ViReport v0.0.1

Niema Moshiri

2020-03-23

1 Input Dataset

The analysis was conducted on a dataset containing 1328 sequences. The average sequence length was 29819.0429, with a standard deviation of 209.491. The earliest sample date was 2019-12-24, the median sample date was 2020-03-02, and the most recent sample date was 2020-03-20.

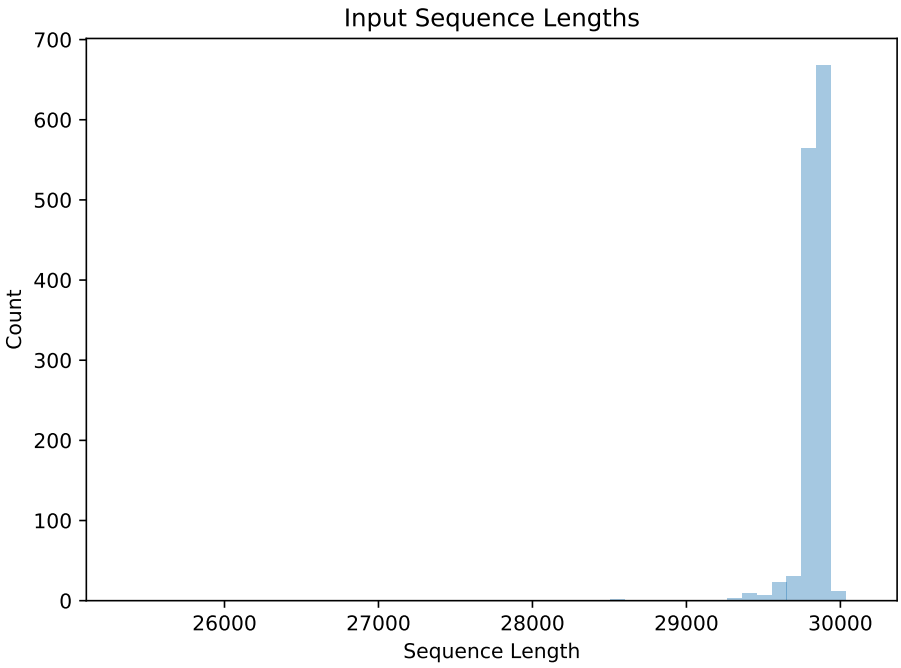


Figure 1: Distribution of input sequence lengths

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 1328 sequences. The average sequence length was 29819.0429, with a standard deviation of 209.491. The earliest sample date was 2019-12-24, the median sample date was 2020-03-02, and the most recent sample date was 2020-03-20.

3 Multiple Sequence Alignment

Multiple sequence alignment was performed using Minimap2 (Li, 2018). Each input sequence was aligned to the reference sequence (MT072688), and the multiple sequence alignment was constructed based on positions in the reference. There were 29811 positions (4614 invariant) and 1079 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 0.000284, with a standard deviation of 0.000131.

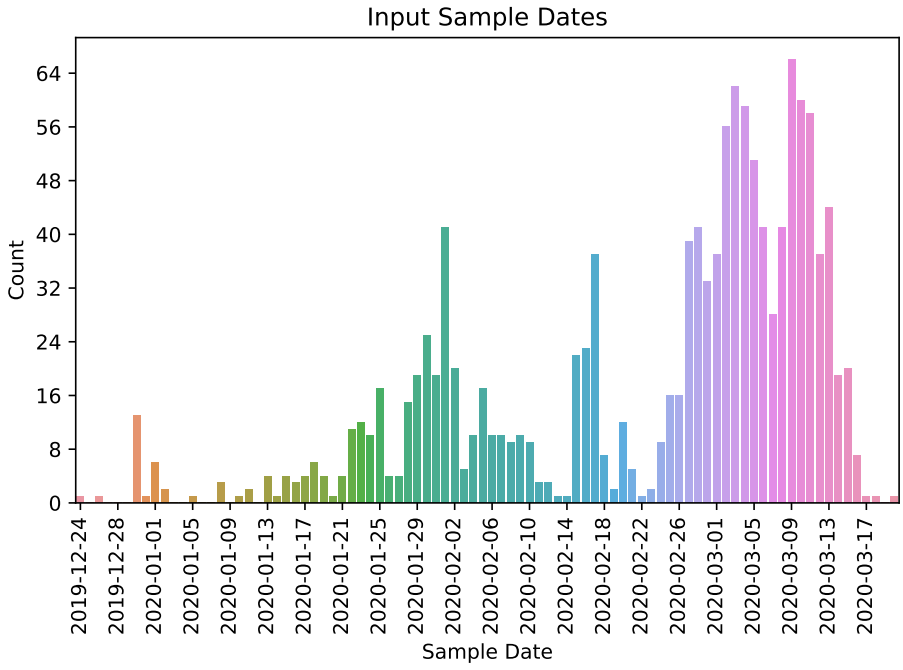


Figure 2: Distribution of input sample dates

Across the positions of the multiple sequence alignment that had non-zero Shannon entropy, the minimum Shannon entropy was 0.00907, the maximum Shannon entropy was 0.998, and the average Shannon entropy was 0.0467, with a standard deviation of 0.0641.

4 Phylogenetic Inference

A maximum-likelihood phylogeny was inferred under the General Time-Reversible (GTR) model (Tavare, 1986) using FastTree 2 (Price et al., 2010) using a Gamma20-based likelihood. 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 0.00123, and the average pairwise phylogenetic distance was 0.000366, with a standard deviation of 0.000178.

5 Phylogenetic Dating

The rooted phylogeny was dated using treedater (Volz & Frost, 2017). The height of the dated tree was 106.648 days, so given that the most recent sample was collected on 2020-03-20, the estimated time of the most recent common ancestor (tMRCA) was 2019-12-04.

6 Ancestral Sequence Reconstruction

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

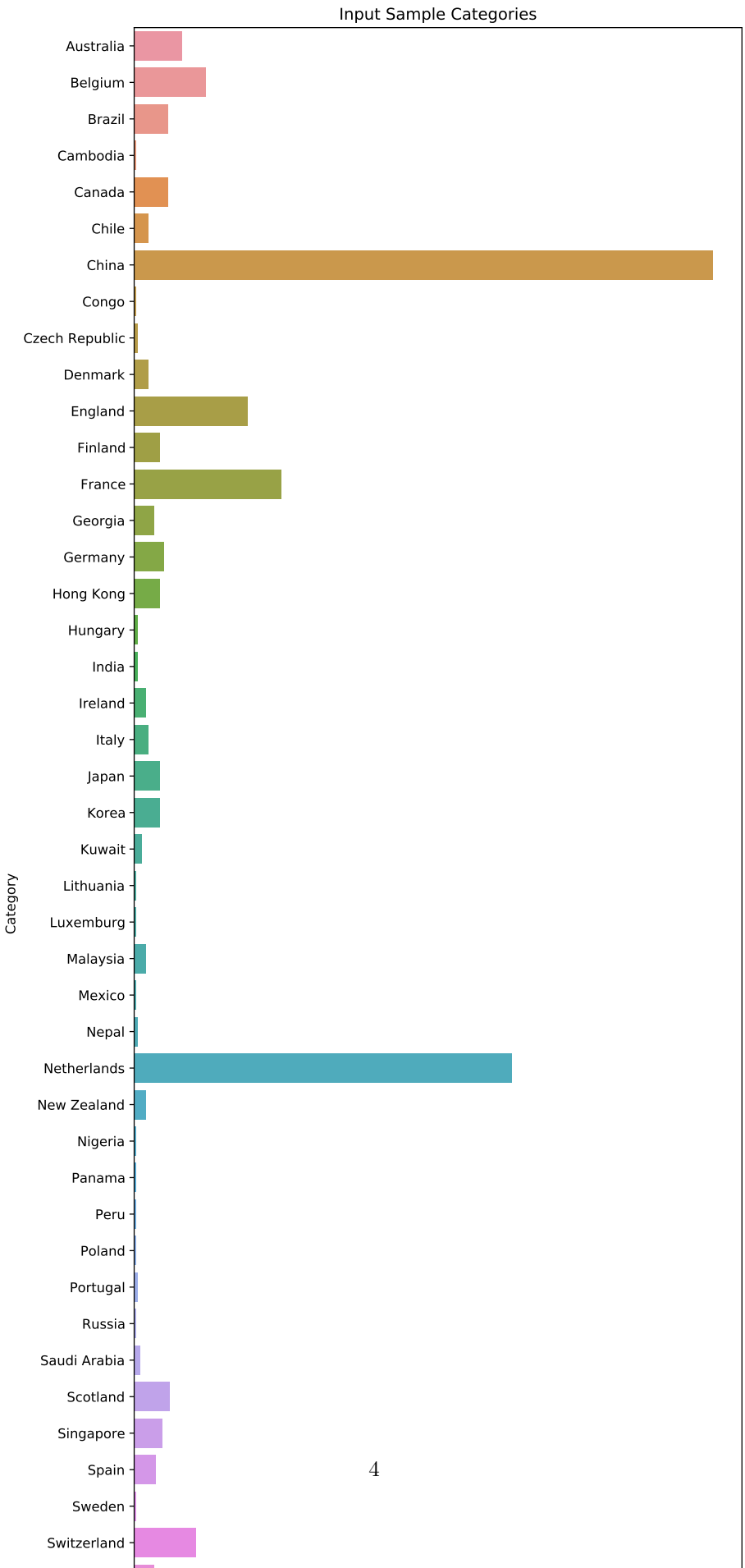
7 Transmission Clustering

Transmission clustering was performed using TreeN93 (Moshiri, 2018) using pairwise phylogenetic distances. The total number of singletons (i.e., non-clustered individuals) was 157, and the total number of clusters

(excluding singletons) was 29. The average cluster size (excluding singletons) was 39.448, with a standard deviation of 124.366, and the maximum and minimum cluster sizes were 648 and 2, respectively.

8 Citations

- Le S.Q., Gascuel O. (2008). "An Improved General Amino Acid Replacement Matrix". Molecular Biology and Evolution. 25(7), 1307-1320.
- Li H. (2018). "Minimap2: pairwise alignment for nucleotide sequences". Bioinformatics. 34(18), 3094-3100.
- 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).
- 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.
- Price M.N., Dehal P.S., Arkin A.P. (2010). "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments". PLoS ONE. 5(3), e9490.
- Sagulenko P., Puller V., Neher R.A. (2018). "TreeTime: Maximum-likelihood phylodynamic analysis". Virus Evolution. 4(1), vex042.
- Tavare S. (1986). ""Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". Lectures on Mathematics in the Life Sciences. 17, 57-86.
- Volz E.M., Frost S.D.W. (2017). "Scalable relaxed clock phylogenetic dating". Virus Evolution. 3(2), vex025.


Input Sample Categories
Category
Australia
Belgium
Brazil
Cambodia
Canada
Chile
China
Congo
Czech Republic
Denmark
England
Finland
France
Georgia
Germany
Hong Kong
Hungary
India
Ireland
Italy
Japan
Korea
Kuwait
Lithuania
Luxemburg
Malaysia
Mexico
Nepal
Netherlands
New Zealand
Nigeria
Panama
Peru
Poland
Portugal
Russia
Saudi Arabia
Scotland
Singapore
Spain
Sweden
Switzerland

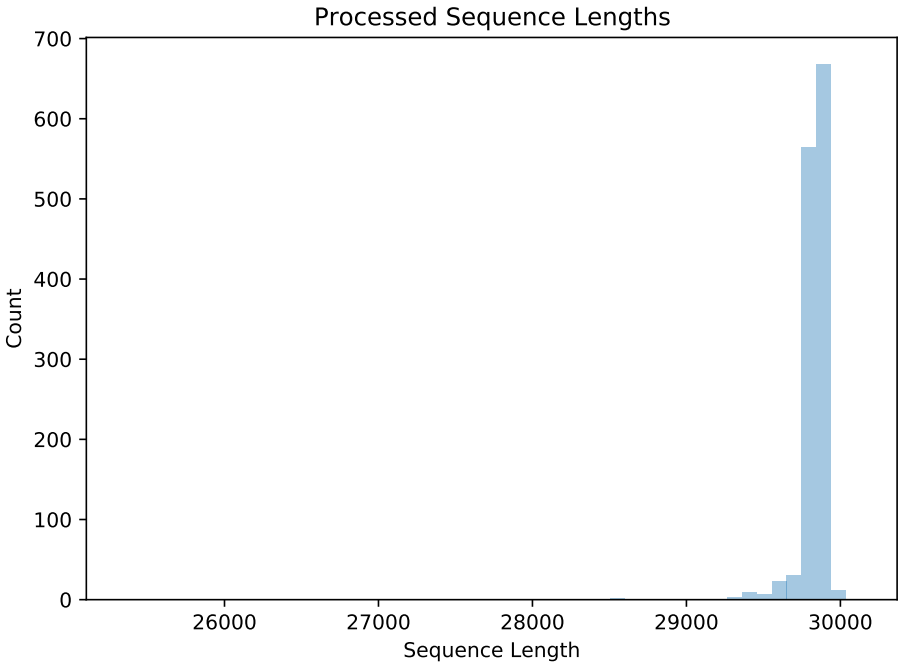


Figure 4: Distribution of preprocessed sequence lengths

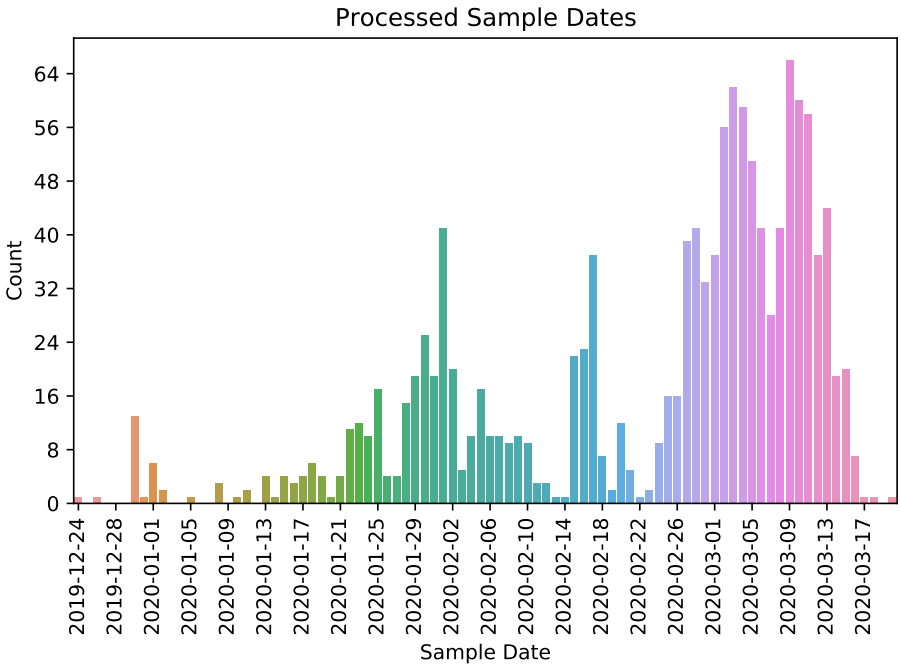
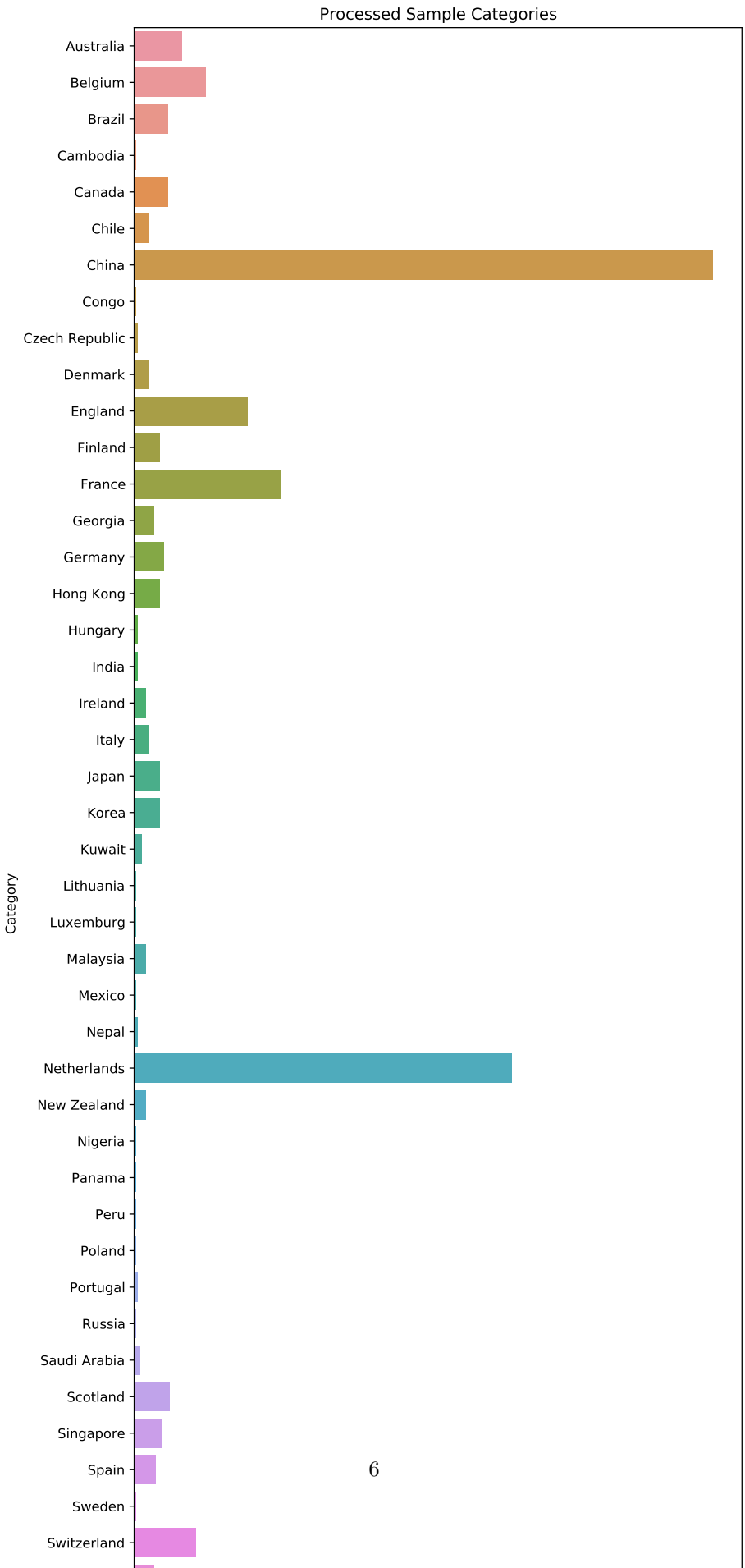


Figure 5: Distribution of preprocessed sample dates


Processed Sample Categories
Category
Australia
Belgium
Brazil
Cambodia
Canada
Chile
China
Congo
Czech Republic
Denmark
England
Finland
France
Georgia
Germany
Hong Kong
Hungary
India
Ireland
Italy
Japan
Korea
Kuwait
Lithuania
Luxemburg
Malaysia
Mexico
Nepal
Netherlands
New Zealand
Nigeria
Panama
Peru
Poland
Portugal
Russia
Saudi Arabia
Scotland
Singapore
Spain
Sweden
Switzerland

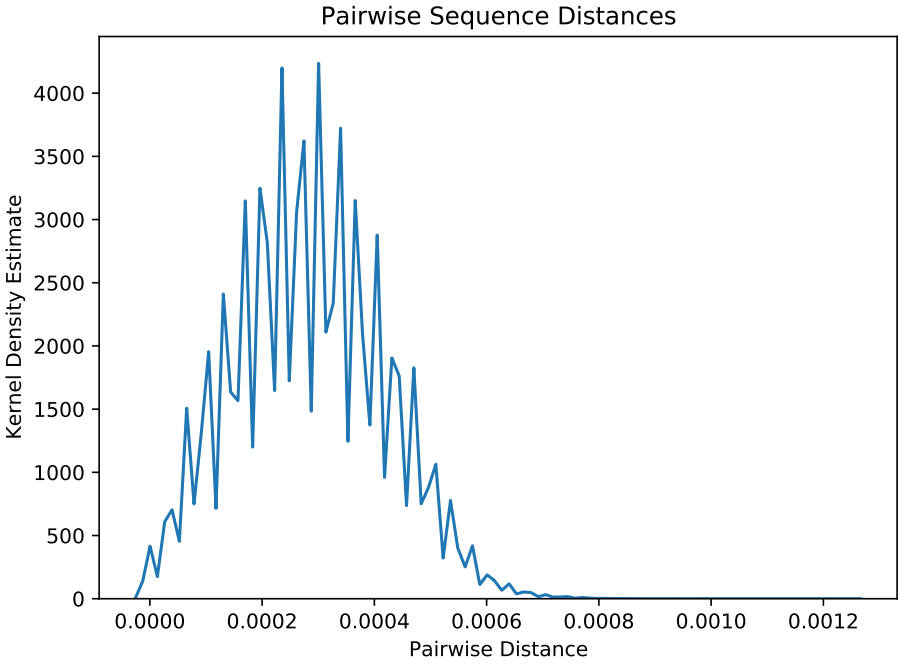


Figure 7: Distribution of pairwise sequence distances

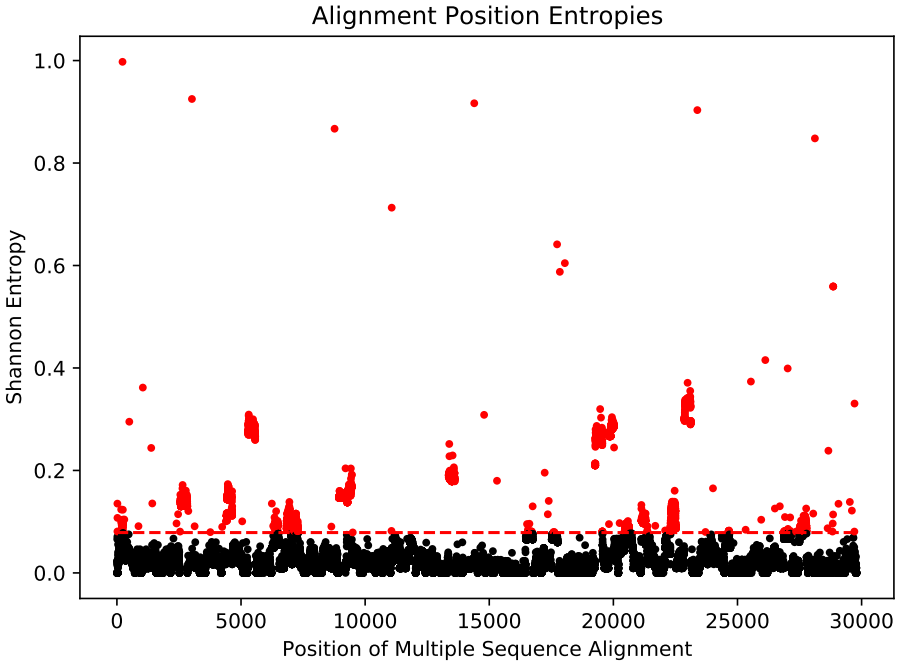


Figure 8: Shannon entropy across the positions of the multiple sequence alignment. A significance threshold was computed using Tukey's Rule: 1.5x the interquartile range added to the third quartile, which was 0.0787. The significance threshold is shown as a red dashed line, and significant points are shown in red.

Figure 9: Rooted phylogenetic tree in unit of expected per-site mutations

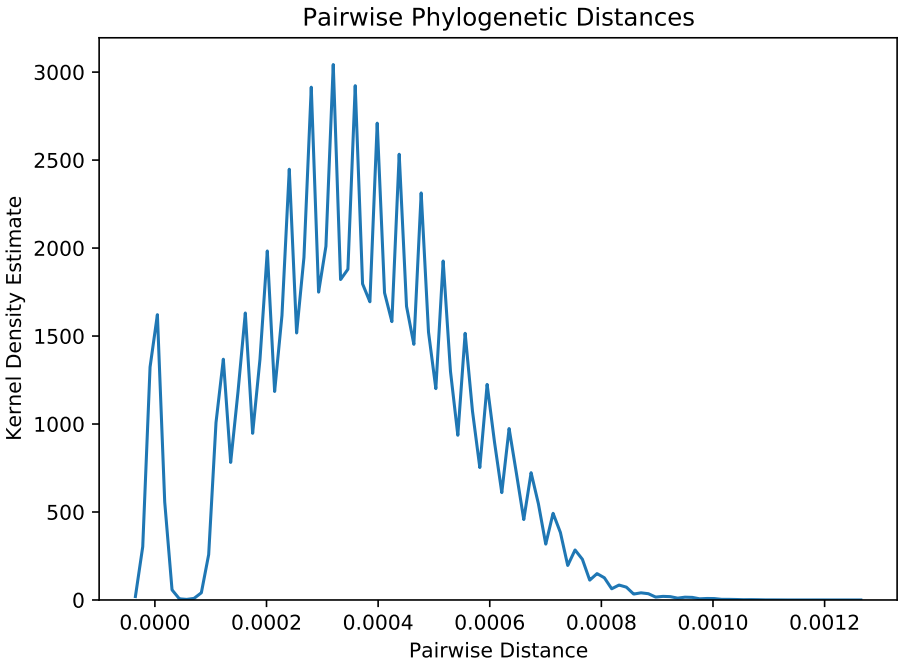


Figure 10: Distribution of pairwise phylogenetic distances

Figure 11: Dated phylogenetic tree in unit of years

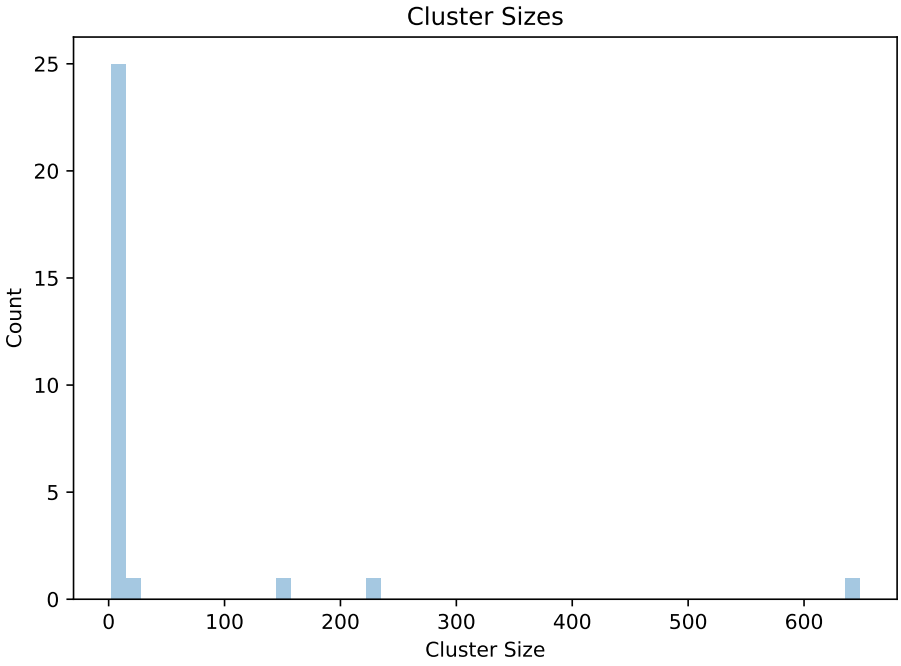


Figure 12: Distribution of cluster sizes (excluding singletons)