Cancer Genomics and Implementation of Data Driven Standards for Genomic Data Exchange

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

University of Zurich[UZH]

elixir

SIB

Global Alliance for Genomics & Health

1992	Heidelberg	Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Lichter) | resident in clinical hematology/oncology | data, clinical studies & cancer systematics
2001	Stanford	Post-doc in hemato-pathology (Michael Cleary) | molecular mechanisms of leukemogenesis | transgenic models | expression arrays | systematic cancer genome data collection | *Progenetix* website
2003	Gainesville	Assistant professor in paediatric haematology | molecular mechanisms of leukemogenesis | focus on bioinformatics for cancer genome data analysis
2006	Aachen	Research group leader in genetics | genomic array analysis for germline alterations | descriptive analysis of copy number aberration patterns in cancer entities
2007	Zürich	Professor of bioinformatics @ IMLS (2015) | systematic assembly of oncogenomic data | databases and software tools | patterns in cancer genomes | *Progenetix* & *arrayMap* resources | GA4GH | SPHN

Michael @ SIB, GA4GH, ELIXIR & SPHN

- member GA4GH since 2014
  - co-lead Discovery WS (2017->)
- ELIXIR Beacon project
  - previous co-chair; now responsible GA4GH liaison
- ELIXIR h-CNV project
- Swiss Personalized Health Network (SPHN)
  - SPHN project champion @ GA4GH
- Swiss Institute of Bioinformatics (SIB) group leader

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

elixir BEACON

Beacon

elixir

Swiss Personalized Health Network

SIB
Swiss Institute of Bioinformatics

Structural Genome Variants in Cancer: Research & Resources

sCNV Frequencies and Patterns :: CNV Databases :: Bioinformatics Tools

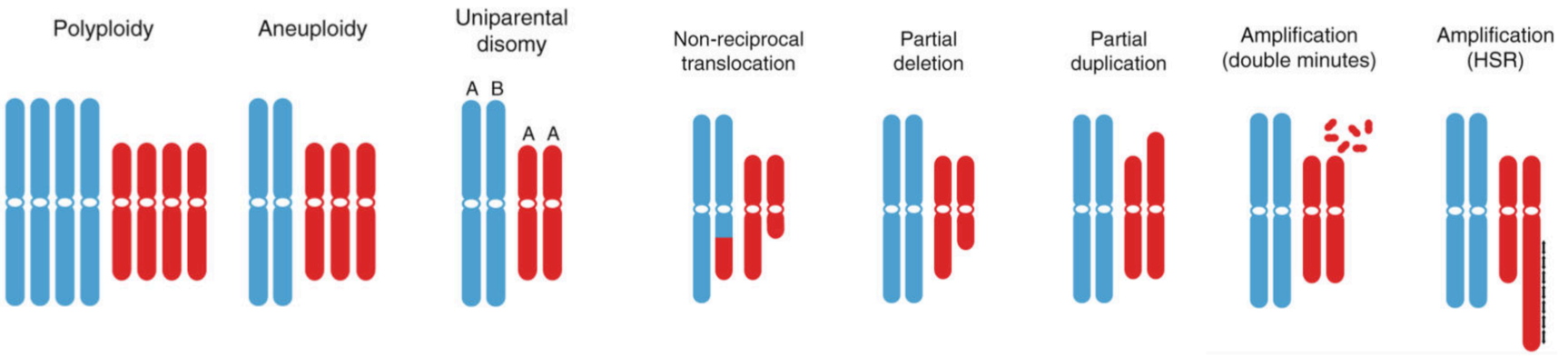
University of Zurich[UZH]
elixir
SIB
Global Alliance for Genomics & Health

Introduction

Types of genomic alterations in Cancer

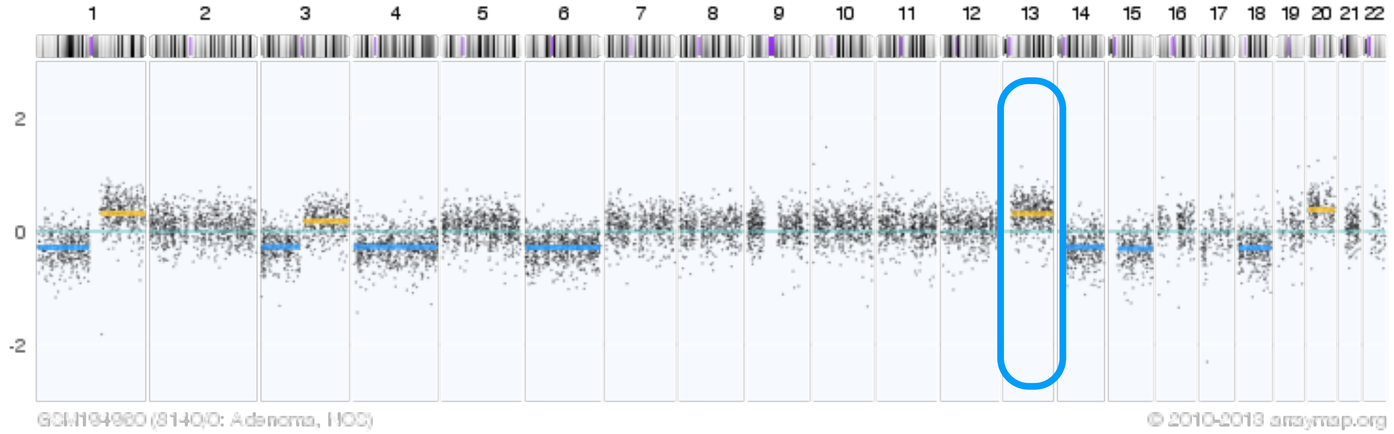
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)

Imbalanced Chromosomal Changes: CNV

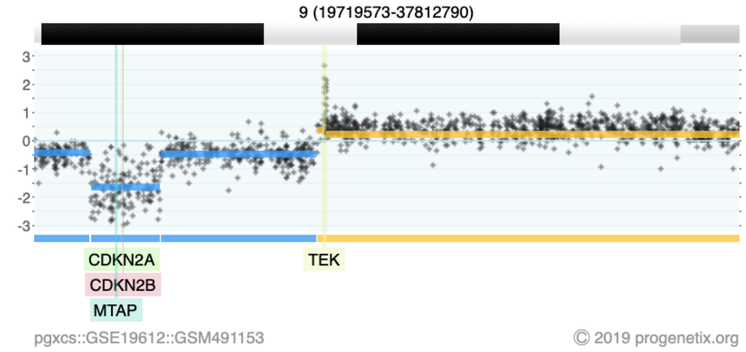


Grade et al., 2015 Recent Results Cancer Res

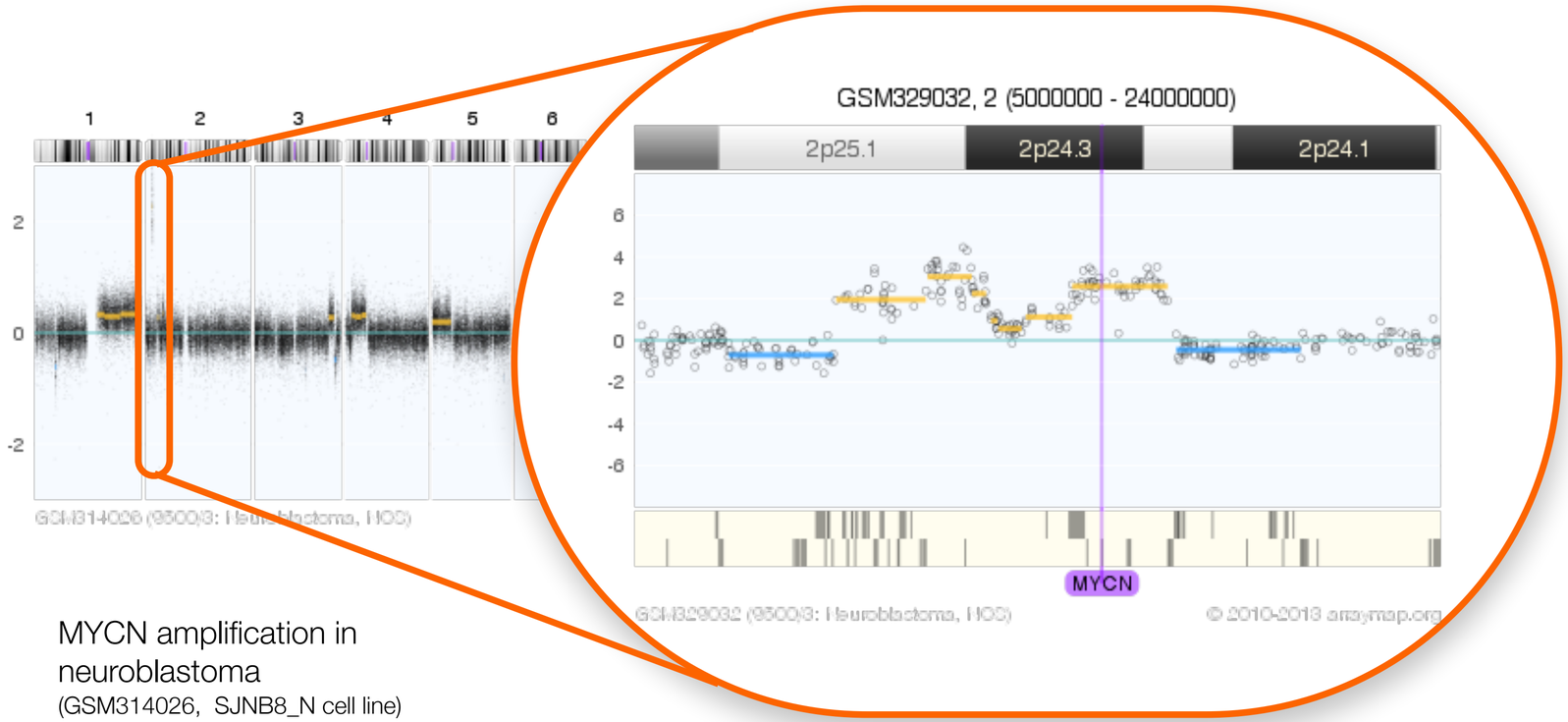
Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma

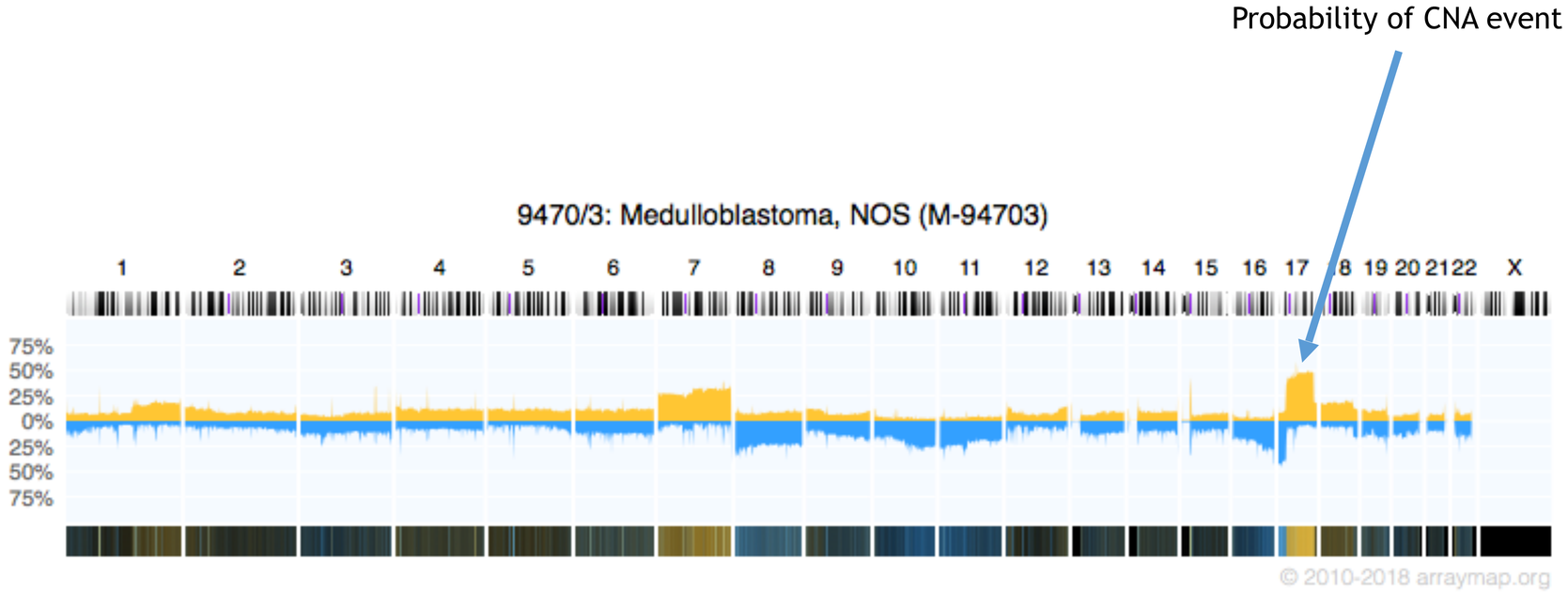


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

low level/**high level** copy number alterations (CNAs)

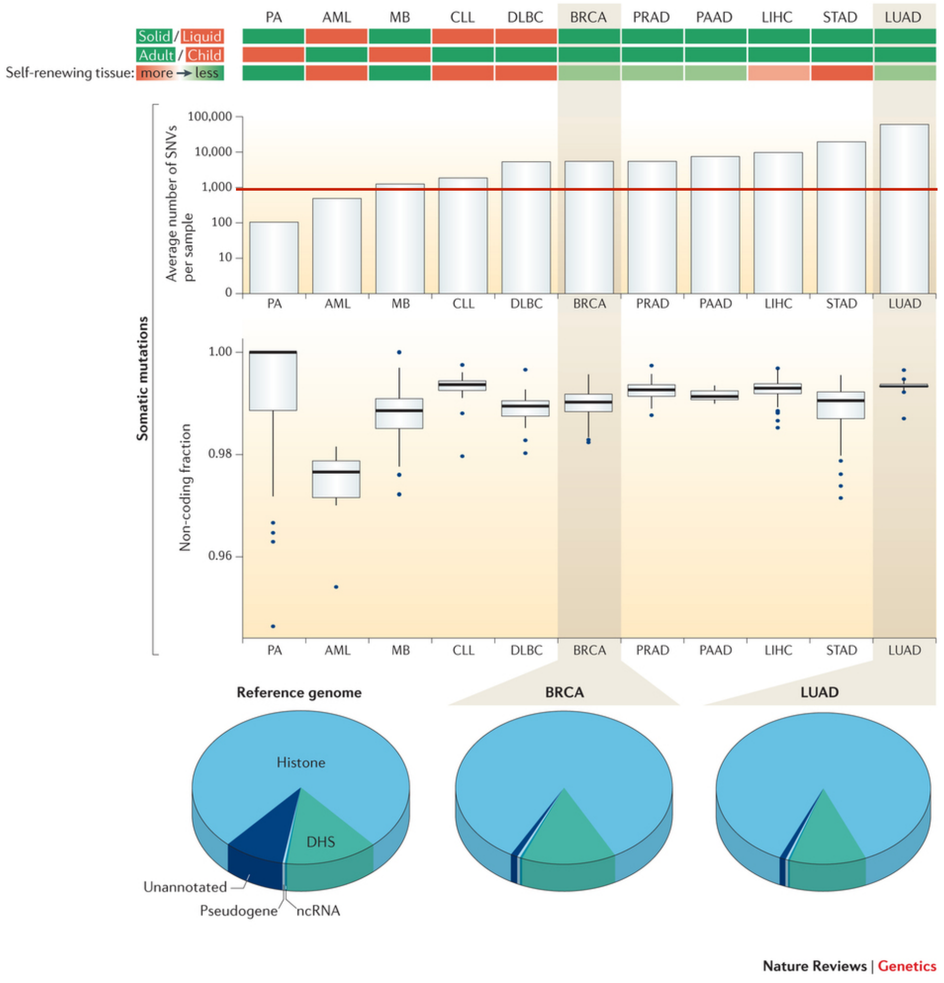
arrayMap

Copy Number Aberrations

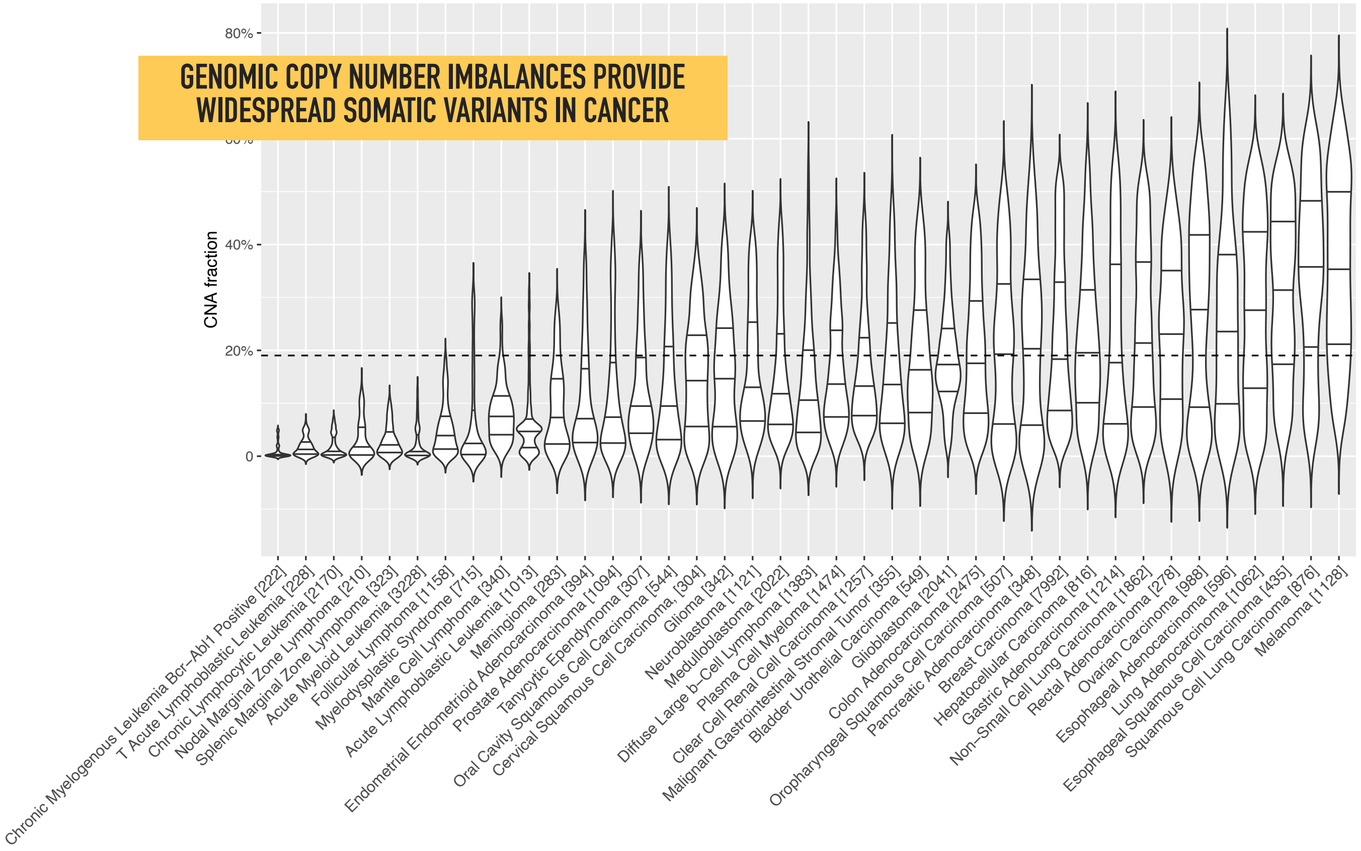


Probability of CNA with the frequency of gains (yellow) and losses (blue) across the chromosomes for 2'021 samples of Medulloblastoma, NOS, extracted from arrayMap database.

Quantifying Somatic Mutations In Cancer



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)

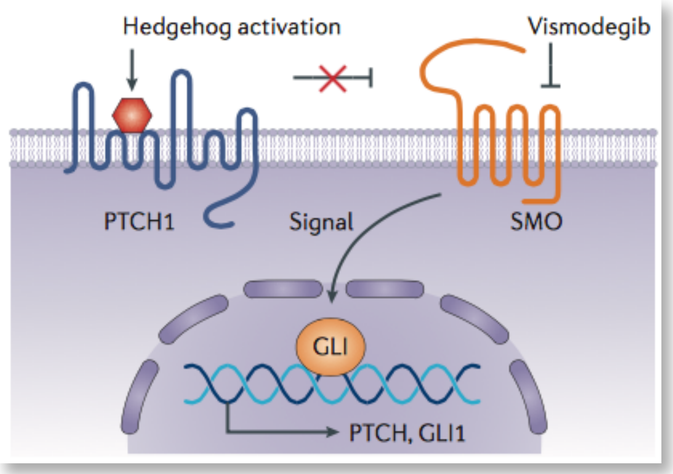
On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from progenetix.org

Rare CNV Events & Hidden Therapeutic Options?
Example: PTCH1 deletions in malignant melanomas

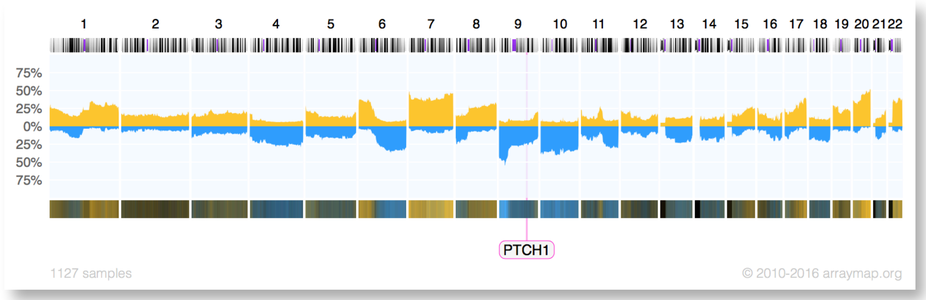
PTCH1 is a actionable tumor suppressor gene, which has been demonstrated in e.g. basaliomas and medulloblastomas

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

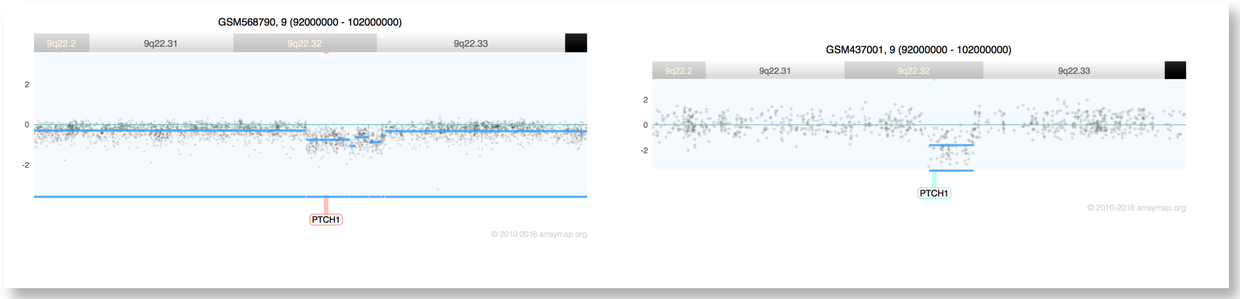
a current project addresses the focal involvement of all mapped genes, in >50'000 cancer genome profiles



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.

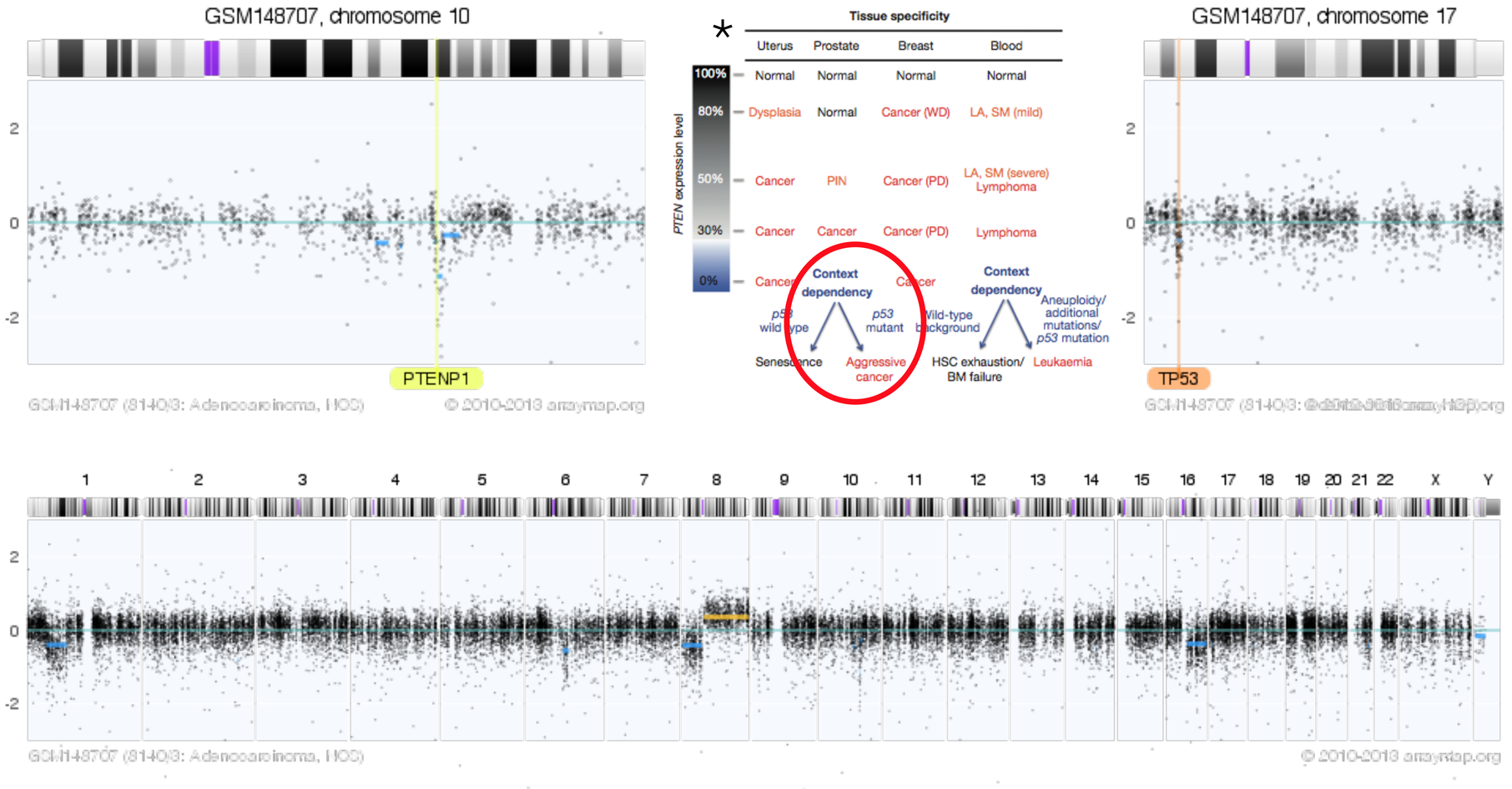


Summary of somatic copy number aberrations from the analysis of 1127 genome profiles of malignant melanomas, collected in our arraymap.org cancer genome resource. While PTCH1 does not represent a deletion hotspot, the genomic locus is part of larger deletions in ~25% of melanoma samples.



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.

Gene dosage phenomena beyond simple on/off effects



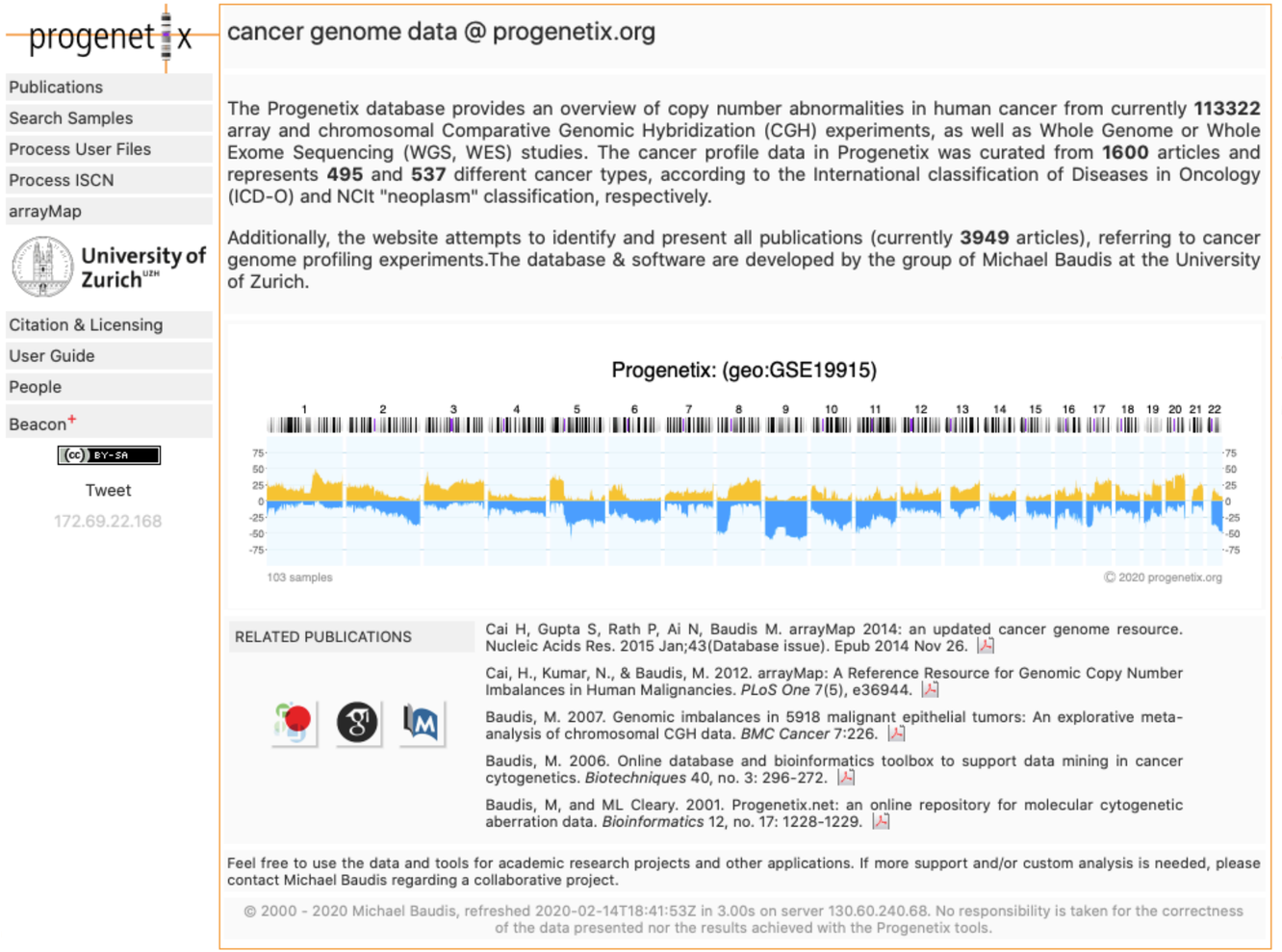
Combined heterozygous deletions involving *PTEN* and TP53 loci in a case of prostate adenocarcinoma
(GSM148707, PMID 17875689, Lapointe *et al.*, CancRes 2007)

* 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

Progenetix - Reference Resource for Oncogenomic Profiling Data

- launched in 2001 as progenetix.net with 999 samples (September 2001)
- curated CNV data from chromosomal CGH studies
- now containing >110000 single sample CNV tracks (90'807 cancer) from ~1600 publications
- **aCGH**, cCGH, WES, WGS
- additionally tracking and annotating of publications reporting compatible original data (more than 3200 articles as of 2019)



arrayMap

Accessing Probe-Level Genomic Array Data in Cancer

arrayMap

- Search Samples
- Search Publications
- Progenetix

University of Zurich[UZH]

- Citation & Licensing
- User Guide
- People
- Beacon+

CC BY-SA

Tweet

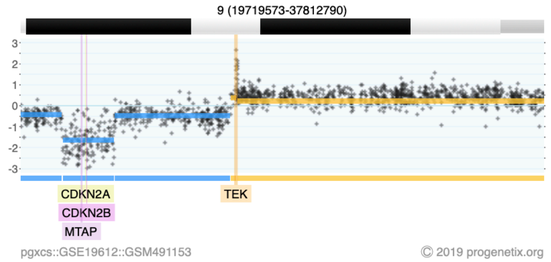
162.158.150.56

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.

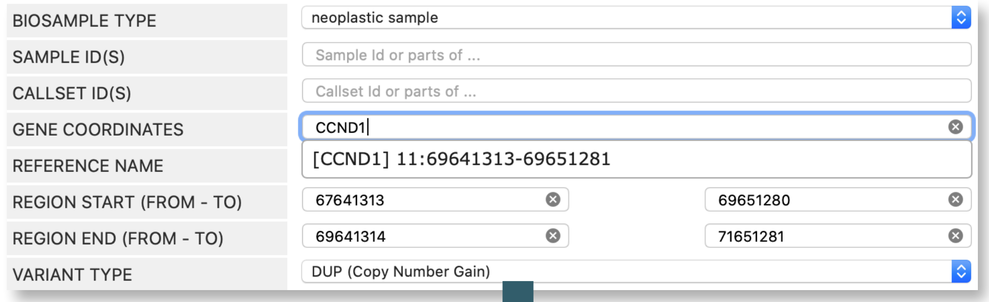
UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	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

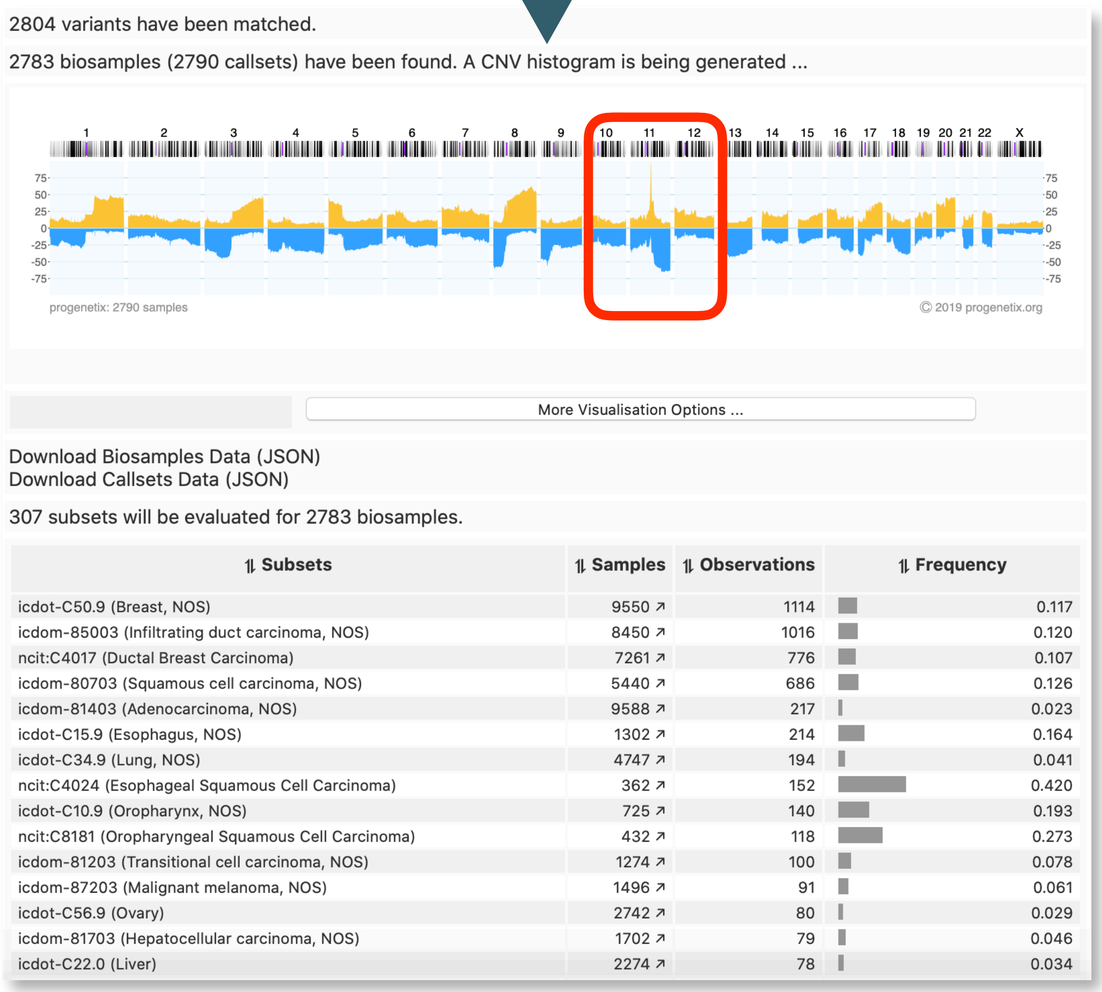
arrayMap

Progenetix - Reference Resource for Oncogenomic Profiling Data

- Progenetix is based on the single-sample CNV tracks of cancer samples from 402/469 (ICD-O/NCIt) diagnostic categories
- typical applications include
  - reference CNV patterns in given diagnoses (e.g. "does my analysis match the diagnosis/ prediction")
  - target gene entity mapping (e.g. "in which tumour type is this gene frequently gained/ lost?")

progenetix





Subsets	Samples	Observations	Frequency
icdot-C50.9 (Breast, NOS)	9550	1114	0.117
icdom-85003 (Infiltrating duct carcinoma, NOS)	8450	1016	0.120
ncit:C4017 (Ductal Breast Carcinoma)	7261	776	0.107
icdom-80703 (Squamous cell carcinoma, NOS)	5440	686	0.126
icdom-81403 (Adenocarcinoma, NOS)	9588	217	0.023
icdot-C15.9 (Esophagus, NOS)	1302	214	0.164
icdot-C34.9 (Lung, NOS)	4747	194	0.041
ncit:C4024 (Esophageal Squamous Cell Carcinoma)	362	152	0.420
icdot-C10.9 (Oropharynx, NOS)	725	140	0.193
ncit:C8181 (Oropharyngeal Squamous Cell Carcinoma)	432	118	0.273
icdom-81203 (Transitional cell carcinoma, NOS)	1274	100	0.078
icdom-87203 (Malignant melanoma, NOS)	1496	91	0.061
icdot-C56.9 (Ovary)	2742	80	0.029
icdom-81703 (Hepatocellular carcinoma, NOS)	1702	79	0.046
icdot-C22.0 (Liver)	2274	78	0.034

"Emerging" Progenetix API

- the Progenetix API provides access to a growing number of database features
  - biosample data listings
  - code translations (ICD-O <-> NCIt)
  - publication data

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 Beacon BeaconPlus JavaScript Perl article code documentation

API::api.cgi Code Documentation

API syntax

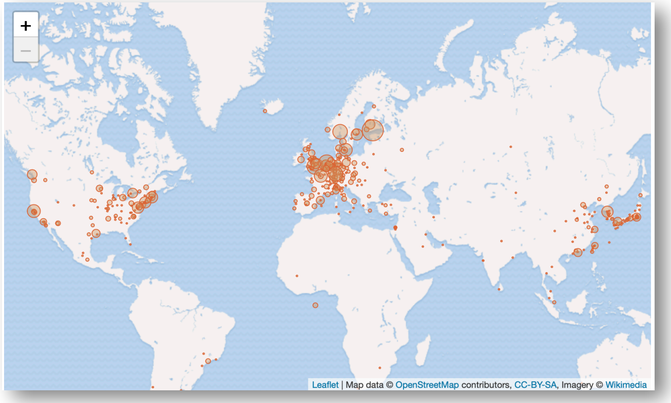
The query elements are ordered in the URL

1. `api`
   - fixed, required; i.e. the request has to start with `/api/`
2. database (parameter `apidb`)
   - mostly “progenetix”
3. collection (`apiscope`)
4. method (`apimethod`)
   - see examples/documentation below
5. filters (`filters`)
   - essentially query parameters in a simplified format
   - comma-concatenated
   - can also be omitted for query string
6. output parameter (`apioutput`)
   - optional
7. query string
   - optional
   - can be used for any parameter; e.g. a query can be formatted completely as standard query string:
     - `progenetix.org/api/?apidb=progenetix&apiscope=publications ...`


```
{
  "label": "Teratoma, malignant, NOS",
  "id": "icdom-90803",
  "count": 70,
  "child_terms": [
    "icdom-90803"
  ]
},
{
  "label": "Adenocarcinoma with neuroendocrine differentiation",
  "id": "icdom-85743",
  "count": 1,
  "child_terms": [
    "icdom-85743"
  ]
},
{
  "label": "Complex epithelial neoplasms",
  "id": "icdom-856-857+",
  "child_terms": [
    "icdom-85623",
    "icdom-85763",
    "icdom-85753",
    "icdom-85603",
    "icdom-85743"
  ],
  "count": 133
},
```

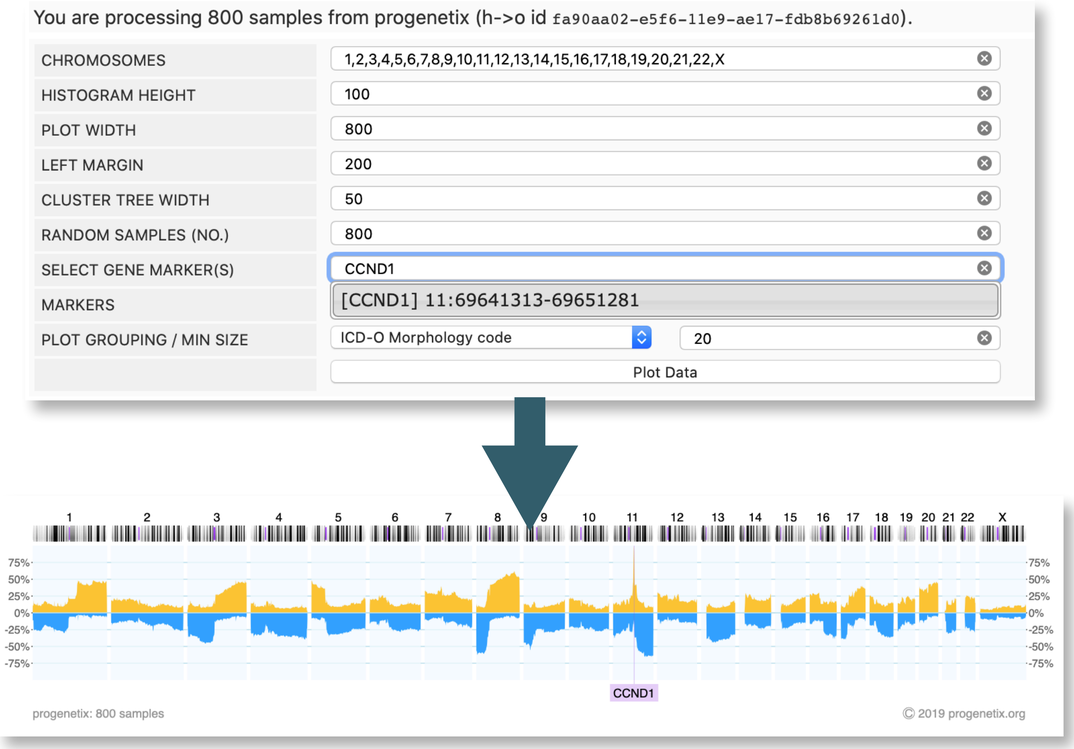


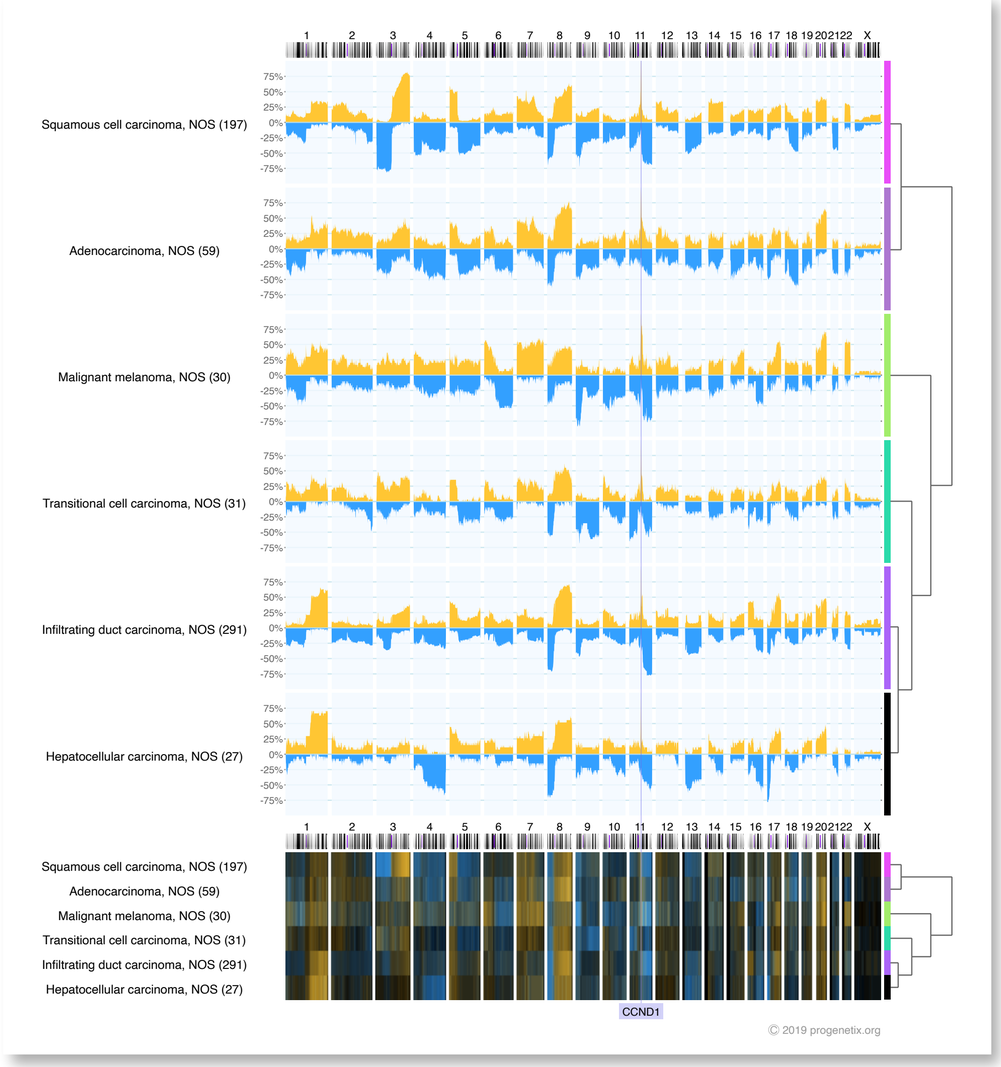
```
        id      description
cellosaurus::id cellosaurus
material::id    material::
1       PGX_AM_BS_9788624_(
P74Y                    Yar
2       PGX_AM_BS_9788624_(
P70Y                    Yar
3       PGX_AM_BS_9788624_(
P66Y                    Yar
4       PGX_AM_BS_9788624_(
P66Y                    Yar
5       PGX_AM_BS_9788624_(
P48Y                    Yar
6       PGX_AM_BS_9788624_(
P37Y                    Yar
7       PGX_AM_BS_9788624_(
P70Y                    Yar
8       PGX_AM_BS_9788624_(
P57Y                    Yar
9       PGX_AM_BS_9788624_(
P67Y                    Yar
10      PGX_AM_BS_9788624_(
P65Y                    Yar
11      PGX_AM_BS_9788624_GB-har-11     glioblastoma    9440/3  Glioblastoma, NOS       C71.0   cerebrum
P58Y                    Yamaguchi-shi   Japan   34.19   131.47  EFO:0009656     neoplastic sample       0.213

id          ncit::label     pubmed::id      pubmed::label
s city      country latitude        longitude
rum         C3058   Glioblastoma    9788624
        0.202
rum         C3058   Glioblastoma    9788624
        0.135
rum         C3058   Glioblastoma    9788624
        0.164
rum         C3058   Glioblastoma    9788624
        0.146
rum         C3058   Glioblastoma    9788624
        0.223
rum         C3058   Glioblastoma    9788624
        0.136
rum         C3058   Glioblastoma    9788624
        0.108
rum         C3058   Glioblastoma    9788624
        0.201
rum         C3058   Glioblastoma    9788624
        0.144
rum         C3058   Glioblastoma    9788624
        0.142
rum         C3058   Glioblastoma    9788624
```

Progenetix - Reference Resource for Oncogenomic Profiling Data

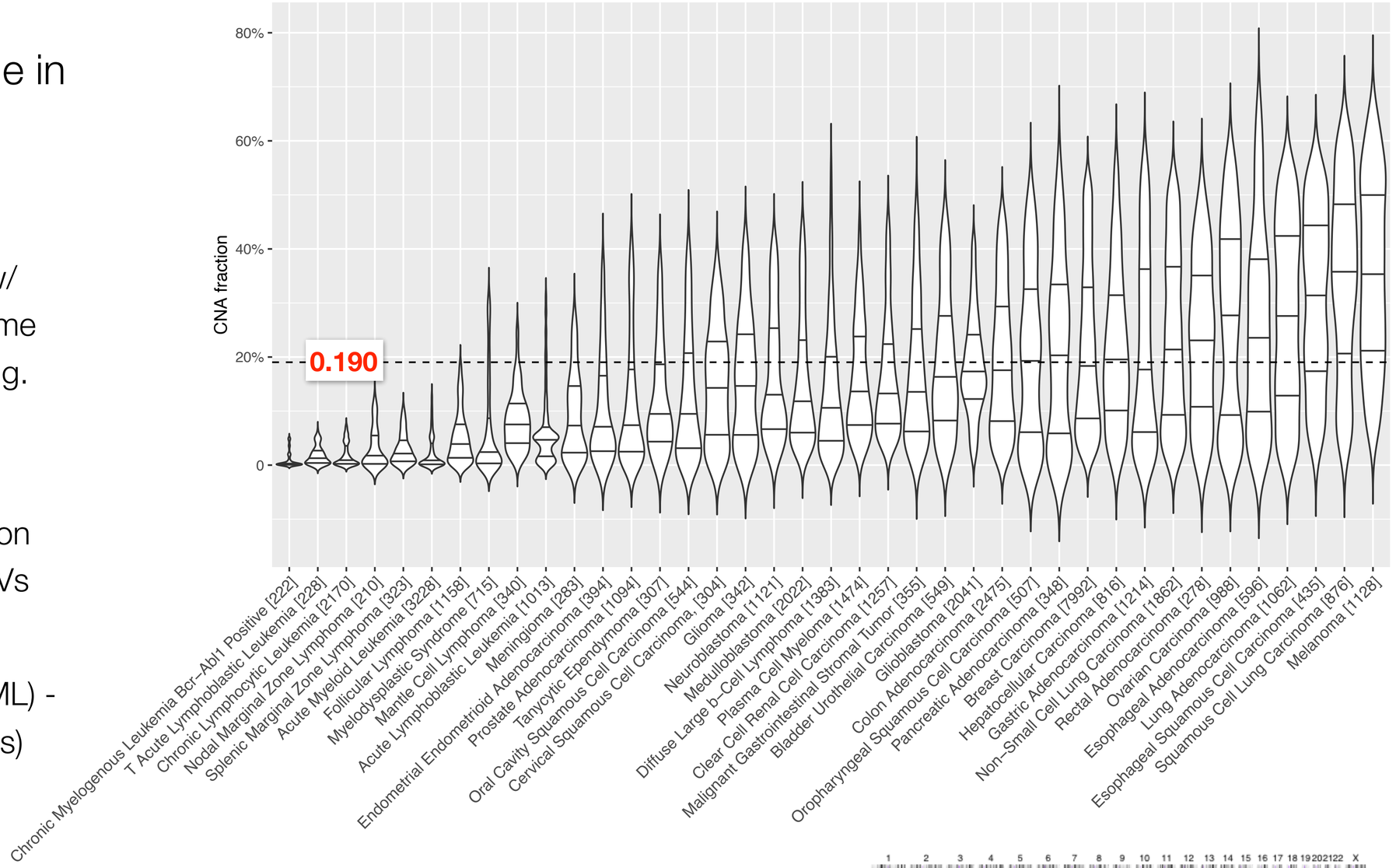
- Group histogram and heatmap representation of CNV profiles by external labels (disease codes, publications ...)



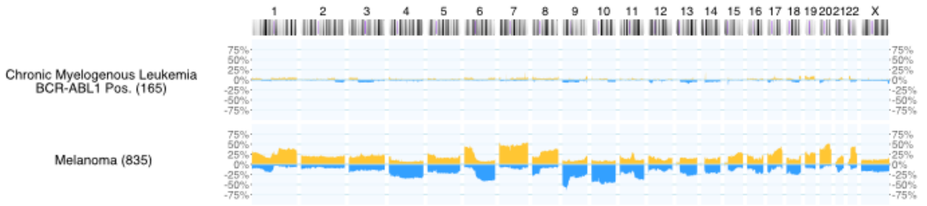


Genome CNV coverage in Cancer Classes

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



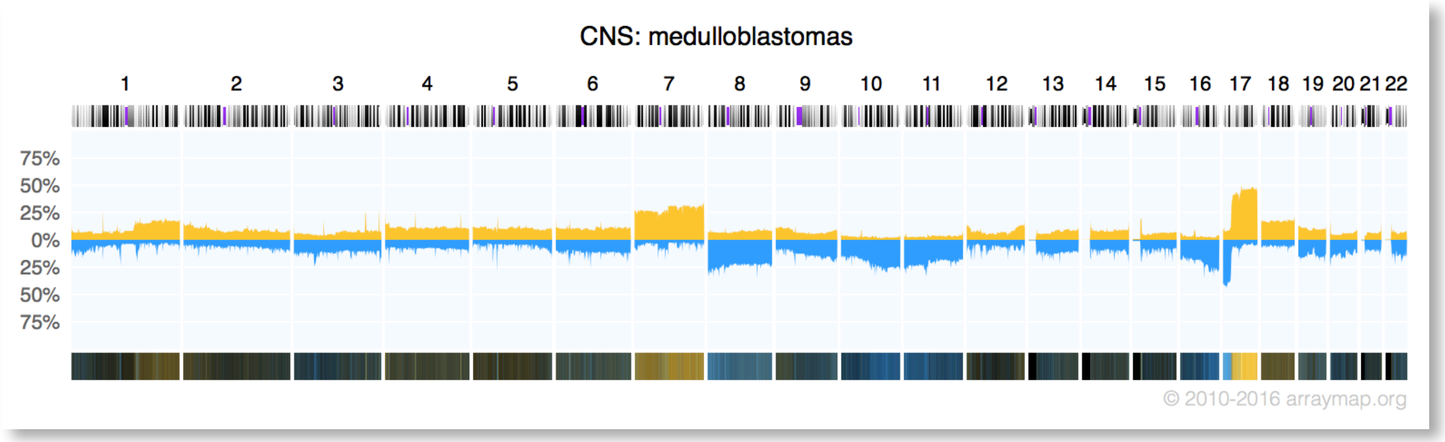
Lowest / Highest CNV fractions =>

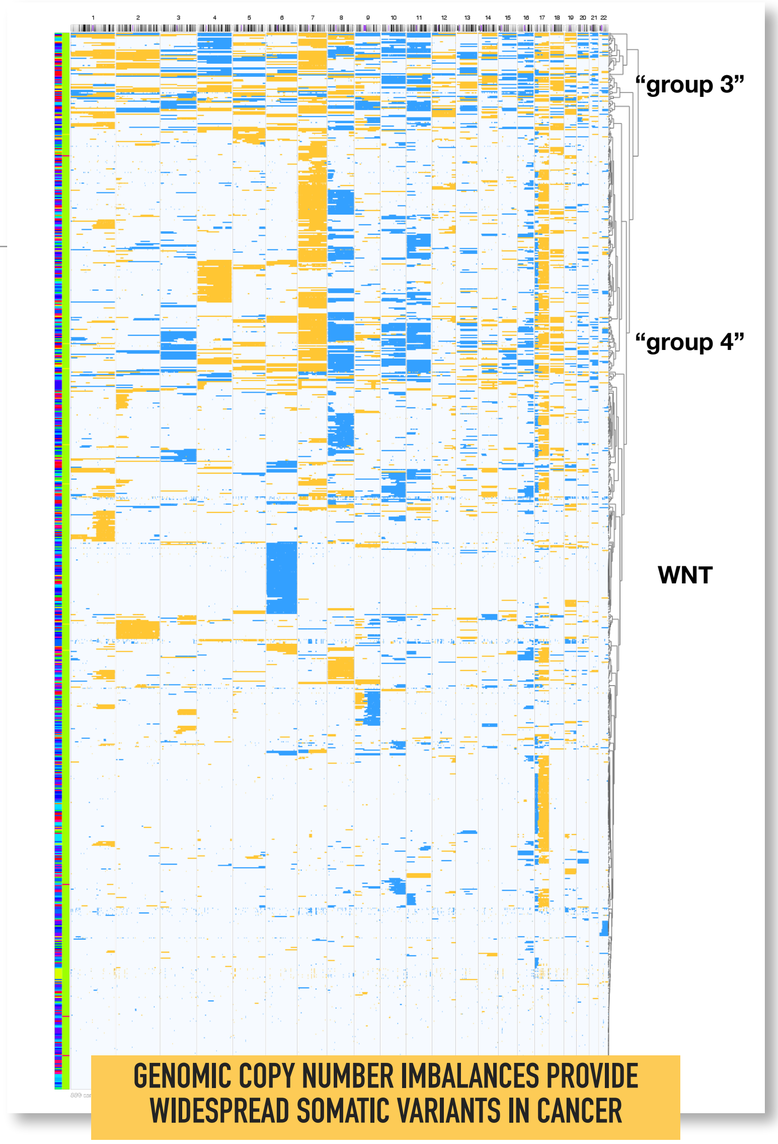


Somatic CNVs In Cancer: Patterns

Many tumor types express **recurrent mutation patterns**

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



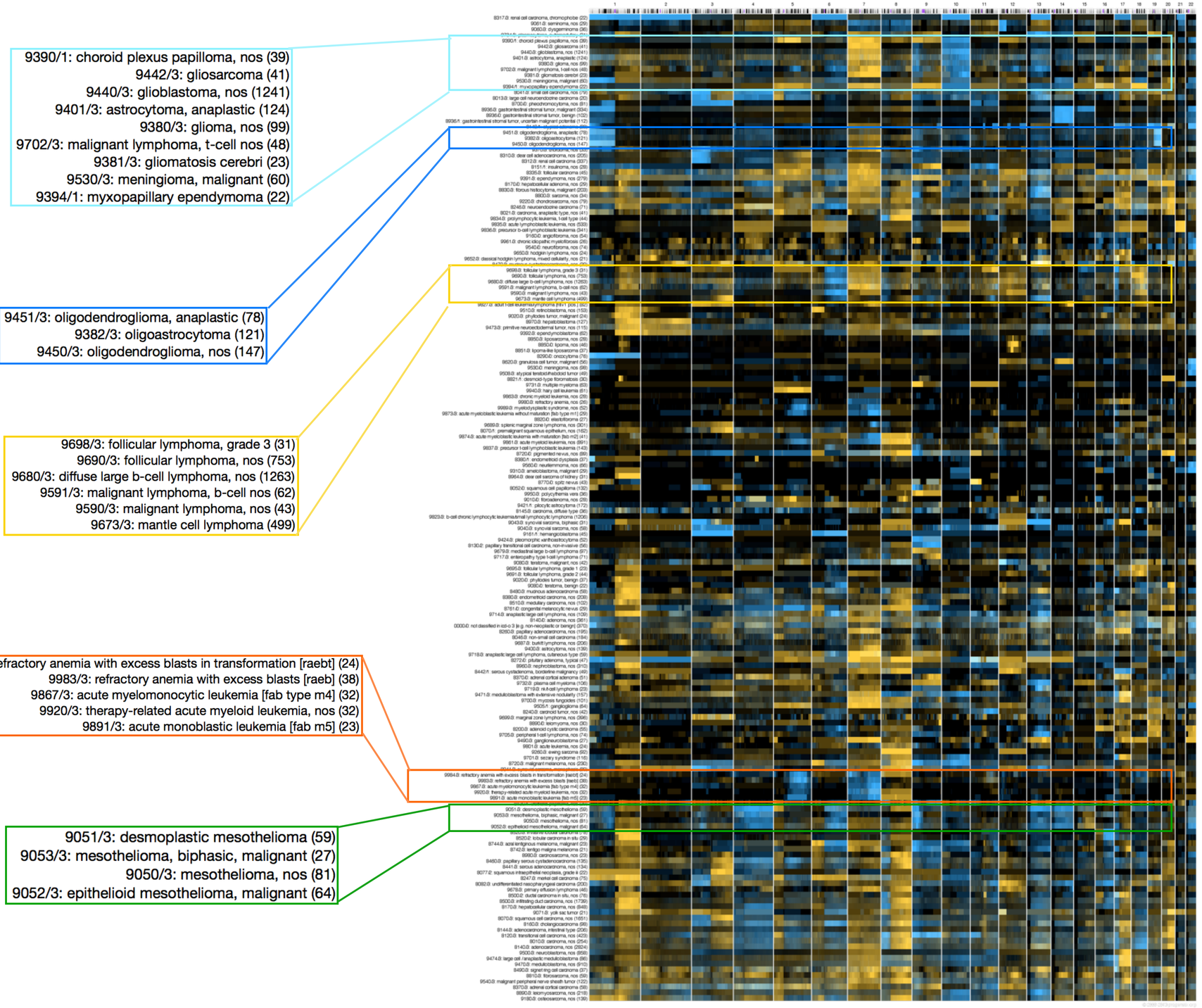


A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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

arrayMap

Somatic Mutations In Cancer: Patterns Making the case for genomic classifications

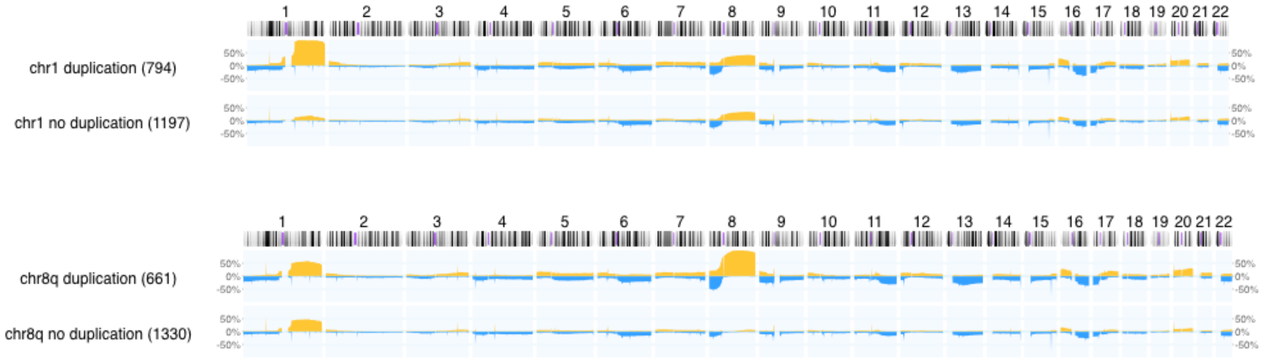
Some related cancer entities show similar copy number profiles



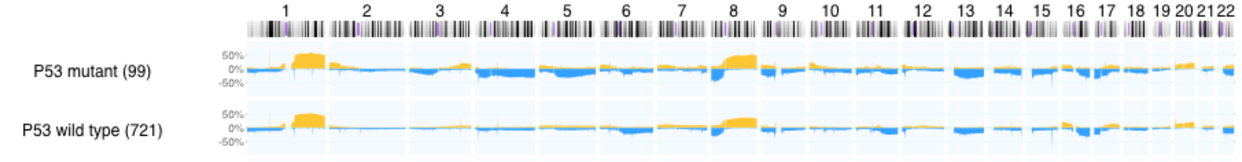
arrayMap

Interpret CNV by Association to Molecular/Clinical Information

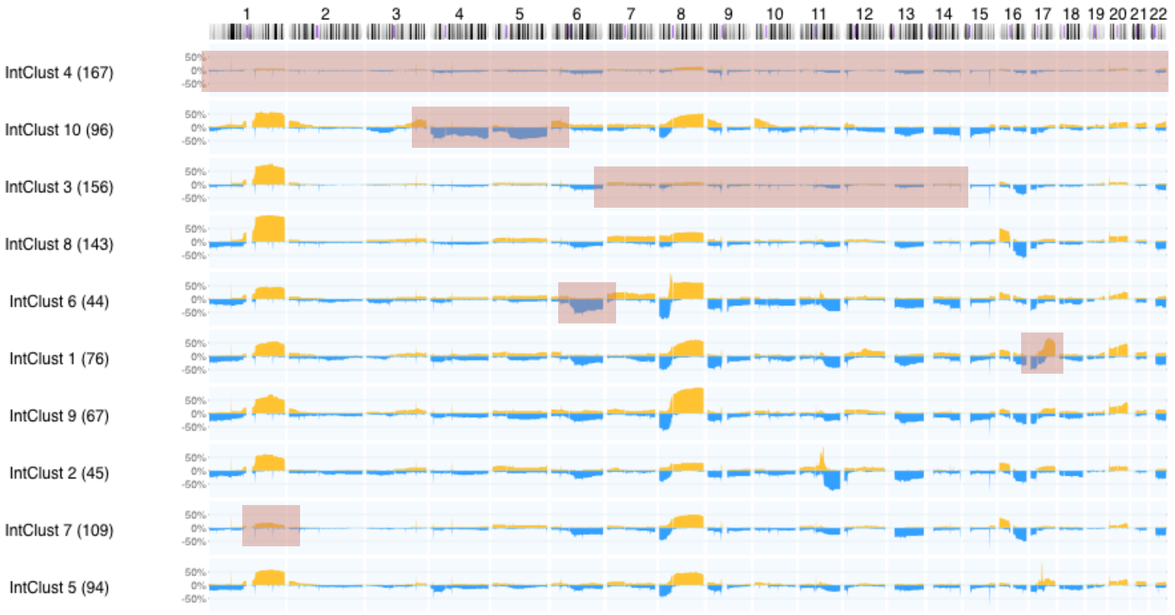
Chr1q and chr8q duplication unlinked



P53 mutant slightly highter CNV



Some IntClust groups define Distinctive CNV patterns



METABRIC

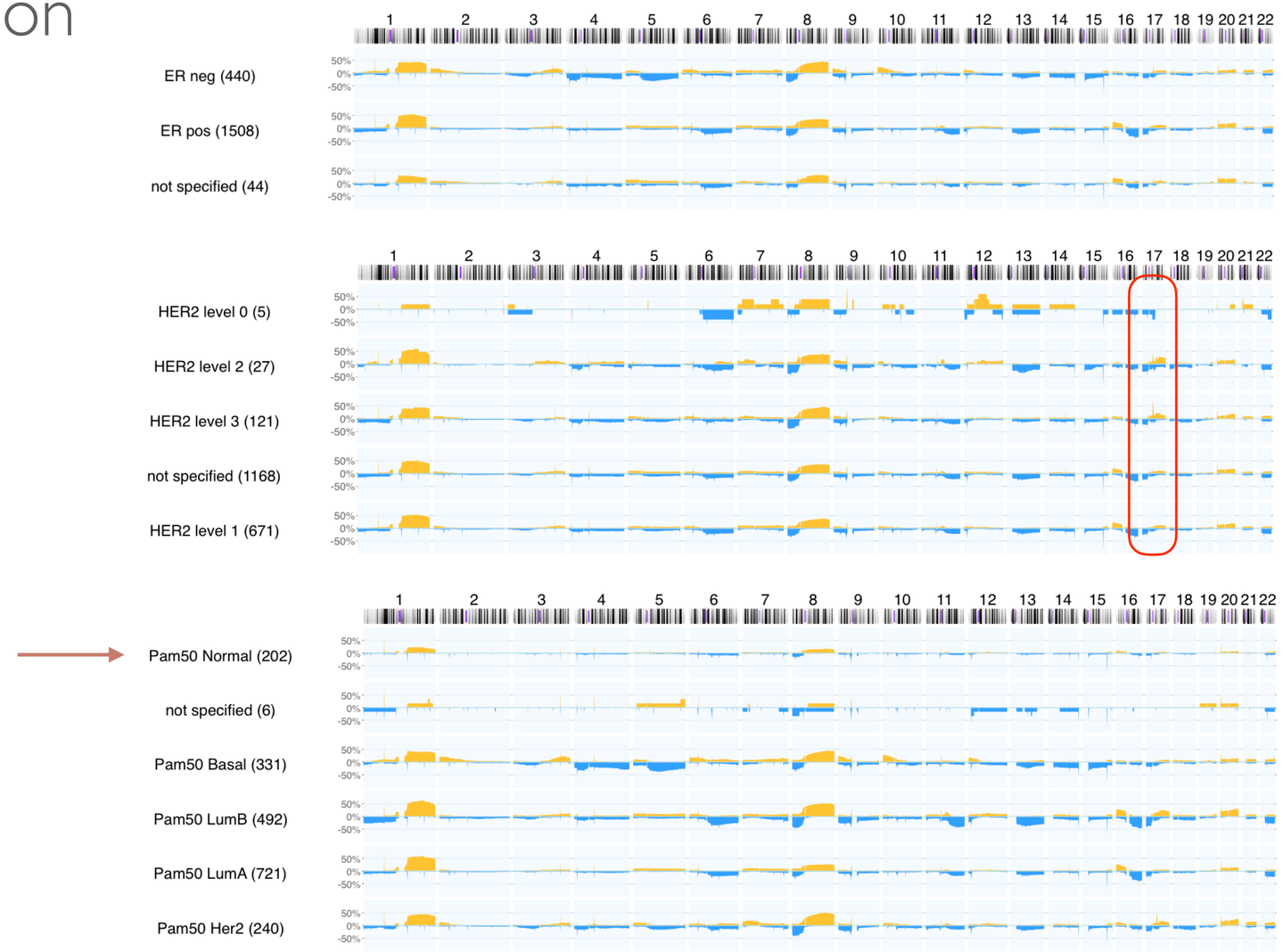
Data source: METABRIC 1992 breast cancer samples, hierarchical clustering, tree not shown

Interpret CNV by Association to Molecular/Clinical Information

ER status

HER2 level

Pam50 category



METABRIC

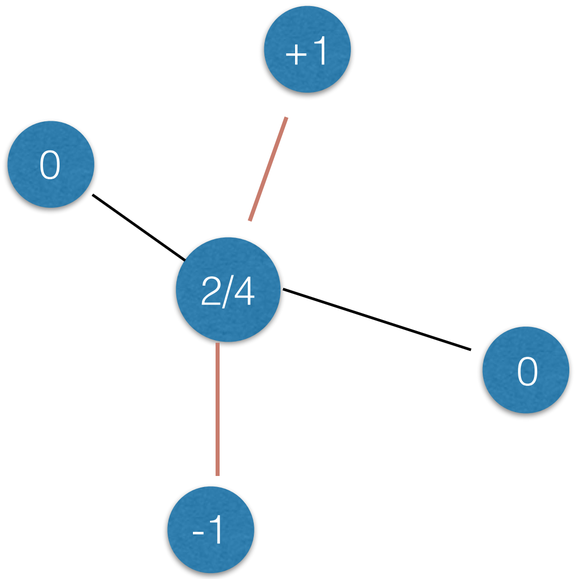
Data source: METABRIC 1992 breast cancer samples, hierarchical clustering, tree not shown

STRING

Interpret Gene-Level CNV With Protein Networks

Highly Connected Genes Have Similar CNV Patterns as Canonical Driver Genes?

Gene CNV score



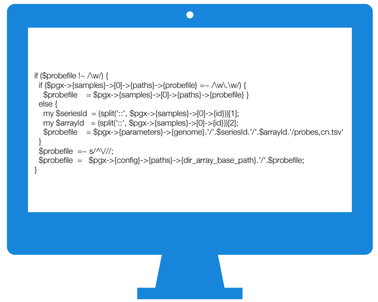
- Incorporate protein network
- Define gene CNV score by partner CNV status
- Driver genes distinguish from rest
- Correct for connectivity → effect is gone

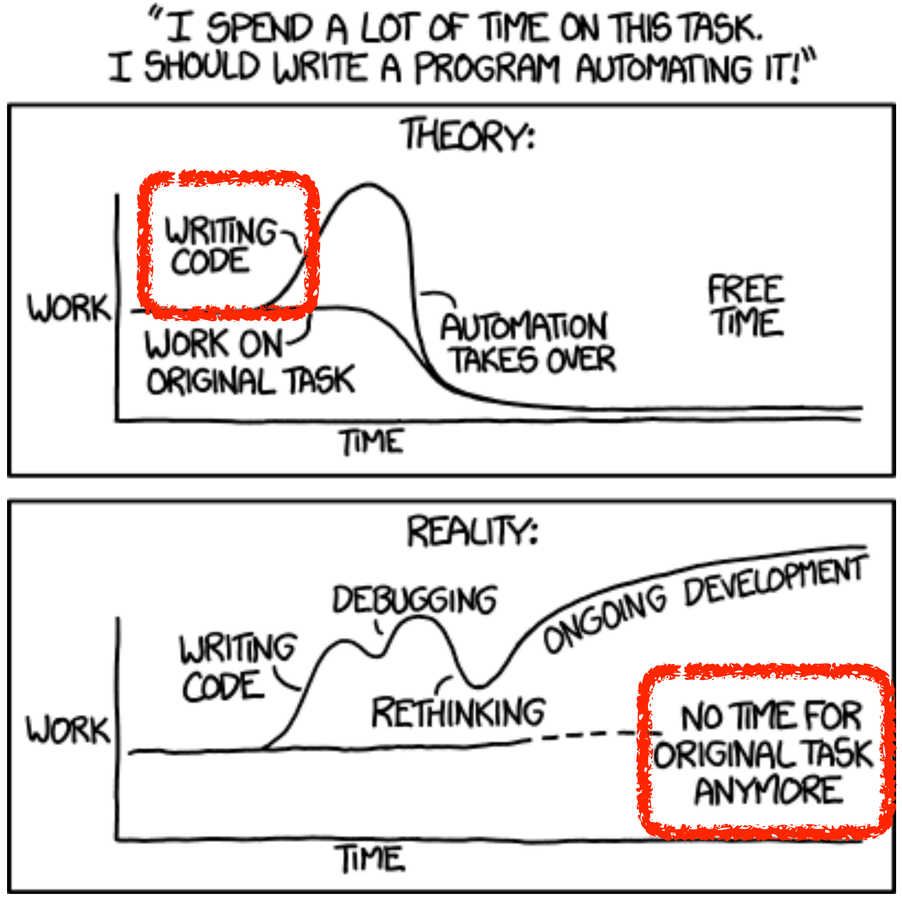
METABRIC

{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;
}
```

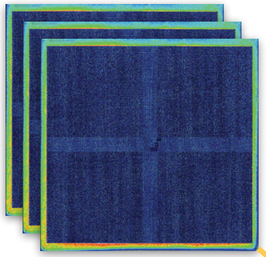
{bio_informatics_science}

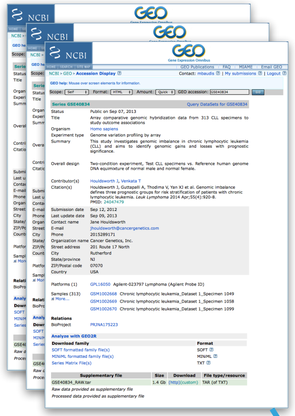


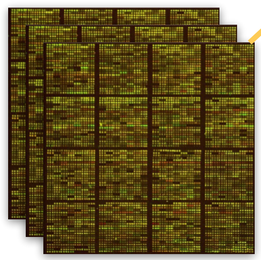


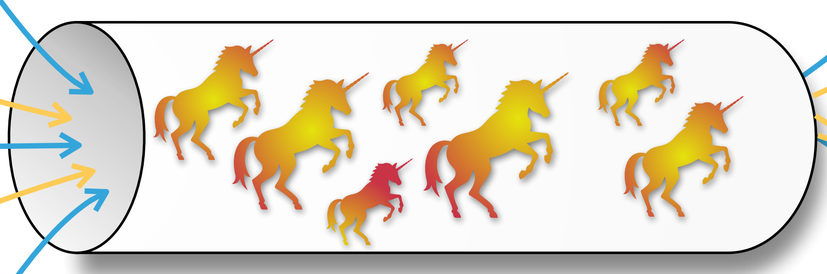
Randall Munroe - XKCD
https://xkcd.com/1319/

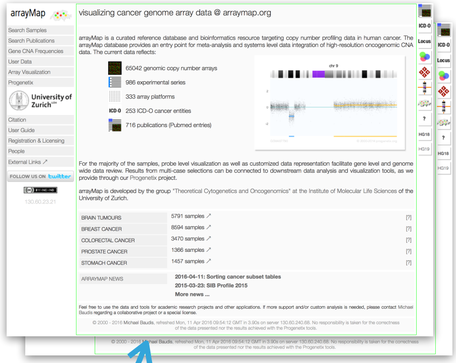
DATA PIPELINE

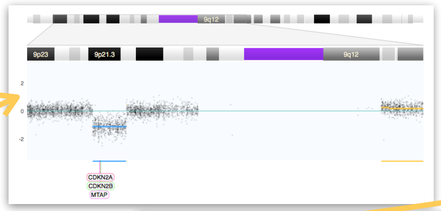


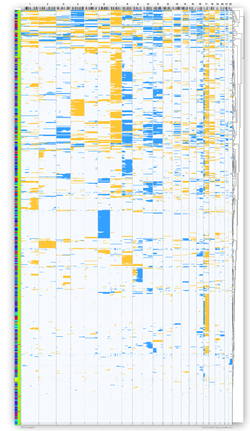


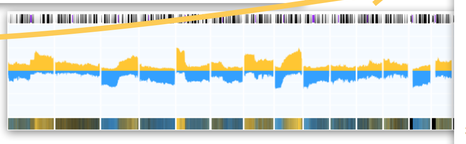


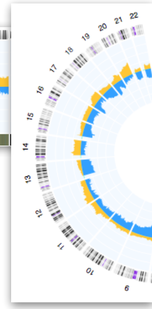


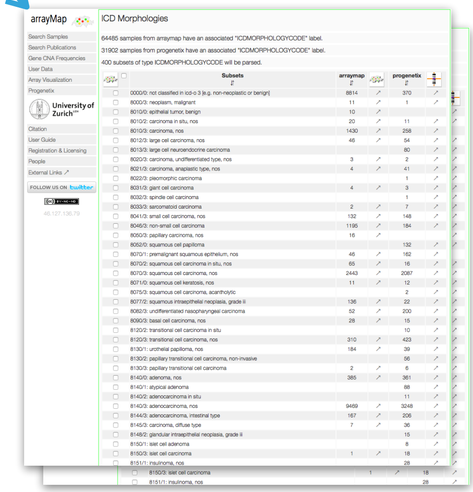








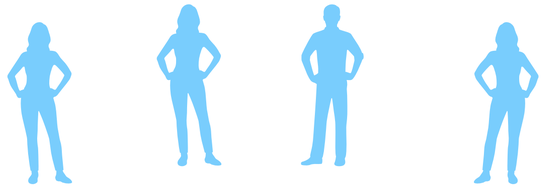




arrayMap

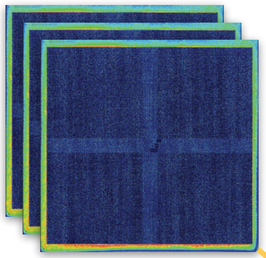
progenetix

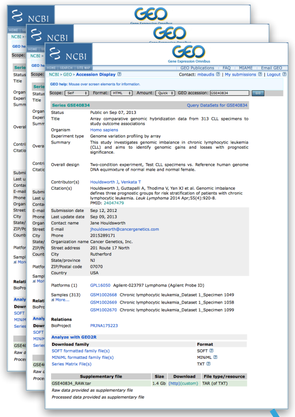
DATA PIPELINE

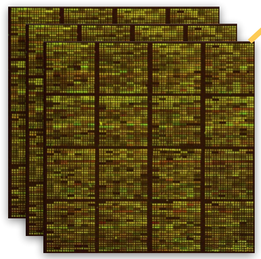


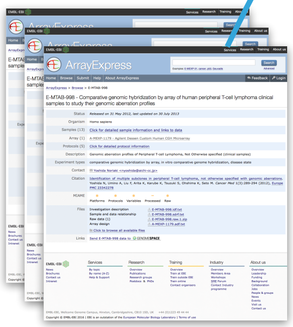
BIOCURATION

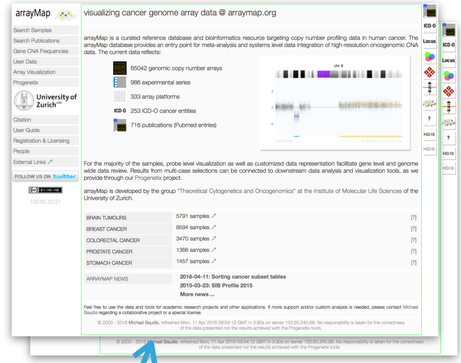
BIOINFORMATICS

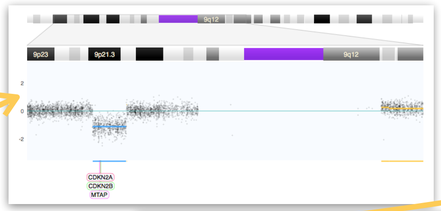


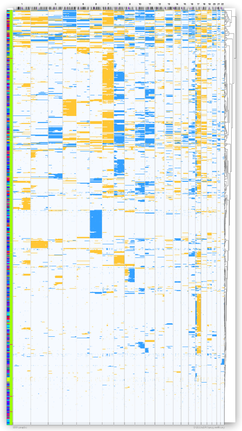


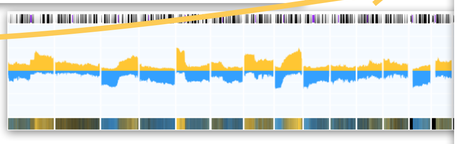


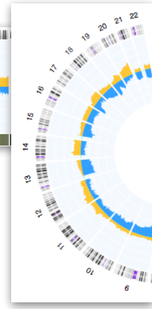


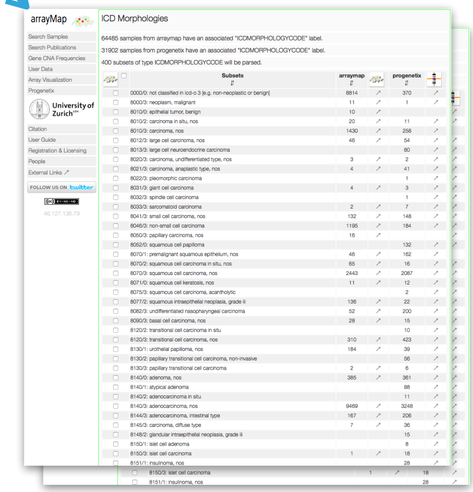












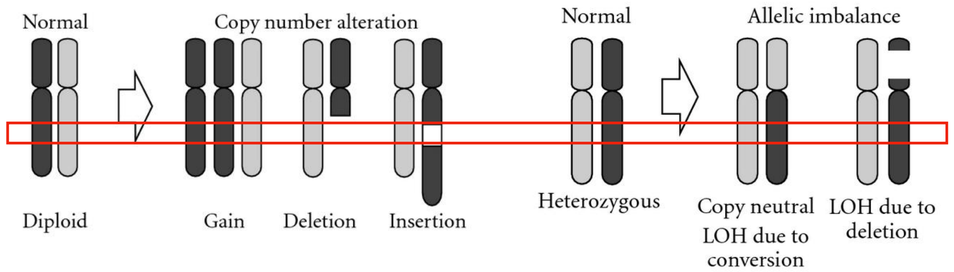
arrayMap

progenetix

Signal noise in copy number profiling

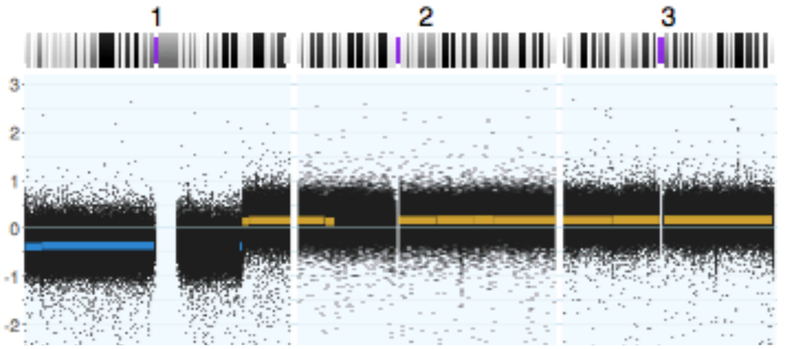
- Actual DNA copy numbers per cell/clone are integers.
- Cancer copy number profiles frequently cannot be interpreted with a simple "integer" model.
- CN profiles difficult to compare:
  - Every sample has its own noise level
  - Every sample has its own signal scale
  - Ambiguity (aneuploidy, subclones)

Regardless mutation types, the copy number of DNA segment should always be an integer

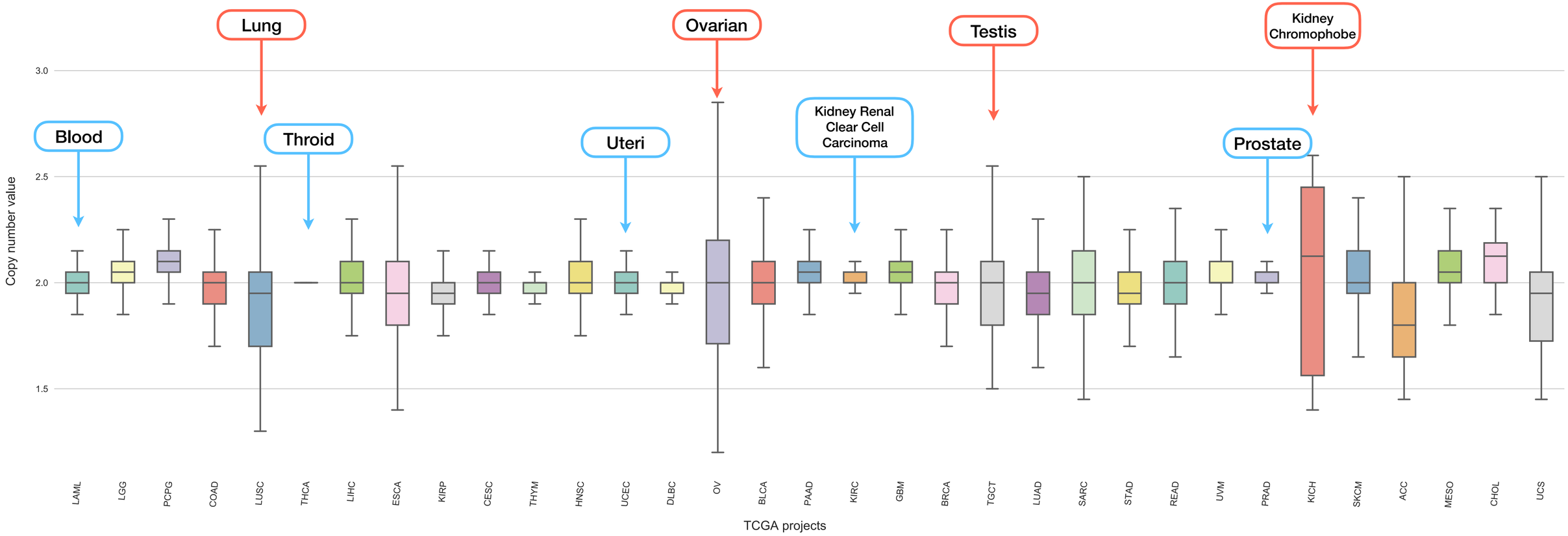


Source: Gibb, Ewan A., et al. "Deciphering squamous cell carcinoma using multidimensional genomic approaches." *Journal of skin cancer* 2011 (2011).

Copy number profile of a gastrointestinal stromal tumor (GSM2443442)



Baseline deviation in TCGA data



Bo Gao

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Minimum Error Calibration and Normalization for Genomic Copy Number Analysis

Bo Gao, Michael Baudis

doi: https://doi.org/10.1101/720854

Mecan4CNA:

Minimum Error Calibration and Normalization for Copy Number Analysis

Goal

Calibrate and normalize copy number datasets

Key feature

Without estimating true copy number levels of each sample

Modeling and deduction

$$x_i = (aN_i + bT_i + \sum_{k=1}^{n} c_i^n S_i^n + \sum_{h=1}^{m} E_i^m) \prod_{j=1}^{l} (1 + e_j)$$

$$= (aN_i + bT_i + cS_i + E_i)(1 + e)$$

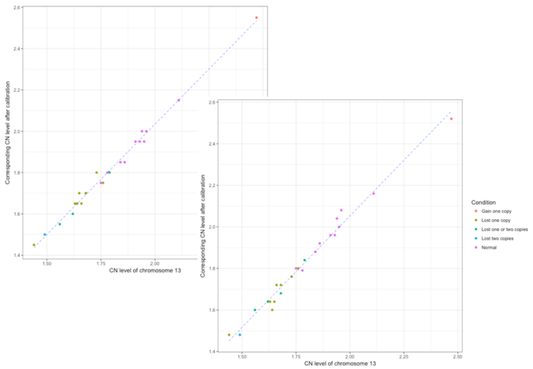
$$= aN_i + bT_i + cS_i + E_i$$

$$R(i,j,k) = \frac{D(i,k)}{D(i,j)}$$

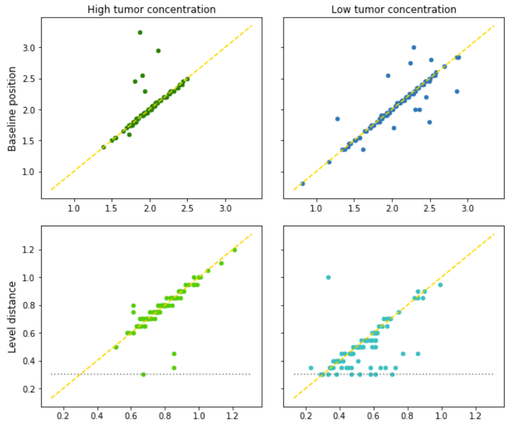
$$R(i,j,k) = \frac{b(T_i - T_k) + c(S_i - cS_k) + E_{i,k}}{b(T_i - T_j) + c(S_i - cS_j) + E_{i,j}}$$

$$= \frac{T_i - T_k}{T_i - T_j}\left(1 + \frac{c(S_i - cS_j) + E_{i,j}}{b(T_i - T_j) + c(S_i - cS_j) + E_{i,j}}\right) + \frac{c(S_i - cS_k) + E_{i,k}}{b(T_i - T_j) + c(S_i - cS_j) + E_{i,j}}$$

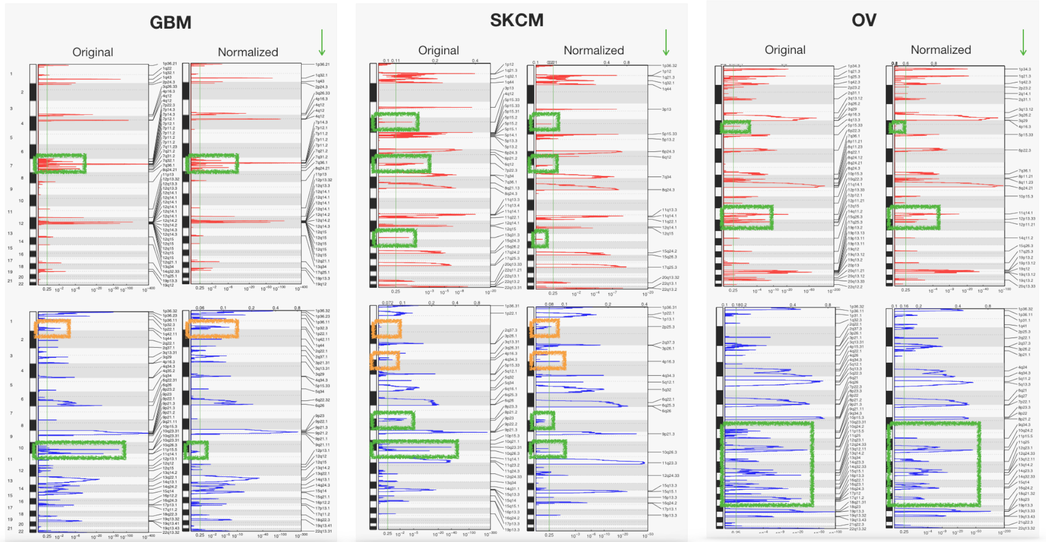
Benchmarking against karyotyping and ABSOLUTE



Benchmarking on simulation data



Application on GISTIC analysis of TCGA data



Bo Gao

intCNA on large dataset

intCNA - variable and parameter learning

- Variational EM (Expectation-Maximisation) - mean field variational inference
  - The E-step calculates the posterior probabilities of the hidden states with current model parameters fixed
  - The M-step maximises the expected log-likelihood of the observations as a function of the model parameters using these posterior probabilities

$$Q(S) = \prod_{t=1}^{T} \prod_{m=1}^{M} q_{t,m}(S_t^{(m)}) \tag{9}$$

$$\log q_{t,m}^{\star}(S_t^{(m)}) = \mathbb{E}_{S \setminus S_t^{(m)}}[\log P(S, Y; d)] + \text{const}$$

- Update variables and parameters
  - Analytical solution - constructing Lagrangian multiplier for $a_{ij}^{(m)}$ and $b_{ij}^{(m)}$
  - Numerical solution - gradient descent for $w^{(m)}$ and θ

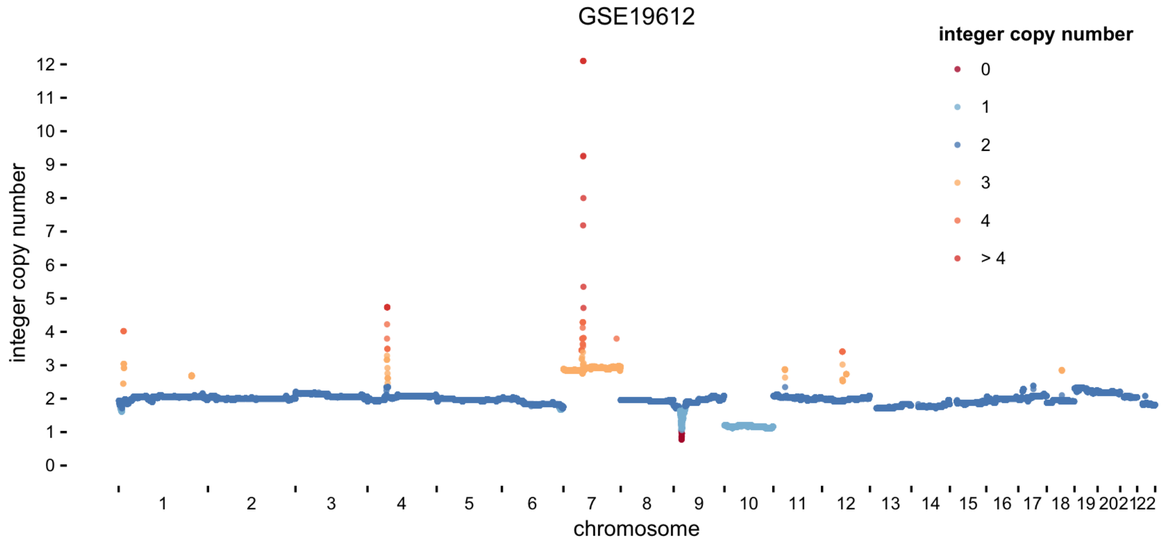


Figure 6: *Mean values of integer copy number for 49 glioblastoma samples containing discernible CNAs from GSE19612 [13]. Recurrent CNAs including chromosome 7 gain, chromosome 10 loss, EGFR amplification at 7p11.2, PDGFRA amplification at 4q12, homozygous and heterozygous focal deletion of 9p21 where CDKN2A, CDKN2B and MTAP are located.*

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]

segmentLiftover:
A tool to re-map segmental genome data between reference genome editions

Challenge

1. Keep the integrity of copy number segments after Liftover.
2. 10% data lost from straight Liftover.
3. 1TB segment and probe data.

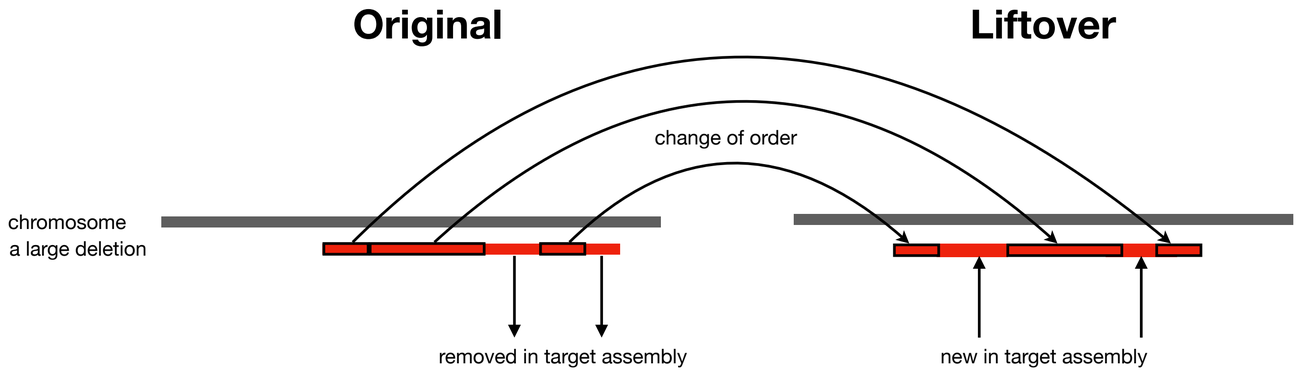
Solution

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

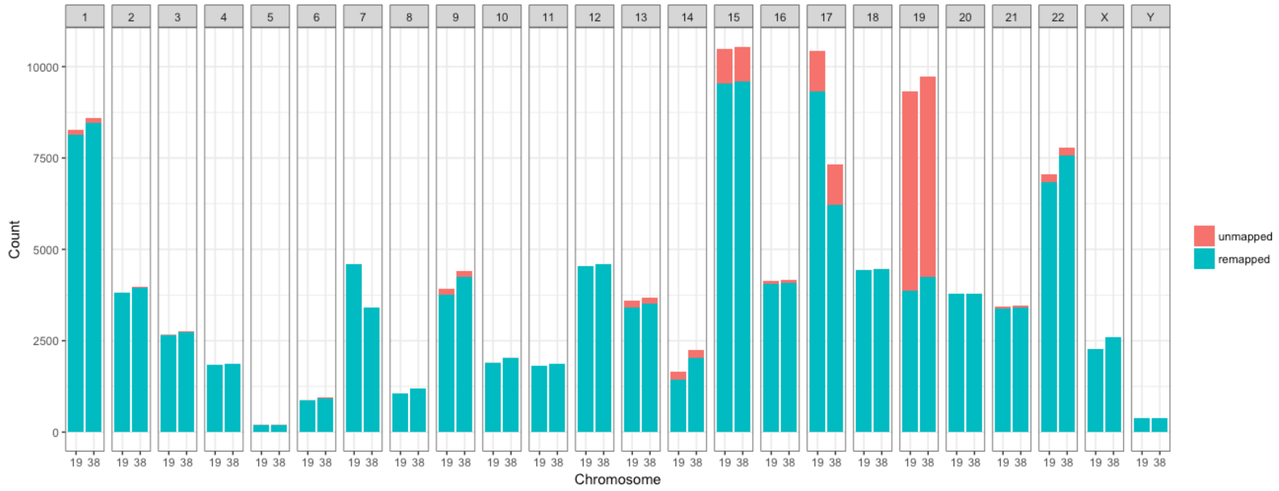
Results

1. Converts hg18 | hg19 | GRCh38
2. Processed 122,788 files, 26,164,205 segments and 28,941,899,671 probes in total
3. straight forward run > 1 week => x4 parallel processes <3 days
4. Reduced data loss: 10% => **0.1%**

The difficulties in copy number segment liftover



Results of segmentLiftover on our data

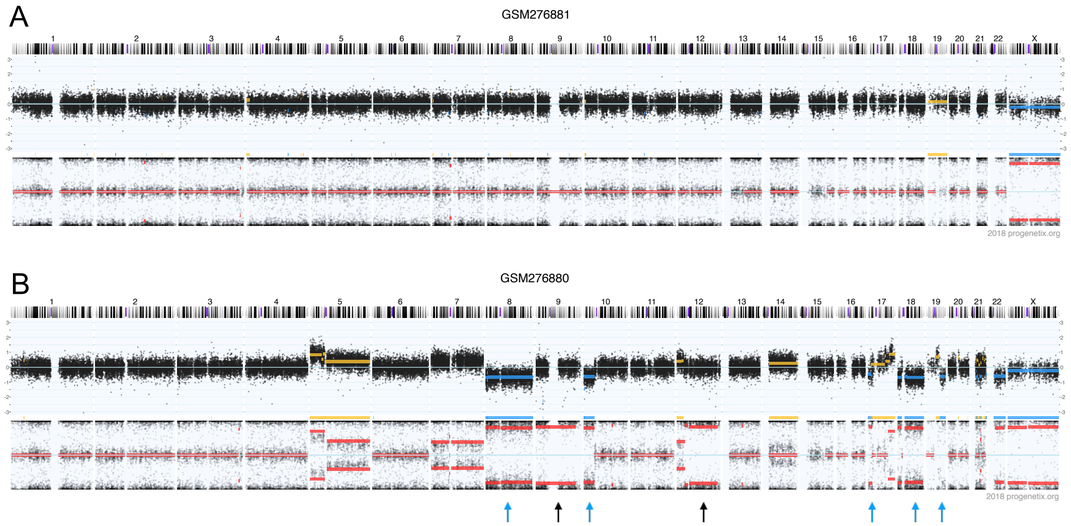


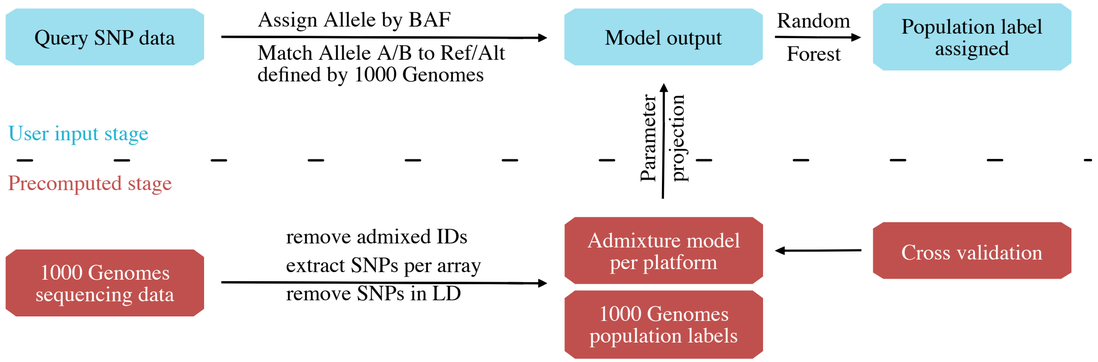
Population stratification in cancer samples based on SNP array data

- 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.

Enabling population assignment from cancer genomes with SNP2pop

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





arrayMap

Qingyao Huang

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

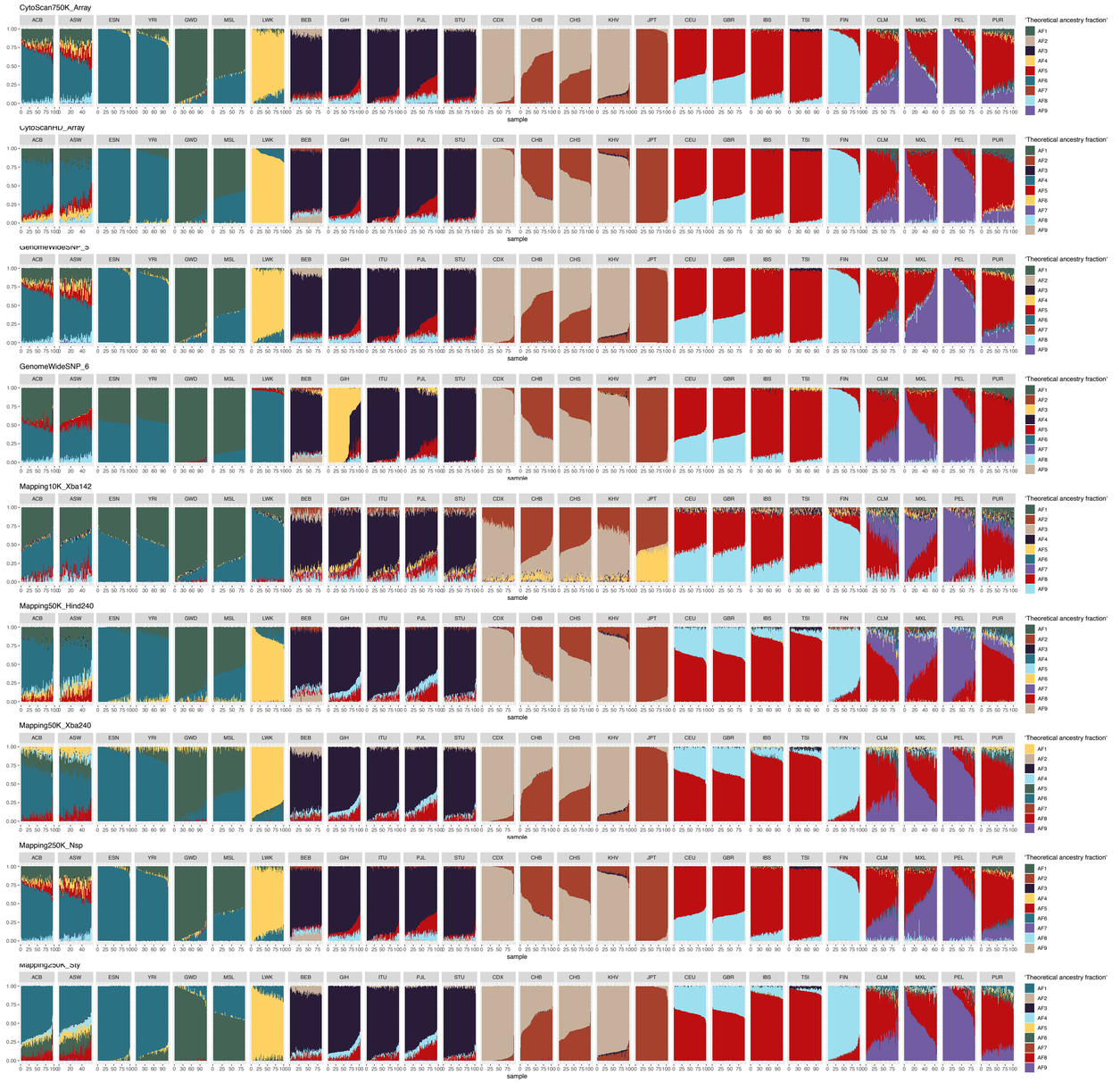


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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

arrayMap

Qingyao Huang

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

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:

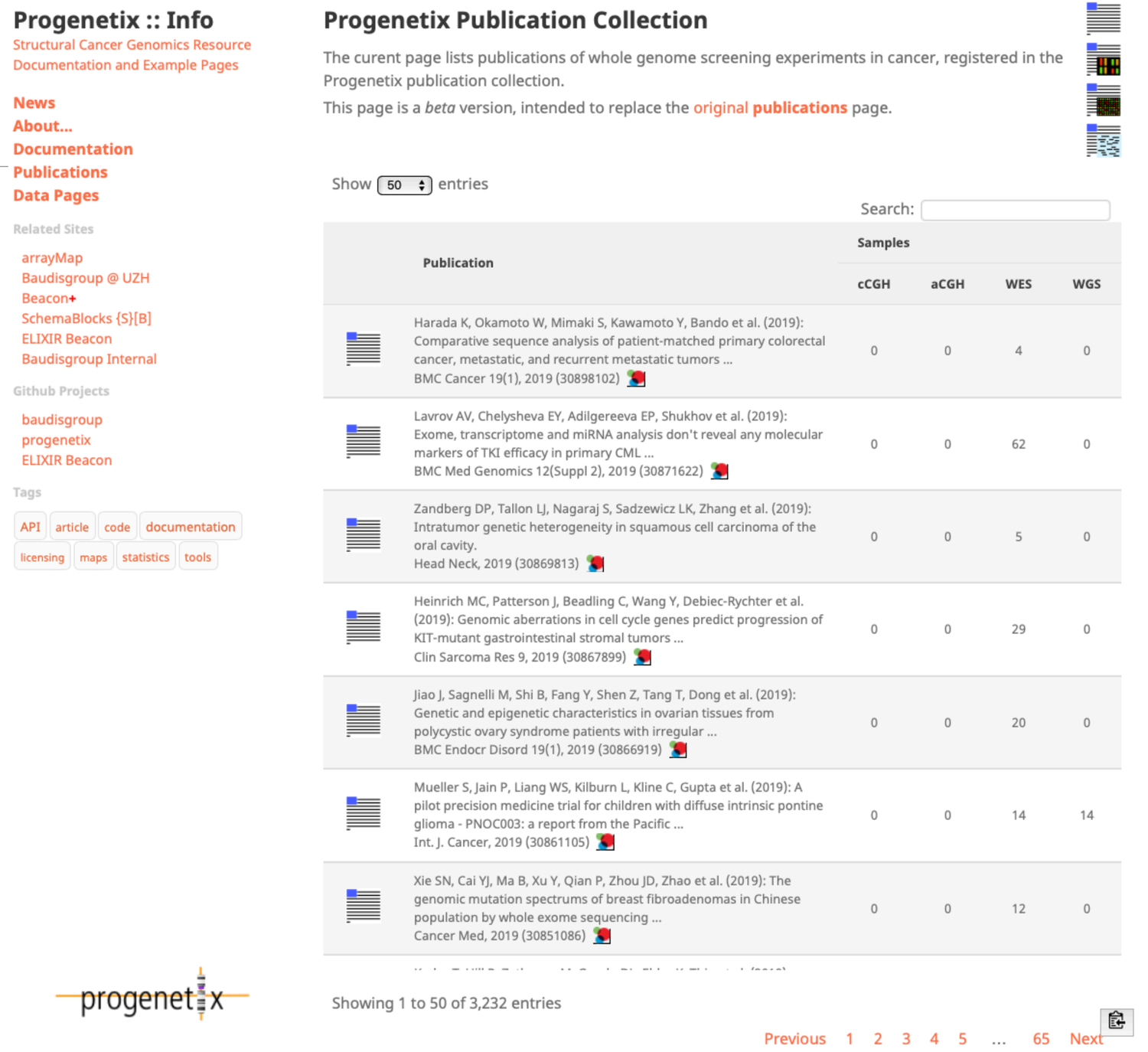
Publication	Samples			
	cCGH	aCGH	WES	WGS
Harada K, Okamoto W, Mimaki S, Kawamoto Y, Bando et al. (2019): Comparative sequence analysis of patient-matched primary colorectal cancer, metastatic, and recurrent metastatic tumors ... BMC Cancer 19(1), 2019 (30898102)	0	0	4	0
Lavrov AV, Chelysheva EY, Adilgereeva 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

Progenetix - Cancer CNV Information Resource

- Progenetix literature collection contains information about articles reporting genome profiling experiments (aCGH, cCGH, WES, WGS) in cancer samples
- continuous collection
- annotation of metadata extracted from the articles
  - cancer type
  - geographic location (by author or from text)
  - sample numbers per technology
  - contact information



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.

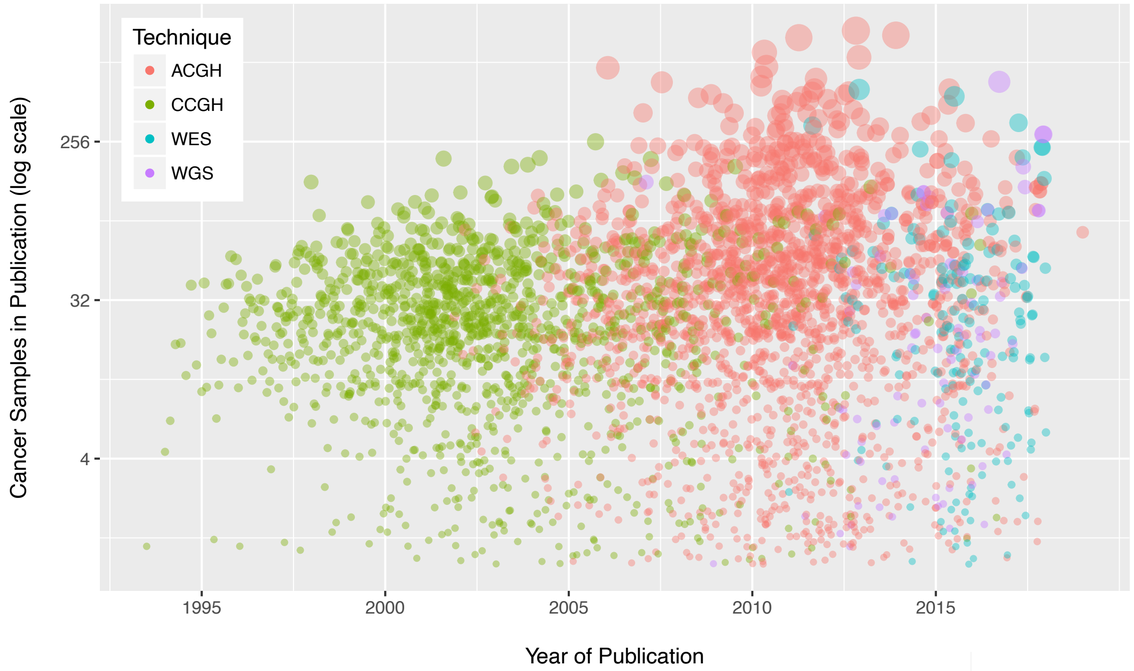
Show 50 entries

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



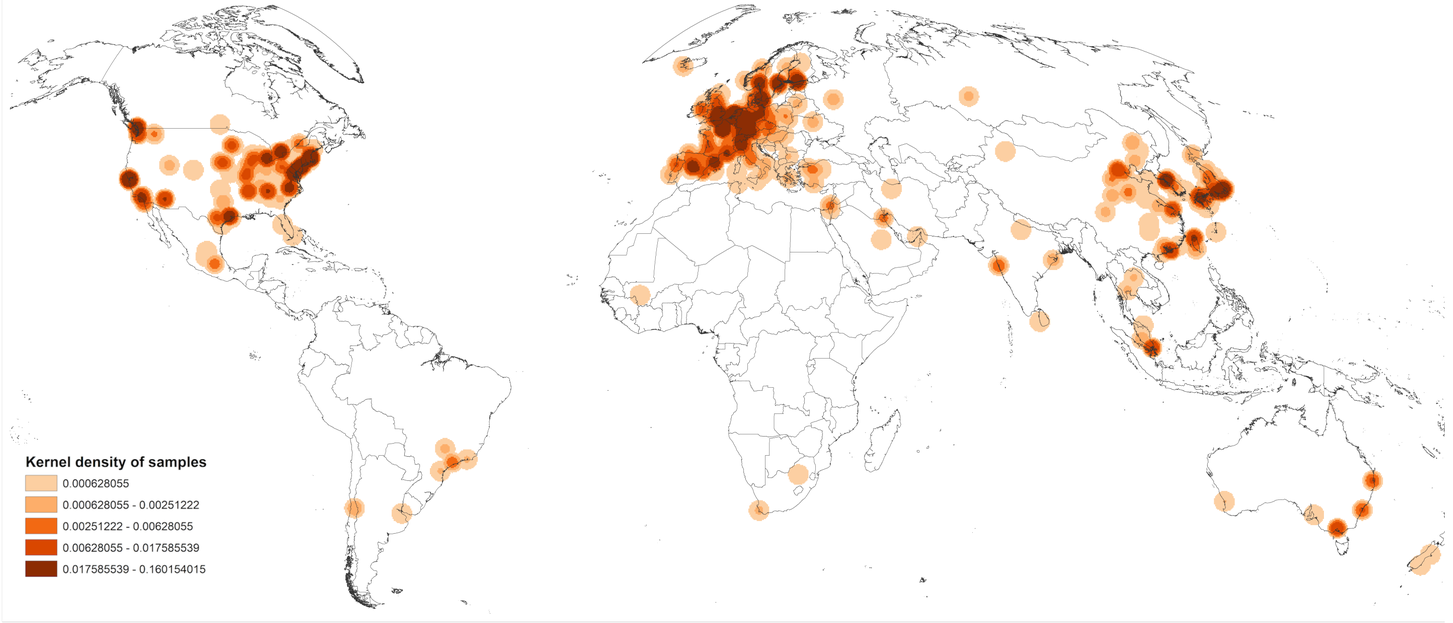
Publication Landscape of Cancer CNV Profiling

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.



progenetix

Data Driven Standards for Genomic Data Exchange

ELIXIR Beacon :: Beacon+ :: SchemaBlocks {S}[B]

University of Zurich[UZH]

elixir

SIB

Global Alliance for Genomics & Health

GENOMICS

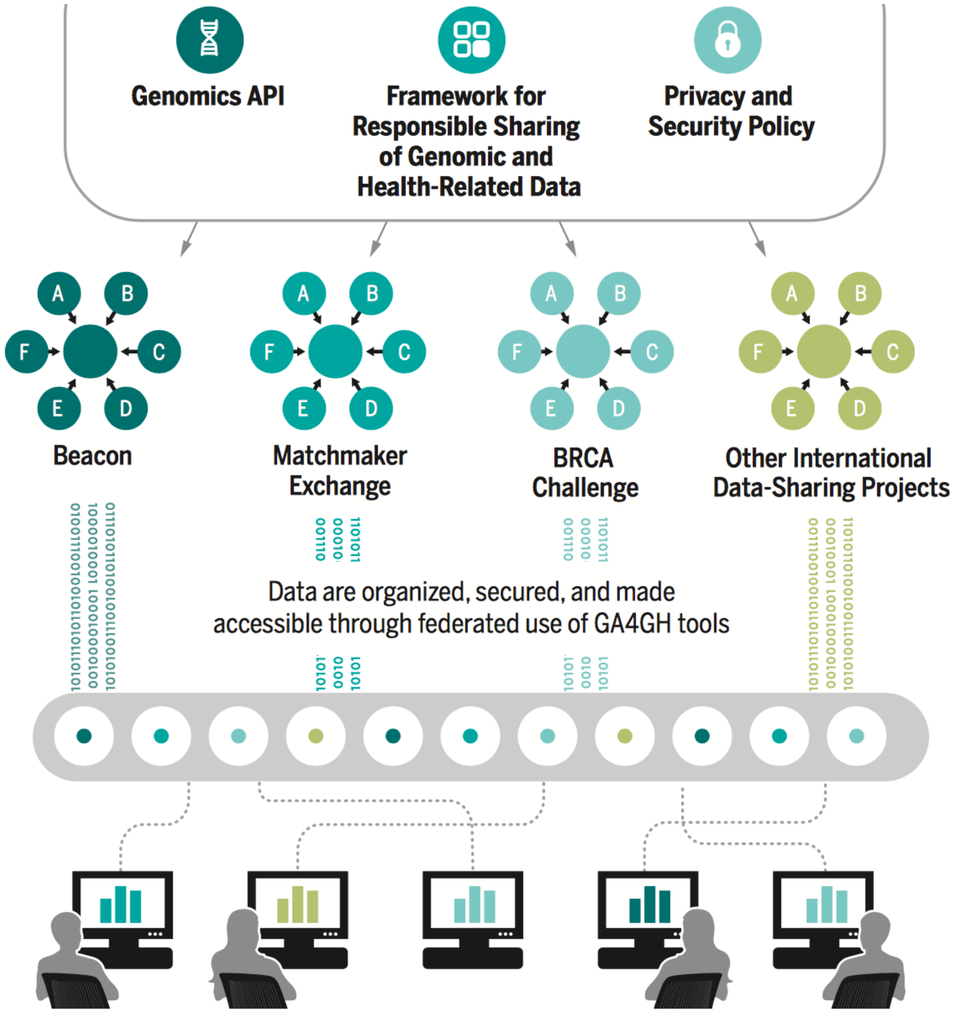
A federated ecosystem for sharing genomic, clinical data

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

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

GENOMICS

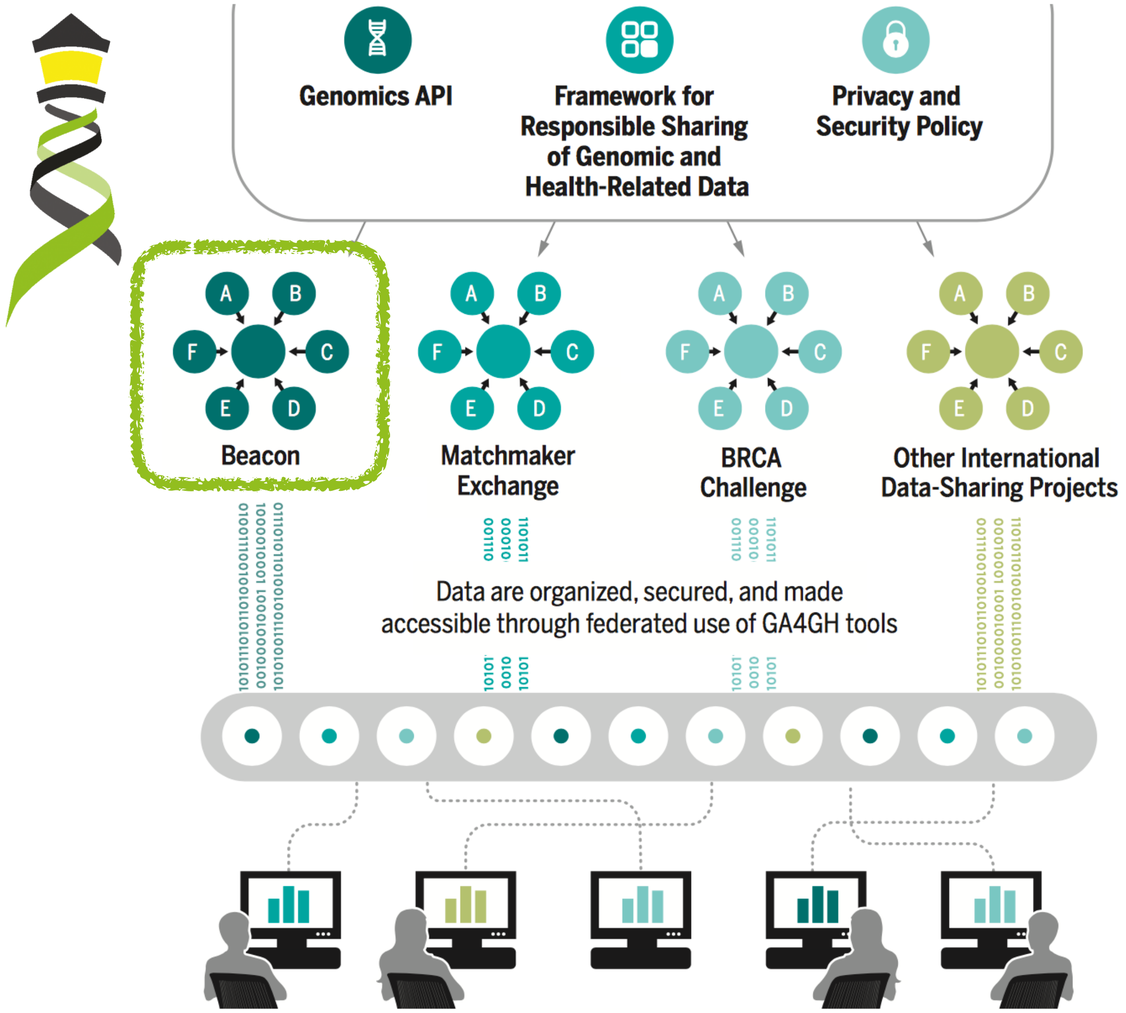
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

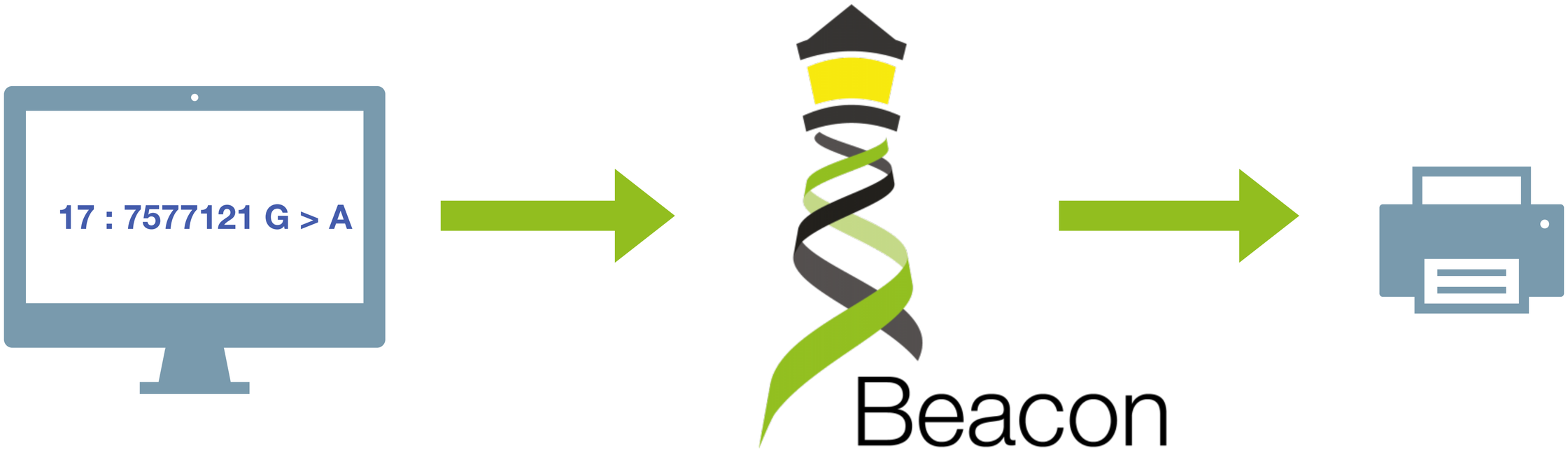
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



DNASTACK

Global Alliance for Genomics & Health



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

YES | **NO** | **\0**

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

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

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

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

ELIXIR - Towards Biomedical Beacons

Needs & Models Beyond Basic Variant Discovery

elixir

Global Alliance
for Genomics & Health

Simplify the way people search for and request access to potentially identifiable data in international and national genomic data resources

Working towards GA4GH standards, APIs and toolkits to be used throughout ELIXIR Nodes for human data discovery and access => GA4GH in Europe

15/23 Nodes connected

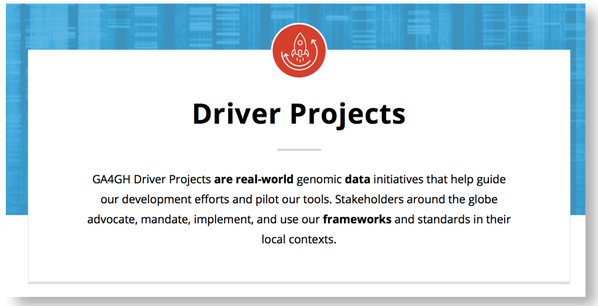
61 connections

Clinical & Phenotypic Data Capture	Large Scale Genomics	Genomic Knowledge Standards	Discovery	Cloud	Data Use & Researcher Identities (DURI)	Regulatory & Ethics	Data Security

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



ELIXIR Beacon
https://www.elixir-europe.org/about/implementation-studies/beacons

Europe

Champions: Jordi Rambla, Juha Tornroos, Gary Saunders

v1.1 and roadmap

- ☑ **structural variations** (DUP, DEL) in addition to SNV
- ☑ … more structural queries (translocations/fusions…)
- ☑ Beacon queries as entry for **data handover** (outside Beacon protocol)
- ☑ layered authentication system using **ELIXIR AAI**
- **v2** **filters** for phenotypic & technical metadata
- **v2** Extended quantitative responses
- ☐ Ubiquitous **deployment** (e.g. throughout ELIXIR network)

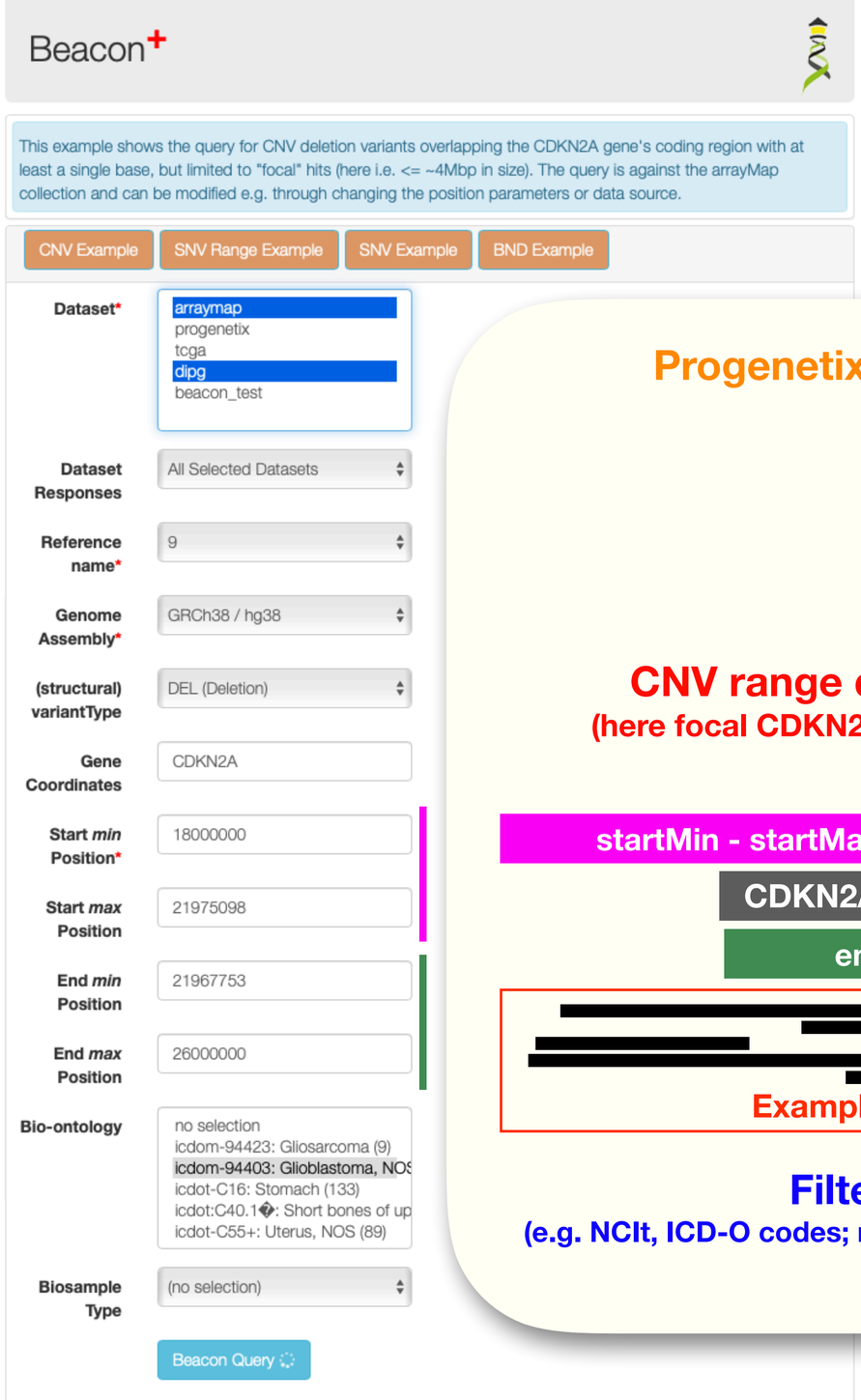
Global Alliance for Genomics & Health

Beacon+ @ UZH

A Beacon Project Technology Demonstrator

- implementing features from roadmap for feasibility testing
  - ***structural variants*** (implemented in v1.0.1)
  - ***handover*** mechanism (implemented in v1.1.0)
  - ***filters*** for phenotypes and other parameters (pre v2)
- runs against complete Progenetix (including TCGA) and arrayMap resources
- backend storage follows GA4GH object model
  - see schemablocks.org

beacon.progenetix.org/ui/



Progenetix datasets

CNV range query example
(here focal CDKN2A/B & MTAP deletion)

startMin - startMax

CDKN2A CDR

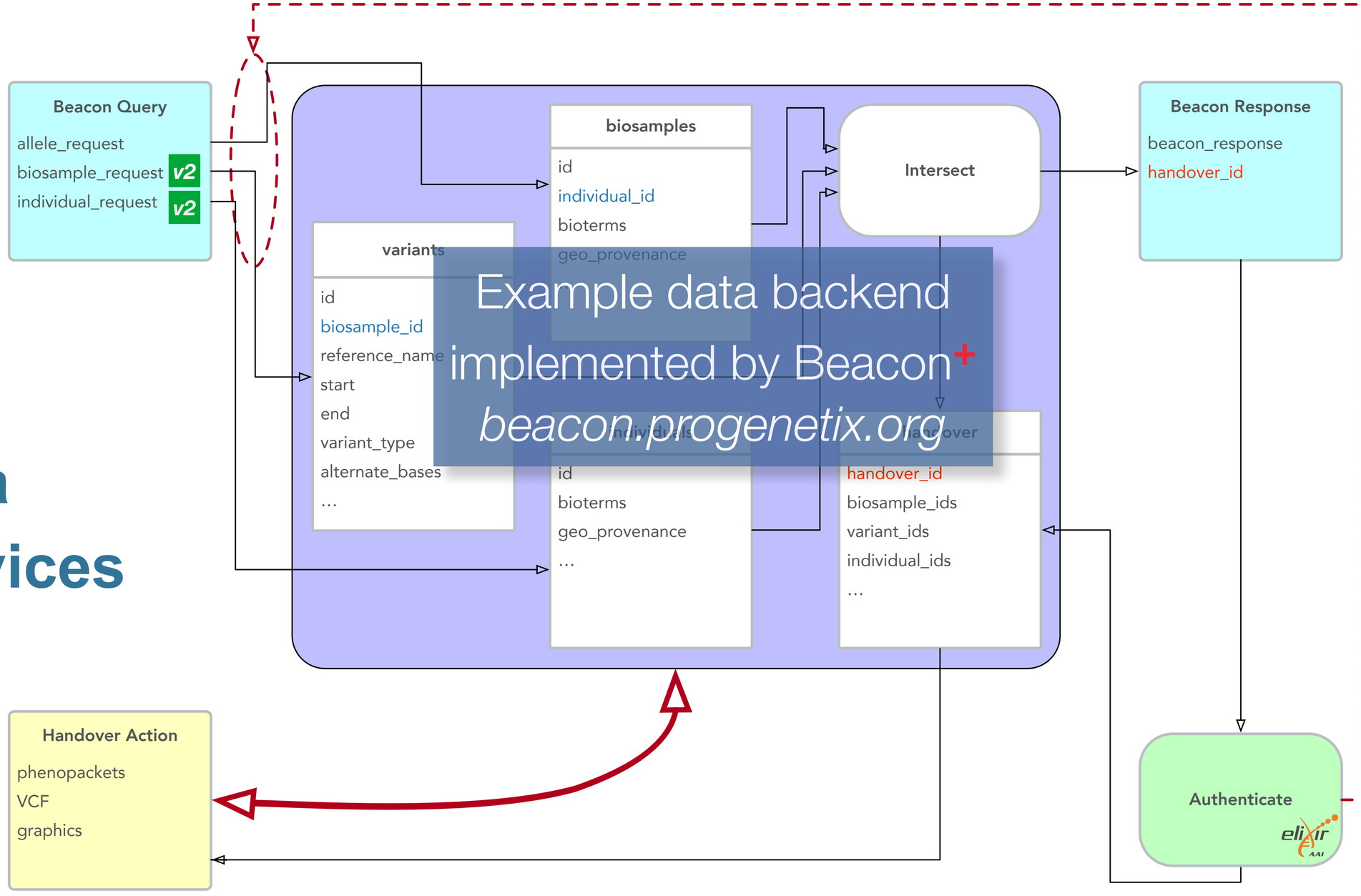
endMin - endMax

Example Hits

Filters
(e.g. NCIt, ICD-O codes; neoplastic/reference ...)

Beacon & Handover

Beacons v1.1 supports **data delivery services**



Beacon I/O

Handover

Authentication

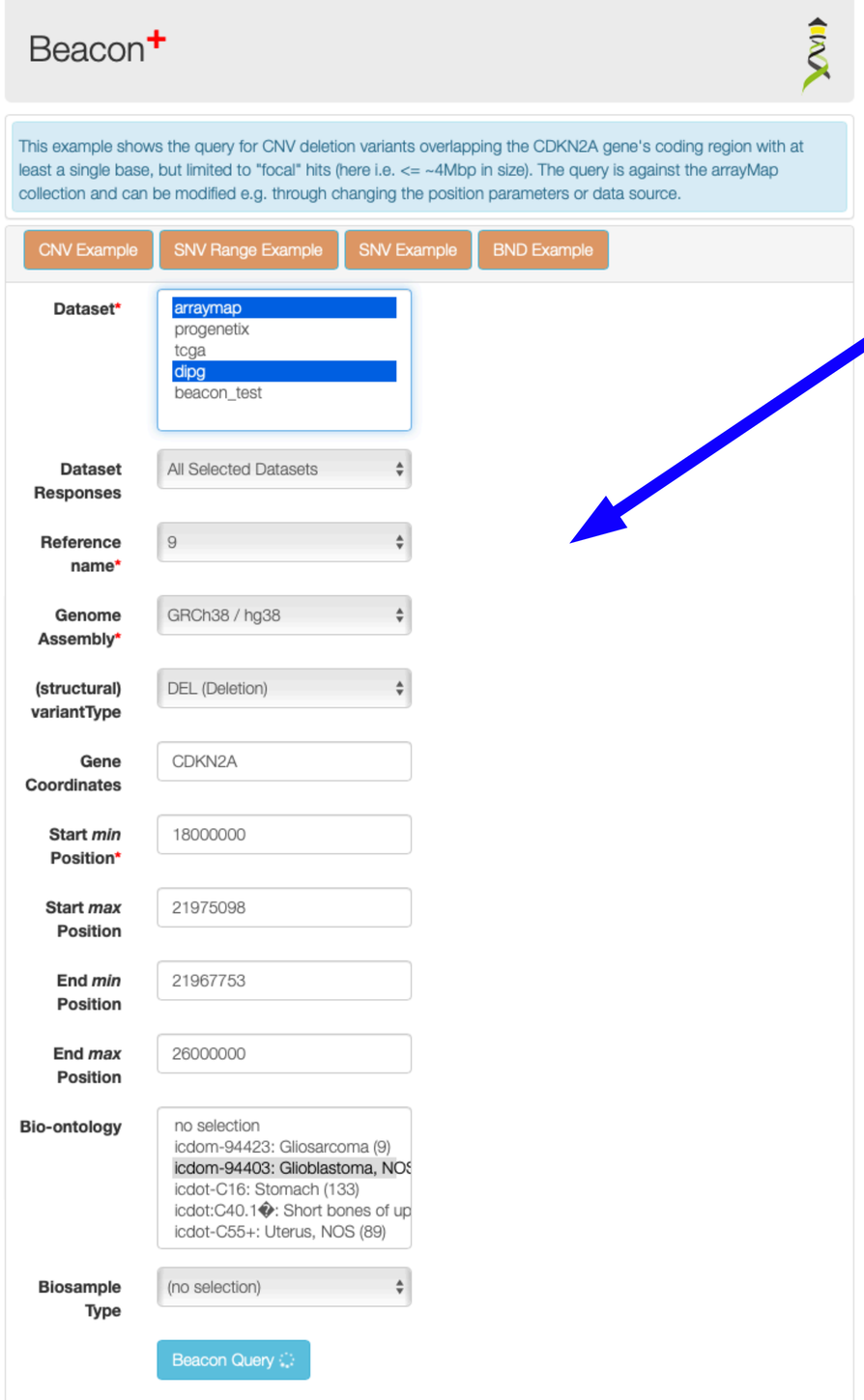
Michael Baudis

Global Alliance for Genomics & Health

SIB

University of Zurich UZH

elixir



https://beacon.progenetix.org/cgi-bin/beaconresponse.cgi?datasetIds=arraymap&datasetIds=dipg&datasetAlleleResponses=ALL&referenceName=9&assemblyId=GRCh38&variantType=DEL&startMin=18000000&startMax=21975098&endMin=21967753&endMax=26000000&filters=icdom-94403

query

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	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
dipg	GRCh38	9	18000000 - 21975098 21967753 - 26000000	N DEL	icdom-94403	0 0 0	0	JSON UCSC

```
"datasetHandover" : [
            {
               "url" : "https://progenetix.org/cgi-bin/beacondeliver.cgi?do=biosamplesdata&accessid=d1ffd548-e68e-11e9-87ed-fcb07b51aec4",
               "description" : "retrieve data of the biosamples matched by the query",
               "handoverType" : {
                  "label" : "Biosamples",
                  "id" : "pgx:handover:biosamplesdata"
               }
            },
```

handover

A Beacon "datasetAlleleResponses" can provide ***handover*** objects to initiate further actions (data retrieval, visualization ...) outside of the Beacon protocol.

ELIXIR Beacons

EMBL-EBI

ELIXIR FINLAND

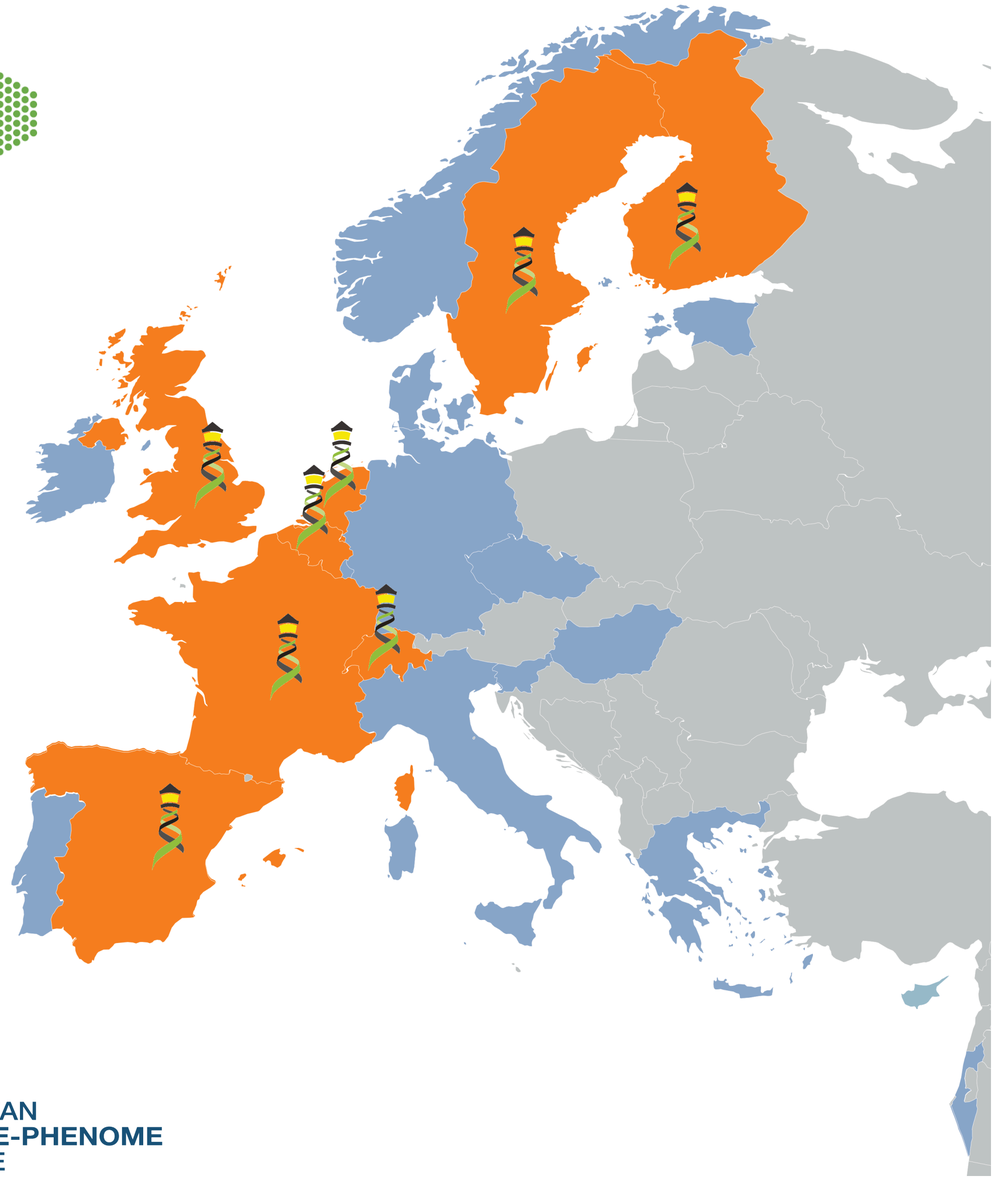
ELIXIR SWEDEN

ELIXIR BELGIUM

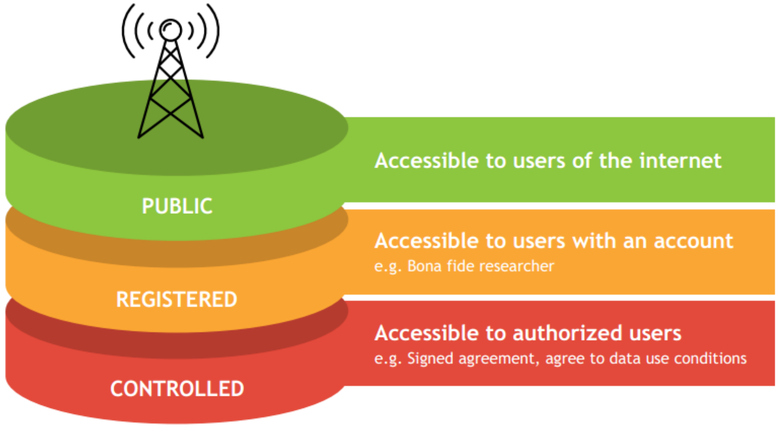
ELIXIR NETHERLANDS

ELIXIR SWITZERLAND

ELIXIR SPAIN



EUROPEAN GENOME-PHENOME ARCHIVE



ELIXIR AAI

- Driving implementation of Beacon technology in ELIXIR Nodes
  - ➡ 9 National Nodes have lit Beacons
- ELIXIR Authentication and Authorization Infrastructure (AAI)

Beacon

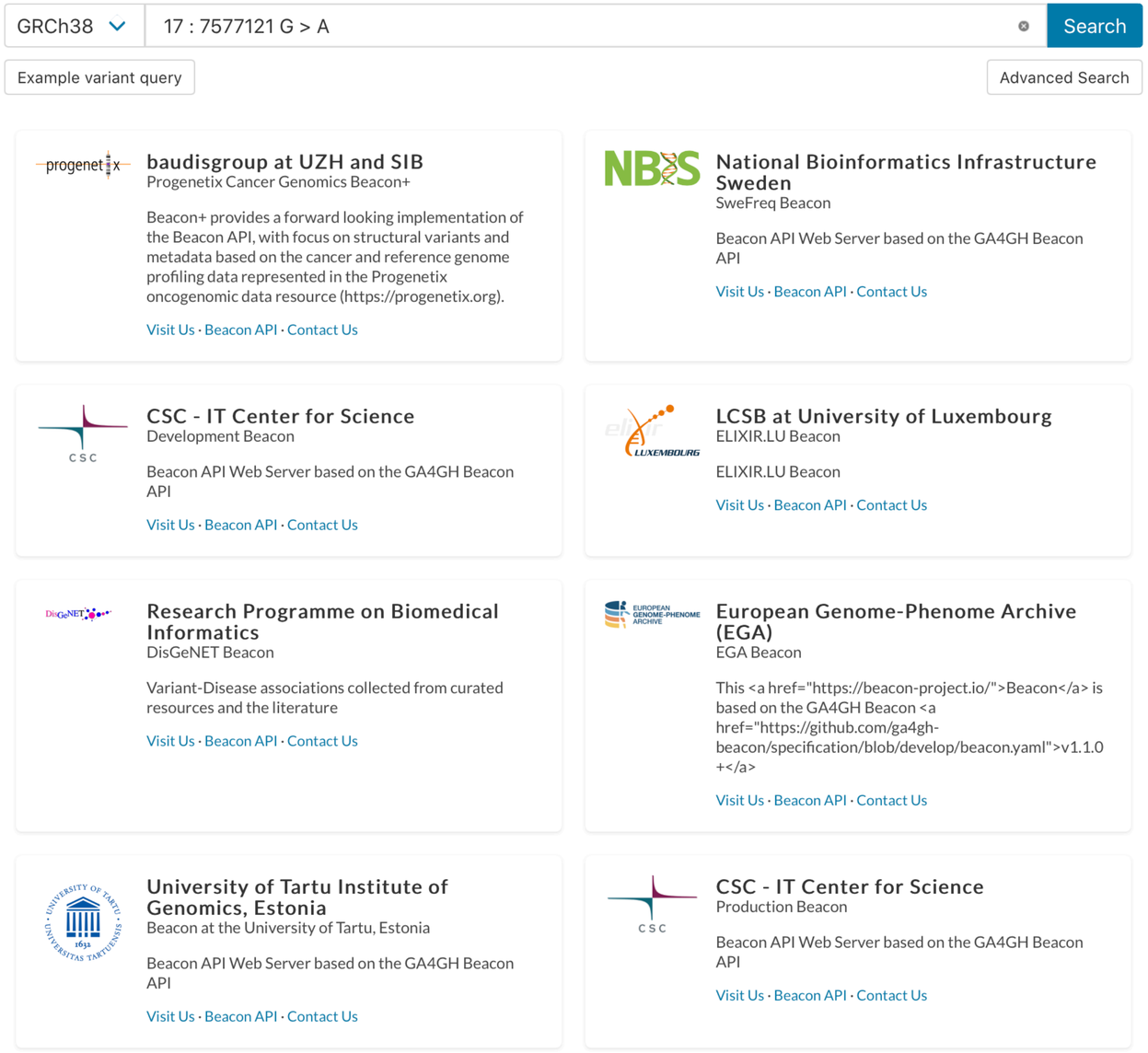
Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

ELIXIR

ELIXIR Beacon Network

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

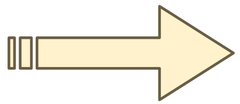


Beacon v2 - Areas of Change

- Separate query types for different genomic variants
  - SNPs ***BeaconSnvRequest***
  - Structural Variants ***BeaconCnvRequest***
  - Region ***BeaconRangeRequest***
  - ...
- Access levels
- Filters
  - Simple general filter schema w/ **scoping through prefixes** (CURIEs, private implementations)
- New types of queries:
  - By sample, patient, variant effect/evidence
  - Complex queries? (stakeholder driven; e.g. EJP-RD, GEL…)
- Schema versions & Service Info
  - Negotiated queries based on individual Beacon capabilities

Tested and already implemented by Beacons

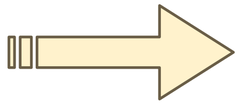
v2.0



GA4GH approval process ("major product update")

Ongoing

v2.n



incremental rollout after v2.0

Code leads: Sabela de la Torre Penas & Jordi Rambla

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

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

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

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**".

ELIXIR Genome Beacons

A Driver Project of the Global Alliance for Genomics and Health

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.

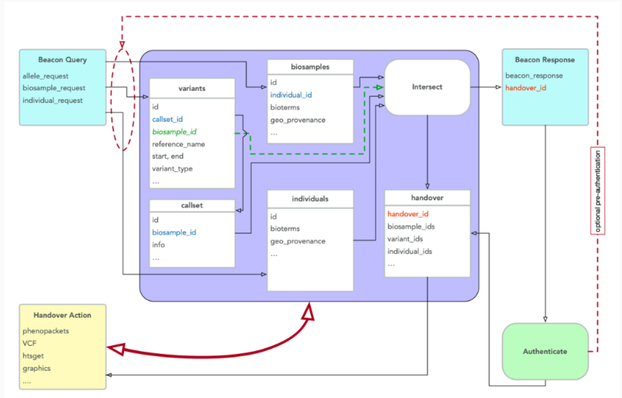
@mbaudis 2019-04-30: more ...

[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

Search or jump to... / Pull requests Issues Marketplace Explore

Beacon

Beacon Project, Global Alliance for Genomics & Health.

http://beacon-project.io/

Repositories 7 People 15 Teams 2 Projects 1 Settings

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

Find a repository... Type: All Language: All New

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

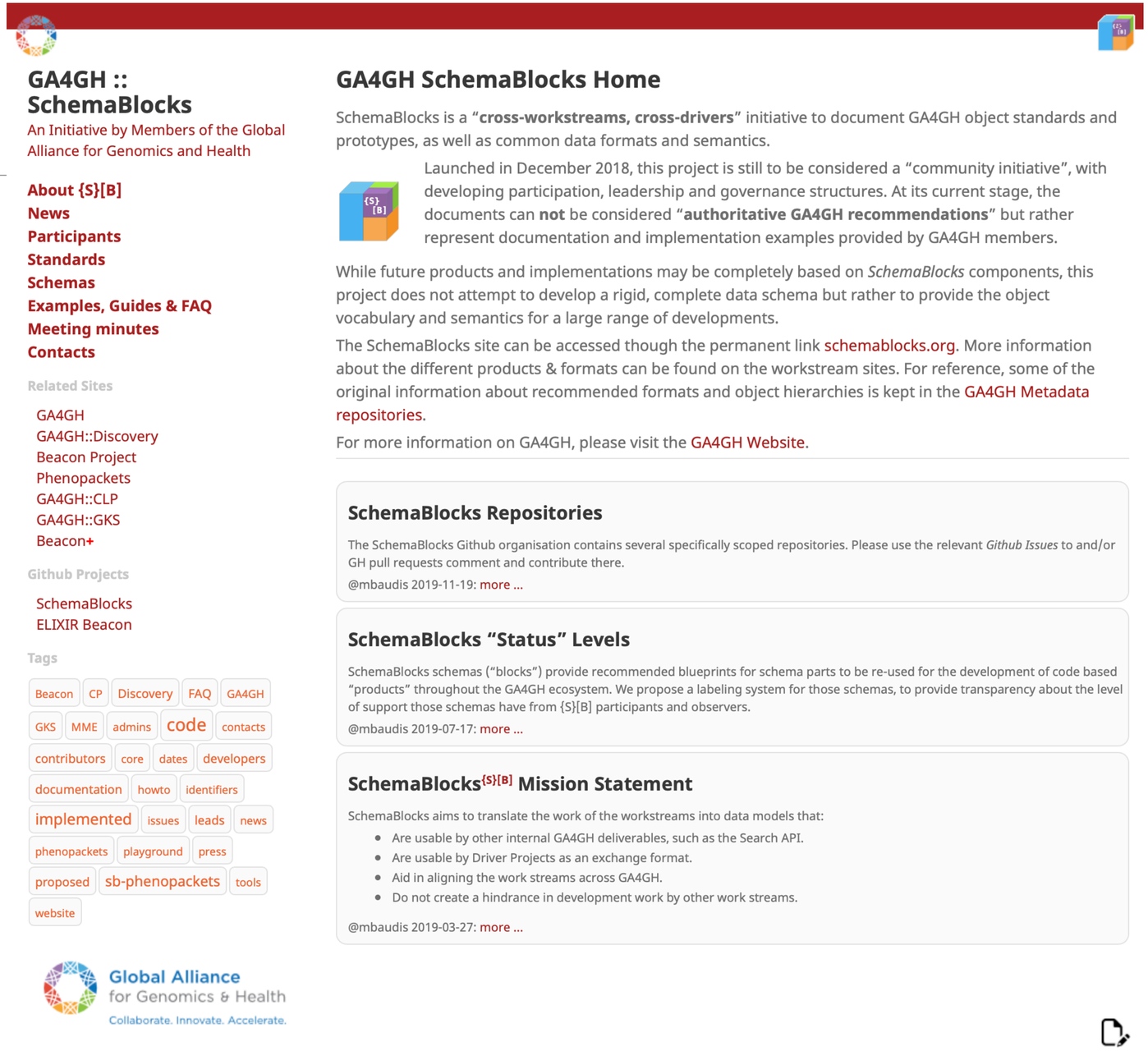
People 15 >

github.com/ga4gh-beacon/

GA4GH {S}[B]
SchemaBlocks

- “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”
- recognized in GA4GH roadmap as element in "TASC" effort

schemablocks.org



GA4GH :: SchemaBlocks

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

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

Related Sites

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

Github Projects

SchemaBlocks
ELIXIR Beacon

Tags

Beacon CP Discovery FAQ GA4GH GKS MME admins code contacts contributors core dates developers documentation howto identifiers implemented issues leads news phenopackets playground press proposed sb-phenopackets tools website

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

GA4GH SchemaBlocks Home

SchemaBlocks is a “**cross-workstreams, cross-drivers**” initiative to document GA4GH object standards and prototypes, as well as common data formats and semantics.

Launched in December 2018, this project is still to be considered a “community initiative”, with developing participation, leadership and governance structures. At its current stage, the documents can **not** be considered “**authoritative GA4GH recommendations**” but rather represent documentation and implementation examples provided by GA4GH members.

While future products and implementations may be completely based on *SchemaBlocks* components, this project does not attempt to develop a rigid, complete data schema but rather to provide the object vocabulary and semantics for a large range of developments.

The SchemaBlocks site can be accessed though the permanent link schemablocks.org. More information about the different products & formats can be found on the workstream sites. For reference, some of the original information about recommended formats and object hierarchies is kept in the GA4GH Metadata repositories.

For more information on GA4GH, please visit the GA4GH Website.

SchemaBlocks Repositories

The SchemaBlocks Github organisation contains several specifically scoped repositories. Please use the relevant *Github Issues* to and/or GH pull requests comment and contribute there.

@mbaudis 2019-11-19: more ...

SchemaBlocks “Status” Levels

SchemaBlocks schemas (“blocks”) provide recommended blueprints for schema parts to be re-used for the development of code based “products” throughout the GA4GH ecosystem. We propose a labeling system for those schemas, to provide transparency about the level of support those schemas have from {S}[B] participants and observers.

@mbaudis 2019-07-17: more ...

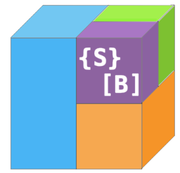
SchemaBlocks{S}[B] Mission Statement

SchemaBlocks aims to translate the work of the workstreams into data models that:

- Are usable by other internal GA4GH deliverables, such as the Search API.
- Are usable by Driver Projects as an exchange format.
- Aid in aligning the work streams across GA4GH.
- Do not create a hindrance in development work by other work streams.

@mbaudis 2019-03-27: more ...

SchemaBlocks - A GA4GH Community Initiative



GA4GH SchemaBlocks Home

SchemaBlocks is a **"cross-workstreams, cross-drivers"** initiative to document GA4GH object standards and prototypes, as well as common data formats and semantics.

Launched in December 2018, this project is still to be considered a "community initiative", with developing participation, leadership and governance structures. At its current stage, the documents can **not** be considered "**authoritative GA4GH recommendations**" but rather represent documentation and implementation examples provided by GA4GH members.

While future products and implementations may be completely based on *SchemaBlocks* components, this project does not attempt to develop a rigid, complete data schema but rather to provide the object vocabulary and semantics for a large range of developments.

The SchemaBlocks site can be accessed though the permanent link schemablocks.org. More information about the different products & formats can be found on the workstream sites. For reference, some of the original information about recommended formats and object hierarchies is kept in the GA4GH Metadata repositories.

For more information on GA4GH, please visit the GA4GH Website.

SchemaBlocks "Status" Levels

SchemaBlocks schemas ("blocks") provide recommended blueprints for schema parts to be re-used for the development of code based "products" throughout the GA4GH ecosystem. We propose a labeling system for those schemas, to provide transparency about the level of support those schemas have from {S}[B] participants and observers.

Proposed {S}[B] Status Levels

The current status level of thiose recommendations is "proposed".

- `playground`
  - early development or import stage, of any quality
  - no recommendation; existence does not mean any current or future {S}[B] support
- `proposed`
  - at least some {S}[B] contributors are in favour of such a block
  - the code may undergo considerable maturation
  - not recommended for integration into products w/o close tracking
  - contributions and discussions are encouraged
- `implemented`
  - mature block which is implemented in one or more {S}[B] aligned schemas
  - may be extended from a core block or be too specific for general ("core") usability
- `core`
  - a schema block with recommended use
  - stable through minor version changes
  - has to be used in at least 2 standards/products approved by the GA4GH Steering Committee

SchemaBlocks{S}[B] Mission Statement

SchemaBlocks aims to translate the work of the workstreams into data models that:

- Are usable by other internal GA4GH deliverables, such as the Search API.
- Are usable by Driver Projects as an exchange format.
- Aid in aligning the work streams across GA4GH.
- Do not create a hindrance in development work by other work streams.

After discussions with stakeholders from GA4GH work streams and driver projects who create data models (such as Phenopackets, Search API) or who would use SchemaBlocks for the development of their APIs and data exchange formats (Beacon, EGA, GeL), the SchemaBlocks team has come up with the following principles for this initiative:

Work Stream Interactions

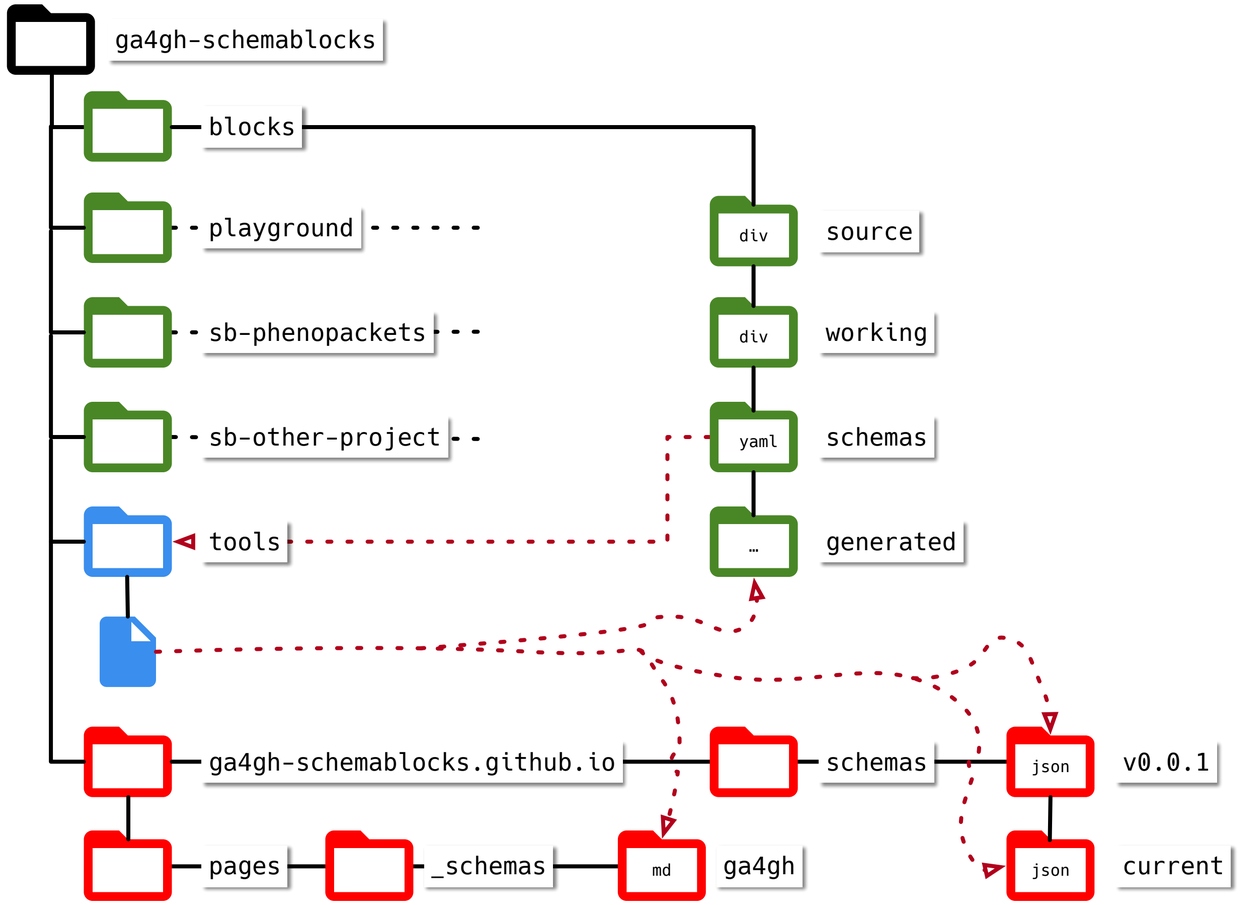
Work streams will continue to create standards proposals and their own coherent project implementations, but will work with the SchemaBlocks group to write the Blocks that will come from their own work and are considered of overarching use. Generally, primary work stream and driver project outputs will live in their own spaces outside of SchemaBlocks, with shareable, mature elements - code, documentation, implementation snapshots - being represented in {S}[B].

{S}[B] SchemaBlocks Github Repository Structure

blocks repositories

conversion/validation tools

website repository
(Markdown w/ YAML for Github Pages)



Dissection & Transformation

```
...
// See http://build.fhir.org/datatypes and http://build.fhir.org/condition-definitions.html#Condition.onset_x_
// In FHIR this is represented as a UCUM measurement - http://unitsofmeasure.org/trac/
message Age {

    // The :ref:`ISO 8601<metadata_date_time>` age of this object as ISO8601
    // duration or time intervals. The use of time intervals makes an additional
    // anchor unnecessary (i.e. DOB and age can be represented as start-anchored
    // time interval, e.g. 1967-11-21/P40Y10M05D)
    string age = 1;
}

message AgeRange {
    Age start = 1;
    Age end = 2;
}

// Message to indicate a disease (diagnosis) and its recorded onset.
message Disease {
    // The identifier of this disease e.g. MONDO:0007043, OMIM:101600, Orphanet:710, DOID:14705 (note these are all equivalent)
    OntologyClass term = 1;

    // The onset of the disease. The values of this will come from the HPO onset hierarchy
    // i.e. subclasses of HP:0003674
    // FHIR mapping: Condition.onset
    oneof onset {
        Age age_of_onset = 2;
        AgeRange age_range_of_onset = 3;
        OntologyClass class_of_onset = 4;
    }

    // Disease staging, the extent to which a disease has developed.
    // For cancers, see https://www.cancer.gov/about-cancer/diagnosis-staging/staging
    // Valid values include child terms of NCIT:C28108 (Disease Stage Qualifier)
    repeated OntologyClass disease_stage = 5;
...
```

32 lines (31 sloc) | 872 Bytes | Raw | Blame | History

```
"$schema": http://json-schema.org/draft-07/sc
"$id": https://schemablocks.org/schemas/sb-ph
title: Age
description: Age
type: object
meta:
  contributors:
    - description: "Michael Baudis"
      id: "orcid:0000-0002-9903-4248"
    - description: "Jules Jacobsen"
      id: "orcid:0000-0002-3265-15918"
    - description: "Peter Robinson"
      id: "orcid:0000-0002-0736-91998"
  provenance:
    - description: Phenopackets
      id: 'https://github.com/phenopackets/ph
  used_by:
    - description: Phenopackets
      id: 'https://github.com/phenopackets/ph
  sb_status: implemented
properties:
  age:
    type: string
    description: Age as ISO8601 period
    examples:
      - 'P12Y'

required:
  - age
additionalProperties: false
examples:
  - age: 'P14Y'
```


```
{
  "$id": "https://schemablocks.org/schemas/sb-phenopackets/Age/v0.0.1",
  "$schema": "http://json-schema.org/draft-07/schema#",
  "additionalProperties": "",
  "description": "Age",
  "examples": [
    {
      "age": "P14Y"
    }
  ],
  "meta": {
    "contributors": [
      {
        "description": "Michael Baudis",
        "id": "orcid:0000-0002-9903-4248"
      },
      {
        "description": "Jules Jacobsen",
        "id": "orcid:0000-0002-3265-15918"
      },
      {
        "description": "Peter Robinson",
        "id": "orcid:0000-0002-0736-91998"
      }
    ],
    "provenance": [
      {
        "description": "Phenopackets",
        "id": "https://github.com/phenopackets/phenopacket-schema/blob/master/docs/age.rst"
      }
    ],
    "sb_status": "implemented",
    "used_by": [
      {
        "description": "Phenopackets",
        "id": "https://github.com/phenopackets/phenopacket-schema/blob/master/docs/age.rst"
      }
    ]
  },
  "properties": {
    "age": {
      "description": "Age as ISO8601 period",
      "examples": [
        "P12Y"
      ],
      "type": "string"
    }
  },
  "required": [
    "age"
  ],
  "title": "Age",
  "type": "object"
}
```


- schema documents are programmatically converted into different outputs
- a versioned JSON document serves as canonical reference for integration into other products/schemas

BeaconAlleleRequest beacon ↗

{S}[B] Status [i]	implemented
Provenance	○ Beacon API
Used by	○ Beacon ○ Progenetix database schema (Beacon+ backend)
Contributors	○ Marc Fiume ○ Michael Baudis ○ Sabela de la Torre Pernas ○ Jordi Rambla ○ Beacon developers...
Source (v1.1.0)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: Allele request as interpreted by the beacon.

Properties

Property	Type
alternateBases	string
assemblyId	string
datasetIds	array of string
end	integer
endMax	integer
endMin	integer
mateName	https://schemablocks.org/schemas/beacon/v1.1.0/Chror [HTML]
referenceBases	string
referenceName	https://schemablocks.org/schemas/beacon/v1.1.0/Chror [HTML]
start	integer (int64)
startMax	integer
startMin	integer
variantType	string

alternateBases

- type: string

The bases that appear instead of the reference bases. Accepted values: [ACGTN]*. N is a wildcard, that denotes the position of any base, and can be used as a standalone base of any type or within a partially known sequence. For example a sequence where the first and last bases are known, but the middle portion can exhibit countless variations of [ACGT], or the bases are unknown: ANNT the Ns can take take any form of [ACGT], which makes both ACCT and ATGT (or any other combination) viable sequences.

Symbolic ALT alleles (DEL, INS, DUP, INV, CNV, DUP:TANDEM, DEL:ME, INS:ME) will be represented in `variantType`.

Optional: either `alternateBases` or `variantType` is required.

`alternateBases` **Value Example**

assemblyId

- type: string

Assembly identifier (GRC notation, e.g. GRCh37).

`assemblyId` **Value Example**

Curie sb-vr-spec ↗

{S}[B] Status [i]	implemented
Provenance	○ vr-spec
Used by	○ vr-spec
Contributors	○ Reece Hart ○ Michael Baudis
Source (v1.0)	○ raw source [JSON] ○ Github

Attributes

Type: string
Pattern: ^\w[^:]+:.+$
Description: A string that refers to an object uniquely. T sender.

VR does not impose any contraints on strings used as ids data, the VR Specification RECOMMENDS that implement

String CURIEs are represented as `prefix:reference` (W `namespace:accession` or `namespace:local id` colloqu

The VR specification also RECOMMENDS that `prefix` be

The `reference` component is an unconstrained string.

A CURIE is a URI. URIs may *locate* objects (i.e., specify wh VR uses CURIEs primarily as a naming mechanism.

Implementations MAY provide CURIE resolution mechan

Using internal ids in public messages is strongly discoura

`Curie` Value Examples

```
"ga4gh:GA.01234abcde"
```

```
"DUO:0000004"
```

```
"orcid:0000-0003-3463-0775"
```

```
"PMID:15254584"
```

Biosample sb-phenopackets ↗

{S}[B] Status [i]	implemented
Provenance	○ Phenopackets
Used by	○ Phenopackets
Contributors	○ GA4GH Data Working Group ○ Jules Jacobsen ○ Peter Robinson ○ Michael Baudis ○ Melanie Courtot ○ Isuru Liyanage
Source (v1.0.0)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: A Biosample refers to a unit of biological material from which the substrate molec genomic DNA, RNA, proteins) for molecular analyses (e.g. sequencing, array hybridisation, mas spectrometry) are extracted.

Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequencii fraction from a gradient centrifugation.

Several instances (e.g. technical replicates) or types of experiments (e.g. genomic array as well experiments) may refer to the same Biosample.

FHIR mapping: Specimen.

Properties

Property	Type
ageOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Age.json [SRC] [HTML]
ageRangeOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/AgeRange.json [SRC] [HTML]
description	string
diagnosticMarkers	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
histologicalDiagnosis	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
htsFiles	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/HtsFile.json [SRC] [HTML]
id	string
individualId	string
isControlSample	boolean
phenotypicFeature	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/PhenotypicFeature.json [SRC] [HTML]
procedure	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Procedure.json [SRC] [HTML]
sampledTissue	https://schemablocks.org/schemas/sb-

Checksum sb-checksum ↗

{S}[B] Status [i]	proposed
Provenance	○ GA4GH DRS (`develop` branch)
Used by	○ GA4GH DRS ○ GA4GH TRS
Contributors	○ Susheel Varma
Source (v0.0.1)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: Checksum

Properties

Property	Type
checksum	string
type	string

checksum

- type: string

The hexadecimal encoded (Base16) checksum for the data

`checksum` **Value Example**

```
"77af4d6b9913e693e8d0b4b294fa62ade6054e6b2f1ffb617ac955dd63fb0182"
```

type

- type: string

The digest method used to create the checksum. The value (e.g. `sha-256`) SHOULD be listed as `Hash Name String` in the GA4GH Hash Algorithm Registry. Other values MAY be used, as long as implementors are aware of the issues discussed in RFC6920.

GA4GH may provide more explicit guidance for use of non-IANA-registered algorithms in the future.

`type` **Value Example**

```
"sha-256"
```

SchemaBlocks {S}[B] - Directions & Contributions

- Recognized need of having a set of recommended standards for integrating into product development
  - ➡ no need to work through complex standards/projects like FHIR, Phenopackets ...
  - ➡ simplification of development
- SchemaBlocks {S}[B] to assume strategic position in GA4GH *TASC system
  - ➡ Inclusion into product approval processes?
  - ➡ Management/Support?
- Wish for participation of (GA4GH affiliated) groups & individuals, to ***expose*** their standards & products
- Most important role is the ***community aspect***, the interactive exchange of concepts, ideas, code, knowledge, resources ...
- Technical to-dos:
  - ➡ Lifecycle: Versioning and representation of donor schemas?
  - ➡ Development of conversion workflows for updated source products?
  - ➡ Alternative/conflicting blocks...: Graded recommendations? Name spacing?

*Technical Alignment Sub Committee

Michael's Hopes for his BBOP time

- Getting our annotations lifted to a higher level ...
  - creating sustainable resources & end points
- initiating research project that make use of our CNV data collections
- get feedback on concepts for the upcoming Beacon developments, specifically "filters" and recommended resources (ontologies...) and query strategies
- Schemablocks direction - feedback, engagement, recommendations...

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

@WORLD

MATTHIAS ALTMEYER
THOMAS EGGERMANN
ROSA NOGUERA
REINER SIEBERT
CAIUS SOLOVAN

University of Zurich[UZH]

SIB

elixir

Global Alliance for Genomics & Health

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

H-CNV

CHRISTOPHE BÉROUD
DAVID SALGADO

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | 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