



progenet X

A Genomics / CNV Resource Built on GA4GH Standards

... also Implementation Driven Standards Development

Michael Baudis | hCNV Community Meeting 2023



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



Cancer CNV Profiles

ICD-O Morphologies
ICD-O Organ Sites
Cancer Cell Lines
Clinical Categories

Search Samples

arrayMap
TCGA Samples
1000 Genomes
Reference Samples
DIPG Samples
cBioPortal Studies
Gao & Baudis, 2021

Publication DB

Genome Profiling
Progenetix Use

Services

NCIt Mappings
UBERON Mappings

Upload & Plot

Beacon⁺

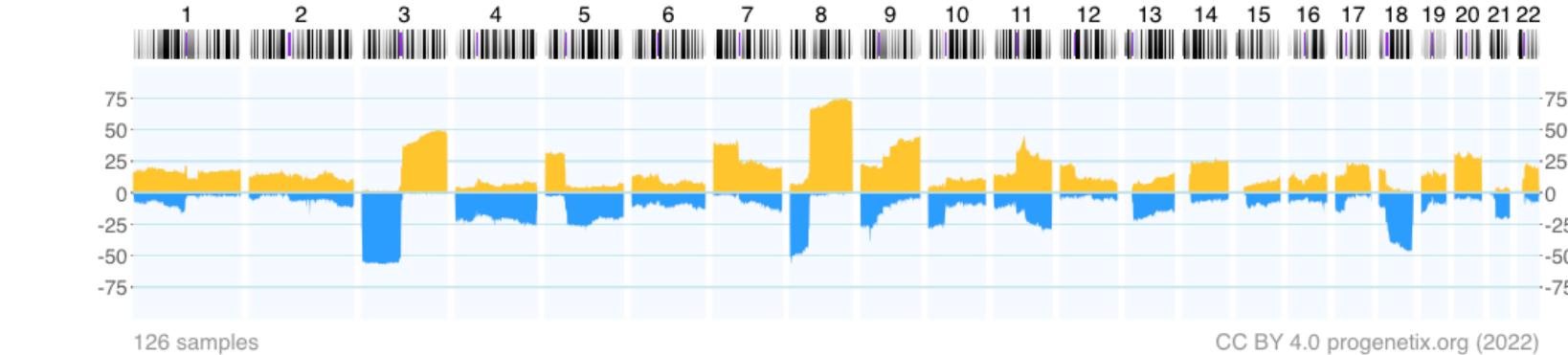
Documentation
News
Downloads & Use
Cases
Sevices & API

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 **142063** samples.

Floor of the Mouth Neoplasm (NCIT:C4401)



[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

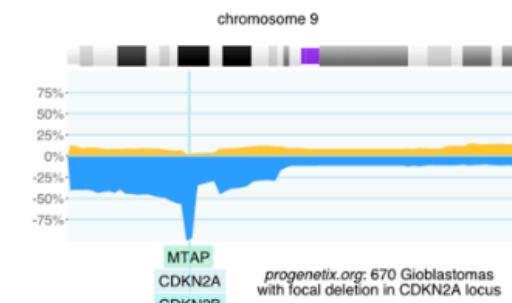
Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.

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 **834** 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 **4164** 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.

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

The screenshot shows the homepage of cancercelllines.org. At the top is a pink header bar. Below it is a navigation menu with the following items: Cancer Cell Lines (with a red circular icon), Search Cell Lines, Cell Line Listing, CNV Profiles by Cancer Type, Documentation, News, Progenetix (which is highlighted in grey), and Baudisgroup @ UZH.

cancercellines.org

Cancer Cell Lines by Cellosaurus ID

The cancer cell lines in [cancercelllines.org](#) are labeled by their parentage hierarchically: Daughter cell lines are displayed below the primary cell line. For example, HeLa is listed as a daughter cell line of **HeLa (CVCL_0030)** and so forth.

Sample selection follows a hierarchical system in which samples are retrieved based on the selected cell line. For example, selecting "HOS" for HeLa will also return the daughter lines by default - but one can also select individual daughter lines.

Cell Lines (with parental/derived hierarchies)

Filter subsets e.g. by prefix

Hierarchy Depth:

No Selection

- > cellosaurus:CVCL_0312: HOS (204 samples)
- > cellosaurus:CVCL_1575: NCI-H650 (6 samples)
- > cellosaurus:CVCL_1783: UM-UC-3 (9 samples)
- > cellosaurus:CVCL_0004: K-562 (28 samples)
- cellosaurus:CVCL_3827: K562/Adr (1 sample)
- > cellosaurus:CVCL_0589: Kasumi-1 (9 samples)
- > cellosaurus:CVCL_XK00: M397 (2 samples)
- > cellosaurus:CVCL_1650: Reh (11 samples)
- cellosaurus:CVCL_8857: EU-1 (1 sample)
- cellosaurus:CVCL_0011: KM-3 (1 sample)
- cellosaurus:CVCL_8462: NOI-90 (1 sample)
- cellosaurus:CVCL_ZV66: Reh/EphA2 (1 sample)
- cellosaurus:CVCL_A049: WSU-CLL (1 sample)
- > cellosaurus:CVCL_2063: HCC827 (27 samples)

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-63ce6a903319d2172d2

Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

- Cellosaurus - a knowledge resource on cell lines [cellosaurus:CVCL_0312](#) ↗

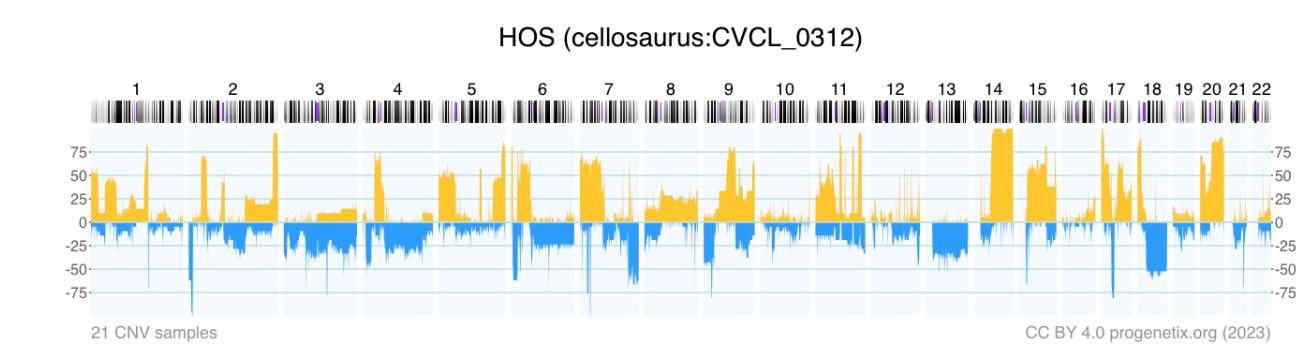
Sample Counts

- 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](#)

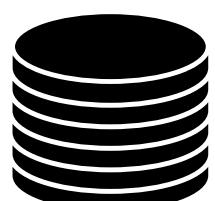
Gene Matches Cytoband Matches Variants

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

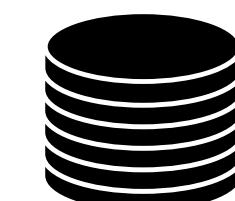
Progenetix Stack



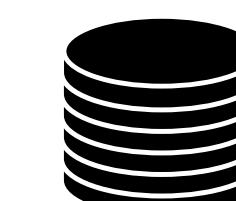
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the **bycon** package
 - ▶ schemas, 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



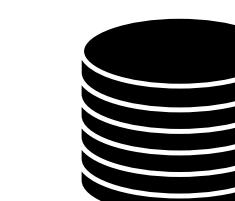
variants



analyses



biosamples

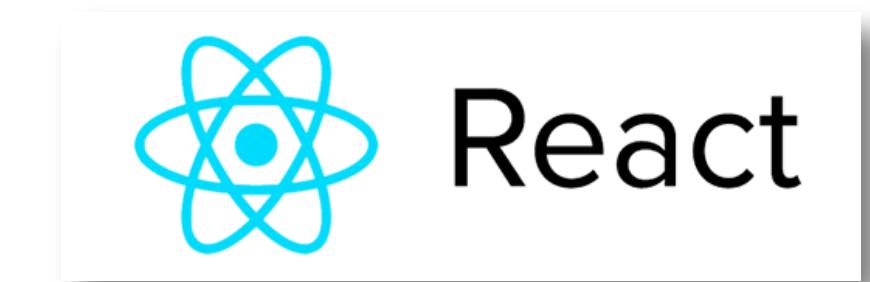


individuals



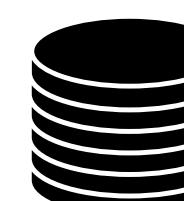
github.com/progenetix/bycon/

Entity collections

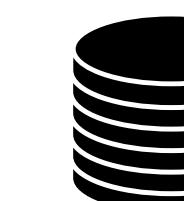


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and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

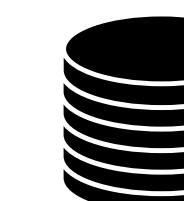
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]
```



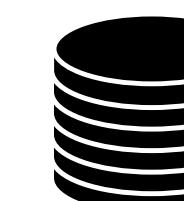
collations



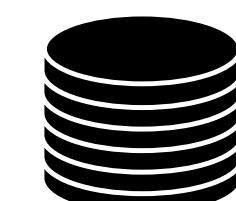
geolocs



genespans



publications



qBuffer

Utility collections



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

Beacon v2 GA4GH Approval Registry

Beacons:  European Genome-Phenome Archive |  progenetix |  cnag |  UNIVERSITY OF LEICESTER

 European Genome-Phenome Archive (EGA)

[Visit us](#) [Beacon API](#) [Contact us](#)

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

 progenetix+

[Visit us](#) [Beacon UI](#) [Beacon API](#) [Contact us](#)

Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

Progenetix Cancer Genomics Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the...

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

 Centre Nacional Analisis Genomica (CNAG-CRG)

[Visit us](#) [Beacon API](#) [Contact us](#)

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

 University of Leicester

[Beacon UI](#) [Beacon API](#) [Contact us](#)

Cafe Variome Beacon v2

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

✓ Matches the Spec ✗ Not Match the Spec ⌚ Not implemented



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, geneld, lengths, EFO & VCF CNV types, pagination
 - ➡ widespread, self-scoping filter use for bio-, technical- 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
- ↵ ↴ ↲ ↳ 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).



Beacon v2

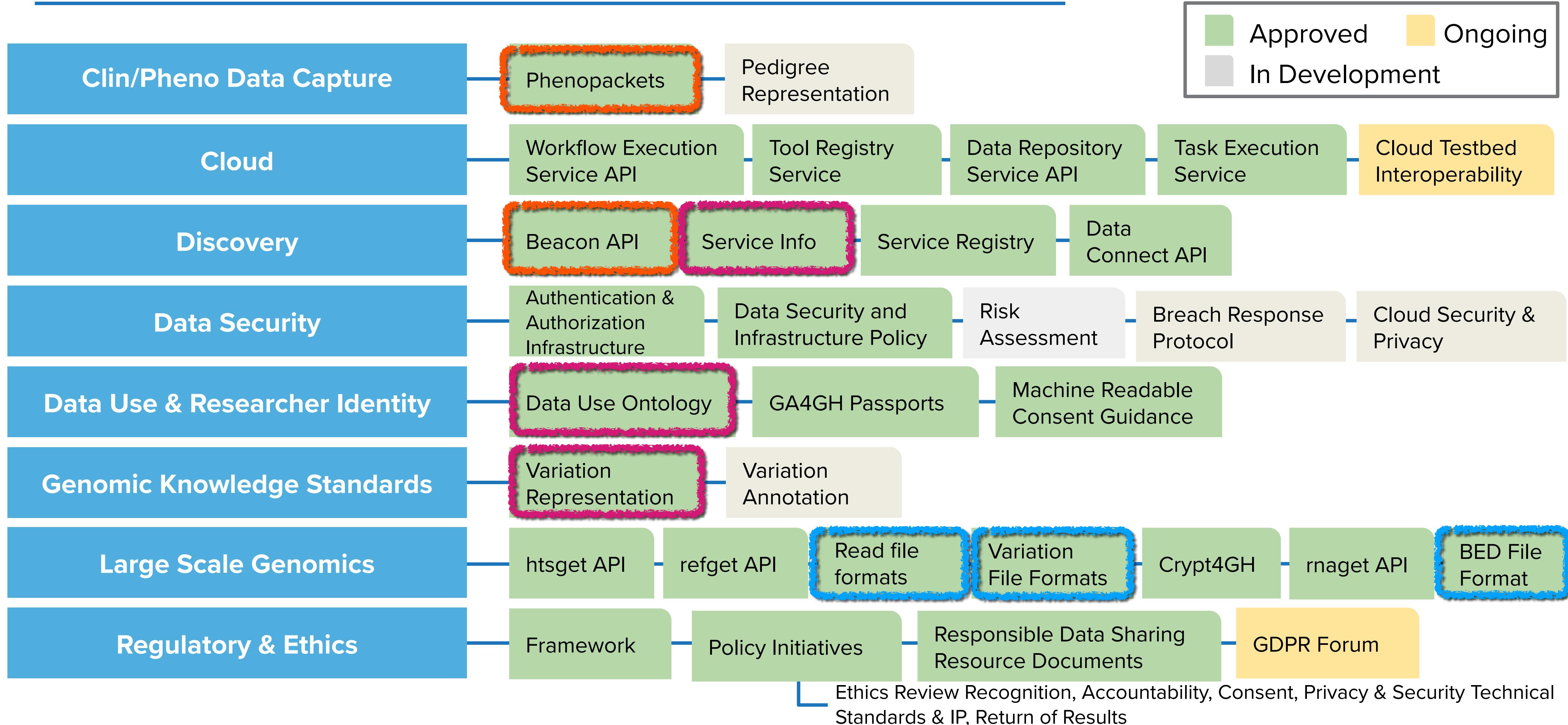
Federated Genomics

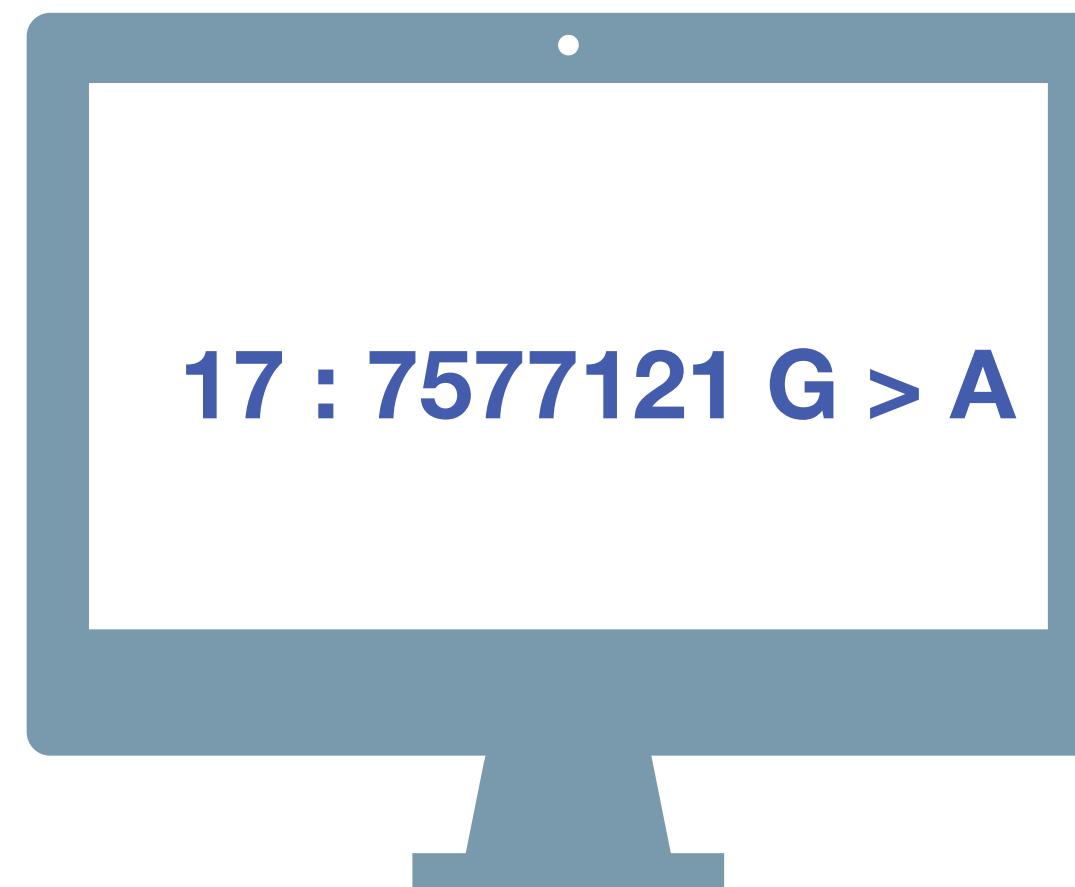


GA4GH 2020-2022 Strategic Roadmap



Global Alliance
for Genomics & Health

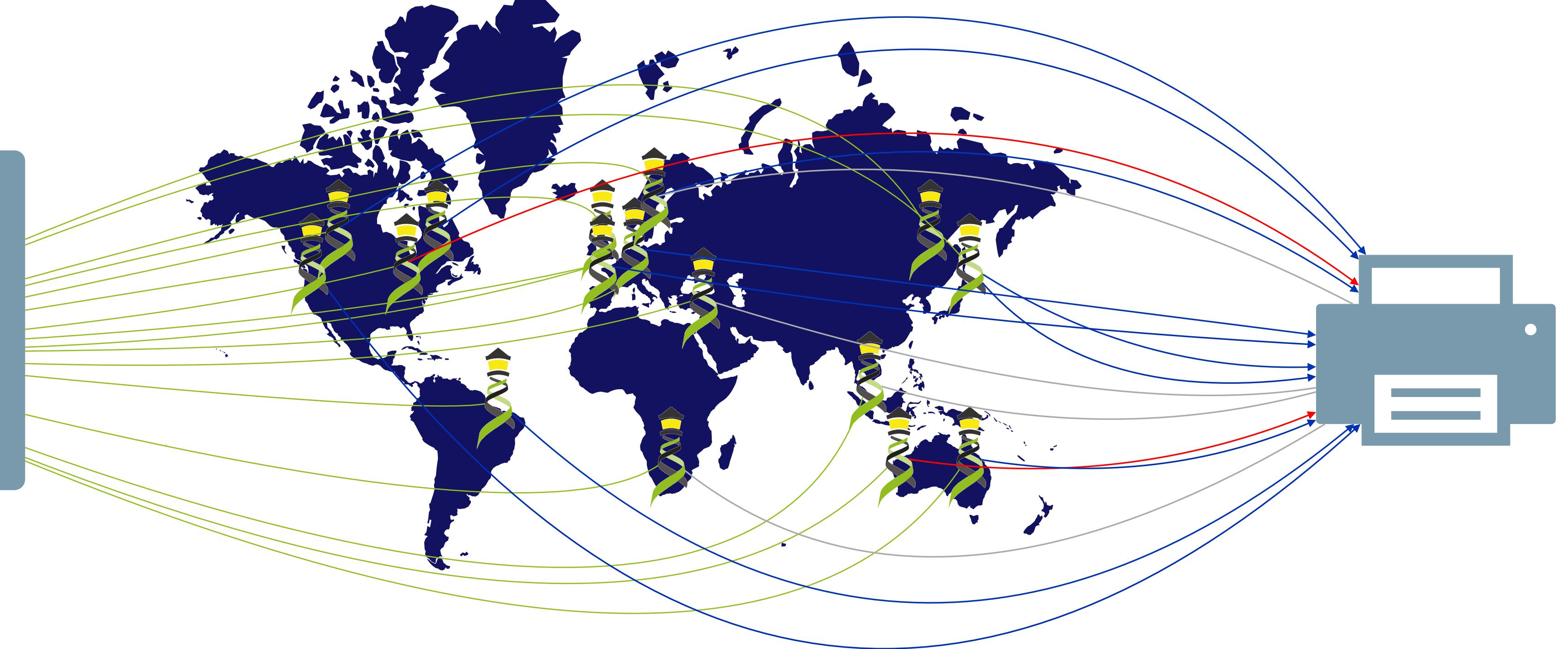
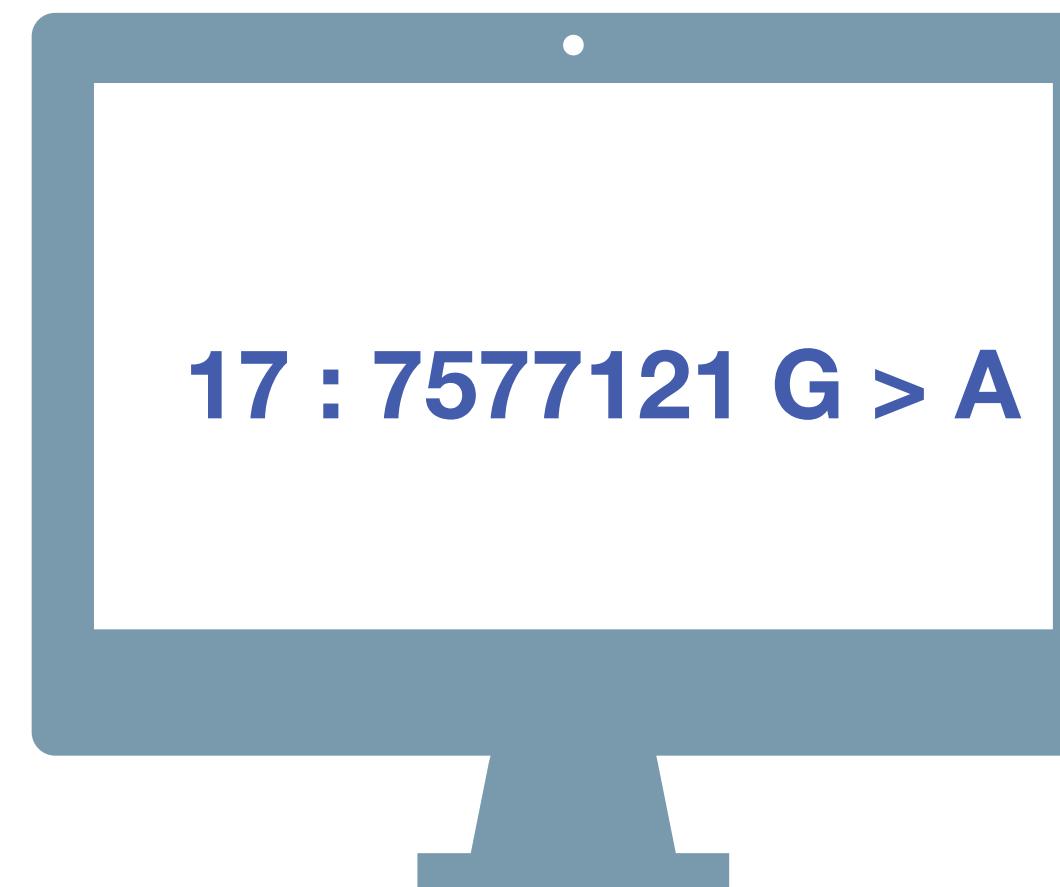




Beacon

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

YES | NO | \0



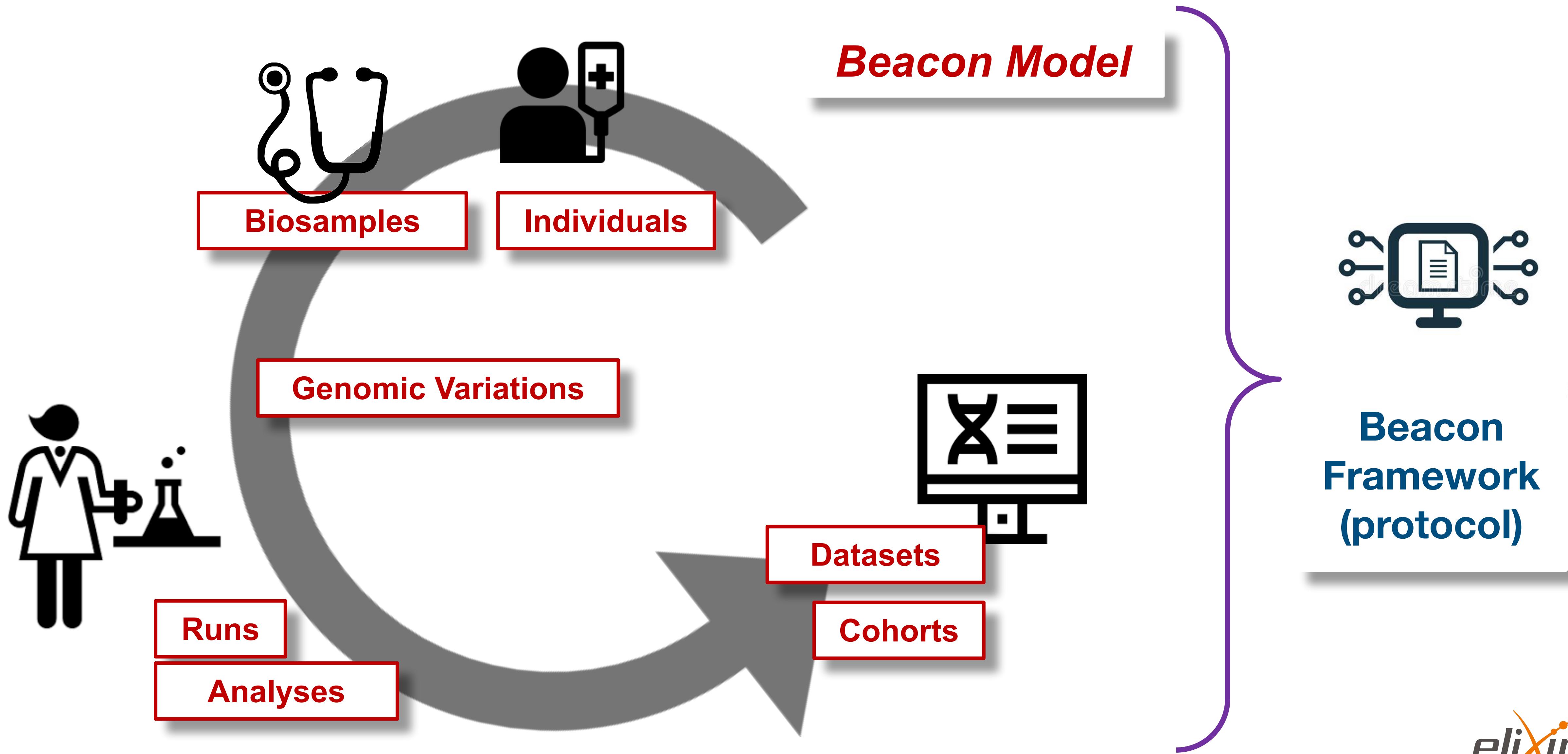
Have you seen this variant?
It came up in my patient
and we don't know if this is
a common SNP or worth
following up.

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 v1 Development

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

- work on queries for structural variants (brackets for fuzzy start and end parameters...)

• OpenAPI implementation

- integrating CNV parameters (e.g. "startMin, statMax")

• Beacon v0.4 release in January; feature release for GA4GH approval process

• GA4GH Beacon v1 approved at Oct plenary

2019 • ELIXIR Beacon Network

2020

2021

2022

Beacon v2 Development

Related ...

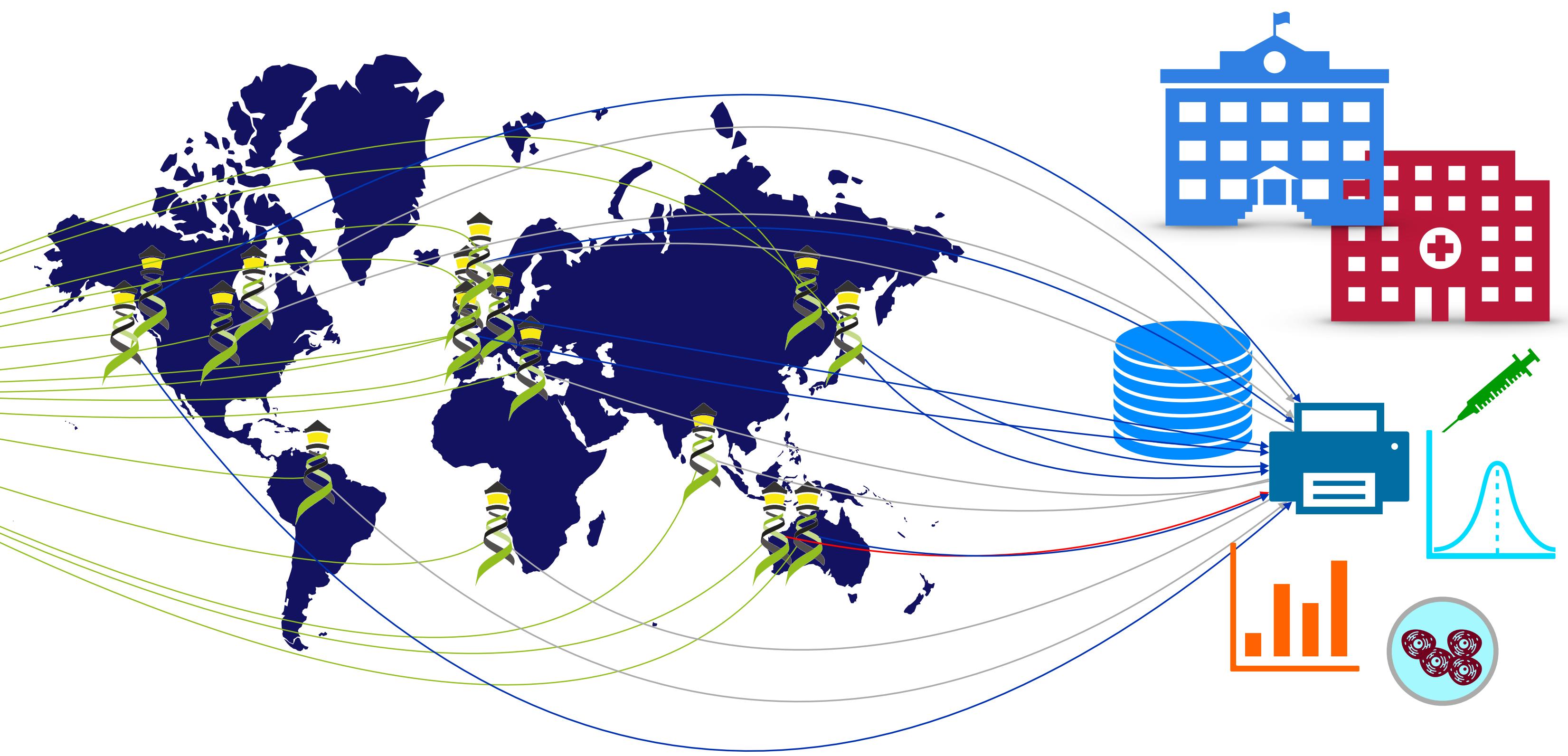
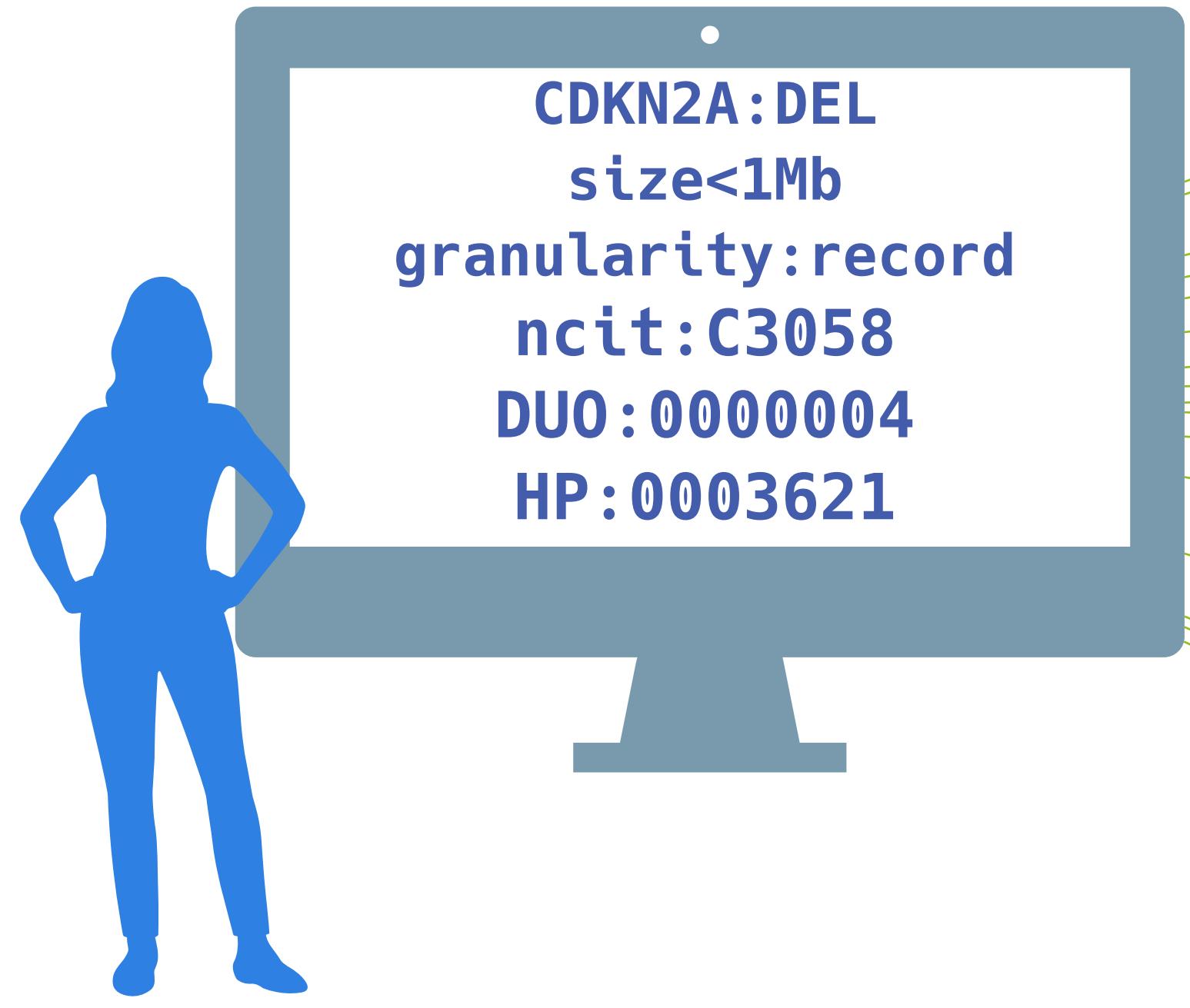
- ELIXIR starts Beacon project support

- GA4GH re-structuring (workstreams...)
- Beacon part of Discovery WS

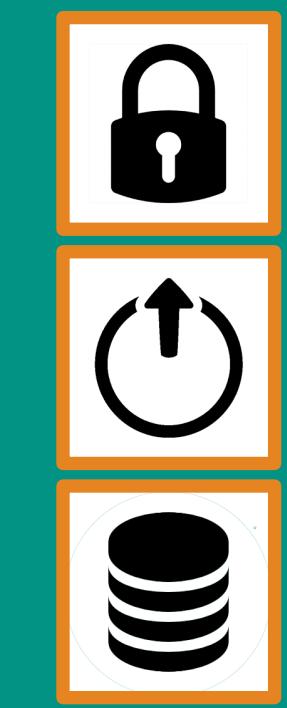
- new Beacon website (March)

- Beacon publication at Nature Biotechnology

- 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

Beacon+ Progenetix Help

Beacon Query Types

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

Dataset: Test Database - examplez

Chromosome: Select... Variant Type: Select...

Start or Position: 19000001-21975098

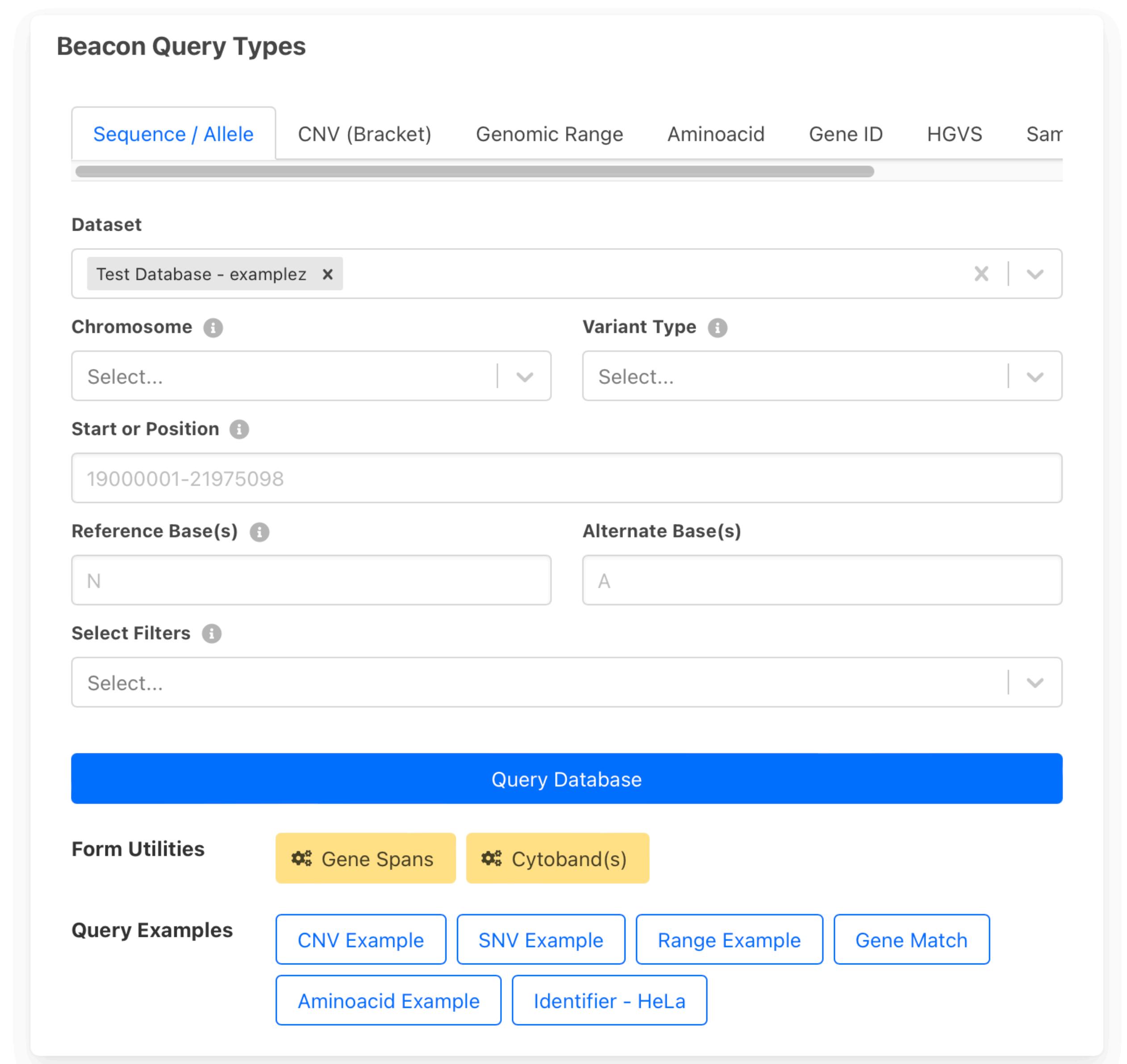
Reference Base(s): N Alternate Base(s): A

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



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

Beacon Query Types

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

Dataset

Progenetix X Test Database - examplez X

Chromosome i Variant Type i
17 (NC_000017.11) SO:0001059 (any sequence alteration - S...)

Start or Position i
7577121

Reference Base(s) i Alternate Base(s)
G A

Select Filters i
Select...

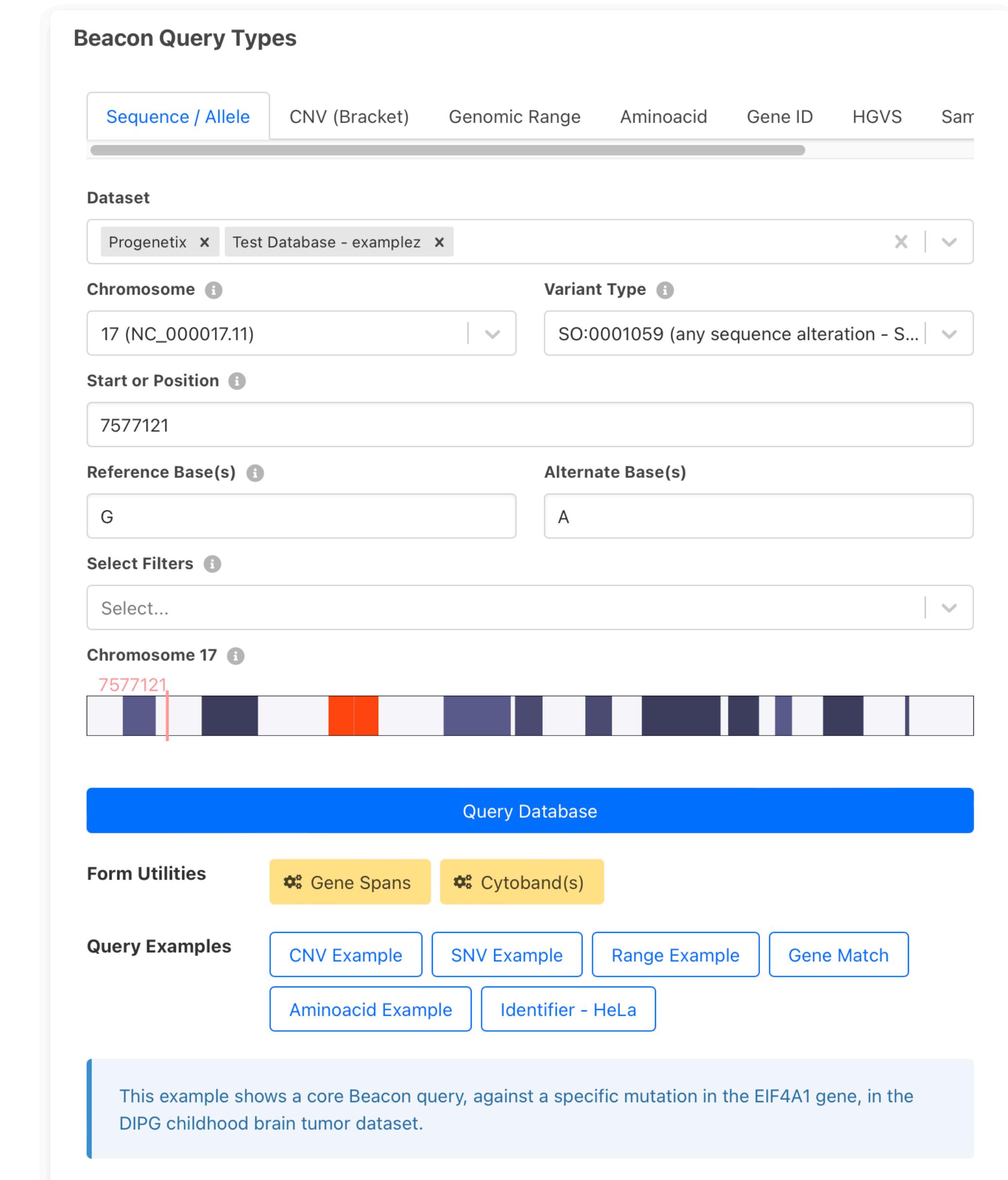
Chromosome 17 i
7577121

Query Database

Form Utilities Gene Spans Cytoband(s)

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

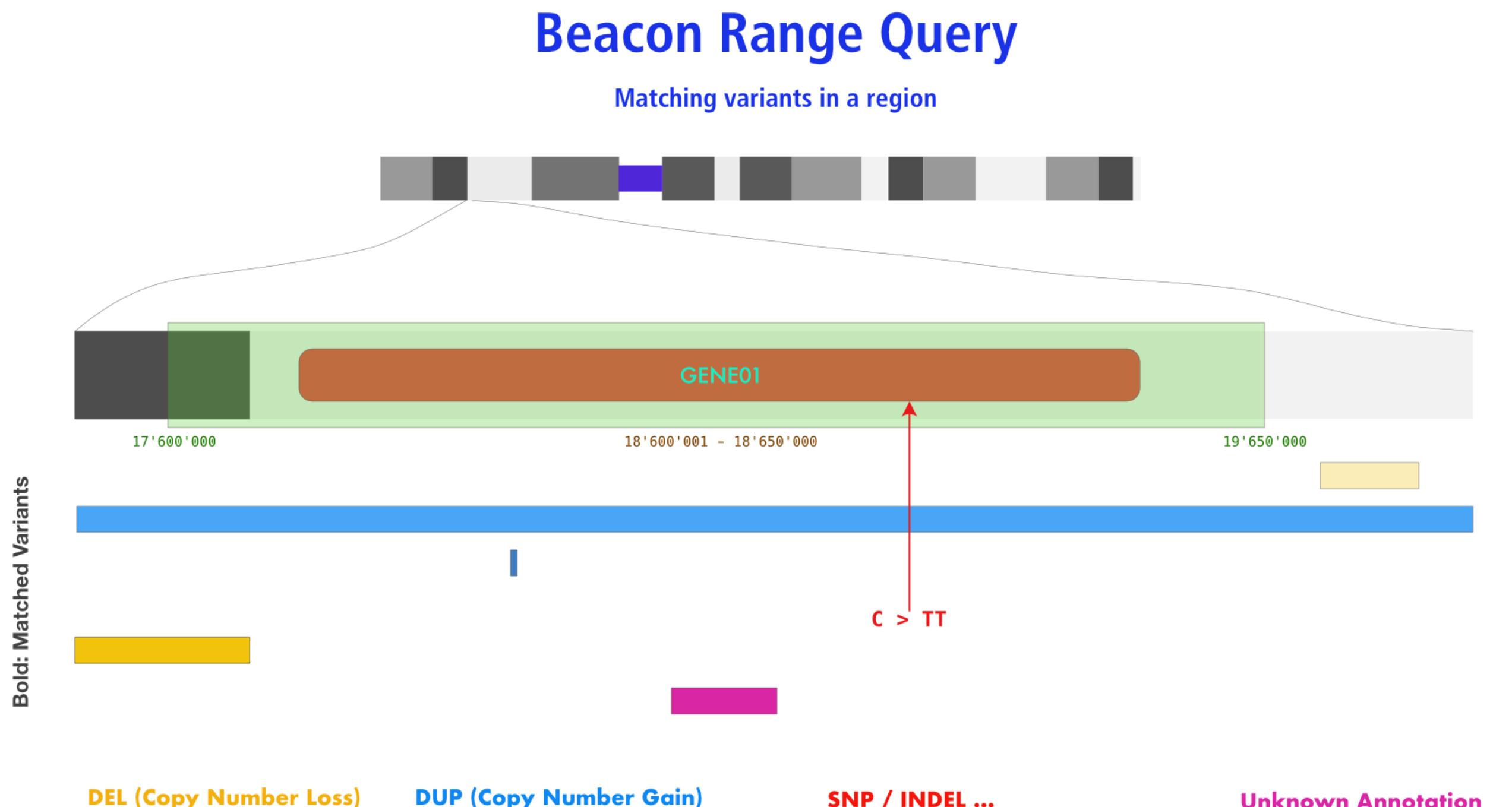
This example shows a core Beacon query, against a specific mutation in the EIF4A1 gene, in the DIPG childhood brain tumor dataset.



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 Query Types

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

Dataset

Test Database - examplez X

Chromosome

17 (NC_000017.11)

Variant Type

SO:0001059 (any sequence alteration - S...)

Start or Position

7572826

End (Range or Structural Var.)

7579005

Reference Base(s)

N

Alternate Base(s)

A

Select Filters

Select...

Chromosome 17

7572826
7579005

Query Database

Form Utilities

Gene Spans Cytoband(s)

Query Examples

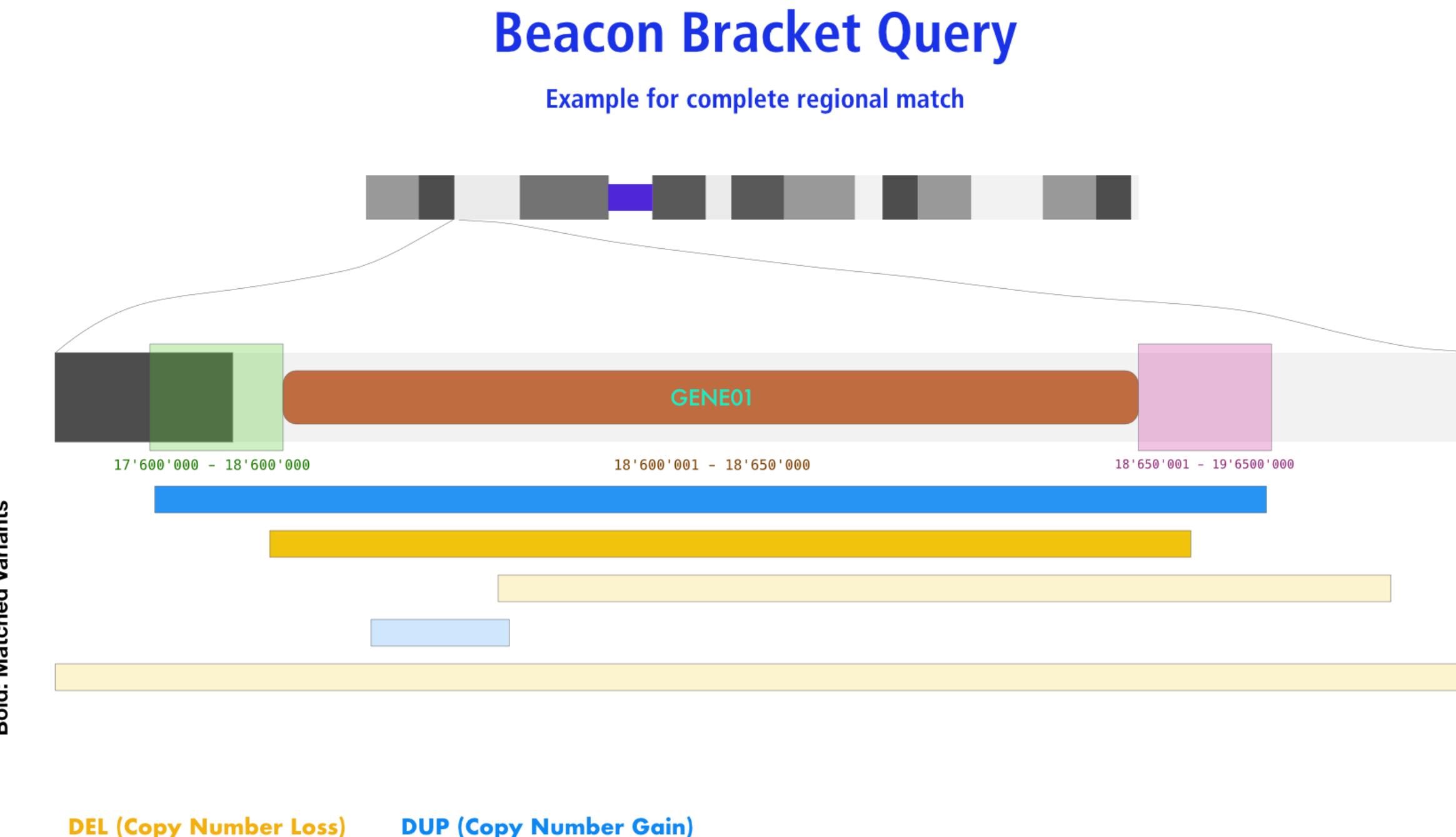
[CNV Example](#) [SNV Example](#) [Range Example](#) [Gene Match](#)
[Aminoacid Example](#) [Identifier - HeLa](#)

As in the standard SNV query, this example shows a Beacon query against mutations in the **EIF4A1** gene in the DIPG childhood brain tumor dataset. However, this range + wildcard query will return any variant with alternate bases (indicated through "N"). Since parameters will be interpreted using an "AND" paradigm, either Alternate Bases OR Variant Type should be specified. The exact variants which were being found can be retrieved through the variant handover [H->O] link.

Beacon Queries

Bracket ("CNV") Query

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



Beacon Query Types

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

Dataset

Test Database - examplez X ▼

Chromosome

9 (NC_000009.12) ▼

Variant Type

EFO:0030067 (copy number deletion) ▼

Start or Position

21000001-21975098

End (Range or Structural Var.)

21967753-23000000

Select Filters

NCIT:C3058: Glioblastoma (100) X ▼

Chromosome 9

21000001-21975098



Query Database

Form Utilities

Gene Spans Cytoband(s)

Query Examples

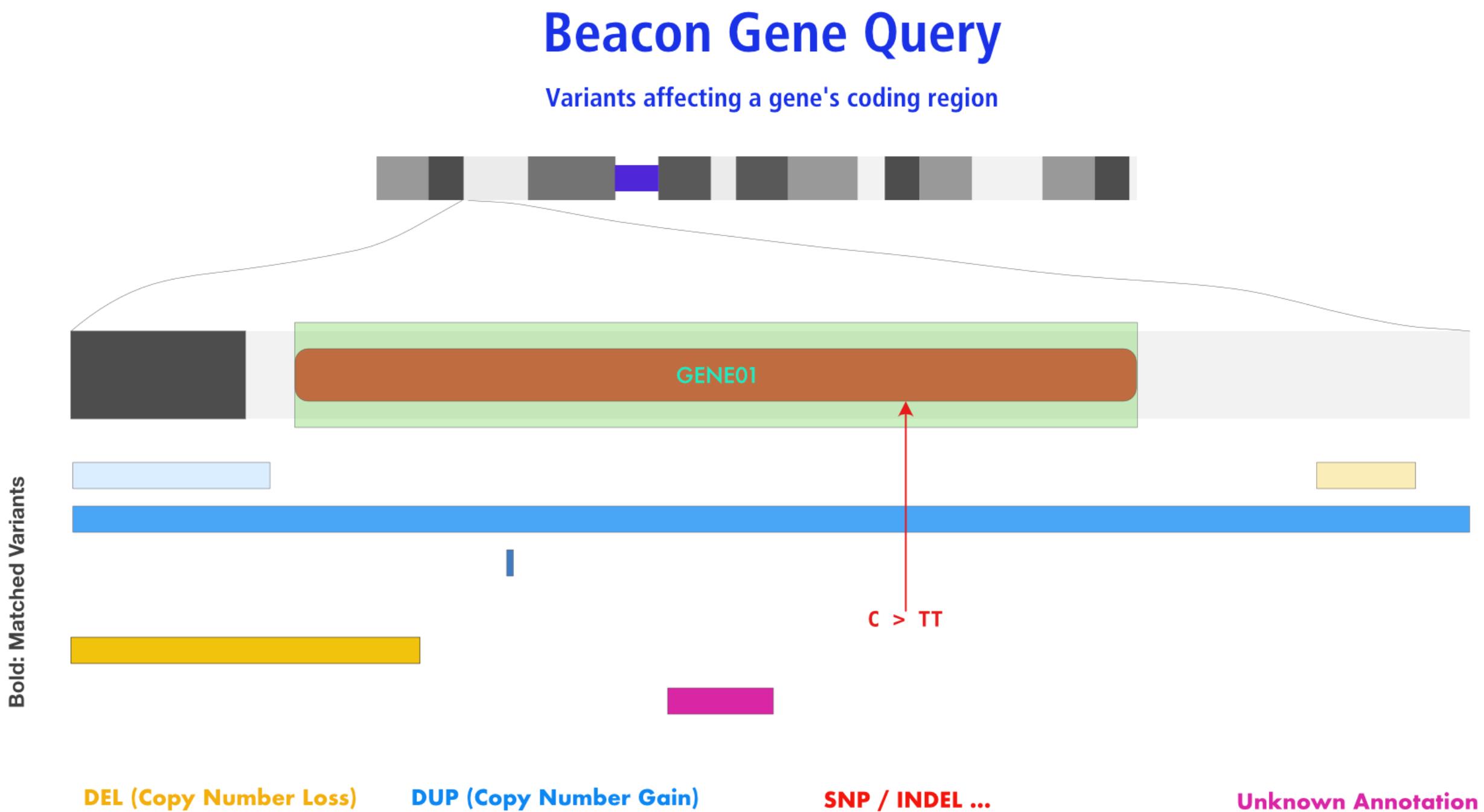
CNV Example SNV Example Range Example Gene Match
Aminoacid Example Identifier - HeLa

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "focal" hits (here i.e. <= ~2Mbp in size). The query is against the examplez collection and can be modified e.g. through changing the position parameters or data source.

Beacon Queries

Gene Request

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



Beacon Query Types

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

Dataset

Cancer Cell Lines Collection x | ▾

Gene Symbol i

CDK2 (12:55966830-55972789) x | ▾

Variant Type i

Select...

Min Variant Length i

Max Variant Length i

Alternate Base(s)

A

Select Filters i

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**

Beacon Query Types

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

Dataset
Test Database - examplez x | ▾

Chromosome i Variant Type i
Select... | ▾ Select... | ▾

Start or Position i
19000001-21975098

Reference Base(s) i Alternate Base(s)
N A

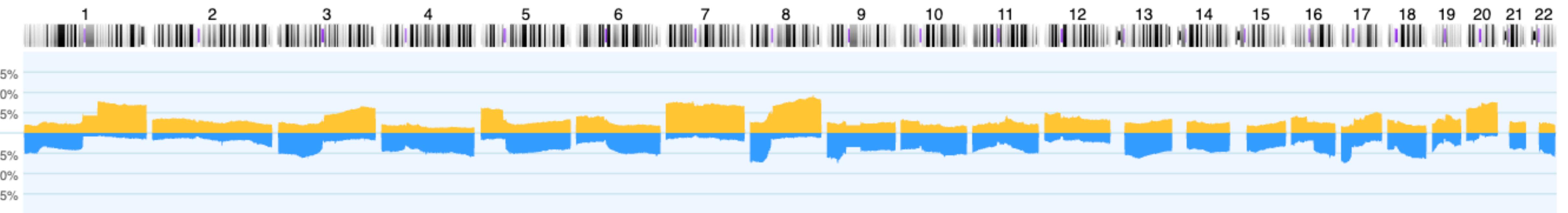
Select Filters i
Select... | ▾

Query Database

Form Utilities Gene Spans Cytoband(s)

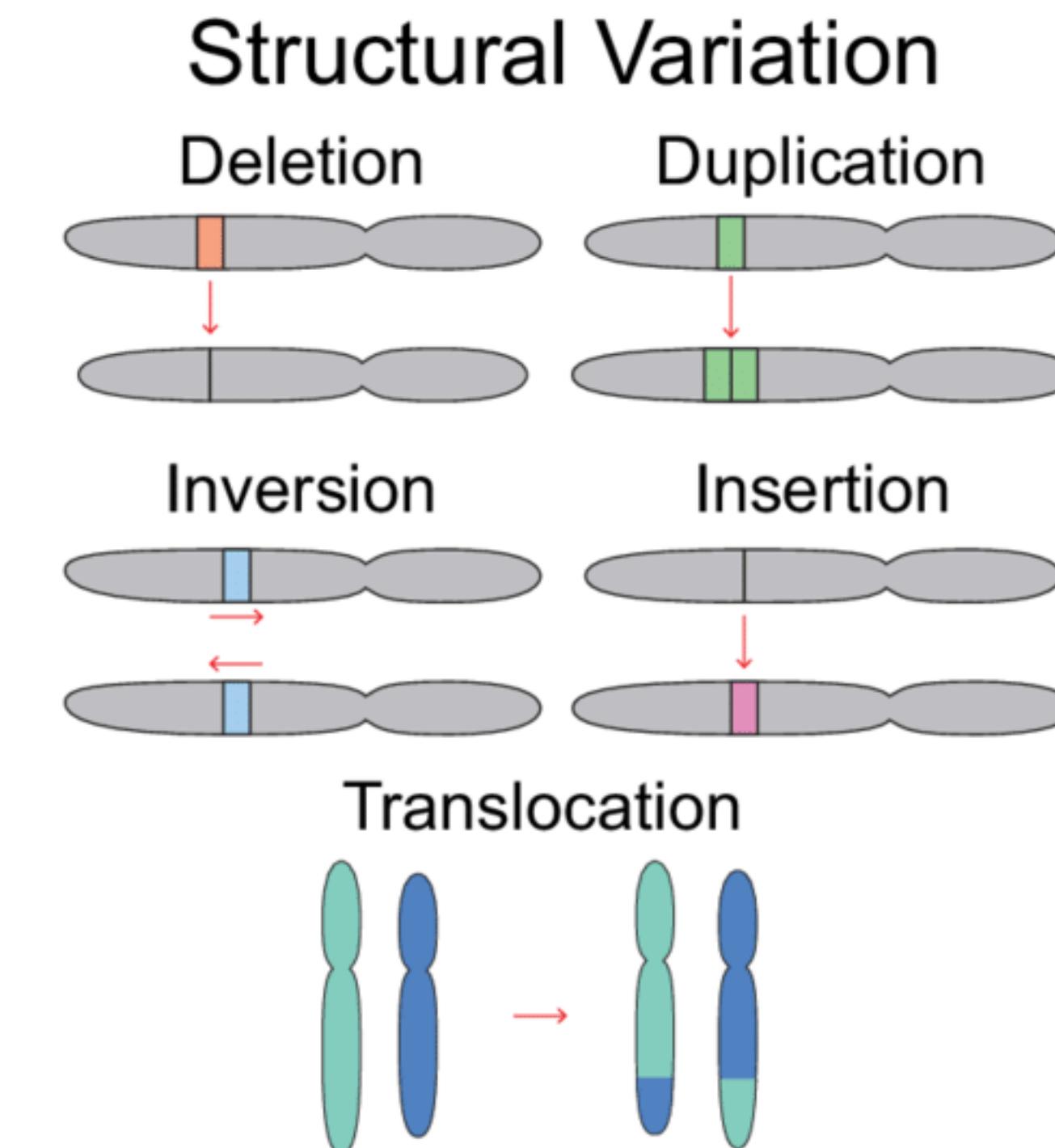
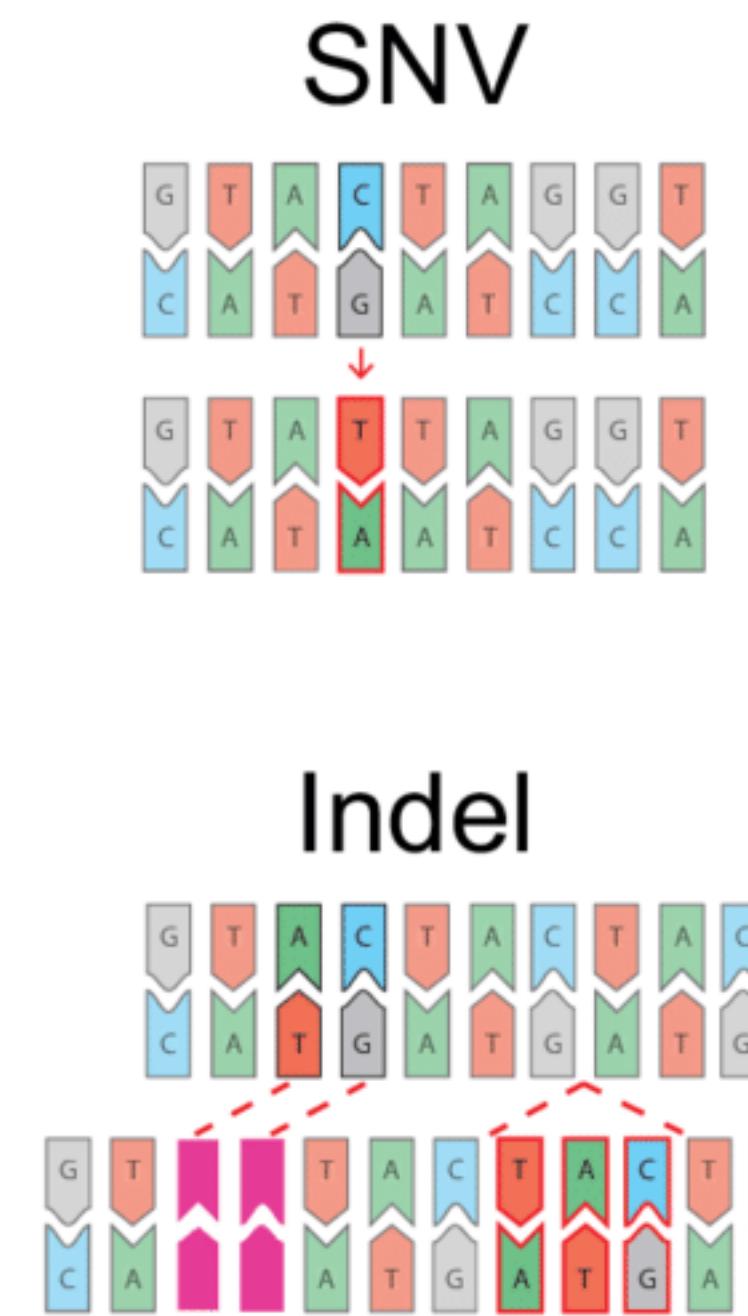
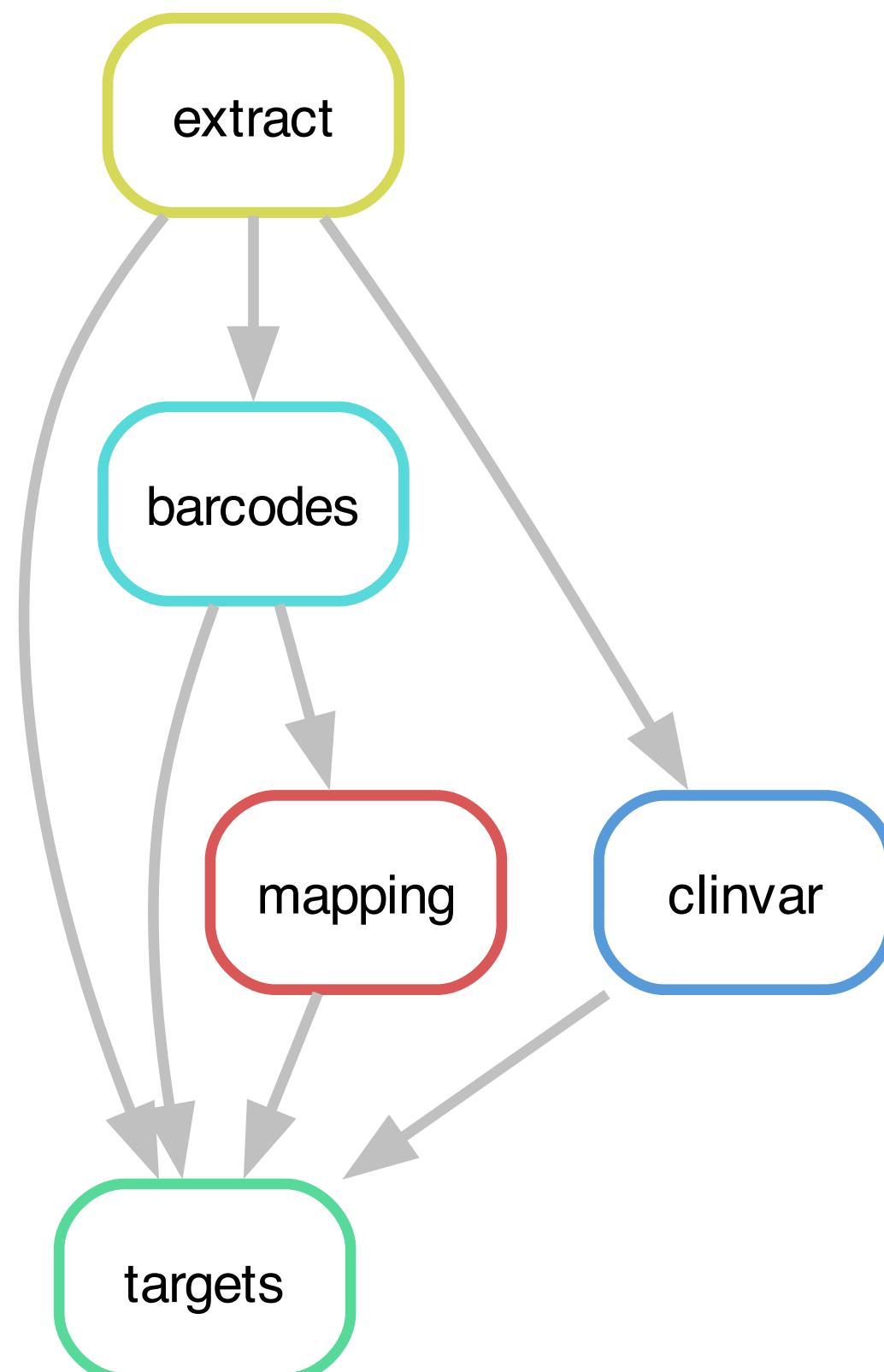
Query Examples CNV Example SNV Example Range Example Gene Match
Aminoacid Example Identifier - HeLa

TCGA Cancer samples (pgx:cohort-TCGAcancers)



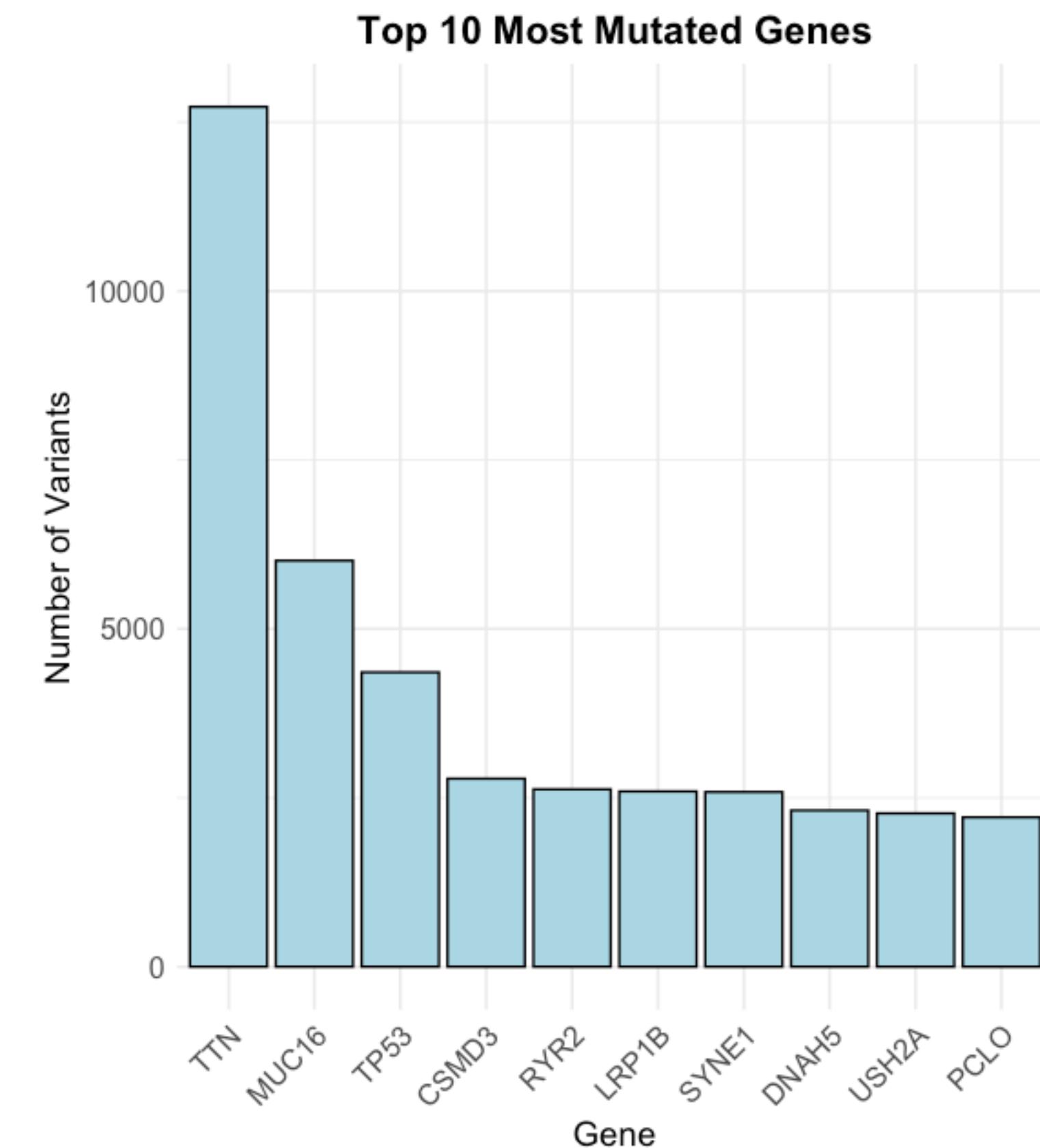
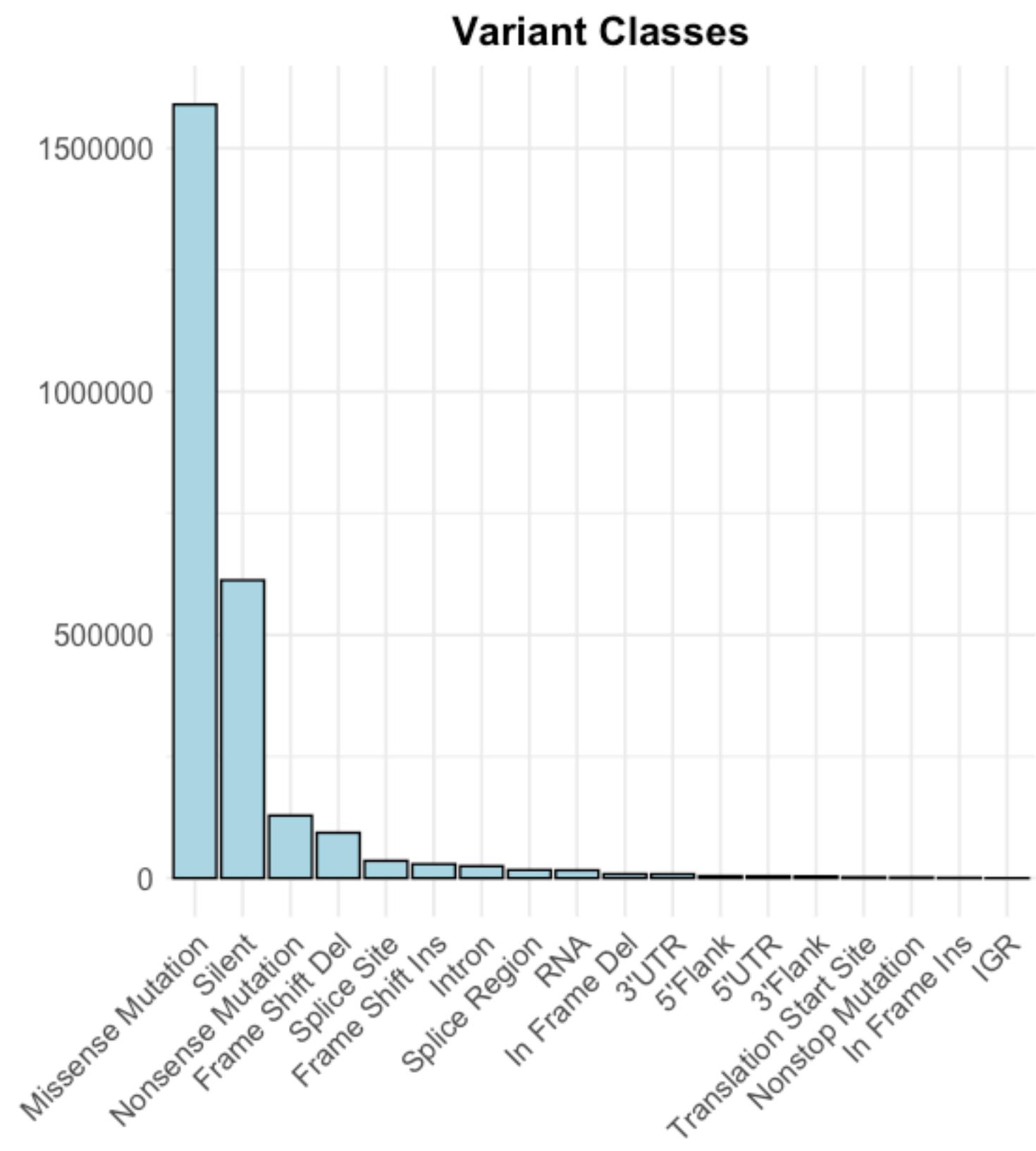
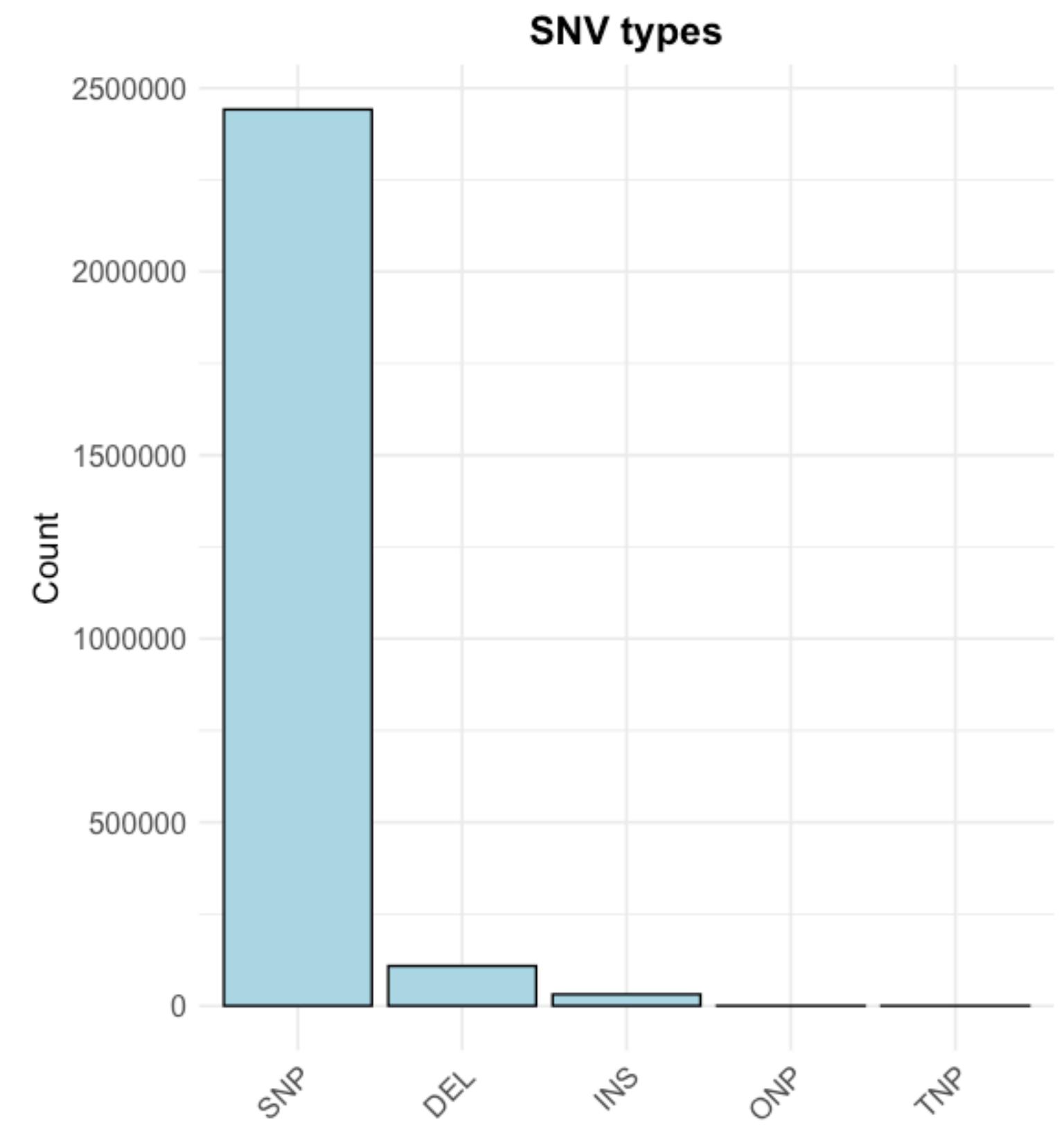
© CC-BY 2001 - 2023 progenetix.org

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

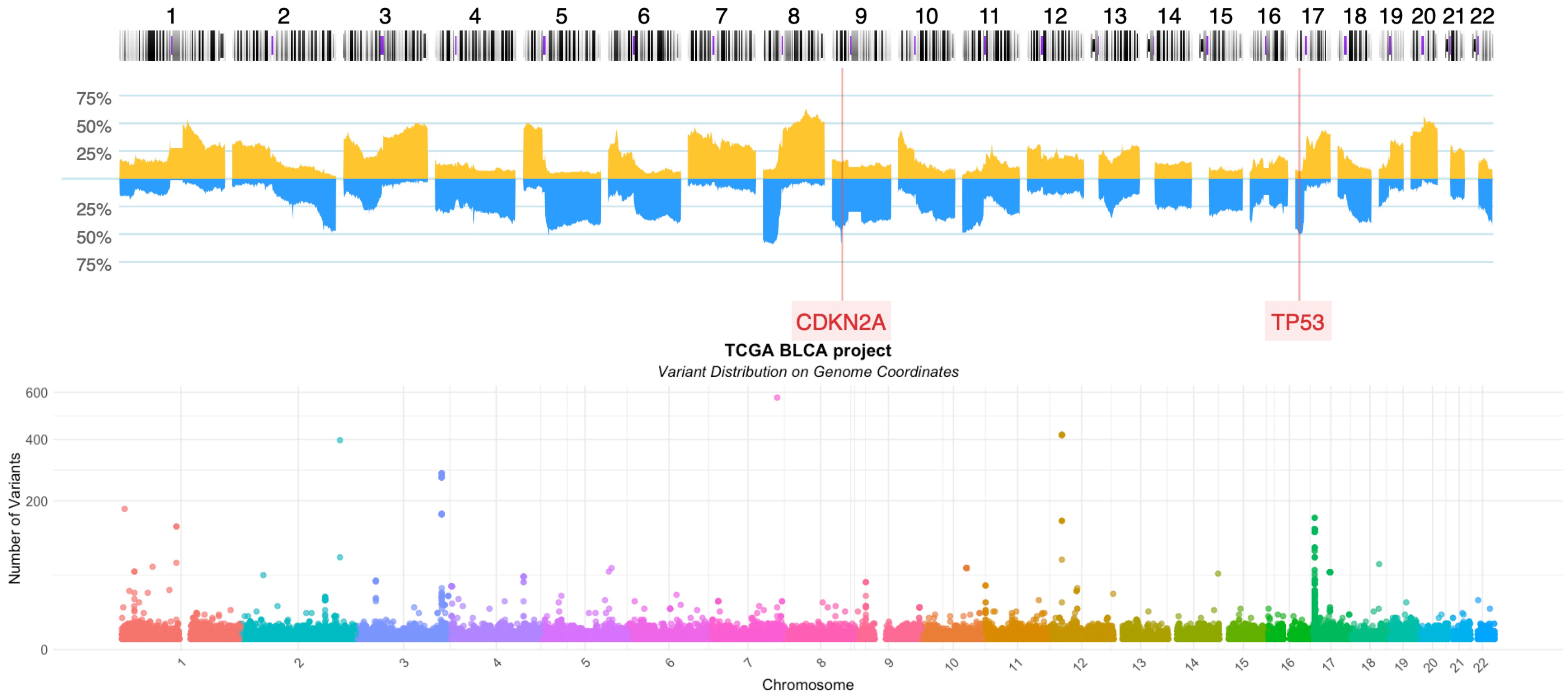


Nesta, Alex & Tafur, Denisse & Beck, Christine. (2020). Hotspots of mutation.

Mutation. Trends in Genetics



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**

Beacon Query Types

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

Dataset
Test Database - examplez x | ▾

Chromosome Variant Type
Select... Select...

Start or Position i
19000001-21975098

Reference Base(s) Alternate Base(s)
N A

Select Filters i
Select... | ▾

Query Database

Form Utilities Gene Spans Cytoband(s)

Query Examples CNV Example SNV Example Range Example Gene Match

Aminoacid Example Identifier - HeLa