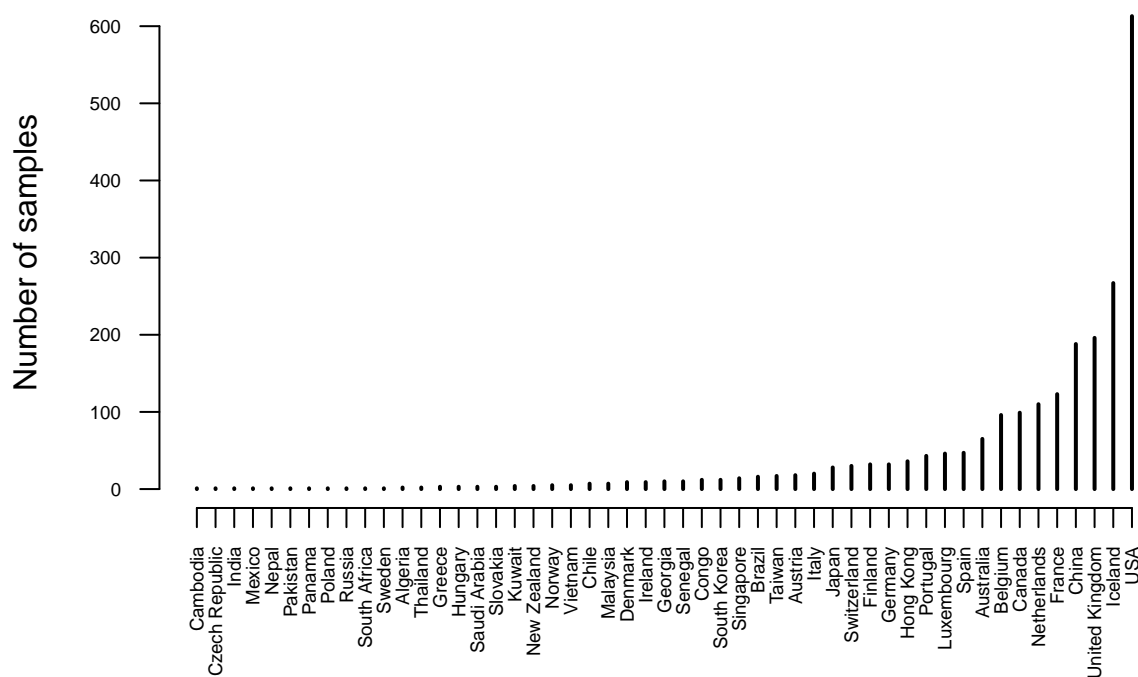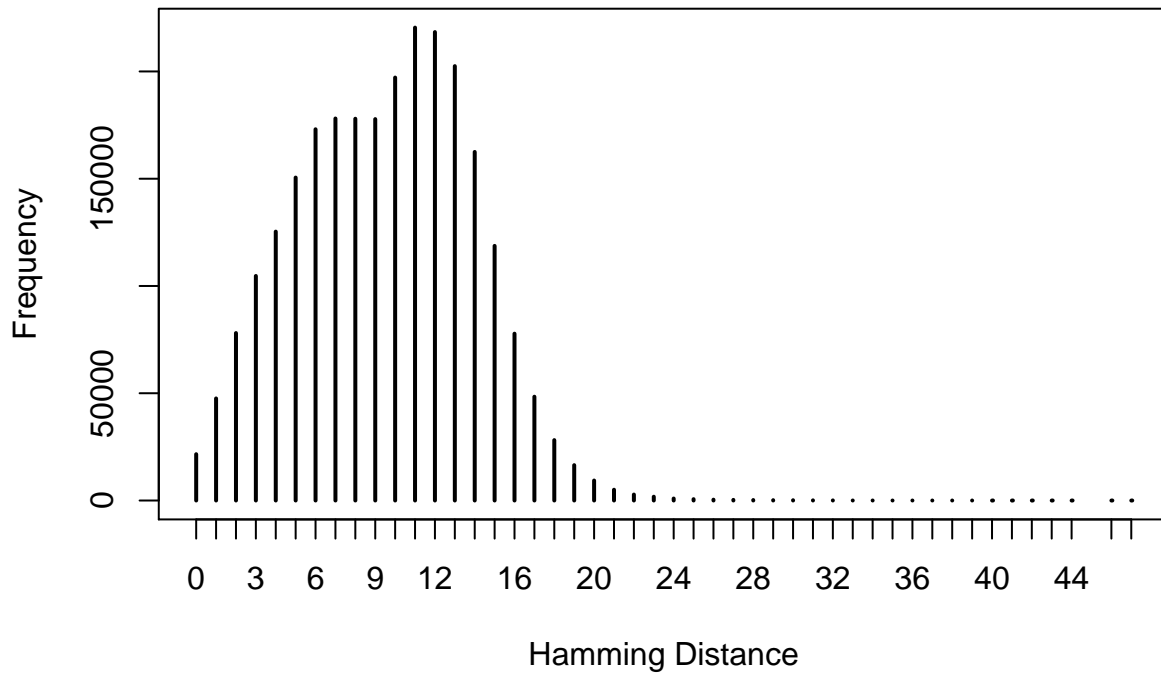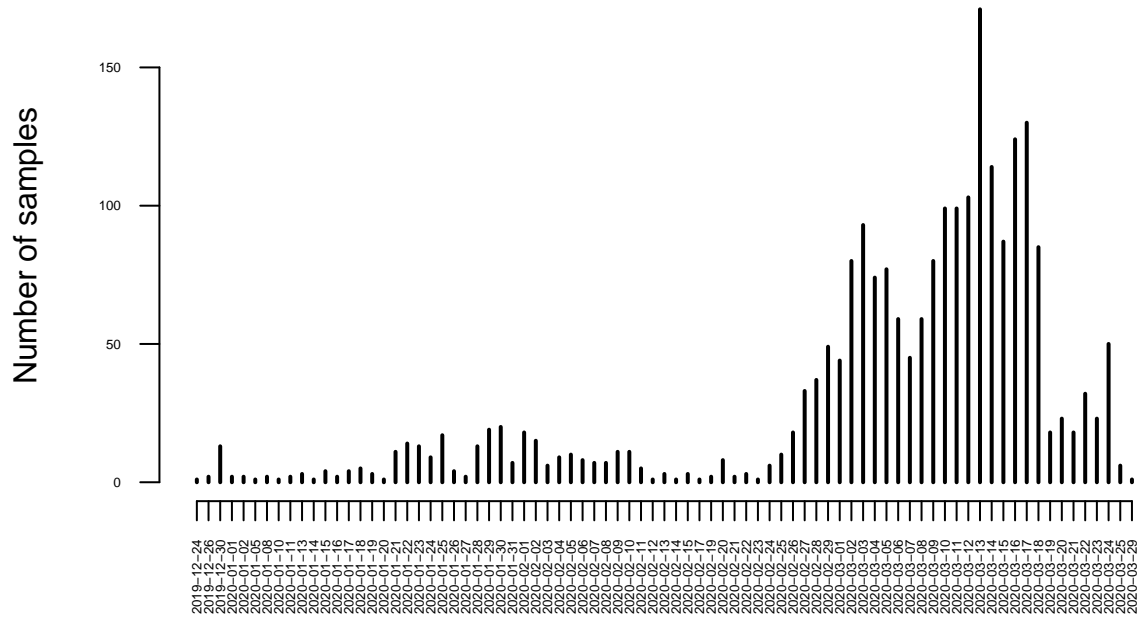# Preliminary Molecular Clock Check and Summary Statistics
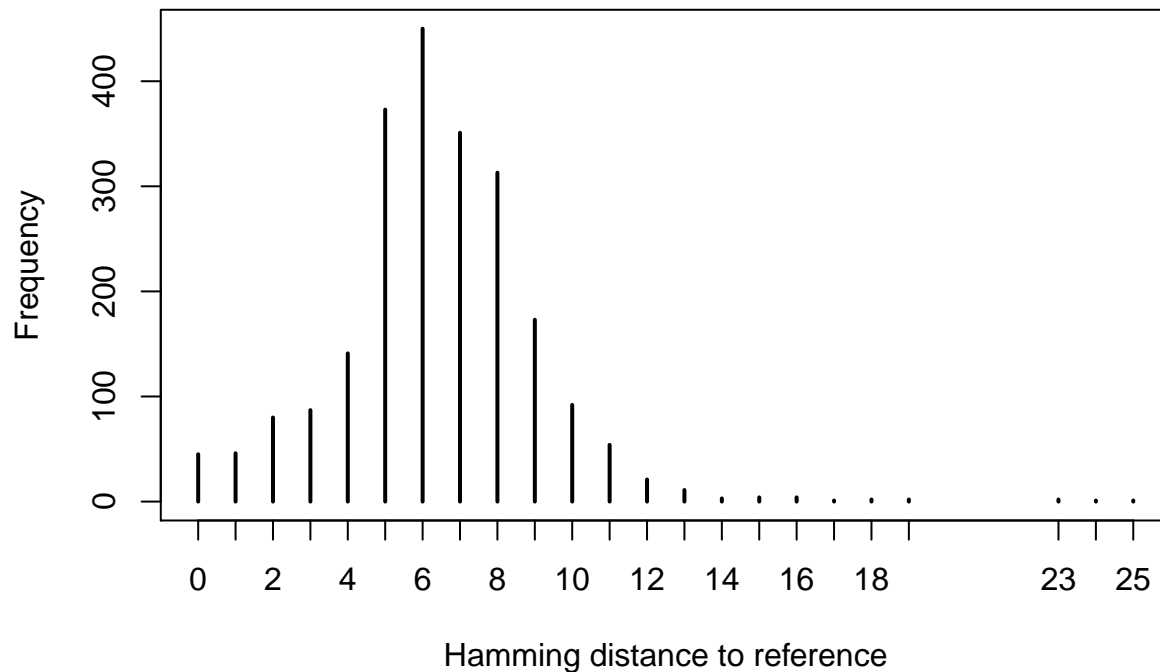
```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

A total of 2257 from 51 countries –– Fri Apr 3 20:14:39 2020

Sequences by date

A rough estimate of the mutation rate can be obtained by regressing genetic divergence on time divergence. This estimate ignores correlation among samples, population structure and pairwise information. A more reliable estimate can be found in the mutation section.

```
##
## Call:
## lm(formula = hamming[ref, ] ~ -1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4988 -1.4084 -0.2277  1.3143 19.8440
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## x 0.0903468  0.0006847   131.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.335 on 2256 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8852
## F-statistic: 1.741e+04 on 1 and 2256 DF,  p-value: < 2.2e-16


##
##  336  738 1076 1077 1159 2036 2171 2172 2173
```

A total of 2.53 of mutations per month
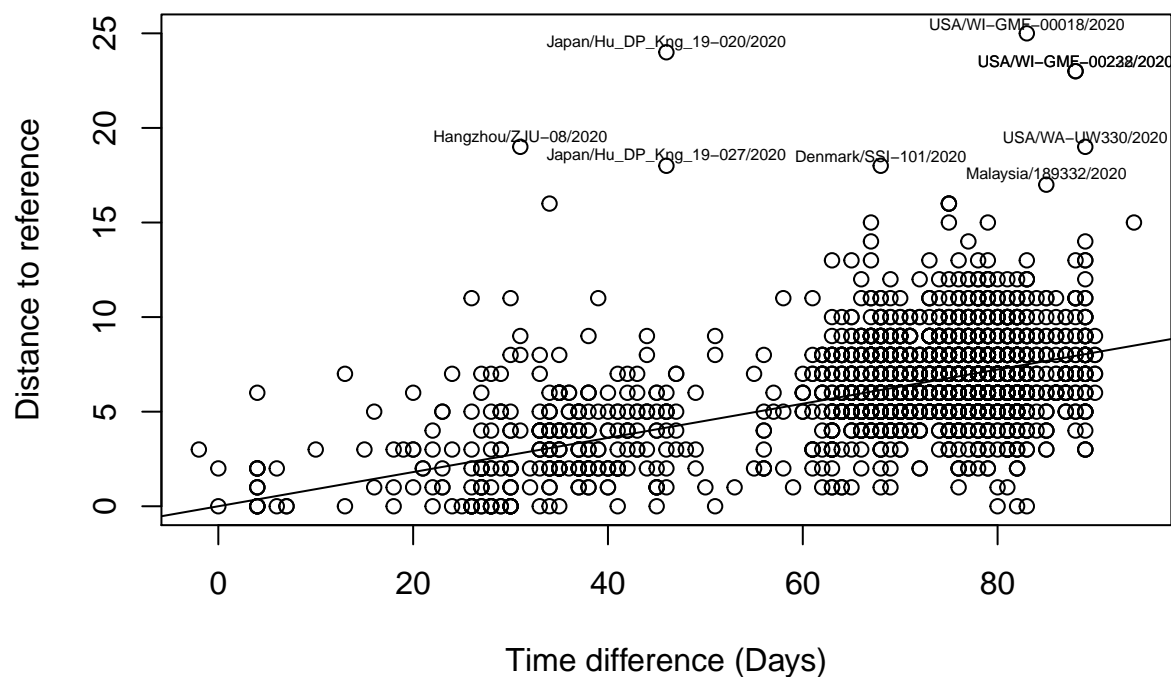
To explore regions and number of polymorphic sites

```
## [1] 29903
```
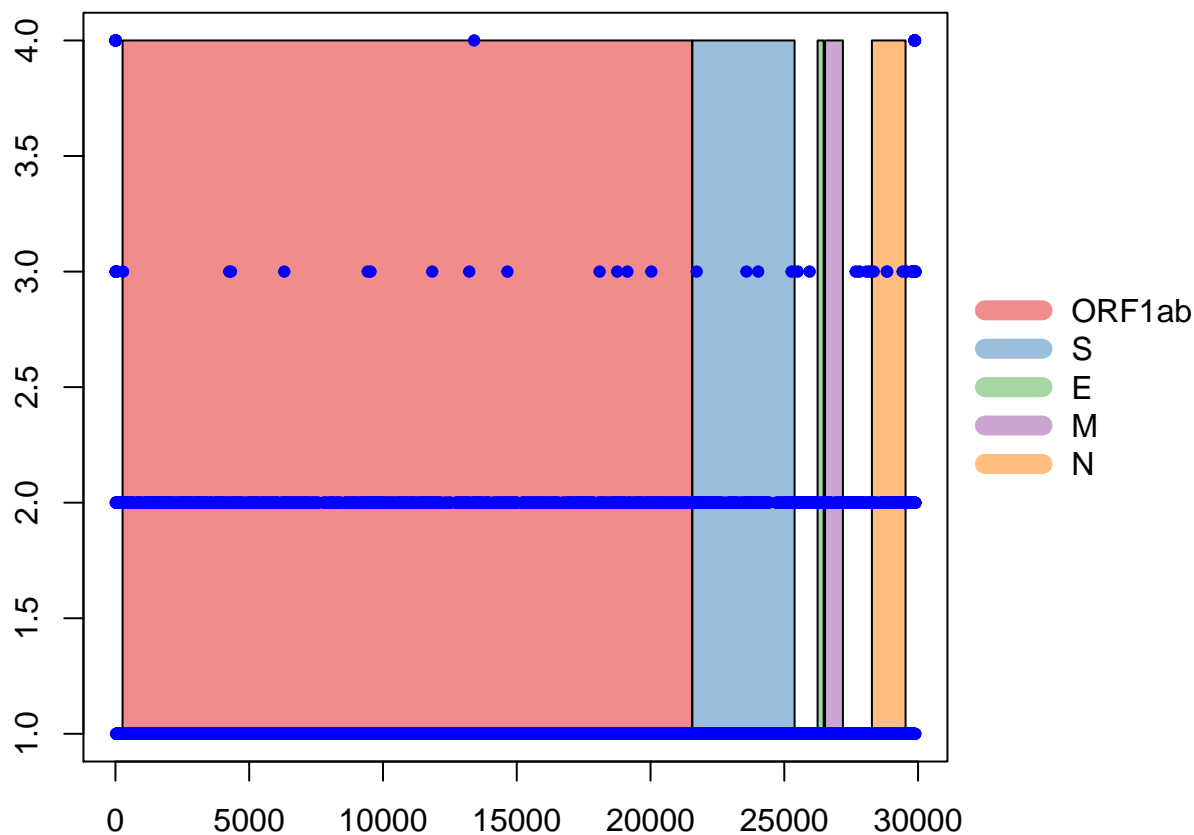
Of the total 29903 sites in the reference sequence, 28280 are monomorphic sites, 1529 are binary, and only 94 have 3 or 4 types. Largest entropy is observed is in the first 285 bases and the last 26447 bases.

Reference:

Rambaut, Andrew, Tommy T. Lam, Luiz Max Carvalho, and Oliver G. Pybus. "Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)." Virus evolution 2, no. 1 (2016).