Ebola virus phylogenetic analysis protocol

Nanopore | bioinformatics

Document:	ARTIC-IturiEBOV-phylogeneticsSOP-v1.0.0
Creation Date:	2019-12-12
Author:	Andrew Rambaut
Licence:	Creative Commons Attribution 4.0 International License

> **Overview:** An analysis protocol for an initial phylogenetic analysis of consensus genomes. Includes alignment, phylogeny estimation and visualization.

This document is part of the Ebola virus Nanopore sequencing protocol package:

http://artic.network/ebov/ ⬈

Related documents:

Ebola virus Nanopore sequencing protocol:

http://artic.network/ebov/ebov-seq-sop.html (/ebov/ebov-seq-sop.html)

Setting up the laptop computing environment using Conda:

http://artic.network/ebov/ebov-it-setup.html ⬈

Ebola virus Nanopore bioinformatics protocol:

http://artic.network/ebov/ebov-bioinformatics-sop.html ⬈

Preparation

Set up the computing environment as described in this document: ebov-it-setup (ebov-it-setup.html)

This protocol also assumes that the setup and installation of the bioinformatics protocol has been performed as described in this document: ebov-bioinformatics-sop (ebov-bioinformatics-sop.html) .

Installing software

Activate the ARTIC Conda environment:

```
source activate artic-ebov
```

Reference genomes

`genomes_462_2019-12-09.aln.fasta` - An alignment of 462 complete or largely-complete genomes from North Kivu and Ituri of DRC.

This data is provided by the INRB on their GitHub.

This is a FASTA format file which contains a multiple alignment of the 462 genomes.

Annotating new sequences with metadata

We recommend adding appropriate metadata to the headers of each of the new sequences at this stage. We suggest following the format of the reference alignment above:

```
>[virus]|[strain]|[accession]|[country]|[admin1]|[admin2]|[date of collection]
```

i.e.:

```
>EBOV|18FHV089|MK007329|COD|Nord-Kivu|Mabalako|2018-07-27
```

Note the date format at the end (ISO 8601) - this is the most unambiguous date format and is the standard in computational biology and data science. If the day of sampling is unknown (or the month) then these can be omitted:

```
>EBOV|18FHV089|MK007329|COD|Nord-Kivu|Mabalako|2018-07
```

or

```
>EBOV|18FHV089|MK007329|COD|Nord-Kivu|Mabalako|2018
```

Other fields can be added but the standard is to have the date of collection at the end.

Building a multiple alignment

Use MUSCLE ⤤ multiple alignment software to align the new genome consensus sequences to the existing reference genome alignment:

```
muscle -profile -in1 genomes_462_2019-12-09.aln.fasta -in2 new_genomes.fasta -
fastaout new_genomes.aligned.fasta
```

This methods keeps the existing alignment and pair-wise aligns the new sequence to it.

> **Note:** The `profile` option is much quicker than doing a full multiple alignment but could be problematic if the new genome is divergent from all the reference genomes. It may be worth doing a full re-alignment.

- Optional step – To re-align an existing alignment:

```
muscle -in new_genomes.aligned.fasta -out new_genomes.re-aligned.fasta -
refine
```

Inferring a phylogenetic tree

We will infer a phylogenetic tree using maximum likelihood (ML) with iqTree ⤤. This will use the default nucleotide model (HKY with gamma distributed site rate heterogeneity):

```
iqtree -m HKY+G -pre new_genomes.iqtree -s new_genomes.aligned.fasta
```

The `-pre new_genomes.iqtree` option defines what the output files will be called. The output goes into six files with various extensions, but the tree will be in : `new_genomes.iqtree.treefile` . The `-fast` option will produce a tree much faster but it will be more approximate.

By default an ML tree is arbitrarily rooted so to help with the interpretation of the tree, so use the Gotree ⤤ utility to re-root the tree so some of the earliest virus of the epidemic are at at the root:

```
gotree reroot outgroup -i new_genomes.iqtree.treefile 'EBOV_18FHV089_MK007329_
COD_Nord-Kivu_Mabalako_2018-07-27' > new_genomes.iqtree.rooted.tree
```

View the phylogenetic tree

We suggest using FigTree ⤤ for interactive tree viewing and interpretation.

The lastest version can be downloaded from here:

```
wget https://github.com/rambaut/figtree/releases/download/v1.4.4/FigTree_v1.4.
4.tgz
tar xfz FigTree_v1.4.4.tgz
```

To open the tree in FigTree:

```
java -jar FigTree_v1.4.4/lib/figtree.jar new_genomes.iqtree.rooted.tree
```