

# Statistically consistent estimation of level-1 phylogenetic networks from SNPs

Tandy Warnow, Yasamin Tabatabaee, and Steven N. Evans

TW and YT at UIUC, SNE at UC Berkeley

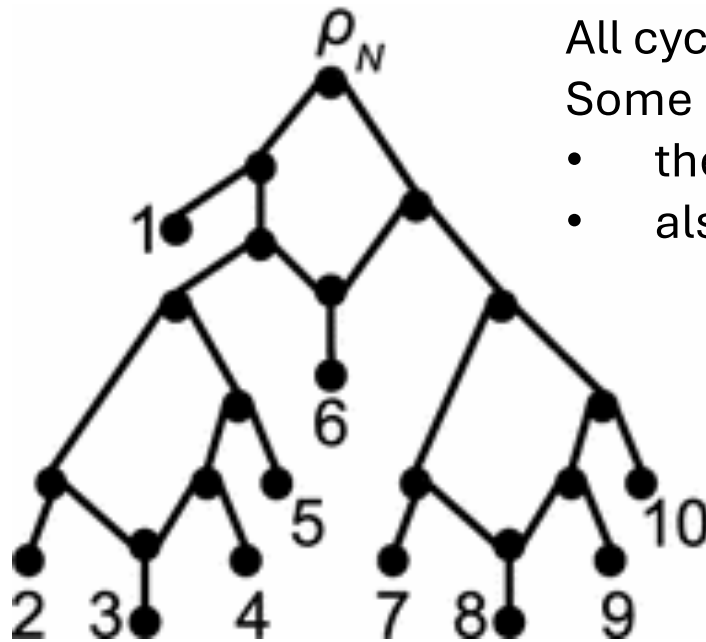
Paper presented in RECOMB-CG 2024

[https://link.springer.com/chapter/10.1007/978-3-031-58072-7\\_1](https://link.springer.com/chapter/10.1007/978-3-031-58072-7_1)

# Phylogenetic networks

- Models evolution of a group of species (or individuals)
- Handles hybridization and HGT, goes beyond trees
- Substantial literature in discrete mathematics
- Major software packages :
  - PhyloNet (Luay Nakhleh et al.)
  - PhyloNetwork (Claude Solís-Lemus et al.)
- This talk: statistical consistency in estimating “Level-1” phylogenetic networks

# Level-1 phylogenetic network (aka “galled tree”)



All cycles are node-disjoint

Some nodes have indegree  $> 1$

- these are called “reticulate nodes”
- also called “bottom nodes” of cycles

Figure from Gambette et al. 2015, On the challenge of reconstructing level-1 phylogenetic networks from triplets and clusters, <https://arxiv.org/abs/1511.08056v2>

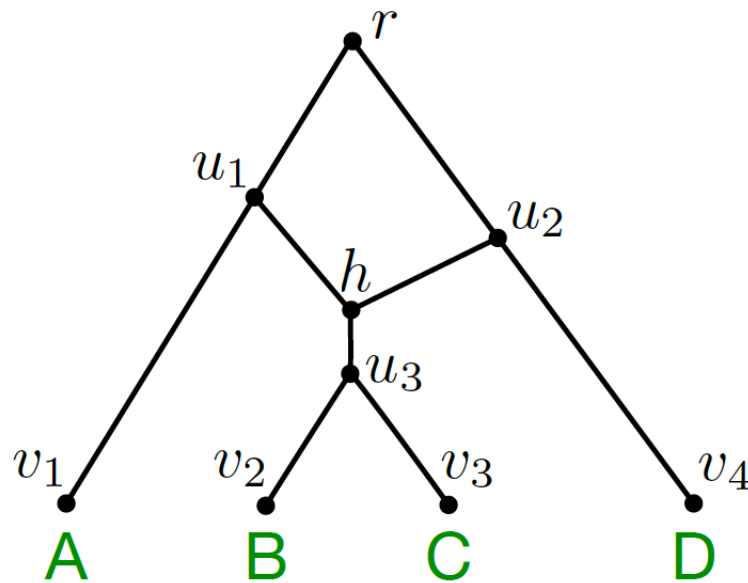
# Constructing rooted level-1 networks

*Prior work (e.g., Jansson et al., SICOMP 2006) has already established the following:*

Assuming  $N$  has **no cycles of size less than 5**, guaranteed correct and polynomial time reconstruction of the **rooted network topology** if given

- the set of all rooted trees in  $N$
- the set of all clades in  $N$
- the set of all rooted triplet trees in  $N$

# Rooted trees, rooted triplet trees, and clades inside rooted networks



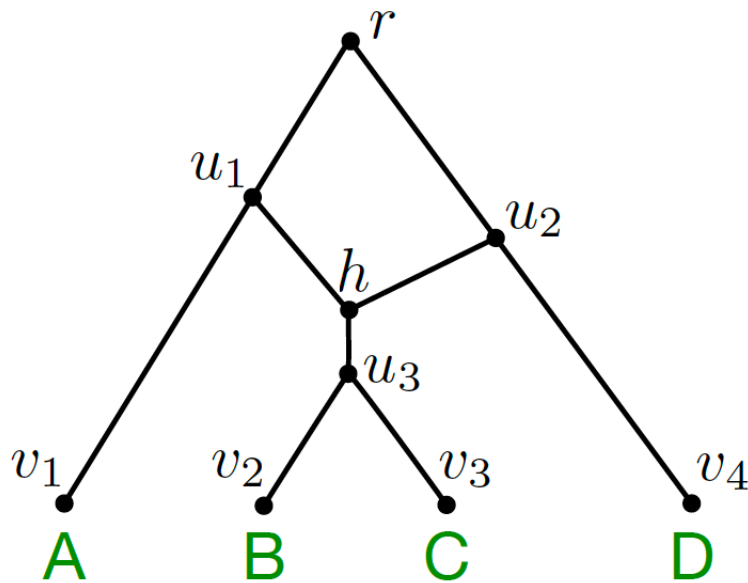
This network  $N$  is rooted at  $r$  and has one reticulate node,  $h$

For each way of removing one edge entering  $h$ , we obtain a rooted tree:

- $(A, (D, (B, C)))$  – if we remove  $(u_1, h)$
- $((A, (B, C)), D)$  – if we remove  $(u_2, h)$

The set of rooted trees is denoted  $T(N_r)$

# Rooted trees, rooted triplet trees, and clades inside rooted networks



The set of rooted trees is denoted  $T(N_r)$   
 Clades( $N_r$ ): clades of trees in  $T(N_r)$   
 =  $\{\{B,C\}, \{A,B,C\}, \{B,C,D\}, \{A,B,C,D\}\}$

Triplet trees: induced from the rooted trees.

- A|BC – but not AB|C or AC|B
- AB|D, A|BD – but not B|AD
- AC|D, A|CD – but not C|AD
- BC|D – but not BD|C or CD|B

# Our problem: Estimating networks from SNPs

- Sequences evolve down a level-1 phylogenetic network  $N$ , some of which are **SNPs** (defined here to be **binary sequences that evolve without homoplasy**)
- Question: Can we infer the topology of  $N$  from the data, under the assumption of access to an Oracle that tells us which sites are SNPs?
- Version 1: We **know the ancestral state** for every SNP
  - We can try to estimate the rooted topology
- Version 2: We **do not know the ancestral state** for any SNP
  - We can only try to estimate the unrooted topology

# Constructing rooted level-1 networks from SNPs?

If we have SNPs with known ancestral state, then we can get clades.

Therefore: given **SNPs with known ancestral state** that evolve down a level-1 network without cycles of size less than 5, we can construct **the correct rooted network topology**, if they cover all the clades of N



## But: if the ancestral state is not known?

- The biologically more realistic case is when we do not know the ancestral state. We just have two states, 1 and 2.
- Hence, we cannot infer trees, clades, or rooted triplet trees
- Remember: each SNP evolves down a tree in  $\mathcal{T}(N_r)$
- Therefore, the SNPs can provide
  - Bipartitions: splits of leafset into state 1 and state 2 for some SNP
  - $Q(N_r)$ : quartet subtrees  $AB|CD$  induced in  $\mathcal{T}(N_r)$
  - These can be inferred from SNPs: A and B exhibit state 1 and C and D exhibit state 2 for some SNP

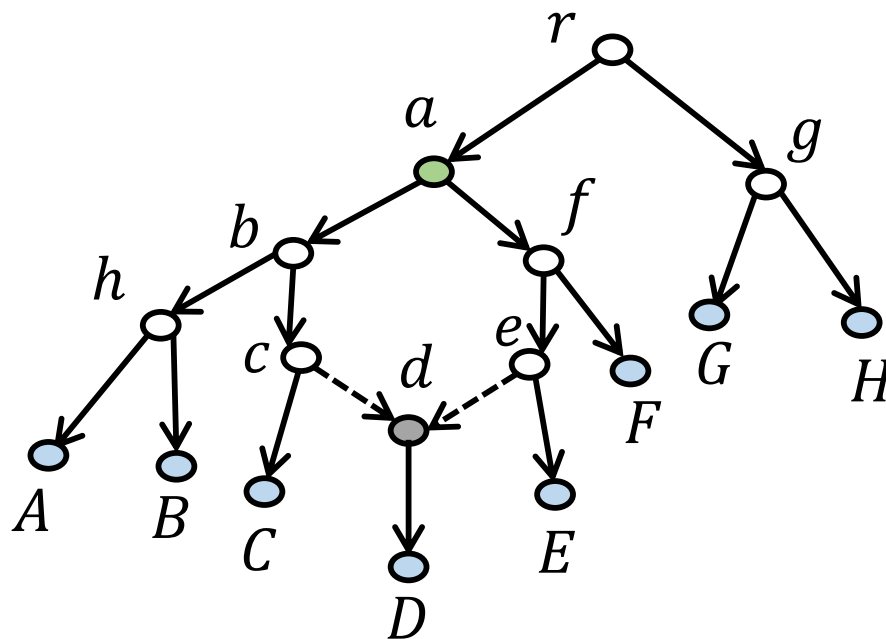
## But: if the ancestral state is not known?

- The biologically more realistic case is when we do not know the ancestral state. We just have two states, 1 and 2.
- Hence, we cannot infer trees, clades, or rooted triplet trees
- But we can use the SNPs to obtain:
  - Bipartitions: splits of leafset into those that exhibit state 1 and those that exhibit state 2
  - $Q(N_r)$ : quartet subtrees of trees in  $\mathcal{T}(N_r)$
- Question: Can we infer the unrooted topology of a level-1 phylogenetic network from SNPs without known ancestral state?

# Prior results relevant to SNPs without known ancestral state

- **Gusfield** (JCSS 2005): polynomial time algorithm to construct an unrooted level-1 phylogenetic network consistent with input set of SNPs (without known ancestral state), **but no proof of uniqueness**
- **Gambette, Berry, and Paul** (JBCB 2012): poly time algorithm to construct the **true** unrooted level-1 phylogenetic network **when given  $Q(N)$** , the set of all quartet trees in  $N$

# Q(N): quartet trees of network N (unrooted)



Ignore the root!

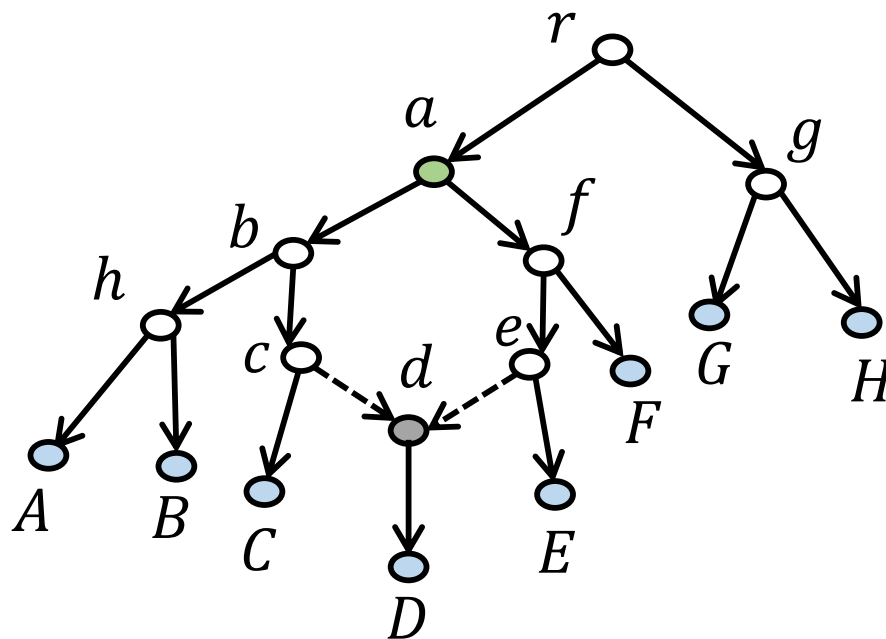
Then we can define  $Q(N)$ :

- The set of all  $UV|XY$  so that there are node-disjoint paths from  $U$  to  $V$  and from  $X$  to  $Y$

Question: which of the following are in  $Q(N)$ ?

- $AC|EF$
- $CD|AE$
- $AF|CE$
- $AD|CE$

# Q(N): quartet trees of network N (unrooted)



Ignore the root!

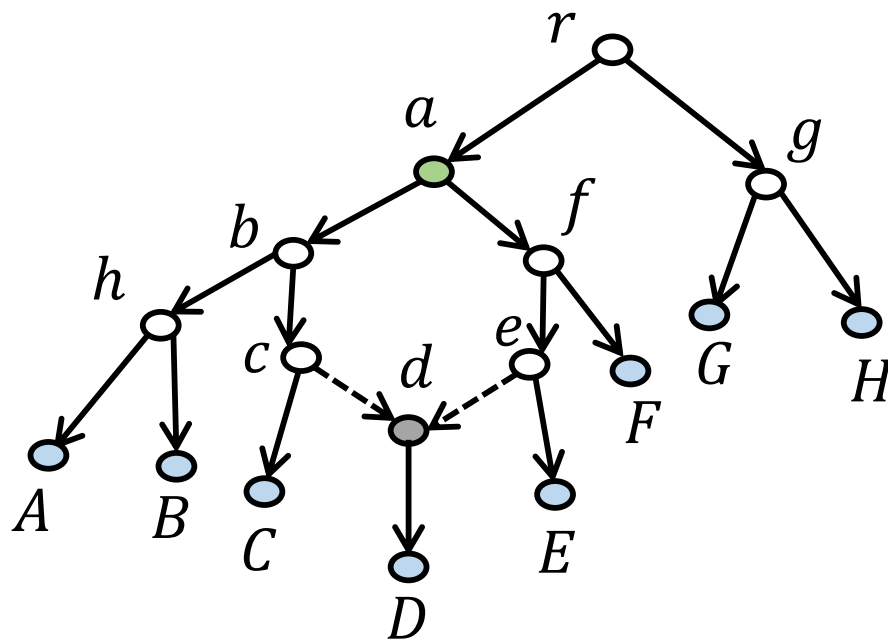
Then we can define Q(N):

- The set of all  $UV|XY$  so that there are node-disjoint paths from U to V and from X to Y

Question: which of the following are in Q(N)?

- AC|EF - yes
- CD|AE - yes
- AF|CE - yes
- AD|CE - no

# $Q(N_r)$ : quartet trees of network $N$ (rooted)



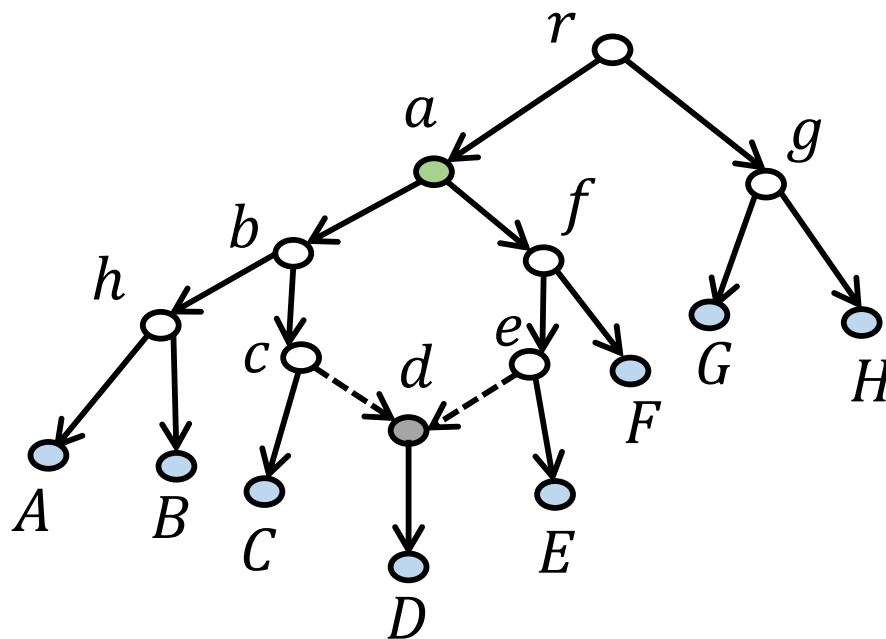
Given the set  $\mathcal{T}(N_r)$  of trees in the network, we can define  $Q(N_r)$ :

- The set of all  $UV|XY$  so that there are node-disjoint paths from  $U$  to  $V$  and from  $X$  to  $Y$  **within some tree in  $\mathcal{T}(N_r)$**

Question: which of the following are in  $Q(N_r)$ ?

- $AC|EF$
- $CD|AE$
- $AF|CE$
- $AD|CE$

# $Q(N_r)$ : quartet trees of network $N$ (rooted)



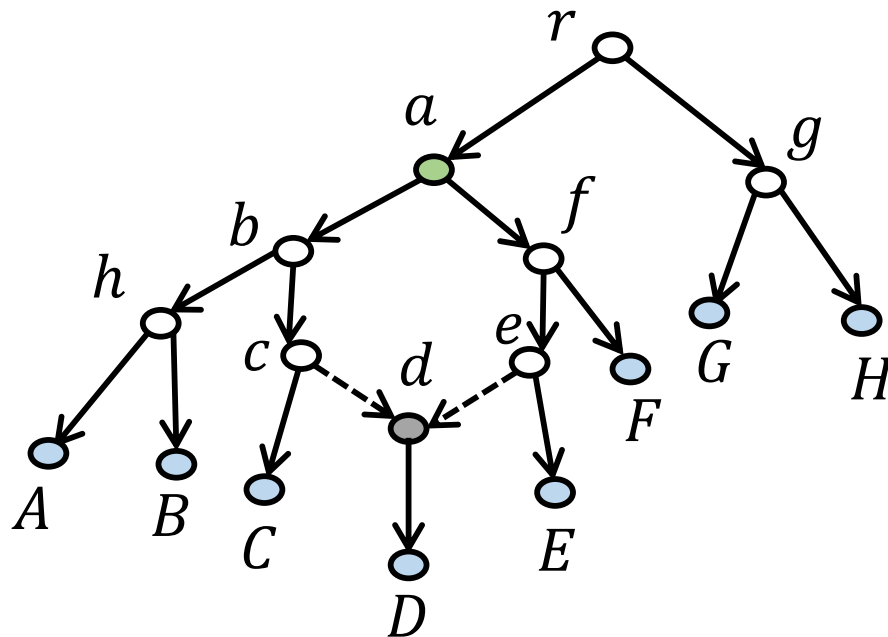
Given the set  $\mathcal{T}(N_r)$  of trees in the network, we can define  $Q(N_r)$ :

- The set of all  $UV|XY$  so that there are node-disjoint paths from  $U$  to  $V$  and from  $X$  to  $Y$  within some tree in  $\mathcal{T}(N_r)$

Question: which of the following are in  $Q(N_r)$ ?

- $AC|EF$  - **yes**
- $CD|AE$  - **yes**
- $AF|CE$  - **no**
- $AD|CE$  - **no**

# Q(N) vs. Q(N<sub>r</sub>)



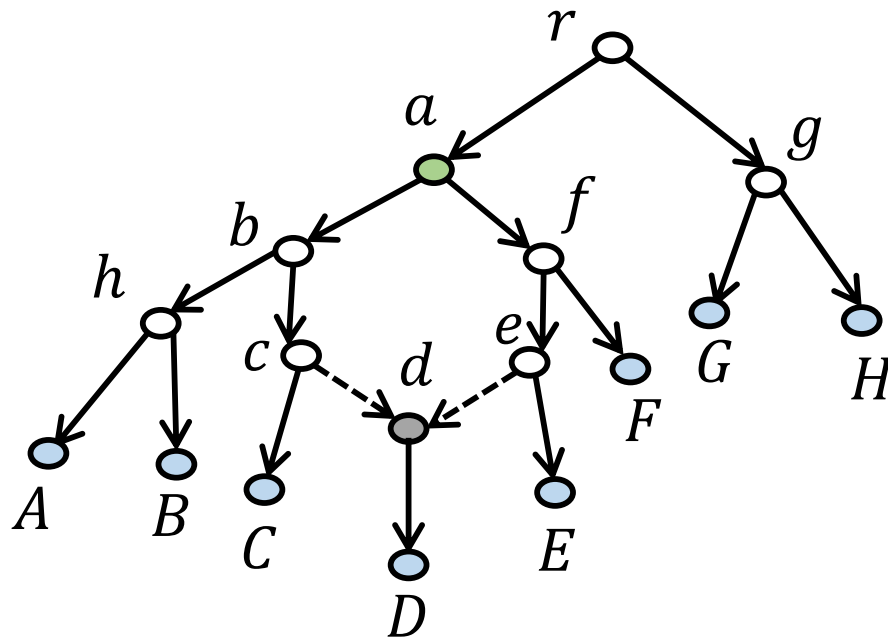
Question: which of these are in  $Q(N)$  and/or  $Q(N_r)$ ?

- $AC|EF$  - in both
- $CD|AE$  - in both
- $AF|CE$  - only in  $Q(N)$ , not in  $Q(N_r)$
- $AD|CE$  - not in either

Thus:  $Q(N_r)$  is a subset of  $Q(N)$ , and can be a proper subset.



# $Q(N)$ vs. $Q(N_r)$ : *Not the same!*



Question: which of these are in  $Q(N)$  and/or  $Q(N_r)$ ?

- $AC|EF$  - in both
- $CD|AE$  - in both
- $AF|CE$  - only in  $Q(N)$ , not in  $Q(N_r)$
- $AD|CE$  - not in either

Thus:  $Q(N_r)$  is a subset of  $Q(N)$ , and can be a proper subset.

# Our contributions: algorithms

- Our problem: unrooted level-1 network topology reconstruction given SNPs without known ancestral state, where cycles are of length at least 5, and we assume the SNPs cover the clades of  $N$ .
- We use quartet trees and bipartitions from SNPs
- We prove:
  - The algorithm from Gambette, Berry, and Paul (GBP) can fail
  - SNPs suffice to identify the unrooted topology of  $N$ .
  - Gusfield's algorithm correctly reconstructs  $N$

# Our contributions: statistical estimation

- We provide a **model of site evolution** down level-1 phylogenetic networks
- Given an **Oracle** that determines which sites are SNPs, we establish that Gusfield's algorithm can be used in a **statistically consistent** and polynomial time pipeline
- We provide a **sample complexity** for this approach

# This talk

I will cover:

- Why GBP doesn't work for constructing networks from SNPs without known ancestral state
- Proof that SNPs suffice to construct unique network if they cover all clades and  $N$  has no cycles of length less than five
- The CUPNS algorithm (Constructing Unrooted Phylogenetic Networks from SNPs) correct for level-1 network construction
- Proof that Gusfield's algorithm is also correct

I will not cover the stochastic model, proof of statistical consistency, or sample complexity (not enough time)

## GBP: Gambette, Berry, and Paul, JBCB 2012

- GBP prove that their algorithm returns the unrooted topology of  $N$  if given  $Q(N)$ , where  $N$  is a level-1 phylogenetic network.

# GBP: Gambette, Berry, and Paul, JBCB 2012

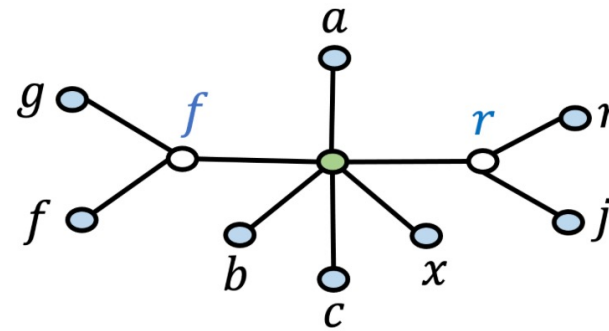
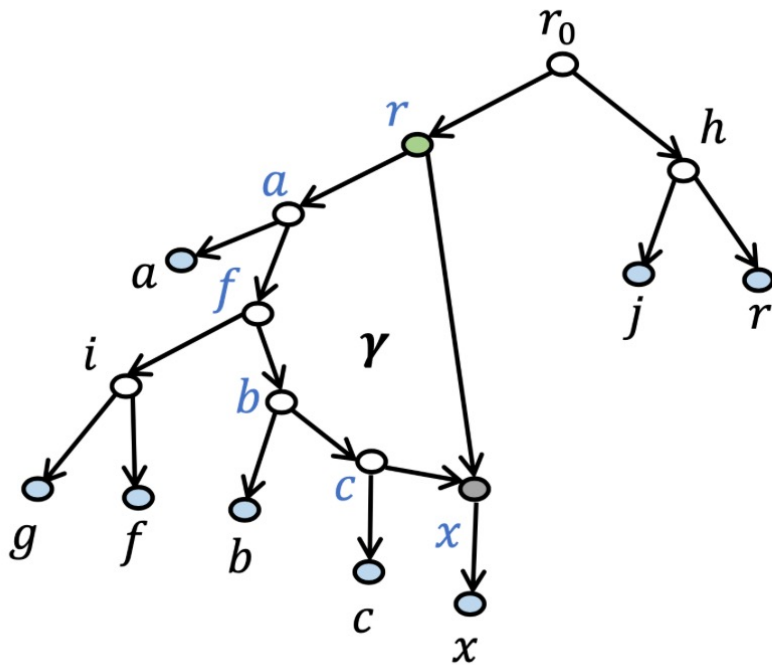
- Given set  $Q$  of quartet trees:
  - Phase 1: Construct “SN-tree” (maximally refined tree  $T$  all of whose resolved quartets are in  $Q$ , and no resolved quartet tree in  $T$  conflicts with any quartet tree in  $Q$ )
  - Phase 2: Replace each polytomy by a cycle, using the “node ordering algorithm”
  - Return the resultant graph

## Phase 2 of GBP

Given a polytomy  $v$  in the SN-tree, and given set  $Q$  of quartet trees:

- Label the neighbors of  $v$  by leaves
- Determine that nodes  $a, b, c$  appear consecutively, in that order, in the cycle expansion of the polytomy **if and only if** there does not exist a label  $z$  such that  $ac|bz$  is a quartet in  $Q$ .
- Return the cycle if and only if information is consistent with a unique cycle.

# Phase 1: Constructing the SN-tree given $Q(N)$



The SN-tree given  $Q(N)$

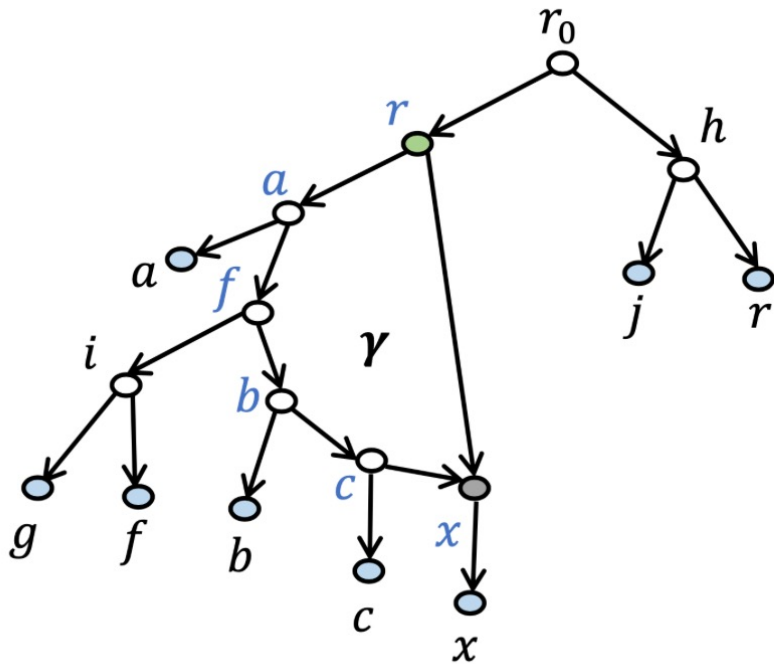
- The cycle has been collapsed to a polytomy,
- The nodes adjacent to the polytomy are labelled by leaves.

Network  $N$  with cycle  $\gamma$

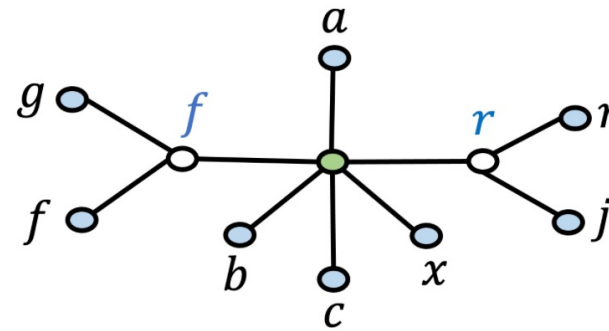
The internal nodes of  $\gamma$  are labelled by leaves that attach to the cycle at those nodes.



# Phase 2: Refining the polytomy given $Q(N)$



Network N with cycle



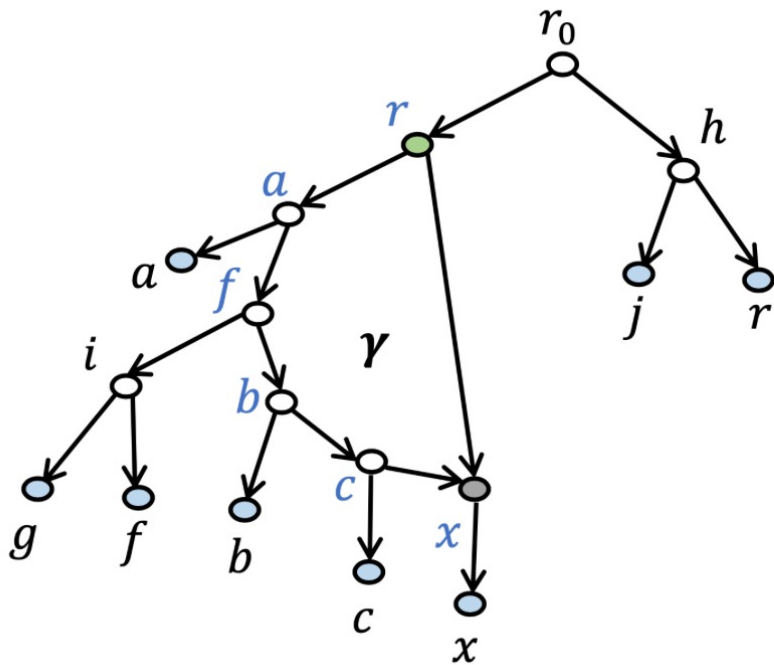
The SN-tree given  $Q(N)$

Given  $Q(N)$ , the node ordering algorithm will replace the polytomy in the SN-tree by the correct cycle—and so returns the unrooted topology of  $N$ .

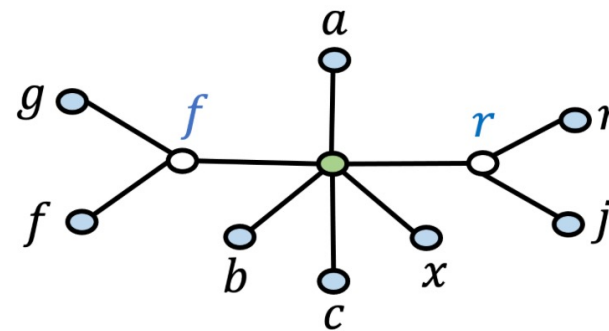
# GBP: usable from SNPs?

- Remember: the SNPs define  $Q(N_r)$ 
  - Each SNP defines a bipartition of the leaves into two sets
  - Hence we can get quartet trees  $AB|CD$  where A and B share one state, and C and D share the other state.
- But we already showed  $Q(N_r)$  can be a proper subset of  $Q(N)$
- So: what happens if we give GBP the set  $Q(N_r)$  instead of  $Q(N)$ ?

We proved the SN-tree given  $Q(N)$  is the same as the SN-tree given  $Q(N_r)$



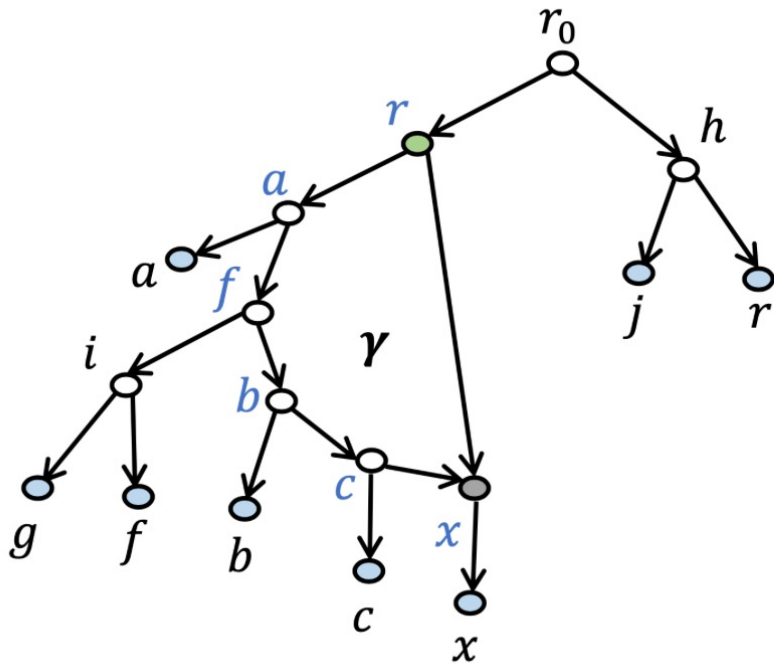
Network N with one cycle.



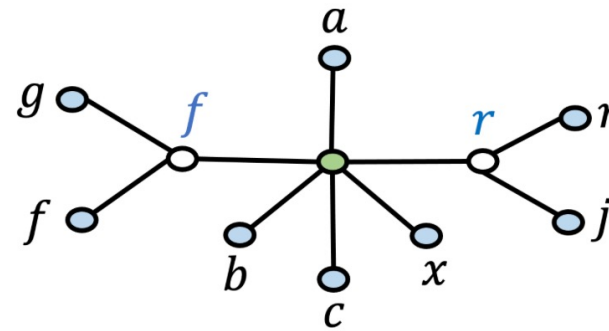
The SN-tree given  $Q(N_r)$  or  $Q(N)$

- The cycle has been collapsed to a polytomy,
- The nodes adjacent to the polytomy are labelled by leaves.

# Phase 2: Refining the polytomy given $Q(N_r)$

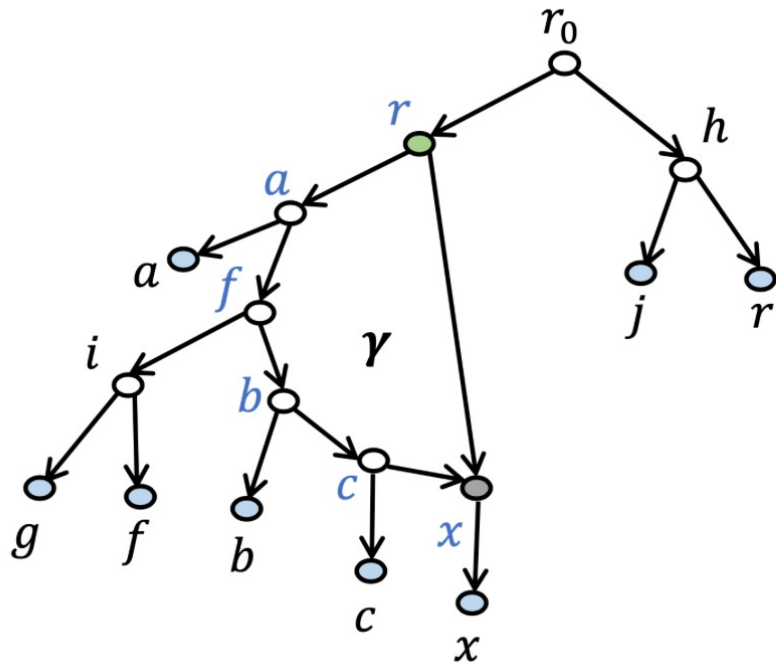


Network N with one cycle.

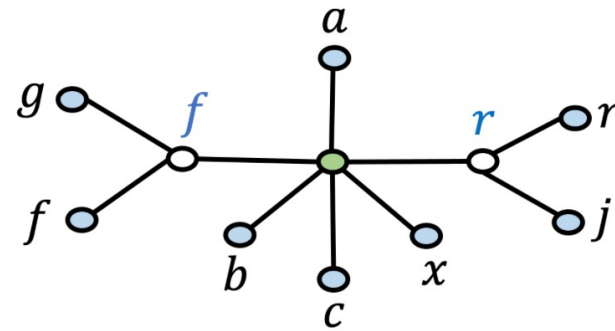


The SN-tree given  $Q(N_r)$

# Phase 2: Refining the polytomy given $Q(N_r)$



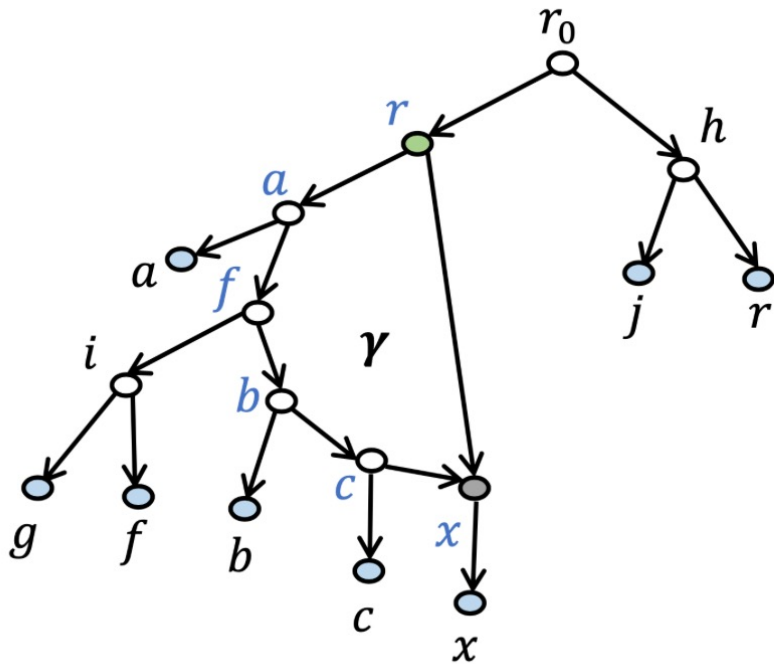
Network N with one cycle.



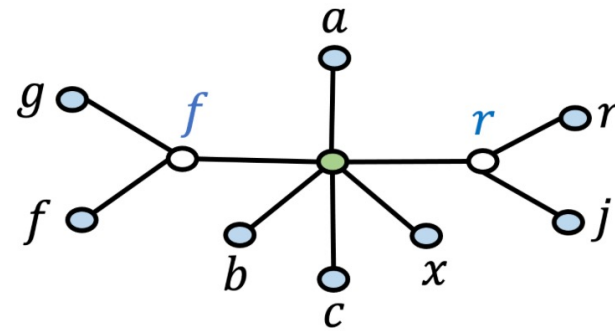
The SN-tree given  $Q(N_r)$

Note:  $bf|ac$  is in  $Q(N)$ , but  $bf|ac$  is NOT in  $Q(N_r)$ !

# Phase 2: Refining the polytomy given $Q(N_r)$



Network N with one cycle

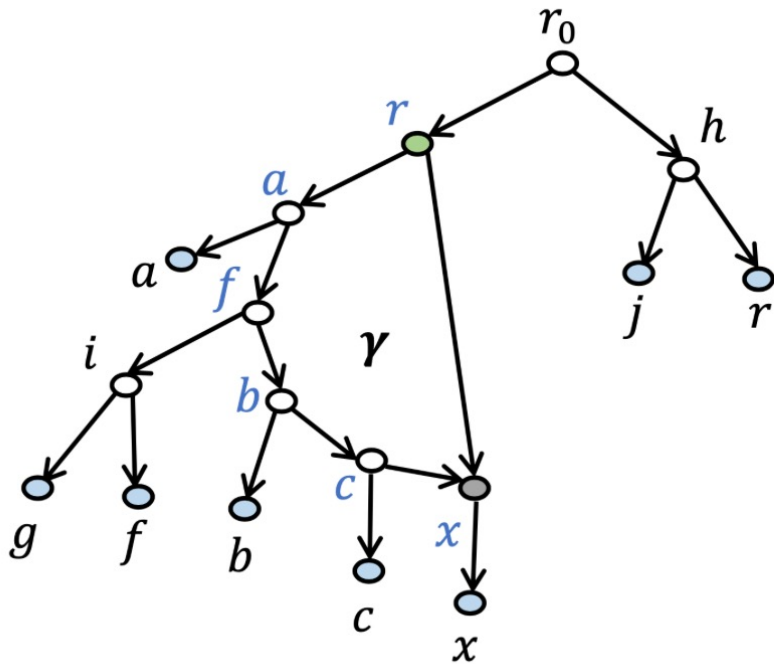


The SN-tree given  $Q(N_r)$

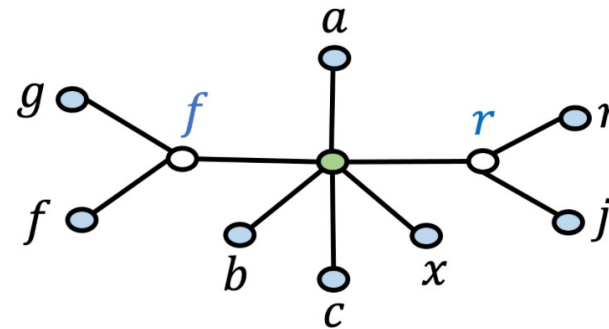
Note:  $bf|ac$  is in  $Q(N)$ , but  $bf|ac$  is NOT in  $Q(N_r)$ !

So: the node ordering algorithm thinks  $a, b, c$  appear consecutively, which is a mistake.

## Phase 2: Refining the polytomy given $Q(N_r)$



Network N with one cycle.



The SN-tree given  $Q(N_r)$

Note:  $bf|ac$  is in  $Q(N)$ , but  $bf|ac$  is NOT in  $Q(N_r)$ !

So: the node ordering algorithm thinks  $a, b, c$  appear consecutively, which is a mistake.

Thus, the node-ordering algorithm can fail on  $Q(N_r)$ !

# Estimating level-1 networks from SNPs

- With known ancestral state, many algorithms are correct (provided cycles not below size 5, SNPs cover the clades)
- Without known ancestral state, GBP can fail
  
- But we can use  $Q(N_r)$  to obtain the level-1 network even without known ancestral state, as we now show (CUPNS method)



## CUPNS: Constructing Unrooted Phylogenetic Networks from SNPs

- Input: SNPs without known ancestral state
- Output: network N if required conditions hold
  - Level-1
  - No cycles of size less than 5
  - The SNPs cover the clades of N

# CUPNS method

- Input: SNPs without known ancestral state
- Steps:
  - Construct set  $Q$  of quartet trees
  - Construct SN-tree for  $Q$ , and label neighbors of each polytomy
  - For each polytomy,
    - Use  $Q$  to find “bottom node” (unique leaf  $x$  for which two quartet trees exist in  $Q$  for any set of four leaves including  $x$ )
    - Let  $Q'$  be result of removing quartets involving  $x$  from  $Q$ .
    - Construct tree  $T$  on  $Q'$ . If  $Q=Q(N_r)$ , this will be a caterpillar tree.
    - Use  $Q$  to close  $T$  into a cycle including  $x$ .

# CUPNS method

- Theorem: Suppose  $N$  is level-1, all cycles are of size at least five, and the SNPs are given without known ancestral state but cover the clades of  $N$ . Then CUPNS will return  $N$ , and uses polynomial time.
- Corollary: There is only one unrooted level-1 network that is consistent with  $Q(N_r)$  when the required conditions hold.
- Corollary: Gusfield's algorithm is correct when the required conditions hold.

# Contributions: algorithms when SNPs are given without known ancestral state

Method	Comments if required conditions hold	Runtime
Gusfield-Construct-Unrooted	Guaranteed correct	$O(mn + n^3)$
CUPNS	Guaranteed correct	$O(mn^4)$
GBP-SNPs	Will fail on some networks	$O(mn^4)$

Notes:  $m$  is the number of SNPs and  $n$  is the number of leaves.

Required conditions:

- (1) the SNPs cover the bipartitions of  $N$ ,
- (2)  $N$  is level-1, and
- (3) every cycle in  $N$  is of length at least 5

# Statistical estimation

- We also provide a parametric model of character evolution down level-1 phylogenetic networks
- We assume existence of a perfect oracle (tells us which sites are SNPs)
- Under these assumptions, we prove that CUPNS and Gusfield's algorithms are statistically consistent estimators of the unrooted level-1 network.
- We also provide sample complexity analysis.

# Summary and Future Work

- We provided a proof that Gusfield's algorithm is statistically consistent for estimating a level-1 phylogenetic network
- We also provided an alternative approach, CUPNS, based on using quartet trees
- Unfortunately, these methods have no guarantees if there is Oracle error or the SNPs evolve down a more complex network (e.g., a level-2 network)
- Much future work is needed!
- [https://link.springer.com/chapter/10.1007/978-3-031-58072-7\\_1](https://link.springer.com/chapter/10.1007/978-3-031-58072-7_1)