A Genomics / CNV Resource Built on GA4GH Standards

... also Implementation Driven Standards Development

Michael Baudis | hCNV Community Meeting 2023

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- 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
- recent addition of SNV data for some series

Universität Zürich[UZH]

progenetix

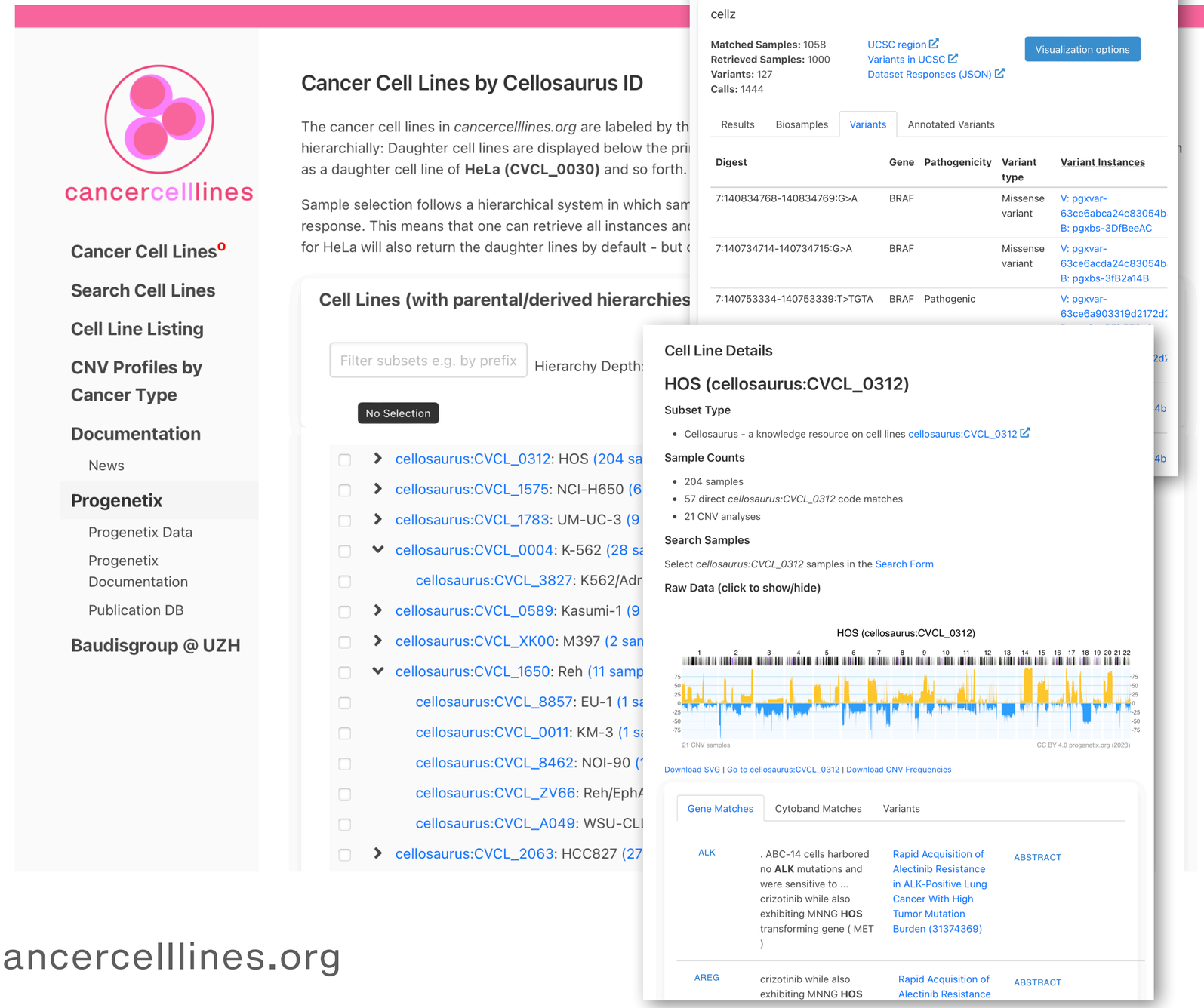
SIB Swiss Institute of Bioinformatics

Cancer Cell Lines

Cancer Genomics Reference Resource

- starting from >5000 cell line CNV profiles
  - 5754 samples | 2163 cell lines
  - 256 different NCIT codes
- genomic mapping of annotated variants and additional data from several resources (ClinVar, CCLE, Cellosaurus...)
  - 16178 cell lines
  - 400 different NCIT codes
- query and data delivery through Beacon v2 API

➡ integration in data federation approaches



cancercelllines.org

Lead: Rahel Paloots

Progenetix Stack

progenetix

- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
  - biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads…
- the complete middleware / CGI stack is provided through the ***bycon*** package
  - schemas, query stack, data transformation (e.g. Phenopackets generation)…
- data collections mostly correspond to the main Beacon default model entities
  - no separate *runs* collection; integrated w/ analyses
  - *variants* are stored per observation instance

React

APACHE HTTP SERVER PROJECT

python™

mongoDB

- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703…
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding access handle for **handover** generation

```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
```

variants analyses biosamples individuals

Entity collections

collations geolocs genespans publications qBuffer

Utility collections

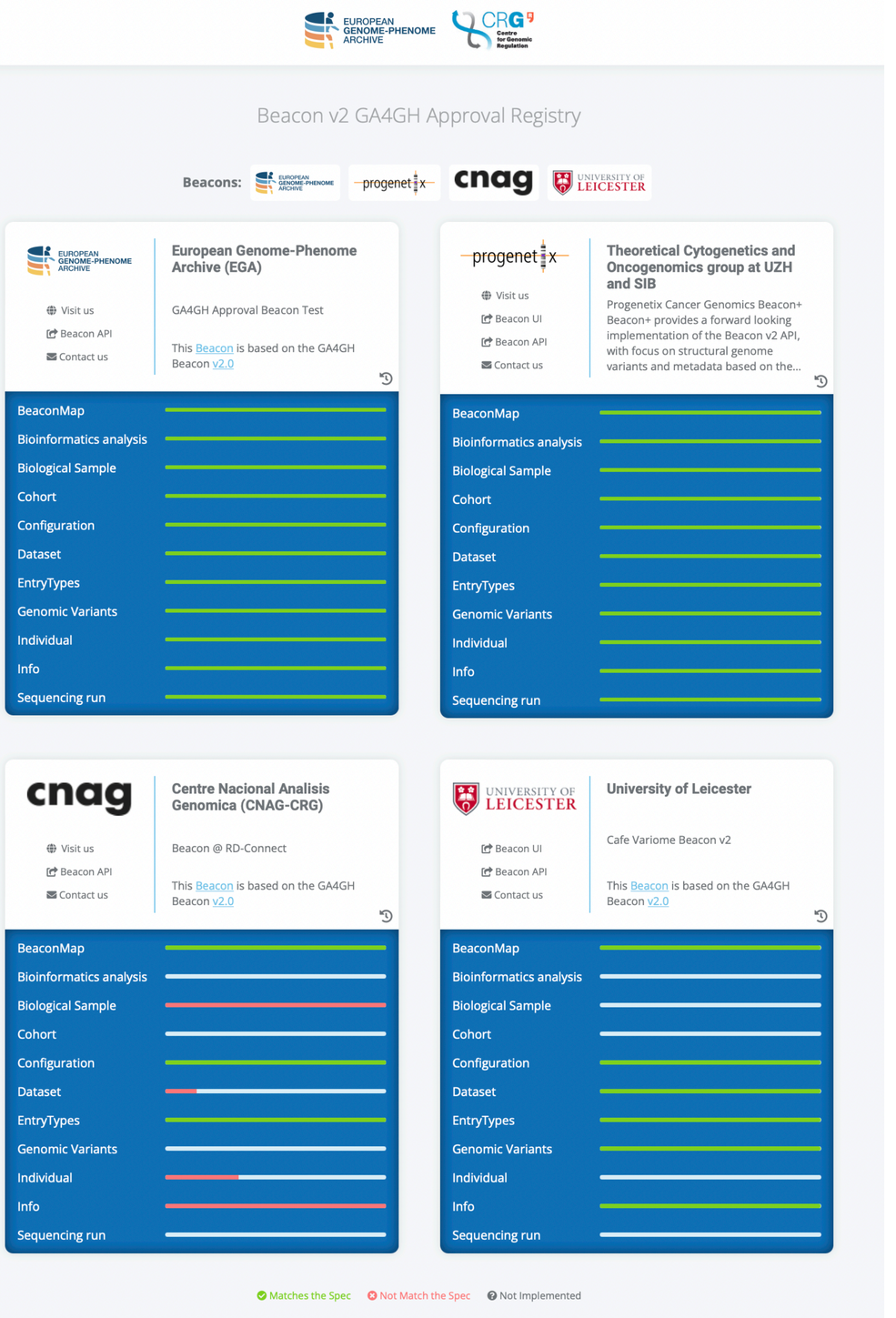
github.com/progenetix/bycon/

ELIXIR ALL HANDS 2022

Onboarding

Demonstrating Compliance

- Progenetix Beacon+ has served as implementation driver since 2016
- Beacon v2 as service with protocol-driven registries for federation
- GA4GH approved Beacon v2 in April 2022



Global Alliance for Genomics & Health

Beacon v2 Conformity and Extensions in Progenetix

Putting the + into Beacon ...

- support & use of standard Beacon v2 PUT & GET variant queries, filters and meta parameters
  - ➡ variant parameters, geneId, lengths, EFO & VCF CNV types, pagination
  - ➡ widespread, self-scoping filter use for bio-, technical- and and id parameters with switch for descending terms use (globally or per term if using POST)
- extensive use of handovers
  - ➡ asynchronous delivery of e.g. variant and sample data, data plots
- + optional use of OR logic for filter combinations (global)
- + extension of query parameters
  - ➡ geographic queries incl. $geonear and use of GeoJSON in schemas
- ┐(ツ ▽ ツ)┌ no implementation of authentication on this open dataset

Progenetix provides a number of additional services and output formats which are initiated over the / services path or provided as request parameters and are not considered Beacon extensions (though they follow the syntax where possible).

cc

progenetix

Beacon v2

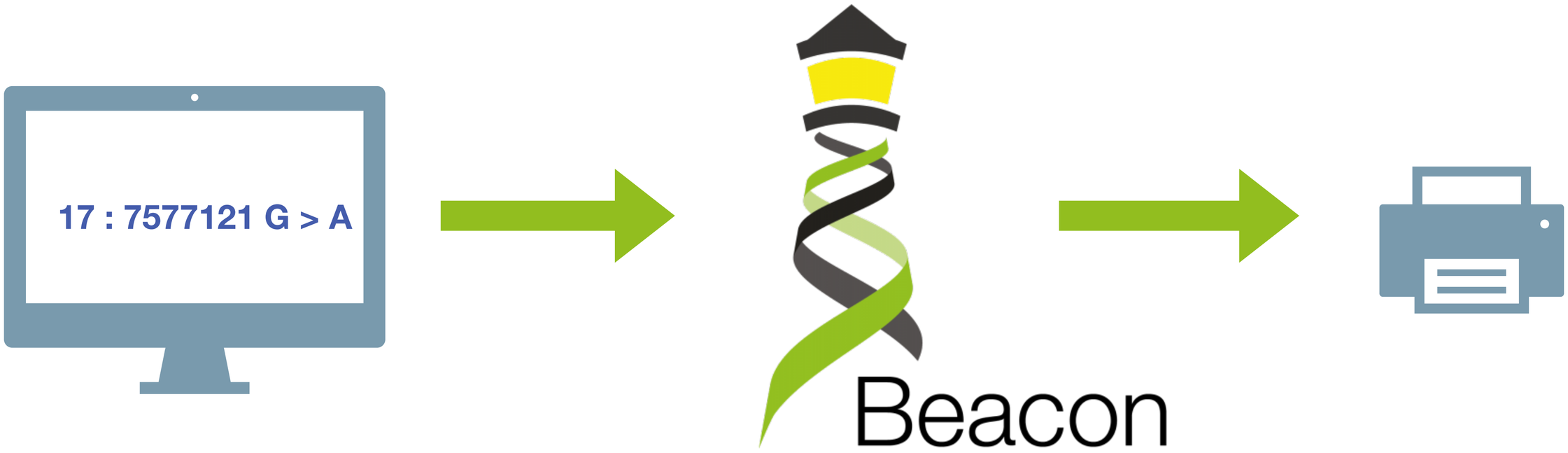
Federated Genomics

GA4GH 2020-2022 Strategic Roadmap

Global Alliance for Genomics & Health

Clin/Pheno Data Capture	Phenopackets	Pedigree Representation				
Cloud	Workflow Execution Service API	Tool Registry Service	Data Repository Service API	Task Execution Service	Cloud Testbed Interoperability	
Discovery	Beacon API	Service Info	Service Registry	Data Connect API		
Data Security	Authentication & Authorization Infrastructure	Data Security and Infrastructure Policy	Risk Assessment	Breach Response Protocol	Cloud Security & Privacy	
Data Use & Researcher Identity	Data Use Ontology	GA4GH Passports	Machine Readable Consent Guidance			
Genomic Knowledge Standards	Variation Representation	Variation Annotation				
Large Scale Genomics	htsget API	refget API	Read file formats	Variation File Formats	Crypt4GH	rnaget API	BED File Format
Regulatory & Ethics	Framework	Policy Initiatives	Responsible Data Sharing Resource Documents	GDPR Forum		

Ethics Review Recognition, Accountability, Consent, Privacy & Security Technical Standards & IP, Return of Results



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

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

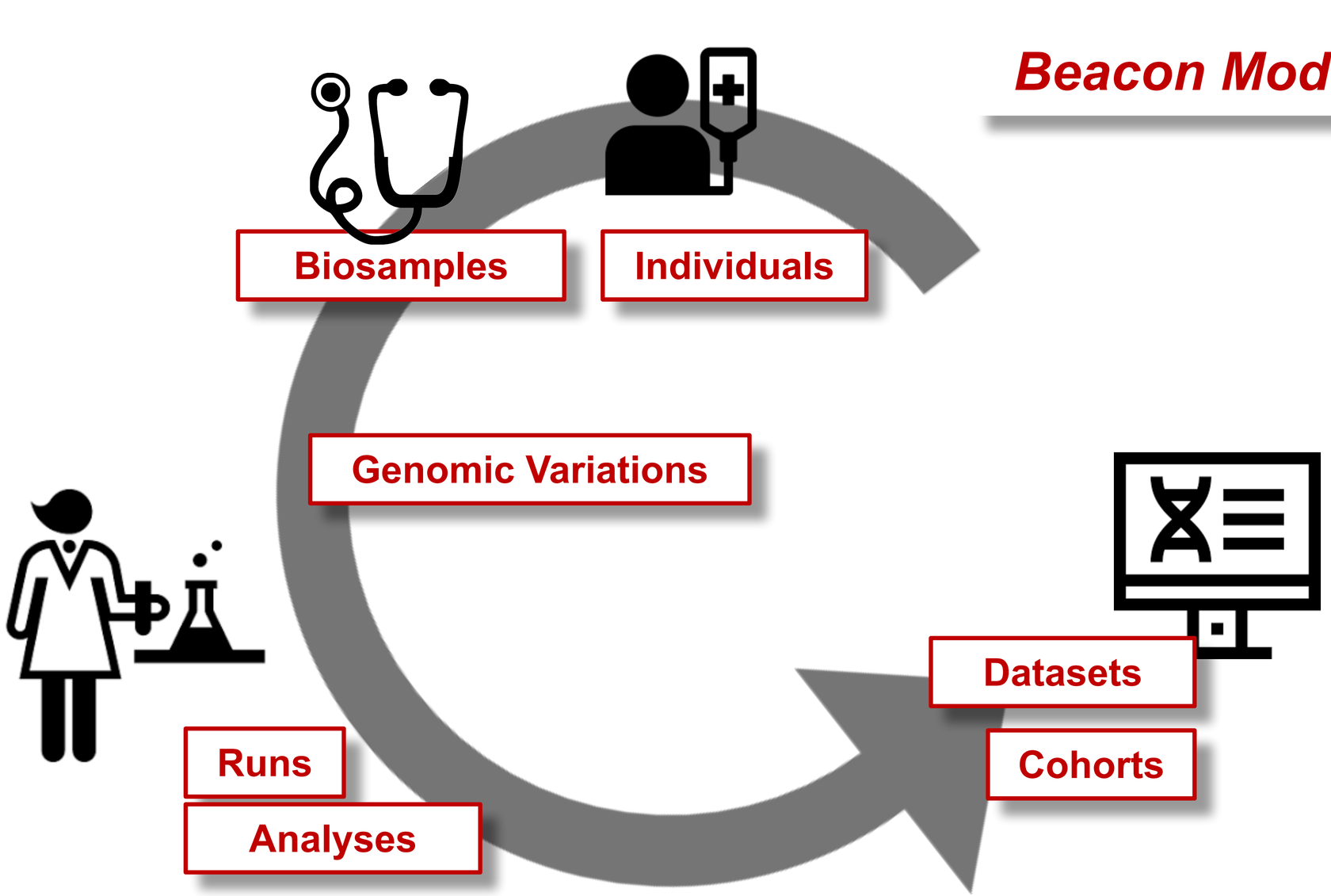
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

Beacon v2

docs.genomebeacons.org

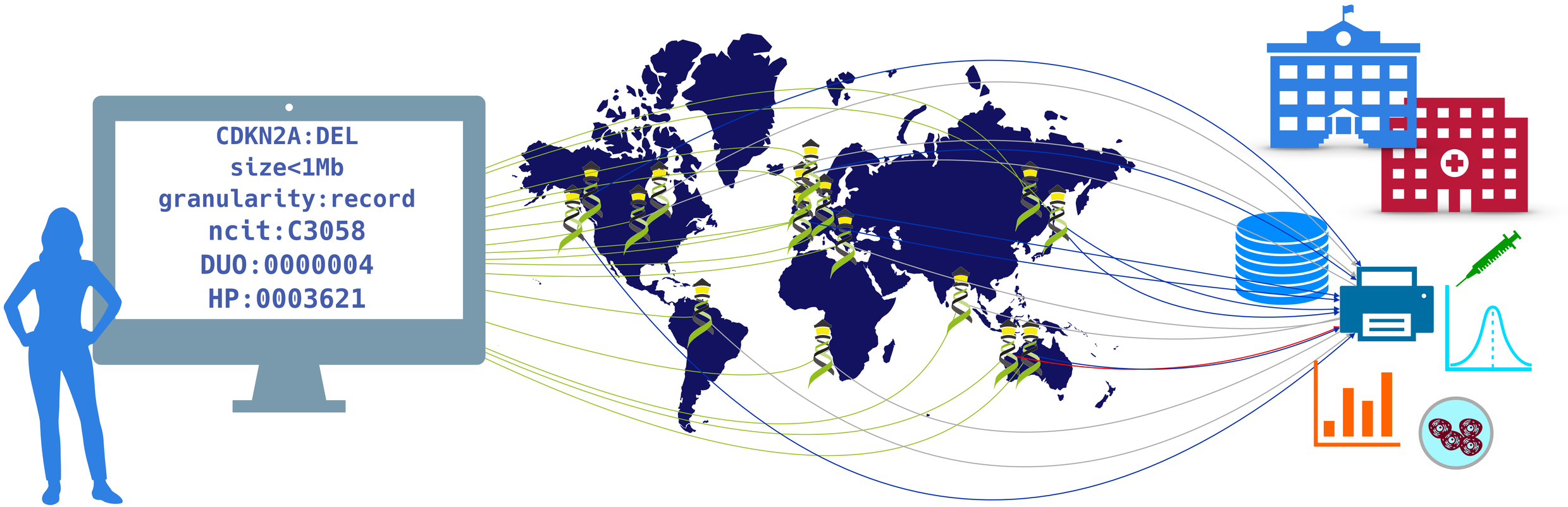


Beacon Model

Beacon Framework (protocol)

slide stolen from Jordi Rambla ...

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	**GA4GH founding event; Jim Ostell proposes Beacon concept with "more features… version 2"**		
2015	• beacon-network.org aggregator created by DNAstack • Beacon v0.3 release	• Beacon+ concept implemented on progenetix.org	• ELIXIR starts Beacon project support
2016	• work on queries for structural variants (brackets for fuzzy start and end parameters...)	• 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	• new Beacon website (March)
2019	• ELIXIR Beacon Network	• adopting "handover" concept • "Scouts" teams working on different aspects - filters, genomic variants, compliance ... • discussions w/ clinical stakeholders	• Beacon publication at Nature Biotechnology
2020		• framework + models concept implemented	
2021		• range and bracket queries, variant length • starting of GA4GH review process	
2022		• further changes esp. in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • **Beacon v2 approved at Apr GA4GH Connect**	• *docs.genomebeacons.org*



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

Beacon *v2* API

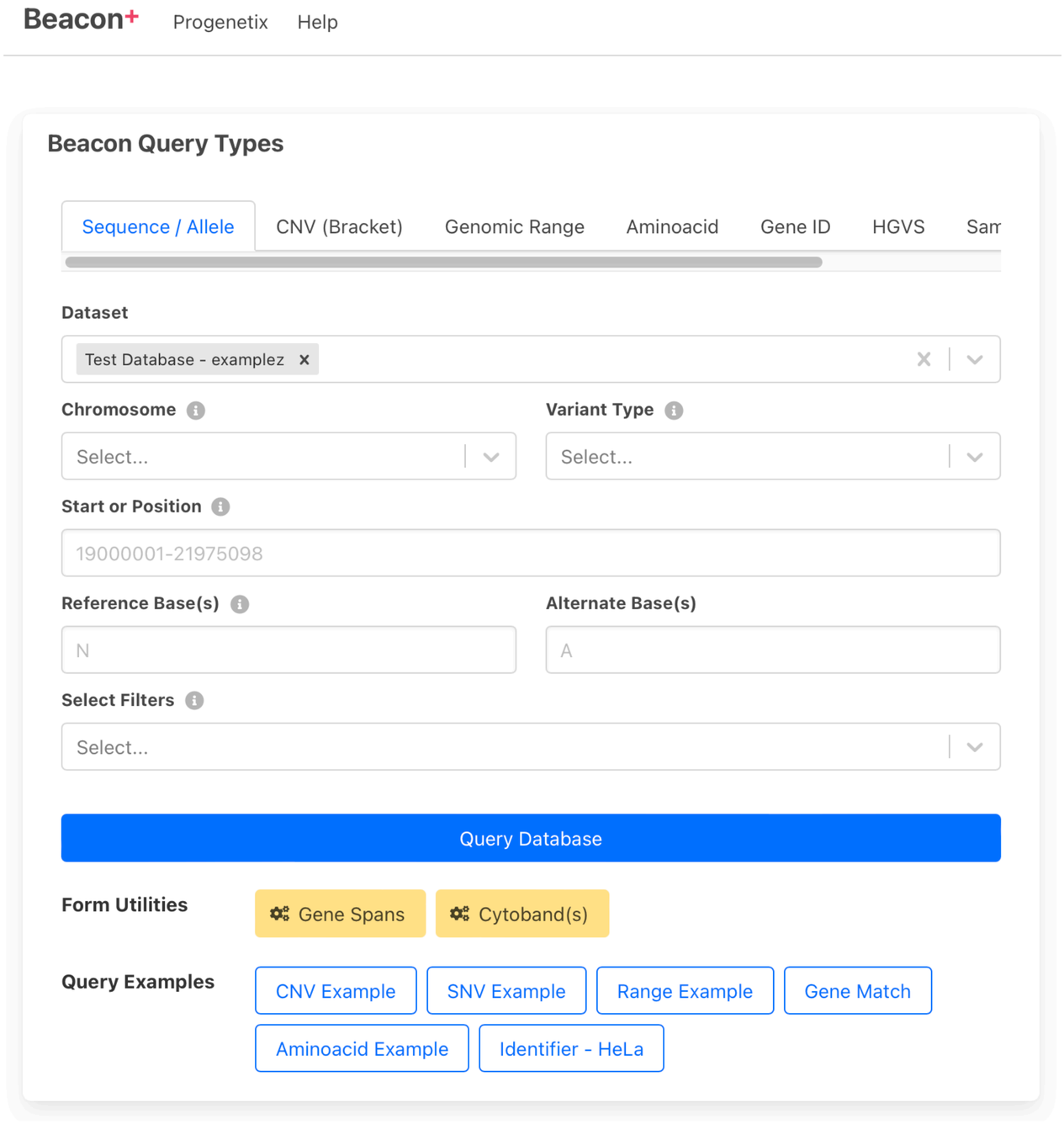
The Beacon API v2 represents a simple but powerful **genomics API** for ***federated*** data discovery and retrieval

Beacon Queries

Implementation of Current Options

- (so far) the Beacon model does not define explicit query types
- disambiguation of parameters is left to implementers
- implicit query types:
  - ➡ allele/sequence query
  - ➡ range query, w/ or w/o additional parameters
  - ➡ bracket query (e.g. sized CNVs)
  - ➡ aminoacid, HGVS, gene

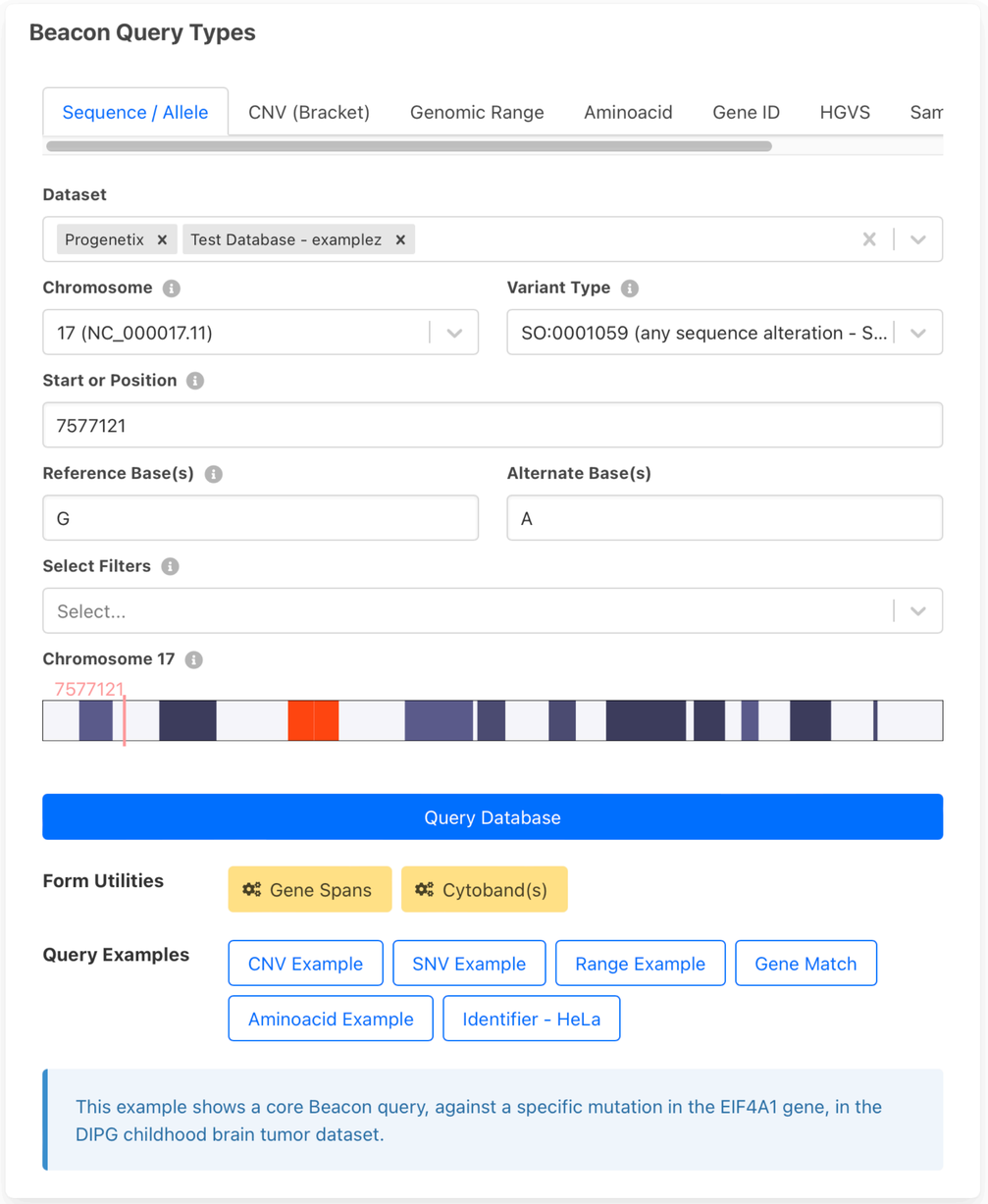
beaconplus.progenetix.org



Beacon Queries

Implementation of Current Options

- (so far) the Beacon model does not define explicit query types
- disambiguation of parameters is left to implementers
- implicit query types:
  - ➡ allele/sequence query
  - ➡ range query, w/ or w/o additional parameters
  - ➡ bracket query (e.g. sized CNVs)
  - ➡ aminoacid, HGVS, gene

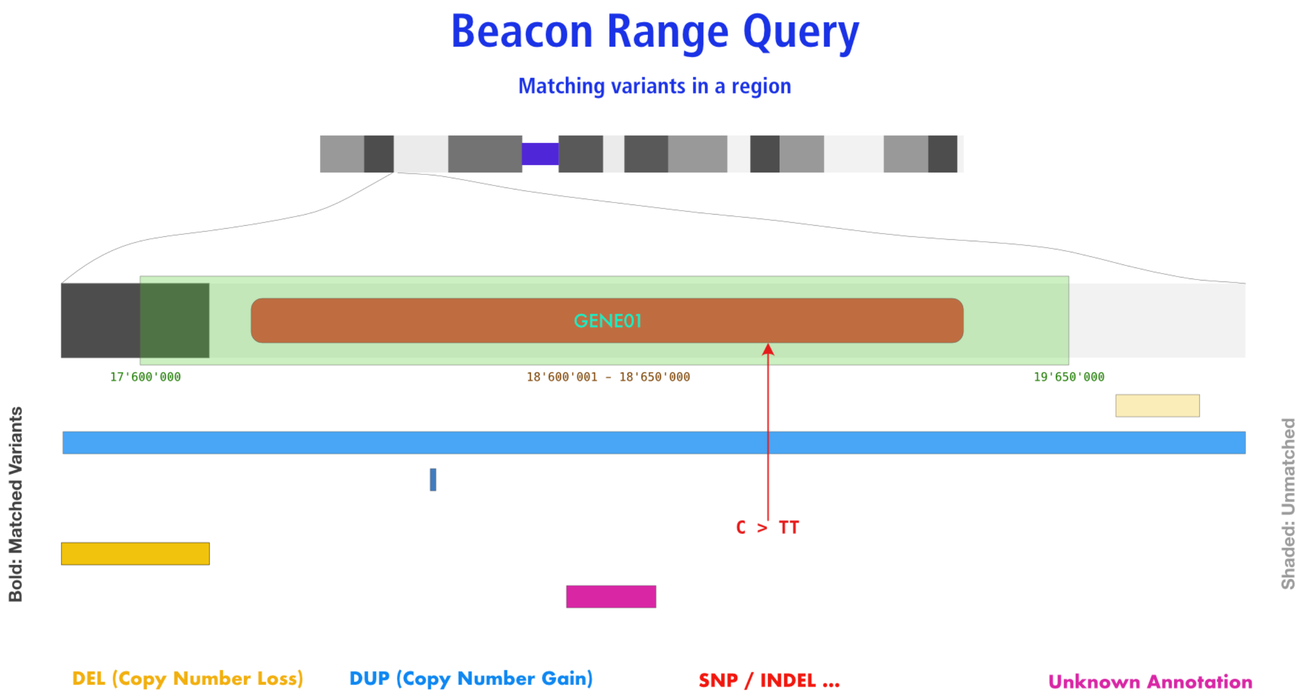


beaconplus.progenetix.org

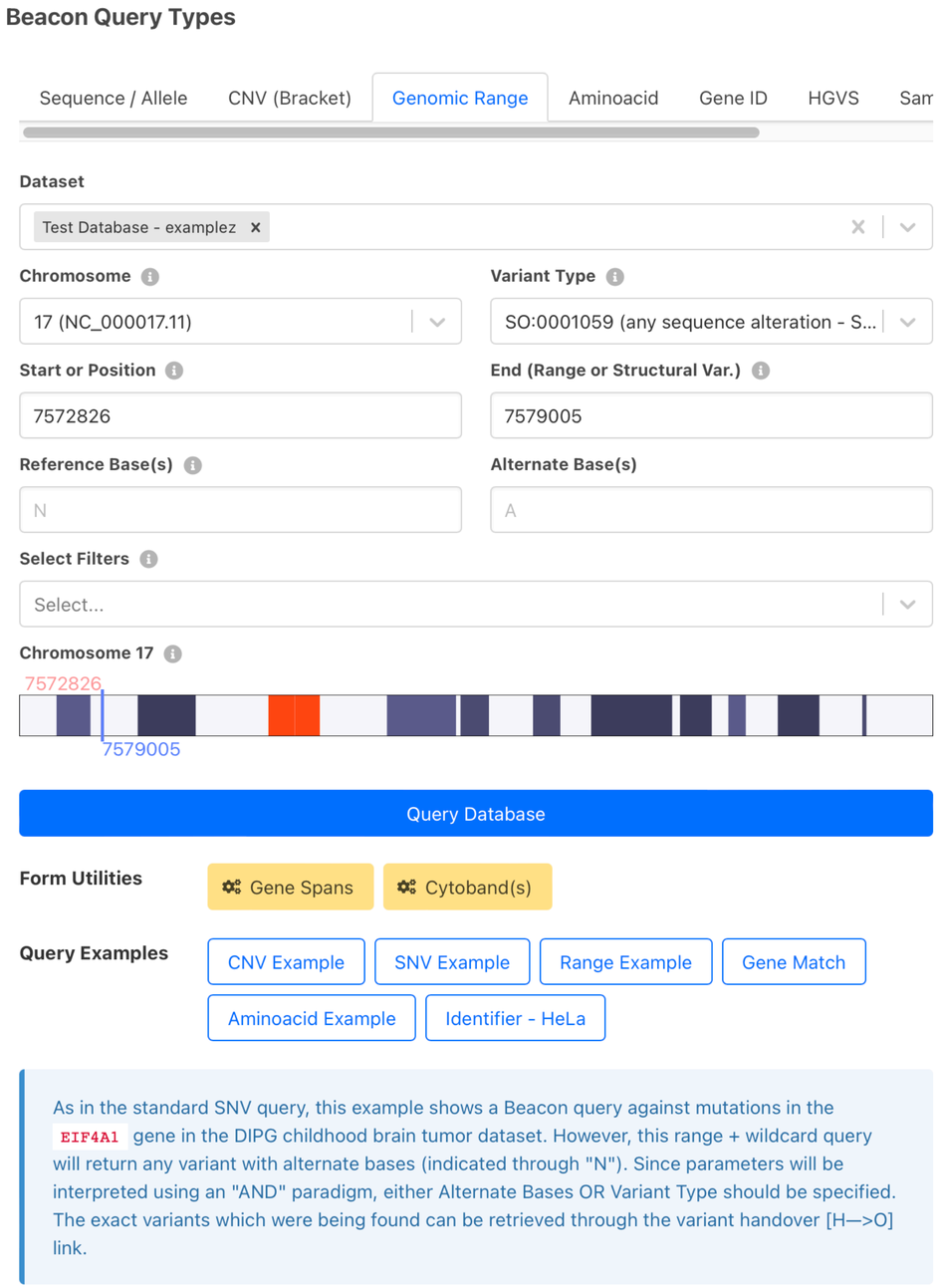
Beacon Queries

Range ("anything goes") Request

- defined through the use of 1 start, 1 end
- any variant... but can be limited by type etc.



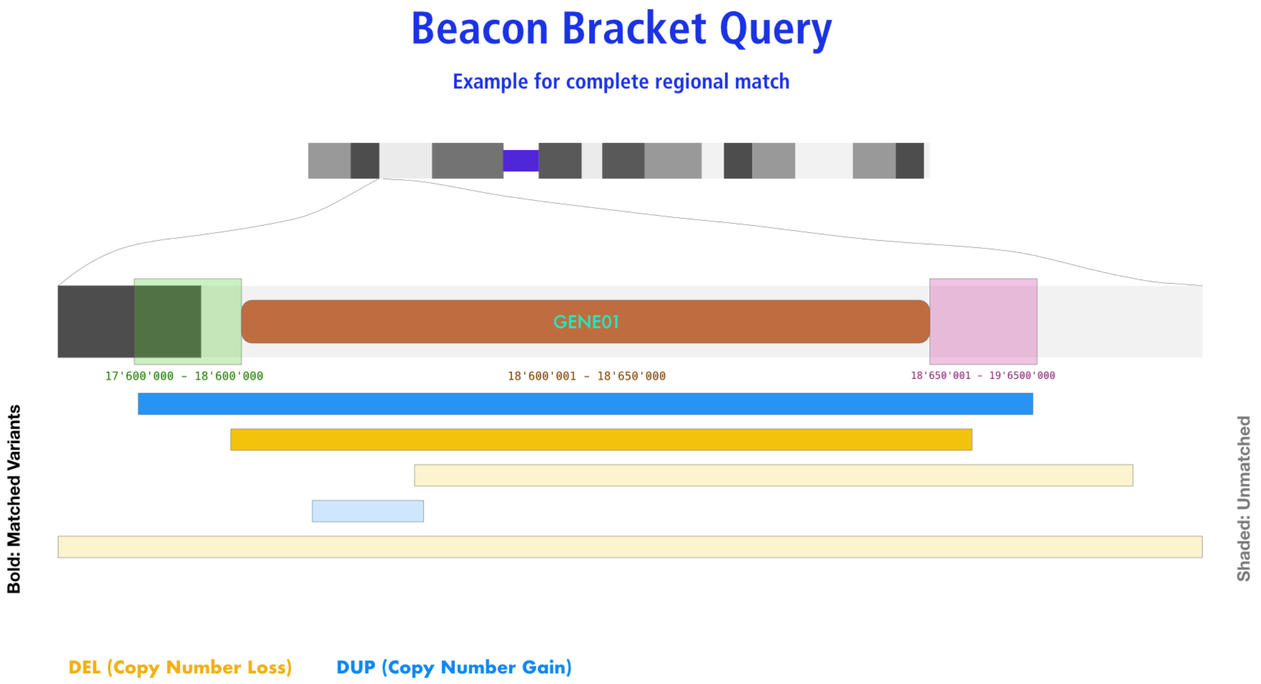
beaconplus.progenetix.org

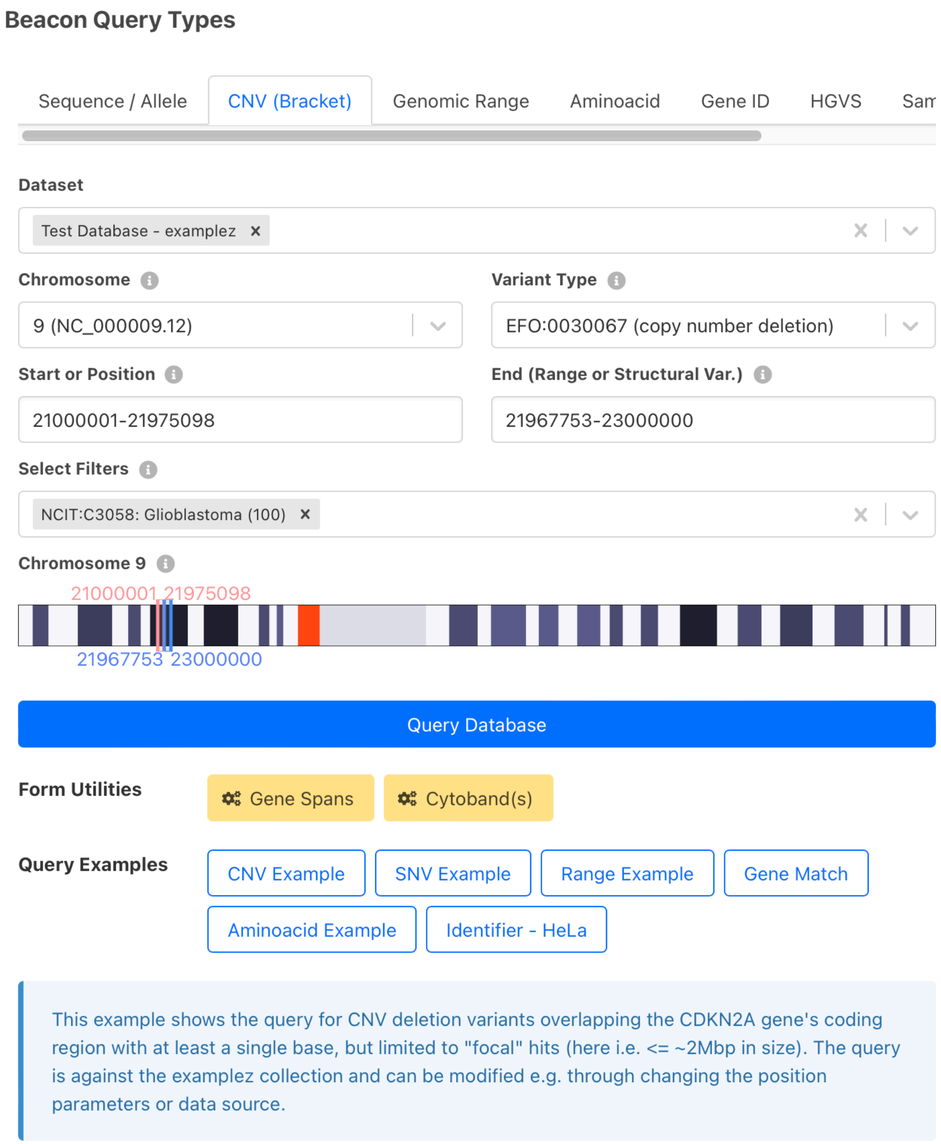


Beacon Queries

Bracket ("CNV") Query

- defined through the use of 2 start, 2 end
- any contiguous variant...



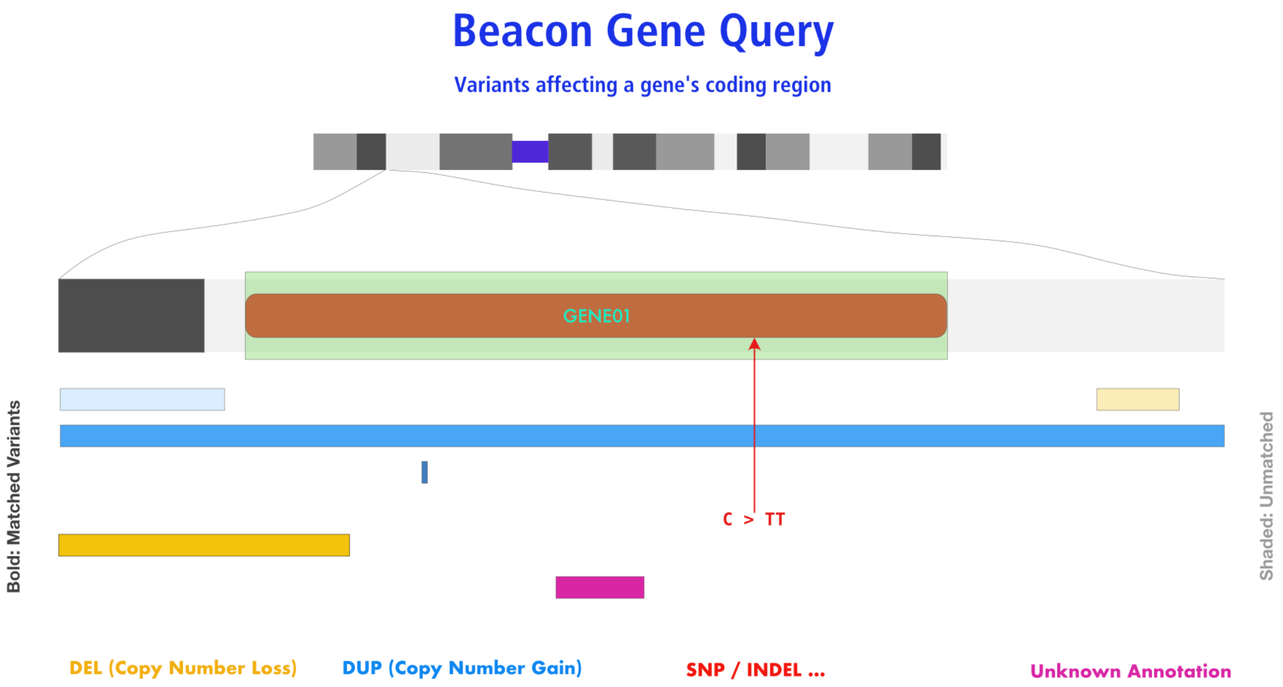


beaconplus.progenetix.org

Beacon Queries

Gene Request

- defined through a (HUGO) gene symbol
- assuming hit on the gene's CDR but YMMV



beaconplus.progenetix.org

Beacon Query Types

Sequence / Allele | CNV (Bracket) | Genomic Range | Aminoacid | Gene ID | HGVS | Sam

Dataset

Cancer Cell Lines Collection ×

Gene Symbol

CDK2 (12:55966830-55972789)

Variant Type

Select...

Min Variant Length | **Max Variant Length** | **Alternate Base(s)**

A

Select Filters

Select...

Query Database

Form Utilities Gene Spans | Cytoband(s)

Query Examples CNV Example | SNV Example | Range Example | Gene Match | Aminoacid Example | Identifier - HeLa

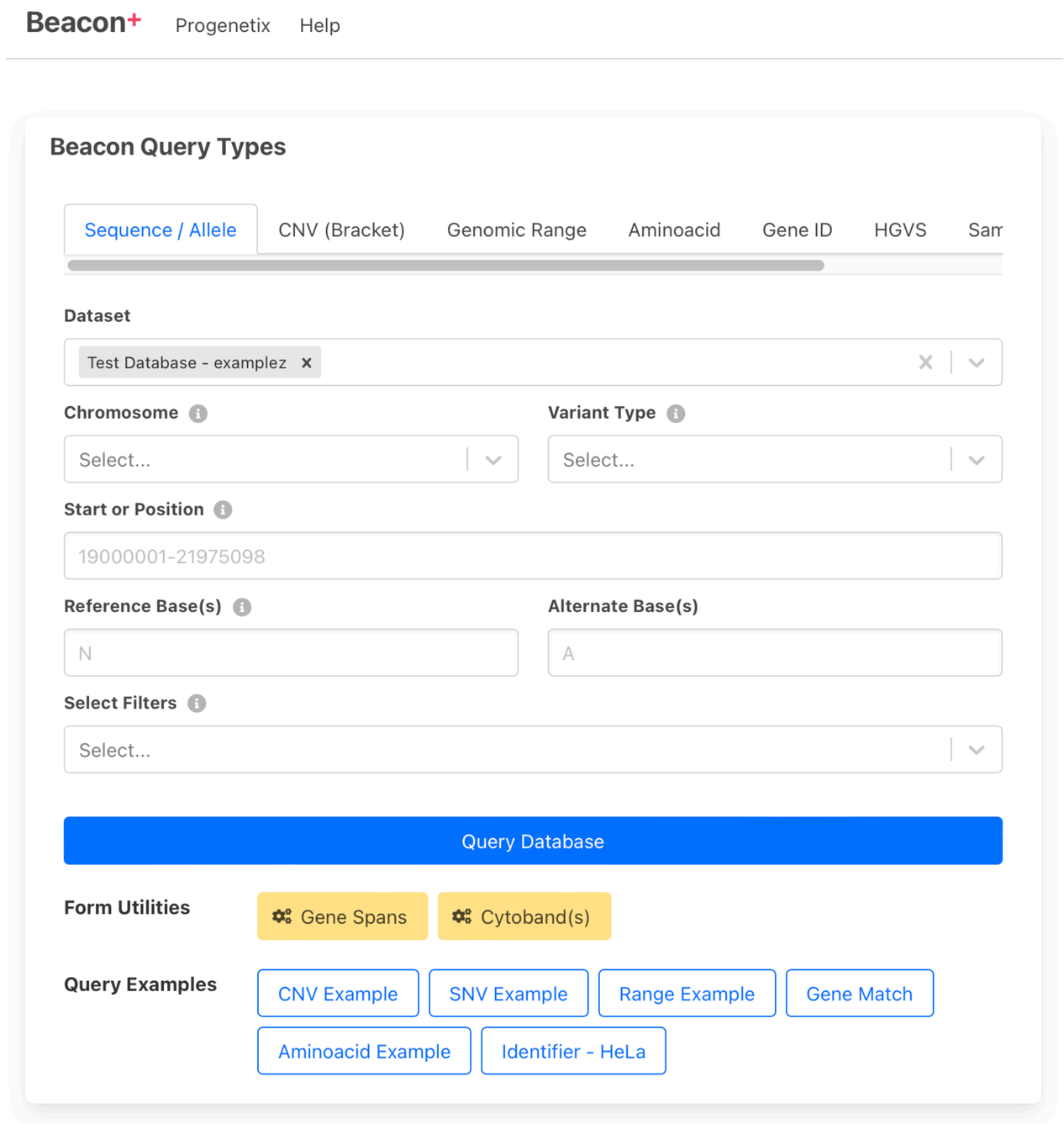
Beacons in v2 can support the discovery of variants with overlap with the genomic location of a gene, indicated by its symbol (e.g. `CDK2`). Additional parameters can *optionally* be used to make matches more specific:

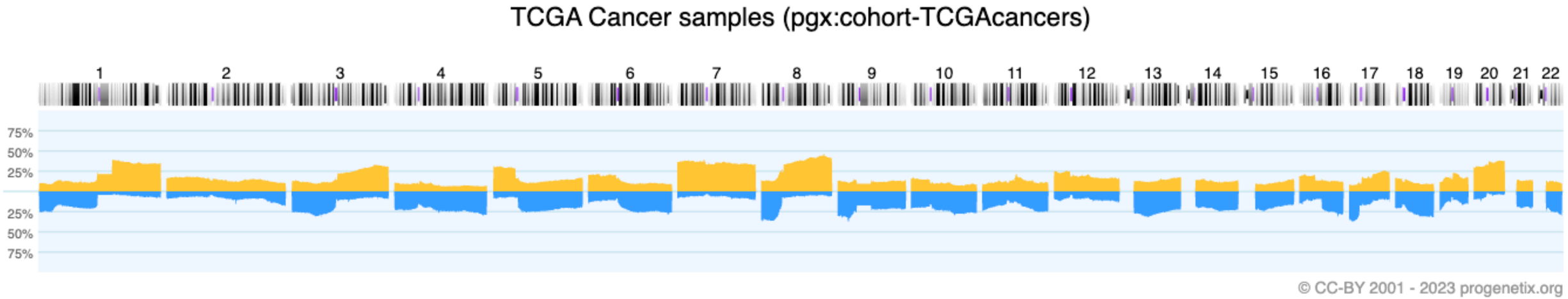
- `variantMinLength` and `variantMaxLength` to limit matched CNV sizes
- `genomicAlleleShortForm` (e.g. `V600E` with `BRAF`)
- `variantType` and `alternateBases` to specify variants

Beacon Queries

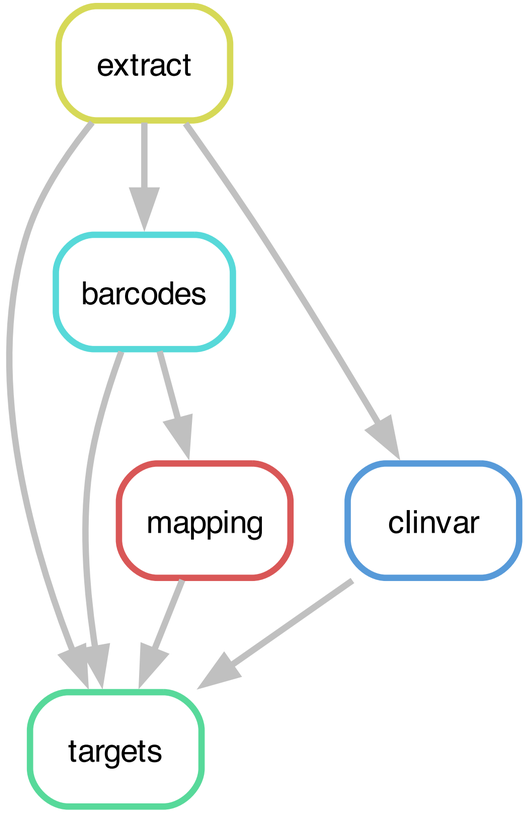
Missing or ill defined options

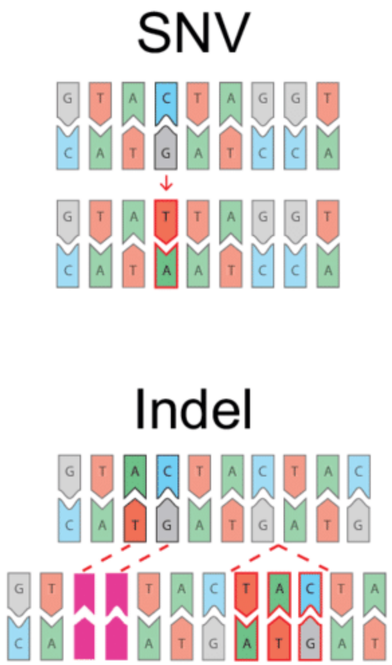
- **translocations** are in principle possible (start bracket with "referenceName" and end bracket with "mateName") but not yet documented / battle tested
- **functional elements**?
- exon hits beyond specifying individual ones by sequence
- tandem dups ...
- genomic **double hits**

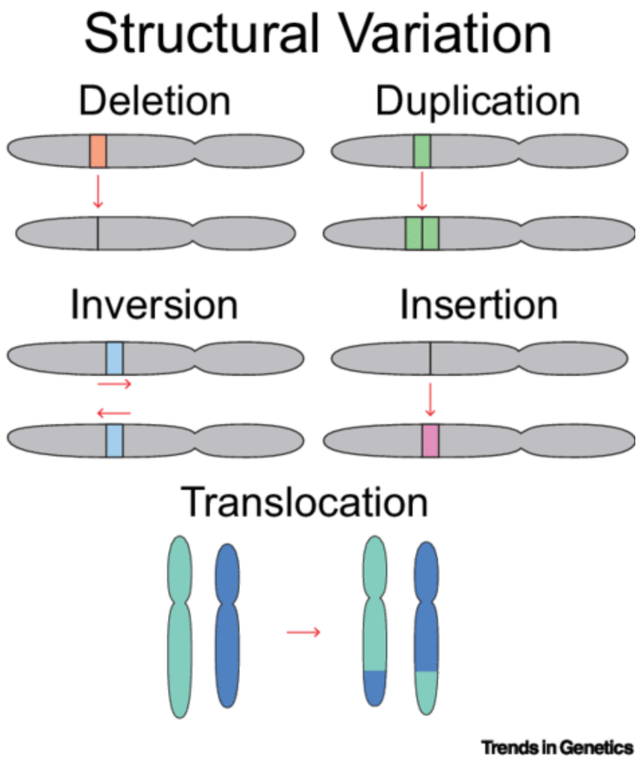



TCGA Cancer samples (pgx:cohort-TCGAcancers)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
75%
50%
25%
25%
50%
75%
© CC-BY 2001 - 2023 progenetix.org

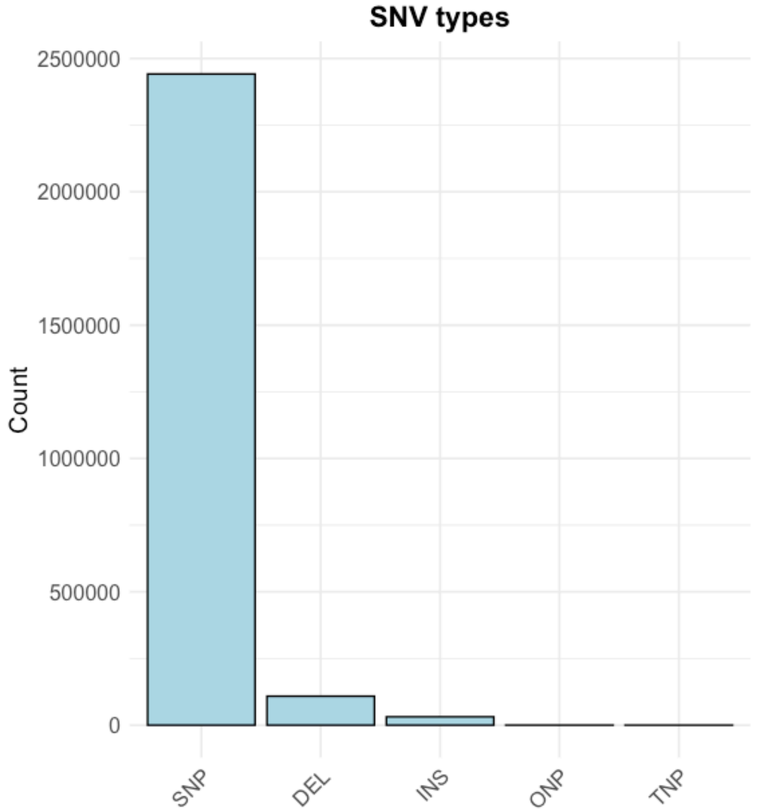
- **Integrating SNVs:**
  - Renewed interest due to technological advantages
  - Allow compound variant queries

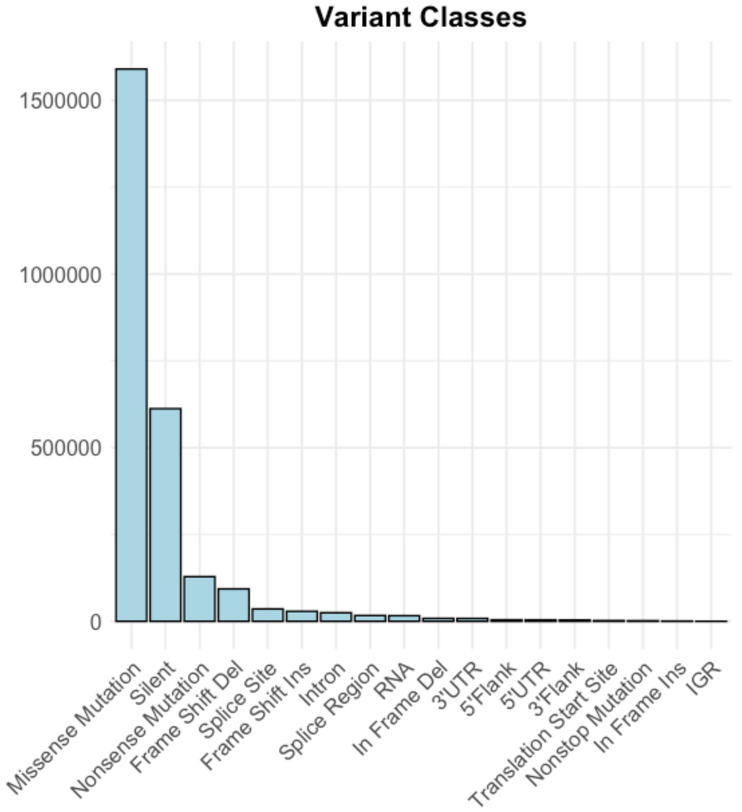


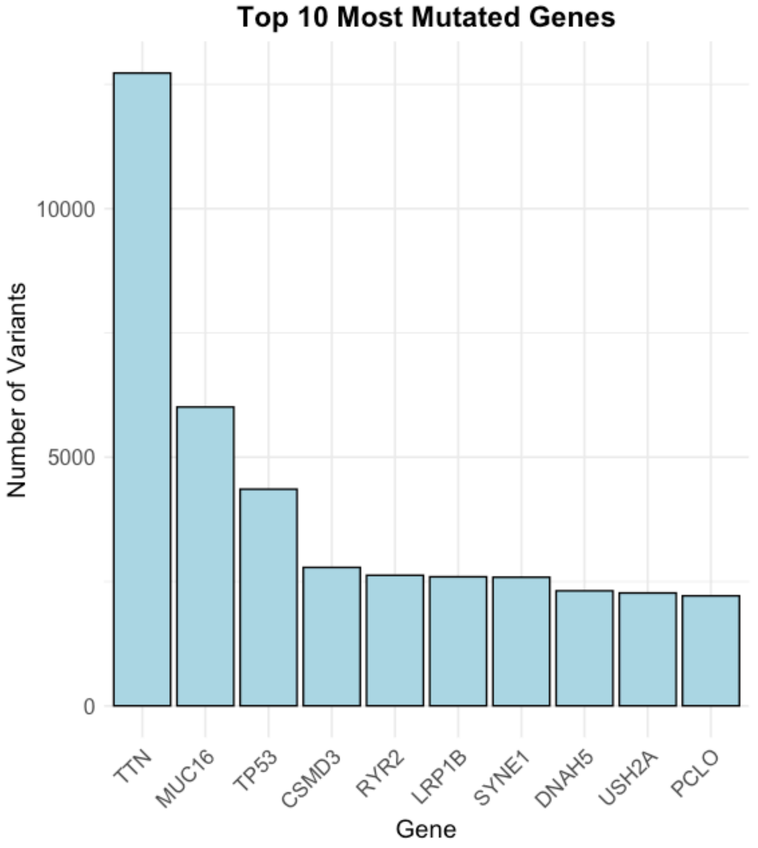




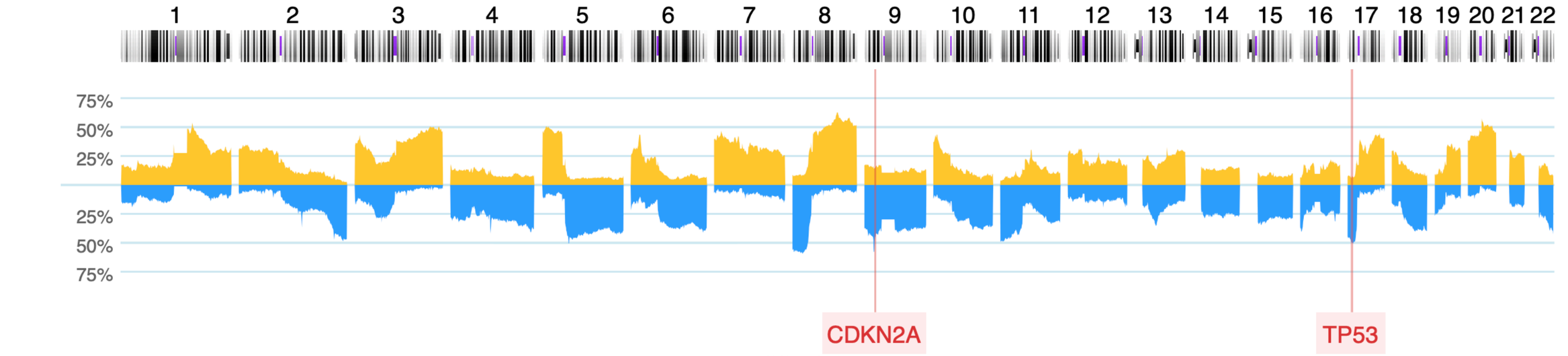
Nesta, Alex & Tafur, Denisse & Beck, Christine. (2020). Hotspc Mutation. Trends in Genetics

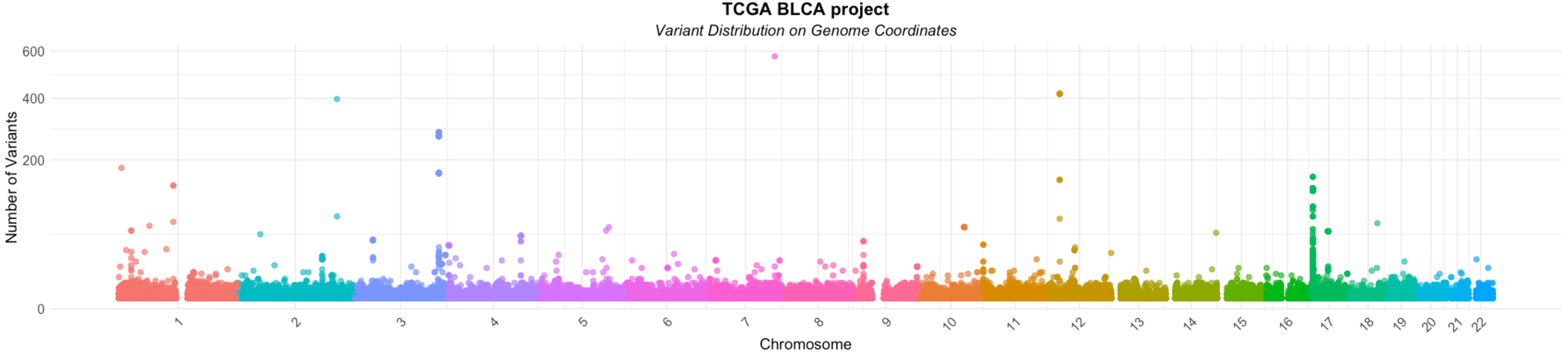

SNV types
Count
2500000
2000000
1500000
1000000
500000
0
SNP
DEL
INS
ONP
TNP


Variant Classes
1500000
1000000
500000
0
Missense Mutation
Silent
Nonsense Mutation
Frame Shift Del
Splice Site
Frame Shift Ins
Intron
Splice Region
RNA
In Frame Del
3'UTR
5'Flank
5'UTR
3'Flank
Translation Start Site
Nonstop Mutation
In Frame Ins
IGR


Top 10 Most Mutated Genes
Number of Variants
10000
5000
0
TTN
MUC16
TP53
CSMD3
RYR2
LRP1B
SYNE1
DNAH5
USH2A
PCLO
Gene

TCGA BLCA project (pgx:TCGA.BLCA)





Beacon Queries

Missing or ill defined options

- **translocations** are in principle possible (start bracket with "referenceName" and end bracket with "mateName") but not yet documented / battle tested
- **functional elements**?
- exon hits beyond specifying individual ones by sequence
- tandem dups ...
- genomic **double hits**

➡ **Beacon & hCNV Scout Team**