Updates on
Progenetix
Oncogenomics resource

2020 Oct 30

Qingyao Huang
Baudis group

Presentation Agenda

01 **Introduction**
Progenetix resource

02 **New meta-data features**
Domain-specific mapping

03 **New data sources**
Sample expansion

04 **Data standards**
CURIE, GA4GH, Phenopackets schema

05 **Beacon protocol**
Features and prospects

06 **New web interface**
Many features…

1

Introduction

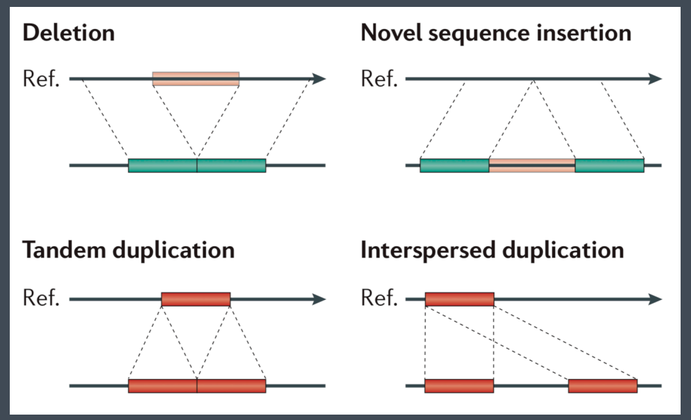
Copy number variation (CNV)

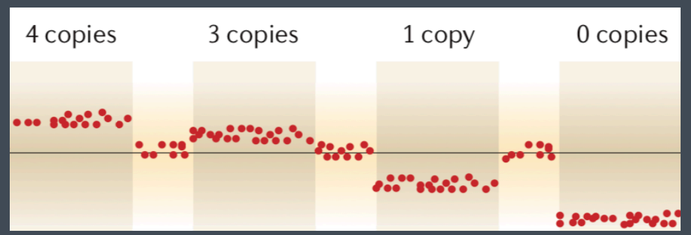
Structural changes in cancer genomes

Exhibit distinct patterns cross cancer types

Marker for prognostic and subtype stratification

Molecular pathways in cancer development





Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011

Progenetix

- Released in 2001, currently the most comprehensive reference resources for copy number aberration in cancer.
- Currently hosts 138’334 copy number samples (incl. 115’158 cancer samples of 788 types) from array-based as well as sequencing platforms.
- Supports development of data standard and exchange protocols through Global Alliance for Genomics and Health (GA4GH)

2

Meta-data

Ontology features

Cancer type classification

ICD-O and NCIt

Uberon anatomy

Tissue origin

Geographical location

Where cancer research is conducted

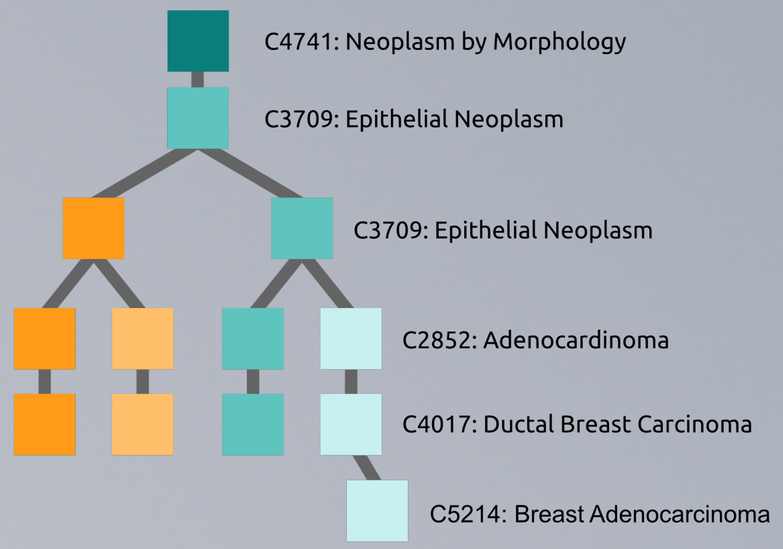
Ancestry background

HANCESTRO

Cancer type classification

C50.9: Breast, NOS

8500/3: Infiltrating duct carcinoma, NOS



ICD-O system

- Classical standard dual coding system for oncology
- Primary site (topography)
- Type of tissue (morphology)

NCI thesaurus (NCIt)

- Logic-based terminology
- Organised in hierarchical structure
- 7’579 terms in current Neoplasm Core (v20.05)
- Relate key terms, molecular characteristics, EVS resourses…

Pro and cons of both systems

ICD-O M+T (1550 pairs)

Location specificity

icdot-C18.5: **Splenic flexure of** colon
icdom-81403: Adenocarcinoma, NOS

NCIT:C4349: Colon Adenocarcinoma

Molecular Marker specificity

Triple-negative breast cancer
Gene translocation
TP53 status

NCIt (788 terms)

mondo

THE WORLD'S DISEASE CONCEPTS, UNIFIED

ICD topography

Clinical and diagnostic aspects of tumor entities

icdot-C53.9 cervix uteri

icdot-C75.5
Aortic body and other paraganglia

icdot-C03.0 Upper Gum

Mapped all **221** ICD topography codes
Text mining + manual curation

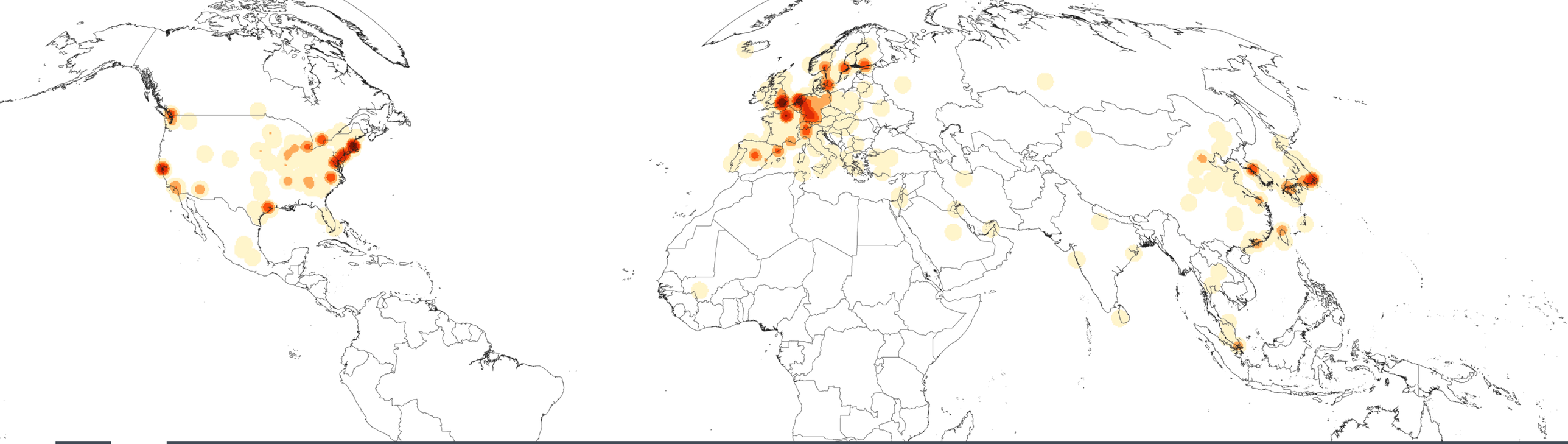
Uberon

Cross-species anatomical ontology
Functional and developmental lineages
Cross-database reference
Spatial relations

UBERON:0000002
uterine cervix

UBERON:0034978
paraganglion (generic)

UBERON:0011601
gingiva of upper jaw



Geographical location of research

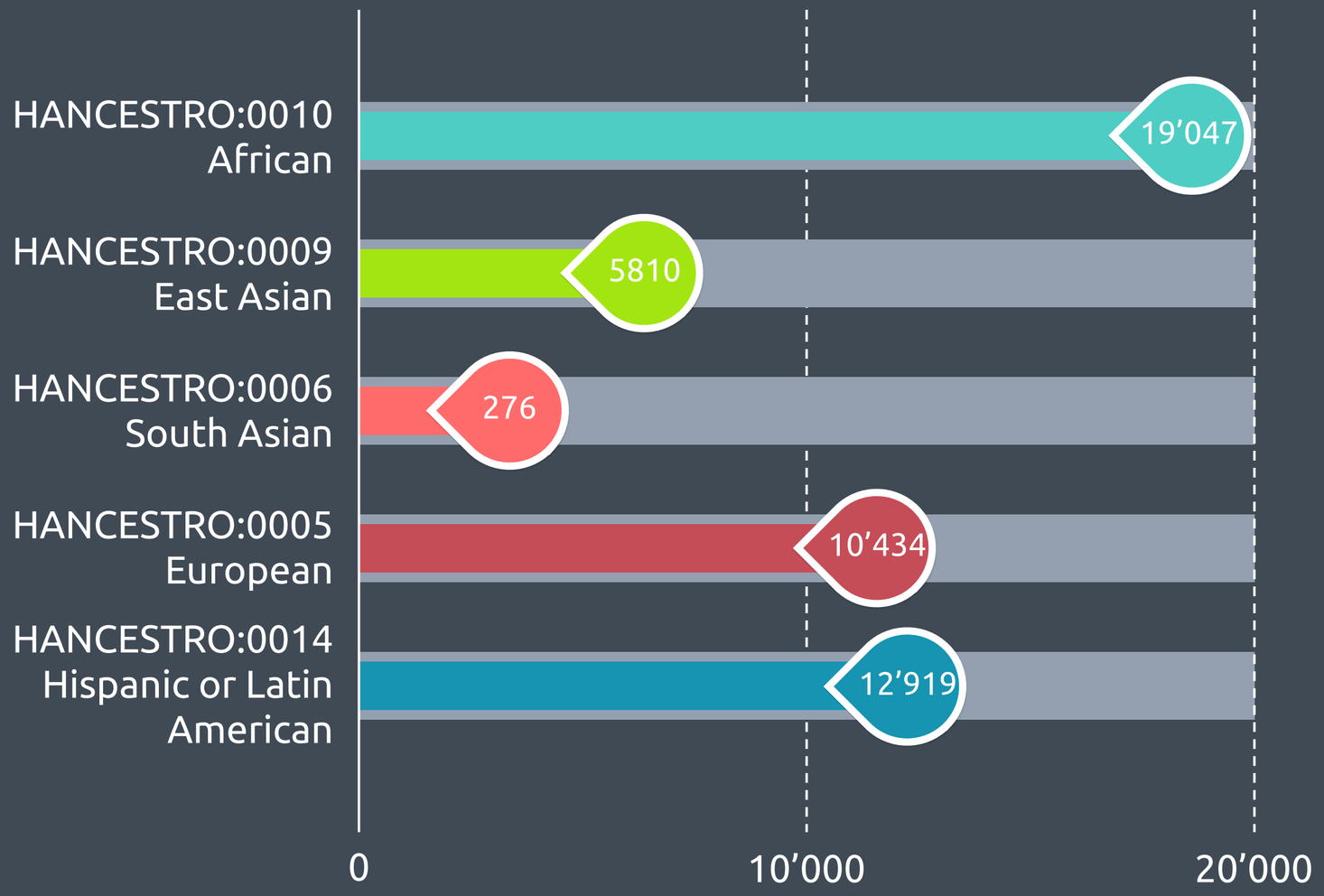
“GeoLocation” attribute from GA4GH SchemaBlocks (v0.0.1)

properties:

label, longitude, latitude, altitude, city, country, precision

Population background classification

Currently estimated individuals, mapping ongoing…



Data-driven estimation of population background

Genome-wide SNP information of cancer genome

Benchmarked on noise from somatic changes

Classification to labels from 1000 Genomes Project

Label mapping to HANCESTRO

5 continent groups and 26 population groups

Mapped to

5 and 24 HANCESTRO terms

Huang, Q., & Baudis, M. (2020). Enabling population assignment from cancer genomes with SNP2pop. *Scientific reports*, *10*(1), 1-9.

3

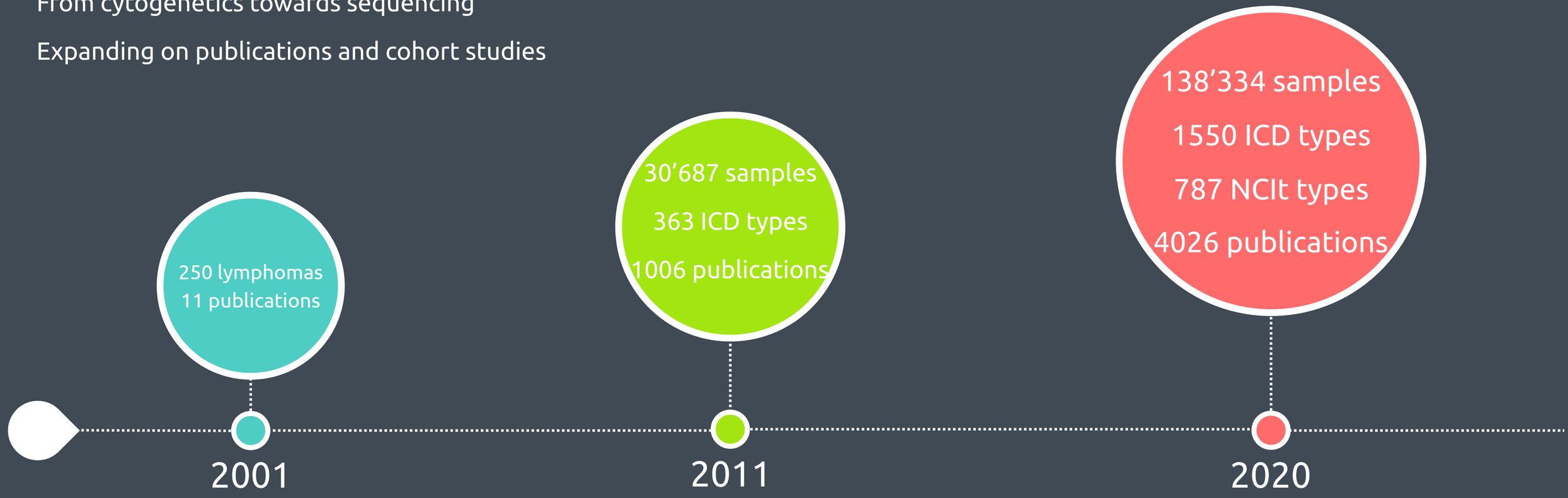
Data expansion

Data growth

From hematological malignancy to most studied cancer types

From cytogenetics towards sequencing

Expanding on publications and cohort studies



Public data repositories

Gene Expression Omnibus (GEO)

63’568 samples
346 NCIt cancer types
cCGH and aCGH

Array Express

4351 samples
148 NCIt cancer types
aCGH

The Cancer Genome Atlas (TCGA)

22’142 samples
182 NCIt cancer types
aCGH

cBioPortal

19712 samples
422 NCIt cancer types
aCGH, WES and WGS

Total: 138’334 samples with 787 NCIt cancer types

Data inclusion Process

01 Data retrieval & analysis

- Total and allelic copy number estimation
- CNV segmentation

02 Data quality evaluation

- Baseline adjustment
- Segment evaluation
- Global CNV fraction

03 Metadata extraction & curation

- Automated text extraction and inference
- Manual curation

4

Data standards & Modeling

Why data standards?

Inter-operable

Exchange

Integration

Accessible

Permission

Protection

Reliable

Consistency

Reduce redundancy

Progenetix-conformed data exchange formats

Progenetix data objects are identified with Compact URI (CURIE) syntax,
e.g. Biosample ID as pgx:pgxbs-kftvgk8h

prefix:reference

NCIT:C4349, PMID:23079654, arrayexpress:E-MEXP-1330, geo:GSE21420…

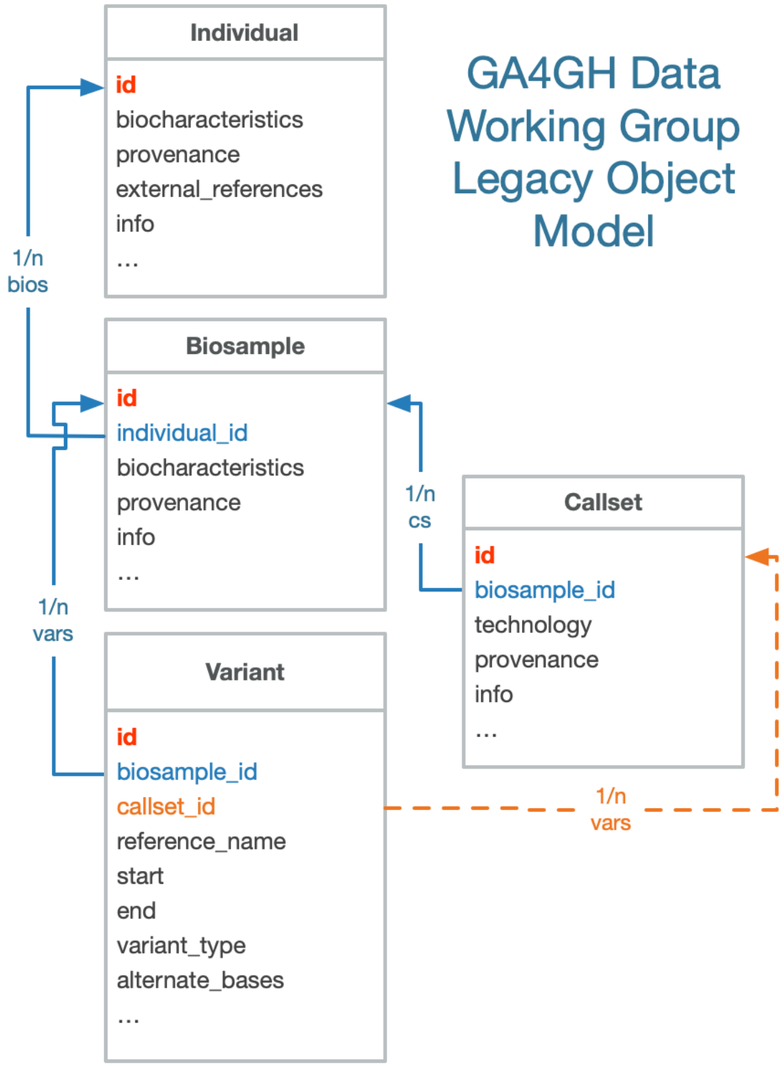
GA4GH specification

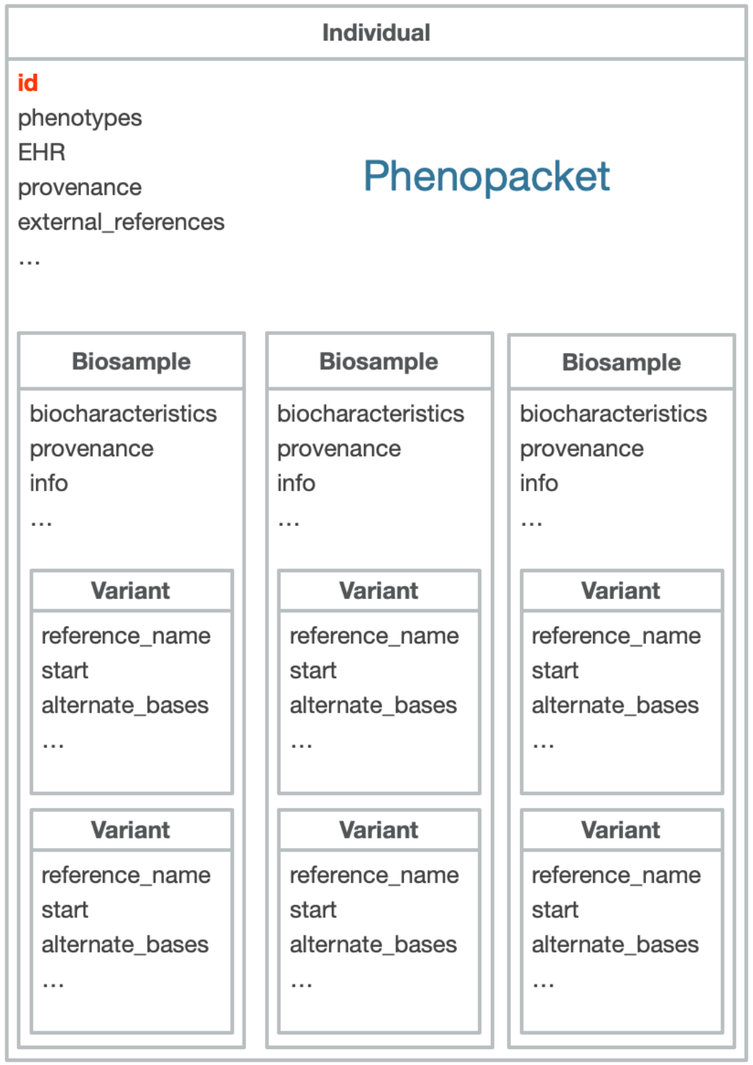
For sequence and variation data

Phenopacket Schema

For clinical phenotypes and environment data

Current data object model





5

Beacon+ Protocol

Beacon project

Global Alliance
for Genomics & Health

Beacon

Facilitate genomic data sharing

Driver Project of GA4GH

Framework of web services and queries

Security standards to protect sensitive data

Features and prospects of Beacon specifications

Query types

Precise (chr17:7577121G>A)

Range (start to end positions with specified tolerance)

Gene element-centered

Filters

CURIE standard prefixes

NCIt, phenotype, experiment factor

Handover object delivery

Anonymous link to external services with own security/ privacy implementations

Authentication

Network authentication empowered by ELIXIR AAI integration

6

Web interface

progenetix

About Progenetix

Cancer CNV Profiles

Search Samples

Publication DB

Services

Upload & Plot

Documentation

Beacon+

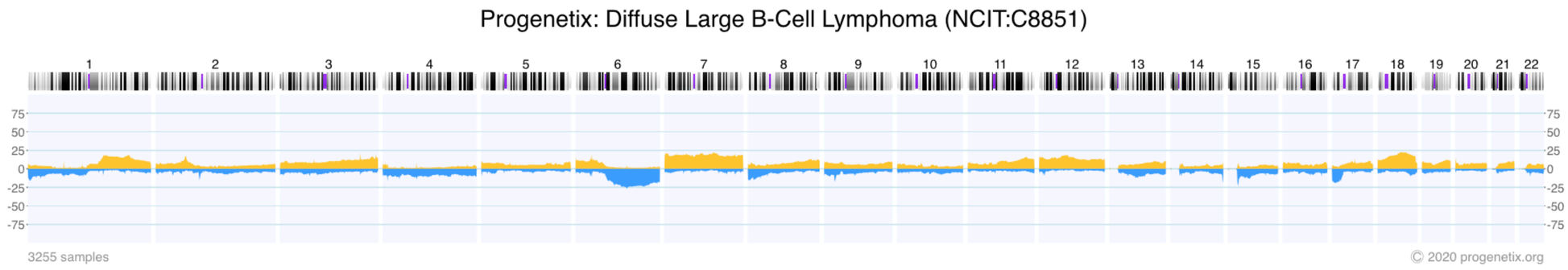
Baudisgroup @ UZH

Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies.

For exploration of the resource it is suggested to either start with:

- Cancer Types
- searching for CNVs in genes of interest



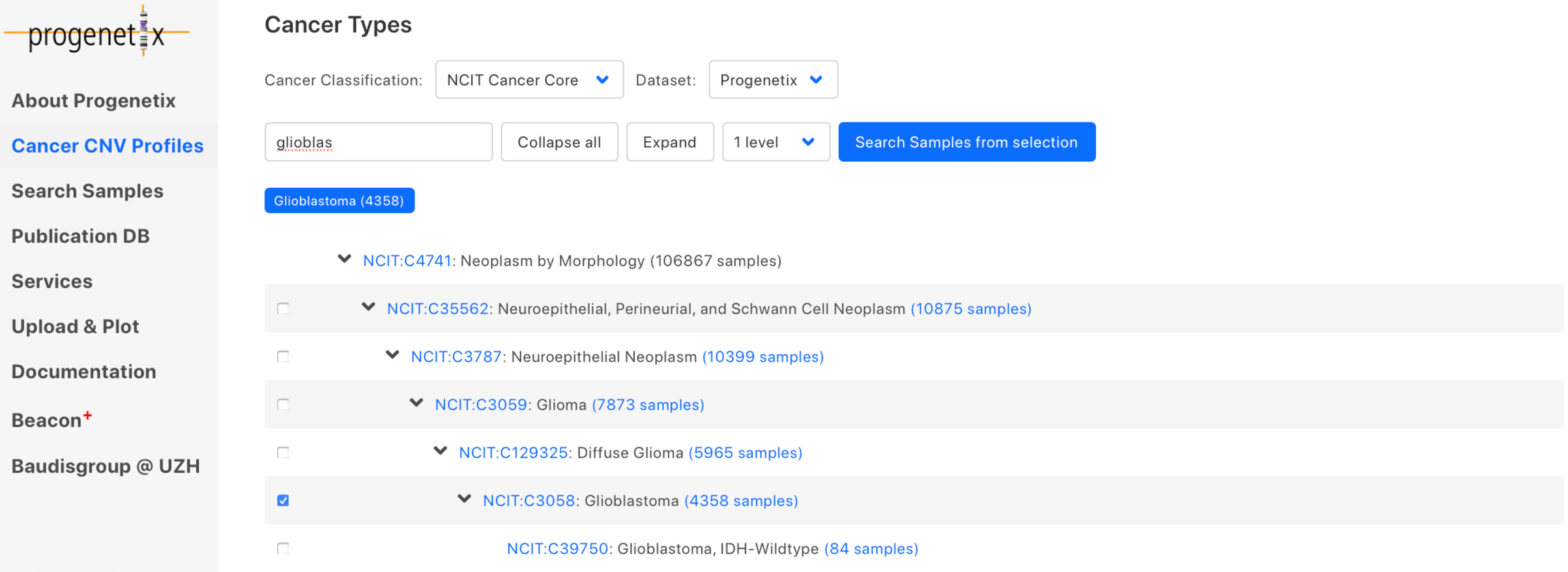
The resource currently contains genome profiles of **138334** individual samples and represents **698** cancer types, according to the NCIt "neoplasm" classification.

Additionally to this genome profiles and associated metadata, the website present information about publications (currently **4026** articles) referring to cancer genome profiling experiments.

Homepage

Intro and summary statistics

Aggregated CNV profile of a randomly chosen set of samples



NCIt term visualisation in hierarchical tree

Search by keywords

Expand/Collapse tree branches to certain level

Select samples for data visualisation and download

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Search Samples

CDKN2A Deletion Example | MYC Duplication | TP53 Del. in Cell Lines

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "highly focal" hits (here i.e. <= ~1Mbp in size). The query is against the Progenetix and arrayMap collections. It can be modified e.g. through changing the position parameters or diagnosis.

Gene Spans | Cytoband(s)

Reference name: 9

(Structural) Variant Type: DEL (Deletion)

Start or Position: 21500001-21975098

End (Range or Structural Var.): 21967753-22500000

Cancer Classification(s): NCIT:C3058: Glioblastoma (4358) ×

Biosample Type:

Filters:

Filter Logic: AND

City: Select...

Query Beacon

Sample search by CNV
Query page

Sample search by CNV Results Tab

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Search Samples

Assembly: GRCh38 **Chro:** 9 **Start:** 21500001-21975098 **End:** 21967753-22500000

progenetix

Samples: 668
Variants: 286
Calls: 675
$f_{alleles}$: 0.000088

Phenopackets
Callsets Variants
Variants in UCSC
UCSC region
JSON Response

Results | Biosamples | Biosamples Map | Variants


```
{
    "data": [
        {
            "biosamples": [
                {
                    "externalReferences": [
                        {
                            "id": "PMID:23079654"
                        }
                    ],
                    "histologicalDiagnosis": {
                        "id": "NCIT:C3058",
                        "label": "Glioblastoma"
                    },
                    "id": "pgxbs-kftvgk8h",
                    "sampledTissue": {
                        "id": "UBERON:0001869",
                        "label": "cerebral hemisphere"
                    },
                    "variants": [
                        {
                            "_id": "5bab578b727983b2e00ca99e",
                            "biosample_id": "pgxbs-kftvgk8h",
                            "callset_id": "pgxcs-kftvmlzx",
                            "digest": "9:21548871-21999595:DEL",
                            "end_max": 21999595.0,
                            "end_min": 21999595.0,
                            "info": {
                                "cnv_length": 450724,
                                "cnv_value": null
                            },
                            "mate_name": null,
                            "reference_name": "9",
                            "start_max": 21548871.0,
                            "start_min": 21548871.0,
                            "updated": "2018-09-26 09:50:58.094031",
                            "variant_type": "DEL",
                            "variantset_id": "AM_VS_GRCH36"
                        }
                    ]
                }
            ],
            "id": "pxf__pgxind-kftx2am8",
            "subject": "pgxind-kftx2am8"
        },
```

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Search Samples

Assembly: GRCh38 **Chro:** 9 **Start:** 21500001-21975098 **End:** 21967753-22500000 **Type:** DEL **Filters:** NCIT:C3058

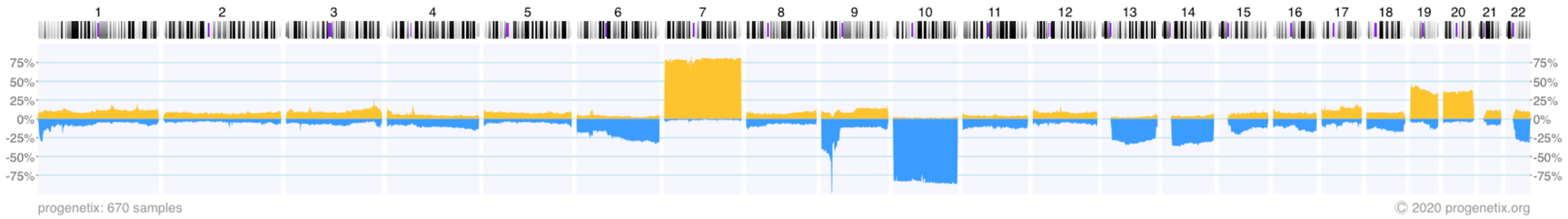
progenetix

Samples: 668
Variants: 286
Calls: 675
$f_{alleles}$: 0.000088

Phenopackets
Callsets Variants
Variants in UCSC
UCSC region
JSON Response

Visualization options

Results | Biosamples | Biosamples Map | Variants



Subsets	Subset Samples	Query Matches	Subset Match Frequencies
icdot-C71.4	4	1	0.250
icdom-94403	4274	664	0.155
NCIT:C3058	4358	664	0.152
icdot-C71.1	14	2	0.143
icdot-C71.9	6684	651	0.097
NCIT:C3796	84	4	0.048
icdom-94423	84	4	0.048
icdot-C71.0	1712	14	0.008

Sample search by CNV Results Tab

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Search Samples

Assembly: GRCh38 **Chro:** 9 **Start:** 21500001-21975098 **End:** 21967753-22500000 **Type:** DEL **Filters:** NCIT:C3058

progenetix

Samples: 668
Variants: 286
Calls: 675
$f_{alleles}$: 0.000088

Phenopackets
Callsets Variants
Variants in UCSC
UCSC region
JSON Response

Visualization options

Results | Biosamples | Biosamples Map | Variants

JSON | Download Response

Id	Description	Classifications	Identifiers	DEL	DUP	CNV
pgxbs-kftvgk8h	Glioblastoma	icdot-C71.0 Cerebrum icdom-94403 Glioblastoma NCIT:C3058 Glioblastoma	PMID:23079654	0.079	0.17	0.249
pgxbs-kftvgk90	Glioblastoma	icdot-C71.0 Cerebrum icdom-94403 Glioblastoma NCIT:C3058 Glioblastoma	PMID:23079654	0.162	0.128	0.29
pgxbs-kftvgka5	Glioblastoma	icdot-C71.9 brain, NOS icdom-94403 Glioblastoma NCIT:C3058 Glioblastoma	PMID:23079654	0.09	0.058	0.148
pgxbs-kftvgkae	Glioblastoma	icdot-C71.9 brain, NOS icdom-94403 Glioblastoma NCIT:C3058 Glioblastoma	PMID:23079654	0.076	0.128	0.204
pgxbs-kftvgkaf	Glioblastoma	icdot-C71.9 brain, NOS icdom-94403 Glioblastoma NCIT:C3058 Glioblastoma	PMID:23079654	0.004	0.018	0.021

Page **1 of 134**

Sample search by CNV
Biosamples Tab

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Search Samples

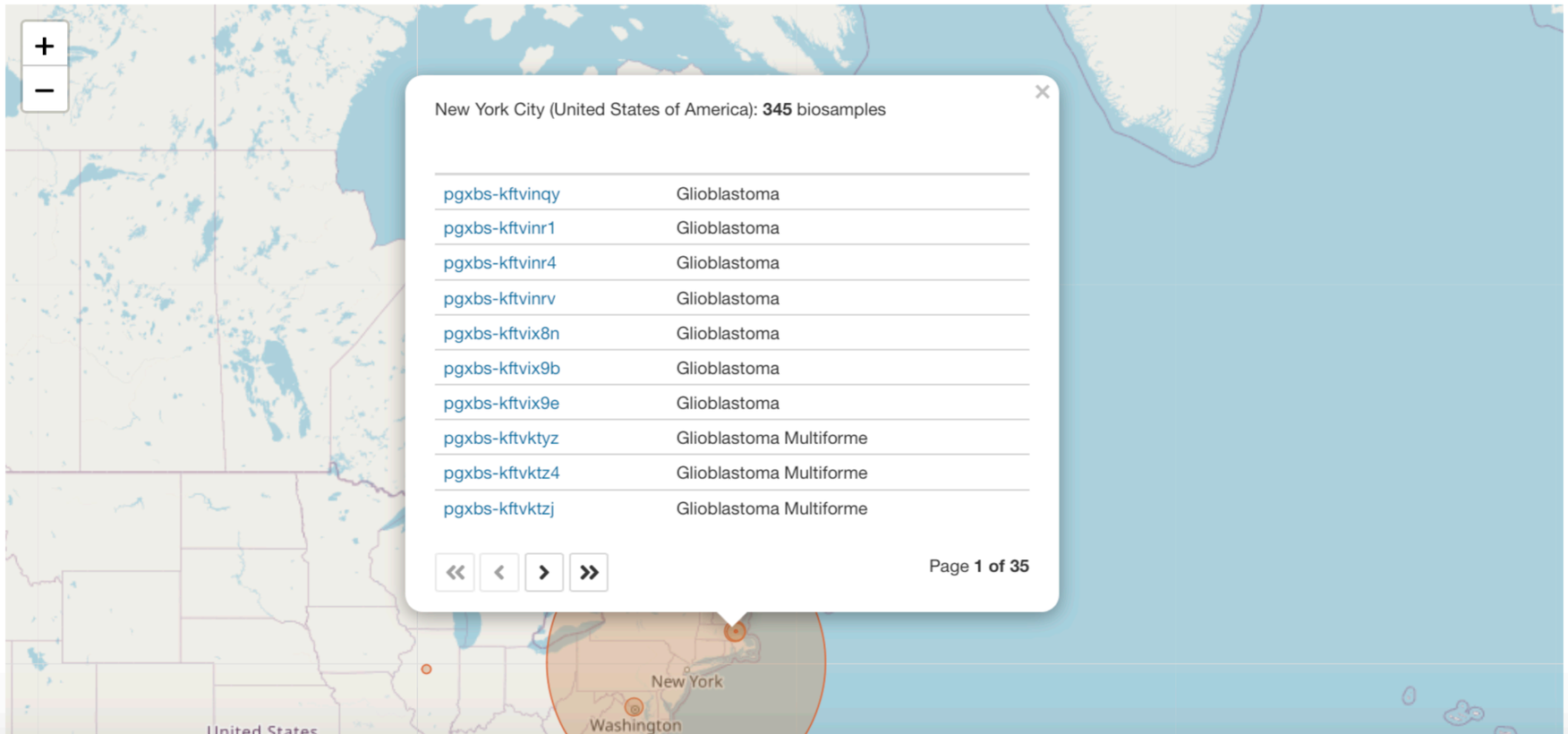
Assembly: GRCh38 **Chro:** 9 **Start:** 21500001-21975098 **End:** 21967753-22500000 **Type:** DEL **Filters:** NCIT:C3058

progenetix

Samples: 668
Variants: 286
Calls: 675
$f_{alleles}$: 0.000088

Phenopackets
Callsets Variants
Variants in UCSC
UCSC region
JSON Response
Visualization options

Results | Biosamples | Biosamples Map | Variants



Sample search by CNV Map Tab

Sample search by CNV
Variants Tab

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Search Samples

Assembly: GRCh38 **Chro:** 9 **Start:** 21500001-21975098 **End:** 21967753-22500000 **Type:** DEL **Filters:** NCIT:C3058

progenetix

Samples: 668
Variants: 286
Calls: 675
$f_{alleles}$: 0.000088

Phenopackets
Callsets Variants
Variants in UCSC
UCSC region
JSON Response

Visualization options

Results | Biosamples | Biosamples Map | Variants

JSON | Download Response

Int. ID	Digest	Callset	Biosample	Chr.	Ref. Base(s)	Alt. Base(s)	Type
5bab578b727983b2e00ca99e	9:21548871-21999595:DEL	pgxcs-kftvmlzx	pgxbs-kftvgk8h	9			DEL
5bab578d727983b2e00cb505	9:21958233-21999595:DEL	pgxcs-kftvmm5j	pgxbs-kftvgk90	9			DEL
5bab5793727983b2e00cdc18	9:21958233-21999595:DEL	pgxcs-kftvmmjj	pgxbs-kftvgka5	9			DEL
5bab5794727983b2e00ce2c6	9:21791897-21999595:DEL	pgxcs-kftvmmlu	pgxbs-kftvgkae	9			DEL
5bab5794727983b2e00ce49a	9:21958233-21999595:DEL	pgxcs-kftvmmmb	pgxbs-kftvgkaf	9			DEL

Page **1 of 135**

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Progenetix Publication Collection

The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in Progenetix and/or arraymap (array source files).

Please contact us to alert us about additional articles you are aware of.

Search: melanoma

City: Zurich (Switzerland)

Range (km): 500

Publications (13)		Samples					
id	Publication	cCGH	aCGH	WES	WGS	pgx	am
PMID:27261508	Zhao F, Sucker A, Horn S, Heeke C, Bielefeld N, Schrörs et al. (2016): Melanoma Lesions Independently Acquire T-cell Resistance during Metastatic Latency. Cancer Res. 76(15), 2016	0	5	0	0	5	5
PMID:8033101	Speicher MR, Prescher G, du Manoir S, Jauch A, Horsthemke et al. (1994): Chromosomal gains and losses in uveal melanomas detected by comparative genomic hybridization. Cancer Res. 54(14), 1994	11	0	0	0	0	0
PMID:23633454	Griewank KG, Westekemper H, Murali R, Mach M, Schilling et al. (2013): Conjunctival melanomas harbor BRAF and NRAS mutations and copy number changes similar to cutaneous and mucosal ... Clin. Cancer Res. 19(12), 2013	0	30	0	0	0	0

Cancer publications collection

Search by keywords, approximate location

Publications with sample&technology count and if present in progenetix

Internal link to summary information and external link to Pubmed

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Services: Ontologymaps

The **ontologymaps** service provides equivalency mapping between ICD-O and other classification systems, notably NCIt. It makes use of the sample-level mappings for NCIT and ICD-O 3 codes developed for the individual samples in the Progenetix collection.

While NCIT treats diseases as **histologic** and **topographic** described entities (e.g. **NCIT:C7700**: **Ovarian adenocarcinoma**), these two components are represented separately in ICD-O, through the **Morphology** and **Topography** coding arms (e.g. here **8140/3** + **C56.9**).

More documentation with focus on the API functionality can be found on the documentation pages.

Code Selection

NCIT:C4349: Colon Adenocarcinoma

Optional: Limit with second selection

Matching Code Mappings {JSON↗}

NCIT:C4349: Colon Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C18.9: Colon, NOS
NCIT:C4349: Colon Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C18.0: Cecum
NCIT:C4349: Colon Adenocarcinoma	icdom-81403: Adenocarcinoma, NOS	icdot-C18.2: Ascending colon

Ontology mapping service

Currently supports mutual mapping between

NCIt and ICD-O M+T pair

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Data visualization Upload

Drag and drop some files here, or click to select files.

File format

Data has to be submitted as tab-delimited **.tsv** segment files. An example file is being provided here.

While the header values are not being interpreted, the column order has to be followed:

1. `sample`
   - please use only word characters, underscores, dashes
   - the `sample` value is used for splitting multi-sample files into their individual profiles
2. `chro`
   - the reference chromosome
   - 1-22, X, Y (23 => X; 24 => Y)
3. `start`
   - base positions according to the used reference genome
4. `end`

User data upload

For customised visualisation for single sample or aggregated summary plots

Thank you!

Any questions?

Acknolwdgement:

Paula Carrio Cordo
Bo Gao
Rahel Paloots
Pierre-Henri Toussai
Prof. Michael Baudis