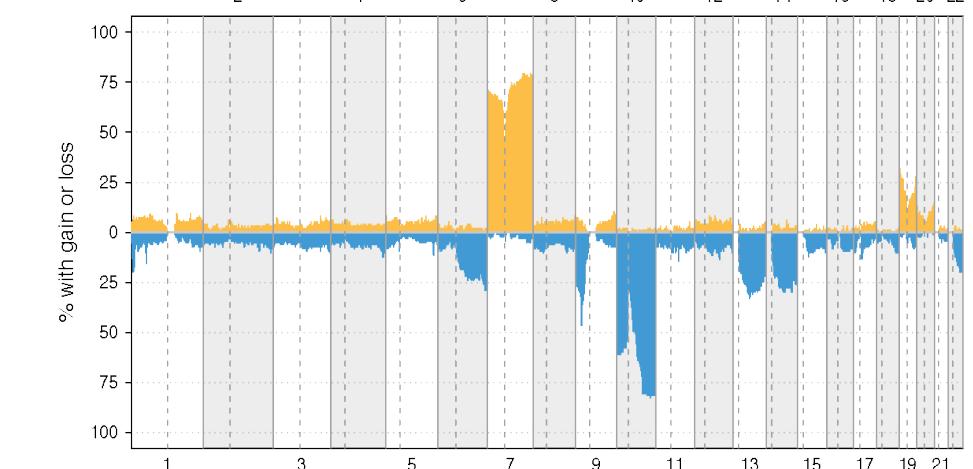
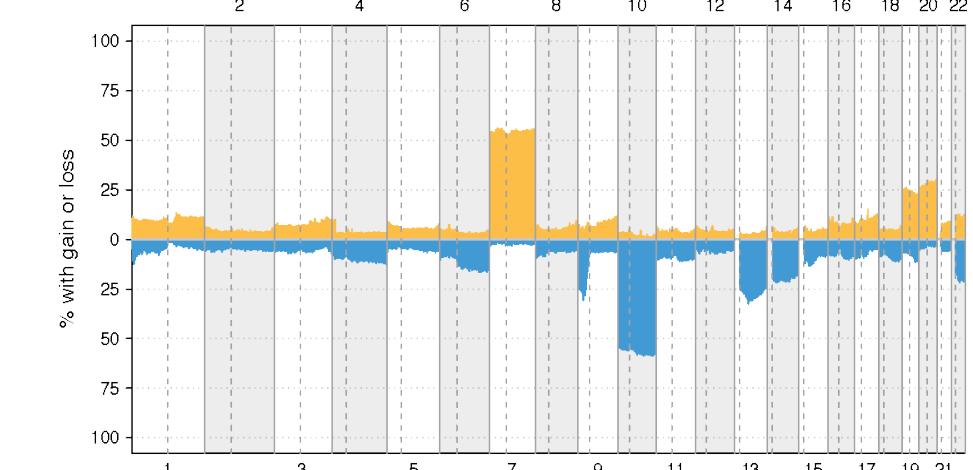
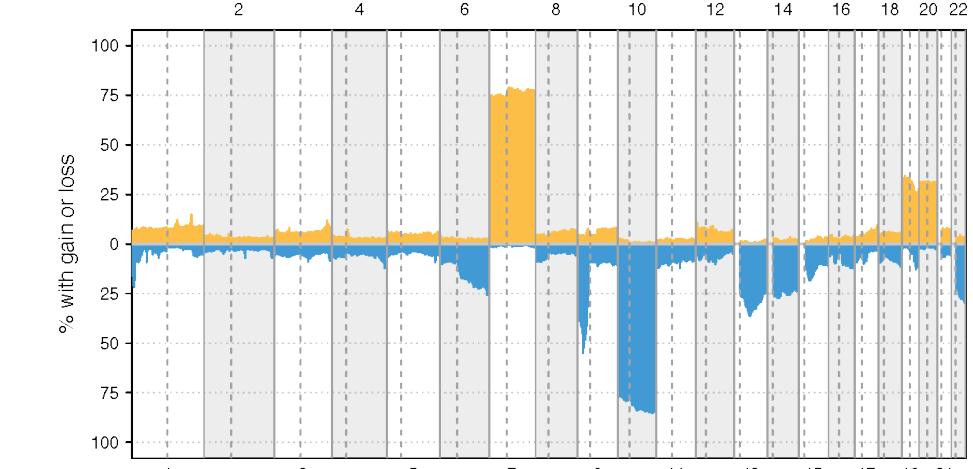
TCGA
566 samples

Progenetix
1390 samples

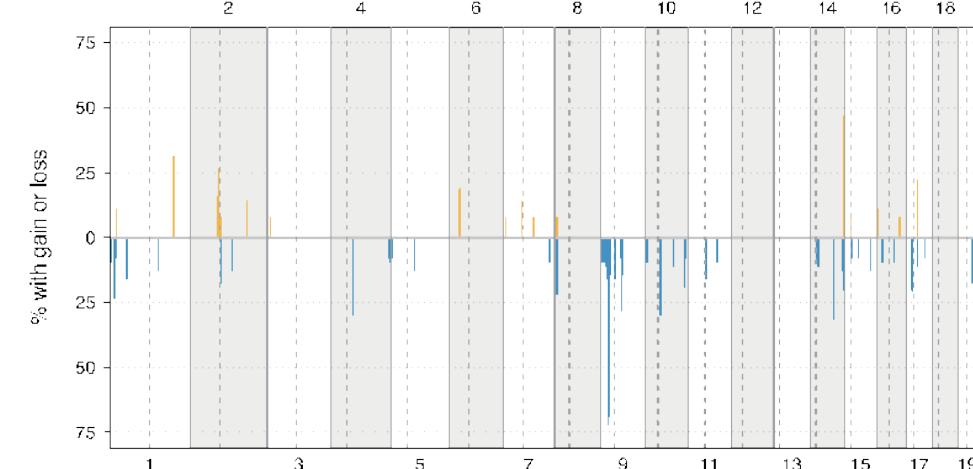
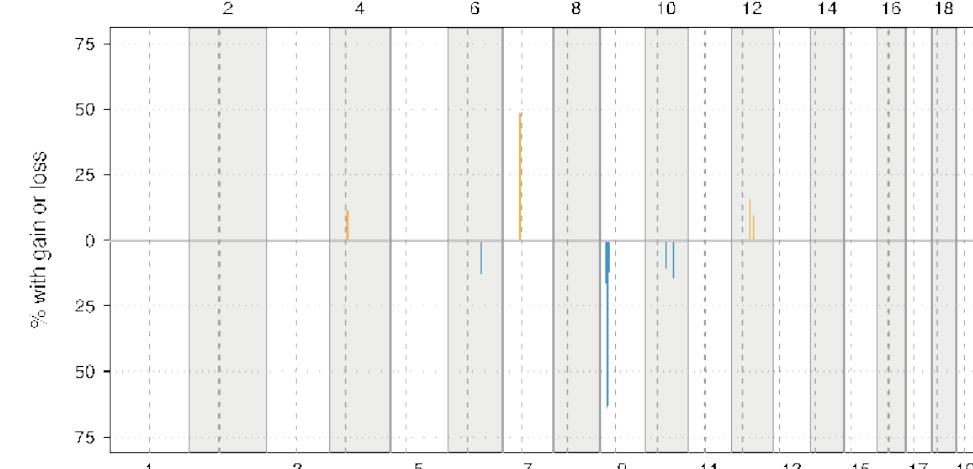
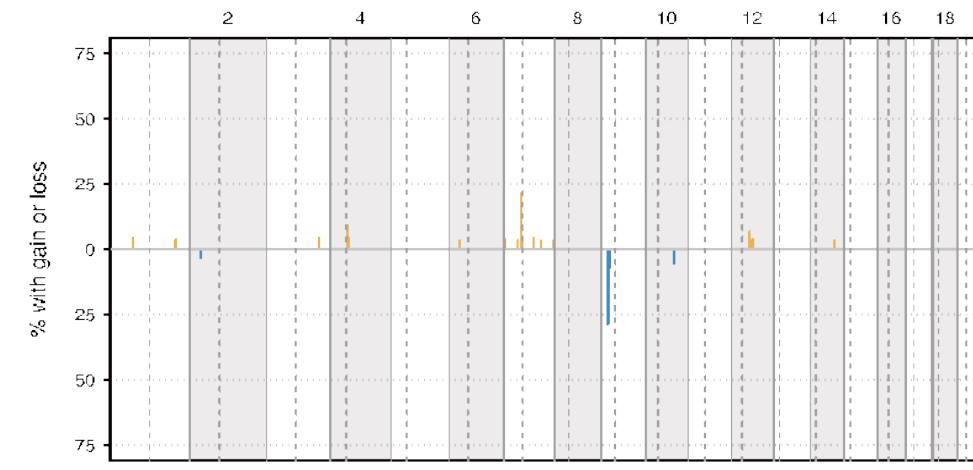
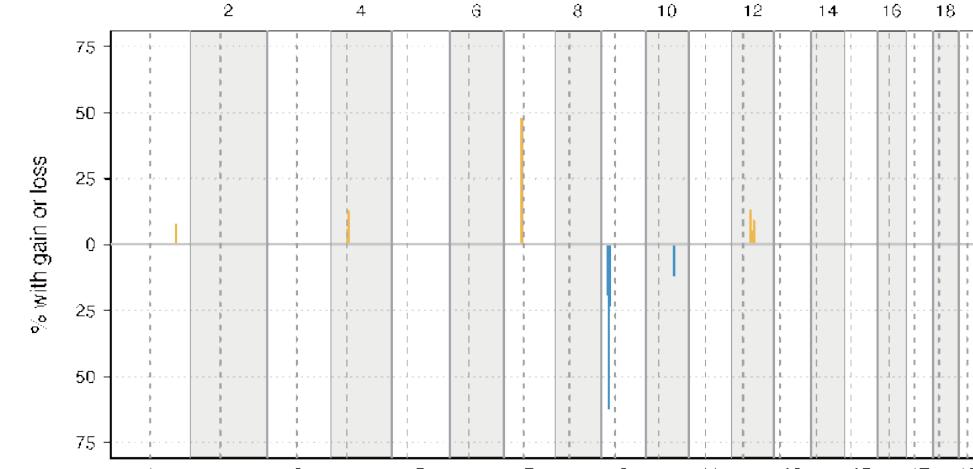
CPTAC
97 samples

CCLE
64 samples

Low-level SCNA



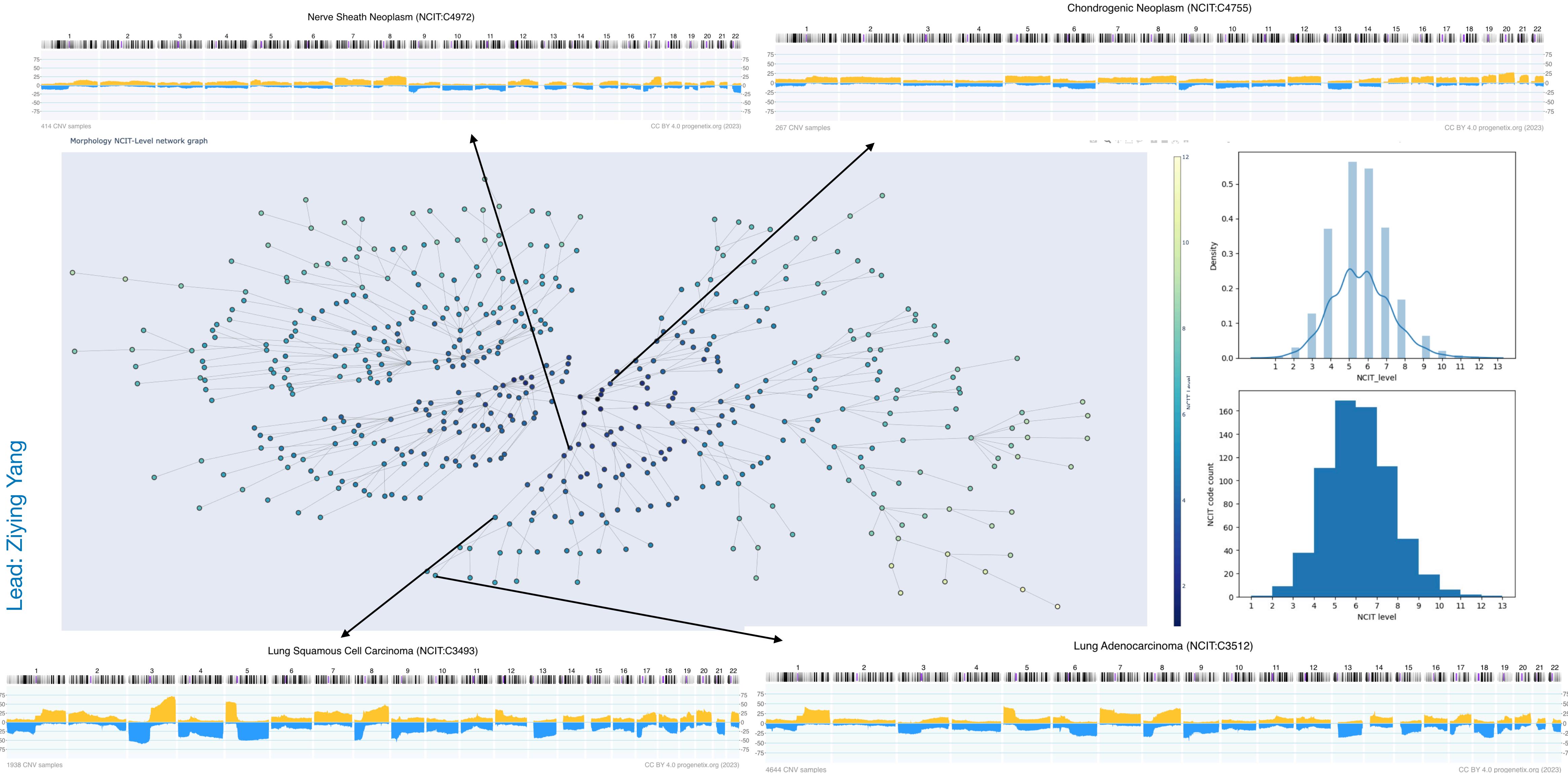
High-level SCNA



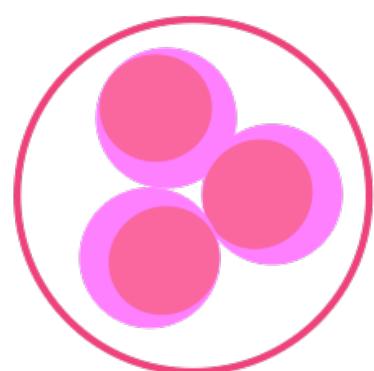
Runtime is just several seconds for a segment dataset of 1000 samples

CNV profiles heterogeneity vs cancer classification

Correspondance of genomic profiles to NCIT cancer hierarchy



Genomic Profiles as Fingerprints of Cancer Types



cancercelllines

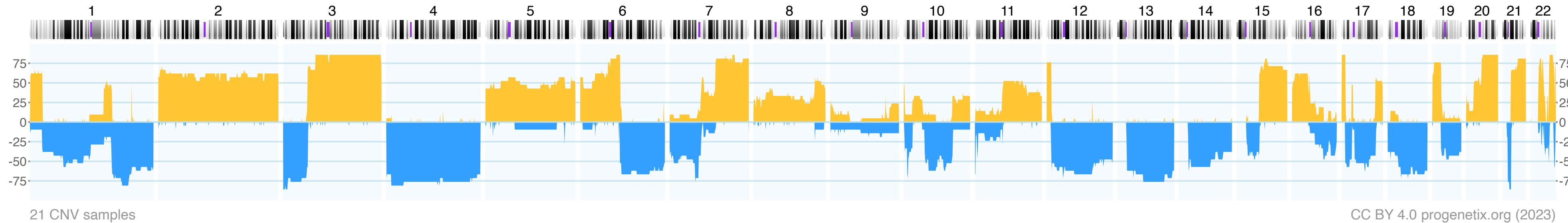
MDA-MB435: A widely used Breast cancer cell line ...

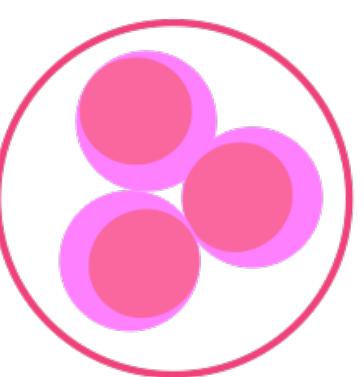
Cell line name	MDA-MB-435
Synonyms	MDA-MB435; MDAMB435; MDA.MB.435; MDA-435; MDA 435; MDA435; MD Anderson-Metastatic Breast-435
Accession	CVCL_0417

Part of: GrayJW breast cancer cell line panel.
Part of: NCI-60 cancer cell line panel.

Mutation; HGNC; [1097](#); BRAF; Simple; p.Val600Glu (c.1799T>A); ClinVar=[VCV000013961](#); Zygosity=Heterozygous (PubMed=[17088437](#); PubMed=[28889351](#)).
Mutation; HGNC; [1787](#); CDKN2A; Simple; c.150+2T>C (IVS1+2T>C); ClinVar=[VCV000406712](#); Zygosity=Heterozygous; Note=Splice donor mutation (PubMed=[17088437](#)).
Mutation; HGNC; [1787](#); CDKN2A; Simple; c.455insCdel26; Zygosity=Heterozygous (PubMed=[17088437](#)).
Mutation; HGNC; [11998](#); TP53; Simple; p.Gly266Glu (c.797G>A); ClinVar=[VCV000161516](#); Zygosity=Heterozygous (PubMed=[17088437](#); PubMed=[28889351](#)).

MDA-MB-435 (cellosaurus:CVCL_0417)



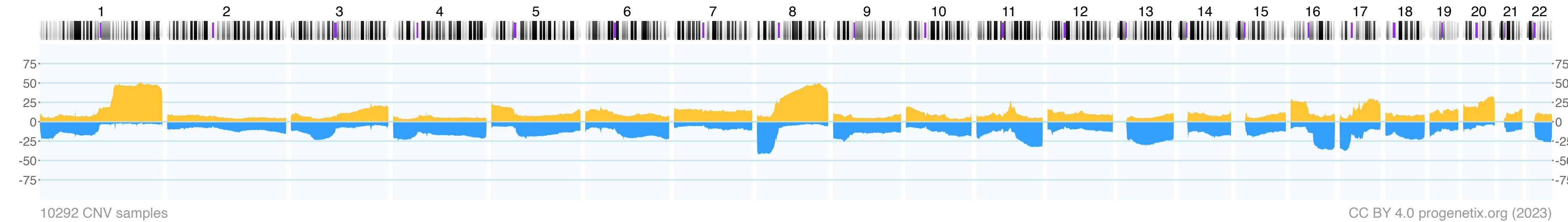


cancer cell lines

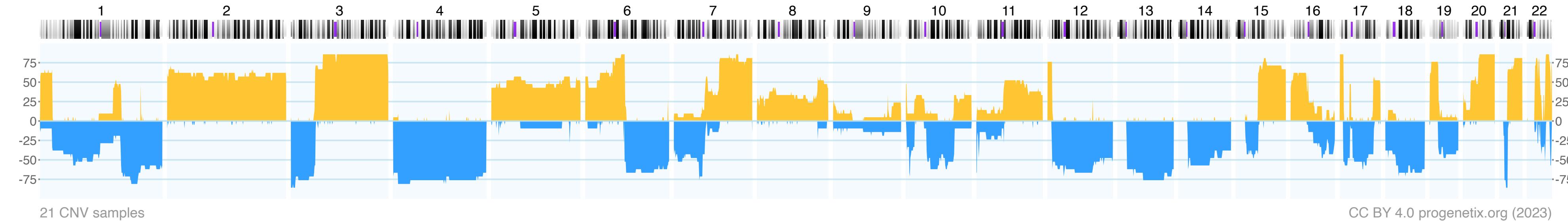
Genomic Profiles as Fingerprints of Cancer Types

An unpleasant surprise...

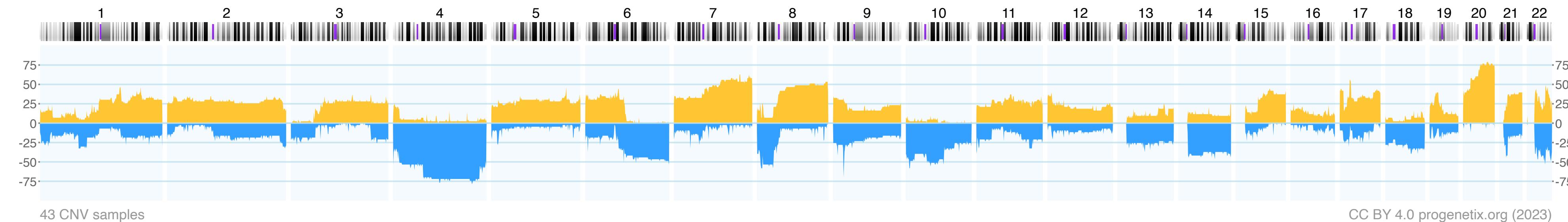
Ductal Breast Carcinoma (NCIT:C4017)



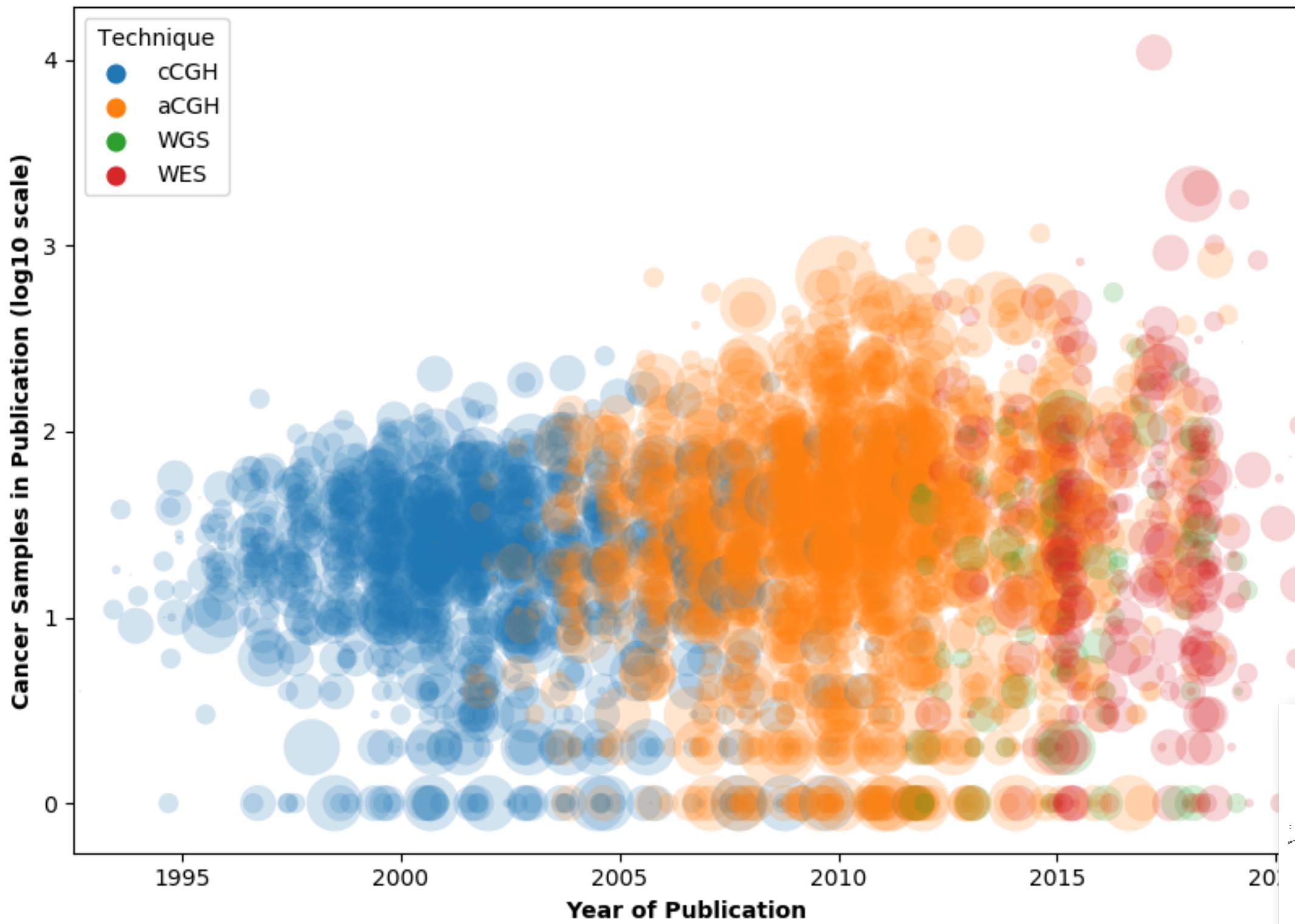
MDA-MB-435 (cellosaurus:CVCL_0417)



Amelanotic Melanoma (NCIT:C3802)



Number of tumor samples for each publication across the years



Database, 2020, 1–9
doi: 10.1093/database/baa009
Articles



Articles

Geographic assessment of cancer genome profiling studies

Paula Carrio-Cordo^{1,2}, Elise Acheson³, Qingyao Huang^{1,2} and Michael Baudis^{1,*}

¹Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland ²Swiss Institute of Bioinformatics, Zurich, Switzerland ³Department of Geography, University of Zurich, Zurich, Switzerland



Cancer CNV Profiles

Search Samples

Studies & Cohorts

- arrayMap
- TCGA Samples
- DIPG Samples
- Gao & Baudis, 2021
- Cancer Cell Lines
- Publication DB
- Services
- NCIt Mappings

Progenetix Publication Collection

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

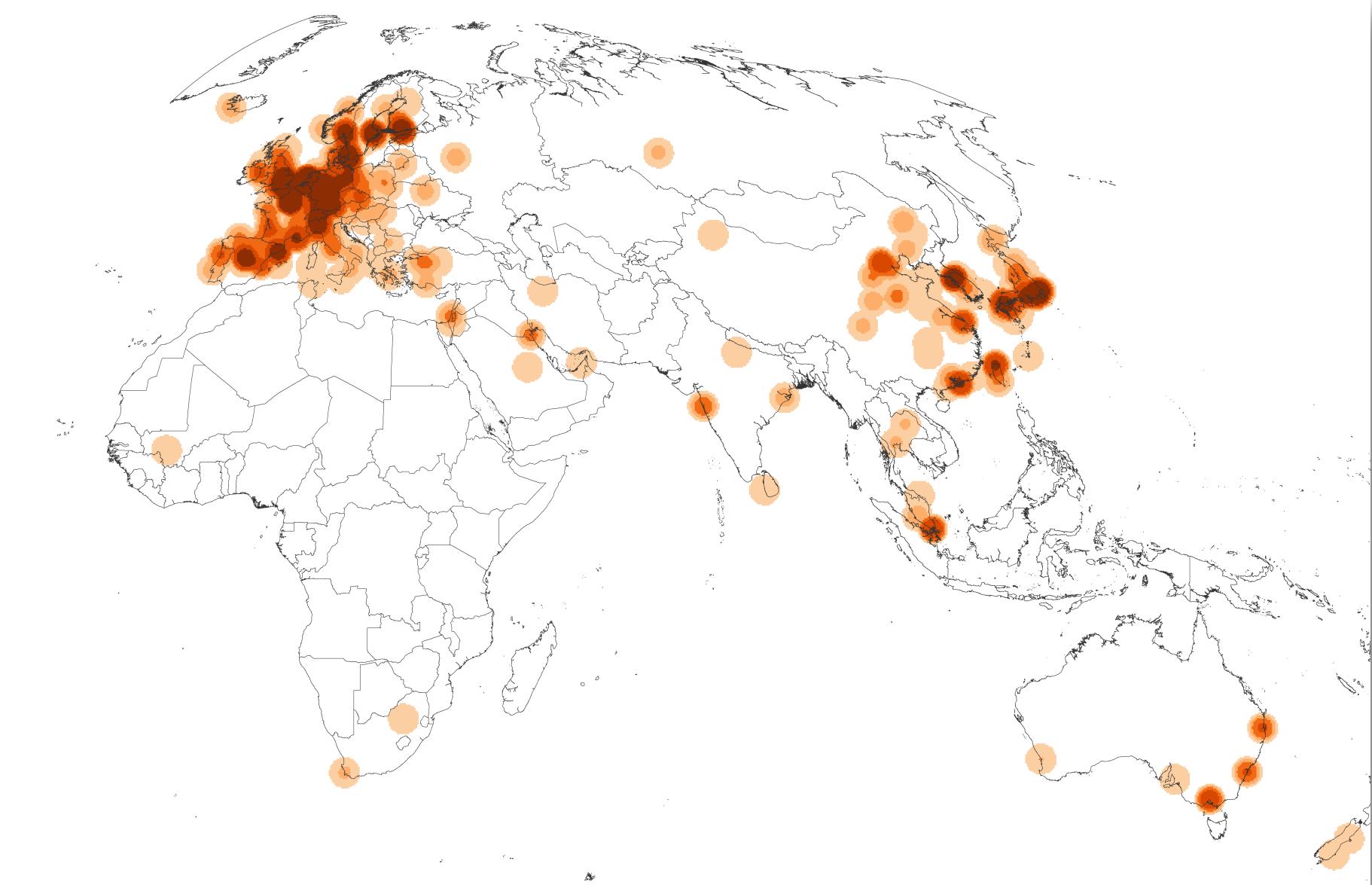
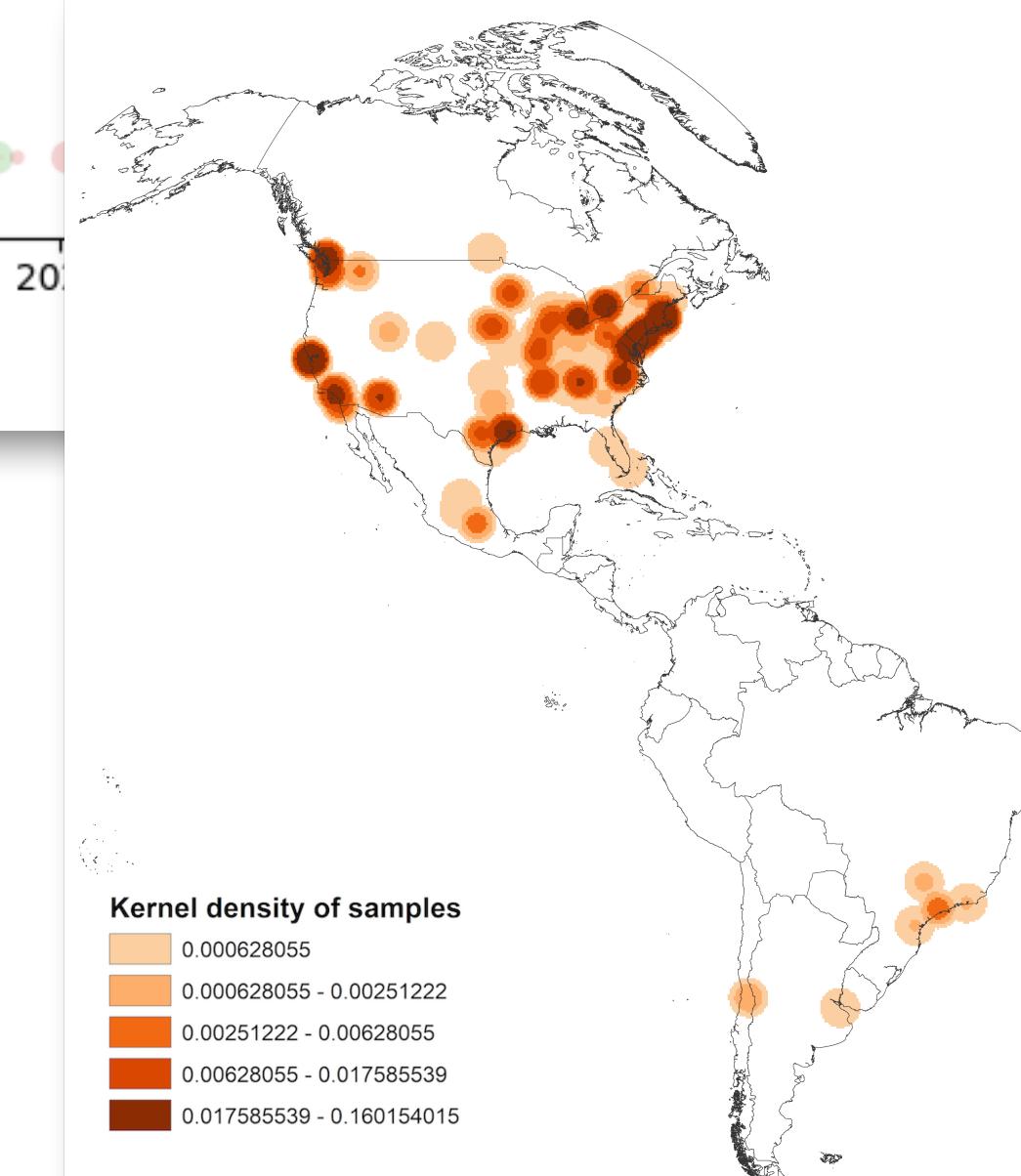
Please [contact us](#) to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation [🔗](#).

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

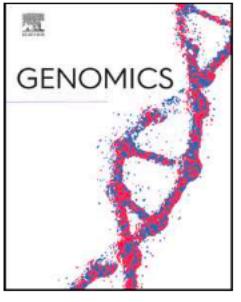
 Type to search... [▼](#)

Publications (3324)

id i ▾	Publication	Samples
		cCGH aCGH WES WGS pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations	0 79 0 0 0



Map of the geographic distribution (by 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 in Progenetix.



Recent Publications

CNV Data Analysis & Methods

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

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

ORIGINAL PAPER

CNARA: reliability assessment for genomic copy number profiles

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

F1000Resea Minimum error calibration and normalization for genomic copy number analysis

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

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

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

Genomic Instability of Osteosarcoma Cell Lines in Culture: Impact on the Prediction of Metastasis Relevant Genes

Roman Muff¹, Prisni Rath², Ram Mohan Ram Kumar¹, Knut Husmann¹, Walter Born¹, Michael Baudis², Bruno Fuchs^{1*}

ORIGINAL RESEARCH article

Front. Genet., 16 January 2023
Sec. Cancer Genetics and Oncogenomics
Volume 13 - 2022 | <https://doi.org/10.3389/fgene.2022.1017657>

This article is part of the Research Topic
Mutational Signatures and Immune Response in Cancer
[View all 8 Articles >](#)

Candidate targets of copy number deletion events across 17 cancer types

Qingyao Huang^{1,2} and Michael Baudis^{1,2*}

SCIENTIFIC REPORTS
nature research

Enabling population assignment from cancer genomes with SNP2pop

Qingyao Huang^{ID 1,2} & Michael Baudis^{ID 1,2*}

JEADV

JOURNAL OF
THE EUROPEAN
ACADEMY OF
DERMATOLOGY &
VENEREOLOGY

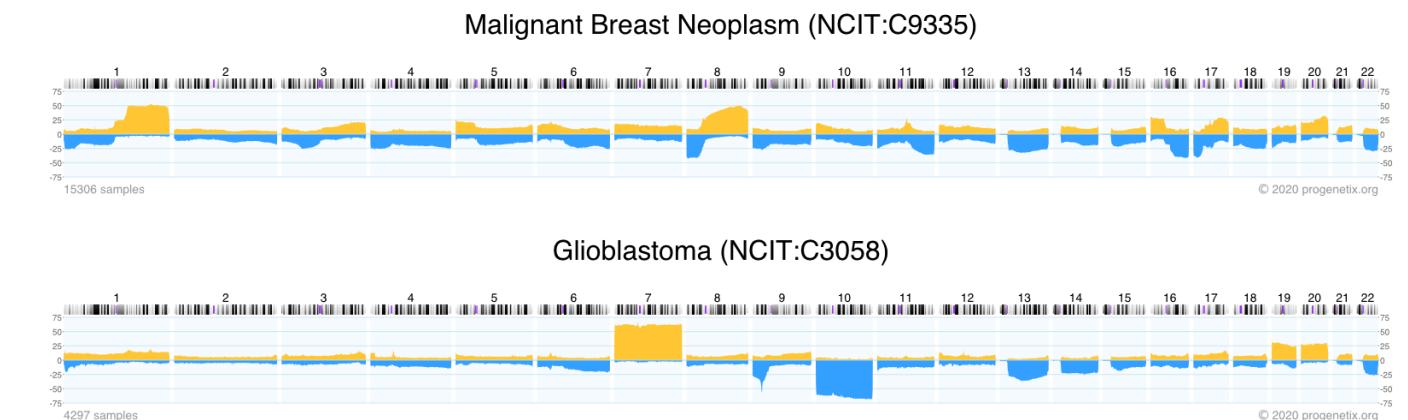
Copy number imbalances in primary cutaneous lymphomas

G. Gug¹, Q. Huang², E. Chiticariu³, C. Solovan⁴, M. Baudis⁵

Theoretical Cytogenetics and Oncogenomics

... but what does this entail @baudisgroup?

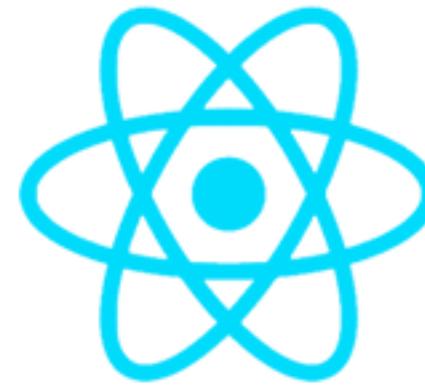
- patterns & markers in cancer genomics, especially somatic structural genome variants
- bioinformatics support in collaborative studies
- reference resources for curated cancer genome variations
- bioinformatics tools & methods
- standards and reference implementations for data sharing in genomics and personalized health
- open research data "ambassadoring"



(Bio)informatics Skill Set

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

text mining



React



regular expressions
s/knowledge/mastery/

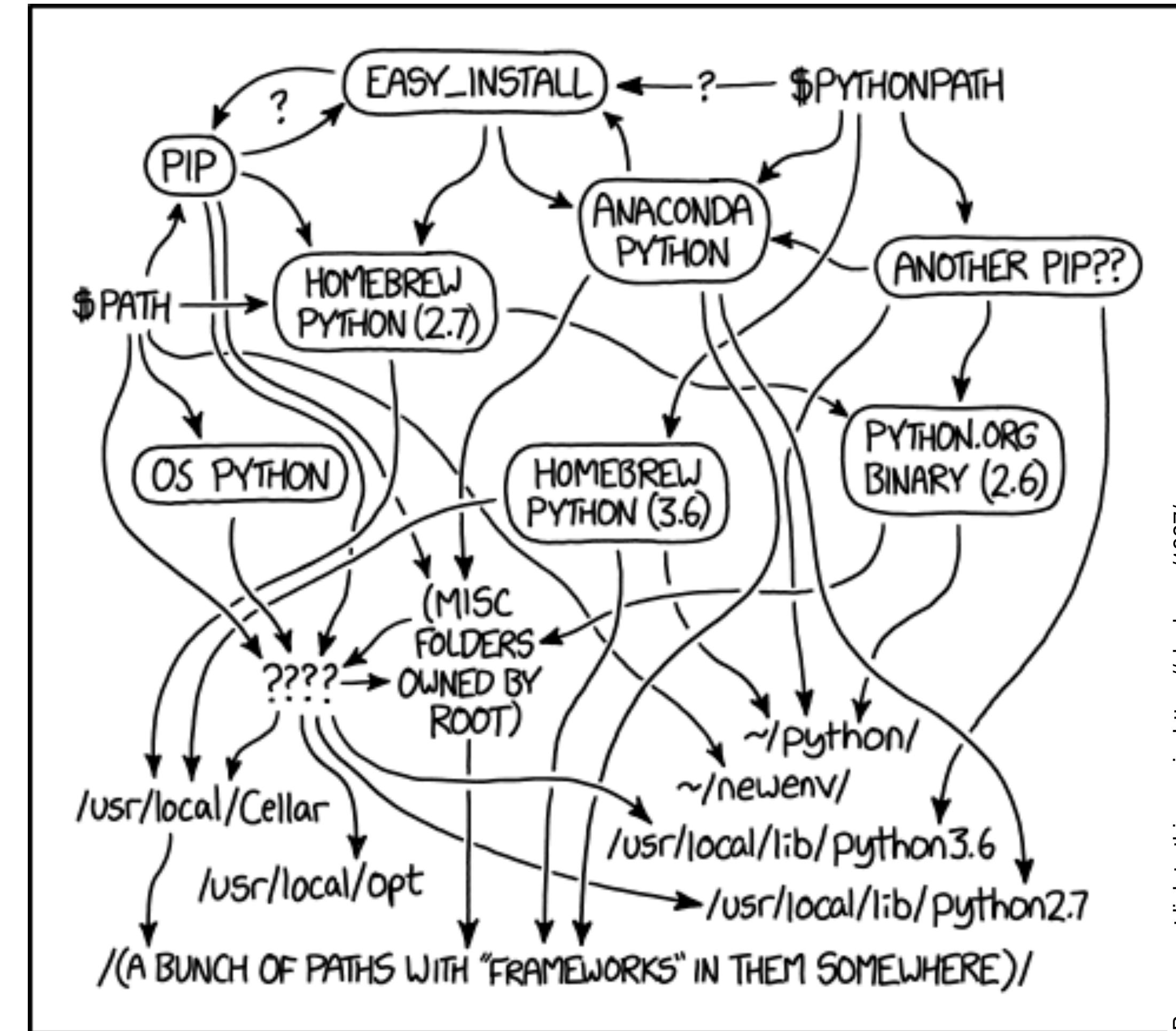


array pipelines

statistics



But in reality that is what bioinformaticians do...



Permanent link to this comic: <https://xkcd.com/1987/>



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.



info.baudisgroup.org