Lecture 5:
Properties of DNA data and assessing robustness of hypotheses

Jadranka Rota and Niklas Wahlberg

Systematic Biology Group

Department of Biology

Lund University

LUND UNIVERSITY

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?**
  - **Include representative of all major clades**
  - **Iterative process: lab work for a set of taxa, preliminary analyses can inform further sampling**
- **Know what gene sequences are available from previous studies**
  - **Databases: GenBank, BOLD (DNA barcodes)**
  - **So you're not duplicating efforts**

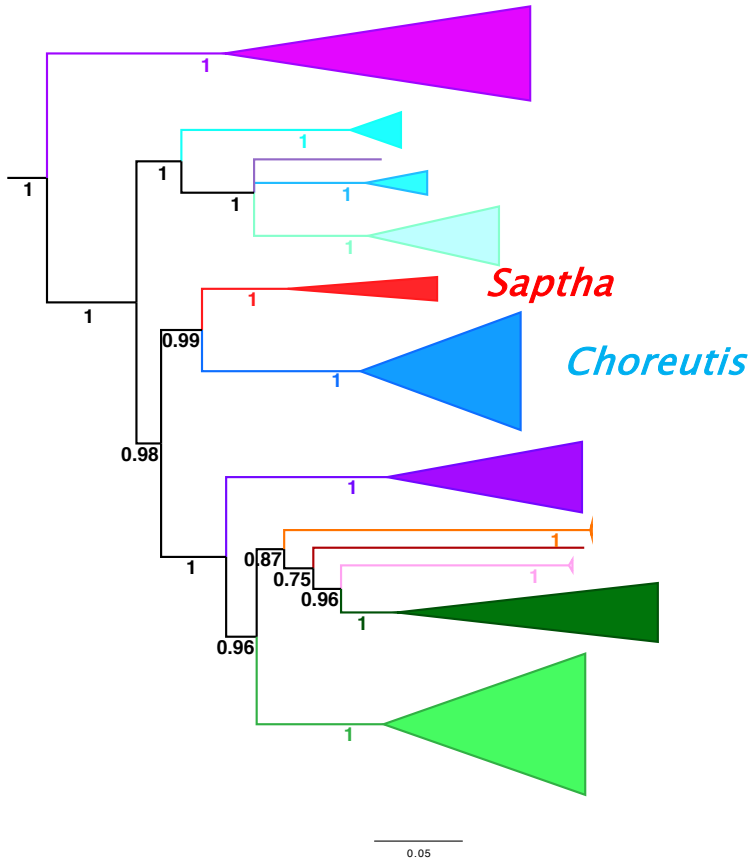
Number of genes

- **Single gene datasets – not very good for resolving phylogenies**
  - **Very rare nowadays**
- **Mitochondrial and chloroplast DNA used to be very 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**

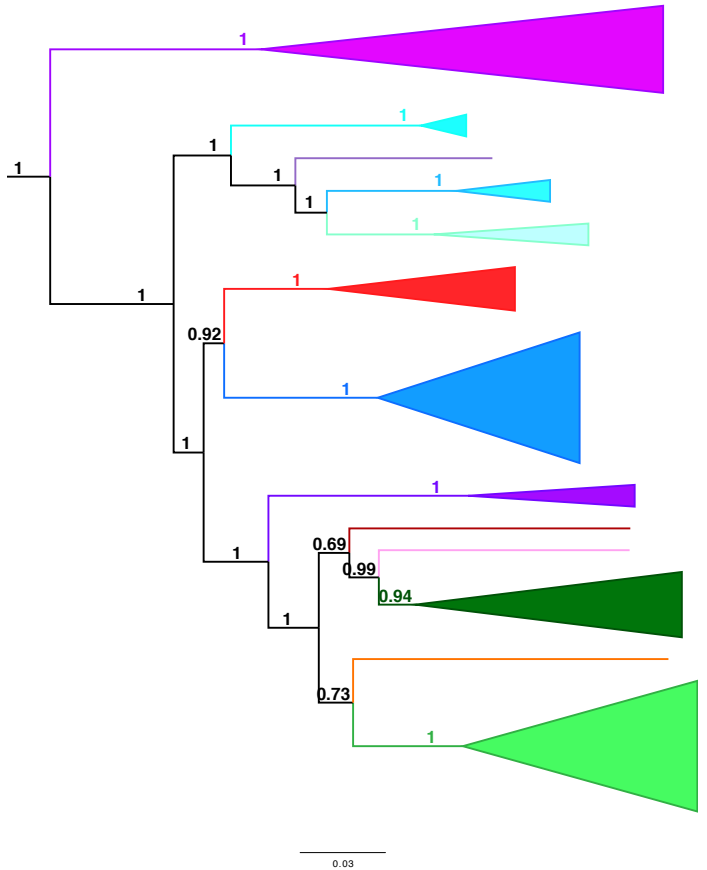
Number of taxa

- **What is good taxon sampling? – 10%, 20%, 50% of extant taxa?**
  - **The more, the better (usually)**
  - **Again, it depends on your question**
  - **Important to sample across your group of interest, not to leave out entire lineages**
- **Level of taxon sampling – different across different groups in the literature**
  - **Dense taxon sampling in well known groups – vertebrates, plants, some insect groups (e.g. butterflies)**
  - **Relatively low taxon sampling – many invertebrate groups**

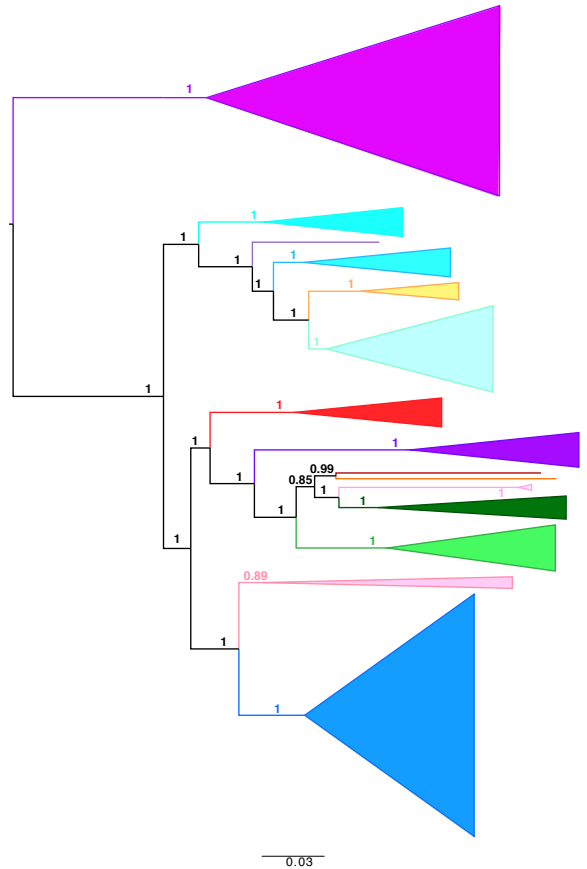
Adding genes and adding taxa helps!



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’**
  - **’rogue taxa’ – taxa that jump around in the phylogeny**

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]

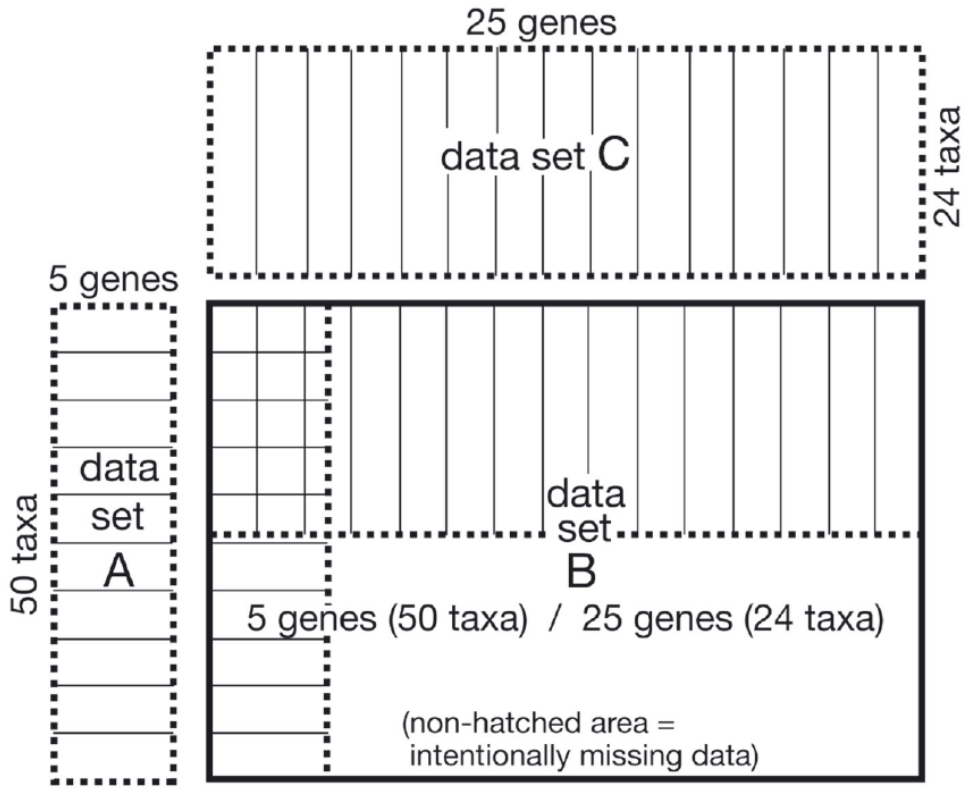


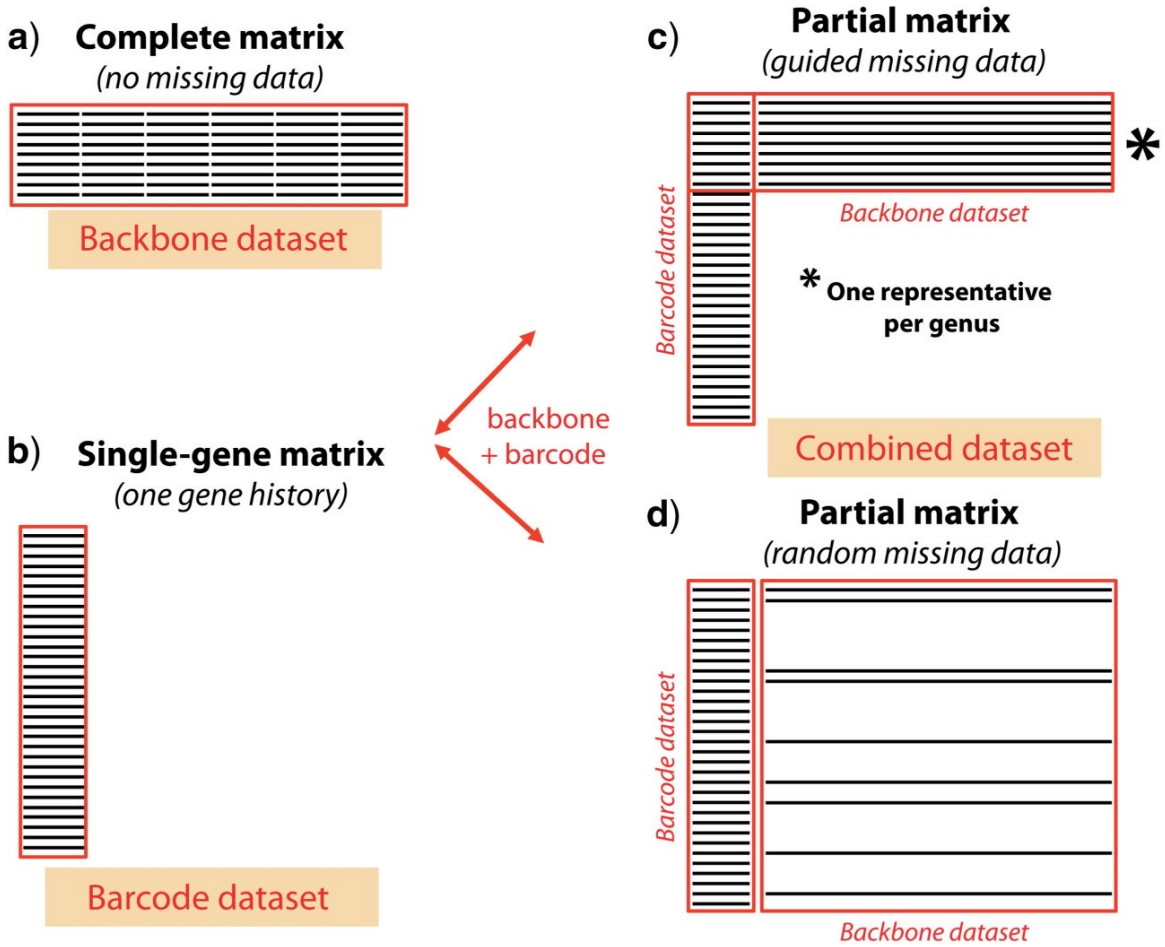
Figure 1. Distribution of missing data in molecular matrices. a) A complete matrix, where no missing data are involved (referred as the backbone data set). b) A single-gene matrix, including only one molecular marker and therefore providing information about only one gene history (referred as the barcode data set). c) The combined matrix, the product of merging a backbone and a barcode data set…

CORRECTED PROOF

DNA Barcodes Combined with Multilocus Data of Representative Taxa Can Generate Reliable Higher-Level Phylogenies

Gerard Talavera ✉, Vladimir Lukhtanov, Naomi E Pierce, Roger Vila

Systematic Biology, syab038, https://doi.org/10.1093/sysbio/syab038
Published: 19 July 2021 **Article history** ▾



Phylogenomics

- **Number of genes sequenced is in hundreds or thousands**
- **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 work well**

Is having thousands of genes enough to resolve phylogenies?

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths

Akito Y. Kawahara, David Plotkin, Marianne Espeland, +16, and Jesse W. Breinholt Authors Info & Affiliations

Edited by Douglas Futuyma, Stony Brook University, Stony Brook, NY, and approved September 16, 2019 (received for review May 29, 2019)

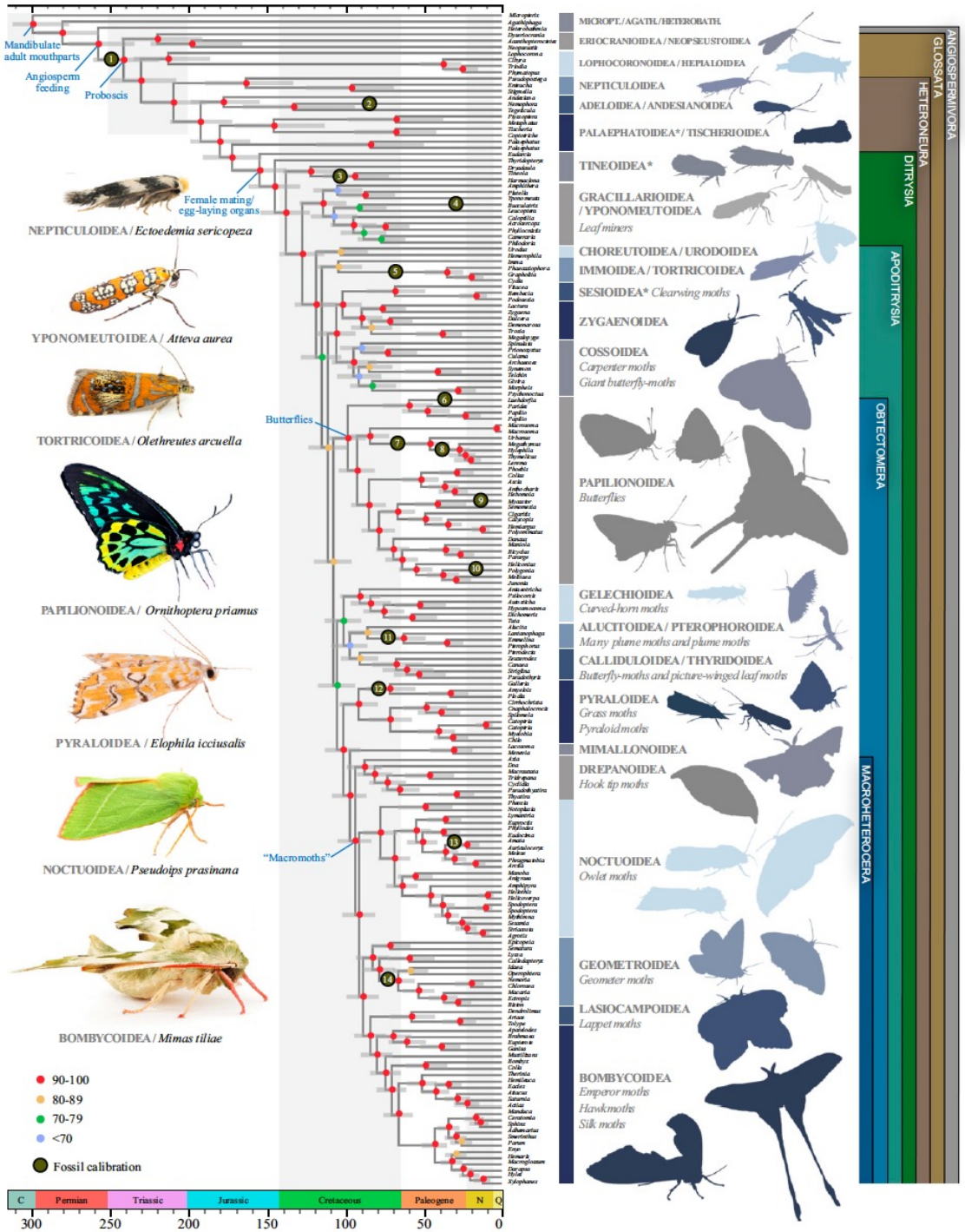
October 21, 2019 | 116 (45) 22657-22663 | https://doi.org/10.1073/pnas.1907847116

Whole-genome analyses resolve early branches in the tree of life of modern birds

ERICH D. JARVIS , SIAVASH MIRARAB, ANDRE J. ABERER, BO LI, PETER HOUDE, CAI LI, SIMON Y. W. HO, BRANT C. FAIRCLOTH, BENOIT NABHOLZ, [...], AND GUOJIE ZHANG

+95 authors Authors Info & Affiliations

SCIENCE • 12 Dec 2014 • Vol 346, Issue 6215 • pp. 1320-1331 • DOI: 10.1126/science.1253451

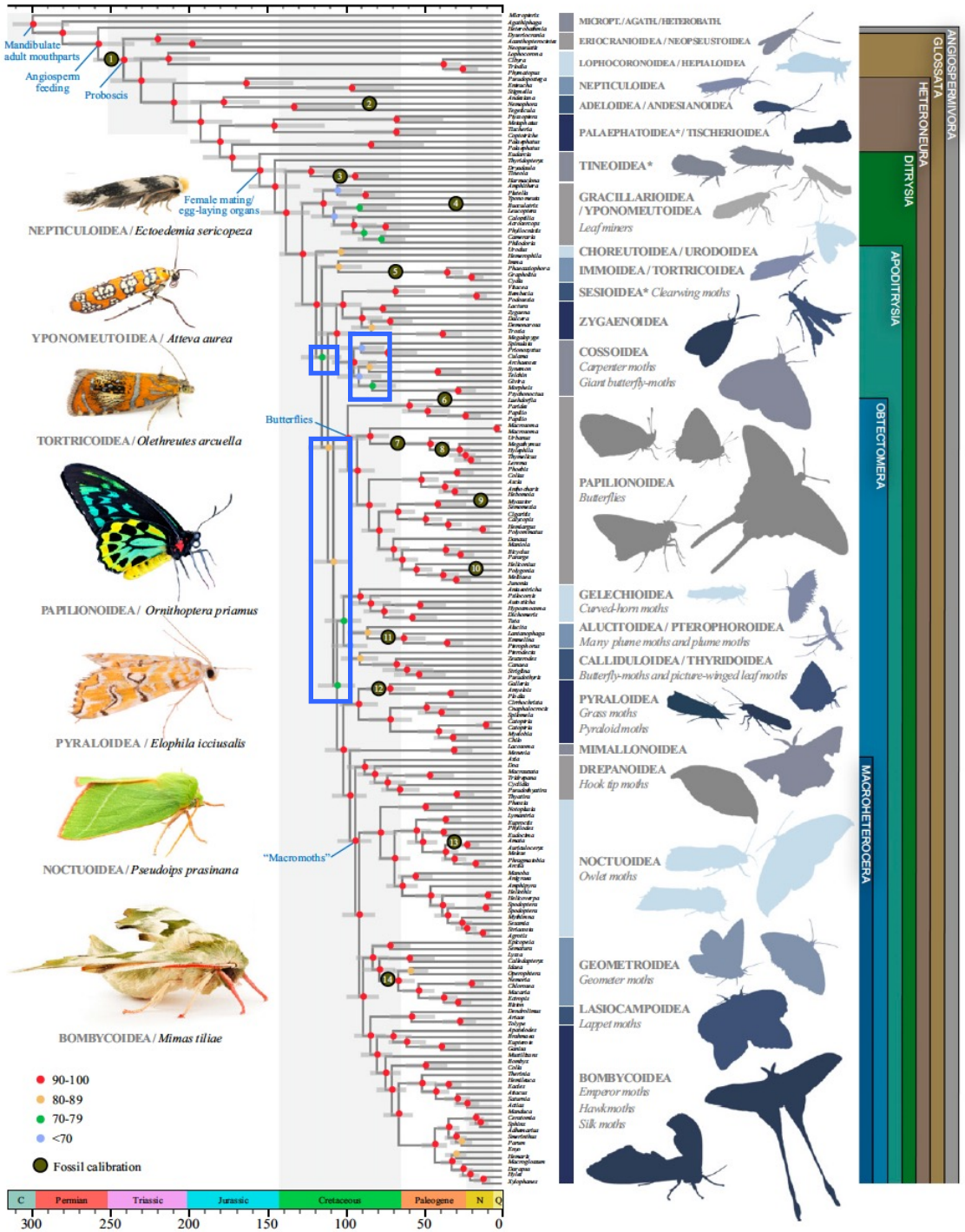


Kawahara et al. 2019 PNAS
Phylogeny of Lepidoptera

Bootstrap support – in red significant support

- 90-100
- 80-89
- 70-79
- <70

Fig. 1. Dated evolutionary tree of butterfly and moth relationships. The tree is derived from a maximum-likelihood analysis of 2098 genes (amino acid alignment).



Kawahara et al. 2019 PNAS
Phylogeny of Lepidoptera

Bootstrap support – in red significant support

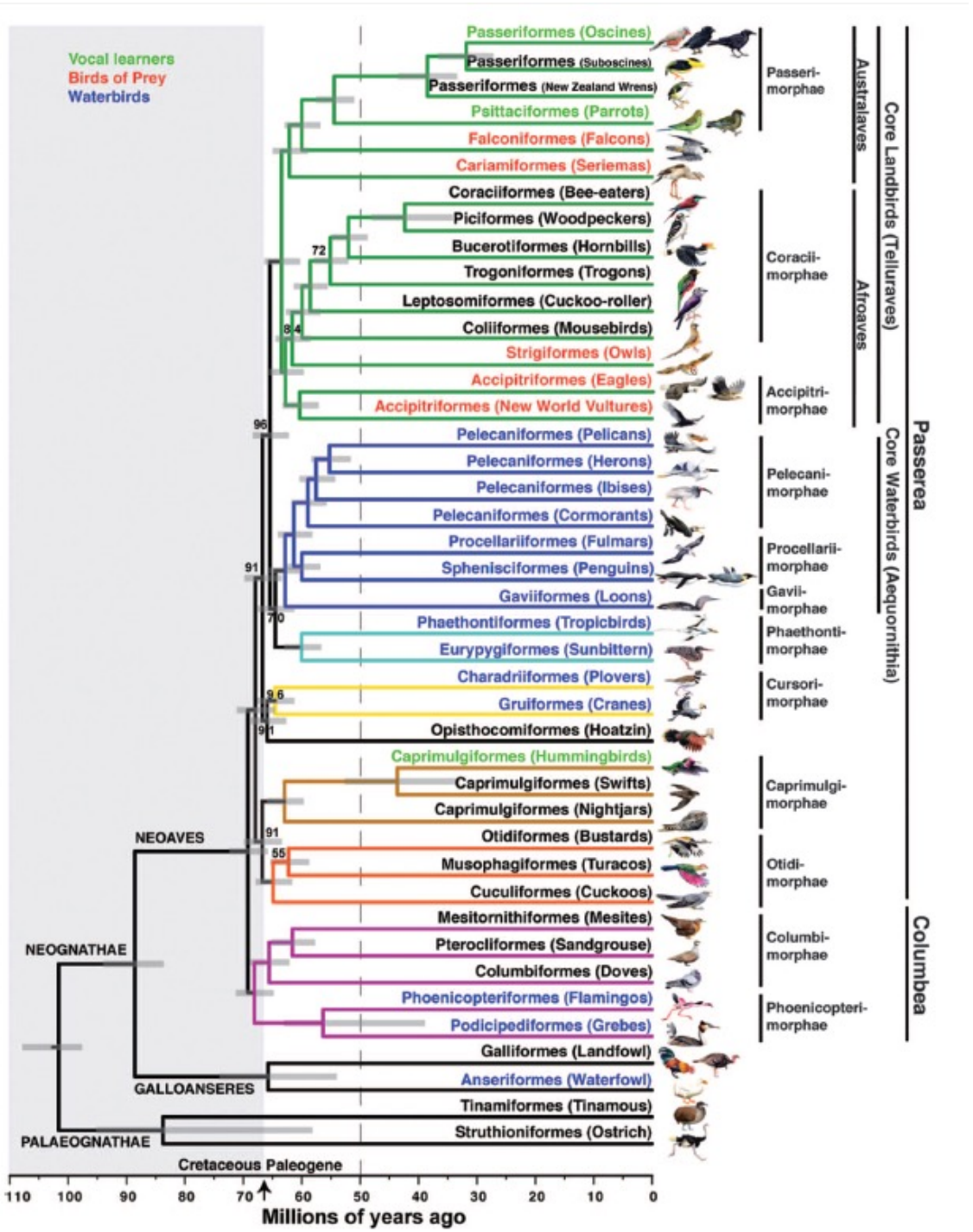
- 90-100
- 80-89
- 70-79
- <70

Fig. 1. Dated evolutionary tree of butterfly and moth relationships. The tree is derived from a maximum-likelihood analysis of 2098 genes (amino acid alignment).

8251 genes
2516 introns
3769 ultraconserved elements

41.8 million bp…

Jarvis et al. 2014: Science 346

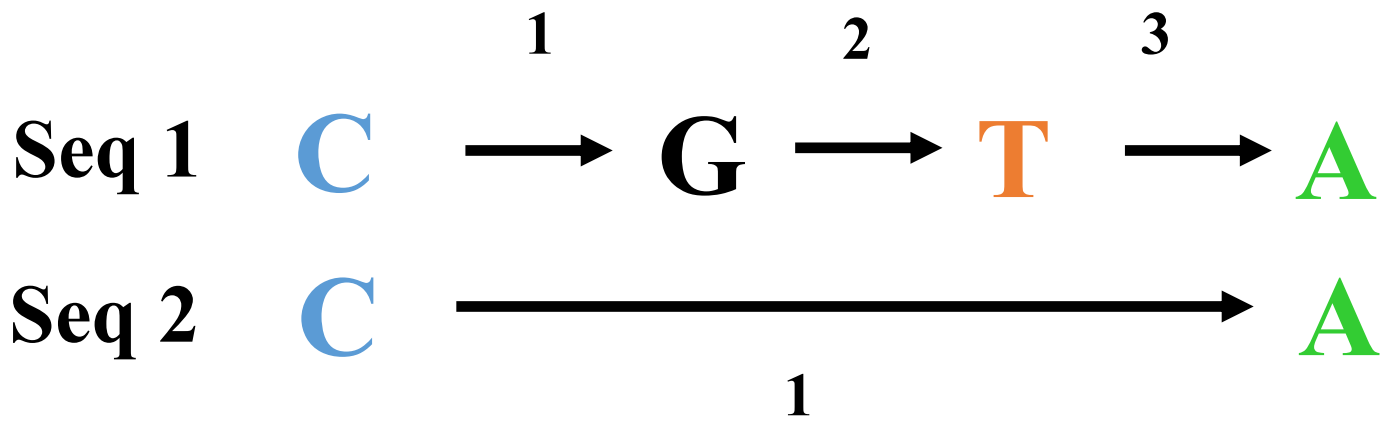


Properties of DNA data important to consider

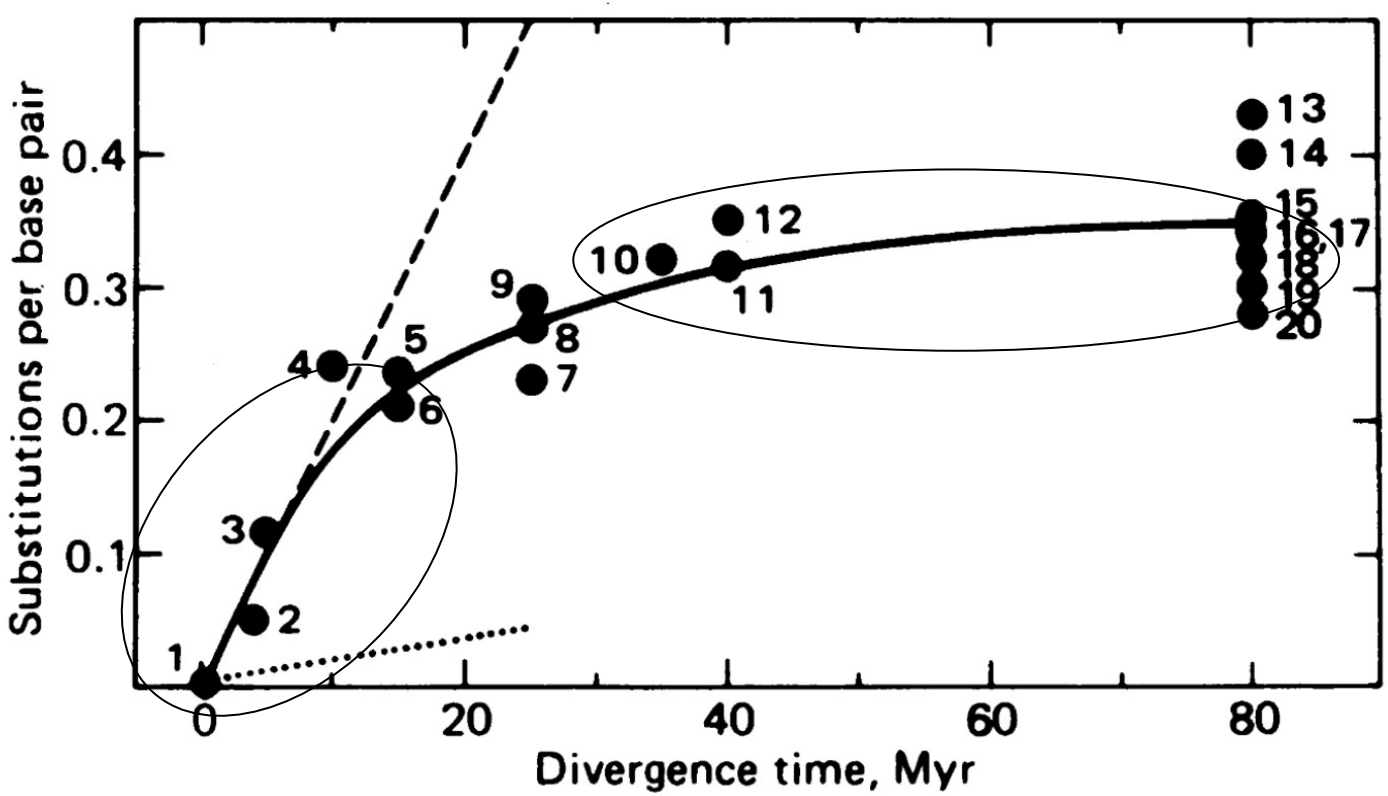
- **Saturation and long-branch attraction**
- **Incomplete lineage sorting**
- **Lateral gene transfer**
- **Mito-nuclear discordance**
- **Biased base composition**

Multiple changes at a single site – hidden changes

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



“Multiple hits” or saturation



Rapid evolution of animal mitochondrial DNA

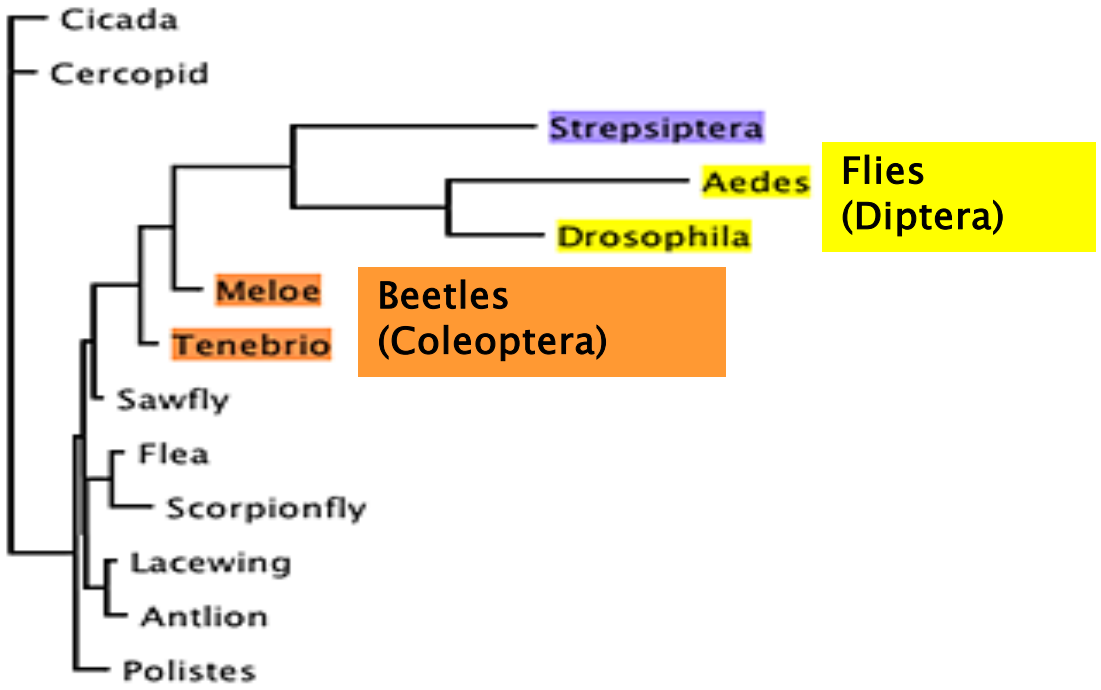
by WM Brown · 1979 · Cited by 4629 — **Mitochondrial DNA** was purified from four species of higher primates (Guinea baboon, rhesus macaque, guenon, and human) and digested with 11…

Brown et al. 1979. PNAS 76:1967

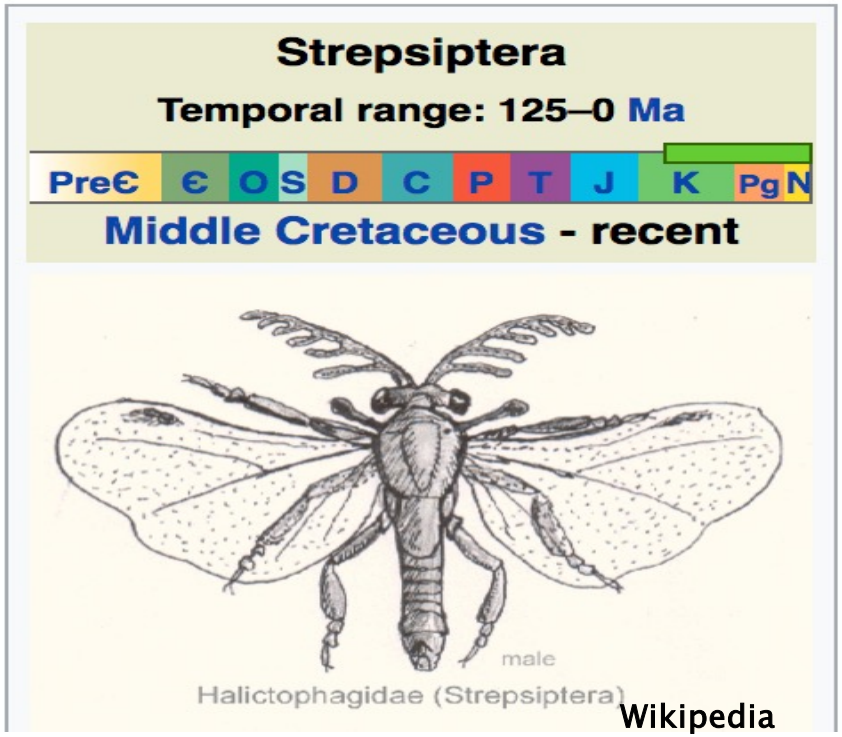
Saturation and long-branch attraction

- **Homoplasy (incorrectly inferred homology)**
  - **One of the issues with molecular data**
  - **Made worse by 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

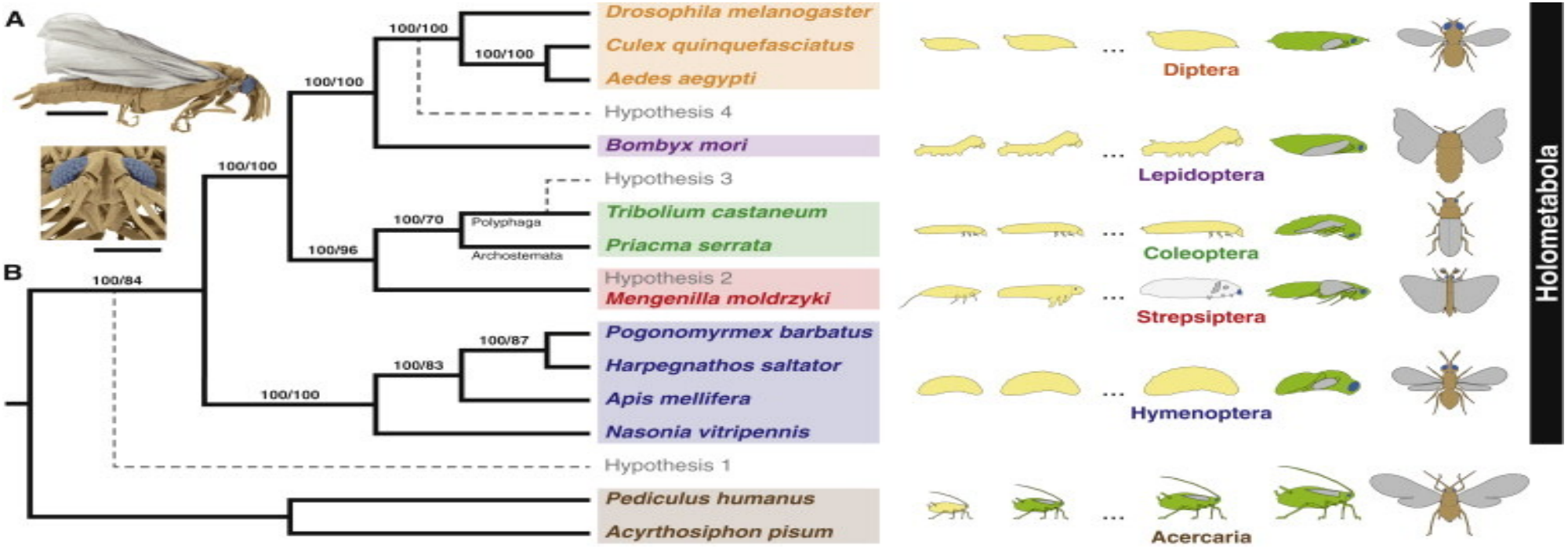


Figure 1. Evolutionary Origin of Twisted-Wing Parasites Inferred from Genomic Evidence(A) Mengenilla moldrzyki male in lateral (top; scale bar represents 1 mm) and frontal (bottom; scale bar represents 500 μm) view (colored SEM micrographs; wings in gray, comp...

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?

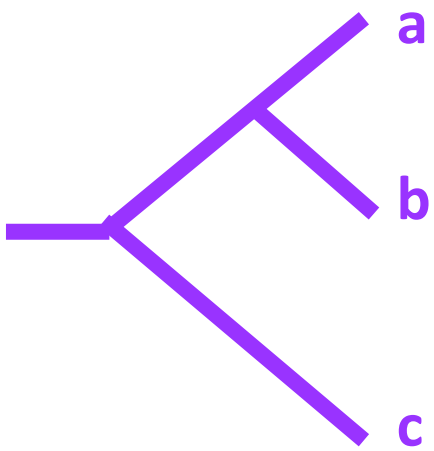
- **Modelling DNA evolution**
- **Taxon sampling is important – whenever possible break up long branches**
- **For divergent taxa with few extant species, this can be a big problem**
  - **BUT branch support is usually low for long branches sticking together in model-based methods – so we should be able 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?

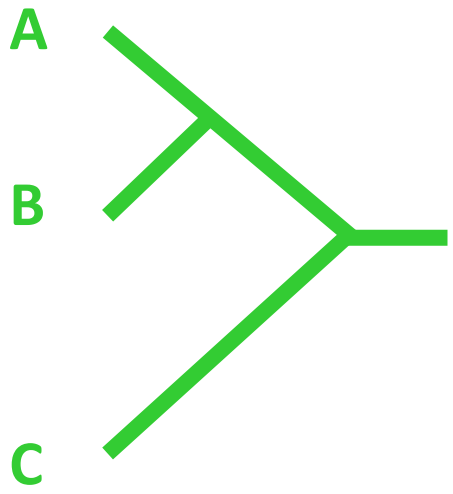
- **Are the genome regions sequenced from different species the same (homologous)?**
- **Gene duplication**
  - **1) duplicate gene degenerates – pseudogene**
  - **2) duplicate gene aquires new function**
- **A problem particularly accute currently as we analyze phylogenomic data**

Orthology: gene trees and species trees

Gene phylogeny



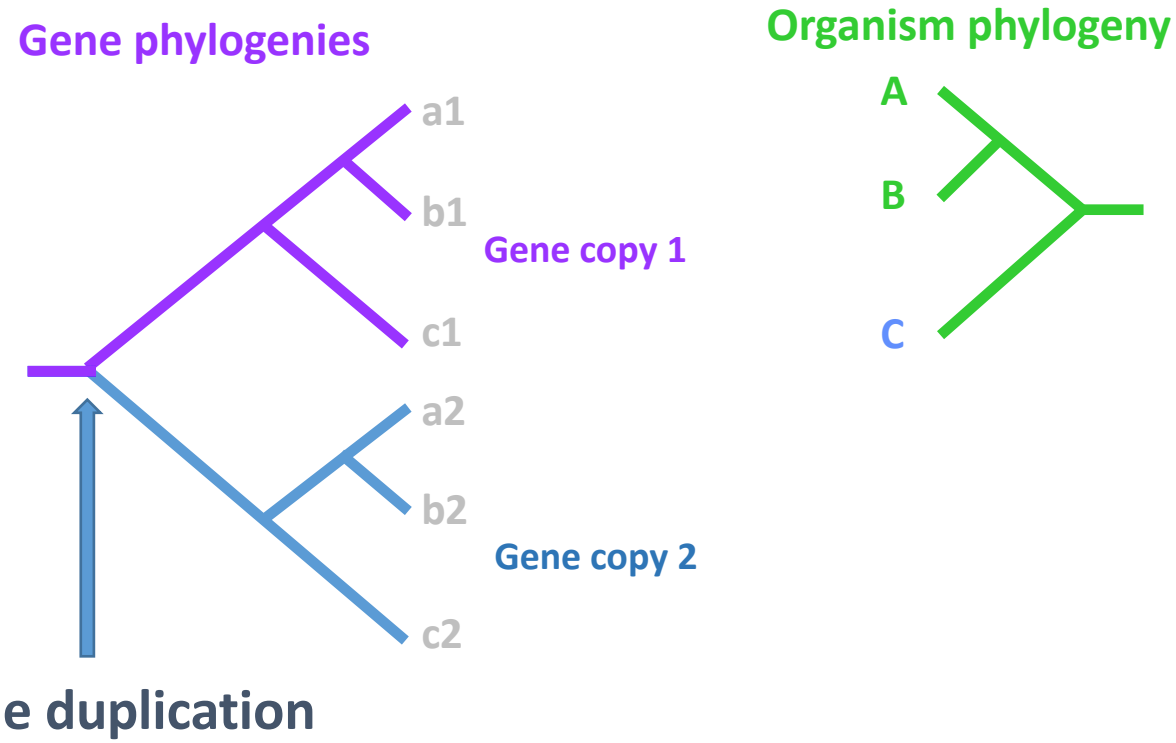
Organism phylogeny



ORTHOLOGY

Orthologs: genes that arose due to speciation

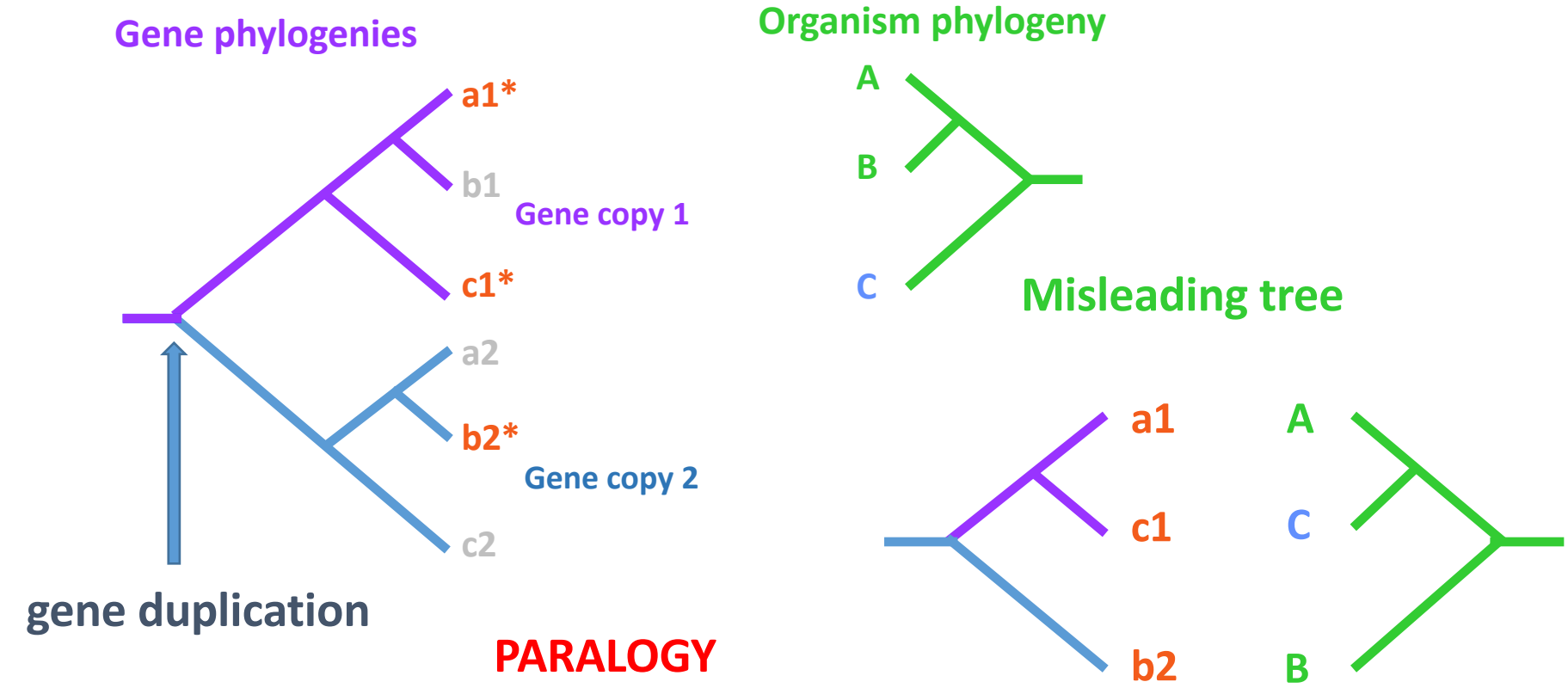
Paralogy: can produce misleading trees



PARALOGY

Paralogs: genes that arose due to duplication events

Paralogy: can produce misleading trees



Paralogs: genes that arose due to duplication events

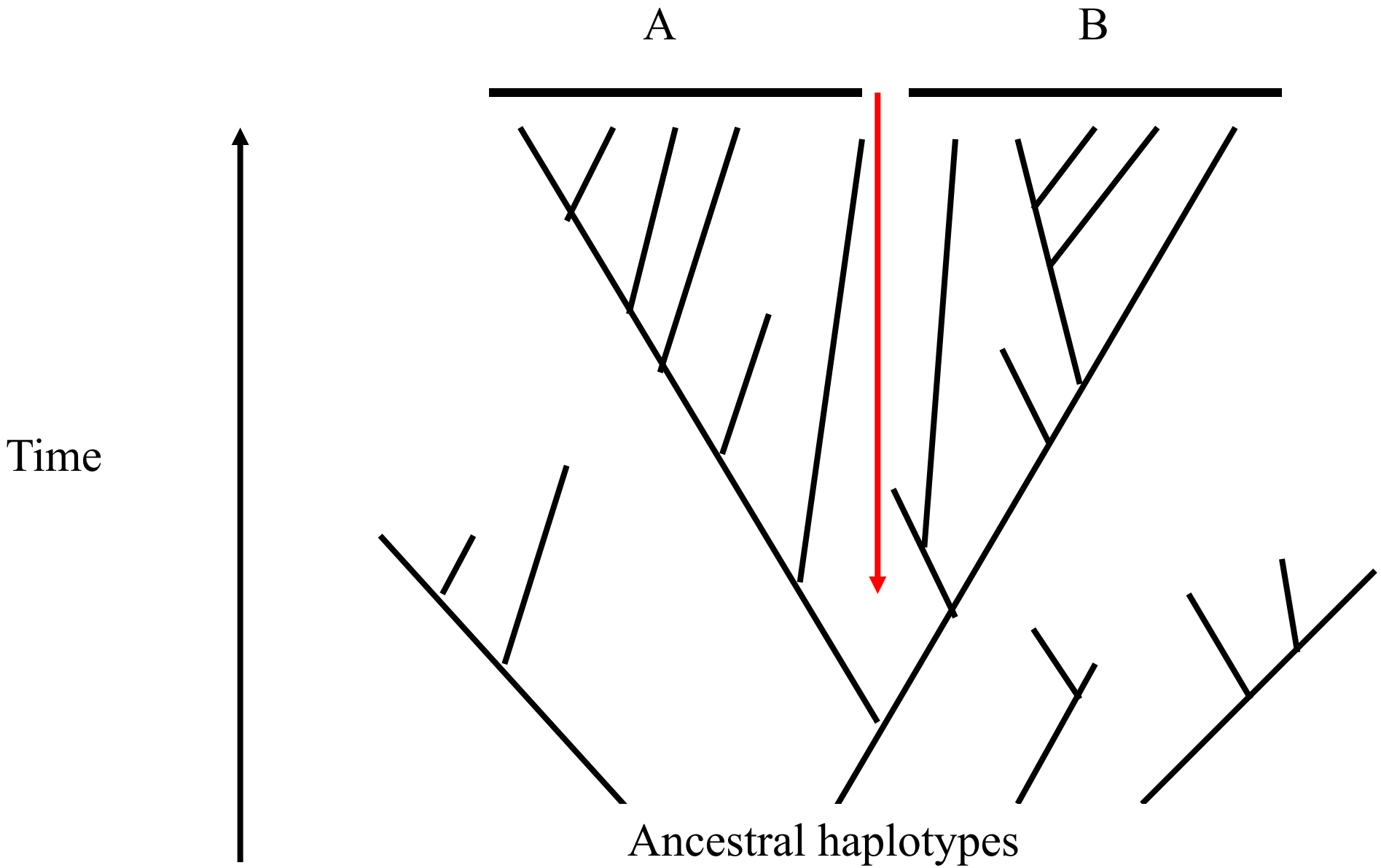
Incomplete lineage sorting (ILS)

- **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**

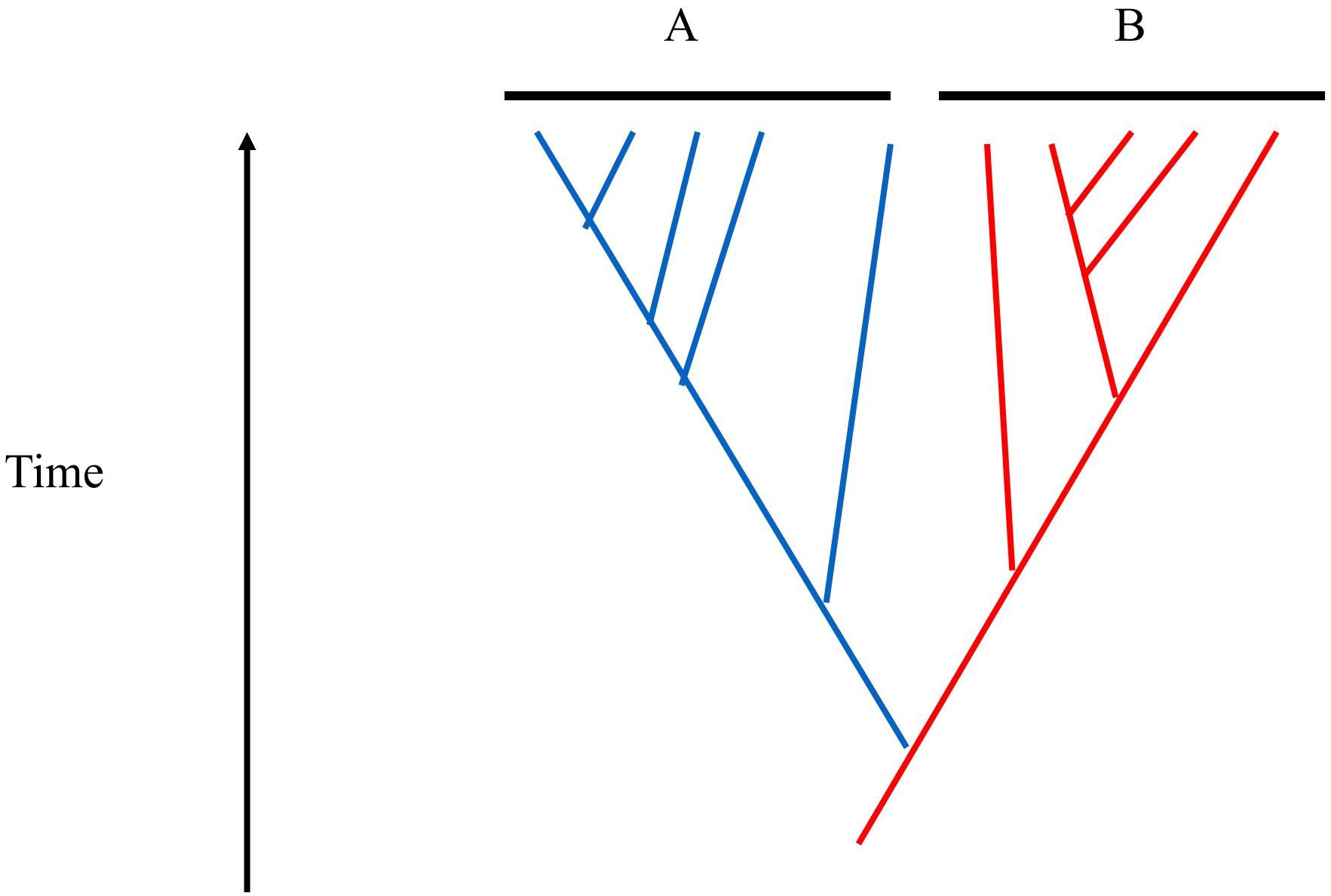
Are species monophyletic?

- **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



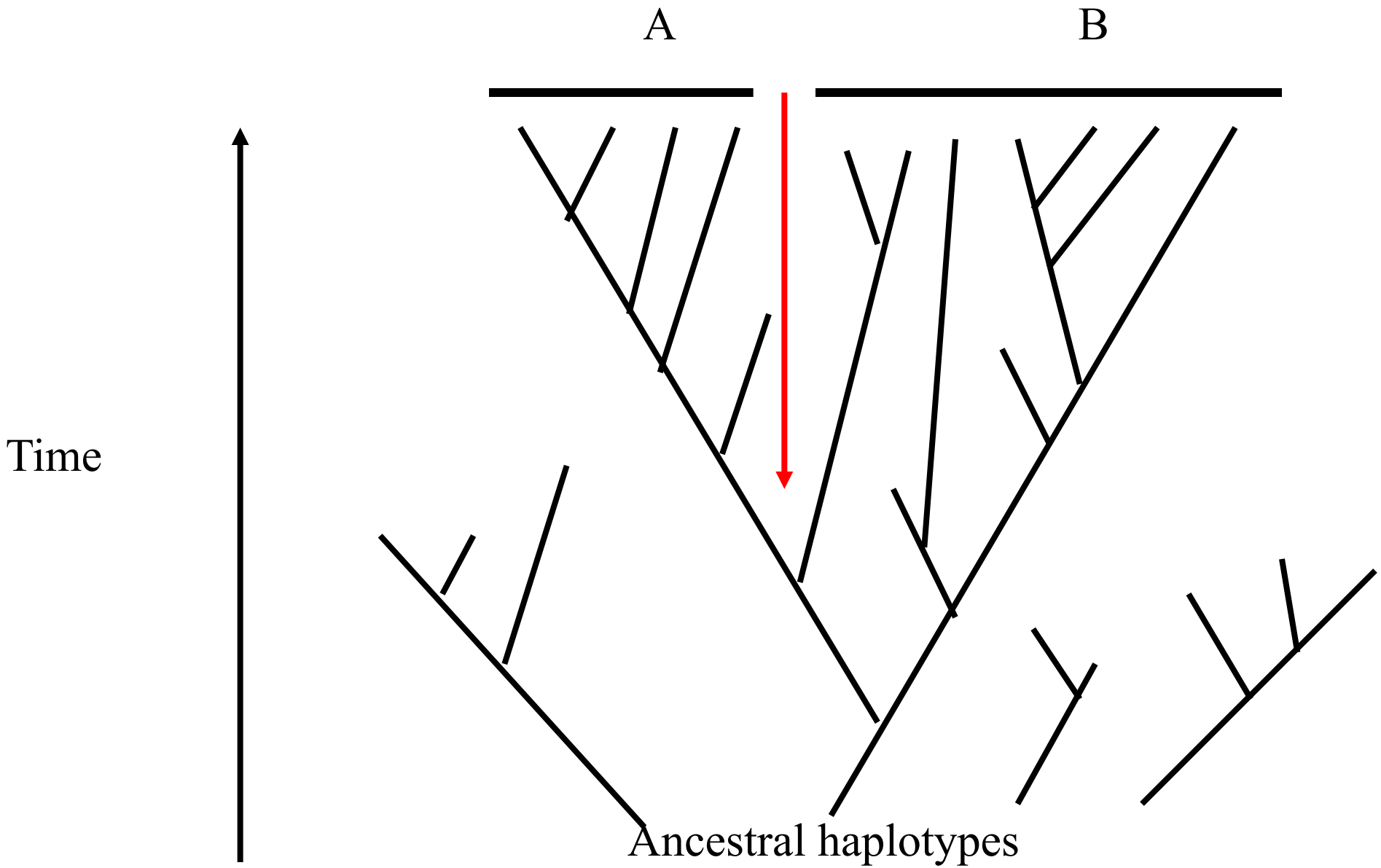
The assumption: monophyly



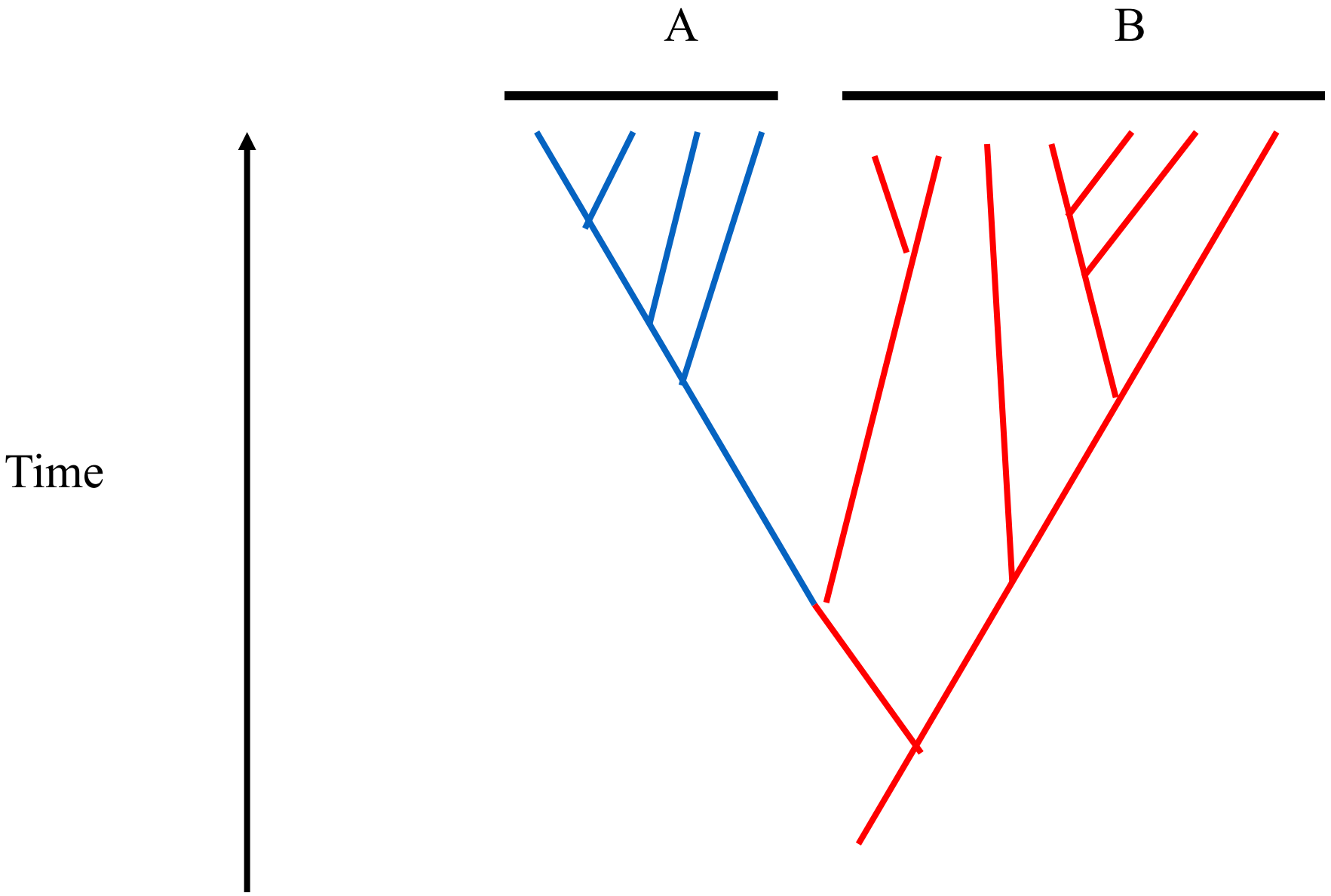
The presence of poly- and paraphyletic lineages

- **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**

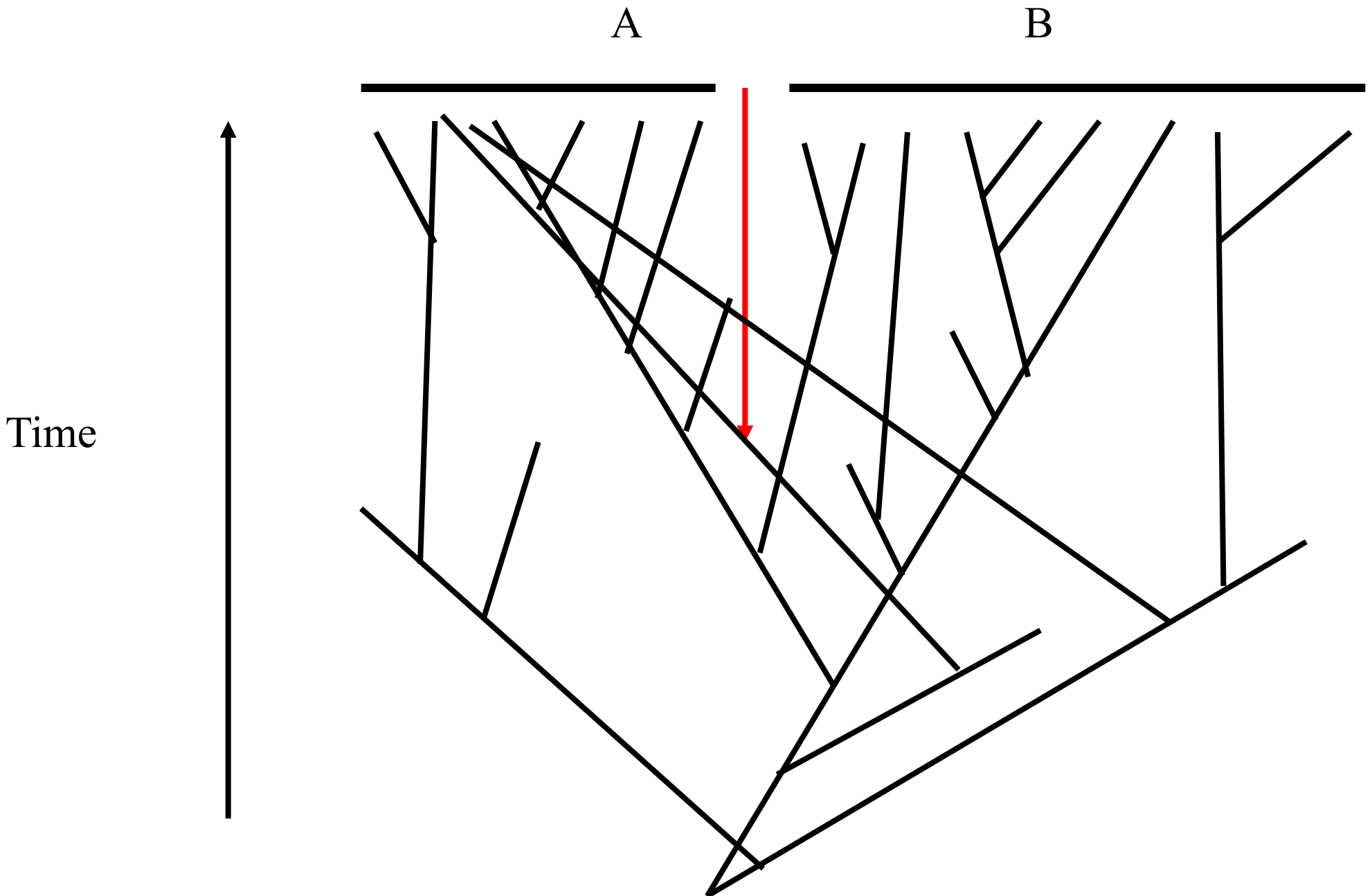
Paraphyly



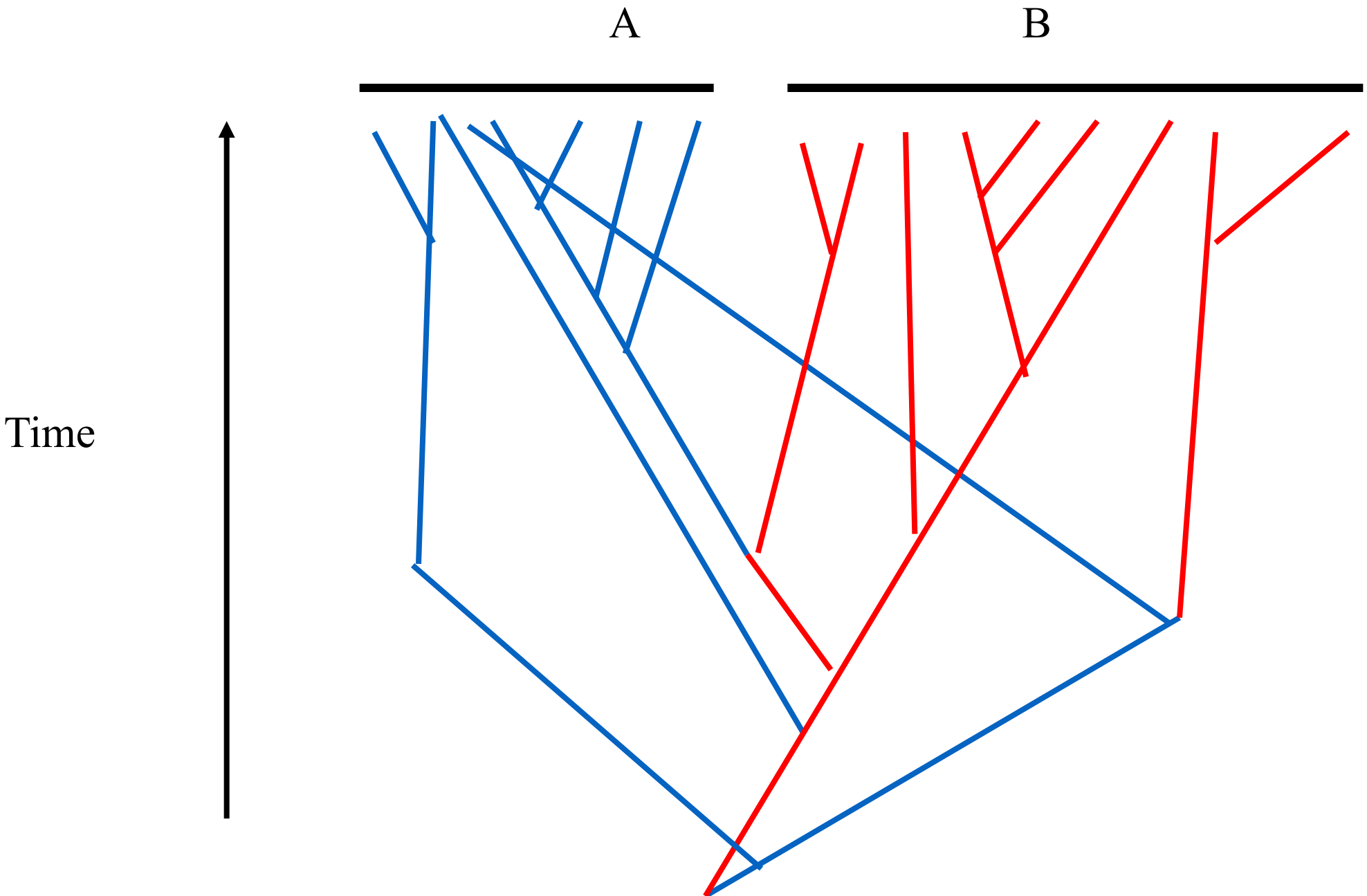
Paraphyly



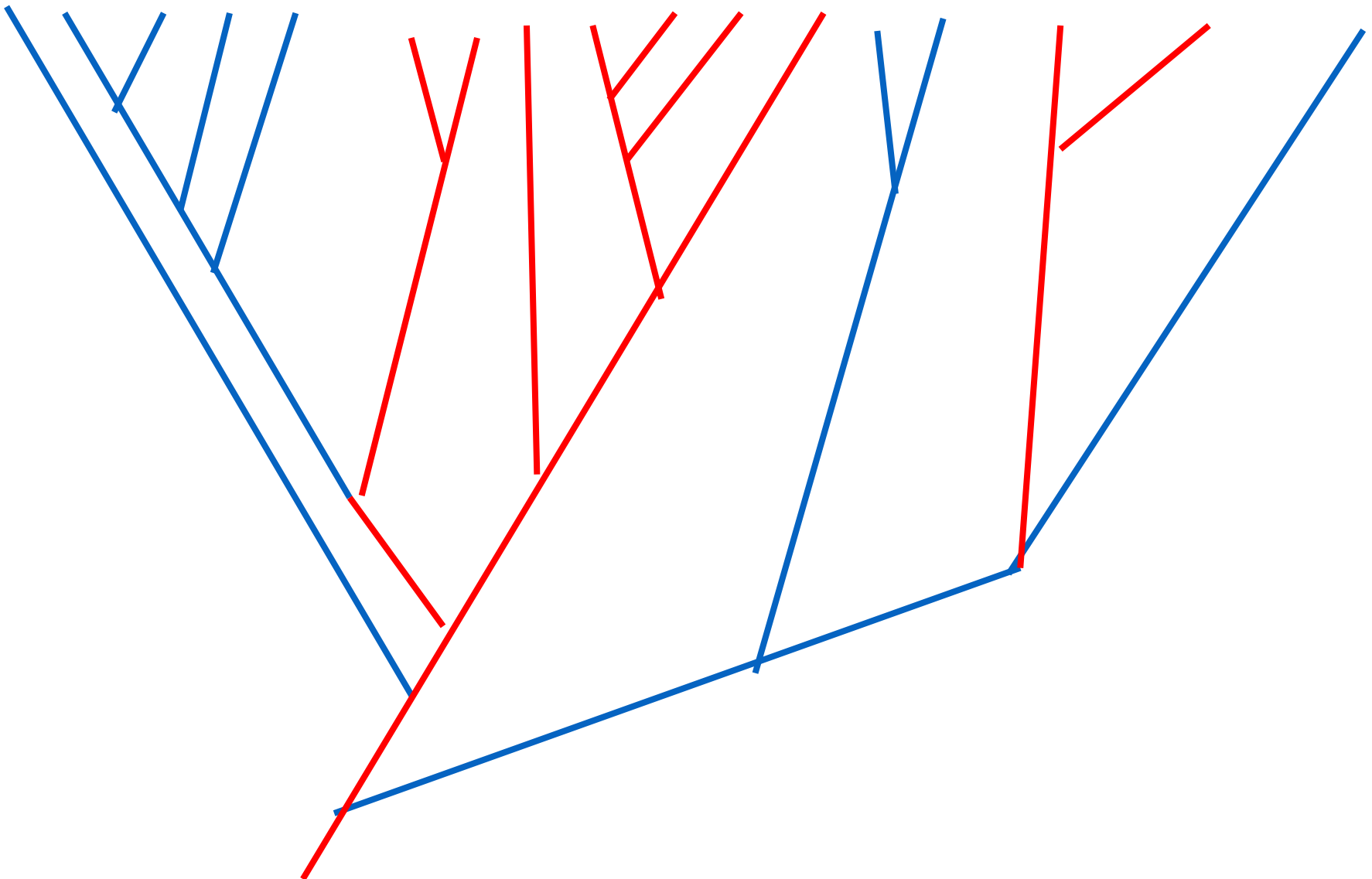
Polyphyly



Polyphyly



Polyphyly



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

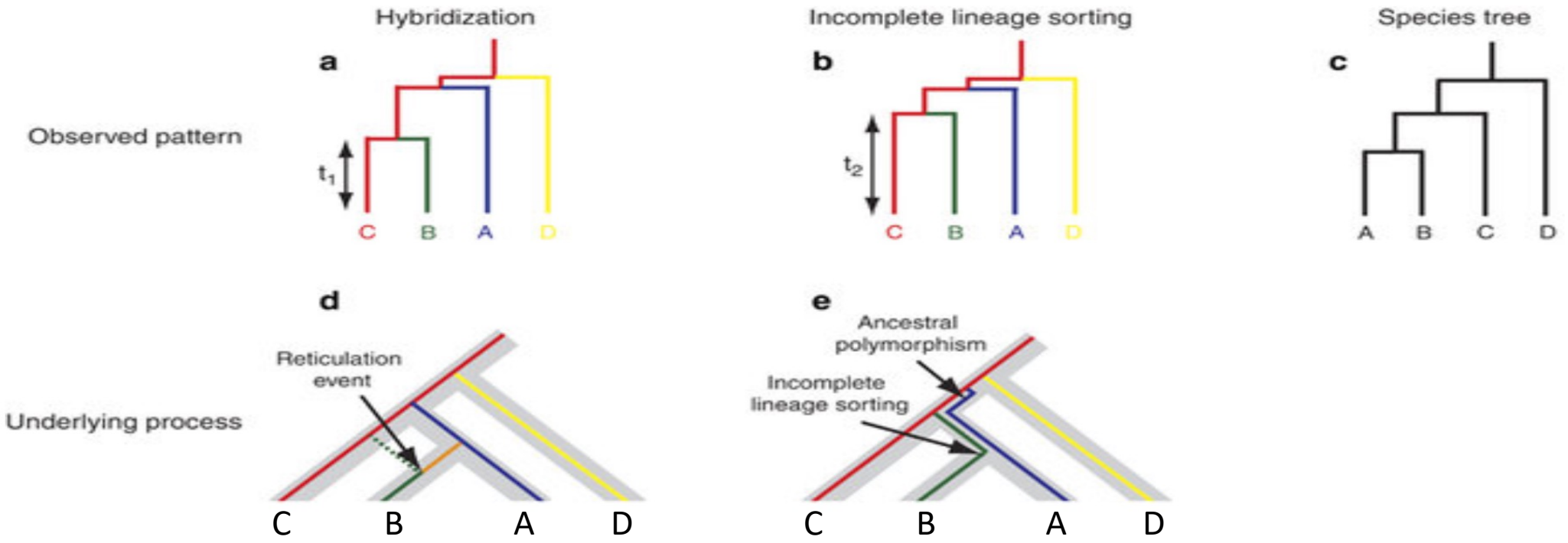
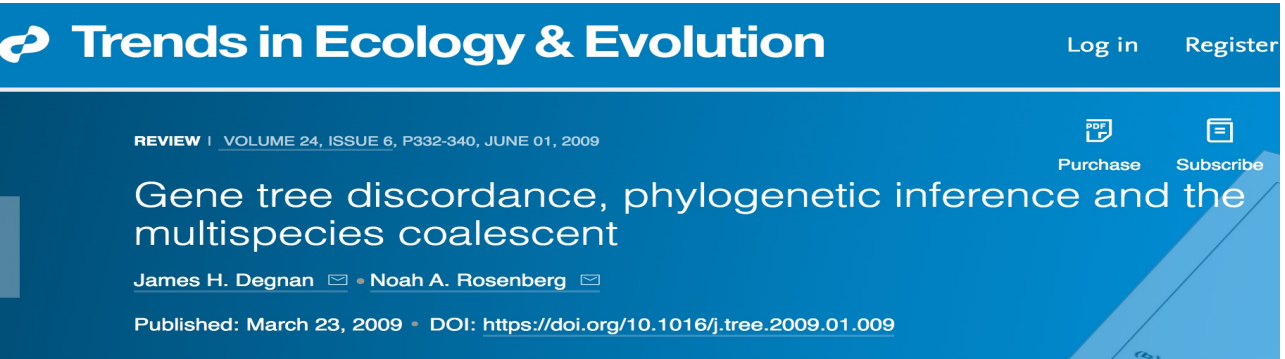
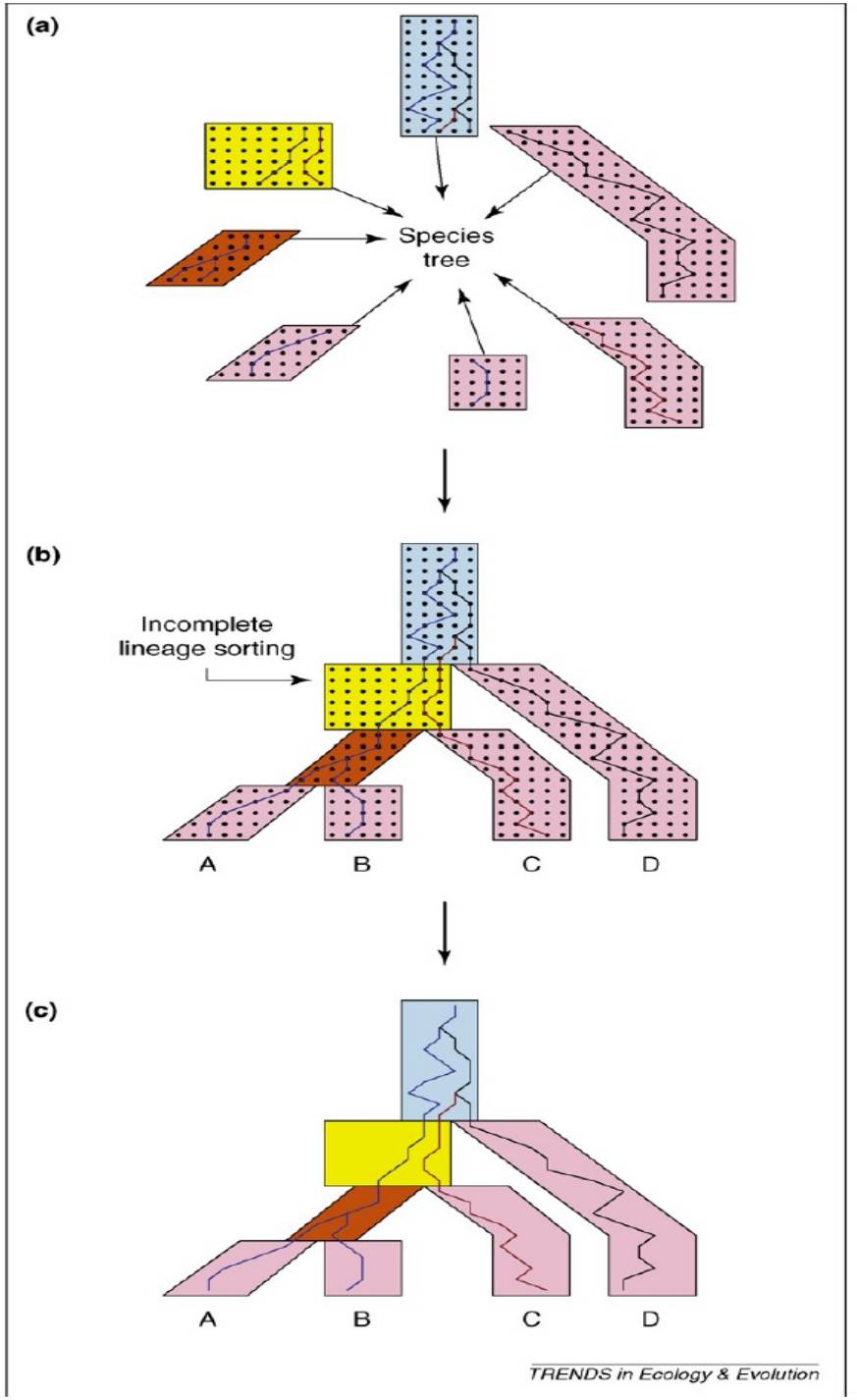


Figure from ResearchGate, uploaded by Richard A Ennos

Multispecies coalescent model (MSC)

- **Gene tree vs. species tree**
- **Model that accommodates gene tree heterogeneity caused by ILS**
- **ASTER package (ASTRAL IV)**





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.

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

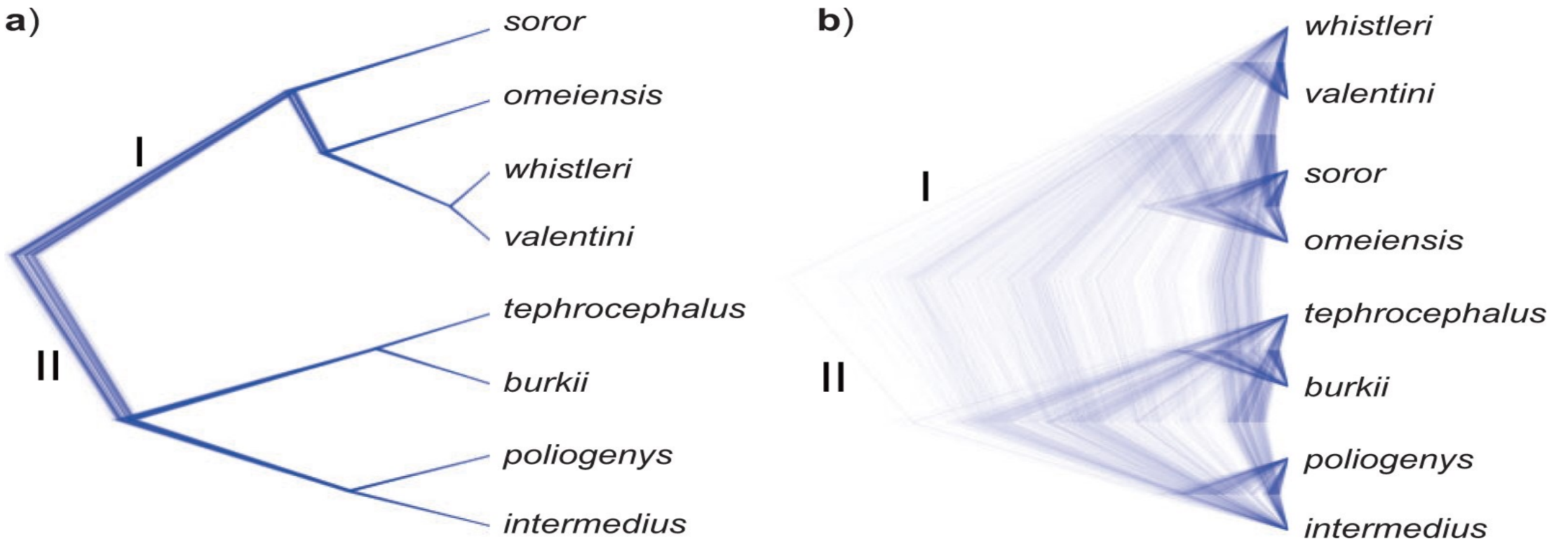
Dezhi Zhang, Frank E Rheindt, Huishang She, Yalin Cheng, Gang Song, Chenxi Jia, Yanhua Qu, Per Alström ✉, Fumin Lei ✉

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

Published: 31 March 2021 **Article history**

PDF Split View Cite Permissions Share

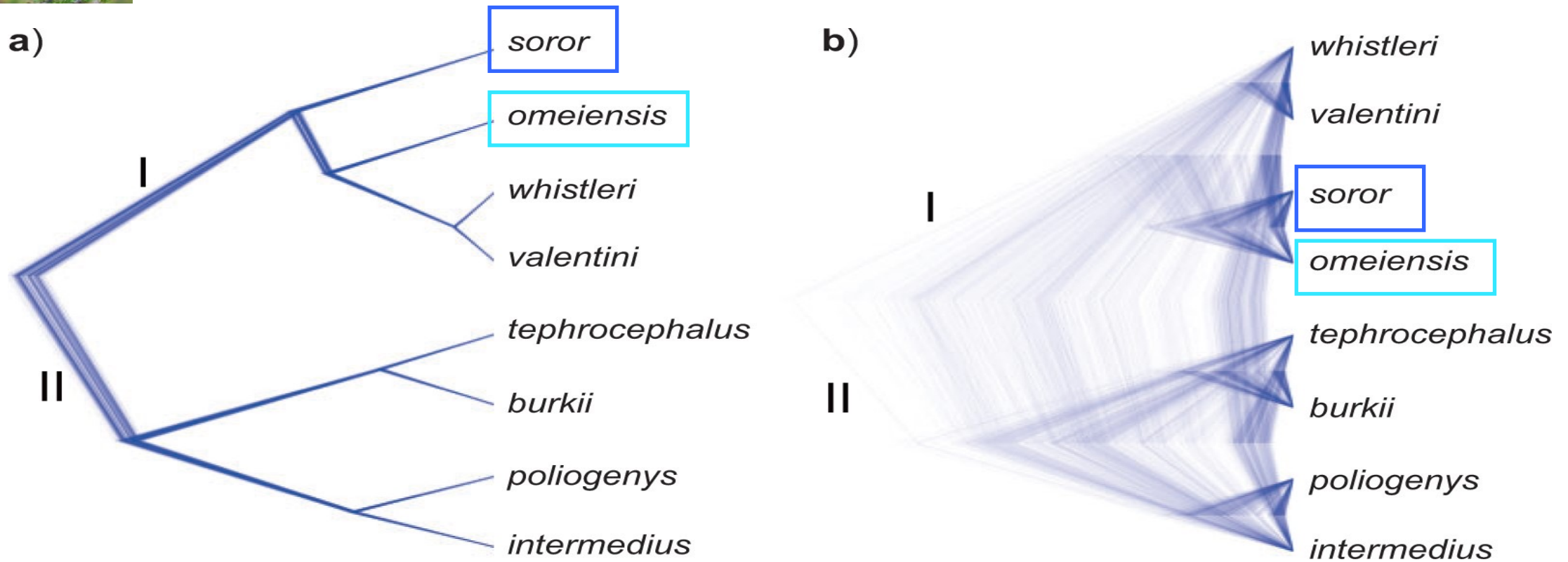
Leaf warblers



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

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

- **Topological differences between well resolved phylogenies from mitochondrial and nuclear genes**
- **Relatively widespread**
- **Biological reasons for it: incomplete lineage sorting and lateral gene transfer**
- **Operational reasons for it:**
  - **faster substitution rate in animal mitochondrial DNA compared to the nuclear**
  - **Too little phylogenetic signal for deeper relationships**

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

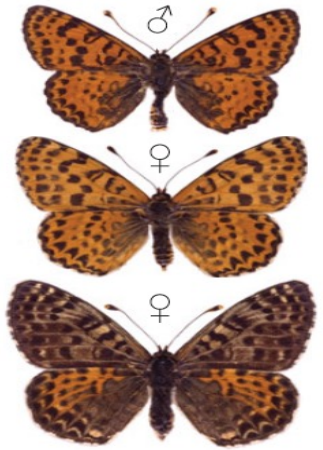
Dinca et al. (2019)

Proc. Roy. Soc B,

286: 20191311

Spotted fritillary (Melitaea didyma)

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

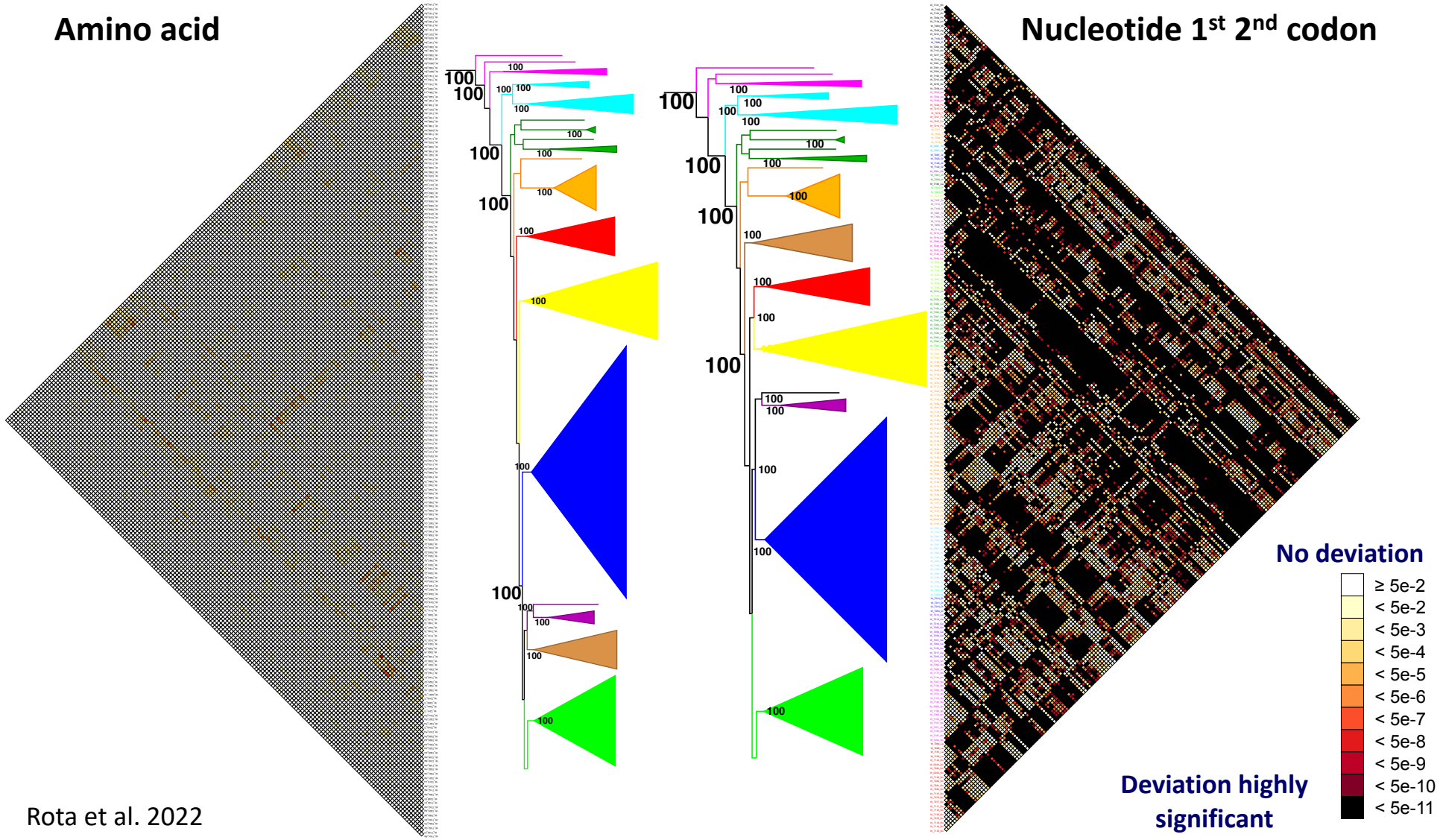
Biased base composition

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/low 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, 332 genes
Bowker's symmetry test



Rota et al. 2022

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

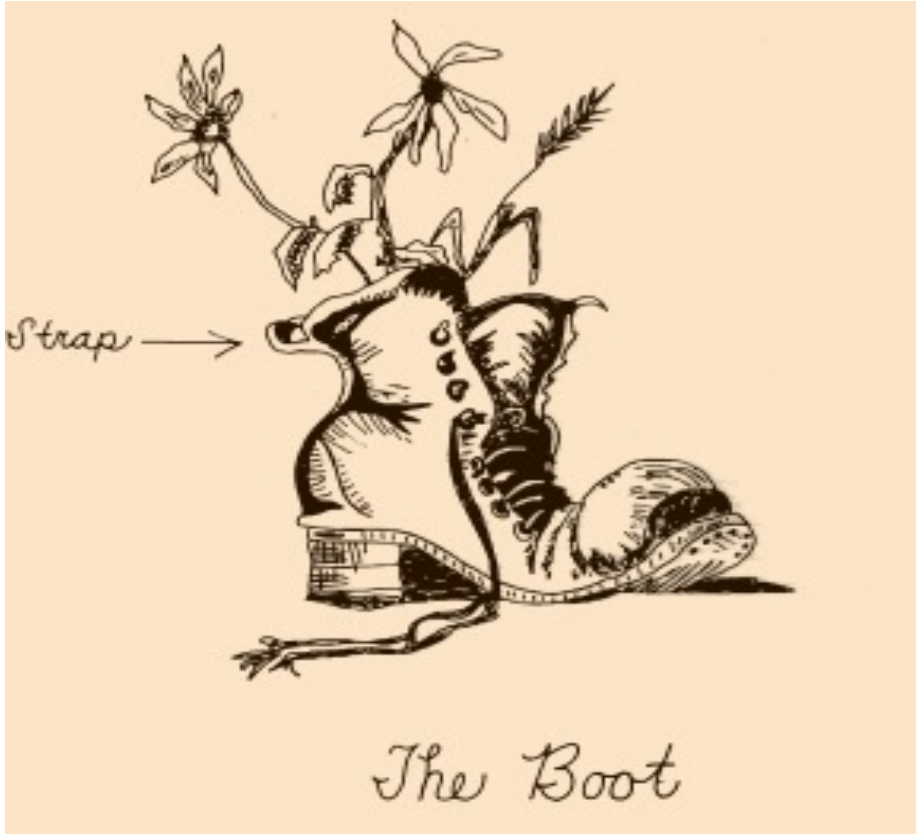
- **How strongly do the data support your phylogenetic hypothesis?**
- **How stable is your phylogenetic hypothesis?**
  - **Is it likely to change with the addition of new data?**
  - **Do you get the same result with different analysis methods?**

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

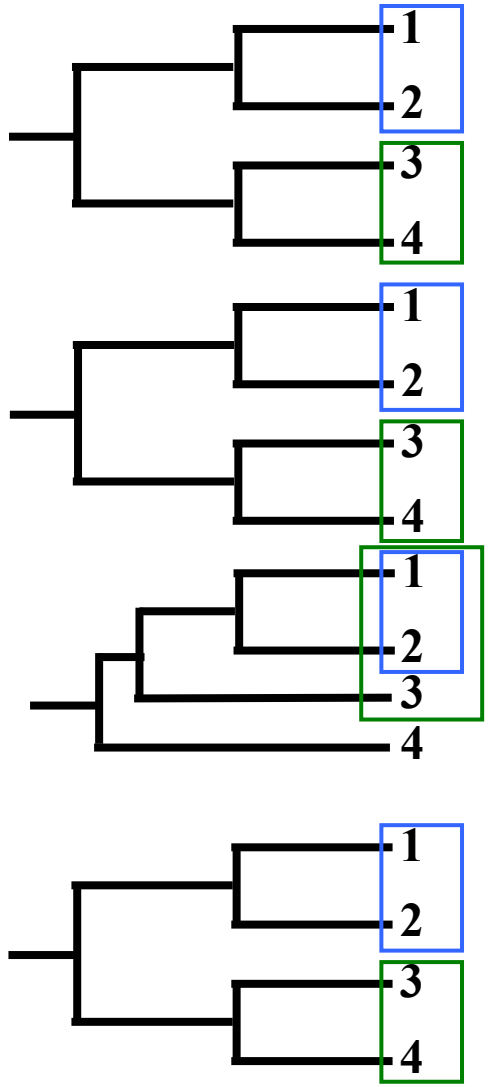
```
1 CGAGAA
2 AGAGAA
3 AGTTTT
4 GGATAA
```

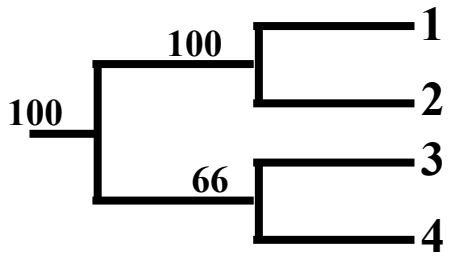
Pseudoreplicate 2

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

Pseudoreplicate n (e.g. 1000)

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

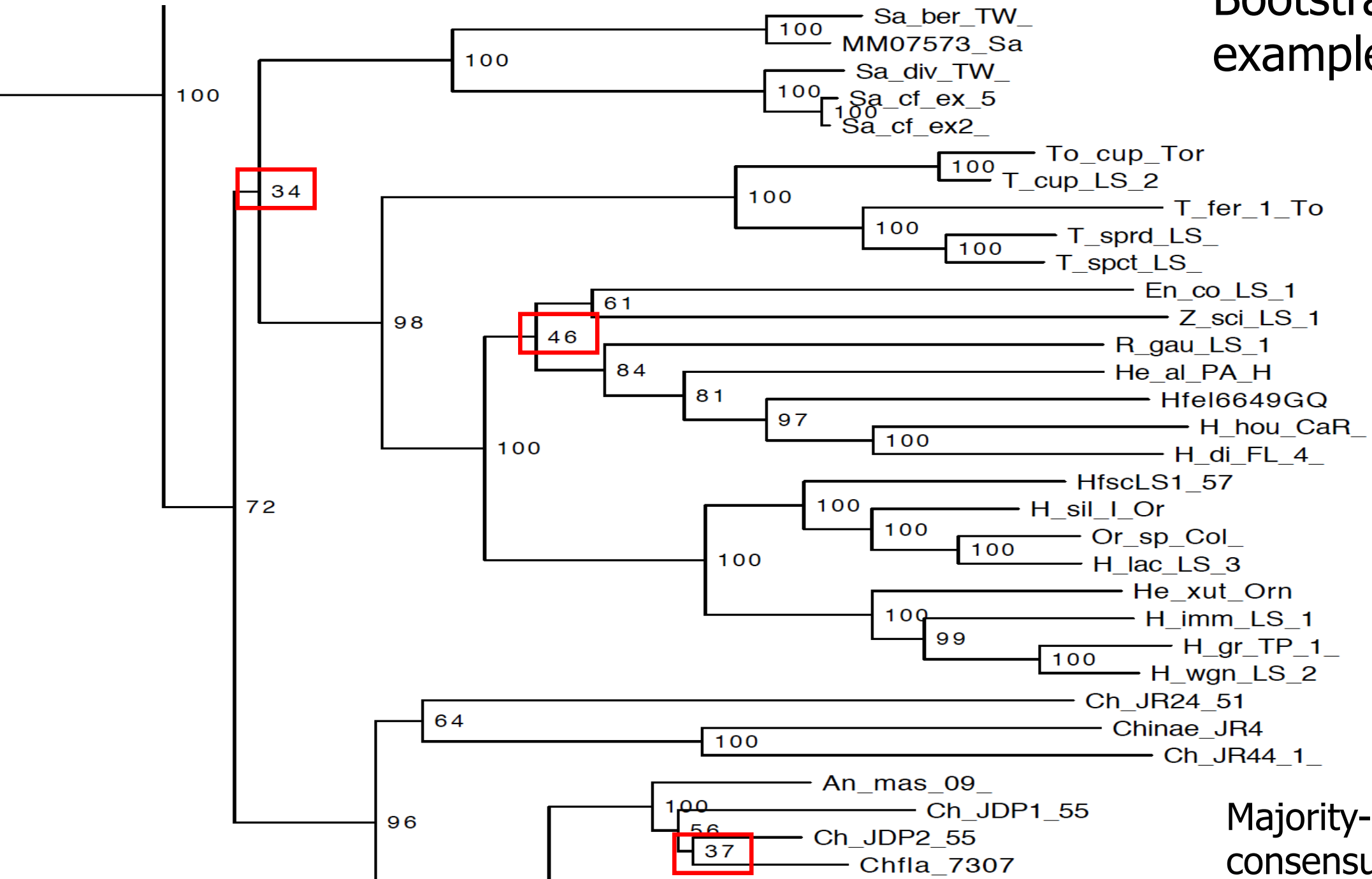




Bootstrap values superimposed on majority-rule consensus tree

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 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)**
  - **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

Concordance factors

New Methods to Calculate Concordance Factors for Phylogenomic Datasets

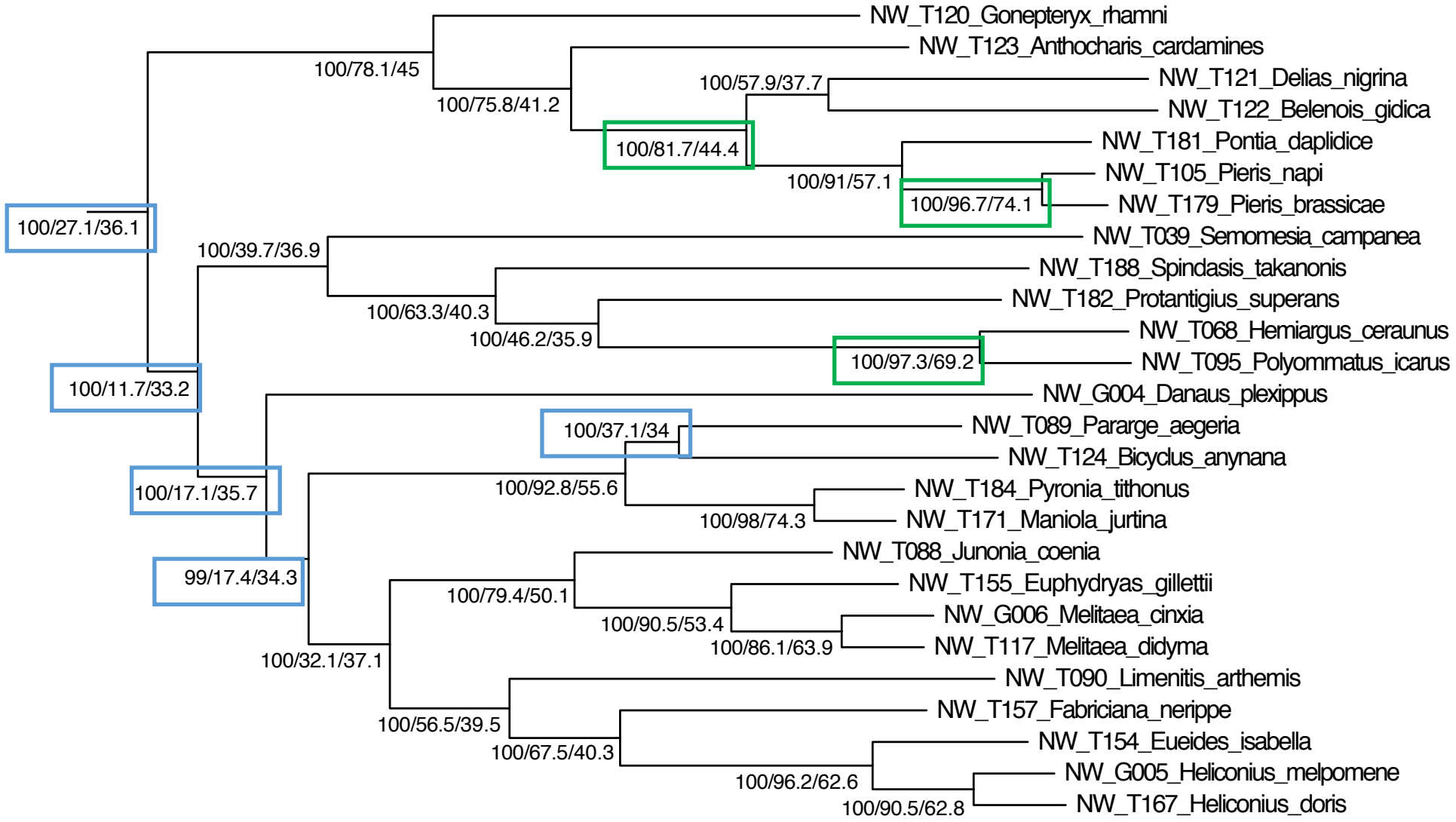
Bui Quang Minh, Matthew W Hahn, Robert Lanfear

Molecular Biology and Evolution, Volume 37, Issue 9, September 2020, Pages 2727–2733, https://doi.org/10.1093/molbev/msaa106

Published: 04 May 2020

- **gCF – gene concordance factor**
  - **Percentage of decisive gene trees containing that branch**
  - **Can be close to 0% if very few genes support a particular topology**
- **sCF – site concordance factor**
  - **Percentage of decisive alignment sites supporting a branch in the reference tree**
  - **If alignment without phylogenetic signal, sCF ≈ 33%**
- **Not necessarily high when bootstraps are high**

Concordance factors example: UFB/gCF/sCF – potential issue: gene tree estimation error



Phylogeny of butterflies (332 genes, IQ-Tree)

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**

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**

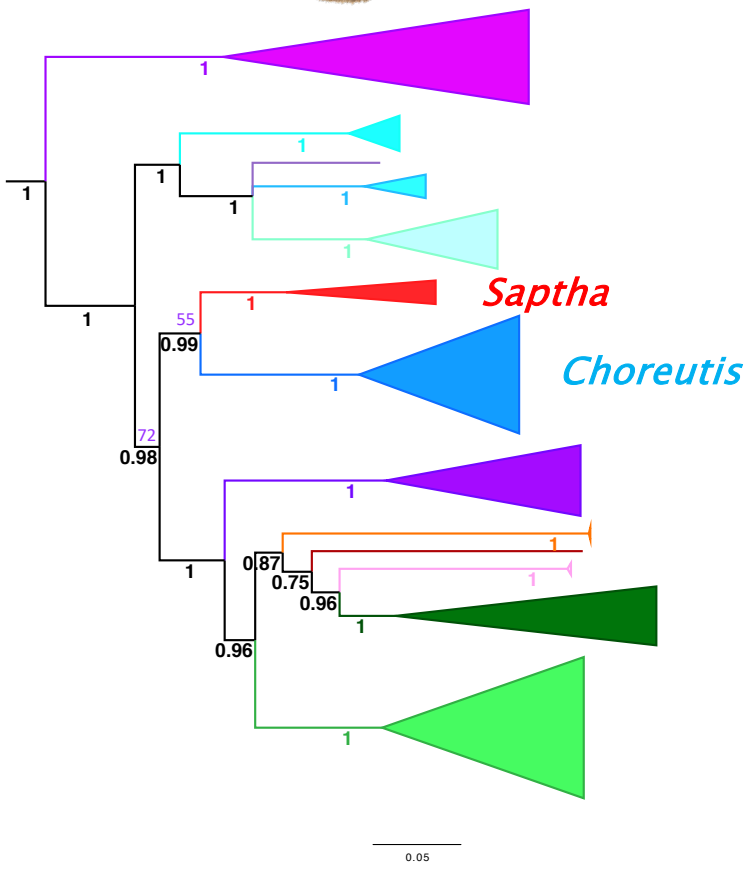
What is significant support?

- **BPs: weak support 50–70%; 70–85% medium; >85% strong**
- **BPs in phylogenomic 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**

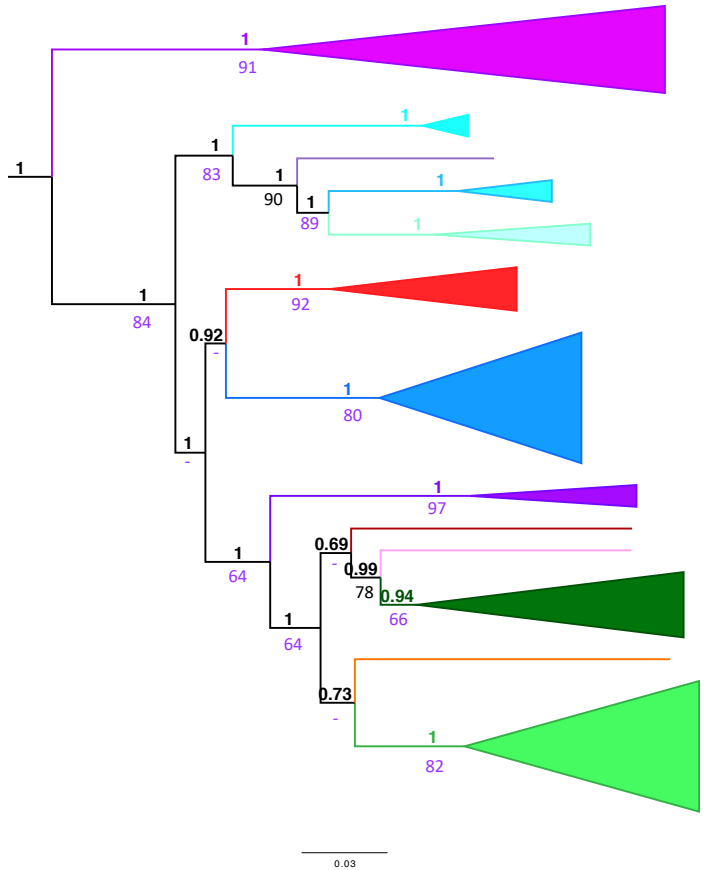
Stability of the hypothesis

- **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?**

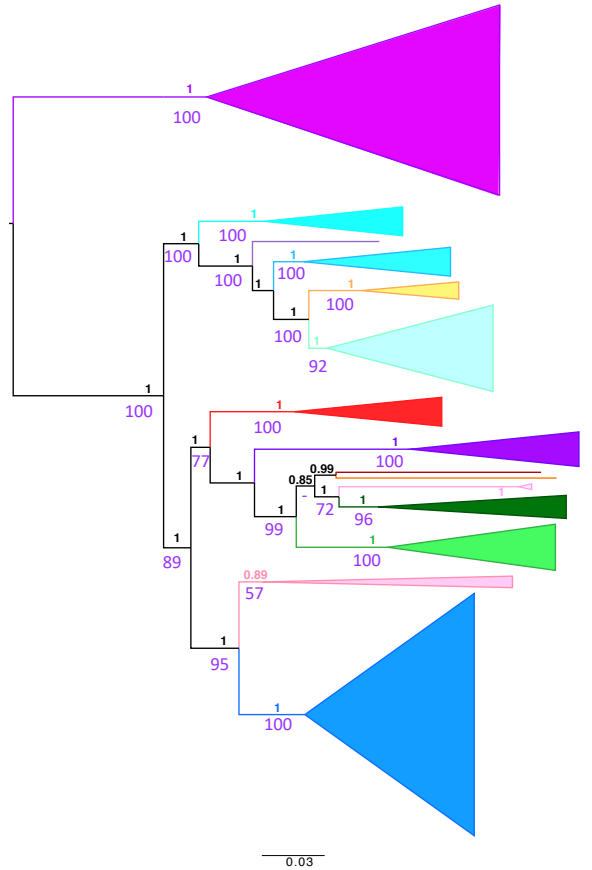
Comparison: BPs and PPs



3 genes, 42 taxa
(Rota 2011)



8 genes, 38 taxa
(Rota & Wahlberg 2012)

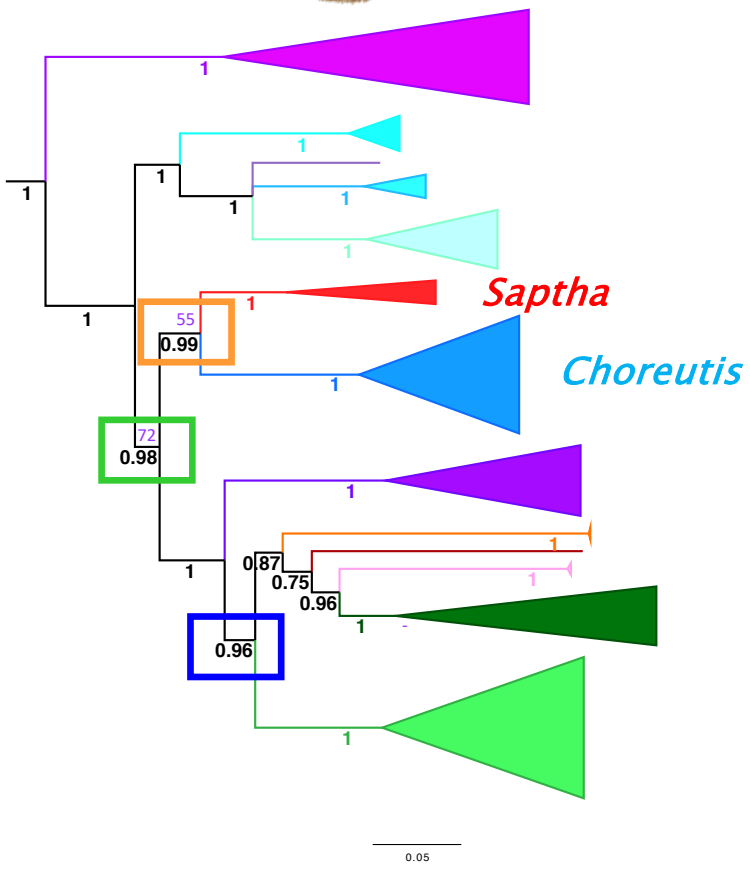


11 genes, 146 taxa
(Rota unpubl.)

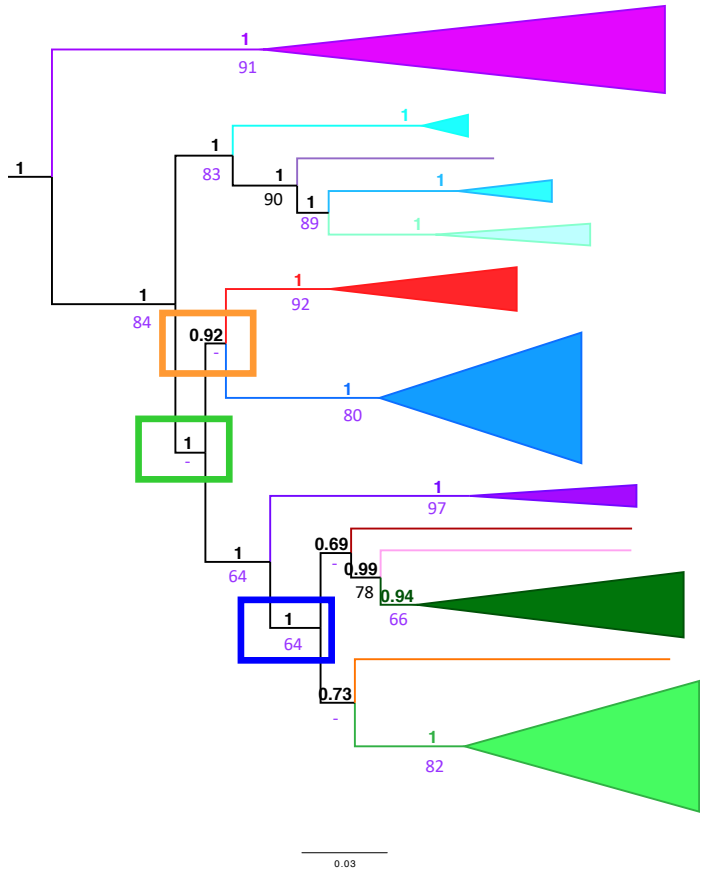
An empirical example: metalmark moths (Lepidoptera, Choreutidae), ca. 600 known species

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

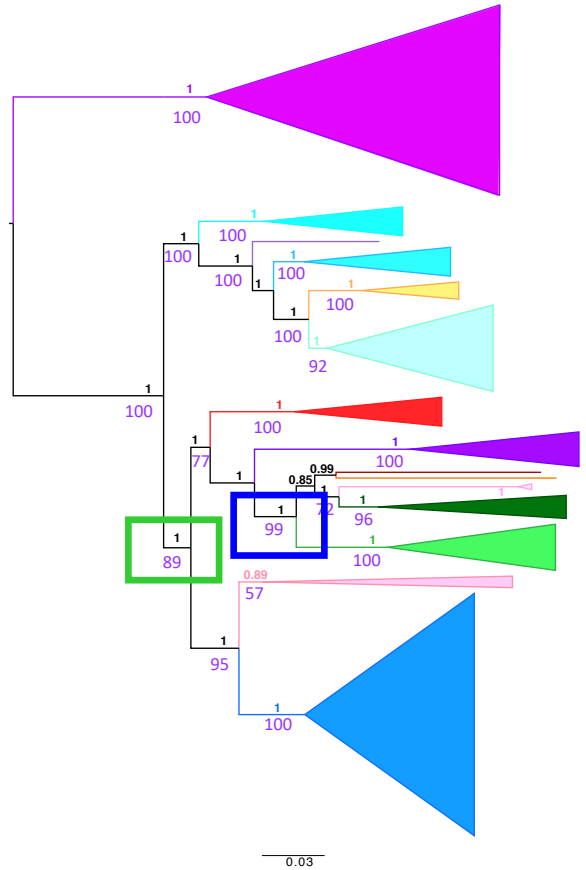
Comparison: BPs and PPs



3 genes, 42 taxa
(Rota 2011)



8 genes, 38 taxa
(Rota & Wahlberg 2012)



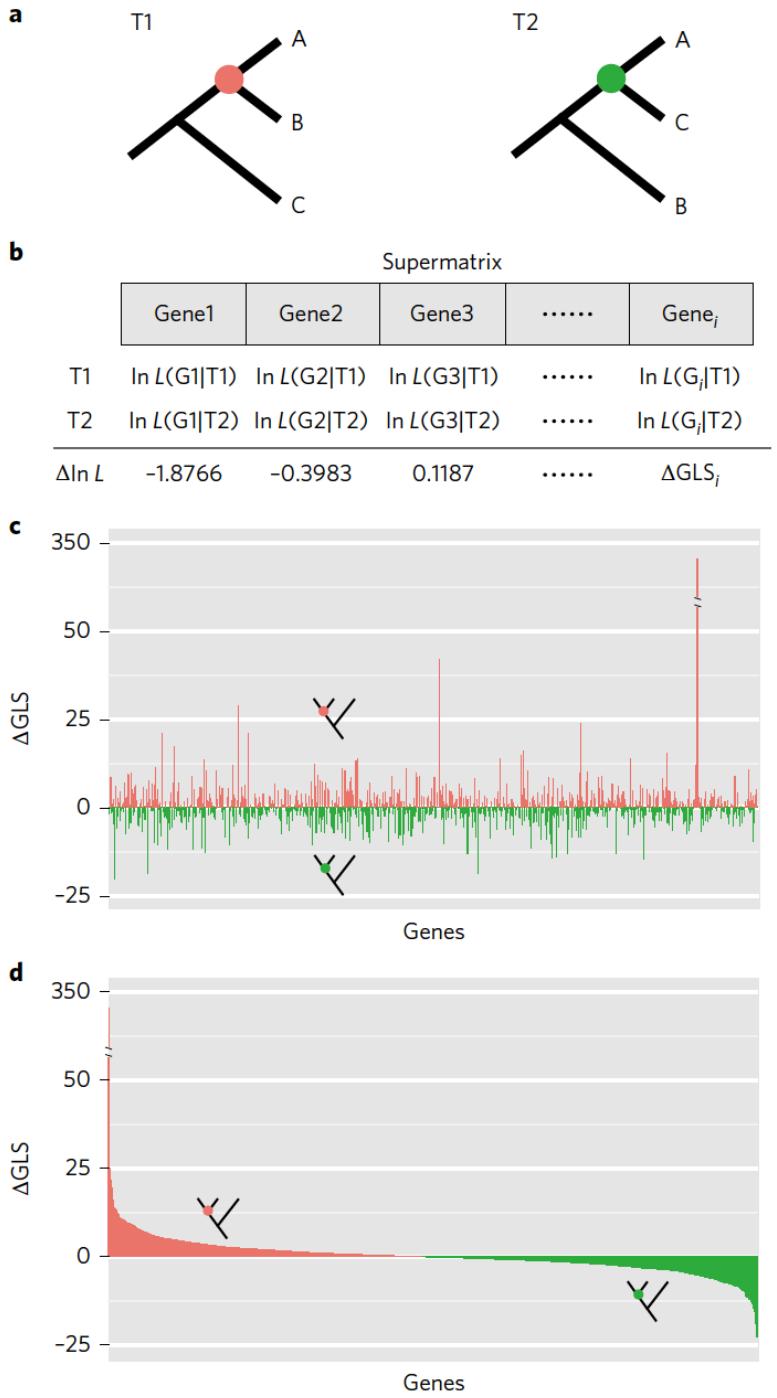
11 genes, 146 taxa
(Rota unpubl.)

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

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

Phylogenetic signal

- **How is phylogenetic signal distributed in your dataset?**
- **Are there a few genes that have an oversized influence?**
- **Shen et al. 2017**



nature ecology & evolution
ARTICLES
PUBLISHED: 10 APRIL 2017 | VOLUME: 1 | ARTICLE NUMBER: 0126

Contentious relationships in phylogenomic studies can be driven by a handful of genes

Xing-Xing Shen[1], Chris Todd Hittinger[2] and Antonis Rokas[1]*

Recommended reading

- **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.**