BV-BRC SARS-CoV-2 Genome Reference Tree: Methods

Last updated: 05/04/2021

Questions, comments: czmasek@jcvi.org

1. Download SARS-CoV-2 complete genome nucleotide sequences from ViPR (https://www.viprbrc.org)
2. Remove sequences with more than 0.01% non-ATCG characters or shorter than 29,400 nucleotides
3. Analyze cleaned up sequences with pangoLEARN
4. Selecte 3 or fewer sequences per PANGO lineage
5. Phylogenetic inference:
    a. Multiple sequence alignment: MAFFT v7.453 (auto option)
    b. Remove multiple sequence alignment columns with more than 50% gap characters
    c. Tree inference: Minimal evolution tree calculated by FastME v 2.1.4 based on ML pairwise distances calculated by TREE-PUZZLE v5.2 using GTR model
6. Used various custom scripts to "decorate" tree with mutation, lineage, host, country, region, and date information (mutation and lineage of concern designation is from https://beta.bv-brc.org/view/VariantLineage).