

Biotechnology Solutions to Infectious Disease

Pathogen Sequences and Phylogenetic Analyses

The following exercises will allow you to understand how pathogen genetics is applied to epidemiologic investigation and studies.

You will need to download the input files for this workshop from Canvas.

Alternatively, you can download it from GitHub https://github.com/fredjaya/pathseq_phylo. GitHub is a version control platform used by many bioinformaticians and developers to track projects and collaborate with others.

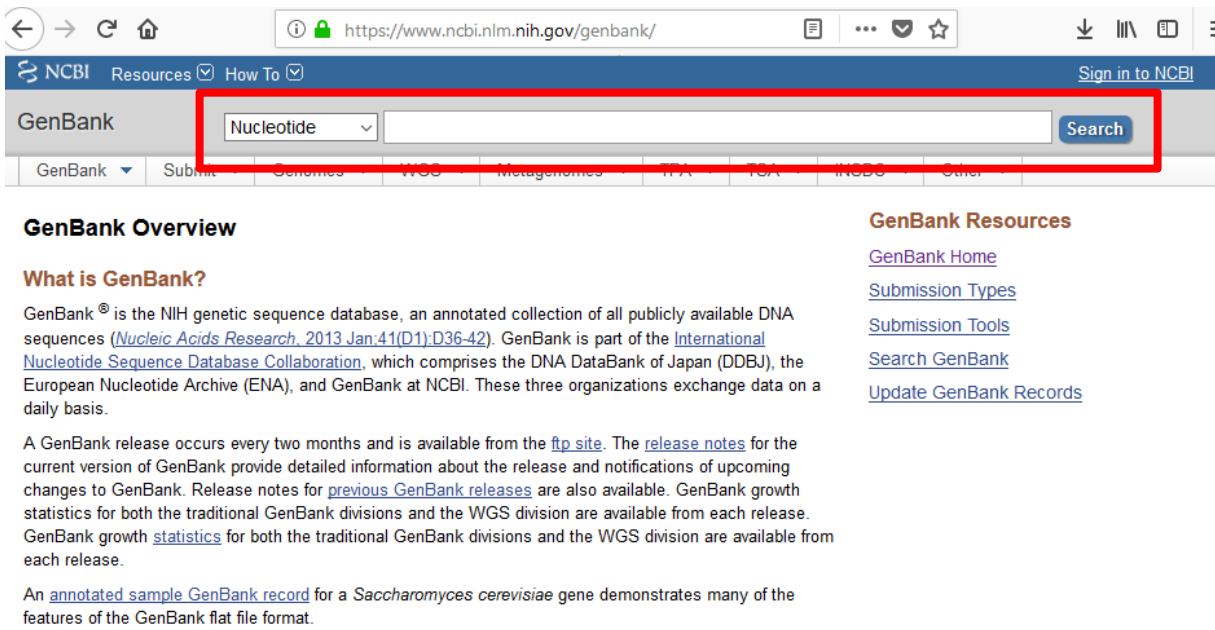
Exercise 1. NCBI GenBank

GenBank is a public repository of genetic sequences, administrated by the National Institutes of Health (NIH, USA). The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information.

GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Here, we can find genetic sequences of any organisms that have been submitted by groups that do genetic sequences for different purposes.

Go to <https://www.ncbi.nlm.nih.gov/genbank/>



The screenshot shows the NCBI GenBank homepage. A red rectangle highlights the search interface, which includes a dropdown menu currently set to 'Nucleotide' and a 'Search' button. Below the search bar, the 'GenBank Overview' section is visible, containing text about the database and links to resources. The 'GenBank Resources' section on the right lists links such as 'GenBank Home', 'Submission Types', 'Submission Tools', 'Search GenBank', and 'Update GenBank Records'.

Let's look at a bacterial whole genome sequence.

Search for the following accession number:

LAPF01000001

This is a whole genome sequence for a *Salmonella enterica* subsp. *Typhimurium* strain ABBSB1189-1 from Cooke et al., (2008).

Now let's look at a viral whole genome sequence. Note that some viruses have segmented genomes, while some others have one large open reading frame which code for different proteins. We will look at an influenza genome.

Search in Genbank for:

A/California/04/2009(H1N1)

This was one of the first sequenced isolates from the 2009 H1N1 'swine flu' pandemic. Note that for segmented viruses, each of the 'genes' has its own accession number.

Open the genetic information for:

A/California/04/2009(H1N1) HA gene (segment 4)

Organisms can be searched by their sequence name or specific accession number!

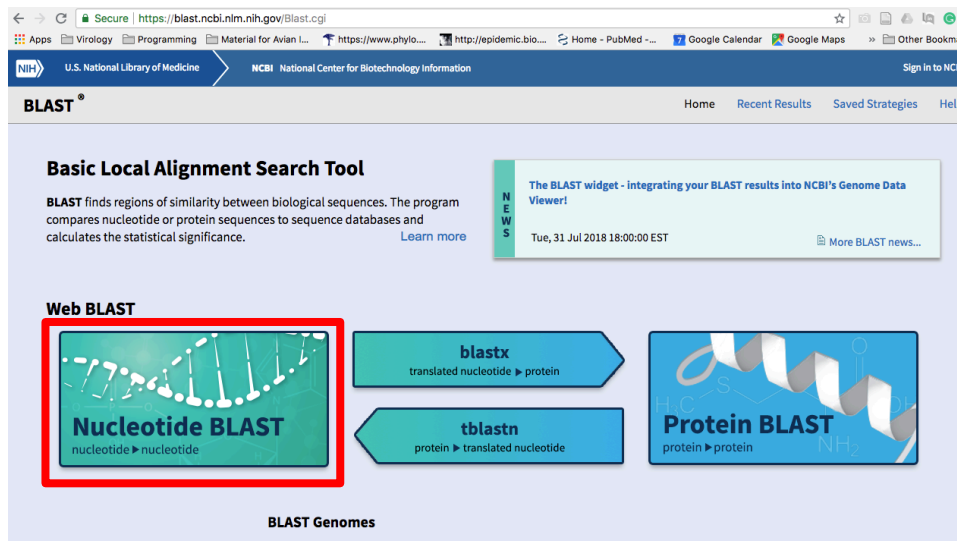
There are also specific tools to analyse and display sequence data from pathogens viruses and bacteria developed by NCBI.

Exercise 2. BLAST (Basic Local Alignment Search Tool)

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST is one most of the widely used tools in genomics. It helps identify your sequence and which is the closest reference by comparing it to all available sequenced organisms, so it's not just limited to microbes. Let's see how it works.

Go to BLAST: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Select Nucleotide Blast



Open the file (in a plain text editor):

ex_2_unknown_sequence.fasta

FASTA format files are one of the most common formats in which phylogenetic sequences are stored. It is a plain text file in which each sequence has to be specified by its name in the first line preceded by a ">" character and the sequence specified in the following line.

Copy the sequence (text) and paste it in the box that says:

'Enter accession number(s), gi(s), or FASTA sequence(s)'

We will not tell the program anything about this sequence (organism, database etc.). So, we will let BLAST sort it out and search for what this sequence is.

Click BLAST

BLAST® » blastn suite
Home
Rece

Standard Nucleotide BLAST

blastn
blastp
blastx
tblastn
tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

>Unknown
CTTTTGTGTGCGAATAACTATGAGGAAGATTATAATTTCTCTCACTGAAATCTATATCG
GAATTAA
ATTGAAATGTACTGTAATCACACCTGGTTTGTTCAGAGCCACATCACAAGATAGAGAA
CAACCTAG
GTCTCCGAAGGGAGCAAGGCATCAGTGTGCCAGTGAATCCCTTGTCAACATCTAGGT

Query subrange

From

To

or, upload me

Browse...
No file selected.

Job Title

Unknown

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

We are beta testing a New Results page

☐ Click here if you would like to see your results in the new format. You can always switch back to the Traditional Results page.

Choose Search Set

Database

☐ Human genomic + transcript
☐ Mouse genomic + transcript
☒ Others (nr etc.):

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id—completions will be suggested

☐ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude
Optional

☐ Models (XM/XP)
☐ Uncultured/environmental sample sequences

Limit to
Optional

☐ Sequences from type material

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

☒ Highly similar sequences (megablast)
☐ More dissimilar sequences (discontiguous megablast)
☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar seq

☐ Show results in a new window

What did you find? Let's discuss the output.

Exercise 3. Cholera outbreak

In this step we will use pathogen genetic data to explore a Cholera outbreak (*Vibrio cholerae*). The data used for this step is described in Hendriksen et al., (2011)

Background

On 12 January 2010, a 7.0 MW earthquake hit Haiti. By 24 January, at least 52 aftershocks had been reported, and an estimated 316,000 people had died, 300,000 were injured and more than one million were homeless. This disaster destroyed the already fragile infrastructure and required international assistance in the form of food, water, and aid workers. Volunteers and help from all over the world arrived quickly.

On 21 October 2010, the Haitian public health authorities confirmed a cholera outbreak. By 7 July 2011, 386,429 cases, including 5,885 deaths have been reported. The outbreak has also spread to the neighboring Dominican Republic and to Florida and the United States where sporadic cases have been observed.

In the early days of the outbreak, rumors spread that the disease was brought to Haiti by a battalion of Nepalese soldiers serving as United Nations peacekeepers. Though not proven definitively, the putative link to United Nations peacekeepers from Nepal gained global media attention.

Cholera occurs in sporadic cases and outbreaks in Nepal each year. In 2010, a 1,400-case outbreak occurred in midwestern Nepal. The outbreak started around 28 July and was controlled by 13 or 14 August, just prior to the time the UN Nepalese soldiers left for Haiti.

Cholera isolated from samples collected in Nepal, and Haiti were sequenced. The closest publicly available reference sequences were also retrieved. From the whole genome sequence 742 nucleotide positions with diversity among sequences were identified and used for phylogenetic analysis.

Open the file (in a plain text editor):

cholera_haiti.fasta

This file contains the sequences that we need to analyse and investigate the source of the cholera outbreak. We will use an online version of the phylogenetic tool 'IQ-TREE' to create a tree (Trifinopoulos et al., 2016).

Go to the following website: <http://iqtree.cibiv.univie.ac.at/>

← → ↻ 🏠 ⓘ iqtree.cibiv.univie.ac.at ... 📧 ☆

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

Server load: 4% Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) Nucl. Acids Res. 44 (W1): W232-W235. doi: 10.1093/nar/gkw256

Tree Inference | Model Selection | Analysis Results

For a quick start, take a look at the [tutorial](#) for the IQ-TREE web server.
Please visit the [IQ-TREE homepage](#) for more information or if you want to download the main software.
Data Privacy Statement: All your personal data are strictly confidential and will not be shared with any third parties. Your data will be automatically deleted after 180 days.

Input Data

Alignment file : Browse... Show example >

Use example alignment: ☐ Yes ?

Sequence type: ☒ Auto-detect ☐ DNA ☐ Protein ☐ Codon
☐ DNA->AA ☐ Binary ☐ Morphology ?

Partition file: This field is optional. Browse... Show example >

Partition type: ☒ Edge-linked ☐ Edge-unlinked ?

IQ-TREE Search Parameters

Perturbation strength: 0.5

IQ-TREE stopping rule: 100

Email (optional, to retrieve results):

SUBMIT JOB

Upload cholera_haiti.fasta and hit “Submit Job”

IQ-TREE will construct a ‘maximum likelihood’ phylogenetic tree. After the job is complete, analysis results will be printed. For now, download the output files:

Select “Download Selected Files”

Tree Inference **Model Selection** **Analysis Results**

User name or Email: guest **QUERY STATUS**

<input checked="" type="checkbox"/>	N...	Submission Time	Status
<input checked="" type="checkbox"/>	1	2019-08-05 08:24	Success

Summary **Run Log** **Full Result**

IQ-TREE 1.6.11 built Jun 6 2019

Input file name: cholera_haiti.fasta
 Type of analysis: ModelFinder + tree reconstruction + ultrafast bootstrap (1000 replicates)
 Random seed number: 304439

REFERENCES

 To cite ModelFinder please use:
 Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler, and Lars S Jermini (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods, 14:587-589.
<https://doi.org/10.1038/nmeth.4285>
 To cite IQ-TREE please use:
 Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol, 32:268-274.
<https://doi.org/10.1093/molbev/msu300>
 Since you used ultrafast bootstrap (UFBoot) please also cite:
 Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh (2017) UFBoot2: Improving the ultrafast bootstrap approximation. Mol Biol Evol, in press.
<https://doi.org/10.1093/molbev/msx281>

SEQUENCE ALIGNMENT

 Input data: 34 sequences with 752 nucleotide sites
 Number of constant sites: 0 (= 0% of all sites)
 Number of invariant (constant or ambiguous constant) sites: 0 (= 0% of all sites)
 Number of parsimony informative sites: 184
 Number of distinct site patterns: 154

ModelFinder

 Best-fit model according to BIC: K2P+ASC

List of models sorted by BIC scores:

Model	LogL	AIC	w-AIC	AICc	w-AICc	BIC	w-BIC
K2P+ASC	-3463.7458	7011.4916	- 0.0033	7016.5861	- 0.0053	7205.6466	+ 0.5337
K3P+ASC	-3461.1564	7008.3129	- 0.0162	7013.6575	- 0.0231	7207.0905	+ 0.2593
TVM+ASC	-3456.1540	7002.3079	+ 0.3260	7008.1719	+ 0.3586	7210.3311	+ 0.0513
TNe+ASC	-3462.9038	7011.8077	- 0.0028	7017.1523	- 0.0040	7210.5854	- 0.0452
TIM3e+ASC	-3459.6578	7007.3156	- 0.0267	7012.9167	- 0.0334	7210.7160	- 0.0423
TI+ASC	-3460.3149	7008.6297	- 0.0138	7014.2308	- 0.0173	7212.0301	- 0.0219
K2P+ASC+G4	-3463.7468	7013.4937	- 0.0012	7018.8383	- 0.0017	7212.2713	- 0.0194

DOWNLOAD SELECTED JOBS

After unzipping the downloaded folder, you should have the following output files:

- cholera_haiti.fasta.contree – consensus tree
- cholera_haiti.fasta.iqtree – analysis results i.e. model selection and trees
- cholera_haiti.fasta.log - detailed report of the analysis
- cholera_haiti.fasta.treefile – maximum likelihood tree

Let's view the maximum likelihood tree. We will use the program "FigTree" to view the constructed tree, located under the Start Menu.

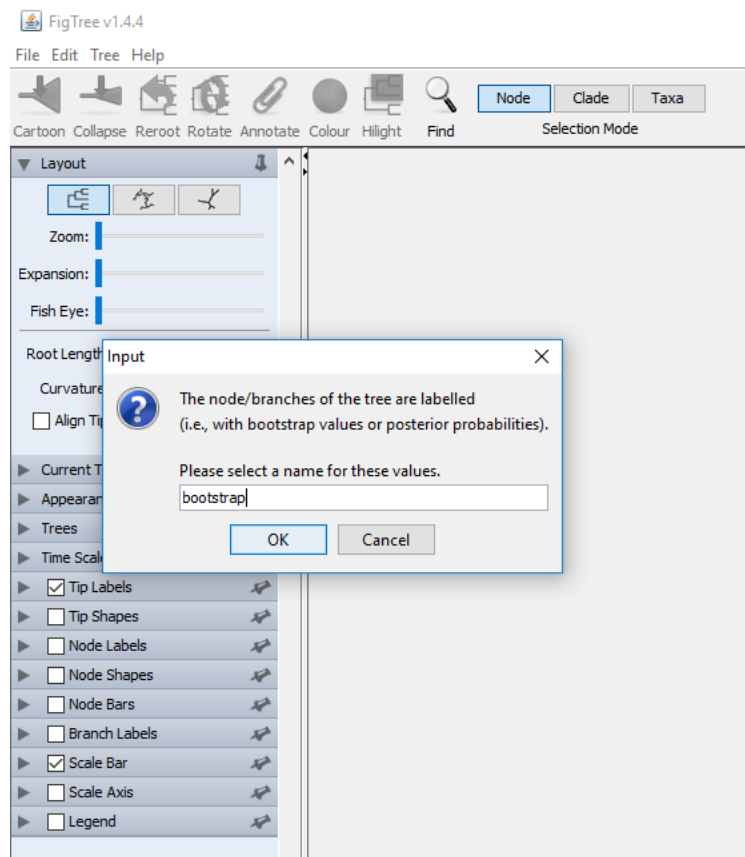
Open FigTree.exe

Then,

Open cholera_haiti.fasta.treefile

You should receive a prompt to input a name for nodes or branches.

Name the node labels as "bootstrap"

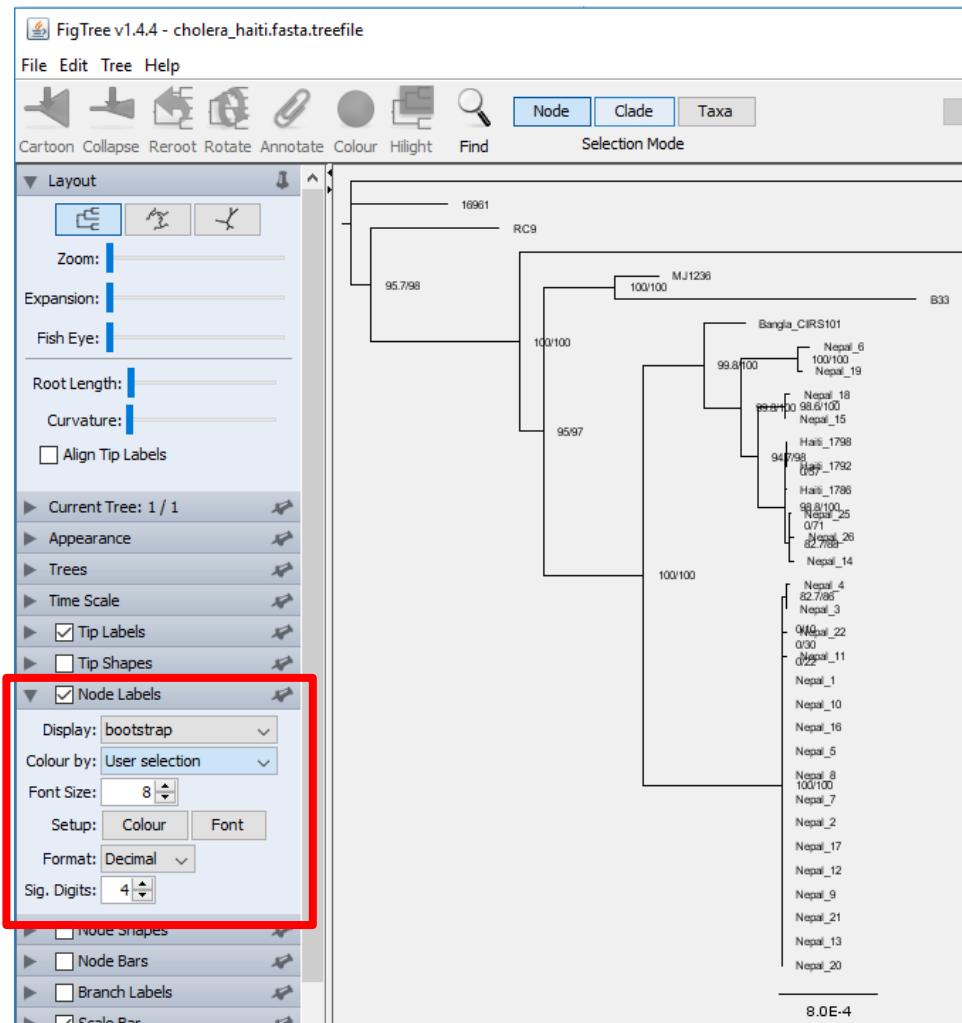


The maximum likelihood method of tree inference uses a statistical technique called the 'bootstrap'. Bootstrapping is the process of 'resampling with replacement', where trees are constructed based on the original sequence alignment with homologous sites re-ordered. For example, with 1000 bootstrap iterations, 1000 trees are constructed based on 1000 different re-arrangements of the sequence alignment. Such

statistical methods are used to measure the variability of an estimate (tree estimation) by providing a score for (un)certainty. In Bayesian tree inference, the statistic used is the posterior probability.

To show the bootstrap supports per node:

Select/tick 'node labels' and select 'bootstrap' under the 'Display' option



Try experimenting with the various sliders to modify the tree visualization! A few options to make the tree more presentable include changing the font size and colours of the tips, branch width, zoom and expansion sliders.

Let's discuss the output.

- What do the horizontal branch lengths represent?
- Vertical distances?
- What are the numbers in the nodes?
- The scale number?
- Is this a reliable tree?

What can you conclude about the genetic sequences of Cholera? Can we confirm the initial hypothesis that the virus came from UN workers from Nepal?

Exercise 4. Ebola - can we vaccinate?

Blasting is one of the ways in which we can identify the similarity of a genetic sequence to the ones available in public databases. Another way in which we can identify the relationship between our sequence and other genetic sequences, is by reconstructing a phylogeny. In this exercise, we will work with filoviruses. These are RNA viruses that cause severe viral hemorrhagic fever in humans and nonhuman primates. Ebola and Marburg are part of this family of viruses.

Filoviruses, which include the genera Ebolavirus and Marburgvirus, cause sporadic outbreaks of severe haemorrhagic disease in humans with case mortality rates between 25% and 90%. At present there are five known species of Ebolavirus: Zaire ebolavirus (Ebola virus, EBOV), Sudan ebolavirus (Sudan virus, SUDV), Taï Forest ebolavirus (Taï Forest virus, TAFV), Reston ebolavirus (Reston virus, RESTV), and Bundibugyo ebolavirus (Bundibugyo virus, BDBV). There are two known viruses in the Marburg marburgvirus species: Marburg virus (MARV) and Ravn virus (RAVV). In addition, Cuevavirus has a single species, Lloviu cuevavirus, which has been genetically isolated from bats.

Outbreaks of filovirus infection start when humans have direct contact with infected animals or with their contaminated body fluids, spreading in the human population by human-to-human transmission. Mapping models of previous outbreaks and reservoir habitats in Africa have identified a population of 22 million people who are at potential risk from Ebolavirus transmission, and 105 million people who are at potential risk from Marburgvirus transmission. The lack of specific treatment, high mortality rates, and substantial social and economic impact of the disease indicate the need for vaccines to prevent infection in a cost-effective manner. Vaccination strategies to combat filovirus infection differentiate between prophylactic vaccination and reactive use during outbreaks. Prophylactic vaccination would be beneficial to populations deemed at risk from geographical or occupational exposure and may be administered on a large scale.

The 2014–2016 West African epidemic caused by the Ebola virus was of unprecedented magnitude, duration and impact. At least 28,646 cases and 11,323 deaths¹ have been attributed to the Makona variant of Ebola virus (EBOV)2 in the two and a half years it circulated in West Africa.

Ring immunization is aimed at providing short-term protection against a specific filovirus species. This strategy preferentially involves a single immunization with antigen(s) relevant to the species circulating in the outbreak and was shown to be effective in a small-scale trial during the 2013–2016 outbreak of EBOV in West Africa. In 2017 clinical disease compatible with hemorrhagic fever disease is observed. We know there has been vaccine development for the Zaire Ebolavirus strain that we can use, if this is the case.

Open the file:

filoviruses.fasta

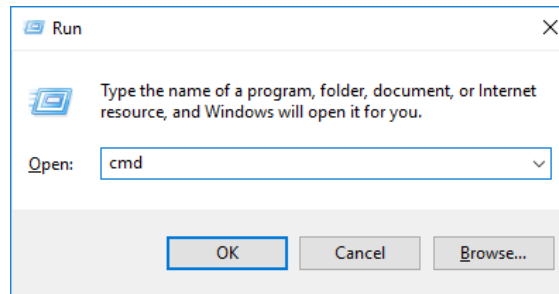
Within the sequences in this file, you will find that one of the sequences is called 'My_sample_DRCongo_2017_outbreak', which belongs to one of the recent cases that was detected from Ebola.

For this exercise, we will use the desktop/command line version of IQ-TREE.

First, open the command-prompt.

Use the shortcut windows+R to access the “run” window

Search for “cmd” and hit OK



Your command prompt should look something like:

```
C:\> Select C:\Windows\system32\cmd.exe
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\13444841>
```

“C:\Users\[ID]>” denotes your current working directory, or the folder you are in.

IQ-TREE, like many bioinformatic tools, is accessed and operated via the command-line. It simply means that we interact with the program by giving it written commands (code), rather than clicking on graphical components (buttons). The learning curve can be steeper, however, CLI tools tend to use less memory, is faster and can be used in automated scripts to analyse large amounts of data.

Some useful commands for navigating the command line:

dir	Shows the files in your current directory
cd	Changes directory
CTRL + C	Clears your line
TAB	Autocomplete

Change directories to where the IQ-TREE executable (program) is located by entering the following command and hitting enter:

cd C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin

```
C:\> C:\WINDOWS\system32\cmd.exe
Microsoft Windows [Version 10.0.17134.523]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\13444841>cd C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin
C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin>
```



To view the contents of your current directory, enter:

`dir`

A list of all the files and folders in this directory will be returned. Note, the `'.'` denotes the current directory you're in and `'..'` denotes the directory directly above. In this case, the current directory `'.'` is the `'bin'` folder, and the directory above `'..'` is the folder `'IQ-Tree'`.

```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [Version 10.0.17134.523]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\13444841>cd C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin

C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin>dir
Volume in drive C is Windows
Volume Serial Number is AA6D-97F2

Directory of C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin

06/08/2019  09:17 AM    <DIR>        .
06/08/2019  09:17 AM    <DIR>        ..
06/06/2019  10:57 PM             7,375,360 iqtrees-click.exe
06/06/2019  10:57 PM             7,375,360 iqtrees.exe
18/05/2017  06:59 PM             1,114,552 libiomp5md.dll
               3 File(s)          15,865,272 bytes
               2 Dir(s)      393,540,722,688 bytes free

C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin>
```

Let's have a look at IQ-TREE. To access and run the program, input the name of the program and hit enter:

`iqtrees.exe`

For most command-line programs, entering the program name will return instructions and example commands. It's also a good way to test that the program is working and installed correctly. Entering the program name followed by `'-h'` or `'--h'` usually returns a manual/help page or information about the program options. i.e. `'iqtrees -h'`.

We will be using example #3 for this run.

```
C:\WINDOWS\system32\cmd.exe

C:\Program Files (x86)\IQ-Tree\IQ-Tree\bin>iqtrees.exe
IQ-TREE multicore version 1.6.11 for Windows 64-bit built Jun  6 2019
Developed by Bui Quang Minh, Nguyen Lam Tung, Olga Chernomor,
Heiko Schmidt, Dominik Schrempf, Michael Woodhams.

Command-line examples (replace 'iqtrees ...' by actual path to executable):

1. Infer maximum-likelihood tree from a sequence alignment (example.phy)
   with the best-fit model automatically selected by ModelFinder:
   iqtrees -s example.phy

2. Perform ModelFinder without subsequent tree inference:
   iqtrees -s example.phy -m MF
   (use '-m TEST' to resemble jModelTest/ProtTest)

3. Combine ModelFinder, tree search, ultrafast bootstrap and SH-aLRT test:
   iqtrees -s example.phy -alrt 1000 -bb 1000
```

Enter the following command, replacing the input alignment file from the example #3 with our own:

```
iqtree.exe -s filoviruses.fasta -alrt 1000 -bb 1000
```

Note that if you input file isn't in the same directory as IQ-TREE, you will need to specify the full path. For example, "iqtree.exe -s C:\Users\13444841\Downloads\filoviruses.fasta". To save time, you can drag and drop the file into the command prompt.

When the program runs, you should have something that looks like the below picture. Real-time information about your analysis will be printed which include all the statistical tests that need to be run.

```
C:\WINDOWS\system32\cmd.exe - iqtree.exe -s C:\Users\13444841\Downloads\filoviruses.fasta -alrt 1000 -bb 1000

C:\Program Files (x86)\IQ-Tree\bin>iqtree.exe -s C:\Users\13444841\Downloads\filoviruses.fasta -alrt 1000 -bb 1000
IQ-TREE multicore version 1.6.11 for Windows 64-bit built Jun  6 2019
Developed by Bui Quang Minh, Nguyen Lam Tung, Olga Chernomor,
Heiko Schmidt, Dominik Schrempf, Michael Woodhams.

Host:      LAB0404331001 (AVX2, FMA3, 7 GB RAM)
Command:  iqtree.exe -s C:\Users\13444841\Downloads\filoviruses.fasta -alrt 1000 -bb 1000
Seed:     349074 (Using SPRNG - Scalable Parallel Random Number Generator)
Time:     Tue Aug 06 12:43:26 2019
Kernel:   AVX+FMA - 1 threads (6 CPU cores detected)

HINT: Use -nt option to specify number of threads because your CPU has 6 cores!
HINT: -nt AUTO will automatically determine the best number of threads to use.

Reading alignment file C:\Users\13444841\Downloads\filoviruses.fasta ... Fasta format detected
Alignment most likely contains DNA/RNA sequences
Alignment has 8 sequences with 21965 columns, 4460 distinct patterns
8963 parsimony-informative, 5255 singleton sites, 7747 constant sites

      Gap/Ambiguity  Composition  p-value
1  FJ750957_Marburg_virus      12.98%    failed    0.00%
2  FJ750953_Ravn_virus         12.98%    failed    0.00%
3  MH121169.1_Sudan_ebolavirus  15.01%    passed    15.73%
4  KT357840_Zaire_Ebolavirus   19.25%    passed     7.19%
5  My_sample_DRCongo_2017_outbreak 14.96%    passed    26.18%
6  KT725373_Zaire_ebolavirus   15.65%    passed    10.69%
7  KC545395_Bundibugyo_ebolavirus 14.59%    failed     0.01%
8  KU182910.1_TaiForest_ebolavirus 14.61%    failed     0.00%
**** TOTAL                    15.00%    4 sequences failed composition chi2 test (p-value<5%; df=3)

Create initial parsimony tree by phylogenetic likelihood library (PLL)... 0.005 seconds
NOTE: ModelFinder requires 28 MB RAM!
ModelFinder will test 286 DNA models (sample size: 21965) ...
No. Model      -LnL      df  AIC      AICc      BIC
1  JC          118983.413  13  237992.826  237992.843  238096.790
2  JC+I        117951.543  14  235931.087  235931.106  236043.048
3  JC+G4       117880.623  14  235789.246  235789.265  235901.207
4  JC+I+G4     117862.928  15  235755.856  235755.878  235875.814
5  JC+R2       117878.782  15  235787.564  235787.586  235907.522
```

Examine your tree in FigTree with node support values.

Differentiate the different filoviruses by colouring in the tips/clades.

Can we use the vaccine for Zaire Ebolavirus?

Exercise 5. Court case - intentional HIV infection?

Because of the rapid rate of HIV-1 evolution, phylogenetic analysis of HIV-1 DNA sequences is a powerful tool for the identification of closely related viral strains, which may be used to investigate putative transmission between individuals. In Lafayette, Louisiana, a gastroenterologist was accused of trying to kill his former lover by injecting her with HIV-infected blood from one of his patients. The former lover said that on the night of 4 August 1994, the gastroenterologist, who had been giving her vitamin shots, came to her house and gave her another injection against her wishes. In December, after the victim began having suspicious symptoms, her obstetrician tested her for HIV. The victim found out she carried the virus in January 1995, and in May of that year, she accused the gastroenterologist of deliberately infecting her. The gastroenterologist has pleaded not guilty, and his lawyers said he was at home with his wife on the night in question.

As part of their investigation, the police obtained samples of blood from the victim and from the gastroenterologist's only HIV-positive patients. They arranged to have Michael Metzker, then a graduate student in the lab of molecular biologist Richard Gibbs at Baylor College of Medicine in Houston, compare the genetic material from those two HIV strains to each other. They were also compared to viral sequences from 30 randomly chosen HIV patients in the Lafayette area and to hundreds of HIV sequences in the national database.

In this exercise, we will perform a phylogenetic analysis based on the data of this investigation and test the hypothesis of HIV transmission. Metzker et al. (2002) amplified and sequenced part of the reverse transcriptase (RT, pol) and part of the envelope gene.

Open the file: HIV.fasta

Note that the sequences named Patient1 - Patient7 are from the gastroenterologist patients with HIV. The sequences named Victim1 - Victim2 are from the victim.

Use the alignment to construct a phylogenetic tree, feel free to use either the web or desktop version from the previous exercises. Edit and visualise the output tree in FigTree.

What do you think, guilty or innocent?

References

- Cooke, F.J., Brown, D.J., Fookes, M., Pickard, D., Ivens, A., Wain, J., Roberts, M., Kingsley, R.A., Thomson, N.R., Dougan, G., 2008. Characterization of the Genomes of a Diverse Collection of *Salmonella enterica* Serovar Typhimurium Definitive Phage Type 104. *J. Bacteriol.* 190, 8155–8162. <https://doi.org/10.1128/JB.00636-08>
- Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., Bortolaia, V., Pearson, T., Waters, A.E., Upadhyay, B.P., Shrestha, S.D., Adhikari, S., Shakya, G., Keim, P.S., Aarestrup, F.M., 2011. Population Genetics of *Vibrio cholerae* from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. *mBio* 2, e00157-11. <https://doi.org/10.1128/mBio.00157-11>
- Metzker, M.L., Mindell, D.P., Liu, X.-M., Ptak, R.G., Gibbs, R.A., Hillis, D.M., 2002. Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14292–14297. <https://doi.org/10.1073/pnas.222522599>
- Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., Minh, B.Q., 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44, W232-235. <https://doi.org/10.1093/nar/gkw256>