Genes and Genomes in Populations and Evolution - BIO00056I

Virus Evolution practical session 2020

Daniel Jeffares and Ville Friman

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

Aims of the practical, learning outcomes and general considerations

The main aim of this practical is to learn to extract and download genetic data from publicly available databases and to visualise and interpret the data using basic phylogenetic analyses. 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.

By the end of this workshop, you should be able to:

- Build phylogenetic trees using publicly available sequence data using MEGA software
- Infer evolutionary relationships based on phylogeny
- Correlate molecular evolutionary changes with time using R

In order to give you some biological background, you can start by reading two articles provided in VLE:

- Baum *et al*. (2005): The Tree-Thinking Challenge
- Shao *et al.* (2017): Evolution of Influenza A Virus by Mutation and Re-Assortment

Also, watching this video is helpful for understanding how to infer evolutionary trees: https://youtu.be/6_XMKmFQ_w8

Why is this important?

Understanding molecular evolutionary changes helps us to understand evolution in general, but it can also help us to understand disease. For example, the global pandemic of the coronavirus SARS-Cov2, which causes the COVID-19 disease is an RNA virus. As with all viruses, this is undergoing evolutionary change.

There is substantial genomic data for SARS-Cov2, which is publicly available: https://nextstrain.org/ncov/europe?branchLabel=none&c=country

This workshop describes an analysis of the influenza virus, but it would be possible to use the same techniques on SARS-Cov2.

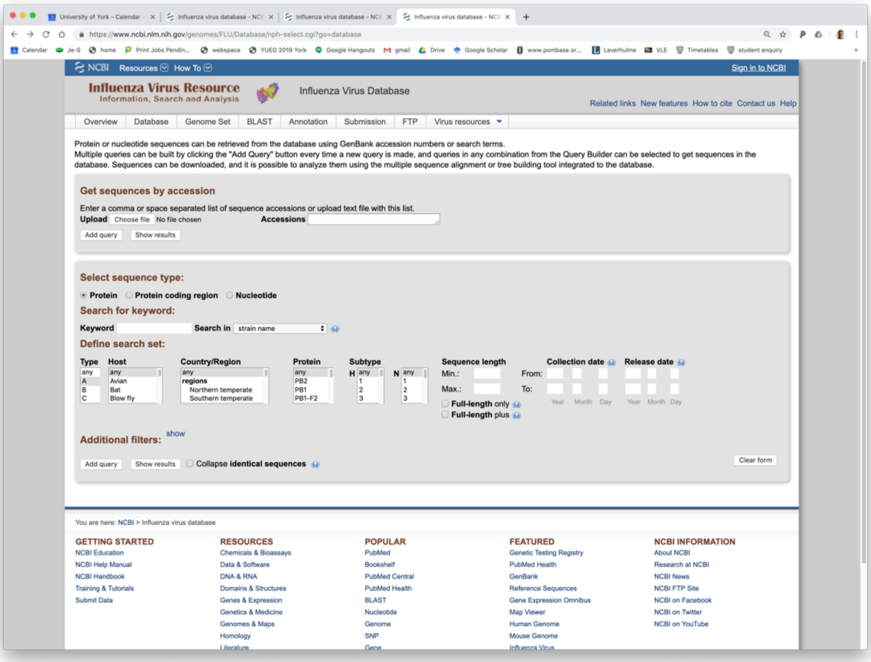
Getting data

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:



https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database

Influenza Virus Resource
Information, Search and Analysis

Influenza Virus Database

Protein or nucleotide sequences can be retrieved from the database using GenBank accession numbers or search terms.
Multiple queries can be built by clicking the "Add Query" button every time a new query is made, and queries in any combination from the Query Builder can be selected to get sequences in the database. Sequences can be downloaded, and it is possible to analyze them using the multiple sequence alignment or tree building tool integrated to the database.

Get sequences by accession

Enter a comma or space separated list of sequence accessions or upload text file with this list.

Upload: No file chosen

Select sequence type: ● Protein ○ Protein coding region ○ Nucleotide

Search for keyword: Keyword ; Search in: strain name

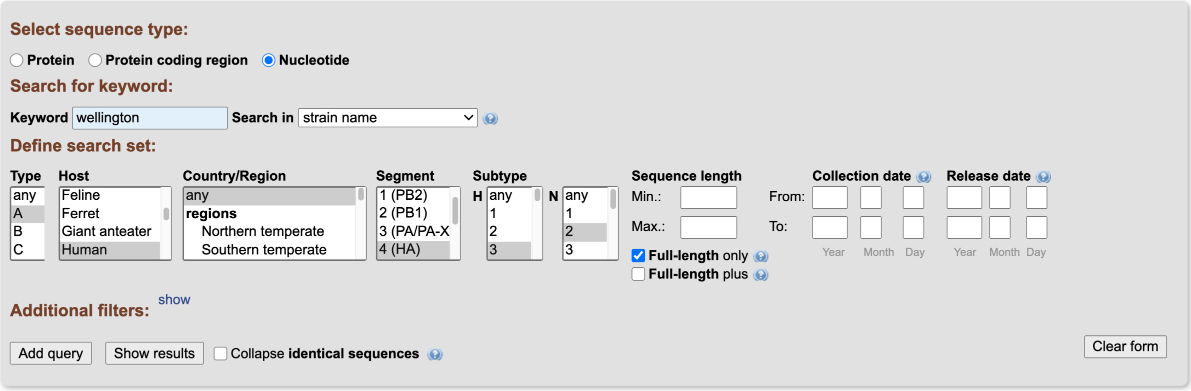
□ Full-length only □ Full-length plus

□ Collapse identical sequences

You are here: NCBI > Influenza virus database

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 (below).**



Select sequence type: ○ Protein ○ Protein coding region ● Nucleotide

Search for keyword:

Keyword: wellington ; Search in: strain name

Define search set:

Type: A ; Host: Human ; Country/Region: any ; Segment: 4 (HA) ; Subtype: H 3, N 2

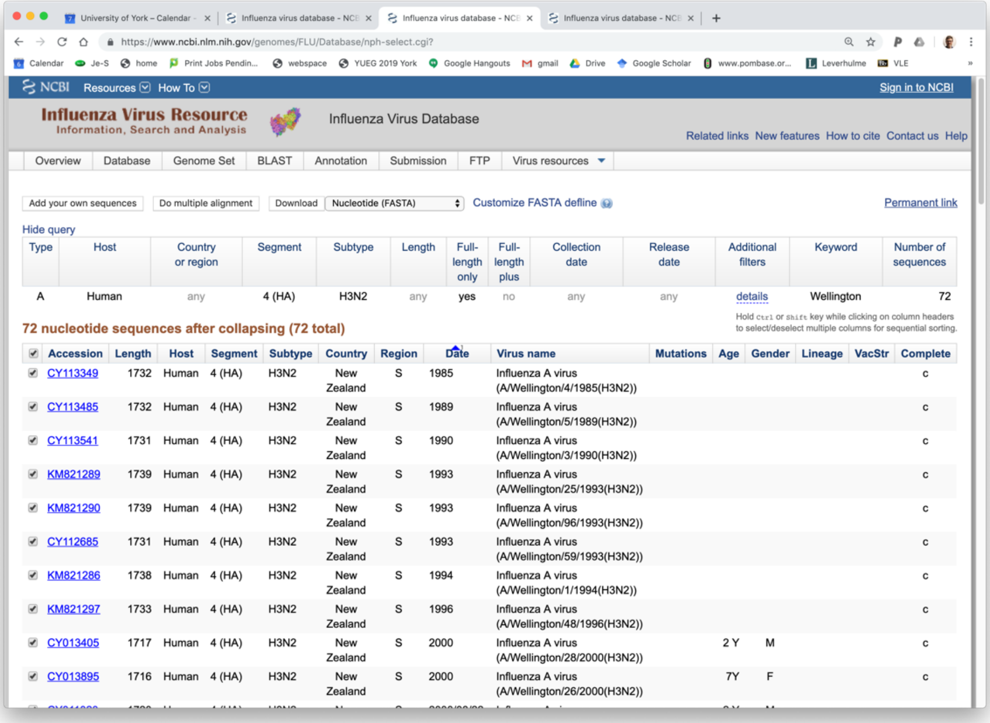
☑ Full-length only □ Full-length plus

□ Collapse identical sequences

Figure 2. Your query should look like this. Select
Select sequence type: Nucleotide.
Keyword: wellington

Type: A
Host: Human
Country/Region: any
Segment: HA
Subtype: H3, N2
Tick full length only

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 **Date**.



Type	Host	Country or region	Segment	Subtype	Length	Full-length only	Full-length plus	Collection date	Release date	Additional filters	Keyword	Number of sequences
A	Human	any	4 (HA)	H3N2	any	yes	no	any	any	details	Wellington	72

Hold ctrl or shift key while clicking on column headers to select/deselect multiple columns for sequential sorting.

72 nucleotide sequences after collapsing (72 total)

☑	Accession	Length	Host	Segment	Subtype	Country	Region	Date	Virus name	Mutations	Age	Gender	Lineage	VacStr	Complete
☑	CY113349	1732	Human	4 (HA)	H3N2	New Zealand	S	1985	Influenza A virus (A/Wellington/4/1985(H3N2))						c
☑	CY113485	1732	Human	4 (HA)	H3N2	New Zealand	S	1989	Influenza A virus (A/Wellington/5/1989(H3N2))						c
☑	CY113541	1731	Human	4 (HA)	H3N2	New Zealand	S	1990	Influenza A virus (A/Wellington/3/1990(H3N2))						c
☑	KM821289	1739	Human	4 (HA)	H3N2	New Zealand	S	1993	Influenza A virus (A/Wellington/25/1993(H3N2))						c
☑	KM821290	1739	Human	4 (HA)	H3N2	New Zealand	S	1993	Influenza A virus (A/Wellington/96/1993(H3N2))						c
☑	CY112685	1731	Human	4 (HA)	H3N2	New Zealand	S	1993	Influenza A virus (A/Wellington/59/1993(H3N2))						c
☑	KM821286	1738	Human	4 (HA)	H3N2	New Zealand	S	1994	Influenza A virus (A/Wellington/1/1994(H3N2))						c
☑	KM821297	1733	Human	4 (HA)	H3N2	New Zealand	S	1996	Influenza A virus (A/Wellington/48/1996(H3N2))						c
☑	CY013405	1717	Human	4 (HA)	H3N2	New Zealand	S	2000	Influenza A virus (A/Wellington/28/2000(H3N2))		2 Y	M			c
☑	CY013895	1716	Human	4 (HA)	H3N2	New Zealand	S	2000	Influenza A virus (A/Wellington/26/2000(H3N2))		7Y	F			c

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/. Use the latest version, Mega X. 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 X program, you will get the following window.

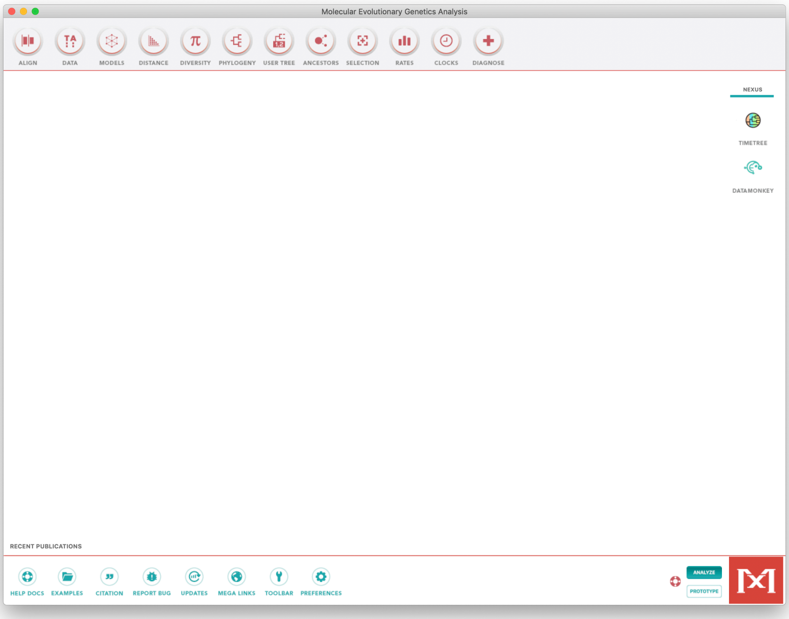


Figure 5. Mega starting window

To explore its features, I recommend you first Mega walk-through, here: https://www.megasoftware.net/web_help_10/index.htm#t=Introduction.htm

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), click 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:

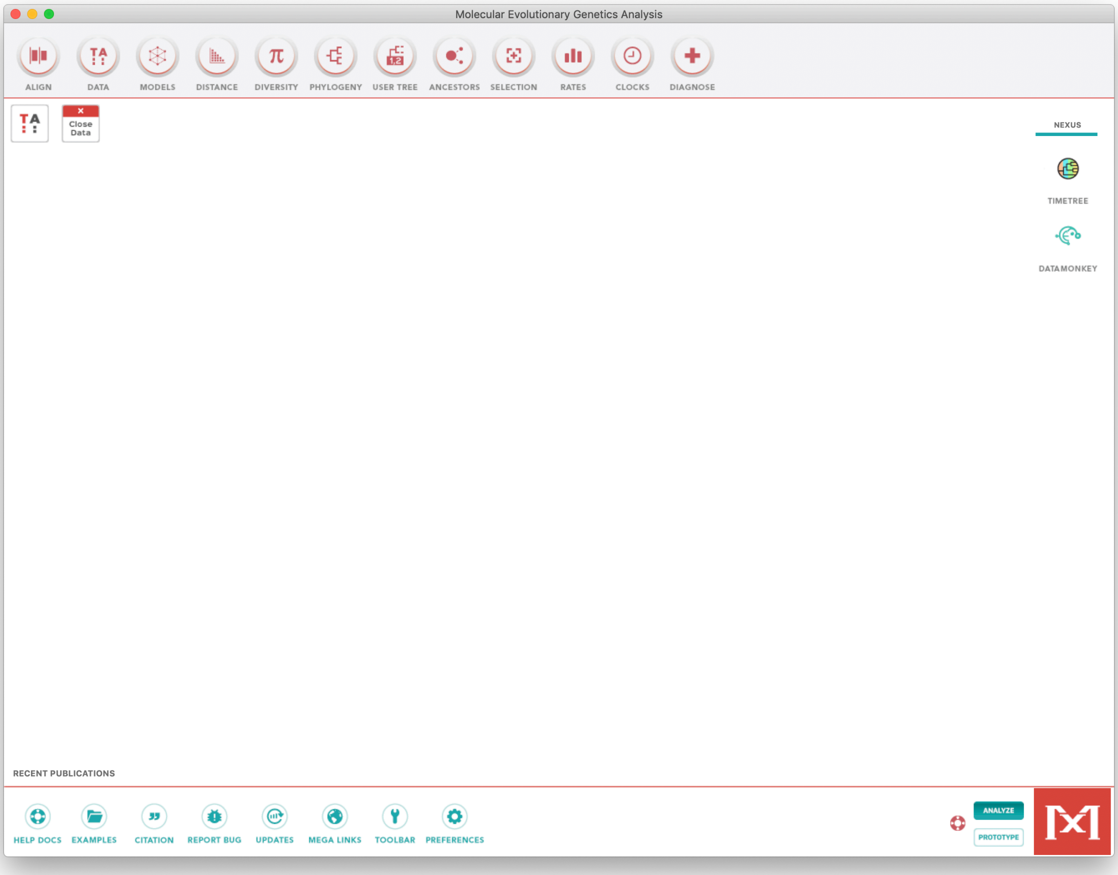


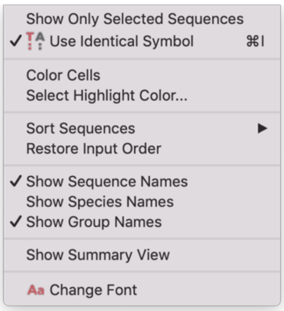
Figure 6. Wellington set: 72 strains from 1985 to 2005.

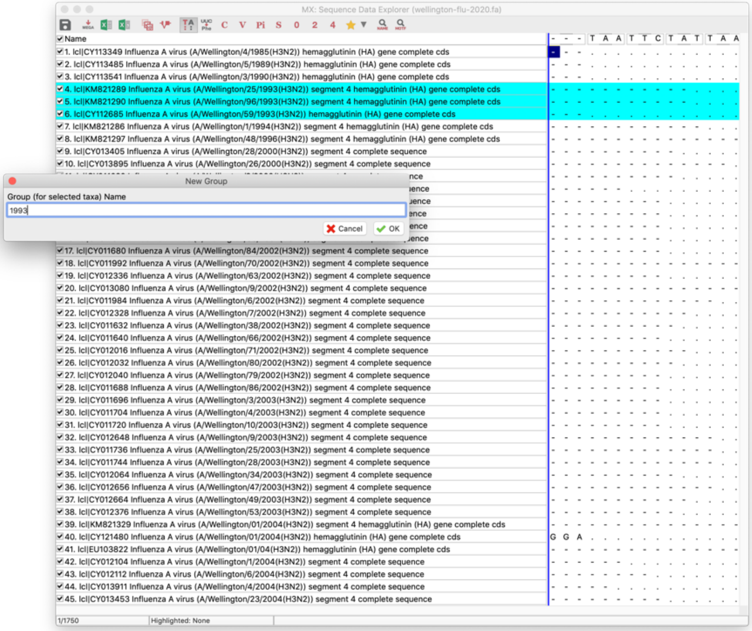
Click on the "TA" window at top left to see the **Sequence Data Explorer**. You can scroll around and resize this window.

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 2004; Influenza A virus (A/Wellington/34/**2004**(H3N2)).

First, go the **Display** menu, and click on **Show Group Names**. This will help you to keep track on what you've done.

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, as below:





MX: Sequence Data Explorer (wellington-flu-2020.fa)

☑ Name

☑ 1. lcl|CY113349 Influenza A virus (A/Wellington/4/1985(H3N2)) hemagglutinin (HA) gene complete cds
☑ 2. lcl|CY113485 Influenza A virus (A/Wellington/5/1989(H3N2)) hemagglutinin (HA) gene complete cds
☑ 3. lcl|CY113541 Influenza A virus (A/Wellington/3/1990(H3N2)) hemagglutinin (HA) gene complete cds
☑ 4. lcl|KM821289 Influenza A virus (A/Wellington/25/1993(H3N2)) segment 4 hemagglutinin (HA) gene complete cds
☑ 5. lcl|KM821290 Influenza A virus (A/Wellington/96/1993(H3N2)) segment 4 hemagglutinin (HA) gene complete cds
☑ 6. lcl|CY112685 Influenza A virus (A/Wellington/59/1993(H3N2)) hemagglutinin (HA) gene complete cds
☑ 7. lcl|KM821286 Influenza A virus (A/Wellington/1/1994(H3N2)) segment 4 hemagglutinin (HA) gene complete cds
☑ 8. lcl|KM821297 Influenza A virus (A/Wellington/48/1996(H3N2)) segment 4 hemagglutinin (HA) gene complete cds
☑ 9. lcl|CY013405 Influenza A virus (A/Wellington/28/2000(H3N2)) segment 4 complete sequence
☑ 10. lcl|CY013895 Influenza A virus (A/Wellington/26/2000(H3N2)) segment 4 complete sequence

New Group

Group (for selected taxa) Name

1993

✖ Cancel ✔ OK

☑ 17. lcl|CY011680 Influenza A virus (A/Wellington/84/2002(H3N2)) segment 4 complete sequence
☑ 18. lcl|CY011992 Influenza A virus (A/Wellington/70/2002(H3N2)) segment 4 complete sequence
☑ 19. lcl|CY012336 Influenza A virus (A/Wellington/63/2002(H3N2)) segment 4 complete sequence
☑ 20. lcl|CY013080 Influenza A virus (A/Wellington/9/2002(H3N2)) segment 4 complete sequence
☑ 21. lcl|CY011984 Influenza A virus (A/Wellington/6/2002(H3N2)) segment 4 complete sequence
☑ 22. lcl|CY012328 Influenza A virus (A/Wellington/7/2002(H3N2)) segment 4 complete sequence
☑ 23. lcl|CY011632 Influenza A virus (A/Wellington/38/2002(H3N2)) segment 4 complete sequence
☑ 24. lcl|CY011640 Influenza A virus (A/Wellington/66/2002(H3N2)) segment 4 complete sequence
☑ 25. lcl|CY012016 Influenza A virus (A/Wellington/71/2002(H3N2)) segment 4 complete sequence
☑ 26. lcl|CY012032 Influenza A virus (A/Wellington/80/2002(H3N2)) segment 4 complete sequence
☑ 27. lcl|CY012040 Influenza A virus (A/Wellington/79/2002(H3N2)) segment 4 complete sequence
☑ 28. lcl|CY011688 Influenza A virus (A/Wellington/86/2002(H3N2)) segment 4 complete sequence
☑ 29. lcl|CY011696 Influenza A virus (A/Wellington/3/2003(H3N2)) segment 4 complete sequence
☑ 30. lcl|CY011704 Influenza A virus (A/Wellington/4/2003(H3N2)) segment 4 complete sequence
☑ 31. lcl|CY011720 Influenza A virus (A/Wellington/10/2003(H3N2)) segment 4 complete sequence
☑ 32. lcl|CY012648 Influenza A virus (A/Wellington/9/2003(H3N2)) segment 4 complete sequence
☑ 33. lcl|CY011736 Influenza A virus (A/Wellington/25/2003(H3N2)) segment 4 complete sequence
☑ 34. lcl|CY011744 Influenza A virus (A/Wellington/28/2003(H3N2)) segment 4 complete sequence
☑ 35. lcl|CY012064 Influenza A virus (A/Wellington/34/2003(H3N2)) segment 4 complete sequence
☑ 36. lcl|CY012656 Influenza A virus (A/Wellington/47/2003(H3N2)) segment 4 complete sequence
☑ 37. lcl|CY012664 Influenza A virus (A/Wellington/49/2003(H3N2)) segment 4 complete sequence
☑ 38. lcl|CY012376 Influenza A virus (A/Wellington/53/2003(H3N2)) segment 4 complete sequence
☑ 39. lcl|KM821329 Influenza A virus (A/Wellington/01/2004(H3N2)) segment 4 hemagglutinin (HA) gene complete cds
☑ 40. lcl|CY121480 Influenza A virus (A/Wellington/01/2004(H3N2)) hemagglutinin (HA) gene complete cds
☑ 41. lcl|EU103822 Influenza A virus (A/Wellington/01/04(H3N2)) hemagglutinin (HA) gene complete cds
☑ 42. lcl|CY012104 Influenza A virus (A/Wellington/1/2004(H3N2)) segment 4 complete sequence
☑ 43. lcl|CY012112 Influenza A virus (A/Wellington/6/2004(H3N2)) segment 4 complete sequence
☑ 44. lcl|CY013911 Influenza A virus (A/Wellington/4/2004(H3N2)) segment 4 complete sequence
☑ 45. lcl|CY013453 Influenza A virus (A/Wellington/23/2004(H3N2)) segment 4 complete sequence

1/1750 Highlighted: None

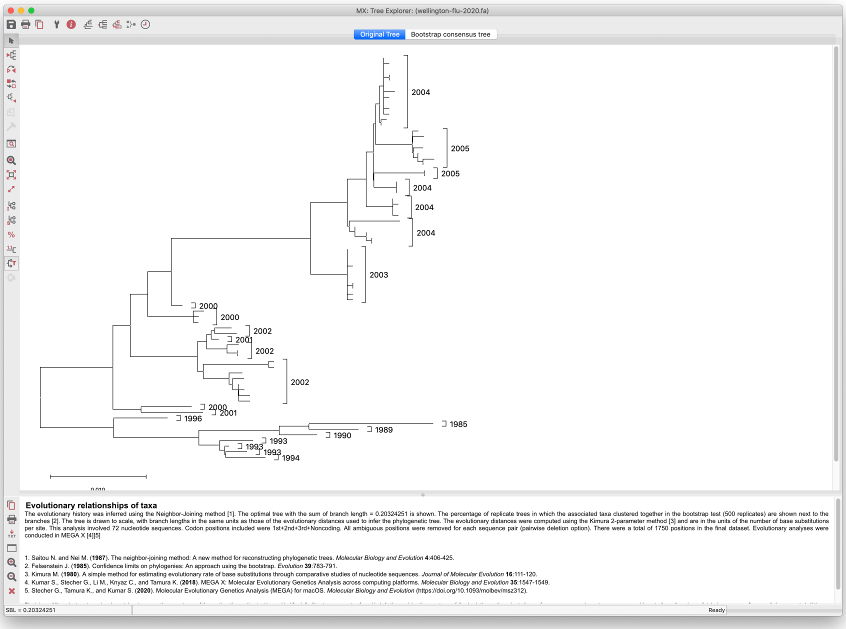
Then to group these strains, choose the **Groups** menu, then **Add/Edit Group name**. It will be very useful later if you name them by year (here I am selecting and naming the 1993 strains).

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 **Phylogeny -> Construct/Test 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. 1000 bootstraps should run relatively fast.

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. Another the **Autosize the tree** button is also useful. You should see a tree like Figure 7.



Evolutionary relationships of taxa

The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 0.20324251 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Kimura 2-parameter method [3] and are in the units of the number of base substitutions per site. This analysis involved 72 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 1750 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [4][5]

1. Saitou N. and Nei M. (**1987**). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406-425.
2. Felsenstein J. (**1985**). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-791.
3. Kimura M. (**1980**). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**:111-120.
4. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* **35**:1547-1549.
5. Stecher G., Tamura K., and Kumar S. (**2020**). Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Molecular Biology and Evolution* (https://doi.org/10.1093/molbev/msz312).

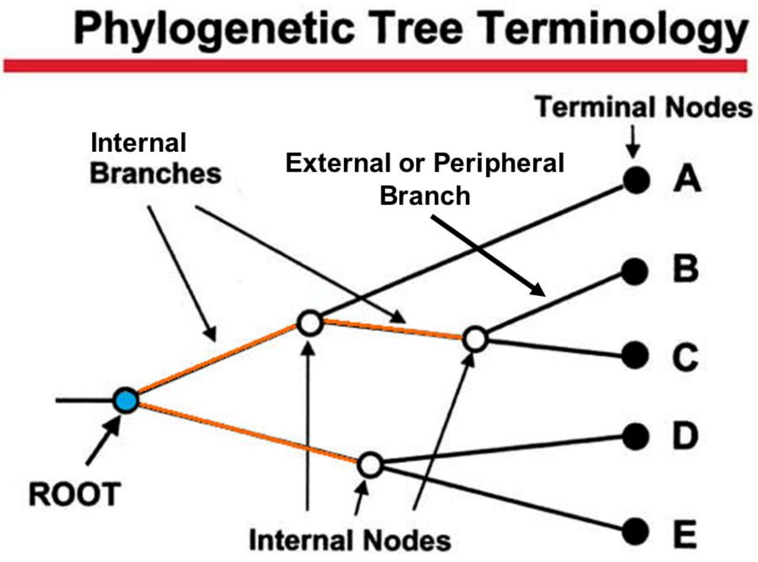


Figure 7. Top, Years shown for Wellington H3N3 HA sequences. **Below**, some terminology we use then describing phylogenetic trees.

Question 1: Look at the main group of branches from 1996 to 2005. What pattern do you notice about this tree?
Tip: Notice how far the 'tips' of the branches are from the from the root of the tree.

Question 2 : Some internal branches are longer than others. What do branch lengths represent, and what does this mean?

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

Question 4: 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 (at the top). Keep Kimura 2 parameter model, and click Compute. You will get the matrix in figure 8.

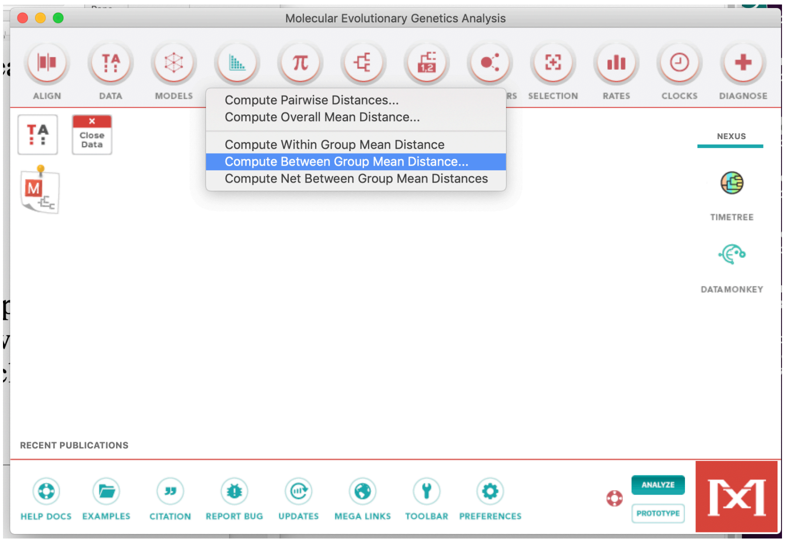
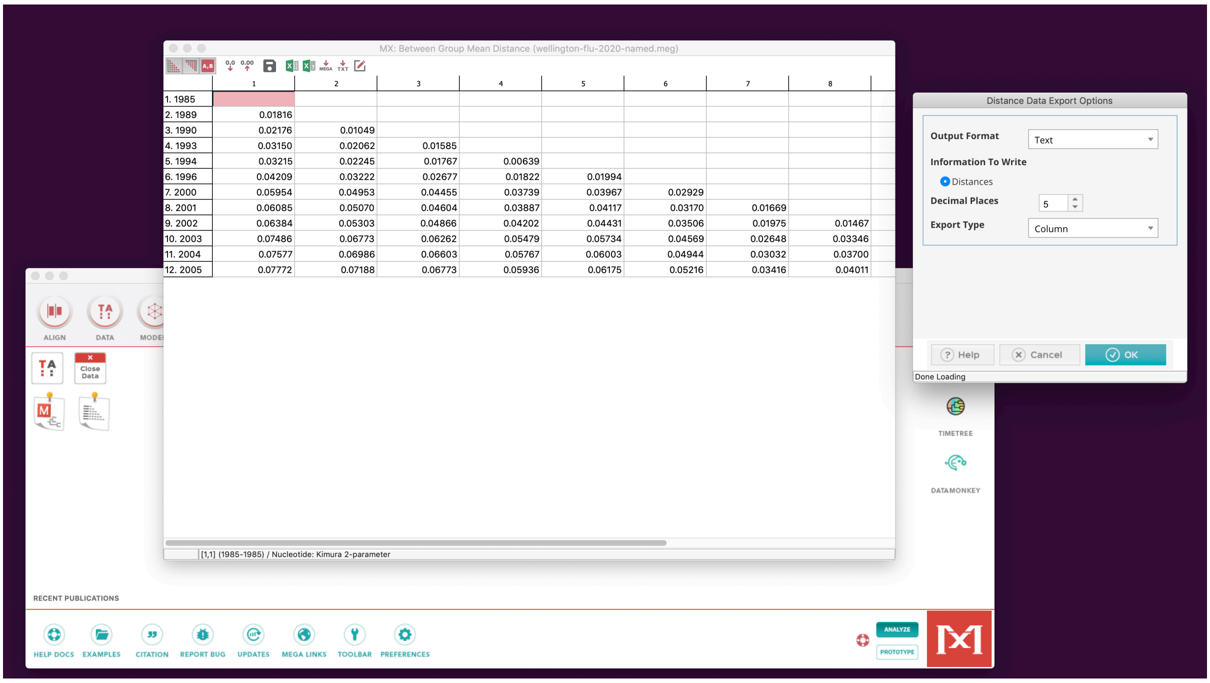

Molecular Evolutionary Genetics Analysis
ALIGN
DATA
MODELS
SELECTION
RATES
CLOCKS
DIAGNOSE
Compute Pairwise Distances...
Compute Overall Mean Distance...
Compute Within Group Mean Distance
Compute Between Group Mean Distance...
Compute Net Between Group Mean Distances
Close Data
NEXUS
TIMETREE
DATAMONKEY
RECENT PUBLICATIONS
HELP DOCS
EXAMPLES
CITATION
REPORT BUG
UPDATES
MEGA LINKS
TOOLBAR
PREFERENCES
ANALYZE
PROTOTYPE

Figure 8. Genetic distance matrix between different years

Now export this data to a file. Click on the **File** menu, then “**Export/Print Distances**, choosing the **text** output format and **column** Export Type options as Figure 10, below.



MX: Between Group Mean Distance (wellington-flu-2020-named.meg)

	1	2	3	4	5	6	7	8
1. 1985								
2. 1989	0.01816							
3. 1990	0.02176	0.01049						
4. 1993	0.03150	0.02062	0.01585					
5. 1994	0.03215	0.02245	0.01767	0.00639				
6. 1996	0.04209	0.03222	0.02677	0.01822	0.01994			
7. 2000	0.05954	0.04953	0.04455	0.03739	0.03967	0.02929		
8. 2001	0.06085	0.05070	0.04604	0.03887	0.04117	0.03170	0.01669	
9. 2002	0.06384	0.05303	0.04866	0.04202	0.04431	0.03506	0.01975	0.01467
10. 2003	0.07486	0.06773	0.06262	0.05479	0.05734	0.04569	0.02648	0.03346
11. 2004	0.07577	0.06986	0.06603	0.05767	0.06003	0.04944	0.03032	0.03700
12. 2005	0.07772	0.07188	0.06773	0.05936	0.06175	0.05216	0.03416	0.04011

[1,1] (1985-1985) / Nucleotide: Kimura 2-parameter

Distance Data Export Options

Output Format: Text

Information To Write

● Distances

Decimal Places: 5

Export Type: Column

Done Loading

Figure 10.
Output genetic distances from MEGA.

Analysis of genetic change with time.

The remainder of this part of the workshop will use R Studio, so **start R Studio now**.
NB: It would be possible to do this work in Excel. But Excel is harder to scale up, harder to describe what to do, and almost impossible to document.

The all the commands you’ll need, and the instructions are available here:
https://www-users.york.ac.uk/~dj757/BIO00056I/BIO00056I-influenza-practical-2020.R
We also provide all the R code at the end of this document.

During this analysis you will produce a scatter plot comparing the between-group genetic distance with between-group time difference.

Question 5: Why does this demonstrate genetic change with time?
Question 6: What is the ‘molecular clock’, and how fast does it ‘tick’ in this case?
Question 7: In the absence of purifying selection to remove mutations at important functional sites, how long would it take for there to be a mutation at every site in the influenza virus genome?

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 flu 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.

Question 8: 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. We hope you felt the practical was interesting.

Daniel and Ville.

R Code:

We do **not** advise copying and pasting code from this pdf. Better to get the code from the link below, where the text won't be mangled by Microsoft Word!

https://www-users.york.ac.uk/~dj757/BIO00056I/BIO00056I-influenza-practical-2020.R

```
##############################################################################################
#Genes and Genomes in Populations and Evolution - BIO00056I
#Influenza Virus Practical 2020
#Ville Friman and Daniel Jeffares (via Francois Balloux)
##############################################################################################

#set your working directory
#Your working directory will be different!
setwd("/Users/ucbtdje/gd/teaching/modules/Genes_and_genomes_in_populations_and_evolution_BIO00
056I/workshops/influenza/workshop-files")

##############################################################################################
#Wellington data
##############################################################################################
#clear al your previous variables
rm(list=ls())

#set your working directory to where the data is
#NB:  your working directory will look different to mine!
setwd("/Users/ucbtdje/gd/teaching/modules/Genes_and_genomes_in_populations_and_evolution_BIO00
056I/workshops/flu2020")

#open your distances file, into a data frame:
#Your distance file will have a different name
flu <-read.csv("distancedata",h=T)

#take a look at your file
#note that the year groups are called Species.1 and Species.2
#not so helpful, but it will do
#and the genetic distance is called Dist
View(flu)

#calculate how many years between each group comparsison
flu$time.passed = flu$ Species.2 - flu$Species.1

#see if distance increases with time, as mutations accumulate
plot(flu$time.passed,flu$Dist,
     xlab ="time (years)",
     ylab ="genetic distance")

#add a line of best fit:
abline(lm(flu$Dist ~ flu$time.passed),col=2,lty=2)

#wow! looks like there is a correlation :)
#let's test it
cor.test(flu$time.passed,flu$Dis)


##############################################################################################
#Tasman sea data
##############################################################################################

#open the tasman data
tas <- read.table("tasman-distance.txt",h=T)

#label NA regions, using grep
#set up AUS as a default for both regions
tas$region1 = "AUS"
tas$region2 = "AUS"

#lable the NZ rows, using grep
tas[grep("NZ", tas$loc1),]$region1 = "NZ"
tas[grep("NZ", tas$loc2),]$region2 = "NZ"

#get the within country data
within = subset(tas, region1 == region2)
between = subset(tas, region1 != region2)
```

```
#make a plot to see if within-country groups are more closely relared than between-country
groups
stripchart(
  list(within$distance,between$distance),
  vert=T,
  method="jitter",
  pch=1,
  group.names=c("within","between"),
  ylab="distance"
)

#test whether there is any significant difference (doesn't look like it, does it)
wilcox.test(within$distance,between$distance)

#############################################################################################
#END
#############################################################################################
```

ANSWERS:

Question 1: Look at the main group of branches from 1996 to 2005. What pattern do you notice about this tree?
Tip: Notice how far the ‘tips’ of the branches are from the from the root.
Answer: Later years are have longer root to tip branch lengths.

Question 2: Some internal branches are longer than others. What do branch lengths represent, and what does this mean?
Answer: Branch lengths are proportional to the number of substitutions that have occurred between nodes. Substitutions are mutations that have been fixed in the clade. This suggests that longer the branch length, the higher the genetic divergence between the strains.

Is the accumulation of mutation (substitution rate) is constant over time?
Question 3: Why do you think this is an important question?
Answer: If mutations occur constantly, this is consistent with the *neutral model of molecular evolution*, where most mutations are neutral or deleterious

Question 4: Why could substitution rates *not* be constant over time?
Answer: If the rates are episodic, or different in different branches, this may indicate some non-neutral evolution, such as *adaptive evolution*. Perhaps due to some differences in the host environment. Alternatively, the population size may have changed (remember that population size influences the proportion of mutations that are neutral).

Question 5: Why does this demonstrate genetic change with time?
Answer: The branch lengths are indicative of mutations that have occurred between the viruses that are circulating between successive years. Each year the virus is different, because mutations accumulate.

Question 6: What is the ‘molecular clock’, and how fast does it ‘tick’ in this case?
Answer: The molecular clock concept refers to the approximately constant accumulation of genetic changes with time. This flu virus data is a good example. See here for more
https://www.nature.com/articles/nrg1020

Question 7: In the absence of purifying selection to remove mutations at important functional

sites, how long would it take for there to be a mutation at every site in the influenza virus genome?
Answer: We observe approximately 0.08 substitutions per site occurring over a 20 year period. A constant clock appears to be reasonable, given the linear slope of the genetic distance vs time distance plot. To achieve 1 change per site, we would require (1/0.08) x 20 years = 250 years. In reality, some sites will change much more slowly, as mutations will be removed due to purifying selection.

Question 8: What would you expect from within- and between-country differences if New Zealanders and Australians had different strains of influenza virus?
Answer: If different strains were circulating between these two locations, we would expect between-country differences to be greater than within-country differences.
This if we observed this, it would be consistent with these countries receiving viruses from different locations. If we observed that within and between-country differences were about the same, it would be consistent with one source for both countries.