

# Phylogenetic Network Inference with PhyloNet

Luay Nakhleh  
Department of Computer Science  
Rice University

Species Tree Estimation Workshop  
Ohio State University  
4 June 2018

*Syst. Biol.* 0(0):1–7, 2018

© The Author(s) 2018. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.

For permissions, please email: journals.permissions@oup.com

DOI:10.1093/sysbio/syy015

## Inferring Phylogenetic Networks Using PhyloNet

DINGQIAO WEN<sup>1</sup>, YUN YU<sup>1</sup>, JIAFAN ZHU<sup>1</sup>, AND LUAY NAKHLEH<sup>1,2,\*</sup>

<sup>1</sup>Computer Science and <sup>2</sup>BioSciences, Rice University, 6100 Main Street, Houston, TX 77005, USA;

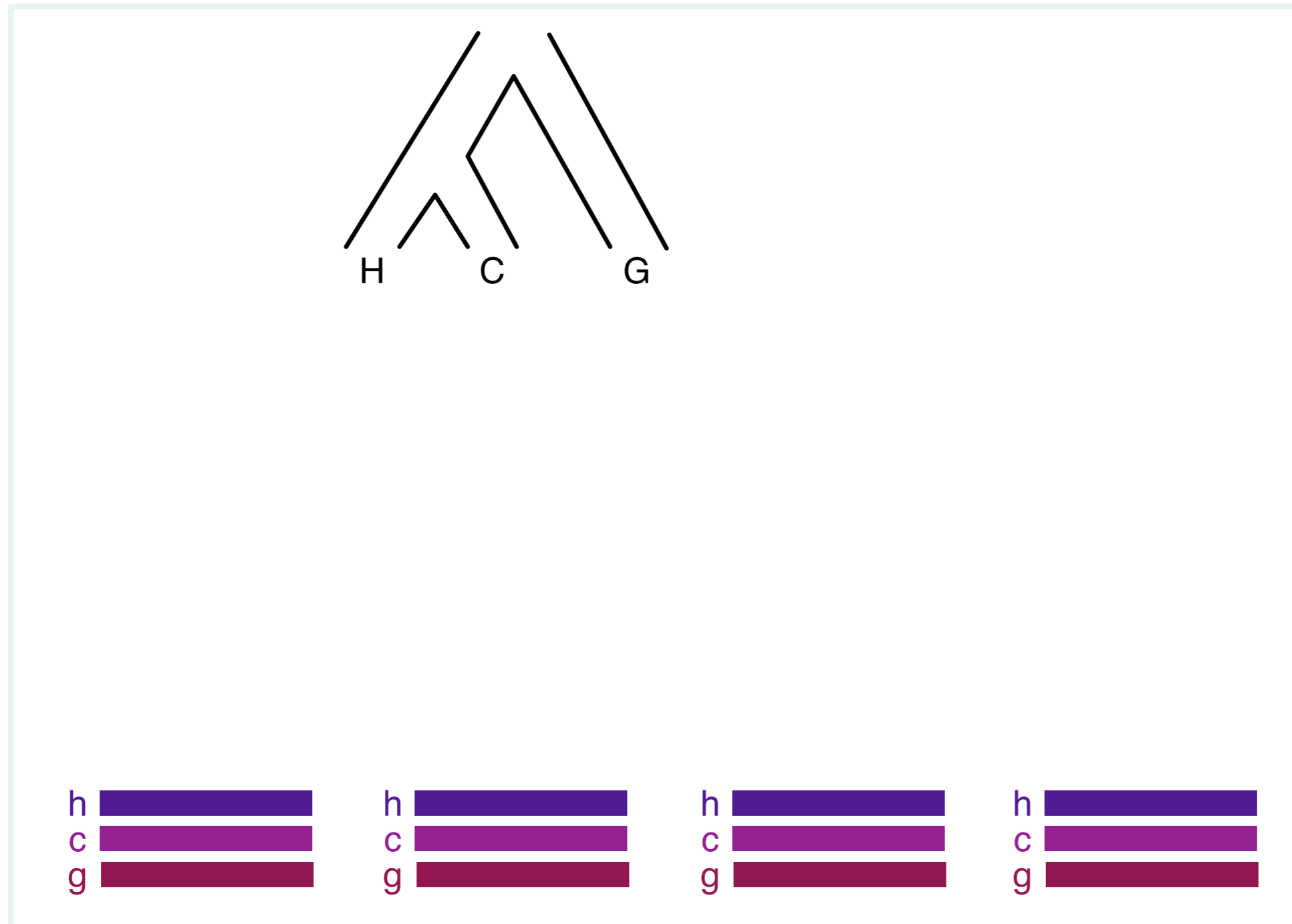
\*Correspondence to be sent to: Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA;  
E-mail: nakhleh@rice.edu.

Received 21 December 2017; reviews returned 20 February 2018; accepted 23 February 2018

Associate Editor: David Posada

*Abstract.*—PhyloNet was released in 2008 as a software package for representing and analyzing phylogenetic networks. At the time of its release, the main functionalities in PhyloNet consisted of measures for comparing network topologies and a single heuristic for reconciling gene trees with a species tree. Since then, PhyloNet has grown significantly. The software package now includes a wide array of methods for inferring phylogenetic networks from data sets of unlinked loci while accounting for both reticulation (e.g., hybridization) and incomplete lineage sorting. In particular, PhyloNet now allows for maximum parsimony, maximum likelihood, and Bayesian inference of phylogenetic networks from gene tree estimates. Furthermore, Bayesian inference directly from sequence data (sequence alignments or biallelic markers) is implemented. Maximum parsimony is based on an extension of the “minimizing deep coalescences” criterion to phylogenetic networks, whereas maximum likelihood and Bayesian inference are based on the multispecies network coalescent. All methods allow for multiple individuals per species. As computing the likelihood of a phylogenetic network is computationally hard, PhyloNet allows for evaluation and inference of networks using a pseudolikelihood measure. PhyloNet summarizes the results of the various analyzes and generates phylogenetic networks in the extended Newick format that is readily viewable by existing visualization software. [Bayesian inference; incomplete lineage sorting; maximum likelihood; maximum parsimony; multispecies network coalescent; phylogenetic networks; reticulation.]

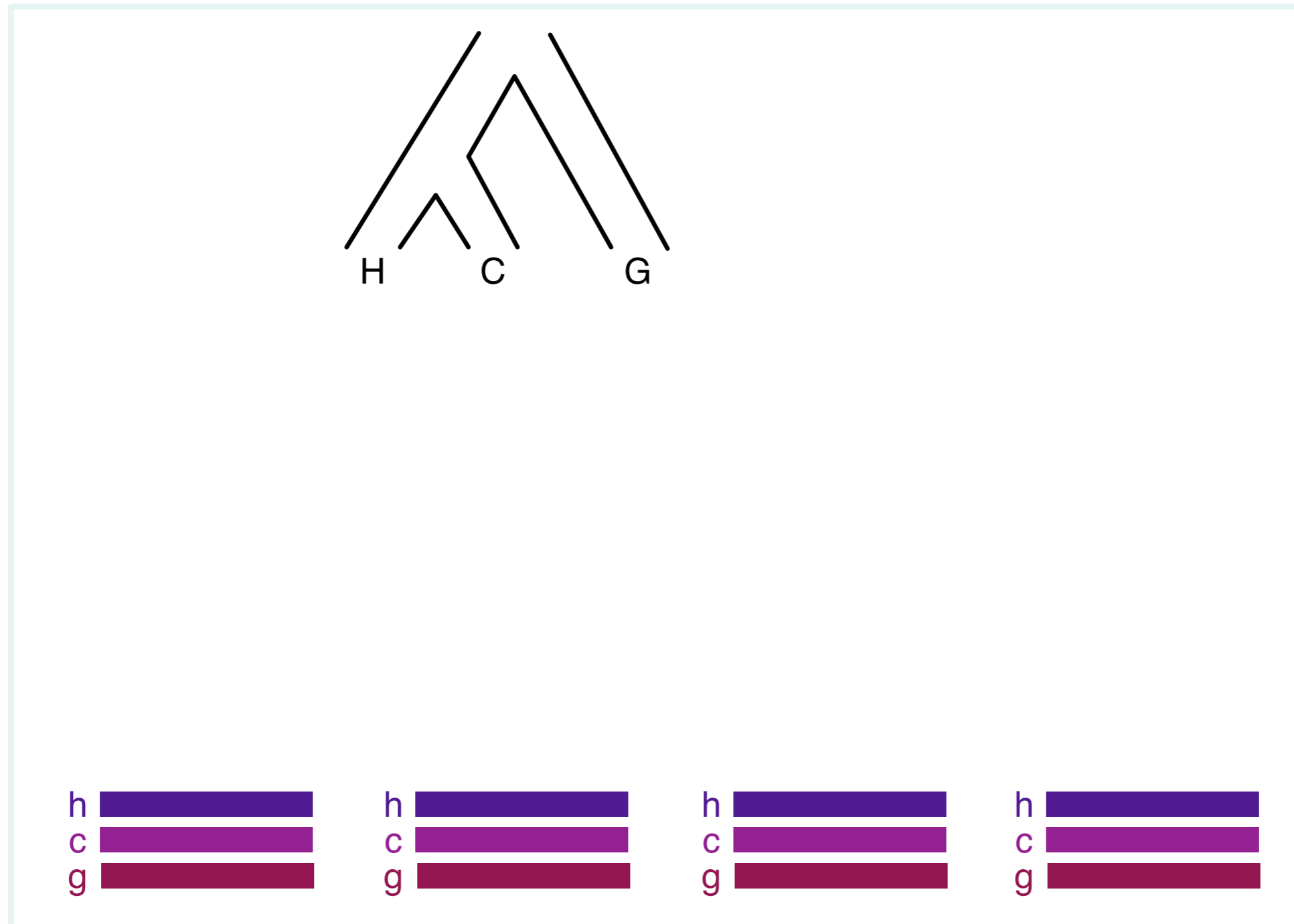
# Phylogenomics



**Generative  
model:  
Species tree  $\Psi$**

**Observed  
data  $S$**

# Phylogenomics



**Generative  
model:  
Species tree  $\Psi$**

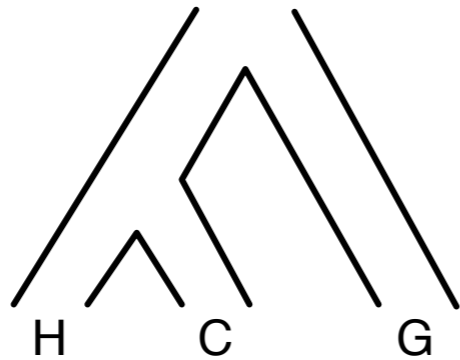
**Observed  
data S**

What is  $p(S | \Psi)$ ?

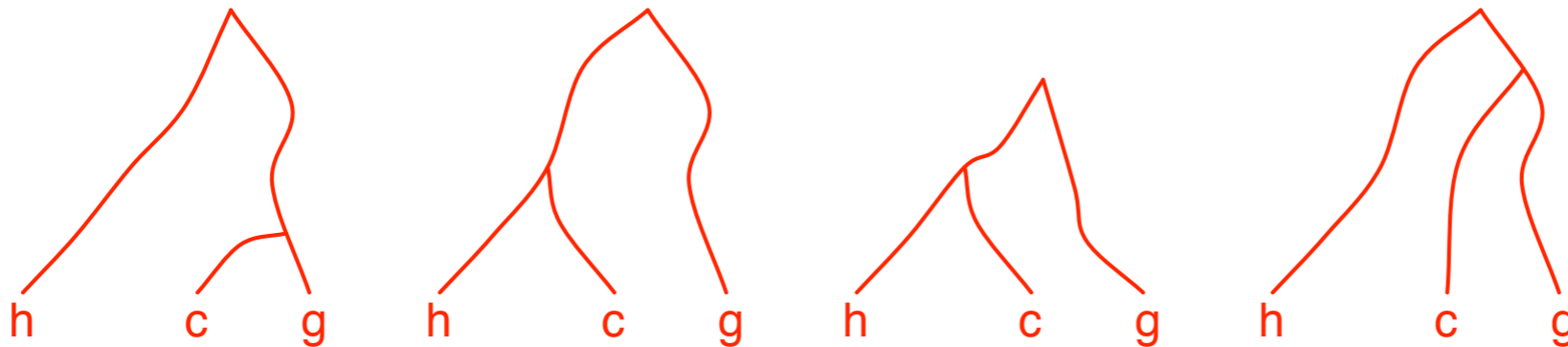
# Phylogenomics

MODEL INFERENCE

DATA GENERATION



Generative  
model:  
Species tree  $\Psi$



Hidden Variables:  
Local gene genealogies  
 $G$

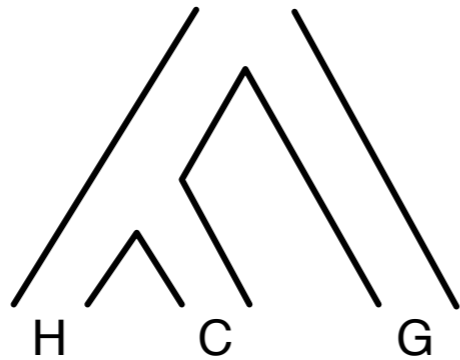


Observed  
data  $S$

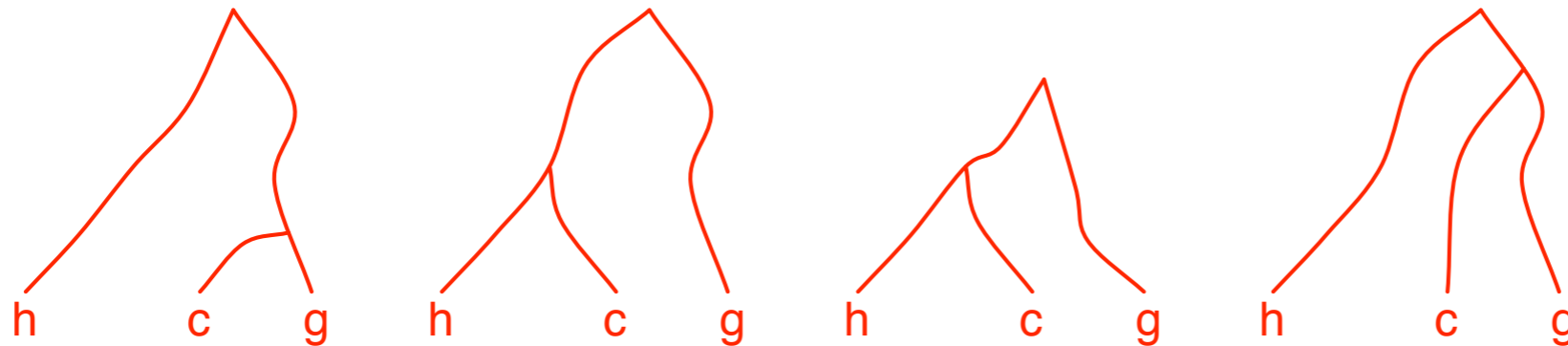
# Phylogenomics

MODEL INFERENCE

DATA GENERATION



Generative  
model:  
Species tree  $\Psi$



Hidden Variables:  
Local gene genealogies  
 $G$



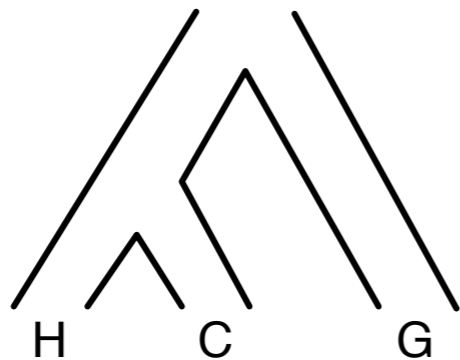
Observed  
data  $S$

$$p(S|\Psi) = \int_G p(S|G)p(G|\Psi)dG$$

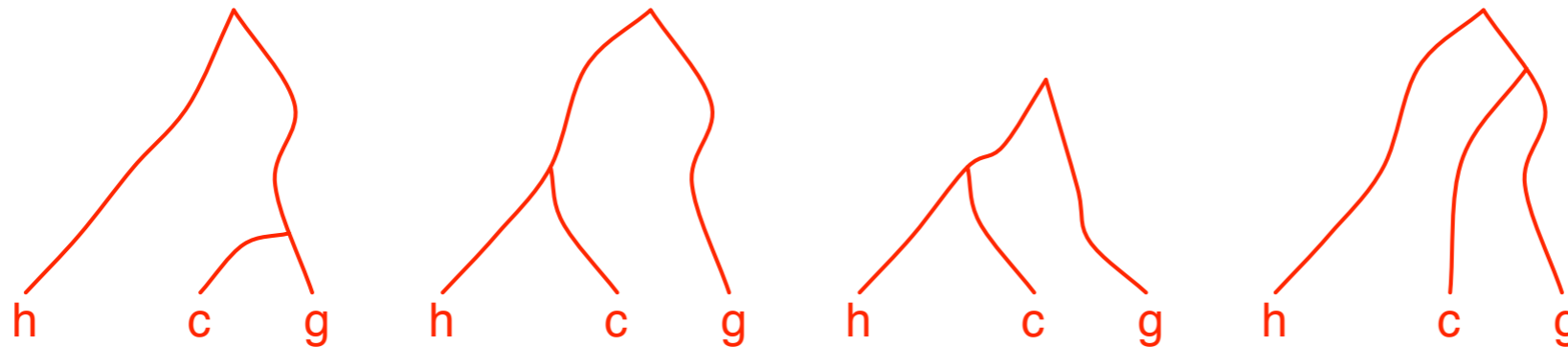
# Phylogenomics

MODEL INFERENCE

DATA GENERATION



Generative  
model:  
Species tree  $\Psi$



Hidden Variables:  
Local gene genealogies  
 $G$



Observed  
data  $S$

Assuming independent loci: 
$$p(S|\Psi) = \prod_i \int_G p(S_i|g)p(g|\Psi)dg$$

# Phylogenomics

$p(S | g)$ : *the “Felsenstein”  
likelihood of gene tree  $g$*

$p(g | \Psi)$ : *the multispecies  
coalescent (MSC)*



- What happens to the model when both **reticulation** (say, hybridization) and **ILS** are **simultaneously** at play?

$$p(S|\Psi) = \prod_i \int_G p(S_i|g)p(g|\Psi)dg$$

$$p(S|\Psi) = \prod_i \int_G p(S_i|g)p(g|\Psi)dg$$

1)  $\Psi$  is now a network, rather than a tree

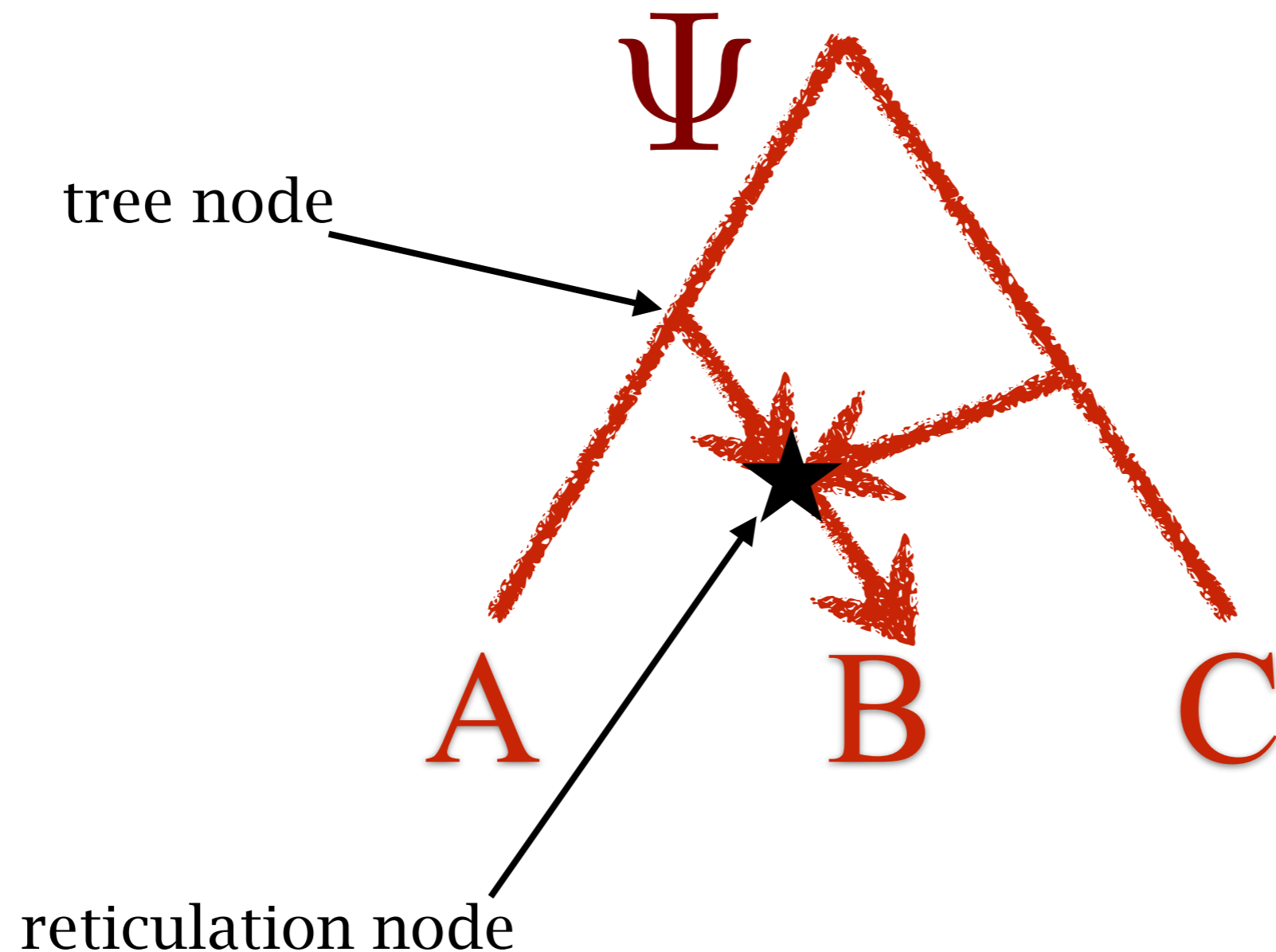
$$p(S|\Psi) = \prod_i \int_G p(S_i|g)p(g|\Psi)dg$$

1)  $\Psi$  is now a network, rather than a tree

2)  $p(g|\Psi)$  is now given by the multispecies network coalescent

# Phylogenetic Networks

A leaf-labeled, rooted, directed, acyclic graph (rDAG)

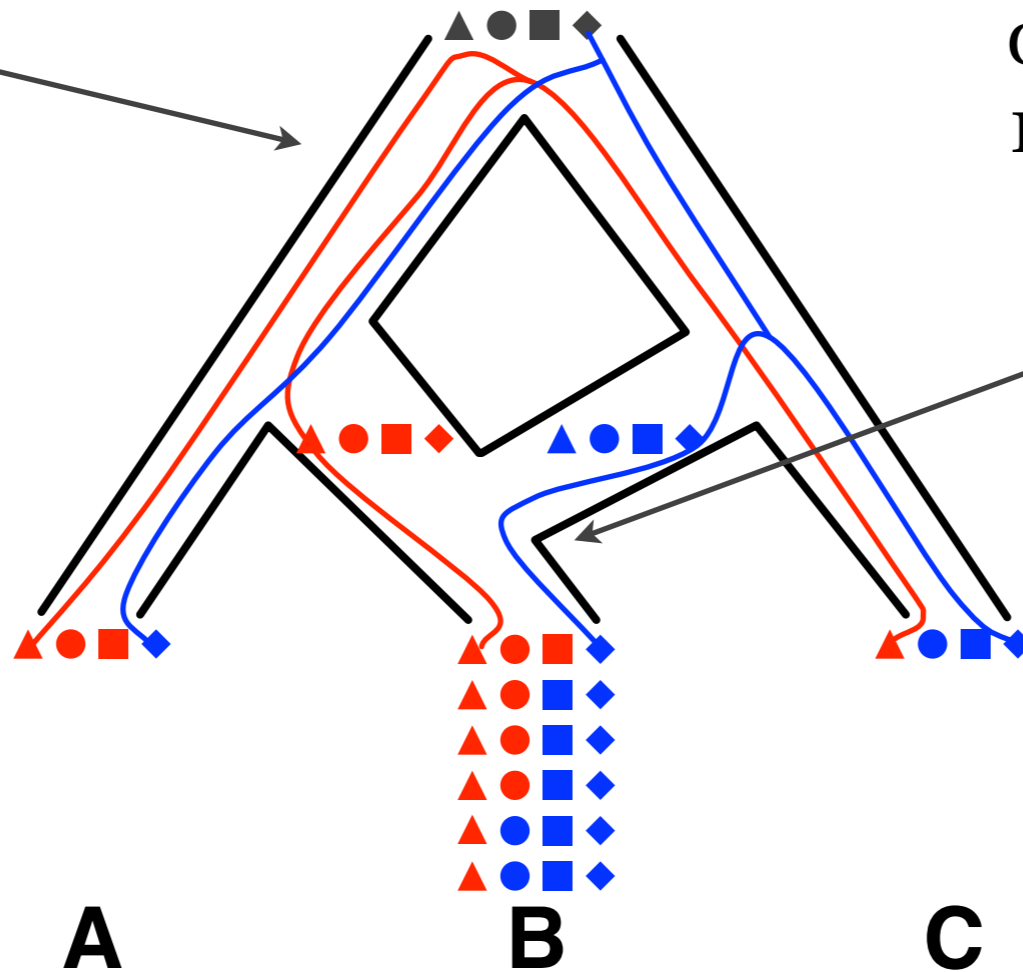


# The Multispecies Network Coalescent

Phylogenetic network with  
branch lengths (in coalescent  
units)

$\Psi$

Inheritance probabilities,  
one per locus per  
reticulation node



# The Multispecies Network Coalescent

OPEN ACCESS Freely available online

PLoS GENETICS

## The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection

Yun Yu<sup>1</sup>, James H. Degnan<sup>2,3</sup>, Luay Nakhleh<sup>1\*</sup>

<sup>1</sup> Department of Computer Science, Rice University, Houston, Texas, United States of America, <sup>2</sup> Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, <sup>3</sup> National Institute of Mathematical and Biological Synthesis, Knoxville, Tennessee, United States of America

### Abstract

Gene tree topologies have proven a powerful data source for various tasks, including species tree inference and species delimitation. Consequently, methods for computing probabilities of gene trees within species trees have been developed

## Maximum likelihood inference of reticulate evolutionary histories

Yun Yu<sup>a,1</sup>, Jianrong Dong<sup>a</sup>, Kevin J. Liu<sup>a,b</sup>, and Luay Nakhleh<sup>a,b,1</sup>

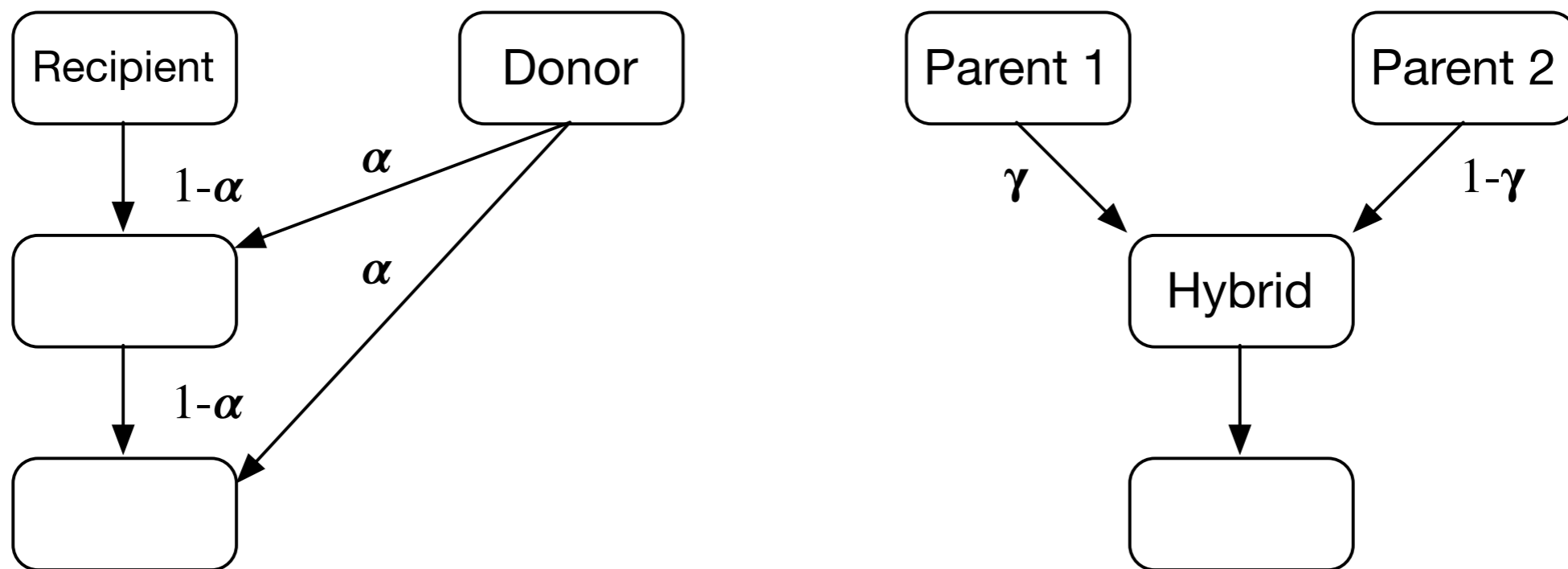
Departments of <sup>a</sup>Computer Science and <sup>b</sup>Ecology and Evolutionary Biology, Rice University, Houston, TX 77005

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved October 7, 2014 (received for review April 30, 2014)

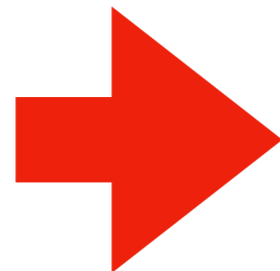
Hybridization plays an important role in the evolution of certain groups of organisms, adaptation to their environments, and diversification of their genomes. The evolutionary histories of such groups are reticulate, and methods for reconstructing them are still in their infancy and have limited applicability. We present a maximum likelihood method for inferring reticulate evolutionary histories while accounting simultaneously for incomplete lineage

To the best of our knowledge, the first method to conduct a search of the phylogenetic network space in search of optimal phylogenies is described in a study by our group (18). However, this method is based on the maximum parsimony criterion: It seeks a phylogenetic network that minimizes the number of “extra lineages” resulting from embedding the set of gene tree topologies within its branches.

- Phylogenetic networks model a continuous epoch of gene flow as one instantaneous event.

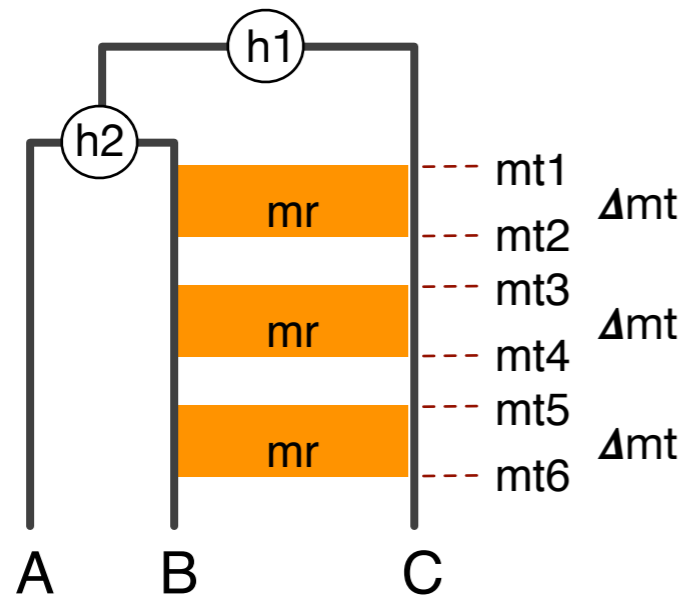


**Reality**

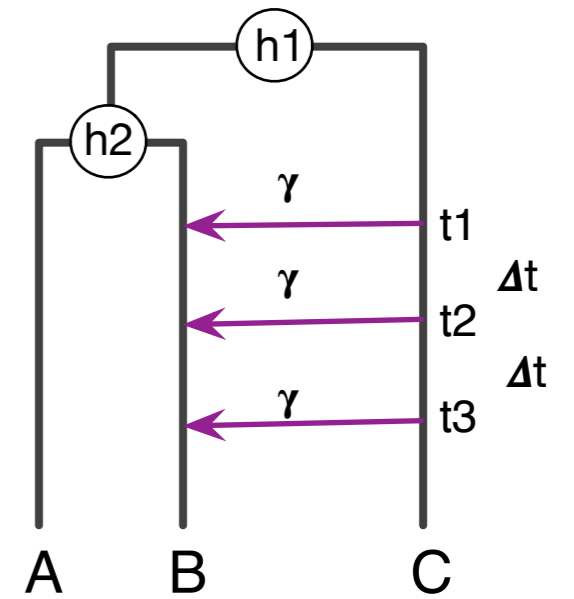


**Model**



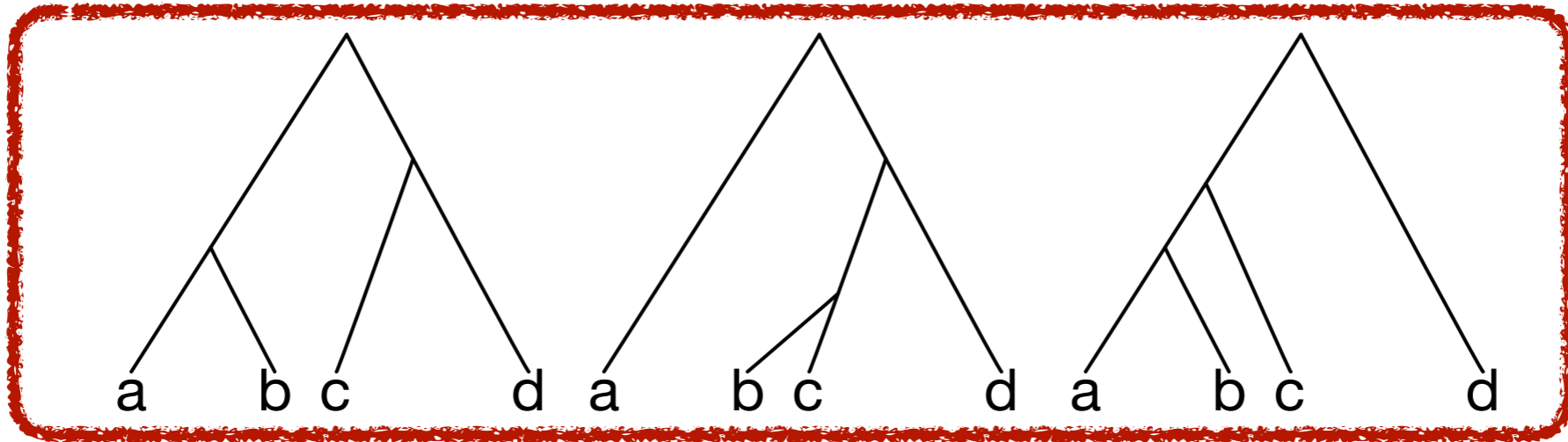


is inferred as

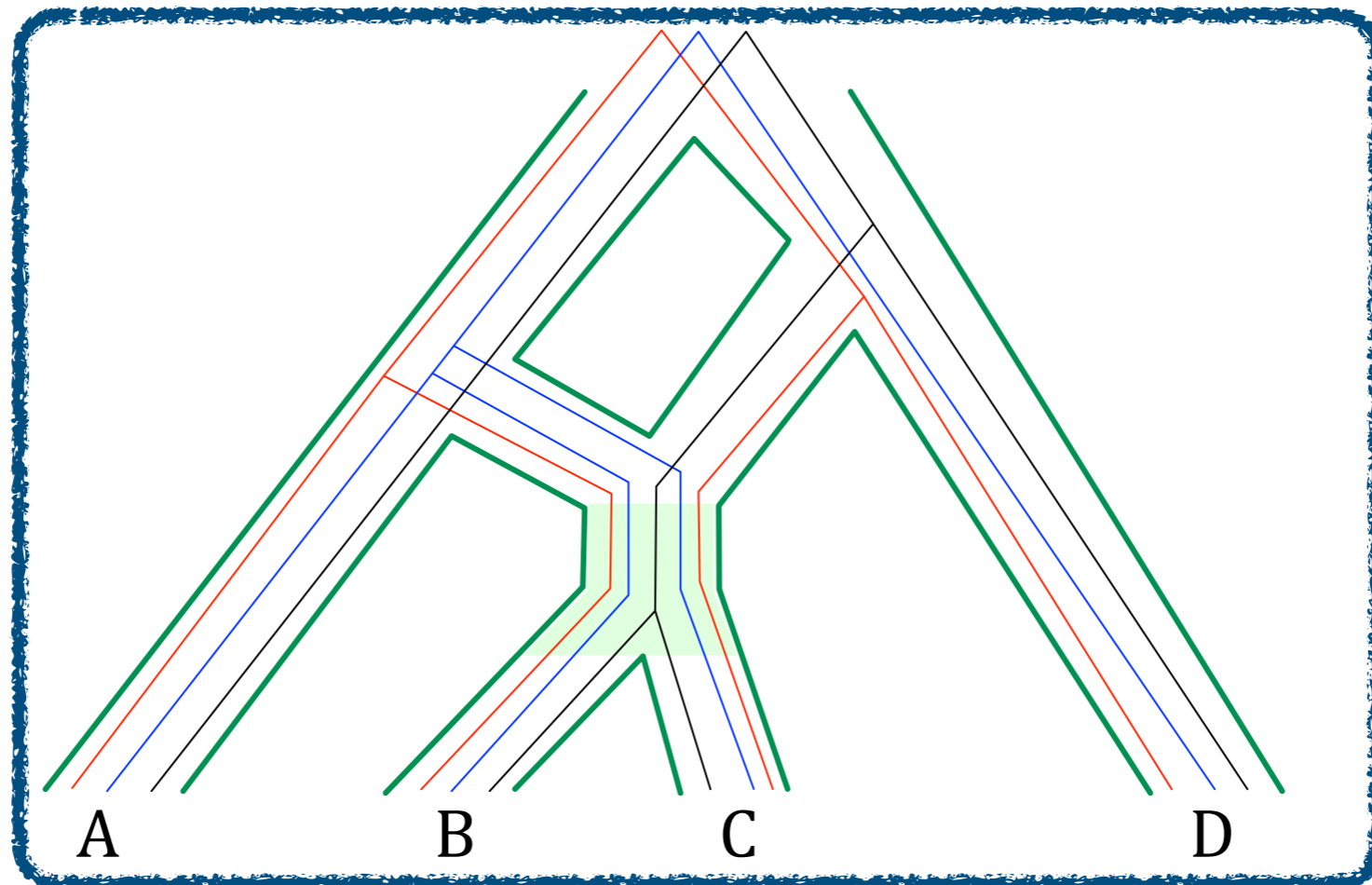


# The MDC (parsimony) Criterion

**Input  
gene  
trees**

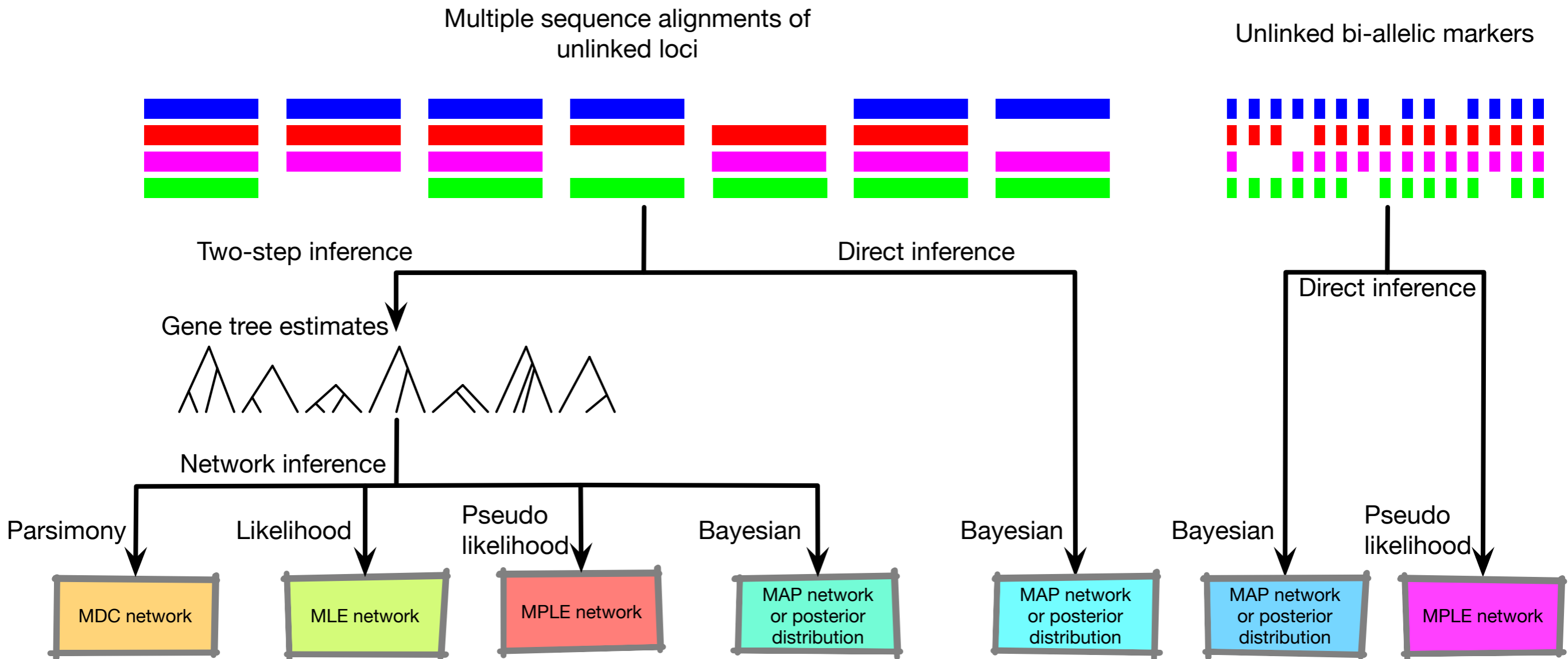


**Network with  
1 reticulation  
and  
2 extra lineages**

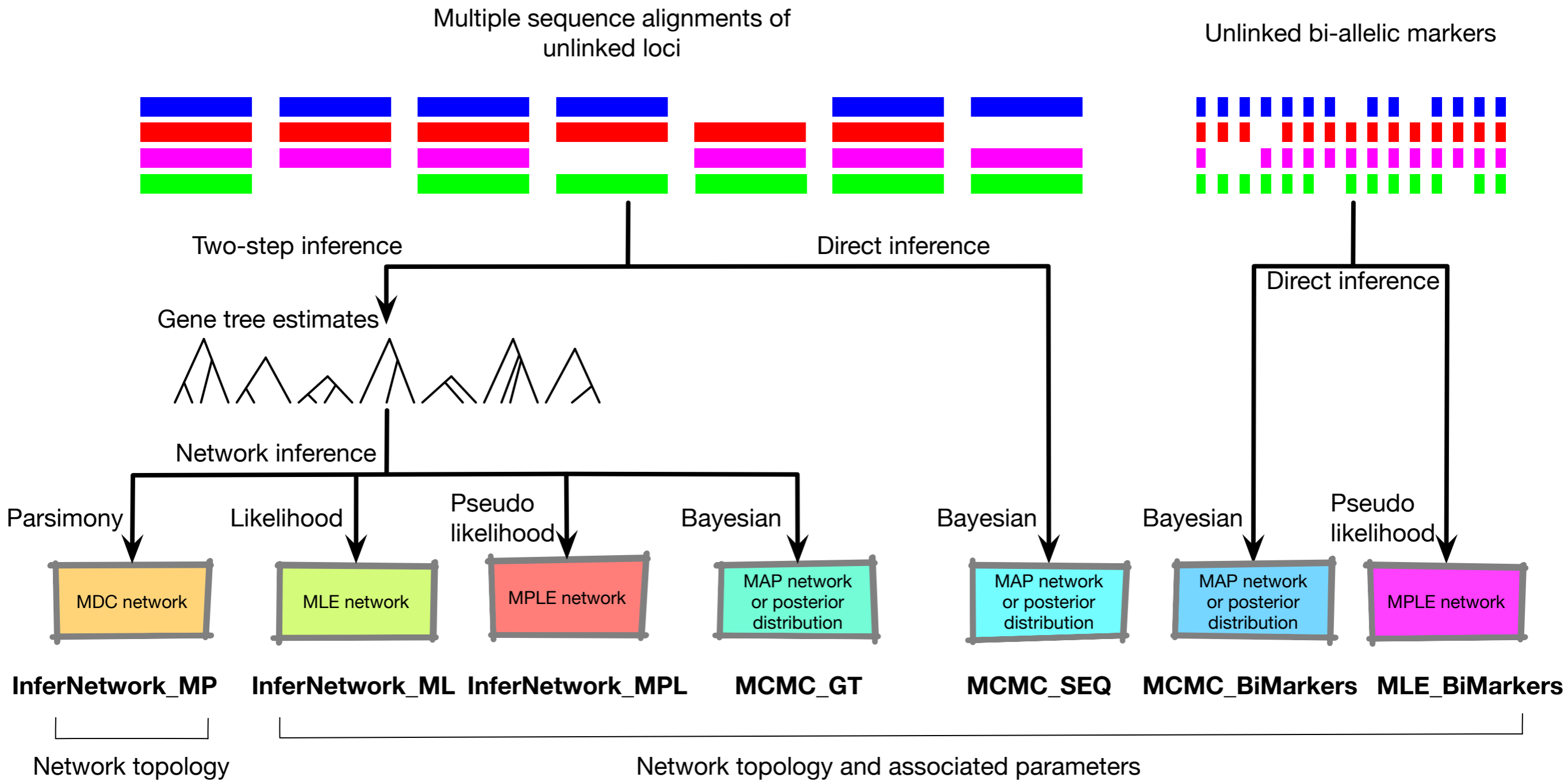


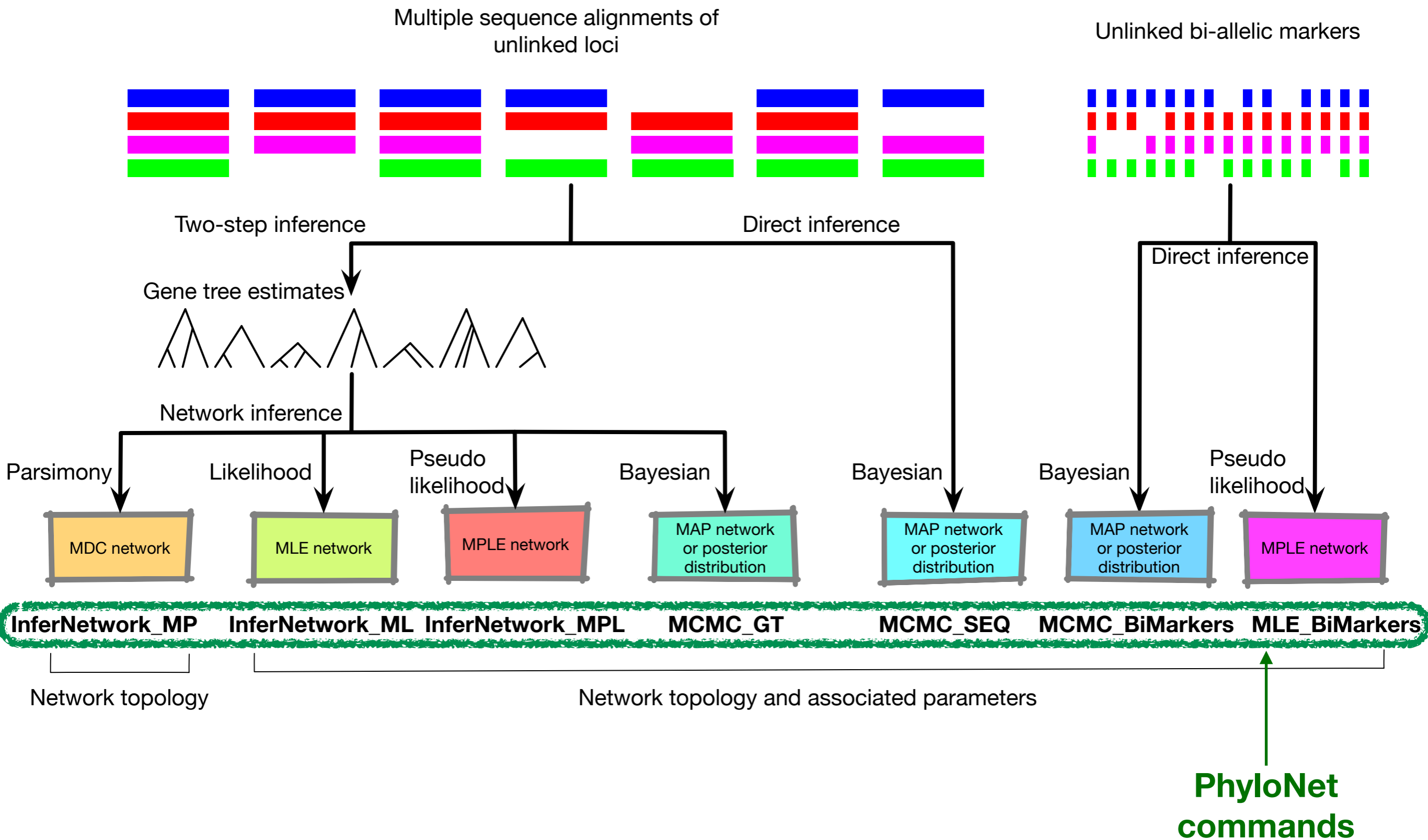
# Phylogenetic Network

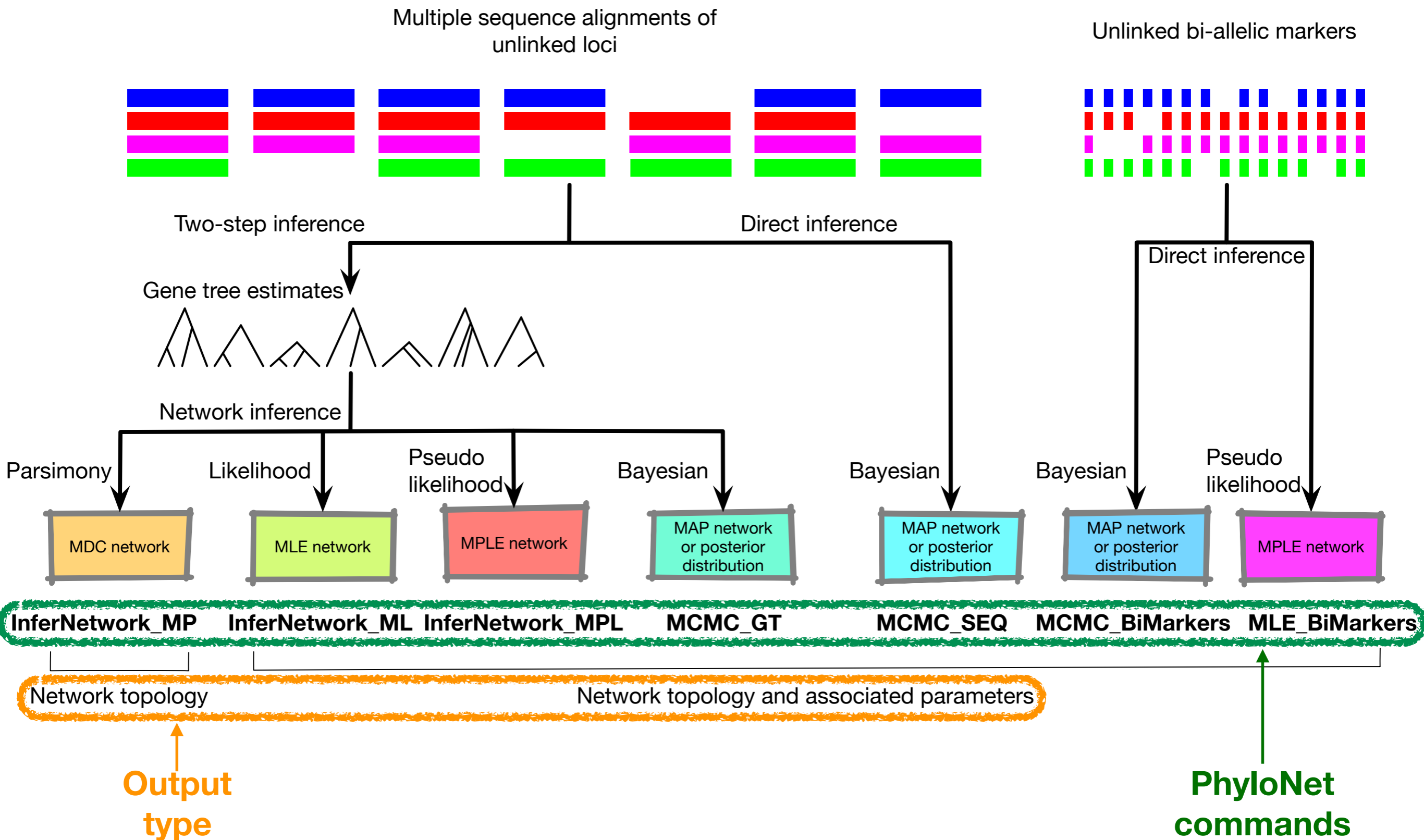
## Inference: Data and Methods



- PhyloNet is a software package that implements all these inference methods.







Name	Description	Ref	Since
<b>Methods for Species Network (and Tree) Inference (all account for ILS)</b>			
<a href="#">MCMC_SEQ</a>	Bayesian MCMC posterior estimation of phylogenetic networks and gene trees on sequences from multiple independent loci.	<a href="#">here</a>	3.6.1
<a href="#">MCMC_BiMarkers</a>	Bayesian estimation of the posterior distribution of phylogenetic networks given bi-allelic genetic markers (SNPs, AFLPs, etc).	<a href="#">here</a>	3.6.1
<a href="#">MCMC_GT</a>	Bayesian MCMC posterior estimation of phylogenetic networks given a list of gene tree topologies.	<a href="#">here</a>	3.6.0
<a href="#">MLE_BiMarkers</a>	Maximum (pseudo-)likelihood estimation of phylogenetic networks given bi-allelic genetic markers (SNPs, AFLPs, etc).	<a href="#">here</a>	3.6.4
<a href="#">InferNetwork_MPL</a>	Infers a phylogenetic network from gene trees under maximum pseudo-likelihood.	<a href="#">here</a>	3.5.5
<a href="#">InferNetwork_ML_Bootstrap</a>	Infers a phylogenetic network from gene trees under maximum likelihood with parametric bootstrap.	<a href="#">here</a>	3.5.2
<a href="#">InferNetwork_ML_CV</a>	Infers a phylogenetic network from gene trees under maximum likelihood with cross-validation.	<a href="#">here</a>	3.5.2
<a href="#">InferNetwork_ML</a>	Infers a phylogenetic network from gene trees under maximum likelihood.	<a href="#">here</a>	3.4.0
<a href="#">InferNetwork_MP</a>	Infers a phylogenetic network from gene trees under the MDC criterion.	<a href="#">here</a>	3.4.0



- Since a tree is a special case of network (a network with zero reticulation nodes), all these methods can be used to **infer species trees**
- Simply **set the maximum number of reticulations to 0** and the methods will search the tree (not network) space!

- But, PhyloNet also has tree-specific methods:

Methods for Species Tree (not Networks) Inference			
<a href="#">Infer_ST_Bootstrap</a>	Infers a species tree using bootstrap with existing <code>Infer_ST</code> commands.		3.0.0
<a href="#">Infer_ST_DV</a>	Infers a species tree from gene trees using the "Democratic Vote" method.		3.0.0
<a href="#">Infer_ST_GLASS</a>	Infers a species tree using the GLASS method of Mossel and Roch.	<a href="#">here</a>	3.0.0
<a href="#">Infer_ST_MC</a>	Infers a species tree from gene trees using greedy consensus (allows for gene trees with multiple alleles in species and for unrooted gene trees).		3.0.0
<a href="#">Infer_ST_MDC</a>	Infers a species tree from gene tree topologies using the "Minimize Deep Coalescence" (MDC) criterion.	<a href="#">here</a>	3.0.0
<a href="#">Infer_ST_MDC_Time</a>	Infers a species tree from gene trees with coalescent times using the MDC criterion.		3.0.0
<a href="#">Infer_ST_MDC_UR</a>	Infers a species tree from unrooted gene tree topologies using the MDC criterion.	<a href="#">here</a>	3.0.0
<a href="#">GenCPLEX</a>	Generates CPLEX input for a species tree and a set of gene trees.	<a href="#">here</a>	3.0.0
<a href="#">GenST</a>	Generates species tree topologies based on maximal sets of compatible clusters.	<a href="#">here</a>	3.0.0

- Phylogenetic network inference is computationally very hard.
- All the methods in PhyloNet are heuristics.
- (This answers the question “Why did different runs return different networks?”)

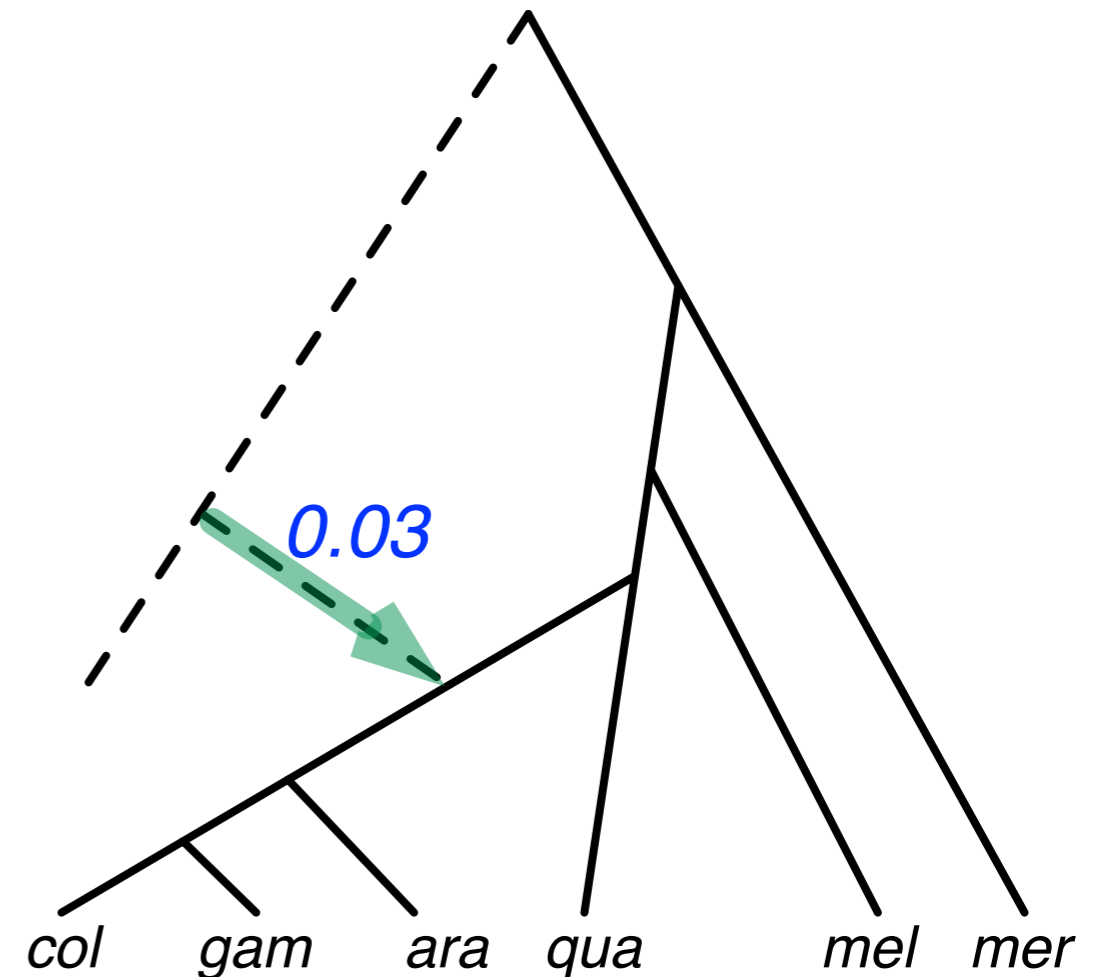
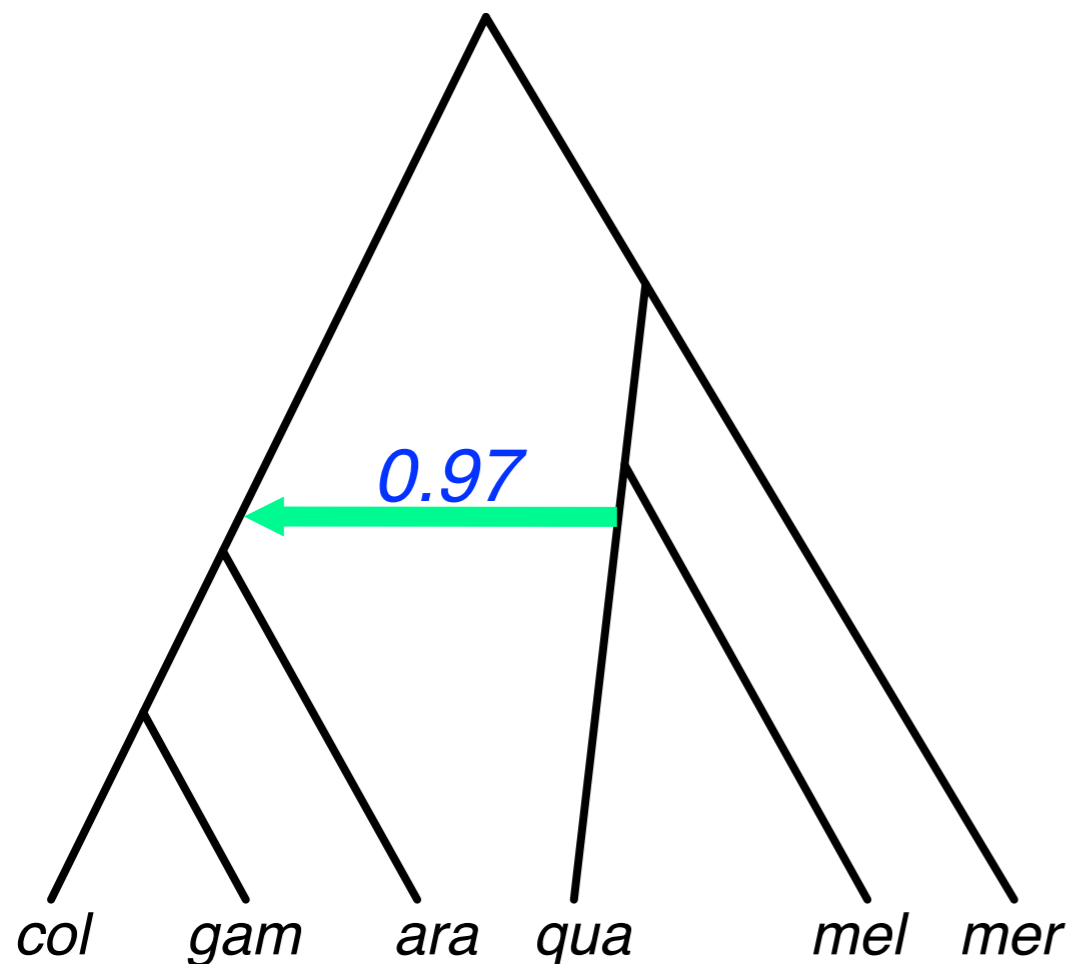
- **NO**, PhyloNet **does not** designate a species tree and search for reticulations to add to it!
- PhyloNet searches the space of phylogenetic networks; the “species tree” inside the network is in the eyes of the beholder.

SPECIAL ISSUE: GENOMICS OF HYBRIDIZATION

## Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis

DINGQIAO WEN,\* YUN YU,\* MATTHEW W. HAHN†‡ and LUAY NAKHLEH\*§

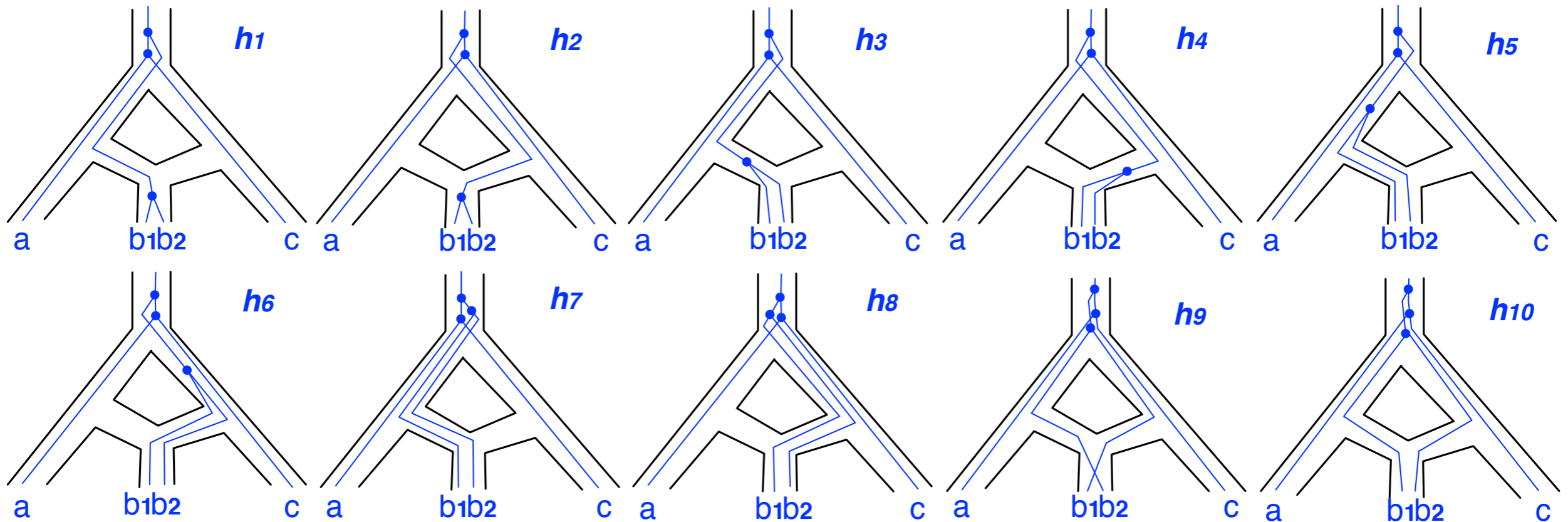
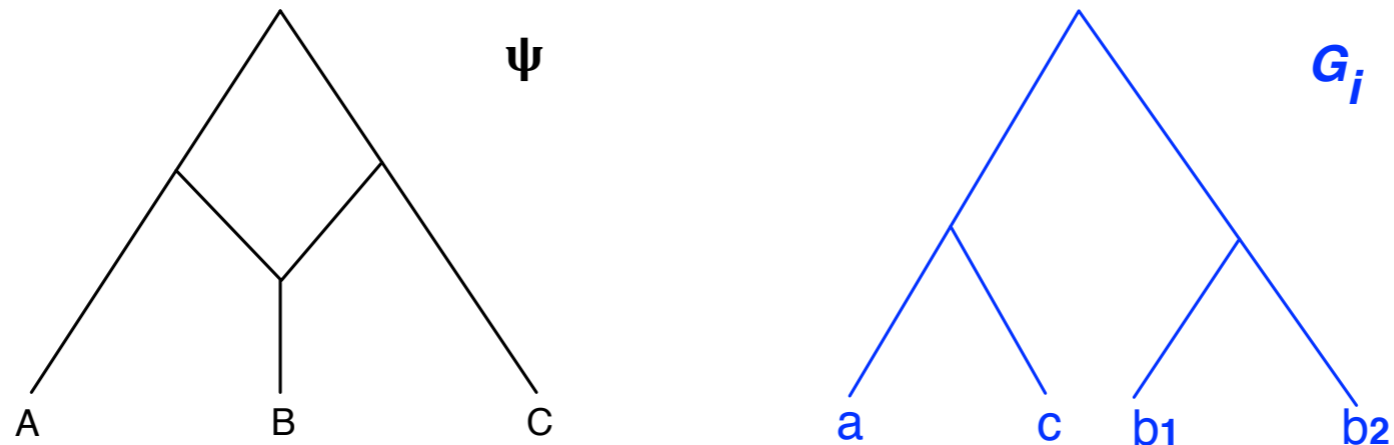
\*Department of Computer Science, Rice University, Houston, TX 77005, USA, †Department of Biology, Indiana University, Bloomington, IN 47405, USA, ‡School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA, §Department of BioSciences, Rice University, Houston, TX 77005, USA



# On the Number of Reticulations

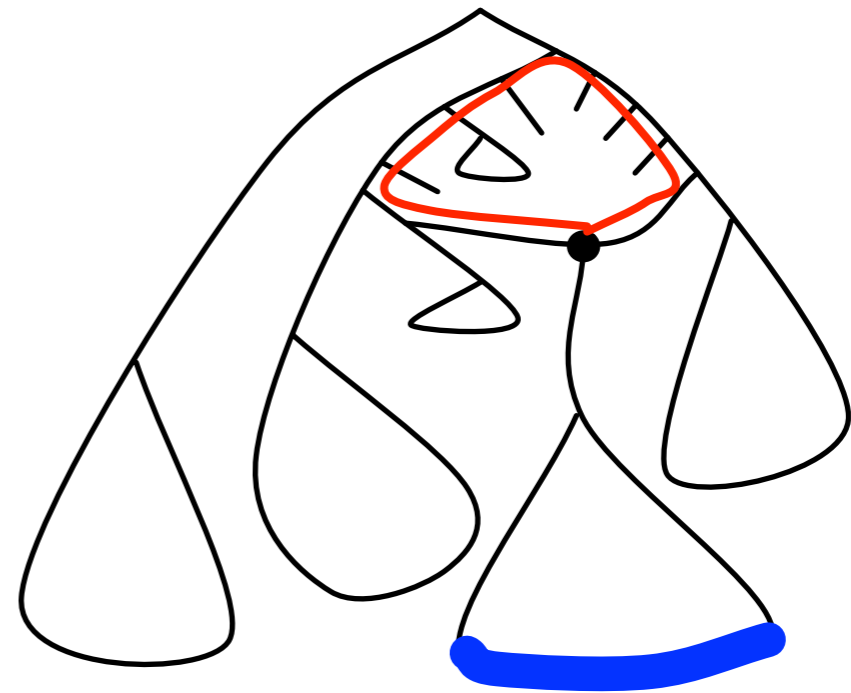
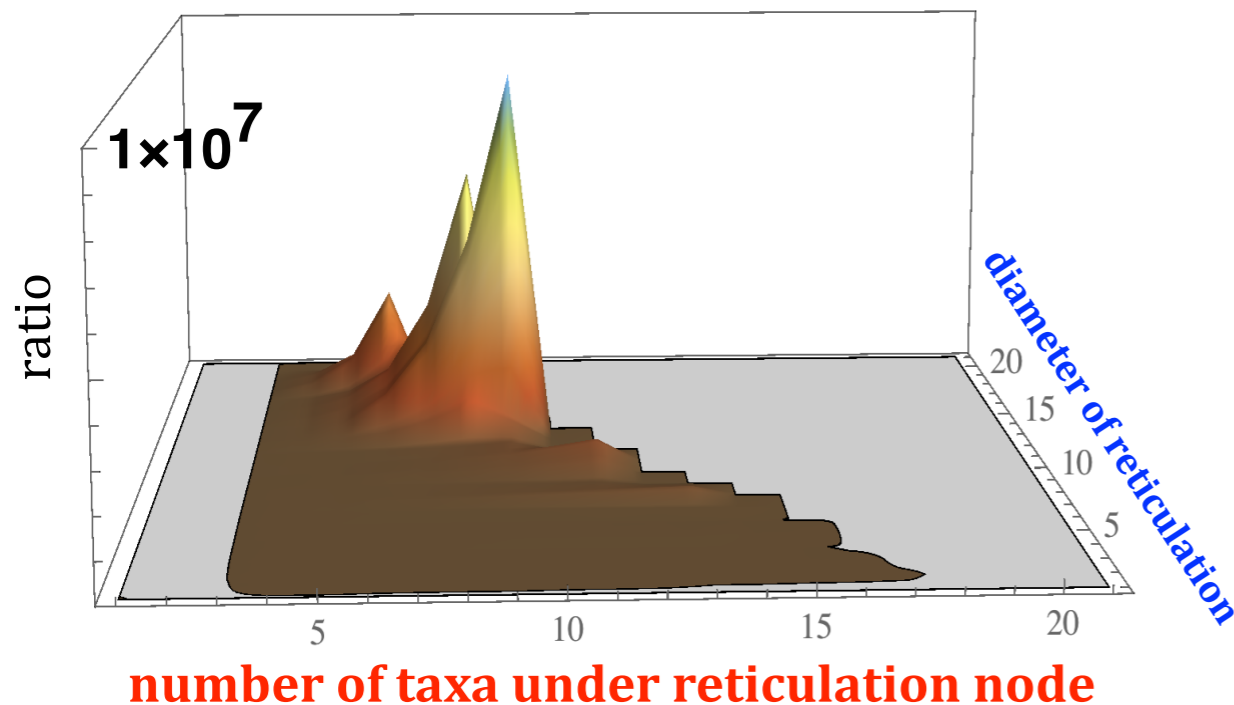
- MDC and MLE based inferences favor more complex networks.
- Recommendation: Search for networks with increasing numbers of reticulations and examine the improvements in the MDC score and likelihood, respectively, of the inferred networks.
- The prior in Bayesian inference penalizes model complexity.

# Why Are Networks Computationally More Challenging than Trees?



$$H_{\psi}(G_i)$$

# Why Are Networks Computationally More Challenging than Trees?



The size of  $H_\psi(g)$  increases by **7 orders of magnitude** by adding just **one reticulation event** to a species tree!!!



# On the Number of Reticulations

- In general, I recommend limiting the number of reticulations in the analysis since it has a huge negative impact on the computational complexity.
- Pseudo-likelihood is hardly sensitive to the number of reticulations, though.


# Individuals Per Species

- All methods allow data from multiple individuals per species (but that further adds to the computational complexity).
- Missing data (as in missing an entire sequence for a certain locus) is also handled.

```
Begin data;
  Dimensions ntax=5 nchar=108;
  Format datatype=dna symbols="ACTG" missing=? gap=-;
  Matrix
[loci1, 53, ...]
a1  ATTGGAGACRAGCGARGACCGAGCTCACGAACCTGAGGAATGGAATCGATTAC
a2  ATTTGAGACRAGCGARGACCGAGCTCACGAACCTGAGGANTGGAATCGATTAC
b1  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTAAGGATTGGAATCGATTGT
b2  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTGAGGATTGGAATCGATTGT
[loci2, 58, ...]
a2  ACTTTGCAAGCCAAAAATGGTATGCGAGACAACGCCTGTCATGGATGATGAACCAGAT
b1  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
b2  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
c1  GCTTTGRAAGRCAAAAAATGATATGCGAAACAACGCCCGTGATGGACGATGAACAGGAT
;End;
BEGIN PHYLONET;
MCMC_SEQ -loci (loci1,loci2) -c1 5000000 -b1 1000000 -tm <A:a1,a2; B:b1,b2; C:c1>;
END;
```

```
Begin data;
  Dimensions ntax=5 nchar=108;
  Format datatype=dna symbols="ACTG" missing=? gap=-;
  Matrix
[loci1, 53, ...]
a1  ATTGGAGACRAGCGARGACCGAGCTCACGAACCTGAGGAATGGAATCGATTAC
a2  ATTTGAGACRAGCGARGACCGAGCTCACGAACCTGAGGANTGGAATCGATTAC
b1  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTAAGGATTGGAATCGATTGT
b2  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTGAGGATTGGAATCGATTGT
[loci2, 58, ...]
a2  ACTTTGCAAGCCAAAAATGGTATGCGAGACAACGCCTGTCATGGATGATGAACCAGAT
b1  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
b2  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
c1  GCTTTGRAAGRCAAAAAATGATATGCGAAACAACGCCCGTGATGGACGATGAACAGGAT
;End;
BEGIN PHYLONET;
MCMC_SEQ -loci (loci1,loci2) -c1 5000000 -b1 1000000 -tm <A:a1,a2; B:b1,b2; C:c1>;
END;
```

**Bayesian inference  
directly from the  
sequence data**



```

Begin data;
  Dimensions ntax=5 nchar=108;
  Format datatype=dna symbols="ACTG" missing=? gap=-;
  Matrix
[loci1, 53, ...]
a1  ATTGGAGACRAGCGARGACCGAGCTCACGAACCTGAGGAATGGAATCGATTAC
a2  ATTTGAGACRAGCGARGACCGAGCTCACGAACCTGAGGANTGGAATCGATTAC
b1  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTAAGGATTGGAATCGATTGT
b2  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTGAGGATTGGAATCGATTGT
[loci2, 58, ...]
a2  ACTTTGCAAGCCAAAAATGGTATGCGAGACAACGCCTGTCATGGATGATGAACCAGAT
b1  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
b2  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
c1  GCTTTGRAAGRCAAAAAATGATATGCGAAACAACGCCCGTGATGGACGATGAACAGGAT
;End;
BEGIN PHYLONET;
MCMC_SEQ -loci (loci1,loci2) -cl 5000000 -bl 1000000 -tm <A:a1,a2; B:b1,b2; C:c1>;
END;

```

**Bayesian inference  
directly from the  
sequence data**

**Mapping individuals  
to species**

```

Begin data;
  Dimensions ntax=5 nchar=108;
  Format datatype=dna symbols="ACTG" missing=? gap=-;
  Matrix
[loci1, 53, ...]
a1  ATTGGAGACRAGCGARGACCGAGCTCACGAACCTGAGGAATGGAATCGATTAC
a2  ATTTGAGACRAGCGARGACCGAGCTCACGAACCTGAGGANTGGAATCGATTAC
b1  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTAAGGATTGGAATCGATTGT
b2  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTGAGGATTGGAATCGATTGT
[loci2, 58, ...]
a2  ACTTTGCAAGCCAAAAATGGTATGCGAGACAACGCCTGTCATGGATGATGAACCAGAT
b1  GCTTTGCAAGCCTAAGATGGTTTGGCGAGACGACGATGGCAGTCGACGATGAATCAGAC
b2  GCTTTGCAAGCCTAAGATGGTTTGGCGAGACGACGATGGCAGTCGACGATGAATCAGAC
c1  GCTTTGRAAGRCAAAAAATGATATGCGAAACAACGCCCGTGATGGACGATGAACAGGAT
;End;
BEGIN PHYLONET;
MCMC_SEQ -loci (loci1,loci2) -cl 5000000 -bl 1000000 -tm <A:a1,a2; B:b1,b2; C:c1>;
END;

```

Locus 1:  
2 individuals from A  
2 individuals from B  
0 individuals from C

Bayesian inference  
directly from the  
sequence data

Mapping individuals  
to species

```

Begin data;
  Dimensions ntax=5 nchar=108;
  Format datatype=dna symbols="ACTG" missing=? gap=-;
  Matrix
[loci1, 53, ...]
a1  ATTGGAGACRAGCGARGACCGAGCTCACGAACCTGAGGAATGGAATCGATTAC
a2  ATTTGAGACRAGCGARGACCGAGCTCACGAACCTGAGGANTGGAATCGATTAC
b1  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTAAGGATTGGAATCGATTGT
b2  TTGGGAGACGAGCGAAGACAGAGCATATGAGCCTGAGGATTGGAATCGATTGT
[loci2, 58, ...]
a2  ACTTTGCAAGCCAAAAATGGTATGCGAGACAACGCCTGTCATGGATGATGAACCAGAT
b1  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
b2  GCTTTGCAAGCCTAAGATGGTTTTCGAGACGACGATGGCAGTCGACGATGAATCAGAC
c1  GCTTTGRAAGRCAAAAAATGATATGCGAAACAACGCCCGTGATGGACGATGAACAGGAT
;End;
BEGIN PHYLONET;
MCMC_SEQ -loci (loci1,loci2) -cl 5000000 -bl 1000000 -tm <A:a1,a2; B:b1,b2; C:c1>;
END;

```

**Locus 1:**  
 2 individuals from A  
 2 individuals from B  
 0 individuals from C

**Locus 3:**  
 1 individual from A  
 2 individuals from B  
 1 individual from C

**Bayesian inference  
 directly from the  
 sequence data**

**Mapping individuals  
 to species**

# Other Useful Functionalities in PhyloNet

## Methods for Simulating Locus Data on Phylogenetic Networks (and Trees)

<a href="#">SimGTinNetwork</a>	Simulates gene trees under the multispecies network coalescent (automates the 'ms' program on an arbitrary phylogenetic network).	<a href="#">here</a>	3.6.1
<a href="#">SimBiMarkersinNetwork</a>	Simulates bi-marker alleles under the multispecies network coalescent.	<a href="#">here</a>	3.6.1

## Characterizing and Comparing Phylogenetic Trees/Networks Based on Their Topologies

<a href="#">Cmpnets</a>	Computes the distance between two phylogenetic networks based on their topologies.		3.0.0
-------------------------	--	--	-------



# References: Methods

- MDC: *Syst Biol* 62(5): 738-751, 2013
- MLE: *PNAS* 111(46): 16448-16453, 2014
- MLE (pseudo / gene trees): *BMC Genomics* 16(Suppl 10): S10, 2015
- Bayesian (gene trees): *PLoS Genetics* 12(5): e1006006, 2016
- Bayesian (sequences): *Syst Biol* 67(3): 439-457, 2018
- Bayesian (bi-allelic markers): *PLoS Comp Bio* 14(1): e1005932, 2018
- MLE (pseudo / bi-allelic markers): *Bioinformatics*, 2018 (to appear)

# References: PhyloNet

- Than et al., BMC Bioinfo 9: 322, 2008
- Wen et al., Syst Biol, 2018 (to appear)

# Thank You



John Simon  
Guggenheim  
Memorial Foundation

## PhyloNet

The current version of PhyloNet is 3.6.4.

- **Download**

- [Binary jar file](#)

- **Download 3.6.5 Beta**

- [Binary jar file](#)

- **Usage**

- [General overview](#)

- [Tutorial: Species phylogeny inference](#)

- [List of PhyloNet commands](#)

- The phylogenetic network format (the Rice Newick format) used in PhyloNet can be readily visualized by [Dendroscope](#).

- **Troubleshooting**

- [Troubleshooting](#) for frequently encountered problems

**M. Barnett, J.H. Degnan,  
J. Dong, K. Liu, H. Ogilvie,  
D. Ruths, C. Than,  
D. Wen, Y. Yu, J. Zhu**

<http://bioinfoccs.rice.edu/phyloNet>