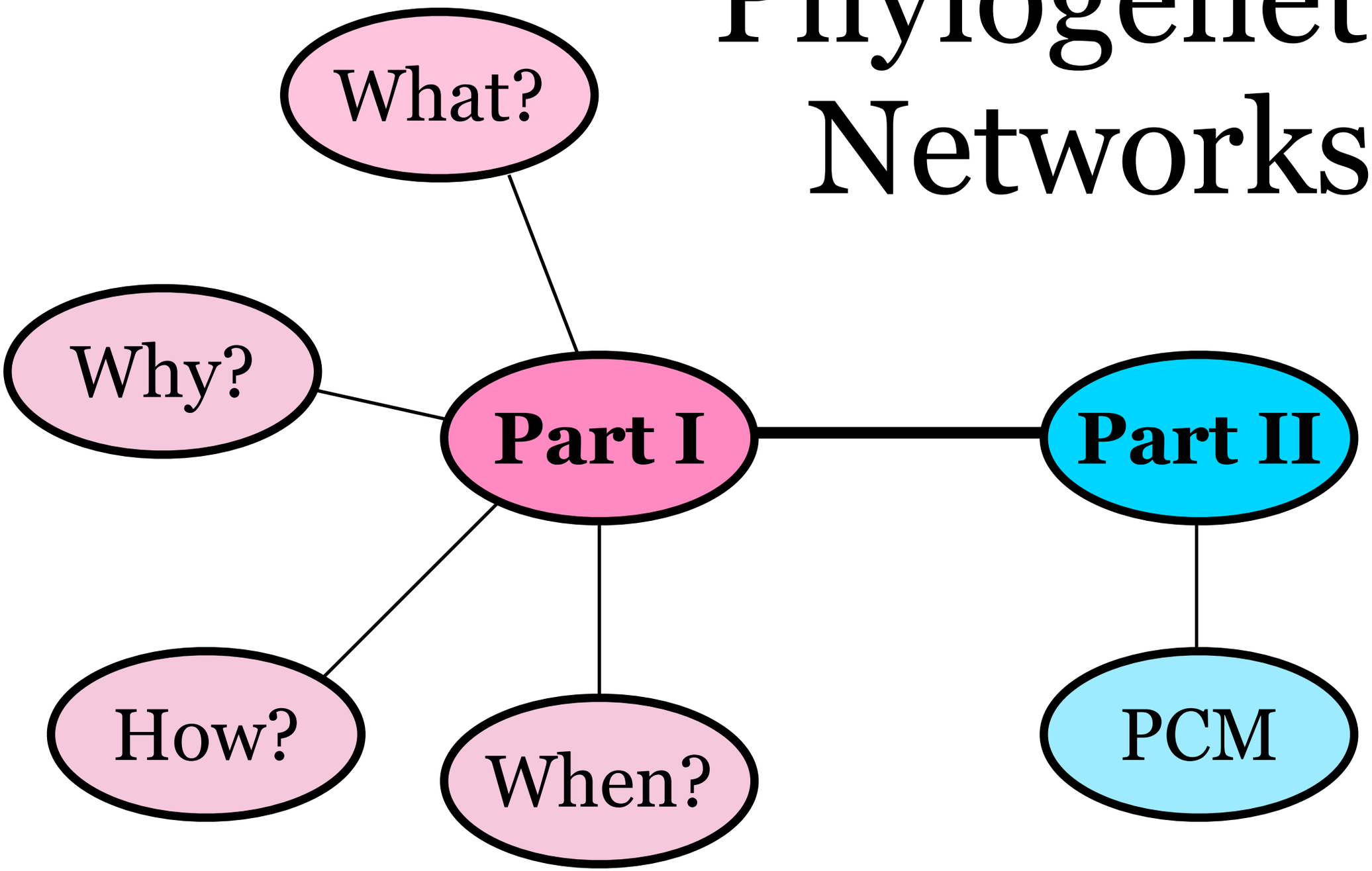
Statistical models on phylogenetic networks

Claudia Solís-Lemus, PhD
University of Wisconsin-Madison
Wisconsin Institute for Discovery
Department of Plant Pathology

June 3, 2022

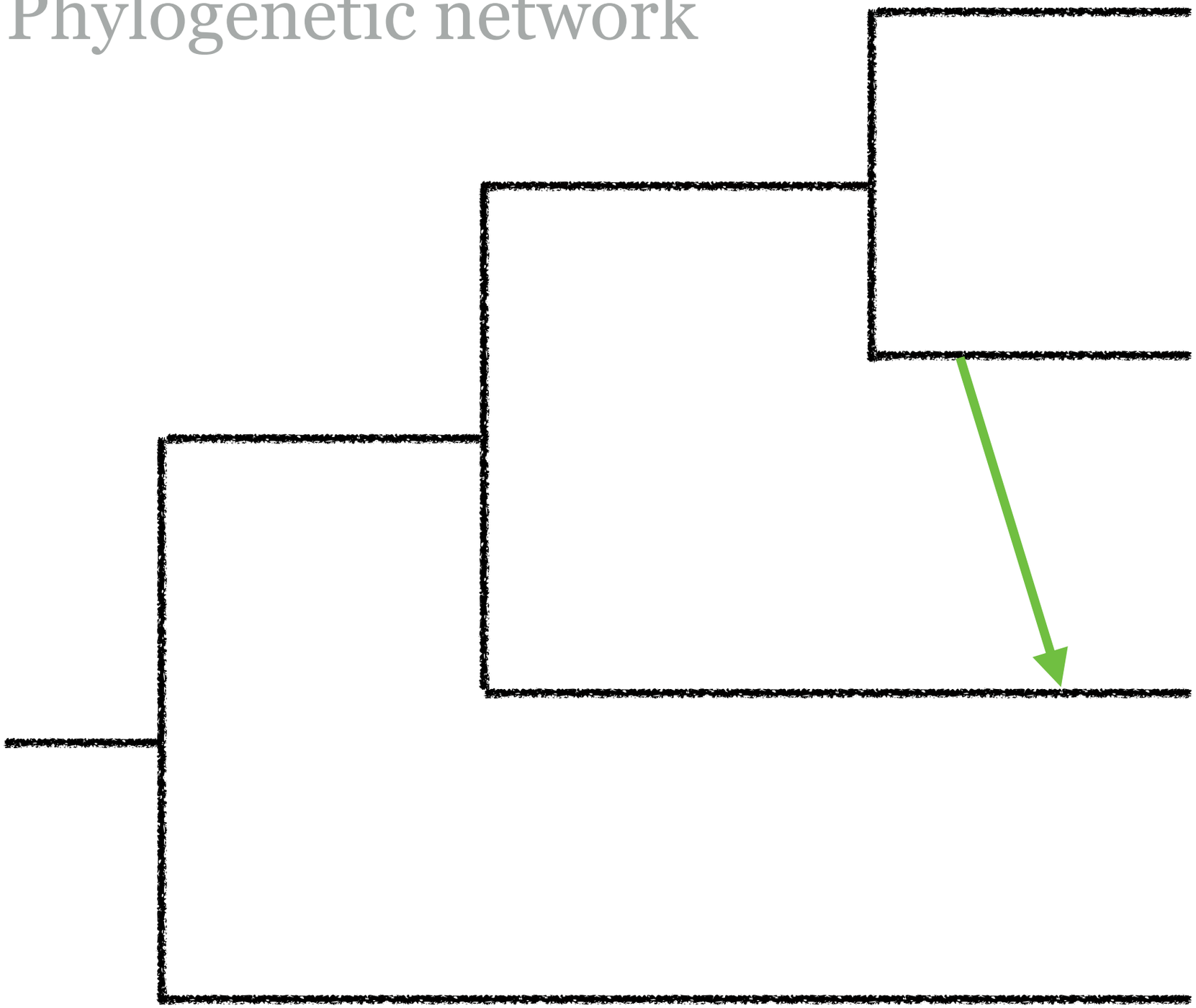
https://solislemuslab.github.io/ @solislemuslab crsl4 @thestatistician

Phylogenetic Networks



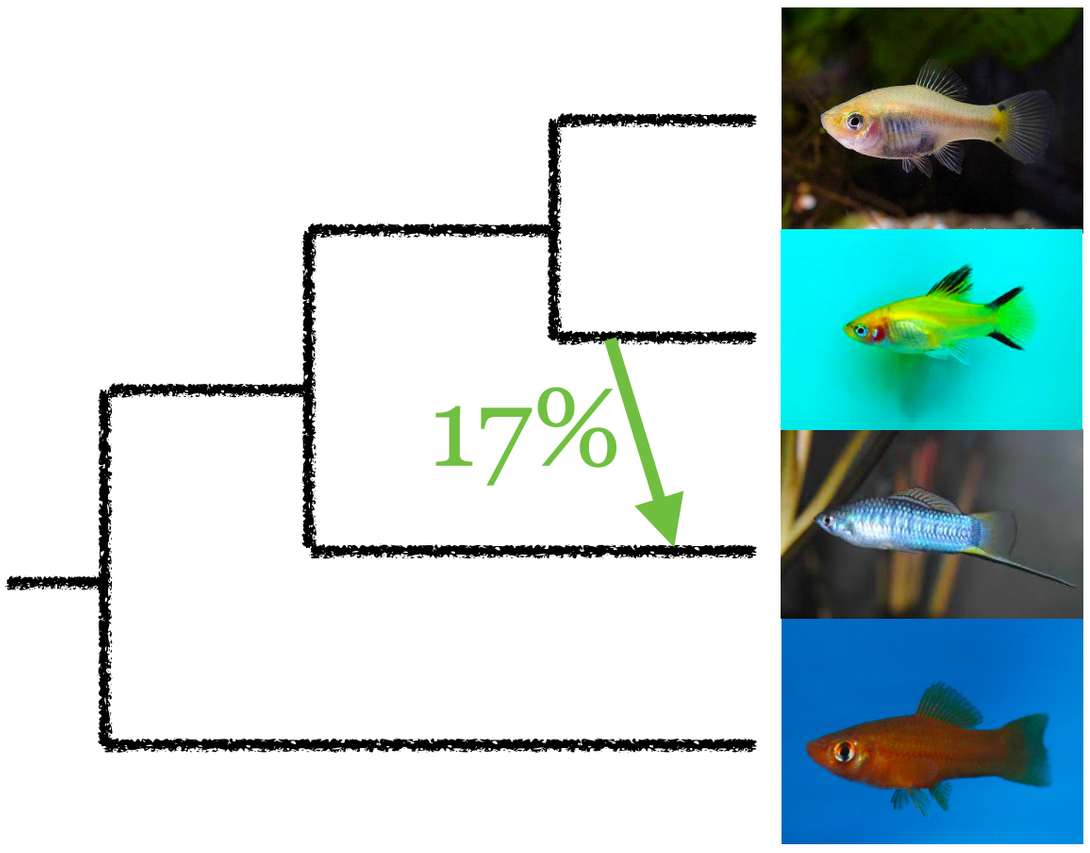
What?

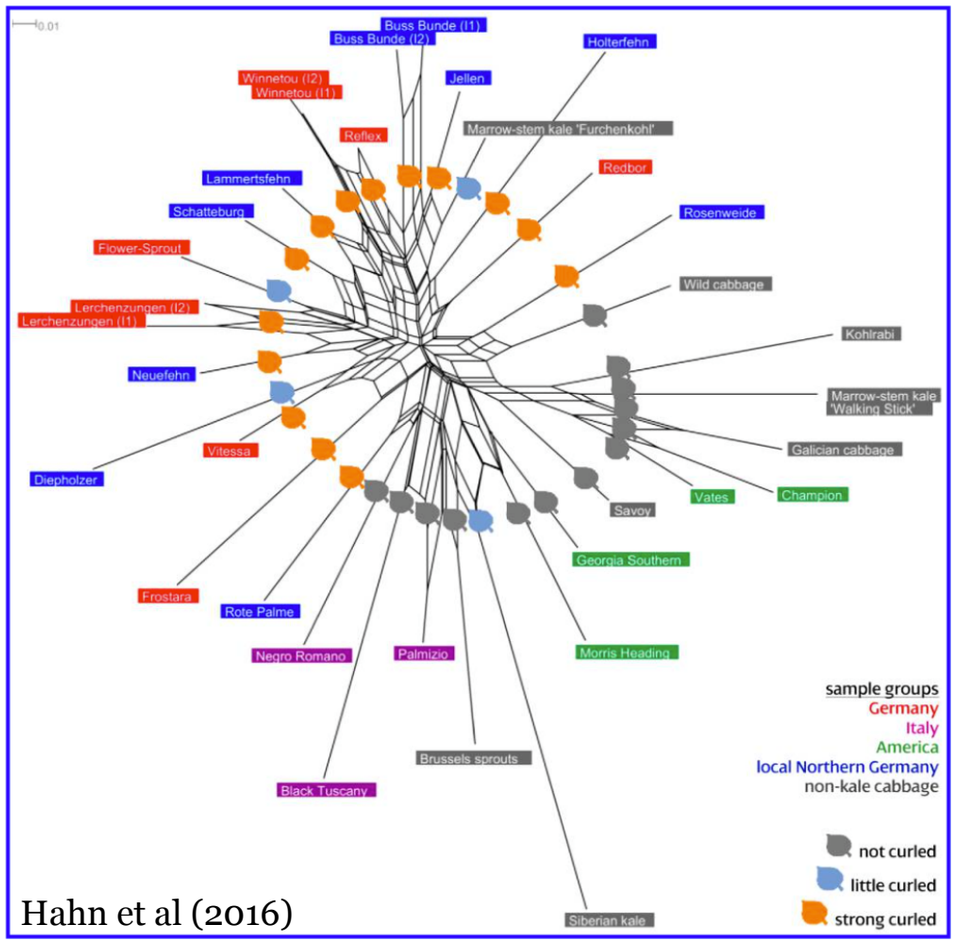
Phylogenetic network



What?

Phylogenetic network





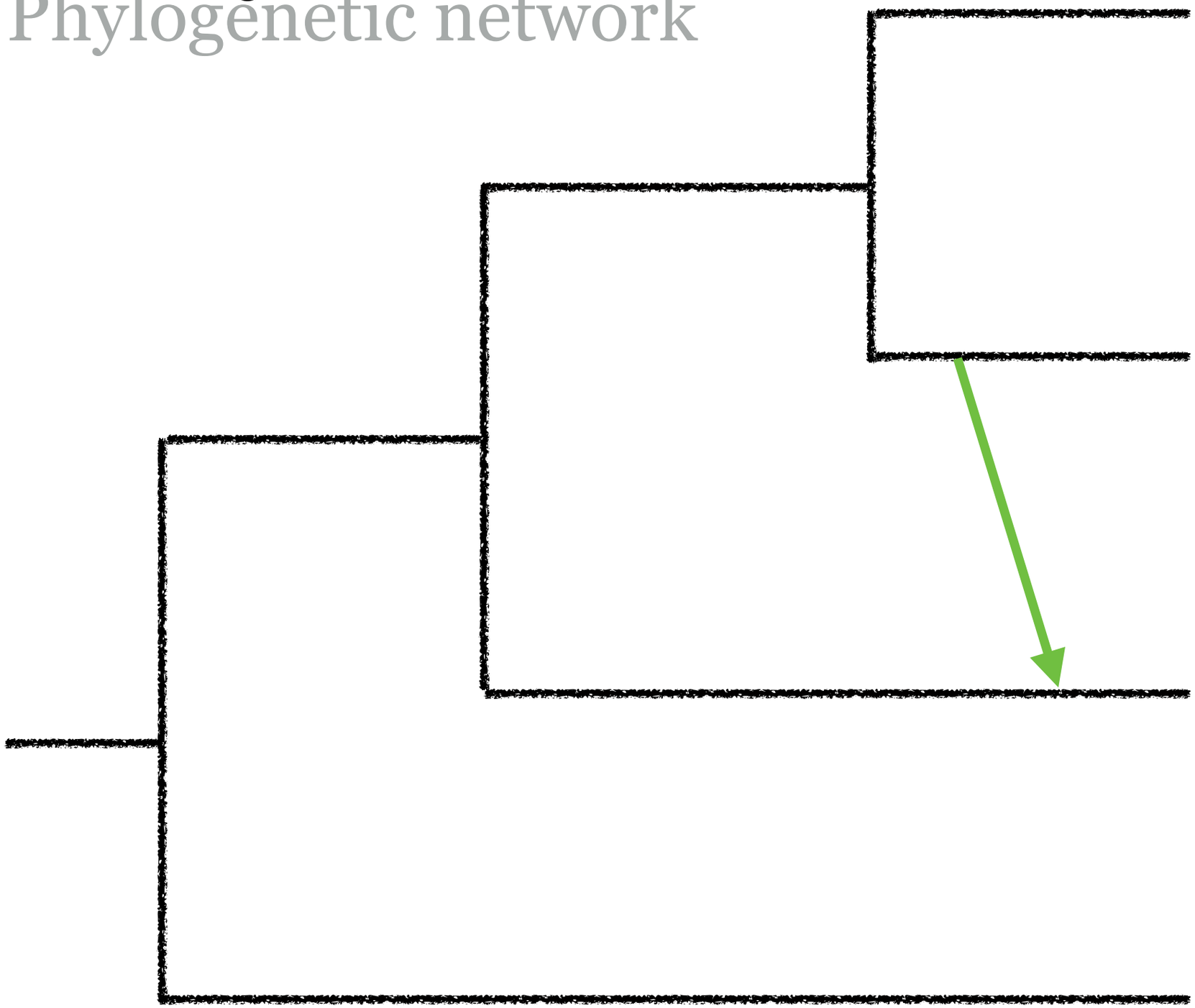
Hahn et al (2016)

Explicit

Implicit

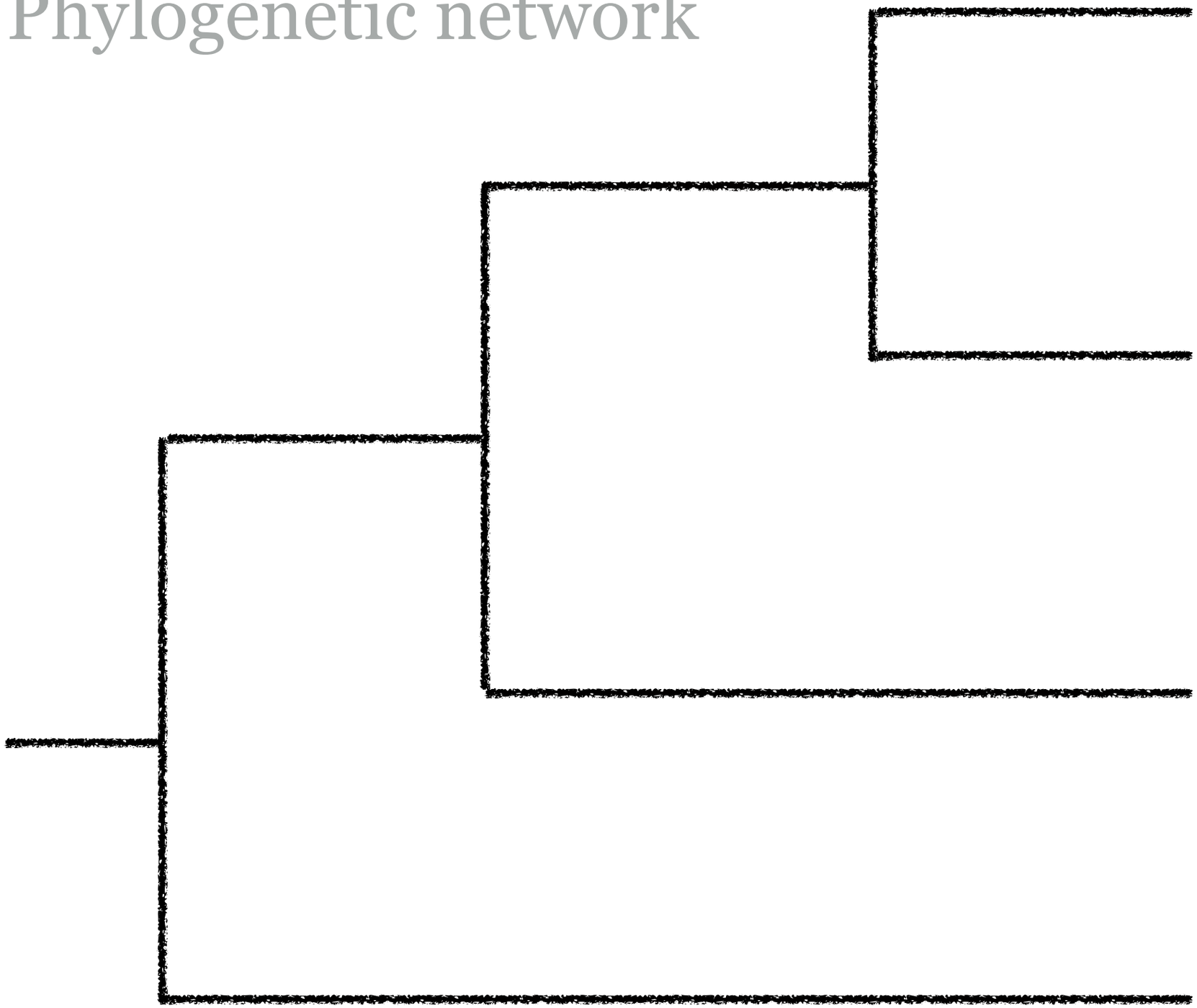
Why?

Phylogenetic network



Why?

Phylogenetic network

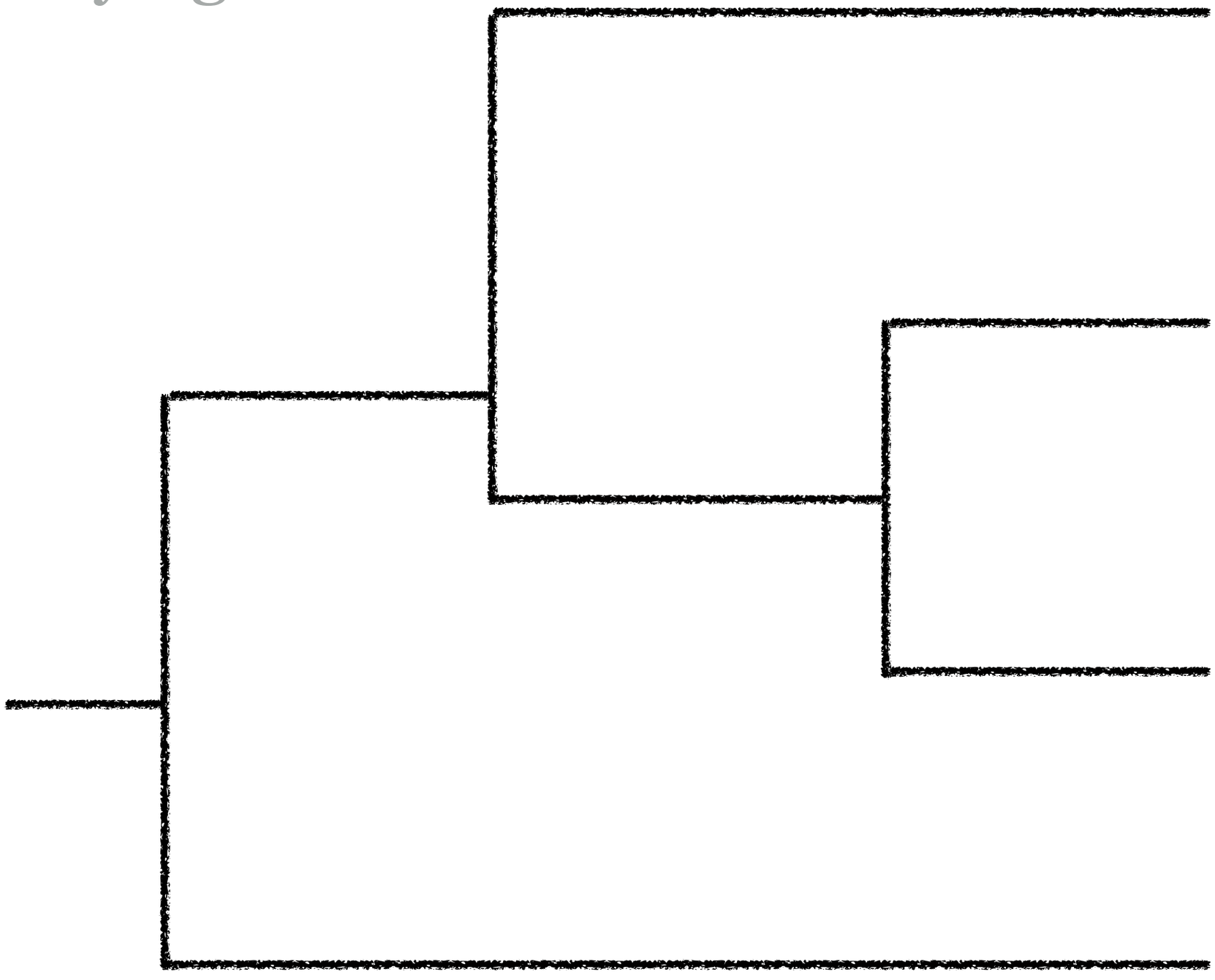


Main tree

Why?

Phylogenetic network

Ignore gene flow
=>Wrong tree!

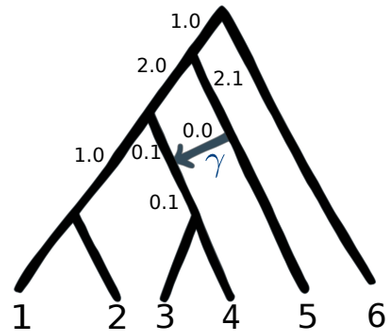


(S.-L., Yang, Ané, 2016, Syst Bio)

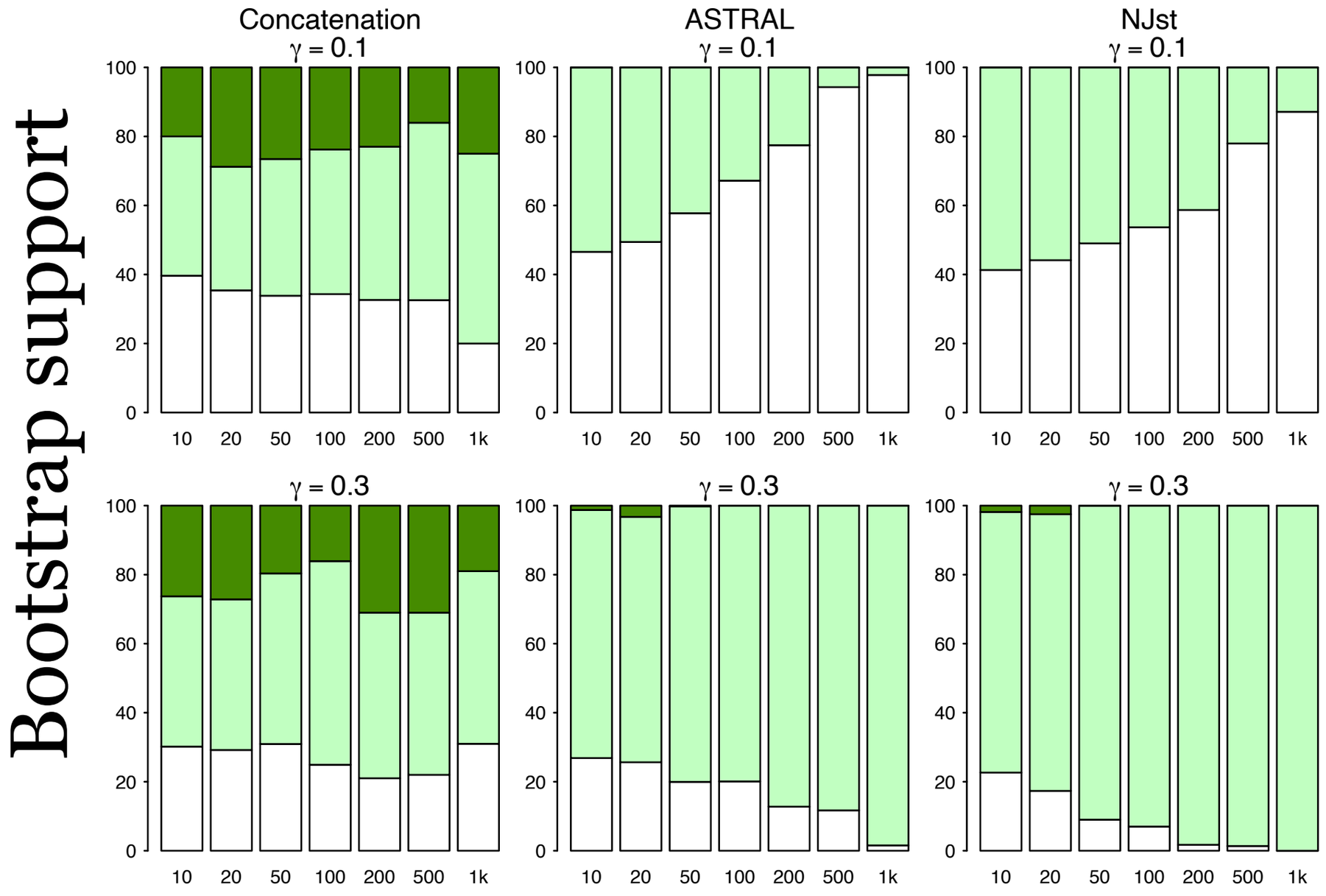
Why?

Phylogenetic network

Coalescent tree methods not robust to gene flow



White: true tree



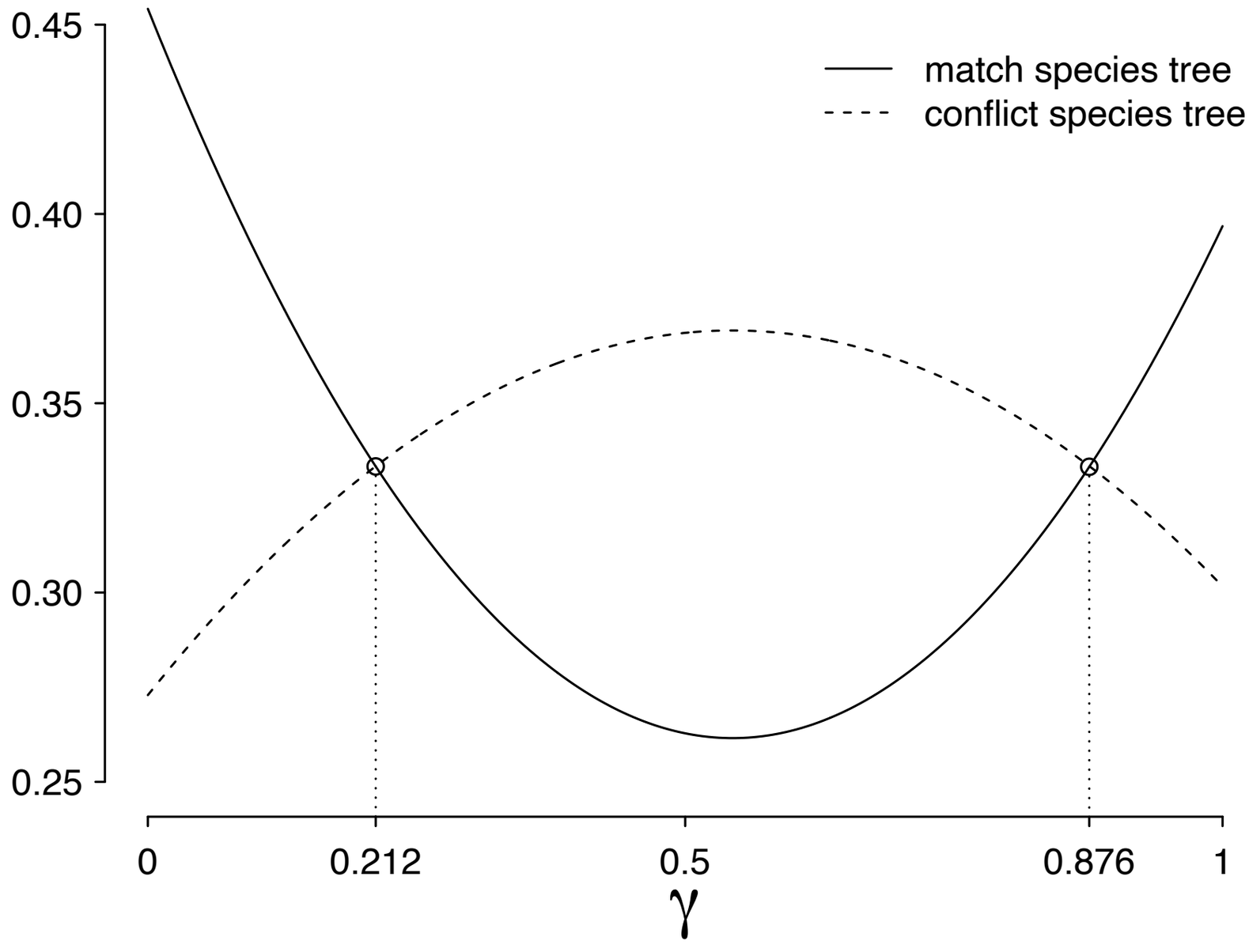
(S.-L., Yang, Ané, 2016, Syst Bio)

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

Why?

Phylogenetic network

Anomaly zone with gene flow

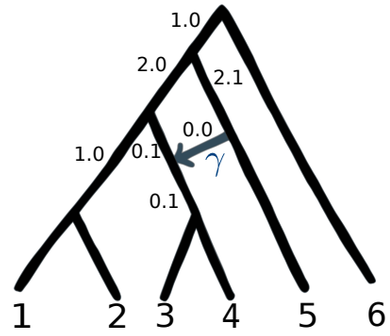


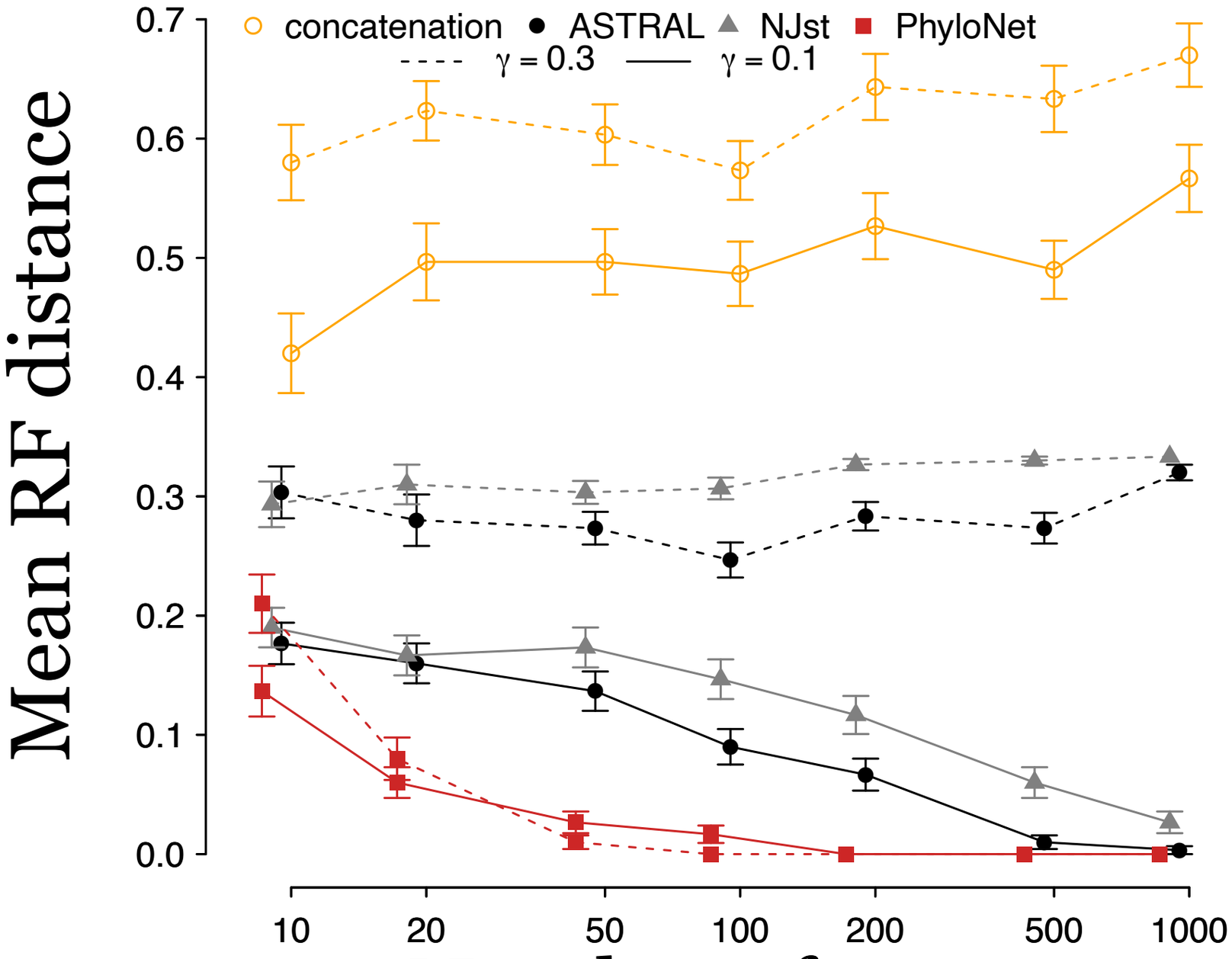
(S.-L., Yang, Ané, 2016, Syst Bio)

Why?

Phylogenetic network

Coalescent tree methods not robust to gene flow





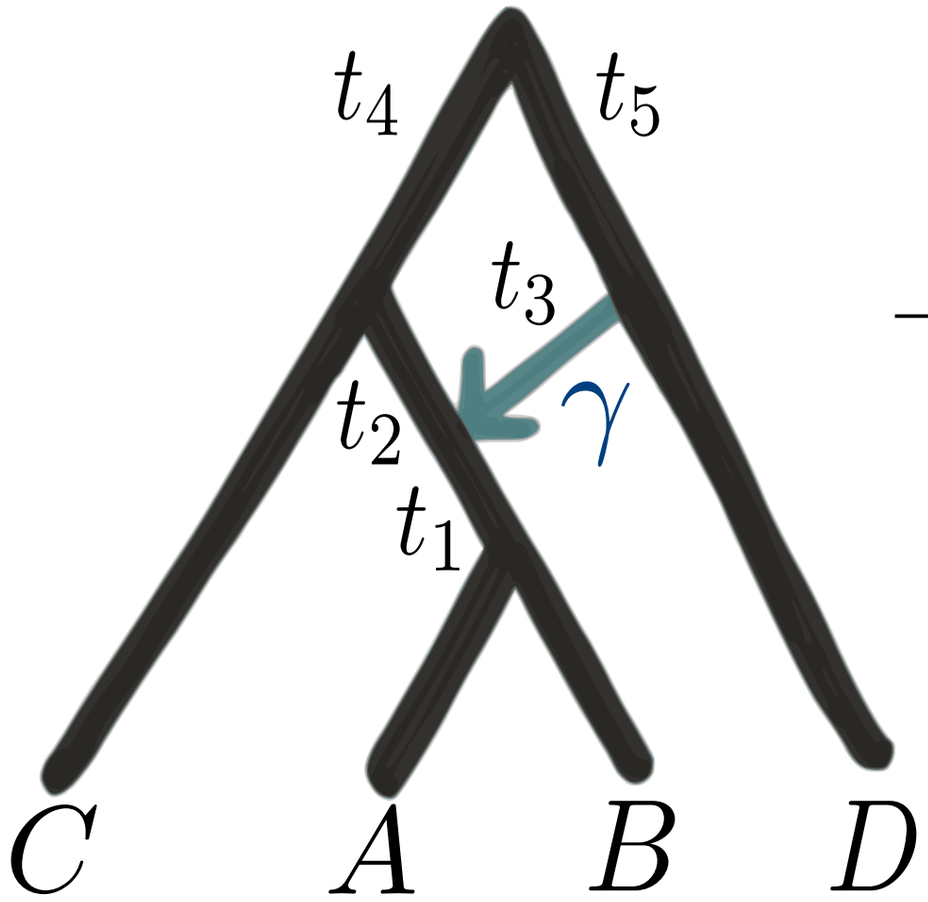
(S.-L., Yang, Ané, 2016, Syst Bio)

ASTRAL (Mirarab et al, 2014)
NJst (Liu&Yu,2011)
PhyloNet (Yu et al 2012, 2014)

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)

See also Long & Kubatko (2018) for AUGT under continuous gene flow between sister species

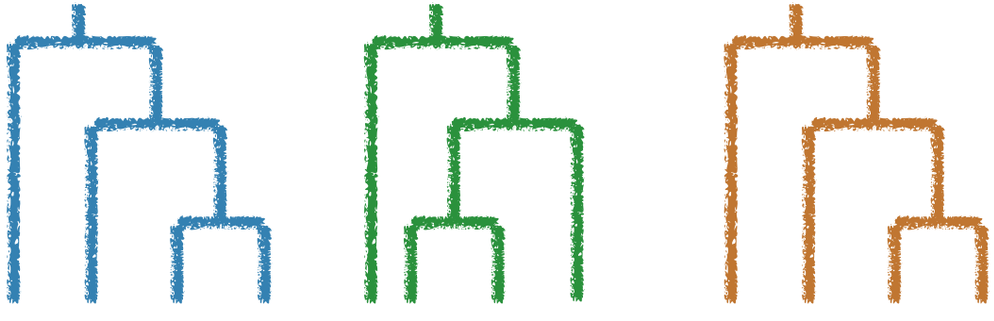
So far…

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

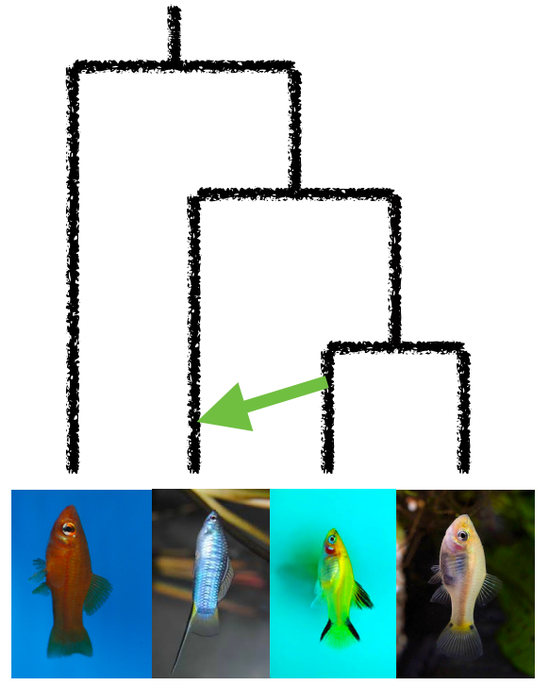
How?

Phylogenetic network

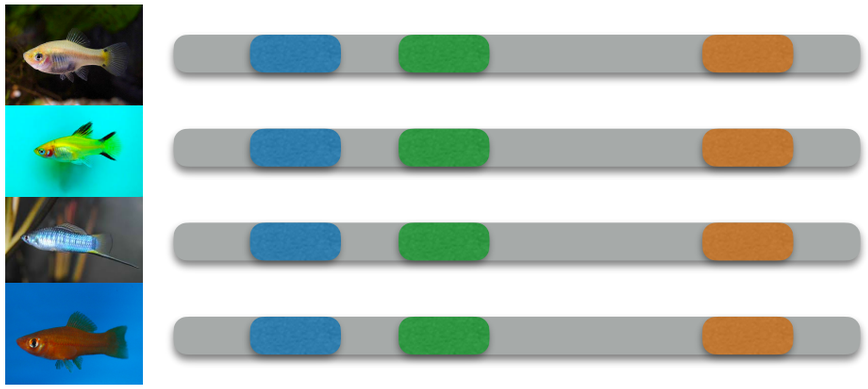
MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
PhyML
(Guindon et al, 2010)
RevBayes
(Hoehna et al, 2016)
IQ-TREE
Nguyen et al. (2015)



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



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



$$P(N, G, \theta | D) \propto \pi(N)\pi(\theta) \prod_{i=1}^{L} P(D_i|G_i)P(G_i|N)$$

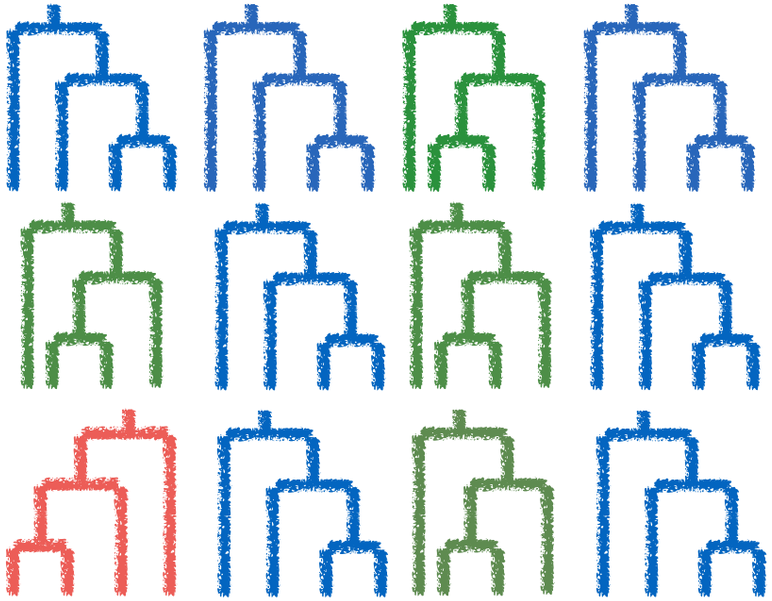
network · gene trees · L loci · **Prior Network** · **Substitution Model** · **Multispecies Coalescent**

Birth-hybridization process

reticulations, cycle diameter

BEAST2
(Zhang et al, 2017)

PhyloNet Bayesian
(Wen et al, 2016)

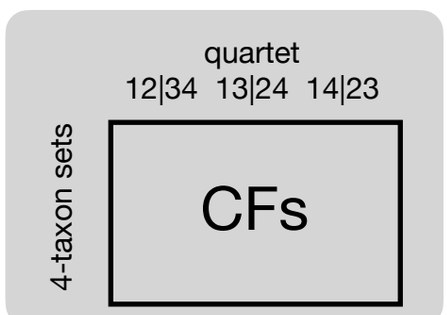


$$P(N, \theta | G) \propto \pi(N)\pi(\theta) \prod_{i=1}^{L} P(G_i|N, \theta)$$

PhyloNet Bayesian
(Wen et al, 2016)

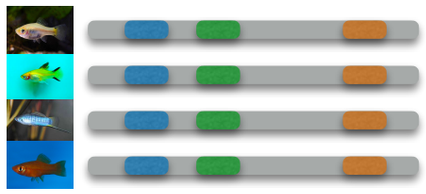
$$L(N, \theta) = \prod_{i=1}^{L} P(G_i|N, \theta)$$

PhyloNet Likelihood
(Yu et al, 2014)



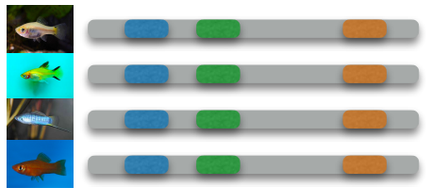
$$\tilde{L}(N, \theta) \propto \prod_{q} L(q|N, \theta)$$

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



BEAST2
(Zhang et al, 2017)

Birth-hybridization process

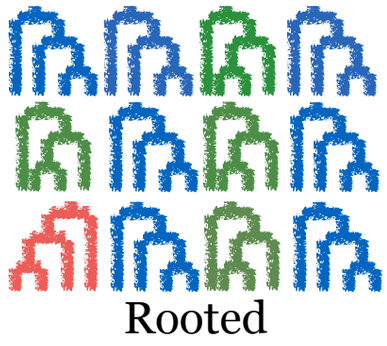


PhyloNet Bayesian
(Wen et al, 2016)

reticulations, cycle diameter

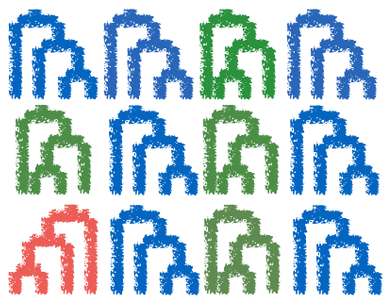
MCMC: Network moves, mixing

Most accurate, not scalable



Rooted

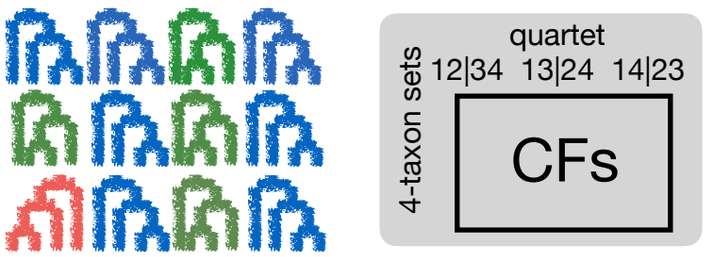
PhyloNet Bayesian
(Wen et al, 2016)



Rooted

PhyloNet Likelihood
(Yu et al, 2014)

Heuristic search: Network moves



Unrooted

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

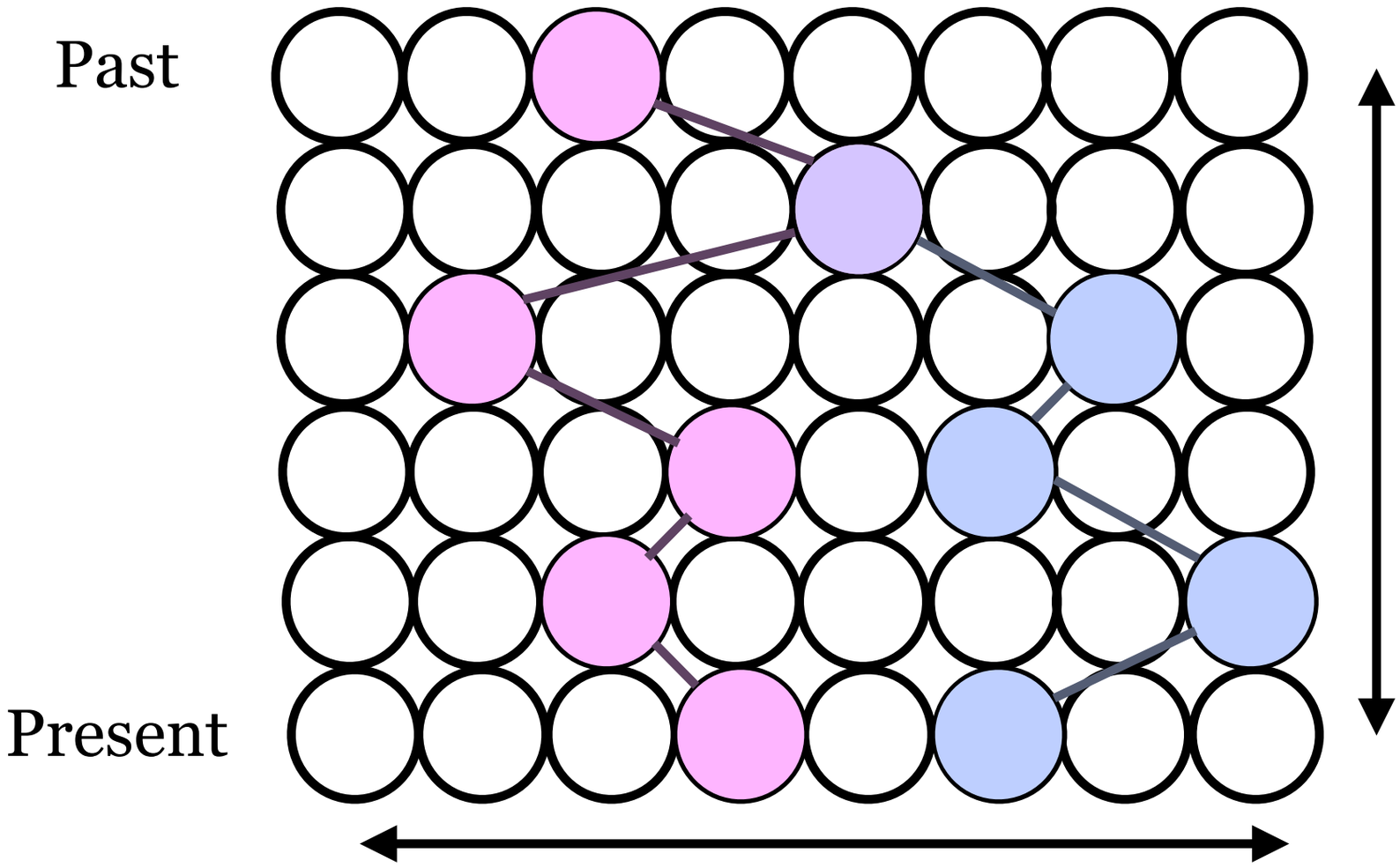
Level-1 networks

More scalable, Robust

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.

Slide by Cecile Ane

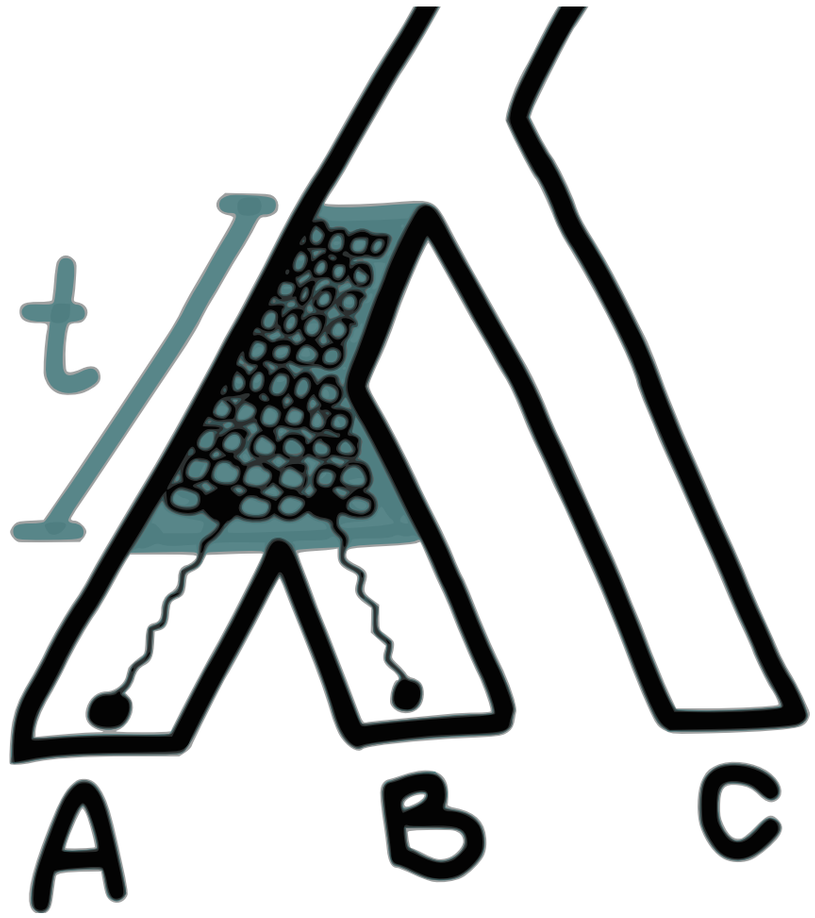
Coalescent model within 1 population



Probability of no coalescence in g generations: $\left(1-\frac{1}{N}\right)^{g}$

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

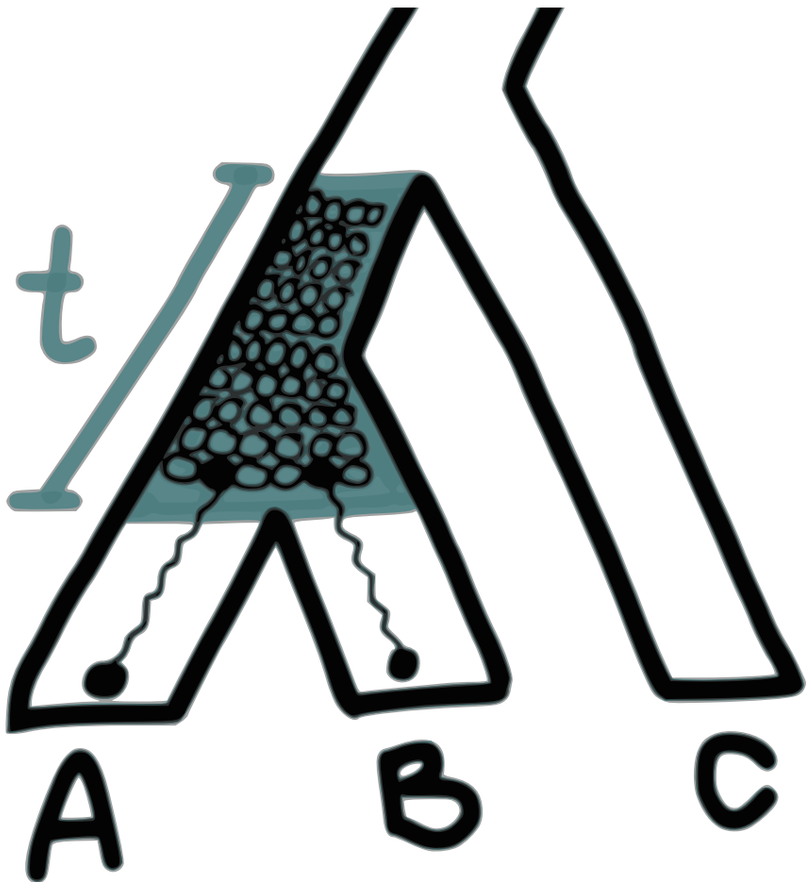
Multispecies coalescent on a tree



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

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

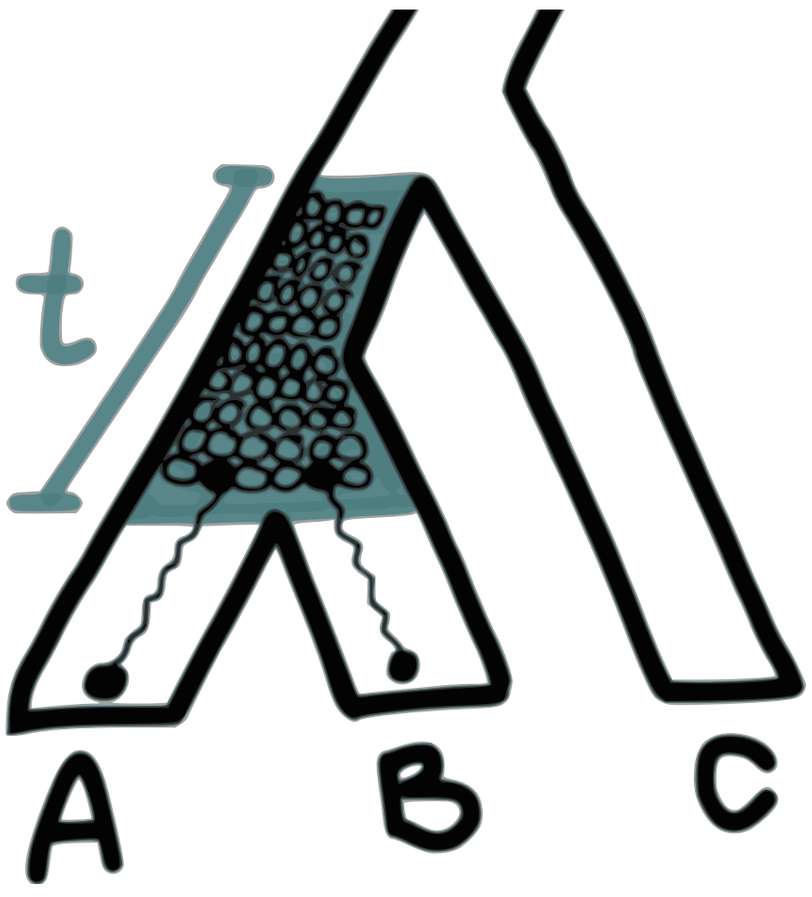
Multispecies coalescent on a tree



$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) =$$

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

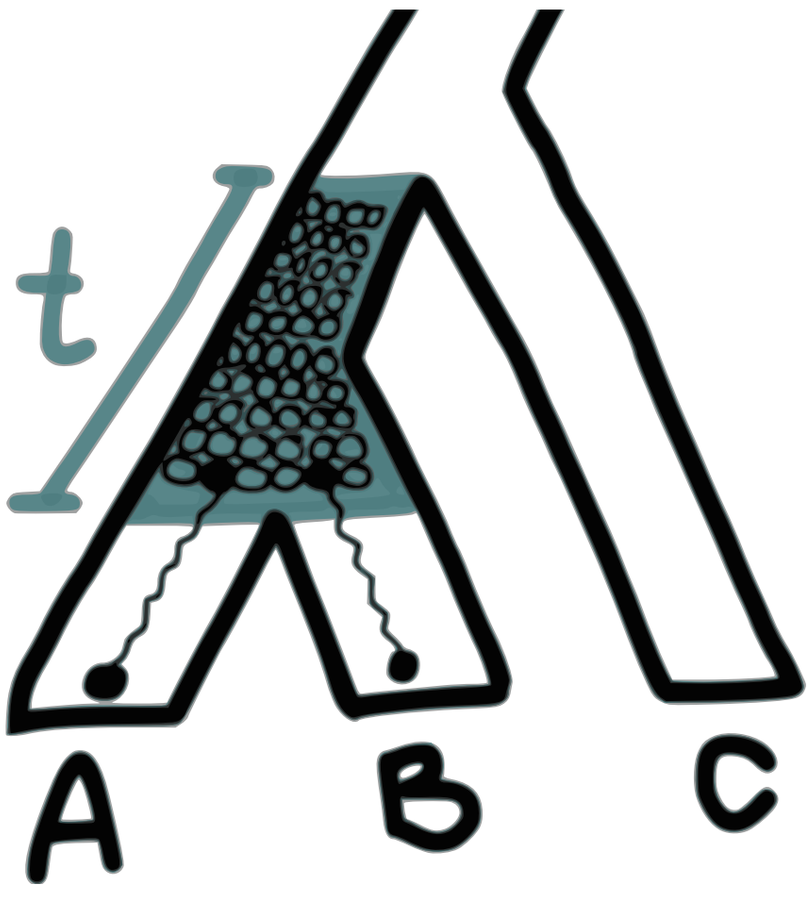
Multispecies coalescent on a tree



$$\mathbb{P}\left(\begin{array}{c}\\ \wedge \\ \text{A B C}\end{array}\right) = 1 - e^{-t}$$

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

Multispecies coalescent on a tree



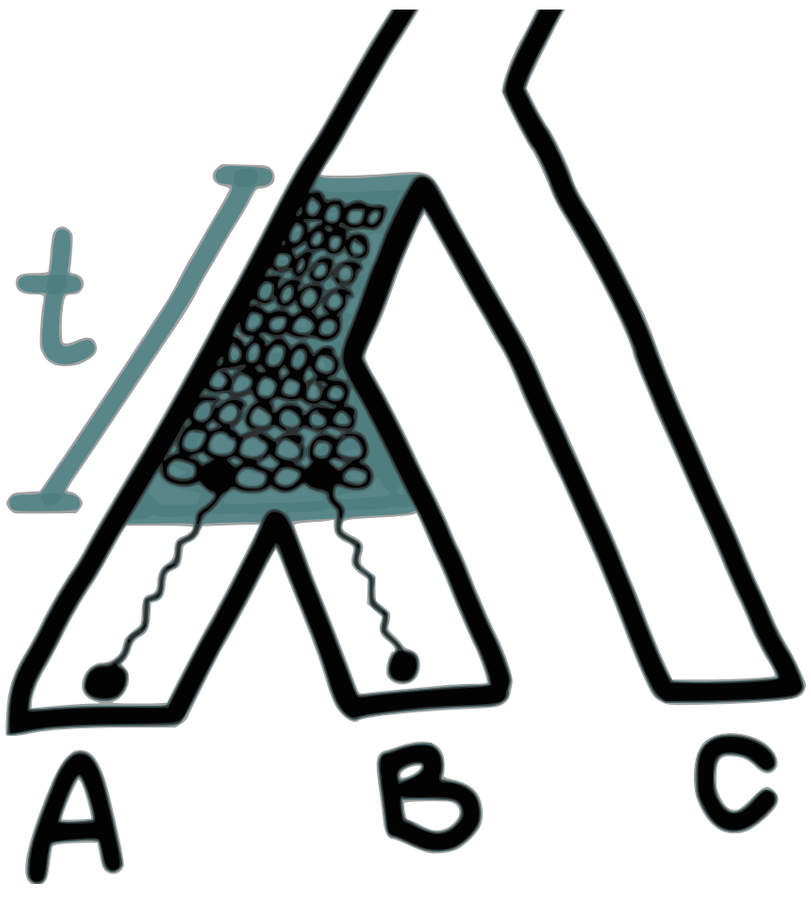
$$P(T > t) = e^{-t}$$

$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) =$$

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

$$+$$

Multispecies coalescent on a tree



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

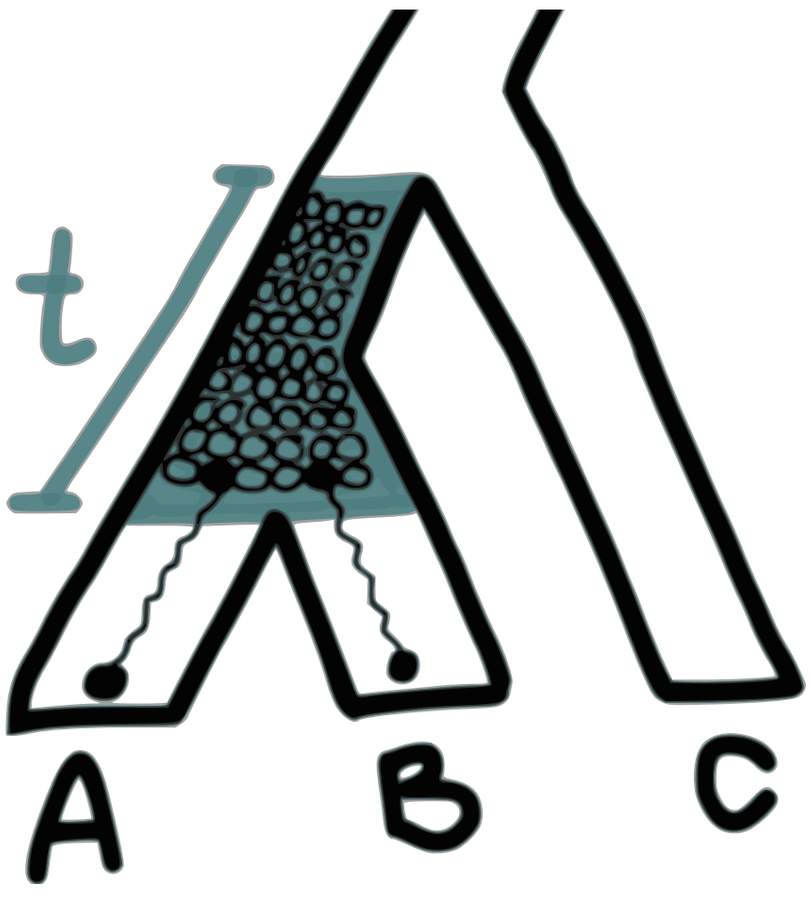
$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) =$$

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

$$+$$

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

Multispecies coalescent on a tree



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

$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) =$$

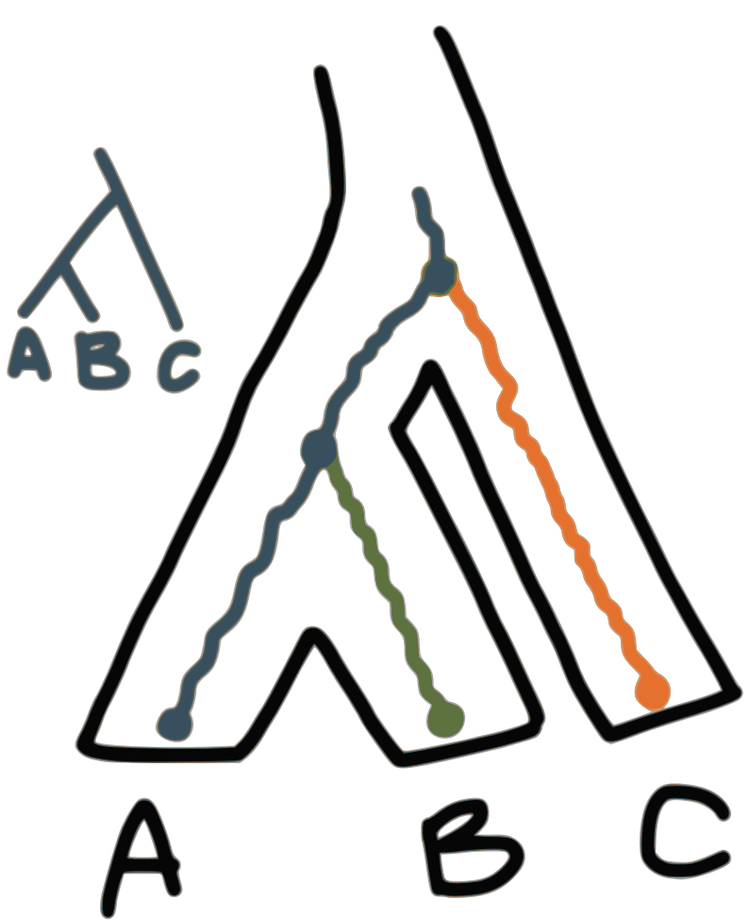
$$1 - e^{-t}$$

$$+$$

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

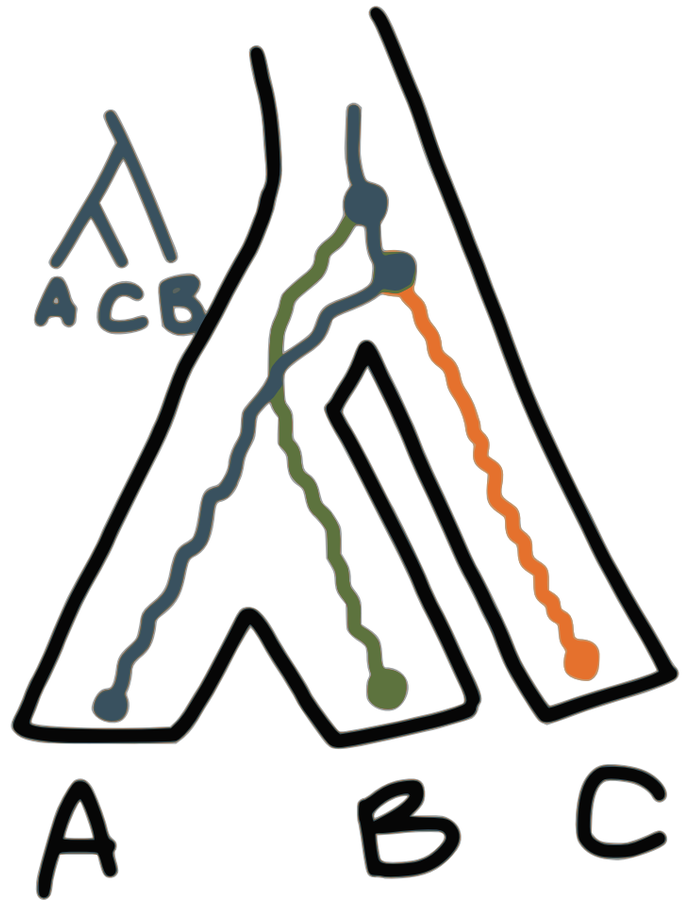
$$= 1 - \frac{2}{3} e^{-t}$$

Multispecies coalescent on a tree



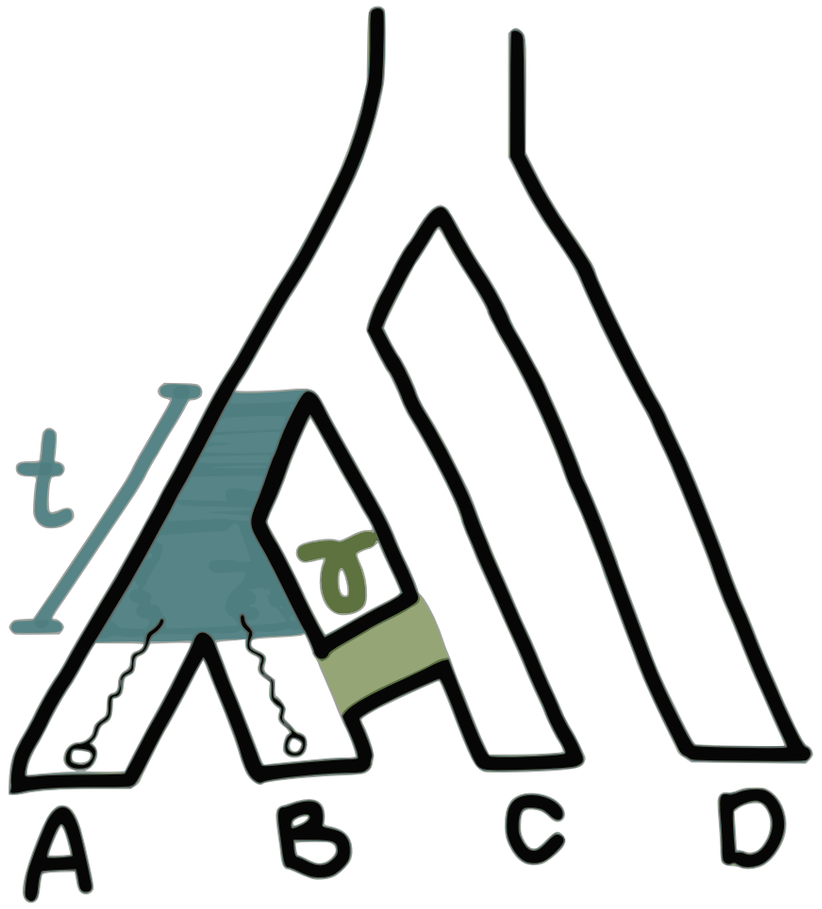
$$1 - \frac{2}{3}e^{-t}$$

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



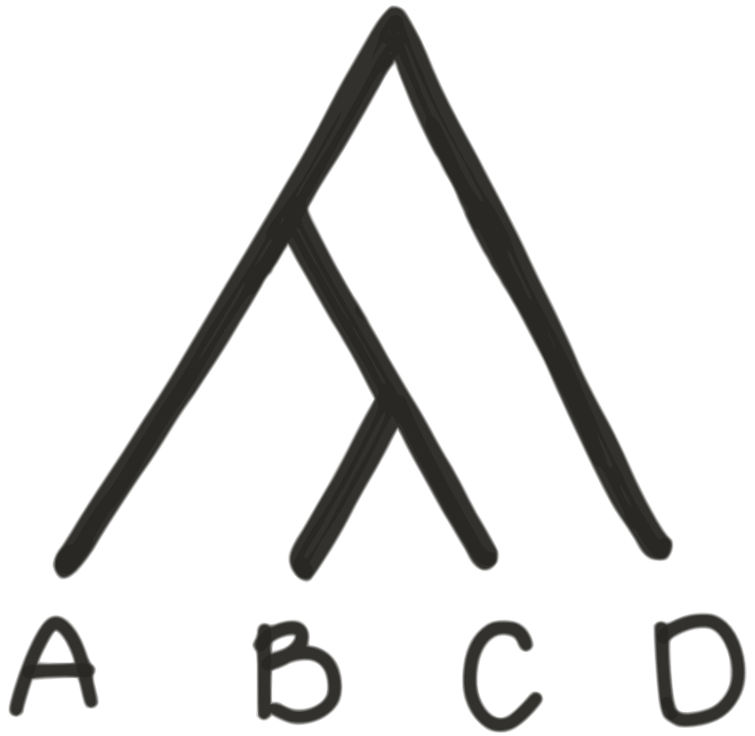
$$\frac{1}{3}e^{-t}$$

Multispecies coalescent on a network



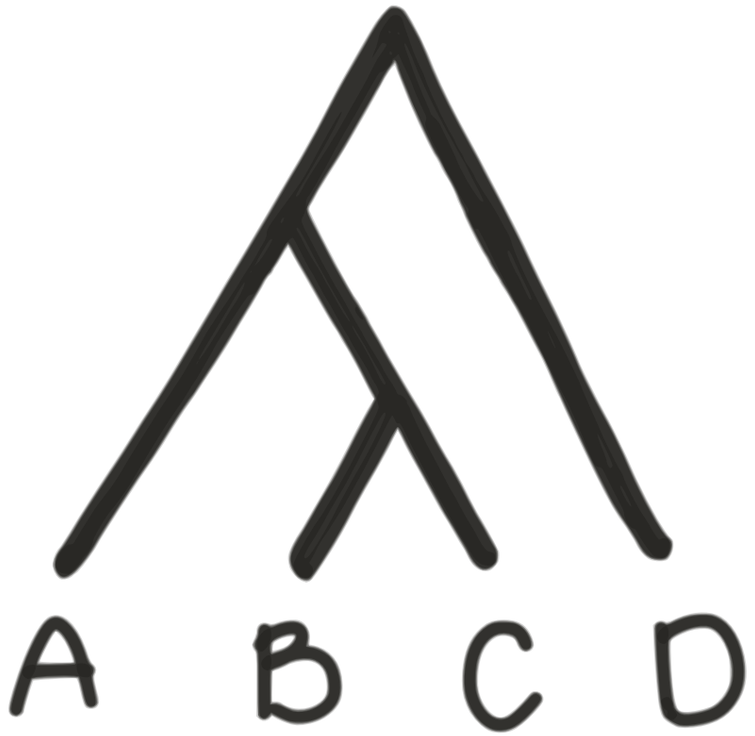
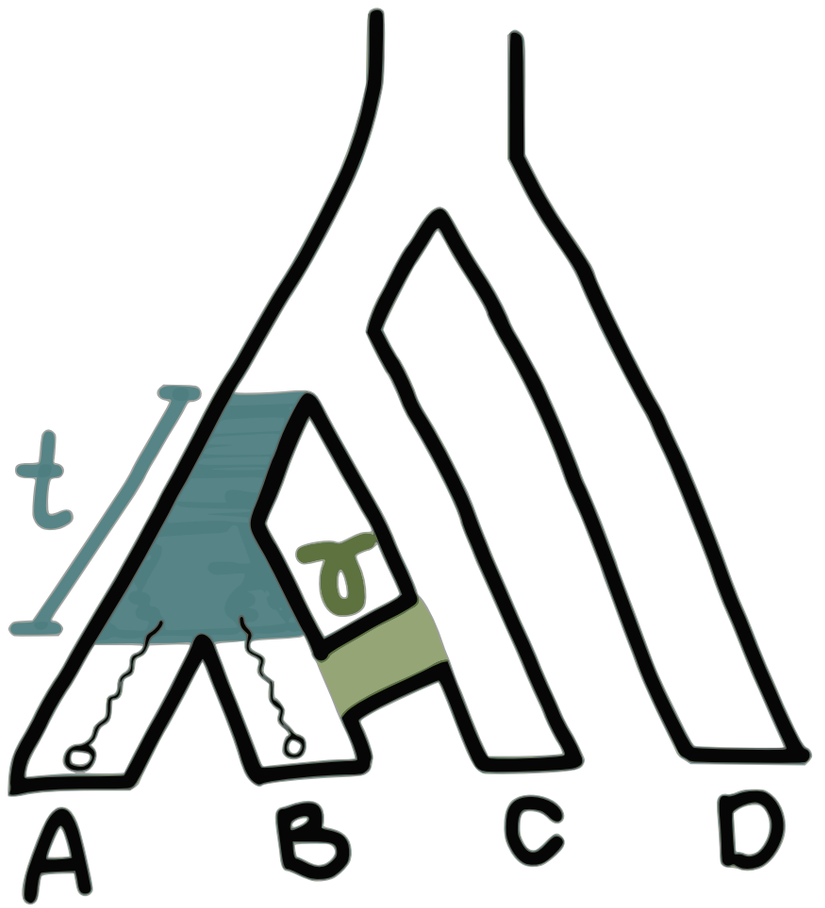
(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

Multispecies coalescent on a network

$P($  $|$  $)$

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

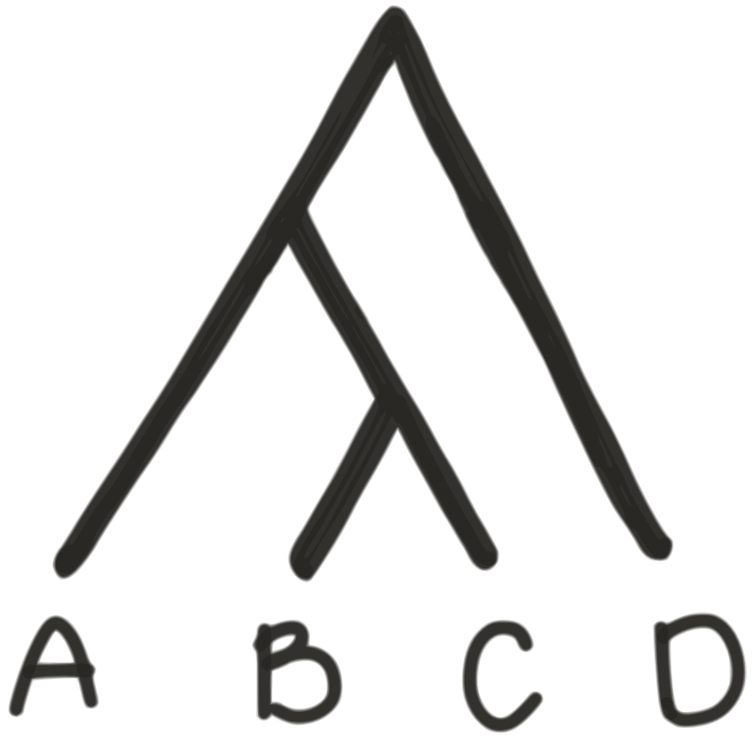
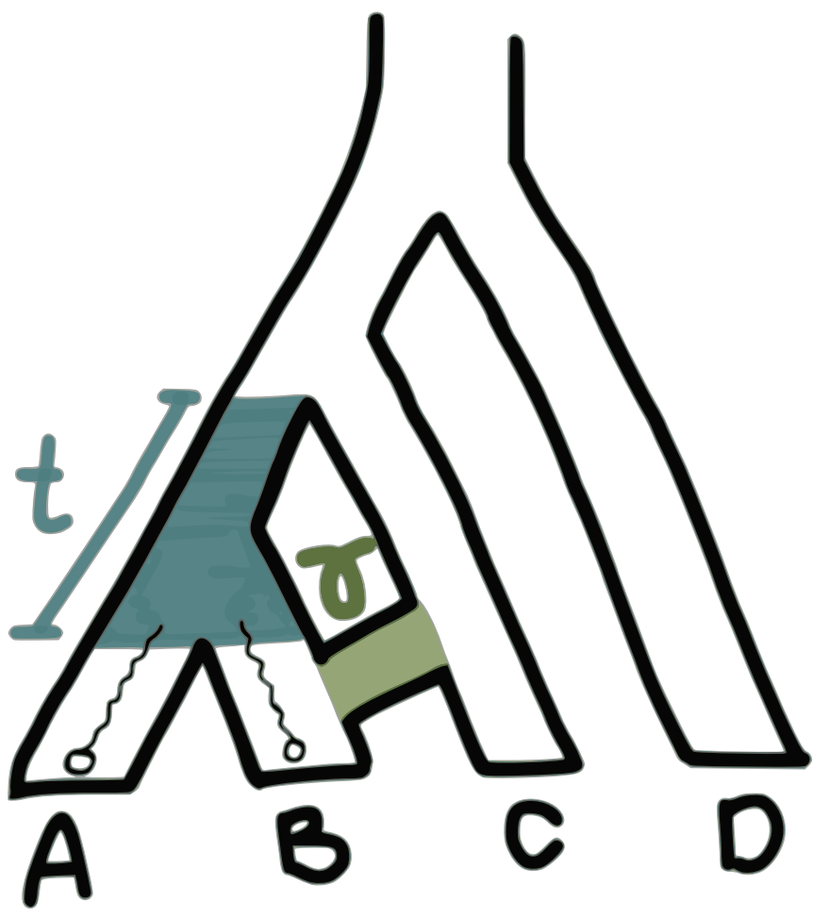
Multispecies coalescent on a network

$P($  $|$  $)$

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

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

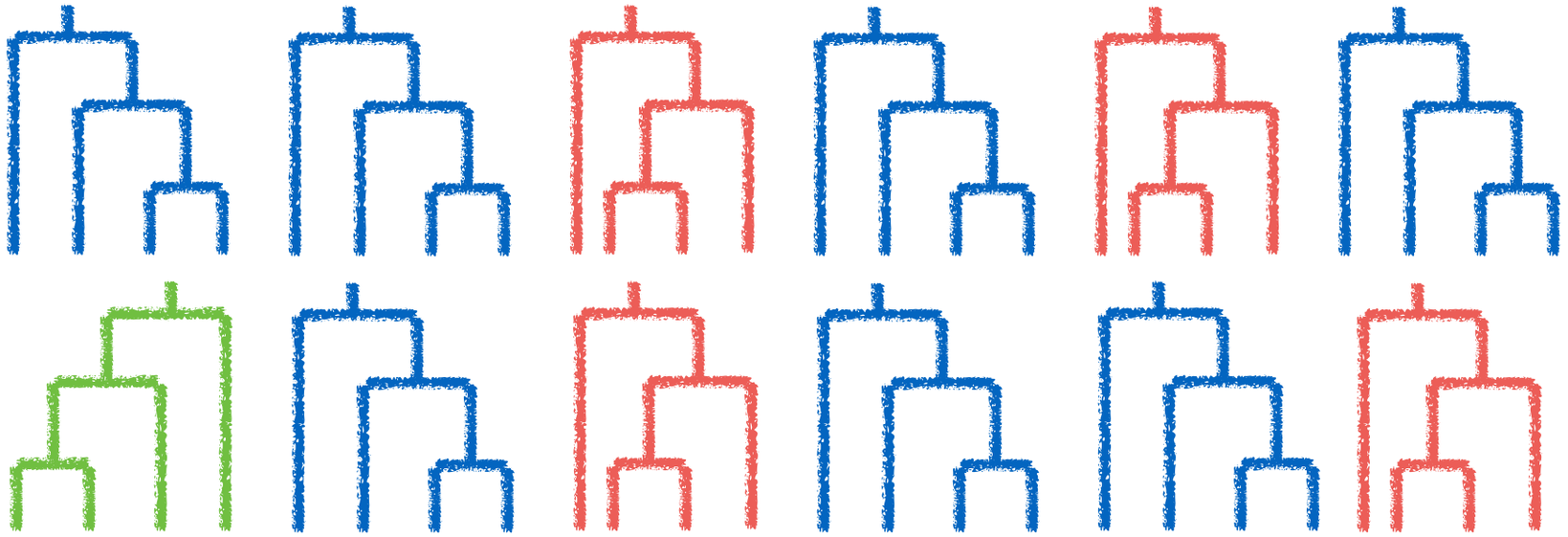
Multispecies coalescent on a network

$P($  $|$  $)$

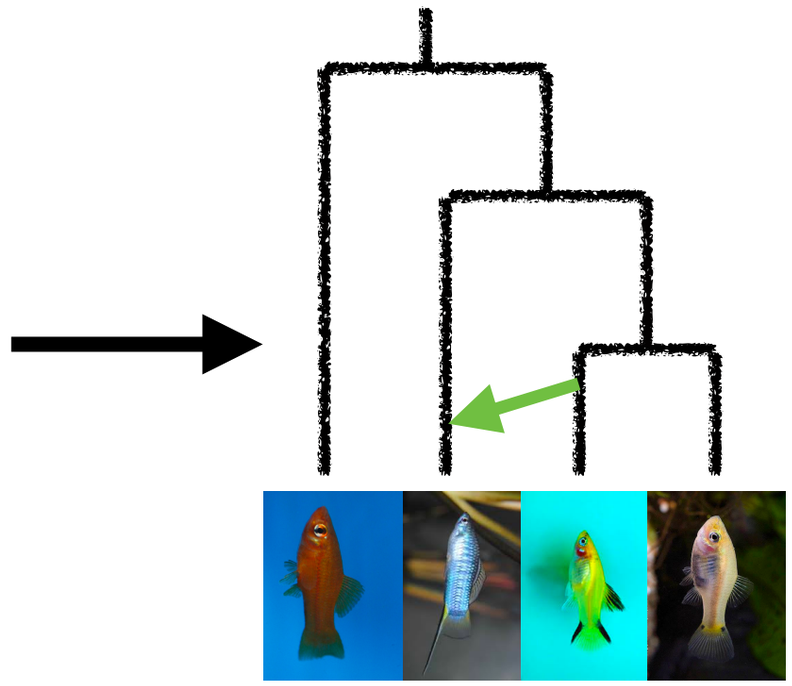
$$CF_{BC|AD}(t, t_2, \gamma) = (1-\gamma)\frac{1}{3}e^{-t} + \gamma(1 - \frac{2}{3}e^{-t_2})$$

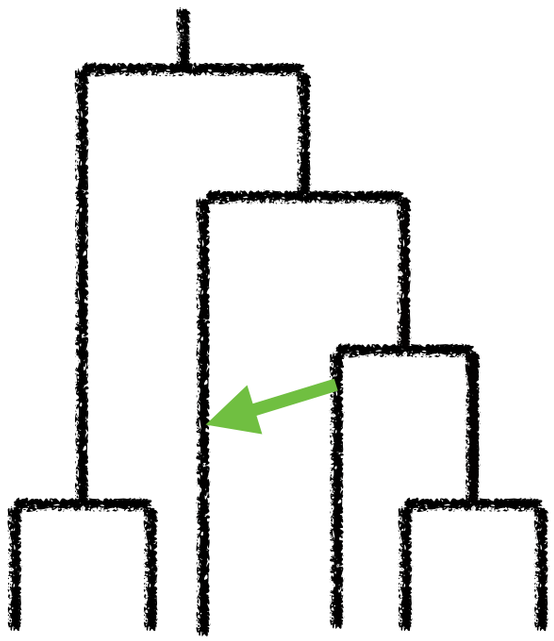
(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

Maximum **pseudo**likelihood



Data





Quartet-based inference

$$\tilde{L}(network) = \prod L(quartet)$$

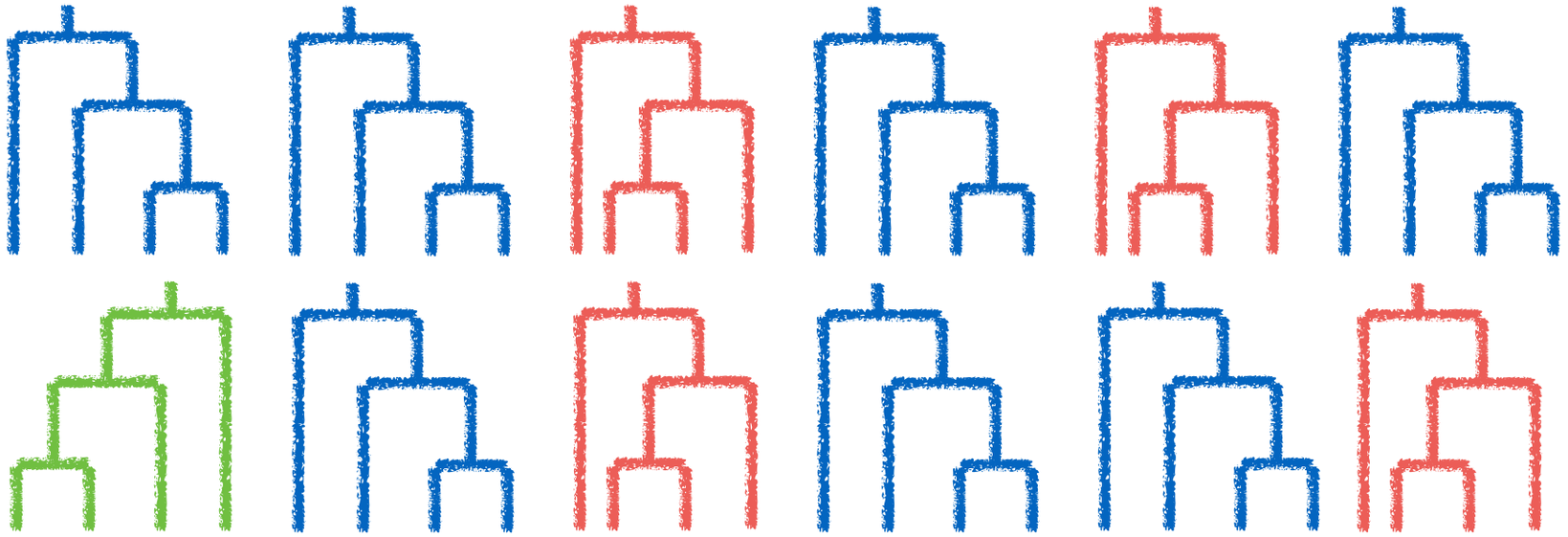
(S-L, Ané, 2016, PLoS Genetics)

`www.github.com/CRSL4/PhyloNetworks`

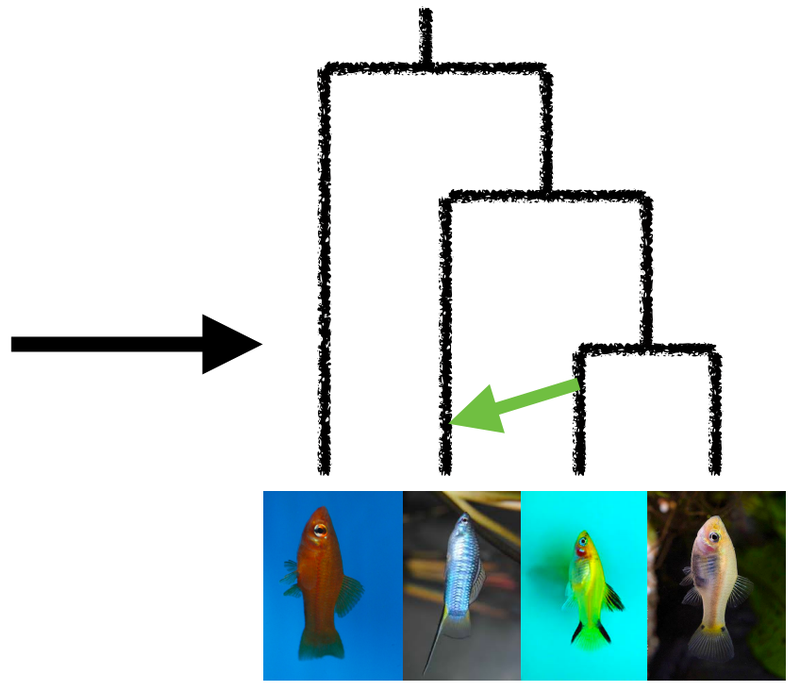
SNAQ julia

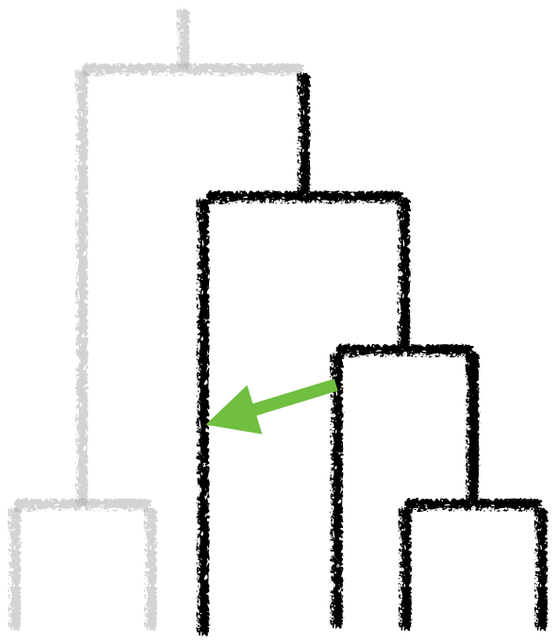
https://solislemuslab.github.io/ @solislemuslab crsl4

Maximum **pseudo**likelihood



Data





Quartet-based inference

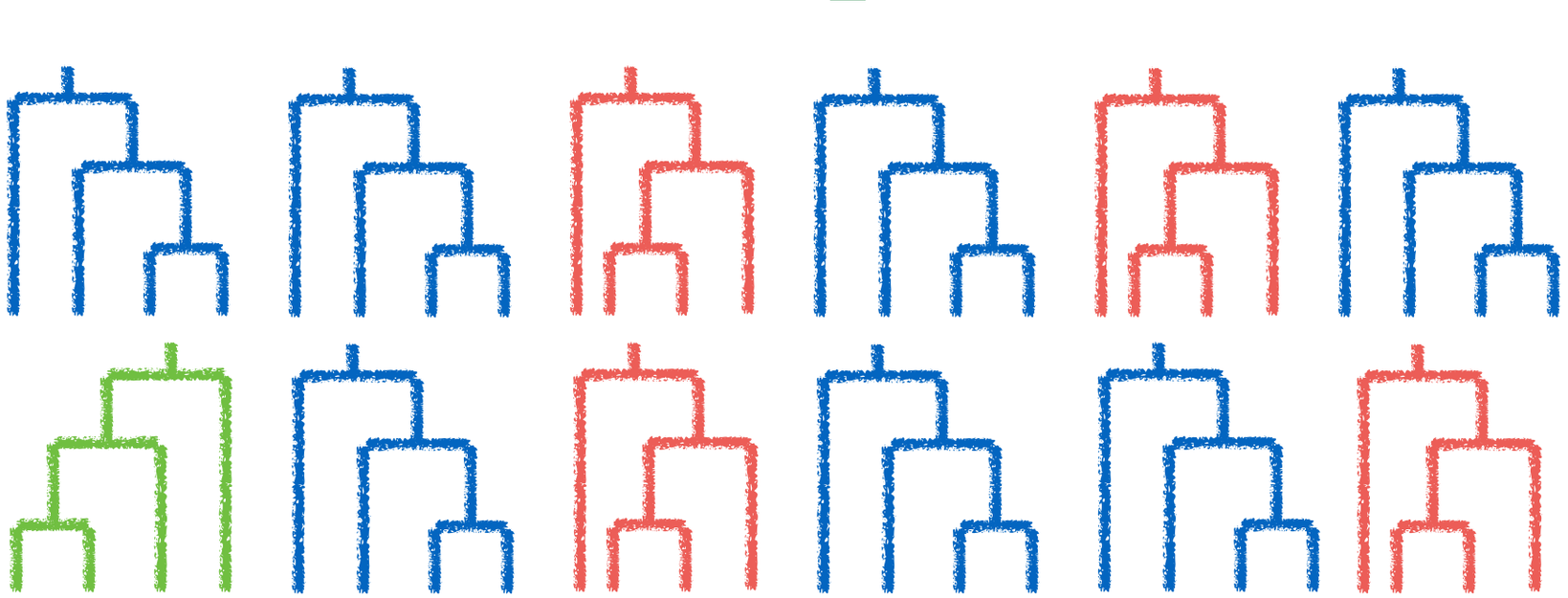
$$\tilde{L}(network) = \prod L(quartet)$$

(S-L, Ané, 2016, PLoS Genetics)

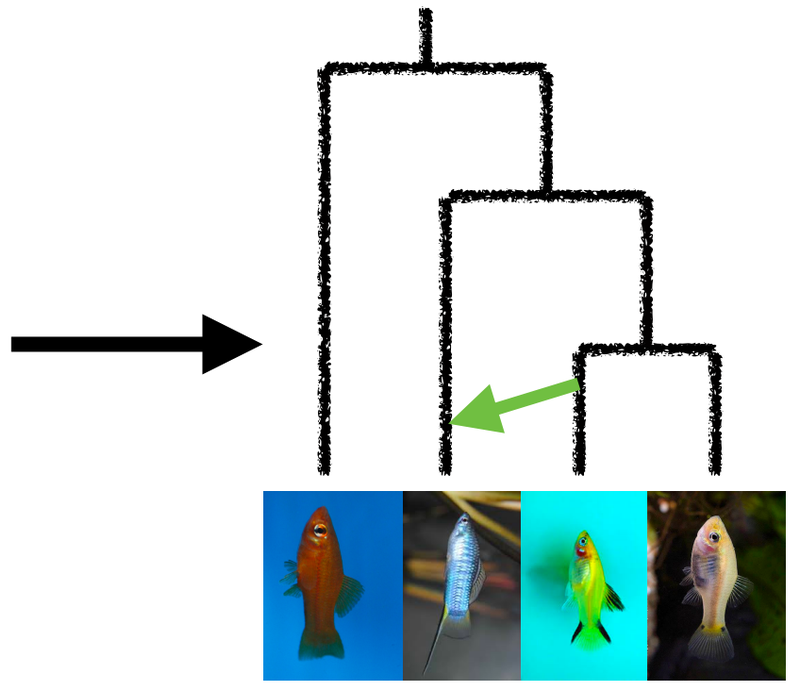
`www.github.com/CRSL4/PhyloNetworks`

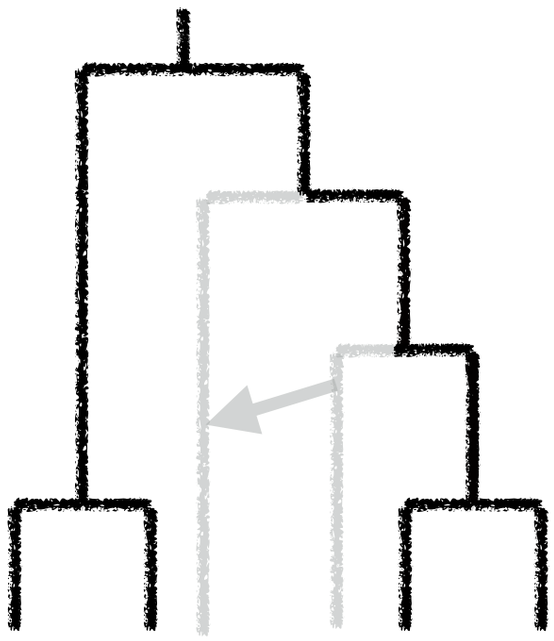
https://solislemuslab.github.io/ @solislemuslab crsl4

Maximum **pseudo**likelihood



Data





Quartet-based inference

$$\tilde{L}(network) = \prod L(quartet)$$

(S-L, Ané, 2016, PLoS Genetics)

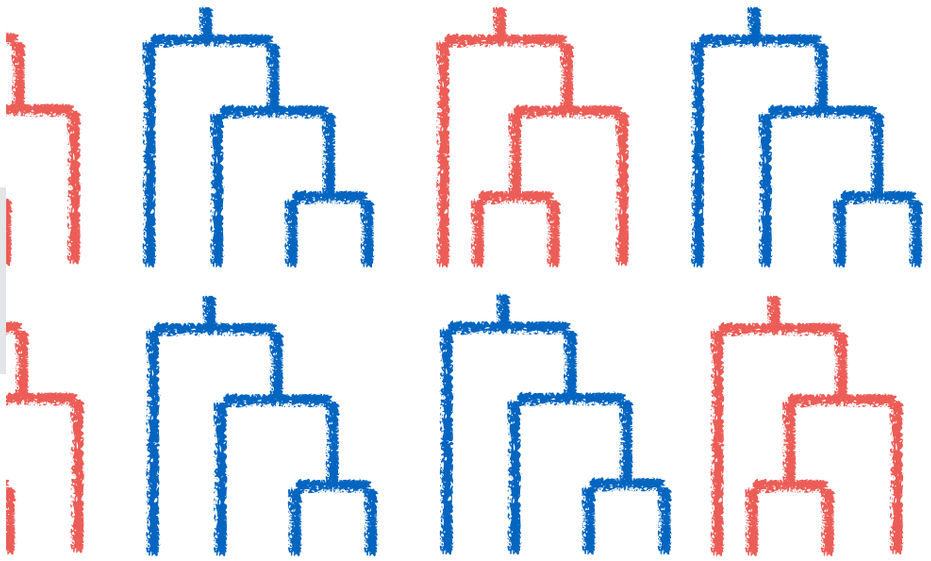
`www.github.com/CRSL4/PhyloNetworks`

SNAQ julia

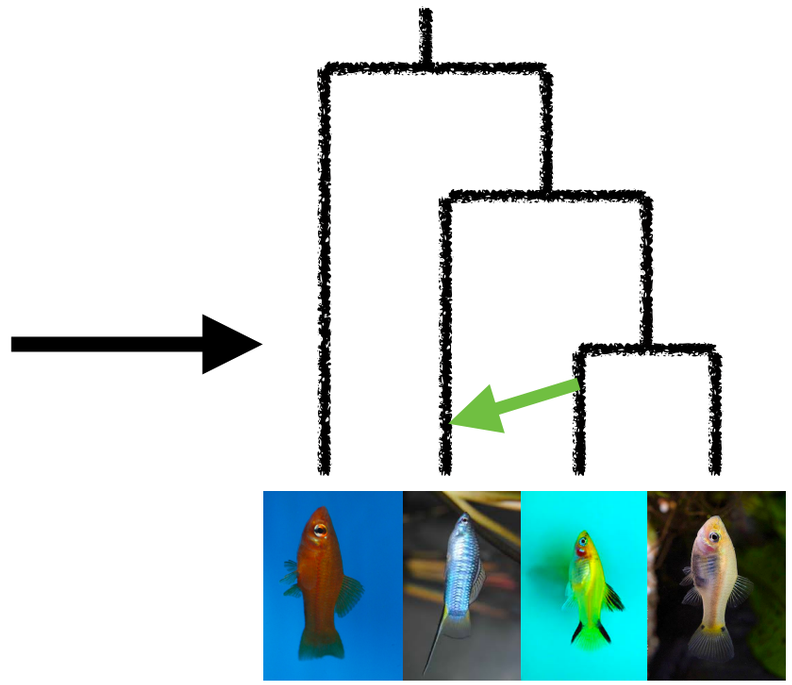
https://solislemuslab.github.io/ @solislemuslab crsl4

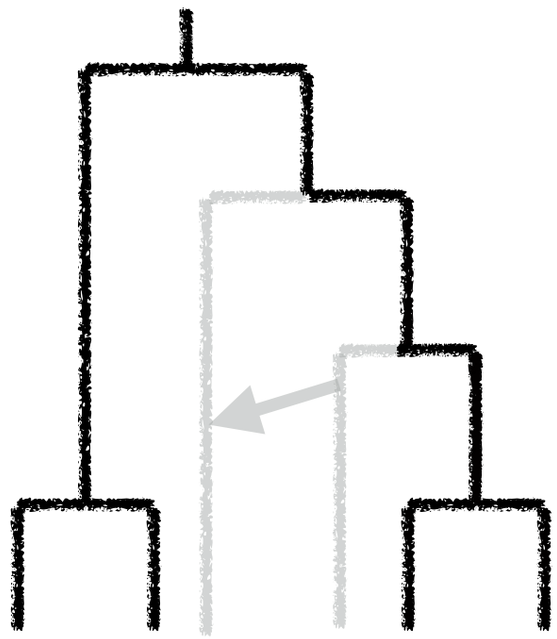
Maximum **pseudo**likelihood

Unrooted gene trees	No rooting error
No branch lengths	No molecular clock assumption
Concordance factors	Account for tree estimation error



Data





Quartet-based inference

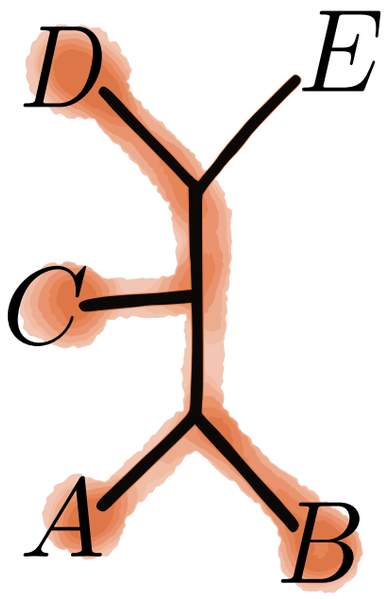
$$\tilde{L}(network) = \prod L(quartet)$$

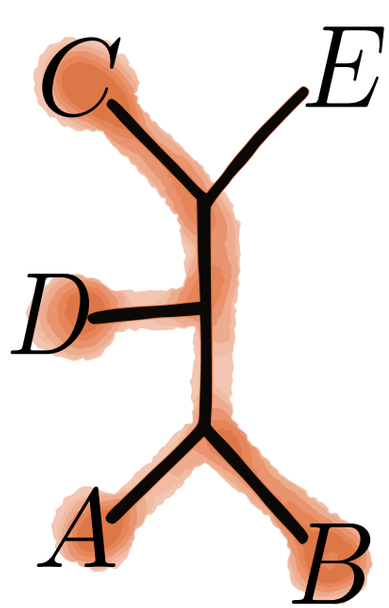
(S-L, Ané, 2016, PLoS Genetics)

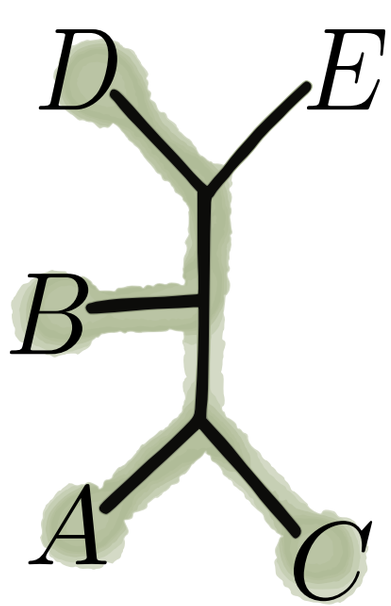
`www.github.com/CRSL4/PhyloNetworks`

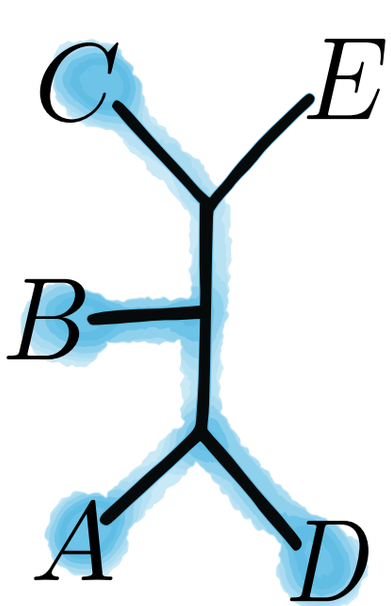
https://solislemuslab.github.io/ @solislemuslab crsl4

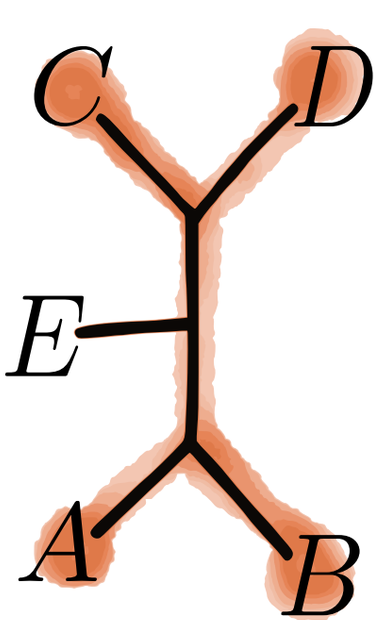
Quartet-based inference



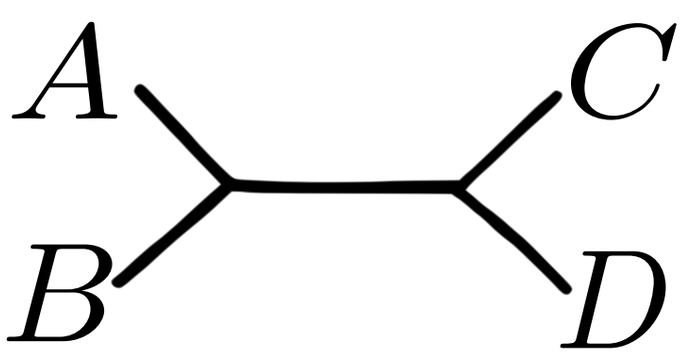




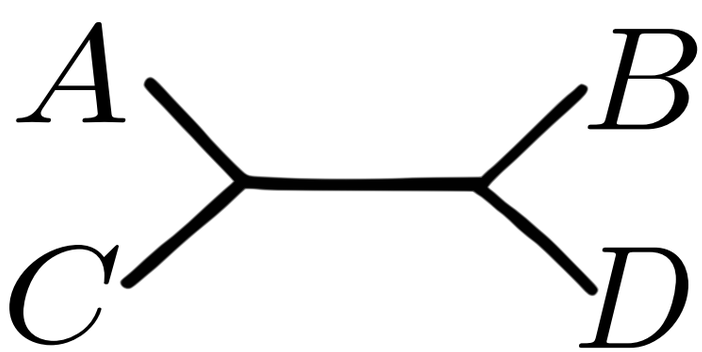




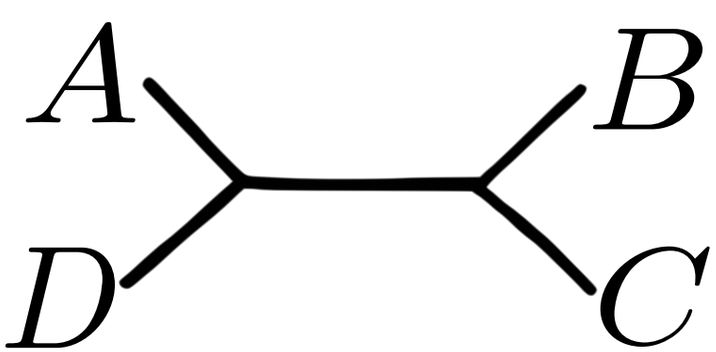
Concordance factors (CF):
% of genes having the quartet in their tree



3/5



1/5



1/5

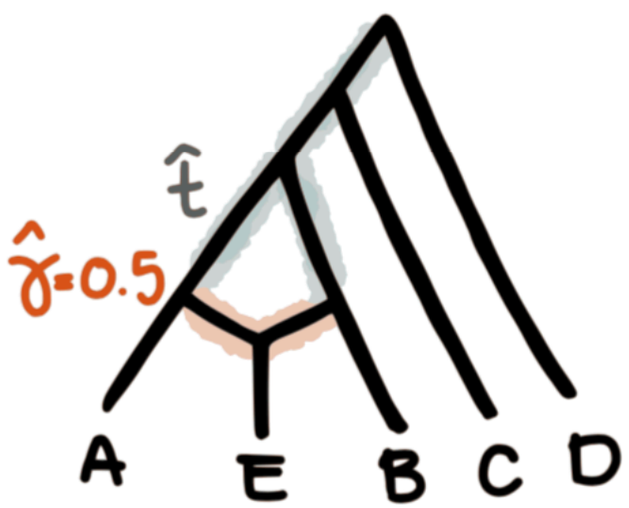
Quartet-based inference

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

$\longrightarrow$

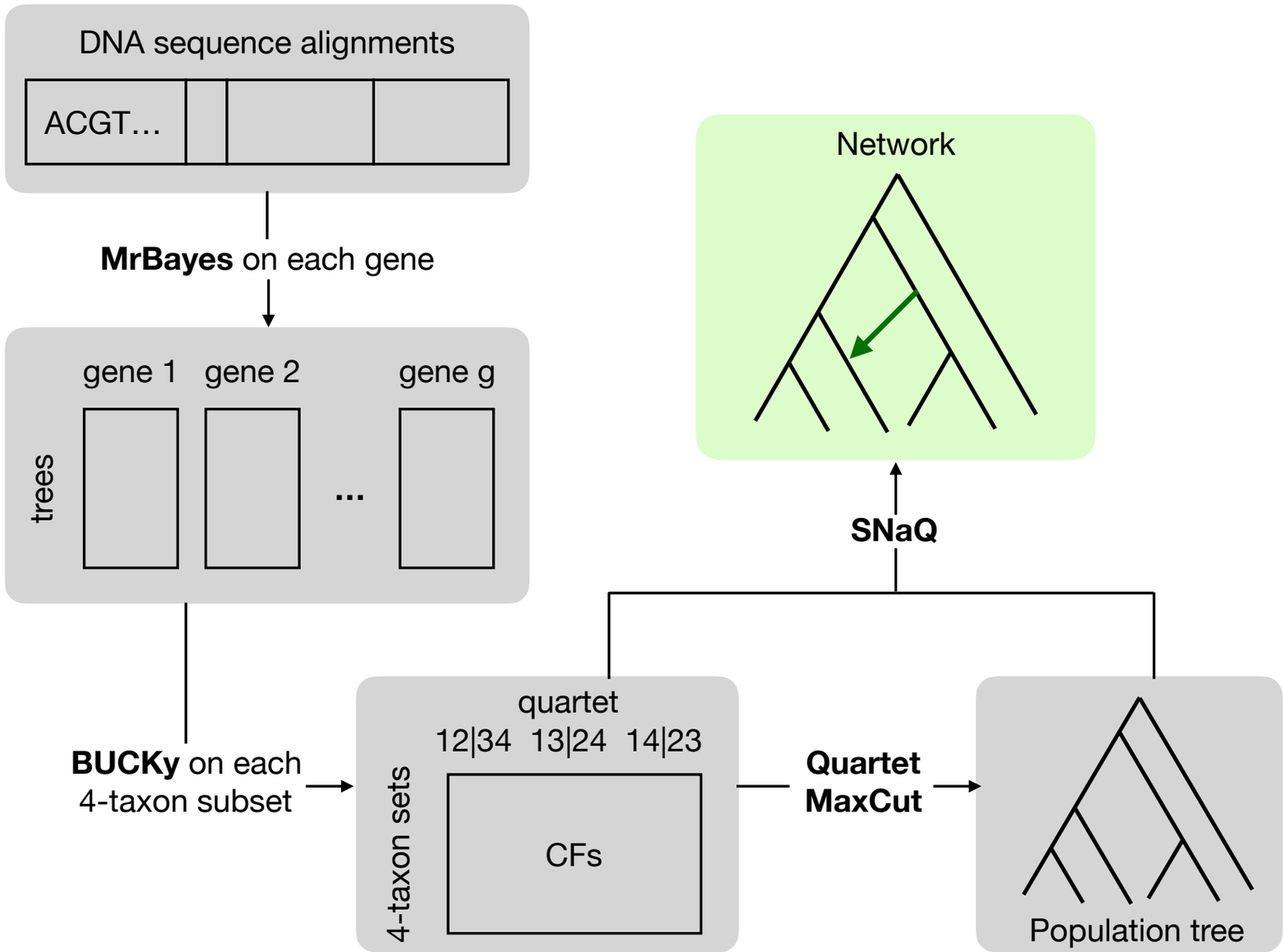
inferred network:



Maximum Pseudo-Likelihood:

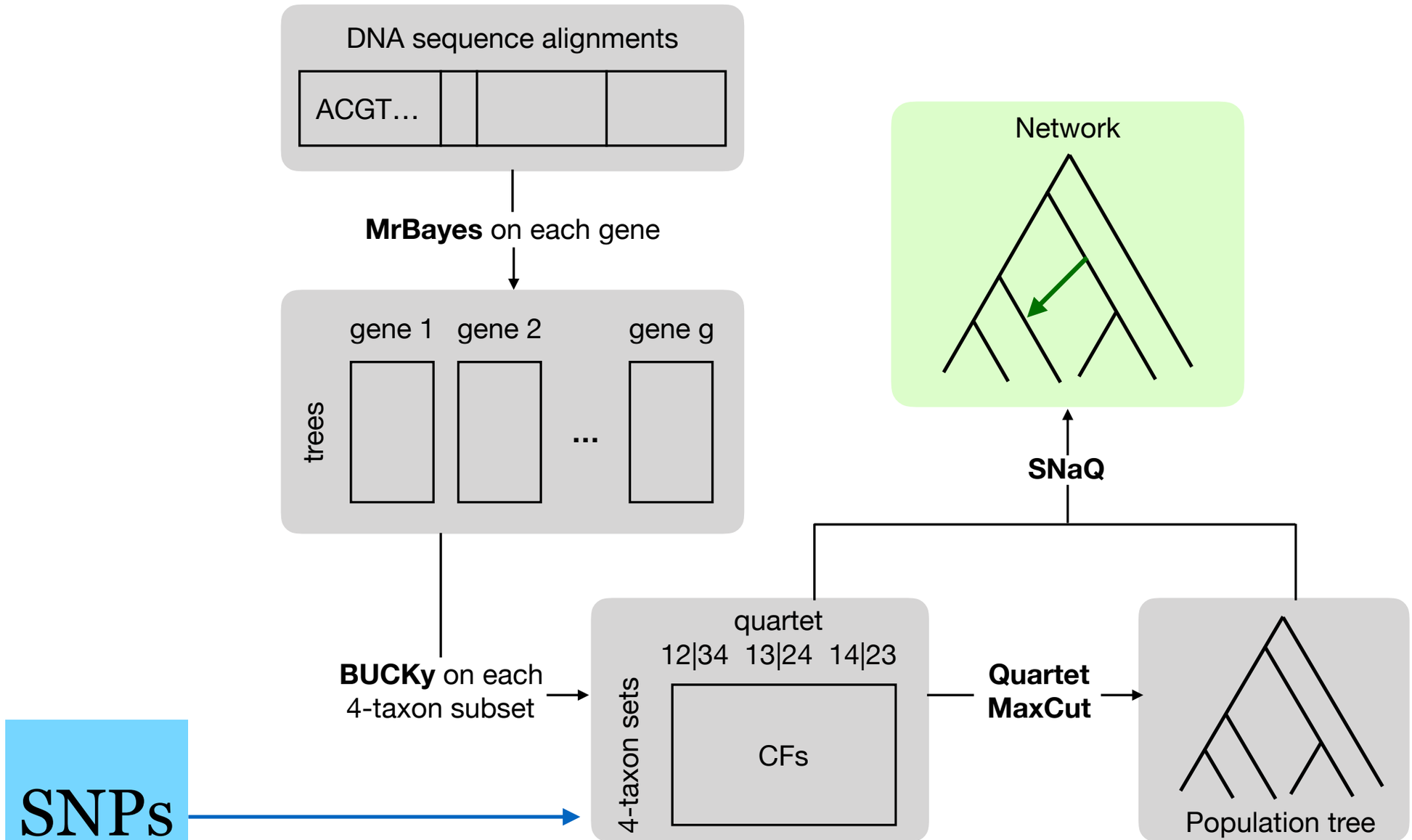
$$\log \tilde{L} = \sum_{q \in Q(N)} \mathrm{CF}_{in,1} \log(\mathrm{CF}_{net,1}) + \mathrm{CF}_{in,2} \log(\mathrm{CF}_{net,2}) + \mathrm{CF}_{in,3} \log(\mathrm{CF}_{net,3})$$

How?
Phylogenetic network



How?

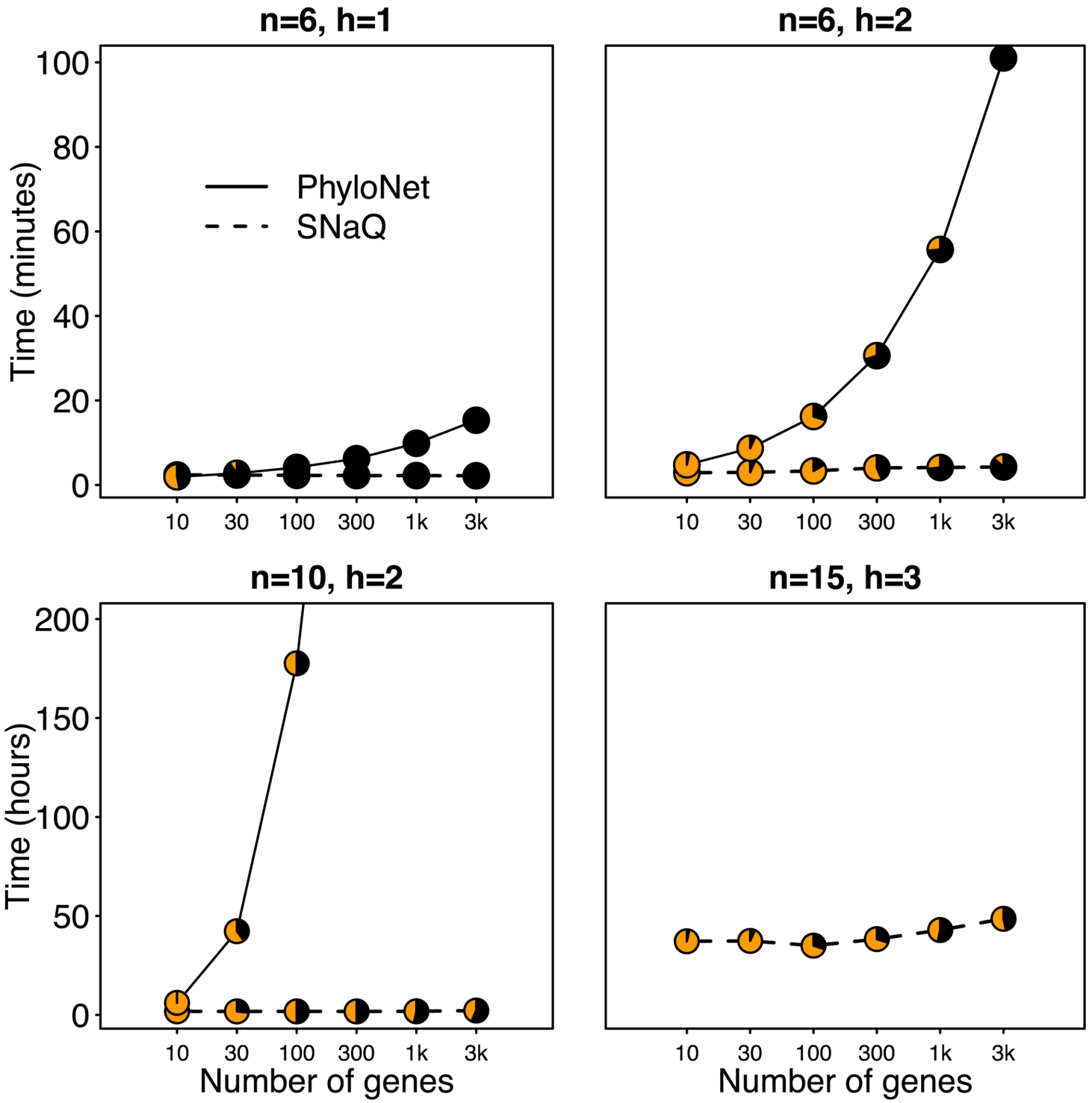
Phylogenetic network



Network challenges

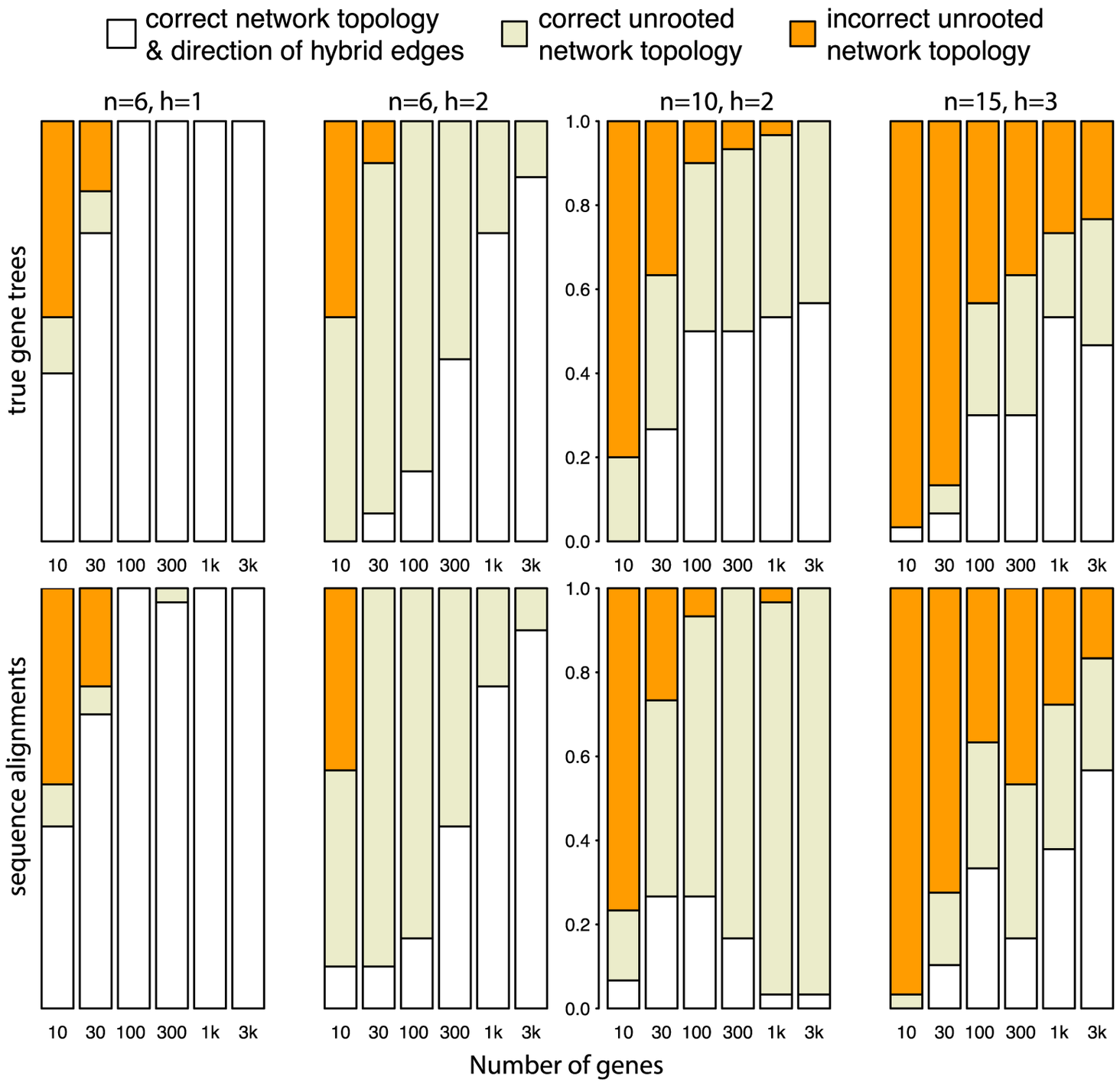
- Scalability
- Identifiability
- Network space
- Network comparison

Scalability gains



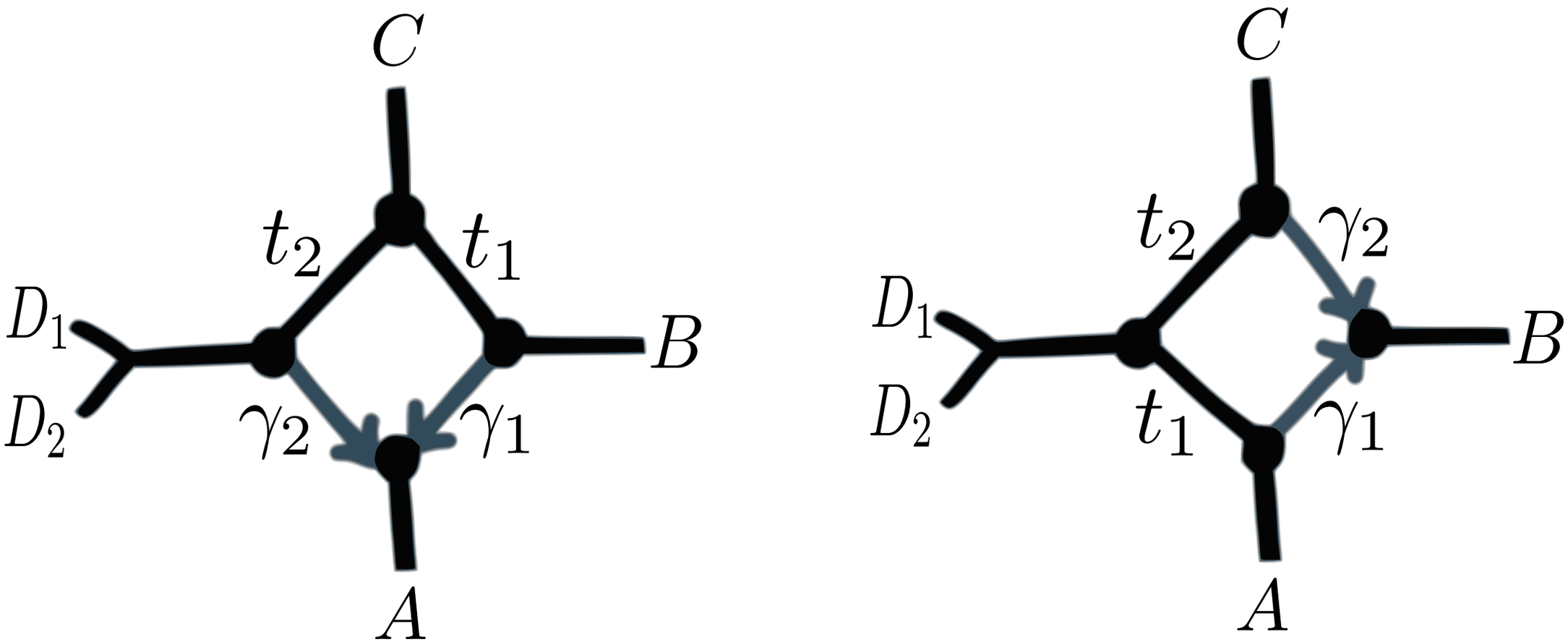
(Solís-Lemus, Ané, 2016, PLoS Genetics)

Accuracy



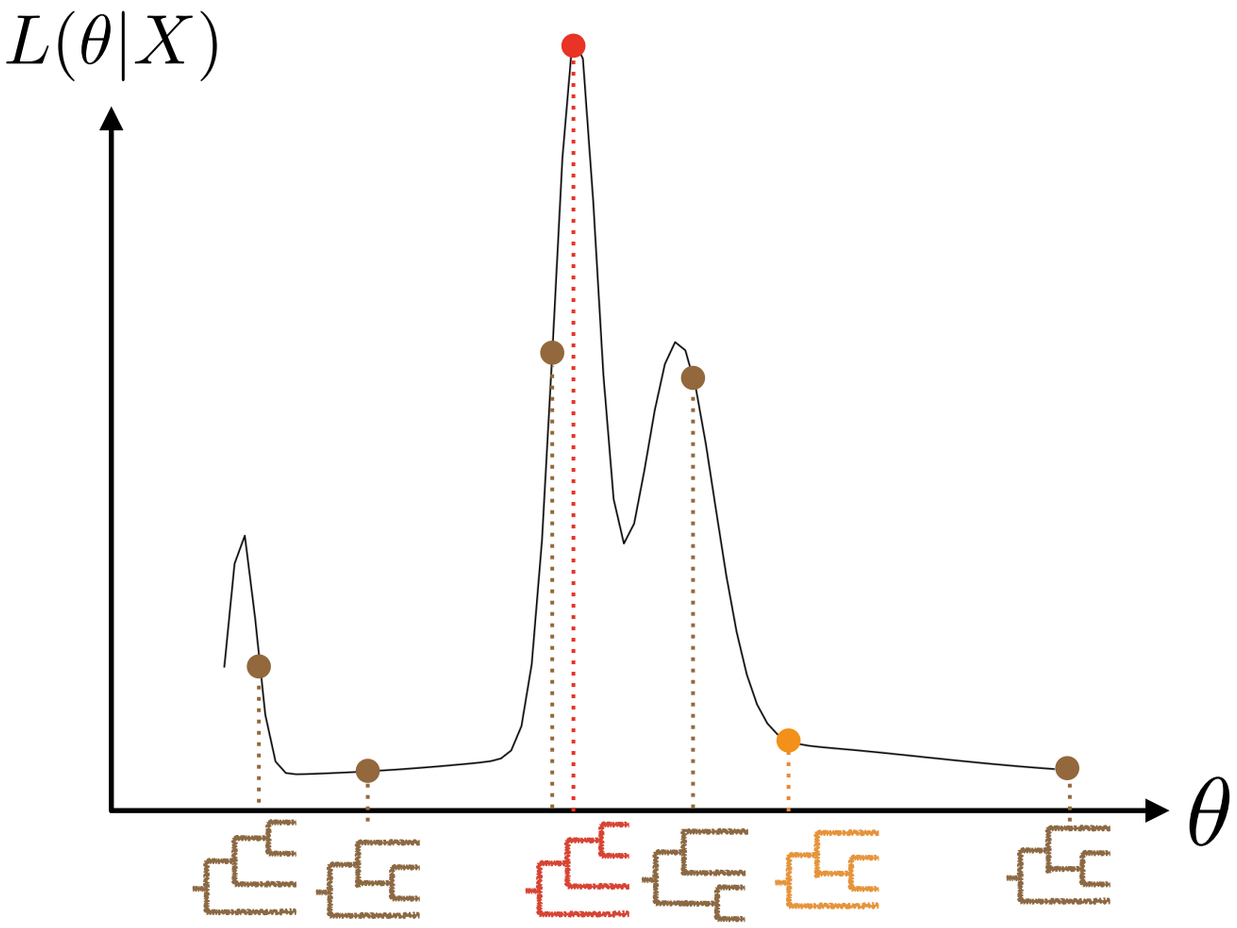
(Solís-Lemus, Ané, 2016, PLoS Genetics)

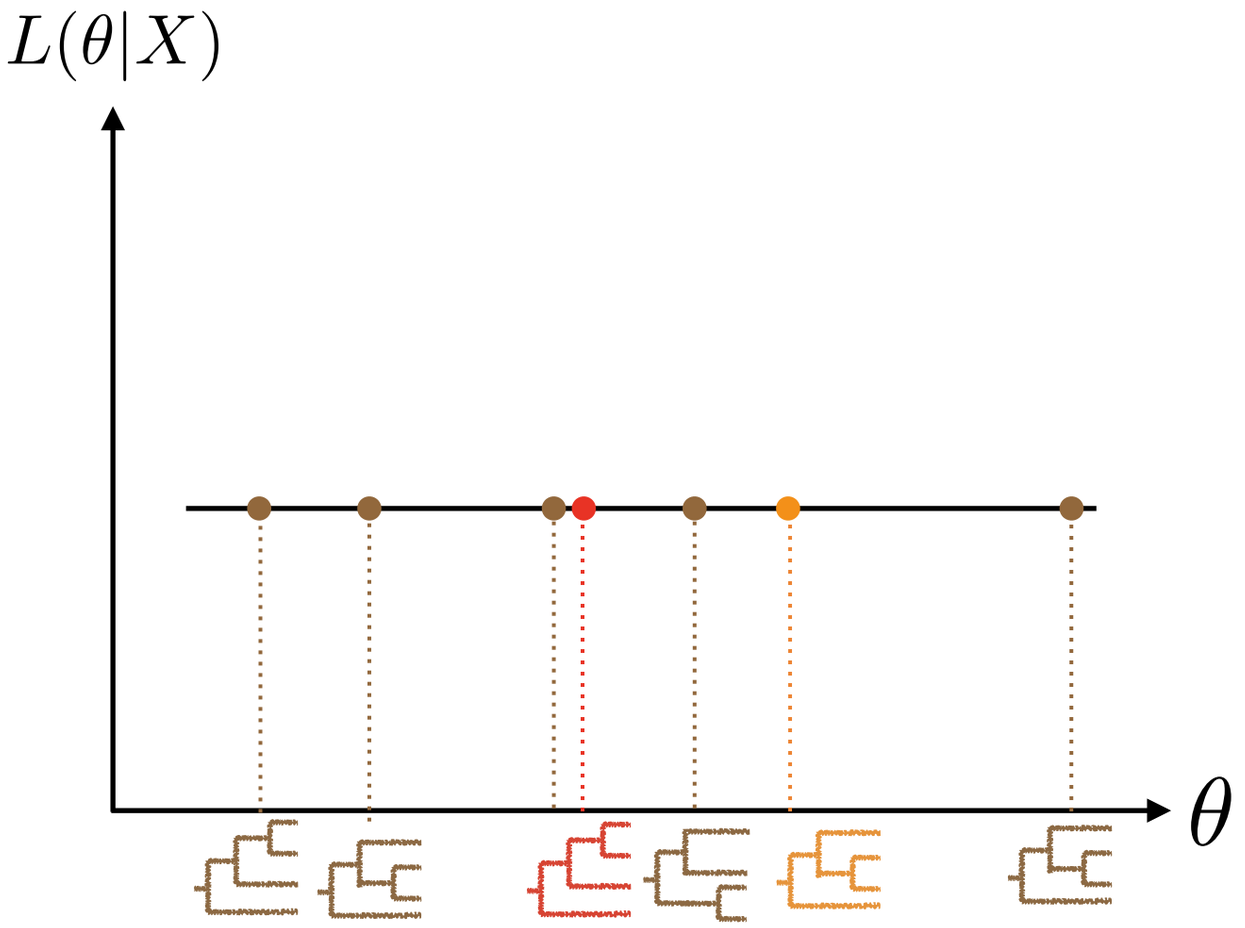
In practice: flat pseudolikelihood



(S.-L., Ané, 2016, PLoS Genetics)

Identifiability





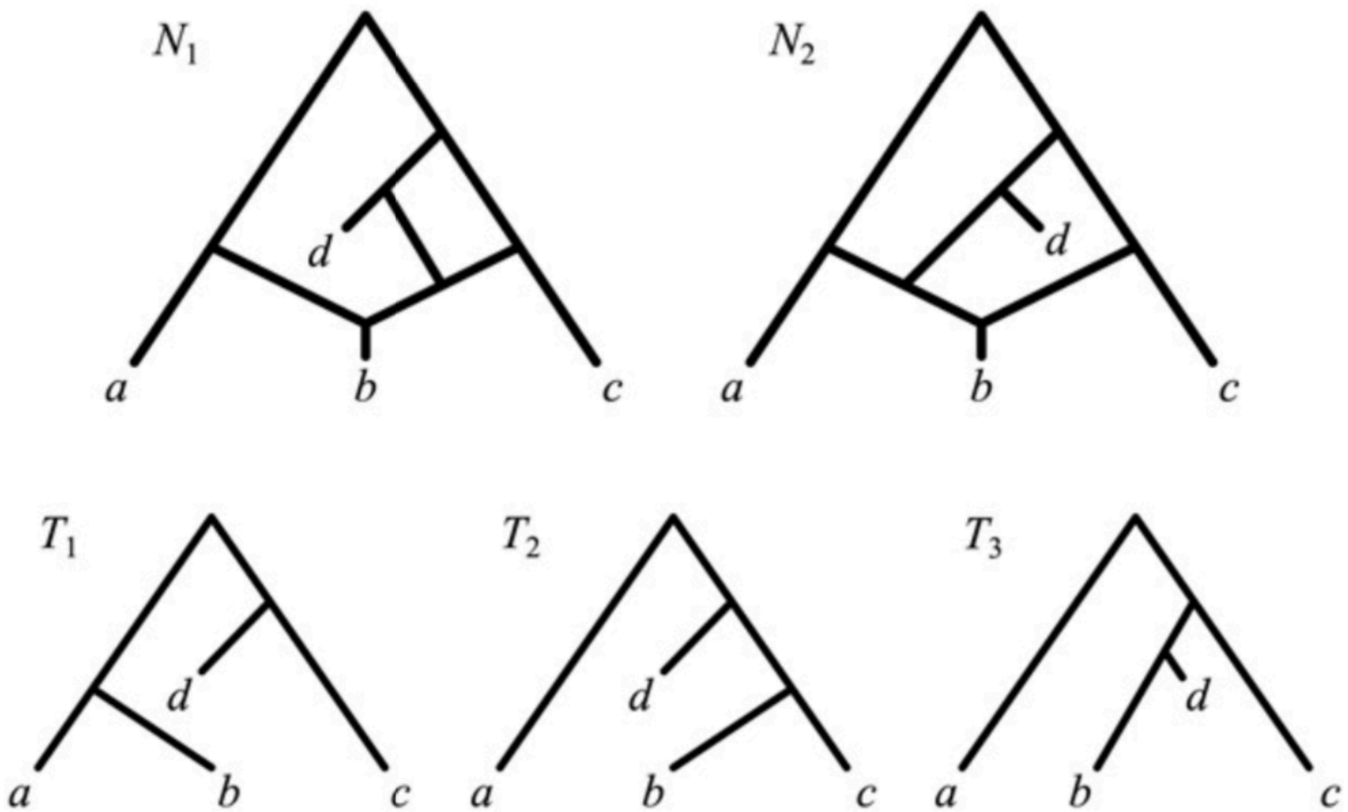
PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

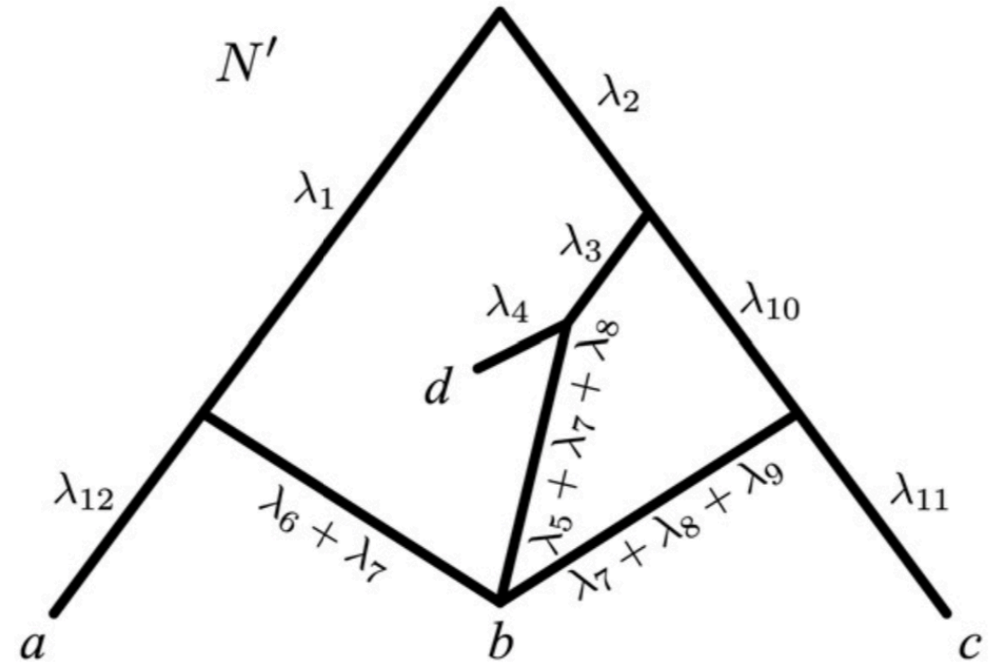
Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable

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 Computationnelle, Montpellier, France



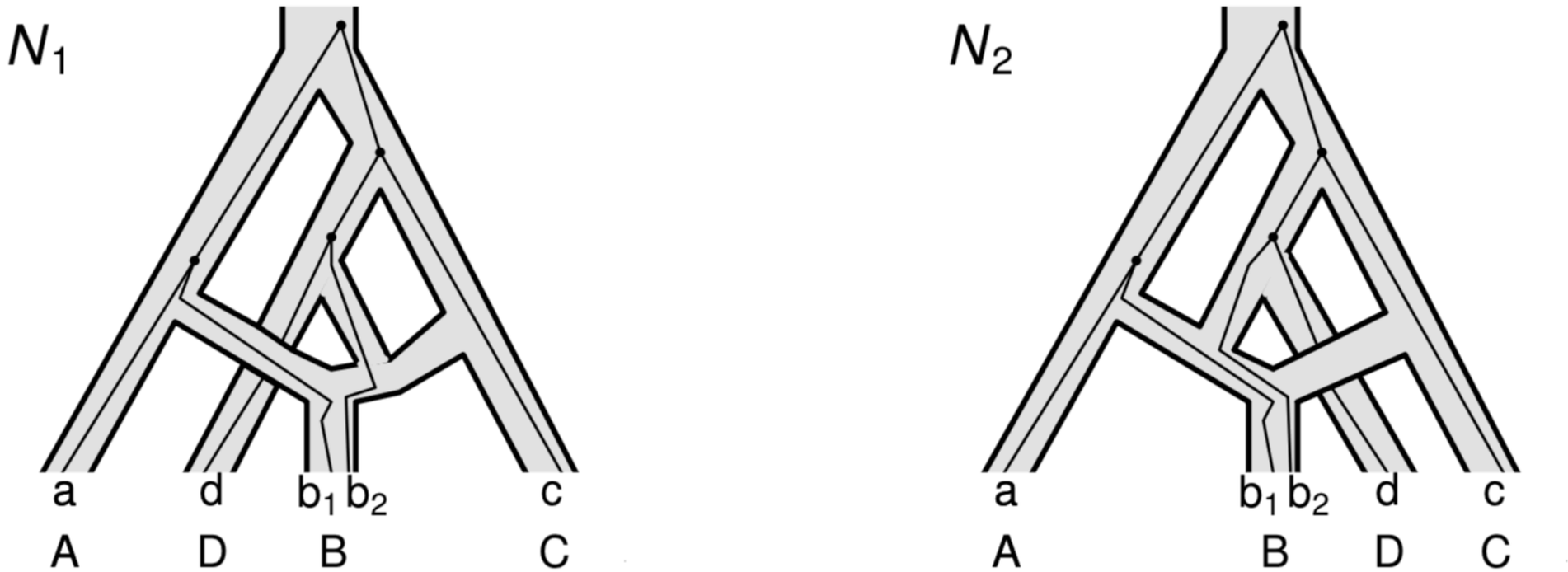
Undistinguishable with the "displayed trees" criterion



Solution: Canonical network ("unzipped")

Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

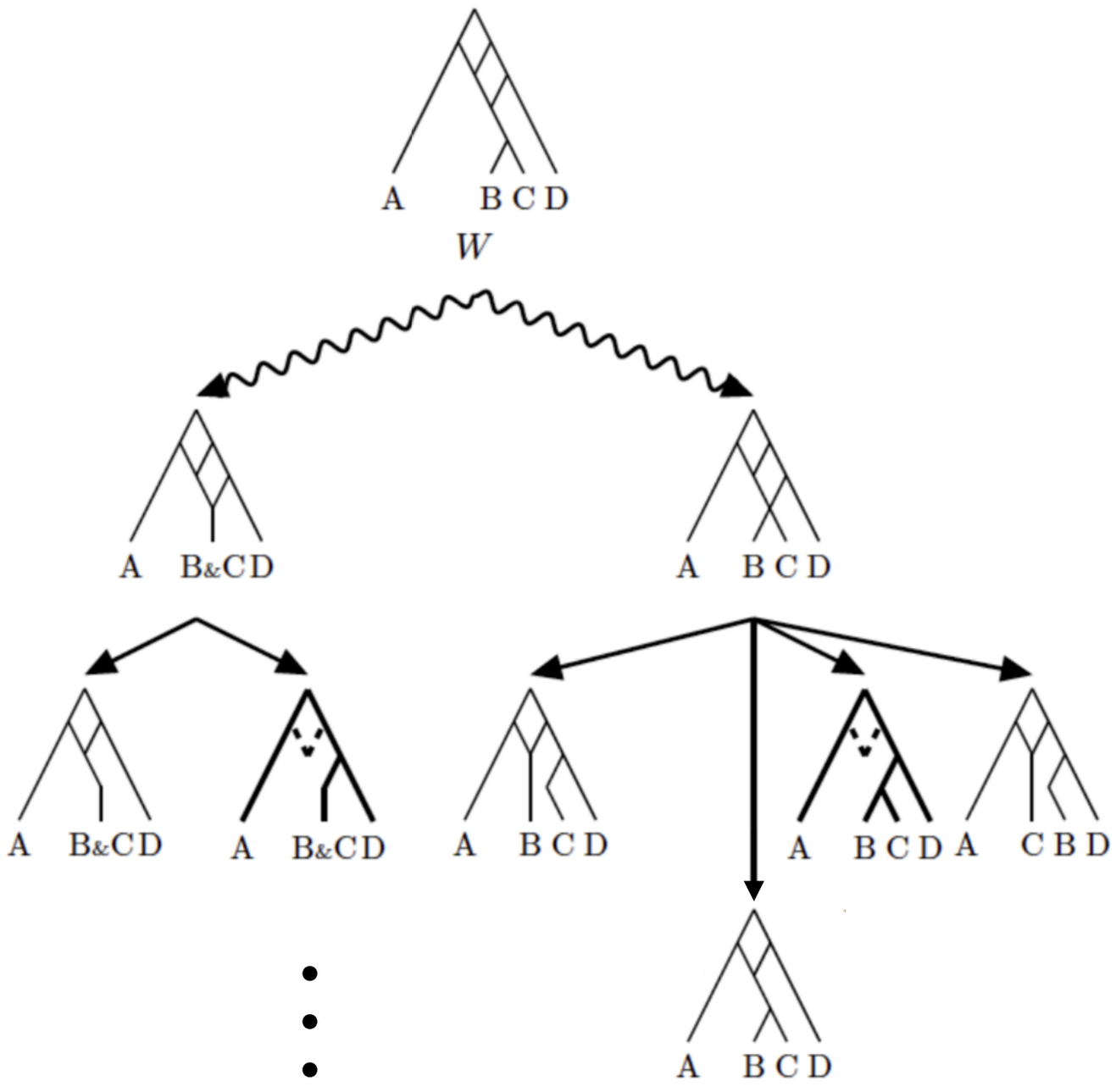
Sha Zhu[1], James H. Degnan[2]



Distinguishable under the MSC

Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

Sha Zhu[1], James H. Degnan[2]



Decomposing network in **parental** trees

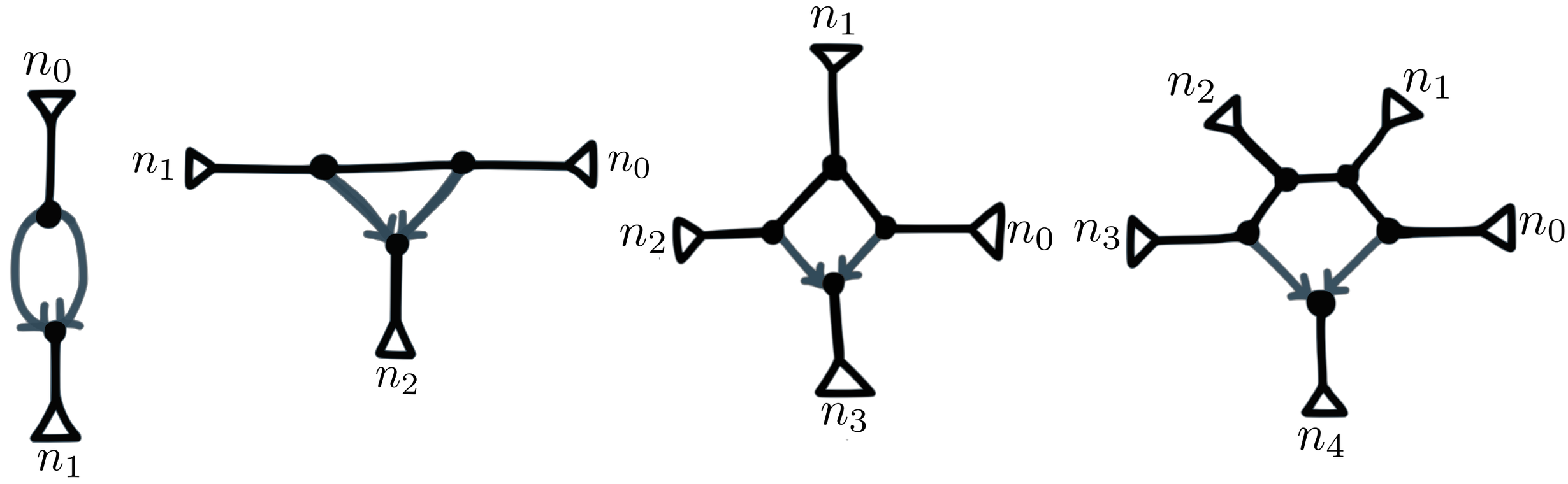
PLOS | GENETICS

RESEARCH ARTICLE

Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting

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

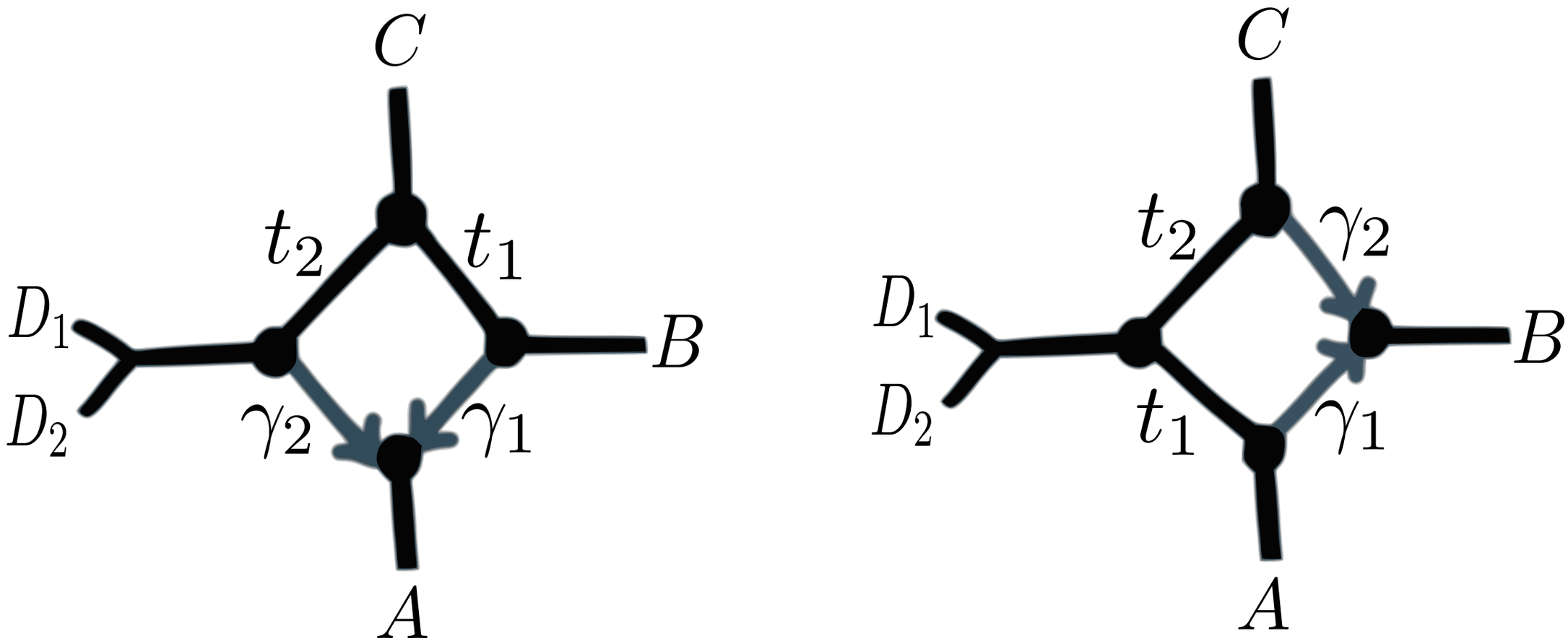
Can we detect the presence of hybridization in level-1 networks?



No **Yes** $(n_i, n_j \geq 2)$ **Yes** $(n_i \geq 2)$ **Yes**

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

In practice: flat pseudolikelihood



(S.-L., Ané, 2016, PLoS Genetics)

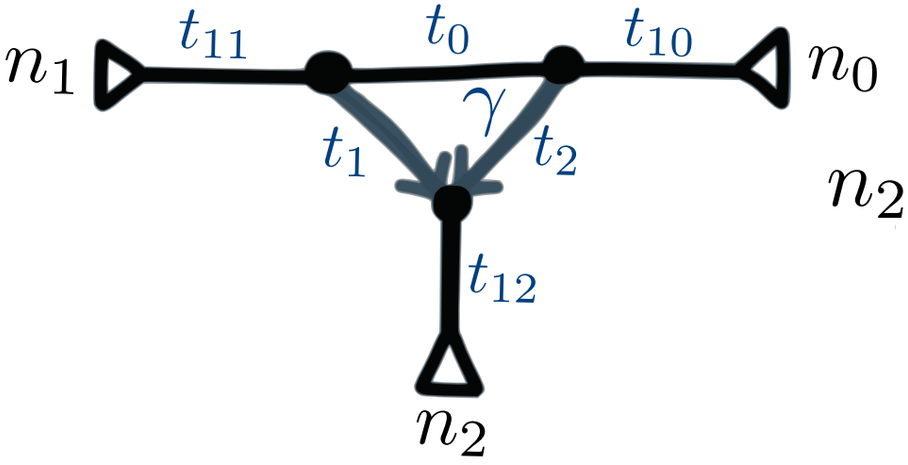
PLOS | GENETICS

RESEARCH ARTICLE

Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting

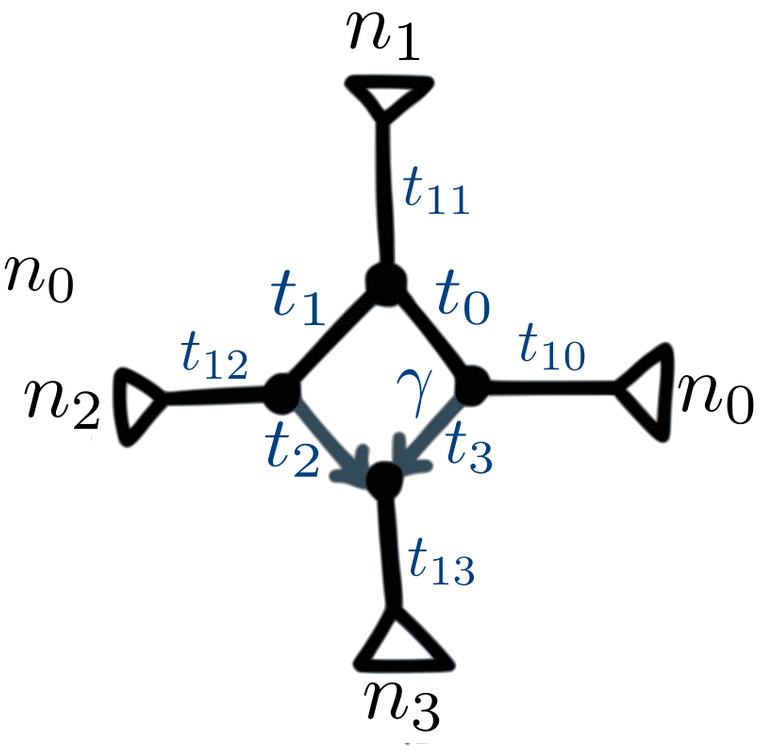
Claudia Solís-Lemus[1]*, Cécile Ané[1,2]

Can we estimate numerical parameters?



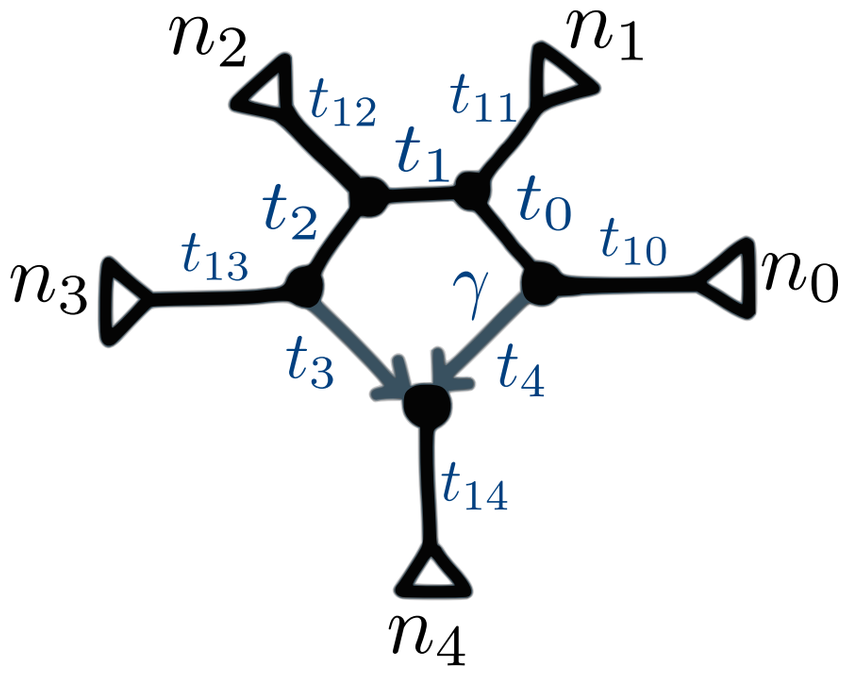
No

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



Yes

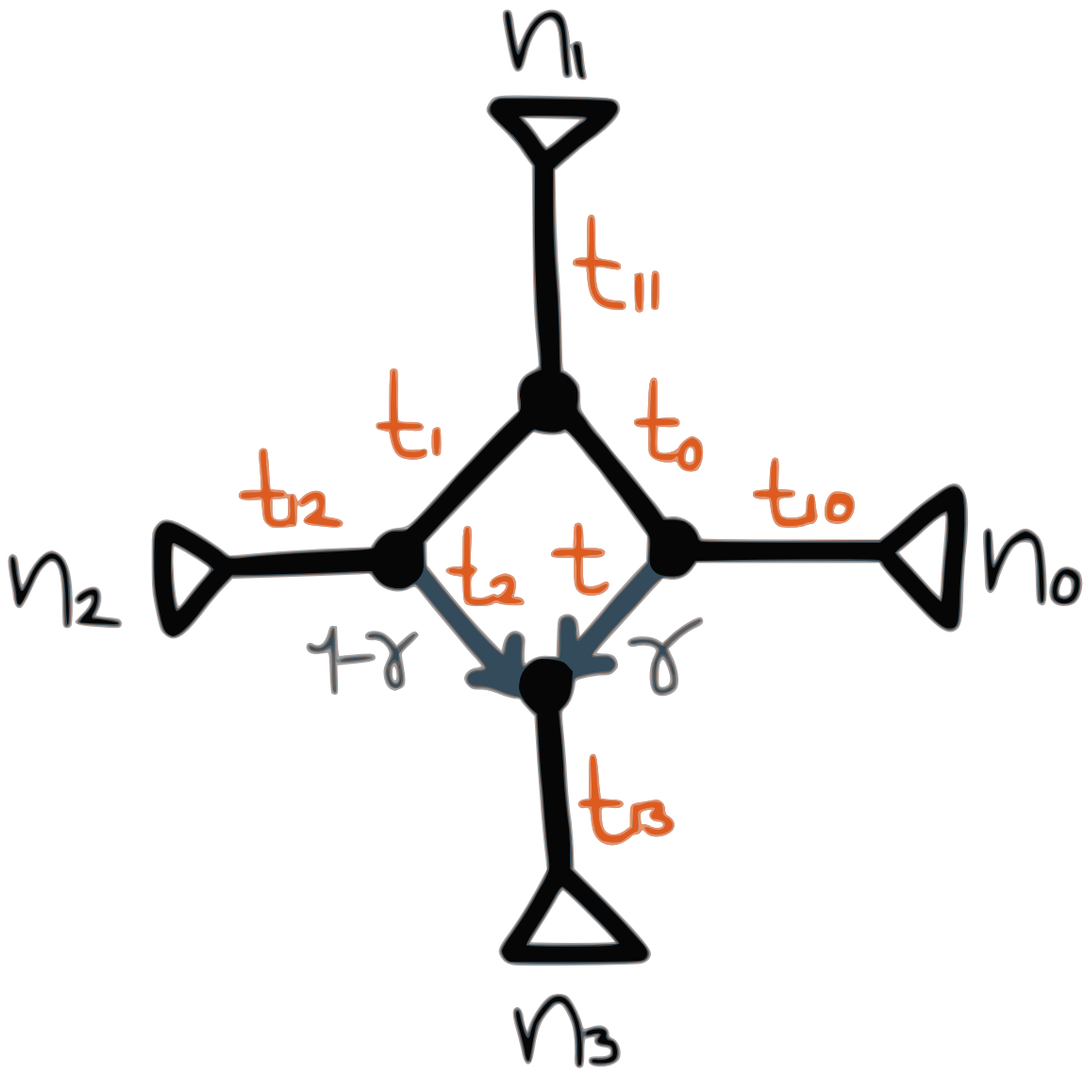
Good diamond
$(n_0, n_2 \geq 2)$

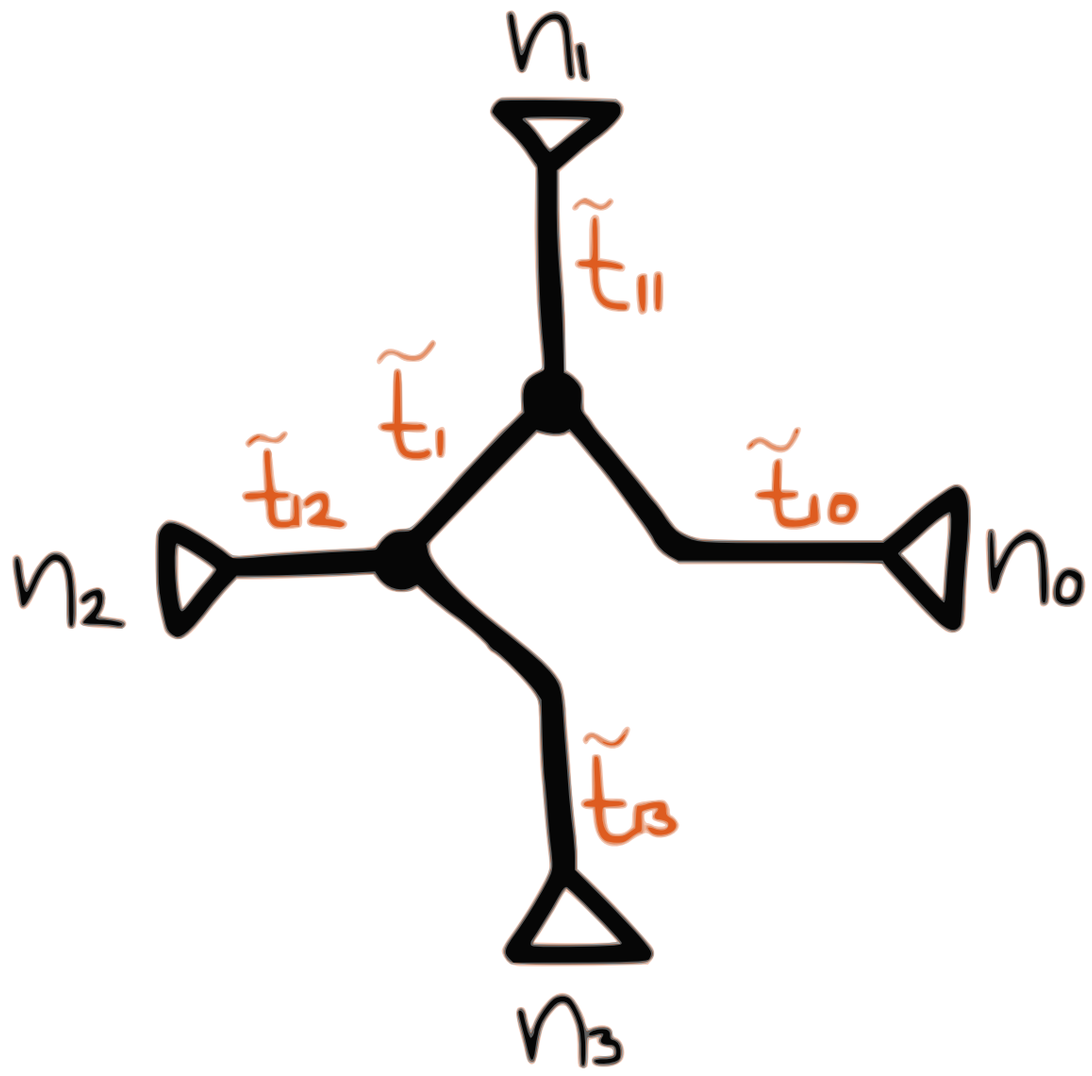


Yes

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

Idea of proof of identifiability: hybridization





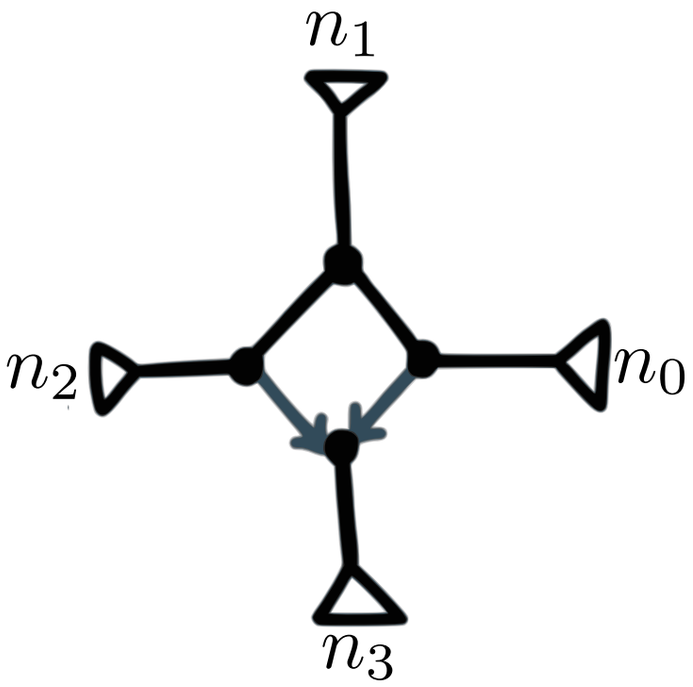
System of equations $\{CF_{network}\}$

(Solís-Lemus & Ané, 2016; Solís-Lemus et al, 2020)

System of equations $\{CF_{tree}\}$

Can we detect the presence of hybridization in level-1 networks?

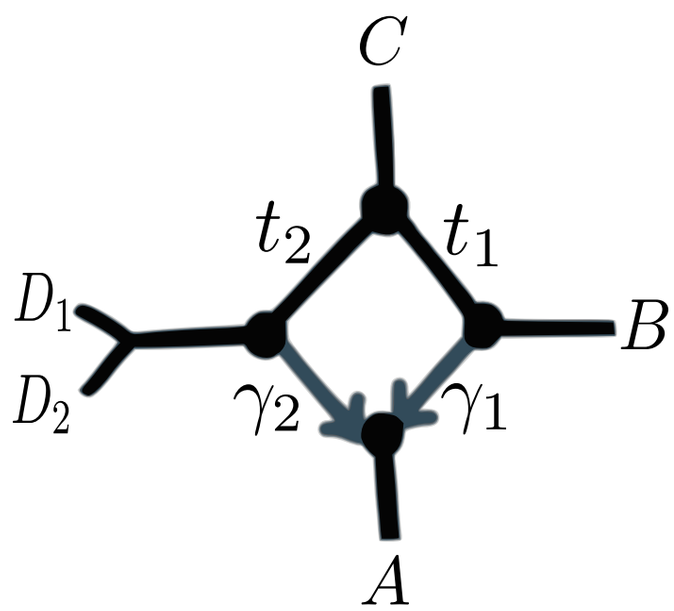
In theory

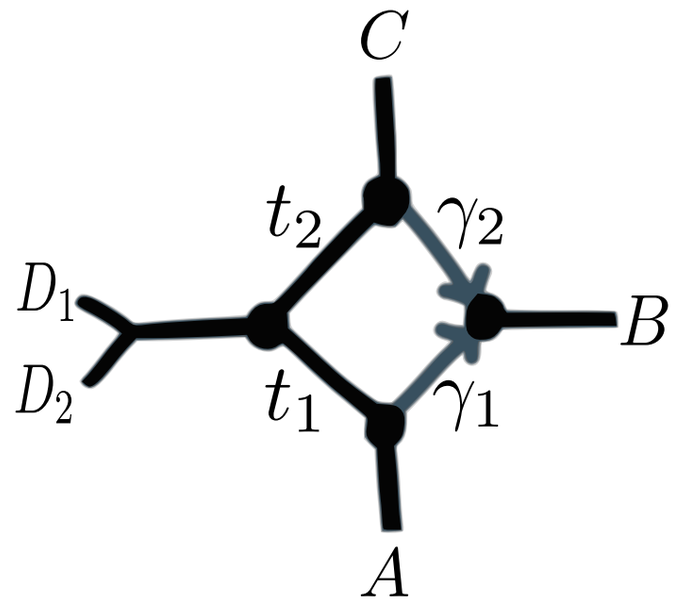


Yes

$(n_i \geq 2)$

In practice

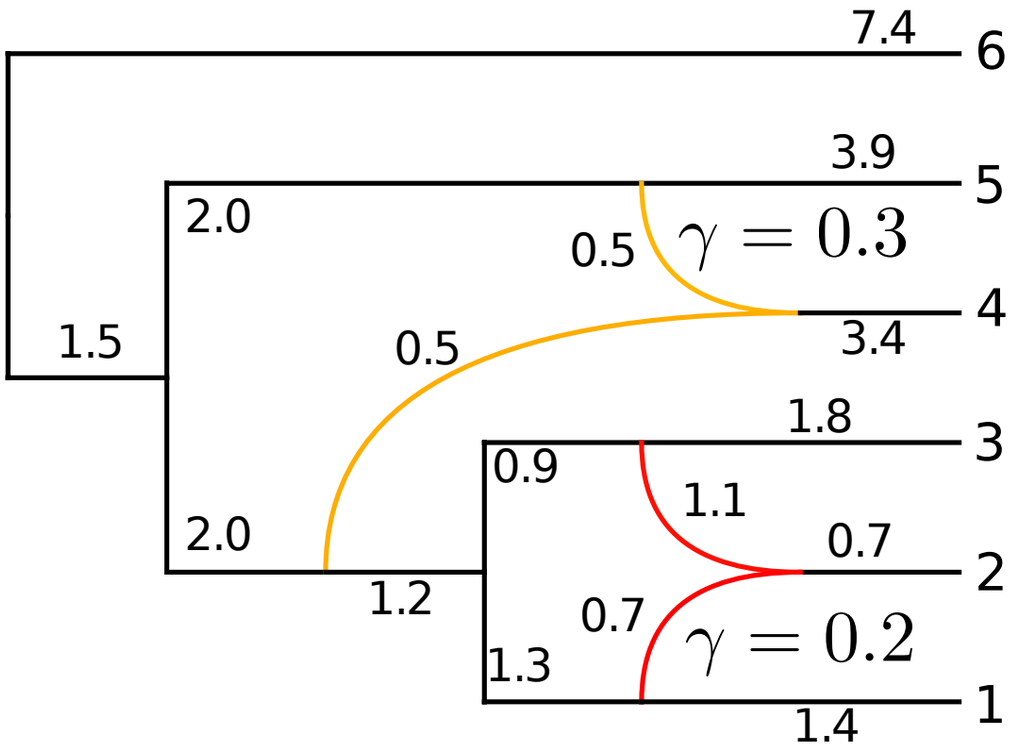




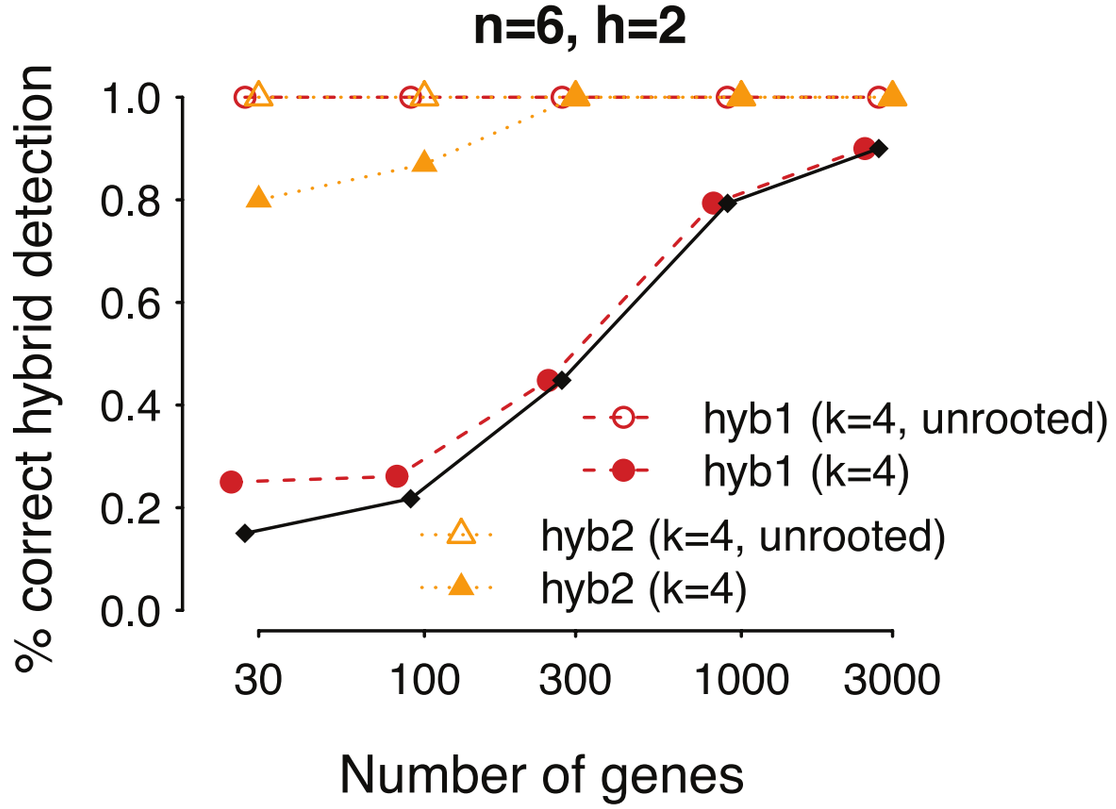
Sometimes

Identifiability matters: SNaQ performance

Good diamond



Bad diamond



(S.-L., Ané, 2016, PLoS Genetics)

Network challenges

- Scalability
- Identifiability
- Network space
- Network comparison

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

Network challenges

- Scalability
- Identifiability
- Network space
- Network comparison

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

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

Network challenges

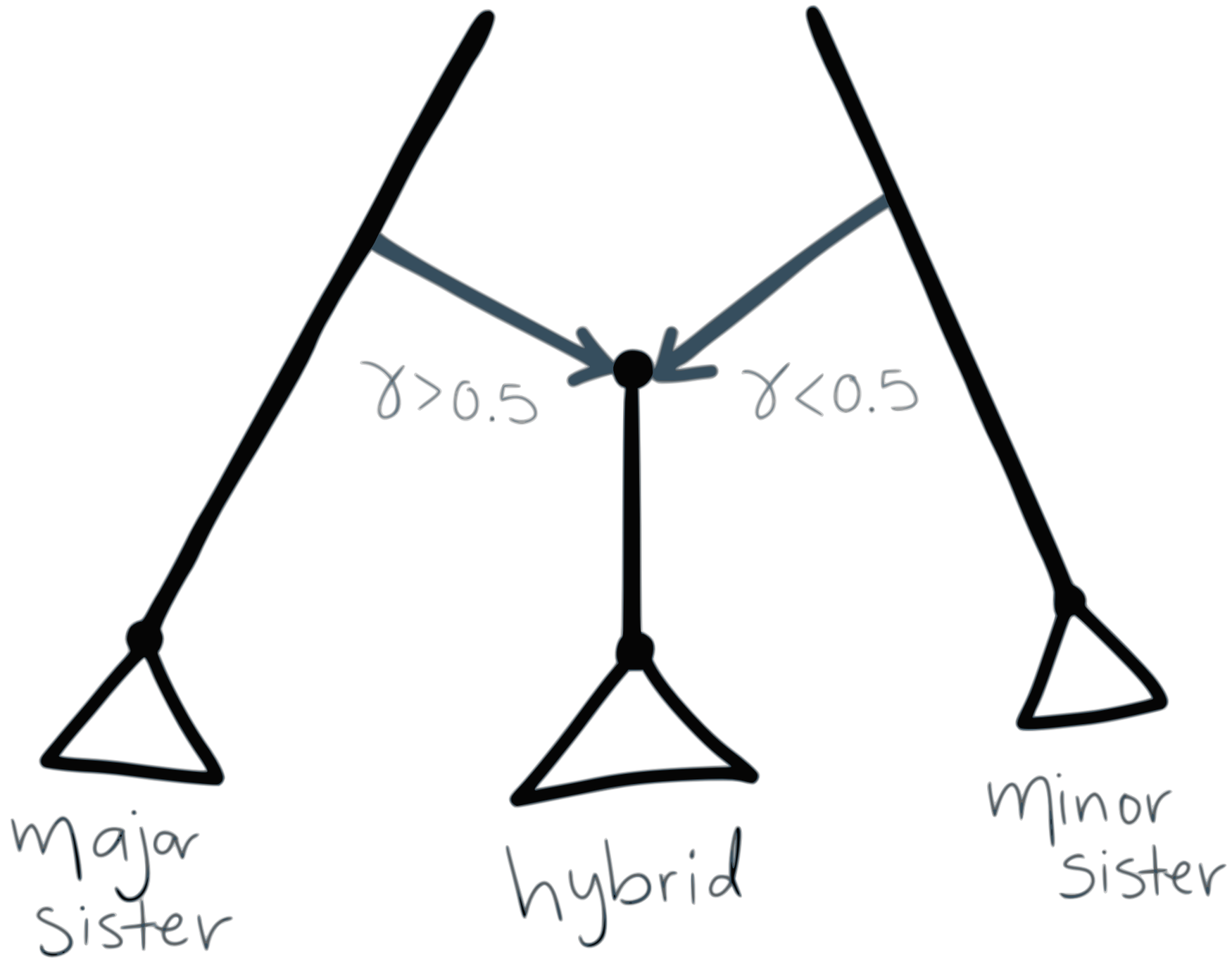
- Scalability
- Identifiability
- Network space
- Network comparison

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

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

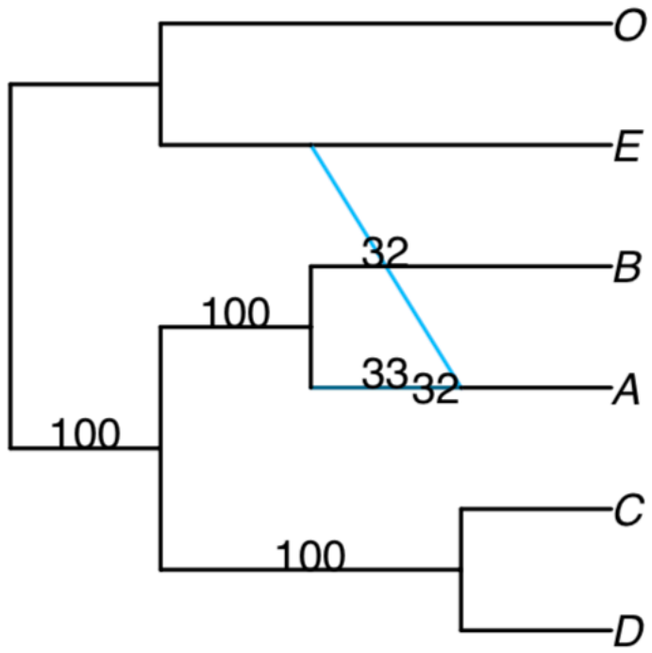
Missing: distance function
Hardwired-cluster distance only for rooted networks
Summary of networks: clades!

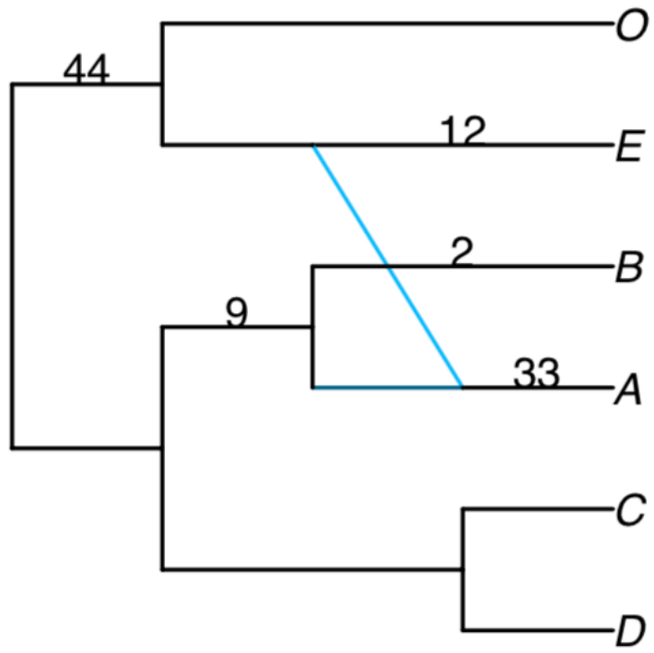
Network summary



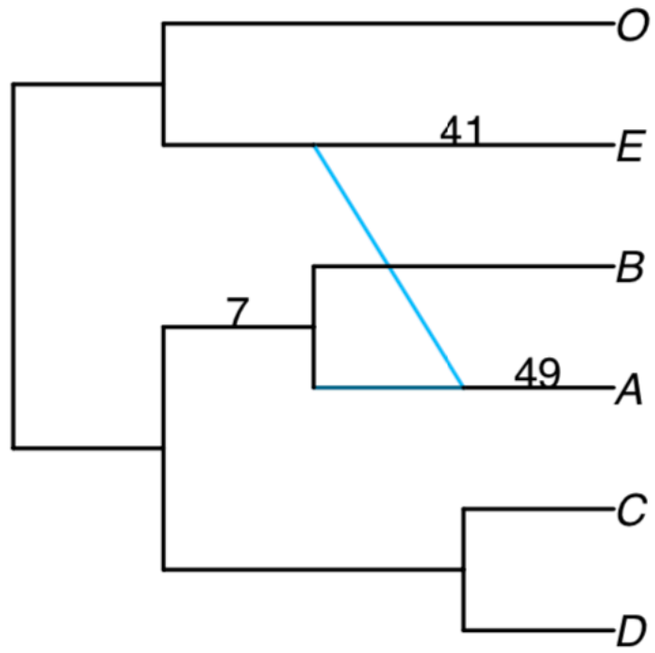
(S.-L. et al, 2017, MBE)

Network summary





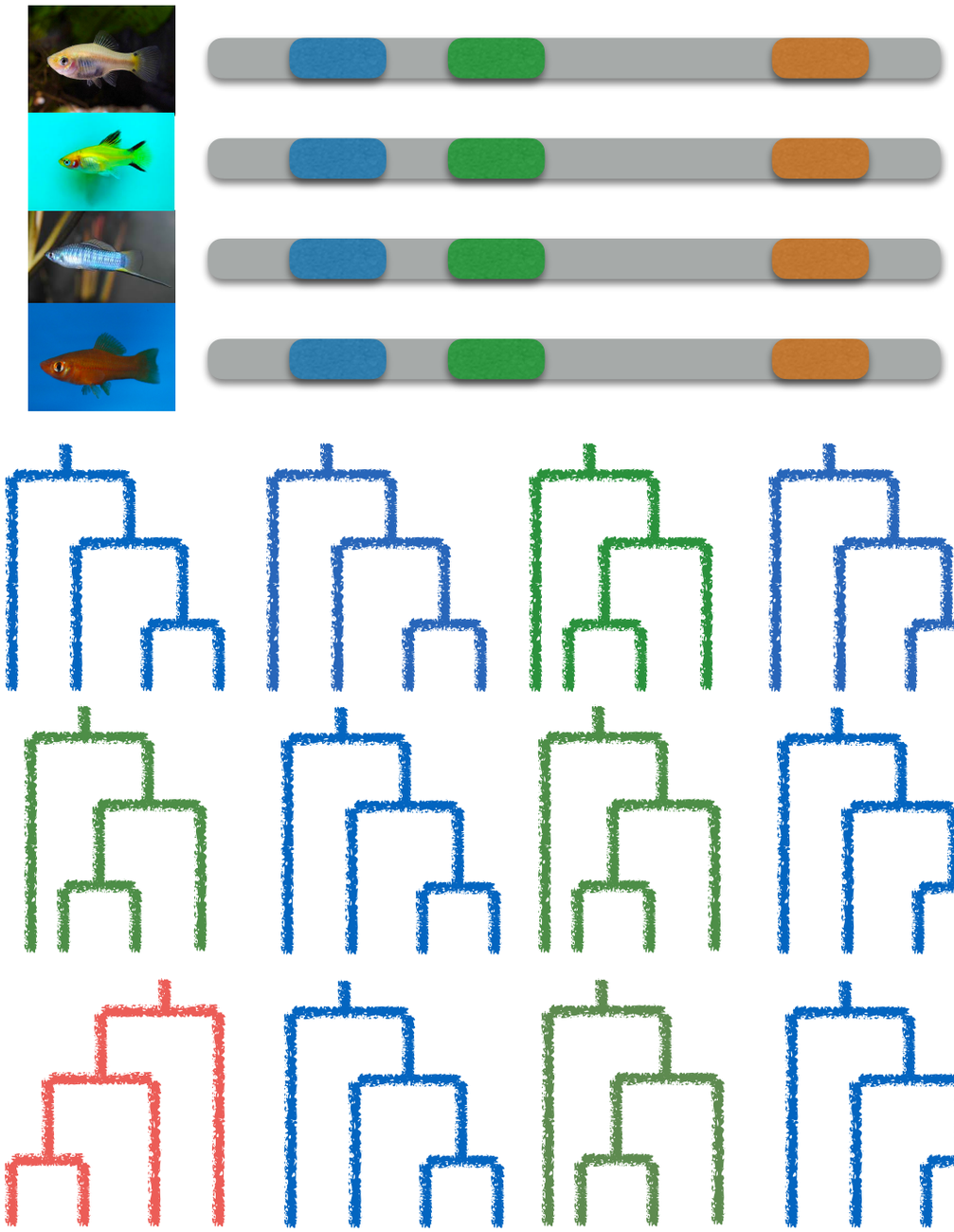
Hybrid clades



Minor sister clades

When?

Phylogenetic network



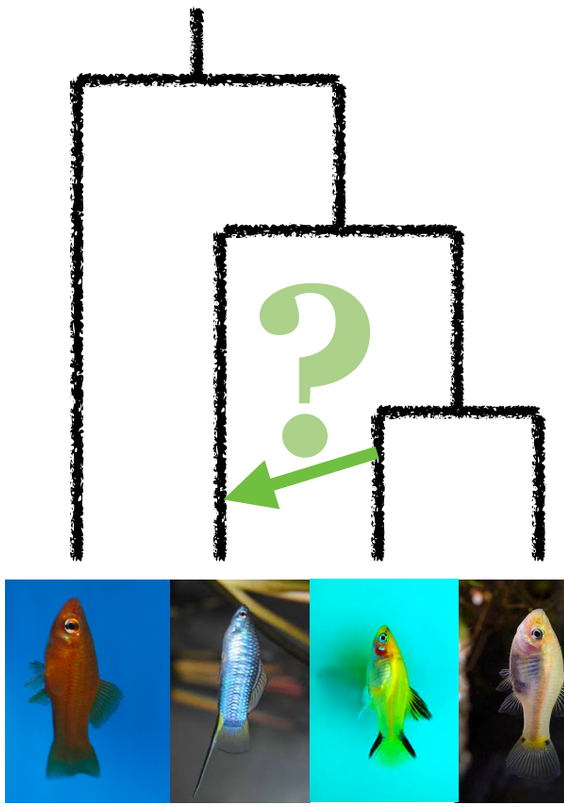
Data

TICR

GitHub

Goodness-of-fit test

Hypothesis test:
Is a tree a good fit?



`https://github.com/nstenz/TICR`

(Stenz et al, 2015, Syst Bio)

Practical advice

- Do multiple runs
- Do bootstrap
- Check the .networks output file (especially if hybridization conflicts with outgroup)
- What is the quality of my input data (gene trees/CFs)?
- Run SNaQ sequentially: h=0, h=1, h=2,…

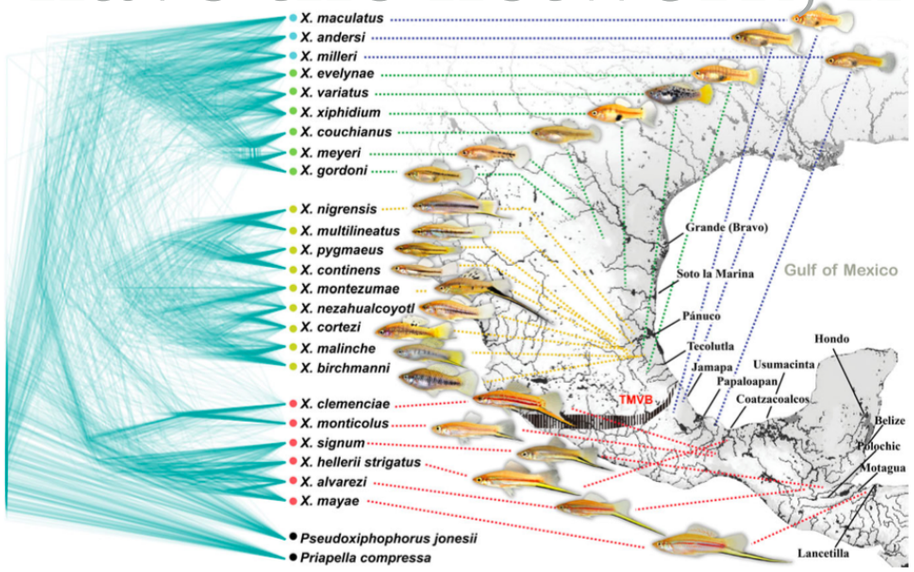
Practical advice

- Do multiple runs
- Do bootstrap
- Check the .networks output file (especially if hybridization conflicts with outgroup)
- What is the quality of my input data (gene trees/CFs)?
- Run SNaQ sequentially: h=0, h=1, h=2,…

When to stop?
(Cai and Ané, 2020)

Part II

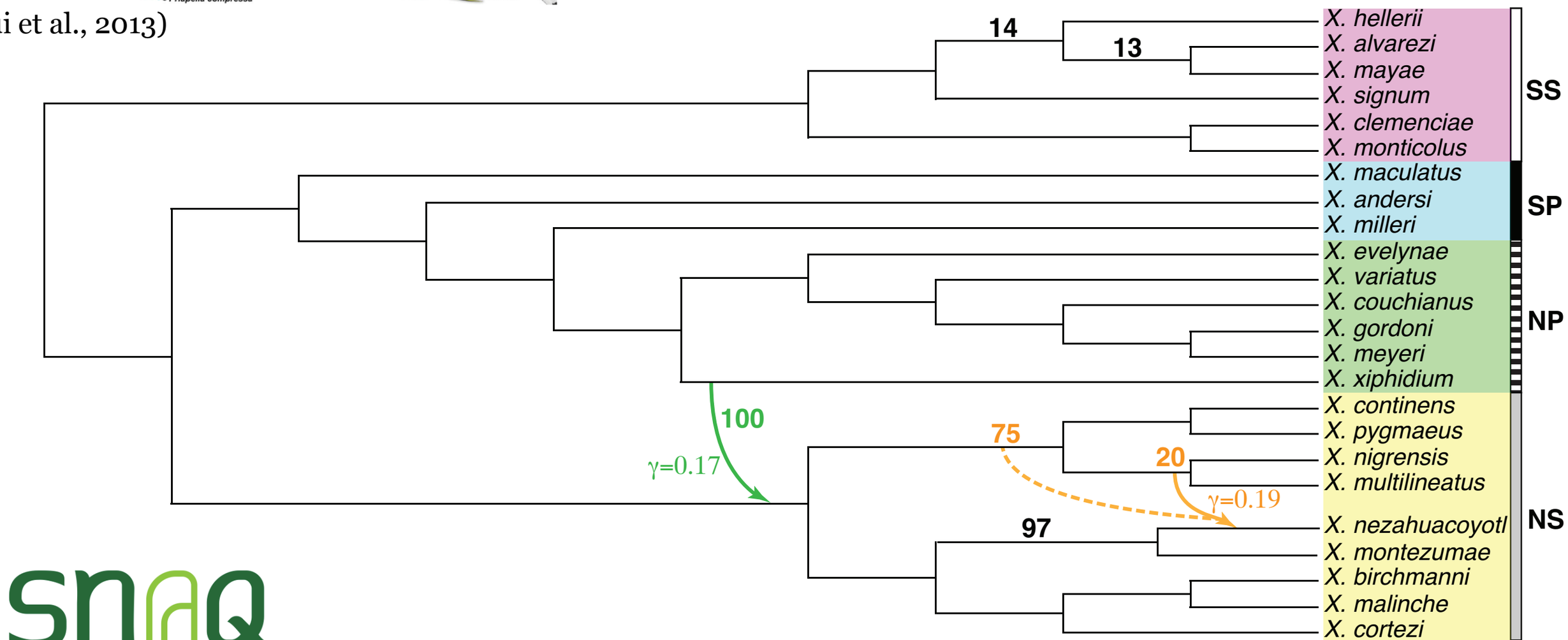
I have the network, now what?



(Cui et al., 2013)

1183 genes,
24 swordtails
and platyfish

Xiphophorus fish data

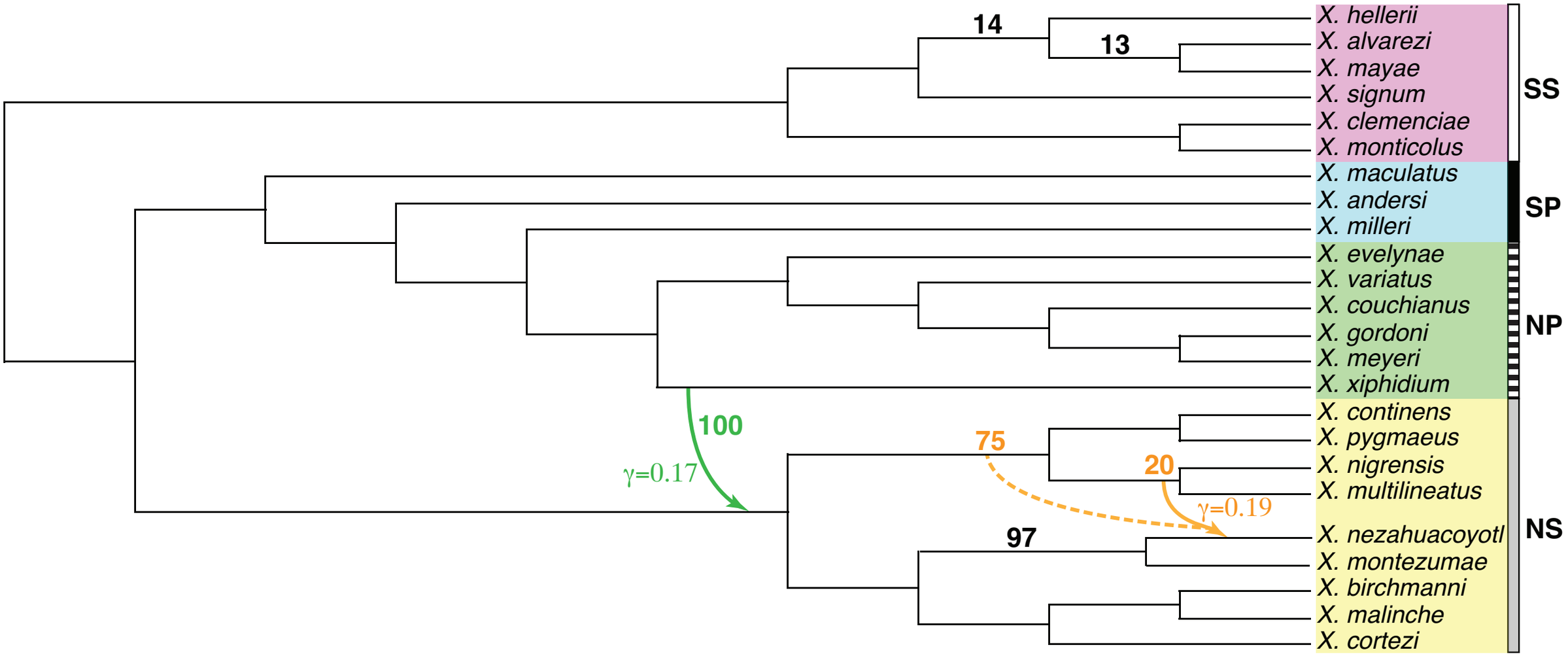


SNAQ

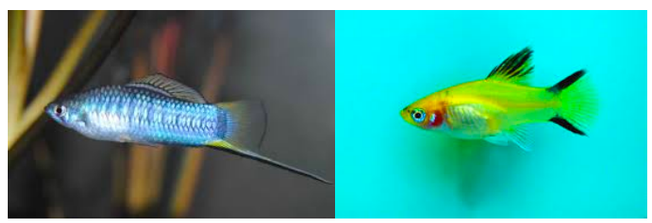
(Solís-Lemus, Ané, 2016, PLoS Genetics)

Part II

I have the network, now what?



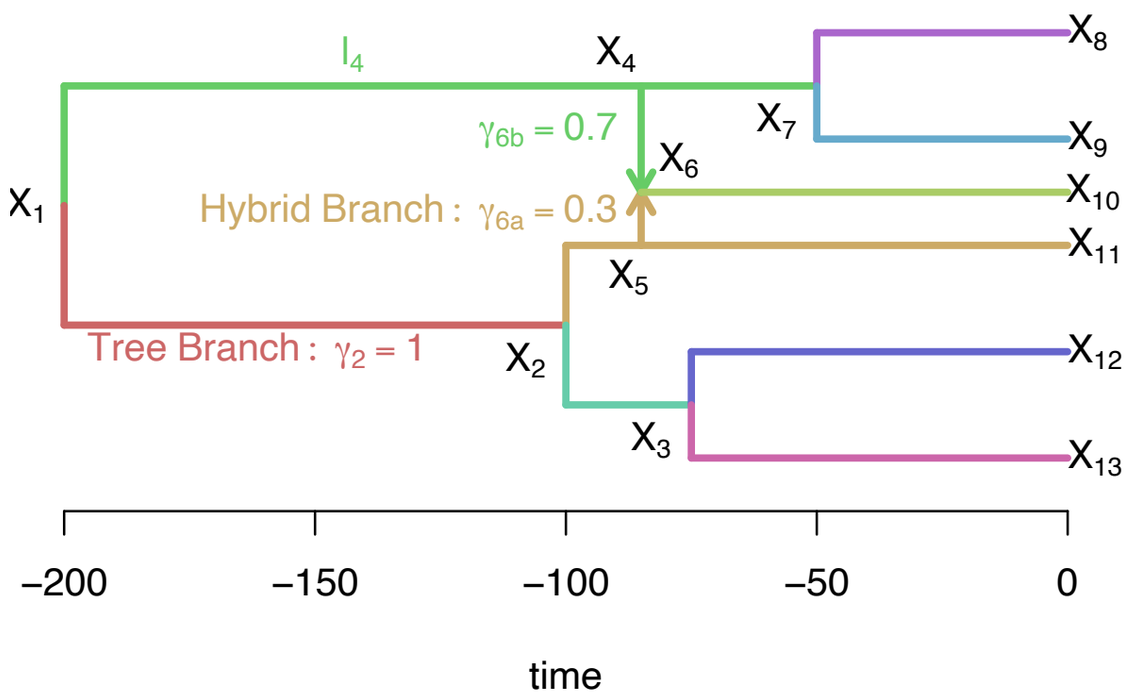
- Sword index
- Female preference

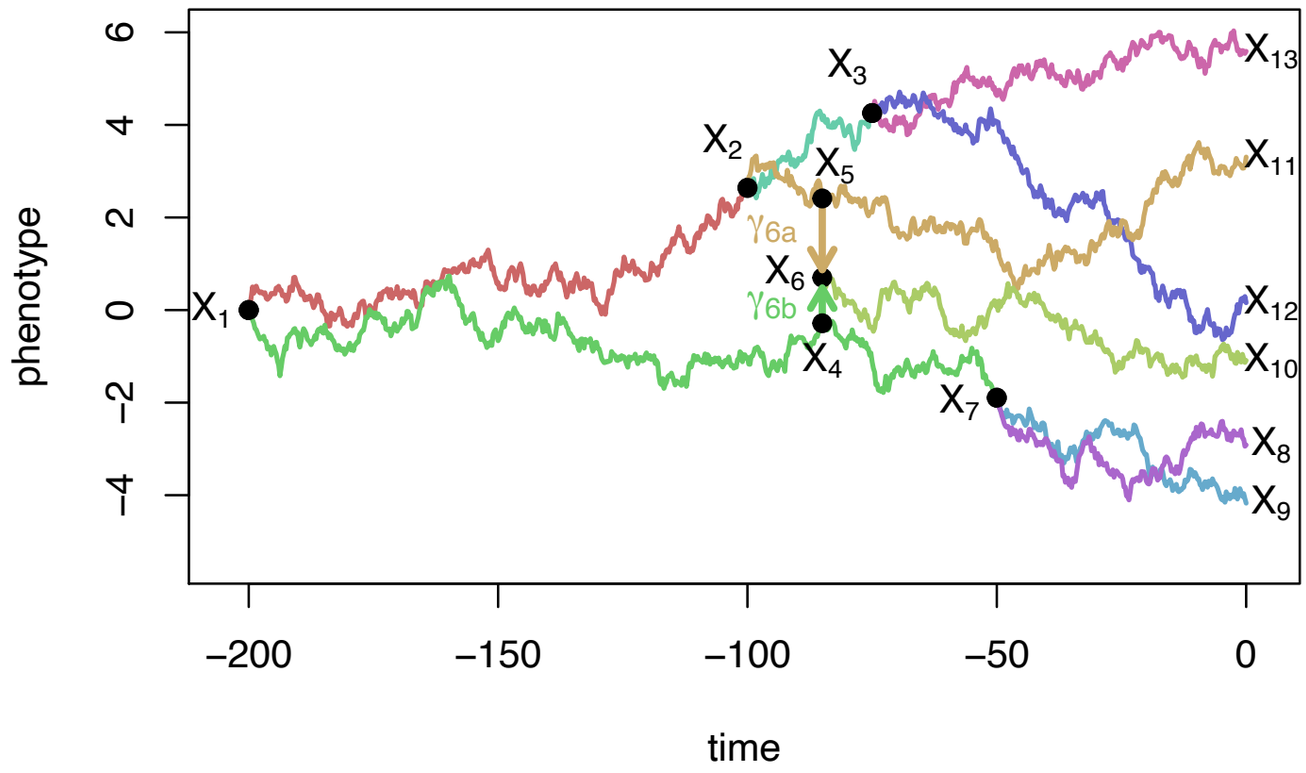


(Cui et al., 2013)

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

Trait models of evolution in networks





Brownian Motion + weighted average in hybrid

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

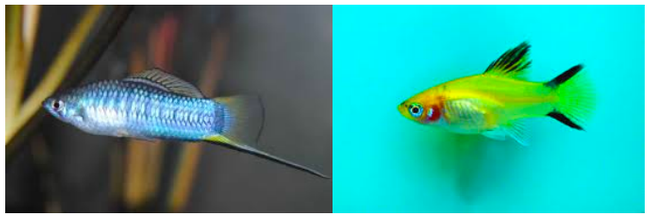
(Bastide et al, 2018, Syst Bio)

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

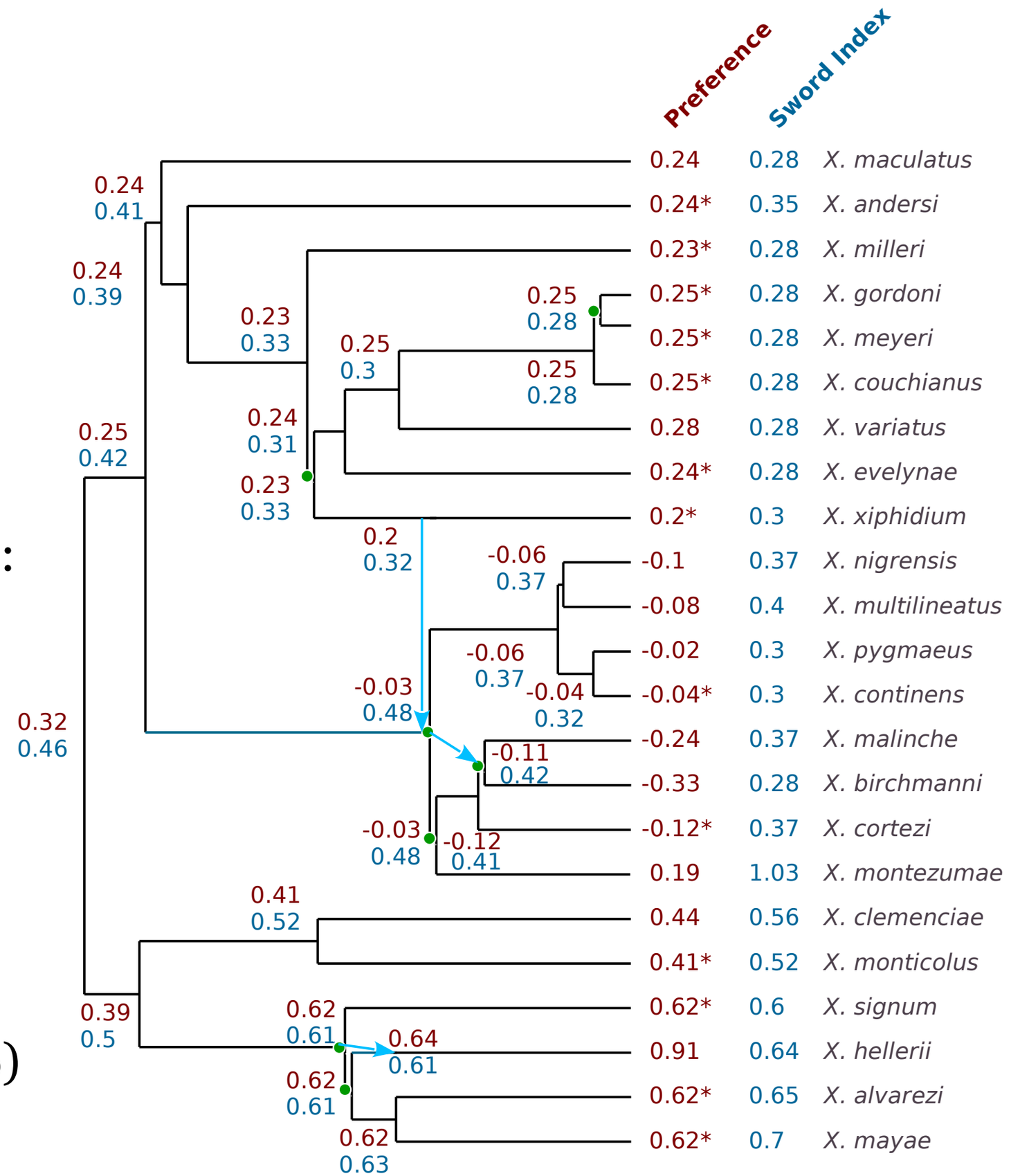
- Phylogenetic signal
- Ancestral reconstruction
- Phylogenetic regression
- Phylogenetic ANOVA

`www.github.com/CRSL4/PhyloNetworks`

- Sword index
- Female preference

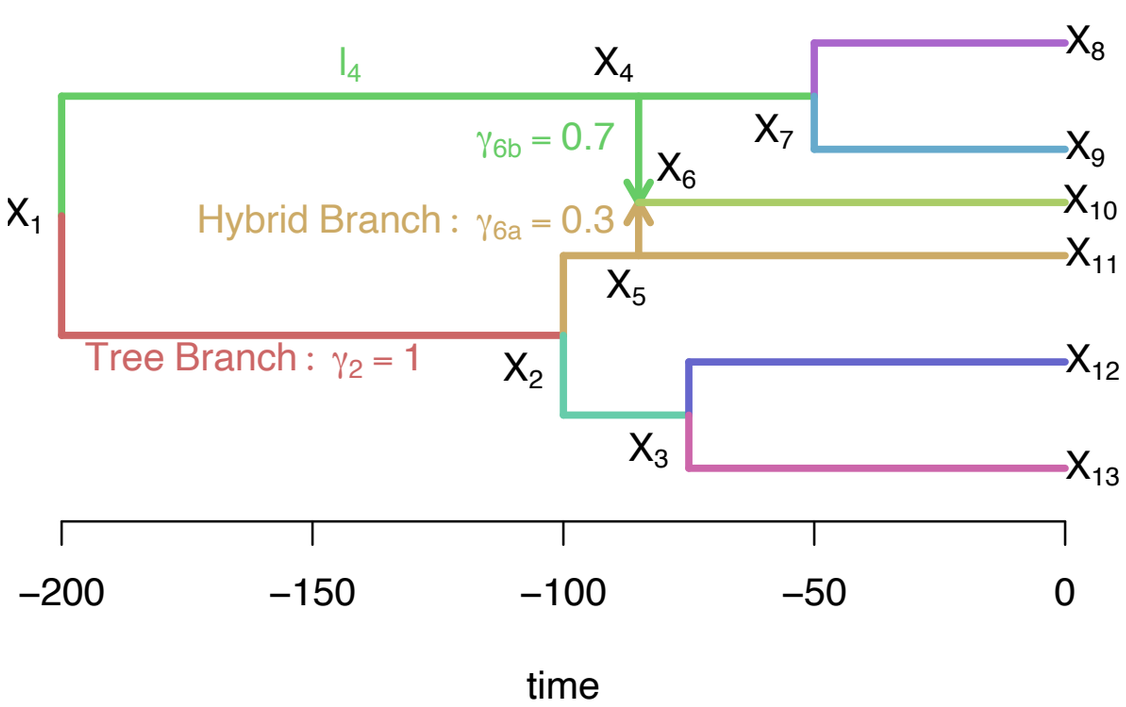


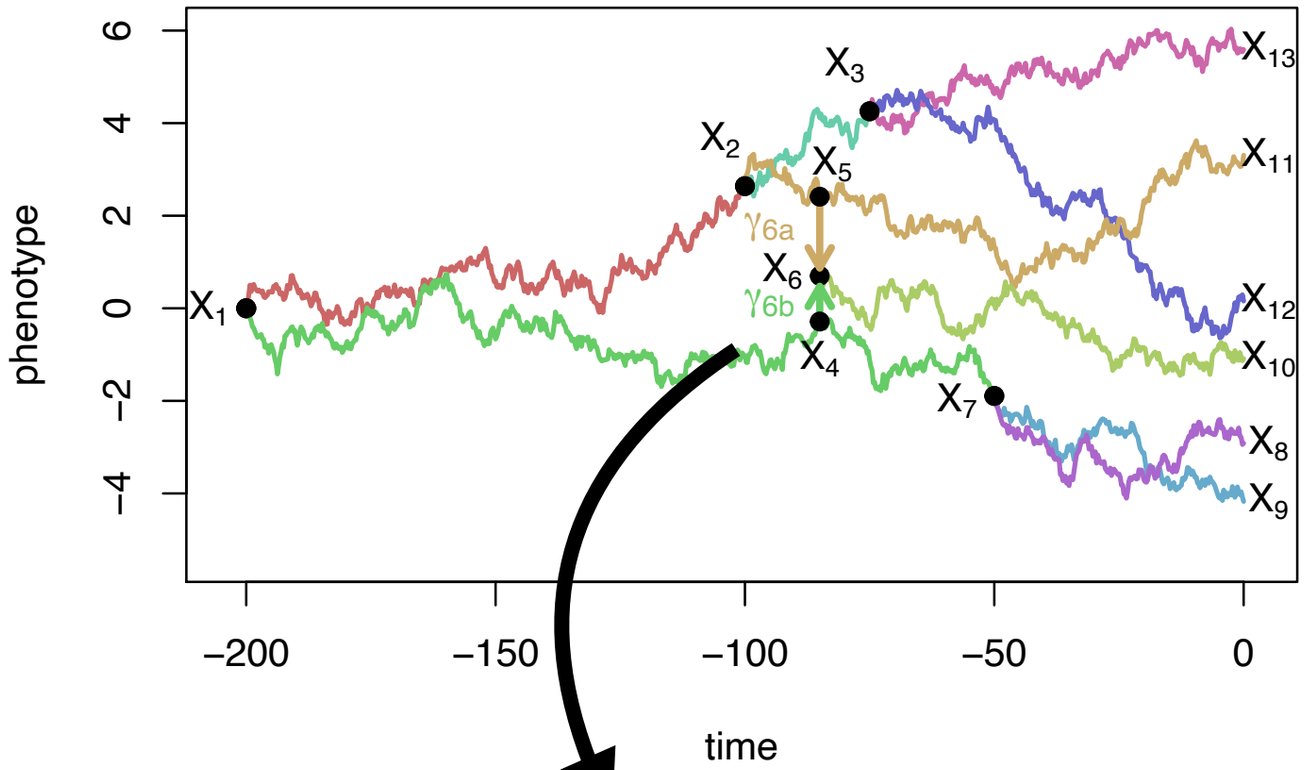
- **Ancestral reconstruction:** common ancestor likely had sword
- **Phylogenetic regression:** positive association between sword index and female preference
  but not significant (p = 0.106)



Test for transgressive evolution

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





$\Delta_h = 0$ No transgressive evolution

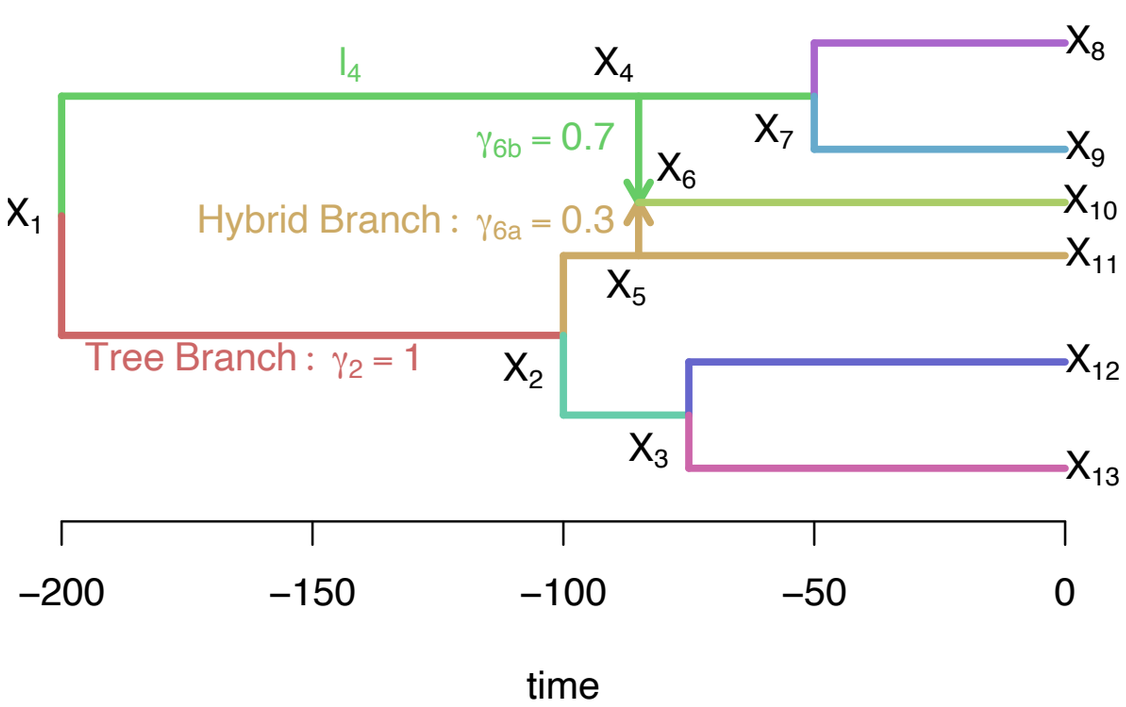
$\Delta_h = \Delta$ Single-effect transgressive evolution

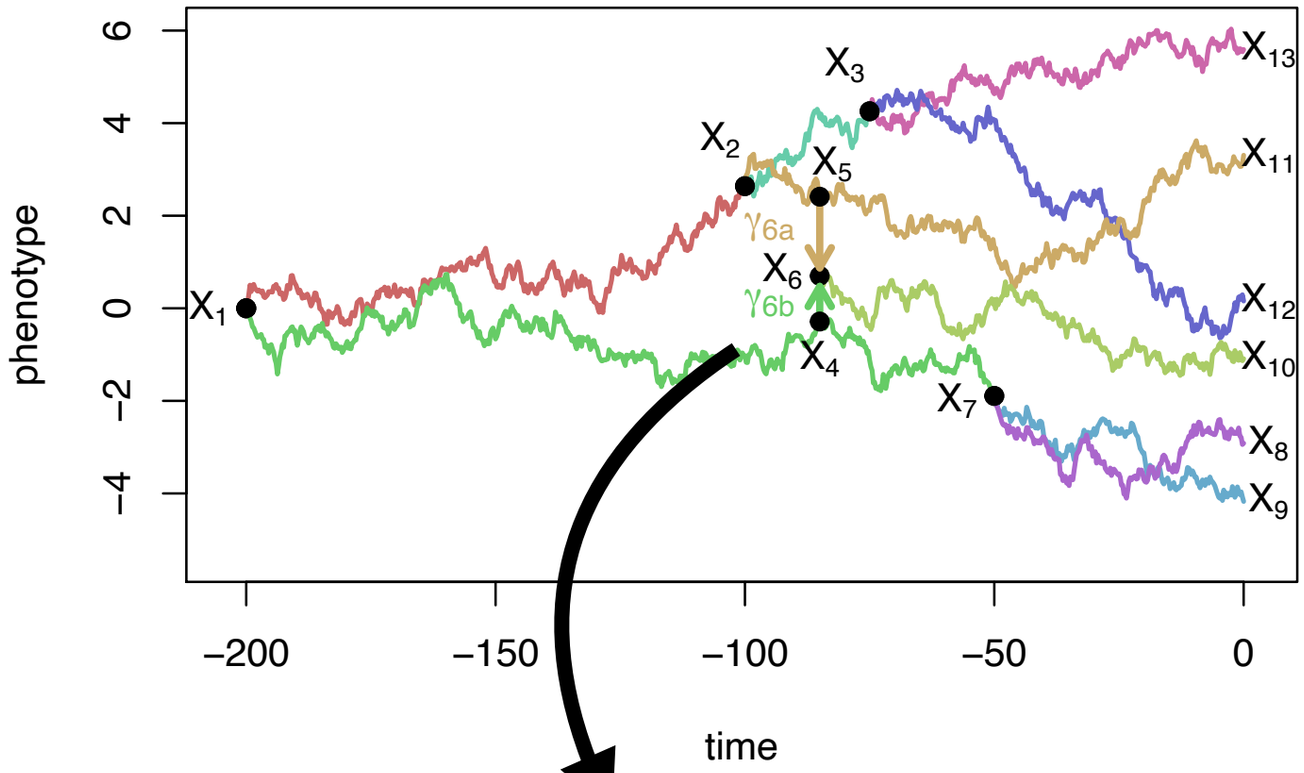
Δ_h Multi-effect transgressive evolution

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

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

(Solis-Lemus & Ane, 2016; Solis-Lemus. et al, 2017)


https://solislemuslab.github.io/


@solislemuslab


crsl4


@thestatistician