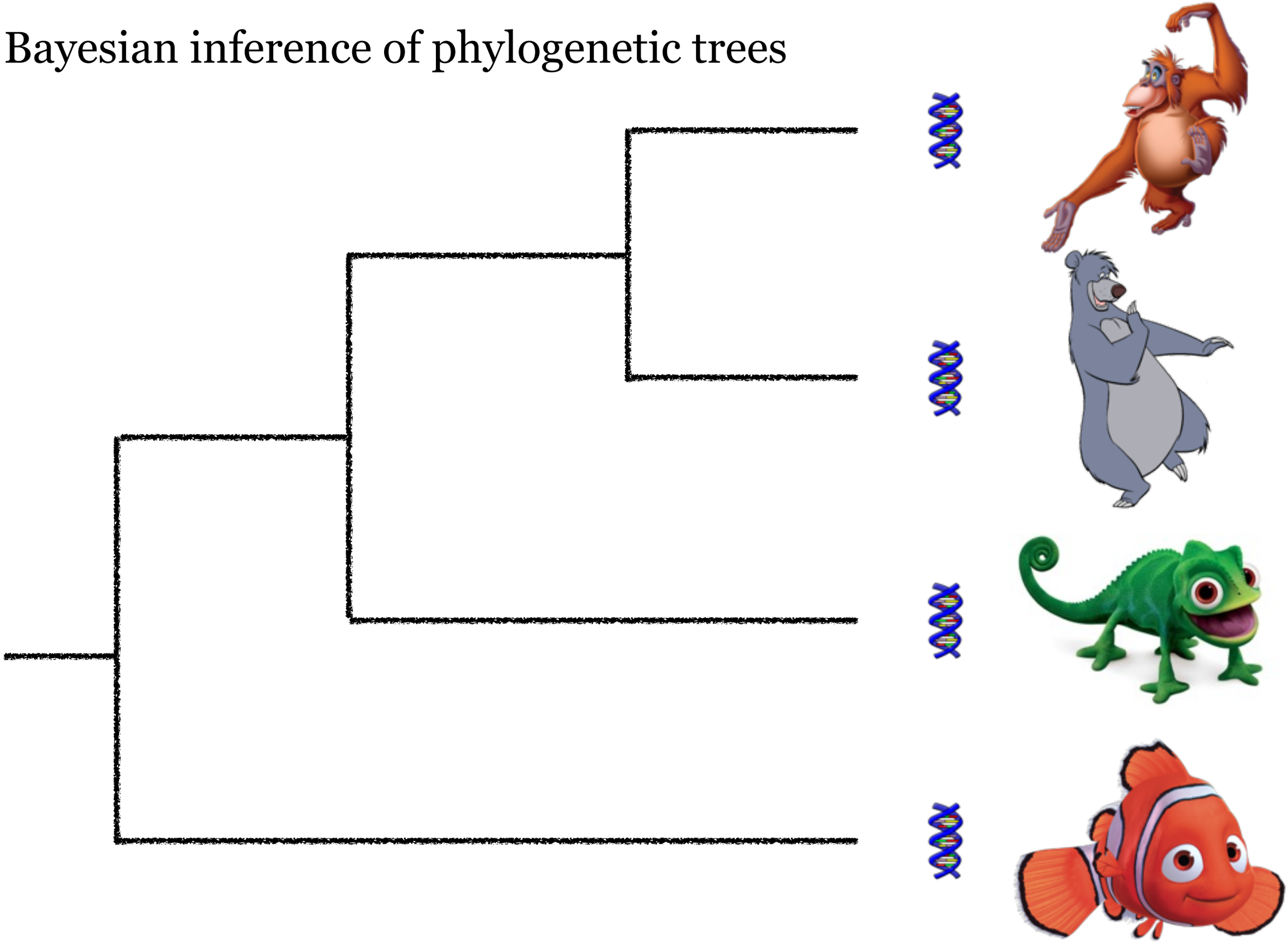
Likelihood challenges for big trees and networks

Claudia Solís-Lemus
University of Wisconsin-Madison

Joint work with Cécile Ané, Bret Larget

Mathematical Approaches to Evolutionary Trees and Networks
Banff International Research Station
February 13, 2017

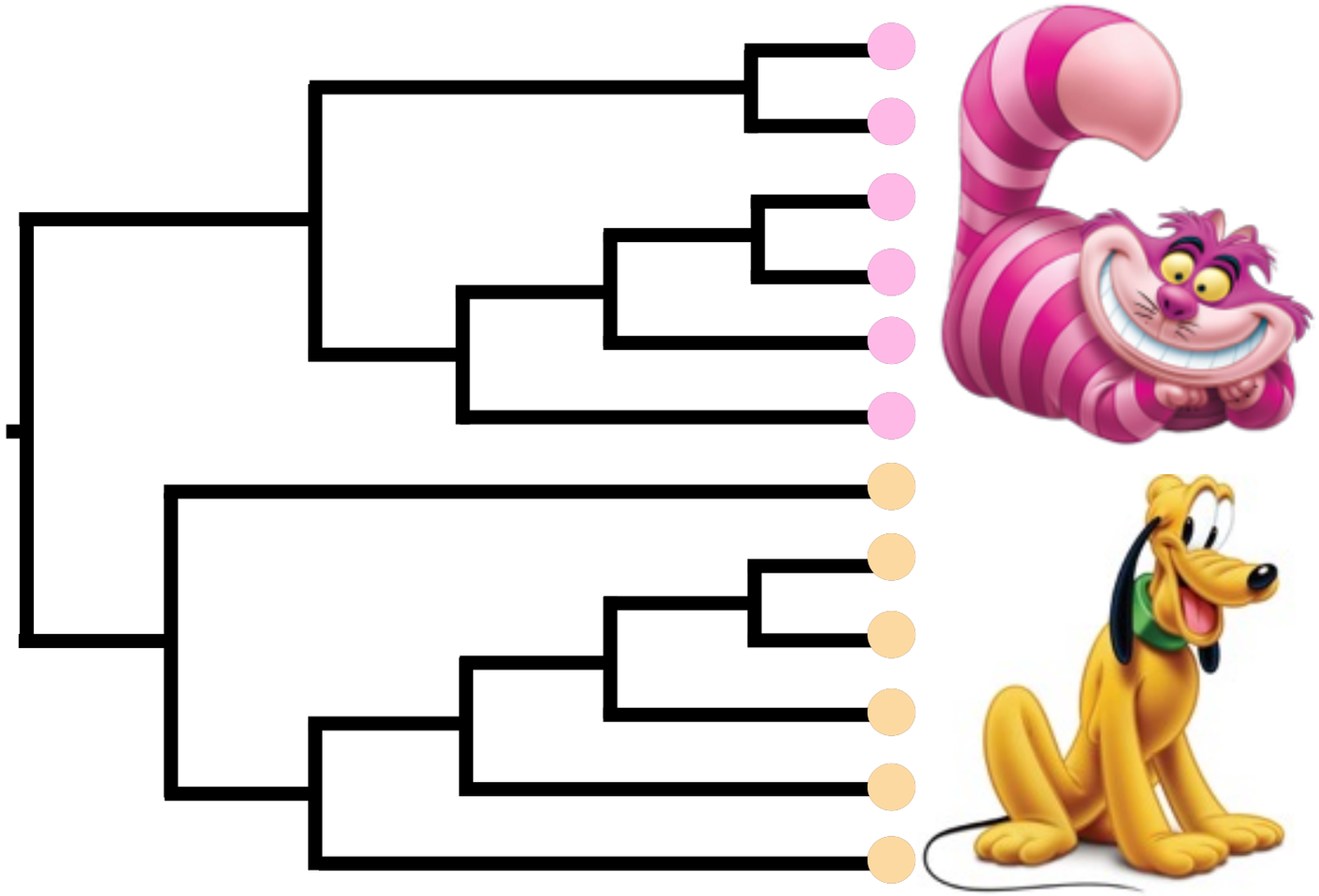
Bayesian inference of phylogenetic trees



Tree inference

Maximum likelihood	**Bayesian inference**
Heuristic search	MCMC
RAxML (Stamatakis, 2014) PhyML (Guindon et al, 2010)	MrBayes (Huelsenbeck, Ronquist, 2001)

# Species	# Unrooted trees	# Rooted trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
11	34,459,425	654,729,075
12	654,729,075	13,749,310,575
13	13,749,310,575	316,234,143,225
$\vdots$	$\vdots$	$\vdots$
52	$>$ # atoms in universe	



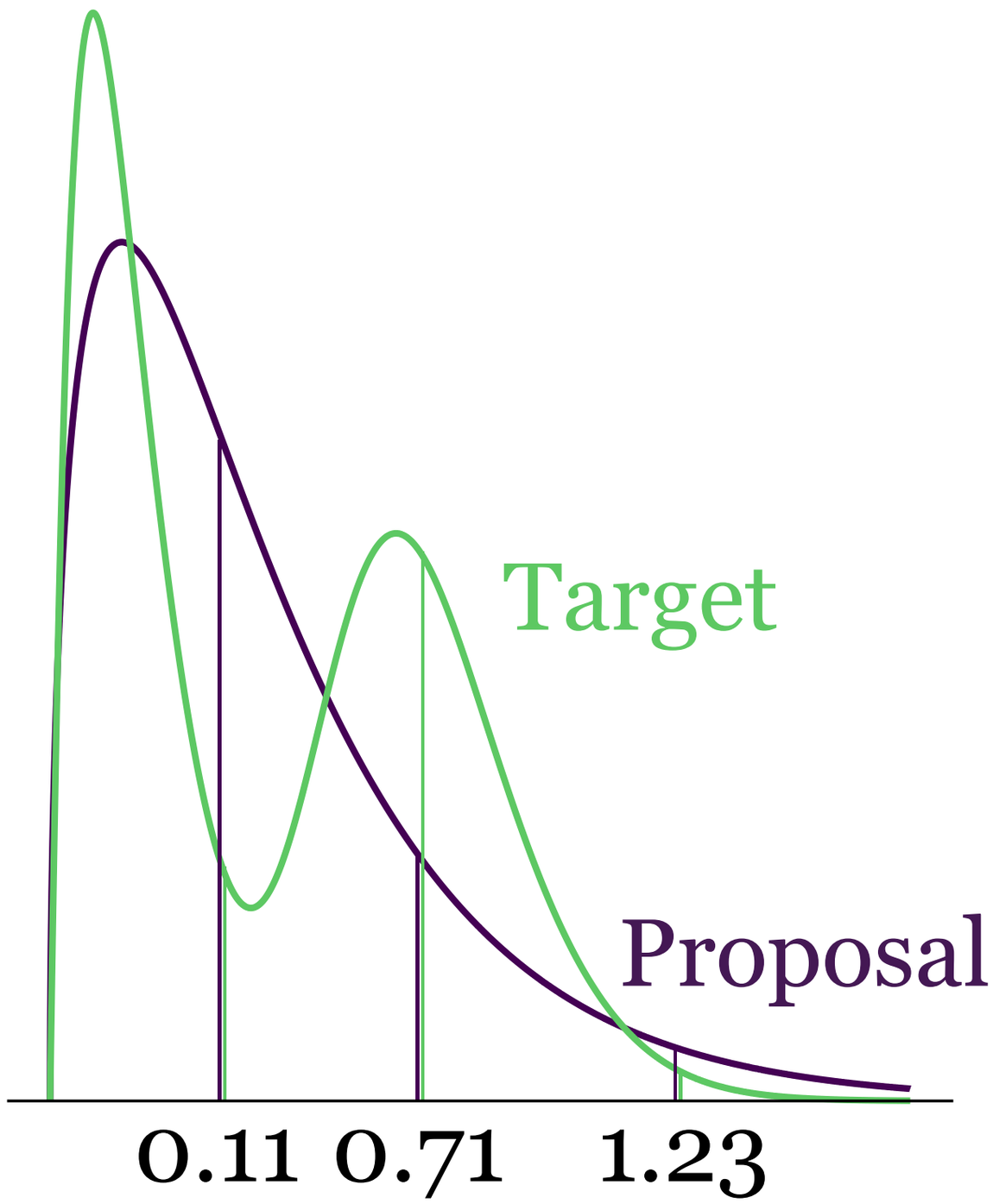
12 taxa *Carnivora*

MCMC efficiency ~0.025%

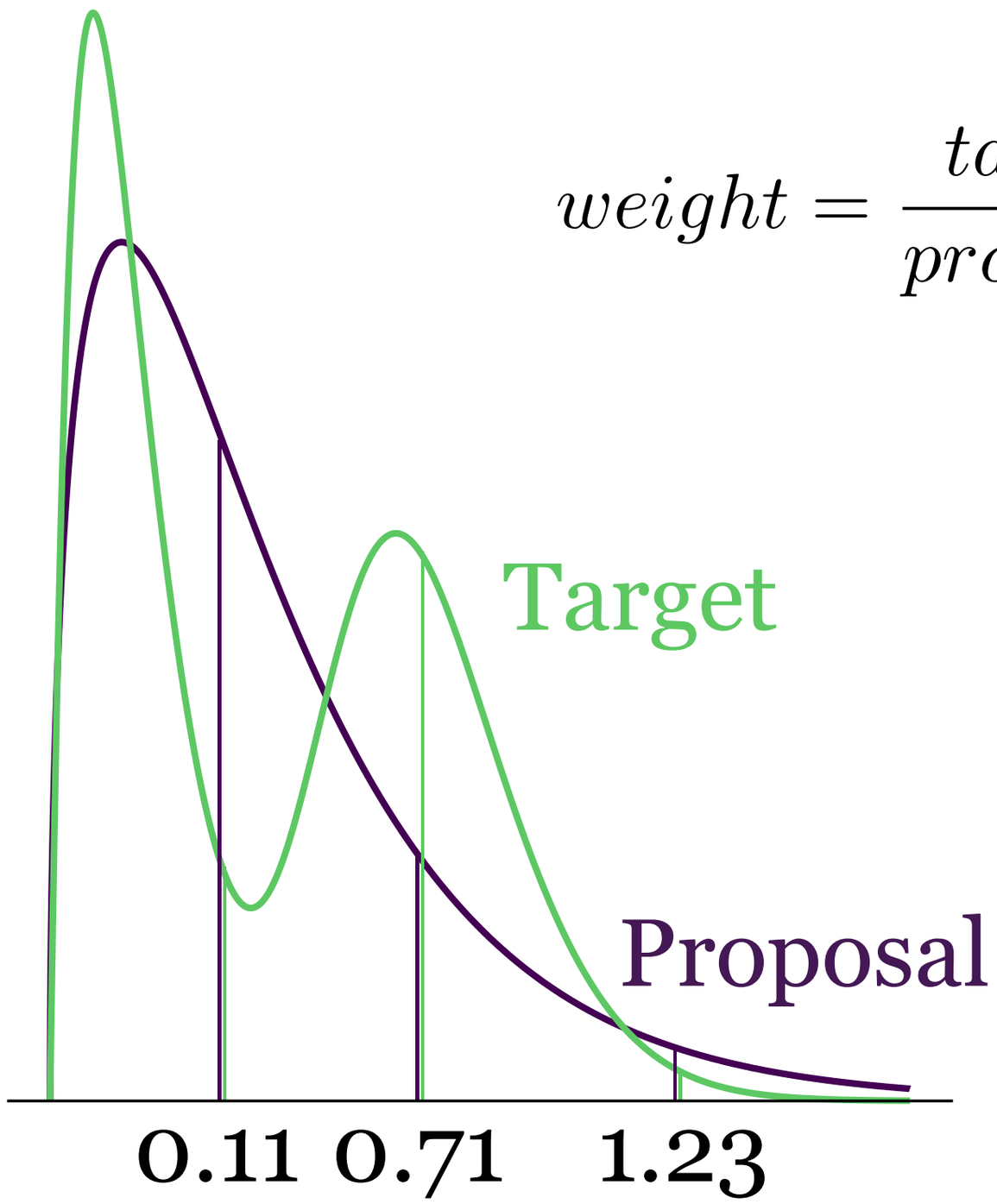
(250 from 1 million post-burnin generations)

What if we could sample from the posterior more efficiently?

Importance sampling

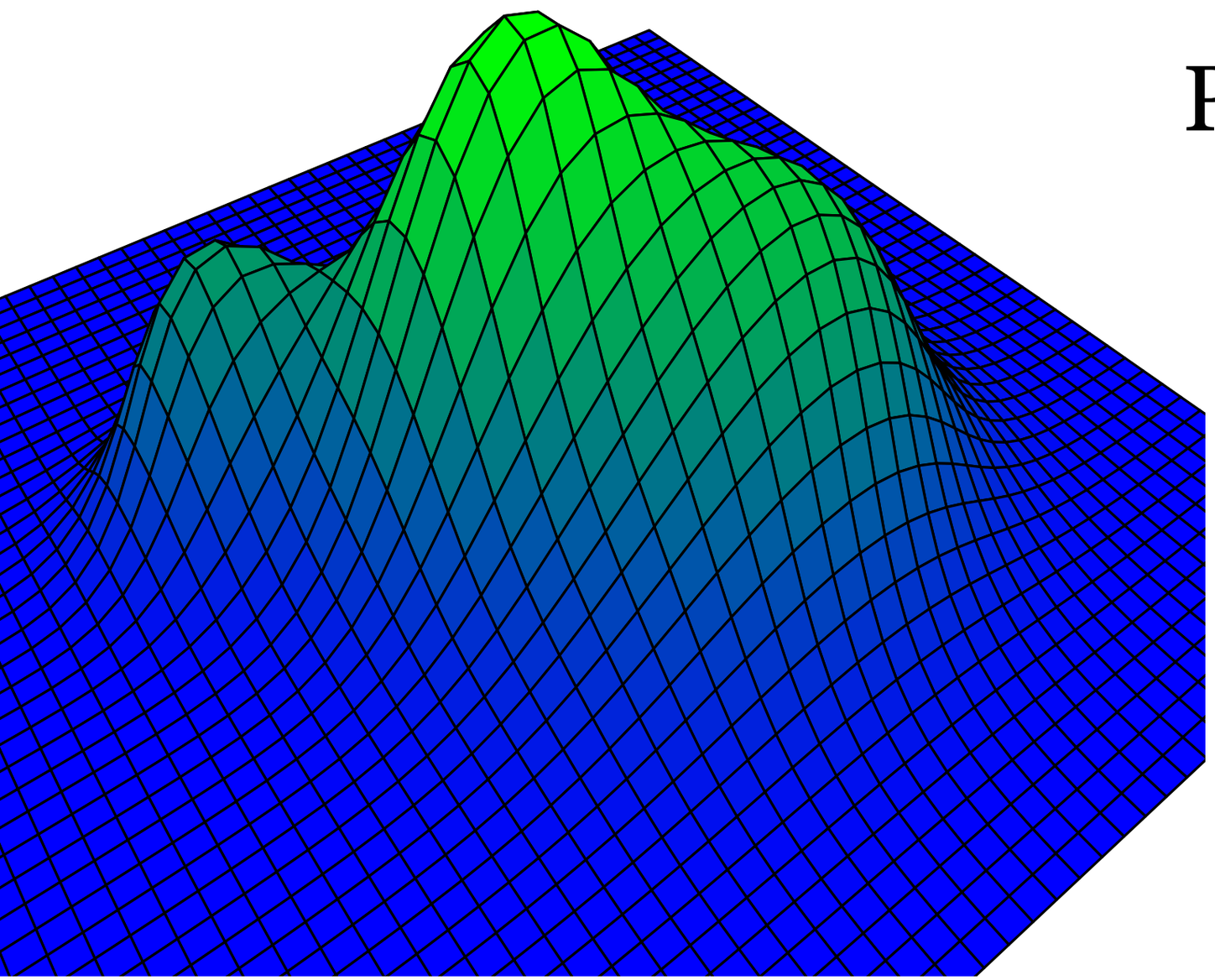


Importance sampling



$$weight = \frac{target}{proposal}$$

Importance sampling in phylogenetics

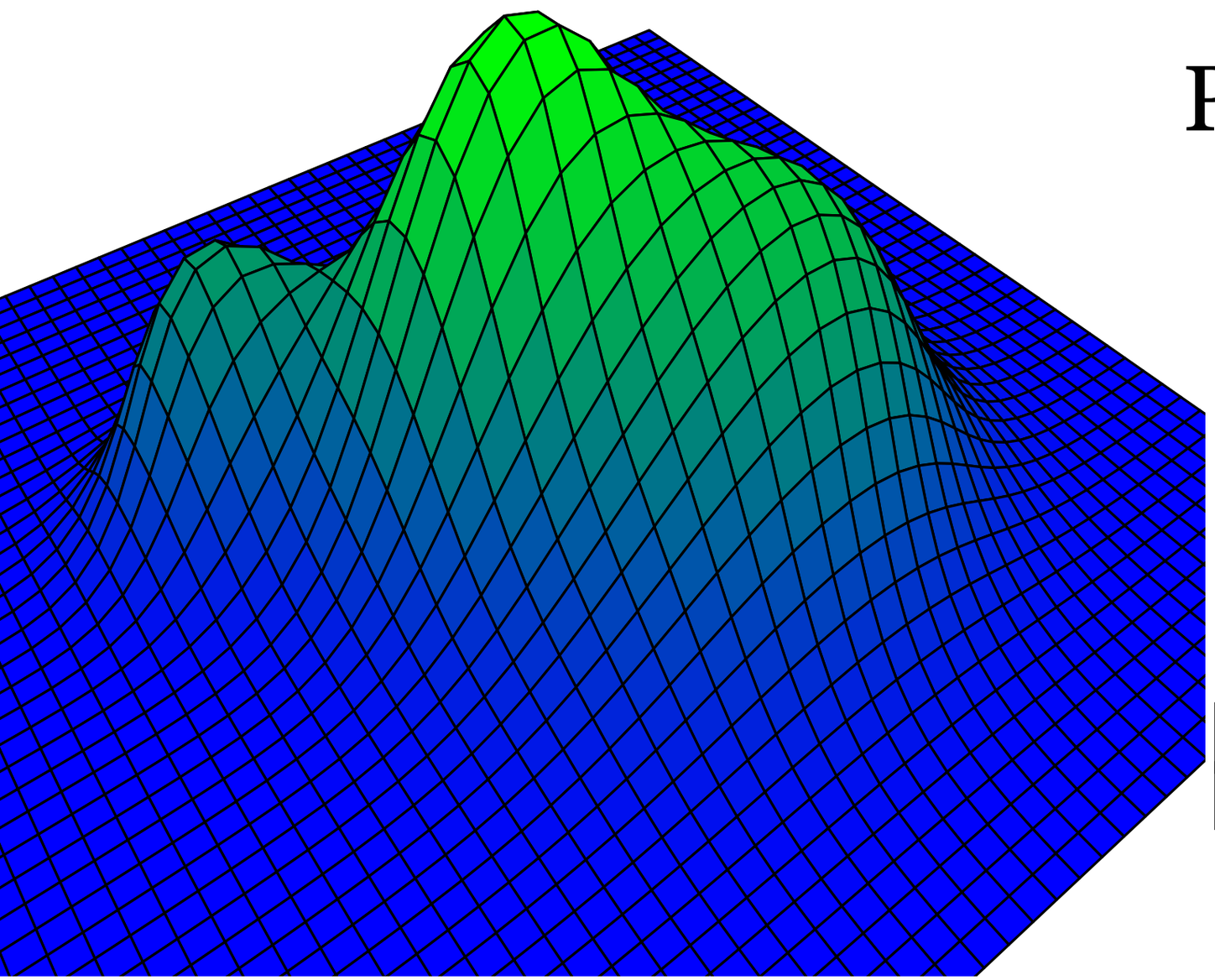


Target:
Posterior distribution

$$(T, t, Q)$$

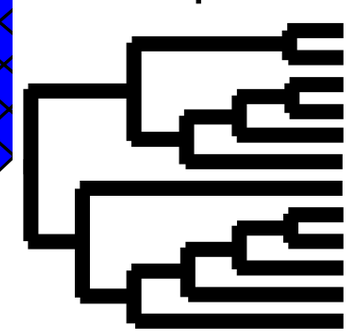
Work in progress with B. Larget

Importance sampling in phylogenetics



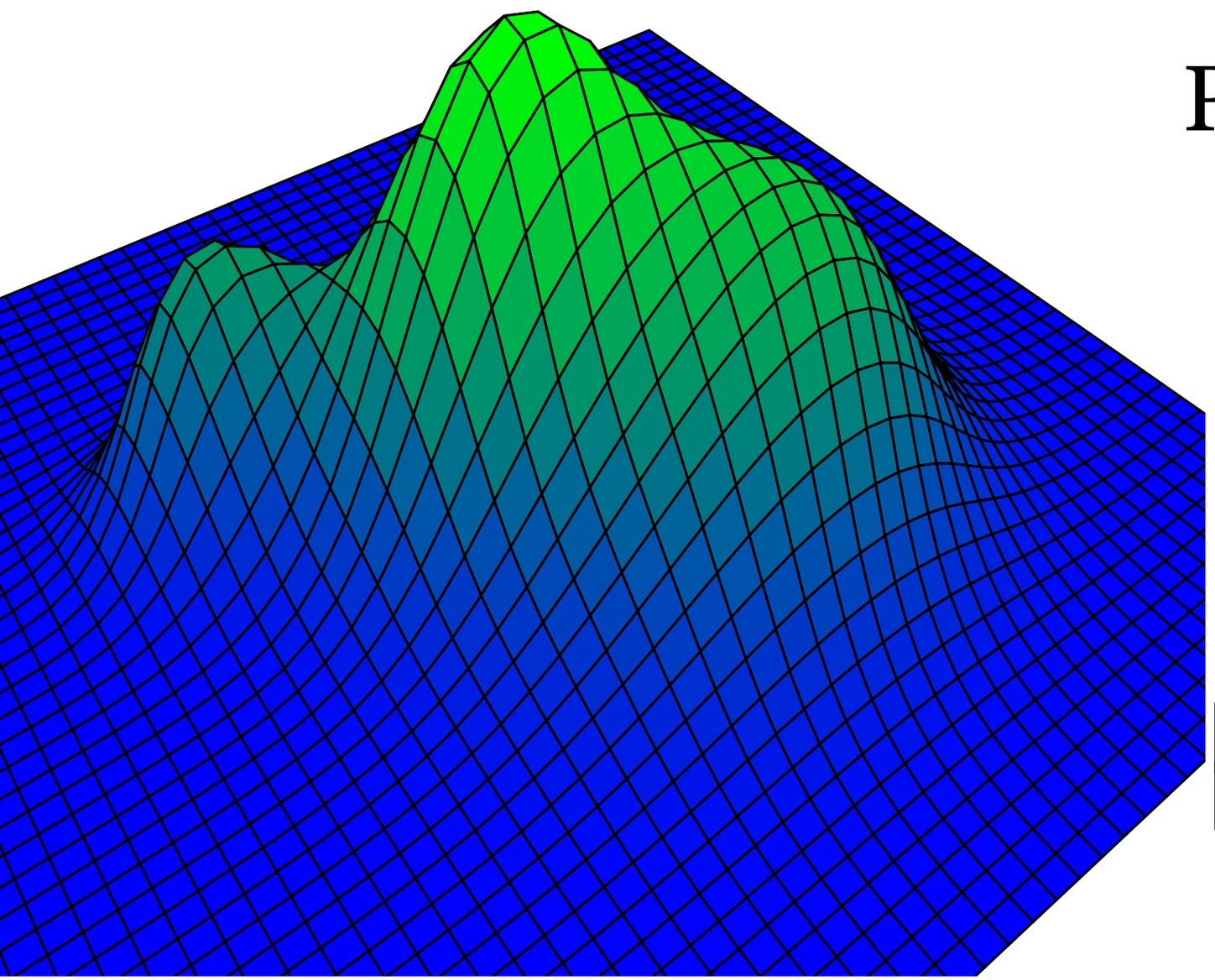
Target:
Posterior distribution

$$(T, t, Q)$$



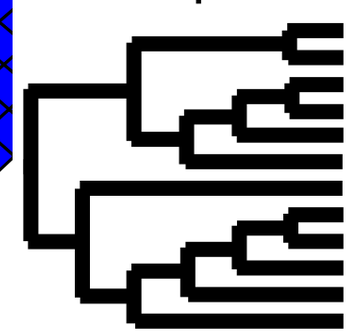
Work in progress with B. Larget

Importance sampling in phylogenetics



Target:
Posterior distribution

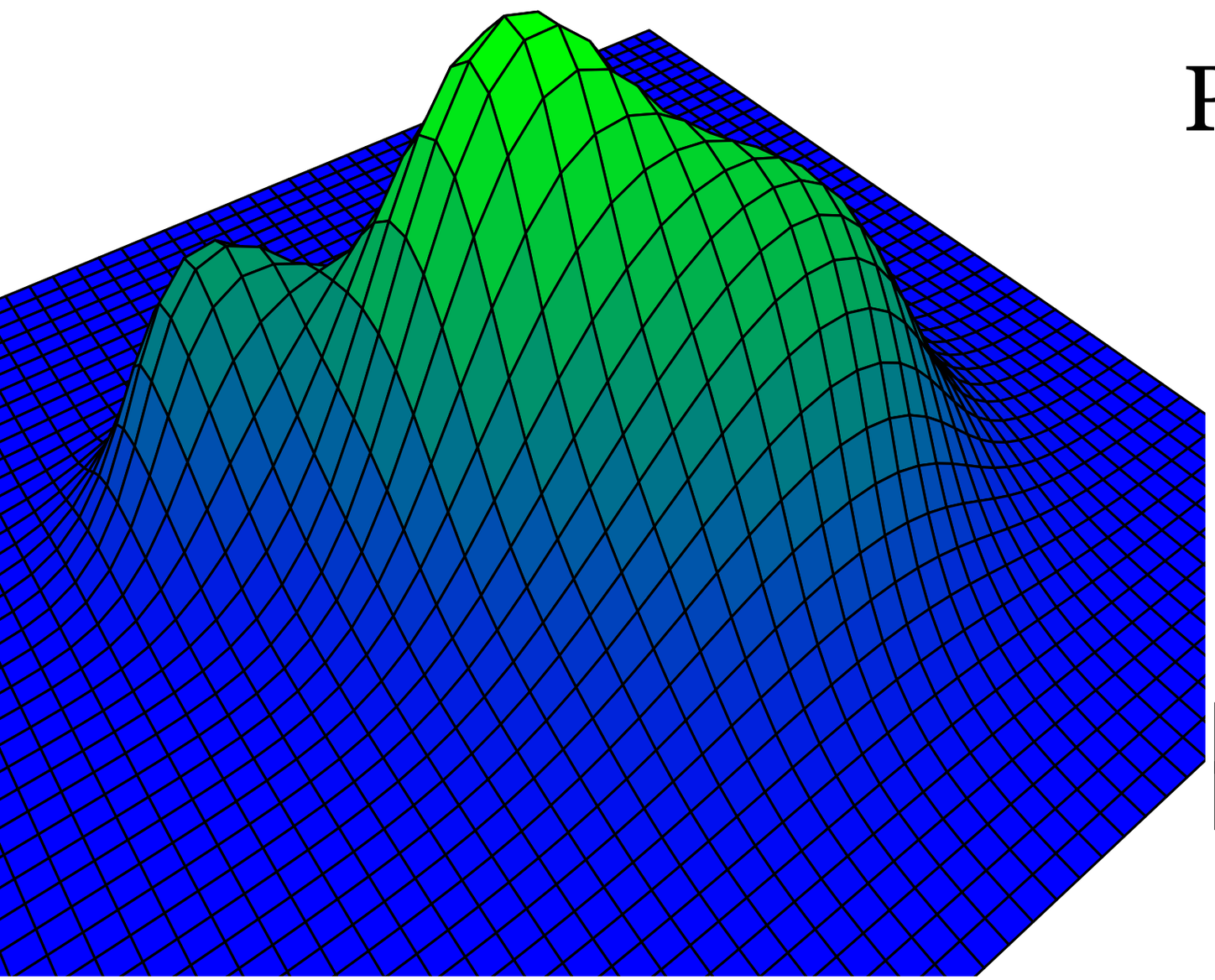
$$(T, t, Q)$$



Branch lengths

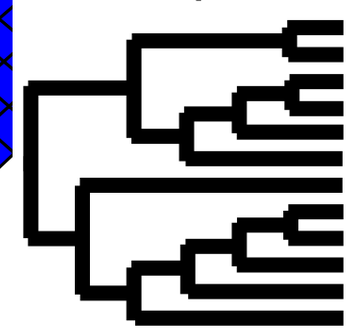
Work in progress with B. Larget

Importance sampling in phylogenetics



Target:
Posterior distribution

$$(T, t, Q)$$

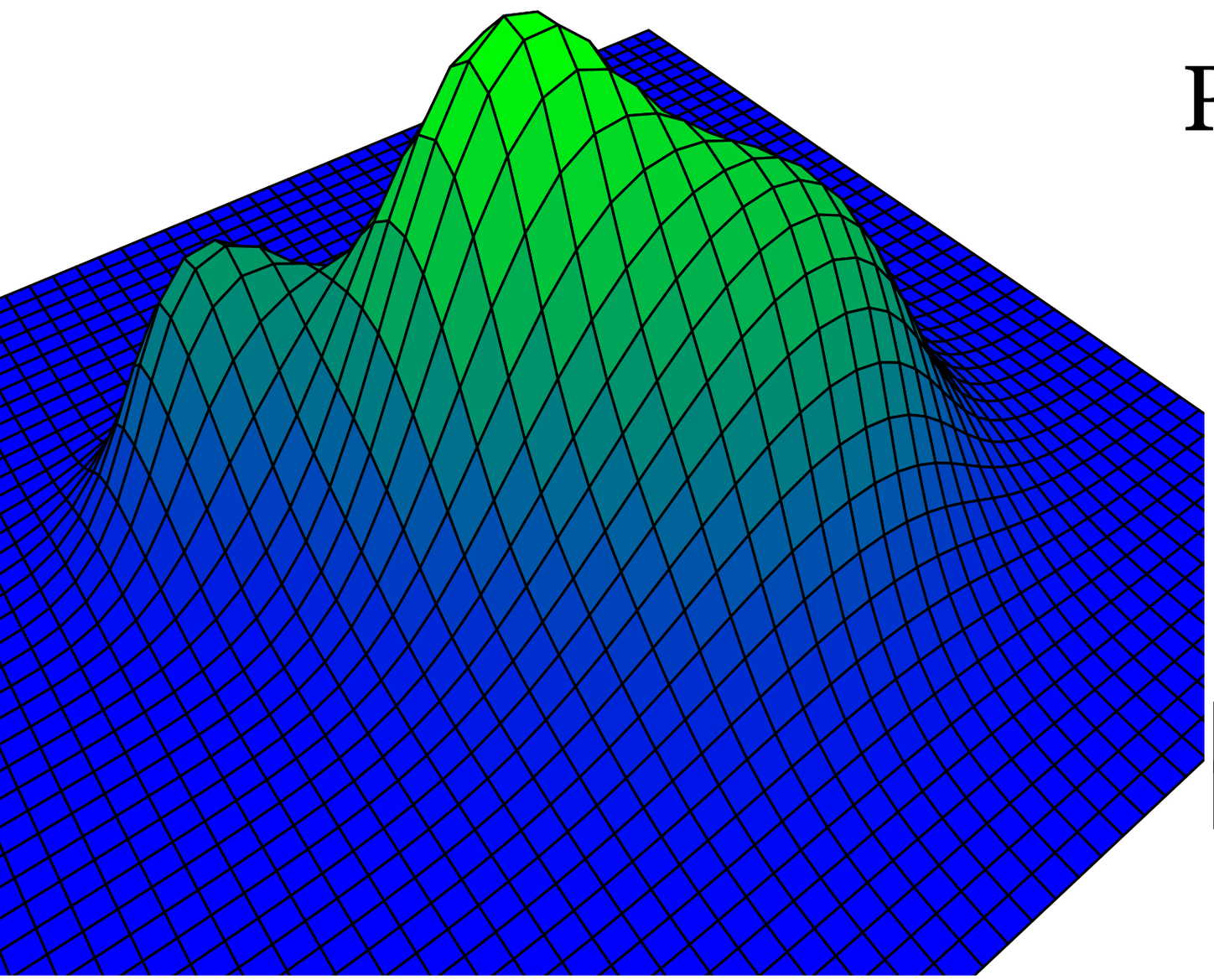


Branch lengths

Rate matrix

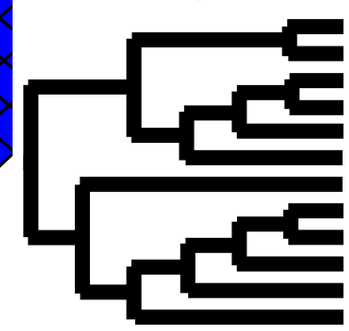
Work in progress with B. Larget

Importance sampling in phylogenetics



Target:
Posterior distribution

$$(T, t)$$

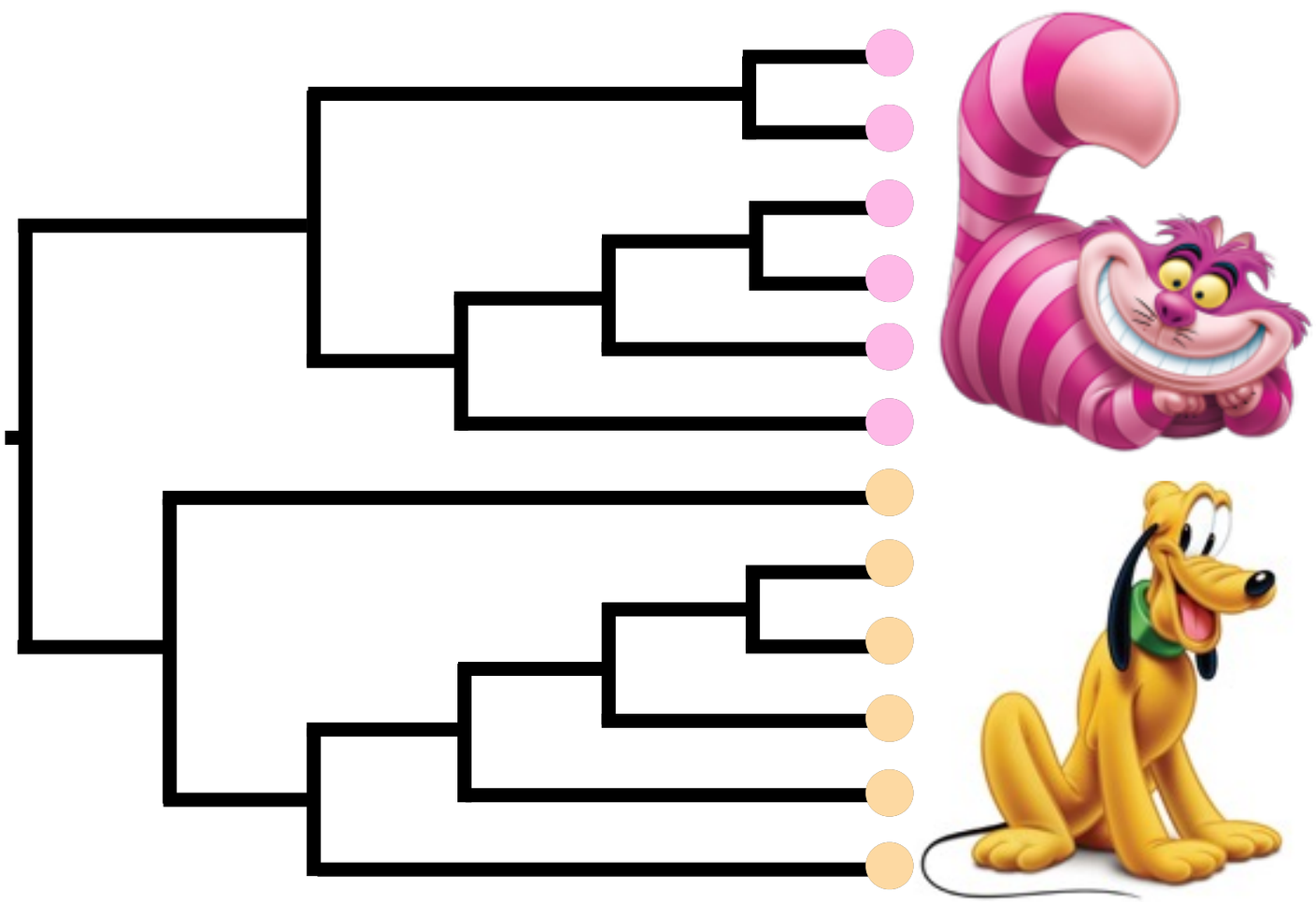


Branch lengths

Work in progress with B. Larget

Proposal density for topology

Conditional clade distribution: Sister clades are approximately conditionally independent

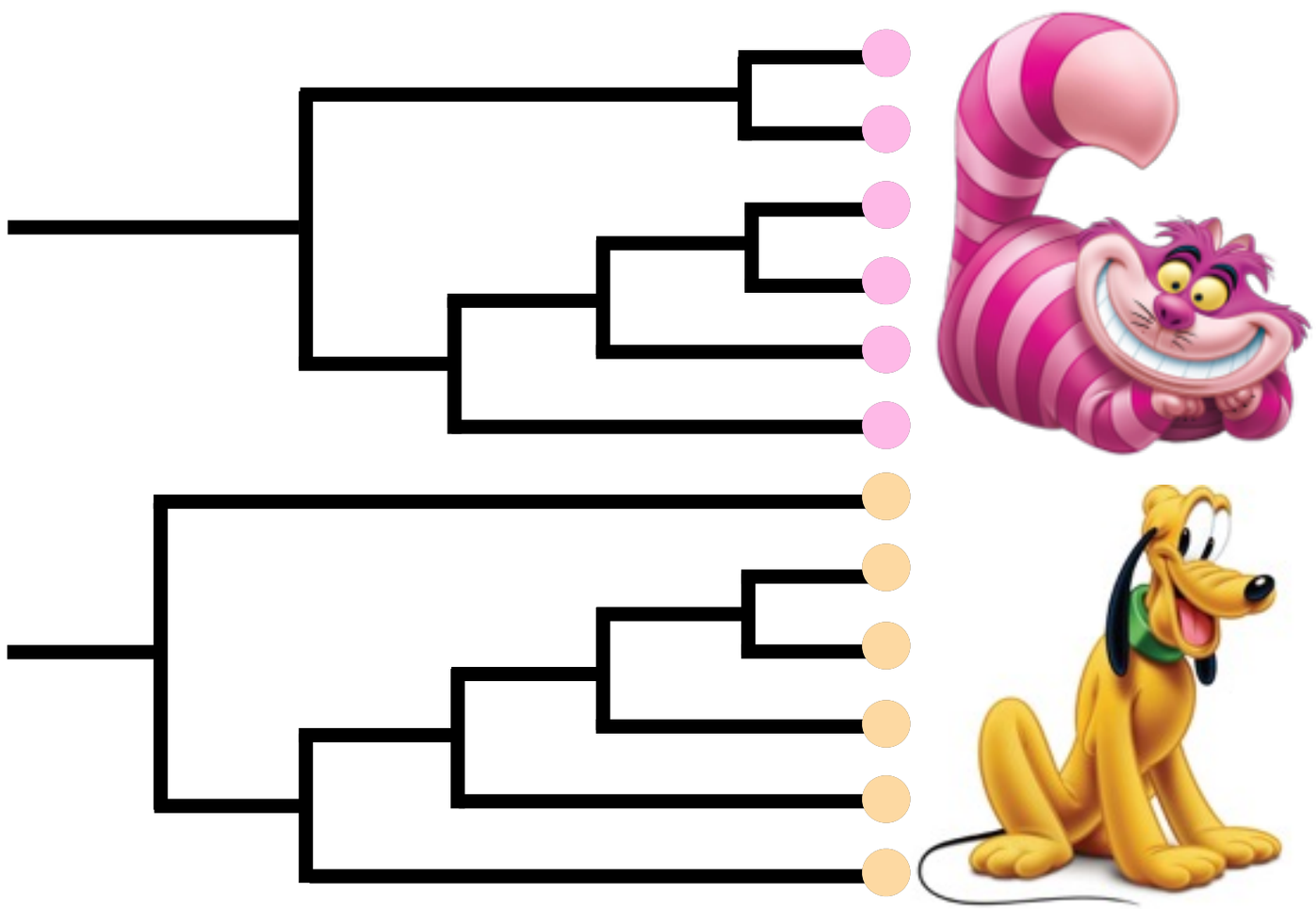


(Larget, 2013)

Work in progress with B. Larget

Proposal density for topology

Conditional clade distribution: Sister clades are approximately conditionally independent

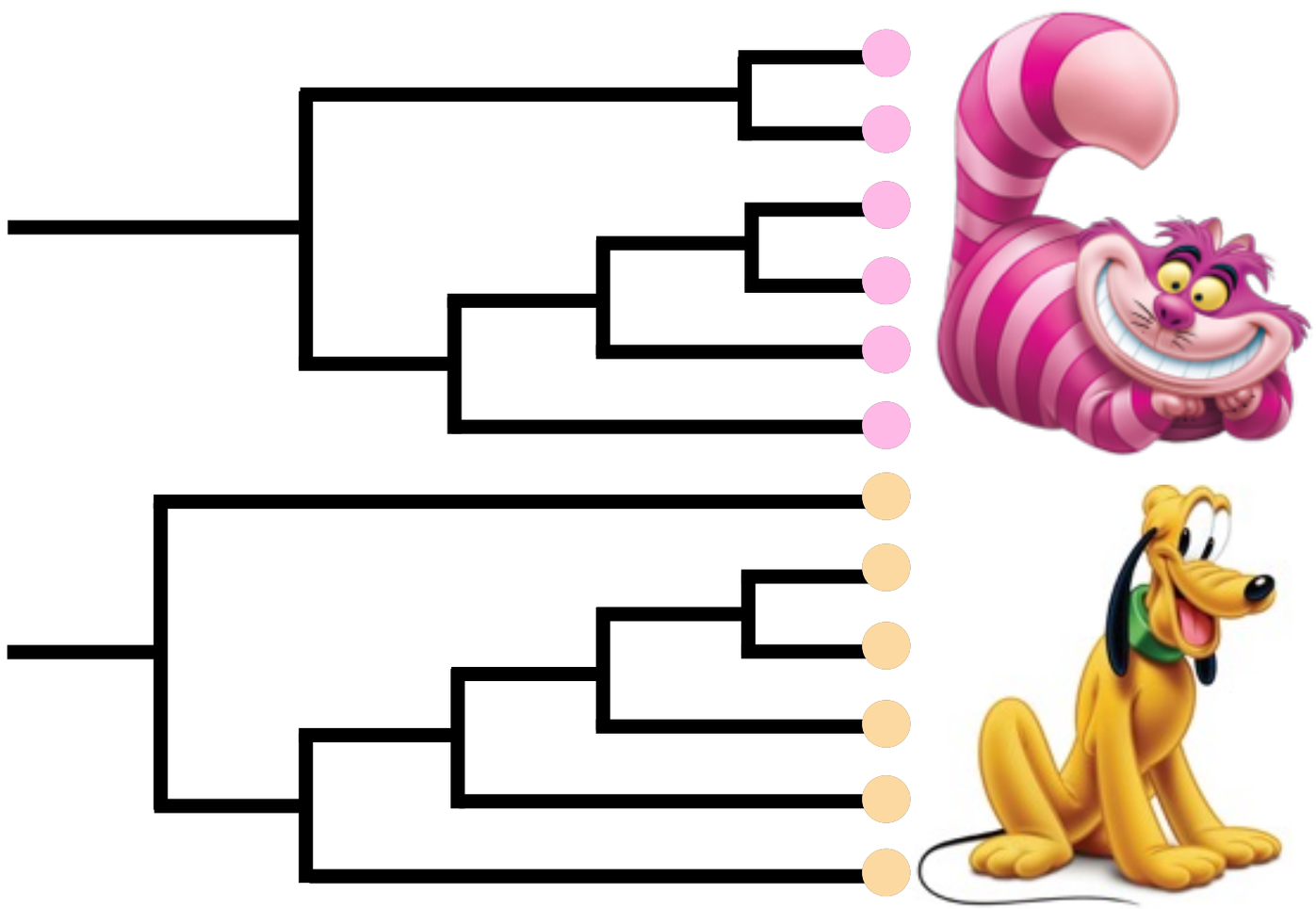


(Larget, 2013)

Work in progress with B. Larget

Proposal density for topology

Conditional clade distribution: Sister clades are approximately conditionally independent



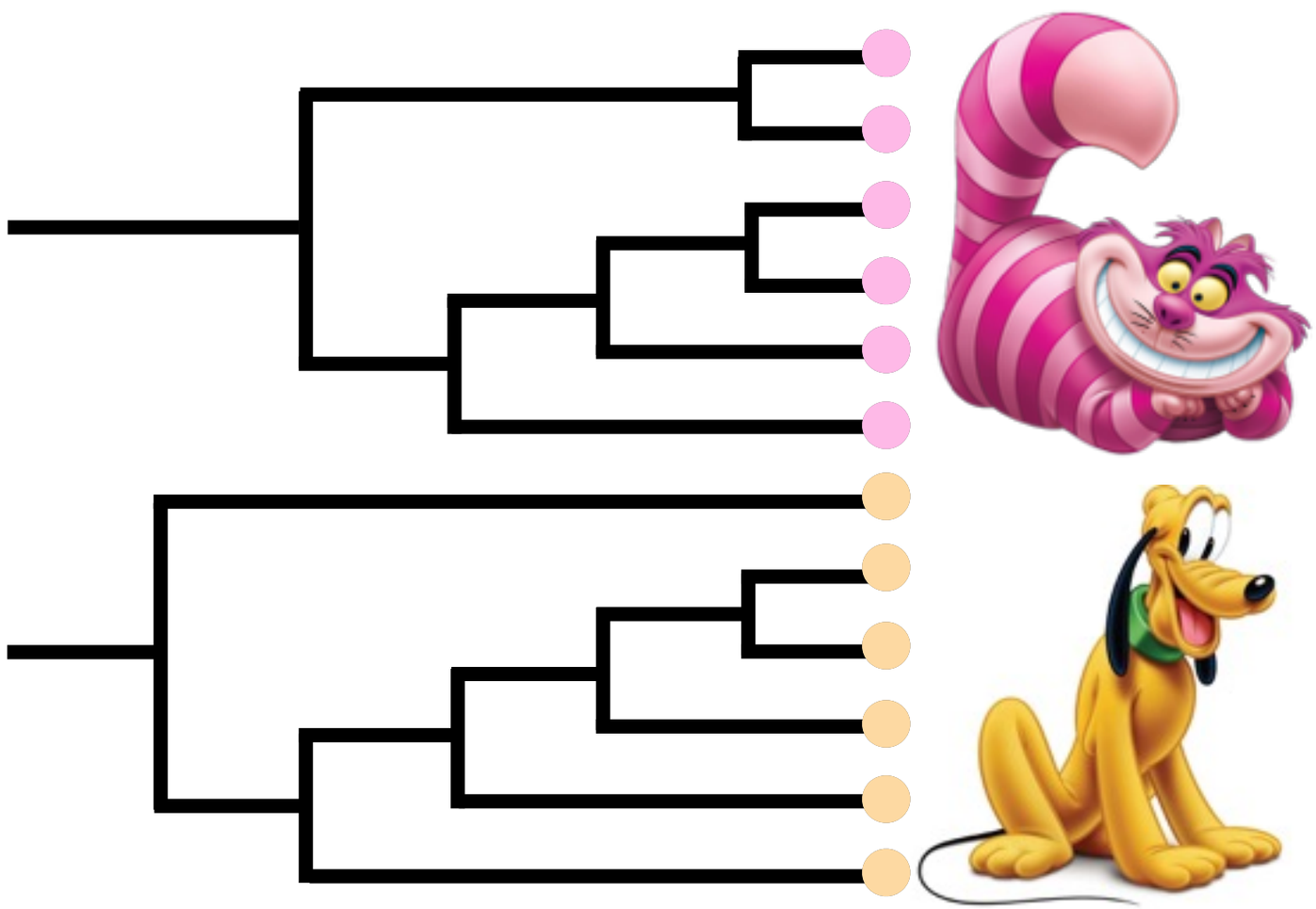
1,2,3,4

(Larget, 2013)

Work in progress with B. Larget

Proposal density for topology

Conditional clade distribution: Sister clades are approximately conditionally independent



1,2,3,4

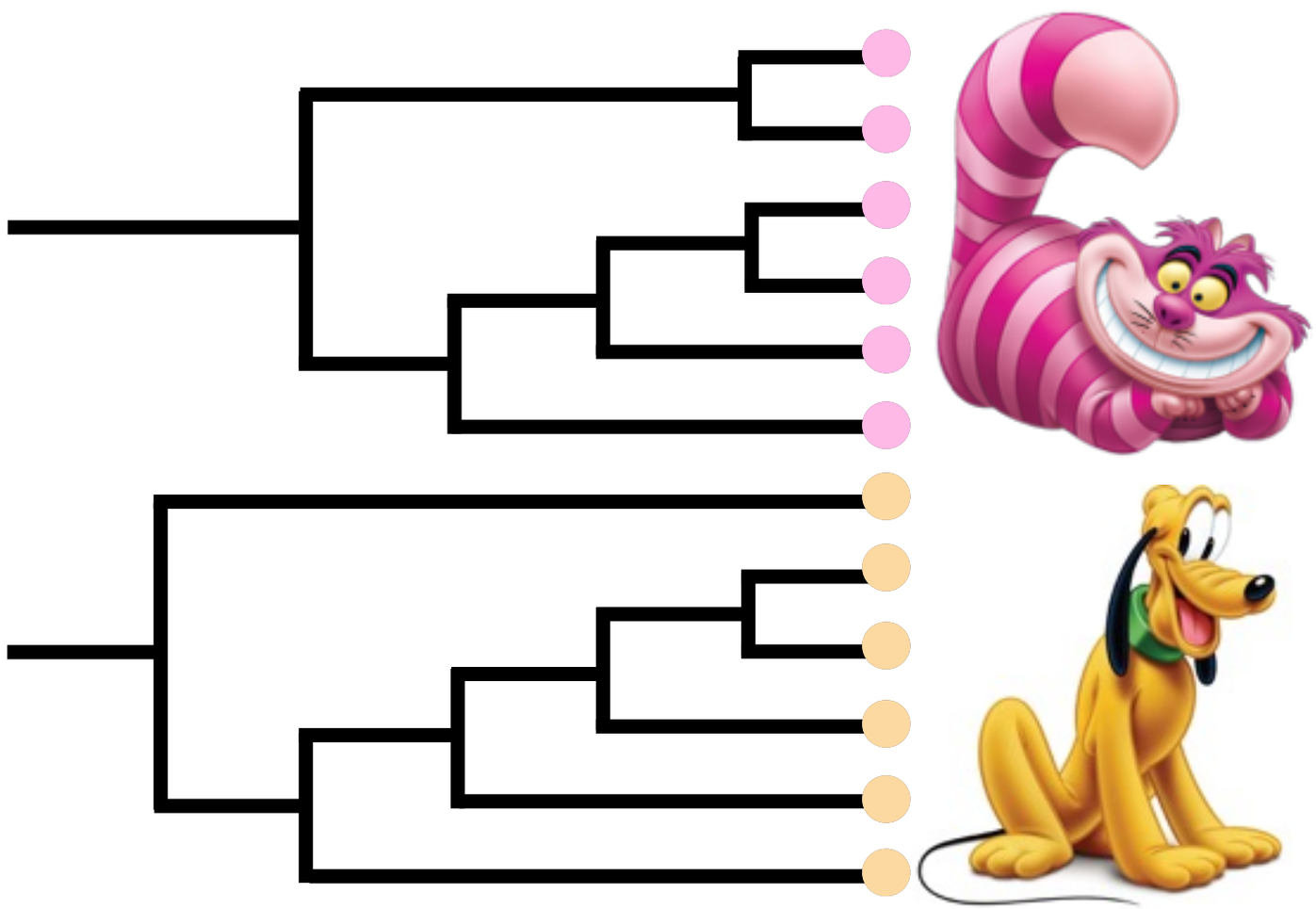
1 2,3,4

(Larget, 2013)

Work in progress with B. Larget

Proposal density for topology

Conditional clade distribution: Sister clades are approximately conditionally independent



1,2,3,4

1 2,3,4

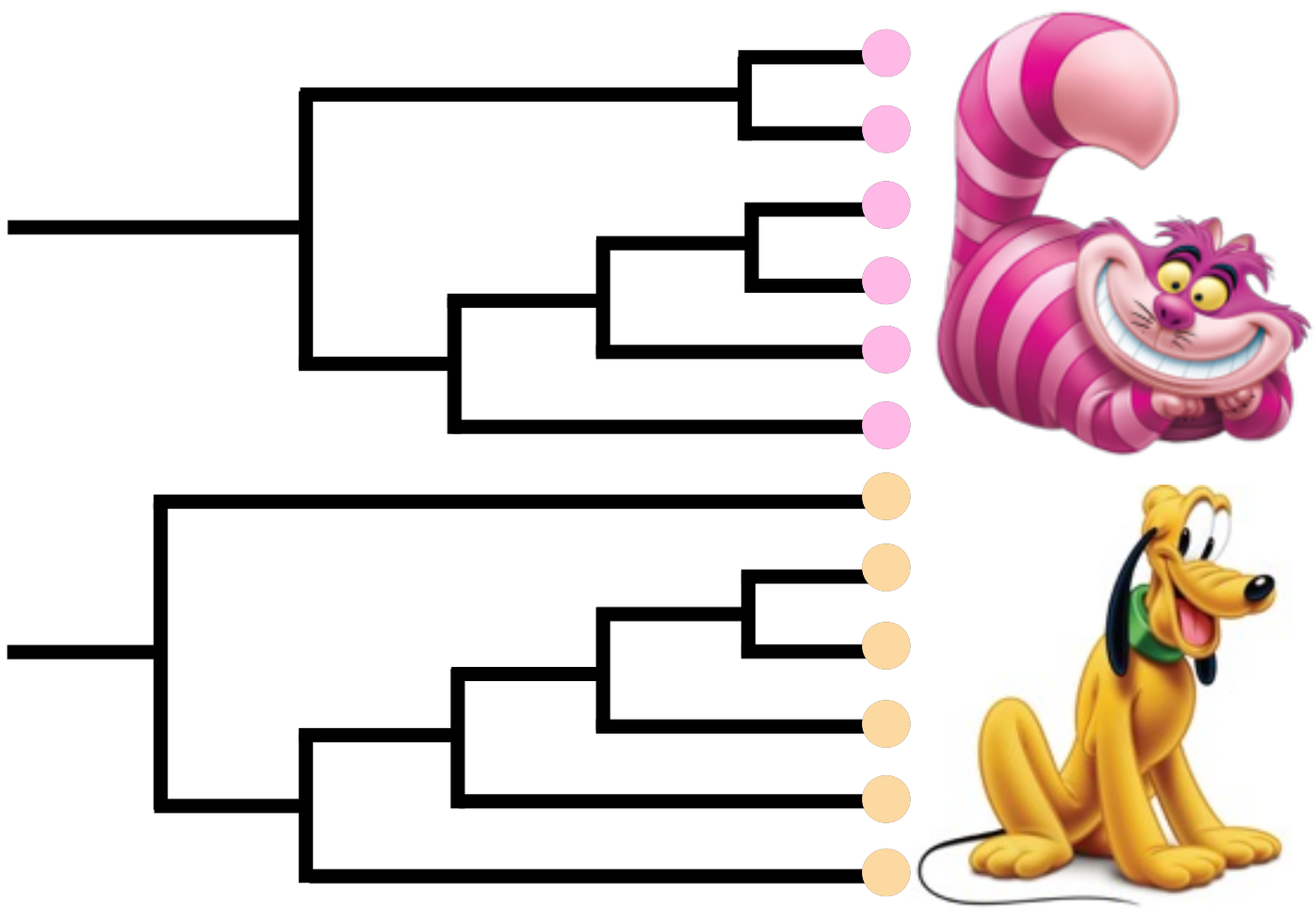
2,3 4

(Larget, 2013)

Work in progress with B. Larget

Proposal density for topology

Conditional clade distribution: Sister clades are approximately conditionally independent



1,2,3,4

1 2,3,4

2,3 4

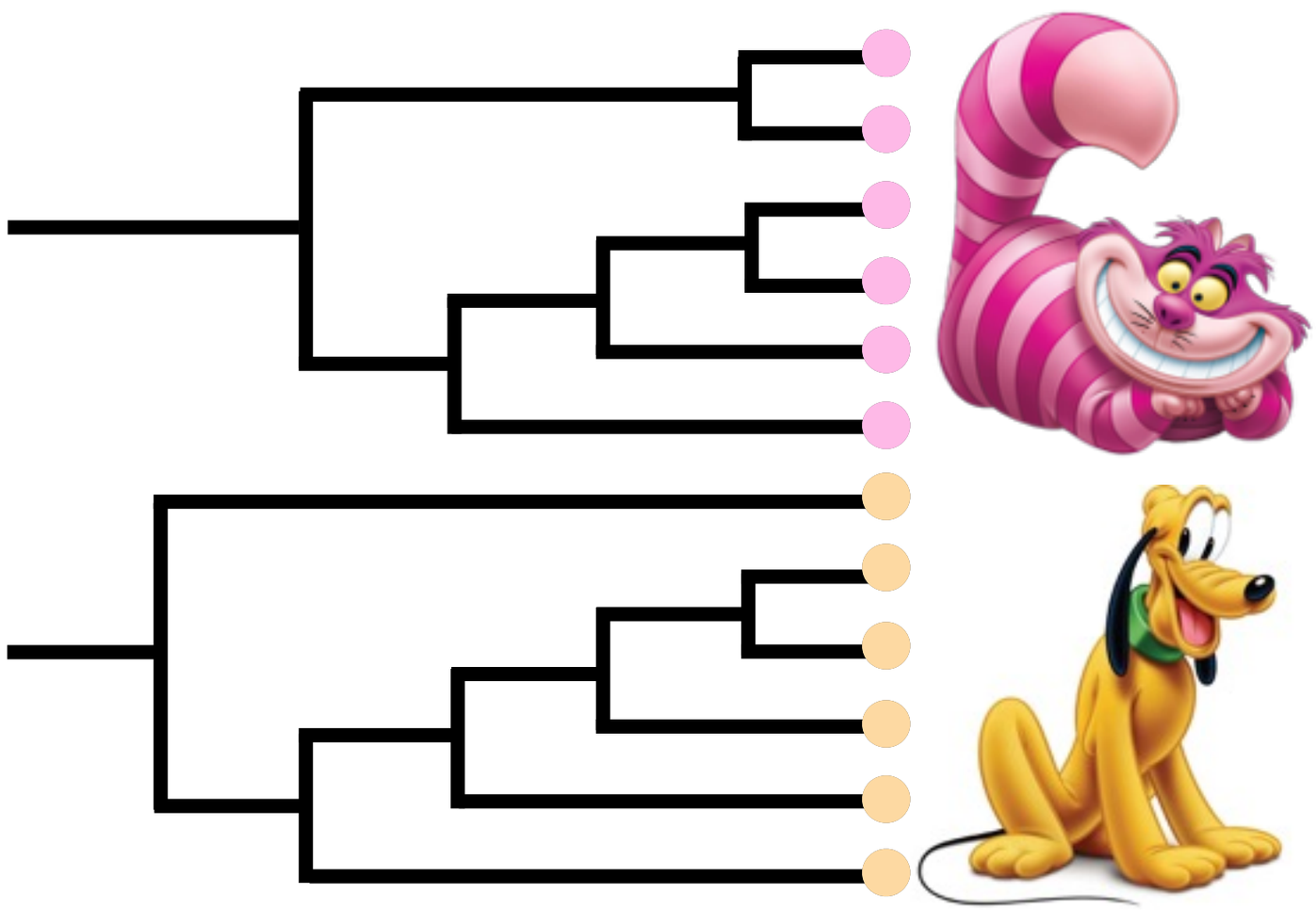
2 3

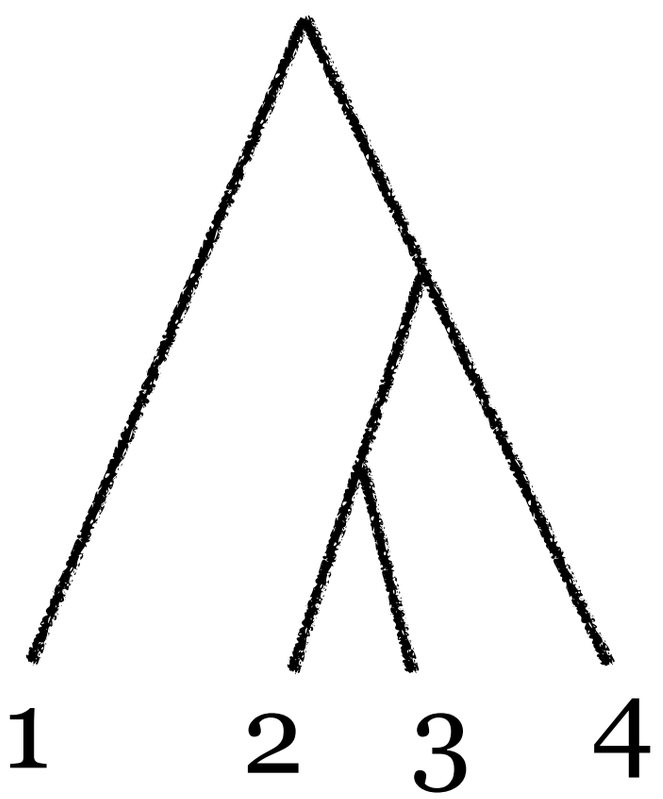
(Larget, 2013)

Work in progress with B. Larget

Proposal density for topology

Conditional clade distribution: Sister clades are approximately conditionally independent





$P(1|2,3,4)P(2,3|4)P(2|3)$

(Larget, 2013)

Work in progress with B. Larget

Proposal density for topology

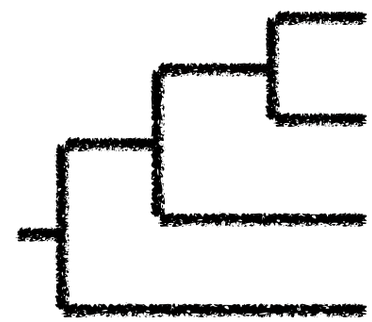
Bootstrap sample
of Neighbor-Joining
trees

AAGTCTAG
AAGTCTAG
AACTCTAG
AATTCTAG

Work in progress with B. Larget

Proposal density for topology

Bootstrap sample
of Neighbor-Joining
trees

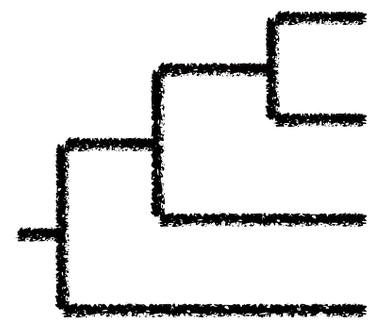


AAGTCTAG
AAGTCTAG
AACTCTAG
AATTCTAG

Work in progress with B. Larget

Proposal density for topology

Bootstrap sample
of Neighbor-Joining
trees

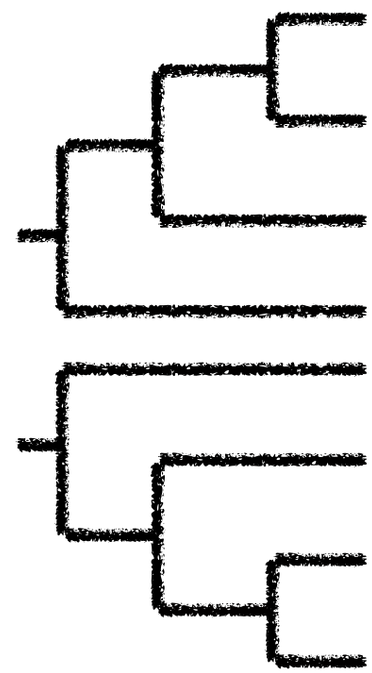


TAGAGCTA
TAGAGCTA
TAGACCTA
TAGATCTA

Work in progress with B. Larget

Proposal density for topology

Bootstrap sample
of Neighbor-Joining
trees



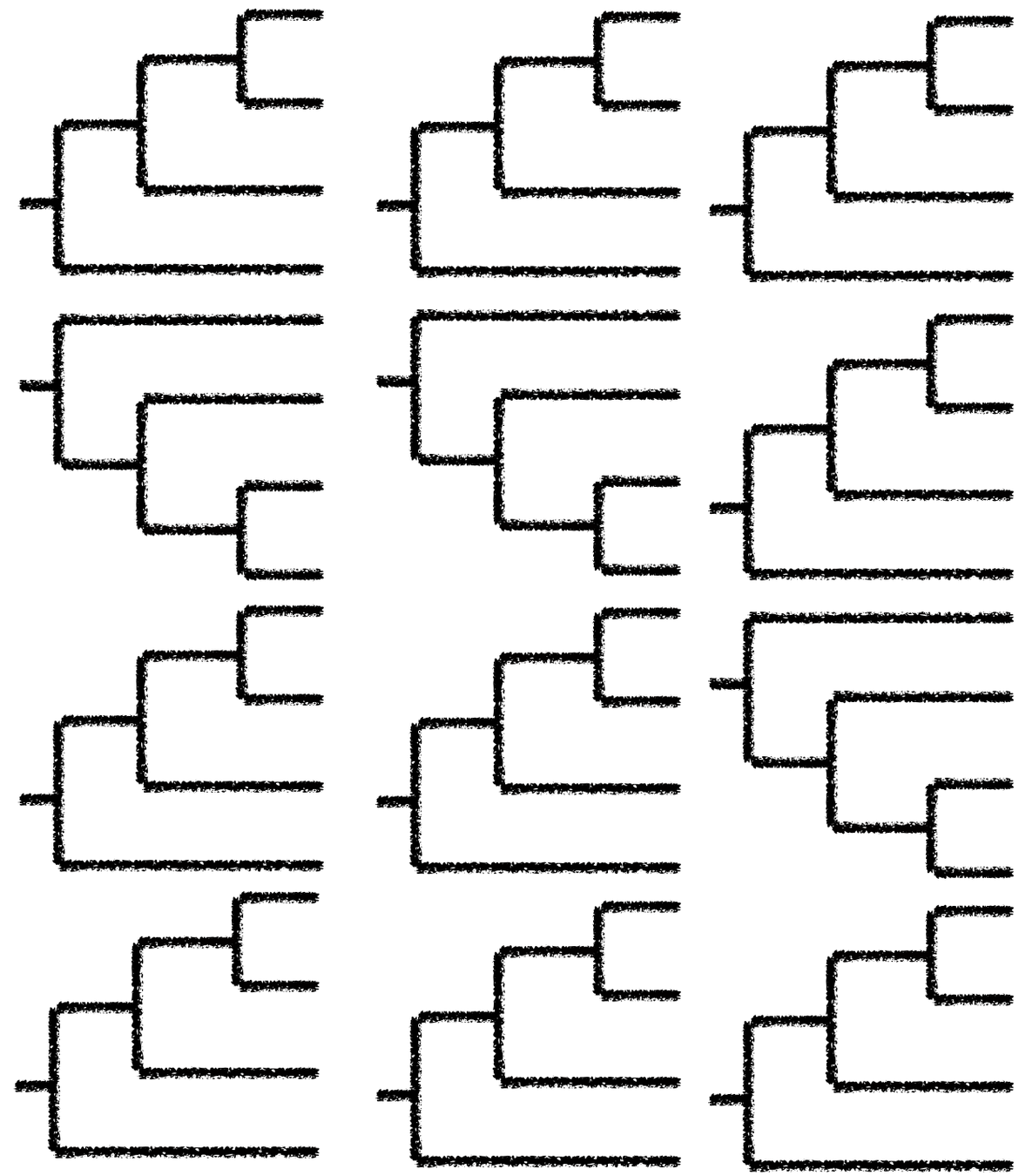
TAGAGCTA
TAGAGCTA
TAGACCTA
TAGATCTA

Work in progress with B. Larget

Proposal density for topology

Bootstrap sample of Neighbor-Joining trees

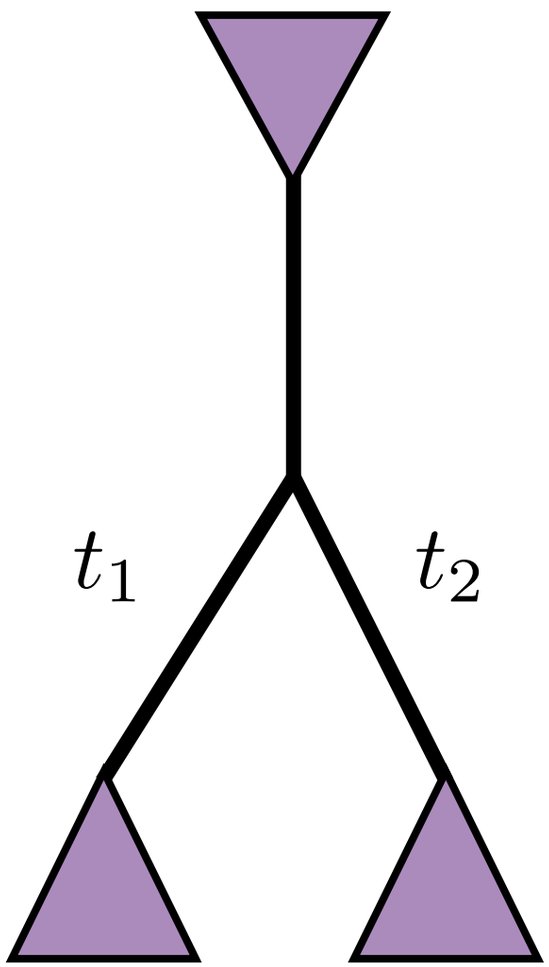
TAGAGCTA
TAGAGCTA
TAGACCTA
TAGATCTA



Work in progress with B. Larget

Proposal density for branch lengths

Correlation of sister edges



$$(t_1, t_2) \sim Gamma$$

$$\mu = MLE$$

$$\Sigma = I^{-1}$$

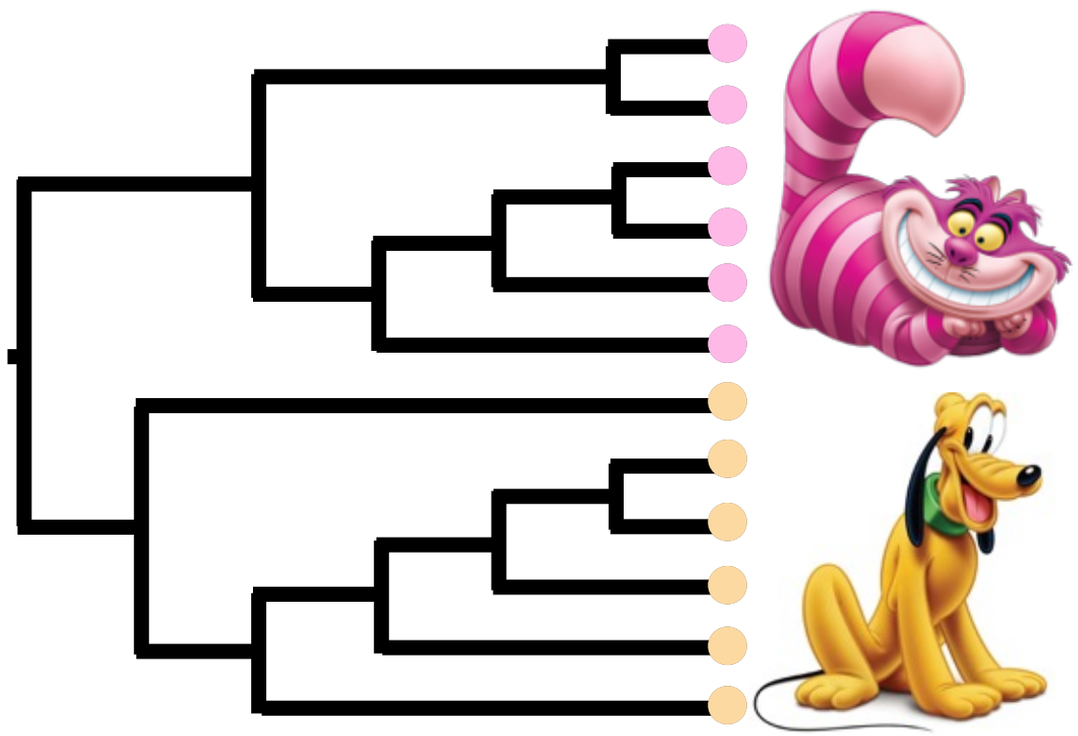
Work in progress with B. Larget

Importance sampling in phylogenetics: Bistro

- Fixed Q
- Sample a topology from clade distribution
- Sample branch lengths from Gamma
- Compute the likelihood of topology with branch lengths, and weight
- Repeat
- Do inference on weighted sample

Work in progress with B. Larget

Results

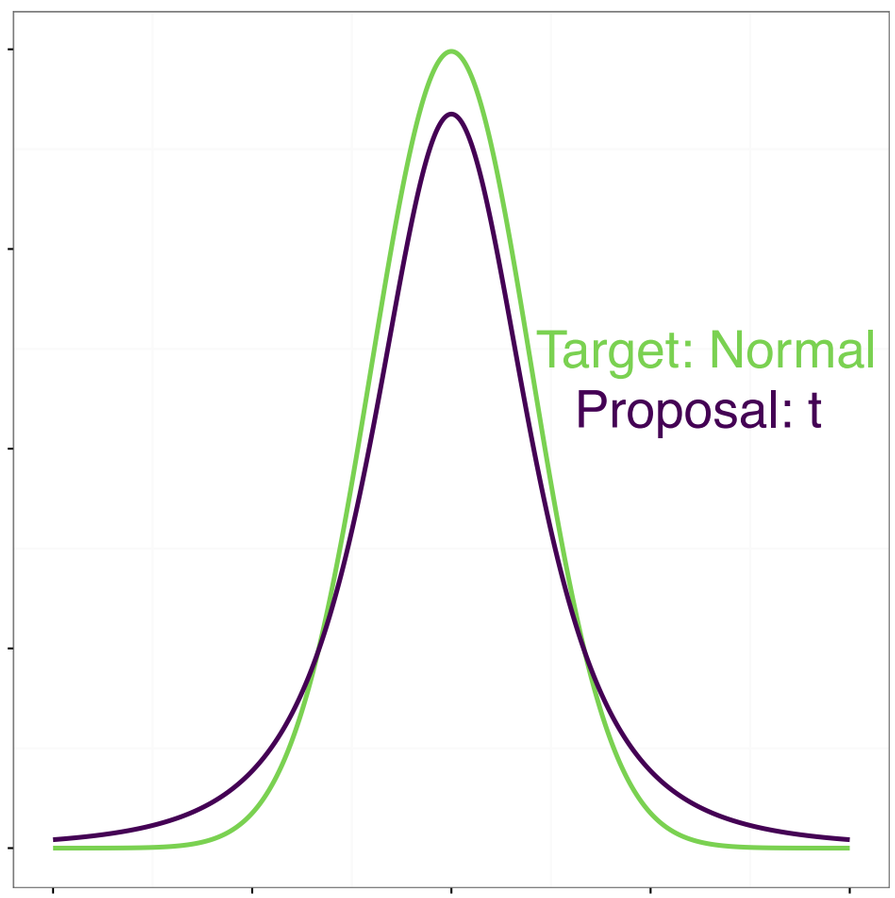


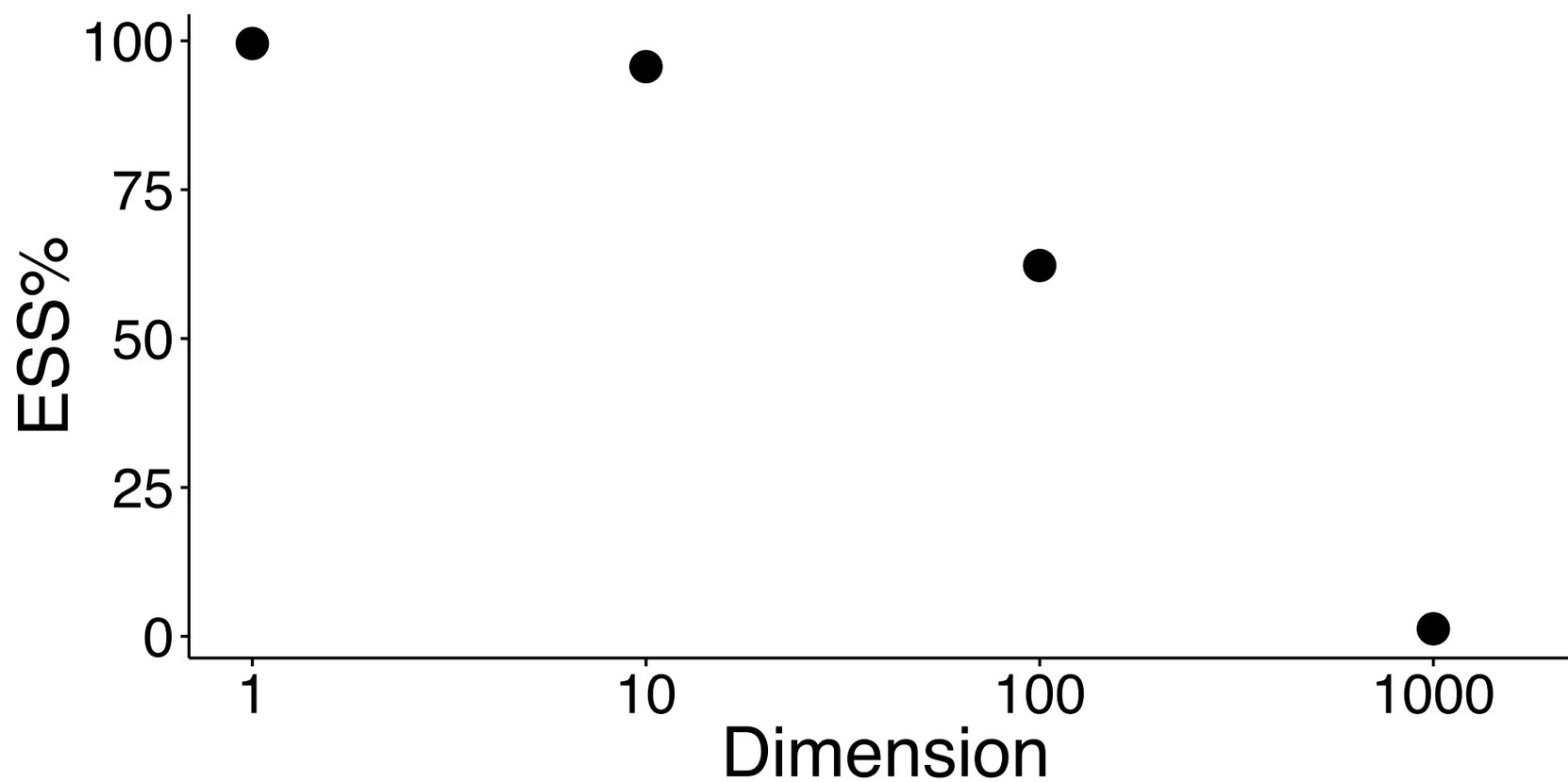
	MrBayes	Bistro
#Trees	1,000,000	1,000
ESS	250	116
Efficiency	0.025%	**12%**

Work in progress with B. Larget

Challenges

Curse of dimensionality





Extra variance 10%

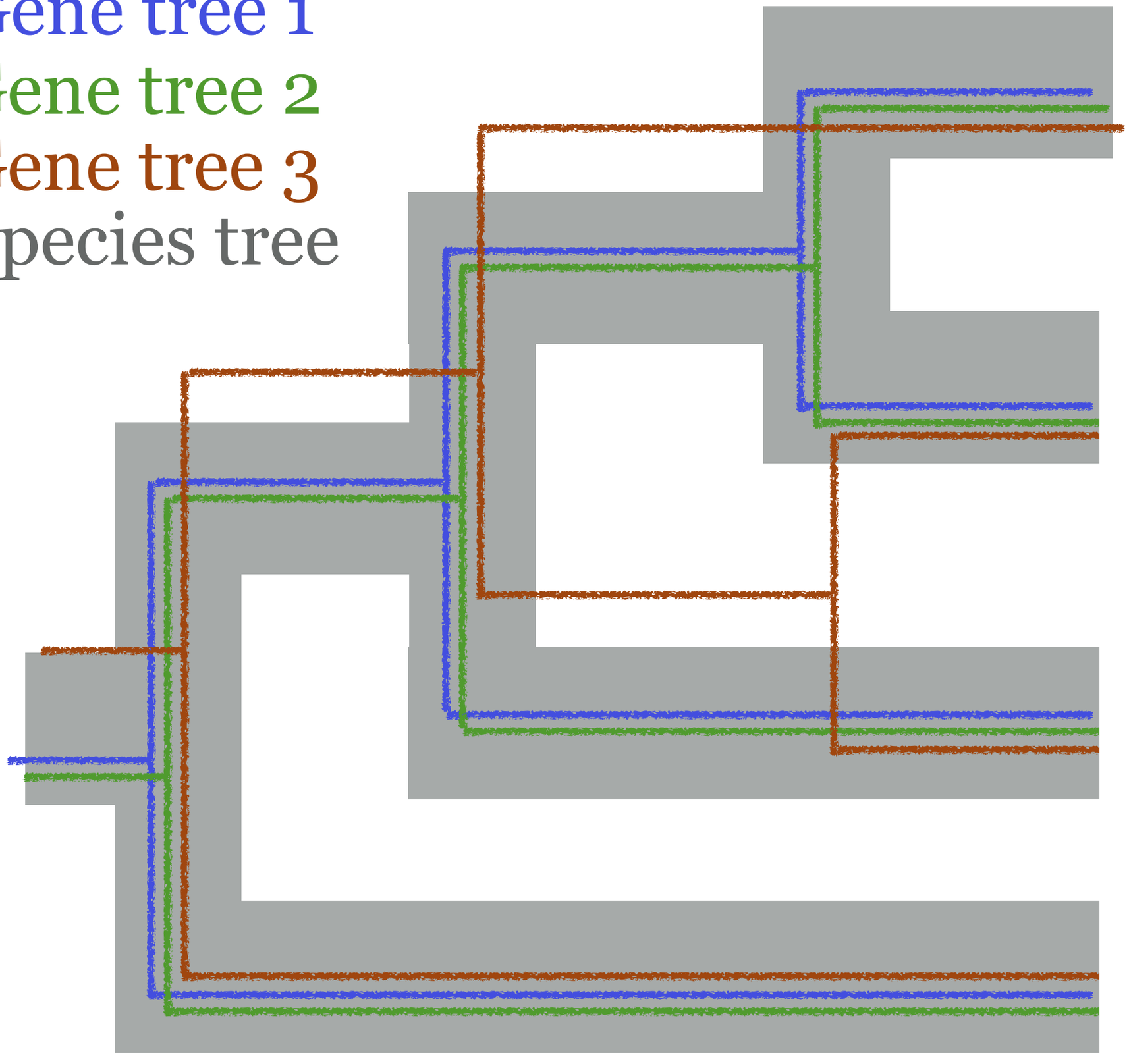
Work in progress with B. Larget

We see efficiency gains, but

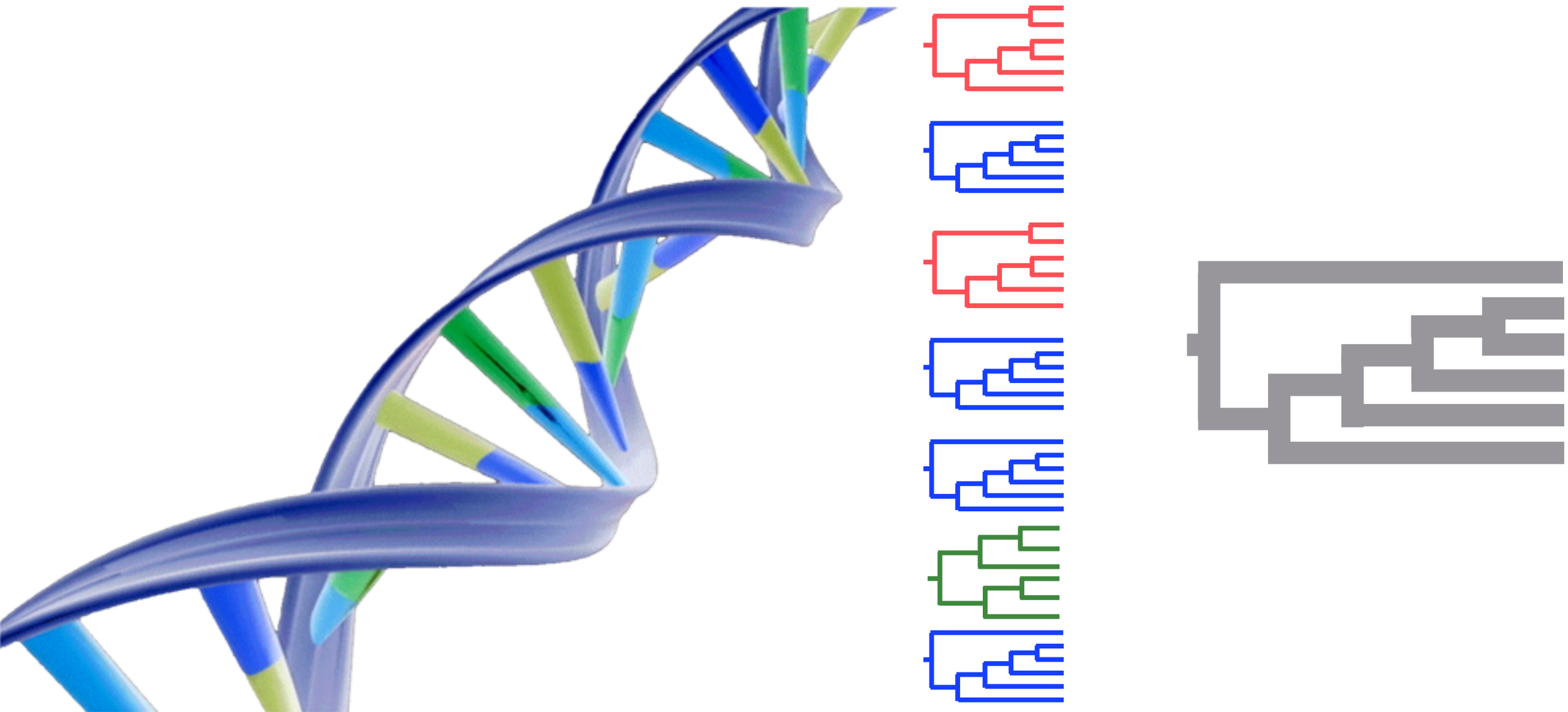
- **Topology:** bootstrap sample does not work for big trees
    - Ideas: Consensus or (Fréchet) mean tree, density with exponential decay
- **Branch lengths:** Dimension and correlation
- **Q:** Dirichlet proposal densities for base frequencies and rates, mean/var estimate?

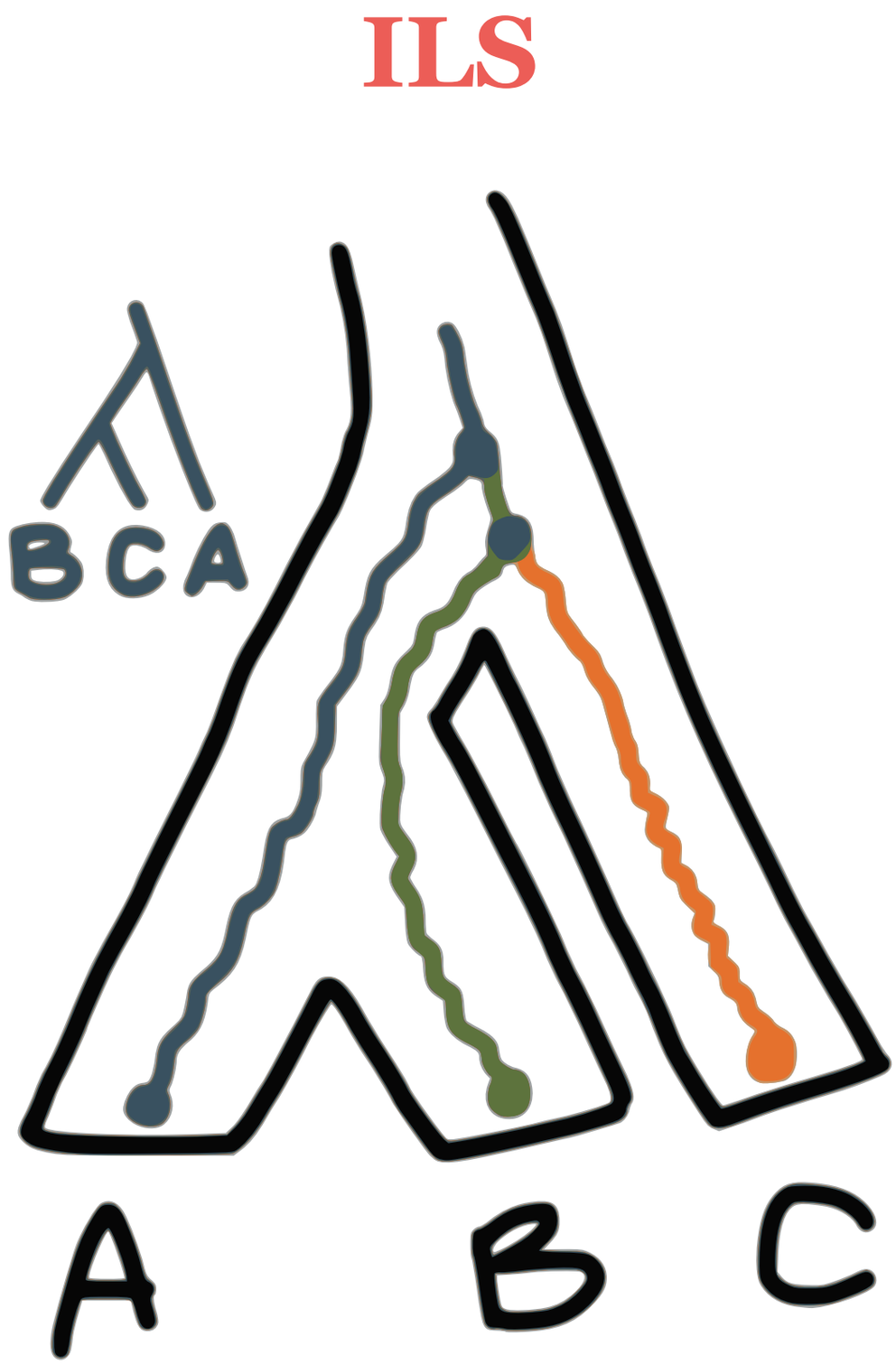
Work in progress with B. Larget

Pseudolikelihood estimation of phylogenetic networks

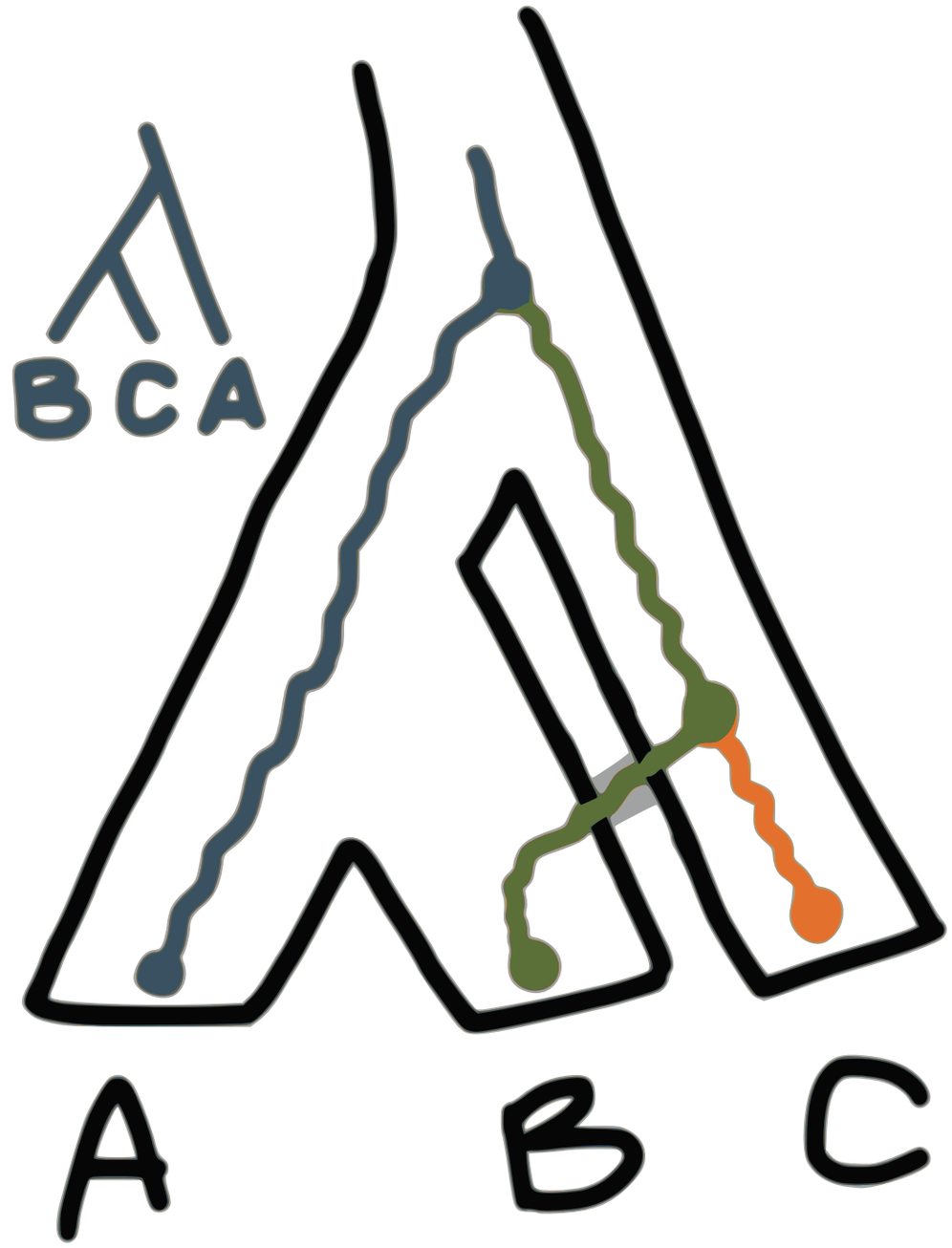


How to estimate the species tree with gene tree discordance?



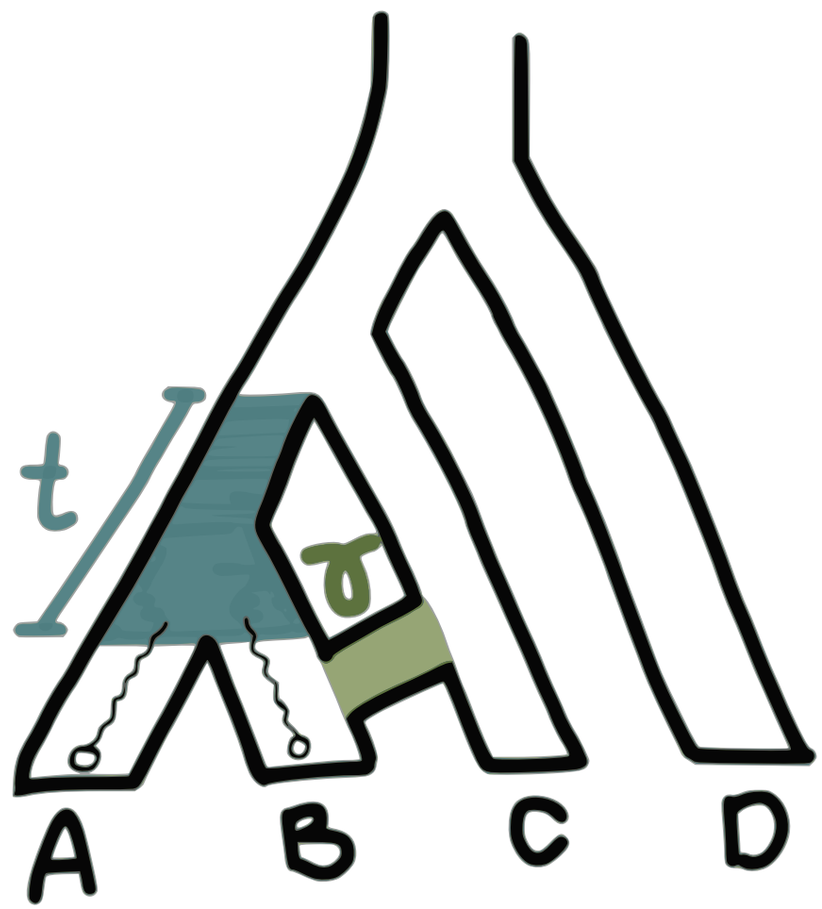


Tree



Network

Multispecies coalescent model on a network

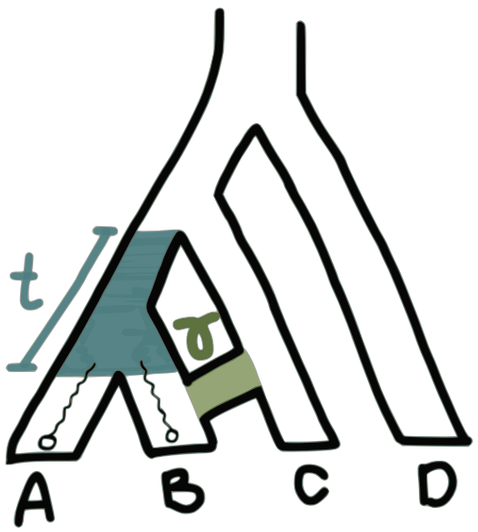


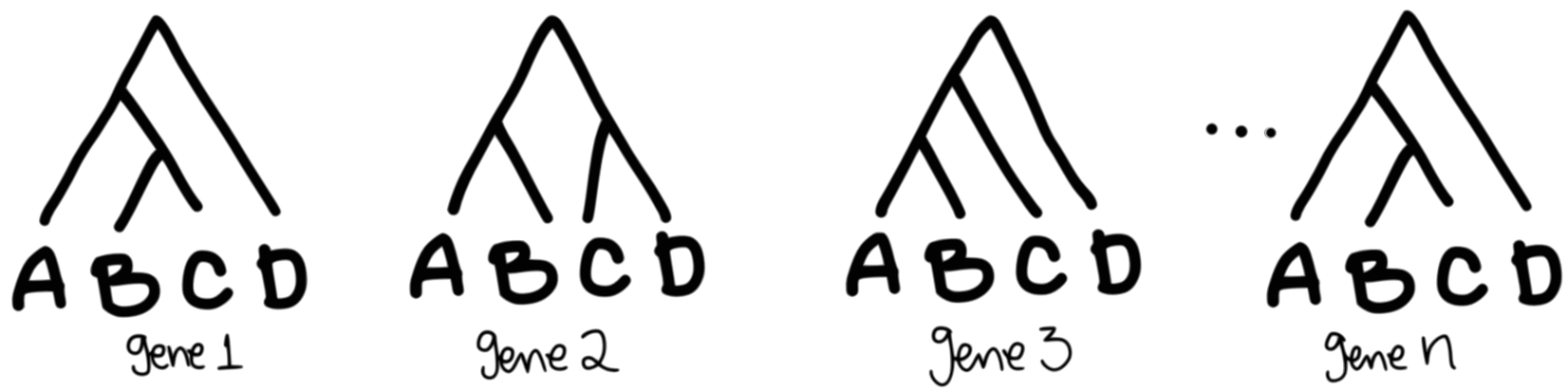
(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

Multispecies coalescent model on a network

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

Maximum likelihood





Model

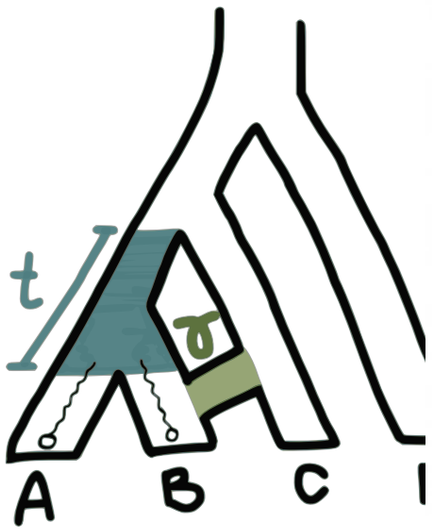
Data

$$L(network, t, \gamma) = \prod_{g} P(g|network, t, \gamma)$$

PhyloNet

(Yu, Dong, Liu, Nakhleh, 2014)

Maximum likelihood



Model

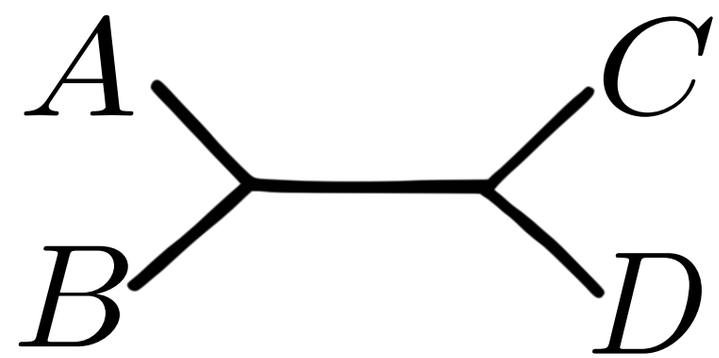
$L($ $)$

PhyloNet

(Yu, Dong, Liu, Nakhleh, 2014)

Maximum **pseudo**likelihood

Quartet-based inference



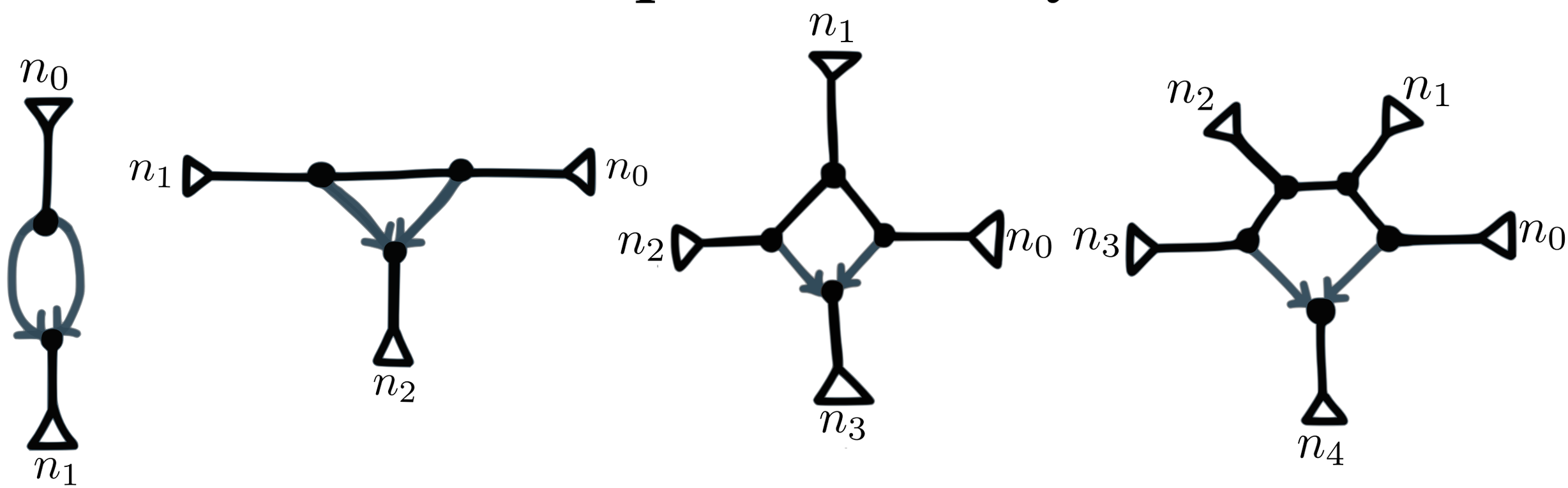
$$\tilde{L}(network, t, \gamma) = \prod_{q \in Q(network)} Likelihood(q, t, \gamma)$$

(S-L, Ané, 2016, PLoS Genetics)

www.github.com/CRSL4/PhyloNetworks

Model identifiability

Can we detect the presence of hybridization?



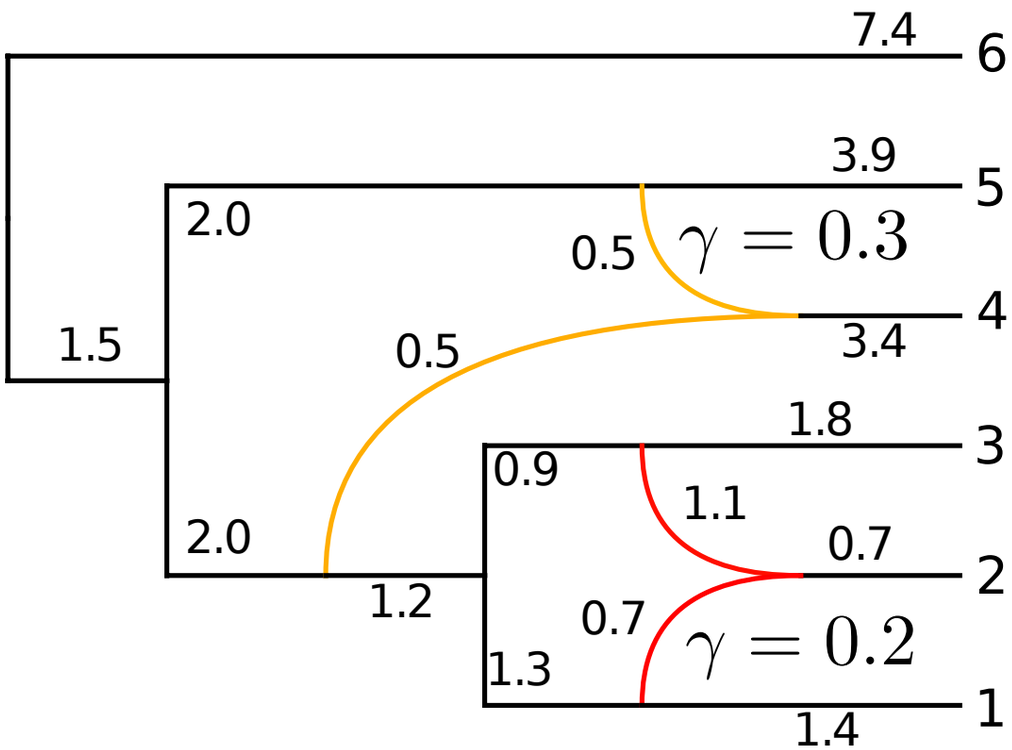
No **Yes** $(n_i, n_j \geq 2)$ **Yes** $(n_i \geq 2)$ **Yes**

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

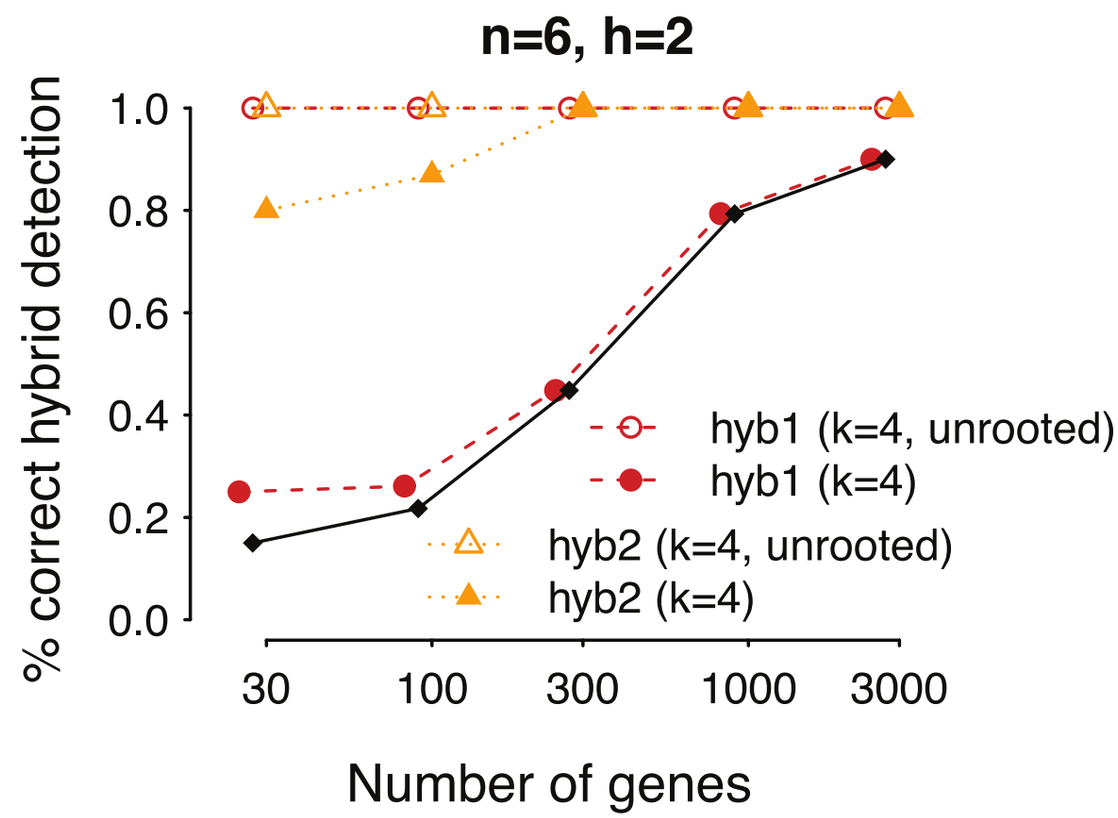
(S-L, Ané, 2016, PLoS Genetics)

SNaQ performance

Good diamond



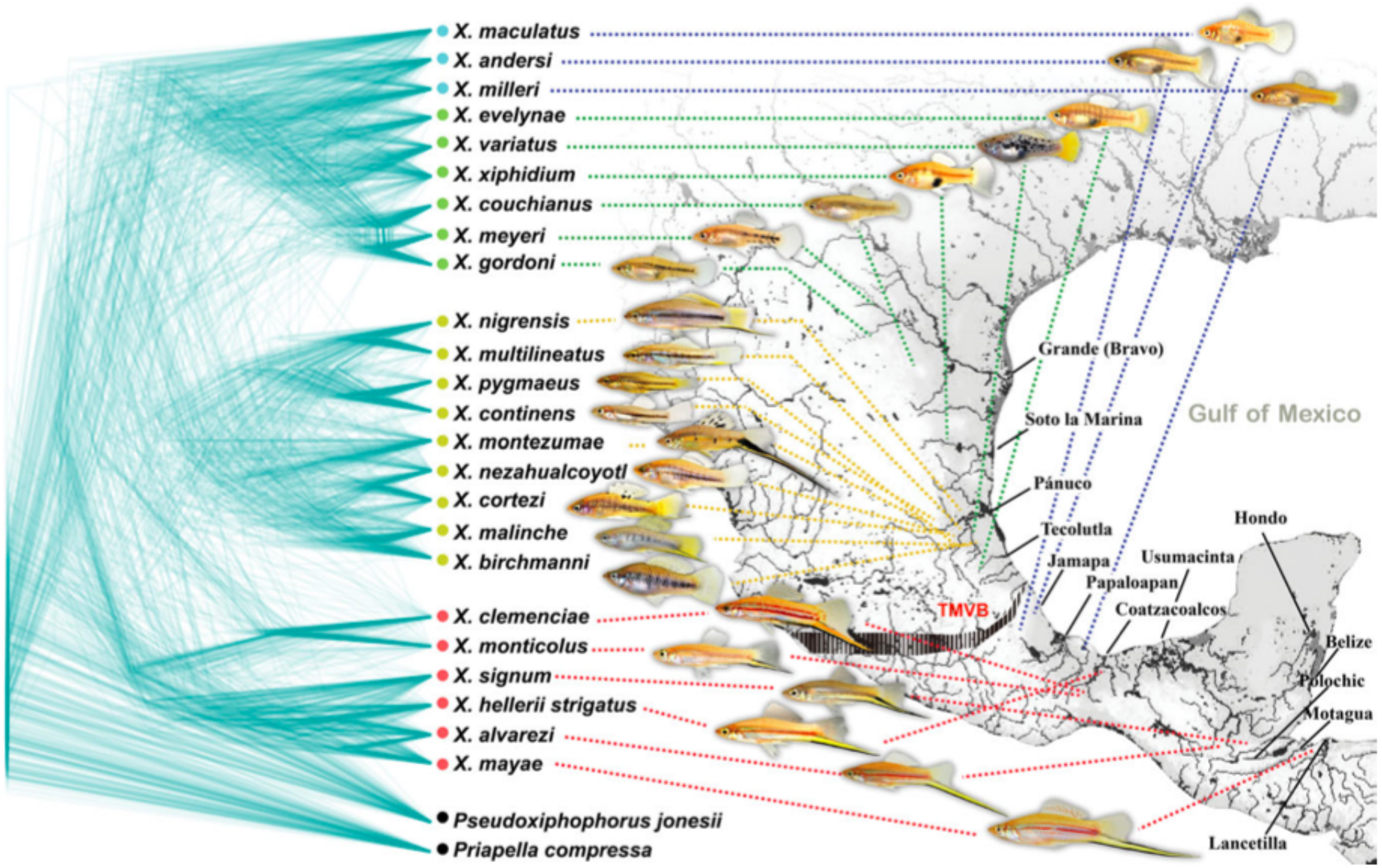
Bad diamond



(S-L, Ané, 2016, PLoS Genetics)

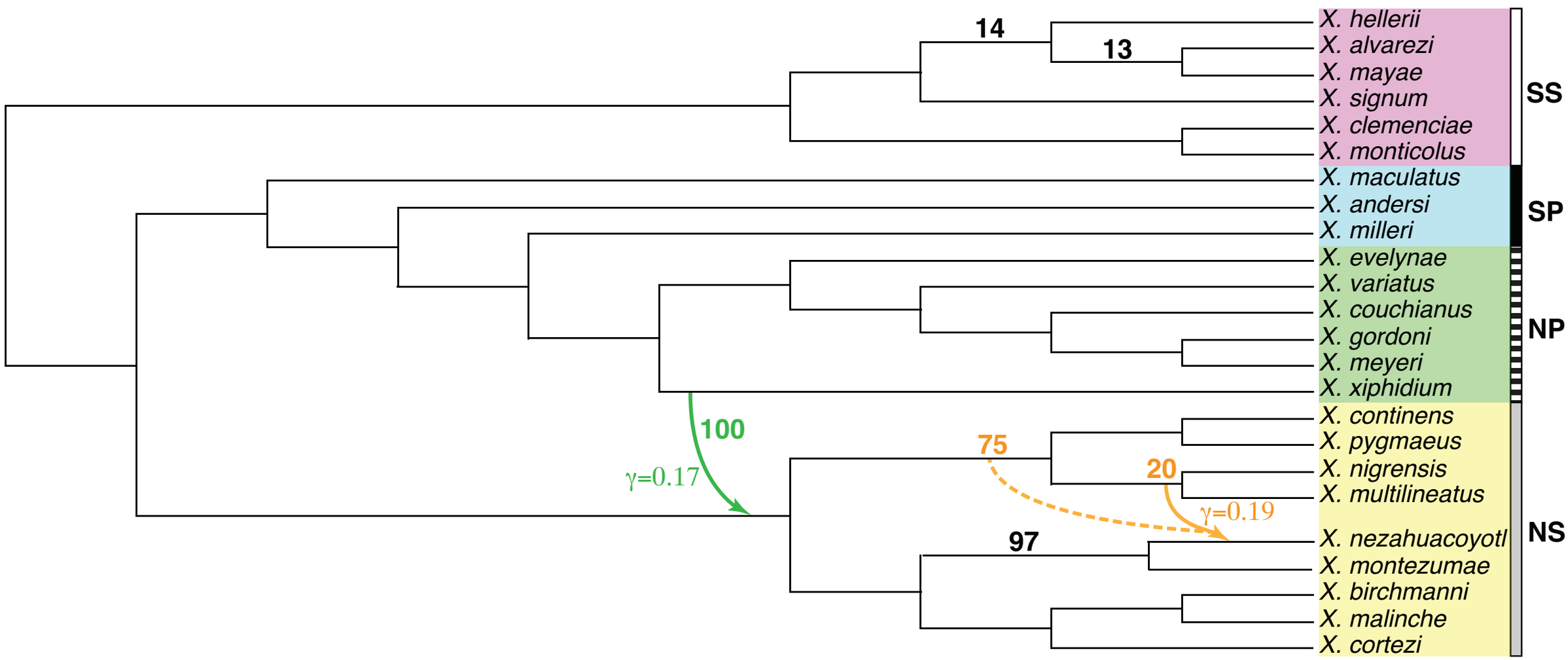
Xiphophorus fish data

1183 genes, 24 swordtails and platyfish



(Cui et al., 2013)

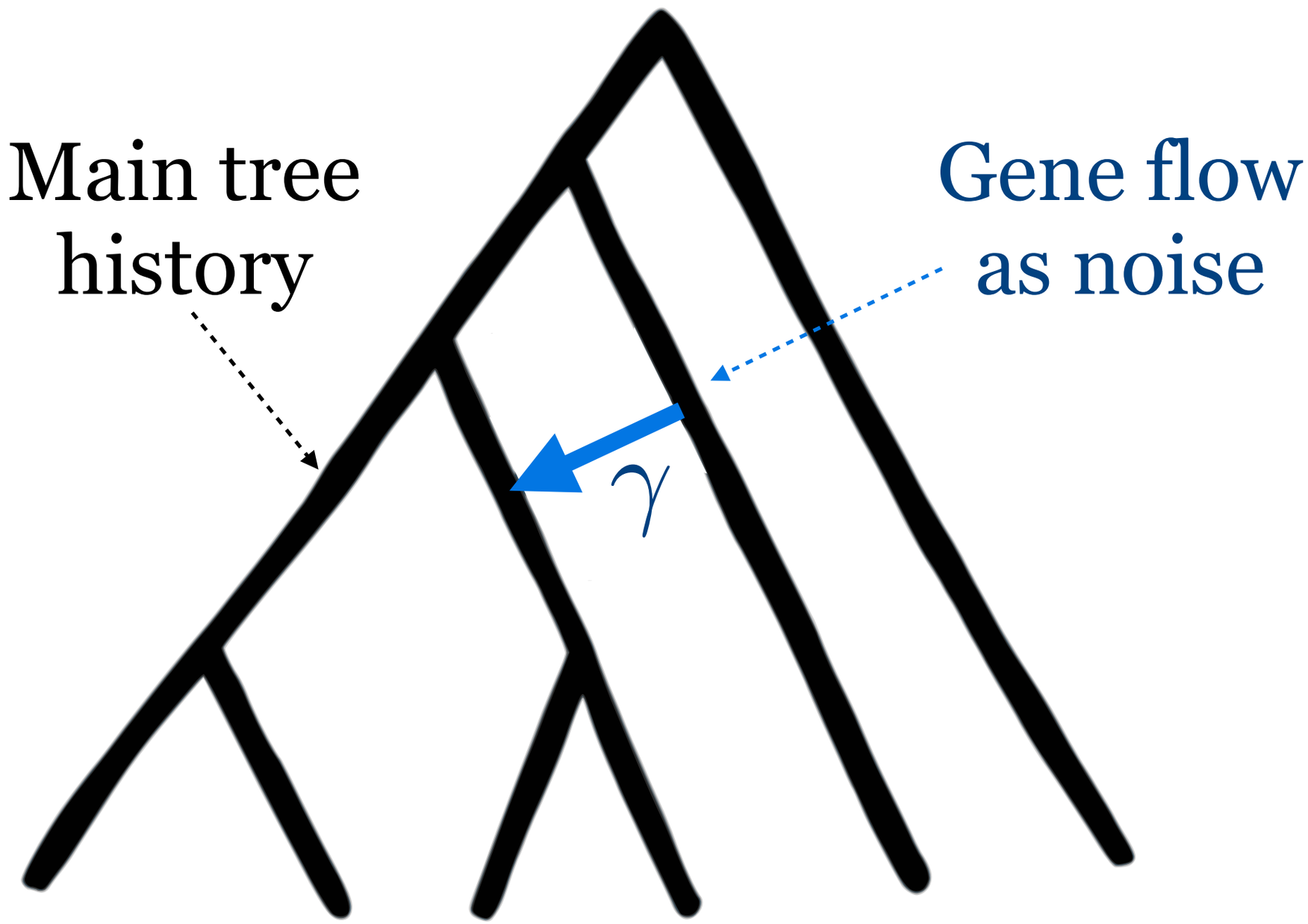
Xiphophorus fish data



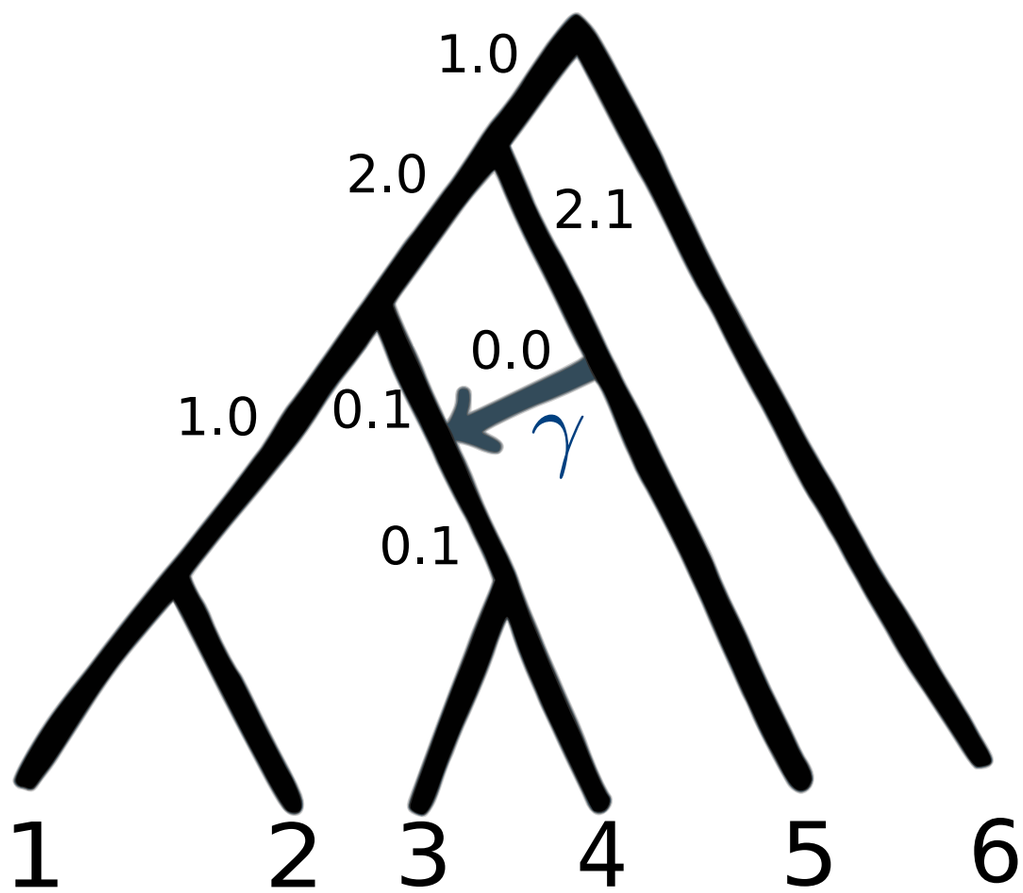
snaq

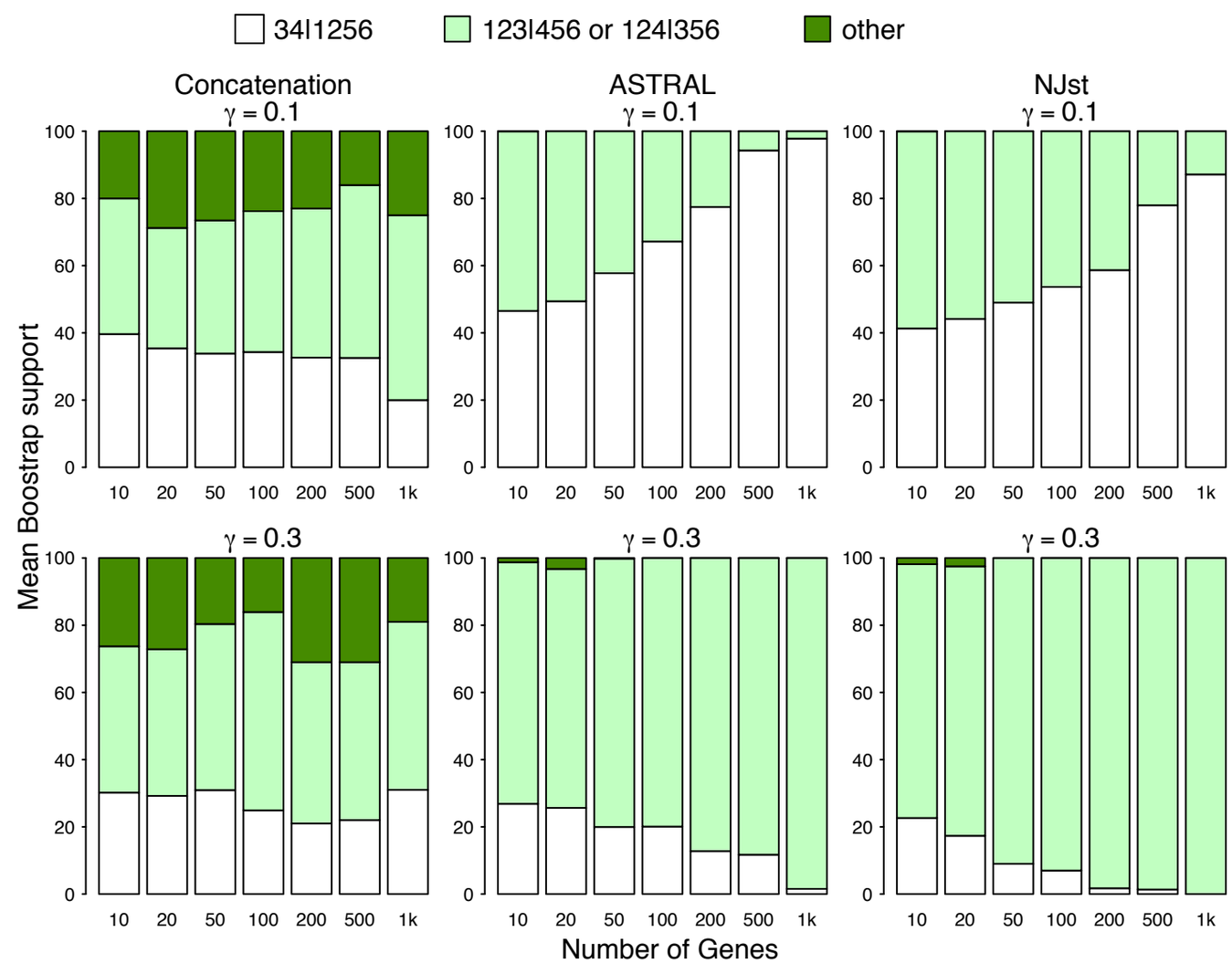
(S-L, Ané, 2016, PLoS Genetics)

Why networks?



Inconsistency with gene flow





(S-L, Yang, Ané, 2016, Syst Bio)

PhyloNetworks: analysis for phylogenetic networks in Julia

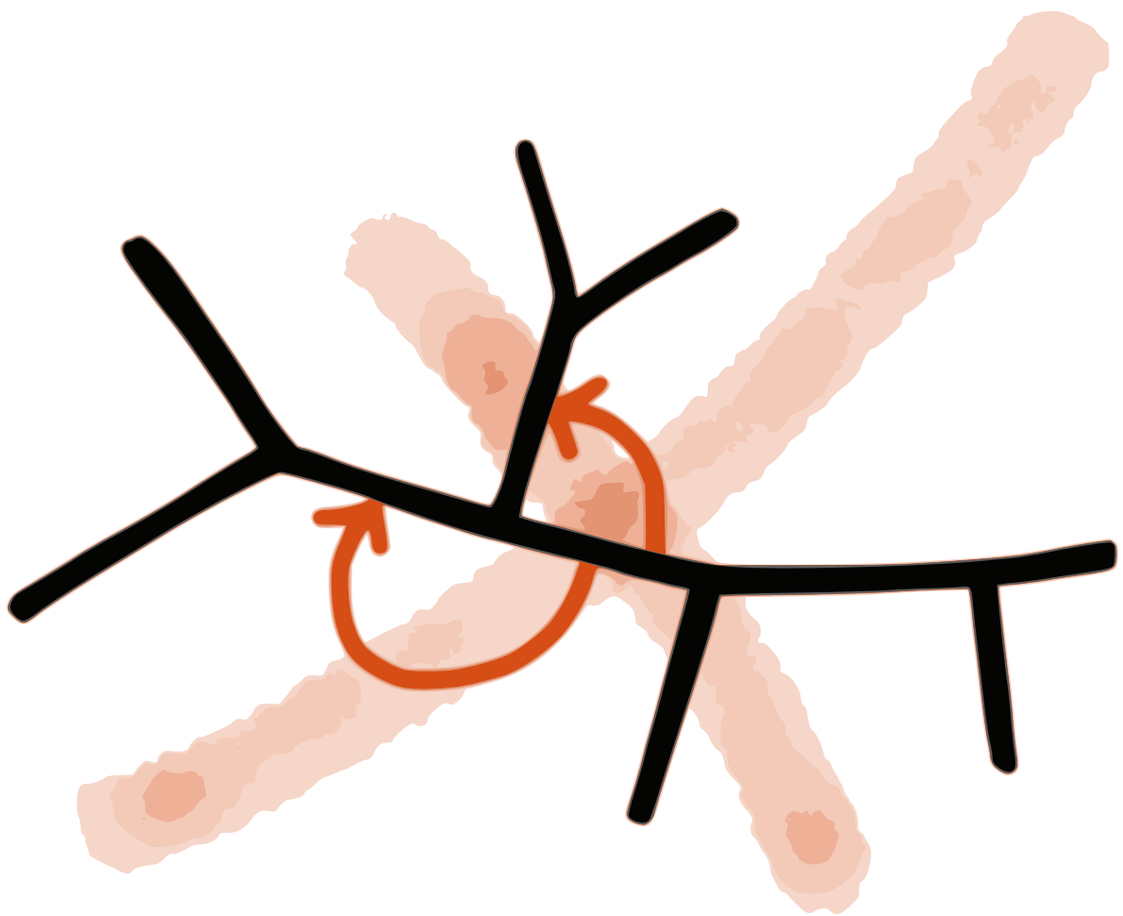
Maximum pseudolikelihood estimation of species network: SNaQ

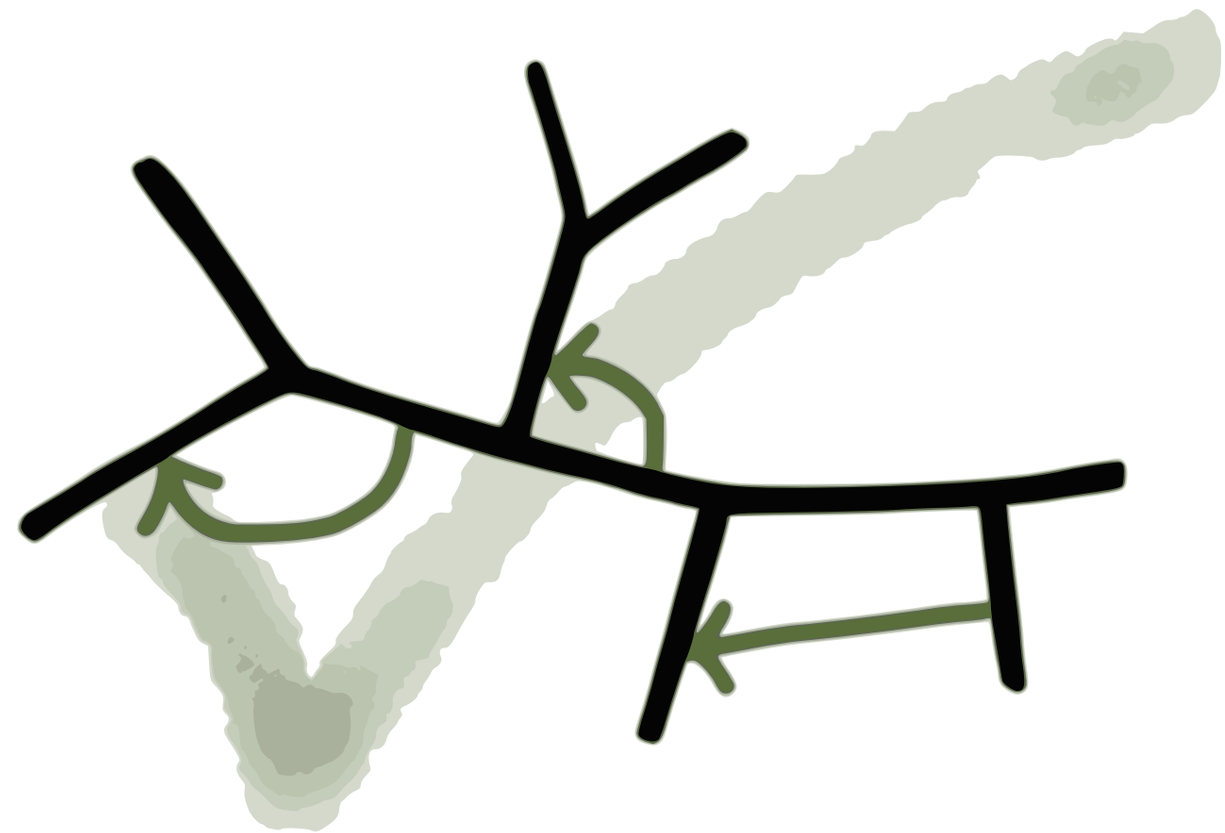
build passing | docs stable | docs latest

SNaQ implements the statistical inference method in Solís-Lemus and Ané (2016, PLoS Genetics). The procedure involves a numerical optimization of branch lengths and inheritance probabilities and a heuristic search in the space of phylogenetic networks.

http://crsl4.github.io/

Level-1 networks





What we have:

- scalable method for **level-1** networks from multilocus data

What we want:

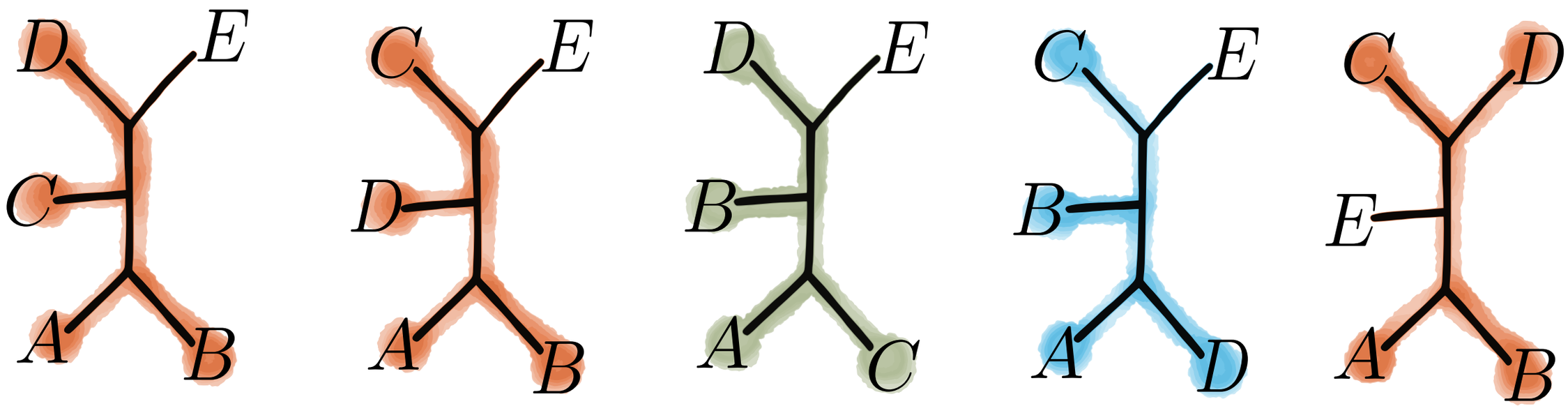
- **level-k** networks: identifiability
- better optimization tools in space of networks
- model selection tools

Acknowledgements

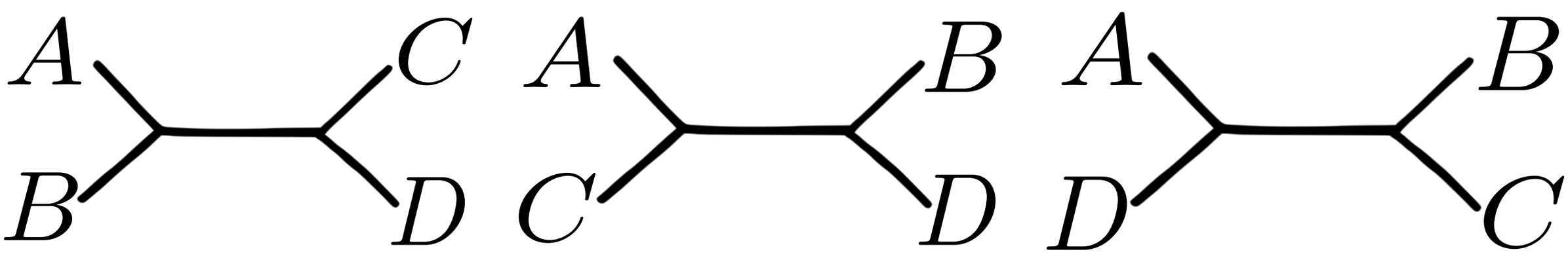
Cécile Ané
Bret Larget
Douglas Bates
David Baum
Mengyao Yang
John Malloy
John Spaw
Noah Stenz
Nan Ji
Jordan Vonderwell
Josh McGrath

http://crsl4.github.io/
claudia@stat.wisc.edu

Genes → Gene trees → Quartet CF

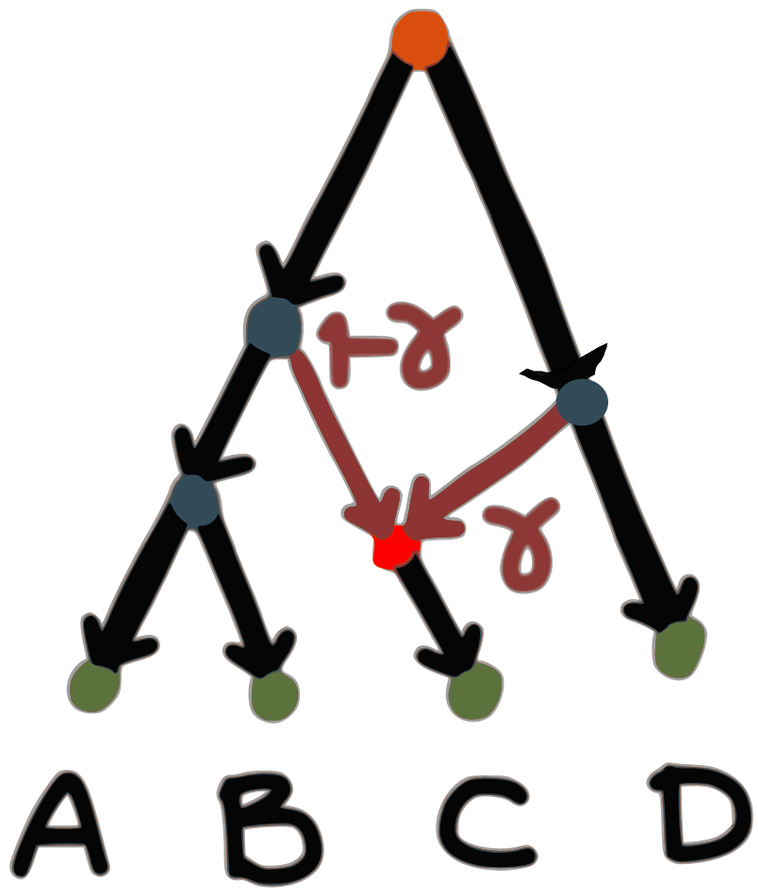


concordance factors (CF):
% of genes having the quartet in their tree

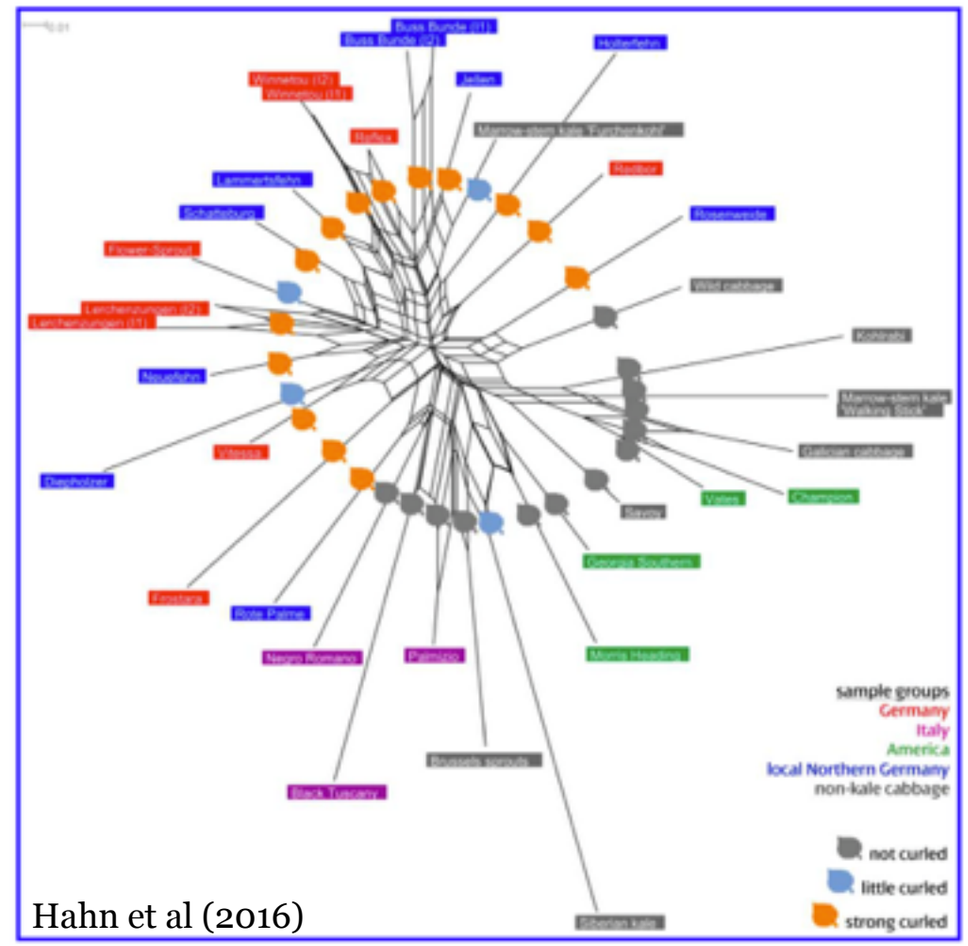


3/5 **1/5** **1/5**

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



Explicit



Hahn et al (2016)

Implicit

no distinction: ILS, HGT

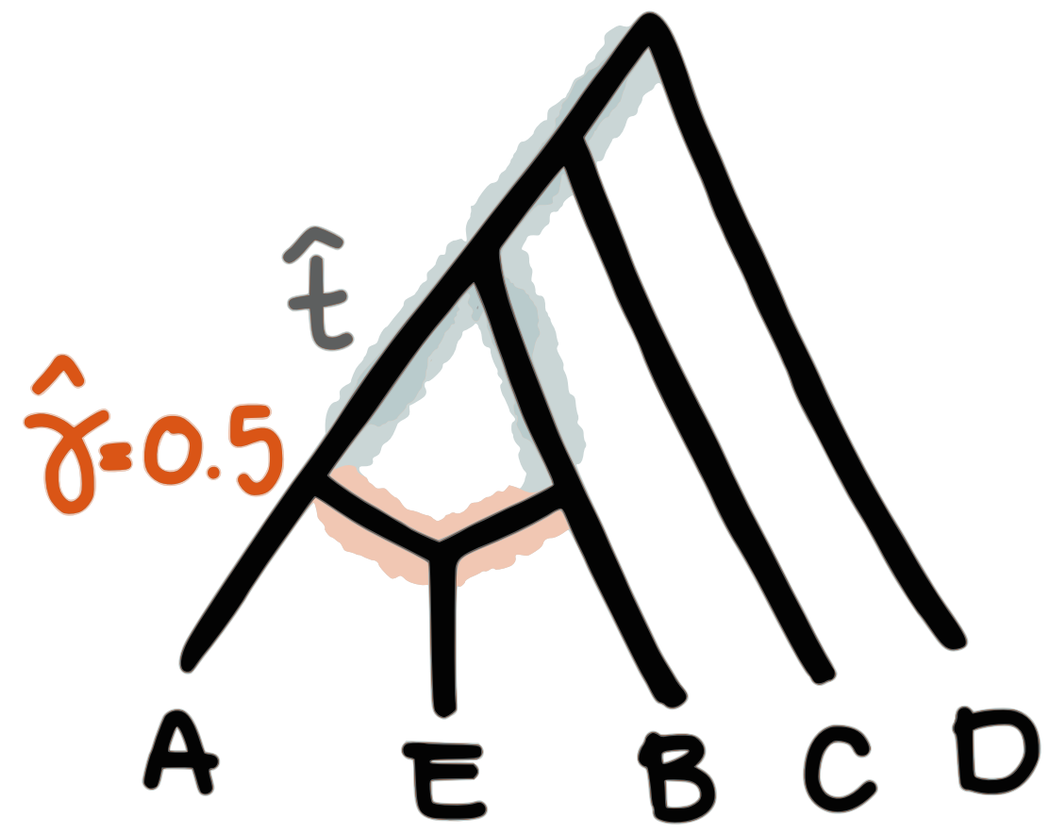
Reasons for gene tree discordance

- Gene tree reconstruction error
- Horizontal gene transfer (HGT)
- Incomplete lineage sorting (ILS)

Observed CF

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

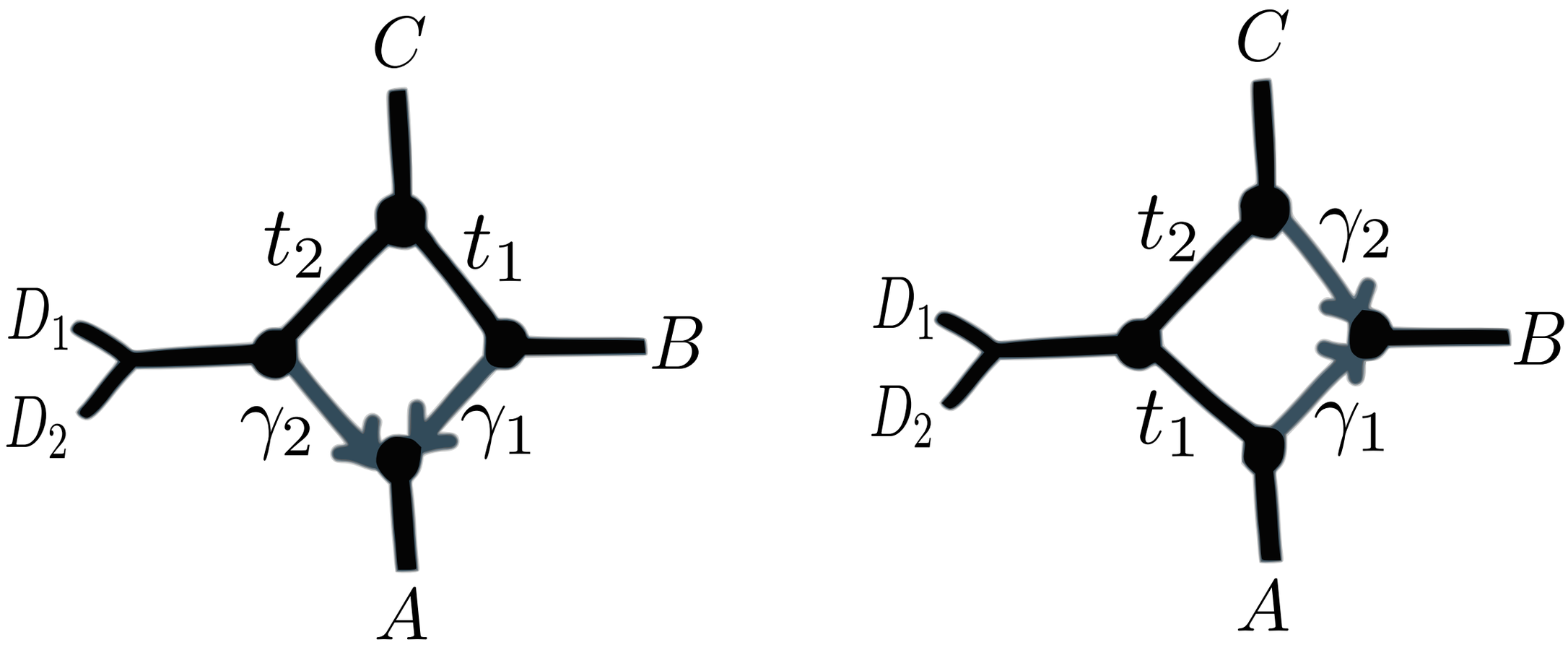
Expected CF



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

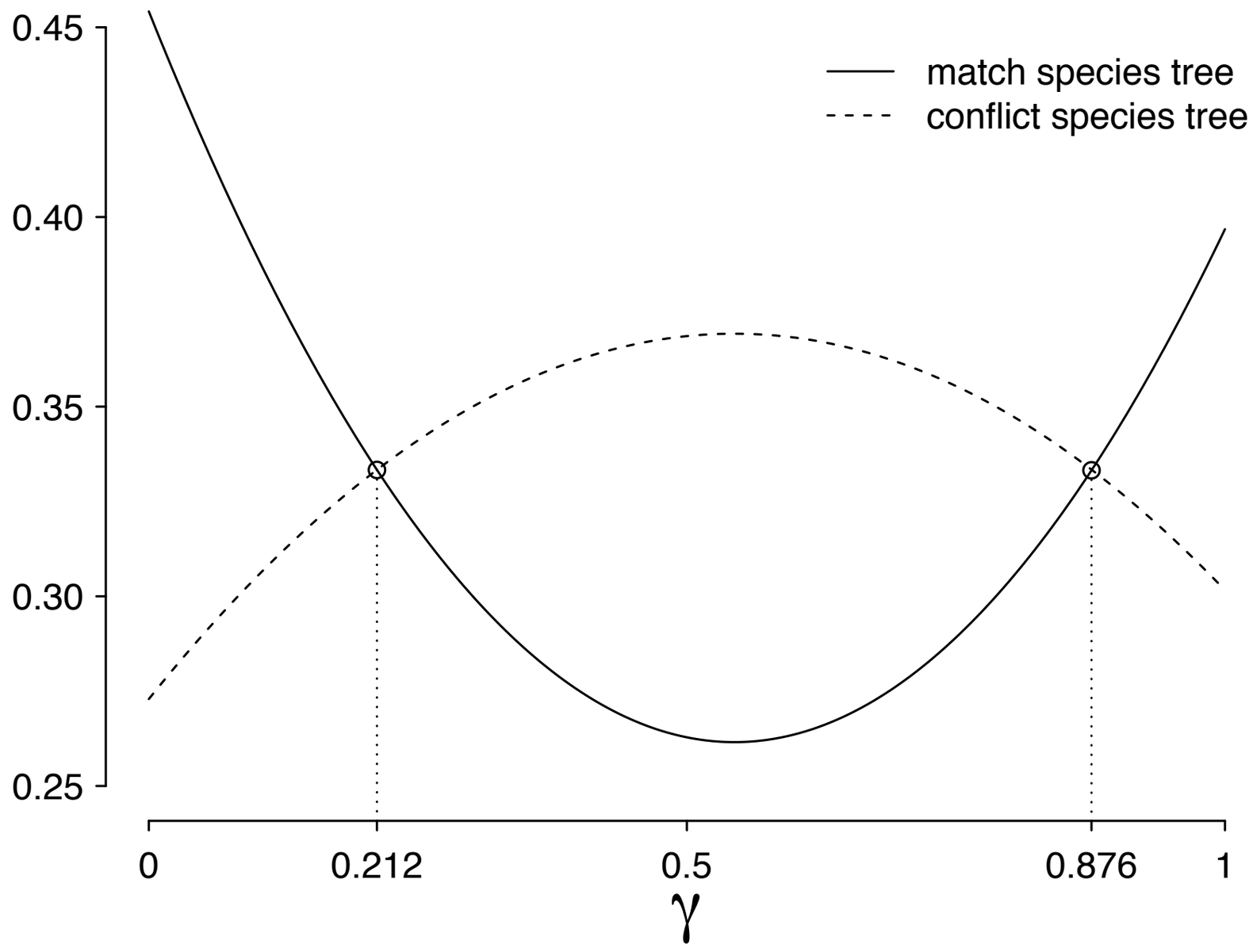
(Solís-Lemus, Ané, 2016, PLoS Genetics)

In practice: flat pseudolikelihood



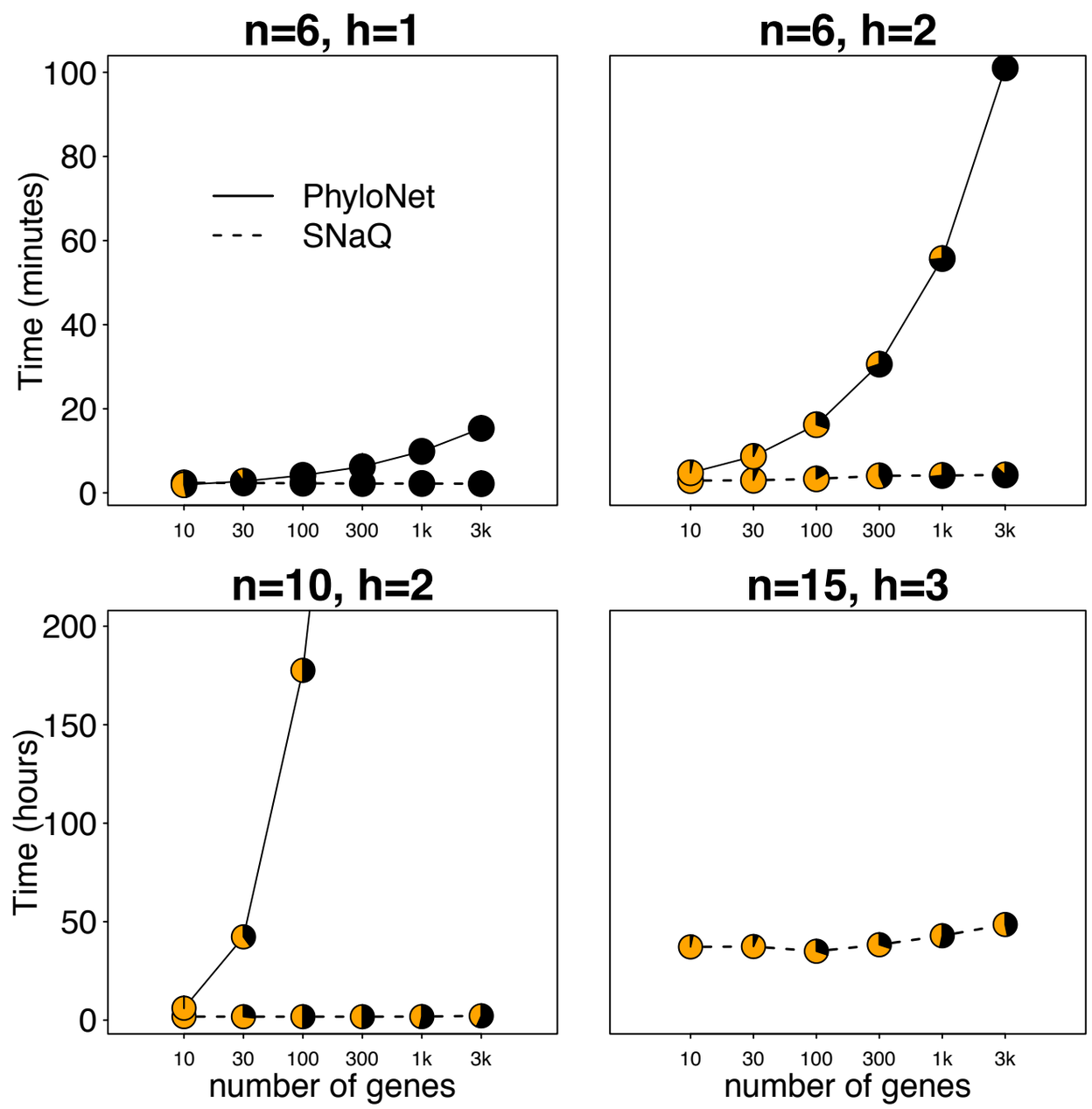
(Solís-Lemus, Ané, 2016, PLoS Genetics)

Anomaly zone with gene flow



(Solís-Lemus, Yang, Ané, 2016, Syst Bio)

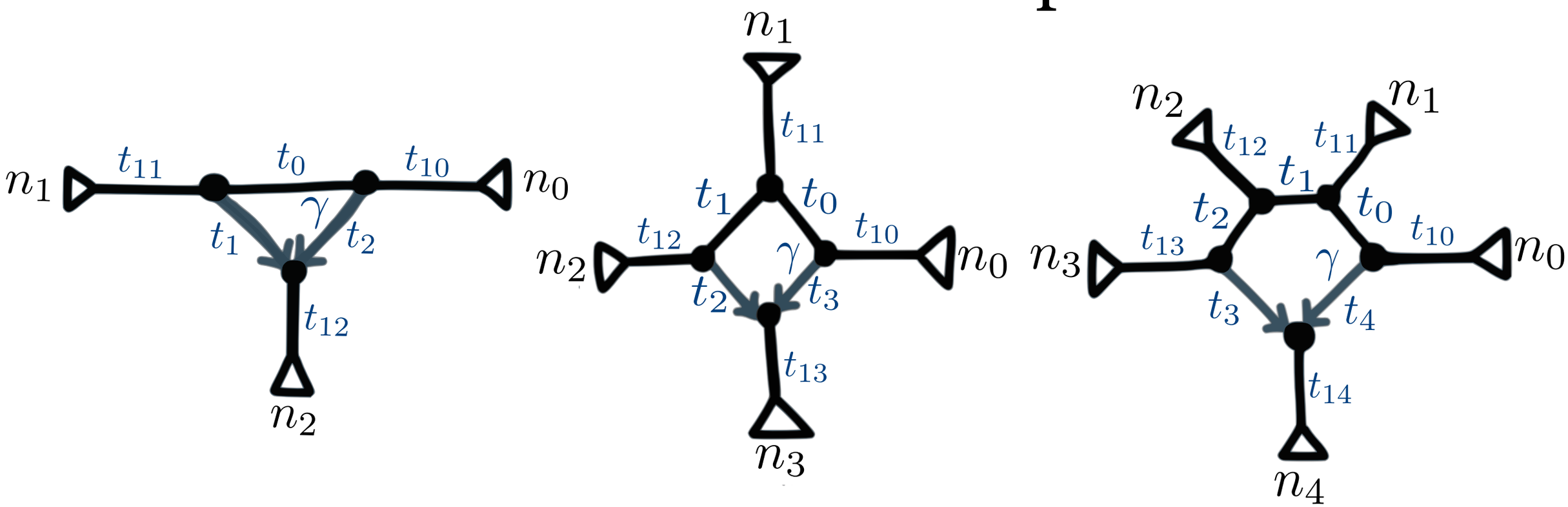
SNaQ performance



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

Model identifiability

Can we estimate numerical parameters?



No

Good triangle

$(t_{12} = 0)$

Yes

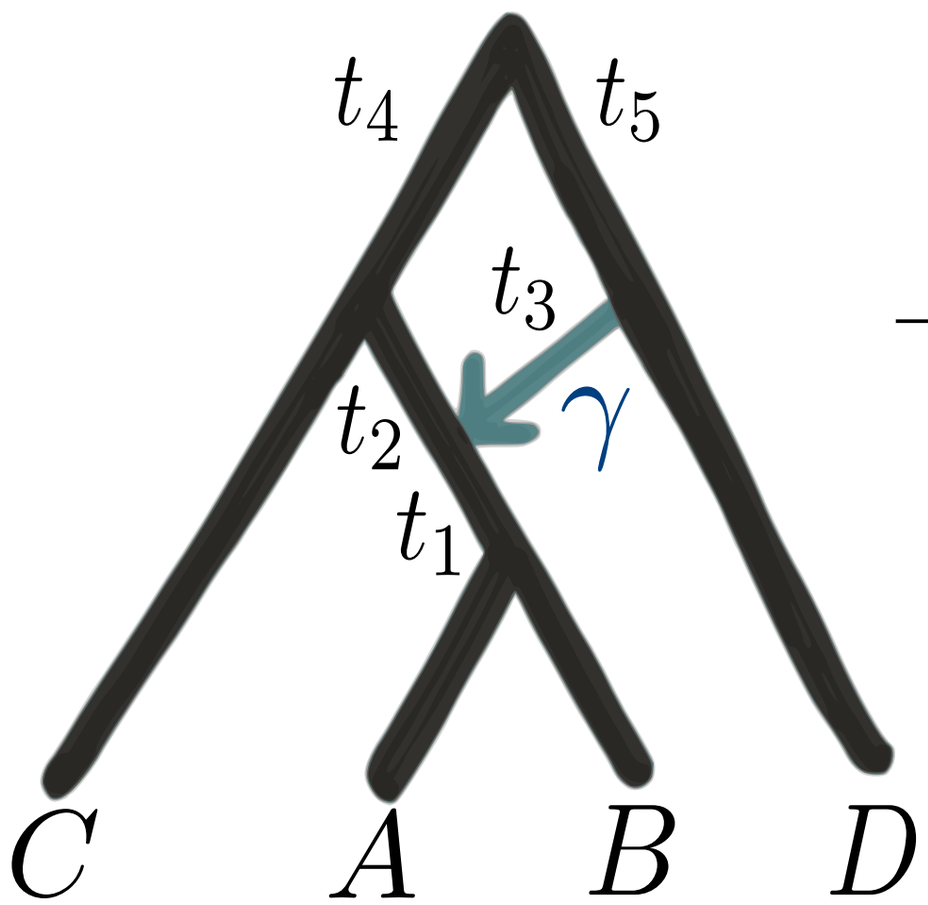
Good diamond

$(n_0, n_2 \geq 2)$

Yes

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

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

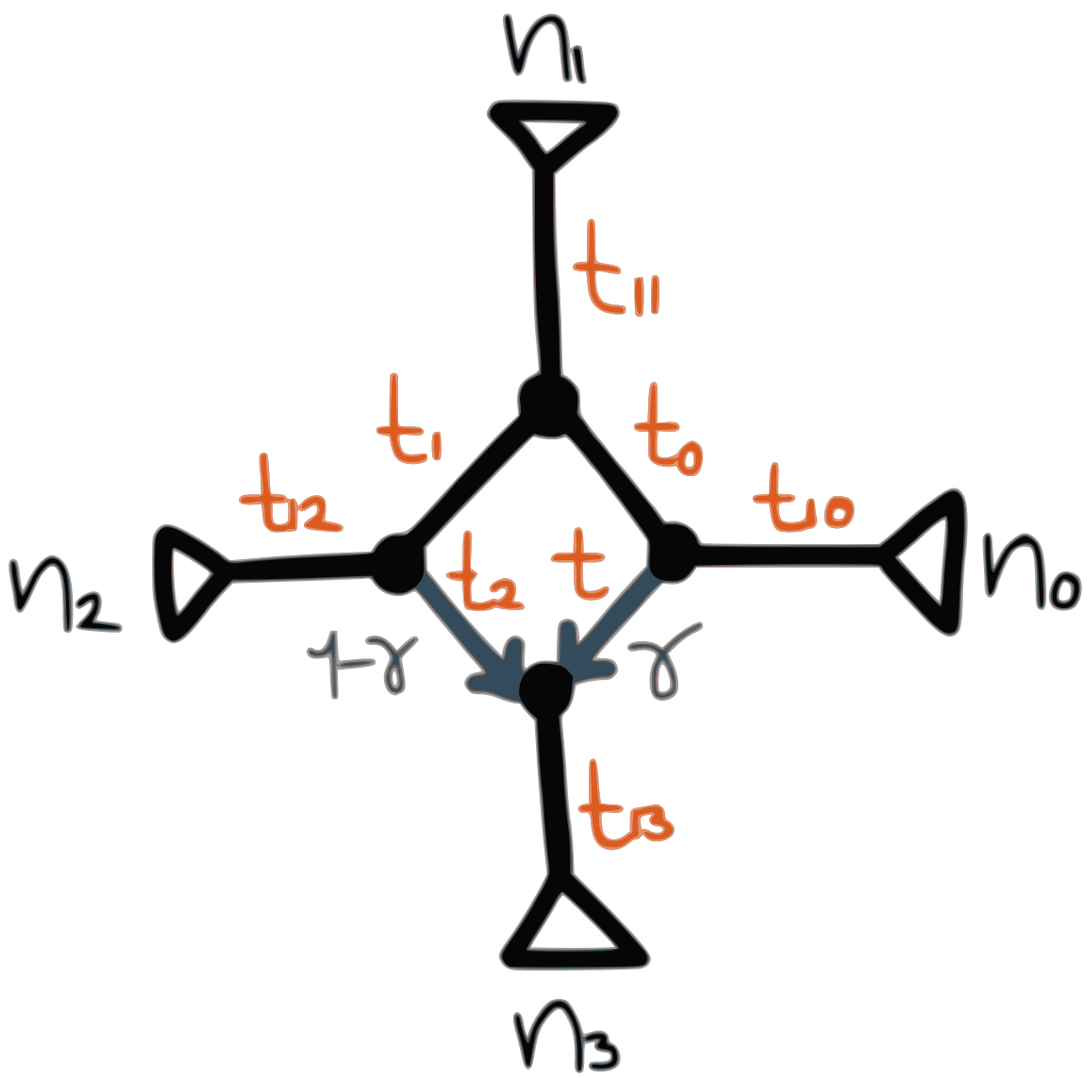
$t_1 = t_2 = 0.01, t_3 = t_4 = t_5 = 1$

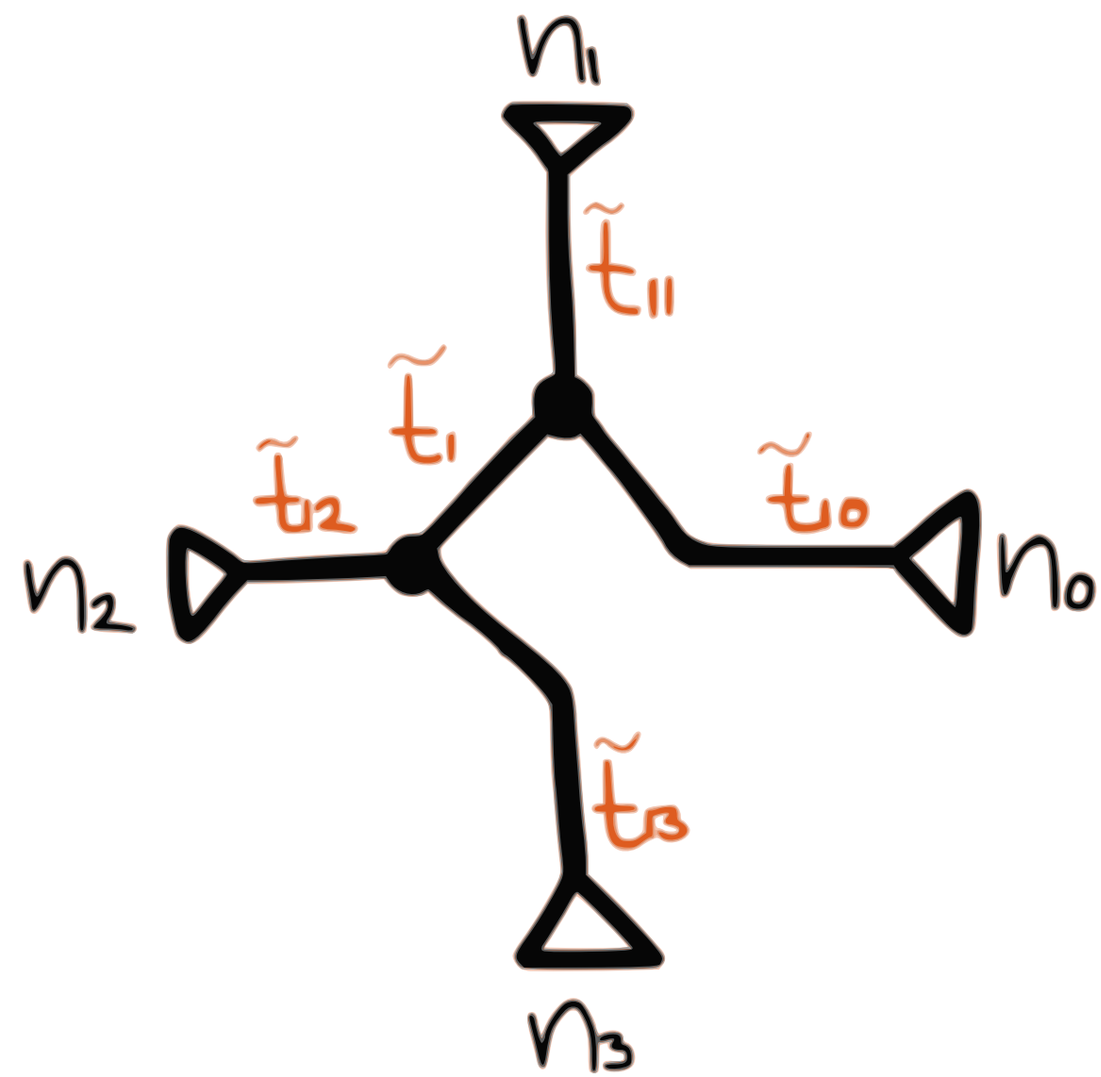
- **ILS**: no AUGT on 4 taxa (Degnan, 2013)
- **ILS**+**HGT**: AUGT on 4 taxa (Solís-Lemus, Yang, Ané, 2016, Syst Bio)

Why networks?

	Concatenation	Coalescent Tree	Coalescent Network
HGT	✗	✗	✓
ILS	✗	✓	✓

Idea of proof of identifiability: hybridization





System of equations

$\{CF_{network}\}$

System of equations

$\{CF_{tree}\}$

Idea of proof of identifiability: hybridization

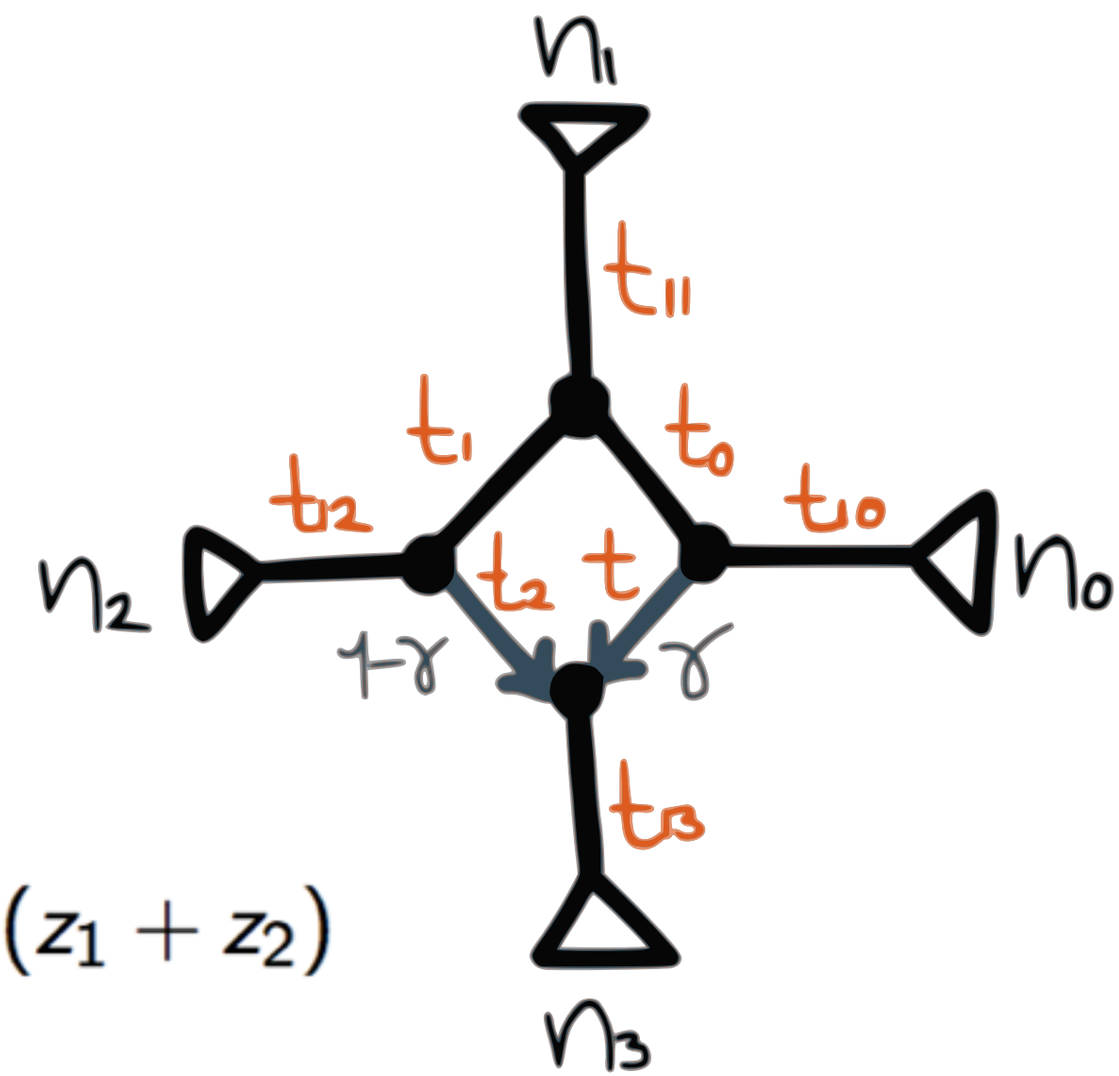
Solution to $CF_{network} = CF_{tree}$ if

$\gamma = 0$

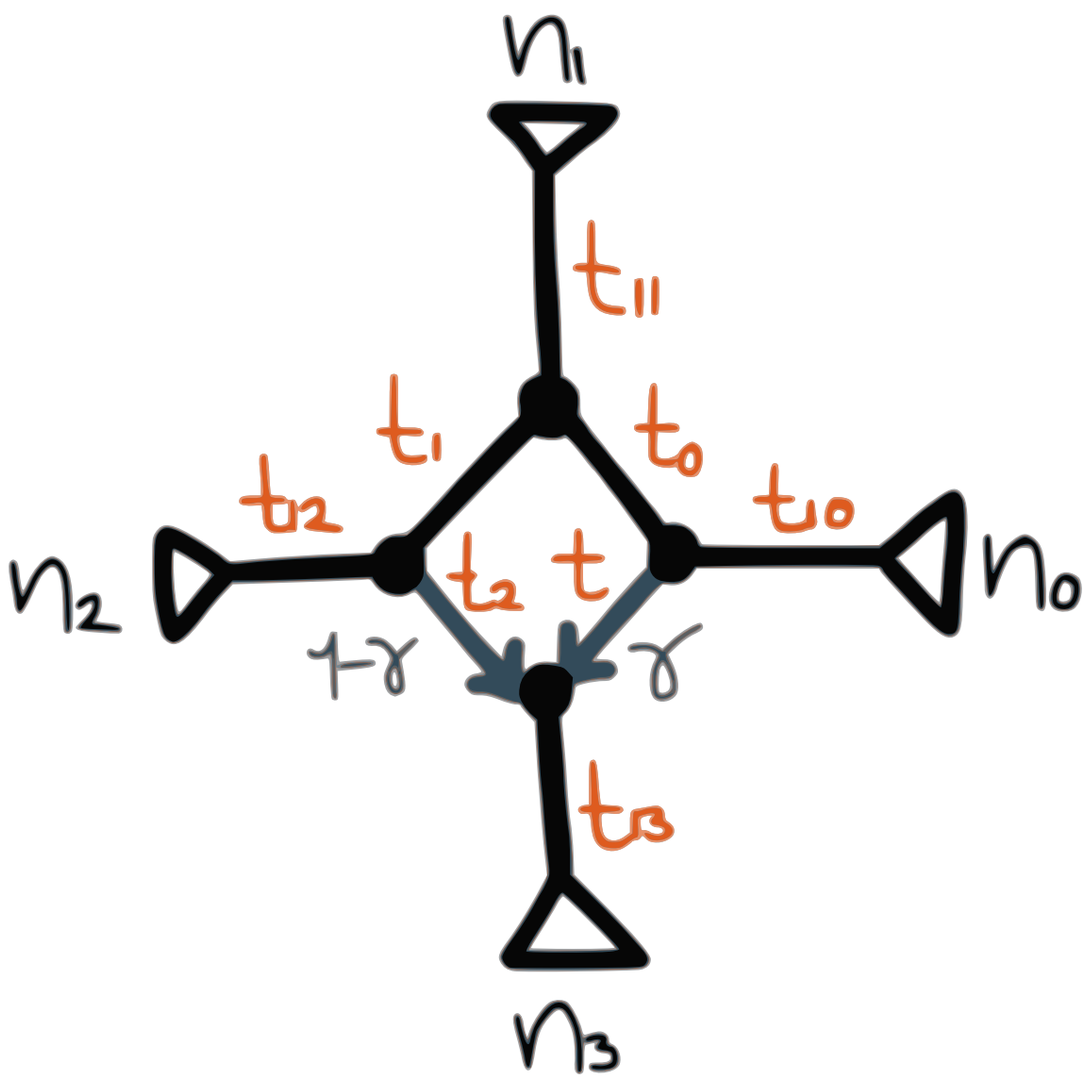
$t_0 = 0$

$\gamma = 1$
$t_{12} = \infty$
$t_1 = 0$

$t_{13} = \infty$
$\gamma z = (1-\gamma)(z_1 + z_2)$



Idea of proof of identifiability: parameters



Unique solution: hard

Finitely many solutions:
alg. indep. eqs ~ # parameters

System of equations

$\{CF_{network}\}$

Coalescent model

- Haploid population: constant size N
- 1 individual = 1 chromosome
- No selection: uniform probability
- Probability of no coalescence in g generations:

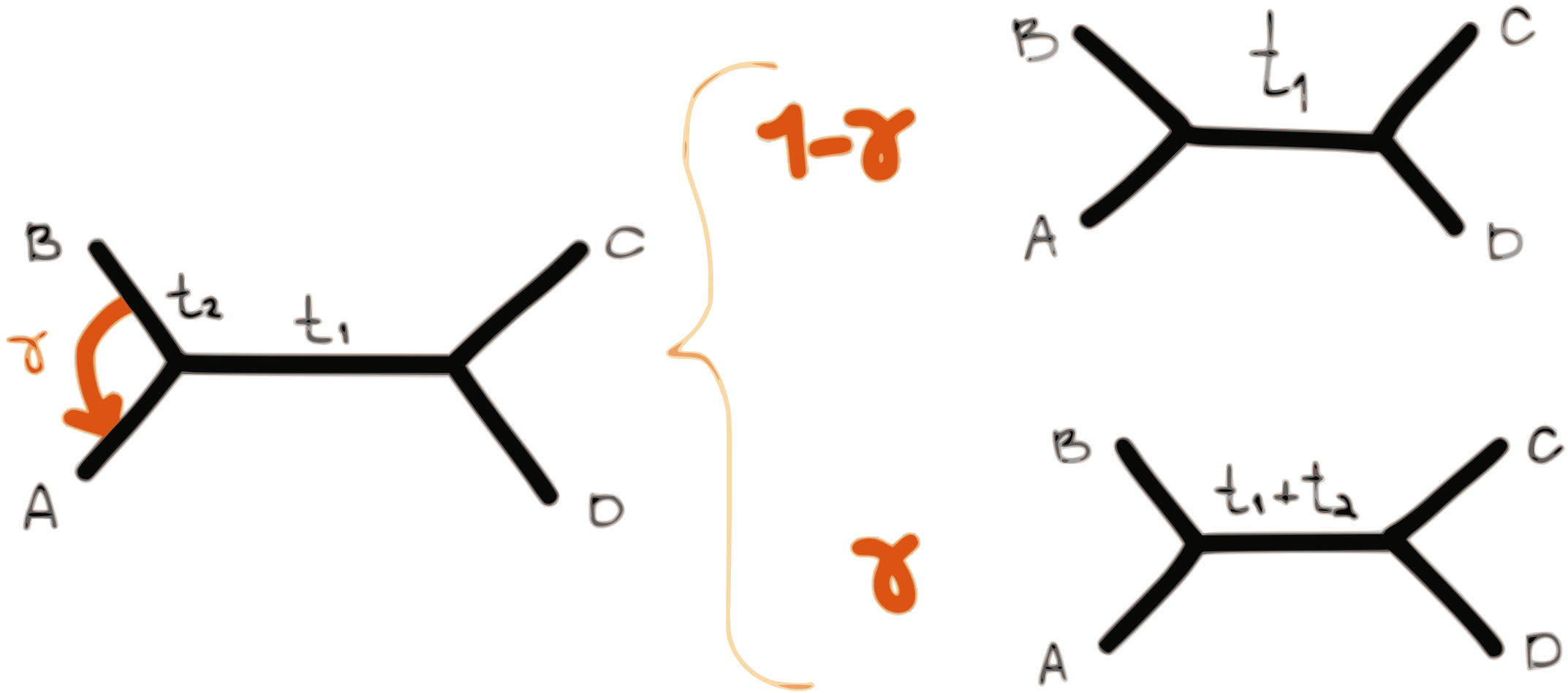
$$\left(1 - \frac{1}{N}\right)^g$$

- Coalescence time $t = \frac{g}{N}$

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

- Exponential distribution with mean 1

Computing expected CF



$$CF_{AB|CD} = (1-\gamma)(1-2/3e^{-t_1}) + \gamma(1-2/3e^{-t_1-t_2})$$
$$CF_{AC|BD} = CF_{AD|BC} = (1-\gamma)(1/3e^{-t_1}) + \gamma(1/3e^{-t_1-t_2})$$