Implementation studies for the Global Alliance for Genomics and Health data schemas

Using the arrayMap data cancer genome data resource to drive schema development

Michael Baudis - #Biocuration2017 - @Stanford

University of Zurich[UZH]

arrayMap

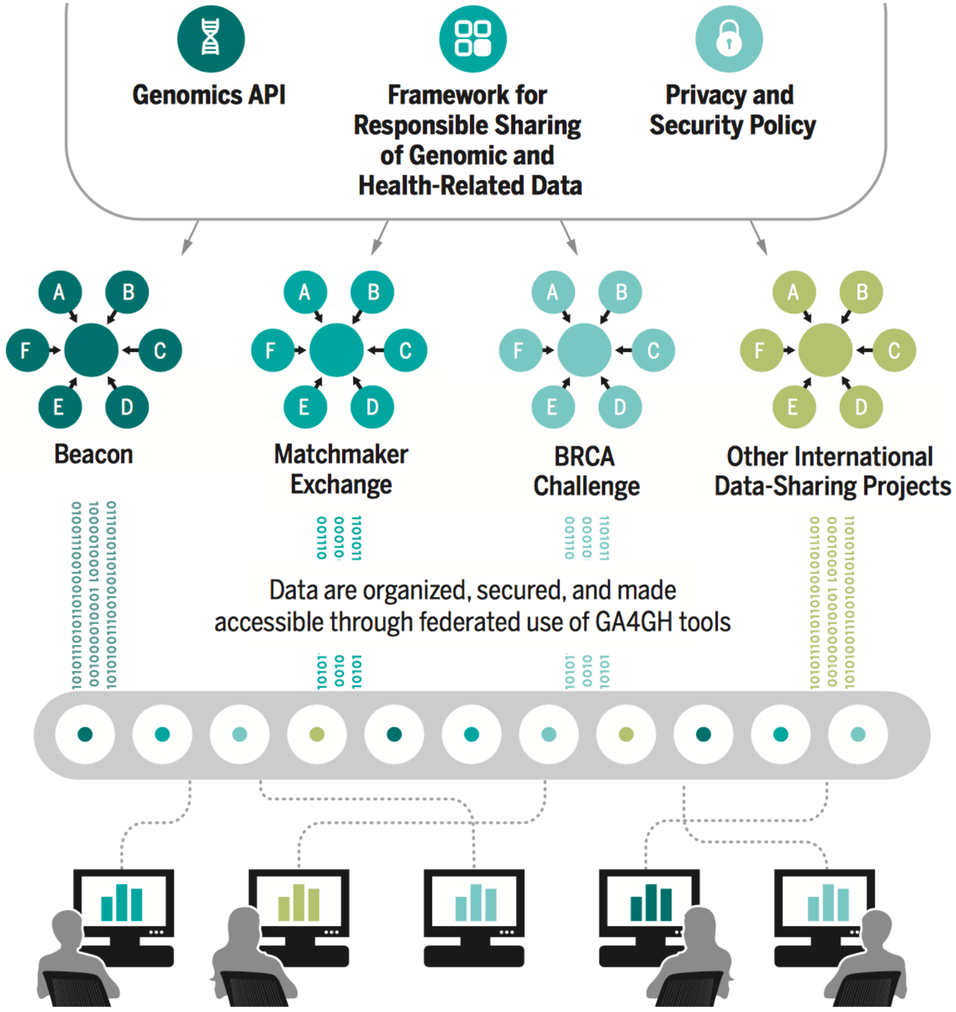
ELIXIR

Global Alliance for Genomics & Health

SIB

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

Global Alliance for Genomics & Health

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

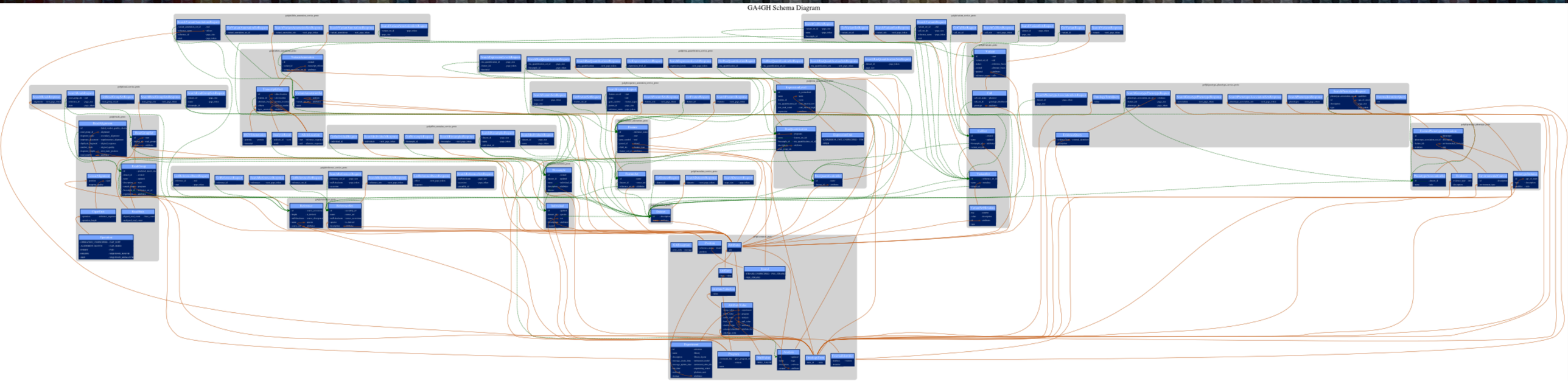
mbaudis Merge branch 'master' into metadata-integration Latest commit `077c2c7` 2 days ago

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
CONTRIBUTING.rst	Convert Avro -> proto3.	10 months ago

The State of the Schema, Feb 2017

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.



Meta Data: Everything but the sequence

Global Alliance for Genomics & Health

May 2014 — Formation

August 2014 — Specification document

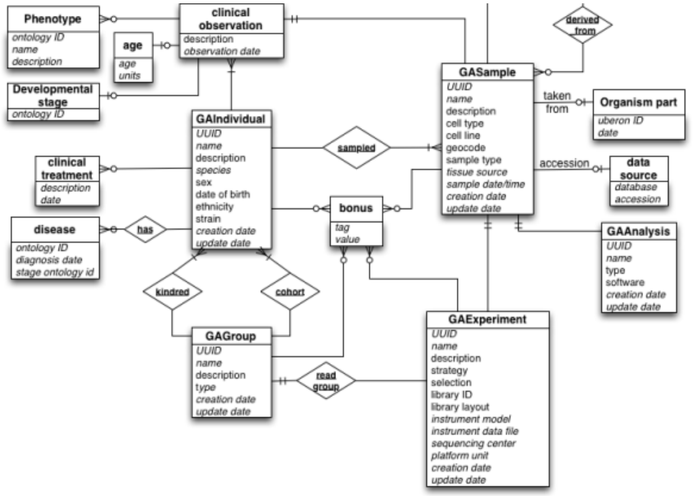
September 2014 — Draft Avro schema

October 2014 — Merged Schema

Frameworks for consistent description of annotated data domains:

- *GAExperiment* Technical data - bridge to Reads Task Team
- *GAIndividual* Clinical data - bridge to Clinical and Regulatory & Ethics
- *GAIndividualGroup* Static or dynamically generated semantic collections
- *GAAnalysis* Interpretation & methodology of one or more
- *GASample* Biological information using common ontologies

- Species neutral but focus on human use cases
- Standardised ontologies for feature annotations
- Next steps:
  - ontology recommendations (with CWG)
  - model refinement through implementation
  - validation and data transformation tools



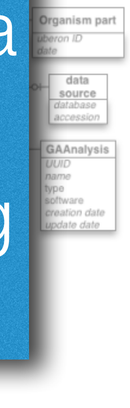
Meta Data: Everything but the sequence

Global Alliance for Genomics & Health

The GA4GH metadata schema aimes at providing a **blueprint** for the development of consistent and accessible APIs and resources.

It is **not** intended to **complete**ly cover areas like e.g. electronic health records, but representation of "research need driven" subsets of such data.

While resources may be developed from the schema specifications, the main use will be an **API layer** to facilitate uniform access to subsets of the underlying data.



“Metadata” Objects

metadata branch, December 2015

Individual
accessions
dateOfBirth
description
developmentalStage
diseases
ethnicity
geographicLocation
guid
id
info
interventions
name
observations
phenotypes
recordCreateTime
recordUpdateTime
sex
species
strain

BioSample
accessions
ageAtSampling
cellLine
cellType
description
geographicLocation
guid
id
individualIds
info
interventions
name
observations
organismPart
preservationMethod
recordCreateTime
recordUpdateTime
samplingDate
sex
species
specimenType

Experiment
description
id
info
instrumentDataFile
instrumentModel
molecule
name
platformId
platformName
preparationId
processingFacility
recordCreateTime
recordUpdateTime
runTime
selection
strategy

```
record OntologyTerm {
    union { null, string } ontologySourceID = null;
    union { null, string } ontologySourceName = null;
    union { null, string } ontologySourceVersion = null;
}
```

```
record GeographicLocation {
    union { null, string } description = null;
    union { null, float } elevation = null;
    union { null, float } latitude = null;
    union { null, float } longitude = null;
```

```
record Observation {
    union { null, string } ageAtObservation = null;
    union { null, string } dateTimeObserved = null;
    union { null, string } id = null;
    OntologyTerm observation;
    union { null, string } unit = null;
    OntologyTerm value;
}
```

```
record Evidence {
    union { null, string } description = null;
    OntologyTerm evidenceType;
```

```
record Dataset {
    array<string> accessions;
    union { null, string } description = null;
    union { null, string } guid = null;
    string id;
    array<string> memberIds;
    union { null, string } name = null;
}
```

```
record Disease {
    union { null, string } ageOfOnset = null;
    union { null, string } dateTimeDiagnosis = null;
    OntologyTerm disease;
    union { null, OntologyTerm } stageAtDiagnosis = null;
}
```

Global Alliance
for Genomics & Health

Biometadata

Assay-metadata

Individual
accessions
dateOfBirth
description
developmentalStage
diseases
ethnicity
geographicLocation
guid
id
info
interventions
name
observations
phenotypes
recordCreateTime
recordUpdateTime
sex
species
strain

BioSample
accessions
ageAtSampling
cellLine
cellType
description
geographicLocation
guid
id
individualIds
info
interventions
name
observations
organismPart
preservationMethod
recordCreateTime
recordUpdateTime
samplingDate
sex
species

Experiment
description
id
info
instrumentDataFile
instrumentModel
molecule
name
platformId
platformName
preparationId
processingFacility
recordCreateTime
recordUpdateTime
runTime
selection
strategy

```
record OntologyTerm {
    union { null, string } ontologySourceID = null;
    union { null, string } ontologySourceName = null;
    union { null, string } ontologySourceVersion = null;
```

```
record GeographicLocation {
    union { null, string } description = null;
    union { null, float } elevation = null;
    union { null, float } latitude = null;
    union { null, float } longitude = null;
```

```
record Observation {
    union { null, string } ageAtObservation = null;
    union { null, string } dateTimeObserved = null;
    union { null, string } id = null;
    OntologyTerm observation;
    union { null, string } unit = null;
    OntologyTerm value;
}
```

```
record Evidence {
    union { null, string } description = null;
    OntologyTerm evidenceType;
```

```
record Dataset {
    array<string> accessions;
    union { null, string } description = null;
    union { null, string } guid = null;
    string id;
    array<string> memberIds;
    union { null, string } name = null;
}
```

```
record Disease {
    union { null, string } ageOfOnset = null;
    union { null, string } dateTimeDiagnosis = null;
    OntologyTerm disease;
    union { null, OntologyTerm } stageAtDiagnosis = null;
}
```

Reduction of named attributes

Global Alliance for Genomics & Health

SIB

progenetix

arrayMap

University of Zurich[UZH]

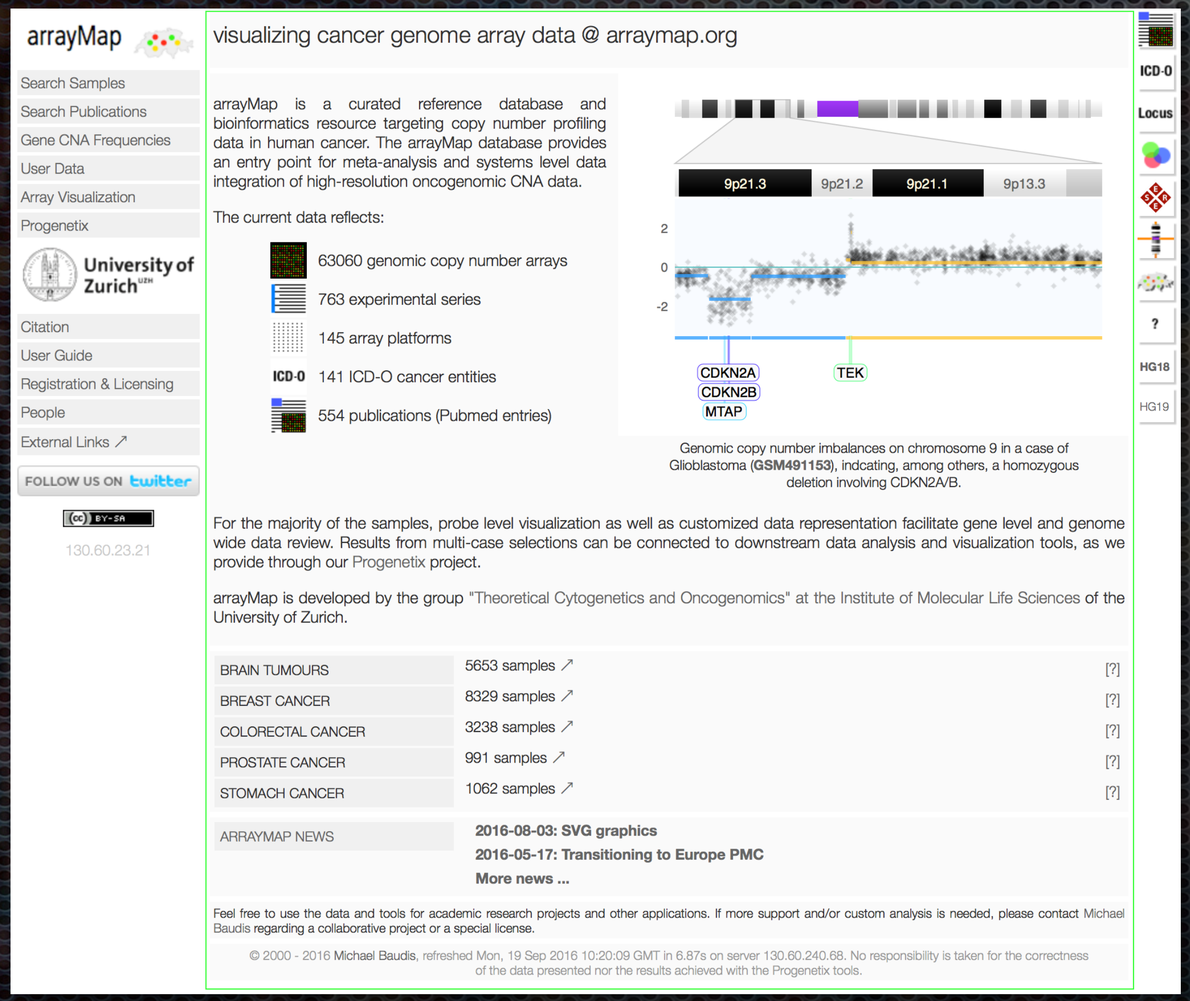
Reference Resources for Cancer Genome Profiling

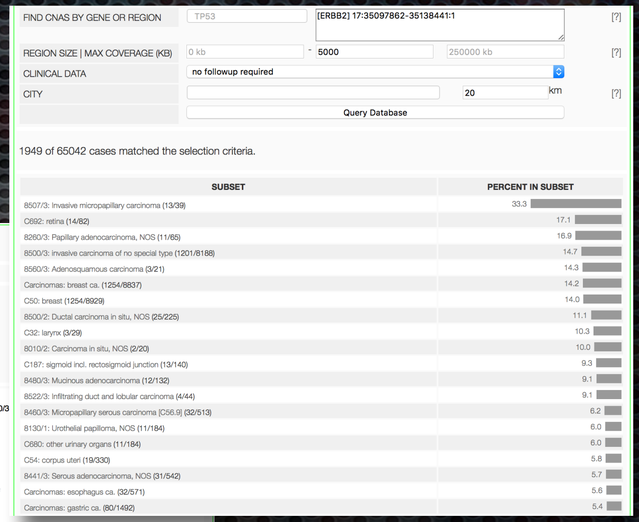
- curated reference resources for cancer genome profiling data and related information
- basis for own research activities, collaborative projects and external use
- structured information serves for implementing GA4GH concepts

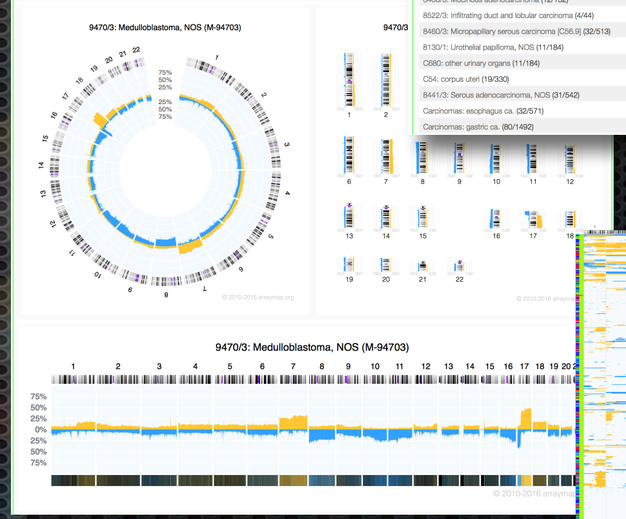
	progenetix	arrayMap
techniques	cCGH, aCGH, WES, WGS	aCGH (+?)
scope	sample (e.g. combination of several experiments)	experiment
content	**>31000** samples	**>60000** arrays
raw data presentation	no (link to sources if available)	yes (raw, log2, segmentation if available)
per sample re-analysis	no; supervised result (mostly as provided through publication)	yes (re-segmentation, thresholding, size filters ...)
final data	annotated/interpreted CN status for GP and cytogenetic regions	unsupervised CN status for GP and cytogenetic regions
main purposes	• Distribution of CNA target regions in most tumor types (>350 ICD-O) • Cancer classification	• Gene specific hits • Genome feature correlation (fragile sites ...)

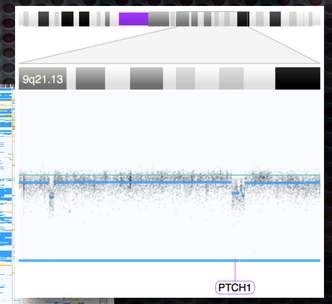
arrayMap

Resource for copy number variation data in cancer





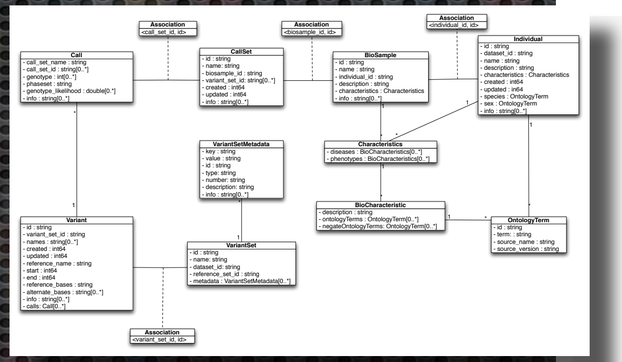




UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	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

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



Driving **Beacon** Development

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


```
  "_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"),
}
```


Global Alliance for Genomics & Health

EMBL-EBI

arrayMap

elixir

SIB Swiss Institute of Bioinformatics

- object model instead of named attributes
- referencing of ontologies instead of text descriptors

State of the schema

Biosample from arrayMap 2017-02-17

- fallback to key:value map for
- unassigned data; this should disappear over time


```
"_id" : ObjectId("589dfa5109d374e4f3655aee"),
"name" : "AM_BS_GSM322223",
"individual_id" : "PGIND_GSM322223",
"id" : "AM_BS_GSM322223",
"characteristics" : {
        "diseases" : [
                {
                        "ontologyTerms" : [
                                {
                                        "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")
```


sanity check, fallback, unmapped

Now that we got this settled … OntologyTerm generic in GA4GH schema

Global Alliance for Genomics & Health

- an OntologyTerm object is used in GA4GH to provide external classifications
- used throughout the schema, not only biomedical metadata

```
// An ontology term describing an attribute. (e.g. the phenotype attribute
// 'polydactyly' from HPO)
message OntologyTerm {
  // Ontology term identifier - the CURIE for an ontology term. It
  // differs from the standard GA4GH schema's :ref:`id <apidesign_object_ids>`
  // in that it is a CURIE pointing to an information resource outside of the
  // scope of the schema or its resource implementation.
  string term_id = 1;

  // Ontology term - the label of the ontology term the termId is pointing to.
  string term = 2;
}
```

- However: This “outsources” data logistics to **external ontology services**

➡mapping of original data to “standard” ontologies through the data provider

➡providing of information about used ontologies through the implementer of the resource

➡development of ontology based querying methods for APIs and web interfaces

- object model instead of named attributes
- referencing of ontologies instead of text descriptors

- need for **ontologies** & mappings
- these are no "real" open ontologies

- curating phenotypic data into ontologies

- fallback to key:value map for
- unassigned data; this should disappear over time


```
"_id" : ObjectId("589dfa5109d374e4f3655aee"),
"name" : "AM_BS_GSM322223",
"individual_id" : "PGIND_GSM322223",
"id" : "AM_BS_GSM322223",
"characteristics" : {
        "diseases" : [
                {
                        "ontologyTerms" : [
                                {
                                        "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")
```

Ontologies need an Einstein to sort them out

University of Zurich[UZH]

ARRAYID	NCIT:CODE	NCIT:TERM	ICD-MORPHOLOGY-CODE	ICD-TOPOGRAPHY-CODE
GSM393758	C2773	Acute Myeloid Leukemia Not Otherwise Specified	9861/3	C42
GSM302275	C2852	Adenocarcinoma	8140/3	C34
GSM918983	C3222	Medulloblastoma	9470/3	C716
GSM551398	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM412374	C3163	Chronic Lymphocytic Leukemia	9823/3	C42
GSM121826	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM714412	C2852	Adenocarcinoma	8140/3	C569
GSM1109923	C9306	Soft Tissue Sarcoma	8800/3	C499
GSM711848	C2852	Adenocarcinoma	8140/3	C25
GSM746794	C89476		8077/2	C53
GSM198528	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM271399	C7949		8500/2	C50
GSM533469	C9349	Plasmacytoma	9731/3	C42

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

example_dx	ICDMORPHOLOGY	ICDOM	ICDTOPOGRAPHY	ICDOT	NCIT:CODE
malignant melanoma [metastatic cell line MaMel19]	Malignant melanoma NOS	8720/3	skin	C44	C3224
malignant melanoma [vagina]	Malignant melanoma NOS	8720/3	vagina and labia	C510	C3224
malignant melanoma [uvea metastasized]	Malignant melanoma NOS	8720/3	retina	C692	C3224
meningioma	Meningioma NOS	9530/0	meninges cerebral spinal	C700	C3230
mesothelioma	Mesothelioma NOS	9050/3	lung and bronchus	C34	C3234
pleural mesothelioma	Mesothelioma NOS	9050/3	pleura	C384	C3234
mesothelioma	Mesothelioma NOS	9050/3	connective and soft tissue NOS	C499	C3234
multiple myeloma	Plasma cell myeloma	9732/3	hematopoietic and reticuloendothelial syste	C42	C3242
Mycosis fungoides	Mycosis fungoides	9700/3	skin	C44	C3246
Myelodysplastic syndrome	Myelodysplastic syndrome NOS	9989/3	hematopoietic and reticuloendothelial syste	C42	C3247
Acute myeloblastic leukemia with maturation [FAB M2]	Acute myeloblastic leukemia with maturation [FAB M2]	9874/3	hematopoietic and reticuloendothelial syste	C42	C3250
neuroblastoma	Neuroblastoma NOS	9500/3	peripheral nervs incl. autonomous	C47	C3270
Cerebral neuroblastoma [cerebral region midline frontal lobe]	Neuroblastoma NOS	9500/3	cerebrum	C710	C3270
neuroblastoma [adrenal gland cell line]	Neuroblastoma NOS	9500/3	adrenal gland	C76	C3270
Cutaneous neurofibroma	Neurofibroma NOS	9540/0	skin	C44	C3272
Plexiform neurofibroma	Neurofibroma NOS	9540/0	Nervous system NOS	C729	C3272
Oligodendroglioma [Supratentorial Frontal Lobe]	Oligodendroglioma NOS	9450/3	cerebrum	C710	C3288
oilgodendroglioma	Oligodendroglioma NOS	9450/3	Brain NOS	C719	C3288
oligodendroglioma	Oligodendroglioma NOS	9450/3	brain nos	c719	C3288
Paraganglioma	Paraganglioma NOS	8680/1	Nervous system NOS	C729	C3308
paraganglioma	paraganglioma NOS	8680/1	adrenal cortex	C740	C3308
Pheochromocytoma	Pheochromocytoma NOS	8700/0	adrenal cortex	C740	C3326
polycythemia vera	Polycythemia vera	9950/3	hematopoietic and reticuloendothelial syste	C42	C3336
pediatric rhabdomyosarcoma	Rhabdomyosarcoma NOS	8900/3	connective and soft tissue NOS	C499	C3359
Sezary syndrome	Sezary syndrome	9701/3	hematopoietic and reticuloendothelial syste	C42	C3366
sezary syndrome	sezary syndrome	9701/3	skin	C44	C3366
Synovial sarcoma	Synovial sarcoma NOS	9040/3	connective and soft tissue NOS	C499	C3400
essential thrombocythemia	Essential thrombocythemia	9962/3	hematopoietic and reticuloendothelial syste	C42	C3407
carcinosarcoma	Carcinosarcoma NOS	8980/3	connective and soft tissue NOS	C499	C34448
Carcinosarcoma [breast cell line HS578T]	Carcinosarcoma NOS	8980/3	breast	C50	C34448
acute monocytic leukemia	Monocytic leukemia NOS	9860/3	hematopoietic and reticuloendothelial syste	C42	C3171
leiomyoblastoma	Epithelioid leiomyoma	8891/0	kidney	C649	C3157
colon mucosa [low grade dysplastic tumor; myhmut]	atypical adenoma	8140/1	large intestine excl. rectum and rectosigmoi	C189	C7559

Global Alliance for Genomics & Health

arrayMap

State of the schema

Biosample from arrayMap

2017-03-20


```
{
        "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" : "SNMI:M-85003",
                                        "term_label" : "invasive carcinoma of no special type"
                                },
                                {
                                        "term_id" : "PGX:ICDOM:8500_3",
                                        "term_label" : "invasive carcinoma of no special type"
                                },
                                {
                                        "term_id" : "PGX:ICDOT:C50",
                                        "term_label" : "breast"
                                },
                                {
                                        "term_id" : "PGX:SEER:26000",
                                        "term_label" : "Breast"
                                }
                        ],
                        "negated_ontology_terms" : [ ],
                }
        ],
        "individual_age_at_collection" : "P47Y",
        "attributes" : {
                "tnm" : {
                        "values" : [
                                {
                                        "string_value" : "T1N0M0"
                                }
                        ]
                },
                "death" : {
                        "values" : [
                                {
                                        "string_value" : "0"
                                }
                        ]
                },
                "country" : {
                        "values" : [
                                {
                                        "string_value" : "Norway"
                                }
                        ]
                },
                "city" : {
                        "values" : [
                                {
                                        "string_value" : "Oslo"
                                }
                        ]
                },
                "geo_lat" : {
                        "values" : [
                                {
                                        "double_value" : 59.91
                                }
                        ]
                },
                "geo_long" : {
                        "values" : [
                                {
                                        "double_value" : 10.75
                                }
                        ]
                },
        },
        "external_identifiers" : [
                {
                        "database" : "Pubmed",
                        "identifier" : "20592421"
                },
                {
                        "database" : "GEO",
                        "identifier" : "GSM510730"
                },
                {
                        "database" : "GEO",
                        "identifier" : "GSE20394"
                }
        ],
        "created" : ISODate("2017-03-20T08:37:07.771Z"),
        "updated" : ISODate("2017-03-20T08:37:07.771Z"),
}
```

Global Alliance for Genomics & Health

arrayMap

- drop of diseases | phenotypes wrapper used to type characteristics (discussion with code integration team at UCSC)
- **TODO**: introduce alternative characteristic_type label?
- introduction of “PGX” prefix for “ontologized” local versions of ICD-O 3 etc.

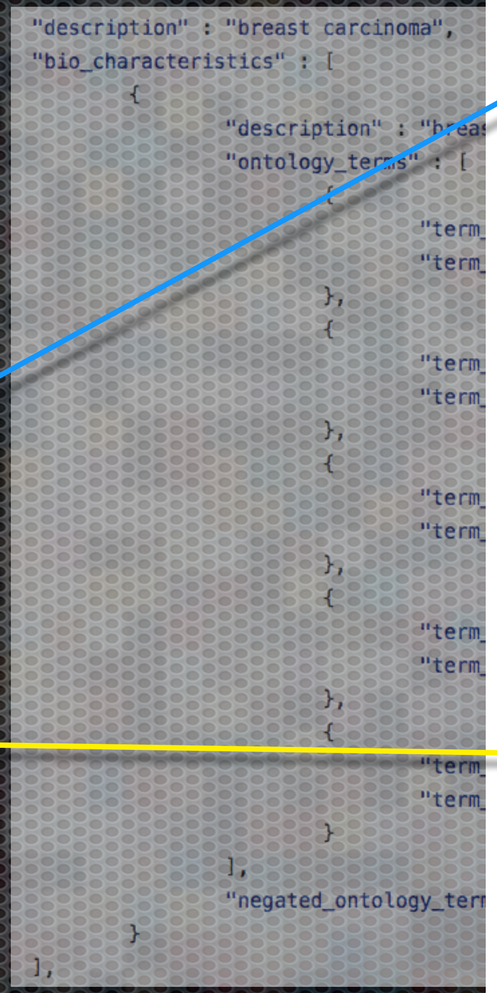

```
"description" : "breast carcinoma",
"bio_characteristics" : [
        {
                "description" : "breast carcinoma",
                "ontology_terms" : [
                        {
                                "term_id" : "NCIT:C4017",
                                "term_label" : "Ductal Breast Carcinoma"
                        },
                        {
                                "term_id" : "SNMI:M-85003",
                                "term_label" : "invasive carcinoma of no special type"
                        },
                        {
                                "term_id" : "PGX:ICDOM:8500_3",
                                "term_label" : "invasive carcinoma of no special type"
                        },
                        {
                                "term_id" : "PGX:ICDOT:C50",
                                "term_label" : "breast"
                        },
                        {
                                "term_id" : "PGX:SEER:26000",
                                "term_label" : "Breast"
                        }
                ],
                "negated_ontology_terms" : [ ],
        }
],
```

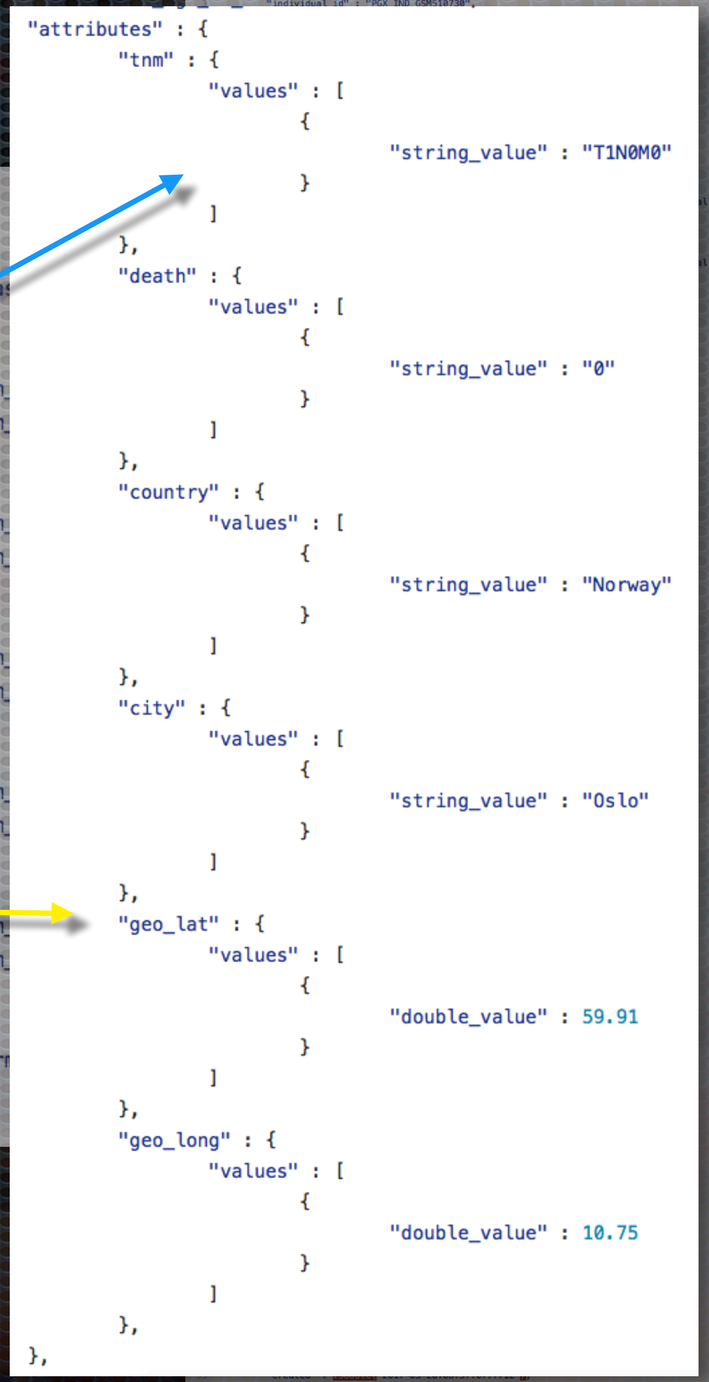

Biosample from arrayMap, 2017-03-20

Global Alliance for Genomics & Health

arrayMap

- drop of diseases | phenotypes wrapper used to type characteristics (discussion with code integration team at UCSC)
- **TODO**: introduce alternative characteristic_type label?
- introduction of “PGX” prefix for “ontologized” local versions of ICD-O 3 etc.
- typed constructors now being used for “arbitrary” attributes
- **TODO**: Finish design of a geographic location object



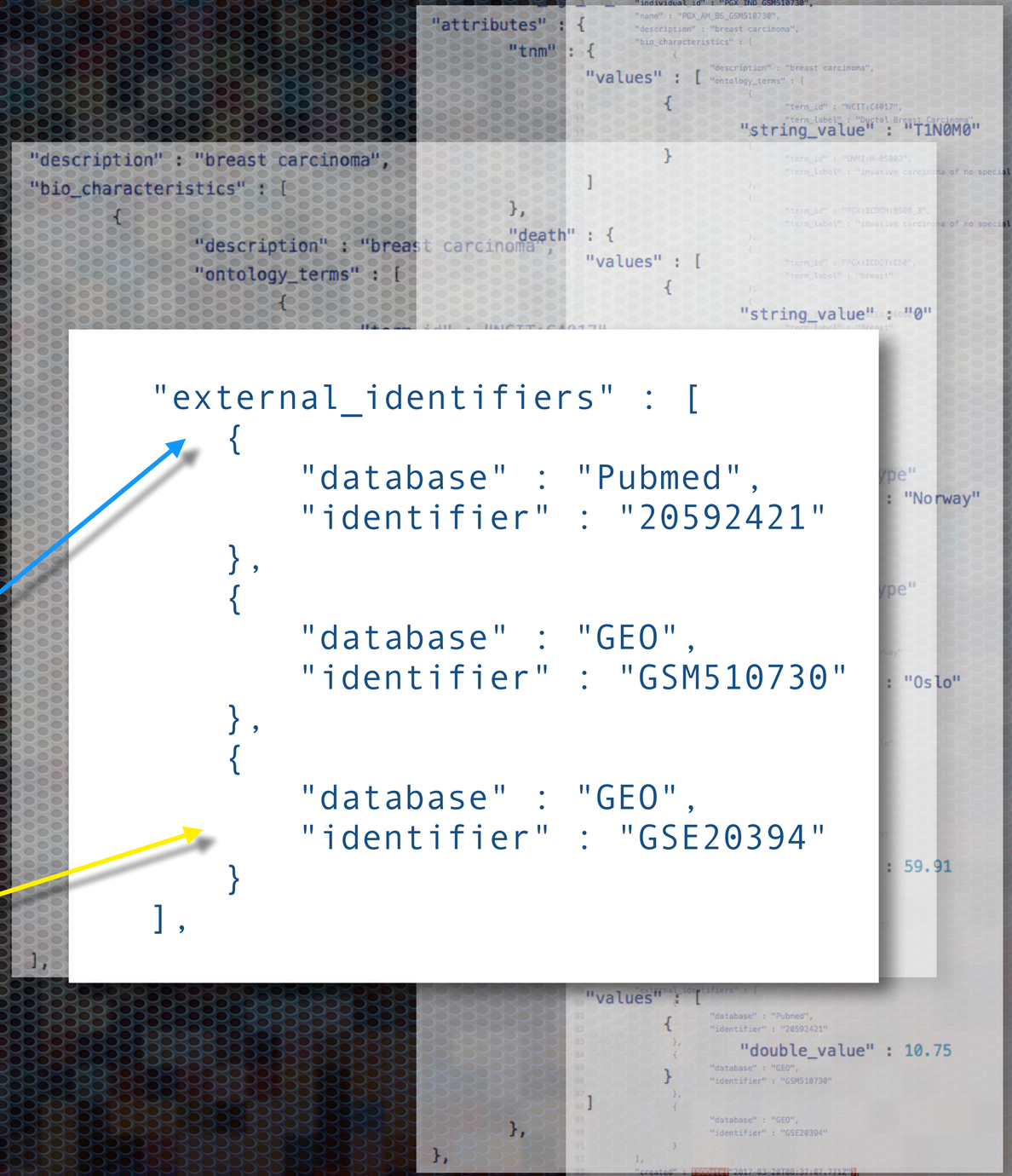


Biosample from arrayMap, 2017-03-20

Global Alliance for Genomics & Health

arrayMap

- drop of diseases | phenotypes wrapper used to type characteristics (discussion with code integration team at UCSC)
- **TODO**: introduce alternative characteristic_type label?
- introduction of “PGX” prefix for “ontologized” local versions of ICD-O 3 etc.
- typed constructors now being used for “arbitrary” attributes
- **TODO**: Finish design of a geographic location object
- External identifiers as pointers to other places where this record is represented
- **TODO**: Add label for type of representation (i.e., contained_in)



Biosample from arrayMap, 2017-03-20

Global Alliance for Genomics & Health

arrayMap

- sex described here, not at Biosample level
- referencing of ontologies instead of free or controlled text descriptors

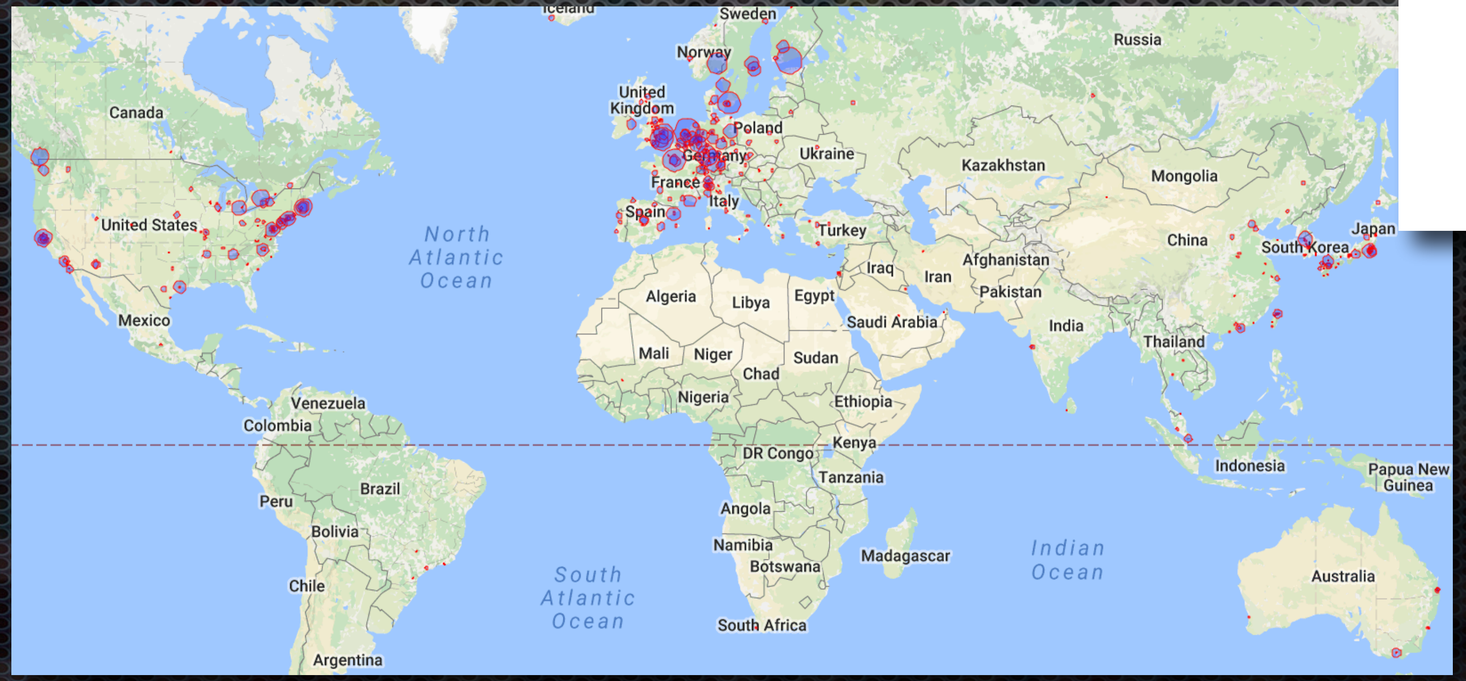
- diseases & phenotypes which are associated with the individual are recorded here
- diseases can also include those for which biosample records exist (independent of the specific annotation there)

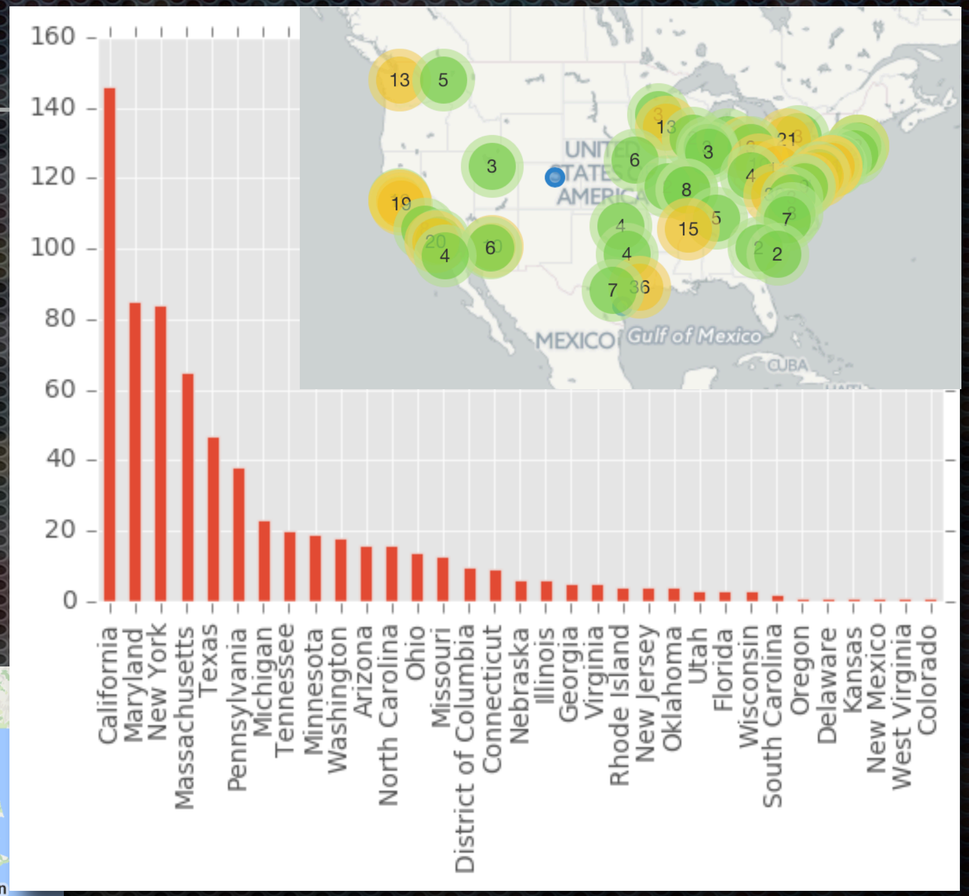
Individual from arrayMap, 2017-03-20


```
{
        "id" : "PGX_IND_GSM847445",
        "name" : "PGX_IND_GSM847445_edited",
        "species" : {
                "term_id" : "NCBITaxon:9606",
                "term_label" : "Homo sapiens"
        },
        "sex" : {
                "term_id" : "PATO:0020002",
                "term_label" : "female genotypic sex"
        },
        "description" : "individual with Li-Fraumeni syndrome",
        "bio_characteristics" : [
                {
                        "description" : "Li-Fraumeni syndrome carrier",
                        "ontology_terms" : [
                                {
                                        "term_id" : "DOID:3012",
                                        "term_label" : "Li-Fraumeni syndrome",
                                },
                        ],
                },
                {
                        "description" : "adrenocortical carcinoma",
                        "ontology_terms" : [
                                {
                                        "term_label" : "Adrenal Cortex Carcinoma",
                                        "term_id" : "NCIT:C9325",
                                },
                        ],
                }
        ],
        "attributes" : null,
        "created" : ISODate("2017-03-20T08:37:07.771Z"),
        "updated" : ISODate("2017-03-20T08:37:07.771Z"),
}
```

Geodata Proposal for GA4GH Schema

- geodata mapping as standard feature of biomedical objects (origin of biosamples or the individuals they were derived from; places of technical analysis) can provide insights and arguments for population matching, epistemological analyses and, importantly, project design (and funding decisions)





Mapping articles out of ~3000 whole cancer genome screening publications (cCGH, aCGH, WES, WGS); data visualization support by Elise Achison & Ross Purves, UZH Geography

University of Zurich[UZH]

progenetix

GeoLocation object proposal

- general consensus about some sort of **geographic attribution**, for a yet to be determined subset of GA4GH records/objects
- GeoJSON itself is not a good option for storing data, but rather for indicating objects on a maplocal **obfuscation** approaches will be needed for privacy protection
- destructive obfuscation can recode addresses to (random) points in higher level administrative boundaries
- the combination of *lat*, *long* with a location name seems a good compromise, with a "**precision level**" providing additional features (e.g. possibility to randomize point locations in a given boundary)

```
message GeoLocation {
  // a text representation, preferably using standard geographic identification
  // elements, of the corresponding latitude,longitude(,altitude)
  // This representation serves the purposes to
  //  - capture standard data entry parameters
  //  - provide a sanity check for latitude,longitude values
  // Example:
  //  - 34 Washington Blvd, Venice Beach, Los Angeles, CA, United States
  //  - Str Marasesti 5, 300077 Timisoara, Romania
  //  - Heidelberg, Deutschland
  string geo_label = 1;

  // an optional indication of the maximum precision to be derived from the
  // latitude,longitude values
  // Example:
  // Given a street address "Winterthurerstrasse 190, 8057 Zürich, Switzerland",
  // a privacy driven (destructive) obfuscation approach could recode this
  // to
  //  "latitude": 47.37, "longitude": 8.54
  //  while providing
  //  "geo_precision":"city", "geo_label": "Zürich, Switzerland"
  // ... indicating that the original location could correspond to any
  // latitude,longitude point value inside the administrative boundaries of
  // the city of Zürich, Switzerland
  string geo_precision = 2;

  // signed decimal degrees (North, relative to Equator)
  double latitude = 3;

  // signed decimal degrees (East, relative to IERS Reference Meridian)
  double longitude = 4;

  // optional, e.g. for environmental samples
  double altitude = 5;
}
```

Global Alliance
for Genomics & Health

Repository for example database files for method testing

progenetix / arraymap2ga4gh

Wiki

Branch: master ▾ | New pull request

mbaudis edited individual entry ···

data	changing the project structure
examples	edited individual entry
README.md	links

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

progenetix / arraymap2ga4gh

Code | Issues 1 | Pull requests 0

Branch: master ▾ arraymap2ga4gh / data / Sample

Bo Gao add a new demo set

..

biosamples.json

callsets.json

individuals.json

variants.json

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap data
- structural variations
- quantitative queries
- metadata
- current version uses GA4GH schema compatible, non-SQL database backend (MongoDB)

Heinz Stockinger, Séverine Duvaud & SIB Technology Group

Global Alliance for Genomics & Health

arrayMap

Beacon arrayMap

Beacon v0.4 implementation for arrayMap.

Reference name	9
Start	42049214
Length	1000
Assembly ID	GRCh36
Dataset Ids	(9440/3) 9440/3: Glioblastoma, NOS (2047)
Alternate bases	DEL (Deletion)
Confidence Interval (Start position)	500
Confidence Interval (End position)	500
Match type	Complete

SIB

Beacon Query | Beacon Info

arrayMap

Structural *Variants* from arrayMap

- name here could be e.g. an rsid
- calls are embedded in the variant set
- alternative option?
- “info” provides an intensity measurement
- could e.g. provide QC, copy number count …
- Housekeeping needed?


```
{
    "id" : "AM_V_3110636",
    "name" : null,
    "reference_name" : "11",
    "reference_bases" : ".",
    "start" : 75085926,
    "alternate_bases" : ".",
    "end" : 75744338,
    "variant_type" : "DEL",
    "svlen" : 658412,
    "info" : {},
    "calls" : [
        {
            "call_set_id" : "AM_CS_GSM902433",
            "info" : {
                "segvalue" : -0.3881
            },
            "genotype" : [ “.”, “.” ]
        },
        {
            "call_set_id" : "AM_CS_GSM902435",
            "info" : {
                "segvalue" : -0.4112
            },
            "genotype" : [ “.”, “.” ]
        }
    ],
    "updated" : ISODate("2017-02-10T17:15:02.380Z"),
    "created" : ISODate("2017-02-10T17:15:02.380Z"),
}
```

GA4GH schema: Beginning support for structural variants in master

- variant_type, svlen now providing essential support for annotation of CNV/CNA
- cipos, ciend (derived from VCF4.2 INFO.CIPOS, INFO.CIEND) enable "imprecise" data (e.g. array based)

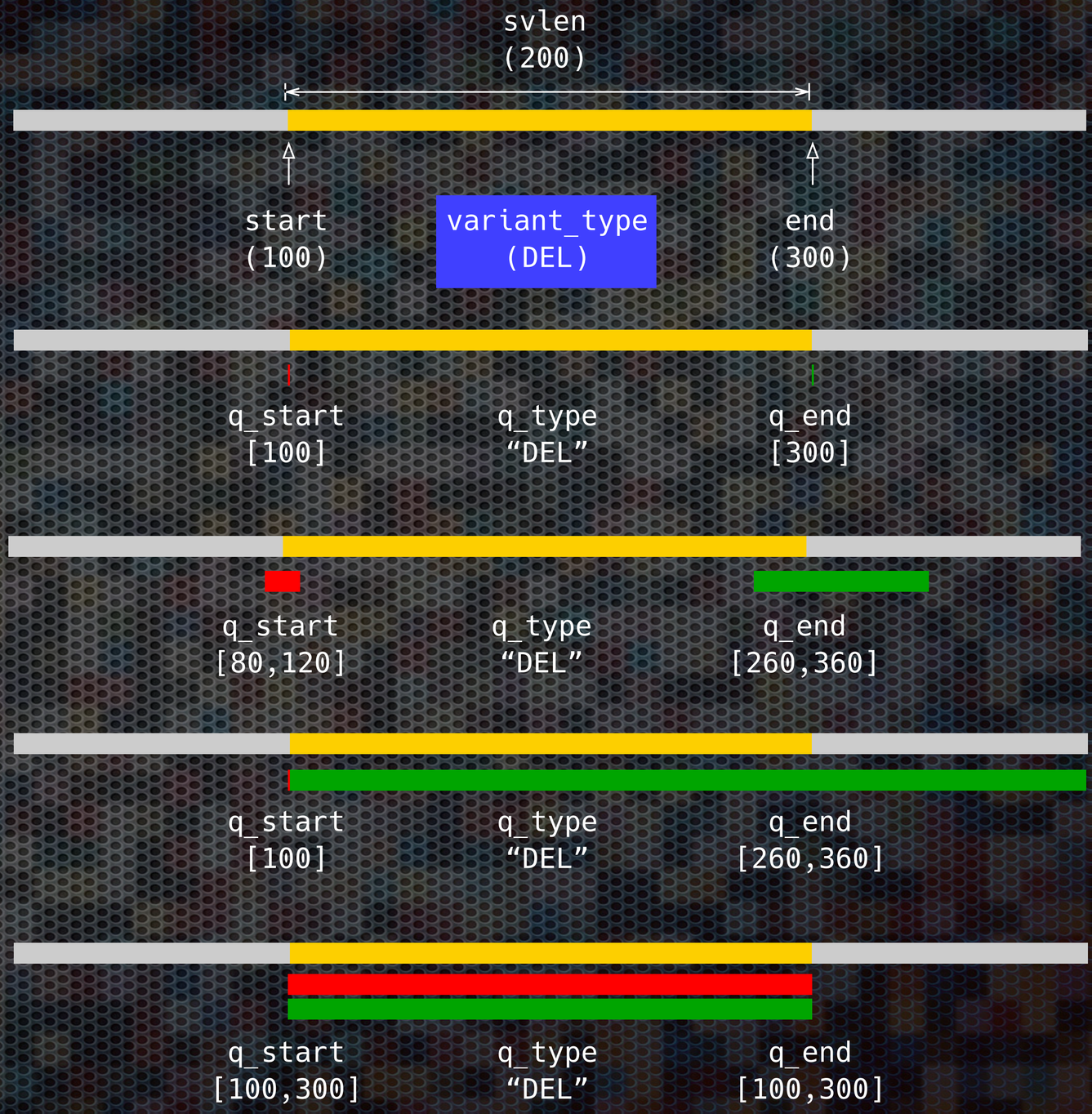
```
195
196     // The "variant_type" is used to denote e.g. structural variants.
197     // Examples:
198     //   DUP  : duplication of sequence following "start"; not necessarily in situ
199     //   DEL  : deletion of sequence following "start"
200     string variant_type = 17;
201
202     // Length of the - if labeled as such in variant_type - structural variation.
203     // Based on the use in VCFv4.2
204     int64 svlen = 18;
205
206     // In the case of structural variants, start and end of the variant may not
207     // be known with an exact base position. "cipos" provides an interval with
208     // high confidence for the start position. The interval is provided by 0 or
209     // 2 signed integers which are added to the start position.
210     // Based on the use in VCFv4.2
211     // Example:
212     //   [ -12000, 1000 ]
213     repeated sint32 cipos = 19;
214
215     // Similar to "cipos", but for the variant's end position (which is derived
216     // from start + svlen).
217     // Example:
218     //   [ -1000, 0 ]
219     repeated sint32 ciend = 20;
220
```

Changing to Ontology?!

GA4GH schemas v0.6.0a10 (March 2017)

Global Alliance for Genomics & Health

Beacon+ Proposal: Range Matching with Bracketed Start and End Positions



GA4GH Schema - Next Steps, and then Some More

- GA4GH schema development is clearly **implementation driven**:
  - ➡ demonstrate the needs with real world data
  - ➡ **engage** others in **discussions**, for fine tuning & optimizaton (or rejection)
  - ➡ develp a the concept into a Github pull request & work with the schema gods
  - ➡ … next
- Additions under way or with “known” need:
  - ➡ representation of meta-genomes | -proteomes | -…omes
  - ➡ geographic location data
  - ➡ object **relations** in substance and time flow
- Development of software tools to ingest & transform “standard” biomedical and technical metadata together with molecular data

Global Alliance
for Genomics & Health

If you can't explain it simply,
you don't understand it well enough.

Albert Einstein

UZH

MICHAEL BAUDIS
PAULA CARRIO CORDO
BO GAO
SAUMYA GUPTA

ELISE ACHISON
ROSS PURVES

EMBL-EBI

MELANIE COURTOT
HELEN PARKINSON

SIB

HEINZ STOCKINGER
SÉVERINE DUVAUD
VASSILIOS IOANNIDIS
DANIEL TEIXEIRA

University of Zurich[UZH]

SIB

EMBL-EBI

Global Alliance for Genomics & Health

ELIXIR

JORDI RAMBLA DE ARGILA
MACHA NIKOLSKI
S. DE LA TORRE PERNAS
SUSANNA REPO
SERENA SCOLLEN

GA4GH DWG + CWG

JACQUI BECKMANN
ANTHONY BROOKES
MARK DIEKHANS
MELISSA HAENDEL
SARAH HUNT
STEPHEN KEENAN
SUZY LEWIS
DAVID LLOYD
CHRIS MUNGALL
MICHAEL MILLER
ELEANOR STANLEY
DAVID STEINBERG

Random Links

- *Github GA4GH*
- *arrayMap implementation example data*
- *Progenetix publication list*
- *Geodata discussion page*
- *arraymap.org*
- *Presentations, including this one*