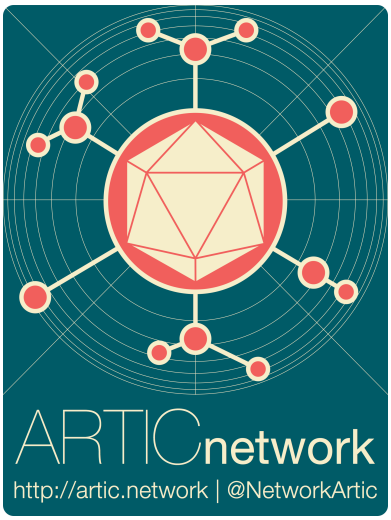

Ebola virus phylogenetic analysis protocol

Nanopore | bioinformatics

Document: ARTIC-EBOV-phylogeneticsSOP-v1.0.0
Creation Date: 2018-05-26
Author: Andrew Rambaut

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

Funded by the Wellcome Trust
Collaborators Award 206298/Z/17/Z --- ARTIC network (artic.network)

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

An alignment of 35 complete or nearly-complete genomes spaning 1976-2014 is available. This has representitives from many of the Middle Africa outbreaks in DRC, Gabon and Republic of Congo and 3 from the West African outbreak in Guinea in 2014. This set is intended to provide a framework in order to place new, previously uncharacterised, outbreak sequences.

35 EBOV genome alignment - FASTA file.

This is a FASTA format file which contains a multiple alignment of the 35 genomes.

Test data

To test these instructions you can find a synthetic genome in the ARTIC repository. This file, called `fake_ebola_genome.fasta` is an artificial genome sequence constructed to fall within the diversity of the 35 provided reference genomes to simulate the discovery of a new lineage of EBOV. This is for testing only and shouldn't be included in any analyses.

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 ebov-reference-genomes-35.fasta -in2 new_genomes.fasta -f
astaout aligned.afa
```

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 aligned.afa -out re-aligned.afa -refine
```

Inferring a phylogenetic tree

We will infer a phylogenetic tree using maximum likelihood (ML) with PhyML. This program uses the PHYLIP alignment format and we can use the Goalign utility to convert from FASTA format:

```
goalign reformat phylip -i aligned.afa > aligned.phy
```

Then build the tree. This will use the default nucleotide model (HKY with gamma distributed site rate heterogeneity):

```
phyml --input aligned.phy --datatype nt
```

The output goes into two files: `aligned.phy_phyml_stats.txt` provides all the estimated parameter values and other information, `aligned.phy_phyml_tree.txt` is the resulting tree in NEWICK format.

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 the 1970s viruses are at the root:

```
gotree reroot outgroup -i aligned.phy_phyml_tree.txt 'KC242791|Bonduni|DRC|1977-06' 'KC242801|deRoover|DRC|1976' 'KM655246|Yambuku-Ecran|DRC|1976' > rooted.tree
```

ETE3 can be used to open a window to view the resulting tree:

```
ete3 view -t rooted.tree
```