VERSION 1
DEC 02, 2022

WORKS FOR ME 1

Methodology to Define the Origin of SARS-CoV-2 V.1

COMMENTS 0

DOI
dx.doi.org/10.17504/protocols.io.x54v9yqz4g3e/v1

David Maison[1,2,3], Sean Cleveland[4,5], Vivek R. Nerurkar[1,2,3]

[1]Department of Tropical Medicine, Medical Microbiology, and Pharmacology;
[2]Pacific Center for Emerging Infectious Diseases Research;
[3]John A. Burns School of Medicine, University of Hawai'i - System, Honolulu, Hawai'i 96813;
[4]Hawai'i Data Science Institute;
[5]Information Technology Services - Cyberinfrastructure, University of Hawai'i - System, Honolulu, Hawai'i 96813

David Maison

ABSTRACT

Using the CDC-classified SARS-CoV-2 VOC (B.1.1.7, B.1.351, B.1.427, B.1.429, and P.1), identified in Hawai'i as an example, we demonstrate a method to define the origin of SARS-CoV-2 lineages and VOC. This method works using either open-source or licensed software with either a personal computer or a supercomputer.

DOI

dx.doi.org/10.17504/protocols.io.x54v9yqz4g3e/v1

EXTERNAL LINK

https://doi.org/10.1371/journal.pone.0278287

PROTOCOL CITATION

David Maison, Sean Cleveland, Vivek R. Nerurkar 2022. Methodology to Define the Origin of SARS-CoV-2. **protocols.io** https://dx.doi.org/10.17504/protocols.io.x54v9yqz4g3e/v1

FUNDERS ACKNOWLEDGEMENT

Pacific Center for Emerging Infectious Diseases Research, COBRE
Grant ID: P30GM114737-05

INBRE, National Institute of General Medical Sciences, NIH
Grant ID: P20GM103466-20S1

NSF grant on the University of Hawai'i MANA High Performance Computing Cluster
Grant ID: #1920304

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Maison DP, Cleveland SB, et al. Genomic Analysis of SARS-CoV-2 Variants of Concern Circulating in Hawai'i to Facilita Public-Health Policies. Res Sq. Published online June 9, 2021. doi: 10.21203/rs.3.rs-378702/v3

KEYWORDS

SARS-CoV-2, phylogenetics, public health, COVID-19

IMAGE ATTRIBUTION

Maison DP, Cleveland SB, et al. Genomic Analysis of SARS-CoV-2 Variants of Concern Circulating in Hawai'i to Facilitate Public-Health Policies. Res Sq. Published online June 9, 2021. doi: 10.21203/rs.3.rs-378702/v3. Created with BioRender.com.

CREATED

Aug 07, 2022

LAST MODIFIED

Dec 02, 2022

PROTOCOL INTEGER ID

68322

1 The lineage-defining sequences of SARS-CoV-2 Lineage A and Lineage B act as the most ancestral roots. Lineage A (EPI_ISL_406801) is from GISAID, and Lineage B (MN908947) is from GenBank.

1.1 Register for a free GISAID account (https://gisaid.org/register/) to obtain EPI_ISL_406801

1.2 GenBank MN908947 (https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3)

2 Identify lineages of interest in an area:
filter GISAID by location (e.g.: North America/USA/Hawai'i) and download all sequences. For VOC with >10,000 sequences, GISAID sequences were downloaded in batches due to GISAID maximum download size. Similarly, all geographically similar sequences reported in GenBank were downloaded using the search term SARS-CoV-2 and state abbreviation (e.g., "SARS-CoV-2 HI") and the sequence length filter (20,000 - 40,000).

3 Combine the GISAID and GenBank sequences into one .fasta file using AliView, Geneious Prime, or a text editor, and assign lineages using Pangolin Lineage Assigner (pangolin.cog-uk.io).

3.1 Geneious Prime (http://www.geneious.com)
AliView (https://ormbunkar.se/aliview/)

4 Determine prevalence of Each Lineage:
Download the results to Microsoft Excel, use advanced filter to copy unique records of lineages to a new column (ex: column M), then use COUNTIF (e.g., =COUNTIF($B$2:$B$1432,M2)) to determine prevalence of each lineage. Alternatively, upload the results to Google Sheets and use the =UNIQUE command (e.g., =UNIQUE(B2:B1432) followed by the above COUNTIF command.

5 Filter GISAID and GenBank by the lineage of interest (e.g., B.1.429) and download all sequences.

6 Combine lineage of interest (B.1.429) GenBank sequences, GISAID (B.1.429) sequences, and EPI_ISL_406801 into one fasta file.

7 Align sequences using multiple alignment using fast Fourier transform (MAFFT) program or server with MN908947 as a reference and do not remove any uninformative sequences and all parameters set as "same as input.

Citation: David Maison, Sean Cleveland, Vivek R. Nerurkar Methodology to Define the Origin of SARS-CoV-2
https://dx.doi.org/10.17504/protocols.io.x54v9yqz4g3e/v1

7.1 MAFFT (https://mafft.cbrc.jp/alignment/server/add_fragments.html?frommanualnov6)

8 Remove the newly added MN908947 sequence that MAFFT places at the beginning of the alignment using AliView, Geneious Prime, or a text editor. If not, the sRNA toolbox will remove the MN908947 sequence during the duplicate removal step, and Lineage B will not serve as an ancestral root in the phylogenetic tree.

9 Import Multiple Sequence Alignment (MSA) file into Geneious Prime or AliView, search for the orf1a 5' start of the entire alignment (5'-atggagagccttgtccctggtttca-3') and remove the 5' untranslated region (UTR) by deleting the upstream region (~265 bp) from the MSA. Next, search for ORF10 3' end (5'-tgtagttaactttaatctcacatag-3') and remove the entire 3' UTR by deleting the downstream region (~229 bp) from the MSA.

10 Create a duplicate file for the MN908947 sequence and remove the 5' UTR and 3' UTR from MN908947 as described above.

11 Using MAFFT, align the trimmed MSA with the trimmed MN908947 as a reference and delete sequences with uncalled nucleotides 'n'. Set the "remove uninformative sequences" parameter in the MAFFT at >0%.

12 Using sRNAtoolbox program or server, load the updated alignment to remove duplicate sequences and merge identifications (also referred to as sequence accession numbers) of duplicates. This merger will create "appendages" in the phylogenetic tree where the sRNA toolbox will line up identical sequences together with equal signs (=).

12.1 sRNAtoolbox ((https://arn.ugr.es/srnatoolbox/helper/removedup/)

13 Import the final alignment into Geneious Prime and create an approximately maximum-likelihood phylogenetic tree using the FastTree program. Alternatively, FastTree can run as standalone software, and FastTreeMP is appropriate when multiple CPU cores/threads are available.

14 Root the tree with Lineage A (EPI_ISL_406801), which should then be the most recent common ancestor (MRCA) to Lineage B (MN908947) if performing phylogenetics on a Lineage B subgroup. Identify the MRCA of each sequence of interest.

Citation: David Maison, Sean Cleveland, Vivek R. Nerurkar Methodology to Define the Origin of SARS-CoV-2 https://dx.doi.org/10.17504/protocols.io.x54v9yqz4g3e/v1

Citation: David Maison, Sean Cleveland, Vivek R. Nerurkar Methodology to Define the Origin of SARS-CoV-2
https://dx.doi.org/10.17504/protocols.io.x54v9yqz4g3e/v1