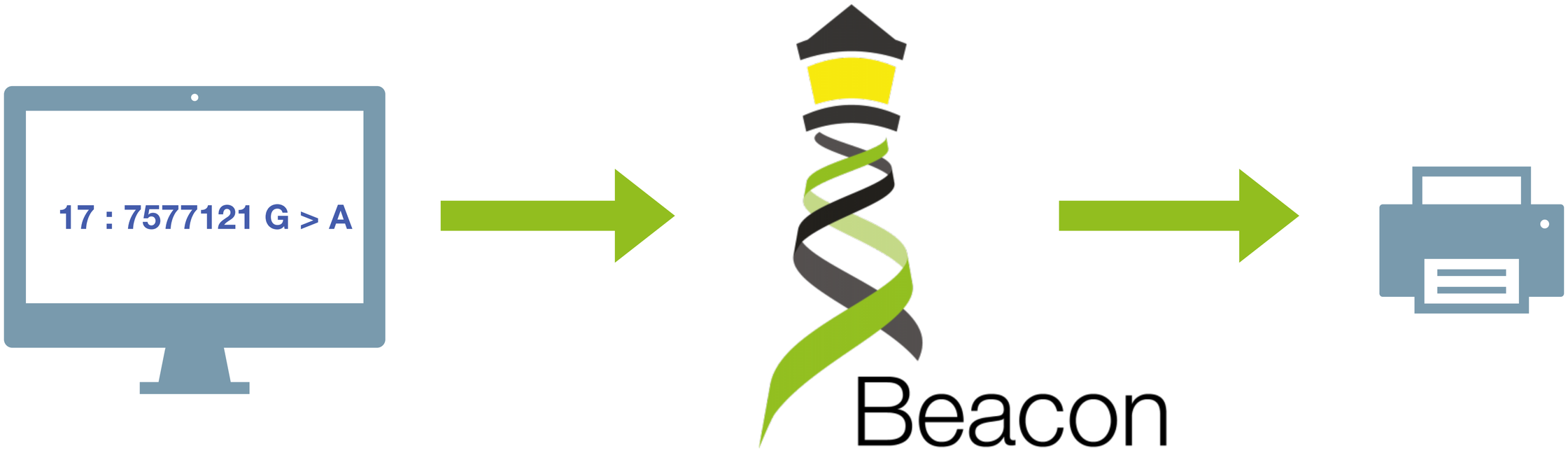
ELIXIR ALL HANDS 2021

hCNV data and the Progenetix Beacon

Implementing a cancer hCNV reference resource on top of Beacon v2

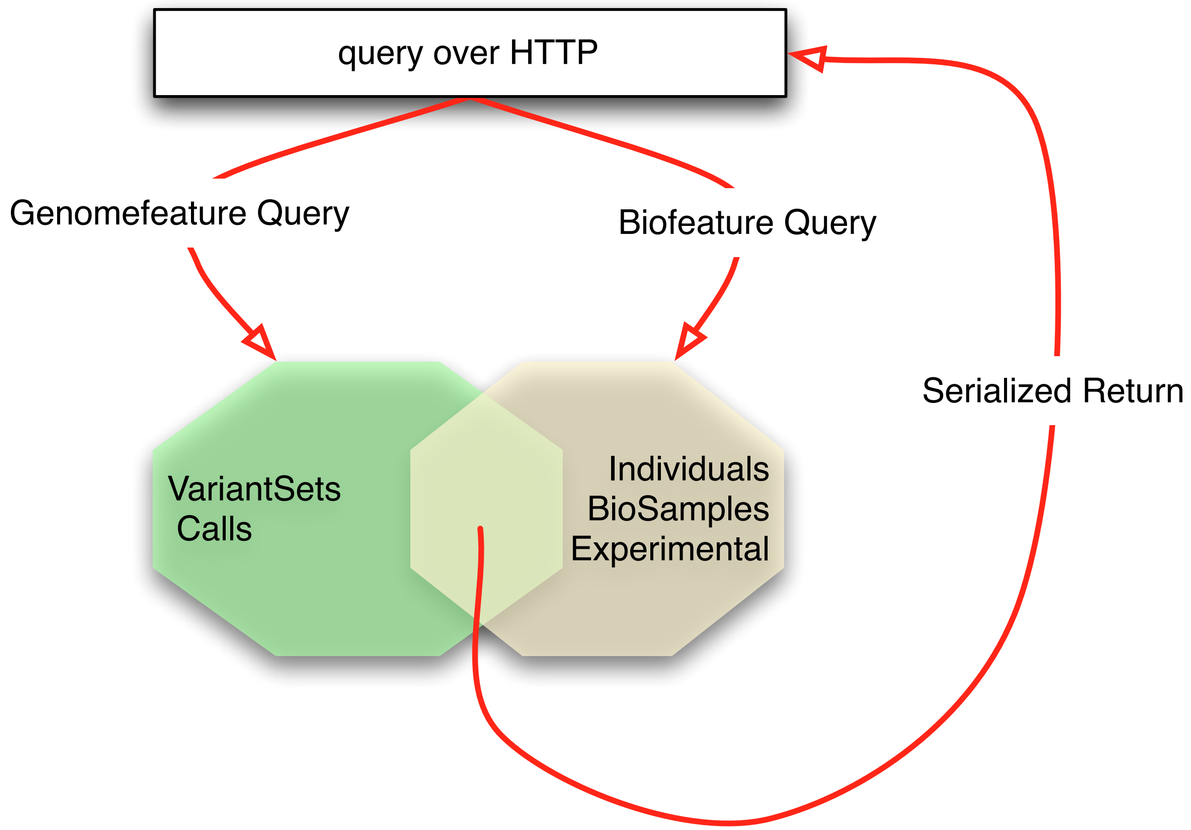
Michael Baudis | Virtual Prague | 2021-06-08



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

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

Minimal GA4GH query API structure



Beacon(+) “Metadata”

Beacon+ Concept

- “quantitative Beacon” with metadata in query and return; especially needed for cancer genomes
- extending GA4GH use cases beyond SNP calls
- server implementation using GA4GH query methods
- needed: structural variants, metadata, context representation …
- test implementation for **cancer** CNA data from arrayMap will be supported through ***SIB***

  => local copy number frequency reference in cancer (+ background CNVs)

arrayMap EMBL-EBI Swiss Institute of Bioinformatics Global Alliance for Genomics & Health

Progenetix in 2021

Cross-platform Oncogenomics

- source data (i.e. array probe data access) and annotation derived (aCGH, WGS, WES, other arrays)
- >130'000 cancer and reference CNA profiles
- systematic metadata annotations following GA4GH standards
- unrestricted access w/o registration
- data access API
- online visualization
- CNA statistics

progenetix

University of Zurich[UZH]

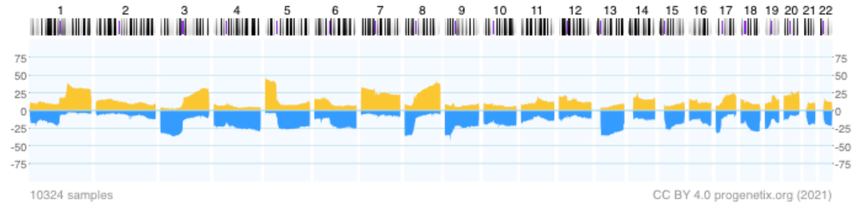
progenetix

- **Cancer CNV Profiles**
- **Search Samples**
- **Studies & Cohorts**
  - arrayMap
  - TCGA Samples
  - DIPG Samples
  - Gao & Baudis, 2021
  - Cancer Cell Lines
- **Publication DB**
- **Services**
  - NCIt Mappings
  - UBERON Mappings
- **Upload & Plot**
- **Beacon+**
- **Progenetix Info**
  - About Progenetix
  - Use Cases
  - Documentation
  - 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. The data is based on *individual sample data* from currently **139448** samples.

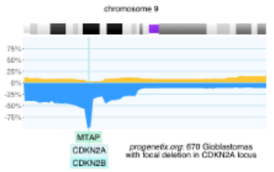
Malignant Thoracic Neoplasm (NCIT:C3576)



Download SVG | Go to NCIT:C3576 | Download CNV Frequencies

Example for aggregated CNV data in 10324 samples in Malignant Thoracic Neoplasm.
Here the frequency of regional copy number gains and losses are displayed for all 22 autosomes.

Progenetix Use Cases



Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [Search Page] provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.

Cancer CNV Profiles

The progenetix resource contains data of **788** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [Cancer Types] page with direct visualization and options for sample retrieval and plotting options.

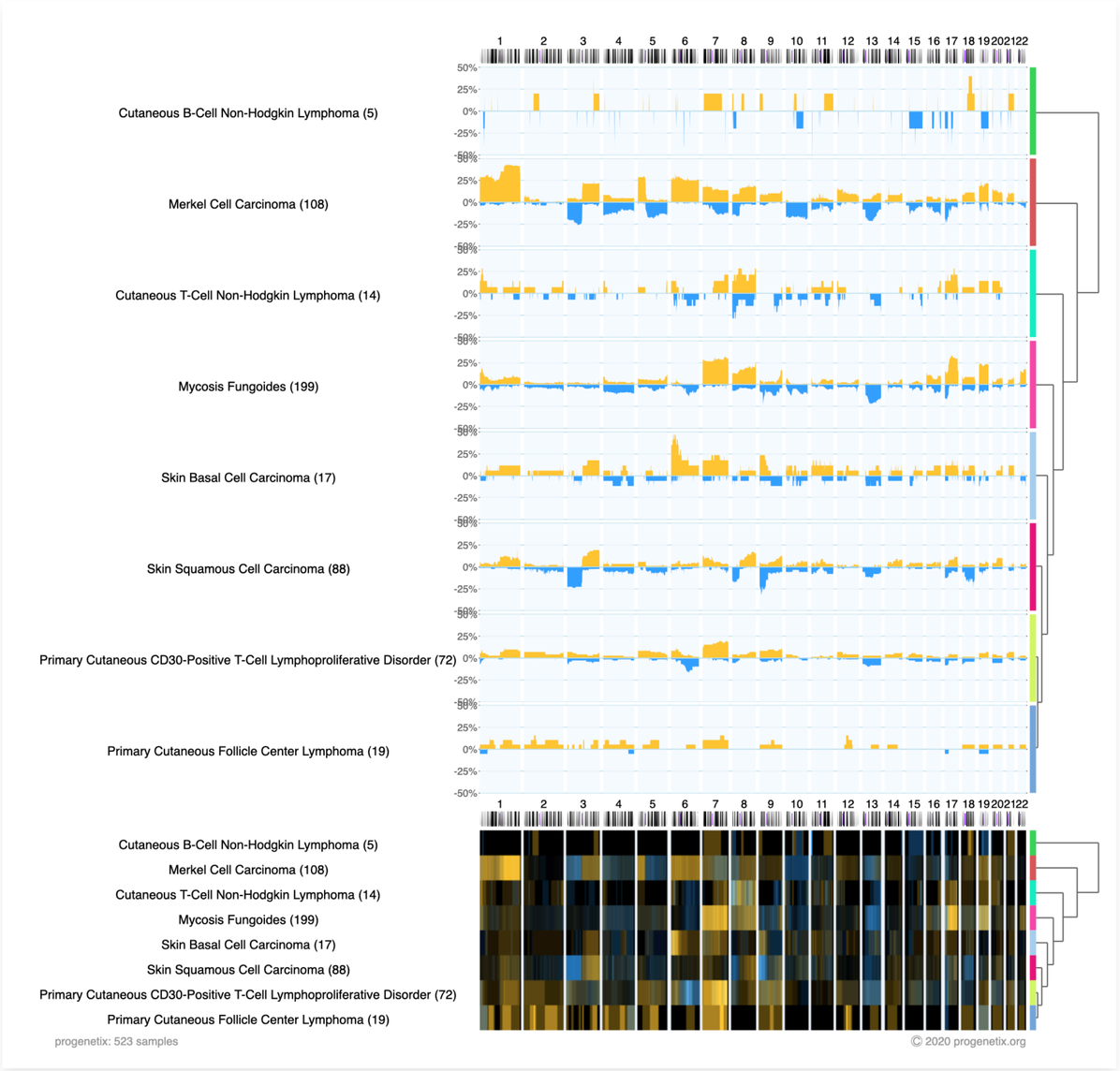
Cancer Genomics Publications

Through the [Publications] page Progenetix provides **4025** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

Beacon+ by Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

- Since 2016 the Progenetix resource has been used to model options for Beacon development
  - 138334 individual samples from 698 cancer types
- The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/clinically scoped queries
- Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services:
  - downloads
  - visualization
  - use of external services (UCSC browser display...)

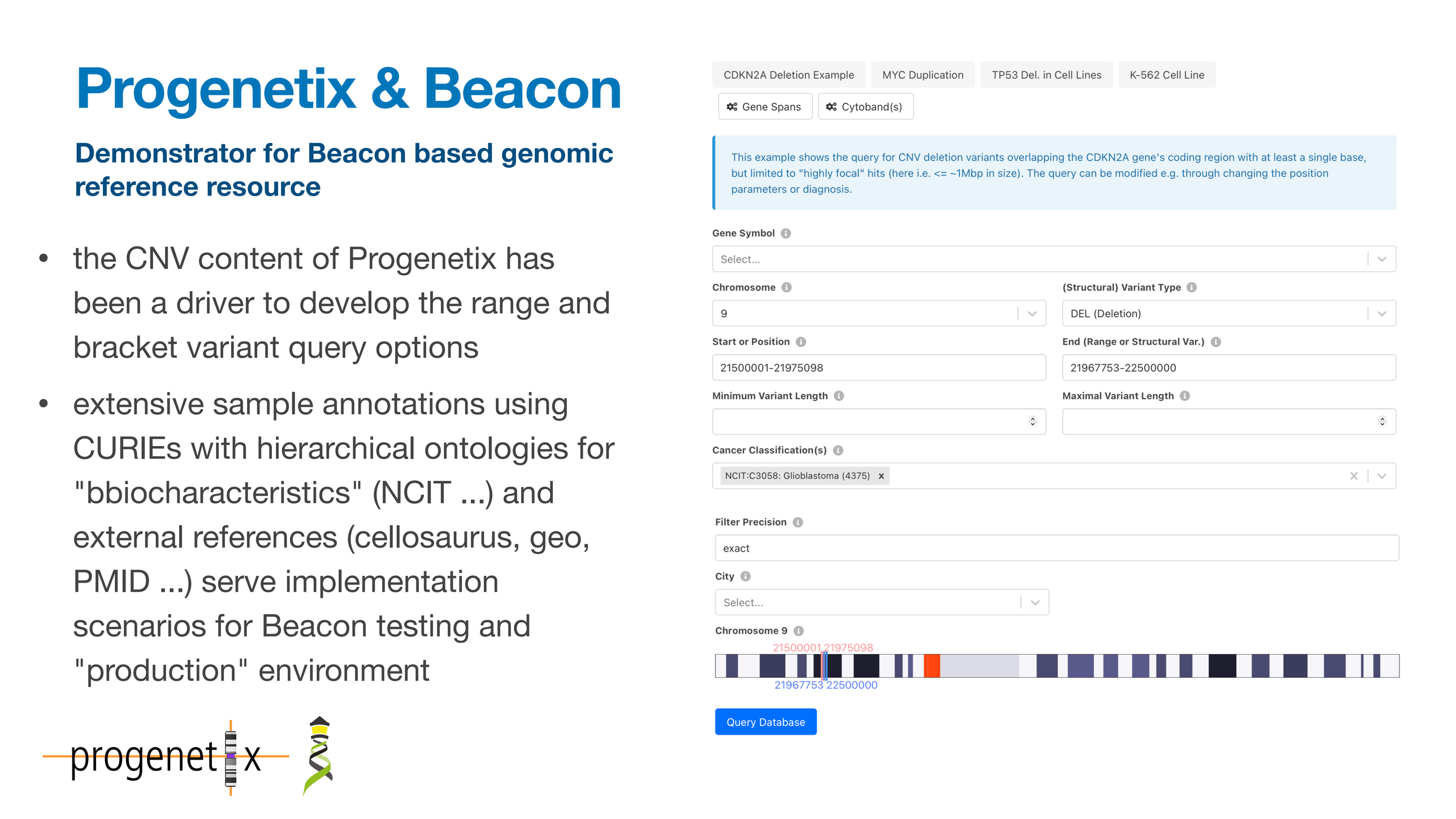


Progenetix & Beacon

Demonstrator for Beacon based genomic reference resource

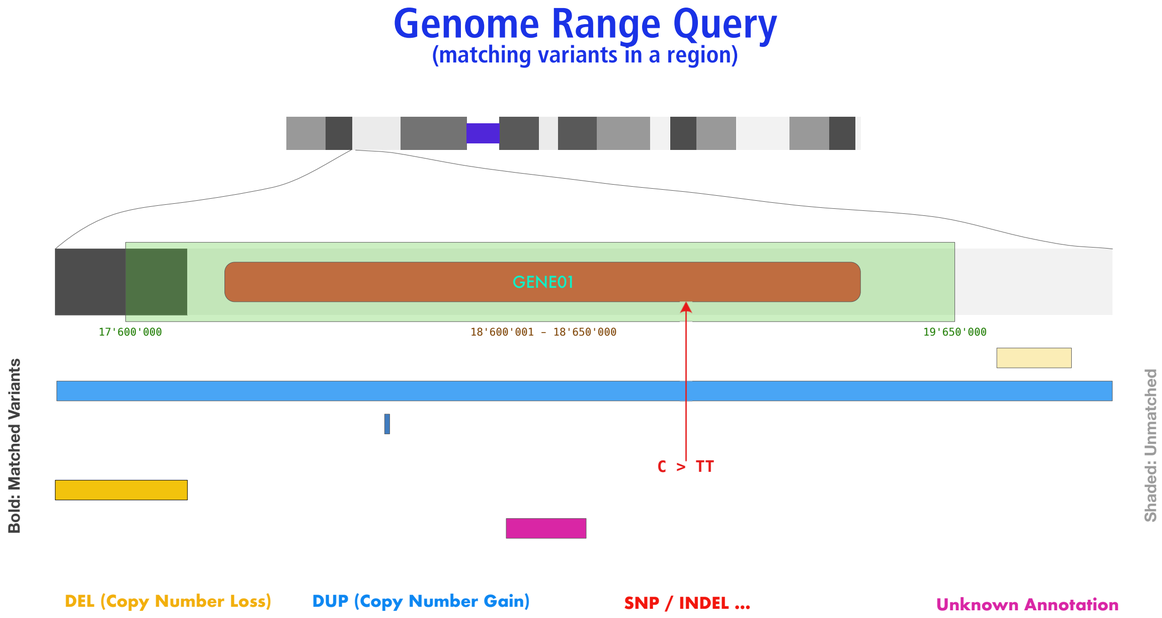
- the CNV content of Progenetix has been a driver to develop the range and bracket variant query options
- extensive sample annotations using CURIEs with hierarchical ontologies for "bbiocharacteristics" (NCIT ...) and external references (cellosaurus, geo, PMID ...) serve implementation scenarios for Beacon testing and "production" environment

progenetix

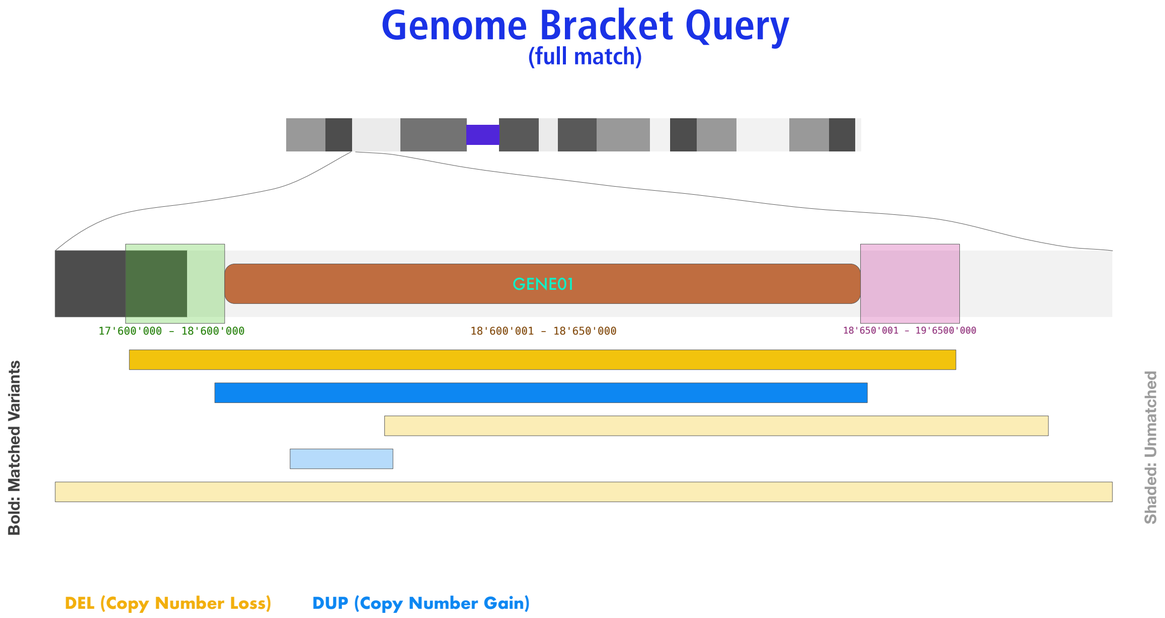


Beacon v2: Extended Variant Queries

Range and Bracket queries enable positional wildcards and fuzziness



- Genome Range Queries provide a way to "fish" for variants overlapping an indicated region, e.g. the CDR of a gene of interest
- Additional parameters (e.g. variant type, regerence or alternate bases) limit the scope of the responses
- new Beacon v2 size parameters to limit structural variants (e.g. "focal" CNVs)



- Genome Bracket Queries allow to search for structural variants with start and end positions falling into defined sequence ranges
- allows to query any contiguous genomic variant (and in principle also can step in for range queries)
- typical use case is e.g the query for variants such as duplications covering the whole CDR of a gene, while limiting the allowed start or end regions

DX Ontologies

Hierarchical NCIt Neoplasm Core replaces heterogeneous primary annotations

- heterogeneous and inconsistent diagnostic annotations are common in clinical reports and research studies ("text", ICD-10, ICD-O 3, OncoTree, domain-specific classifications)
- highly variable granularity of annotations is a major road block for comparative analyses and large scale data integration
  - "Colorectal Cancer" or "Rectal Mucinous Adenoca."
- initiatives and services such as Phenopackets, MONDO, OXO ... rely on and/or provide mappings to hierarchical ontologies

progenetix

NCIt Neoplasm Core coded display (excerpt) for samples in the Progenetix cancer genome data resource allows sample selection on multiple hierarchy levels →

Collapse all | Expand | All

	Subsets	Samples
☐	NCIT:C3262: Neoplasm	88844
☐	NCIT:C3263: Neoplasm by Site	84747
☐	NCIT:C156482: Genitourinary System Neoplasm	11616
☐	NCIT:C156483: Benign Genitourinary System Neoplasm	219
☐	NCIT:C4893: Benign Urinary System Neoplasm	90
☐	NCIT:C4778: Benign Kidney Neoplasm	90
☐	NCIT:C159209: Kidney Leiomyoma	1
☐	NCIT:C4526: Kidney Oncocytoma	82
☐	NCIT:C8383: Kidney Adenoma	7
☐	NCIT:C7617: Benign Reproductive System Neoplasm	129
☐	NCIT:C4934: Benign Female Reproductive System Neoplasm	129
☐	NCIT:C2895: Benign Ovarian Neoplasm	58
☐	NCIT:C4510: Benign Ovarian Epithelial Tumor	58
☐	NCIT:C40039: Benign Ovarian Mucinous Tumor	58
☐	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
☐	NCIT:C4060: Ovarian Cystadenoma	58
☐	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
☐	NCIT:C3609: Benign Uterine Neoplasm	71
☐	NCIT:C3608: Benign Uterine Corpus Neoplasm	71
☐	NCIT:C3434: Uterine Corpus Leiomyoma	71
☐	NCIT:C156484: Malignant Genitourinary System Neoplasm	11171
☐	NCIT:C157774: Metastatic Malignant Genitourinary System Neoplasm	2
☐	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
☐	NCIT:C8946: Metastatic Prostate Carcinoma	2
☐	NCIT:C164141: Genitourinary System Carcinoma	10561
☐	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
☐	NCIT:C8946: Metastatic Prostate Carcinoma	2
☐	NCIT:C3867: Fallopian Tube Carcinoma	19

Beacon v2 Filters

Example: Use of hierarchical classification systems (here NCIt neoplasm core)

- Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications

  ➡ implicit *OR* with otherwise assumed *AND*

- implementation of hierarchical annotations overcomes some limitatiions of "fuzzy" disease annotations

- data *handover* (Beacon v1.1+) enables further data exploration and export scenarios

Beacon+ specific: Multiple term selection with OR logic

☑	NCIT:C4914: Skin Carcinoma	213
☐	NCIT:C4475: Dermal Neoplasm	109
☑	NCIT:C45240: Cutaneous Hematopoietic and Lymphoid Cell Neoplasm	310

Filters: NCIT:C4914, NCIT:C4819, NCIT:C9231, NCIT:C2921, NCIT:C45240, NCIT:C6858, NCIT:C3467, NCIT:C45340, NCIT:C7195, NCIT:C3246, NCIT:C7217

progenetix

Variants: 0
Calls: 0
Samples: 523

$f_{alleles}$: 0

Callsets Variants | Legacy Interface | UCSC region

Show JSON Response

Results | Biosamples

Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.154	0.056	0.21
PGX_AM_BS_MCC03	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.137	0.21	0.347
PGX_AM_BS_MCC04	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.158	0.056	0.214
PGX_AM_BS_MCC05	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.107	0.327	0.434

Page **1 of 105**

Beacon v2 Paths

Progenetix utilizes Beacon v2 REST paths

- Beacon v2 paths are usedd in the Beacon specification to scope query and delivery
- Progenetix uses a default `/biosamples/` + query path for its front end queries, and then collection specific methods for data retrieval (see next)
- current implementation addresses a core subset of all options, and evaluates some still moving targets
  - ➡ variants_interpretations
  - ➡ variant instances versus prototypes
  - ➡ ...

progenetix

Base `/biosamples`

`/biosamples/` + query

- /biosamples/?filters=cellosaurus:CVCL_0004
  - this example retrieves all biosamples having an annotation for the Cellosaurus *CVCL_0004* identifier (K562)

`/biosamples/{id}/`

- /biosamples/pgxbs-kftva5c9/
  - retrieval of a single biosample

`/biosamples/{id}/variants/` & `/biosamples/{id}/variants_in_sample/`

- /biosamples/pgxbs-kftva5c9/variants/
- /biosamples/pgxbs-kftva5c9/variants_in_sample/
  - retrieval of all variants from a single biosample
  - currently - and especially since for a mostly CNV containing resource - `variants` means "variant instances" (or as in the early v2 draft `variantsInSample`)

Base `/variants`

There is currently (April 2021) still some discussion about the implementation and naming of the different types of genomic variant endpoints. Since the Progenetix collections follow a "variant observations" principle all variant requests are directed against the local `variants` collection.

If using `g_variants` or `variants_in_sample`, those will be treated as aliases.

`/variants/` + query

- /variants/?assemblyId=GRCh38&referenceName=17&variantType=DEL&filterLogic=AND&start=7500000&start=7676592&end=7669607&end=7800000
  - This is an example for a Beacon "Bracket Query" which will return focal deletions in the TP53 locus (by position).

`/variants/{id}/` or `/variants_in_sample/{id}` or `/g_variants/{id}/`

- /variants/5f5a35586b8c1d6d377b77f6/
- /variants_in_sample/5f5a35586b8c1d6d377b77f6/

`/variants/{id}/biosamples/` & `variants_in_sample/{id}/biosamples/`

- /variants/5f5a35586b8c1d6d377b77f6/biosamples/
- /variants_in_sample/5f5a35586b8c1d6d377b77f6/biosamples/

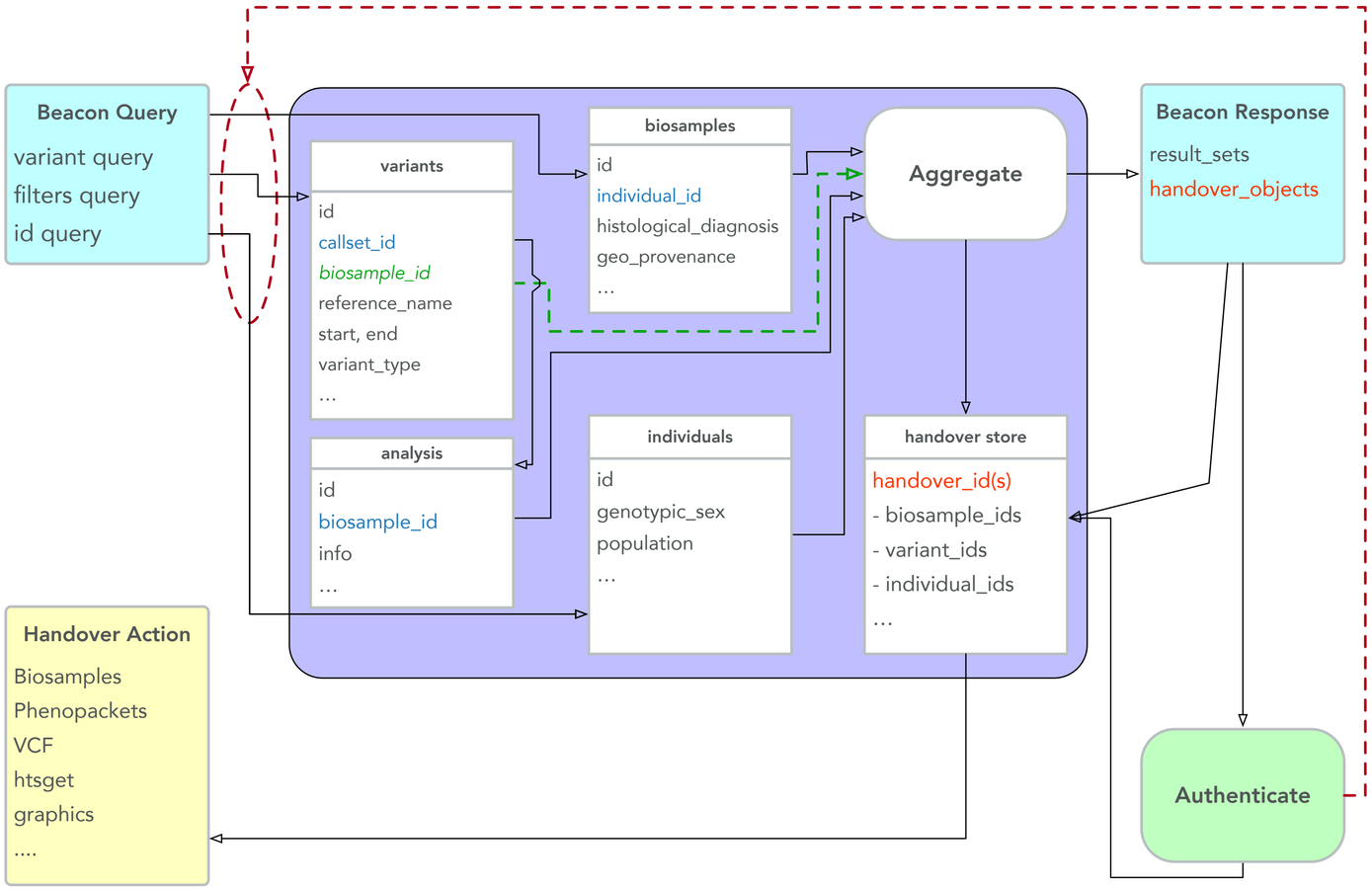
https://github.com/ga4gh-beacon/implementations-v2/

Progenetix & Beacon v1->2

Handover elements in Beacon responses

- Progenetix utilizes handovers to deliver data matched by the Beacon queries
- These handovers are interpreted by the front end to populate different parts of the UI, w/o the need of active selection
- Handovers are either standard Beacon v2 paths or dedicated custom functions

Handover Concept



Progenetix & Beacon v1->2

Handover elements in Beacon responses

- Progenetix utilizes handovers to deliver data matched by the Beacon queries
- These handovers are interpreted by the front end to populate different parts of the UI, w/o the need of active selection
- Handovers are either standard Beacon v2 paths or dedicated custom functions


```
  "results_handovers": [
    {
      "description": "create a CNV histogram from matched callsets",
      "handoverType": {"id": "pgx:handover:cnvhistogram", "label": "CNV
Histogram"},
      "url": "https://progenetix.org/cgi-bin/PGX/cgi/samplePlots.cgi?
method=cnvhistogram&accessid=aff0f73f-6dbf-45e5-91ba-04f19e3621bb"
    },
    {
      "description": "retrieve data of the biosamples matched by the query",
      "handoverType": {"id": "pgx:handover:biosamples", "label":
"Biosamples"},
      "url": "https://progenetix.org/beacon/biosamples/?
accessid=61b68a59-2160-41e4-a17d-0cf128841a57"
    },
    {
      "description": "retrieve variants matched by the query",
      "handoverType": {"id": "pgx:handover:variants", "label": "Found
Variants (.json)"},
      "url": "https://progenetix.org/beacon/variants/?
method=variants&accessid=5cced529-3acf-4156-b121-6ae7e5e63d0c"
    },
    {
      "description": "Download all variants of matched samples - potentially
huge dataset...",
      "handoverType": {"id": "pgx:handover:callsetsvariants", "label": "All
Sample Variants (.json)"},
      "url": "https://progenetix.org/beacon/variants/?
method=callsetsvariants&accessid=61b68a59-2160-41e4-a17d-0cf128841a57"
    },
    {
      "description": "map variants matched by the query to the UCSC
browser",
      "handoverType": {"id": "pgx:handover:bedfile2ucsc", "label": "Show
Variants in UCSC"},
      "url": "http://genome.ucsc.edu/cgi-bin/hgTracks?
org=human&db=hg38&position=chr9:21531306-22492891&hgt.customText=https://
progenetix.org/tmp/5cced529-3acf-4156-b121-6ae7e5e63d0c.bed"
    }
  ]
```

Progenetix & Beacon v1->2

Handover elements in Beacon responses

- Progenetix utilizes handovers to deliver data matched by the Beacon queries
- These handovers are interpreted by the front end to populate different parts of the UI, w/o the need of active selection
- Handovers are either standard Beacon v2 paths or dedicated custom functions

progenetix

Cancer CNV Profiles
Search Samples
Studies & Cohorts
arrayMap
TCGA Samples
DIPG Samples
Gao & Baudis, 2021
Cancer Cell Lines
Publication DB
Services
NCIt Mappings
UBERON Mappings
Upload & Plot
Beacon+
Progenetix Info
About Progenetix
Use Cases
Documentation
Baudisgroup @ UZH

Search Samples | Modify Query

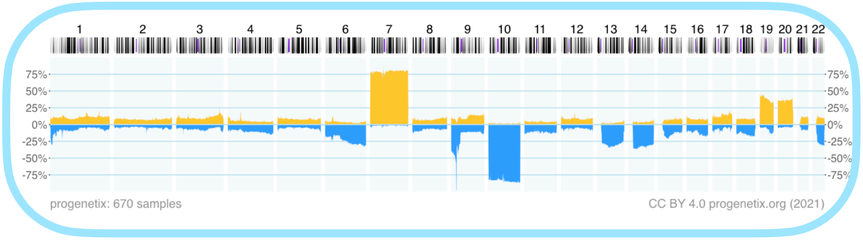
Assembly: GRCh38 **Chro:** 9 **Start:** 21500001-21975098 **End:** 21967753-22500000
Type: DEL **Filters:** NCIT:C3058

progenetix

Samples: 668
Variants: 286
Calls: 675

Found Variants (.pgxseg)
All Sample Variants (.json)
All Sample Variants (.pgxseg)
Show Variants in UCSC
UCSC region
JSON Response
Visualization options

Results | Biosamples | Biosamples Map | Variants

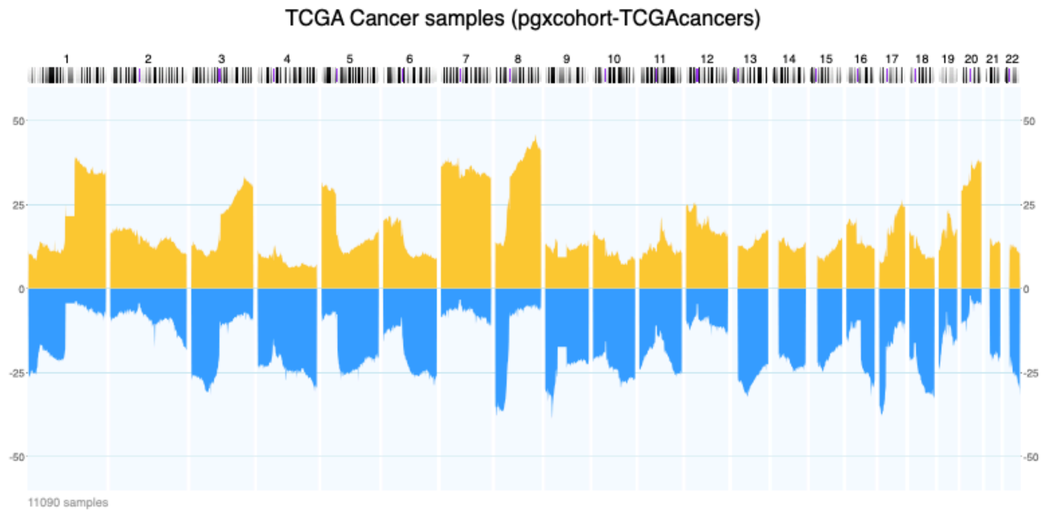


Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
icdot-C71.4	4	1	0.250
UBERON:0002021	4	1	0.250
icdom-94403	4291	664	0.155
NCIT:C3058	4375	664	0.152
icdot-C71.1	14	2	0.143
UBERON:0016525	14	2	0.143
UBERON:0000955	6700	651	0.097
icdot-C71.9	7282	651	0.089

Progenetix API

Data & Plots

- "all" of the data can be accessed using API calls
- segmented CNV data in .pgxsg (columnar) and JSON format
- histograms for disease, study or cohort from precomputed frequencies - live generated as SVG for embedding with plot options



*https://progenetix.org/cgi/PGX/cgi/collationPlots.cgi?datasetIds=progenetix&id=**pgxcohort-TCGAcancers**&-size_plotimage_w_px=**800**&-size_plotarea_h_px=**300**&-value_plot_y_max=**60***

Beacon & Phenopackets

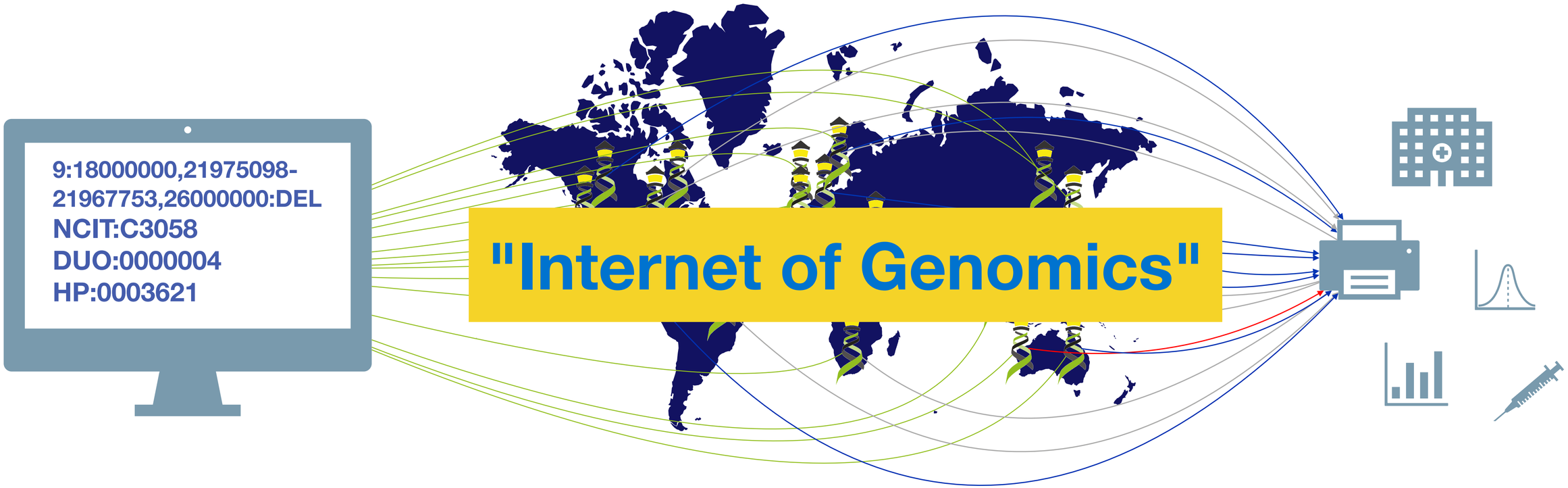
Data *discovery* and *delivery* using standardized GA4GH formats and schemas

- modern standards and protocols such as Beacon & Phenopackets are essential for federation and exchange of biomedical data
- emerging / established principles are the use of hierarchical coding systems and with widespread use of CURIEs
- other formats based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)
- these standards become pervasive throughout GA4GH's ecosystem

➡ Beacon query **filters** correspond well to Phenopackets data

➡ Phenopackets as supported protocol for Beacon data delivery

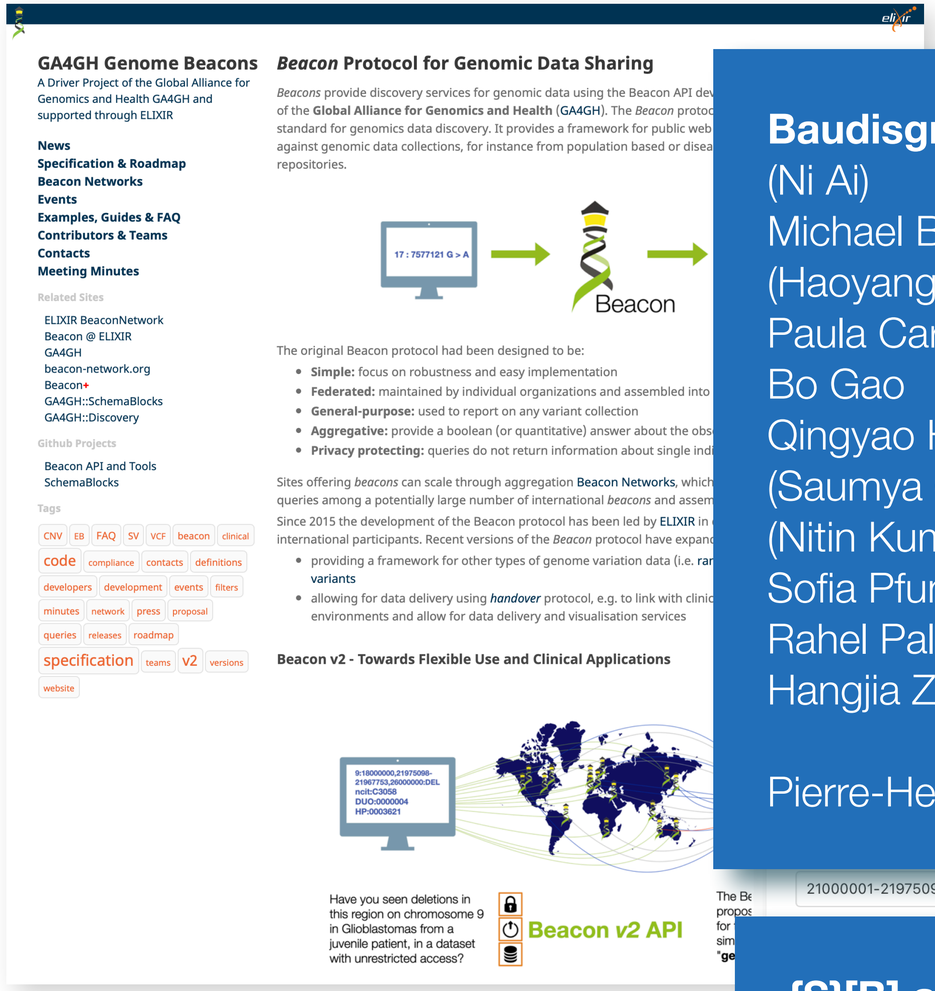
```
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},
"material" : {
  "id" : "EFO:0009656",
  "label" : "neoplastic sample"
},
{
  "ageAtDiagnosis": "P25Y3M2D"
},
"sampled_tissue" : {
    "id" : "UBERON:0002037",
    "label" : "cerebellum"
},
"histological_diagnosis" : {
    "id" : "NCIT:C3222",
    "label" : "Medulloblastoma"
},
```



Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon *v2* API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".



beacon-project.io

Baudisgroup @ UZH

(Ni Ai)
Michael Baudis
(Haoyang Cai)
Paula Carrio Cordo
Bo Gao
Qingyao Huang
(Saumya Gupta)
(Nitin Kumar)
Sofia Pfund
Rahel Paloots
Hangjia Zhao

Pierre-Henri Toussaint

{S}[B] and GA4GH

Melanie Courtot
Helen Parkinson
many more ...

Beacon API Leads

Jordi Rambla
Anthony Brooks
Juha Törnroos

Discovery WS

Michael Baudis (Beacon)
Marc Fiume (Networks)

ELIXIR

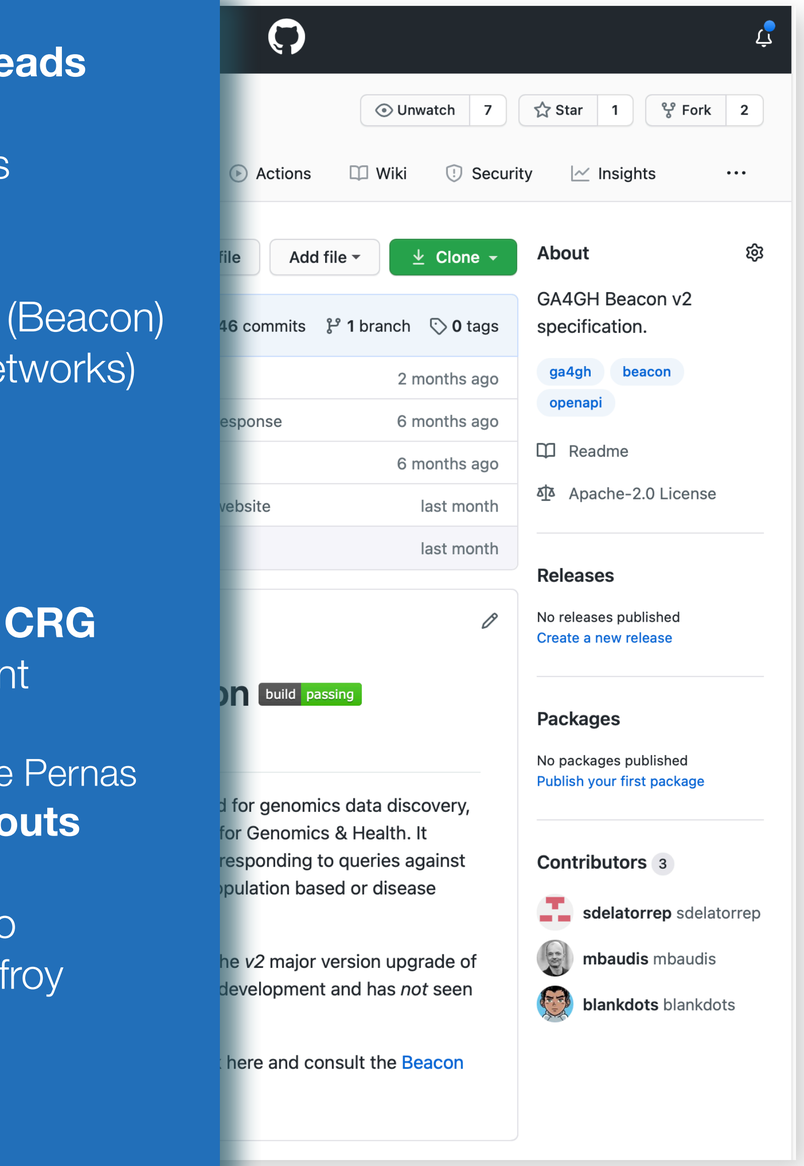
Gary Saunders
David Lloyd
Serena Scollen

Beacon Team CRG

Laureen Fromont
Babita Singh
Sabela de la Torre Pernas

Beacon v2 Scouts

Tim Beck
Joaquin Dopazo
Veronique Geoffroy
Jean Muller
David Salgado
Alex Wagner
...



beacon.progenetix.org/ui/

github.com/ga4gh-beacon/