# Introduction to phylogenetics

**Thibaut Jombart**

15 December 2017

MRC Centre for Outbreak Analysis and Modelling
Imperial College London

# Outline

Phylogenies...

Distance trees

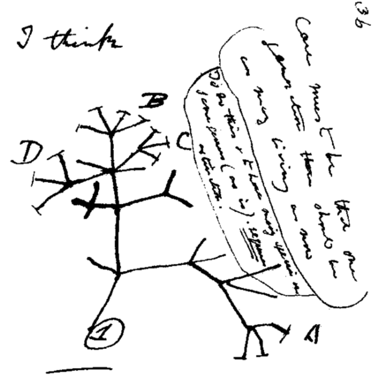
Parsimony

Likelihood/Bayesian

Uncertainty

Pitfalls & best practices

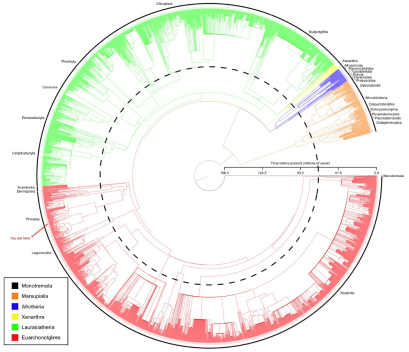
# Phylogenetics: from the origins...



C. Darwin, Notebook, 1837.

*'From the first growth of the tree, many a limb and branch has decayed and dropped off; and these fallen branches of various sizes may represent those whole orders, families, and genera which have now no living representatives, and which are known to us only in a fossil state.'*

# Phylogenetics: ...to the present



Bininda-Emonds *et al.*, 2007, Nature.

- phylogenetic trees are part of the standard toolbox of genetic data analysis
- represent the evolutionary history of a set of (sampled) taxa

# And the main difference is...

# And the main difference is...

Current trees look better!

# And the main difference is...

**Current trees look better!**

(and some other minor differences)

# About the minor differences...

- DNA sequencing revolution
- huge data banks freely available (e.g. GenBank)
- easier, cheaper, faster to obtain DNA sequences
- increasing number of full genomes available

# About the minor differences...

- DNA sequencing revolution
- huge data banks freely available (e.g. GenBank)
- easier, cheaper, faster to obtain DNA sequences
- increasing number of full genomes available

**Different ways to exploit this information.**

# Outline

Phylogenies...

Distance trees

Parsimony

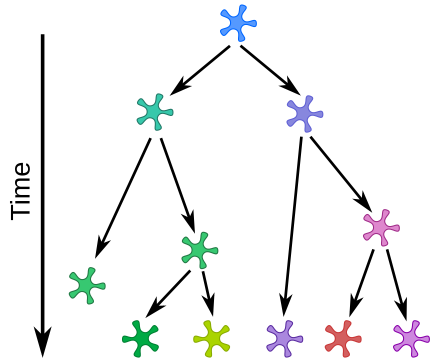
Likelihood/Bayesian

Uncertainty

Pitfalls & best practices

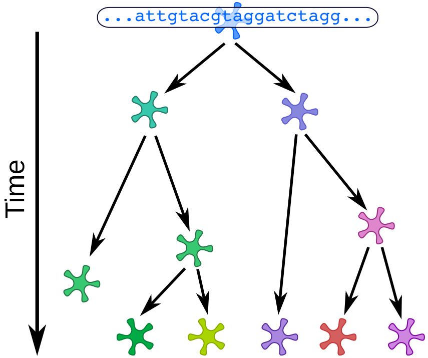
# Genetic changes (substitution) accumulate over time

**Substitution**: replacement of a nucleotide (e.g. a → t)



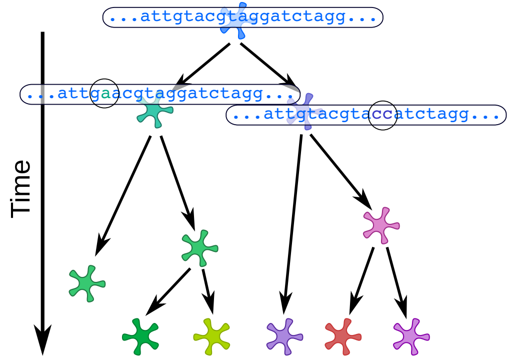
# Genetic changes (substitution) accumulate over time

**Substitution**: replacement of a nucleotide (e.g. a → t)



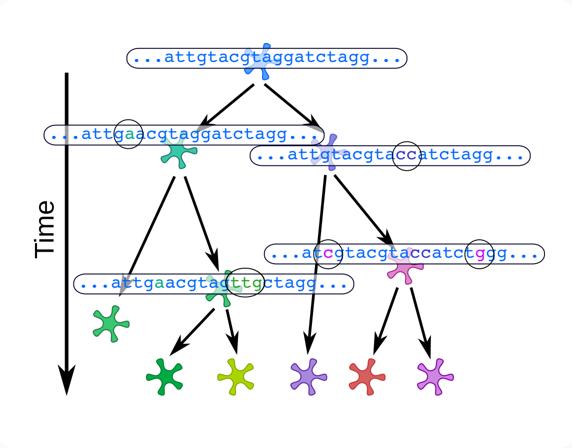
# Genetic changes (substitution) accumulate over time

**Substitution**: replacement of a nucleotide (e.g. a → t)



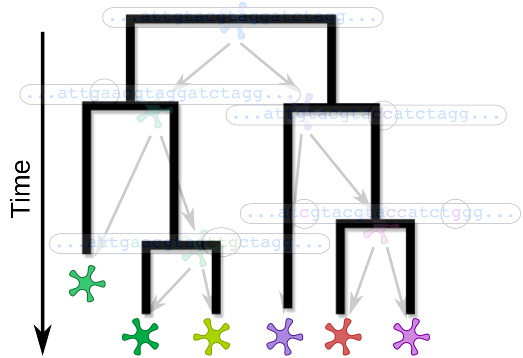
# Genetic changes (substitution) accumulate over time

**Substitution**: replacement of a nucleotide (e.g. a → t)



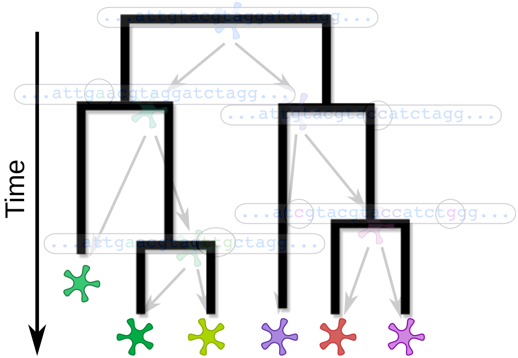
# Genetic changes (substitution) accumulate over time

**Substitution**: replacement of a nucleotide (e.g. a → t)



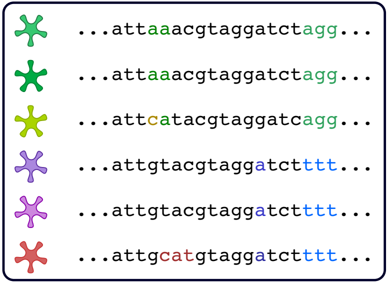
# Genetic changes (substitution) accumulate over time

**Substitution**: replacement of a nucleotide (e.g. a → t)



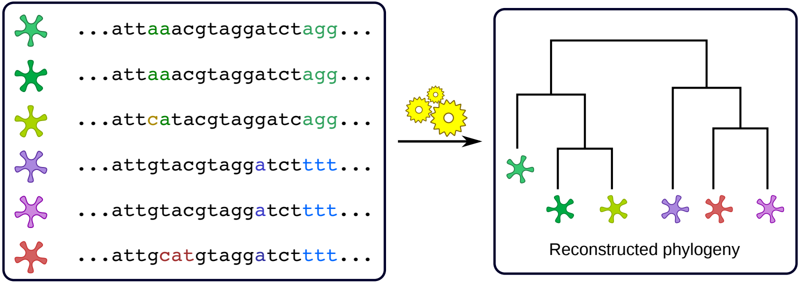
Substitution patterns reflect the evolutionary history

# Using substitution patterns to reconstruct the evolutionary history



```
...attaaacgtaggatctagg...
...attaaacgtaggatctagg...
...attcatacgtaggatcagg...
...attgtacgtaggatctttt...
...attgtacgtaggatctttt...
...attgcatgtaggatctttt...
```

# Using substitution patterns to reconstruct the evolutionary history



...attaaacgtaggatctagg...

...attaaacgtaggatctagg...

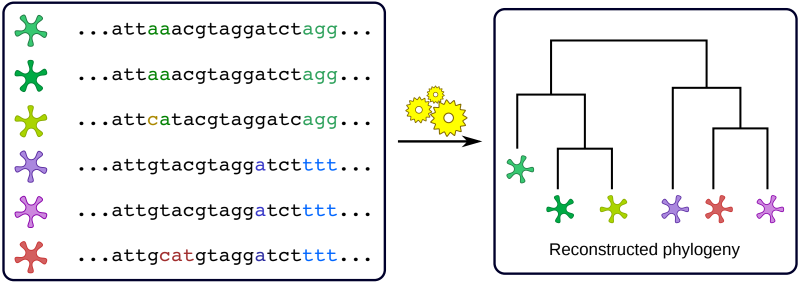
...attcatacgtaggatcagg...

...attgtacgtaggatctttt...

...attgtacgtaggatctttt...

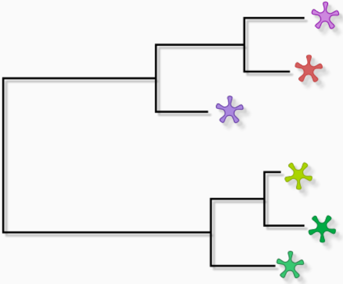
...attgcatgtaggatctttt...

# Using substitution patterns to reconstruct the evolutionary history

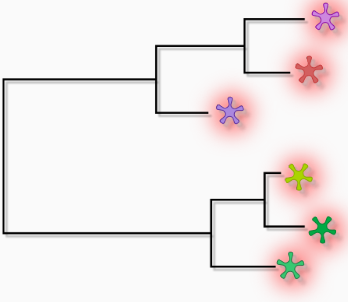


Phylogenetics aim to reconstruct evolutionary trees (*phylogenies*) from genetic sequence data.

# Using trees to represent the evolutionary history



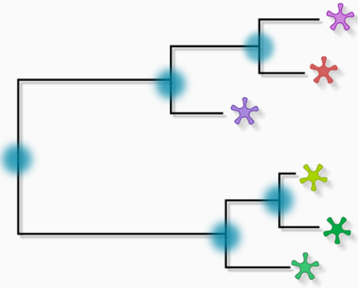
# Using trees to represent the evolutionary history



analysed sequences

- *"taxa"*
- *"tips"*
- *"leaves"*
- *"Operational Taxonomic Units (OTUs)"*

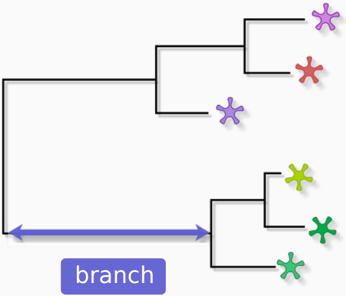
# Using trees to represent the evolutionary history



Most Recent Common Ancestors (MRCA)

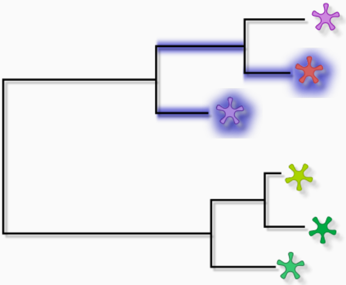
- *"nodes"*
- *"Hypothetical Taxonomic Units (HTUs)"*

# Using trees to represent the evolutionary history



- "*edge*"
- length = amount of evolution (**not time**, as a rule)
- length is optional

# Using trees to represent the evolutionary history



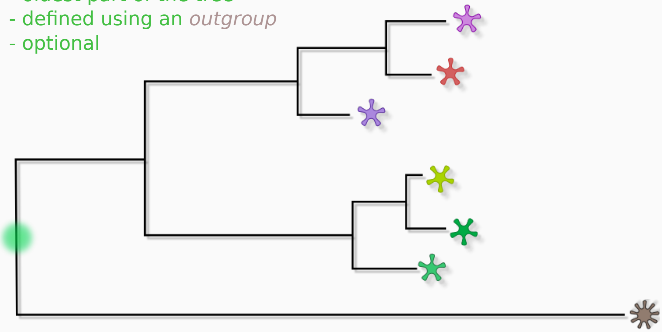
**distances between tips**

- "*patristic*" distance: sum of branch lengths
- other measures of distance/dissimilarity
- vertical axis meaningless

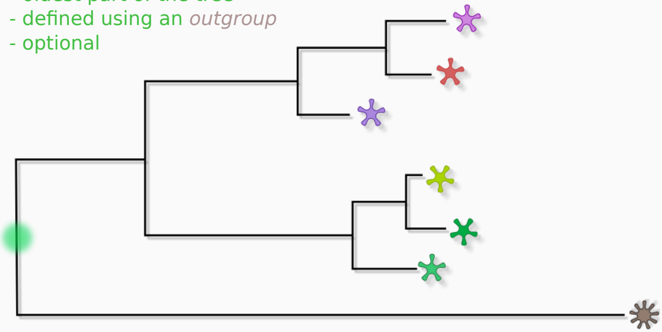
# Using trees to represent the evolutionary history

Root

- oldest part of the tree
- defined using an *outgroup*
- optional



# Using trees to represent the evolutionary history



Root

- oldest part of the tree
- defined using an *outgroup*
- optional

How to we build them?

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

## Build the tree

- distance-based methods
- maximum parsimony
- likelihood-based methods (ML, Bayesian)

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

## Build the tree

- distance-based methods
- maximum parsimony
- likelihood-based methods (ML, Bayesian)

## Analyse the tree

- assess uncertainty
- test phylogenetic signal
- model trait evolution
- ...

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

## Build the tree

- **distance-based methods**
- **maximum parsimony**
- **likelihood-based methods (ML, Bayesian)**

## Analyse the tree

- assess uncertainty
- test phylogenetic signal
- model trait evolution
- ...

# Outline

Phylogenies...

Distance trees

Parsimony

Likelihood/Bayesian

Uncertainty

Pitfalls & best practices

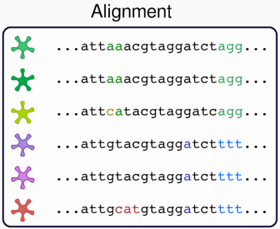
# Distance-based phylogenetic reconstruction

Approaches relying on **agglomerative clustering** algorithms (e.g. Single linkage, UPGMA, Neighbor-Joining)

**Rationale**

1. compute pairwise genetic distances $\mathbf{D}$
2. group closest sequences
3. update $\mathbf{D}$
4. go back to 2) until all sequences are grouped

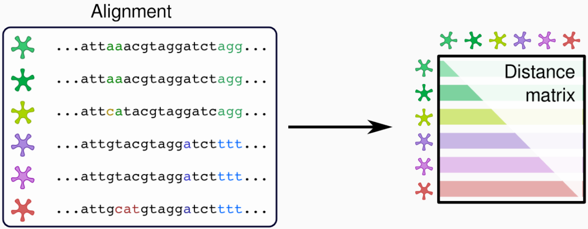
# Distance-based phylogenetic reconstruction



Alignment

```
...attaaacgtaggatctagg...
...attaaacgtaggatctagg...
...attcatacgtaggatcagg...
...attgtacgtaggatctttt...
...attgtacgtaggatctttt...
...attgcatgtaggatctttt...
```

# Distance-based phylogenetic reconstruction

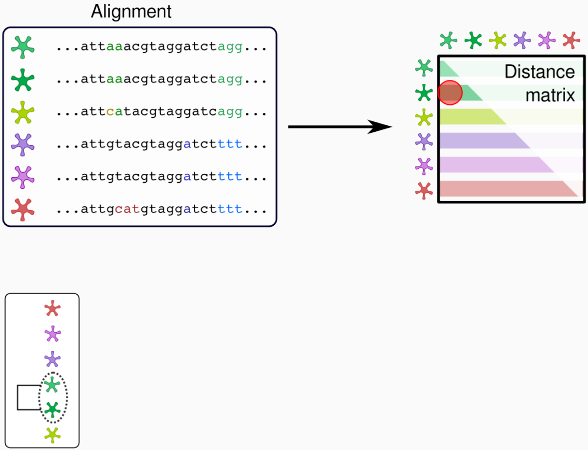


Alignment

```
...attaaacgtaggatctagg...
...attaaacgtaggatctagg...
...attcatacgtaggatcagg...
...attgtacgtaggatctttt...
...attgtacgtaggatctttt...
...attgcatgtaggatctttt...
```

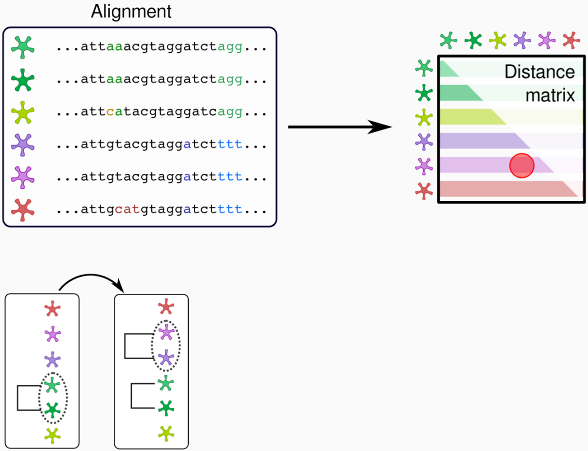
Distance matrix

# Distance-based phylogenetic reconstruction

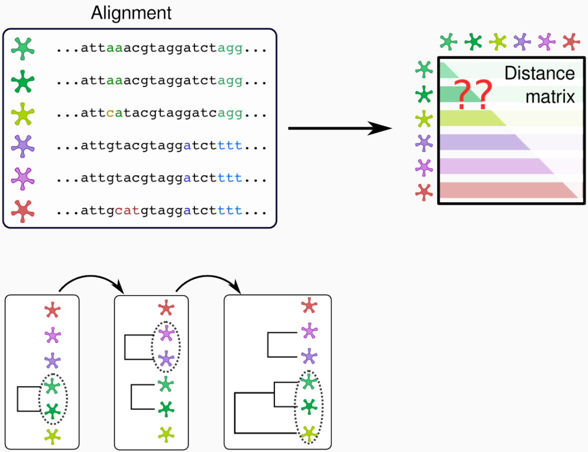


```
...attaaacgtaggatctagg...
...attaaacgtaggatctagg...
...attcatacgtaggatcagg...
...attgtacgtaggatctttt...
...attgtacgtaggatctttt...
...attgcatgtaggatctttt...
```

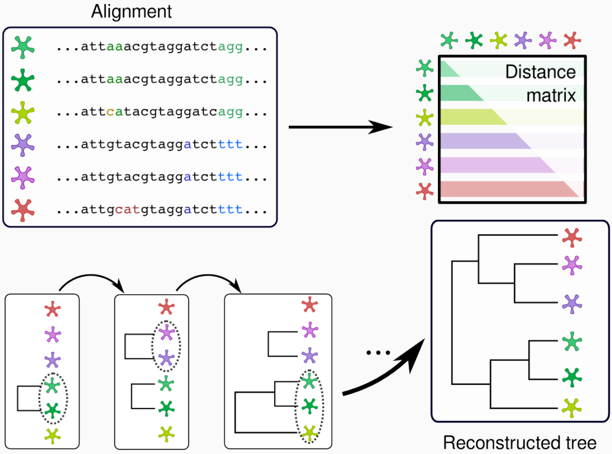
# Distance-based phylogenetic reconstruction



# Distance-based phylogenetic reconstruction



# Distance-based phylogenetic reconstruction



Alignment

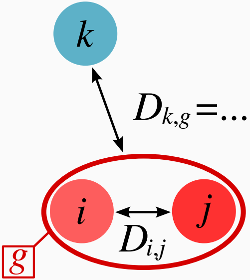
```
...attaaacgtaggatctagg...
...attaaacgtaggatctagg...
...attcatacgtaggatcagg...
...attgtacgtaggatctttt...
...attgtacgtaggatctttt...
...attgcatgtaggatctttt...
```

Distance matrix

Reconstructed tree

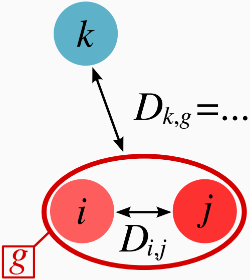
# What is the distance between a node and tips?

Hierarchical clustering:



# What is the distance between a node and tips?

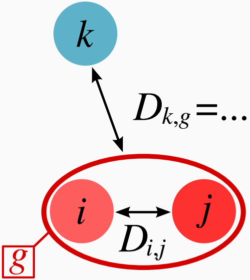
Hierarchical clustering:



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$

# What is the distance between a node and tips?

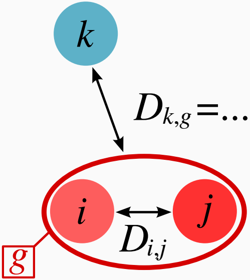
Hierarchical clustering:



- single linkage: $D_{k,g} = \text{min}(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \text{max}(D_{k,i}, D_{k,j})$

# What is the distance between a node and tips?

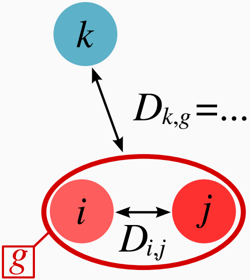
Hierarchical clustering:



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

# What is the distance between a node and tips?

**Hierarchical clustering:**



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

**Neighbor joining:**

Transforms original distances to account for heterogeneous rates of evolution.

# Distance-based phylogenetic reconstruction

**Advantages**

- simple
- flexible (many distances and clustering algorithms)
- fast and scalable (applicable to large datasets)

# Distance-based phylogenetic reconstruction

**Advantages**

- simple
- flexible (many distances and clustering algorithms)
- fast and scalable (applicable to large datasets)

**Limitations**

- sensitive to distance/clustering chosen
- evolutionary rates are not estimated
- no measure of uncertainty for the tree obtained

# Outline

Phylogenies...

Distance trees

Parsimony

Likelihood/Bayesian

Uncertainty

Pitfalls & best practices

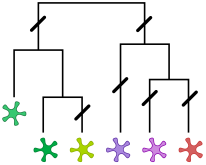
# Maximum parsimony phylogenies

Approaches relying on finding the tree with the **smallest number of character changes (substitutions)**

**Rationale**

1. start from a pre-defined tree
2. compute initial parsimony score
3. permute branches and compute parsimony score
4. accept new tree if the parsimony score is improved
5. go back to 3) until convergence

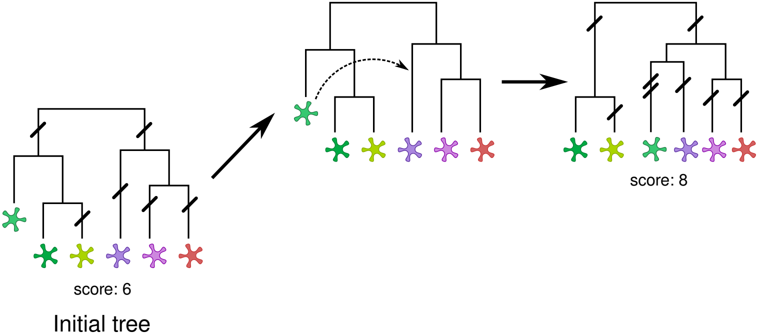
# Maximum parsimony phylogenies



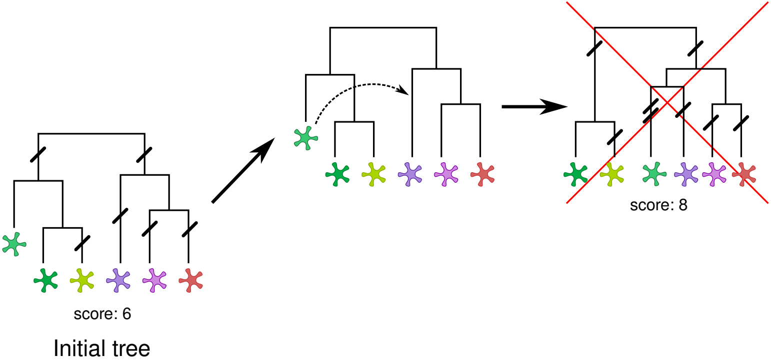
score: 6

Initial tree

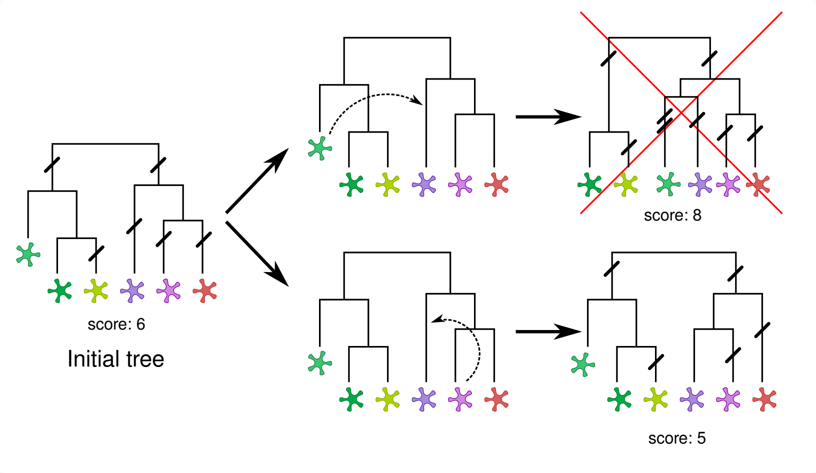
# Maximum parsimony phylogenies



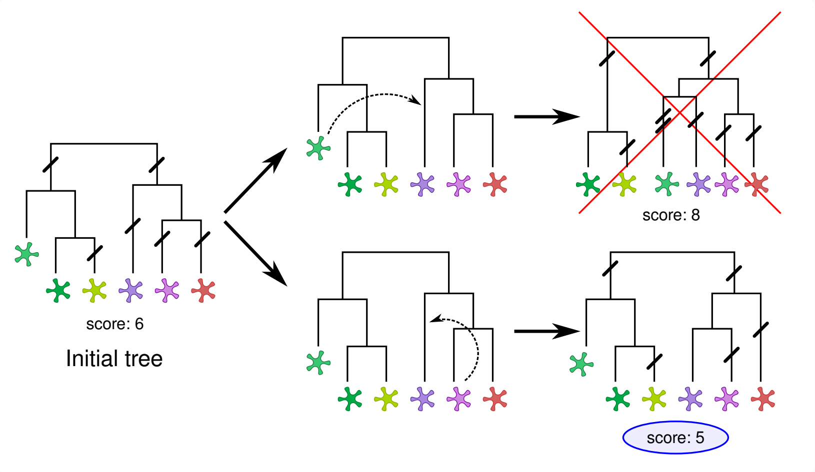
# Maximum parsimony phylogenies



# Maximum parsimony phylogenies



# Maximum parsimony phylogenies



# Maximum parsimony phylogenies

## Advantages

- applicable to any discontinuous characters (not just DNA)
- intuitive explanation: ‘simplest’ evolutionary scenario

# Maximum parsimony phylogenies

## Advantages

- applicable to any discontinuous characters (not just DNA)
- intuitive explanation: ‘simplest’ evolutionary scenario

## Limitations

- evolutionary rates are not estimated
- no measure of uncertainty for the tree obtained
- computer-intensive
- different types of substitutions ignored
- evolution not necessarily parsimonious
- sensitive to heterogeneous rates of evolution (*long branch attraction*)

# Outline

Phylogenies...

Distance trees

Parsimony

Likelihood/Bayesian

Uncertainty

Pitfalls & best practices

# Likelihood-based phylogenies (ML / Bayesian)

Approaches relying on a **model of sequence evolution**:

- **ML**: find tree and evolutionary rates with highest likelihood
- **Bayesian**: find tree and evolutionary rates to posterior probability

# Likelihood-based phylogenies (ML / Bayesian)

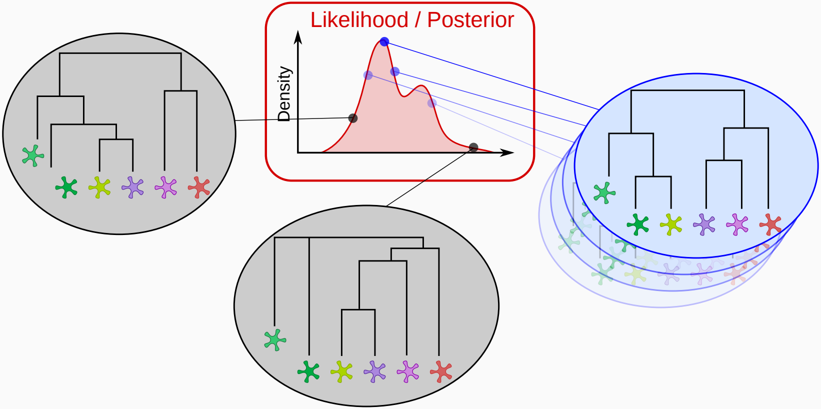
Approaches relying on a **model of sequence evolution**:

- **ML**: find tree and evolutionary rates with highest likelihood
- **Bayesian**: find tree and evolutionary rates to posterior probability

**Rationale**

1. start from a pre-defined tree
2. compute initial likelihood/posterior
3. permute branches, sample new parameters and compute likelihood/posterior
4. accept new tree and parameters based on likelihood/posterior improvement
5. go back to 3) until convergence

# Likelihood-based phylogenies (ML / Bayesian)



# Likelihood-based phylogenies (ML / Bayesian)

**Advantages**

- very flexible
- consistent with a model of evolution
- statistically consistent (model comparison)
- parameter estimation
- (Bayesian) several trees → measure of uncertainty

# Likelihood-based phylogenies (ML / Bayesian)

## Advantages

- very flexible
- consistent with a model of evolution
- statistically consistent (model comparison)
- parameter estimation
- (Bayesian) several trees → measure of uncertainty

## Limitations

- computer-intensive
- choice of the model of evolution
- (ML) no measure of uncertainty for the tree obtained
- (Bayesian) need to find a consensus tree

# Outline

Phylogenies...

Distance trees

Parsimony

Likelihood/Bayesian

Uncertainty

Pitfalls & best practices

# How do we know the tree is robust?

**Main issue:** assess the uncertainty of the tree topology / individual nodes

# How do we know the tree is robust?

**Main issue:** assess the uncertainty of the tree topology / individual nodes

## Approaches

- ML: model selection to compare trees (whole tree)

# How do we know the tree is robust?

**Main issue:** assess the uncertainty of the tree topology / individual nodes

**Approaches**

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)

# How do we know the tree is robust?

**Main issue:** assess the uncertainty of the tree topology / individual nodes

## Approaches

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)
- any method: bootstrap (individual nodes)

# How do we know the tree is robust?

**Main issue:** assess the uncertainty of the tree topology / individual nodes

## Approaches

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)
- **any method: bootstrap (individual nodes)**

# Bootstrapping phylogenies

- assess **variability due to sampling the genome** and **conflicting signals**
- relies on analysing **resampled datasets**

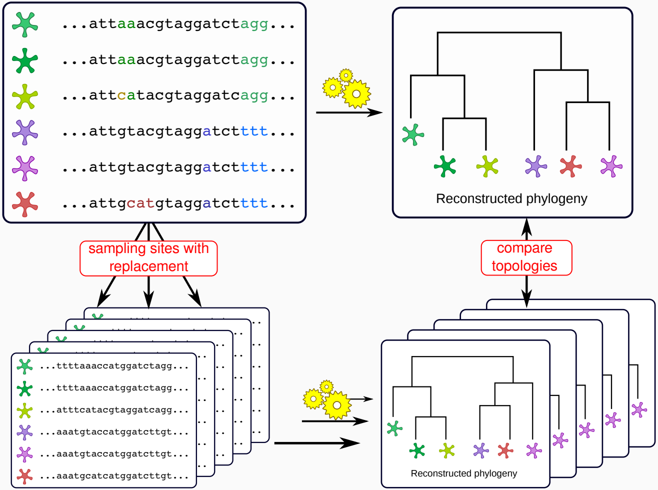
# Bootstrapping phylogenies

- assess **variability due to sampling the genome** and **conflicting signals**
- relies on analysing **resampled datasets**

## Rationale

1. obtain a reference tree
2. resample the sites with replacement
3. obtain a tree for the resampled dataset
4. go back to 2) until the desired number of bootstrapped trees is attained
5. compute the frequency of each bifurcation of the reference tree occuring in bootstrapped trees

# Bootstrapping phylogenies



...attaaacgtaggatctagg...

...attaaacgtaggatctagg...

...attcatacgtaggatcagg...

...attgtacgtaggatctttt...

...attgtacgtaggatctttt...

...attgcatgtaggatctttt...

Reconstructed phylogeny

sampling sites with replacement

compare topologies

...ttttaaaccatggatctagg...

...ttttaaaccatggatctagg...

...atttcatacgtaggatcagg...

...aaatgtaccatggatcttgt...

...aaatgtaccatggatcttgt...

...aaatgcatcatggatcttgt...

Reconstructed phylogeny

# Bootstrapping phylogenies

**Advantages**

- standard
- simple to implement

# Bootstrapping phylogenies

**Advantages**

- standard
- simple to implement

**Limitations**

- possibly computer-intensive
- assumes that the genome has been sampled randomly (often wrong)

# Outline

Phylogenies...

Distance trees

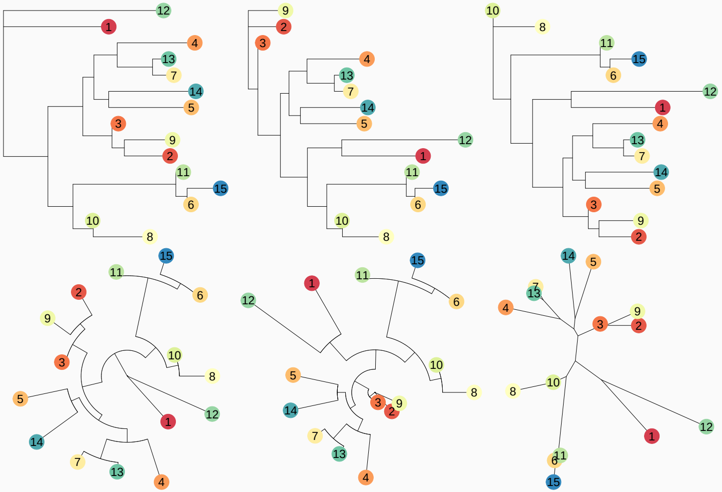
Parsimony

Likelihood/Bayesian

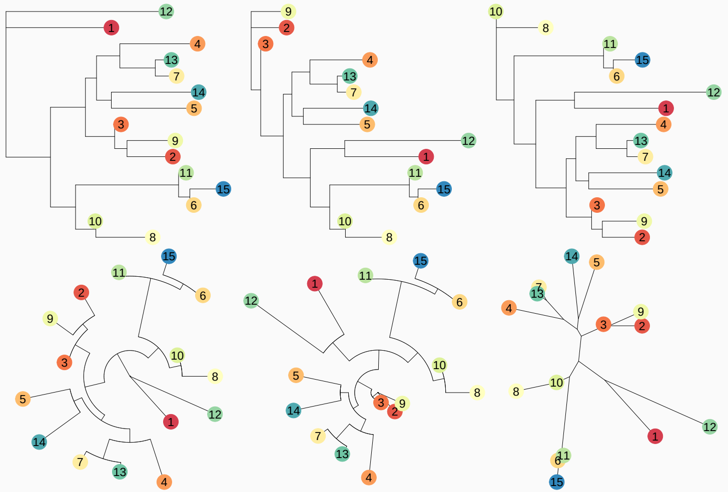
Uncertainty

Pitfalls & best practices

# Plotting trees as rooted

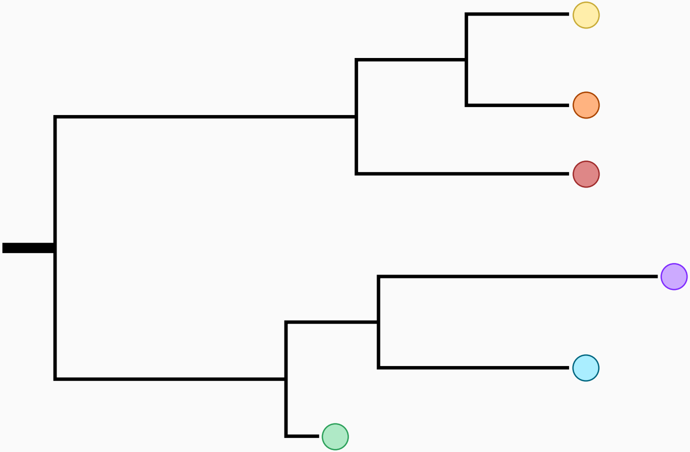


# Plotting trees as rooted

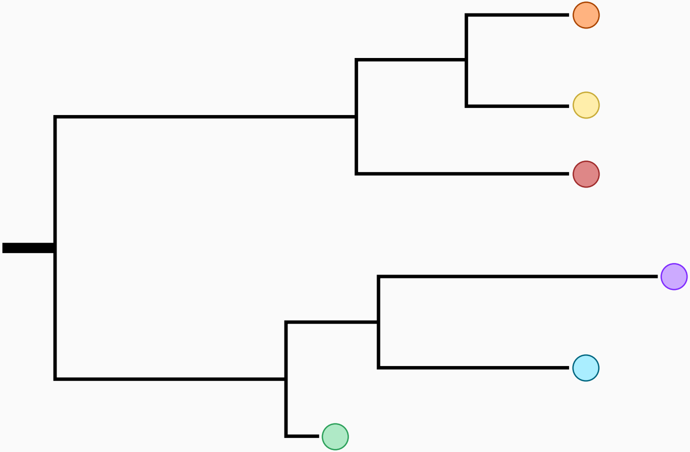


Never plot an unrooted tree as rooted.

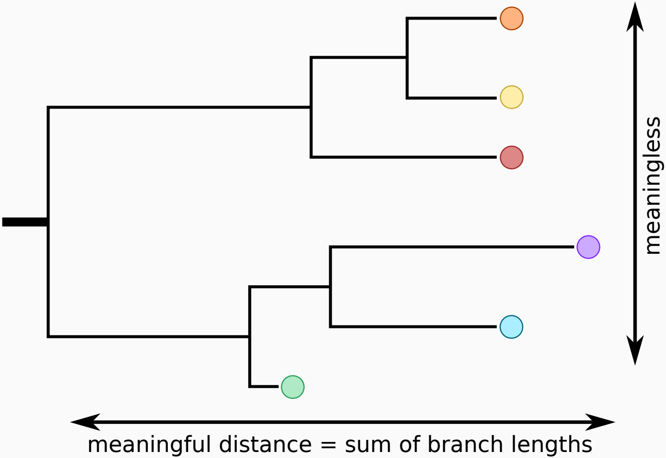
# Interpreting distances



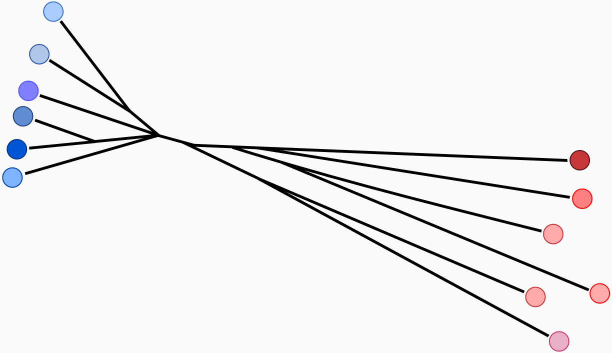
# Interpreting distances



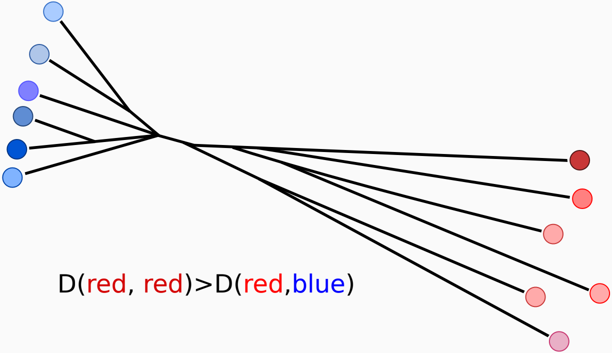
# Interpreting distances



# The paradox of divergent clusters

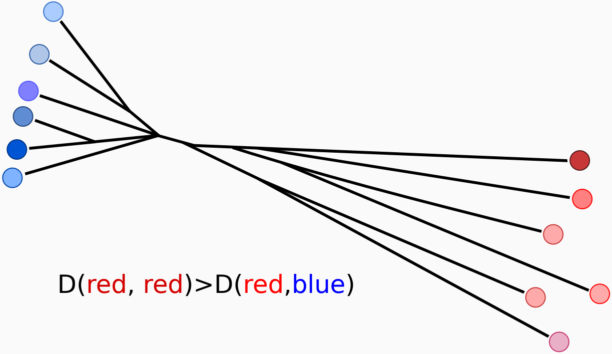


# The paradox of divergent clusters



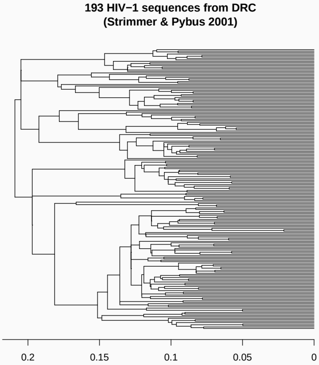
D(red, red)>D(red,blue)

# The paradox of divergent clusters

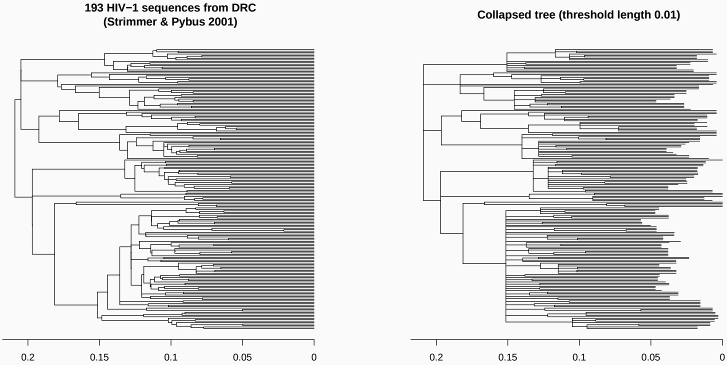


MRCA and genetic distances may give different information.

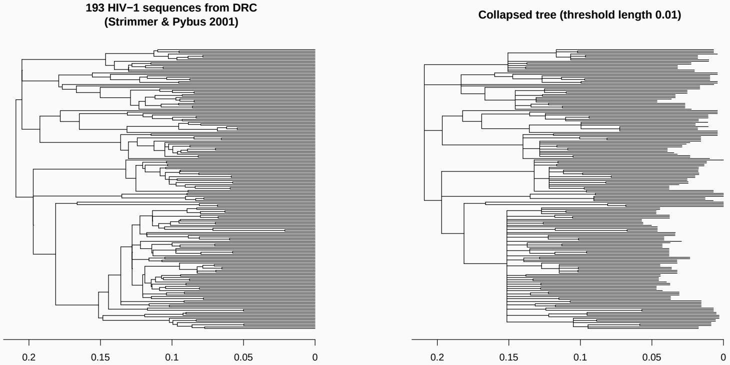
# Taking uncertainty into account



# Taking uncertainty into account

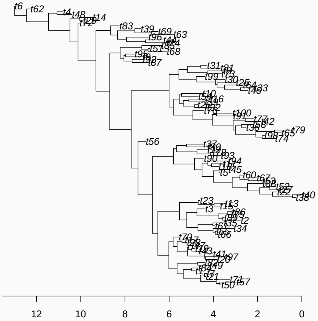


# Taking uncertainty into account

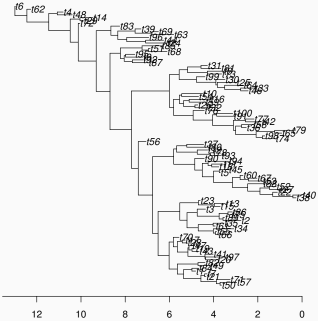


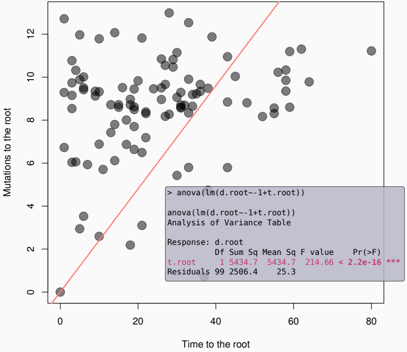
At best, the tree is an estimate of the likely evolutionary history of the taxa studied.

# (Over, Mis)Interpreting temporal trends



# (Over, Mis)Interpreting temporal trends



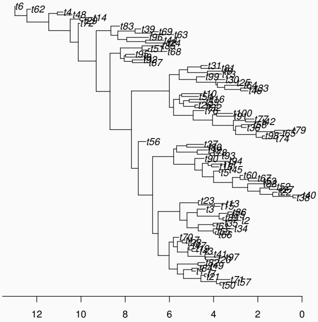


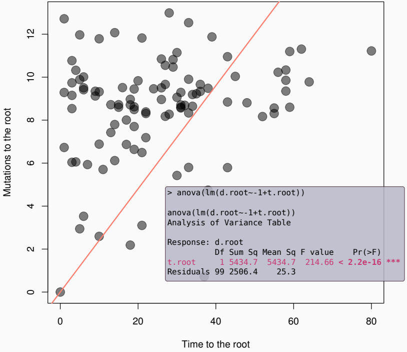
```
> anova(lm(d.root~-1+t.root))

anova(lm(d.root~-1+t.root))
Analysis of Variance Table

Response: d.root
          Df Sum Sq Mean Sq F value    Pr(>F)
t.root     1 5434.7  5434.7  214.66 < 2.2e-16 ***
Residuals 99 2506.4    25.3
```

# (Over, Mis)Interpreting temporal trends





```
> anova(lm(d.root~-1+t.root))

anova(lm(d.root~-1+t.root))
Analysis of Variance Table

Response: d.root
          Df Sum Sq Mean Sq F value    Pr(>F)
t.root     1 5434.7  5434.7  214.66 < 2.2e-16 ***
Residuals 99 2506.4    25.3
```

“Time trees” only make sense under a near-perfect molecular clock.

# Outline

Phylogenies...

Distance trees

Parsimony

Likelihood/Bayesian

Uncertainty

Pitfalls & best practices

# This is only the beginning

Many things can be done with trees

- estimate divergence time

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...
- **and also, other approaches than phylogenetics to analyse genetic data**


The End