Variants Statistics in Beacon

1 General Statistics

By using `Query by region -by names/aliases` and `Filters` and we can calculate and show in Statistics in Beacon mainpage the following statistics:

1.1 Stats on sequences (maybe table only, 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 geographic region: Australia, China, USA, Singapur, etc

option Split by collection date (month)

1.2 Stats on variants

1. Number of genomic positions: 29903
2. Number of positions with variants: 28073
3. Variance per position: fig Needle plot (counts of unique sequences/haplotypes or counts of runs having them?)
4. Number of unique variants in database: 34951
5. 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 (quartiles)

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

option Split by genomic region field: 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.

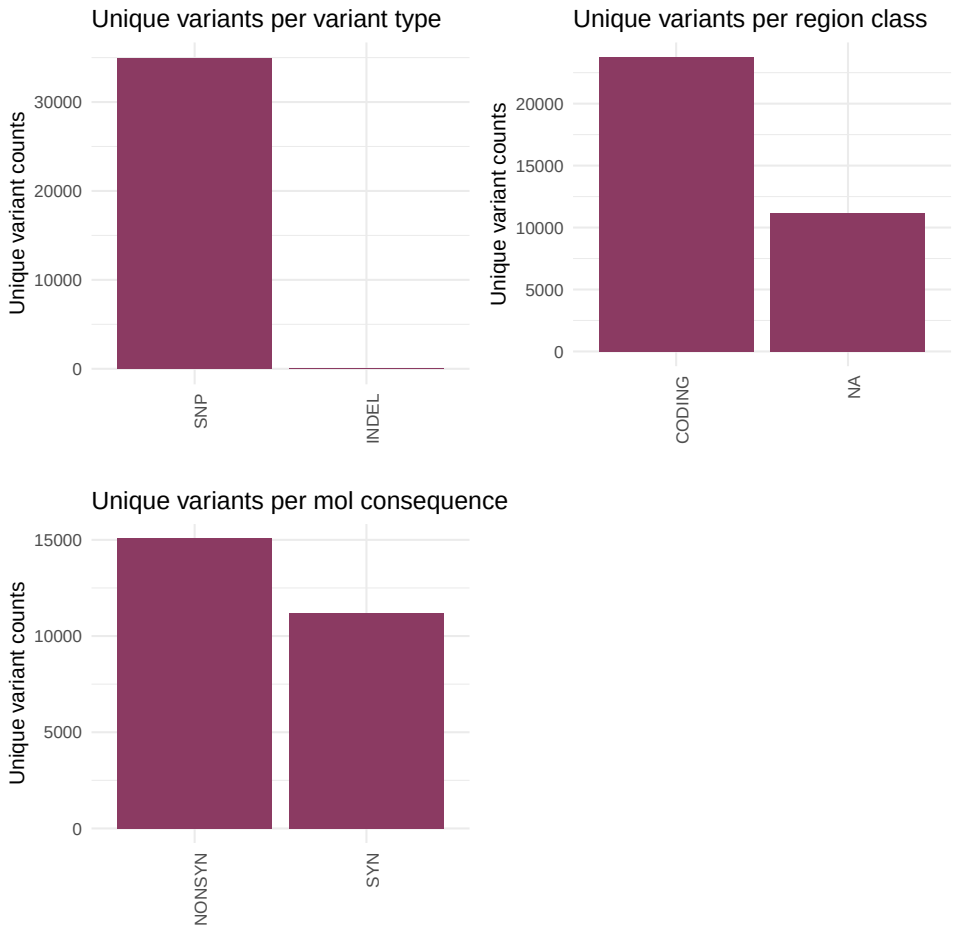

Figure 1: General statistics: Number of unique variants split by variant type, region class and mol consequence

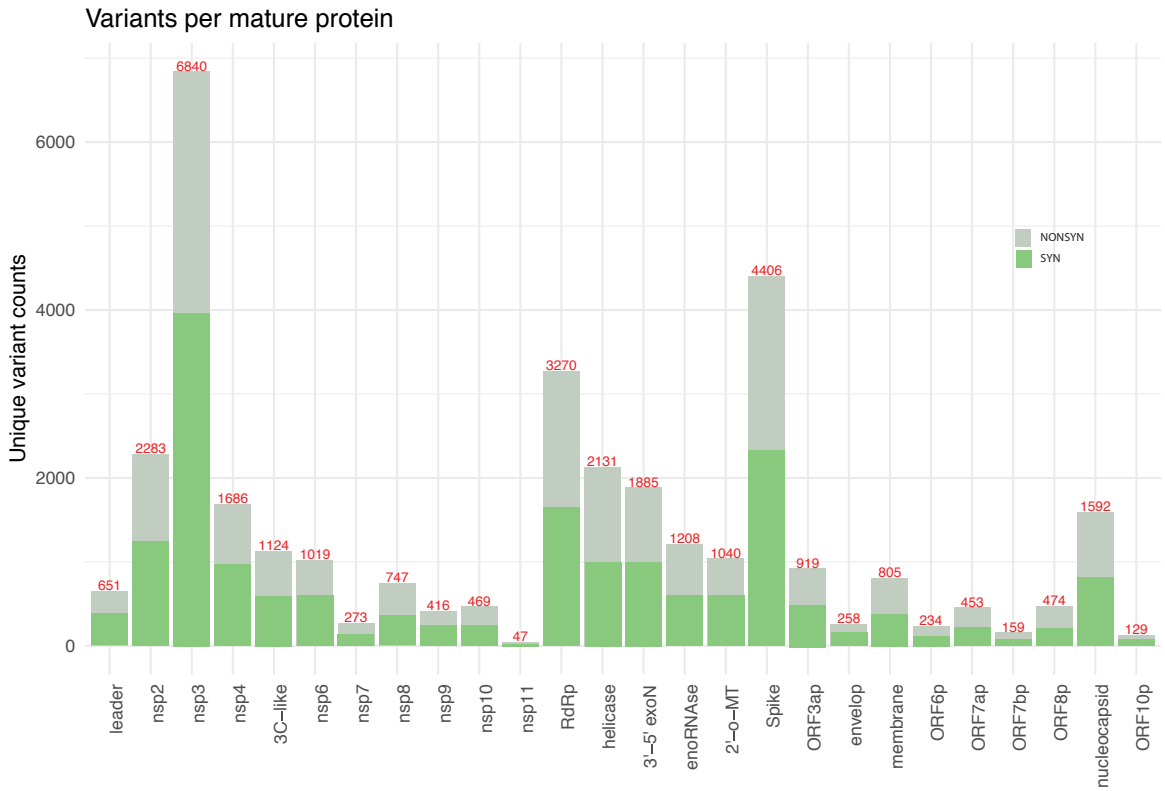

Figure 2: Unique variants in coding regions of SARS-CoV2, shown per mature proteins, shown on clicking cds/mature peptide.