ViReport v0.0.1

Niema Moshiri

2020-03-10

1 Input Dataset

The analysis was conducted on a dataset containing 363 sequences. The average sequence length was 29801.774, with a standard deviation of 347.964. The earliest sample date was 2013-07-24, the median sample date was 2020-02-01, and the most recent sample date was 2020-03-05.

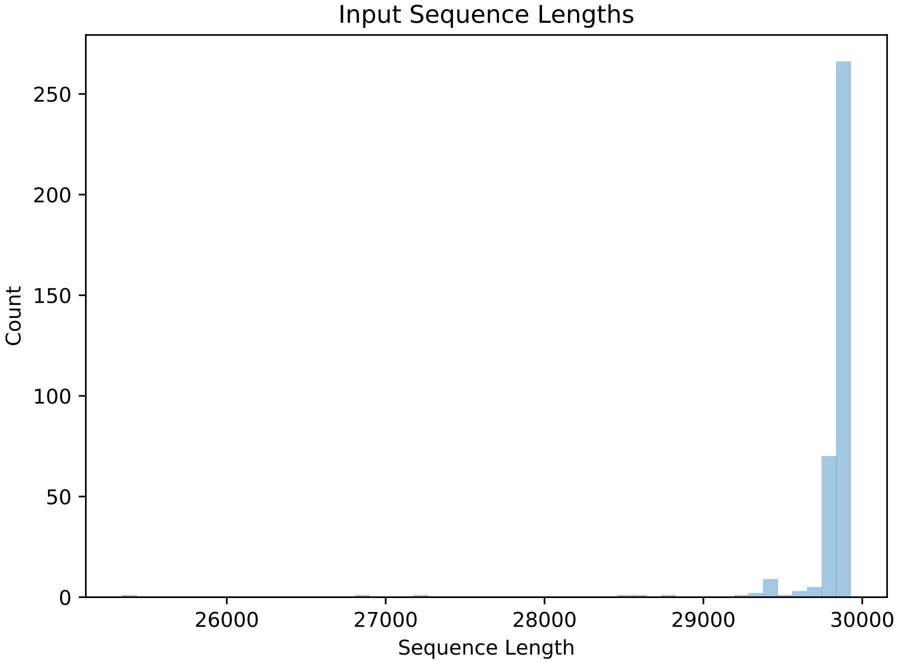

Figure 1: Distribution of input sequence lengths

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 363 sequences. The average sequence length was 29801.774, with a standard deviation of 347.964. The earliest sample date was 2013-07-24, the median sample date was 2020-02-01, and the most recent sample date was 2020-03-05.

3 Multiple Sequence Alignment

Multiple sequence alignment was performed using MAFFT (Katoh & Standley, 2013) in automatic mode. There were 30208 positions (11950 invariant) and 325 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 0.000254, with a standard deviation of 0.000197.

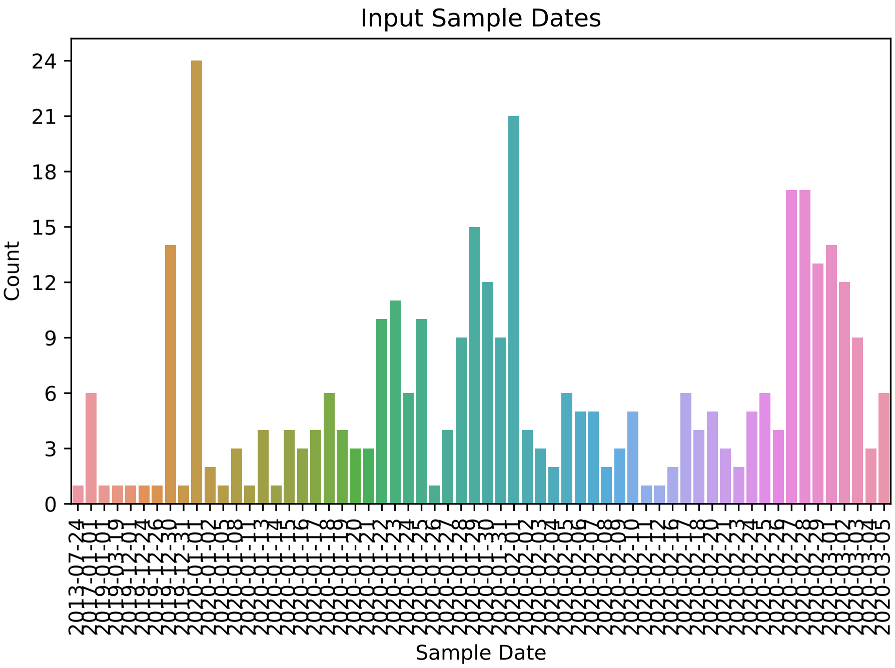

Figure 2: Distribution of input sample dates

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 0.0032, and the average pairwise phylogenetic distance was 0.000419, with a standard deviation of 0.000315.

5 Phylogenetic Dating

The rooted phylogeny was dated using treedater (Volz & Frost, 2017). The height of the dated tree was 109.636 days, so given that the most recent sample was collected on 2020-03-05, the estimated time of the most recent common ancestor (tMRCA) was 2019-11-16.

6 Ancestral Sequence Reconstruction

Ancestral sequence reconstruction was performed using TreeTime (Sagulenko et al., 2018).

7 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 57, and the total number of clusters (excluding singletons) was 17. The average cluster size (excluding singletons) was 17.412, with a standard deviation of 46.956, and the maximum and minimum cluster sizes were 204 and 2, respectively.

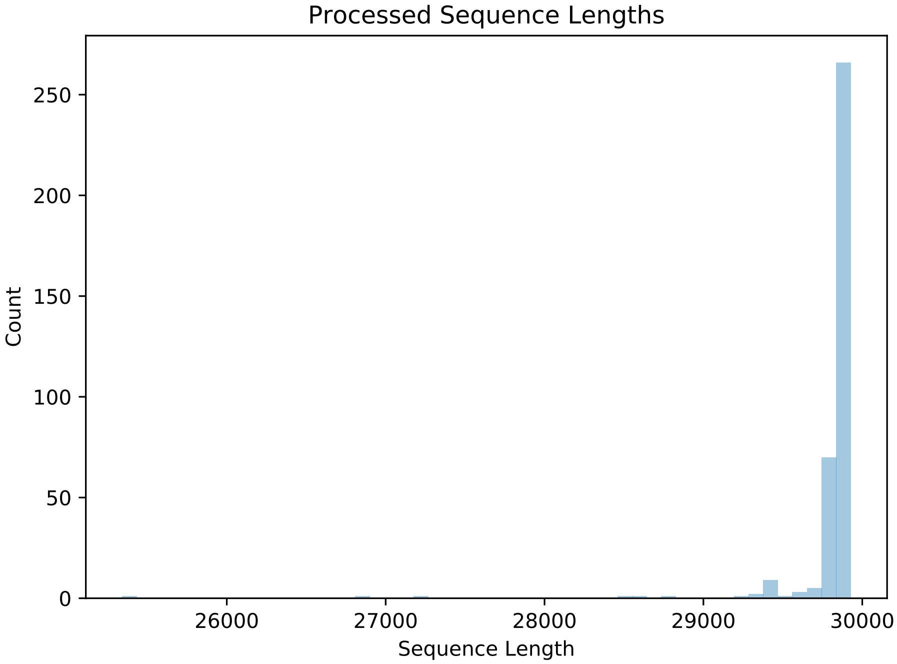

Figure 3: Distribution of preprocessed sequence lengths

8 Citations

- Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. (2017). ”ModelFinder: Fast model selection for accurate phylogenetic estimates”. Nature Methods. 14, 587-589.

- Katoh K., Standley D.M. (2013). ”MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability”. Molecular Biology and Evolution. 30(4), 772-780.

- Mai U., Sayyari E., Mirarab S. (2017). ”Minimum Variance Rooting of Phylogenetic Trees and Implications for Species Tree Reconstruction”. PLoS ONE. 12(8), e0182238.

- Moshiri N. (2018). ”TreeN93: a non-parametric distance-based method for inferring viral transmission clusters”. bioRxiv.

- Moshiri N. (2020). ”TreeSwift: a massively scalable Python tree package”. SoftwareX. In press.

- Moshiri N. (2020). ”ViReport” (https://github.com/niemasd/ViReport).

- Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. (2015). ”IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies”. Molecular Biology and Evolution. 32(1), 268-274.

- Pond S.L.K., Weaver S., Leigh Brown A.J., Wertheim J.O. (2018). ”HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens”. Molecular Biology and Evolution. 35(7), 1812-1819.

- Sagulenko P., Puller V., Neher R.A. (2018). ”TreeTime: Maximum-likelihood phylodynamic analysis”. Virus Evolution. 4(1), vex042.

- Volz E.M., Frost S.D.W. (2017). ”Scalable relaxed clock phylogenetic dating”. Virus Evolution. 3(2), vex025.

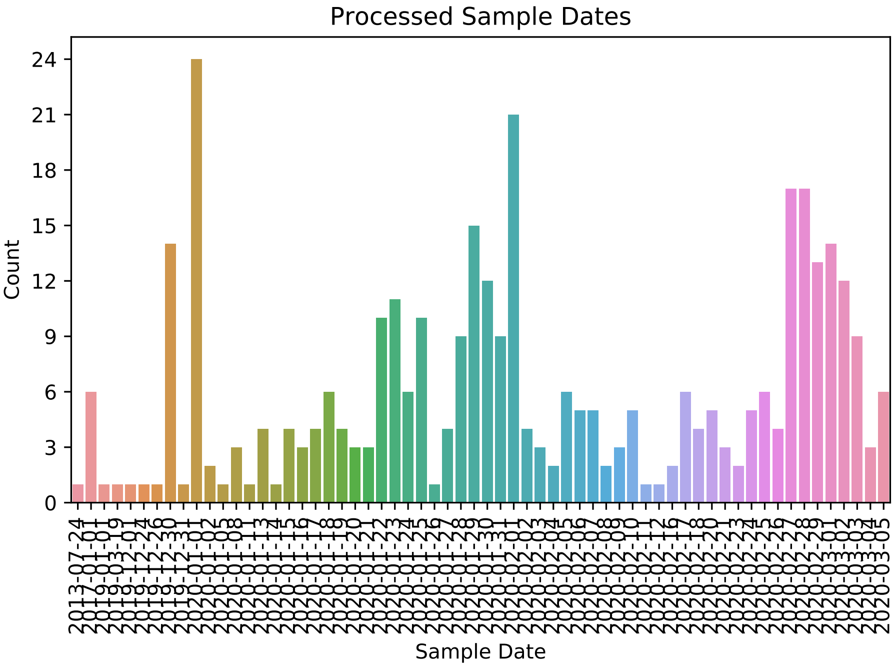

Figure 4: Distribution of preprocessed sample dates

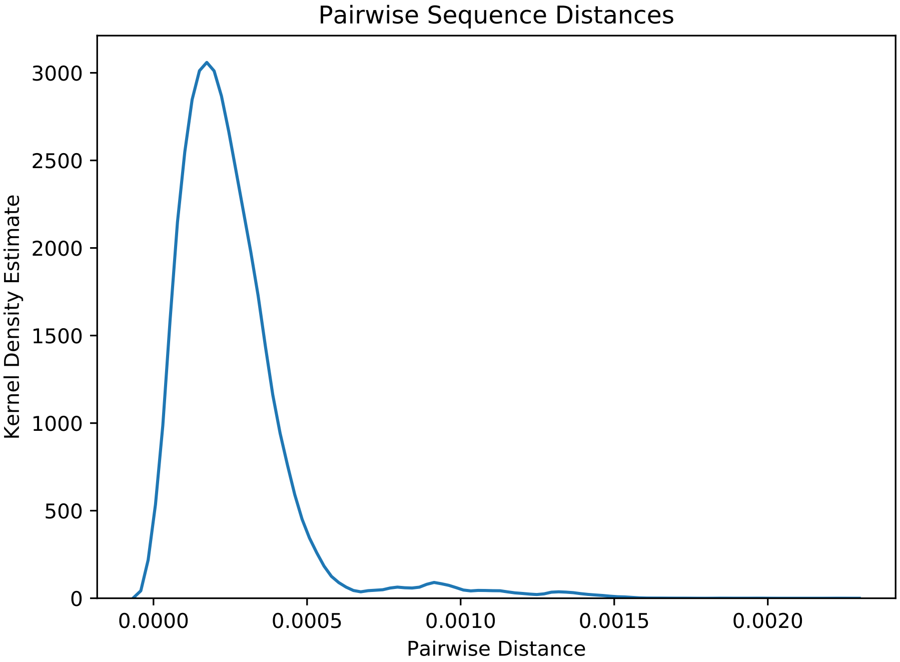

Figure 5: Distribution of pairwise sequence distances

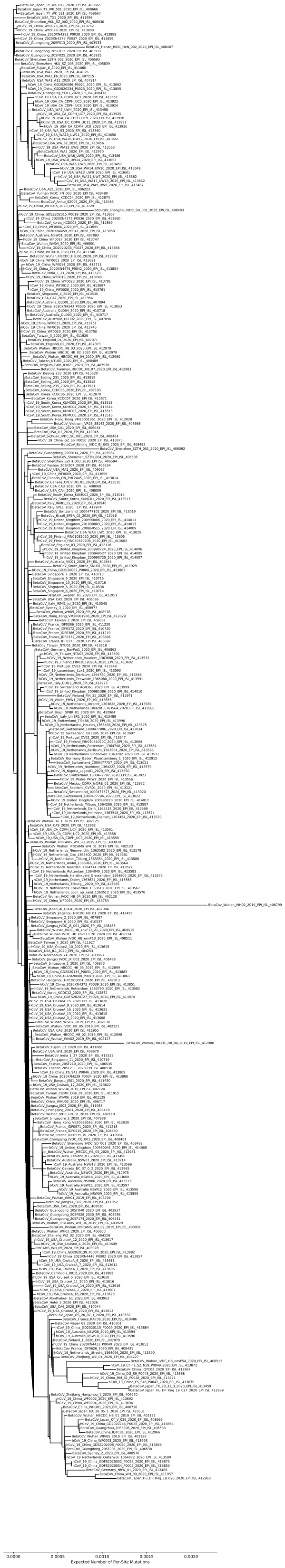

Figure 6: Rooted phylogenetic tree in unit of expected per-site mutations

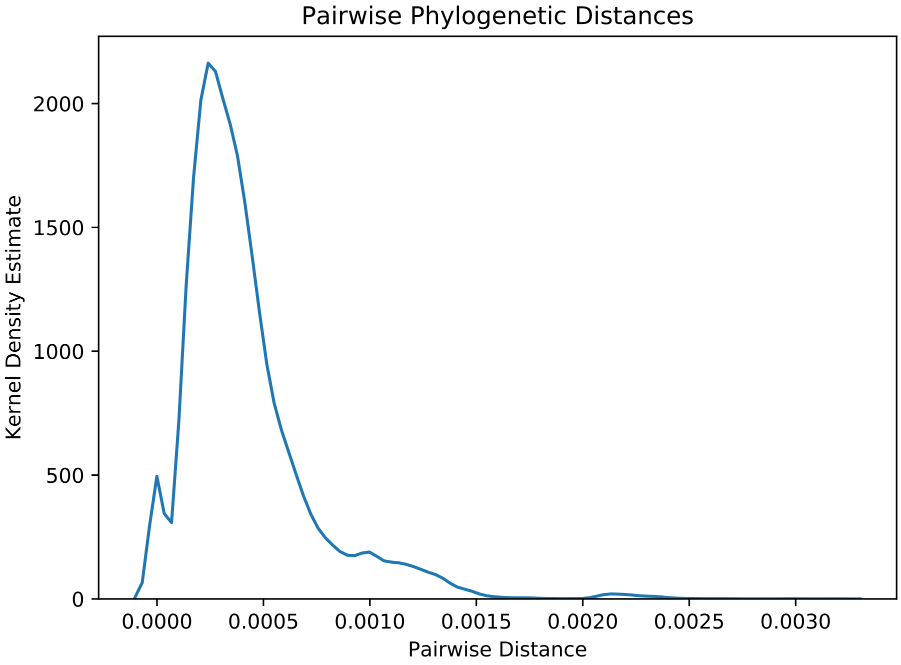

Figure 7: Distribution of pairwise phylogenetic distances

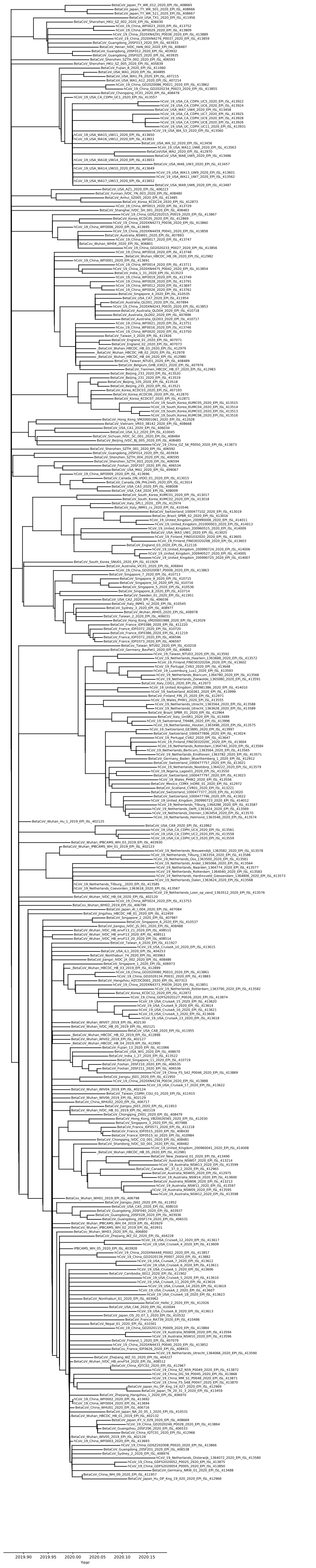

Figure 8: Dated phylogenetic tree in unit of years

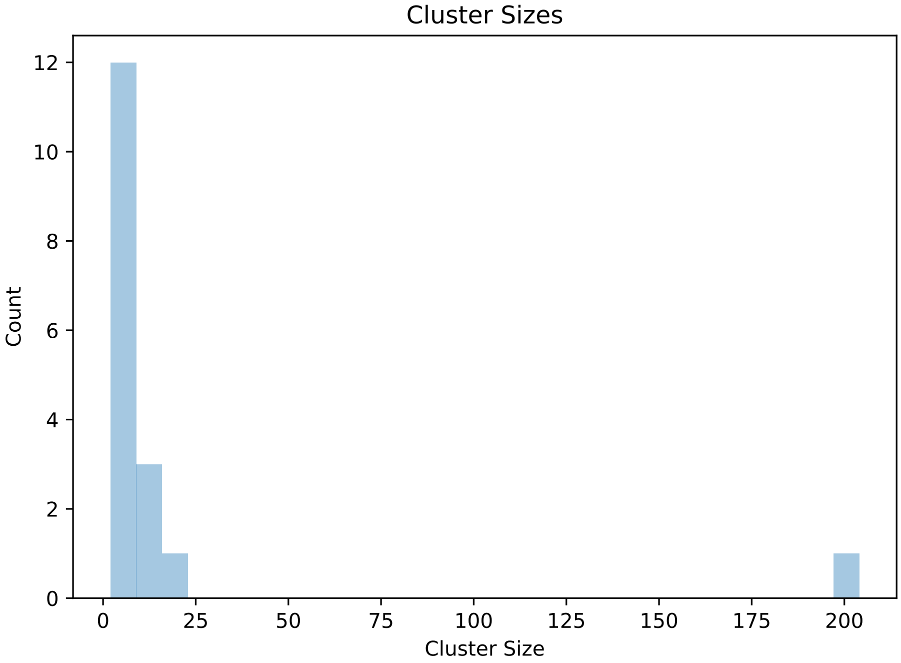

Figure 9: Distribution of cluster sizes (excluding singletons)