Genomic Copy Number Variation Resources Powered by Genomic Beacons

Michael Baudis
Professor of Bioinformatics
University of Zürich
Swiss Institute of Bioinformatics **SIB**
Member GA4GH Strategic Leadership Committee
GA4GH Workstream Co-lead *DISCOVERY*
Co-lead ELIXIR Beacon API Development
Co-lead ELIXIR hCNV Community

Universität Zürich[UZH]

SIB Swiss Institute of Bioinformatics

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

elixir

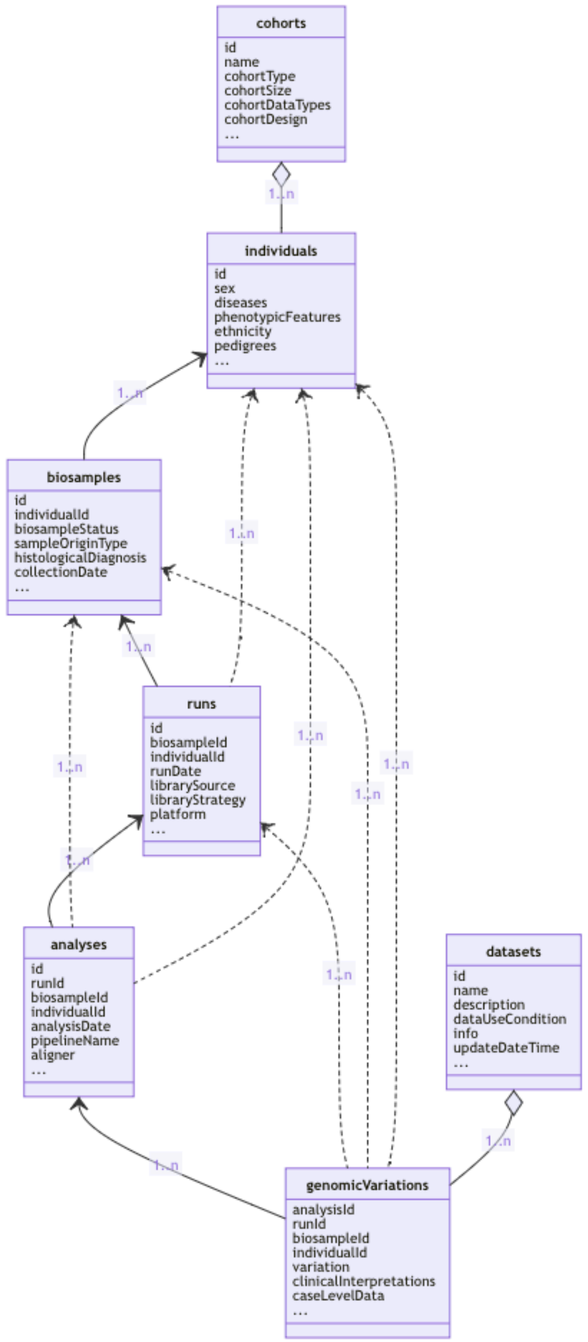
A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

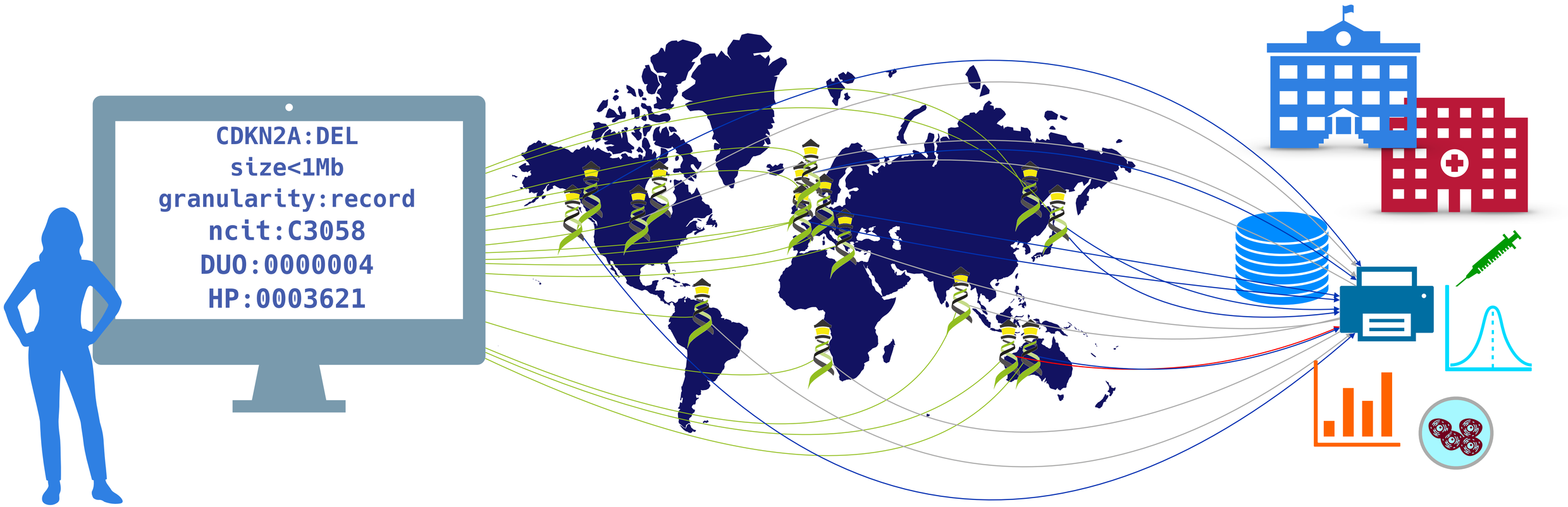
YES | **NO** | **\0**

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"		
2015	• beacon-network.org aggregator created by DNAstack		• ELIXIR starts Beacon project support
2016	• Beacon v0.3 release • work on queries for structural variants (brackets for fuzzy start and end parameters...)	• Beacon+ concept implemented @ **progenetix.org** • concepts from GA4GH Metadata (ontologies…) • entity-scoped query parameters ("individual.age")	
2017	• OpenAPI implementation • integrating **CNV parameters** (e.g. "startMin, statMax")	• Beacon+ demos "**handover**" concept	• GA4GH re-structuring (workstreams...) • Beacon part of Discovery WS
2018	• Beacon v0.4 release in January; feature release for GA4GH approval process • **GA4GH Beacon v1 approved** at Oct plenary	• Beacon hackathon Stockholm; settling on **filters** • Barcelona goes Zurich developers meeting • Beacon API v2 Kick off • adopting "handover" concept	• new Beacon website (March)
2019	• ELIXIR Beacon Network		• Beacon publication at Nature Biotechnology
2020		• "**Scouts**" teams working on different aspects - filters, genomic variants, compliance ...	
2021		• **framework** + **models** concept implemented • range and bracket queries, variant length • starting of GA4GH review process	• Phenopackets v2 approved
2022		• changes in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • **Beacon v2 approved** at April GA4GH Connect	• *docs.genomebeacons.org*

Beacon Default v2 Model

- The Beacon ***framework*** describes the overall structure of the API requests, responses, parameters, the common components, etc.
- Beacon ***models*** describe the set of concepts included in a Beacon, like individual or biosample, and also the relationships between them.
- Besides logical concepts, the Beacon ***models*** represent the schemas for data delivery in “record” granularity
- Beacon explicitly allows the use of *other models* besides its *version specific default*.
- Adherence to a shared ***model*** empowers federation
- Use of the ***framework*** w/ different models extends adoption





Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

Beacon API

The Beacon API v2 represents a simple but powerful **genomics API** for ***federated*** data discovery and retrieval

progenetix.org

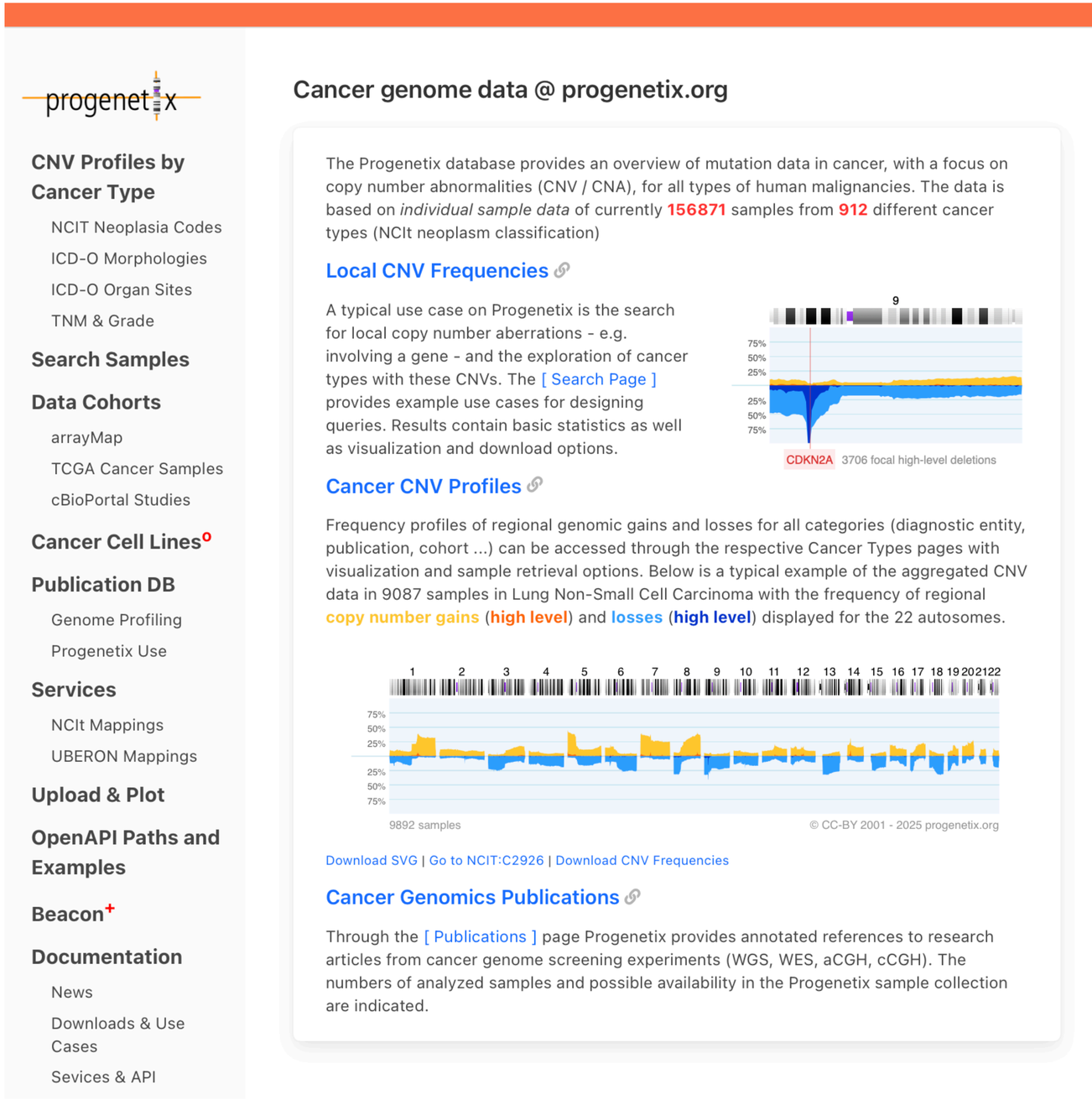
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **150'000 cancer CNV profiles**
- more than 9**00 diagnostic types**
- runs on a **Beacon API**
- inclusion of reference datasets (e.g. TCGA)
- support for SNV data (TCGA, cell lines...)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services

Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics



Cancer Cell Lines

&&

refCNV

cancercelllines.org

refcnv.org

Cancer Cell Lines by Cellosaurus ID

The cancer cell lines in *cancercelllines.org* are labeled by th
hierarchially: Daughter cell lines are displayed below the pri
as a daughter cell line of **HeLa (CVCL_0030)** and so forth.

Sample selection follows a hierarchical system in which sam
response. This means that one can retrieve all instances an
by default - but

refCNV

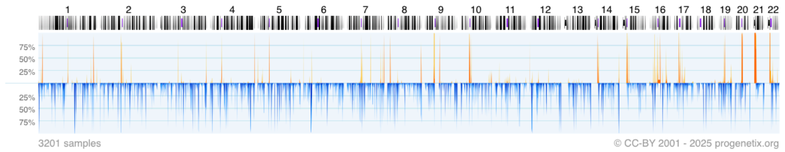
- CNV Profiles by Platform
- CNV Profiles by Analysis Pipeline
- Search Samples
- Beacon+
- Documentation
- Baudisgroup @ UZH

Genomic Copy Number Variation (CNV) data from reference samples

Under Construction

This site is currently under construction, with contributions by groups from the University of Zurich and Erasmus MC. Neither data content nor representation have been finalized. PLEASE DO NOT USE FOR ANY RESEARCH OR REFERENCE PURPOSES!

Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the respective Cancer Types pages with visualization and sample retrieval options. Below is a typical example of the aggregated CNV data in 3201 samples of the 1000 Genomes Dragen CNV analysis set. The frequency of regional **copy number gains** (**high level**) and **losses** (**high level**) displayed for the 22 autosomes as occurrence of any of these CNVs in the 1Mb binned intervals.



Download SVG | Go to DRAGEN-CNV | Download CNV Frequencies

The repository contains CNV tracks for many of the 1000 Genomes samples, analyzed by different platforms or data pipelines and therefore allows to compare private analysis data to results from these different call sets, to avoid interpretation biases from using reference data with a different analysis profile from the one used in your study. The plot below shows analysis specific CNV tracks for chromosome 13 in the HG01572 sample from the 1000 Genomes set, for several calling pipelines.

Please be aware that the small size of most CNVs is not correctly represented at this zoom level (overplotting due to limited resolution).

Assembly: GRCh38 **Chro:** NC_000007.14 **Start:** 140713328 **End:** 140924929
Type: SNV

cellz

Matched Samples: 1058
Retrieved Samples: 1000
Variants: 127
Calls: 1444

UCSC region
Variants in UCSC
Dataset Responses (JSON)

Visualization options

Results | Biosamples | Variants | Annotated Variants

Digest	Gene	Pathogenicity	Variant type	Variant Instances
7:140834768-140834769:G>A	BRAF		Missense variant	V: pgxvar-63ce6abca24c83054b B: pgxbs-3DfBeeAC
7:140734714-140734715:G>A	BRAF		Missense variant	V: pgxvar-63ce6acda24c83054b B: pgxbs-3fB2a14B
7:140753334-140753339:T>TGTA	BRAF	Pathogenic		V: pgxvar-

Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

- Cellosaurus - a knowledge resource on cell lines cellosaurus:CVCL_0312

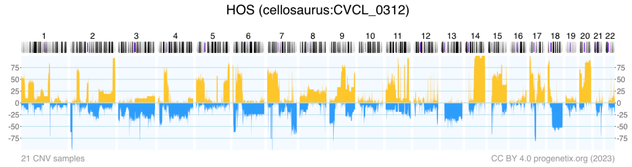
Sample Counts

- 204 samples
- 57 direct *cellosaurus:CVCL_0312* code matches
- 21 CNV analyses

Search Samples

Select *cellosaurus:CVCL_0312* samples in the Search Form

Raw Data (click to show/hide)



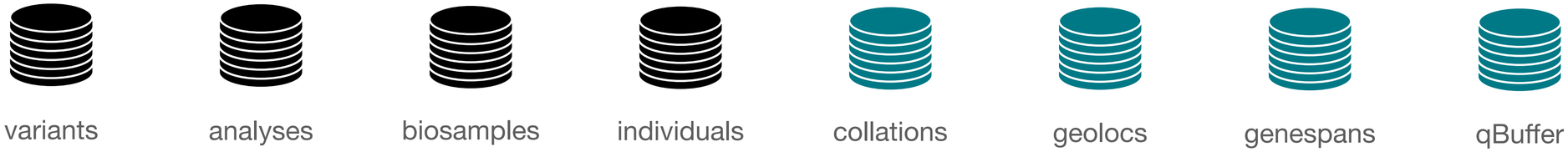
Download SVG | Go to cellosaurus:CVCL_0312 | Download CNV Frequencies

Gene Matches | Cytoband Matches | Variants

ALK	. ABC-14 cells harbored no **ALK** mutations and were sensitive to ... crizotinib while also exhibiting MNNG **HOS** transforming gene (MET)	Rapid Acquisition of Alectinib Resistance in ALK-Positive Lung Cancer With High Tumor Mutation Burden (31374369)	ABSTRACT
AREG	crizotinib while also exhibiting MNNG **HOS**	Rapid Acquisition of Alectinib Resistance	ABSTRACT

bycon based Beacon+ Stack

- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - pubmed:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703…
  - precomputed frequencies per collection informative e.g. in form autofills
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding **accessid** for **handover** generation
- complete query aggregation; i.e. individual queries are run against the corresponding entities and ids are intersected
  - ➡ retrieval of any entity, e.g. all individuals which have queried variants analyzed on a given platform
  - ➡ allows multi-variant queries, i.e. all bio samples or individuals which had matches of all of the individual variant queries



Entity collections **Utility collections**

github.com/progenetix/bycon GitHub

Beacon+: Phenopackets

Testing alternative response schemas...

https://progenetix.org/beacon/phenopackets/pgxind-kftx26j0

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon+ this is done through *ad hoc* handover URIs


```
"id": "pgxpxf-kftx3tl5",
"metaData": {
  "phenopacketSchemaVersion": "v2",
  "resources": [
    {
      "id": "NCIT",
      "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
      "name": "NCIt Plus Neoplasm Core",
      "namespacePrefix": "NCIT",
      "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.c
      "version": "2022-04-01"
    },

"subject": {
  "dataUseConditions": {
    "id": "DUO:0000004",
    "label": "no restriction"
  },
  "diseases": [
    {
      "clinicalTnmFinding": [],
      "diseaseCode": {
        "id": "NCIT:C3099",
        "label": "Hepatocellular Carcinoma"
      },
      "onset": {
        "age": "P48Y9M26D"
      },
      "stage": {
        "id": "NCIT:C27966",
        "label": "Stage I"
      }
    }
  ],
  "id": "pgxind-kftx3tl5",
  "sex": {
    "id": "PATO:0020001",
    "label": "male genotypic sex"
  },
  "updated": "2018-12-04 14:53:11.674000",
  "vitalStatus": {
    "status": "UNKNOWN_STATUS"
  }
}

"biosamples": [
  {
    "biosampleStatus": {
      "id": "EFO:0009656",
      "label": "neoplastic sample"
    },
    "dataUseConditions": {
      "id": "DUO:0000004",
      "label": "no restriction"
    },
    "description": "Primary Tumor",
    "externalReferences": [
      {
        "id": "pgx:TCGA-0004d251-3f70-4395-b175-c94c2f5b1b81",
        "label": "TCGA case_id"
      },
      {
        "id": "pgx:TCGA-TCGA-DD-AAVP",
        "label": "TCGA submitter_id"
      },
      {
        "id": "pgx:TCGA-9259e9ee-7279-4b62-8512-509cb705029c",
        "label": "TCGA sample_id"
      },
      {
        "id": "pgx:TCGA-LIHC",
        "label": "TCGA LIHC project"
      }
    ],
    "files": [
      {
        "fileAttributes": {
          "fileFormat": "pgxseg",
          "genomeAssembly": "GRCh38"
        },
        "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
      }
    ],
    "histologicalDiagnosis": {
      "id": "NCIT:C3099",
      "label": "Hepatocellular Carcinoma"
    },
    "id": "pgxbs-kftvhyvb",
    "individualId": "pgxind-kftx3tl5",
    "pathologicalStage": {
      "id": "NCIT:C27966",
      "label": "Stage I"
    },
    "sampledTissue": {
      "id": "UBERON:0002107",
      "label": "liver"
    },
    "timeOfCollection": {
      "age": "P48Y9M26D"
    },
```

pgxRpi: an R/Bioconductor package

Client for Accessing Beaconized Data

pgxRpi

This is the **development** version of pgxRpi; for the stable release version, see pgxRpi.

R wrapper for Progenetix

platforms all | rank 2178 / 2266 | support 0 / 0 | in Bioc < 6 months | build unknown | updated < 1 month | dependencies 137

DOI: 10.18129/B9.bioc.pgxRpi

Bioconductor version: Development (3.20)

The package is an R wrapper for Progenetix REST API built upon the Beacon v2 protocol. Its purpose is to provide a seamless way for retrieving genomic data from Progenetix database—an open resource dedicated to curated oncogenomic profiles. Empowered by this package, users can effortlessly access and visualize data from Progenetix.

Author: Hangjia Zhao [aut, cre], Michael Baudis [aut]

Maintainer: Hangjia Zhao <hangjia.zhao at uzh.ch>

- **Query and export variants**

https://progenetix.org/beacon/biosamples/pgxbs-kftvh94d/g_variants

```
> variants <- pgxLoader(type="variant",biosample_id="pgxbs-kftvh94d")
```

- **Query metadata of biosamples and individuals by filters (e.g. NCIt, PMID)**

http://progenetix.org/services/sampletable/?filters=NCIT:C3697

```
> biosamples <- pgxLoader(type="biosample",filters="NCIT:C3697")
```

- **Query and visualize CNV frequency by filters**

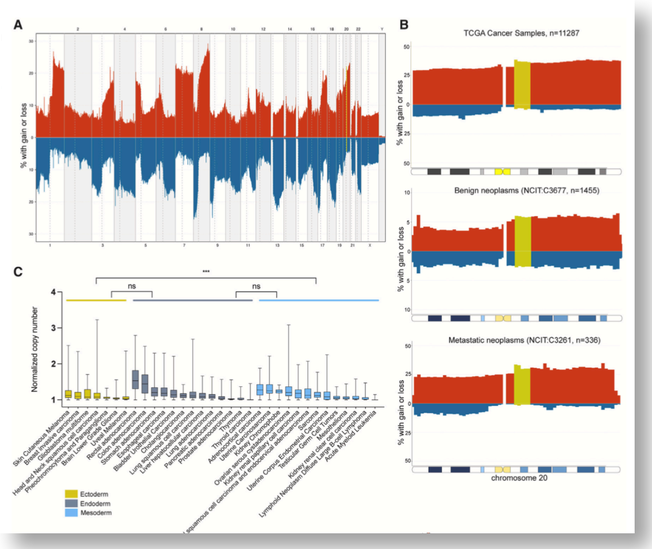
http://www.progenetix.org/services/intervalFrequencies/?filters=NCIT:C3512

```
> freq <- pgxLoader(type="frequency",output="pgxfreq",filter
> pgxFreqplot(freq)
```

- **Process local .pgxseg files**

```
> info <- pgxSegprocess(file=file, show_KM_plot = T,
return_seg = T,  return_metadata = T, return_frequency = T)
```

Use case: 2024 article using Progenetix' ***pgxRpi*** to retrieve & visualize 117'587 cancer CNV profiles for a study into pluripotent stem cells' genomics



Stem Cell Reports
Review
ISSCR
OPEN ACCESS

Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research

Nuša Krivec,[1,2] Manjusha S. Ghosh,[1,2] and Claudia Spits[1,2,*]
[1]Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium
[2]These authors contributed equally
*Correspondence: claudia.spits@vub.be
https://doi.org/10.1016/j.stemcr.2023.11.013

Figure 2. Copy-number alterations of human chromosome 20q11.21 in cancers
(A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079-35871578 is marked in moss green. NCIT, National Cancer Institute Thesaurus.
(B) Top to bottom: Aggregated CNV data of 11,287 TCGA cancer samples, 336 metastatic neoplasms (NCIT: C3261), and 1,455 benign neoplasms (NCIT: C3677) from the Progenetix database (Huang et al., 2021), respectively, were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green.

Changes & To Do

Changes Tracker

While changes are documented for individual point versions we actually do not push releases out for all of them; they serve more as internal development milestones.

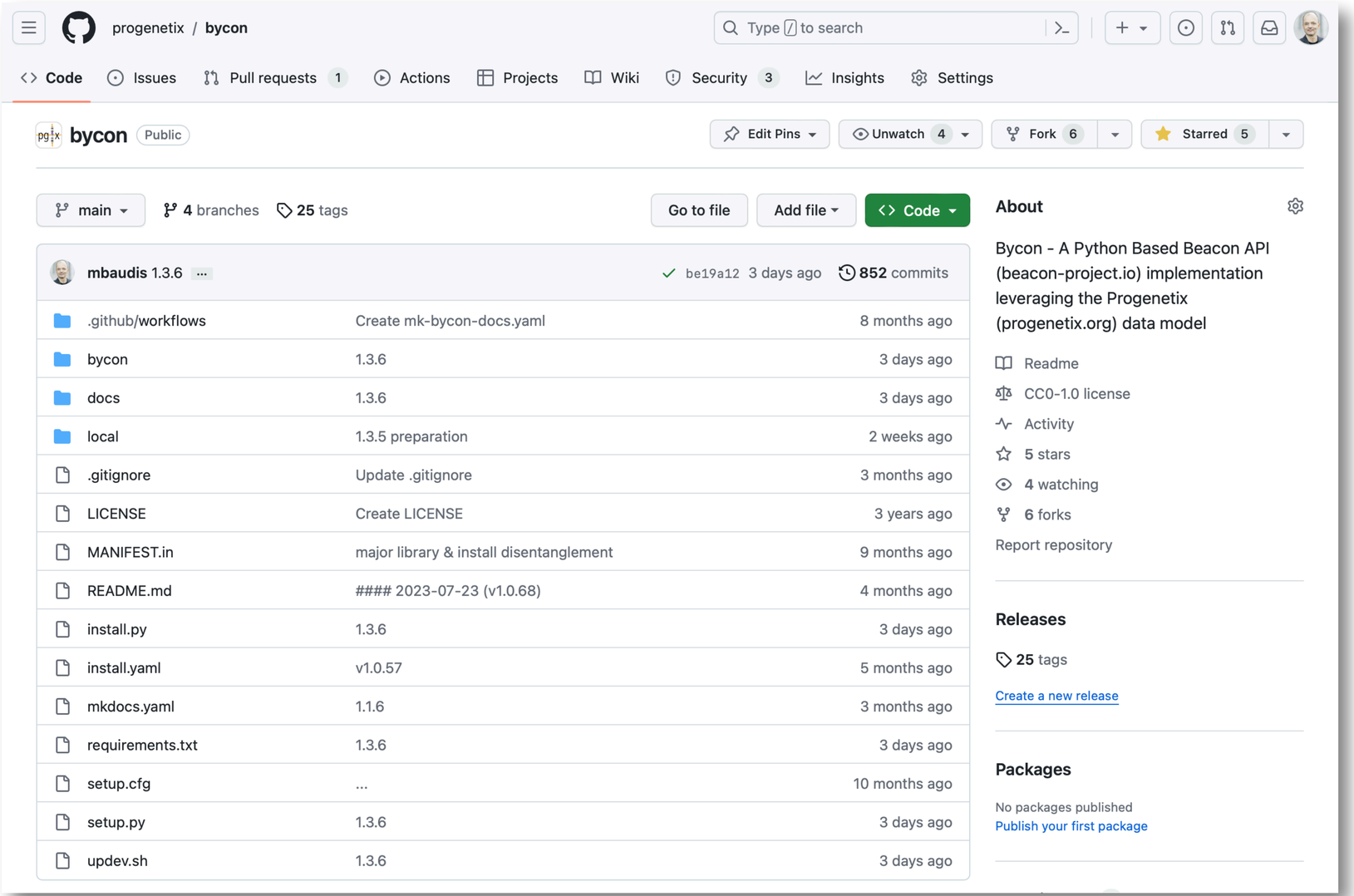
2025-05-15: (v2.4.3 "Bologna")

- expanded `NCITsex` ontology to have hierarchical terms with the current NCIT terms at the tip of the branches
  - e.g. `pgx:sex` => `pgx:sex-female` => `PATO:0020001` => `NCIT:C16576`
  - allows for query expansion & use of alternate terms (e.g. PATO)
  - not strictly correct since the NCIT terms are for "any description of biological sex or gender", wherease PATO is for genotypic sex; so might be flipped later w/ annotations in the databas switched accordingly (this was the orriginal state but Beacon docs used NCIT ...)
- changed `byconautServiceResponse` to `byconServiceResponse`
- added a new subset / cancer type histogram multi-selection to the `beaconplusWeb` front-end (at beaconplus.progenetix.org/subsetsSearch/)

Table of contents

Looking for implementers and contributors

- containerization
- data I/O ...
- standard library integration (VRSification of variants...)



bycon.progenetix.org
github.com/progenetix/bycon/

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

What Can You Do?

- find a way to make your (patients') **data discoverable** - through adding *at least* the relevant metadata to national or project centric repositories
- use forward looking consent and data protection models (**ORD** principle "*as secure as necessary, as open as possible*")
- **support** and/or get involved with international **data standards** efforts and projects
- **... talk to us**

bycon.progenetix.org
github.com/progenetix/bycon/