Modeling the Evolution of Genes and Genomes in the Presence of ILS and Hybridization

Luay Nakhleh
Department of Computer Science
Rice University

RICE
Unconventional Wisdom

New Directions in Probabilistic Models of Evolution
University of California, Berkeley
28 April 2014

SIMONS
INSTITUTE
for the Theory of Computing

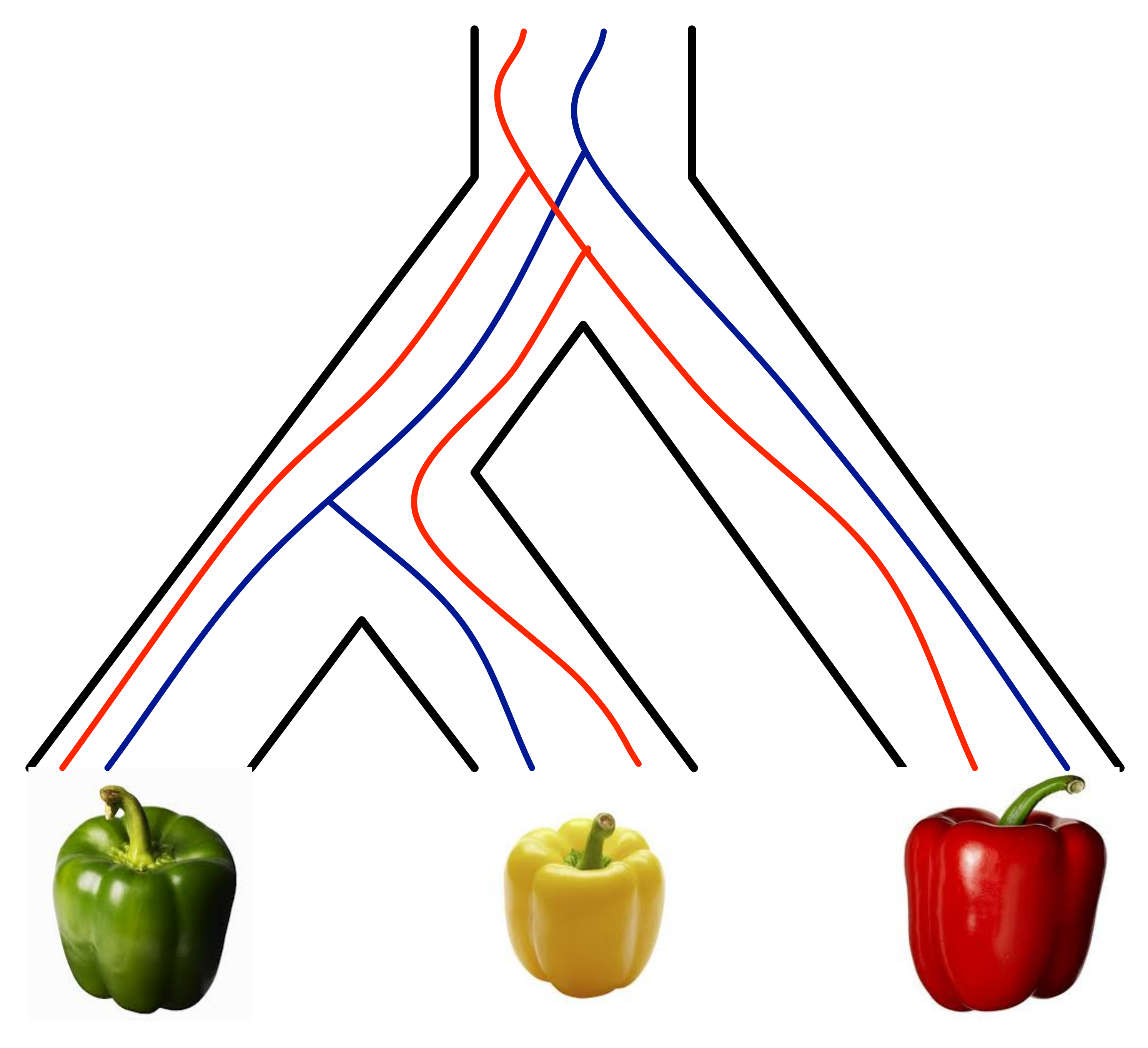
OUTLINE

(1) From gene trees to phylogenetic networks

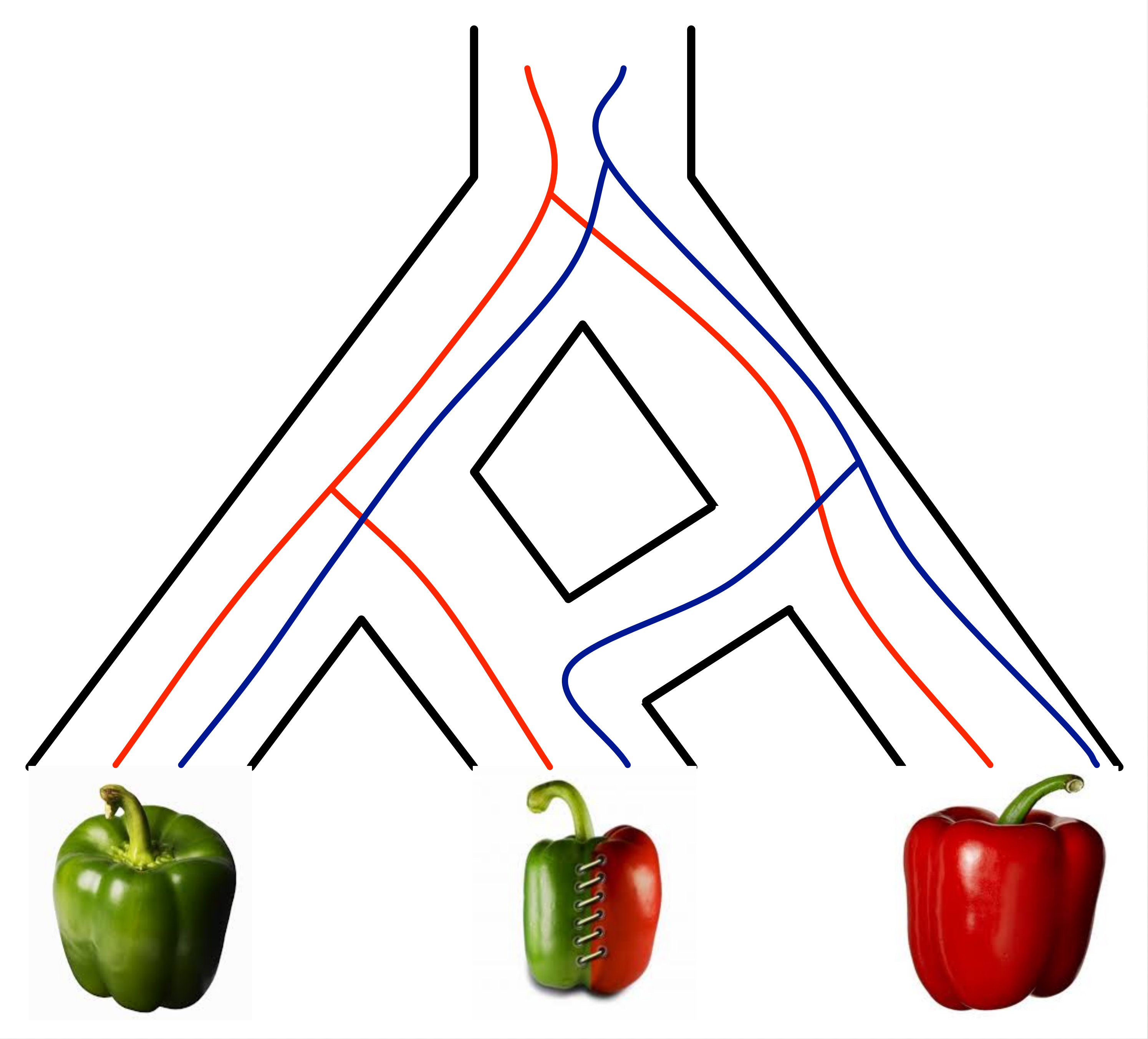
(2) From phylogenetic networks to genome annotation with introgression

From Gene Tree to Phylogenetic Networks

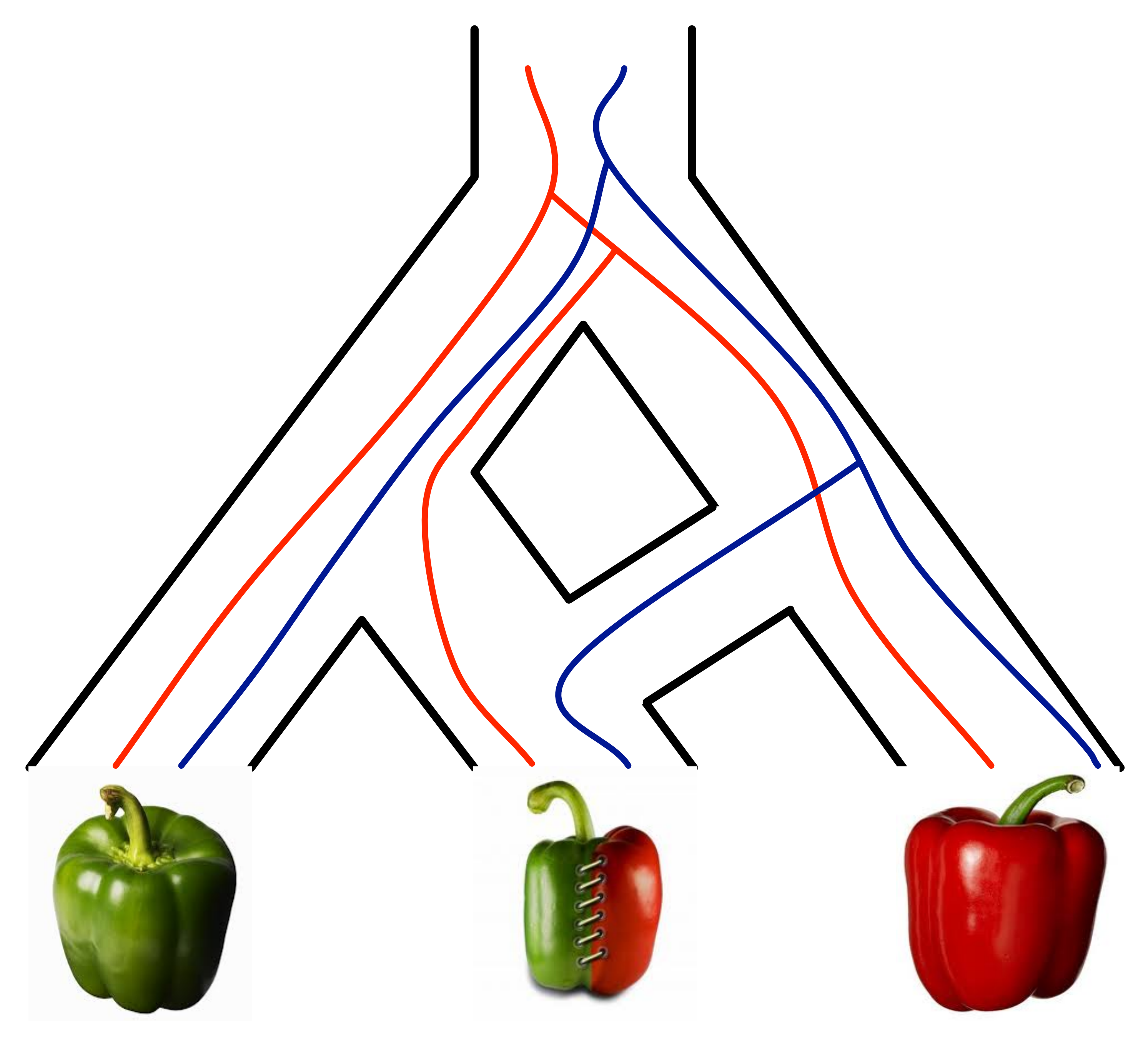
INCOMPLETE LINEAGE SORTING (ILS)

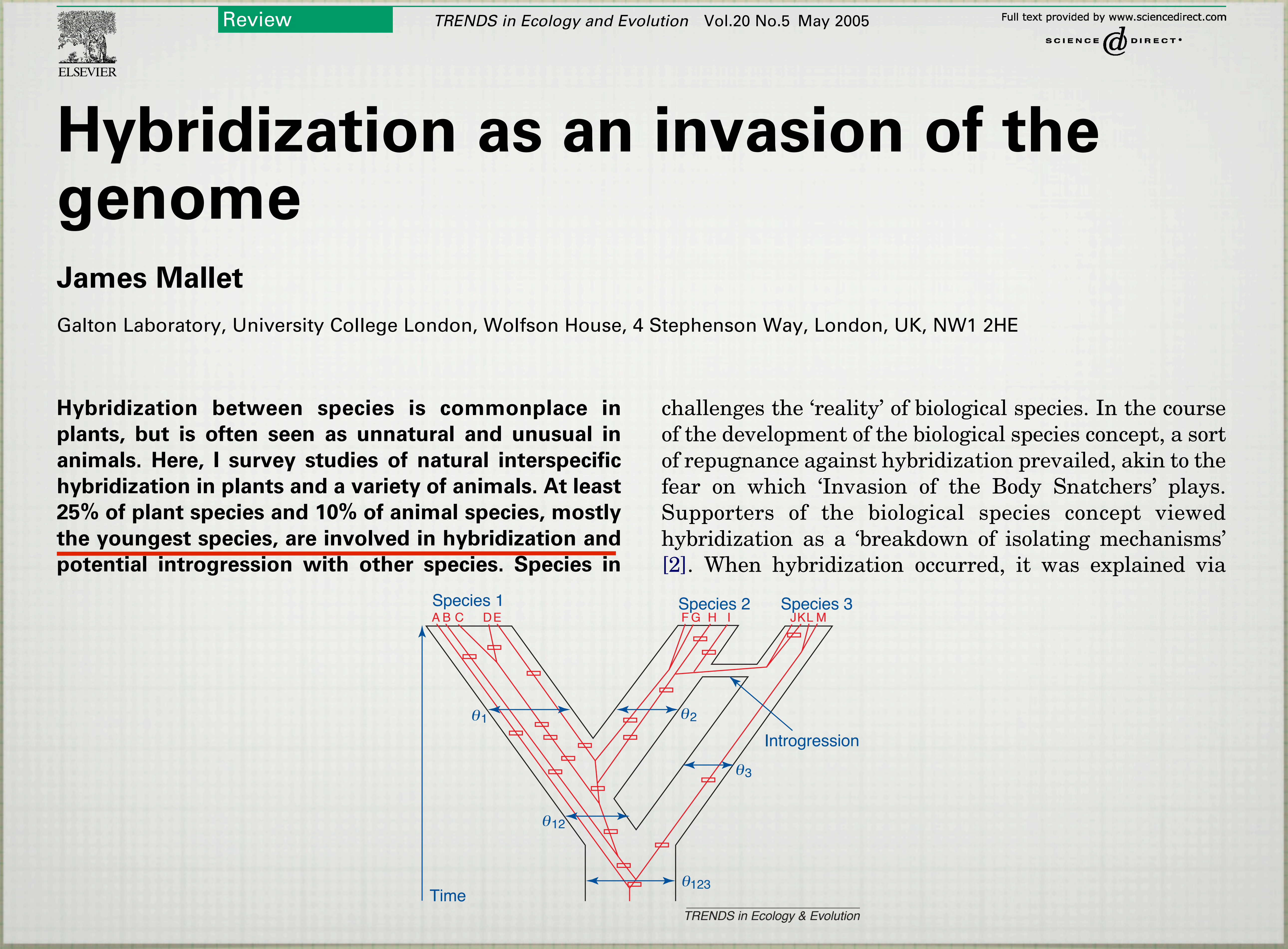


HYBRIDIZATION



ILS + HYBRIDIZATION





ELSEVIER

Review

TRENDS in Ecology and Evolution Vol.20 No.5 May 2005

Full text provided by www.sciencedirect.com

SCIENCE DIRECT

Hybridization as an invasion of the genome

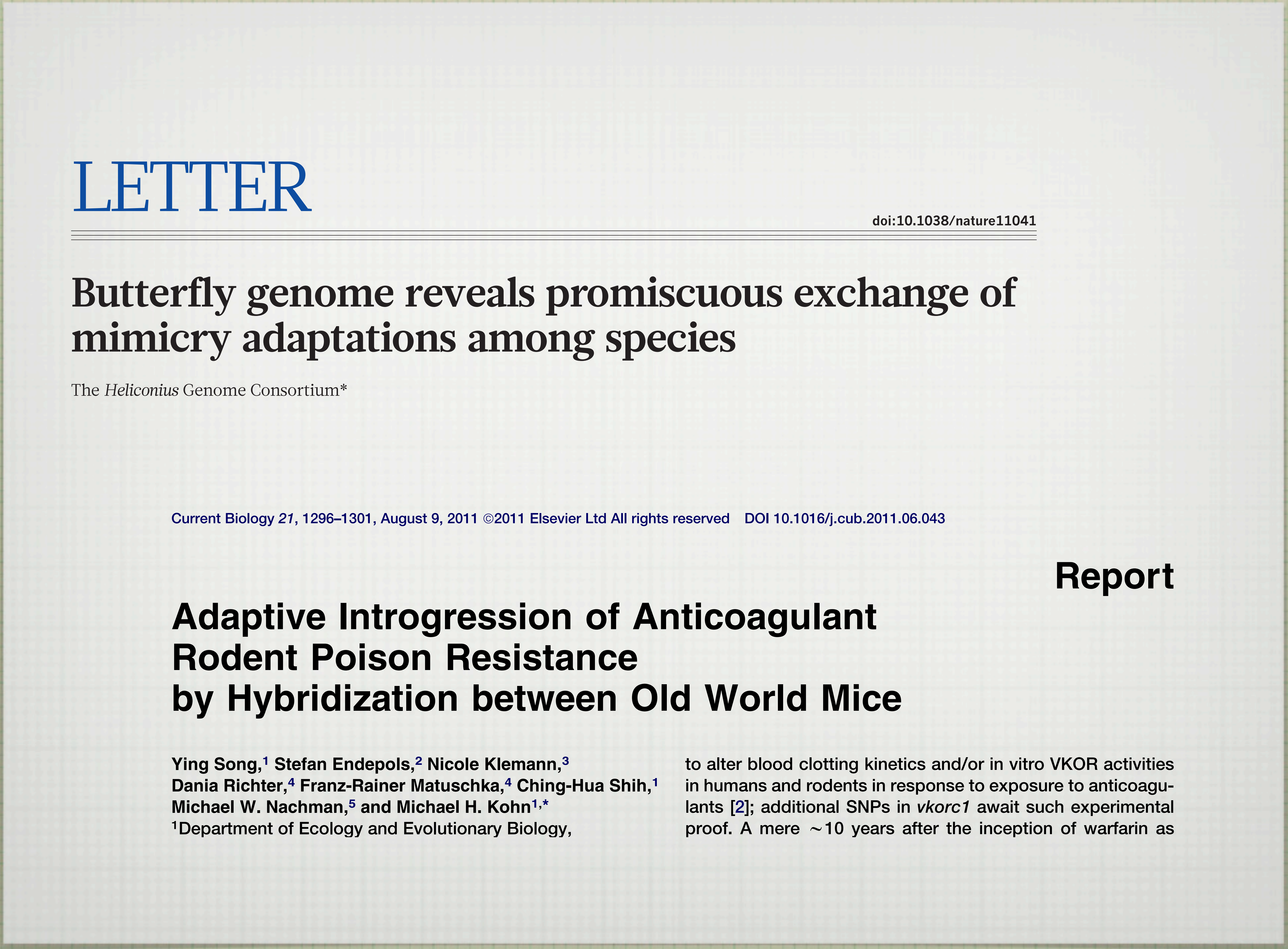
James Mallet

Galton Laboratory, University College London, Wolfson House, 4 Stephenson Way, London, UK, NW1 2HE

Hybridization between species is commonplace in plants, but is often seen as unnatural and unusual in animals. Here, I survey studies of natural interspecific hybridization in plants and a variety of animals. At least 25% of plant species and 10% of animal species, mostly the youngest species, are involved in hybridization and potential introgression with other species. Species in

challenges the 'reality' of biological species. In the course of the development of the biological species concept, a sort of repugnance against hybridization prevailed, akin to the fear on which 'Invasion of the Body Snatchers' plays. Supporters of the biological species concept viewed hybridization as a 'breakdown of isolating mechanisms' [2]. When hybridization occurred, it was explained via

TRENDS in Ecology & Evolution



LETTER

doi:10.1038/nature11041

Butterfly genome reveals promiscuous exchange of mimicry adaptations among species

The *Heliconius* Genome Consortium*

Current Biology *21*, 1296–1301, August 9, 2011 ©2011 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2011.06.043

Report

Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice

Ying Song,[1] Stefan Endepols,[2] Nicole Klemann,[3] Dania Richter,[4] Franz-Rainer Matuschka,[4] Ching-Hua Shih,[1] Michael W. Nachman,[5] and Michael H. Kohn[1],*
[1]Department of Ecology and Evolutionary Biology,

to alter blood clotting kinetics and/or in vitro VKOR activities in humans and rodents in response to exposure to anticoagulants [2]; additional SNPs in *vkorc1* await such experimental proof. A mere ~10 years after the inception of warfarin as

A MAXIMUM LIKELIHOOD APPROACH

$$L(\Psi|\mathcal{S}) = \prod_{S \in \mathcal{S}} \left[\sum_{T} \left[\mathbf{P}(S|T) \cdot \mathbf{P}(T|\Psi) \right] \right]$$

species phylogeny
and its parameters

sequences of
gene families

If a gene tree has been inferred for each gene family, then:

$$L(\Psi|\mathcal{G}) = c \cdot \prod_{gt \in \mathcal{G}} \mathbf{P}(gt|\Psi)$$

A MAXIMUM LIKELIHOOD APPROACH

$$L(\Psi|\mathcal{S}) = \prod_{S\in\mathcal{S}} \left[\sum_{T} [\mathbf{P}(S|T) \cdot \mathbf{P}(T|\Psi)] \right]$$

species phylogeny
and its parameters

sequences of
gene families

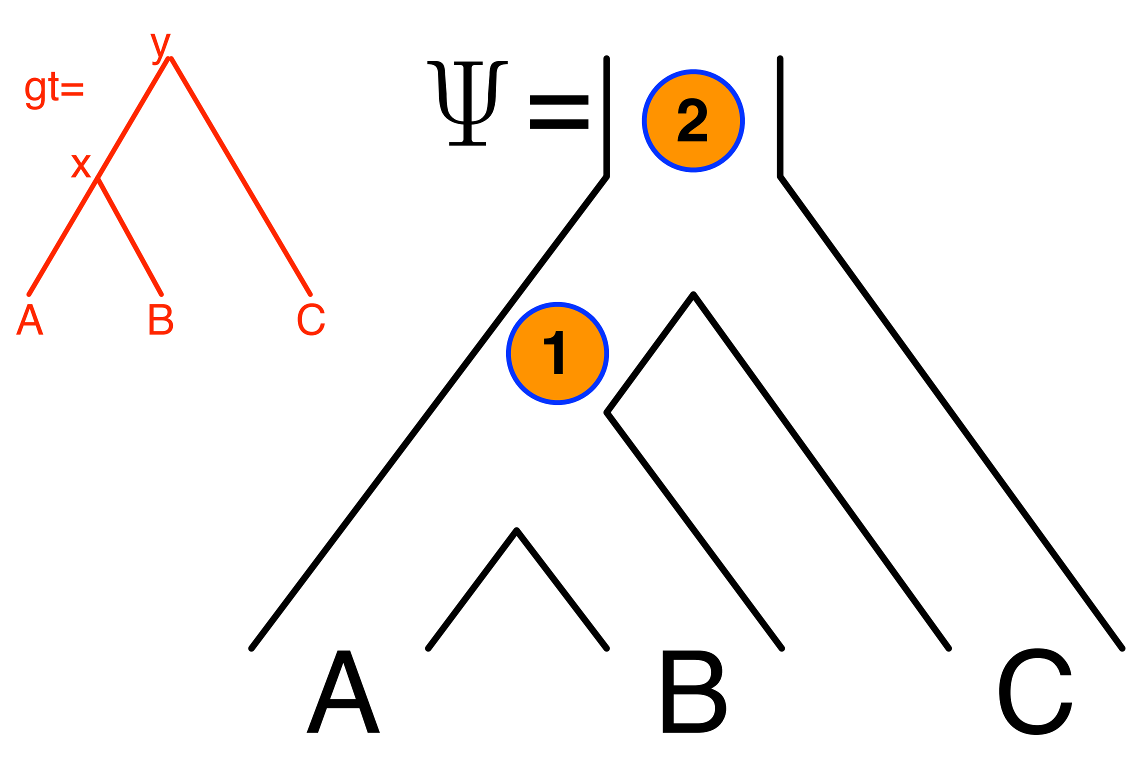
If a gene tree has been inferred for each gene family, then:

$$L(\Psi|\mathcal{G}) = c \cdot \prod_{gt\in\mathcal{G}} \mathbf{P}(gt|\Psi)$$

How do we compute $\mathbf{P}(gt|\Psi)$?

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

- Denote by H_Ψ(gt) the set of all coalescent histories of species tree Ψ and gene tree topology gt



$$H_\Psi(gt) = \{(1,2),(2,2)\}$$

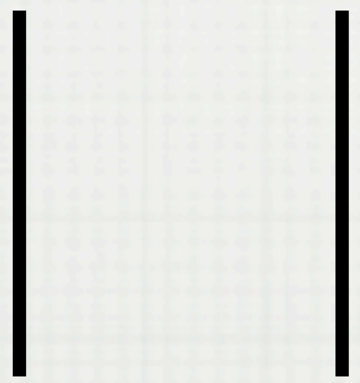
$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

□ Degnan and Salter (Evolution, 2005) gave the mass probability function of a gene tree topology gt for a given species tree with topology Ψ and vector of branch lengths **λ:**

$$P_{\Psi,\lambda}(gt) = \sum_{h \in H_\Psi(gt)} \frac{w(h)}{d(h)} \prod_{b=1}^{n-2} \frac{w_b(h)}{d_b(h)} p_{u_b(h)v_b(h)}(\lambda_b)$$

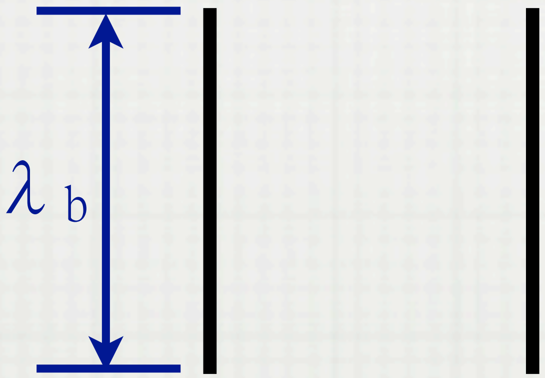
$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b



$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

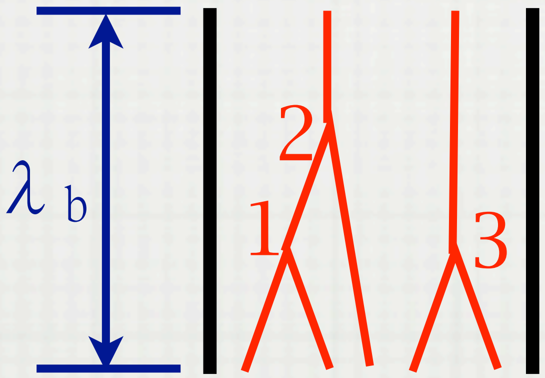
branch b



$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

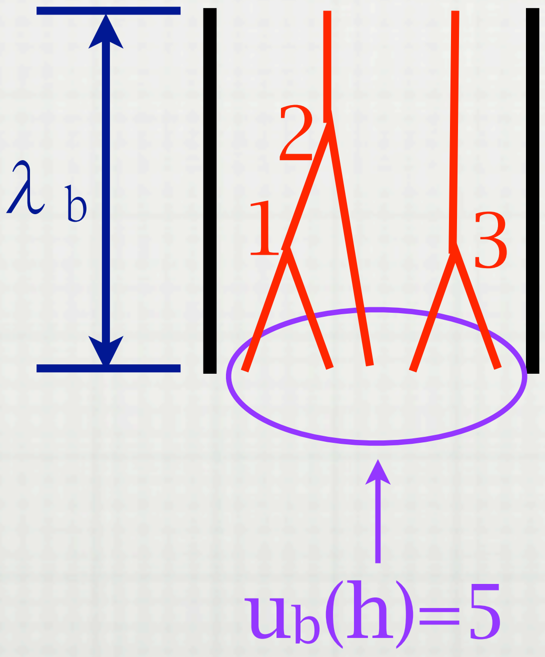
coalescent history h



$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

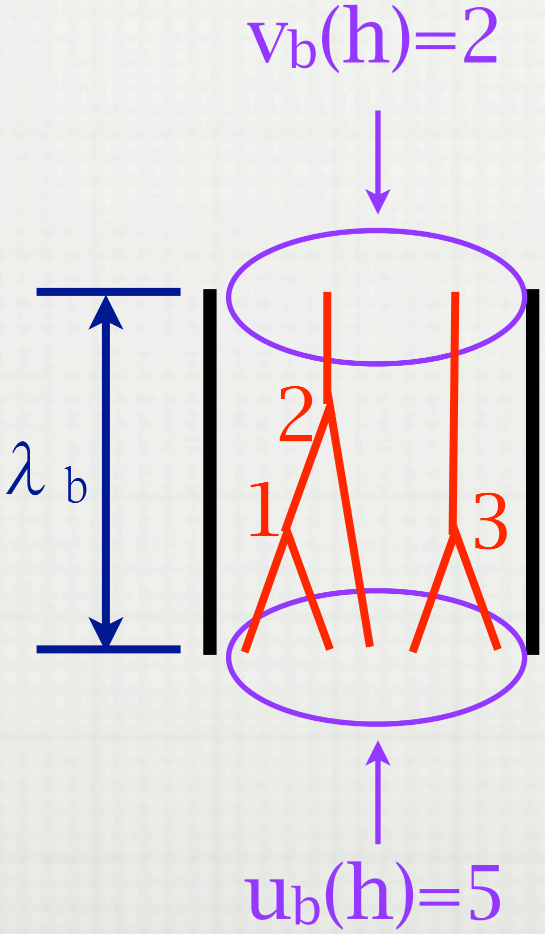
coalescent history h



$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

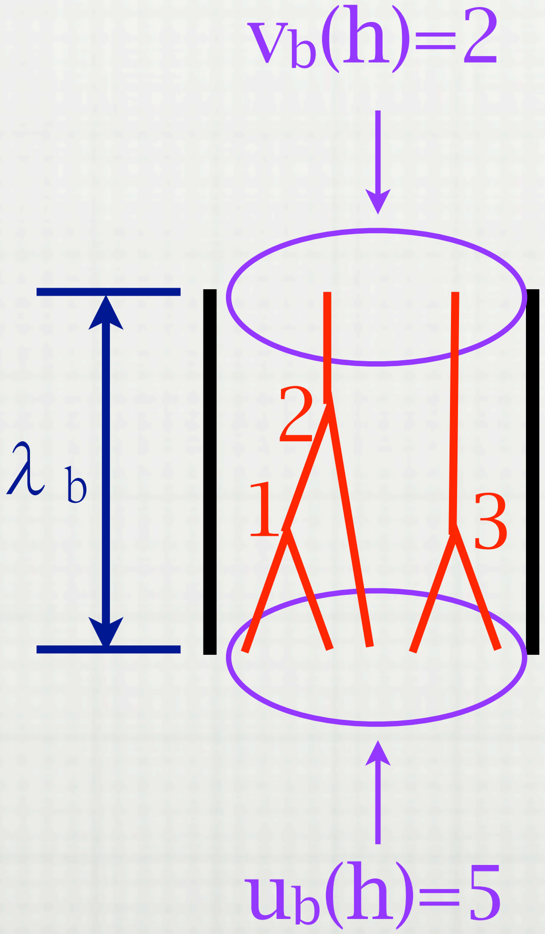
coalescent history h



$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

coalescent history h



$v_b(h)=2$

λ_b

2

1

3

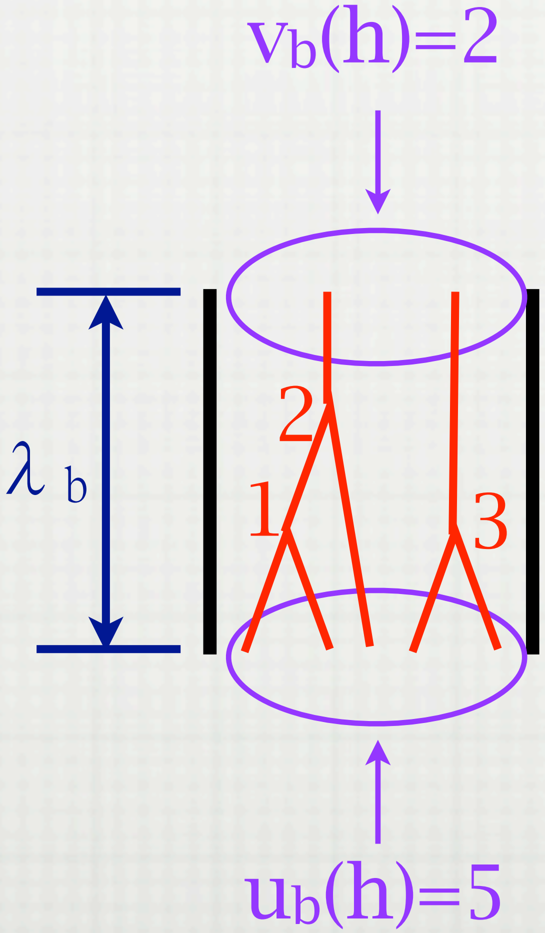
$u_b(h)=5$

$w_b(h)=3$

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

coalescent history h



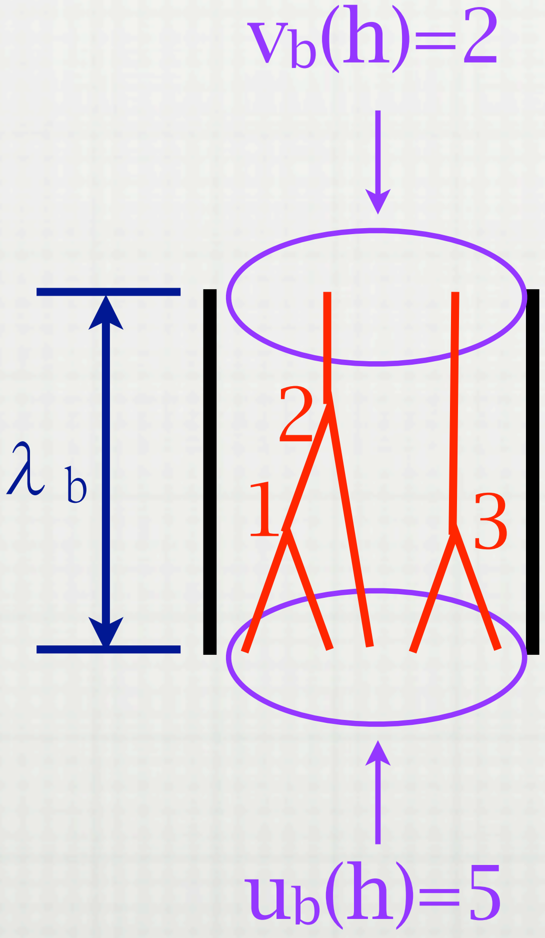
$w_b(h)=3$

3<1<2
1<3<2
1<2<3

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

coalescent history h



$v_b(h)=2$

$u_b(h)=5$

$w_b(h)=3$

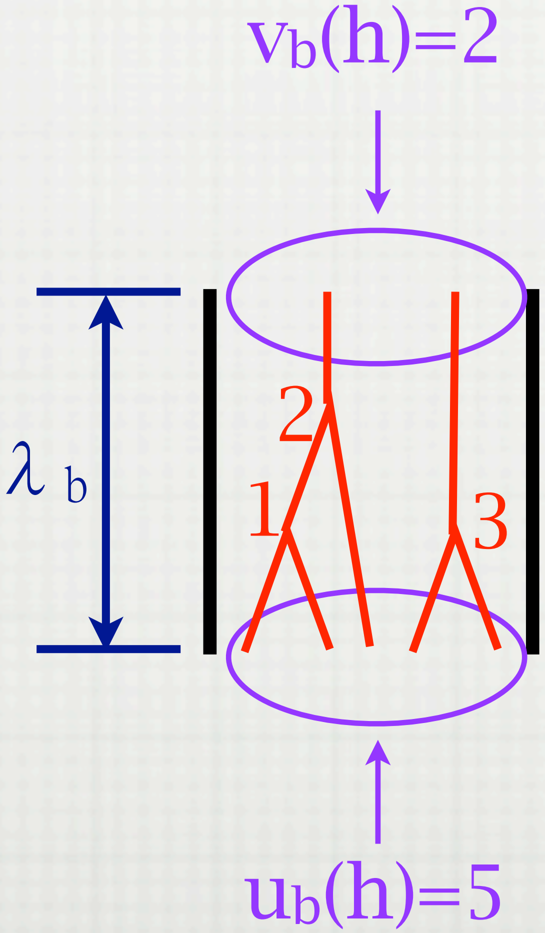
3<1<2
1<3<2
1<2<3

$d_b(h)=180$

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

coalescent history h



$w_b(h)=3$

3<1<2
1<3<2
1<2<3

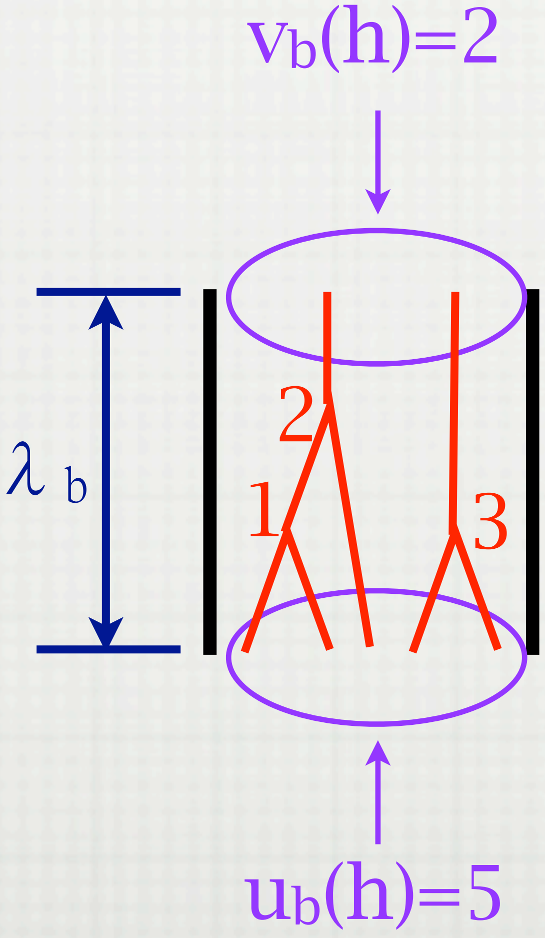
$d_b(h)=180$

$$\binom{5-0}{2}\binom{5-1}{2}\binom{5-2}{2} = 180$$

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

branch b

coalescent history h



$w_b(h)=3$

3<1<2
1<3<2
1<2<3

$$d_b(h) = \prod_{y=0}^{c_b-1} \binom{u_b - y}{2}$$

$d_b(h)=180$

$$\binom{5-0}{2}\binom{5-1}{2}\binom{5-2}{2} = 180$$

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

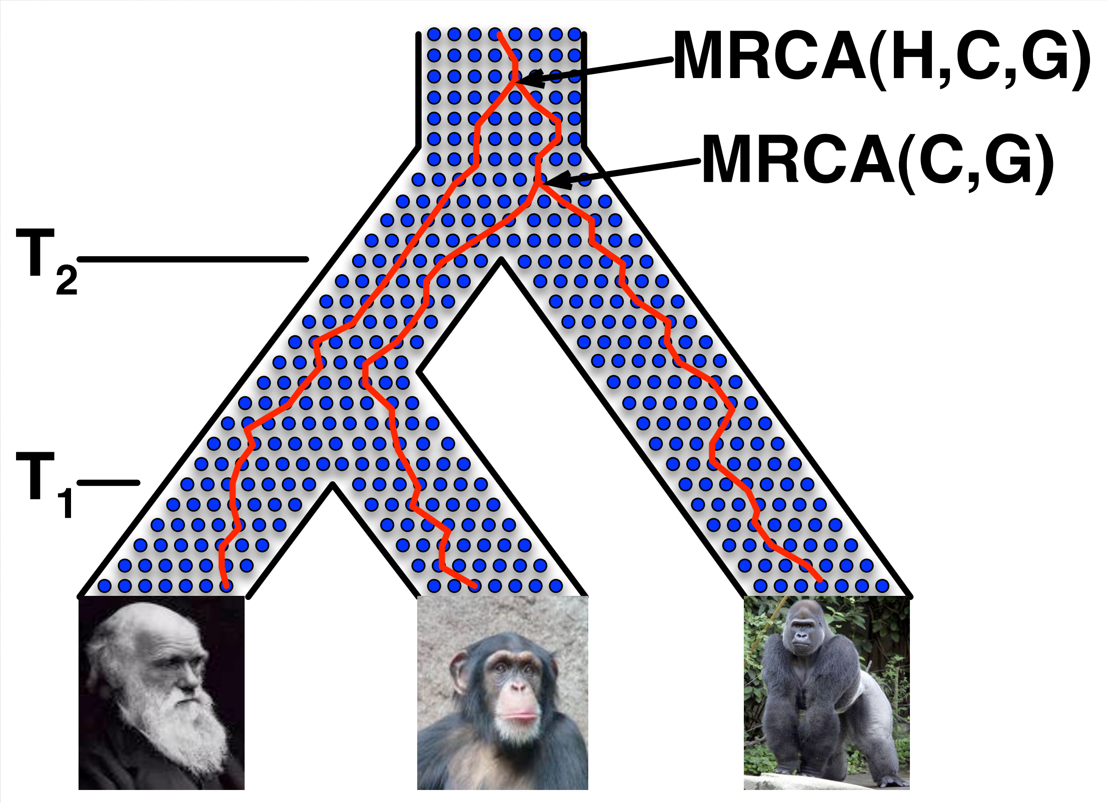
$$p_{uv}(t) = \sum_{k=v}^{u} \left[e^{-\frac{k(k-1)t}{2}} \frac{(2k-1)(-1)^{k-v}}{v!(k-v)!(v+k-1)} \prod_{y=0}^{k-1} \frac{(v+y)(u-y)}{u+y} \right]$$

Tavaré (Theoretical Population Biology, 1984)

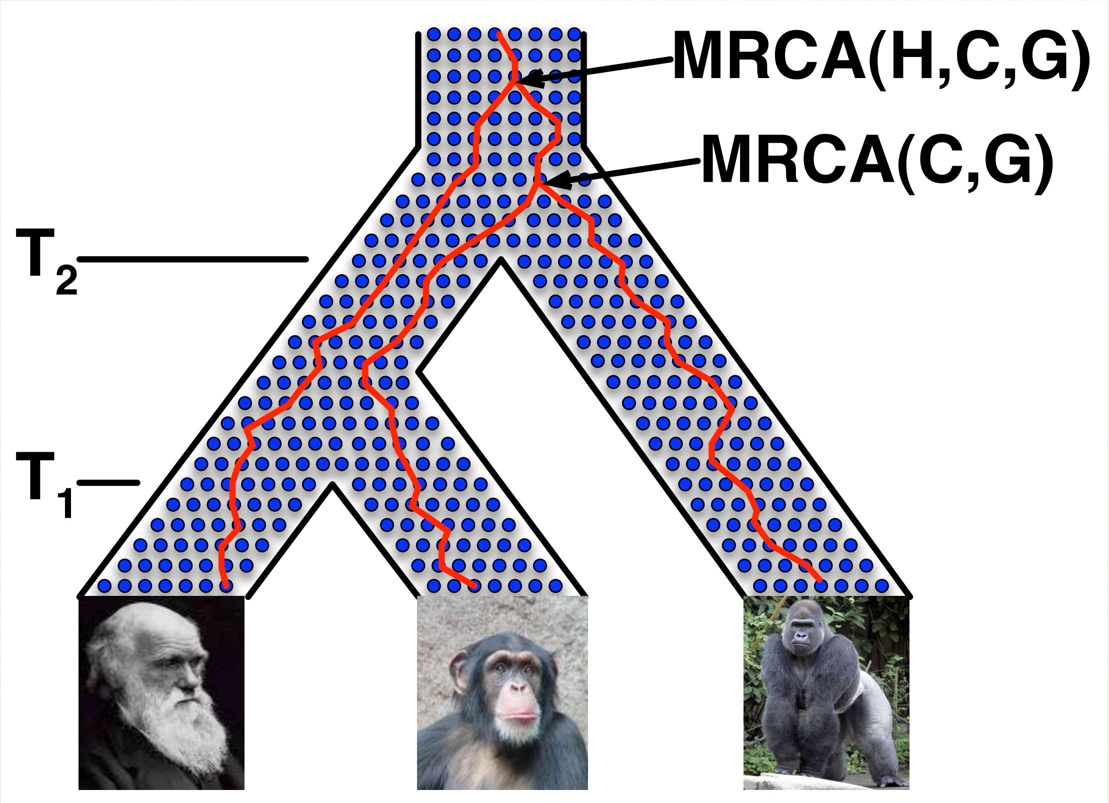
Watterson (Theoretical Population Biology, 1984)

Takahata and Nei (Genetics, 1985)

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

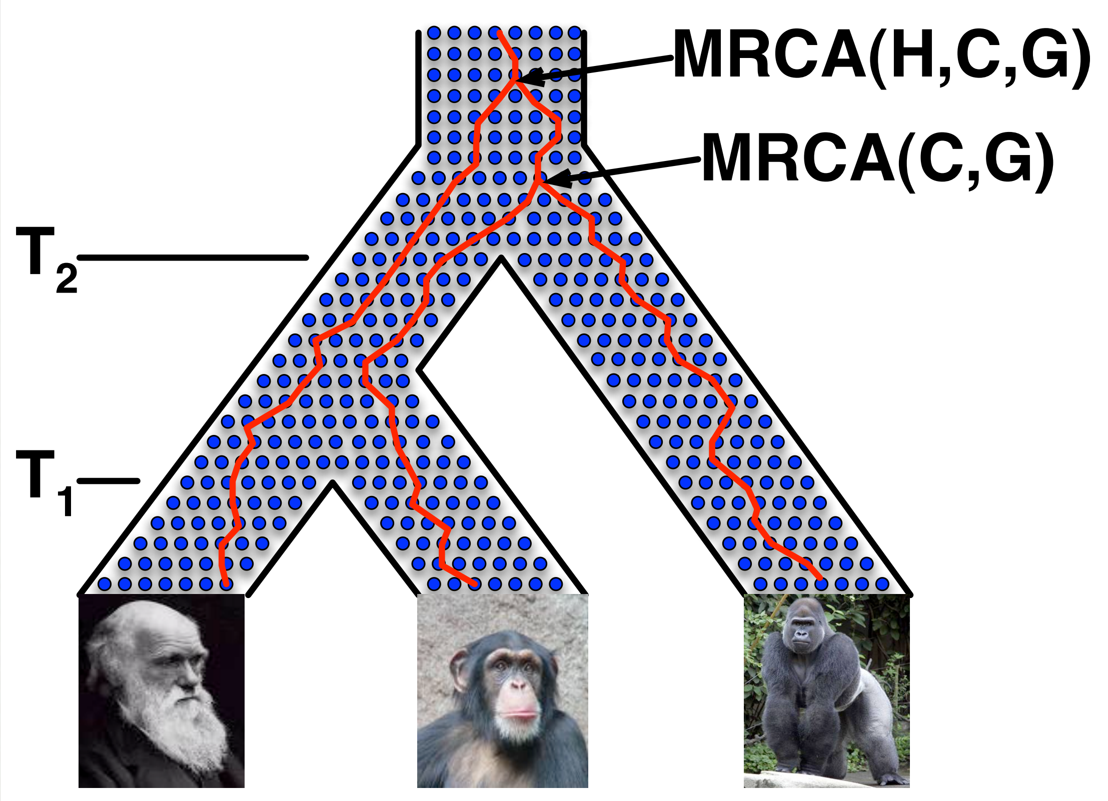


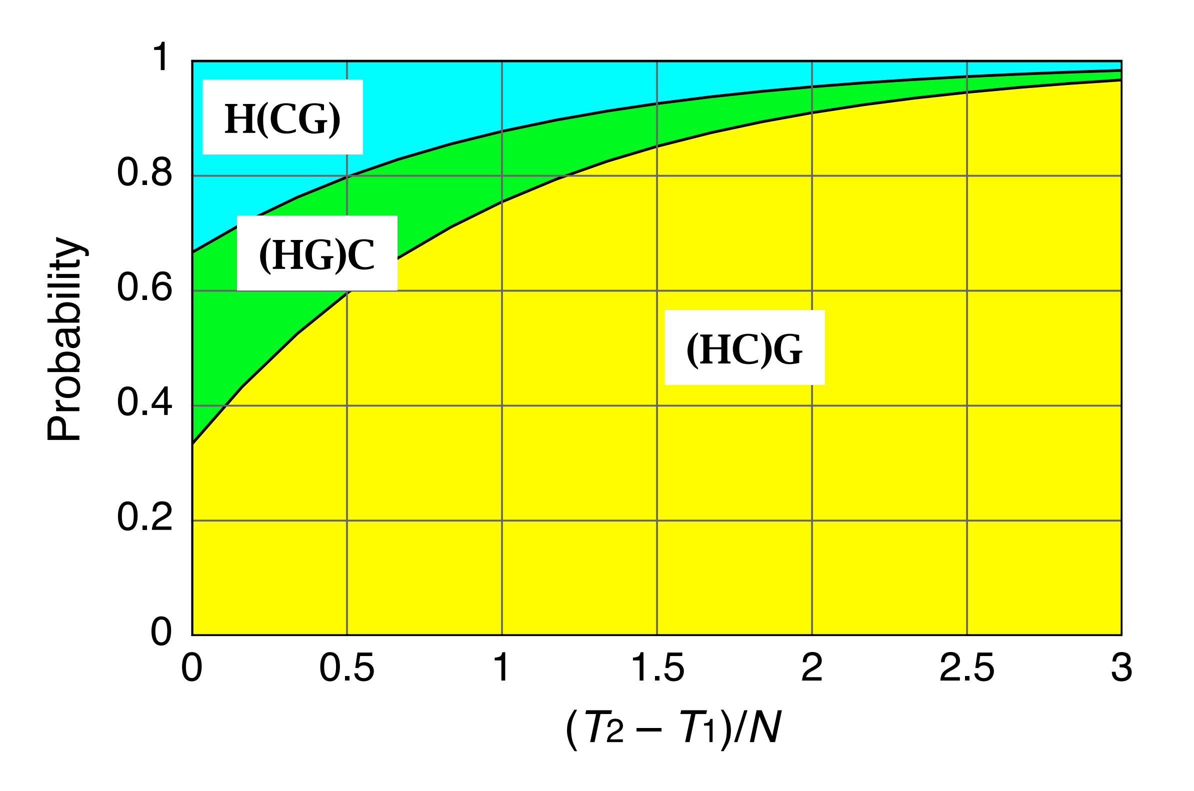
$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT



$$P[((HC)G)] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}$$
$$P[((HG)C)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$
$$P[((CG)H)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT



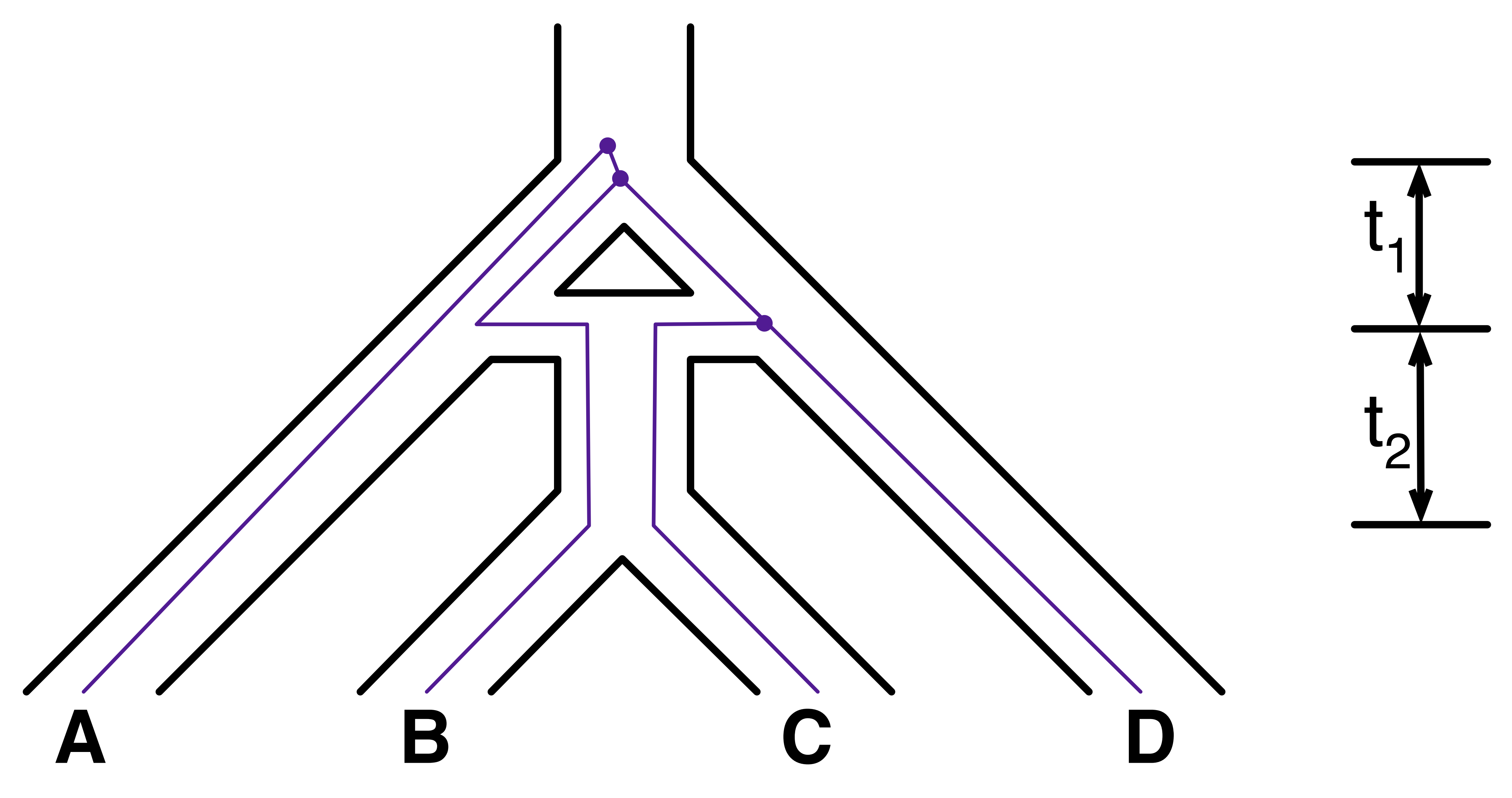


$$P[((HC)G)] = 1 - \frac{2}{3}e^{-(T_2 - T_1)/N}$$

$$P[((HG)C)] = \frac{1}{3}e^{-(T_2 - T_1)/N}$$

$$P[((CG)H)] = \frac{1}{3}e^{-(T_2 - T_1)/N}$$

$\mathbf{P}(gt|\Psi)$ UNDER BOTH SCENARIOS

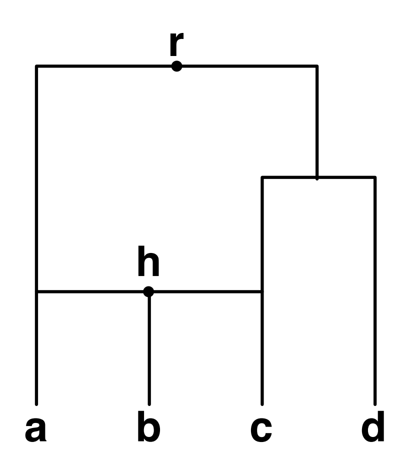


PHYLOGENETIC NETWORKS

A *phylogenetic network* N on set $\mathcal{X}$ of taxa is an ordered pair (G, f), where

- $G = (V, E)$ is a directed, acyclic graph (DAG) with $V = \{r\} \cup V_L \cup V_T \cup V_N$, where
  - $indeg(r) = 0$ (r is the *root* of N);
  - $\forall v \in V_L$, $indeg(v) = 1$ and $outdeg(v) = 0$ (V_L are the *leaves* of N);
  - $\forall v \in V_T$, $indeg(v) = 1$ and $outdeg(v) \geq 2$ (V_T are the *tree nodes* of N); and,
  - $\forall v \in V_N$, $indeg(v) = 2$ and $outdeg(v) = 1$ (V_N are the *reticulation nodes* of N),

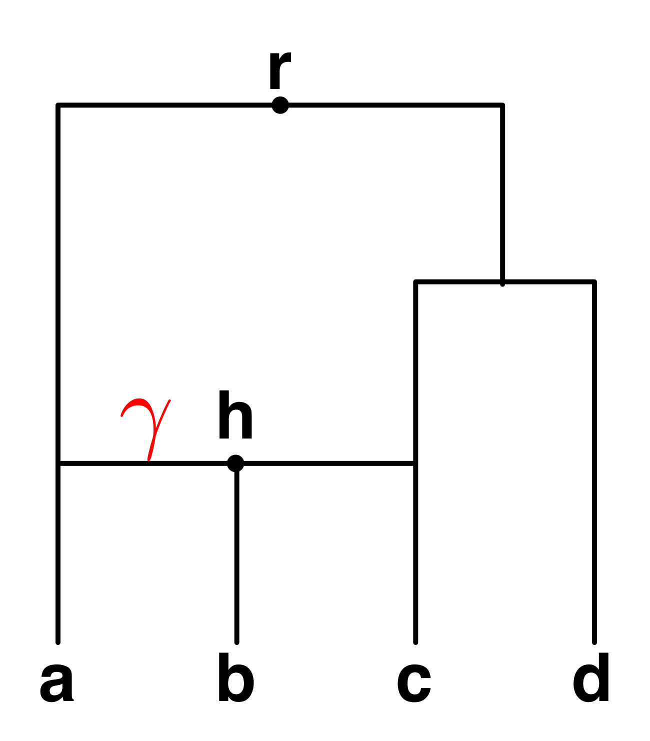
  and $E \subseteq V \times V$ are the network's edges (we distinguish between *reticulation edges*, edges whose heads are reticulation nodes, and *tree edges*, edges whose heads are tree nodes.
- $f : V_L \rightarrow \mathcal{X}$ is the *leaf-labeling* function, which is a bijection from V_L to $\mathcal{X}$.



PHYLOGENETIC NETWORKS

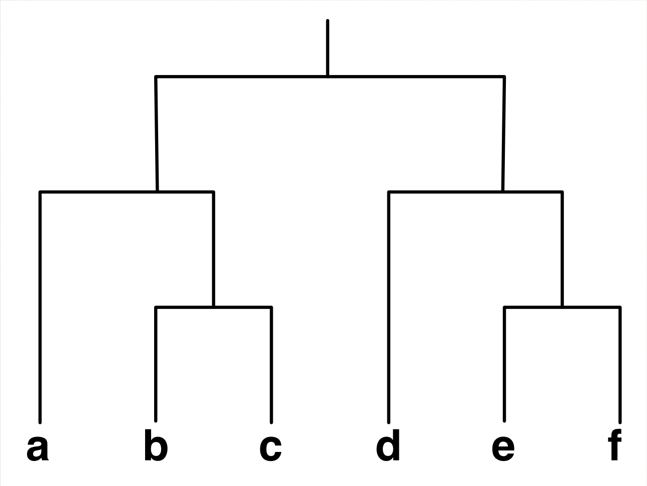
In addition to the topology, the network has

- branch lengths (in coalescent units), and
- inheritance probabilities

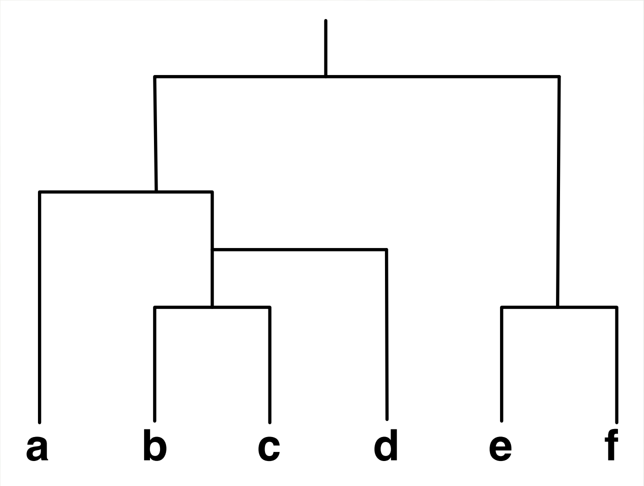


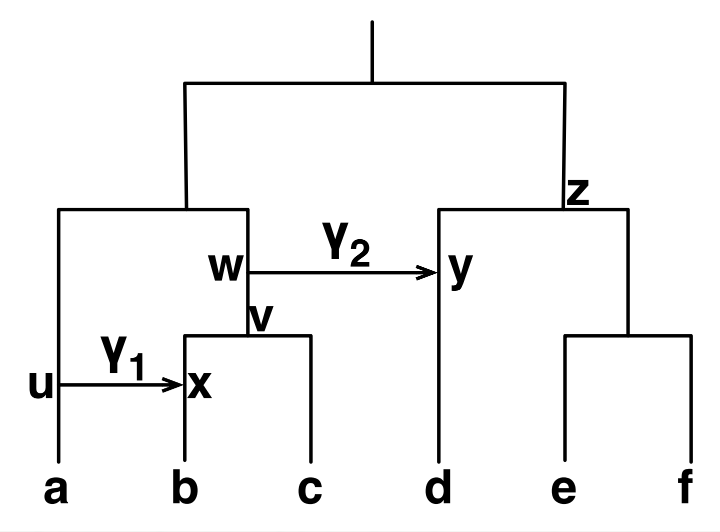
TREES INDUCED BY NETWORKS

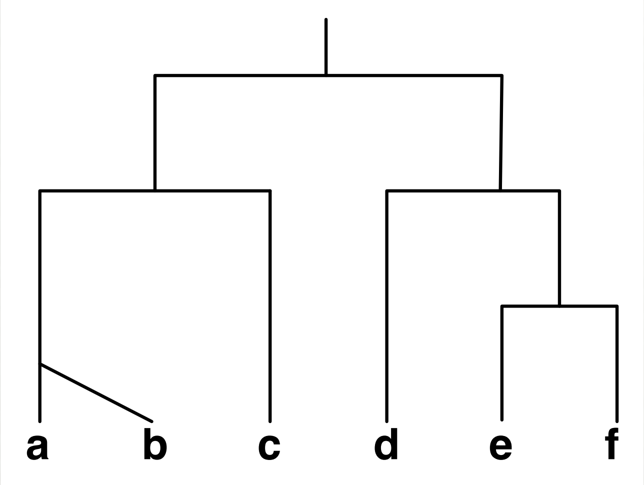
$$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)(1-\gamma_2)$$



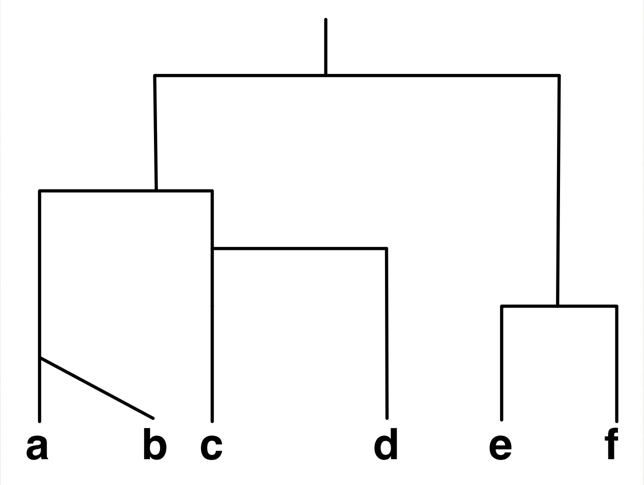
$$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)\gamma_2$$





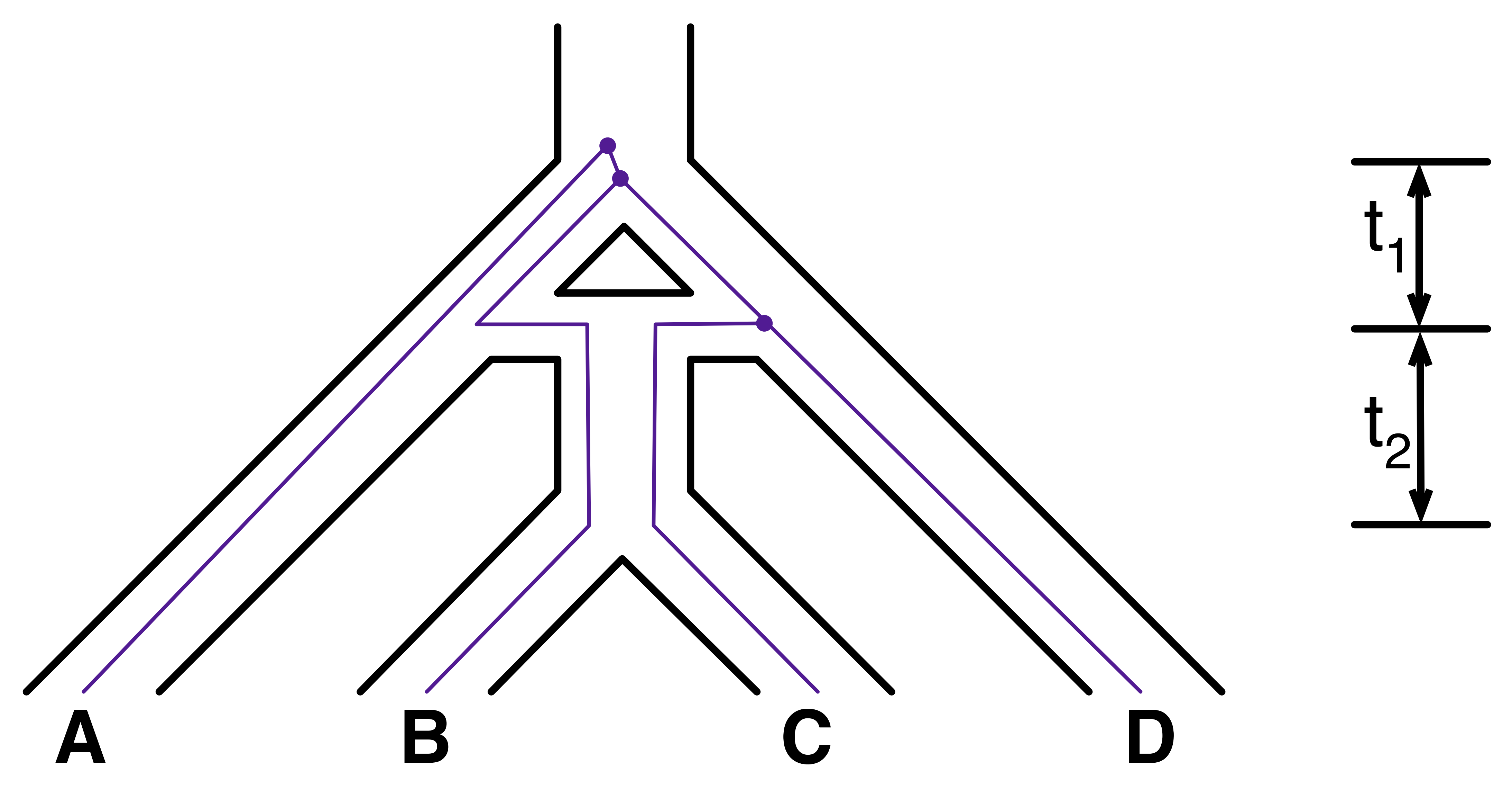


$$P_{N,\gamma_1,\gamma_2}(gt) = \gamma_1(1-\gamma_2)$$



$$P_{N,\gamma_1,\gamma_2}(gt) = \gamma_1\gamma_2$$

INDUCED TREES DON'T CAPTURE THE FULL STORY WHEN ILS IS PRESENT

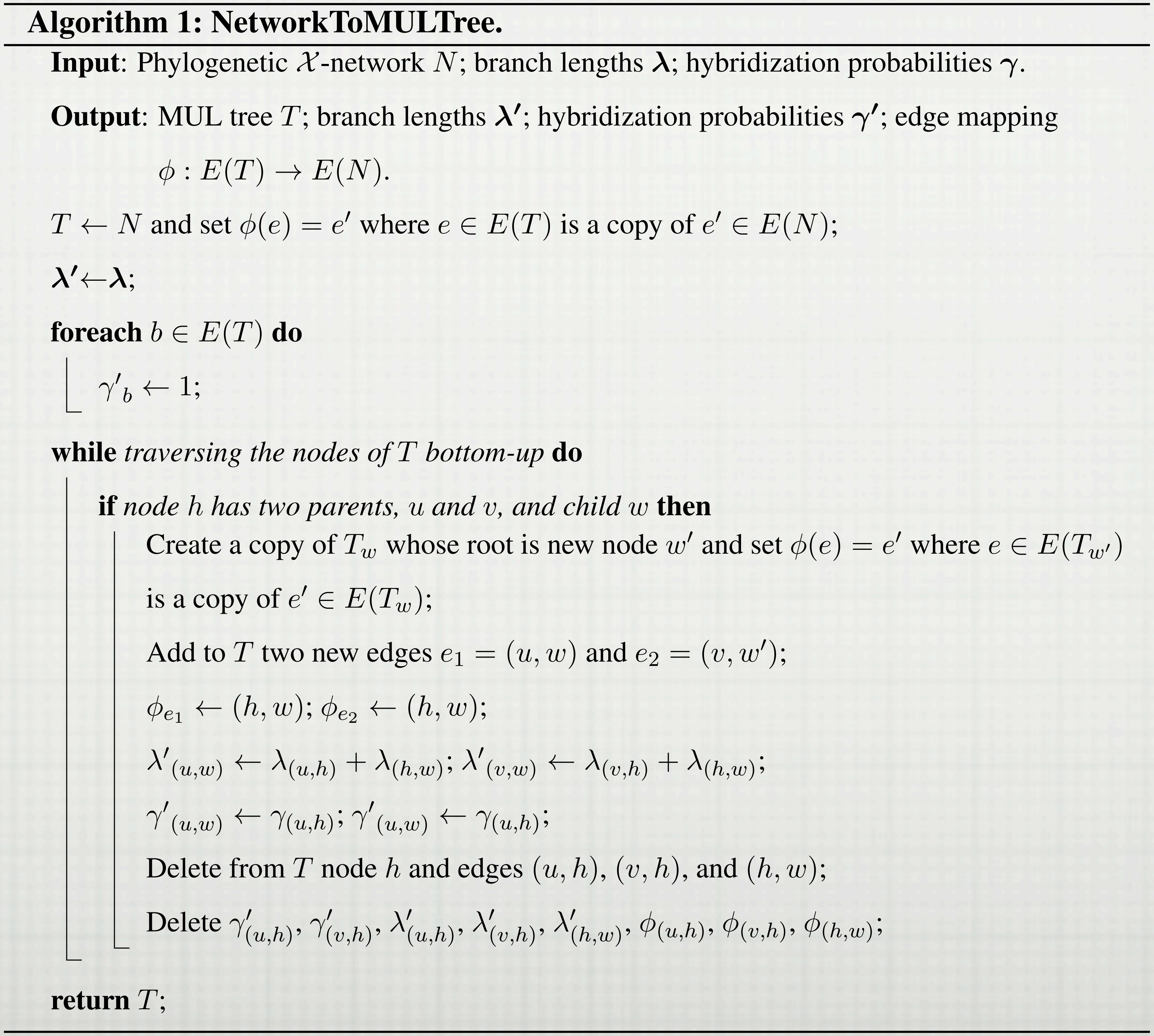


A SOLUTION

1. Convert the phylogenetic network N into a MUL-tree T
2. Consider all allele mappings from the leaves of gt to the leaves of T
3. For each allele mapping, compute the probability of observing gt, given T, and sum the probabilities.

[Yu, Degnan, Nakhleh, PLoS Genetics, 2012.]

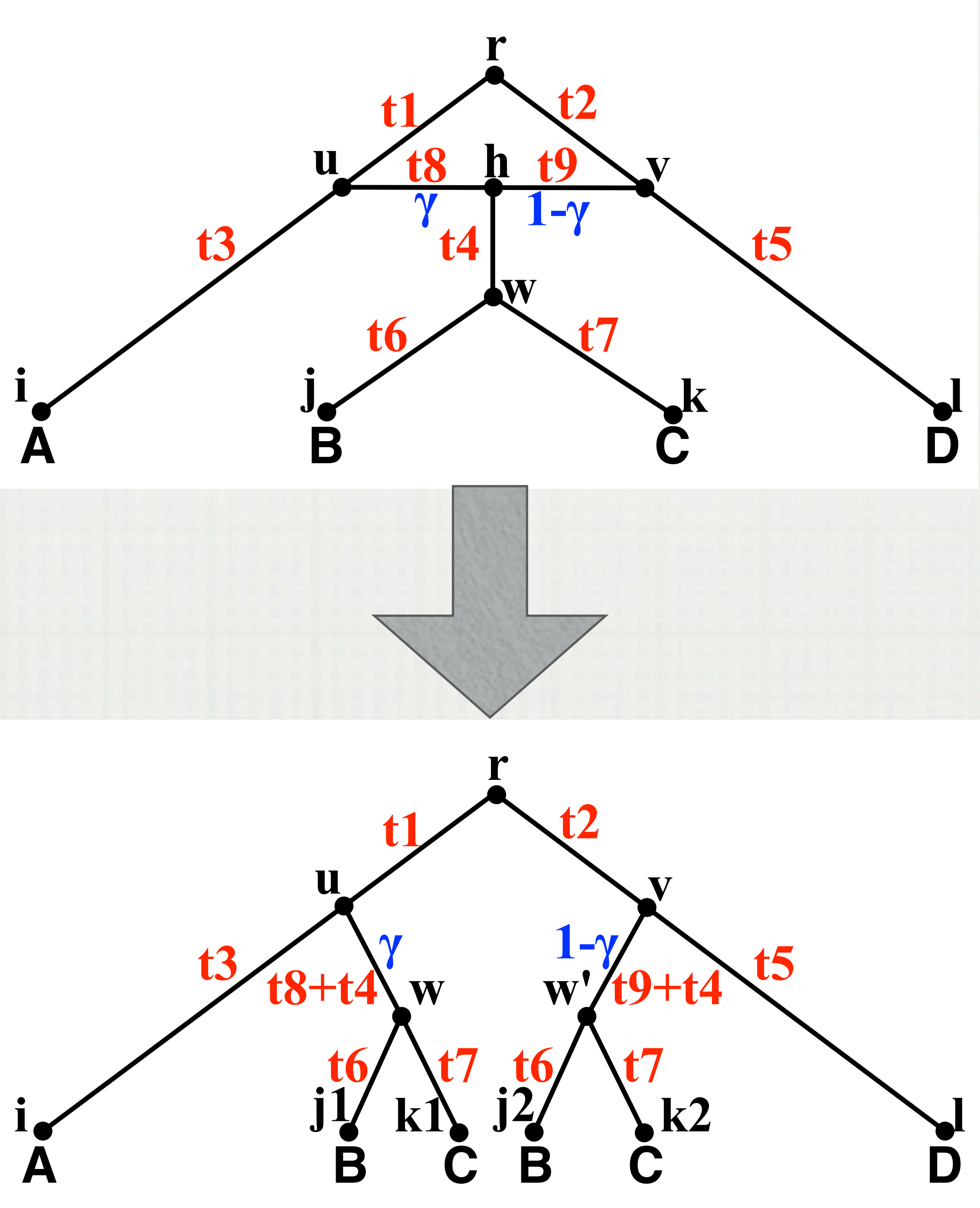
1. FROM A NETWORK TO A MUL-TREE



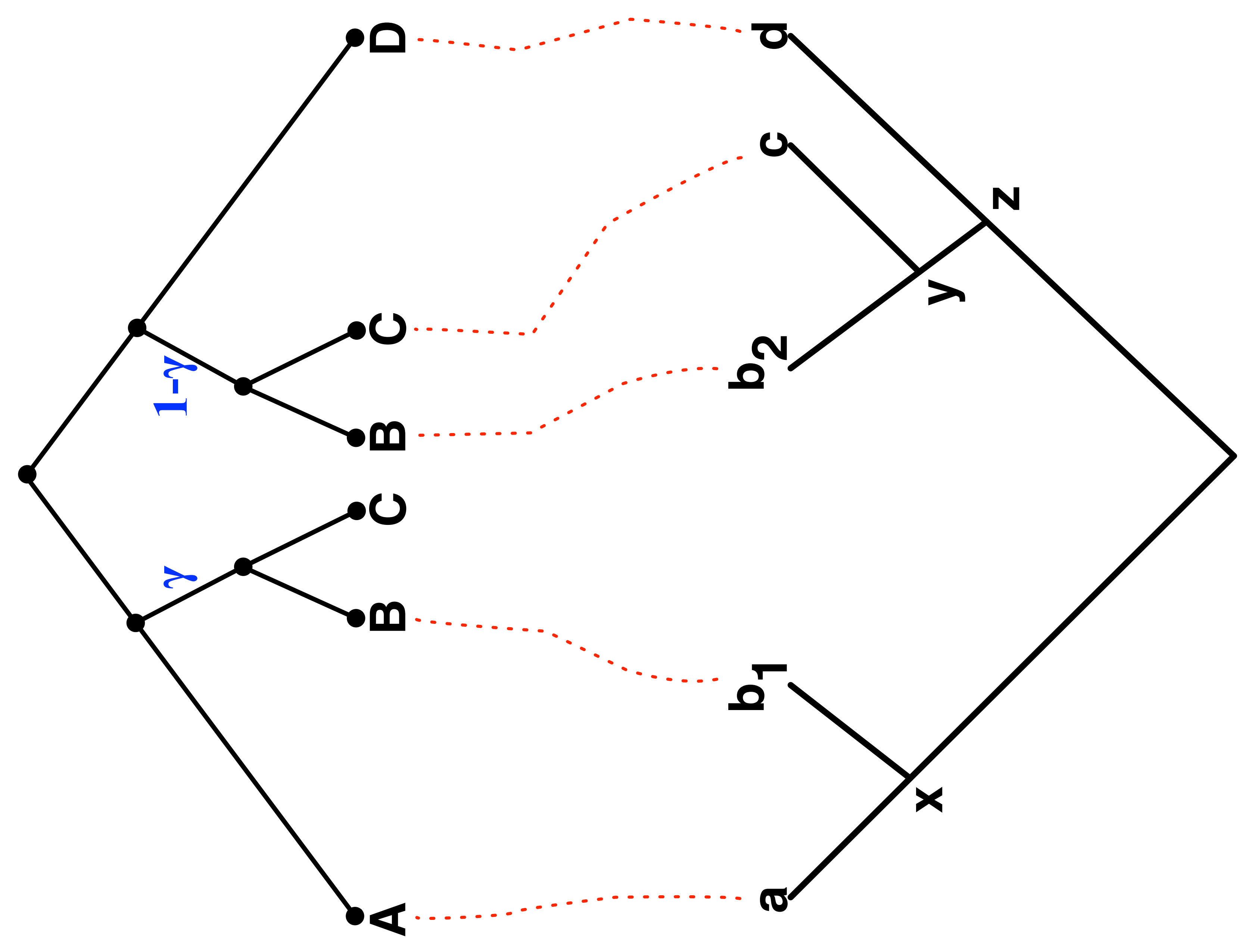
Algorithm 1: NetworkToMULTree.

```
Input: Phylogenetic $\mathcal{X}$-network $N$; branch lengths $\boldsymbol{\lambda}$; hybridization probabilities $\boldsymbol{\gamma}$.
Output: MUL tree $T$; branch lengths $\boldsymbol{\lambda'}$; hybridization probabilities $\boldsymbol{\gamma'}$; edge mapping
        $\phi : E(T) \rightarrow E(N)$.
$T \leftarrow N$ and set $\phi(e) = e'$ where $e \in E(T)$ is a copy of $e' \in E(N)$;
$\boldsymbol{\lambda'} \leftarrow \boldsymbol{\lambda}$;
foreach $b \in E(T)$ do
    ${\gamma'}_b \leftarrow 1$;
while traversing the nodes of $T$ bottom-up do
    if node $h$ has two parents, $u$ and $v$, and child $w$ then
        Create a copy of $T_w$ whose root is new node $w'$ and set $\phi(e) = e'$ where $e \in E(T_{w'})$
        is a copy of $e' \in E(T_w)$;
        Add to $T$ two new edges $e_1 = (u, w)$ and $e_2 = (v, w')$;
        $\phi_{e_1} \leftarrow (h, w)$; $\phi_{e_2} \leftarrow (h, w)$;
        ${\lambda'}_{(u,w)} \leftarrow \lambda_{(u,h)} + \lambda_{(h,w)}$; ${\lambda'}_{(v,w)} \leftarrow \lambda_{(v,h)} + \lambda_{(h,w)}$;
        ${\gamma'}_{(u,w)} \leftarrow \gamma_{(u,h)}$; ${\gamma'}_{(u,w)} \leftarrow \gamma_{(u,h)}$;
        Delete from $T$ node $h$ and edges $(u, h)$, $(v, h)$, and $(h, w)$;
        Delete $\gamma'_{(u,h)}$, $\gamma'_{(v,h)}$, $\lambda'_{(u,h)}$, $\lambda'_{(v,h)}$, $\lambda'_{(h,w)}$, $\phi_{(u,h)}$, $\phi_{(v,h)}$, $\phi_{(h,w)}$;
return $T$;
```

1. FROM A NETWORK TO A MUL-TREE



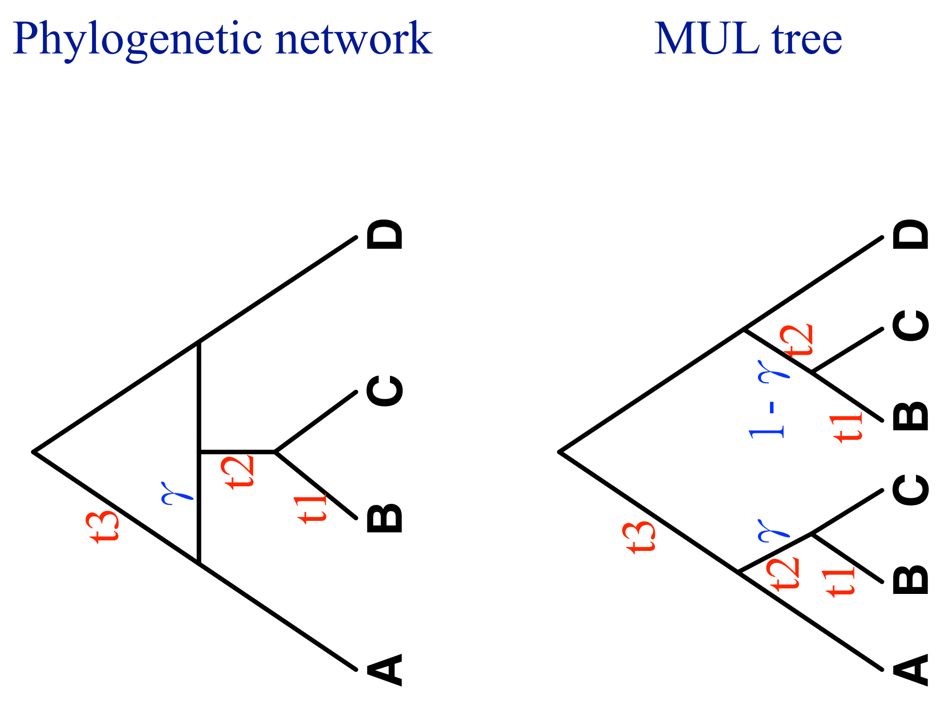
2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T



2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T

Phylogenetic network

MUL tree



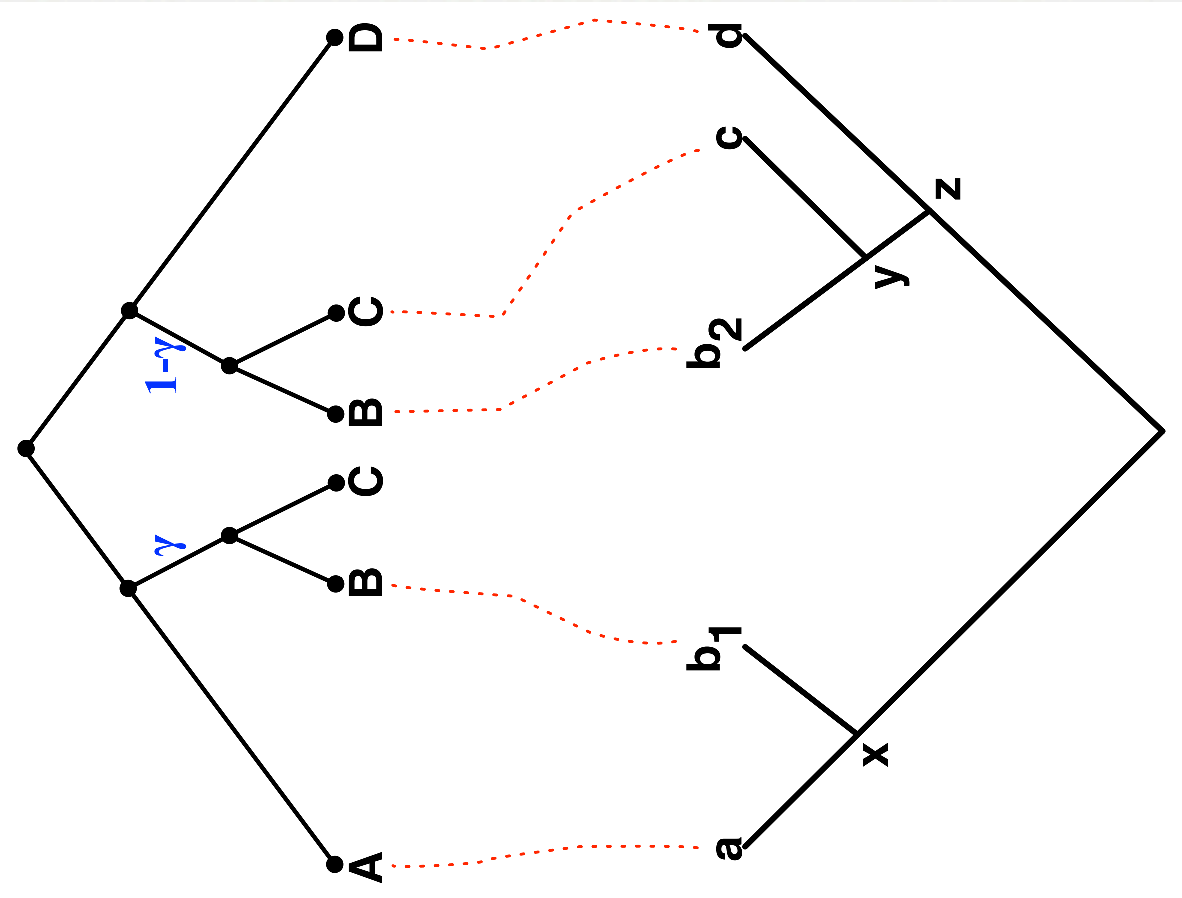
Valid allele mappings

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
d	d	d	d	d	d	d	d
—	c	—	c	—	c	—	c
—	—	b_2	b_2	b_1	b_1	b_1 b_2	b_1 b_2
c	—	c	—	c	—	c	—
b_1 b_2	b_1 b_2	b_1	b_1	b_2	b_2	—	—
a	a	a	a	a	a	a	a

$$P_{N,\boldsymbol{\lambda},\boldsymbol{\gamma}}(gt) = \sum_{f\in\mathcal{F}} P_{T,\boldsymbol{\lambda}',\boldsymbol{\gamma}',f}(gt)$$

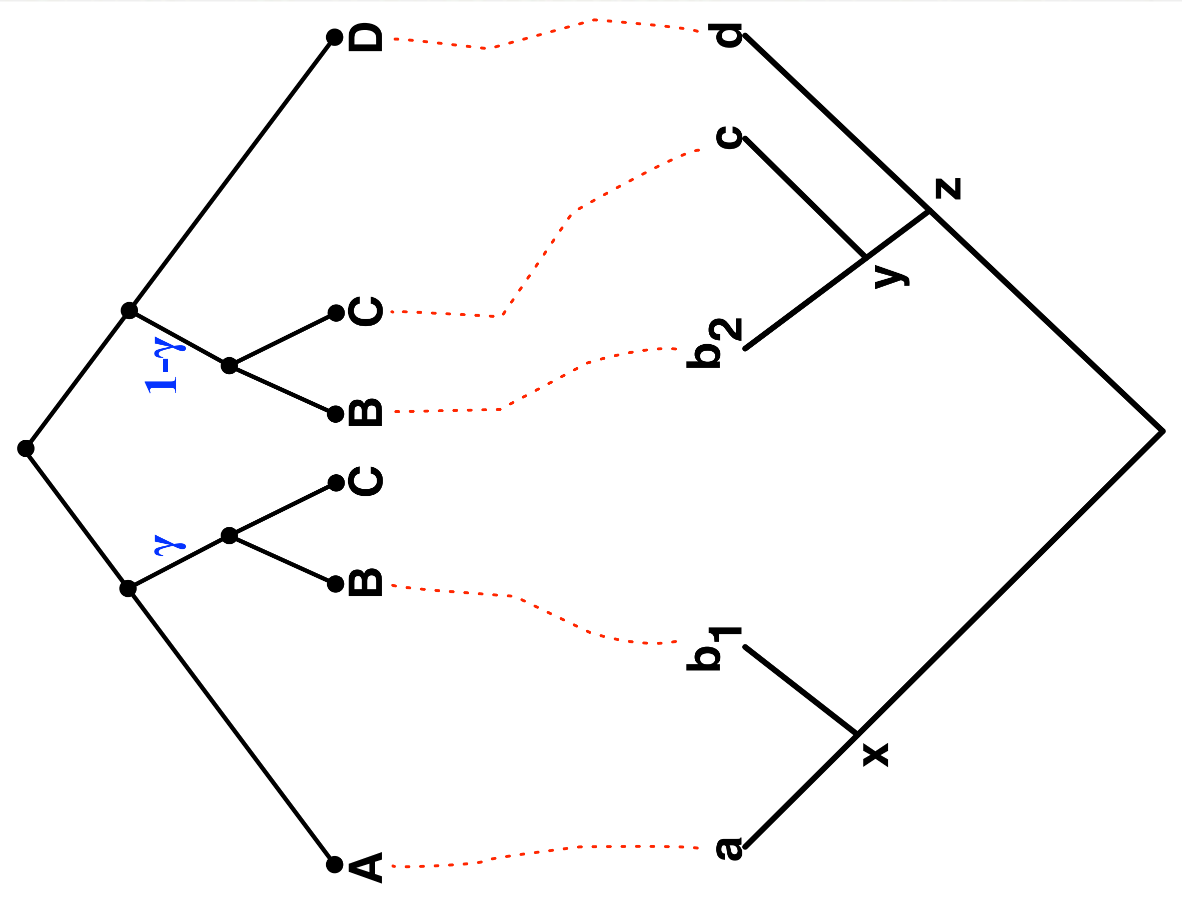
2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T

- □ We need to account for dependence among the branches of the MUL-tree



2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T

□ We need to account for dependence among the branches of the MUL-tree



□ The edge-mapping ϕ solves this problem.

3. THE PROBABILITY OF gt GIVEN MUL-TREE T

$$P_{T,\boldsymbol{\lambda}',\gamma',f}(gt) = \sum_{h \in H_{T,f}(gt)} \frac{w(h)}{d(h)} \prod_{b=1}^{n-2} {\gamma'_b}^{v_b(h)} P'_b(h)$$

$$\prod_{b \in \phi^{-1}(b')} P'_b(h) = \left[\frac{1}{d_{b'}(h)} p_{u_{b'}(h) v_{b'}(h)}(\lambda_{b'}) \left[(u_{b'}(h) - v_{b'}(h))! \prod_{b \in \phi^{-1}(b')} \frac{w_b(h)}{(u_b(h) - v_b(h))!} \right] \right]$$

$$u_{b'}(h) = \sum_{b \in \phi^{-1}(b')} u_b(h) \qquad v_{b'}(h) = \sum_{b \in \phi^{-1}(b')} v_b(h)$$

ACCOUNTING FOR UNCERTAINTY IN GENE TREES

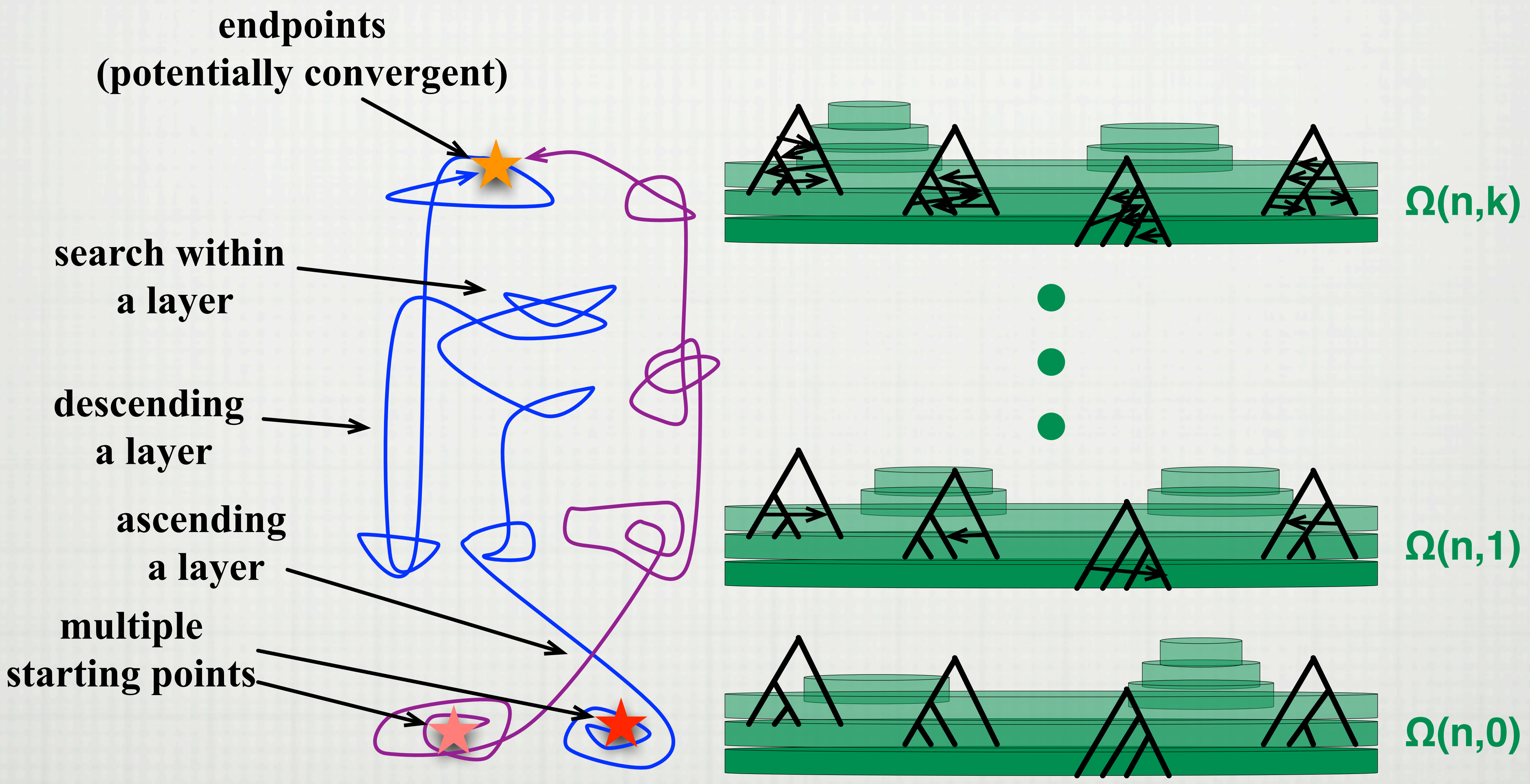
- □ We have implemented two methods for accounting for uncertainty in the estimated gene trees:
  - □ Using gene tree distributions: $L(N,\boldsymbol{\lambda},\boldsymbol{\gamma}|\mathscr{G}) = \prod_{g\in\mathscr{G}} [\mathbf{P}_{N,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G=g)]^{p_g}$
  - □ Using non-binary trees:

$$L(N,\boldsymbol{\lambda},\boldsymbol{\gamma}|\mathscr{G}) = \prod_{g\in\mathscr{G}} \max_{g'\in b(g)} \{\mathbf{P}_{N,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G=g')\}$$

$$L(\Psi|\mathcal{G}) = c \cdot \prod_{gt \in \mathcal{G}} \mathbf{P}(gt|\Psi)$$

$$Objective : \ \mathrm{argmax}_{\Psi} L(\Psi|\mathcal{G})$$

SOLUTION



endpoints
(potentially convergent)

search within
a layer

descending
a layer

ascending
a layer

multiple
starting points

$\Omega(n,k)$

$\Omega(n,1)$

$\Omega(n,0)$

[Yu, Dong, Liu, Nakhleh, Under Review, 2014.]

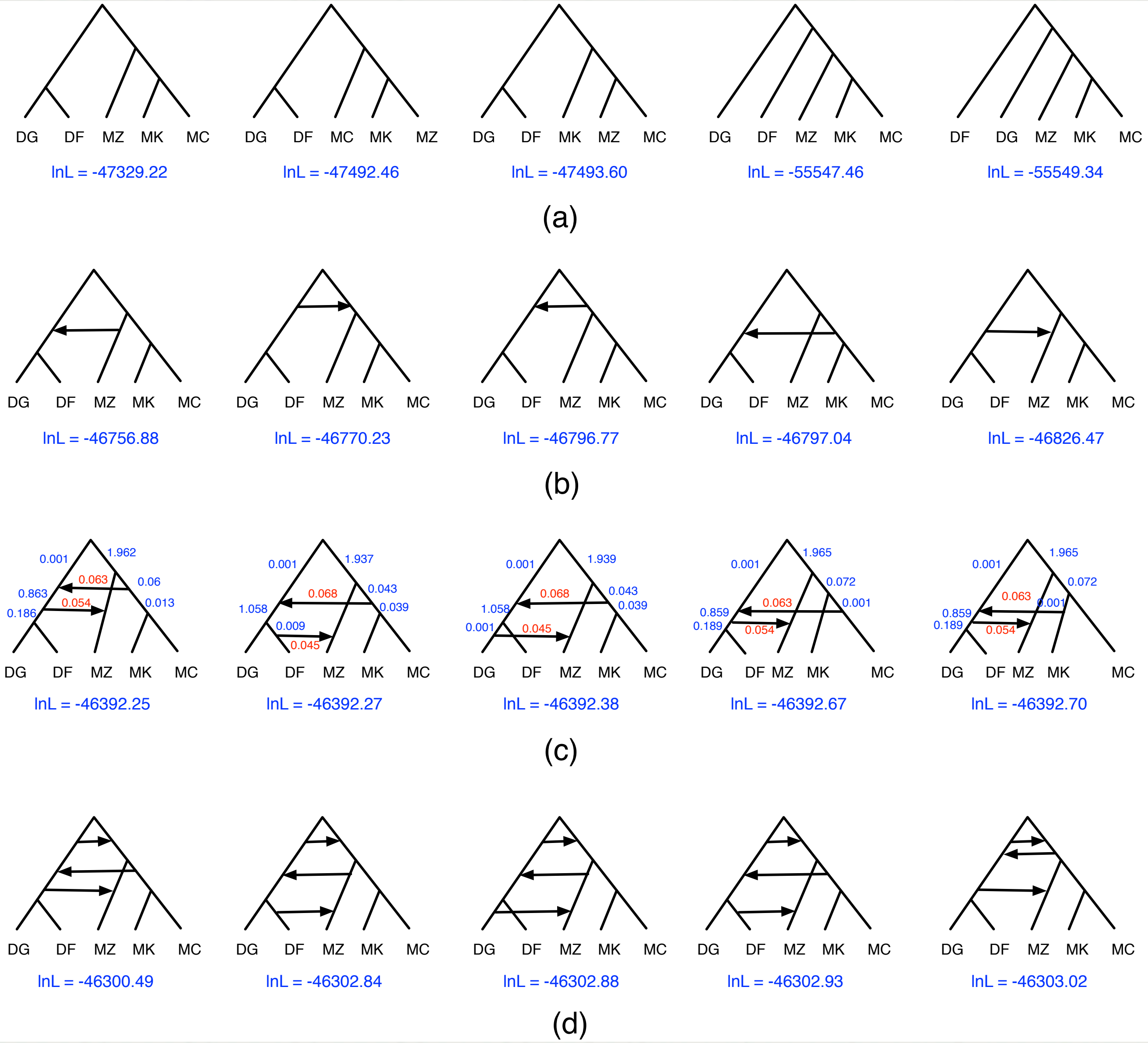
SOLUTION

□ We have a much faster algorithm for computing gene tree probabilities that neither converts the network to a MUL-tree nor does an explicit summation over coalescent histories.

□ [Yu, Ristic, Nakhleh, BMC Bioinformatics, 2013]

SOLUTION

□ To account for model complexity, we considered information criteria (which were used before in this context), and introduced an implementation with cross-validation.



(a)

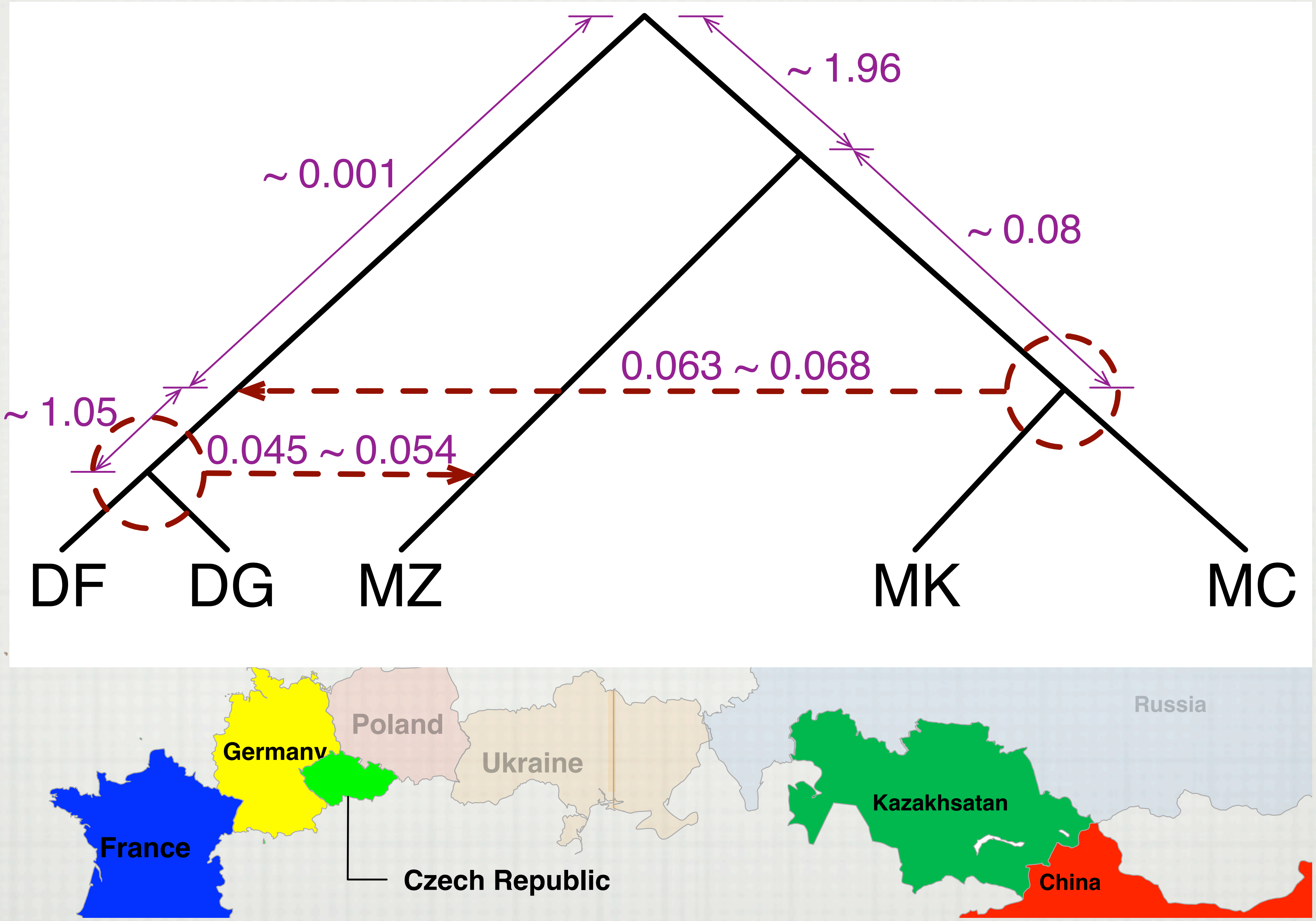
(b)

(c)

(d)

	lnL	AIC	AICc	BIC	Error of cross-validation
$N(0)$	-47329	94664	94664	94688	7.69×10^{-5}
$N(1)$	-46756	93527	93527	93583	5.36×10^{-5}
$N(2)$	-46392	92806	92806	92893	4.03×10^{-5}
$N(3)$	-46300	92635	92635	92754	4.13×10^{-5}

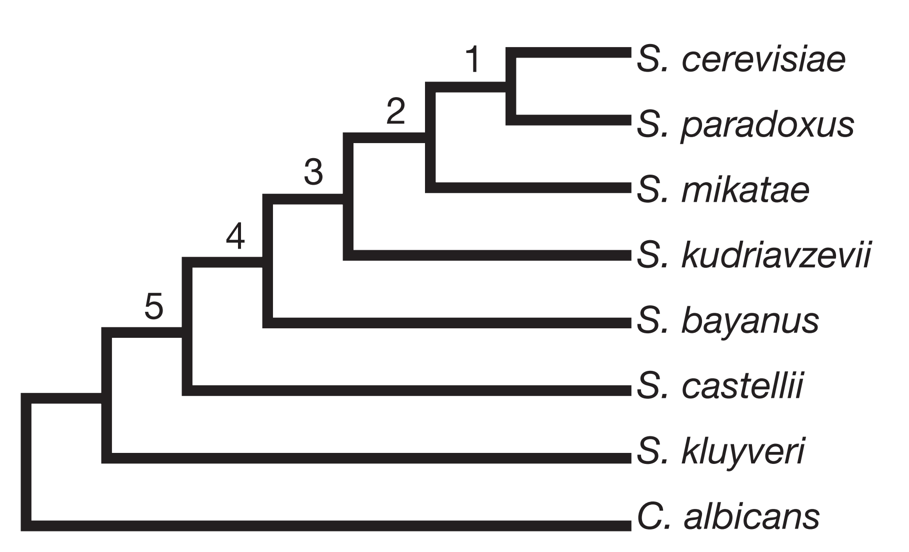
[Yu, Dong, Liu, Nakhleh, Under Review, 2014.]



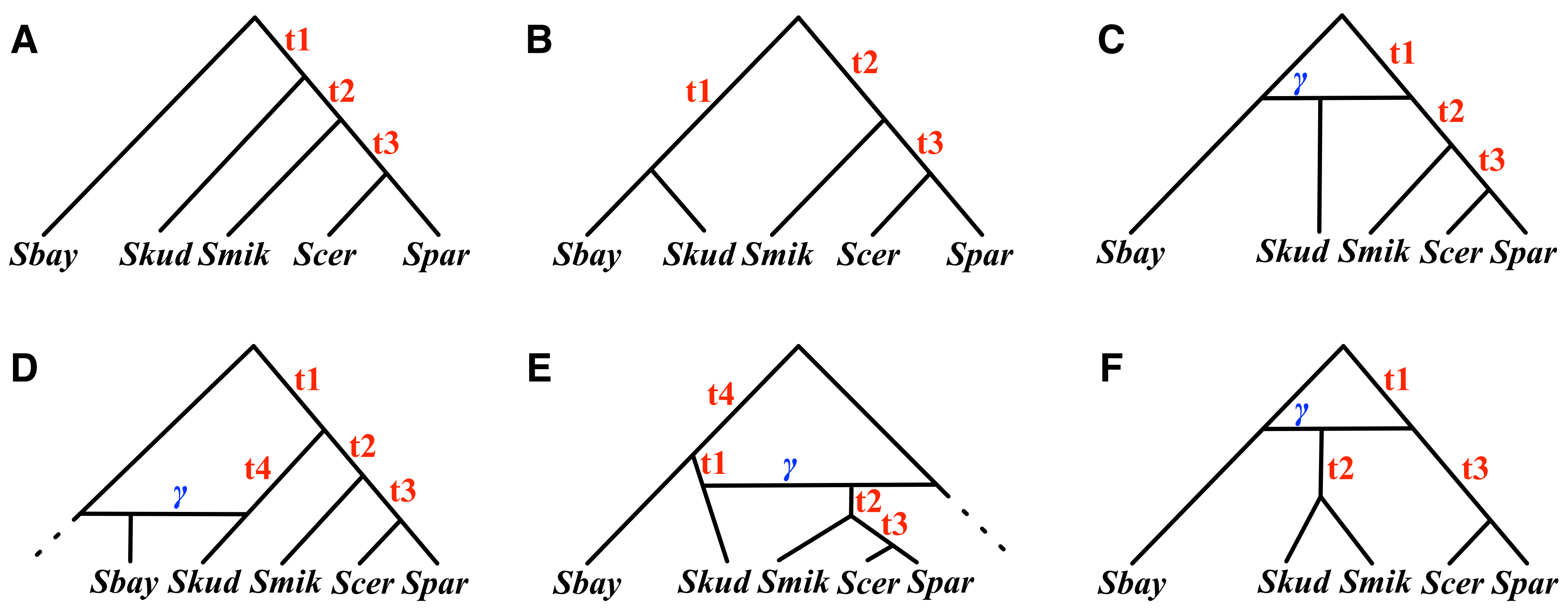
[Yu, Dong, Liu, Nakhleh, Under Review, 2014.]

THE YEAST DATA SET OF ROKAS ET AL. (NATURE 2003)

- The authors concatenated the sequences of 106 genes, and inferred a single species tree, which had 100% bootstrap support of all branches



REANALYSIS OF THE YEAST DATA



Species phylogeny	t_1	t_2	t_3	t_4	γ	$-lnL$	AIC	AICc	BIC
Fig. 3(**A**)	0.3	1.25	3.6	N/A	N/A	205	416	417	424
Fig. 3(**B**)	0.2	1.35	3.6	N/A	N/A	208	423	423	431
Fig. 3(**C**)	1.1	1.05	3.6	N/A	0.34	188	384	385	395
Fig. 3(**D**)	3.45	1.15	3.6	3.05	0.34	157	325	326	338
Fig. 3(**E**)	0.3	1.25	3.6	N/A	1.0	205	420	421	434
Fig. 3(**F**)	1.55	0.05	3.7	N/A	0.18	252	512	512	523

[Yu, Degnan, Nakhleh, PLoS Genetics, 2012.]

□ For a gene tree with its coalescence times, we also have a solution:

$$P(ht|N_{\boldsymbol{\lambda},\boldsymbol{\gamma}}) = \prod_{b=(u,v)\in E(N_{\boldsymbol{\lambda},\boldsymbol{\gamma}})} \left[\prod_{k=1}^{|T_b(ht)|-1} e^{-\binom{u_b(ht)-k+1}{2}(T_b(ht)_{k+1}-T_b(ht)_k)} \right] \times e^{-\binom{v_b(ht)}{2}(\tau_{N_{\boldsymbol{\lambda},\boldsymbol{\gamma}}}(u)-T_b(ht)_{|T_b(ht)|})} \times \gamma_b^{u_b(ht)}$$

$$P(g_{\boldsymbol{\lambda}'}|N_{\boldsymbol{\lambda},\boldsymbol{\gamma}}) = \sum_{ht\in H_{N_{\boldsymbol{\lambda},\boldsymbol{\gamma}}}(g_{\boldsymbol{\lambda}'})} P(ht|N_{\boldsymbol{\lambda},\boldsymbol{\gamma}})$$

[Yu, Dong, Liu, Nakhleh, Under Review, 2014.]

- Our models and solutions allow for inference of networks directly from sequences when independent loci are used:

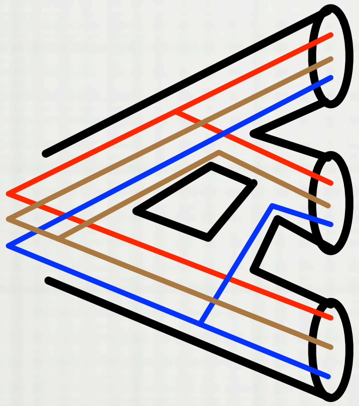
$$L(N_{\boldsymbol{\lambda},\boldsymbol{\gamma}}|\mathcal{S}) = \prod_{s\in\mathcal{S}} \left[\sum_{g} \int_{\boldsymbol{\tau}} \mathbf{P}(s|g_{\boldsymbol{\tau}}) \cdot \mathbf{P}(g_{\boldsymbol{\tau}}|N_{\boldsymbol{\lambda},\boldsymbol{\gamma}}) \right]$$

From Phylogenetic Networks to Genome Annotation with Introgression

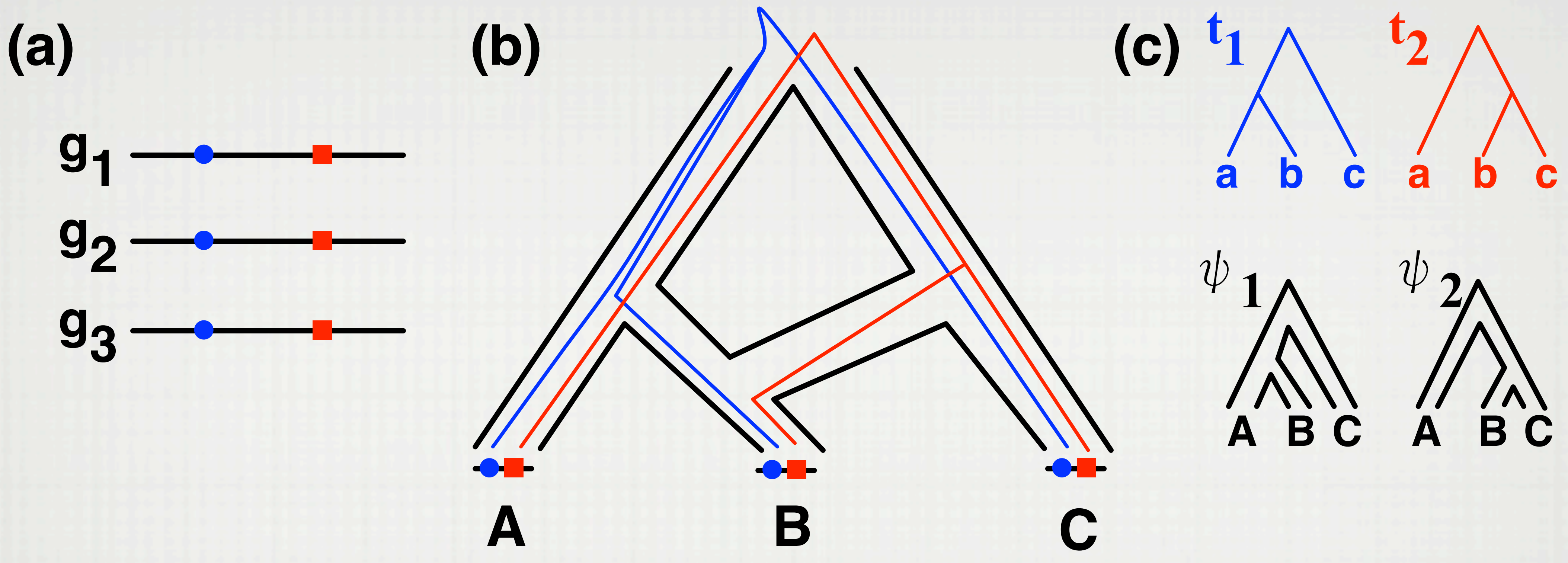
Phylogenetic network

+

Local (gene) genealogies



Genomes



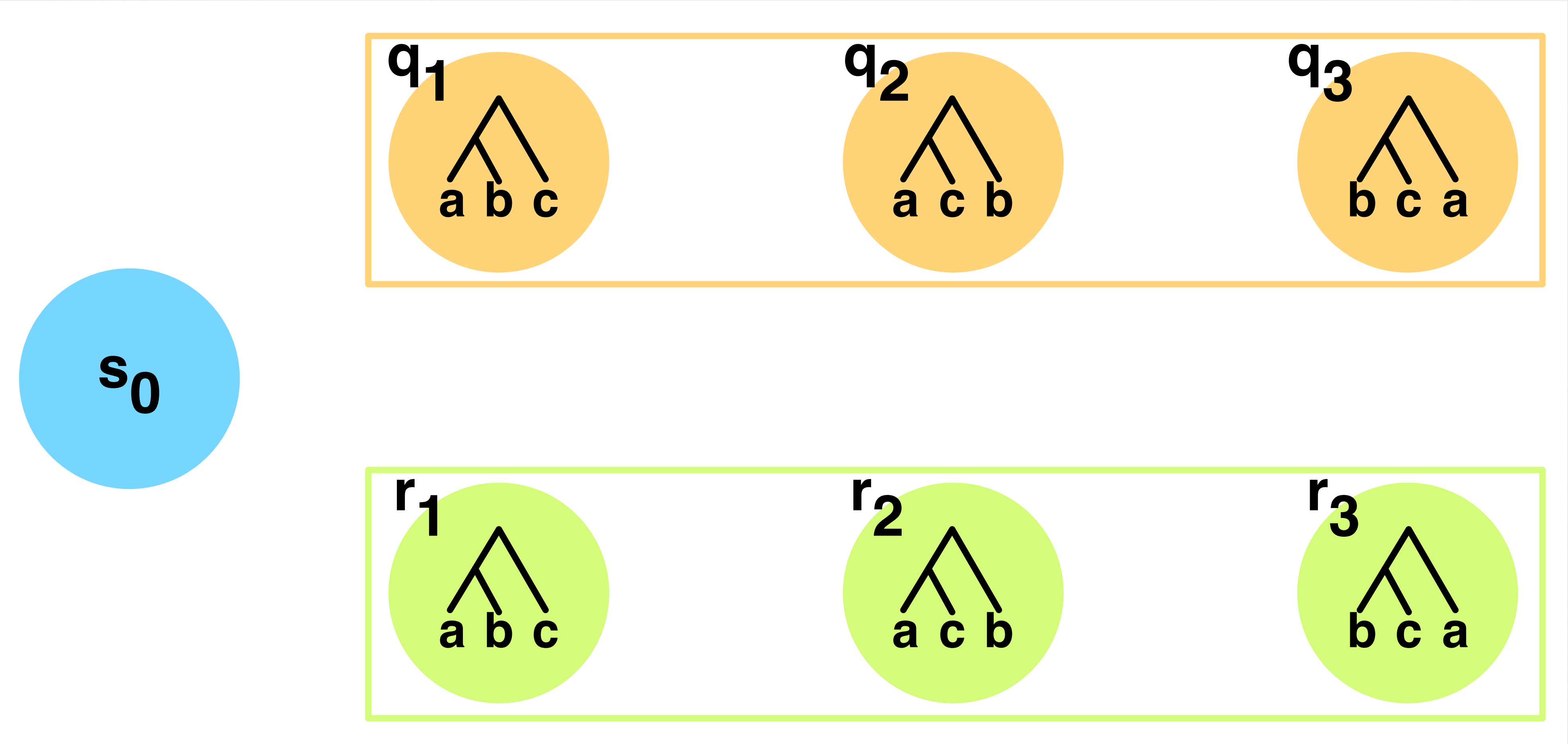
Input: A set $\mathcal{G}$ of m aligned genomes, each of length n, and a set Ψ of parental species trees.

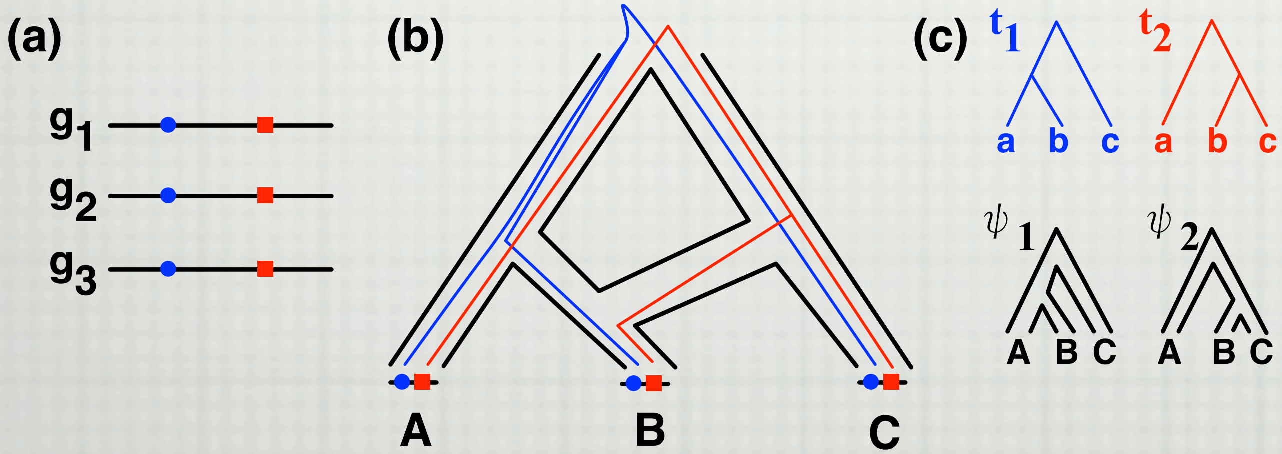
Output: For each site $1 \leq j \leq n$, the probability

$$\mathbf{P}(\pi_j = (t_x, \psi_y)|\mathcal{G})$$

for every $t_x \in \Delta(m)$ and $\psi_y \in \Psi$.

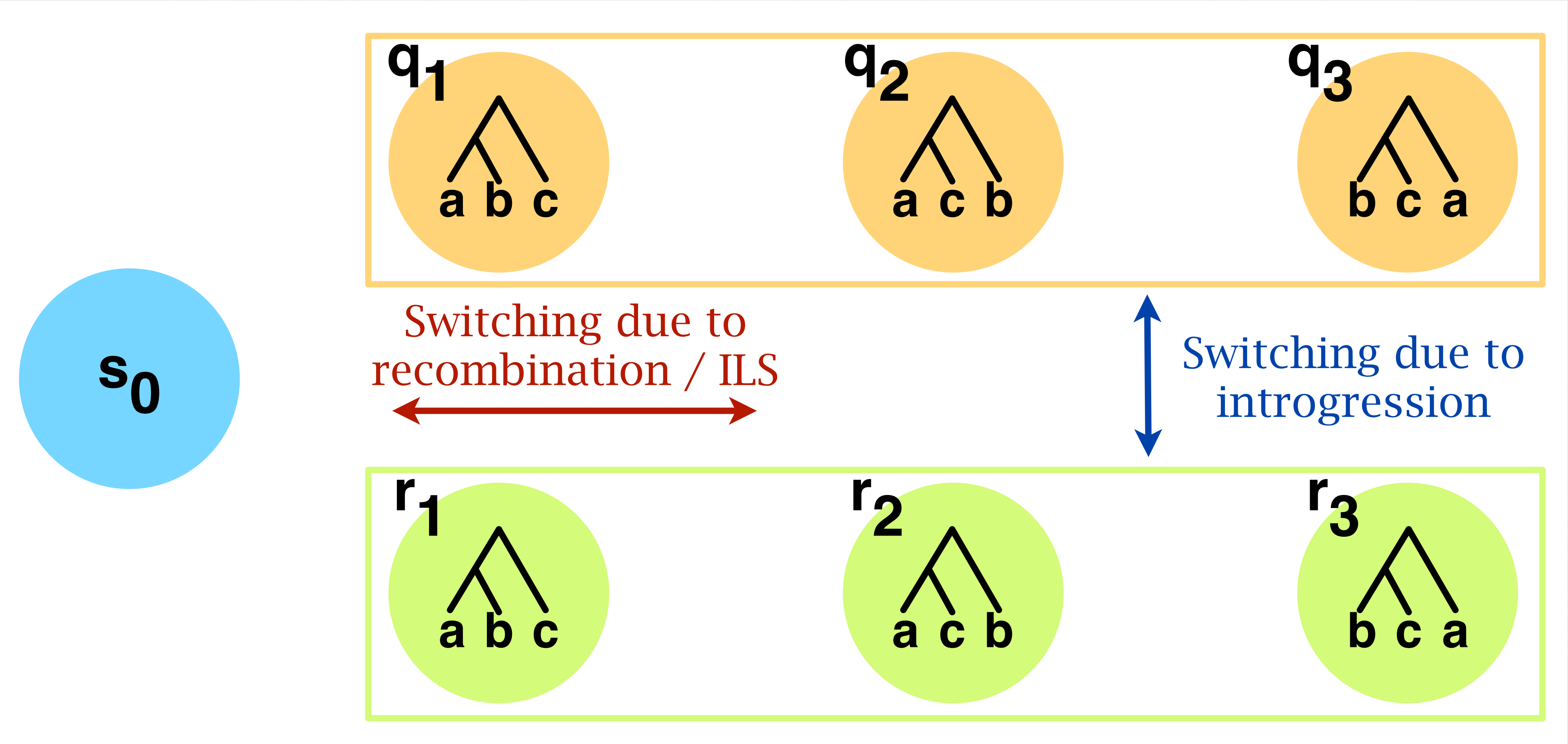
SOLUTION: PHYLONET-HMM





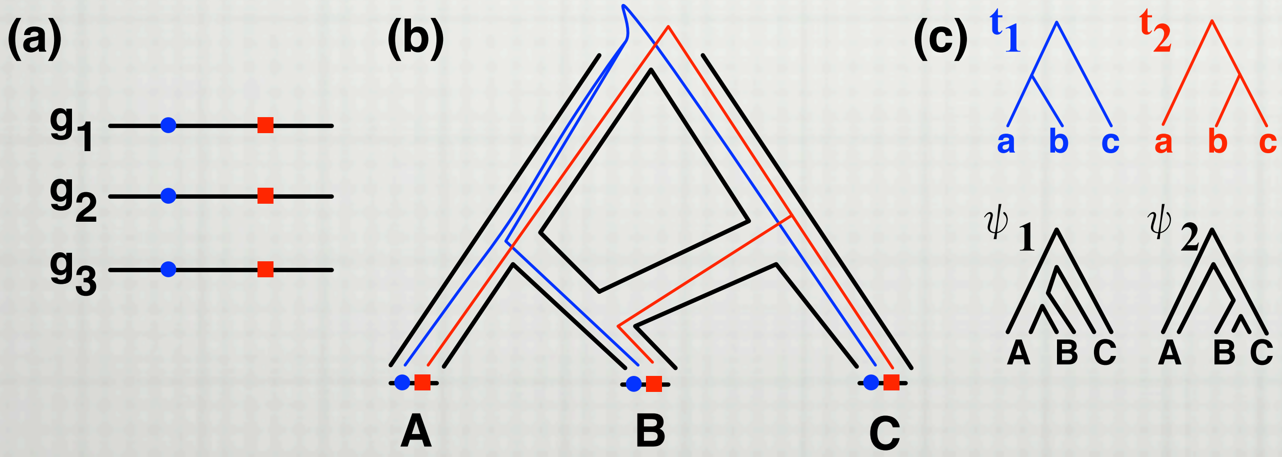
[Liu, Dai, Truong, Song, Kohn, Nakhleh, PLoS Comp Bio, 2014.]

SOLUTION: PHYLONET-HMM



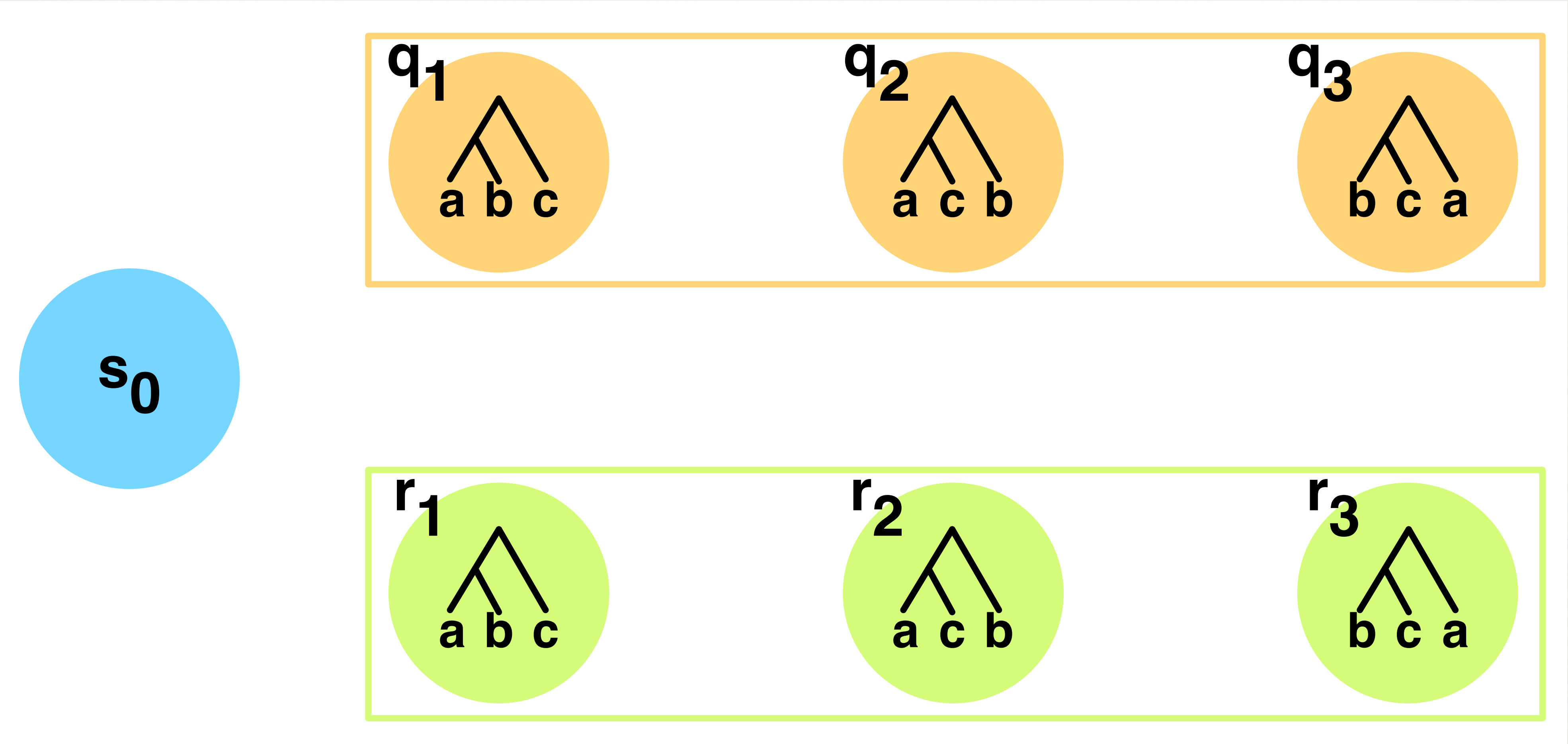
Switching due to recombination / ILS

Switching due to introgression



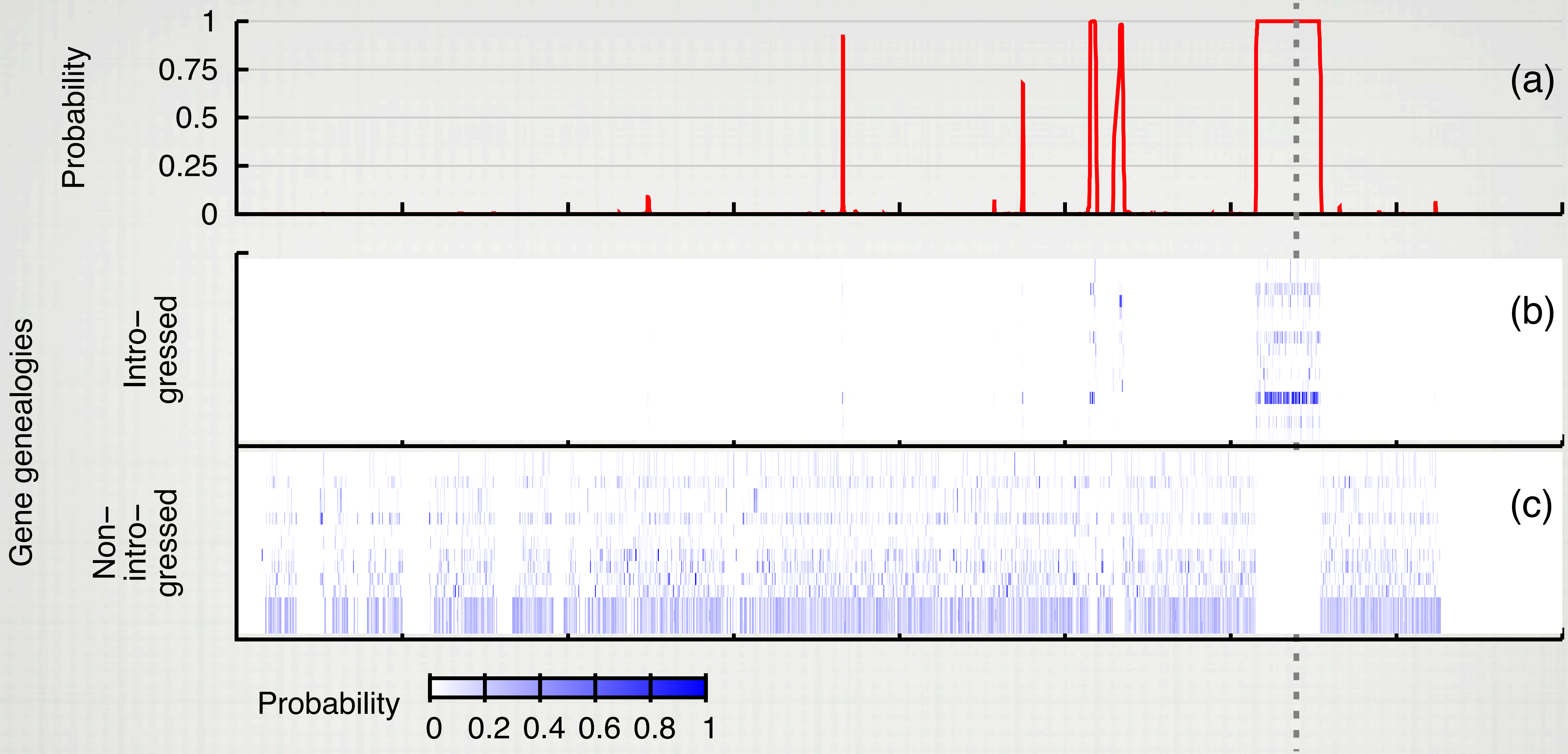
[Liu, Dai, Truong, Song, Kohn, Nakhleh, PLoS Comp Bio, 2014.]

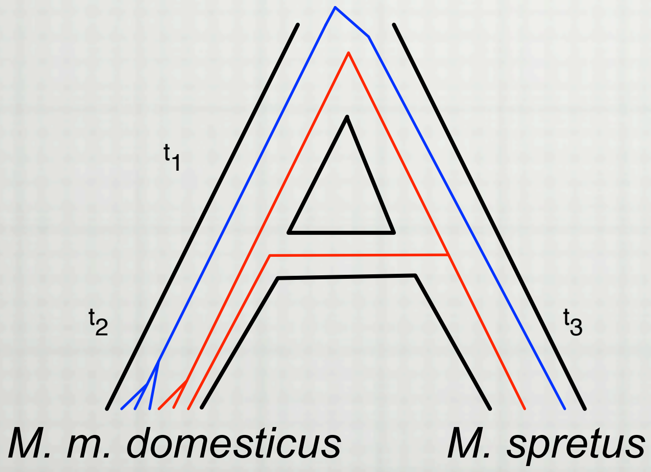
SOLUTION: PHYLONET-HMM



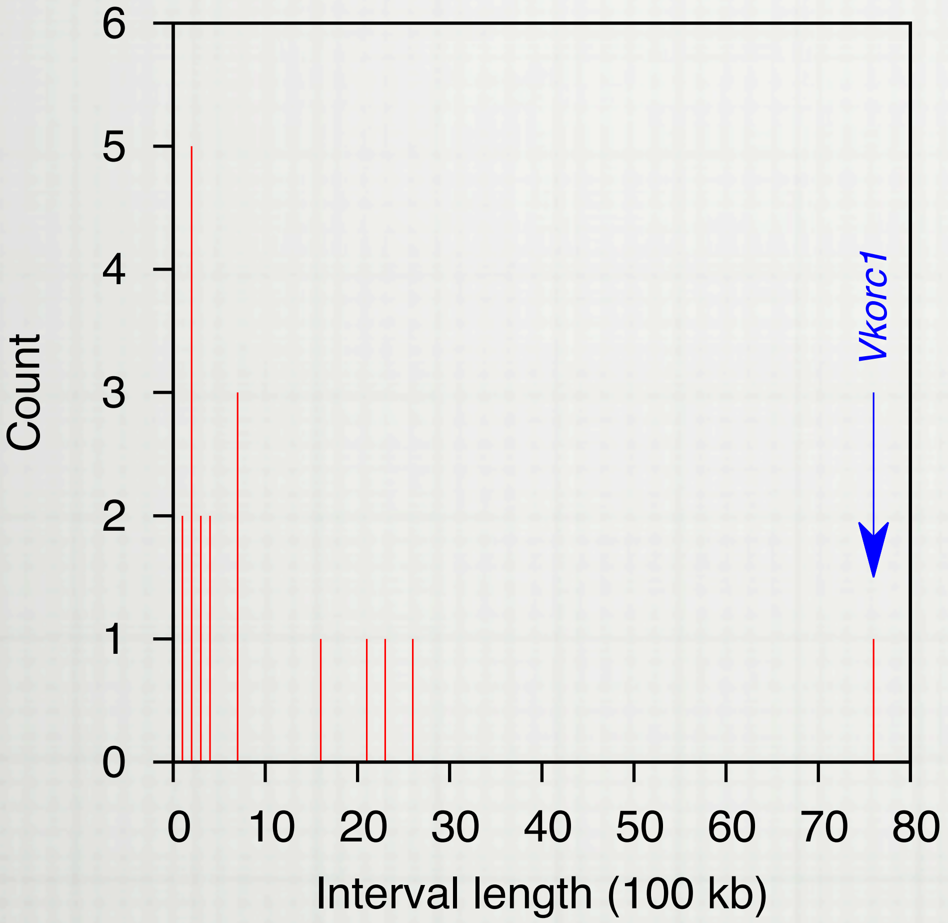
$$\mathbf{P}(\pi_j = (t_x, \psi_y) | \mathcal{G}) = \frac{f_{(t_x,\psi_y)}(j) b_{(t_x,\psi_y)}(j)}{\mathbf{P}(\mathcal{G})}$$

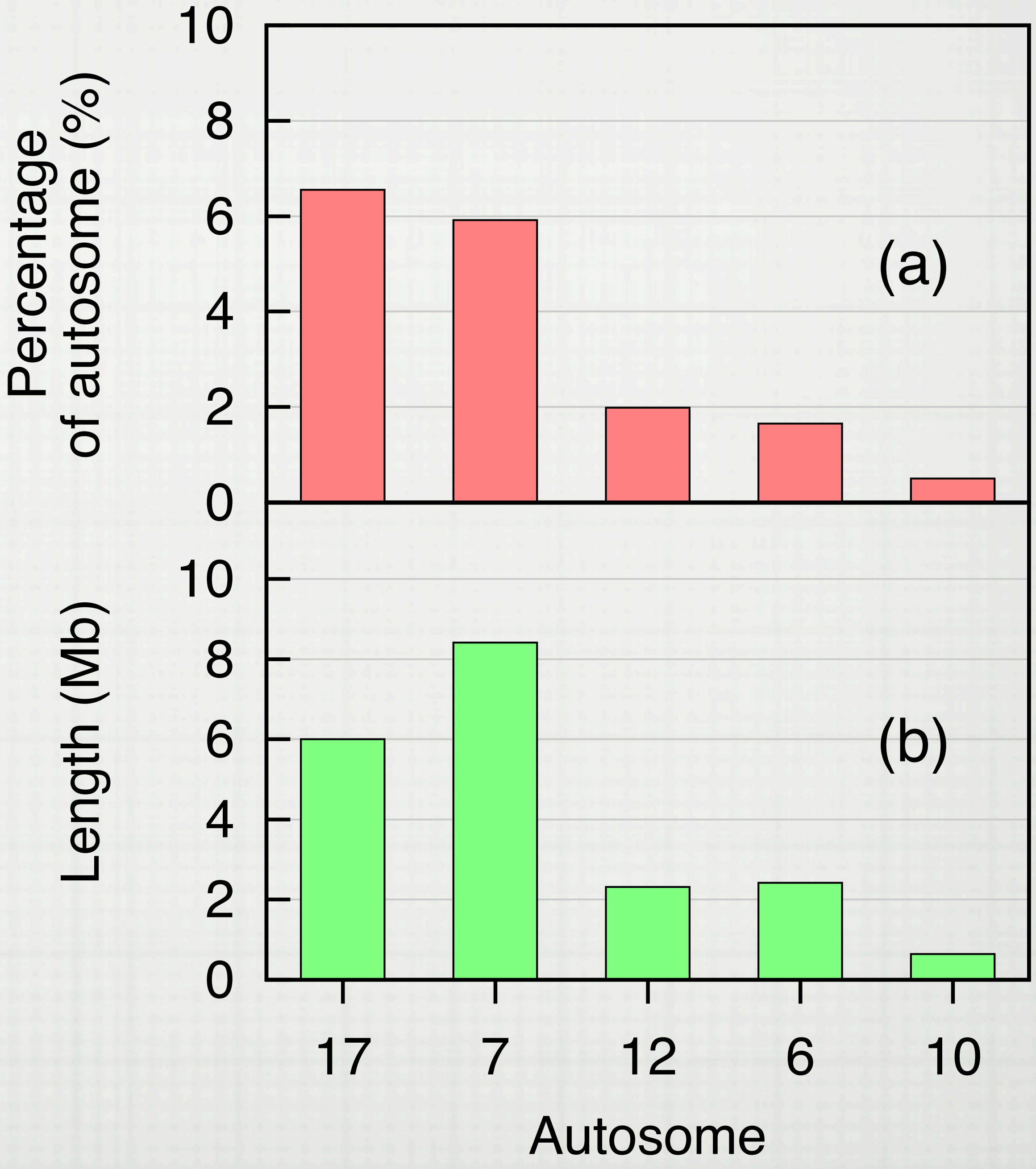
[Liu, Dai, Truong, Song, Kohn, Nakhleh, PLoS Comp Bio, 2014.]





[Liu, Song, Kohn, Nakhleh, Under review, 2014.]





[Liu, Song, Kohn, Nakhleh, Under review, 2014.]

SUMMARY

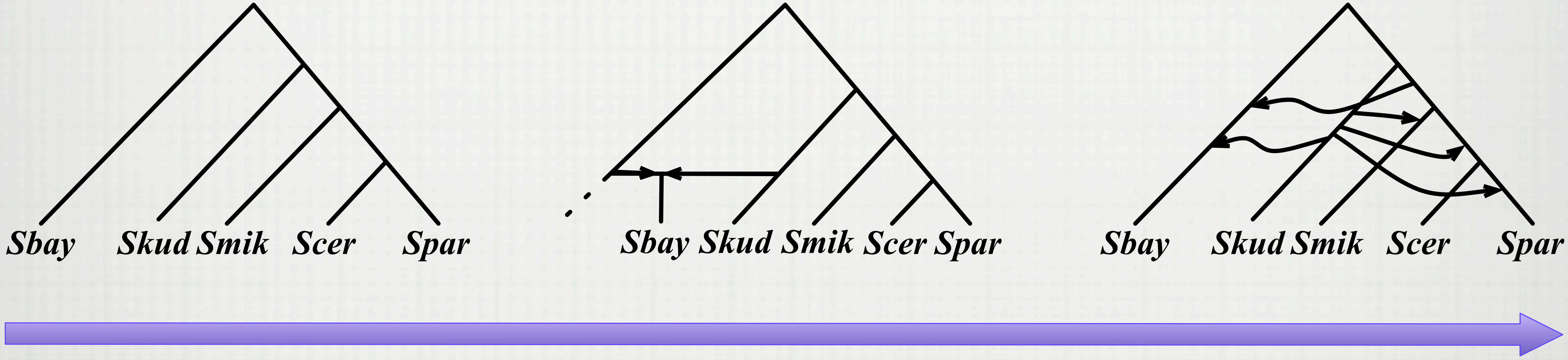
□ Viewing a phylogenetic network as a collection of (MUL-tree,allele mapping) pairs provides a natural way to extend the multi-species coalescent and allows for computing gene tree probabilities in the presence of both ILS and hybridization.

□ This view also allows for extending HMMs to annotate genomes in the presence of introgression.

□ Major challenge: Computational requirements!

□ All methods are implemented in PhyloNet and publicly available in open-source (Java): http://bioinfo.cs.rice.edu/phylonet

SUMMARY



lineage sorting is the sole explanation of all gene tree incongruence

both hybridization and lineage sorting explain gene tree incongruence

hybridization is the sole explanation of all gene tree incongruence

ACKNOWLEDGMENTS

- Collaborators:
  - **R.M. Barnett** (Rice), **J. Dai** (Rice), **J.H. Degnan** (Canterbury), **J. Dong**, (Rice), **K. Liu** (Rice), **K. Truong** (Rice), **Y. Yu** (Rice)
- Funding:
  - NSF, NIH, Alfred P. Sloan Foundation, Guggenheim Foundation

THANK YOU

HTTP://WWW.CS.RICE.EDU/~NAKHLEH