

# Implementing Data Models for the Global Alliance for Genomics and Health



Universität Zürich UZH



Global Alliance for Genomics & Health

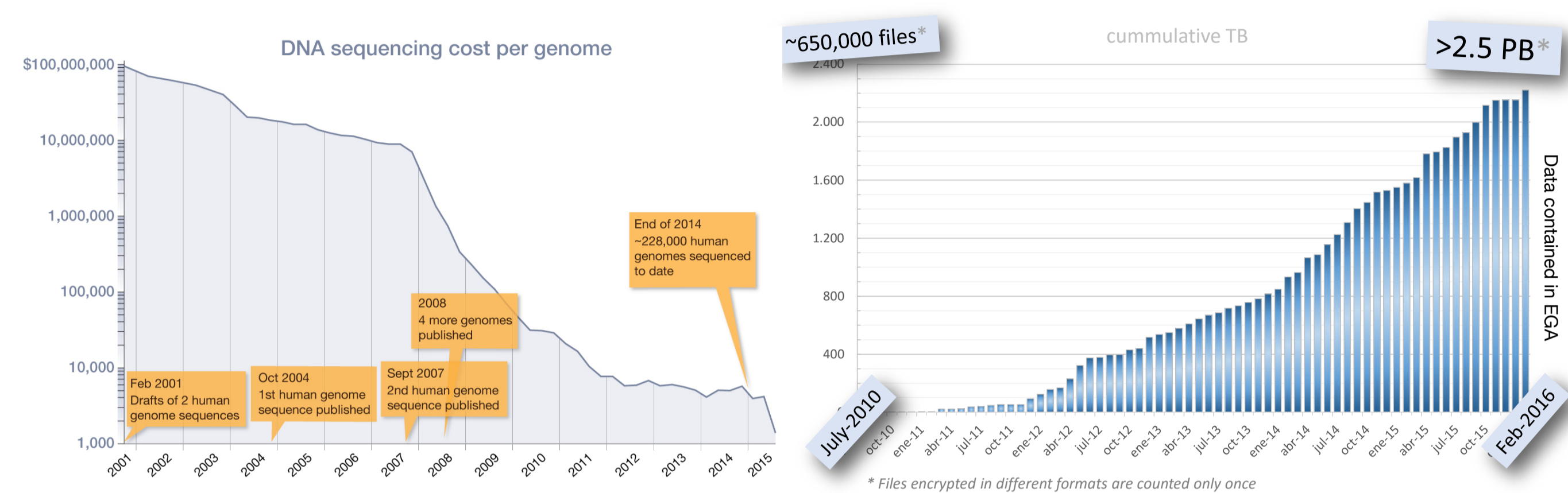
Bo Gao

Baudis Group, Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zürich

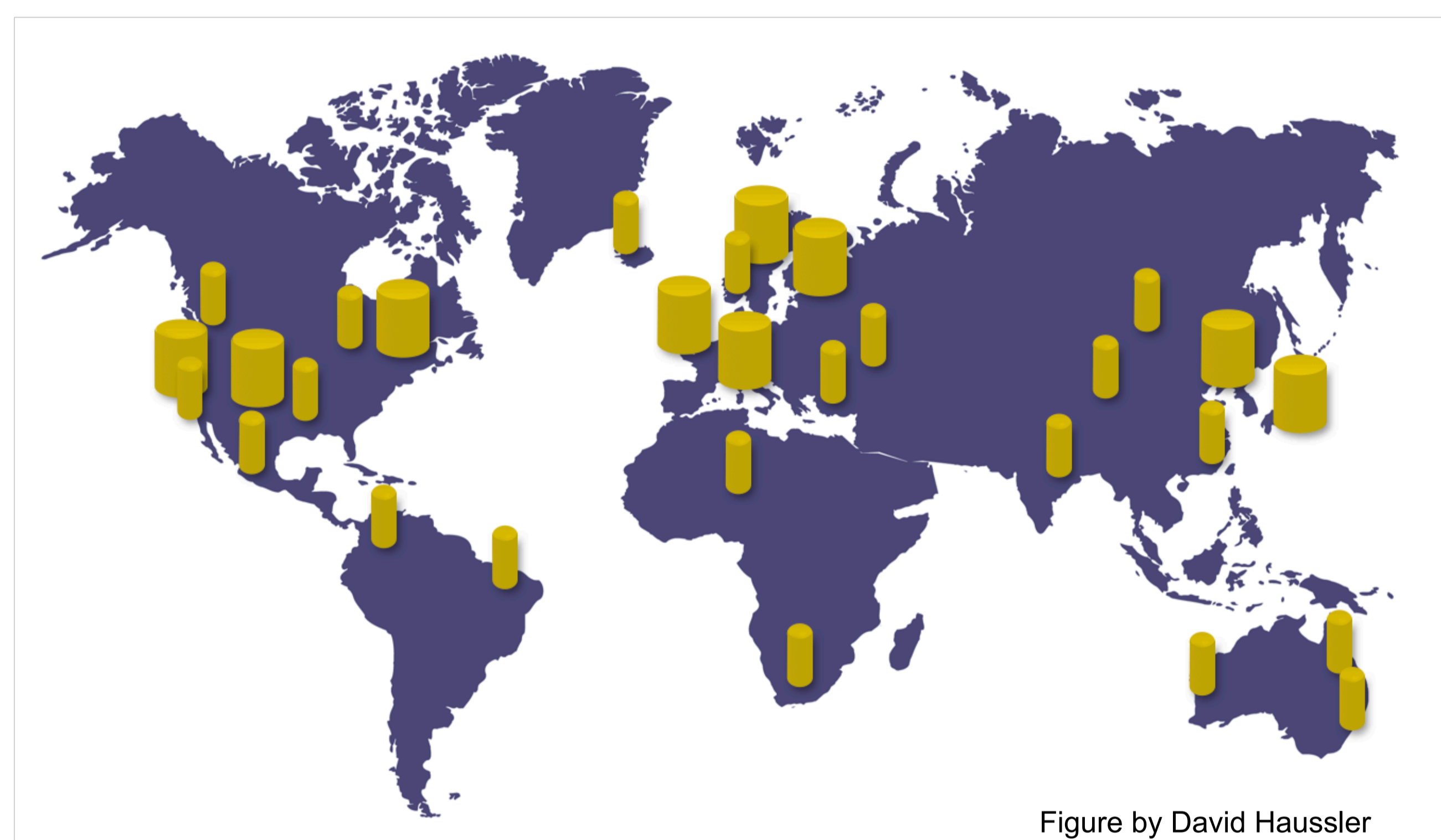
## Overview

- The advent of sequencing age has enriched our understanding of human malignancies.
- As the cost comes down, the amount of data expands exponentially.
- Large scale comparative study of genome variations is crucial for modern biomedical research.
- However, data resources are scattered behind firewalls.

## Cost comes down & Data goes up



## Data held in silos & unshared

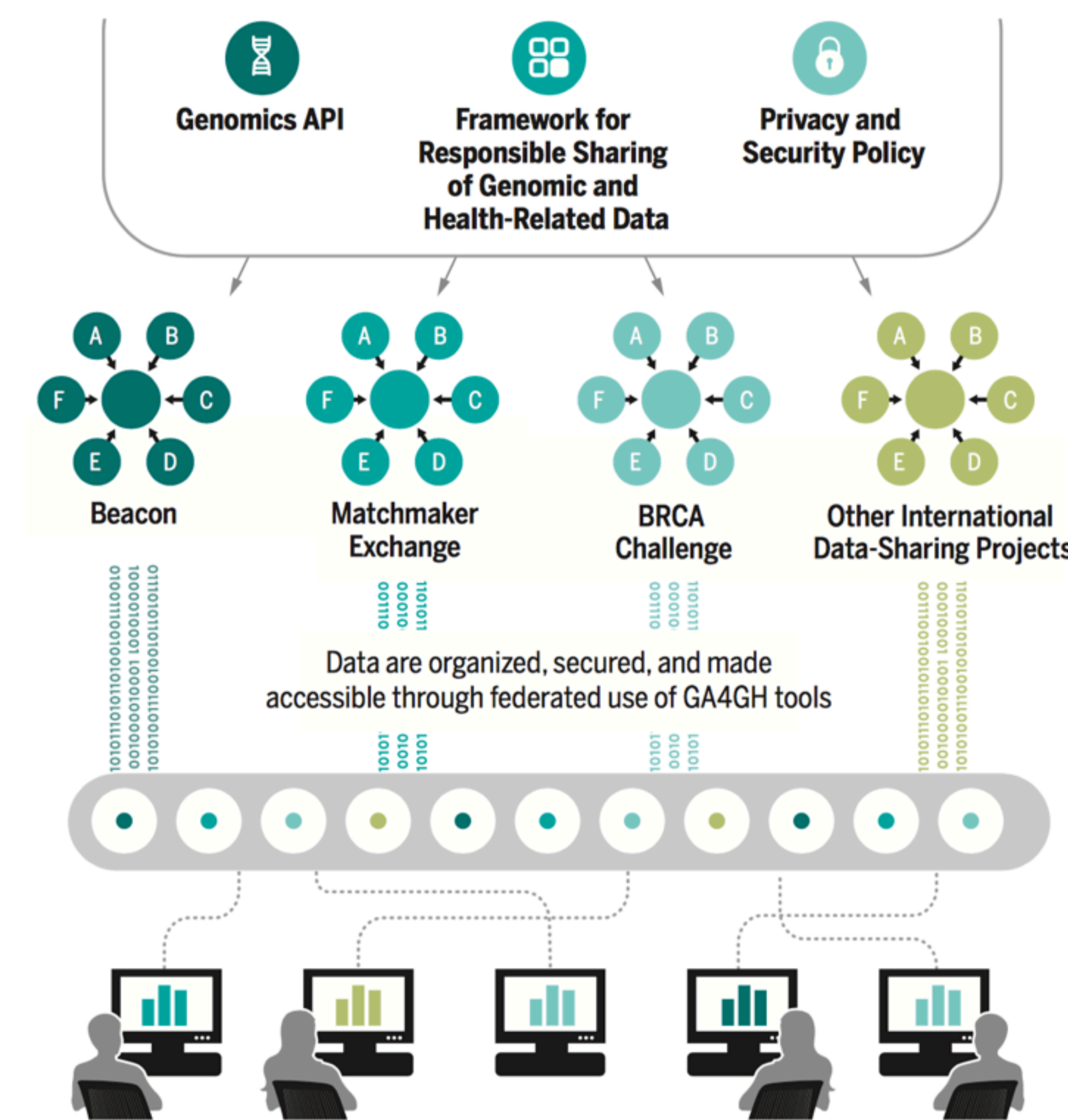


## The Global Alliance for Genomics and Health

GA4GH was founded by leading scientists in biology, medicine, computational research, data security as well as law and ethics. The aim is:

- to develop standards for the representation and exchange of genome data and supporting information,
- to promote the implementation of legal and ethics frameworks and procedures related with the use of this data for research purposes.

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



## Schema

- The arrayMap to GA4GH development pioneers on the definition of data formats and reference software implementations for genomic and associated metadata.
- We are developing modern, standardised data schemas, to facilitate unambiguous annotation and mining of biological or biomedical attributes as well as provenance associated with physical or procedural objects, related to genomic data.
- Since the first prototype version of the GA4GH metadata schema in 2014, considerable progress has been made in the schema's refinement and especially the integration of ontologies for standardised attribute coverage.

## arrayMap

- The arrayMap resource has been established as curated oncogenomic resource, focussing on genomic arrays and copy number aberration (CNA) profiles.
- The underlying data is being extracted from NCBI's Gene Expression Omnibus (GEO), EBI's ArrayExpress, and, importantly, through targeted mining of publication data.
- It is ideal for cancer related genome data and clinical use, such as the diagnostic validations as well as target evaluation for personalized therapeutic approaches.

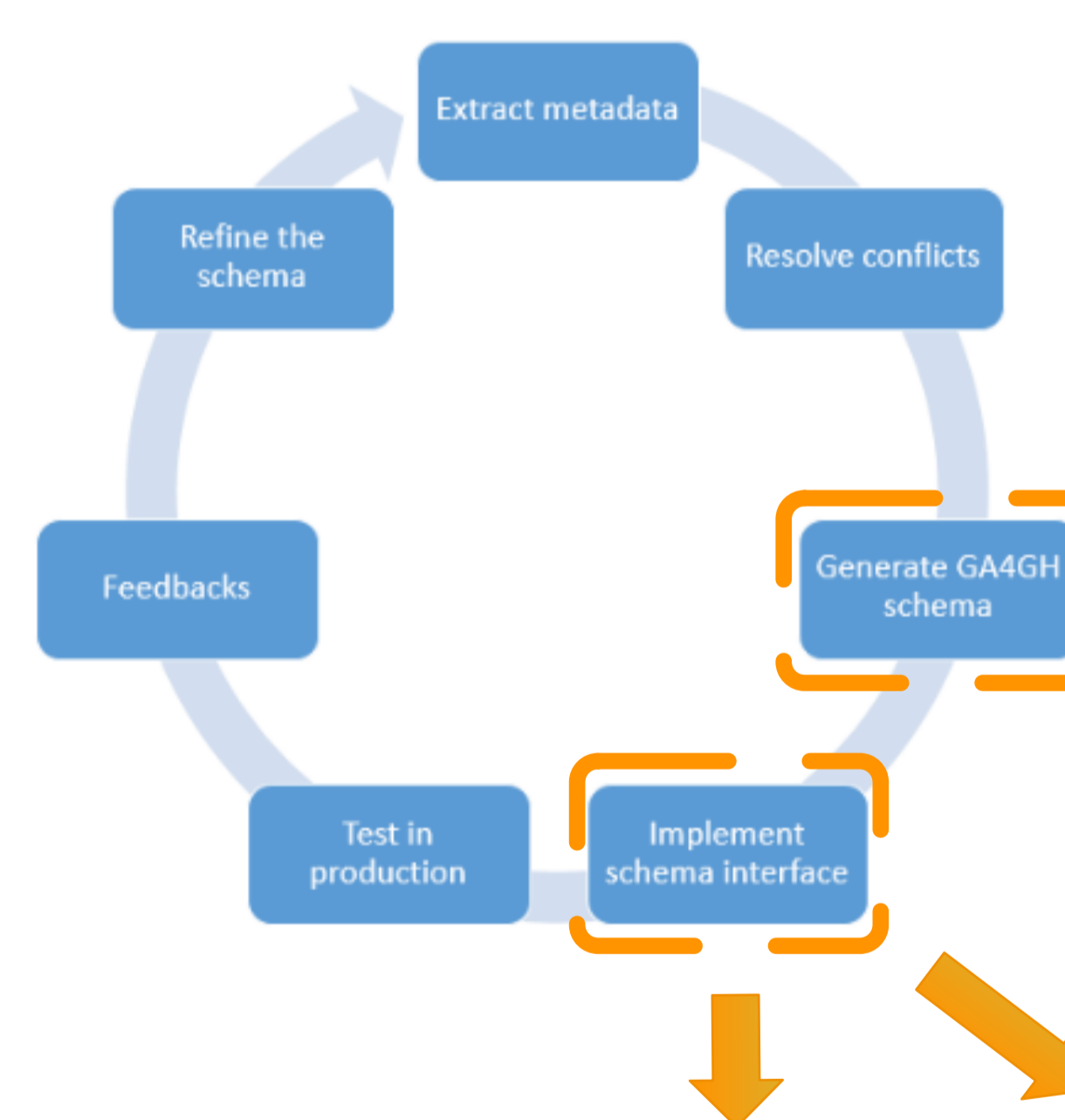
## A Cancer Genome Resource with 60,000+ aCGH arrays

BRAIN TUMOURS	5593 samples	62977 genomic array profiles
BREAST CANCER	8329 samples	914 experimental series
COLORECTAL CANCER	3157 samples	267 array platforms
PROSTATE CANCER	991 samples	ICD-O 245 ICD-O cancer entities
STOMACH CANCER	1062 samples	

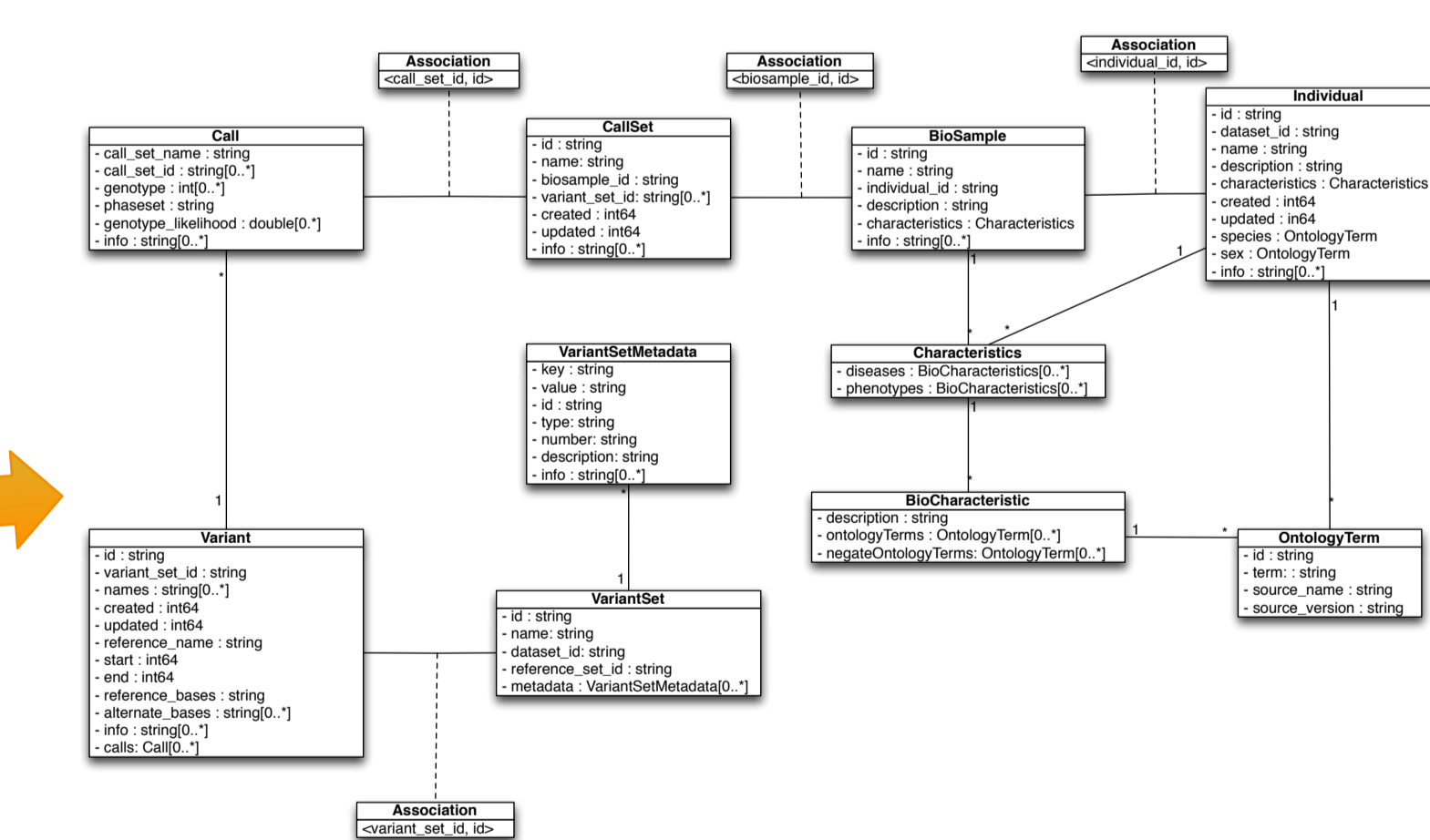
## Visualization of Cancer Genome Profiling



## Iterative Efforts



## arrayMap Data Translation



## arrayMap-ga4gh Implementation

Key	Value	Type
Objectid("584823a...")	{ 11 fields }	Object
id	Objectid("584823a88514...")	Objectid
id	AM_V_3001960	String
end	214426141	Int32
created	2016-12-07 14:33:10.626Z	Date
variant_set_id	AM_VS_HG18	String
reference_bases	null	String
alternate_bases	DEL [ 1 element ]	Array
calls	{ 9 fields }	Object
Objectid("584196f...")	{ 10 fields }	Object
id	Objectid("584196f28514e986936cee73")	Objectid
created	2016-12-02 15:20:33.881Z	Date
characteristics	null	String
updated	2016-12-02 15:20:33.881Z	Date
redirected_to	null	String
species	{ 4 fields }	Object
id	http://purl.obolibrary.org/obo/NCBITaxon...	String
source_version	null	String
source_name	NCBITaxon	String
term	Homo sapiens	String
id	PGIND_GSM255272	String
description	lymph node negative breast cancer	String
name	null	String
sex	{ 4 fields }	Object
id	null	String
source_version	http://purl.obolibrary.org/obo/pato/releas...	String
source_name	PATO	String
term	genotypic sex	String

## arrayMap-beacon Implementation

### Beacon arrayMap

Beacon v0.4 implementation for arrayMap.

Reference name:

Start:

Length:

Assembly ID:

Dataset ID:

Alternate bases:

Confidence Interval (Start position):

Confidence Interval (End position):

Match type: