Viral Beacon Variants Statistics - GISAID

Contents

1 Summary Stats

1. Number of positions with variants: 13987 (46.8%)
2. Number of genomic positions without variants: 15910 (53.2%)) coding: 13494, non-coding: 493
3. Frequency of runs with variants per position, figure 1

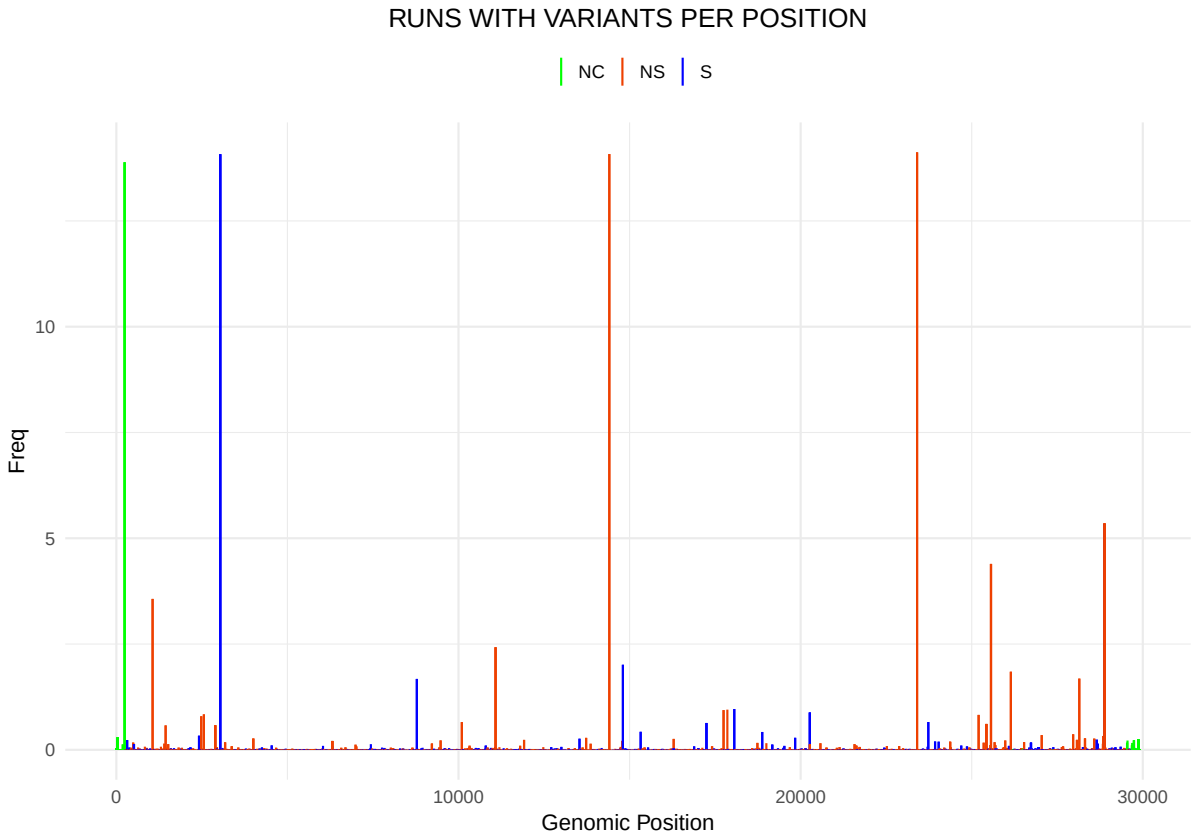

Figure 1: Needle plot: Frequency (proportion of runs) with variants per position

4. Number of variants in dataset: 20288

   by variant type: Number of variants by variant type: SNP: 100 %

   by genomic region: coding: 19107 (94.2%), intergenic: 112 (0.6%), 5UTR: 534 (2.6%), 3UTR: 535 (2.6%)

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

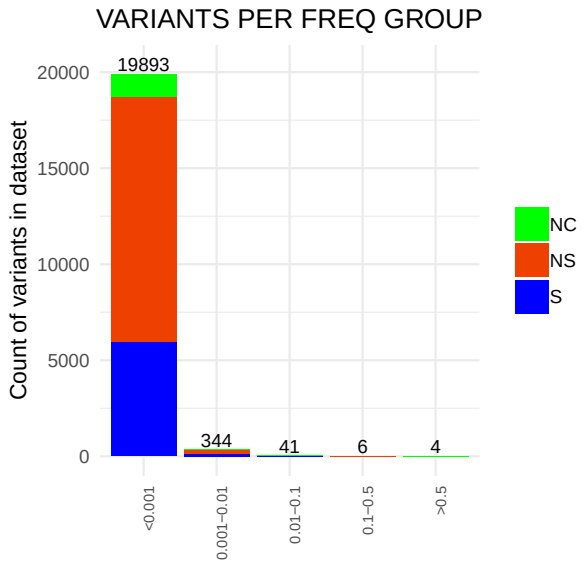

Figure 2: Number of variants per frequency (AF in dataset) group, by functional class

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)

1. Number of positions by number of non-iupac alternates per position: 0: 15919 (53.2%), 1: 9145 (30.6%), 2: 3725 (12.5%) 3: 876 (2.9%), figure 3

   Note: This info is not very useful. We should calculate IUPAC as their corresponding alternates.

   (a) Number of polymorphic positions at dataset level (either more than one alternate in different samples or samples with IUPAC codes): 8139 (27.2%)

2. Number of polymorphic positions at sample-level (IUPAC codes): 7362 (24.6%)
3. Number of shared polymorphic positions (polymorphic in more than 1 sample) at sample level: 1627 (5.4%)
4. Number of samples with variants at polymorphic positions: pending
5. Max vs Min dataset frequency per position, figure 4

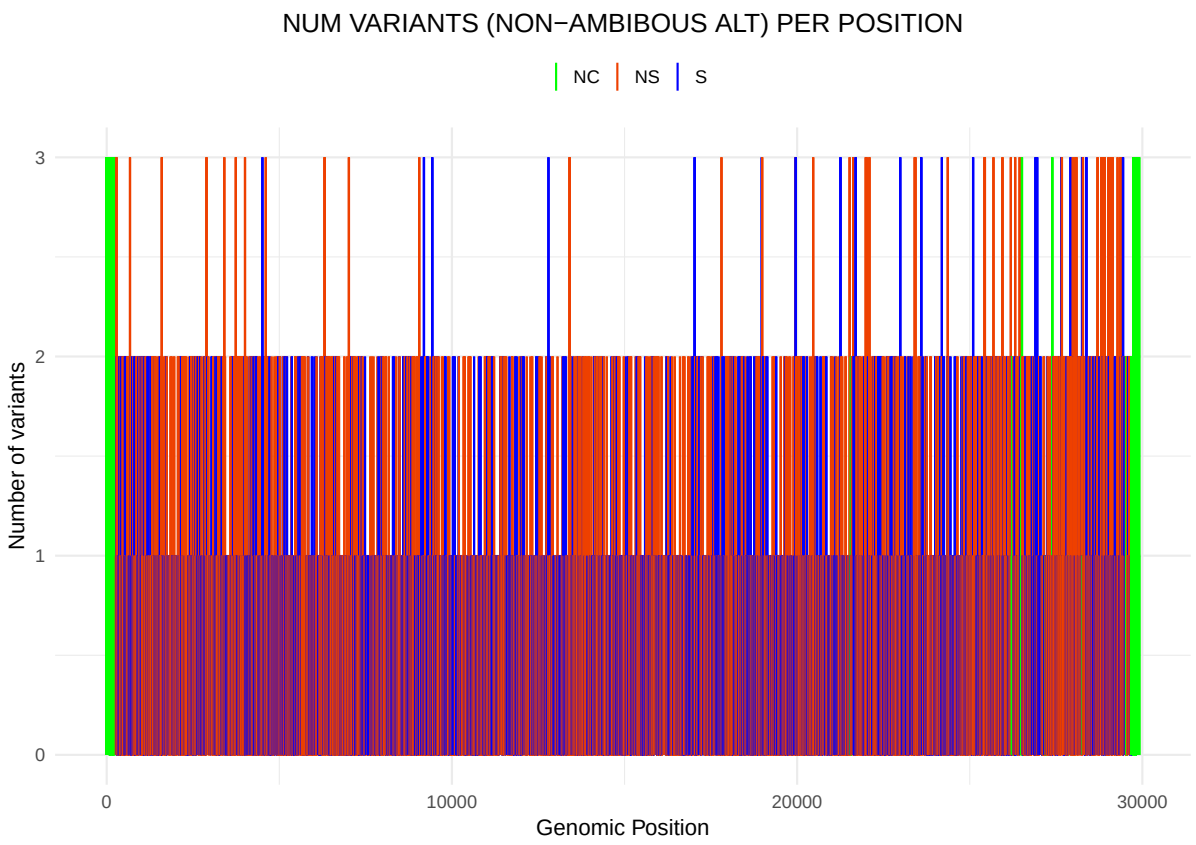

Figure 3: Number of variants per position

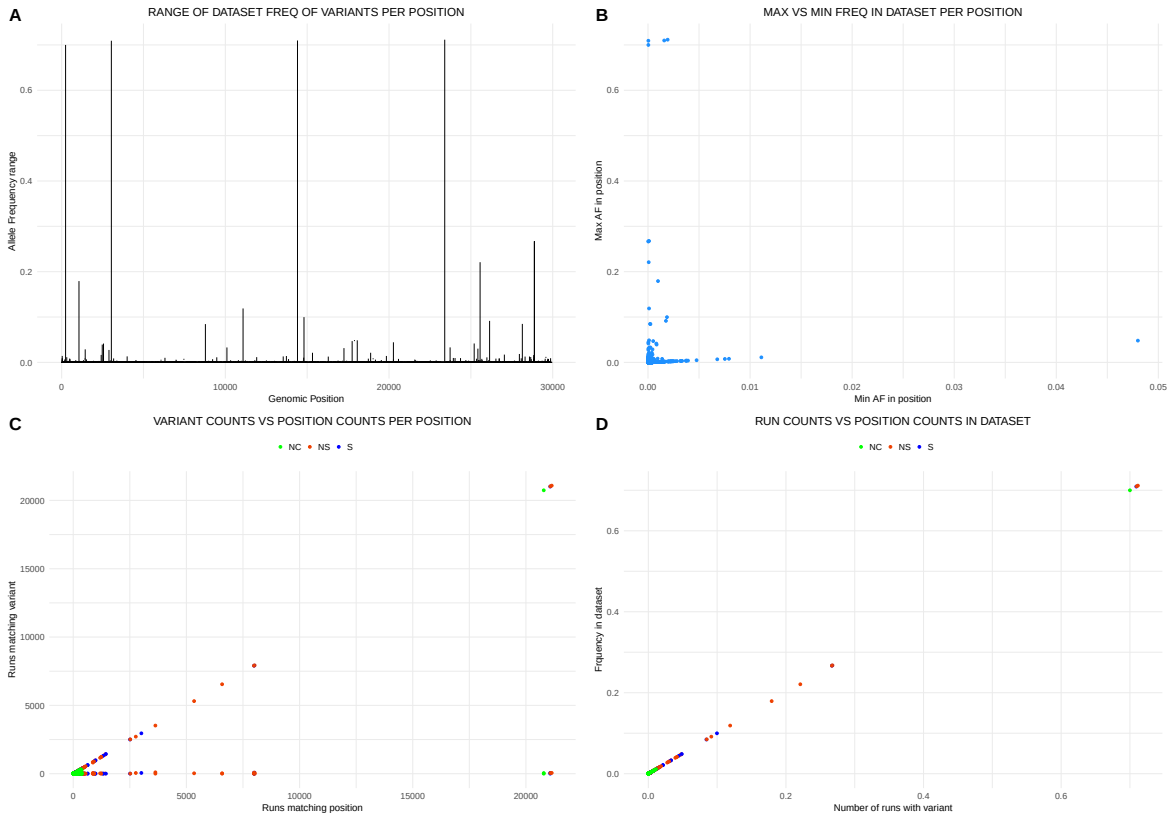

Figure 4: A) Range of dataset frequency of variants at each genomic position. B) Relation between max and min Frequency in dataset in dataset per position, variants outside the rect represent variants in polymorphic sites. C) Relation between the number of runs bearing a variant and their maximum Frequency in dataset. D) Relation between the Frequency in dataset and the number of runs bearing the variant.

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

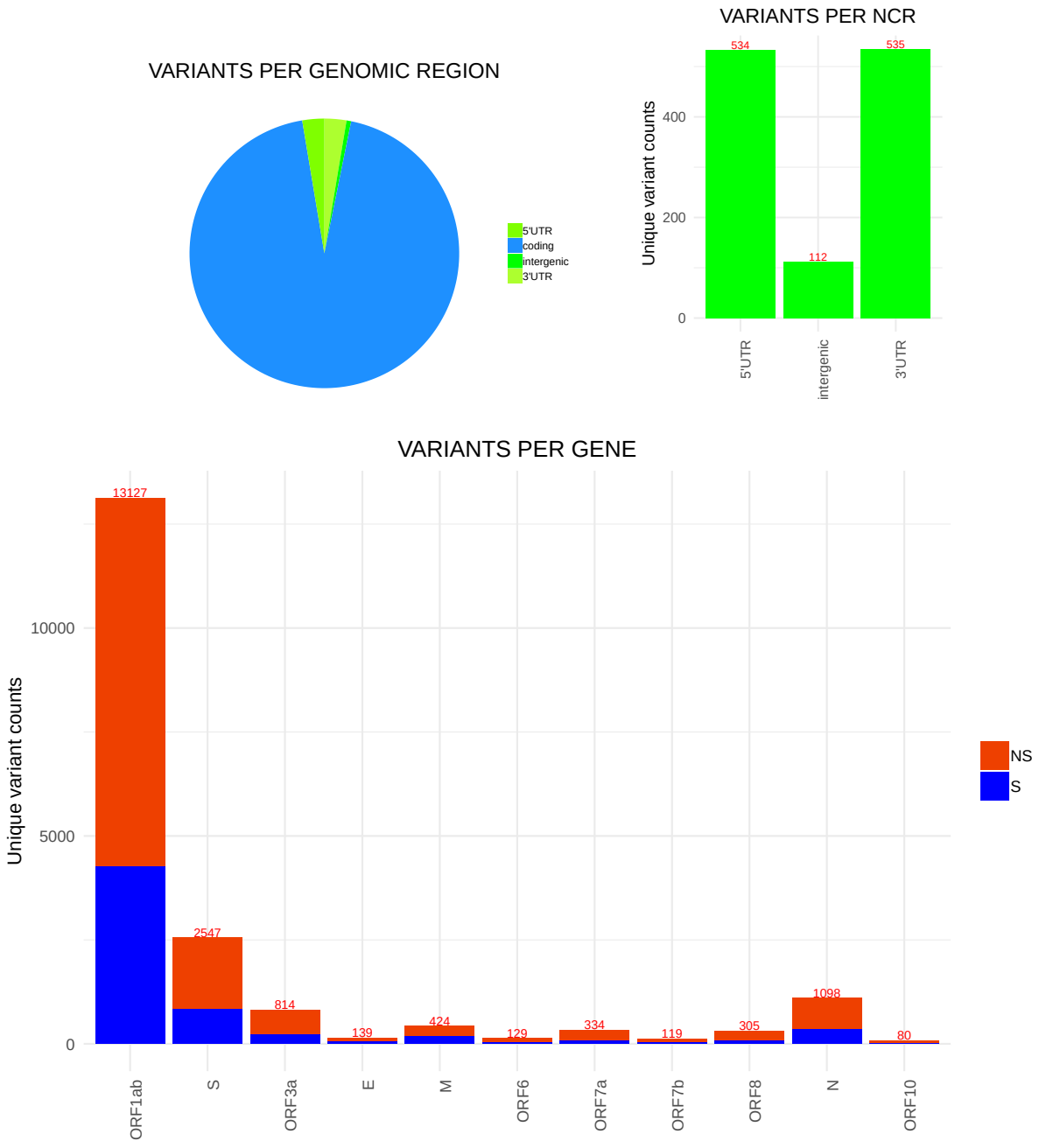

Figure 5: Distribution of genomic variants per genomic region

3. Number of unique aminoacid changes per protein 7

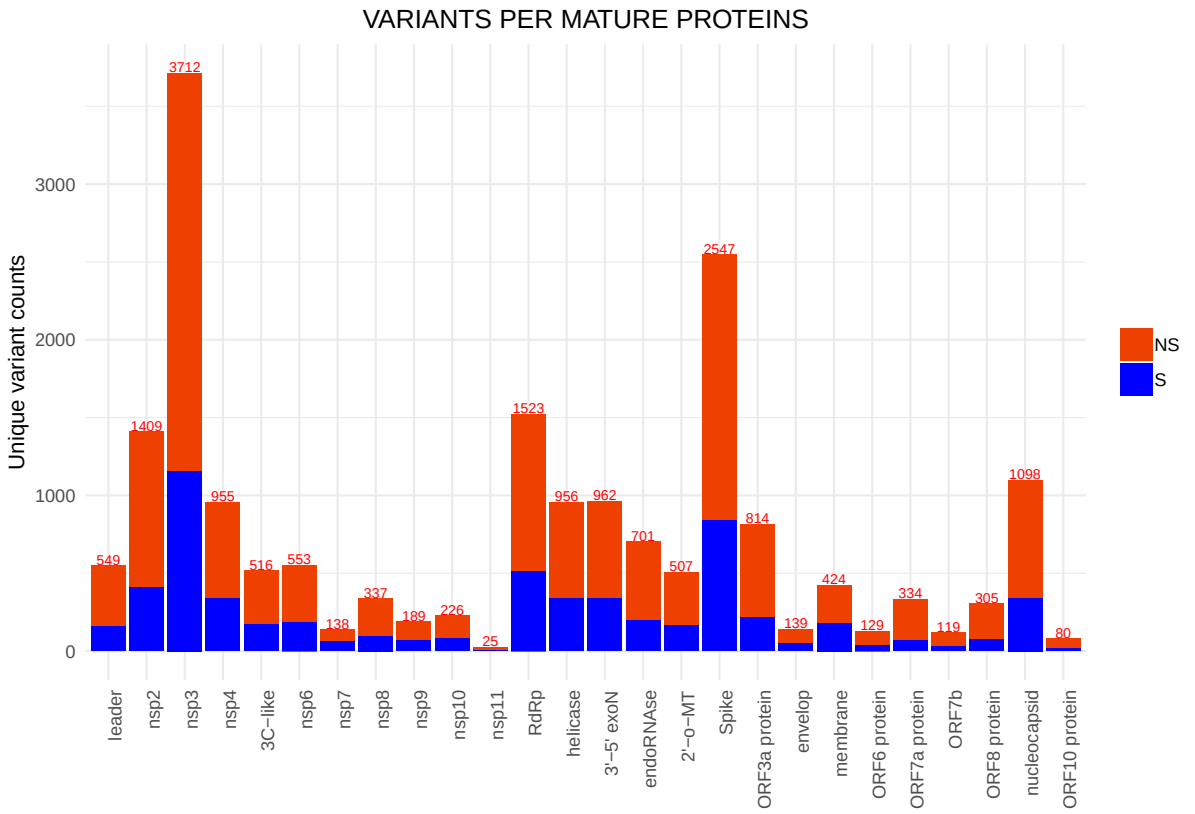

Figure 6: Distribution of genomic variants per mature protein

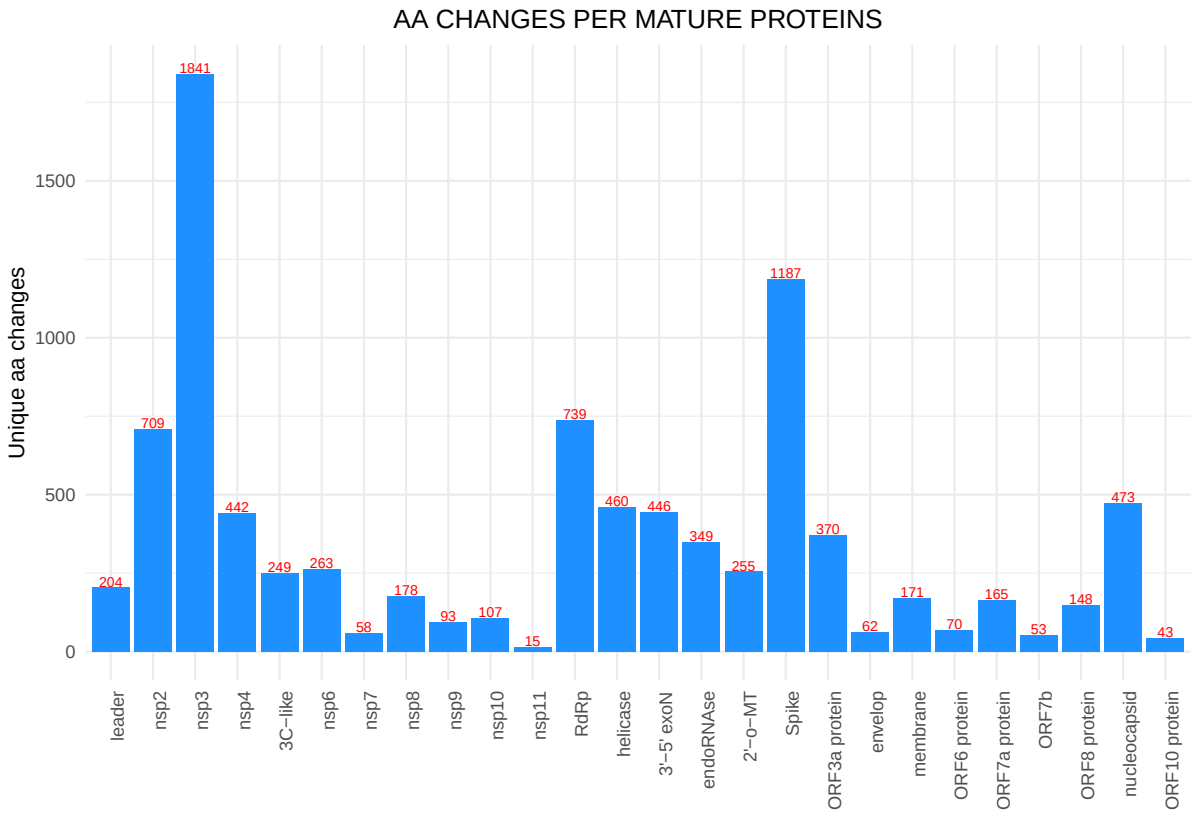

Figure 7: Distribution of genomic variants per mature protein