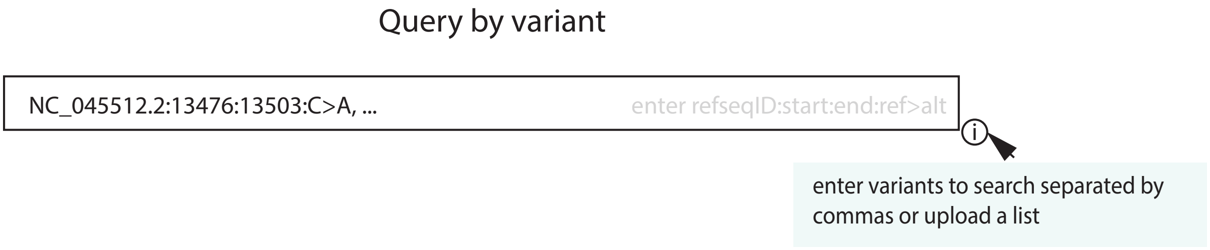
Spec for query by types query&responses

1 Query by variant

1.1 Query

Currently, in basic query by variant user enters a ”descriptor” of query by variant basic query parameters i.e, ref_genome, start and end positions ref and alt values.

If user were to add more than one, ie, a list of variants to query, she/he would have to enter a list of descriptors separated by commas.



NC_045512.2:13476:13503:C>A, ...

enter variants to search separated by commas or upload a list

Figure 1: Query by variant - current

As an alternative to this, user could use dedicated menus/boxes to add a taxon id (species), assembly ID and variantType(s) from those available in this 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.) so those can be default for now. Start, End should be given, but maybe in separate boxes. Leaving one empty would allow more flexible queries (finding all variants that start or end in a given certain position, not query by region).

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 (yes, there is only one so far), 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 2.

Response Babita

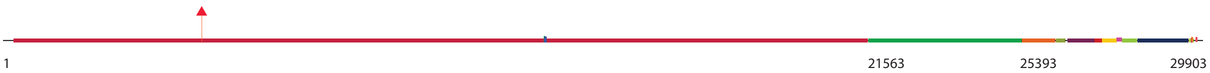


Figure 2: Region queries represented on genome line

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.

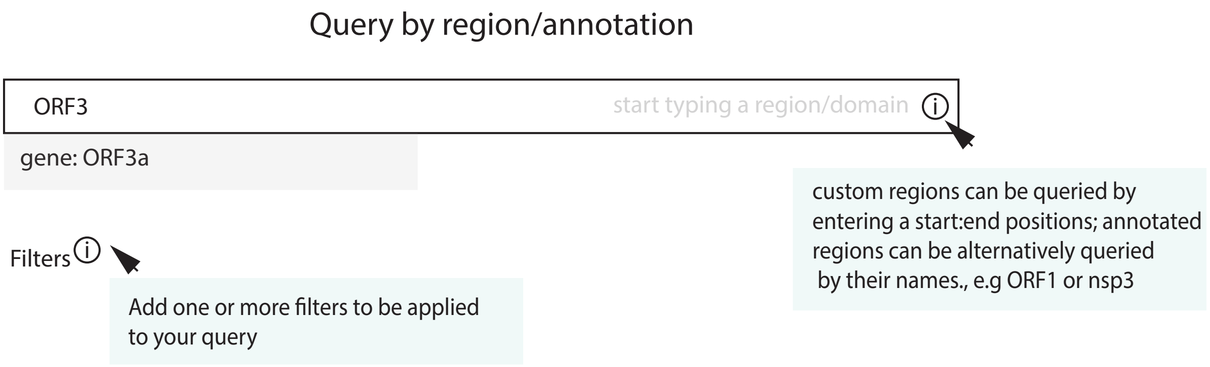
Start and end positions of queried position/region(s) could be shown on top of at-scale viral genome position line for the selected reference, as in fig 1.

2.2 Query by region name/alias

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

2.2.1 Query

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



Query by region/annotation

ORF3

gene: ORF3a

Filters

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

Add one or more filters to be applied to your query

Figure 3: Region queries represented on genome line

2.2.2 What is necessary

- A table/dictionary mapping names/aliases/accessions accepted in this field to genomic region coordinates (a plain text file looking like tables below, containing reference genome annotation-coordinates for genes, UTRs, stem loops, CDSs and mature proteins)
- 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

2.3 Response

- Needle for region freq alternate per position
- Freq variants per variantType
- dN/dS...

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.

proteins	start	end	protein_name	protein_id	parent_gene_name
1	266	13468	ORF1ab polyprotein	YP_009724389.1	ORF1ab
2	13468	21555	ORF1ab polyprotein	YP_009724389.1	ORF1ab
3	266	13483	ORF1a polyprotein	YP_009725295.1	ORF1ab
4	21563	25384	surface glycoprotein	YP_009724390.1	S
5	25393	26220	ORF3a protein	YP_009724391.1	ORF3a
6	26245	26472	envelope protein	YP_009724392.1	E
7	26523	27191	membrane glycoprotein	YP_009724393.1	M
8	27202	27387	ORF6 protein	YP_009724394.1	ORF6
9	27394	27759	ORF7a protein	YP_009724395.1	ORF7a
10	27756	27887	ORF7b	YP_009725318.1	ORF7b
11	27894	28259	ORF8 protein	YP_009724396.1	ORF8
12	28274	29533	nucleocapsid phosphoprotein	YP_009724397.2	N
13	29558	29674	ORF10 protein	YP_009725255.1	ORF10

proteins	start	end	protein_name	protein_id	parent_gene_name
1	206	805	leader protein, nsp1	YP_009725297.1, YP_009742608.1	ORF1ab
2	806	2719	nsp2	YP_009725298.1, YP_009742609.1	ORF1ab
3	2720	8554	nsp3	YP_009725299.1, YP_009742610.1	ORF1ab
4	8555	10054	nsp4	YP_009725300.1, YP_009742611.1	ORF1ab
5	10055	10972	3C-like proteinase	YP_009725301.1, YP_009742612.1	ORF1ab
6	10973	11842	nsp6	YP_009725302.1, YP_009742613.1	ORF1ab
7	11843	12091	nsp7	YP_009725303.1, YP_009742614.1	ORF1ab
8	12092	12685	nsp8	YP_009725304.1, YP_009742615.1	ORF1ab
9	12686	13024	nsp9	YP_009725305.1, YP_009742616.1	ORF1ab
10	13025	13441	nsp10	YP_009725306.1, YP_009742617.1	ORF1ab
11	13442	13480	nsp11	YP_009725312.1	ORF1ab
12	13442	16236	RNA-dependent RNA polymerase	YP_009725307.1	ORF1ab
13	16237	18039	helicase	YP_009725308.1	ORF1ab
14	18040	19620	3'-to-5' exonuclease	YP_009725309.1	ORF1ab
15	19621	20658	endoRNAse	YP_009725310.1	ORF1ab
16	20659	21552	2-O-ribose methyltransferase	YP_009725311.1	ORF1ab