hCNV data and the Progenetix Beacon

Implementing a cancer hCNV reference resource on top of Beacon v2
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)
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
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
- custom implementation
  - "bycon" code on GH > /progenetix/bycon/
  - MongoDB backend
  - React front-end on GH > /progenetix/progenetix-web/
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, reference 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
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 Adenocar.
- initiatives and services such as Phenopackets, MONDO, OXO ... rely on and/or provide mappings to hierarchical ontologies

NCIt Neoplasm Core coded display (excerpt) for samples in the Progenetix cancer genome data resource allows sample selection on multiple hierarchy levels
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 limitations of "fuzzy" disease annotations

- data handover (Beacon v1.1+) enables further data exploration and export scenarios
Beacon v2 Paths
Progenetix utilizes Beacon v2 REST paths

- Beacon v2 paths are used 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
  ➡...

Base /biosamples

/biosamples/ + query
- /biosamples/?filters=cellulosaurus:CVCL_0004
  - this example retrieves all biosamples having an annotation for the Cellulosaurus CVCL_0004 identifier (K562)
- /biosamples/{id}/
  - /biosamples/pgxbis-kfva5c9/
    - retrieval of a single biosample
  - /biosamples/{id}/variants/ & /biosamples/{id}/variants_in_sample/
    - /biosamples/pgxbis-kfva5c9/variants/
    - /biosamples/pgxbis-kfva5c9/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 /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=7669696&end=7800000
    - This is an example for a Beacon "Bracket Query" which will return focal deletions in the TPS3 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/
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 & 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.

```json
"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 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
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

```json
"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 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

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 ...