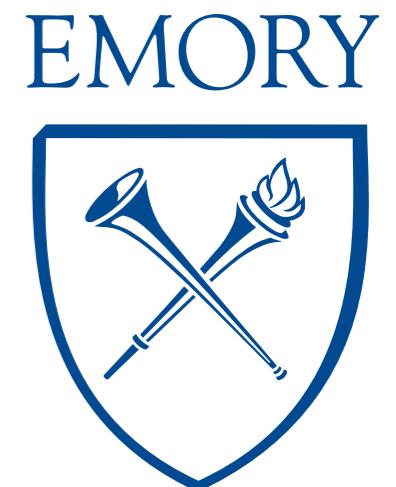


Statistical models on phylogenetic networks

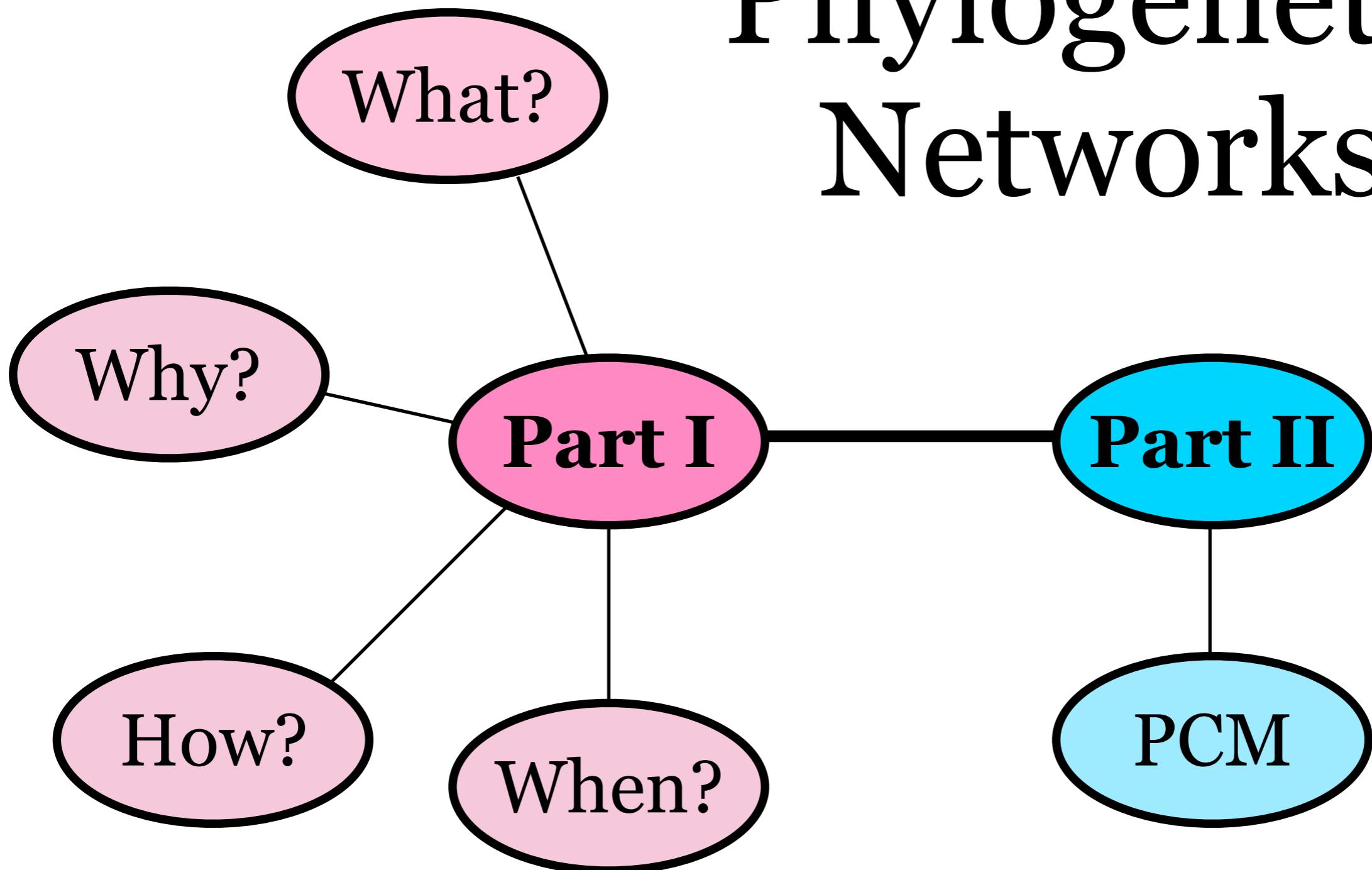
Claudia Solís-Lemus, PhD
Emory University



August 6, 2019

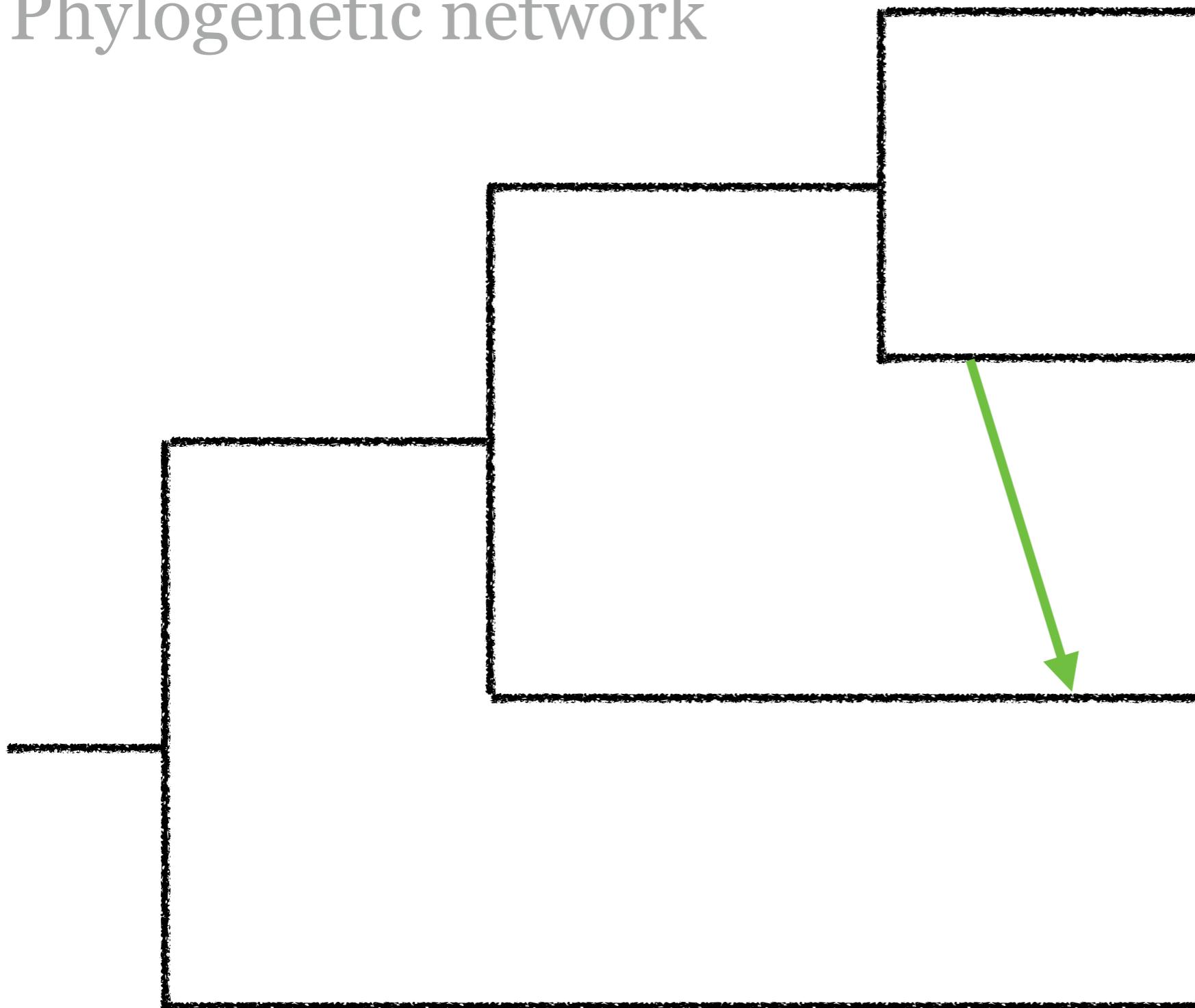


Phylogenetic Networks



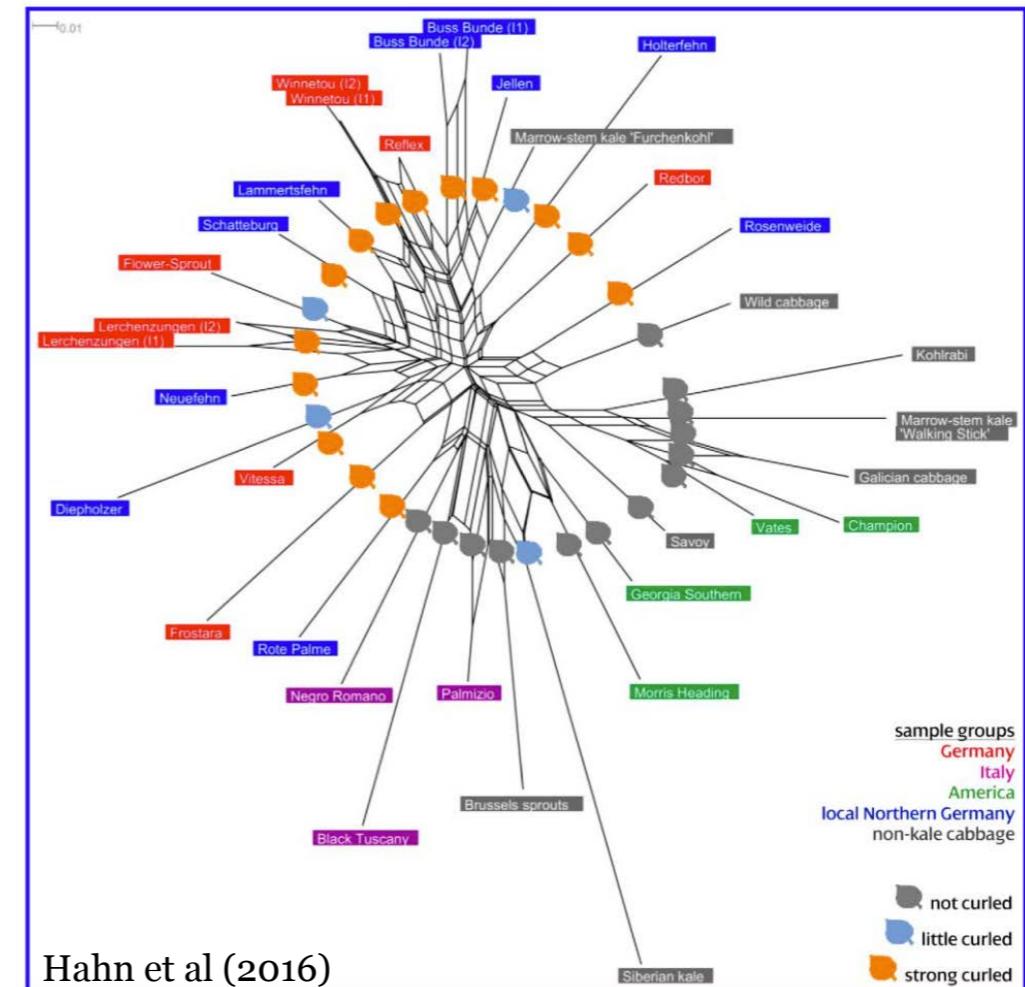
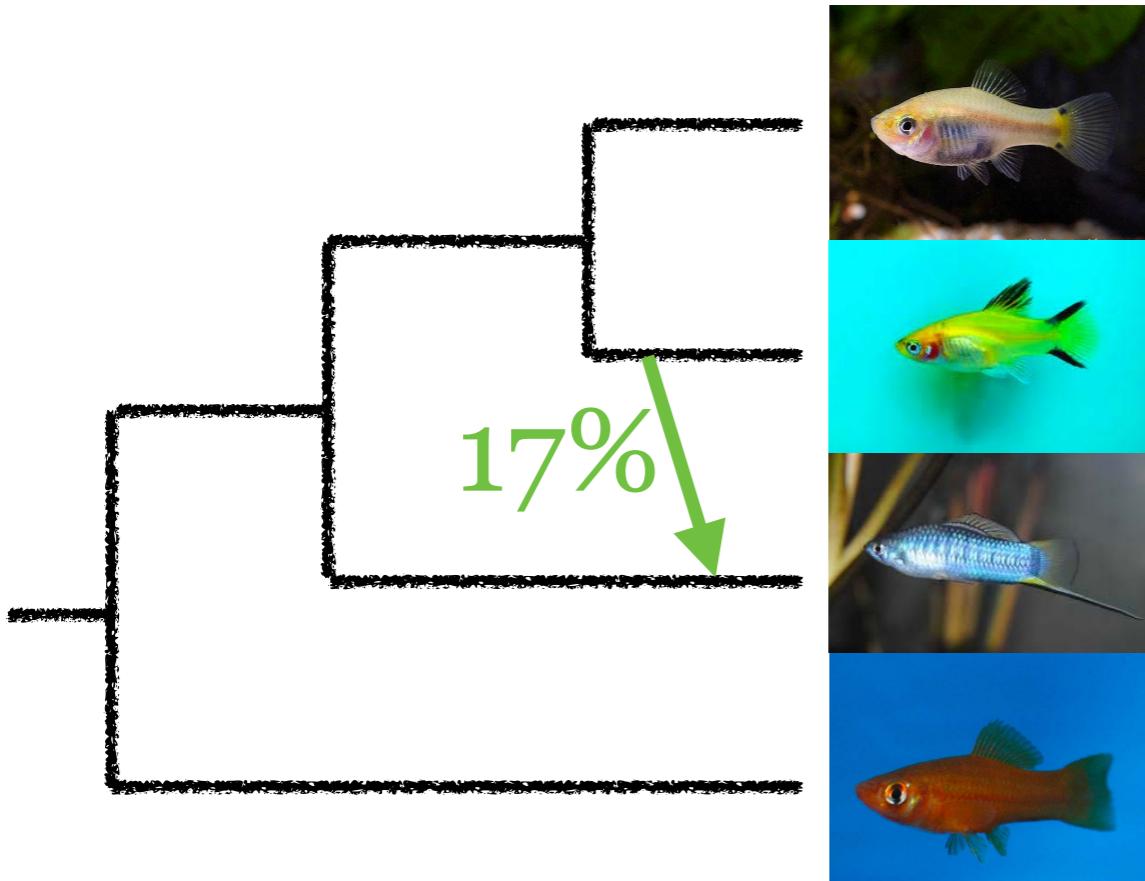
What?

Phylogenetic network



What?

Phylogenetic network

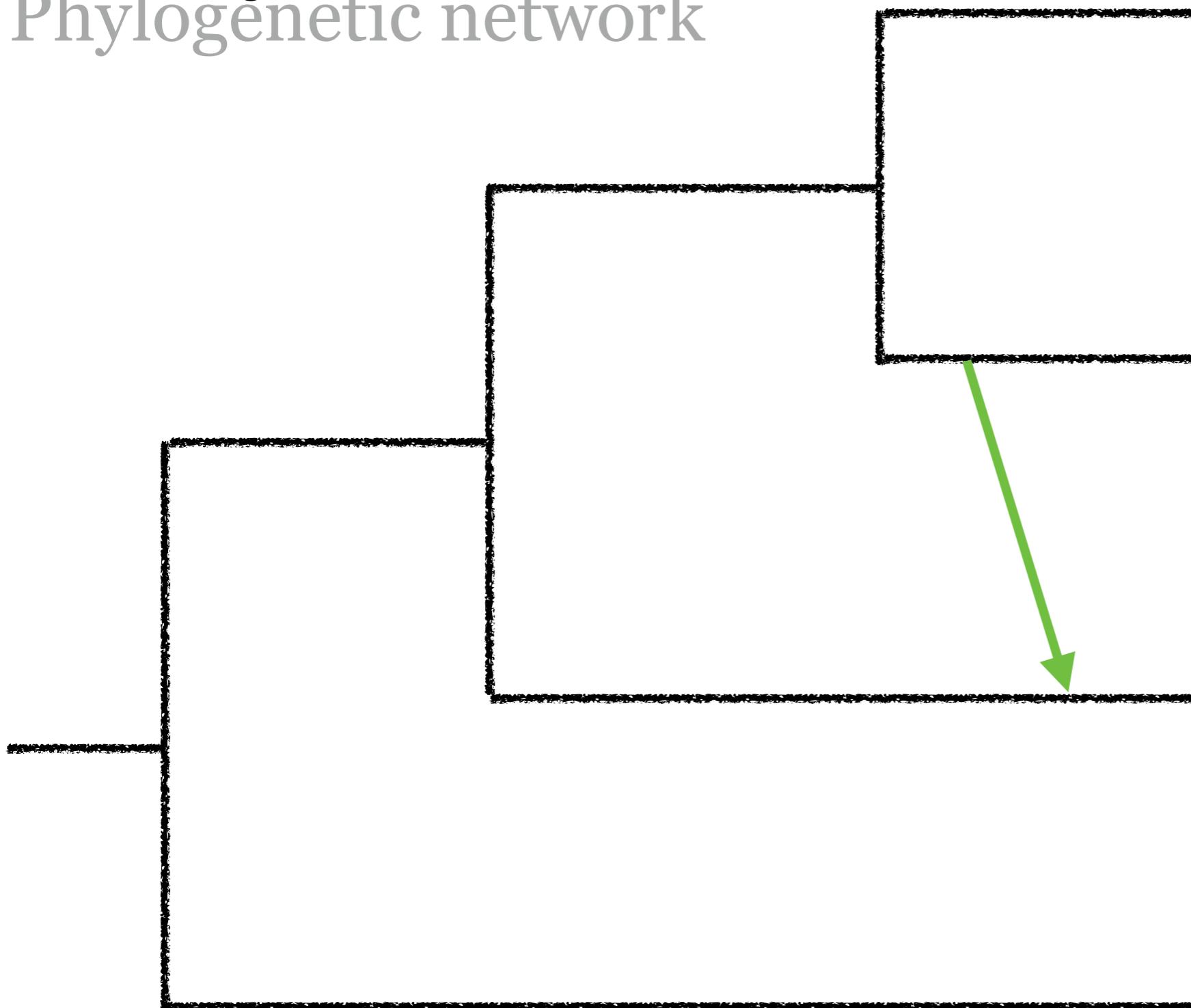


Explicit

Implicit

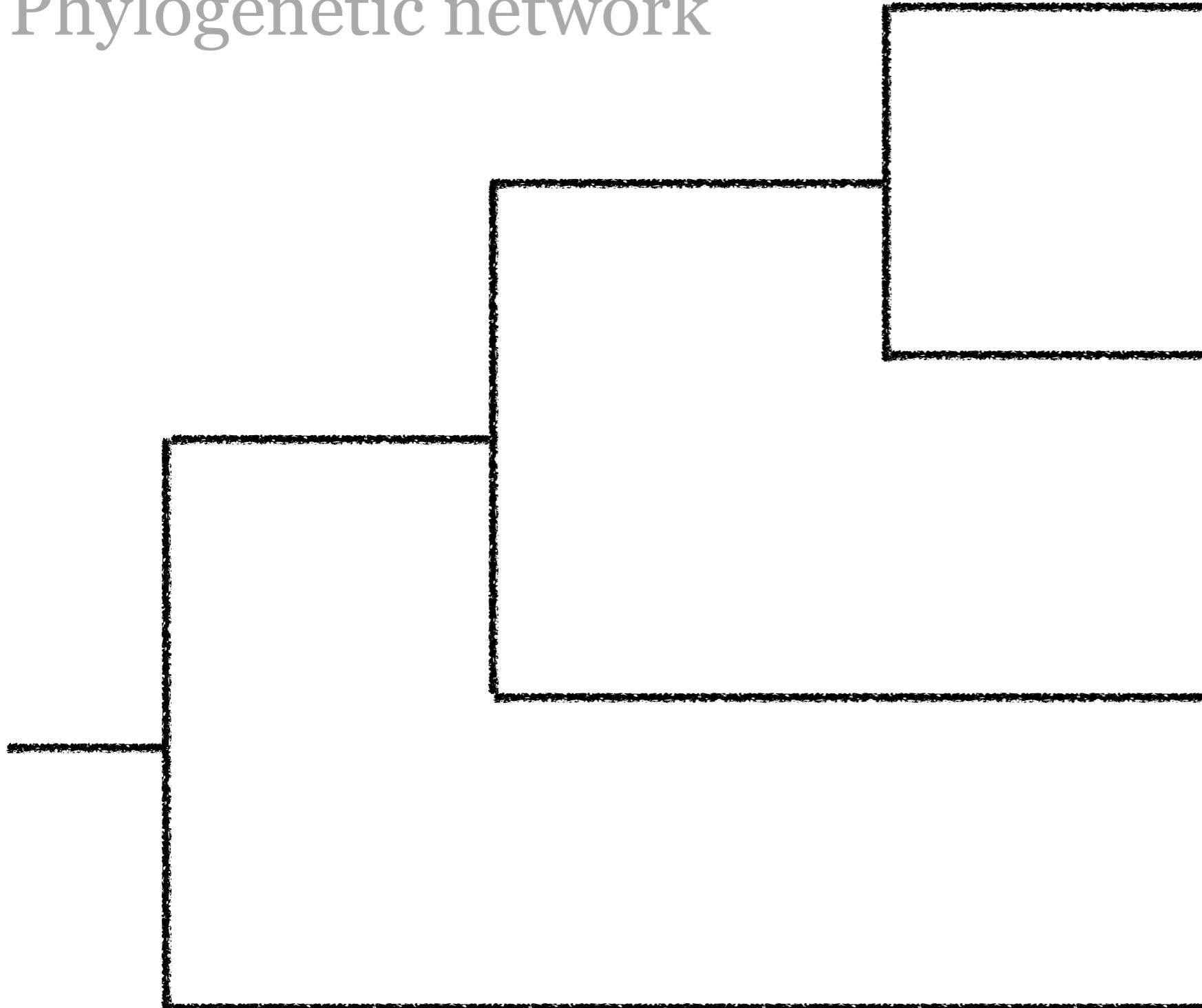
Why?

Phylogenetic network



Why?

Phylogenetic network



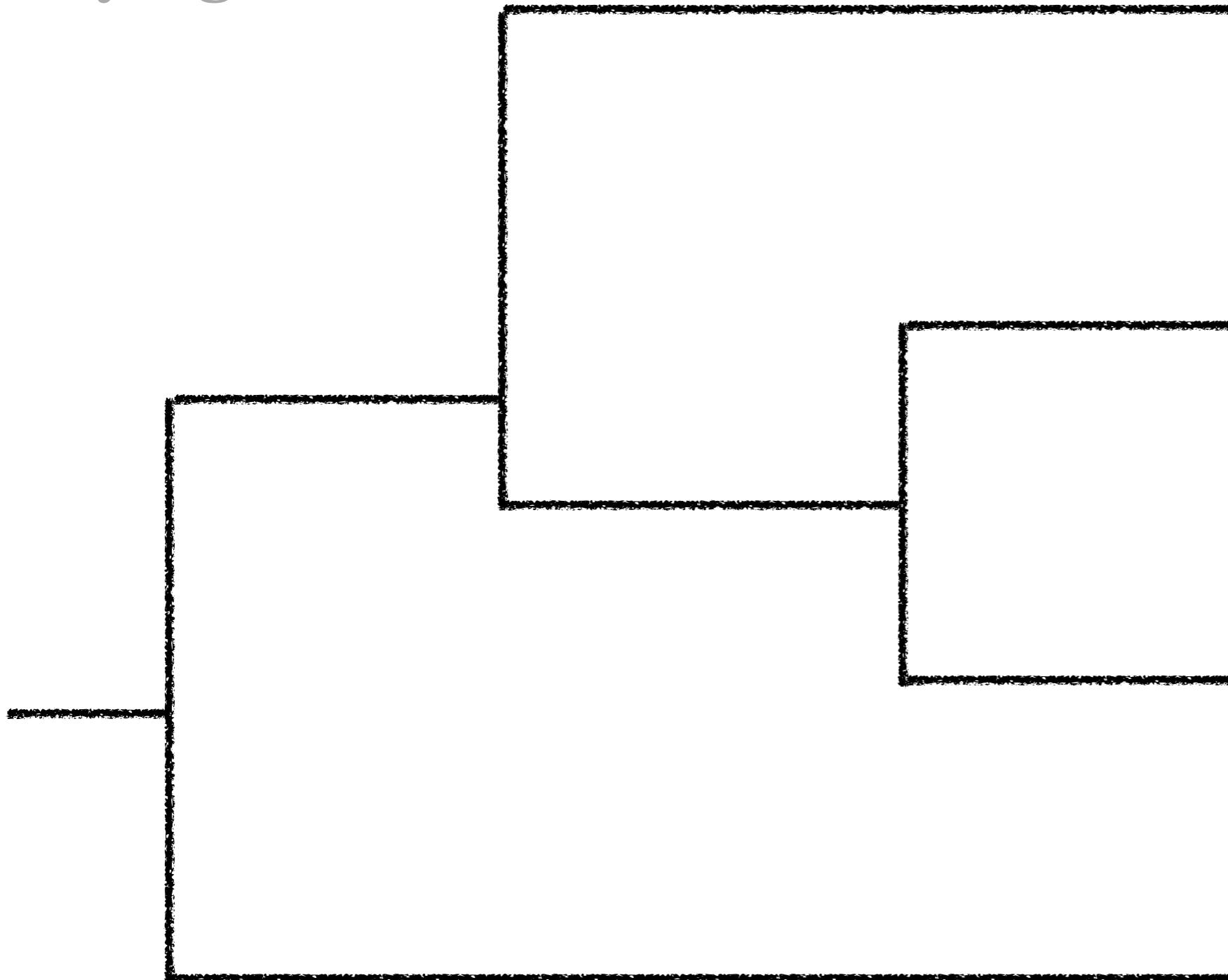
Main tree



Why?

Phylogenetic network

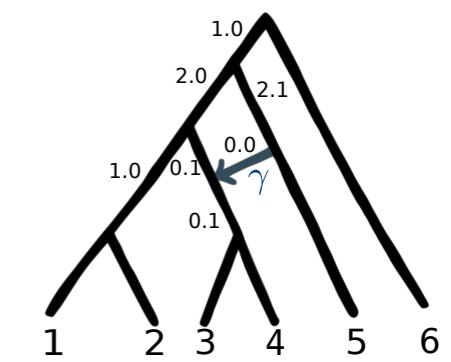
Ignore gene flow
=>Wrong tree!



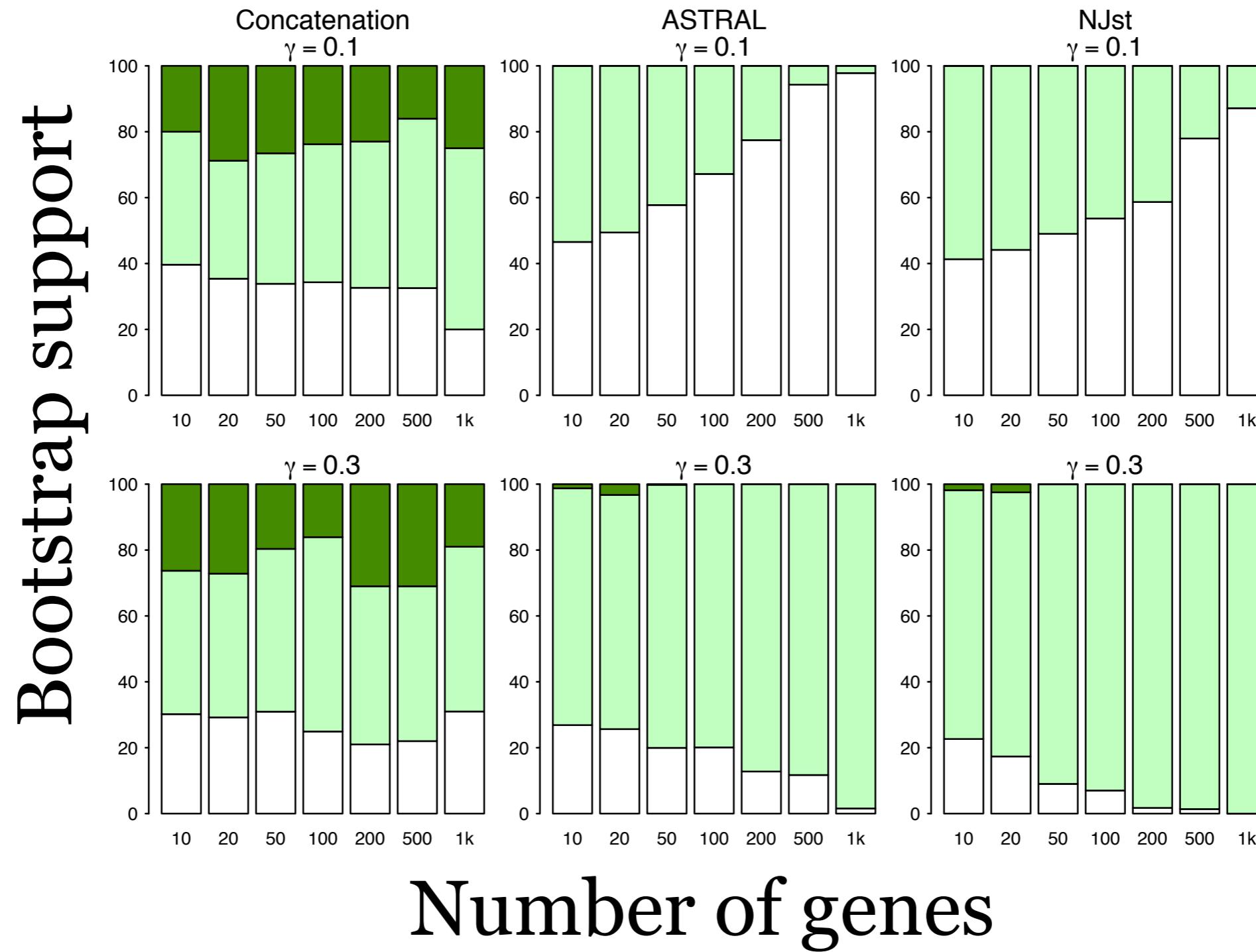
Why?

Phylogenetic network

Coalescent tree methods
not robust to gene flow



White:
true tree



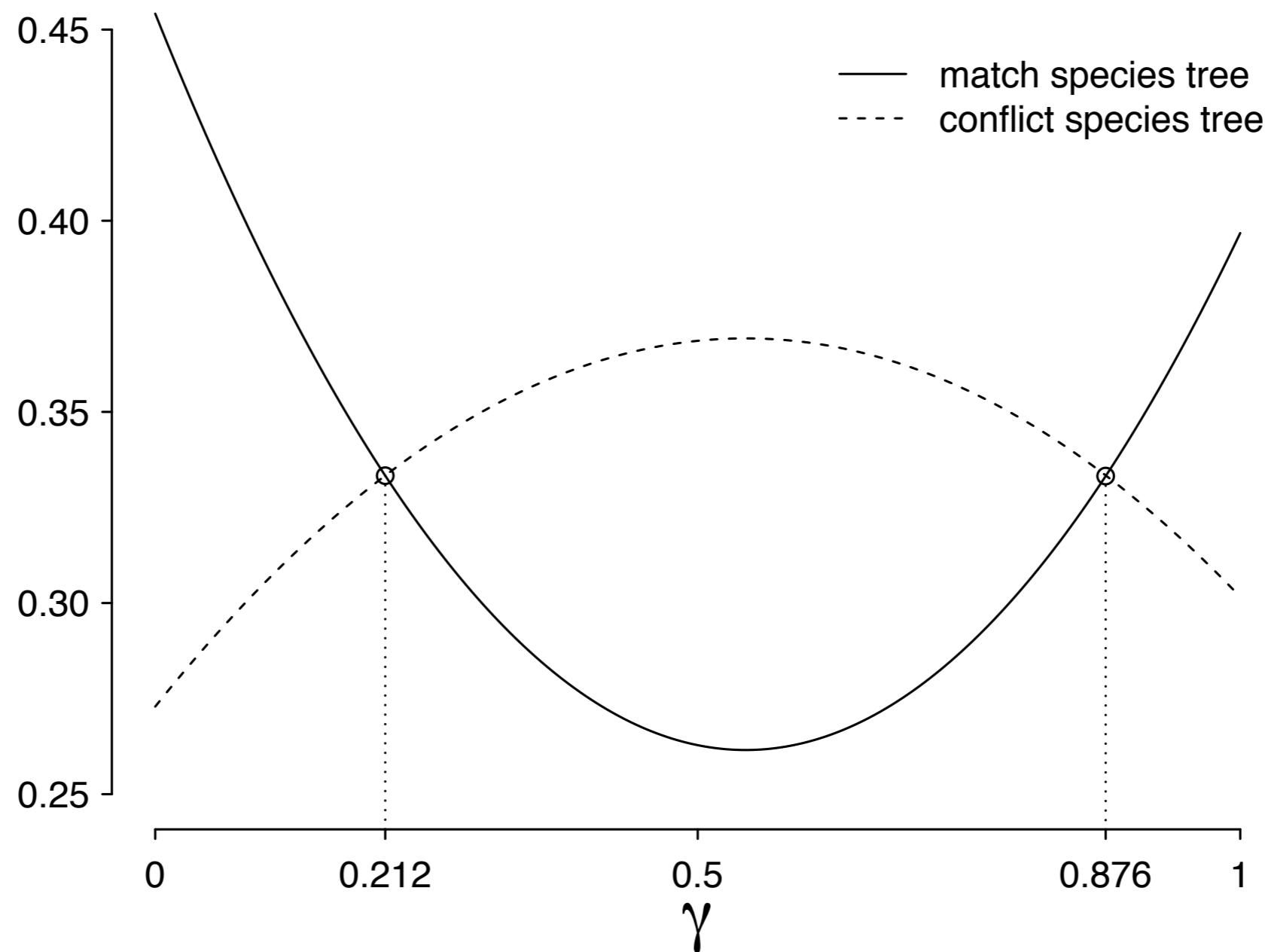
Number of genes

(S.-L., Yang, Ané, 2016, Syst Bio)

ASTRAL (Mirarab et al, 2014)
NJst (Liu&Yu, 2011)

Why? Phylogenetic network

Anomaly zone with gene flow

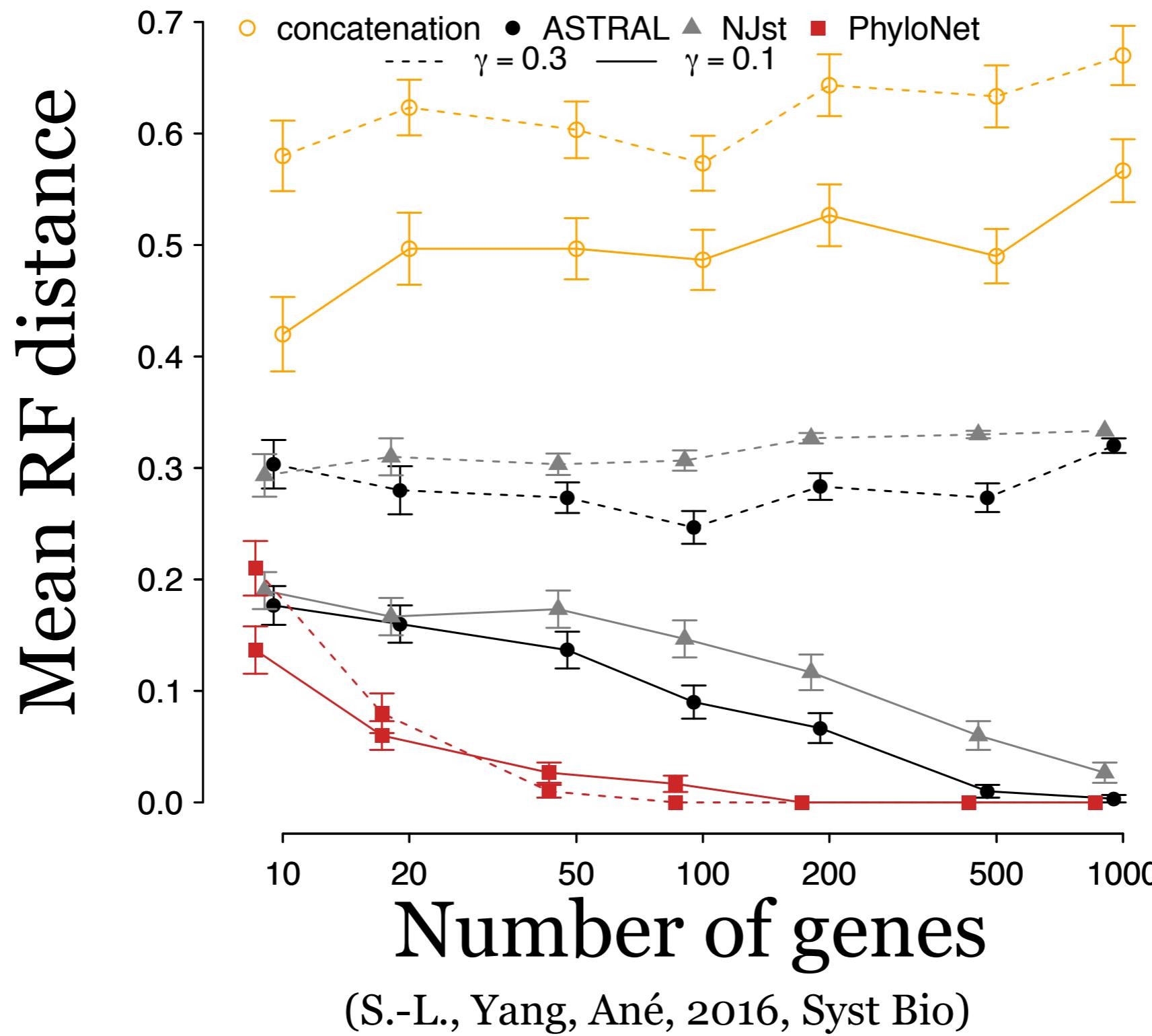
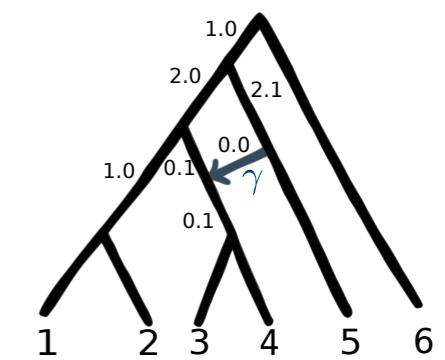


(S.-L., Yang, Ané, 2016, Syst Bio)

Why?

Phylogenetic network

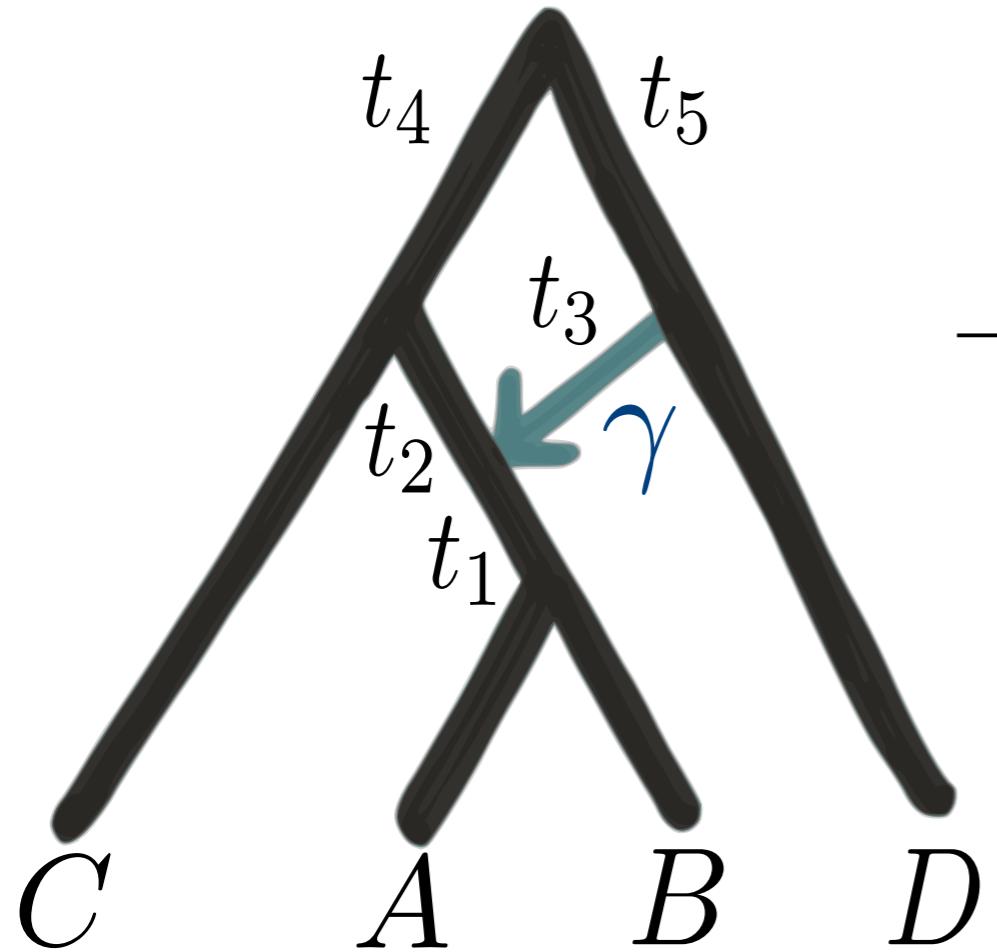
Coalescent tree methods
not robust to gene flow



Why?

Phylogenetic network

Anomalous unrooted
gene trees with gene flow



Frequency among 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

$$t_1 = t_2 = 0.01, t_3 = t_4 = t_5 = 1$$

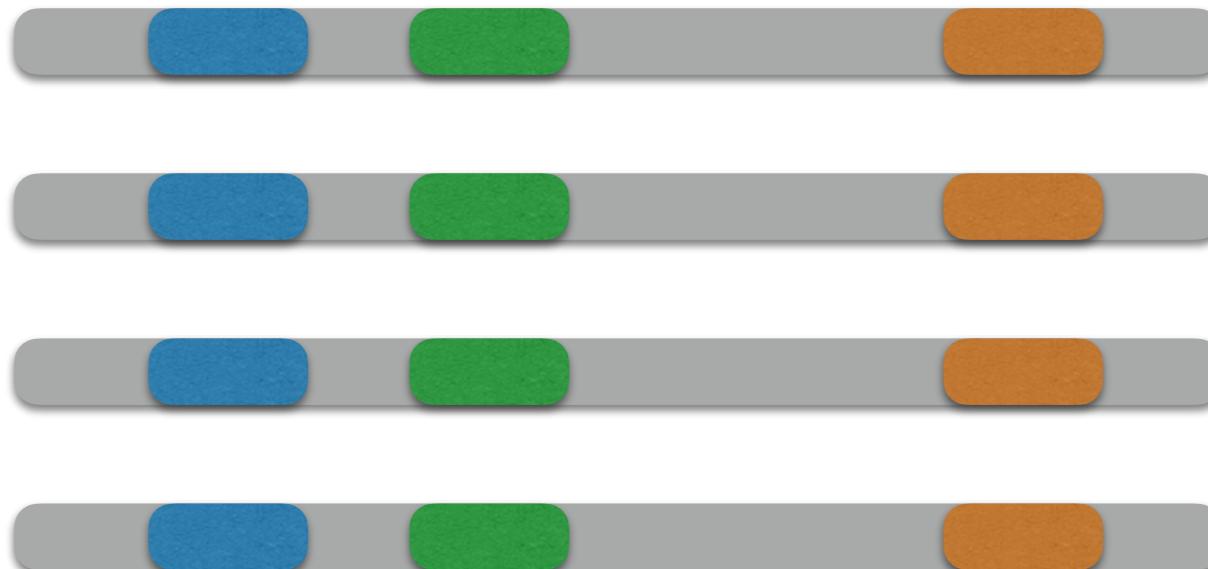
- **ILS**: no AUGT on 4 taxa (Degnan, 2013)
- **ILS+HGT**: AUGT on 4 taxa (S.-L., Yang, Ané, 2016, Syst Bio)

So far...

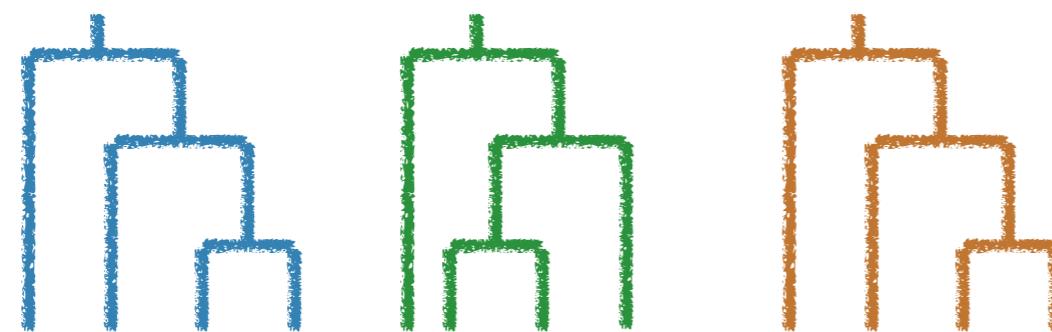
- Networks are good
- Explicit networks are better
- If you ignore gene flow, you can estimate the wrong tree

How?

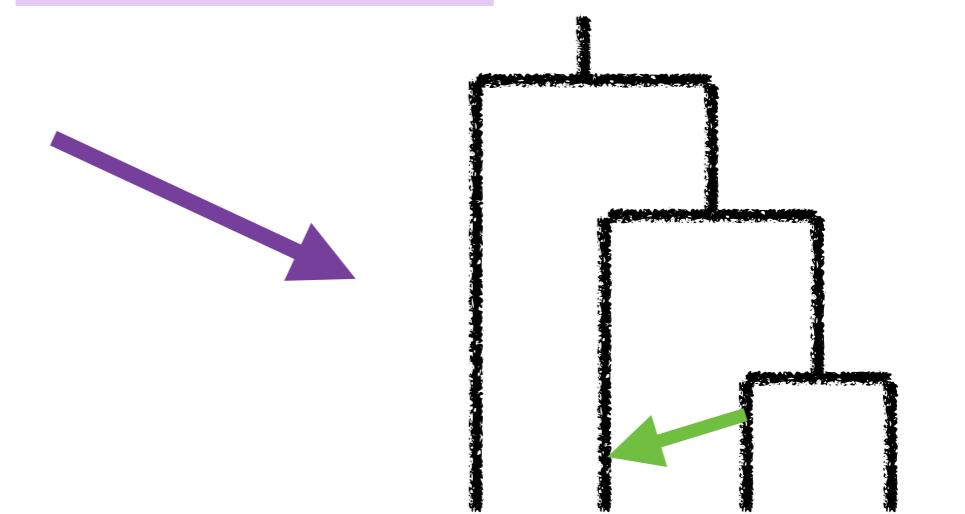
Phylogenetic network



MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
PhyML
(Guindon et al, 2010)



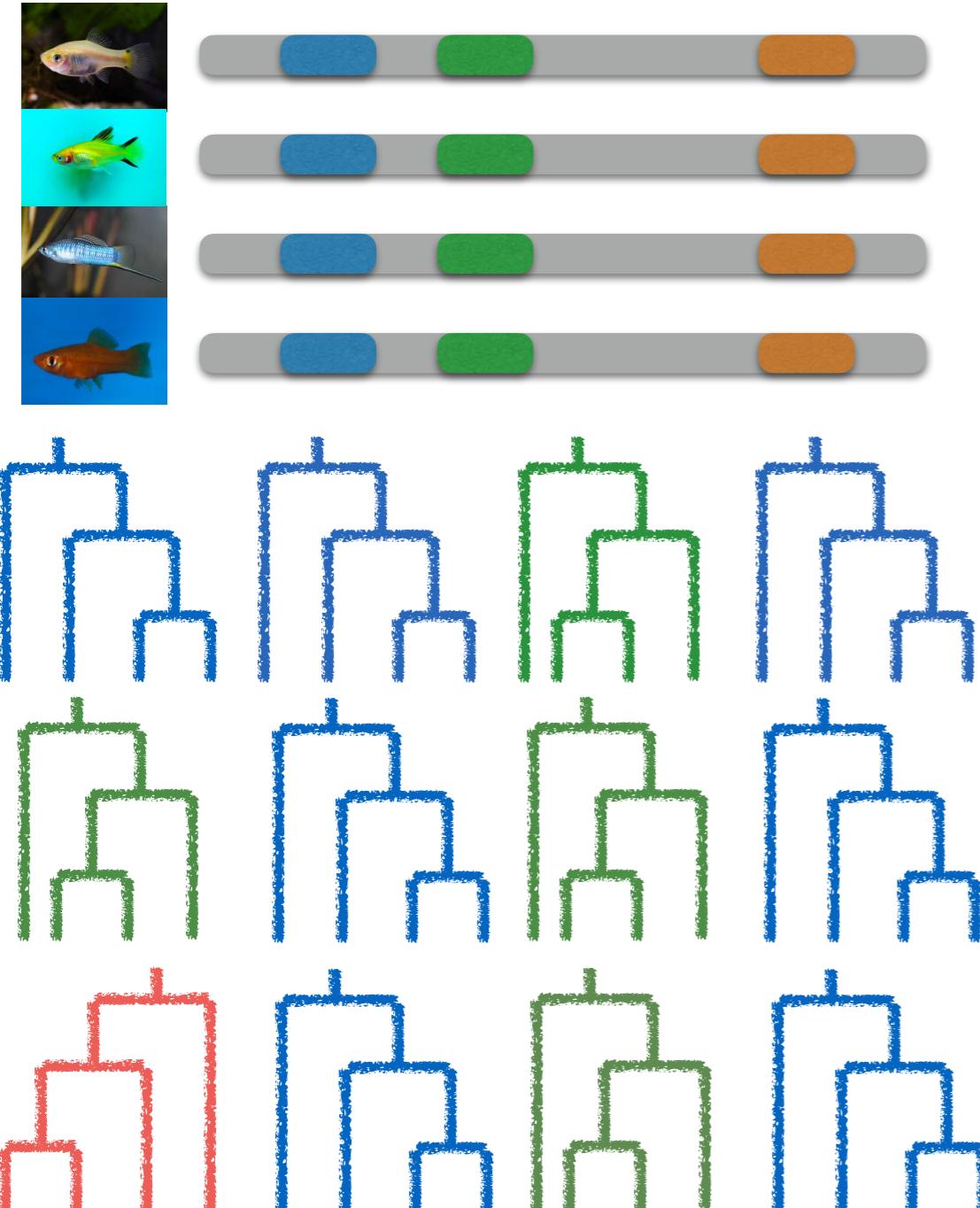
BEAST2
(Zhang et al, 2017)
PhyloNet
(Wen et al, 2016)



SNaQ
(S.-L., Ane, 2016)
PhyloNet
(Yu et al, 2014)

How?

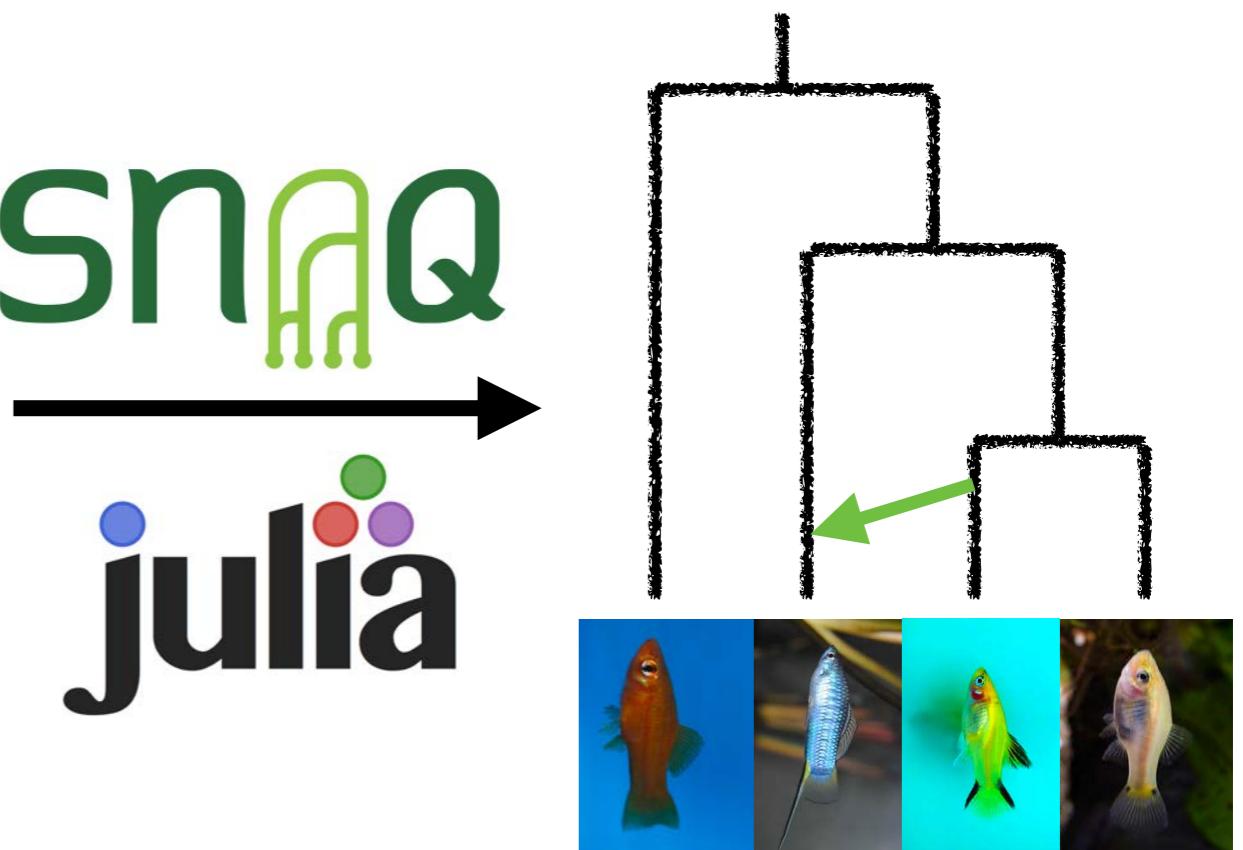
Phylogenetic network



Data

Maximum
pseudolikelihood

snaQ
→
julia



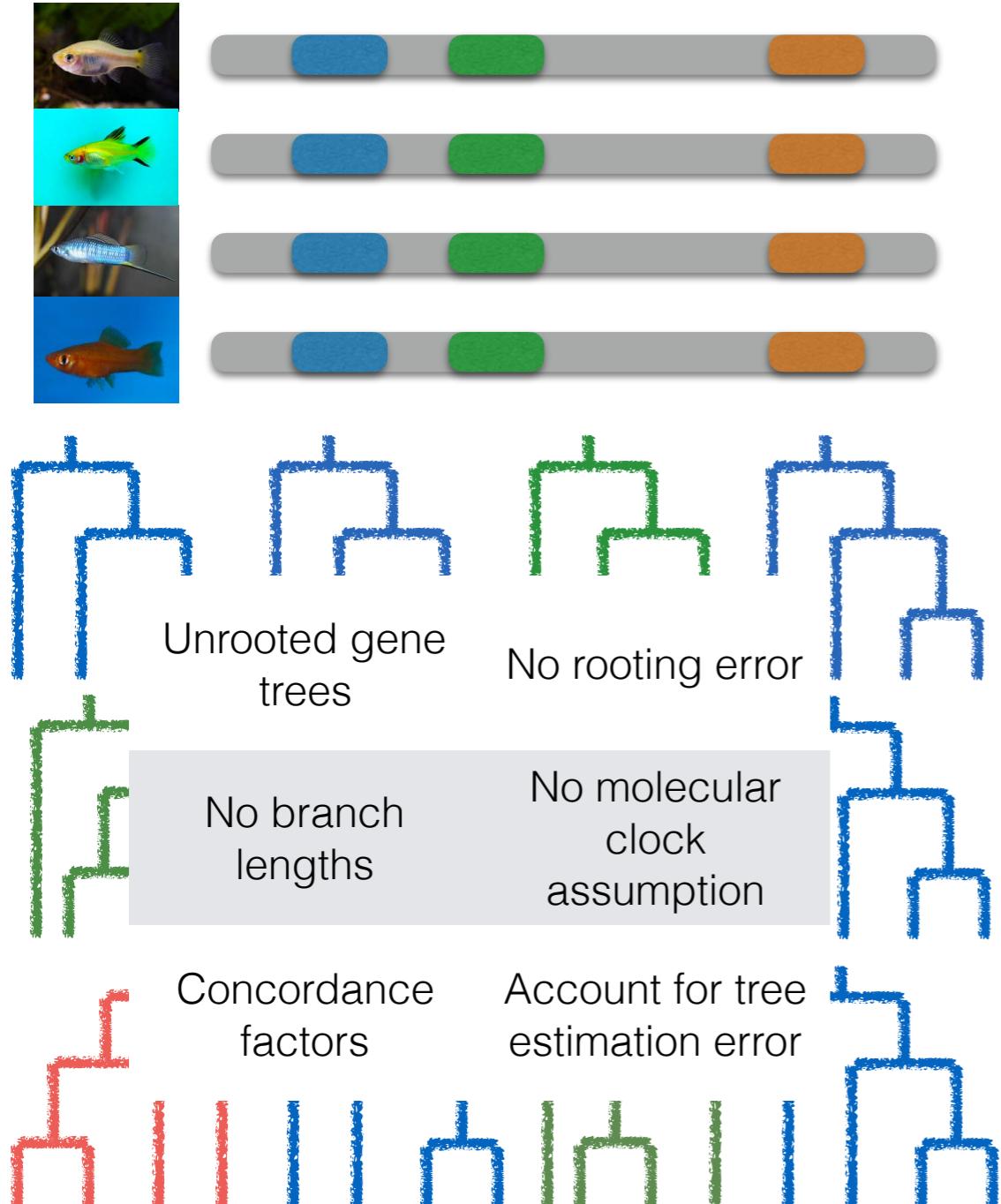
(Solís-Lemus, Ané, 2016, PLoS Genetics)

(Solís-Lemus et al, 2017, MBE)

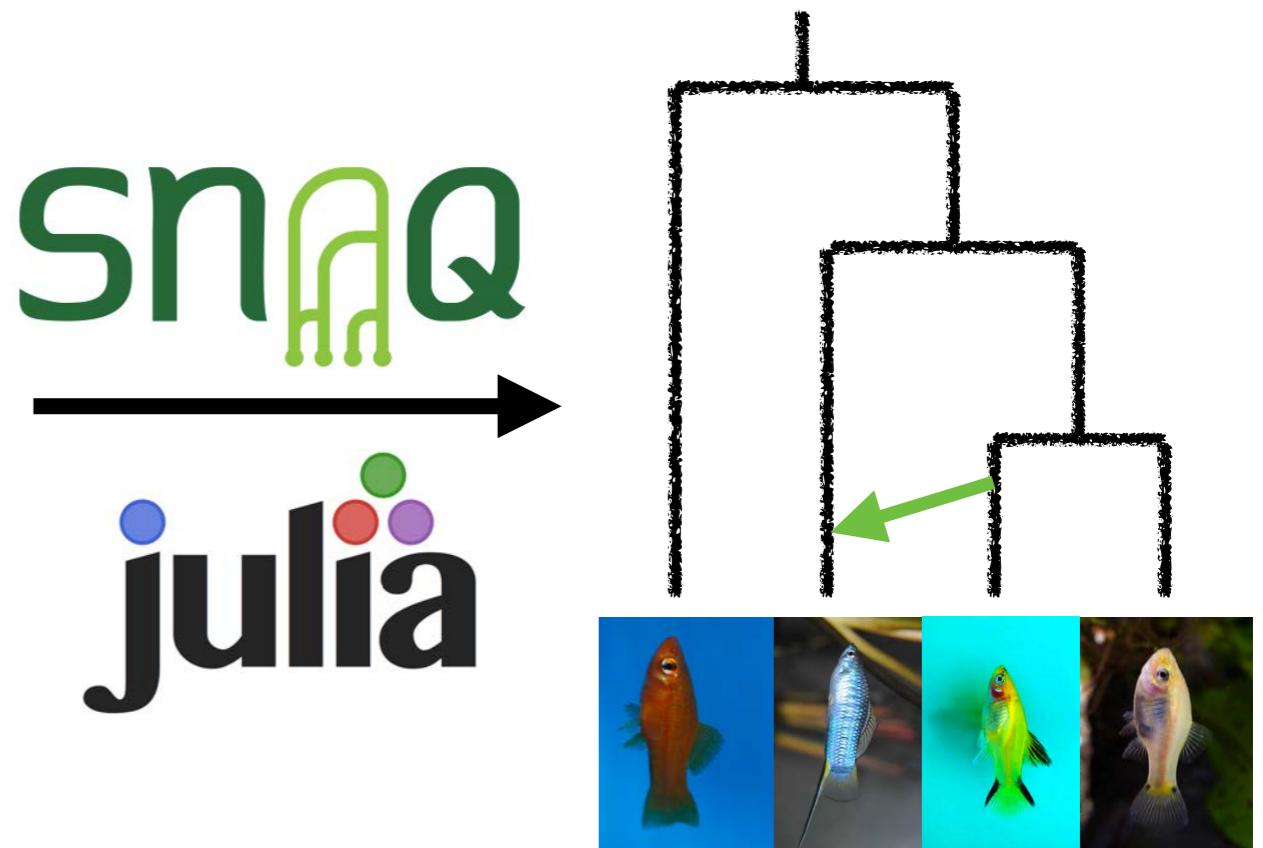
www.github.com/CRSL4/PhyloNetworks

How?

Phylogenetic network



Maximum
pseudolikelihood

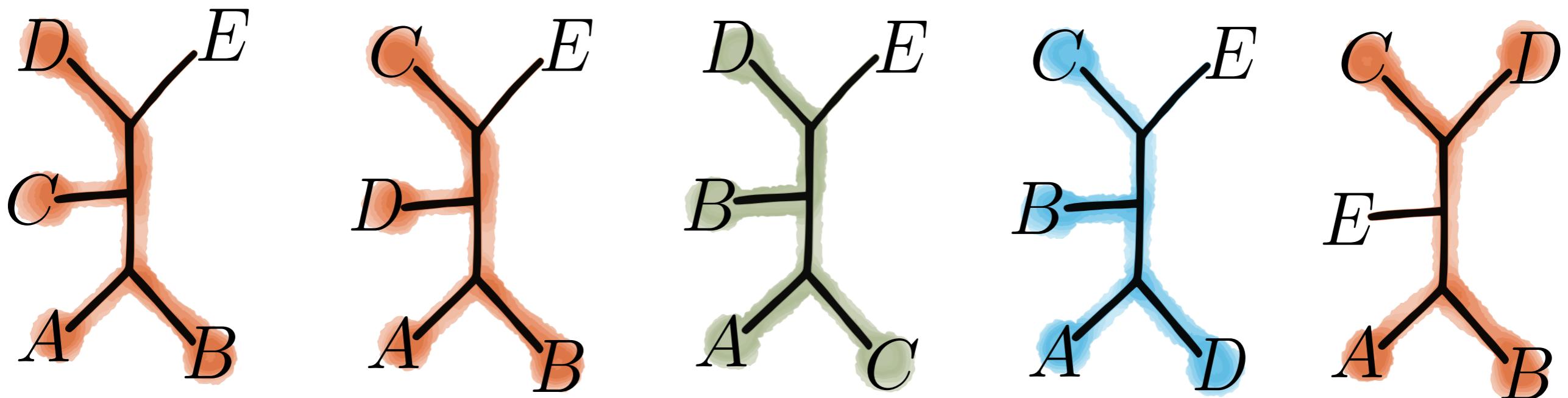


(Solís-Lemus, Ané, 2016, PLoS Genetics)

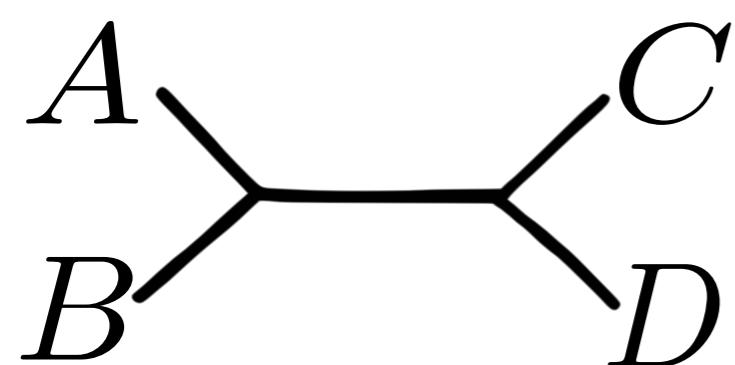
(Solís-Lemus et al, 2017, MBE)

www.github.com/CRSL4/PhyloNetworks

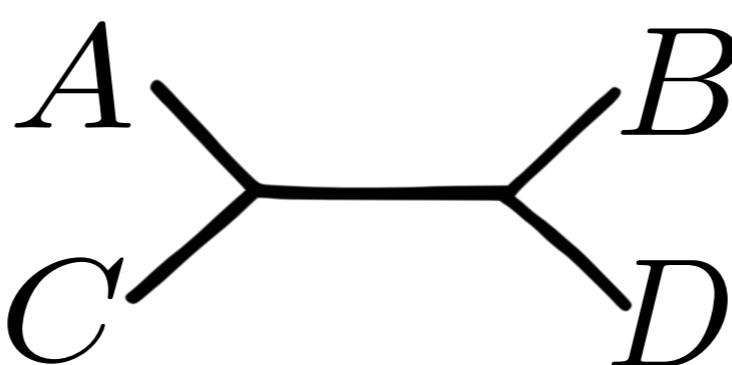
Quartet-based inference



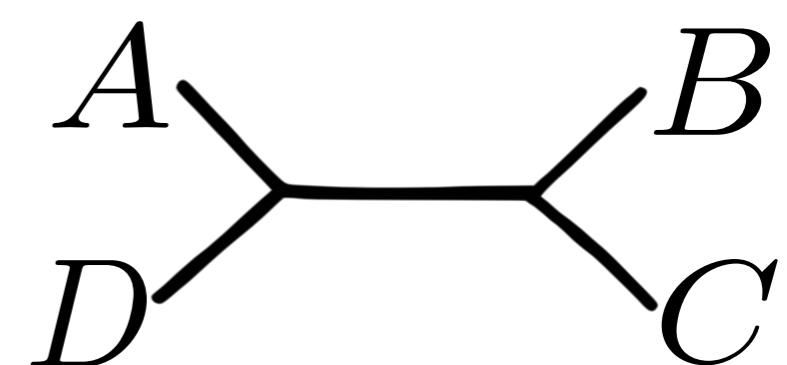
Concordance factors (CF):
% of genes having the quartet in their tree



3/5



1/5



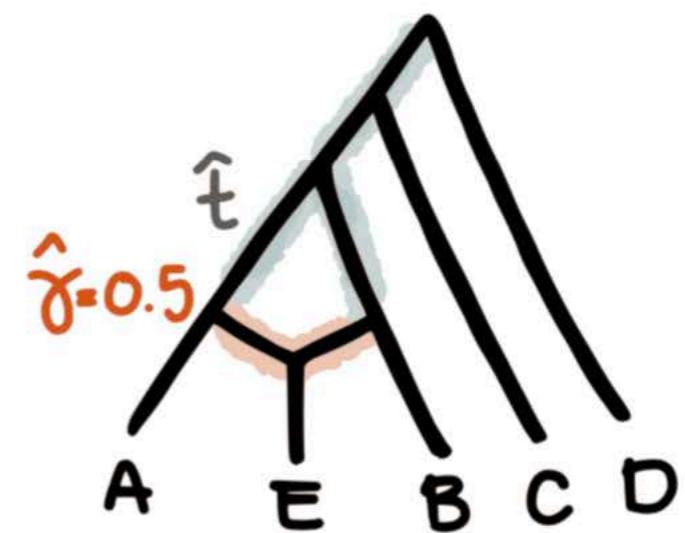
1/5

Quartet-based inference

Observed **quartet** CFs:

4 taxon set	CF_1	CF_2	CF_3
A B C D	.80	.10	.10
A B C E	.40	.40	.20
A B D E	.40	.40	.20
A C D E	.84	.08	.08
B C D E	.82	.10	.08

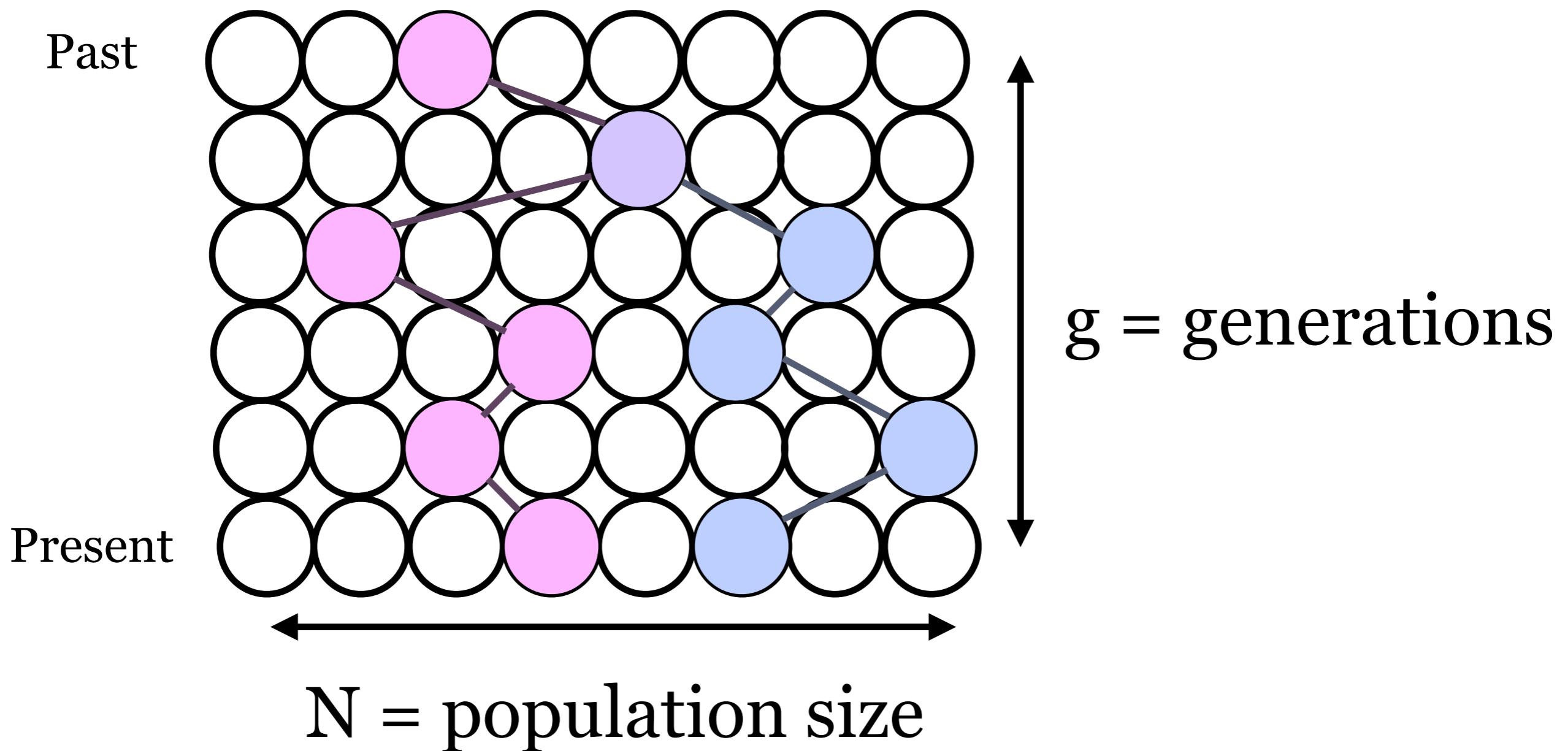
inferred network:



Maximum Pseudo-Likelihood:

$$\log \tilde{L} = \sum_{q \in Q(N)} CF_{in,1} \log(CF_{net,1}) + CF_{in,2} \log(CF_{net,2}) + CF_{in,3} \log(CF_{net,3})$$

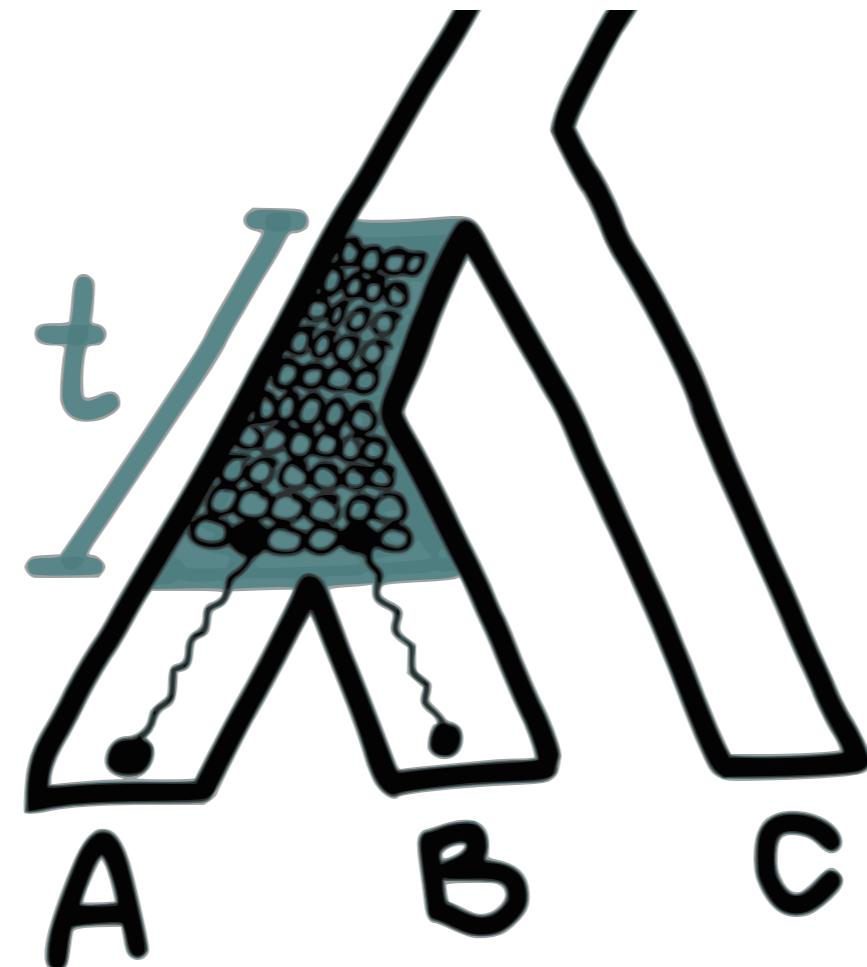
Coalescent model within 1 population



Probability of no coalescence in g generations:

$$\left(1 - \frac{1}{N}\right)^g$$
$$t = g/N \Rightarrow \left(1 - \frac{t}{Nt}\right)^{Nt} \xrightarrow[N \rightarrow \infty]{} e^{-t}$$

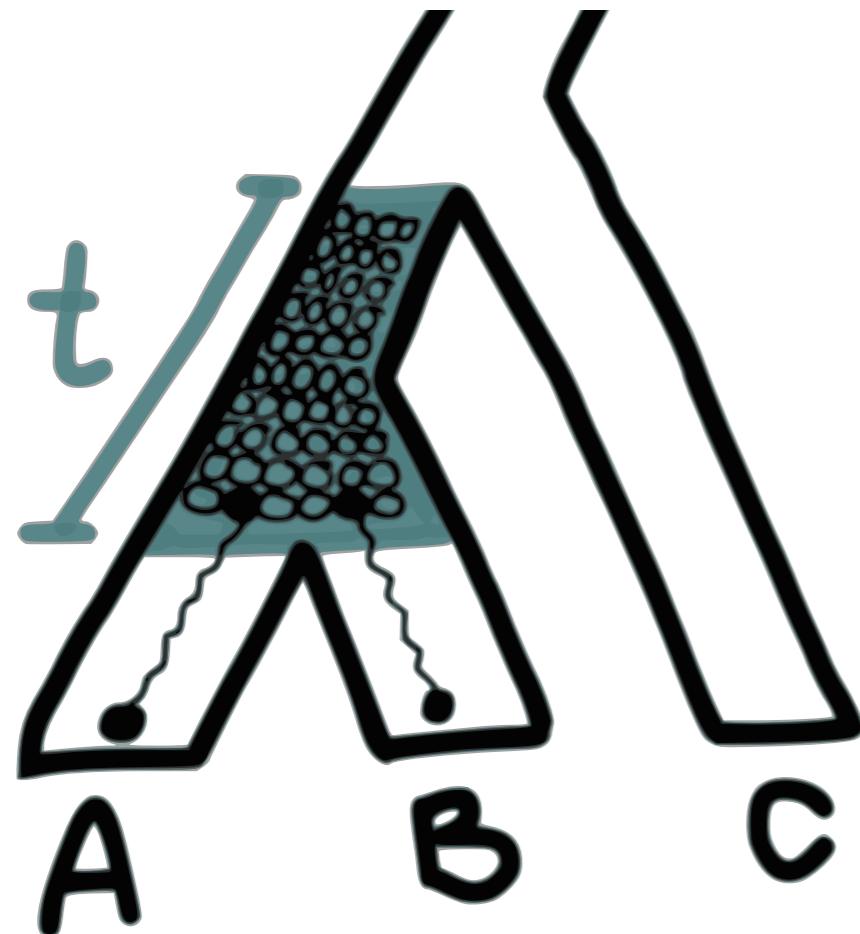
Multispecies coalescent on a tree



$$P(T > t) = e^{-t}$$

$$T = \frac{g}{N} \text{ coalescent units} \sim \text{Exp}(1)$$

Multispecies coalescent on a tree

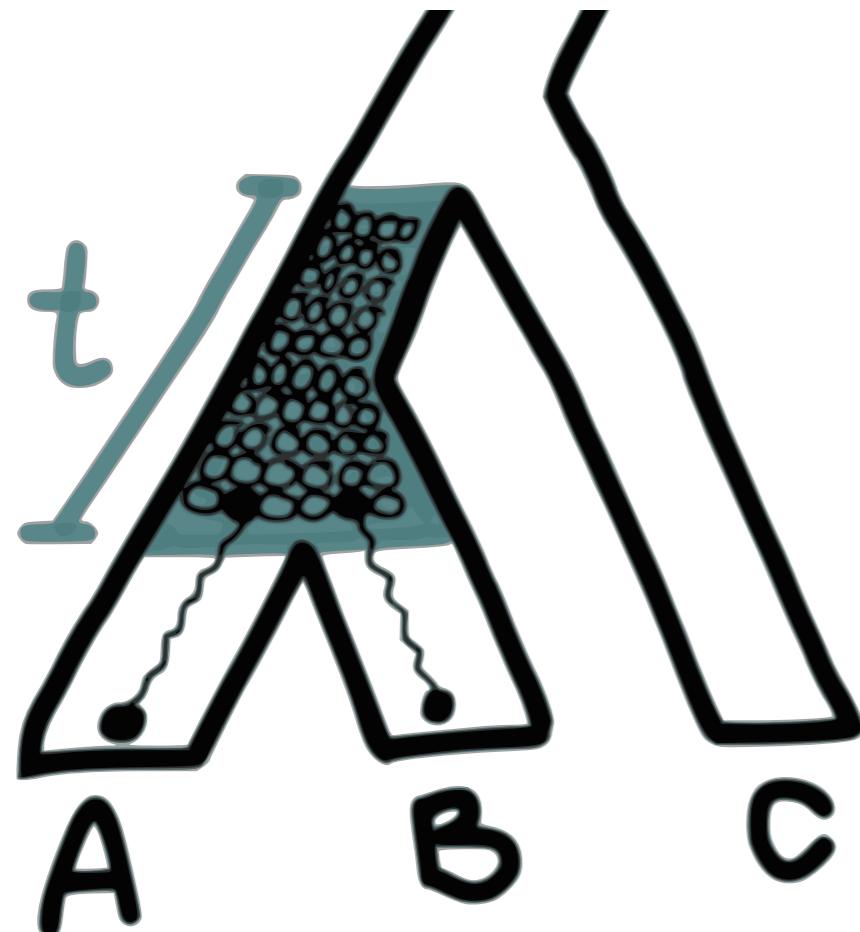


$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ }) =$$

A B C

$$P(T > t) = e^{-t}$$

Multispecies coalescent on a tree

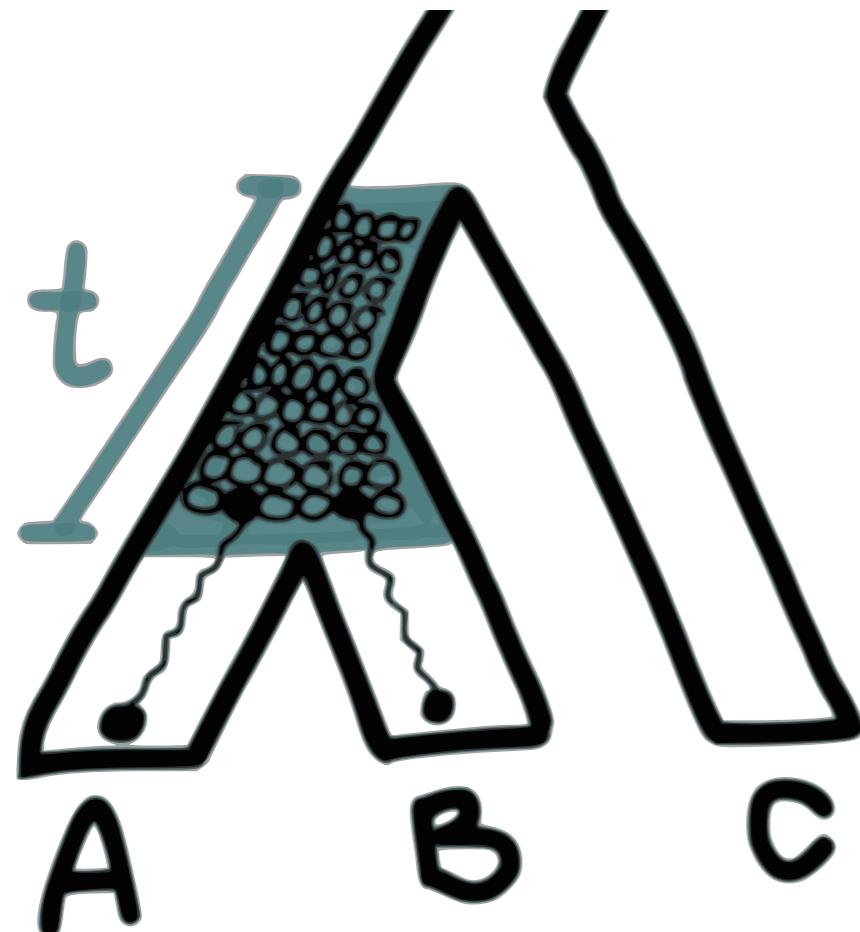


$$P(\text{ } \text{ } \text{ } \text{ }) = 1 - e^{-t}$$

The probability of finding a specific tree topology (in this case, a star-like tree where all species A, B, and C share a single common ancestor) is given by the formula $P(\text{ } \text{ } \text{ } \text{ }) = 1 - e^{-t}$, where t is the time since the most recent common ancestor.

$$P(T > t) = e^{-t}$$

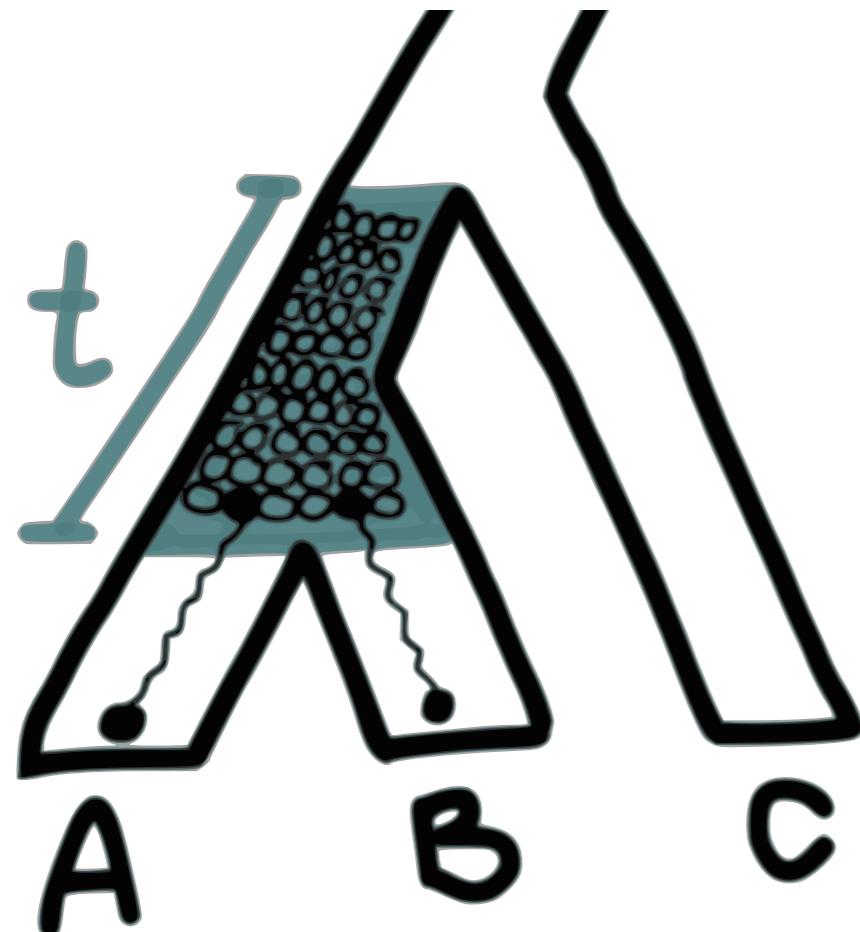
Multispecies coalescent on a tree



$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ }) = \\ 1 - e^{-t} \\ +$$

$$P(T > t) = e^{-t}$$

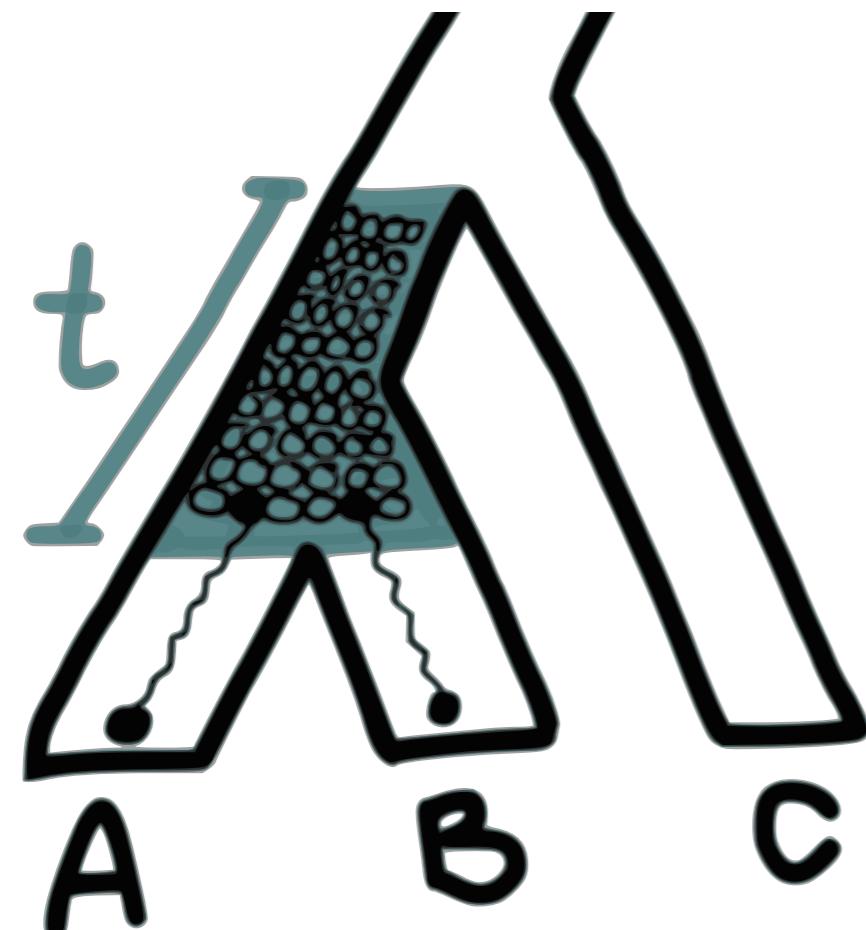
Multispecies coalescent on a tree



$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ }) = \\ 1 - e^{-t} \\ + \\ e^{-t} \times 1/3$$

$$P(T > t) = e^{-t}$$

Multispecies coalescent on a tree



$$P(\text{ } \text{ } \text{ } \text{ } \text{ } \text{ }) =$$

$$1 - e^{-t}$$

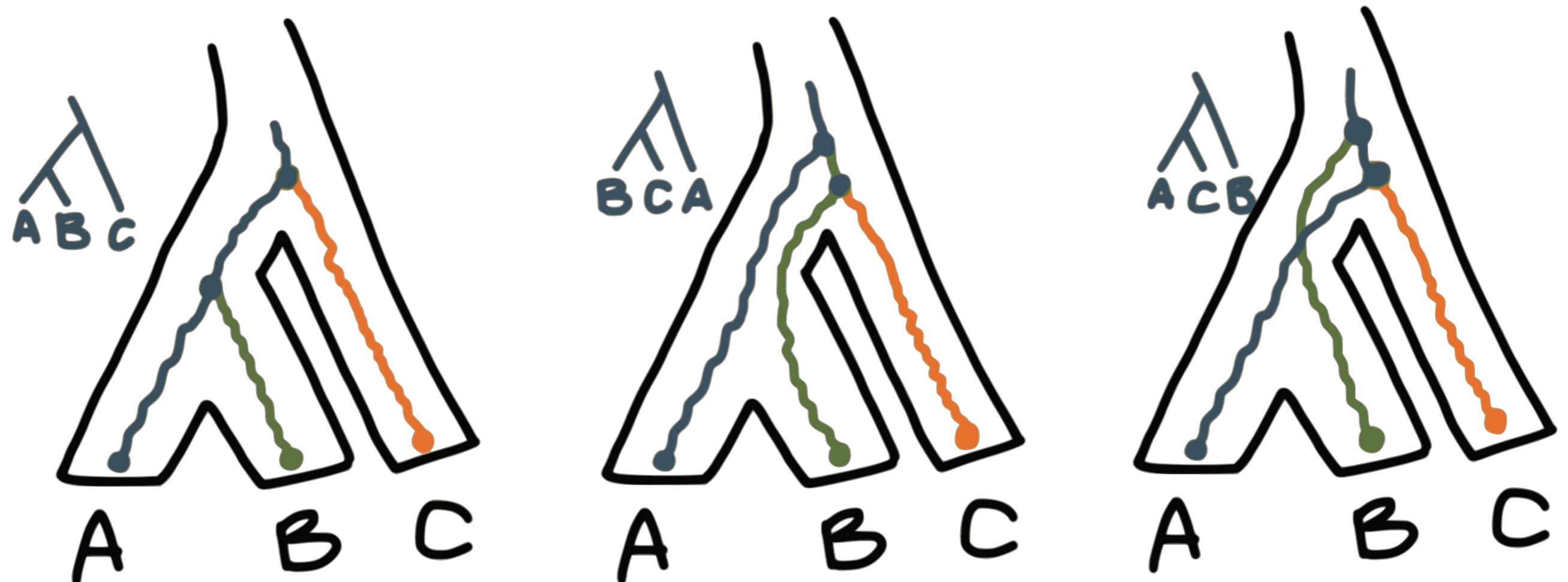
+

$$e^{-t} \times 1/3$$

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

$$P(T > t) = e^{-t}$$

Multispecies coalescent on a tree

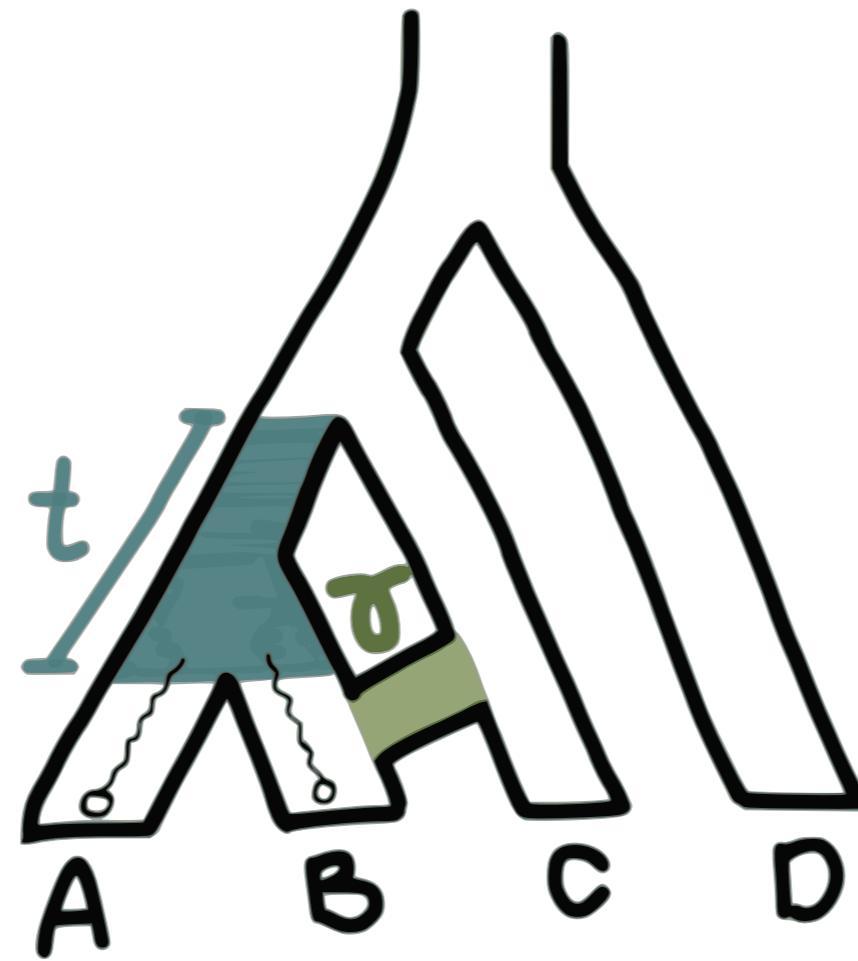


$$1 - \frac{2}{3}e^{-t}$$

$$\frac{1}{3}e^{-t}$$

$$\frac{1}{3}e^{-t}$$

Multispecies coalescent on a network



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

Multispecies coalescent on a network



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

Multispecies coalescent on a network



$$(1 - \gamma) \frac{1}{3} e^{-t} + \gamma (1 - \frac{2}{3} e^{-t_2})$$

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

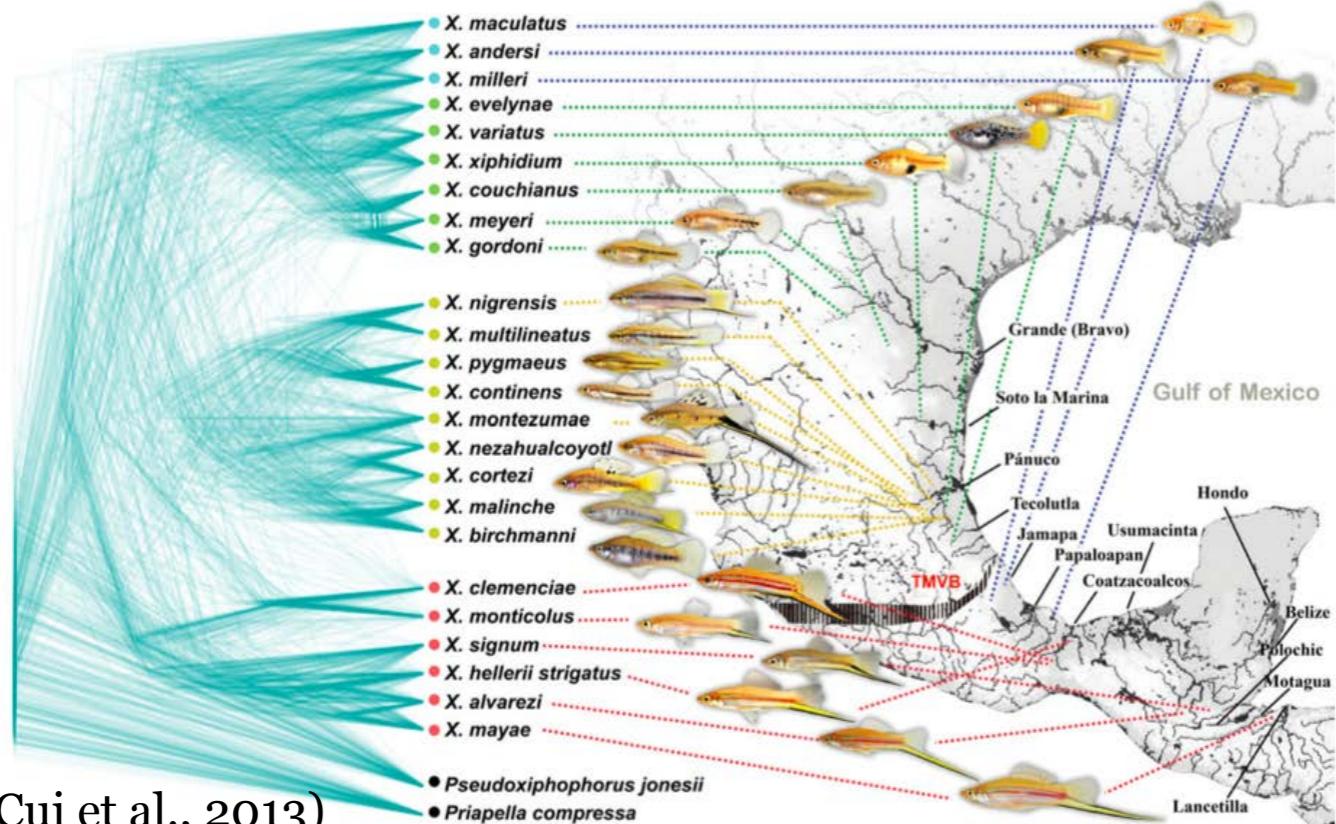
Multispecies coalescent on a network



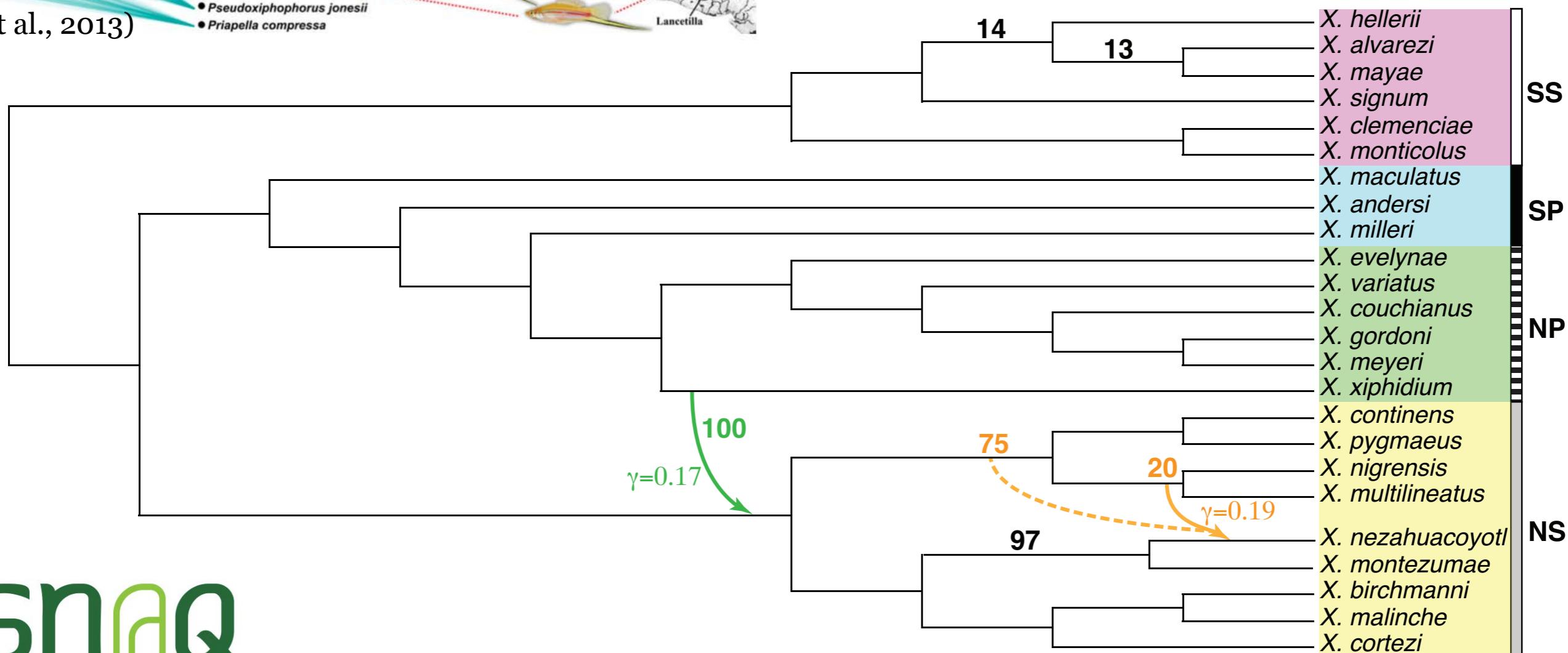
$$CF_{BC|AD}(t, t_2, \gamma) = (1 - \gamma) \frac{1}{3} e^{-t} + \gamma (1 - \frac{2}{3} e^{-t_2})$$

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

1183 genes, 24 swordtails and platyfish



Xiphophorus fish data



snaQ

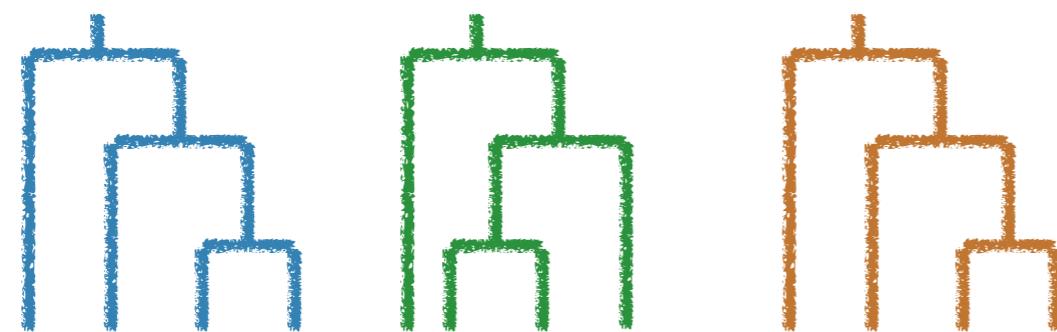
(Solís-Lemus, Ané, 2016, PLoS Genetics)

How?

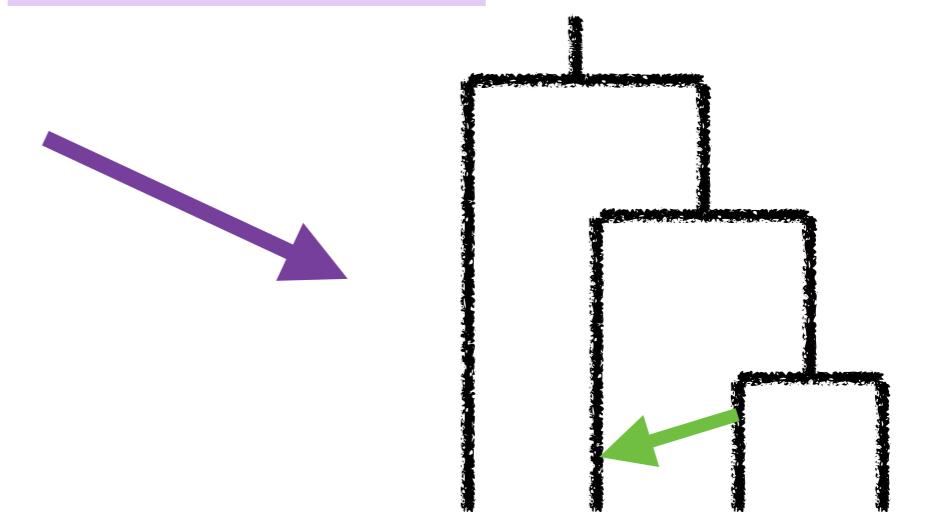
Phylogenetic network



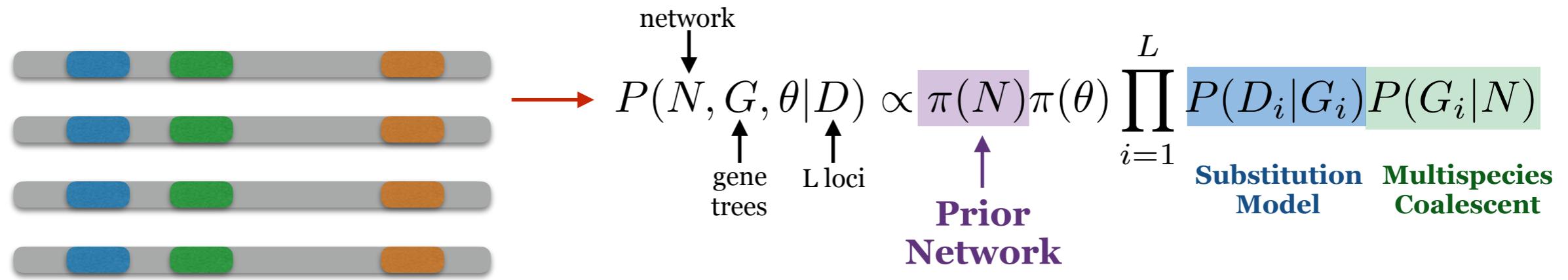
MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
PhyML
(Guindon et al, 2010)



BEAST2
(Zhang et al, 2017)
PhyloNet
(Wen et al, 2016)



SNaQ
(S.-L., Ane, 2016)
PhyloNet
(Yu et al, 2014)

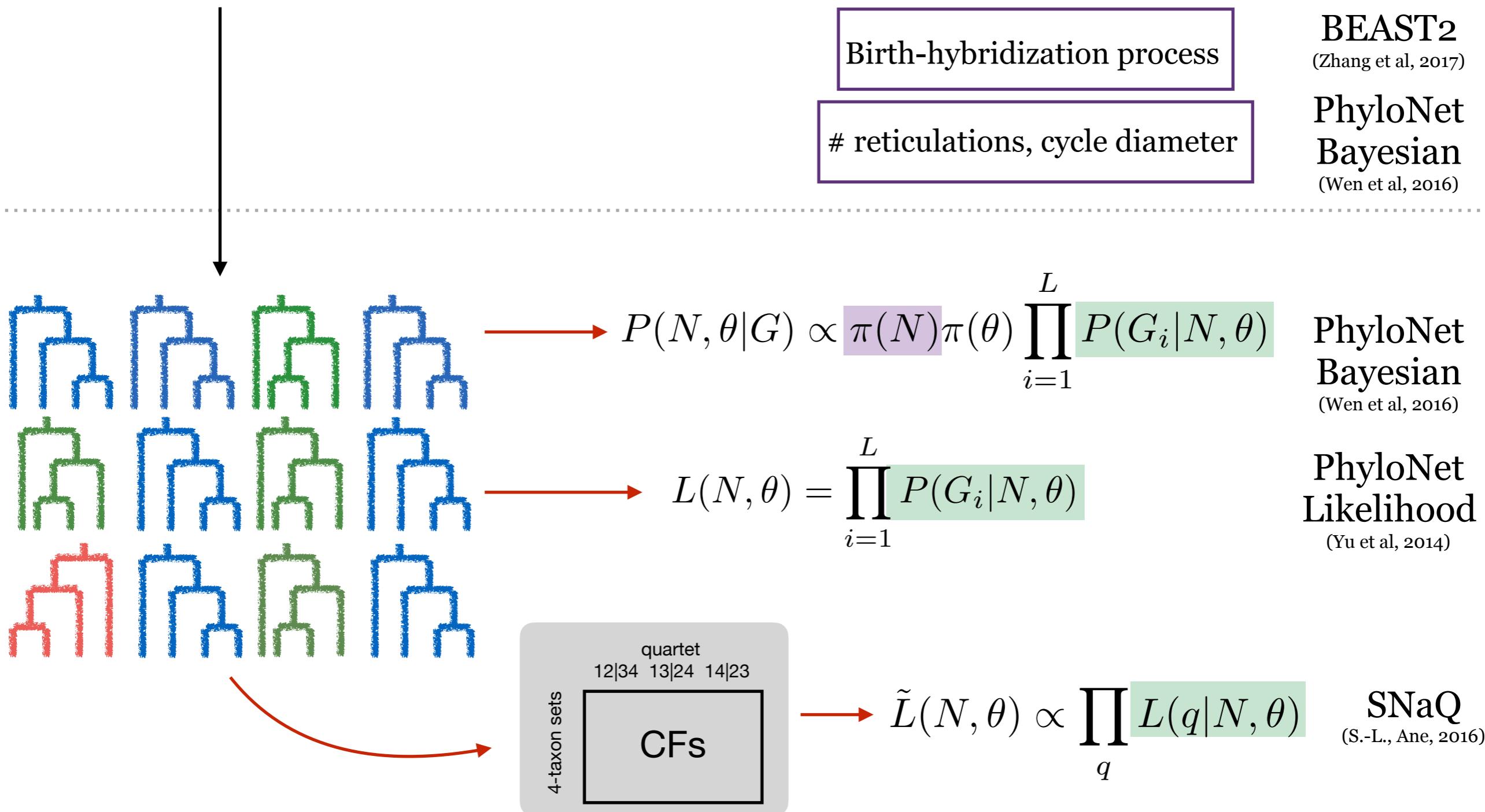
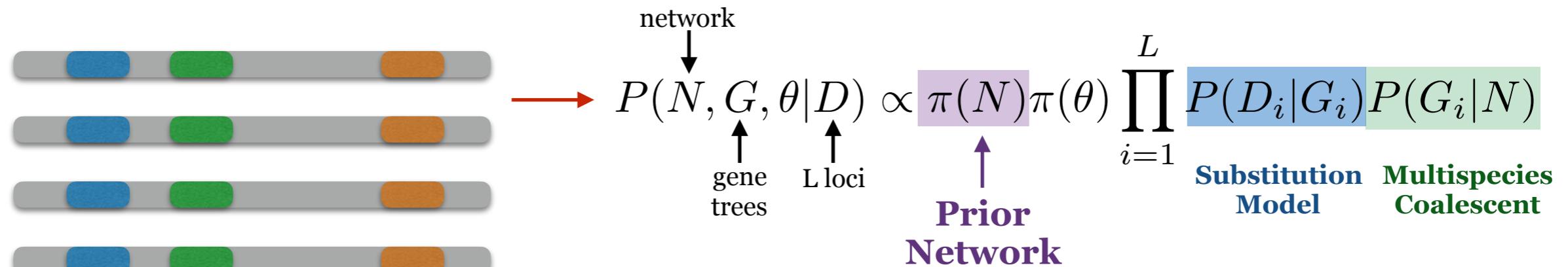


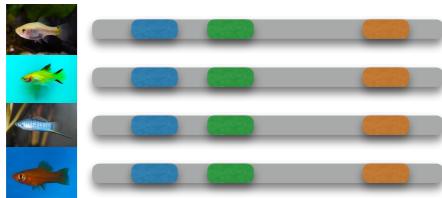
Birth-hybridization process

reticulations, cycle diameter

BEAST2
(Zhang et al, 2017)

**PhyloNet
Bayesian**
(Wen et al, 2016)

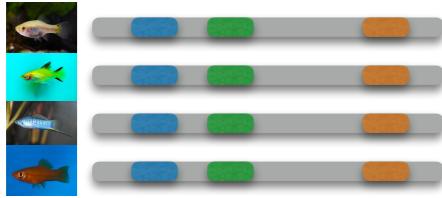




BEAST2
(Zhang et al, 2017)

Birth-hybridization process

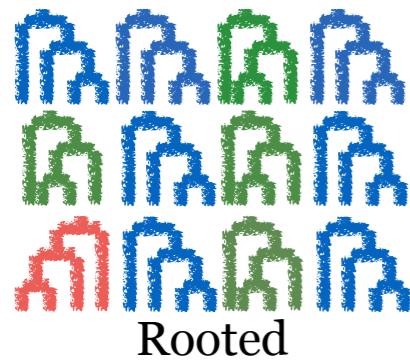
Most accurate,
not scalable



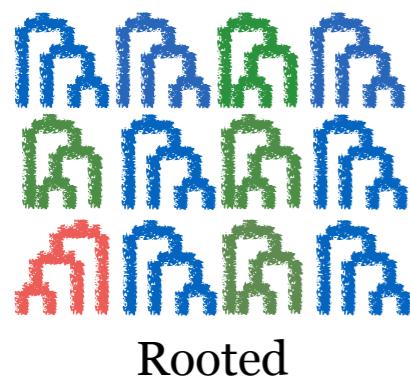
PhyloNet
Bayesian
(Wen et al, 2016)

MCMC:
Network moves,
mixing

reticulations,
cycle diameter

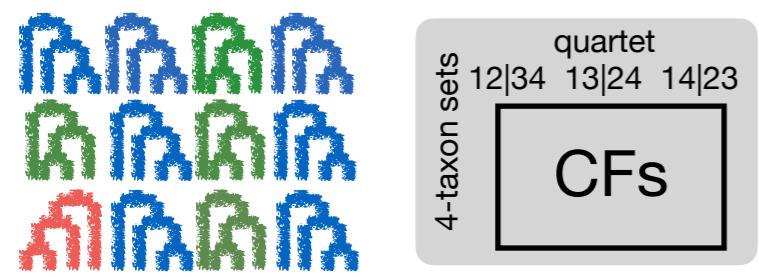


PhyloNet
Bayesian
(Wen et al, 2016)



PhyloNet
Likelihood
(Yu et al, 2014)

Heuristic search:
Network moves



SNaQ
(S.-L., Ane, 2016)

Level-1
networks

More scalable,
Robust

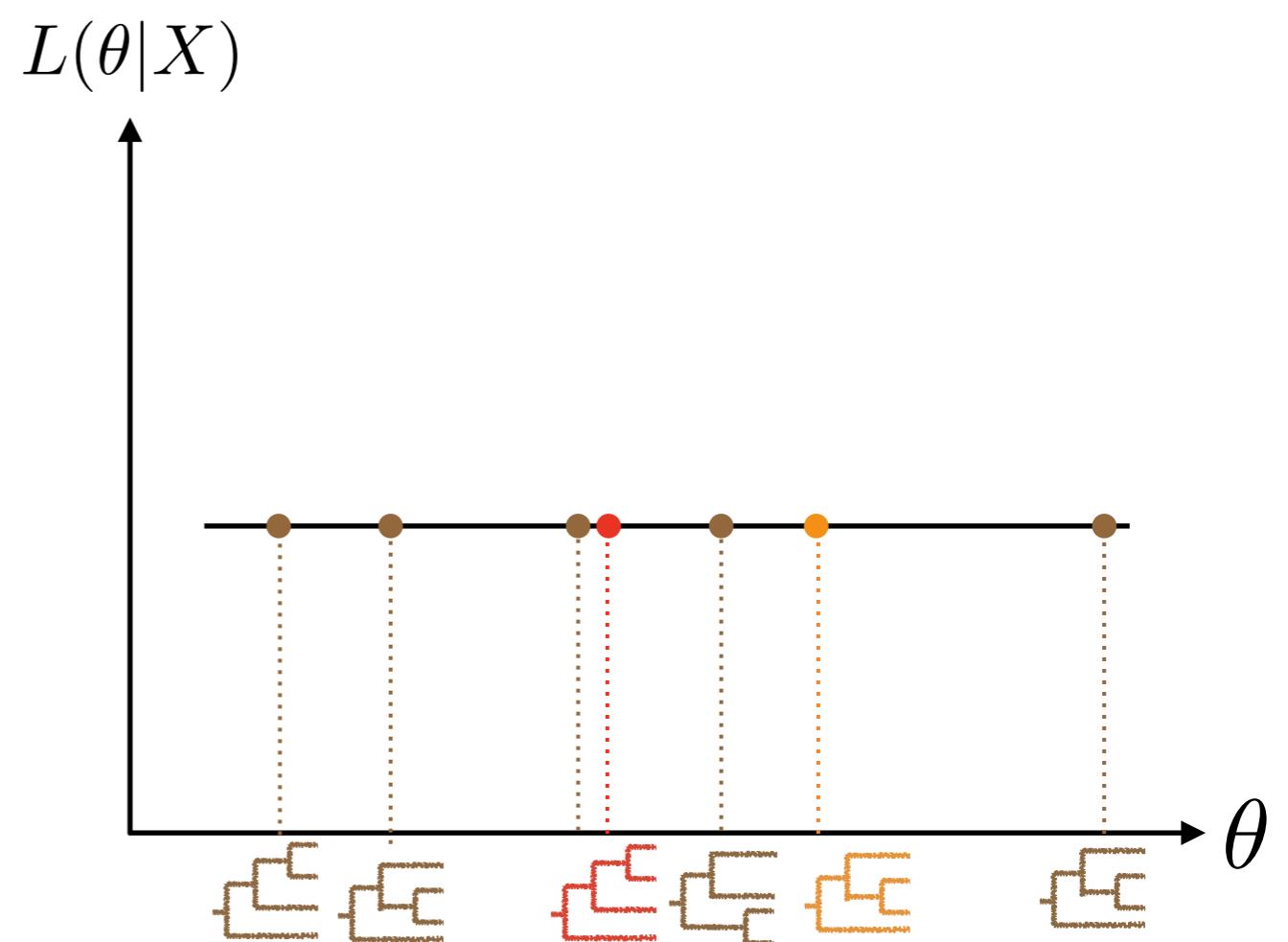
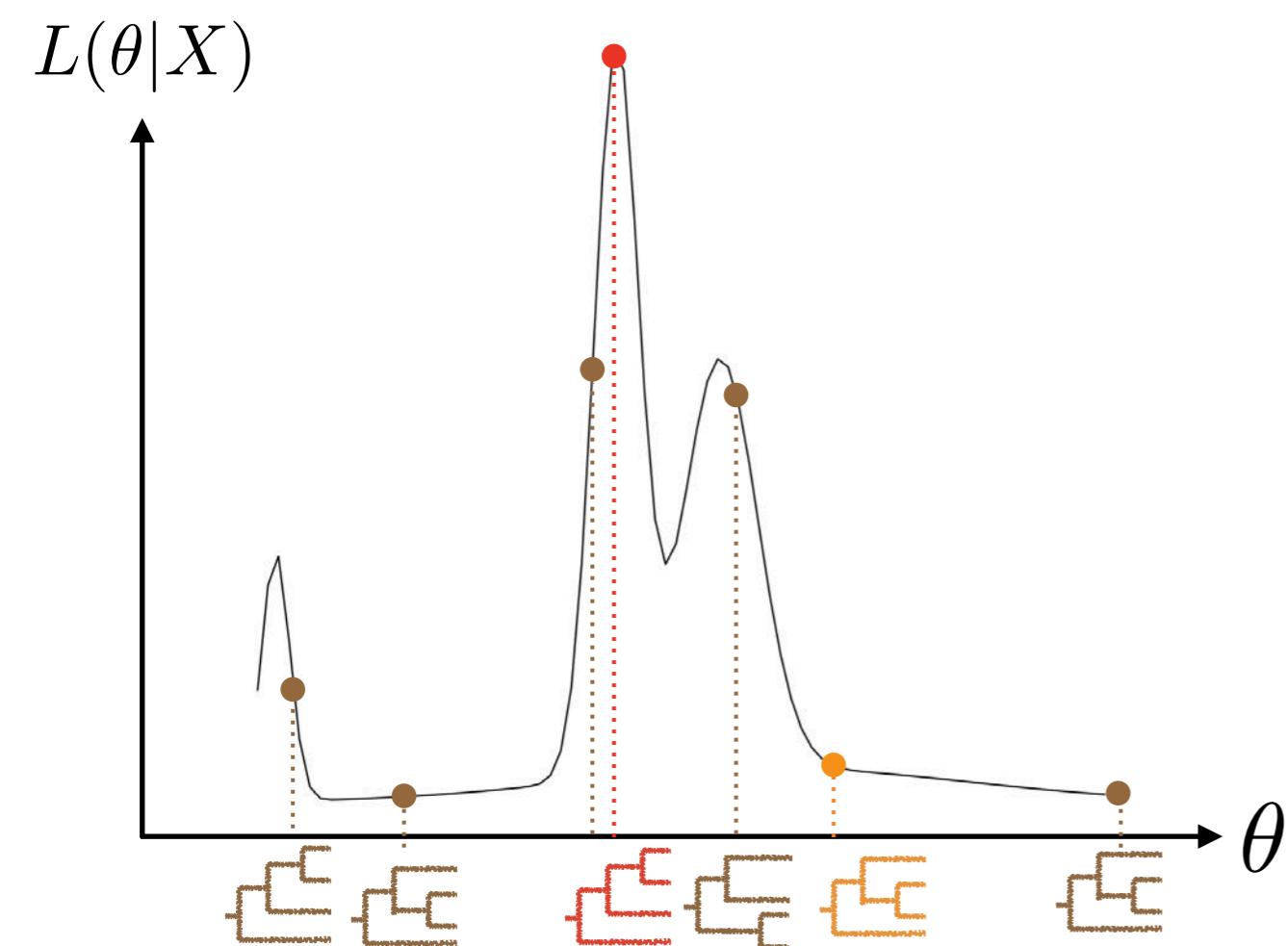
Unrooted

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	
Phylogenetworks 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.

Challenges

- Network space
- Identifiability
- Network comparison

Identifiability

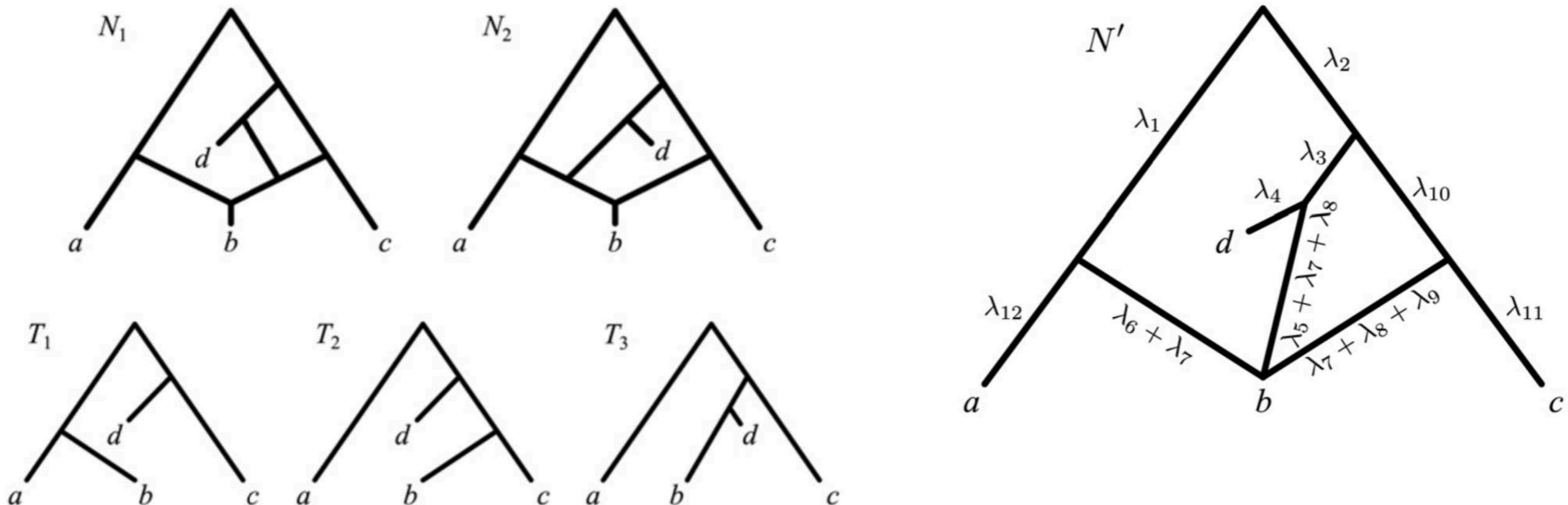


RESEARCH ARTICLE

Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable

Fabio Pardi^{1,3*}, Celine Scornavacca^{2,3}

1 Laboratoire d’Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM, UMR 5506) CNRS, Université de Montpellier, France, **2** Institut des Sciences de l’Evolution de Montpellier (ISE-M, UMR 5554) CNRS, IRD, Université de Montpellier, France, **3** Institut de Biologie Computationalle, Montpellier, France

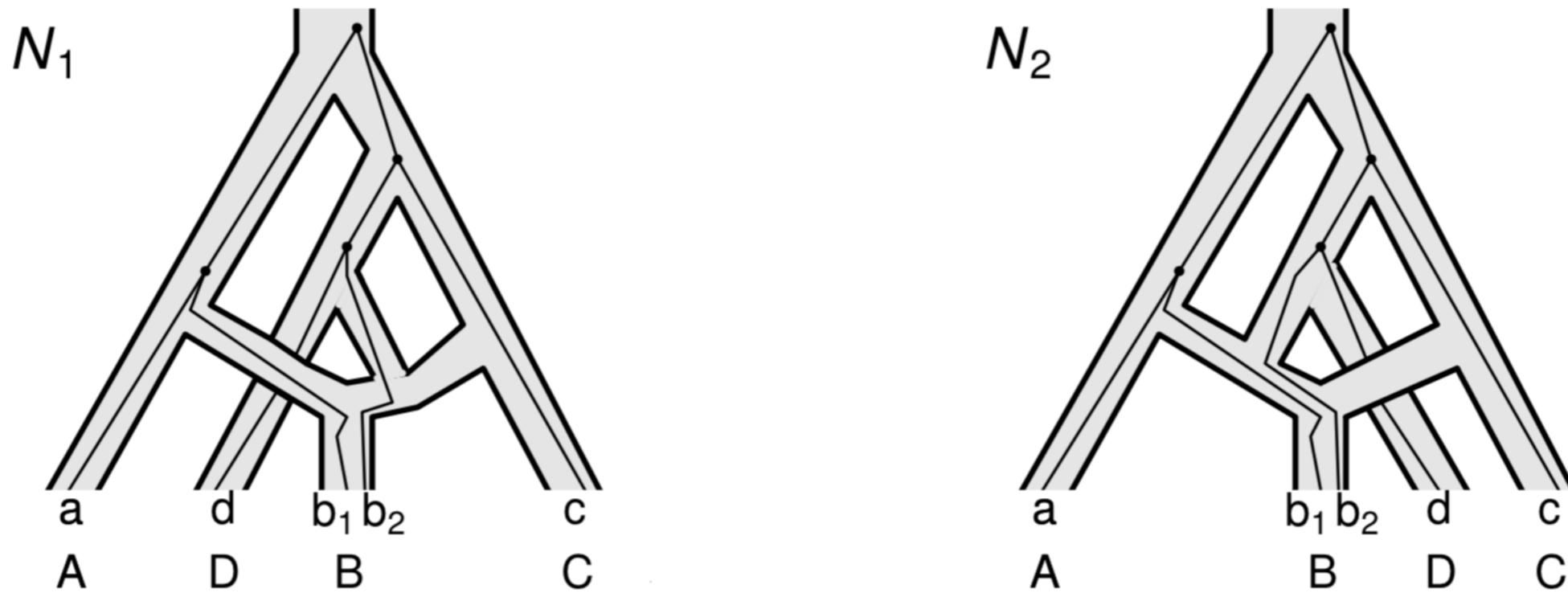


Undistinguishable with the
“displayed trees” criterion

Solution: Canonical
network (“unzipped”)

Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

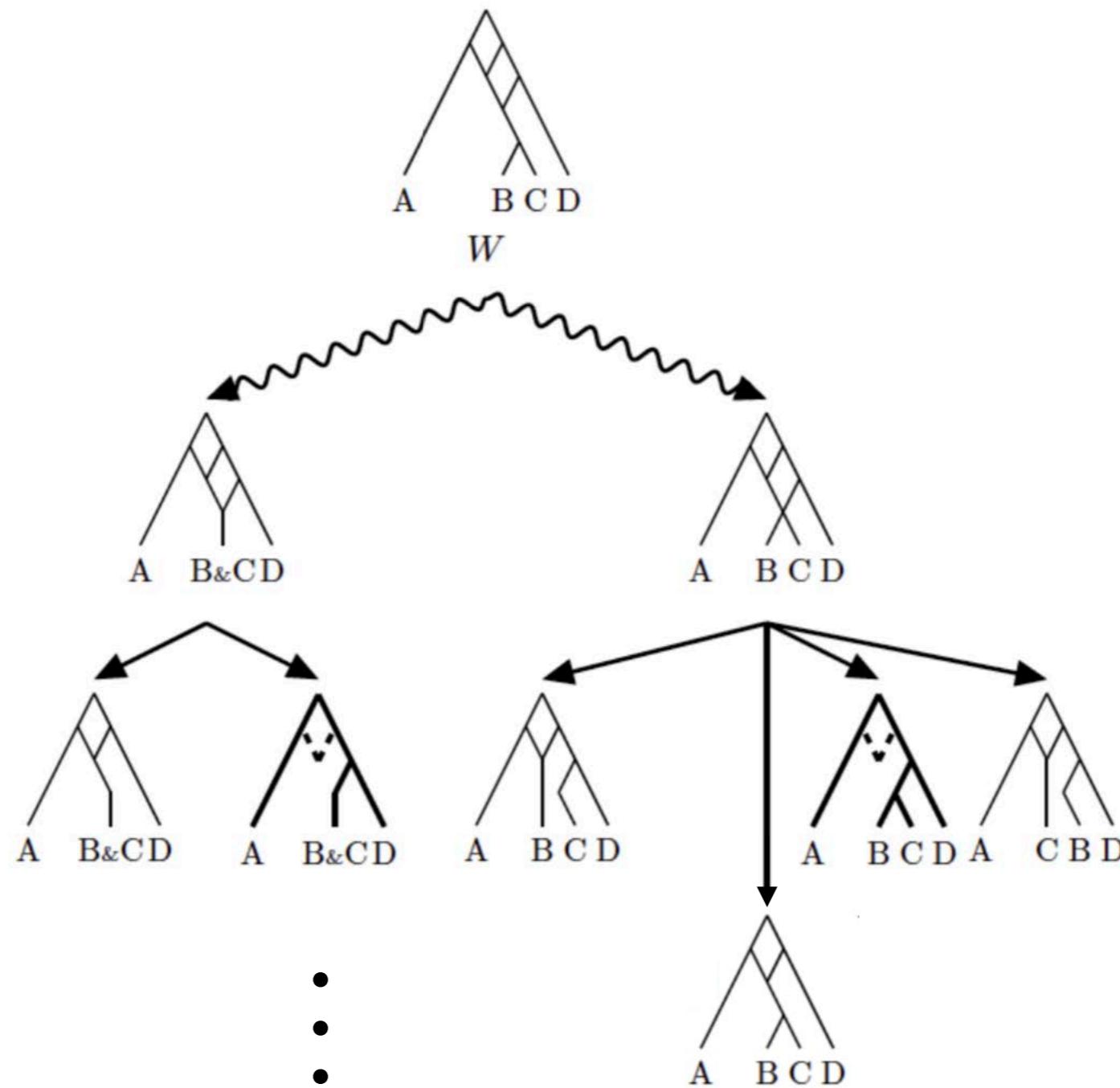
Sha Zhu¹, James H. Degnan²



Distinguishable under the MSC

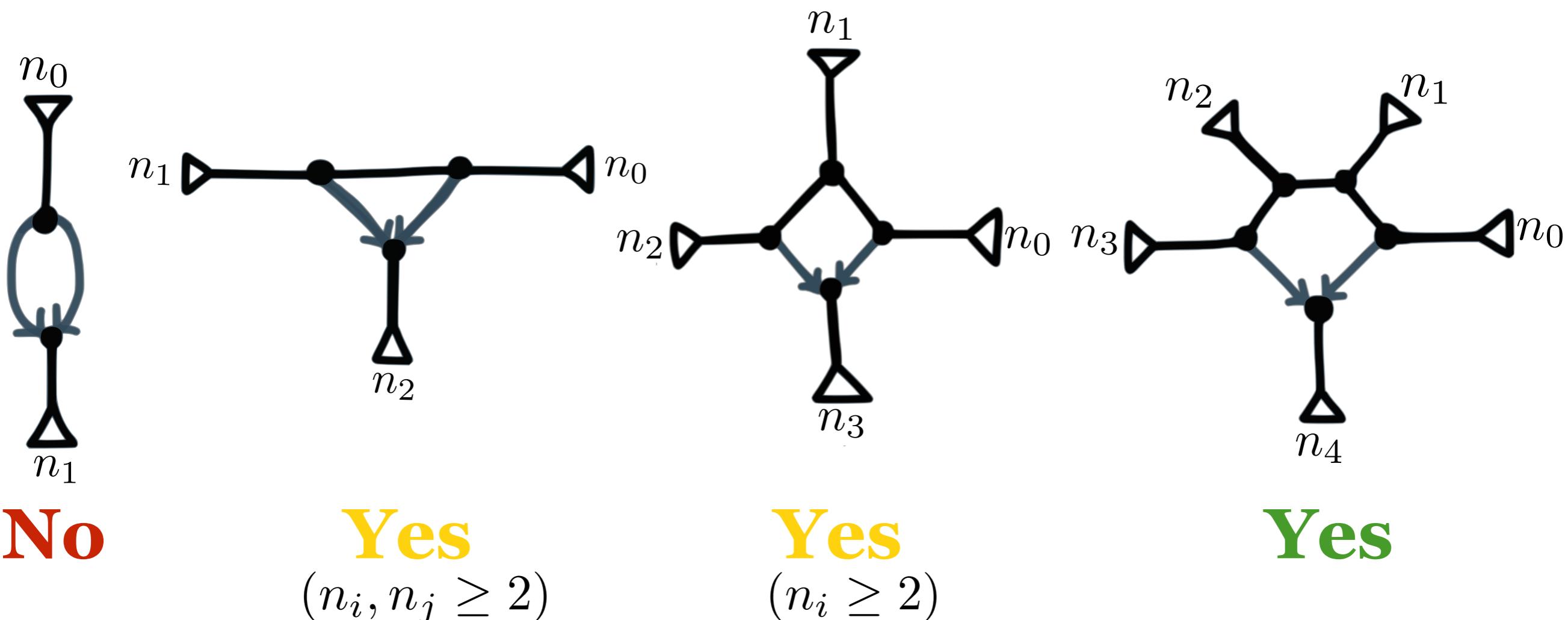
Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

Sha Zhu¹, James H. Degnan²



Decomposing network in **parental** trees

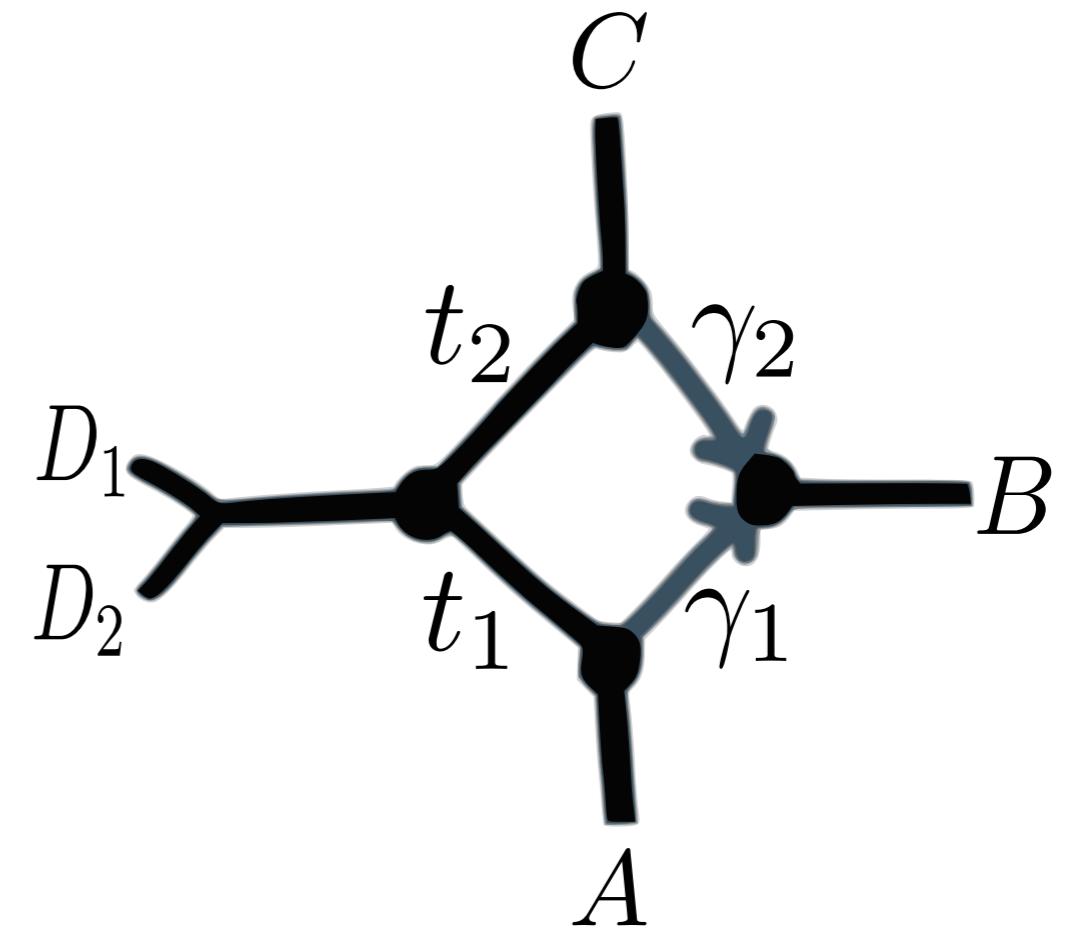
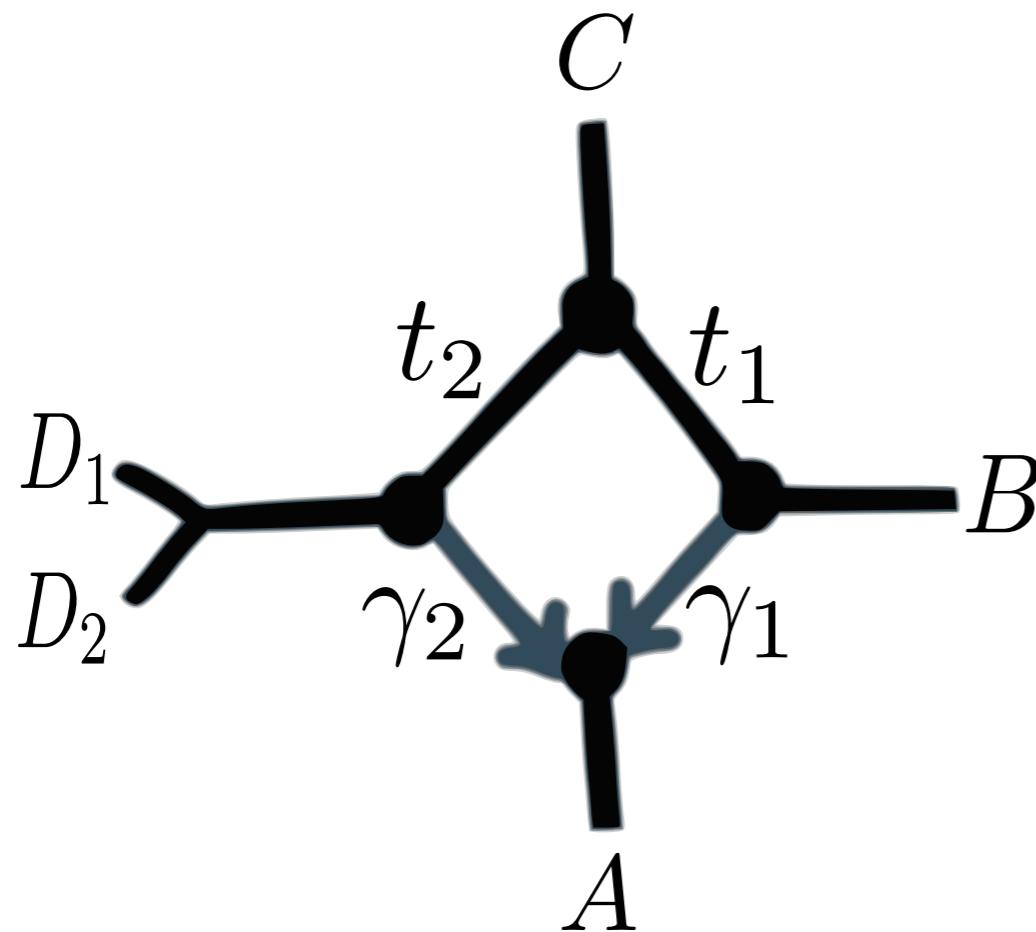
RESEARCH ARTICLE

Inferring Phylogenetic Networks with
Maximum Pseudolikelihood under
Incomplete Lineage SortingClaudia Solís-Lemus^{1*}, Cécile Ané^{1,2}Can we detect the
presence of
hybridization in level-1
networks?

Generic Identifiability

 $t_i \in (0, \infty), \gamma \in (0, 1)$

In practice: flat pseudolikelihood

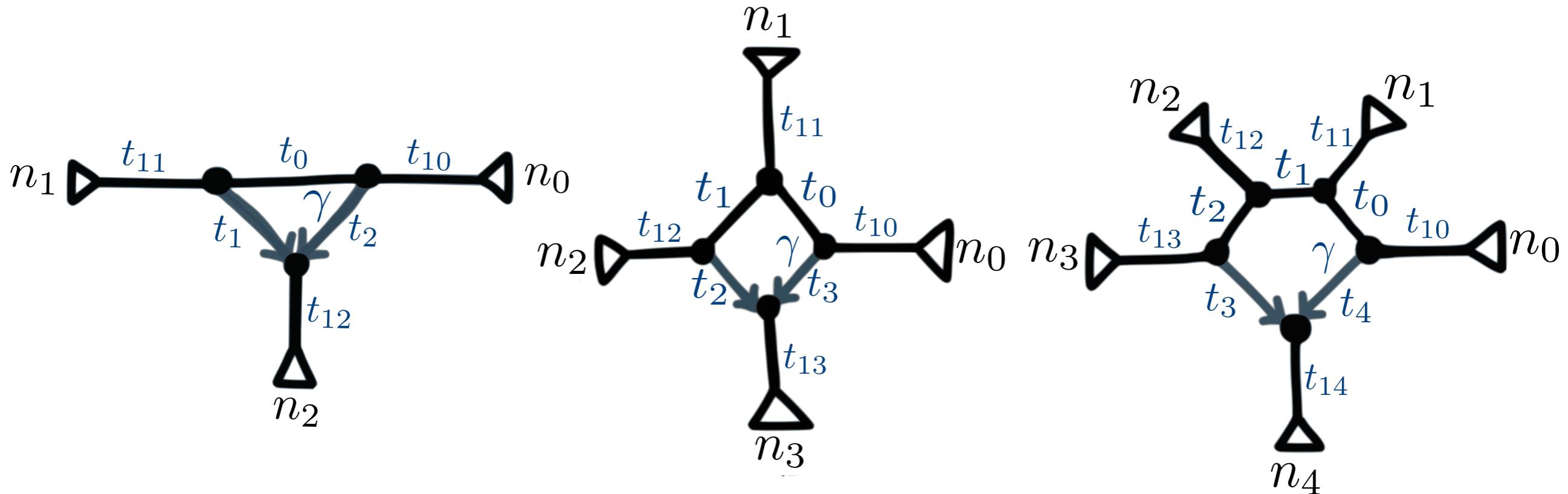


Can we estimate numerical parameters?

RESEARCH ARTICLE

Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting

Claudia Solís-Lemus^{1*}, Cécile Ané^{1,2}



No

Good triangle
($t_{12} = 0$)

Yes

Good diamond
($n_0, n_2 \geq 2$)

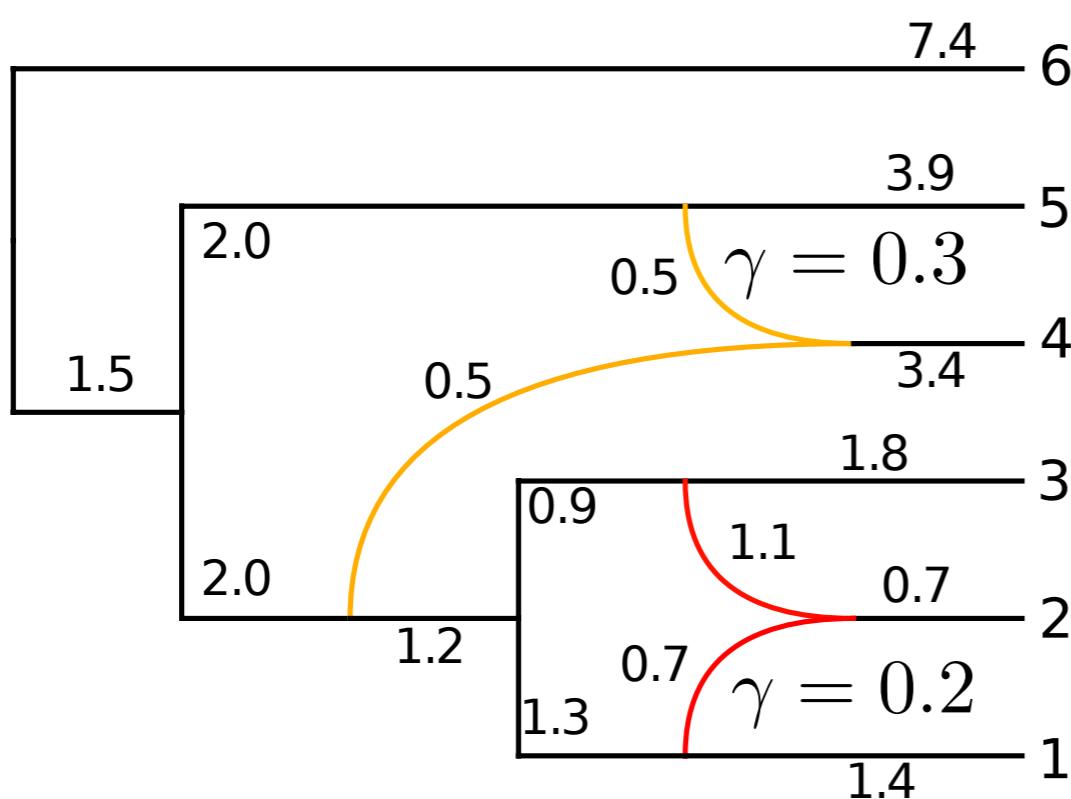
Yes

Generic Identifiability

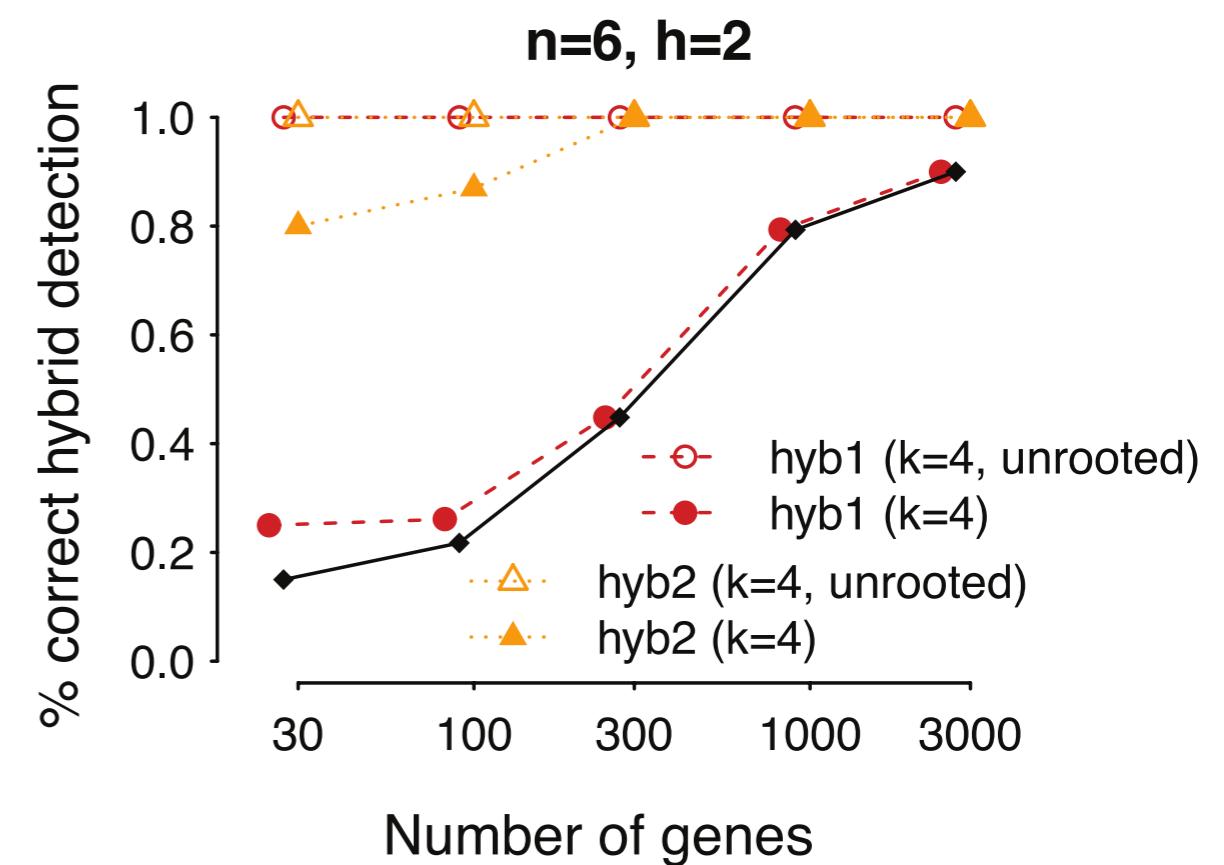
$t_i \in (0, \infty), \gamma \in (0, 1)$

Identifiability matters: SNaQ performance

Good diamond



Bad diamond



Challenges

- Network space

- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

Challenges

- Network space

K. Huber, V. Moulton, C. Scornavacca,...
Missing: path through tree space, semi-directed

- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

Challenges

- Network space

K. Huber, V. Moulton, C. Scornavacca,...
Missing: path through tree space, semi-directed

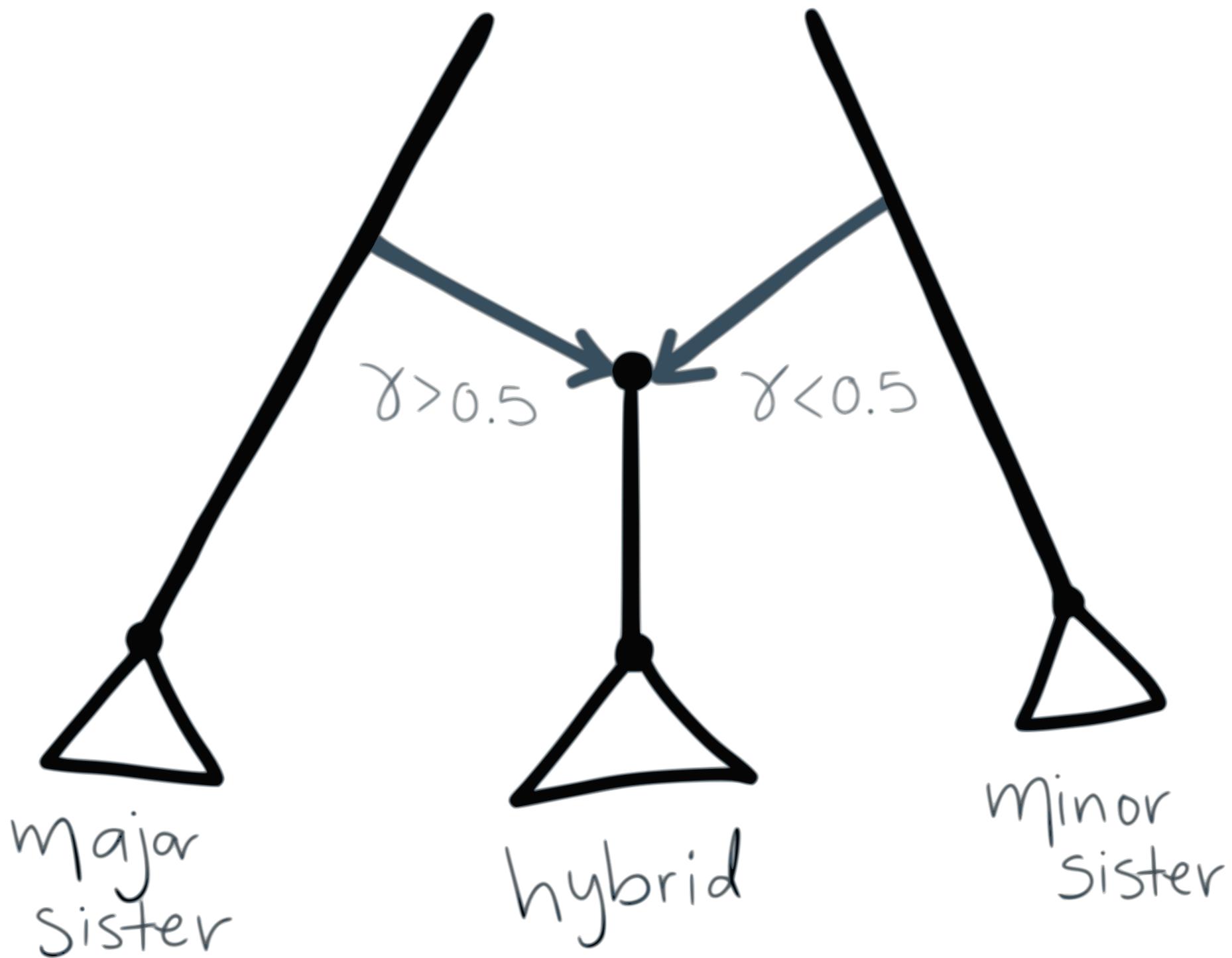
- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

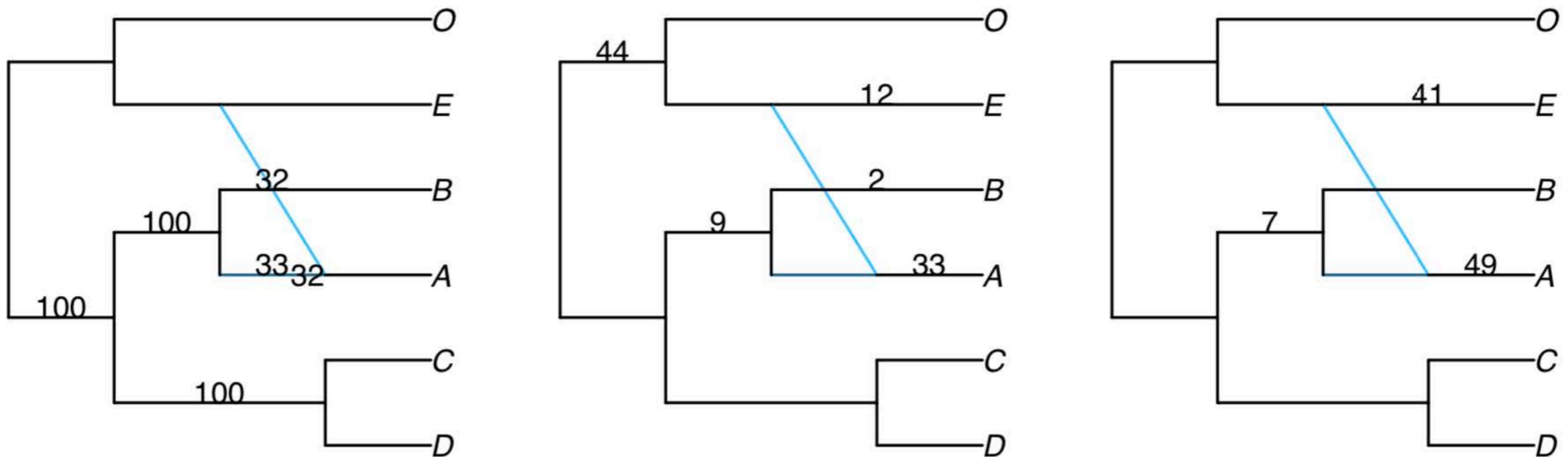

Missing: distance function
Hardwired-cluster distance only for rooted networks
Summary of networks: clades!

Network summary



(S.-L. et al, 2017, MBE)

Network summary

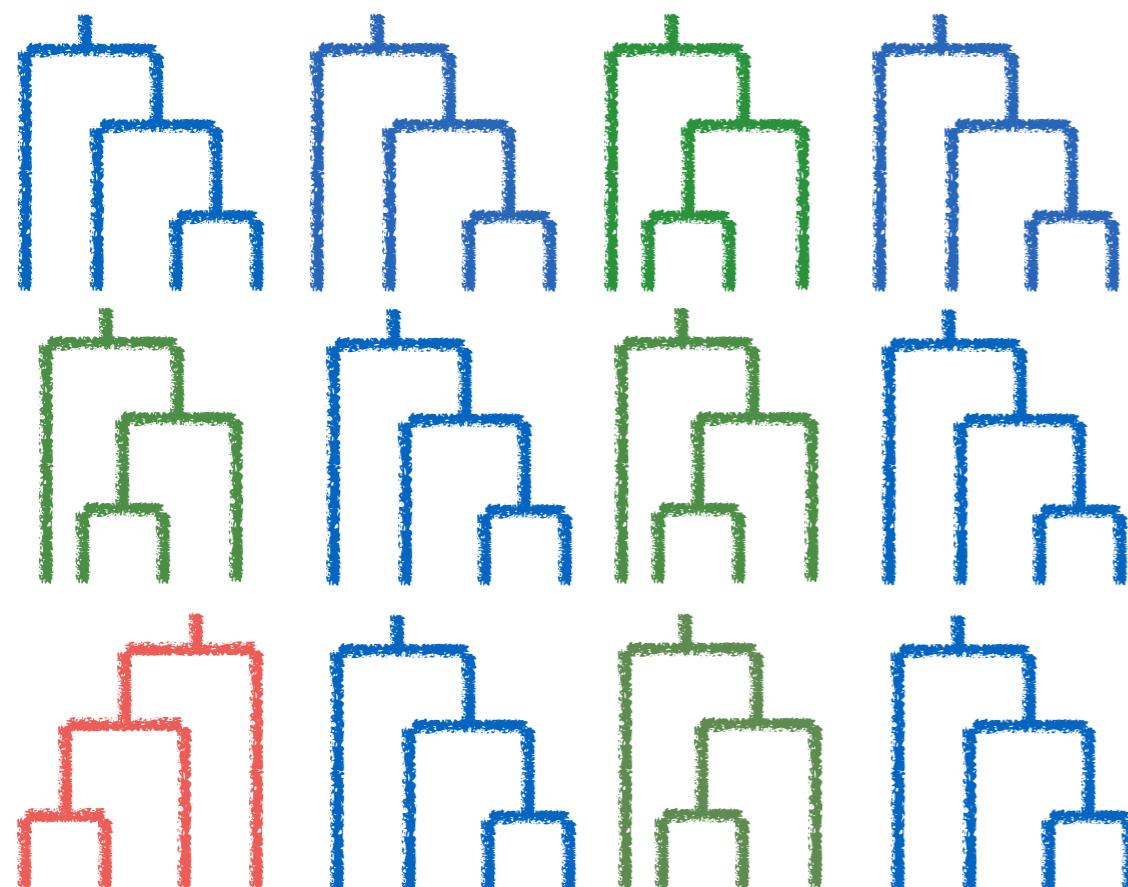
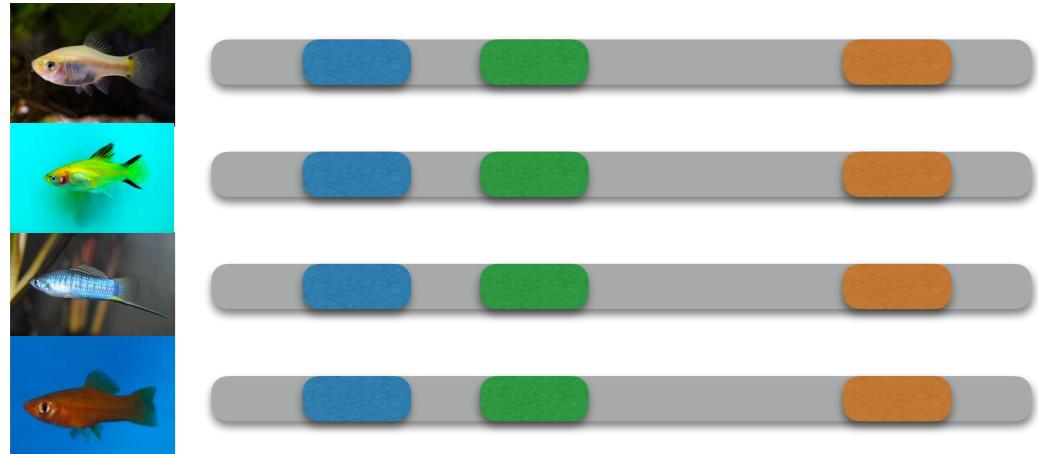


Hybrid
clades

Minor
sister
clades

When?

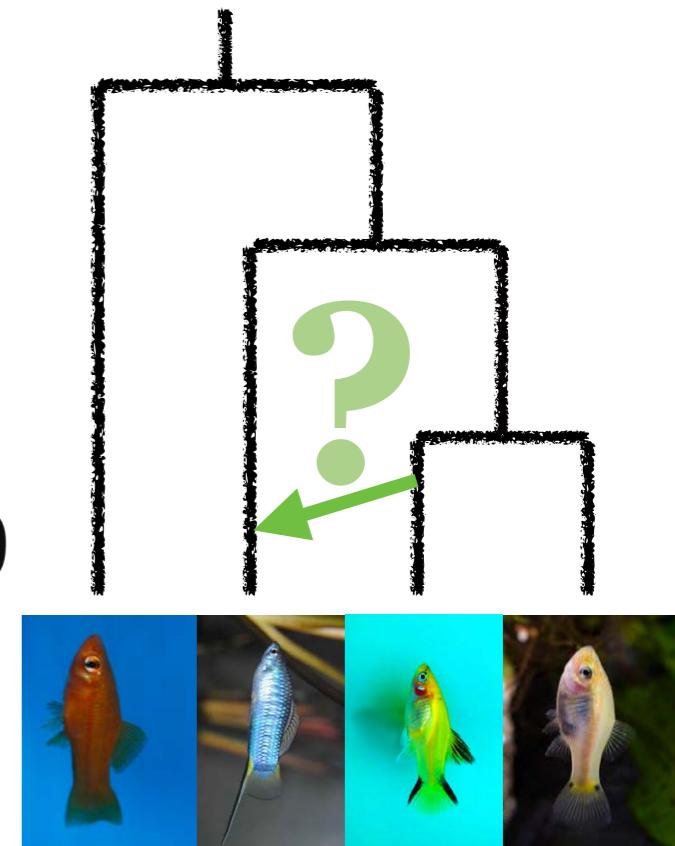
Phylogenetic network



Data

Goodness-of-fit test
Hypothesis test:
Is a tree a good fit?

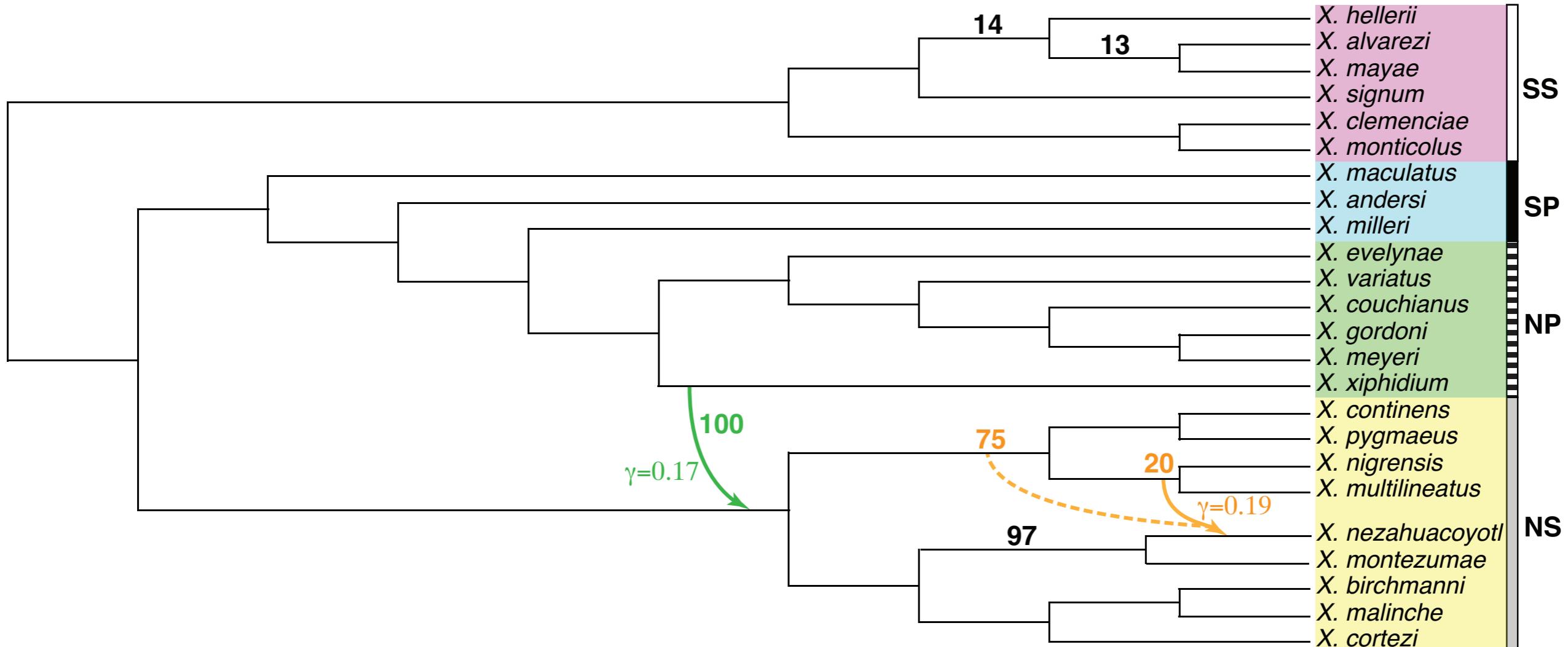
TICR
→
GitHub



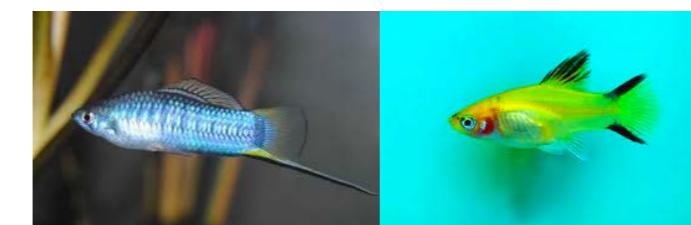
<https://github.com/nstenz/TICR>
(Stenz et al, 2015, Syst Bio)

Part II

I have the network, now what?



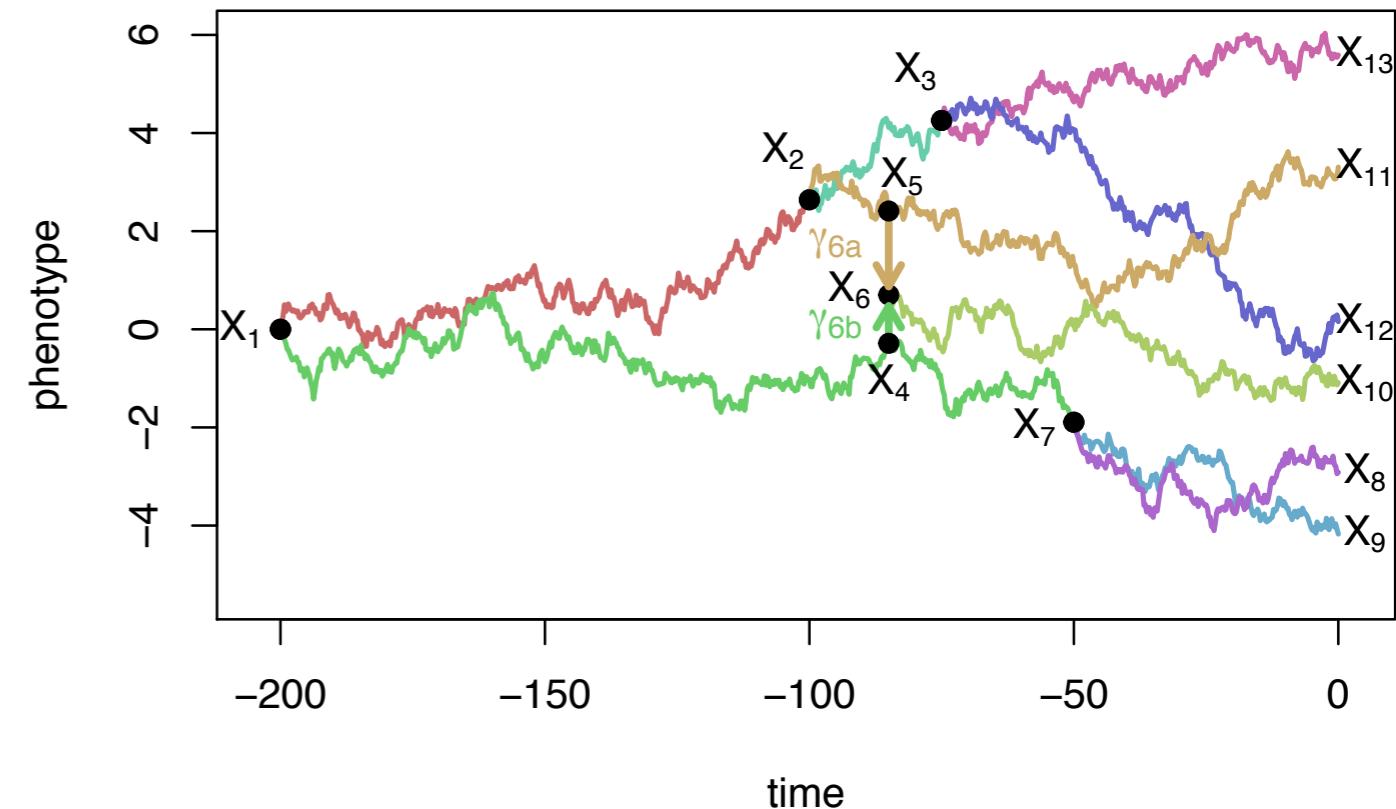
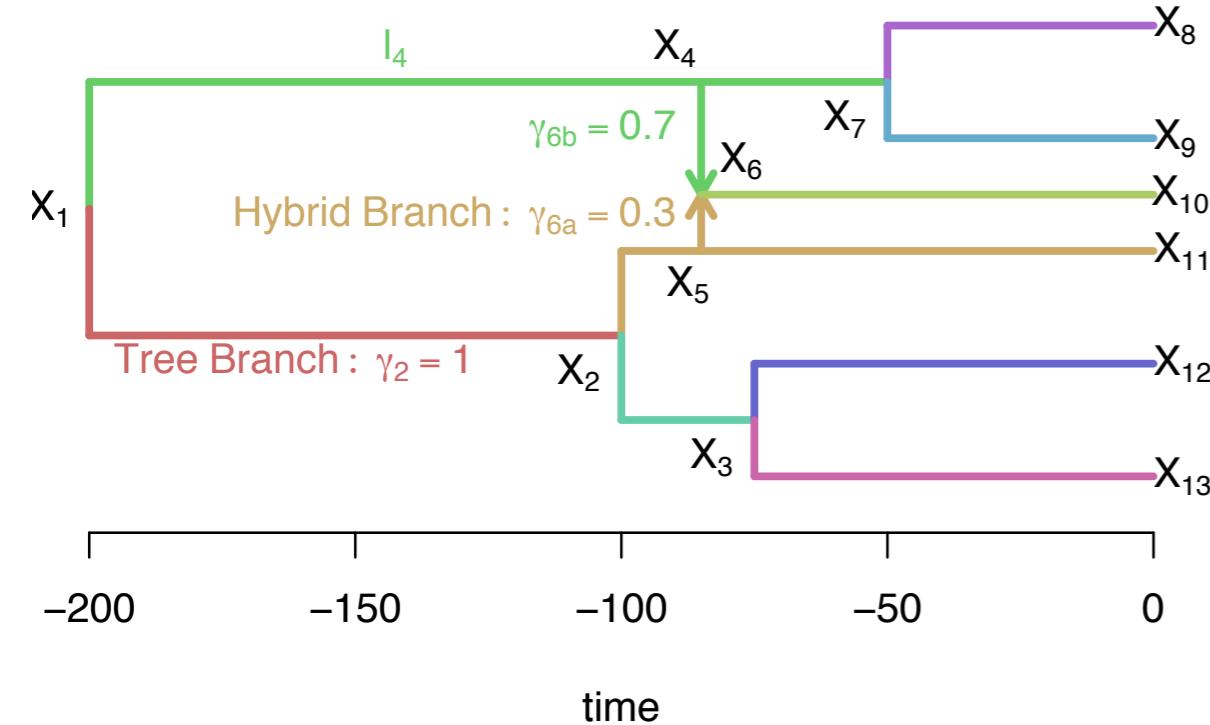
- Sword index
- Female preference



(Cui et al., 2013)

(Solís-Lemus, Ané, 2016, PLoS Genetics)

Trait models of evolution in networks



Brownian Motion
+ weighted
average in hybrid

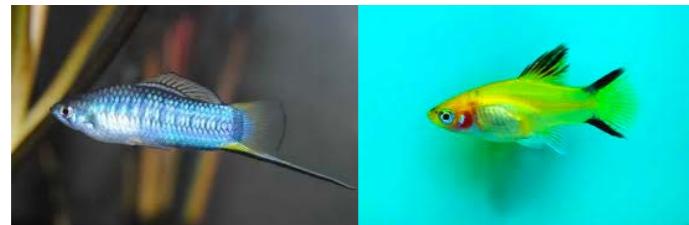
$$X_h = \gamma_1 X_{p_1} + \gamma_2 X_{p_2}$$

(Bastide et al, 2018, Syst Bio)

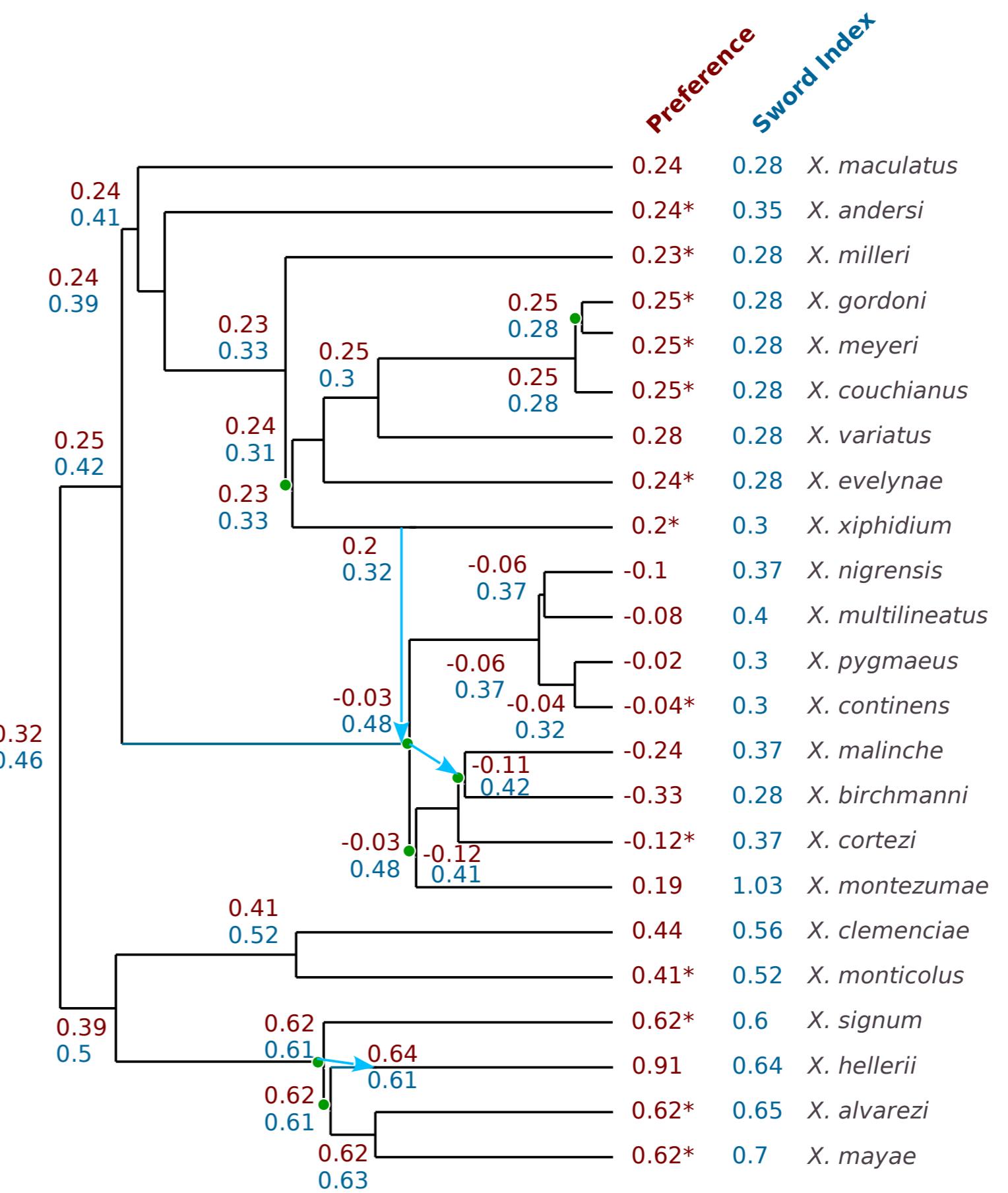
$$\mathbf{X} \sim N(X_{root}, \sigma^2 \mathbf{V})$$

- Phylogenetic signal
- Ancestral reconstruction
- Phylogenetic regression
- Phylogenetic ANOVA

- Sword index
- Female preference

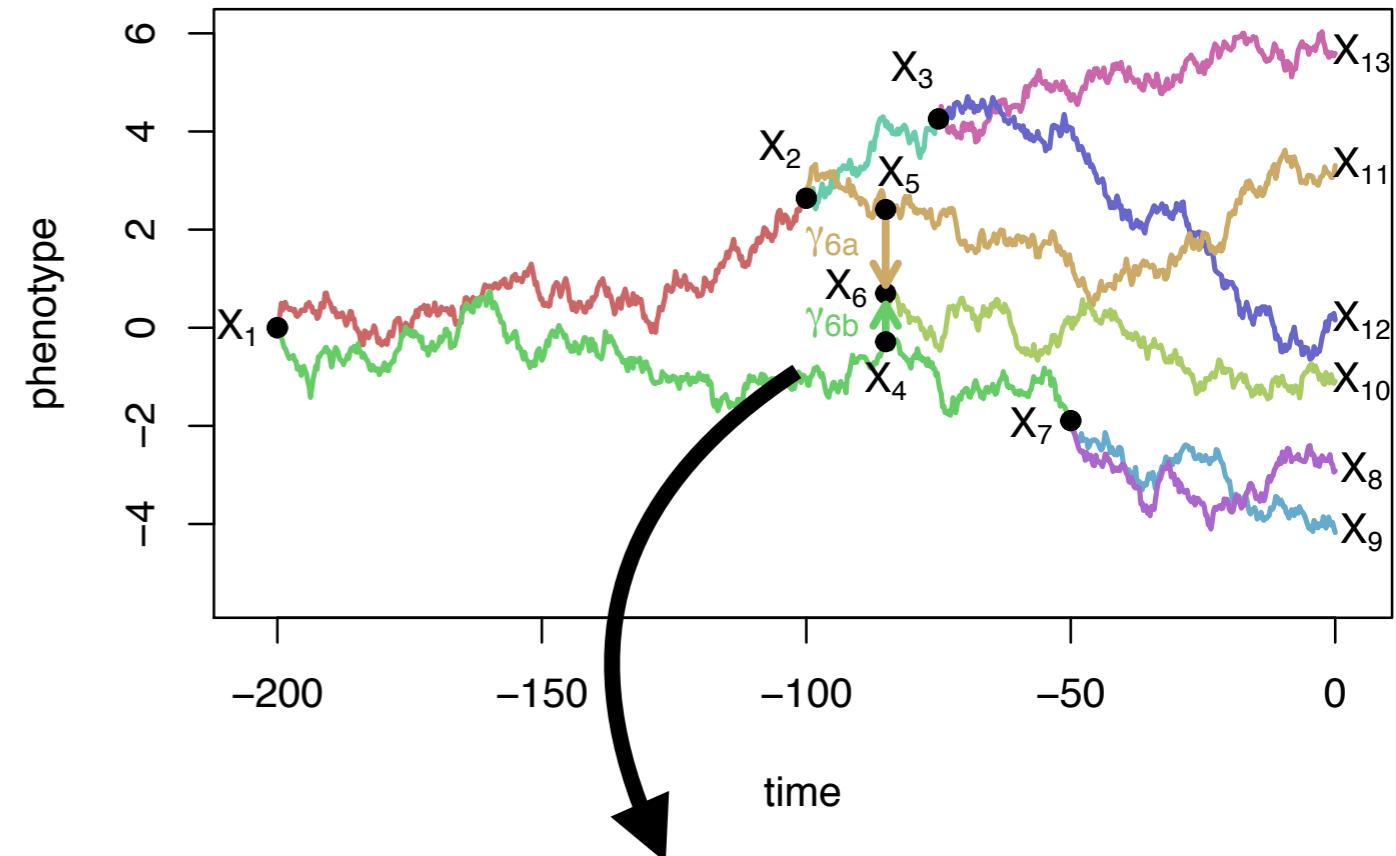
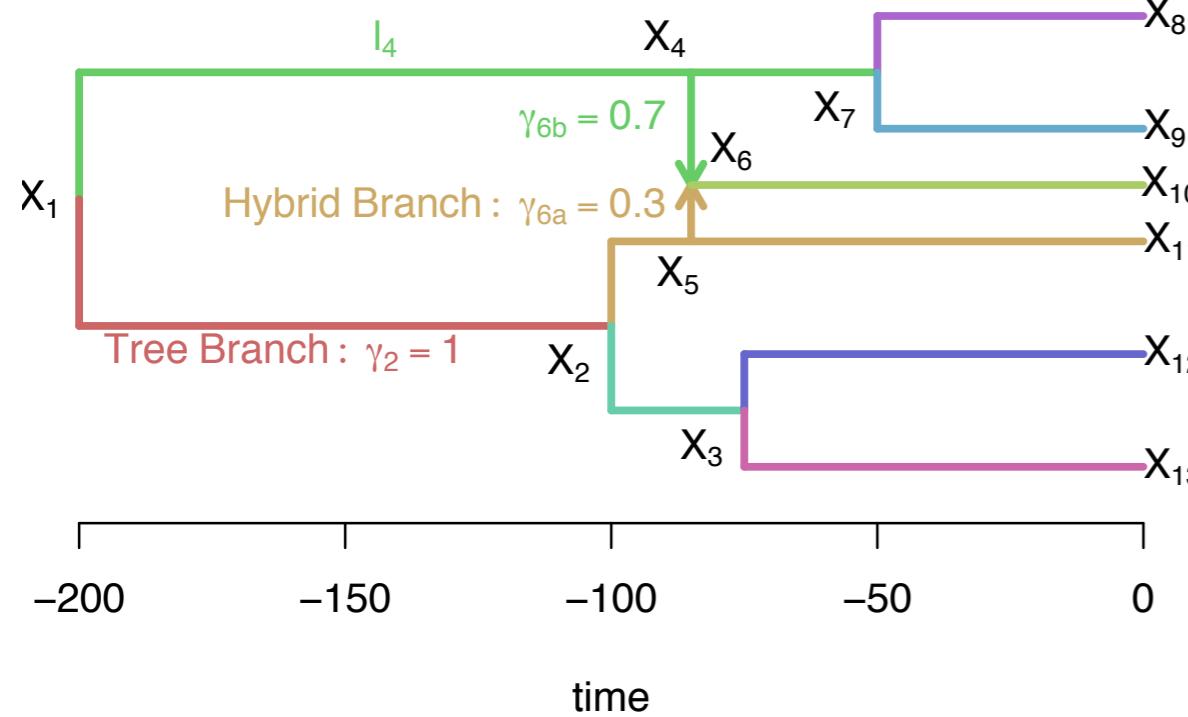


- **Ancestral reconstruction:** common ancestor likely had sword
- **Phylogenetic regression:** positive association between sword index and female preference but not significant ($p = 0.106$)



Test for transgressive evolution

$$X_h = \gamma_1 X_{p_1} + \gamma_2 X_{p_2} + \Delta_h$$



$\Delta_h = 0$ No transgressive evolution

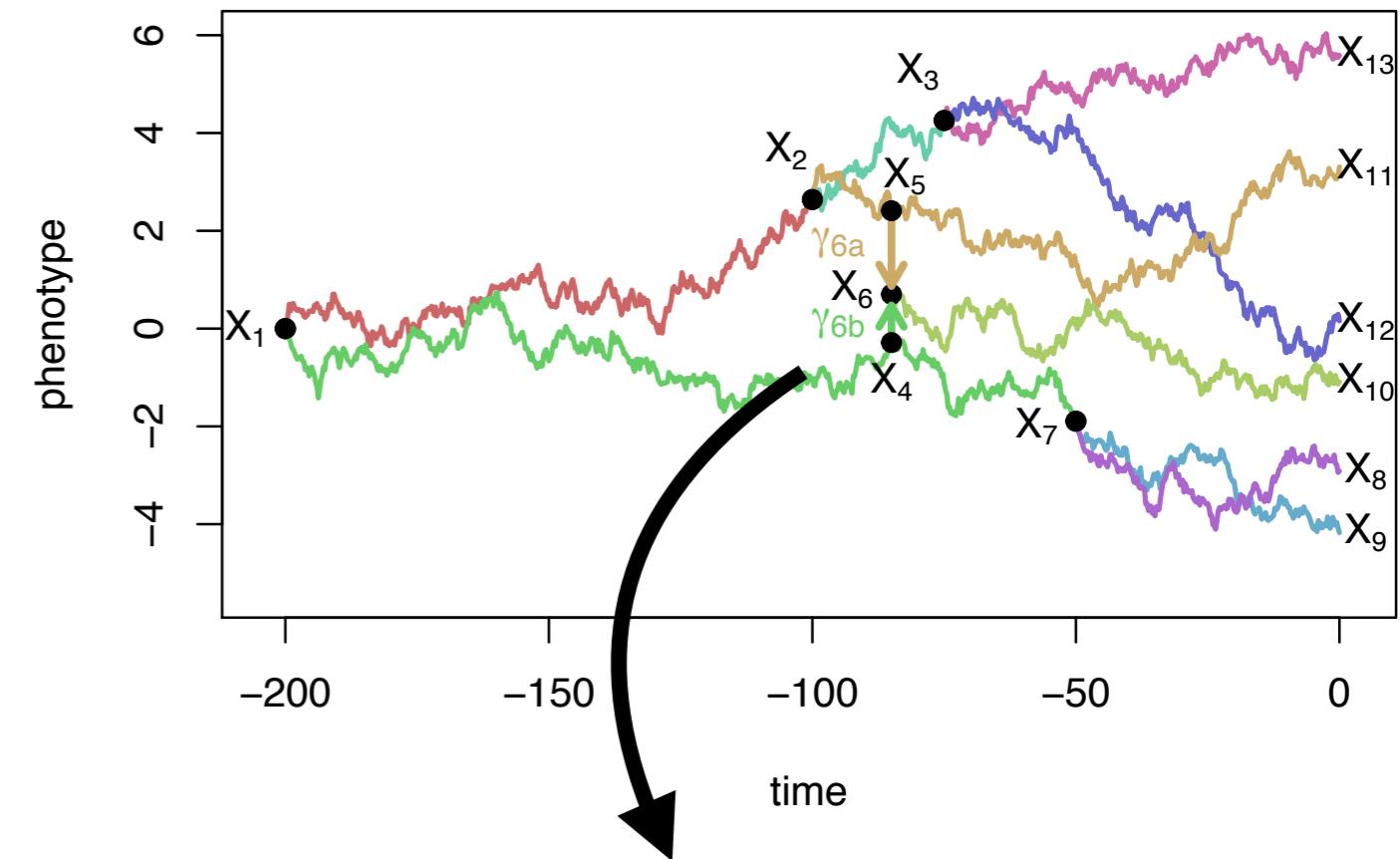
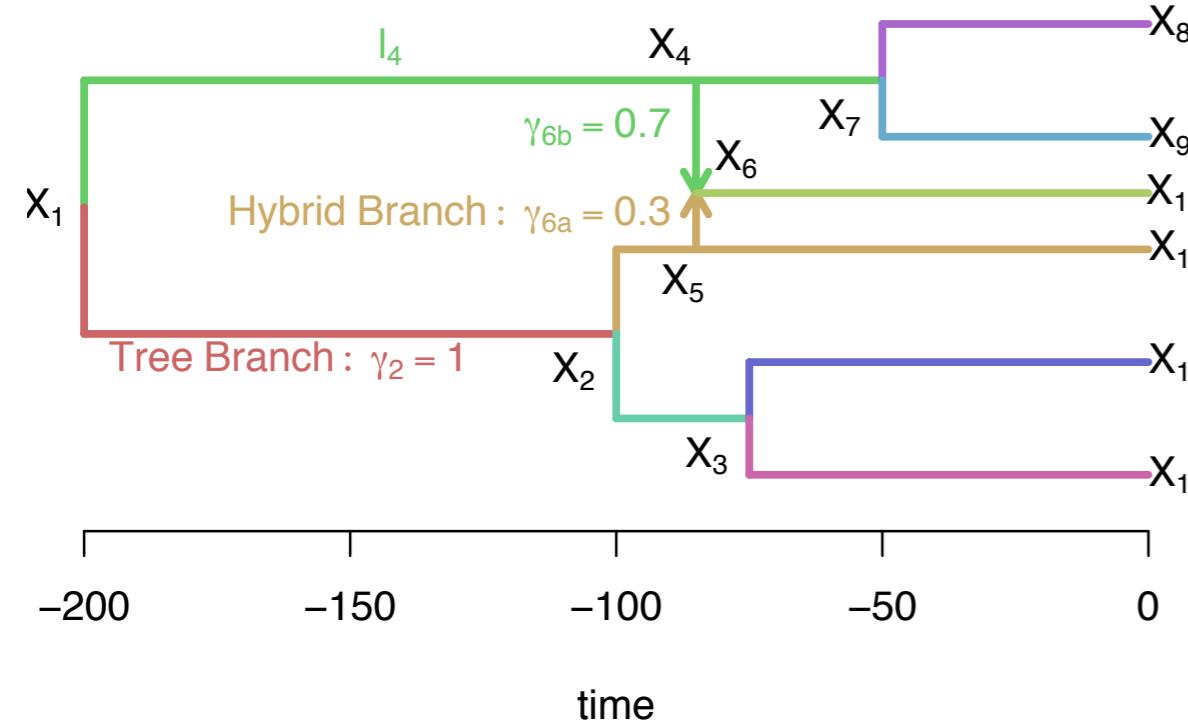
$\Delta_h = \Delta$ Single-effect transgressive evolution

Δ_h Multi-effect transgressive evolution

F tests

Hybrid value:
shift from
parents range

Test for transgressive evolution



- Sword index: $p=0.55$
- Female preference: $p=0.0064$

Hybrid value:
shift from
parents range

PhyloNetworks: analysis for phylogenetic networks

build passing docs stable docs dev codecov 81% coverage 67%



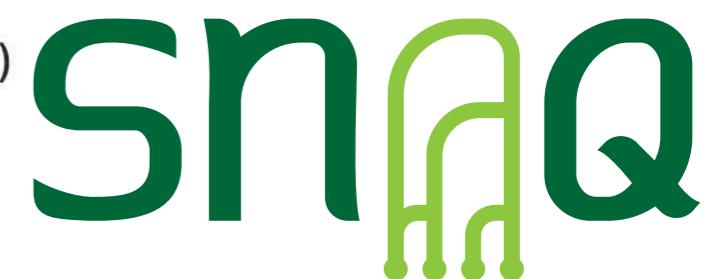
Overview

PhyloNetworks is a [Julia](#) package with utilities to:

- read / write phylogenetic trees and networks, in (extended) Newick format. Networks are considered explicit: nodes represent ancestral species. They can be rooted or unrooted.
- manipulate networks: re-root, prune taxa, remove hybrid edges, extract the major tree from a network, extract displayed networks / trees
- compare networks / trees with dissimilarity measures (Robinson-Foulds distance on trees)
- summarize samples of bootstrap networks (or trees) with edge and node support
- estimate species networks from multilocus data (see below)
- phylogenetic comparative methods for continuous trait evolution on species networks / trees



- Step-by-step tutorial
- Online documentation
- Google user group



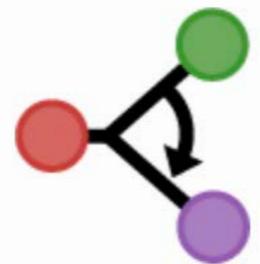
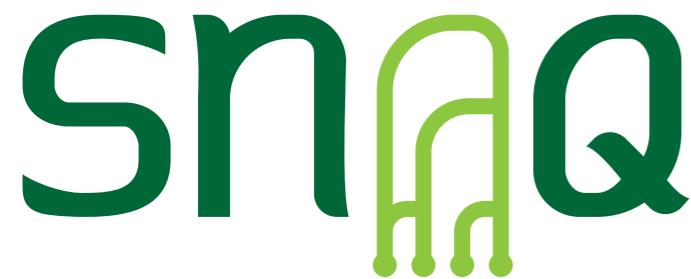
<http://crsl4.github.io/>

(S.-L. et al, 2017, MBE)

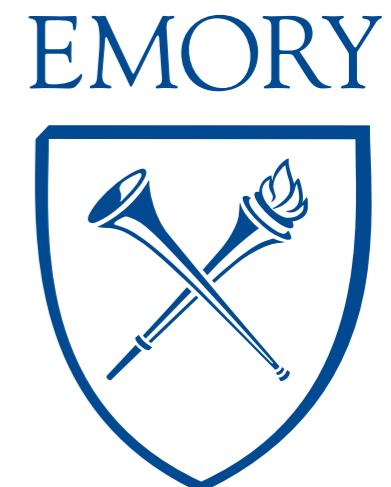
Acknowledgements

Cécile Ané (UW)
Paul Bastide (KU-Leuven)
Bret Larget (UW)
Douglas Bates (UW)
David Baum (UW)
Sarah Friedrich (UW)
Michael Epstein (Emory)

Mengyao Yang
John Malloy
John Spaw
Noah Stenz
Nan Ji
Jordan Vonderwell
Josh McGrath



PhyloNetworks.jl



<http://crsl4.github.io/>