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

2

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

- man and a second s	r/genomes/FLU/Database/nph-select.cgi?go	=database		Q. \$
Je-S	🥝 webspace 🔇 YUEG 2019 York 🤤	Google Hangouts M gmail 🙆 Drive 🔶	Google Scholar 🚺 www.pombase.or 🚺 Lever	rhuime 📫 VLE 🦁 Timetables 🦁 student enquiry
S NCBI Resources 🕑 Ho	w To 🖂			Sign in to NCBI
Influenza Virus	Resource 45 Influer	nza Virue Databaso		
Information, Search	and Analysis		Related	links New features How to cite Contact us Hel
Ourselaw Database	Conomo Col DI ACT Apposizion	- Submission ETD Mass re		
Overview Database	Genome Set BLAST Annotation	a Submission FTF Virus re	sources +	
Protein or nucleotide sequence Multiple queries can be built by database. Sequences can be c	is can be retrieved from the database us clicking the "Add Query" button every to townloaded, and it is possible to analyze	sing GenBank accession numbers or s- ime a new query is made, and queries a them using the multiple sequence alig	earch terms. in any combination from the Query Builder c ament or tree building tool integrated to the	an be selected to get sequences in the database.
Get seguences by acc	ession			
Get sequences by acc	6651011			
Upload Choose file No file d	arated itst of sequence accessions or up tosen Accessions	road text file with this list.		
Add guery Show results				
How query Show results				
Select sequence type:				
Dretein Oretein eedle	n malan 🔿 Nucleatida			
Oraceh fan brotein codi	ig region O Hucieotide			
Search for keyword:				
Keyword	Search in strain name \$			
Keyword Define search set:	Search in strain name \$	9		
Keyword Define search set:	Search in strain name +	Gubburg Samu	anna lanath Collection data G	Palazza data ()
Keyword Define search set: Type Host C any any	Search in strain name Country/Region Protein any any	Subtype Sequ	ence length Collection date 📦 From:	Release date 😡
Keyword Define search set: Type Host A Avian	Search in strain name Country/Region Protein any egions P2	Subtype Sequ	ence length Collection date 📦 From: To:	Release date 😖
Keyword Define search set: Type Host C any any any A Avian B Bat C Biow fly	Search in strain name Country/Region Protein any egions Page Northern temperate PB1-F2	Subtype Sequ H any N any Min.: 2 2 Max: 3 3 Max:	ence length Collection date G From: To: Var. Month Day	Release date 😠
Keyword Define search set: Type Host C arry arry arry Arry Arvian B B Bat C Blow fly	Search in strain name Country/Region Protein any southern temperate Southern temperate Southern temperate PB1-F2	Subtype Sequ H any N any Min.: 1 2 2 Max: 3 3 Fu Fu	ence length Collection date From: To: It-length torky Year Month Day Hength torky	Release date 📦 Year Month Day
Keyword Define search set: Type Host any my A Avian B Bat C Blow fly	Search in strain name C Country/Region Protein nry regions Anothern temperate Southern temperate	Subtype Sequ Harry Narry Min: 2 3 3 Fu Fu Fu	ence length Collection date From: To: Il-length only & Year Month Day I-length plus &	Reinase date 🐲 Year: Month Day
Keyword Define search set: Type Host A Avian t B Bat B Bat Bat B Bat B Bat Bat Bat B Bat Bat Bat Bat Bat Bat Bat Bat Bat Bat	Search In stain name : Country/Region Protein Mary Search In Sea	Subtype Sequ H arry N arry Mn: 2 2 3 3 Fu Fu Fu	ence length Collection date From: To: H-length plus	Release date 😖 Year Month Day
Keyword Define search set: Type Host A Avian B Bat C Bow fly Additional filters: ^{show}	Search In strain name 2 Country(Region Protein my egion environmentemperate Southern temperate Southern temp	Subtype Sequ H av N av Mn.: 2 3 3 Fu Fu	ence length Collection date From: To: Hength orty Hength plus	Release date as
Keyword Define search set: Type Hoat C Bow Hy Additional filters: show Add query Show results	Search in strain name a Country/Region Protein Mary Morthern temperate Southern temperate Collapse Identical sequences @	Subtype H World N Wy Mn.: 2 2 3 Maxi B 9 Fu	ence length Collection date S From: To: U-longth only S Hength plus S	Release date and a second seco
Keyword Define search set: Type Host are are are are are are are been fly are are are been fly are are are are are are Add quary Show results	Search In strain name 2 Country(Region Protein signers southern temporals Bouthern temporals Collapse Identical sequences @	Subtype Sequ H and N and Mna 2 2 3 A Fu Fu Fu	ence length Collection date a From: To: Beength only a Ware Month Day Heength plus a	Release date to Year Month Day Clear form
Keyword Define search set: Type Heat A Ada a Bat C Boor hy Additional filters: Show results	Search in strain name s Country/Region Protein Wo Wo Southern temperate Southern temperate Collapse Identical sequences @	Subtype H H m N Mm 2 2 2 FH FW FW	ence length Collection date so From: Too: Hength city so Hength plus so	Release date to War Moth Day
Keyword Define search set: Type float average float average float Additional filters: show Additional filters:	Search in strain name 5 Country/Region Protein Pry Search in strain name 5 Protein Pry	Subtype Sequ H PY N Y 3 3 Max Fu Fu Fu	ence length Collection date a From: Hength city (a) Year Month Day Hength plus (a)	Release date a Year Much Day Clear form
Kywod Define seach set: Type Hoat Mark and a set a set and a set and a set a set and a	Search In Exten name Contry/Region Contry/Region Control Cont	Subhype Sequ Harry N Wry Mnz 2 2 Mnz 2 2 Mnz 2 2 Mnz 2 2 Mnz Fu Fu	ance length Collection date is From: Hength only is View More Day Hength plus is	Release date to Year Month Day Clear form
Keyword Define sectors bet: Type foot of the sector bet: """"""""""""""""""""""""""""""""""""	Search In strain name s Country/Region Protein WY Protein Prot	Subtype Sequ H (1) N (1) N (1) Max 3 3 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ance length Collection date and From: From: Historyth only and the from the	Release date a Year Month Day Clear form
Kyword Define seach set: Type Host With a law and a a law and a distance of the second second Additional filters: show Add gury Show results You are here. NGB > influence word CETTING STARTED NGB Second	Search in strain name a country/Region Protein monor Strain Str	Softype Sequence H H H Mm: 2 2 Mm: H H 2 2 2 Mm: Fill P P Mm: Fill Fill 2 2 2 Mm: Fill P P Mm: Fill Fill 2 3 2 Mm: Fill Fill P P Mm: Fill Fill Fill	Ance length Collection date as From: Hength only as Hength plus as FEATURED Constit: tening Regard	Release date to View Moren Day Clear form Clear form Add MOR
Keyword Define sectors bet: Type Nost WWW A data and the sector bet Add quary Show results Add quary Show results You are here: NCB > Informat vice CETTING STARTED NCB High Manual	Search in strain name Country(Region Protein Protei	Subtype Sequence H B 2 2 3 2 9 2 9 2 9 2 9 2 9 2 9 2 9 2 9 9 9	ence length Collection date as From: Too: Hength only as Hength plus as FEATURED Converts Tening Regary Public Hens	Release date Year Much: Day Clear from Clear from
Kywood Define seach set: Type Host With a law of the second set a law of the second set a law of the second set Additional filters: show Additional filters: show Additional filters: show Show second set Castle Second set NGE Host Second set NGE Host Second set	Search in strain name Contry/Region Potein Potein Potein Potein Potein Potein	Softype Sequ H Service Market 2 2 2 Market 2 2 2 Market 2 2 2 Finite Finite PopPLLAR Pablet Booland Pablet Control	ence length Collection date as From: Hength only as Ware More Day Hength plus as FEATURED Original Control Factors Devicts Taxing Regary Public Hardin Ordinal	Release date to View Month Day Clear form Clear form Clear form Man MOB Research at ACC Mich Nos to Mich Nos to M
Keyword Define seach set: Type Neat With And With And Comparison of the Second Additional filters: Add guary Show results Volu ein here: IXCBI > Industria Vide CETTNO STARTED NGBI (disarbion NGBI Handhoot NGBI Handhoot	Search in strain name Country(Region W P P P P P P P P P P P P	So Subtype	Ance length Collection date as From: Term: Heinigh crivy as Heinigh crivy as Heinigh crivitation (Region Contribution (Region Contribution (Region Contribution (Region Contribution) Region (Region Region Region (Region Contribution) Region (Region Contribution) Region Contribution) Region (Region Contribution) Region (R	Release date Ware Much: Day User Much: Day Clear from
Kyword Define seach set: Type Host Market State Barket State Additional filters: show Additional filters: show Additional filters: show Additional filters: Show Show State Certhor State Additional Market State Market State Mar	Search in strain name Contry/Region Potein Potein Potein Potein Potein Potein Potein Potein Potein Potein Potein Potein Potein Potein Potein Po	Softype Sequ H Suffype Sequ 2 2 2 Mark 2 2 2 Mark 2 2 2 Mark 4 Fin 4 Fin	ence length Collection date an From: Hength only as Were More Day Hength plus as FEATURED Oversite: Instig Regary Public Hardh Orders Regenome Rege	Release date to View Merch Day Clear form Clear for
Keyword Define sectors etc. Type Nost WWW WWW WWW WWW Adjust Adjust Vou an Area Kall > Information Adjust Vou an Area Kall > Information Cettino StARTED Notil Kosalin Cettino StarteD Notil Kosalin Sudi Russi Sudi Russi Sudi Russi Sudi Russi Sudi Russi Sudi Russi Sudi Russi	Search in strain name Country(Region Weights Protein Protein Pro	So Subtype H Subtype Subtyp	Ance length Collection date an Prom: Prom: Prom: Prom:	Release date Year Much: Day Lear form Celer form
Kywood Define seach set: Type Host With a law of the second set a law of the second second second second a law of the second second second second second a law of the second seco	Search in strain name Contry/Region Contry/Regio	Softype Sequ Harry N WW Max 2 2 Max 2 2 2 Max Fu Pable Booked Pable Booked Pable Come Pable Booked Pable Booked Pable Booked Pable Booked Pable Booked Pable Pable Pable Booked Pable Pabl	Ance length Collection date and From: From: Heingrith only and the second secon	Release date Ver Work Day Clear form Clear form Clear form Research at NCB NCBI IN FORMATION MCBI IN Formation NCBI IN Formation

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.

Intract/www.ncbi.nlm.nb.gov/genomes/FUUDatabase/nph-select.cg/#maintorm (🔎 🗧 📅 University of York – Calendar - 🛛 😣 Influenza virus	database - NCB 🗙 🗧 Influenza virus databa	ase - NCB 🗙 📔 🔀 Influenza virus d	iatabase - NCE × +	
And and set of the	→ C A https://www.ncbi.nlm.nih.aov/aenomes	/FLU/Database/nph-select.cgi#mainform		,	Q 🛧 🕨 🛆 🌘
NCBI Resources (i) How To (ii) Sign in to NCBI Influenza Virus Resource Information, Search and Analysis Influenza Virus Database Related links New features How to cite. Contact us Ho Overview Database Genome Set BLAST Annotation Submission FTP Virus resources • Related links New features How to cite. Contact us Ho Overview Database Genome Set BLAST Annotation Submission FTP Virus resources • Influenza Overview Database Genome Set BLAST Annotation Submission FTP Virus resources • Influenza Underwise Detabase Genome Set BLAST Annotation Submission FTP Virus resources • Influenza Influenza Misci FTP Virus resources • Influenza Influenza <td>Calendar 👁 Je-S 🐼 home 📮 Print Jobs Pendin 🐼 w</td> <td>abspace 🚱 YUEG 2019 York 💿 Google Ha</td> <td>ingouts M gmail 🔥 Drive ┥</td> <td>Google Scholar</td> <td>or 🚺 Leverhulme 📧 VLE</td>	Calendar 👁 Je-S 🐼 home 📮 Print Jobs Pendin 🐼 w	abspace 🚱 YUEG 2019 York 💿 Google Ha	ingouts M gmail 🔥 Drive ┥	Google Scholar	or 🚺 Leverhulme 📧 VLE
Influenza Virus Resource Influenza Virus Database Related links New features How to cite Contact us He Overview Database Genome Set BLAST Annotation Submission FTP Virus resources Orient or nucleotide sequences can be obtilived from the database using GenBark accession numbers or search terms. Litiple 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 tabase. Sequences can be downloaded, and it is possible to analyze them using the multiple sequence alignment or tree building tool integrated to the database. Setters accession Enter a comma or space separated list of sequence accessions or upload text file with this list. Ipload Toole file Not file chosen Accessions Control for Regulated is to subtrole the database Securities Country/Region Securities Protein Protein Country/Region Securities Subtrom Subtread Subtrom Subtrom Subtrom	NCBI Resources V How To V				Sign in to NCB
Overview Database Genome Set BLAST Annotation Submission FTP Virus resources oblin or nucleotide sequences can be oblit 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 tatabase. Sequences can be downloaded, and it is possible to analyze them using the multiple sequence alignment or tree building tool integrated to the database. Sequences by accession Set sequences by accessions or upload text file with this list. Ipload Choose the Notifie choose Accessions Add query Show results Show results Sequence type: Protein Protein Protein coding region Nucleotide Sequences test: Yee Notifient sequence Sequence length Collection date Release date Min: From: Southern temperate Collapse identical sequences and Collapse identical sequences Collapse identical sequences Southern temperate Southern temperate Collapse identical sequences Clear form	Influenza Virus Resource	Influenza Virus Database		Related links New	features How to cite Contact us H
oble in or nucleotide sequences can be retrieved from the database using GenBank accession numbers or search terms. altiple 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 tabase. Sequences can be downloaded, and it is possible to analyze them using the multiple sequence alignment or tree building tool Integrated to the database. Set sequences by accession Enter a comma or space separated list of sequence accessions or upload text file with this list. lpload Choose file No file chosen Accessions Select sequence fype: Protein O Protein coding region Nucleotide Search for keyword: regvord Wellington Search in strain name 1 [1923] 2 Notifier Terment 1 [1923] 2 Notifier Terment Sequence length Collection date @ Release date @ Maxie: Notifier Terment Notifier Terme	Overview Database Genome Set BLAST	Annotation Submission FT	P Virus resources 🔻		
Set sequences by accessions The reaction of space separated list of sequence accessions or upload text file with this list. Upload Chooses file No file chosen Accessions Add query Show results Sequence fype: Protein Protein coding region Nucleotide Search in strain name Image: Search in strain name Sequence length Collection date Release date Image: The results Northern temperate Sequence length Collection date Release date Image: The results Northern temperate Sequence length Collection date Release date Image: The results Northern temperate Sequence length Collection date Release date Image: The results Northern temperate Sequence length Collection date Release date Image: The results Sequence length Collection date Image: The results Sequence length Northern temperate Sequence length Sequence length	otein or nucleotide sequences can be retrieved from th ultiple queries can be built by clicking the "Add Query" tabase. Sequences can be downloaded, and it is poss	e database using GenBank accession button every time a new query is made ible to analyze them using the multiple	numbers or search terms. , and queries in any combina sequence alignment or tree b	tion from the Query Builder can I puilding tool integrated to the dat	be selected to get sequences in the abase.
Inter a comma or space separated list of sequence accessions or upload text file with this list. Jpload Choose file Notice chosen Add query Show results	Set sequences by accession				
upload Choose file No file chosen Accessions Add query Show results	nter a comma or space separated list of sequence ac	cessions or upload text file with this list			
Add query Show results	Ipload Choose file No file chosen Acc	essions	1		
Belect sequence type: Protein Protein coding region Nucleotide Search for keyword: Keyword Wellington Search in strain name : Profese search set: ype Host Seguence length Particular to the Southern temperate Southern temperat	Add query Show results				
Select sequence type: Protein Protein coding region Nucleotide Search for Keyword:					
Select sequence type: Protein Protein coding region Nucleotide Search for keyword: Nother is strain name Image: Segment Segment Very Host Country/Region Segment Subtype N any Printe Signs Segment 1(P82) N any N any 3 Giant anteater Southern temperate 3 (PA/PA-) 1 2 3 2 Human Southern temperate 3 (PA/PA-) 2 3 3 Year Month Day Year Year Year Month Day Year Year Month Day Year					
Protein Protein Protein Protein Search for keyword: Search in strain name Image: Search set: Vpe Host Country/Region Segment Subtype Inv Feline Image: Segment Subtype Image: Segment Subtype Inv Feline Image: Segment Subtype Image: Segment Segment Image: Segment Segment Name Inv Feline Image: Segment Subtype Image: Segment Segment Image: Segment Segment Name Inv Feline Image: Segment Subtype Image: Name Image: Segment Segment Name Inv Feline Image: Segment Image: Segment Segment Name Image: Segment Name Inv Feline Image: Segment Image: Segment Image: Segment Name Image: Segment Name Inv Feline Image: Segment Image: Segment Image: Segment Name Image: Segment Name Signal anteater Southern temperate Image: Segment Image: Segment Image: Seg					
Search for Keyword: Geyword Wellington Search for Keyword: Define search set: Ype Host Country/Region Segment 1 (P82) H any N any 3 Giant antester Southern temperate 3 (PA/PA-) 3 (PA/PA-) 3 (PA/PA-) 3 (P44) 4 (P4A) 4 (P4A) 3 (PA/PA-) 3 (PA/PA-) 3 Southern temperate 4 (P4A) 4 (P4A) 3 (PA/PA-) 3 (PA/PA-) 3 Southern temperate 4 (P4A) 4 (P4A) 3 (PA/PA-) 3 (PA/PA-) 3 Southern temperate 4 (P4A) 4 (P4A) 3 (PA/PA-) 3 (PA/PA-) 4 (P4A) 3 (PA/PA-) 3 (PA/PA-) 5 Bold Southern temperate 4 (P4A) 6 Giant antester Show Year Month Day Year Month Collapse identical sequences (Particul Sequences) Clear form	elect sequence type:				
Sequence length Collection date @ Release date @ Sequence length Collection date @ Release date @ Marry Feline Marry N any Giont anteater Southern temperate 3 (PA/PA-> 3 (PA/PA-> 3 (PA/PA-> 3 (PA/PA-> Add query Show Show results Collapse identical sequences @ Collapse identical sequences @ Clear form	Select sequence type:				
Operation Country/Region Segment Namy Namy Sequence length Collection date Release date Release date Max Image: Southern temperate 2 (PBI) 2 (PBI) <t< th=""><th>Select sequence type: Protein O Protein coding region ® Nucleotide Search for Keyword</th><th>,</th><th></th><th></th><th></th></t<>	Select sequence type: Protein O Protein coding region ® Nucleotide Search for Keyword	,			
Yper Host Country/Region Segment Subtype Any I (PB2) N any N any Segment N any N any I (PB2) Yes Min: From: Max: To: Water Month Day Year Mo	Select sequence type: Protein O Protein coding region ® Nucleotide Search for keyword: Connor Indiana				
Vype Host Country/Region Segment Subtype Sequence length Collection date () Release date () any Ferret any regions 1 (P62) 1 (P42) 1 Min.: From: Image: From:	Select sequence type: Protein O Protein coding region © Nucleotide Search for keyword: Keyword Wellington Search in strain name Define concept codi	¢) 😡			
B Giant antester Northern temperate 3 (PA/PA-) 2 2 3 Max: To: B Southern temperate 4 (HA) 3 3 3 Image: Construction of the state of t	Select sequence type: Protein O Protein coding region © Nucleotide Search for keyword: Keyword Wellington Search In strain name Define search set:	¢ 😡			
Add query Show results Collapse identical sequences @	Select sequence type: Protein Protein coding region Protein Protein Search for keyword: Search In strain name Ceyword Wellington Search In strain name Define search set: Ype Host A Ferret endows any Peline any A Ferret endows any		Sequence length	Collection date <table-cell> R From:</table-cell>	elease date 🤢
Additional filters: show Add query Show results Collapse identical sequences 🔬	Select sequence type: Protein O Protein coding region O Nucleotide Search for keyword: Keyword Wellington Search In strain name Define search set: Type Host Country/Region The function of the set of the	Segment Subtype 1(PB2) 2(P01) 1 2(P01) 2 1 1	Sequence length	Collection date <table-cell> R From: To:</table-cell>	elease date 😡
Add query Show results Clear form	Select sequence type: Protein O Protein coding region Nucleotide Search for keyword: Search in strain name Define search set: ype Host Genrat In Search in strain name Perine Search Set: Southern temperate Southern temperate	• • • • • • • • • • • • • • • • • • •	Sequence length Min.: Max.: Ø Full-length only Full-length plus	Collection date <table-cell> R From: To: Year Month Day Y</table-cell>	elease date 😖 Year Month Day
Add query Snow results Conapse ruentical sequences w	Select sequence type: Protein Protein coding region Nucleotide Search for keyword: Keyword Wellington Search in strain name Define search set: Type Host Country/Region any Feline Sorthern temperate Southern temperate Southern temperate	Segment Subtype 1(P82) 2(P81) 3(PAPA-2) 4(HA) 3 3	Sequence length Min.: Max.: ✓ Full-length only s Full-length plus s	Collection date 💀 R From: To: Year Month Day Y	elease date 😡 rear Month Day
	Select sequence type: Protein Protein coding region Nucleotide Search for Keyword: Keyword Wellington Search in strain name Define search set: Type Host Country/Region any Feine Bine And anteater C Human Northern temperate Southern temperate Southern temperate	Segment Subtype I(P82) I(P81) 3(PA/PA) 4(HA) N	Sequence length Min.: Max.: ♥ Full-length only ♥ Full-length plus §	Collection date 💀 R From: To: Year Month Day Y	elease date <table-cell></table-cell>
	Select sequence type: Protein Protein coding region Nucleotide Search for Keyword: Search in strain name Define search set: Ype Host Northern temperate Southern temperate Souther		Sequence length Min.: Max.: S Full-length only Full-length plus	Collection date R From: To: Year Month Day Y	elease date 😠 Year Month Day
	Select sequence type: Protein Protein coding region Protein Protein coding region iearch for keyword: Search in strain name verifie search set: Search in strain name verifie search set: Country/Region mv Feline Giant antester Northern temperate Southern temperate Southern temperate Additional filters: Show Add query Show results		Sequence length Min.: Max.: Full-length only Full-length plus §	Collection date <table-cell> R From: To: Year Month Day Y</table-cell>	elease date 😠 Year Month Day Clear form

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?

\rightarrow	CÓ	A https://	www.ncbi.r	nlm.nih.gov/	genomes/FLU	/Database/npl	h-select.c	gi?					Q	☆ P	P & 🔮
Cale	endar 👄 J	s-S 🙆 home	e 👂 Print	Jobs Pendin.	. 📀 webspa	ice 🕝 YUEG	2019 York	Google Han	gouts M	gmail 💧 Drive 🔶 G	oogle Scholar 🌘	www.pombase.or	r 🚺 Leverhulme	U VI	LE
3 N	NCBI Re	sources 🖂	How To (J										Sign	n in to NCBI
1	Influer	za Viru	s Res	OUITCE Analysis	1	Influenz	a Virus	Database			Relate	ed links New f	eatures How to	cite Co	intact us He
0	verview	Database	Genor	me Set	BLAST	Annotation	Submi	ssion FTF	Viru	s resources 🔻					
Add	l your own s	equences	Do multipl	e alignment	Download	Nucleotide	(FASTA)	¢ Cu	stomize F	ASTA defline 😡				Per	rmanent link
ide Typ	query be l	Host	Court or reg	itry jion	Segment	Subtype	Leng	gth Full- length only	Full- length plus	Collection date	Release date	Additional filters	Keyword	N	lumber of equences
А	н	uman	any	/	4 (HA)	H3N2	an	y yes	no	any	any	details	Wellington		72
2 r	nucleoti	le sequer	nces aft	er collap	sing (72 t	otal)						Hold ctrl or to select/des	shift key while click select multiple column	ing on co is for seq	olumn headers juential sorting
	Accessio	n Length	Host	Segment	t Subtype	Country	Region	Date	Virus n	ame	Mutations	Age Gende	er Lineage Va	icStr	Complete
	CY113349	1732	Human	4 (HA)	H3N2	New Zealand	S	1985	Influenz (A/Welli	za A virus ington/4/1985(H3N2))					с
	CY11348	1732	Human	4 (HA)	H3N2	New Zealand	S	1989	Influenz (A/Welli	za A virus ington/5/1989(H3N2))					с
	CY11354	1731	Human	4 (HA)	H3N2	New Zealand	s	1990	Influenz (A/Welli	za A virus ington/3/1990(H3N2))					с
	KM82128	1739	Human	4 (HA)	H3N2	New Zealand	s	1993	Influenz (A/Welli	a A virus ington/25/1993(H3N2))				с
	KM82129	1739	Human	4 (HA)	H3N2	New Zealand	s	1993	Influenz (A/Welli	za A virus ington/96/1993(H3N2))				с
	CY11268	1731	Human	4 (HA)	H3N2	New Zealand	S	1993	Influenz (A/Welli	za A virus ington/59/1993(H3N2))				с
/	KM82128	1738	Human	4 (HA)	H3N2	New Zealand	s	1994	Influenz (A/Welli	a A virus ington/1/1994(H3N2))					с
	KM82129	1733	Human	4 (HA)	H3N2	New Zealand	S	1996	Influenz (A/Welli	za A virus ington/48/1996(H3N2))				с
	CY01340	1717	Human	4 (HA)	H3N2	New Zealand	S	2000	Influenz (A/Welli	a A virus ington/28/2000(H3N2))	2 Y M			с
2	CY01389	1716	Human	4 (HA)	H3N2	New	s	2000	Influenz	a A virus		7Y F			с

Figure 3. List of queried sequences

4

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 <u>http://www.megasoftware.net/</u>. 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), 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.

			XI	MEGA 7.0.26	(7180411-i	386)				
Eile Analysis Help										
≣ • <mark>™</mark> Align • Daţa	s ™ Models	• Listance •	<u>₹</u> Di⊻ersity •	ୟୁ Phylogeny	ট User Tree `	Ancestors	₹ Selection	₩ <u>R</u> ates	 ⊆locks	† Diagnose
TA Close Data										
¢b.	đ			a	۵		y	6		
Help Docs	Examples	Citation	Report a <u>B</u> ug	Updates	2 MEGAL	inks T	ojbar F	references	•	
MEGA release #71	80411-i386									tasman.fa

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:

O O M7: Sequence Data Explorer																
Data Display Search Groups Highlight Statistics Help																
🖬 🚠 🧏 🖓 👽 Tt 🚾 C V Pi S # 2 4 Special - 🎎 🛤																
✓Name				_					_		_		_			•
I. Icl CY113349 Influenza A virus (AWVellington/4/1985(H3N2)) hemagglutinin (HA) gene complete cds		-	Т	A J	ΑT	т	С	ΤĂ	λT	Т	А	A (сс	А	ΤI	3
		-	Т	A /	ΑT	т	С	ΤA	ΑT	т	Α	A (сс	А	ΤΙ	3
✓3. Icl CY113541 Influenza A virus (AWVellington/3/1990(H3N2)) hemagglutinin (HA) gene complete cds		-	Т	A J	ΑT	т	С	ΤA	λT	т	A	A (сс	А	Τ	3
		-	-	-		-	-		T	т	A	A (сс	А	Τ	3
€ 5. Icl KM821290 Influenza A virus (AVVellington/96/1993(H3N2)) segment 4 hemagglutinin (HA) gene complete cds		-	-	-		-	-		- T	т	А	A (сс	А	ΤΙ	3
✓ 6. Icl(CY112685 Influenza A virus (A/Wellington/59/1993(H3N2)) hemagglutinin (HA) gene complete cds		-	Т	A,	ΑT	т	С	ΤA	λT	Т	A	A (сс	А	Τ	3
7. Icl KM821286 Influenza A virus (AVVellington/1/1994(H3N2)) segment 4 hemagglutinin (HA) gene complete cds	· ·	-	-	-		-	-		-	Т	А	A (сс	А	Τ	3
8. Icl[KM821297 Influenza A virus (AVVellington/48/1996(H3N2)) segment 4 hemagglutinin (HA) gene complete cds		-	-	-		-	-			-	-			Α	Τι	3
✓9. Icl(CY013405 Influenza A virus (AWellington/28/2000(H3N2)) segment 4 complete sequence		-	-	-		-	-	- 4	ΑT	т	Α	A (сс	Α	ΤI	3
✓10. Icl/CY013895 Influenza A virus (A/Wellington/26/2000(H3N2)) segment 4 complete sequence		-	-	-		-	-			-	-	Α (сс	Α	τı	3
11. Icl/CY011020 Influenza A virus (A/Wellington/9/2000(H3N2)) segment 4 complete sequence		-	-			-	-	- 4	ΑT	т	А	Α (сс	Α	τı	3
✓12. Icl/CY019989 Influenza A virus (A/Wellington/21/2000(H3N2)) segment 4 complete sequence			-		- Т	т.	С	ΤĂ	ΑT	т	A	Α (сс	Α	τı	3
13. Icl/CY013903 Influenza A virus (A/Wellington/30/2000(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	ΤĂ	ΑT	т	A	A (сс	Α	ΤJ	3
214. Icl/CY011616 Influenza A virus (A/Wellington/22/2001(H3N2)) segment 4 complete sequence		-	-	-		-	-	ΤA	ΑT	т	А	A (сс	Α	ΤI	3
✓ 15. Icl/CY012312 Influenza A virus (A/Wellington/34/2001(H3N2)) segment 4 complete sequence		-	-	-		-	-	ΤĂ	ΑT	Т	A	Α (сс	А	Τ	3
✓ 16. Icl/CY011672 Influenza A virus (A/Wellington/83/2002(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	ΤĂ	ΑT	т	А	A (сс	А	Τ	3
✓ 17. Icl/CY011680 Influenza A virus (A/Wellington/84/2002(H3N2)) segment 4 complete sequence		-	-	-		-	-	ΤĂ	ΑT	т	A	Α (сс	А	Τ	3
✓ 18. IcI(CY011992 Influenza A virus (A/Wellington/70/2002(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-			-	-	- (сс	А	Τ	3
III. ICI/CY012336 Influenza A virus (A/Wellington/63/2002(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	ΤA	ΑT	т	А	A (сс	А	Τι	3
✓ 20. Icl(CY013080 Influenza A virus (A/Wellington/9/2002(H3N2)) segment 4 complete sequence		-	-	-		-	-			-	-	-		А	Τı	3
	· ·	-	-	-		-	-		-	-	А	A (сс	А	Τı	3
✓ 22. Icl/CY012328 Influenza A virus (A/Wellington/7/2002(H3N2)) segment 4 complete sequence		-	-	-		-	-		- T	т	А	A (сс	А	Τι	3
✓23. Icl/CY011632 Influenza A virus (A/Wellington/38/2002(H3N2)) segment 4 complete sequence		-	-	-		-	-	- 4	λT	Т	A	A (сс	А	Τı	3
24. Icl/CY011640 Influenza A virus (A/Wellington/66/2002(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	ΤA	λT	т	А	A (сс	А	Τι	3
✓25. Icl/CY012016 Influenza A virus (A/Wellington/71/2002(H3N2)) segment 4 complete sequence		-	-	-		-	-		- т	Т	А	A (сс	А	Τı	3
✓26. Icl(CY012032 Influenza A virus (A/Wellington/80/2002(H3N2)) segment 4 complete sequence		-	-	-		-	-	- 4	λ T	Т	А	A (сс	А	Τı	ا د
27. Icl/CY012040 Influenza A virus (A/Wellington/79/2002(H3N2)) segment 4 complete sequence		-	-	-		-	-	ΤA	A Τ	Т	А	A (сс	А	Τı	3
28. Icl/CY011688 Influenza A virus (A/Wellington/86/2002(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	ΤA	ΑT	т	А	A (сс	Α	Τι	3
✓29. IcI(CY011696 Influenza A virus (A/Wellington/3/2003(H3N2)) segment 4 complete sequence		-	-	-		-	-	ΤĂ	ΑT	Т	А	Α (сс	А	Τ	3
✓ 30. IcI(CY011704 Influenza A virus (A/Wellington/4/2003(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	ΤA	ΑT	Т	А	A (сс	A	Τι	ا د
31. Icl/CY011720 Influenza A virus (A/Wellington/10/2003(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	ΤA	ΑT	т	А	A (сс	Α	Τι	3
✓ 32. Icl(CY012648 Influenza A virus (A/Wellington/9/2003(H3N2)) segment 4 complete sequence		-	-	-		-	-			-	А	Α (сс	А	Τı	3
✓ 33. Icl/CY011736 Influenza A virus (A/Wellington/25/2003(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-			-	А	Α (сс	А	Τ	3
✓ 34. Icl(CY011744 Influenza A virus (A/Wellington/28/2003(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-		-	-	A	Α (СС	А	Τ	3
✓ 35. IcI(CY012064 Influenza A virus (A/Wellington/34/2003(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-	- 4	ΑT	т	А	Α (сс	А	Τ	3
✓ 36. Icl/CY012856 Influenza A virus (A/Wellington/47/2003(H3N2)) segment 4 complete sequence	· ·	-	-	-		-	-			-	А	A (СС	А	Τ	3
		-						• •	- т	Т	A	A (сс	A	т	} ▼
1/1750 Highlighted: None Data																

6

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.

File Display Caption Help Image: Applie (A,B) Image: Applie (A,B) <th></th> <th colspan="10">M7: Between Group Mean Distance (C:\users\ucbtdje\Desktop\wellington.mdsx)</th>		M7: Between Group Mean Distance (C:\users\ucbtdje\Desktop\wellington.mdsx)											
Image: A,B Q Q Q Image: A	Eile [Elle Display Caption Help											
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$		L (A	,B)	.0 0.00			. csv	MELA	TXT	Capt			
1.1985 <t< td=""><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></t<>		1	2	3	4	5	6	7	8	9	10	11	12
2.1989 0.019	1.1985												
3. 1990 0.022 0.011	2, 1989	0.019											
4. 1993 0.032 0.021 0.016 -	3, 1990	0.022	0.011	0.017									
3. 1997 0.033 0.023 0.016 0.007 <	4, 1993	0.032	0.021	0.016	0.007								
0.11222 0.001a 0.002a 0.002a 0.002a 0.002a 0.002a 7. 2000 0.060 0.050 0.045 0.038 0.040 0.029 <t< td=""><td>6 1999</td><td>0.033</td><td>0.023</td><td>0.018</td><td>0.007</td><td>0.020</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></t<>	6 1999	0.033	0.023	0.018	0.007	0.020							
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 0.042 0.012	7. 2000	0.060	0.052	0.020	0.038	0.040	0.029						
9. 2002 0.064 0.053 0.049 0.042 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	8. 2001	0.061	0.051	0.046	0.039	0.041	0.031	0.016					
10.2003 0.067 0.069 0.064 0.056 0.046 0.027 0.033 0.035	9. 2002	0.064	0.053	0.049	0.042	0.044	0.034	0.020	0.015				
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	10. 200	3 0.076	0.069	0.064	0.056	0.058	0.046	0.027	0.033	0.035			
<u>12. 2005</u> 0.079 0.073 0.069 0.060 0.063 0.052 0.035 0.040 0.042 0.016 0.008	11.200	4 0.077	0.071	0.067	0.058	0.061	0.049	0.031	0.037	0.039	0.012		
	12, 200	5 0.079	0.073	0.069	0.060	0.063	0.052	0.035	0.040	0.042	0.016	0.008	
		[1,1] (1985-1985) / Nucleotide: Kimura 2-parameter											

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.

	M7: Distance Write-out Options
Output Format Unformatted	i text
Information to <u>W</u> rite	Decimal <u>P</u> laces 3
	Column
	✓ <u>P</u> rint/Save Matrix
	X Cancel ? Help

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.