

# ViReport v0.0.1

Niema Moshiri

2020-03-06

## 1 Input Dataset

The analysis was conducted on a dataset containing 214 sequences. The average sequence length was 29806.644859813085, with a standard deviation of 375.74831883734686. The earliest sample date was 2013-07-24, the median sample date was 2020-01-25, and the most recent sample date was 2020-03-03.

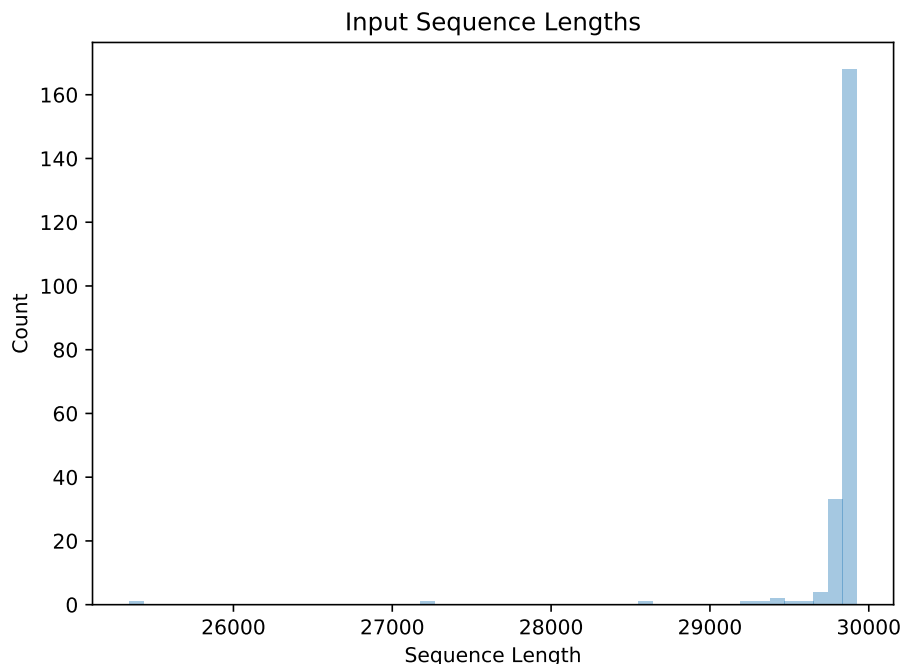


Figure 1: Distribution of input sequence lengths

## 2 Preprocessed Dataset

The input dataset was preprocessed such that sequences were given safe names: non-letters/digits in sequence IDs were converted to underscores. After preprocessing, the dataset contained 214 sequences. The average sequence length was 29806.644859813085, with a standard deviation of 375.74831883734686. The earliest sample date was 2013-07-24, the median sample date was 2020-01-25, and the most recent sample date was 2020-03-03.

## 3 Multiple Sequence Alignment

Multiple sequence alignment was performed using MAFFT (Katoh & Standley, 2013) in automatic mode. There were 30372 positions (16889 invariant) and 190 unique sequences in the multiple sequence alignment. Pairwise distances were computed from the multiple sequence alignment using the tn93 tool of HIV-TRACE (Pond et al., 2018). The average pairwise sequence distance was 0.0002624832123973732, with a standard deviation of 0.0002347743243505385.

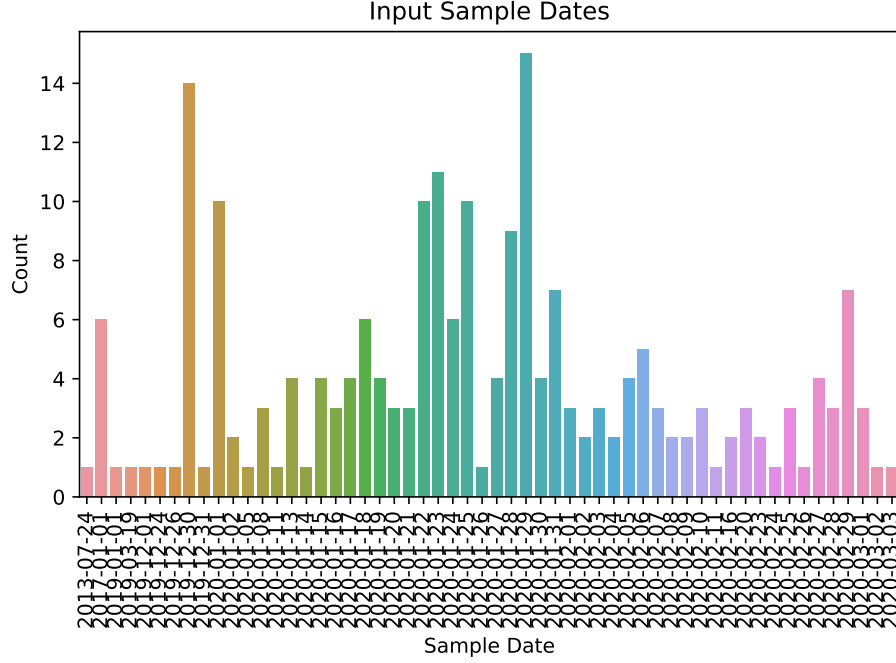


Figure 2: Distribution of input sample dates

## 4 Phylogenetic Inference

A maximum-likelihood phylogeny was inferred using IQ-TREE (Nguyen et al., 2015) in ModelFinder Plus mode (Kalyaanamoorthy et al., 2017). The inferred phylogeny was MinVar-rooted using FastRoot (Mai et al., 2017). Pairwise distances were computed from the phylogeny using TreeSwift (Moshiri, 2020). The maximum pairwise phylogenetic distance (i.e., tree diameter) was 0.0029801924999999993, and the average pairwise phylogenetic distance was 0.00037968161967545644, with a standard deviation of 0.00032592330778371745.

## 5 Phylogenetic Dating

The rooted phylogeny was dated using treedater (Volz & Frost, 2017). The height of the dated tree was 0.3045444379597261 days, so given that the most recent sample was collected on 2020-03-03, the estimated time of the most recent common ancestor (tMRCA) was 2020-03-02.

## 6 Transmission Clustering

Transmission clustering was performed using TreeN93 (Moshiri, 2018) using pairwise phylogenetic distances. The total number of singletons (i.e., non-clustered individuals) was 131, and the total number of clusters (excluding singletons) was 24. The average cluster size (excluding singletons) was 3.0416666666666665, with a standard deviation of 1.1718634258687695, and the maximum and minimum cluster sizes were 5 and 2, respectively.

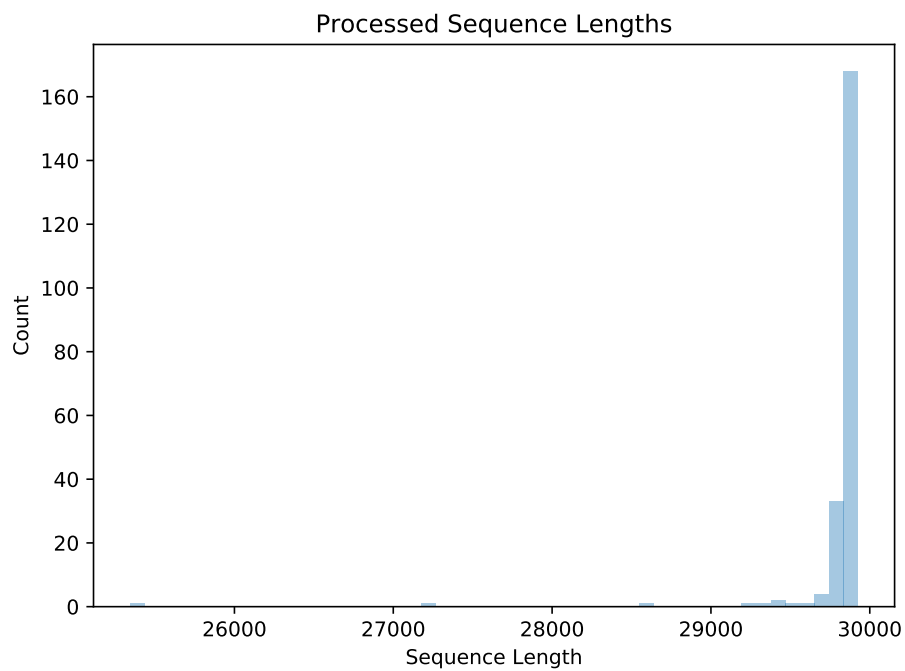


Figure 3: Distribution of preprocessed sequence lengths

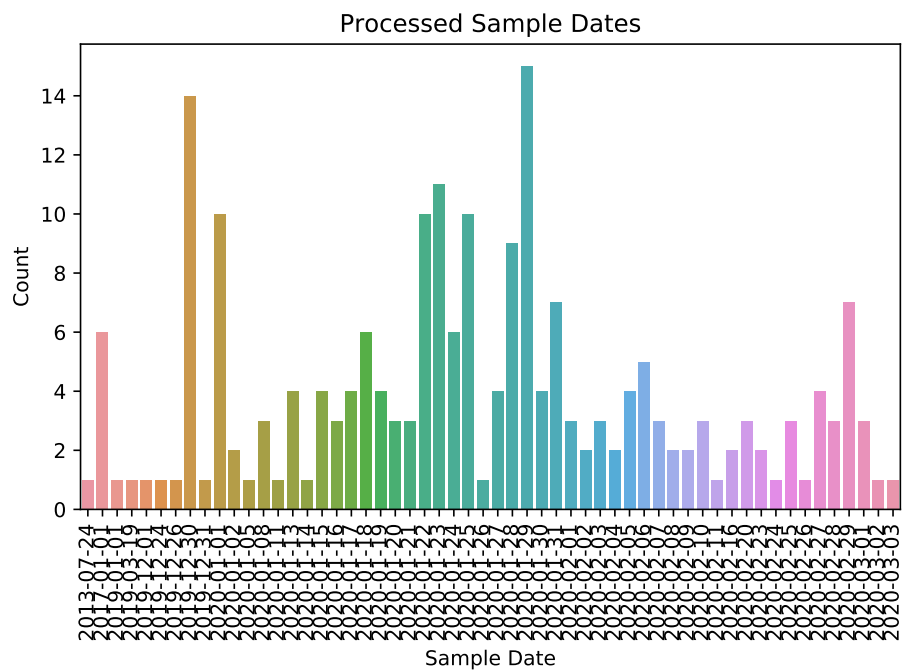


Figure 4: Distribution of preprocessed sample dates

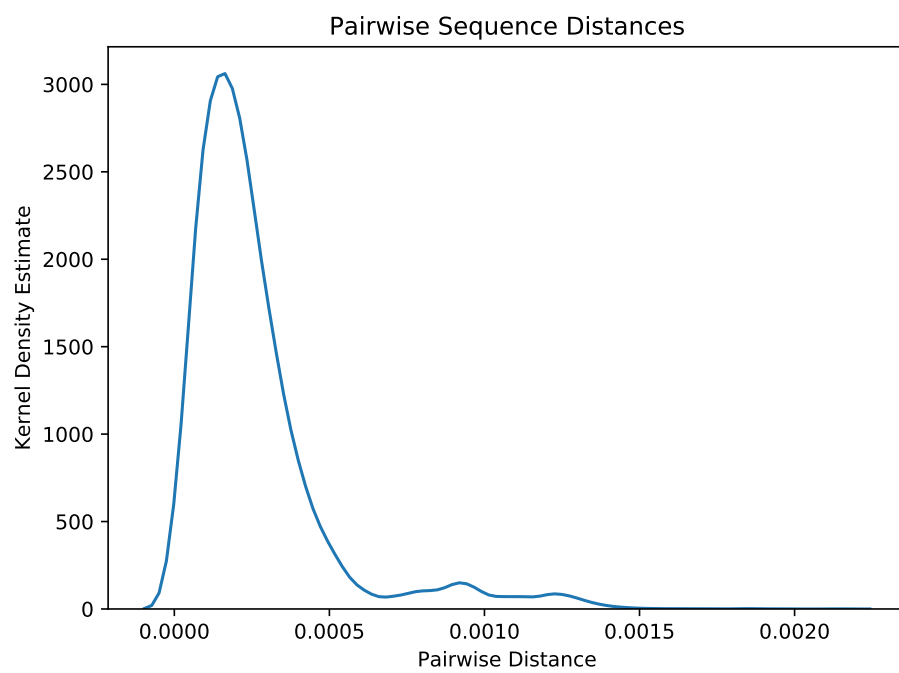


Figure 5: Distribution of pairwise sequence distances



Figure 6: Rooted phylogenetic tree in unit of expected per-site mutations

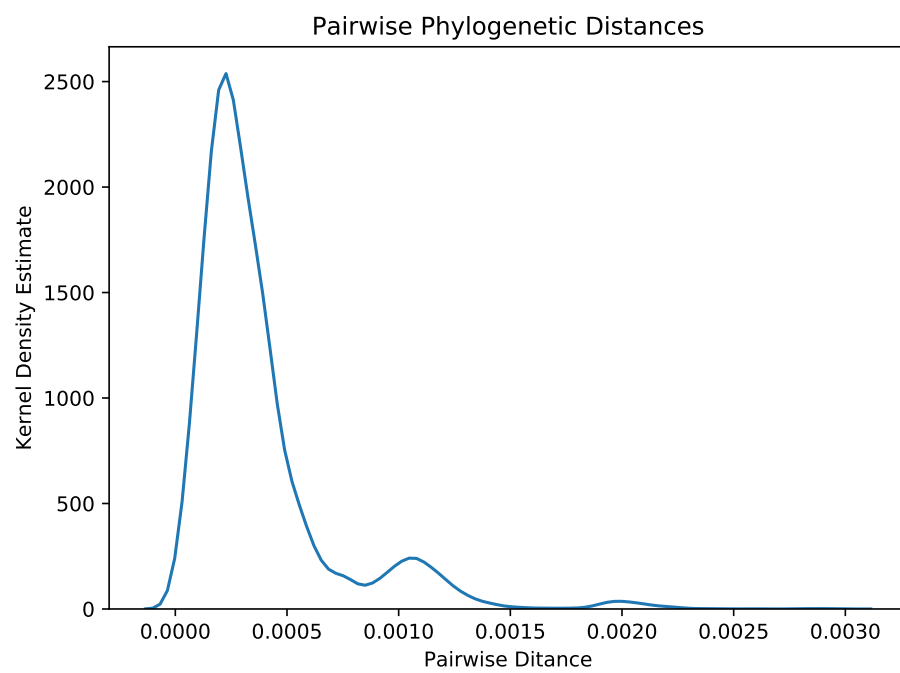


Figure 7: Distribution of pairwise phylogenetic distances

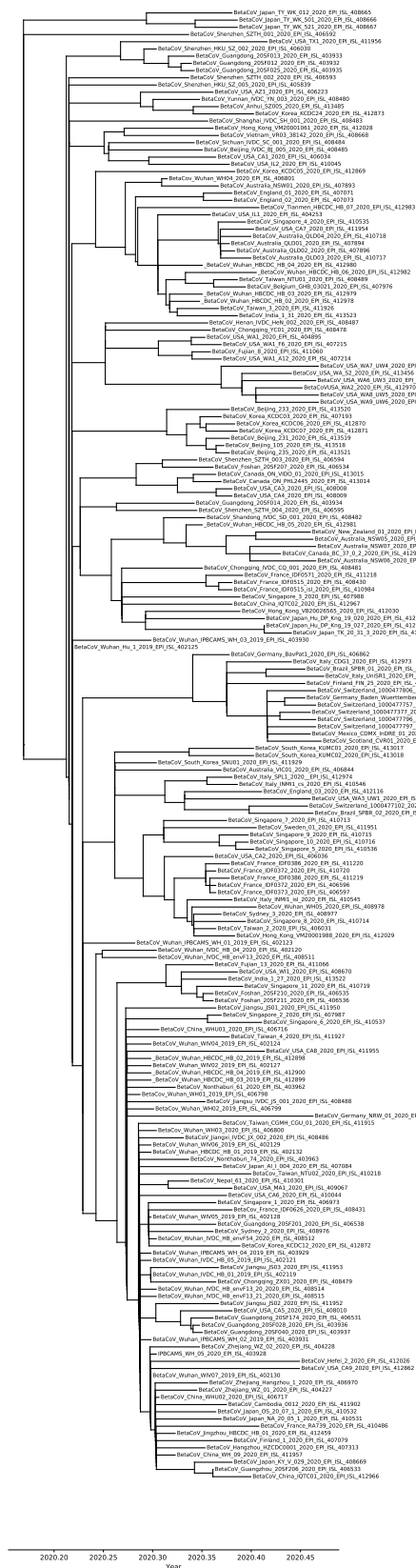


Figure 8: Dated phylogenetic tree in unit of years

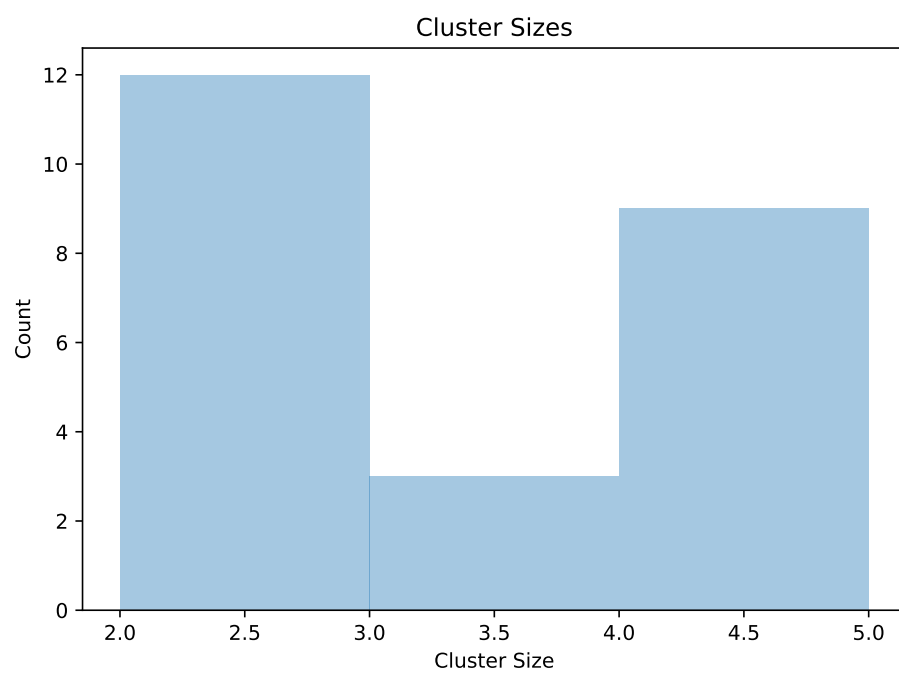


Figure 9: Distribution of cluster sizes (excluding singletons)