

Genome Data for Personalised Medicine

Advancing the Global Alliance for Genomics and Health data schemas through *data-driven* implementations

Michael Baudis - [BC]² 2017



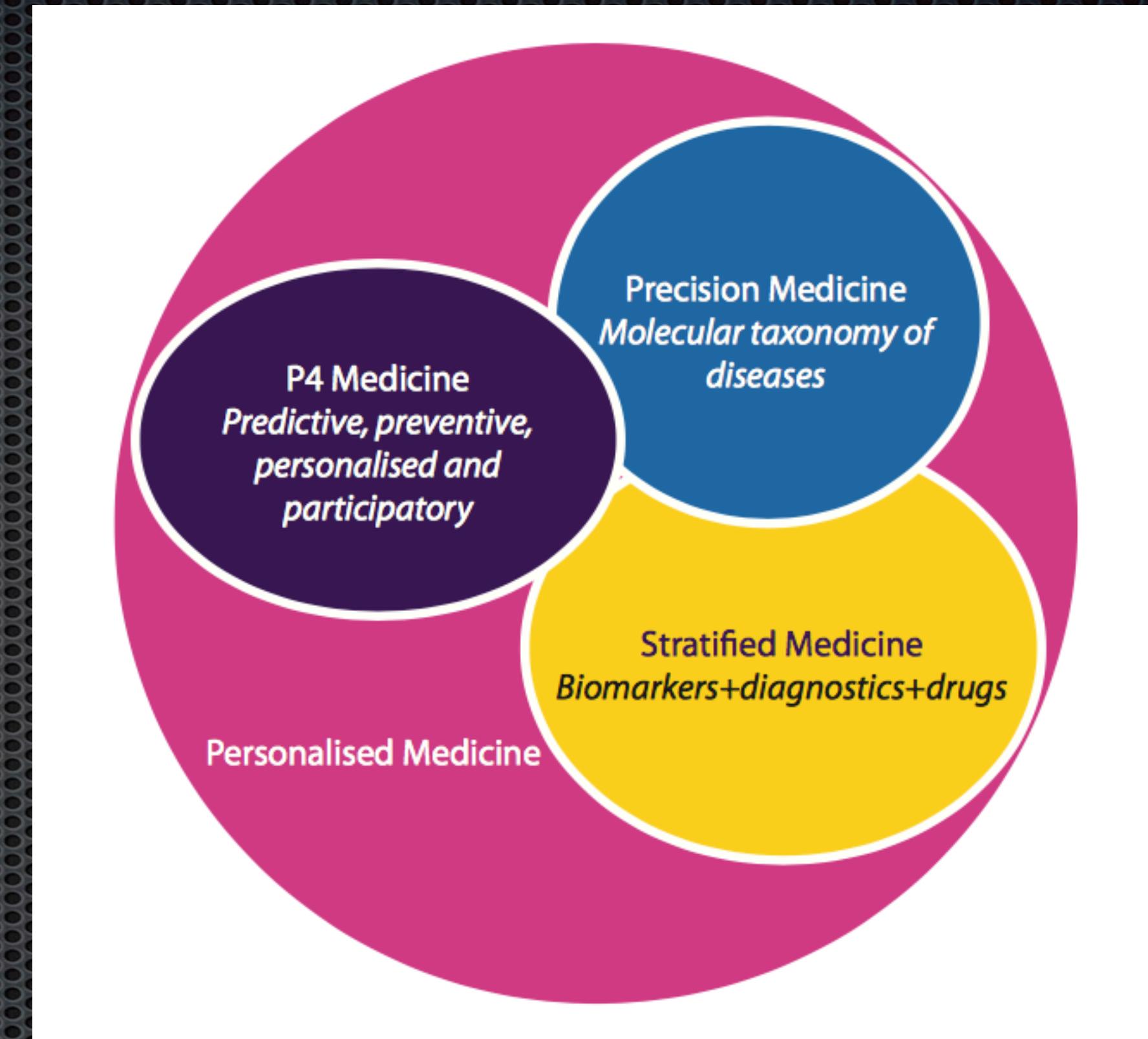
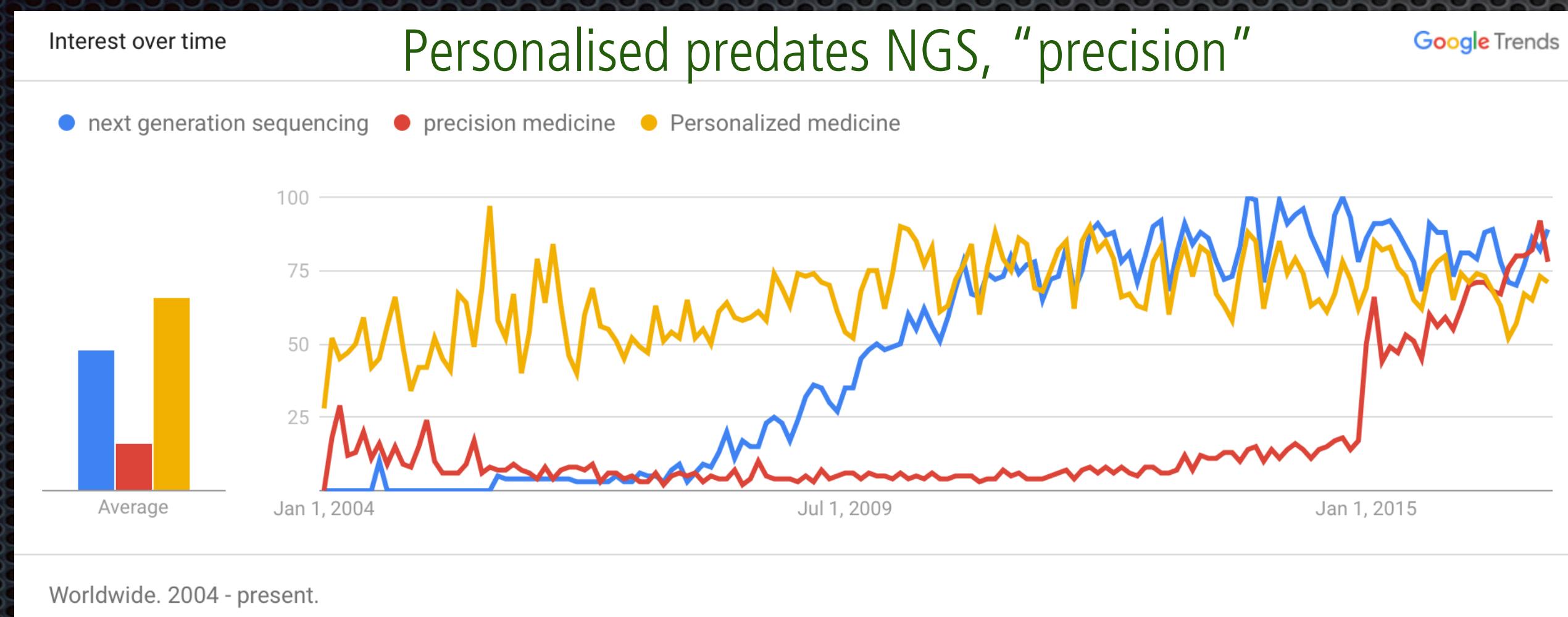
University of
Zurich^{UZH}



Genomic Background + Disease Parameters

Personalised Medicine **Precision Medicine**

...

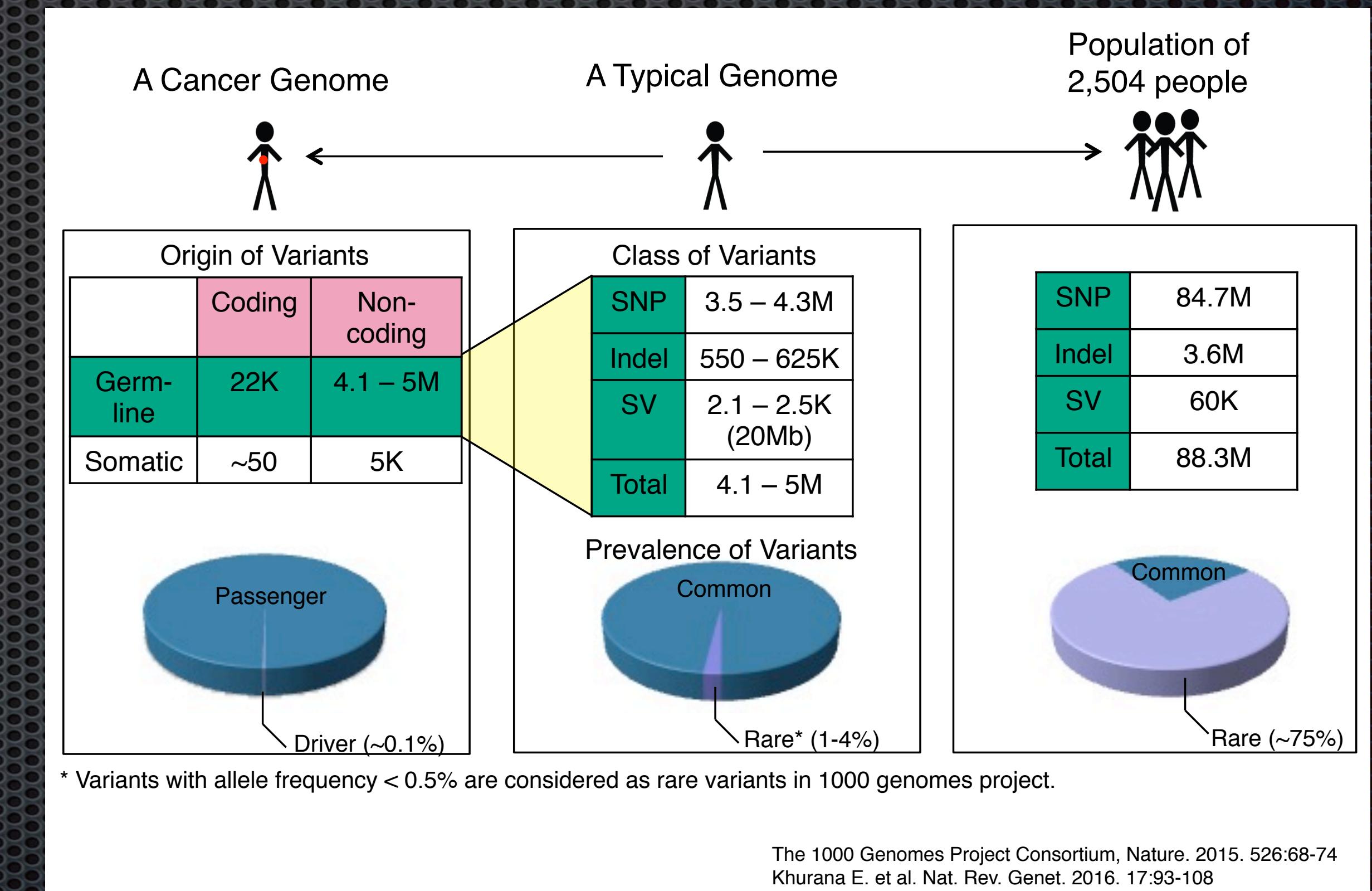


Source: PHG Foundation

While medicine has always been "personal" and "precise" in the given context of available knowledge and technologies, the concept of "**Personalised Medicine**" describes **the use of individual genome information and individually targeted therapies**.

Finding Somatic Mutations In Cancer: Many Needles in a Large Haystack

- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "rare"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease



Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

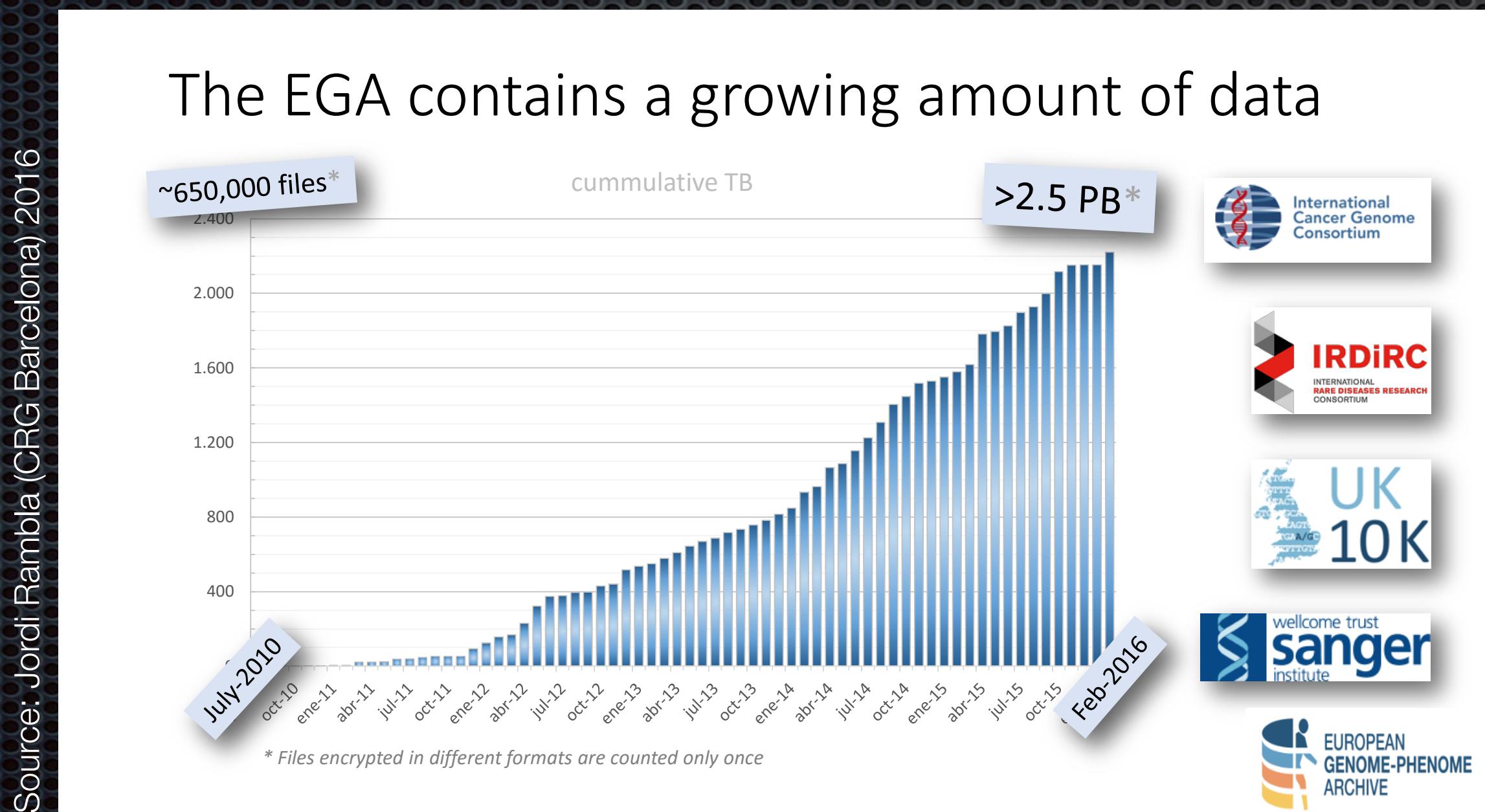
Genomes Everywhere

Large Genome Data Generation, Analysis & Sharing Initiatives

Organization / Initiative: Name	Organization / Initiative: Category	Cohort	
100K Wellness Project	Research Project	107 unaffected individuals (scaling up to 100,000)	>1'000'000
23andMe	Organization	>1 million customers (>80% consented to research)	
Actionable Cancer Genome Initiative (ACGI)	Data-Sharing Project	Goal: 100,000 individuals	
Ancestry.com	Organization	1.4 million customer DNA samples (what % consented to research?)	100'000
BioBank Japan	Repository	Specimens from >200,000 patients and unaffected controls	
Cancer Moonshot2020	Consortium	Phase 1: 20,000 cancer patients	20'000+
Children's Hospital of Philadelphia Biorepository	Repository	Capacity for 8.6 million samples	
China Kadoorie Biobank	Repository	>512,000 participants (general population, China). Genotyping data available for ~100,000.	
CIMBA	Consortium	>15,000 BRCA1 carriers, >8,000 BRCA2 carriers	
Clinical Sequencing Exploratory Research (CSER)	Consortium	~4,000 patients and healthy controls	
DECIPHER	Repository	19,014 patients (international)	
deCode Genetics	Organization	500,000 participants (international)	
East London Genes & Health	Research Project	100,000 unaffected individuals (East London, Pakistani or Bangladeshi heritage)	
Electronic Medical Records and Genomics (eMERGE) Network	Repository, Consortium, Research Project	55,028 patients	
European Network for Genetic and Genomic Epidemiology (ENGAGE)	Research Project	80,000 GWAS scans, and DNA and serum/plasma from >600,000 individuals	
Exome Aggregation Consortium (ExAC)	Consortium	60,706 individuals	
GENIE/AACR	Data-Sharing Project	>17,000 cancer patients (international)	17'000+
Genome Asia 100K	Consortium	Goal: 100,000 individuals (Asia)	
Genomics England	Organization	Goal: 100,000 genomes from 70,000 individuals (rare disease & cancer patients, and their relatives)	
GoT2D	Consortium, Data-Sharing Project	Multiple case-control cohorts	
International Cancer Genome Consortium (ICGC)	Consortium	currently data from >16'000 samples	16'000+
International Genomics of Alzheimer's Project (IGAP)	Consortium	40,000 patients with Alzheimer's disease	
International Multiple Sclerosis Genetics (IMSG) Consortium	Consortium	Goal: >50,000 patients with MS	
Kaiser Permanente: Genes, Environment, and Health (RPGEH)	Repository, Research Project	200,000 DNA samples (scaling up to 500,000)	>2-500'000
Leiden Open Variation Database (LOVD)	Repository	>170,000 individuals	
Million Veteran Program	Research Project	Goal: 1 million individuals; first 200,000 is complete.	
MyCode® Community Health Initiative	Repository, Research Project	Goal: >250,000 patients	
Precision Medicine Initiative	Research Project	Goal: >1 million participants, starting in 2016 (US)	>1'000'000
Psychiatric Genomics Consortium (PGC)	Consortium	>170,000 subjects	
Resilience Project	Research Project	589,306 individuals	
Saudi Human Genome Program	Research Project	Goal: ~100,000 patients and controls (Saudi Arabia)	
Scottish Genomes Partnership (SGP)	Research Project	>3,000 individuals (Scotland)	
T2D-GENES	Consortium, Data-Sharing Project	10,000 patients and controls (five ethnicities); 600 individuals (Mexican American)	
TBResist	Consortium	>2,600 samples	
UK Biobank	Repository, Consortium, Research Project	500,000 individuals (age 40-69 years; UK)	500'000
UK10K	Research Project	10,000 participants (6,000 patients and 4,000 controls)	
Vanderbilt's BioVU	Repository	>215,000 samples	

Genome Datasets: Rapid Growth, Limited Access

population based and cancer research studies produce a rapidly increasing amount of genome sequence data



genome data is stored in an increasing number of institutional and core repositories, with **incompatible data** structures and **access** policies

GA4GH to solve genome
data access....

GA4GH HISTORY & MILESTONES

- January - June 2013 - Meeting & White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
 - March 2014 - **Working groups** established in Hinxton & 1st plenary in London
 - October 2014 - Plenary meeting, San Diego; interaction with ASHG meeting
 - June 2015 - 3rd Plenary meeting, Leiden
 - September 2015 - GA4GH at ASHG, Baltimore
 - October 2015 - DWG / New York Genome Centre
 - April 2016 - Global Workshop @ ICHG 2016, Kyoto
 - October 2016 - 4th Plenary Meeting, Vancouver
 - May 2017 - Strategy retreat, Hinxton
 - October 2017: 5th Plenary Meeting, Orlando
- “GA4GH II”: **Work Streams** && **Driver Projects**

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



Global Alliance
for Genomics & Health

GA4GH API promotes sharing

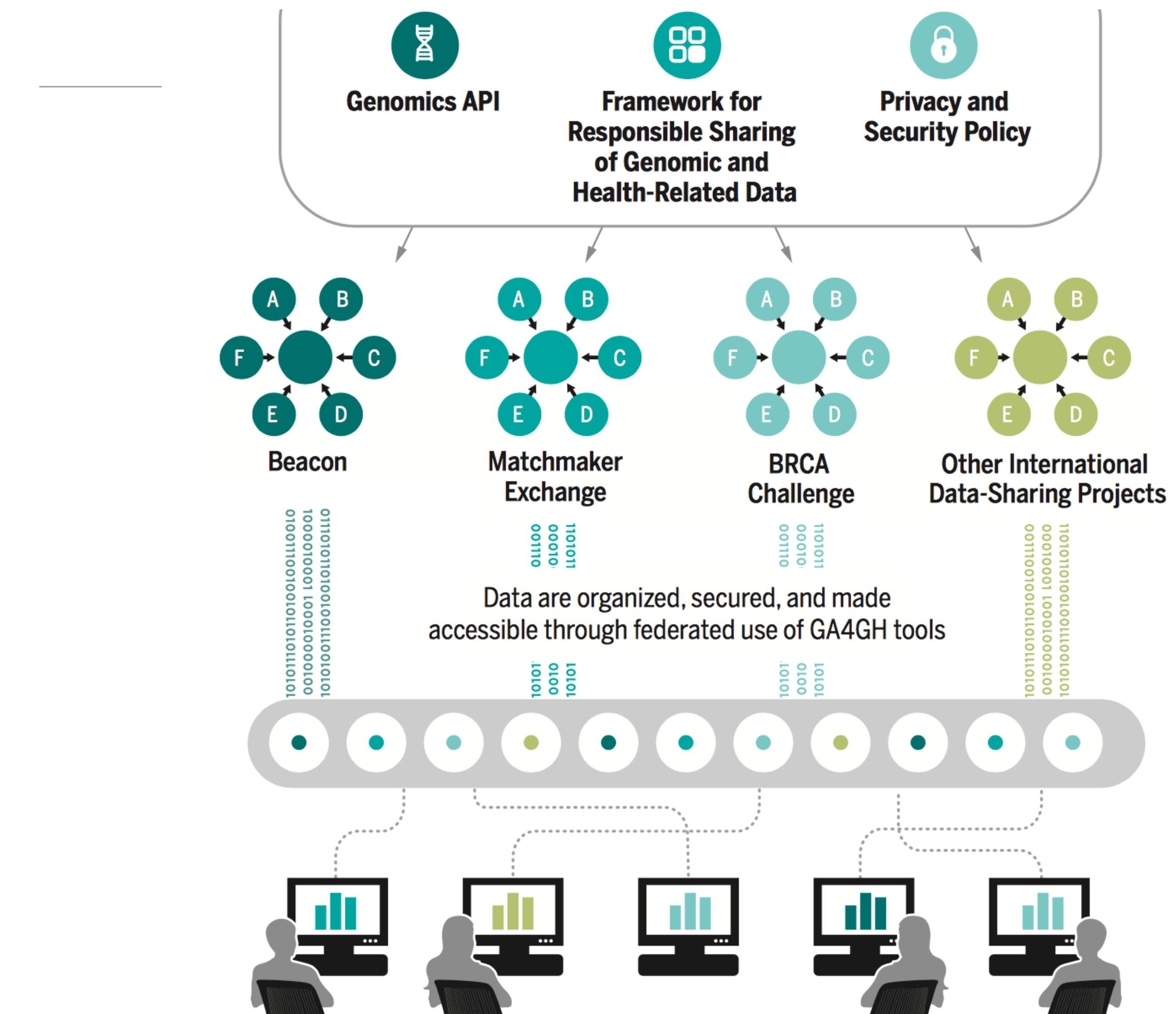
A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



GENOMICS

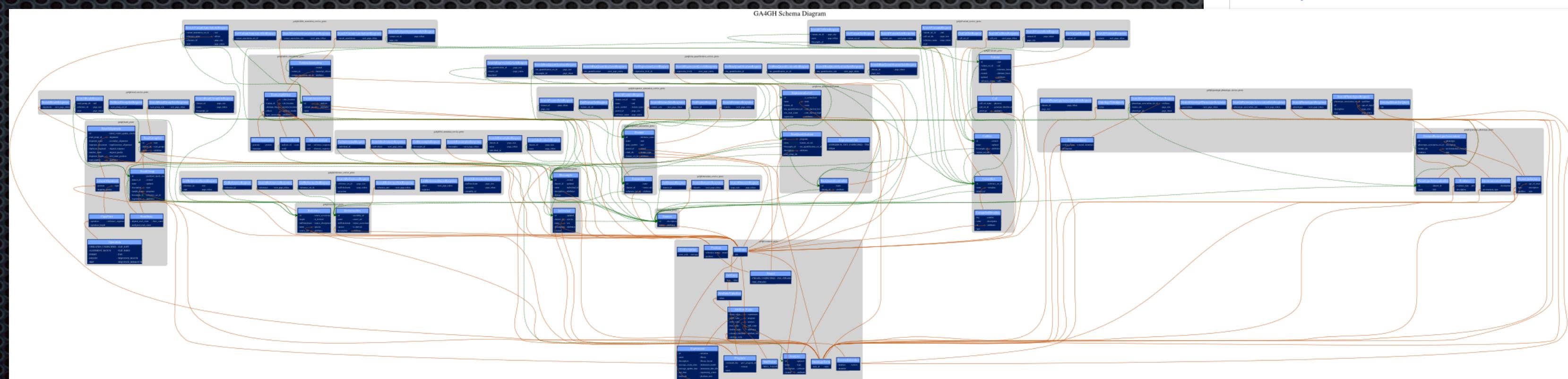
A federated ecosystem for sharing genomic, clinical data

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



GA4GH Data **Working** Group

- development of schemas for genome data & metadata representation
- reference server, documentation
- “all in one” master schema



This repository Search Pull requests Issues Gist

ga4gh / schemas

Unwatch 115 Star 196 Fork 110

Code Issues 152 Pull requests 29 Projects 1 Wiki Pulse Graphs

Work on data models and APIs for Genomic data. <http://ga4gh.org/#/api>

1,102 commits 17 branches 16 releases 46 contributors Apache-2.0

Branch: metadata-integ... New pull request Create new file Upload files Find file Clone or download

This branch is 15 commits ahead, 3 commits behind master. Pull request Compare

mbaudis Merge branch 'master' into metadata-integration Latest commit 077c2c7 2 days ago

doc Merge branch 'master' into metadata-integration 2 days ago

python Add constraints file 2 days ago

scripts Utilize new common methods in schemas 2 days ago

src/main/proto Merge branch 'master' into metadata-integration 13 days ago

tests Utilize new common methods in schemas 2 days ago

tools Merge branch 'master' into metadata-integration 13 days ago

.gitignore Remove protoc call from install path (#781) 7 days ago

.travis.yml Add constraints file 2 days ago

proto3. 10 months ago



Global Alliance
for Genomics & Health

GA4GH Driver Projects

BRCA Challenge

The BRCA Challenge aims to advance understanding of the genetic basis of breast and other cancers using data from around the world.



Beacon Project

Beacon Project is an open web service that tests the willingness of international sites to share genetic data. It is being implemented on the websites of the world's top genomic research organizations.



Matchmaker Exchange

Matchmaker Exchange is a federated network of databases whose goal is to find genetic causes of rare diseases by matching similar phenotypic and genotypic profiles.



The Cancer Gene Trust

proposes to aggregate somatic cancer mutation data and some clinical data in order to improve the genomic landscape of actionability in some cancers and to enable greater personalized clinical care for individuals with rare cancer mutations.



Beacon Project

An open web service that tests the willingness of international sites to share genetic data.



Beacon Network

Search Beacons

A global search engine for genetic mutations.

GRCh37 ▾ e.g. 1: 100,000 A>C Search

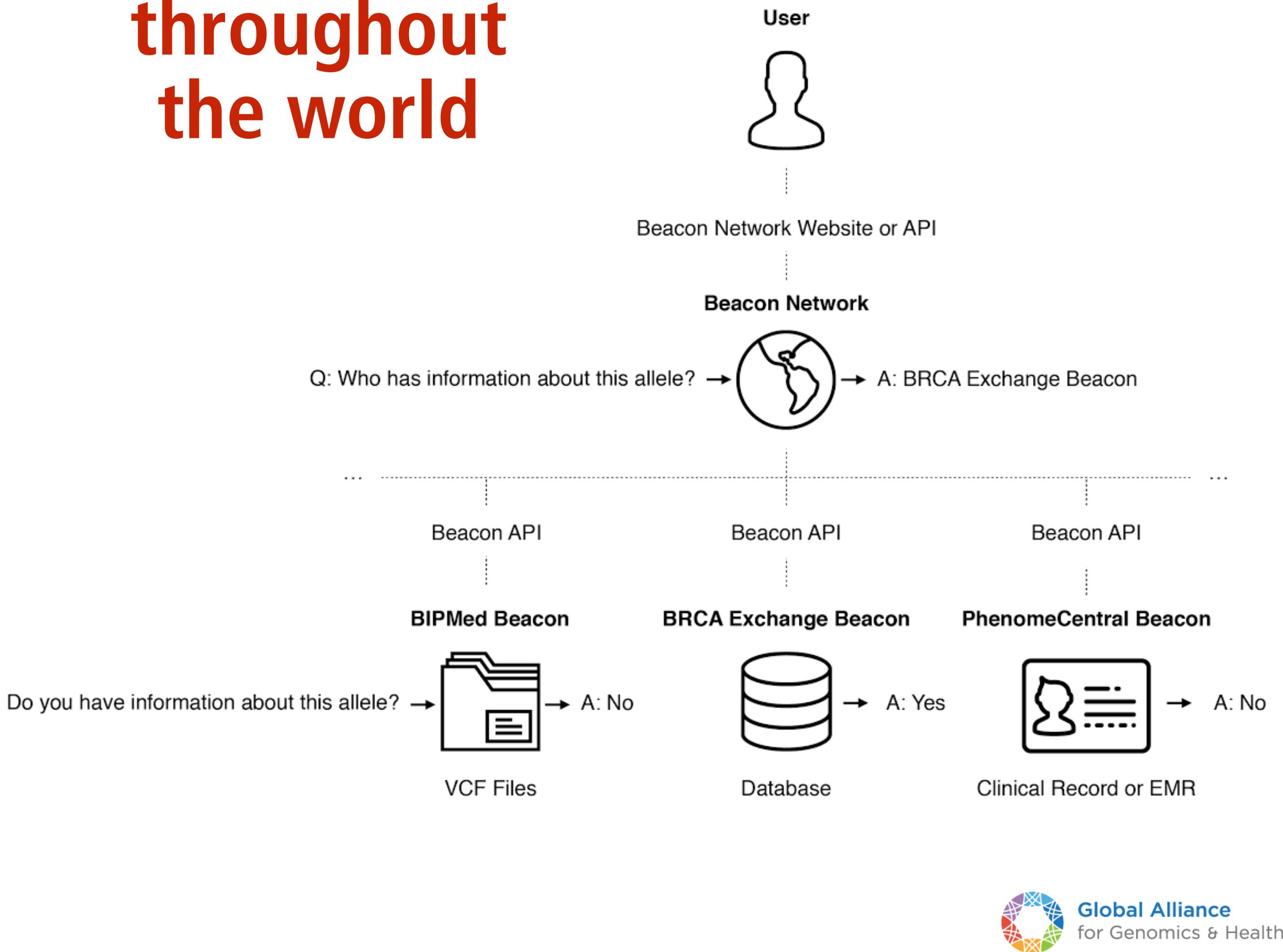
Quickstart: Search for a BRCA2 variant

Find genetic mutations shared by these organizations

- Global Gene Corp
- BRCA EXCHANGE
- Google
- BIPMed Beacon
- PC
- PhenomeCentral Beacon
- Clinical Record or EMR

Browse Beacons »

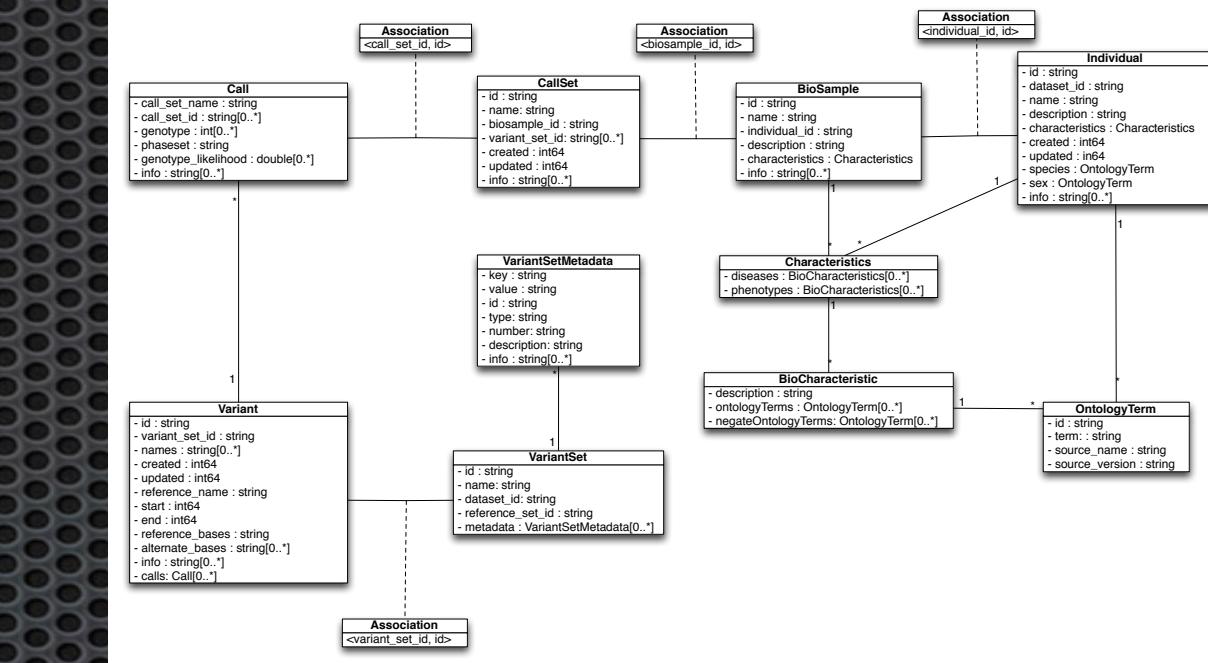
> 50 Beacons throughout the world



Developing the GA4GH Metadata Schema

▶ arrayMap for GA4GH

- metadata schema development through implementation of arrayMap resource data
- OntologyTerm objects for biodata
- implementation w/ ontology services



```

{
  "_id" : ObjectId("58297ca32ca4591e5a0df054"),
  "id" : "AM_V_1778741",
  "variant_set_id" : "AM_VS_HG18",
  "reference_name" : "10"
  "start" : 579049,
  "end" : 17236099,
  "alternate_bases" : "DUP",
  "reference_bases" : ".",
  "info" : {
    "svlen":16657050,
    "cipos": [
      -1000,
      1000
    ],
    "ciend": [
      -1000,
      1000
    ]
  },
  "calls" : [
    {
      "genotype" : [
        ".",
        "."
      ],
      "call_set_id" : "AM_CS_TCGA-61-1917-01A-01D-0648-01",
      "info" : {
        "segvalue" : 0.5491
      }
    },
    {
      "created" : ISODate("2016-11-14T08:33:58.202Z"),
      "updated" : ISODate("2016-11-14T08:33:58.202Z"),
      ...
    }
  ]
}
  
```

Driving Beacon Development

▶ Beacon⁺

- CNV/CNA as first type of structural variants
- disease specific queries
- quantitative reporting

arrayMap

Resource for copy number variation data in cancer

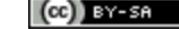
arrayMap 

- [Search Samples](#)
- [Search Publications](#)
- [Gene CNA Frequencies](#)
- [User Data](#)
- [Array Visualization](#)
- [Progenetix](#)

 **University of Zurich** 

- [Citation](#)
- [User Guide](#)
- [Registration & Licensing](#)
- [People](#)
- [External Links ↗](#)

[FOLLOW US ON !\[\]\(58b939658e30b77fcda4f0badcc2af08_img.jpg\) !\[\]\(7a826f8dea09f13e2fee55c90bb7dad2_img.jpg\)](#)

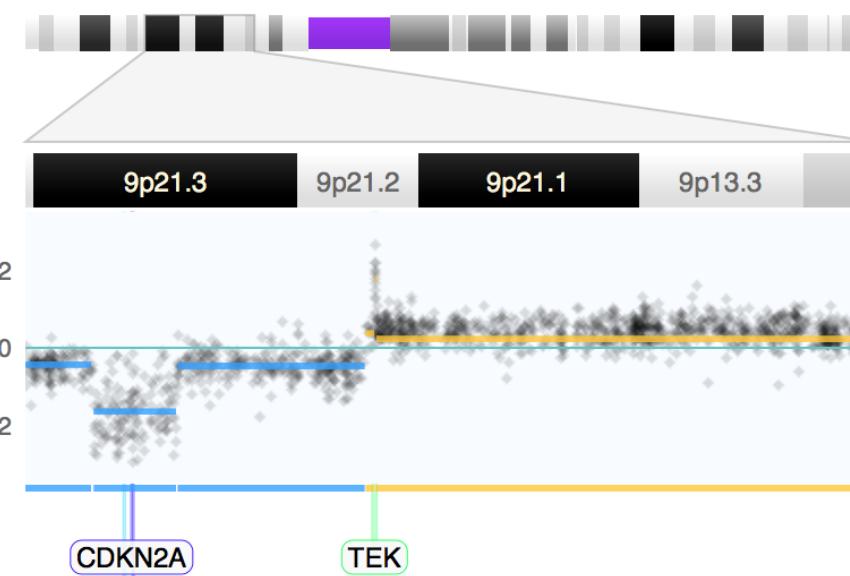
 130.60.23.21

visualizing cancer genome array data @ 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 data integration of high-resolution oncogenomic CNA data.

The current data reflects:

-  63060 genomic copy number arrays
-  763 experimental series
-  145 array platforms
-  141 ICD-O cancer entities
-  554 publications (Pubmed entries)



Genomic copy number imbalances on chromosome 9 in a case of Glioblastoma ([GSM491153](#)), indicating, among others, a homozygous deletion involving CDKN2A/B.

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.

BRAIN TUMOURS	5653 samples ↗	[?]
BREAST CANCER	8329 samples ↗	[?]
COLORECTAL CANCER	3238 samples ↗	[?]
PROSTATE CANCER	991 samples ↗	[?]
STOMACH CANCER	1062 samples ↗	[?]

ARRAYMAP NEWS

- [2016-08-03: SVG graphics](#)
- [2016-05-17: Transitioning to Europe PMC](#)
- [More news ...](#)

Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project or a special license.

© 2000 - 2016 Michael Baudis, refreshed Mon, 19 Sep 2016 10:20:09 GMT in 6.87s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.

ICD-O
Locus
S E
HG18
HG19

ICD Morphologies

2021 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.

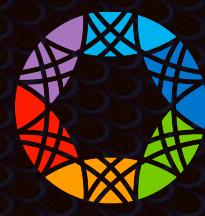
9470/3: Medulloblastoma, NOS (M-94703)

Synonyms

- Medulloblastoma, NOS
- Melanotic medulloblastoma



UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	GSE40834	24047479	9823/3	C42
GSM1002669	GSE40834	24047479	9823/3	C42
GSM1002670	GSE40834	24047479	9823/3	C42
GSM1002671	GSE40834	24047479	9823/3	C42
GSM1002672	GSE40834	24047479	9823/3	C42
GSM1002673	GSE40834	24047479	9823/3	C42
GSM1002674	GSE40834	24047479	9823/3	C42
GSM1002675	GSE40834	24047479	9823/3	C42
GSM1002676	GSE40834	24047479	9823/3	C42
GSM1002677	GSE40834	24047479	9823/3	C42
GSM1002678	GSE40834	24047479	9823/3	C42
GSM1002679	GSE40834	24047479	9823/3	C42
GSM1002680	GSE40834	24047479	9823/3	C42



- object model instead of columnar
- referencing of ontologies instead of text descriptors
- need for **ontologies** & mappings
- **these** are no “real” open ontologies
- data standards use (e.g. ISO)
- fallback to generic object map for unassigned data; this should disappear over time

GA4GH DWG Metadata Task Team: Schema Wranglers

```
"id" : "PGX_AM_BS_GSM510730",
"individual_id" : "PGX_IND_GSM510730",
"name" : "PGX_AM_BS_GSM510730",
"description" : "breast carcinoma",
"bio_characteristics" : [
  {
    "description" : "breast carcinoma",
    "ontology_terms" : [
      {
        "term_id" : "ncit:C4017",
        "term_label" : "Ductal Breast Carcinoma"
      },
      {
        "term_id" : "pgx:icdom:8500_3",
        "term_label" : "invasive carcinoma of no special type"
      },
      {
        "term_id" : "pgx:icdot:C50",
        "term_label" : "breast"
      }
    ],
    "negated_ontology_terms" : [ ],
  }
],
"individual_age_at_collection" : "P47Y",
"attributes" : {
  "tnm" : {
    "values" : [
      {
        "string_value" : "T1N0M0"
      }
    ],
    "location" : {
      "geo_label" : "Oslo, Norway",
      "latitude" : 59.91,
      "longitude" : 10.75,
      "geo_precision" : "city"
    },
    "external_identifiers" : [
      {
        "database" : "Pubmed",
        "identifier" : "20592421",
        "relation" : "part_of"
      },
    ],
    "updated" : ISODate("2017-03-20T08:37:07.771Z"),
  }
}
```



Do classifications & ontologies need an Einstein to sort them out?!

ARRAYS NOT OTHERWISE SPECIFIED

GSM393858 C2853 Acute Myeloid Leukemia Not Otherwise Specified 9861/3 C42

GSM302285 C2852 Adenocarcinoma 8140/3 C34

GSM918983 C3222 Medulloblastoma 9480/3 C816

GSM551398 C4017 Ductal Breast Carcinoma 8500/3 C50

GSM412384 C3163 Chronic Lymphocytic Leukemia 9823/3 C42

GSM1218276 C4017 Ductal Breast Carcinoma 8500/3 C50

GSM214412 C2852 Adenocarcinoma 8140/3 C569

GSM1109923 C9306 Soft Tissue Sarcoma 8800/3 C499

GSM211848 C2852 Adenocarcinoma 8140/3 C25

GSM246294 C89426 8087/2 C53

GSM1981528 C4017 Ductal Breast Carcinoma 8500/3 C50

GSM281399 C8949 8500/2 C50

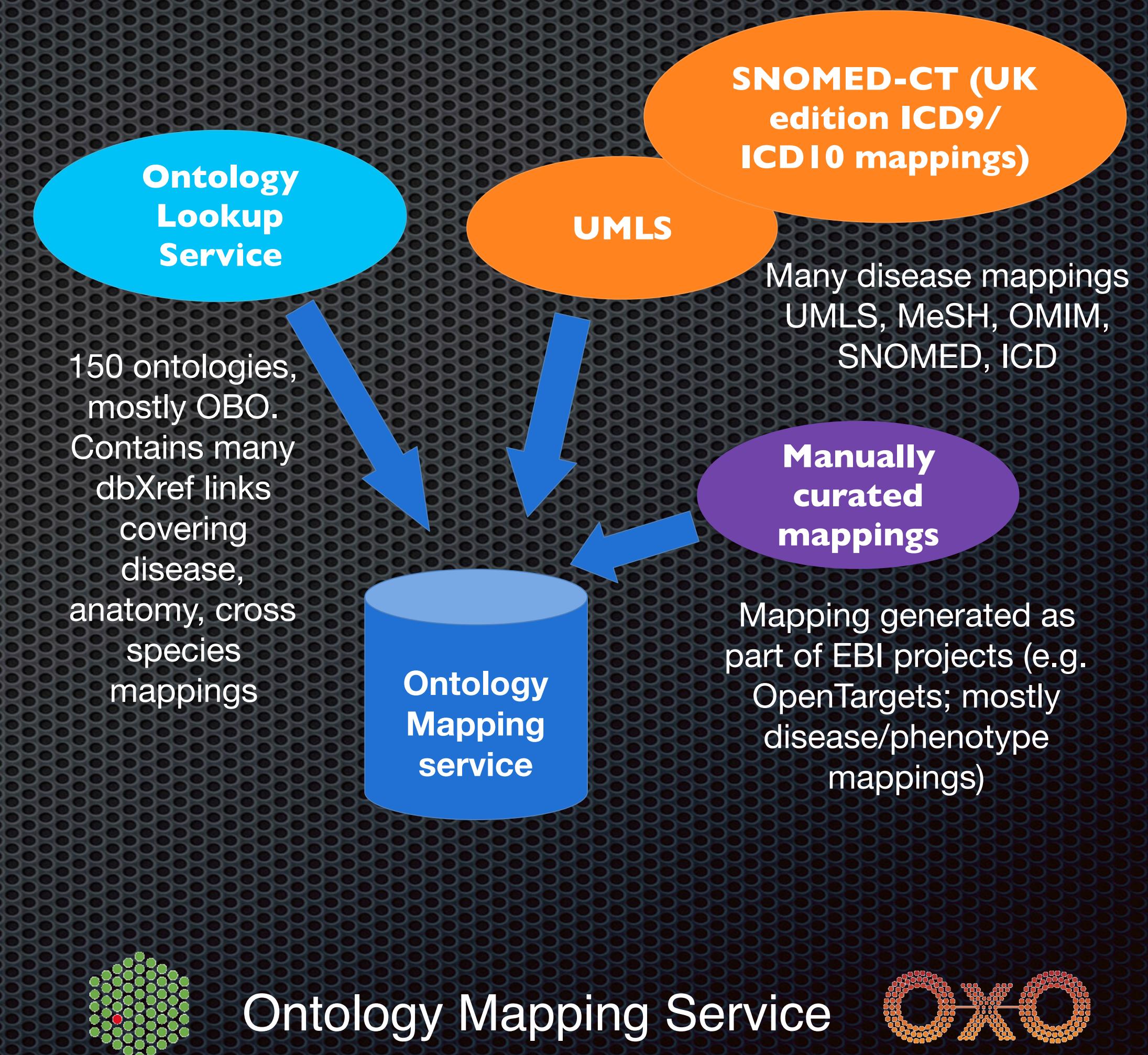
GSM533469 C9349 Plasmacytoma 9831/3 C42



Making Ontologies Work for GA4GH Implementation Studies

- biomedical "metadata" in different resources frequently follows incompatible classification systems
- medical coding systems are driven by different paradigms compared to biological ontologies (e.g. for cross-species comparisons)
- frequently used classifications (ICD, Snomed...) are either not "ontologised" or cannot be referenced in open resources

Federated queries across resources need **curated mappings** of classifications/ontologies



Working towards ontologies w/ arrayMap: Mapping >55'000 samples from ICD-O to NCIt neoplasm core

ICDM	ICDMORPHOLOGY
8021/3	Carcinoma anaplastic type
9451/3	Oligodendrogloma anaplastic
9051/3	Desmoplastic mesothelioma
9732/3	Plasma cell myeloma
8070/3	Squamous cell carcinoma
8380/3	Endometrioid adenocarcinoma
8070/3	Squamous cell carcinoma
8430/3	Mucoepidermoid carcinoma
9680/3	Diffuse large B-cell lymphoma
8800/3	Sarcoma
8441/3	Serous adenocarcinoma
9689/3	splenic marginal zone lymphoma nos
8077/2	Squamous intraepithelial neoplasia grade III
8140/0	Adenoma
8272/3	Pituitary carcinoma
8500/2	Ductal carcinoma in situ
8200/3	Adenoid cystic carcinoma
9370/3	Chordoma
9717/3	Enteropathy type T-cell lymphoma
9698/3	Follicular lymphoma grade 3
9863/3	Chronic myeloid leukemia
8852/3	Liposarcoma myxoid
9080/3	Teratoma malignant
8530/3	Inflammatory carcinoma
8140/3	Adenocarcinoma
8200/3	Adenoid cystic carcinoma

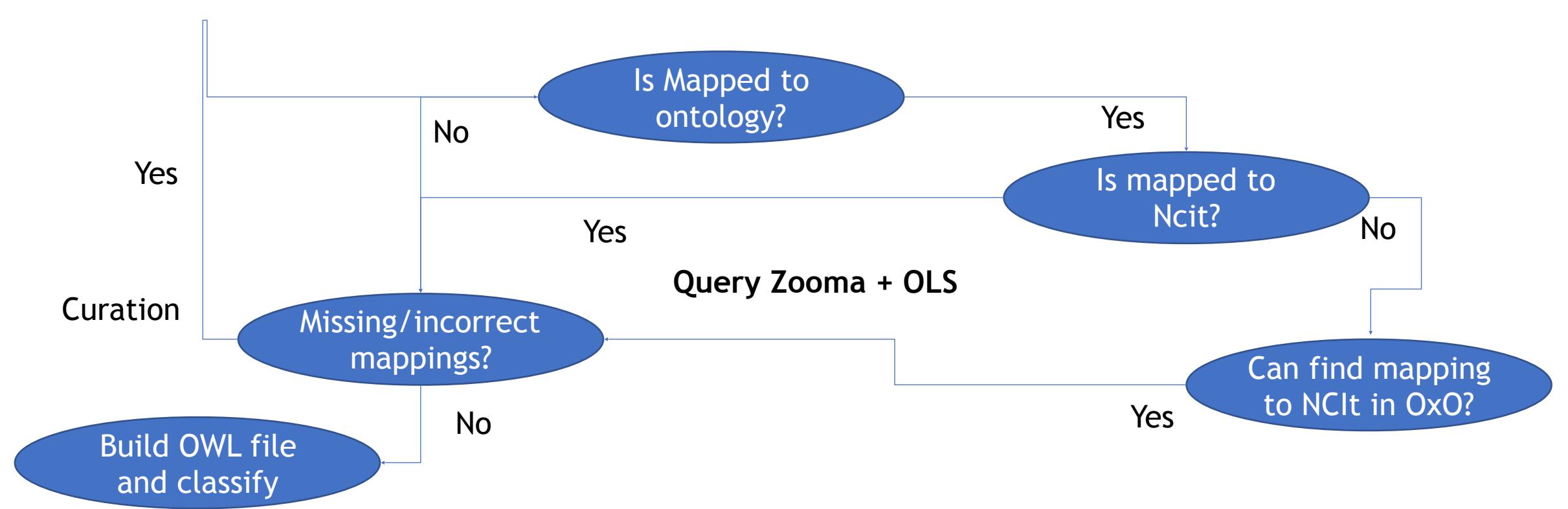
NCItcode	NCItlabel
C4326	anaplastic oligodendrogloma
C6747	
C3242	multiple myeloma
C2926	non-small cell lung carcinoma
C3769	endometrioid carcinoma
C2926	non-small cell lung carcinoma
C45544	pulmonary mucoepidermoid carcinoma
C8851	diffuse large B-cell lymphoma
C9118	sarcoma
C7550	ovarian serous adenocarcinoma
C4196	adenoma
C4536	Pituitary carcinoma
C3641	ductal carcinoma in situ
C2970	adenoid cystic carcinoma
C2947	Chordoma
C3177	chronic myelogenous leukemia
C3735	myxoid liposarcoma
C4872	breast carcinoma
C27745	lung adenocarcinoma
C2670	

ICDT	ICDTOPOGRAPHY
C739	thyroid gland
C719	Brain
C499	connective and soft tissue
C42	hematopoietic and reticuloendothelial systems
C140	pharynx
C54	corpus uteri
C44	skin
C089	salivary gland
C42	hematopoietic and reticuloendothelial systems
C559	uterus nos
C570	fallopian tube
C422	spleen
C53	cervix uteri
C189	large intestine excl. rectum and rectosigmoid junction
C751	pituitary gland
C50	breast
C32	larynx
C419	bone
C17	small intestine
C42	hematopoietic and reticuloendothelial systems
C42	hematopoietic and reticuloendothelial systems
C499	connective and soft tissue
C809	unknown
C50	breast
C809	unknown
C12	uterus nos

NCItcode	NCItlabel
C12400	thyroid gland
C12439	brain
C12316	
C12470	zone of skin
C12426	saliva-secreting gland
C12403	fallopian tube
C12432	spleen
C12311	
C12399	pituitary gland
C12971	breast
C12420	larynx
C13076	bone tissue
C12386	small intestine
C35882	Hereditary elliptocytosis
C12971	breast
C35882	Hereditary elliptocytosis
C12762	oropharynx
C12415	kidney
C12499	internal ear
C12683	bronchus
C12343	retina
C12393	pancreas
C12422	tongue
C12390	rectum
C12404	female gonad
C12391	

NCIt_mapped	NCIt_mapped_ICDM_T_label
C3878	Thyroid Gland Undifferentiated (Anaplastic) Carcinoma
C4326	Anaplastic Oligodendrogloma
C6747	Desmoplastic Mesothelioma
C3242	Plasma Cell Myeloma
C102872	Pharyngeal Squamous Cell Carcinoma
C6287	Endometrial Endometrioid Adenocarcinoma
C4819	Skin Squamous Cell Carcinoma
C5953	Minor Salivary Gland Mucoepidermoid Carcinoma
C8851	Diffuse Large B-Cell Lymphoma
C9306	Soft Tissue Sarcoma
C40101	Serous Adenocarcinoma
C4663	Splenic Marginal Zone Lymphoma
C89476	Grade III Vaginal Intraepithelial Neoplasia
C4349	Colon Adenocarcinoma
C4536	Pituitary Gland Carcinoma
C2924	Ductal Breast Carcinoma In Situ
C2970	Adenoid Cystic Carcinoma
C2947	Chordoma
C4737	Enteropathy-Associated T-Cell Lymphoma
C3460	Grade 3 Follicular Lymphoma
C3174	Chronic Myelogenous Leukemia BCR-ABL1 Positive
C27781	Myxoid Liposarcoma
C3403	Tetrotoma
C4001	Inflammatory Breast Carcinoma
C2852	Adenocarcinoma
C2970	Adenoid Cystic Carcinoma
C3158	Leiomyosarcoma
C2970	Adenoid Cystic Carcinoma
C2923	Bronchioloalveolar Carcinoma
C3224	Melanoma
C8459	Hepatosplenic T-Cell Lymphoma
C8294	Pancreatic Adenocarcinoma
C3996	Monoclonal gammopathy of Undetermined Significance
C4817	Ewing Sarcoma
C3288	Oligodendrogloma
C4648	Tongue Squamous Cell Carcinoma
C2862	Primary Myelofibrosis
C4833	Oral Cavity Squamous Cell Carcinoma
C9383	Rectal Adenocarcinoma
C3158	Leiomyosarcoma
C3898	Extranodal Marginal Zone Lymphoma of Mucosa-Associated Lymphoid Tissue
C4512	Ovarian Mucinous Cystadenoma
C5519	Other

- From 456 pairs of ICD-O terms Morphology and Topography representative of cancer entities in arrayMap
- Develop Python script to take ICD-O Morphology and Topography labels separately QUERY ZOOMA, Oxo and OLS to find mapping to NCIt



From 456 pairs of ICD-O
70% ICD-O Morphology - NCIt
65% ICD-O Topography - NCIt

45% ICD-O-3 Pairs mapped to NCIt terms

=> MANUAL CURATION of >50%

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set
(MacKay *et al.*, Cancer Cell 2017, in print)
- structural variations (DUP, DEL) in addition to SNV
- diagnosis queries using ontology codes (NCIT, ICD-O)
- quantitative responses
- current version uses **GA4GH schema compatible** database

Beacon+

This forward looking Beacon interface implements additional, planned features beyond the current GA4GH specifications. [Info](#)

Query

Dataset: DIPG (CNV + selected SNV)

Reference name*: 17

Genome Assembly*: GRCh36 / hg18

Variant type*: SNV / indel

Position*: 7577121

Ref. Base(s)*: G

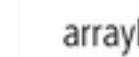
Alt. Base(s)*: A

Bio-ontology: pgx:icdom:9380_3

[Beacon Query](#)

Response

Dataset	Chro.	Assembly	Var. Type	Start Min	Start Max	End Min	End Max	Pos.	Ref.	Alt.	Bio Query	Call Count	Samples	f	Query
arraymap	9	GRCh36	DEL	19000000	21984490	21900000	25000000				pgx:icdom:8140_3	3781	403	0.0065	show JSON
dipg	17	GRCh36	SNV			7577121		G	A	pgx:icdom:9380_3	21	20	0.0187	show JSON	

arrayMap  University of Zurich UZH  This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.   

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set
(MacKay *et al.*, Cancer Cell 2017, in print)
- structural variations (DUP, DEL) in addition to SNV
- diagnosis queries using ontology codes (NCIT, ICD-O)
- quantitative responses
- current version uses **GA4GH schema compatible** database

Beacon+

This forward looking Beacon interface implements additional, planned features beyond the current GA4GH specifications. [Info](#)

Query [SNV Example](#) [CNV Example](#)

Dataset: arrayMap (CNV only)

Reference name*: 9

Genome Assembly*: GRCh36 / hg18

Variant type*: DEL (Deletion)

Start min Position*: 19000000

Start max Position: 21984490

End min Position: 21900000

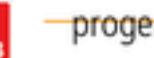
End max Position: 25000000

Bio-ontology: ncit:C3059

[Beacon Query](#)

Response

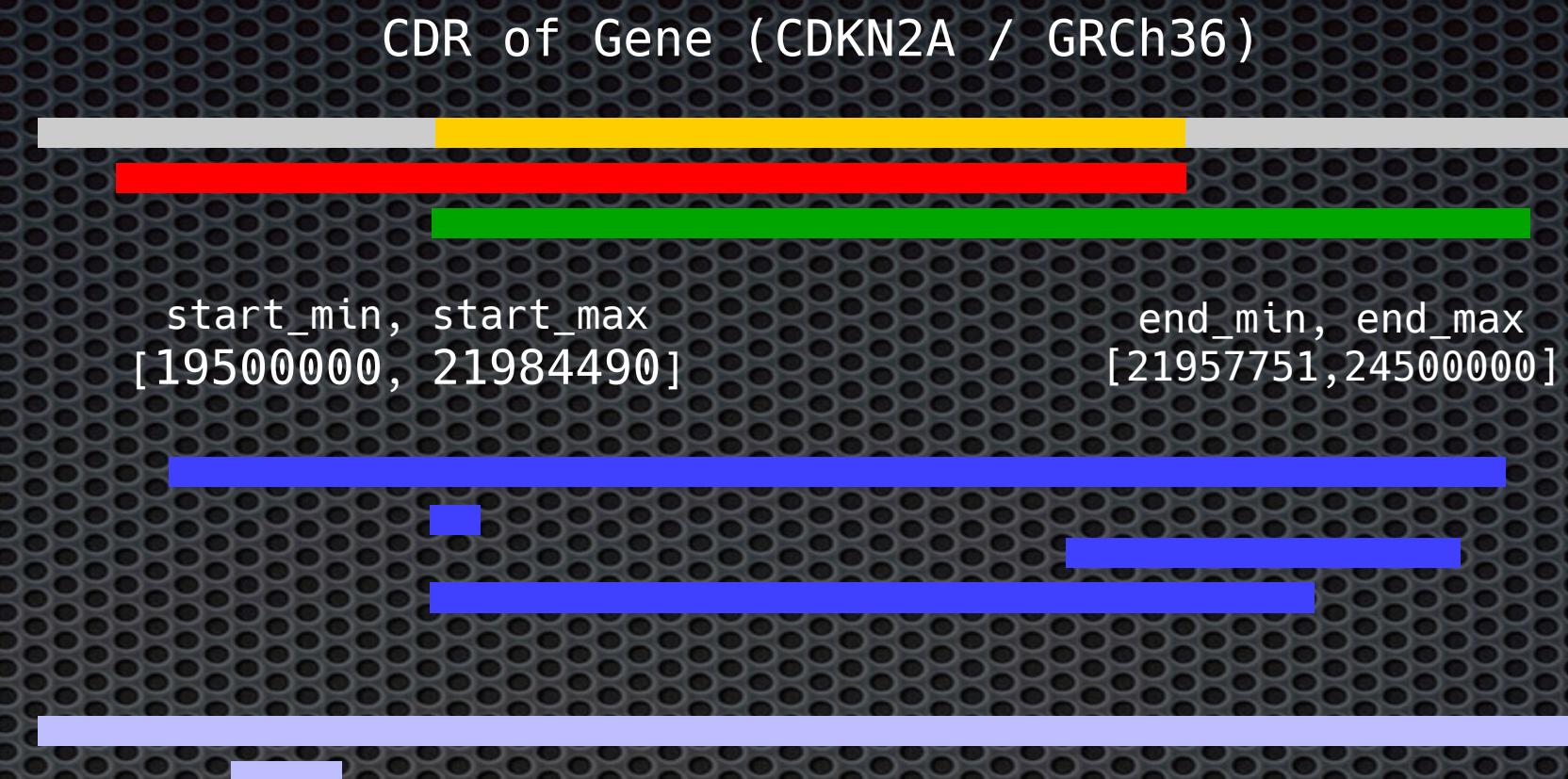
Dataset	Chro.	Assembly	Var. Type	Start Min	Start Max	End Min	End Max	Pos.	Ref.	Alt.	Bio Query	Call Count	Samples	f	Query
dipg	17	GRCh36	SNV					7577121	G	A	pgx:icdom:9380_3	21	20	0.0187	show JSON
arraymap	9	GRCh36	DEL	19000000	21984490	21900000	25000000				ncit:C3059	3781	59	0.001	show JSON

arrayMap  University of Zurich UZH  This Beacon implementation is developed by the Computational Oncogenomics Group at the [University of Zurich](#), with support from the [SIB Technology group](#) and [ELIXIR](#).   

```

  "reference_name" : "0",
  {
    "variant_type" : "DEL" },
    { "start" : { "$gte" : 19500000 } },
    { "start" : { "$lte" : 21984490 } },
    { "end" : { "$gte" : 21957751 } },
    { "end" : { "$lte" : 24500000 } }
  ],
  "api_version" : "0.4",
  "beacon_id" : "org.progenetix.progenetix.beacon",
  "exists" : true,
  "info" : {
    "query_string" :
"dataset_id=arraymap&variants.reference_name=chr9&assembly_id=GRCh36&variants.variant_type=DEL&variants.start_max=19000000&variants.start_min=21984490&variants.end_min=21900000&variants.end_max=25000000&biosamples.bio_characteristics.ontology_terms.term_id=pgx:icdom:9440_3",
    "version" : "Beacon+ implementation based on a development branch of the beacon-team project: https://github.com/ga4gh/beacon-team/pull/94"
  },
  "url" : "http://progenetix.org/beacon/info/",
  "dataset_allele_responses" : [
    {
      "dataset_id" : "arraymap",
      "error" : null,
      "exists" : true,
      "external_url" : "http://arraymap.org",
      "sample_count" : 584,
      "call_count" : 3781,
      "variant_count" : 3244,
      "frequency" : 0.0094,
      "info" : {
        "description" : "The query was against database \\\"arraymap_ga4gh\\\", variant collection \\\"variants_cnv_grch36\\\". 3781 / 59428 matched callsets for 3602919 variants. Out of 62105 biosamples in the database, 2047 matched the biosample query; of those, 584 had the variant."
      },
      "ontology_ids" : [
        "ncit:C3058",
        "pgx:icdom:9440_3",
        "pgx:icdot:C71.9",
        "pgx:icdot:C71.0"
      ]
    }
  ]
}

```



Match using query ranges “at least one base in interval affected”

Example “focal” matches

Mismatches

- Beacon⁺ **range queries** allow the definition of a genome region of interest, containing a specified variant or potentially other position related feature
- “fuzzy” matching of region ends essential for inexact features
- current Beacon implementation addresses CNV (<DUP>,), as are specified in VCF && GA4GH variant schema



```

    },
    "reference_name" : "9" ,
    "variant_type" : "DEL" ,
    "start" : { "$gte" : 19500000 } ,
    "start" : { "$lte" : 21984490 } ,
    "end" : { "$gte" : 21957751 } ,
    "end" : { "$lte" : 24500000 } }

],
"api_version" : "0.4",
"beacon_id" : "org.progenetix:progenetix-beacon",
"exists" : true,
"info" : {
    "query_string" :
"dataset_id=arraymap&variants.reference_name=chr9&assembly_id=GRCh36&variants.variant_type=DEL&variants.start_max=19000000&variants.start_min=21900000&variants.end_max=21984490&variants.end_min=21957751&biosamples
.bio_characteristics.ontology_terms.term_id=pgx:icdom:9440_3"
"dataset_id": "arraymap" implementation based on development branch
of the beacon-team project: https://github.com/ga4gh/beacon-team/pull/94
},
"url" : "http://progenetix.org/beacon/info/",
"dataset_allele_responses" : [
{
    "dataset_id" : "arraymap",
    "error" : null,
    "exists" : true,
    "external_url" : "http://arraymap.org",
    "sample_count" : 584,
    "call_count" : 3781,
    "variant_count" : 3244,
    "frequency" : 0.0094,
    "this": 1
}
]
    "description" : "The query was against database
\"arraymap_ga4gh\", variant collection \"variants_cnv_grch36\". 3781 /
59428 matched callsets for 3602919 variants. Out of 62105 biosamples in
the database, 2047 matched the biosample query; of those, 584 had the
variant.",
    "ontology_ids" : [
        "ncit:C3058",
        "pgx:icdom:9440_3",
        "pgx:icdot:C71.9",
        "pgx:icdot:C71.0"
    ]
}
]
}

```




Beacon+ Concept

Testing Beacons for Data Discovery

- standard Beacon payload (e.g. “exists”)
- testing GA4GH metadata “biocharacteristics” ontology term ids
- multiple datasets can be returned (only one shown here)
- quantitative reporting
- additional information about query & dataset(s)



- GA4GH will be restructured (**work streams**; “real-world” **driver projects**) - to be announced at GA4GH plenary in Orlando, October 15-17, 2017
- developing, distributing and harmonising (meta-)data schema among different driver projects and implementations (clinical/phenotypic data - Phenopackets, genome standards...)
- Beacon as part of “**discovery**” work stream will be expanded, with ELIXIR as project driver
- GA4GH discovery work stream to harmonise APIs, possibly based on Beacon
- We will contribute to both Beacon & schema development, including implementation of necessary tools & examples:
 - example datasets
 - implement Oxo ⇔ arrayMap ontology searches (using the Beacon+ as test/implementation case)
 - schema extension for metagenomic data representation
 - implementing more bio- and other metadata (geolocation data, complex time ...)



Check it Out!

- managed, participation driven projects living on Github: **ga4gh**
 - test datasets & code available through our **progenetix** repositories

- test
 - comment
 - suggest
 - propose
 - complain ...

ga4gh / ga4gh-schemas

Unwatch 108 Star 212 Fork 113

progenetix / beaconplus-server

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights

Unwatch 3 Star 0 Fork 0

progenetix / beaconplus-ui

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights

Unwatch 5 Star 0 Fork 0

ga4gh / beacon-team

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights

Unwatch 36 Star 15 Fork 15

progenetix / arraymap2ga4gh

Code Issues 2 Pull requests 0 Projects 0 Wiki Insights

Unwatch 5 Star 2 Fork 1

No dependencies

GA4GH

Branch: master

This branch

Branch: master

New pull request

85 commits 2 branches 0 releases 3 contributors

Create new file Upload files Find file Clone or download

KyleGao Multi genome editions Latest commit 5db07db 6 days ago

data Multi genome editions 6 days ago

examples update DIPG examples 2 months ago

tools remove the per scripts => beaconplus-server 2 months ago

README.md link 2 months ago

schema.pdf updated schema diagram 6 months ago

README.md

Implementation of the GA4GH schema based on genome profiles and metadata from arrayMap

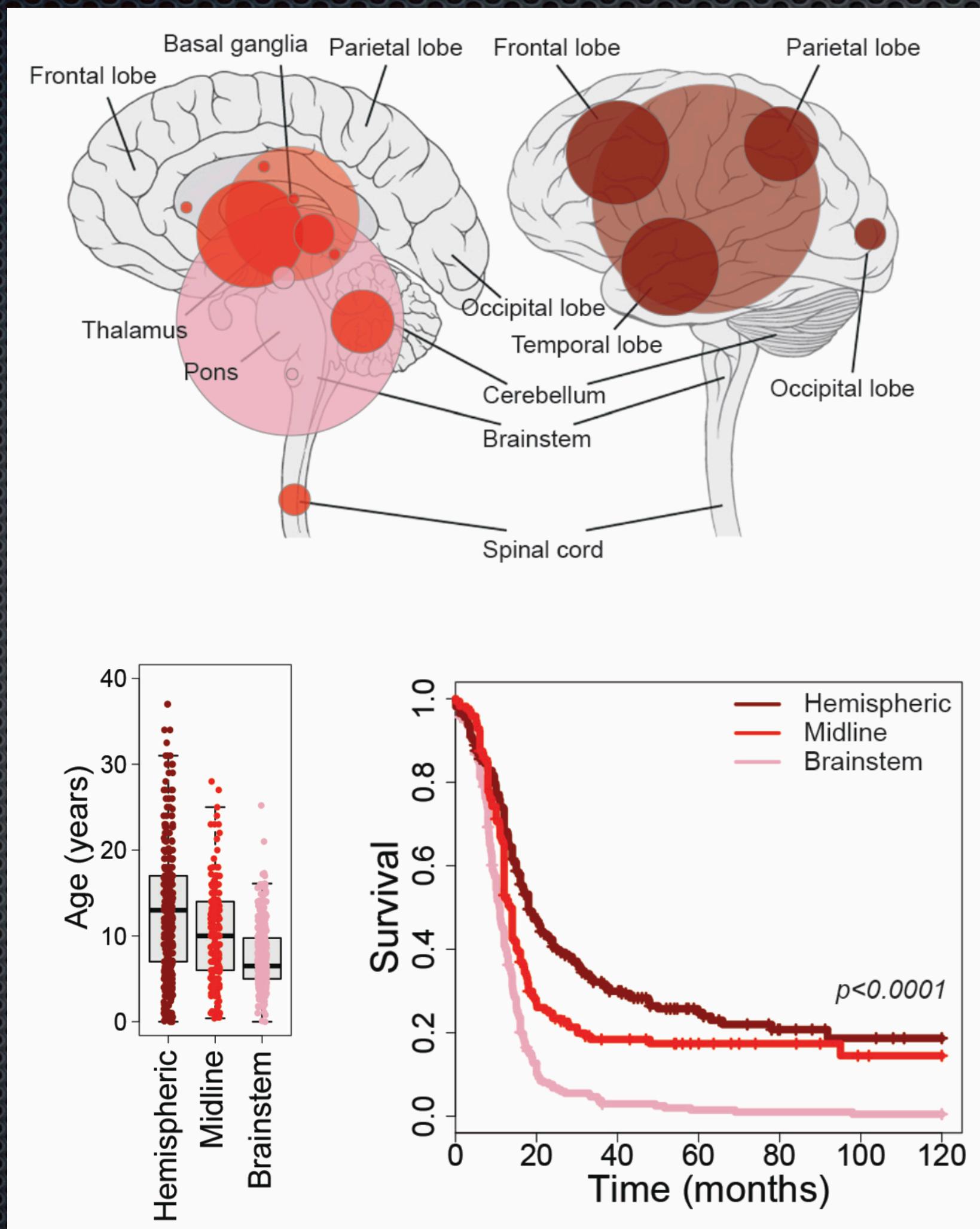
This repository will contain data and information regarding the [arrayMap](#) based implementation of a GA4GH schema structure. While it is not expected that GA4GH compliant resources mirror the schema in their internal structure, this project is aimed at showing the principle feasibility of such an approach, mainly to test & drive schema development.

Data & schemas represented here are not kept in a stable/versioned status, but are updated together with or anticipating GA4GH schema changes.



Implementing real-world datasets for federated access using GA4GH schema specifications: pHGG

Mackay A, Jones C, Baudis M and many, many others:
Integrated molecular meta-analysis of 1000 paediatric high grade and diffuse intrinsic pontine glioma (2017, Cancer Cell, in press)



```
_id : "objectID_591eb7370903744421cc02a",
"individual_id" : "DIPG_IND_0809",
"id" : "DIPG_BS_0809",
"name" : "pHGG_META_0809",
"description" : "glioma, paediatric, high grade",
"individual_age_at_collection" : {
    "age_class" : {
        "term" : "Adult onset",
        "term_id" : "HP:0003581"
    },
    "age" : "P17Y0M"
},
"bio_characteristics" : [
    {
        "ontology_terms" : [
            {
                "term_label" : "Glioma",
                "term_id" : "ncit:C3059"
            },
            {
                "term_label" : "Brain NOS",
                "term_id" : "pgx:icdot:C71.9"
            }
        ],
        "description" : "Juvenile high grade glioma"
    }
],
"external_identifiers" : [
    {
        "database" : "Pubmed",
        "identifier" : "25752754",
        "relation" : "reported_in"
    }
],
"attributes" : {
    "grade" : { "values" : [ { "string_value" : "4" } ] },
    "histone" : { "values" : [ { "string_value" : "wt" } ] }
}
```

BAUDISGROUP @ UZH

NI AI
MICHAEL BAUDIS
(HAOYANG CAI)
PAULA CARRIO CORDO
BO GAO
(LINDA GROB)
SAUMYA GUPTA
(ROMAN HILLJE)
QINGYAO HUANG
(NITIN KUMAR)
(ALESSIO MILANESE)

SIB

HEINZ STOCKINGER
SÉVERINE DUVAUD
VASSILIOS IOANNIDIS
DANIEL TEIXEIRA

THOMAS EGGERMANN
ROSA NOGUERA
REINER SIEBERT
CAIUS SOLOVAN



University of
Zurich UZH



Global Alliance
for Genomics & Health

GA4GH DWG + CWG

JACQUI BECKMANN
ANTHONY BROOKES
MARK DIEKHANS
MARC FIUME
MELISSA HAENDEL
DAVID HAUSSLER
SARAH HUNT
STEPHEN KEENAN
SUZY LEWIS
DAVID LLOYD
MICHAEL MILLER
HELEN PARKINSON
GUNNAR RÄTSCH
DAVID STEINBERG
JULIA WILSON

ELIXIR, CRG, EBI

JORDI RAMBLA DE ARGILA
MELANIE COURTOT
S. DE LA TORRE PERNAS
SUSANNA REPO
SERENA SCOLLEN
TRISH WHETZEL