Genes and Genomes in Populations and Evolution - BIO00056I

Virus Evolution practical session 2019

Daniel Jeffares and Ville Friman

NB: This workshop was adapted (very slightly) from material to us by Francois Balloux.

General considerations

Phylogenetics is a relatively complex field, with a variety of tools adapted to very specific questions and analyses. It is also a field which has not clear-cut boundaries with related disciplines such as population genetics, statistics and bioinformatics. As such, it is not reasonable to expect an in depth treatment of any aspect of phylogenetics within an afternoon. This practical has been devised to give a glimpse of the vast publicly available sequence resources and illustrate the kind of research questions that can be addressed. While the practical is (hopefully) easy to follow as a cookbook recipe, the program should be sufficiently light to leave some time to explore the databases and the Mega software.

Getting data

There is a wealth of freely available sequence data available on Genbank. The downside is that most data is often poorly annotated and misses the critical associated information (e.g. date and place of collection). Here we will take advantage of the new influenza virus resource database, which is relatively well curated:
https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database

Generating a dataset

If you go to the link above, you will get to the following window:

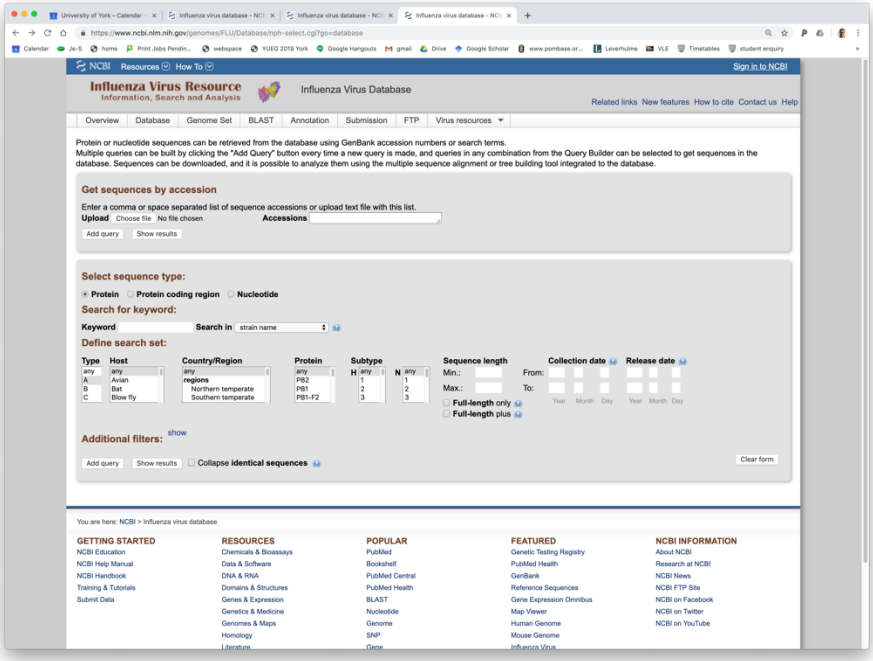


Figure 1: Influenza virus resource website frontpage

As an exercise, we can download all the human full-length H3N2 hemagglutinin (HA) sequences from Wellington (New Zealand). To do so you have to fill the query site as in figure 2.

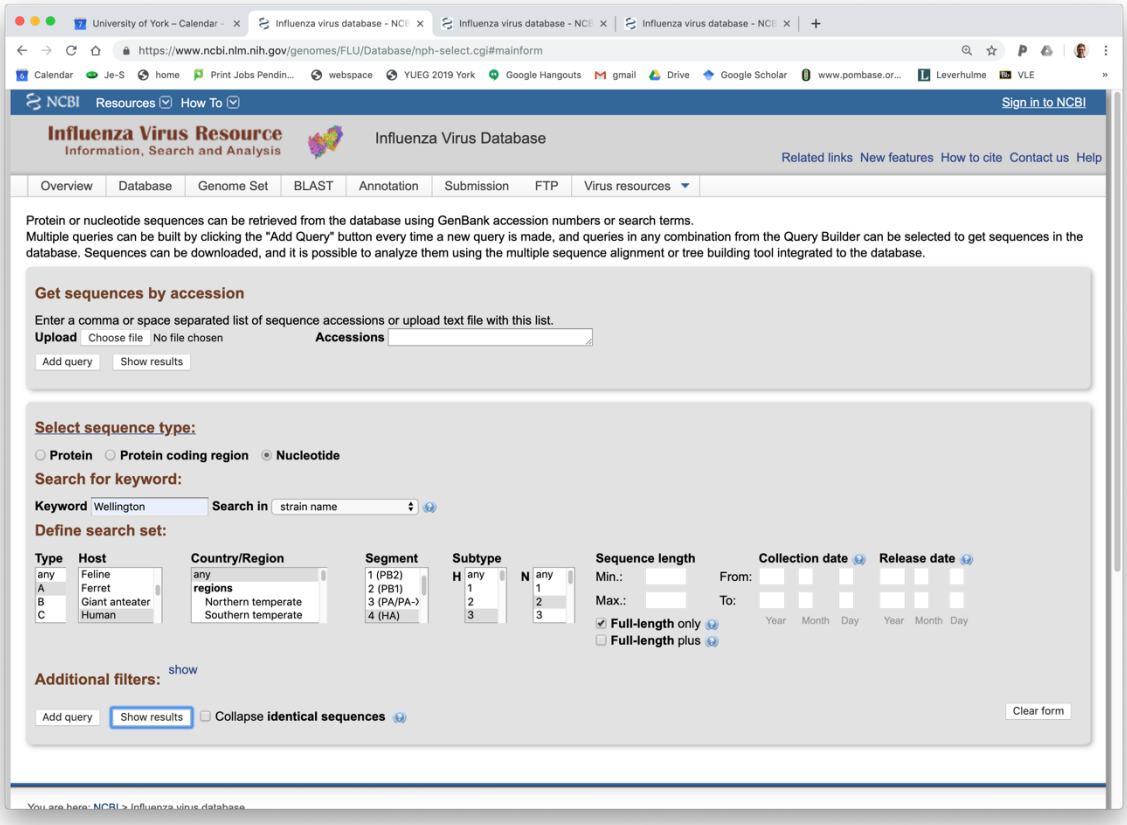


Figure 2. Selecting Type: A, Host: Human, Country/Region: any, Segment: HA, Subtype: H3N2. And tick full length only. Remember to click **Nucleotide**.

If you now click on “Show results”, you will open a new page with a list of sequences (figure 3). You should get 72 strains from 1985 to 2005. First, sort these by year.

Question: Why do you think we chose influenza isolates from the southern hemisphere?

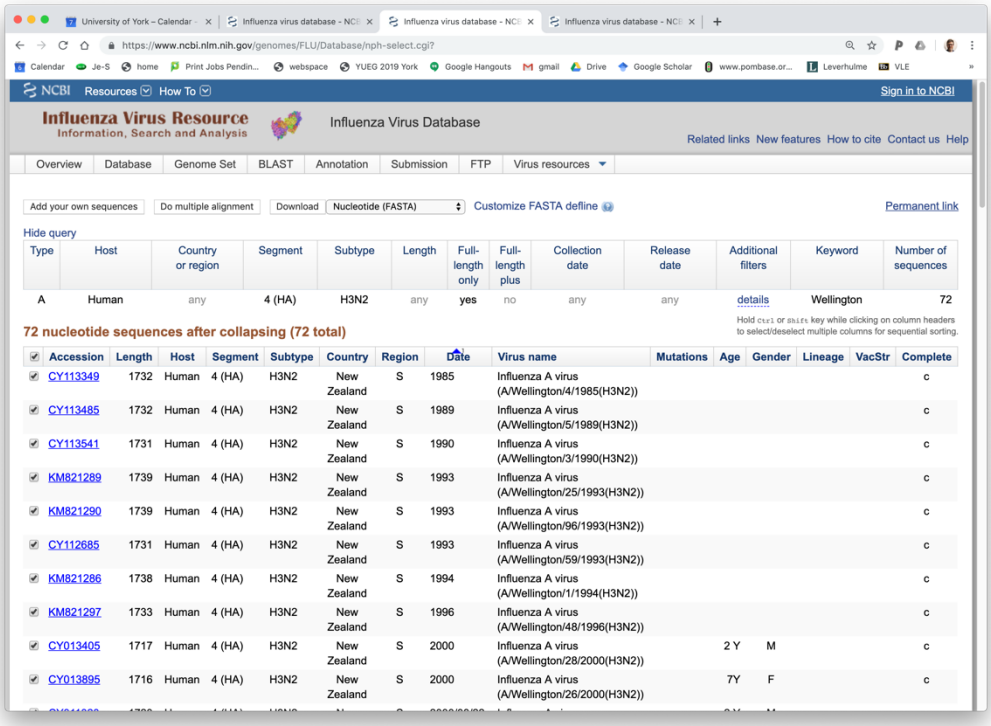


Figure 3. List of queried sequences

You can now align your sequences. This is usually done with dedicated programs but the online tool on the influenza virus resource site is remarkably accurate and fast. Thus, click the "Do multiple alignment" tab. About a minute later or so, the alignment should be finished. Download it and give it a meaningful name (e.g. H3N2_Wellington_seq.fa). This is a Fasta file (.fa extension), which is one of the most standard format for sequence data. We can load this aligned sequence file into MEGA.

Visualising and analysing data

The Mega software

We will use the Mega software for all analyses. It should be installed on your computers. If not download it from http://www.megasoftware.net/. All the screenshots refer to Mega version 7.0. Mega is a reasonably neat package, even if somewhat counterintuitive at first. It does miss some crucial tools such as Maximum Likelihood and Bayesian tree reconstruction, but otherwise includes a fair number of useful features and options. In particular, it offers three distance-based tree reconstruction methods (Neighbour Joining, Minimum Evolution and UPGMA) and also does Maximum Parsimony.

If you start the Mega program, you will get the following window.

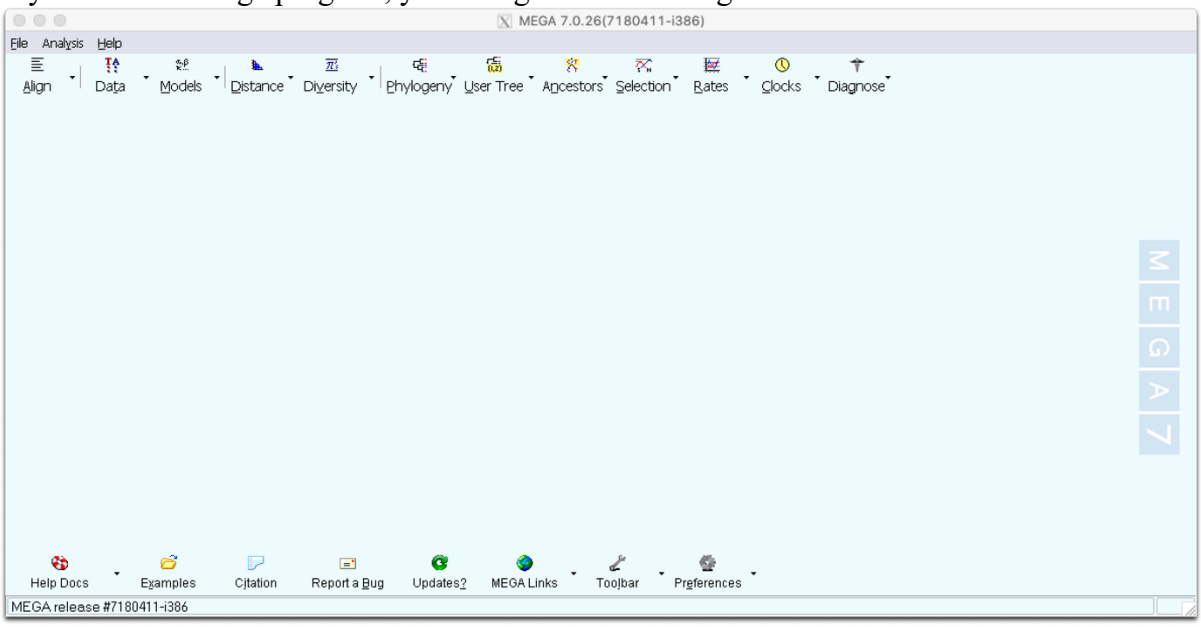


Figure 5. Mega starting window

To explore its features, I recommend you first go to the Tutorial (under Help Docs, bottom left of screen). Skip the "Sequence alignment section", but try to go briefly through the "Building Trees from Distance Data" and the "Computing Statistical Quantities for Nucleotide Sequences" tutorials by using the inbuilt examples.

Once you got somewhat accustomed to the idiosyncrasy of the multiple windows, load the *H3N2_Wellington_seq.fa* file through the **Data > Open a File/Session Menu (Ctrl+O)**.

Then click **Analyse** (not align)**, c**lick OK to "nucleotide sequences" to Protein coding" and to "Standard Code". You will get the window displayed in figure 6. This datafile comprises 72 complete HA sequences from influenza H3N2 sampled between 1985 to 2005. Sequences were also been labelled in groups according to their year of collection.

Now we'll have a look at the alignments, and define some groups of strains, defined by the year they were collected.

When you have loaded up the data, you should see a window like this:

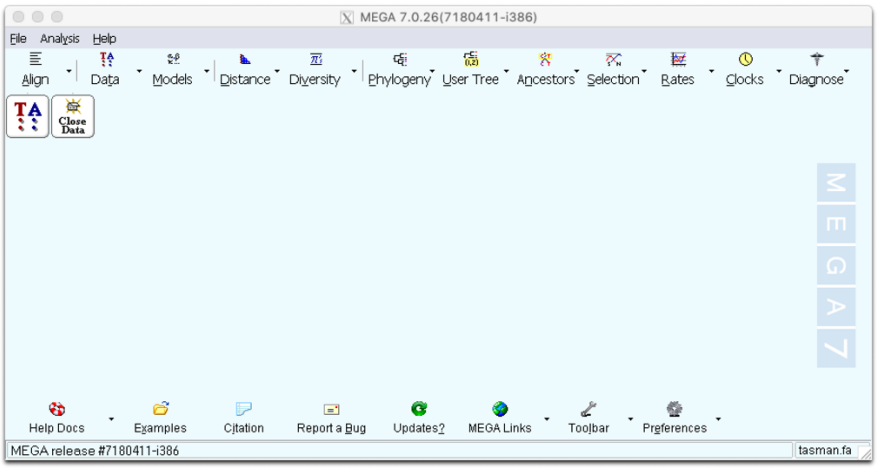
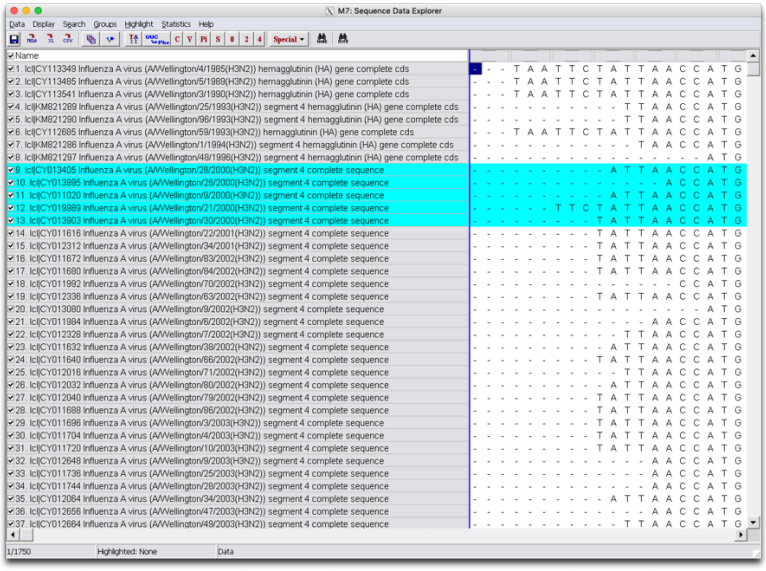


Figure 6. Wellington set: 72 strains from 1985 to 2005.

Click on the "TA" window at top left to see the **Sequence Data Explorer**. You can scroll around and resize this window.

Now let's group the strains by year. This will be fairly easy because we sorted the data by year before we downloaded it. Samples have the year embedded in the same. For example, this one is from 1985; Influenza A virus (A/Wellington/4/**1985**(H3N2)). And this from 2004l; Influenza A virus (A/Wellington/34/2004(H3N2))

To group strain, select one or more strains from a year by clicking on the name of the first strain in the grey part of the window, and shift+click to select the last one, like so:



Then to group these strains, choose the **Groups** menu, then **Add/Edit Group name**. We suggest naming them by year.

Building a tree

As you will have seen, there are a large number of possible options for analysis in MEGA. Feel free to explore the various tools and methods. However, we will build one tree along a single, robust methodology. Go to Phylogeny menu on the main menu. Go to "Bootstrap Test of Phylogeny -> Neighbor-Joining" (Bootstrap represents statistical support for individual clades, values in excess of 70-80% are considered as indicative of well supported). Choose a Kimura 2-parameter model of evolution in Model, otherwise leave all other default options. After less than a minute, you will get a phylogenetic tree. Have a look at it. The different menus allow you to change the presentation. You can also define a root. The clade with the sequences from 1985 at the bottom makes a biologically reasonable root. To define a root, simply right click with your mouse on the chosen node and select "Place root".

One obvious problem with the graphical representation is that the number of taxa is very large. Nicer trees can be generated switching off the tames of the strains, and just showing the group names. You can do this by clicking on **Toggle the display of taxa names**, the lowest button the left hand button bar. You should see a tree like Figure 7.

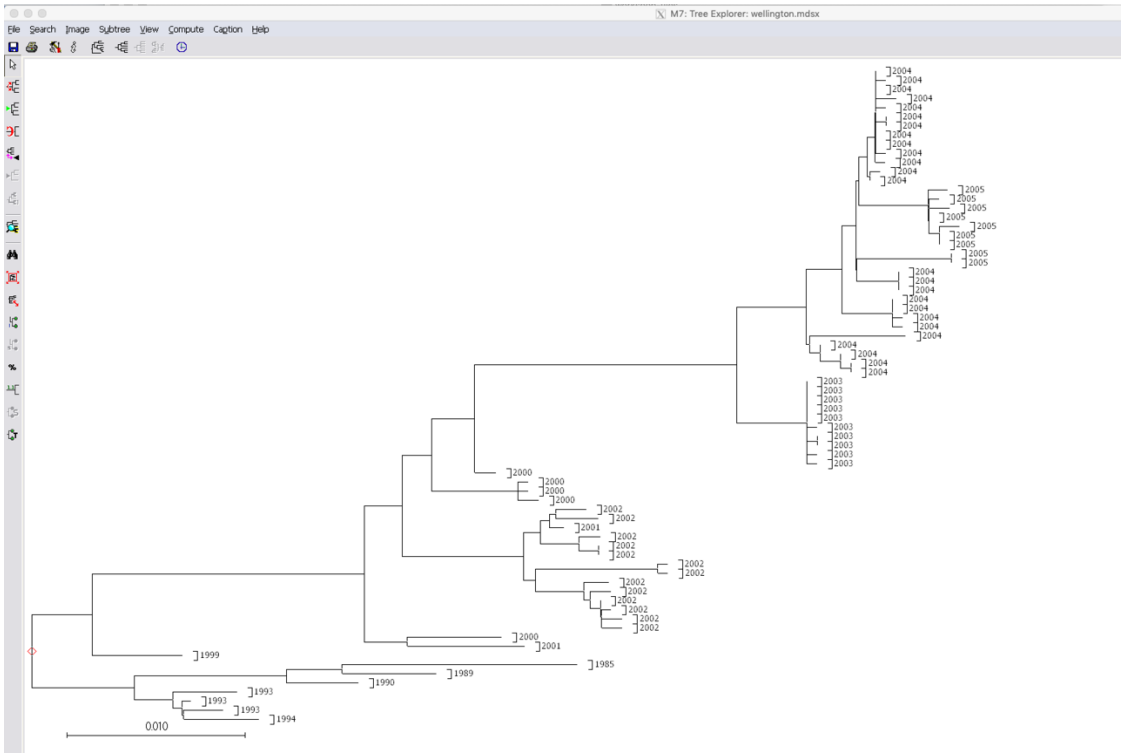


Figure 7. Years shown for Wellington H3N3 HA sequences.

Question: What pattern do you notice about this tree?

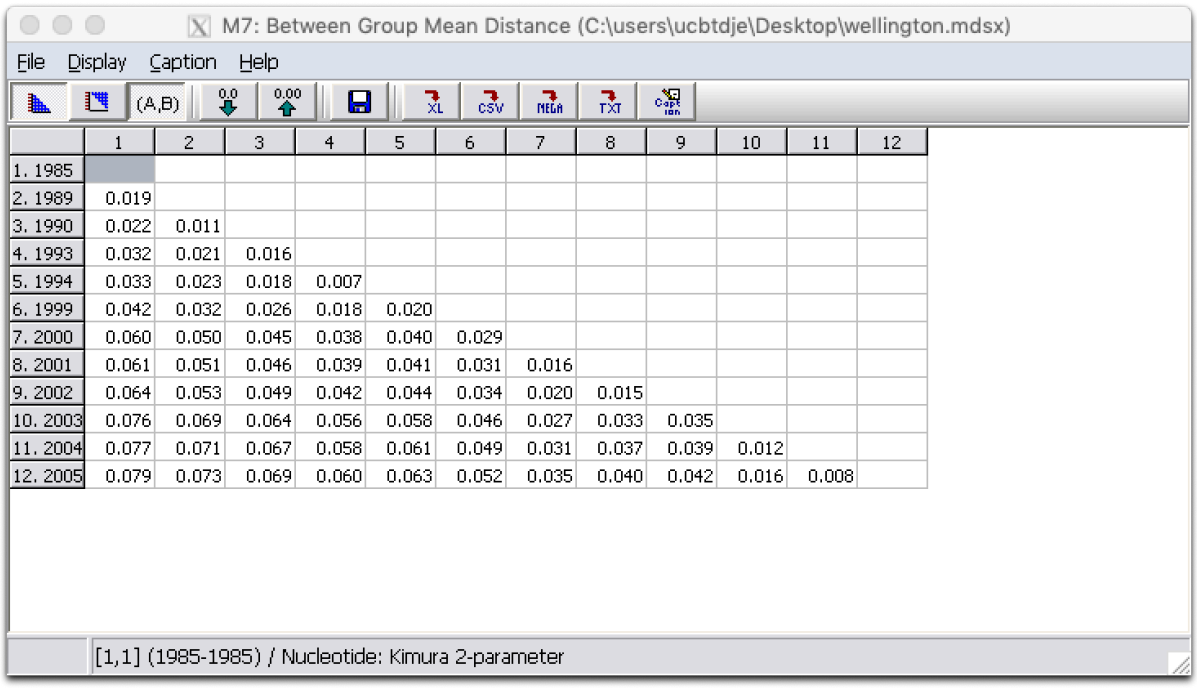
Measuring molecular change over time

An important question in phylogenetics is whether the accumulation of mutation (substitution rate) is constant over time. One way to test for this would be to count the number of mutations from the root to each tip. Estimating the most likely root is not completely straightforward and counting the number of mutations from root to tip would require some scripting well beyond the scope of this practical. However, we can test whether genetic distances increase linearly with time.

Question: Why do you think this is an important question?

Question: Why could substitution rates *not* be constant over time?

To address the issue of linearity between substitution rates and time, we can first compute genetic distances between the sequences grouped by years. To do this, click "Compute Between Group Means", in the Distance menu (light blue master window). Keep Kimura 2 parameter model, and click Compute. You will get the matrix in figure 8.



	1	2	3	4	5	6	7	8	9	10	11	12
1. 1985												
2. 1989	0.019											
3. 1990	0.022	0.011										
4. 1993	0.032	0.021	0.016									
5. 1994	0.033	0.023	0.018	0.007								
6. 1999	0.042	0.032	0.026	0.018	0.020							
7. 2000	0.060	0.050	0.045	0.038	0.040	0.029						
8. 2001	0.061	0.051	0.046	0.039	0.041	0.031	0.016					
9. 2002	0.064	0.053	0.049	0.042	0.044	0.034	0.020	0.015				
10. 2003	0.076	0.069	0.064	0.056	0.058	0.046	0.027	0.033	0.035			
11. 2004	0.077	0.071	0.067	0.058	0.061	0.049	0.031	0.037	0.039	0.012		
12. 2005	0.079	0.073	0.069	0.060	0.063	0.052	0.035	0.040	0.042	0.016	0.008	

Figure 8. Genetic distance matrix between different years

Click on the **File** menu, then "**Export/Print Distances**, choosing the **unformatted text** and column options as Figure 10, below.

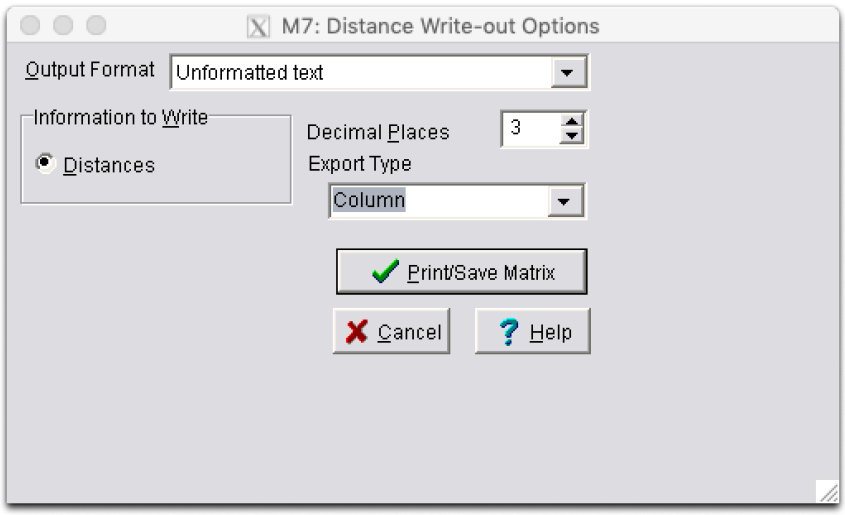


Figure 10. Output genetic distances from MEGA.

The remainder of this part of the workshop is we'll carry out using R Studio.

The all the commands are in the file: **BIO00056I-influenza-practical-2019.R**, which you can find on the VLE.

Part 2: Does the influenza virus spread across the Tasman Sea?

Another use of virus genetic data is to determine how fast pathogens spread. Here, we'll show an example where we examine whether people in New Zealand share the same strains as their neighbours across the Tasman Sea.

If you like, you can choose any two countries, and test these.

To do this, go back to the flue data base:
https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database

Then:

1. Search for strains from Southern Temperate region, HA, H3N2 from year 2000 to year 2000. This will find strains from Australia and New Zealand.
2. So that you don't give MEGA too much data, choose about 100 nucleotide sequences, from this subset in some way (eg: choose all from females).
3. Sort the samples by **virus name**, which happens to sort by country and region. We'll use this to make groups in MEGA.
4. Align, as before
5. Output fasta file, as before. And save the file with a sensible name (tasman.fa).

Then open MEGA, and:

1. Add groups to sets of sequences from the same regions within Australia and NZ. Be sure to name groups to that they include country prefixes, like NZ- and AU-. We'll use these later.
2. Compute inter-group distances, as before.
3. Output the distance file, as before.

We will then examine the between-country differences and compare these to within-country differences. Instructions for this are in the R script.

What would you expect from within- and between-country differences if New Zealanders and Australians had different strains of influenza virus?

That is it for today. I hope you felt the practical was interesting.

Daniel and Ville.