Molecular Phylogenetics Course

1. How much data to use?
2. Problems inherent to molecular data
3. Assessing hypotheses

Jadranka Rota

Data: how much is needed?

more sequence or more individuals, tens of genes or thousands of gene?

How much data?

- All extant species?
- The whole genome?
- Impractical?
- Trade-off between more genes and more taxa
- Think about your study/question
  - How deep in time does your phylogeny go?
    - Deeper phylogenies require more sequence data
  - What are you going to do with the phylogeny?
    - Change classification, infer historical biogeography, study character evolution, …?

Choosing taxa or data

- Know your group – which taxa are the most relevant for your study?
- Know what gene sequences are available from previous studies
  - Databases: GenBank, BOLD
  - So you're not duplicating efforts

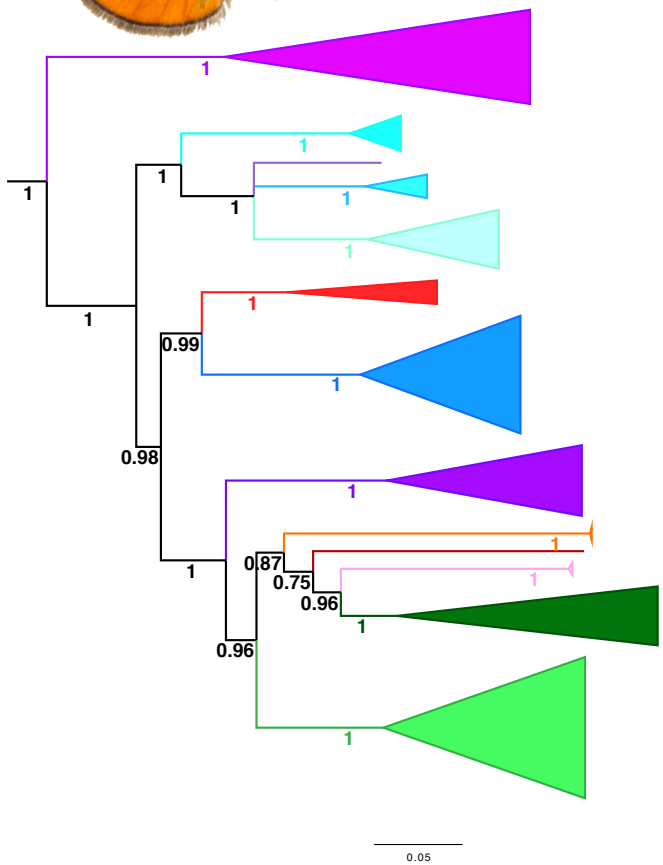
Number of genes

- **Single gene datasets – not very good for resolving phylogenies**
  - **Very rare nowadays**
- **Mitochondrial and chloroplast DNA popular because easy to amplify and sequence**
  - **But they have some inherent problems**
- **Nuclear genes - worth increasing their number**
  - **Can evolve independently from each other**
  - **When different nuclear genes give the same phylogeny, our confidence in the hypothesis grows**

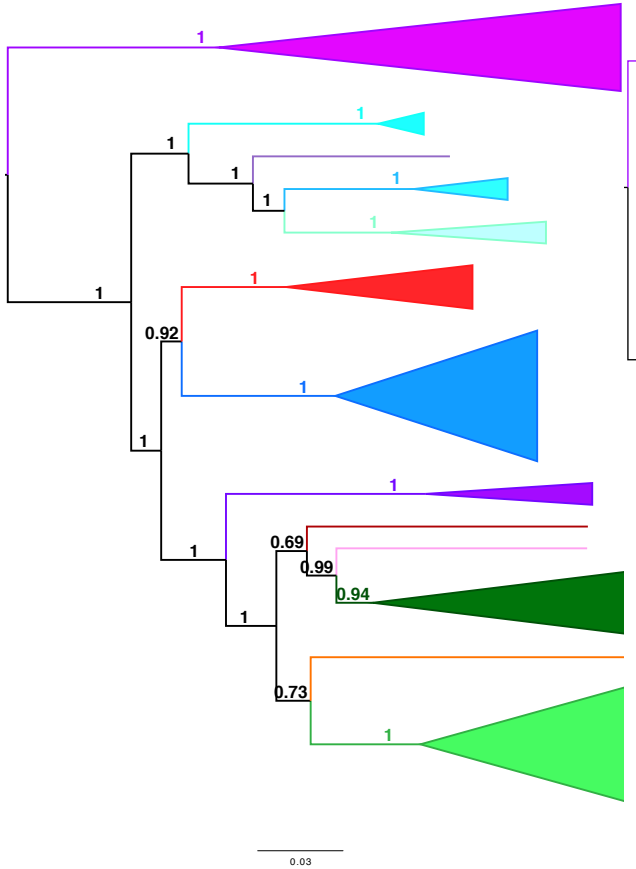
Number of taxa

- **What is good taxon sampling? – 10%, 20%, 50% of extant taxa?**
  - **Again, it depends on your question**
- **Taxon sampling – different across different groups**
  - **Dense taxon sampling in well known groups – vertebrates, plants, some insect groups (e.g. butterflies)**
  - **Relatively low taxon sampling – many invertebrate groups**

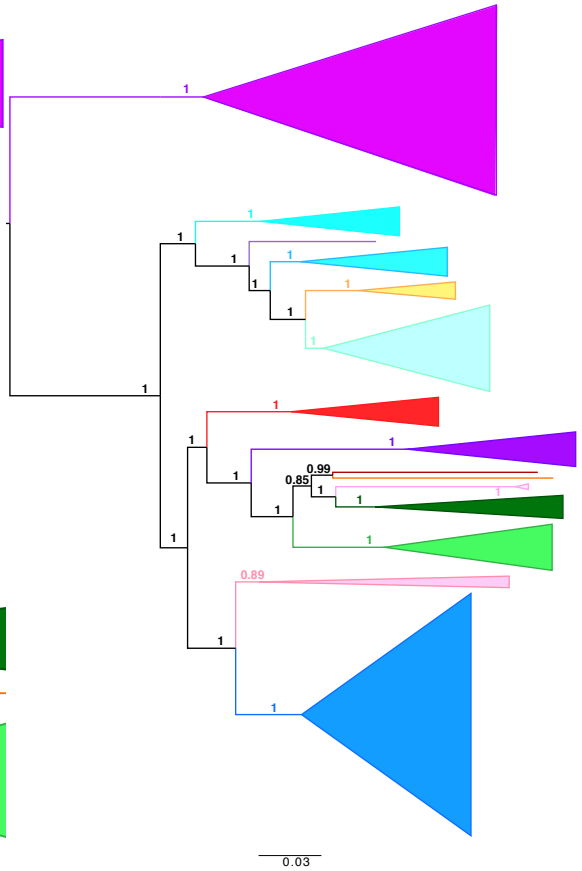
Adding genes and adding taxa helps!



3 genes,
42 taxa (Rota 2011)



8 genes,
38 taxa (Rota & Wahlberg 2012)



11 genes,
146 taxa (Rota unpubl.)

An empirical example: metalmark moths (Lepidoptera, Choreutidae), ca. 600 known species

Analysis: MrBayes; Branch support: Bayesian posterior probability

Missing data?

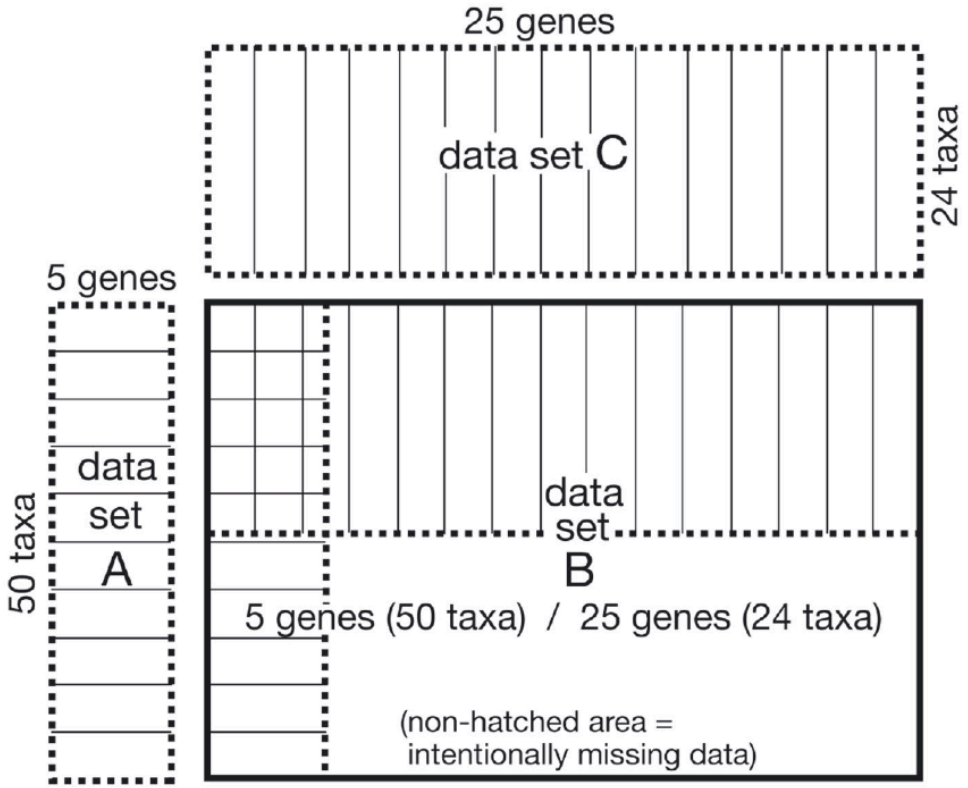
- **Sometimes not all genes amplify/are found from all samples**
  - **Should these samples be discarded?**
- **No - increased taxon sampling, despite missing data, *usually* increases resolution**
  - **As long as missing data are spread out across the phylogeny**
- **Start with using all available data in your data exploration**
  - **And perhaps drop some taxa if they are behaving as 'rogues'**

Systematic Entomology

Systematic Entomology (2010), DOI: 10.1111/j.1365-3113.2010.00543.x

Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera)

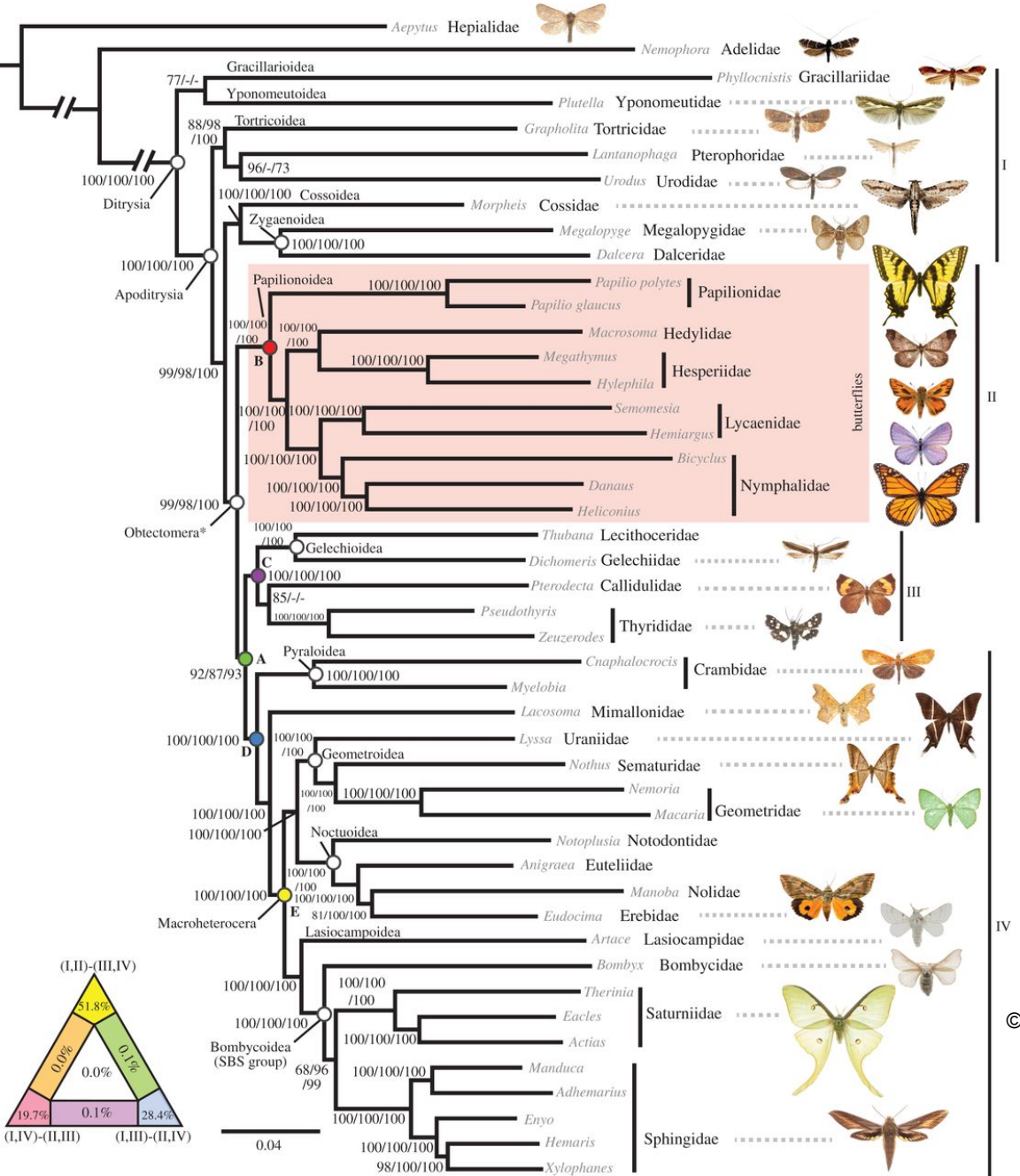
ANDREAS ZWICK[1,2], JEROME C. REGIER[1,3,4], CHARLES MITTER[3] and MICHAEL P. CUMMINGS[5]



Phylogenomics

- **Number of genes sequenced is going up by orders of magnitude**
- **Whole genome analyses allow us to understand:**
  - **Intron-exon boundary dynamics**
  - **Gene duplication-deletion dynamics**
  - **Gene transfer dynamics**
  - **We are getting a good understanding of the regions of the genome that are most suitable for systematics**
  - **Single copy, protein-coding nuclear genes seem to be the best**

Maximum-likelihood tree estimated in RAxML from the 2696-gene nucleotide dataset.

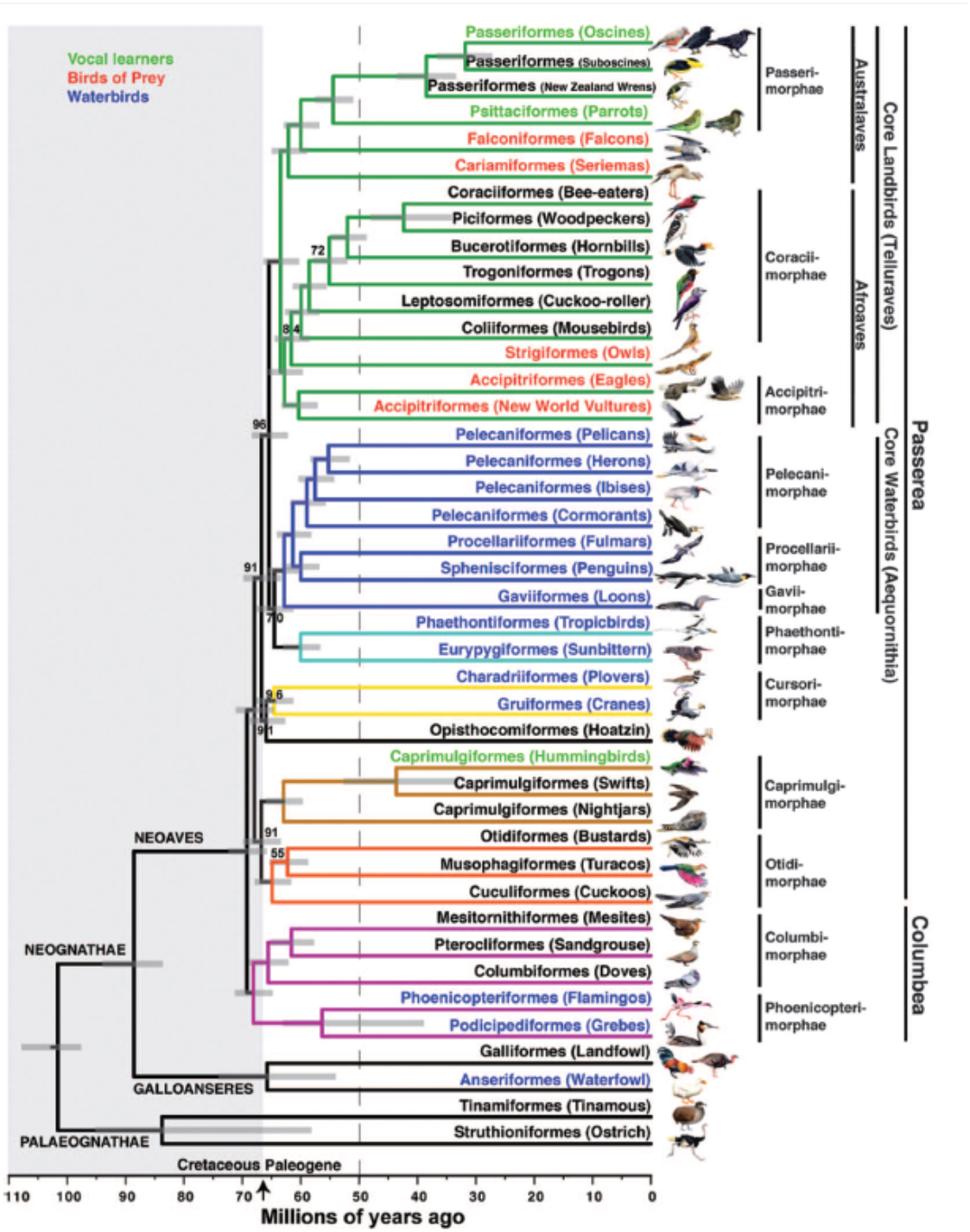


Akito Y. Kawahara, and Jesse W. Breinholt Proc. R. Soc. B 2014;281:20140970

PROCEEDINGS OF THE ROYAL SOCIETY B

8251 genes
2516 introns
3769 ultraconserved elements

41.8 million bp…



Jarvis et al. 2014: Science 346

Problems inherent to molecular data

What are the problems?

- **Saturation**
- **Bias in nucleotide composition**
- **Orthology *vs*. paralogy**
- **Lineage sorting**
- **Lateral gene transfer**
- **Mito-nuclear discordance**

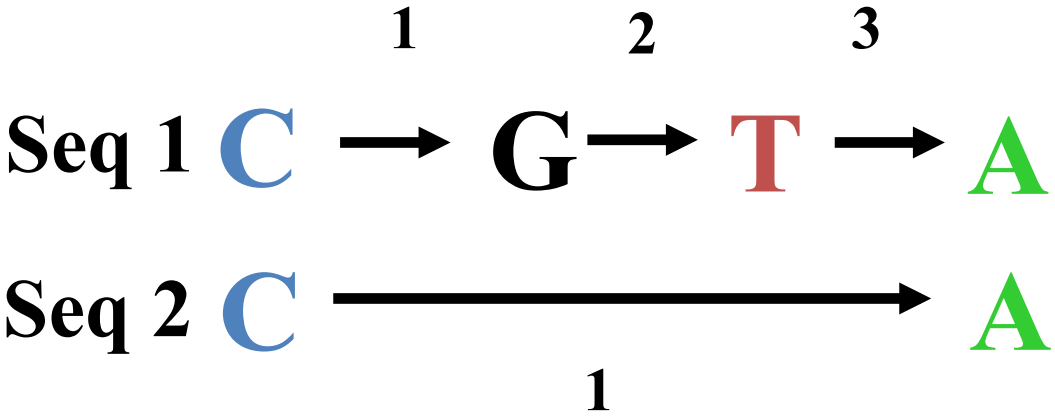
Saturation in sequence data

- **Saturation is due to multiple changes at the same site subsequent to lineage splitting**
- **Models of evolution attempt to infer the missing information through correcting for "multiple hits"**
- **Most data will contain some fast evolving sites which are potentially saturated (e.g. in proteins often codon position 3)**
- **In severe cases the data become essentially random and all information about relationships can be lost**

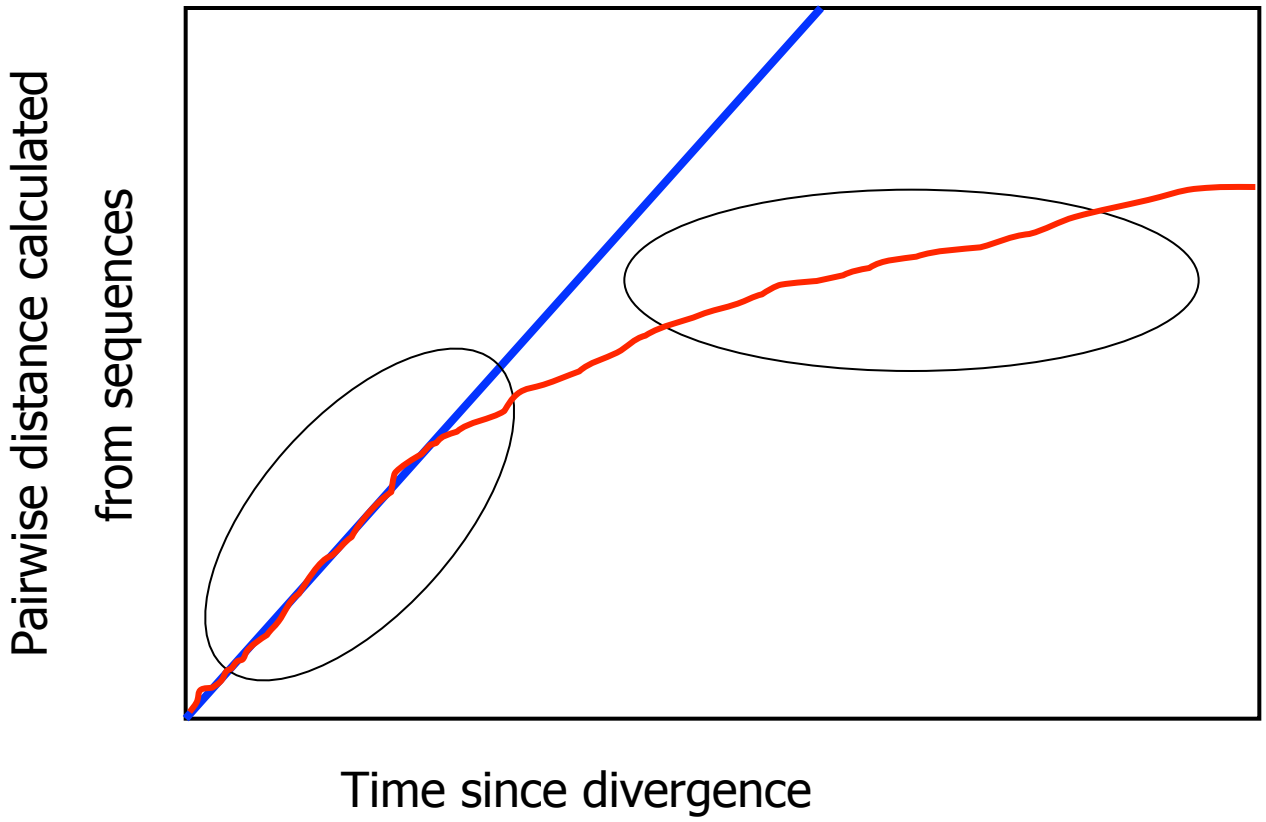
Multiple changes at a single site - hidden changes

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

Number of changes



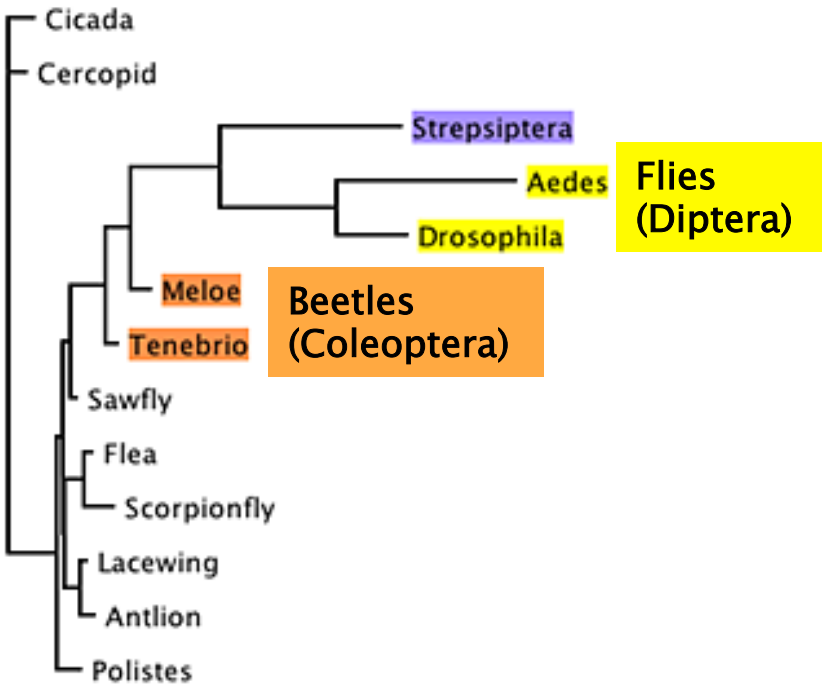
Saturation


Pairwise distance calculated
from sequences
Time since divergence

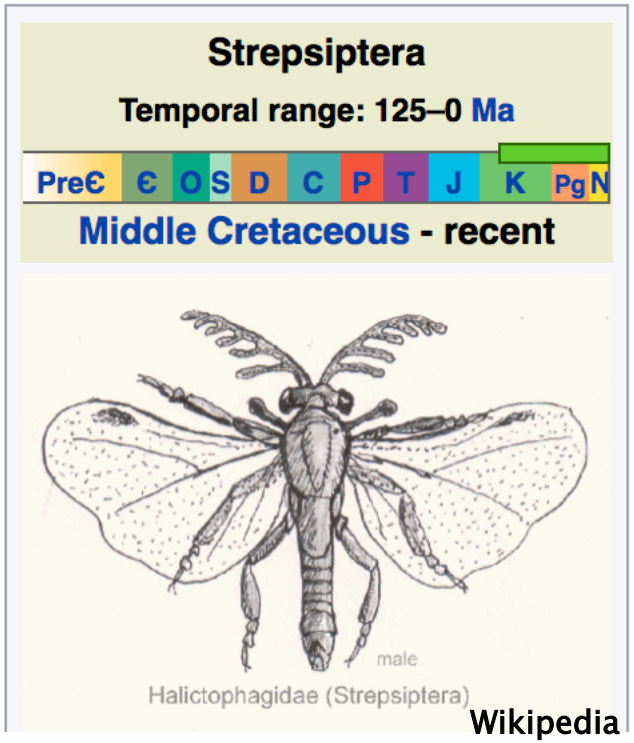
Saturation and long-branch attraction

- **Homoplasy is a 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

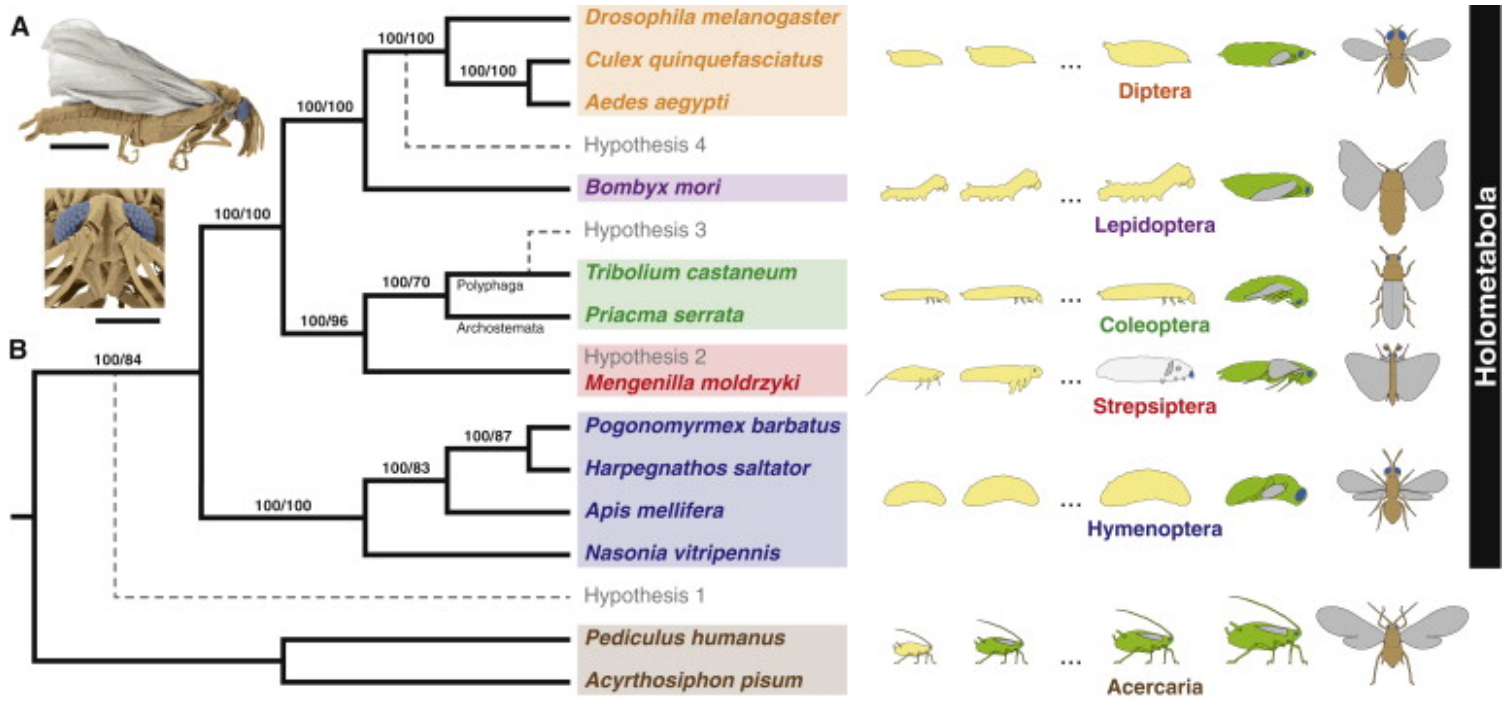


Figure 1. Evolutionary Origin of Twisted-Wing Parasites Inferred from Genomic Evidence(A) Mengenilla moldrzyki male in lateral (top; scale bar represents 1 mm) and frontal (bottom; scale bar represents 500 μm) view (colored SEM micrographs; wings in gray, comp...

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?

- **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 not strong for long branches sticking together in model-based methods – so we should be able to recognize it!**
- **More data from different sources**
  - **Could be that molecular data are not able to resolve the position of some taxa**
  - **Morphological data!**

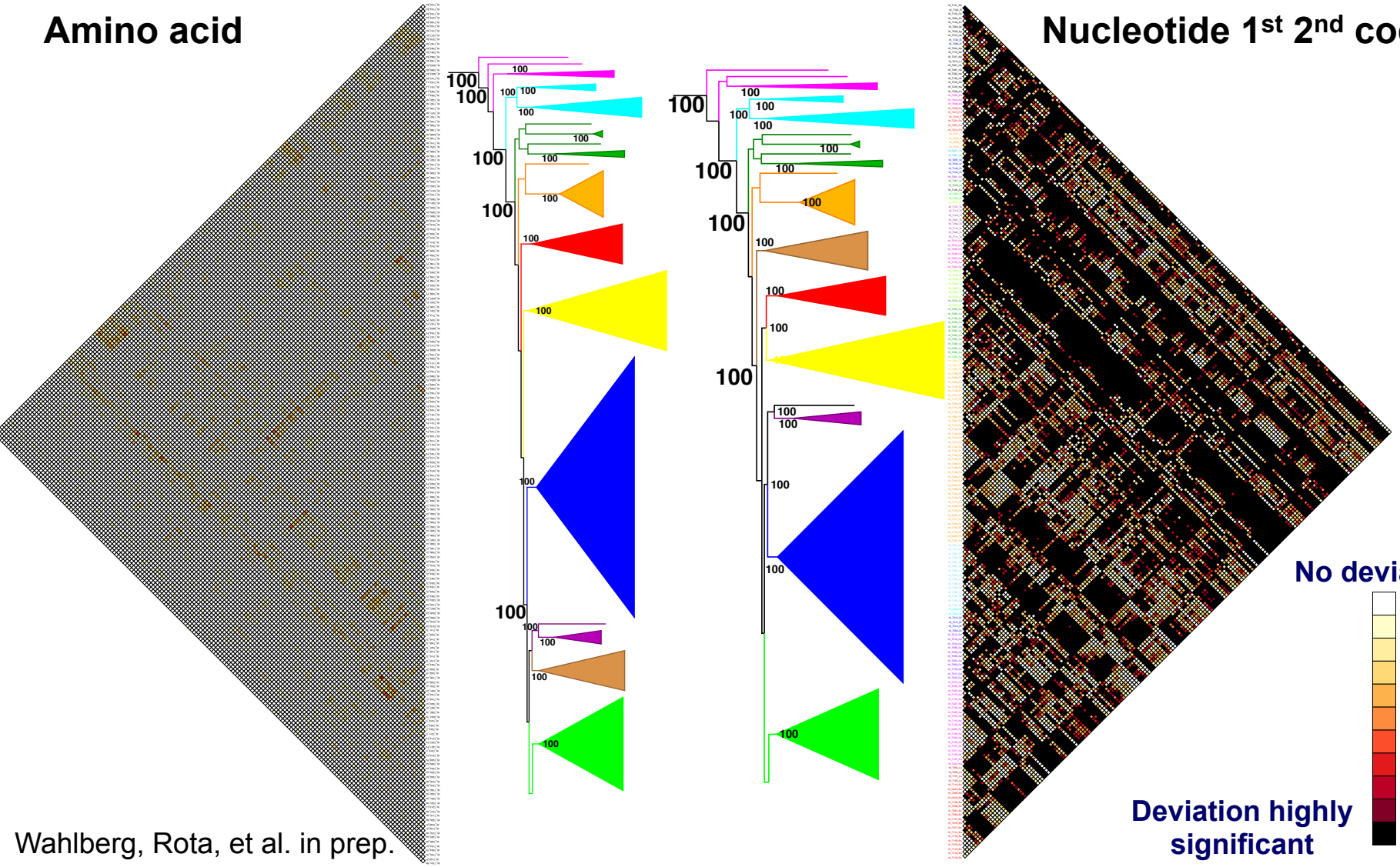
Biased base composition

Biased base compositions?

- **Do sequences manifest biased base compositions or biased codon usage patterns, which may obscure phylogenetic signal?**
  - **E.g. some taxa have a high GC content, if they are inferred to belong to the same clade – is this real or is it because of their base composition?**

Compositional hetereogenety

Lepidoptera, 200 taxa representing most superfamilies, 338 gene fragments



Wahlberg, Rota, et al. in prep.

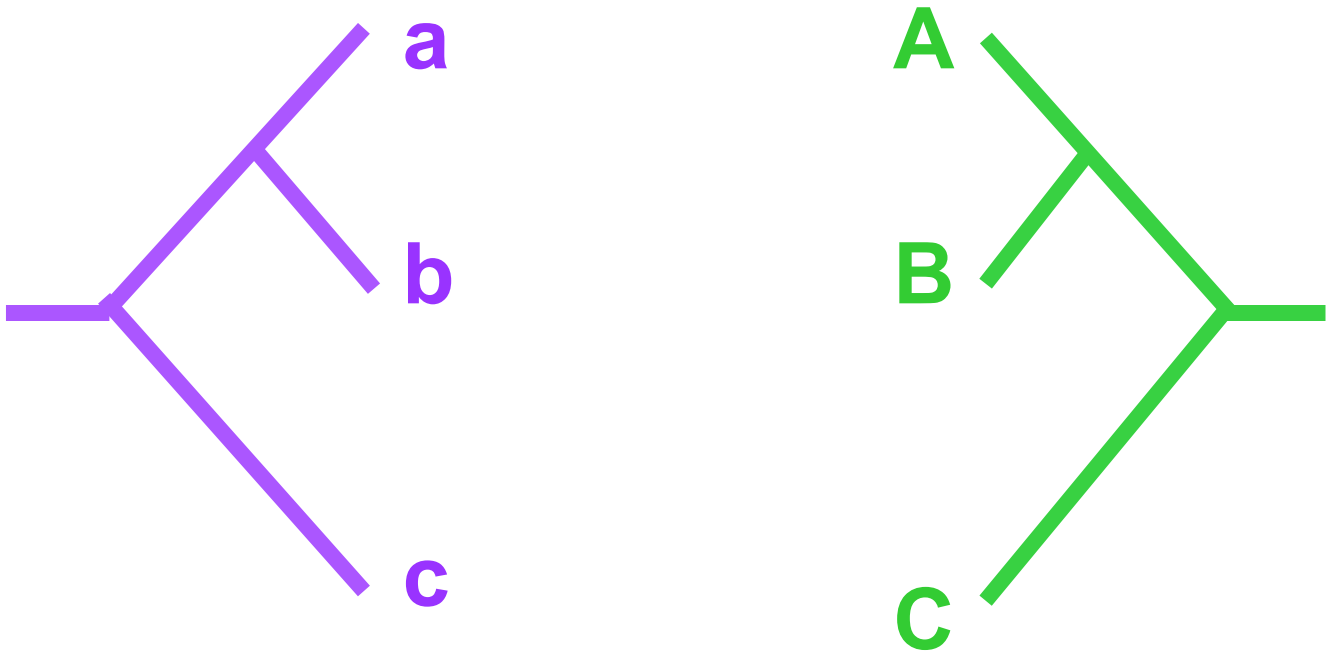
Orthology or paralogy?

- Are the data 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 and paralogy

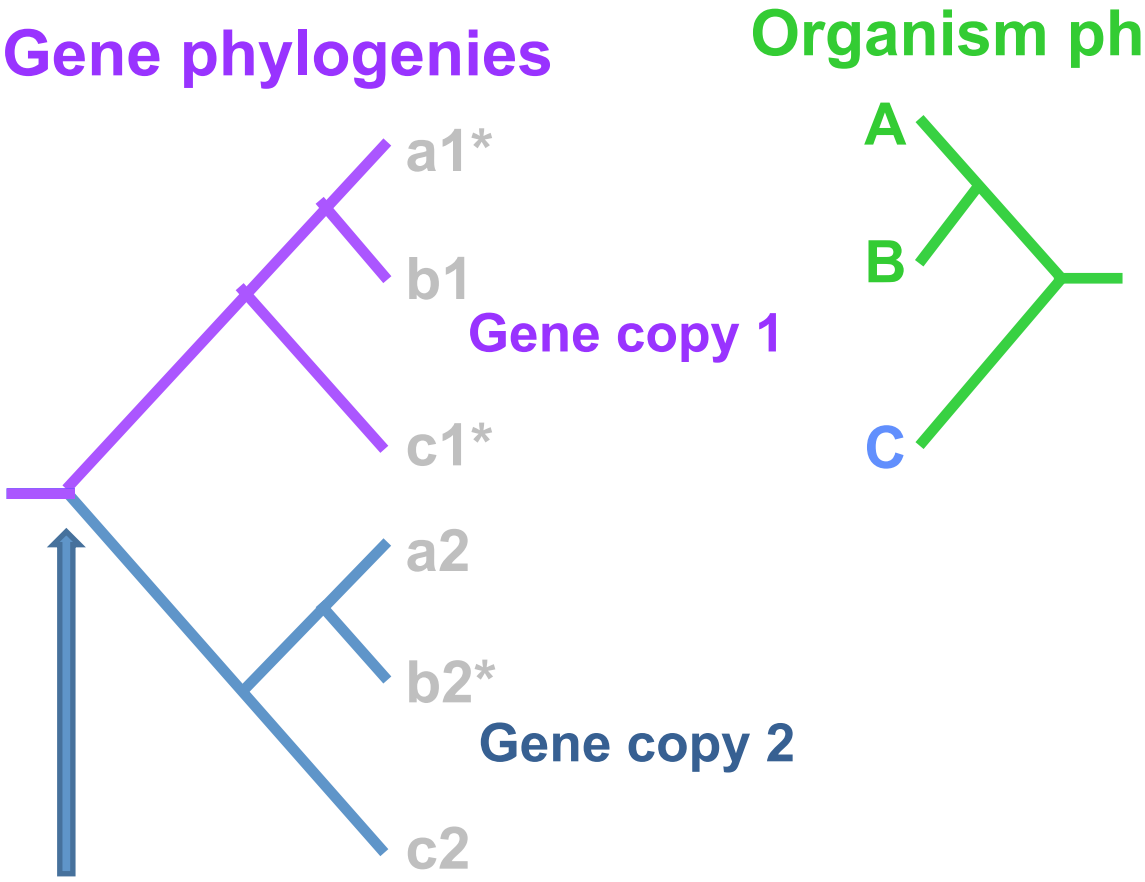
Orthology:
gene trees and species trees

Gene phylogeny Organism phylogeny



ORTHOLOGY

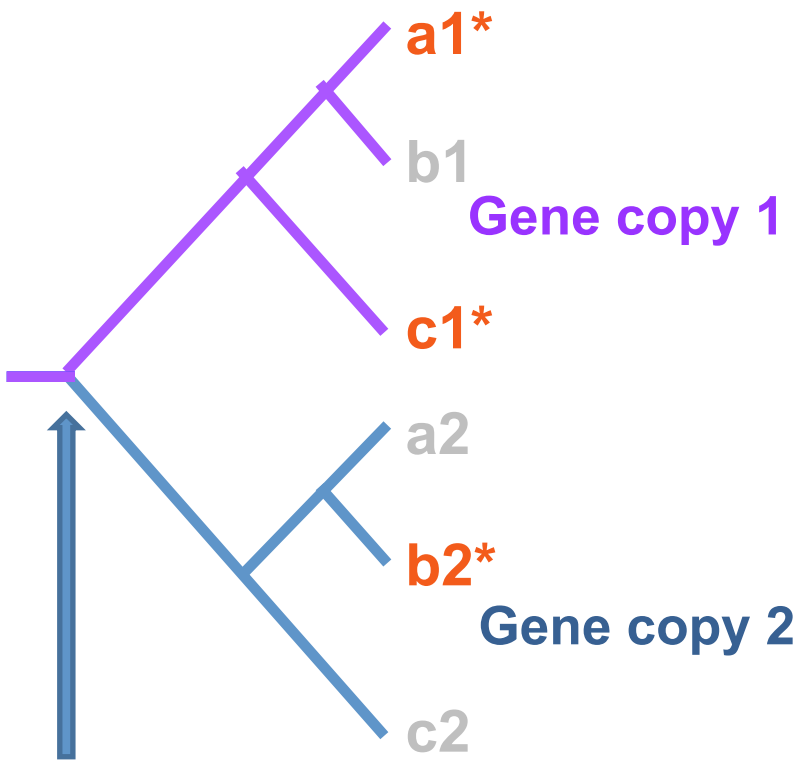
Paralogy: can produce misleading trees



PARALOGY

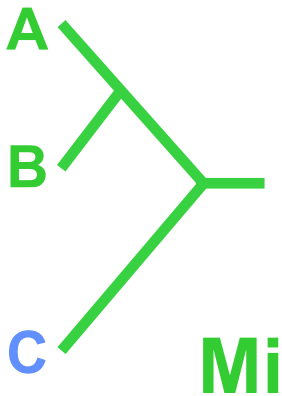
Paralogy: can produce misleading trees

Gene phylogenies

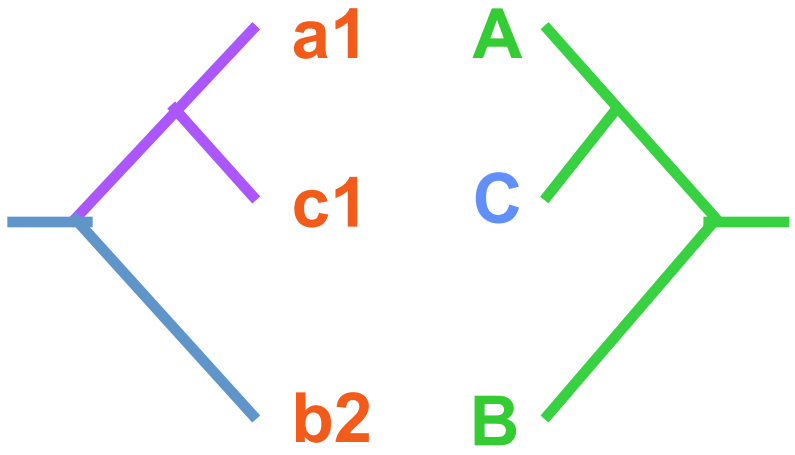


gene duplication

Organism phylogeny



Misleading tree



PARALOGY

Lineage sorting

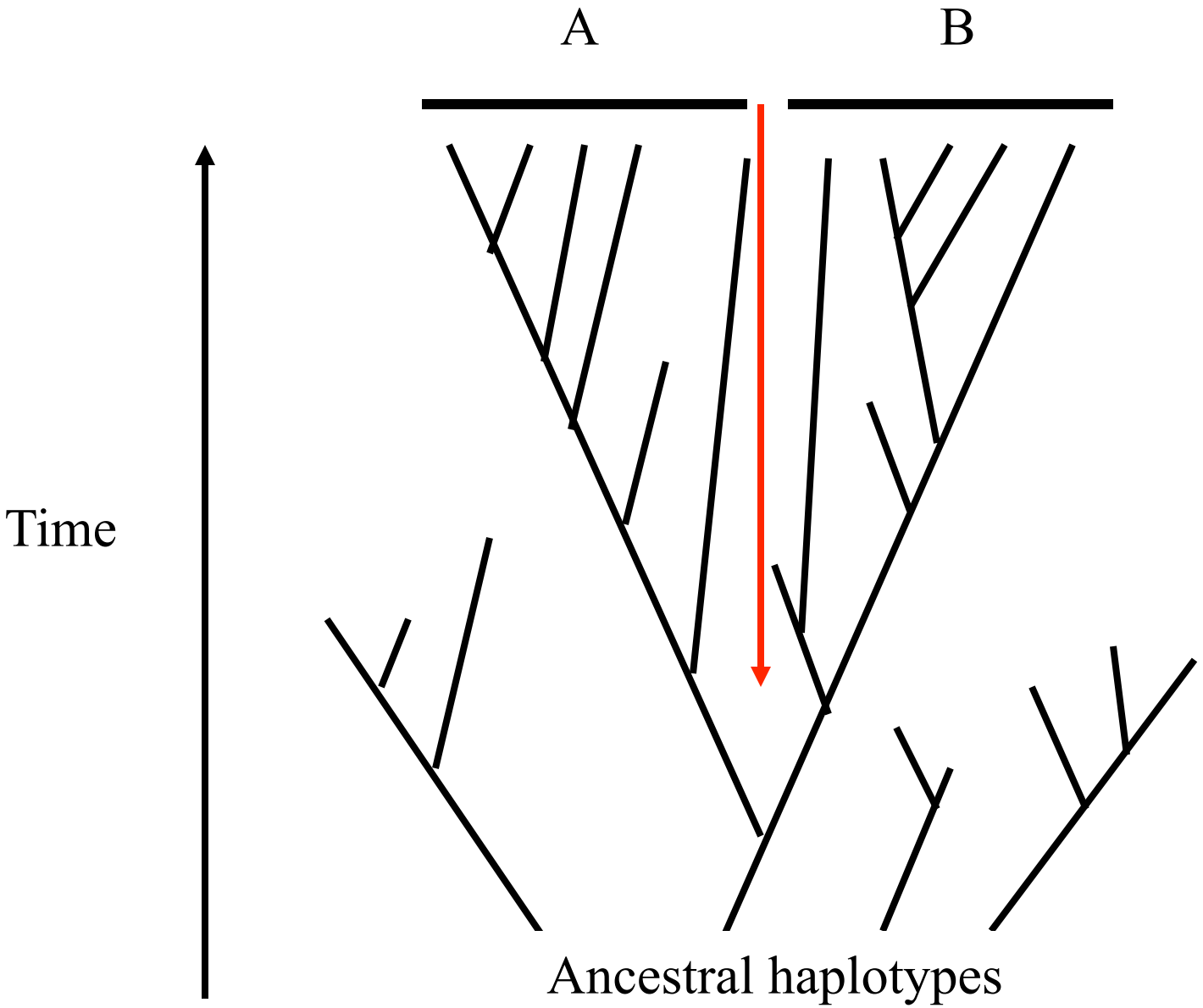
Lineage sorting

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

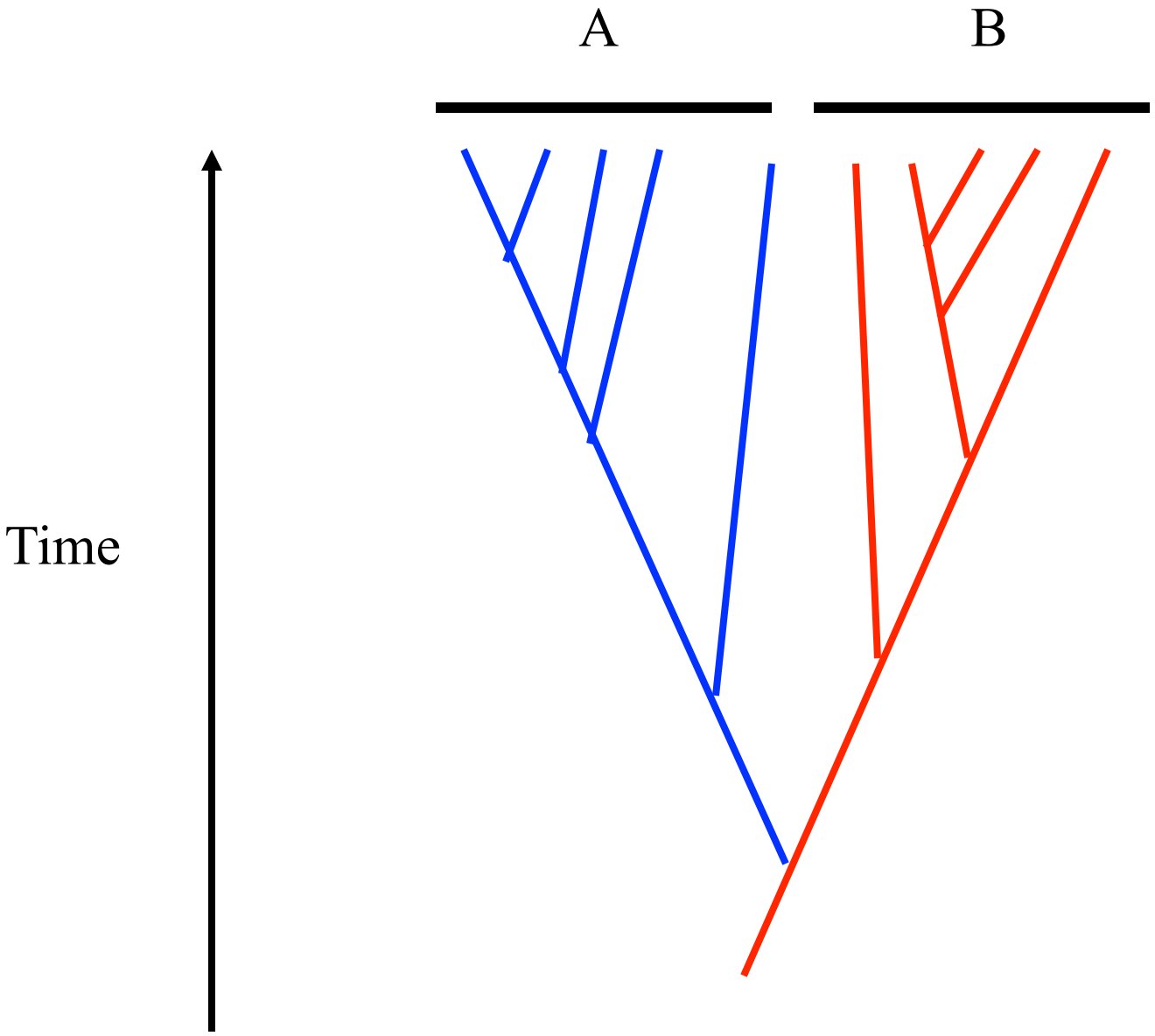
Are species monophyletic?

- **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



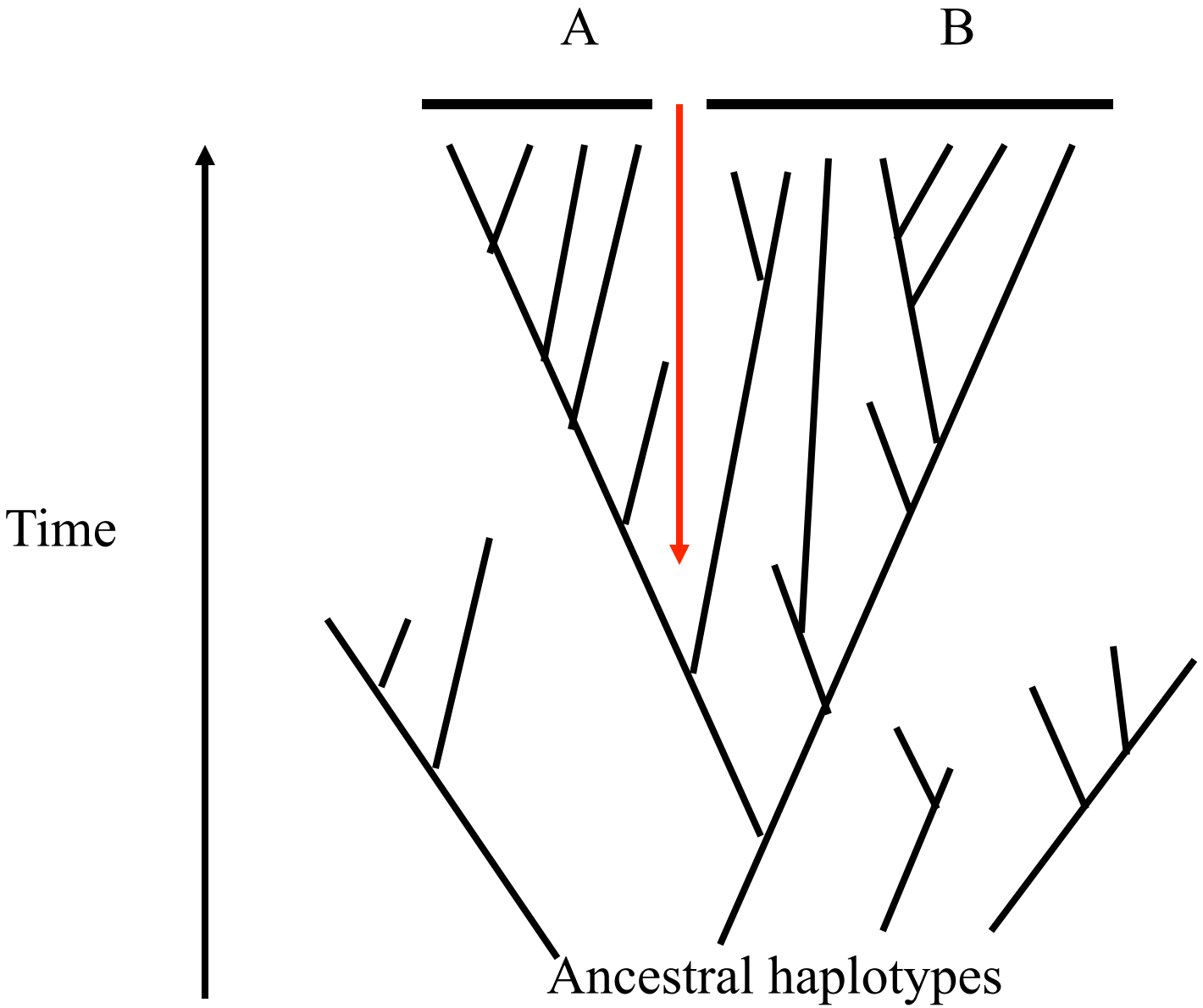
The assumption: monophyly



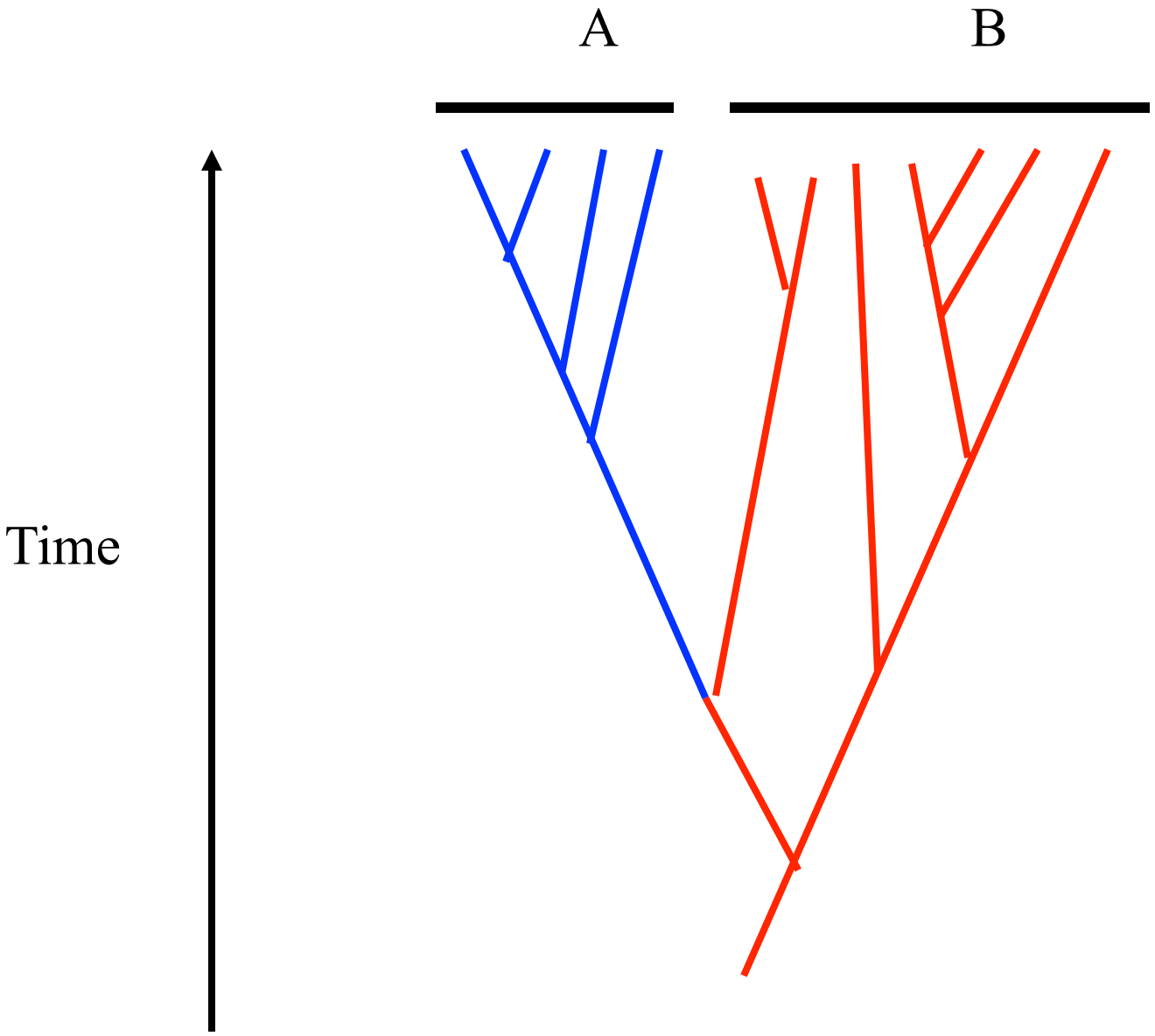
The presence of poly- and paraphyletic lineages

- **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**

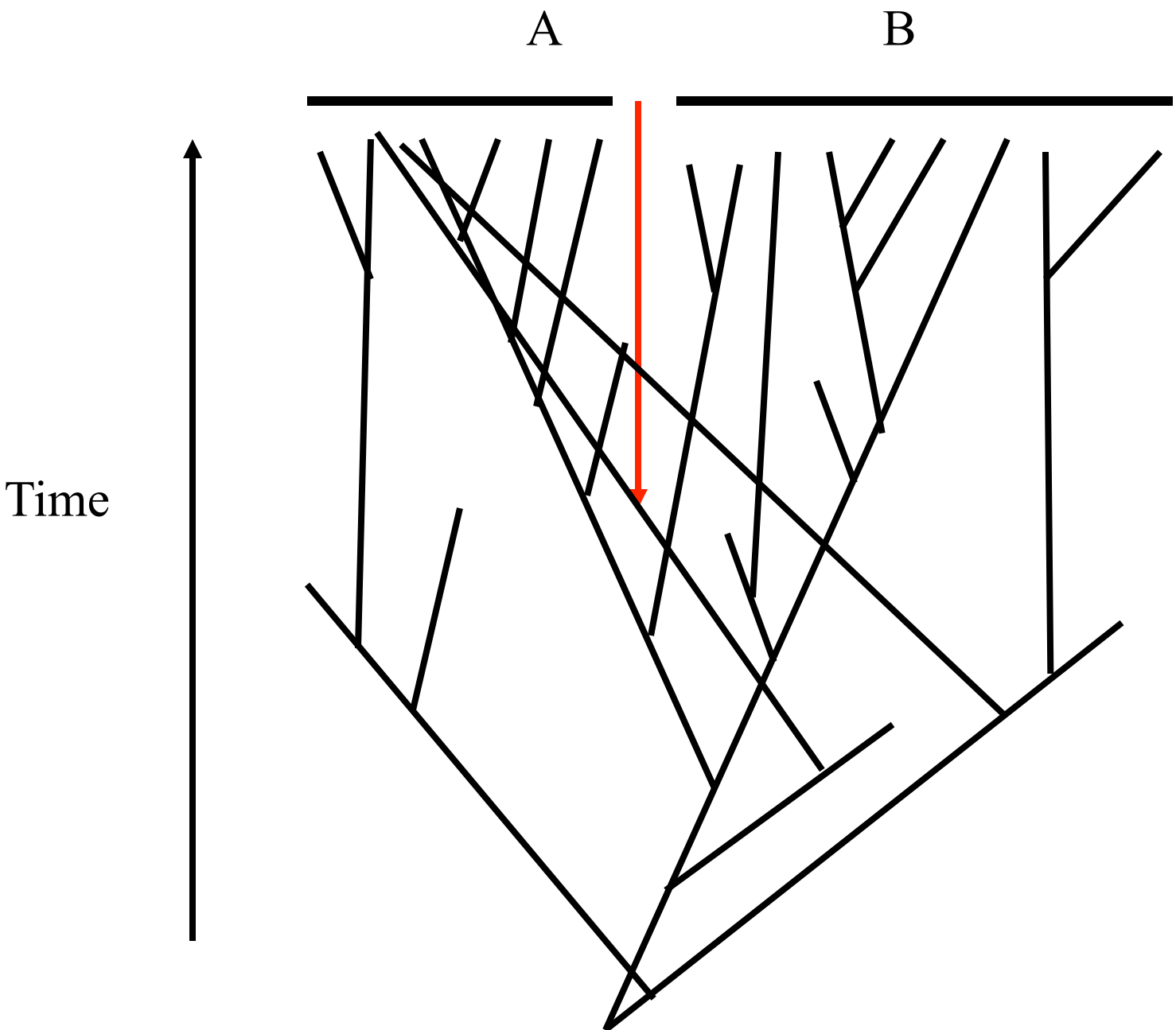
Paraphyly



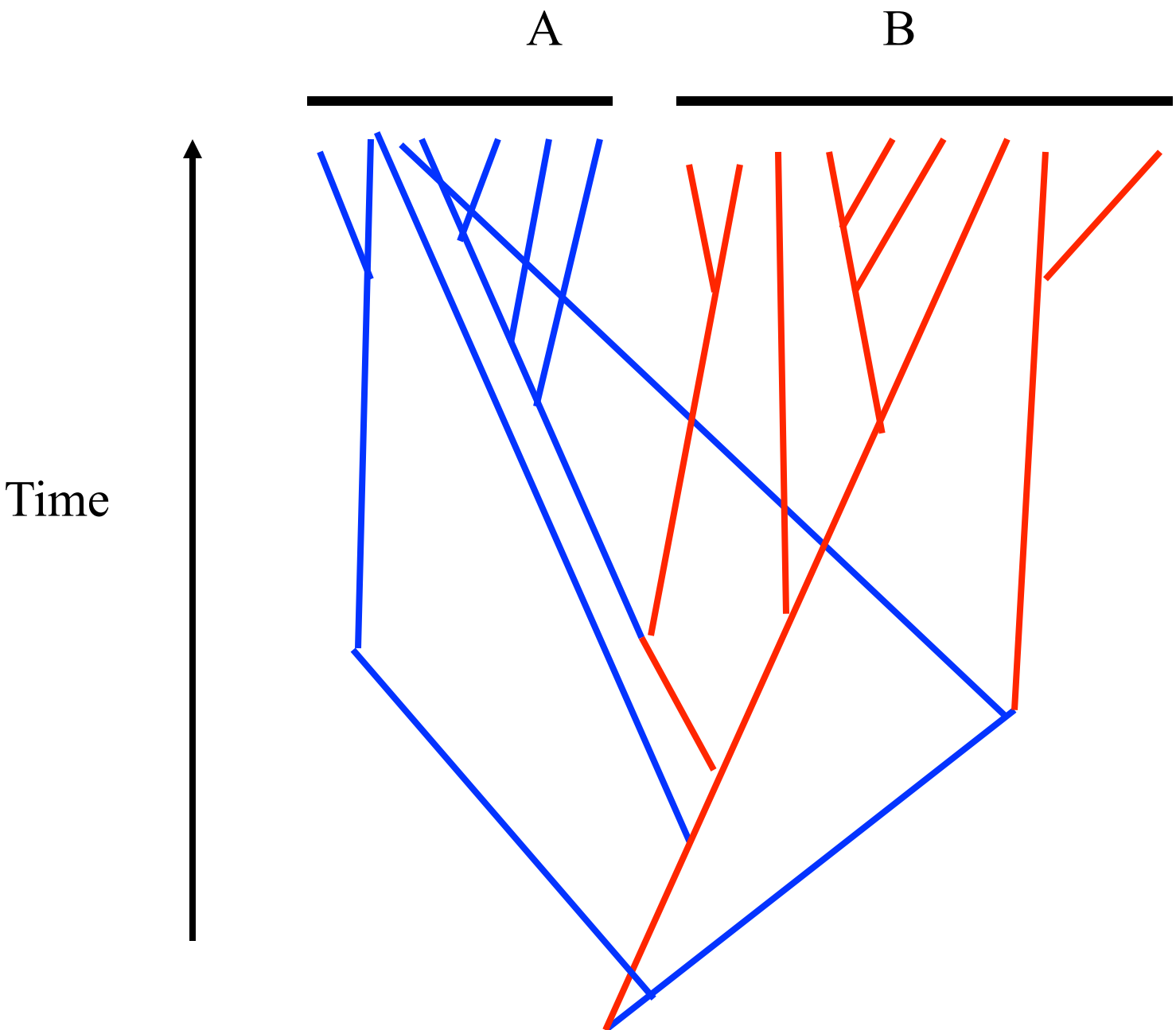
Paraphyly



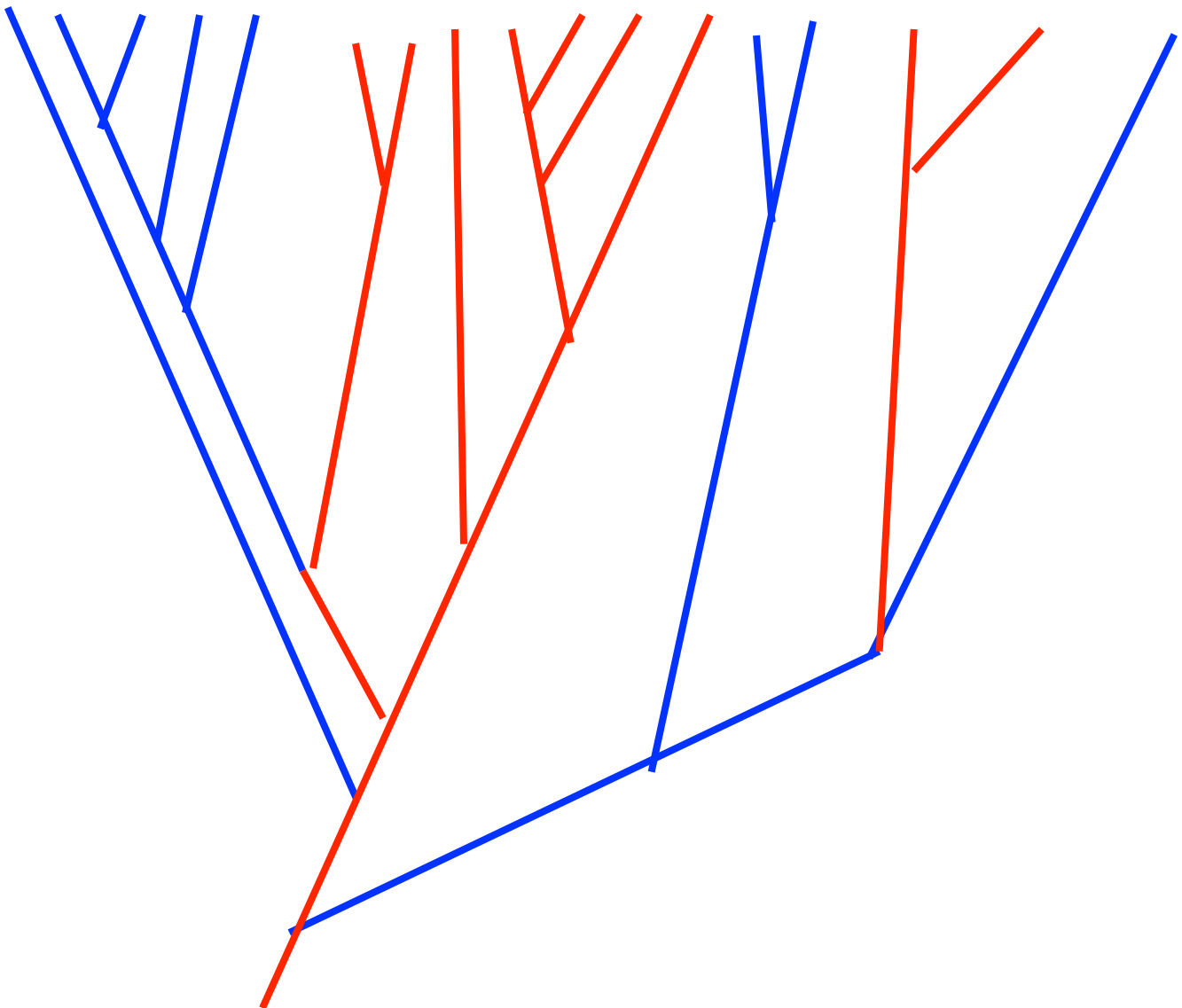
Polyphyly



Polyphyly



Polyphyly



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

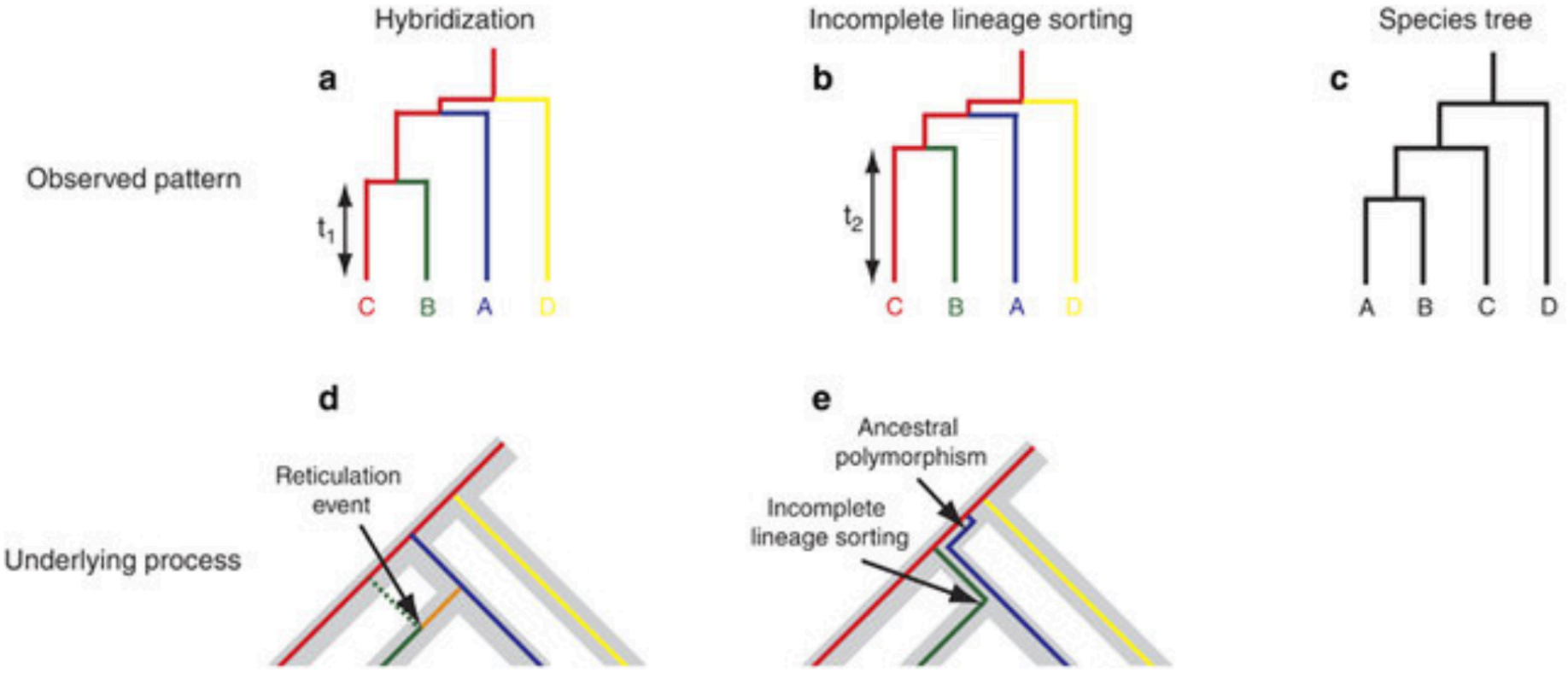
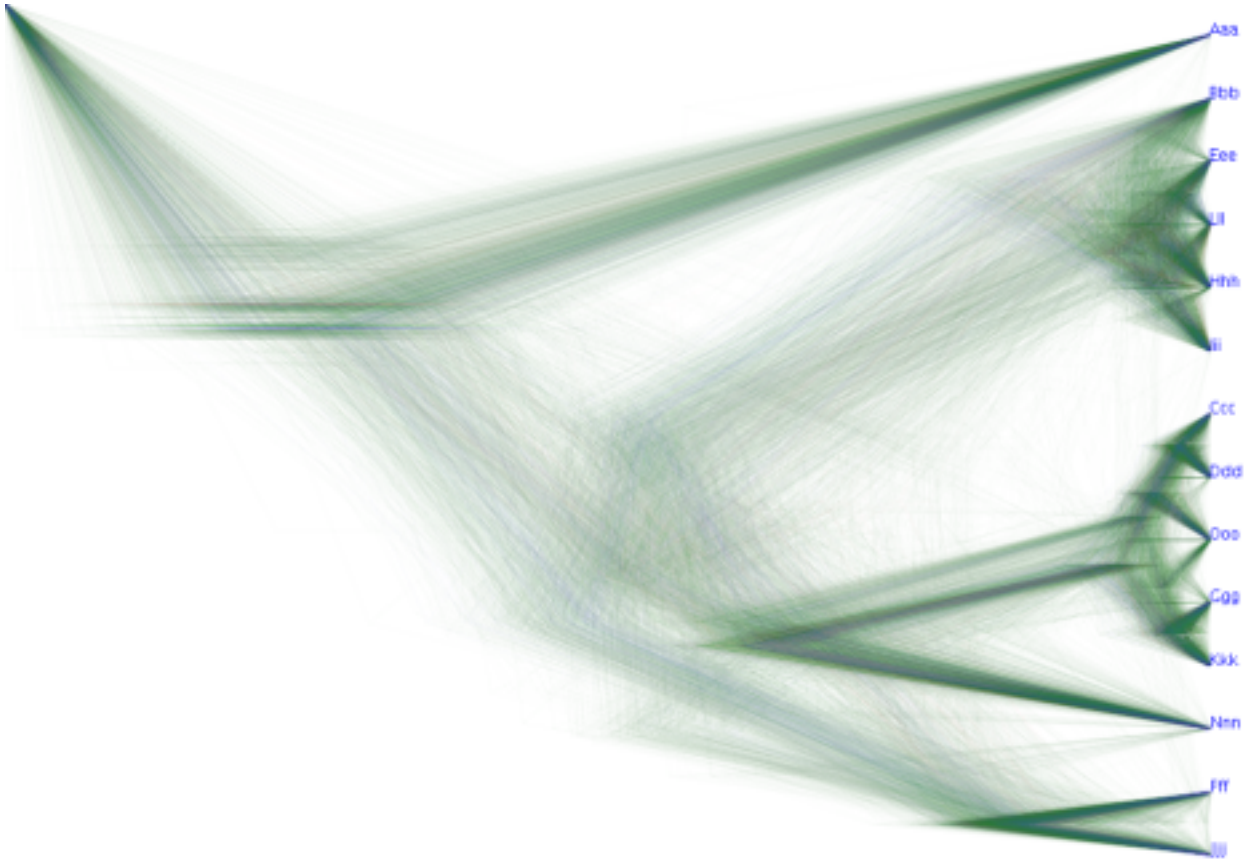


Figure from ResearchGate, uploaded by Richard A Ennos

Gene tree/species tree – Densitree

Many thousands of genes will give the general story



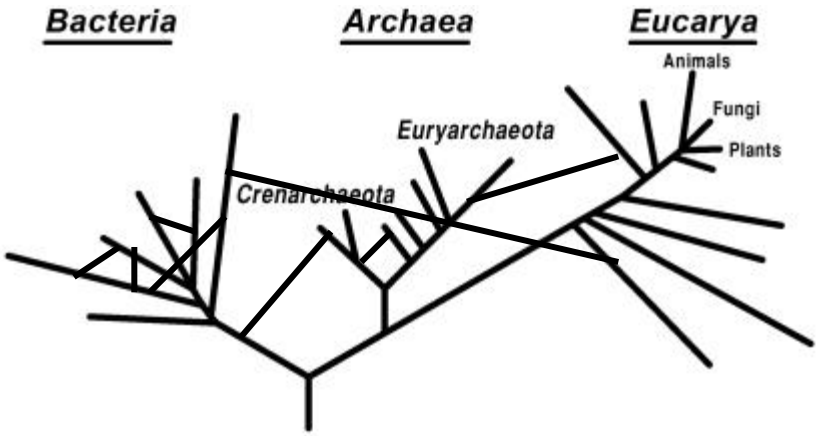
Lateral gene transfer

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**

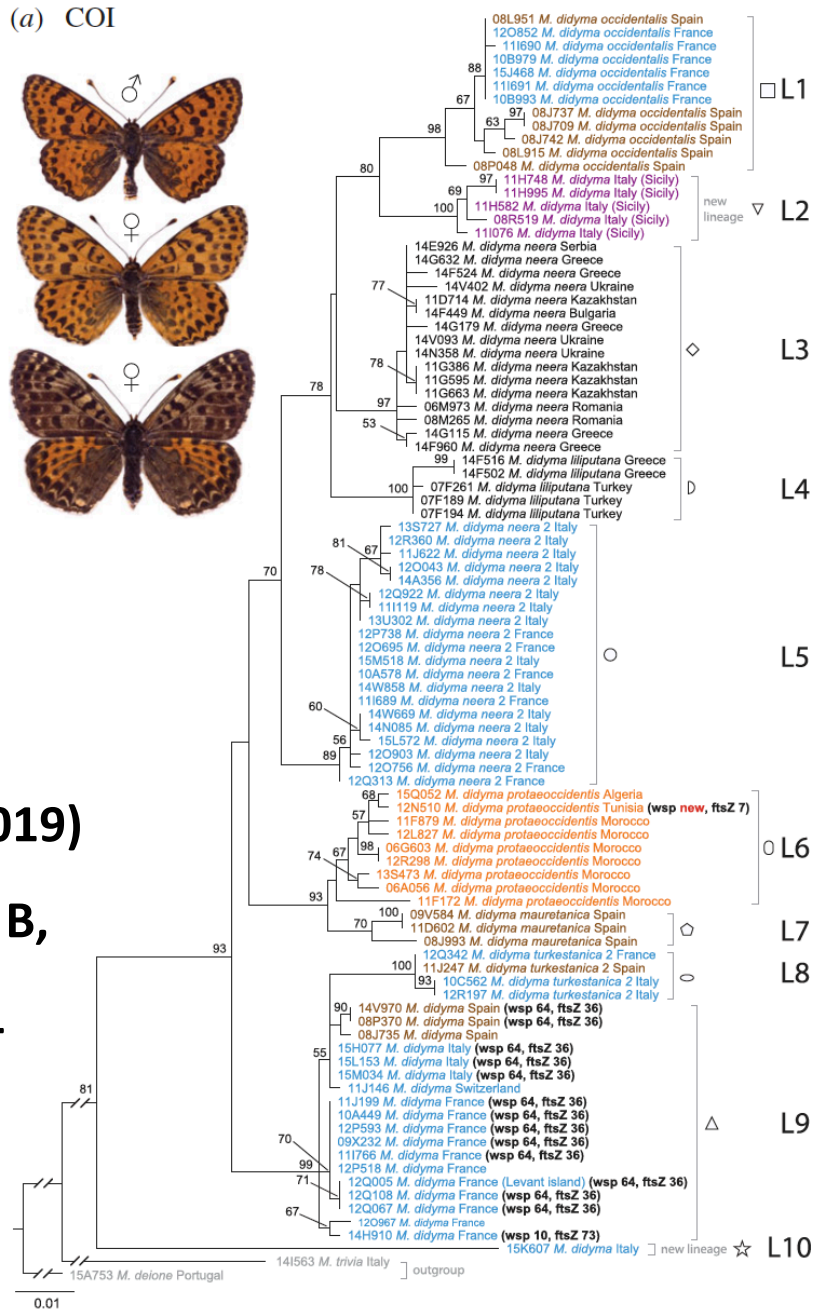
Lateral Gene Transfer

- **Is the Tree of Life really a Web of Life?**



Mito-nuclear discordance

An empirical example: *Melitaea* butterflies

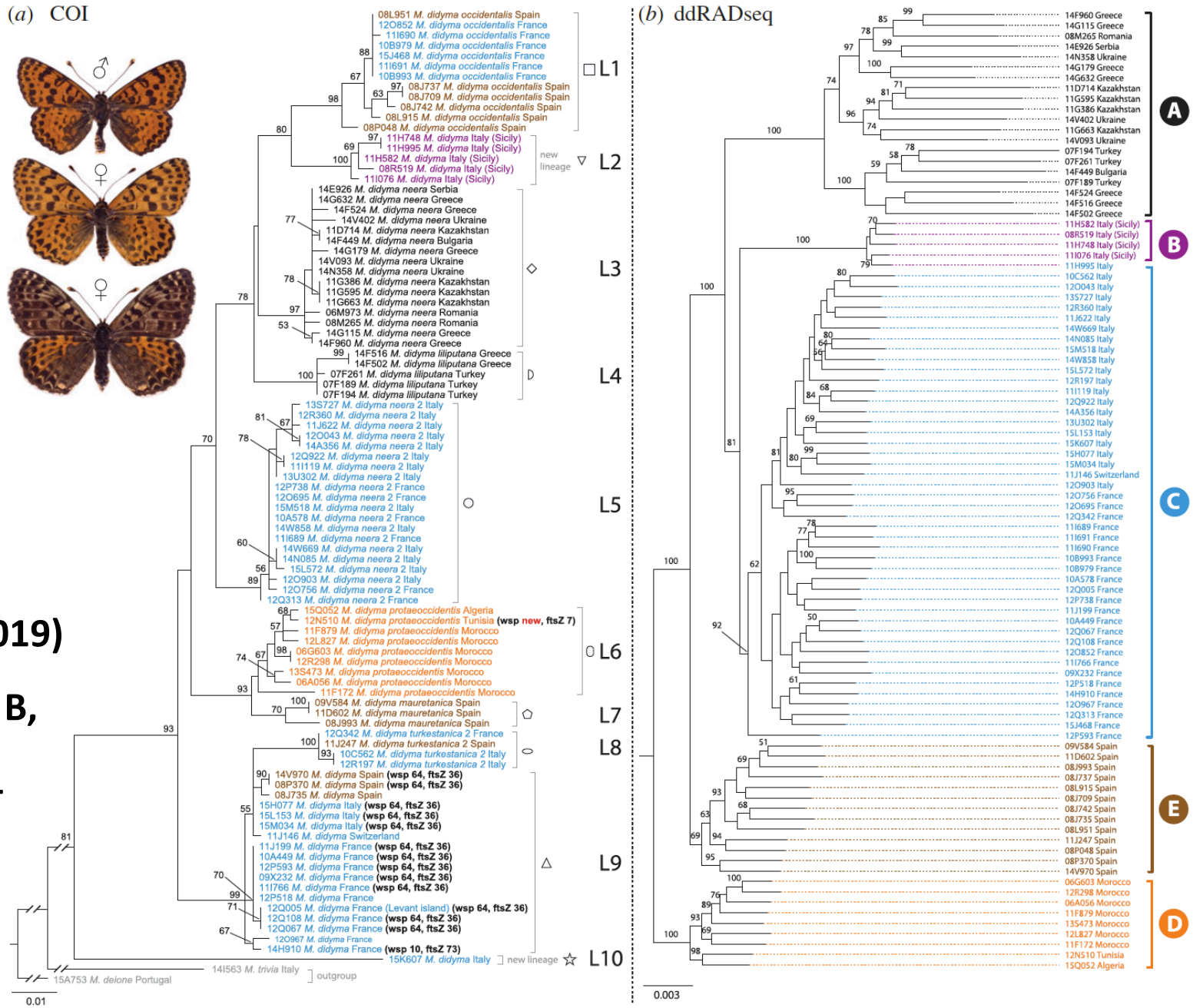


Dinca et al. (2019)

Proc. Roy. Soc B,

286: 20191311

An empirical example: *Melitaea* butterflies



Dinca et al. (2019)

Proc. Roy. Soc B,

286: 20191311

Problems inherent to molecular data?

- These ”problems” 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

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**

Bootstrapping phylogenies

- Characters are resampled with replacement to create many bootstrap replicate data sets
  - Often 1000 replicates done
- Each bootstrap replicate 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

Bootstrapping

Original data matrix

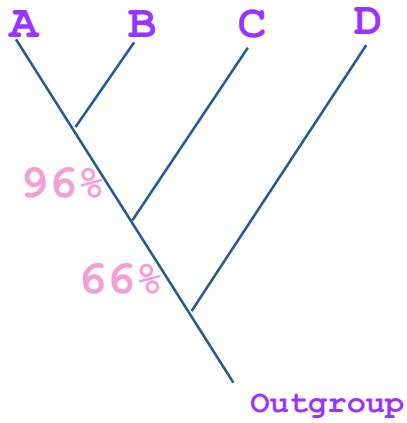
	Characters							
Taxa	1	2	3	4	5	6	7	8
A	R	R	Y	Y	Y	Y	Y	Y
B	R	R	Y	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	R	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R

Resampled data matrix

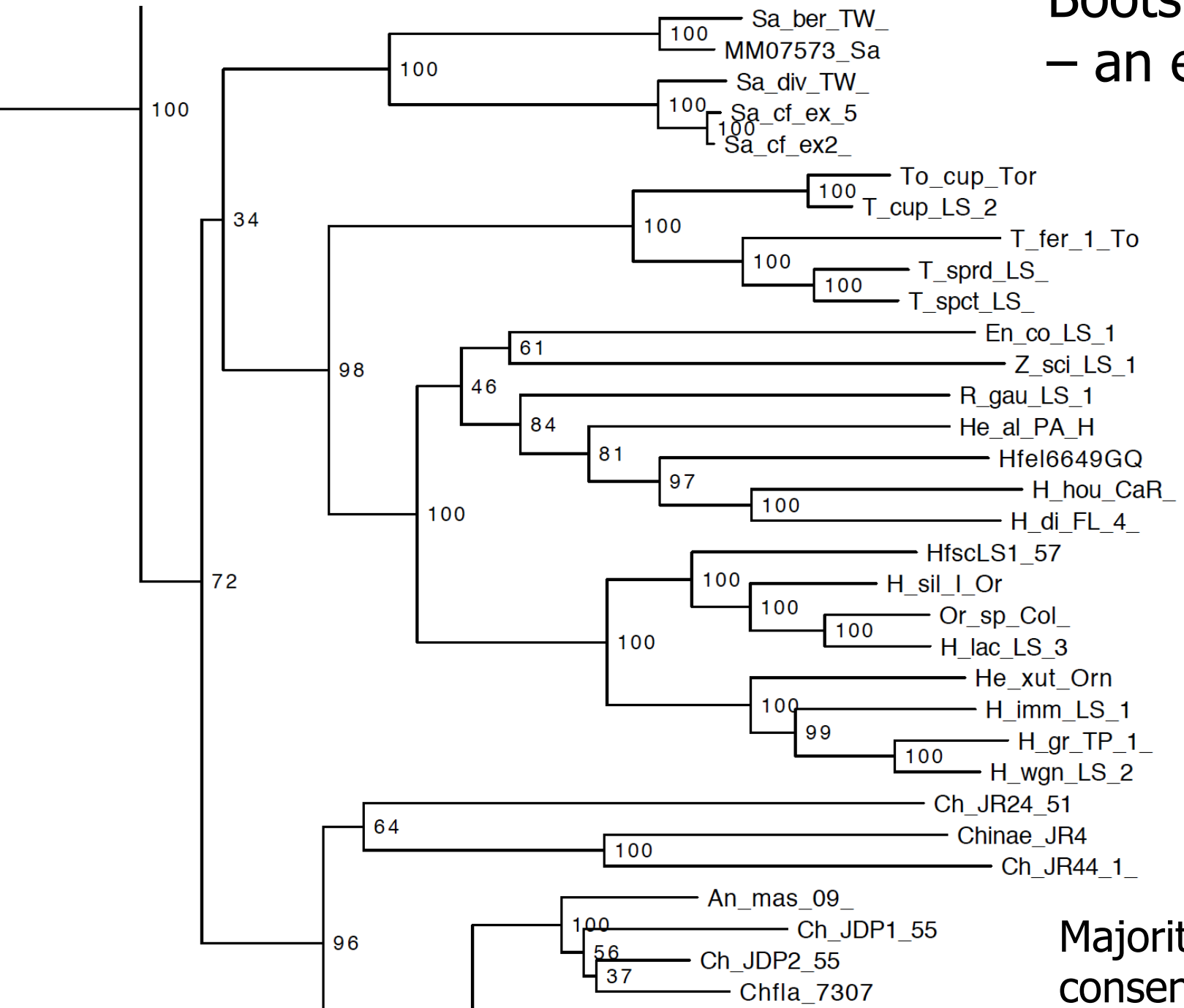
	Characters							
Taxa	1	2	2	5	5	6	6	8
A	R	R	R	Y	Y	Y	Y	Y
B	R	R	R	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	Y	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R

Randomly resample with replacement characters from the original data to build many replicate data sets of the same size as the original. Analyse each replicate data set.

Summarise the results of multiple analyses with a majority-rule consensus tree Bootstrap proportions (BPs) are the frequencies with which groups are encountered in analyses of replicate data sets



Bootstrapping – an example



Majority-rule consensus tree

Bootstrap - interpretation

- **Felsenstein 1985:**
  - **“A measure of repeatability” or “the probability that a specific internal branch would be found in an analysis of a new, independent sample of characters.”**
- **Felsenstein & Kishino 1993:**
  - **“A measure of accuracy” or “the probability that a specified branch is contained in the true tree” (assuming consistency)**
- **Swofford et al. 1996:**
  - **“The frequency with which a group appears in replicate trees is better thought of as a measure of support rather than a statistical statement. Statistical validity would require that the node of interest was specified in advance.”**

Bootstrap - interpretation

- **It gets complicated…**
  - **Some systematists reject the assumptions of bootstraping, which complicates the interpretation…**
    - **Sanderson 1995. Objections to bootstrapping phylogenies: a critique. Syst. Biol. 44:299-320**

Bootstrap - interpretation

- Hillis & Bull 1993
  - Examined interpretation of BP using simulated data & known phylogenies
  - Conclusions:
    - Low BPs overestimate accuracy
    - High BPs underestimate accuracy - BP = 70% was statistically significant support (only applies to their simulated data)

Bootstrap - interpretation

- Bootstrapping is a very valuable and widely used technique (it is demanded by some journals), but requires a pragmatic interpretation
- BPs thus provide a good index of the relative support for groups provided by a set of data
- BPs widely used for 20+ years in phylogenetics
- Somewhat unclear how useful BPs are in phylogenomics
  - At congresses one hears ”weakly supported with only 98% BP”

Jackknifing

- Jackknifing is very similar to bootstrapping and differs only in the character resampling strategy, tends to produce similar results to bootstrapping
- Resampling without replacement
- Some proportion of characters (e.g. 50%) are randomly selected and deleted
- Replicate data sets are analysed and the results summarised with a majority-rule consensus tree
- Jackknifing is not as widely used as bootstrapping

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-like approximate likelihood ratio test**
  - **>=80% is significant**
  - **Robust to various model assumption violations**
- **aBayes − a Bayesian like transformation of aLRT**

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 posteriror probabilities**
- **Ronquist & Deans 2010. Ann. Rev. Ent.**

Bayesian posterior probabilities (PPs)

- **Branch support in Bayesian analysis**
  - **Frequency of each clade in the posterior distribution of the trees**
  - **Interpreation: an estimate (approximation) of the probability that a certain branch exists, given the data, the model, the priors**
- **PPs usually shown as 50% majority-rule consensus tree**
- **PPs more sensitive to violations of the model, especially under-parameterization**

Bayesian Inference - Err on side of complexity

- PP sensitive to violations of model
  - Buckley 2002, Erixon et al. 2003, Huselsenbeck & Rannala 2004
- Slight over-parameterization not a problem for PP (slightly increased variance)
  - Cunningham et al. 1998
- Under-parameterization can inflate PP
  - Erixon et al. 2003, Huelsenbeck & Rannala 2004, Lemmon & Moriarty 2004

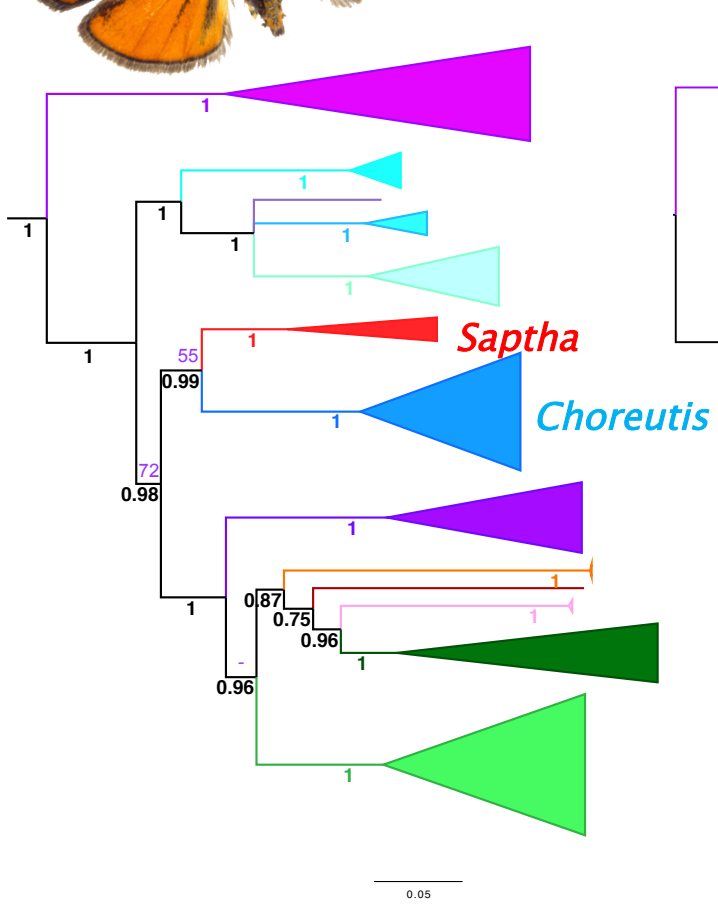
Branch support: summary

- **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 BPs and *vice versa* and such differences in branch support may indicate that data are misleading in some way**

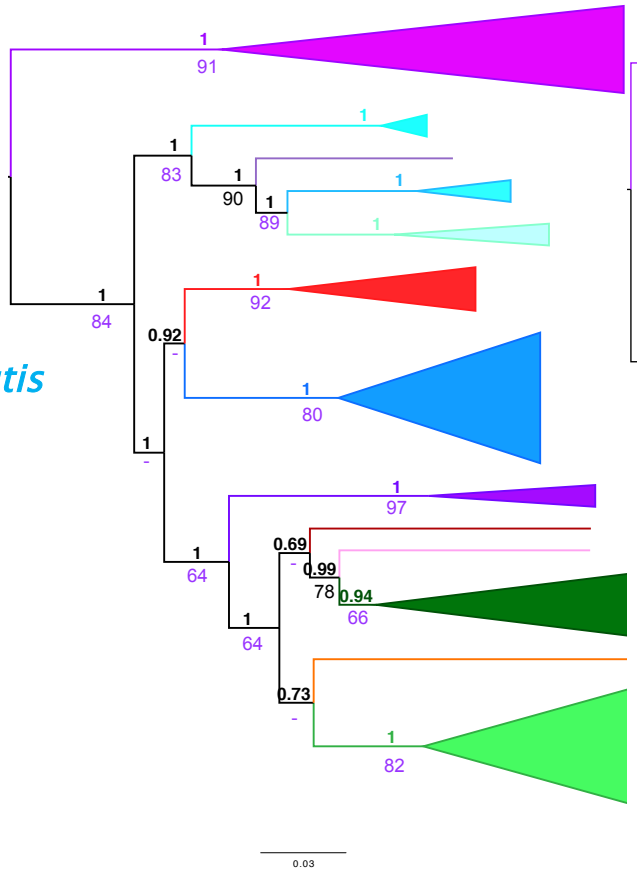
What is significant support?

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

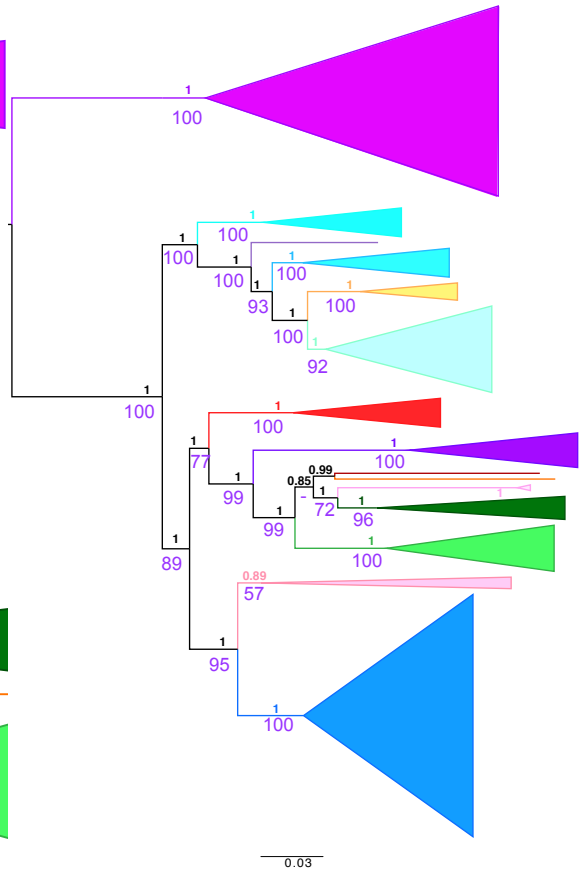
Comparison: BPs and PPs



3 genes,
42 taxa (Rota 2011)



8 genes,
38 taxa (Rota & Wahlberg 2012)

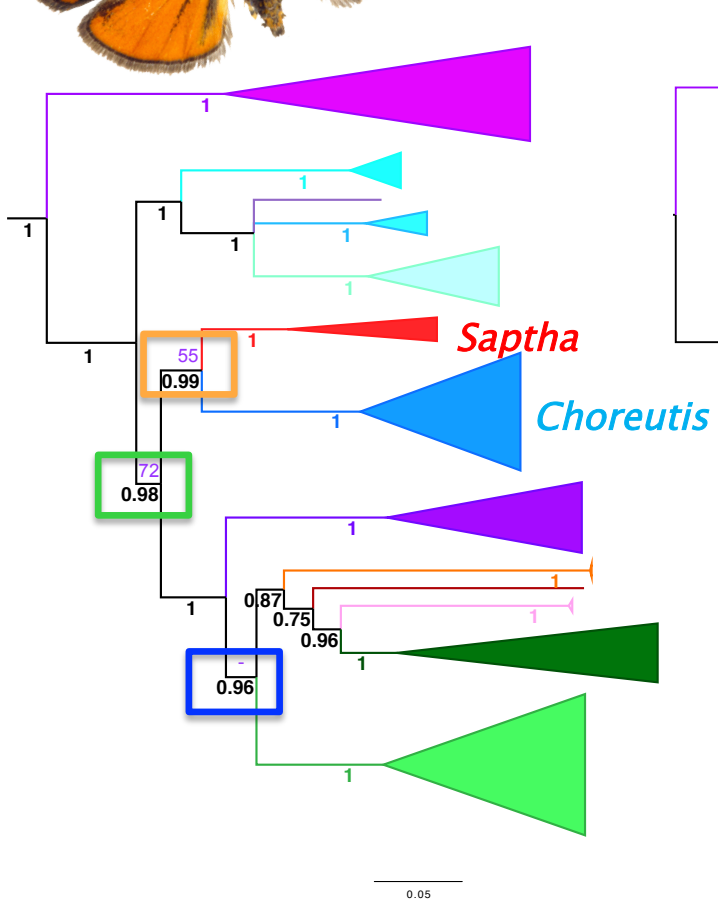


11 genes,
146 taxa (Rota unpubl.)

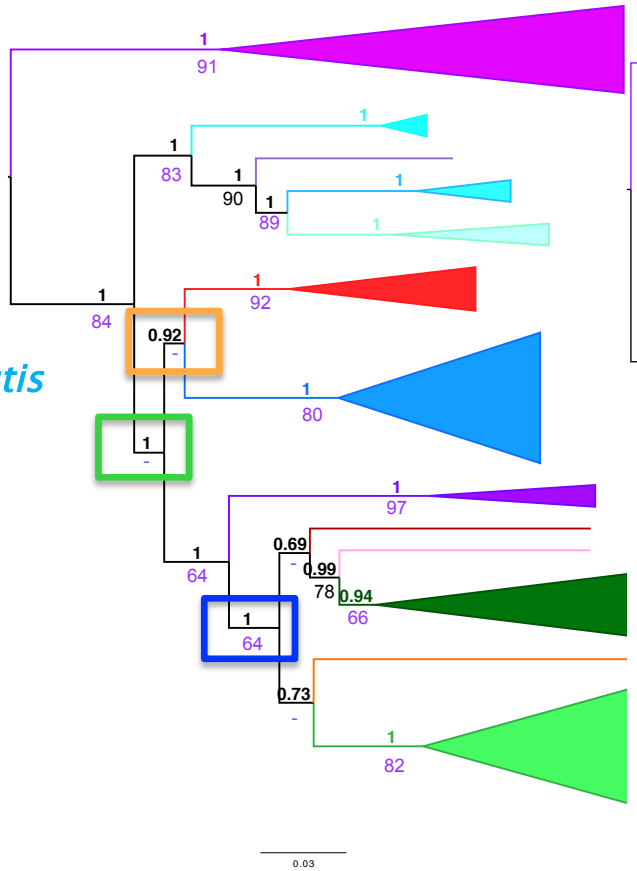
An empirical example: metalmark moths (Lepidoptera, Choreutidae), ca. 600 known species

Analysis: MrBayes, RAxML
Branch support: Bayesian PP, bootstrap

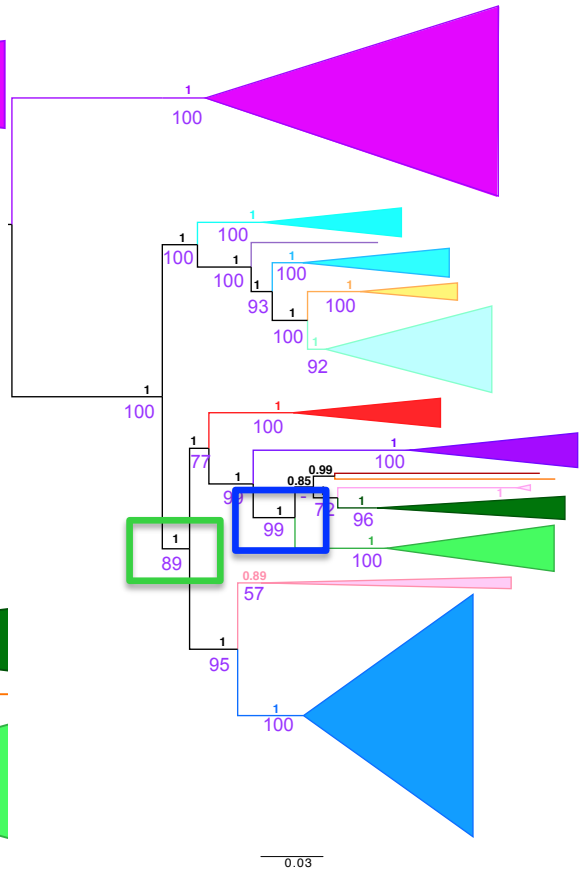
Comparison: BPs and PPs



3 genes,
42 taxa (Rota 2011)



8 genes,
38 taxa (Rota & Wahlberg 2012)



11 genes,
146 taxa (Rota unpubl.)

An empirical example: metalmark moths (Lepidoptera, Choreutidae), ca. 600 known species

Analysis: MrBayes, RAxML
Branch support: Bayesian PP, bootstrap

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

- 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
- Xu & Yang. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. GENETICS 204(4): 1353-1368. doi.org/10.1534/genetics.116.190173
- Yang & Rannala. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13, 303-314. doi:10.1038/nrg3186