Reconstructing species trees and testing the coalescent model (and placing polyploid species)

Cécile Ané

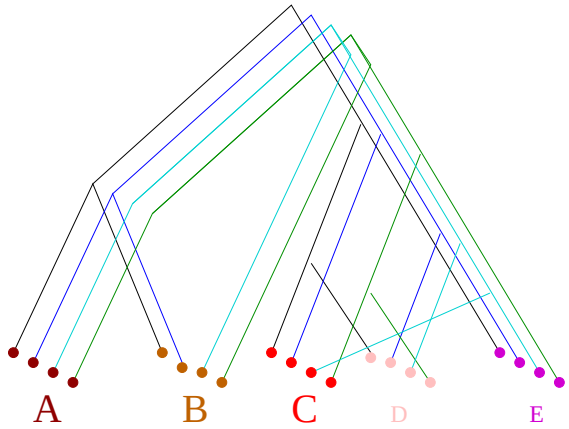
Departments of Statistics and of Botany
University of Wisconsin - Madison

MCEB conference,
June 18-22, 2012

1. **Gene Tree / Species Tree** methods and the coalescent

2. **Concordance** analysis, and testing the coalescent model

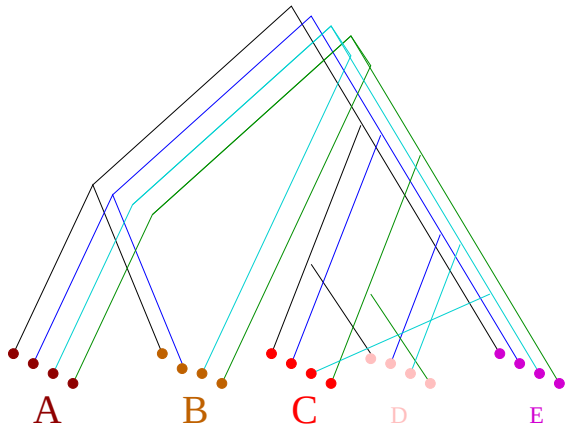
3. **Polyploids**: gene tree discordance and reticulate evolution



1. **Gene Tree / Species Tree** methods and the coalescent

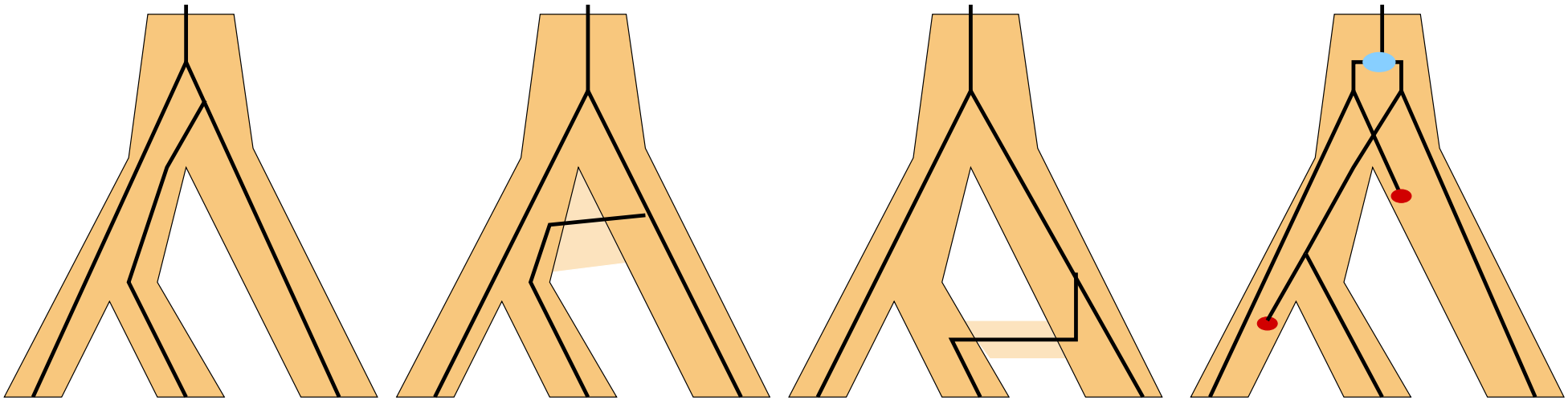
2. **Concordance** analysis, and testing the coalescent model

3. **Polyploids**: gene tree discordance and reticulate evolution



Reasons for gene tree discordance

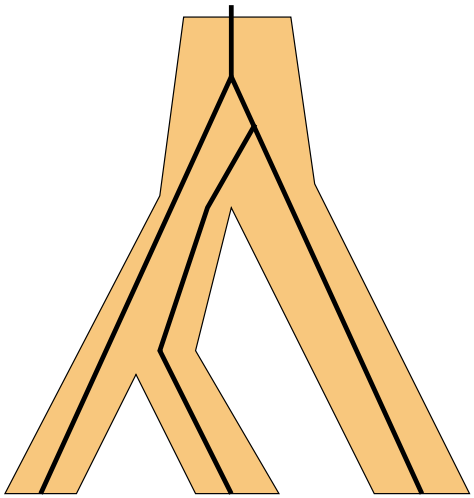
The conflict between gene trees can be real:



Incomplete lineage sorting | Gene flow | Horizontal gene transfer Hybridization | Undetected paralogy

or simply estimation error, including systematic biases.

Mathematical model for incomplete lineage sorting

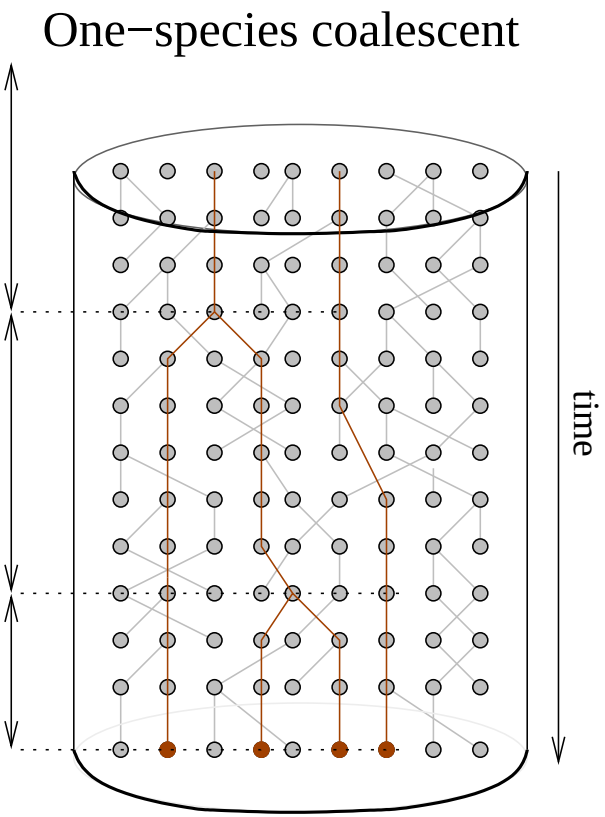


Incomplete lineage sorting

Kingman's coalescent model: backward in time

Wright-Fisher model: forward in time.

The coalescent model



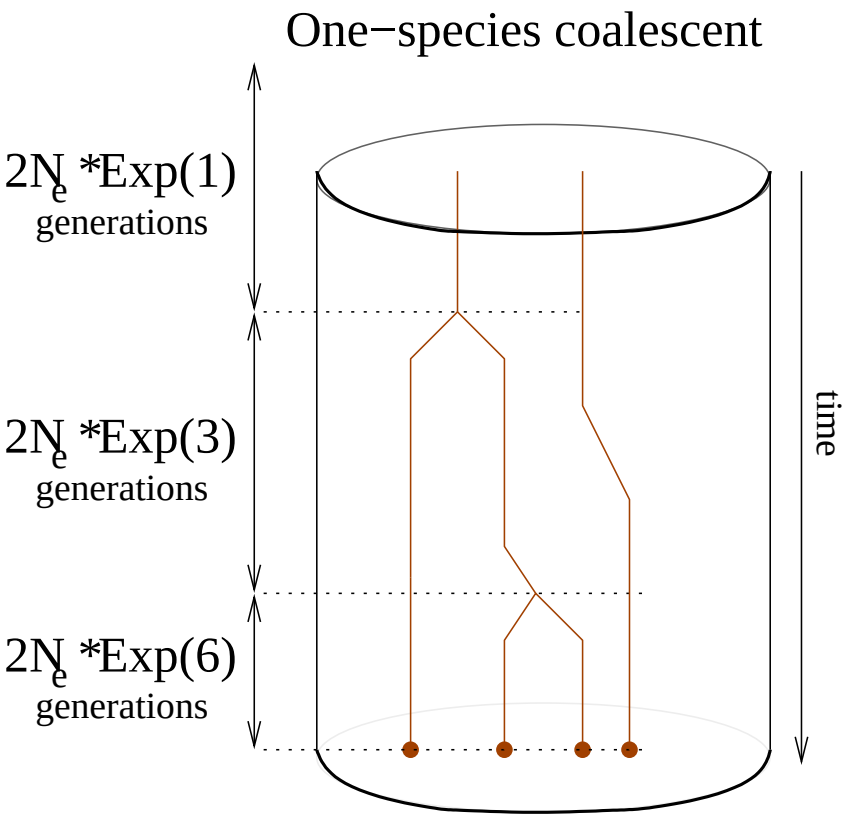
Assume: **panmictic** population of N_e **diploid** individuals, **no selection** at the locus, large N_e.

T = time to the next coalescence among k copies:

$$T/(2N_e) \sim \mathcal{E}\left(\frac{k(k-1)}{2}\right)$$

$k = 2$ genes do *not* coalesce during t generations with probability $\exp(-\frac{t}{2N_e})$

The coalescent model



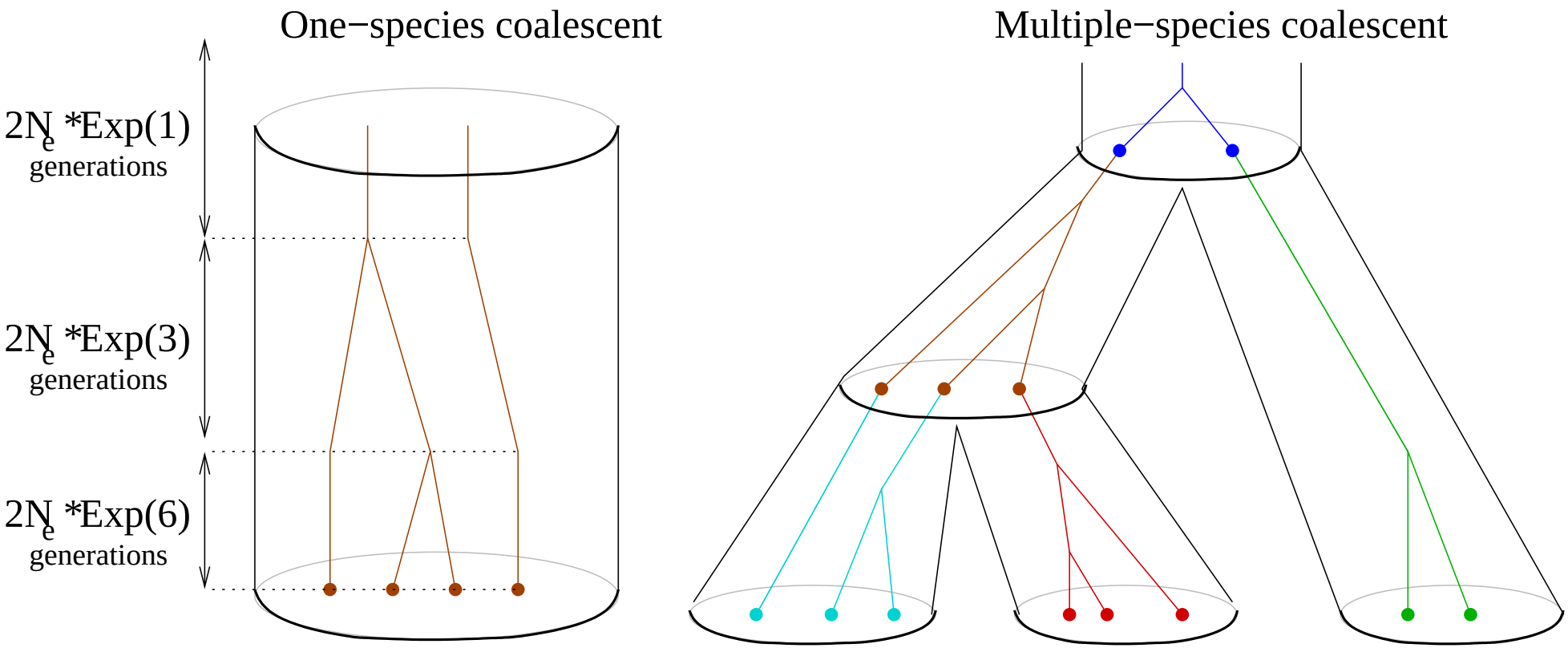
Assume: **panmictic** population of N_e **diploid** individuals, **no selection** at the locus, large N_e.

T = time to the next coalescence among k copies:

$$T/(2N_e) \sim \mathcal{E}\left(\frac{k(k-1)}{2}\right)$$

$k = 2$ genes do *not* coalesce during t generations with probability $\exp(-\frac{t}{2N_e})$

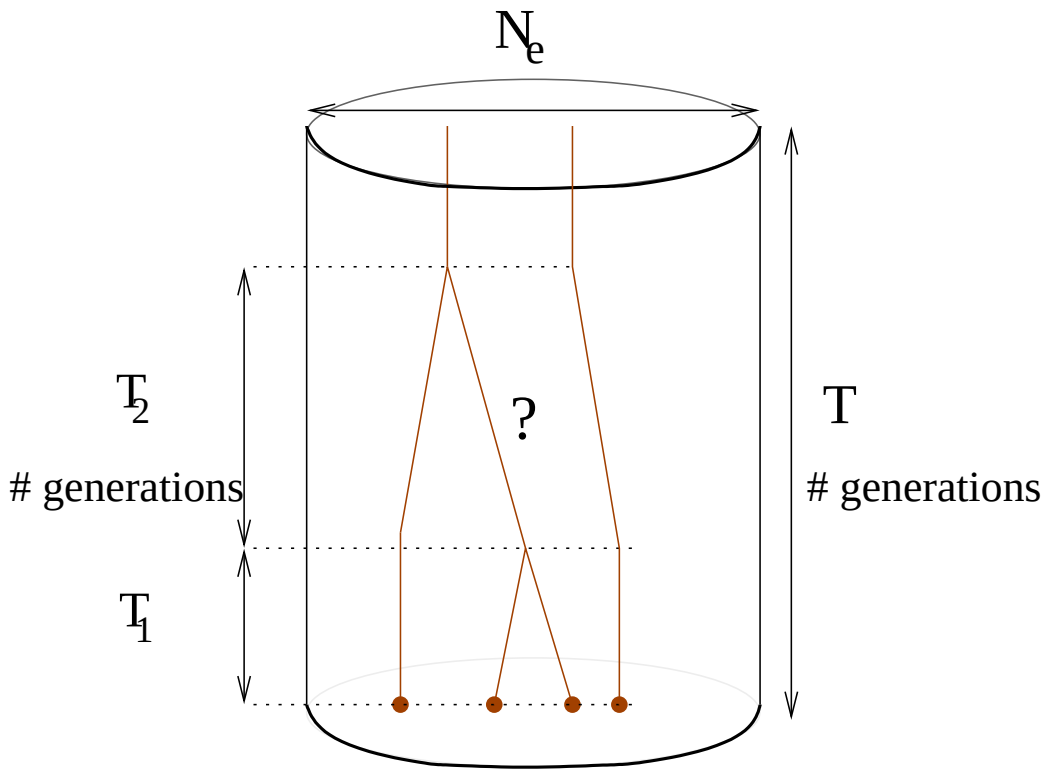
The multispecies coalescent model



Branch lengths: Assumptions and Issues

- In gene trees: observe **substitutions** / site
- In species tree: coalescent model needs **coalescent units**: $u = \#\text{ generations } / 2N_e$.

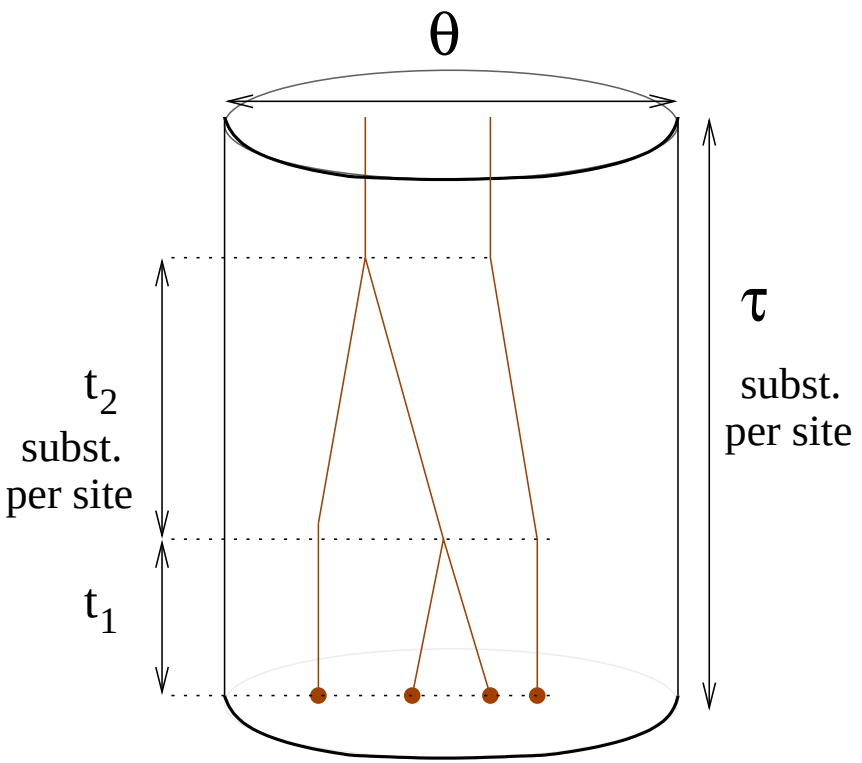
Issue 1: **Non-identifiability** of # generations and N_e separately.



Branch lengths: Assumptions and Issues

Fix 1:

- rescale # generations: $\tau = \mu *$ # generations, μ in subst/site/gen.
- rescale pop size: $\theta = 4\mu * N_e$, or use instead coalescent units $u = 2\tau/\theta$.



$$\mathbb{P}(\text{ gene tree and lengths } t_i) =$$

$$\frac{2}{\theta}\exp\left(-\frac{12\,t_1}{\theta}\right) * \frac{2}{\theta}\exp\left(-\frac{6\,t_2}{\theta}\right) * \exp\left(-\frac{2(\tau-(t_1+t_2))}{\theta}\right)$$

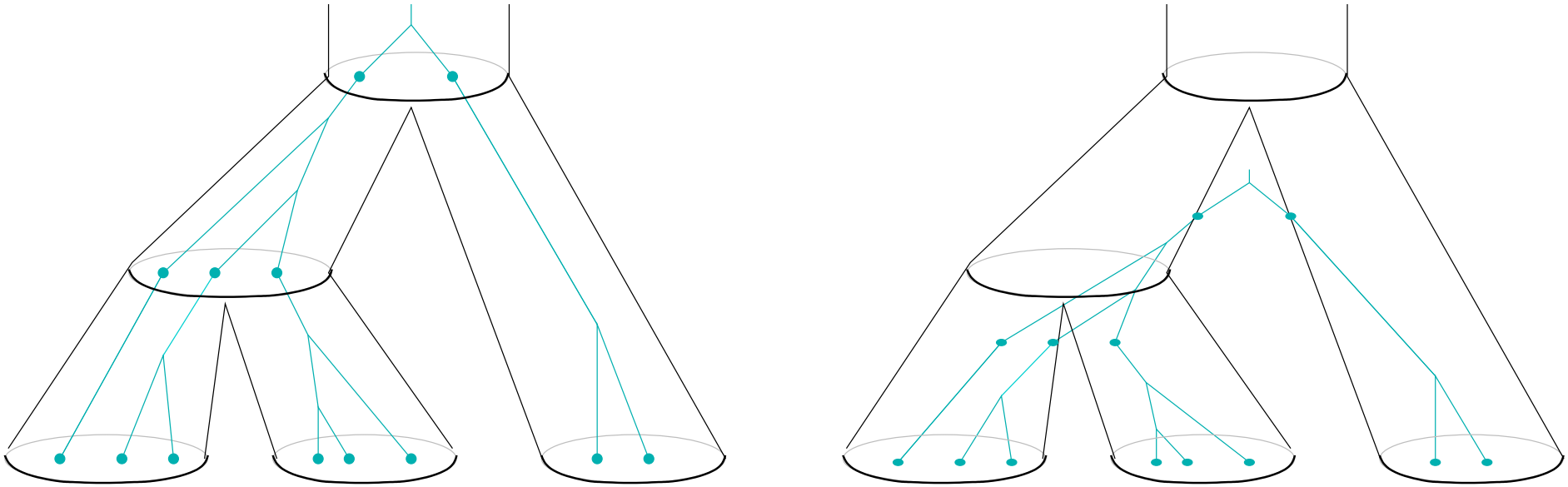
$\mathbb{P}(\text{gene topology}) = \text{function of } u \text{ only.}$

Yang & Rannala (2003)

Degnan & Salter (2005)

Branch lengths: Assumptions and Issues

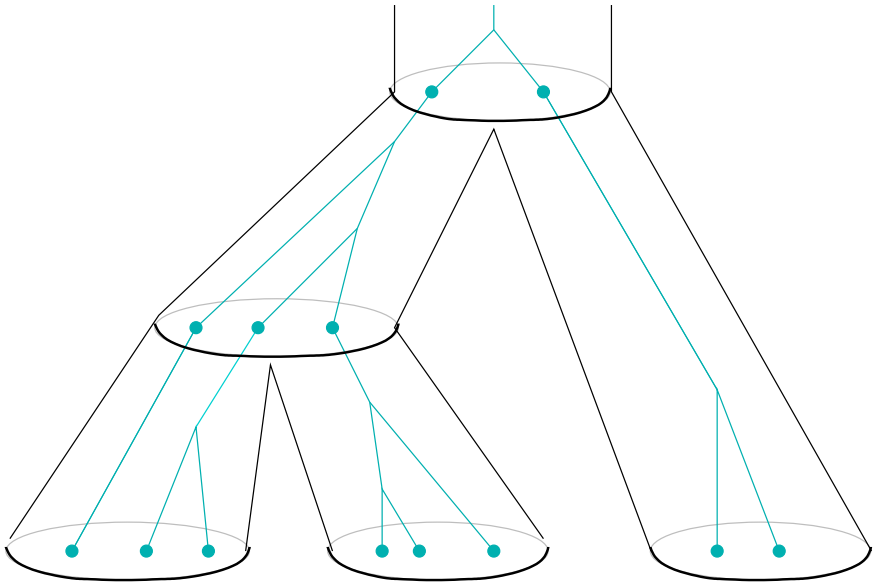
Issue 2: **gene-specific rate variation**.
slow genes force early divergence times,
fast genes force deep coalescent events.

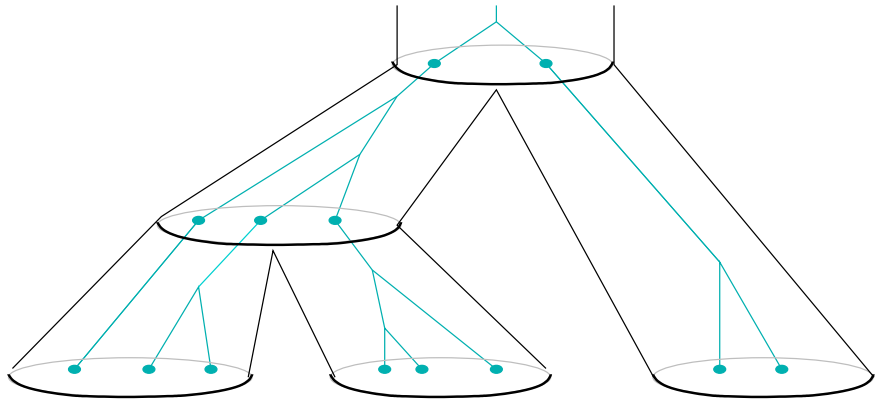


Fix 2: **ad-hoc rescaling** of branch lengths in gene trees, or use information from gene **topologies only**, i.e. ranks of coalescences, triples / quartets, or whole gene topologies.

Branch lengths: Assumptions and Issues

Issue 2: **gene-specific rate variation**.
slow genes force early divergence times,
fast genes force deep coalescent events.

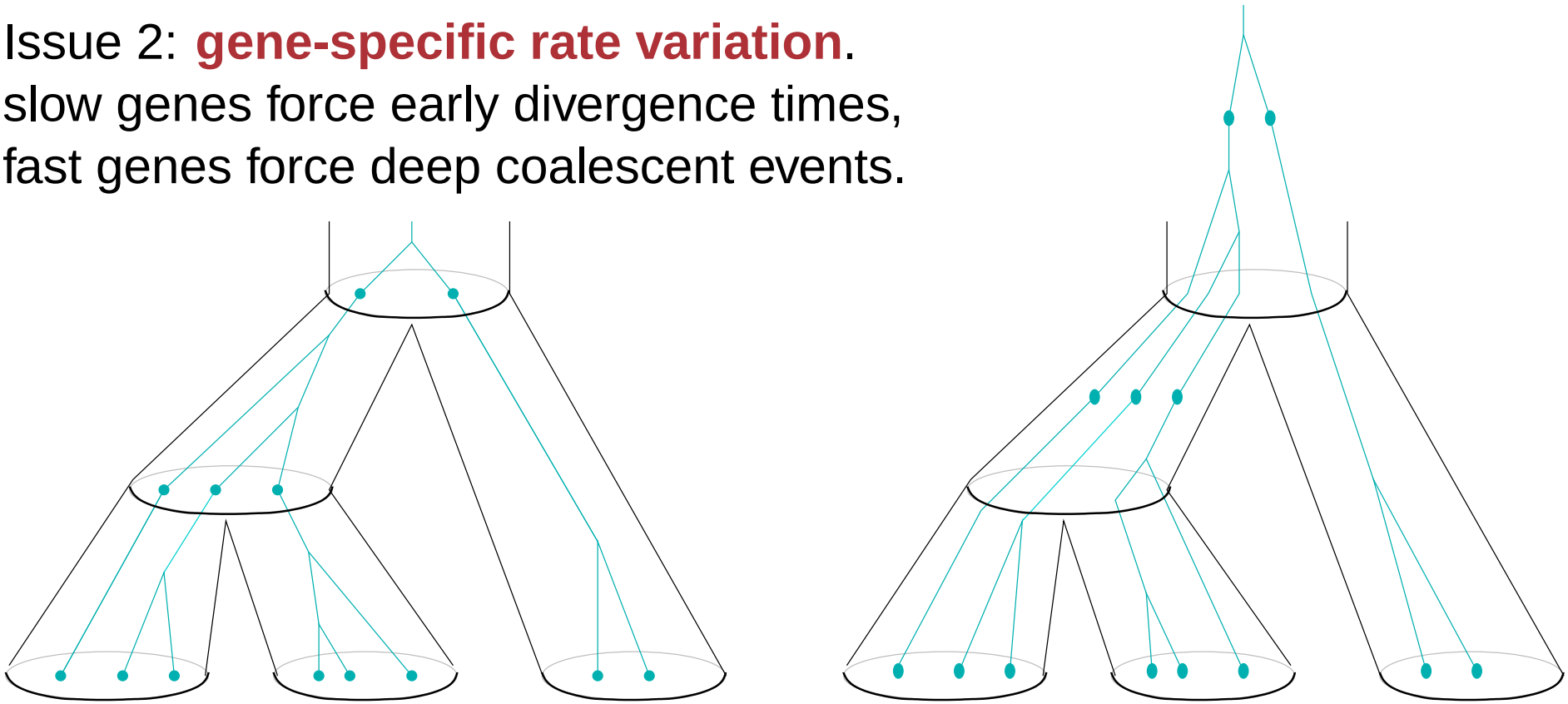




Fix 2: **ad-hoc rescaling** of branch lengths in gene trees, or use information from gene **topologies only**, i.e. ranks of coalescences, triples / quartets, or whole gene topologies.

Branch lengths: Assumptions and Issues

Issue 2: **gene-specific rate variation**.
slow genes force early divergence times,
fast genes force deep coalescent events.



Fix 2: **ad-hoc rescaling** of branch lengths in gene trees, or use information from gene **topologies only**, i.e. ranks of coalescences, triples / quartets, or whole gene topologies.

Method	Need perfect gene trees?	Assume coalescent?	branch lengths with clock?	Need rooted gene trees?
STEM	yes	yes	yes	yes
GLASS	yes	yes	yes	yes
STEAC	yes	yes	yes	yes
STAR	yes	yes	no	yes
NJst	yes	yes	no	no
MP-EST	yes	yes	no	yes
STELLS	yes	yes	no	yes
RT-con	yes	no	no	yes
MDC	yes	no	no	no
BEST	no	yes	yes, ad-hoc rate variation	yes
*BEAST	no	yes	yes	no
BUCKy	no	no	no	no
HybTree	yes	yes+hybr.	no	yes
Yu et al.	yes	yes+hybr.	no	yes

To delimit species: BPP and O'Meara method.

STEM, MP-EST, STELLS: likelihood-based

Kubatko et al (2009), Liu et al (2010), Wu (2011)

Input: 1 tree for each gene.
Assume the coalescent.

STEM: ML using branch lengths (rescaled). User specifies θ.
Species tree S and τ (scaled # generations) to maximize:

$$\prod_{\text{genes } i} \mathbb{P}(T_i, t_i | S, \tau, \theta)$$

MP-EST: Pseudo-ML: combines likelihood of all triples.
Species tree S and $u = 2\tau/\theta$ (coal. units) to maximize:

$$\prod_{\text{genes } i} \prod_{\text{triples } \{a,b,c\}} \mathbb{P}\left(T_{i|\{a,b,c\}} | S, u\right)$$

STELLS: ML of gene topologies. Species tree S and $u = 2\tau/\theta$ to maximize:

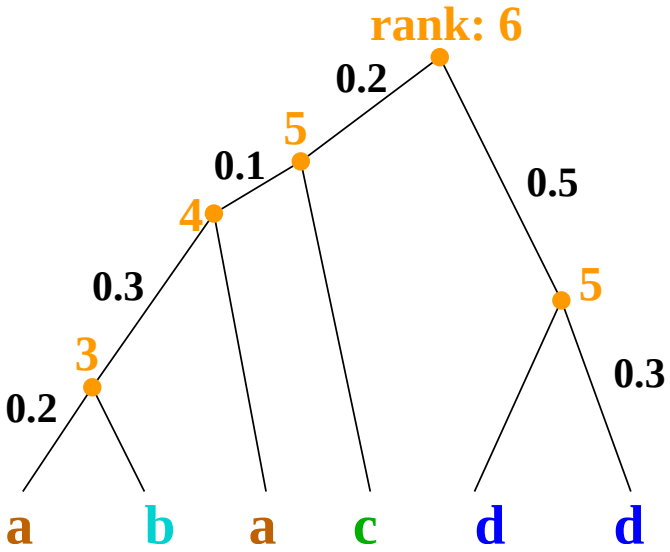
$$\prod_{\text{genes } i} \mathbb{P}(T_i | S, u)$$

GLASS, STEAC, STAR, NJst: from summary statistics

Liu et al (2009), Liu & Yu (2011)

Input: 1 tree for each gene.

1. Form a distance matrix between species, from: minimum coalescent times (GLASS/MT), Average Coalescent times (STEAC), Average Ranks of coalescences (STAR), or average internode distance (NJst)



$$
\begin{aligned}
d_{\mathrm{GLASS}}(A,B) &= 0.2 \\
d_{\mathrm{STEAC}}(A,B) &= (0.2+0.5)/2 = 0.35 \\
d_{\mathrm{STAR}}(A,B) &= (3+4)/2 = 3.5 \\
d_{\mathrm{NJ_{st}}}(A,B) &= (1+2)/2 = 1.5
\end{aligned}
$$

2. Use Neighbor Joining to get a tree from this distance matrix

BEST and *BEAST

Liu (2008), Heled & Drummond (2010)

Input: 1 alignment for each gene.

Posterior probability of gene trees and species tree:

$$\underbrace{\mathbb{P}(\text{sequences} \mid \text{gene trees}, t)}_{\text{substitution model}} * \underbrace{\mathbb{P}(\text{ gene trees}, t \mid \text{ species tree}, \theta, \tau)}_{\text{coalescent model}}$$

Bayesian framework, MCMC, computationally intensive,
different population sizes θ on different branches in the species tree

BEST: ad-hoc rate variation between lineages to evaluate $\mathbb{P}(\text{sequences}|\text{gene trees}, t)$ on non-ultrametric trees.

*BEAST: θ can vary continuously at speciation nodes.

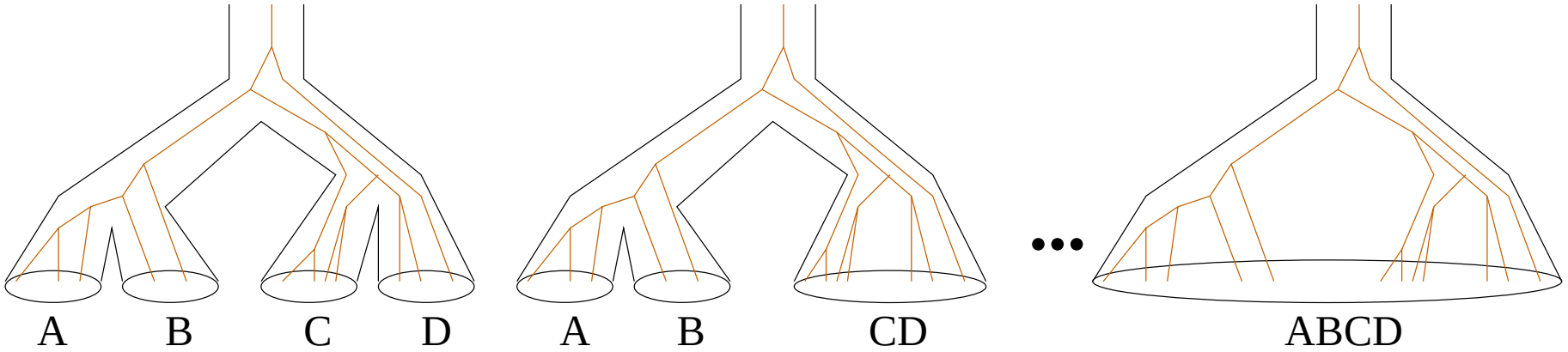
BPP: species delimitation under coalescent

Yang & Rannala (2010)

Input: 1 alignment for each gene, also:

- assignment of individuals to **putative** species
- **guide tree** on these putative species.

Bayesian framework, MCMC, search through limited set of species trees:



Posterior probability of gene trees and species tree:

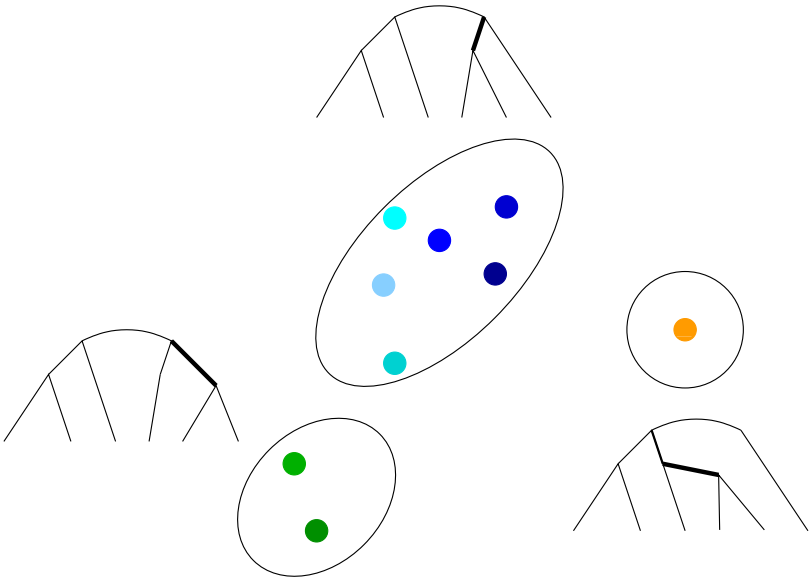
$$\underbrace{\mathbb{P}(\text{sequences} \mid \text{gene trees}, t)}_{\text{substitution model}} * \underbrace{\mathbb{P}(\text{ gene trees}, t \mid \text{ species tree}, \theta, \tau)}_{\text{coalescent model}}$$

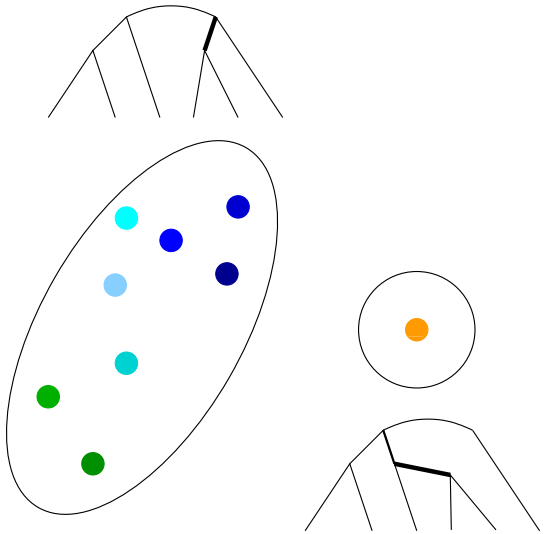
BUCKy: Bayesian concordance analysis

Ané et al (2007), Larget et al (2010)

Bayesian framework, non-parametric prior on gene trees, based on clustering genes. Posterior probability of gene trees:

$$\underbrace{\mathbb{P}(\text{sequences} \mid \text{gene trees})}_{\text{substitution model}} * \underbrace{\mathbb{P}(\text{ gene trees}|\alpha)}_{\text{Dirichlet prior}}$$



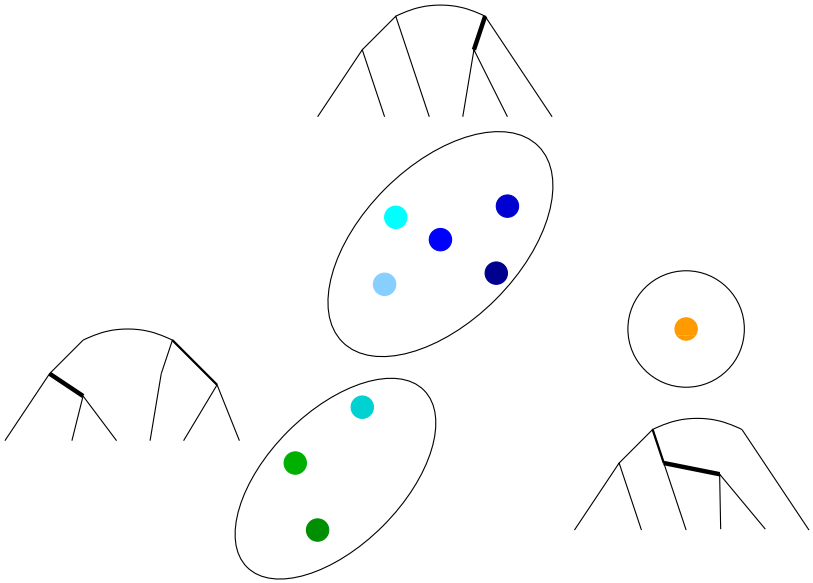


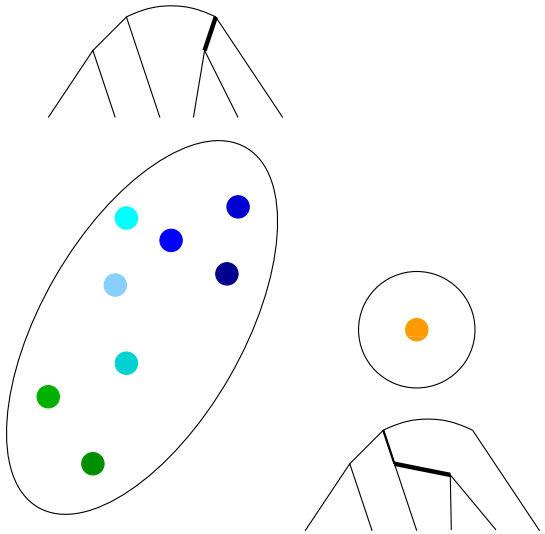
BUCKy: Bayesian concordance analysis

Ané et al (2007), Larget et al (2010)

Bayesian framework, non-parametric prior on gene trees, based on clustering genes. Posterior probability of gene trees:

$$\underbrace{\mathbb{P}(\text{sequences} \mid \text{gene trees})}_{\text{substitution model}} * \underbrace{\mathbb{P}(\text{ gene trees}|\alpha)}_{\text{Dirichlet prior}}$$





BUCKy: Bayesian concordance analysis

Ané et al (2007), Larget et al (2010)

Summarize the sampled sets of gene trees by the clades' **concordance factors** (CF):

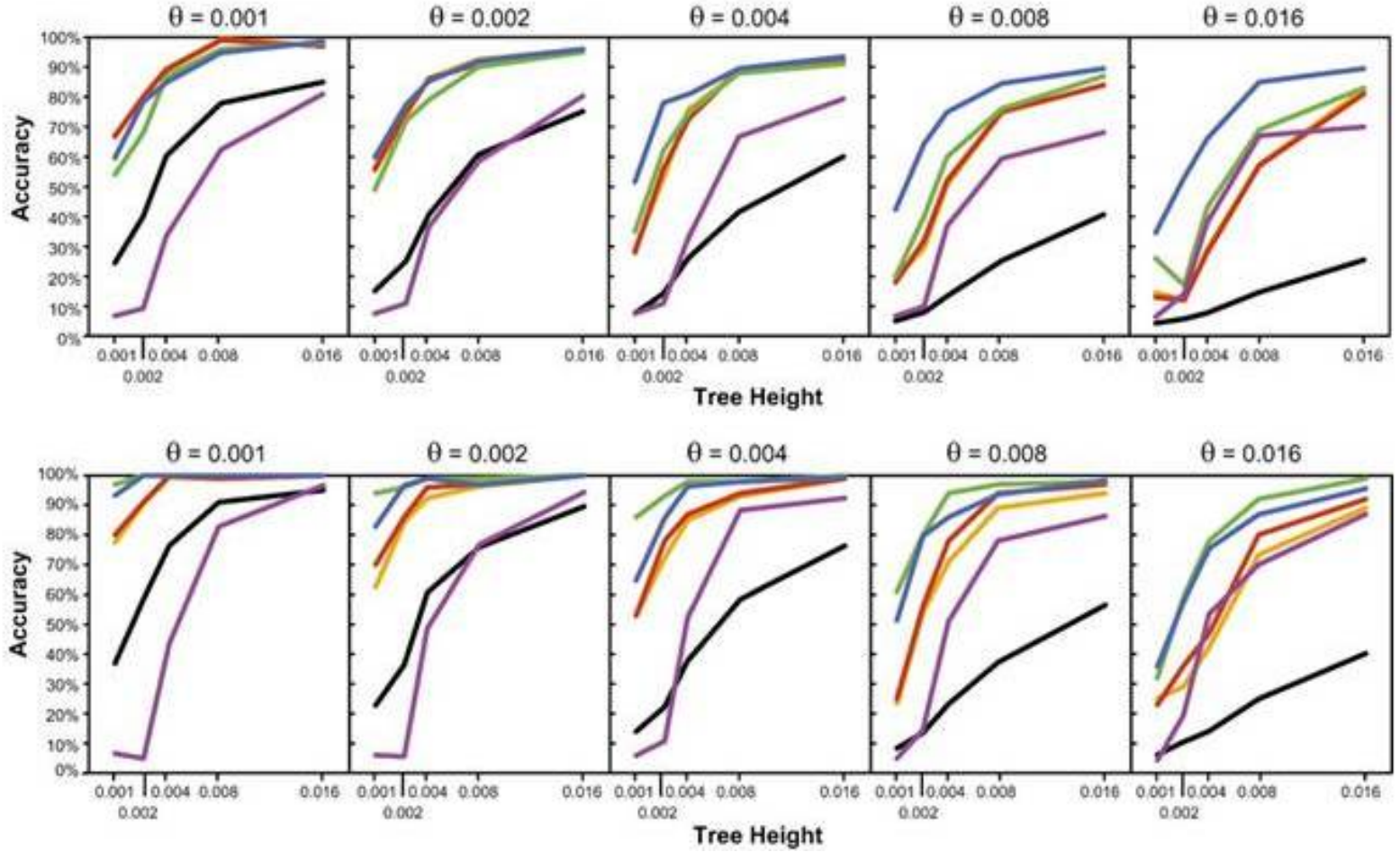
$$\begin{aligned}\text{CF of clade } \mathcal{C} &= \text{proportion of genes having } \mathcal{C} \text{ in their trees} \\ &= \text{genomic support for clade } \mathcal{C}\end{aligned}$$

- Concordance tree: from clades with highest concordance factors, like greedy consensus.
- Population tree: from quartets with highest concordance factors. Quartet $ab|cd$ favored if it has higher CF than $ac|bd$ and $ad|bd$. Then supertree algorigthm to get tree on full taxon set.

How well is the species tree estimated?

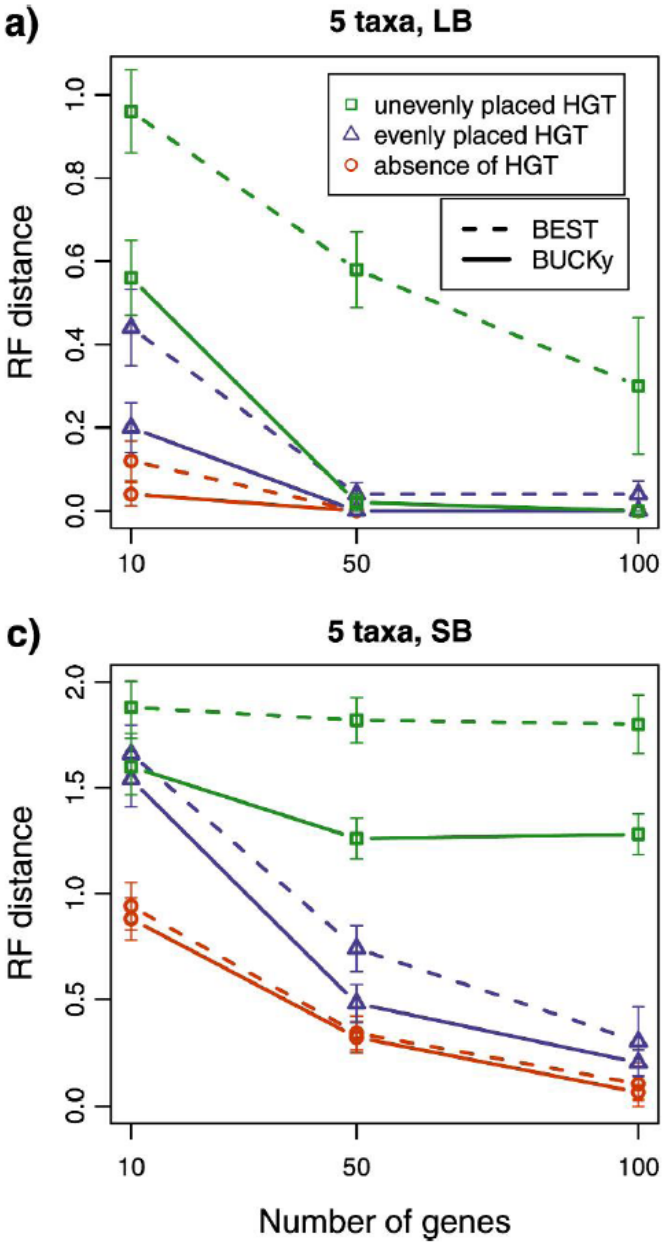
Leaché & Rannala(2011)

Asym/symmetric trees (top/bottom), under coalescent, no rate variation:

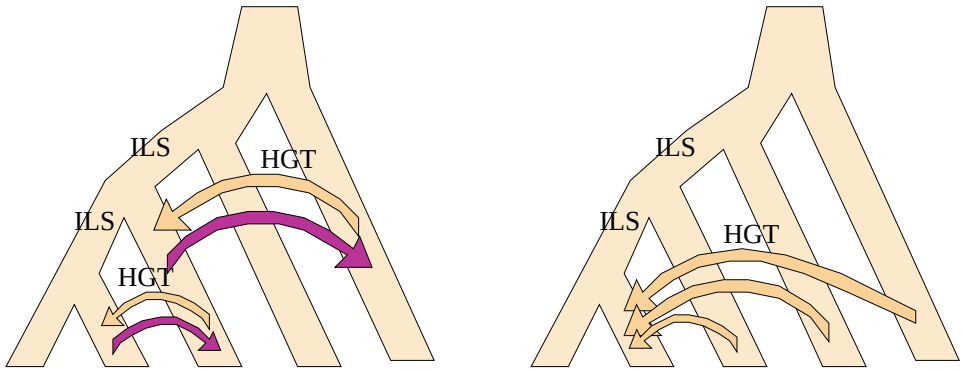


How well is the species tree estimated?

Chung & Ané (2010)

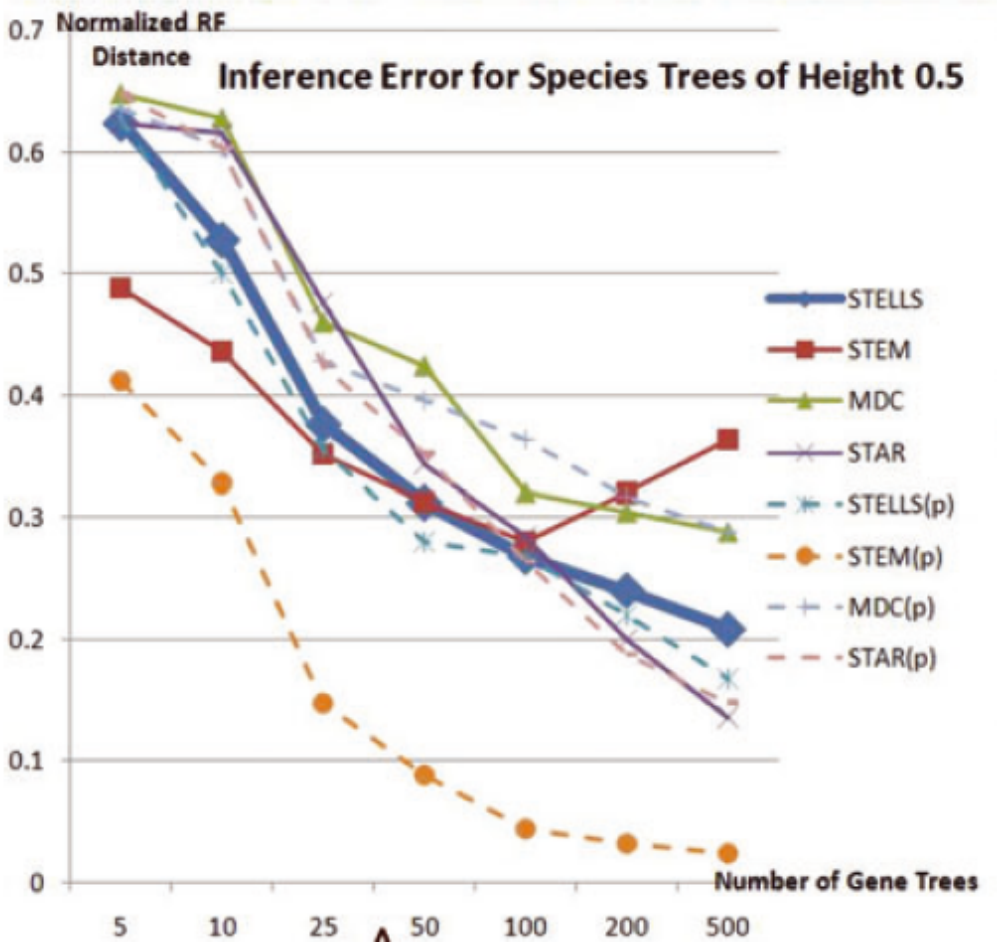


under the coalescent + HGT (gene flow).
HGT **absent**, **evenly**, or **unevenly** distributed on the tree.



How well is the species tree estimated?

Wu (2011)



Under the coalescent. From

- true gene trees (- - -), or
- estimated gene trees (—)

Other simulation studies under the Coalescent

***BEAST more accurate than BEST**

might be a result of faster mixing. Under no rate variation.

Heled & Drummond (2010)

STEM tends to be more accurate than MDC

Various sampling efforts. Under no rate variation.

McCormack, Huang & Knowles (2009)

BEST $>$ STAR $\approx$ NJst

Liu & Yu (2011)

STAR $>$ STEAC $>$ GLASS

especially with lineage-specific rate variation. Huge effect of using inferred gene trees rather than true gene trees. Liu et al. (2009)

STAR $\approx$ MP-EST $>$ RT consensus

also did well under symmetric gene flow. Liu, Yu & Edwards (2010)

BUCKy $>$ greedy consensus

or other fast methods (but computational cost). on 17- or 100-taxa.

Yang & Warnow (2011)

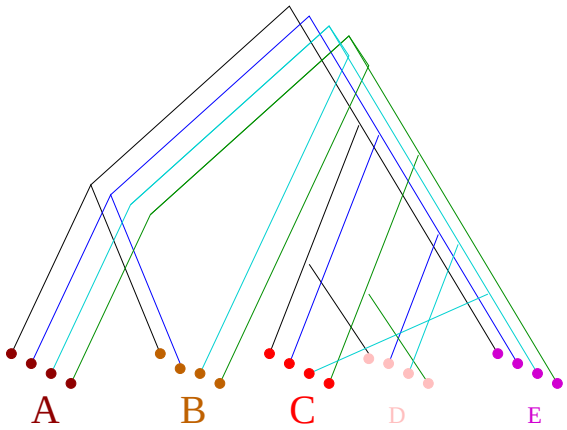
Method comparisons

1. Computationally demanding methods are more accurate:
   no free lunch!
2. Methods accounting for gene tree uncertainty are more accurate
3. Methods not using branch lengths are robust to
   rate variation: between genes and between lineages
   symmetric gene flow
4. My own favorite:
   NJst for computational speed,
   *BEAST for accuracy under the coalescent,
   BUCKy for accuracy and robustness.

1. **Gene Tree / Species Tree** methods and the coalescent

2. **Concordance** analysis, and testing the coalescent model

3. **Polyploids**: gene tree discordance and reticulate evolution



The concept of Concordance trees

Primary **Concordance tree** displays:

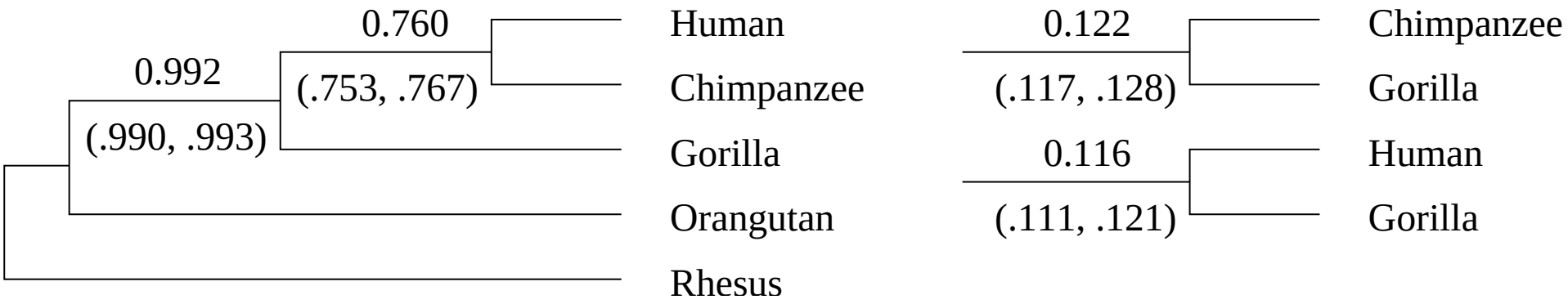
- clades with highest concordance factors, built using greedy consensus
- clades with highest genomic support
- dominant vertical inheritance pattern in the actual history of present-day genes, regardless of how this signal came to be.

Example: 76% of our genes have a human-chimp clade, 24% don't, regardless of the historical process.

The concordance tree **summarizes the phylome**: displays the relationships with highest genomic support.

Concordance analysis of 30,040 gene fragments

Ané (2010), data from Ebersberger et al (2007)



Genomic support for Human-Chimp clade: 76%,
versus 12.2%, 11.6% for conflicting clades

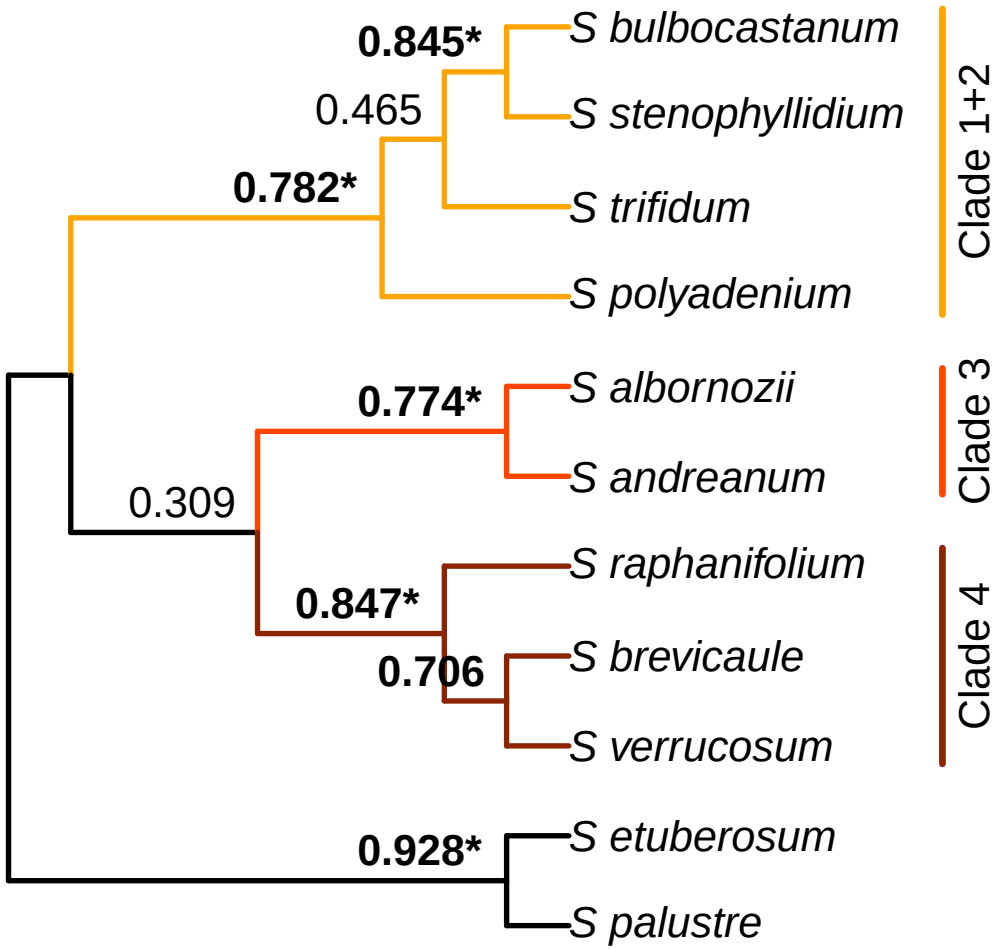
Statistical support for Human-Chimp clade in species tree: 1.0 PP that $>$ 50% genes have this clade.

- 1.0 posterior prob. that tree on the left = species tree
- 1.0 posterior prob. that there is true conflict among gene trees.

4 days * 3 CPUs (MrBayes) + 5h (BUCKy).

Concordance tree: dominant vertical signal

Rodriguez et al (2009)



wild potatoes, 12 nuclear markers

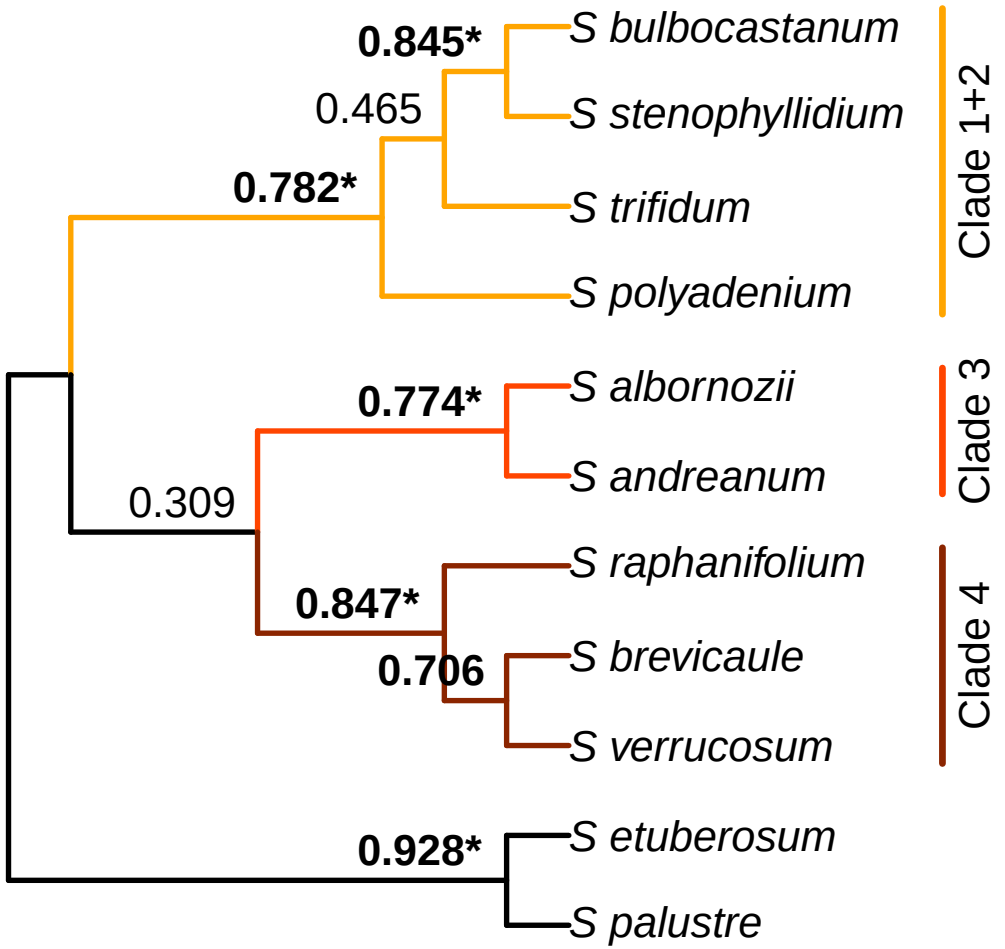
values = concordance factors

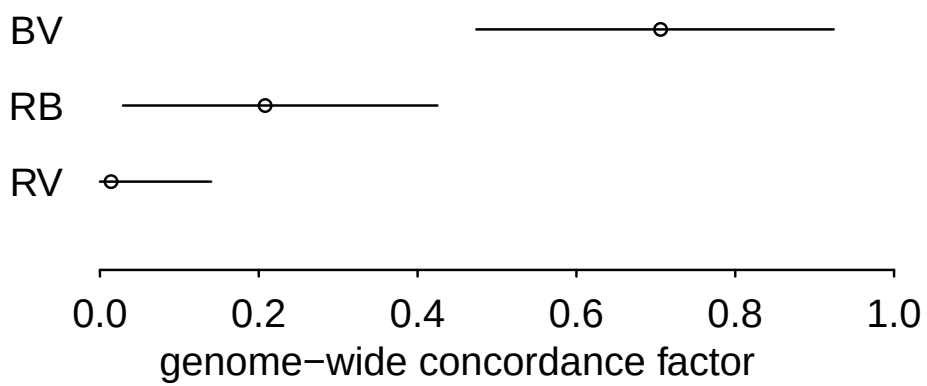
***** = CF>.50 with >.99 credibility

bold = in concordance tree with >.99 credibility

Statistical support for the Genomic support

Compare 99% credibility intervals of CFs on alternative resolutions

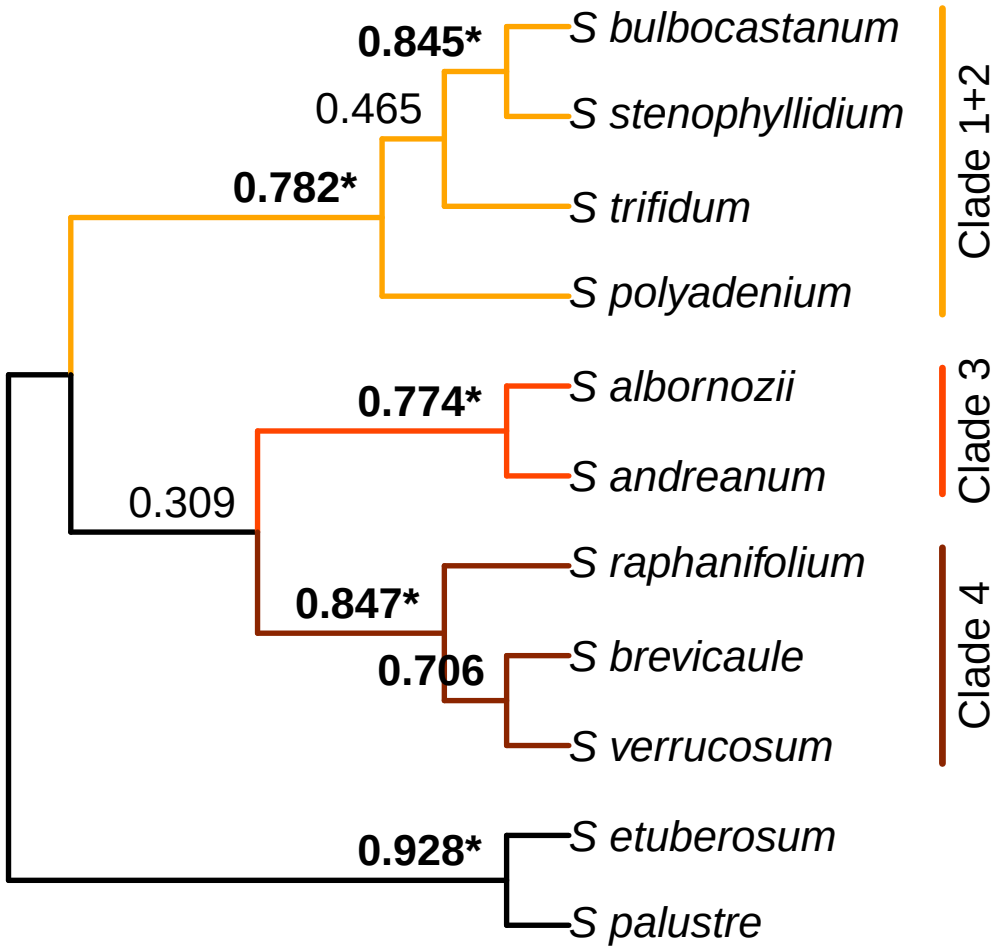


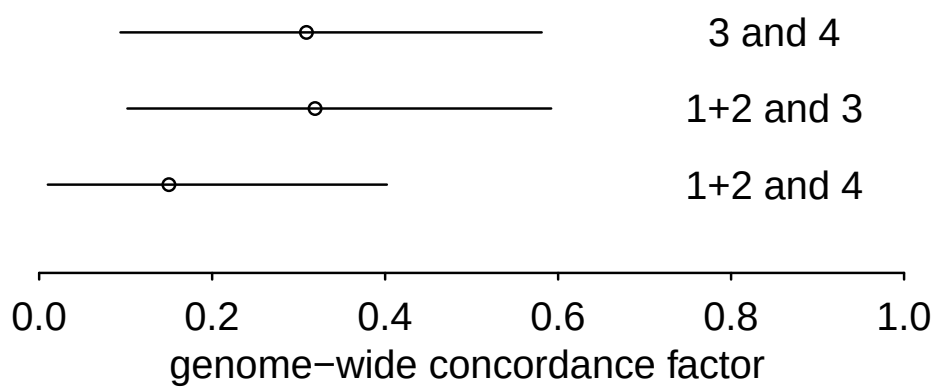


S. brevicaule - *S. verrucosum* in the concordance tree with high PP

Statistical support for the Genomic support

Compare 99% credibility intervals of CFs on alternative resolutions



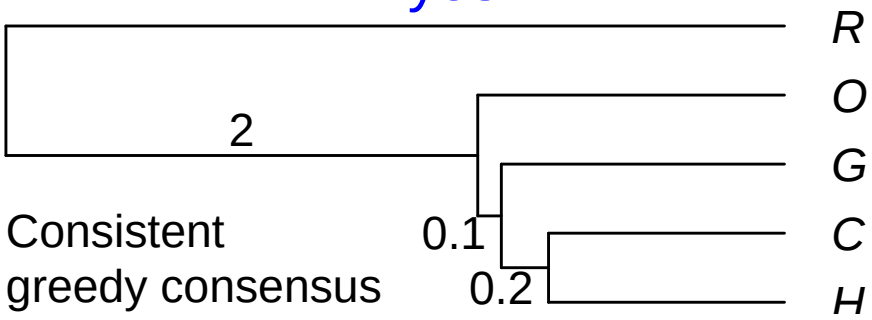


No evidence that clades 3 and 4 are sister in the concordance tree.

Concordance tree vs. Species tree, under coalescent

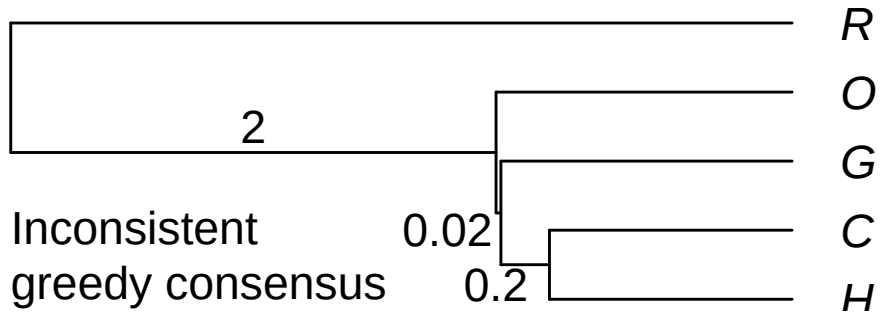
Concordance tree $\stackrel{?}{\neq}$ Species tree

yes



Consistent
greedy consensus

no



Inconsistent
greedy consensus

Concordance tree has (GO)

Splits, Full Taxon Set	Tree 1 (%)	Tree 2 (%)
HC|GOR	**38.7**	**36.8**
HCG|OR	**25.1**	20.7
GO|HCR	21.2	**23.6**
HG|COR	20.5	18.7
CG|HOR	20.5	18.7

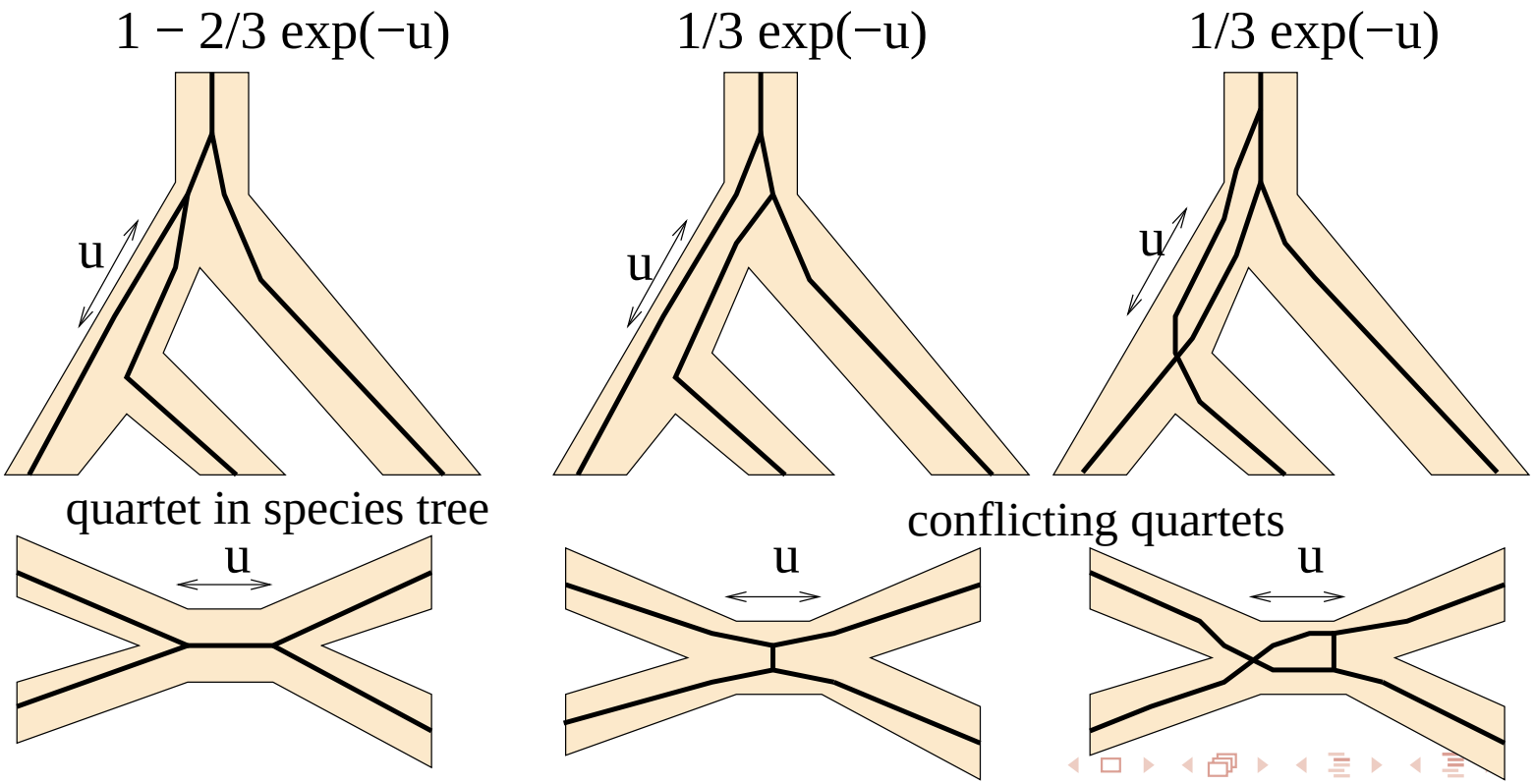
Population tree in BUCKy

Larget et al. (2010)

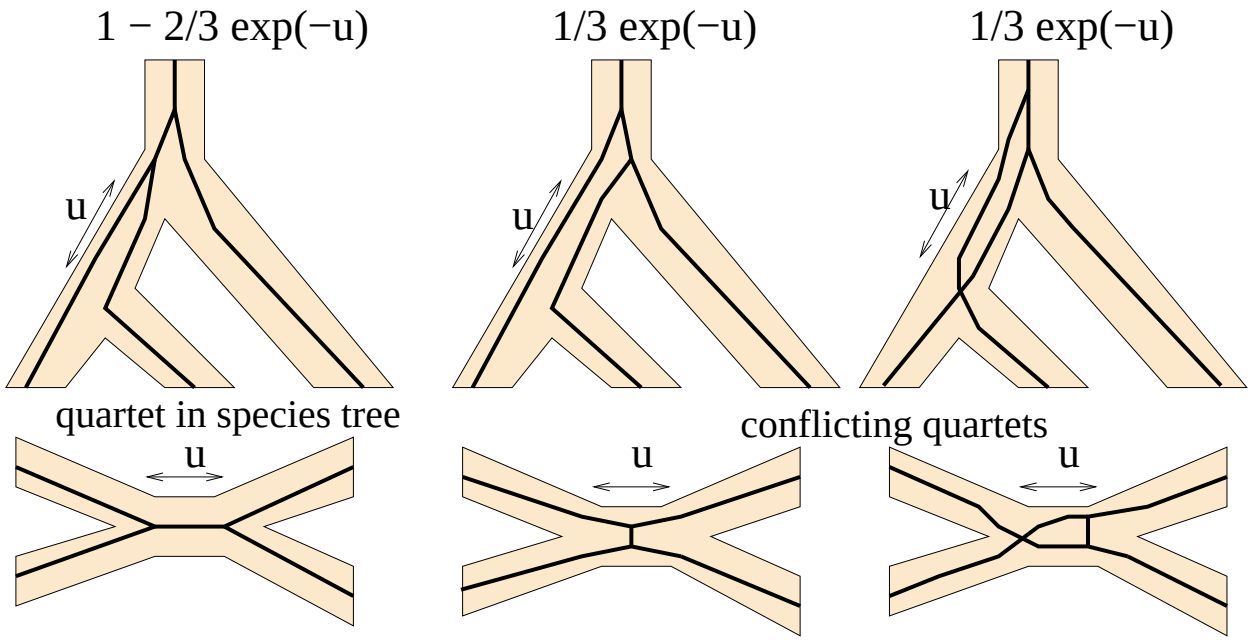
2 tree estimates: Population tree alongside concordance tree.

1. estimate **concordance factors** as before
2. for 4 taxa a, b, c and d, favor the quartet $ab|cd$ with maximum concordance factor $CF_{ab|cd}$

Expected CF under the coalescent:



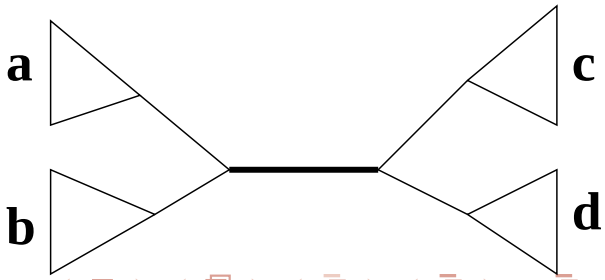
Population tree in BUCKy



3. build species tree from these **favored quartets**.
4. estimate branch lengths in species tree in coalescent units:

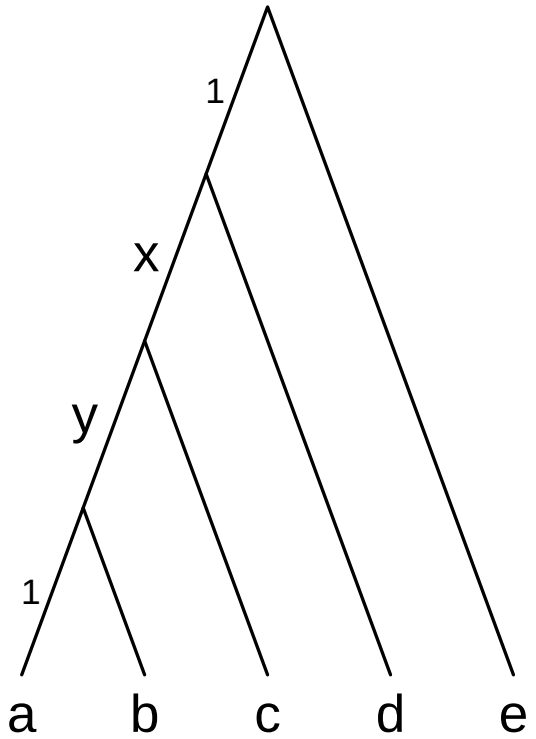
$$\hat{u} = \log(3/2 * (1 - \widehat{\text{CF}}))$$

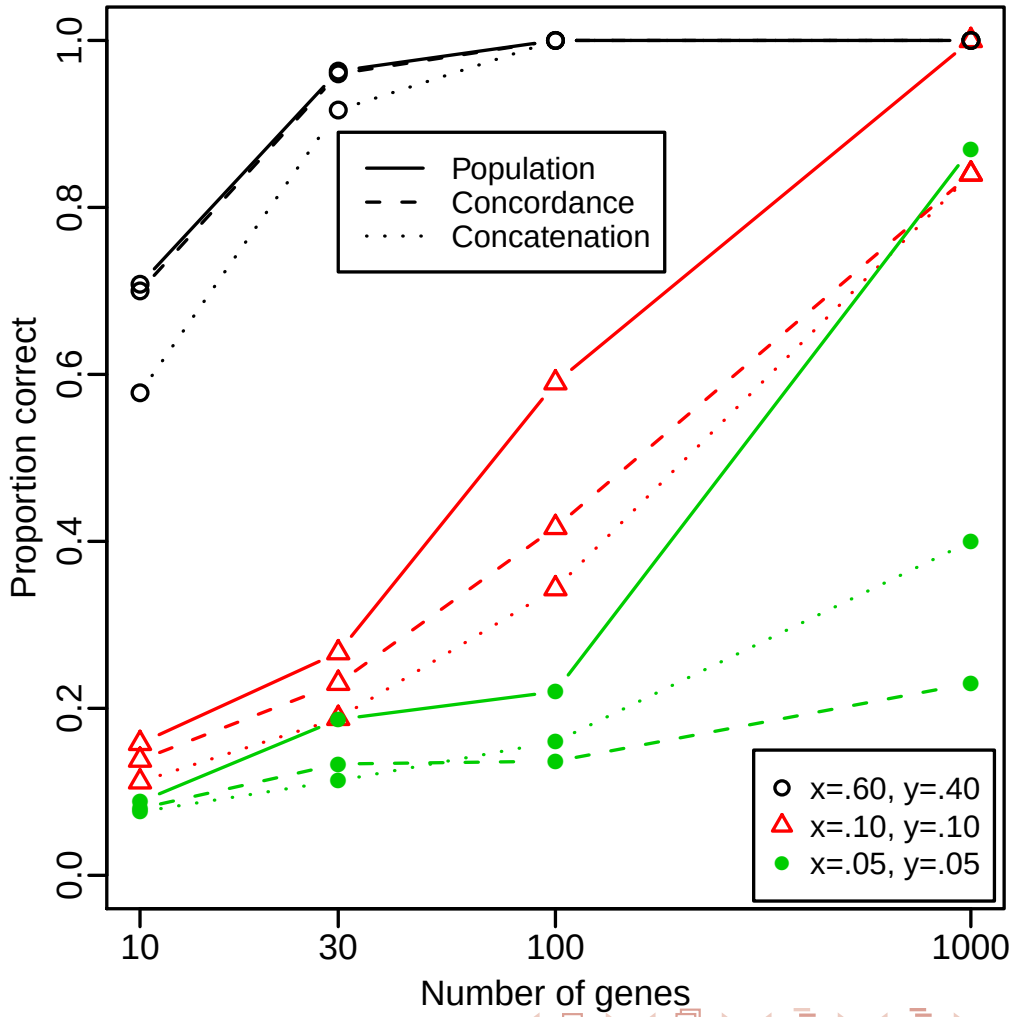
$\widehat{\text{CF}}$ = average CF across all quartets defining the branch.



How well is the species tree estimated?

Coalescent simulations, x, y = branch lengths in true species tree.
Discordance level: **medium**, high, very high





What biological **process** caused the conflict?

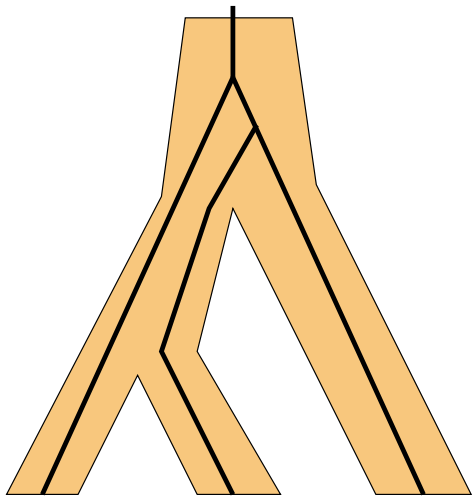
- estimate gene trees with minimal assumptions

  species tree = **vertical** signal.

- test null hypothesis of the coalescent, as sole explanation for gene tree discordance.

  Information about the **process** is in the **horizontal** signal.

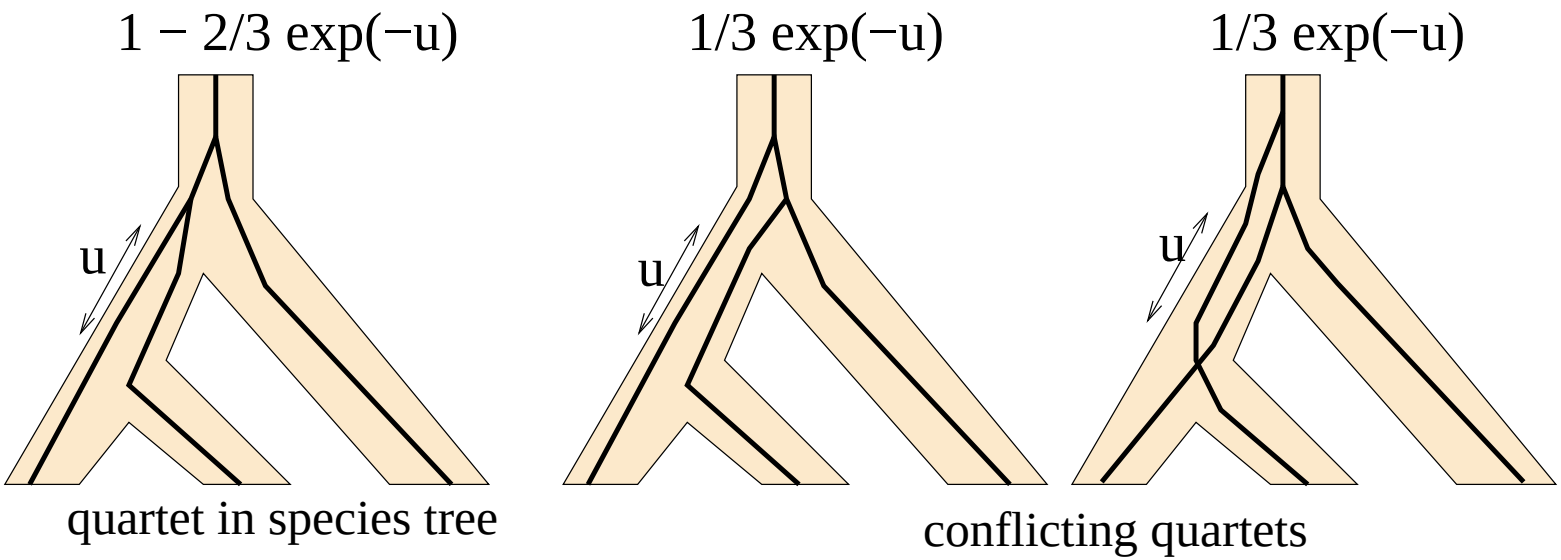
  also in branch lengths, but confounding with rate variation.



Incomplete lineage sorting

Testing the adequacy of the coalescent

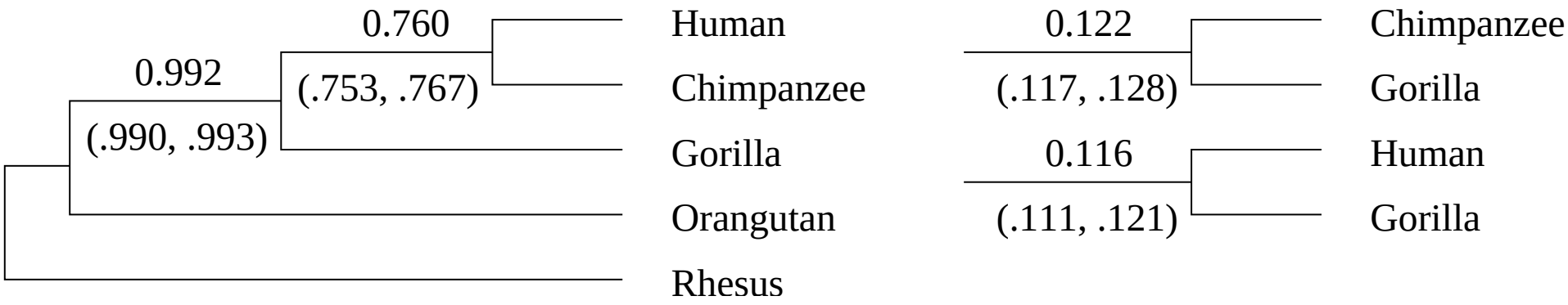
using the symmetric signature of the coalescent on genomic support.
Expected concordance factors (proportion of genes):

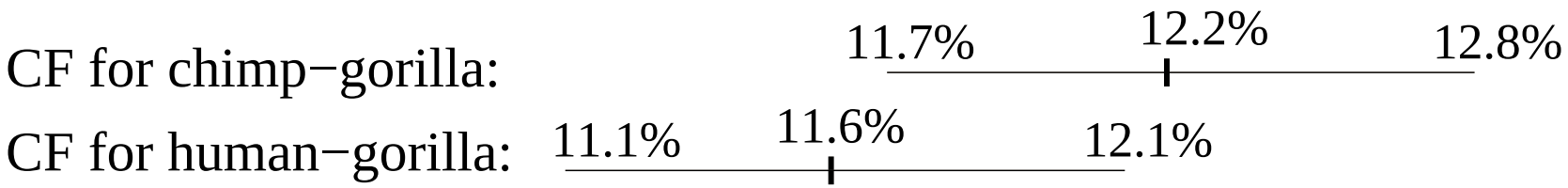


$u =$ internal branch length in species tree, coalescent units:
generations $/2 * N_e$.

Testing the adequacy of the coalescent

Along a given branch: test equality of the concordance factors for the two minor resolutions.



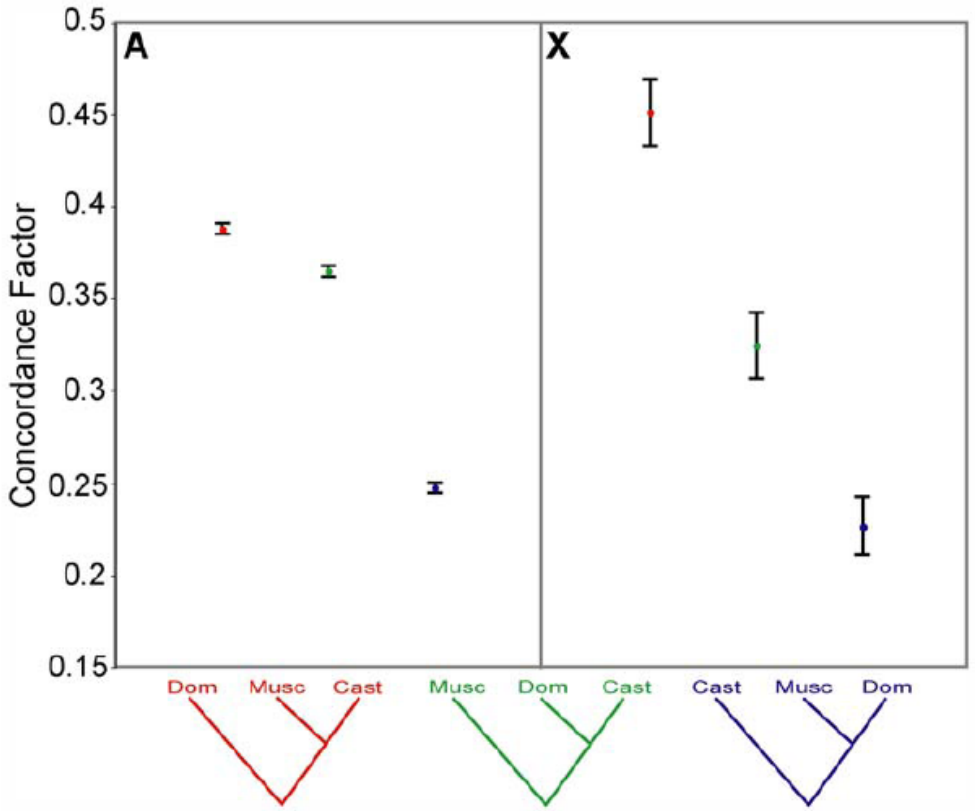


The 95% credibility interval **overlap**: Accept the coalescent model as a sufficient explanation for the discordance.

Discordance across the House Mouse genome

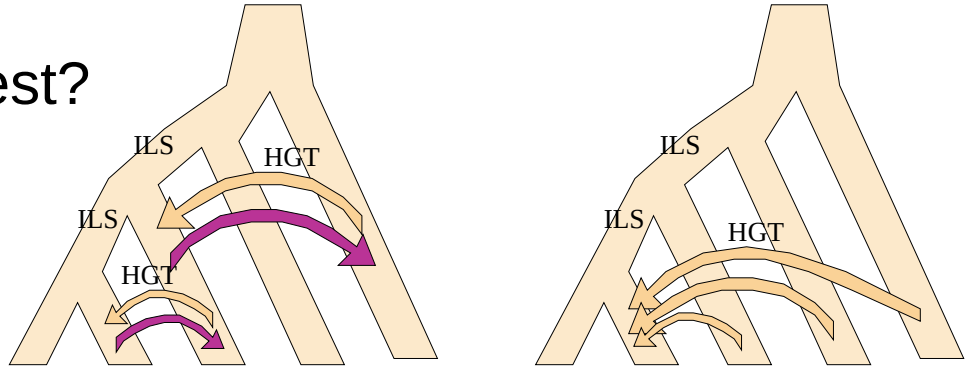
White et al (2009)

- *Mus musculus*, *M. castaneus*, *M. domesticus* and rat. Whole genome-alignments: complete genome + perlegen SNP data. X chromosome + 19 autosomes: 1.8 billion sites.
- Loci first defined with MDL (parsimony-based), then BUCKy.



Testing the adequacy of the coalescent

How reliable and powerful is this test?



- Coalescent simulations
- HGT: absent, **evenly** or **unevenly** distributed on the tree.

Proportion of times the coalescent-only hypothesis was rejected:

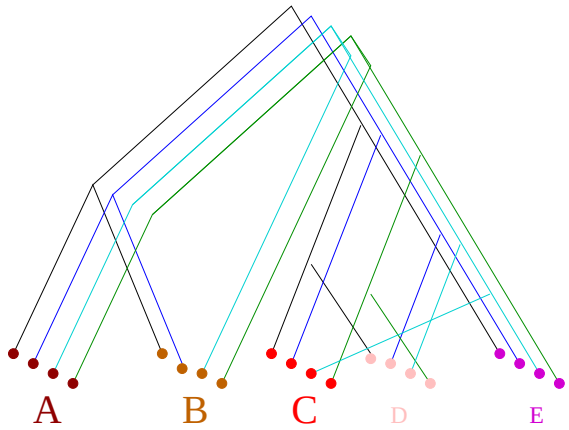
	No HGT		Evenly dist. HGT		Unevenly HGT	
ILS	weak	strong	weak	strong	weak	strong
# genes						
10	0	0.01	0.46	0.15	0.80	0.12
50	0.02	0.04	0.52	0.18	1.0	0.54
100	0	0.07	0.43	0.15	1.0	0.58

Chung & Ané (2010)

1. **Gene Tree / Species Tree** methods and the coalescent

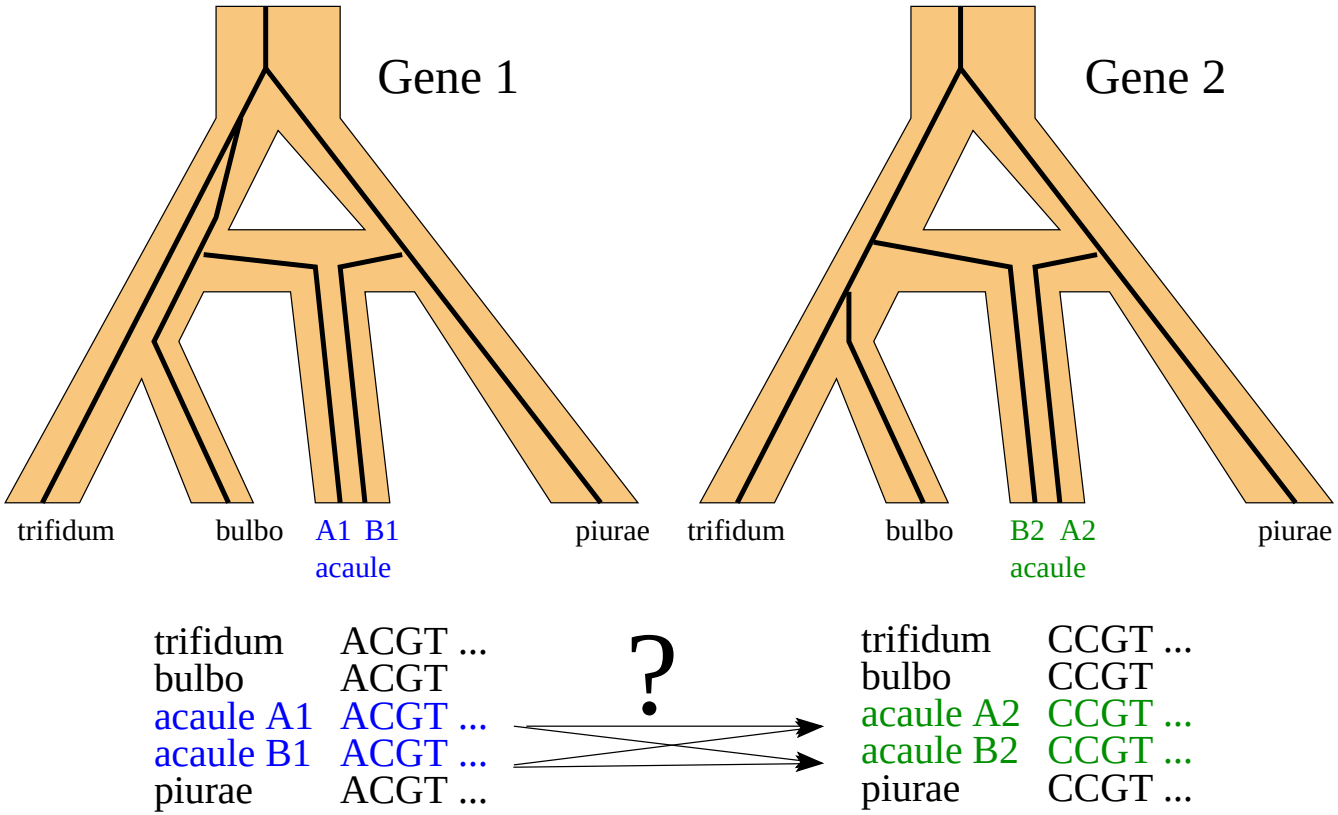
2. **Concordance** analysis, and testing the coalescent model

3. **Polyploids**: gene tree discordance and reticulate evolution



The challenge of polyploid species

Their history is **not reticulate**: cannot be represented by a tree.
Cannot concatenate alignments from multiple genes



Conflict among gene trees in wild potatoes

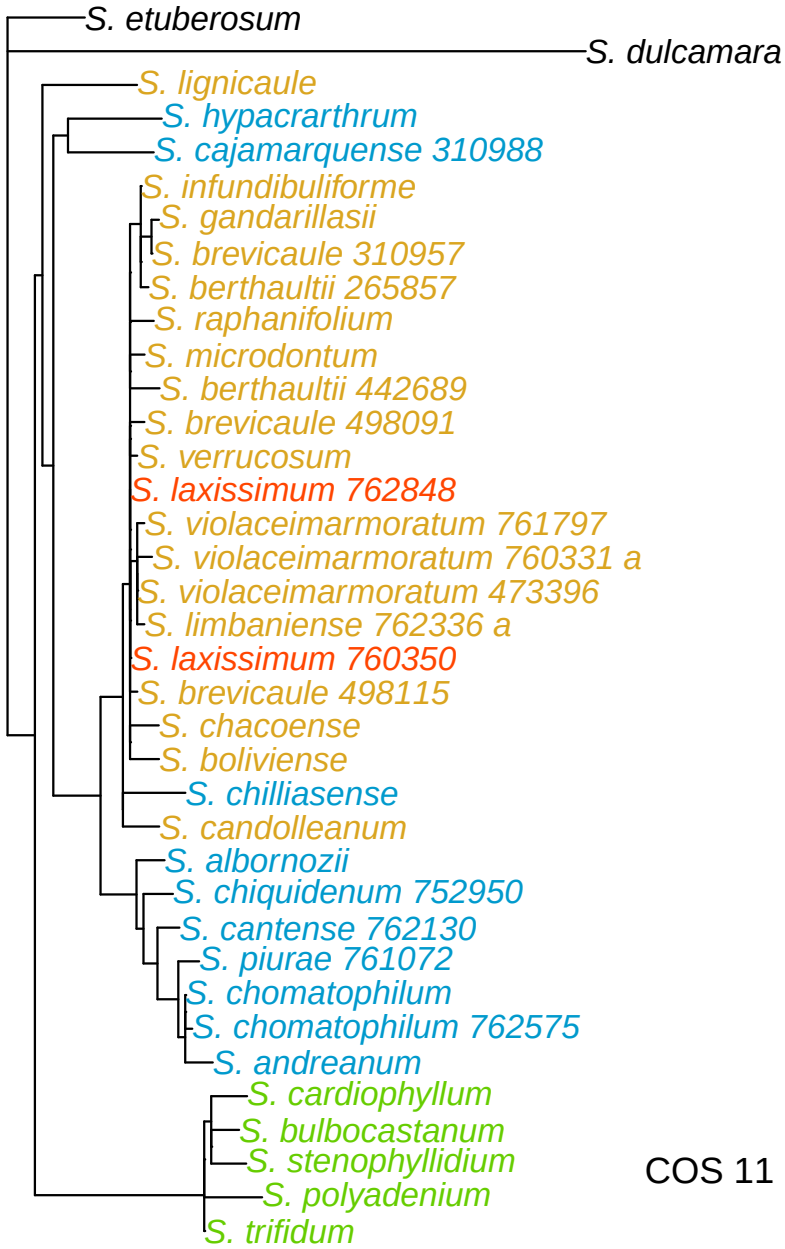
Species hard to delimit:

232 in section *petota* by Hawkes (1990)
190 by Spooner & Salas (2006)

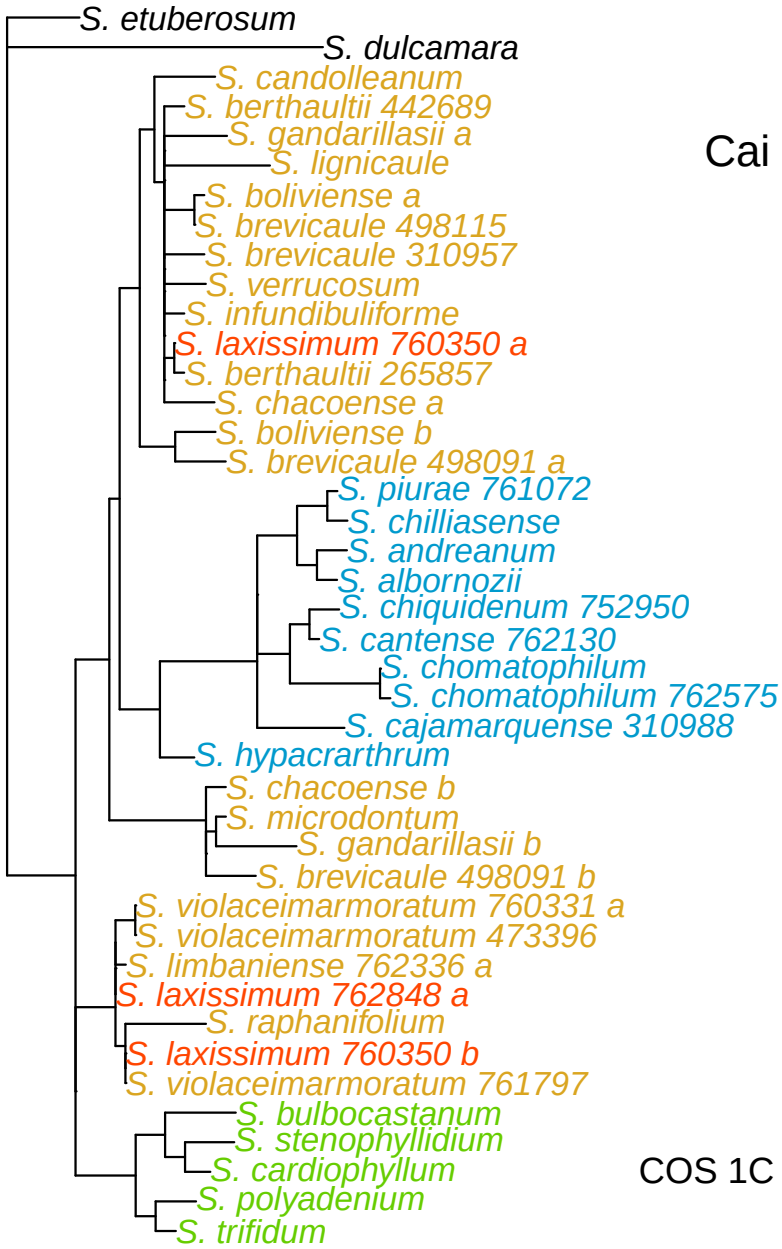
Plastid DNA: 4 clades, little relationship to Hawkes's series ('92, '97)
Nuclear DNA: 3 main clades, but much discordance among loci ('08, '09)

This is just about diploids. $\sim$30% species are polyploids.

Next: 2 nuclear markers, wild potatoes, 1 polyploid species

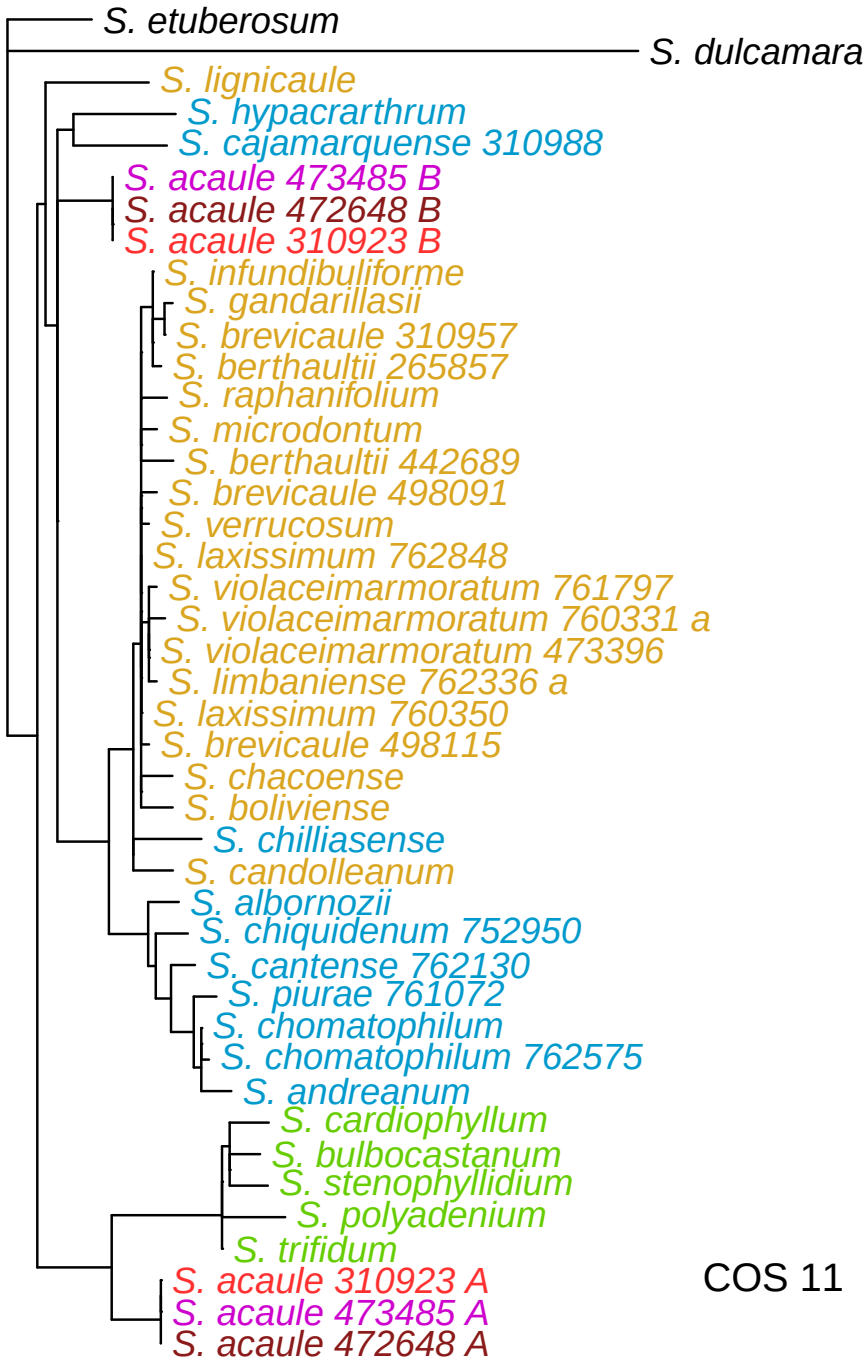


COS 11

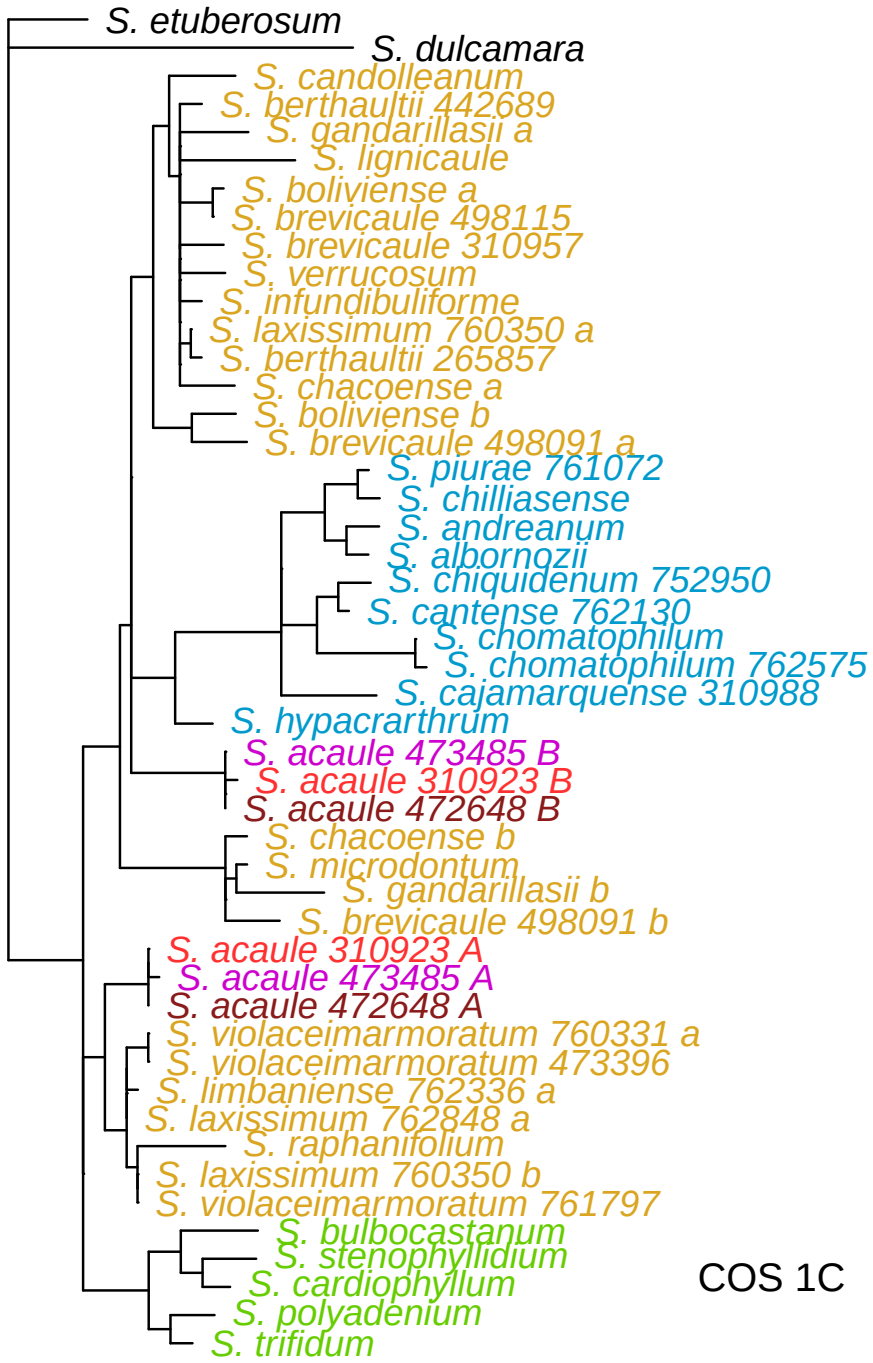


Cai et al (2012)

COS 1C


S. etuberosum
S. dulcamara
S. lignicaule
S. hypacrarthrum
S. cajamarquense 310988
S. acaule 473485 B
S. acaule 472648 B
S. acaule 310923 B
S. infundibuliforme
S. gandarillasii
S. brevicaule 310957
S. berthaultii 265857
S. raphanifolium
S. microdontum
S. berthaultii 442689
S. brevicaule 498091
S. verrucosum
S. laxissimum 762848
S. violaceimarmoratum 761797
S. violaceimarmoratum 760331 a
S. violaceimarmoratum 473396
S. limbaniense 762336 a
S. laxissimum 760350
S. brevicaule 498115
S. chacoense
S. boliviense
S. chilliasense
S. candolleanum
S. albornozii
S. chiquidenum 752950
S. cantense 762130
S. piurae 761072
S. chomatophilum
S. chomatophilum 762575
S. andreanum
S. cardiophyllum
S. bulbocastanum
S. stenophyllidium
S. polyadenium
S. trifidum
S. acaule 310923 A
S. acaule 473485 A
S. acaule 472648 A

COS 11


S. etuberosum
S. dulcamara
S. candolleanum
S. berthaultii 442689
S. gandarillasii a
S. lignicaule
S. boliviense a
S. brevicaule 498115
S. brevicaule 310957
S. verrucosum
S. infundibuliforme
S. laxissimum 760350 a
S. berthaultii 265857
S. chacoense a
S. boliviense b
S. brevicaule 498091 a
S. piurae 761072
S. chilliasense
S. andreanum
S. albornozii
S. chiquidenum 752950
S. cantense 762130
S. chomatophilum
S. chomatophilum 762575
S. cajamarquense 310988
S. hypacrarthrum
S. acaule 473485 B
S. acaule 310923 B
S. acaule 472648 B
S. chacoense b
S. microdontum
S. gandarillasii b
S. brevicaule 498091 b
S. acaule 310923 A
S. acaule 473485 A
S. acaule 472648 A
S. violaceimarmoratum 760331 a
S. violaceimarmoratum 473396
S. limbaniense 762336 a
S. laxissimum 762848 a
S. raphanifolium
S. laxissimum 760350 b
S. violaceimarmoratum 761797
S. bulbocastanum
S. stenophyllidium
S. cardiophyllum
S. polyadenium
S. trifidum

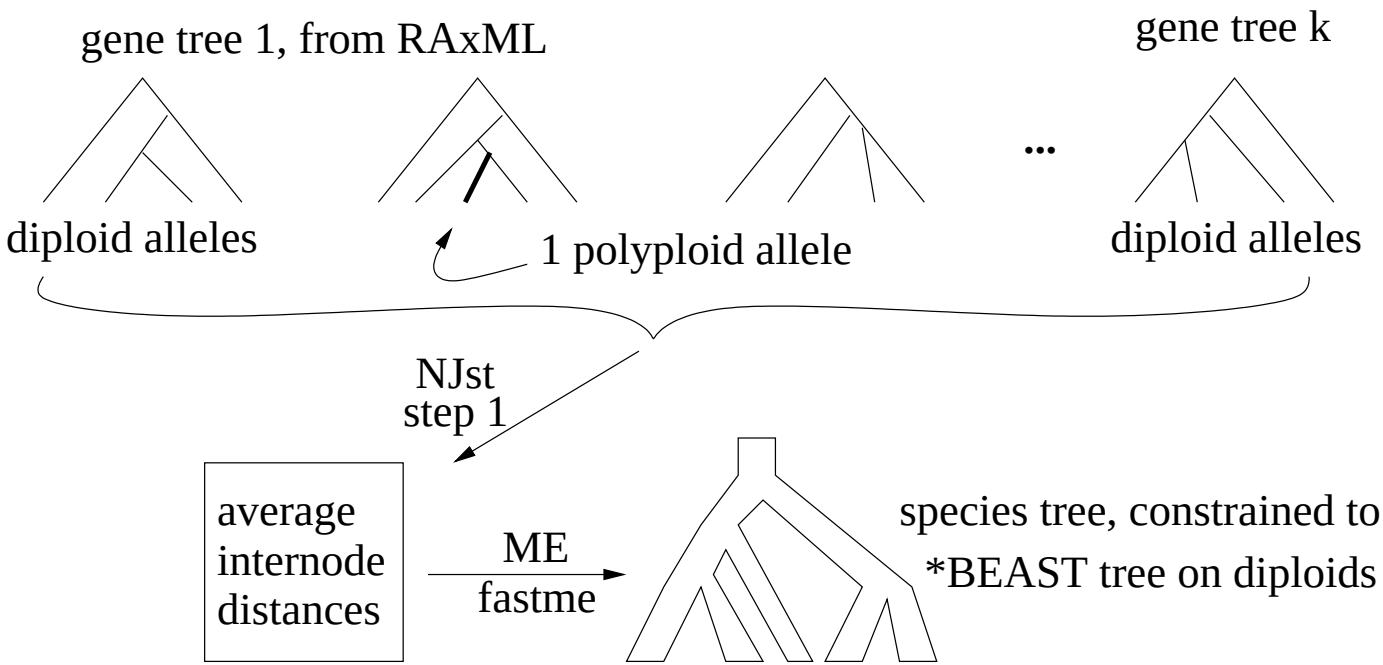
COS 1C

Strategy for polyploid species with tree discordance

Cai et al (2012)

1. Estimate a robust species tree for the diploids:
   *BEAST
2. Estimate individual gene trees with polyploids:
   RAxML
3. Estimate the placement of each polyploid allele: 11 polyploids, 54 accessions, 6 nuclear markers, 823 alleles total.
   NJst species tree method, modified to use the constraint of the backbone diploid species tree.
4. Repeat step 3 100 times, on bootstrap trees from RAxML pipeline to do this automatically 823*100 times.
5. Summarize the placements that receive high bootstrap support.

Step 3: placement of one polyploid allele



"MEst": modified NJst to use the Minimum Evolution criterion instead of neighbor-joining. Estimated length of candidate species tree S, based on observed distances:

$$\hat{\ell}(T) = \sum_{\text{taxa } i,j} 2^{-n_{ij}(S)} d_{\text{internode}}(i,j)$$

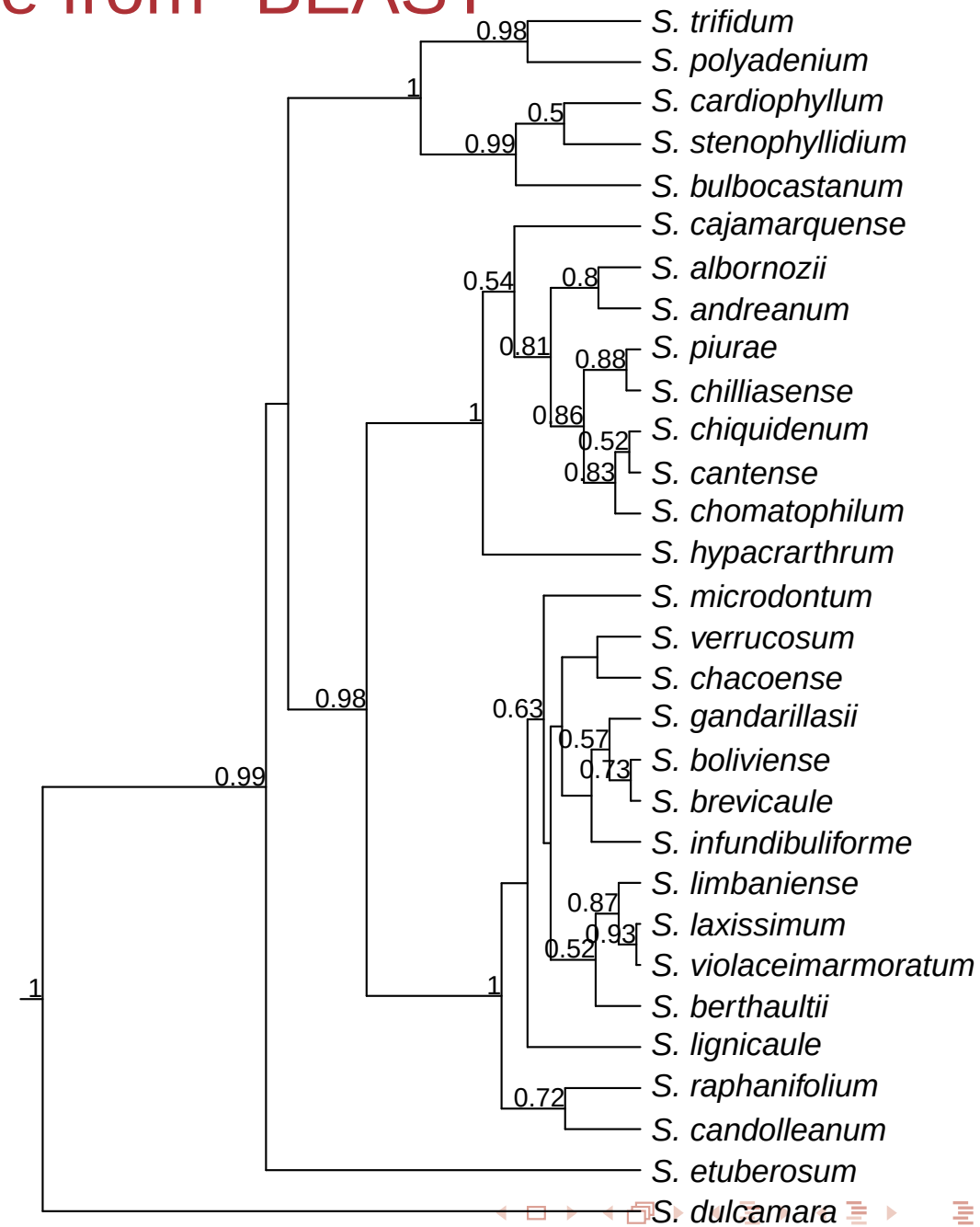
Gascuel & Steel (2006), Desper & Gascuel (2004)

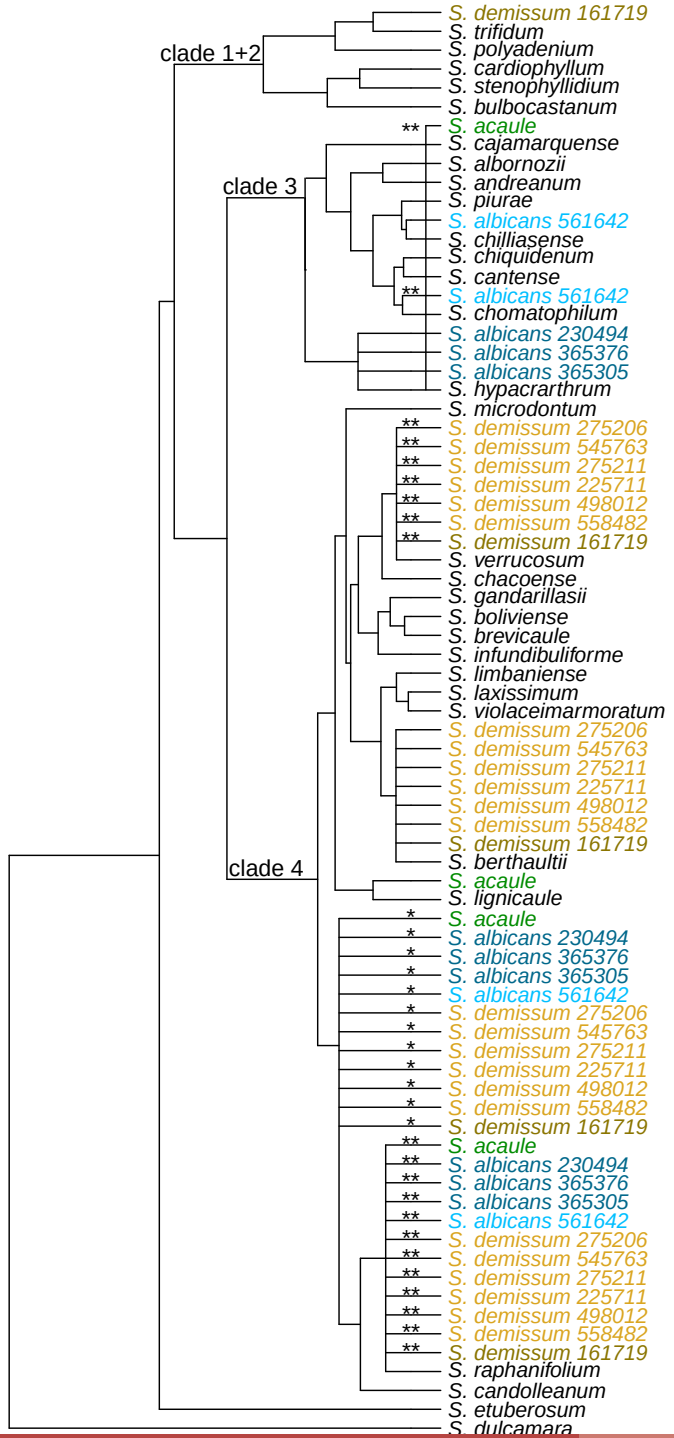
Strategy for polyploid species with tree discordance

Cai et al (2012)

1. Estimate a robust species tree for the diploids:
   *BEAST
2. Estimate individual gene trees with polyploids:
   RAxML
3. Estimate the placement of each polyploid allele: 11 polyploids, 54 accessions, 6 nuclear markers, 823 alleles total.
   NJst species tree method, modified to use the constraint of the backbone diploid species tree.
4. Repeat step 3 100 times, on bootstrap trees from RAxML pipeline to do this automatically 823*100 times.
5. Summarize the placements that receive high bootstrap support.

Diploid backbone tree from *BEAST





Polyploids: Acaulia group

black = diploids

S. albicans: hexaploid, *S. acaule* (tetraploid) as putative parent.
1st time clade 3 genome documented for *S. albicans*.

1 accession (of 4) shows separate origin in clade 3 –different geographic location, was recognized as *S. acaule* subsp. *palmirense* by Kardolus (1998).

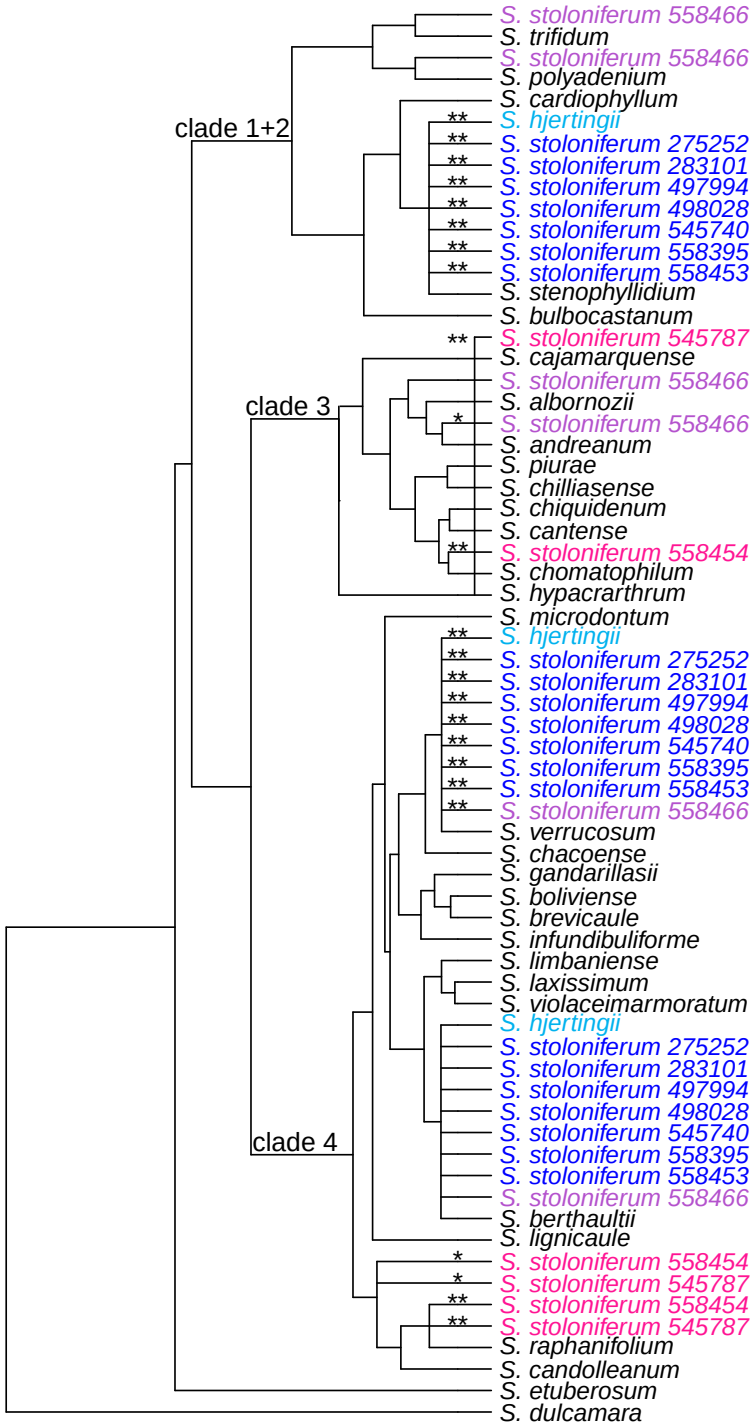
S. demissum: hexaploid. 1 accession has alleles in clade "1+2", no others do.

Possible explanations:

- Multiple origins of polyploid species
- Introgressive hybridization subsequent to speciation
- Loss of alleles

Evidence for:

- more precise phylogenetic placement of polyploid origins
- variation between accessions (species definition?)
- variation between loci



Longipedicellata group

S. stoloniferum: tetraploid

7 accessions with similar origins.

3 accessions: show alleles in clade 3 (no others do), and different origins in clade 1+2 and clade 4.

Conclusions

Polyploidy adds more challenges to gene tree discordance:

reticulate non-tree like history
sometimes hard to match alleles from different loci
new tools are needed
I showed a strategy using currently available methods.

Collaborators

Bret Larget (Statistics & Botany)

David Baum (Botany)

Colin Dewey (Comp.Sc & Biostatistics)

Satish Kotha (Computer Sc.)

Yujin Chung (Statistics)
Bret Payseur (Genetics)
Mike White (Genetics)
David Spooner (USDA, Horticulture)
Flor Rodriguez, Danying Cai (Horticulture)