# Variants Statistics in Beacon

## 1 General Statistics

Statistics are calculated using `Query by region -by names/aliases` and `Filters`.

### 1.1 Stats on sequences (already in help page?)

1. Number of runs: total runIds
2. Number of runs with variants (we are keeping WT?)
3. Number of unique sequence/haplotypes: ?

*option* Split by sequencing technology (platform): Illumina, Nanopore (for now only Illumina)

*option* Split by sequence database: SRA, GISIAD, GWH

*option* Split by Individual geographic region: Australia, China, USA, Singapur, etc

*option* Split by Biosample collection date (month)

*option* Split by Biosample sample.type

*option* Split by Individual sex

### 1.2 Stats on variants

1. Number of genomic positions with variants: 28073/ Number of genomic positions: 29903
2. Variance per position: fig Needle plot (counts of unique sequences/haplotypes or counts of runs having them?)
3. Number of unique variants in database: 34951
4. Graph: Number of unique variants, split by options, e.g fig 1

   *option* Split by sequencing technology (platform): Illumina, Nanopore (for now only Illumina)

   *option* Split by variant frequency (percentiles)

   *option* Split by variant type field: SNP, indels (although for now there are only SNPs?)

   *option* Split by genomic region: coding: all with genomic region=CODING, non-coding: the rest

   *option* Split by molecular consequence (grouped: SYN: SILENT, NON-SYN: MISSENSE+NONSENSE, NONCODING:the rest)

   *option* See Per region Statistics: Distribution in genomic regions: non-coding, gene, cds/mature peptide, stem loops. Number of unique variants are aggregated in regions, show also split by syn/non syn, as in fig 2.

5. Number of positions with aminoacid substitutions in database/ Number of coding positions
6. Number of variants producing unique aminoacid substitutions in database: 18308

*option* See Per region Statistics: Distribution in genomic regions: non-coding, gene, cds/mature peptide, stem loops. Number of unique aminoacid substitutions (aminoacid change) (eg. ”G507C”) are aggregated in regions

# positions with variants: 28073/ # genomic positions: 29903

needle plot here

# unique variants: 34951

by  frequency (percentiles)   by variant type    by region class    by molecular consequence

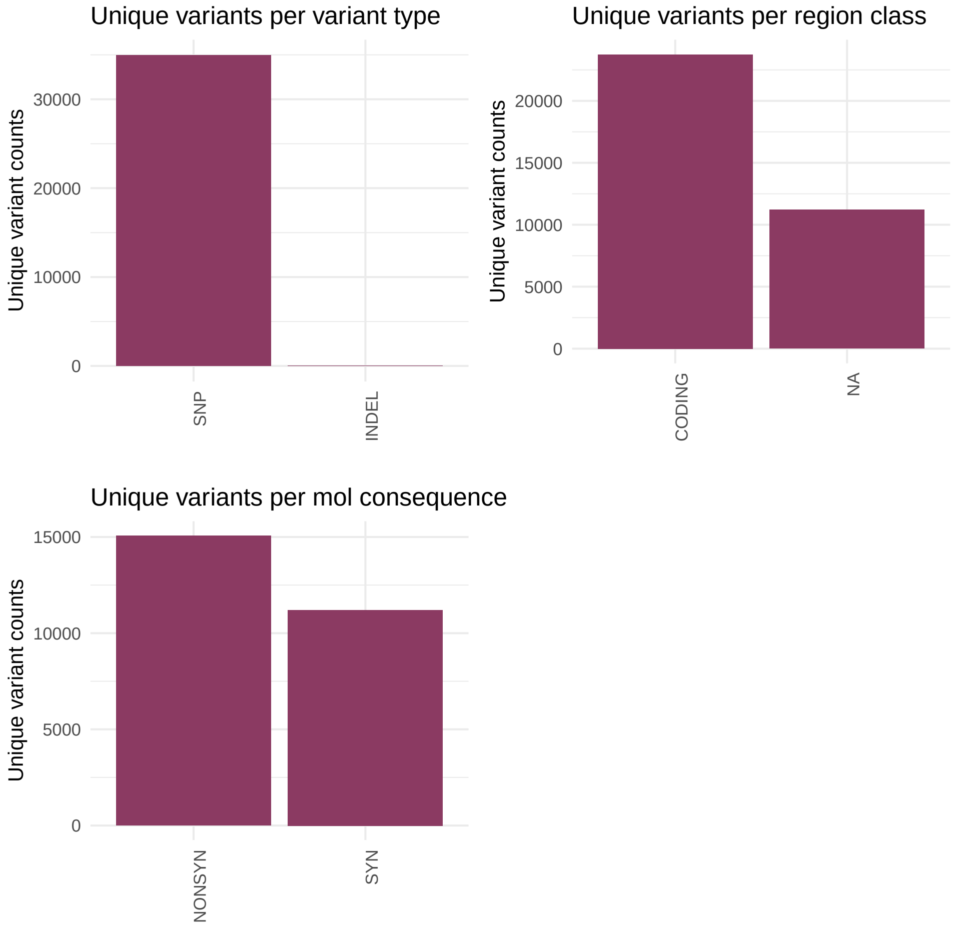

Figure 1: General statistics: Number of unique variants, split by variant type, region class and mol consequence are selected

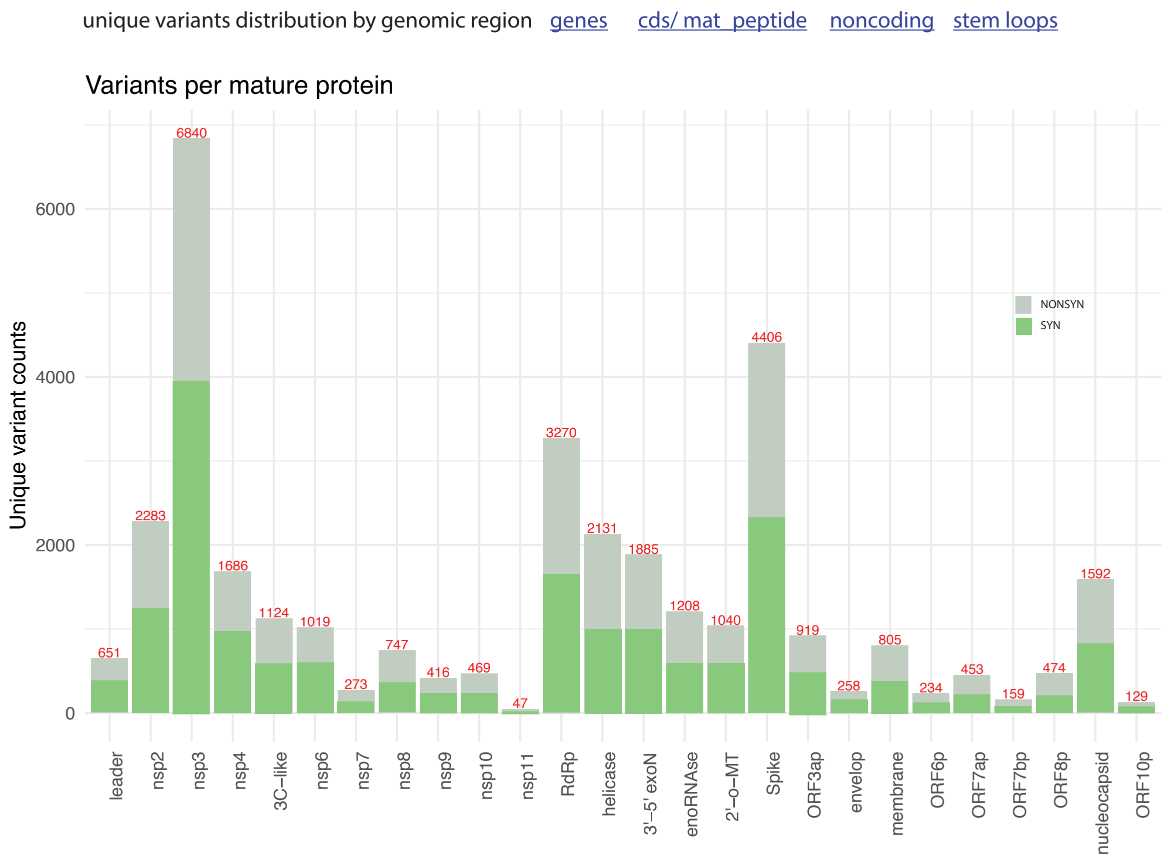

Figure 2: Unique variants in coding regions of SARS-CoV2, shown per mature proteins, shown on clicking cds/mature peptide.