Lecture 3: Properties of DNA data and assessing robustness of hypotheses

Jadranka Rota and Niklas Wahlberg

Systematic Biology Group

Department of Biology

Lund University

LUND UNIVERSITY

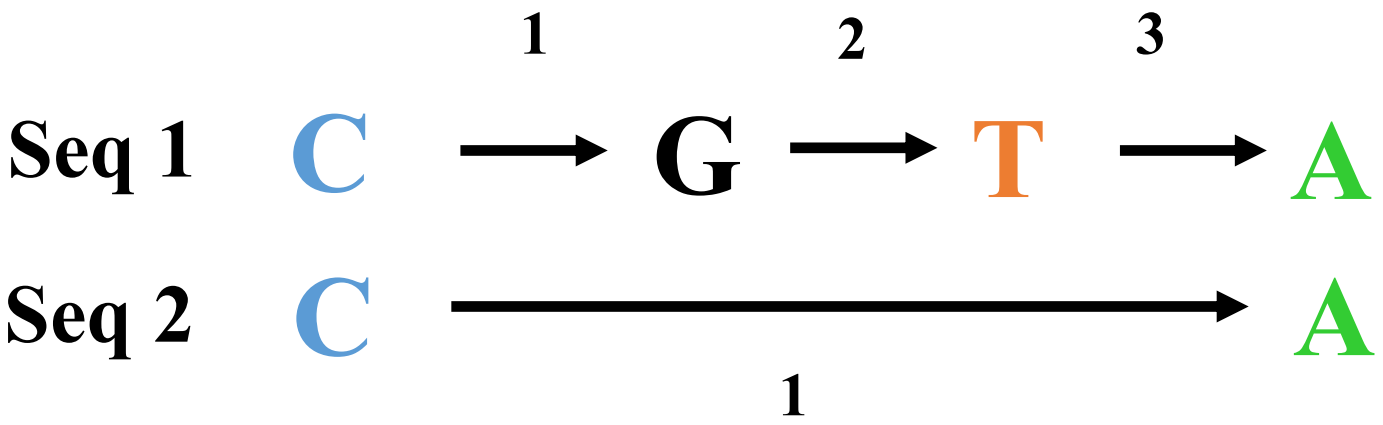
Some slides from Paul Lewis (University of Connecticut, USA)

What are the properties we are interested in?

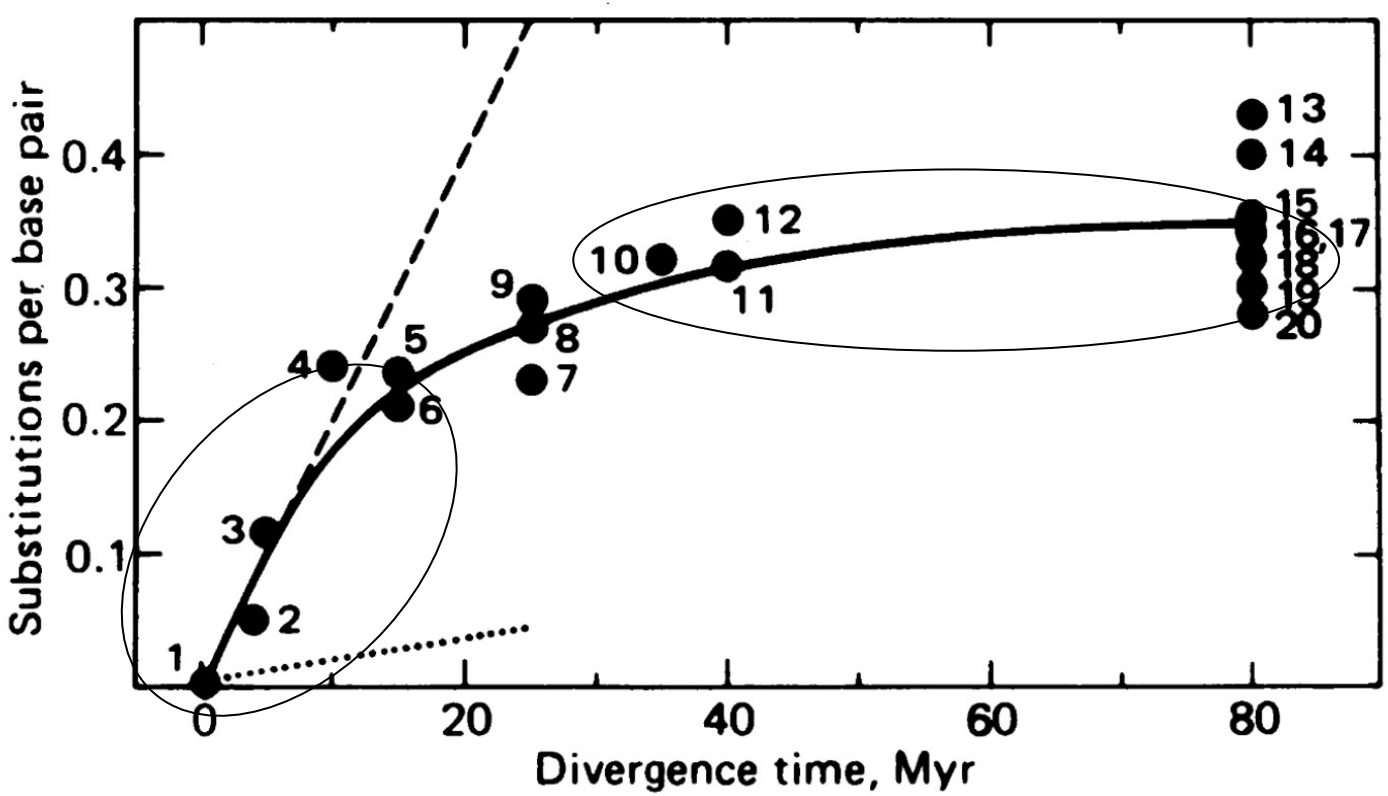
- **Saturation and long-branch attraction**
- **Incomplete lineage sorting**
- **Lateral gene transfer**
- **Mito-nuclear discordance**

Multiple changes at a single site - hidden changes

```
Ancest GGCGCG
Seq 1  AGCGAG
Seq 2  GCGGAC
```



“Multiple hits” or saturation



https://www.pnas.org › doi › pnas.76.4.1967

Rapid evolution of animal mitochondrial DNA. - PNAS

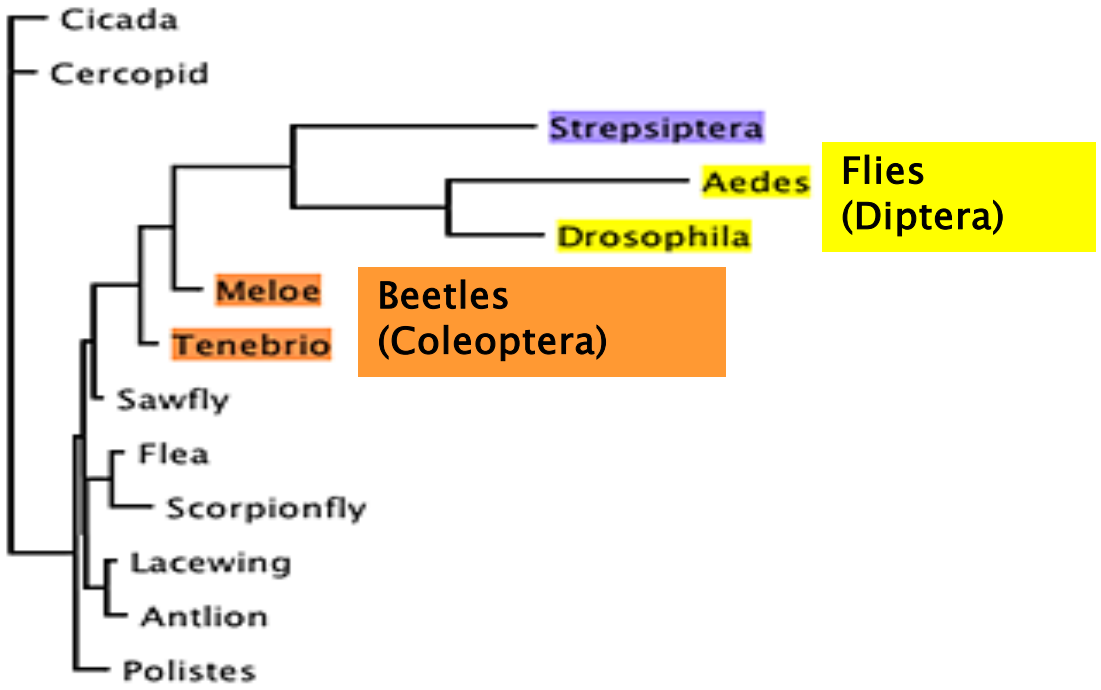
by WM Brown · 1979 · Cited by 4306 — Rapid evolution of animal mitochondrial DNA. W M **Brown**, M George, Jr, and A C WilsonAuthors Info & Affiliations. April 1, **1979**. **76** (4) **1967**-1971.

Brown et al. 1979. PNAS 76:1967

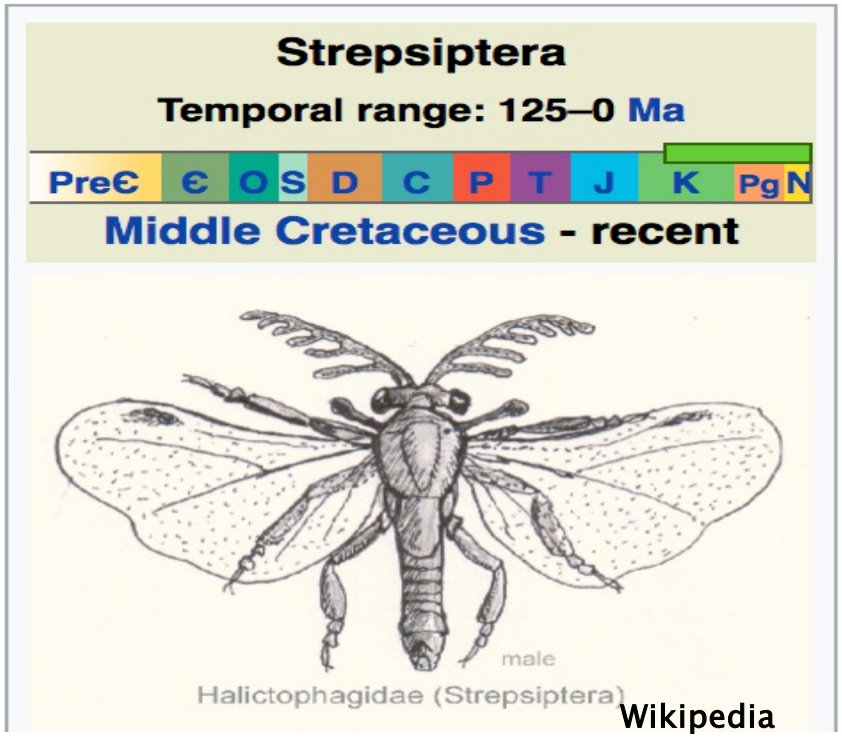
Saturation and long-branch attraction

- **Homoplasy (incorrectly inferred homology)**
  - **Problem with molecular data**
  - **Results from having only four characters (A, C, G, T)**
- **Long-branch attraction (LBA)**
  - **Elevated rates of molecular evolution in unrelated lineages**
  - **Sparse taxon sampling leading to long branches**

Classical LBA example



Based on 18S, 28S, and morphology
(Whiting & Wheeler 1994)



Wikipedia

Photo by C. Fägerström

In 2012, question finally resolved with data from 13 insect genomes (18 mill. nucleotides)

Strepsiptera are sister to beetles

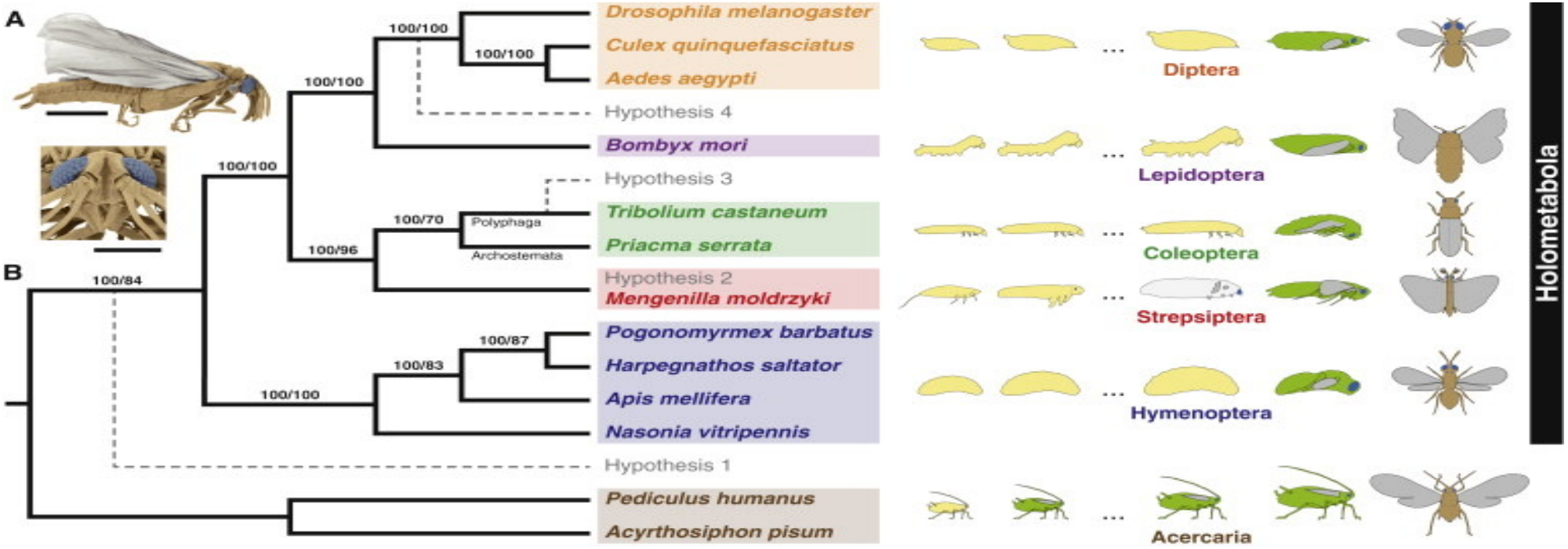


Figure 1. Evolutionary Origin of Twisted-Wing Parasites Inferred from Genomic Evidence(A) Mengenilla moldrzyki male in lateral (top; scale bar represents 1 mm) and frontal (bottom; scale bar represents 500 μm) view (colored SEM micrographs; wings in gray, comp...

Oliver Niehuis, Gerrit Hartig, Sonja Grath, Hans Pohl, Jörg Lehmann, Hakim Tafer, Alexander Donath, Veiko Krauss, Carina Eisenhardt, Jana Hertel, Malte Petersen, Christoph Mayer, Karen Meusemann, Ralph S. Peters...

Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera

Current Biology, Volume 22, Issue 14, 2012, 1309–1313

What can we do about saturation/LBA?

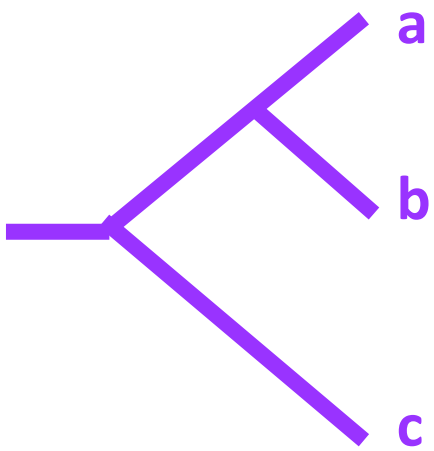
- **Modelling DNA evolution**
- **Taxon sampling is important – whenever possible break up long branches**
- **For divergent taxa with few extant species, this can be a big problem**
  - **BUT branch support is usually low for long branches sticking together in model-based methods – so we should be able to recognize it!**
  - **”sticky” long branches – a bigger problem in parsimony**
- **More data from different sources**
  - **Could be that molecular data are not able to resolve the position of some taxa**
  - **Morphological data!**

Orthology or paralogy?

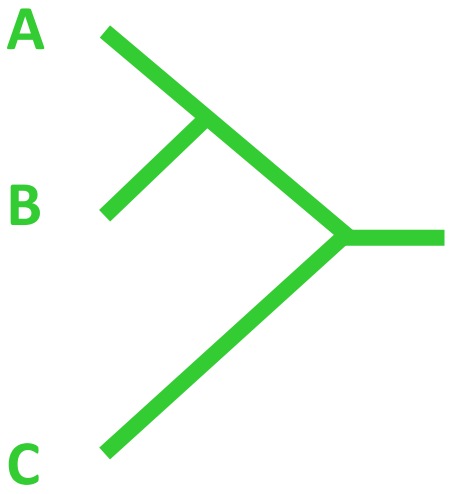
- **Are the genome regions sequenced from different species the same (homologous)?**
- **Gene duplication**
  - **1) duplicate gene degenerates - pseudogene**
  - **2) duplicate gene aquires new function**
- **A problem particularly accute currently as we analyze phylogenomic data**

Orthology:
gene trees and species trees

Gene phylogeny



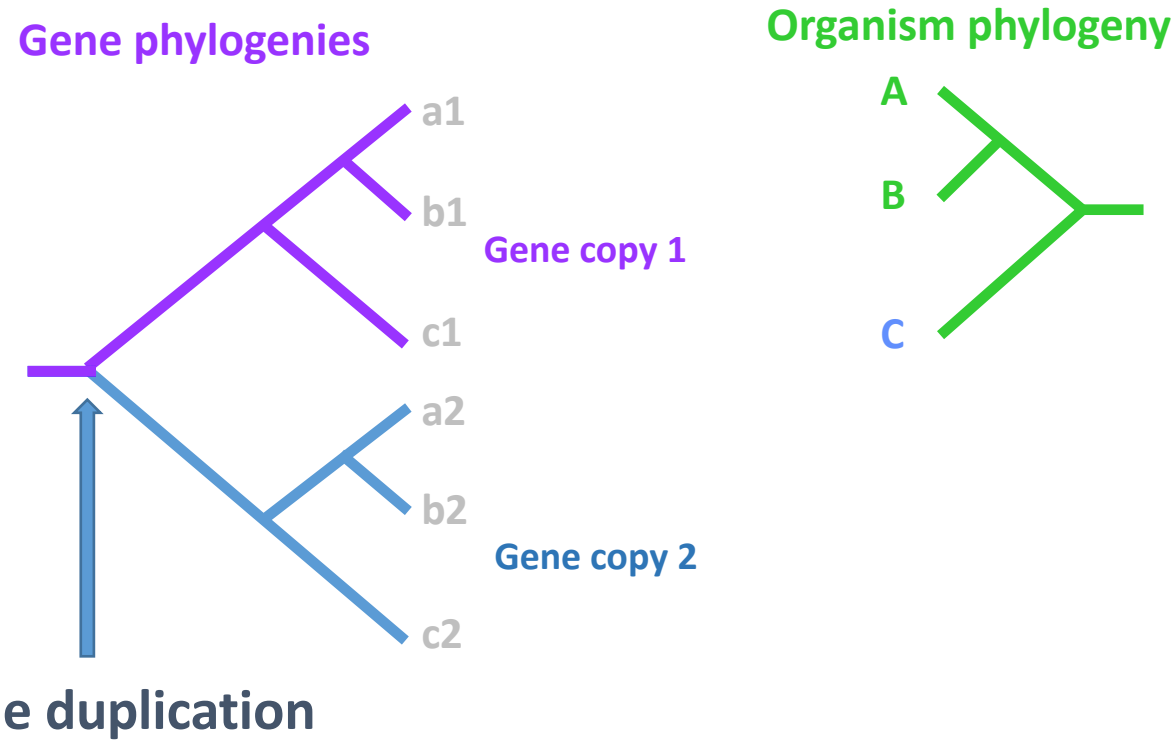
Organism phylogeny



ORTHOLOGY

Orthologs: genes that arose due to speciation

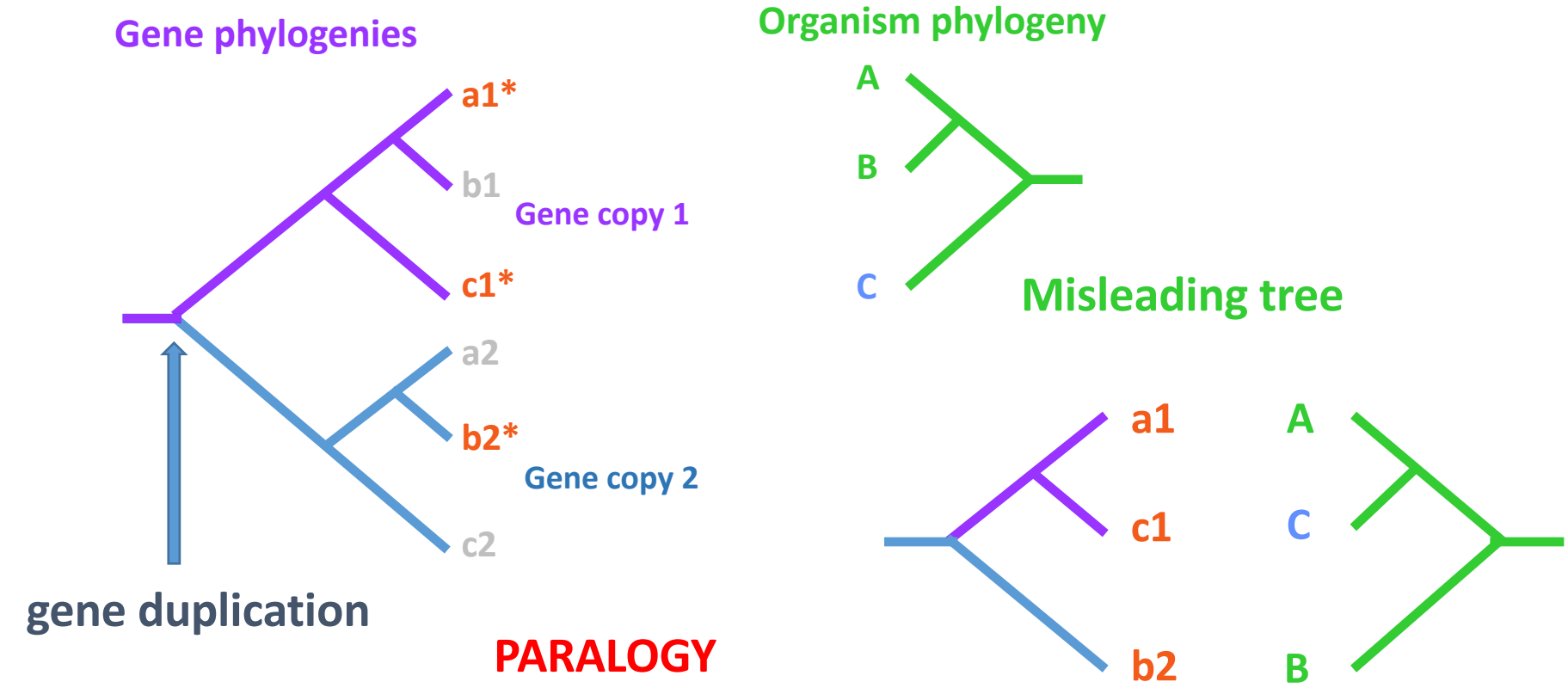
Paralogy: can produce misleading trees



PARALOGY

Paralogs: genes that arose due to duplication events

Paralogy: can produce misleading trees



Paralogs: genes that arose due to duplication events

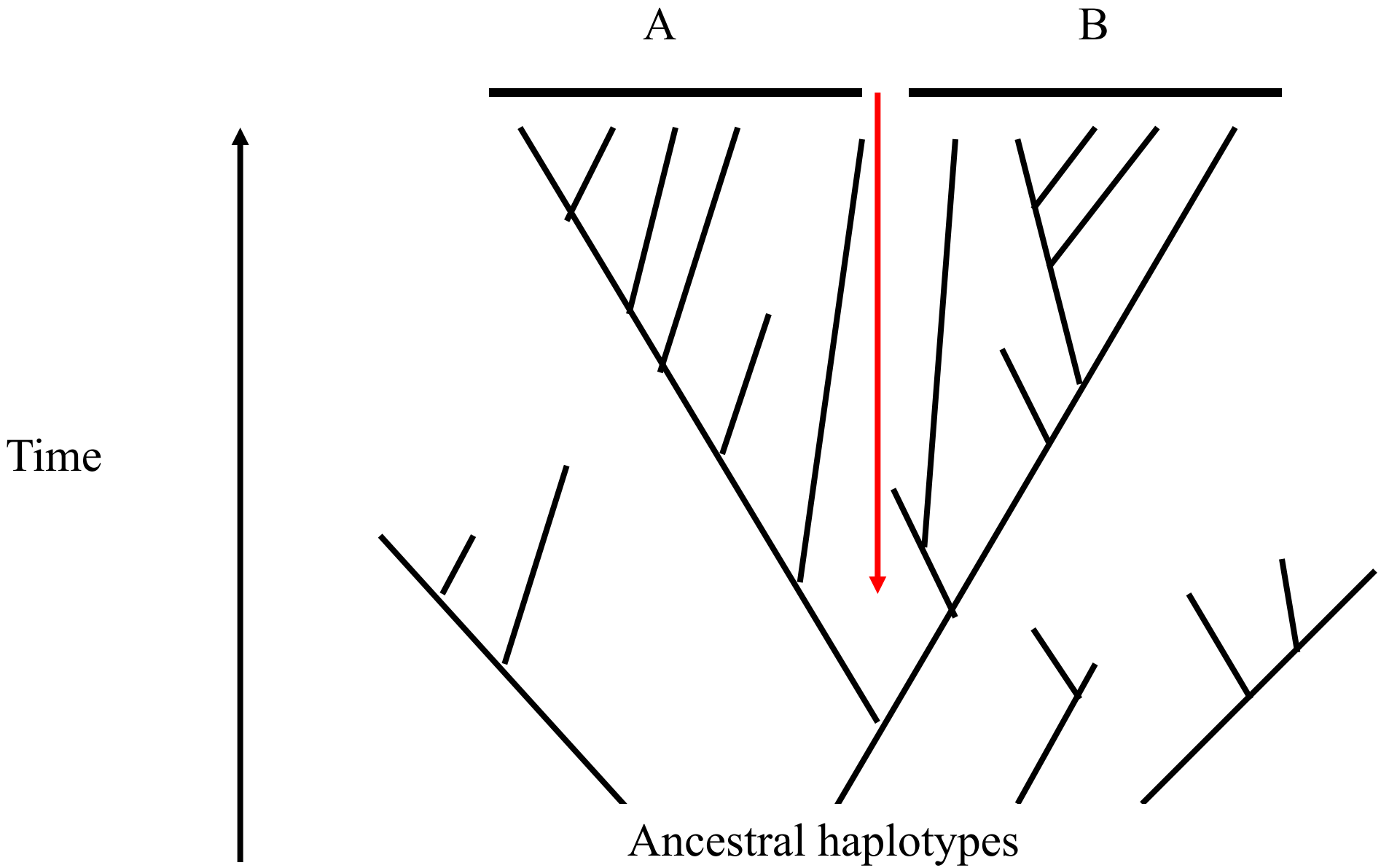
Incomplete lineage sorting (ILS)

- **Gene trees may not be the same as species trees**
- **Usually not a problem for deep phylogenies BUT…**
- **Extant populations may retain ancestral polymorphisms**
- **Species level phylogenies should never sample single individuals of different species**
  - **Sample several individuals from across the range**

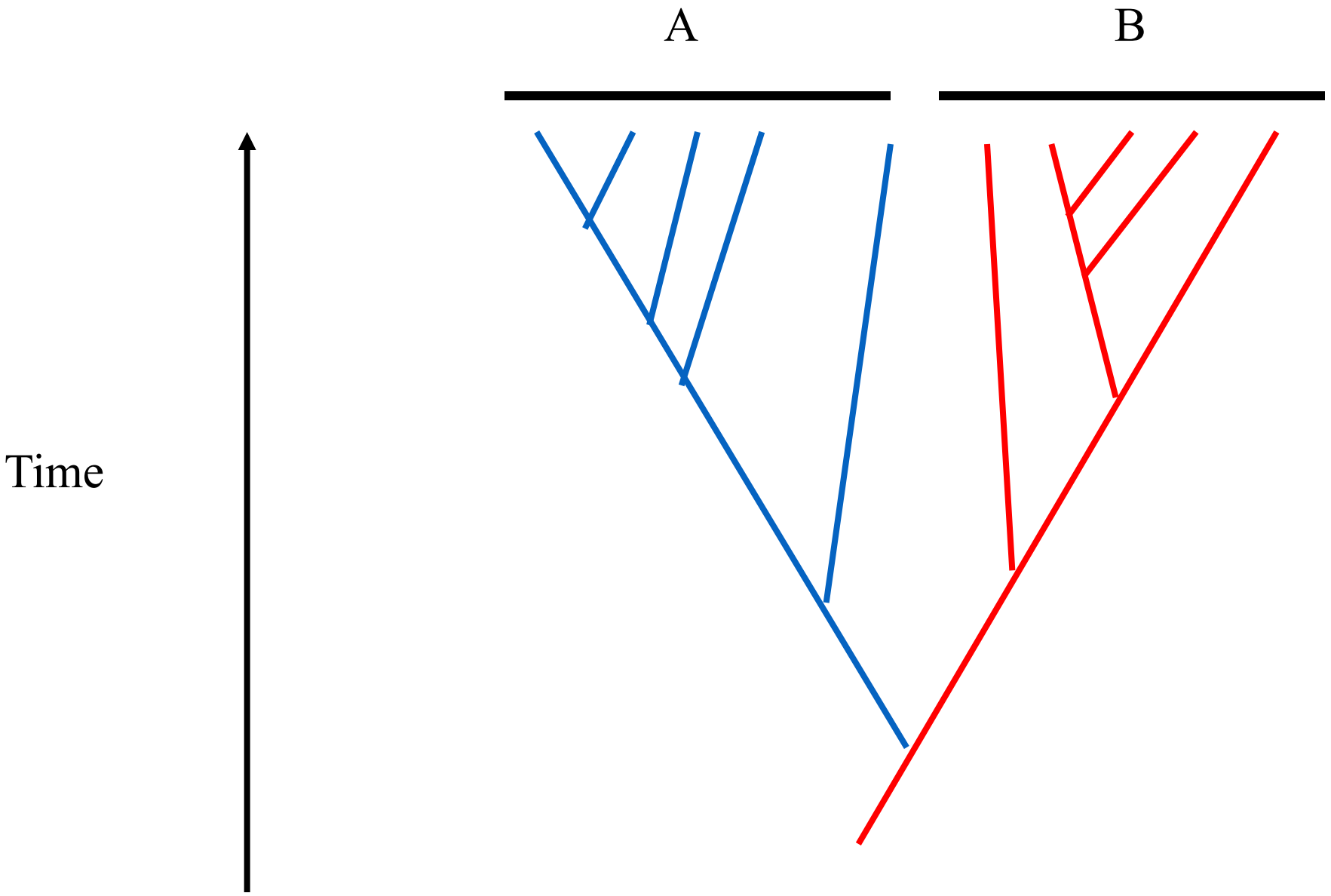
Are species monophyletic?

- **Implicit assumption in many studies using mtDNA – DNA barcoding**
- **Theoretical studies predict that DNA lineages pass through several phases in evolution of a species**

The assumption: monophyly



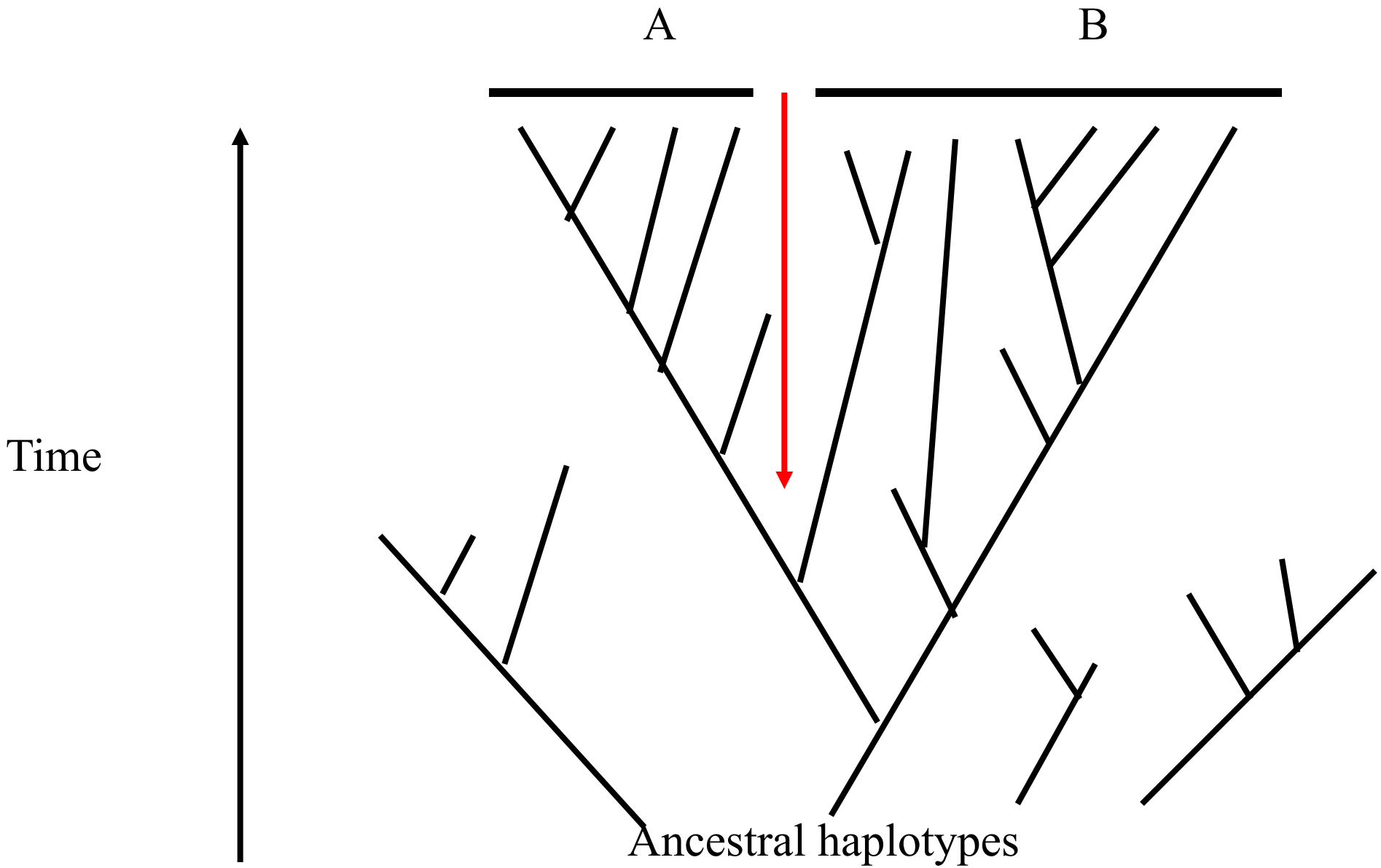
The assumption: monophyly



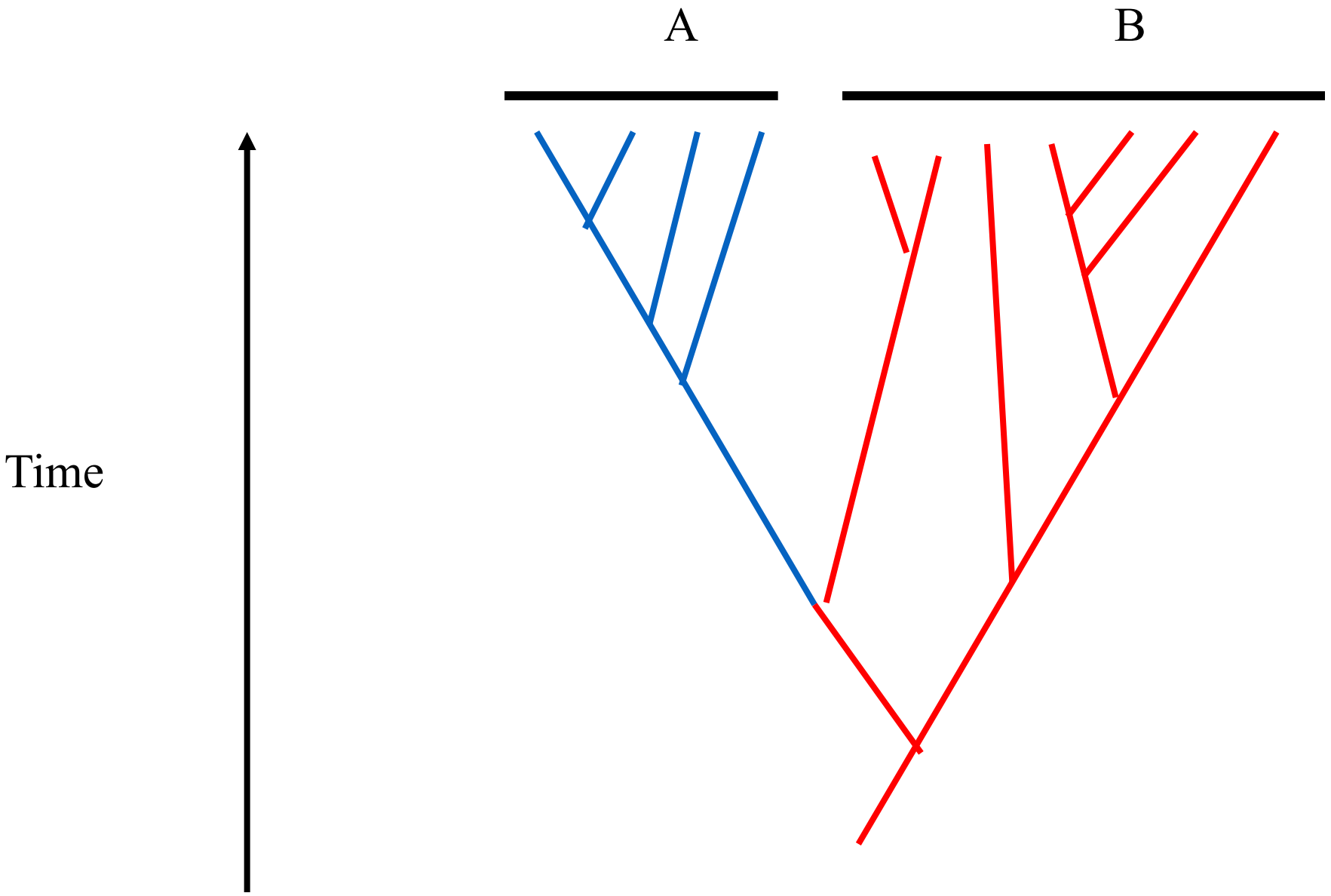
The presence of poly- and paraphyletic lineages

- **Paraphyly can occur when one population in a set of locally panmictic populations speciates**
- **Polyphyly occurs when a highly polymorphic population is subdivided**
- **Can be highly informative of the history of divergence**
  - **i.e., how speciation occurred**

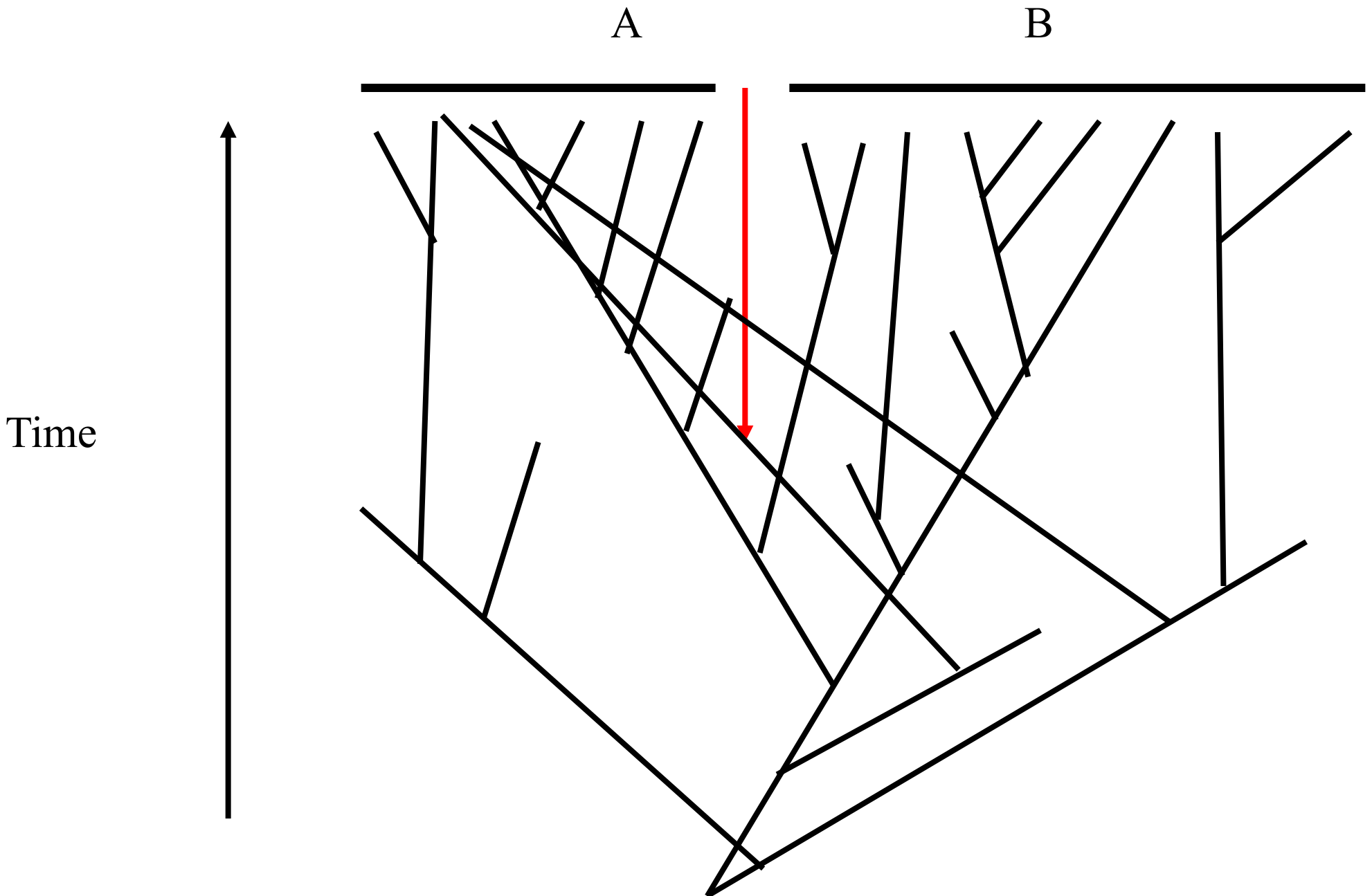
Paraphyly



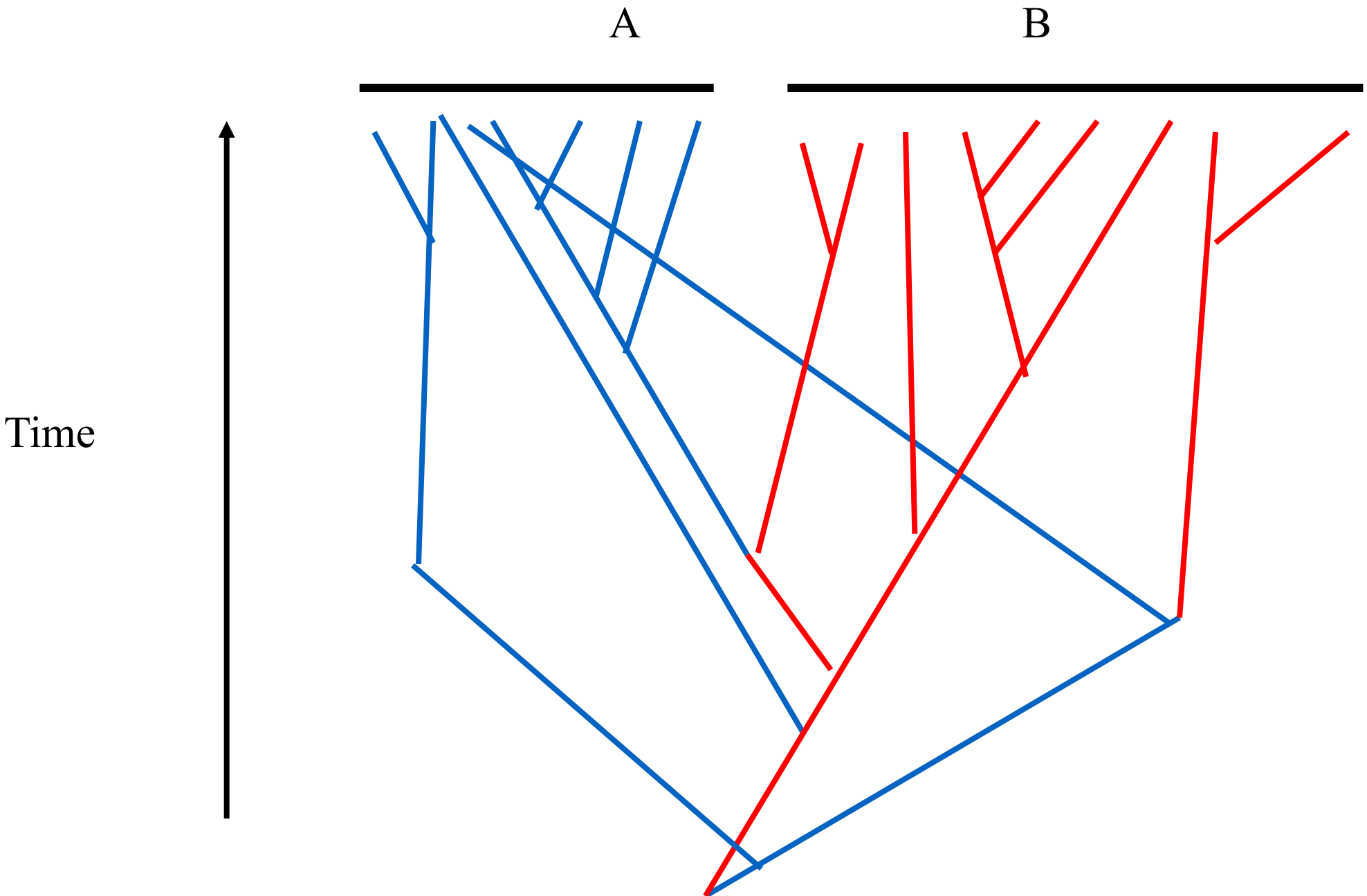
Paraphyly



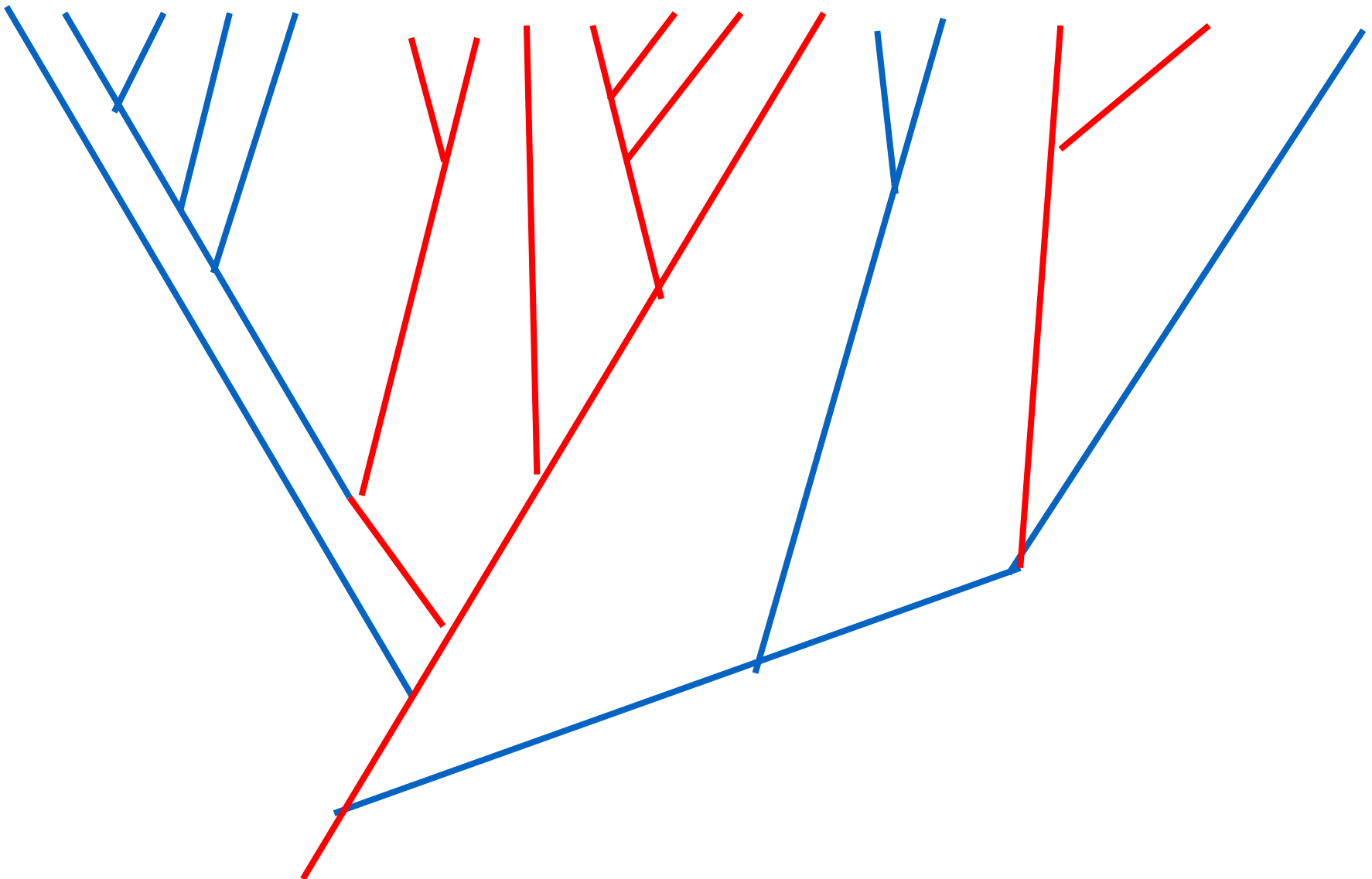
Polyphyly



Polyphyly



Polyphyly



Paraphyly of a species can be due to incomplete lineage sorting and/or secondary gene flow

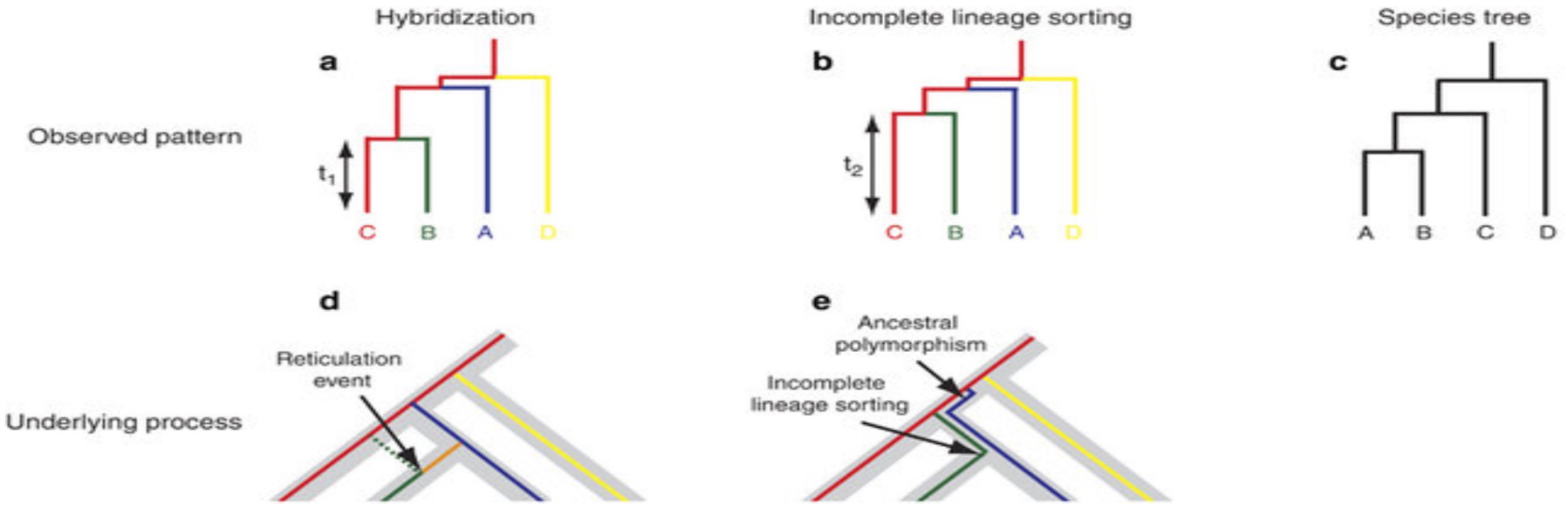
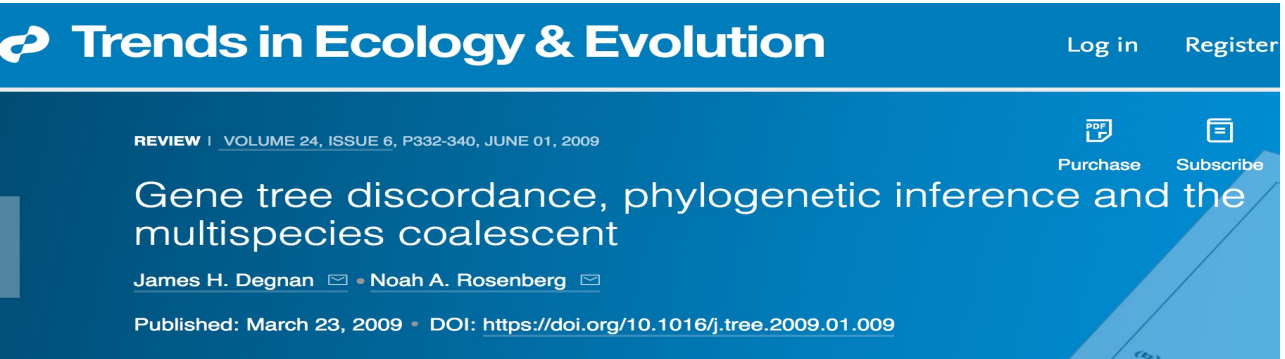
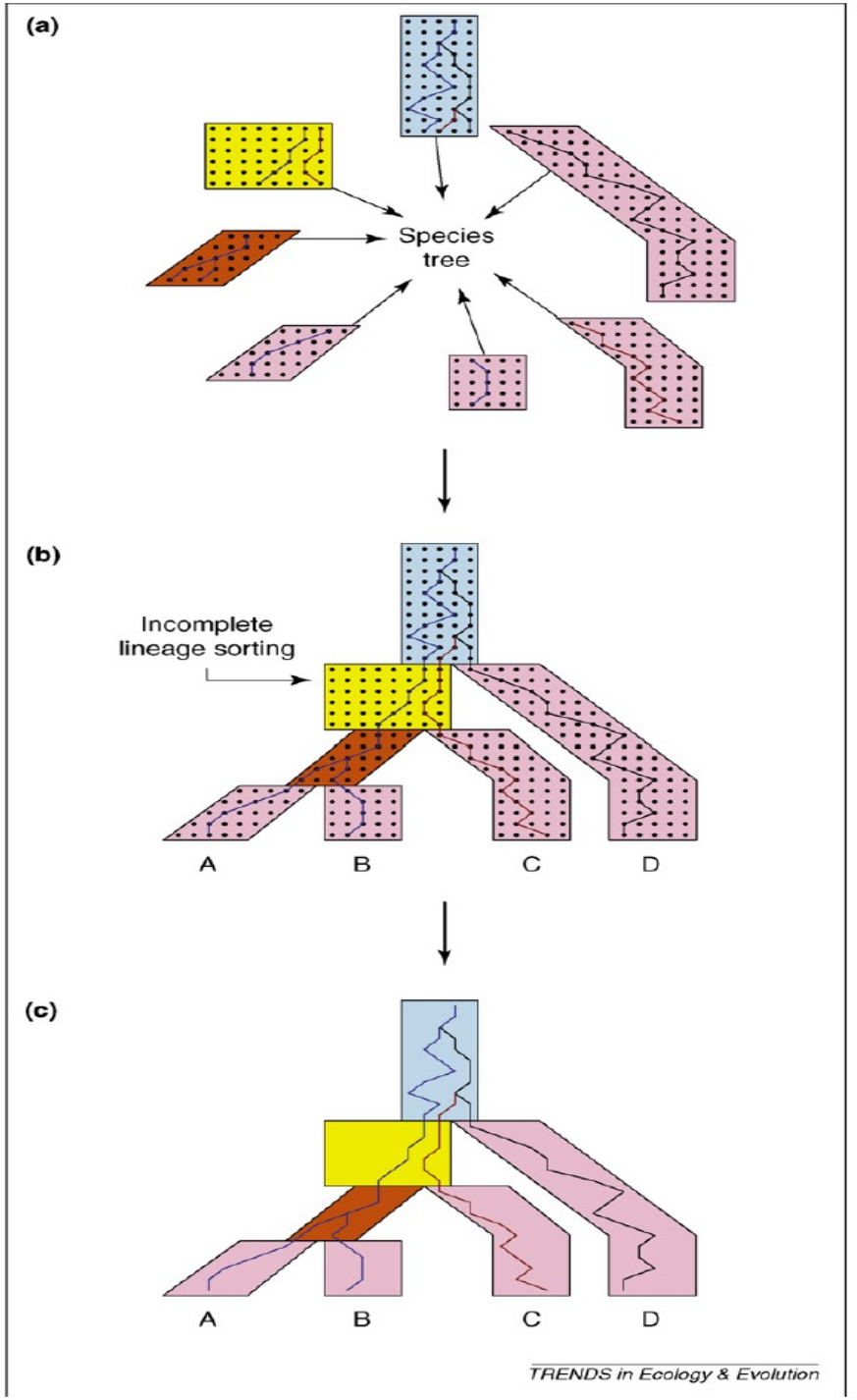


Figure from ResearchGate, uploaded by Richard A Ennos

Multispecies coalescent model (MSC)

- **Gene tree vs. species tree**
- **Model that accommodates gene tree heterogeneity caused by ILS**



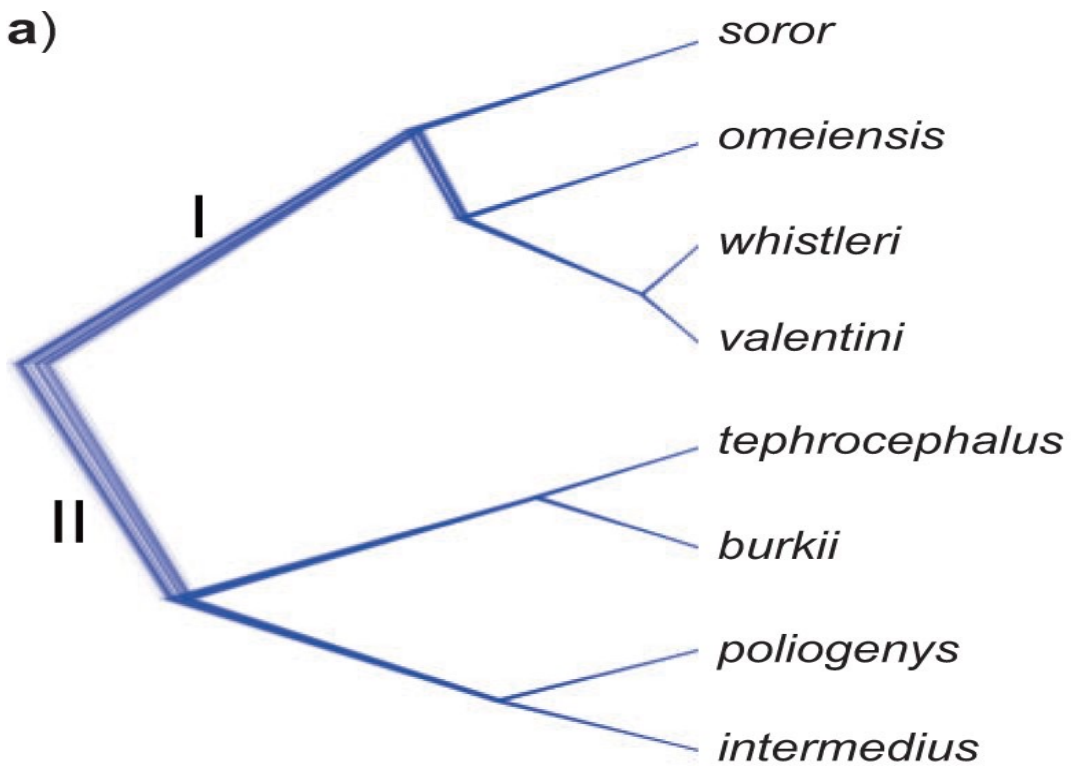


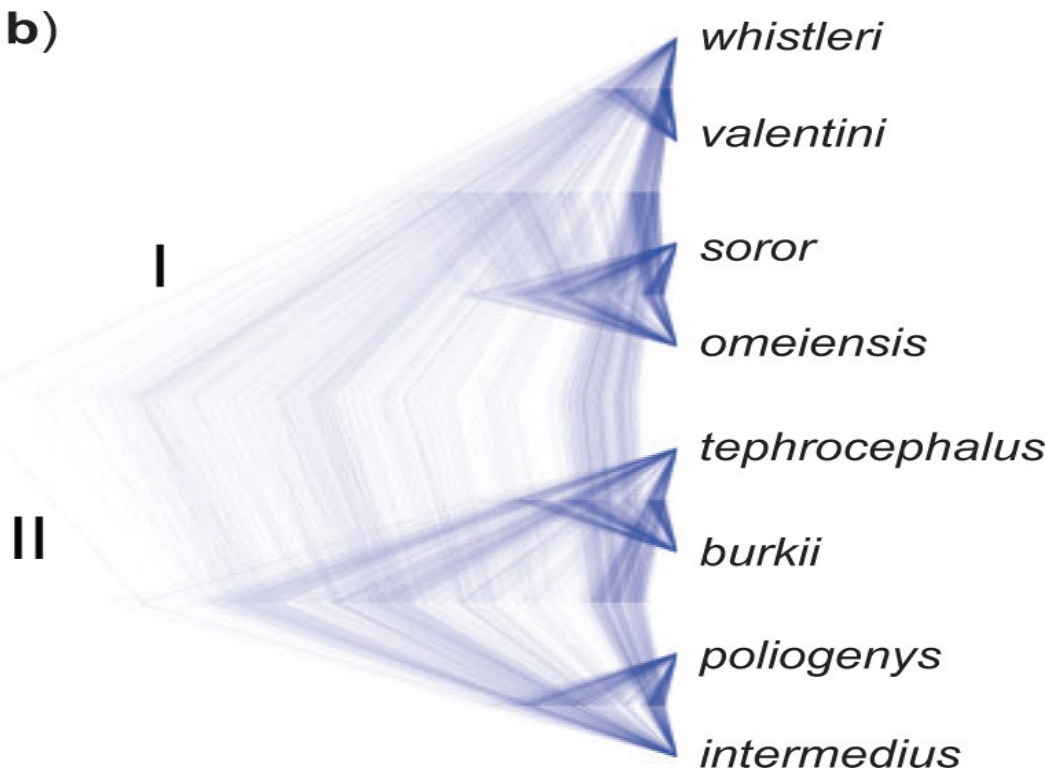
Lateral (=Horizontal) Gene Transfer

- **Widespread in single-celled organisms**
  - **Even between distantly related lineages**
- **In multi-celled organisms more a problem in closely related species**
  - **It happens through hybridization**
  - **Some estimates suggest that 25% of plant species and 10% of animal species hybridize (Mallet 2005 TREE 20(5):229-237)**

Figure 4. Species trees estimated using SNAPP (including five samples per species). a) Species tree based on [one subset of sequence data and b) species tree based on [a different subset]… All nodes are supported by a posterior probability of 1.00.

Leaf warblers





Syst Biol, Volume 70, Issue 5, September 2021, Pages 961–975, https://doi.org/10.1093/sysbio/syab024

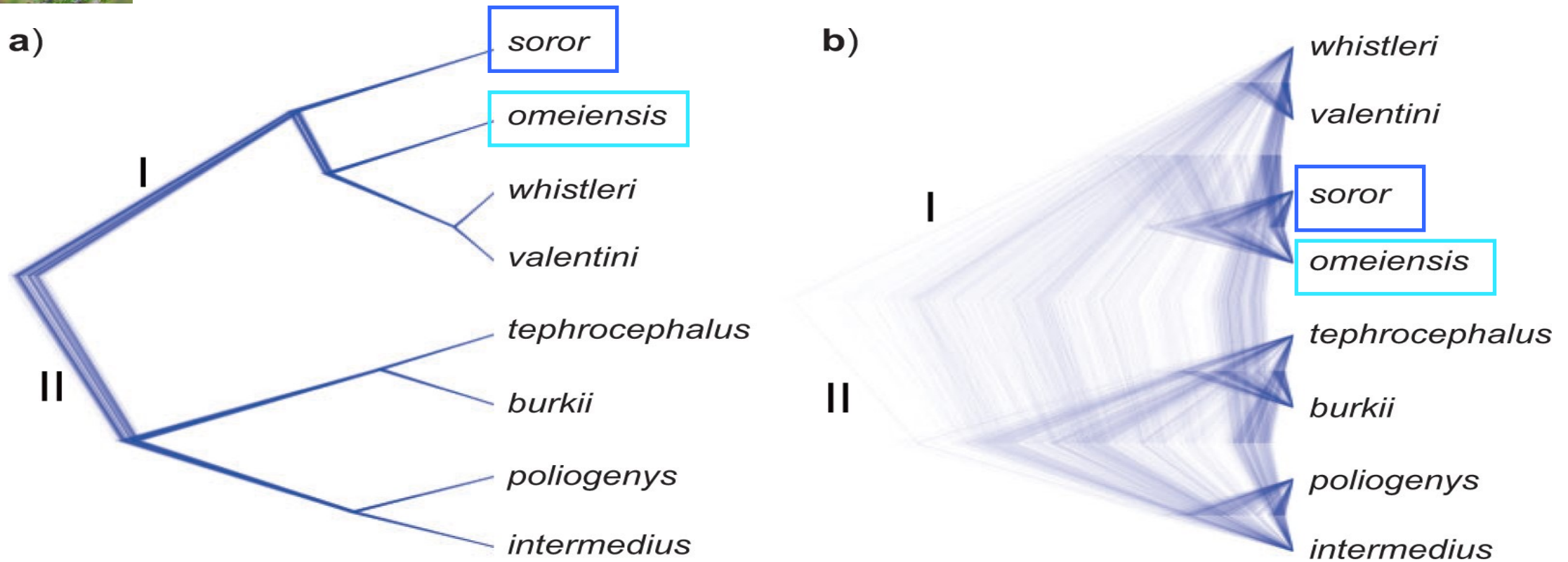
Most Genomic Loci Misrepresent the Phylogeny of an Avian Radiation Because of Ancient Gene Flow

Dezhi Zhang, Frank E Rheindt, Huishang She, Yalin Cheng, Gang Song, Chenxi Jia, Yanhua Qu, Per Alström, Fumin Lei

Systematic Biology, Volume 70, Issue 5, September 2021, Pages 961–975, https://doi.org/10.1093/sysbio/syab024
Published: 31 March 2021 **Article history**

Figure 4. Species trees estimated using SNAPP (including five samples per species). a) Species tree based on [one subset of sequence data and b) species tree based on a different subset]… All nodes are supported by a posterior probability of 1.00.

Leaf warblers



Syst Biol, Volume 70, Issue 5, September 2021, Pages 961–975, https://doi.org/10.1093/sysbio/syab024

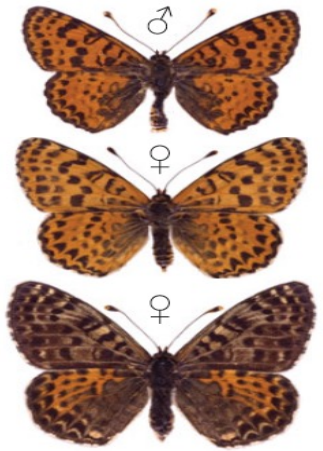
Most Genomic Loci Misrepresent the Phylogeny of an Avian Radiation Because of Ancient Gene Flow

Dezhi Zhang, Frank E Rheindt, Huishang She, Yalin Cheng, Gang Song, Chenxi Jia, Yanhua Qu, Per Alström ✉, Fumin Lei ✉

Systematic Biology, Volume 70, Issue 5, September 2021, Pages 961–975, https://doi.org/10.1093/sysbio/syab024
Published: 31 March 2021 **Article history**

Mito-nuclear discordance

An empirical example: *Melitaea* butterflies



The conundrum of species delimitation: a genomic perspective on a mitogenetically super-variable butterfly

Vlad Dincă†, Kyung Min Lee†, Roger Vila and Marko Mutanen

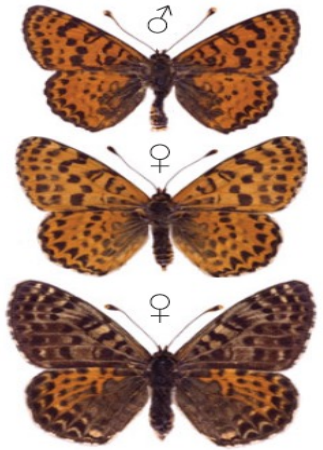
Published: 18 September 2019
https://doi.org/10.1098/rspb.2019.1311

Dinca et al. (2019)

Proc. Roy. Soc B,

286: 20191311

An empirical example: *Melitaea* butterflies



The conundrum of species delimitation: a genomic perspective on a mitogenetically super-variable butterfly

Vlad Dincă†, Kyung Min Lee†, Roger Vila and Marko Mutanen

Published: 18 September 2019
https://doi.org/10.1098/rspb.2019.1311

Dinca et al. (2019)

Proc. Roy. Soc B,

286: 20191311

Properties of molecular data?

- **These properties are highly interesting phenomena in themselves!**
- **When taking the different factors into account, can be informative about evolutionary history**
- **”When in doubt, get more data”**
  - **Brooks and McLennan 2002**
- **And then think about how to analyse your data given these properties**

How good is our phylogenetic hypothesis?

Support and stability

Assessing phylogenetic hypotheses and signal in phylogenetic data

- **Inferring a tree is not enough**
  - **We also need to know how much support there is for our phylogenetic hypothesis in the data**
  - **How much confidence can we place in the phylogenetic hypothesis?**
  - **Do the data strongly support the relationships?**
  - **If not, we may end up drawing wrong conclusions about how evolution proceeded**

Support and stability

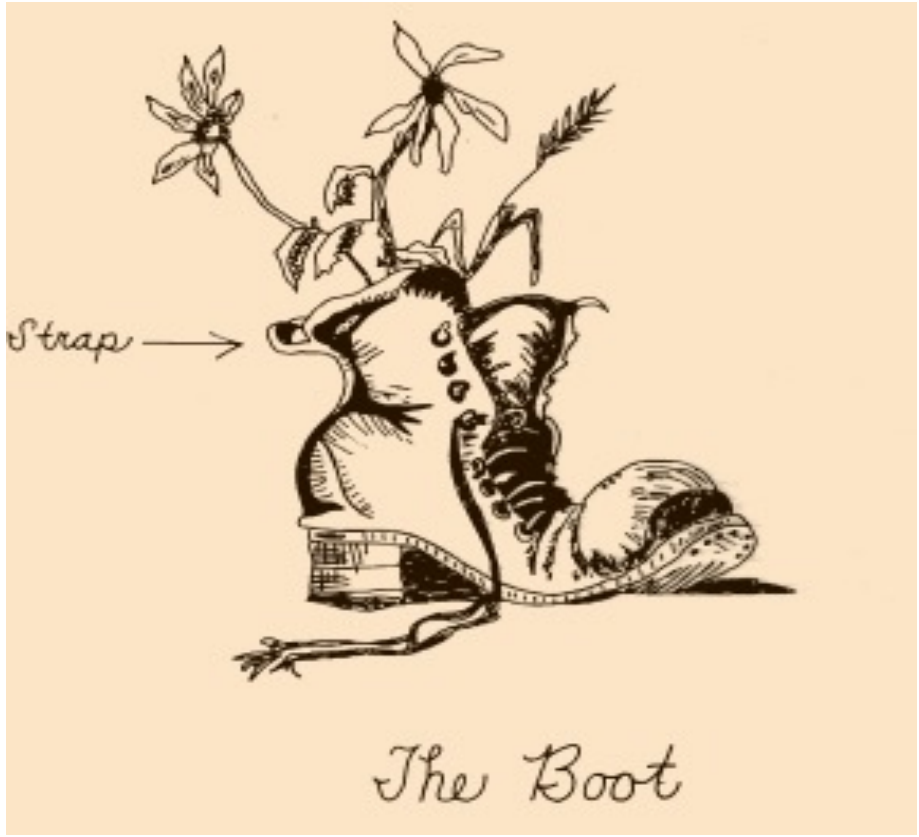
- **How strongly do the data support your phylogenetic hypothesis?**
- **How stable is your phylogenetic hypothesis?**
  - **Is it likely to change with the addition of new data?**
  - **Do you get the same result with different analysis methods?**

Assessing phylogenetic hypotheses - Support

- **Several methods provide some measure of the strength of support for tree nodes**
  - **Nodal or branch support**
- **These methods include:**
  - **Character resampling methods - bootstrap and jackknife**
  - **Posterior probability in Bayesian analysis**

Bootstrapping

- **Statistical technique that uses random resampling of data to determine sampling error or confidence intervals for some estimated parameter**
- **Introduced into phylogenetics by Felsenstein (1985)**



Bootstrapping phylogenies

- **Characters are resampled with replacement to create many bootstrap pseudoreplicate data sets**
  - **Often 1000 pseudoreplicates done**
- **Each bootstrap data set is analysed**
- **Agreement among the resulting trees is summarized with a majority-rule consensus tree**
- **Frequencies of occurrence of groups, bootstrap proportions (BPs), are a measure of support for those groups**

Bootstrap

Original alignment

```
1 CGAGAC
2 AGCGAC
3 AGATTC
4 GGATAG
```

Pseudoreplicate 1

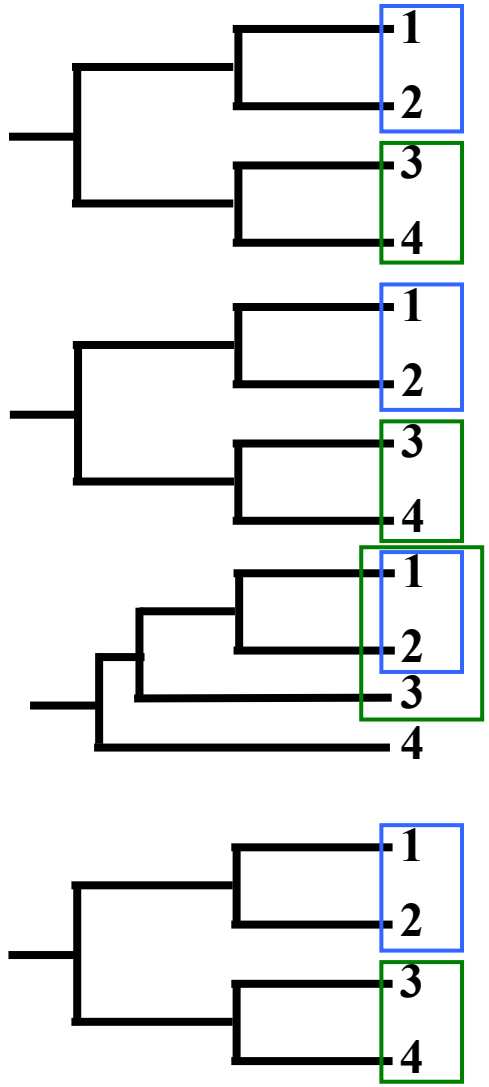
```
1 CGAGAA
2 AGAGAA
3 AGTTTT
4 GGATAA
```

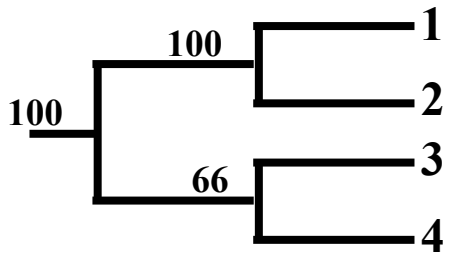
Pseudoreplicate 2

```
1 AGAGAC
2 AGCGCC
3 TGATAC
4 AGATAG
```

Pseudoreplicate n (e.g. 1000)

```
1 CCAGAC
2 ACCGAC
3 ACAGTC
4 GGAGAG
```

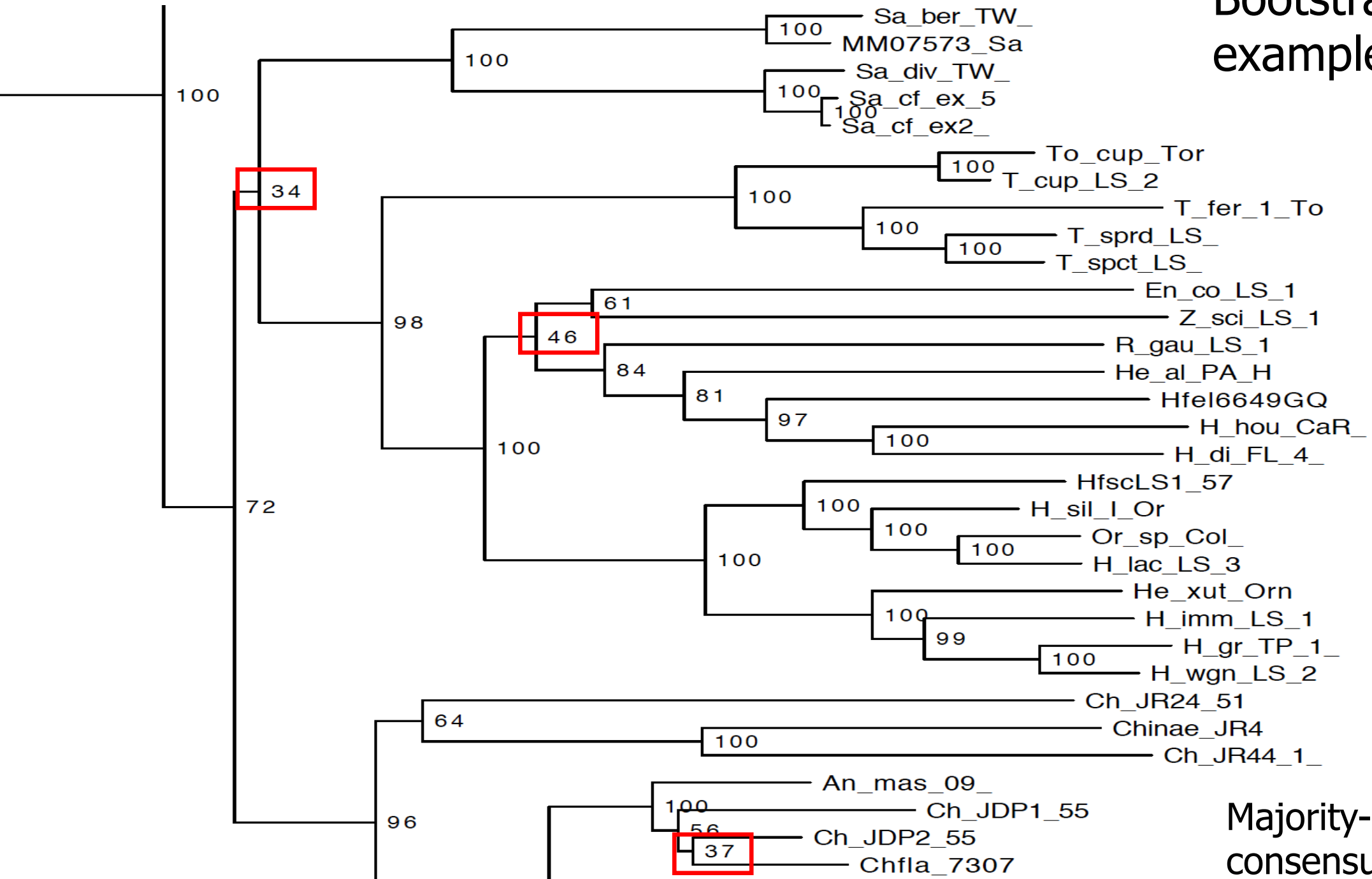




Bootstrap values superimposed on majority-rule consensus tree

Bootstrapping doesn't really assess the accuracy of a tree, it only indicates the consistency of the data

Bootstrapping – an example



Majority-rule consensus tree

Bootstrap - interpretation

- **Hillis & Bull 1993**
  - **Examined interpretation of bootstraps using simulated data & known phylogenies**
  - **Conclusions:**
    - **Low bootstraps overestimate accuracy**
    - **High bootstraps underestimate accuracy - bootstrap = 70% was statistically significant support (only applies to their simulated data)**
  - **Done on small datasets (few genes); phylogenomic datasets seem to inflate bootstraps – a higher number is needed to be considered significant**

Other branch support measures

- **Ultrafast bootstrap (Nguyen et al. 2015)**
  - **10 to 40 times faster than RAxML rapid bootstrap and obtains less biased support values**
  - **Different interpretation from the usual bootstrap**
  - **These support values are more unbiased: 95% support correspond roughly to a probability of 95% that a clade is true**
  - **>=95% is significant**

L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, and B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol., 32:268-274. DOI: 10.1093/molbev/msu300
D.T. Hoang, O. Chernomor, A. von Haeseler, B.Q. Minh, and L.S. Vinh (2018) UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol., 35:518–522. DOI: 10.1093/molbev/msx281

Other branch support measures

- **SH-aLRT branch test**
  - **Shimodaira-Hasegawa approximate likelihood ratio test**
  - **>=80% is significant**
  - **Robust to various model assumption violations**

Guindon et al. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. 59:307–321. DOI: 10.1093/sysbio/syq010
Anisimova et al. (2011) Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Syst. Biol. 60(5):685–699. DOI:10.1093/sysbio/syr041

ML *vs*. Bayesian view

- **ML maximizes probability of data, given the model/parameter values (incl. toplogy and branch lengths).**
  - **Confidence is measured by bootstrap**
- **Bayesian inference - probability of topology with branch lengths and other parameters, given the data and the model**
  - **Confidence given by posterior probabilities**
- **Ronquist & Deans 2010. Ann. Rev. Ent.**

Bayesian Inference: summarizing posterior trees

- **50% Majority consensus tree**
  - **Contains all clades occurring in at least 50% of the trees in the posterior distribution (=stationary distribution)**
  - **Branch support = frequency of each clade in the posterior distribution of the trees – posterior probabilities (PPs)**
  - **Interpreation: an estimate (approximation) of the probability that a certain branch exists, given the data, the model, the priors**

Branch support: summary

- **Support for your phylogenetic hypothesis**
  - **Quantitative measures from the data that you have**
- **Measures of support tend to be correlated with each other**
  - **But PPs can sometimes be much higher than bootstraps and *vice versa***
  - **Such differences in branch support may indicate that the signal in the data is misleading in some way**

What is significant support?

- **Bootstrap: weak support 50–70%; 70–85% medium; >85% strong**
- **Boostraps in phylogenomic datasets – 100% good support, not so sure about anything below 100%**
- **UFBS >=95% significant**
- **SH-aLRT >=80% significant**
- **PPs: ≈0.95 weak support; 1.00 strong support**

Stability of the hypothesis

- **How stable is your phylogenetic hypothesis to changing the assumptions of the analysis?**
- **Does choice of model have an effect on your results?**
  - **Simple models *vs*. more complex models**
  - **Unpartitioned *vs*. partitioned**
  - **How sensitive is your hypothesis to the parameter values estimated (precise *vs*. imprecise estimates)**
- **Does choice of method have an effect – e.g. ML *vs*. Bayesian?**

Recommended reading

- **Christoph Bleidorn (2017) Phylogenomics: An Introduction (DOI: 10.1007/978-3-319-54064-1)**
- **Yang & Rannala. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13, 303-314. doi:10.1038/nrg3186**
- **Nascimento, dos Reis & Yang. 2017. A biologist's guide to Bayesian phylogenetic analysis. Nature Ecology & Evolution 1, 1446–1454. doi:10.1038/s41559-017-0280-x**
- **Baum & Smith. 2013. Tree Thinking: An Introduction to Phylogenetic Biology. W.H. Freeman, New York.**