Ideas for Variants Statistics in Beacon

1 Number of variants per functional region (proteins, utrs, stem loops)

Once we have annotation on protein products on VCF, or we load coordinates from refseq annotation file, in addition to being able to filter by dropdown many in web, we could add info of number of variants that have been found per protein/other genomic regions and their frequencies.

1.1 Data on mutations on coding regions

Following graphs show number of variants on coding regions based on data available in http://cov-glue.cvr.gla.ac.uk/. We could do this (also for non-coding regions) if we had good annotated VCFs or we used RefSeq annotation file coordinates to calculate these on the fly in beacon.

Additionally, it would be interesting to calculate and show maybe behind calculated expected number of variants per region under null model, so possible selection preassure on regions is easily spotted.

1.1.1 Nonsynonymous mutations derived aminoacid replacements

Nonsynonymous mutations have been found in all 26 mature proteins, including coding region of new ORF10.

It would

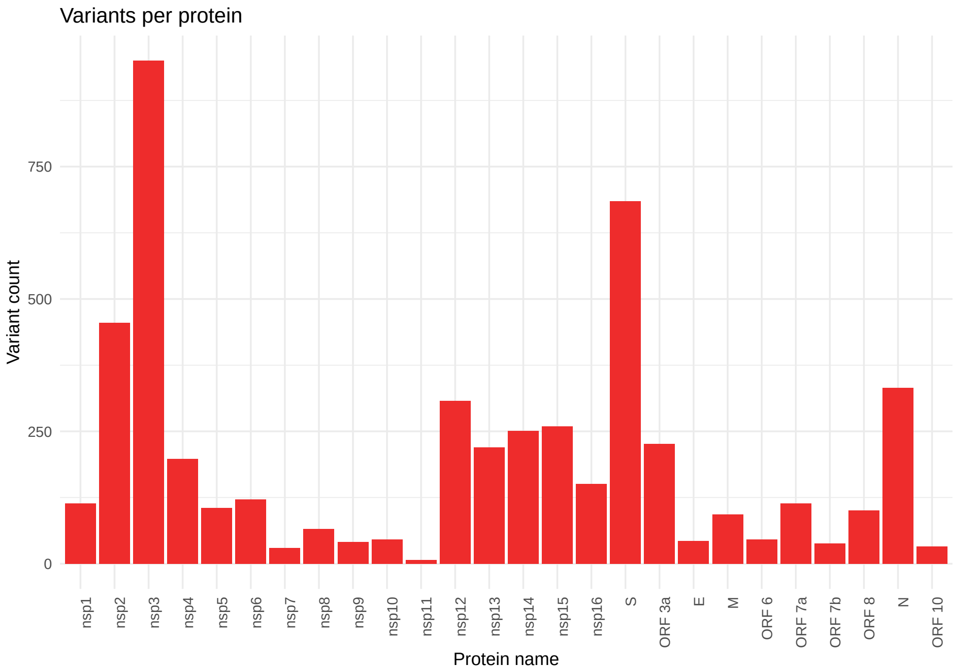

Figure 1: Aminoacid replacements per protein in SARS-CoV2. Note that here all mutations leading to the same change are counted as one.

Frequency (counts) of aminoacid changes.

It is interesting to find co-evolving mutations among those with similar frequencies (co-ocurring mutations). For isntance, nonsynonymous mutations in nsp2, ORF 3a and N appear in virtually the same number of sequences. Are those the same?

For example, some authors have found one mutation in RdRp that could have lead to faster mutation rate to be co-ocurring with presumably later mutations.

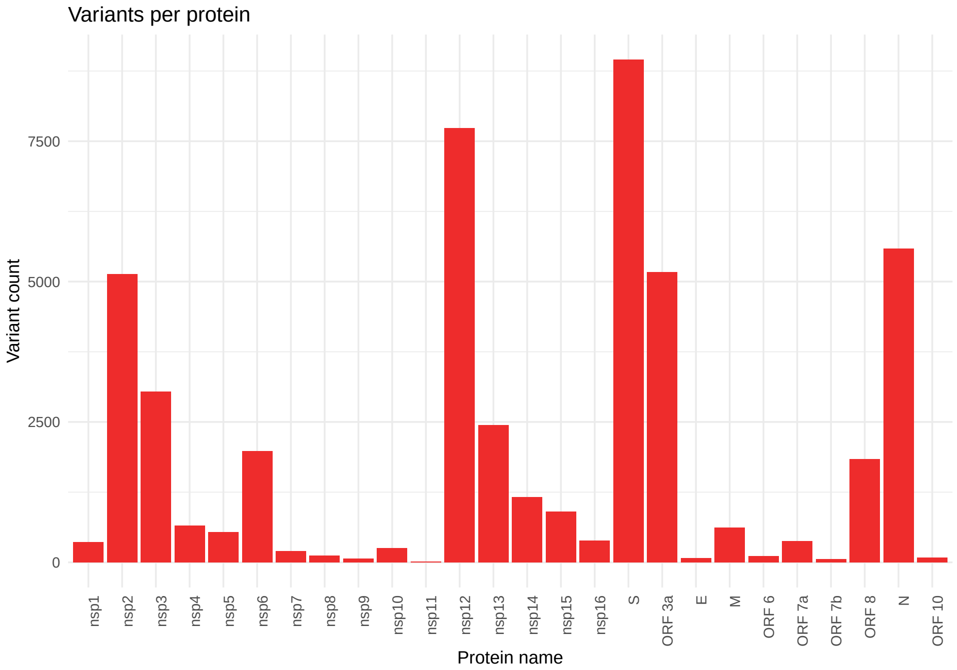

Figure 2: Frequency (counts) of sequences with aminoacid replacements per protein in SARS-CoV2. Note that here all mutations leading to same change are counted as one.

Additional statistics for coevolution of nonsynonymous mutations would include