

# Spec for query by types query&responses

## 1 Query by variant

### 1.1 Query

In basic query by variant user could enter region as a set of variant basic query parameters separated by commas, i.e, ref\_genome, start and end positions.

Alternatively, user could use dedicated menus to add a taxon id (species), assembly ID and variant-Type(s) from those available in Beacon implementation. (For this, a table with the taxon IDs available in beacon and their corresponding available assembly/version IDs and RefSeqs (Variant Basic **assemblyId**) should be available in backend, as table or dictionary). Bioteam to provide. For Viral Beacon so far only one species, SARS-CoV2, taxon id " " and one RefSeq "NC\_045512.2" are available.) Start, End should be given.

Fig with dropdown menus?

### 1.2 Filters

Most variant metadata/data (see those used in Viral Beacon highlighted in [Beacon V2 spec](#)) should be used as filters to narrow down query to only results meeting certain conditions, e.g, only a certain variant type: "indels", or a certain sequencing platform "Nanopore", or certain variant molecular consequence "nonsynonymous", or certain geographic origin "Spain", or lists thereof. Most important filters for Viral Beacon would be: Species taxonId (not now), Variant Basic variantType, Variant Annotation molecularEffect, and molecularConsequence, Biosample sample collectionDate and sampleOriginDetail, Individual sex, geographicOrigin, diseases.ageOfOnset and Run platform. These filters could appear as menus allowing select one or more (AND, XOR,...?).

For this, what is necessary - Variant annotation from VCF - Harmonization of values (Harmonization rules to Sabela) - Backend..

### 1.3 Response

Start and end positions of queried position(s) could be shown on top of at-scale viral genome position line for the selected reference, as in fig 1. This representation is based on coordinates for genomic regions in reference sequence can be obtained from Genbank annotation file that can be fetched and parsed (by Bioteam) to look like this:



Figure 1: Region queries represented on genome line

Response table figs same Babita

## 2 Query by region - custom coordinates or annotation/based

### 2.1 Query by region coordinates

This query is based on POS and length from variants in VCF.

### 2.1.1 Query

By entering (a list of) start:end position(s). Query should retrieve all variants in given stretch. One or more regions can be added in query.

Start and end positions of queried position/region(s) could be shown on top of at-scale viral genome position line for the selected reference, as in fig 1.

## 2.2 Query by region name/alias

This query uses aliases to refer to genomic regions. Mapping of these aliases to region start:end positions come from genome annotation file.

### 2.2.1 Query

By entering a (list of) start:end name(s)/alia(s).

**Query by region/annotation**

start typing a region/domain ⓘ

gene: ORF3a

Filters ⓘ

Add one or more filters to be applied to your query

custom regions can be queried by entering a start:end positions; annotated regions can be alternatively queried by their names., e.g ORF1 or nsp3

Figure 2: Region queries represented on genome line

What is necessary: - A table/dictionary mapping names/aliases/accessions accepted in this field to genomic region coordinates A plain text file looking like this, containing reference genome annotation-coordinates (by Bioteam): - Convert this in options to appear as suggestions maybe while user types? or menu, eg. gene:ORF8, stem\_loop:Coronavirus 3' UTR pseudoknot stem-loop 1, locus: GU280 gp08. - Do a (list of) query by region coordinates

region	start	end	name	locus_id
five_prime_UTR	1	265	5'UTR	NC.045512.2:1..265
gene	266	21555	ORF1ab	GU280_gp01
stem_loop	13476	13503	Coronavirus frameshifting stimulation element stem-loop 1	GU280_gp01
stem_loop	13488	13542	Coronavirus frameshifting stimulation element stem-loop 2	GU280_gp01-2
gene	21563	25384	S	GU280_gp02
gene	25393	26220	ORF3a	GU280_gp03
gene	26245	26472	E	GU280_gp04
gene	26523	27191	M	GU280_gp05
gene	27202	27387	ORF6	GU280_gp06
gene	27394	27759	ORF7a	GU280_gp07
gene	27756	27887	ORF7b	GU280_gp08
gene	27894	28259	ORF8	GU280_gp09
gene	28274	29533	N	GU280_gp10
gene	29558	29674	ORF10	GU280_gp11
stem_loop	29609	29644	Coronavirus 3' UTR pseudoknot stem-loop 1	GU280_gp11
stem_loop	29629	29657	Coronavirus 3' UTR pseudoknot stem-loop 2	GU280_gp11-2
three_prime_UTR	29675	29903	3'UTR	NC.045512.2:29675.
stem_loop	29728	29768	Coronavirus 3' stem-loop II-like motif (s2m)	NC.045512.2:29728.

## 2.3 Response