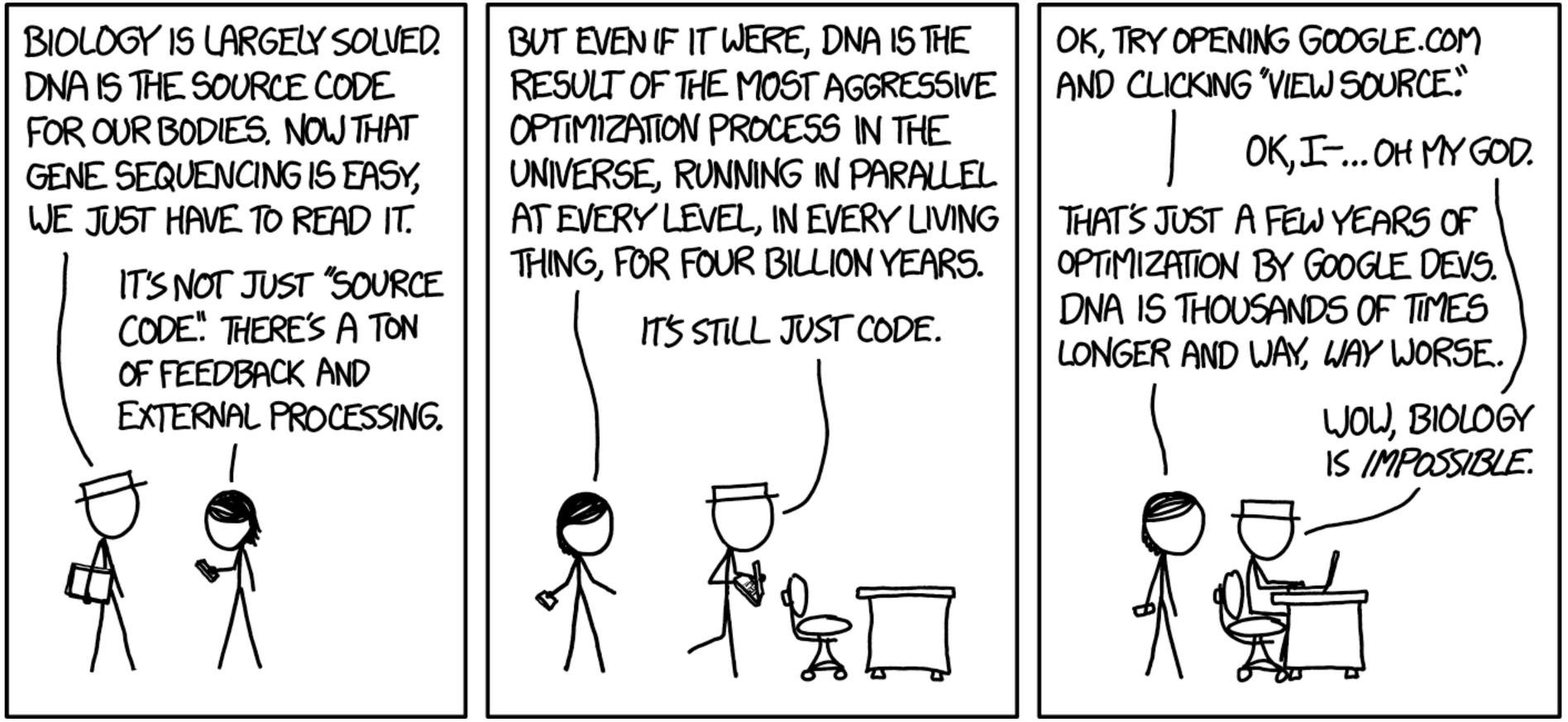
Theoretical Cytogenetics and Oncogenomics

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

Michael Baudis | 2023-03-24

Biology is has complex source code - Bioinformatics might help

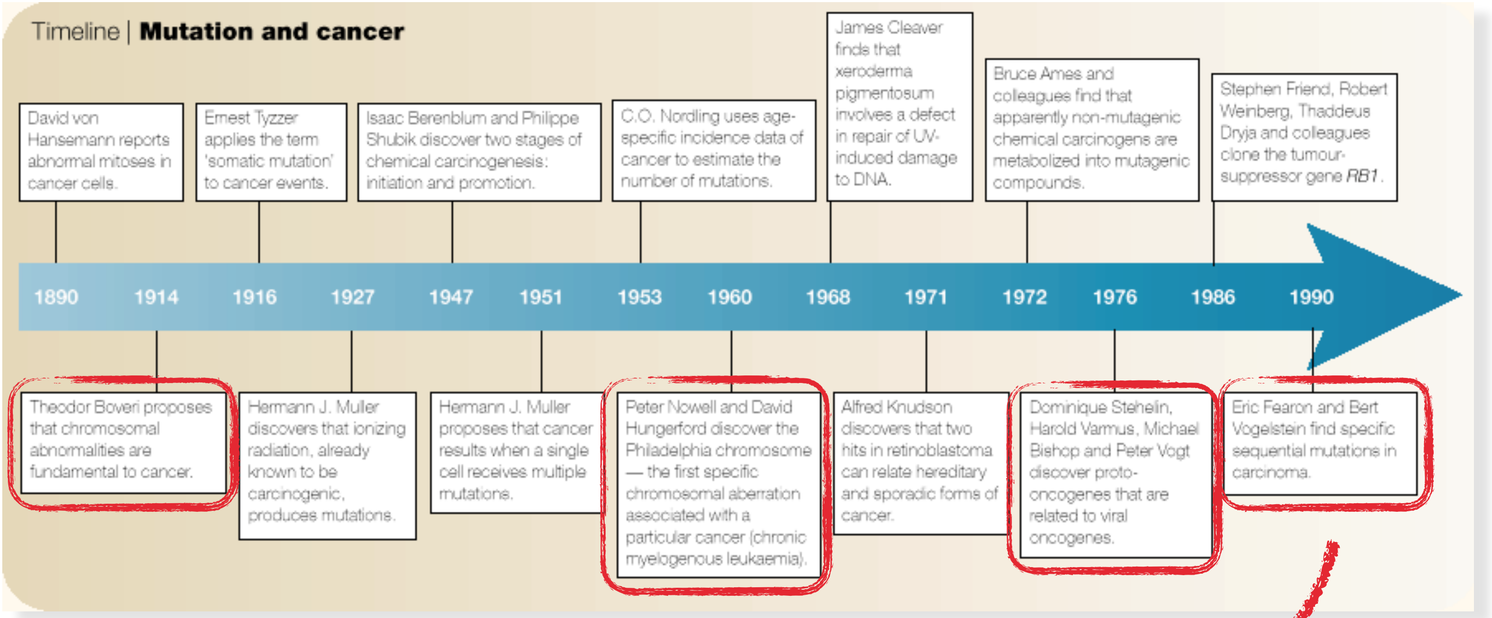


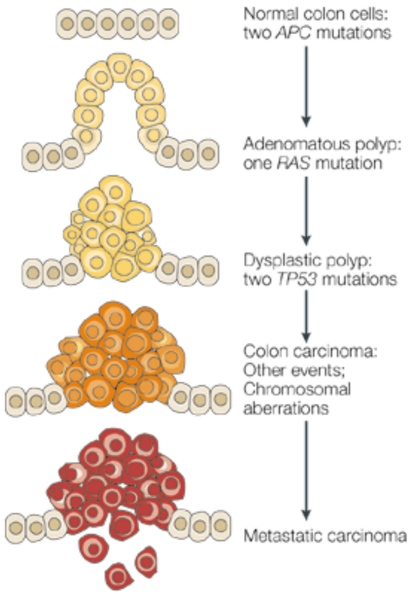
Randall Munroe: https://xkcd.com/1605/

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"





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.*)

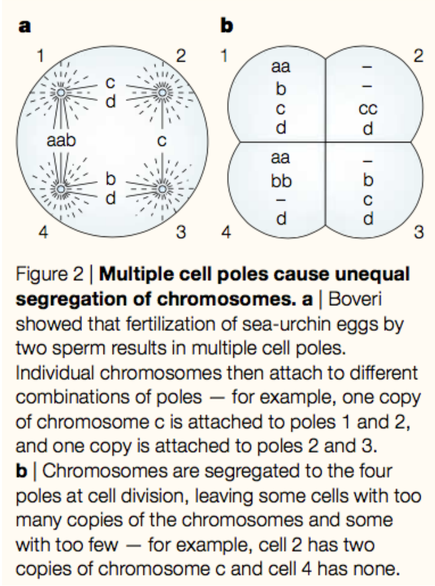


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes. a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached to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

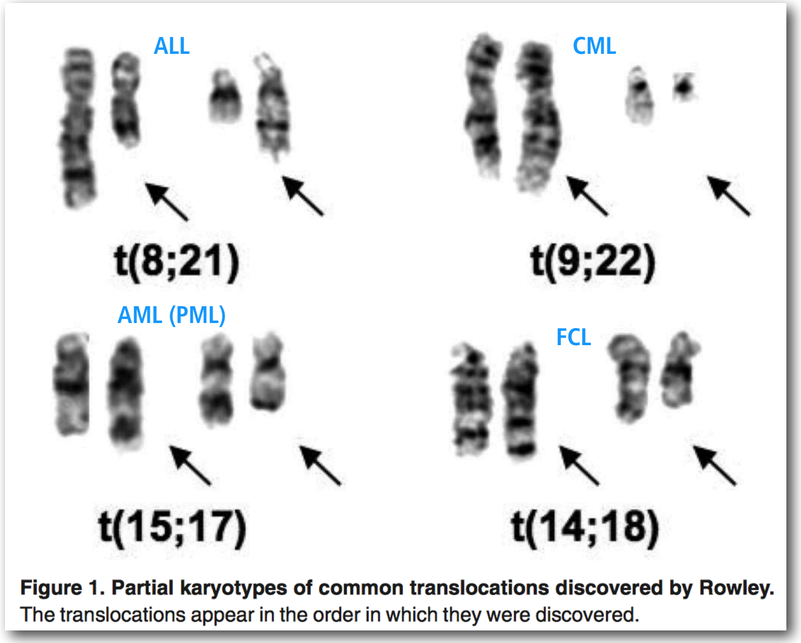
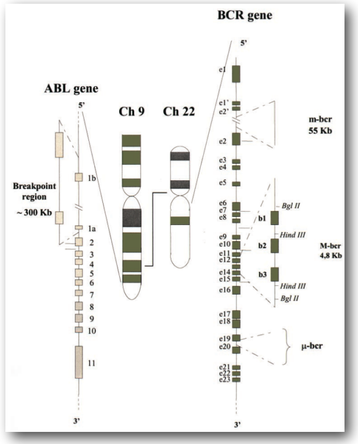
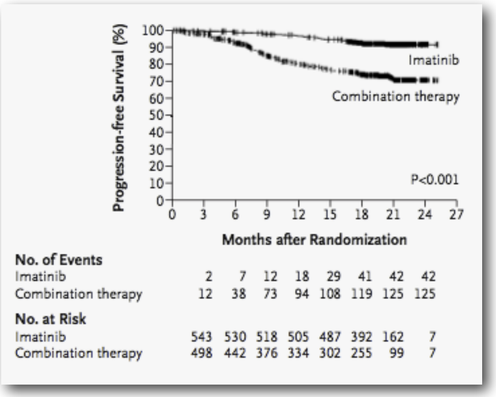


Figure 1. Partial karyotypes of common translocations discovered by Rowley. The translocations appear in the order in which they were discovered.

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



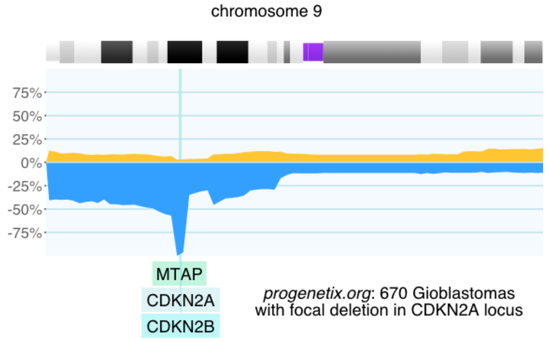
Pane et al. BCR/ABL genes …. Oncogene (2002), 21 (56)



Event free Survival in first large Imatinib Trials

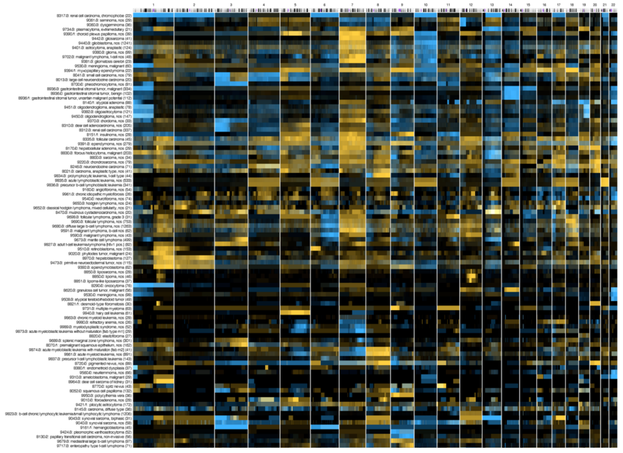
O'Brien et al. Imatinib compared with interferon and low-dose cytarabine… NEJM (2003) vol. 348 (11)

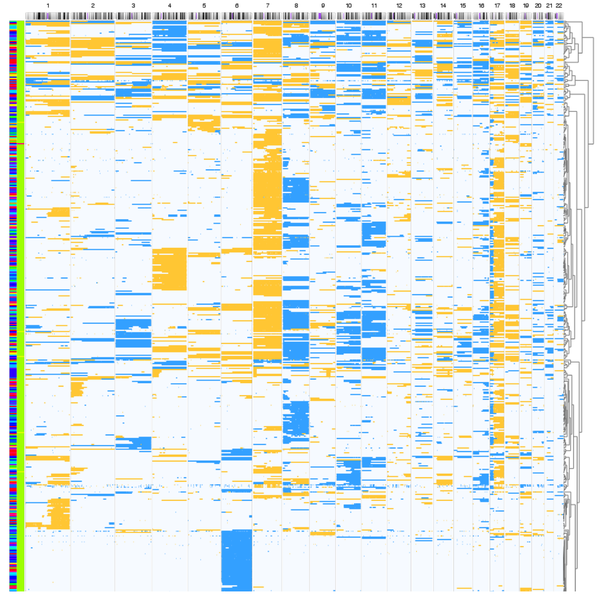
Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

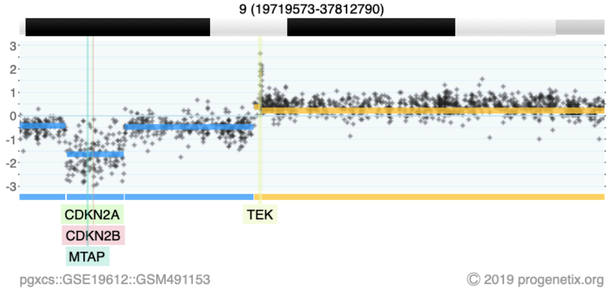


Genomic Imbalances in Cancer - Copy Number Variations (CNV)

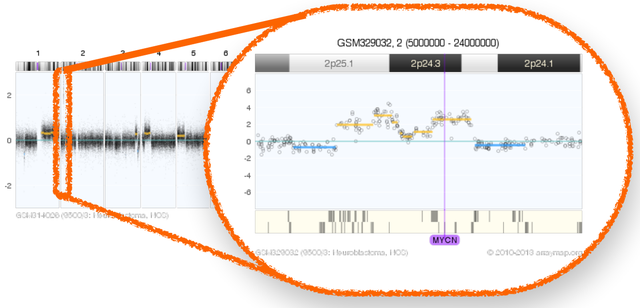
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)







2-event, homozygous deletion in a Glioblastoma



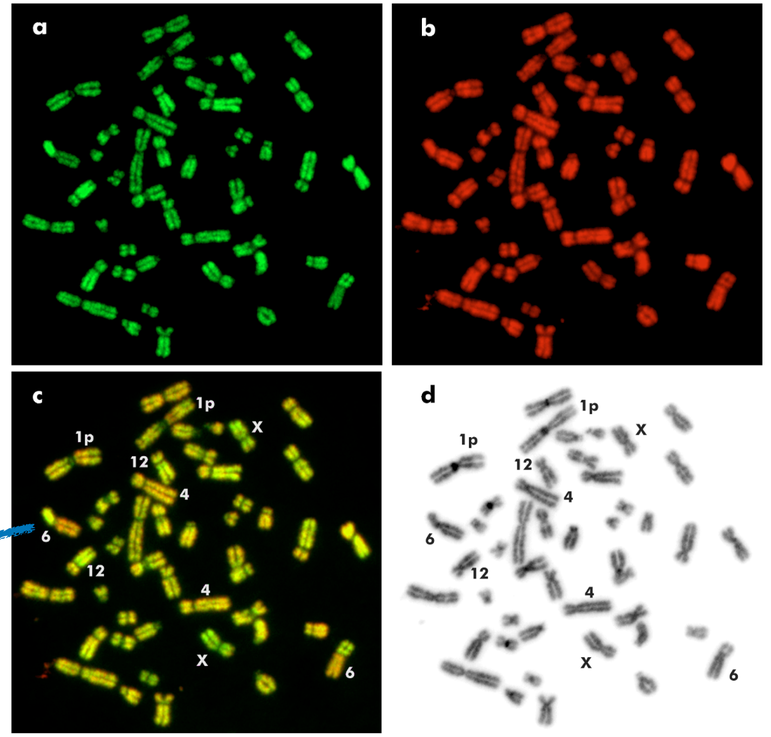
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

Comparative Genomic Hybridization

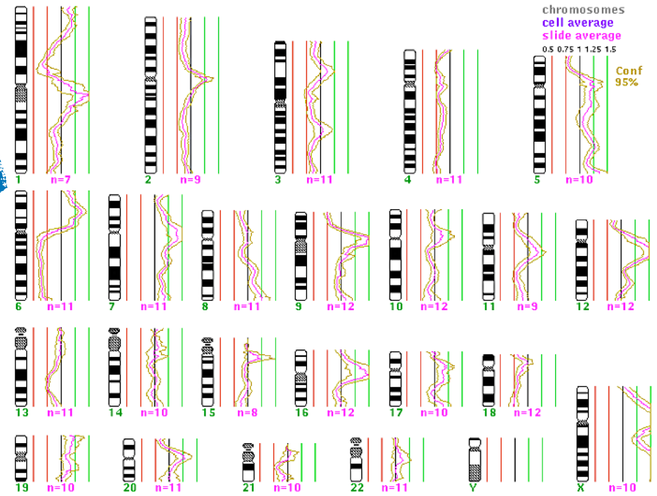
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- 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)

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

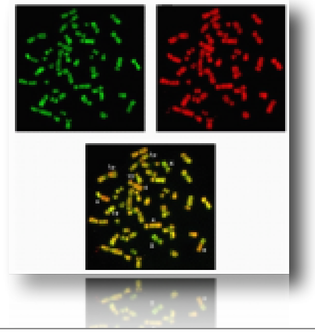


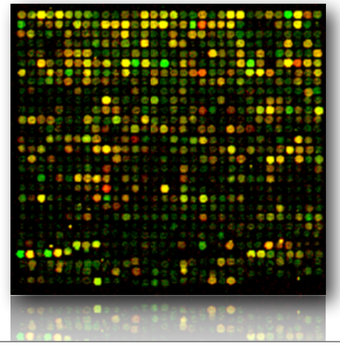
CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen

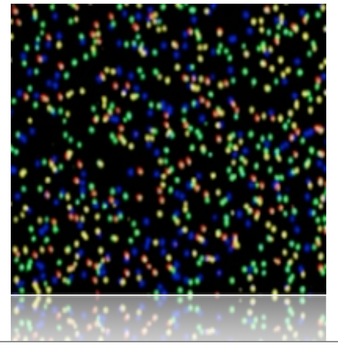


Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

WHOLE GENOME SCREENING IN CANCER

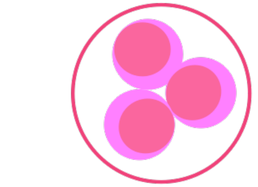






	chromosomal CGH	genomic arrays	"NGS" genome sequencing (WES, WGS)
1st application report	**1992**	**1997**	**2010**
source	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)
main source problems	mixed/degraded source tissue	mixed/degraded source 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 rax 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

arrayMap

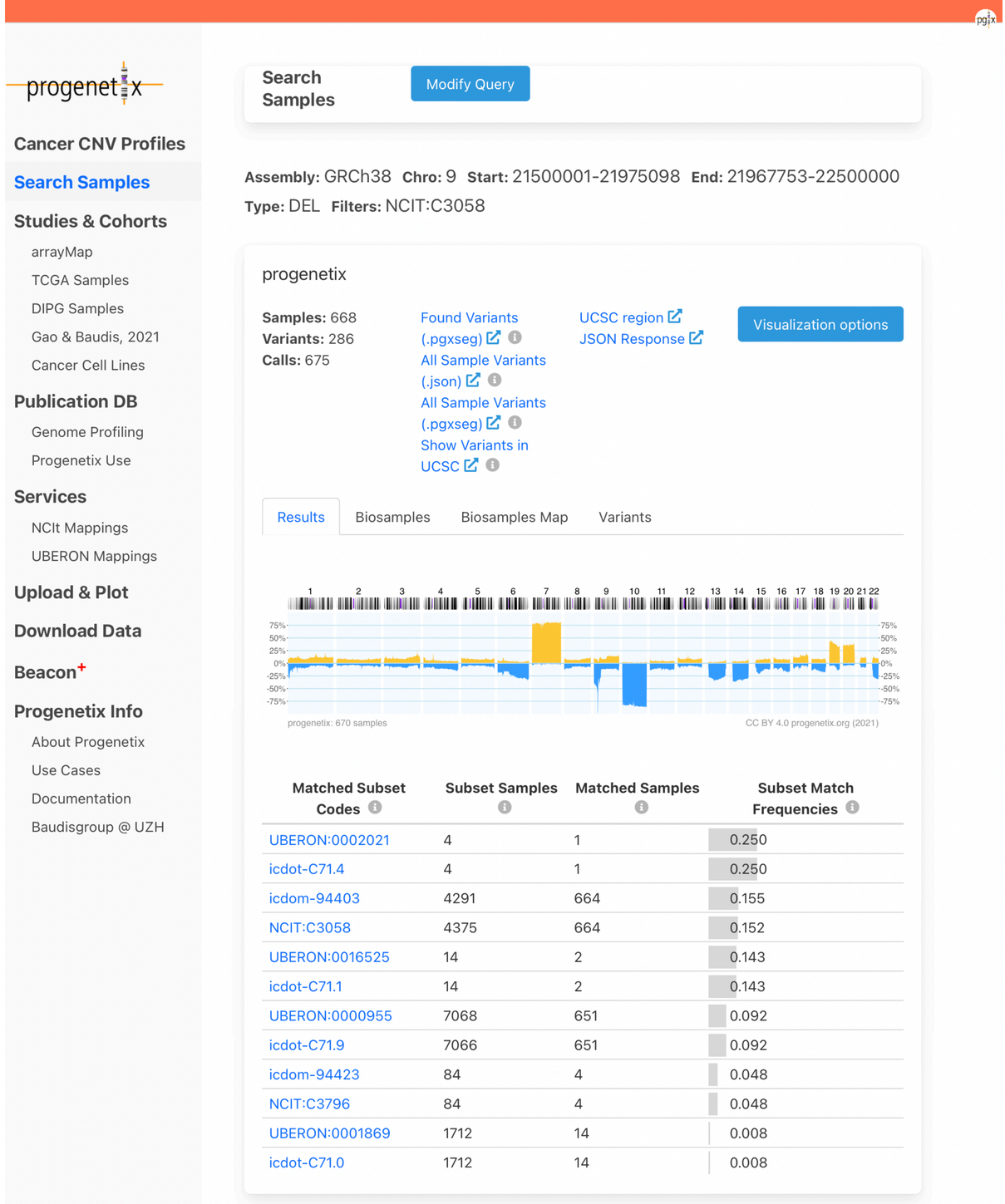
progenetix

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 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

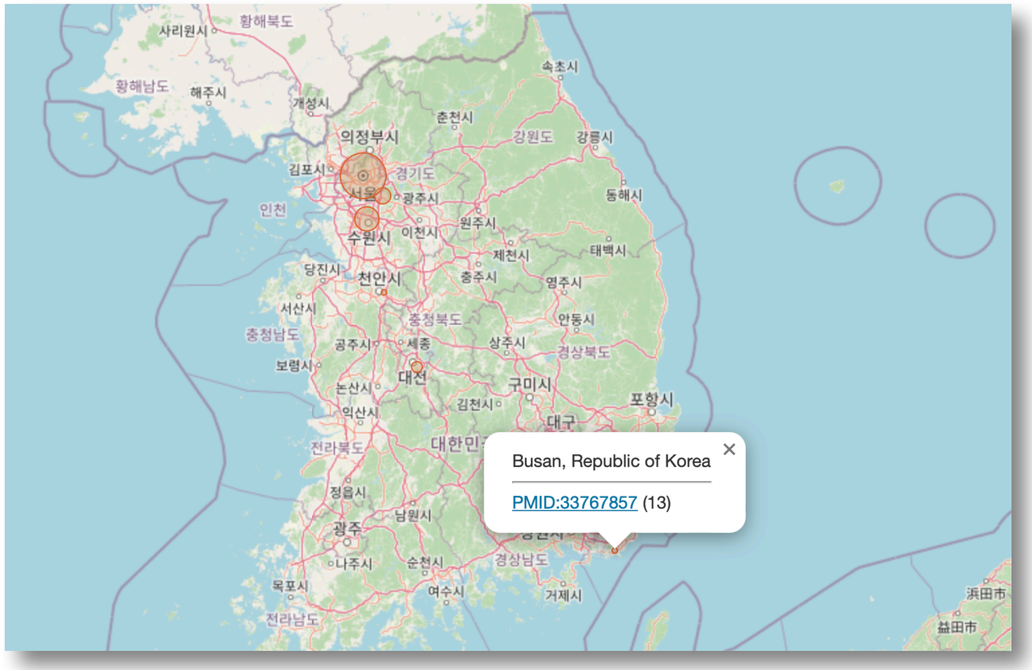
progenetix



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

http://progenetix.org/services/publications/?filters=genomes:>0&ISO3166alpha2=kr&output=map

Database, 2021, **2021(0)**, 1–9
DOI: https://doi.org/10.1093/database/baab043
Database update

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,*]

[1]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland
[2]Swiss 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

[a]set of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Thesaurus.

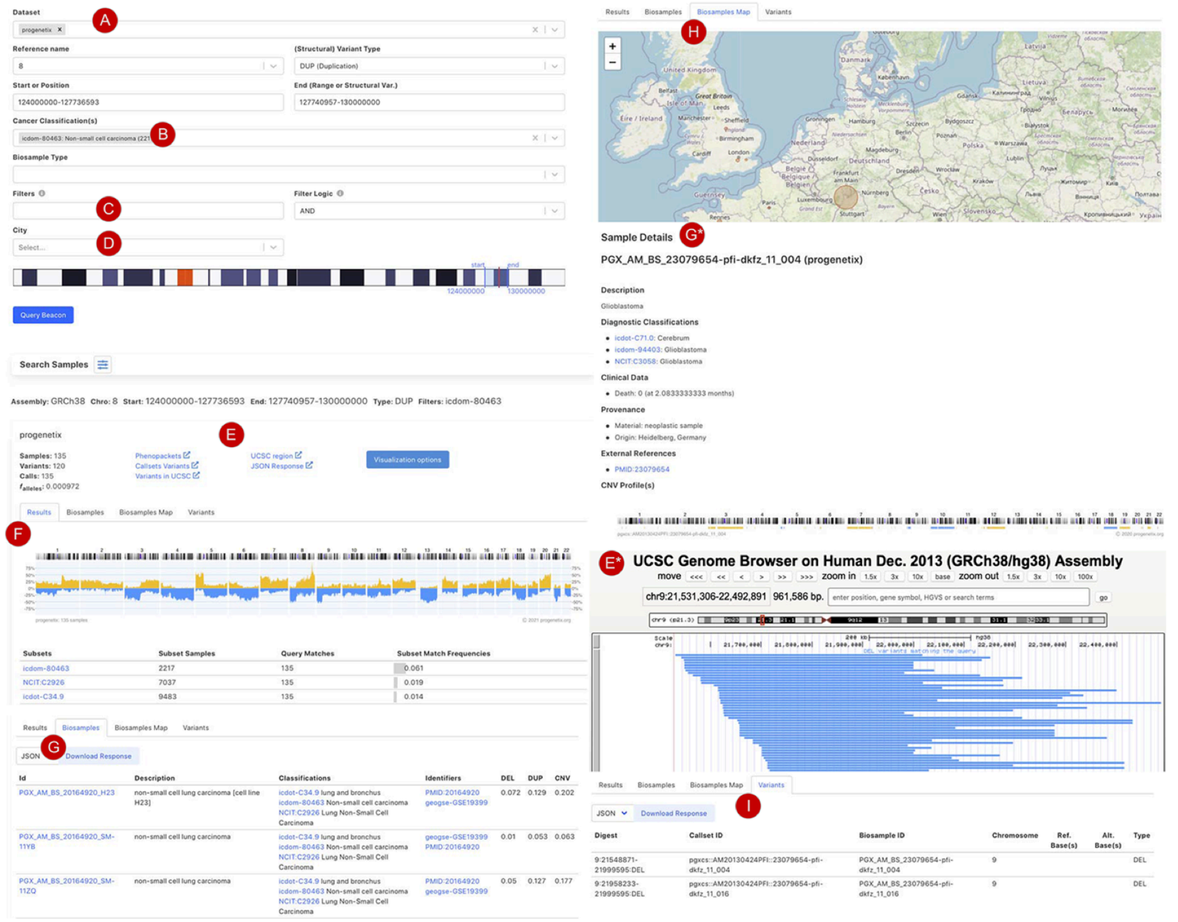


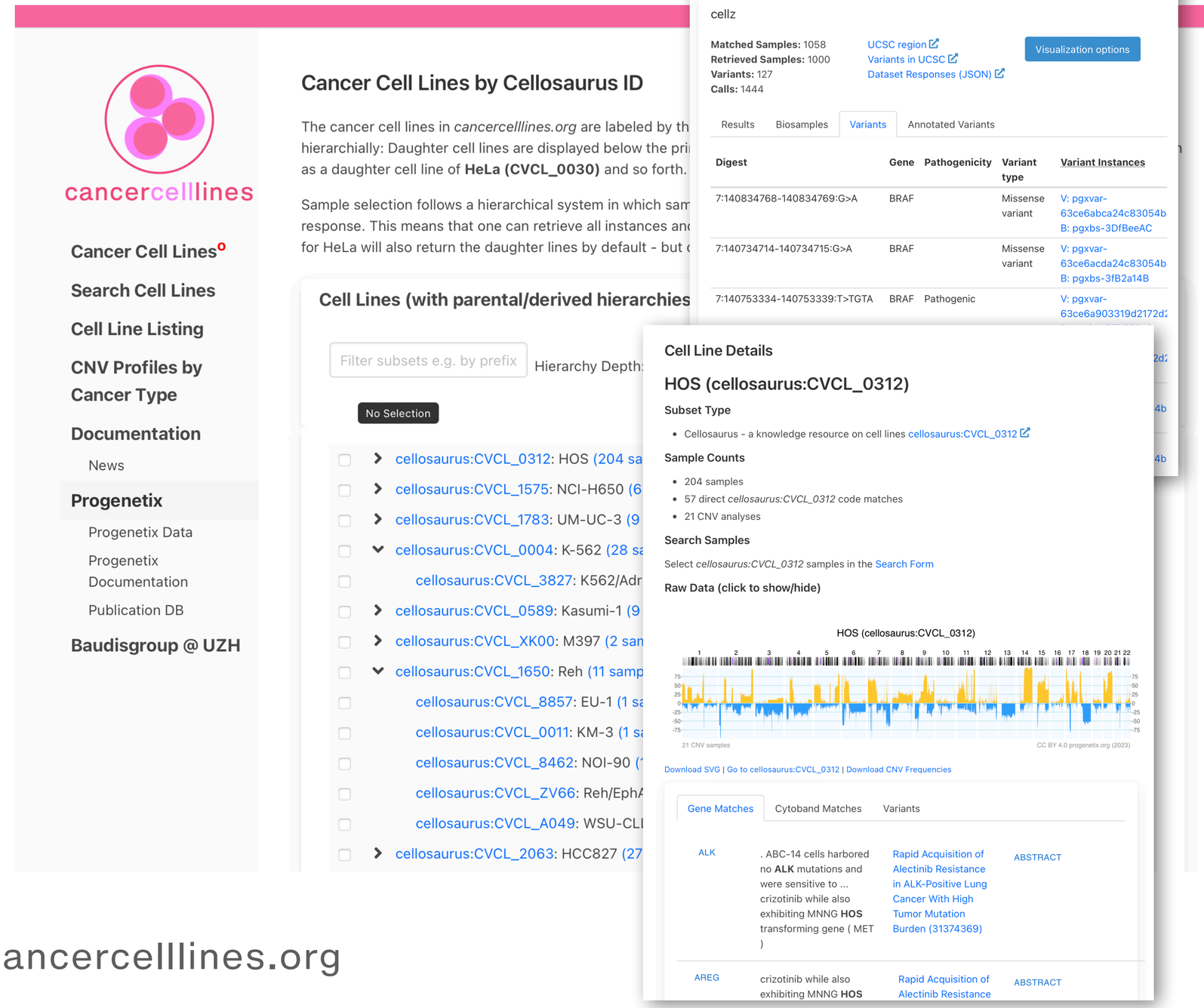
Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and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



cancercelllines.org

Lead: Rahel Paloots

Resource Infrastructure - *pgx* Stack

progenetix

- 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

React

APACHE HTTP SERVER PROJECT

python™

mongoDB

- *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"),
```

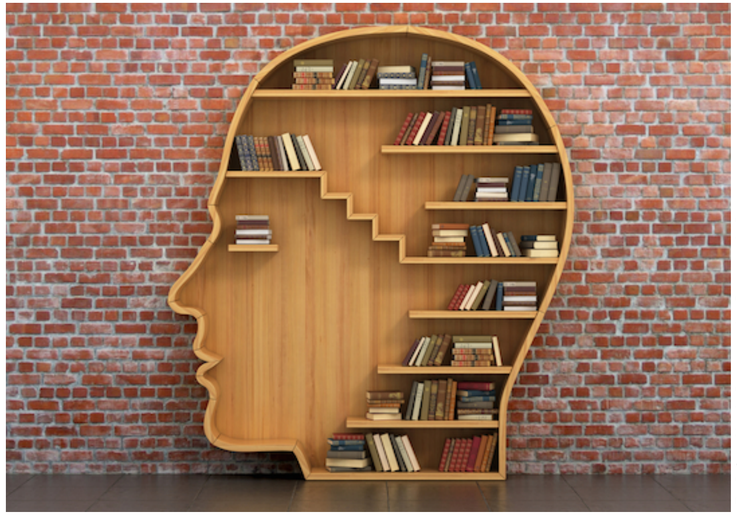
Entity collections

Utility collections

Data Parasites

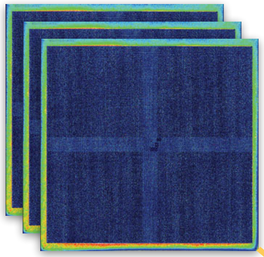
Data Parasites

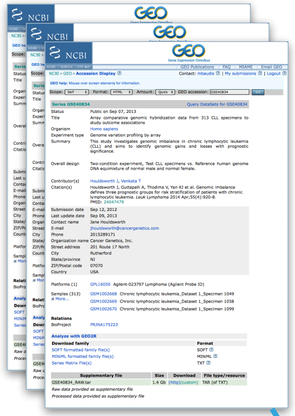
... or let's call us CURATORS

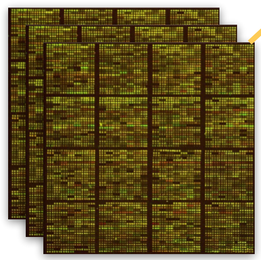


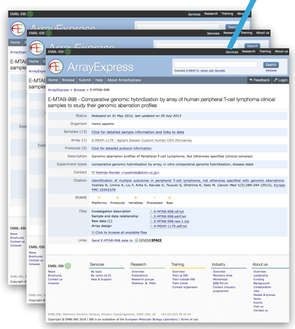
Belén Nadales - belennadales.com

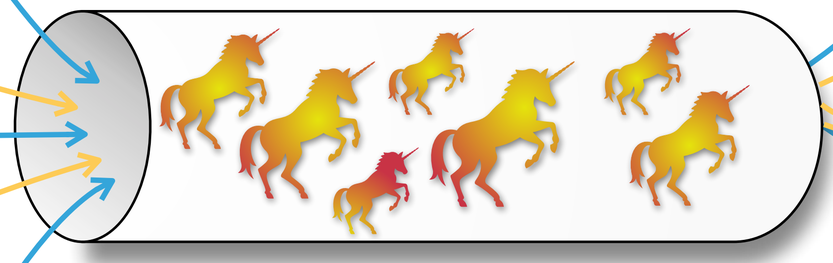
DATA PIPELINE

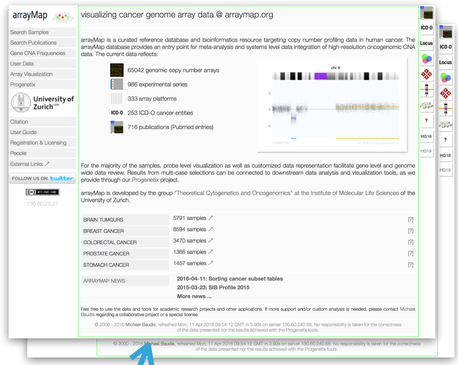


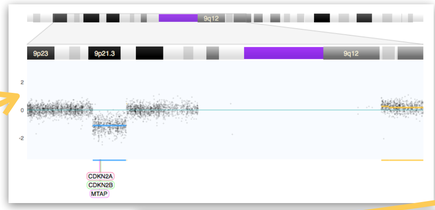


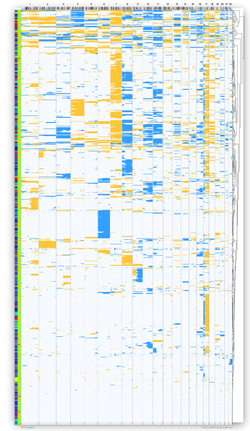


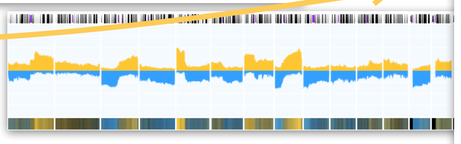


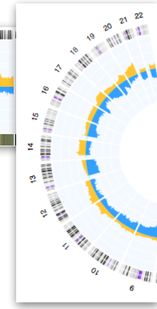


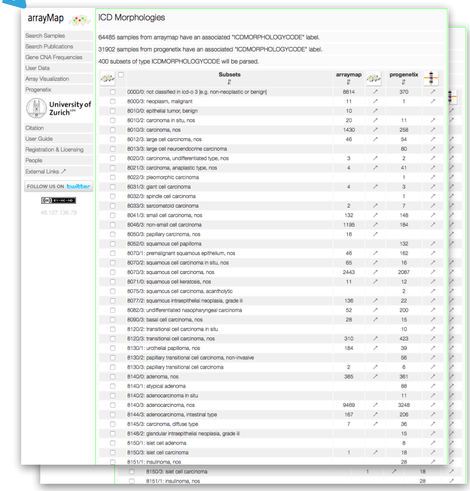








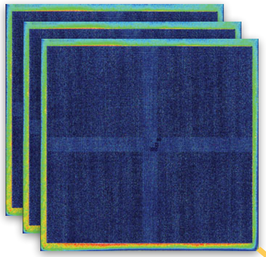


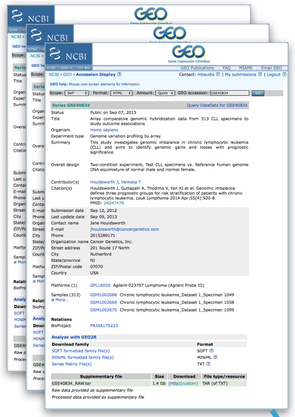


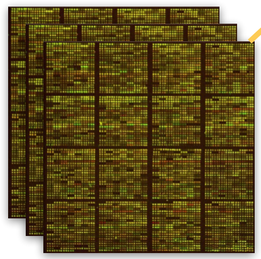
DATA PIPELINE

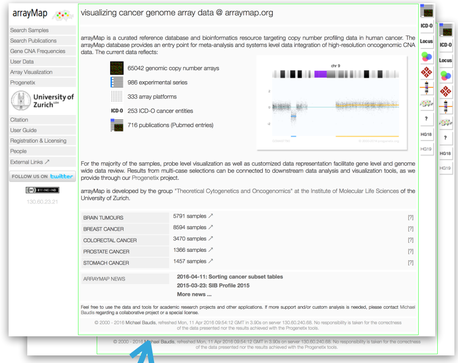
BIOCURATION

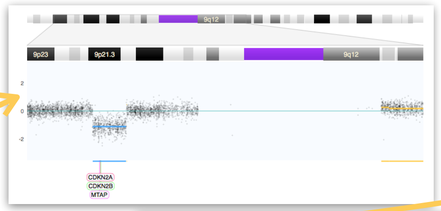
BIOINFORMATICS

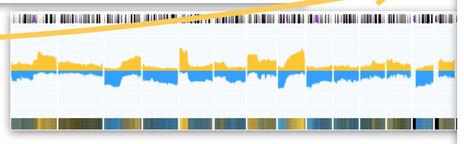


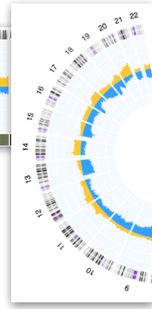


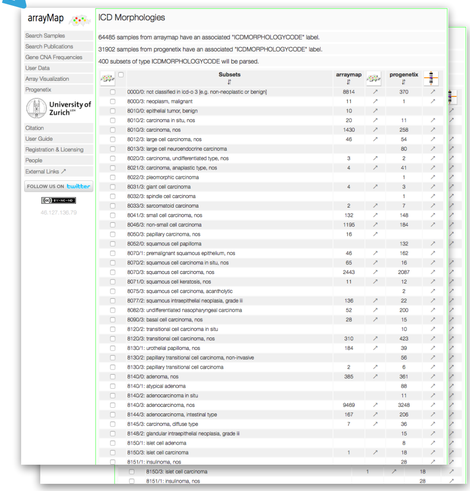












cancercelllines

arrayMap

progenetix


EVER
CLEANING
THE DATA
FITTING
THE MODEL
@kareem_carr

Data sets in tutorials

Data sets in the wild

Cancer Classifications need an Einstein to sort them out

University of Zurich[UZH]

ARRAYID	NCIT:CODE	NCIT:TERM	ICD-MORPHOLOGY-CODE	ICD-TOPOGRAPHY-CODE
GSM393758	C2773	Acute Myeloid Leukemia Not Otherwise Specified	9861/3	C42
GSM302275	C2852	Adenocarcinoma	8140/3	C34
GSM918983	C3222	Medulloblastoma	9470/3	C7169
GSM551398	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM412374	C3163	Chronic Lymphocytic Leukemia	9823/3	C42
GSM121876	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM714412	C2852	Adenocarcinoma	8140/3	C569
GSM110923	C9306	Soft Tissue Sarcoma	8800/3	C499
GSM711848	C2852	Adenocarcinoma	8140/3	C25
GSM746794	C89476		8077/2	C53
GSM198528	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM271399	C7949		8500/2	C50
GSM533469	C9349	Plasmacytoma	9731/3	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

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

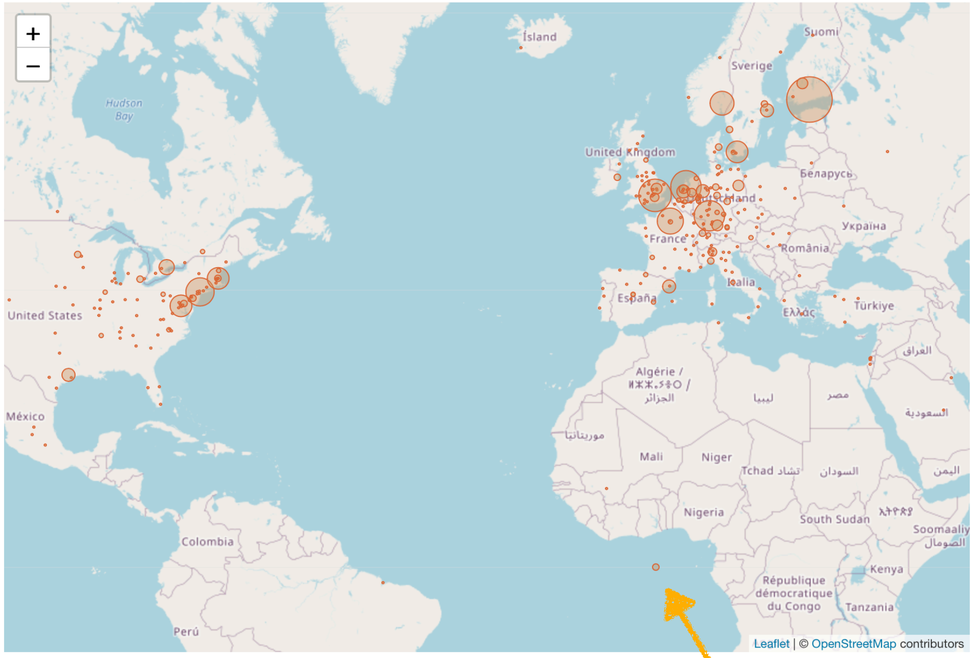
Filter prefix / local part	Code/Ontology	Example
pgx:icdom-...	ICD-O 3[7] Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3[7] Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix)[8]	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort-...	Progenetix cohorts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

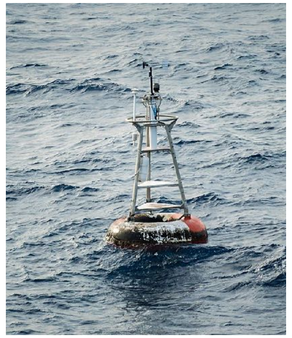
Progenetix publication collection
progenetix.org/publications/list
2020-11-28

25 / 3306 publications

Data Curation - Geolocations

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

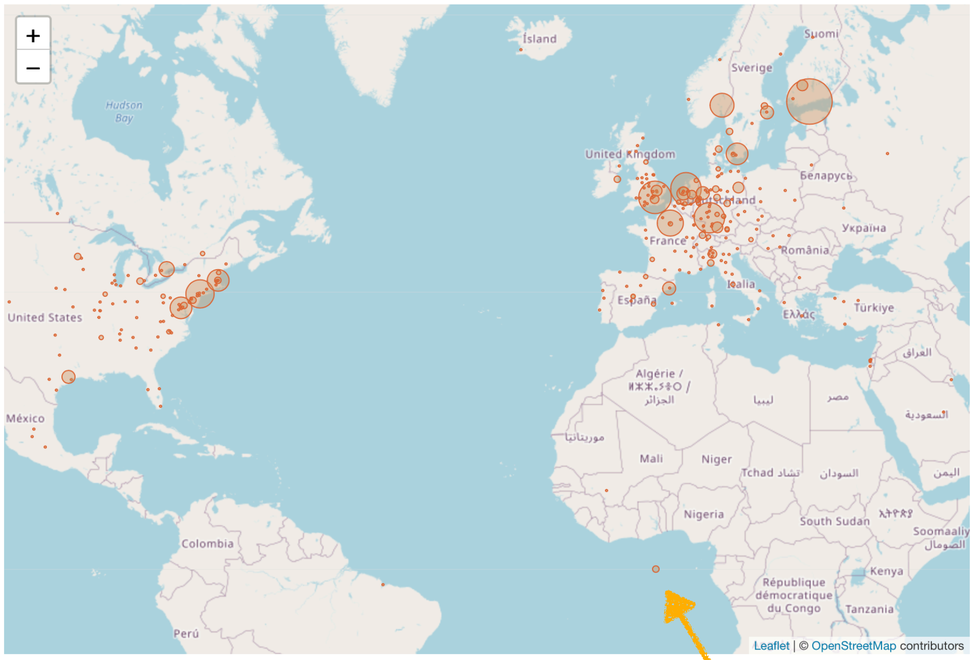
The most geo-tagged place on earth is Null Island



*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



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

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=pgxbs-kftvh94d&output=pgxseg

> variants <- pgxLoader(type="variant",biosample_id="pgxbs-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:

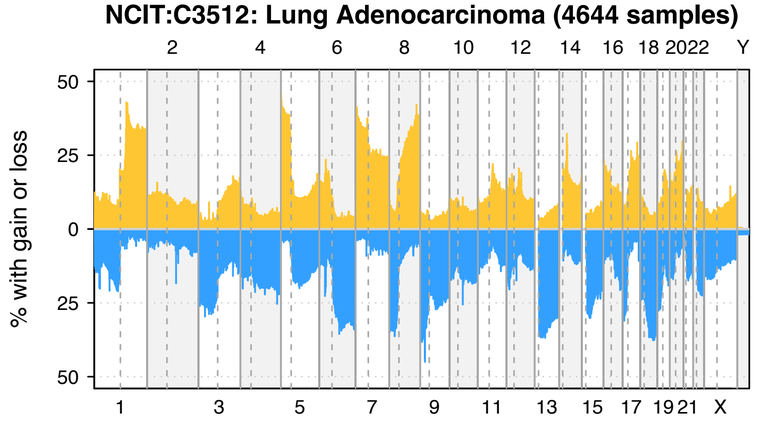
```
install.packages("devtools")
devtools::install_github("progenetix/pgxRpi")
```

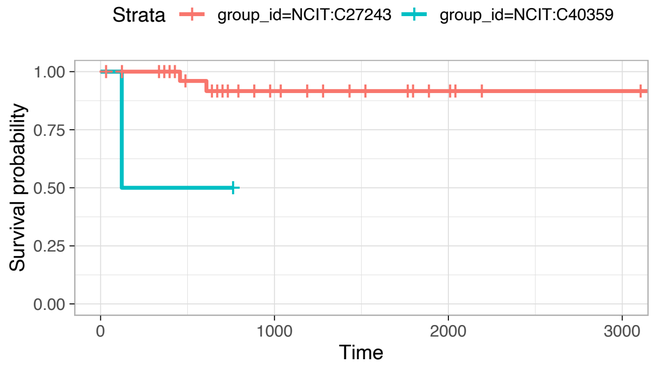
If you are interested in accessing CNV variant data, get started from this vignette

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.

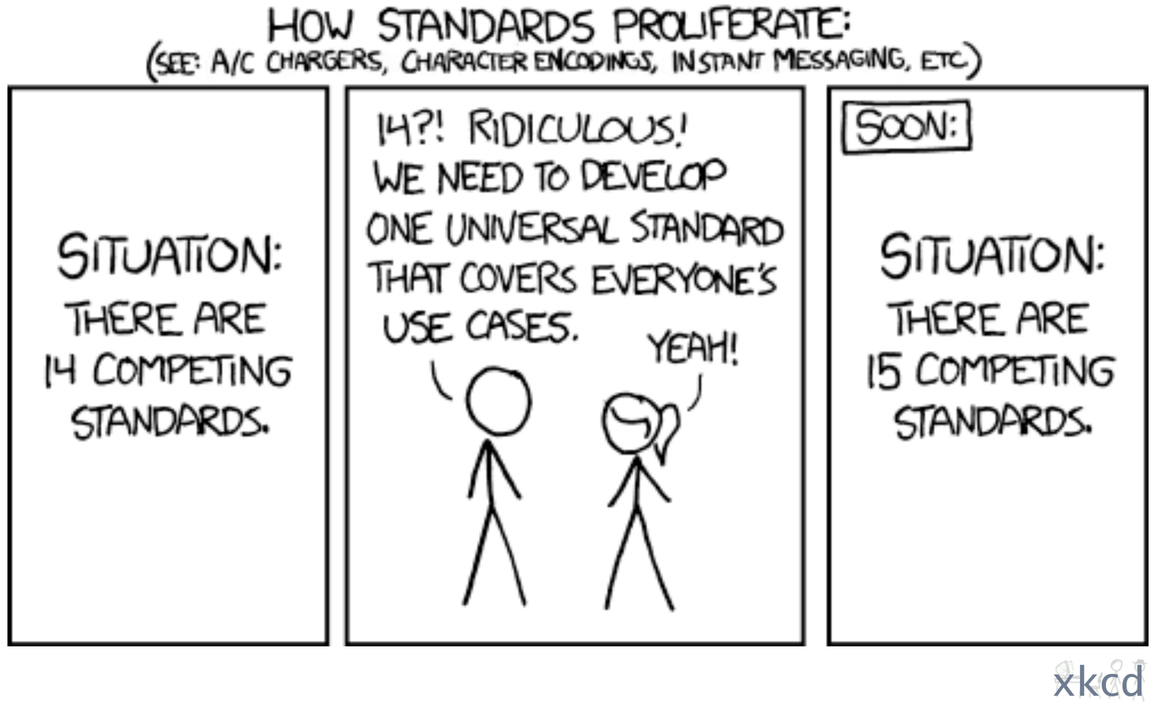




Lead: Hangjia Zhao

Standardized Data

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




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

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

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

GA4GH Standards Development

GENOMICS

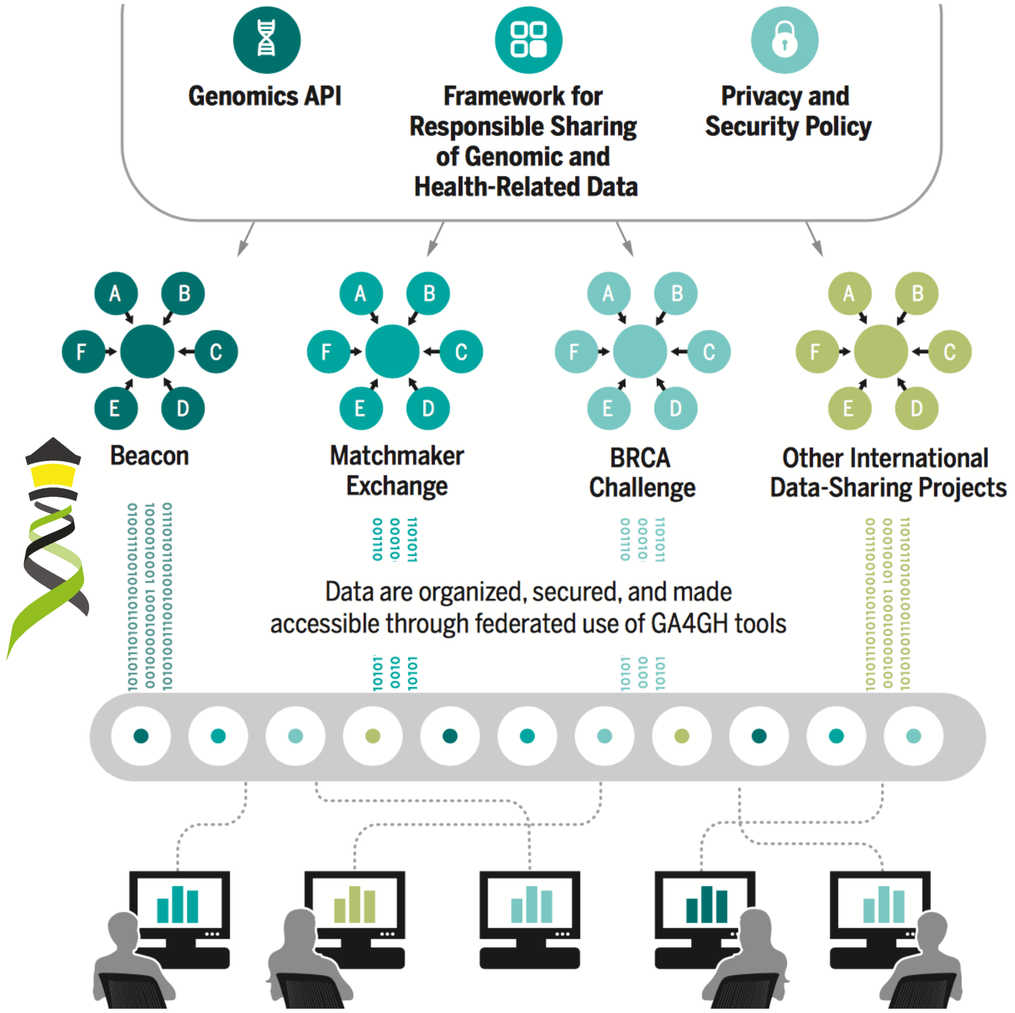
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



Global Alliance for Genomics & Health

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

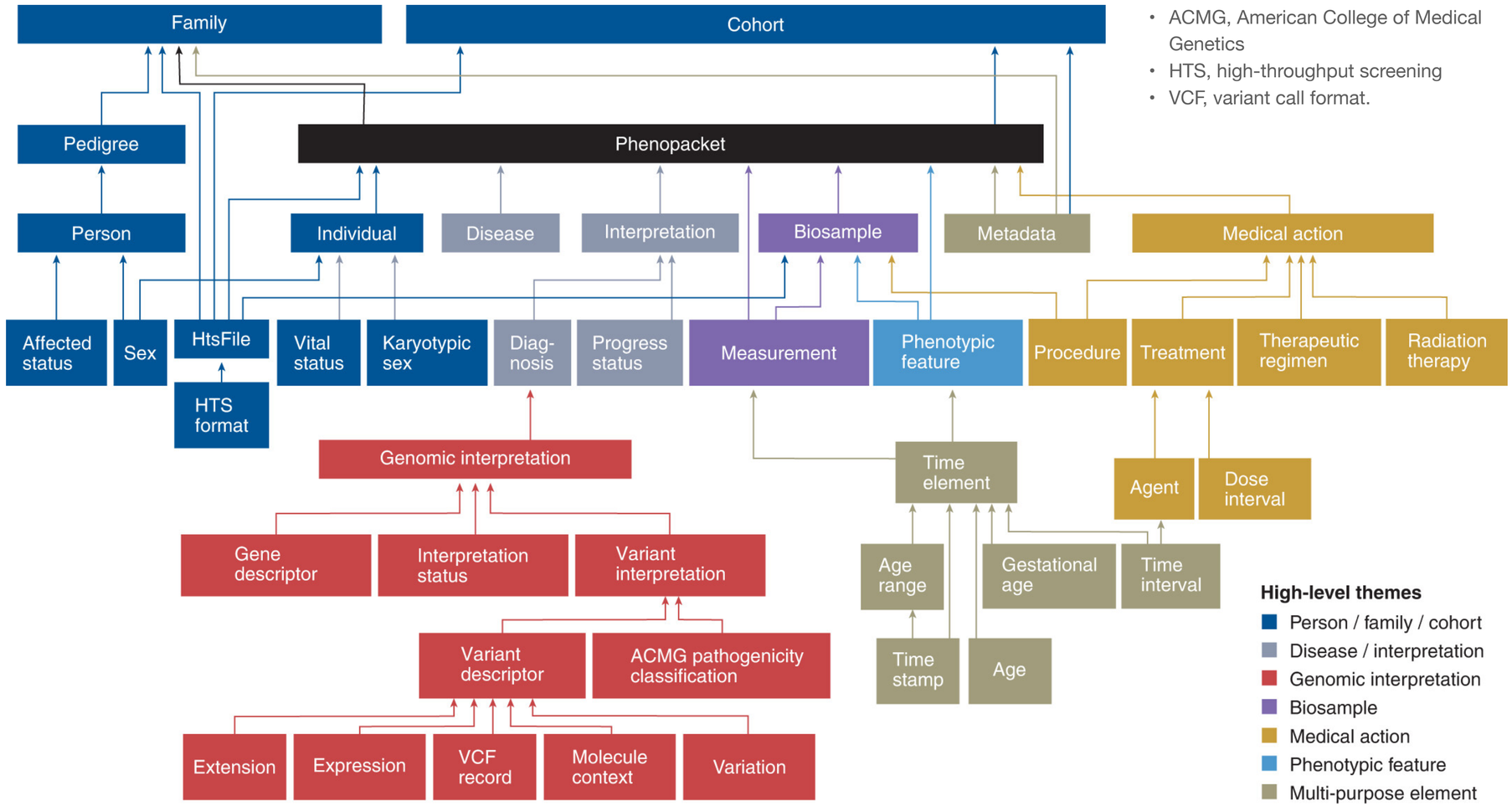
- Progenetix Beacon+ has served as implementation driver since 2016
- Beacon v2 as service with protocol-driven registries for federation
- GA4GH approved Beacon v2 in April 2022

Global Alliance for Genomics & Health

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 optio elements, each containir information about a certa such as phenotype, varia pedigree. An element ca other elements, which al hierarchical representatic

For instance, Phenopack contains elements of typ *Individual*, *PhenotypicFe Biosample* and so on. In elements can therefore b regarded as **building blc** larger structures.

Jacobsen JOB, Baudis M, Beckmann JS, Beltran S, Bus TJ, *et al.* 2022.
The GA4GH Phenopacket Scl Computable Representation of Clinical Data.
Nature Biotechnology 40 (6): 817–20.

Family | Cohort

Extension | Expression | VCF record | Molecule context | Variation

ISO Standards About us News Taking part **Store** Search

← ICS ← 35 ← 35.240 ← 35.240.80

ISO 4454:2022

Genomics informatics — Phenopackets: A format for phenotypic data exchange

Abstract

Preview

This document specifies a uniform, machine-readable, phenotypic description of an individual, patient or sample in the context of rare disease, common/complex disease or cancer.

It is applicable to academic, clinical and commercial research, as well as clinical diagnostics. While intended for human data collection, it can be used in other areas (e.g. mouse research). It does not define the phenotypic information that needs to be collected for a particular use but represents that information in an appropriately descriptive manner that allows it to be computationally exchanged between systems.

General information

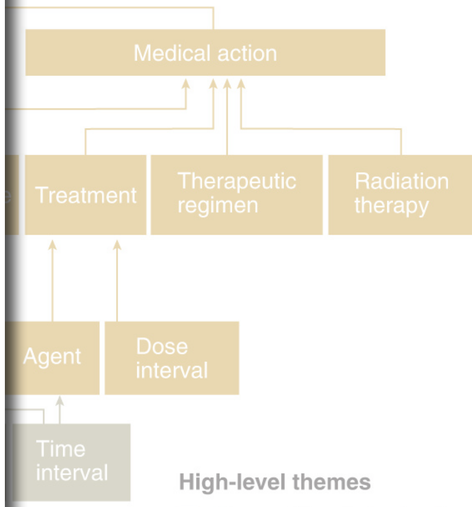
Status : Published
Publication date : 2022-07
Edition : 1
Number of pages : 86
Technical Committee : ISO/TC 215/SC 1 Genomics Informatics

Buy this standard

Format	Language
PDF + ePub	English
Paper	English

CHF 198 Buy

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



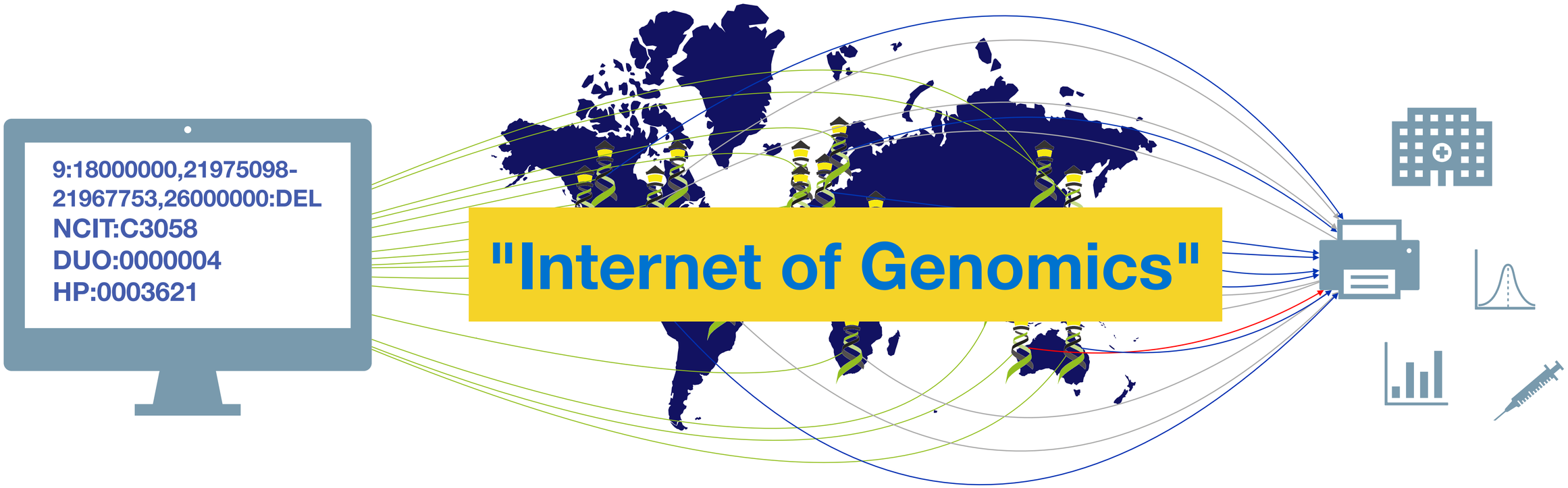
Maintaining some Standards

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 SVCLAIM=D [3]	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 SVCLAIM=D [3]	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 SVCLAIM=D [3]	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 SVCLAIM=D [3]	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 SVCLAIM=D [3]	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 SVCLAIM=D [3]	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 SVCLAIM=D [3]	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 SVCLAIM=D [3]	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

CellPress
OPEN ACCESS

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 Birney[19]

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 Drysda
Flicek, Francesco Florindi, Peter Goodhand, Ivo Gut, Jaap Hering
Juty, Thomas M. Keane, Jan O. Korbel, Ilkka Lappalainen, Brane
Mayrhofer, Andres Metspalu, Arcadi Navarro, Steven Newhouse
Persson, Aarno Palotie, Helen Parkinson, Jordi Rambla, David Sa
Swertz, Alfonso Valencia, Susheel Varma, Niklas Blomberg & Serena Scollen

— Show fewer authors

Nature Reviews Genetics **20**, 693–701 (2019) | Cite this article

ADVANCED GENETICS

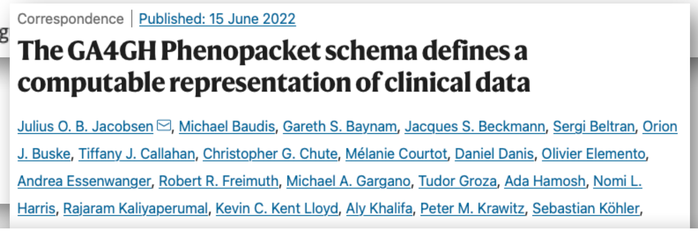
Open Access

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

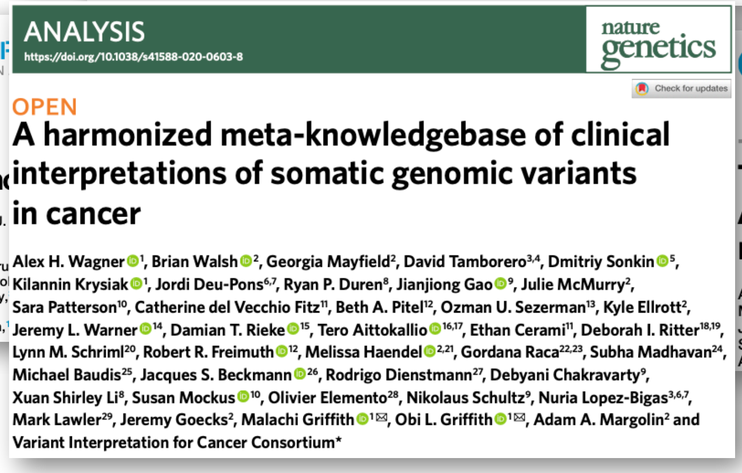
Julius O. B. Jacobsen, Michael Baudis, Gareth S. Baynam, Jacques S. Beckmann, Sergi Beltran, Orion J. Buske, Tiffany J. Callahan, Christopher G. Chute, Mélanie Courtot, Daniel Danis, Olivier Elemento, Andrea Essenwanger, Robert R. Freimuth, Michael A. Gargano, Tudor Groza, Ada Hamosh, Nomi L. Harris,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

Cell Genomics

Perspective

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

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,[3,9] 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,[19] Michael Brudno,[18,19,20,21,38] Matthew H. Bru David Bujold,[9,18,38] Tony Burdett,[23] Orion J. Buske,[24] Moran N. Cabili,[1] Daniel L. Cameron,[25,26] Robert J. Carro Esmeralda Casas-Silva,[123] Debyani Chakravarty,[29] Bimal P. Chaudhari,[30,31] Shu Hui Chen,[32] J. Michael Cherry, Justina Chung,[3,4] Melissa Cline,[34] Hayley L. Clissold,[15] Robert M. Cook-Deegan,[35] Mélanie Courtot,[23] Fiona Cunningham,[23] Miro Cupak,[6] Robert M. Davies,[15] Danielle Denisko,[19] Megan J. Doerr,[36] Lena I. Dolman,



ANALYSIS

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

nature genetics

Check for updates

OPEN

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

Alex H. Wagner[1], Brian Walsh[2], Georgia Mayfield[2], David Tamborero[3,4], Dmitriy Sonkin[5], Kilannin Krysiak[1], Jordi Deu-Pons[6,7], Ryan P. Duren[8], Jianjiong Gao[9], Julie McMurry[2], Sara Patterson[10], Catherine del Vecchio Fitz[11], Beth A. Pitel[12], Ozman U. Sezerman[13], Kyle Ellrott[2], Jeremy L. Warner[14], Damian T. Rieke[15], Tero Aittokallio[16,17], Ethan Cerami[11], Deborah I. Ritter[18,19], Lynn M. Schriml[20], Robert R. Freimuth[12], Melissa Haendel[2,21], Gordana Raca[22,23], Subha Madhavan[24], Michael Baudis[25], Jacques S. Beckmann[26], Rodrigo Dienstmann[27], Debyani Chakravarty[9], Xuan Shirley Li[8], Susan Mockus[10], Olivier Elemento[28], Nikolaus Schultz[9], Nuria Lopez-Bigas[3,6,7], Mark Lawler[29], Jeremy Goecks[2], Malachi Griffith[1✉], Obi L. Griffith[1✉], Adam A. Margolin[2] and Variant Interpretation for Cancer Consortium*

Cell Genomics

CellPress
OPEN ACCESS

Technology

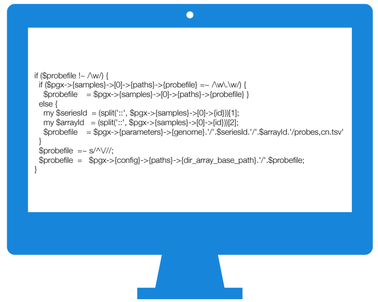
The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification

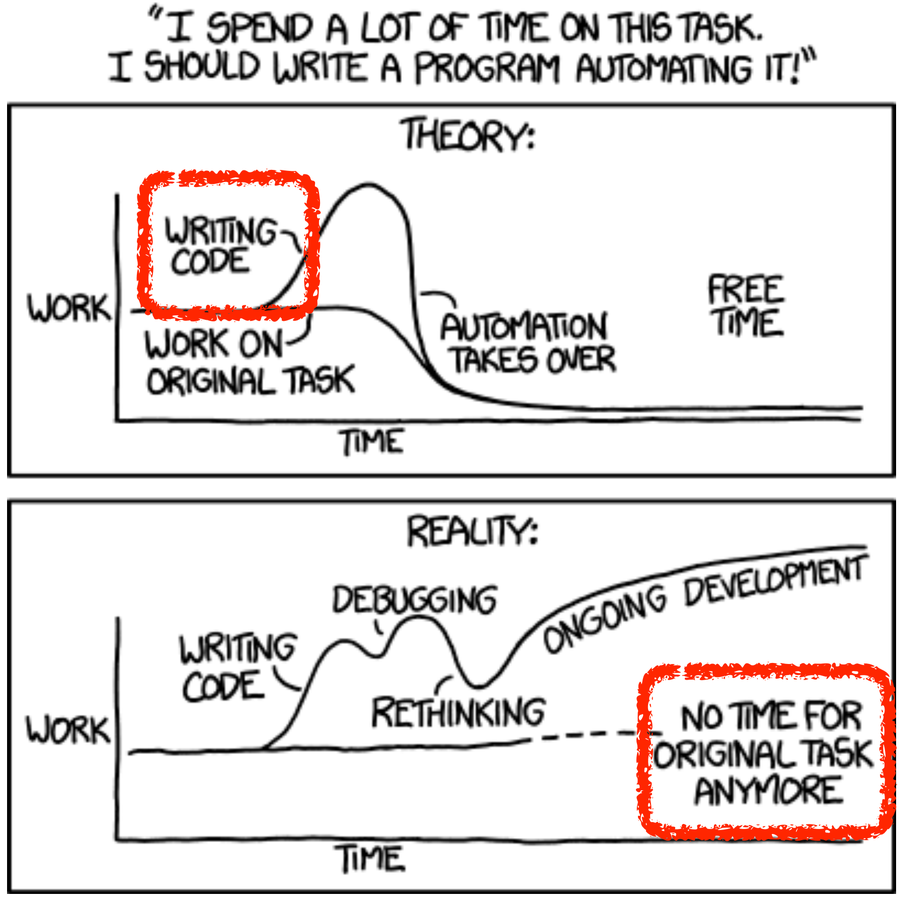
Alex H. Wagner,[1,2,25,*] Lawrence Babb,[3,*] Gil Alterovitz,[4,5] Michael Baudis,[6] Matthew Brush,[7] Daniel L. Cameron,[8,9] Melissa Cline,[10] Malachi Griffith,[11] Obi L. Griffith,[11] Sarah E. Hunt,[12] David Kreda,[13] Jennifer M. Lee,[14] Stephanie Li,[15] Javier Lopez,[16] Eric Moyer,[17] Tristan 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}

```
if ($probefile !~ /\w/) {
  if ($pgx->{samples}->[0]->{paths}->{probefile} =~ /\w\.\w/) {
    $probefile    = $pgx->{samples}->[0]->{paths}->{probefile} }
  else {
    my $seriesId  = (split('::', $pgx->{samples}->[0]->{id}))[1];
    my $arrayId   = (split('::', $pgx->{samples}->[0]->{id}))[2];
    $probefile    = $pgx->{parameters}->{genome}.'/'.$seriesId.'/'.$arrayId.'/
probes,cn.tsv'
  }
  $probefile  =~ s/^\///;
  $probefile  =   $pgx->{config}->{paths}->{dir_array_base_path}.'/'.$probefile;
}
```

{bio_informatics_science}





Randall Munroe - XKCD
https://xkcd.com/1319/

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

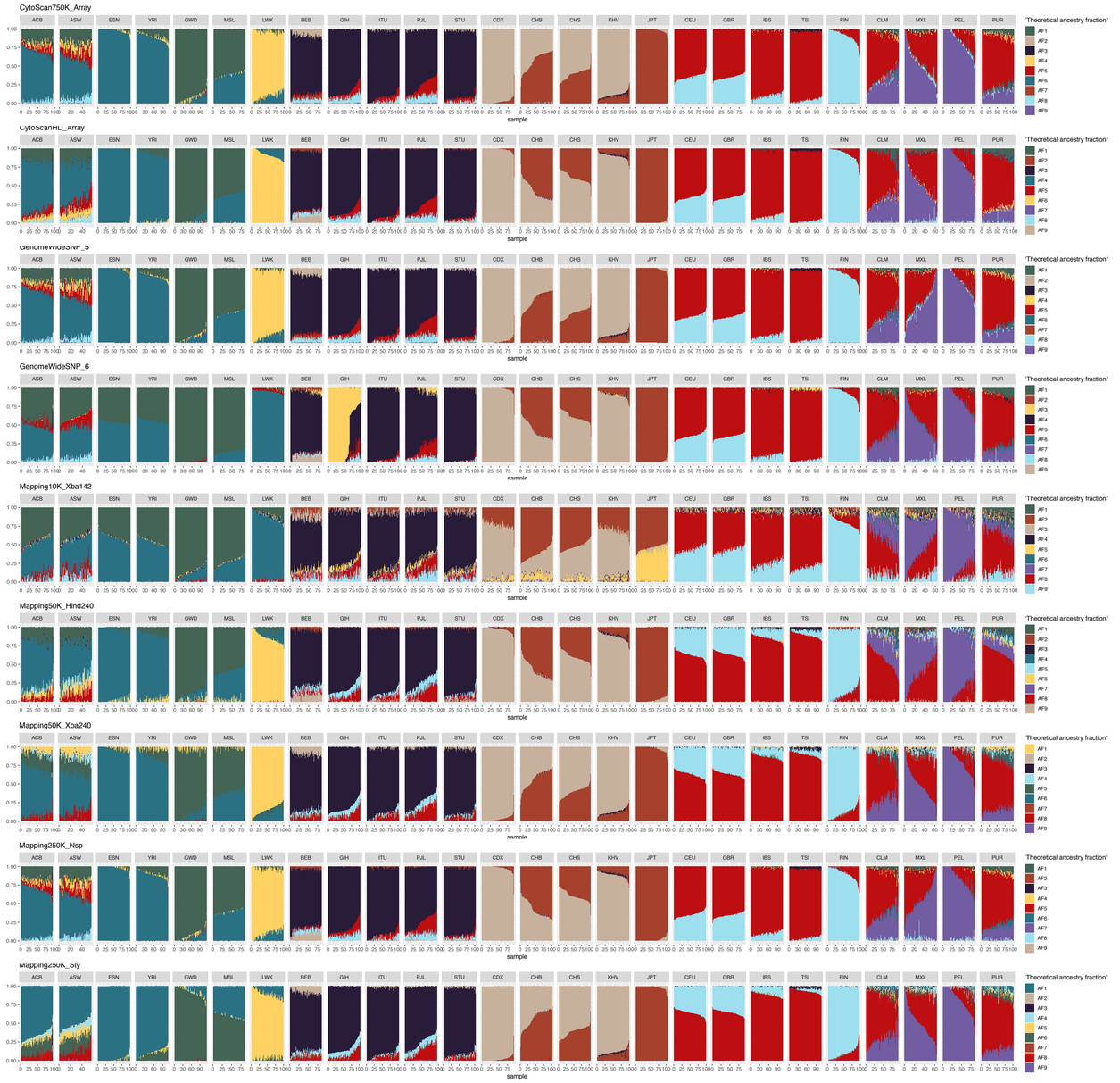


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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

arrayMap

Qingyao Huang

frontiers in Genetics

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

Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes

Bo Gao [1,2] *and Michael Baudis* [1,2*]

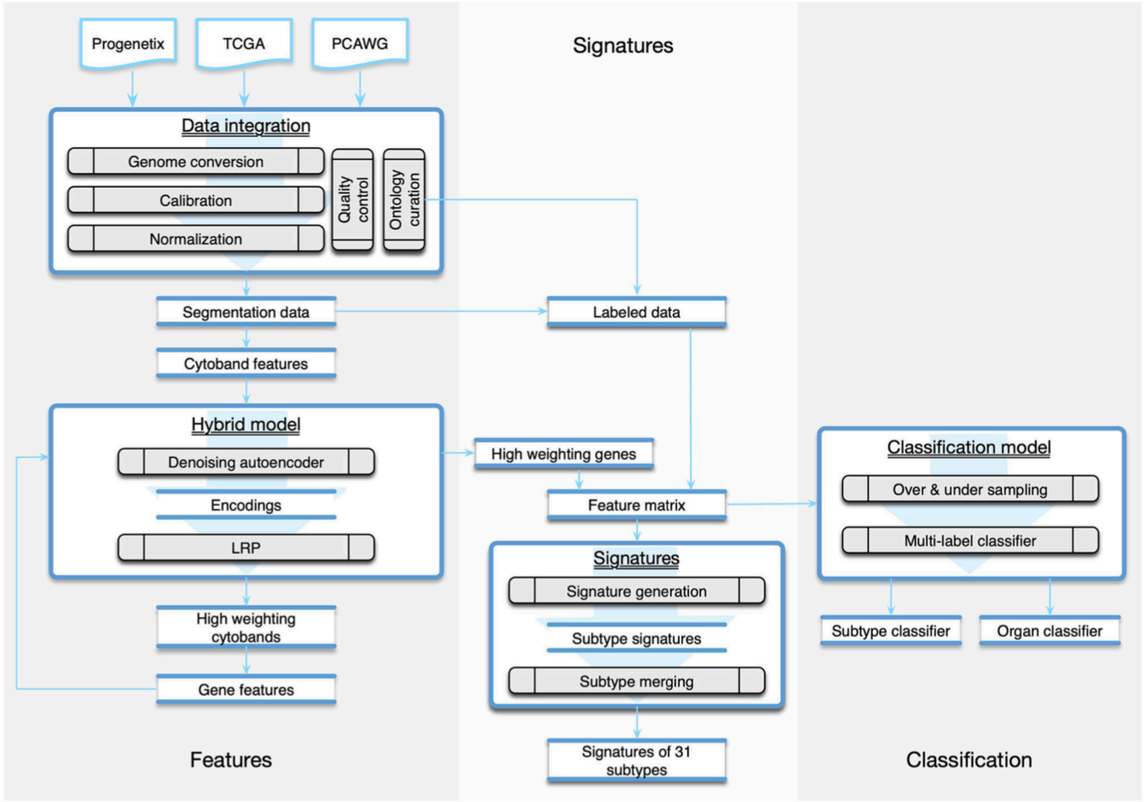


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine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

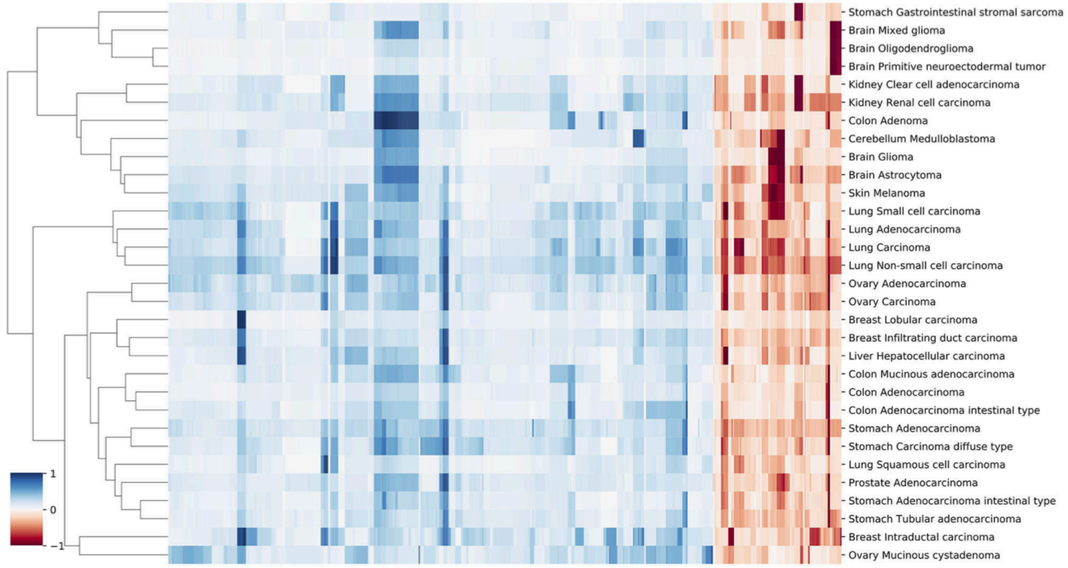


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

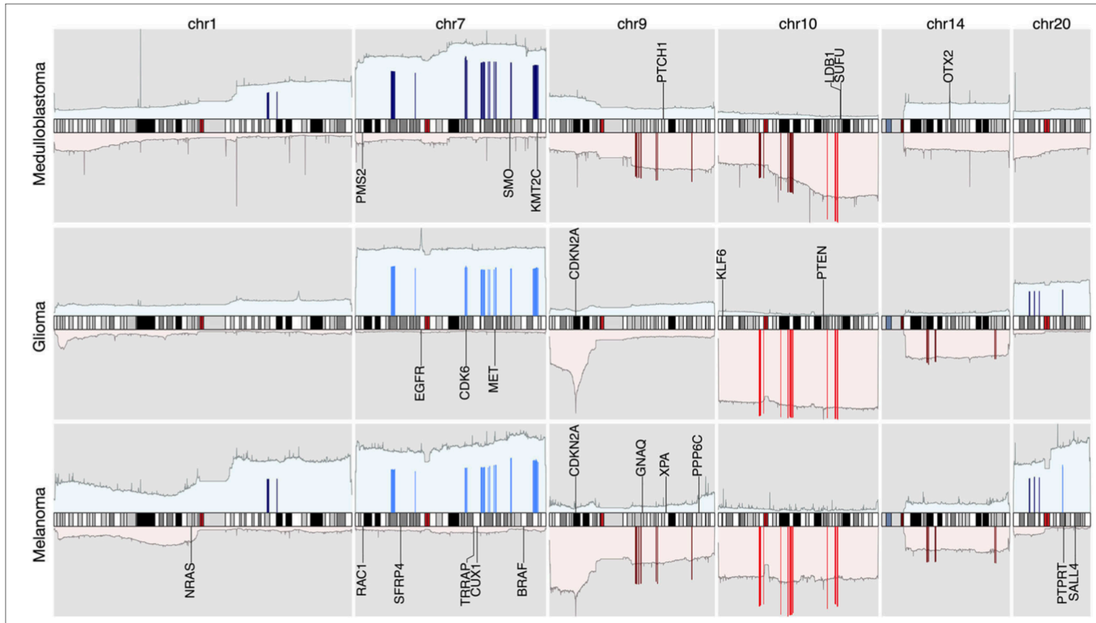
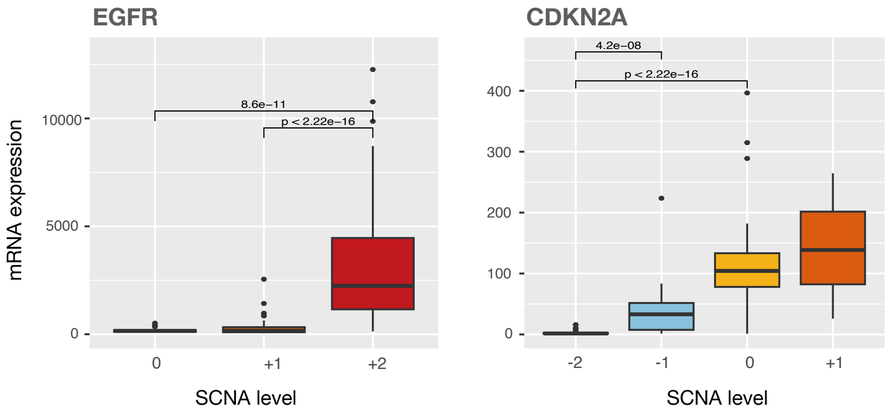


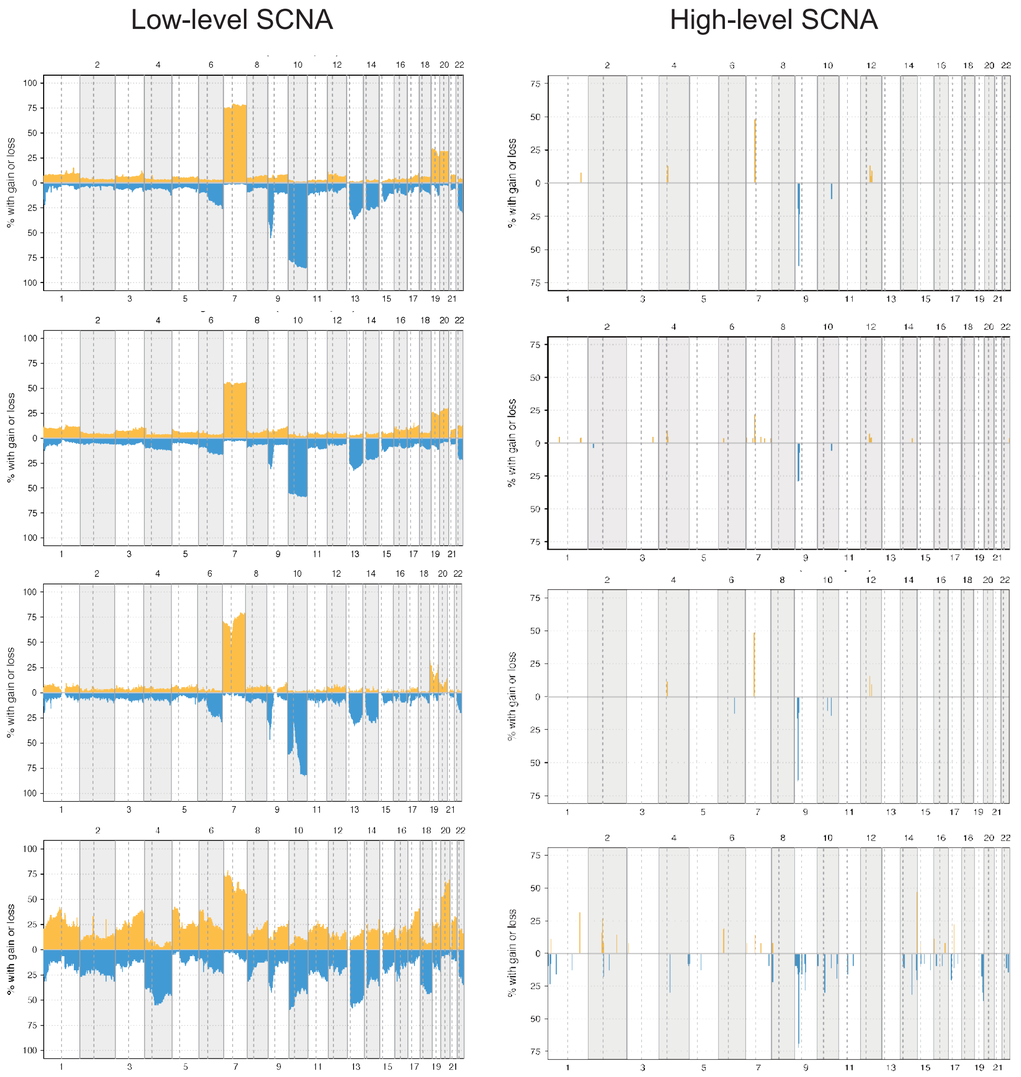
FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger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

LabelSeg

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



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

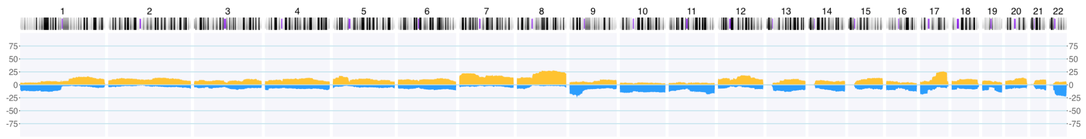


Hangjia Zhao

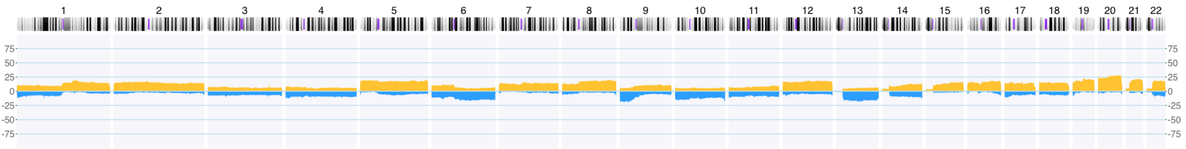
CNV profiles heterogeneity vs cancer classification

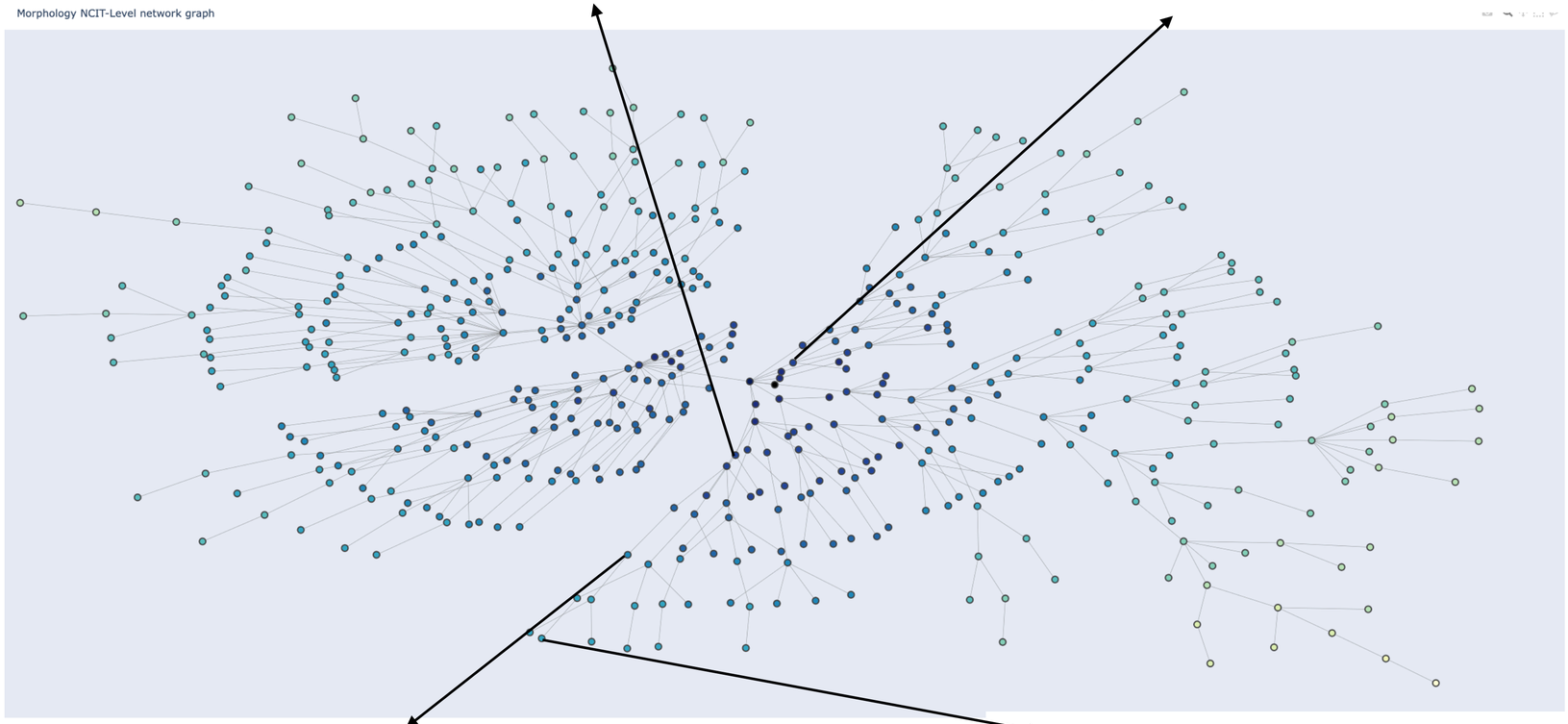
Correspondance of genomic profiles to NCIT cancer hierarchy

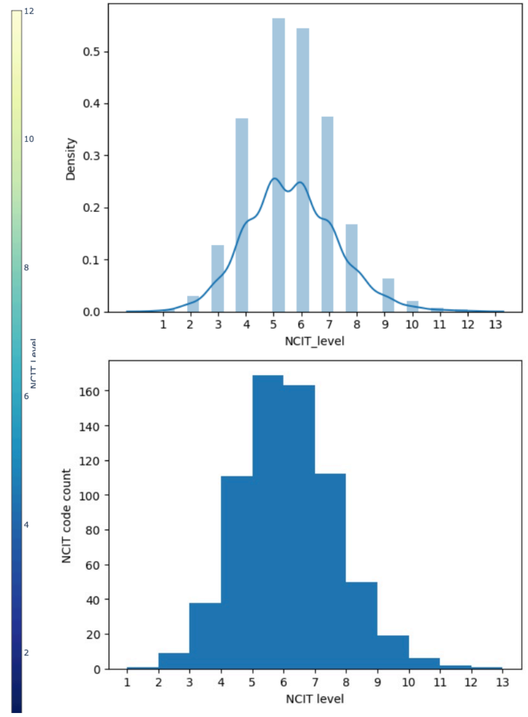
Nerve Sheath Neoplasm (NCIT:C4972)



Chondrogenic Neoplasm (NCIT:C4755)

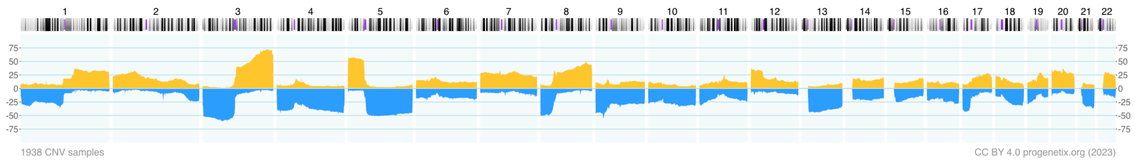




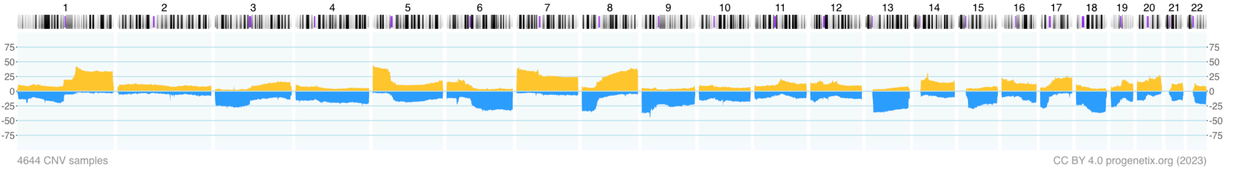


Lead: Ziying Yang

Lung Squamous Cell Carcinoma (NCIT:C3493)



Lung Adenocarcinoma (NCIT:C3512)



Genomic Profiles as Fingerprints of Cancer Types

cancercelllines

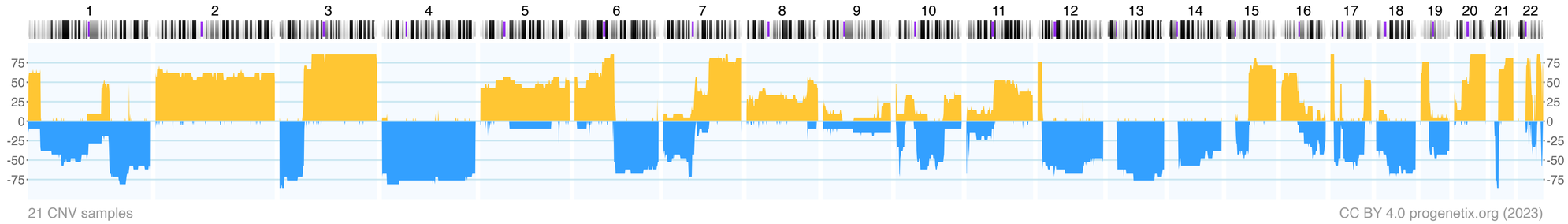
MDA-MB435: A widely used Breast cancer cell line ...

Cellosaurus

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)



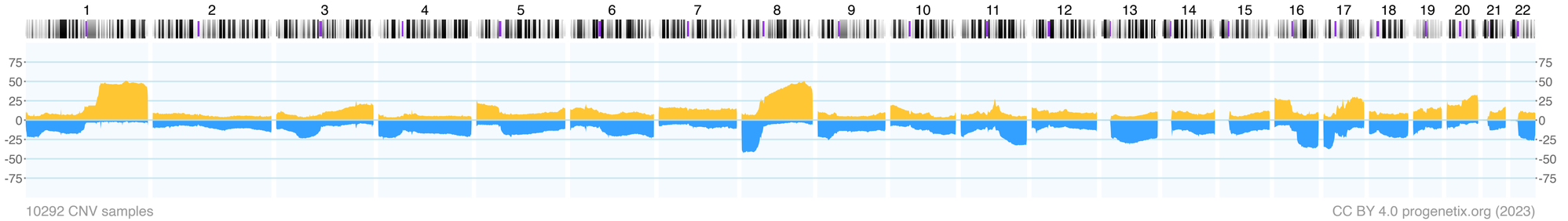
Lead: Rahel Paloots

Genomic Profiles as Fingerprints of Cancer Types

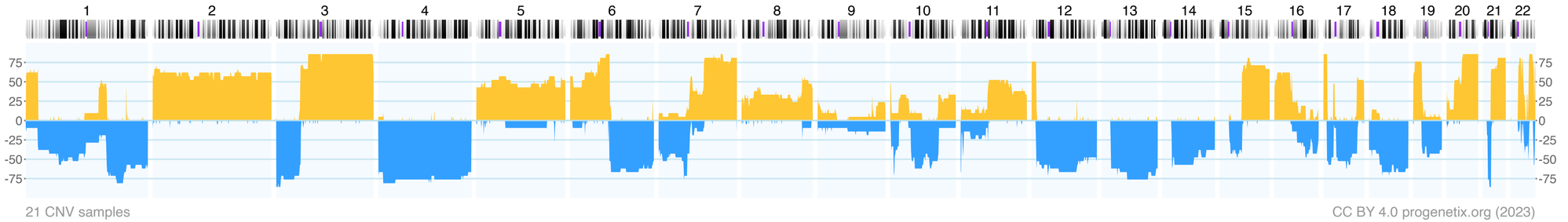
cancercelllines

An unpleasant surprise...

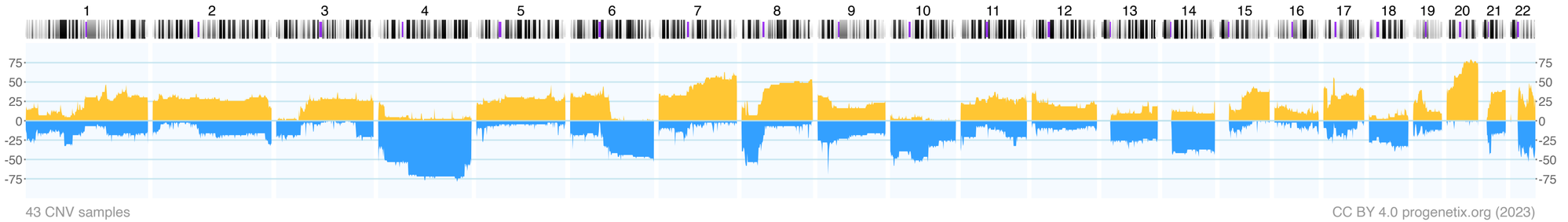
Ductal Breast Carcinoma (NCIT:C4017)



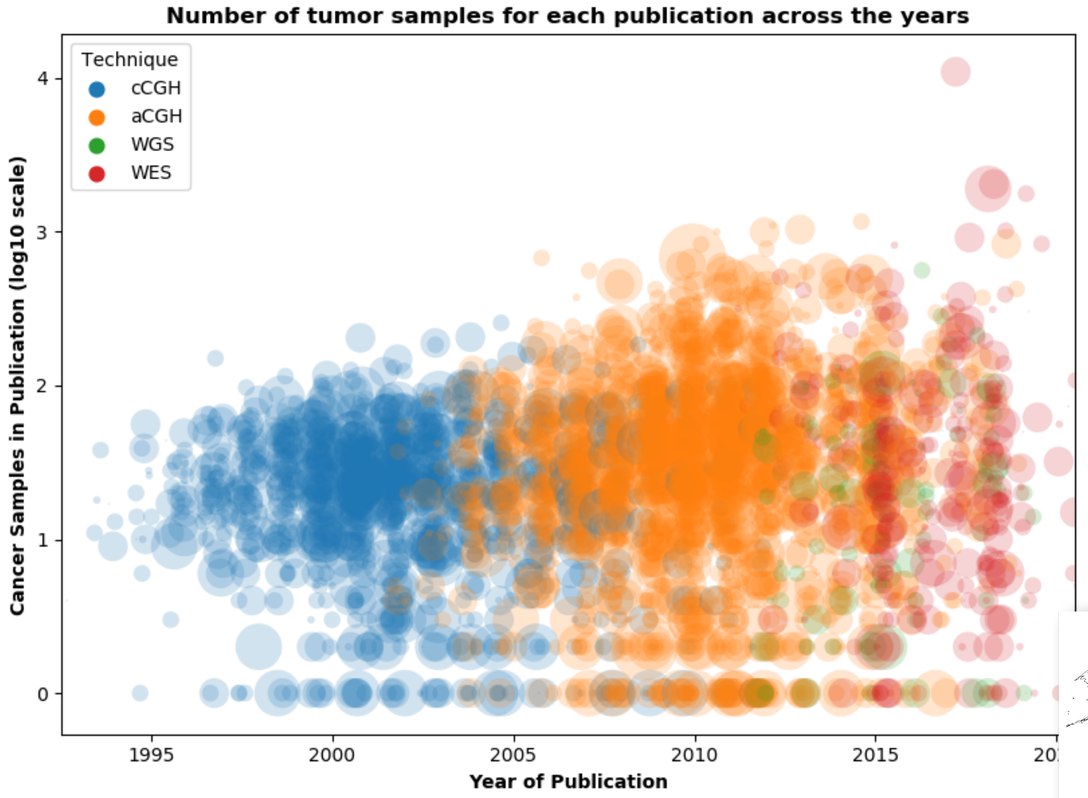
MDA-MB-435 (cellosaurus:CVCL_0417)



Amelanotic Melanoma (NCIT:C3802)



Lead: Rahel Paloots



progenetix

- 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.

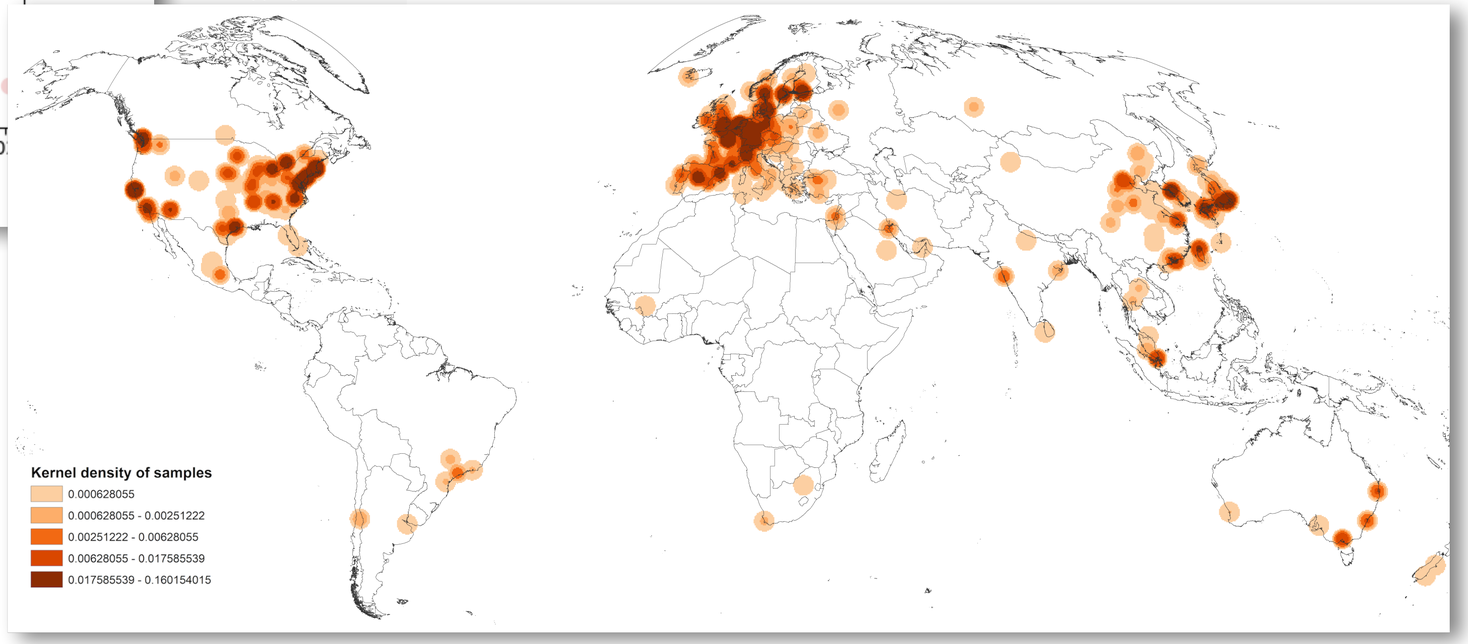
Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.

Filter

City

Type to search...

Publications (3324)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations	0	79	0	0	0



DATABASE
The Journal of Biological Databases and Curation

Database, 2020, 1–9
doi: 10.1093/databa/baaa009
Articles

Articles

Geographic assessment of cancer genome profiling studies

Paula Carrio-Cordo[1,2], Elise Acheson[3], Qingyao Huang[1,2] and Michael Baudis[1,*]

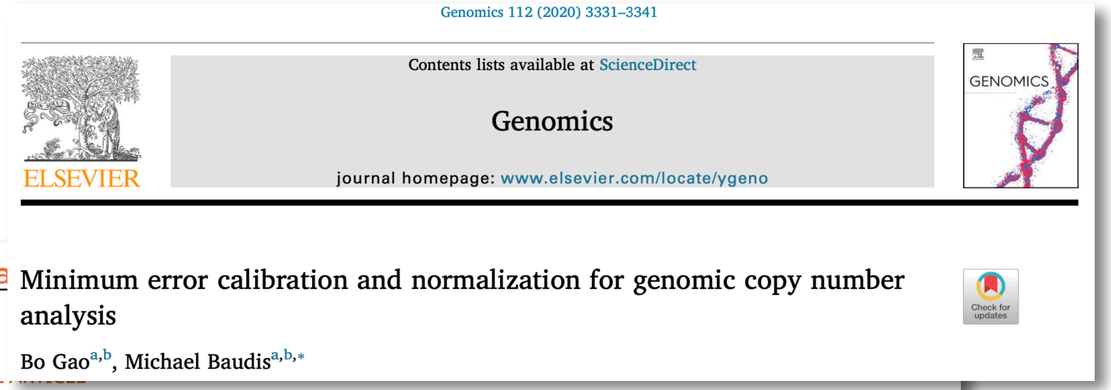
[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland [2]Swiss Institute of Bioinformatics, Zurich, Switzerland [3]Department of Geography, University of Zurich, Zurich, Switzerland

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


Genomics 112 (2020) 3331–3341

Contents lists available at ScienceDirect

Genomics

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

Minimum error calibration and normalization for genomic copy number analysis

Bo Gao[a,b], Michael Baudis[a,b,*]

SOFTWARE TOOL ARTICLE

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

Bo Gao[1,2], Qingyao Huang[1,2], Michael Baudis[1,2]

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

Roman Muff[1], Prisni Rath[2], Ram Mohan Ram Kumar[1], Knut Husmann[1], Walter Born[1], Michael Baudis[2], Bruno Fuchs[1]*

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

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]*

ORIGINAL PAPER

CNARA: reliability assessment fo genomic copy number profiles

Ni Ai[1]*, Haoyang Cai[2], Caius Solovan[3] and Michael Baudis[1]*

SCIENTIFIC REPORTS
natureresearch

Enabling population assignment from cancer genomes with SNP2pop

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

JEADV JOURNAL OF THE EUROPEAN ACADEMY OF DERMATOLOGY & VENEREOLOGY

Full Access

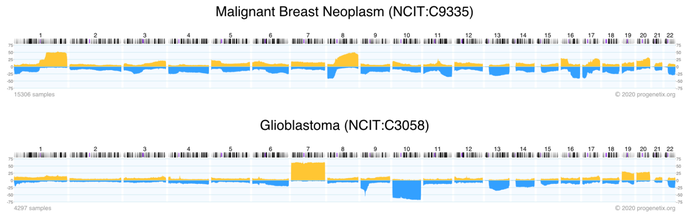
umber imbalances in primary cutaneous lymphomas

G. Gug, Q. Huang, E. Chiticariu, C. Solovan, M. Baudis

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"



progenetix

(Bio)informatics Skill Set

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

GitHub

text mining

regular expressions

```
s/knowledge/mastery/
```

React

mongoDB

APACHE
HTTP SERVER PROJECT

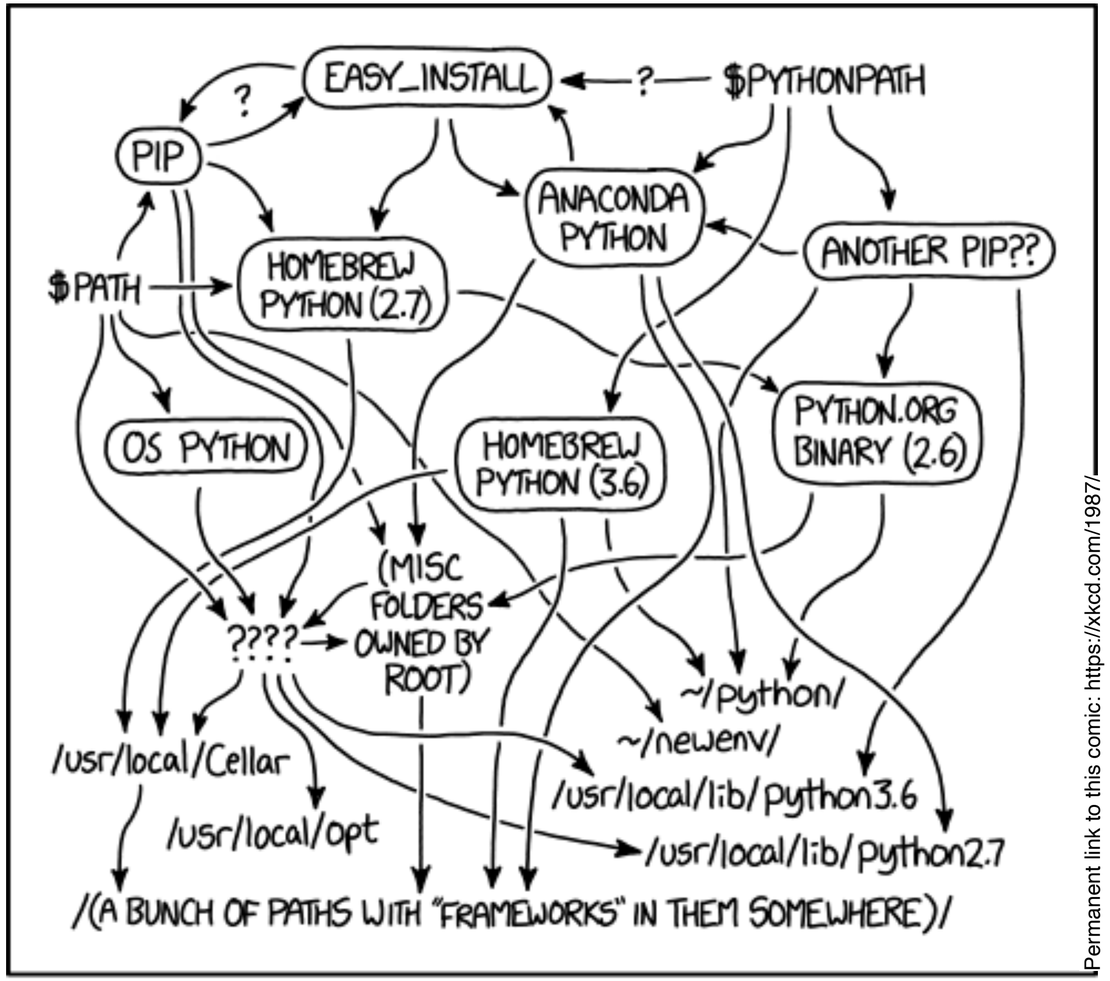
array pipelines

statistics

python™

HTML CSS JS

But in reality that is what bioinformaticians do...



xkcd

info.baudisgroup.org