MBL workshop on molecular evolution

network models

Cécile Ané

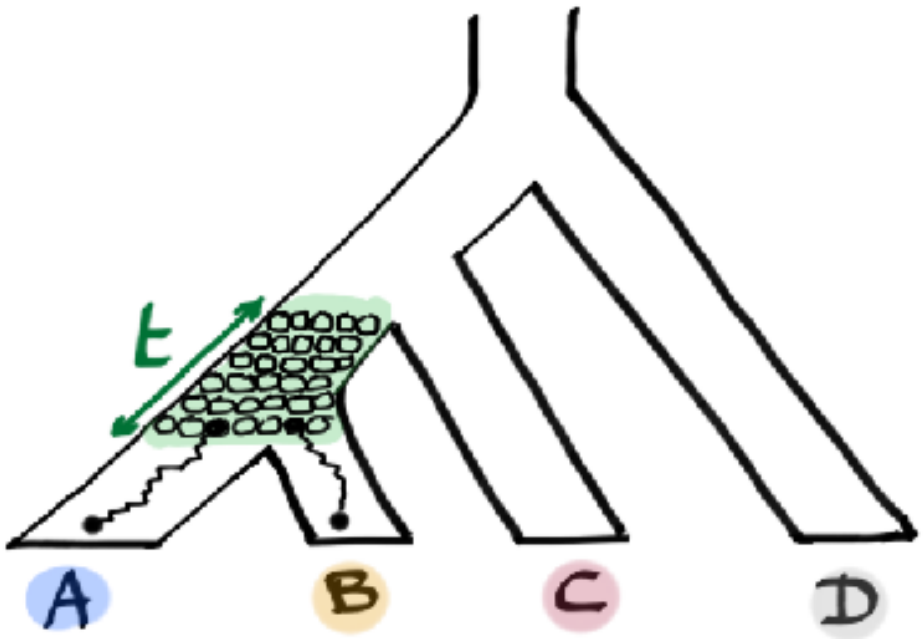
UW - Madison, Departments of Statistics and of Botany

MBL 2018

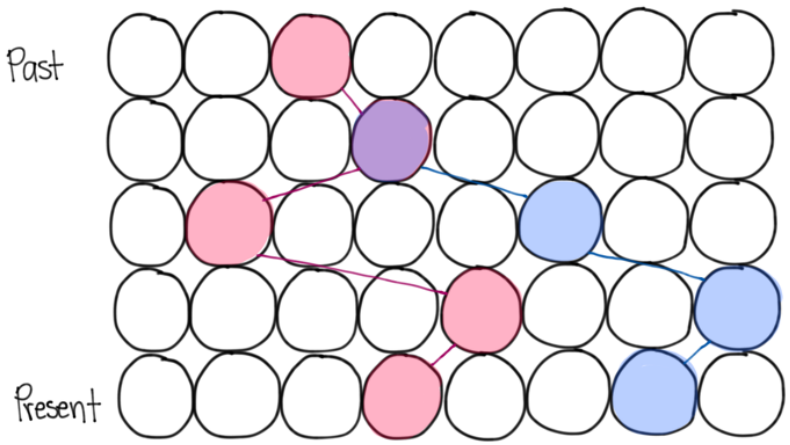
1. impact of introgression / hybridization?
2. is a tree sufficient, or do we need a network?
3. network models

can we ignore incomplete lineage sorting?

Does **incomplete lineage sorting** impact tree reconstruction?

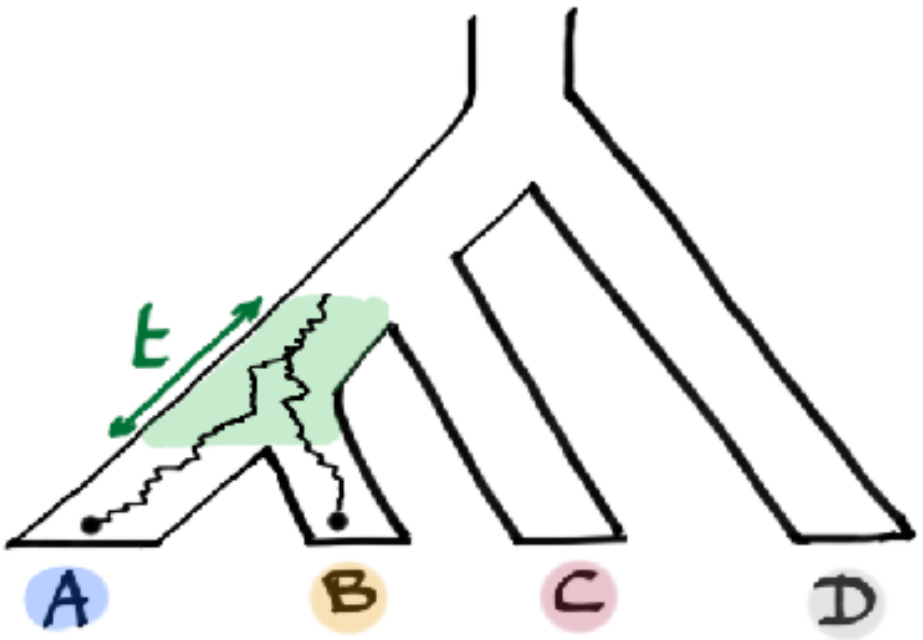


within 1 population:

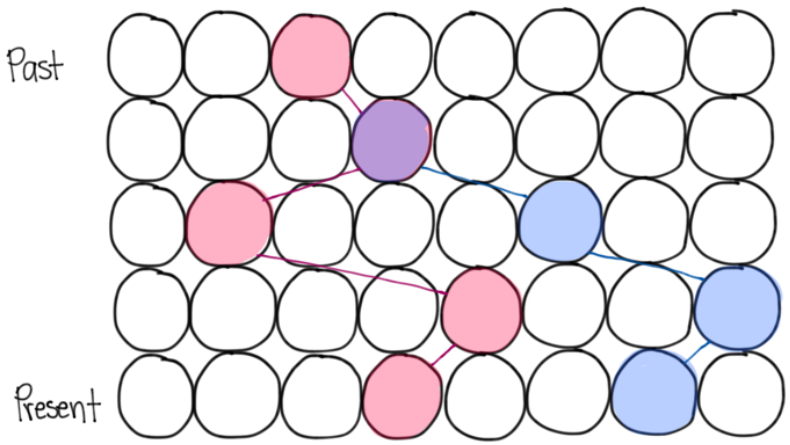


can we ignore incomplete lineage sorting?

Does **incomplete lineage sorting** impact tree reconstruction?

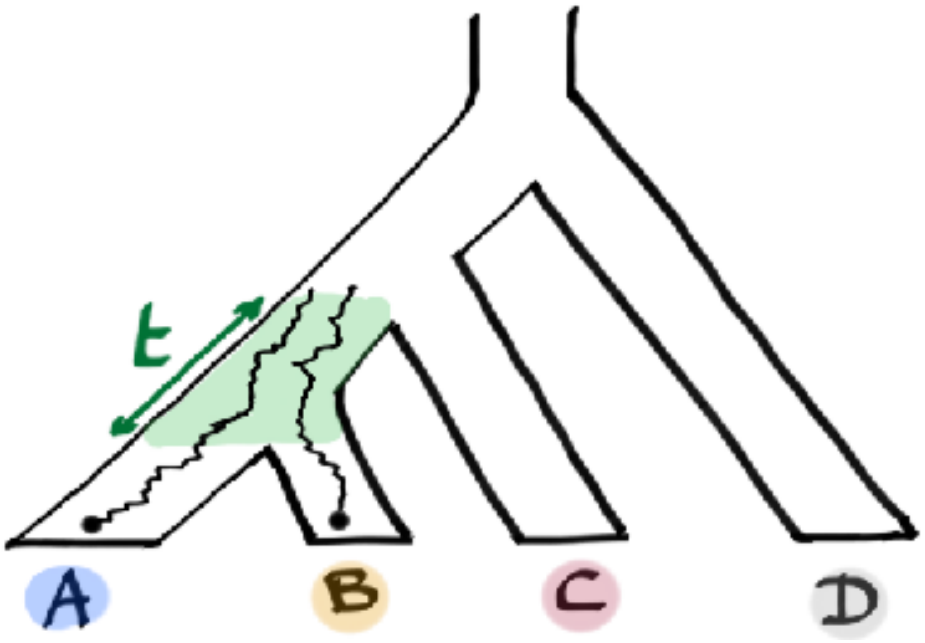


within 1 population:

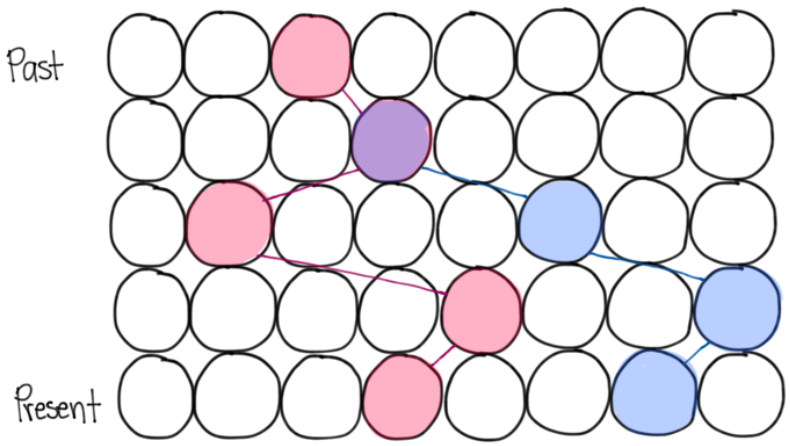


can we ignore incomplete lineage sorting?

Does **incomplete lineage sorting** impact tree reconstruction?



within 1 population:



can we ignore incomplete lineage sorting?

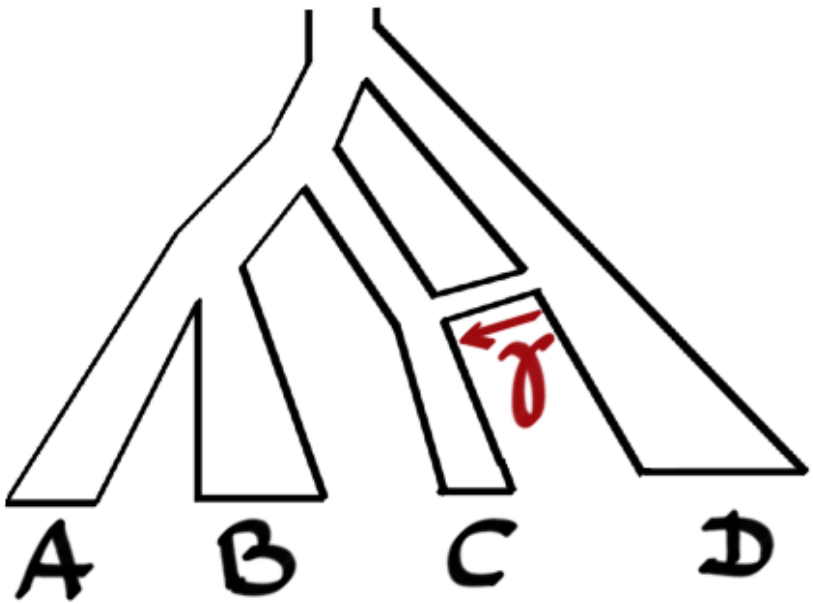
Does **incomplete lineage sorting** impact tree reconstruction?

yes! concatenation is not robust to ILS. (Kubatko & Degnan 2007)

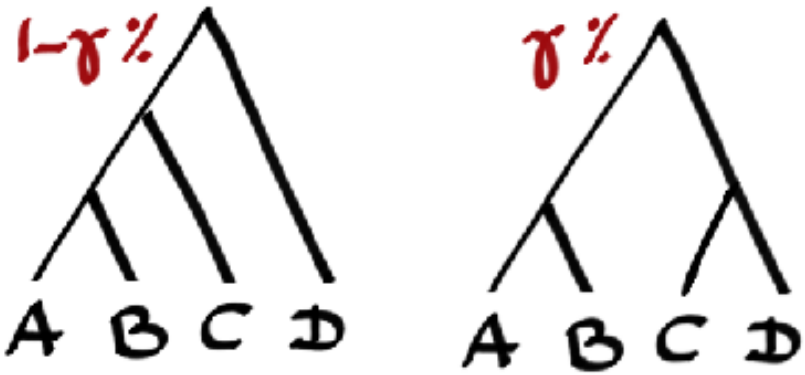
anomalous genes trees
coalescent methods: *BEAST, MP-EST, ASTRAL, SDVquartet, etc.

can we ignore introgression?

Does **gene flow** / **introgression** impact coalescent methods?



gene tree discordance



$\gamma =$ inheritance, e.g. Neanderthals - modern humans: $\gamma \sim 2\%$

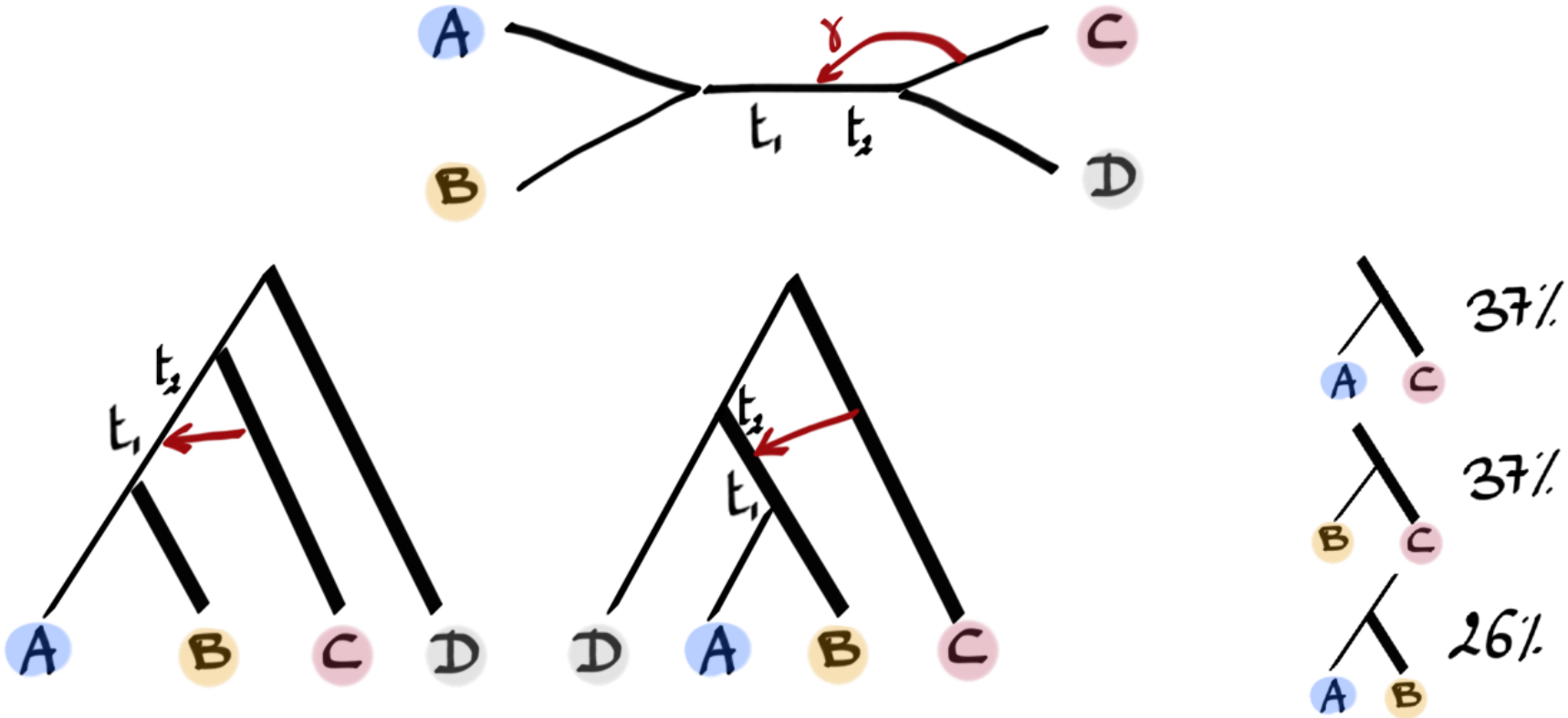
can we ignore introgression?

Does **gene flow** / **introgression** impact coalescent methods?

yes! some coalescent-based methods are not robust to gene flow.

(Solís-Lemus, Yang & Ané 2016)

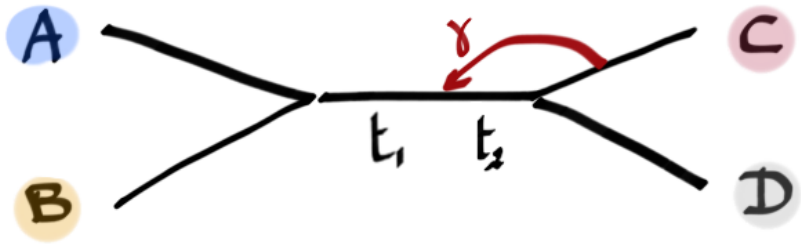
anomalous unrooted gene trees: AuGT



under network model (Solís-Lemus, Yang & Ané 2016)

under continuous gene flow between sister species (Long & Kubatko 2018)

anomalous unrooted gene trees: AuGT



frequency of gene trees:

Quartet	$\gamma = 0.0$	$\gamma = 0.1$	$\gamma = 0.3$
$AB\|CD$	**0.347**	**0.298**	**0.260**
$CA\|BD$	0.327	0.351	0.370
$CB\|AD$	0.327	0.351	0.370

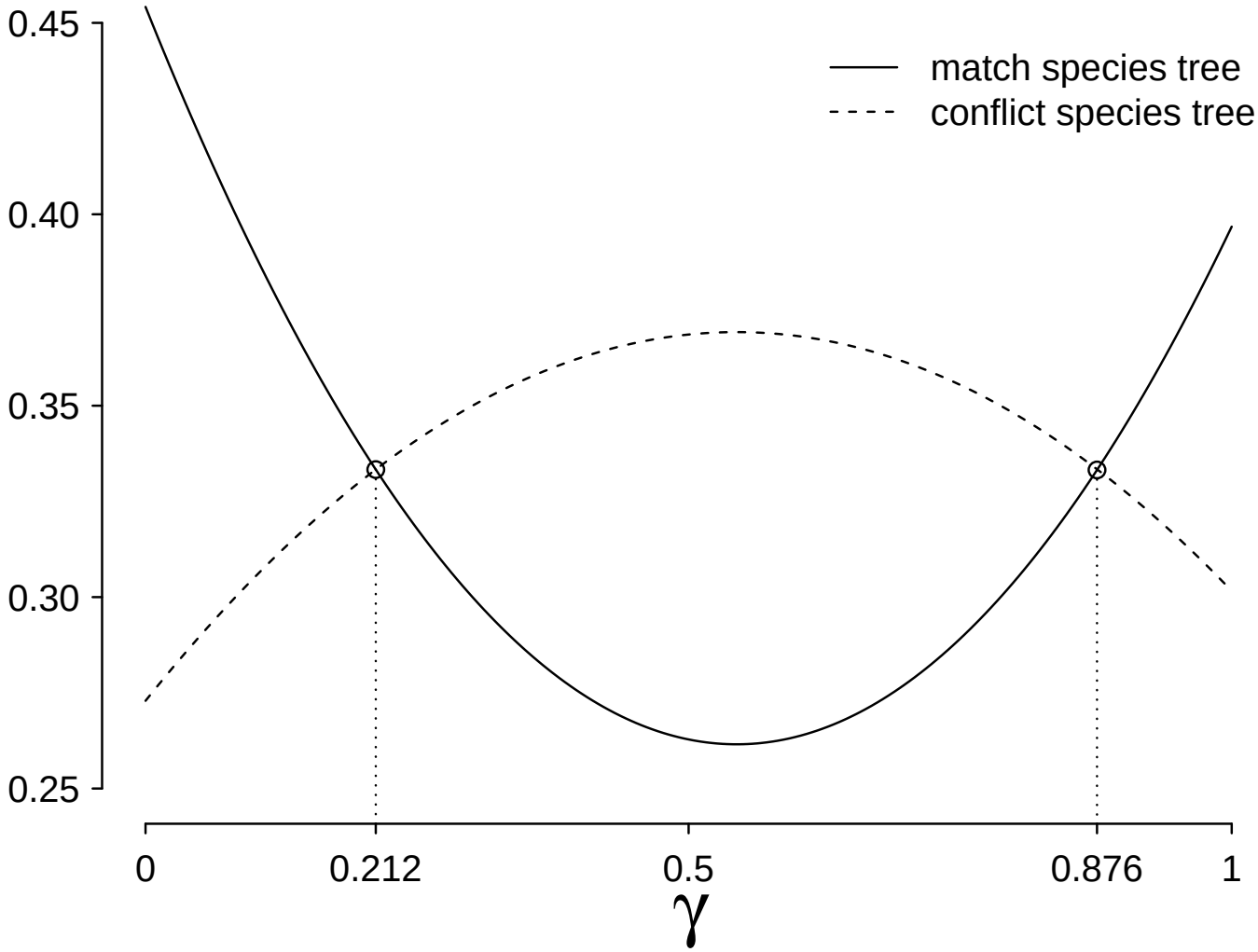
$t_1 = t_2 = 0.01$

ILS only: **no AuGT** on 4 taxa (Degnan 2013)

ILS + gene flow: **AuGT** on 4 taxa (Solís-Lemus, Yang & Ané 2016)

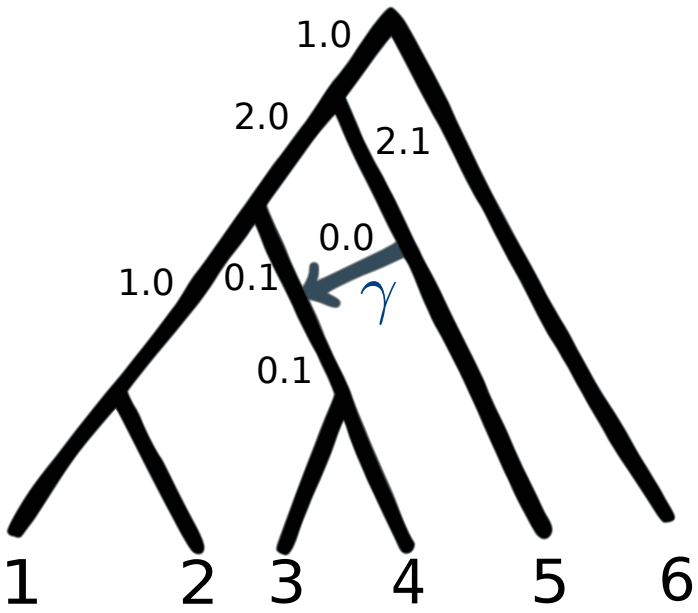
rooted gene trees, 3 taxa: same pattern

anomaly zone with gene flow (4 taxa)



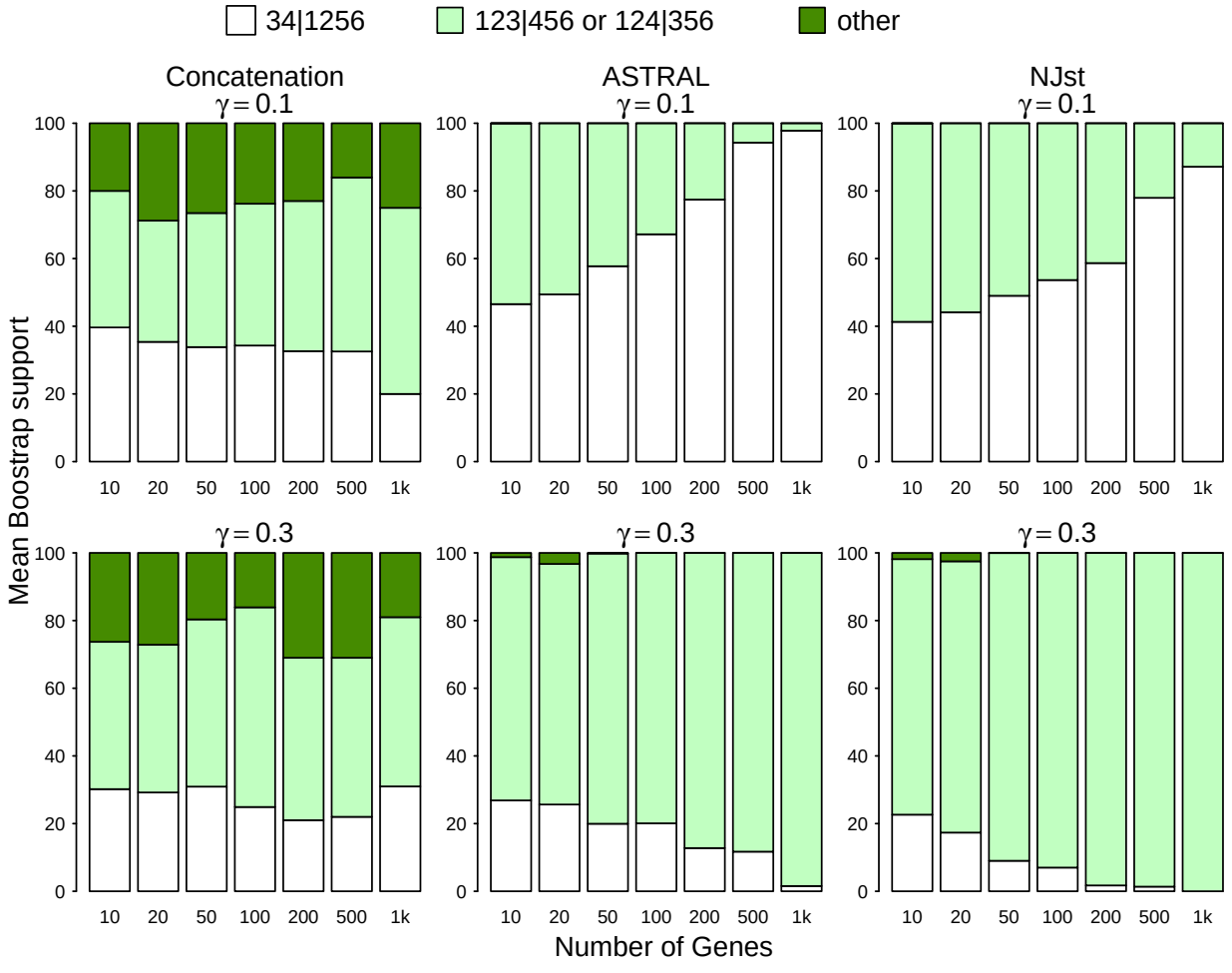
$t_1 = t_2 = 0.1, t_3 = 0, t_4 + t_5 = 4.1$

inconsistent methods: concatenation, ASTRAL, NJst

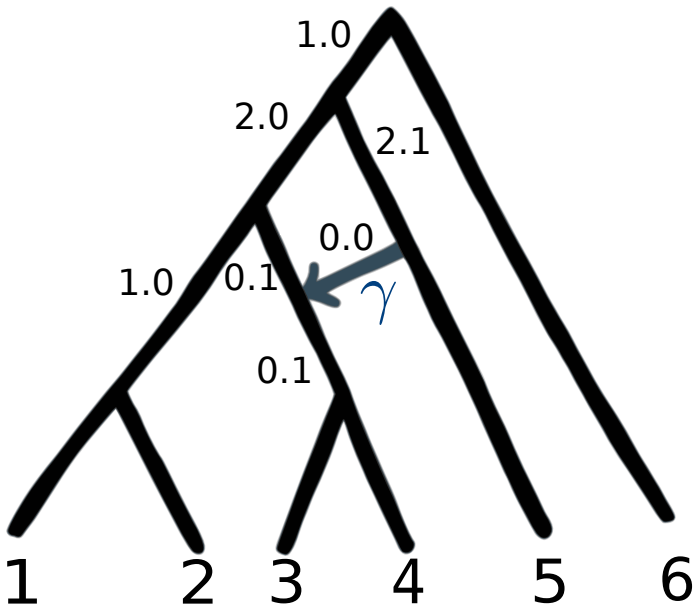


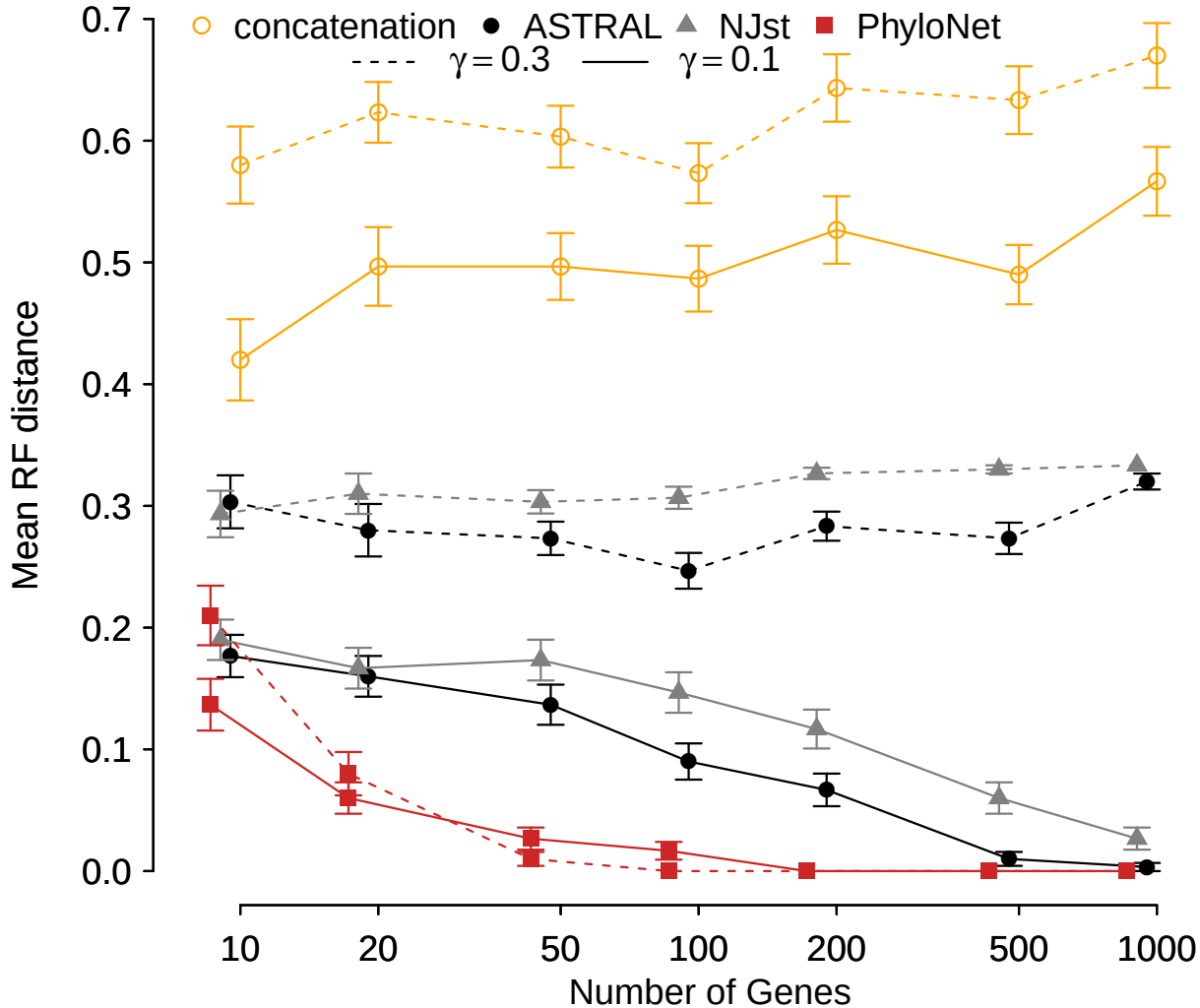
3 & 4 sister

3 & 4 not sister



inconsistent methods: concatenation, ASTRAL, NJst





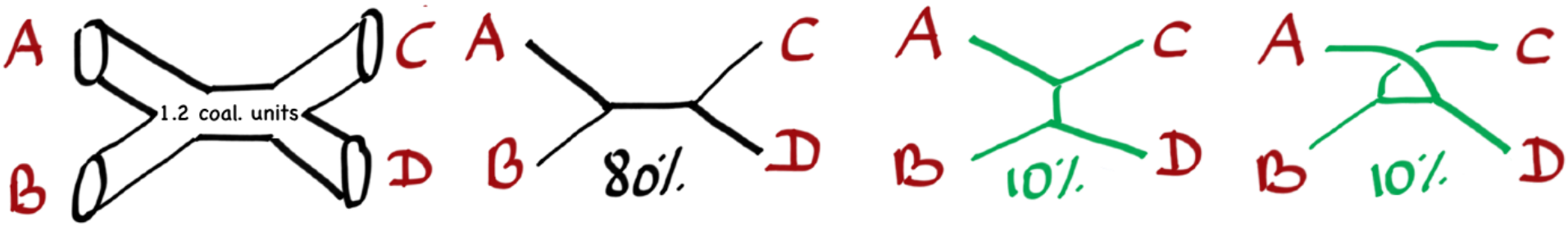
(Solís-Lemus, Yang & Ané 2016)

1. impact of introgression / hybridization?
2. is a tree sufficient, or do we need a network?
3. network models

does a tree fit the data well? or network needed?

TICR: goodness-of-fit test of ILS on a population tree Stenz et al. (2015)

expectation from ILS: equal % genes (CF) with minor resolutions



t $p_1 = 1 - \frac{2}{3}e^{-t}$ $p_2 = \frac{1}{3}e^{-t} = p_3$

- similar idea to ABBA-BABA test on SNPs Green et al. (2010) Durand et al. (2011)
- combine all 4-taxon sets in a single test

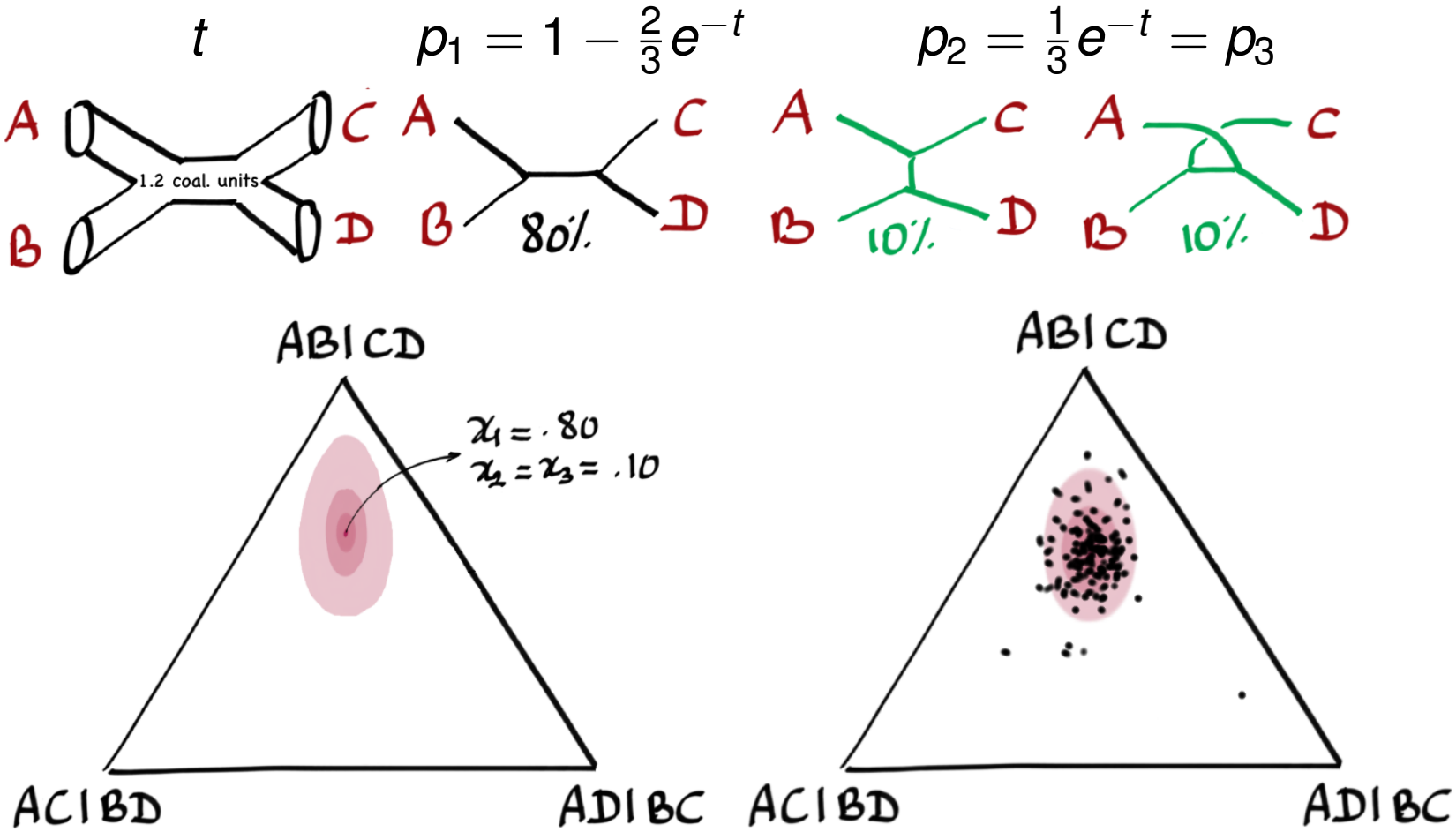
data: quartet concordance factors

4-taxon subset				**proportion of genes** with		
1	2	3	4	12|34	13|24	14|23
				x_1	x_2	x_3
A.gre	A.dig	A.gran	A.za	0.38	0.30	0.32
A.gre	A.dig	A.gran	A.mad	0.42	0.28	0.30
⋮				⋮		
A.gran	A.za	A.per	A.mad	0.25	0.35	0.40

Stenz et al. (2015): 3,595 loci; 30 taxa so 27,405 four-taxon sets

x_1, x_2, x_3: % genes for 3 quartet trees, one 4-taxon set

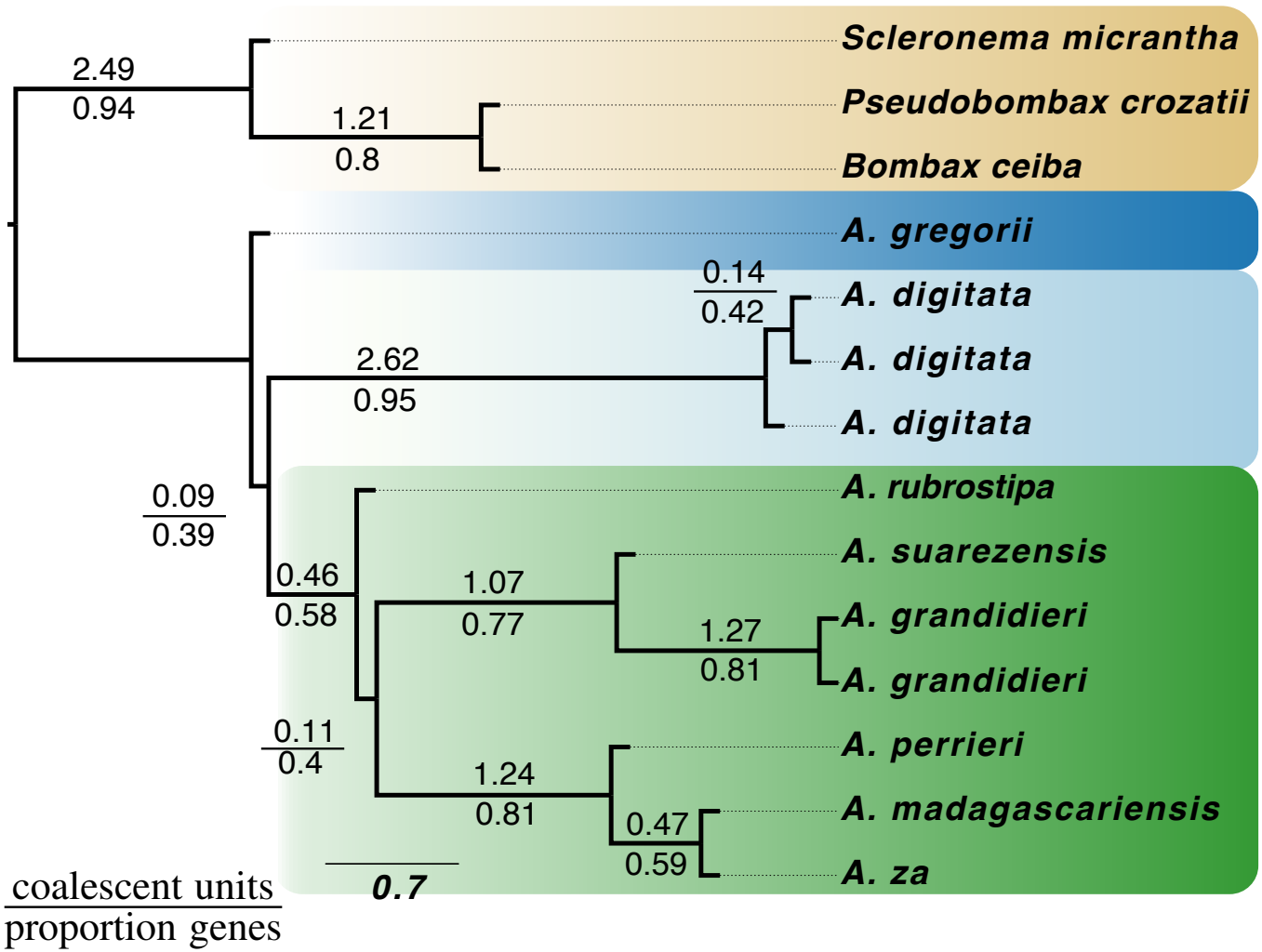
$\sim$ Dirichlet, precision α, centered at p_1, p_2, p_3 from the tree:



- p-value for each 4-taxon set
- overall test: based on proportion of outlier 4-taxon sets

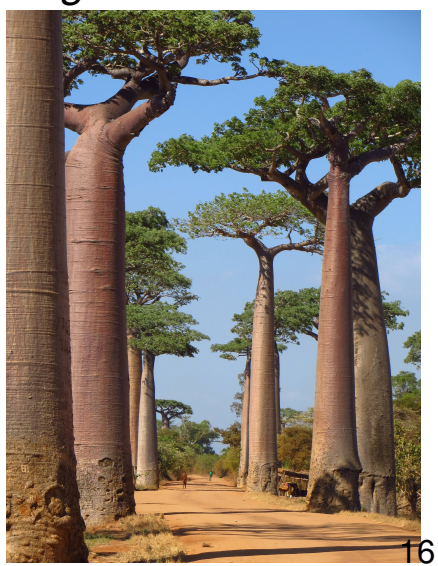
baobabs (*Adansonia*): tree with ILS rejected (p=0.04)

14 individuals, 282 orthologous genes (targeted sequence capture)



A. rubrostipa

A. grandidieri



1. impact of introgression / hybridization?
2. is a tree sufficient, or do we need a network?
3. network models

network models

early work:

- based on parsimony
- no gene tree error
- no ILS (except MDC)

focus for today:

- the multispecies network coalescent model
- network thinking: interpretation issues
- available methods: pros and cons

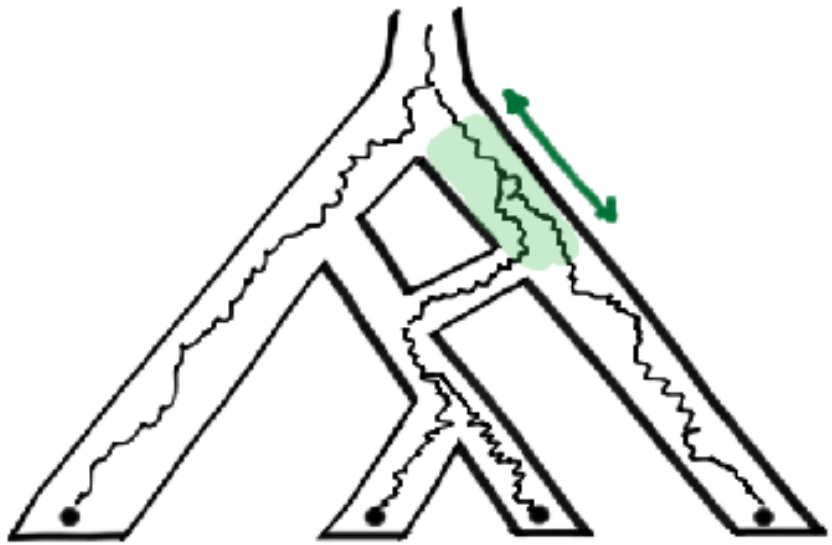
coalescent for ILS: extended to network

network coalescent:

- branch lengths: coalescent units for ILS
- network topology: extra edges for gene flow, hybridization or HGT
- inheritance γ, $1-\gamma$ on hybridization edges

Meng & Kubatko (2009), Yu Degnan & Nakhleh (2012)

coalescent for ILS: extended to network

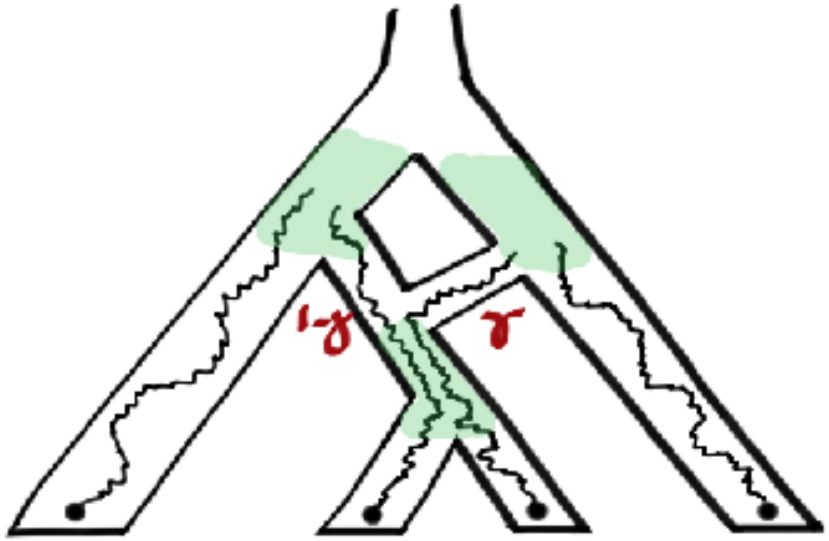


network coalescent:

- branch lengths: coalescent units for ILS
- network topology: extra edges for gene flow, hybridization or HGT
- inheritance γ, $1-\gamma$ on hybridization edges

Meng & Kubatko (2009), Yu Degnan & Nakhleh (2012)

coalescent for ILS: extended to network



network coalescent:

- branch lengths: coalescent units for ILS
- network topology: extra edges for gene flow, hybridization or HGT
- inheritance γ, $1-\gamma$ on hybridization edges

Meng & Kubatko (2009), Yu Degnan & Nakhleh (2012)

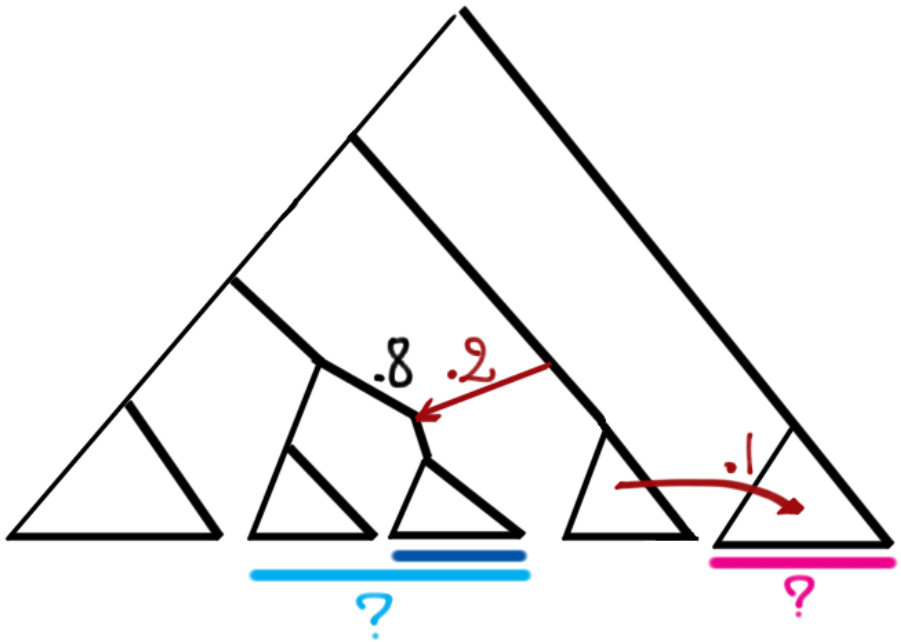
network thinking

- model: simplified one-time events to summarize episodes of continuous gene flow

- blurred "sister" relationship, half-sibs

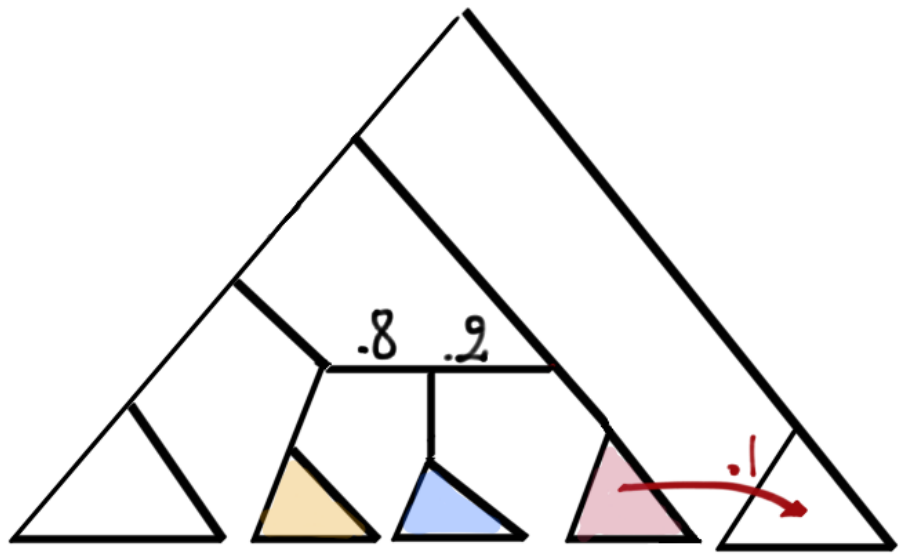
  clade concept?

  classification more difficult

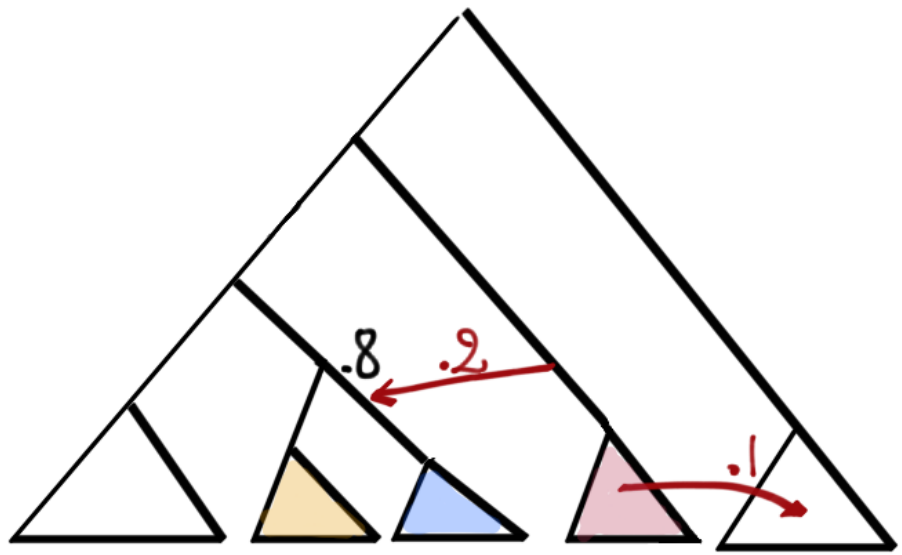


- "major" tree concept: drop each minor hybrid edge ($\gamma < 0.5$)
  meaning of species tree?

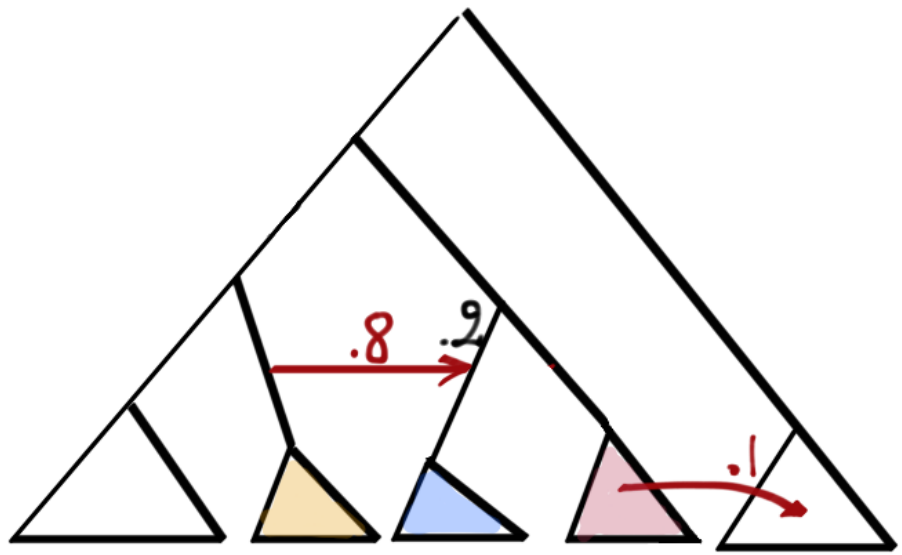
- the network model does not say anything about the process: resulting genetic contributions only
- visual artifacts: can mislead interpretation



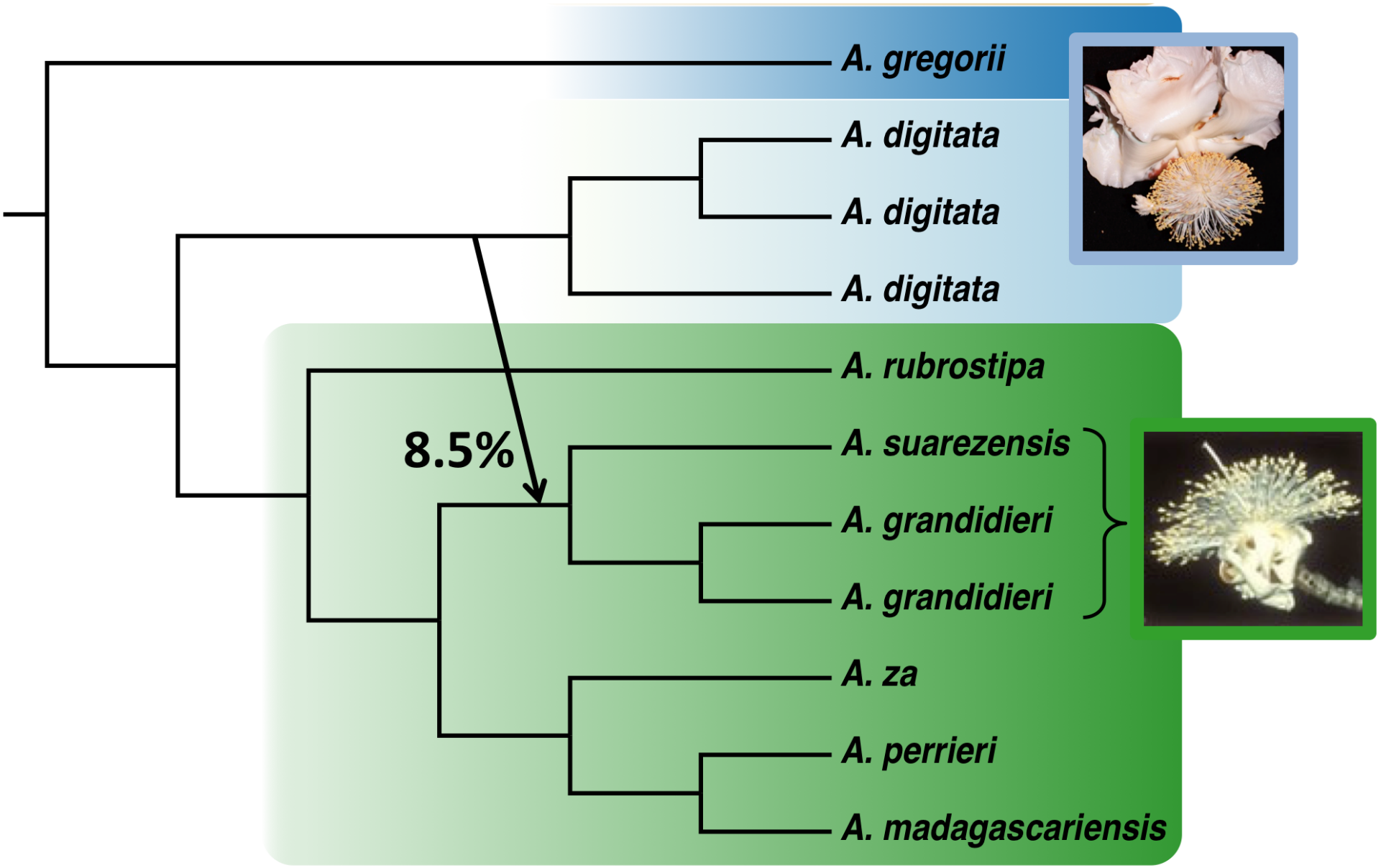
- the network model does not say anything about the process: resulting genetic contributions only
- visual artifacts: can mislead interpretation



- the network model does not say anything about the process: resulting genetic contributions only
- visual artifacts: can mislead interpretation



baobabs: 1 reticulation event



network coalescent: maximum (pseudo) likelihood

Fast algorithms and heuristics for phylogenomics under ILS and hybridization

Yun Yu*, Nikola Ristic*, Luay Nakhleh*

(2013, BMC Bioinformatics)

Maximum likelihood inference of reticulate evolutionary histories

Yun Yu[a,1], Jianrong Dong[a], Kevin J. Liu[a,b], and Luay Nakhleh[a,b,1]

(2014, PNAS)

Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting

Claudia Solís-Lemus[1]*, Cécile Ané[1,2]

(2016, PLOS Genetics)

- identifiability: what can (or cannot) we learn from data?

network coalescent: Bayesian

Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent

Dingqiao Wen[1]*, Yun Yu[1], Luay Nakhleh[1,2]*

(2016, PLOS Genetics)

Syst. Biol. 67(3):439–457, 2018

DOI:10.1093/sysbio/syx085
Advance Access publication October 27, 2017

Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data

DINGQIAO WEN[1] AND LUAY NAKHLEH[1,2,*]

Bayesian Inference of Species Networks from Multilocus Sequence Data

Chi Zhang,*[,1,2,3] Huw A. Ogilvie,[4,5] Alexei J. Drummond,[5,6] and Tanja Stadler*[,1,2]

(2018, MBE)

Bayesian inference of phylogenetic networks from bi-allelic genetic markers

Jiafan Zhu[1], Dingqiao Wen[1], Yun Yu[1], Heidi M. Meudt[2], Luay Nakhleh[1,3]*

(2018)

how to estimate a species network?

Complex task!

- PhyloNet — gene trees, multiple alignments, biallelic SNPs
- PhyloNetworks — gene trees, quartet concordance factors
- BEAST2 — multiple alignments

None of these methods scale well to many **species**

PhyloNet and BEAST2 methods: do **not** scale well to many **loci**

STEM-hy	gene trees rooted, BL	likelihood	hybridization b/w sister lineages
PhyloNet `InferNetwork_ML`	gene trees rooted	likelihood	
PhyloNet `InferNetwork_MPL`	gene trees rooted	triplet likelihood	
PhyloNetworks `SNaQ`	gene trees or quartet CFs	quartet likelihood	level-1 network
PhyloNet `MCMC_GT`	gene trees rooted	Bayesian	compound prior
PhyloNet `MCMC_SEQ`	alignments	Bayesian	compound prior no rate variation
BEAST2 `SpeciesNetwork`	alignments	Bayesian	birth-hyb prior
PhyloNet `MLE_BiMarkers`	biallelic sites	likelihood	compound prior
PhyloNet `MCMC_BiMarkers`	biallelic sites	Bayesian	compound prior
HyDe	sites	invariants	4 taxa, 1 hyb.

gene trees versus sequences?

- alignments: slower, but more accurate (if rate assumptions met)
- gene trees: faster, but less accurate
  data summary, gene tree error
- quartet concordance factor: data summary,
  but gene tree error can be accounted for

use branch lengths in gene trees?

if so, dangerous assumptions of no rate variation typically:

- all genes evolve at the same rate
- same rate on all gene lineages: molecular clock
  or same departure from a molecular clock across all genes

For reconstructing species *trees*, methods that ignore branch lengths in gene trees are more robust.

If rate variation suspected, favor

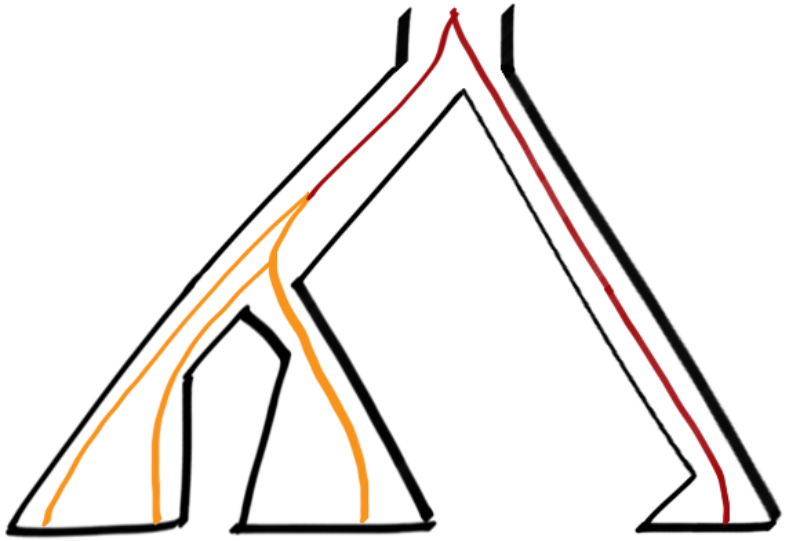
- methods based on gene tree topologies, or
- BEAST2 with gene multipliers and relaxed clock.

use rooted or unrooted gene trees?

danger of rooting all gene trees with an outgroup:

outgroup involved in ILS

or saturation, or long branch attraction

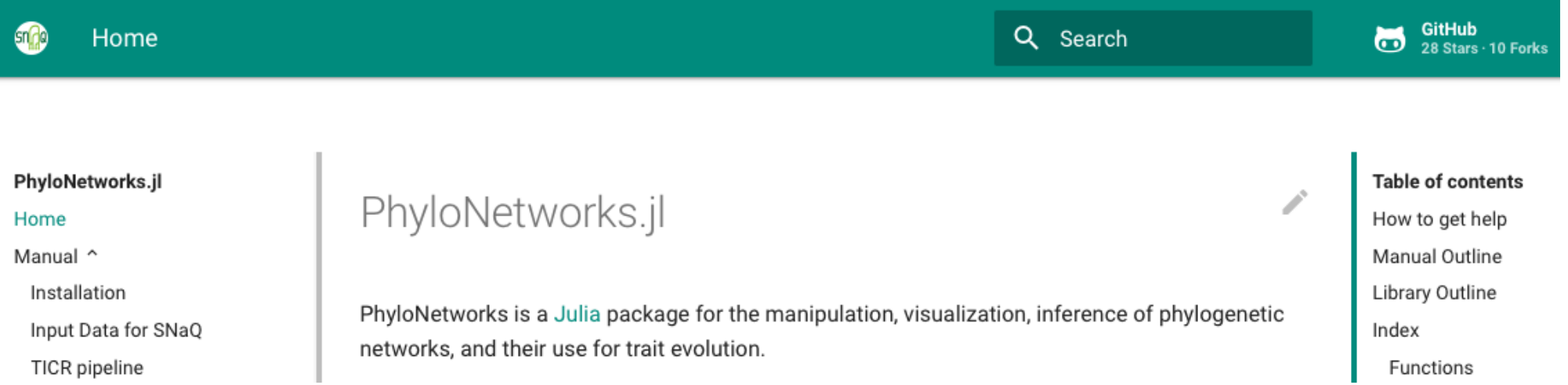


Gatesy, DeSalle & Wahlberg (2007): rooting errors explain incongruence in yeast dataset (Rokas et al. 2003)

we rarely check the root of 1000 gene trees...

My own preference (but the field is moving fast):

- BEAST2-SpeciesNetwork or PhyloNet-Bayesian
- PhyloNetworks-SNaQ for more species and/or more loci
- HyDe and ABBA-BABA tests to confirm on specific taxon subsets



PhyloNetworks package:

- SNaQ: gene trees or quartet CFs $\rightarrow$ species network
  bootSNaQ: bootstrap gene trees $\rightarrow$ bootstrap networks
  bootstrap support: for tree edges, gene flow recipient, donor

- trait evolution on networks: continuous response

- plot, root, re-root networks
  extract the major tree, extract all displayed trees

Acknowledgements

Claudia Solís-Lemus

Paul Bastide

Mengyao Yang
John Malloy
John Spaw
Doug Bates
Sarah Friedrich

baobabs:
David Baum Nisa Karimi

Jonathan Wendel, Joe Gallagher,
Corrinne Grover, Noah Stenz

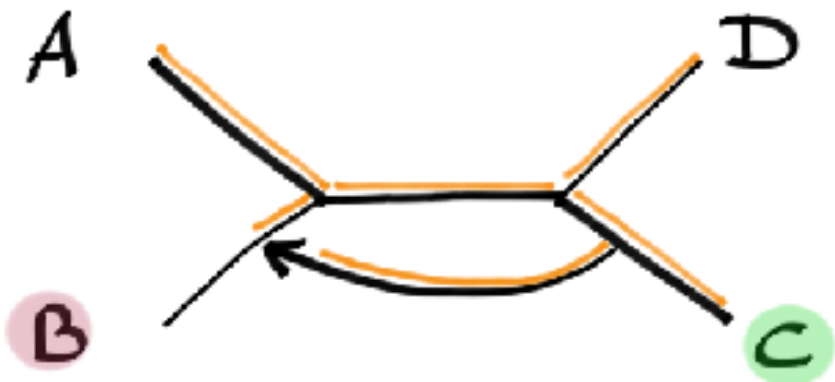
DEB 0949121
DEB 1354793

is the root identifiable?

no

same quartet proportions ($\widehat{\text{CF}}$s) from these networks, provided same parameters (γ, branch lengths t)

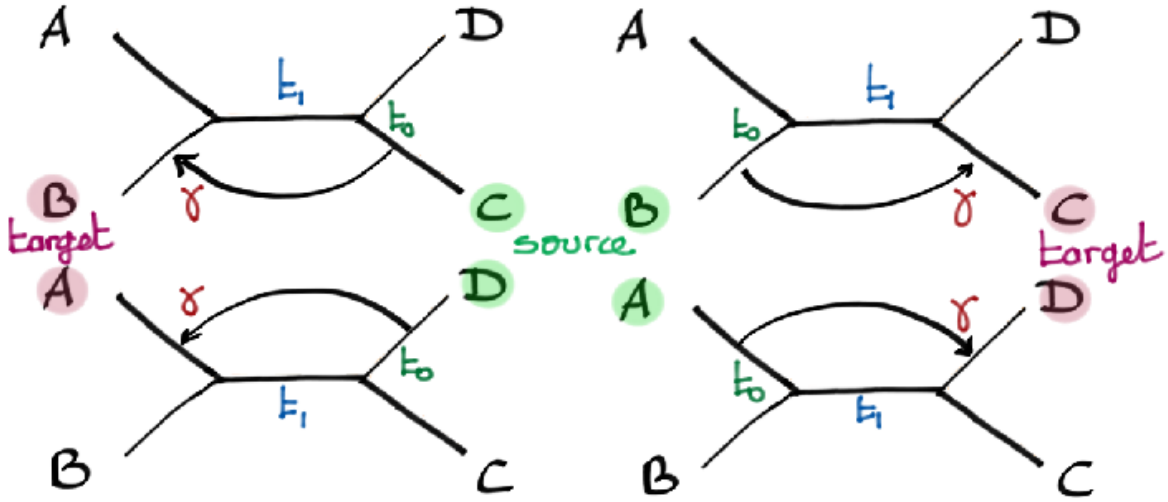
infer semi-directed network, root if after with an outgroup



can we identify the gene flow placement and direction?

4 taxa: no. we can detect its presence only.

same with ABBA-BABA test: not enough info



Same quartet probabilities:

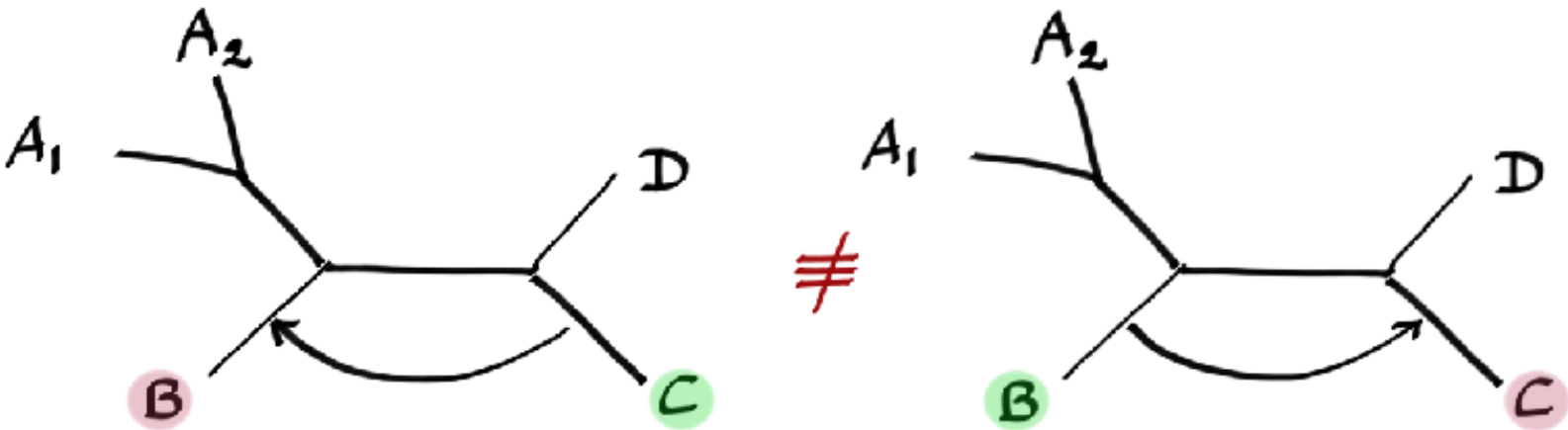
$$\begin{aligned}
\widehat{\mathrm{CF}}(AB|CD) &= (1-\gamma)(1-2/3\,e^{-t_1}) + \gamma\, e^{-t_0}/3 \\
\widehat{\mathrm{CF}}(AD|BC) &= (1-\gamma)\, e^{-t_1}/3 + \gamma(1-2/3\,e^{-t_0}) \\
\widehat{\mathrm{CF}}(AC|BD) &= (1-\gamma)\, e^{-t_1}/3 + \gamma\, e^{-t_0}/3
\end{aligned}$$

can we identify the gene flow placement and direction?

5+ taxa: yes, for most networks

same with D_{FOIL}: 5-taxon version of ABBA-BABA test Pease & Hahn 2015

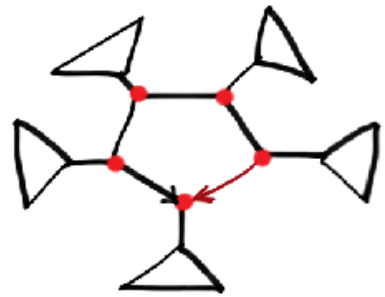
this network is identifiable (presence and placement of gene flow) from the 15 quartet CFs: 3 on A_1BCD, 3 on A_2BCD, 3 on A_1A_2BC, etc.

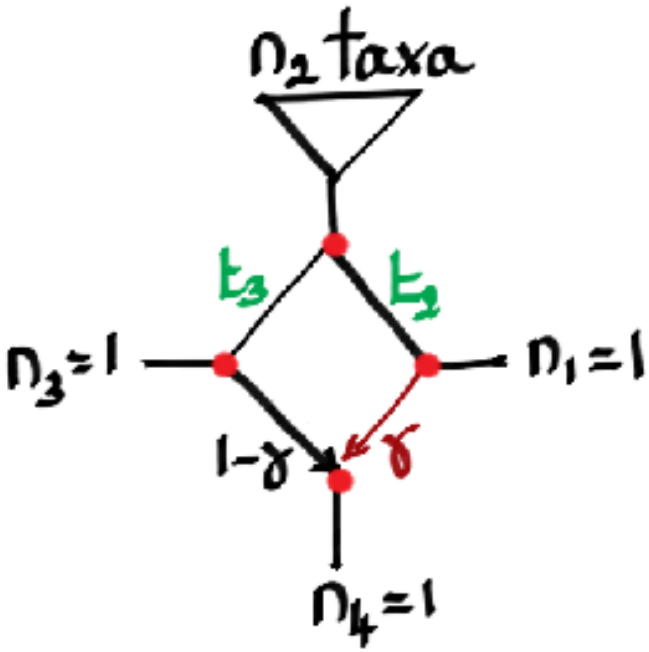


but *not all* networks are identifiable.

are branch lengths and inheritance γ's identifiable?

- $k \geq 5$: **yes**
- $k = 4$: **yes** if $n_1 \geq 2$ or $n_3 \geq 2$ ("good" diamond), **no** otherwise.
- $k = 3$: **no**



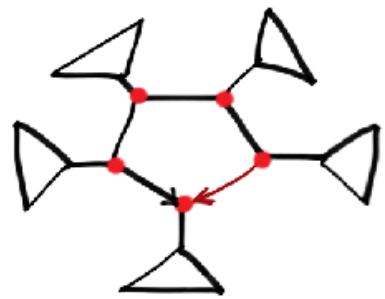


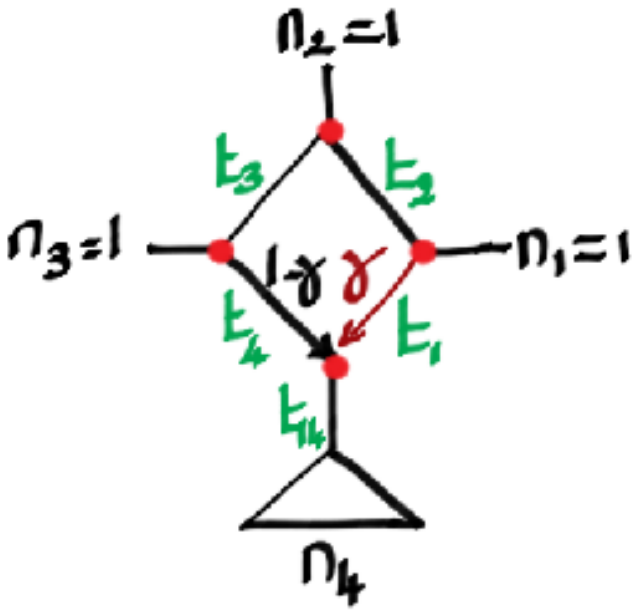
"bad" diamond I:
γ, t_2, t_3 *not* identifiable, but
$\gamma(1 - e^{-t_2})$ and $(1 - \gamma)(1 - e^{-t_3})$ are.

"bad" diamond II:
6 parameters, 5 independent equations only.
in practice: assume $t_{14} = 0$.

are branch lengths and inheritance γ's identifiable?

- $k \geq 5$: **yes**
- $k = 4$: **yes** if $n_1 \geq 2$ or $n_3 \geq 2$ ("good" diamond), **no** otherwise.
- $k = 3$: **no**





"bad" diamond I:
γ, t_2, t_3 *not* identifiable, but
$\gamma(1 - e^{-t_2})$ and $(1 - \gamma)(1 - e^{-t_3})$ are.

"bad" diamond II:
6 parameters, 5 independent equations only.
in practice: assume $t_{14} = 0$.