University of Zurich[UZH]
Department of Molecular Life Sciences

Data Mining in Genomics

Resources | Standards | Protocols | Tools | Discourse for Genomic Research and Personalised Health Strategies

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

2020-03-19

Global Alliance for Genomics & Health

Genome screening at the core of “Personalised Health”

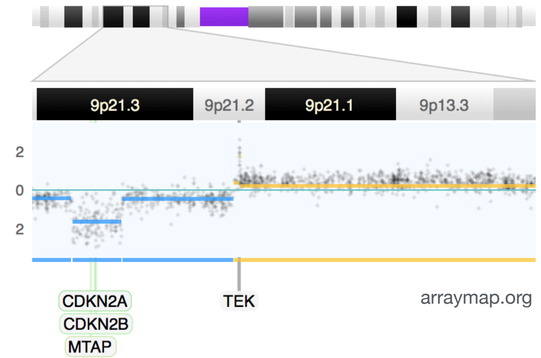
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**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

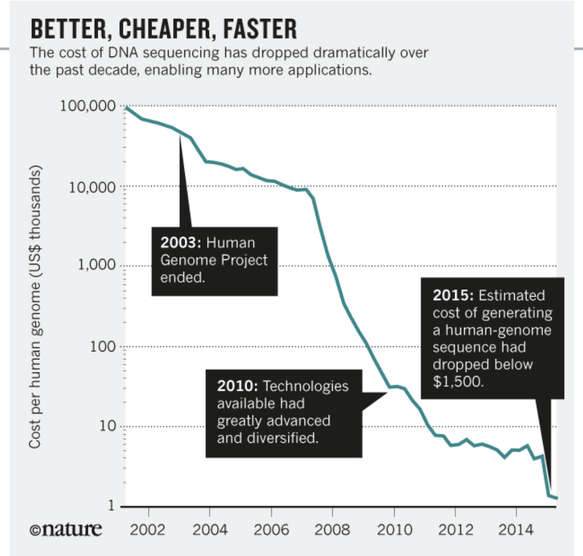
arrayMap

progenetix

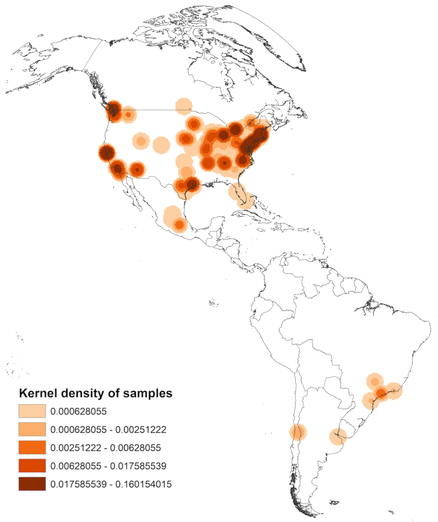
elixir

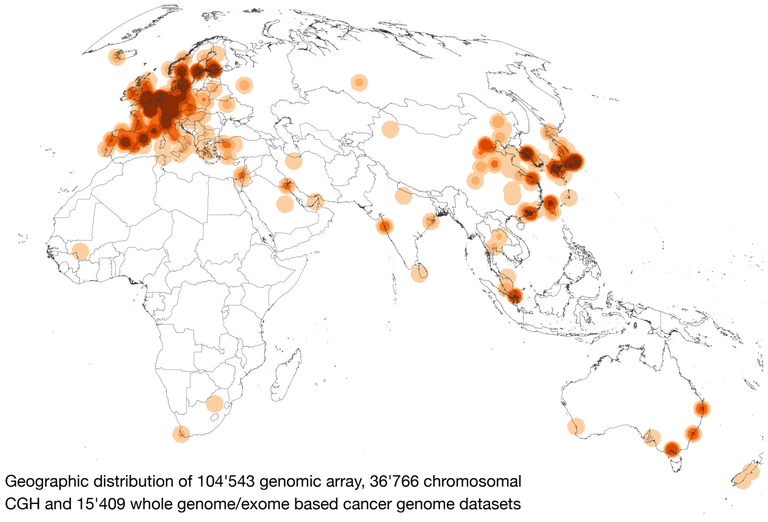
Global Alliance for Genomics & Health





The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)



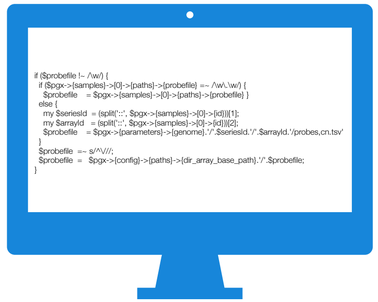


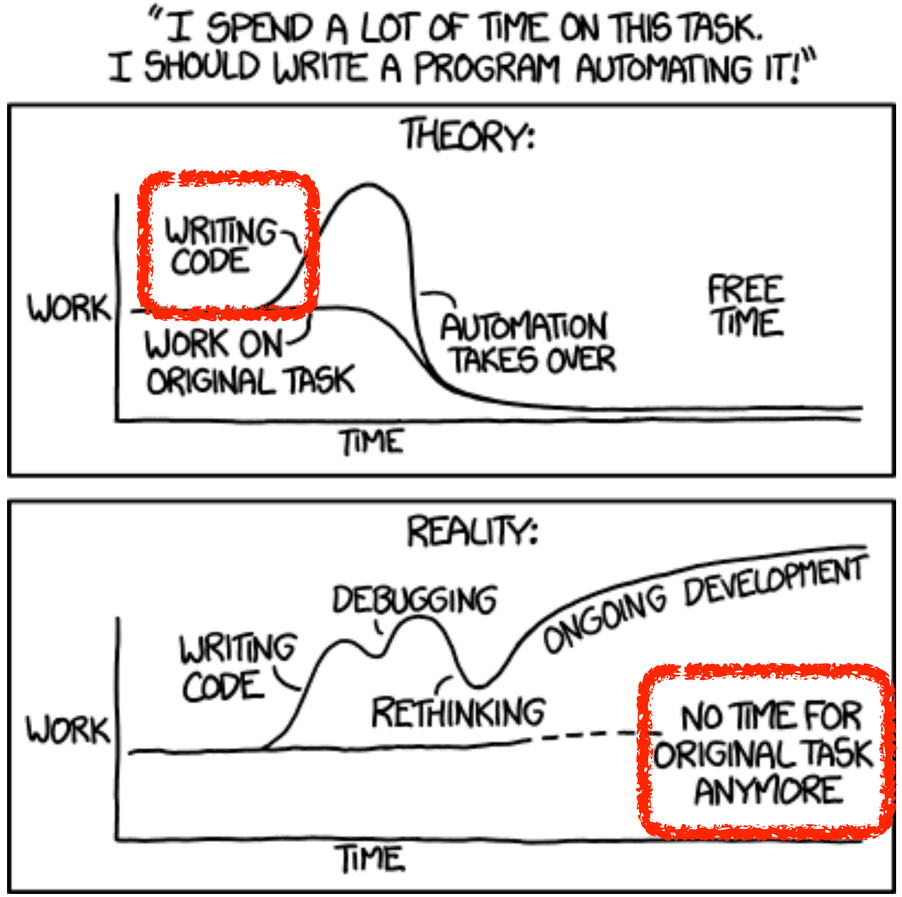
Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

{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/

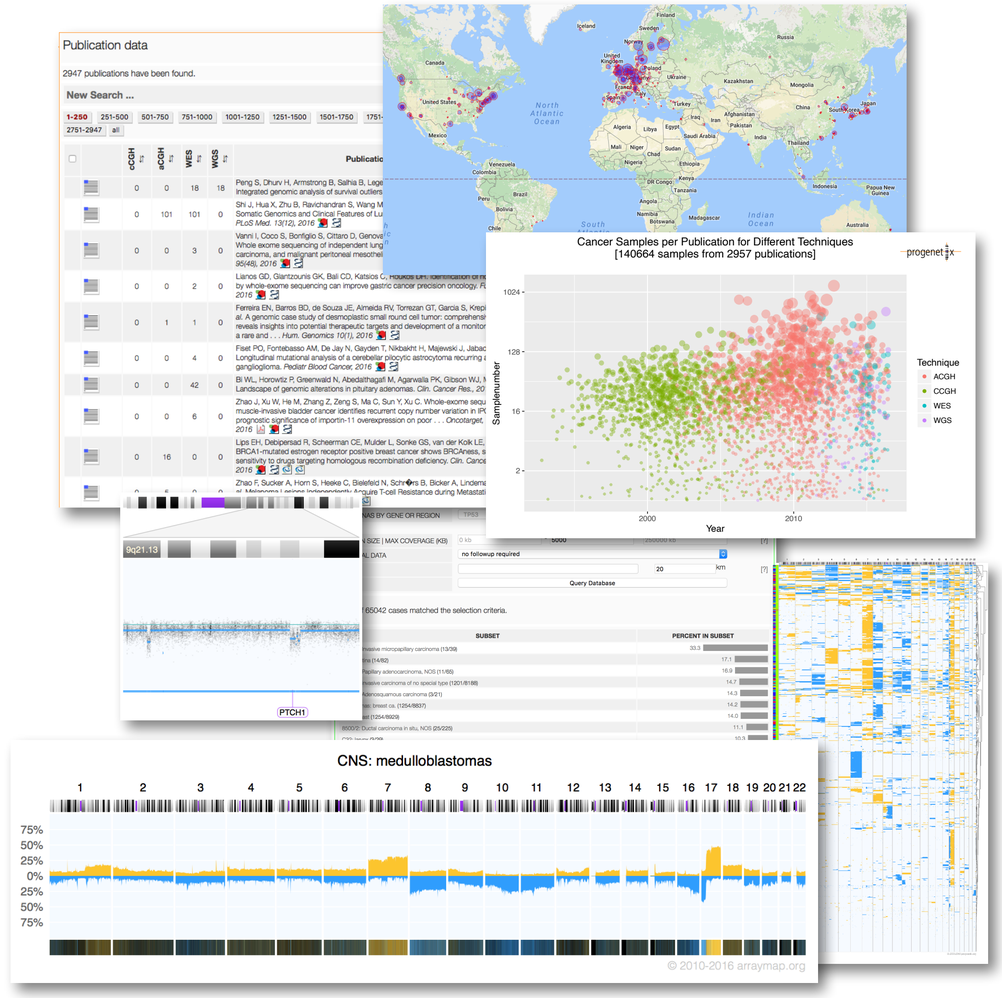
Our contributions I: Cancer genome knowledge resources and research

progenetix

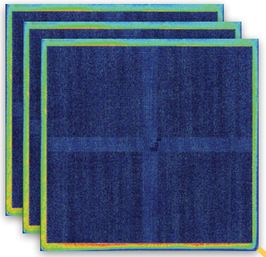
- curated cancer genome publication resource (more than 3200 manually curated articles)
- article metadata
- annotated genome profiles of >100'000 samples
- ontology mappings; clinical data where available
- epistemology: geographic and histologic sampling biases

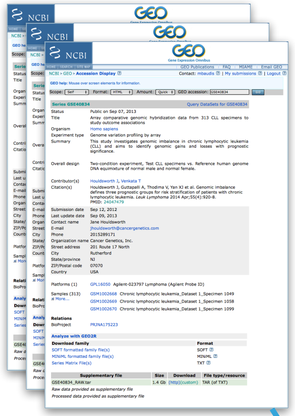
arrayMap

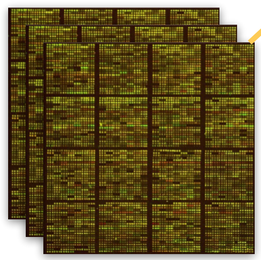
- more than 70'000 array based genome profiles
- probe level, copy number, metadata
- completely open data access through web interface, downloads and API calls
- re-annotated metadata (diagnostic coding, basic clinical) for all samples

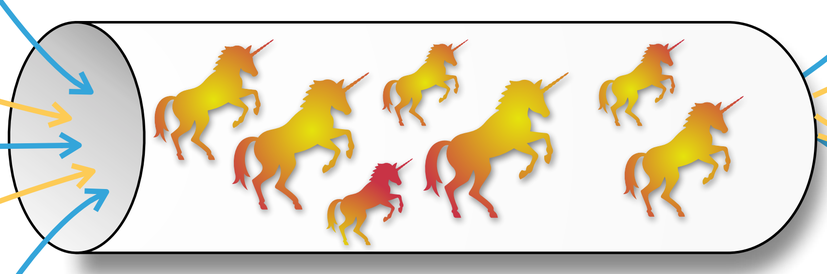


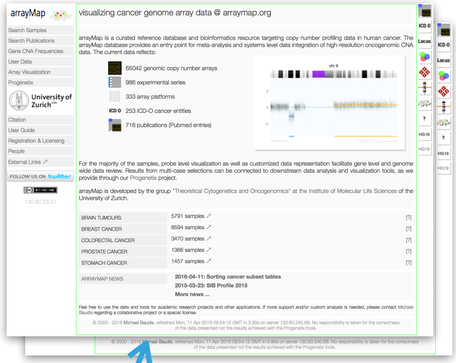
DATA PIPELINE

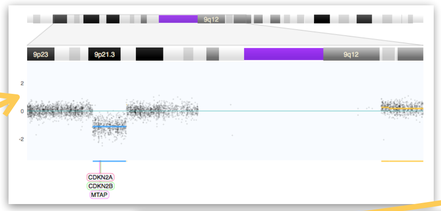


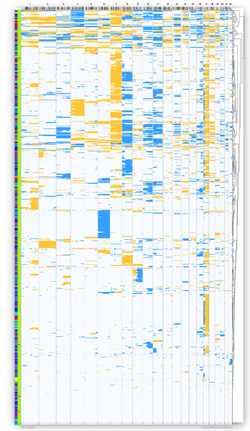


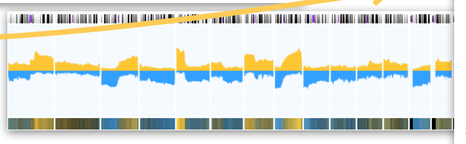


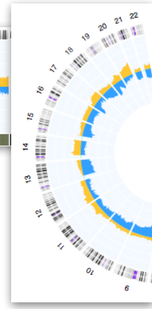


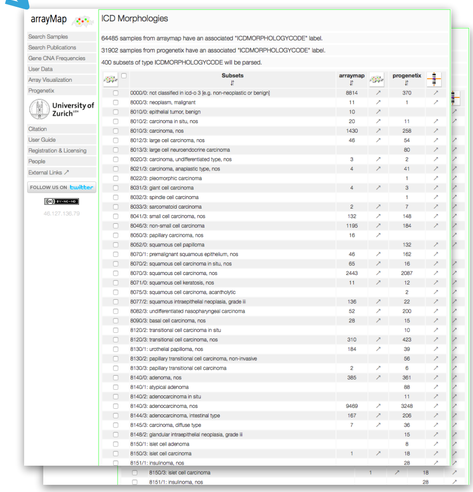








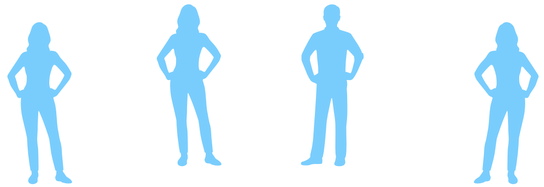




arrayMap

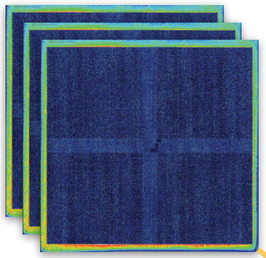
progenetix

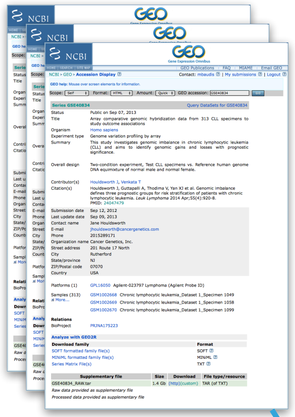
DATA PIPELINE

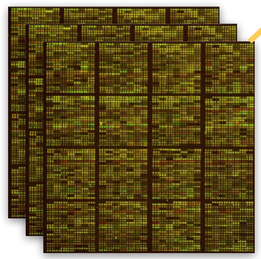


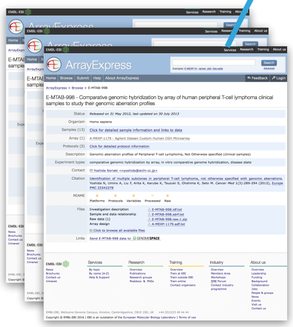
BIOCURATION

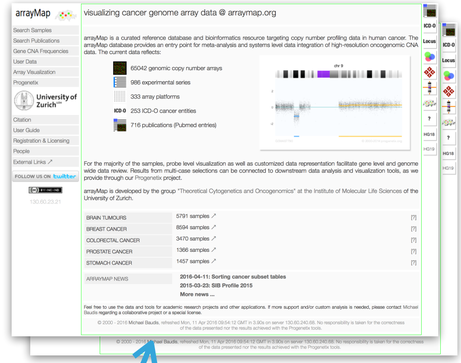
BIOINFORMATICS

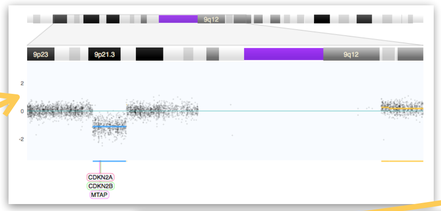


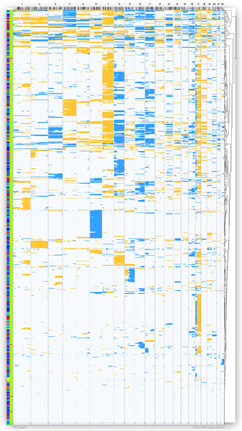


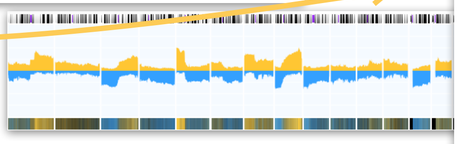


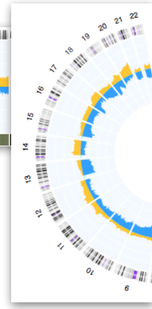


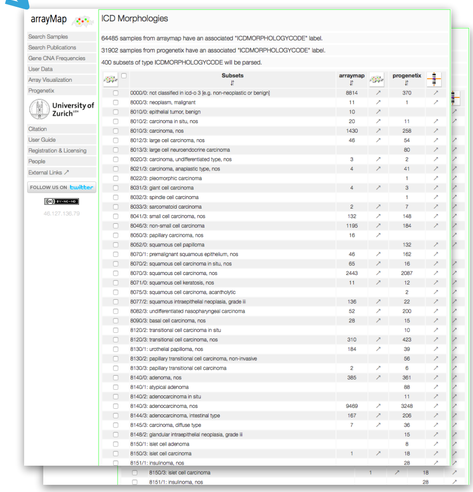












arrayMap

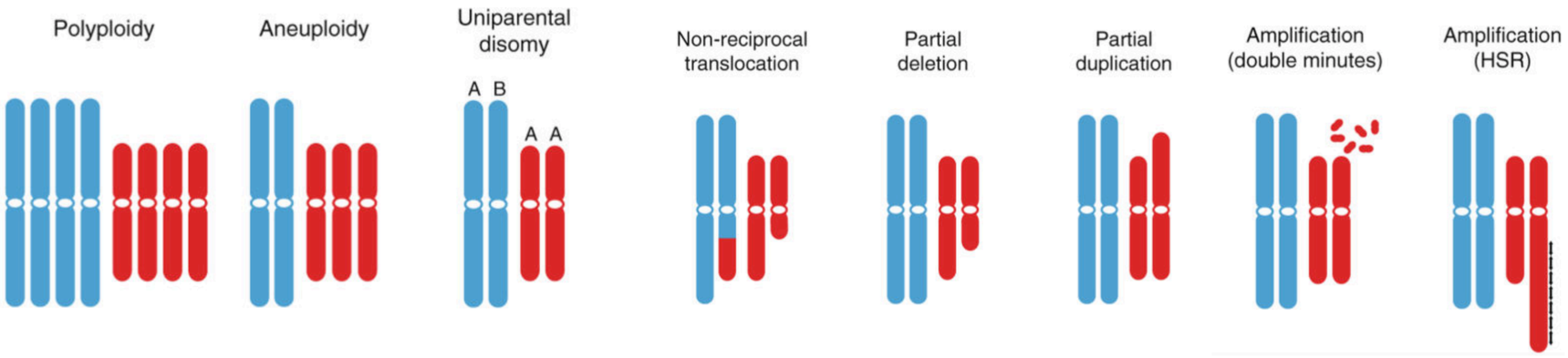
progenetix

Introduction

Types of genomic alterations in Cancer

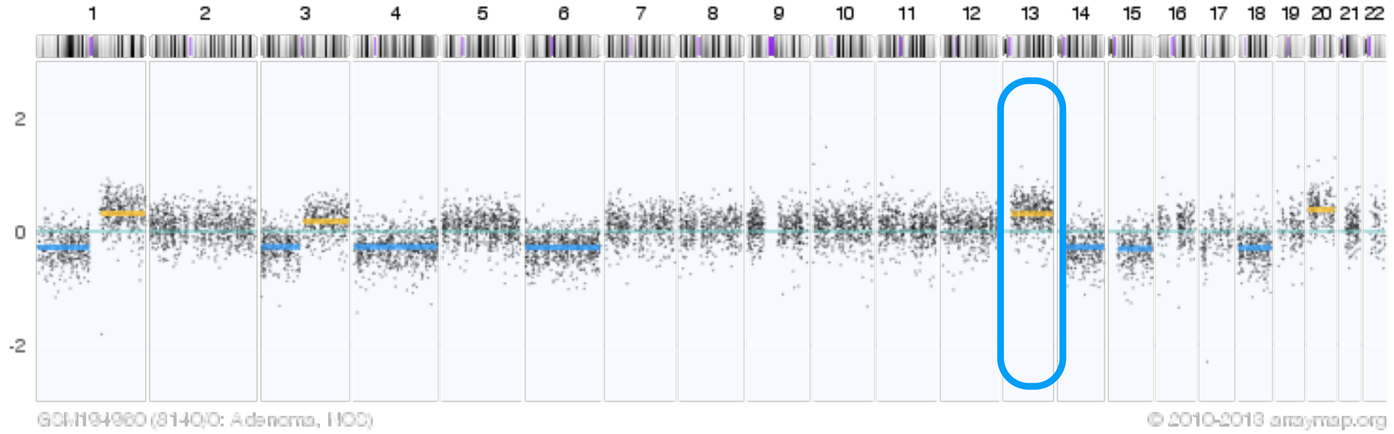
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)

Imbalanced Chromosomal Changes: CNV

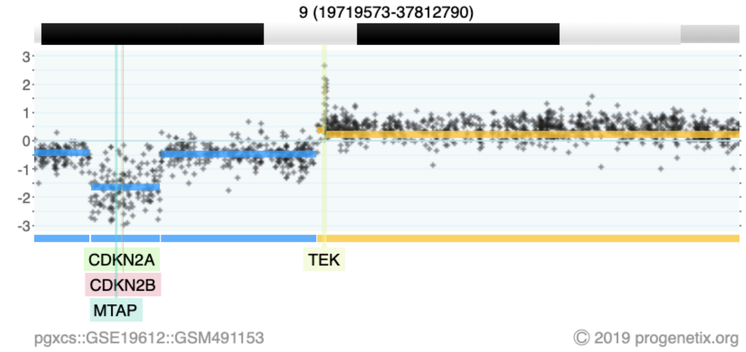


Grade et al., 2015 Recent Results Cancer Res

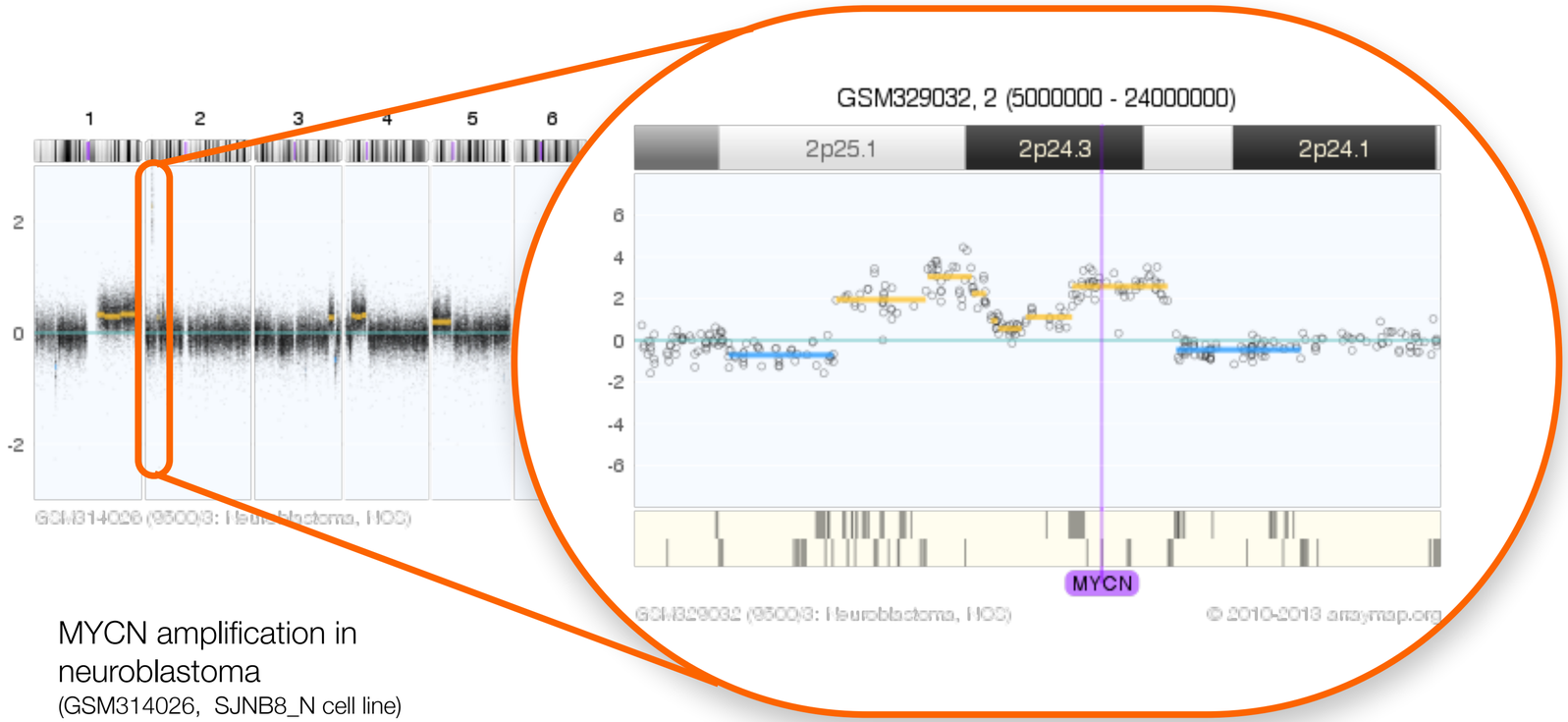
Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma

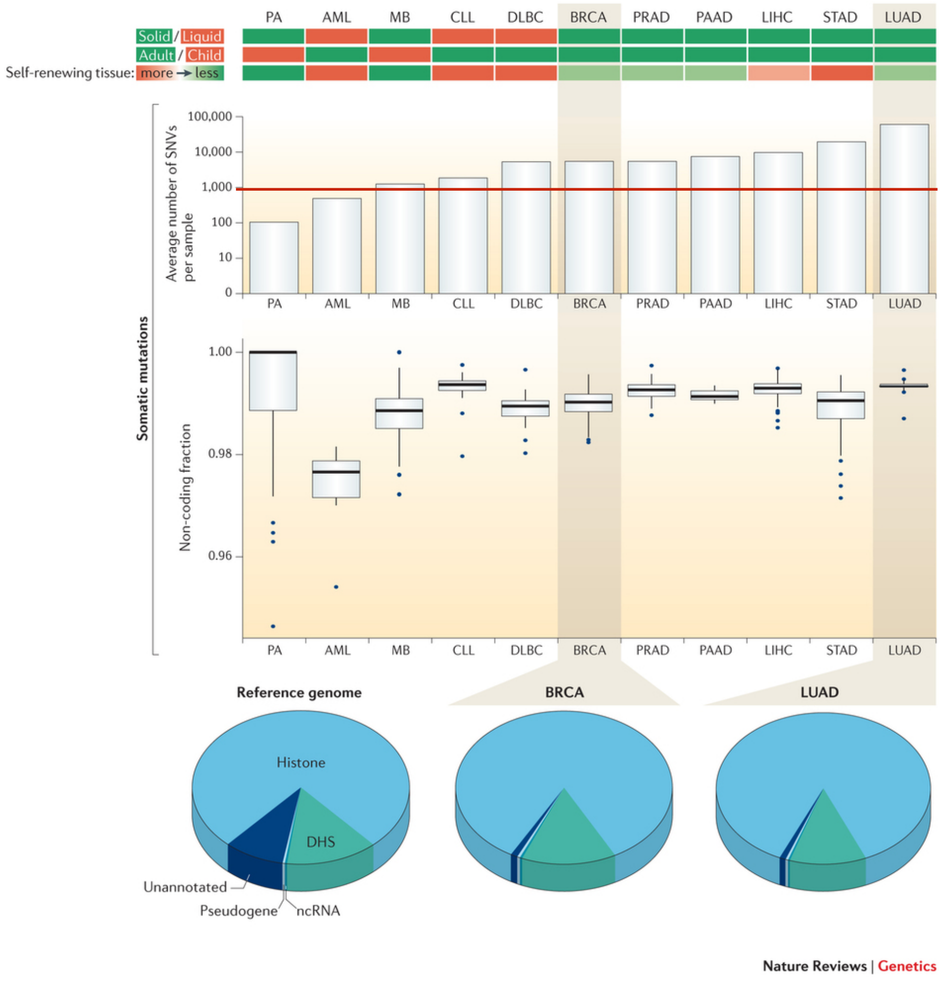


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

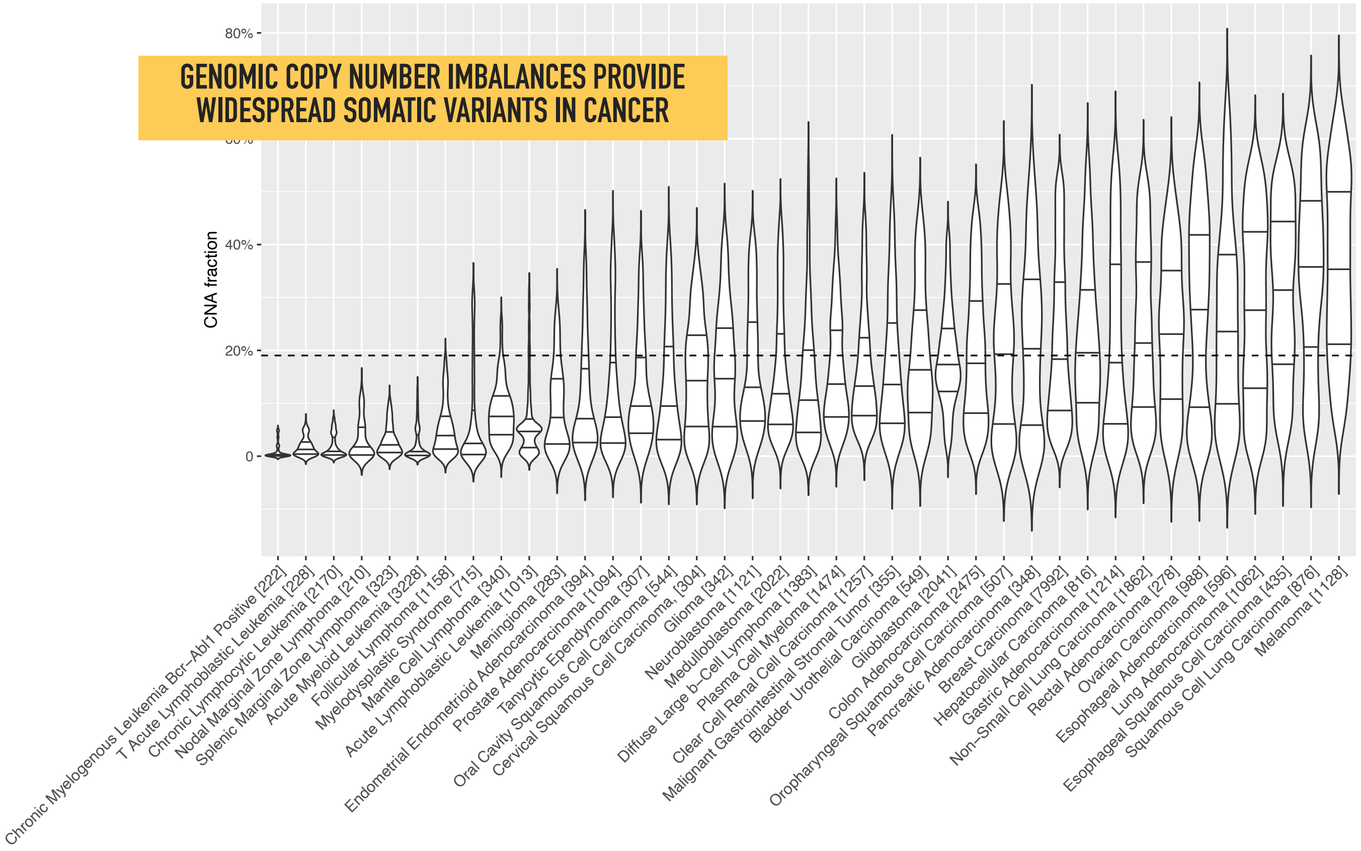
low level/**high level** copy number alterations (CNAs)

arrayMap

Quantifying Somatic Mutations In Cancer



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



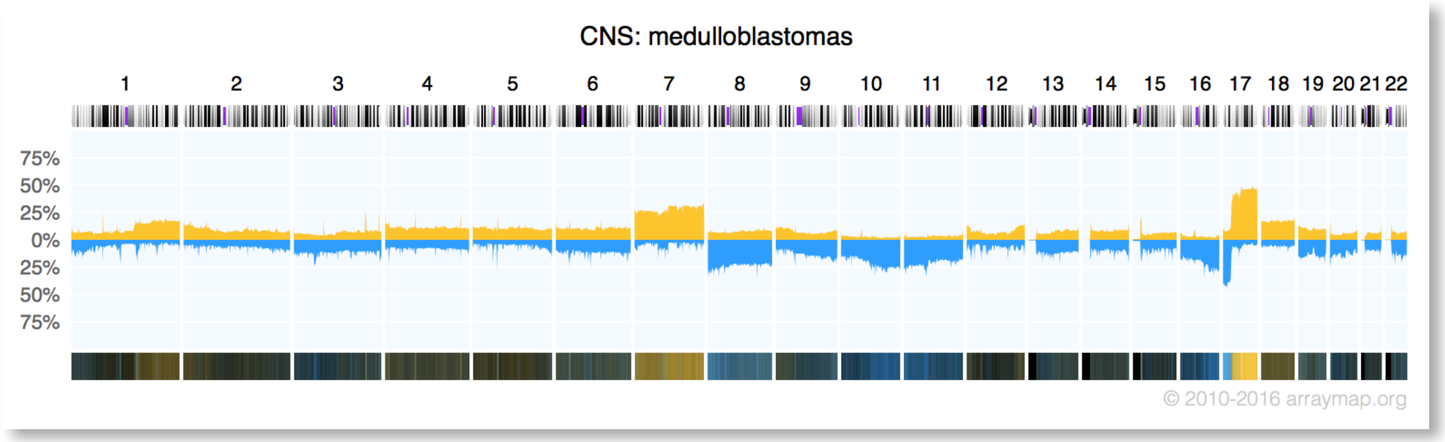
Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)

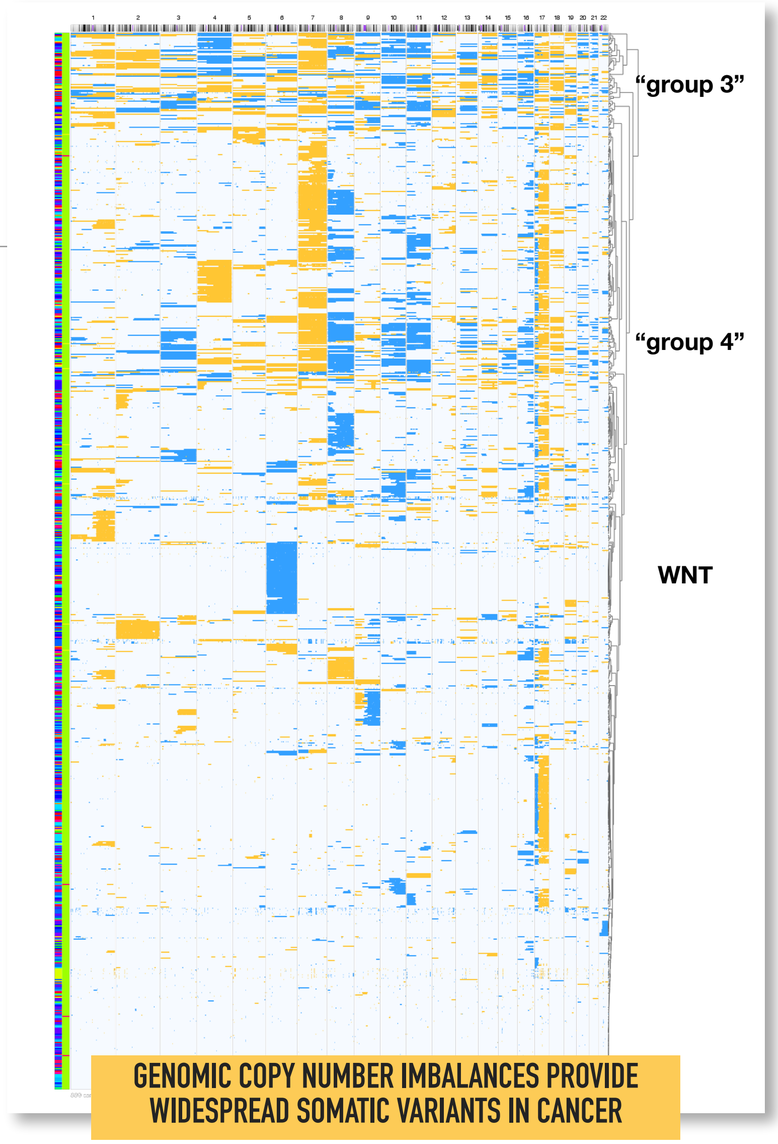
On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from progenetix.org

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?



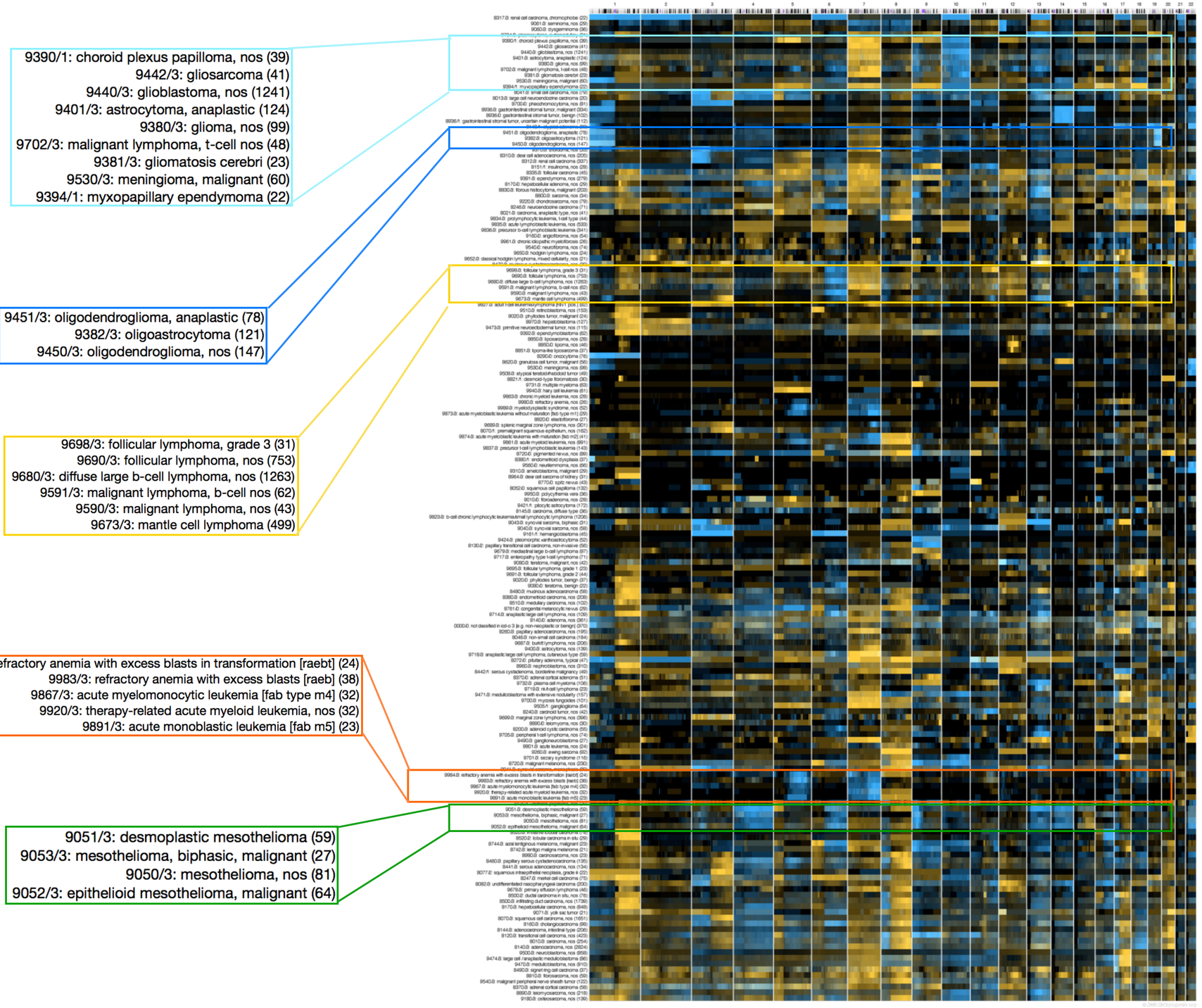


A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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

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

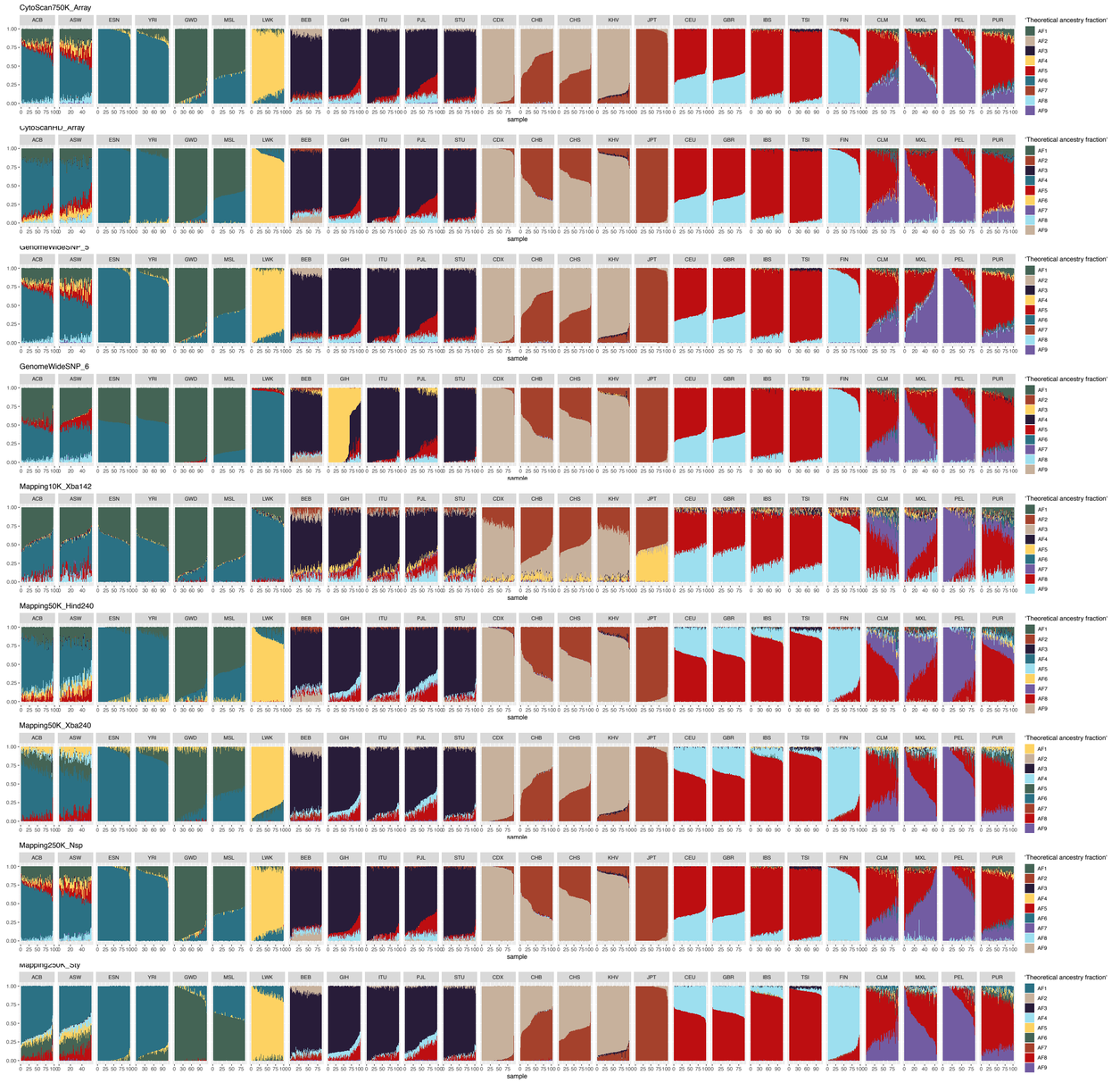


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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

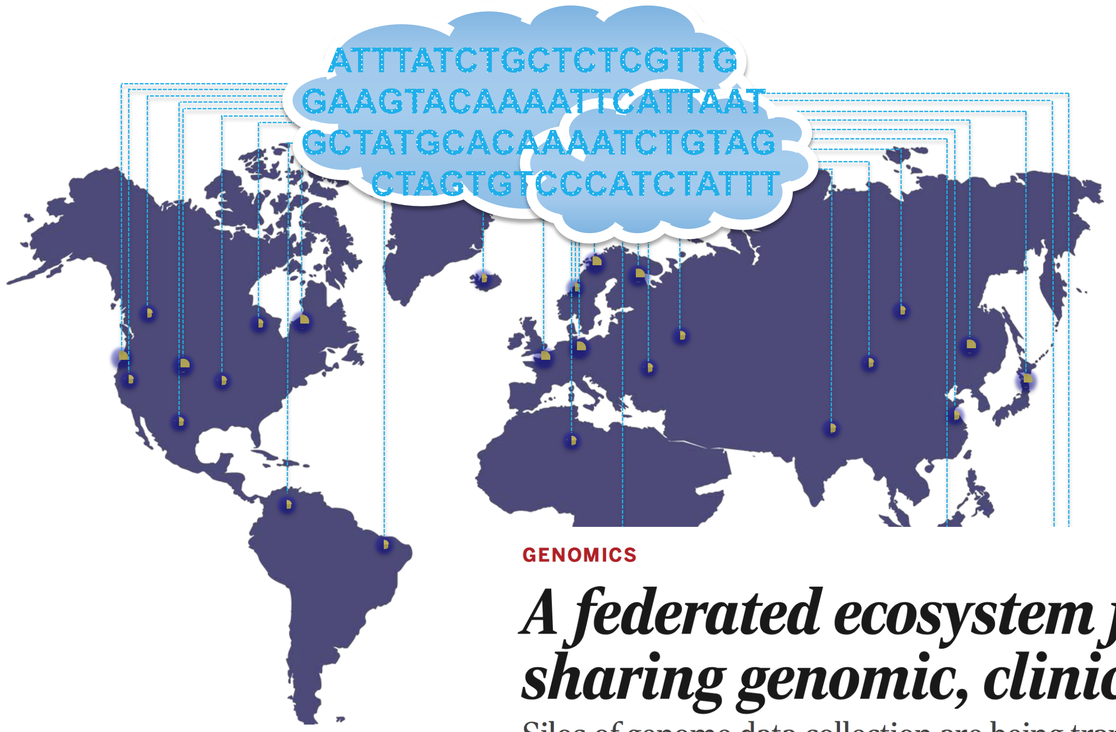
arrayMap

Qingyao Huang

University of Zurich[UZH]

Department of Molecular Life Sciences

Genome Data Sharing: The Global Alliance for Genomics and Health (GA4GH)



GENOMICS

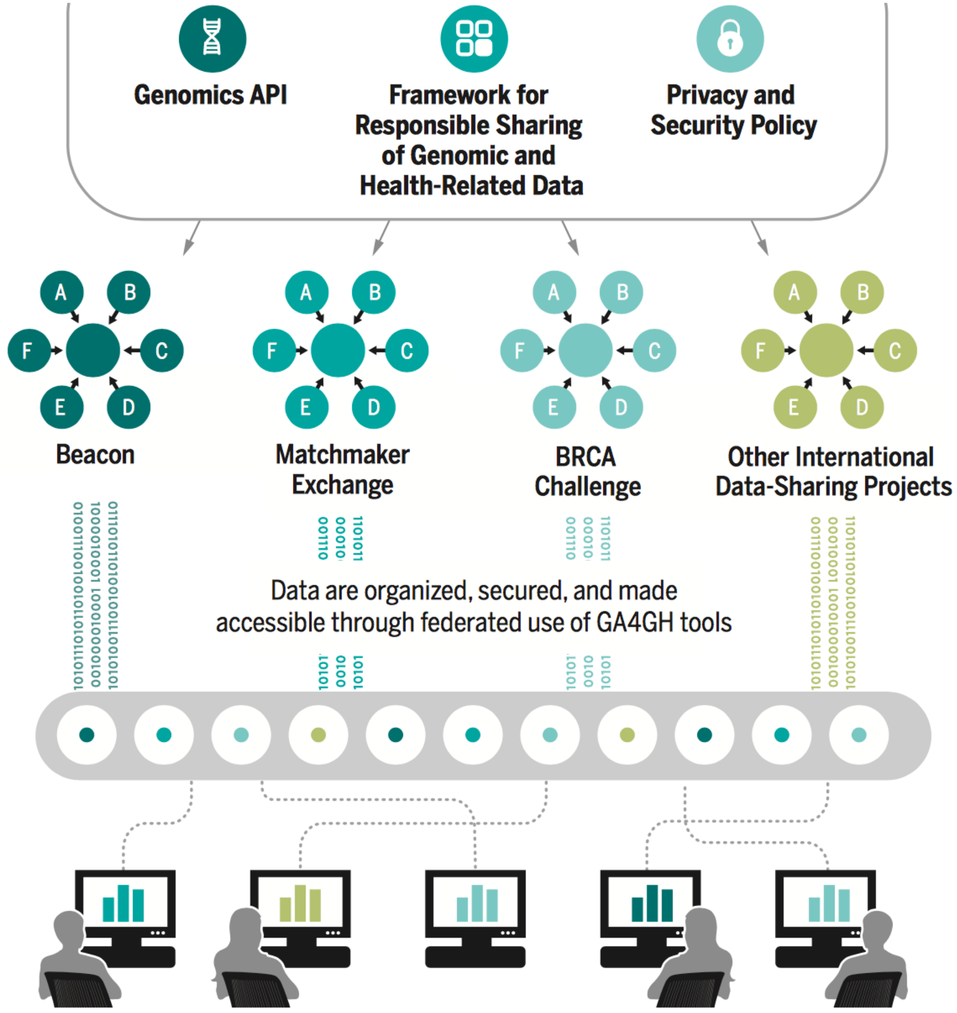
A federated ecosystem for sharing genomic, clinical data

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.



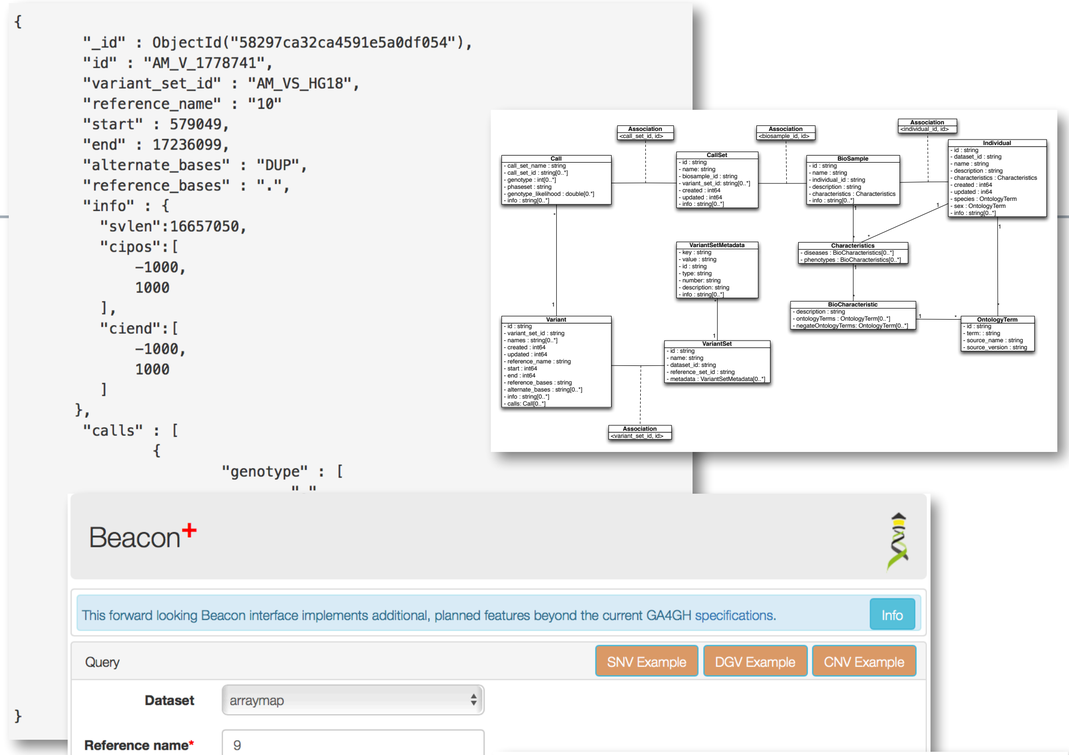
Global Alliance for Genomics & Health

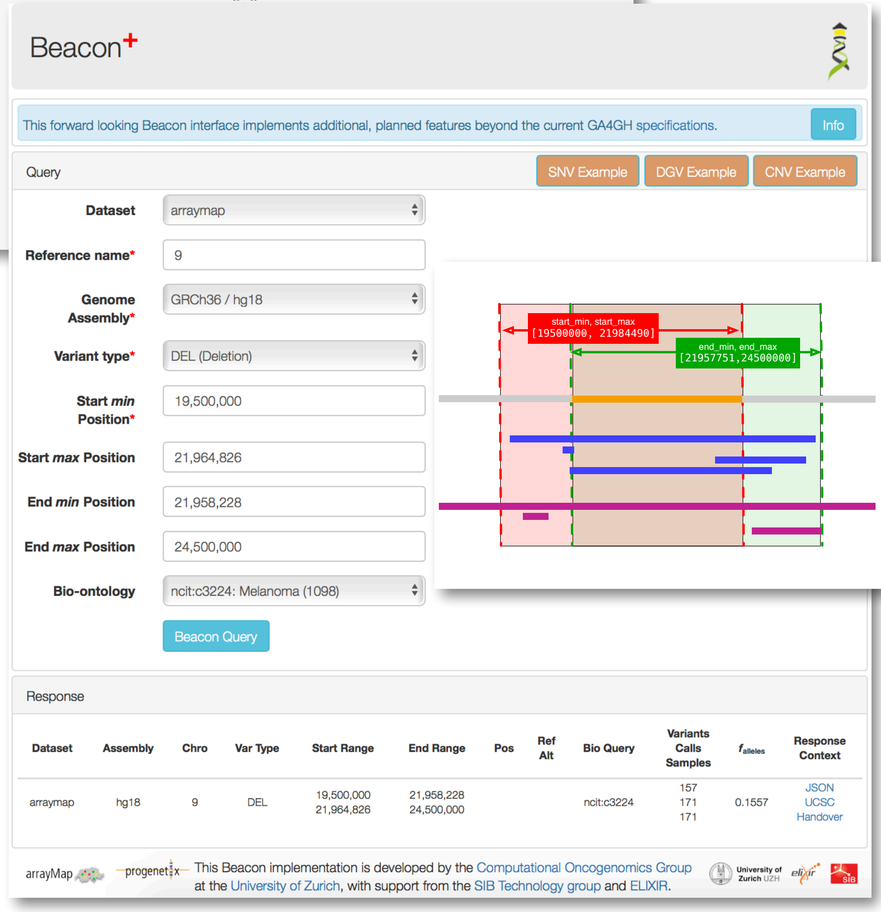
University of Zurich[UZH]

Department of Molecular Life Sciences

Our contributions II: Developing Genomic Knowledge Standards & Implementing Driver Projects

- Progenetix for data driven GA4GH development
  - **Metadata** schema development through implementation of Progenetix and arrayMap data (former ELIXIR project)
  - developing standards for structural genome variants as part of the **ELIXIR h-CNV** and **Beacon** initiatives
  - **SchemaBlocks** objects(w/ EBI, BBOP @ LBNL ...)
- Beacon+
  - Beacons - a GA4GH driver project - are **web services** for single-entry genome variant queries over >200 genome resources worldwide.
  - We are developing Beacon+ protocols for **structural variants**, **metadata** and "**handoff**" retrieval protocols (co-lead ELIXIR Beacon).
  - Enabling **Implementation** driven development using our and external datasets (arrayMap, TCGA, DIPG …)





Global Alliance for Genomics & Health

SIB Swiss Institute of Bioinformatics

arrayMap

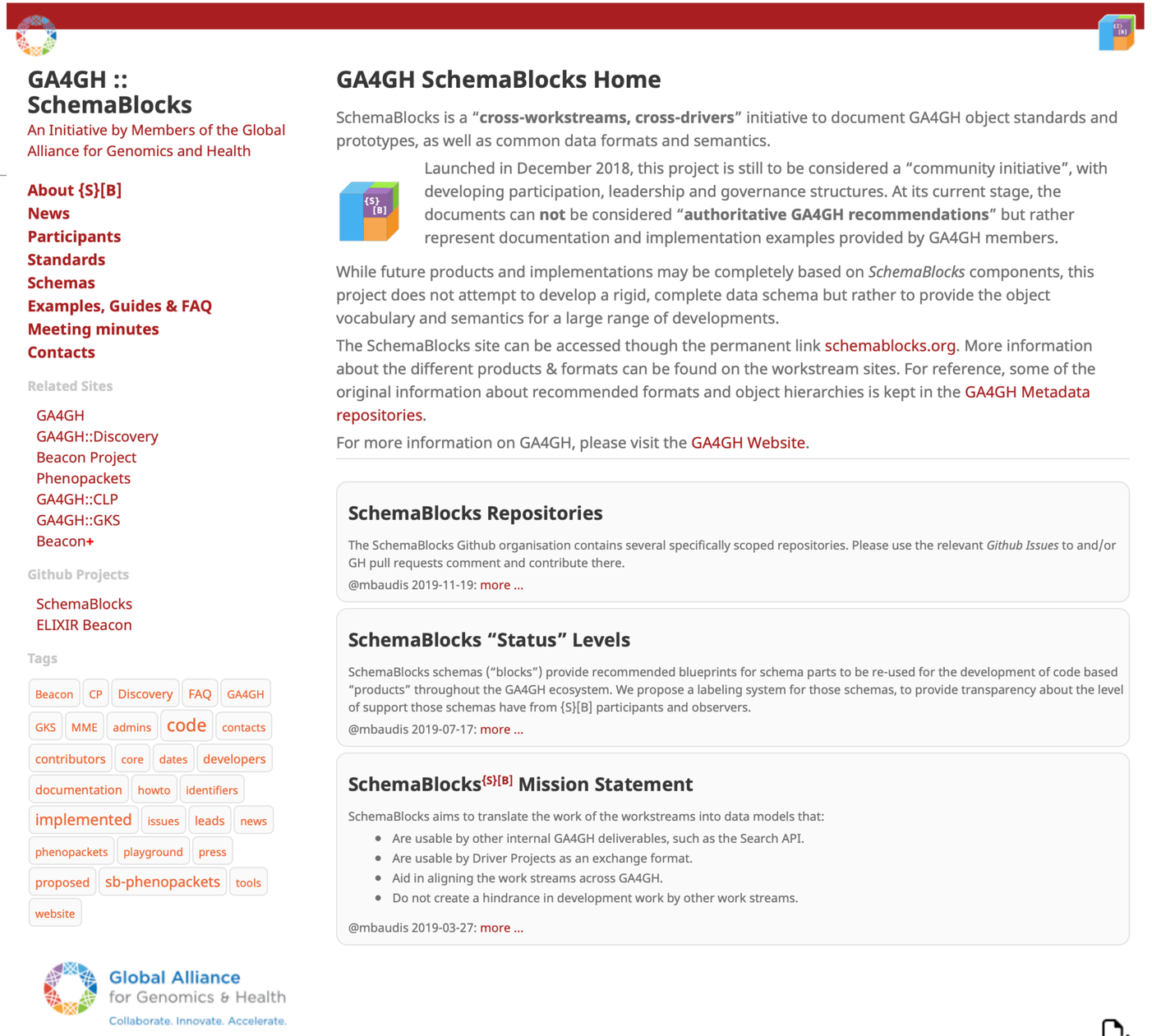
elixir

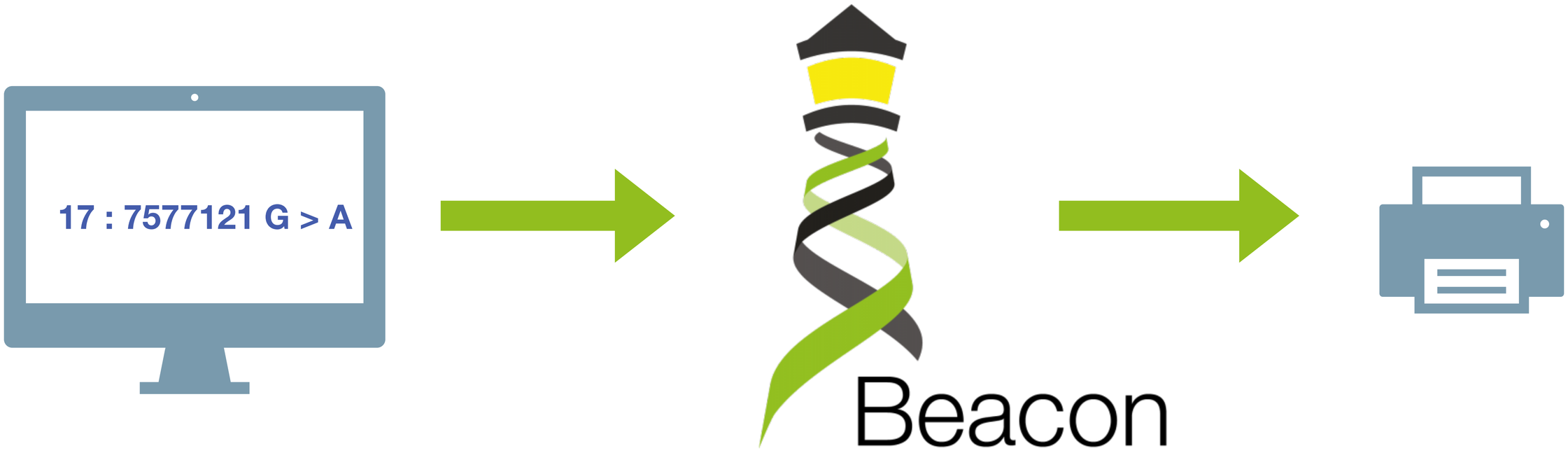
GA4GH {S}[B] *SchemaBlocks*

Standardized formats and data schemas for developing an "Internet of Genomics"

- "cross-workstreams, cross-drivers" initiative to document GA4GH object ***standards*** and ***prototypes***
- launched in December 2018
- documentation and implementation examples provided by GA4GH members
- not a rigid, complete data schema
- object ***vocabulary*** and ***semantics*** for a large range of developments
- recognized in **GA4GH roadmap** as element in "TASC" effort

schemablocks.org

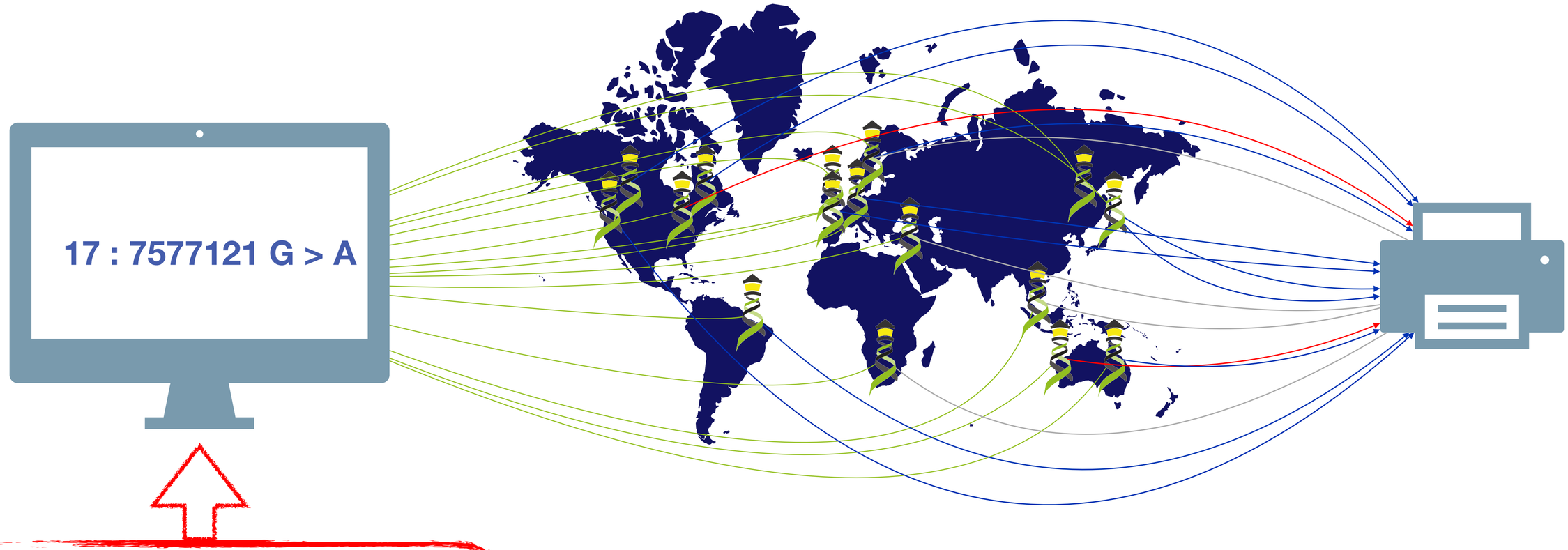




A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

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

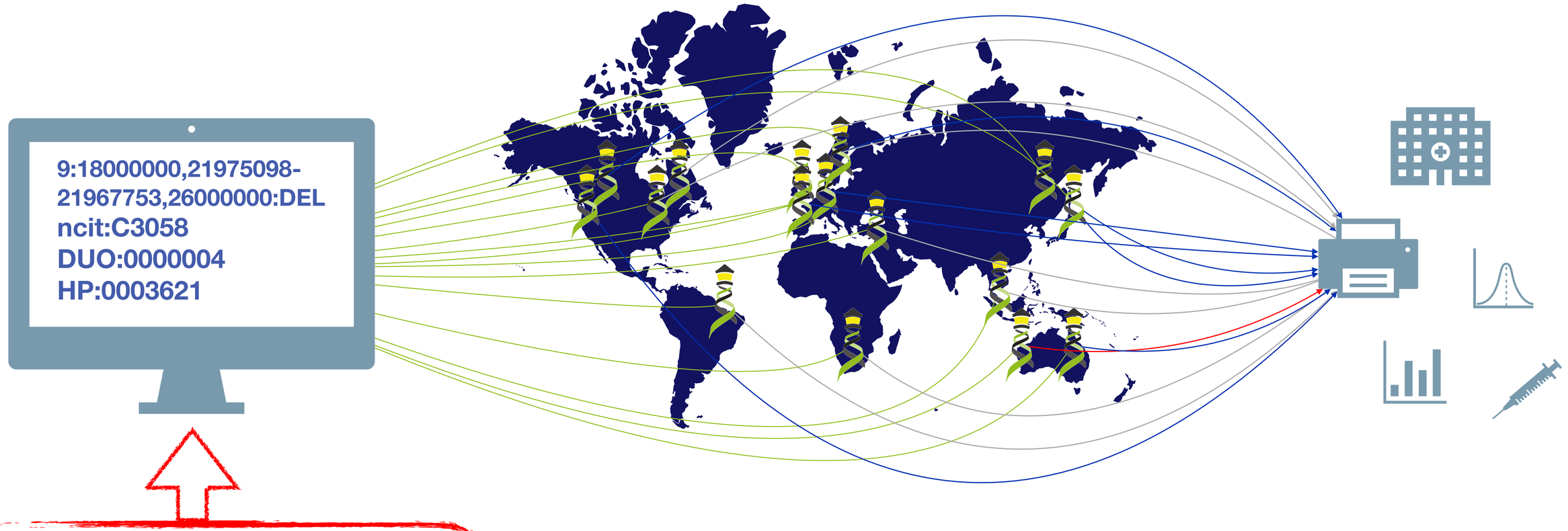
A Beacon network federates
genome variant queries
across databases that
support the ***Beacon API***



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***

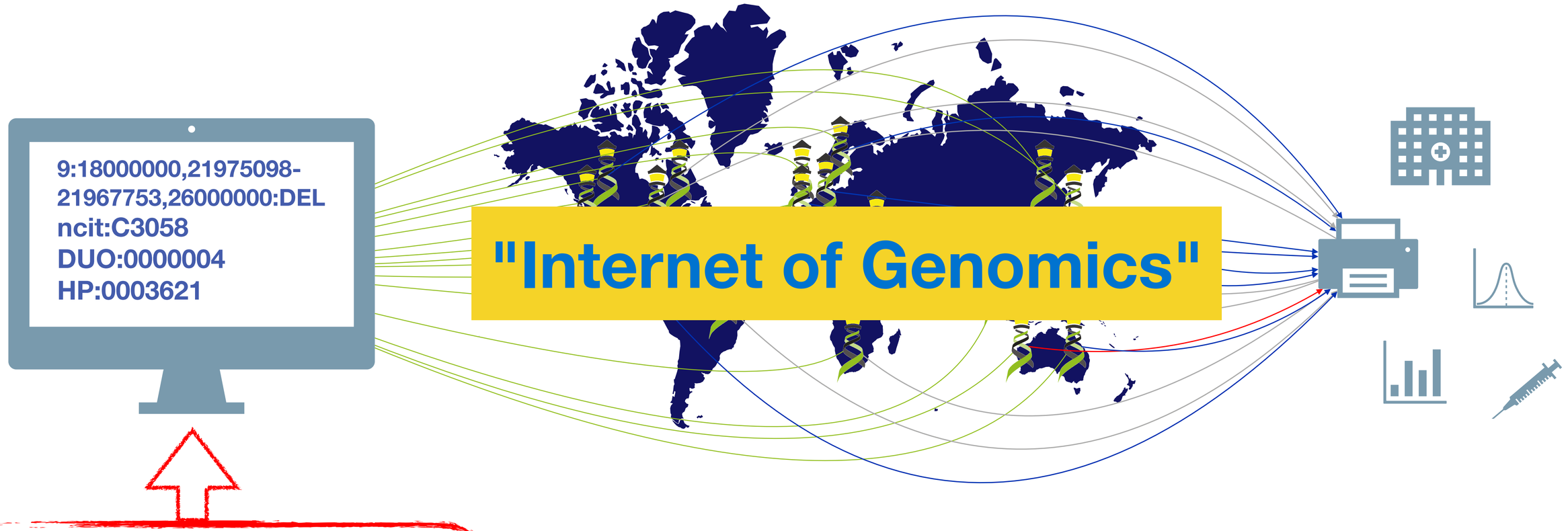
Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.



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

Beacon *v2* API

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



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

Beacon *v2* API

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

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

@WORLD

MATTHIAS ALTMEYER
THOMAS EGGERMANN
ROSA NOGUERA
REINER SIEBERT
CAIUS SOLOVAN

University of Zurich[UZH]

SIB

elixir

Global Alliance for Genomics & Health

BBOP GROUP @ LNBL

CHRIS MUNGALL
JUSTIN REESE
DEEPAK UNNI

GA4GH

MELANIE COURTOT
MELISSA HAENDEL
HELEN PARKINSON

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

University of Zurich[UZH]

Department of Molecular Life Sciences

Prof. Dr. Michael Baudis
Department 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

Global Alliance
for Genomics & Health