



University of  
Zurich<sup>UZH</sup>

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



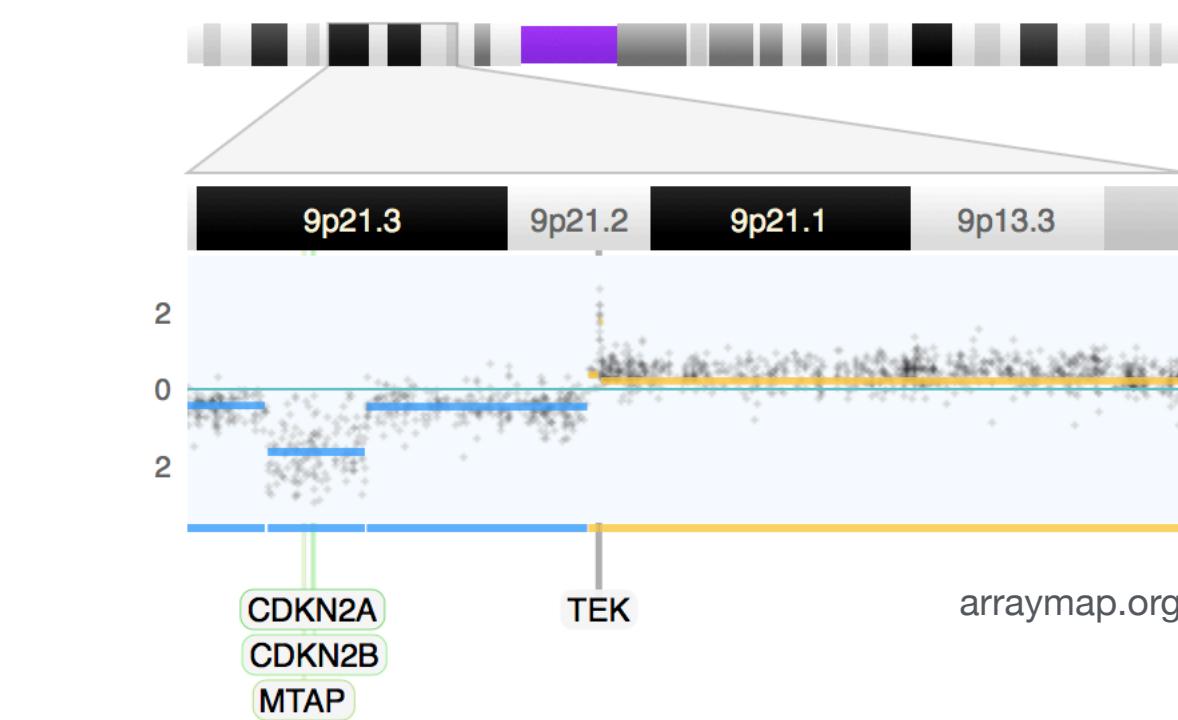
**Global Alliance**  
for Genomics & Health





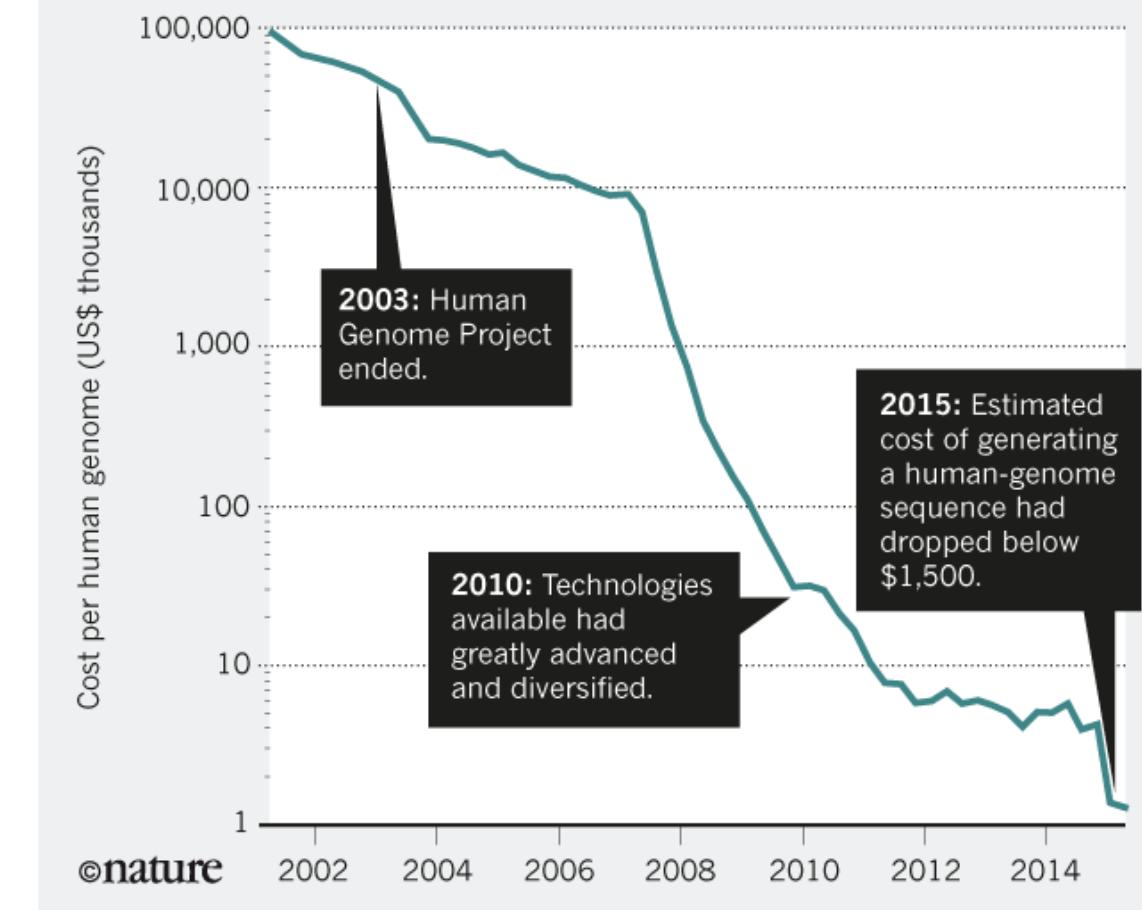
## Genome screening at the core of “Personalised Health”

- ▶ **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**

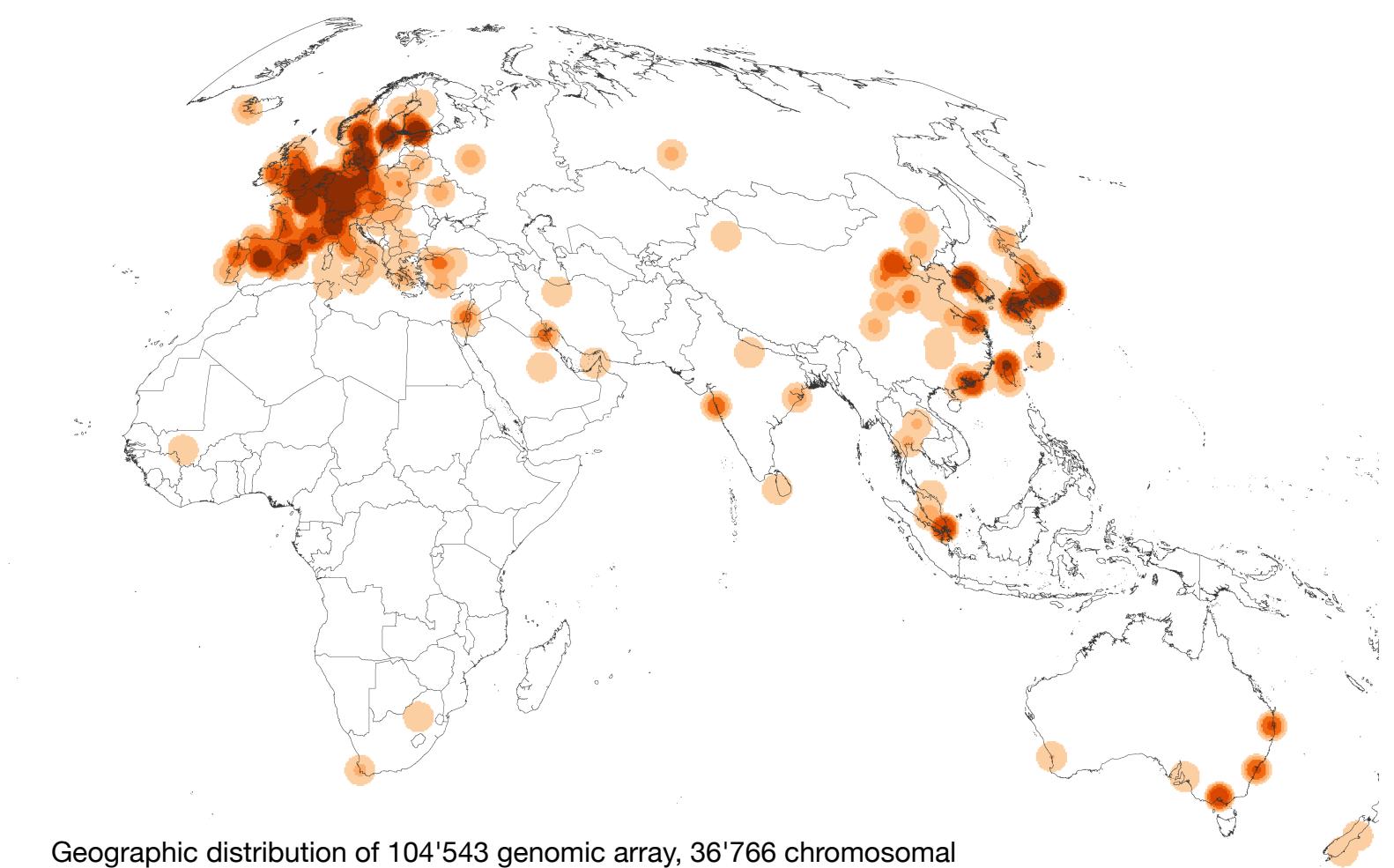
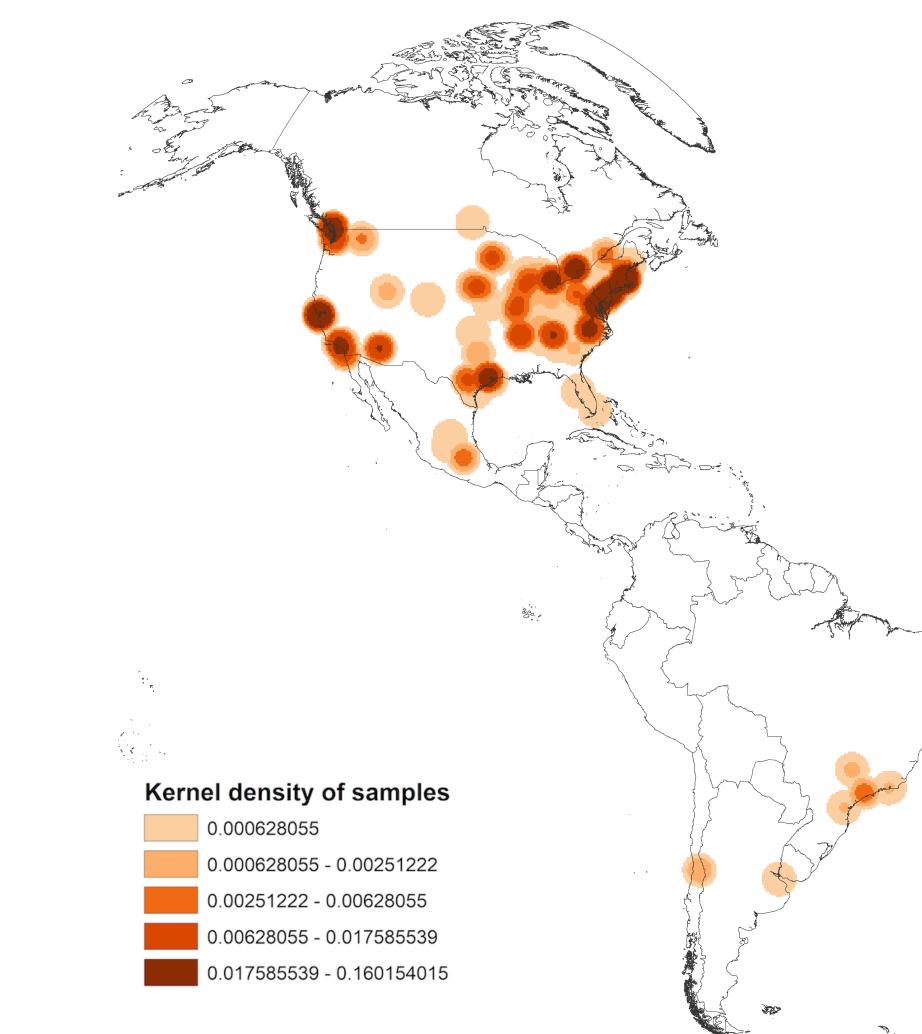


### BETTER, CHEAPER, FASTER

The cost of DNA sequencing has dropped dramatically over the past decade, enabling many more applications.



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



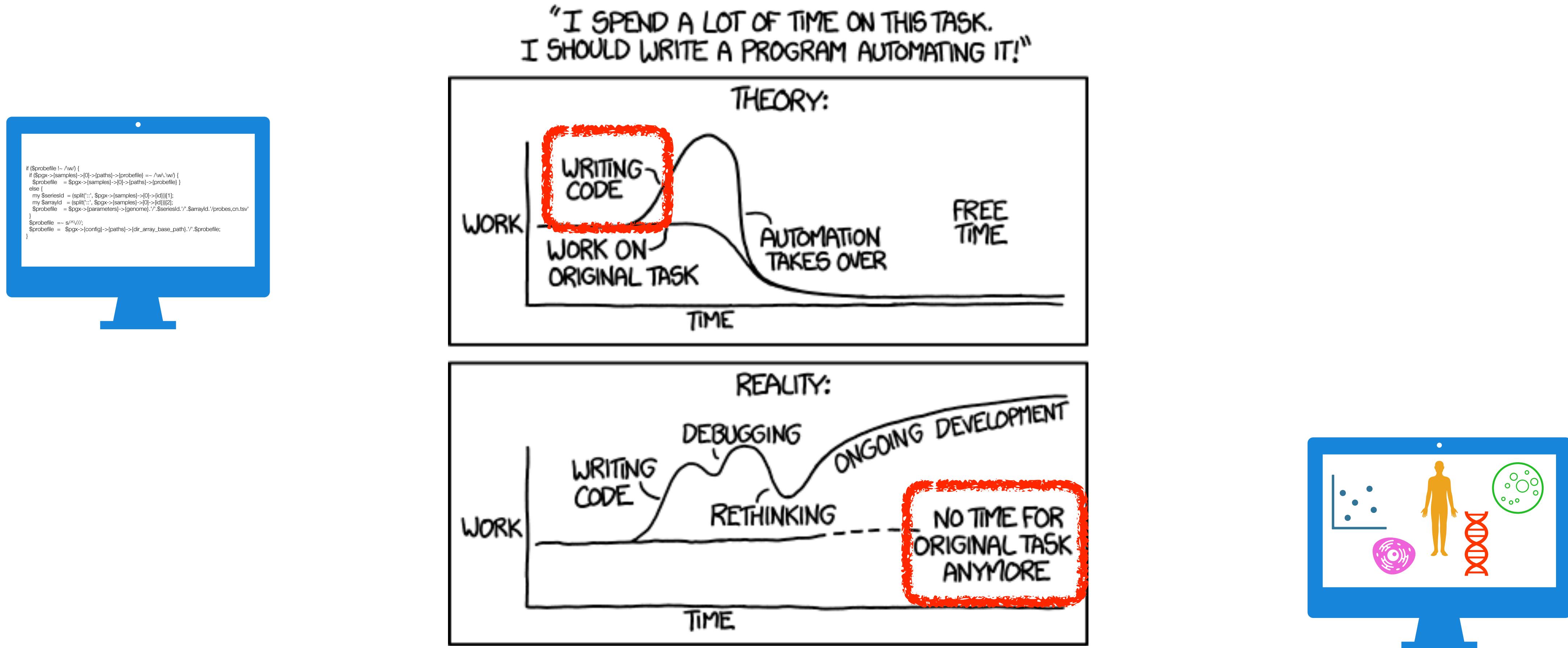
Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

# {bio\_informatics\_science}

---



# {bio\_informatics\_science}





## Our contributions I: Cancer genome knowledge resources and research



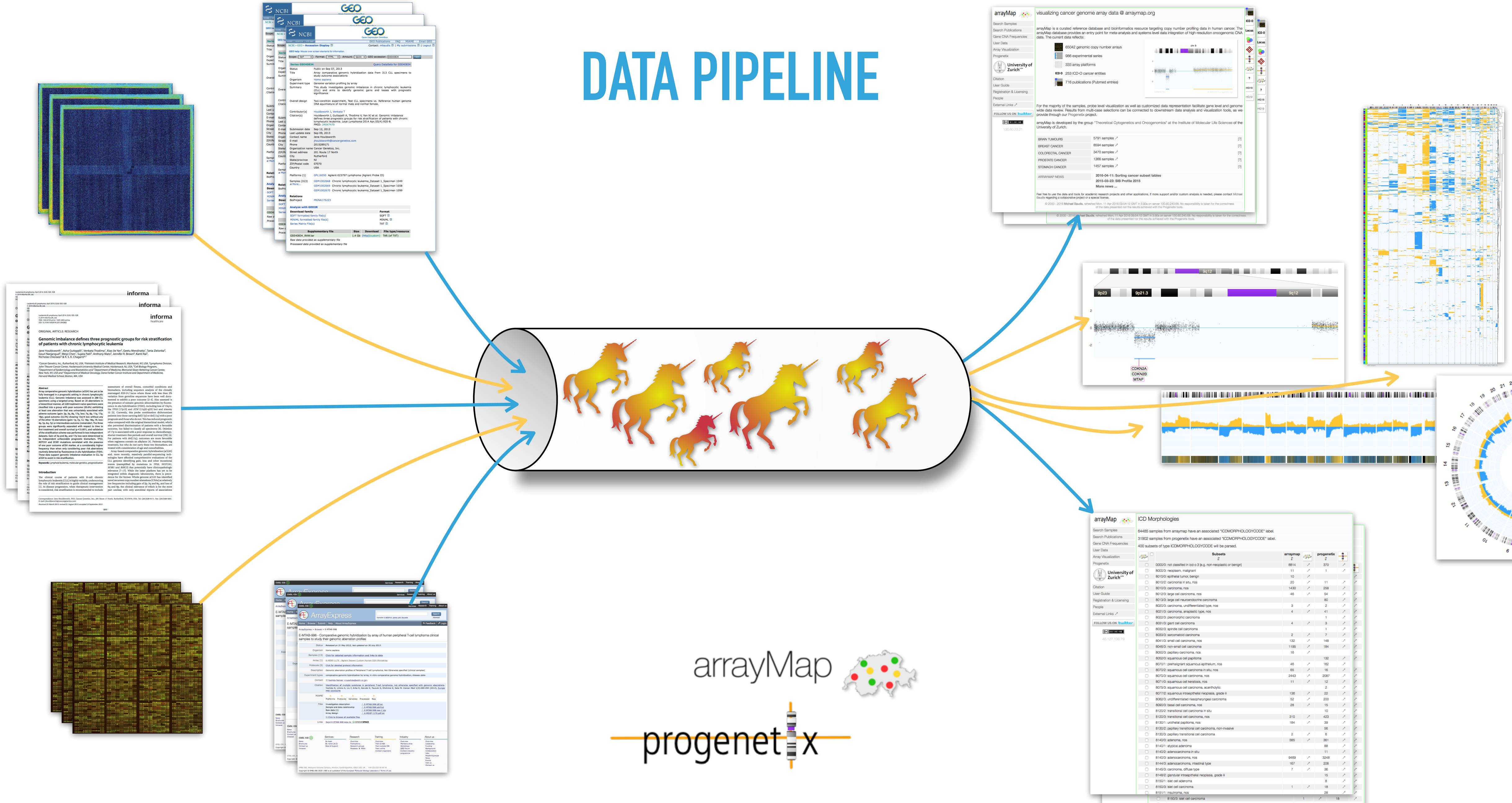
- ▶ curated cancer genome publication resource (more than 3200 manually curated articles)
- ▶ article metadata
- ▶ annotated genome profiles of >100'000 samples
- ▶ ontology mappings; clinical data where available
- ▶ epistemology: geographic and histologic sampling biases



- ▶ 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



# DATA PIPELINE

## BIOCURATION BIOINFORMATICS



NCBI GEO Accession Display

Series GSE640034 Public on Sep 07, 2013

Organism: Human

Experiment type: Genomic variation profiling by array

Summary: This study investigates genomic variation in chronic lymphocytic leukemia (CLL) specimens with prognostic significance.

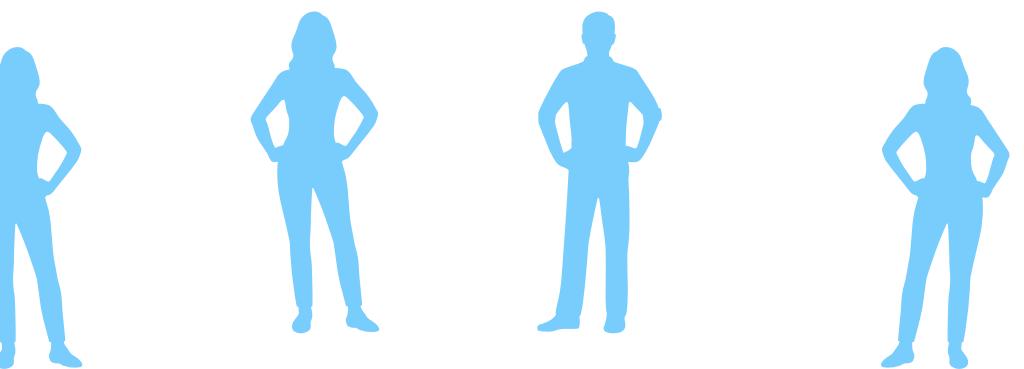
Overall design: Overall design experiment, Test vs. Specimens vs. Reference human genome

Contributor(s): Houldsworth J, Venkata T, Guttagji A, Thoduri V, Yan XI et al.

Sample ID: GSE640034

Platform: Agilent G1317P Lymphoma (Agilent Probe ID)

Supplementary file: GSE640034.RAW.tar



arrayMap

visualizing cancer genome array data at 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 integration of high-resolution oncogenomic DNA data. The current data reflects:

- 65024 genomic copy number arrays
- 985 experimental series
- 333 array platforms
- 253 ICD-O cancer entities
- 716 publications (PubMed entries)

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

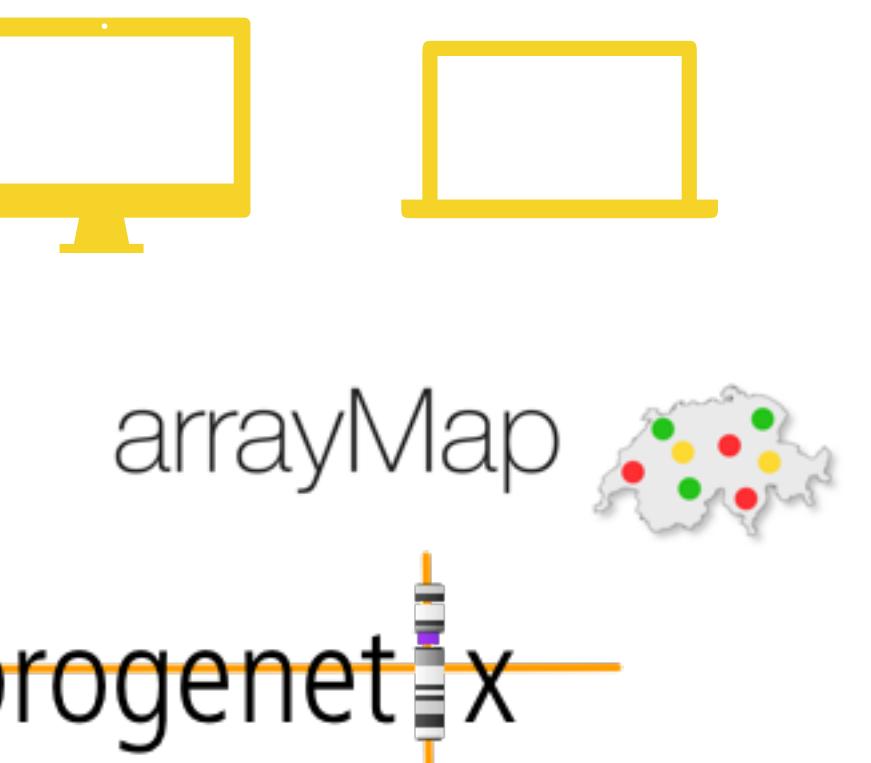
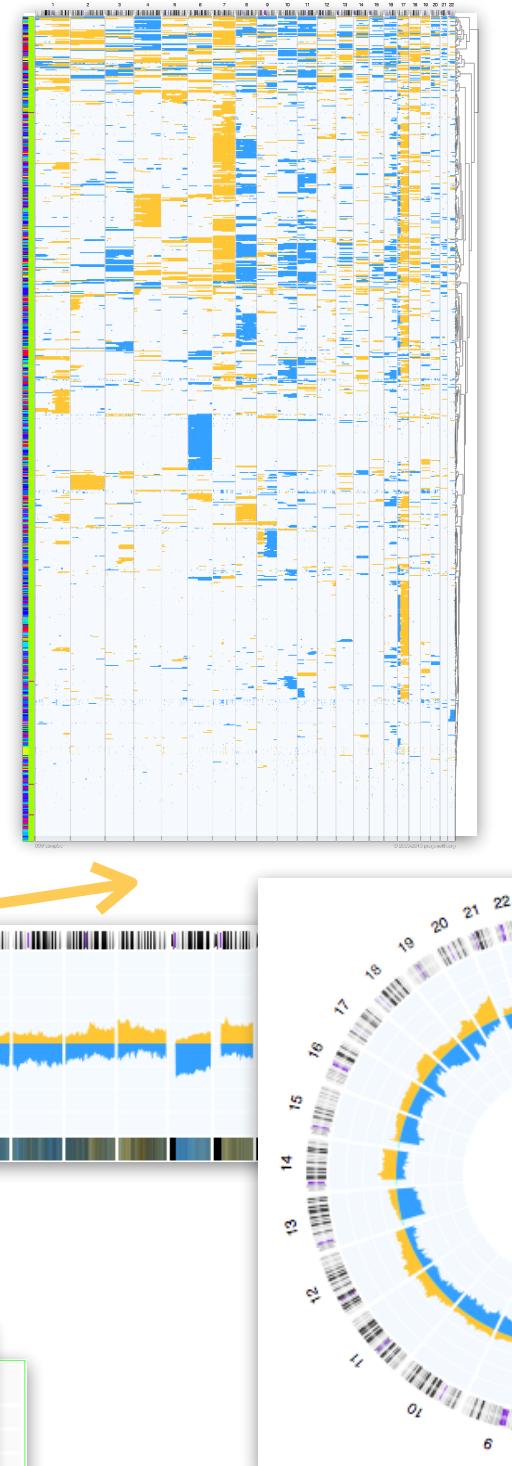
arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

Platforms (1): GPR100, Agilent G1317P Lymphoma (Agilent Probe ID)

Supplementary file: GSE640034.RAW.tar

Raw data provided as supplementary file

Processed data provided as supplementary file



informa healthcare

Leukemia & Lymphoma April 2014 95(4): 500-508

ORIGINAL ARTICLE RESEARCH

Genomic imbalance defines three prognostic groups for risk stratification of patients with chronic lymphocytic leukemia

Jane Houldsworth<sup>1</sup>, Asha Guttapalli<sup>1</sup>, Venkata Thoduri<sup>1</sup>, Xiao Jie Yan<sup>1</sup>, Geeta Mendrekar<sup>1</sup>, Tamja Zelenka<sup>2</sup>, Gouri Nangisetty<sup>2</sup>, Wei Chen<sup>2</sup>, Supratik Pati<sup>2</sup>, Anthony Mato<sup>2</sup>, Jennifer R. Brown<sup>2</sup>, Kari Rar<sup>2</sup>

<sup>1</sup>Cancer Genetics, Inc., Rutherford, NJ, USA; <sup>2</sup>Weinstein Institute of Medical Research, Manhattan, NY, USA; <sup>3</sup>Lymphoma Division, Department of Hematology and Oncology, Department of Epidemiology and Biostatistics, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY, USA; <sup>4</sup>Department of Pathology, Department of Oncology, David Hahn Cancer Institute and Department of Medicine, Harvard Medical School, Boston, MA, USA

Abstract

Genomic imbalance (GIM) has been fully leveraged in a prognostic setting in chronic lymphocytic leukemia (CLL). We used array-based genomic hybridization to identify genomic imbalances in CLL specimens using a targeted array. Based on 20 aberrations in each specimen, we identified 100% of patients with CLL who had a gain or loss of at least 4 Mb. Gain or loss of 4 Mb was associated with first treatment and overall survival ( $P < 0.001$ ), and validation of these findings in an independent cohort of CLL specimens showed that gains of 4 Mb and 17 Mb loss were determined to be prognostic factors. In addition, we found that patients with CLL with TP53 (17q11) and ATM (11q13) gain and loss, respectively, had a shorter median free period and overall survival (OS) ( $P < 0.001$ ). TP53 and ATM mutations correlated with the presence of TP53 gain and ATM loss, respectively. These findings were confirmed when regressed on an allelic LOH. Patients requiring further risk stratification were identified by multivariate analysis, which included age and constitutional genes (ATM, TP53, and CDKN2A). These results support genomic imbalance evaluation in CLL by using a targeted array. The clinical utility of GIM in CLL is supported by the fact that TP53 and ATM are tumor-suppressor genes (TSGs) that potentially have cliniopathologic relevance. In addition, the prognostic value of GIM can be integrated within diagnostic laboratories, since it provides prognostic information without the need for specialized equipment and trained personnel. GIM can be used to predict survival in CLL and the clinical relevance of which is the most important factor in determining its prognostic value.

Keywords: chronic lymphocytic leukemia, molecular genetics, prognostication

Introduction

The clinical course of patients with B-cell chronic lymphocytic leukemia (CLL) is highly variable, undergoing disease progression, stabilization, or regression. At disease diagnosis, when therapeutic intervention is initiated, the disease is often asymptomatic and may progress slowly. These patients have a long life expectancy, and many recently, massively parallel sequencing techniques have been used to identify the location of the CLL genome, identifying genes, loci, and other mutations that are associated with the disease. The genes TP53, ATM, and CDKN2A (TSGs) that potentially have cliniopathologic relevance have been identified. These genes can be integrated within diagnostic laboratories, since it provides prognostic information without the need for specialized equipment and trained personnel. GIM can be used to predict survival in CLL and the clinical relevance of which is the most important factor in determining its prognostic value.

Array Express

E-MTAB-998 Comparative genomic hybridization array of human peripheral T-cell lymphoma clinical samples to study their genomic aberration profiles

Organism: Homo sapiens

Sample (13): Click for detailed sample information and links to data

Array (1): Agilent G1317P Human Custom Microarray

Description: Genomic aberration profiles of Peripheral T-cell Lymphoma, not otherwise specified (clinical sample)

Experiment type: comparative genomic hybridization array, ex vivo comparative genomic hybridization, disease state

Contact: Dr. Harita Kaur, christiansenlab@uab.edu

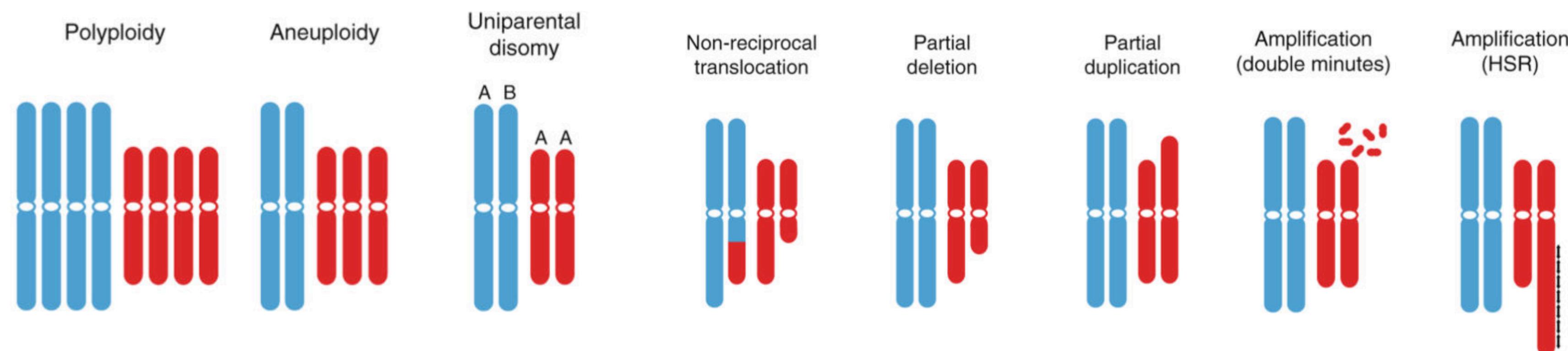
Links: E-MTAB-998-A.zip, E-MTAB-998-B.zip

# Introduction

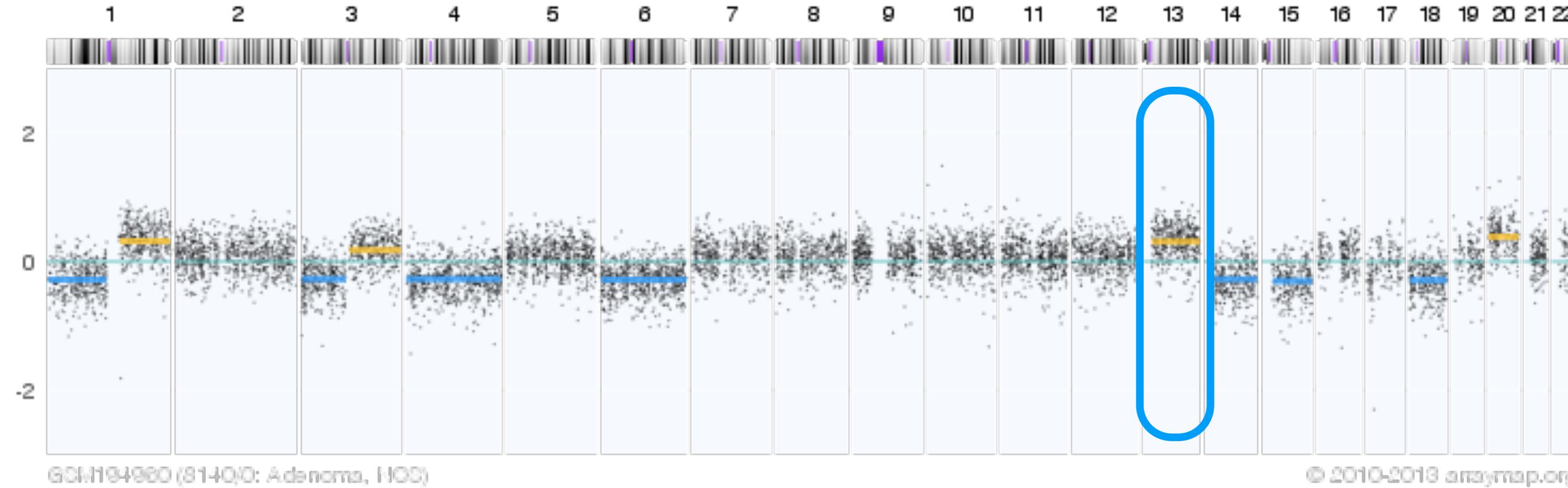
## Types of genomic alterations in Cancer

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

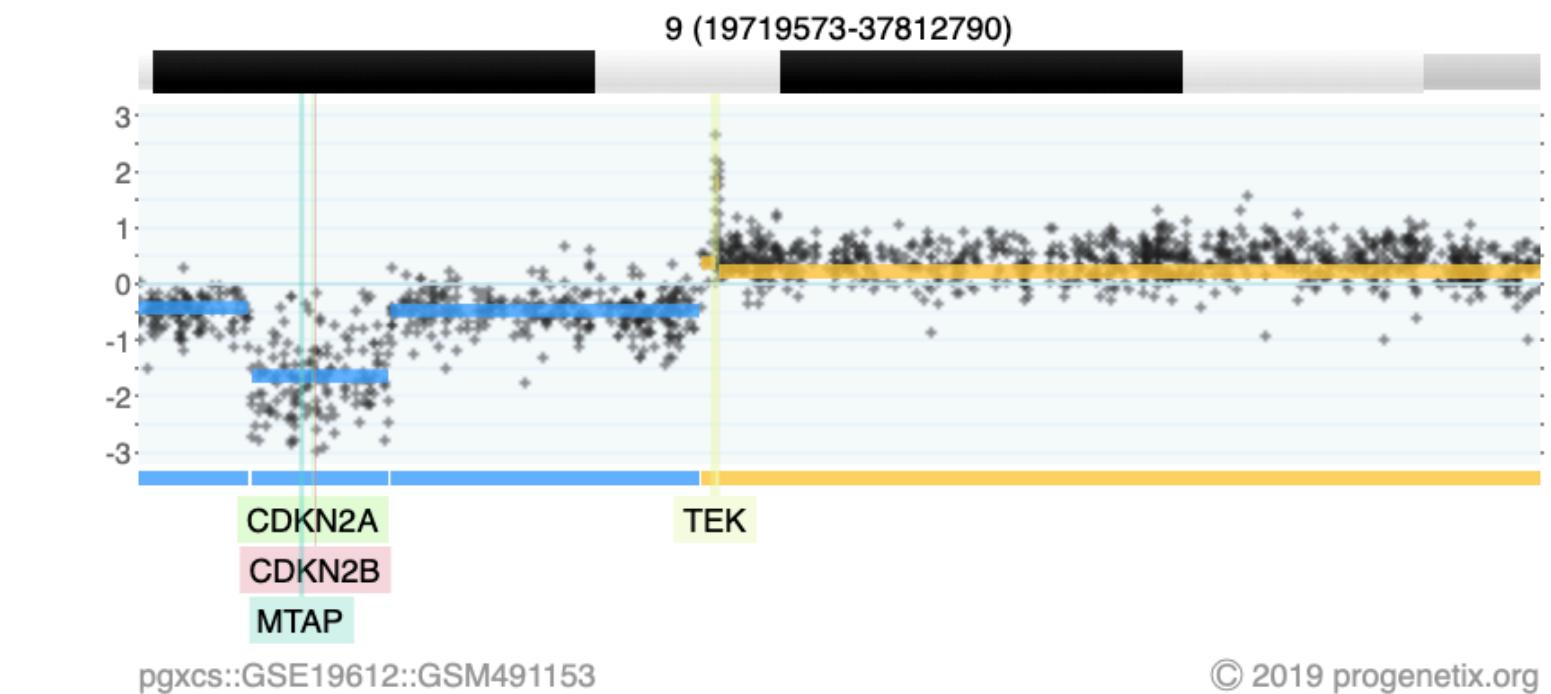
## Imbalanced Chromosomal Changes: CNV



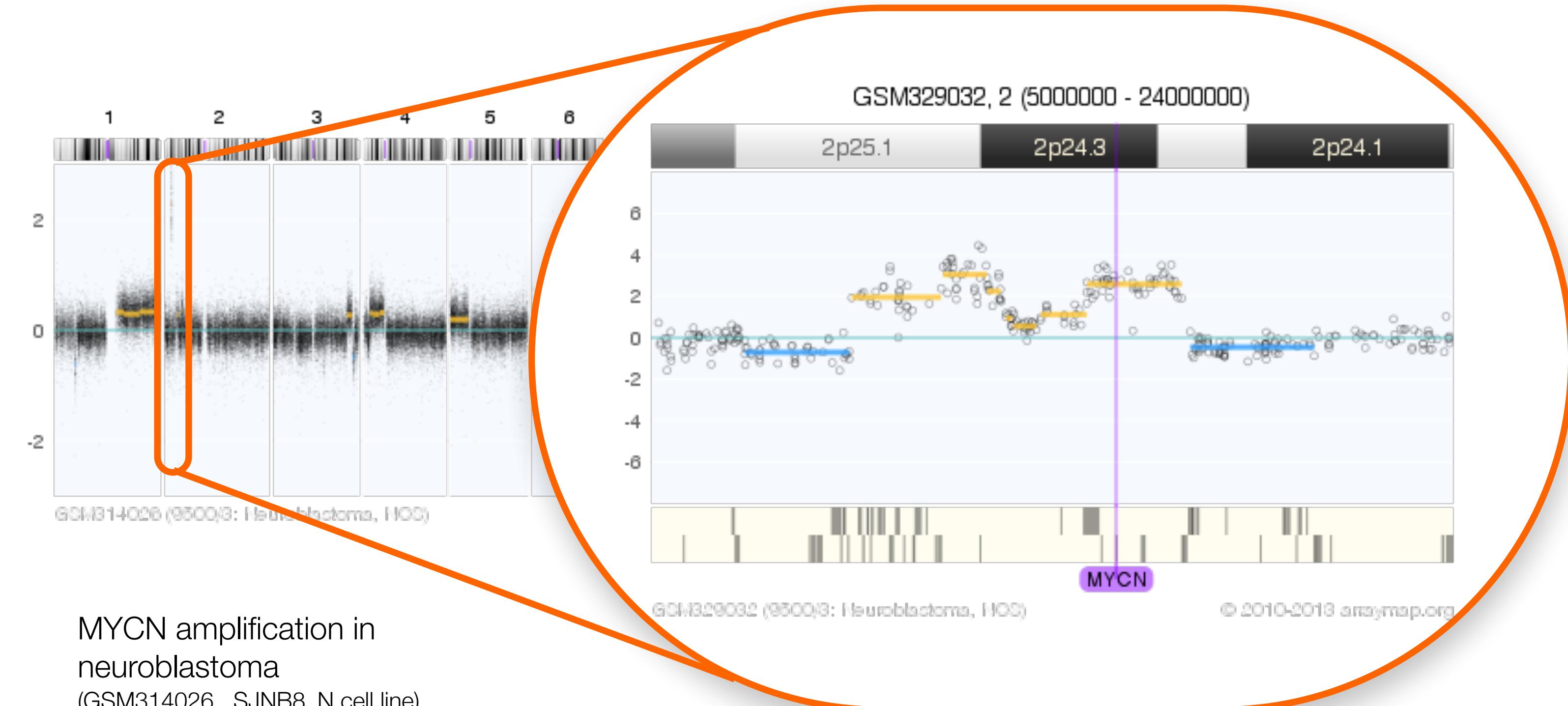
# Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma

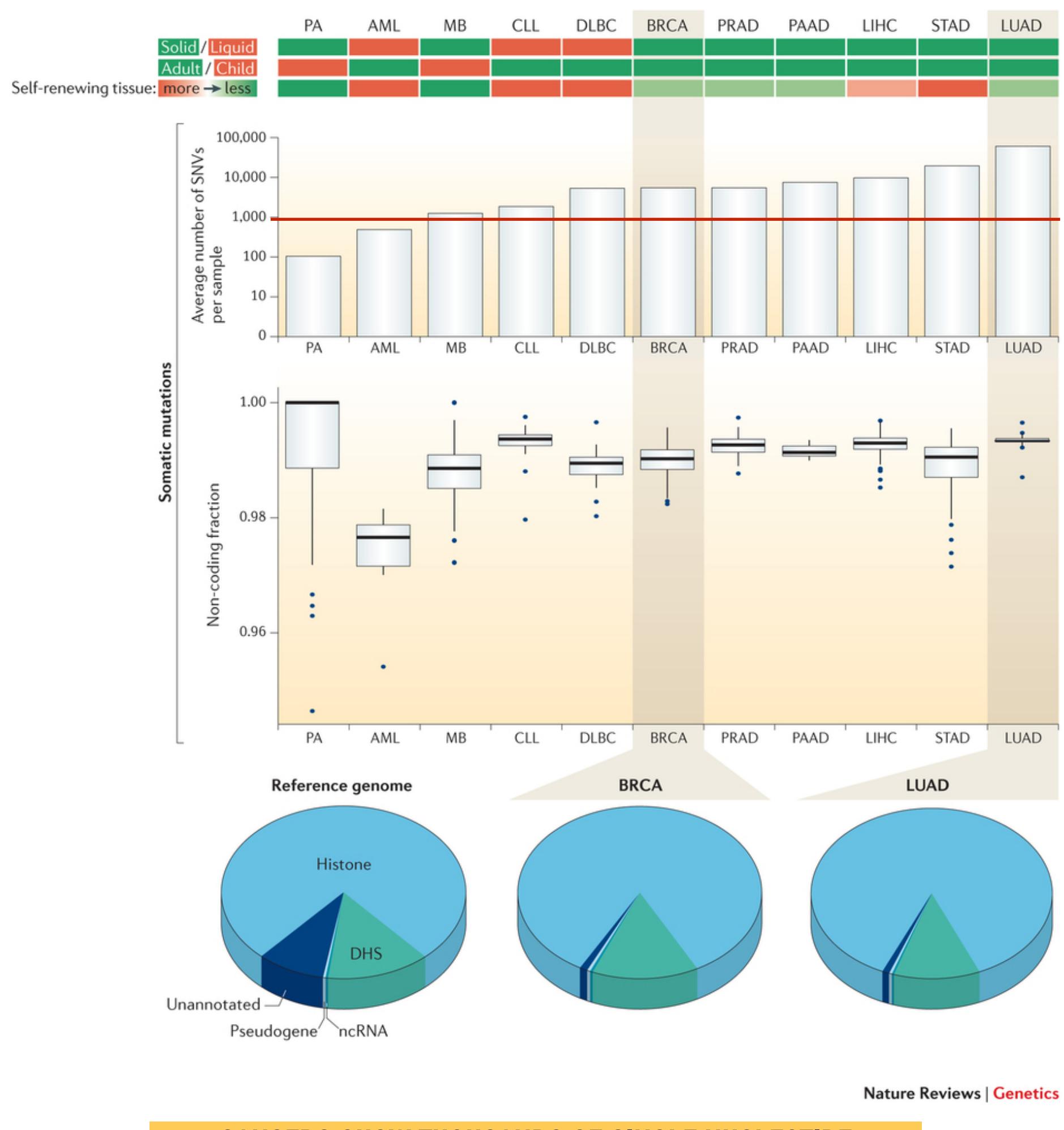


**low level/high level** copy number alterations (CNAs)

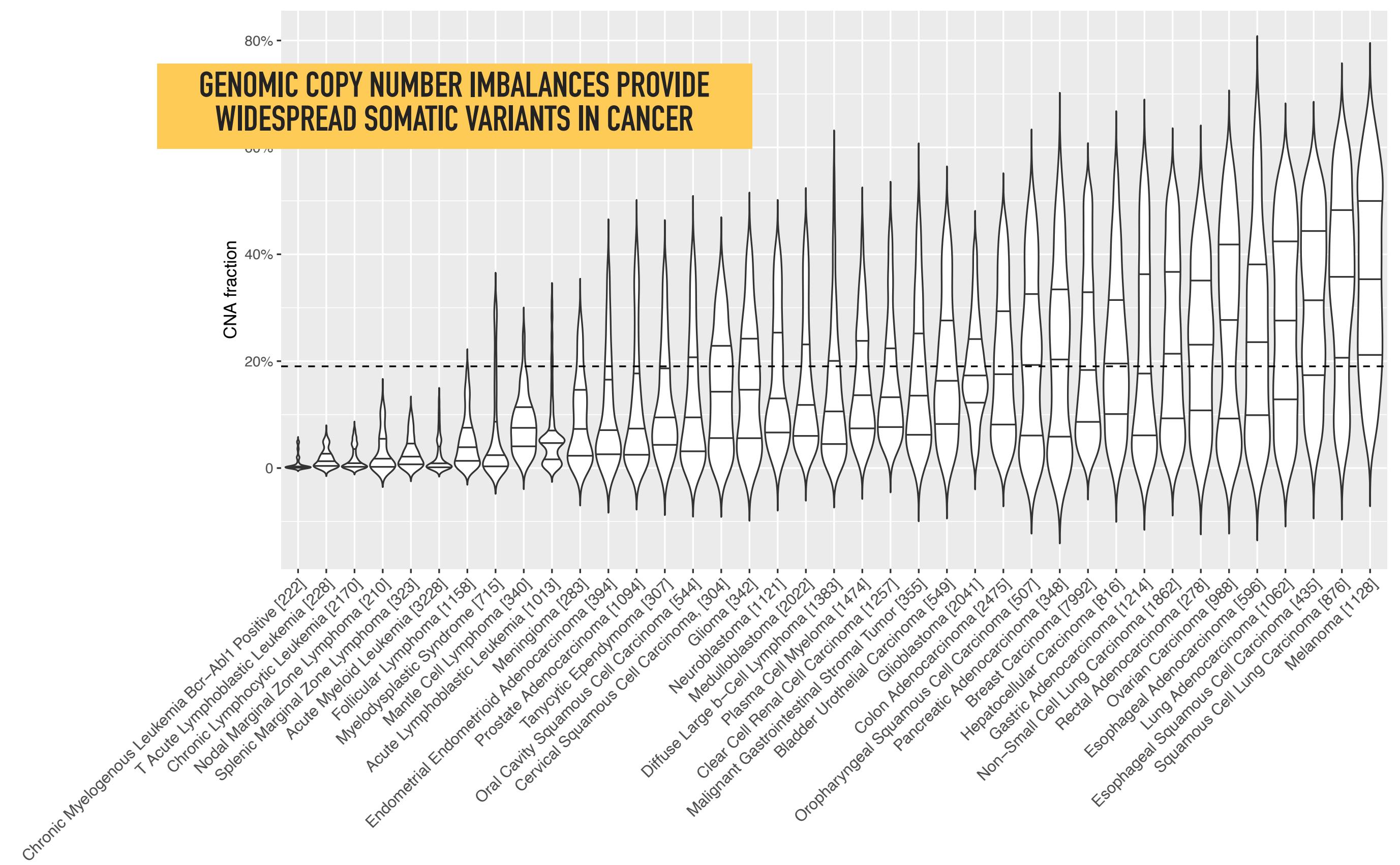
arrayMap



# Quantifying Somatic Mutations In Cancer



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))

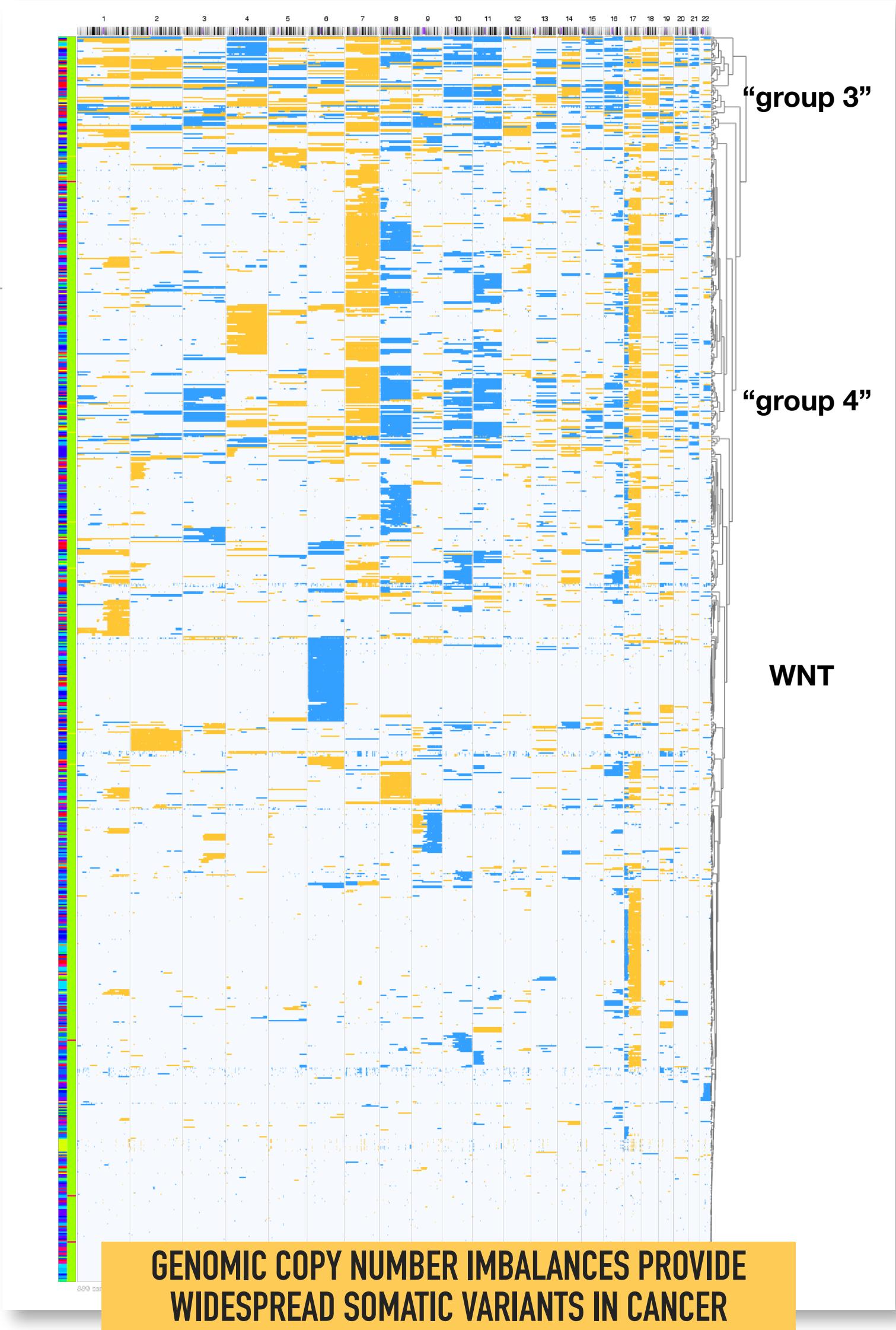
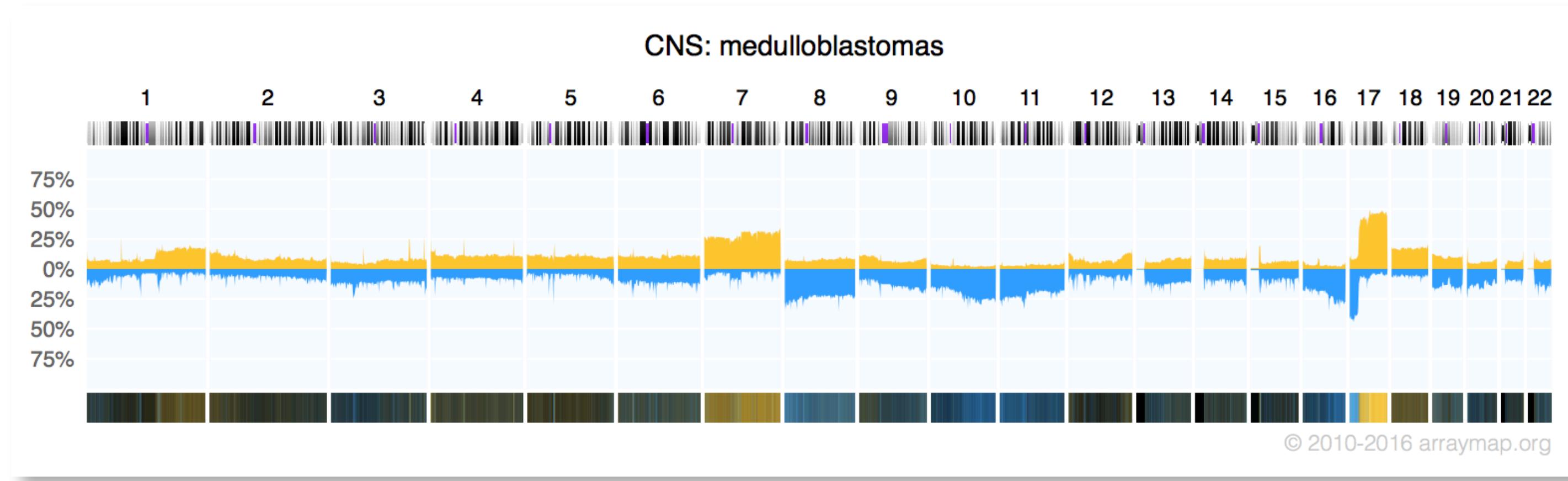


On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from [progenetix.org](http://progenetix.org)

# 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?



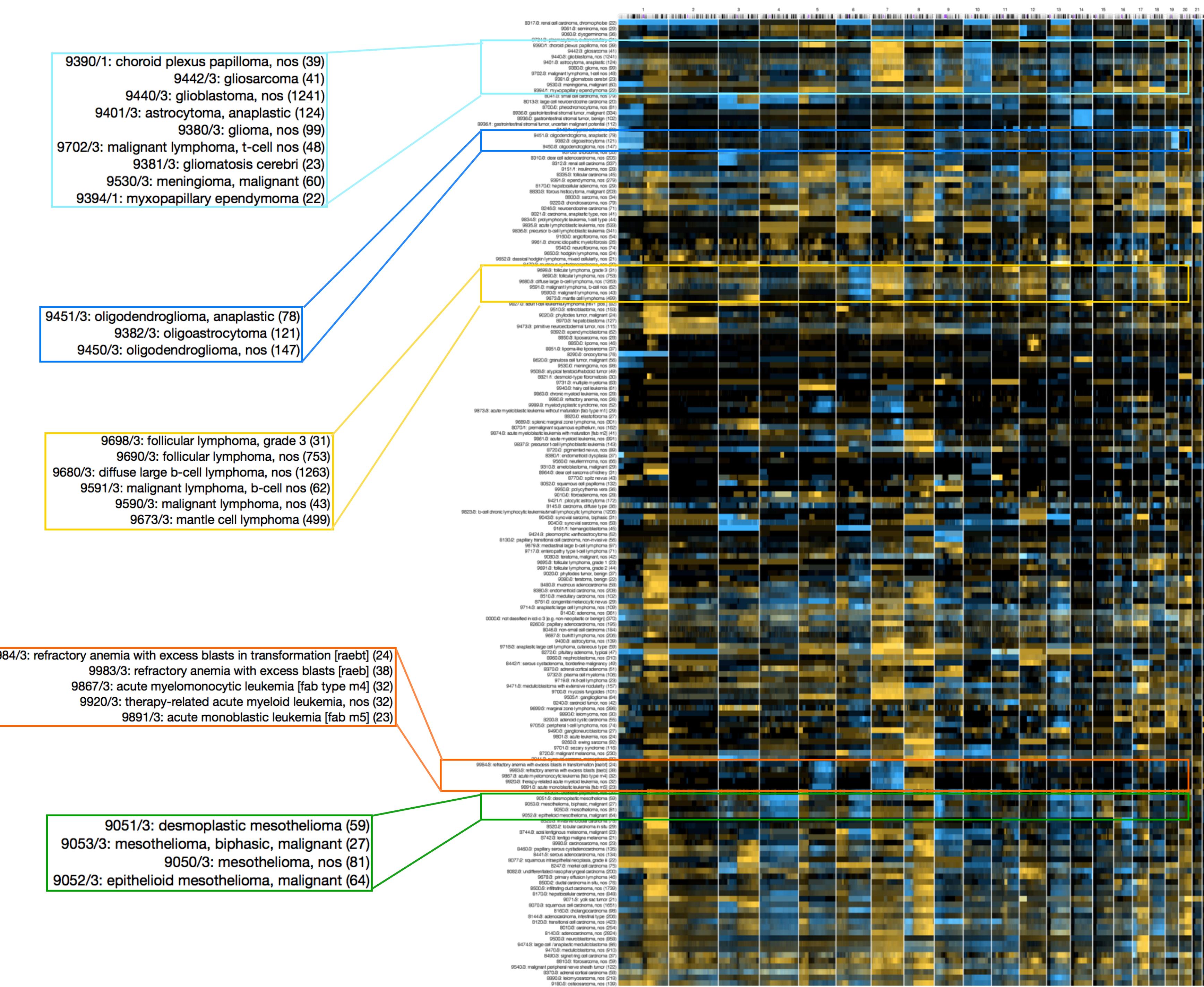
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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



# Somatic Mutations In Cancer: Patterns

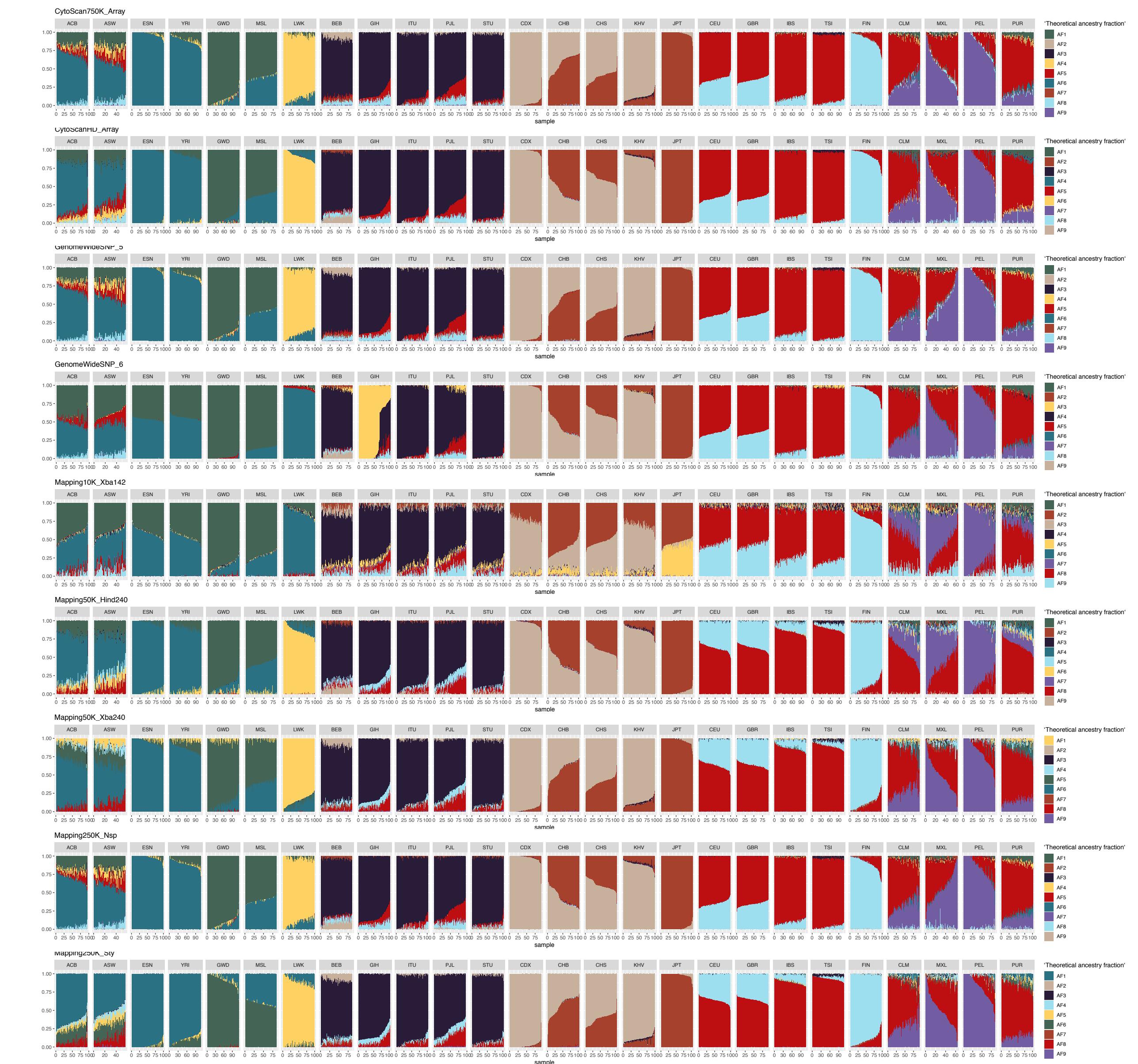
## Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



# 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



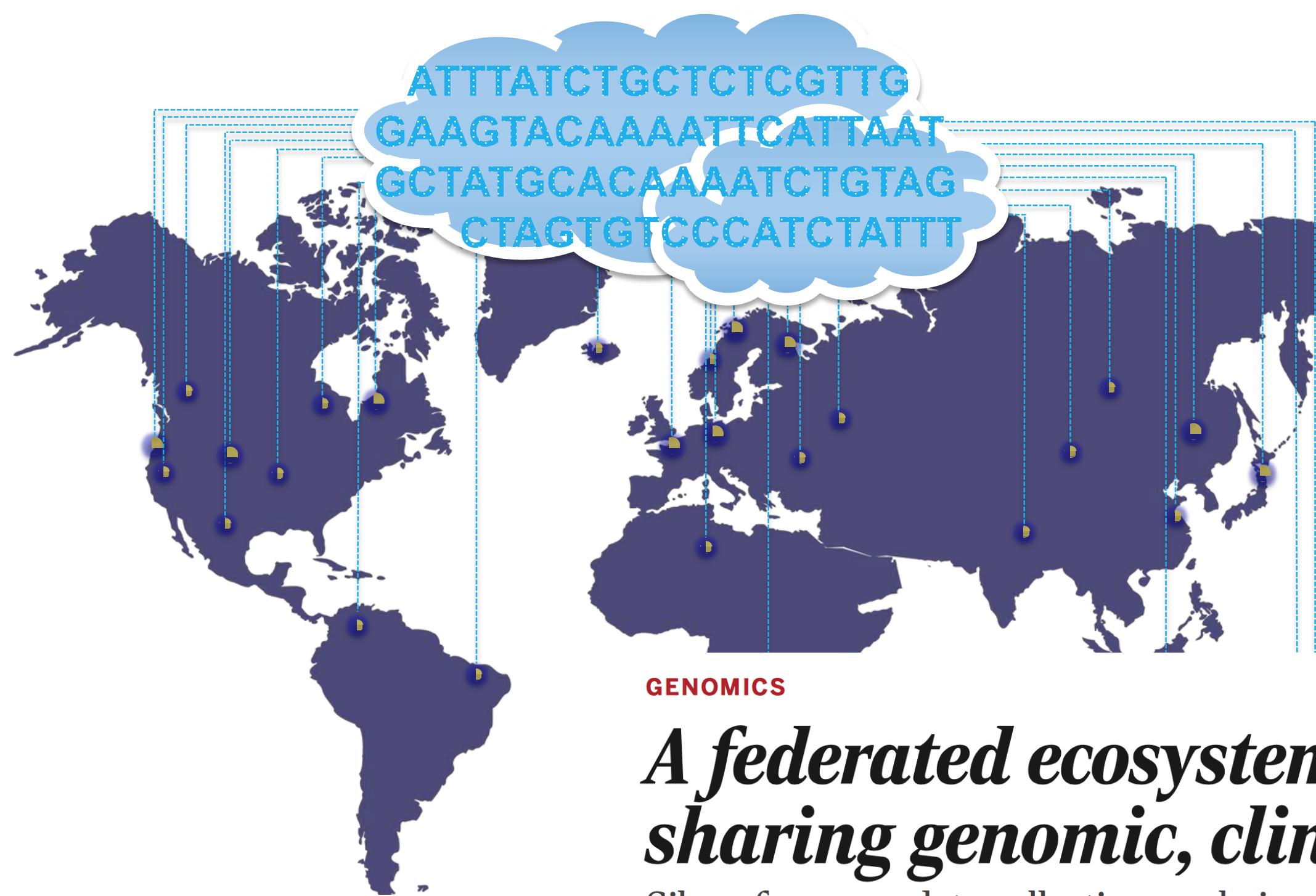
**Figure S1** The fraction or contribution of theoretical ancestors ( $k=9$ ) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are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



University of  
Zurich<sup>UZH</sup>

Department of Molecular Life Sciences

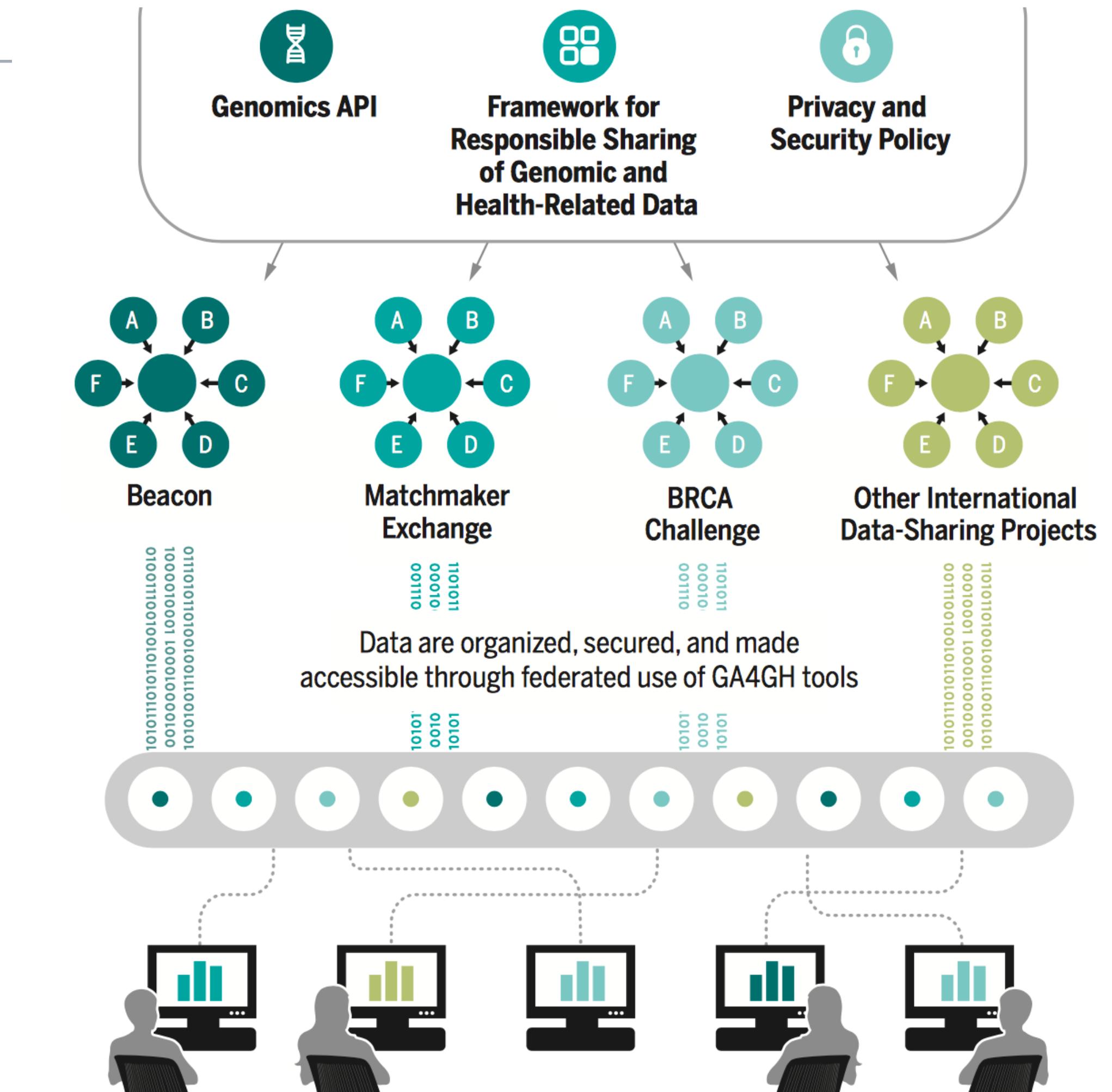
## Genome Data Sharing: The Global Alliance for Genomics and Health (GA4GH)



*A federated ecosystem for sharing genomic, clinical data*

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

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



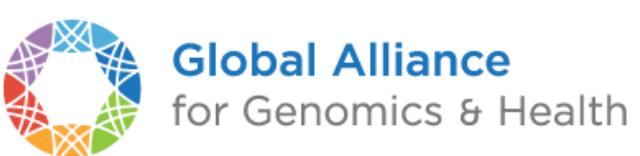


# University of Zurich<sup>UZH</sup>

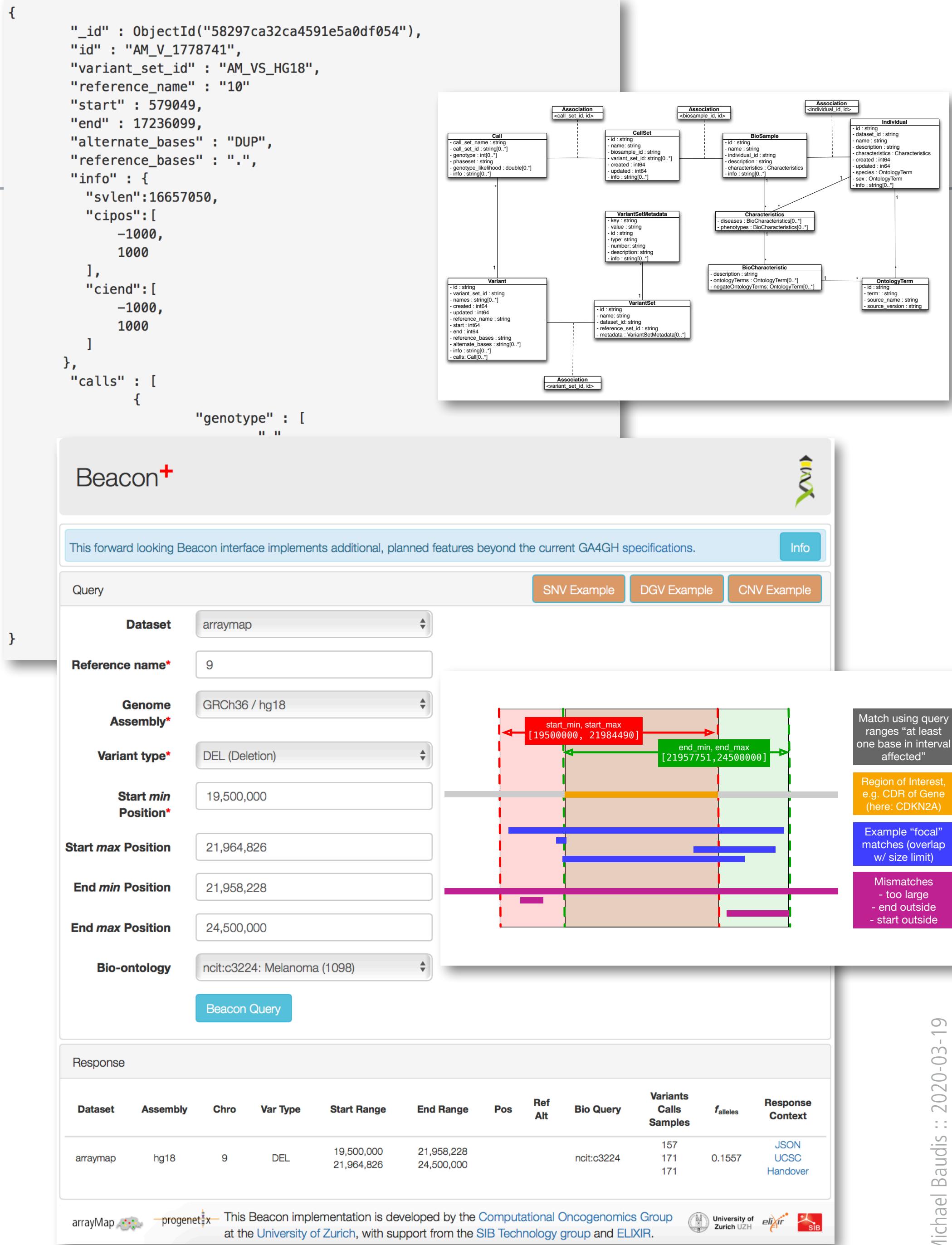
## Department of Molecular Life Sciences

### 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<sup>+</sup>**
  - Beacons - a GA4GH driver project - are **web services** for single-entry genome variant queries over >200 genome resources worldwide.
  - We are developing Beacon<sup>+</sup> 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 ...)



arrayMap



# 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](https://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				

## 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 through the permanent link [schemablocks.org](https://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 ...](#)

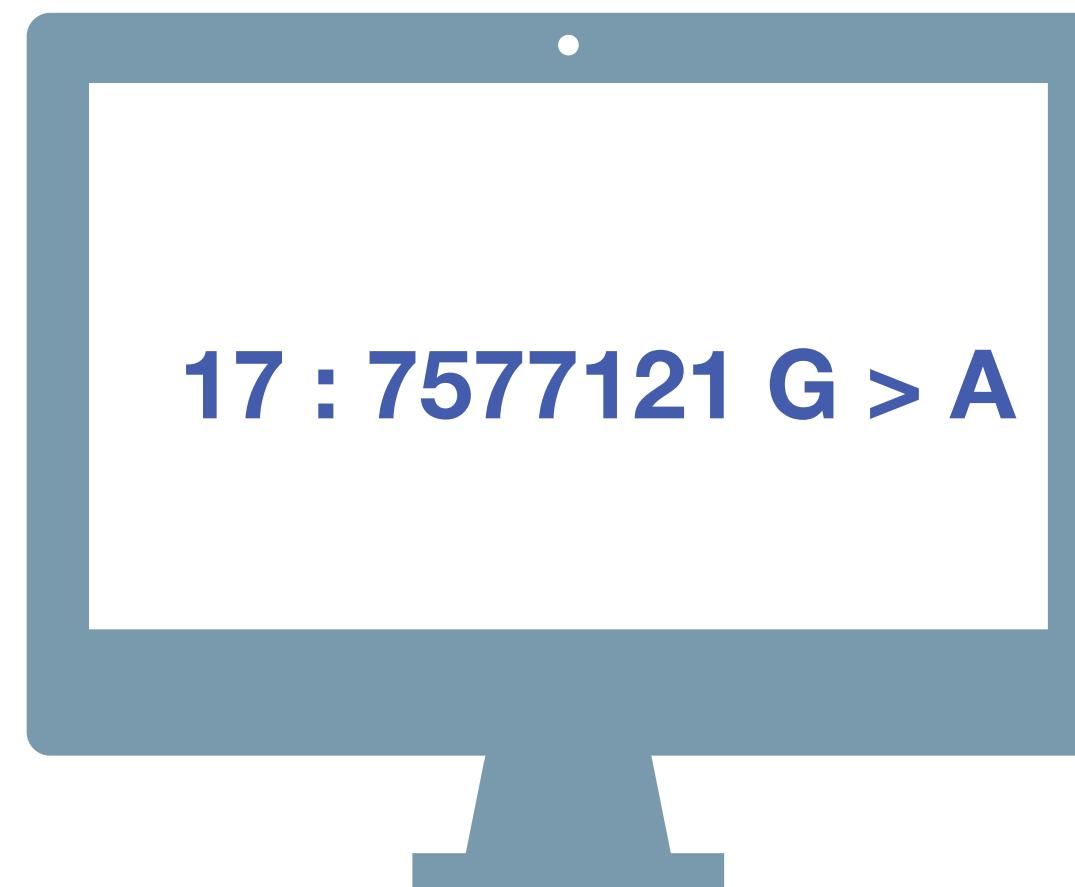
## SchemaBlocks {S}[B] Mission Statement

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 ...](#)

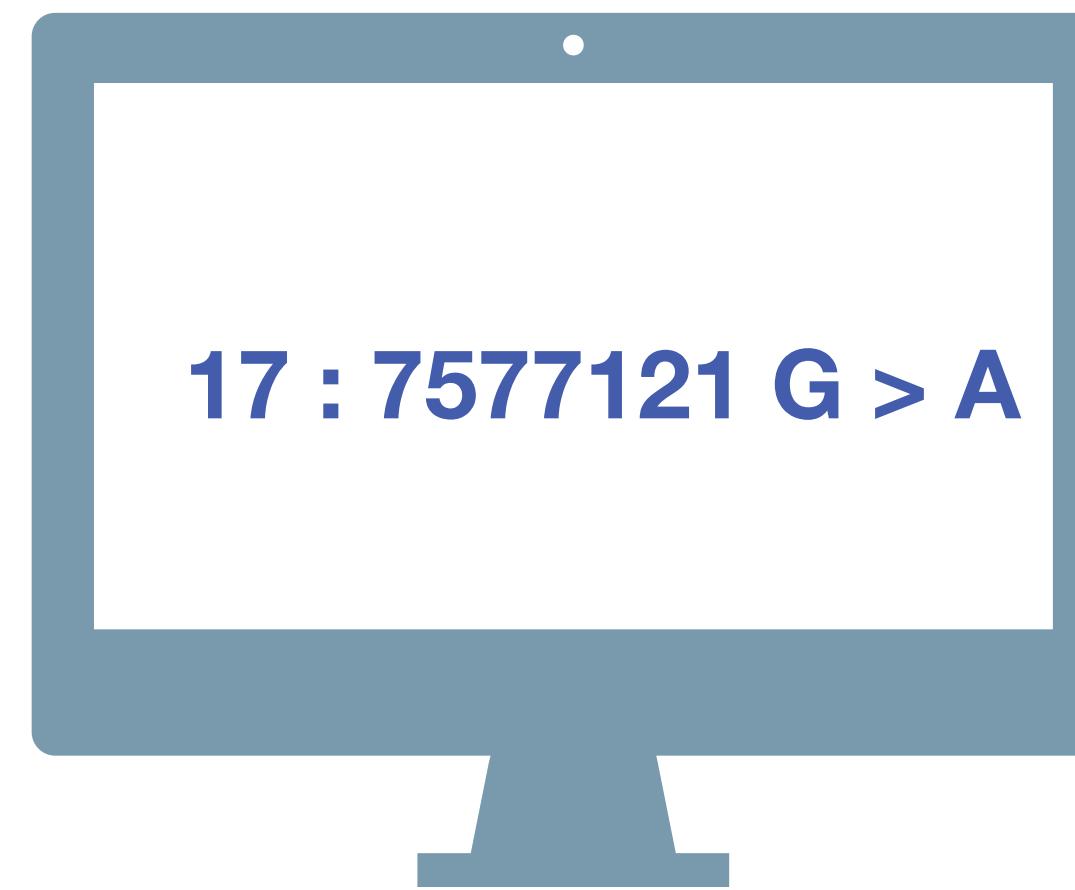




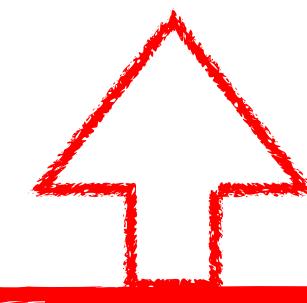
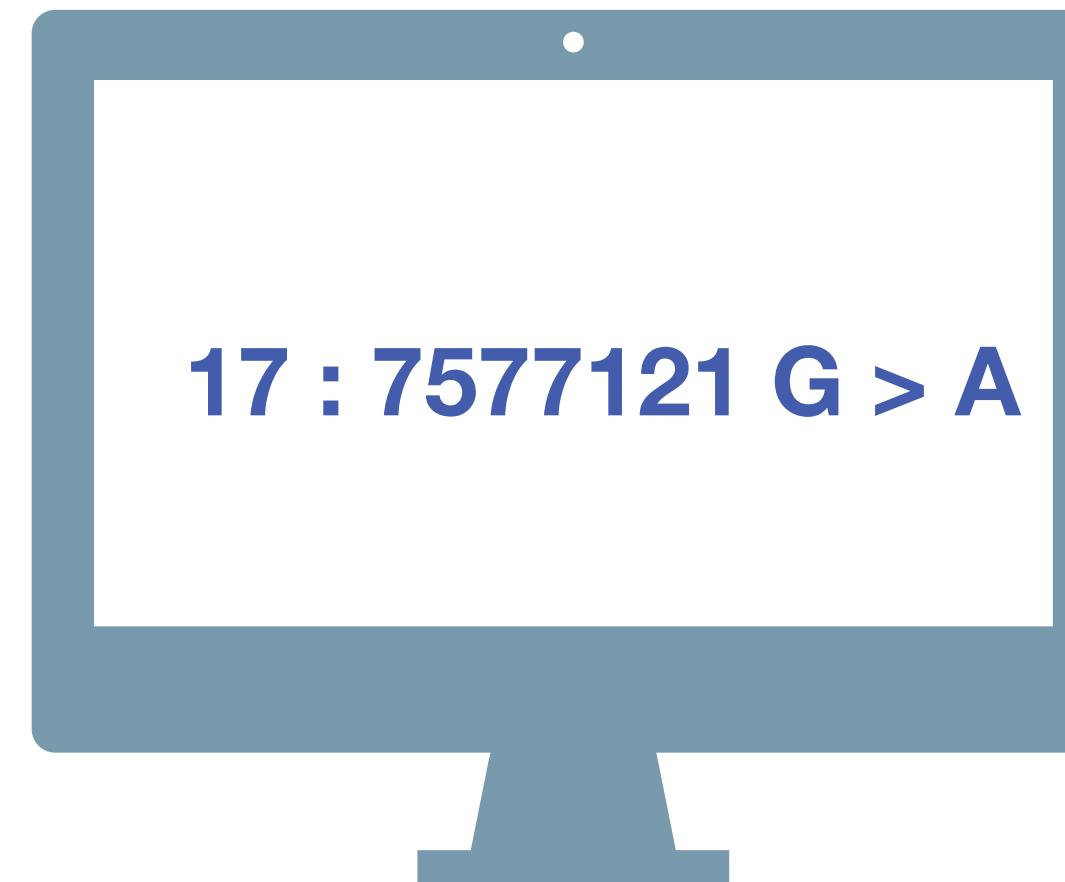
# Beacon

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***

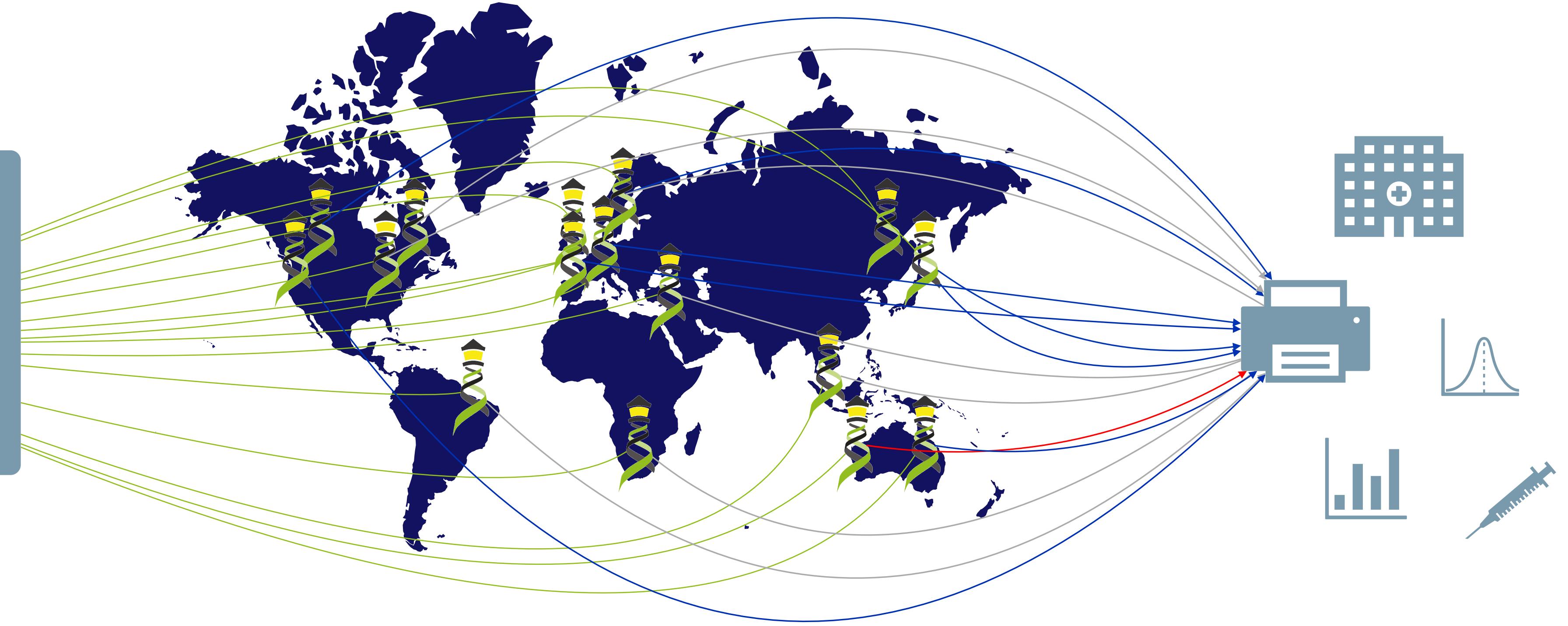
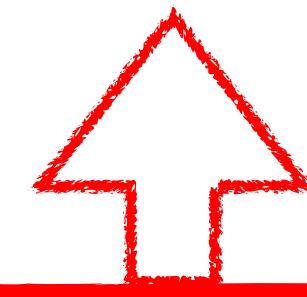
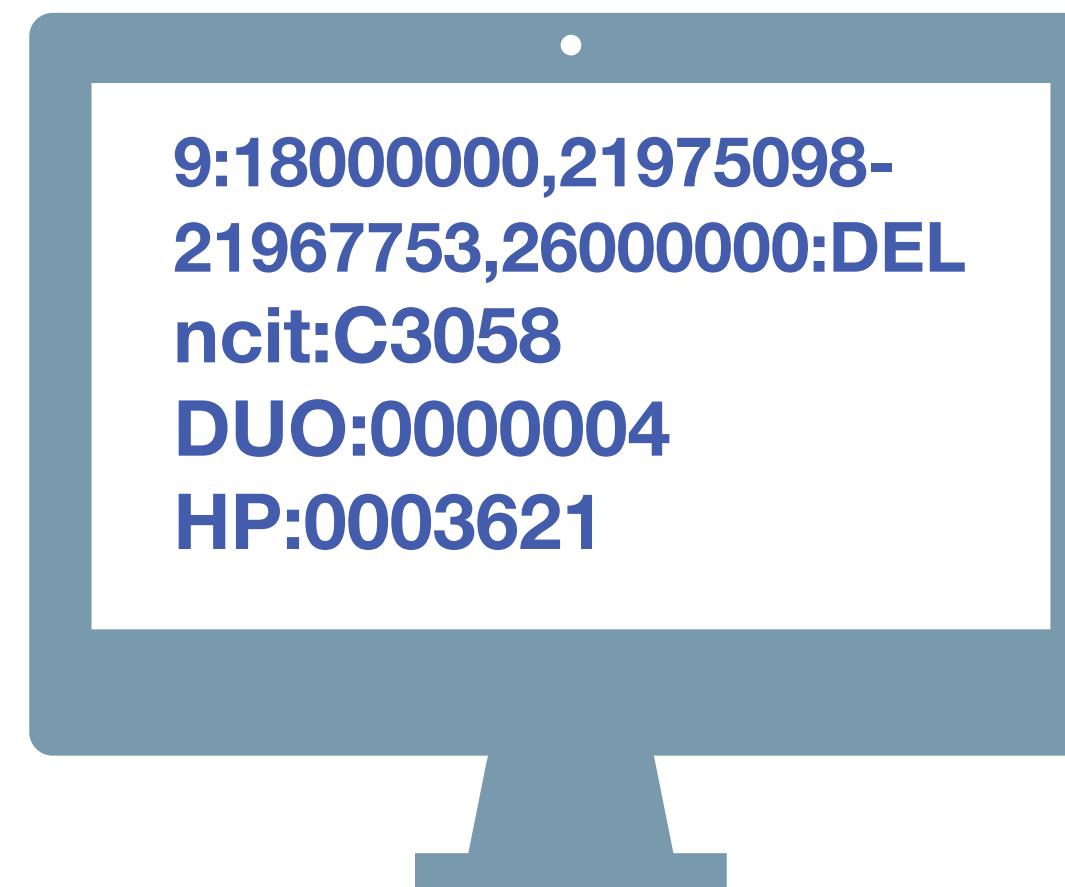


Have you seen this variant?  
It came up in my patient  
and we don't know if this is  
a common SNP or worth  
following up.

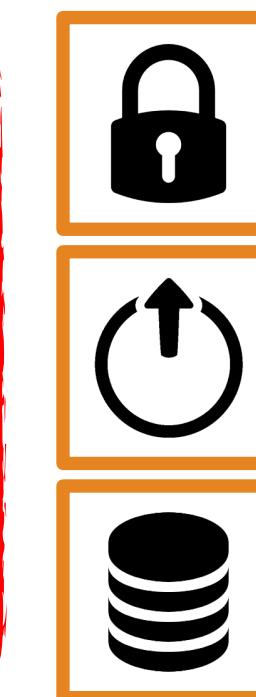


A Beacon network federates  
genome variant queries  
across databases that  
support the **Beacon API**

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

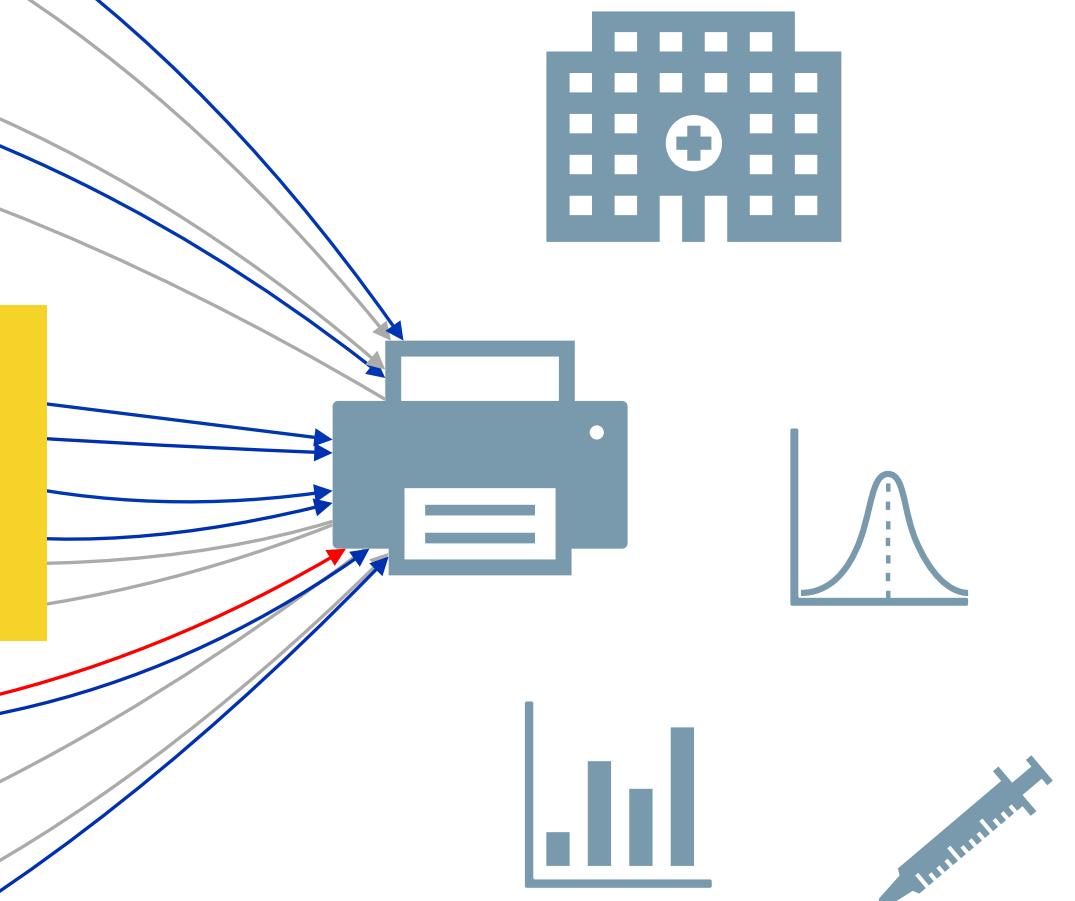
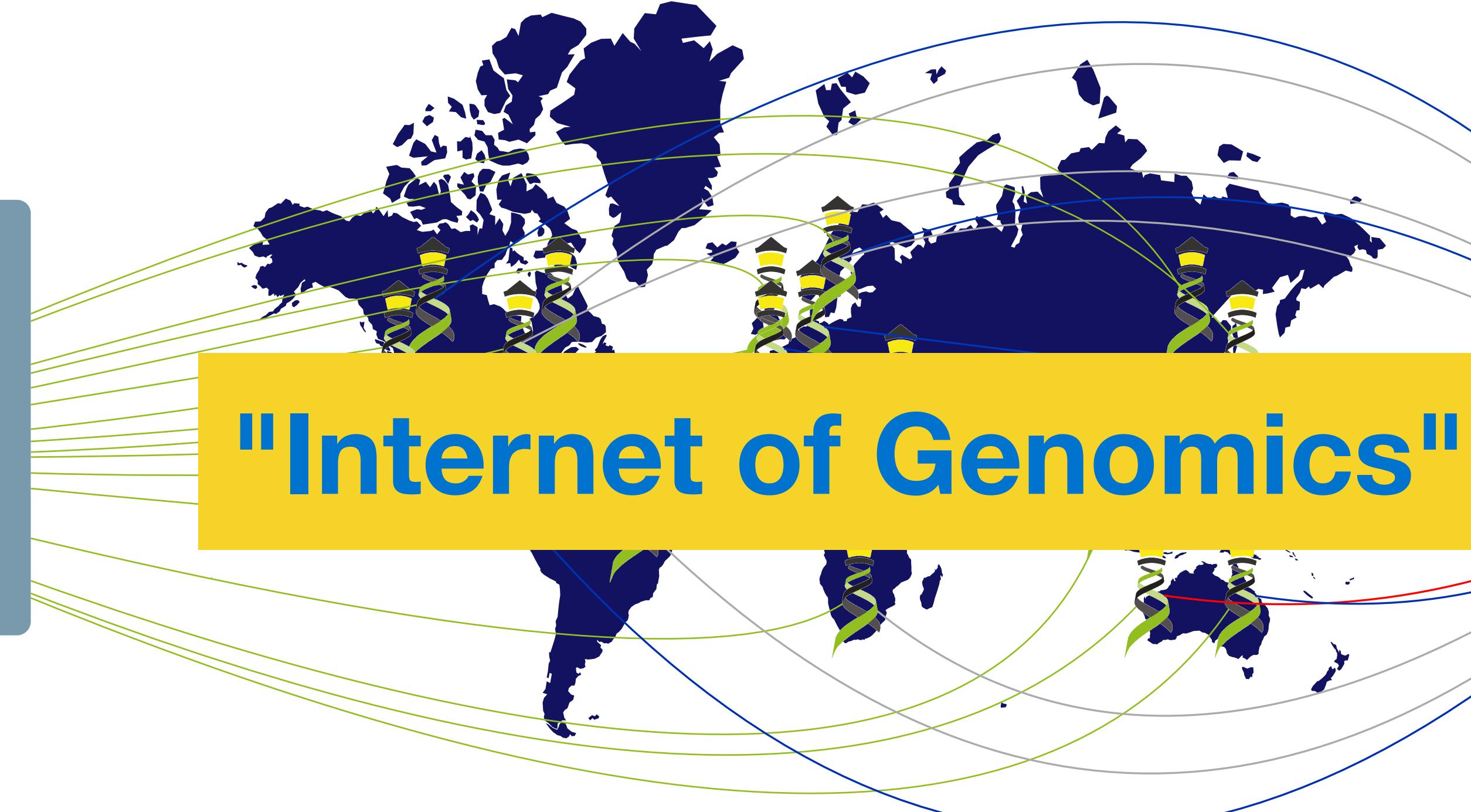
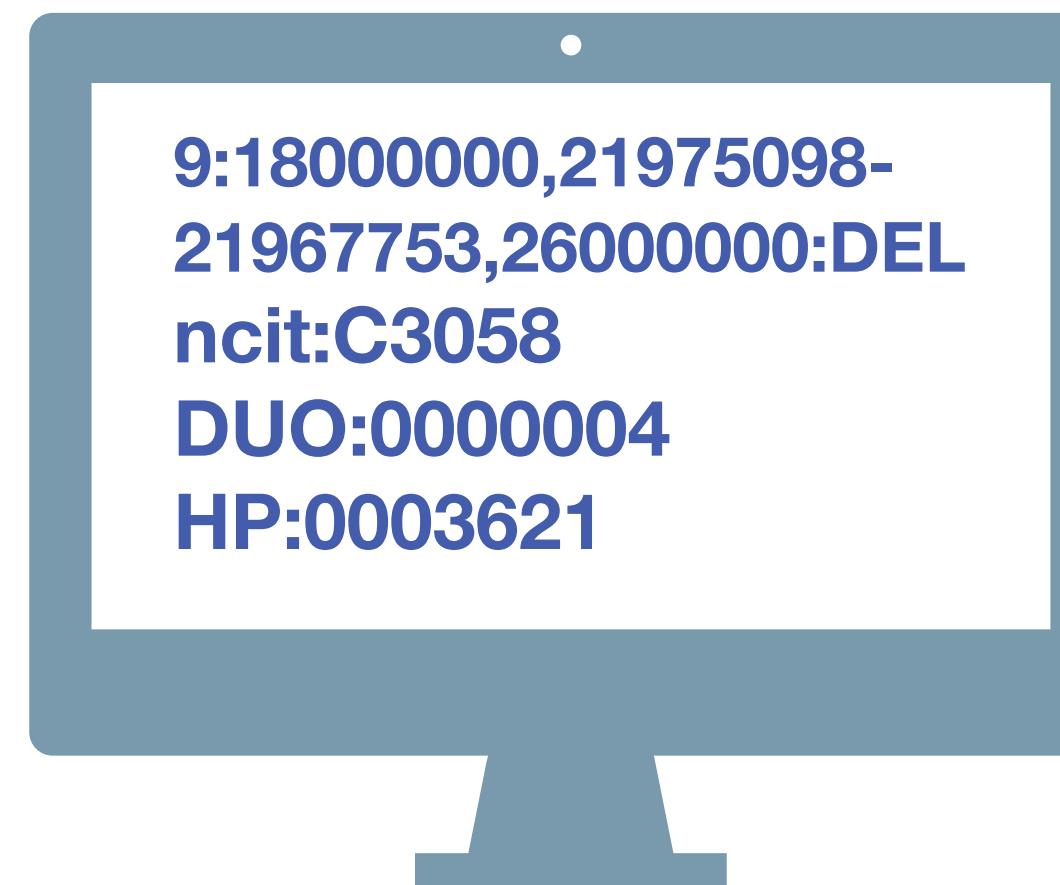


Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

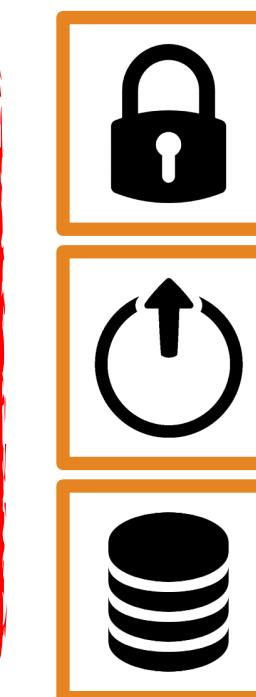


## Beacon v2 API

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 ALTMAYER  
THOMAS EGGERMANN  
ROSA NOGUERA  
REINER SIEBERT  
CAIUS SOLOVAN



University of  
Zurich<sup>UZH</sup>



Global Alliance  
for Genomics & Health

## BBOP GROUP @ LNBL

CHRIS MUNGALL  
JUSTIN REESE  
DEEPAK UNNI

## GA4GH

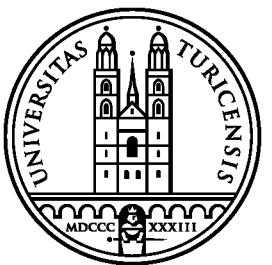
MELANIE COURTOT  
MELISSA HAENDEL  
HELEN PARKINSON

## 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<sup>UZH</sup>

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](http://arraymap.org)  
[progenetix.org](http://progenetix.org)  
[info.baudisgroup.org](mailto:info.baudisgroup.org)  
[sib.swiss/baudis-michael](http://sib.swiss/baudis-michael)  
[imls.uzh.ch/en/research/baudis](http://imls.uzh.ch/en/research/baudis)  
[beacon-project.io](http://beacon-project.io)  
[schemablocks.org](http://schemablocks.org)



Global Alliance  
for Genomics & Health

