

# Sequence Alignments and Phylogeny

Genomics and Clinical Virology

Sunando Roy

[sunando.roy@ucl.ac.uk](mailto:sunando.roy@ucl.ac.uk)

**Scenario:** There has been a recent spike in Influenza cases in your local hospital. The clinicians are worried there may be an outbreak and ward to ward transmission. They have asked you to sequence the virus from clinical samples to determine if they are related.

# Outline

- Sequence retrieval from GenBank
- Multiple Sequence Alignments
- Model Testing and Maximum Likelihood based tree building
- Viewing and modifying a Tree file

# Software

**Mafft** – Alignment tool

**MEGA** – Alignment viewer and editor

**Modeltest-ng** – Model testing

**IQ-TREE** – Tree building

**Figtree** – Tree viewer and editor

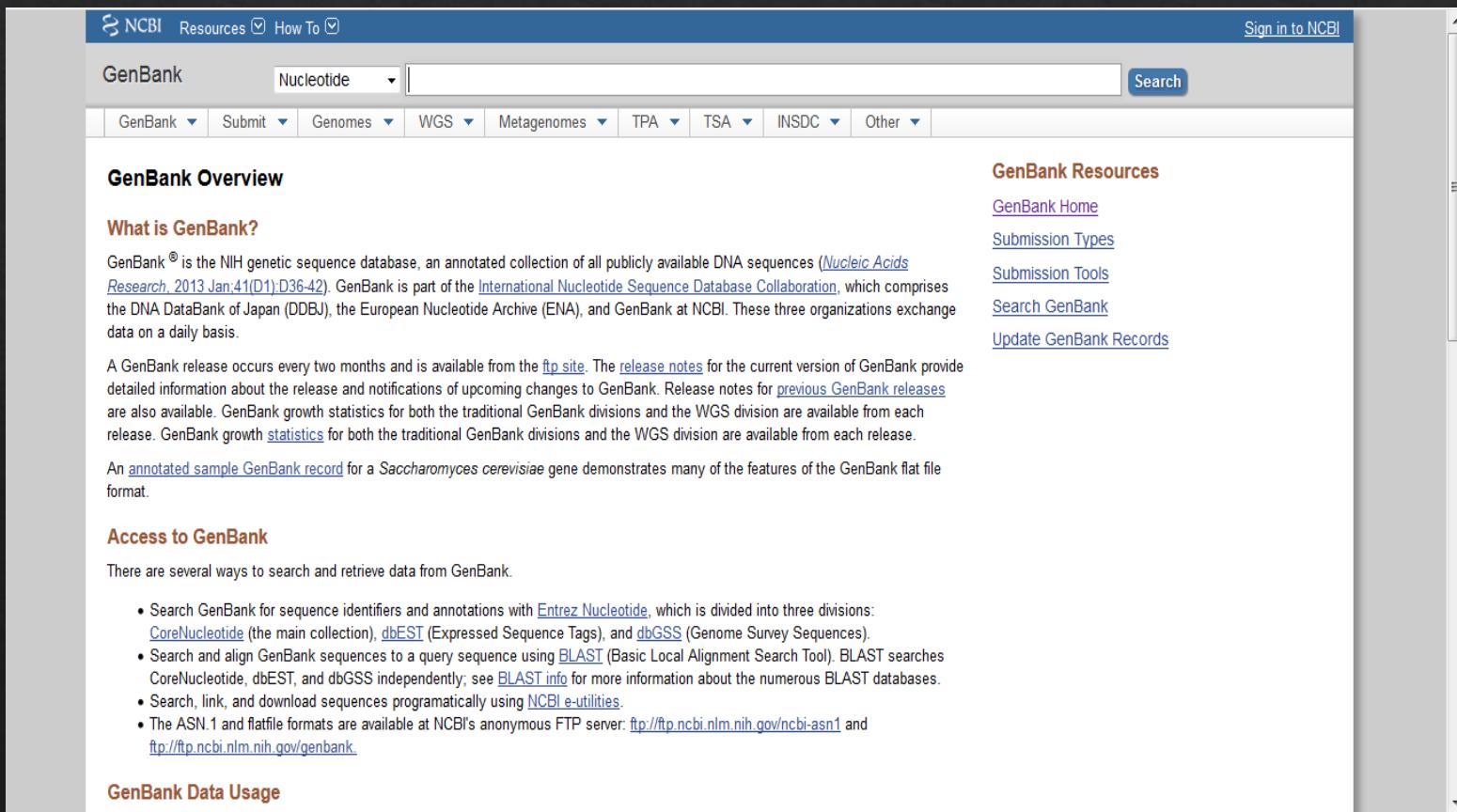
# Sequence Retrieval

- Where?
  - GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>)
  - ENA (<https://www.ebi.ac.uk/ena>)
  - DDBJ (<http://www.ddbj.nig.ac.jp/>)

For all other databases  
([https://en.wikipedia.org/wiki/List\\_of\\_biological\\_databases](https://en.wikipedia.org/wiki/List_of_biological_databases))

There are now pathogen specific databases like  
GISAID (Influenza/SARS-CoV-2) and NoroNet  
(Norovirus)

- How?
  - GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>)



## GenBank Overview

### What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan;41\(D1\):D36-42](#)). 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.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

### Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided into three divisions: [CoreNucleotide](#) (the main collection), [dbEST](#) (Expressed Sequence Tags), and [dbGSS](#) (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

### GenBank Data Usage

- Alternatives
  - Batch Entrez  
(<https://www.ncbi.nlm.nih.gov/sites/batchentrez>)

The screenshot shows the NCBI Batch Entrez web interface. At the top, there's a navigation bar with links for All Databases, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, and Books. Below the navigation bar, a sub-navigation bar shows 'Database' set to 'Nucleotide', a 'File:' dropdown with 'Browse...', and a 'Retrieve' button. The main content area is titled 'Batch Entrez' and contains the following sections:

- Batch Entrez**: A brief description stating that given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records.
- Instructions**: A numbered list of six steps:
  1. Start with a local file containing a list of accession numbers or identifiers
  2. Select the database corresponding to the type of accession numbers or identifiers in your input file
  3. Use the **Browse** or **Choose File...** button to select the input file
  4. Press the **Retrieve** button to see a list of document summaries
  5. Select a format in which to display the data for viewing, and/or saving
  6. Select 'Send to file' to save the file.
- Tips**: A bulleted list of three items:
  - To download entire genome records, check the NCBI FTP site, instead of using Batch Entrez.
  - Some lists of record identifiers can be tens of thousands of lines long, so Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems.
  - When loading large numbers of genome records, put several thousand record identifiers

- Alternatives
  - Entrez E-utilities  
(<ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/>)
  - Command – Browser example

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=AY278488, AY304486, MN908947, MT782115&rettype=fasta&retmode=text`
  - Command – Terminal example

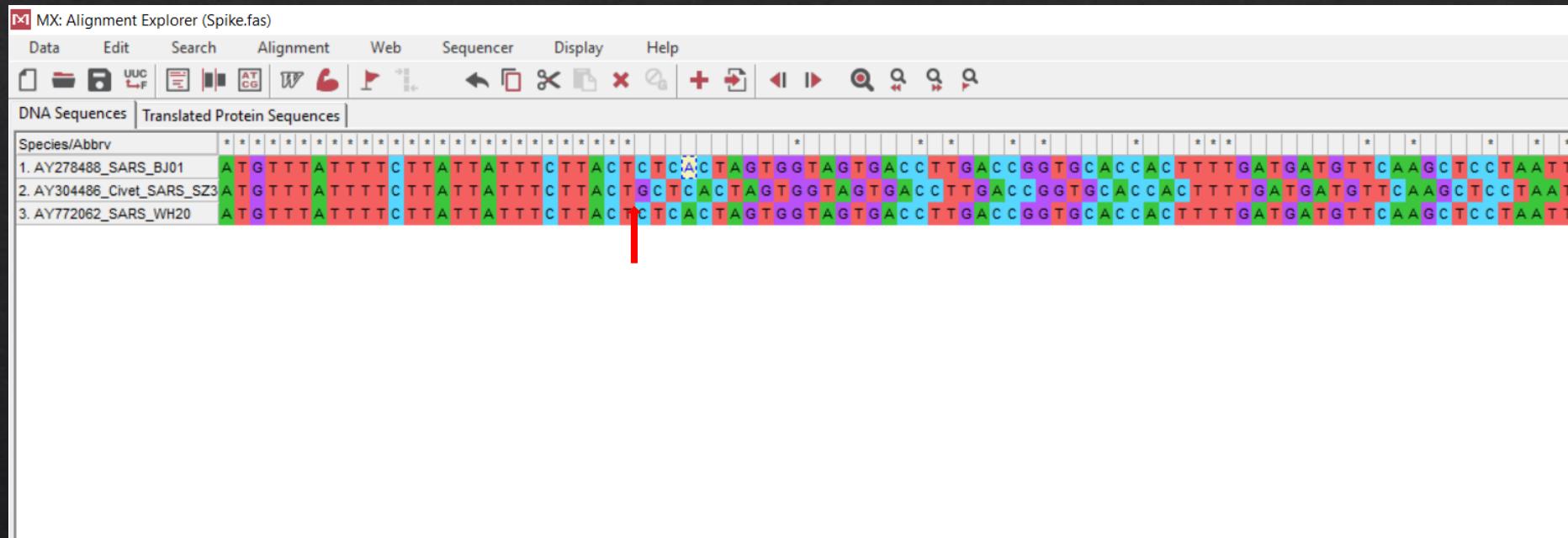
`esearch -db "protein" -query "txid11270[Organism] AND L Protein Complete AND refseq[filter]" | efetech -format fasta > outputfile.fasta`

# Dataset Selection

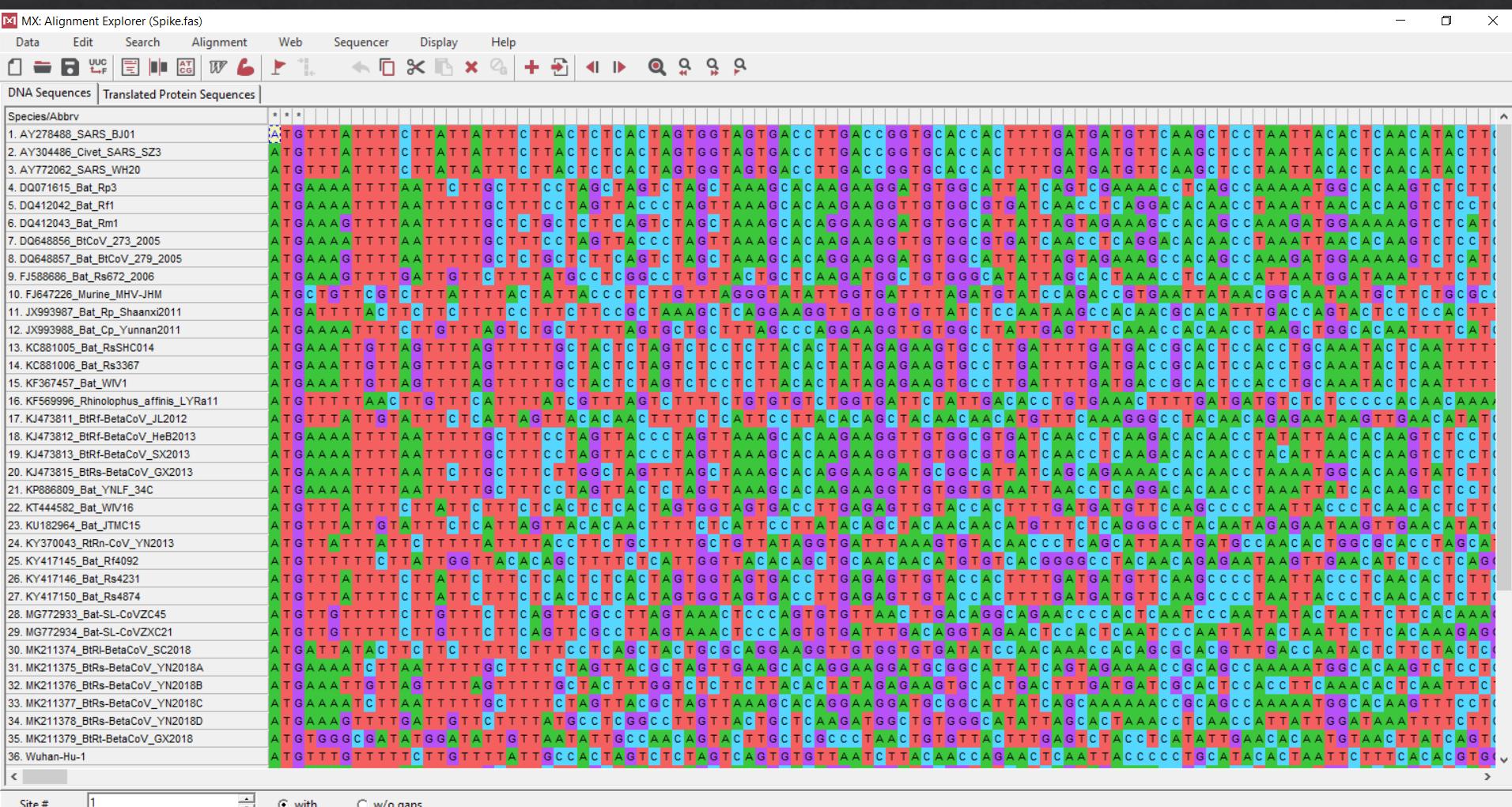
- Relatedness with query strain.
- Time/Location based filtering of strains.
- Size/Completeness of dataset to process.
- Choosing the correct outgroup.

# Multiple Sequence Alignments

- Why?
  - Necessary for every sequence analysis







- How?
  - Maximize identity between sequences in your alignment.
  - Uses scores for matches (+), mismatches (-), gap penalty (-).
  - Changing scoring parameters change alignments.
  - Visual inspection is always necessary.
  - Bad alignment = Bad phylogenetic inferences.

<https://mafft.cbrc.jp/alignment/server/>

MAFFT version 7  
Multiple alignment program for amino acid or nucleotide sequences

Download version  
[Mac OS X](#)  
[Windows](#)  
[Linux](#)  
[Source](#)

Online version  
[Alignment](#)  
[mafft --add](#)  
[Merge](#)  
[Phylogeny](#)  
[Rough tree](#)  
[Merits / limitations](#)  
[Algorithms](#)  
[Tips](#)  
[Benchmarks](#)  
[Feedback](#)

[Follow](#)



For a large number of short sequences, try [an experimental service](#) (2017/Jul).  
**This service will be unavailable for maintenance, Feb.10 6:00PM – Feb.11 (JST).**

Multiple sequence alignment and NJ / UPGMA phylogeny

**Input:**  
Paste protein or DNA sequences in fasta format. [Example](#)

or upload a plain text file:  No file selected.

Use structural alignment(s)  
 Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)

mafft Combined\_HA\_genbank.fas > outputfile.fas

# Model Testing and Maximum Likelihood based tree building

- Why?
  - To infer evolutionary relationships and identify novel pathogens.
  - Identify geographical location for the source of infection.
  - Identify potential host species.
  - Infer transmission events and outbreaks.

# Tree Building

Seq 1 – ATTGCAAT

Seq 2 – ATTGCAAT

Seq 3 – TTTGCTAT

Seq 4 – TTTGCTAT

Seq 5 – ATTCCTAC

# Tree Building

Seq 1 – ATT**GCAAT**

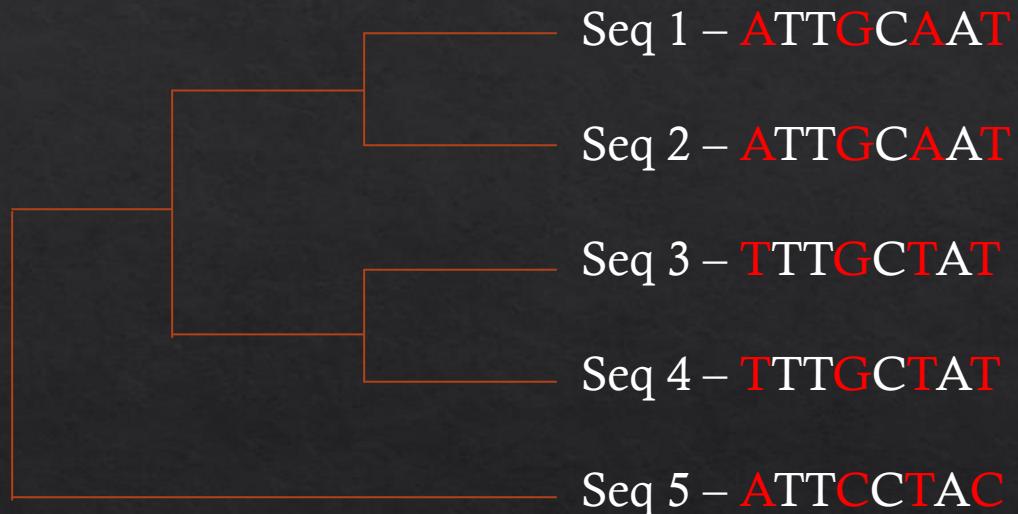
Seq 2 – ATT**GCAAT**

Seq 3 – TTT**GCTAT**

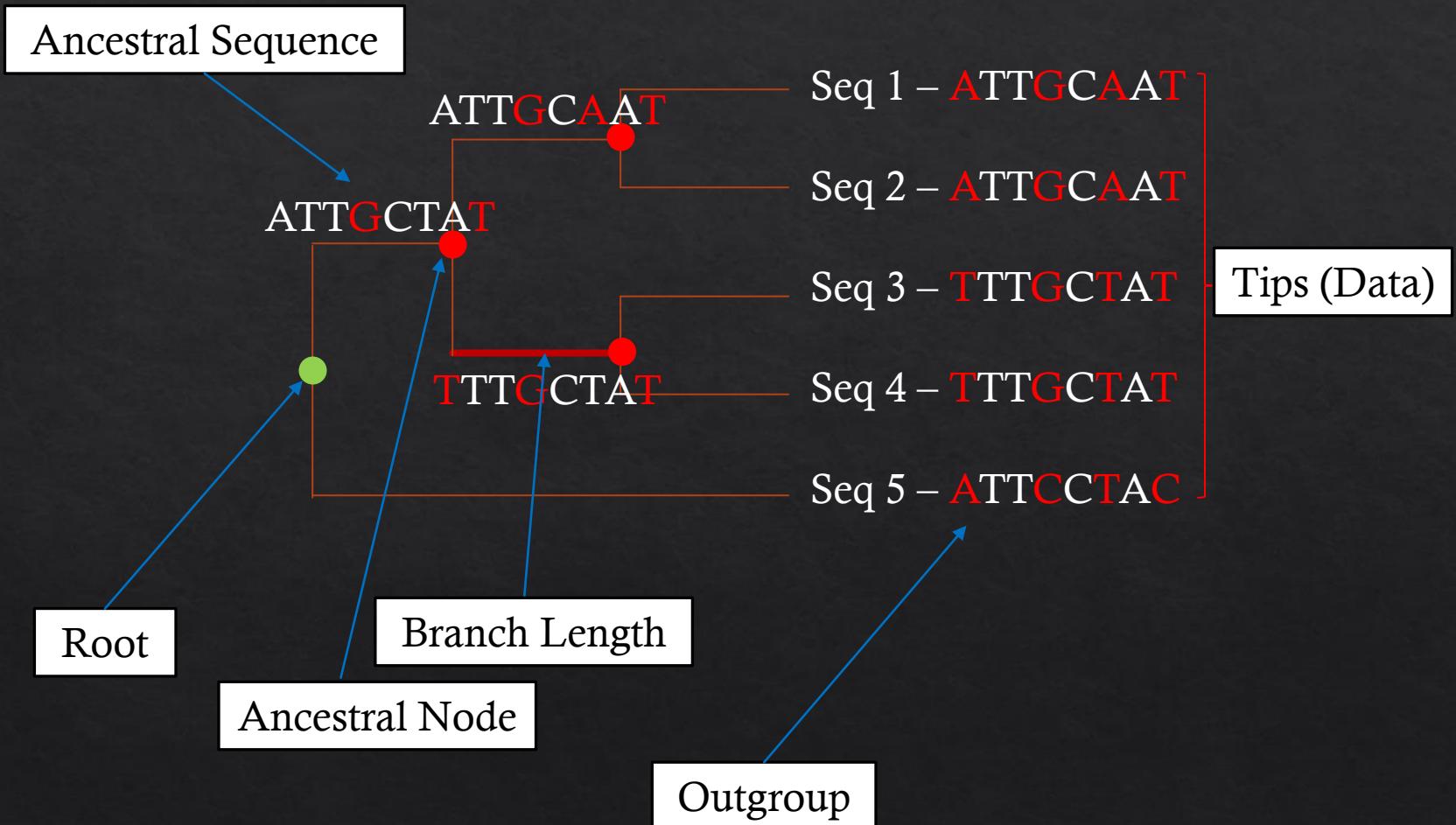
Seq 4 – TTT**GCTAT**

Seq 5 – ATT**CCTAC**

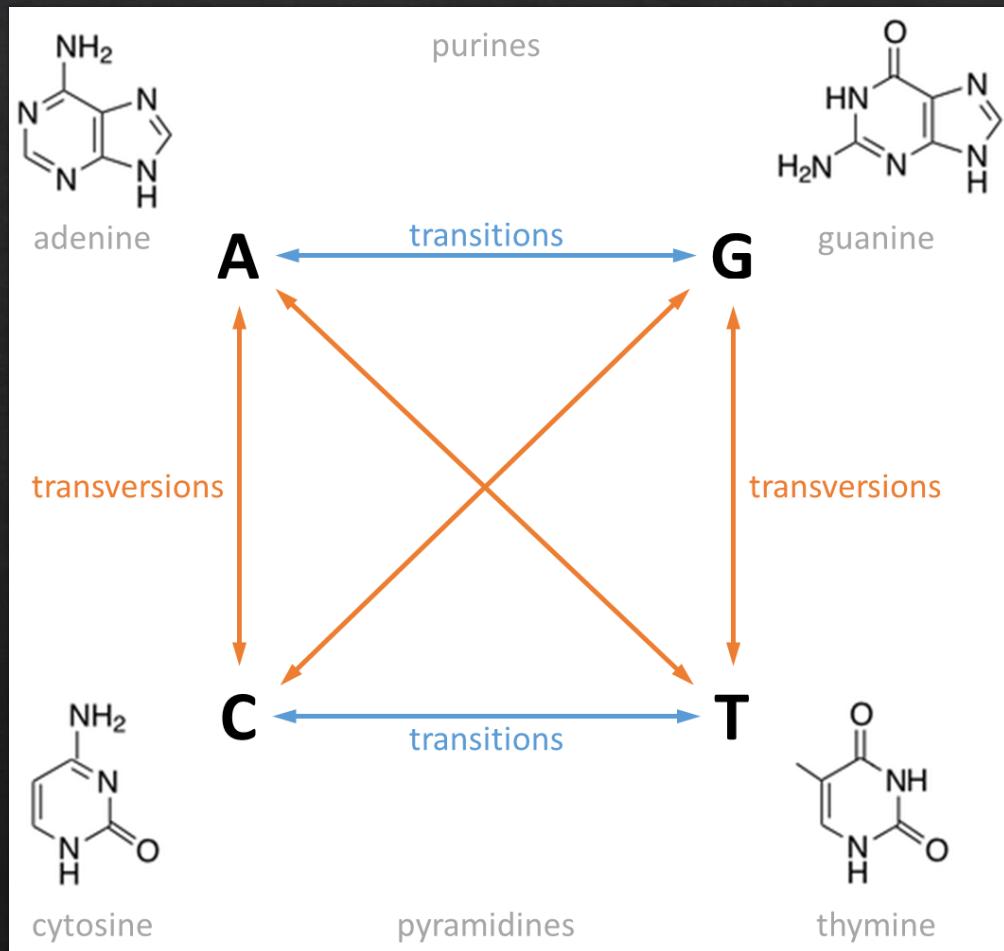
# Tree Building



# Tree Building



# Evolutionary models - DNA



Jukes Cantor

Complexity

General Time Reversible

# Evolutionary models - Proteins

	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																			C	
S	-1	4																		S	
T	-1	1	5																	T	
A	0	1	0	4																A	
G	-3	0	-2	0	6															G	
P	-3	-1	-1	-1	-2	7														P	
D	-3	0	-1	-2	-1	-1	6													D	
E	-4	0	-1	-1	-2	-1	2	5												E	
Q	-3	0	-1	-1	-2	-1	0	2	5											Q	
N	-3	1	0	-2	0	-2	1	0	0	6										N	
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8									H	
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5								R	
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5							K	
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5						M	
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4					I	
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4				L	
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4			V	
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11		W	
Y	-2	-2	-2	-3	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7	Y	
F	-2	-2	-2	-3	-4	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

- Tree Building Tools
  - Neighbour Joining – BioNJ, MEGA
  - Maximum Likelihood – PhyML, RAxML, IQTREE
  - Bayesian Inference – Mr Bayes, BEAST

# Neighbour Joining

- Estimates relationships based on a pairwise distance matrix.
- Distance matrix calculation does take into consideration evolutionary substitution models.
- Collapses closest distance pair into one taxa and repeats steps until all tips are clustered.
- Samples only one possible tree out of all possible outcomes.
- Fast but struggles in estimating relationships over longer evolutionary times.

# Maximum Likelihood

- Calculates the probability of a tree topology at each individual site across the sequence and a final product across sites is computed.
- Can use independent rate measurement across each site.
- Evaluates multiple tree topologies using branch swaps, nearest neighbour interchange etc. to find trees with the best probability.
- The probability is presented as a log likelihood thus less negative the number the greater the probability.

# Bayesian Inference

- Trees built by estimating posterior probability from a set of user defined priors.
- Used in Phylodynamics and Phylogeography analysis.
- Computationally intensive.
- Sensitive to priors and evolutionary assumptions.

# Bootstrap

- Start from a reference tree.
- Alignment sites are sampled with replacement.
- Trees are built for each resampled dataset.
- The frequency of each node occurring in the bootstrapped trees is computed.
- Gives a statistical confidence value to each node.
- Bootstrap values of 70-75 is used as an indicator for good support.

- **Caveats**
  - Model testing – Modeltest-ng
  - Recombination Detection – GARD, Simplot

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

Server load: 33%

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**Substitution Model Options**Substitution model:  Auto [?](#)FreeRate heterogeneity:  Yes [+R] [?](#)Rate heterogeneity:  Gamma [+G]  Invar. sites [+I] [?](#)#rate categories:  4 [?](#)State frequency:  Empirical (from data)  AA model (from matrix)  ML-optimized  
 Codon F1x4  Codon F3x4 [?](#)Ascertainment bias correction:  Yes [+ASC] [?](#)**Branch Support Analysis**Bootstrap analysis:  None  Ultrafast  Standard [?](#)Number of bootstrap alignments:  1000 [?](#)<http://iqtree.cibiv.univie.ac.at/>

```
iqtree -s Combined_HA_genbank_aln.fas -bb 1000 -st DNA -nt 4 -alrt 1000 -pre treefileout
```

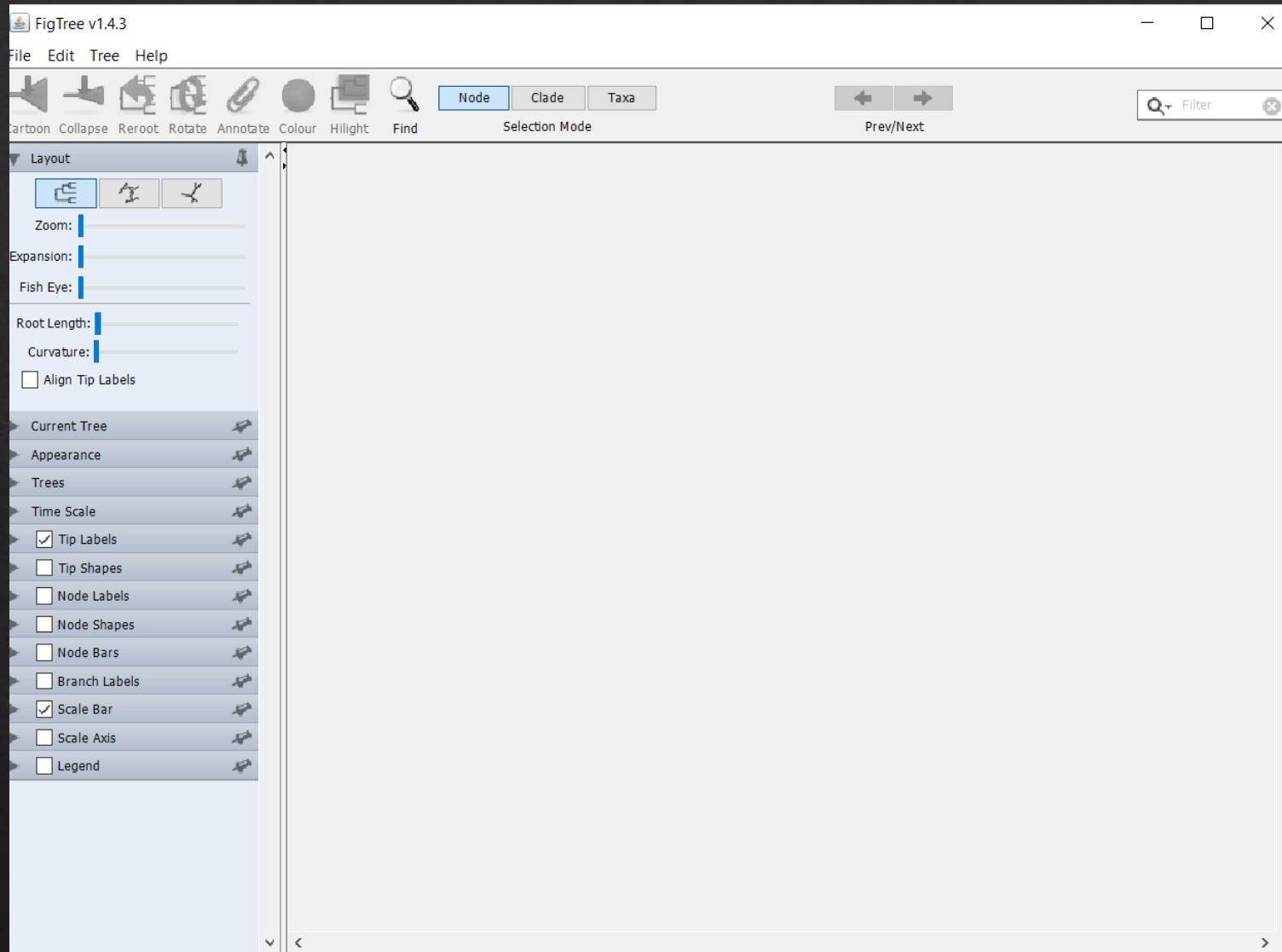
**-s** : Input File  
**-bb** : ultrafast bootstrap  
**-st** : data type  
**-nt** : Number of threads  
**-alrt** : SH-like approximate likelihood ratio test  
**-pre** : Prefix for output file

### Note:

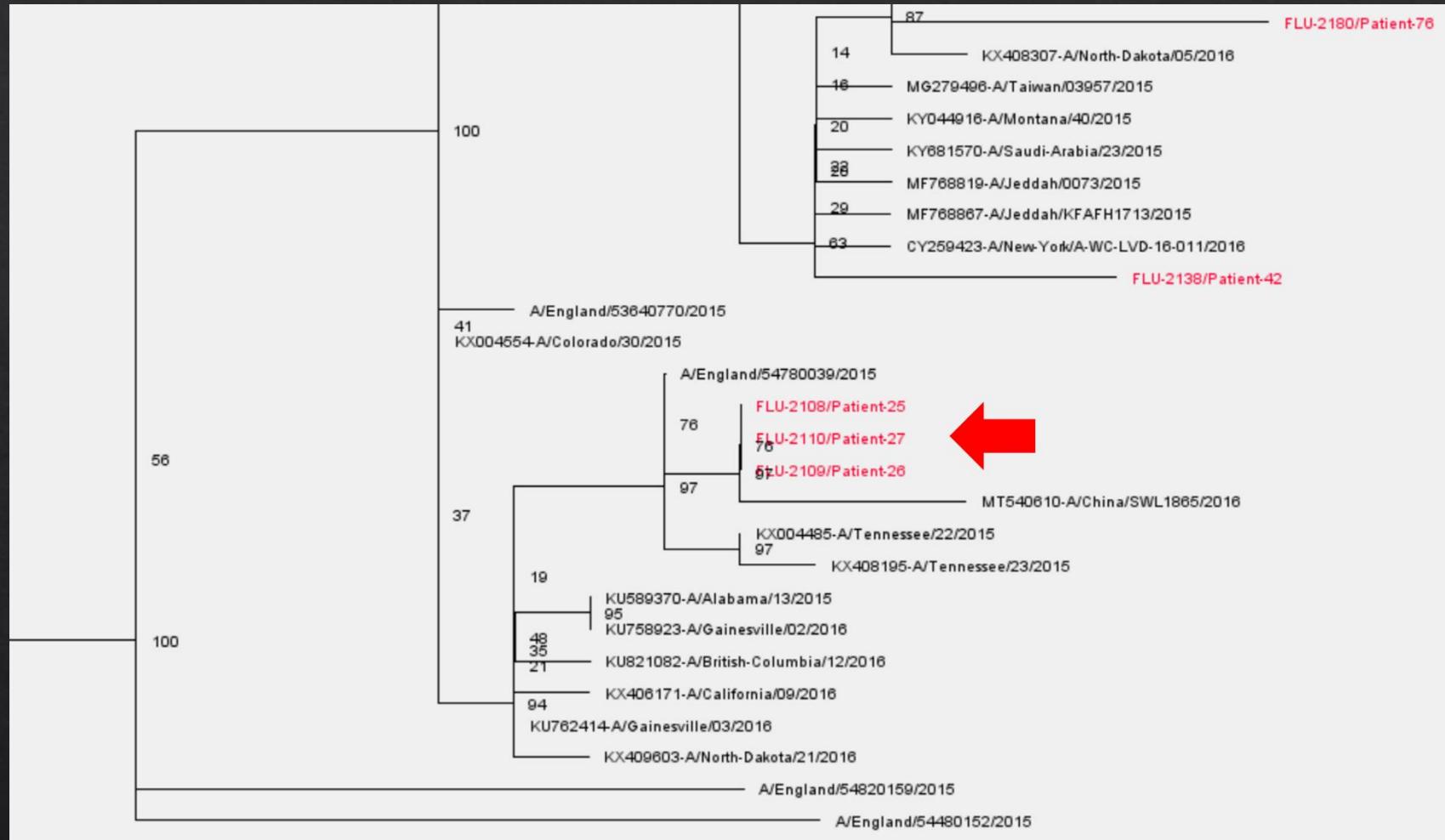
- This will run with model testing included
- Output will have .treefile that has both the UF bootstrap and alrt values and the .contree consensus tree
- For UF bootstrap values >95 and aLRT values >80 considered as strong support
- Traditional bootstrap can be done using –b option
- Models can be specified using –m option

# Viewing and modifying a Tree file

- Why?
  - To visualize final phylogenetic relationships and draw inferences.
  - To create final figures for publications.



figtree



Based on this data one would infer that there were multiple introductions into the hospital and only a small transmission cluster of three patients could be identified.

# Questions