Viral Beacon Use Case: Exploring Intrahost variants in S protein and its domains

Contents

1 Introduction **1**

2 Use case 1: Finding missense variants in S protein **1**
2.1 Recurrent variants within S protein 1

3 Use case 2: Finding missense variants in S subunits or key domains **2**
3.1 Recurrent intrahost missense variants in Receptor-binding Domain 3

4 Use case 3: Finding missense variants whose population frequency and/or within-host AF have increased over time **3**

5 Use case 4: Finding co-ocurring missense variants in S proteins **5**

1 Introduction

Being SARS COV2 main entry factor and the target of neutralizing antibodies and CTL response, Spike protein or its subdomains are the only vaccine candidates currently in clinical trials as well as of many antiviral inhibitors candidates targeting and viral entry and fusion.

The exploration of S variants may be important for awareness of potential limitations of current vaccine candidates over geographic locations or time as the virus evolves, and for future vaccine design.

With the readily availability of variation data from raw NGS data and dedicated pipelines (by Galaxy Project) aimed at detecting low frequency variants along with harmonized metadata and annotations, COVID19 Viral Beacon might be a useful platform for exploring intrahost variation. In this use case exercise we will use Beacon V2.0 features such as filters, region query and feature query to explore intrahost variation within S protein.

2 Use case 1: Finding missense variants in S protein

By using *Query by feature* specifying gene:S and using *Filters* to filter by Functional Class to "MISSENSE" the user can get the result of all missense variants on S gene, along with their frequency at population level (dataset frequency), their withing host AF and metadata including the sample types, geographic locations and collection dates of samples harboring them.

As a brief exploration, the user can use the *Sorting* feature to explore the dataset frequencies and AF of variants, and then use *Filters* filter to select the desired ranges of dataset frequencies and AF.

Figure **??** panel A shows the distribution of positions in S1 protein harboring missense variants in the dataset (2216) and the mutation frequency (number of samples with variants at position) at each of them. Some positions are rarely mutated (negative selection has been described to prevail in key domains) while a few are more frequently mutated. User can focus on frequently mutated positions and explore whether variants at these positions are yielding same or chemically similar aminoacid changes, whether these land variants appear associated to one or many geographic locations or dates.

As a summary, Panel B shows the distribution of variants by population frequency value (number of variants at each population frequency value). Only 1 missense variant in S gene is present in over 30 % of samples and only 10 variants over 10 %. These might me interesting to explore further.

2.1 Recurrent variants within S protein

By using *Filter* user can explore recurrent variants.

Just to illustrate this, for example, selecting only variants with over 0.3 of dataset frequency, only one variant ("23403>A>G", "D614G") is left, which is present in 0.46 %, i.e 677/1473 samples).

User can access metadata for this variant so discover more about it. For example, that it is found in samples from different geographic locations (UK, USA, Australia, China, Peru and Egypt), it lands on Spike S1 chain, the main subunit for receptor interaction (Annotation from Uniprot) and that earliest date in which the variant appeared in database is "2020-02-07" and it has appeared in samples from February to April. Mean AF of this variant is 0.9498617 (i.e, fixed) and has increased over time (not shown), suggestive of positive selection.

Actually, this variant has been described and functional studies suggest that it might be increasing the viral infectivity (Korber, 2020).

Curiously, at same position there are 3 samples with another minor variant giving different aminoacid substitution ("23403>A>G", "D614V", mean AF 0.0086), that appeared in Australia study only in "2020-03-25", which has not been described yet.

Likewise, by using *Filters* to a range of dataset frequency 0.1 - 0.3, ten variants at then different positions appear ("23273>G>T" ("D571Y"), "23294>G>T" ("D578Y"), "23302>G>T" ("Q580H"), "23318>G>T" ("D586Y") ,"24497>G>T"("D979Y"), "24557>G>T" ("G999C"), "24737>G>T" ("G1059C"), "24933>G>T" ("G1124V"), "25340>G>T" ("D1260N")).

Applying and additional filter of within host AF > 0.3 to be on the safe side, two variants are left ("24557>G>T" ("G999C"), "24933>G>T" ("G1124V")). These residues land on extracellular domain of Spike S2 subunit (Uniprot annotation).

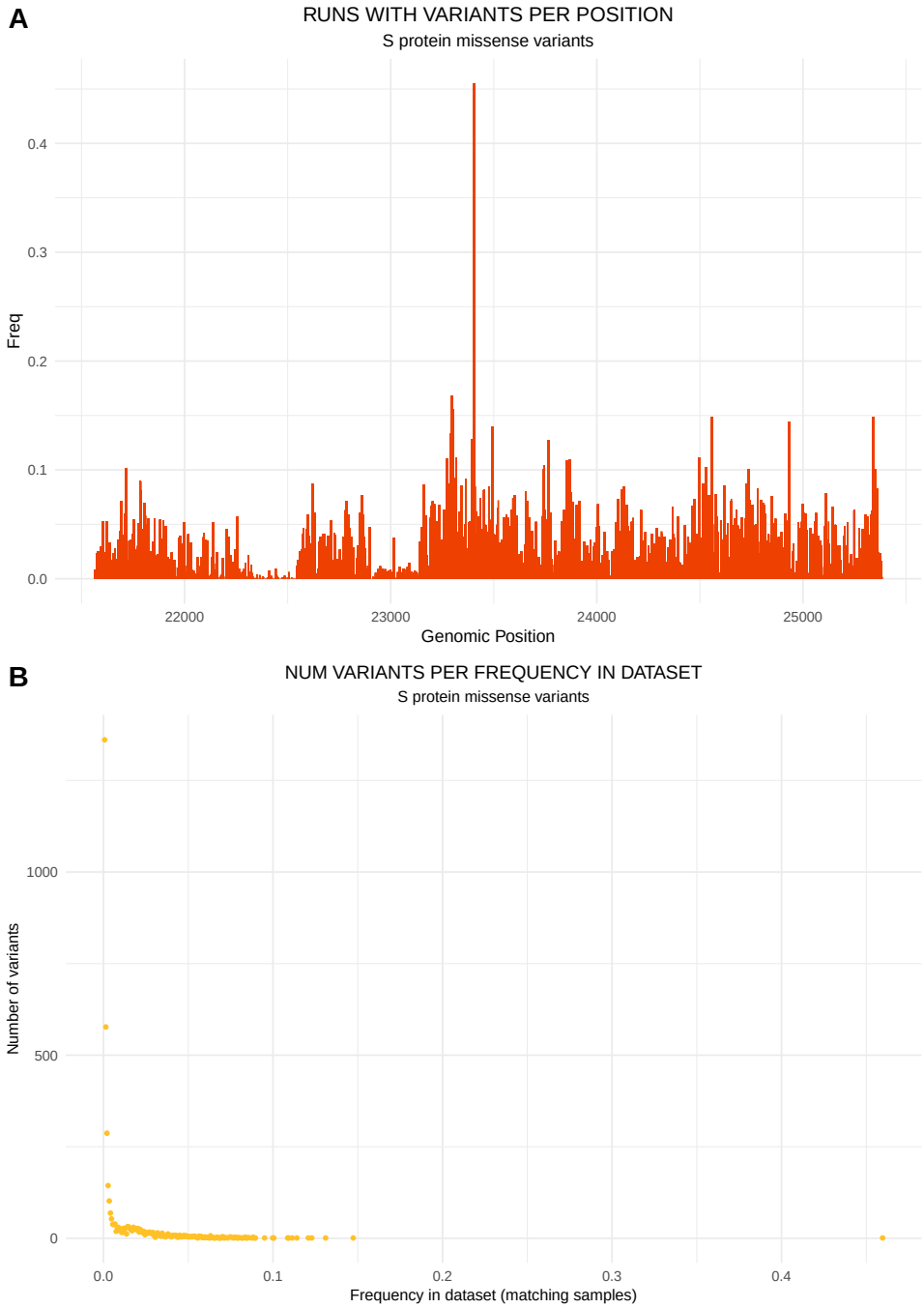

Figure 1: A) Frequency in dataset (matching samples) per genomic position. B) Number of variants by population frequency (frequency in dataset)

3 Use case 2: Finding missense variants in S subunits or key domains

Alternatively to exploring data in the whole S protein, user can narrow down the search to functional domains of interest, by searching by genomic positions in *Query by region* or by name/alias for key features that have Uniprot annotation available e.g S1 or S2 subunits, the receptor binding domain (RBD), the receptor binding motif therein, the fusion peptide, heptad repeat 1 or heptad repeat 2, etc.

As en example, exploration of the RBD (genomic positions 22517-23185, residues 319-541) is shown in Figure .

Missense variants are found in 341 out of 669 positions of RBD, with 538 unique variants giving a

total of 361 unique aminoacid substitutions.

3.1 Recurrent intrahost missense variants in Receptor-binding Domain

Ten missense variants in RBD are found in > 5 % of samples in dataset. ("22599>G>T" ("R346I"), "22620>G>T" ("W353L"), "22621>G>T" ("W353C"), "22630>G>T" ("K356N"), "22785>G>T" ("R408I"), "22851>C>T" ("T430I"), "22859>G>T" ("V433F"), "22865>G>T" ("A435S"), "23161>G>T" ("L533F"), "23162>G>T" ("V534F")).

It would be interersting to study these variants functionally.

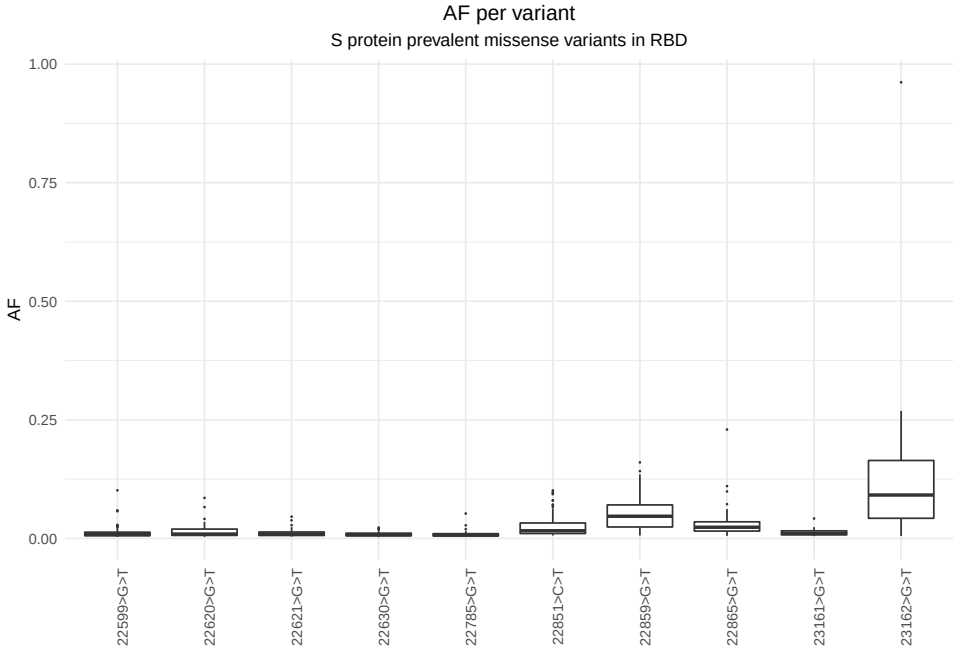

Figure 2: Distribution of AF in runs of prevalent missense variants (>5 %) samples

4 Use case 3: Finding missense variants whose population frequency and/or within-host AF have increased over time

User can use filters of population frequency (split by date and country) and AF over some confidence threshold (e.g 0.3) to discover variants whose frequency or within host AF have significantly changed over time.

Figure 5 summarized the distribution of missense variants by collection date and country, where samples showing variants are represented by genomic position and collection date, country origin of samples are categorized in colors and number of samples within each category are counted and represented as dots of size corresponding to counts.

This would allow user to discover positions hit by variants which are prevalent over time and those whose population frequency might be increasing.

For example, in **??** two spots show most interesting, one is and the other one is position 21998, variant "21998>C>T", whose population frequency seems to have increased in April with respect to March when it first appeared. Although in figure it shows mostly in USA (gray), this variant is present in samples from USA and Australia.

Likewise, grouping AF per run of variant by date would allow user to spot variants that might be increasing its predominance within host, which is a signal of evolution within host that might reveal population bottlenecks.

For example, "21998>C>T" has increased in population frequency (27/785 vs 25/144) and dramatically in AF since its first appearance on March (4).

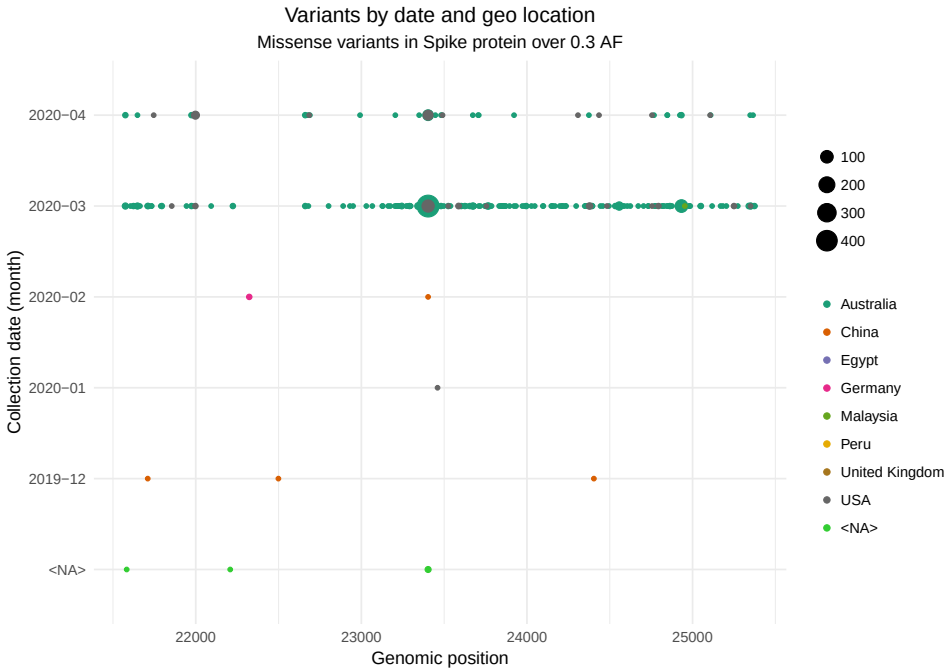

Figure 3: Number of variants in Spike protein per sample across collection dates and countries

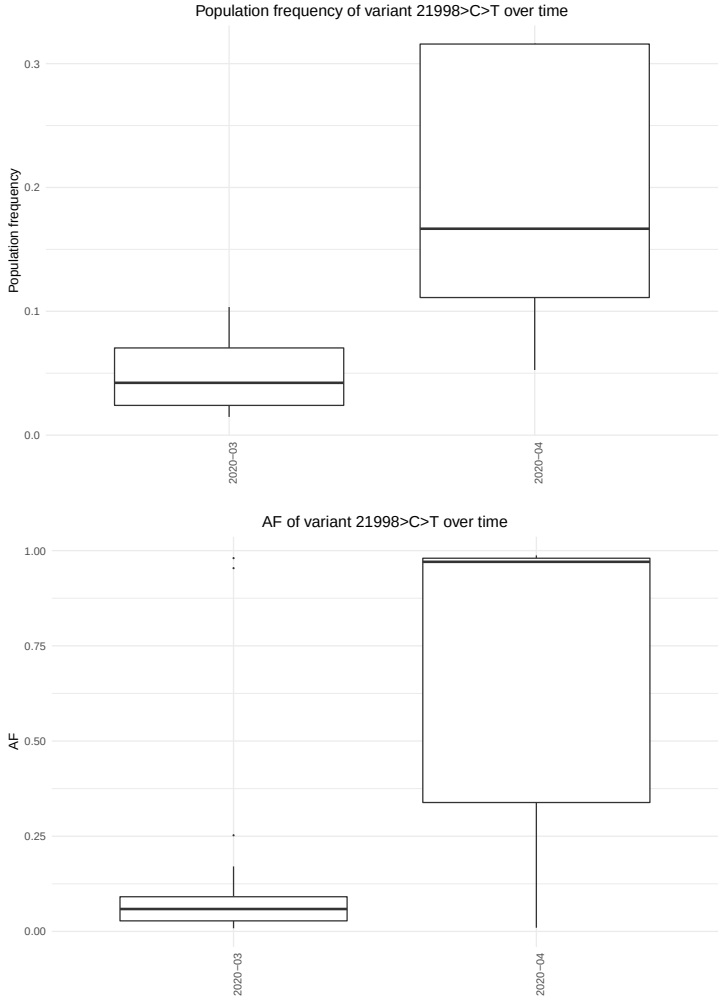

Figure 4: Population frequency and Mean AF of variants "21988C>T" by month

5 Use case 4: Finding co-ocurring missense variants in S proteins

Looking into the number of co-existing missense variants in S protein might reveal whether there is a pattern/constriction in the variants found in the gene so far, such as some signal of co-ocurrence or mutual exclusivity indicating some functional effect.

Figure 5 shows missense variants that are found with AF in run >0.3, which ampunt to 210, whcih are distributed across geographic locations 5. The maximum number of such mutations co-existing in one sample is 24, occurring in one sample from Australia.

Figure 5: Number of variants in Spike protein per sample across collection dates and countries

User can download this data to further analyze statistically whether some of this co-ocurring variants are found recurently co-ocurring.