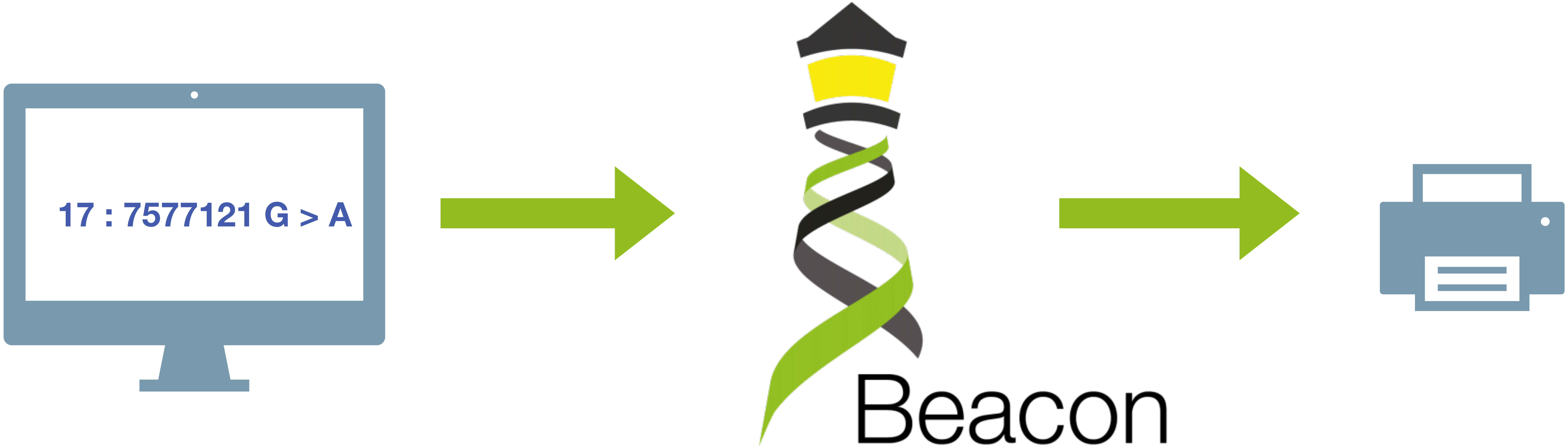
Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Beacon Variant Queries | GA4GH Connect | Michael Baudis | 2021-03-02



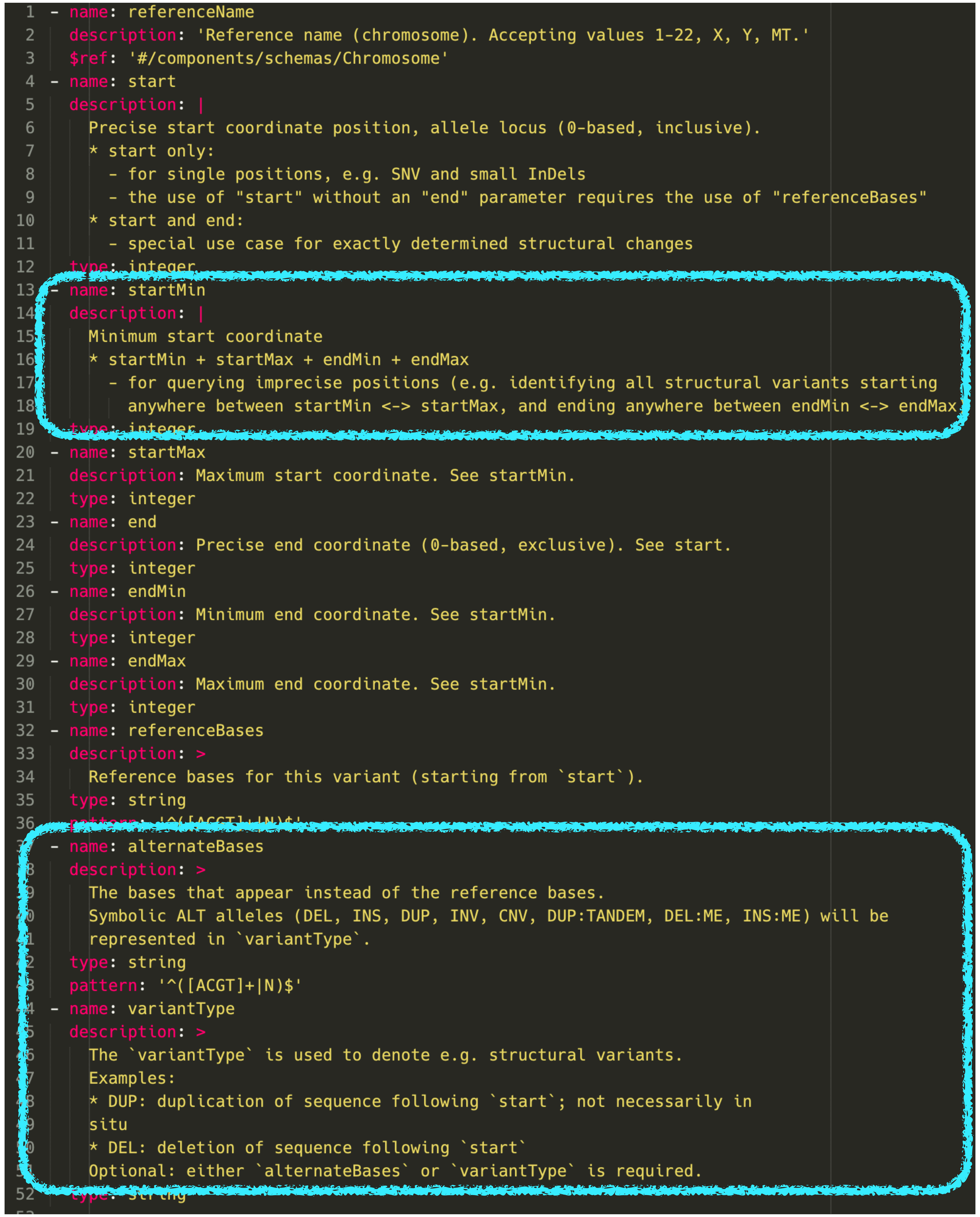
A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

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

Current Status of Variant Queries in Beacon

Lots of possibilities but limited definitions

- precise variant queries (chr17: 7673767 C>T)
- range queries ("any variant between here to there")
- structural genome variants, e.g. CNVs ("any deletion overlapping CDKN2A CDR coordinates")
- But no clear definition:
  - how those queries should be interpreted
  - which variant types are supported

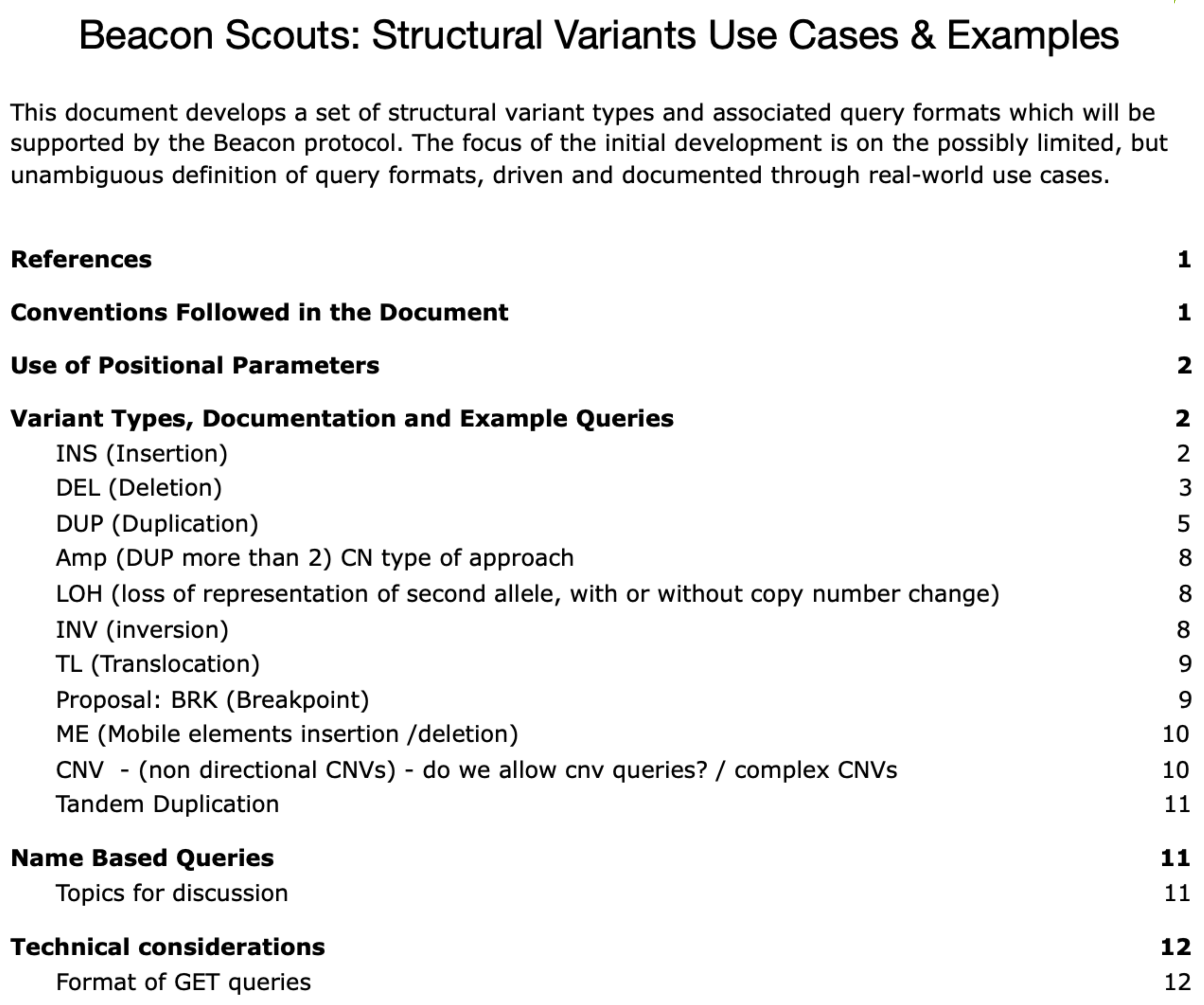


```
- name: referenceName
  description: 'Reference name (chromosome). Accepting values 1-22, X, Y, MT.'
  $ref: '#/components/schemas/Chromosome'
- name: start
  description: |
    Precise start coordinate position, allele locus (0-based, inclusive).
    * start only:
      - for single positions, e.g. SNV and small InDels
      - the use of "start" without an "end" parameter requires the use of "referenceBases"
    * start and end:
      - special use case for exactly determined structural changes
  type: integer
- name: startMin
  description: |
    Minimum start coordinate
    * startMin + startMax + endMin + endMax
      - for querying imprecise positions (e.g. identifying all structural variants starting
        anywhere between startMin <-> startMax, and ending anywhere between endMin <-> endMax)
  type: integer
- name: startMax
  description: Maximum start coordinate. See startMin.
  type: integer
- name: end
  description: Precise end coordinate (0-based, exclusive). See start.
  type: integer
- name: endMin
  description: Minimum end coordinate. See startMin.
  type: integer
- name: endMax
  description: Maximum end coordinate. See startMin.
  type: integer
- name: referenceBases
  description: >
    Reference bases for this variant (starting from `start`).
  type: string
  pattern: '^([ACGT]+|N)$'
- name: alternateBases
  description: >
    The bases that appear instead of the reference bases.
    Symbolic ALT alleles (DEL, INS, DUP, INV, CNV, DUP:TANDEM, DEL:ME, INS:ME) will be
    represented in `variantType`.
  type: string
  pattern: '^([ACGT]+|N)$'
- name: variantType
  description: >
    The `variantType` is used to denote e.g. structural variants.
    Examples:
    * DUP: duplication of sequence following `start`; not necessarily in
    situ
    * DEL: deletion of sequence following `start`
    Optional: either `alternateBases` or `variantType` is required.
  type: string
```

Beacon Scouts: Structural Variants

Re-defining & scoping variant queries

- contributors from different "stakeholder" areas
  - clinical genomics / rare diseases
  - variant repository (Ensembl)
  - cancer research resource
  - cancer variant annotation repositories
- close integration with ELIXIR h-CNV group
- process involved discussions about semantics of variant types, e.g.
  - DUP as CNV or in place
  - DEL as CNV from which size
- general attempt to use Sequence Ontology classes as guidance, but no still ambiguities / lack of terms



Beacon Scouts: Structural Variants Use Cases & Examples

This document develops a set of structural variant types and associated query formats which will be supported by the Beacon protocol. The focus of the initial development is on the possibly limited, but unambiguous definition of query formats, driven and documented through real-world use cases.

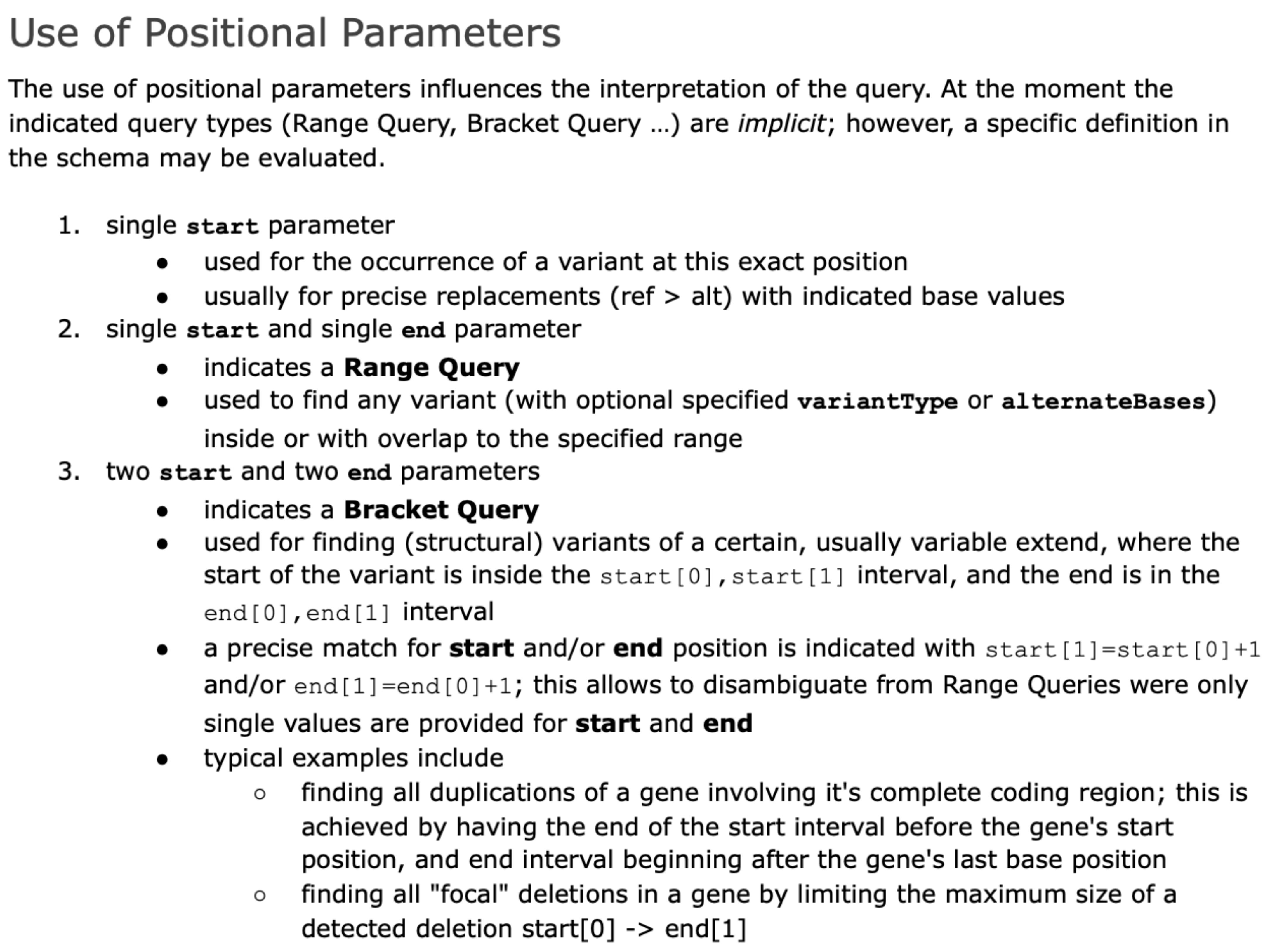
Beacon Scouts: Structural Variants

Positional Parameters & Querytypes

- Beacon v2 has slightly modified parameters
  - a list of 1 or 2 "`start`" parameters replaces `start` + `startMin` + `startMax`
  - a list of 0, 1 or 2 "`end`" parameters replaces `end` + `endMin` + `endMax`
  - (this was first proposed in 2016 but dropped tue to legacy format)
- the use of `start[0]`, `start[1]` && `end[0]`, `end[1]` parameters allows the match of any contiguous genome variant

  => **Bracket Query**
- however, most common use case besides specific "precise" variant is "someting in this region"

  => **Range Query**
  - one `start` and one `end` parameter
  - optional use of `variantType` OR `alternateBases`
  - any Range Query can in principle be expressed as Bracket Query



Use of Positional Parameters

The use of positional parameters influences the interpretation of the query. At the moment the indicated query types (Range Query, Bracket Query ...) are *implicit*; however, a specific definition in the schema may be evaluated.

1. single **`start`** parameter
   - used for the occurrence of a variant at this exact position
   - usually for precise replacements (ref > alt) with indicated base values
2. single **`start`** and single **`end`** parameter
   - indicates a **Range Query**
   - used to find any variant (with optional specified **`variantType`** or **`alternateBases`**) inside or with overlap to the specified range
3. two **`start`** and two **`end`** parameters
   - indicates a **Bracket Query**
   - used for finding (structural) variants of a certain, usually variable extend, where the start of the variant is inside the `start[0],start[1]` interval, and the end is in the `end[0],end[1]` interval
   - a precise match for **start** and/or **end** position is indicated with `start[1]=start[0]+1` and/or `end[1]=end[0]+1`; this allows to disambiguate from Range Queries were only single values are provided for **start** and **end**
   - typical examples include
     - finding all duplications of a gene involving it's complete coding region; this is achieved by having the end of the start interval before the gene's start position, and end interval beginning after the gene's last base position
     - finding all "focal" deletions in a gene by limiting the maximum size of a detected deletion start[0] -> end[1]

Beacon Scouts: Structural Variants

Example: Is an insertion described in a region ranging from chromosome ZZ from 1 to 10000 ?

Provided by: David Salgado
Notes:

- This is an application of a **Range Query**.

Query structure:

```
referenceName: "ZZ"
start: 0
end: 10000
variantType: "SO:0000667"

?referenceName=ZZ&start=0&end=10000&variantType=SO:0000667
```

Example: Find any "focal" deletion involving the CDKN2A locus

Provided by: Michael Baudis
Notes:

- This is an application of a **Bracket Query**
- A "focal" CNV event is (in cancer genomics) smaller than 1-5Mb; therefore, deletion events in this example should begin less than 1Mb 5' and end less than 1Mb 3' of the CDKN2A CDS
- This is the standard Progenetix Beacon+ DEL example query (with added NCIT disease filter and other required Beacon parameters).

```
referenceName: "9"
start:[21500000,21975098]
end:[21967752,22500000]
variantType: "DEL"

?referenceName=9&variantType=DEL&start=21500000,21975098&end=21967752,225
00000
```

Example: Has this specific region (chrXX: 10005-20005) been already found inserted in a genomic region?

Provided by: David Salgado
Notes:

- very rare query
- Michael: Such a query cannot be created for Beacon at the moment since the event of ***what*** is inserted doesn't have a way of being expressed. There is a partial overlap with duplication, translocation and insertion events; e.g. if the copy number of the inserted sequence has changed or if the fusion partners at the insertion points have been determined.

Query structure: NA

DUP (Duplication)

Definitions:

- SO:0001742 - A sequence alteration whereby the copy number of a given region is greater than the reference sequence (copy number gain).
- Beacon: Any quantitative increase of the number of alleles compared to the reference genome, without necessary indication about the physical location of the additional copies. The minimal size of what is matched as "DUP" (or other CNV) is left to the resource provider (some literature uses e.g. 50bp cut-off [1,2])

Examples below based on a specific study (https://doi.org/10.1016/j.tjog.2018.06.018)

Example: Find duplications involving the whole locus (chr2:54,700,000-63,900,000)

Provided by: David Salgado & Michael Baudis
Notes:

- This is an application of a **Bracket Query**
- Here, matched duplication events start 5` of the region and end 3` of it.
- Besides the positions, this requires knowledge about the maximum value of the reference base (or use of a very large one exceeding chromosome size; this example here uses a lazy "just bigger than chr2" value).

Query structure:

```
referenceName: "2"
start:[0,54700000]
end:[54700000,245000000]
variantType: "SO:0001742"

?referenceName=17&start,7669607&end=7676593,83257441&variantType=SO:00017
43
```

DEL (Deletion)

Definitions:

- SO: The point at which one or more contiguous nucleotides were excised.
- Beacon: Any variation of copy number of a genomic segment w/o size limitation imposed by the protocol[1,2] resulting in a net loss of copies compared to the expected allelic number at this locus.

Example: Find deletion events in gene *TP53*, *excluding* those that extend beyond the gene's CDS

Provided by: Diana Lemos
Notes:

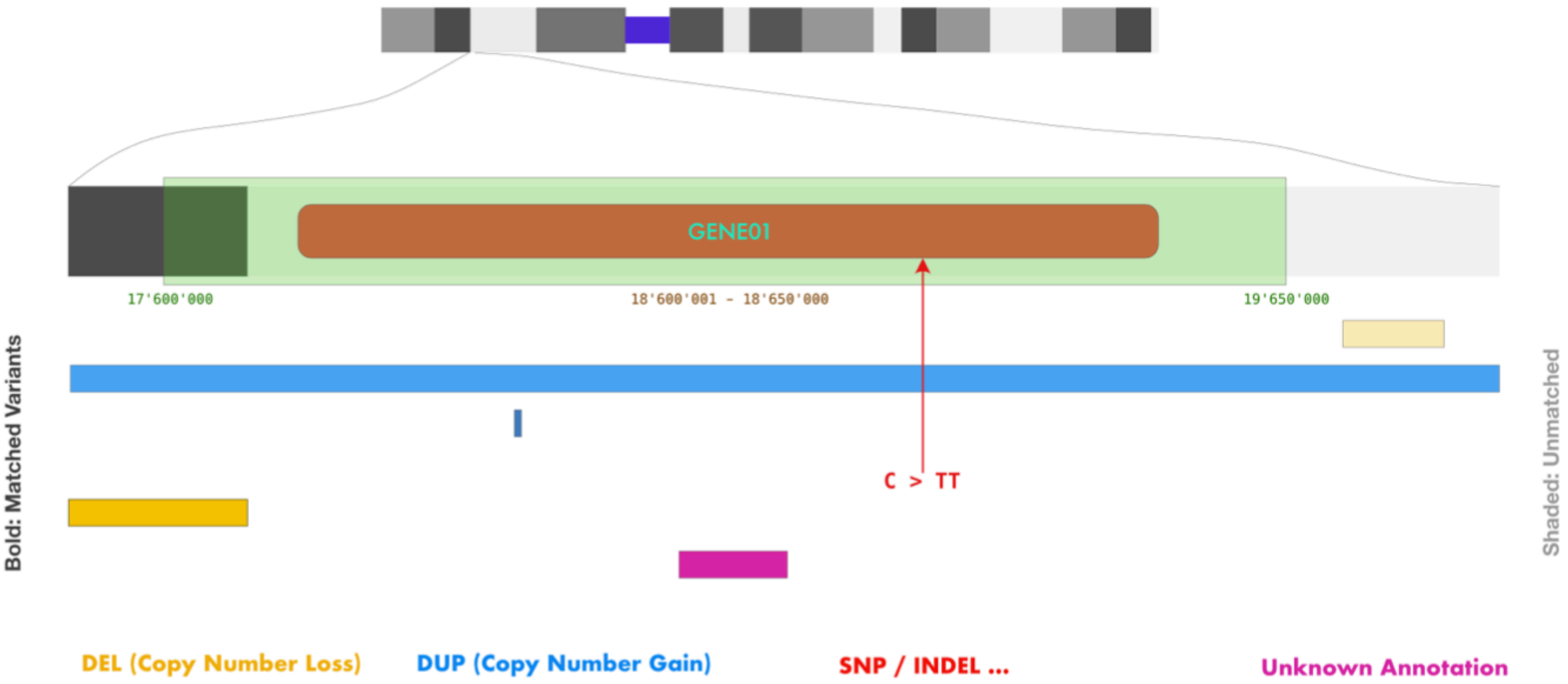
- This is an application of a **Bracket Query**.
- Any DEL that starts and ends inside the region is matched (the parameters would allow for single base deletions, which is correct from the "we do not define cutoff lengths", but also an example why additional `minLength` / `maxLength` parameters would meka sense)

Query structure:

```
referenceName: "17"
start: [7668402, 7687537]
end: [7668403, 7687538]
variantType: "DEL"

?assemblyId=GRCh38&referenceName=17&start=7668402,7687537&end=7668403,768
7538&variantType=DEL
```

BeaconRangeRequest

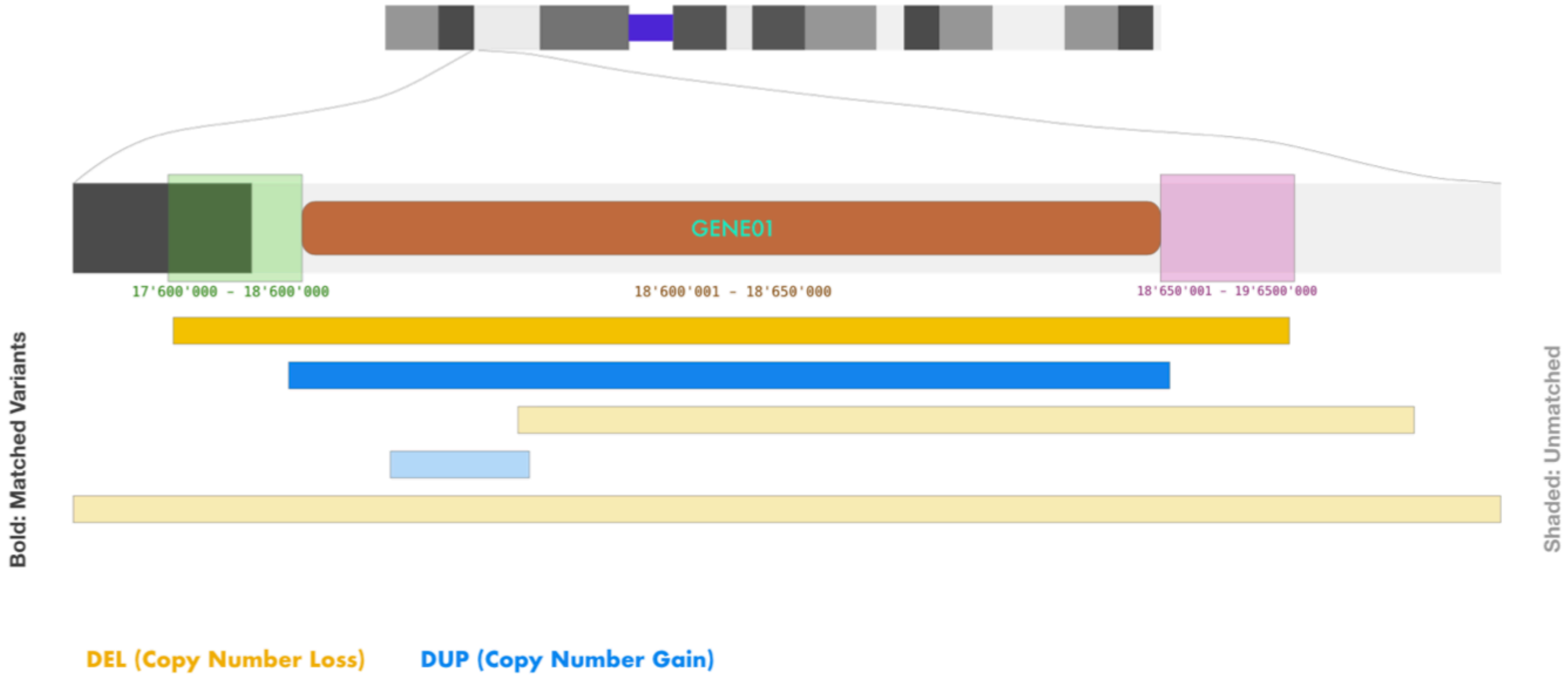


A BeaconRangeRequest answers to any variant overlapping the specified `start - end` range with a match. Responses can be limited by specifying `variantType`, `alternateBases` or `referenceBases` parameters. For limiting the size of matched CNVs a BeaconCnvRequest has to be used.

- `referenceName: 9`
- `start: [ 17600000, 18600000 ]`
- `end: [ 18650001, 19650000 ]`
- `variantType: SO:0001019`

BeaconCnvRequest

(full match)



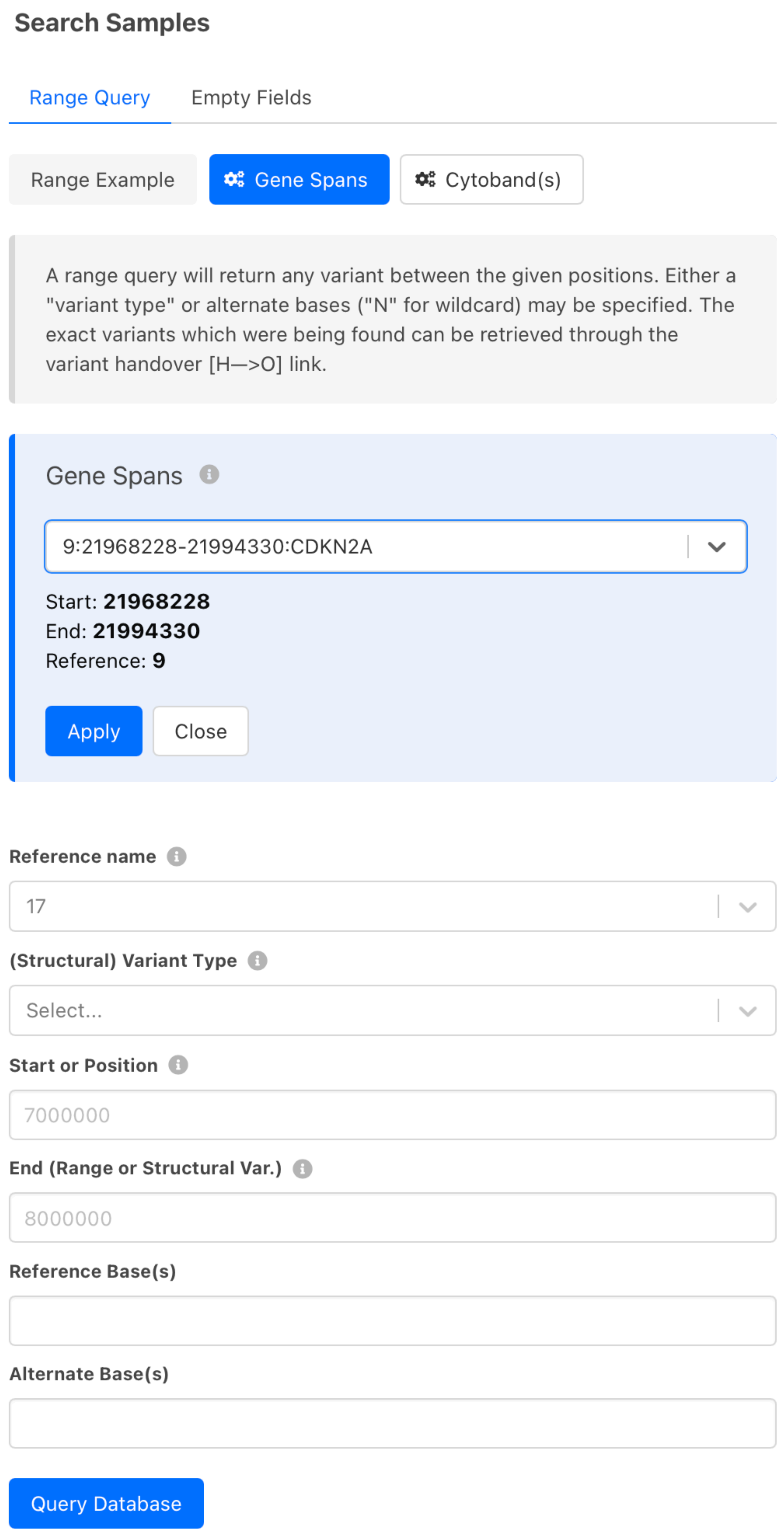
A “full match” BeaconCnvRequest is a typical scenario for e.g. matching CNVs in which the whole CDR of a gene has been duplicated. Here, both start and end search intervals lie outside of the region of interest. The maximum size of matched CNVs can be limited through the extend of the outer bounds (start[0], end[1]).

Named Element Search

Front-end service for finding variants in a given gene or other annotated element

- interactive selection and modification of positional query parameters through user interface
- relies on service (local or remote) to provide coordinate mapping
- positional parameters can be modified for range extension or conversion into bracketed search
- does not require any modification of Beacon parameters compared to standard v1/v2

Beacon+



Search Samples

A range query will return any variant between the given positions. Either a "variant type" or alternate bases ("N" for wildcard) may be specified. The exact variants which were being found can be retrieved through the variant handover [H—>O] link.

Gene Spans

9:21968228-21994330:CDKN2A

Start: **21968228**
End: **21994330**
Reference: **9**

Reference name: 17

(Structural) Variant Type: Select...

Start or Position: 7000000

End (Range or Structural Var.): 8000000

Reference Base(s):

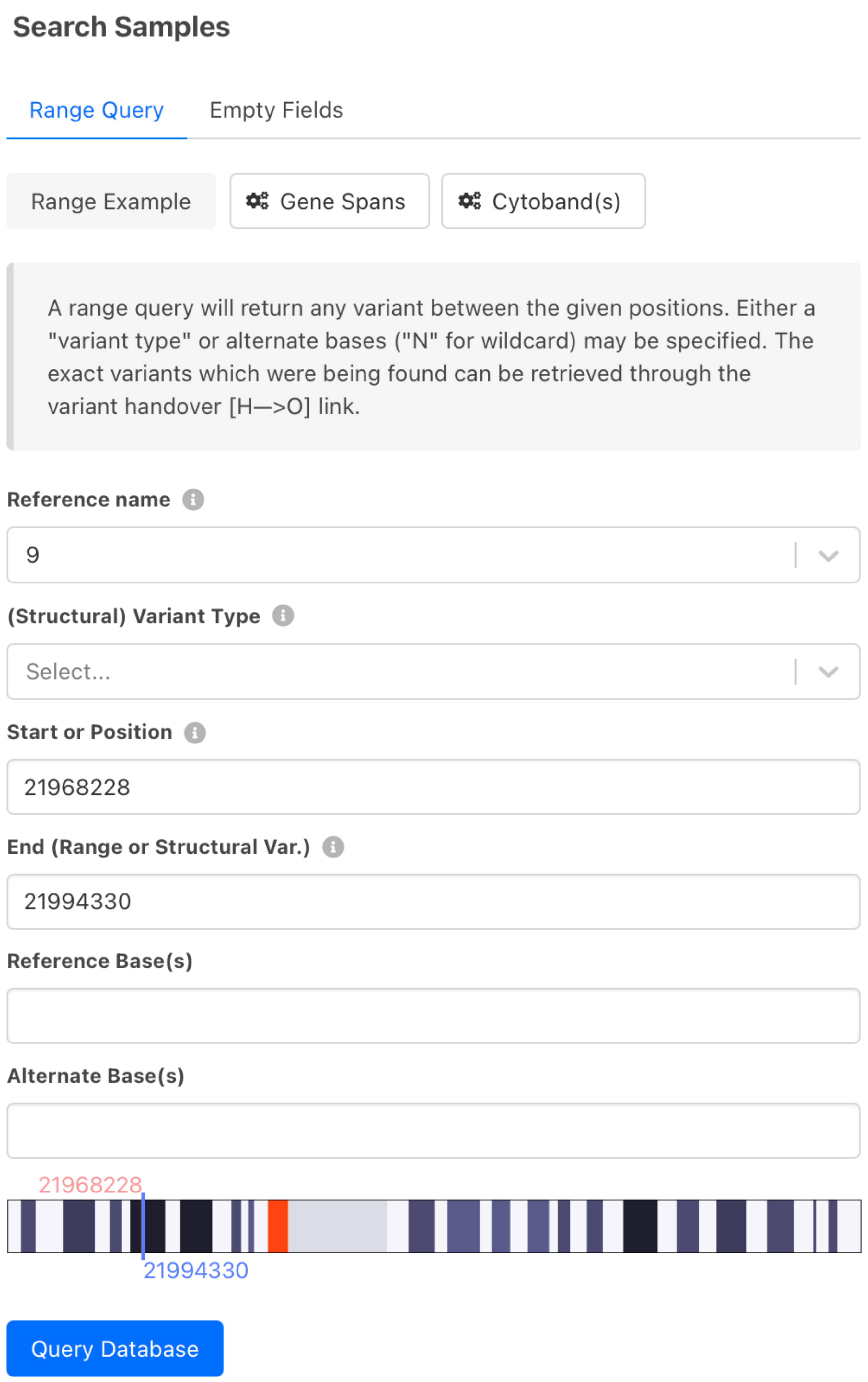
Alternate Base(s):

Named Element Search

Front-end service for finding variants in a given gene or other annotated element

- interactive selection and modification of positional query parameters through user interface
- relies on service (local or remote) to provide coordinate mapping
- positional parameters can be modified for **range** extension or conversion into bracketed search
- does not require any modification of Beacon parameters compared to standard v1/v2

Beacon+



A range query will return any variant between the given positions. Either a "variant type" or alternate bases ("N" for wildcard) may be specified. The exact variants which were being found can be retrieved through the variant handover [H—>O] link.

Reference name: 9

(Structural) Variant Type: Select...

Start or Position: 21968228

End (Range or Structural Var.): 21994330

Reference Base(s):

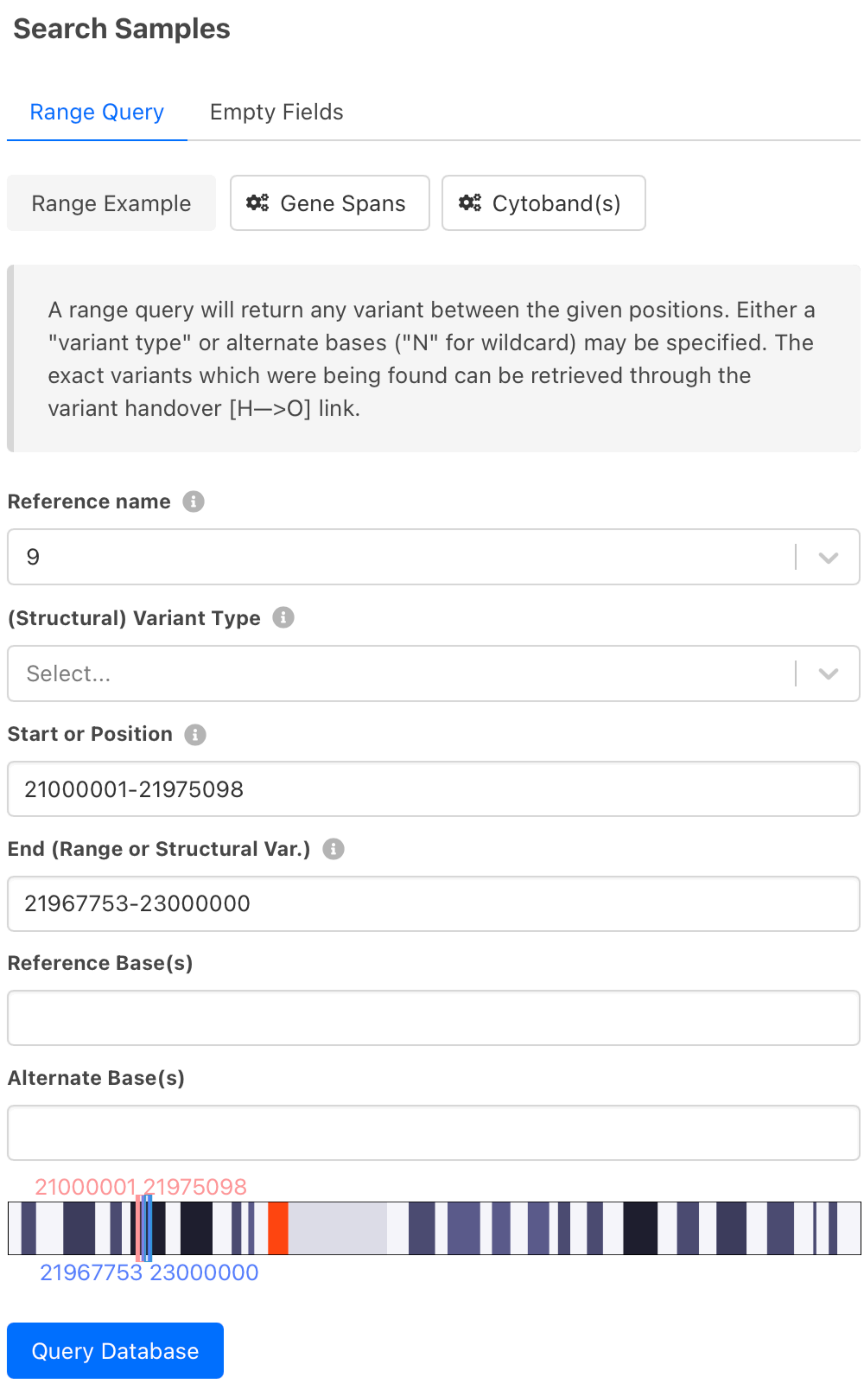
Alternate Base(s):

Named Element Search

Front-end service for finding variants in a given gene or other annotated element

- interactive selection and modification of positional query parameters through user interface
- relies on service (local or remote) to provide coordinate mapping
- positional parameters can be modified for range extension or conversion into **bracketed search**
- does not require any modification of Beacon parameters compared to standard v1/v2

Beacon+



Search Samples

A range query will return any variant between the given positions. Either a "variant type" or alternate bases ("N" for wildcard) may be specified. The exact variants which were being found can be retrieved through the variant handover [H—>O] link.

Reference name: 9

(Structural) Variant Type: Select...

Start or Position: 21000001-21975098

End (Range or Structural Var.): 21967753-23000000

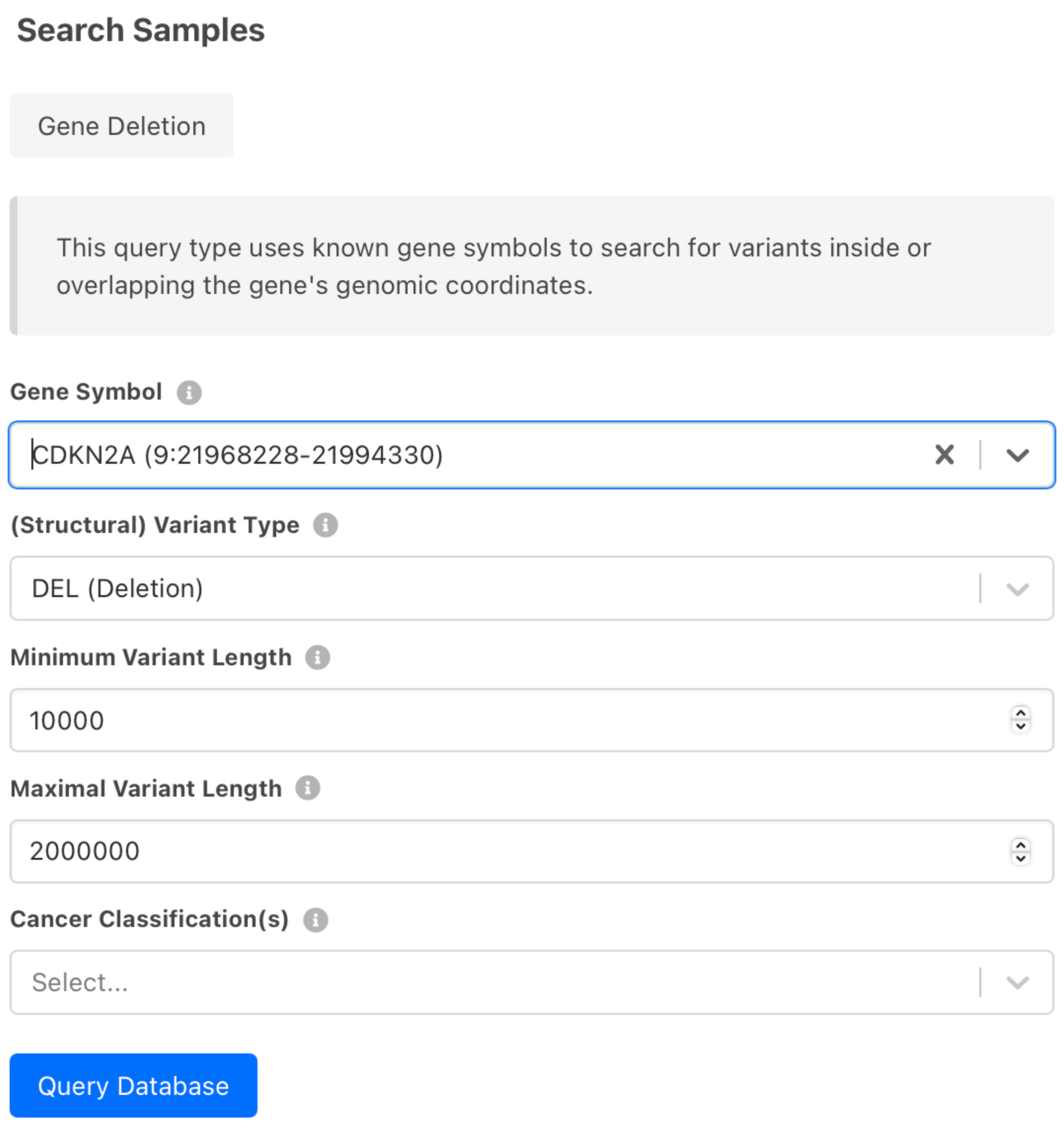
Reference Base(s):

Alternate Base(s):

Named Element Search

Back-end resolution for finding variants in a given gene or other annotated element

- use of a backend-provided resolver for coordinate mapping
- needs additional Beacon parameters compared to standard v1/v2
  - gene symbol, variant id in a common (?) format ...
  - additional positional modifiers - qualitative or quantitative (region expansin, size filters ...)



This query type uses known gene symbols to search for variants inside or overlapping the gene's genomic coordinates.

Gene Symbol: CDKN2A (9:21968228-21994330)

(Structural) Variant Type: DEL (Deletion)

Minimum Variant Length: 10000

Maximal Variant Length: 2000000

Cancer Classification(s): Select...

Beacon+

Beacon Scouts: Structural Variants

Status & To-Dos

- most queries can be expressed through a combination of existing parameters
  - referenceName && start[0],start[1]? && end[0]?, end[1]? &? (variantType || alternateBases)
- any positional variant query can be exopressed as Bracket Query
  - referenceName && start[0],start[1] && end[0], end[1] &? (variantType || alternateBases)
- two special types are documented for convenience
  - precise prositional query - referenceName && start && alternateBases
    - comparison of sequence at a specific position
  - Range Query - referenceName && start && end &? (variantType || alternateBases)
    - anything **overlapping** the interval start <=> end (and matching the optional type or basese parameters)
    - powerful "fishing" queries w/ option to disabiguate on parsing the retrieved variant data
- most use cases of "finding a variant" can be solved by one of
  - precise query like original BeaconAlleleRequest
  - ange Queries with optional post-filtering of the returned variants
- Do we need additional parameters for server-side resolution of "named" elements?
  - gene symbol, Ensemble ID, rsid ...

Beacon Scouts: Structural Variants

Status & To-Dos

- the scouting process has found that there are many "logical" variant types which are either described in conflicting concepts, could be interpreted ambiguously or cannot easily be queried if provided in non-normalized formats
  - duplication - CNV (duplicated material anywhere on the genome or extrachromosomal) versus in place "tandem"
  - VCF style "INDEL" - net-result (e.g. deletion indicated through shortened sequence can only be found by counting reference and alternative bases ...)
- we evaluate the option for a well-documented "mini ontology" for variant matching purposes, extending Sequence Ontology concepts for variant query use cases
- next steps are
  - providing the final documentation for the set of core queries
  - continue discussion on named queries, quantitative parameters ...

➡Watch *beacon-project.io*

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Babita Singh
Laureen Fromont
David Salgado
Michael Baudis
Diana Lemos
Alex H. Wagner
Joaquin Dopazo
Jordi Rambla
Anthony Brooks
Mauricio Moldes

... and many, many others

elixir

https://progenetix.org/beacon-genes/search?

https://progenetix.org/beacon-plus/search