Global Alliance
for Genomics & Health

Developing Beacons for Data Discovery

Advancing Beacons through *data-driven* implementations

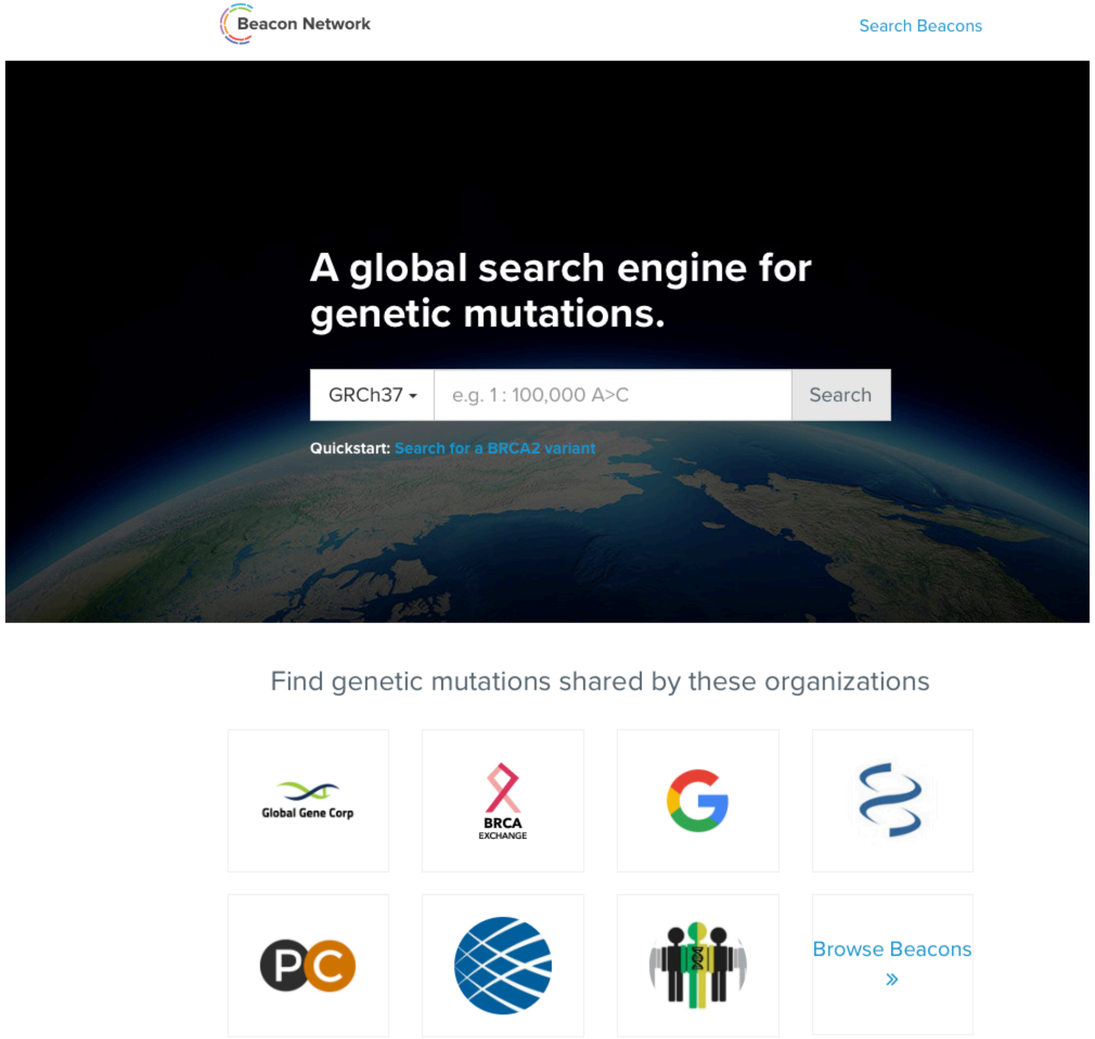
Michael Baudis - #GA4GH2017

University of Zurich[UZH]

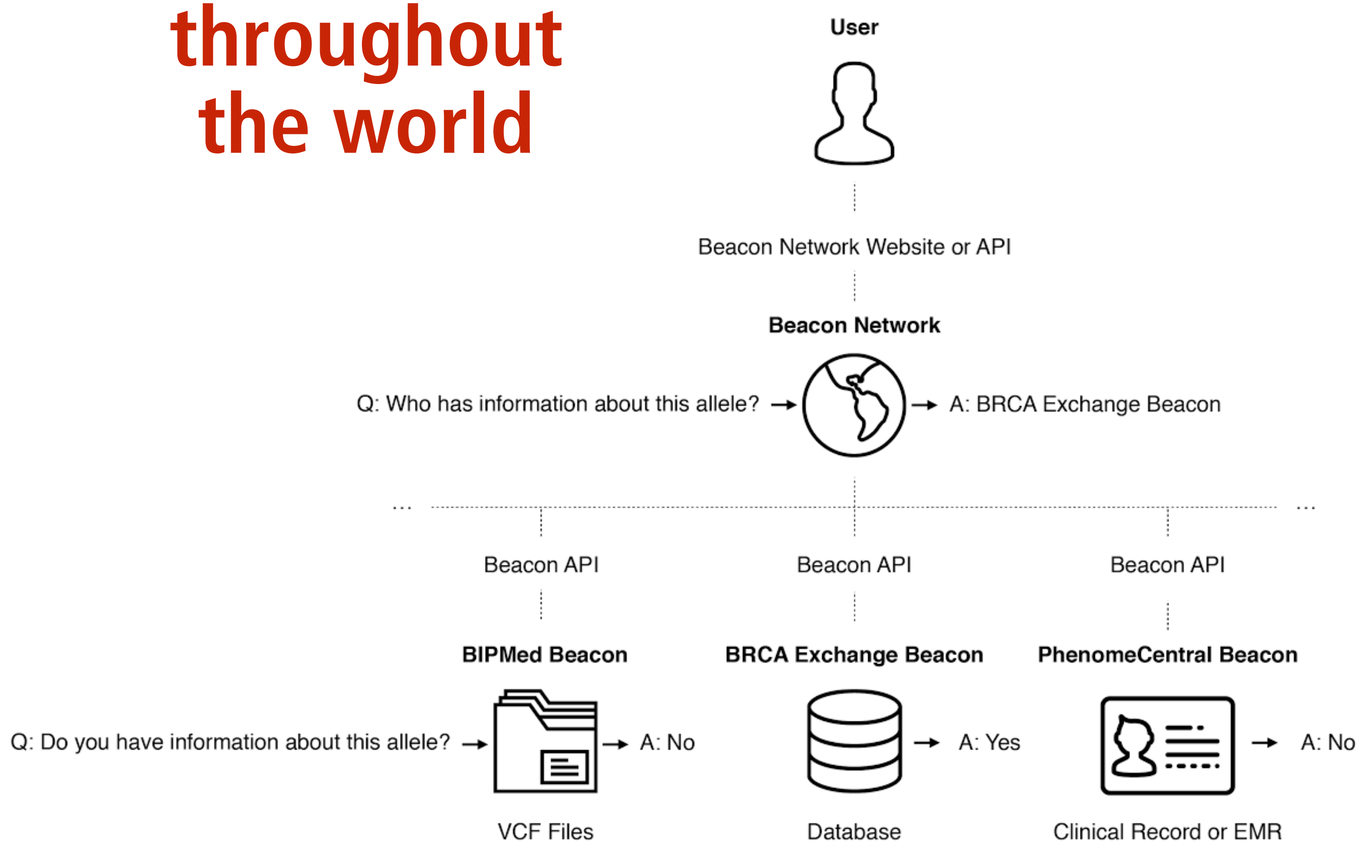
SIB

Beacon Project

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



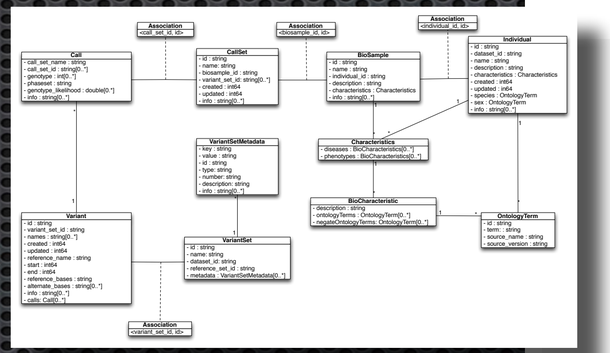
> 50 Beacons throughout the world



Global Alliance for Genomics & Health

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

Global Alliance for Genomics & Health

EMBL-EBI

arrayMap

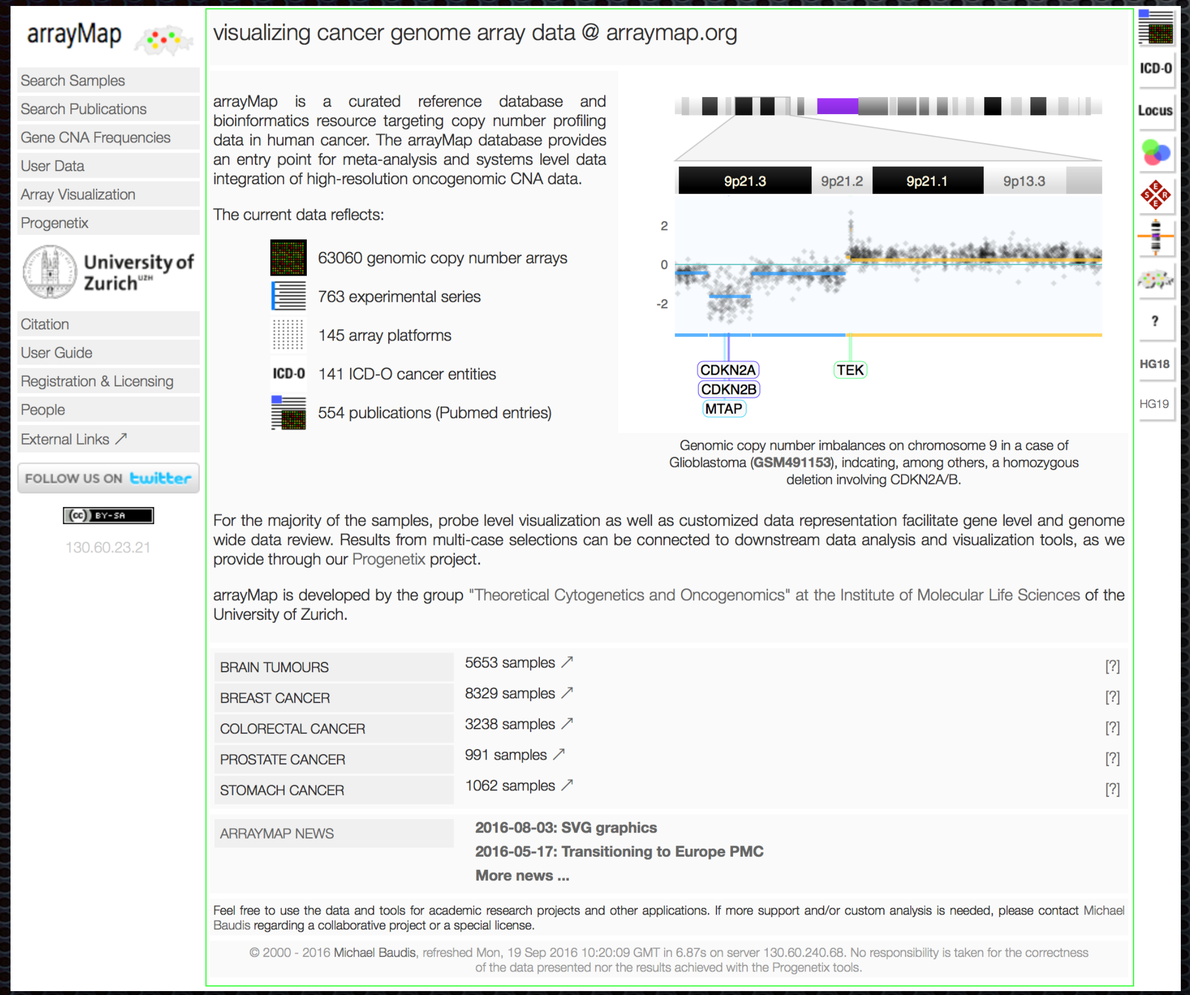
elixir

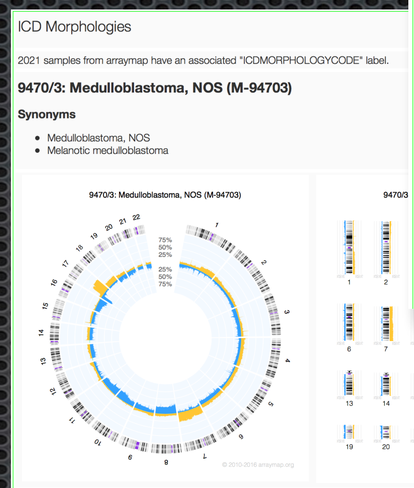
SIB Swiss Institute of Bioinformatics

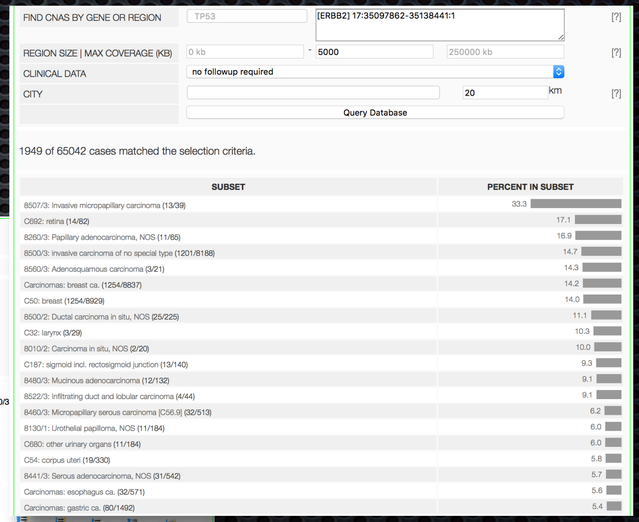

```
"_id" : ObjectId("58297ca32ca4591e5a0df054"),
"id" : "AM_V_1778741",
"variant_set_id" : "AM_VS_HG18",
"reference_name" : "10"
"start" : 579049,
"end" : 17236099,
"alternate_bases" : "DUP",
"reference_bases" : ".",
"info" : {
  "svlen":16657050,
  "cipos":[
      -1000,
      1000
  ],
  "ciend":[
      -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"),
}
```

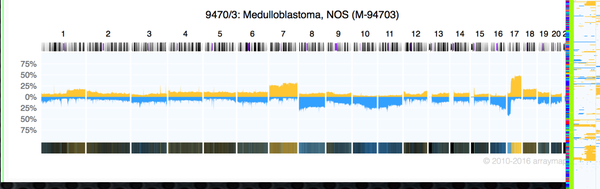
arrayMap

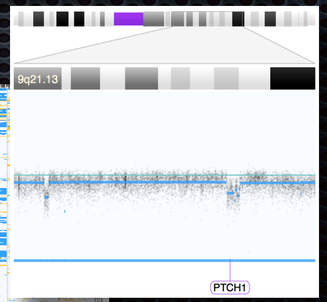
Resource for copy number variation data in cancer









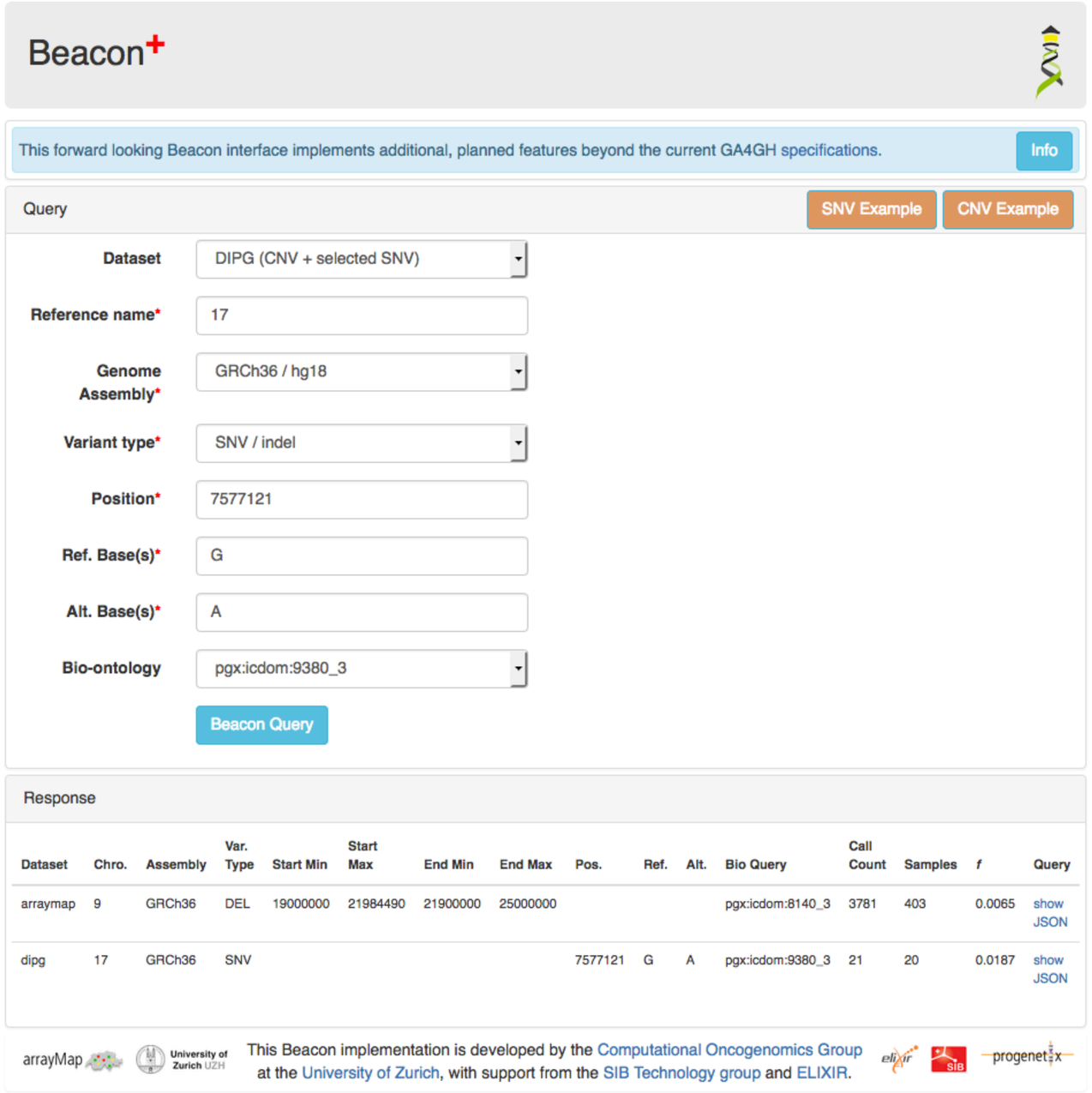


UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668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

arrayMap

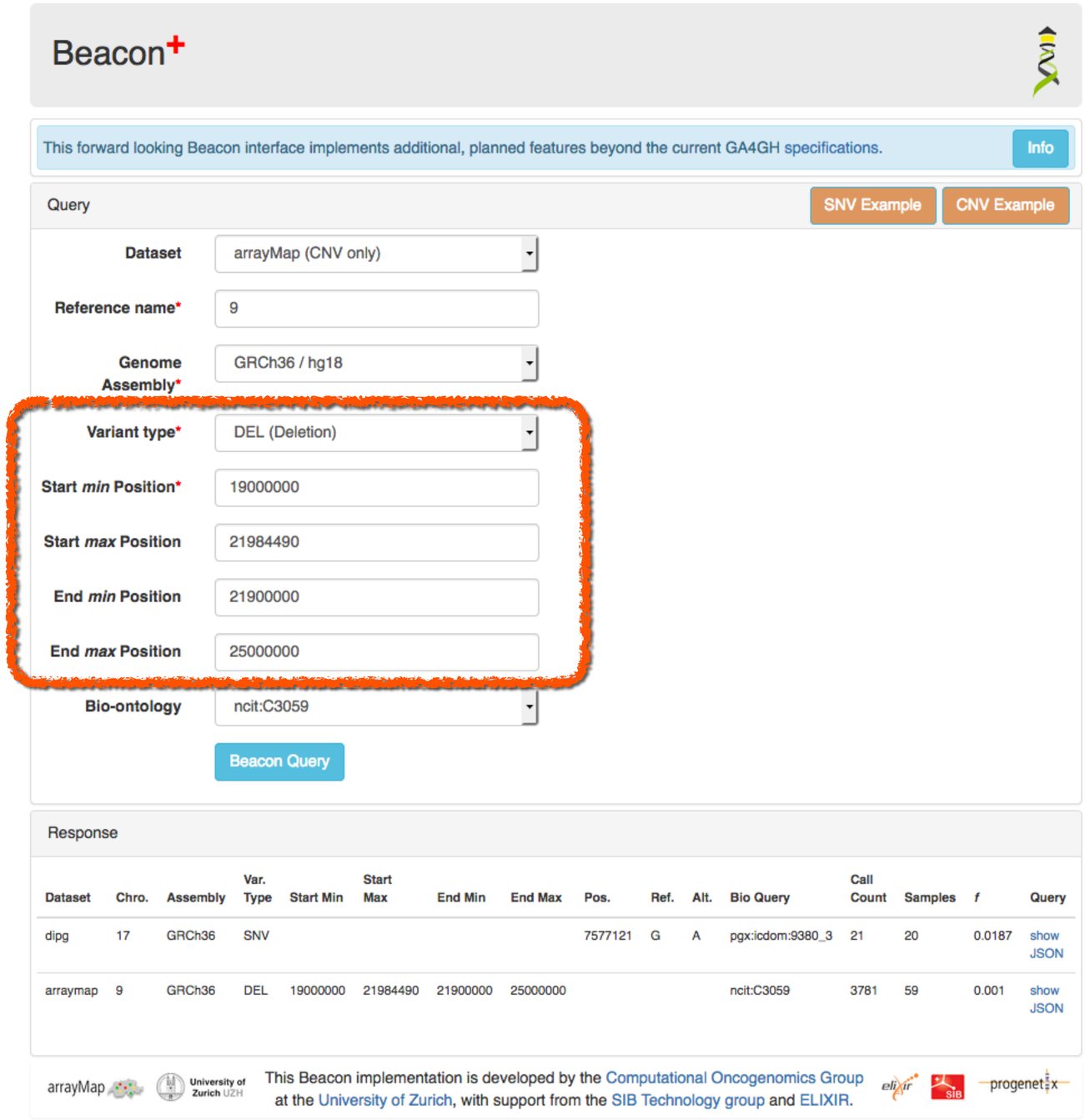
Beacon+ Concept

- 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

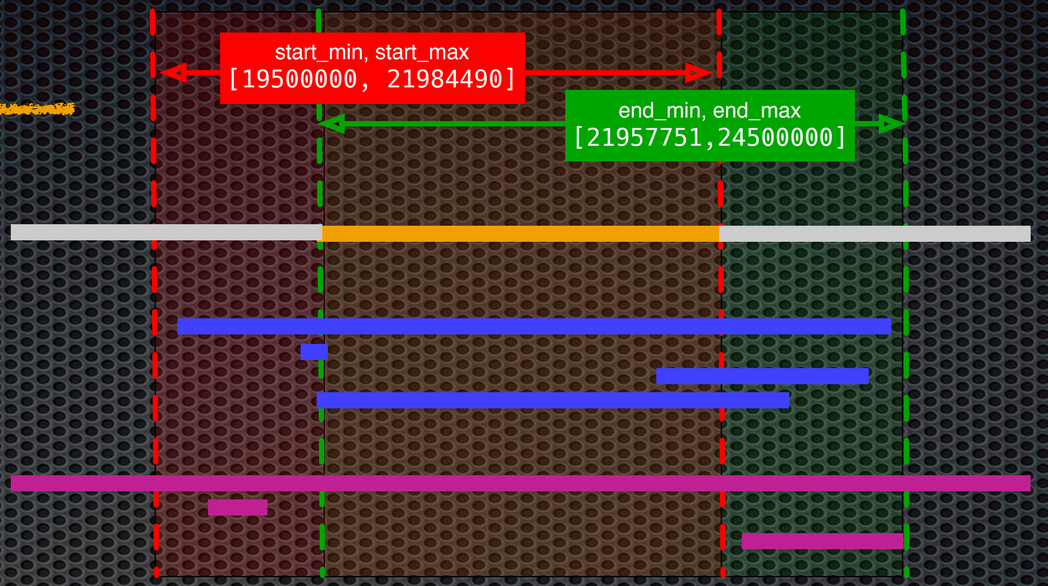


Beacon+ Concept

- 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



```
        { "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&va
riants.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”

Region of Interest, e.g. CDR of Gene (here: CDKN2A)

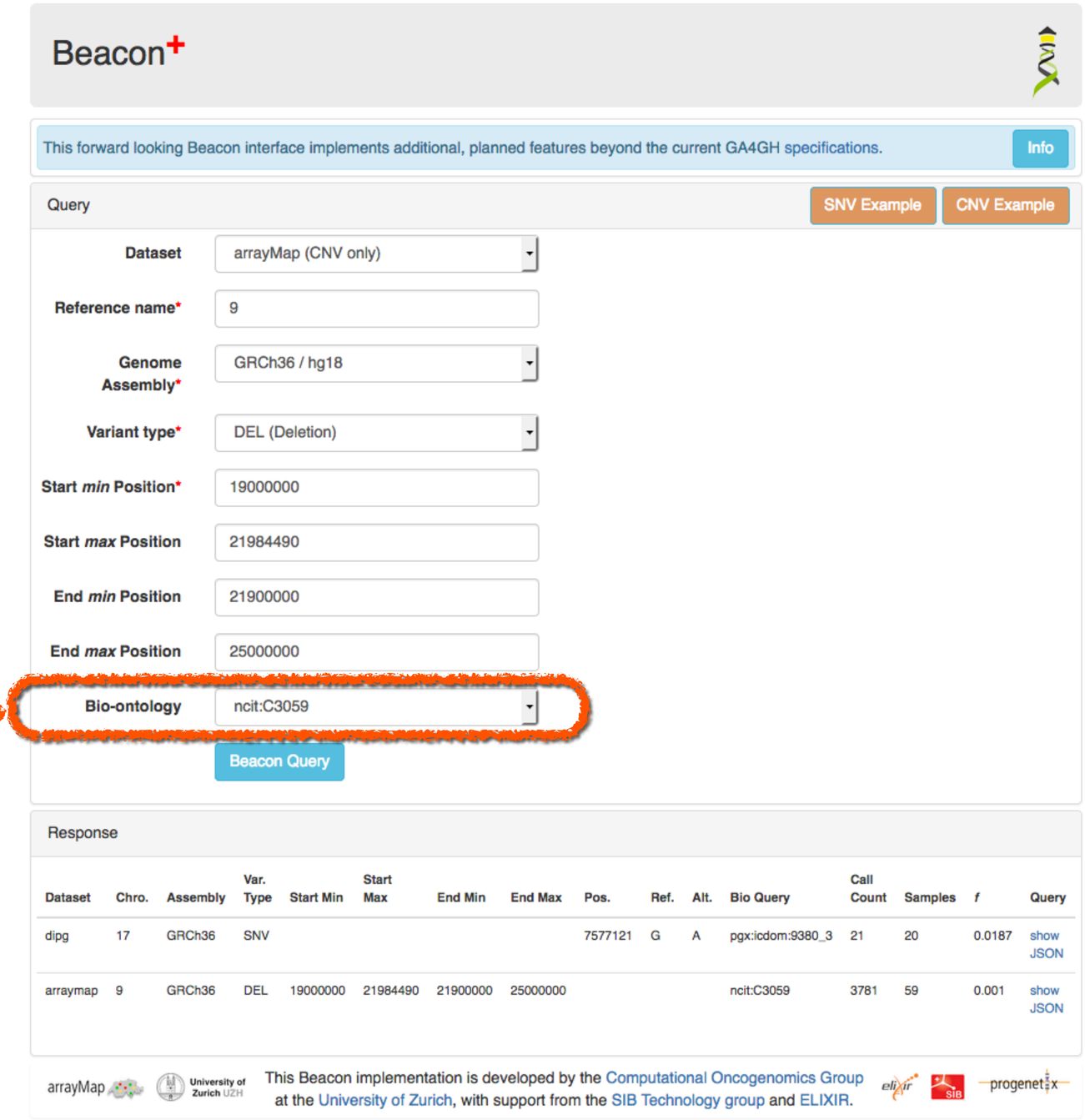
Example “focal” matches (overlap w/ size limit)

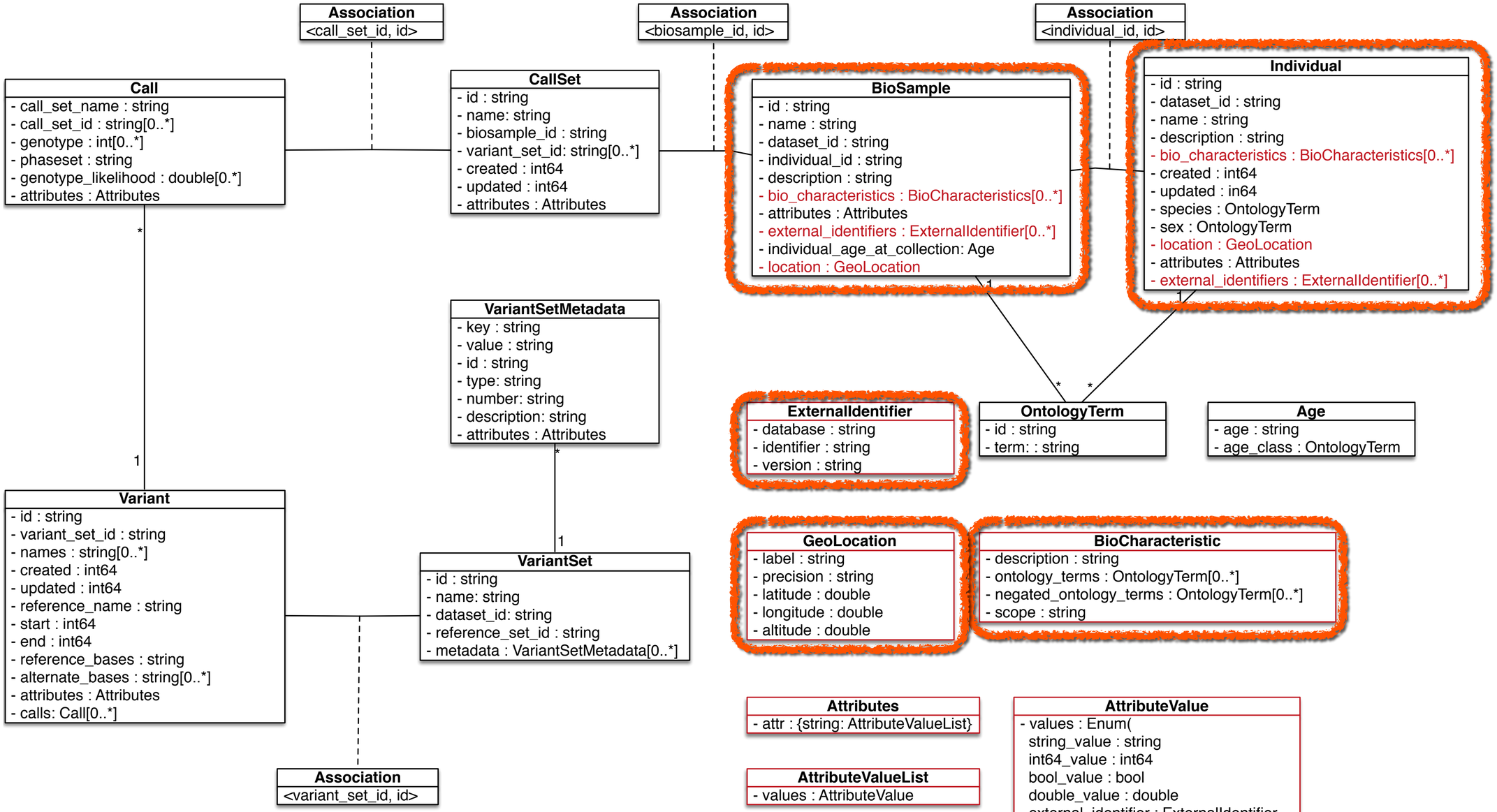
Mismatches
- too large
- end outside
- start outside

- 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>, <DEL>), as are specified in VCF && GA4GH variant schema

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



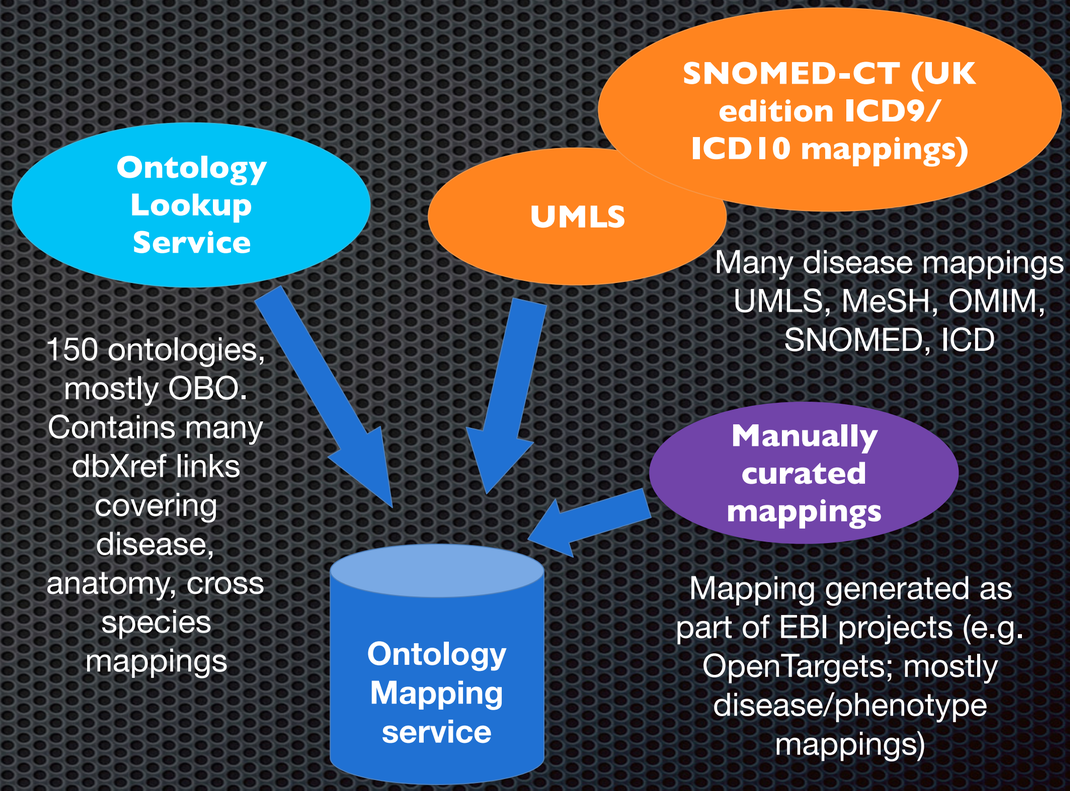


GA4GH DWG Metadata Task Team
Wrangling the Schema

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

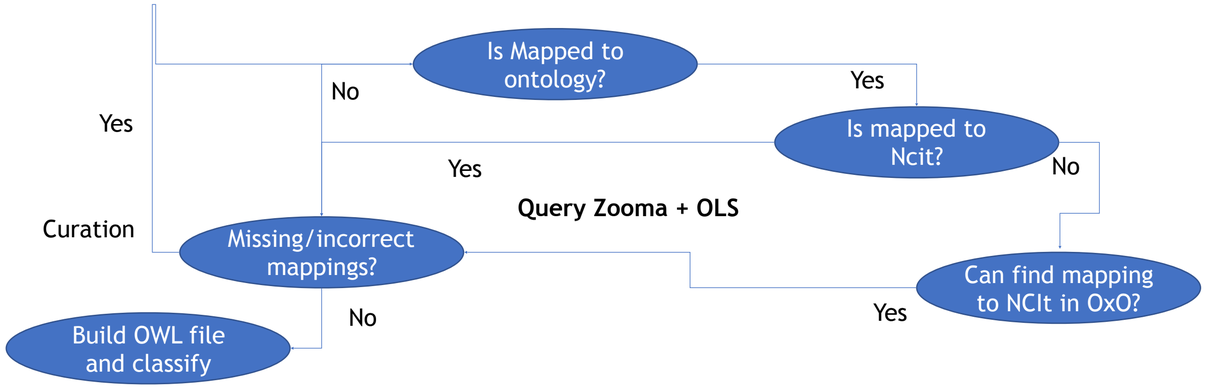


Ontology Mapping Service

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

ICDM	ICDMORPHOLOGY	NCItcode	NCItlabel	ICDT	ICDTOPOGRAPHY	NCItcode	NCItlabel	NCIt_mapp	NCIt_mapped_ICDM_T_label
8021/3	Carcinoma anaplastic type			C739	thyroid gland	C12400	thyroid gland	C3878	Thyroid Gland Undifferentiated (Anaplastic) Carcinoma
9451/3	Oligodendroglioma anaplastic	C4326	anaplastic oligodendroglioma	C719	Brain	C12439	brain	C4326	Anaplastic Oligodendroglioma
9051/3	Desmoplastic mesothelioma	C6747		C499	connective and soft tissue			C6747	Desmoplastic Mesothelioma
9732/3	Plasma cell myeloma	C3242	multiple myeloma	C42	hematopoietic and reticuloendothelial systems			C3242	Plasma Cell Myeloma
8070/3	Squamous cell carcinoma	C2926	non-small cell lung carcinoma	C140	pharynx			C102872	Pharyngeal Squamous Cell Carcinoma
8380/3	Endometrioid adenocarcinoma	C3769	endometrioid carcinoma	C54	corpus uteri	C12316		C6287	Endometrial Endometrioid Adenocarcinoma
8070/3	Squamous cell carcinoma	C2926	non-small cell lung carcinoma	C44	skin	C12470	zone of skin	C4819	Skin Squamous Cell Carcinoma
8430/3	Mucoepidermoid carcinoma	C45544	pulmonary mucoepidermoid carcinoma	C089	salivary gland	C12426	saliva-secreting gland	C5953	Minor Salivary Gland Mucoepidermoid Carcinoma
9680/3	Diffuse large B-cell lymphoma	C8851	diffuse large B-cell lymphoma	C42	hematopoietic and reticuloendothelial systems			C8851	Diffuse Large B-Cell Lymphoma
8800/3	Sarcoma	C9118	sarcoma	C559	uterus nos			C9306	Soft Tissue Sarcoma
8441/3	Serous adenocarcinoma	C7550	ovarian serous adenocarcinoma	C570	fallopian tube	C12403	fallopian tube	C40101	Serous Adenocarcinoma
9689/3	splenic marginal zone lymphoma nos			C422	spleen	C12432	spleen	C4663	Splenic Marginal Zone Lymphoma
8077/2	Squamous intraepithelial neoplasia grade III			C53	cervix uteri	C12311		C89476	Grade III Vaginal Intraepithelial Neoplasia
8140/0	Adenoma	C4196	adenoma	C189	large intestine excl. rectum and rectosigmoid junction			C4349	Colon Adenocarcinoma
8272/3	Pituitary carcinoma	C4536	Pituitary carcinoma	C751	pituitary gland	C12399	pituitary gland	C4536	Pituitary Gland Carcinoma
8500/2	Ductal carcinoma in situ	C3641	ductal carcinoma in situ	C50	breast	C12971	breast	C2924	Ductal Breast Carcinoma In Situ
8200/3	Adenoid cystic carcinoma	C2970	adenoid cystic carcinoma	C32	larynx	C12420	larynx	C2970	Adenoid Cystic Carcinoma
9370/3	Chordoma	C2947	Chordoma	C419	bone	C13076	bone tissue	C2947	Chordoma
9717/3	Enteropathy type T-cell lymphoma			C17	small intestine	C12386	small intestine	C4737	Enteropathy-Associated T-Cell Lymphoma
9698/3	Follicular lymphoma grade 3			C42	hematopoietic and reticuloendothelial systems			C3460	Grade 3 Follicular Lymphoma
9863/3	Chronic myeloid leukemia	C3177	chronic myelogenous leukemia	C42	hematopoietic and reticuloendothelial systems			C3174	Chronic Myelogenous Leukemia BCR-ABL1 Positive
8852/3	Liposarcoma myxoid	C3735	myxoid liposarcoma	C499	connective and soft tissue			C27781	Myxoid Liposarcoma
9080/3	Teratoma malignant			C809	unknown	C35882	Hereditary elliptocytosis	C3403	Tetratoma
8530/3	Inflammatory carcinoma	C4872	breast carcinoma	C50	breast	C12971	breast	C4001	Inflammatory Breast Carcinoma
8140/3	Adenocarcinoma	C27745	lung adenocarcinoma	C809	unknown	C35882	Hereditary elliptocytosis	C2852	Adenocarcinoma
						C12762	oropharynx	C2970	Adenoid Cystic Carcinoma
						C12415	kidney	C3158	Leiomyosarcoma
						C12499	internal ear	C2970	Adenoid Cystic Carcinoma
						C12683	bronchus	C2923	Bronchioloalveolar Carcinoma
						C12343	retina	C3224	Melanoma
					ial systems			C8459	Hepatosplenic T-Cell Lymphoma
						C12393	pancreas	C8294	Pancreatic Adenocarcinoma
					ial systems			C3996	Monoclonal Gammopathy of Undetermined Significance
								C4817	Ewing Sarcoma
								C3288	Oligodendroglioma
						C12422	tongue	C4648	Tongue Squamous Cell Carcinoma
					ial systems			C2862	Primary Myelofibrosis
								C4833	Oral Cavity Squamous Cell Carcinoma
						C12390	rectum	C9383	Rectal Adenocarcinoma
								C3158	Leiomyosarcoma
					ial systems			C3898	Extranodal Marginal Zone Lymphoma of Mucosa-Associate
						C12404	female gonad	C4512	Ovarian Mucinous Cystadenoma

- 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%

```
        { "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&va
riants.variant_type=DEL&variants.start_max=19000000&variants.start_min=
                                                              biosamples
.bio_characteristics.ontology_terms.term_id=pgx:icdom:9440_3"
                                                              ent 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"
```

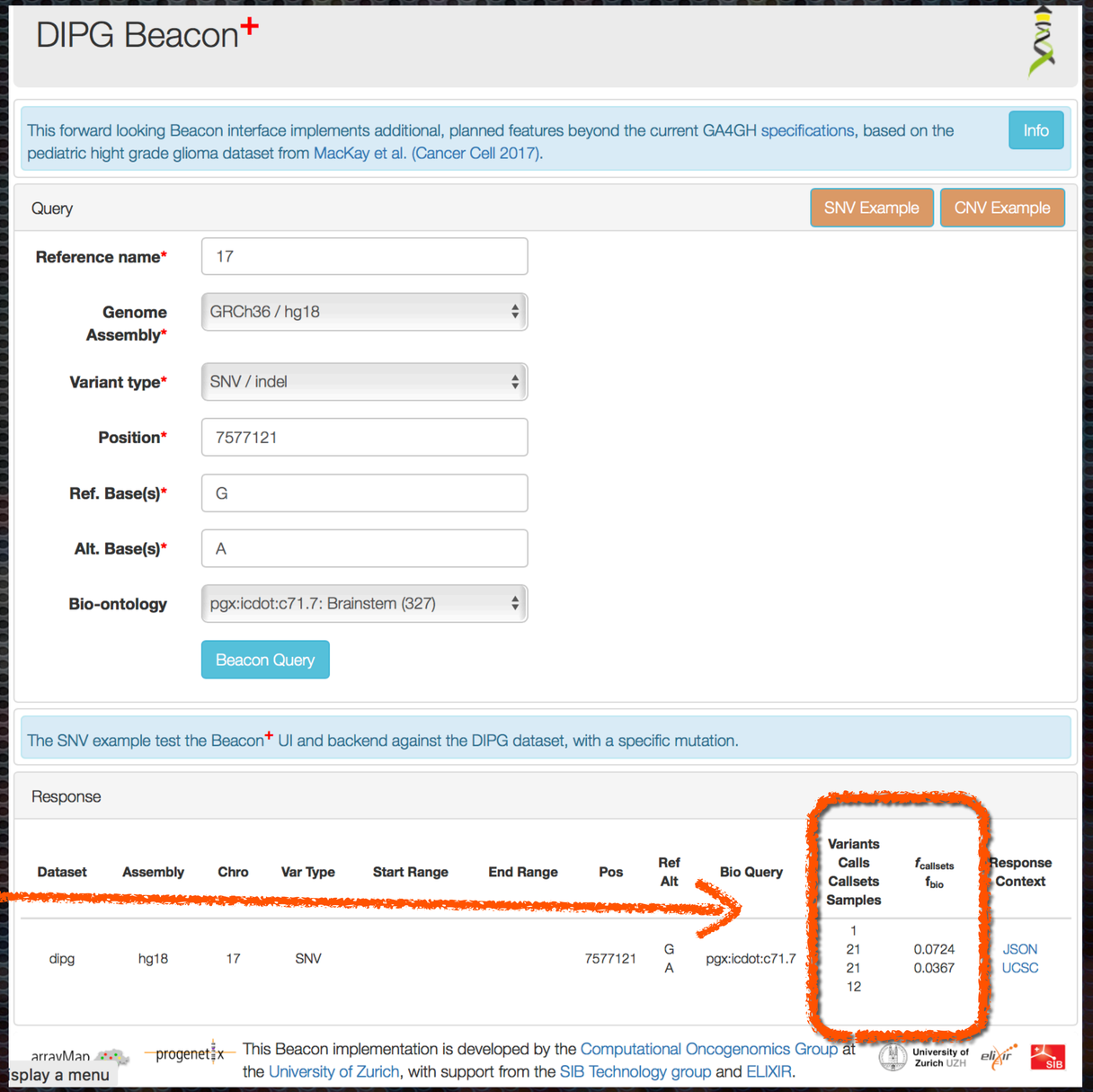

Beacon+ Concept

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)

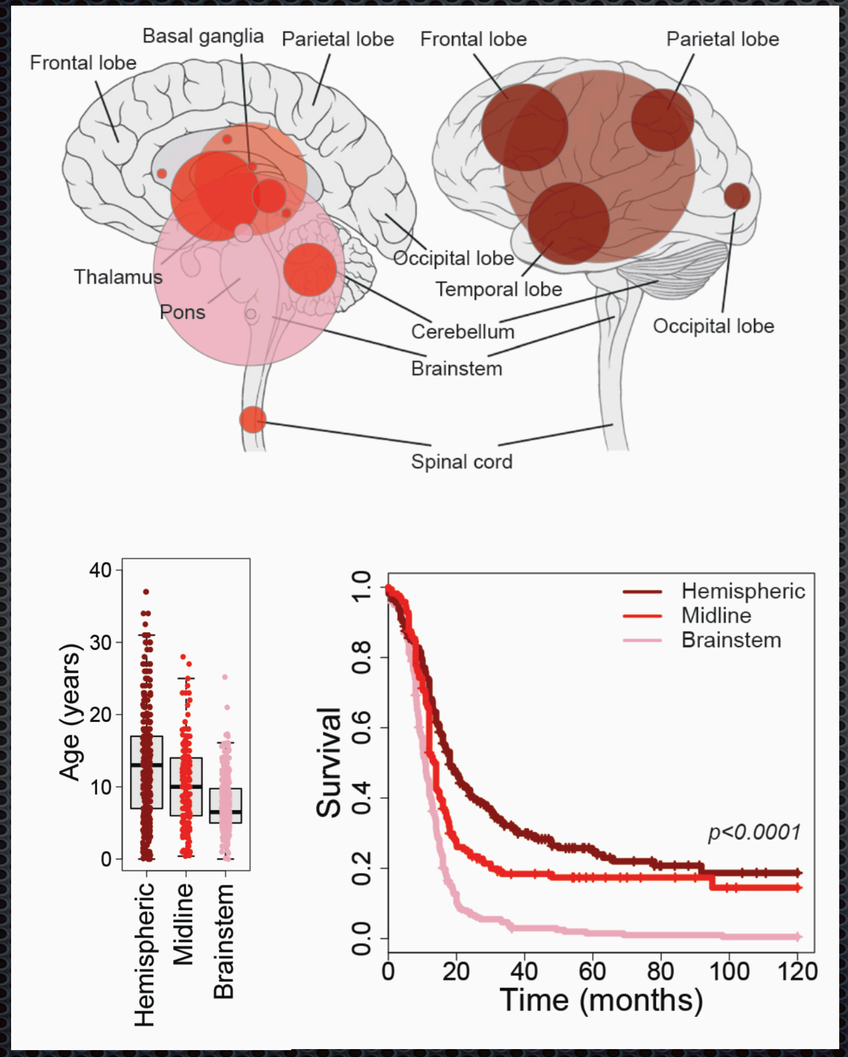
Implementing real-world datasets for federated access using GA4GH schema specifications: pHGG

- Study in >1000 rare aggressiive childhood brain tumors
- 157 of those not published previously
- copy number aberration data and selected gene panel represented in DIPG Beacon+
- interface with quantitative retiurns



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

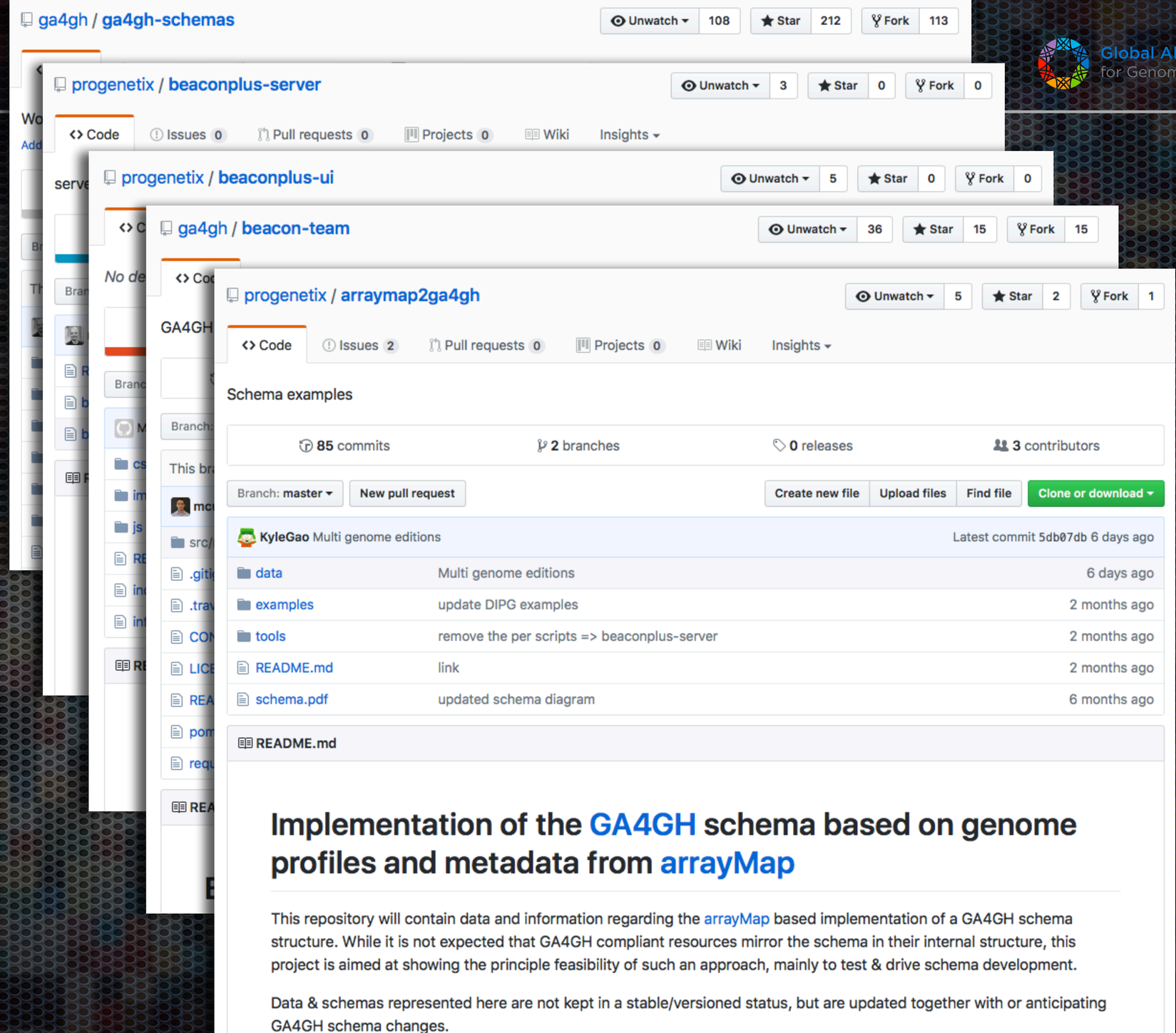
Mackay A, Jones C, Baudis M and many, many others: Integrated molecular meta-analysis of 1000 paediatric high grade and diffuse intrinsic pontine glioma (2017, Cancer Cell, in press)




```
"_id" : ObjectId("591eb73709d374421fc0b2a"),
"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" } ] },
}
```

Check it Out!

- managed, participation driven projects living on Github: ***ga4gh***
- *beacon.arraymap.org*
- *dipg.progenetix.org*
- test datasets & code available through our ***progenetix*** repositories
  - ➡ test
  - ➡ comment
  - ➡ suggest
  - ➡ propose
  - ➡ complain …



Global Alliance for Genomics & Health

University of Zurich[UZH] SIB elixir

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]

SIB

elixir

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