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

Contents

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 Missense variants in S protein

By using *Query by feature* specifying gene:S and using filters to filter Functional Class to "MISSENSE" the user can get the result of all missense variants on S gene, along with their dataset frequency 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 1 panel A shows the distribution of positions in S1 protein harboring missense variants in the dataset (2216) and the prevalence of variants (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). Panel B shows the number of variants at each prevalence value. Only 1 variant s present in over 30 % of samples and only 10 variants over 10 %.

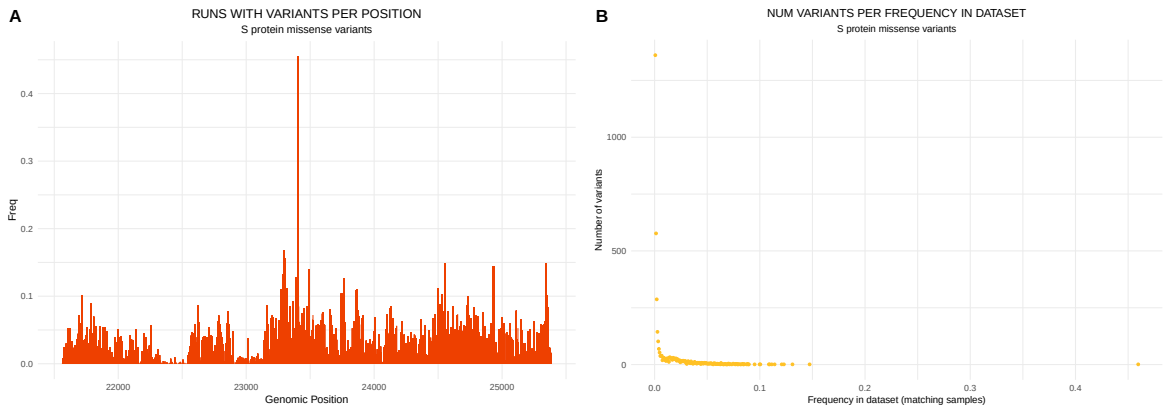

Figure 1: A) Frequency in dataset (matching samples) per genomic position. B) Number of variants by prevalence (frequency in dataset)

2.1 Prevalent variants within S protein

By using *Filter* to select only variants over 0.3 of dataset frequency, 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 that it is found in samples from different geographic locations (UK, USA, Australia, China, Peru and Egypt), it lands on Spike S1, the main subunit for receptor interaction (Annotation from Uniprot) and the earlier date in which the variant appeared in dataset is "2020-02-07" and is present in samples from February to April. Mean AF of this variant is 0.9498617 (i.e, fixed) and has been in increase over time (Figure 2), suggestive of selective pressure.

This variant has been described on paper of July 3 as increasing the viral infectivity (Korber, 2020).

Curiously, at same position there are 3 samples with minor another variant giving different aminoacid substitution ("23403>A>G", "D614V", mean AF 0.0086), that appeared in Australia study only in

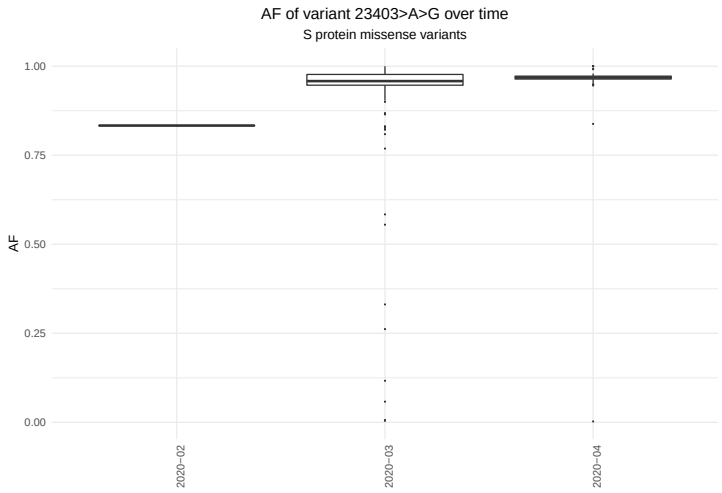

Figure 2: A) Frequency in dataset (matching samples) per genomic position. B) Number of variants by prevalence (frequency in dataset)

”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 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 (Unitprot annotation).

...

3 Missense variants in 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 Prevalent missense variants in RBD

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 could be interesting to explore further those variants having a higher AF (Figure 3)

...

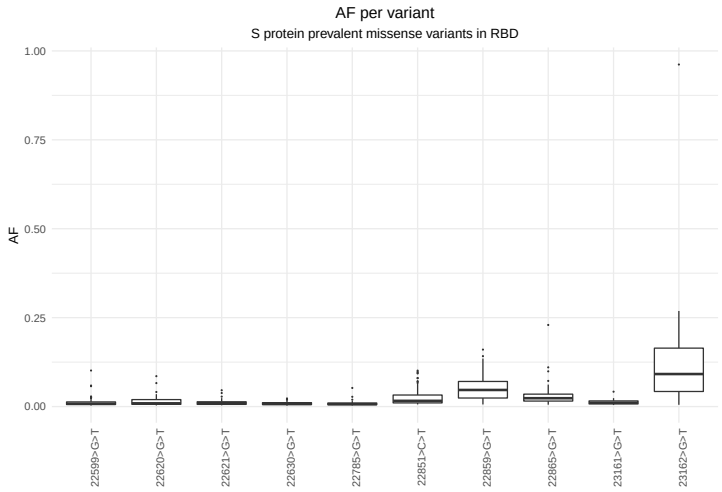

Figure 3: Distribution of AF in runs of prevalent missense variants (>5 %) samples