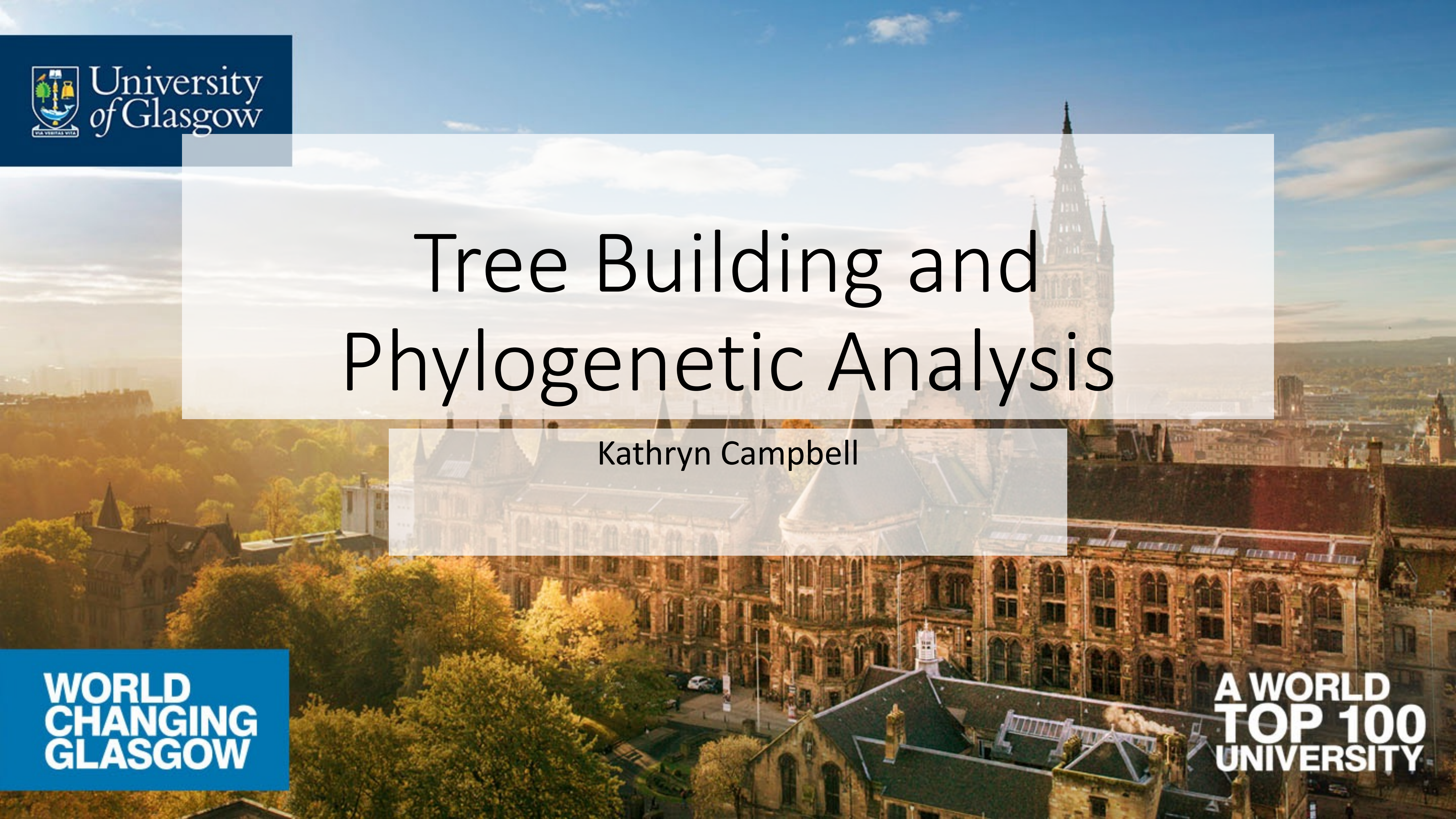


University of Glasgow

Tree Building and Phylogenetic Analysis

Kathryn Campbell

WORLD CHANGING GLASGOW

A WORLD TOP 100 UNIVERSITY

Schedule

Alignments & Understanding Trees

- Alignments
- **Practical:** Viewing and aligning sequences
- Terminology
- Tree building methods
- **Practical:** Interpreting phylogenetic trees

Tree Building

- Substitution models
- Statistical support (bootstrapping, posterior probabilities)
- **Practical:** Using modeltest and IQTREE2 to find the best substitution model and generate a maximum likelihood tree
- Visualisation using FigTree
- Annotating using metadata
- **Practical:** Preparation of publication ready trees with appropriate annotation and bootstrapping in FigTree

Substitution models

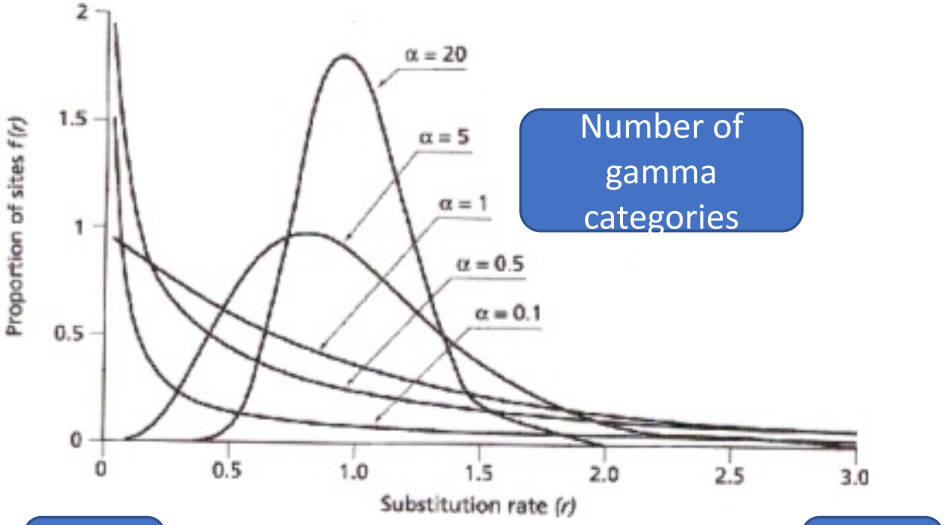
- Markov models that describe changes over evolutionary time

- Attempt to predict the rate of substitution for nucleotides or amino acids at a given site, and also the distribution of substitutions across the entire sequence

- Jukes-Cantor (JC)
  - all nucleotides occur at the same frequency and undergo change at the same rate
- **General Time Reversible (GTR)**
  - **Each type of nucleotide change occurs at its own rate (e.g. A->C =/= C->A)**
- Kimura 2-Parameter (K2P) and Kimura 3-Parameter (K3P)
- Hasegawa-Kishono-Yano (HKY)

Variation among sites

Some sites undergo changes more frequently than others - can be expressed using a gamma distribution



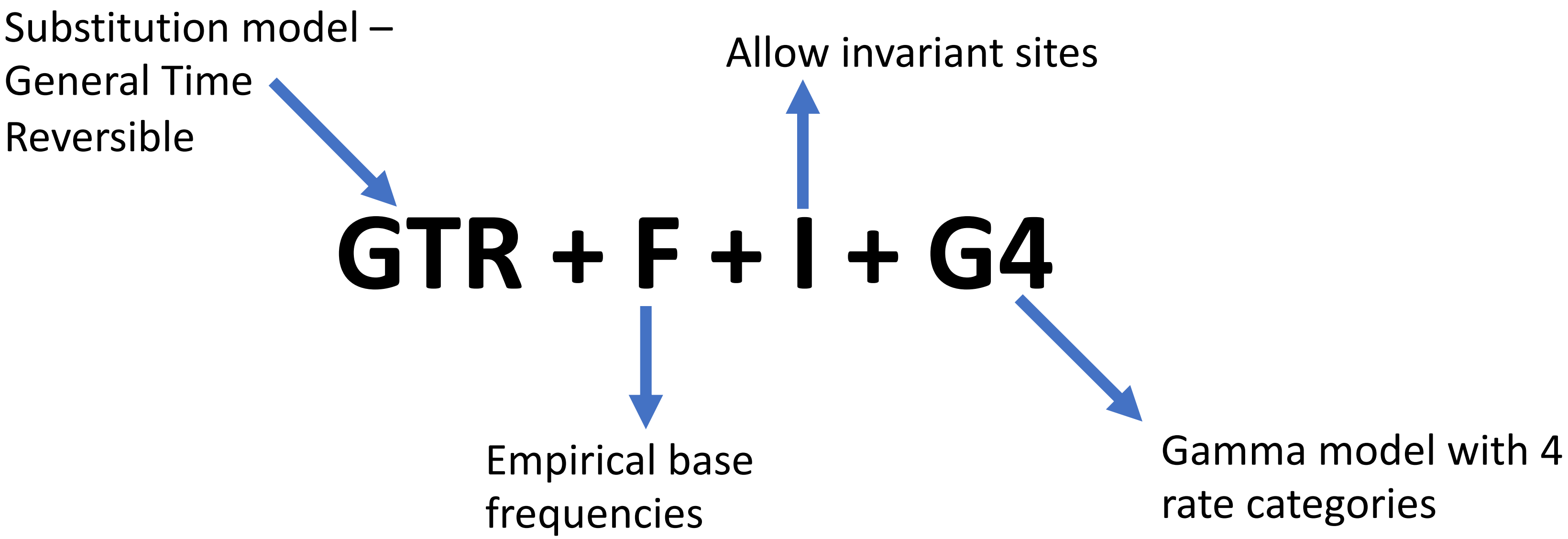
Number of gamma categories

Slow

Fast

Some sites are are not allowed to change, as they have essential roles – these are called invariant sites

Interpreting Substitution Models



Substitution model – General Time Reversible

Allow invariant sites

GTR + F + I + G4

Empirical base frequencies

Gamma model with 4 rate categories

Tree Building Methods

- Distance based
- Maximum parsimony
- **Maximum likelihood**
- **Bayesian methods**

Maximum Likelihood

- Maximize the probability of the sequences, given a tree and its branch lengths and an evolutionary model and its parameters

- Important features
  - Allows full use of evolutionary models
  - Relies heavily on model chosen => can be misleading if there is much variation in the substitution process among lineages
  - Computationally much more demanding

  - **IQTree**

Bayesian Methods

- Objective: determine the posterior distribution of trees given the sequence data
- Based on this distribution, ‘best’ tree can be identified

- Important features:
  - Allows full use of evolutionary models
  - Need to include priors
  - Posterior probabilities are approximated through Markov Chain Monte Carlo (MCMC) methods that sample from the posterior
  - Clade probabilities provide measure of uncertainty

  - **BEAST**

Bootstrapping

- Artificial dataset of same size is generated by picking columns with replacement

```
ATGCAGGTA   AAGCCGGTA  GCGCAGGAA  ATAAGGTTT
ATGCTGCTA   AAGCCGCTA  GCGCTGCAA  ATTTGCTTT
ATGCAGCTC   AAGCCGCTC  GCGCAGCCC  ATAAGCTTT
TAGCAGGAC   TTGCCGGAC  GCGCAGGCC  TAAAGGAAA
ORIGINAL
```

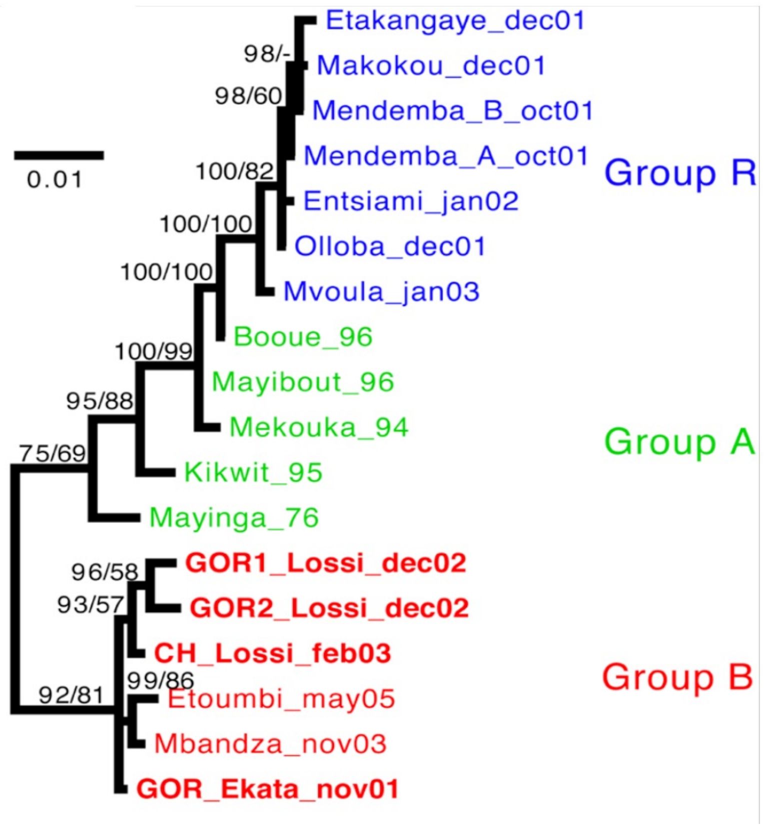
- Tree building applied to these bootstrap matrices
- The frequency with which a node appears is taken as a measure of confidence for that node

Posterior Probabilities

- Count the frequency of a clade within the posterior distribution of trees

- Less conservative: tend to be much higher than bootstrap values

- Strong support:
  - Bootstrap >0.7
  - Posterior prob. >0.95



Phylogenetic Analysis Steps

1) Collect homologous sequences

2) Conduct multiple alignment

3) Fit an appropriate substitution model

4) Estimate tree(s) under that model

5) Test the reliability of the estimated tree(s)

6) Interpret and apply the phylogenetic tree

7) Potentially repeat steps 4-6 using different tree building methods and/or additional data

Phylogenetic Analysis Steps

1) Collect homologous sequences
2) Conduct multiple alignment
3) Fit an appropriate substitution model
4) Estimate tree(s) under that model
5) Test the reliability of the estimated tree(s)
6) Interpret and apply the phylogenetic tree
7) Potentially repeat steps 4-6 using different tree building methods and/or additional data

Follow the instructions in
day4_tree_building.pdf

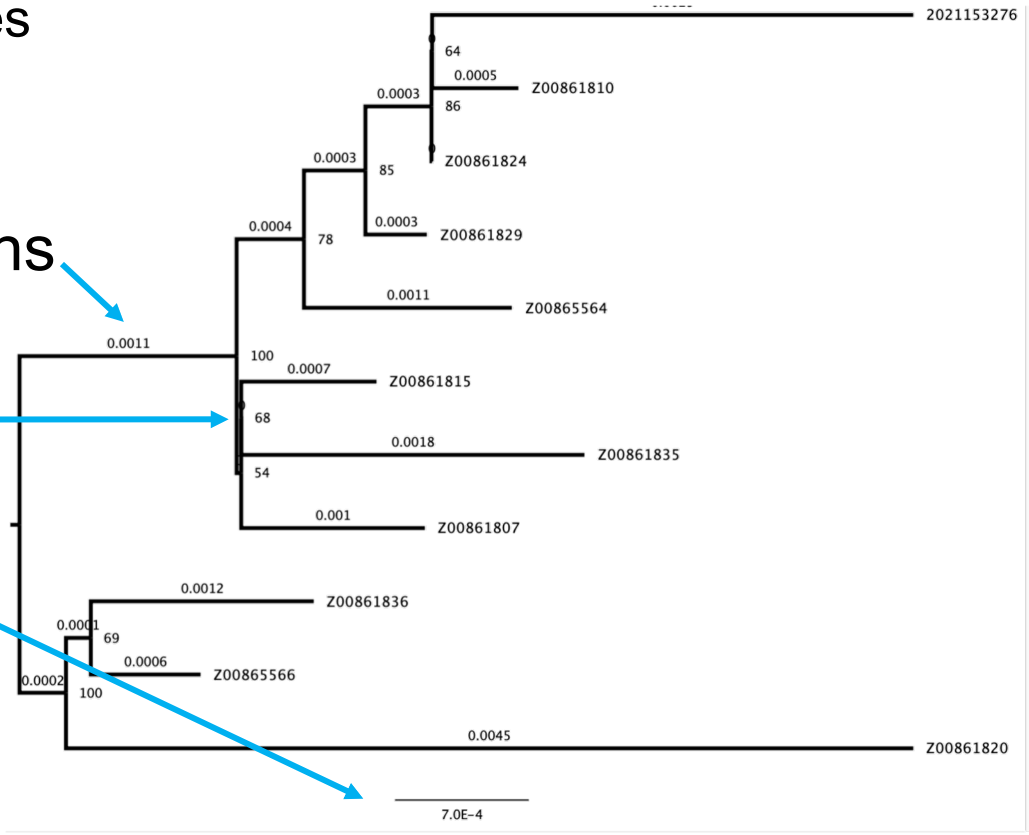
Visualizing Trees

Topology Changes:

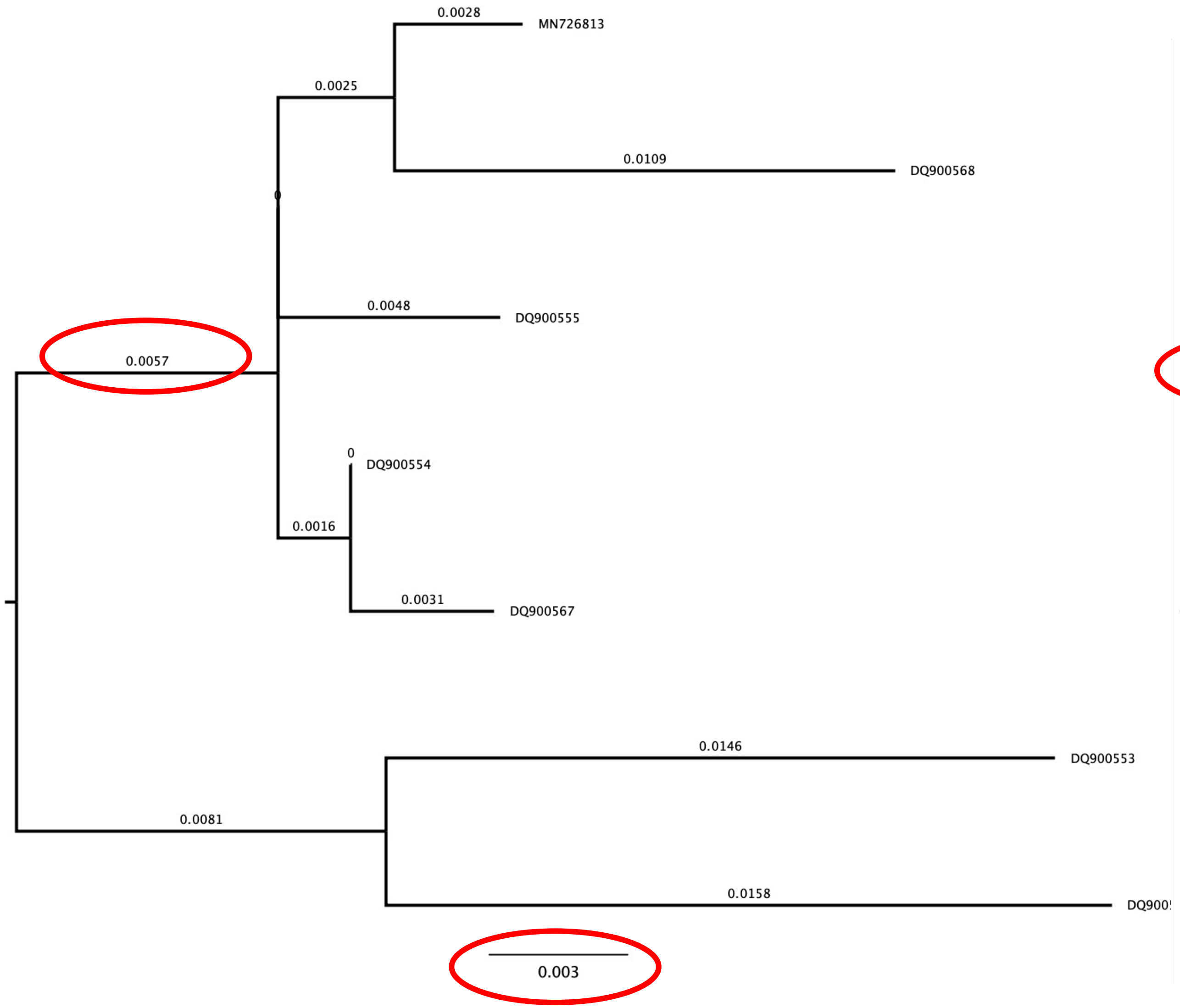
- Position of the root
- “Swiveling” the branches
- Ordering the nodes

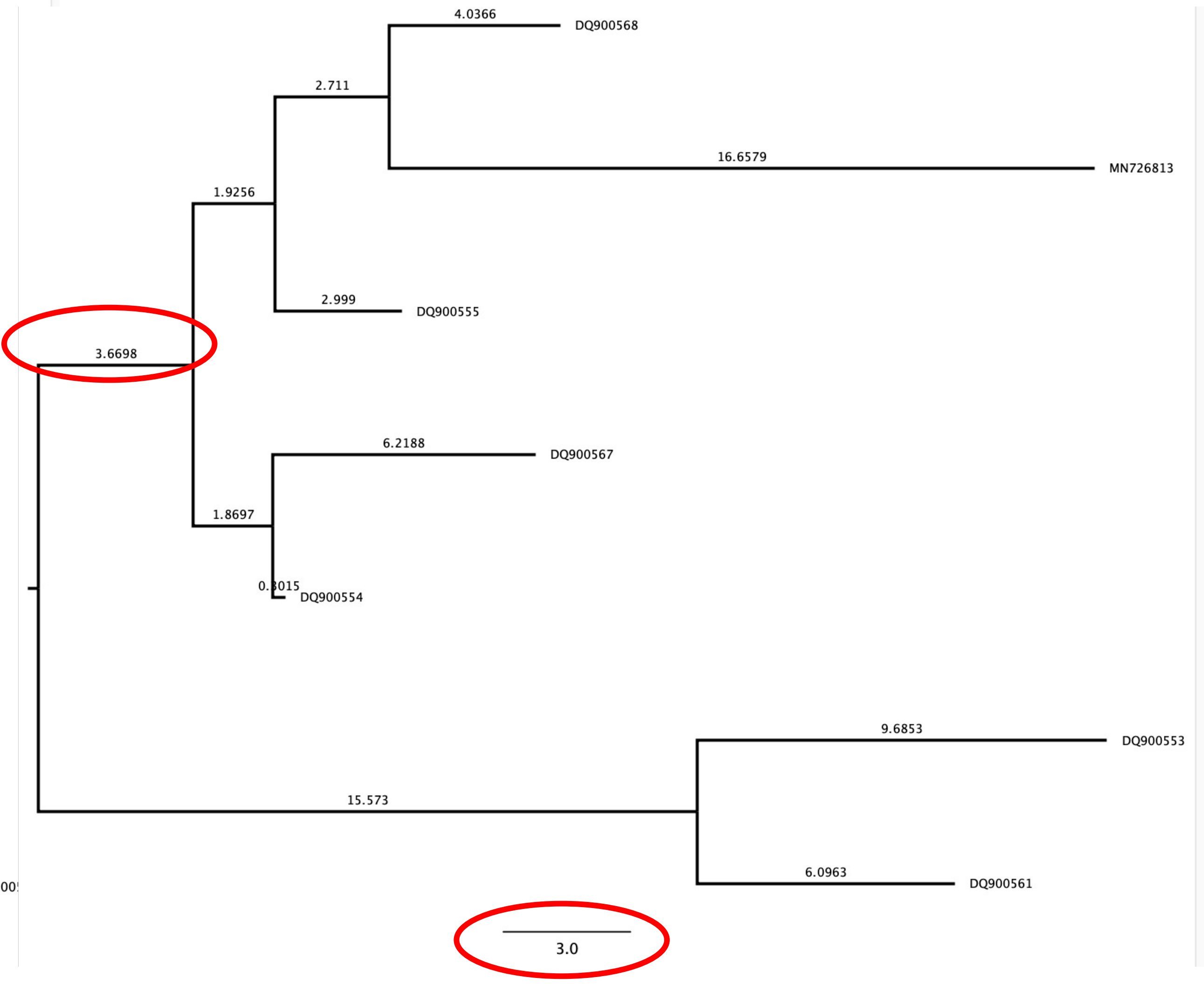
Additional information:

- Branch lengths
- Node labels
  - Bootstraps
  - Node ages
- Scale bar



Units


0.0028
MN726813
0.0025
0.0109
DQ900568
0
0.0048
DQ900555
0.0057
0
DQ900554
0.0016
0.0031
DQ900567
0.0146
DQ900553
0.0081
0.0158
DQ900
0.003

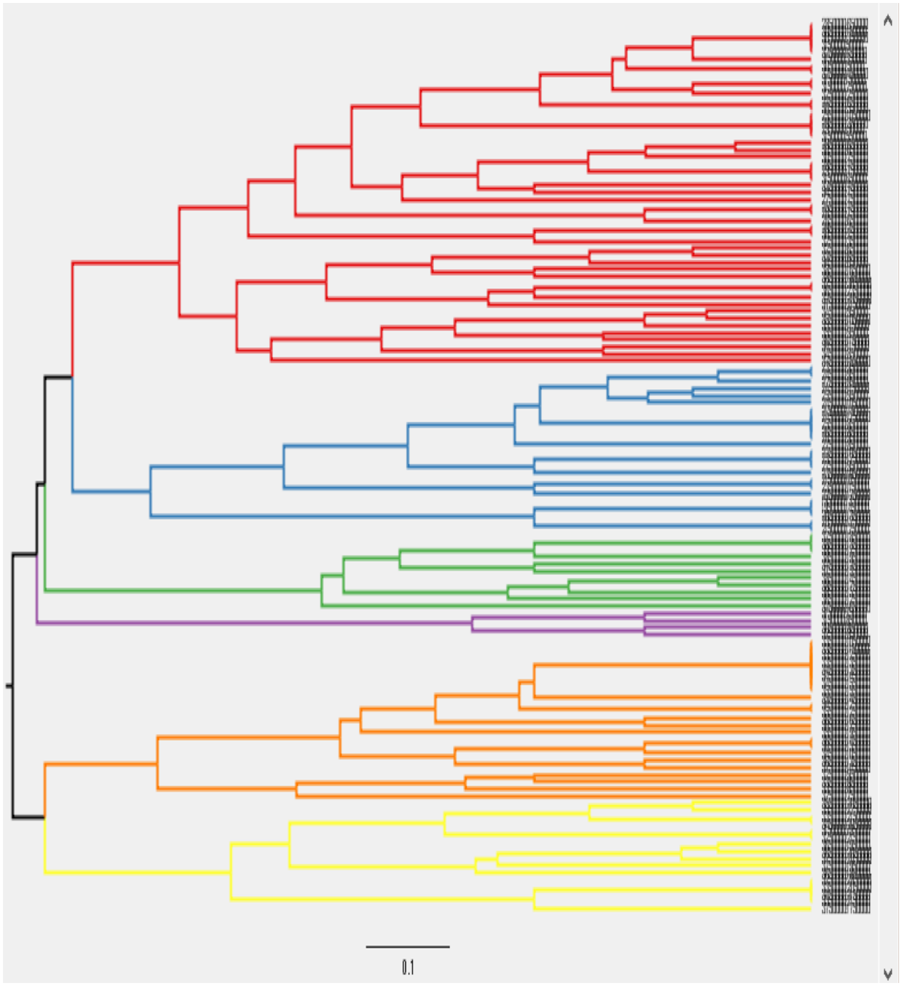

4.0366
DQ900568
2.711
16.6579
MN726813
1.9256
2.999
DQ900555
3.6698
6.2188
DQ900567
1.8697
0.3015
DQ900554
9.6853
DQ900553
15.573
6.0963
DQ900561
3.0

Adding Metadata

Can add useful information such as:

- Species/host
- Location
- Lineage

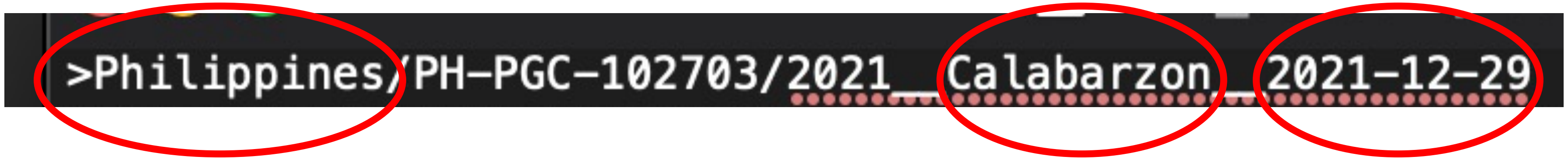
Can help predict the location/host origin of a cluster



Adding Metadata

Ideally, we have a nicely formatted metadata table with information about all of our sequences. This isn’t always the case!

Sometimes the sequence ID contains useful information that needs to be extracted!



```
>Philippines/PH-PGC-102703/2021__Calabarzon__2021-12-29
```

If all the IDs follow the same format, an automated pipeline to extract the information can be made!

Phylogenetic Analysis Steps

1) Collect homologous sequences

2) Conduct multiple alignment

3) Fit an appropriate substitution model

4) Estimate tree(s) under that model

5) Test the reliability of the estimated tree(s)

6) Interpret and apply the phylogenetic tree

7) Potentially repeat steps 4-6 using different tree building methods and/or additional data

Phylogenetic Analysis Steps

For the practical:
Follow instructions in **day4_annotation.pdf**

1) Collect homologous sequences
2) Conduct multiple alignment
3) Fit an appropriate substitution model
4) Estimate tree(s) under that model
5) Test the reliability of the estimated tree(s)
6) Interpret and apply the phylogenetic tree
7) Potentially repeat steps 4-6 using different tree building methods and/or additional data

University of Glasgow

Good luck, and happy tree building!

#UofGWorldChangers

@UofGlasgow

@ThatKatC