Implementation Driven Development of Standards for Genomic Data Exchange from Cancer Genome Data Collections

CNV Databases :: Variant Representation & Query Formats :: ELIXIR Beacon :: GA4GH

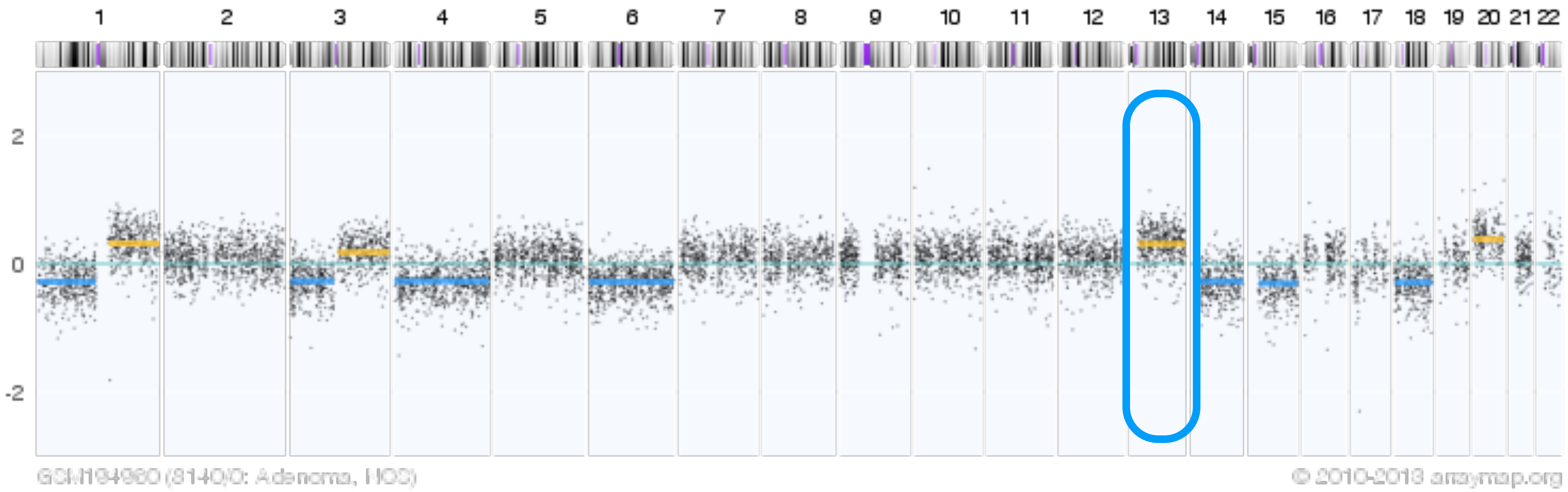
University of Zurich UZH

elixir

SIB

Global Alliance for Genomics & Health

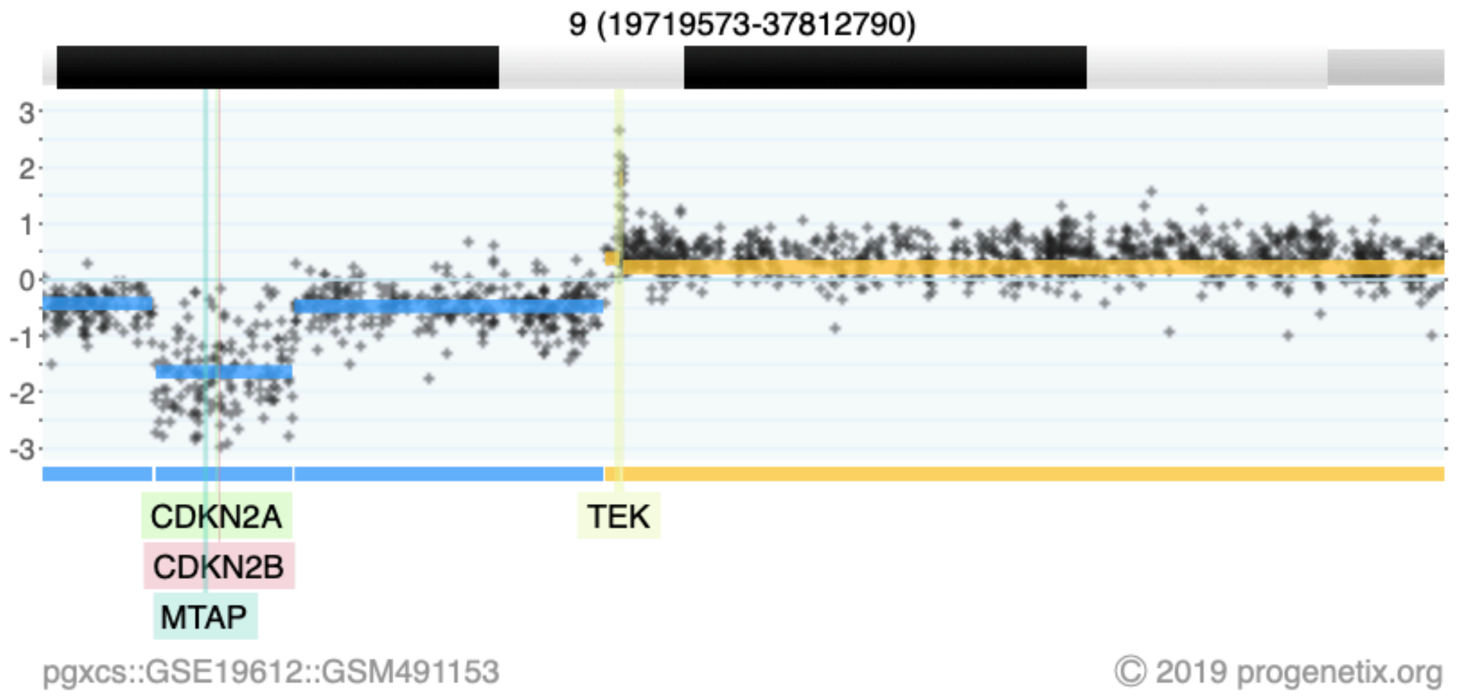
Somatic Copy Number Variations



GSM164980 (8140/0: Adenoma, NOS)

© 2010-2013 arraymap.org

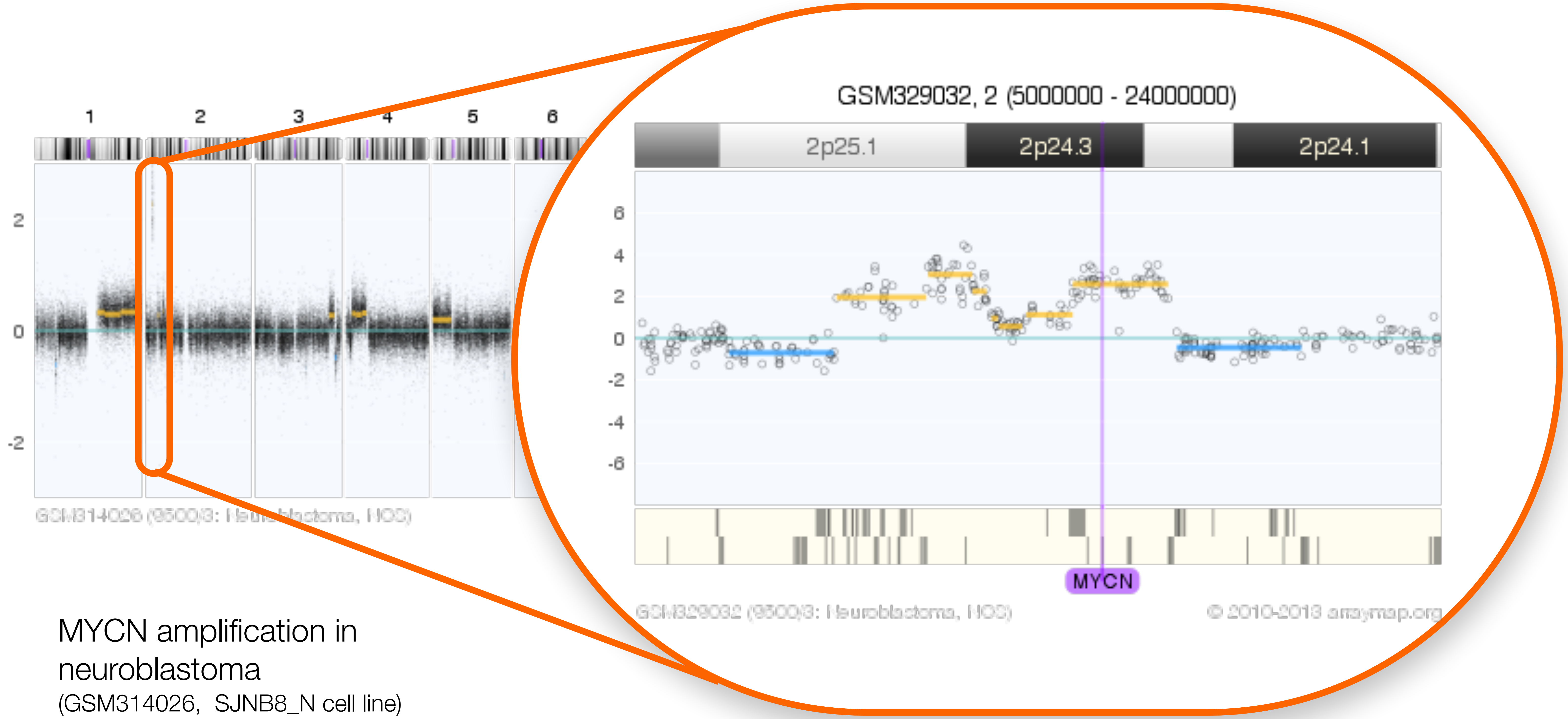
Gain of chromosome arm 13q in colorectal carcinoma



pgxcs::GSE19612::GSM491153

© 2019 progenetix.org

2-event, homozygous deletion in a Glioblastoma

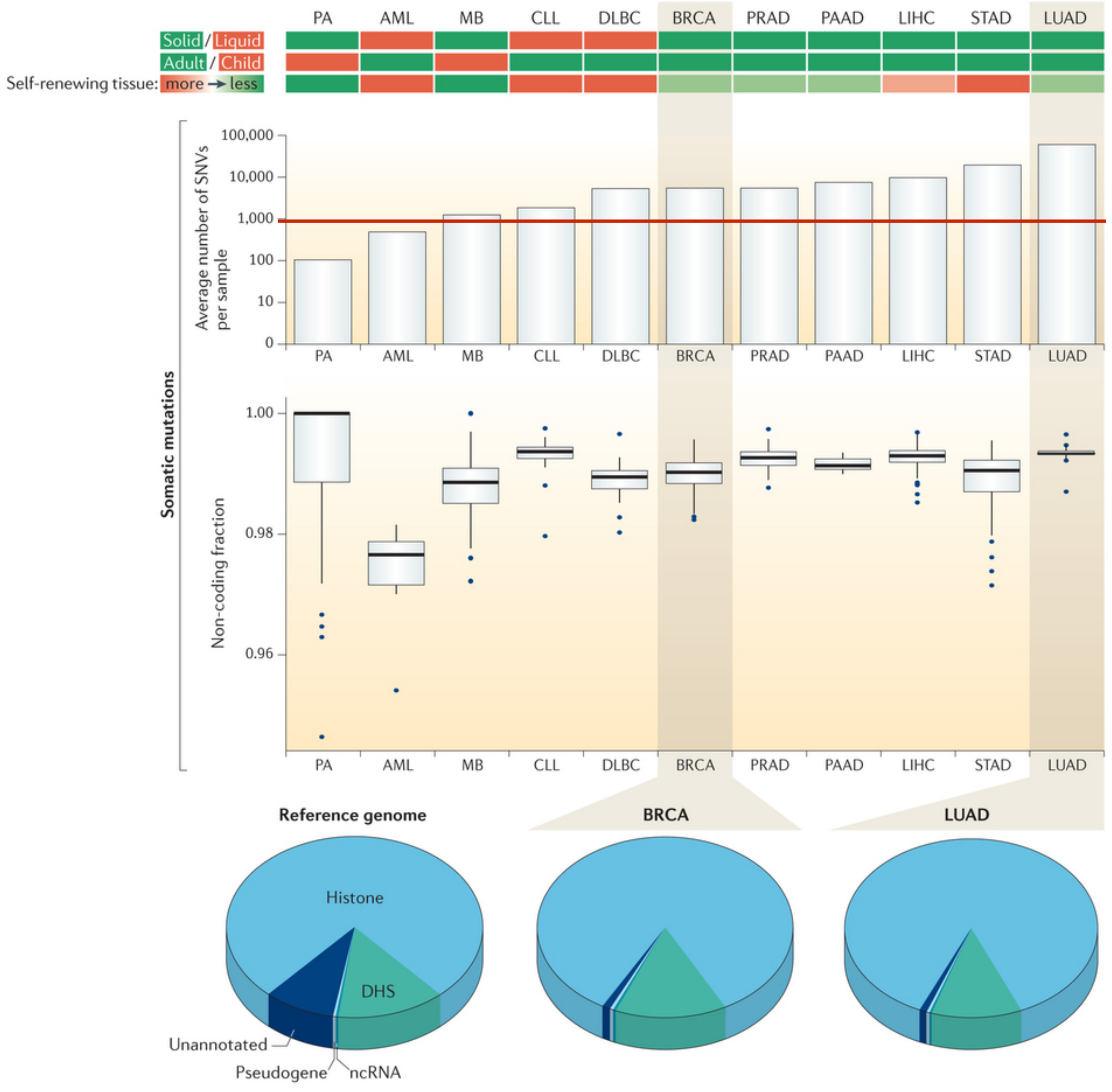


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

low level/high level copy number alterations (CNAs)

arrayMap

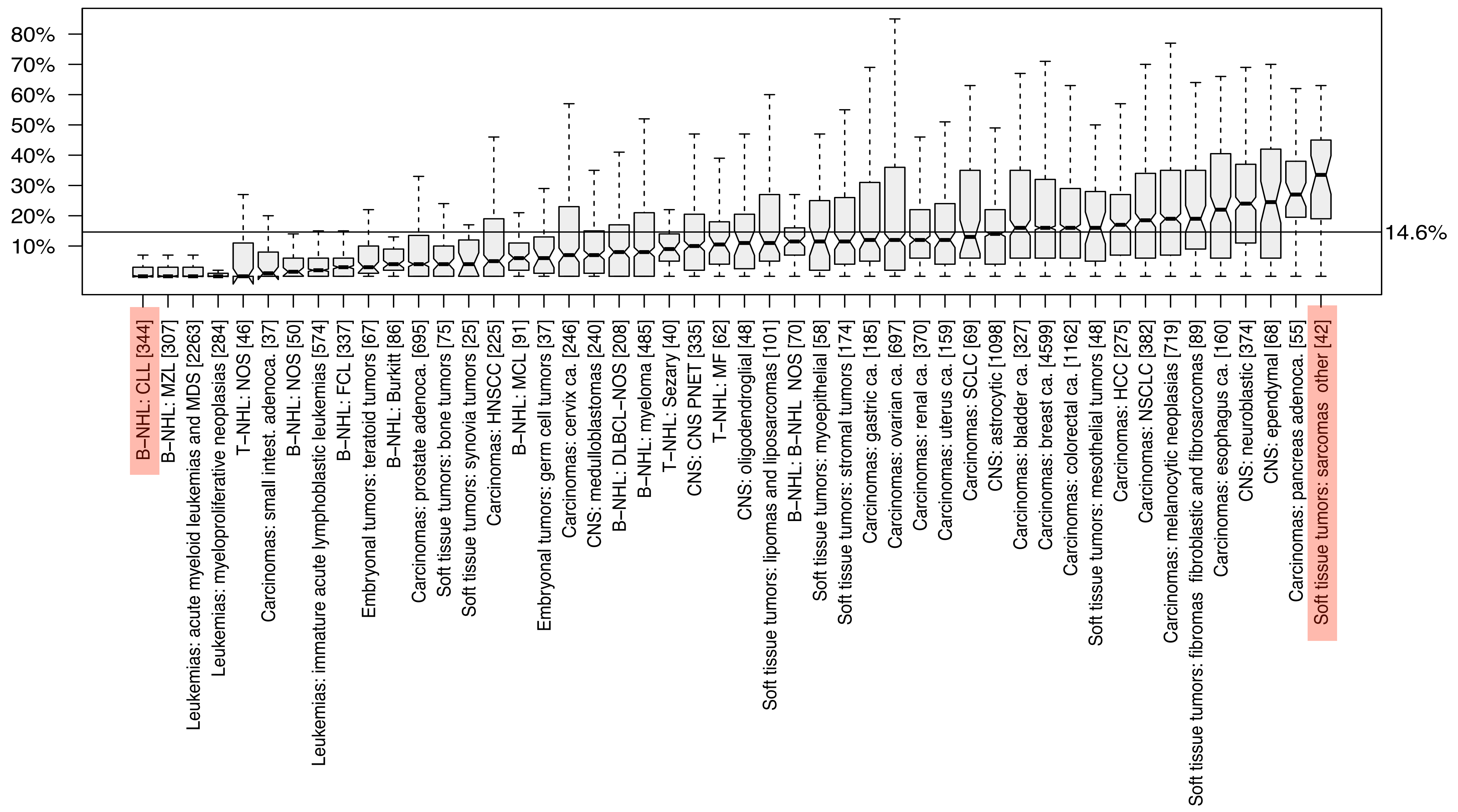
Quantifying Somatic Mutations In Cancer



Nature Reviews | Genetics

CANCERS SHOW THOUSANDS OF SINGLE NUCLEOTIDE VARIANTS PER SAMPLE, MOSTLY IN NON-CODING REGIONS

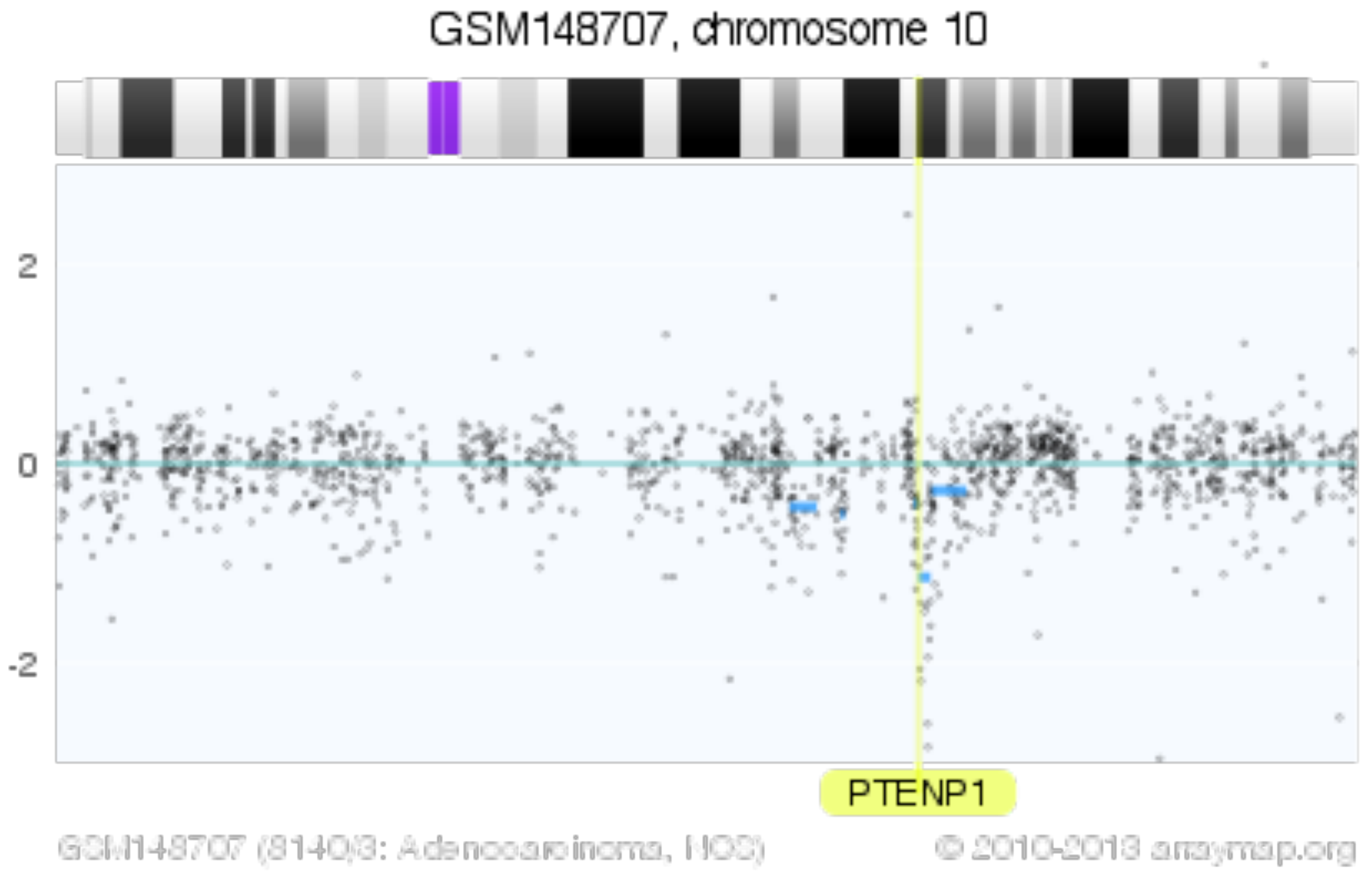
Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)



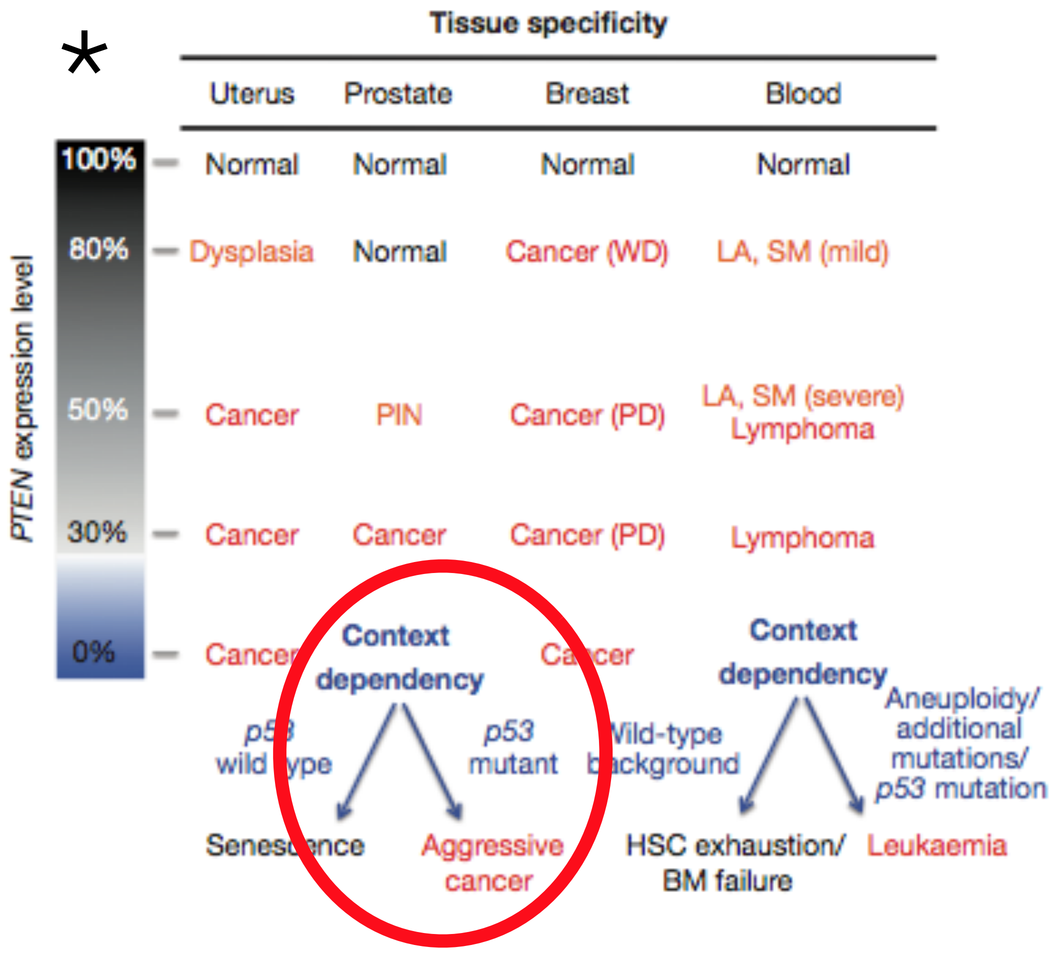
GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

On average ~15% of a cancer genome are in an imbalanced state (more/less than 2 alleles);
Original data based on >30’000 cancer genomes from arraymap.org

Gene dosage phenomena beyond simple on/off effects

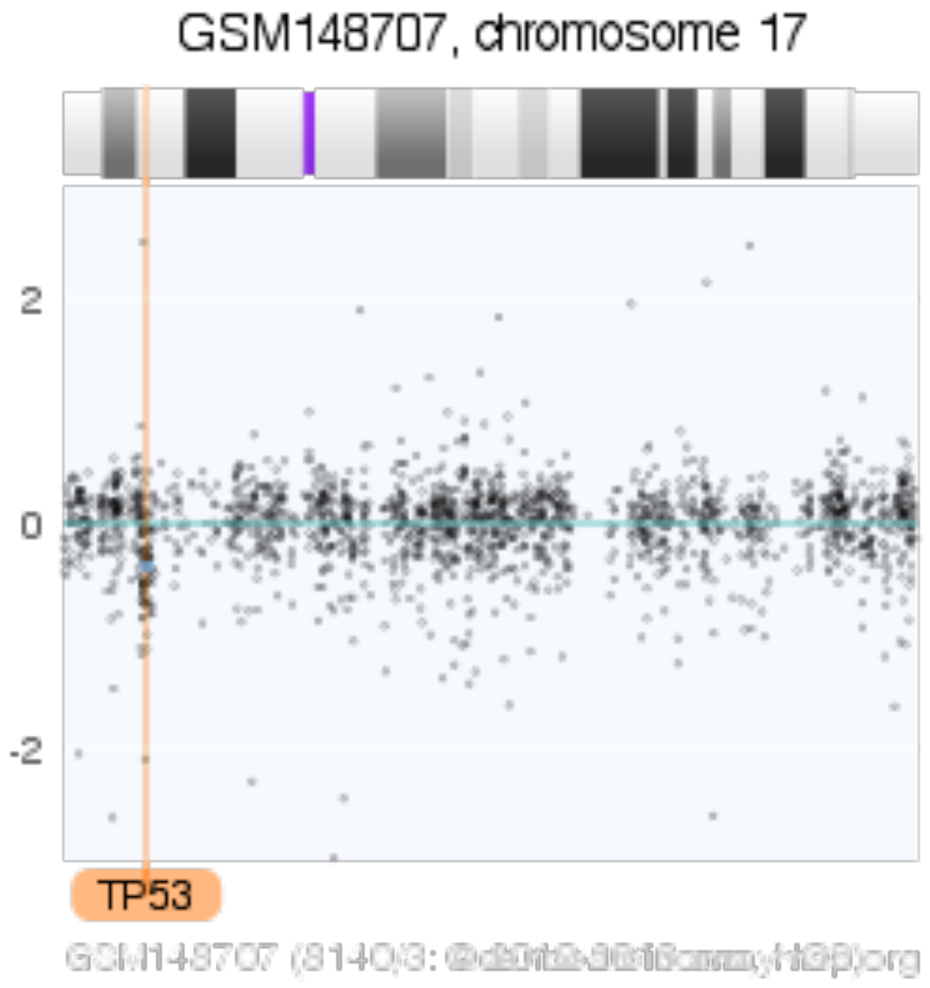


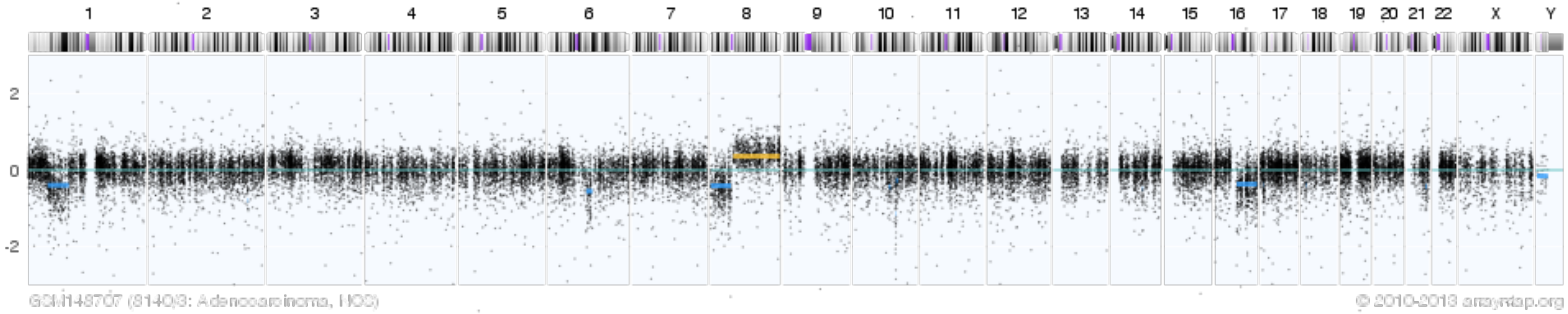
*



Tissue specificity

PTEN expression level	Uterus	Prostate	Breast	Blood
100%	Normal	Normal	Normal	Normal
80%	Dysplasia	Normal	Cancer (WD)	LA, SM (mild)
50%	Cancer	PIN	Cancer (PD)	LA, SM (severe) Lymphoma
30%	Cancer	Cancer	Cancer (PD)	Lymphoma
0%	Cancer	Context dependency	Cancer	Context dependency





Combined heterozygous deletions involving *PTEN* and TP53 loci in a case of prostate adenocarcinoma
(GSM148707, PMID 17875689, Lapointe *et al.*, CancRes 2007)

* A. H. Berger, A. G. Knudson, and P. P. Pandolfi, "A continuum model for tumour suppression," *Nature*, vol. 476, no. 7359, pp. 163–169, Aug. 2011.

arrayMap

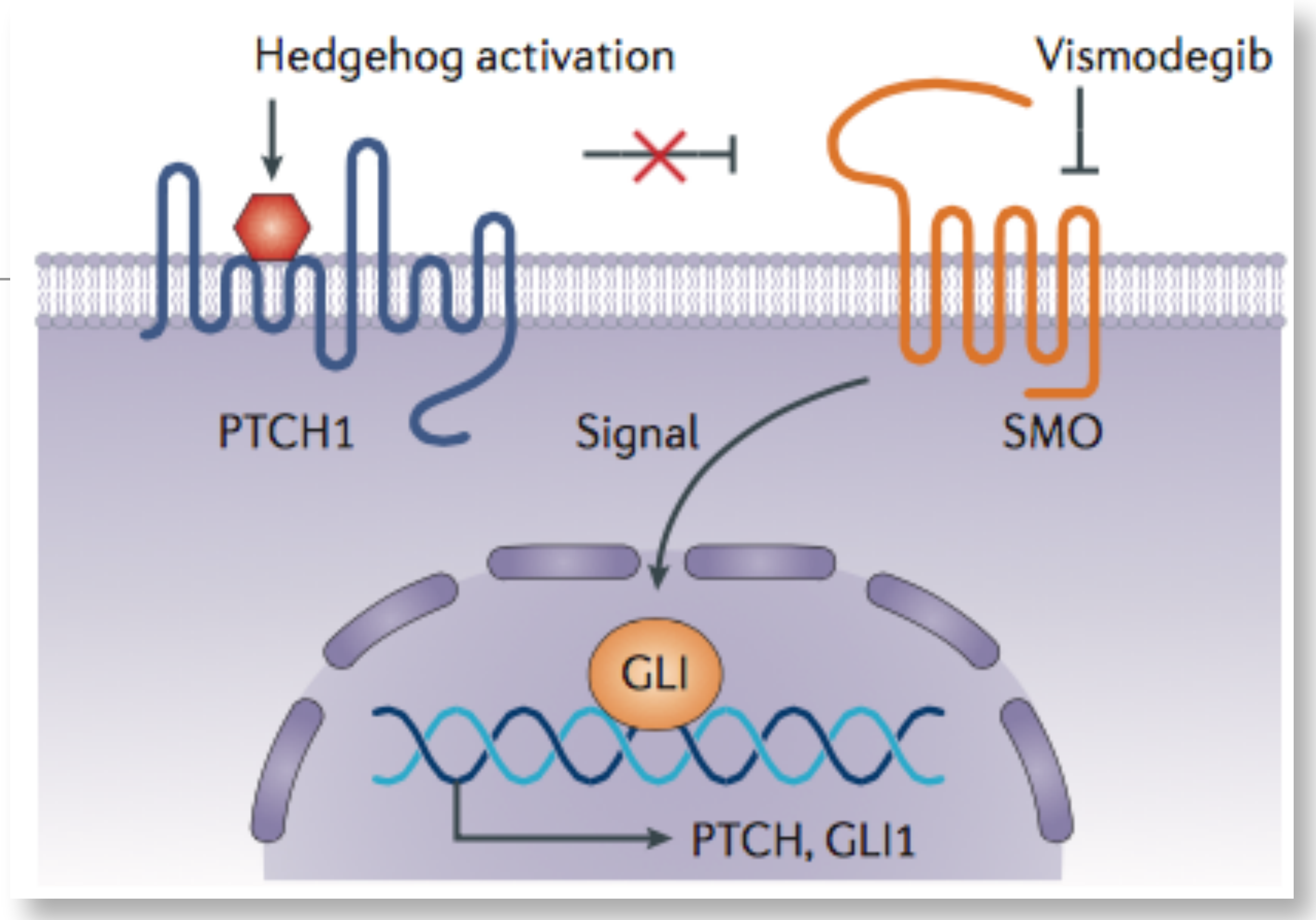
Rare CNV Events & Hidden Therapeutic Options?

Example: PTCH1 deletions in malignant melanomas

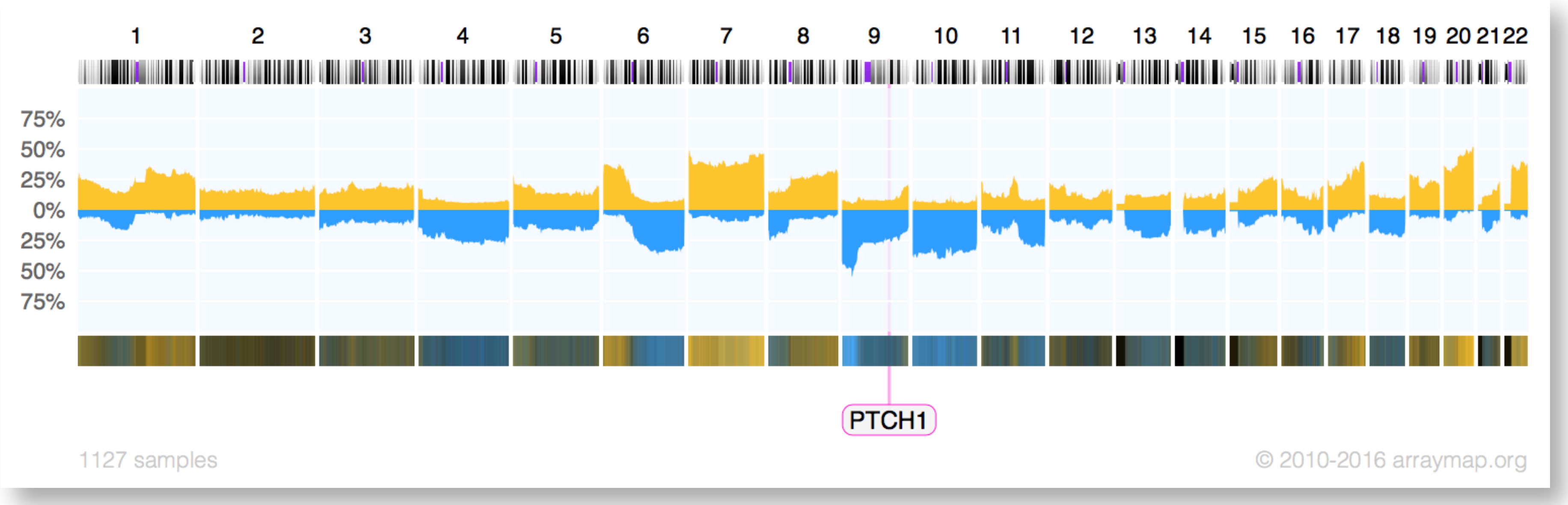
PTCH1 is a actionable tumor suppressor gene, which has been demonstrated in e.g. basaliomas and medulloblastomas

analysis of 1127 samples from 26 different publications could identify **focal** deletions in 4 samples

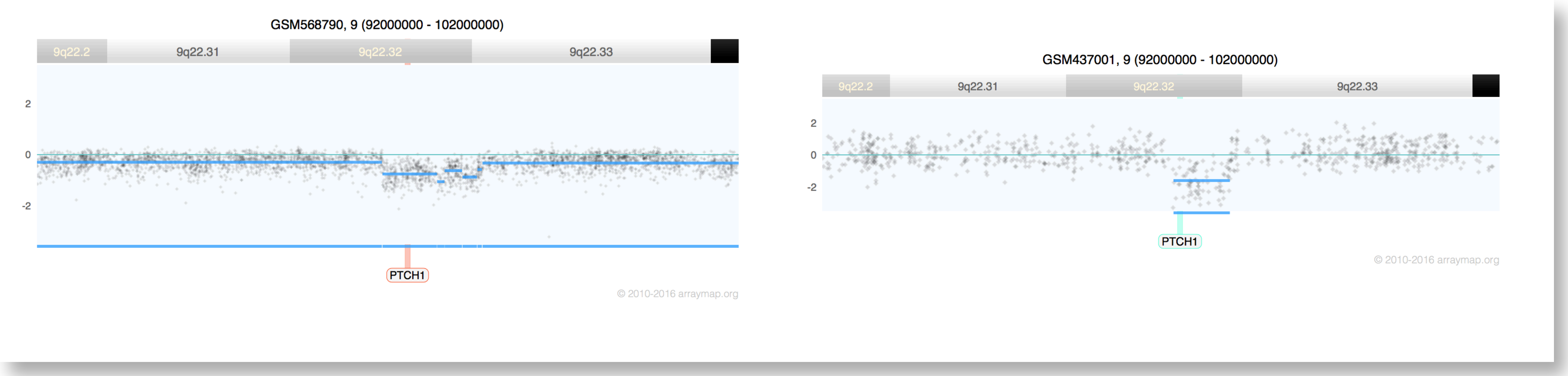
a current project addresses the focal involvement of all mapped genes, in >50'000 cancer genome profiles



In its normal function, PTCH1 is a tumor suppressor gene in the sonic hedgehog pathway and inhibits SMO driven transcriptional activation. A loss of PTCH1 function (mutation, deletion) can be mitigated through drugs antagonistic to SMO activation.



Summary of somatic copy number aberrations from the analysis of 1127 genome profiles of malignant melanomas, collected in our arraymap.org cancer genome resource. While PTCH1 does not represent a deletion hotspot, the genomic locus is part of larger deletions in ~25% of melanoma samples.

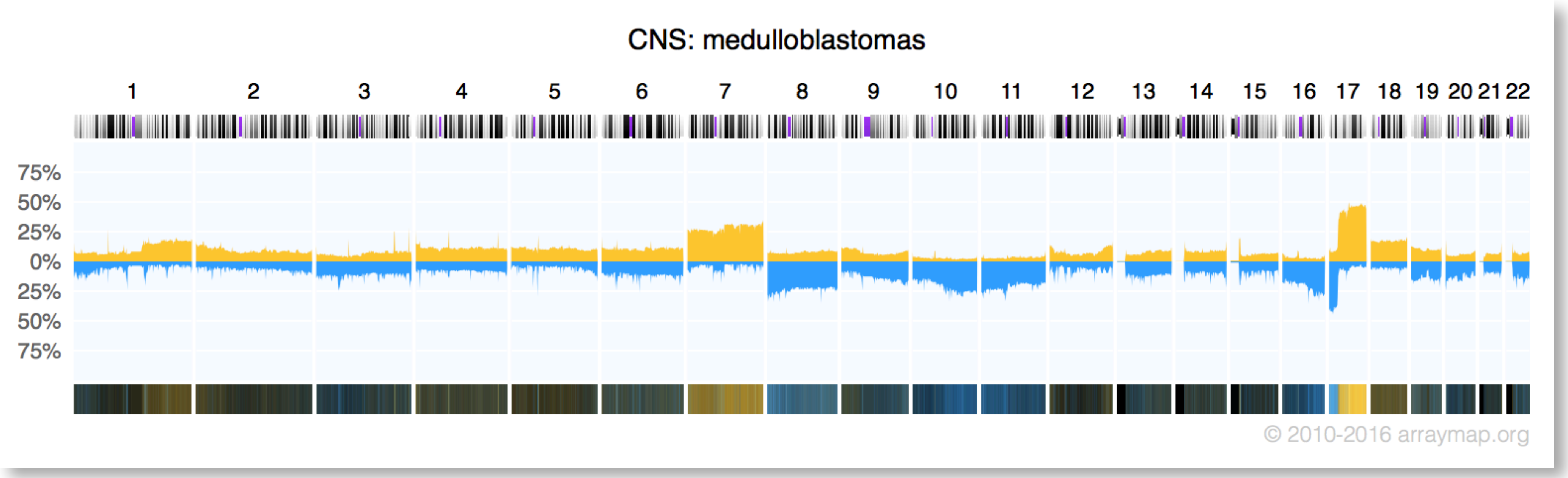


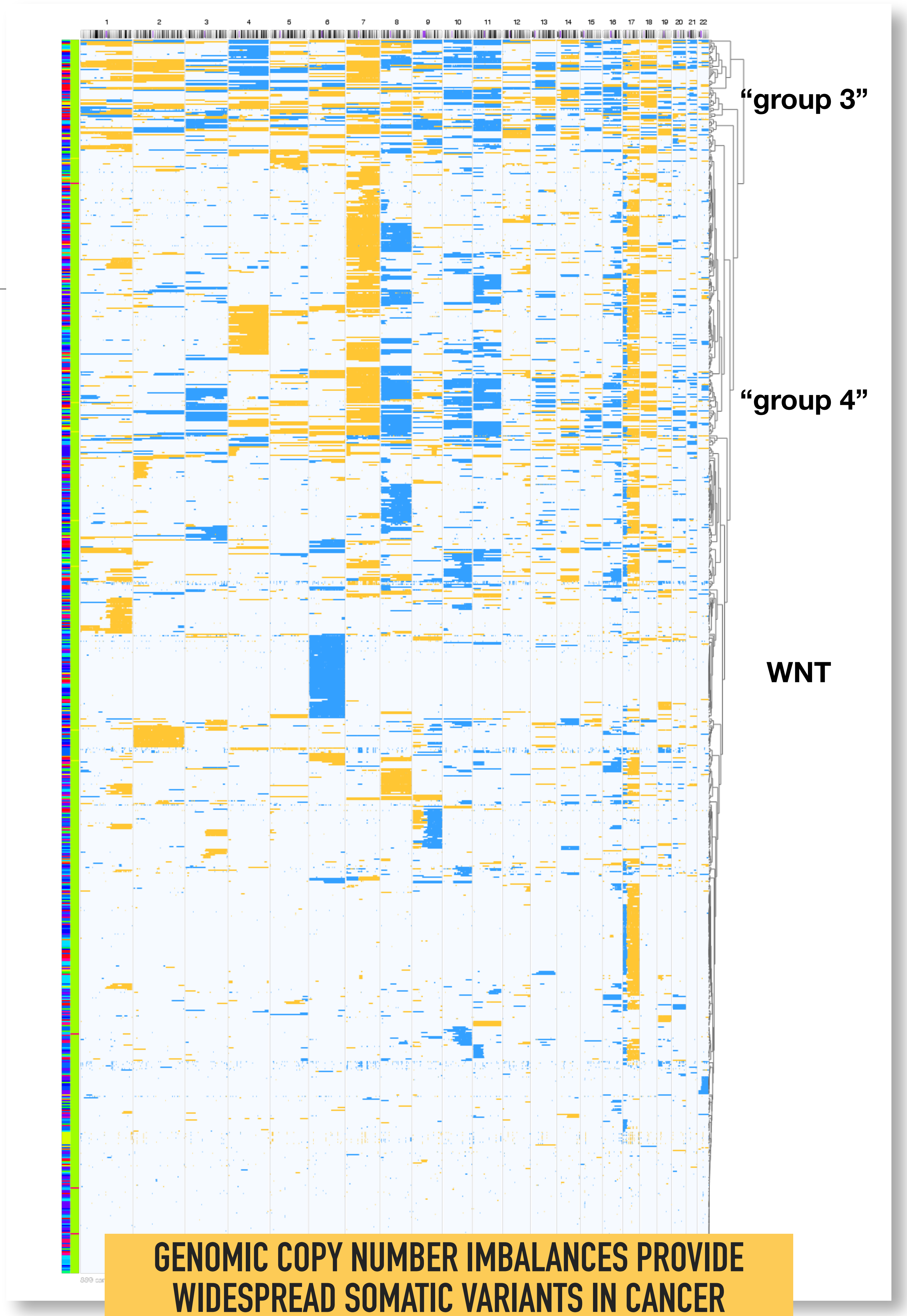
Examples of focal / homozygous PTCH1 deletions detected in the analysis of 1127 genomic array datasets. Focal somatic imbalance events are considered an indicator for oncogenic involvement of the affected target genes.

Somatic CNVs In Cancer: Patterns

Many tumor types express **recurrent mutation patterns**

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





GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

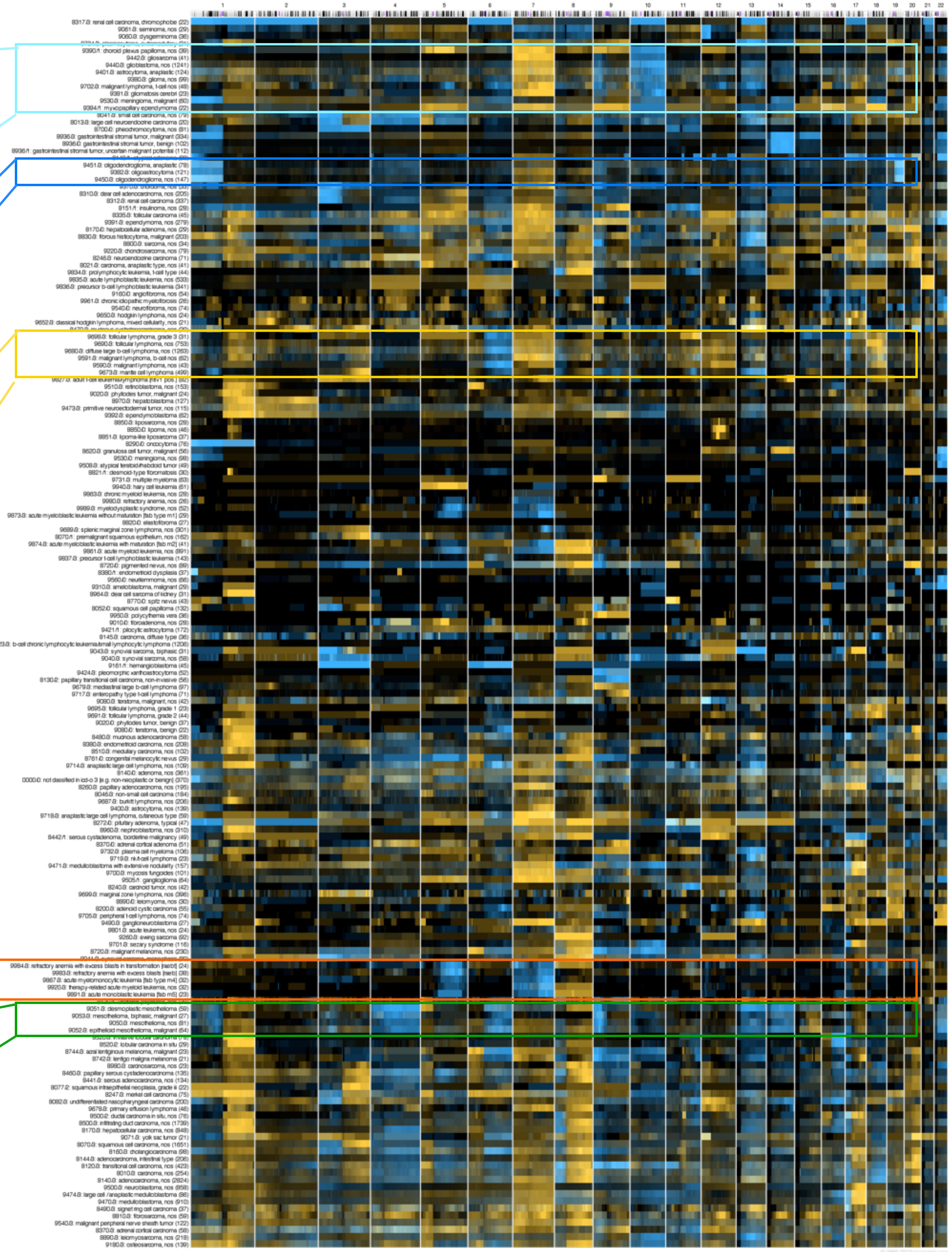
A genomic copy number histogram for malignant medulloblastomas, the most frquent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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

arrayMap

Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma (22)

9451/3: oligodendroglioma, anaplastic (78)
9382/3: oligoastrocytoma (121)
9450/3: oligodendroglioma, nos (147)

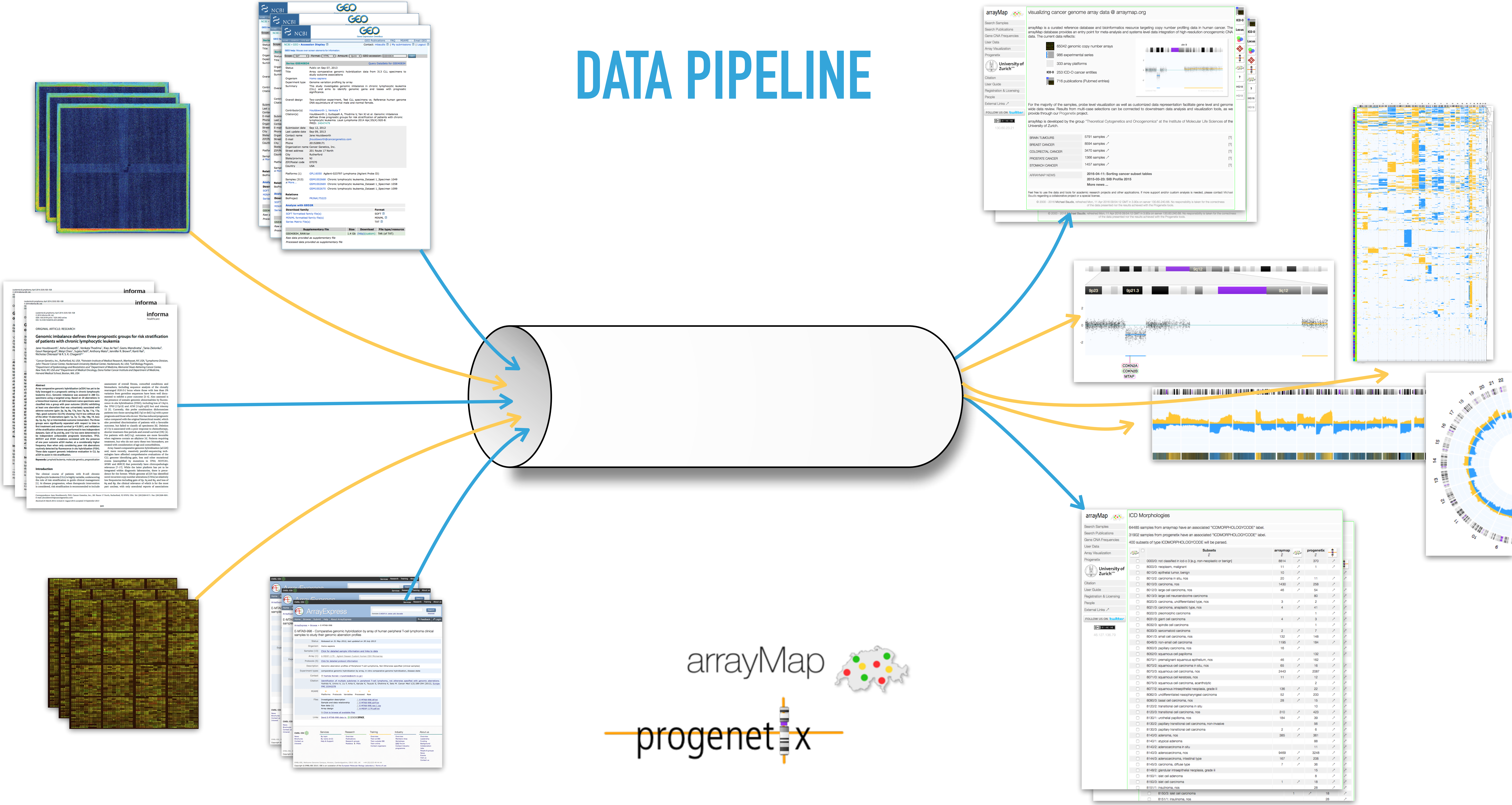
9698/3: follicular lymphoma, grade 3 (31)
9690/3: follicular lymphoma, nos (753)
9680/3: diffuse large b-cell lymphoma, nos (1263)
9591/3: malignant lymphoma, b-cell nos (62)
9590/3: malignant lymphoma, nos (43)
9673/3: mantle cell lymphoma (499)

9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
9983/3: refractory anemia with excess blasts [raeb] (38)
9867/3: acute myelomonocytic leukemia [fab type m4] (32)
9920/3: therapy-related acute myeloid leukemia, nos (32)
9891/3: acute monoblastic leukemia [fab m5] (23)

9051/3: desmoplastic mesothelioma (59)
9053/3: mesothelioma, biphasic, malignant (27)
9050/3: mesothelioma, nos (81)
9052/3: epithelioid mesothelioma, malignant (64)

arrayMap

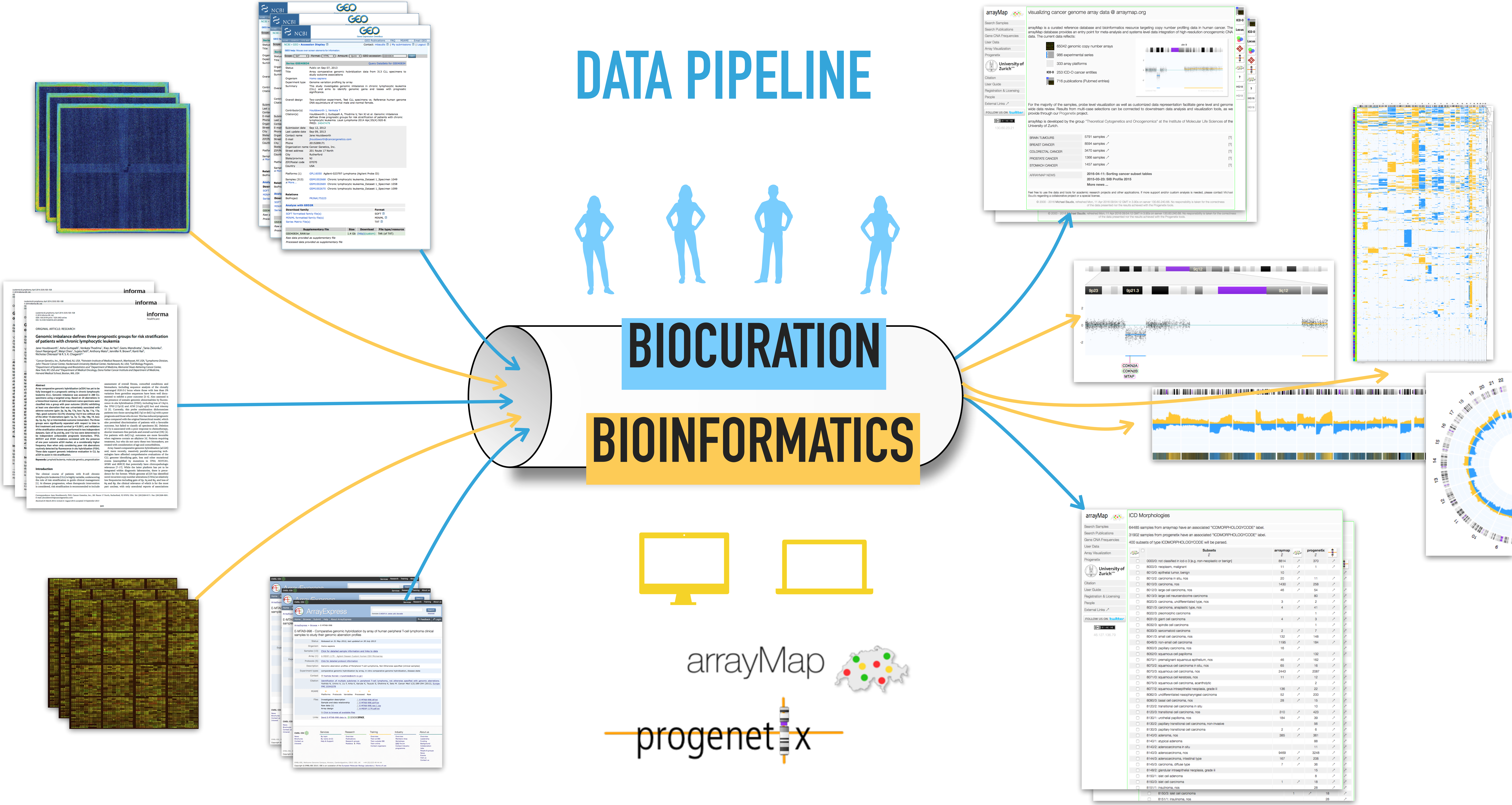
DATA PIPELINE


NCBI
GEO
informa
Genomic imbalance defines three prognostic groups for risk stratification of patients with chronic lymphocytic leukemia
ArrayExpress
arrayMap
visualizing cancer genome array data @ arraymap.org
9p23
9p21.3
9q12
CDKN2A
CDKN2B
MTAP
ICD Morphologies

arrayMap

progenetix

DATA PIPELINE



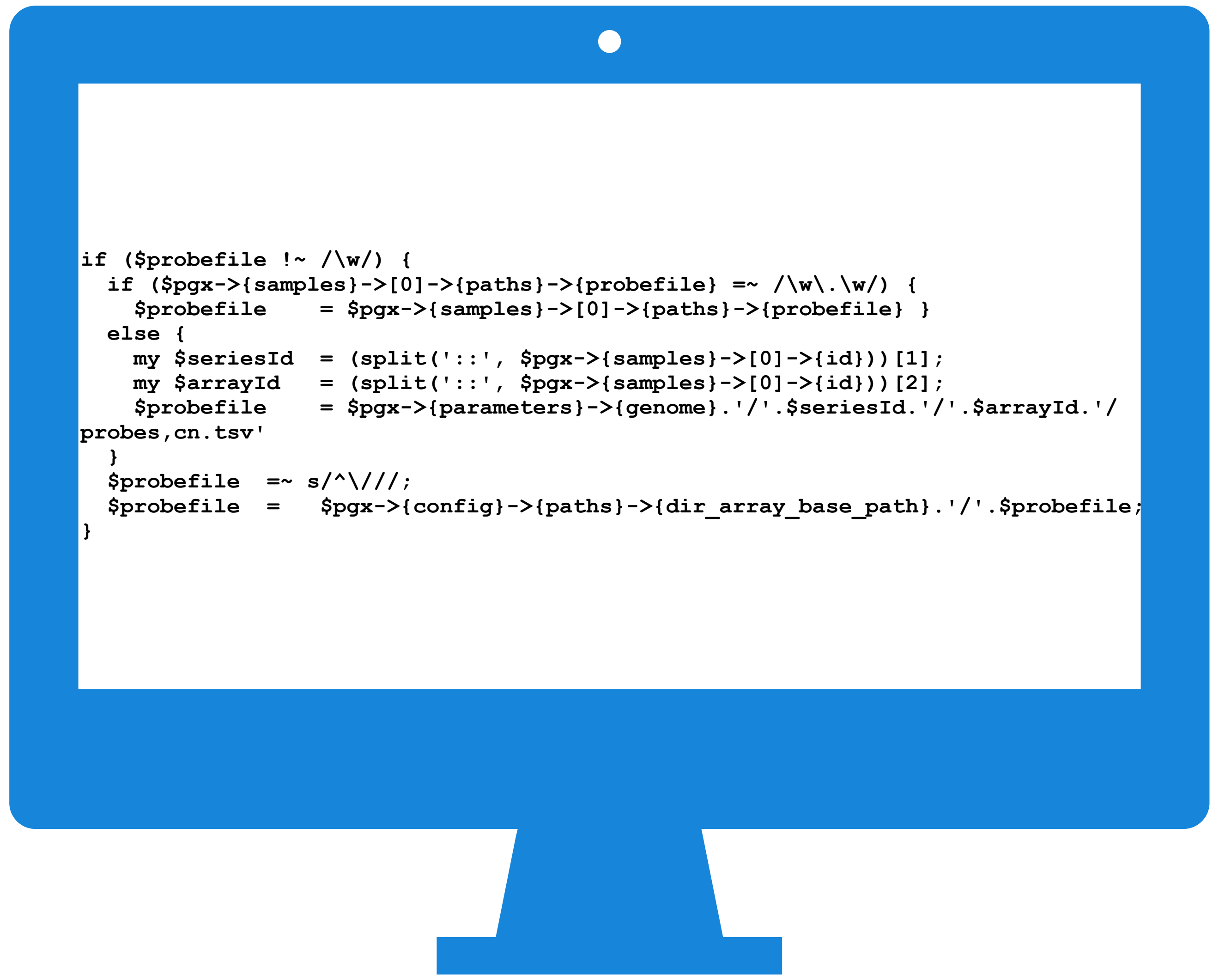
BIOCURATION

BIOINFORMATICS

arrayMap

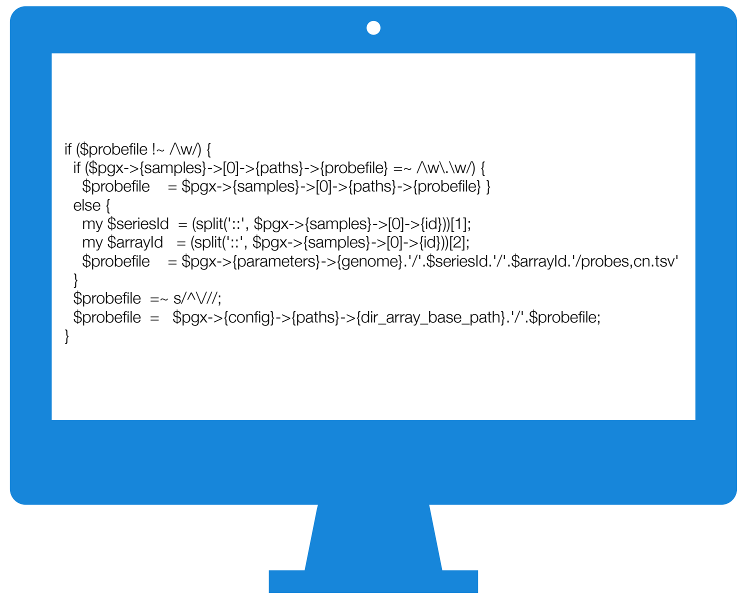
progenetix

{bio_informatics_science}

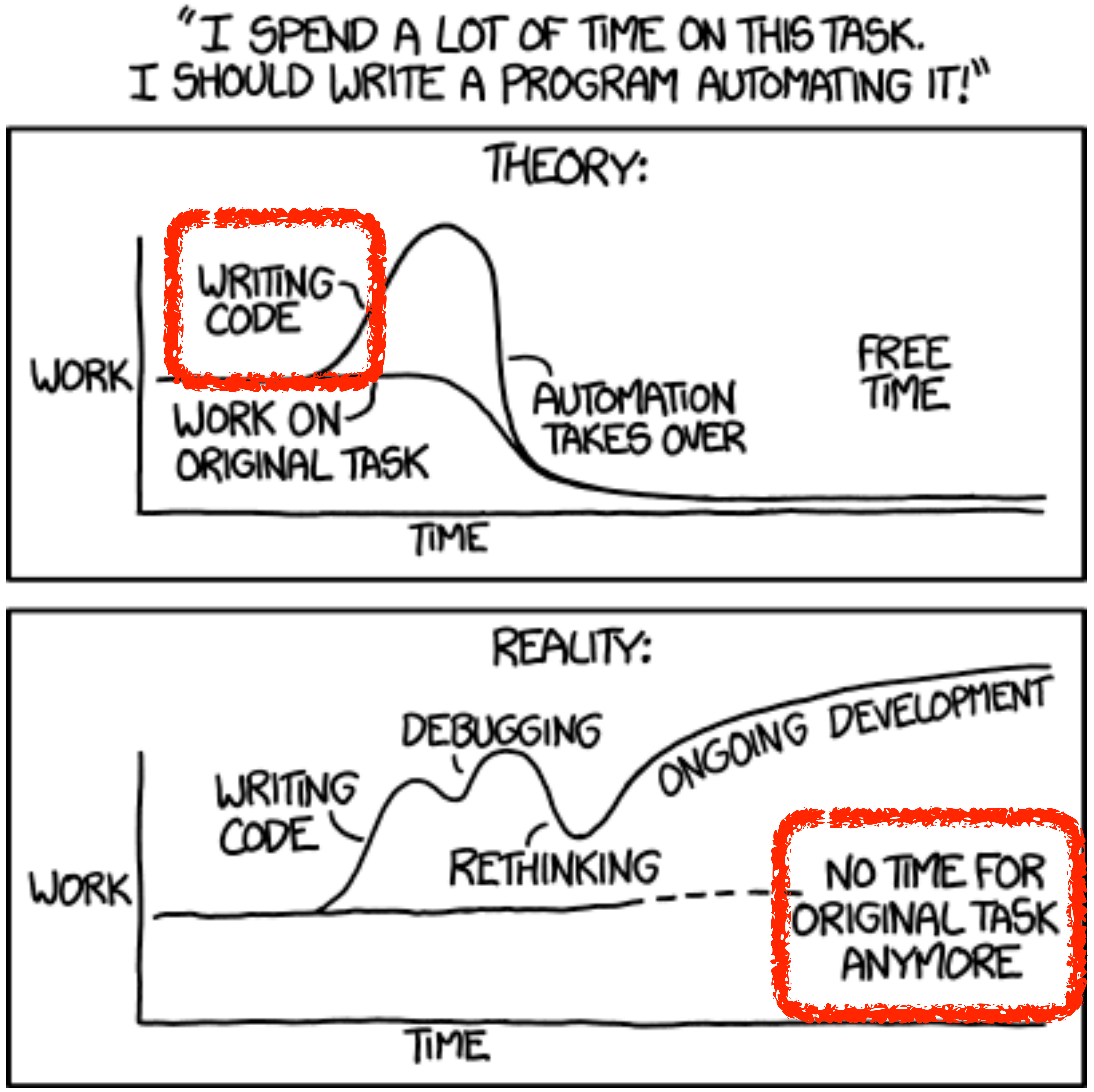


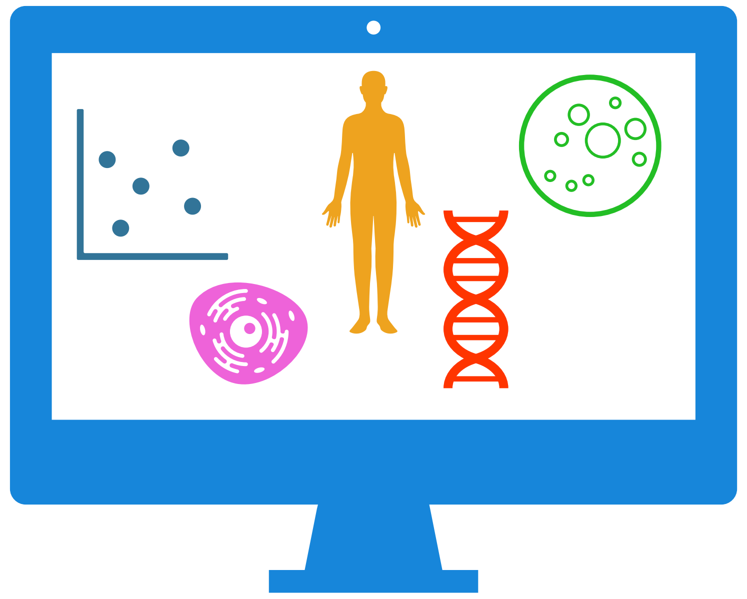
```
if ($probefile !~ /\w/) {
  if ($pgx->{samples}->[0]->{paths}->{probefile} =~ /\w\.\w/) {
    $probefile    = $pgx->{samples}->[0]->{paths}->{probefile} }
  else {
    my $seriesId  = (split('::', $pgx->{samples}->[0]->{id}))[1];
    my $arrayId   = (split('::', $pgx->{samples}->[0]->{id}))[2];
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

{bio_informatics_science}



```
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;
}
```





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

A tool for genome (CNV) batch liftover

Situation

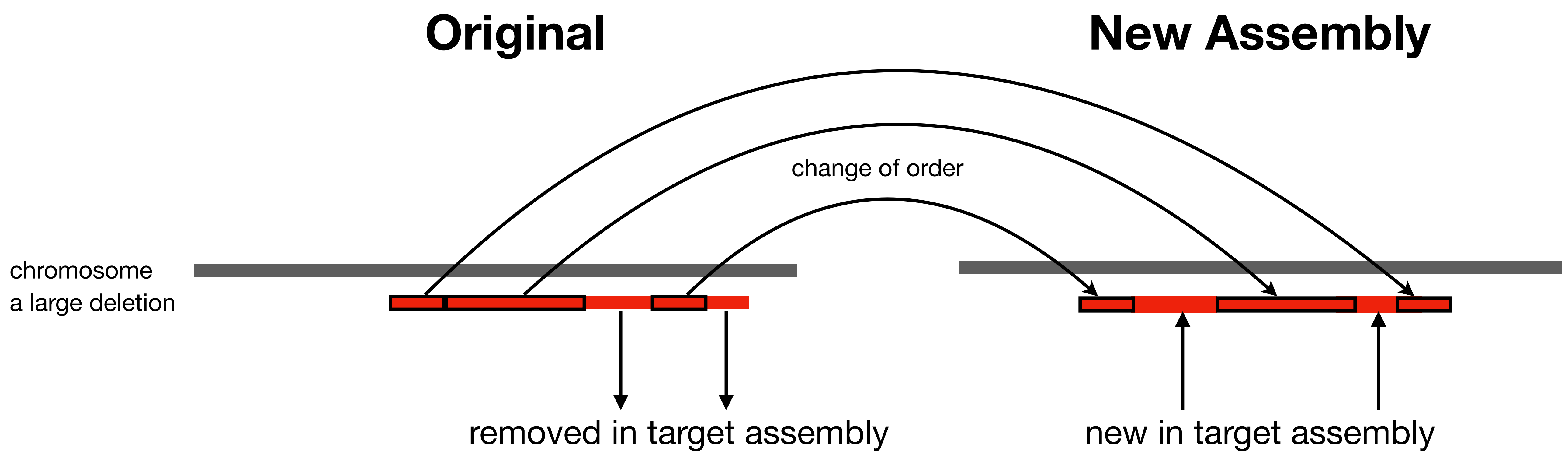
- A continuous genome segment with **real start and end positions** has been duplicated/deleted
- the reported **edge positions are statistically derived** and their real equivalent may be removed/repositioned
- however, CNV segments are determined from **many measurements** - reporting edges is just a convenience

Challenge

1. Keep the "integrity" of copy number segments after liftover
2. improve on the 10% CNV data lost from straight liftover applications
3. process 1TB segment and probe data buried in over 2,000 nested directories

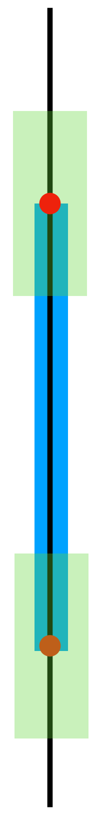
Solution

1. Algorithm to lift segments.
2. Algorithm for fuzzy remapping.
3. Parallel processing and failure recovery mechanism

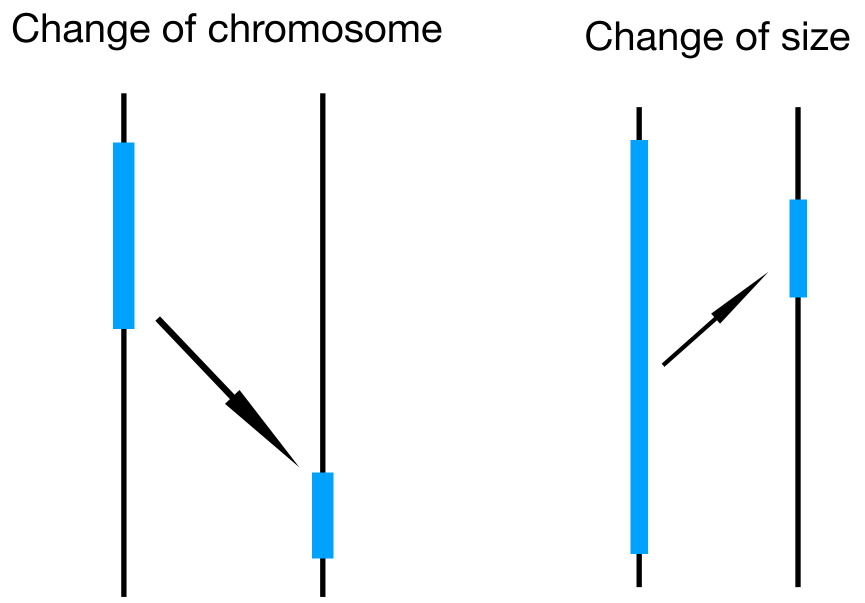


The difficulties in copy number segment liftover

Fuzzy search



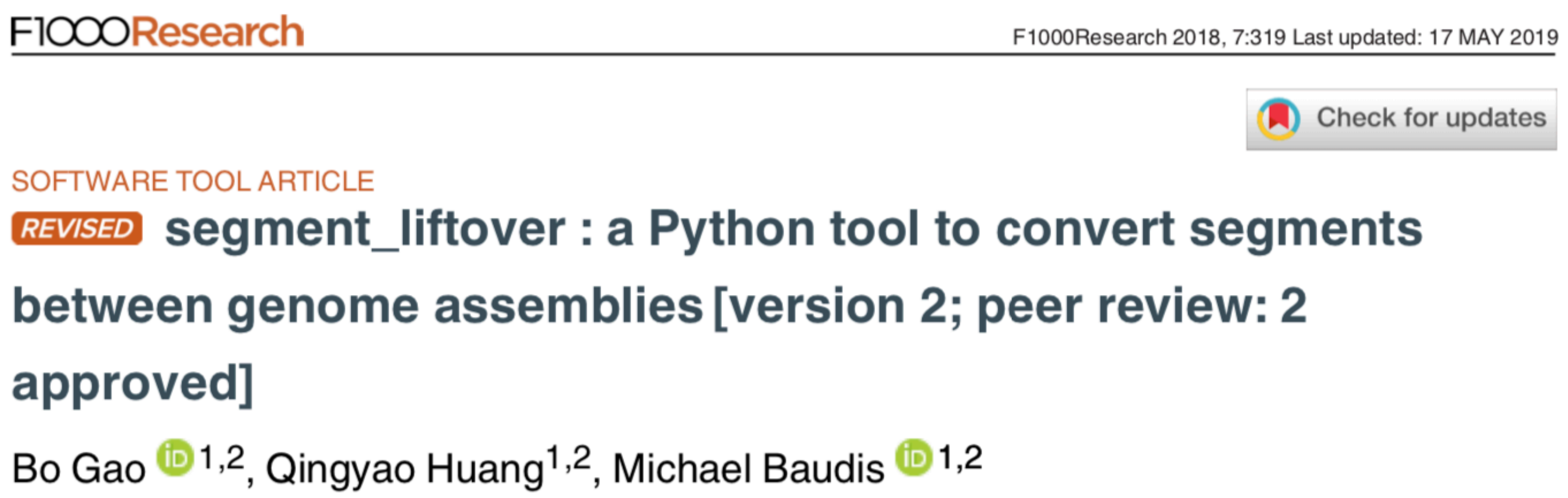
Quality control



Results of ***segment_liftover***

- Convert hg18 | hg19 | GRCh38
- Processed 122,788 files, 26,164,205 segments and 28,941,899,671 probes in total
- A straight forward run took more than a week
- parallel run of 4 processes took less than 3 days
- Reduced data loss: **10% => 0.1%**

https://github.com/baudisgroup/segment-liftover


F1000Research

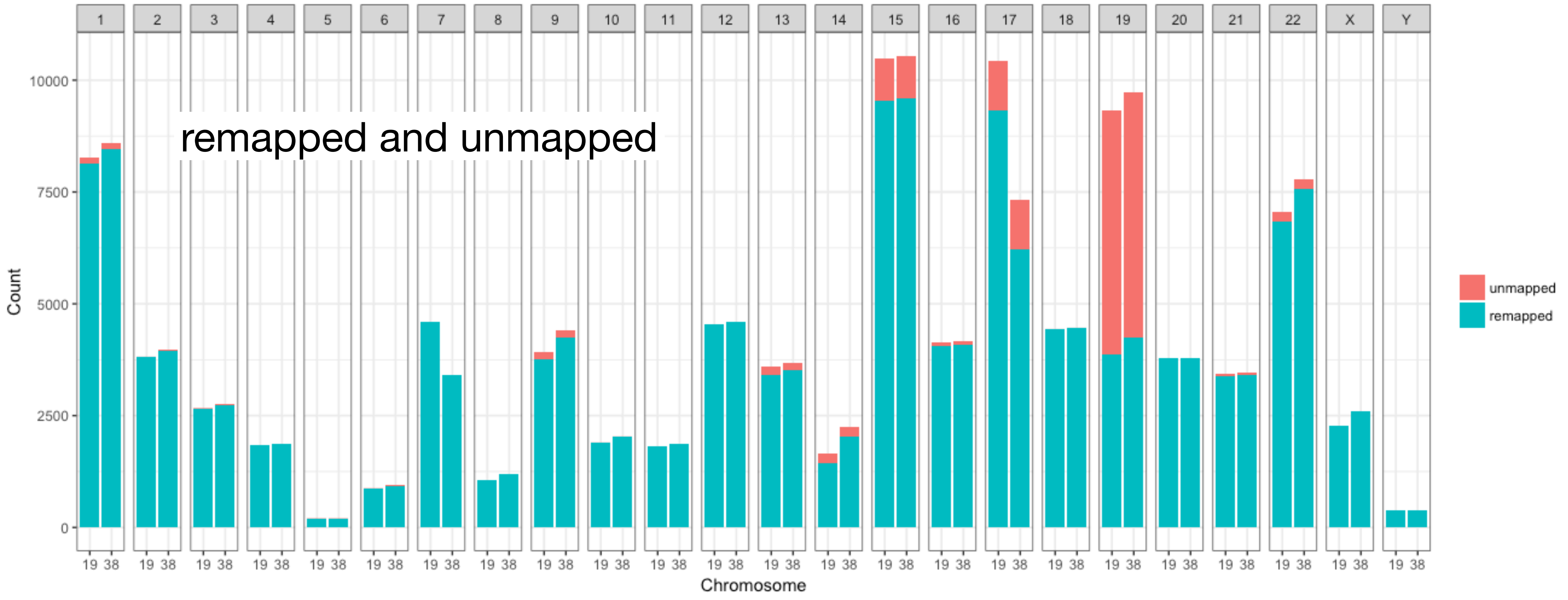
F1000Research 2018, 7:319 Last updated: 17 MAY 2019

Check for updates

SOFTWARE TOOL ARTICLE

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

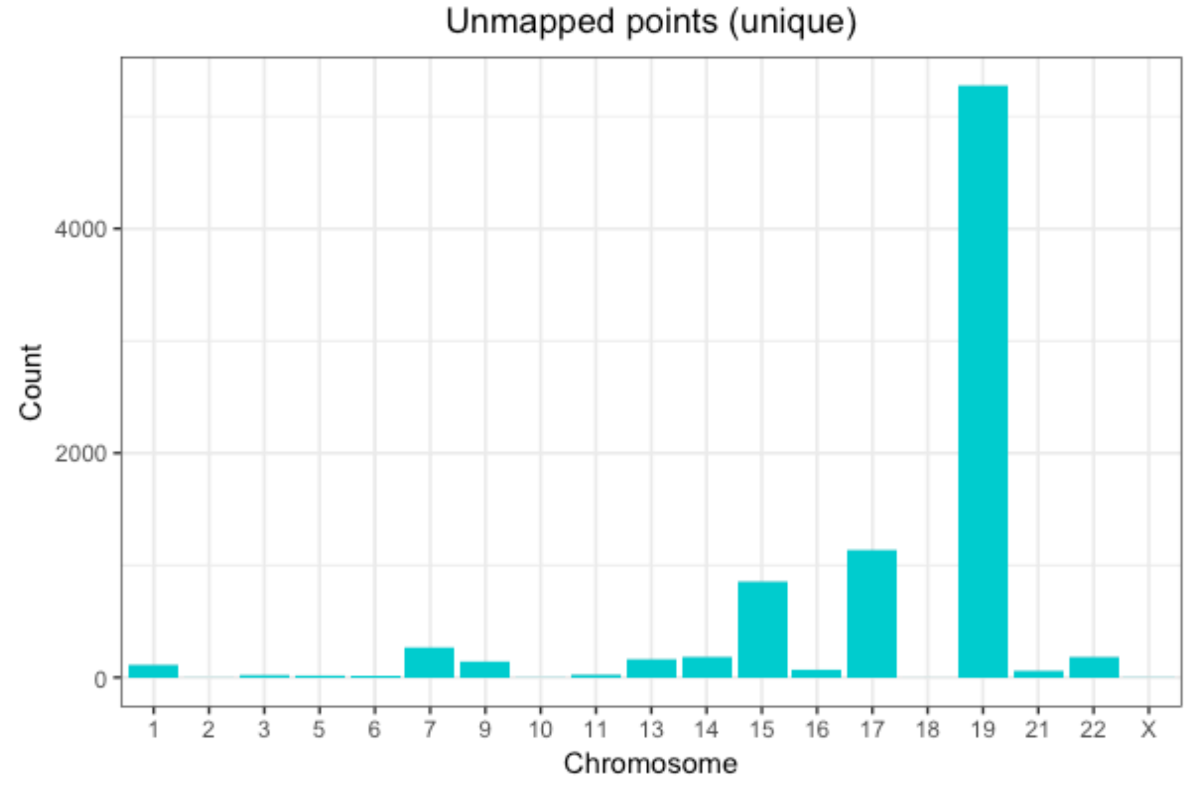
Bo Gao[1,2], Qingyao Huang[1,2], Michael Baudis[1,2]

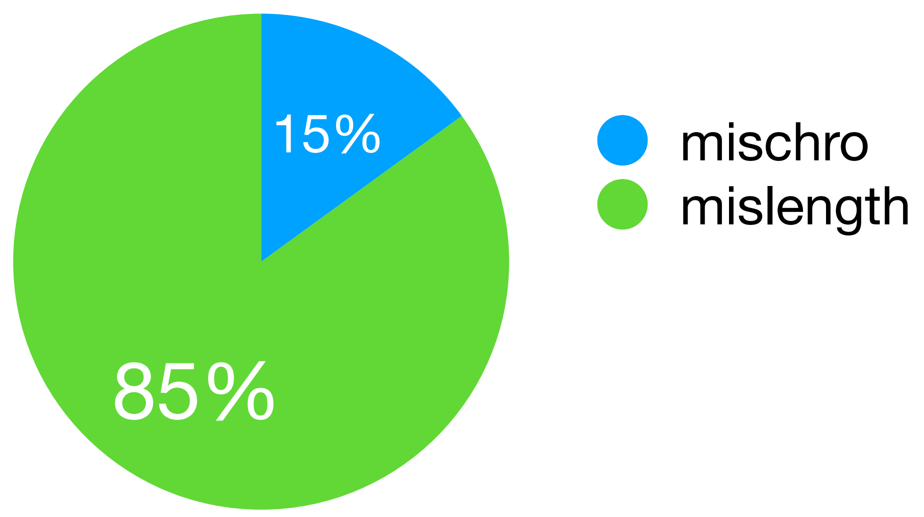


remapped and unmapped

Distribution of unmapped probes

Reason of unmapped segments





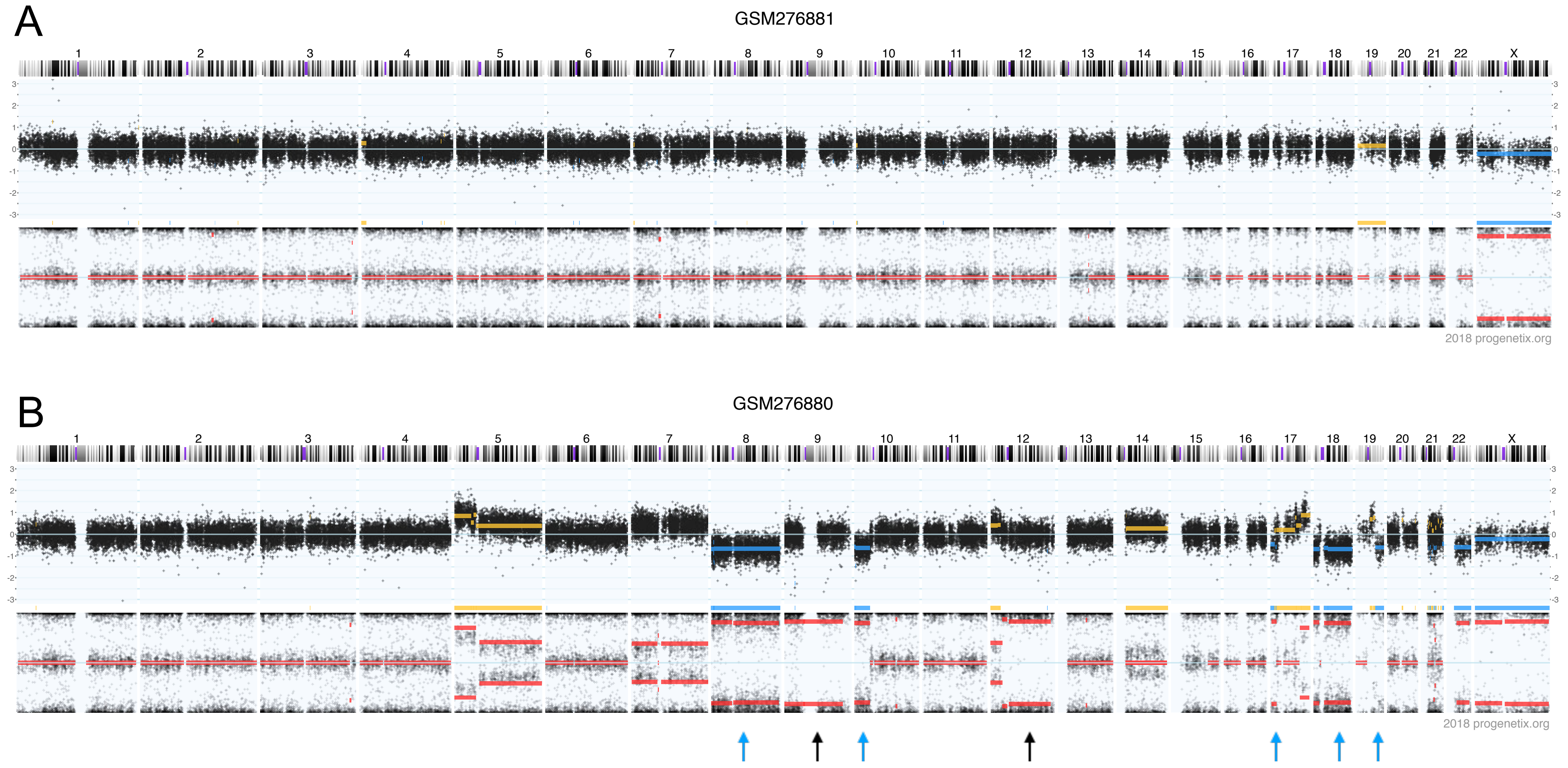
Population stratification in cancer samples based on SNP array data

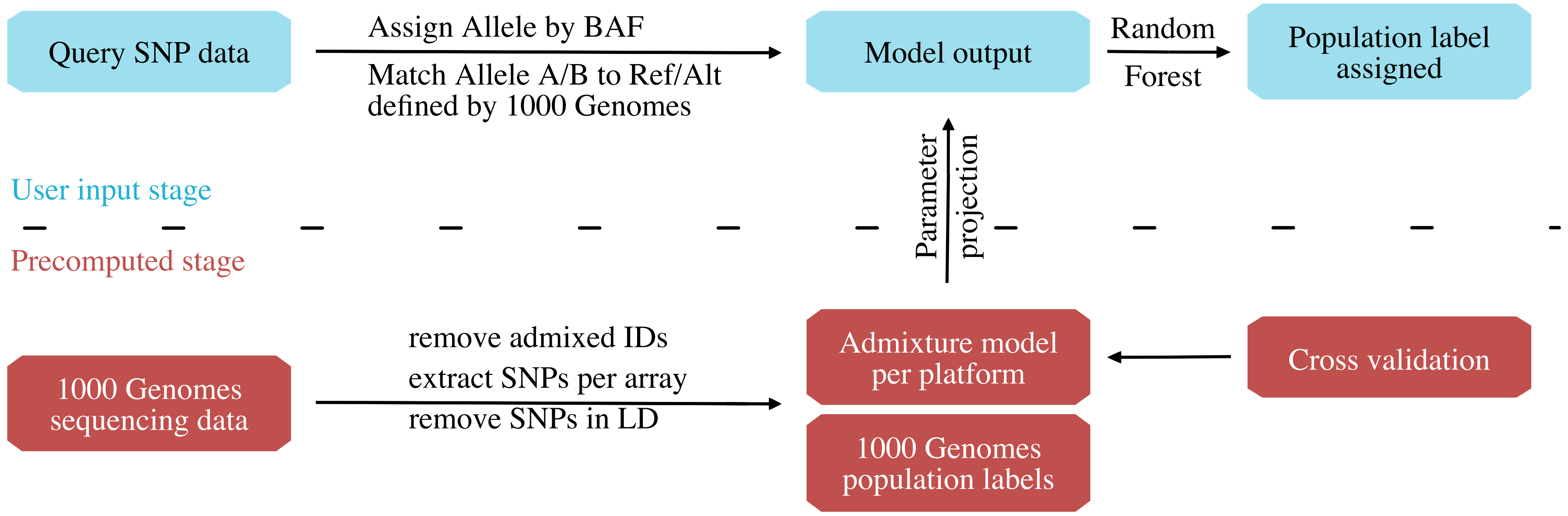
- 2504 genome profiles from 1000 Genome project phase 1 as reference
- 5 (or 26) superpopulations: South Asia, Europe, South America, East Asia and Africa.
- SNP positions used in 9 Affymetrix SNP arrays are extracted to train a population admixture model.

arrayMap

Enabling population assignment from cancer genomes with SNP2pop

Qingyao Huang[1,2] and Michael Baudis[1,2]✉




Query SNP data

Assign Allele by BAF

Match Allele A/B to Ref/Alt defined by 1000 Genomes

Model output

Random Forest

Population label assigned

User input stage

Parameter projection

Precomputed stage

1000 Genomes sequencing data

remove admixed IDs
extract SNPs per array
remove SNPs in LD

Admixture model per platform

1000 Genomes population labels

Cross validation

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

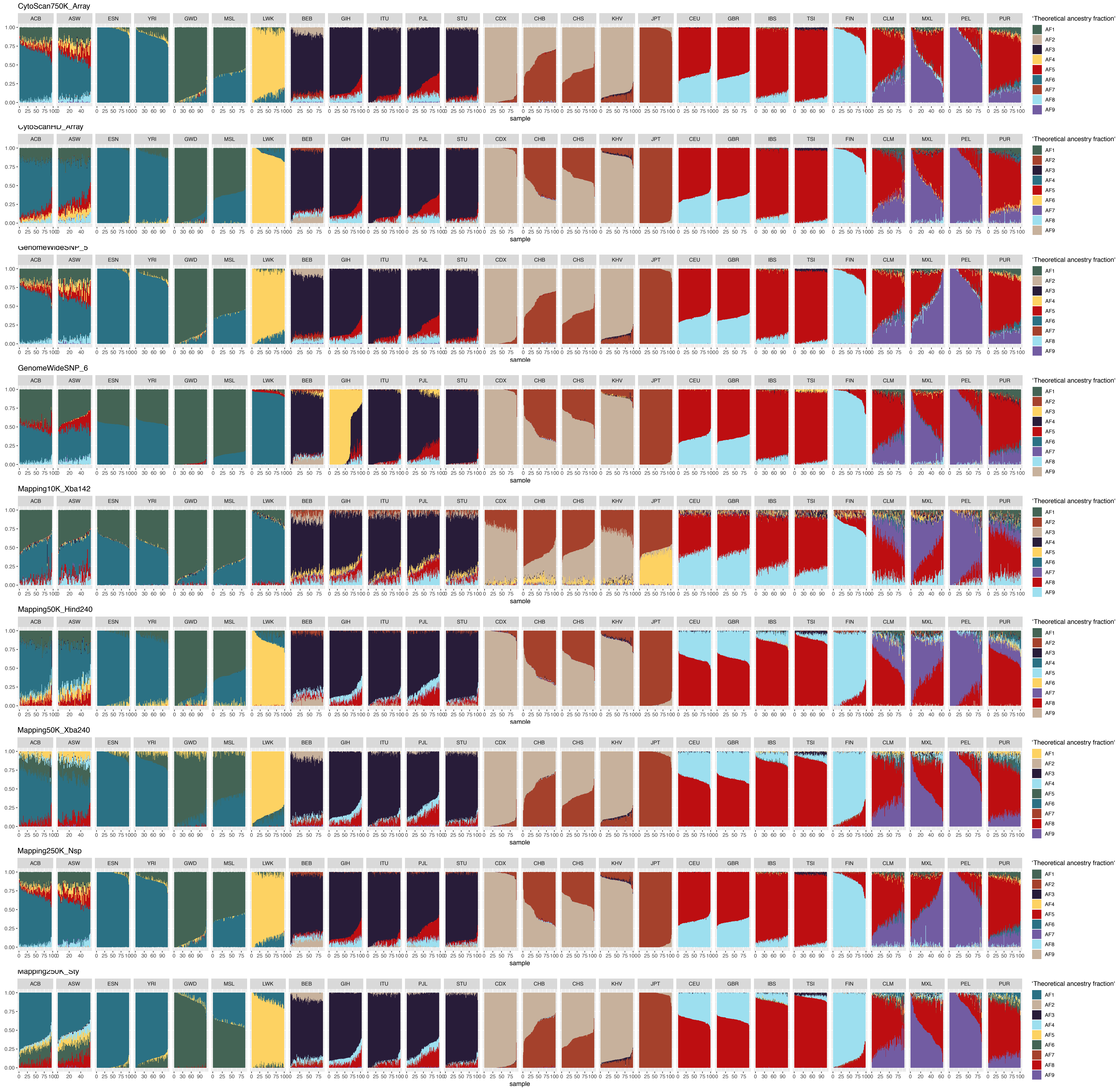
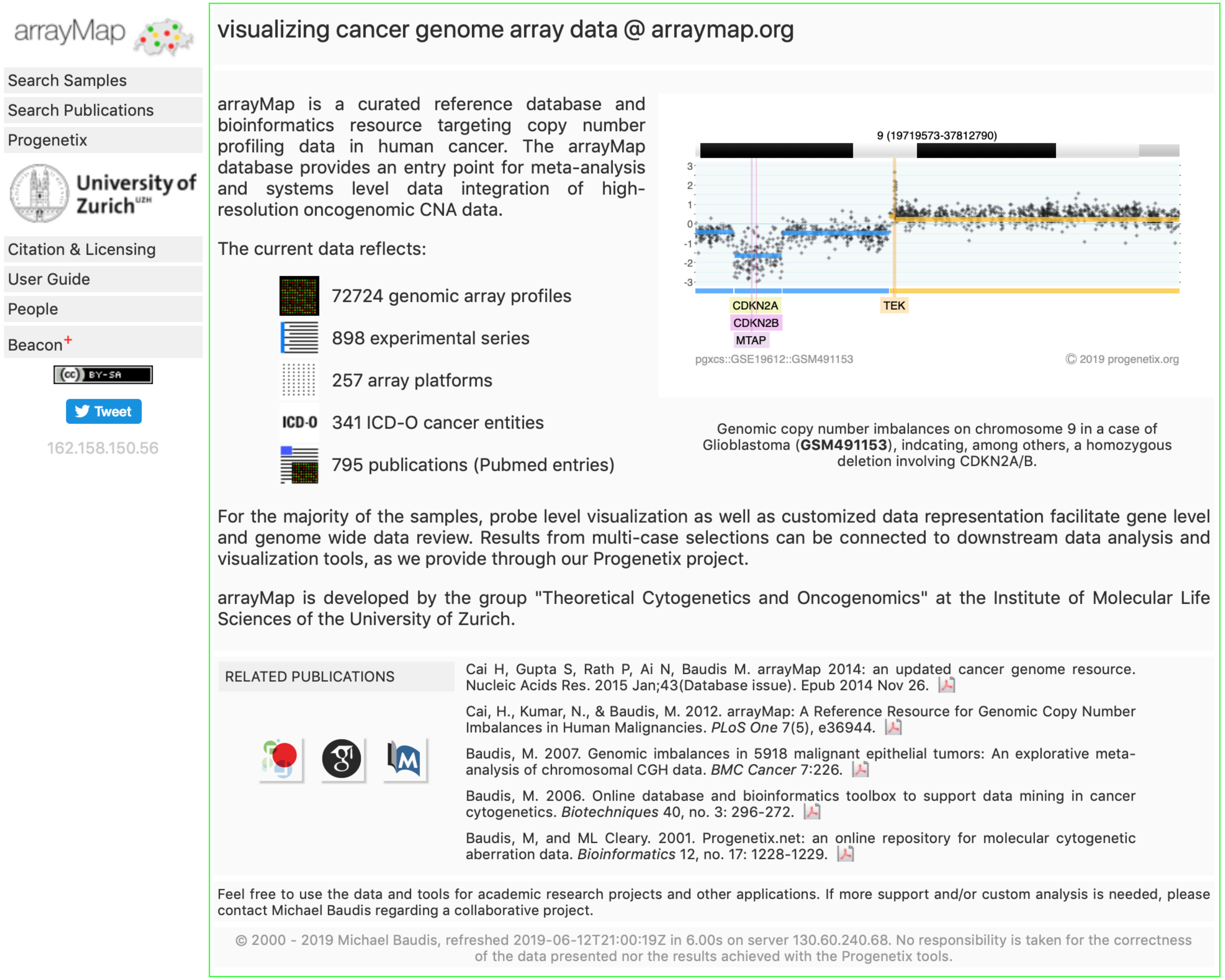


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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

arrayMap

arrayMap

Reference resource for copy number variation data in cancer



arrayMap

- Search Samples
- Search Publications
- Progenetix
- University of Zurich
- Citation & Licensing
- User Guide
- People
- Beacon+

BY-SA

Tweet

162.158.150.56

visualizing cancer genome array data @ arraymap.org

arrayMap is a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides an entry point for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data.

The current data reflects:

- 72724 genomic array profiles
- 898 experimental series
- 257 array platforms
- ICD-O 341 ICD-O cancer entities
- 795 publications (Pubmed entries)

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

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

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

RELATED PUBLICATIONS

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

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

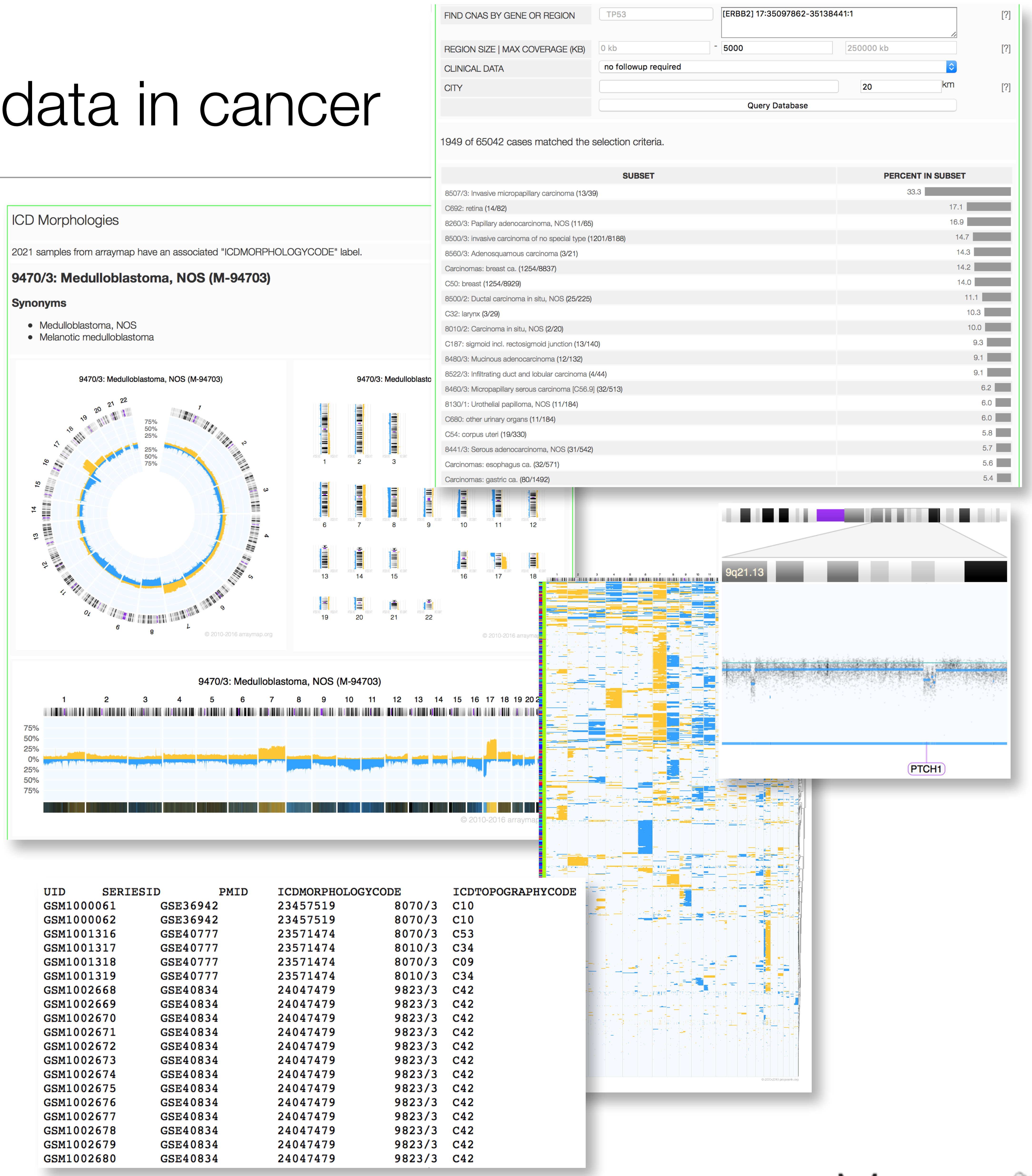
Baudis, M. 2007. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer* 7:226.

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

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

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

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



FIND CNAS BY GENE OR REGION	TP53	[ERBB2] 17:35097862-35138441:1	[?]
REGION SIZE | MAX COVERAGE (KB)	0 kb - 5000	250000 kb	[?]
CLINICAL DATA	no followup required		
CITY		20 km	[?]
	Query Database		

1949 of 65042 cases matched the selection criteria.

SUBSET	PERCENT IN SUBSET
8507/3: Invasive micropapillary carcinoma (13/39)	33.3
C692: retina (14/82)	17.1
8260/3: Papillary adenocarcinoma, NOS (11/65)	16.9
8500/3: invasive carcinoma of no special type (1201/8188)	14.7
8560/3: Adenosquamous carcinoma (3/21)	14.3
Carcinomas: breast ca. (1254/8837)	14.2
C50: breast (1254/8929)	14.0
8500/2: Ductal carcinoma in situ, NOS (25/225)	11.1
C32: larynx (3/29)	10.3
8010/2: Carcinoma in situ, NOS (2/20)	10.0
C187: sigmoid incl. rectosigmoid junction (13/140)	9.3
8480/3: Mucinous adenocarcinoma (12/132)	9.1
8522/3: Infiltrating duct and lobular carcinoma (4/44)	9.1
8460/3: Micropapillary serous carcinoma [C56.9] (32/513)	6.2
8130/1: Urothelial papilloma, NOS (11/184)	6.0
C680: other urinary organs (11/184)	6.0
C54: corpus uteri (19/330)	5.8
8441/3: Serous adenocarcinoma, NOS (31/542)	5.7
Carcinomas: esophagus ca. (32/571)	5.6
Carcinomas: gastric ca. (80/1492)	5.4

ICD Morphologies

2021 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.

9470/3: Medulloblastoma, NOS (M-94703)

Synonyms

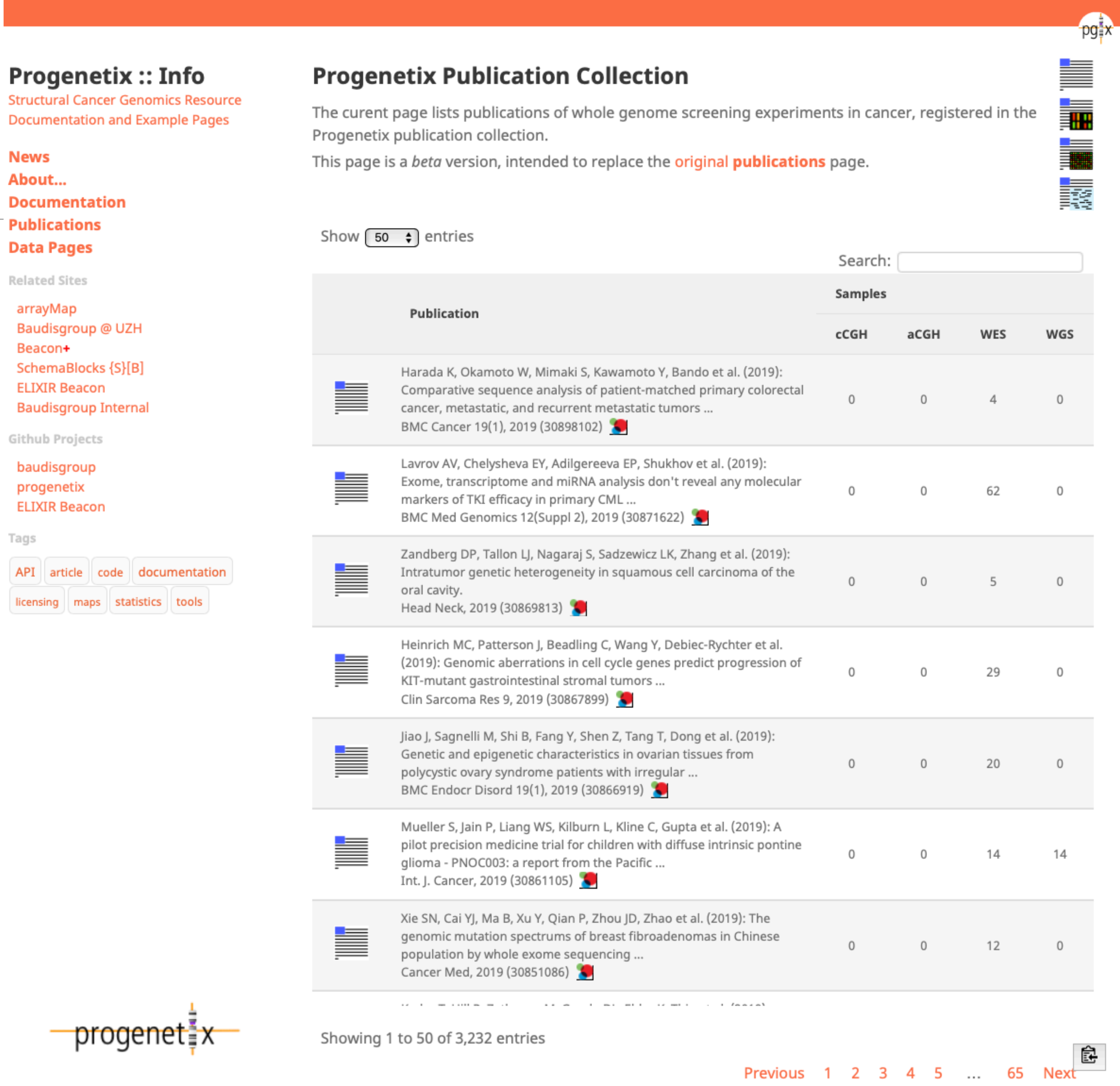
- Medulloblastoma, NOS
- Melanotic medulloblastoma

```
UID         SERIESID         PMID        ICDMORPHOLOGYCODE       ICDTOPOGRAPHYCODE
GSM1000061       GSE36942         23457519          8070/3  C10
GSM1000062       GSE36942         23457519          8070/3  C10
GSM1001316       GSE40777         23571474          8070/3  C53
GSM1001317       GSE40777         23571474          8010/3  C34
GSM1001318       GSE40777         23571474          8070/3  C09
GSM1001319       GSE40777         23571474          8010/3  C34
GSM1002668       GSE40834         24047479          9823/3  C42
GSM1002669       GSE40834         24047479          9823/3  C42
GSM1002670       GSE40834         24047479          9823/3  C42
GSM1002671       GSE40834         24047479          9823/3  C42
GSM1002672       GSE40834         24047479          9823/3  C42
GSM1002673       GSE40834         24047479          9823/3  C42
GSM1002674       GSE40834         24047479          9823/3  C42
GSM1002675       GSE40834         24047479          9823/3  C42
GSM1002676       GSE40834         24047479          9823/3  C42
GSM1002677       GSE40834         24047479          9823/3  C42
GSM1002678       GSE40834         24047479          9823/3  C42
GSM1002679       GSE40834         24047479          9823/3  C42
GSM1002680       GSE40834         24047479          9823/3  C42
```

arrayMap

Progenetix - Cancer CNV Information Resource

- launched online in 2001 as *progenetix.net*
- **curation** of published CNV profiling data
  - originally cCGH and CNV extraction from Mitelman database
  - + aCGH, WES, WGS; - karyotype data
- increasingly focused on representing the "publication landscape" of cancer genome screening - What? Where?
- Genomes:
  - 93640 CNV profiles (cCGH, aCGH, WES, WGS) from 469 cancer types (NCIt & ICD-O mapping)
  - 6'817'645 "CNVs" (i.e. called segments)
- Articles:
  - 3229 registered articles
    - geographic mapping
    - "cancer type" labelling
  - represent 174'530 reported samples



Progenetix :: Info

Structural Cancer Genomics Resource
Documentation and Example Pages

News
About...
Documentation
Publications
Data Pages

Related Sites

arrayMap
Baudisgroup @ UZH
Beacon+
SchemaBlocks {S}[B]
ELIXIR Beacon
Baudisgroup Internal

Github Projects

baudisgroup
progenetix
ELIXIR Beacon

Tags

API article code documentation licensing maps statistics tools

Progenetix Publication Collection

The curent page lists publications of whole genome screening experiments in cancer, registered in the Progenetix publication collection.

This page is a *beta* version, intended to replace the original publications page.

Show 50 entries

Search:

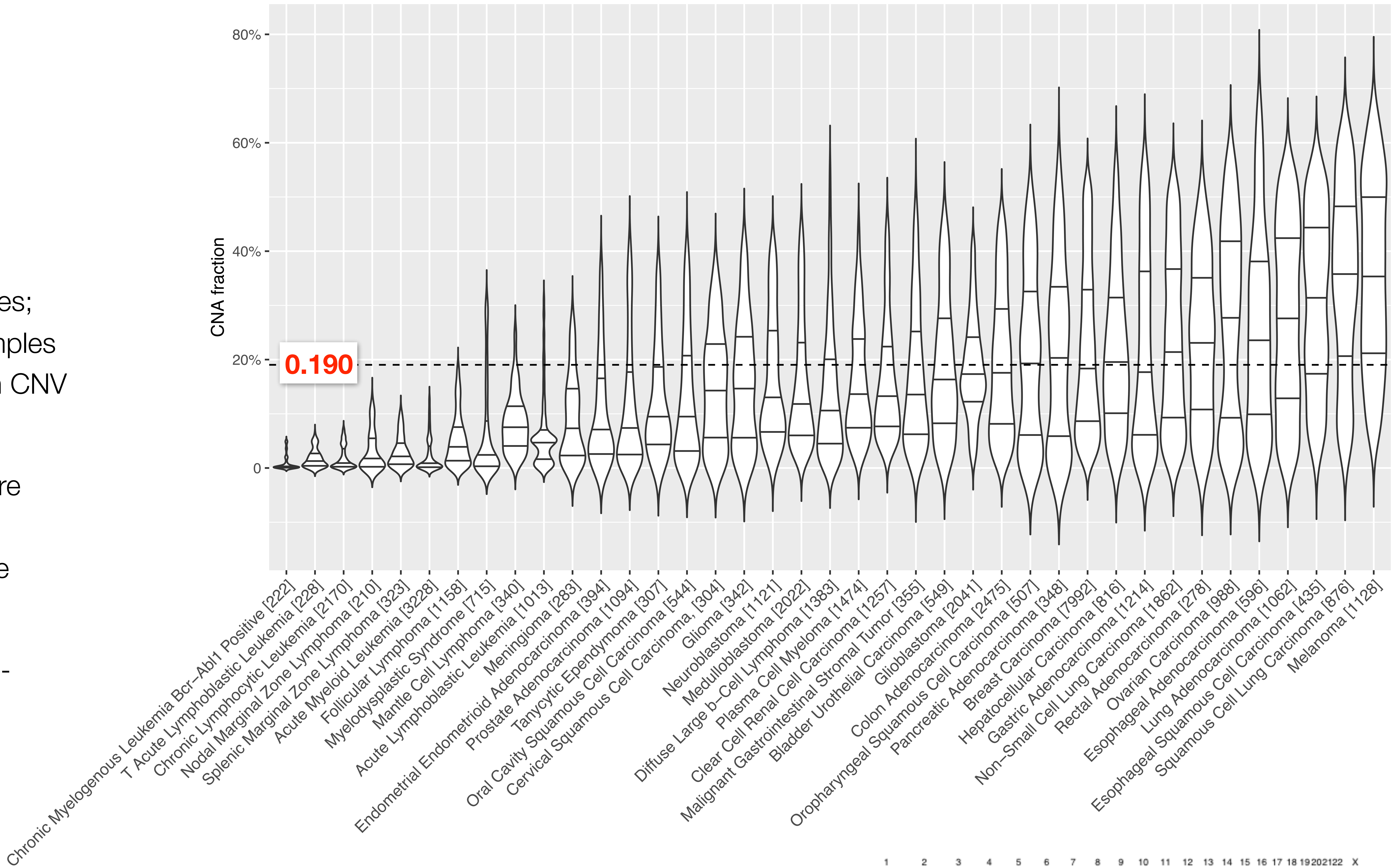
Publication	Samples			
	cCGH	aCGH	WES	WGS
Harada K, Okamoto W, Mimaki S, Kawamoto Y, Bando et al. (2019): Comparative sequence analysis of patient-matched primary colorectal cancer, metastatic, and recurrent metastatic tumors ... BMC Cancer 19(1), 2019 (30898102)	0	0	4	0
Lavrov AV, Chelysheva EY, Adilgereeva EP, Shukhov et al. (2019): Exome, transcriptome and miRNA analysis don't reveal any molecular markers of TKI efficacy in primary CML ... BMC Med Genomics 12(Suppl 2), 2019 (30871622)	0	0	62	0
Zandberg DP, Tallon LJ, Nagaraj S, Sadzewicz LK, Zhang et al. (2019): Intratumor genetic heterogeneity in squamous cell carcinoma of the oral cavity. Head Neck, 2019 (30869813)	0	0	5	0
Heinrich MC, Patterson J, Beadling C, Wang Y, Debiec-Rychter et al. (2019): Genomic aberrations in cell cycle genes predict progression of KIT-mutant gastrointestinal stromal tumors ... Clin Sarcoma Res 9, 2019 (30867899)	0	0	29	0
Jiao J, Sagnelli M, Shi B, Fang Y, Shen Z, Tang T, Dong et al. (2019): Genetic and epigenetic characteristics in ovarian tissues from polycystic ovary syndrome patients with irregular ... BMC Endocr Disord 19(1), 2019 (30866919)	0	0	20	0
Mueller S, Jain P, Liang WS, Kilburn L, Kline C, Gupta et al. (2019): A pilot precision medicine trial for children with diffuse intrinsic pontine glioma - PNOC003: a report from the Pacific ... Int. J. Cancer, 2019 (30861105)	0	0	14	14
Xie SN, Cai YJ, Ma B, Xu Y, Qian P, Zhou JD, Zhao et al. (2019): The genomic mutation spectrums of breast fibroadenomas in Chinese population by whole exome sequencing ... Cancer Med, 2019 (30851086)	0	0	12	0

Showing 1 to 50 of 3,232 entries

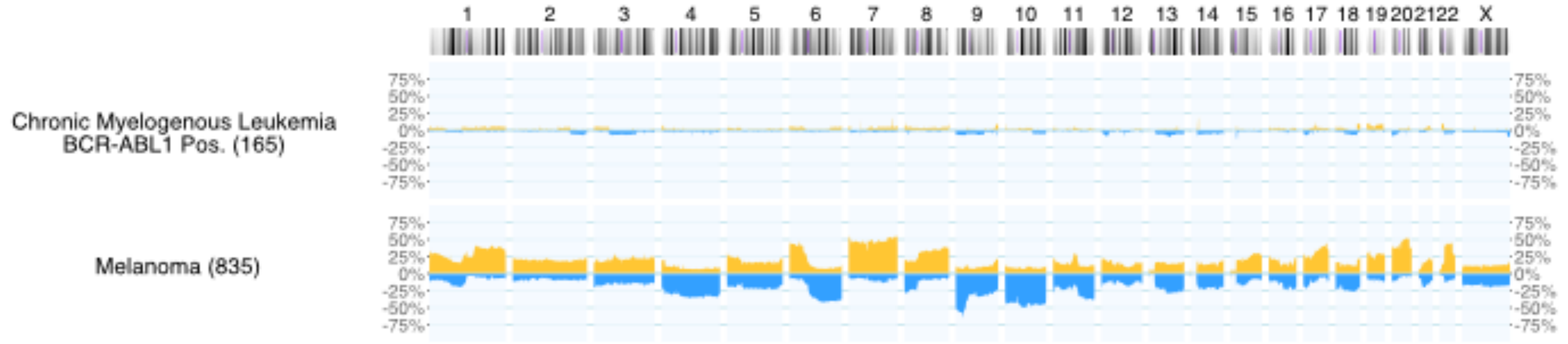
Previous 1 2 3 4 5 ... 65 Next

Genome CNV coverage in Cancer Classes

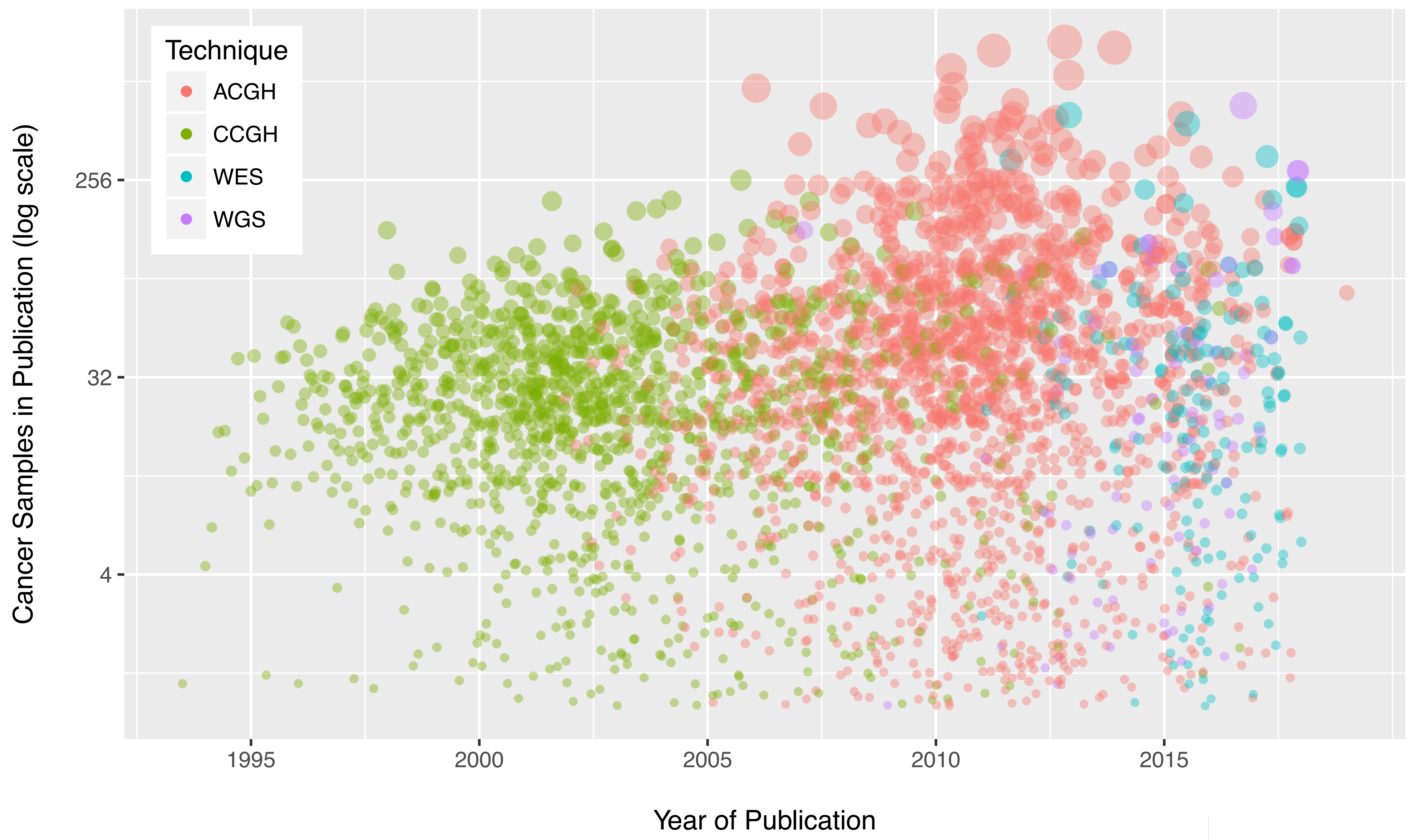
- 43654 out of 93640 CNV profiles; filtered for entities w/ >200 samples (removed some entities w/ high CNV rate, e.g. sarcoma subtypes)
- Single-sample CNV profiles were assessed for the fraction of the genome showing CNVs (relative gains, losses)
- range of medians 0.001 (CML) - 0.358 (malignant melanomas)



Lowest / Highest CNV fractions =>



Publication Landscape of Cancer CNV Profiling

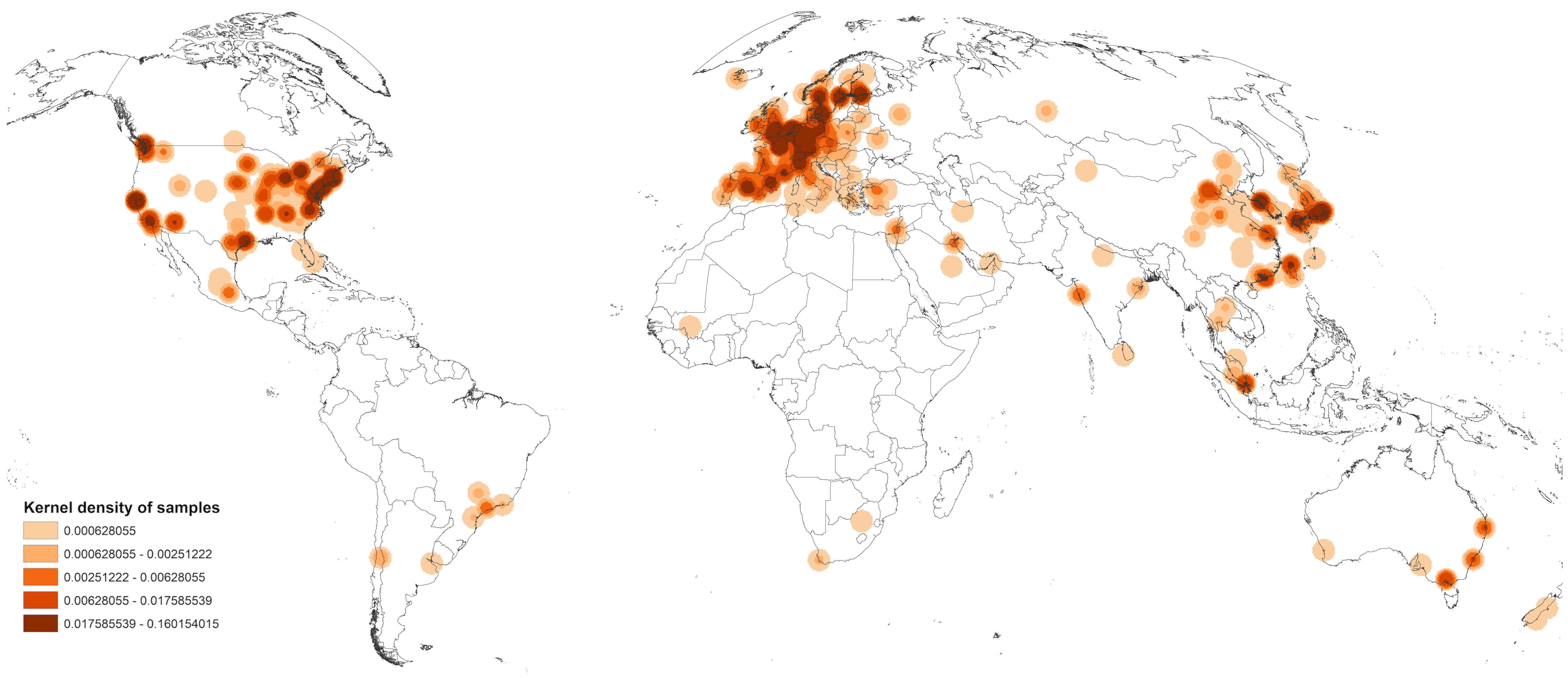


Publication statistics for cancer genome screening studies. The graphic shows our as- sessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

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

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

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



GA4GH API promotes sharing

GENOMICS

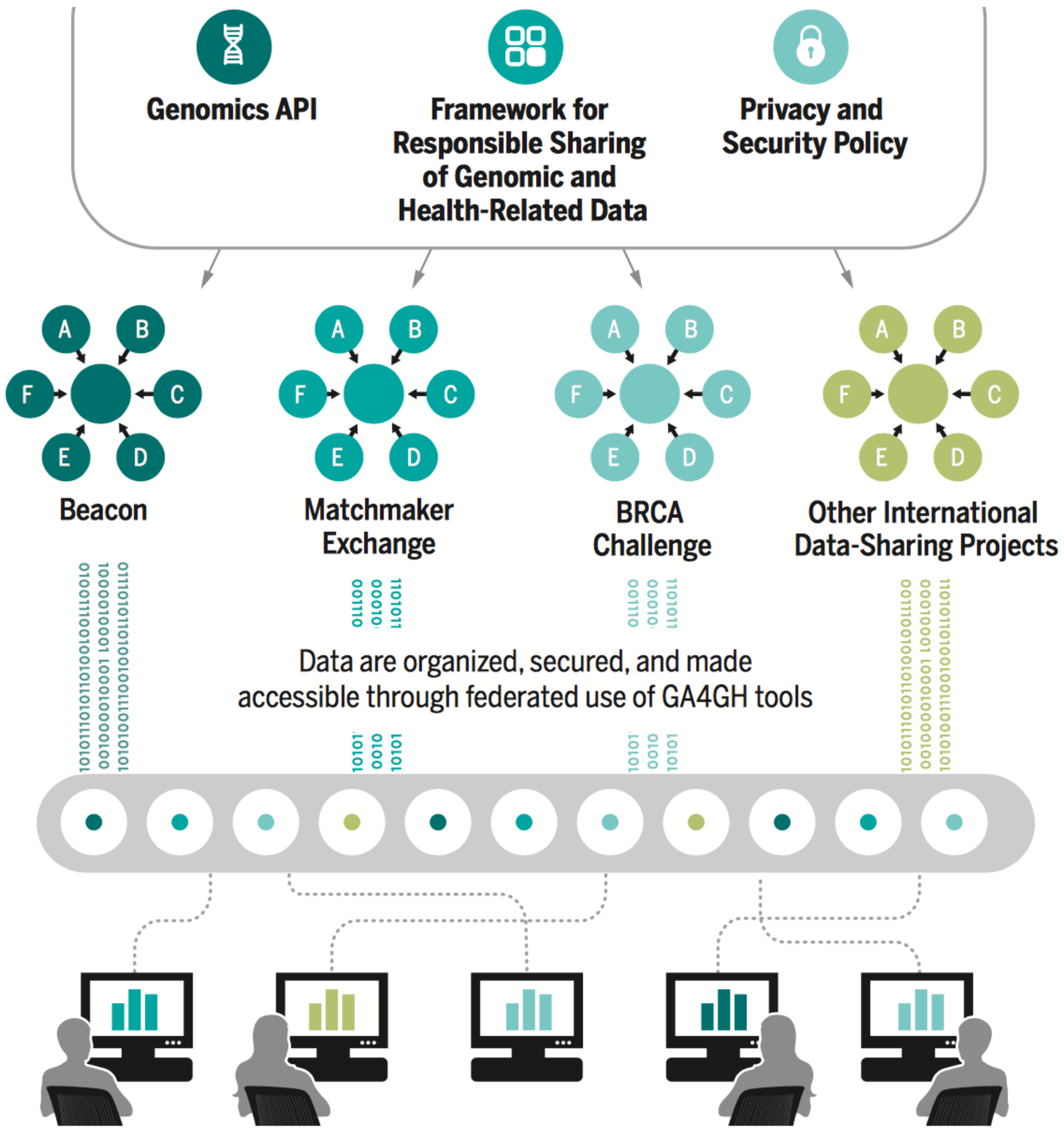
A federated ecosystem for sharing genomic, clinical data

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.



Genomics API

Framework for Responsible Sharing of Genomic and Health-Related Data

Privacy and Security Policy

Beacon

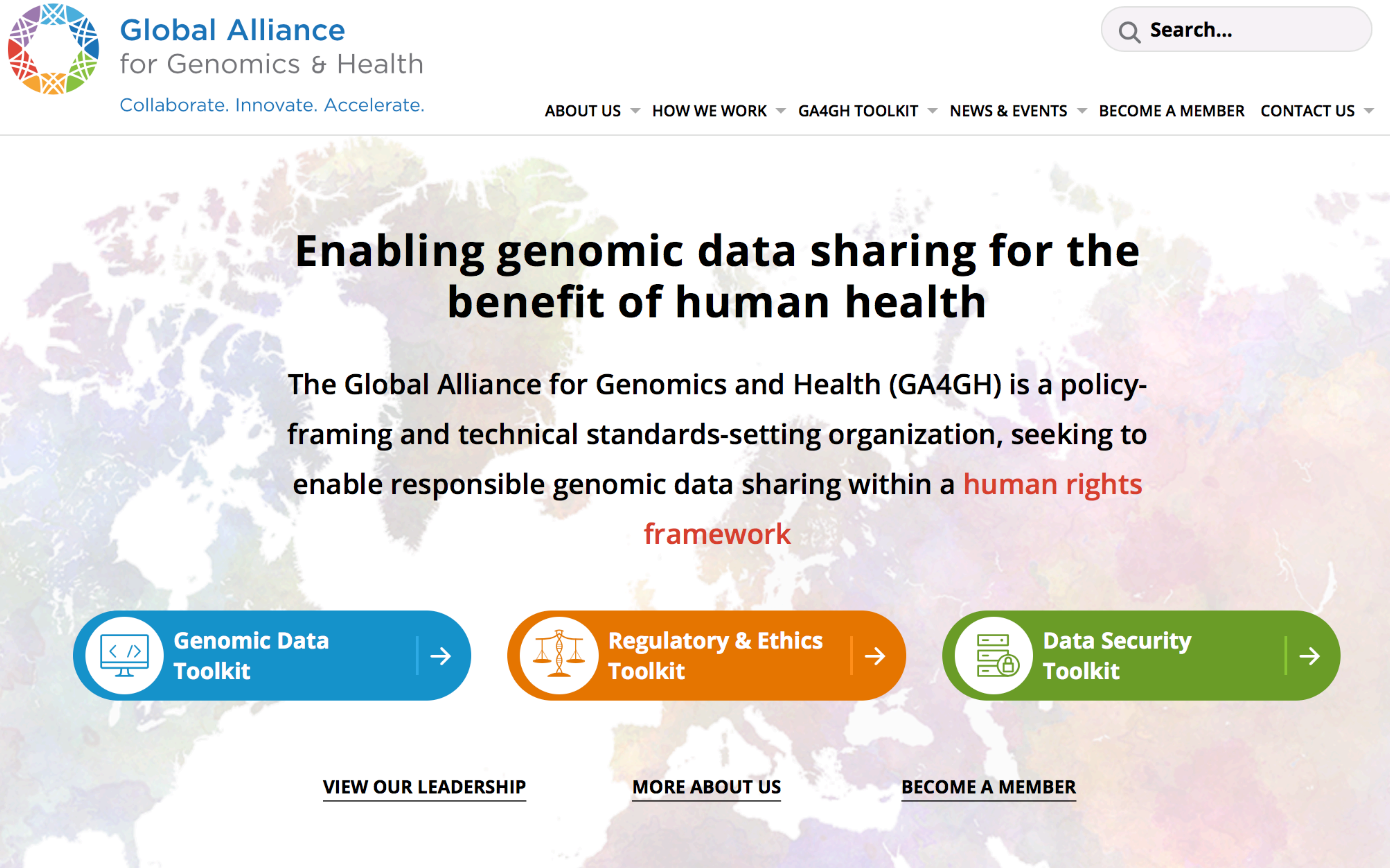
Matchmaker Exchange

BRCA Challenge

Other International Data-Sharing Projects

Data are organized, secured, and made accessible through federated use of GA4GH tools

Global Alliance
for Genomics & Health



Enabling genomic data sharing for the benefit of human health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework

Global Alliance
for Genomics & Health

GA4GH

VISION FOR GENOMIC & HEALTH RELATED DATA SHARING IN 2022

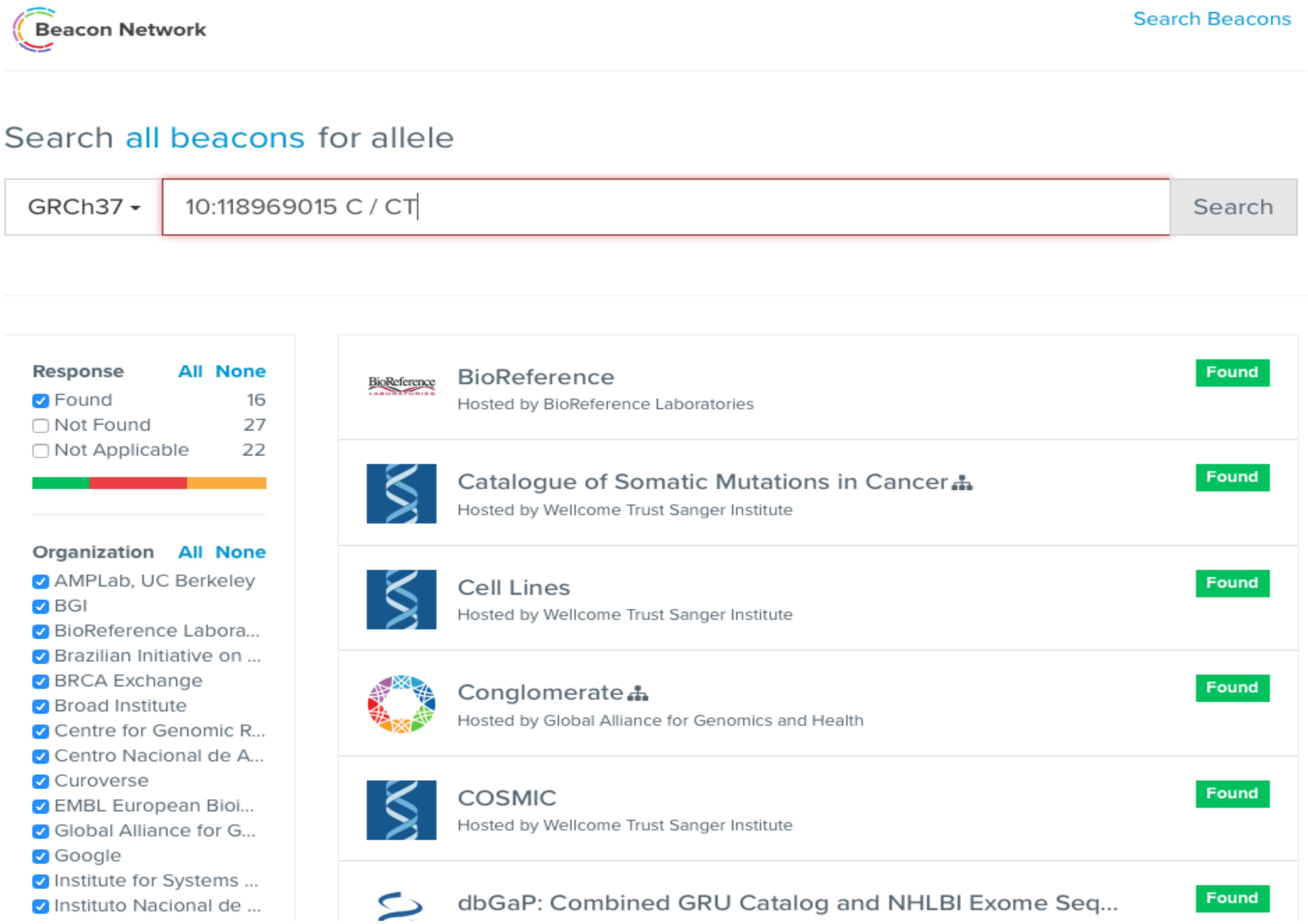
Summary:

- **In 2022, genomic data on tens of millions of individuals are responsibly accessible via GA4GH standards**.
  - Vast majority of this data has been generated due to healthcare approaches rather than research commissioned genomes.
  - Both research-commission genomes and secondary use of healthcare genomes for research is accessible due to the consistent application of the GA4GH APIs, SOPs and tools.
- **Genomics data that can be shared responsibly, are shared responsibly,** meaning every qualified clinician, researcher, and corporate entity around the globe, shares and has access to, the maximal dataset that is privacy preserving within the context of the relevant and localised consent and authorization policies.
- **Genomic and phenotypic are integrated in clinical records** and form a “healthcare learning system”.
- **GA4GH collaborates and coordinates** with the many other global, national, regional, and enterprise activities within the genomics and health ecosystem and regularly engages policymakers to ensure ongoing funding of genomic testing and sustainability

Katryn North, 2017

Beacon Project

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



Beacon Network

Search Beacons

Search all beacons for allele

GRCh37 ▾ | 10:118969015 C / CT | Search

Response All None

- ☑ Found 16
- □ Not Found 27
- □ Not Applicable 22

Organization All None

- ☑ AMPLab, UC Berkeley
- ☑ BGI
- ☑ BioReference Labora...
- ☑ Brazilian Initiative on ...
- ☑ BRCA Exchange
- ☑ Broad Institute
- ☑ Centre for Genomic R...
- ☑ Centro Nacional de A...
- ☑ Curoverse
- ☑ EMBL European Bioi...
- ☑ Global Alliance for G...
- ☑ Google
- ☑ Institute for Systems ...
- ☑ Instituto Nacional de ...

BioReference — Hosted by BioReference Laboratories — Found

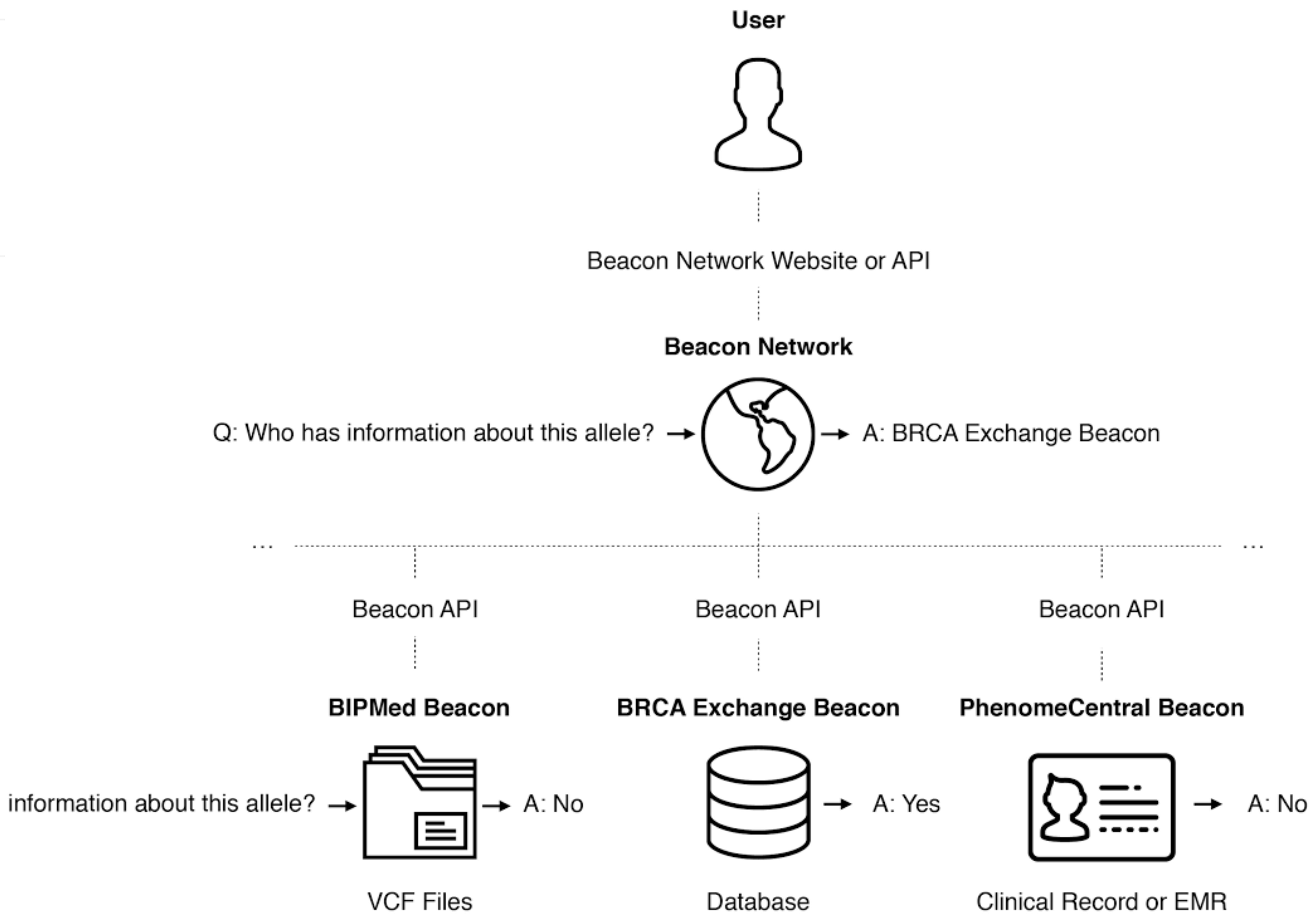
Catalogue of Somatic Mutations in Cancer — Hosted by Wellcome Trust Sanger Institute — Found

Cell Lines — Hosted by Wellcome Trust Sanger Institute — Found

Conglomerate — Hosted by Global Alliance for Genomics and Health — Found

COSMIC — Hosted by Wellcome Trust Sanger Institute — Found

dbGaP: Combined GRU Catalog and NHLBI Exome Seq... — Found



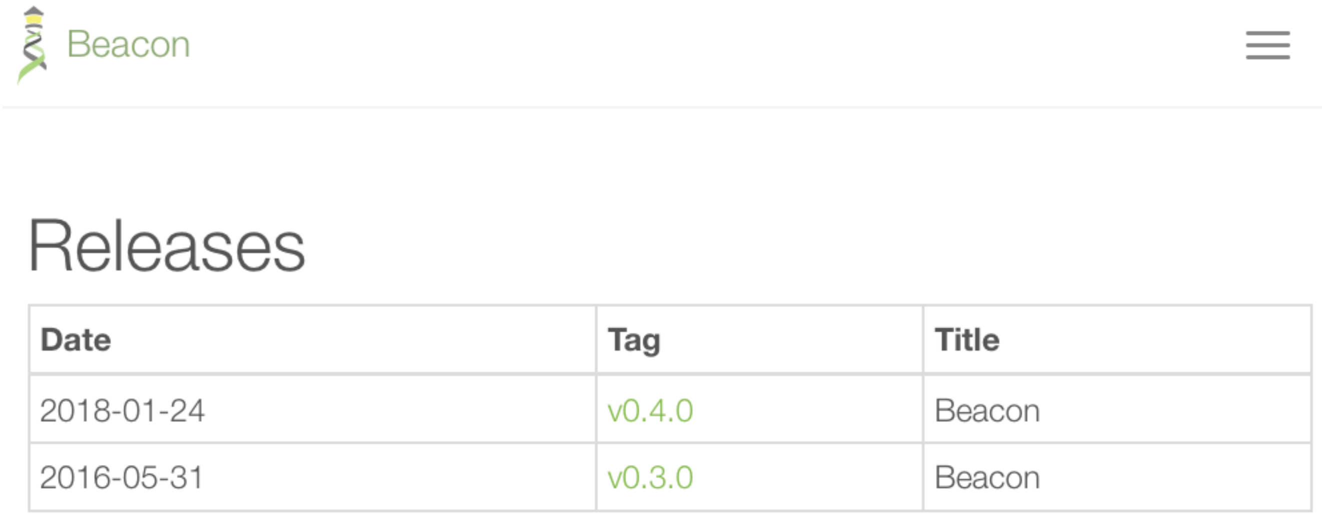
Global Alliance for Genomics & Health

35+ Organizations

90+ Beacons

200+ Datasets

100K+ Individuals



Beacon

Releases

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

ELIXIR - Towards Biomedical Beacons

Needs & Models Beyond Basic Variant Discovery

elixir

Global Alliance
for Genomics & Health

ELIXIR Beacon Project

- Driver project on GA4GH roadmap
- aligns with Discovery Work Stream
- strong impact on GA4GH developments as a concrete, funded project

Driver Projects

GA4GH Driver Projects **are real-world** genomic **data** initiatives that help gu
our development efforts and pilot our tools. Stakeholders around the glol
advocate, mandate, implement, and use our **frameworks** and standards in
local contexts.

ELIXIR Beacon

www.elixir-europe.org

Europe

Champions: Serena Scollen, Ilkka Lappalainen, Michael Baudis

Global Alliance for Genomics & Health

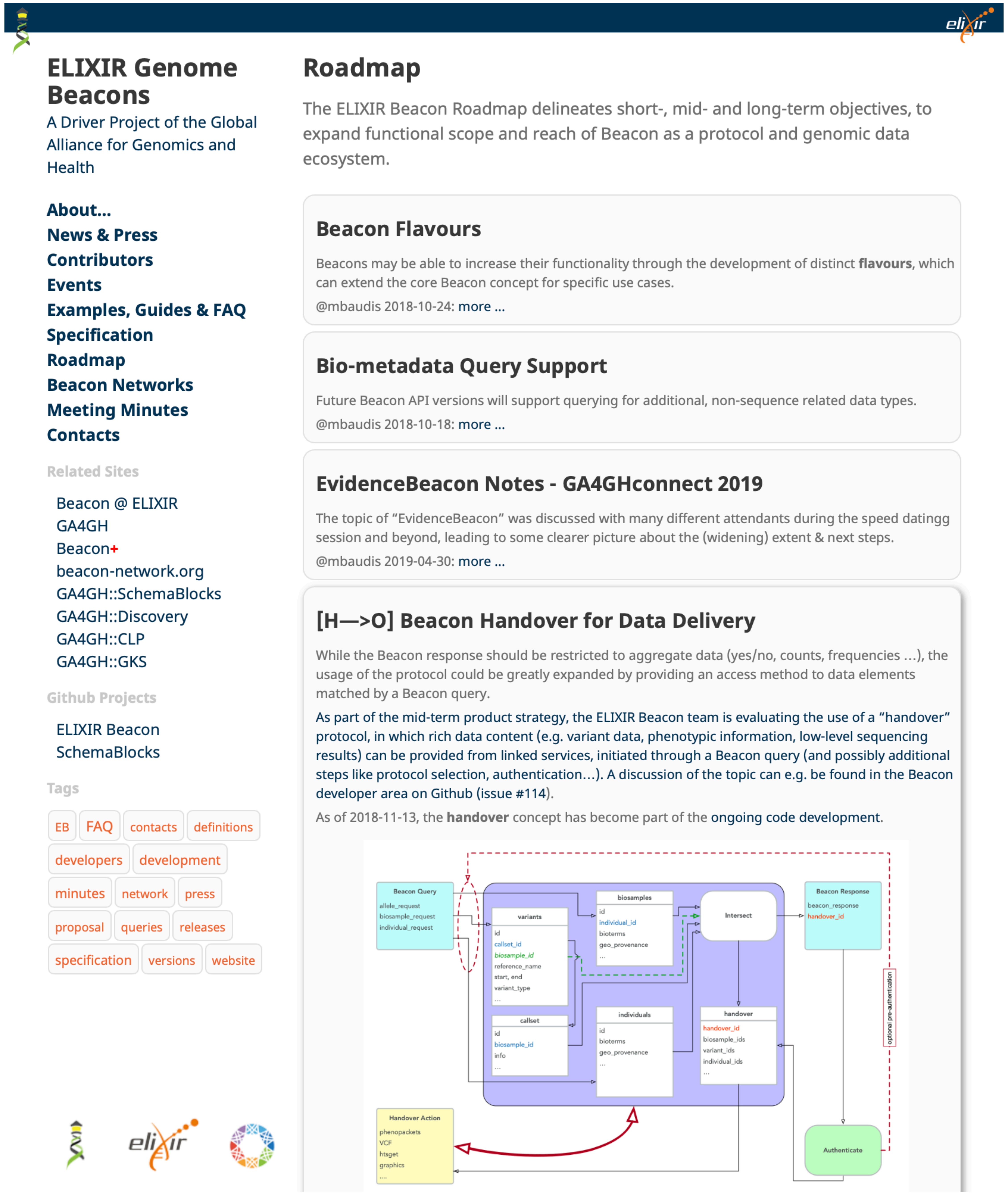
Beacon *forward*

☑ **structural variations** (DUP, DEL) in addition to SNV

- … more structural queries (translocations/fusions…)
- (bio-) **metadata** queries

☑ layered authentication system using **ELIXIR AAI**

- quantitative responses
- Beacon queries as entry for **data delivery** (outside Beacon protocol)
- Ubiquitous **deployment** (e.g. throughout ELIXIR network)



ELIXIR Genome Beacons

A Driver Project of the Global Alliance for Genomics and Health

About...
News & Press
Contributors
Events
Examples, Guides & FAQ
Specification
Roadmap
Beacon Networks
Meeting Minutes
Contacts

Related Sites

Beacon @ ELIXIR
GA4GH
Beacon+
beacon-network.org
GA4GH::SchemaBlocks
GA4GH::Discovery
GA4GH::CLP
GA4GH::GKS

Github Projects

ELIXIR Beacon
SchemaBlocks

Tags

EB FAQ contacts definitions developers development minutes network press proposal queries releases specification versions website

Roadmap

The ELIXIR Beacon Roadmap delineates short-, mid- and long-term objectives, to expand functional scope and reach of Beacon as a protocol and genomic data ecosystem.

Beacon Flavours

Beacons may be able to increase their functionality through the development of distinct **flavours**, which can extend the core Beacon concept for specific use cases.

@mbaudis 2018-10-24: more ...

Bio-metadata Query Support

Future Beacon API versions will support querying for additional, non-sequence related data types.

@mbaudis 2018-10-18: more ...

EvidenceBeacon Notes - GA4GHconnect 2019

The topic of “EvidenceBeacon” was discussed with many different attendants during the speed datingg session and beyond, leading to some clearer picture about the (widening) extent & next steps.

@mbaudis 2019-04-30: more ...

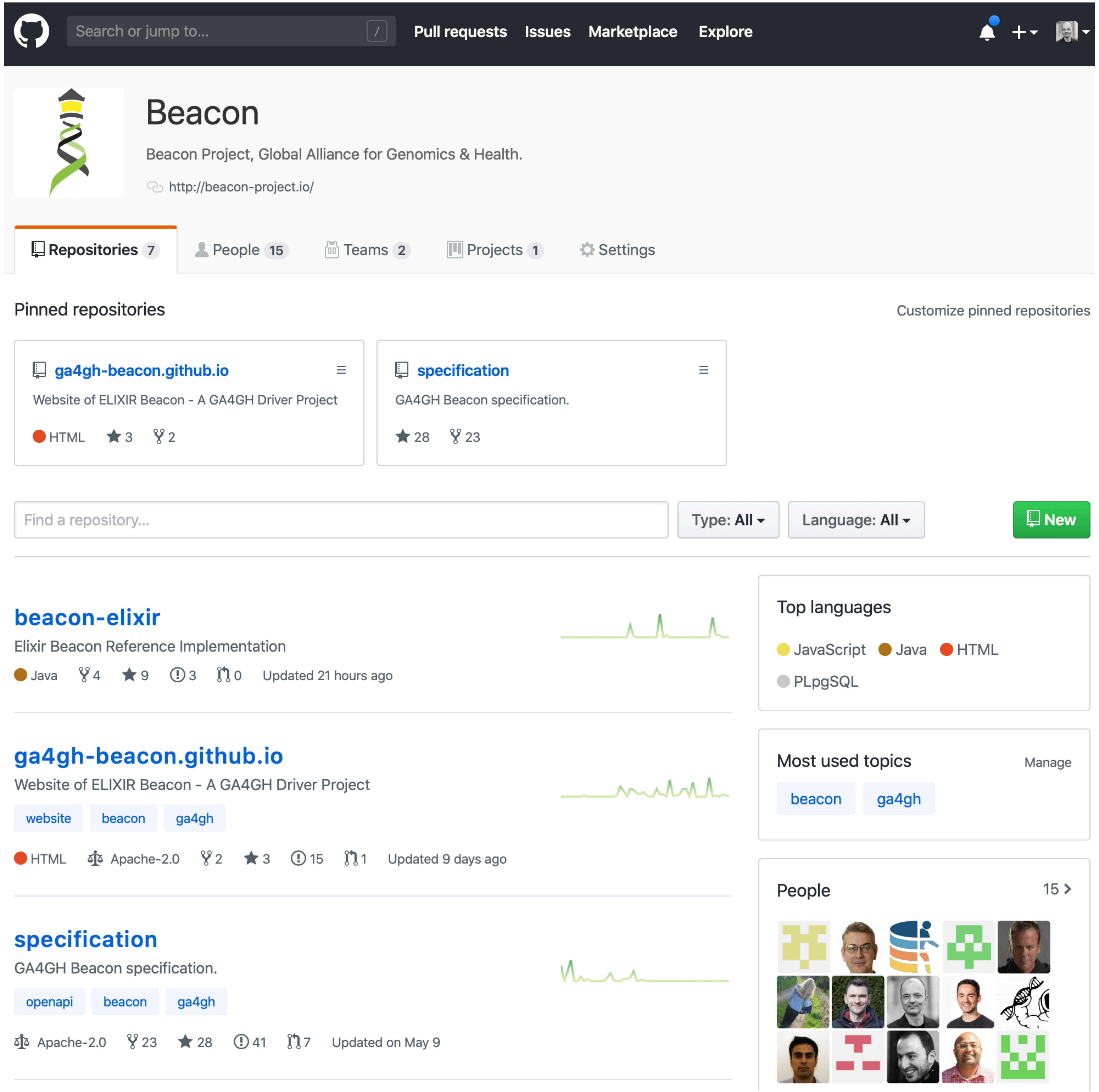
[H—>O] Beacon Handover for Data Delivery

While the Beacon response should be restricted to aggregate data (yes/no, counts, frequencies ...), the usage of the protocol could be greatly expanded by providing an access method to data elements matched by a Beacon query.

As part of the mid-term product strategy, the ELIXIR Beacon team is evaluating the use of a “handover” protocol, in which rich data content (e.g. variant data, phenotypic information, low-level sequencing results) can be provided from linked services, initiated through a Beacon query (and possibly additional steps like protocol selection, authentication...). A discussion of the topic can e.g. be found in the Beacon developer area on Github (issue #114).

As of 2018-11-13, the **handover** concept has become part of the ongoing code development.

beacon-project.io



Beacon

Beacon Project, Global Alliance for Genomics & Health.

http://beacon-project.io/

Pinned repositories

Customize pinned repositories

ga4gh-beacon.github.io
Website of ELIXIR Beacon - A GA4GH Driver Project
● HTML ★ 3 ⑂ 2

specification
GA4GH Beacon specification.
★ 28 ⑂ 23

beacon-elixir
Elixir Beacon Reference Implementation
● Java ⑂ 4 ★ 9 ⓘ 3 0 Updated 21 hours ago

ga4gh-beacon.github.io
Website of ELIXIR Beacon - A GA4GH Driver Project
website beacon ga4gh
● HTML Apache-2.0 ⑂ 2 ★ 3 ⓘ 15 1 Updated 9 days ago

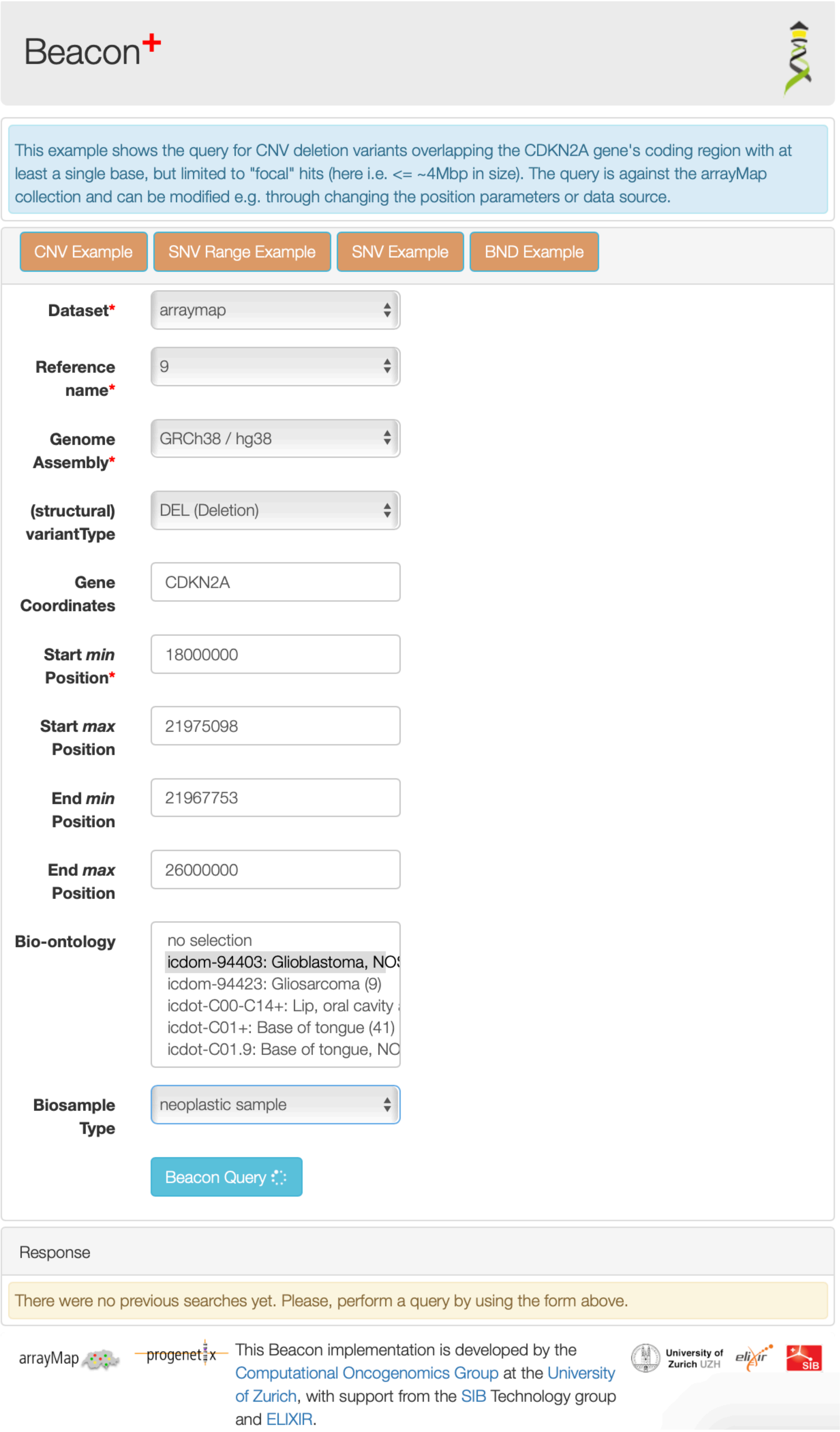
specification
GA4GH Beacon specification.
openapi beacon ga4gh
Apache-2.0 ⑂ 23 ★ 28 ⓘ 41 7 Updated on May 9

Top languages
● JavaScript ● Java ● HTML ● PLpgSQL

Most used topics
Manage
beacon ga4gh

People
15

github.com/ga4gh-beacon/



Beacon+

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "focal" hits (here i.e. <= ~4Mbp in size). The query is against the arrayMap collection and can be modified e.g. through changing the position parameters or data source.

Dataset*	arraymap
Reference name*	9
Genome Assembly*	GRCh38 / hg38
(structural) variantType	DEL (Deletion)
Gene Coordinates	CDKN2A
Start *min* Position*	18000000
Start *max* Position	21975098
End *min* Position	21967753
End *max* Position	26000000
Bio-ontology	no selection icdom-94403: Glioblastoma, NOS icdom-94423: Gliosarcoma (9) icdot-C00-C14+: Lip, oral cavity icdot-C01+: Base of tongue (41) icdot-C01.9: Base of tongue, NO
Biosample Type	neoplastic sample

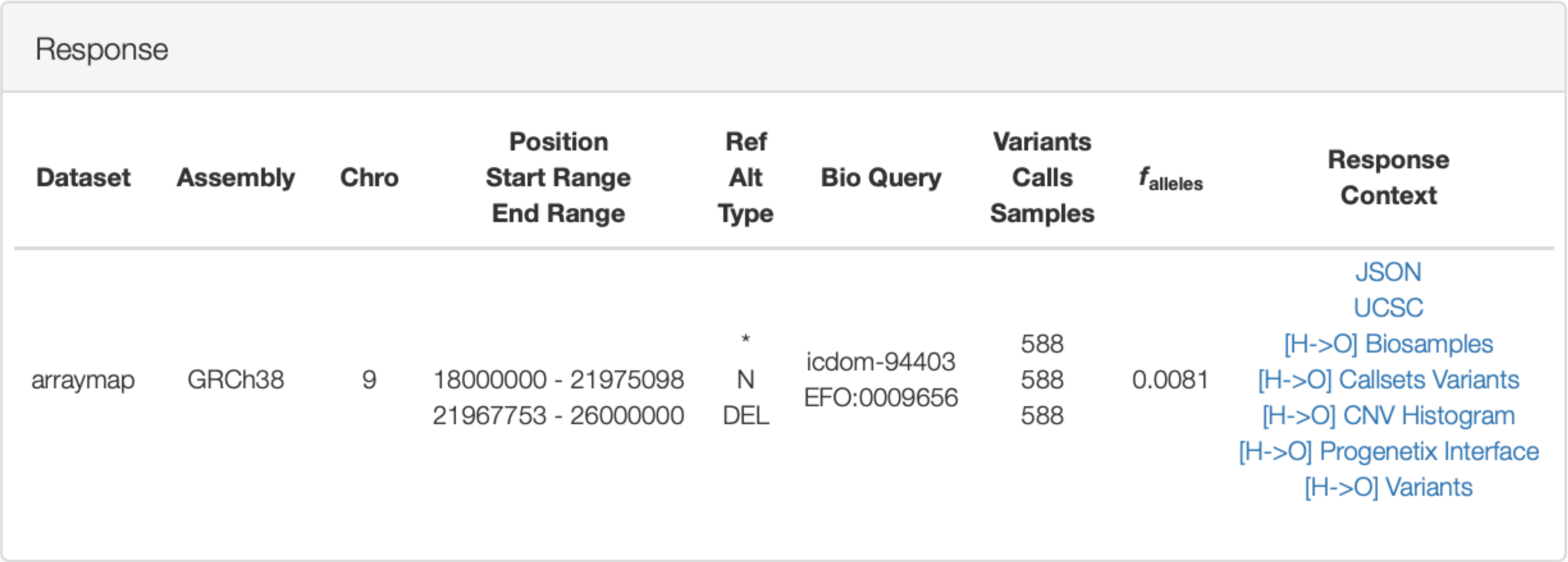
Response

There were no previous searches yet. Please, perform a query by using the form above.

This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.

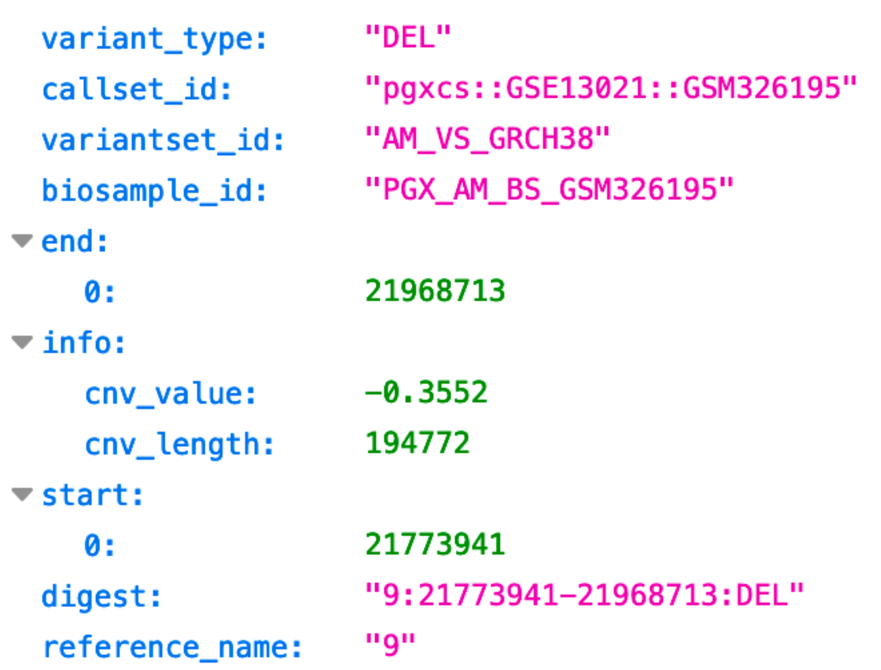
Beacon 2019

- ✓ Handover
- ✓ Filters
- ✓ Renge Queries

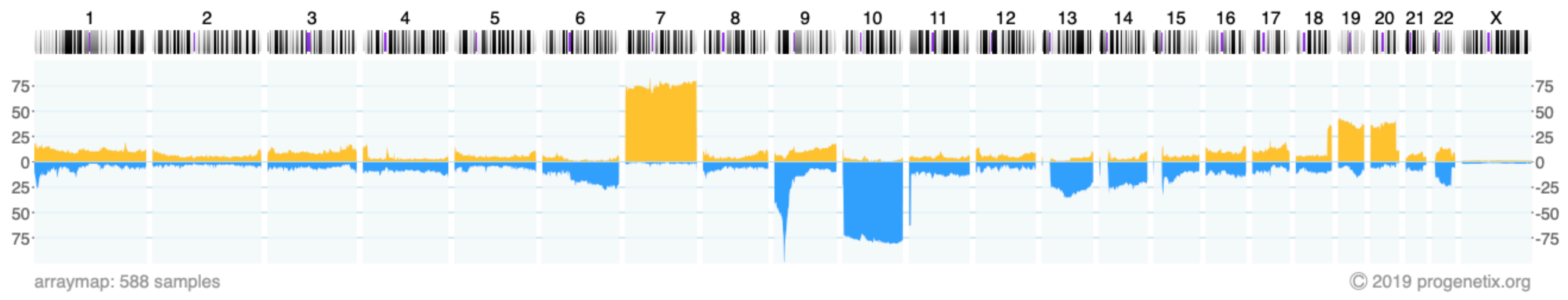


Response

Dataset	Assembly	Chro	Position Start Range End Range	Ref Alt Type	Bio Query	Variants Calls Samples	$f_{alleles}$	Response Context
arraymap	GRCh38	9	18000000 - 21975098 21967753 - 26000000	* N DEL	icdom-94403 EFO:0009656	588 588 588	0.0081	JSON UCSC [H->O] Biosamples [H->O] Callsets Variants [H->O] CNV Histogram [H->O] Progenetix Interface [H->O] Variants

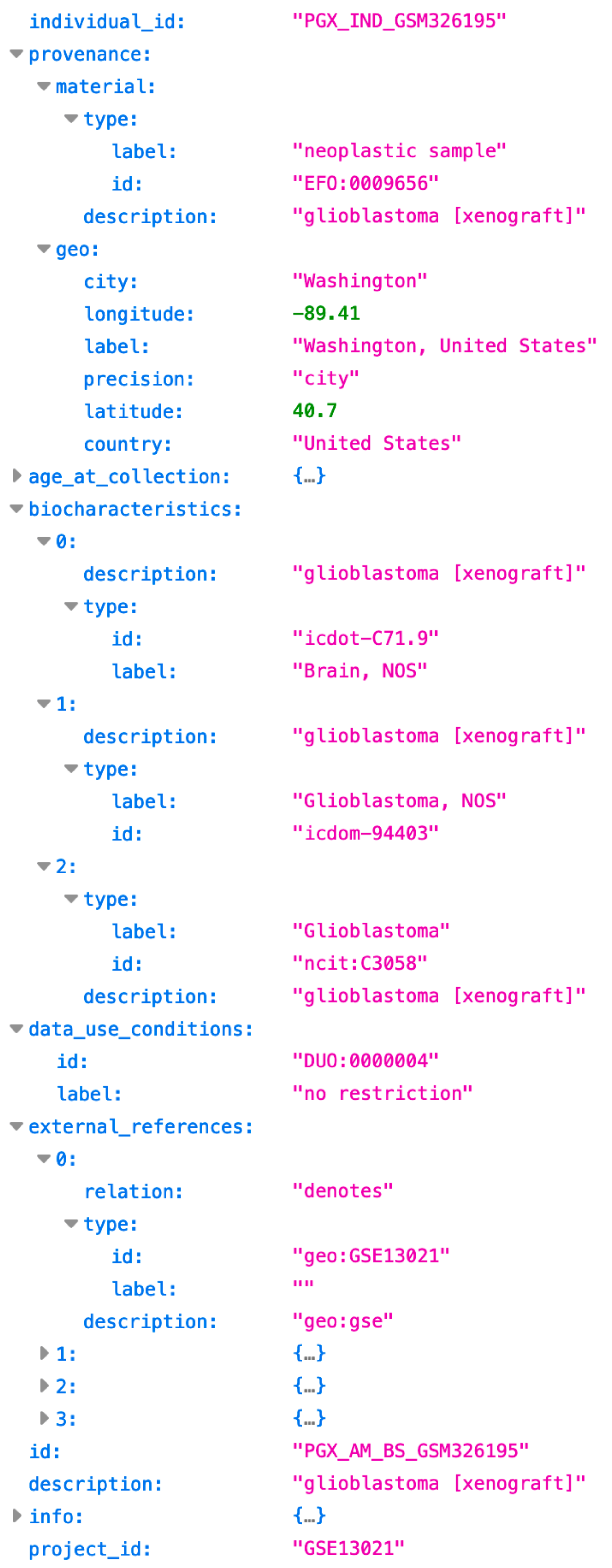


```
variant_type:       "DEL"
callset_id:         "pgxcs::GSE13021::GSM326195"
variantset_id:      "AM_VS_GRCH38"
biosample_id:       "PGX_AM_BS_GSM326195"
▼ end:
    0:              21968713
▼ info:
    cnv_value:      -0.3552
    cnv_length:     194772
▼ start:
    0:              21773941
digest:             "9:21773941-21968713:DEL"
reference_name:     "9"
```



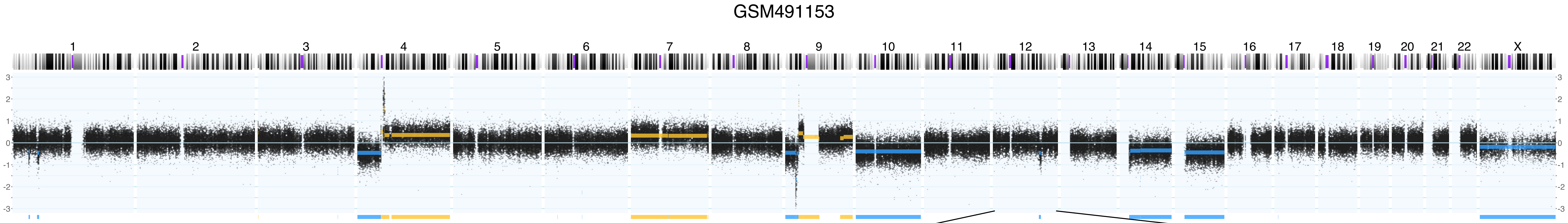
arraymap: 588 samples

© 2019 progenetix.org

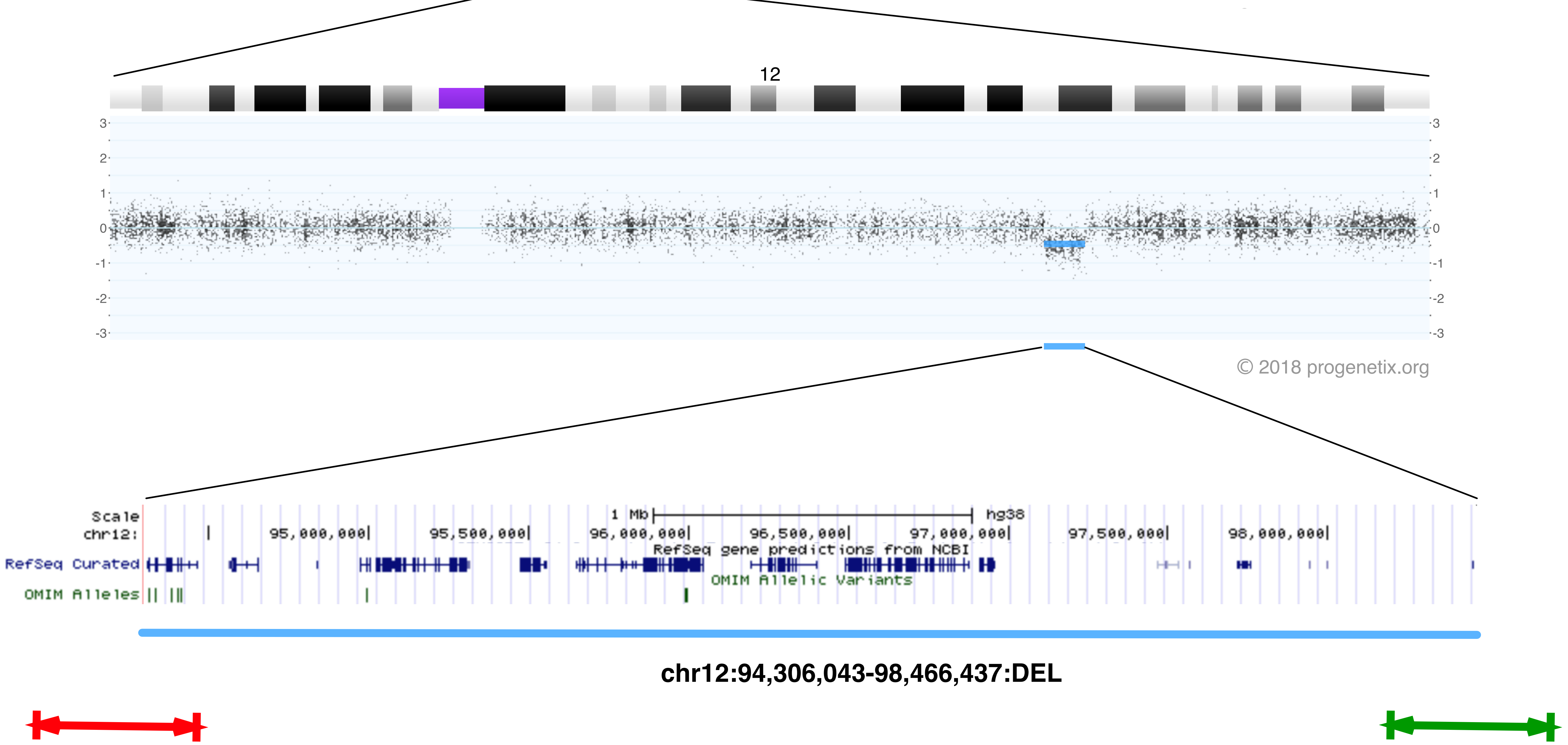


```
individual_id:              "PGX_IND_GSM326195"
▼ provenance:
  ▼ material:
    ▼ type:
        label:              "neoplastic sample"
        id:                 "EFO:0009656"
      description:          "glioblastoma [xenograft]"
  ▼ geo:
      city:                 "Washington"
      longitude:            -89.41
      label:                "Washington, United States"
      precision:            "city"
      latitude:             40.7
      country:              "United States"
▶ age_at_collection:        {…}
▼ biocharacteristics:
  ▼ 0:
      description:          "glioblastoma [xenograft]"
    ▼ type:
        id:                 "icdot-C71.9"
        label:              "Brain, NOS"
  ▼ 1:
      description:          "glioblastoma [xenograft]"
    ▼ type:
        label:              "Glioblastoma, NOS"
        id:                 "icdom-94403"
  ▼ 2:
    ▼ type:
        label:              "Glioblastoma"
        id:                 "ncit:C3058"
      description:          "glioblastoma [xenograft]"
▼ data_use_conditions:
    id:                     "DUO:0000004"
    label:                  "no restriction"
▼ external_references:
  ▼ 0:
      relation:             "denotes"
    ▼ type:
        id:                 "geo:GSE13021"
        label:              ""
      description:          "geo:gse"
  ▶ 1:                      {…}
  ▶ 2:                      {…}
  ▶ 3:                      {…}
id:                         "PGX_AM_BS_GSM326195"
description:                "glioblastoma [xenograft]"
▶ info:                     {…}
project_id:                 "GSE13021"
```

beacon.progenetix.org



- Beacon⁺ **range queries** allow the definition of a genome region of interest, containing a specified variant (or other mappable feature)
- “fuzzy” matching of region ends is essential for features without base specific positions
- current Beacon implementation addresses CNV (<DUP>, <DEL>), as are specified in VCF && GA4GH variant schema



© 2018 progenetix.org

chr12:94,306,043-98,466,437:DEL

start_min: 94,000,000
start_max: 94,500,000

variant_type: “BND”

reference_name: “9”

variant_type: “DEL”

end_min: 98,200,000
end_max: 98,700,000

variant_type: “BND”

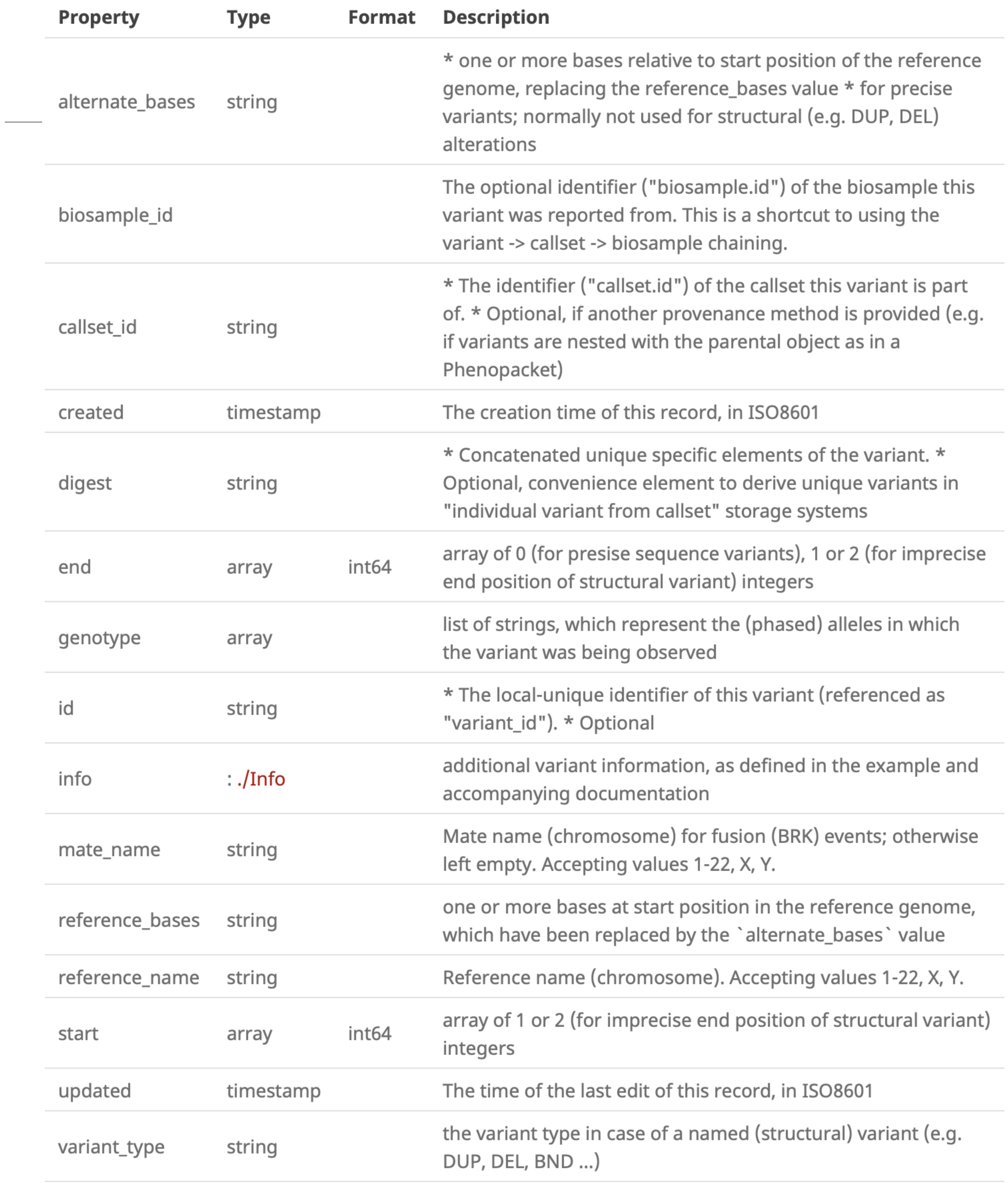
GA4GH transitional ***Variant***

- Derived from original GA4GH data schema developed by the Data Working Group
- based on the VCF file format
- representation of precise sequence alterations, copy number variants and single fusion events
- primary goals
    - sample based data storage
    - object model for query APIs (Beacon…)
- not attempting to provide reference variant, equivalence functionality
- parallel development of complete object model (allele | haplotype …, equivalence) by the GA4GH GKS work stream, based on VMC

```
{
   "biosample_id" : "structdb-bs-nhl-0009876",
   "callset_id" : "structdb-cs-nhl-0009876",
   "created" : "2019-01-22T03:06:45Z",
   "digest" : "6:63450000,63550000-63450000,63550000:DEL",
   "end" : [
      63450000,
      63550000
   ],
   "id" : "structdb-var-123456790",
   "info" : {
      "cnv_length" : 85500000,
      "cnv_value" : -0.294
   },
   "reference_bases" : "N",
   "reference_name" : 6,
   "start" : [
      63450000,
      63550000
   ],
   "updated" : "2019-02-01T12:40:21Z",
   "variant_type" : "DEL"
}
```

```
{
   "alternate_bases" : "AC",
   "callset_id" : "DIPG_CS_0290",
   "created" : "2018-11-06T11:46:30.028Z",
   "digest" : "2:203420136:A>AC",
   "genotype" : [
      "1",
      "."
   ],
   "id" : "5be1840772798347f0ed9e8b",
   "reference_bases" : "A",
   "reference_name" : "2",
   "start" : [
      203420136
   ],
   "updated" : "2018-11-06T11:46:30.028Z"
}
```

Variant Class (*schemablocks.org*)



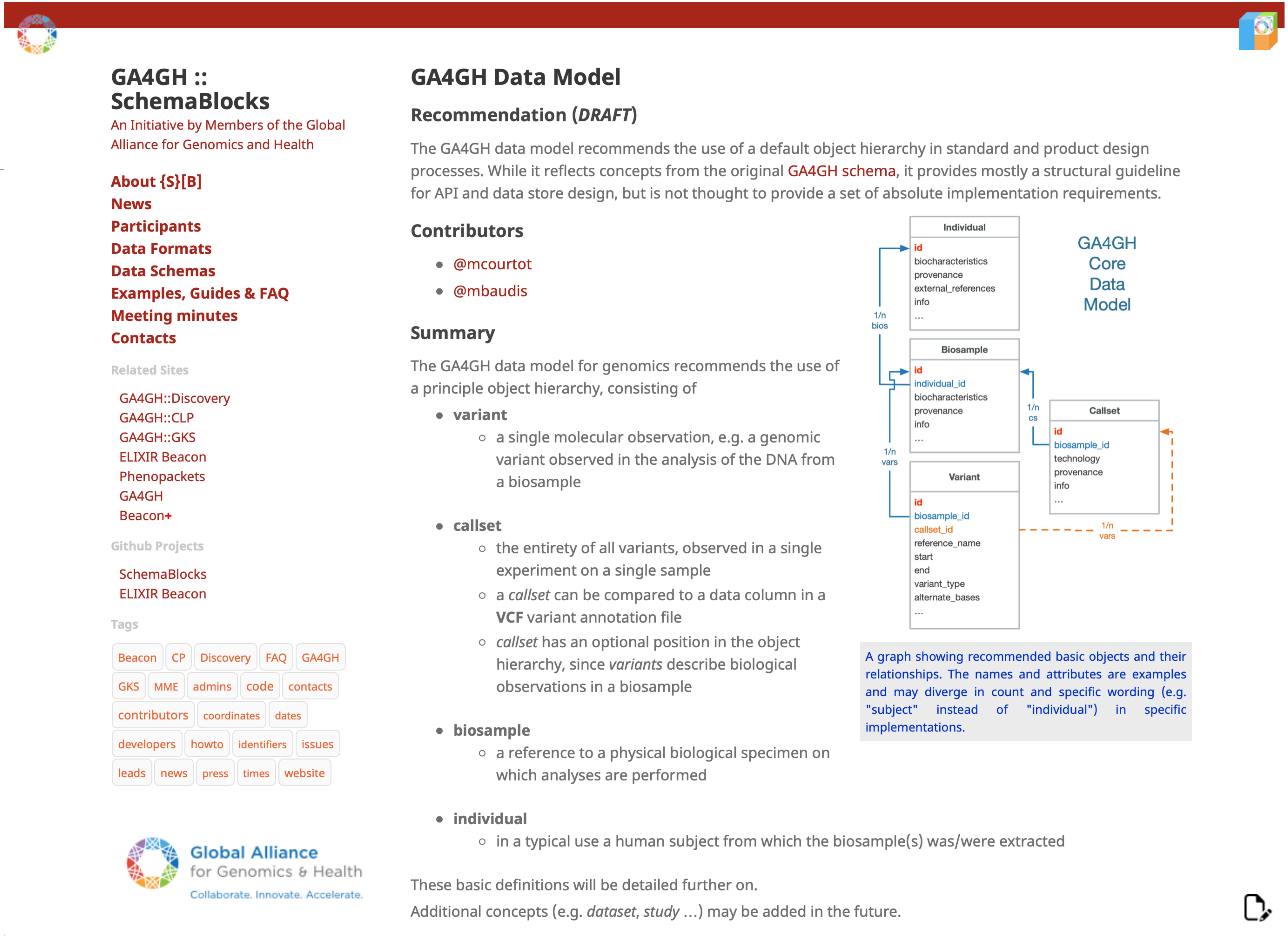
Property	Type	Format	Description
alternate_bases	string		* one or more bases relative to start position of the reference genome, replacing the reference_bases value * for precise variants; normally not used for structural (e.g. DUP, DEL) alterations
biosample_id			The optional identifier ("biosample.id") of the biosample this variant was reported from. This is a shortcut to using the variant -> callset -> biosample chaining.
callset_id	string		* The identifier ("callset.id") of the callset this variant is part of. * Optional, if another provenance method is provided (e.g. if variants are nested with the parental object as in a Phenopacket)
created	timestamp		The creation time of this record, in ISO8601
digest	string		* Concatenated unique specific elements of the variant. * Optional, convenience element to derive unique variants in "individual variant from callset" storage systems
end	array	int64	array of 0 (for presise sequence variants), 1 or 2 (for imprecise end position of structural variant) integers
genotype	array		list of strings, which represent the (phased) alleles in which the variant was being observed
id	string		* The local-unique identifier of this variant (referenced as "variant_id"). * Optional
info	: ./Info		additional variant information, as defined in the example and accompanying documentation
mate_name	string		Mate name (chromosome) for fusion (BRK) events; otherwise left empty. Accepting values 1-22, X, Y.
reference_bases	string		one or more bases at start position in the reference genome, which have been replaced by the `alternate_bases` value
reference_name	string		Reference name (chromosome). Accepting values 1-22, X, Y.
start	array	int64	array of 1 or 2 (for imprecise end position of structural variant) integers
updated	timestamp		The time of the last edit of this record, in ISO8601
variant_type	string		the variant type in case of a named (structural) variant (e.g. DUP, DEL, BND ...)

```
{
   "biosample_id" : "structdb-bs-nhl-0009876",
   "callset_id" : "structdb-cs-nhl-0009876",
   "created" : "2019-01-22T03:06:45Z",
   "digest" : "6:63450000,63550000-63450000,63550000:DEL",
   "end" : [
      63450000,
      63550000
   ],
   "id" : "structdb-var-123456790",
   "info" : {
      "cnv_length" : 85500000,
      "cnv_value" : -0.294
   },
   "reference_bases" : "N",
   "reference_name" : 6,
   "start" : [
      63450000,
      63550000
   ],
   "updated" : "2019-02-01T12:40:21Z",
   "variant_type" : "DEL"
}
```

```
{
   "alternate_bases" : "AC",
   "callset_id" : "DIPG_CS_0290",
   "created" : "2018-11-06T11:46:30.028Z",
   "digest" : "2:203420136:A>AC",
   "genotype" : [
      "1",
      "."
   ],
   "id" : "5be1840772798347f0ed9e8b",
   "reference_bases" : "A",
   "reference_name" : "2",
   "start" : [
      203420136
   ],
   "updated" : "2018-11-06T11:46:30.028Z"
}
```

GA4GH {S}[B]

- "cross-workstreams, cross-drivers" initiative to document GA4GH object standards and prototypes, data formats and semantics
- launched in December 2018
- documentation and implementation examples provided by GA4GH members
- no attempt to develop a rigid, complete data schema
- object vocabulary and semantics for a large range of developments
- currently not "authoritative GA4GH recommendations"



GA4GH :: SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

About {S}[B]
News
Participants
Data Formats
Data Schemas
Examples, Guides & FAQ
Meeting minutes
Contacts

Related Sites

GA4GH::Discovery
GA4GH::CLP
GA4GH::GKS
ELIXIR Beacon
Phenopackets
GA4GH
Beacon+

Github Projects

SchemaBlocks
ELIXIR Beacon

Tags

Beacon | CP | Discovery | FAQ | GA4GH | GKS | MME | admins | code | contacts | contributors | coordinates | dates | developers | howto | identifiers | issues | leads | news | press | times | website

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

GA4GH Data Model

Recommendation (*DRAFT*)

The GA4GH data model recommends the use of a default object hierarchy in standard and product design processes. While it reflects concepts from the original GA4GH schema, it provides mostly a structural guideline for API and data store design, but is not thought to provide a set of absolute implementation requirements.

Contributors

- @mcourtot
- @mbaudis

Summary

The GA4GH data model for genomics recommends the use of a principle object hierarchy, consisting of

- **variant**
  - a single molecular observation, e.g. a genomic variant observed in the analysis of the DNA from a biosample
- **callset**
  - the entirety of all variants, observed in a single experiment on a single sample
  - a *callset* can be compared to a data column in a **VCF** variant annotation file
  - *callset* has an optional position in the object hierarchy, since *variants* describe biological observations in a biosample
- **biosample**
  - a reference to a physical biological specimen on which analyses are performed
- **individual**
  - in a typical use a human subject from which the biosample(s) was/were extracted

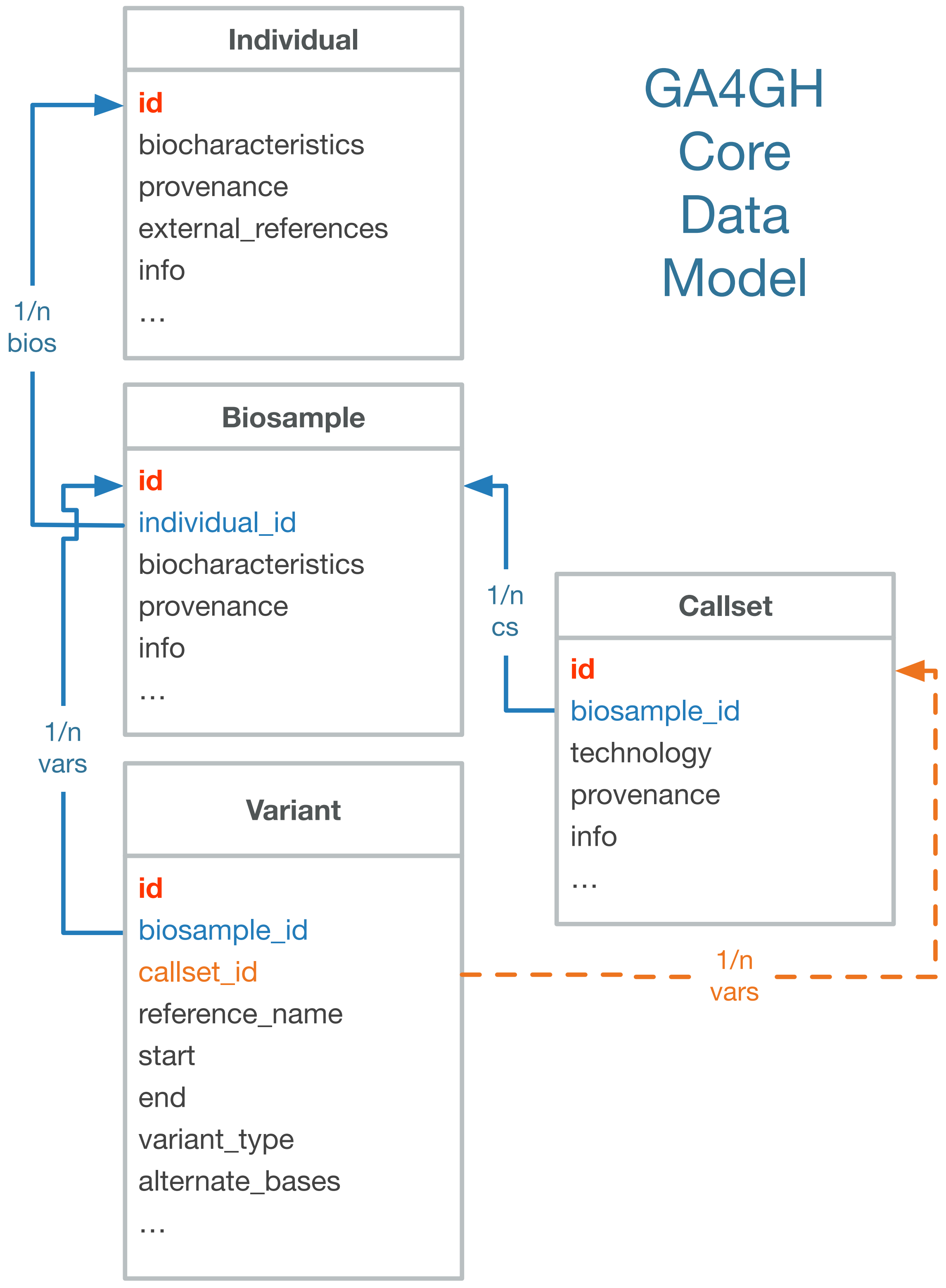
These basic definitions will be detailed further on.

Additional concepts (e.g. *dataset*, *study* …) may be added in the future.

A graph showing recommended basic objects and their relationships. The names and attributes are examples and may diverge in count and specific wording (e.g. "subject" instead of "individual") in specific implementations.

Standardized Data Model for Consistent Schema Development

- A consistent high-level data model is essential for the development of reliable schemas and tools for
  - genomic and clinical, metadata storage
  - development of genomic query and data delivery APIs
  - distributed/federated access across separate (geographic, logistic) data repositories using consistent logical structure:
    - "BRCA1 ***variant*** in germline ***sample*** from a male ***individual*** with a diagnosis of breast carcinoma (ncit:C5214)
- The abstract data model can be expressed in different types of implementations
  - Phenopackets data exchange standard
  - Progenetix database model
    - schema-derived object storage datacollections for individuals, biosamples, callsets and variants



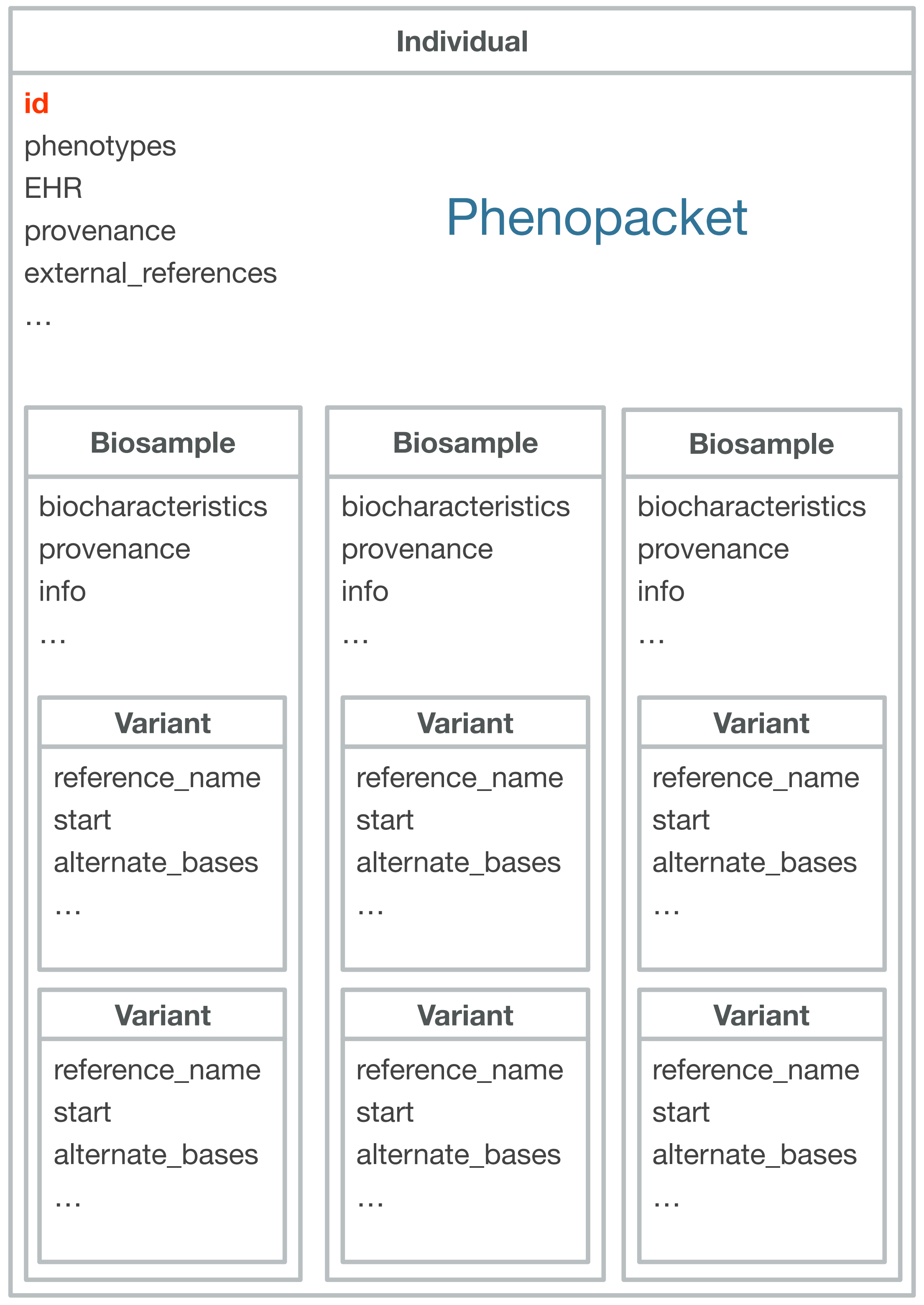
GA4GH Core Data Model

Individual
id
biocharacteristics
provenance
external_references
info
…

Biosample
id
individual_id
biocharacteristics
provenance
info
…

Variant
id
biosample_id
callset_id
reference_name
start
end
variant_type
alternate_bases
…

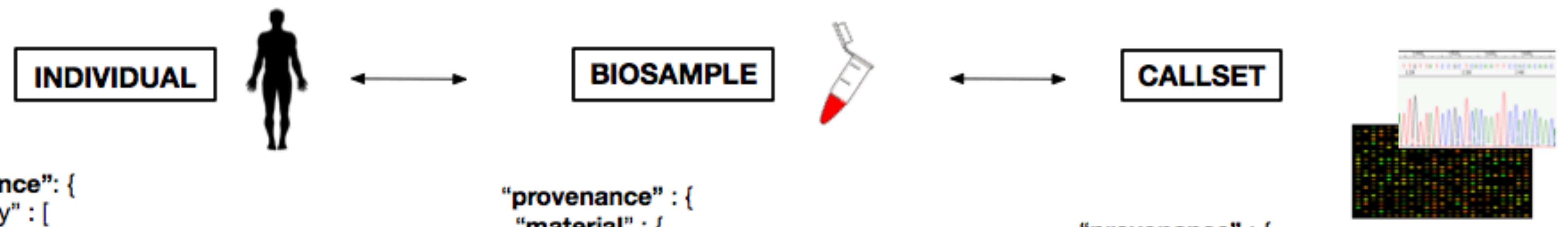
Callset
id
biosample_id
technology
provenance
info
…



Phenopacket

Individual
id
phenotypes
EHR
provenance
external_references
…

Biosample	Biosample	Biosample
biocharacteristics provenance info …	biocharacteristics provenance info …	biocharacteristics provenance info …
Variant reference_name start alternate_bases …	**Variant** reference_name start alternate_bases …	**Variant** reference_name start alternate_bases …
Variant reference_name start alternate_bases …	**Variant** reference_name start alternate_bases …	**Variant** reference_name start alternate_bases …



```
"provenance": {
  "ancestry" : [
  {
    "description":"South African ancestry",
    "type": {
      "id":"ancestro:0519",
      "label":"South African"
    }
  },
 ],
 "geo": {
  "label" : "Cape Town, South Africa, Africa",
  "precision" : "city",
  "city" : "Cape Town",
  "country" : "South Africa",
  "latitude" : 18.42,
  "longitude" : -33.93
 }
}
```

```
"provenance" : {
 "material" : {
   "description" : "neuroblastoma",
   "type" : {
     "id" : "EFO:0009656"
     "label" :  "neoplastic sample"
   }
 },
 "geo" : {
   "label" : "London, United Kingdom",
   "precision" : "city",
   "city" : "London",
   "country" : "United Kingdom",
   "latitude" : -0.13,
   "longitude" : 51.51
  }
}
```

```
"provenance" : {
 "geo" : {
   "label" : "New York City, United States",
   "precision" : "city",
   "city" : "New York City",
   "country" : "United States",
   "latitude" : 40.71,
   "longitude" : -74.01
  }
}
```

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

Random Thoughts on "Big Data" CNVs for Cancer Genomics

- Data accessibility - **quantity**
  - open data w/ "just in time" access & active work to open repositories, archives
  - data curation and long term storage has to be promoted and supported
- New technologies for **qualitatively** new possibilities
  - deep WGS with molecular reconstruction of complex events (chromothripsis / kategesis / chromoplexis...)
- Annotation and exchange formats have to move towards extensible models
  - referring reference genome positions, w/ remapping, provenance
  - technology agnostic (but provenance...)
- Search and exchange APIs have to accommodate distributed and/or federated data access models
  - modular object design, independent from backend structure
  - common interfaces/service APIs/registries

BAUDISGROUP @ UZH

NI AI
MICHAEL BAUDIS
(HAOYANG CAI)
PAULA CARRIO CORDO
BO GAO
QINGYAO HUANG
SAUMYA GUPTA
(NITIN KUMAR)
(RAHEL PALOOTS)

SIB

AMOS BAIROCH
HEINZ STOCKINGER
DANIEL TEIXEIRA

THOMAS EGGERMANN
ROSA NOGUERA
REINER SIEBERT
CAIUS SOLOVAN

GA4GH

LARRY BABB
ANTHONY BROOKES
MELANIE COURTOT
MELISSA HAENDEL
MICHAEL MILLER
HELEN PARKINSON
GUNNAR RÄTSCH
ANDY YATES

ELIXIR & CRG

JORDI RAMBLA DE ARGILA
GARY SAUNDERS
ILKKA LAPPALAINEN
S. DE LA TORRE PERNAS
SERENA SCOLLEN
JUHA TÖRNROOS

University of Zurich UZH

SIB

elixir

Global Alliance for Genomics & Health

University of Zurich^UZH^

SIB

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

arraymap.org
progenetix.org
info.baudisgroup.org
sib.swiss/baudis-michael
imls.uzh.ch/en/research/baudis
beacon-project.io
schemablocks.org

Global Alliance
for Genomics & Health

elixir