The ELIXIR Beacon Project: Future developments

Michael Baudis, University of Zurich | **SIB**

ELIXIR All Hands 2018, 4-7 June 2018, Berlin, Germany

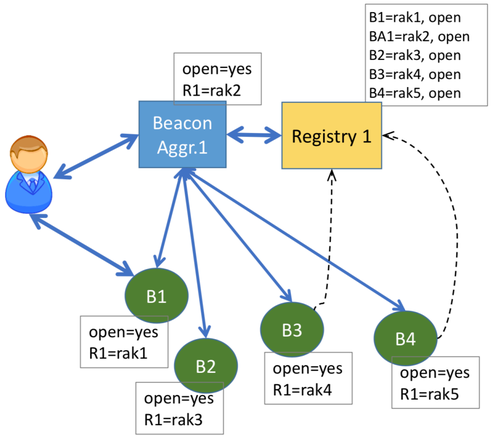
Global Alliance for Genomics & Health

University of Zurich[UZH]

www.elixir-europe.org

Towards a shining future...

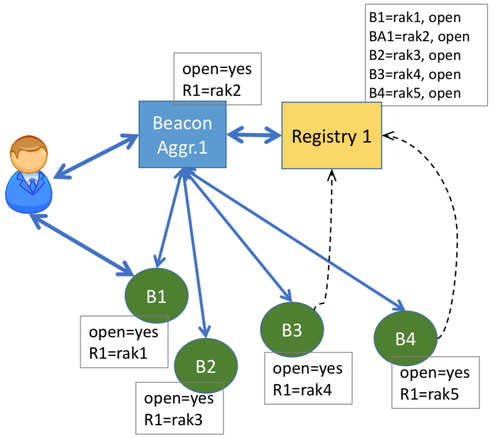
- extensive/complete representation of genome **variant types** in query
  - close coordination with **GA4GH::GKS** and **GA4GH::Discovery** work streams
- providing a tested model for layered **registered access**
  - ELIXIR AAI
- implementing **Beacon network**(s) throughout ELIXIR
  - open protocols for extension and external implementations
  - Registry/ies for ELIXIR resources and external Beacons
- extending Beacon query protocol (**metadata**...)
  - keeping "aggregate response" model
- Beacon queries as entry points for data delivery, using "**handover**" scenarios



Global Alliance
for Genomics & Health

Towards a shining future...

☑ extensive/complete representation of genome **variant types** in query
  - close coordination with **GA4GH::GKS** and **GA4GH::Discovery** work streams
- providing a tested model for layered **registered access**
  - ELIXIR AAI
- implementing **Beacon network**(s) throughout ELIXIR
  - open protocols for extension and external implementations
  - Registry/ies for ELIXIR resources and external Beacons



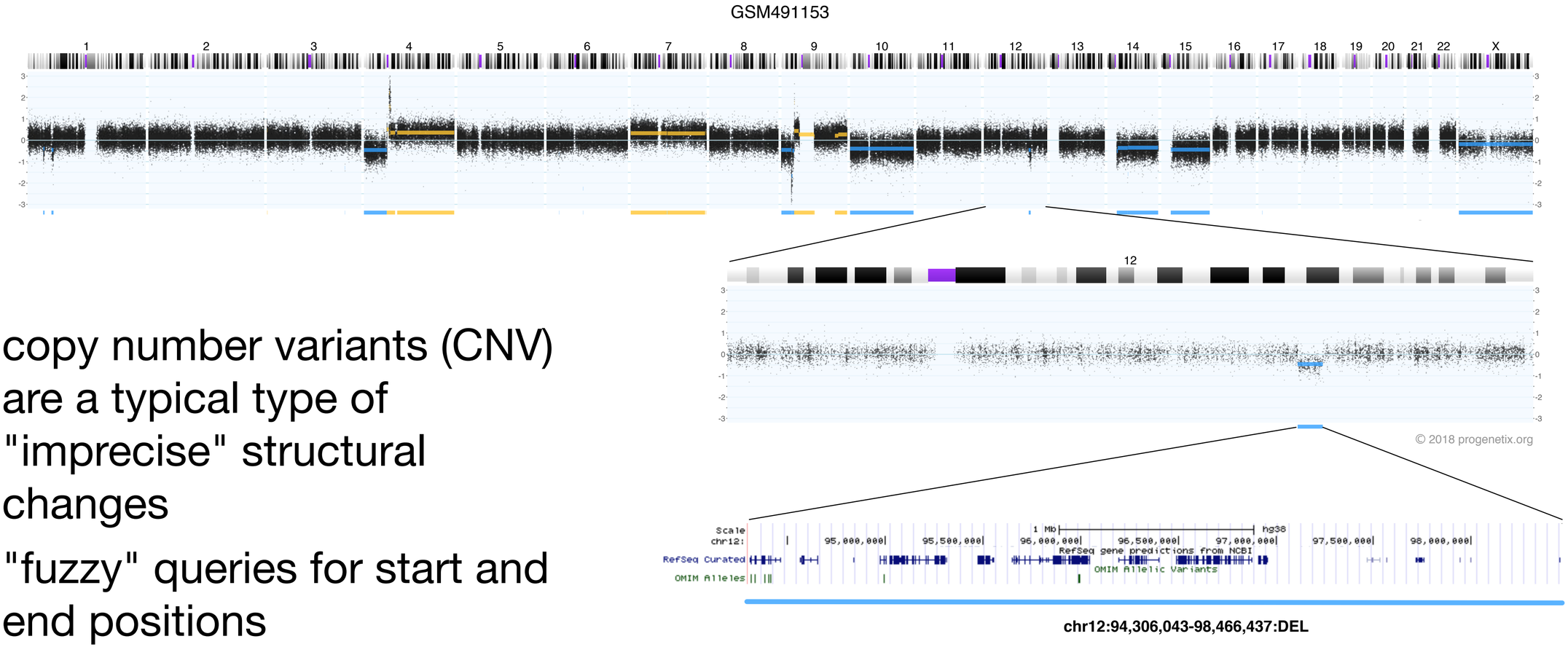
☑ extending Beacon query protocol (**metadata**...)
  - keeping "aggregate response" model

☑ Beacon queries as entry points for data delivery, using "**handover**" scenarios

Global Alliance
for Genomics & Health

Querying Copy Number Variants

- copy number variants (CNV) are a typical type of "imprecise" structural changes
- "fuzzy" queries for start and end positions



chr12:94,306,043-98,466,437:DEL

start_min: 94,000,000
start_max: 94,500,000

reference_name: “9”

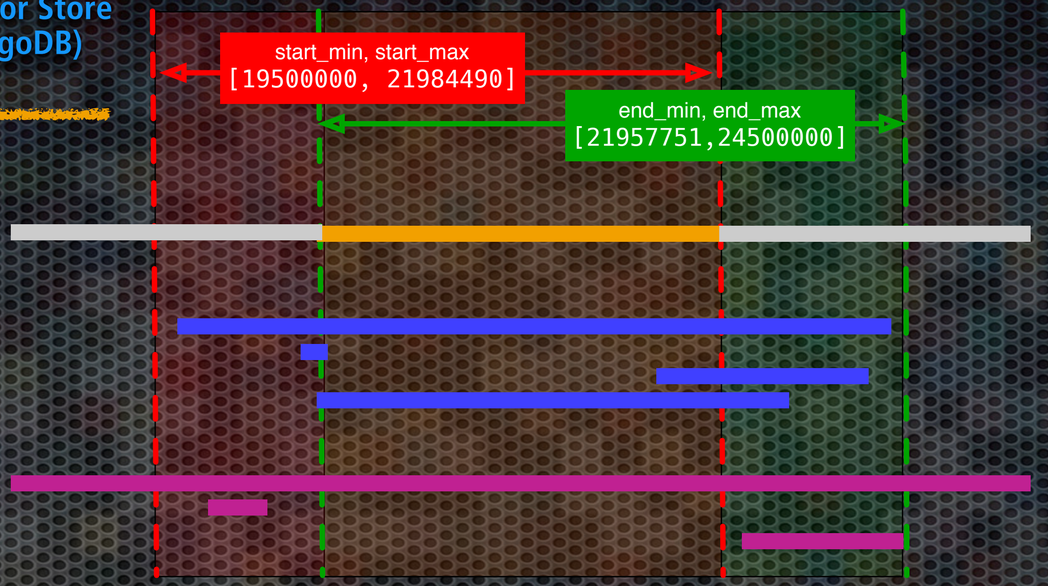
variant_type: “**DEL**”

end_min: 98,200,000
end_max: 98,700,000

```
{
   "allele_request" : {
      "$and" : [
         { "reference_name" : “9" },
         { "variant_type" : “DEL" },
         { "start" : { "$gte" : 19500000 } },
         { "start" : { "$lte" : 21984490 } },
         { "end" : { "$gte" : 21957751 } },
         { "end" : { "$lte" : 24500000 } }
      ]
   },
   "api_version" : "0.4",
   "beacon_id" : "org.progenetix:progenetix-beacon",
   "exists" : true,
   "info" : {
      "query_string" :
"datasetId=arraymap&referenceName=chr9&assemblyId=GRCh38&variantType=DE
L&startMax=19000000&startMin=21984490&endMin=21900000&endMax=25000000&b
iosamples.bio_characteristics.ontology_terms.term_id=icdom:9440_3",
      "version" : "Beacon+ implementation based on a development branch
of the beacon-team project: https://github.com/ga4gh/beacon-team/pull/
94"
   },
   "url" : “http://progenetix.org/beacon/info/“,
   "dataset_allele_responses" : [
      {
         “datasetId" : "arraymap",
         "error" : null,
         "exists" : true,
         "external_url" : “http://arraymap.org",
         "sample_count" : 584,
         "call_count" : 3781,
         "variant_count" : 3244,
         "frequency" : 0.0094,
         "info" : {
           "description" : "The query was against database
\"arraymap_ga4gh\", variant collection \"variants_cnv_grch36\". 3781 /
59428 matched callsets for 3602919 variants. Out of 62105 biosamples in
the database, 2047 matched the biosample query; of those, 584 had the
variant.",
            "ontology_ids" : [
              "ncit:C3058",
              "pgx:icdom:9440_3",
              "pgx:icdot:C71.9",
              “pgx:icdot:C71.0"
            ]
         },
      }
   ],
```


Translation for Store (here MongoDB)

start_min
start_max
end_min
end_max



Match using query ranges “at least one base in interval affected”

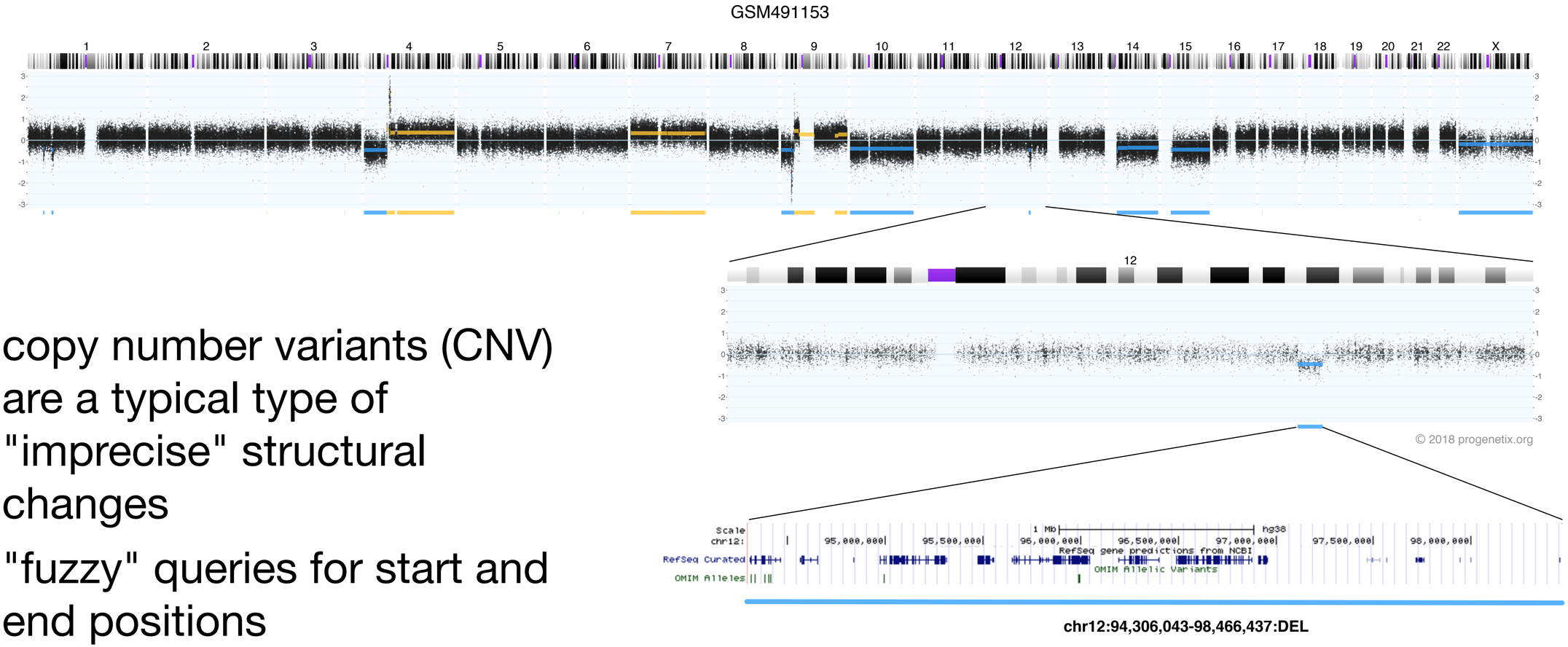
Region of Interest, e.g. CDR of Gene (here: CDKN2A)

Example “focal” matches (overlap w/ size limit)

Mismatches
- too large
- end outside
- start outside

- Beacon+ **range queries** allow the definition of a genome region of interest, containing a specified variant (or other mappable feature)
- “fuzzy” matching of region ends is essential for features without base specific positions
- current Beacon implementation addresses CNV (DUP, DEL), as are specified in VCF && GA4GH variant schema

Querying Copy Number Variants (& breaks)



- copy number variants (CNV) are a typical type of "imprecise" structural changes
- "fuzzy" queries for start and end positions
- alternative hits on ***break ends (BND)***

start_min: 94,000,000
start_max: 94,500,000

variant_type: “**BND**”

reference_name: “9”

variant_type: “**DEL**”

end_min: 98,200,000
end_max: 98,700,000

variant_type: “**BND**”

Beacon - Breaks | Fusions | Translocations

- Minimal unit is a **break** (w/o indication of associated second quality):
  - can already be represented through variant_type: "BND" - but not documented yet
  - a BND query could e.g. deliver all matching DUP/DEL end points
- A fusion event requires definition of a 2nd "**mate_name**" chromosome
  - could be the same as the reference_name, e.g. for large indel or inversion...)

Structural variants: mateName for fusions #176

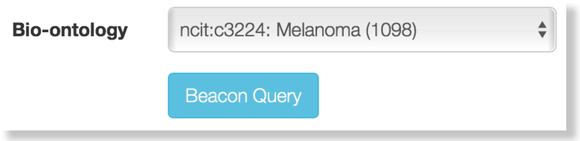
Open **mbaudis** wants to merge 3 commits into `develop` from `develop-structural-variants`

```
* DUP
    * duplication of sequence following `start` or beginning
      in the `startMin` => `startMax` interval and ending at `end`
      or in the `endMin` => `endMax` interval; not necessarily in situ
* DEL
    * deletion of sequence following `start` or beginning
      in the `startMin` => `startMax` interval and ending at `end`
      or in the `endMin` => `endMax` interval; not necessarily in situ
* BND
    * breakend, i.e. termination of the allele at position
      `start` or in the `startMin` => `startMax` interval, or fusion
      of the sequence to distant partner
```

```
- name: mateName
  description: |
    Second chromosome for fusion events. This can be
    * empty (no fusion or unknown partner)
    * identical to `referenceName` (e.g. one side of an inversion)
    * a different chromosome
    Accepting values 1-22, X, Y.
  in: query
  required: false
  schema:
    $ref: '#/components/schemas/Chromosome'
```

Bio-Metadata: Query ontology classes

- Emerging consensus about representation of "bio-characteristics" (phenotypes, diseases, observations...) through ontology classes as robust classifiers
- Convergence around *Phenopackets* (PXF) as implementation for data exchange
- Use of (one or multiple) namespace mapped ontology classes as first implementation of a "bio-query" for Beacon?




```
// A class (aka term, concept) in an ontology
message OntologyClass {
    // a CURIE-style identifier e.g. HP:0100024, MP:0001284, UBERON:0001690.
    // This is the primary key for the ontology class
    // REQUIRED?
    string id = 1;

    // class label, aka name. E.g. "Abnormality of cardiovascular system"
    string label = 2;
}
```

Phenopackets

```
// An ontology term describing an attribute. (e.g. the phenotype attribute
// 'polydactyly' from HPO)
message OntologyTerm {
  // Ontology term identifier - the CURIE
  // differs from the standard GA4GH schema
  // in that it is a CURIE pointing to an information resource outside of the
  // scope of the schema or its resource implementation.
  string term_id = 1;

  // Ontology term - the label of the ontology term the termId is pointing to.
  string term_label = 2;
}
```

GA4GH::Metadata

```
"bio_character
  {
    "descripti
    "ontology_
      {
        "term_
        "term_id
      },
      {
        "term_label" : "Glioma NOS",
        "term_id" : "pgx:icdom:9380_3"
      },
      {
        "term_label" : "Brain NOS",
        "term_id" : "pgx:icdot:C71.9"
      }
    ],
    "negated_ontology_terms" : [ ]
  }
],
```

...wrapping it up

Beacon: Data delivery - *Handover* scenario

- Paradigm: Beacon can be extended in functionality through **advancing query options**, but should only provide **aggregate results**
- Extending Beacon queries will implement a reference **Genomic Query API**
- Broadening of utility and uptake through support for data delivery through a "***Handover***" scenario
- Options

➡ Just an access handle w/o information

➡ BeaconInfo with detailes about delivered data types, formats in the delivery part

Beacon+ Concept

Implementing Data | Structural | Handover | Ontologies

Michael Baudis, 2018-03-19

beacon.progenetix.org

Beacon Implementations

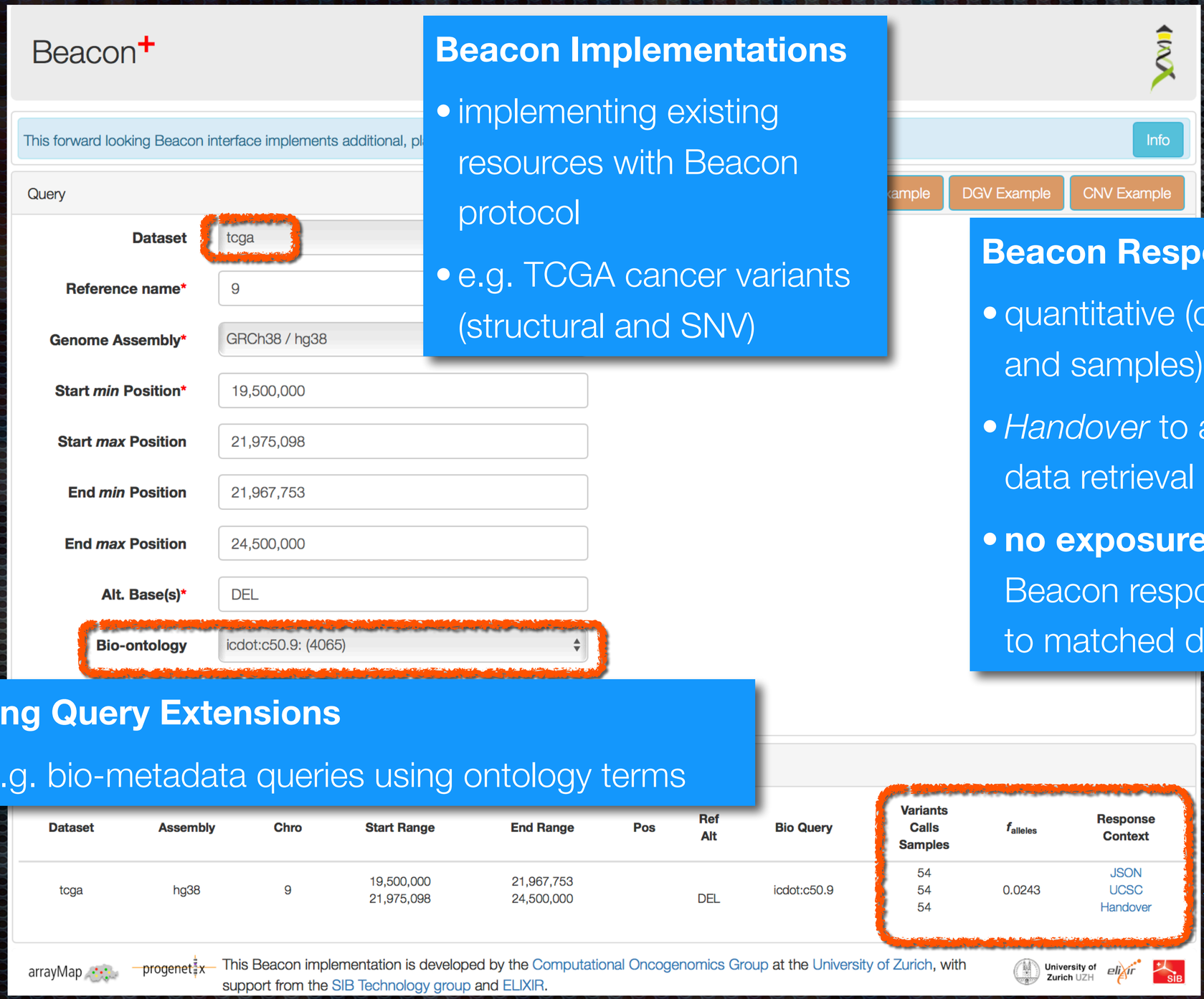
- implementing existing resources with Beacon protocol
- e.g. TCGA cancer variants (structural and SNV)

Beacon Response

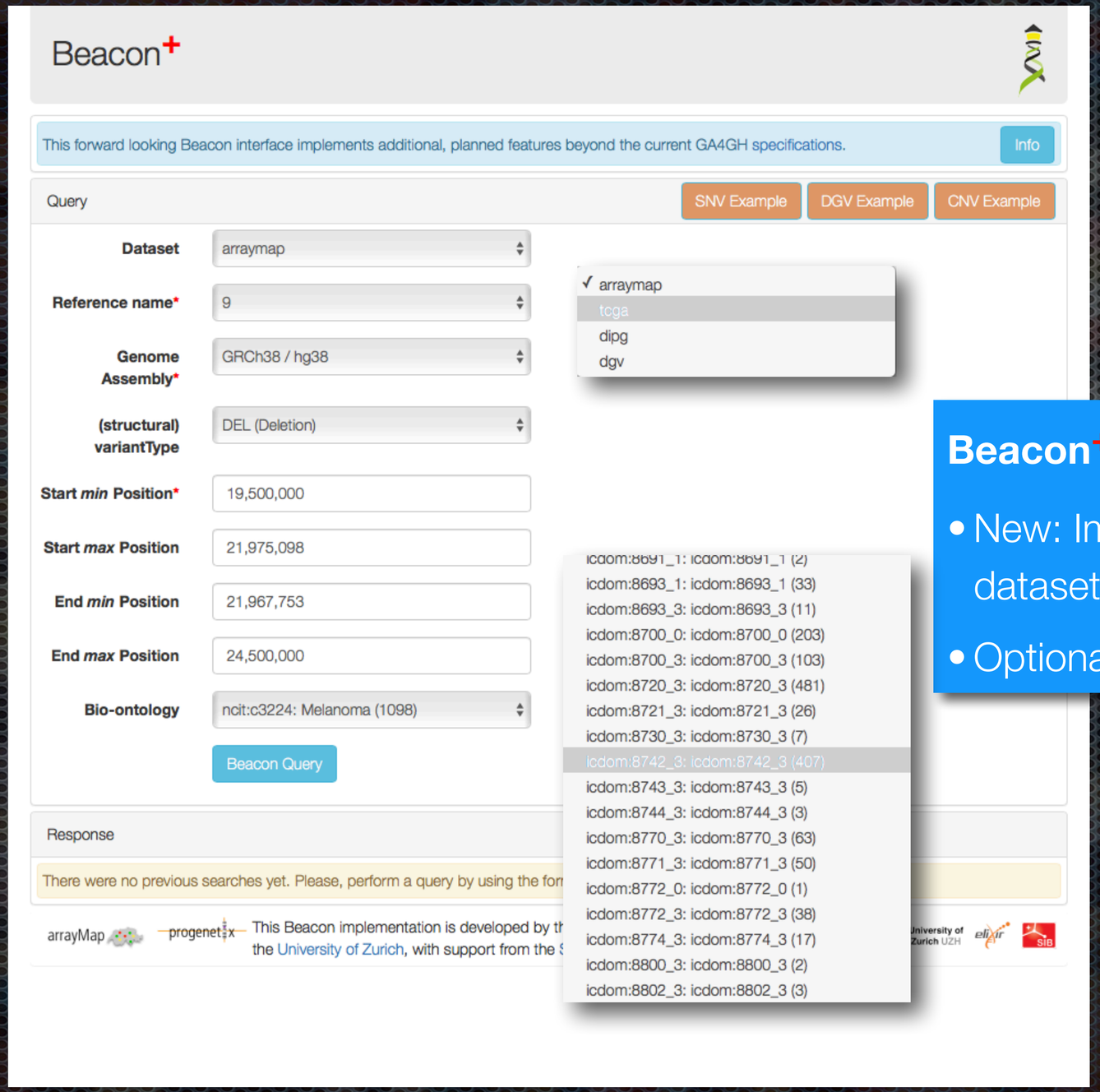
- quantitative (counts for variants, callsets and samples)
- *Handover* to authentication system for data retrieval
- **no exposure** of data beyond standard Beacon response and additional pointer to matched data

Prototyping Query Extensions

- testing e.g. bio-metadata queries using ontology terms



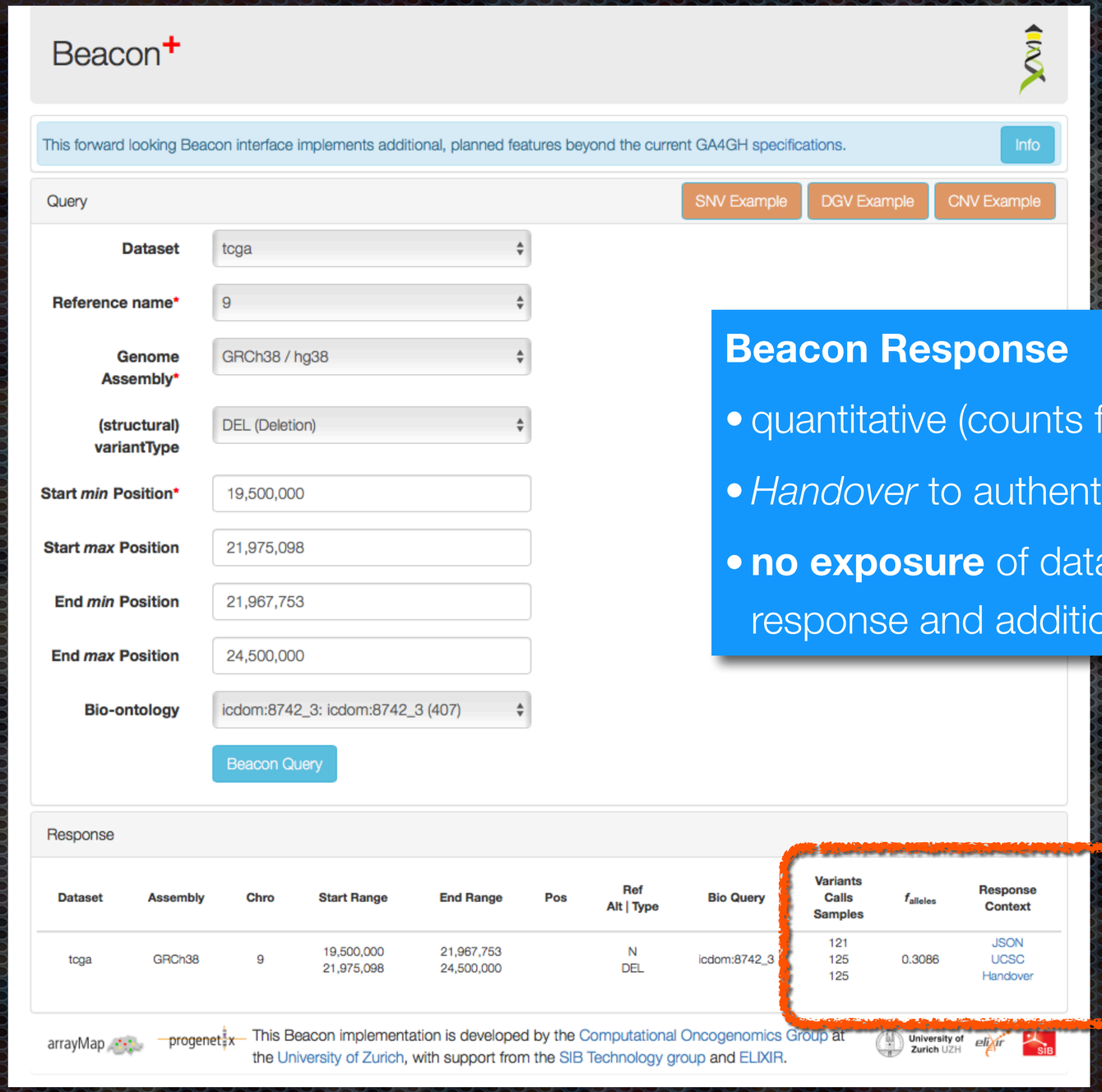
beacon.progenetix.org



Beacon+: Additional Features

- New: Implementation of TCGA dataset (SNV & CNV)
- Optional code based disease scoping

beacon.progenetix.org



Beacon Response

- quantitative (counts for variants, callsets and samples)
- *Handover* to authentication system for data retrieval
- **no exposure** of data beyond standard Beacon response and additional pointer to matched data

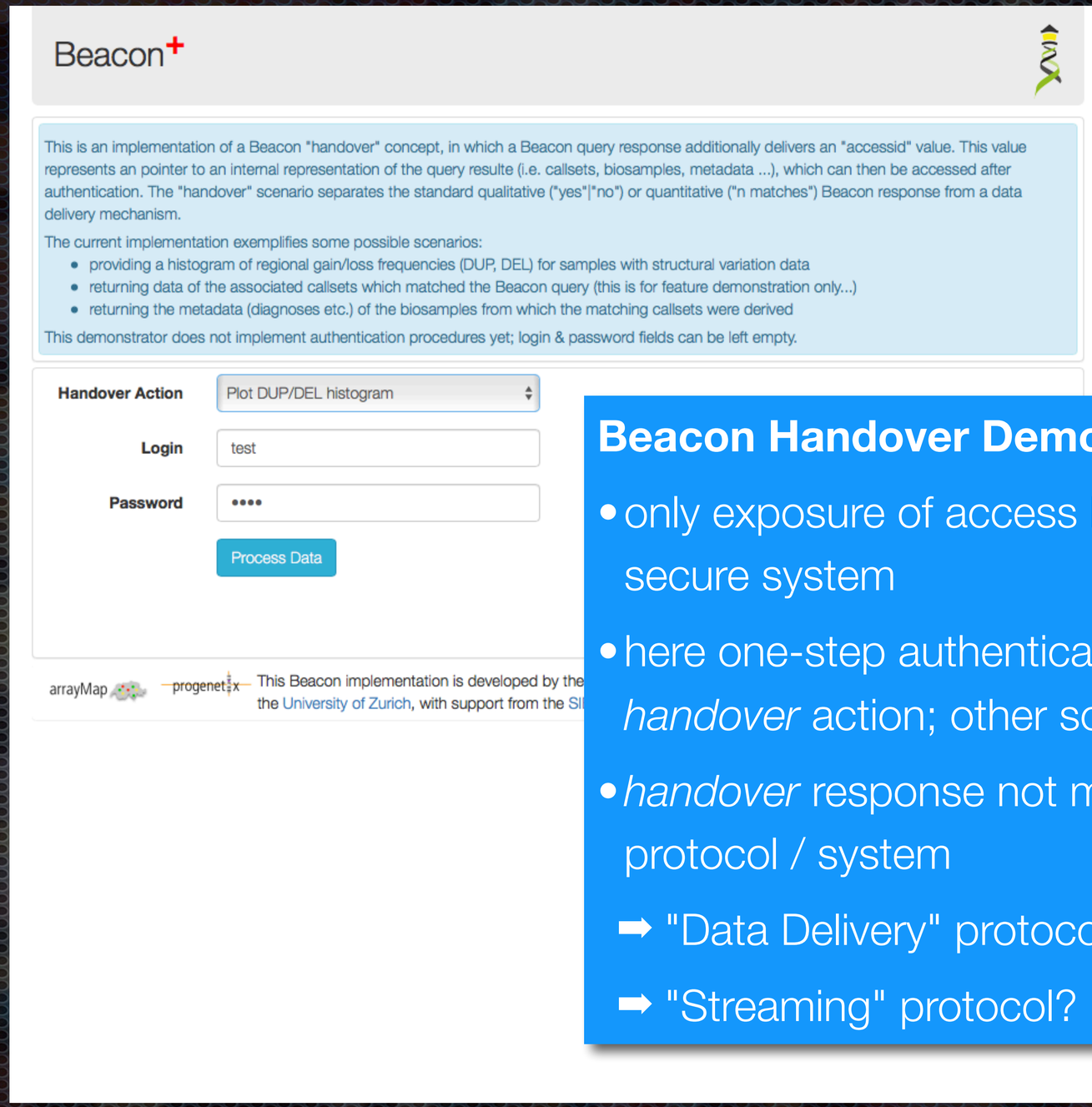
beacon.progenetix.org


```
{
   "alleleRequest" : {
      "assemblyId" : "GRCh38",
      "bio_characteristics.ontology_terms.term_id" : "icdom:8742_3",
      "datasetId" : "tcga",
      "endMax" : 24500000,
      "endMin" : 21967753,
      "referenceBases" : "N",
      "referenceName" : "9",
      "startMax" : 21975098,
      "startMin" : 19500000,
      "variantType" : "DEL"
   },
   "apiVersion" : "0.4",
   "beaconId" : "org.progenetix:progenetix-beacon",
   "datasetAlleleResponses" : [
      {
         "callCount" : 125,
         "datasetId" : "tcga",
         "error" : null,
         "exists" : true,
         "externalUrl" : "http://beacon.arraymap.org",
         "frequency" : 0.3086
         "info" : {
            "callset_access_handle" : "661709f3-2888-11e8-a2e8-aeb232cb87eb",
            "description" : "The query was against database \"tcga_ga4gh\" variant collection \"variants\". 125 matched
callsets for 121 distinct variants. Out of 41672 biosamples in the database, 405 matched the biosample query; of those, 125
had the variant.",
            "ontology_selection" : [],
            "payload" : null,
            "phenotype_response" : []
         },
         "note" : "",
         "sampleCount" : 125,
         "variantCount" : 121
      }
   ],
   "exists" : true,
   "info" : {
      "queryString" :
"datasetId=tcga&referenceName=9&assemblyId=GRCh38&variantType=DEL&startMin=19,500,000&startMax=21,975,098&endM
ndMax=24,500,000&referenceBases=N&biosamples.bio_characteristics.ontology_terms.term_id=icdom:8742_3",
      "version" : "Beacon+ implementation based on the development branch of the beacon-team project:
https://github.com/ga4gh/beacon-team/blob/develop-proto/src/main/proto/ga4gh/beacon.proto"
   },
   "url" : "http://progenetix.org/beacon/info/"
}
```


Beacon Handover Demonstrator

- only exposure of access handle to data stored in secure system
- here one-step authentication and selection of *handover* action; other scenarios possible / likely
- *handover* response outside of Beacon protocol / system

beacon.progenetix.org



Beacon Handover Demonstrator

- only exposure of access handle to data stored in secure system
- here one-step authentication and selection of *handover* action; other scenarios possible / likely
- *handover* response not managed by Beacon protocol / system
  - ➡ "Data Delivery" protocol?
  - ➡ "Streaming" protocol?

beacon.progenetix.org

Beacon+

This is an implementation of a Beacon "handover" concept, in which a Beacon query response additionally delivers an "accessid" value. This value represents an pointer to an internal representation of the query resulte (i.e. callsets, biosamples, metadata ...), which can then be accessed after authentication. The "handover" scenario separates the standard qualitative ("yes"|"no") or quantitative ("n matches") Beacon response from a data delivery mechanism.

The current implementation exemplifies some possible scenarios:

- providing a histogram of regional gain/loss frequencies (DUP, DEL) for samples with structural variation data
- returning data of the associated callsets which matched the Beacon query (this is for feature demonstration only...)
- returning the metadata (diagnoses etc.) of the biosamples from which the matching callsets were derived

This demonstrator does not implement authentication procedures yet; login & password fields can be left empty.

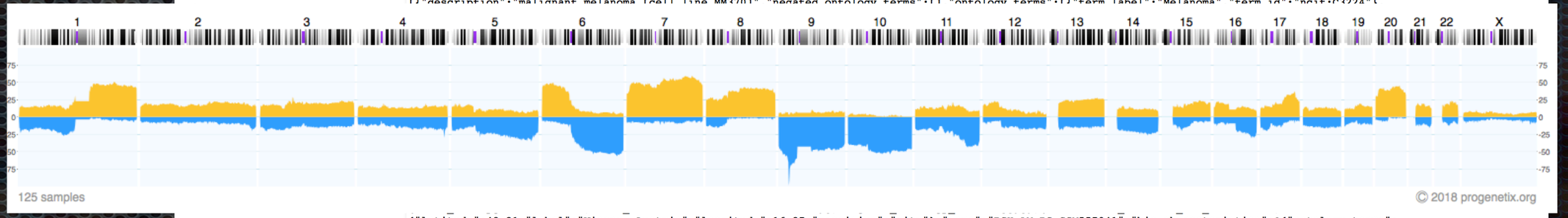
Handover Action Plot DUP/DEL histogram

Login test

Password ••••

Proce

arrayMap progenetix Th



[▼ 25620 items, 13 MB
{ ▼ 14 properties, 528 bytes
"callset_id": "PGX_AM_CS_GSM557941",
"start": 13851025,
"ciend": [],
"reference_name": "4",
"cipos": [],
"info": { ▼ 2 properties, 62 bytes
"svlen": 31963,
"value": -2.4205
},
"alternate_bases": [▼ 1 item, 26 bytes
"<DEL>"
],
"end": 13882988,
"variant_type": "DEL",
"variantset_id": "arraymap_ga4gh_vs_GRCh38",
"reference_bases": ".",
"digest": "4:13851025-13882988:DEL",
"biosample_id": "PGX_AM_BS_GSM557941",
"genotype": []
},

[{"location":{"longitude":-118.49,"precision":"city","label":"Santa Monica, United States","latitude":34.02},"individual_id":"PGX_IND_GSM1076715","name":"PGX_AM_BS_GSM1076715","bio_characteristics":[{"negated_ontolo
[{"term_label":"Melanoma","term_id":"ncit:C3224"},{"term_id":"pgx:icdom:8720_3","term_label":"Malignant melanoma, NOS"},{"term_id":"
{"term_label":"Melanoma of the Skin","term_id":"pgx:seer:25010"}],"description":"malignant melanoma [Lymph node metastasis]"}],"exte
[{"identifier":"geo:GPL6801","relation":"denotes"},{"identifier":"geo:GSM1076715","relation":"denotes"},{"relation":"denotes","ident
{"relation":"denotes","identifier":"geo:GSE44019"}],"individual_age_at_collection":{"age":"","age_class":"null"},"description":"malignant melanoma [Lymph node metastasis]","id":"PGX_AM_BS_GSM1076715"},{"bio_characteristics":[{"description":"malignant melanoma [cell line WM3311]","negated_ontology_terms":[],"ontology_terms":
[{"term_label":"Melanoma","term_id":"ncit:C3224"},{"term_label":"Malignant melanoma, NOS","term_id":"pgx:icdom:8720_3"},{"term_id":"pgx:icdot:C44","term_label":"skin"},
{"term_id":"pgx:seer:25010","term_label":"Melanoma of the Skin"}]}],"name":"PGX_AM_BS_GSM952606","location":{"precision":"city","longitude":-75.16,"label":"Philadelphia, United States","latitude":39.95},"individual_id":"PGX_IND_GSM952606","id":"PGX_AM_BS_GSM952606","individual_age_at_collection":
{"age":"","age_class":"null"},"description":"malignant melanoma [cell line WM3311]","external_identifiers":[{"identifier":"cellosaurus:CVCL_0B72","relation":"denotes"},
{"relation":"denotes","identifier":"geo:GSM952606"},{"identifier":"geo:GPL9777","relation":"denotes"},{"relation":"denotes","identifier":"geo:GSE38946"}]},
{"id":"PGX_AM_BS_GSM750824","individual_age_at_collection":{"age_class":"null","age":""},"description":"malignant melanoma","external_identifiers":
[{"identifier":"cellosaurus:CVCL_0069","relation":"denotes"},{"relation":"denotes","identifier":"geo:GSM750824"},{"relation":"denotes","identifier":"geo:GSE30291"},
{"relation":"denotes","identifier":"geo:GPL13786"}],"bio_characteristics":[{"description":"malignant melanoma","negated_ontology_terms":[],"ontology_terms":
[{"term_label":"Melanoma","term_id":"ncit:C3224"},{"term_id":"pgx:icdom:8720_3","term_label":"Malignant melanoma, NOS"},{"term_id":"pgx:icdot:C44","term_label":"skin"},
{"term_id":"pgx:seer:25010","term_label":"Melanoma of the Skin"}]}],"name":"PGX_AM_BS_GSM750824","location":{"label":"Bethesda, United States","latitude":38.98,"longitude":-77.1,"precision":"city"},"individual_id":"PGX_IND_GSM750824"},{"location":{"label":"Brisbane, Australia","latitude":-27.47,"longitude":153.03,"precision":"city"},"individual_id":"PGX_IND_GSM226658","name":"PGX_AM_BS_GSM226658","bio_characteristics":

{"latitude":48.21,"label":"Vienna, Austria","longitude":16.37,"precision":"city"},"name":"PGX_AM_BS_GSM557941","bio_characteristics":[{"ontology_terms":
[{"term_id":"ncit:C3224","term_label":"Melanoma"},{"term_label":"Malignant melanoma, NOS","term_id":"pgx:icdom:8720_3"},{"term_id":"pgx:icdot:C44","term_label":"skin"},
{"term_id":"pgx:seer:25010","term_label":"Melanoma of the Skin"}],"negated_ontology_terms":[],"description":"malignant melanoma [cell line MM1]"}],"external_identifiers":
[{"identifier":"geo:GSM557941","relation":"denotes"},{"identifier":"cellosaurus:CVCL_2075","relation":"denotes"},{"identifier":"pubmed:21584902","relation":"denotes"},
{"relation":"denotes","identifier":"geo:GPL6801"},{"relation":"denotes","identifier":"geo:GSE22461"}],"description":"malignant melanoma [cell line MM1]","individual_age_at_collection":{"age":"","age_class":"null"},"id":"PGX_AM_BS_GSM557941"},{"bio_characteristics":[{"negated_ontology_terms":[],"ontology_terms":
[{"term_id":"ncit:C3224","term_label":"Melanoma"},{"term_label":"Malignant melanoma, NOS","term_id":"pgx:icdom:8720_3"},{"term_label":"skin","term_id":"pgx:icdot:C44"},
{"term_label":"Melanoma of the Skin","term_id":"pgx:seer:25010"}],"description":"malignant melanoma [cell line WM3438]"}],"location":
{"longitude":-75.16,"precision":"city","latitude":39.95,"label":"Philadelphia, United

Next steps

- expanding **variant types** for more structural types
  - coordination with **GA4GH::GKS**
  - pick up of "Structural variants: mateName for fusions #176"
- Introducing first type of "bio-metadata" query option
  - coordination with **GA4GH::CP**, **GA4GH::Metadata** and **Phenopackets**
- Registry/ies for ELIXIR resources and external Beacons
- Discuss & prototype "**handover**" scenarios
  - security considerations...

Thank You!

... and many more

Global Alliance
for Genomics & Health