Pathogen sequences and phylogenetic analysis

Biotechnology Solutions for Infectious Disease

frederick.r.jaya@student.uts.edu.au
github.com/fredjaya
@fredjaya1

Previous lecture

STEC and raw milk

Outbreak of STEC O157:H7 associated with contaminated salad leaves

WGS of MRSA

Objective: To understand how pathogen genetics is applied to epidemiological investigation and studies

Assessment 2

- Journal article (1500 words)
- Presenting your results on ARGs in bacteria
- **Phylogenetic tree**
    - Construction
    - Presentation

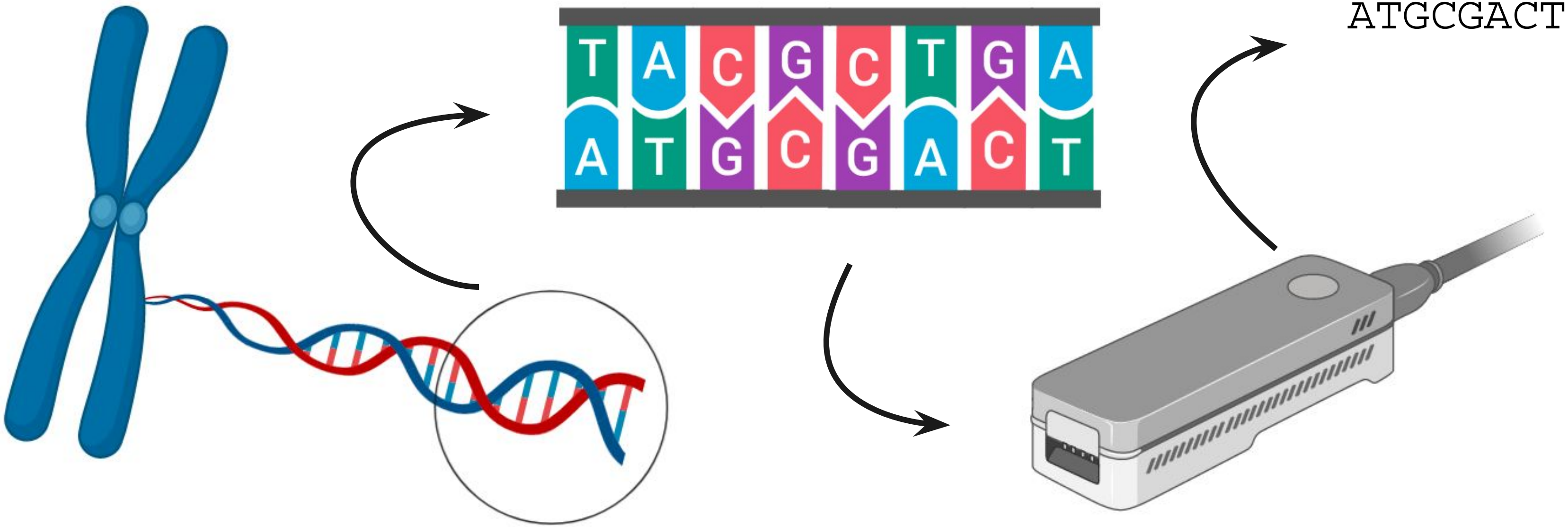
Exercises

- Finding/identifying gene sequences
- Applications of phylogeny

1. GenBank
2. BLAST
3. Haiti outbreak
4. Vaccinating ebola
5. HIV transmission

Pathogen sequences

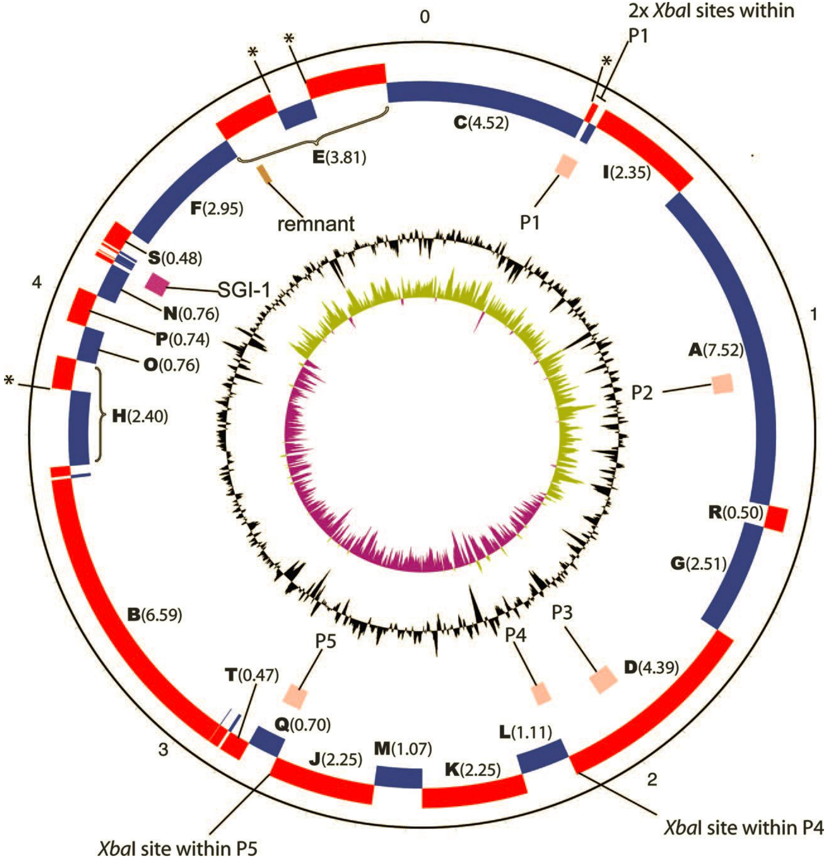
Genetic sequences



```
>seq1_F
TACGCTGA
>seq1_R
ATGCGACT
```

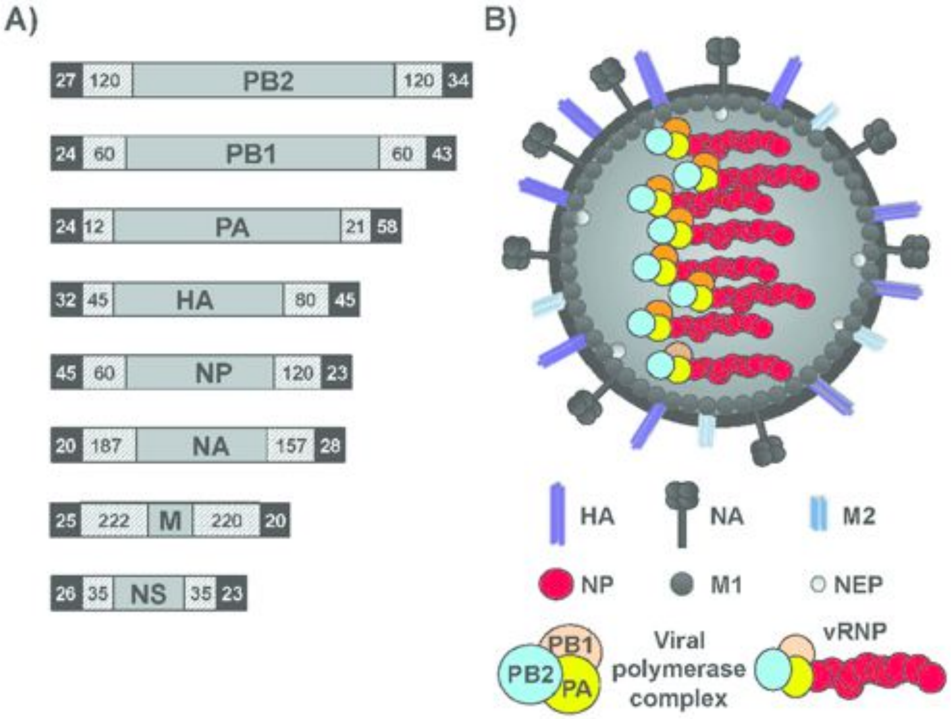
Bacterial whole genome

- Physical map of the Salmonella serovar Typhimurium NCTC113348 genome
- circular

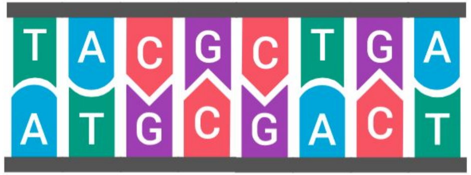


Viral whole genome

- Influenza A virus genome organisation and virion structure
- Linear, segmented

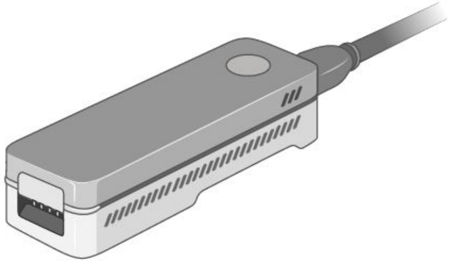


Genetic sequences



Biochemistry

DNA/RNA, Protein



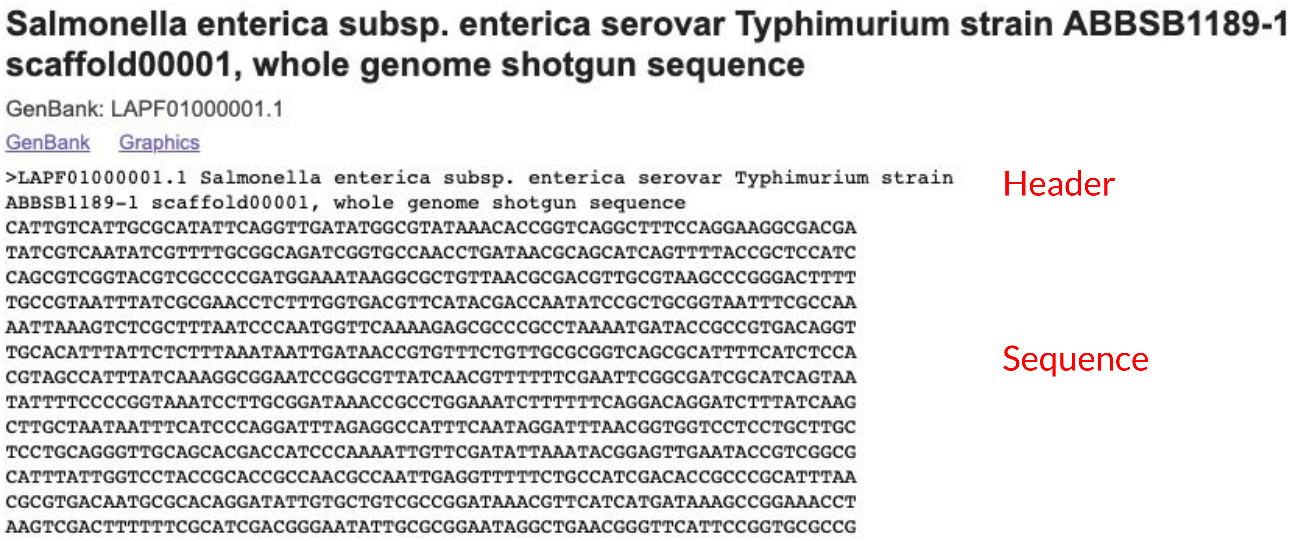
Sequencing

Whole genome, SNPs

File format

.FASTA, .FASTQ, .NEX

Genetic sequences (FASTA)



Salmonella enterica subsp. enterica serovar Typhimurium strain ABBSB1189-1 scaffold00001, whole genome shotgun sequence

GenBank: LAPF01000001.1

GenBank Graphics

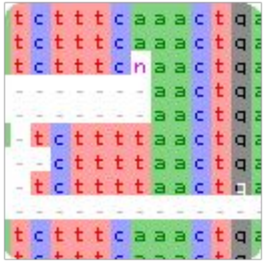
```
>LAPF01000001.1 Salmonella enterica subsp. enterica serovar Typhimurium strain
ABBSB1189-1 scaffold00001, whole genome shotgun sequence
CATTGTCATTGCGCATATTCAGGTTGATATGGCGTATAAACACCGGTCAGGCTTTCCAGGAAGGCGACGA
TATCGTCAATATCGTTTTGCGGCAGATCGGTGCCAACCTGATAACGCAGCATCAGTTTTACCGCTCCATC
CAGCGTCGGTACGTCGCCCCGATGGAAATAAGGCGCTGTTAACGCGACGTTGCGTAAGCCCGGGACTTTT
TGCCGTAATTTATCGCGAACCTCTTTGGTGACGTTCATACGACCAATATCCGCTGCGGTAATTTCGCCAA
AATTAAAGTCTCGCTTTAATCCCAATGGTTCAAAAGAGCGCCCGCCTAAAATGATACCGCCGTGACAGGT
TGCACATTTATTCTCTTTAAATAATTGATAACCGTGTTTCTGTTGCGCGGTCAGCGCATTTTCATCTCCA
CGTAGCCATTTATCAAAGGCGGAATCCGGCGTTATCAACGTTTTTTCGAATTCGGCGATCGCATCAGTAA
TATTTTCCCCGGTAAATCCTTGCGGATAAACCGCCTGGAAATCTTTTTTCAGGACAGGATCTTTATCAAG
CTTGCTAATAATTTCATCCCAGGATTTAGAGGCCATTTCAATAGGATTTAACGGTGGTCCTCCTGCTTGC
TCCTGCAGGGTTGCAGCACGACCATCCCAAAATTGTTCGATATTAAATACGGAGTTGAATACCGTCGGCG
CATTTATTGGTCCTACCGCACCGCCAACGCCAATTGAGGTTTTTCTGCCATCGACACCGCCCGCATTTAA
CGCGTGACAATGCGCACAGGATATTGTGCTGTCGCCGGATAAACGTTCATCATGATAAAGCCGGAAACCT
AAGTCGACTTTTTTCGCATCGACGGGAATATTGCGCGGAATAGGCTGAACGGGTTCATTCCGGTGCGCCG
```

Header

Sequence

Tools - Sequence alignment viewers

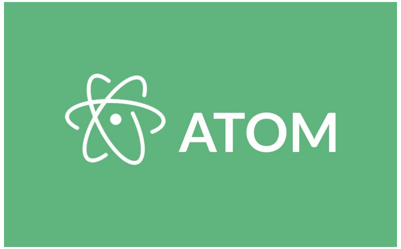
geneious



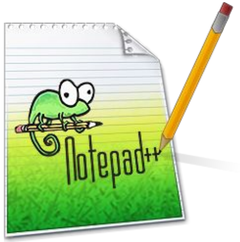
AliView

Seqotron

Tools - text editors



Sublime
Text 3



Notepad++

Notepad

Exercise 1 - NCBI GenBank

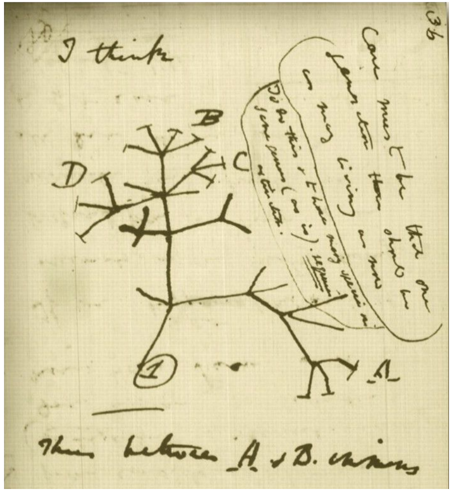
Exercise 2 - BLAST

Phylogenetics

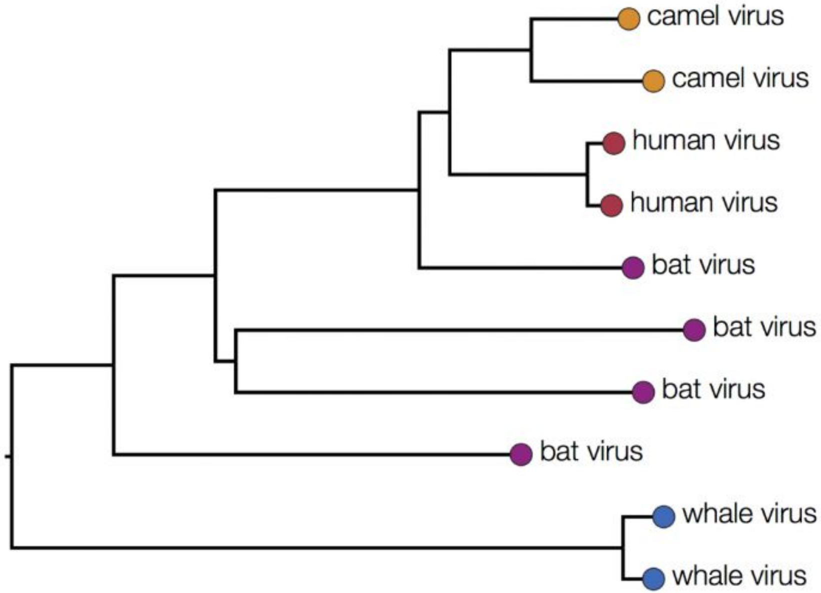
1. Relationships between species

2. Evolutionary changes and history

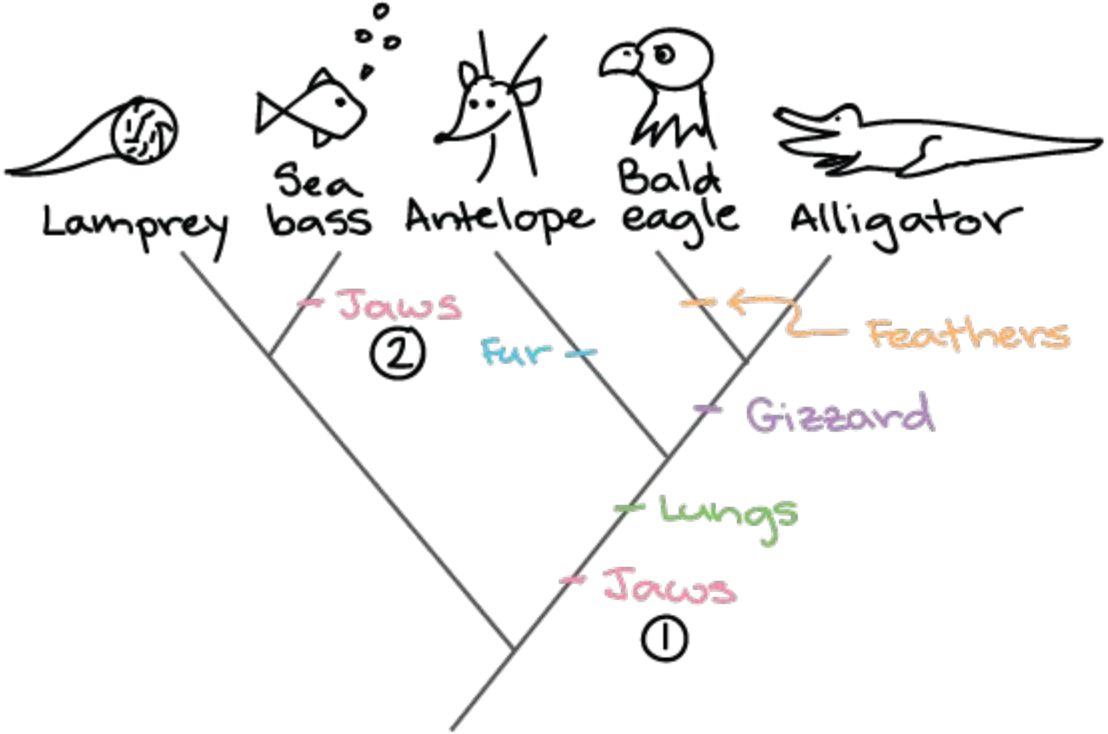
Phylogenetic tree


I think
B
C
D
1
A

Darwin's notes (1837)



"Modern" phylogeny


Lamprey
Sea bass
Antelope
Bald eagle
Alligator
Jaws
2
Fur
Feathers
Gizzard
Lungs
Jaws
1

(Cladogram)

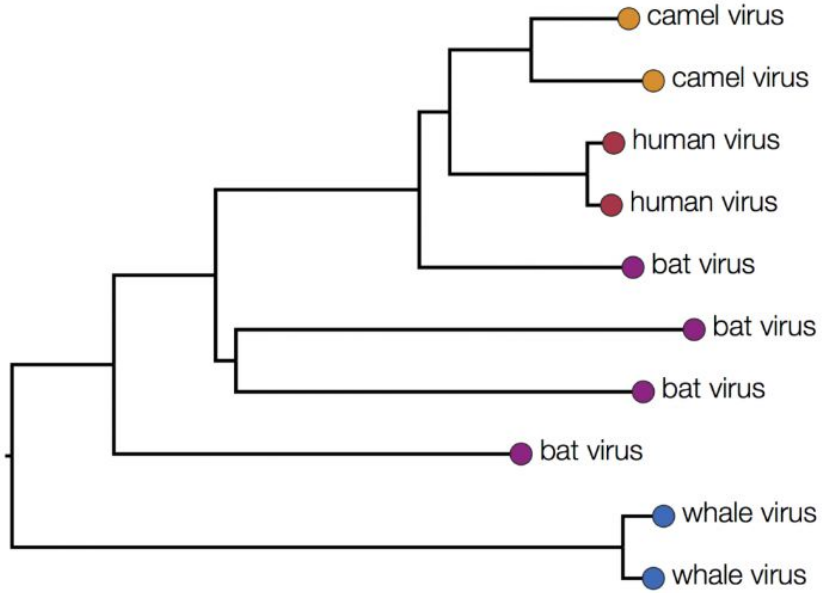
1. Relationships between species

TOPOLOGY

2. Evolutionary changes and history

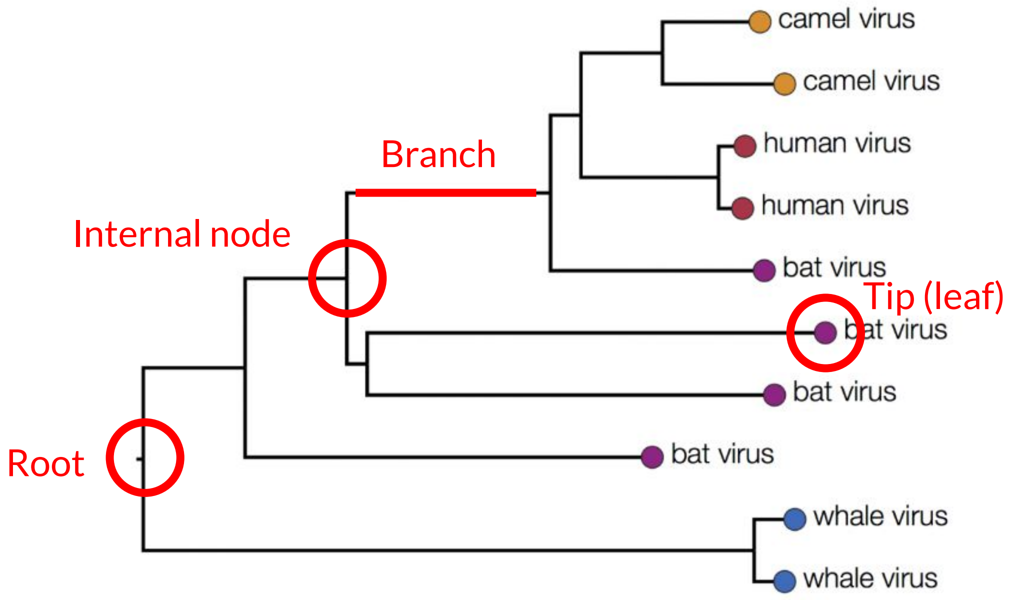
BRANCH LENGTH

BRANCH LENGTH



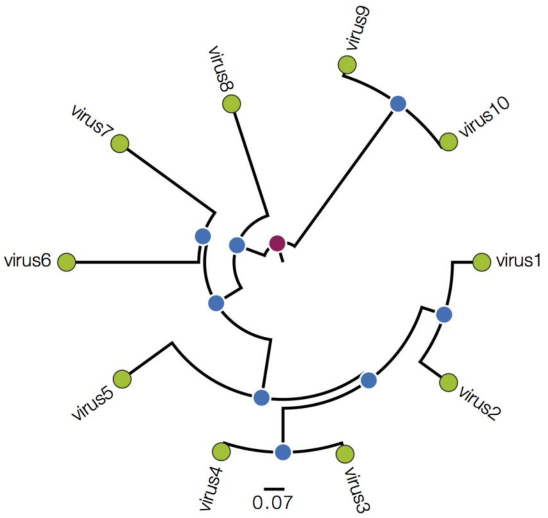
TOPOLOGY

Components of a phylogenetic tree

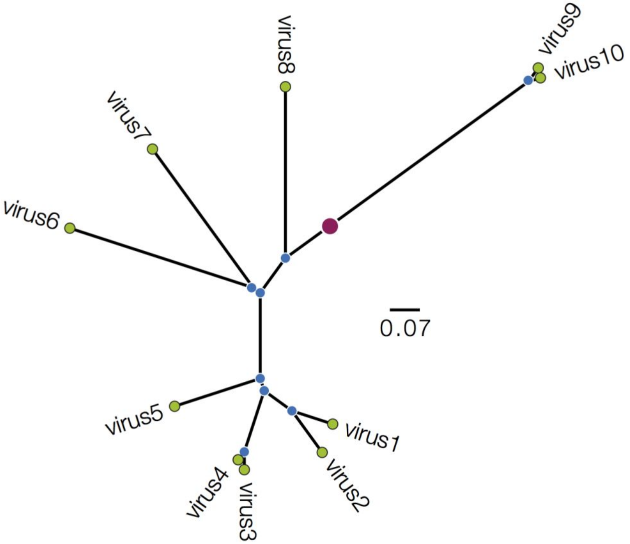


Components of a phylogenetic tree

Variety of tree formats...



Circular/polar

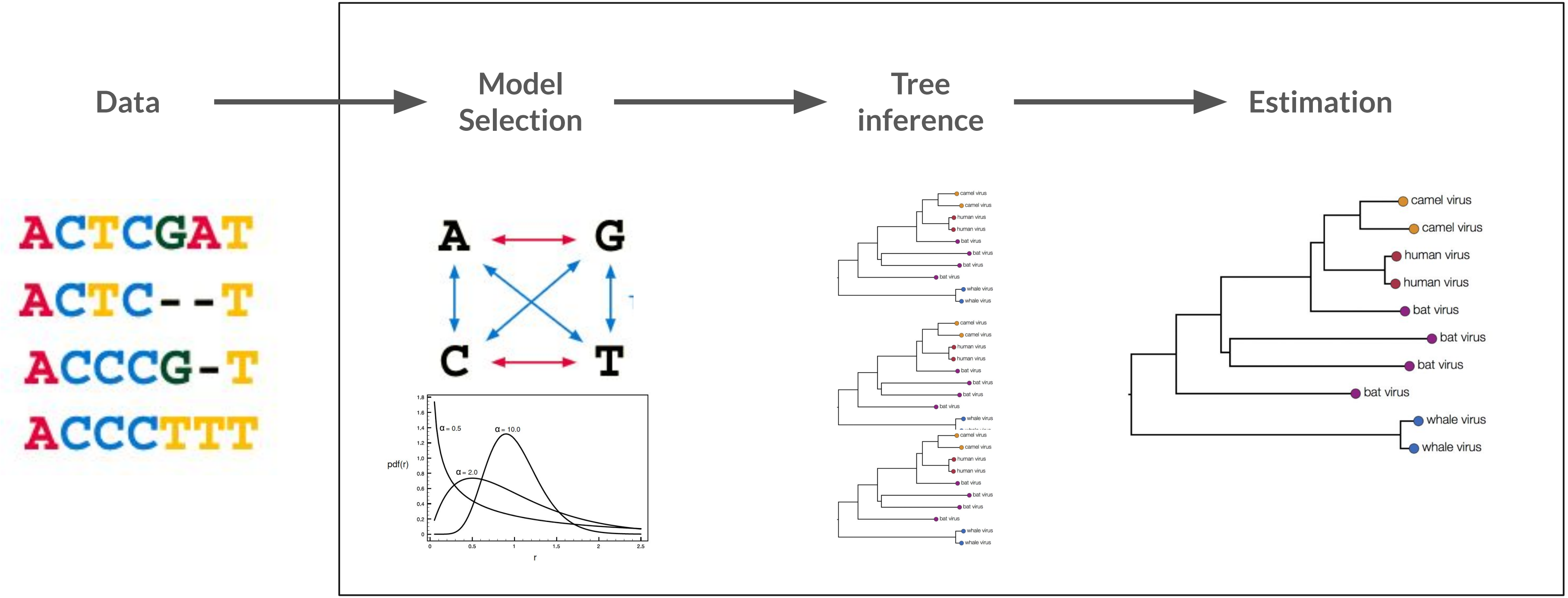


Unrooted

Applications

- Wide range of applications
- Epidemiology - origin, transmission, monitor outbreaks
- Clinical - Drug vaccine, design
- Conservation - identify diversity hotspots
- Cultural - Evolution of art, music, linguistics

PHYLOGENETICS PROGRAMS

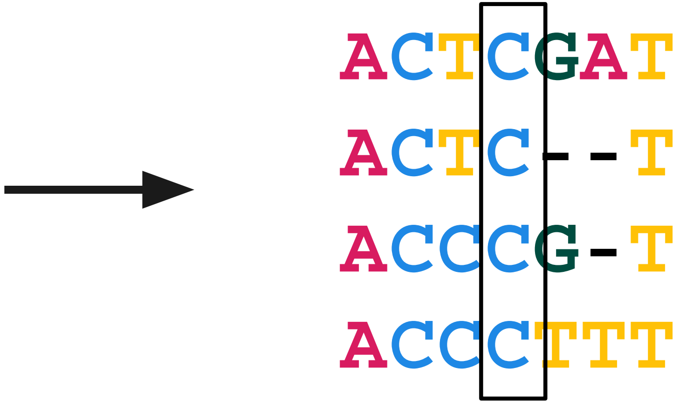


Data → Model Selection → Tree inference → Estimation

Inference workflow

Sequence alignment

Camel	ACTCGAT
Human	ACTCT
Bat	ACCCGT
Whale	ACCCTTT



```
ACTCGAT
ACTC--T
ACCCG-T
ACCCTTT
```

Homologous sites aligned

Inference methods

- Parsimony
- Distance
- Maximum likelihood
- Bayesian

BEAST

Bayesian Evolutionary Analysis Sampling Trees

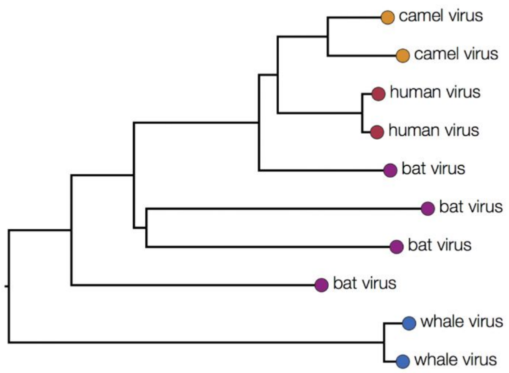
Mr Bayes

Bayesian Inference of Phylogeny

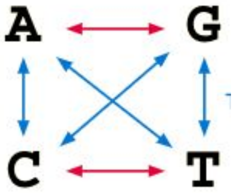
RAxML

IQ-TREE

Efficient software for phylogenomic inference

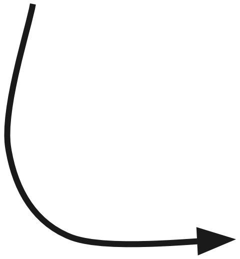


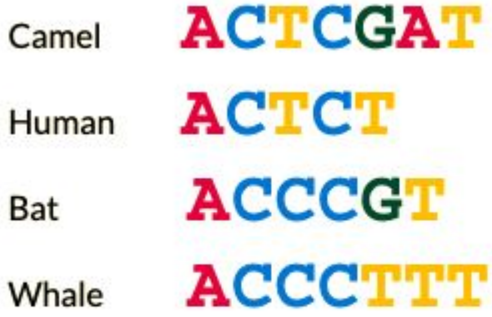
+



P(D|H)

"Which phylogeny and evolutionary model best explains our data?"





Camel	ACTCGAT
Human	ACTCT
Bat	ACCCGT
Whale	ACCCTTT

Maximum likelihood estimation

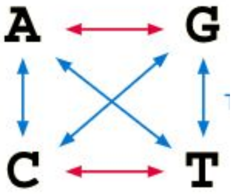
Evolutionary models

Base frequencies

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

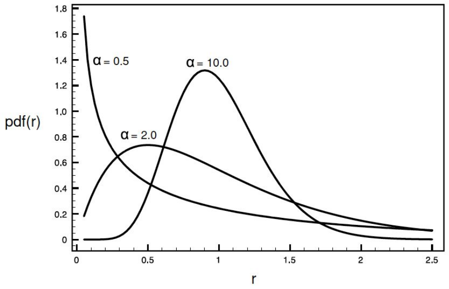
Substitution rates

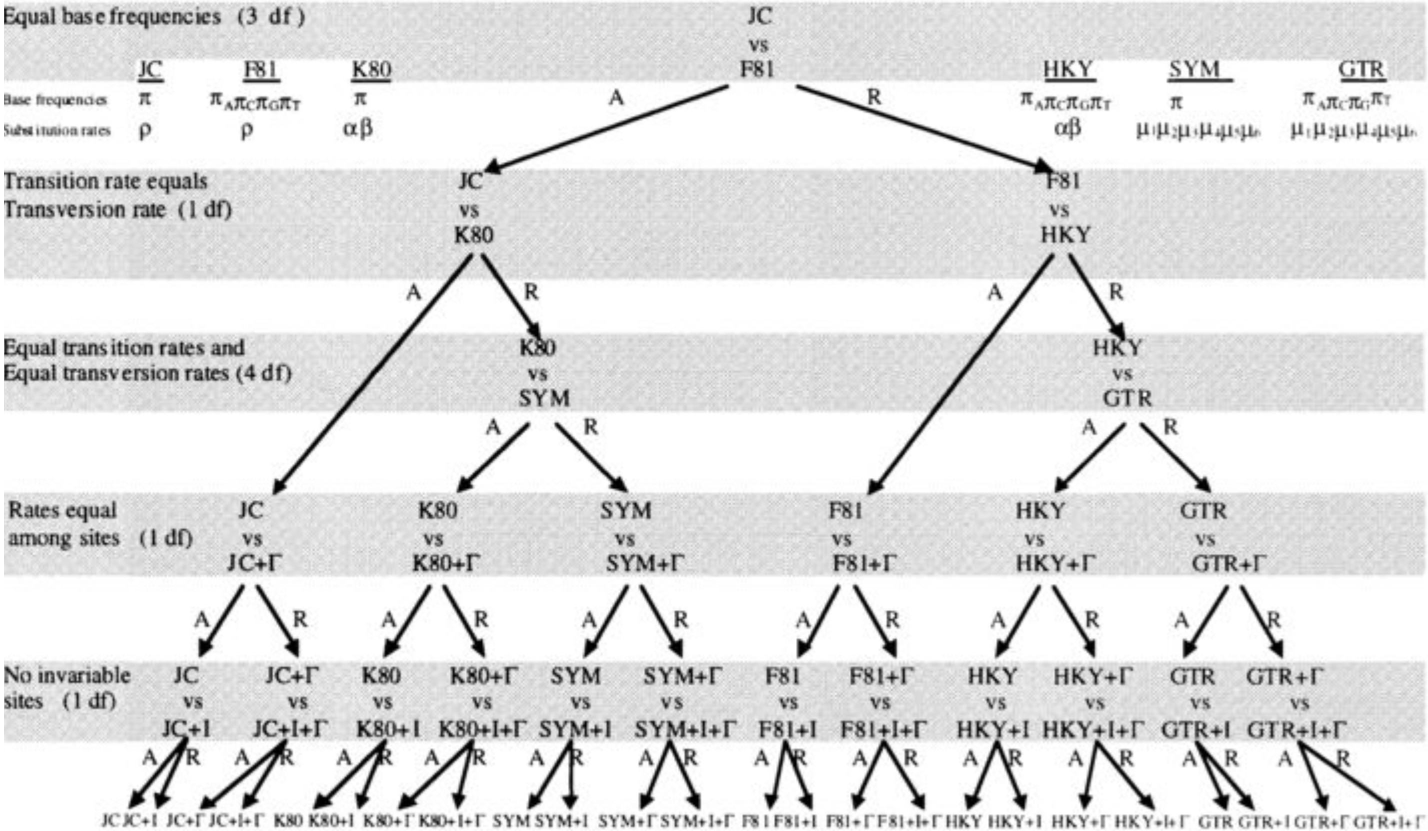
ρ, α, β, μ



Rate variation

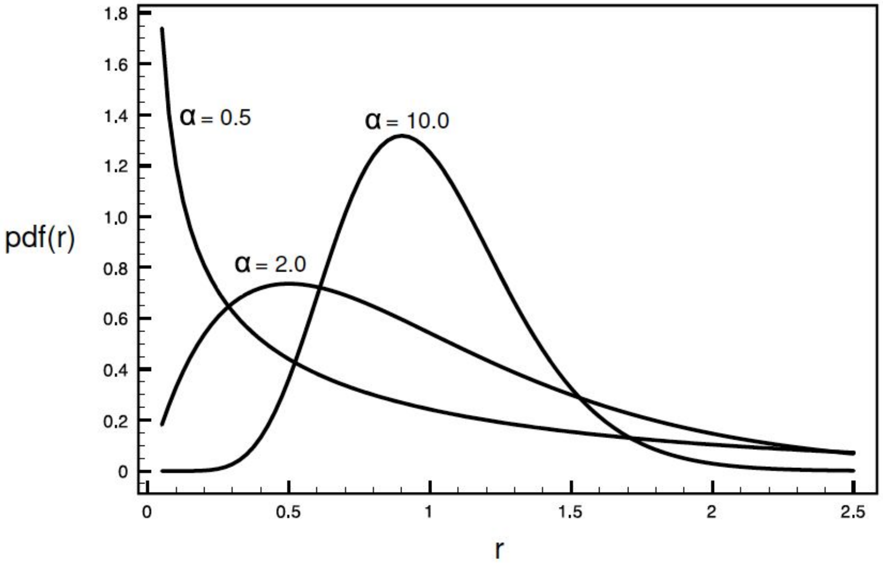
+I +G





Hierarchical model test (Posada and Crandall, 1998)

Rate variation



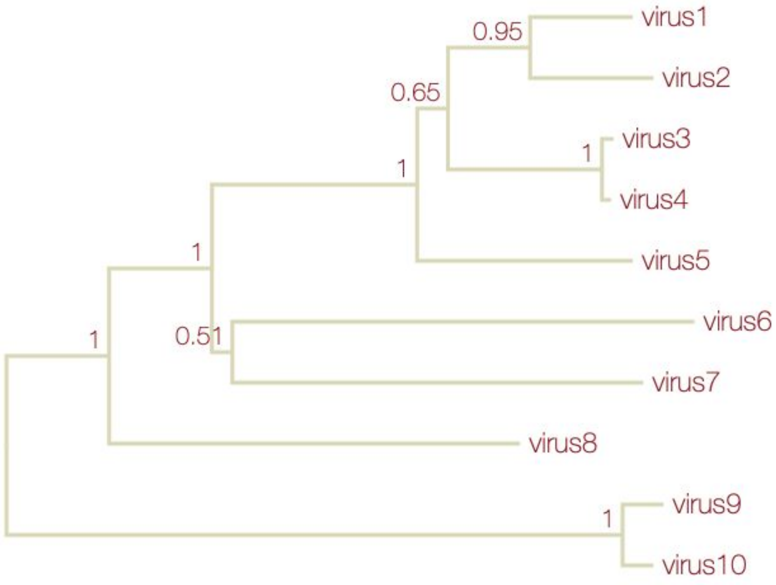
Codon positions mutate at different rates (3rd > 2nd > 1st)

(**+G**) The gamma distribution models rate heterogeneity

(**+I**) Invariant sites assumes “sites do not vary”

Statistical tests

- Bootstrap - measures the certainty of a tree estimate
- Model selection tests:
  - Likelihood ratio tests
  - Akaike Information Criteria (AIC)
  - BIC



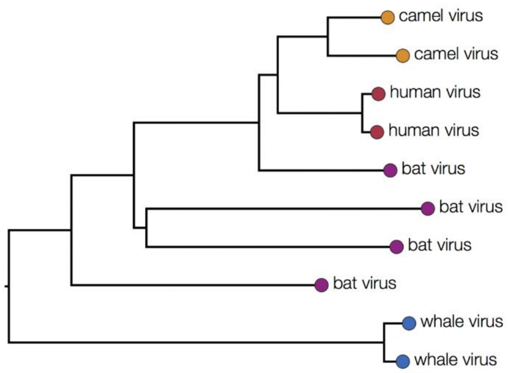
Not all sites evolve at the same rate

+

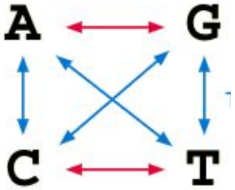
P(D|H)

“Which mode of coin toss and type of coin best explains our data?”

A conceptual example

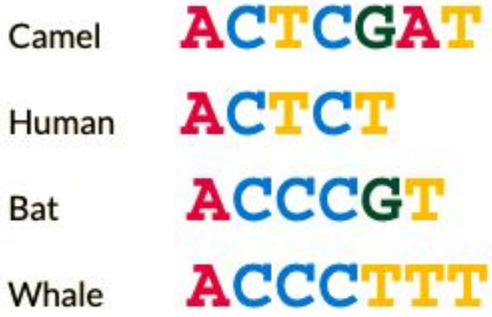


+



P(D|H)

"Which phylogeny and evolutionary model best explains our data?"



Camel	ACTCGAT
Human	ACTCT
Bat	ACCCGT
Whale	ACCCTTT

Maximum likelihood estimation

Exercise 3. Haiti outbreak

CUERVO

Exercise 4. Vaccinating ebola

Exercise 5. HIV transmission