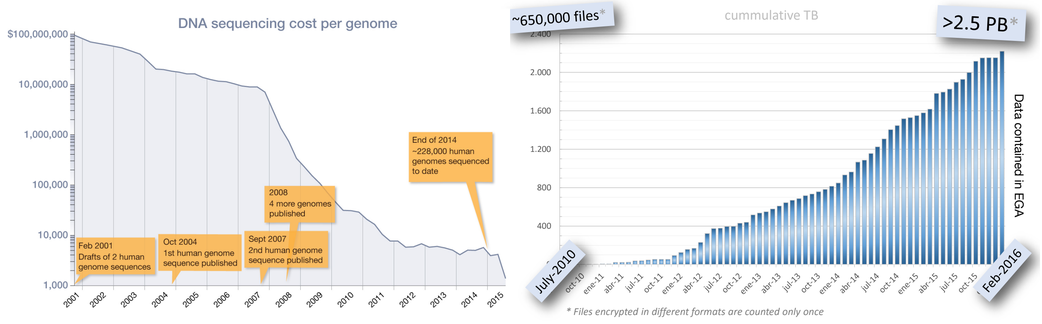
Implementing Data Models for the Global Alliance for Genomics and Health

ELIXIR All Hands 2017, 21-23 March, Rome, Italy

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 biomedical research.
- However, data resources are scatted behind firewalls.

Cost comes down & Data goes up



Data held in silos & unshared

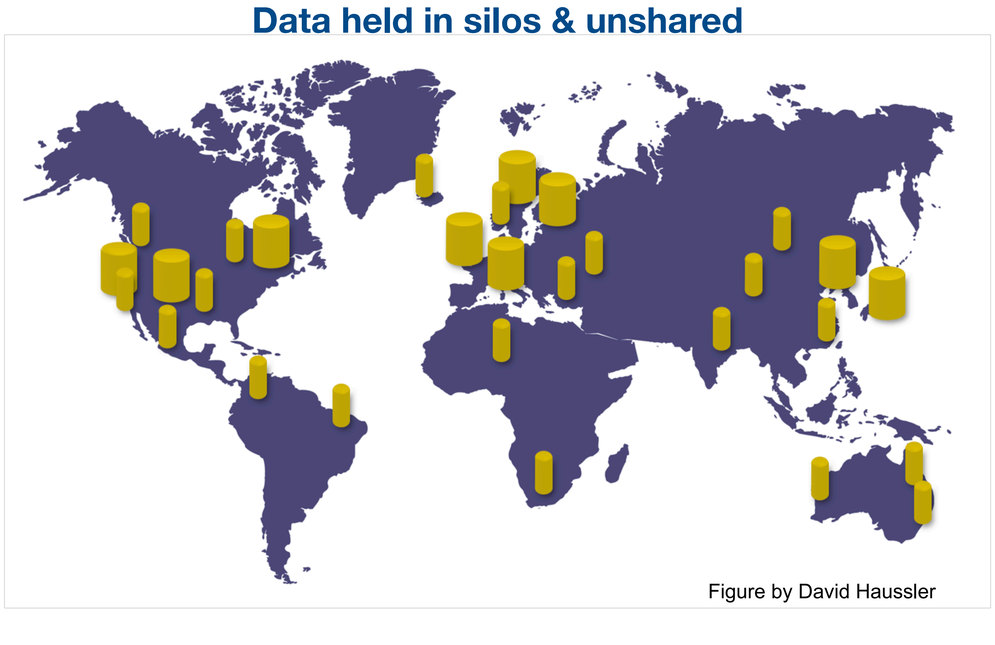


Figure by David Haussler

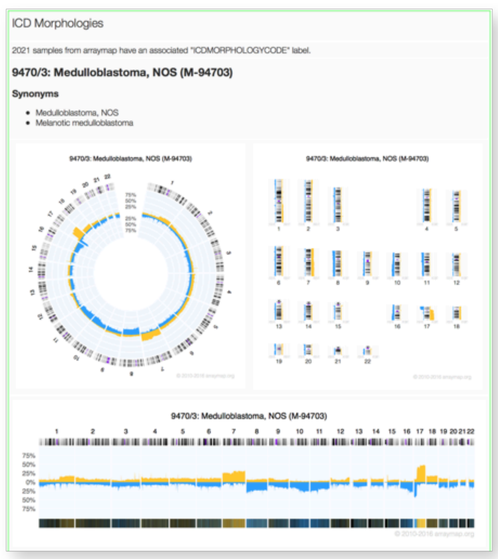
The arraymap.org Cancer Genome Resource

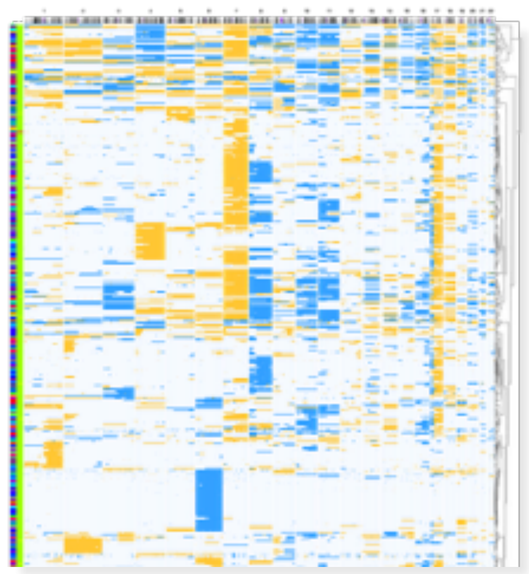
- The arrayMap resource has been established as curated oncogenomic resource, focussing on genomic arrays and copy number aberration (CNA) profiles.
- The underlying data is being extracted from NCBI's Gene Expression Omnibus (GEO), EBI's ArrayExpress, and, importantly, through targeted mining of publication data.
- It is 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 ↗	245 ICD-O cancer entities
STOMACH CANCER	1062 samples ↗	

Visualization of Cancer Genome Profiling

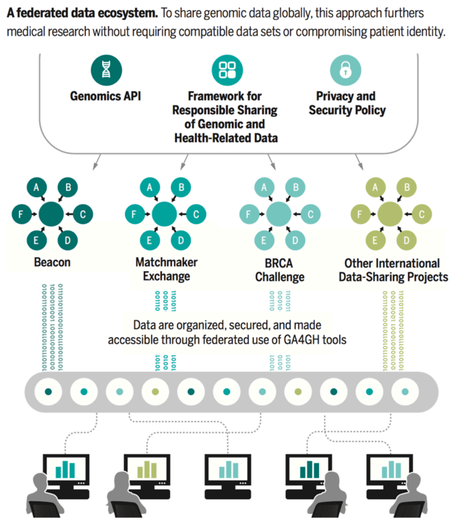




Global Alliance for Genomic and Health

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.



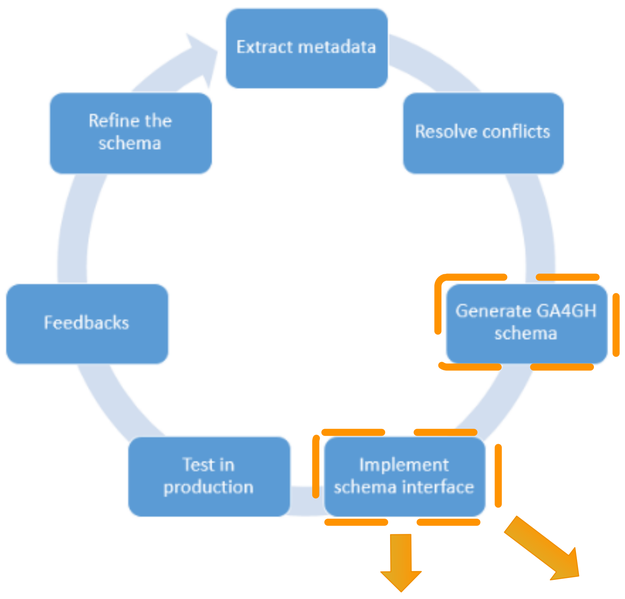
GA4GH task teams

REFERENCE GENOMES TASK TEAM	G2P TASK TEAM
RNASEQ TASK TEAM	VARIANT ANNOTATION TASK TEAM
META DATA TASK TEAM	FILEFORMATS TASK TEAM
CONTAINERS AND WORKFLOWS TASK TEAM	BENCHMARKING TASK TEAM
BEACON TASK TEAM	MATCHMAKER TASK TEAM
READS TASK TEAM	CANCER GENE TRUST
DIRECTORY AND STREAMING API	VICC

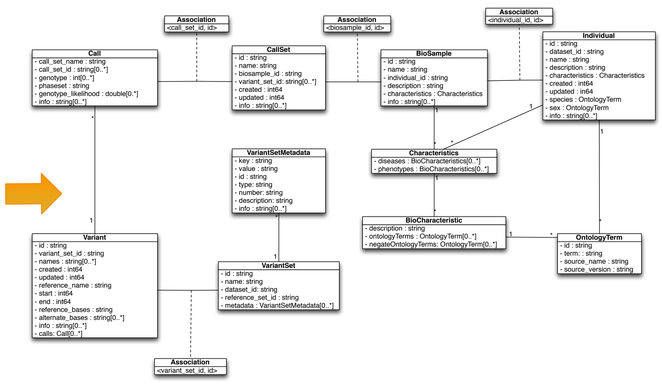
Schema Development

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

Iterative Efforts



arrayMap Data Translation

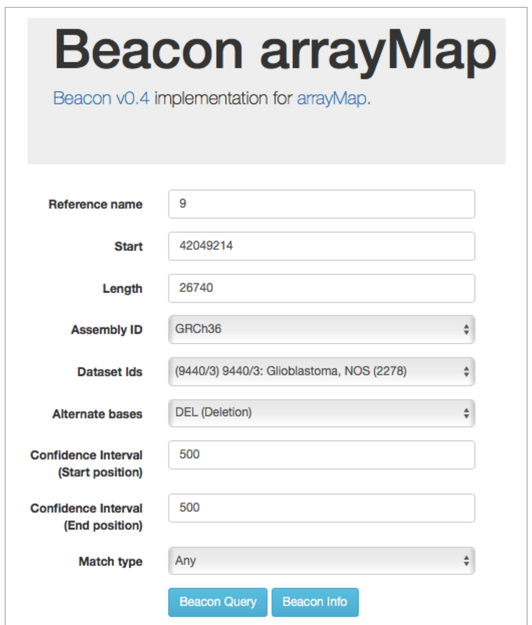


arrayMap-beacon

arrayMap-ga4gh Implementation

@github.com/progenetix/arraymap2ga4gh

Implementation




```
{
        "_id" : ObjectId("589dfa5109d374e4f3655aee"),
        "name" : "AM_BS_GSM322223",
        "individual_id" : "PGIND_GSM322223",
        "id" : "AM_BS_GSM322223",
        "characteristics" : {
                "diseases" : [
                        {
                                "ontologyTerms" : [
                                        {
                                                "termLabel" : "Chronic Lymphocytic Leukemia",
                                                "termId" : "NCIT:C3163"
                                        },
                                        {
                                                "termLabel" : "B-cell chronic lymphocytic leukemia/small lymphocytic lymphoma",
                                                "termId" : "SNMI:M-98233"
                                        },
                                        {
                                                "termLabel" : "B-cell chronic lymphocytic leukemia/small lymphocytic lymphoma",
                                                "termId" : "ICDOM:9823_3"
                                        },
                                        {
                                                "termLabel" : "hematopoietic and reticuloendothelial systems",
                                                "termId" : "ICDOT:C42"
                                        }
                                ],
                                "negatedOntologyTerms" : [ ],
                                "description" : "Chronic Lymphocytic Leukemia"
                        }
                ],
                "phenotypes" : [ ]
        },
        "description" : "Chronic Lymphocytic Leukemia",
        "info" : {
                "tnm" : "T1",
                "death" : "0",
                "country" : "Sweden",
                "geo_long" : 17.64,
                "redirected_to" : "null",
                "followup_months" : 68,
                "geo_lat" : 59.86,
                "pubmed_id" : "18484635",
                "sex" : "female",
                "age" : 59,
                "city" : "uppsala"
        },
        "updated" : ISODate("2017-02-10T17:15:02.380Z"),
        "created" : ISODate("2017-02-10T17:15:02.380Z")
}
```


Contact:

Bo Gao, Michael Baudis
Institute of Molecular Life Sciences,
Swiss Institute of Bioinformatics,
University of Zürich, Switzerland
Email: bo.gao@imls.uzh.ch

Universität Zürich[UZH]

Swiss Institute of Bioinformatics

Global Alliance for Genomics & Health

@ELIXIREurope