**Introduction to Pathogen Phylogenetic Analysis**

**Ebola virus genome data: a phylogenetic study**

**Background**

Ebola virus (EBOV), with species name *Zaire ebolavirus*, is a negative-sense single-stand RNA virus of the family *Filoviridae* that causes severe disease in humans. This virus was first described in humans in 1976, from an outbreak that lasted two months and infected 318 people, with a 88% case-fatality rate. Since its description, there have been 12 smaller outbreaks. It is believed that bats are the primary reservoir host. The recent Ebola virus epidemic in West Africa has been the most devastating by far. This epidemic started in at least December 2013 and lasted until 2016, with a total of 28,646 confirmed cases and 11,323 deaths (Holmes et al. 2016). Evolutionary studies revealed important insights about this outbreak. For example, phylogenetic methods were used to infer when and where the outbreak started, and how quickly it had been spreading. Importantly, such insight was possible as early as 2014 due to the recent surge of large-scale methods to generate real-time genomic data.

Throughout the workshop we will analyse two data sets of EBOV sequences. The first includes historical samples from 1976, and it will be useful to estimate the evolutionary rate and time of origin of the virus. The second data set consists in complete genomes sampled from 72 individuals in Sierra Leone from end of May to mid June 2014. These correspond to 70% of all confirmed cases in this period (Gire et al. 2014).

**Practical 1a: Sequence alignment in MEGA**

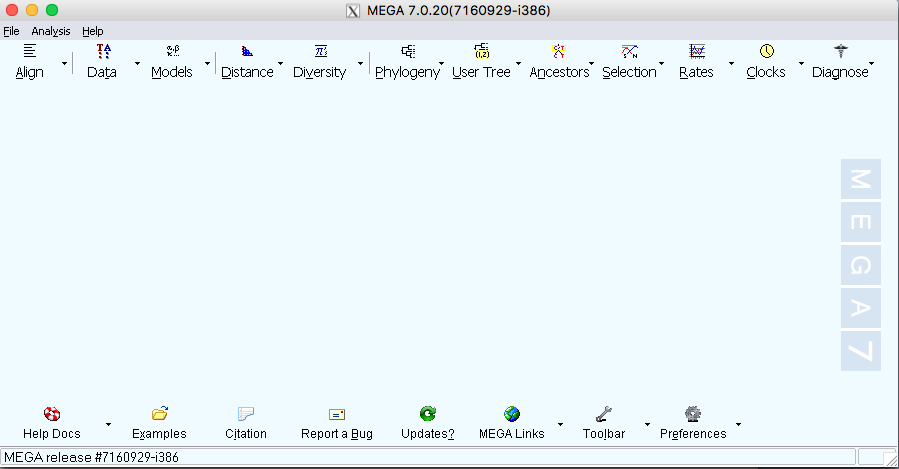
**Data set**

* Sequence data in fasta format of samples collected before the 2013-2016 epidemic: EBOV\_N2\_unaligned\_early\_samples.fasta

**Software**

* MEGA v7 (megasoftware.net)

Open MEGA v7 in your computer, you should see this window appear (Fig 1):



**Fig 1.** MEGA graphical interface

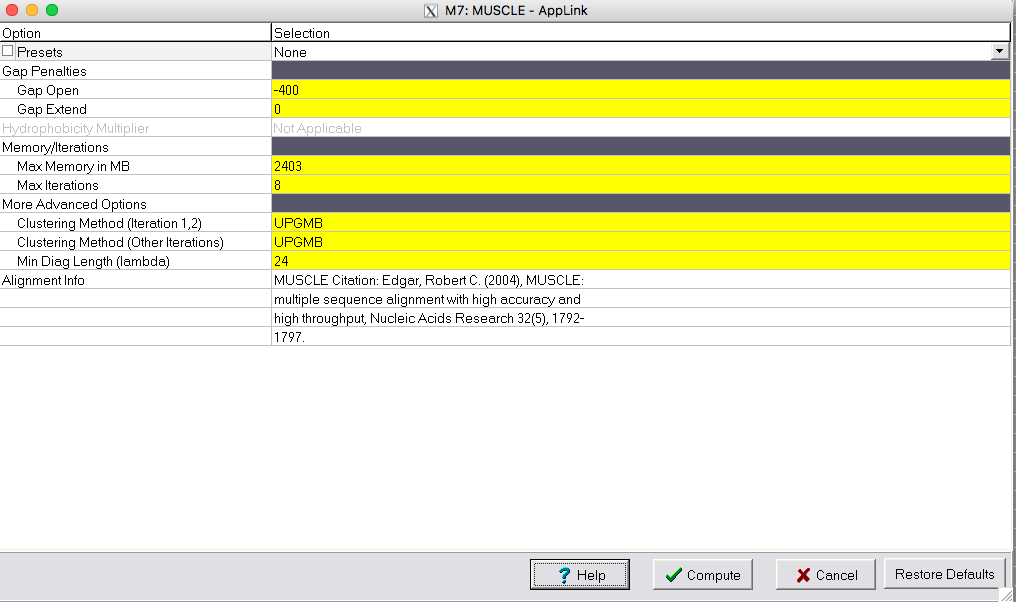
Click on *Data* and then on *Open A File/Session…* Find the file EBOV\_N2\_unaligned\_early\_samples.fasta. A window will come up, asking you whether you want to analyse the data or align it, select *Align*:



**Fig 2.** Ebola virus nucleotide sequence data in MEGA.

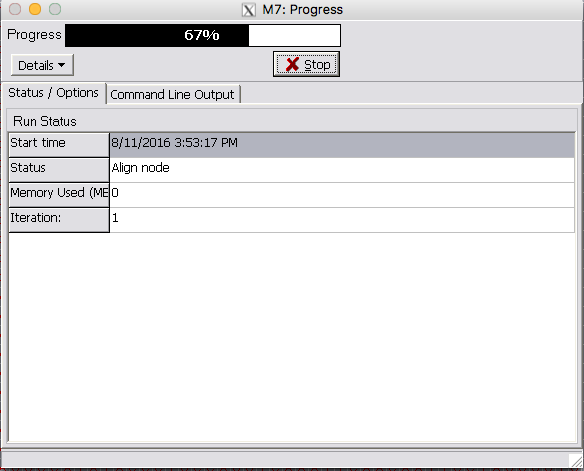
You should see the sequence data, as shown in Fig 2. Each row corresponds to a sequence, and the columns are the different sites. Note that individual sites do not appear aligned. That is, we could not consider them homologous. To align them, click on the symbol and select *Align DNA*. This will conduct a sequence alignment using MUSCLE. If a window pops up, telling you whether it should select all the sequences, click *OK*.

Next, a window with alignment options will appear (Fig 3).



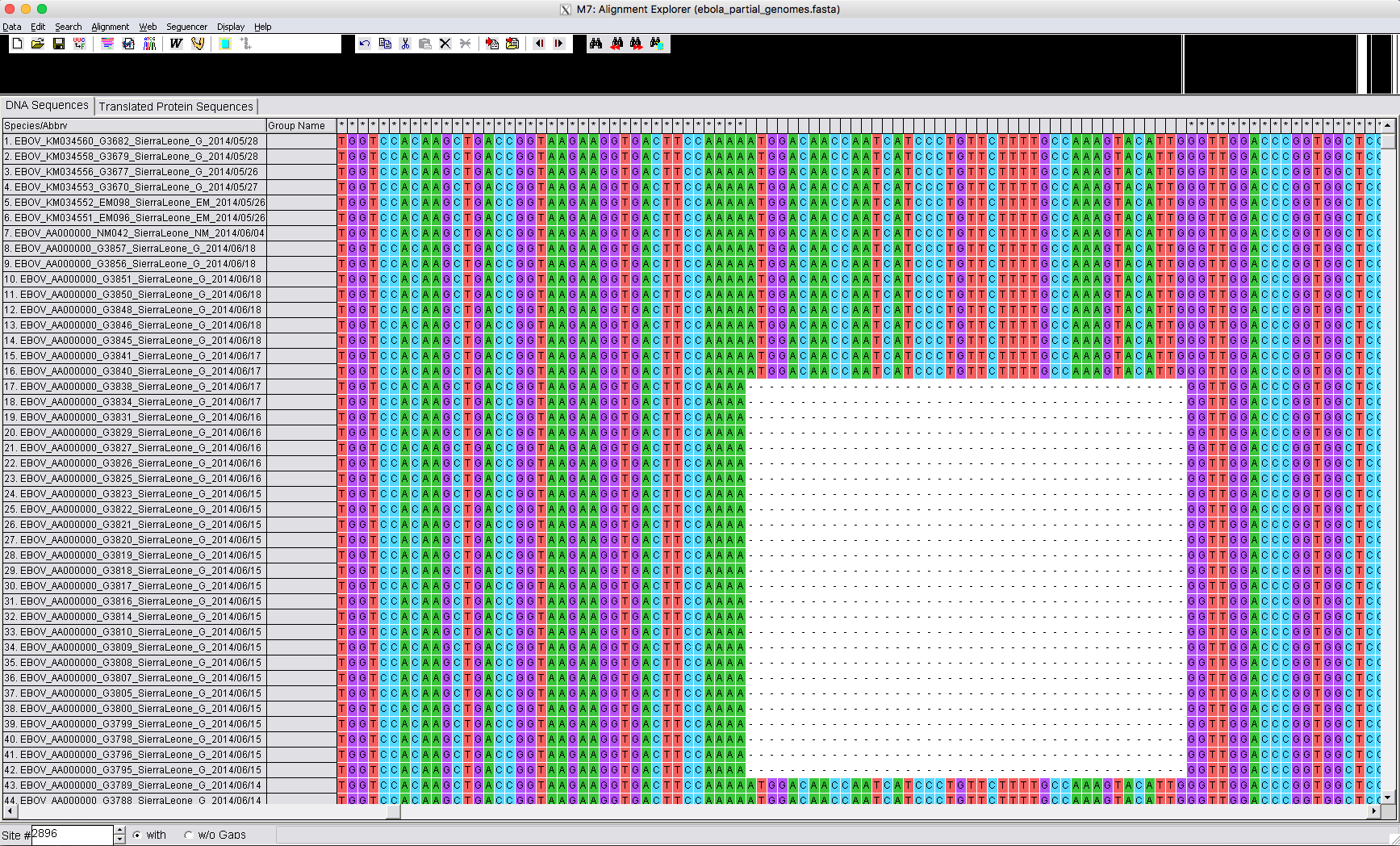
**Fig 3.** Alignment options for the MUSCLE algorithm.

Click *Compute*. The default options are fine for this analysis. In practice you can compare the results between different alignment options. For example, if the resulting alignment has too many gaps, you can increase the penalties for opening and extending gaps. You will see a window appear (Fig 4), which shows the progress of the alignment iterations.



**Fig 4.** Progress of sequence alignment in MEGA.

To see more details, click on *Command Line Output*. The analysis will run for a few minutes. Once it is complete, scroll horizontally to see the resulting alignment. Note that there are a few gaps, but the sequences now look aligned (Fig 5).



**Fig 5.** Sequence alignment in MEGA with gaps.

We want to save the alignment. To do this, click on *Data*, and on *Export* *Alignment*. Select FASTA format and name it: EBOV\_N2\_aligned\_early\_samples.fasta

Repeat this step to save the file in Phylip format, and name it:

EBOV\_N2\_aligned\_early\_samples.phy

Close MEGA, do not save the working session - we only need the alignment that we saved in the previous step.

Open the alignment saved in fasta and phylip format in a text editor and get some familiarity with how they differ. Some phylogenetics programs nowadays accept many common sequence alignment formats, but some programs, like PhyML, are very specific about the format they accept.

**Practical 1b: Model selection in MEGA**

**Data set**

* Sequence alignment in fasta format of samples collected before the 2013-2016 epidemic:

EBOV\_N2\_aligned\_early\_samples.fasta

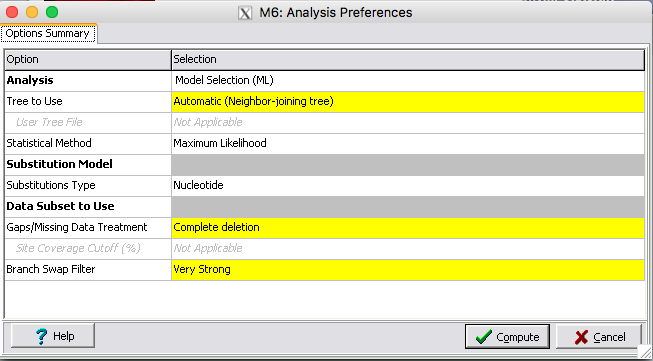
**Software**

* MEGA v7 (megasoftware.net)

Open MEGA and click on *Data*, *Open A File/Session*, and find the alignment from the previous prac (EBOV\_N2\_aligned\_early\_samples.fasta). The program will ask you whether you want to analyse or align the data. Select *Analyze*.

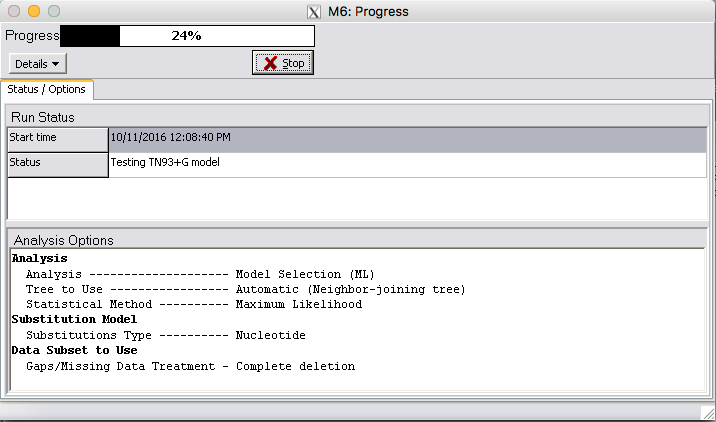
MEA will then ask you whether the data correspond to nucleotide or protein sequences. Select *Nucleotide Sequences* and click OK. It will also ask whether the data are protein coding sequences. Click on *No*. Although they correspond to a protein-coding gene, codon models are very computationally intensive, so we will use nucleotide models instead.

Once the data are loaded in MEGA, click the  button and select *Find Best DNA/Protein Models (ML)…* If it asks you whether you want to use the active data, select Yes. The window in Fig 1 will appear.



**Fig 1.** Substitution model selection menu.

The default settings for the model selection procedure are fine. Click on *Compute*. A window with the progress of the analysis will appear (Fig 2).



**Fig 2.** Progress of model selection in MEGA.

MEGA performs a very thorough model selection. It tests 24 substitution models and calculates the likelihood, and the AICc and BIC scores.

**Question 2.1:** What is the optimal substitution model for these data? What assumptions does it make about the evolution of these sequences?

**Question 2.1:** How many parameters does the optimal model have? What do these parameters represent?

**Question 2.3:** Do the BIC and AICc agree on the optimal model chosen?

**Practical 1c: Maximum likelihood analysis in PhyML**

**Data set**

* Sequence alignment in phylip format of samples collected before the 2013-2016 epidemic:

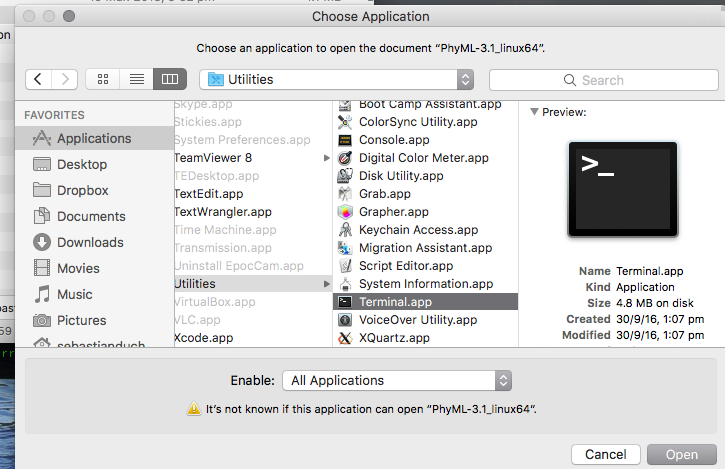
EBOV\_N2\_aligned\_early\_samples.phy

**Software**

* PhyML
* FigTree

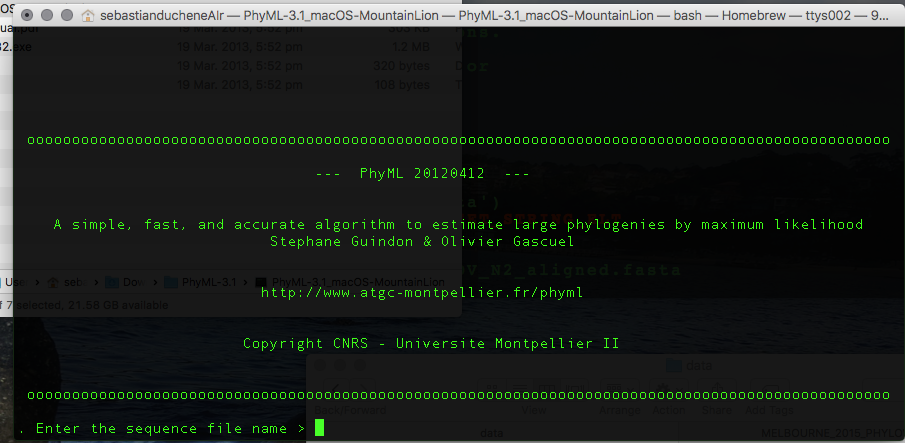
PhyML is typically used through the command-line. We will learn how to use it interactively. Please refer to the user manual, or ask me later about running PhyML in batch mode.

Open the PhyML folder (PhyML-3.1). It should contain executable versions for different operating systems. In windows, double click the .txt file (PhyML-3.1\_win32.exe). In OSX and Linux machines the system might prompt you to choose a program to open this file. If this happens, click on *Choose Application...* and find a folder called *Utilities*. Make sure that at the bottom of the window you select *Enable: All Applications* (Fig 1):



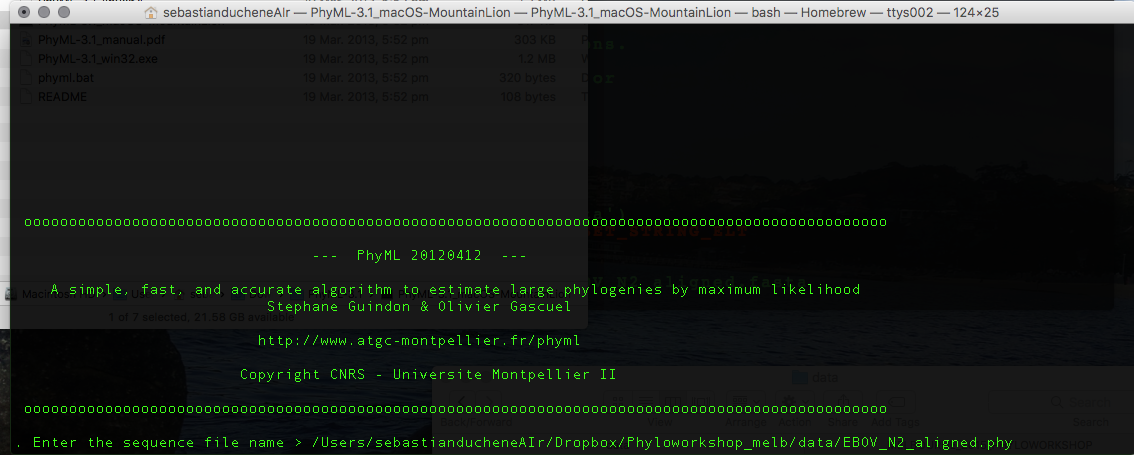
**Fig 1.** Choosing Terminal to open PhyML.

In all operating systems, a command-line window will appear (Fig 2).



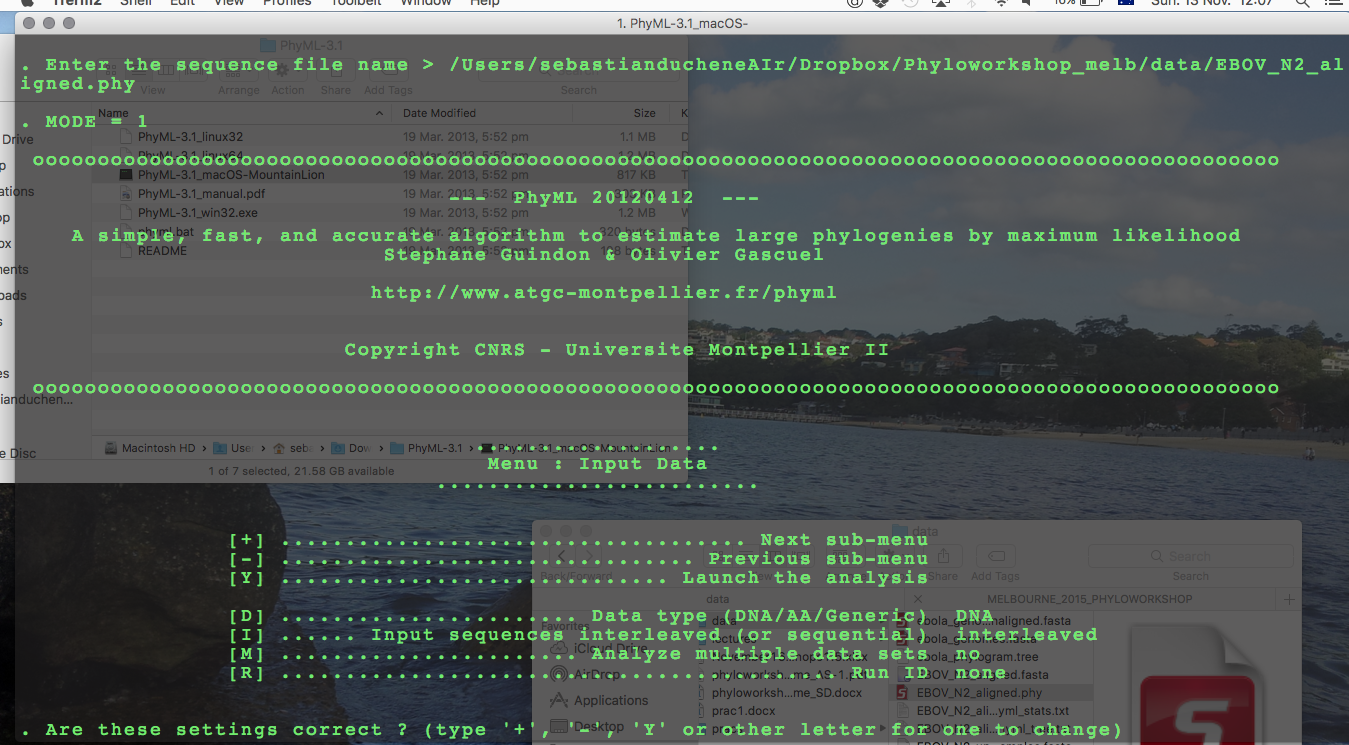
**Fig 2.** Command-line interface for PhyML.

It will ask for the sequence alignment file. Drag the EBOV\_N2\_aligned\_early\_samples.phy. The full path to the file will appear in the window. **Sometimes doing this adds a space at the end of the file path.** Delete it and hit Enter. If you run into an error, it might be that some of the folder names have spaces. In this case, change the location of the alignment, or delete all spaces from file and folder names.



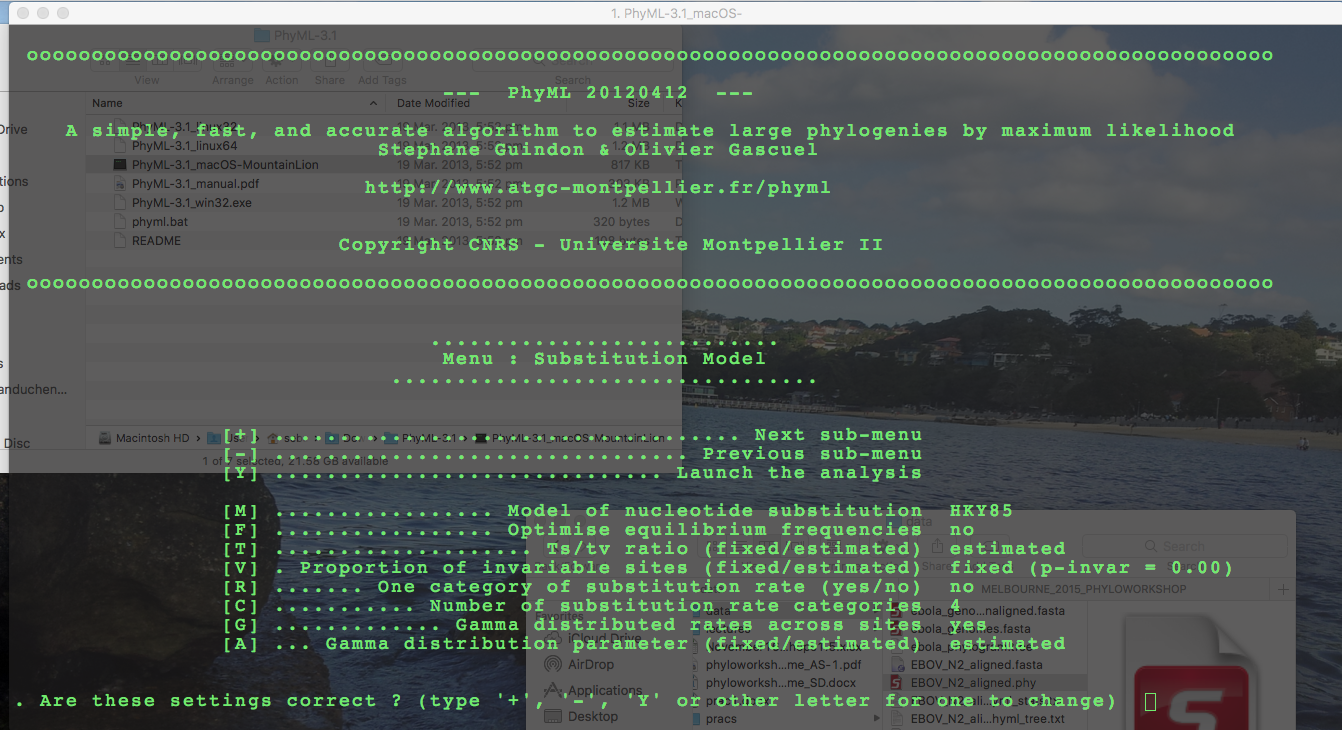
**Fig 3.** File path in PhyML. Note that there should be no spaces at the end of the file path.

You should now see a menu with some options (Fig 4).

****

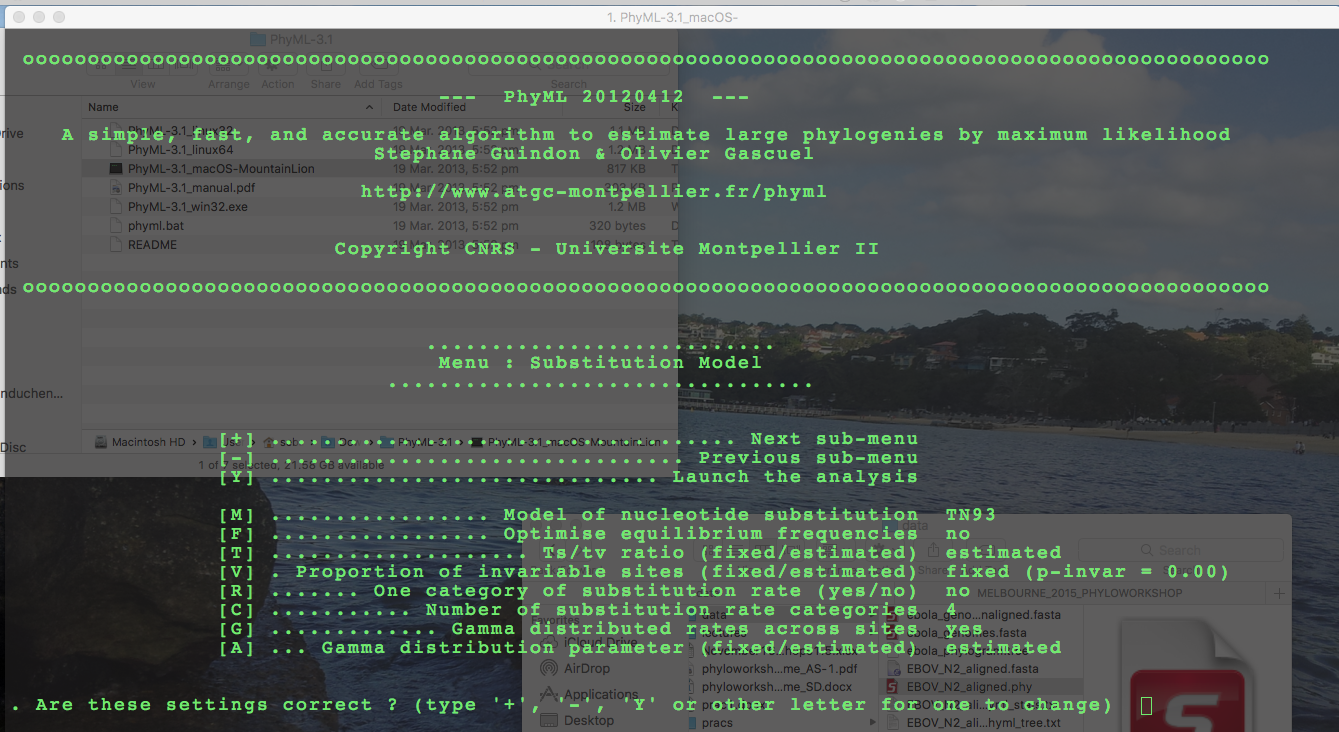
**Fig 4.** Options in PhyML.

These settings are OK. Type + to go to the next menu. This will present some options for the substitution model (Fig 5).



**Fig 5.** Substitution model options in PhyML.

Type M and hit enter a few times until you see the TN93 model, which was selected in MEGA. The resulting set up is in Fig 6.

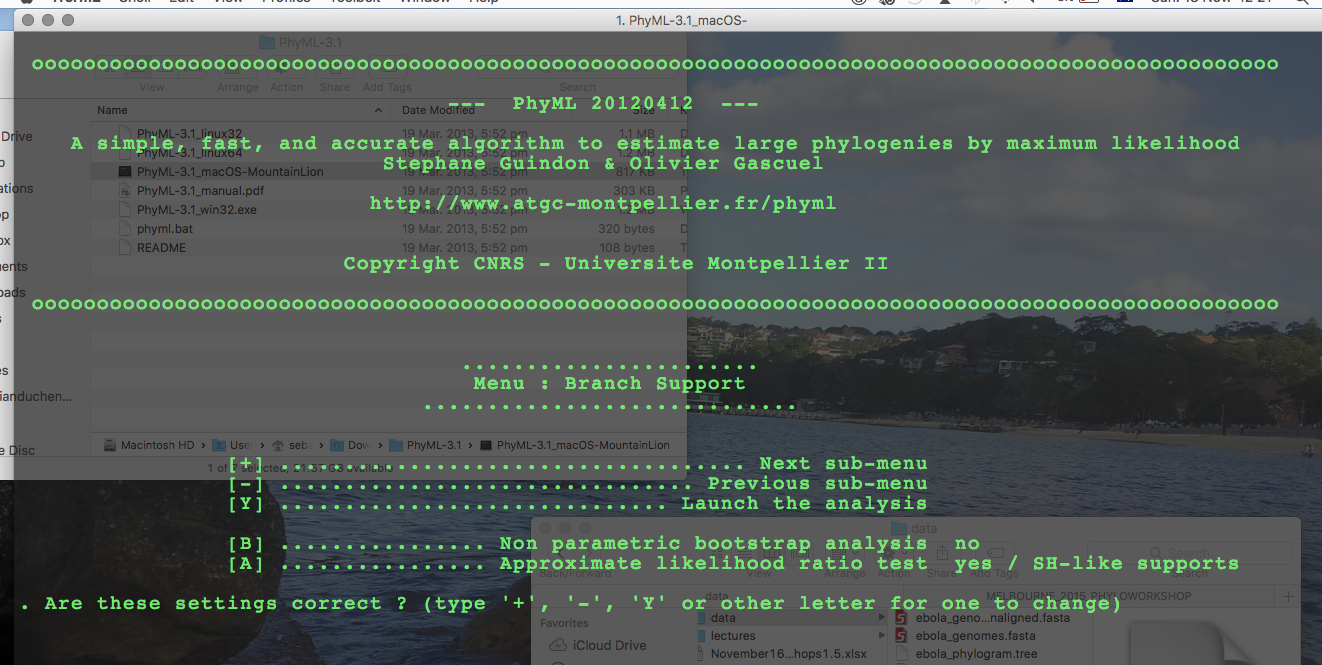


**Fig 6.** Substitution model set up in PhyML.

The substitution model should match the optimal model from MEGA. In this case, we are using a TN93 substitution matrix, which allows for transition-to-transversion bias, a gamma distribution with four categories, and different frequencies for nucleotides. We are assuming no invariable sites.

To go to the next submenu, type + and hit enter. We can set up other options about optimising the tree topology. These are fine for this analysis, but to get more accurate results, change the Tree topology search operations from NNI to SPR (this might not make a difference in this analysis because the data are informative, and are not many sequences). Type + and hit enter.

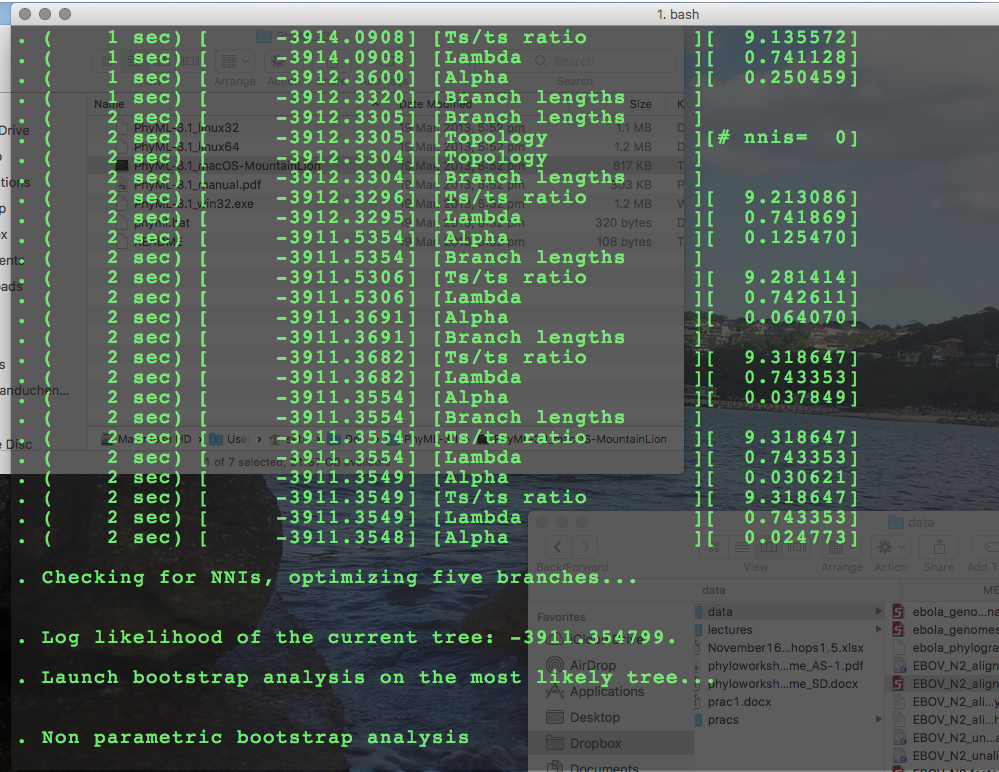
The next submenu is for options of assessing branch support (Fig 7).



**Fig 7.** Options for branch support in PhyML.

PhyML can perform a local topology test to assess branch support, which is very fast and performs similarly to the non-parametric bootstrap. For comparison, however, we will use the non-parametric bootstrap here.

Type A and hit enter to disable the Approximate likelihood ratio test (you might need to do this twice), and then type B and hit enter. The program will ask you how many bootstrap replicates it should perform. Type 100 and hit enter. It will then ask you to whether it should print bootstrap trees and statistics, which we do not require here. Select N and click enter. Out analysis in PhyML is now ready to run. Type Y and hit enter. The analysis will start. You should see the screen in Fig 8.



**Fig 8.** Progress of maximum likelihood analysis in PhyML.

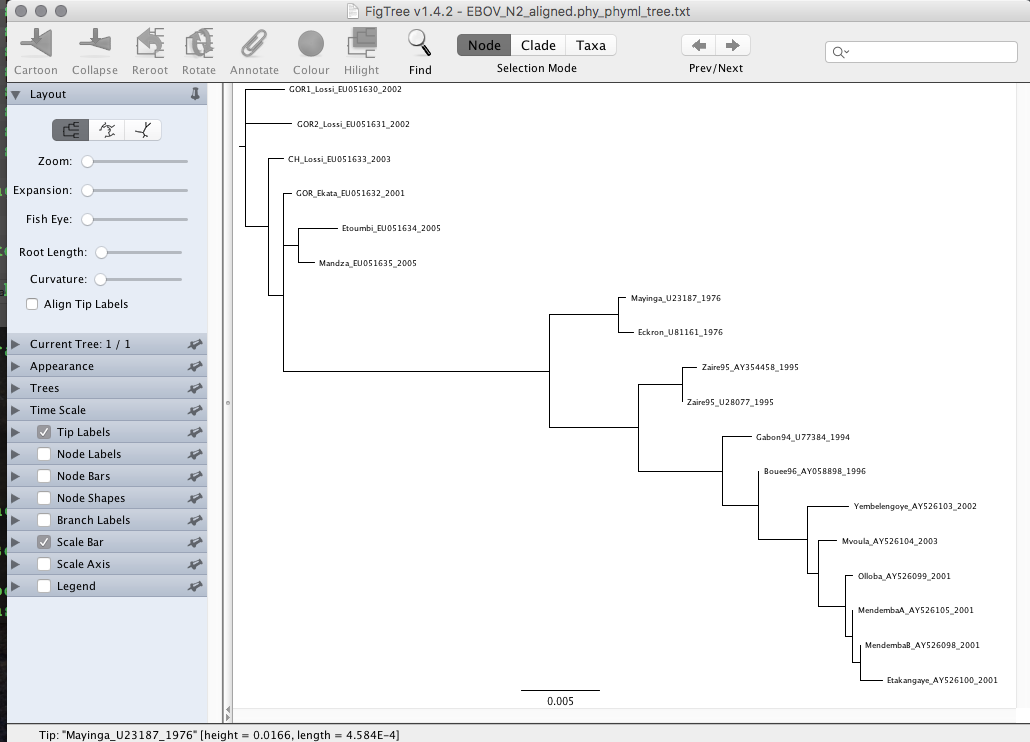
The output from PhyML is a tree file (EBOV\_N2\_aligned\_early\_samples.phy\_phyml\_tree.txt) and a file with the maximum likelihood parameter estimates (EBOV\_N2\_aligned\_early\_samples.phy\_phyml\_stats.txt). Open both in a text editor.

**Question 3.1:** Is there a strong transition-to-transversion bias in these data? Does it appear to differ between purines and pirimidines?

**Question 3.2:** Do you think that it is reasonable to assume that all sites evolve at the same rate in these data?

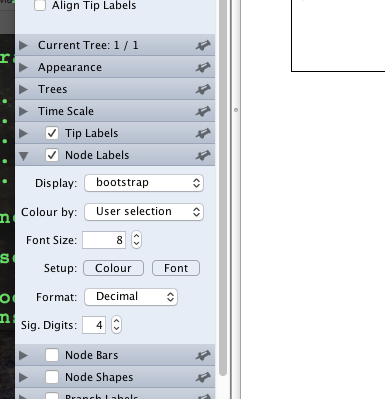
**Question 3.3:** Why do we expect the bootstrap analysis to take longer than an ordinary analysis without branch support?

To visualise the tree we will use FigTree. Click on the program icon . The to the *File* menu and select *Open*. Find the tree file from PhyML (EBOV\_N2\_aligned\_early\_samples.phy\_phyml\_stats.txt). The program will ask us about some branch labels in the tree, which we know to be the bootstrap values. Type ‘bootstrap’ in lieu of ‘label’, and click OK. A window with the tree will appear (Fig 9).



**Fig 9.** Phylogenetic tree displayed in FigTree.

Select *Node Labels* in the menu at the left. Then click on the arrow at the right and select *Display: bootstrap* (Fig 10).



**Fig 10.** Displaying bootstrap values in nodes in FigTree.

The values that appear at the nodes represent the number of bootstrap replicates that contained a particular clade. Use the options in the window to display the tree in different ways. Try selecting different branches and rerooting it.

**Question 3.4:** Sequences starting with CH were obtained from chimps, those starting with GOR were sampled from gorillas, and the rest are from humans. Do human samples all form a monophyletic group? Does this support the notion that Ebola virus has an animal reservoir? If so, can we establish what this reservoir would be?

**Question 3.5:** What does the scale bar at the bottom mean?

**Optional Exercise:** Conduct the same analysis in PhyML, but instead of conducting non-parametric bootstrap, use the SH-like branch support. These values range between 0 and 1. Do they produce similar support as the non-parametric bootstrap?

**Optional Exercise**: Open the Virus Pathogen Resource data base (<https://www.viprbrc.org>) in a web browser and check the location of the strains in this analyses, which are shown for each sequence. For example, Mayinga\_U23187\_1976 means that the sample is from the Mayinga strain, it has the GenBank accession number U23187, and it was sampled in 1976. Is there evidence for geographic clustering in these data?