Bodega Bay Workshop in Applied Phylogenetics

Introductory "State of the Union" Lecture

Mark T. Holder

Department of Ecology and Evolutionary Biology, University of Kansas

March 8, 2014

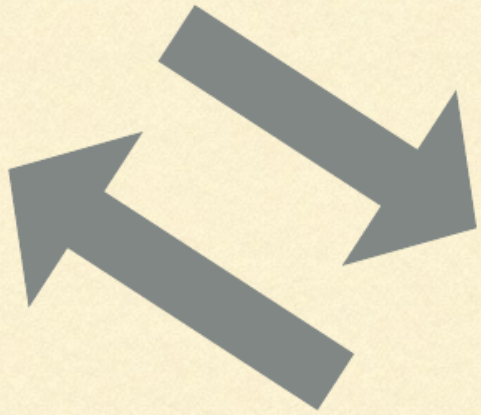
about me

I'm mainly interested in statistical and computational aspects of phylogenetic inference:

- models of sequence evolution.
- multiple sequence alignment.
- improving integration of fossil information into phylogenies.
- "phyloinformatics" and consolidating phylogenetic information across the tree of life

Inputs:
Published phylogenies
Taxonomies

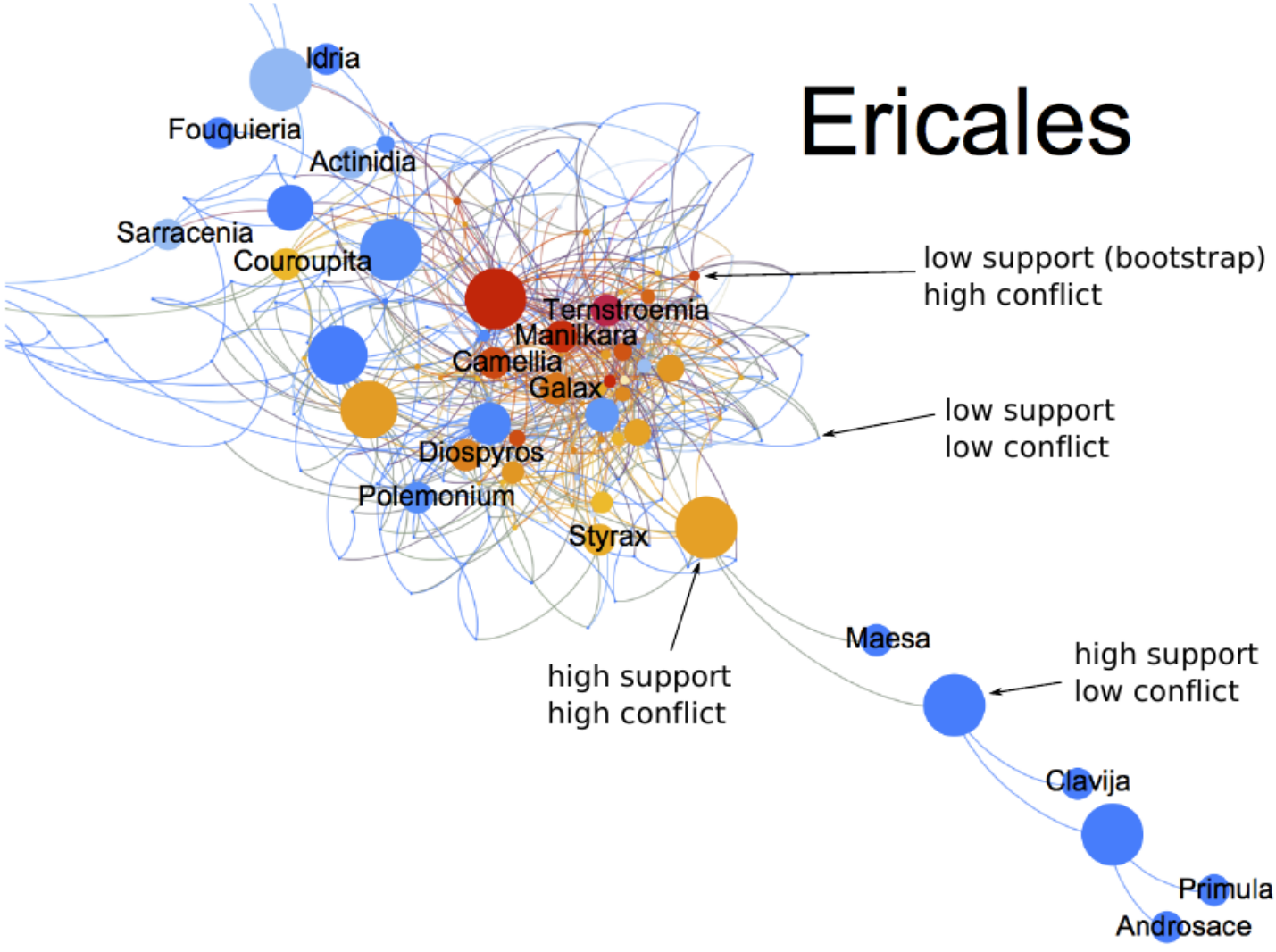
- *filter / weight input trees*
- *synthesize into single data structure*



- *process feedback*
- *input new data sets*

complete tree of life

Karen Cranston: lead PI


Ericales
Idria
Fouquieria
Actinidia
Sarracenia
Couroupita
Ternstroemia
Manilkara
Camellia
Galax
Diospyros
Polemonium
Styrax
Maesa
Clavija
Primula
Androsace
low support (bootstrap)
high conflict
low support
low conflict
high support
high conflict
high support
low conflict

Phylogenetics "State of the union"

Main points:

1. It's Joe's world, we just work here.
2. phylogeny *vs* coalescent *vs* "gene family" tree ...
   - terms matters (somewhat)
   - we probably should not be factoring these distinctions into different boxes
3. Making convincing, rigorous statistical statements in phylogenetics is (still) tough.

then a bit of advice.

During the 1970's and 80's Felsenstein provided a firm foundation:

- pruning algorithm (makes likelihood feasible),
- models of sequence evolution,
- non-parametric bootstrapping for tree inference (1985),
- independent contrasts (1985),
- demographic inference from the coalescent (collab. with J. Yamato, M. Kuhner, and P. Beerli)
- indel model (Thorne, Kishino, Felsenstein 1991, 1992)

Joe's world:

Likelihood-based inference:

- A model with parameter values, θ, predicts what type of data we should see.
- We assess fit of the model to data (X) by the likelihood of the model:

$$\mathbb{P}(X \mid \theta, T)$$

- The original set of phylogenetic models was very simple. Now they are bewilderingly varied and complex.
- Most of us were taught "traditional" statistics that focuses on test statistics (not X), and null distributions (not θ).
- Note that Joe jointly infers $\hat{\theta}$ and T.

"Crust of the earth is a vast museum" - C. Darwin

The history of phylogenetics is a vast museum of abandoned tests/statistics that were not formulated as parameters in a model:

- consistency index,
- retention index,
- permutation tail probability tests
- g_1 statistics for the distribution of tree scores
- ...

The "meta-lesson" I take from the Felsenstein revolution in systematics

Model-based approaches seem cartoonishly simplistic at first, but...

- they are extensible,
- they are transparent,
- using all of the data $\rightarrow$ maximizing statistical power,
- it has taken decades for these methods to be implemented efficiently, but in the long run they seem to win out

Common practice in the 1990's...

- assume that the gene tree = species tree,
- treat all sites as independent and identically distributed,
- infer unrooted trees; root by outgroup if rooted trees are needed
- data collection $\rightarrow$ alignment $\rightarrow$ tree estimation $\rightarrow$ "post-tree analyses"

In the past 15-20 years

1. Massive computational improvements (RAxML, PhyML, FastTree, GARLI, MrBayes, BEAST...)
2. Better treatment of the difference between a gene tree estimate and a species tree
3. Statistical phylogeography
4. Much richer models of sequence evolution along a gene tree
5. Better methods for estimating the timing of events
6. Improved modeling of diversification rates
7. Much improved analysis of character evolution on a tree (stochastic character mapping/ robust counting...)
8. **genomics**

Research foci of the past 15-20 years and this workshop

1. better computational tools	Jeremy, Brian
2. gene tree/species tree	Bob (Mon.)
3. phylogeography	Michael (Tues.)
4. modeling/model selection	John H., Jeremy (Sun).
5. time estimation	Tracy (Mon)
6. diversification	Brian, Mike (Thurs.)
7. character analysis	Rich, Luke, Peter, and Sam
8. next gen sequencing	Jonathan E. (Sun)

Earlier I said that practice in the 1990's was. . .

- assume that the gene tree = species tree,
- treat all sites as independent and identically distributed,
- infer unrooted trees; root by outgroup if rooted trees are needed
- data collection $\rightarrow$ alignment $\rightarrow$ tree estimation $\rightarrow$ "post-tree analyses"

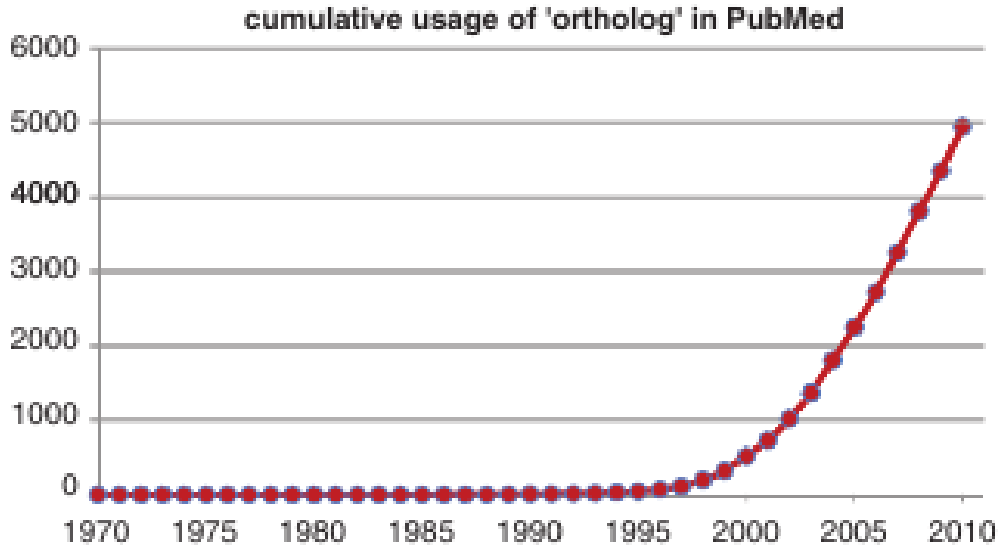


Figure 2: The usage of the term 'ortholog' in the title or abstract of scientific publications. The usage data were from PubMed (http://www.ncbi.nlm.nih .gov/pubmed/).

Figure 2 from Atchley (2011) obituary for Walter Fitch

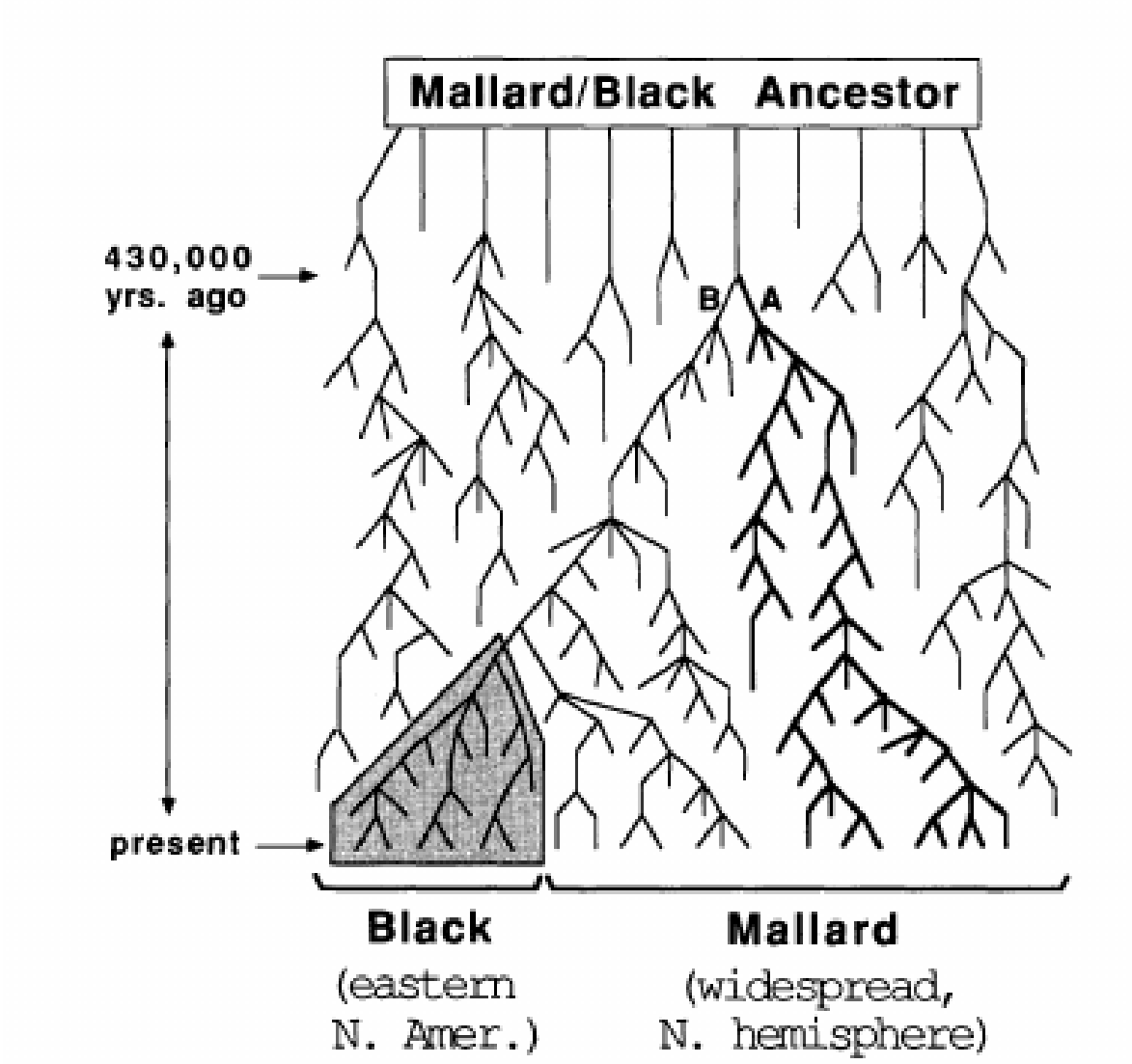


Figure 4 from Avise et al. (1990)

"one is provoked to reconsider precisely what is phylogeny. Perhaps it is misleading to view some gene trees as agreeing and other gene trees as disagreeing with the species tree; rather, all of the gene trees are part of the species tree, which can be visualized like a fuzzy statistical distribution, a cloud of gene histories. Alternatively, phylogeny might be (and has been) viewed not as a history of what happened, genetically, but as a history of what could have happened, i.e., a history of changes in the probabilities of interbreeding."
from the abstract of Wayne Maddison's classic paper
Maddison (1997) (emphasis added)

Let's vote! Which perspective is most helpful?

1. Phylogeny is the sum of gene trees,
2. Phylogeny is "a history of what could have happened"
3. Phylogeny is the "modal" (most common) set of gene tree relationships

There is no right answer to a"which is more helpful" question.

But I vote for option #2. Instead of:

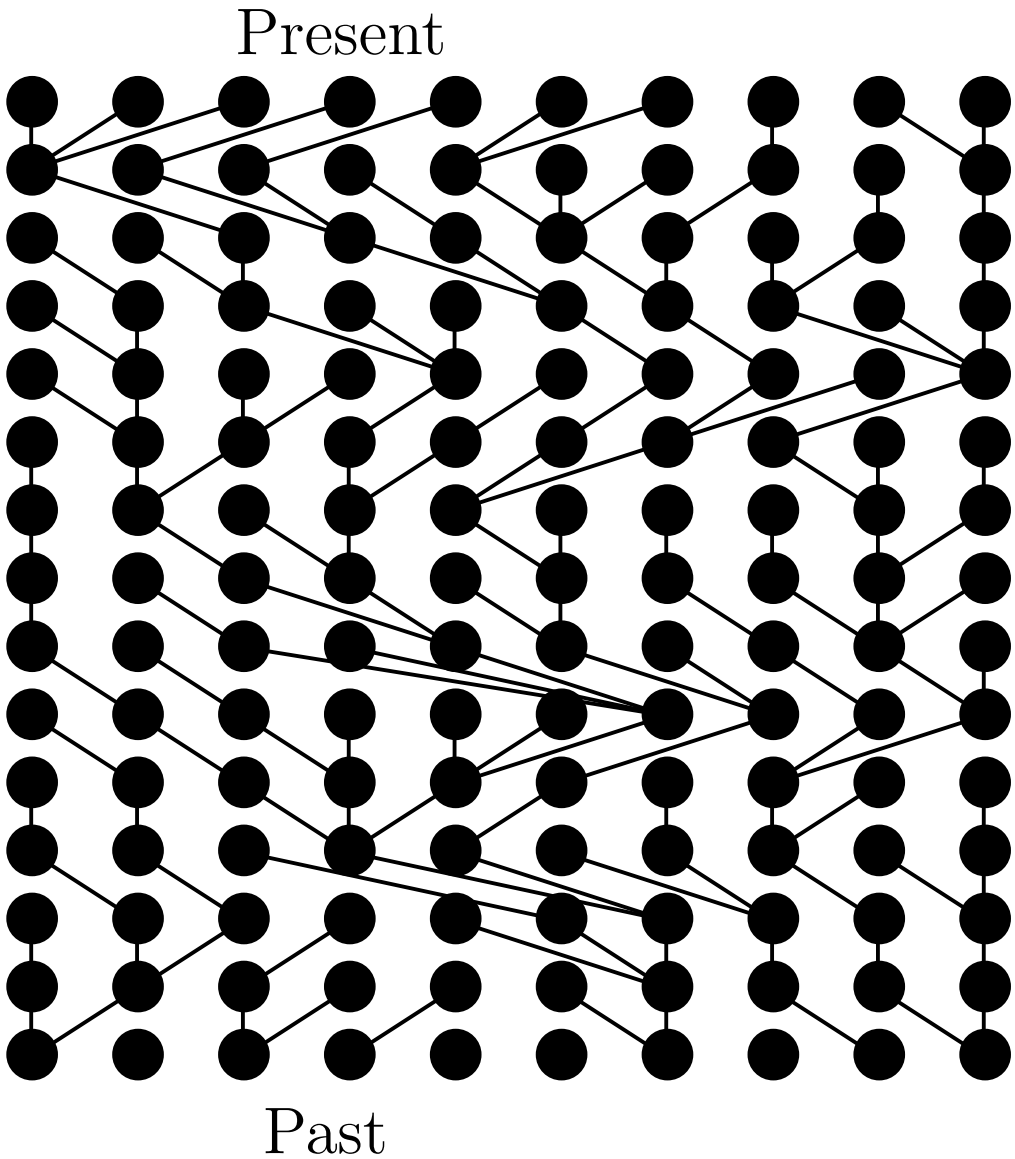
$$\mathbb{P}(X \mid T)$$

where X is the data, and T is the phylogeny,
"a history of changes in the probabilities of inter-breeding" implies:

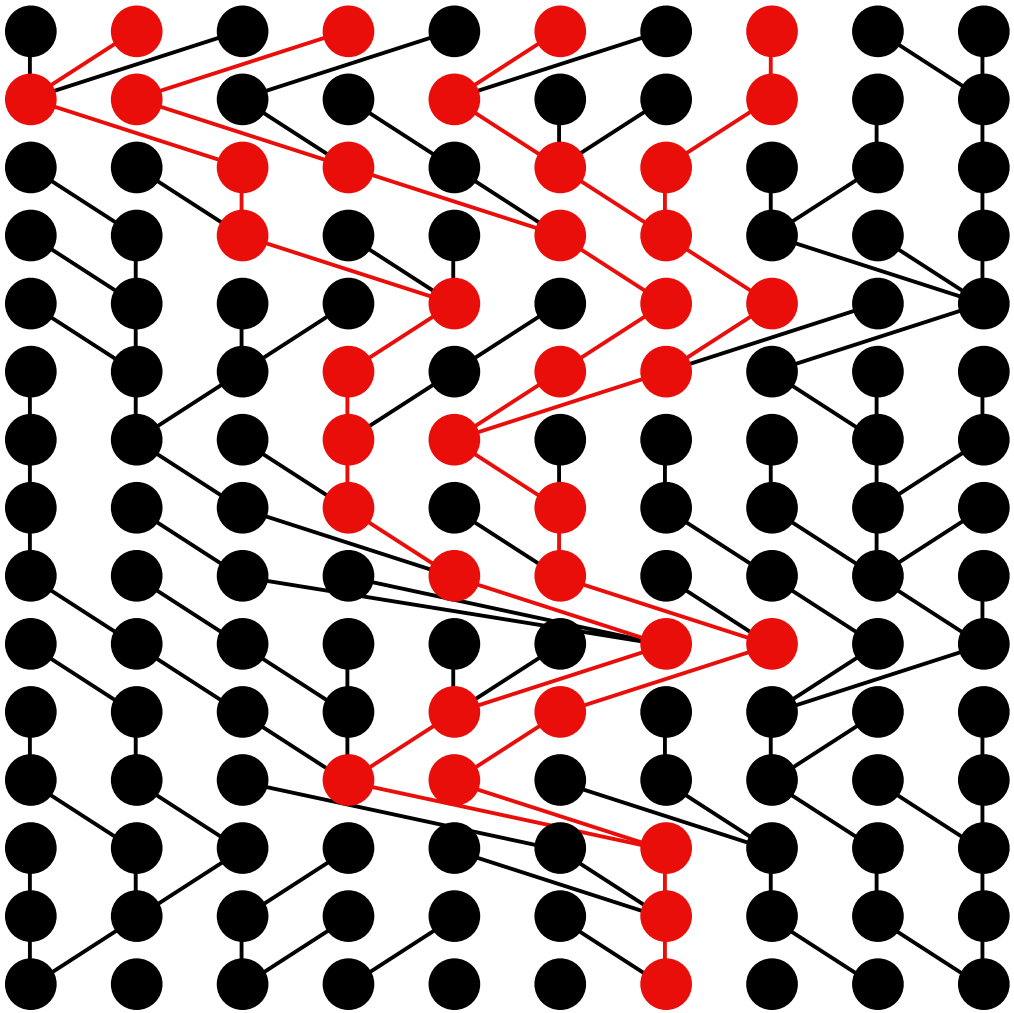
$$\mathbb{P}(X \mid G)\mathbb{P}(G \mid T)$$

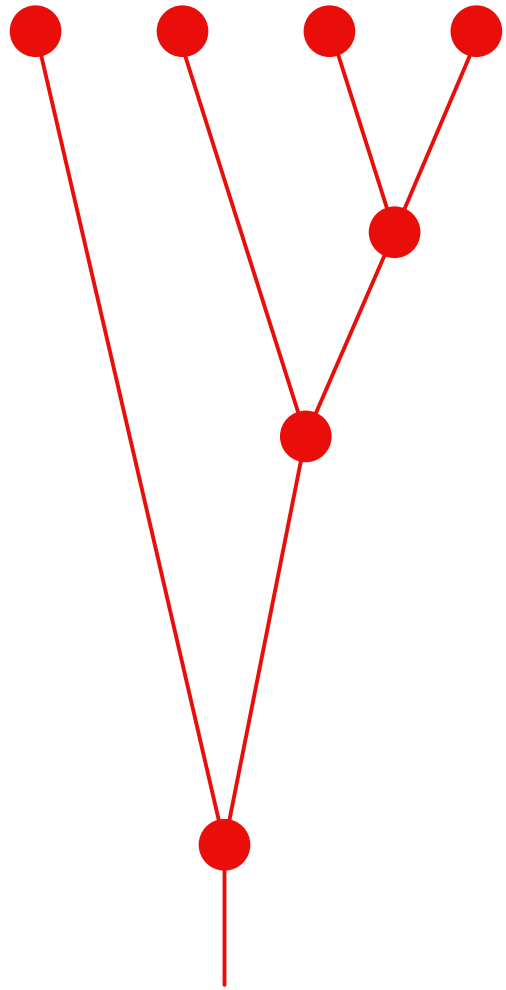
where G is a gene tree.

Genealogies within a population



Genealogies within a population





Biparental inheritance would make the picture messier, but the genealogy would still form a tree (if there is no recombination).

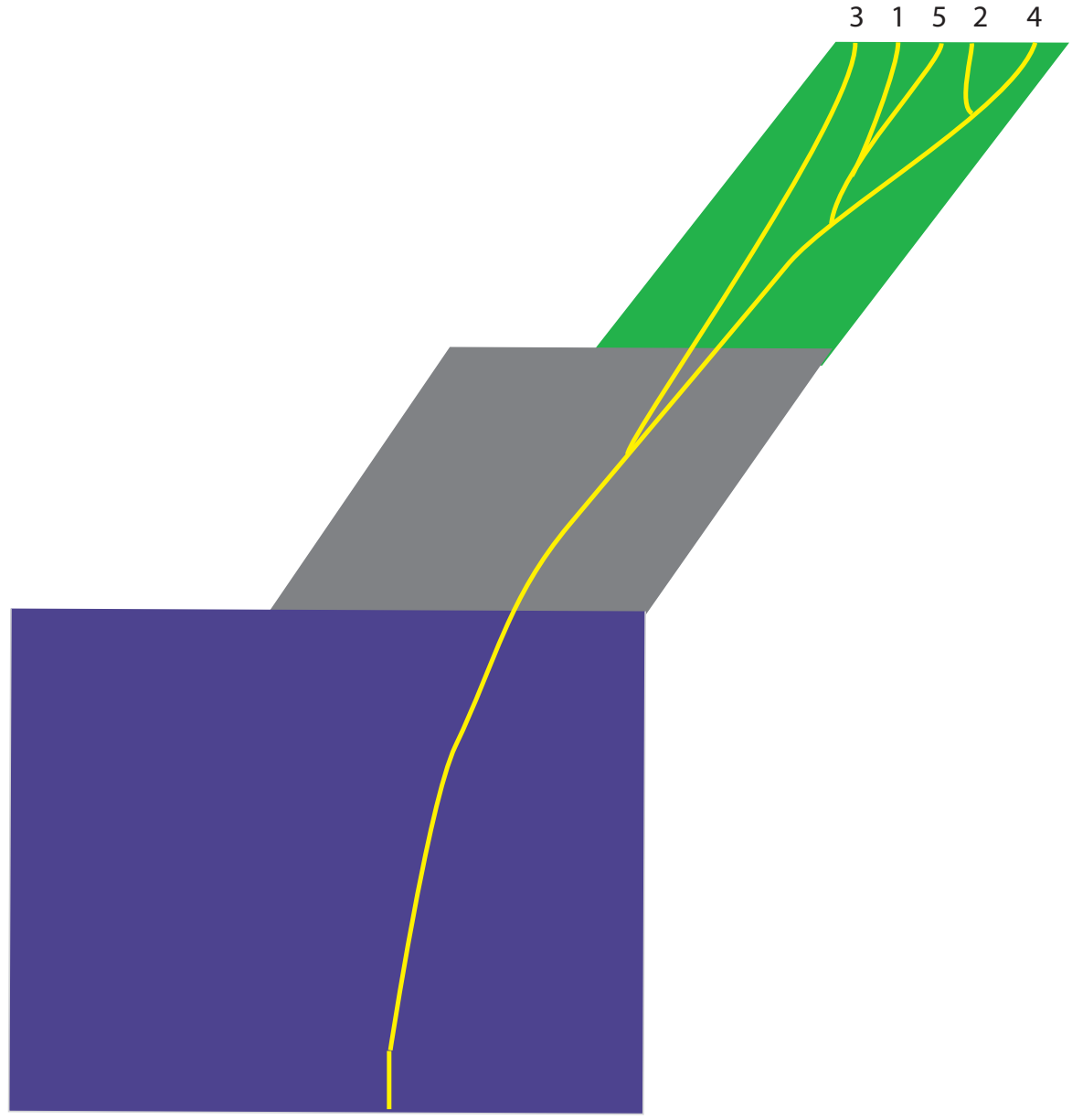
terminology: genealogical trees within population or species trees

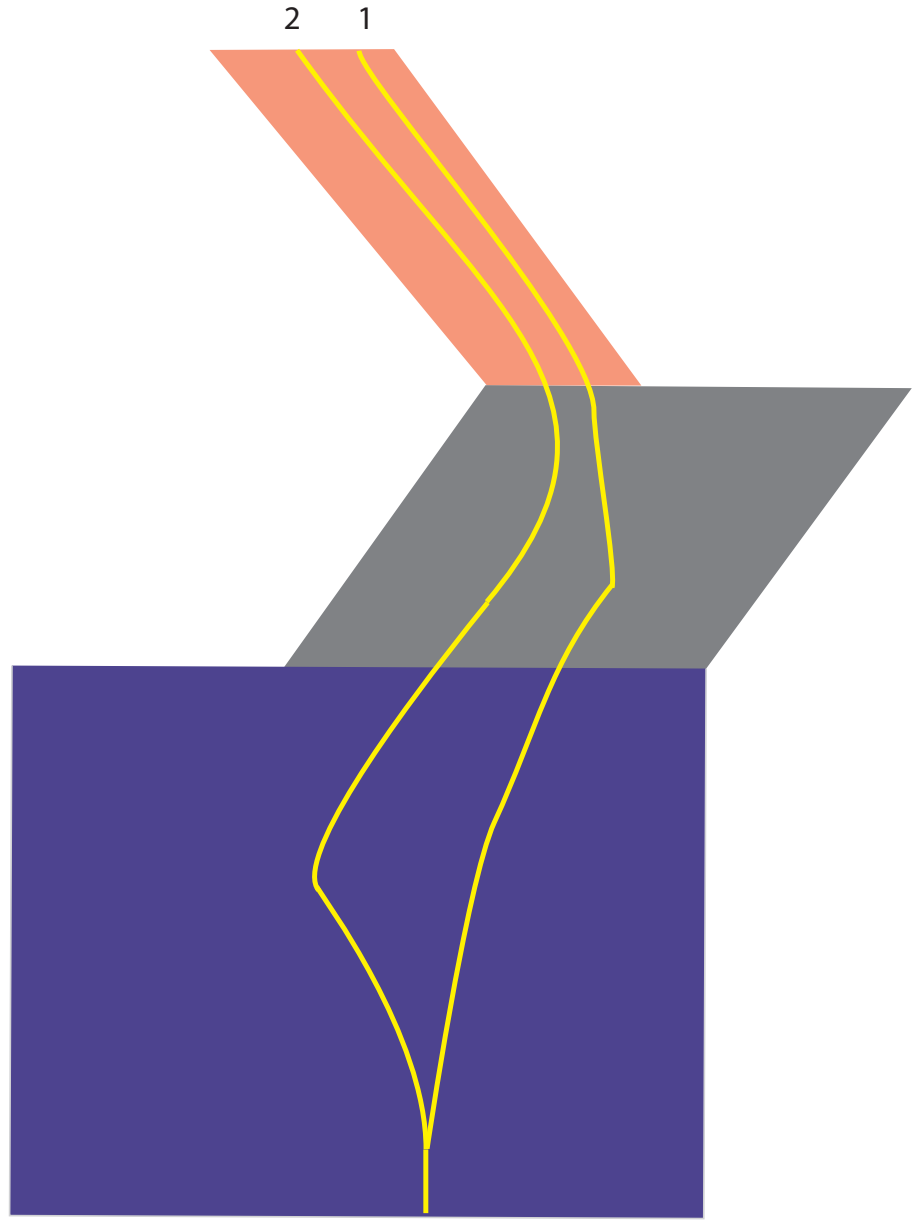
It is tempting to refer to the tips of these gene trees as alleles or haplotypes.

- allele – an alternative form a gene.
- haplotype – a linked set of alleles

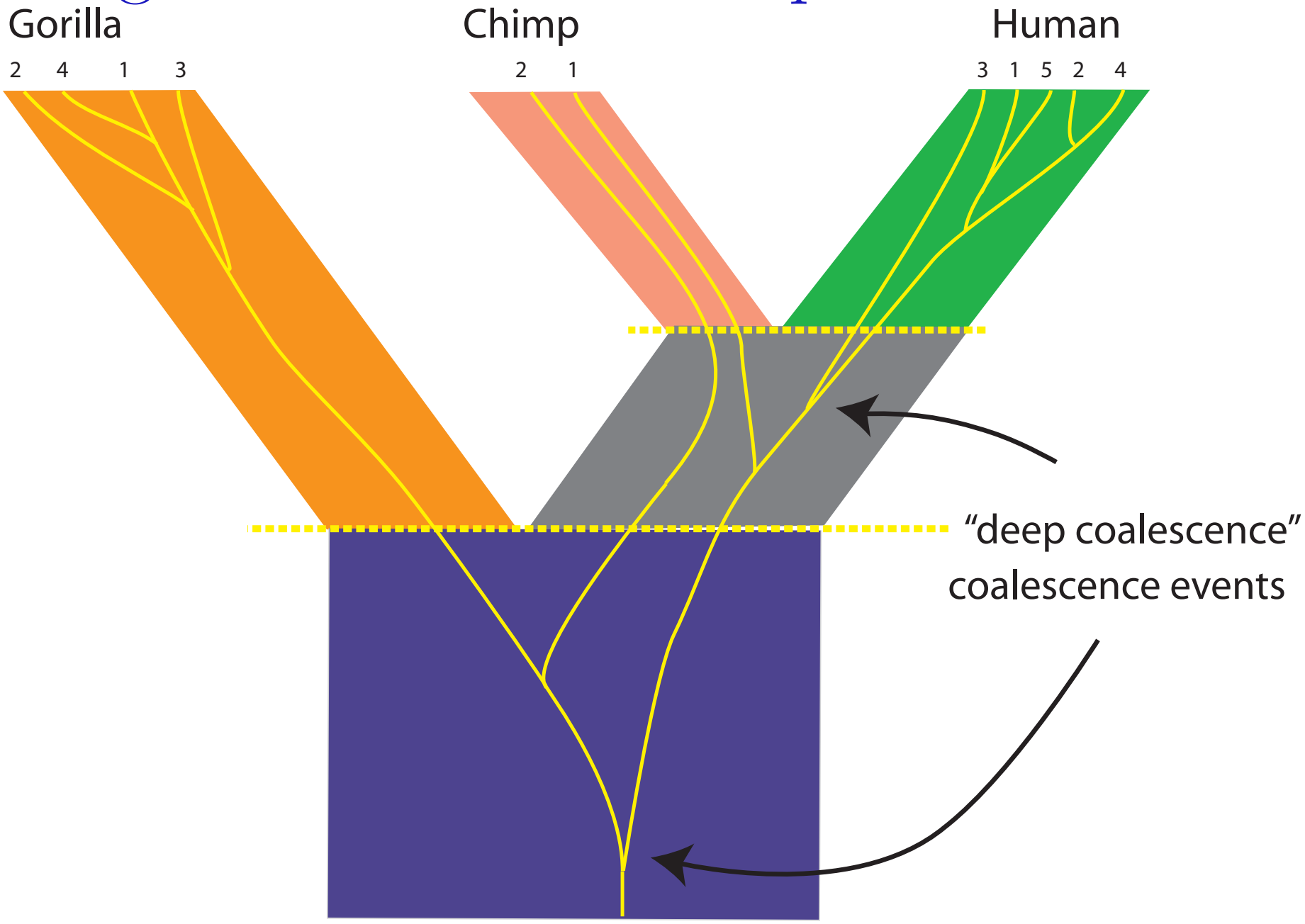
But both of these terms require a differences in sequence.

The gene trees that we draw depict genealogical relationships – regardless of whether or not nucleotide differences distinguish the "gene copies" at the tips of the tree.


3
1
5
2
4


2
1

A "gene tree" within a species tree



terminology: genealogical trees within population or species trees

- coalescence – merging of the genealogy of multiple gene copies into their common ancestor. "Merging" only makes sense when viewed *backwards in time*.
- "deep coalescence" or "incomplete lineage sorting" refer to the *failure* of gene copies to coalesce within the duration of the species – the lineages coalesce in an ancestral species

Coalescents ("gene trees") in species trees

Species tree inference accounting for coalescence

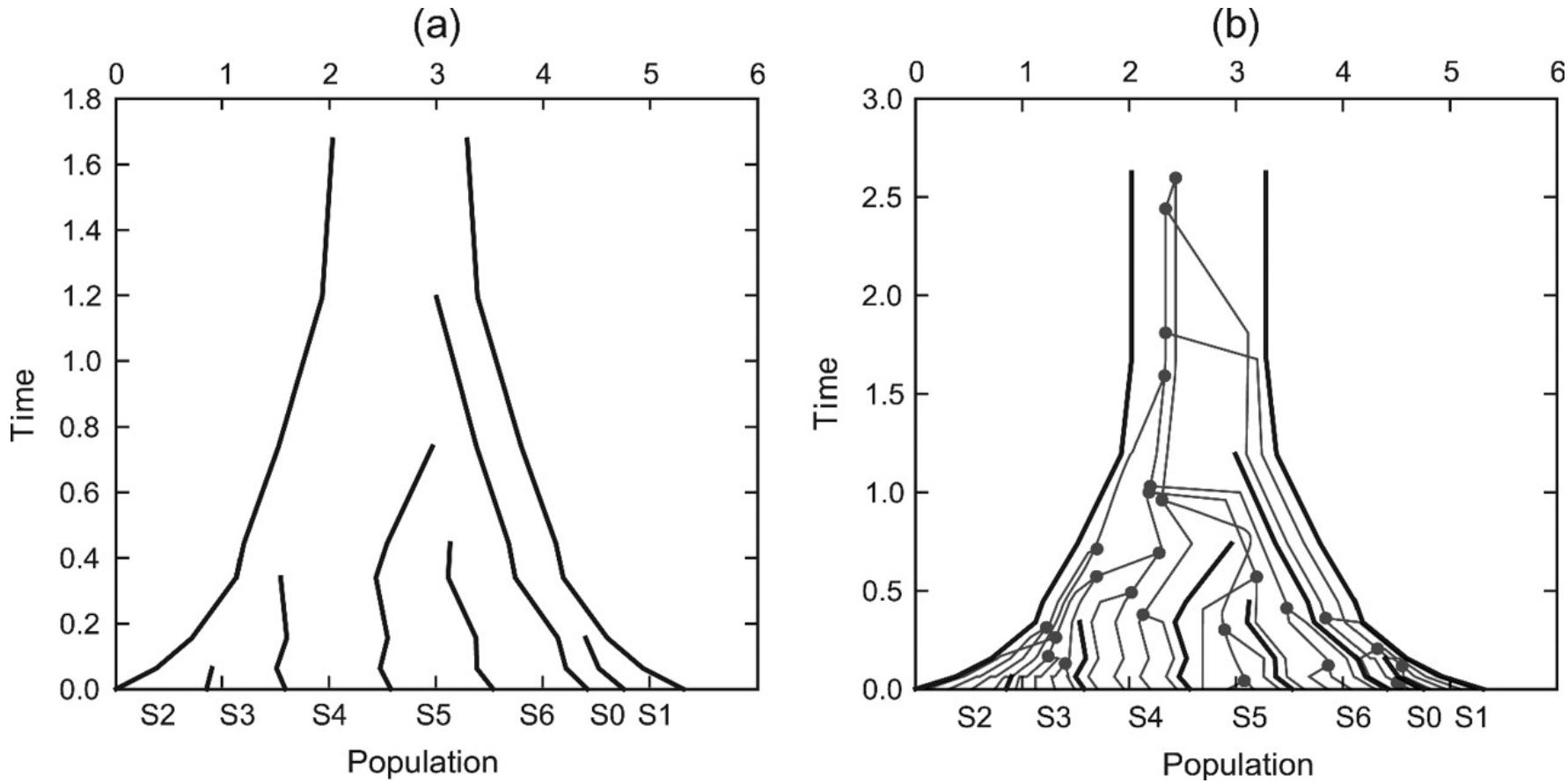


Figure 2 from Heled and Drummond (2010)

Coalescents ("gene trees") in species trees

1. gene tree divergences are older than the species tree divergence;
2. the difference can be big if population sizes are large

But, if you have a tree of 3 species, the most common coalescent topology agrees with the species tree...

Anomalous Gene Trees: most common coalescent topology $\neq$ species tree

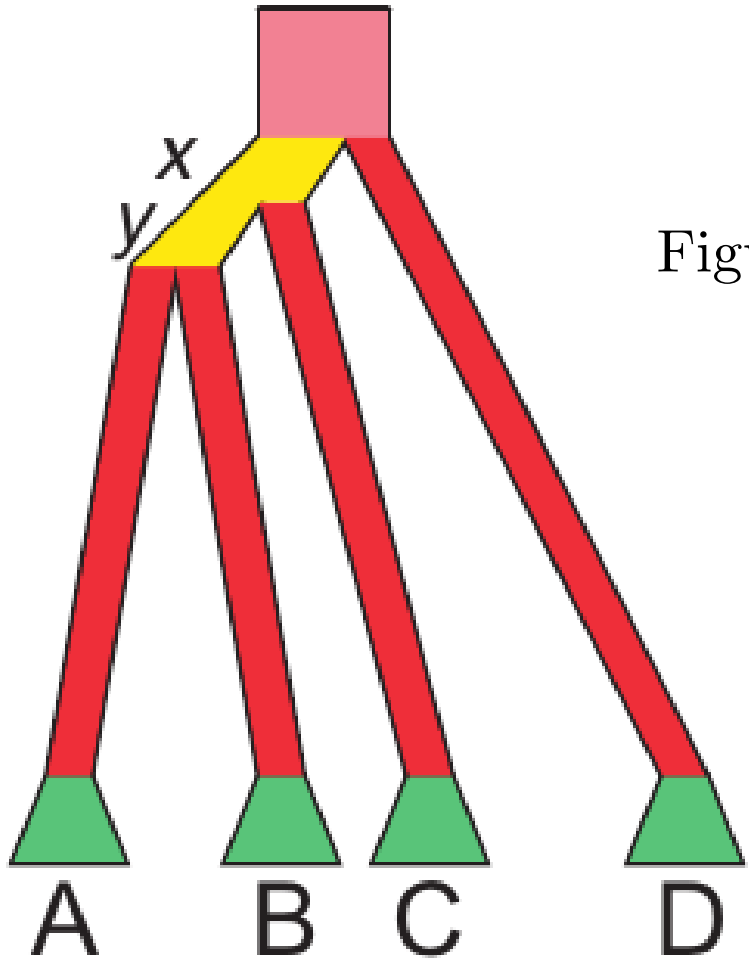


Figure 2 from Rosenberg (2013 MBE)

Anomalous Gene Trees: most common coalescent topology $\neq$ species tree

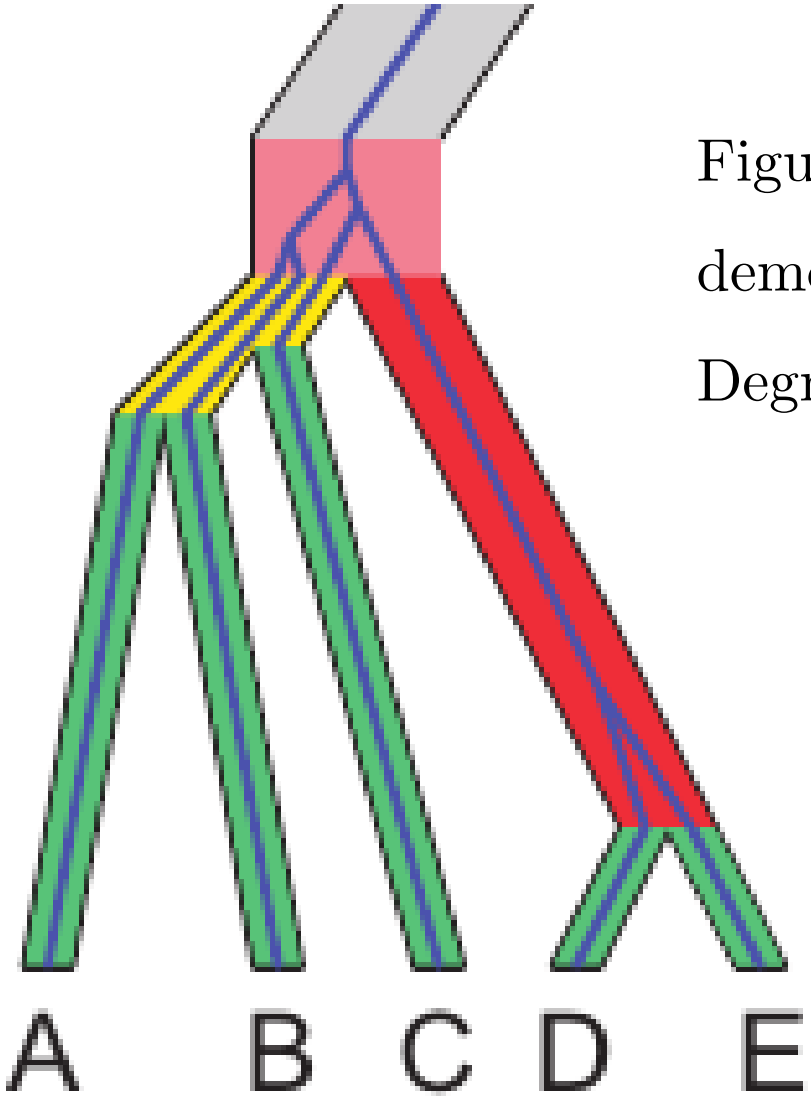
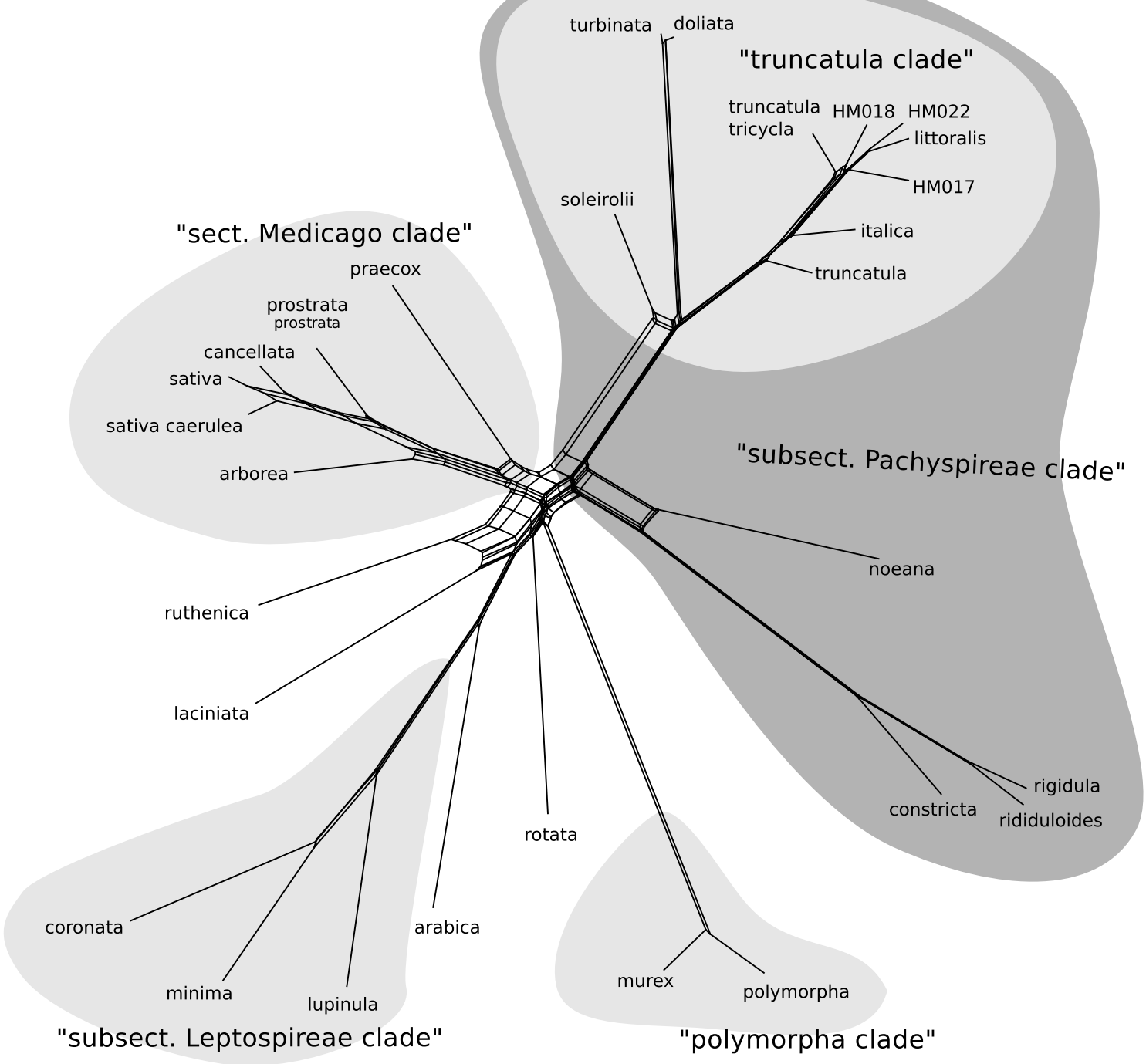


Figure 1 from Rosenberg (2013 MBE) demonstrating result of Degnan+Rosenberg (2006)

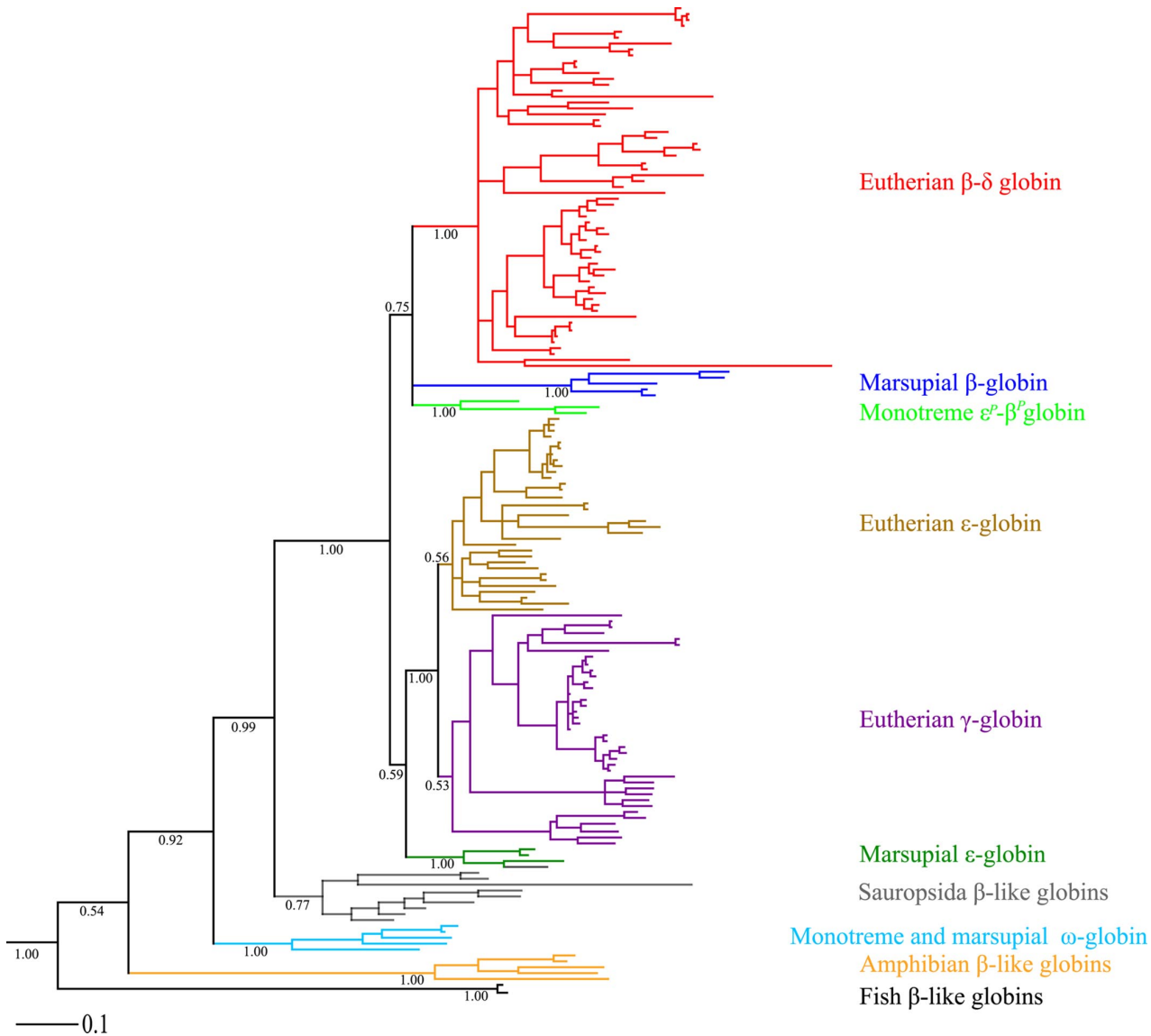


Yoder et al. (2013) genomic analysis of 29 *Medicago* sp.

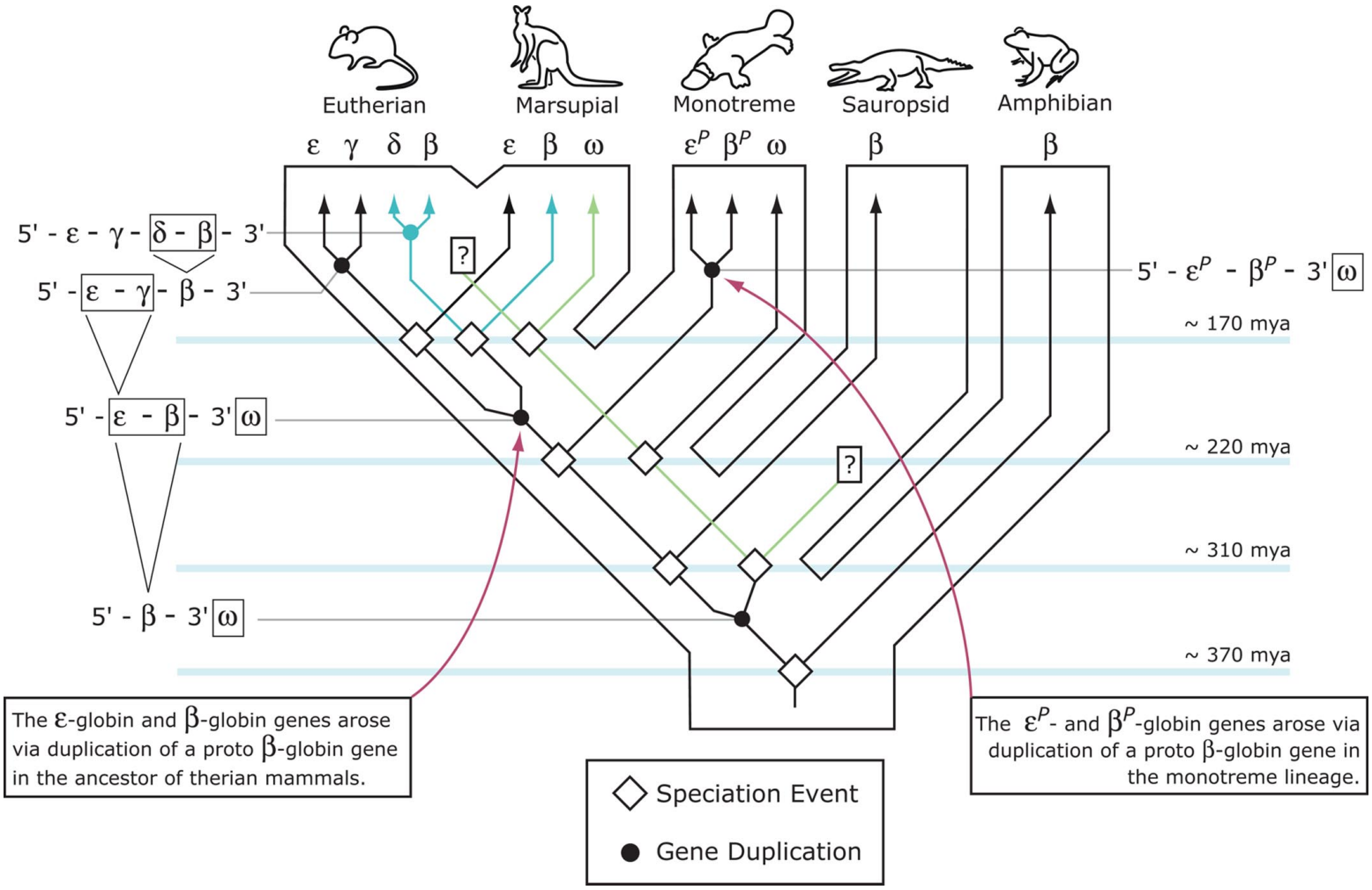
Yoder et al. (2013) summary

- Over 87,000 variable sites (aligned to the reference genome of *M. trunculata*),
- The dataset exceeds to range of the inferring species trees using *BEAST
- lots of deep coalescence, but many previously difficult relationships were resolved,
- randomly sampling 5000 of the variable sites does not yield the full dataset's tree!

A "gene family tree"



Opazo, Hoffmann and Storz(2008)



Opazo, Hoffmann and Storz (2008)

Joint estimation of gene duplication, loss, and species trees using PHYLDOG

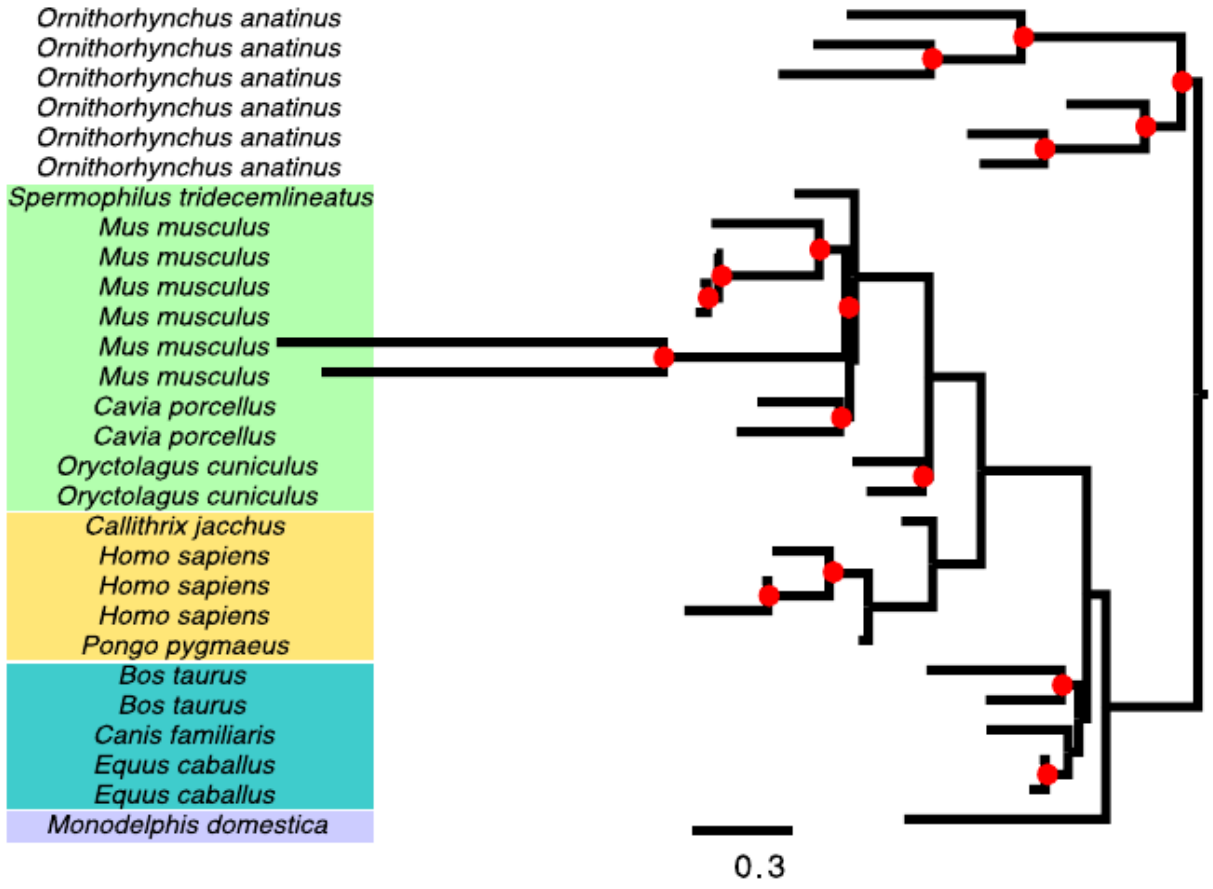
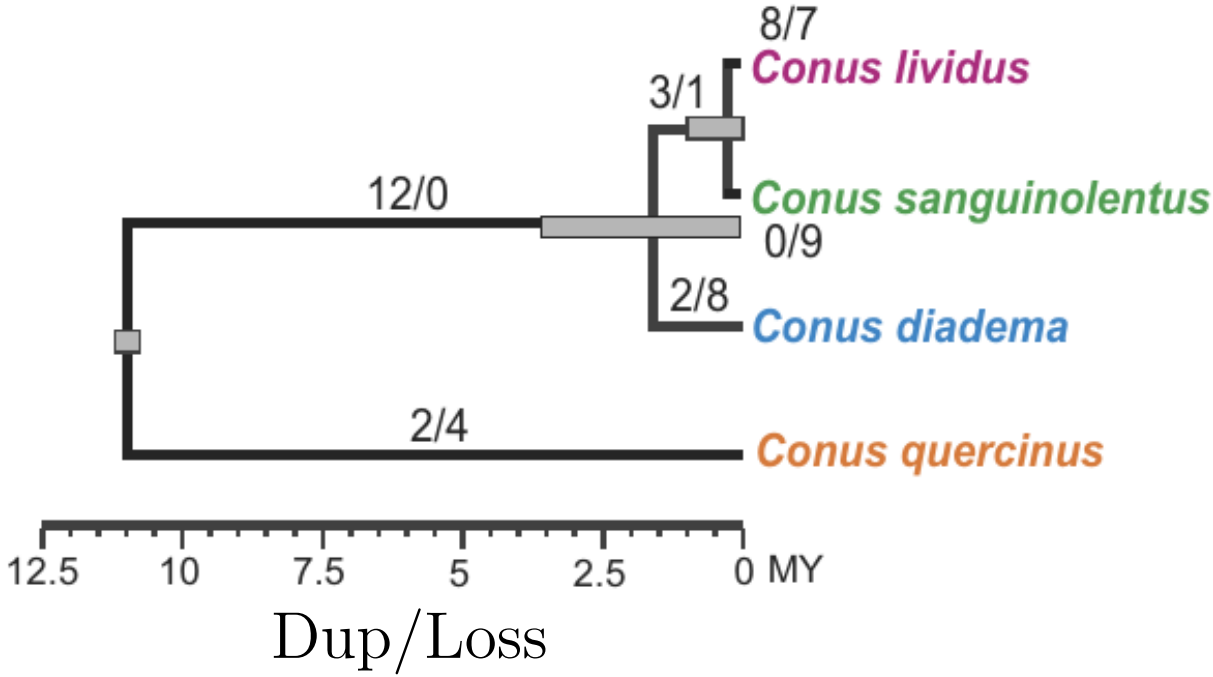


Figure 2A from Boussau et al. (2013)

Very Rapid Turnover of A-superfamily conotoxin genes in *Conus*



Rates of duplications estimated by Notung (Vernot et al., 2008) and PrIME-GSR (Åkerborg et al., 2009)

Figure 1 from Chang and Duda (2012)

Gene Duplication on a fixed species tree

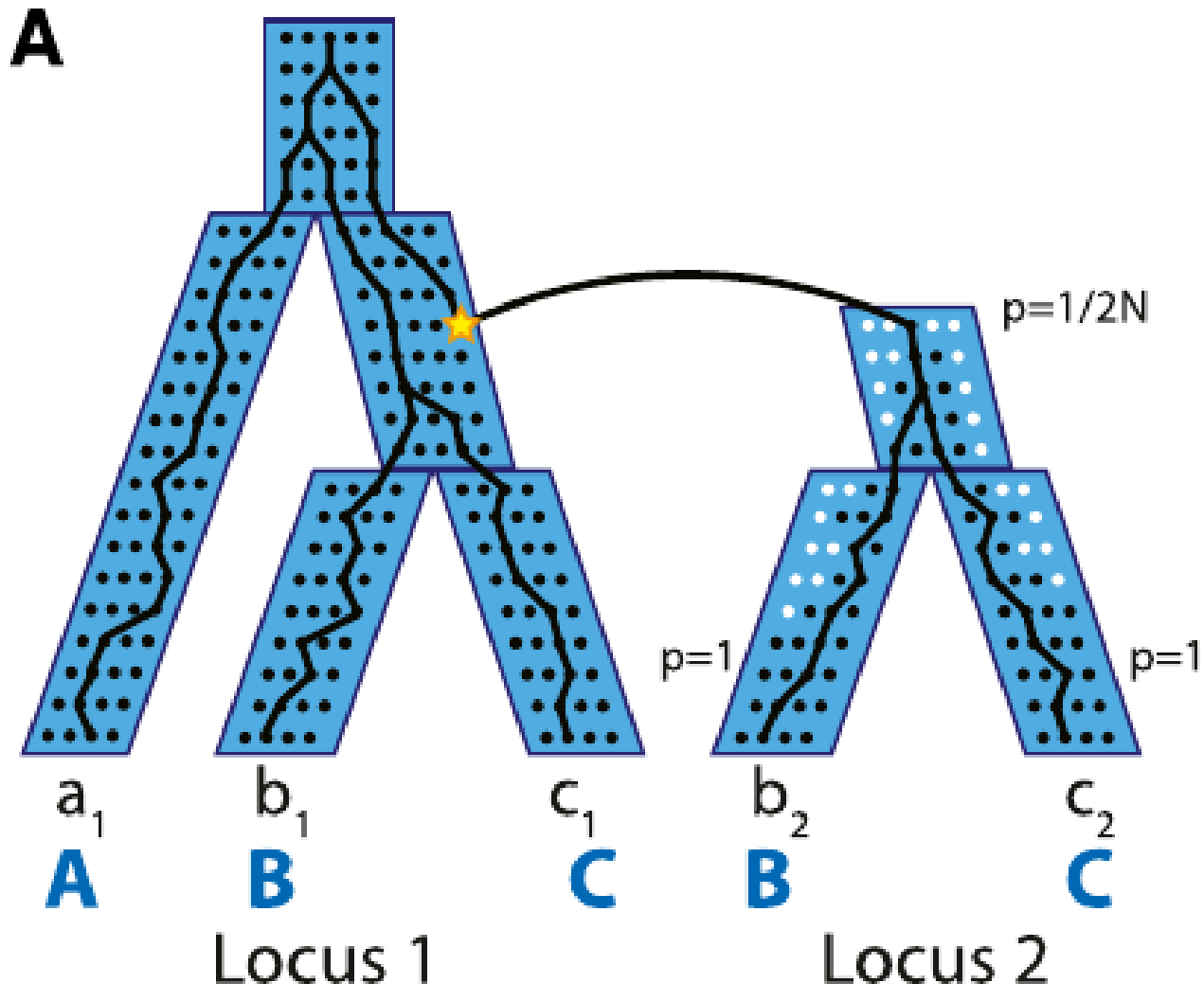


Figure 2A from Rasmussen and Kellis (2012)

Instead of:

$$\mathbb{P}(X \mid T)$$

where X is the data, and T is the phylogeny,
a separation into:

$$\mathbb{P}(X|G)\mathbb{P}(G|L)\mathbb{P}(L|T)$$

where G is a gene tree, L is a "locus tree" (Rasmussen and Kellis, 2012).

Mapping gene/locus/species trees

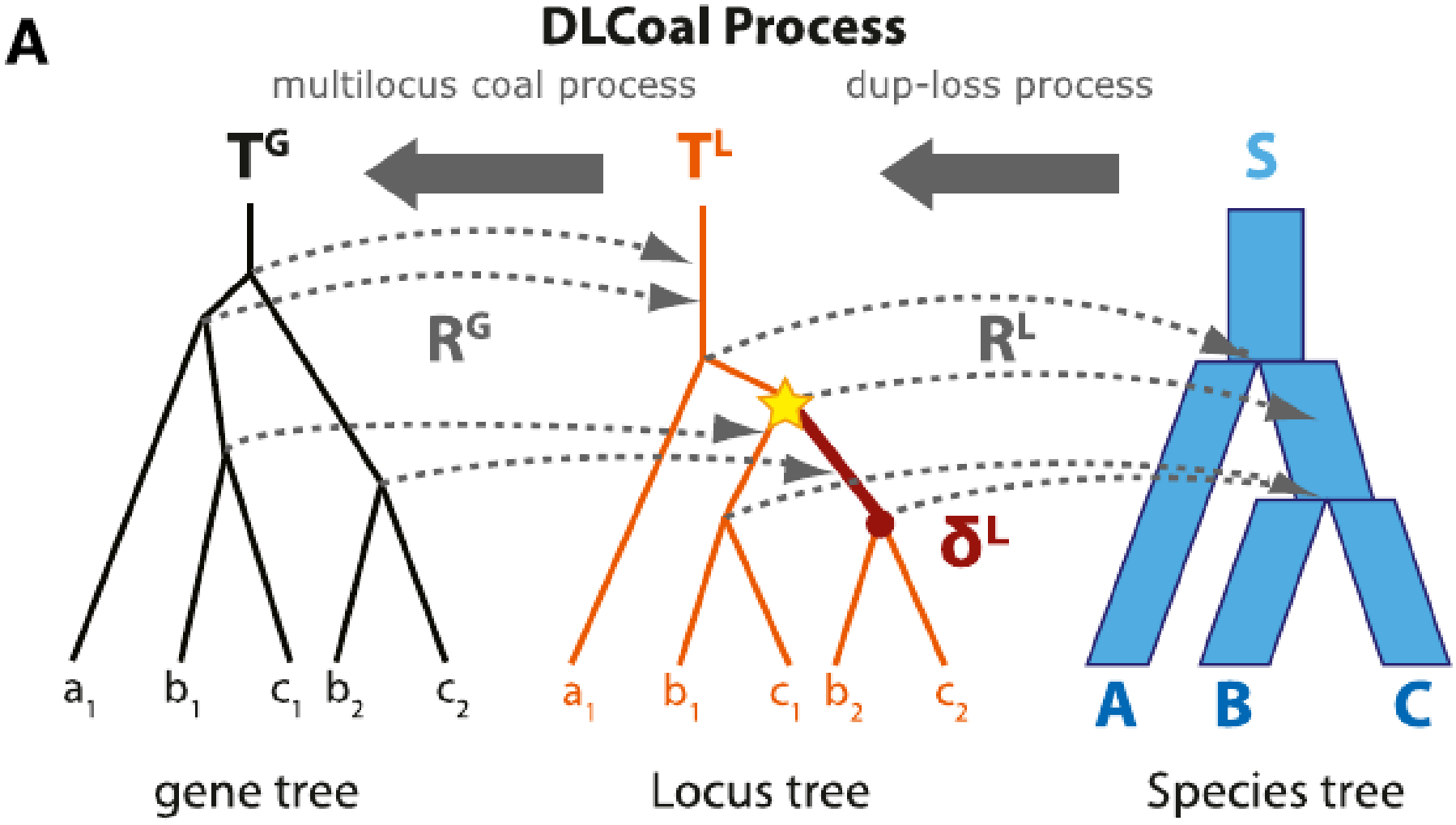


Figure 3A from Rasmussen and Kellis (2012)

Joint estimation of gene duplication, loss, and coalescence with DLCoalRecon

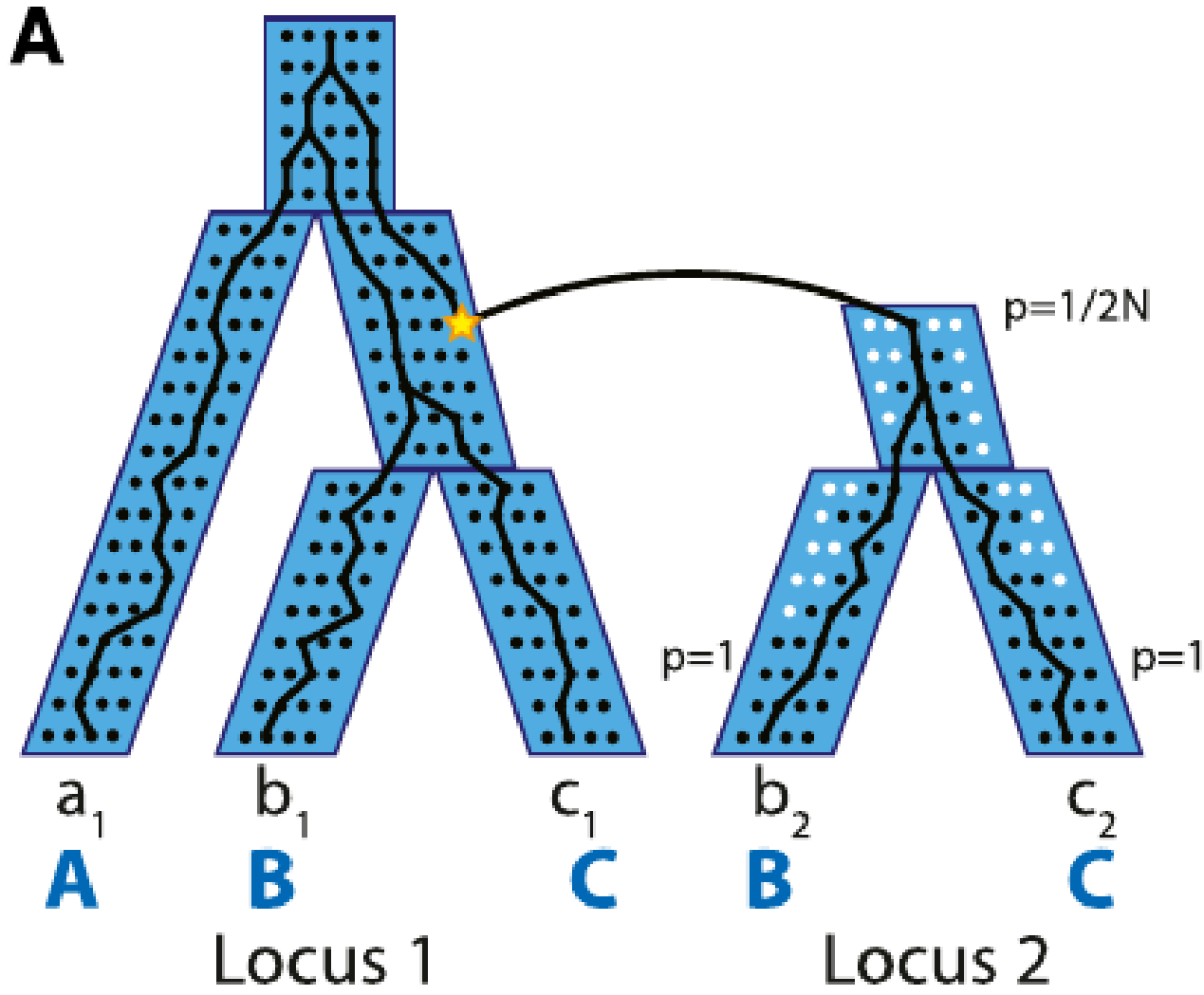


Figure 2A from Rasmussen and Kellis (2012)

Future: improved integration of DL models and coalescence

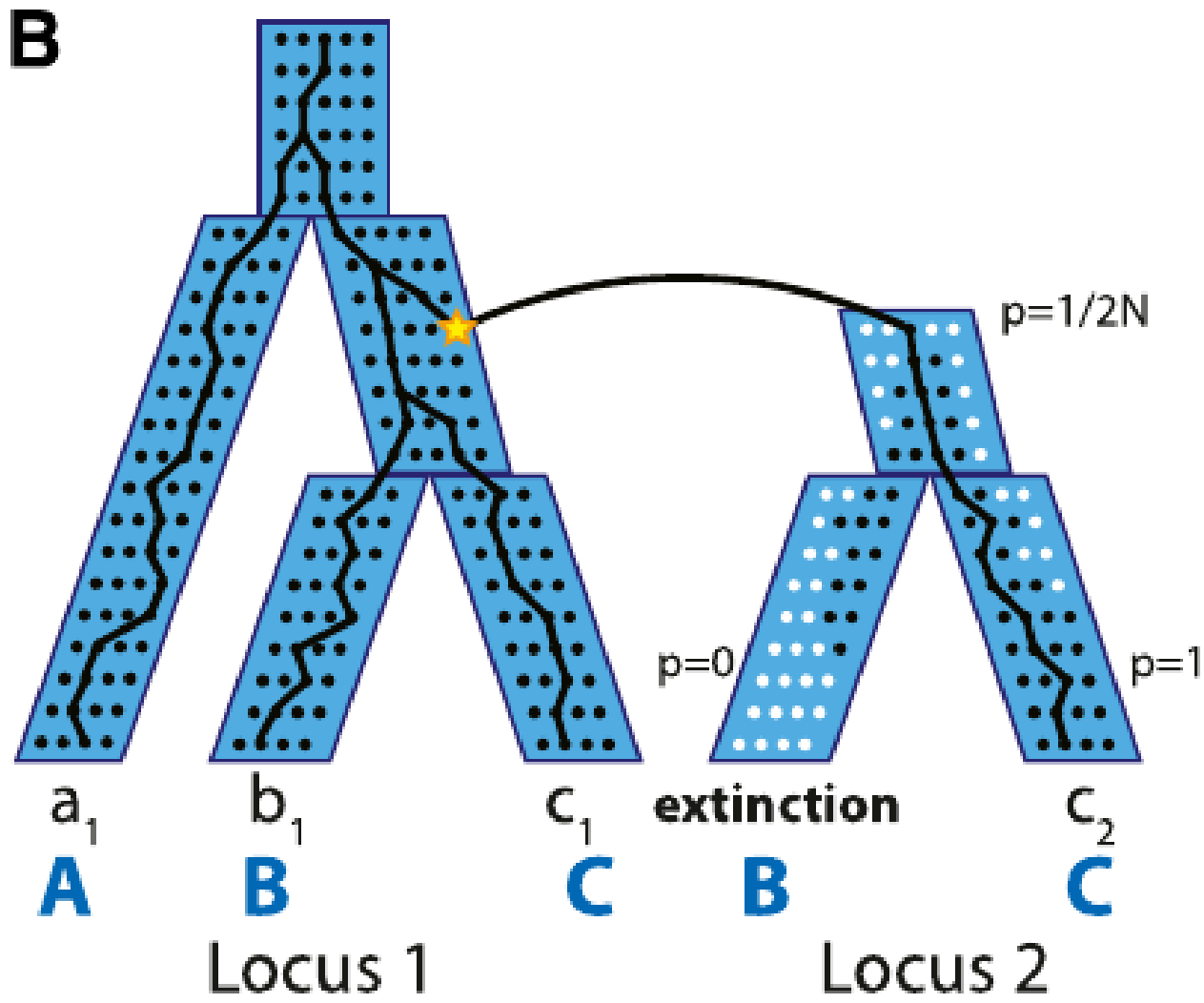


Figure 2B from Rasmussen and Kellis (2012)

Modeling Allopolyploidization

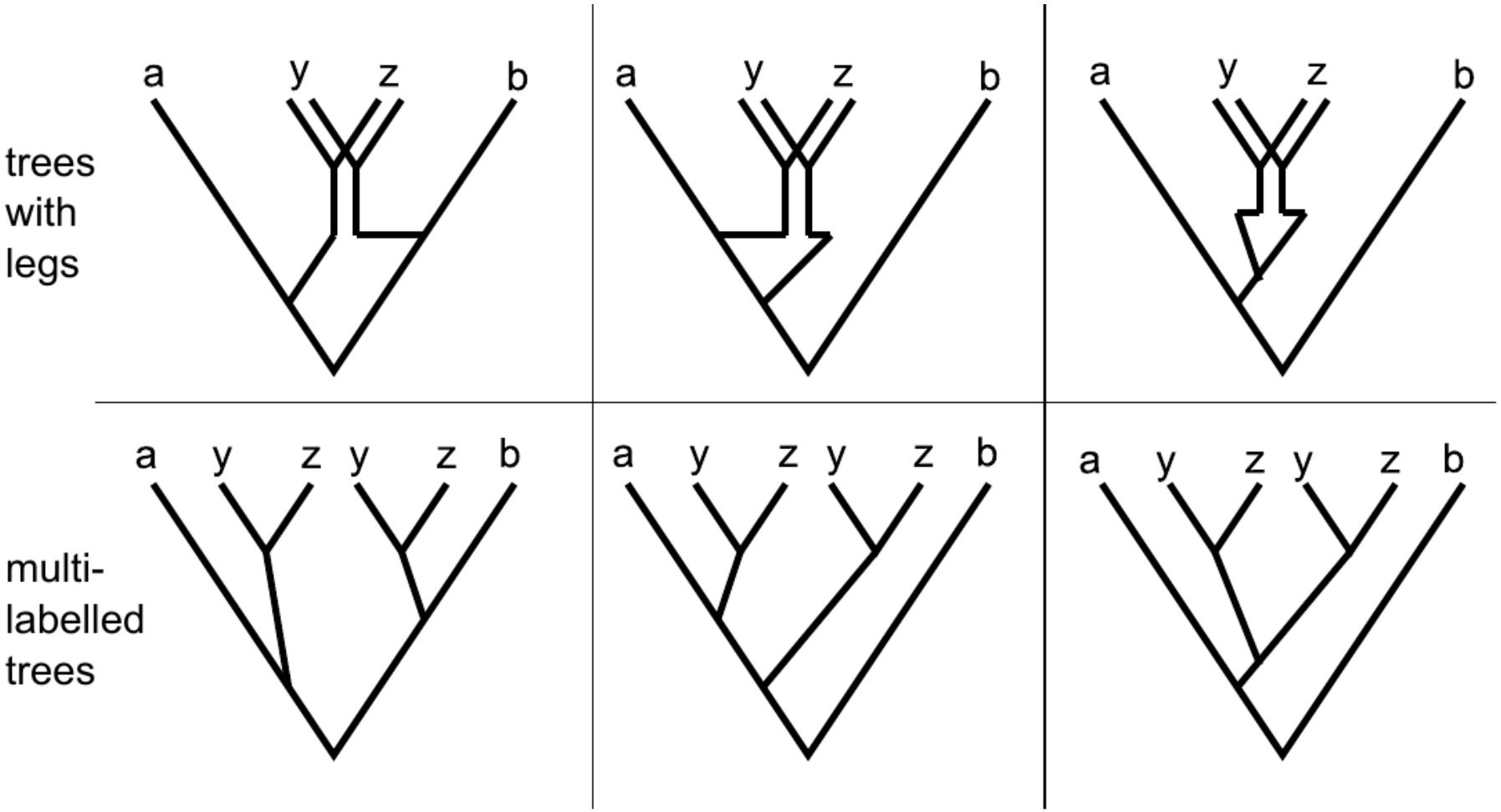
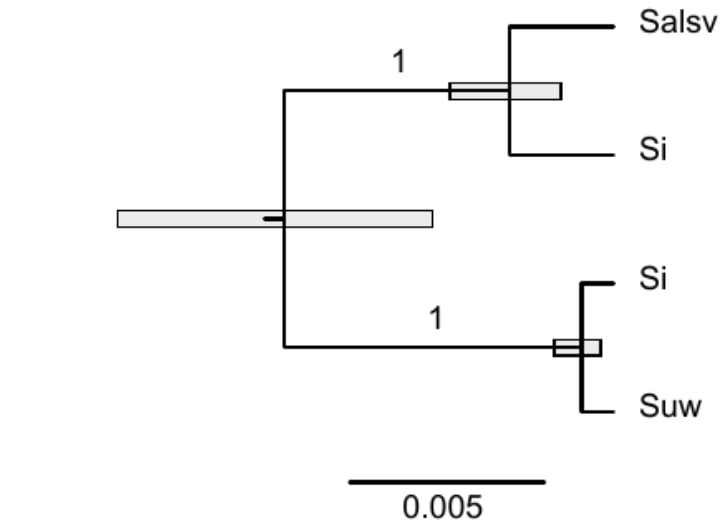
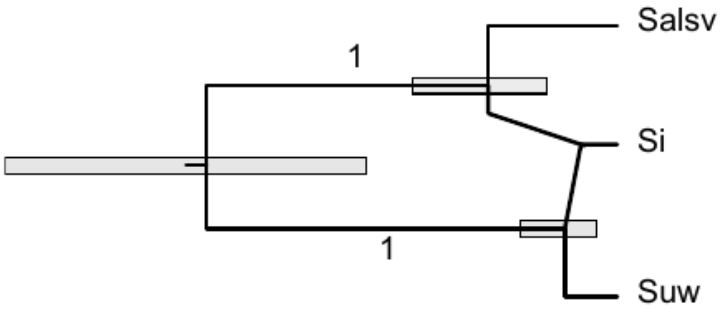


Figure 1 from Jones et al. (2013)

Example of inferring allopolyploidization in *Silene*



AlloppMUL



AlloppNET

Models implemented as add-ons to BEAST

Figure 7 from Jones et al. (2013)

AN EVALUATION OF THE HYBRID SPECIATION HYPOTHESIS FOR *XIPHOPHORUS CLEMENCIAE* BASED ON WHOLE GENOME SEQUENCES

Molly Schumer,[1,2] Rongfeng Cui,[3,4] Bastien Boussau,[5,6] Ronald Walter,[7] Gil Rosenthal,[3,4] and Peter Andolfatto[1,8]

Schumer et al. (2013) use synteny information and size of introgressed blocks to reject hybridization in favor of admixture.

Tools: PhyML_multi (Boussau et al., 2009) and windows of seq analyzed with AU test.

Lateral Gene Transfer

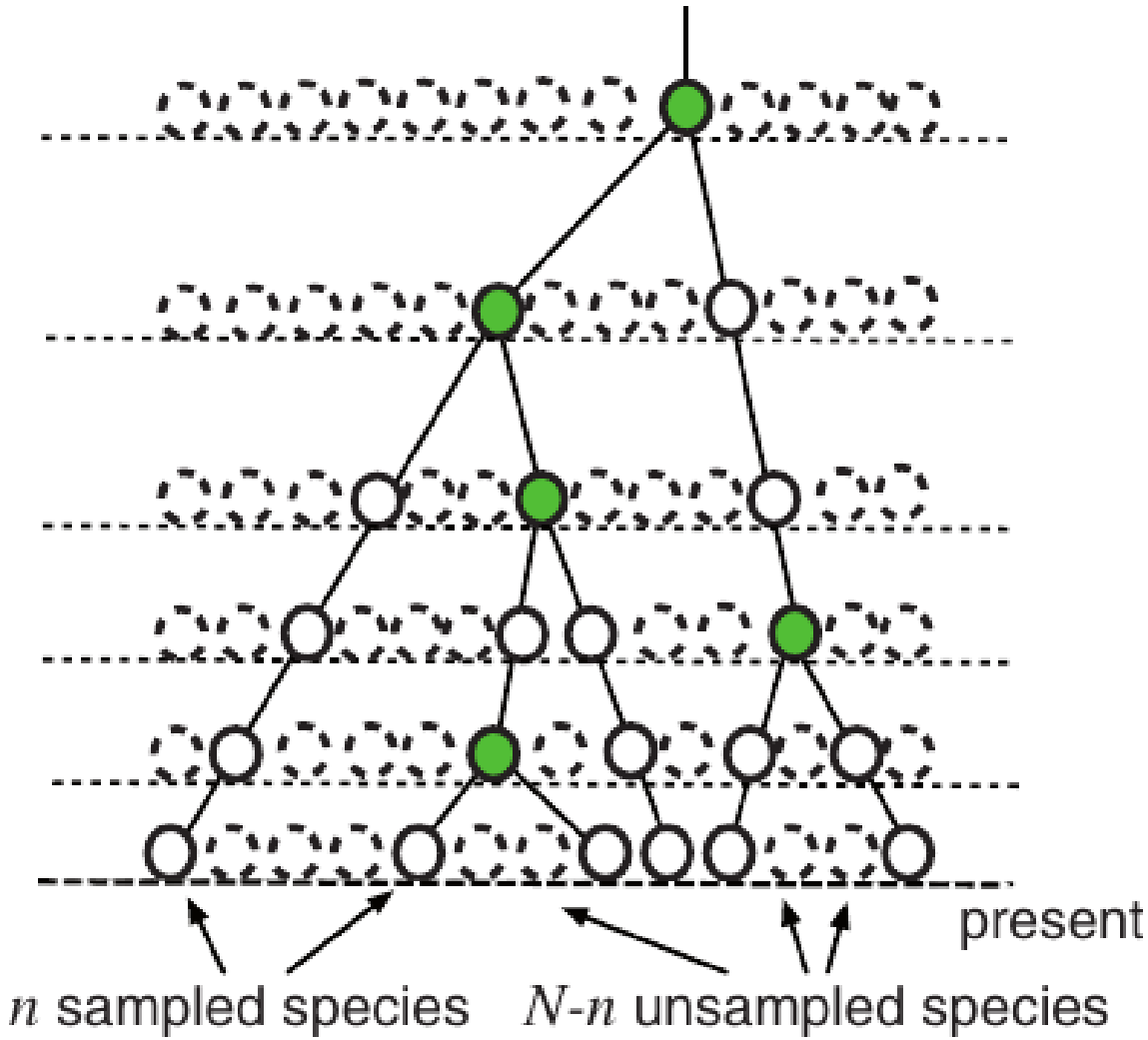


Figure 2c from Szöllősi et al. (2013)

Lateral Gene Transfer

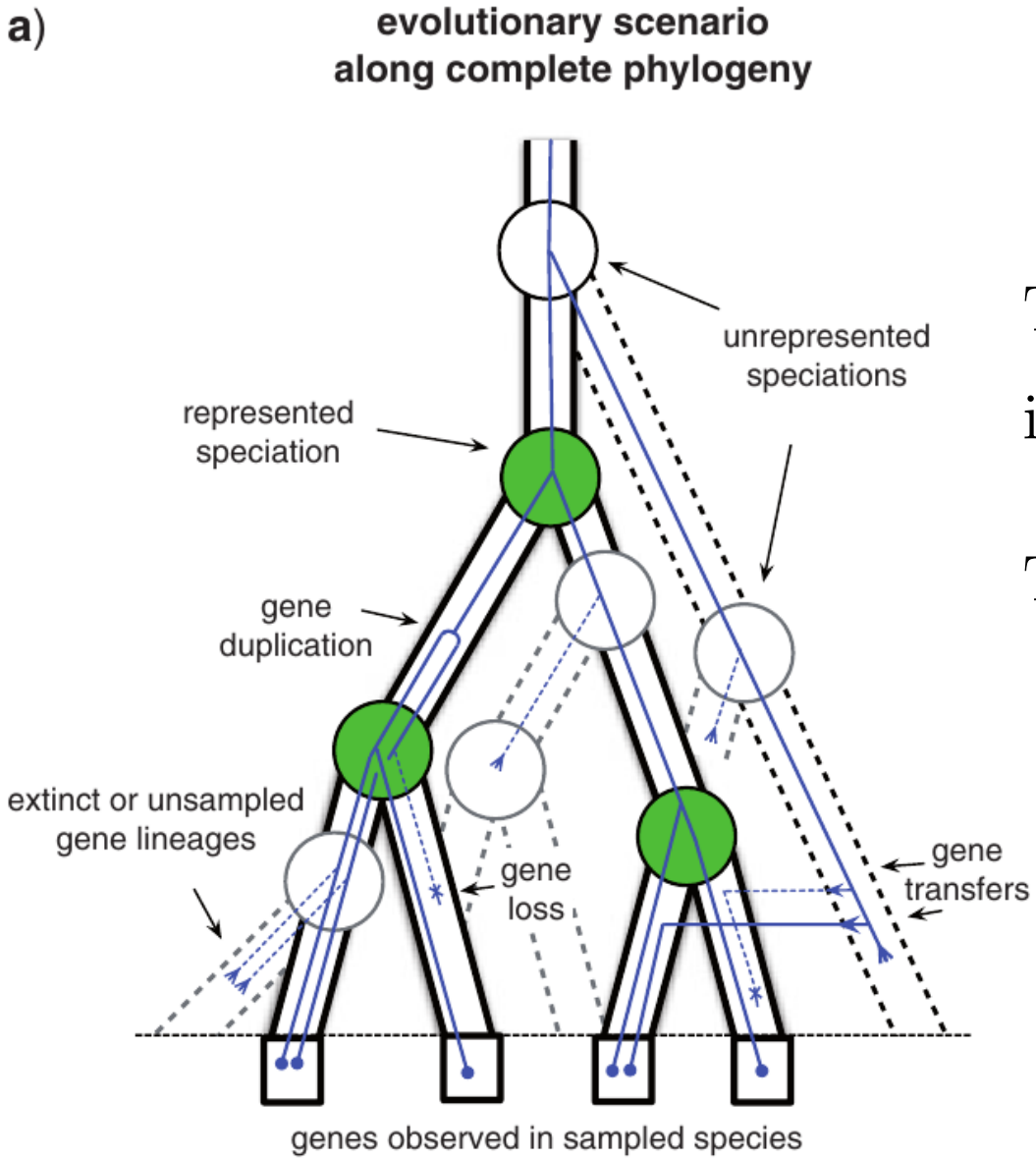


Figure 3 from Szöllősi et al. (2013)

They 423 single-copy genes

in $\geq$ 34 of 36 cyanobacteria

They estimate:

2.56 losses/family

2.15 transfers/family

$\approx$ 28% of transfers between non-overlapping branches

Separating gene and species trees

X = sequence data
G = a gene tree
T = a species tree

$$\mathbb{P}(X \mid T) = \sum_G \mathbb{P}(X \mid G)\mathbb{P}(G \mid T)$$

If the data strongly prefer one gene tree $\hat{G}$, then

$$\begin{aligned} \mathbb{P}(X \mid T) &\approx \mathbb{P}(X \mid \hat{G})\mathbb{P}(\hat{G} \mid T) \\ &\propto \mathbb{P}(\hat{G} \mid T) \end{aligned}$$

Unfortunately, nature is not cooperating. . .

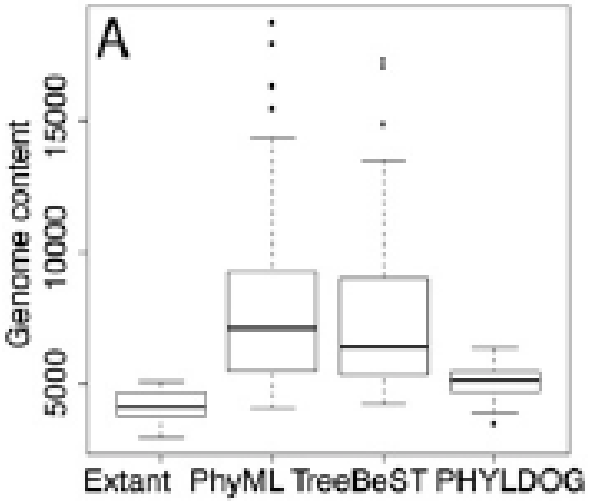


Fig. 4 from Boussau et al. (2013)

- "We also show that the common practice of disregarding reconcilability in gene tree inference overestimates the number of LGT and duplication events." from Sjöstrand *et al* "A Bayesian Method for Analyzing Lateral Gene Transfer" *Syst. Biol.* 2014

Separating gene and species tree estimation

Jointly estimate gene trees and species tree $\rightarrow$ less incongruence.
This implies that we need:

- to "integrate out" uncertainty in gene trees.
- software implementing both
  - state-of-the-art substitution models and
  - gene-tree/species-tree reconciliation.

Part 2 summary

- Several research groups are tackling the disentangling different pieces of the:
  - coalescent history,
  - locus history,
  - hybridization history,
  - lateral gene transfer history
  - recombination history
- These problems overlap.
- It seems optimistic to think that we'll be able to jointly infer all of these processes in one piece of software (any time soon)

Part 3: statistics on trees are tough

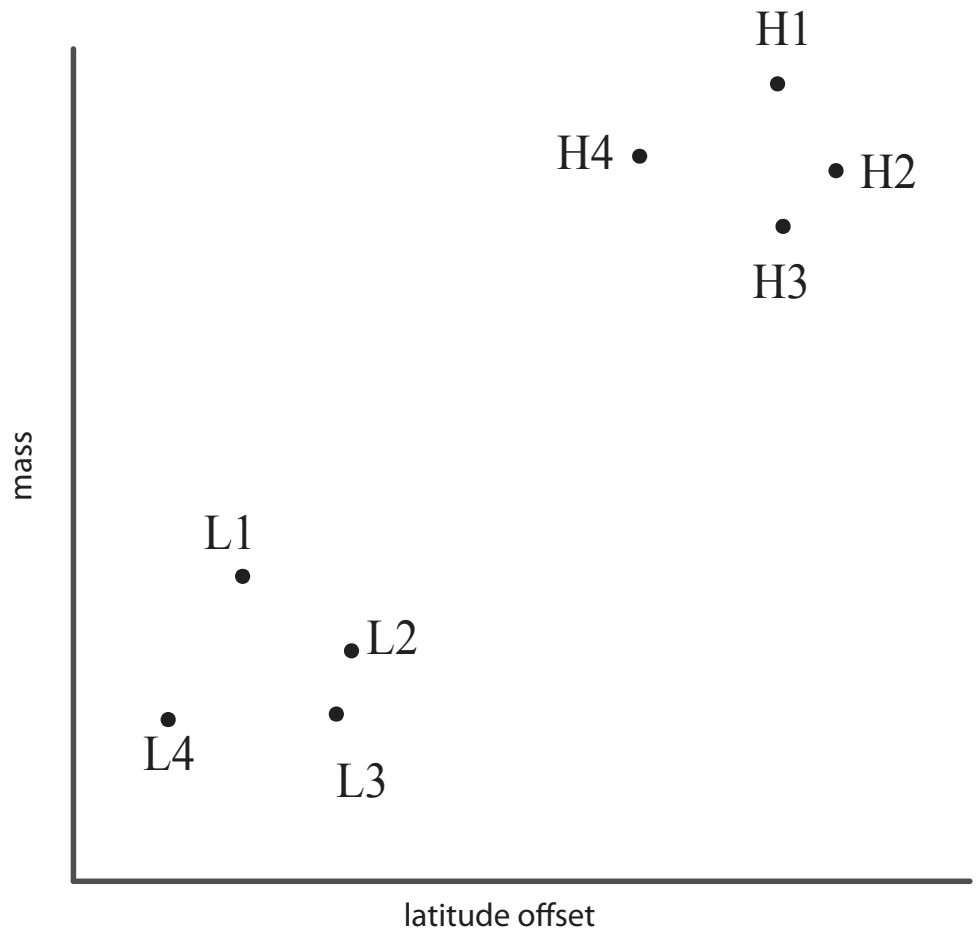
- It is very difficult to create a confidence set of trees
- Very nice adjusted bootstrap proportion work by Susko focused on single branch support,
- AU and SH tests (Shimodaira) require specification of a candidated set of trees
- Susko (2014) just corrected the null distribution for the KH test

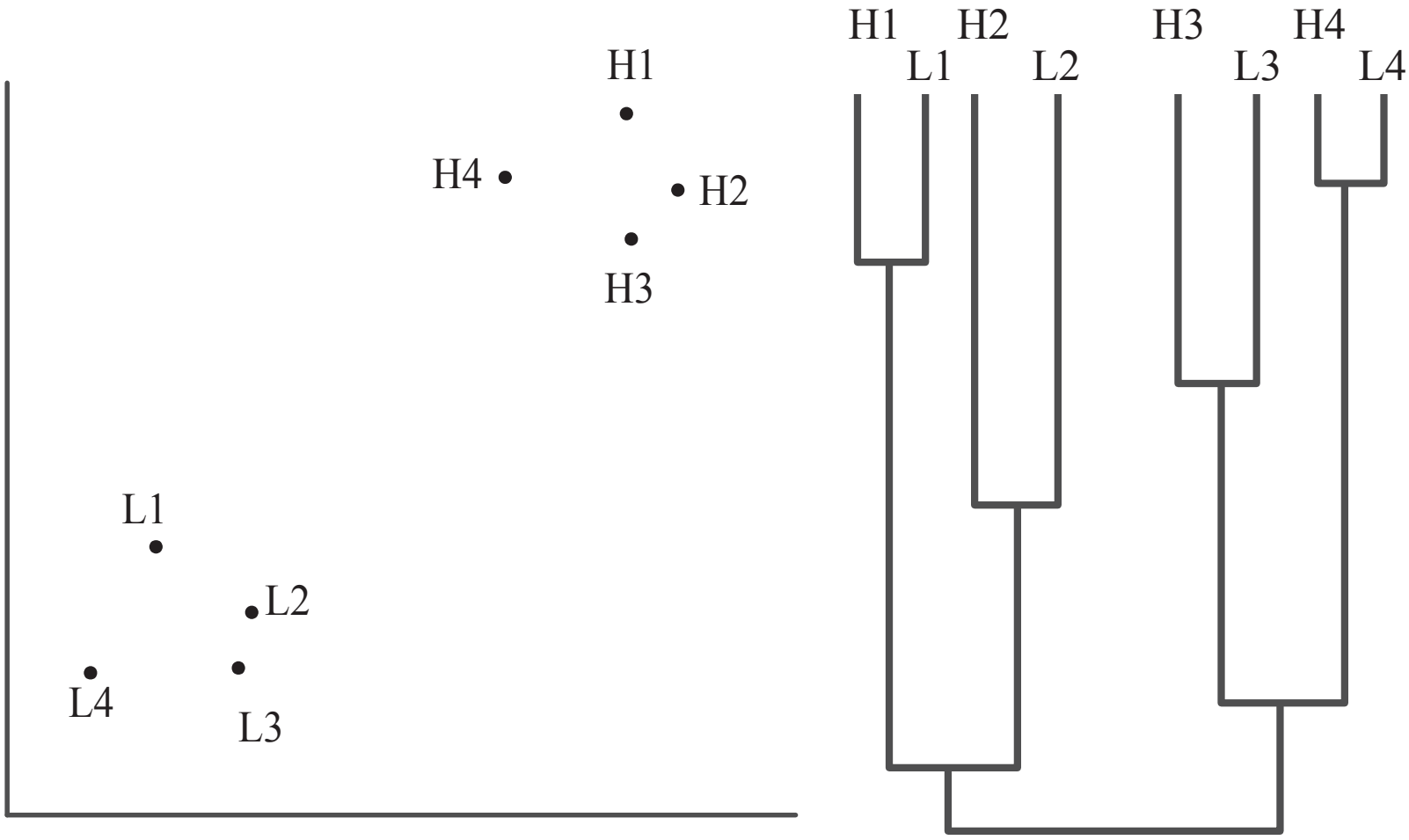
What if we just want to test for correlated evolution between a couple of characters?

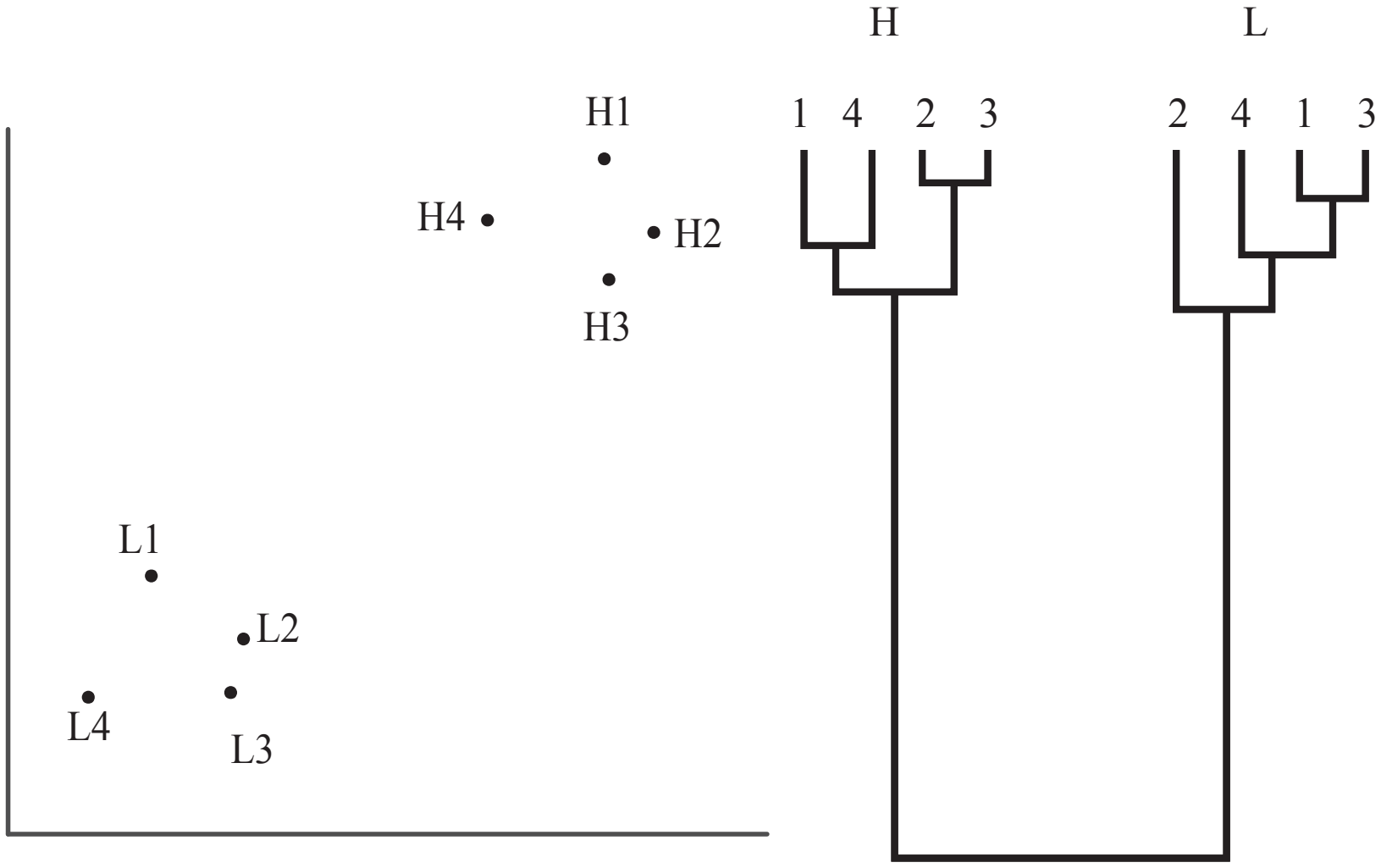
The standard approach:

1. Calculate a test statistic of correlation, $Z(X)$
2. Consider the null distribution: values of $Z(X)$ given no correlated evolution.
3. Calculate $P = \mathbb{P}(z \geq Z(X) \mid H_0)$
4. if $P < 0.05$, reject the null

Testing for correlated evolution between a latitude and mass using data from 8 species:




H1
H4
H2
H3
L1
L2
L4
L3
H1
L1
H2
L2
H3
L3
H4
L4


H
L
H1
1 4 2 3
2 4 1 3
H4
H2
H3
L1
L2
L4
L3

What if we don't know the tree?

The evidence for correlation depends on T, but we don't know what T is correct.

What if we don't know the tree?

We could integrate out the tree:

$$P = \mathbb{P}(z \geq Z(X) \mid H_0) = \int \mathbb{P}(z \geq Z(X) \mid T)\mathbb{P}(T \mid X, H_0)dT$$

What if we don't know the tree?

We could integrate out the tree:

$$p = \mathbb{P}(z \geq Z(X) \mid H_0) = \int \mathbb{P}(z \geq Z(X) \mid T)\mathbb{P}(T \mid X, H_0)dT$$

Unfortunately, this is not a valid p value.

In frequentist statistics, the p value is suppose to be the largest p attainable under the null (this is called the "least favorable condition" for the test)

Berger and Boos. 1994. "P Values Maximized Over a Confidence Set for the Nuisance Parameter." *Journal of the American Statistical Association.* **89(427)**. 1012–1016.

To calcuate a P value, when there is an unknown, nuisance parameter, θ:

1. Calculate a $(1 - \beta)$ confidence set for θ (e.g for a 99% confidence set, $\beta = 0.01$)
2. Calculate a P value for every θ in the confidence set: call this vector $p(\theta)$
3. $P = \max[p(\theta)] + \beta$

In phylogenetics, if we used Berger and Boos' method, we would need to:

1. Get a 99% confidence set. The AU test could help, but this could be a very large set trees
2. Conduct the comparative method assuming each of the trees, and store the highest P value
3. Report 0.01 + the highest P value

We tend to simply perform the comparative method over a collection of trees (from bootstrapping or MCMC) and report a mean.

It is not clear (to me) whether we *should* be using the Berger and Boos method, instead.

My clashing main points

1. Felsenstein showed us the way: conduct full, likelihood based inference of models.
2. We are starting to be able to deal with the intricate connections between levels of inference (gene tree, locus tree, species tree...), but it is hard to do this simultaneously.
3. We don't even have access to easy frequentist probabilities

The unstated, but vital point about the "state of the phylogenetics union"

It is a really cool time to be working in phylogenetics.

1. Thanks to incredible advances in hardware and software, very rich analyses are now possible
2. Phylogenetics continues to be a fertile meeting place for biologists, statisticians, mathematicians, and computer scientists.
3. Even as our image of the "tree" of life changes, genealogical relationships remain central to analyzing comparative data.

A bit of (unsolicited) advice . . .

Think about making a convincing arguments rather than just running the recommended tests/analyses.

"Politicians use statistics in the same way that a drunk uses lamp-posts — for support rather than illumination." Andrew Lang

- We scientists should be using statistics for our own illuminaton,

"Politicians use statistics in the same way that a drunk uses lamp-posts — for support rather than illumination."
Andrew Lang

- We should be using statistics for our own illuminaton, but we should also think of statistics as a communication tool.
- Don't perform *only* the analyses that would convince *you*, make sure that your claims are convincing to a wide range of viewpoints.

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument).

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument).
2. If we used this procedure a lot, we'd bracket the true value 95% of the time (the confidence interval)

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument).
2. If we used this procedure a lot, we'd bracket the true value 95% of the time (the confidence interval).
3. If $\mathbb{P}(\theta)$ describes our beliefs about plausible models before seeing the data, after we see the data our beliefs should be $\mathbb{P}(\theta|X)$. The Bayesian posterior probability.

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument),
2. If we used this procedure a lot, we'd bracket the true value 95% of the time (the confidence interval)
3. If $\mathbb{P}(\theta)$ describes our beliefs about plausible models before seeing the data, after we see the data our beliefs should be $\mathbb{P}(\theta|X)$. The Bayesian posterior probability.
4. If we use this threshold for evidence, we would expect 5% of our positives to be false positives (the false discovery rate)

The p-value argument

These data would be unlikely if the null were true.
Pros:

1. Very conservative. Gives the null the benefit of the doubt by focusing on the "least favorable" conditions.
2. If you don't reject the null, then your conclusions resistant to the addition of more models.

Cons:

1. Can be surprisingly difficult to correctly calculate p
2. Does not give you inference to the best model.
3. Failure to reject null often caused by lack of data.

The confidence interval argument

Pros:

1. Somewhat intuitive.
2. Identifies a plausible set of answers.
3. Not dependent on prior knowledge of parameters.

Cons:

1. Can be surprisingly difficult to correctly calculate
2. Ignores prior information
3. Does not fully condition on the data you observed.

The Bayesian posterior probability argument. $\mathbb{P}(\theta|X)$.

Pros:

1. Very intuitive statement of the best range of parameters
2. The only coherent statement of knowledge that uses all of the information in the data

Cons:

1. Unconvincing to people with different priors
2. Its unclear what would happen if you consider another model
3. Relies on MCMC (which is cool, but dangerous)

The False discovery rate argument.

Pros:

1. Nice mixture of frequentist and emprical Bayesian behavior
2. Check out Nicolas Lartillot's blog `http://bayesiancook.blogspot.com/` for convincing arguments that evolutionary genomics has an empirical Bayesian future.

Cons:

1. You're not going to see a lot of software in phylogenetics that spits out FDR values.

For each analysis/test that we talk about in the course:

1. make sure that you understand the "signal":
    - sketch out cases of the tree or data that would look uninteresting,
    - sketch what interesting data would look like
2. What confounding factors could lead to similiar signal?

    - *e.g.* analyses of diversification times and rates depend crucially on branch lengths.
    - models that are too simplistic distort branch lengths (mainly they underestimate deep and long branches)

For each analysis/test that we talk about in the course:

1. make sure that you understand the "signal":
   - sketch out cases of the tree or data that would look uninteresting,
   - sketch what interesting data would look like
2. What confounding factors could lead to similiar signal?
3. If a confounding factor is not included in your analysis, consider doing some simulations to assess whether your result could be an artifact.
4. What kind of statistical arguments (p value, CI, Bayesian, FDR) would be most convincing to a skeptic?

References I

Åkerborg, Örjan., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, 106(14):5714–5719.

Atchley, W. R. (2011). Walter m. fitch (19292011). *Science*, 332(6031):804.

Avise, J. C., Ankney, C. D., and Nelson, W. S. (1990). Mitochondrial gene trees and the evolutionary relationship of mallard and black ducks. *Evolution*, 44(4):pp. 1109–1119.

Boussau, B., Guéguen, L., and Guoy, M. (2009). A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. *Evolutionary Bioinformatics Online*, 5:67–79.

References II

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330.

Chang, D. and Duda, T. F. (2012). Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Molecular Biology and Evolution*, 29(8):2019–2029.

Dunn, K. A., Jiang, W., Field, C., and Bielawski, J. P. (2013). Improving evolutionary models for mitochondrial protein data with site-class specific amino acid exchangeability matrices. *PLoS ONE*, 8(1):e55816.

Dutheil, J. Y., Galtier, N., Romiguier, J., Douzery, E. J., Ranwez, V., and Boussau, B. (2012). Efficient selection of branch-specific models of sequence evolution. *Molecular Biology and Evolution*, 29(7):1861–1874.

References III

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580.

Jones, G., Sagitov, S., and Oxelman, B. (2013). Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Systematic Biology*.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7):1307–1320.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

References IV

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765.

Schumer, M., Cui, R., Boussau, B., Walter, R., Rosenthal, G., and Andolfatto, P. (2013). An evaluation of the hybrid speciation hypothesis for xiphophorus clemenciae based on whole genome sequences. *Evolution*, pages no–no.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*.

Vernot, B., Stolzer, M., Goldman, A., and Durand, D. (2008). Reconciliation with non-binary species trees. *Journal of Computational Biology*, 15(8):981–1006.

References V

Yoder, J. B., Briskine, R., Mudge, J., Farmer, A., Paape, T., Steele, K., Weiblen, G. D., Bharti, A. K., Zhou, P., May, G. D., Young, N., and Tiffin, P. (2013). Phylogenetic signal variation in the genomes of medicago (fabaceae). *Systematic Biology*.

Zoller, S. and Schneider, A. (2013). Improving phylogenetic inference with a semiempirical amino acid substitution model. *Molecular Biology and Evolution*, 30(2):469–479.

Better methods for estimating the timing of events

- richer models for relaxing the molecular clock
- use of fossil taxa as tips
- Tracy will cover divergence time estimation

Much richer models of sequence evolution

- Model-averaging during tree inference (MrBayes 3.2; RBS-add-on to BEAST; CAT model in PhyloBayes)
- Empirically-derived matrices. See: Zoller and Schneider (2013) for a nice semi-empirical approach and Dunn et al. (2013) for a model that groups sites based on physiochemical properties
- Models that change over the tree

Analysis of mutational mapping to identify model changepoints

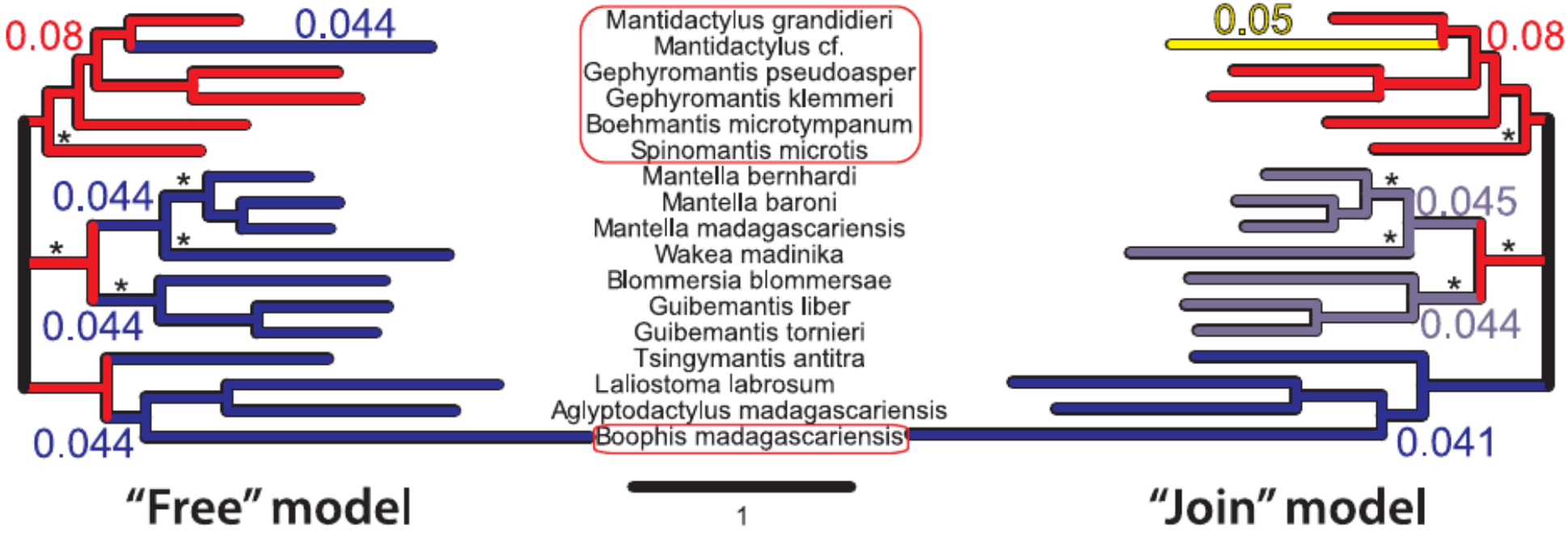


Figure 7 form Dutheil et al. (2012)

Improved modeling of diversification rates

- Better ways of dealing with estimating speciation and extinction from trees with unsampled taxa (see the work of Tanja Stadler)
- Models of characters that affect diversification rates (BiSSE)
- Brian will cover analyses of diversification rates

Stochastic character mapping/ robust counting

- Better assessment of uncertainty than an ancestral character state reconstruction,
- mappings easier to analyze with other character data,
- inferring mappings under an simple model is often quite robust.