Viral Beacon Variants Statistics - SRA Illumina-Galaxy

Contents

1 Summary Stats

1. Number of positions with variants: 26366 (88.2%)
2. Number of positions without variants: 3540 (11.8%)
3. Frequency of runs with variants per position, figure 1 Note: This is based on runs matching variant as major one, not all runs with the variant (pending).

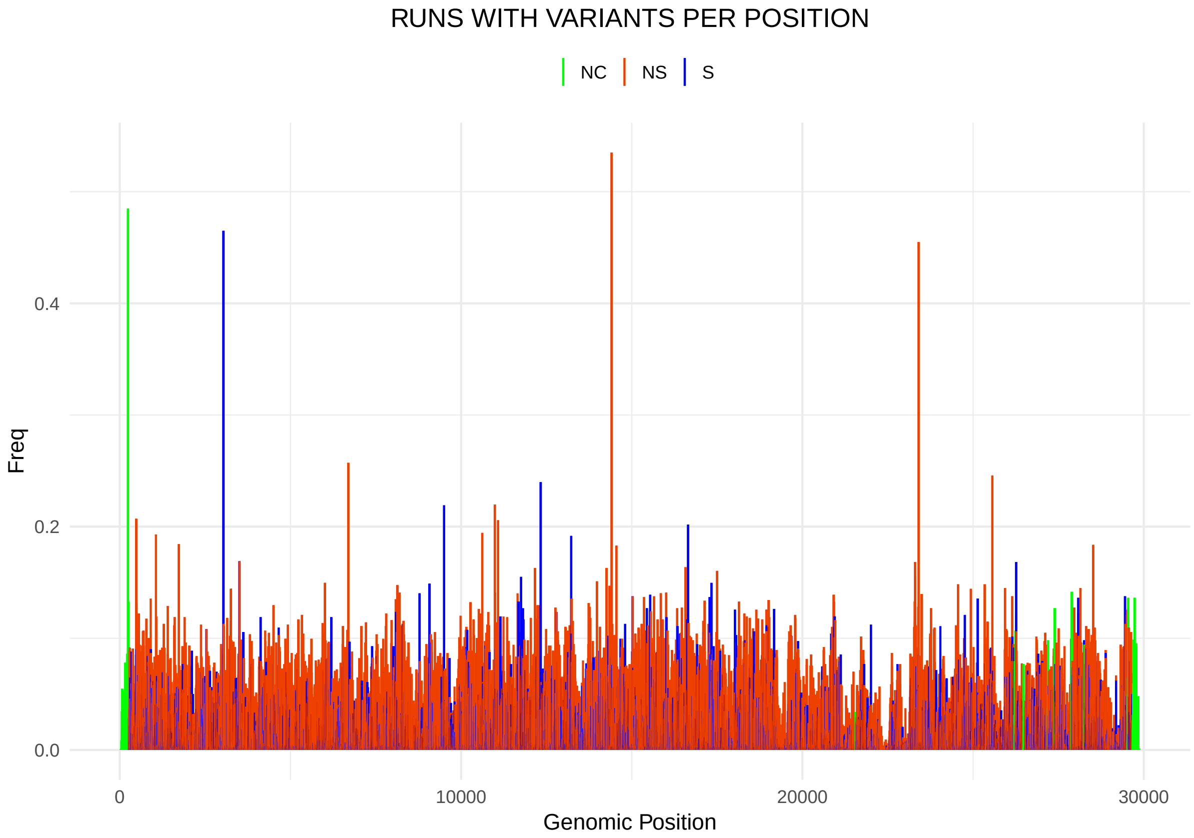

Figure 1: Needle plot: Frequency (proportion of runs) with variants per position

Dataset shows regions with high variability as well as regions (notably, a continuous region between 23000 and 24000) with a lack of variability, visible also in figure 3 and panel A figure 4. It would be interesting to address whether this is a technical problem or could be an effect of negative selection.

4. Number of variants in dataset: 46359

   by variant type: Number of variants by variant type: SNP: 100 %, MNP: 0%, INDEL: 0%, OTHER. 0 %
   Note: Galaxy is updating pipeline to get other variant types.

   by genomic region: coding: 45484 (98.1%), intergenic: 268 (0.58%), 5UTR: 336 (0.72%), 3UTR: 271 (0.58%)

   by dataset AF group: Number of variants by AF groups, figure 2

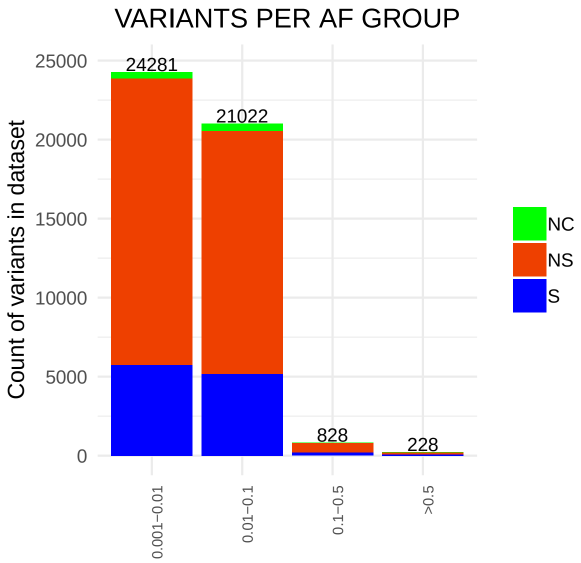

Figure 2: Number of variants per frequency (AF in dataset) group, by functional class

2 Invariant, Non-Polymorphic and Polymorphic sites

There are some positions showing no variants so far, while others show a unique alternate and many positions polymorphism at dataset level.

PMS: positions with more than 1 variant (alternate)

Polymorphism at sample (run) level would be more interesting but is not available for now.

Whether this numbers are compatible with neutral evolution or are indication of positive or negative selection should be explored.

1. Number of positions by number of alternates per position: 0: 3540 (11.8%), 1: 11135 (37.2%), 2: 10469 (35.0%) 3: 4762 (15.9%), figure 3

   (a) Number of polymorphic positions (dataset level): 15231 (50.9%)

2. Number of polymorphic positions at sample-level: not available (sample matching for low freq variants is pending)

3. Max AF vs Min AF per position, figure 4

4. Number of shared polymorphic positions (polymorphic in ¿ 1 sample) at sample level : not available (sample matching)

5. Number of samples with variants at polymorphic positions: this are samples with major ones only for now in. Polymorphic sites in samples will be calculated when AF in run is available)

Intrahost variants are variants in polymorphic sites that are present in a minor subpopulation of the viral quasiespecies.

We don't have at the moment sample matching for non-major variants or AF in run.

However, there are variants that have a low frequency in population (dataset), but relatively many samples bear it (figure) panel D, which should represent intrahost variants.

Also, there are positions with major and minor variants.

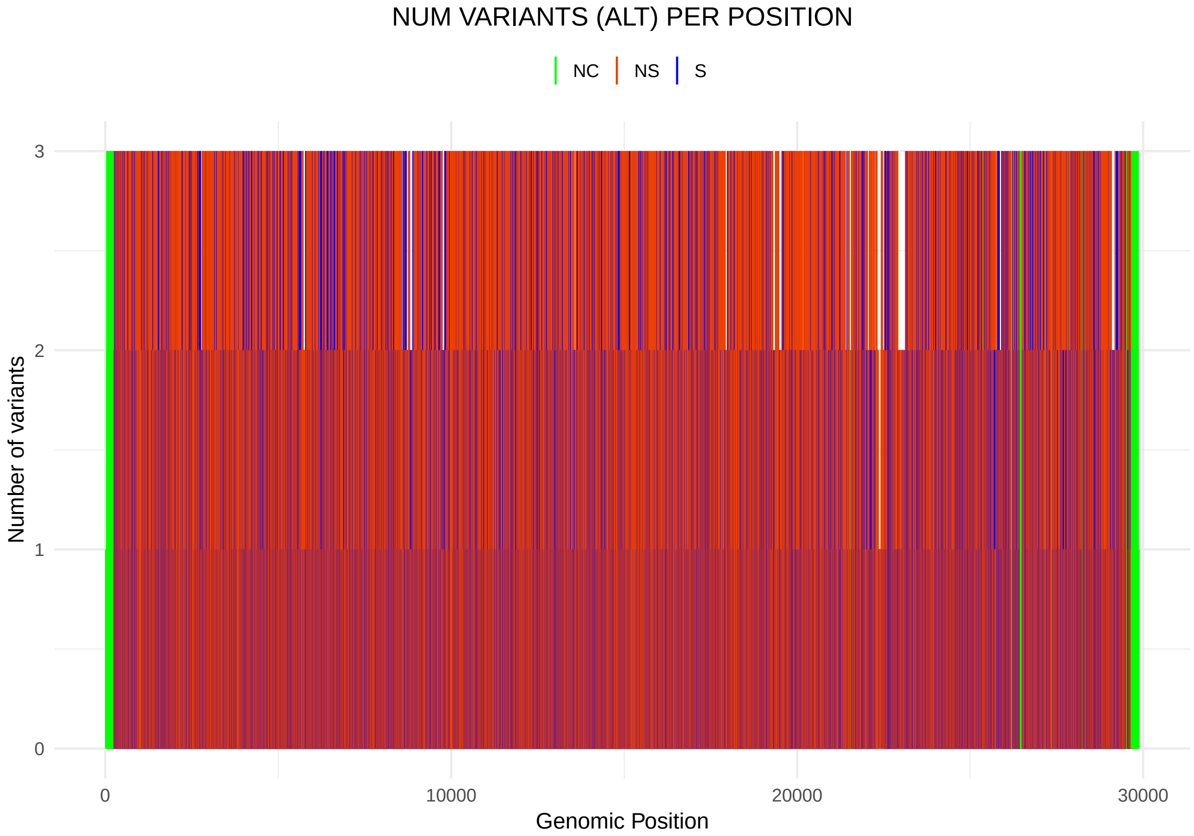

Figure 3: Number of variants per position

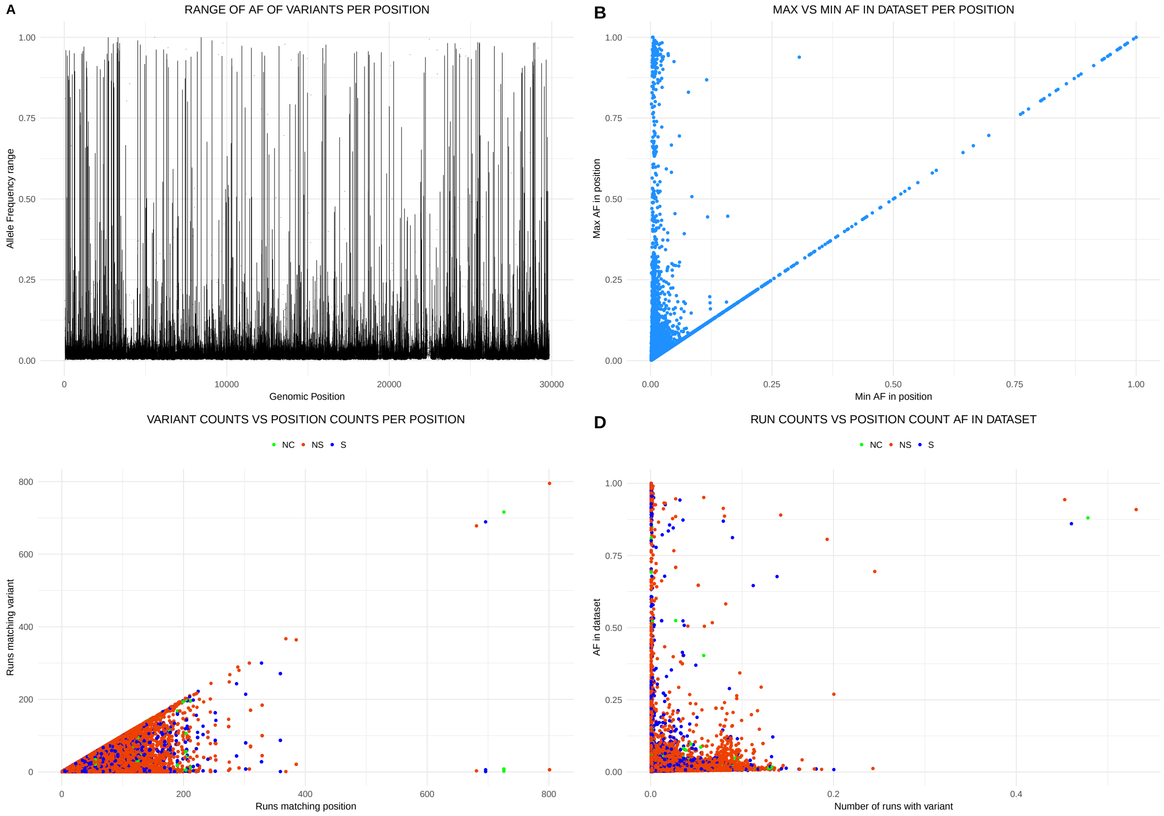

Figure 4: A) Range of AF in dataset of variants at each genomic position. Long lines imply that positions have high frequency and low frequency variants in dataset. B) Relation between max and min AF in dataset per position, variants outside the rect represent variants in polymorphic sites. C) Relation between the number of runs bearing a variant and their maximum AF in dataset. Variants under the the rect represent variants at polymorphic sites, hitting more often a small fraction of runs. D) Relation between the AF in dataset and the number of runs bearing the variant, showing variants with low frequency in dataset that are present in a high number of runs (but also the opposite: see Warnings)

.

3 Per region statistics

The distribution of variants by genomic regions would allow researchers to search for evidence of natural selection.

In particular, variants in coding regions allow to assess positive and negative selection by comparing the observed NS/S ratio in the functional regions, where natural selection acts, with the expected for a region of similar size and composition under the neutral model.

1. Number of variants per genomic regions, non-coding regions and genes: figure 5

2. Number of variants per mature proteins: figure 6

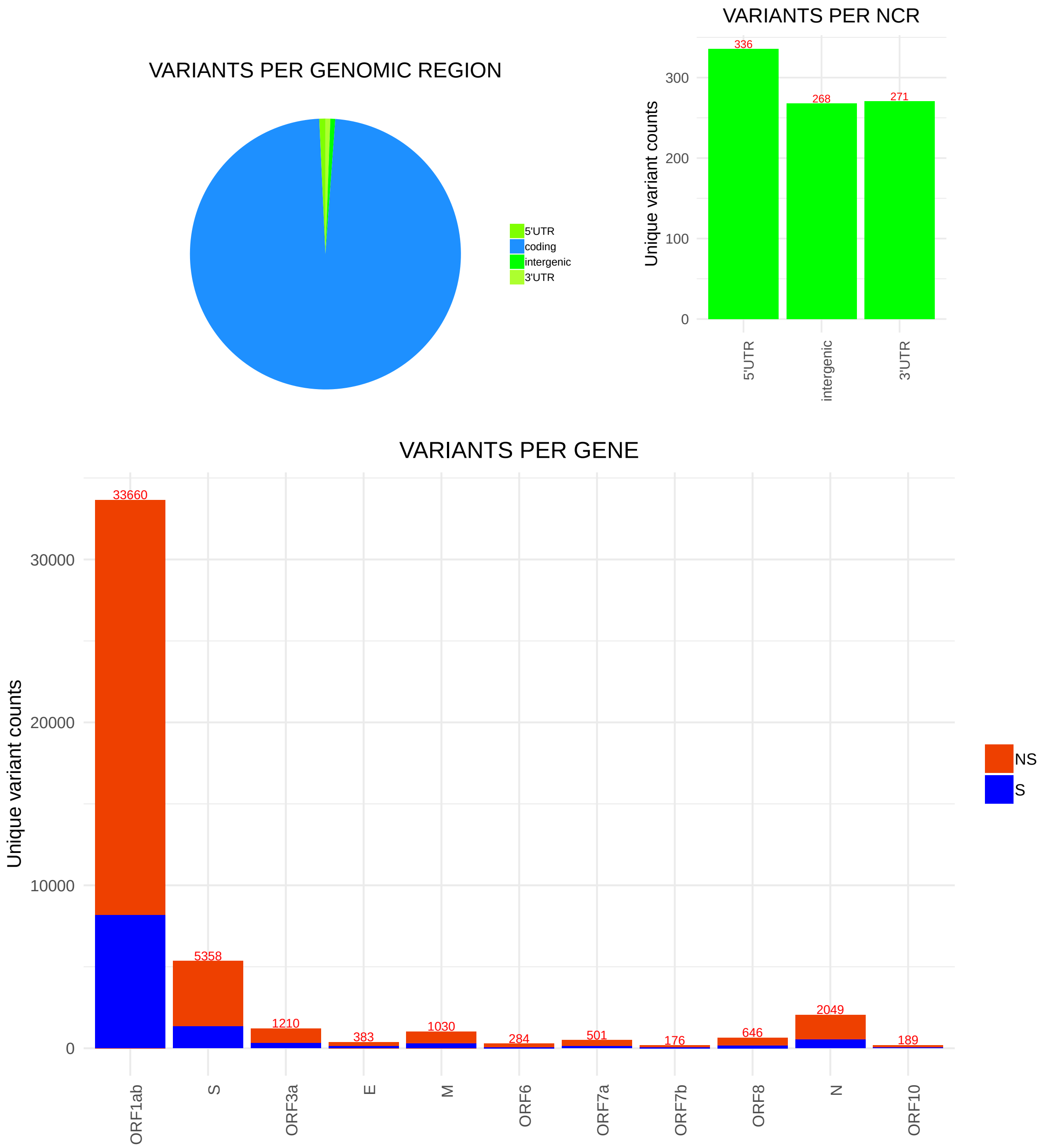

Figure 5: Distribution of genomic variants per genomic region

3. Number of unique aminoacid changes per protein 7

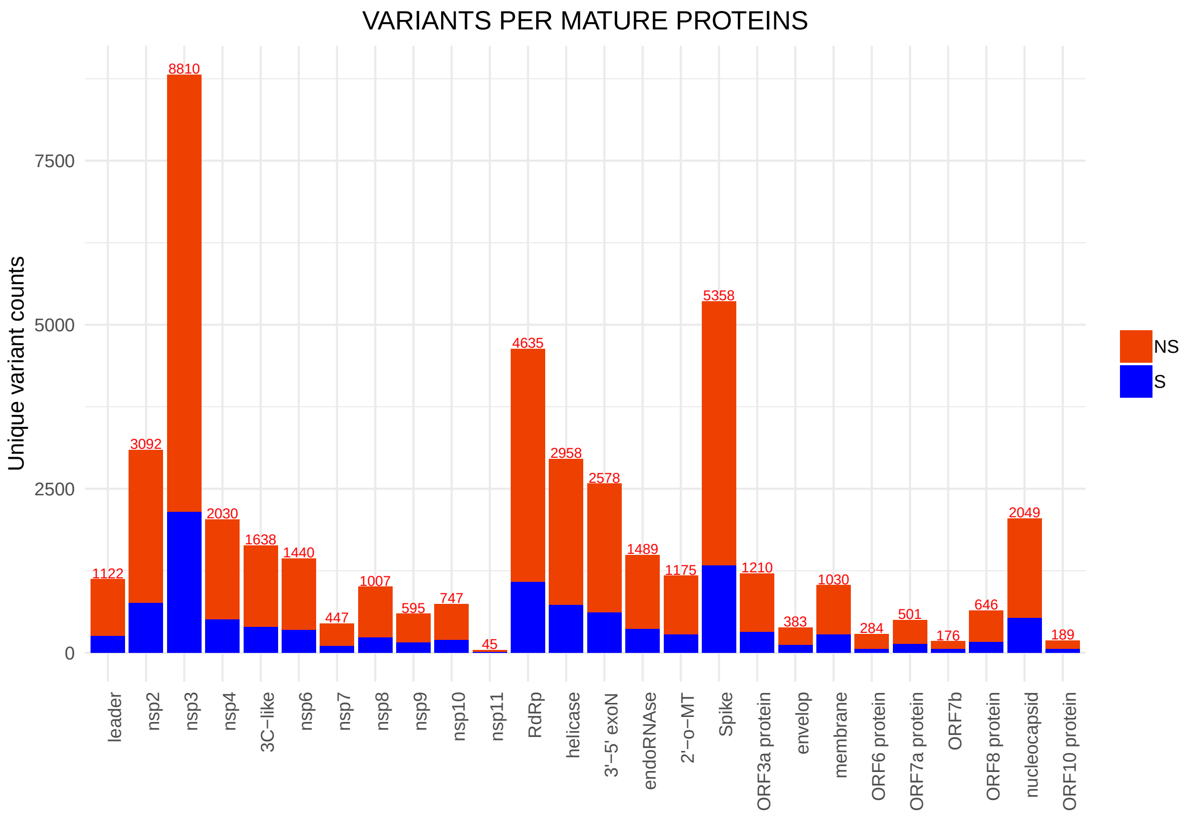

Figure 6: Distribution of genomic variants per mature protein

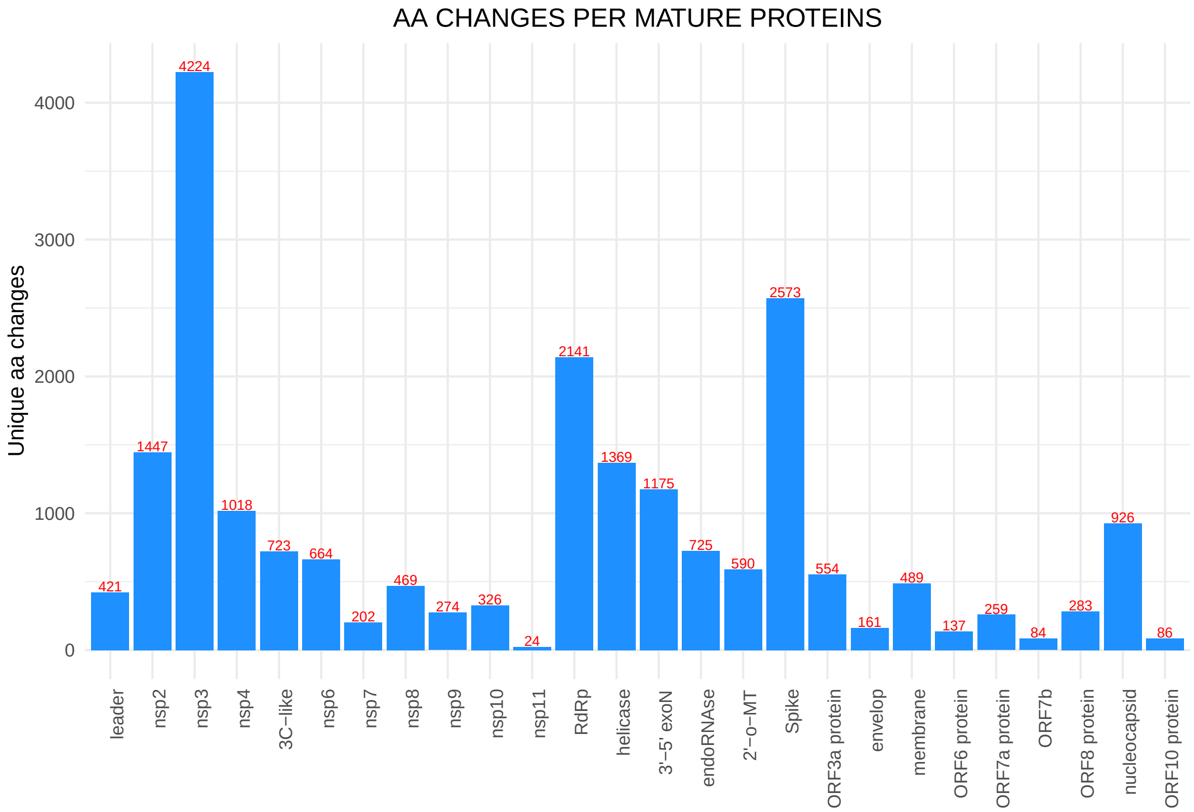

Figure 7: Distribution of genomic variants per mature protein

4 Analysis - to do?

Variants data could be exploited by researchers to assess signatures of evolution, the existance of strains, etc. natural selection: dns/s, spectrum of mutations, convergent evolution (some homoplasies), co-ocurrence, sites with frequency changing over time.

- intrahost reasoning for study and what is needed smc and AF per run only - statistical analysis of natural selection - association with metadata Geo Loc, Sex, Age, Sample site - Positions with no mutations, are they also unmutated in Gisaid/Genbank? Where are them? - Polymorphic sites, are they also so in G/G? - Non-polymorphic, ibidem?

- Are there exclusive variants associated to Geographic Location

5 Warnings

- 3/4 positions with absolute concordance seem wrong, with only 1 sample bearing it and position being polymorphic in dataset. This is weird. I think AF in dataset needs to take into account the sum of DP for all reads covering position in all VCVs and not only the DP in VCFs with the variant.