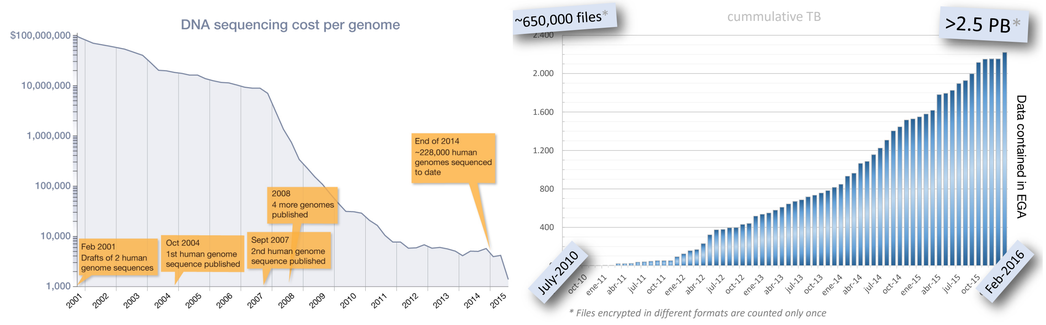
Implementing Data Models for the Global Alliance for Genomics and Health

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 expends exponentially.
- Large scale comparative study of genome variations is crucial for modern biomedical research.
- However, data resources are scatted behind firewalls.

Cost comes down & Data goes up



Data held in silos & unshared

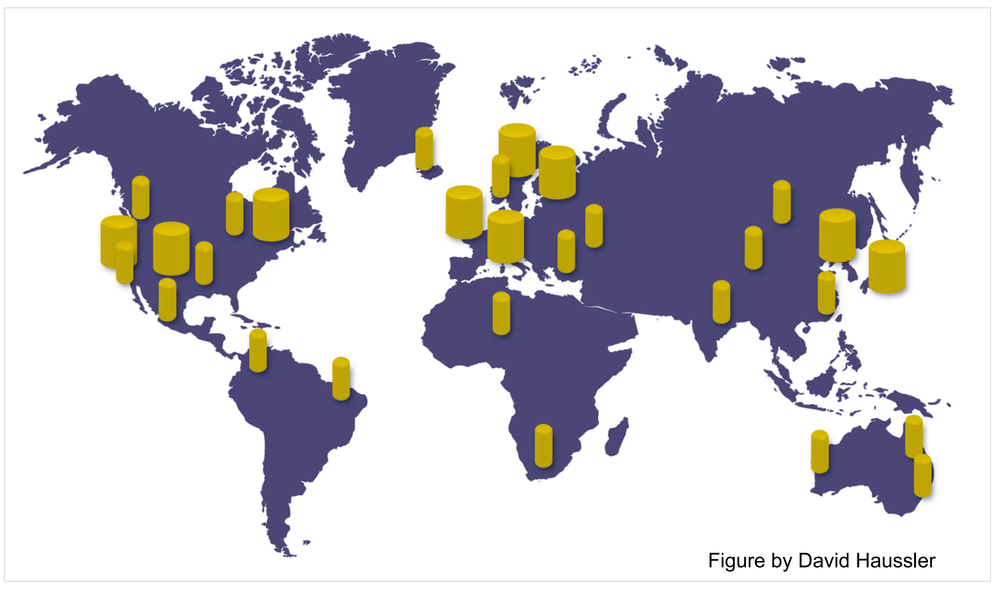


Figure by David Haussler

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.

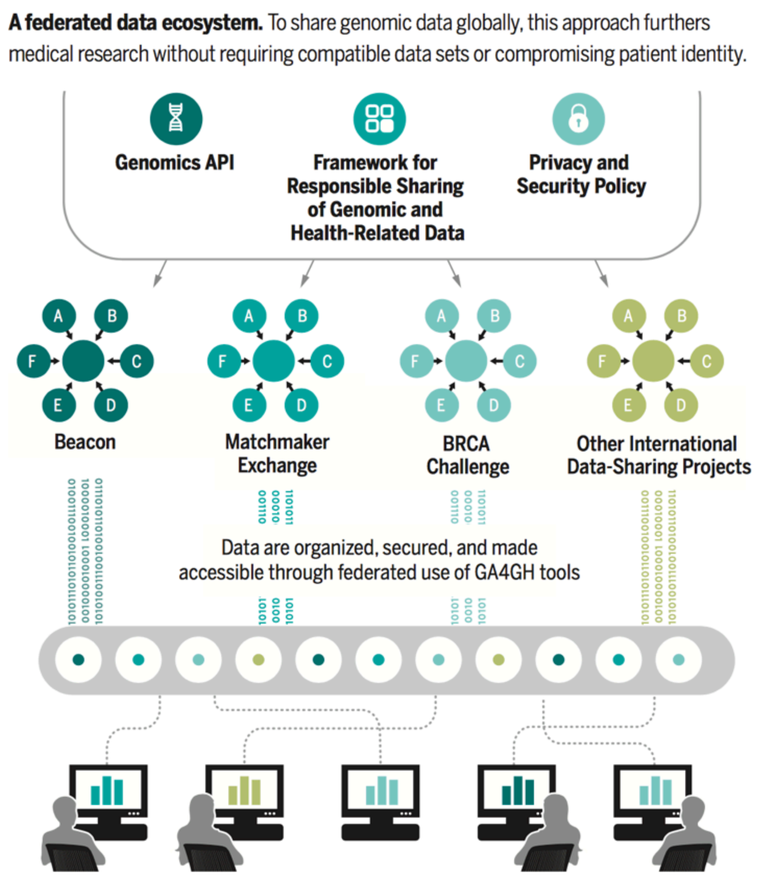
A Cancer Genome Resrouce with 60,000+ aCGH arrays

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 ↗		

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:

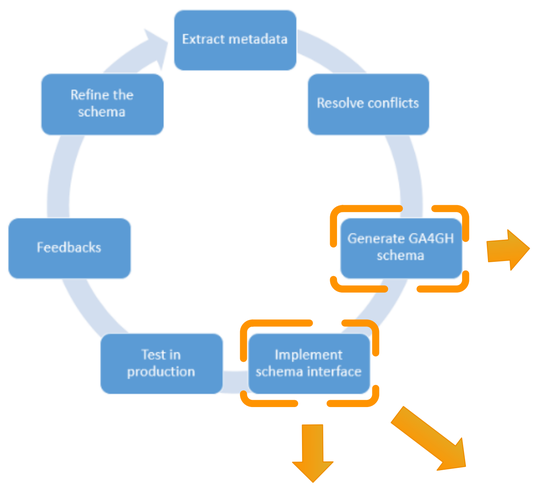
- to develop standards for the representation and exchange of genome data and supporting information,
- to promote the implementation of legal and ethics frameworks and procedures related with the use of this data for research purposes.



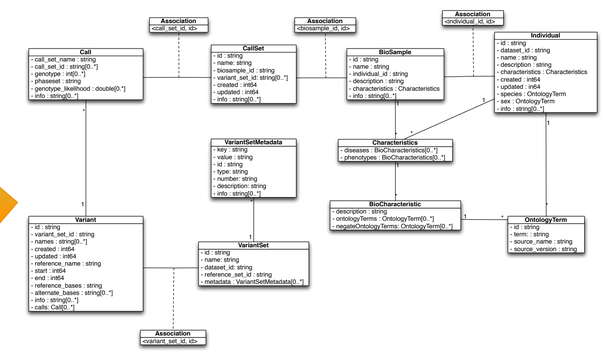
Schema

- 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.

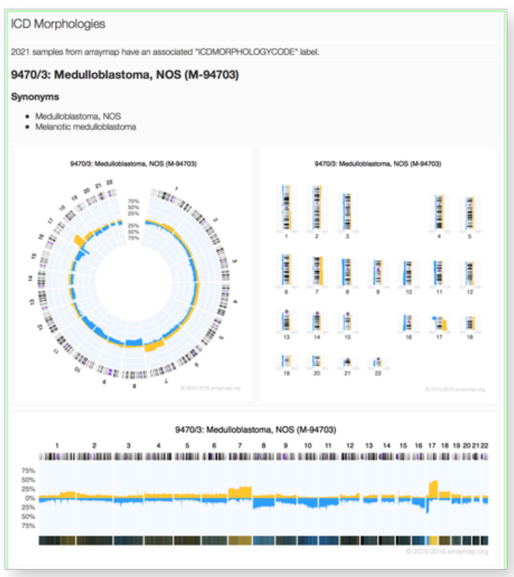
Iterative Efforts

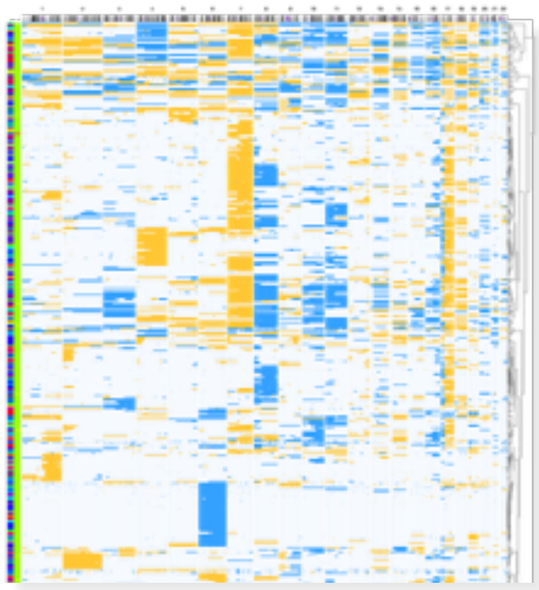


arrayMap Data Translation

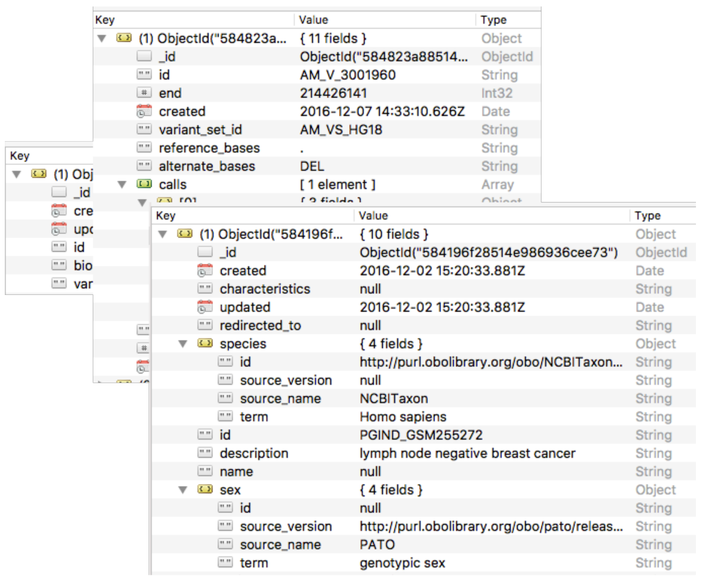


Visualization of Cancer Genome Profiling





arrayMap-ga4gh Implementation



arrayMap-beacon Implementation