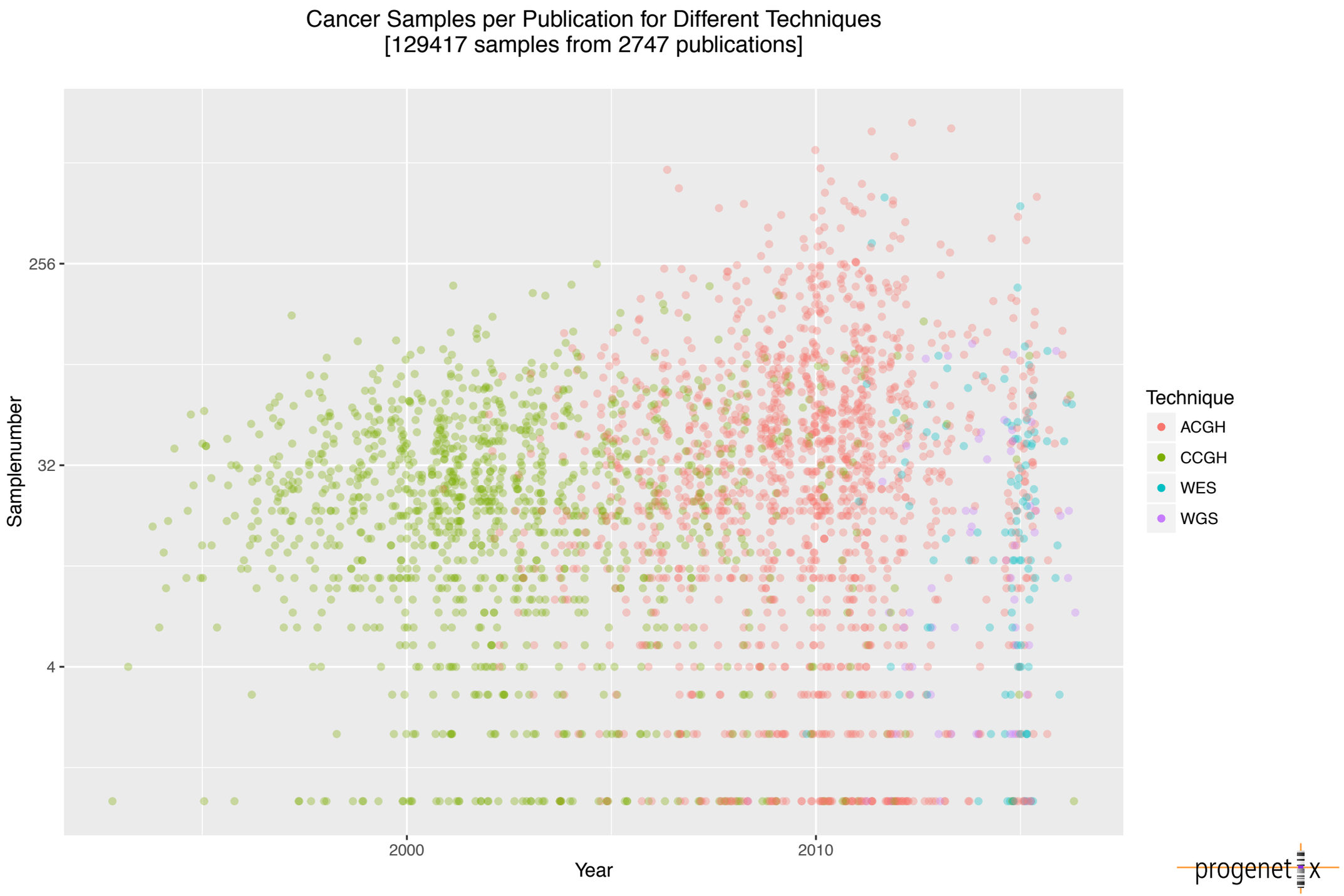
CURRENT OPPORTUNITIES AND FUTURE CHALLENGES

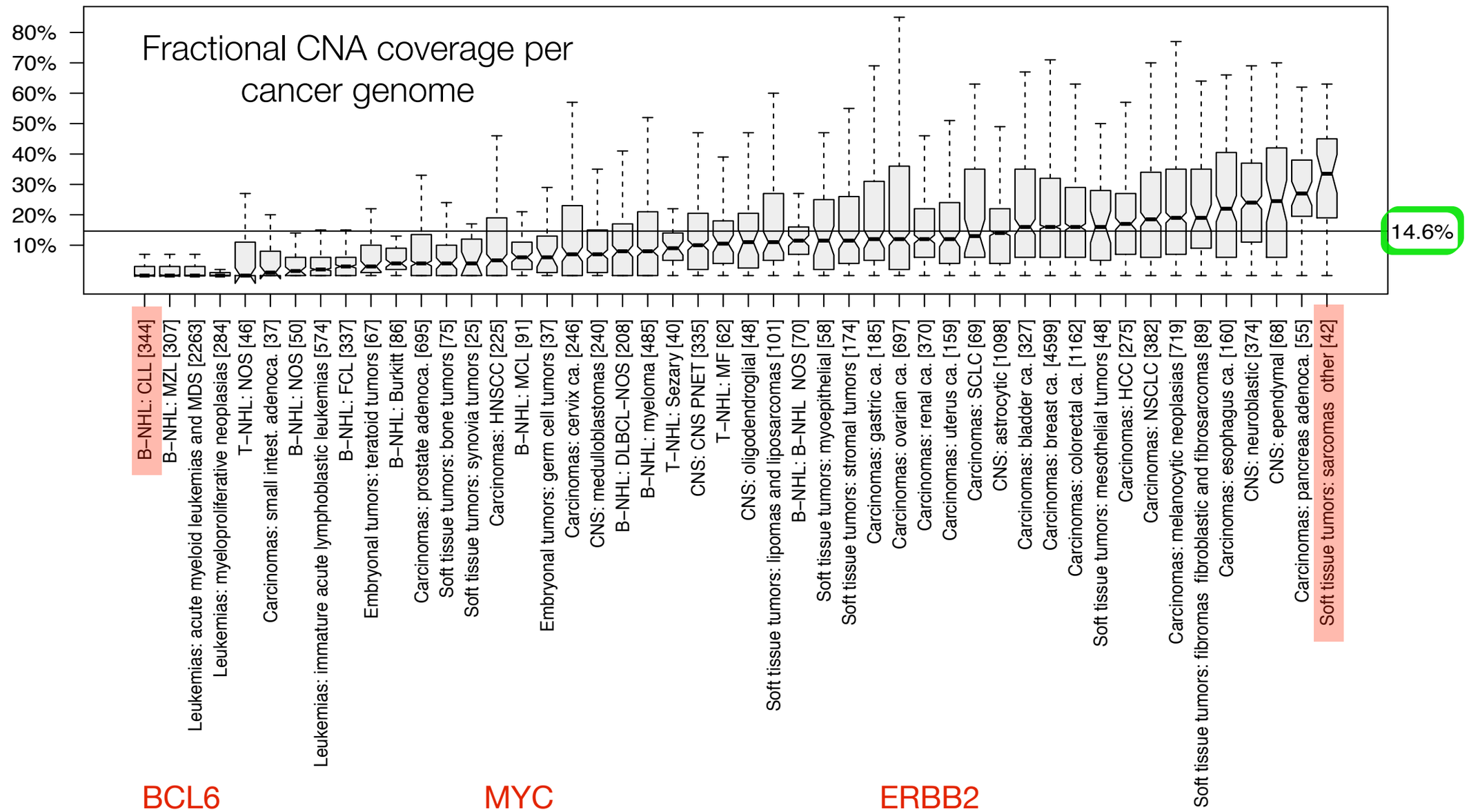
HARVESTING CANCER GENOME DATA

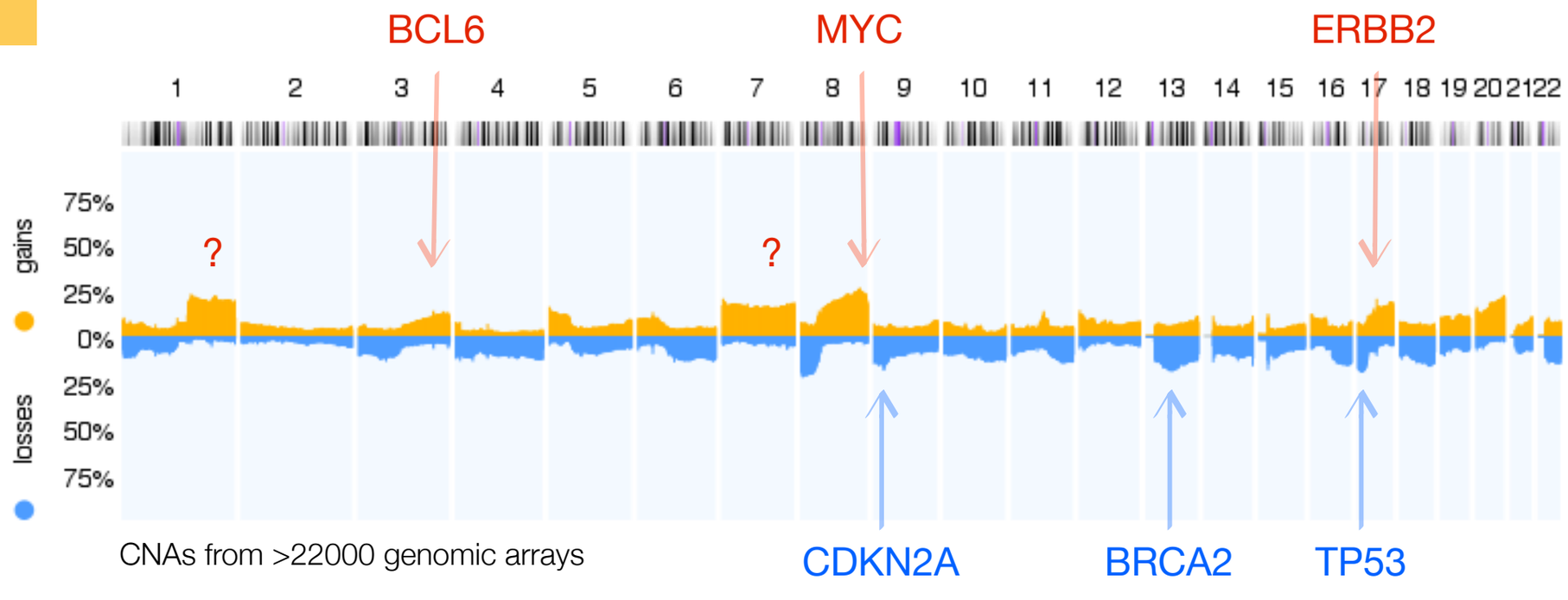
MOLECULAR CYTOGENETICS & SEQUENCING STUDIES FOR WHOLE GENOME PROFILING



progenetix

GENOMIC COPY NUMBER IMBALANCED PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER





CNAs from >22000 genomic arrays

Gain => Oncogenes

Deletion => Tumor Suppressor Genes

arrayMap

Bioinformatics 2001

progenetix

arrayMap

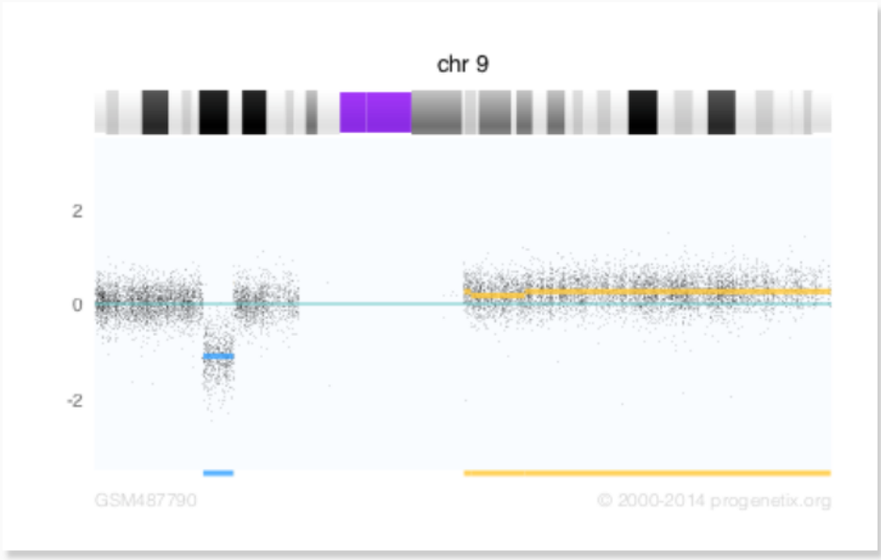
PLOS ONE 2012

	progenetix	arrayMap
techniques	cCGH, aCGH, WES, WGS	aCGH (+?)
scope	**sample** (e.g. combination of several experiments); **literature** tracking	**experiment**
content	**>31000** samples **>2700** publications	**>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 ...)

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:

- 65042 genomic copy number arrays
- 986 experimental series
- 333 array platforms
- ICD-O 253 ICD-O cancer entities
- 716 publications (Pubmed entries)



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	5791 samples ↗	[?]
BREAST CANCER	8594 samples ↗	[?]
COLORECTAL CANCER	3470 samples ↗	[?]
PROSTATE CANCER	1366 samples ↗	[?]
STOMACH CANCER	1457 samples ↗	[?]

ARRAYMAP NEWS

2016-04-11: Sorting cancer subset tables

2015-03-23: SIB Profile 2015

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.

THE ARRAYMAP CANCER GENOME RESOURCE

FIND CNAS BY GENE OR REGION	TP53	[ERBB2] 17:35097862-35138441:1	[?]
REGION SIZE | MAX COVERAGE (KB)	0 kb -	5000	250000 kb [?]
CLINICAL DATA	no followup required		
CITY		20 km	[?]
	Query Database		

1949 of 65042 cases matched the selection criteria.

SUBSET	PERCENT IN SUBSET
8507/3: Invasive micropapillary carcinoma (13/39)	33.3
C692: retina (14/82)	17.1
8260/3: Papillary adenocarcinoma, NOS (11/65)	16.9
8500/3: invasive carcinoma of no special type (1201/8188)	14.7
8560/3: Adenosquamous carcinoma (3/21)	14.3
Carcinomas: breast ca. (1254/8837)	
C50: breast (1254/8929)	
8500/2: Ductal carcinoma in situ, NOS (25/225)	
C32: larynx (3/29)	
8010/2: Carcinoma in situ, NOS (2/20)	
C187: sigmoid incl. rectosigmoid junction (13/140)	
8480/3: Mucinous adenocarcinoma (12/132)	
8522/3: Infiltrating duct and lobular carcinoma (4/44)	
8460/3: Micropapillary serous carcinoma [C56.9] (32/513)	
8130/1: Urothelial papilloma, NOS (11/184)	
C680: other urinary organs (11/184)	
C54: corpus uteri (19/330)	
8441/3: Serous adenocarcinoma, NOS (31/542)	
Carcinomas: esophagus ca. (32/571)	
Carcinomas: gastric ca. (80/1492)	

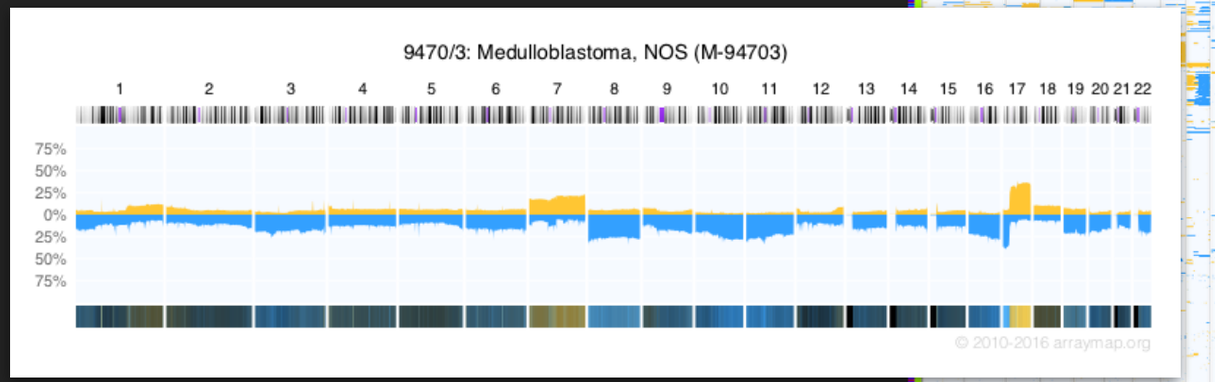
ICD Morphologies

64485 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.

31902 samples from progenetix have an associated "ICDMORPHOLOGYCODE" label.

400 subsets of type ICDMORPHOLOGYCODE will be parsed.

Subsets	arraymap
8140/3: adenocarcinoma, nos	9469
0000/0: not classified in icd-o 3 [e.g. non-neoplastic or benign]	8814
8500/3: invasive carcinoma of no special type	8188
9861/3: acute myeloid leukemia, nos	2831
8070/3: squamous cell carcinoma, nos	2443
9440/3: glioblastoma, nos	2294
9823/3: b-cell chronic lymphocytic leukemia/small lymphocytic lymphoma	2114
9470/3: medulloblastoma, nos	2052
9835/3: acute lymphoblastic leukemia, nos	1789
8010/3: carcinoma, nos	1430
8720/3: malignant melanoma, nos	1405
9500/3: neuroblastoma, nos	1333
small cell carcinoma	1195
cell renal cell carcinoma	1095
se large b-cell lymphoma, nos	1044
ular lymphoma, nos	867

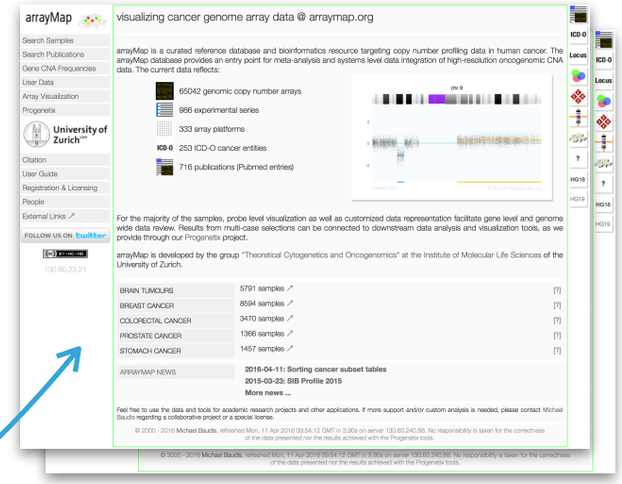


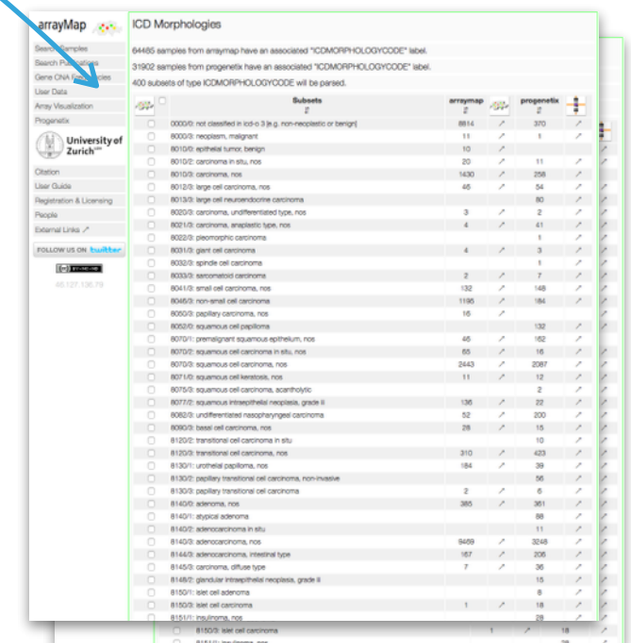
UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668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

ARRAYMAP (META)DATA PIPELINE

Automagical Processing Engine™



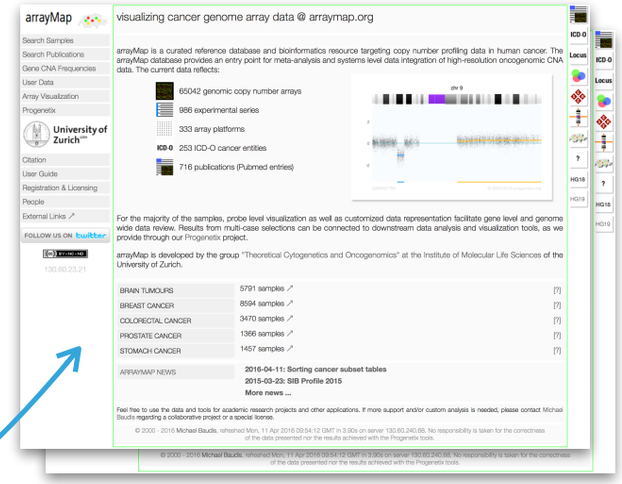


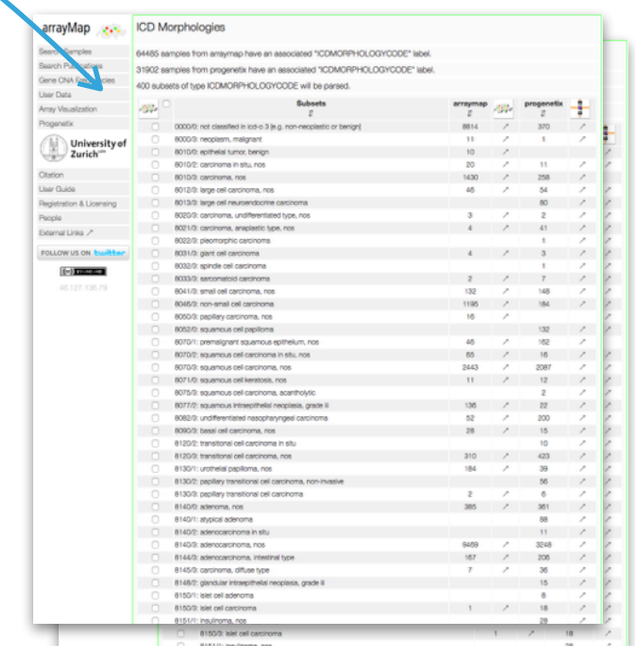
arrayMap

ARRAYMAP (META)DATA "PIPELINE"

Automagical Processing Engine™

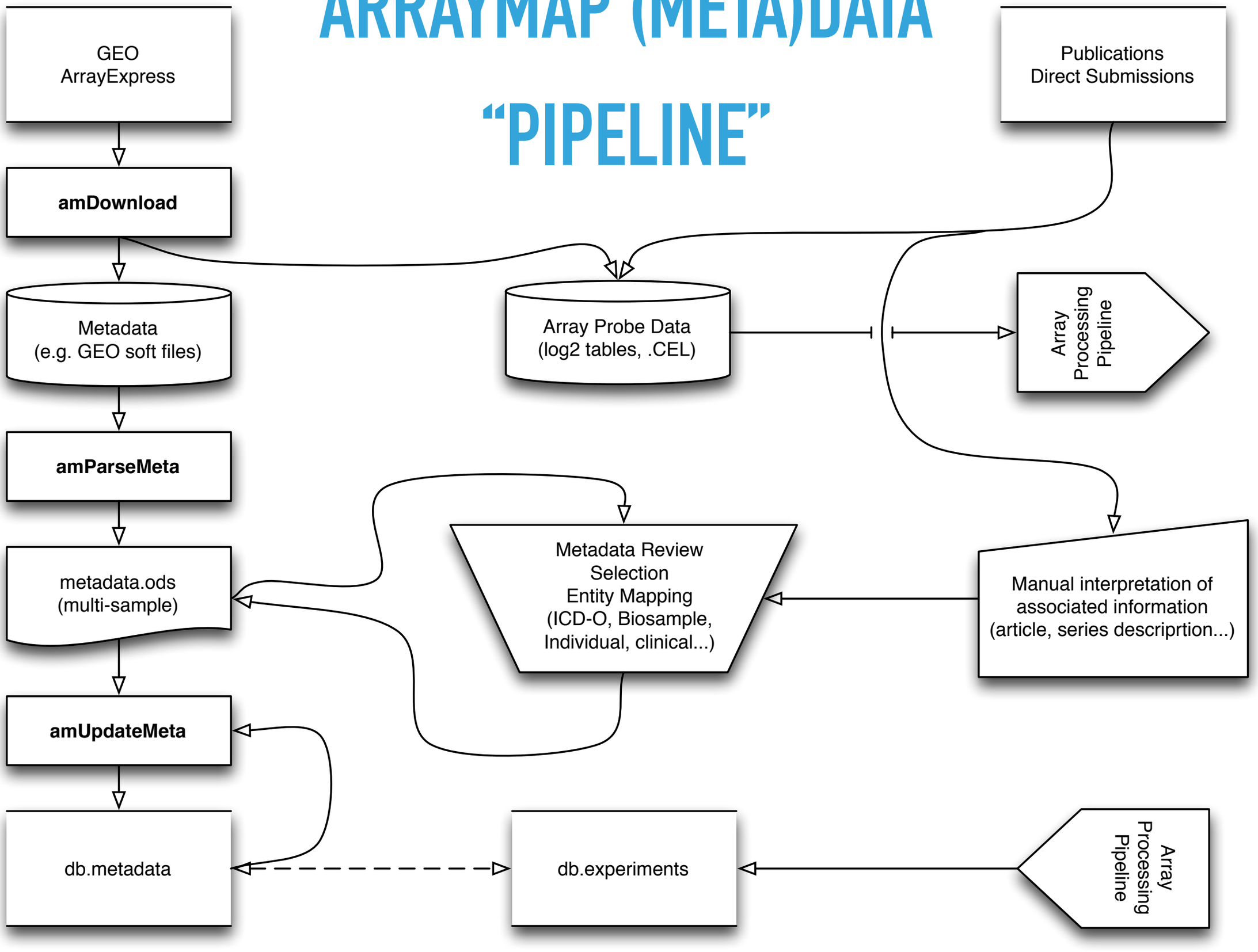
BIOCURATION





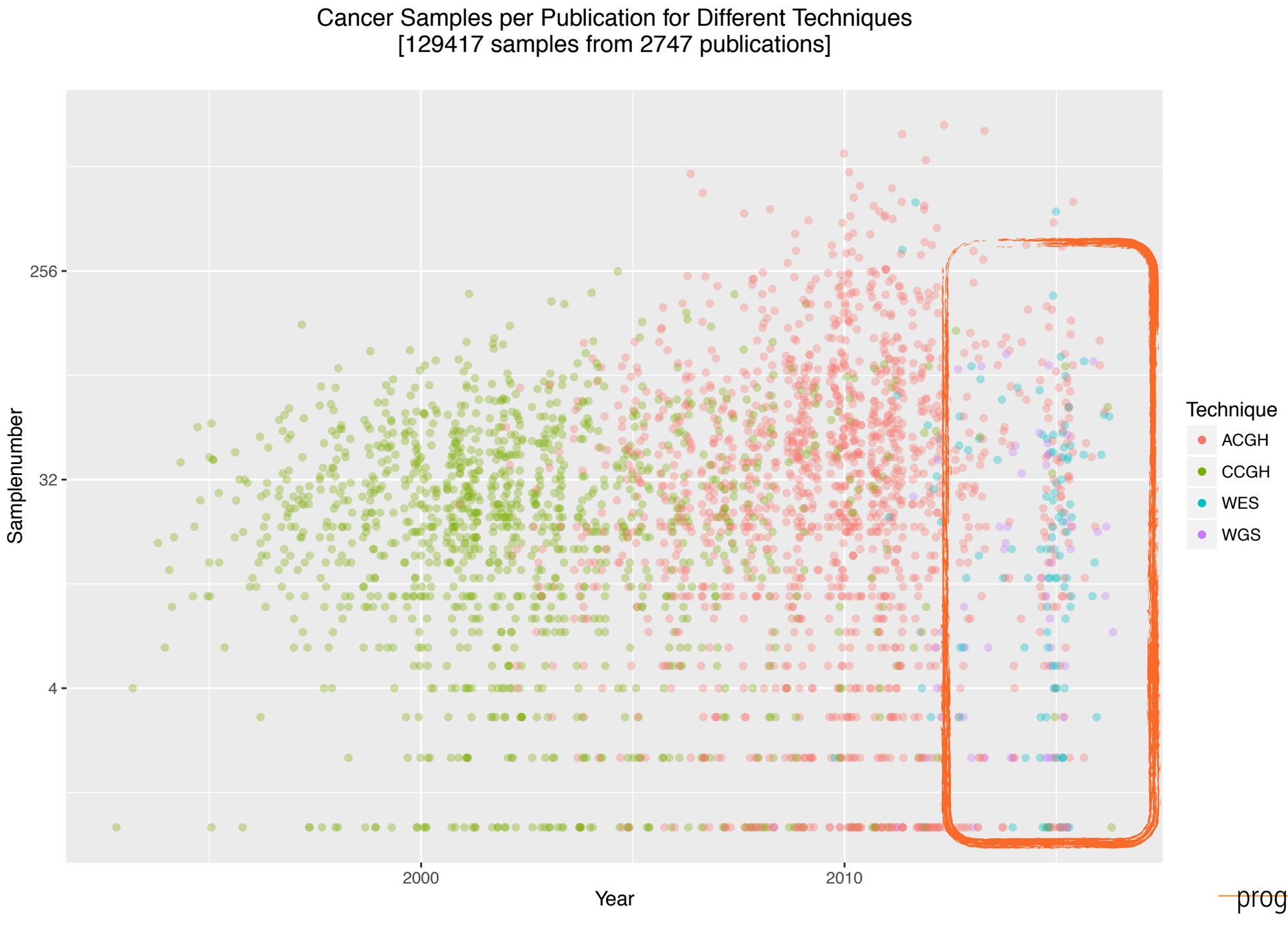
arrayMap

ARRAYMAP (META)DATA "PIPELINE"



arrayMap

SHIFT TO SEQUENCING BASED TECHNIQUES LEADS TO SEVERELY LIMITED DATA ACCESSIBILITY

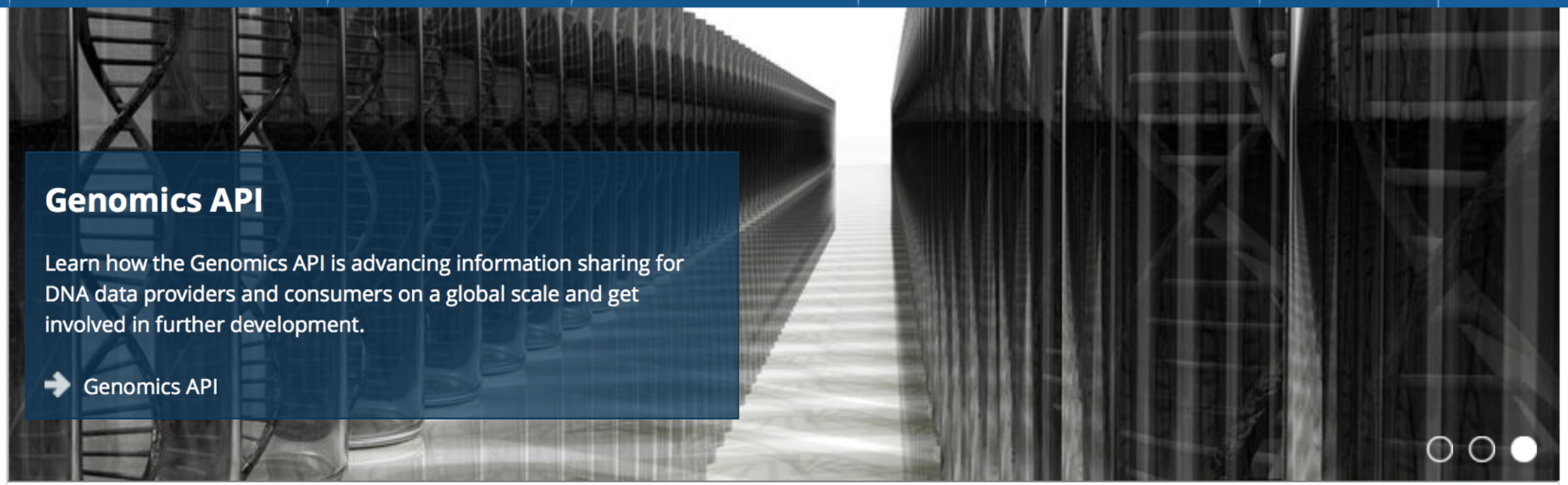


progenetix

Global Alliance
for Genomics & Health

Become a Member

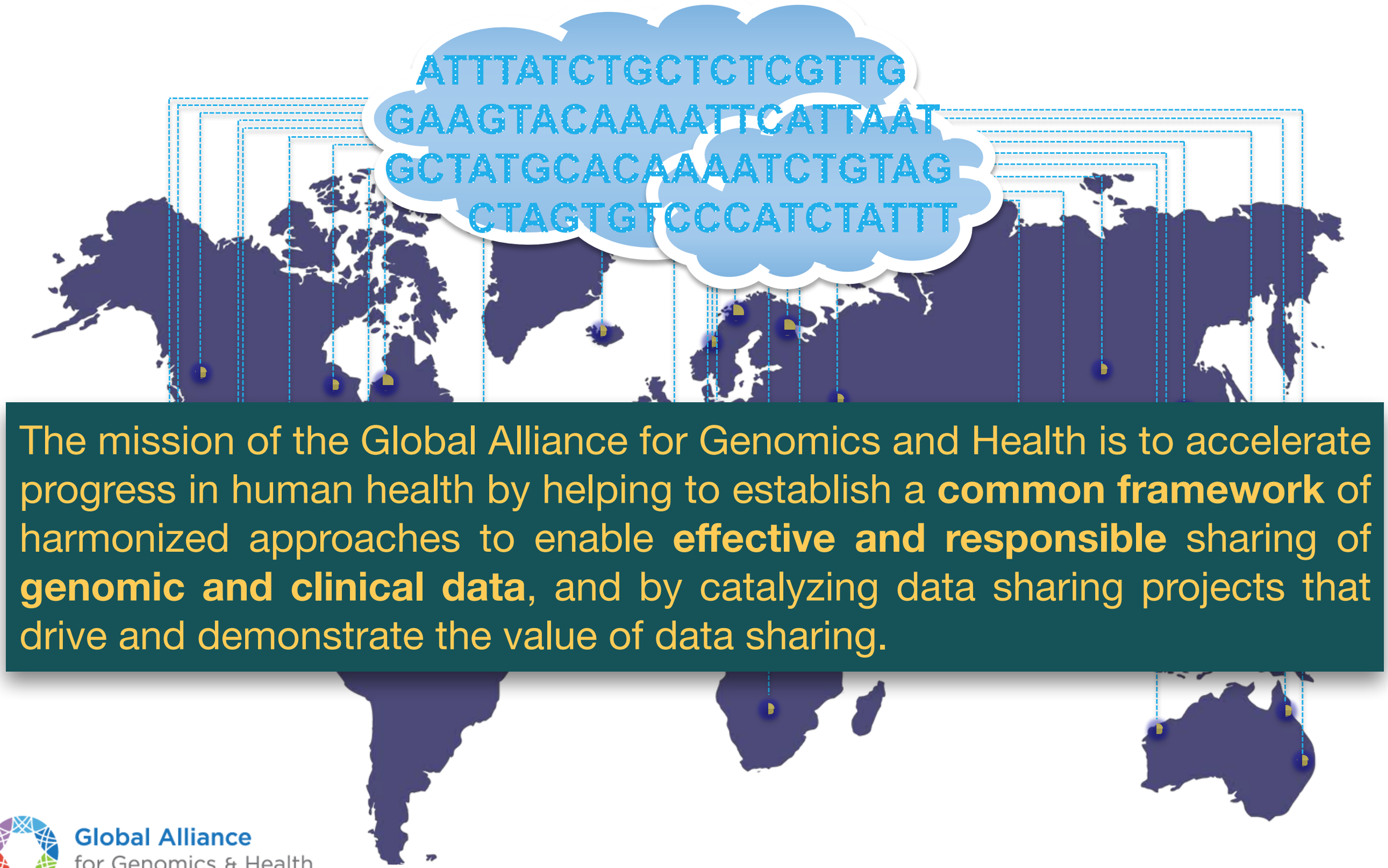
ABOUT GLOBAL ALLIANCE | WORKING GROUPS | PRODUCTS & PROJECTS | MEMBERSHIP | NEWS & EVENTS | CONTACT US



Our Work

The diverse members of the Global Alliance are working together to create interoperable approaches to catalyze projects that will help unlock the great potential of genomic data. Our four Working Groups advance Initiatives that develop key Work Products.

GA4GH API promotes sharing



The mission of the Global Alliance for Genomics and Health is to accelerate progress in human health by helping to establish a **common framework** of harmonized approaches to enable **effective and responsible** sharing of **genomic and clinical data**, and by catalyzing data sharing projects that drive and demonstrate the value of data sharing.

Global Alliance
for Genomics & Health

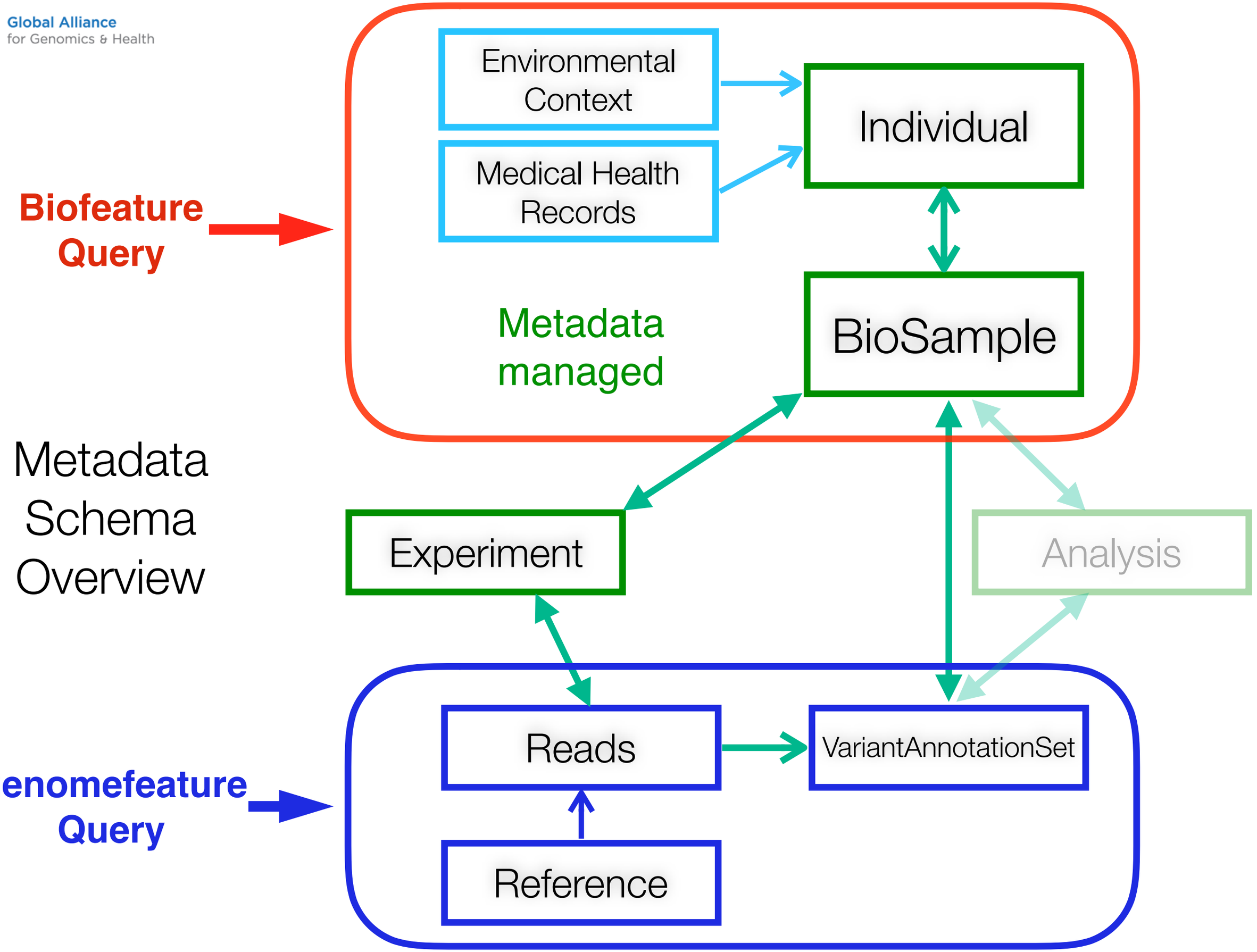
David Haussler

GA4GH DATA ANNOTATION PRINCIPLES

- Ontologies
  - limited use of named attributes
  - “OntologyTerm” object type as core for biological and experimental features
- External standards
  - ISO8601 time formats
- GA4GH managed standards
  - maintenance of core sequence file and data formats (VCF, BAM …) through affiliate partners/members of the Data Working Group

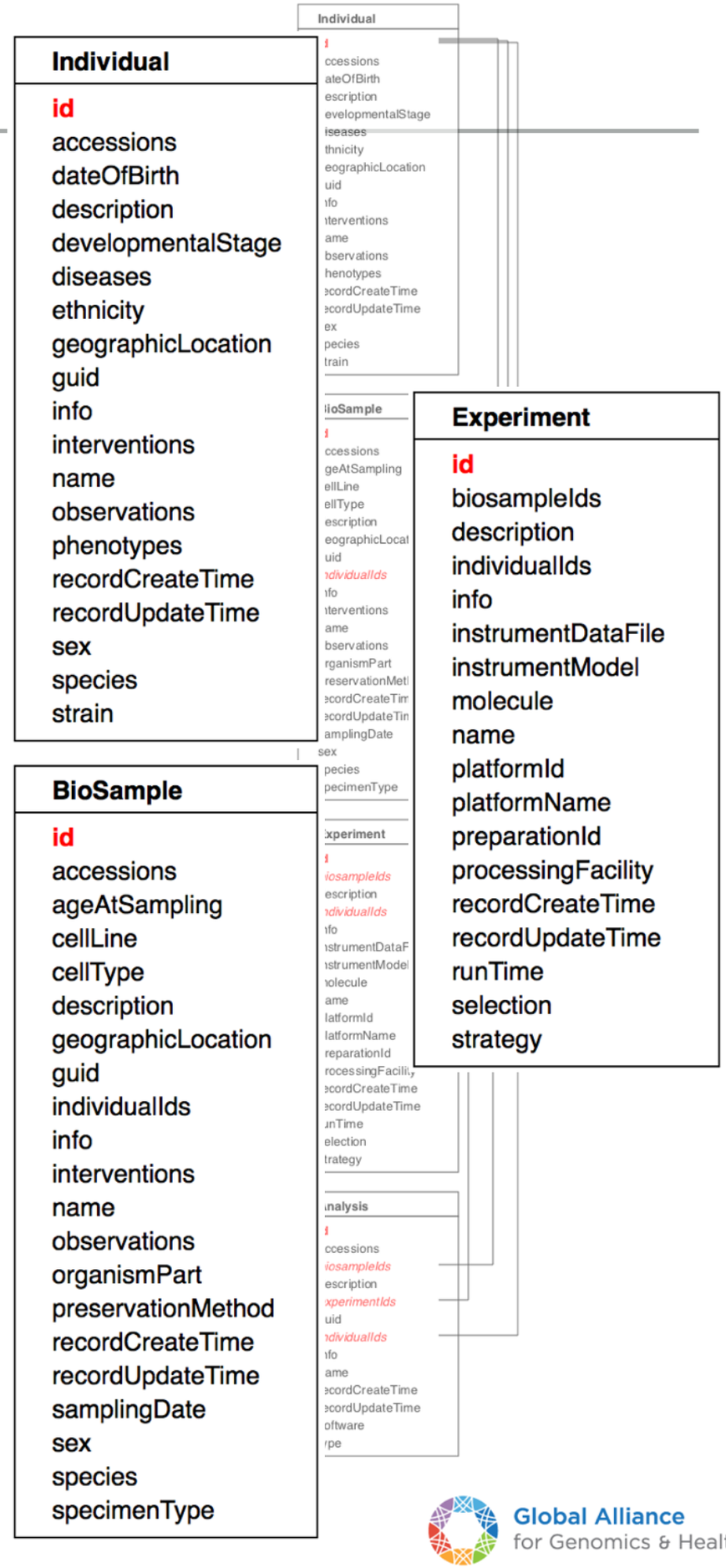
Global Alliance for Genomics & Health

Metadata Schema Overview



METADATA SCHEMA – "BIODATA" OBJECTS

- ***Individual*** (i.e., basic biological entity) and ***BioSample*** (e.g. micro dissected part of a tissue biopsy, environmental sample) are the basic "Bioobjects"
- ***Experiment*** describes the technical procedures used in the analysis of (an aliquot) of the *BioSample*
- ***Analysis*** may be used to store "interpreted" results of an experiment
- ***VariantAnnotationSet*** and other low-level analyses reference *Individual*, *Biosample*, *Experiment* for feature inheritance



Global Alliance for Genomics & Health

ONTOLOGYTERM IN THE GA4GH SCHEMA – CURRENT STATUS

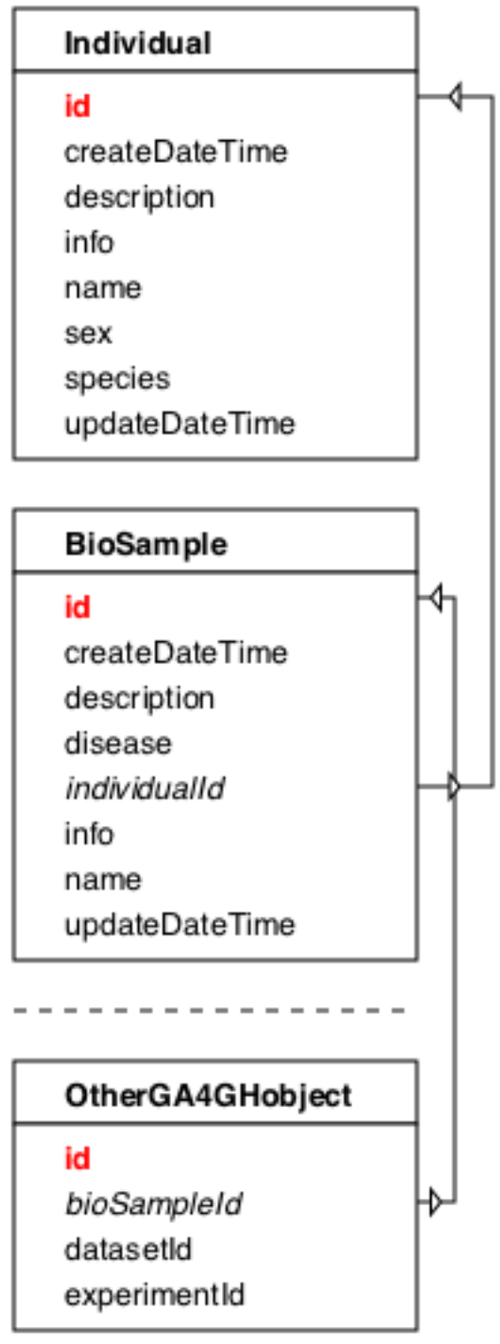
```
/**
An ontology term describing an attribute. (e.g. the phenotype attribute
'polydactyly' from HPO)
*/
record OntologyTerm {
  /**
  Ontology source identifier - the identifier, a CURIE (preferred) or PURL for an
  ontology source e.g. http://purl.obolibrary.org/obo/hp.obo
  It differs from the standard GA4GH schema's “id” in that it is a URI pointing to
  an information resource outside of the scope of the schema or its implementation.
  */
  string id;

  /* Ontology term - the representation the id is pointing to.*/
  union { null, string } term = null;

  /**
  Ontology source name - the name of ontology from which the term is obtained
  e.g. 'Human Phenotype Ontology'
  */
  union { null, string } sourceName = null;

  /**
  Ontology source version - the version of the ontology from which the OntologyTerm
  is obtained; e.g. 2.6.1. There is no standard for ontology versioning and some
  frequently released ontologies may use a datestamp, or build number.
  */
  union { null, string } sourceVersion = null;

}
```

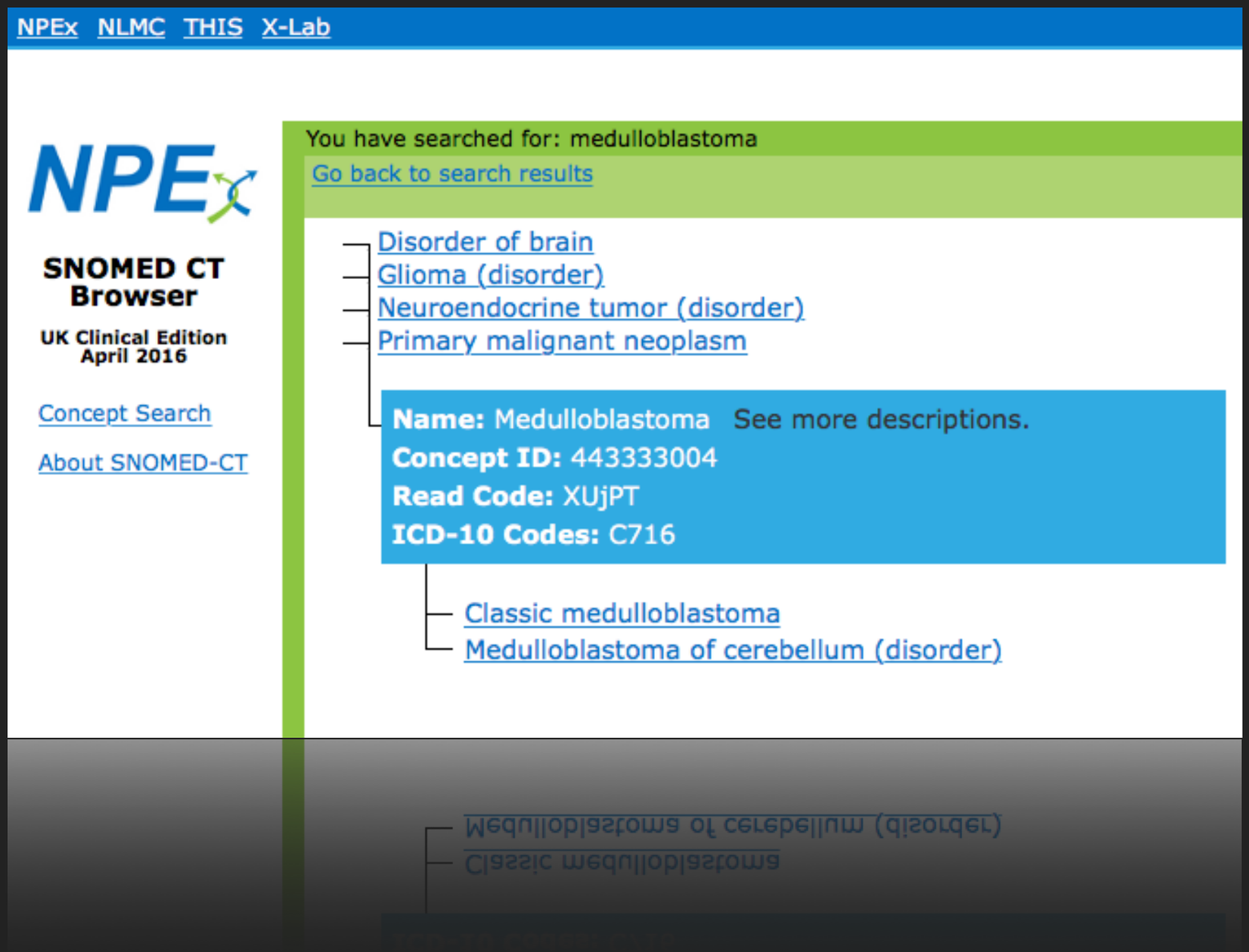


BIOFEATURE OBJECTS AS ONTOLOGY WRAPPERS?

```
"bioFeatures": [
    {
        "description":          "squamous cell carcinoma, base of tongue, stage 2",
        "ageAtObservation":     "P56Y6M",
        "timeOfObservation":    "2015-03-24T15:23:00Z",
        “updateDateTime”:       "2016-04-14T09:02:00Z",
        "ontologyTerms": [
            {
                "ontologyId":       "http://purl.obolibrary.org/obo/DOID_0050865",
                "term":             "tongue squamous cell carcinoma",
                "sourceName":       "disease_ontology",
                "sourceVersion":    "2016-01-25"
            },
            {
                "ontologyId":       "http://purl.obolibrary.org/obo/UBERON_0006919",
                "term":             "tongue squamous epithelium",
                "sourceName":       "Uberon multi-species anatomy ontology",
                "sourceVersion":    "2016-01-25"
            },
            {
                "ontologyId":       "http://purl.obolibrary.org/obo/UBERON_0010033",
                "term":             "posterior part of tongue",
                "sourceName":       "Uberon multi-species anatomy ontology",
                "sourceVersion":    "2016-01-25"
            },
        ],
    }
]
```

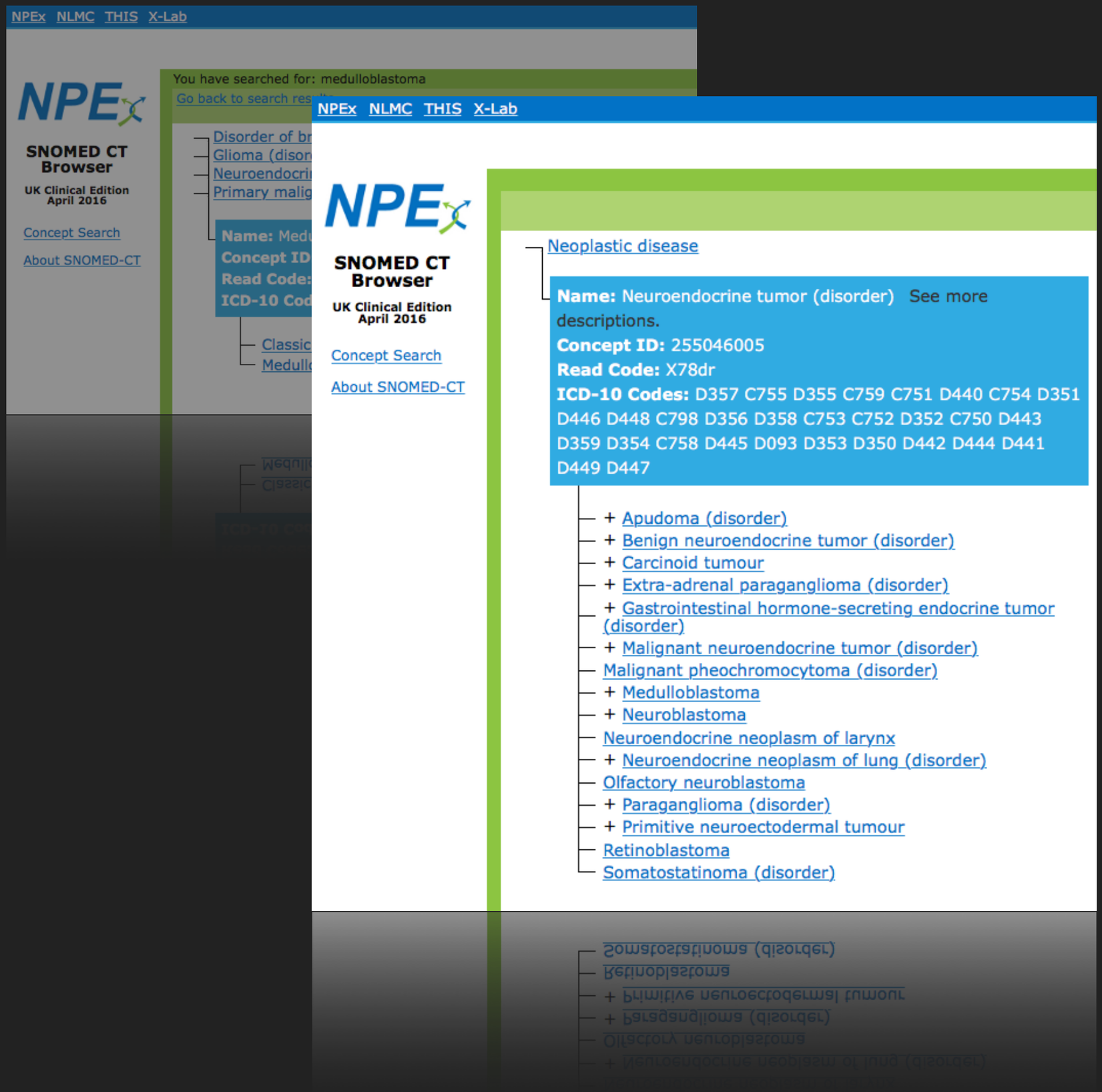
We're sorry but something has gone wrong. We have been notified of this error.

ONTOLOGIES YOU CAN TRUST?



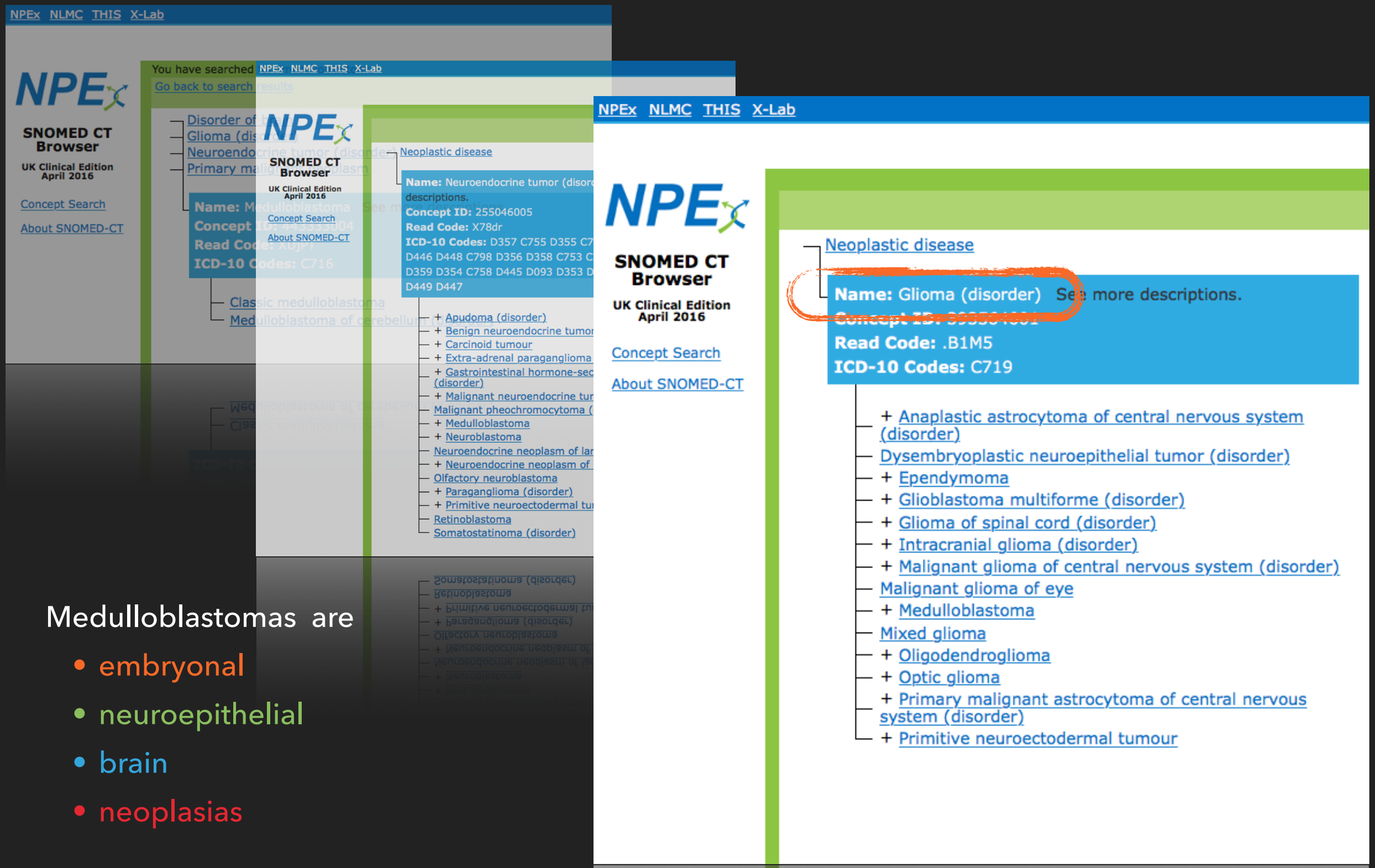
http://www.snomedbrowser.com/Codes/Details/276826005

ONTOLOGIES YOU CAN TRUST?



Source:www.snomedbrowser.com

ONTOLOGIES YOU CAN TRUST?



Medulloblastomas are

- embryonal
- neuroepithelial
- brain
- neoplasias

. . . BUT NOT GLIOMAS & MOST CERTAINLY NOT A NEUROENDOCRINE TUMOURS

Source:www.snomedbrowser.com

ONTOLOGIES ARE RARELY USED IN CASE REPORTING

- Medical practice relies on established, slow moving classification systems.
- Medical diagnoses consist of an abundance of observations and classification items.
- We do not have (never will?) enough ontology concepts for detailed disease descriptions (Where to stop?)
- Relationships may help - but how to do them uniformly?

ONTOLOGIES ARE RARELY USED IN CASE REPORTING

- Medical practice relies on established, slo ving classification systems.
- Medical diagnoses consist abundance of observations and cla tion items.
- We do not h ever will?) enough ontology concepts for de disease descriptions (Where to stop?)
- ationships may help - but how to do them uniformly?

THE GOLD STANDARD FOR A MEDICAL DIAGNOSIS IS STILL WELL WRITTEN PROSE.

CORE PROBLEMS AND CONCEPTS TO BE ADDRESSED

- suitability of **ontologies as core of metadata** features for federated data mining
- mapping ontologies: **WHO** and **HOW**
- identification of essential non-OT attributes and stable definition using internationally accepted standards
- development of a strategy for implementation of **ontology based data annotations for reference data resources**, e.g. Elixir, EBI, SIB …

DRIVING GA4GH METADATA SCHEMA

- arrayMap for GA4GH
  - metadata schema development through implementation of arrayMap resource data
  - testing of OntologyTerm object for covering biodata
  - implementation using EMBL-EBI ontology services

DRIVING BEACON DEVELOPMENT

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

DRIVING GA4GH METADATA SCHEMA

- arrayMap for GA4GH
  - metadata schema development through implementation of arrayMap resource data
  - testing of OntologyTerm object for covering biodata
  - implementation using EMBL-EBI ontology services

DRIVING BEACON DEVELOPMENT

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

MAPPING ARRAYMAP TO GA4GH TRANSITIONAL


```
[
   {
      "individuals" : [
         {
            "createDateTime" : "2015-01-01T12:00:00Z",
            "id" : "AM_IND__18769486_MB0262",
            "sex" : {
               "ontologyId" : "http://purl.obolibrary.org/obo/PATO_0020001",
               "sourceName" : "Ontology: PATO (Phenotypic quality)"
            },
            "species" : {
               "ontologyId" : "http://purl.obolibrary.org/obo/NCBITaxon_9606",
               "sourceName" : "NCBITaxon Ontology"
            },
            "updateDateTime" : "2016-04-13T18:51:01Z"
         }
      ]
   },
   {
      "biosamples" : [
         {
            "ageAtSampling" : "P14Y",
            "createTime" : "2015-01-01T12:00:00Z",
            "diagnosis" : {
               "description" : "medulloblastoma [classic]",
               "ontologyTerms" : [
                  {
                     "ontologyId" : "http://purl.bioontology.org/ontology/SNMI/M-94703",
                     "sourceName" : "Systematized Nomenclature of Medicine, International Version",
                     "term" : "Medulloblastoma, NOS"
                  },
```

Beacon ArrayMap

First Prototype of a Beacon v0.2 implementation for ArrayMap.

See documentation and open questions.

Chromosome 11

Position 34439881

Reference GRCh38

Dataset (8070/3) squamous cell carcinoma, nos

Variant Class DEL (Deletion)

http://beacon-arraymap.vital-it.ch/v0.2/query?
chromosome=11&position=34439881&reference=GRCh38&dataset=8070/3&variantClass=DEL

Beacon Query

http://beacon-arraymap.vital-it.ch/info

Beacon Info

http://beacon-arraymap.vital-it.ch/v0.3/query?
referenceName=11&start=34439881&assemblyId=GRCh38&datasetIds=8070/3&variantClass=DEL

Beacon v0.3

Heinz Stockinger, Séverine Duvaud & SIB Technology Group


```
{
  "query": {
    "referenceName": "11",
    "start": "34439881",
    "assemblyId": "GRCh38",
    "datasetIds": "8070/3"
  },
  "response": {
    "exists": "overlap",
    "info": "ok",
    "error": null,
    "observed": 57,
    "NOT_BEACON_ARRAYMAP_DEBUG_INFO": {
      "matchedSegments": [
        {
          "matchedSampleUID": "GSM675751",
          "matchedDataSet": "8070/3",
          "matchedSegment": {
            "SEGSTART": 232718,
            "SEGSOURCE": "aCGH",
            "CHRO": "11",
            "SEGSIZE": 134462435,
            "SEGVALUE": -0.2637,
            "SEGTYPE": -1,
            "SEGSTOP": 134695153
          }
        },
  ...
  "NOT_BEACON_totalInDataSet": 0
}
```


Swiss Institute of Bioinformatics

arrayMap

elixir

BAUDISGROUP @ UZH

NI AI
MICHAEL BAUDIS
(HAOYANG CAI)
PAULA CARRIO CORDO
LINDA GROB
SAUMYA GUPTA
(NITIN KUMAR)
ALESSIO MILANESE

GA4GH DWG

ANTHONY BROOKES
MARK DIEKHANS
MELISSA HAENDEL
SARAH HUNT
STEPHEN KEENAN
SUZY LEWIS
DAVID LLOYD
MICHAEL MILLER
HELEN PARKINSON
ELEANOR STANLEY
DAVID STEINBERG

SIB

HEINZ STOCKINGER
SÉVERINE DUVAUD
VASSILIOS IOANNIDIS
DANIEL TEIXEIRA

ELIXIR & CRG

JORDI RAMBLA DE ARGILA
SABELA DE LA TORRE PERNAS
SUSANNA REPO

RANDOM LINKS

- *arraymap.org*
- *beacon-arraymap.vital-it.ch*
- *github.com/ga4gh/schemas/tree/metadata*
- *www.progenetix.org/publications/*
- *wiki.progenetix.org*
- *ga4ghdata.org*
- *ga4ghdata.org/Sites/github/ga4gh/biosample/html/api/biodata.html*