A divide-and-conquer approach to phylogenetic network inference

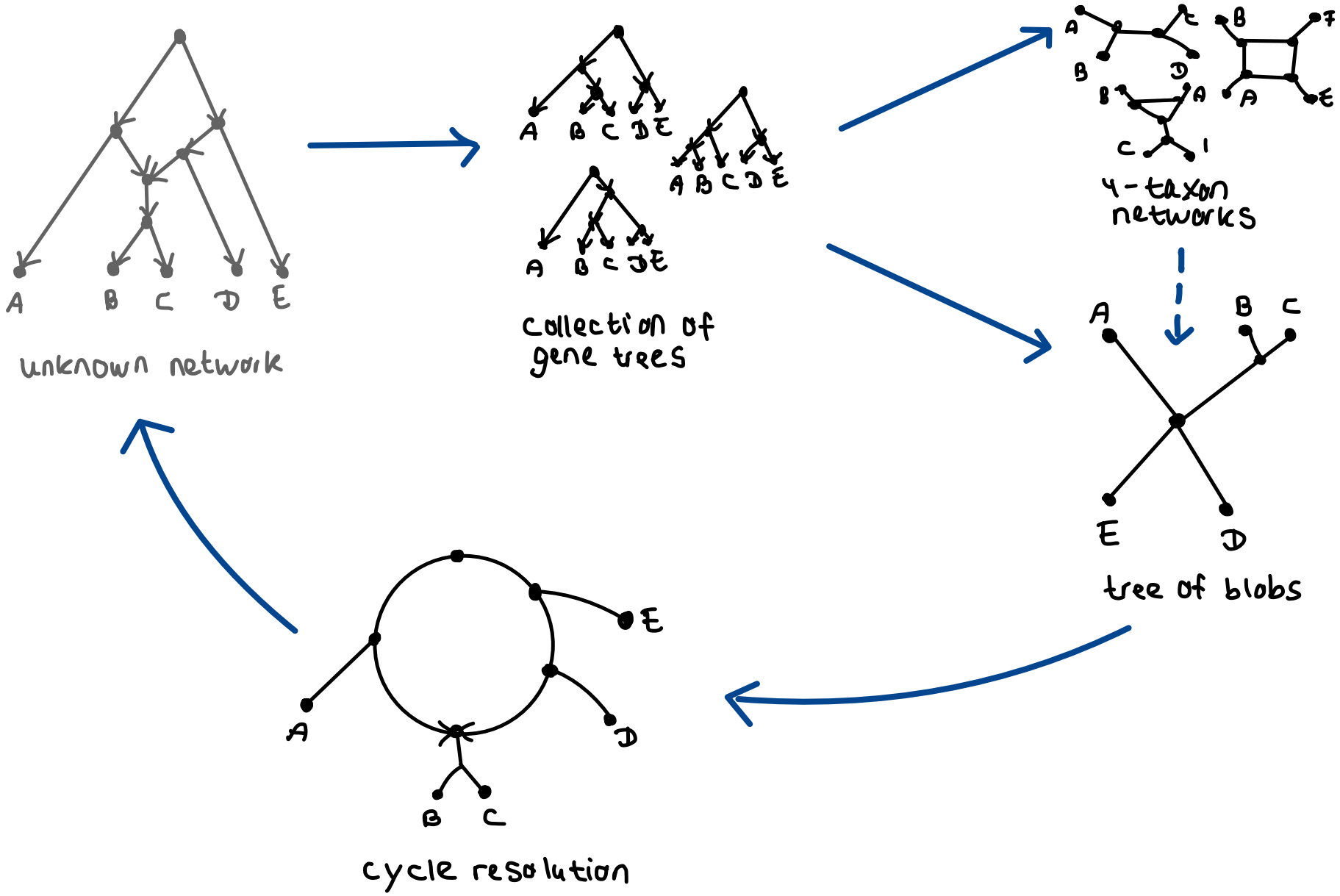
Kristina Wicke
kristina.wicke@njit.edu

Department of Mathematical Sciences
New Jersey Institute of Technology

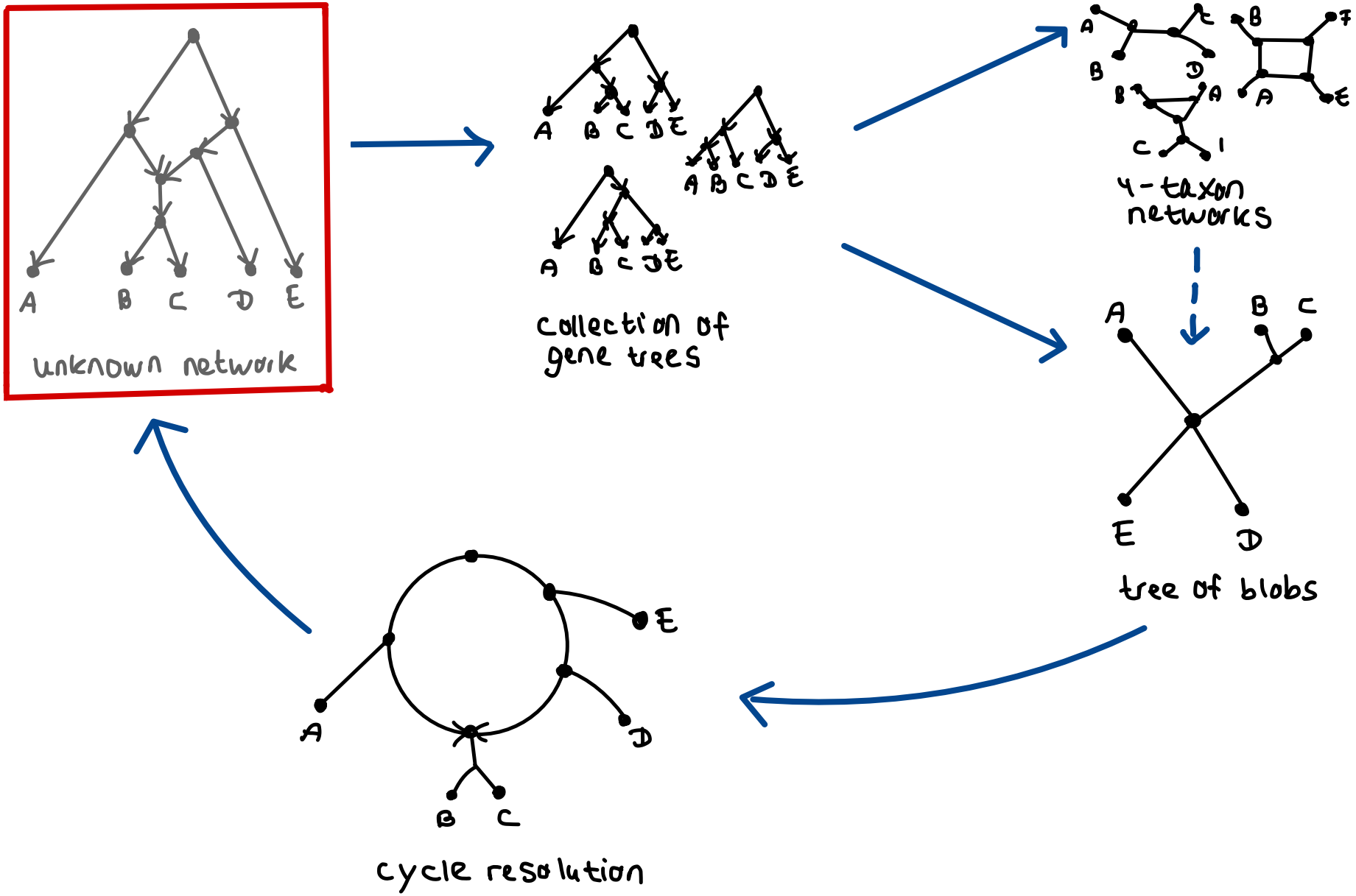
Joint work with Elizabeth Allman, Hector Baños, and John Rhodes

September 18, 2024

Big picture

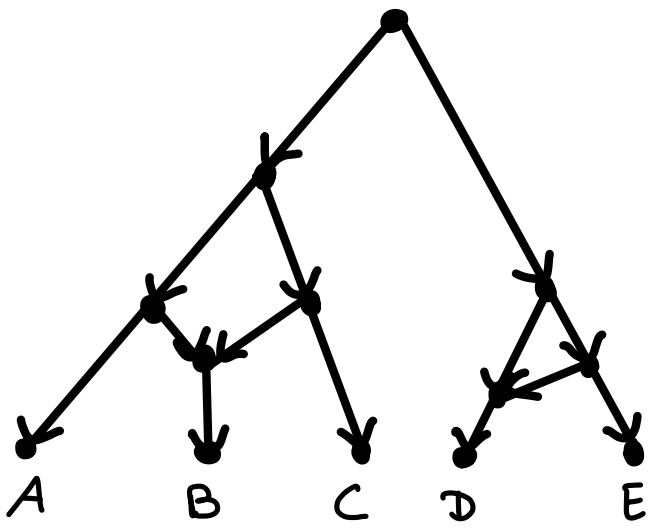

A B C D E
unknown network
A B C D E
A B C D E
A B C D E
collection of gene trees
4-taxon networks
A B C
E D
tree of blobs
E
A
D
B C
cycle resolution

Big picture

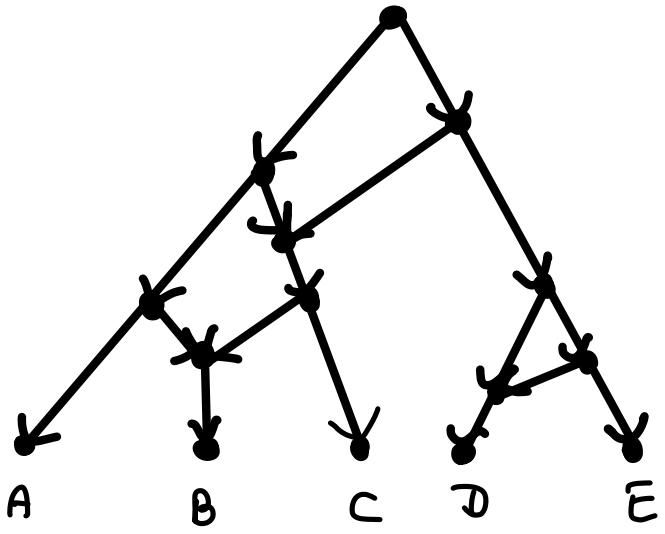

A B C D E
unknown network
A B C D E
A B C D E
A B C D E
collection of gene trees
4-taxon networks
A B C
E D
tree of blobs
E
A
D
B C
cycle resolution

Level-1 networks

A rooted binary phylogenetic network N^+ is called **level**-1 if no two cycles share an edge.



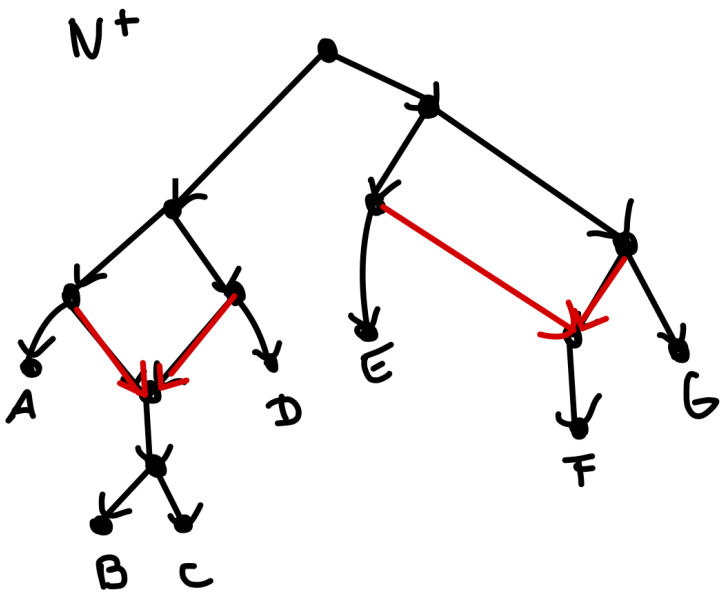
level-1

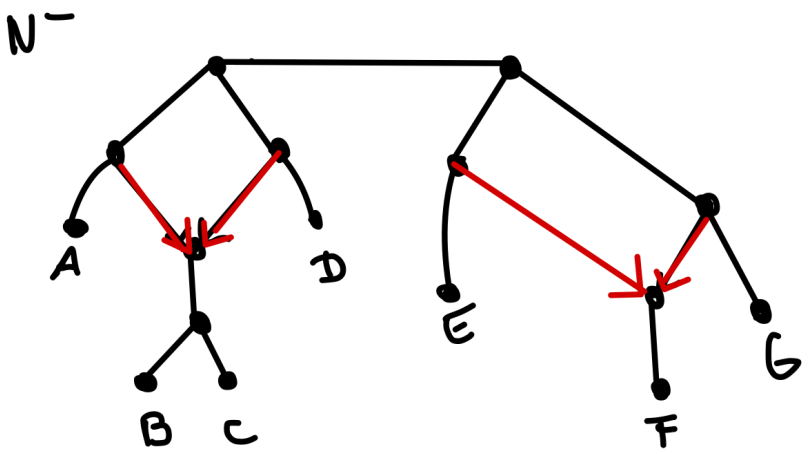


not level-1

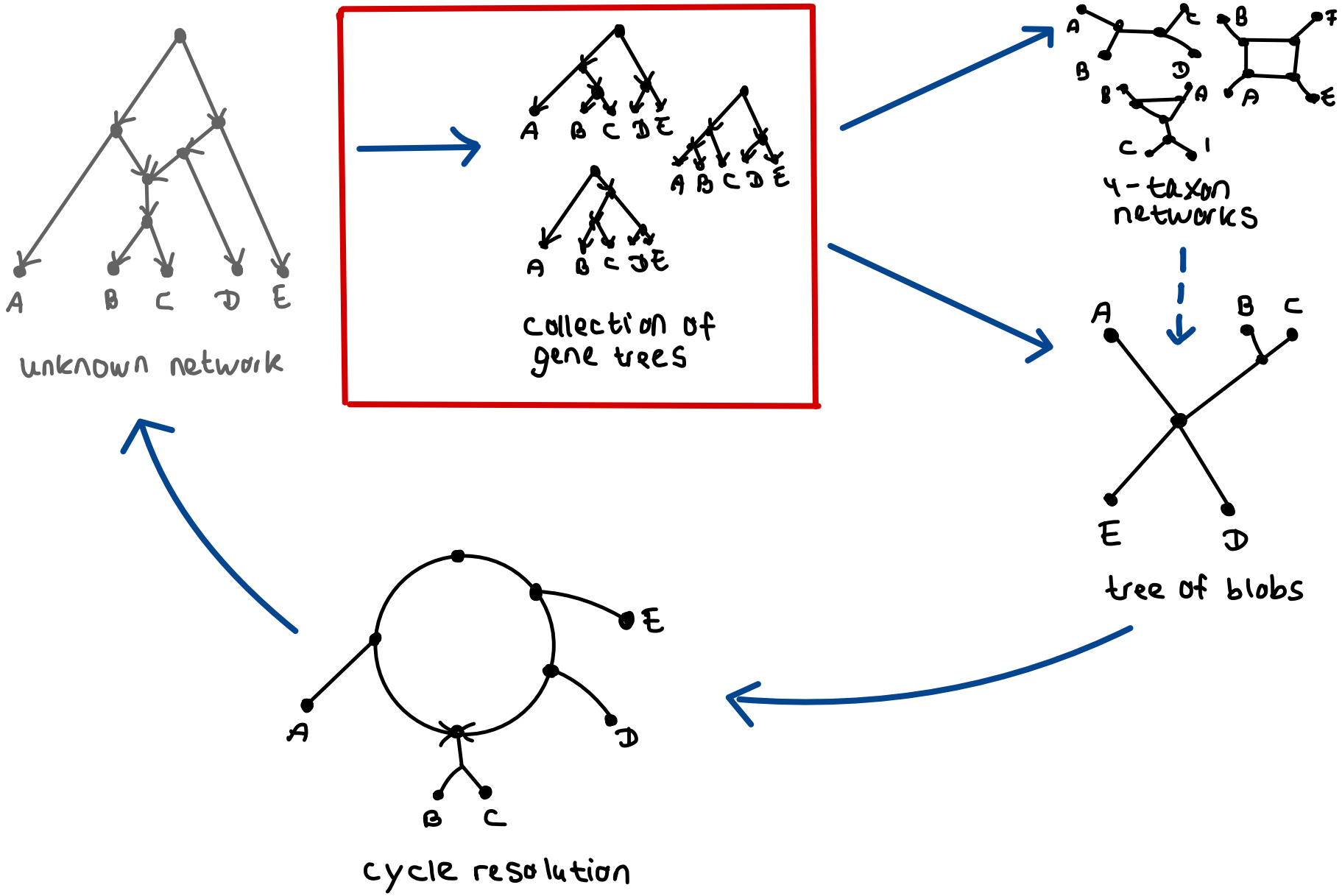
Semi-directed networks

A network is a **semi-directed network** on X if it can be obtained from a rooted network on X by suppressing its root and undirecting all tree edges.

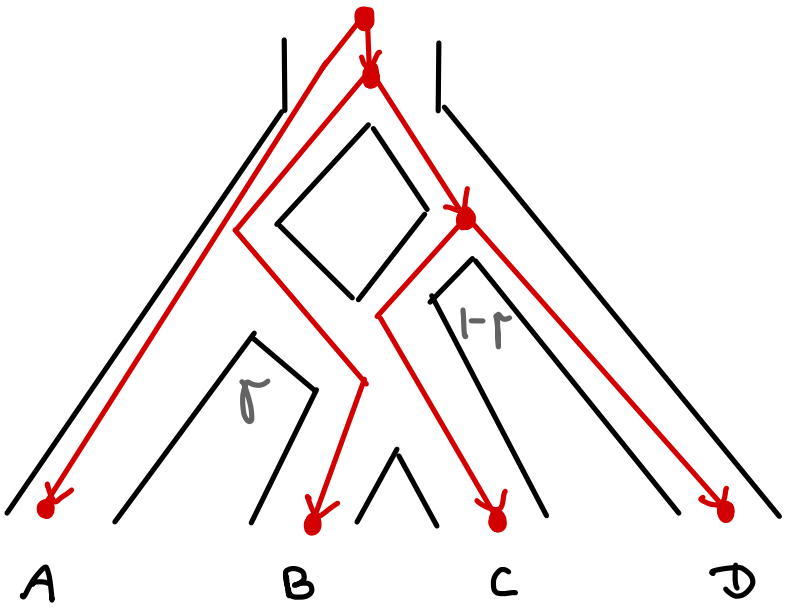


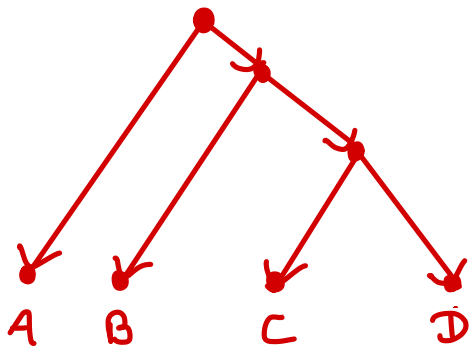


Big picture


A B C D E
unknown network
A B C D E
A B C D E
A B C D E
collection of gene trees
4-taxon networks
A B C
E D
tree of blobs
E
A
D
B C
cycle resolution

Network multispecies coalescent model (NMSC)

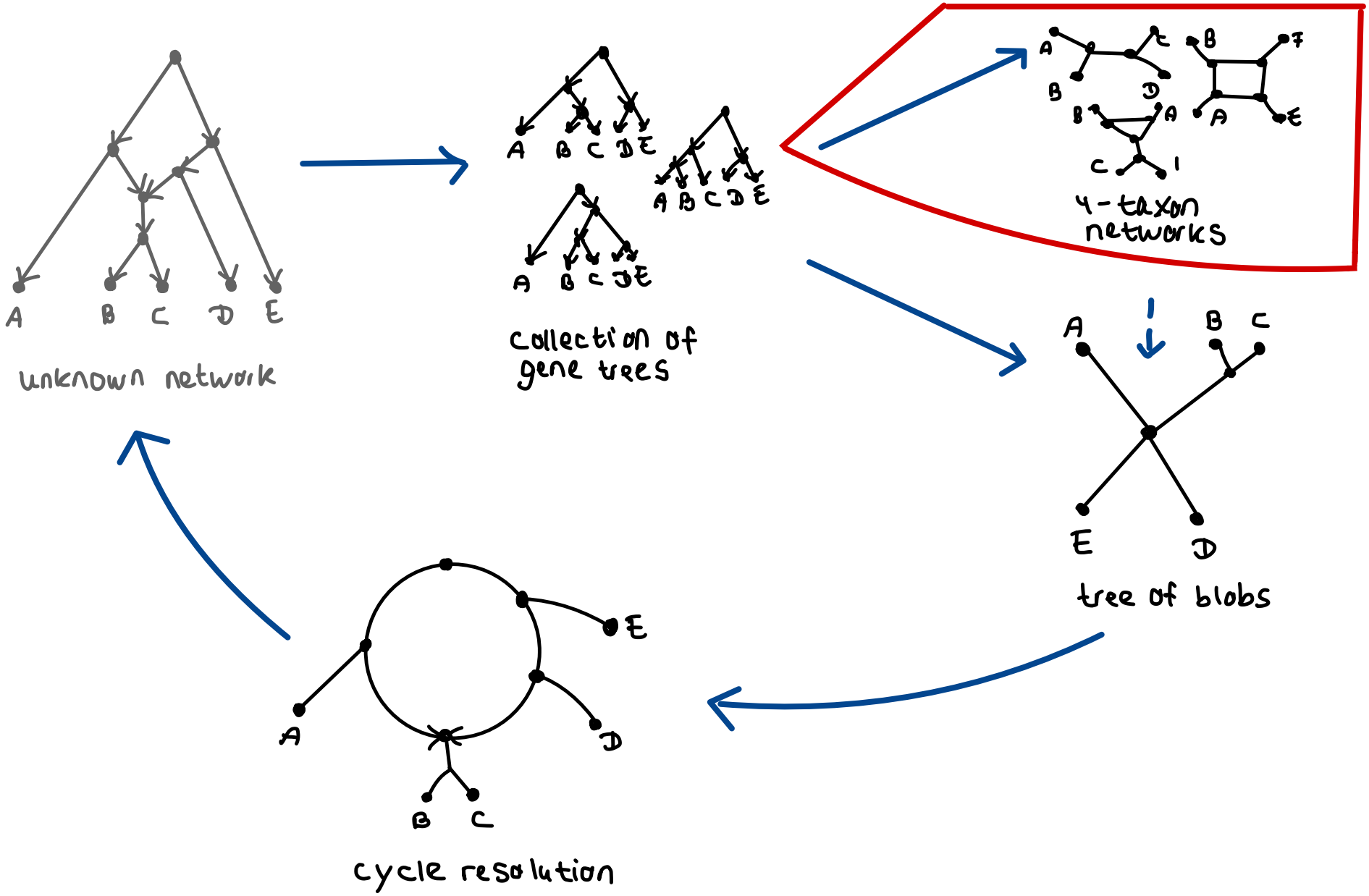




The **network multispecies coalescent model (NMSC)** is a stochastic model of gene tree generation incorporating

- hybridization (or other forms of lateral gene transfer);
- incomplete lineage sorting (ILS).

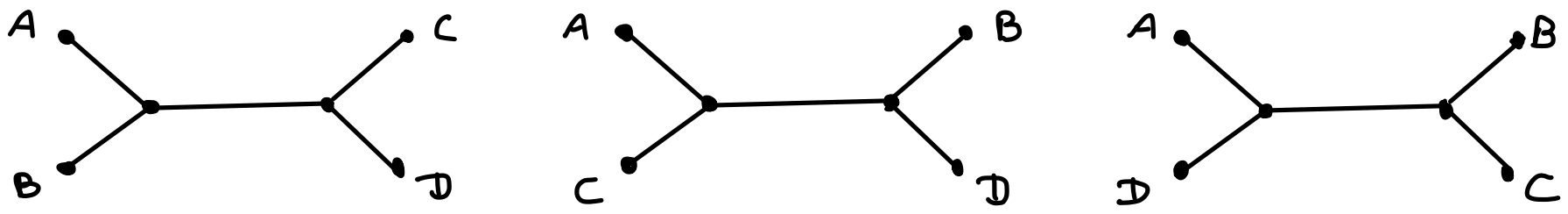
Big picture


A B C D E
unknown network
A B C D E
A B C D E
A B C D E
collection of gene trees
4-taxon networks
A B C D E
tree of blobs
A B C D E
cycle resolution

Quartets and concordance factors

For each 4-taxon set $\{A, B, C, D\}$, the probability that a gene tree induces each of the unrooted quartets $AB|CD$, $AC|BD$, $AD|BC$ is called the **quartet concordance factor (CF)**, denoted

$$CF_{ABCD} = (CF_{AB|CD}, CF_{AC|BD}, CF_{AD|BC}).$$



Empirical CFs

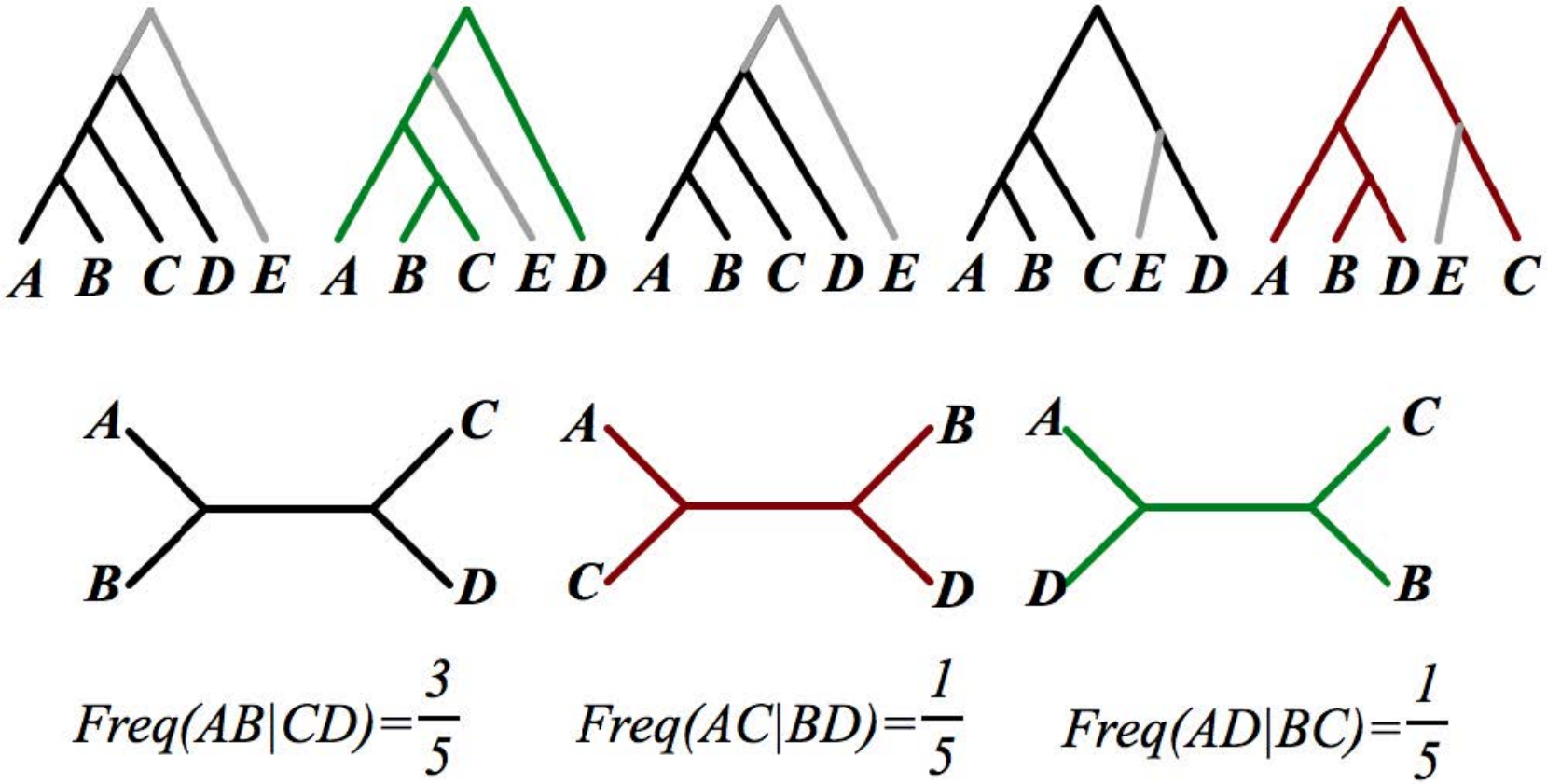


Image credit: Hector Baños

Quartets and concordance factors

We use quartet concordance factors and statistical hypothesis tests to infer the relationships between sets of 4 species.

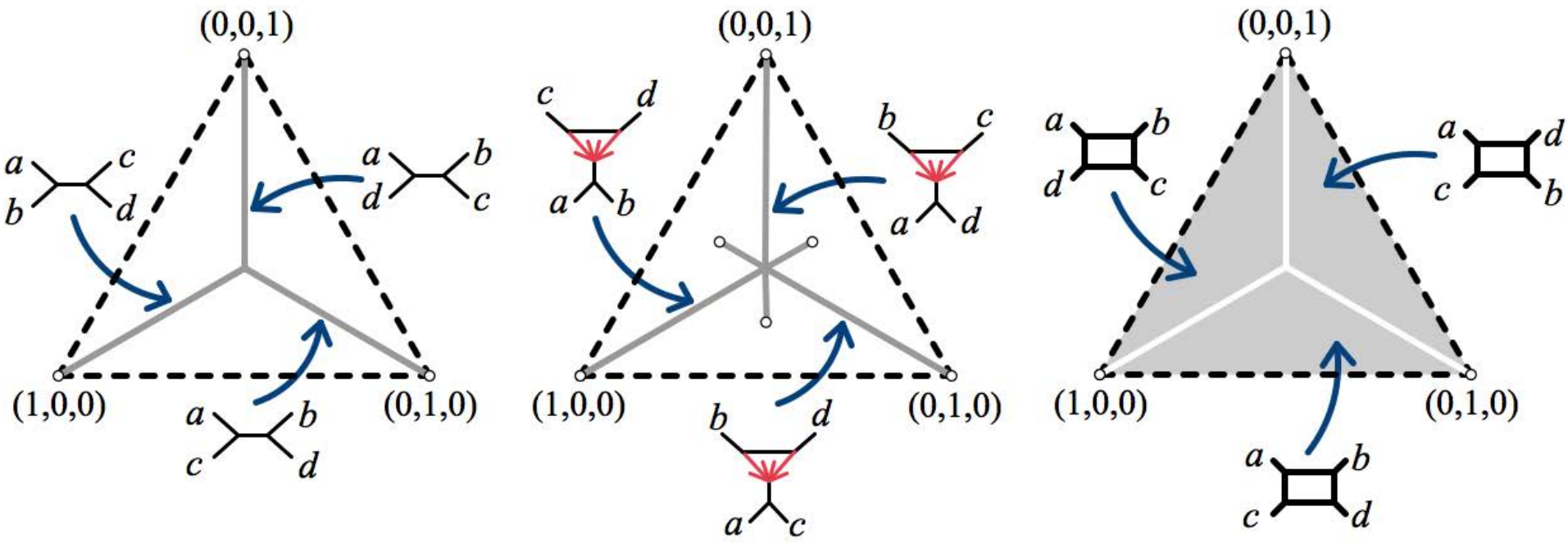
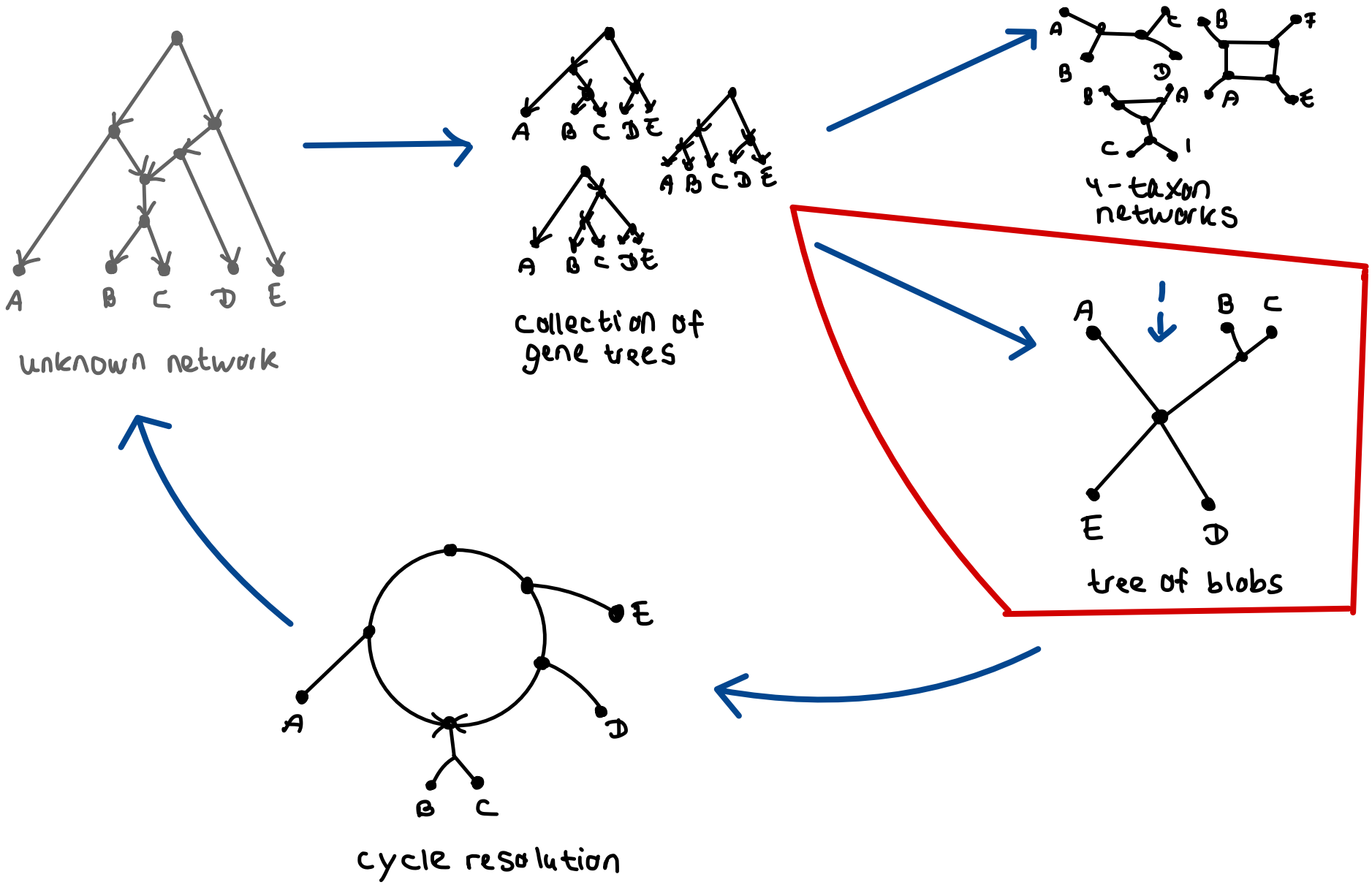


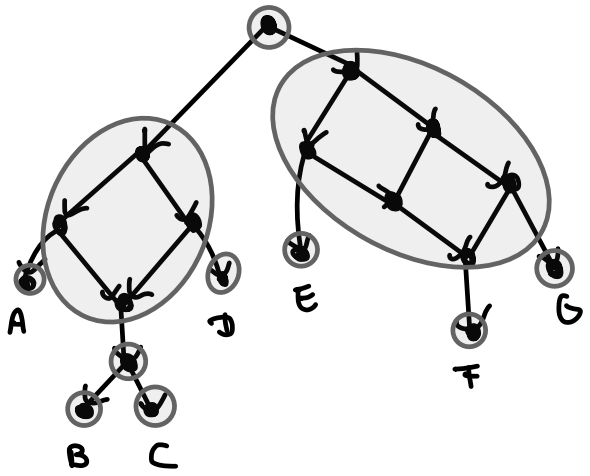
Image credit: Hector Baños

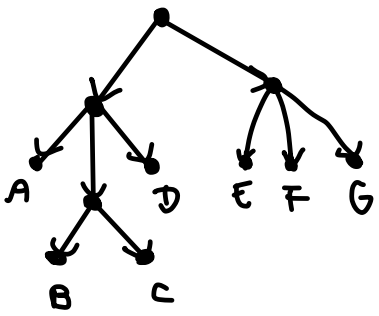
Big picture

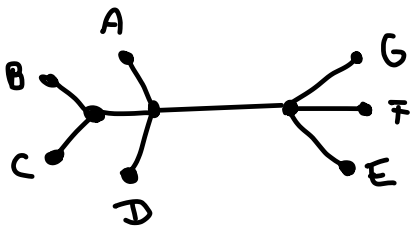

A B C D E
unknown network
A B C D E
A B C D E
A B C D E
collection of gene trees
4-taxon networks
A B C
E D
tree of blobs
E
A
D
B C
cycle resolution

Tree of blobs

- A **blob** of a network is a maximal connected subnetwork that has no cut edges.
- The (reduced) **tree of blobs** of a network is obtained by contracting each blob to a node and suppressing non-root degree-2 vertices.







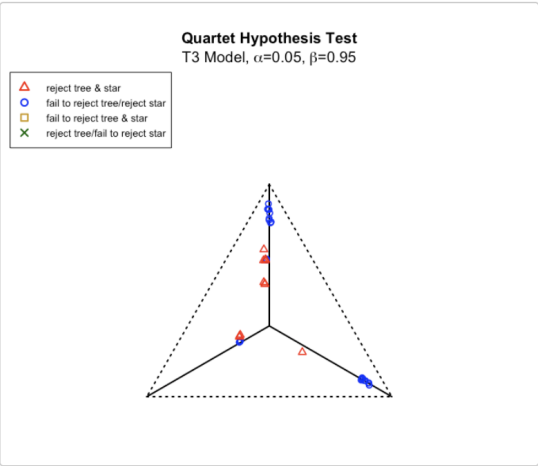
Theorem (Allman, Baños, Mitchell, Rhodes (2023))

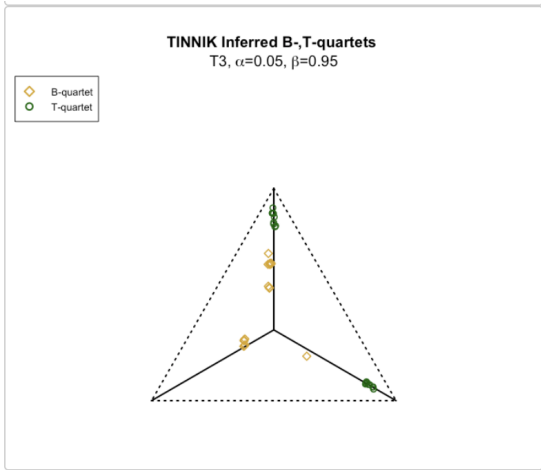
For generic numerial parameters, the reduced unrooted tree of blobs is identifiable from the distribution of gene quartet topologies under the NMSC model.

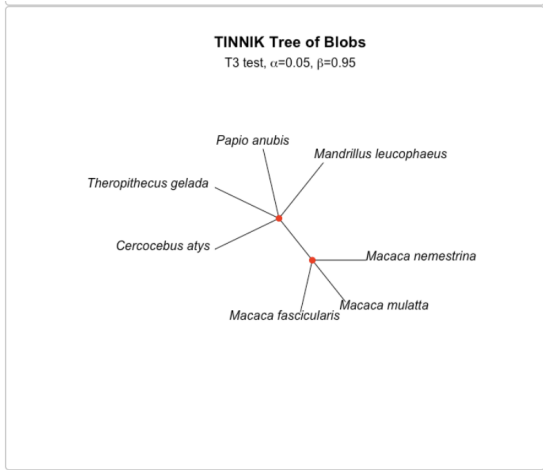
TINNiK: **T**ree of blobs **IN**ference for a species **N**etwor**K**

Allman, Baños, Mitchell, Rhodes (2024)

- Algorithm for the statistically consistent inference of the tree of blobs based on the analysis of gene quartet CFs and a combinatorial inference rule.
- Implemented in the `MSCquartets 2.0` R package.

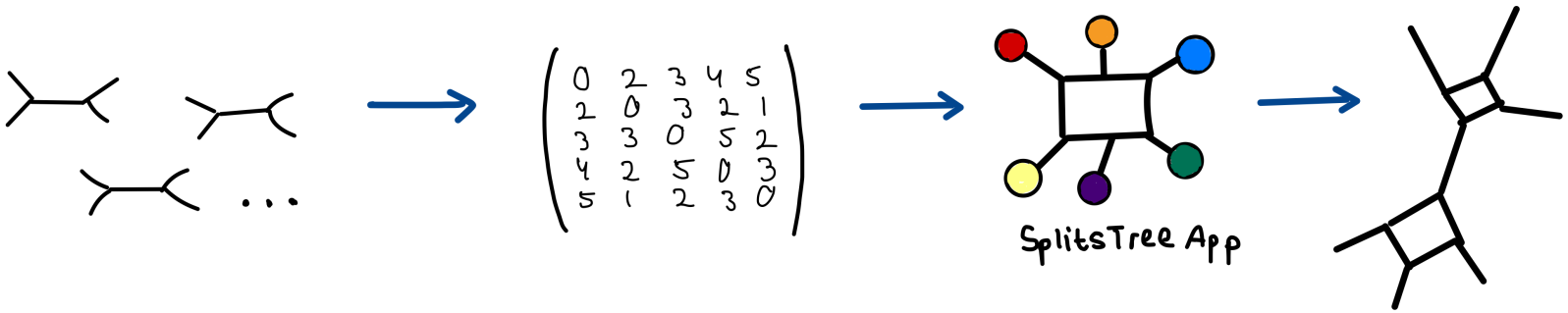






Other ways of obtaining the tree of blobs

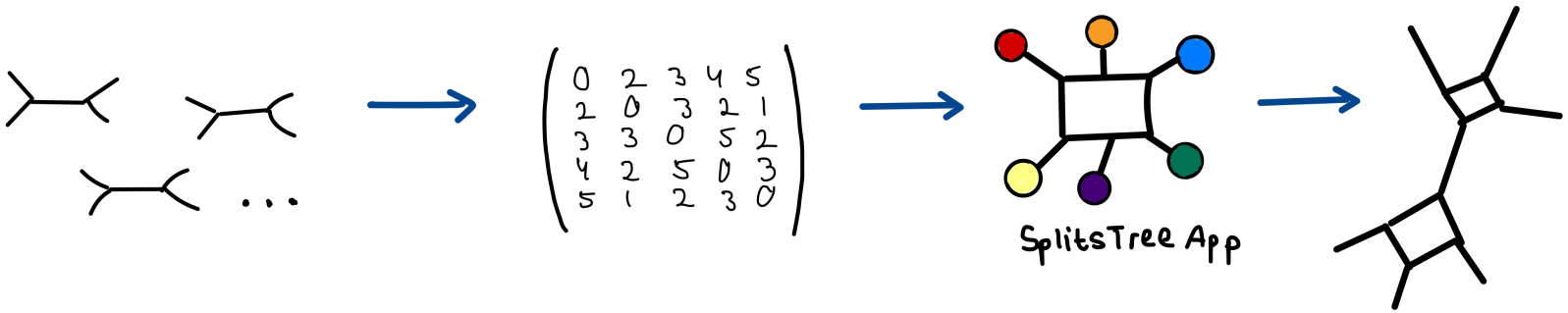
- NANUQ: **N**etwork inference **A**lgorithm via **N**eighbourNet **U**sing **Q**uartet distance (Allman, Baños, Rhodes (2019))



$\longrightarrow$ Obtain the tree of blobs from the NANUQ splits graph by contracting cycles

Other ways of obtaining the tree of blobs

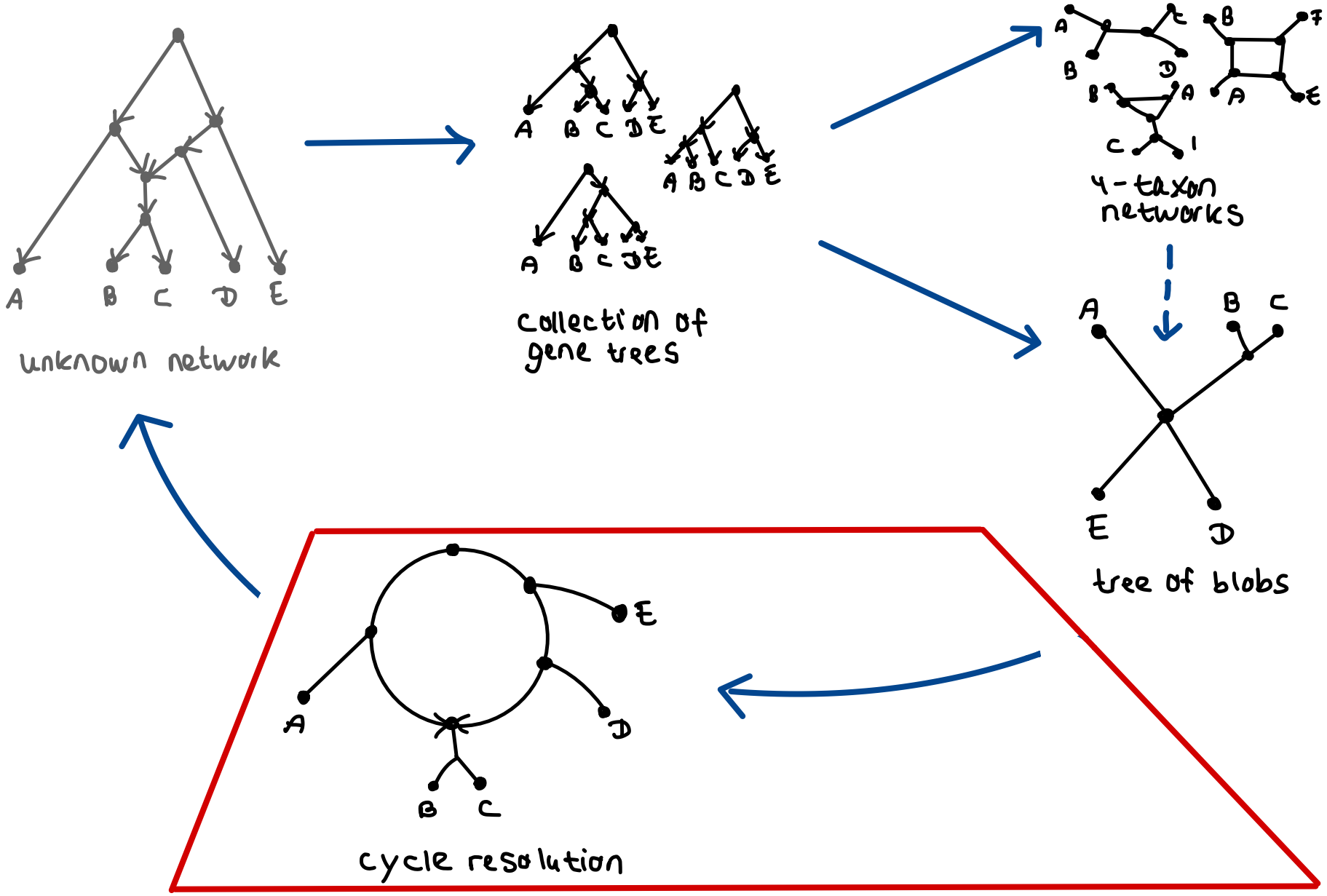
- NANUQ: **N**etwork inference **A**lgorithm via **N**eighbourNet **U**sing **Q**uartet distance (Allman, Baños, Rhodes (2019))



$\longrightarrow$ Obtain the tree of blobs from the NANUQ splits graph by contracting cycles

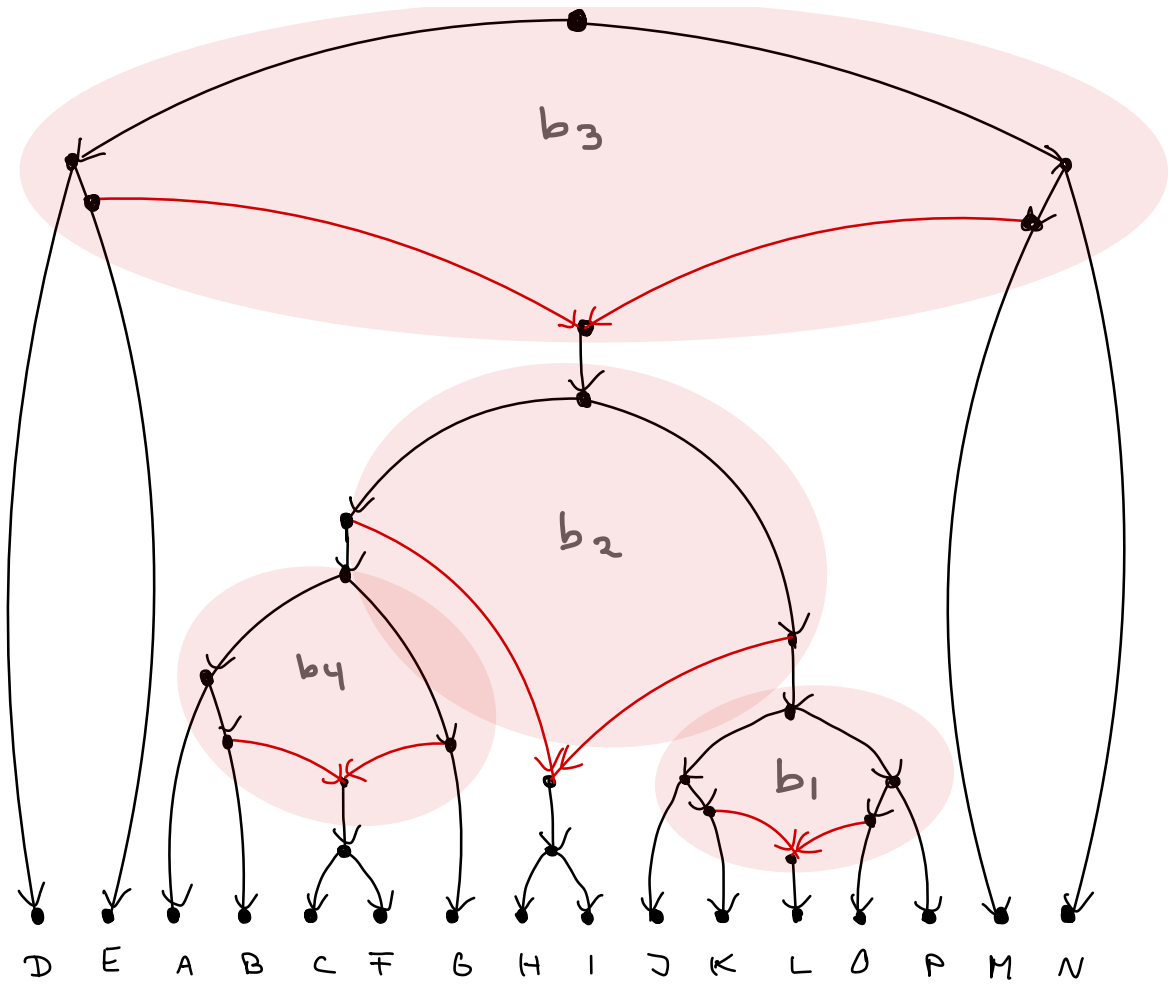
- Your method!

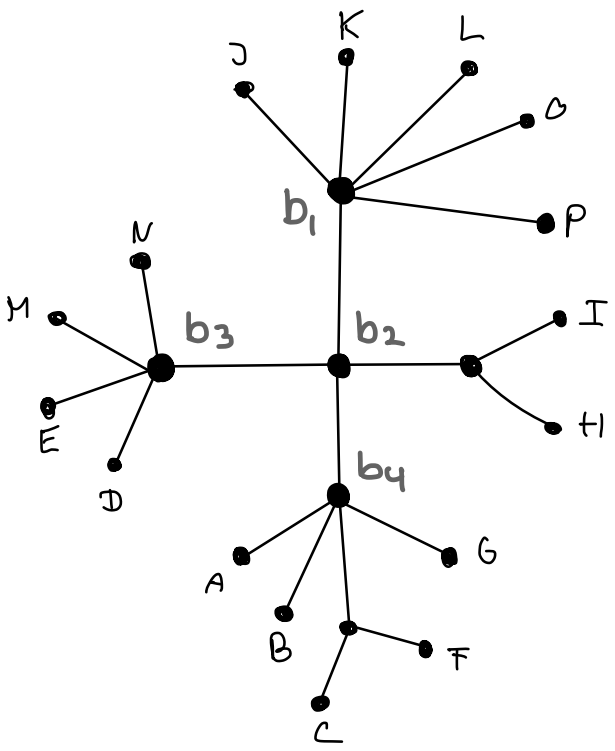
Big picture


A B C D E
unknown network
A B C D E
A B C D E
A B C D E
collection of gene trees
4-taxon networks
A B C
E D
tree of blobs
E
A
D
B C
cycle resolution

From the tree of blobs to a level-1 network

Each multifurcation in the unrooted tree of blobs of a level-1 network corresponds to a simple cycle in the network.

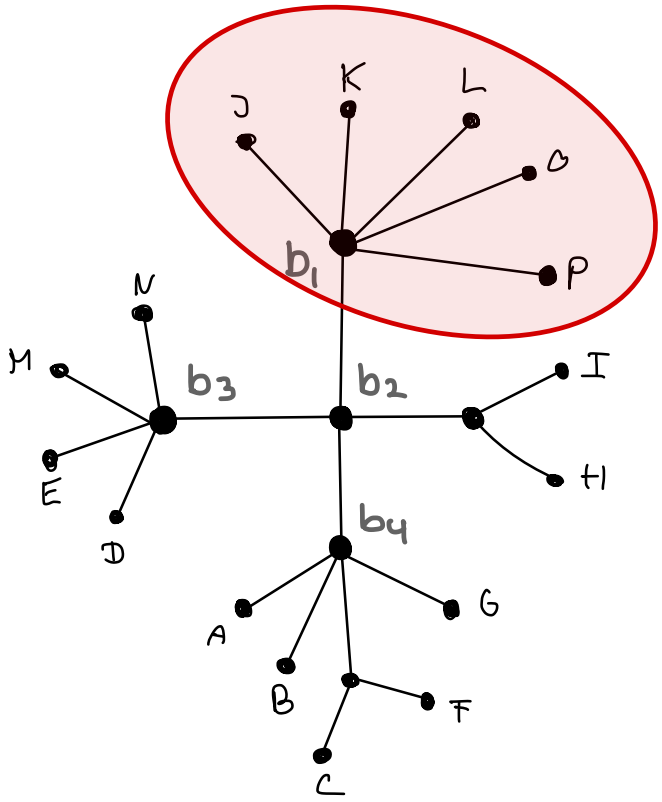


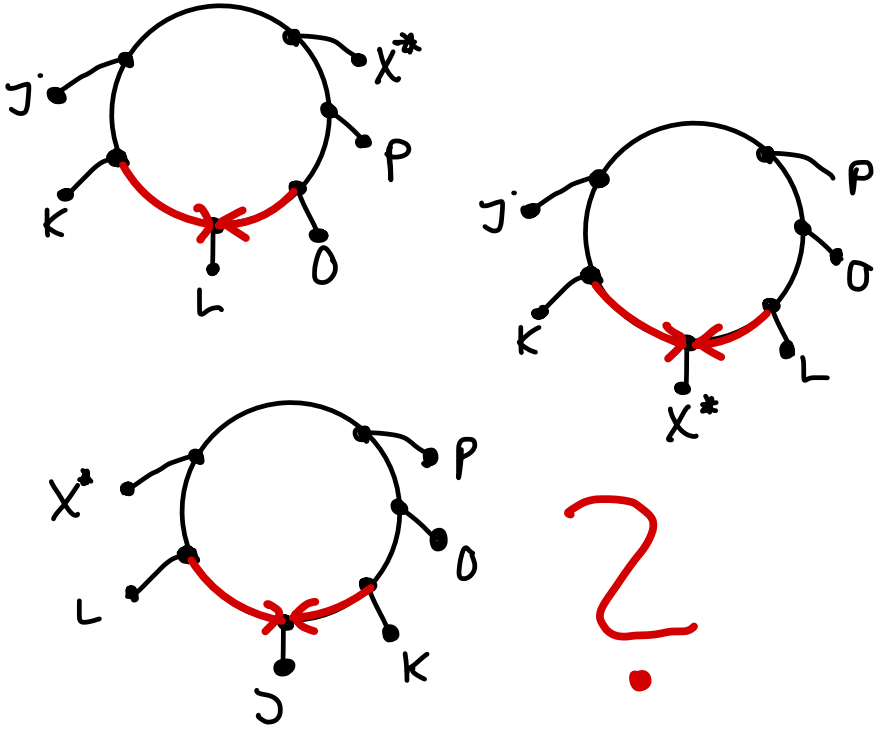


From the tree of blobs to a level-1 network

Idea:

- Focus on one multifurcation at a time and find an optimal cycle resolution for it.
- Repeat this for all multifurcations.
- Combine the cycle resolutions into a full level-1 network (if possible).



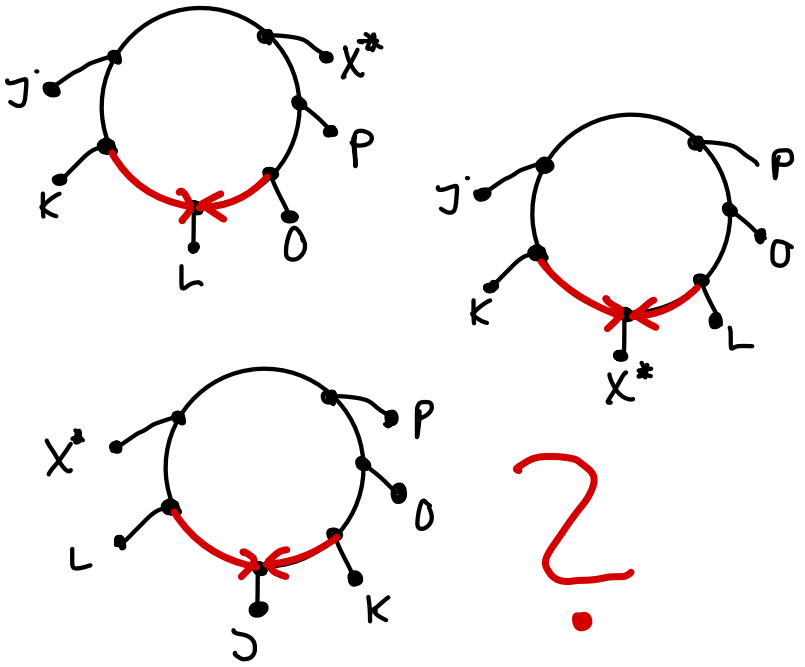


Let $X^* = \{A, B, C, D, E, F, G, H, I, M, N\}$

Inferring an optimal cycle structure

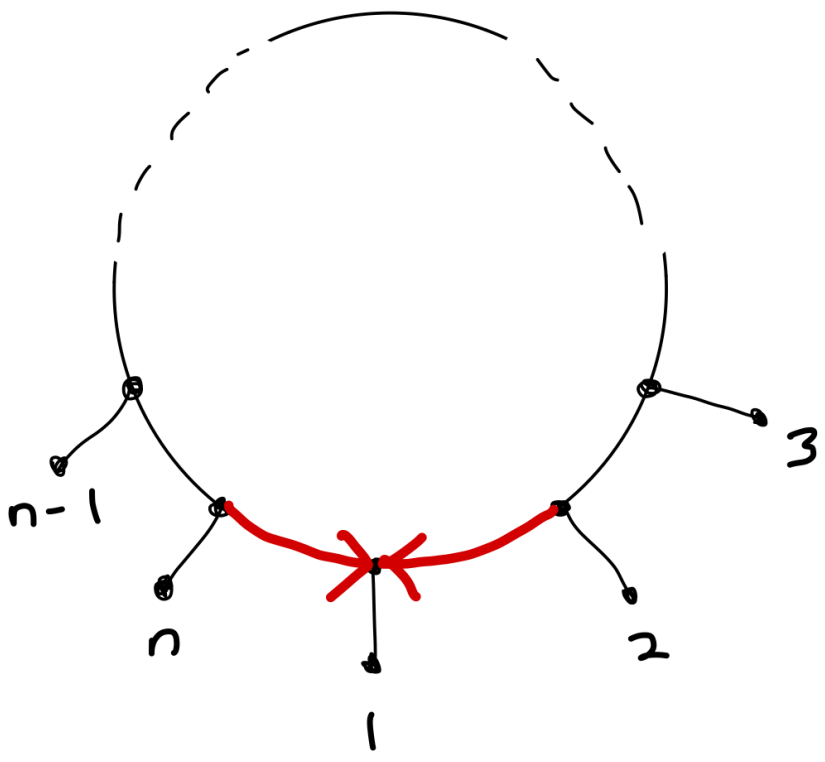
We use a least-squares approach, comparing an empirical quartet-based distance relating groups of taxa around the multifurcation to an expected one for each possible ordering of the groups and choice of hybrid node:

- Exhaustive search can be done quickly for cycles of size ≤ 10;
- Heuristic method for larger cycles.



Sunlet networks

A semi-directed level-1 network on X with $|X| = n$ is called an ***n*-sunlet** if it contains precisely one cycle such that (i) all nodes in the cycle are adjacent to precisely on element of X; and (ii) each element of X is adjacent to precisely one node in the cycle.



A parametric family of quartet distances

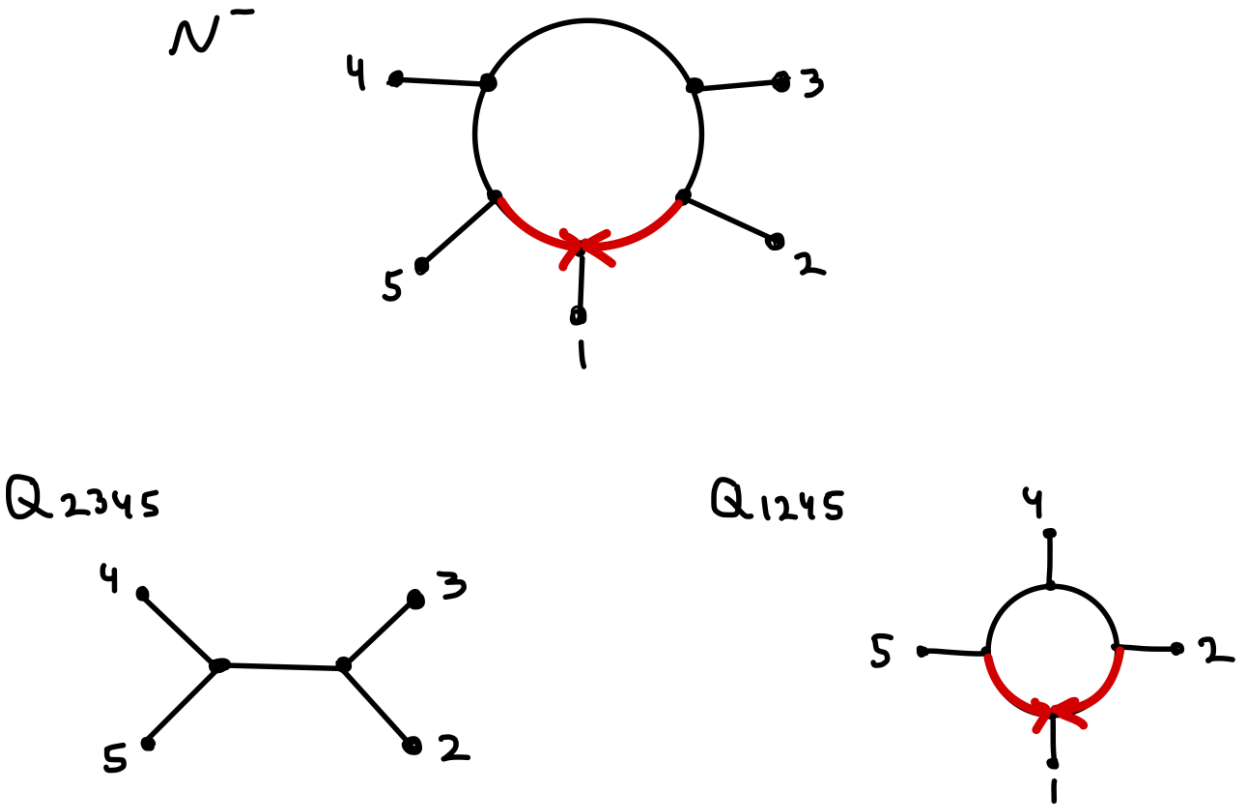
Let Q_{xyzw} be an induced quartet of a semi-directed level-1 network N^-, and let $\widetilde{Q}_{xyzw}$ be the network obtained from it by contracting all 2- and 3-cycles, and suppressing degree-2 nodes, so $\widetilde{Q}_{xyzw}$ is either a tree or has a single 4-cycle. Fix any $\boldsymbol{\rho} \in (\mathbb{R}^{\geq 0})^4$. With

$$\rho_{xy}(Q_{xyzw}) = \begin{cases} \rho_c & \text{if } \widetilde{Q}_{xyzw} \text{ has form } xy|zw, \\ \rho_s & \text{if } \widetilde{Q}_{xyzw} \text{ has form } xz|yw \text{ or } xw|yz, \\ \rho_a & \text{if } \widetilde{Q}_{xyzw} \text{ has a 4-cycle with } x, y \text{ adjacent}, \\ \rho_o & \text{if } \widetilde{Q}_{xyzw} \text{ has a 4-cycle with } x, y \text{ not adjacent}, \end{cases}$$

the **quartet distance** $d_{\boldsymbol{\rho}}^{N^-}$ is defined as

$$d_{\boldsymbol{\rho}}^{N^-}(x, y) = 2 \sum_{z,w \neq x,y} \rho_{xy}(Q_{xyzw}) + 2n - 4. \tag{1}$$

A parametric family of quartet distances



For example, consider taxa 4 and 5. Then, $\rho_{45}(Q_{2345}) = \rho_c$ and $\rho_{45}(Q_{1245}) = \rho_a$.

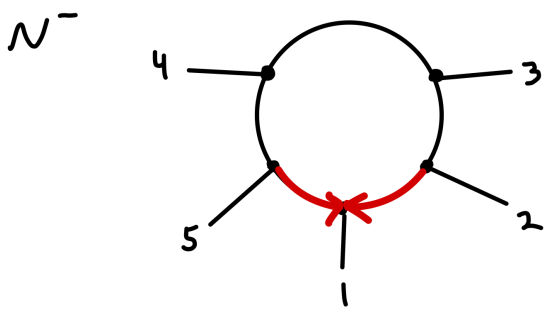
The NANUQ and modified NANUQ distances

- **NANUQ distance** (Allman et al., 2019):

$$\boldsymbol{\rho}_{\mathrm{NQ}} = (\rho_c, \rho_s, \rho_a, \rho_o) = (0, 1, 0.5, 1)$$

- **Modified NANUQ distance**:

$$\boldsymbol{\rho}_{\mathrm{MN}} = (\rho_c, \rho_s, \rho_a, \rho_o) = (0.5, 1, 0.5, 1)$$



$$d^{N^-}_{\boldsymbol{\rho}_{\mathbf{NQ}}} = \begin{pmatrix} 0 & 9 & 11 & 11 & 9 \\ 9 & 0 & 8 & 11 & 12 \\ 11 & 8 & 0 & 10 & 11 \\ 11 & 11 & 10 & 0 & 8 \\ 9 & 12 & 11 & 8 & 0 \end{pmatrix} \quad \text{and} \quad d^{N^-}_{\boldsymbol{\rho}_{\mathbf{MN}}} = \begin{pmatrix} 0 & 9 & 11 & 11 & 9 \\ 9 & 0 & 9 & 11 & 12 \\ 11 & 9 & 0 & 10 & 11 \\ 11 & 11 & 10 & 0 & 9 \\ 9 & 12 & 11 & 9 & 0 \end{pmatrix}$$

The NANUQ and modified NANUQ distances

Both the NANUQ and modified NANUQ distances allow us to recover n-sunlets:

Theorem (Allman, Baños, Rhodes (2019))

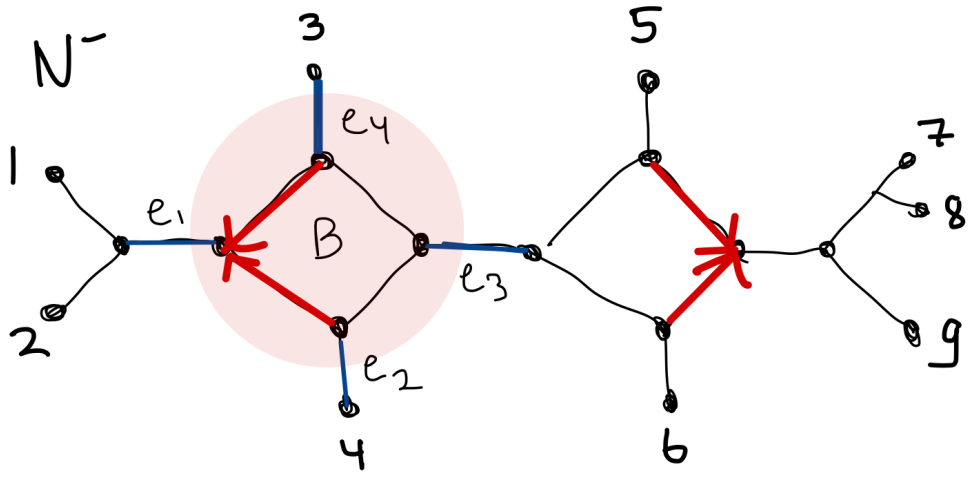
Let N^- be an m-sunlet network. Then from $d^{N^-}_{\rho_{\mathrm{NQ}}}$ the circular order of the taxa around the cycle are identifiable. If $m > 4$, then the hybrid taxon is also identifiable.

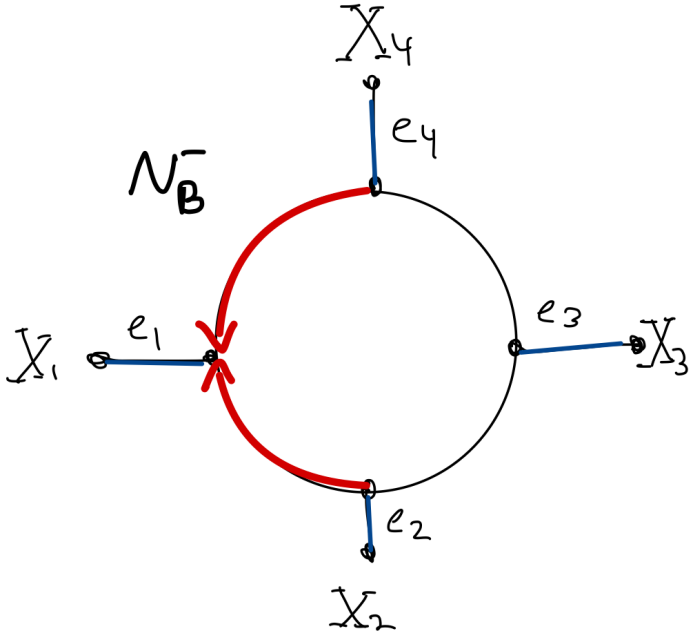
Theorem (Allman, Baños, Rhodes, W (2024+))

Let N^- be an m-sunlet network. Then from $d^{N^-}_{\rho_{\mathrm{MN}}}$ the circular order of the taxa around the cycle is identifiable. If $m > 4$, then the hybrid taxon is also identifiable.

Generalized sunlets

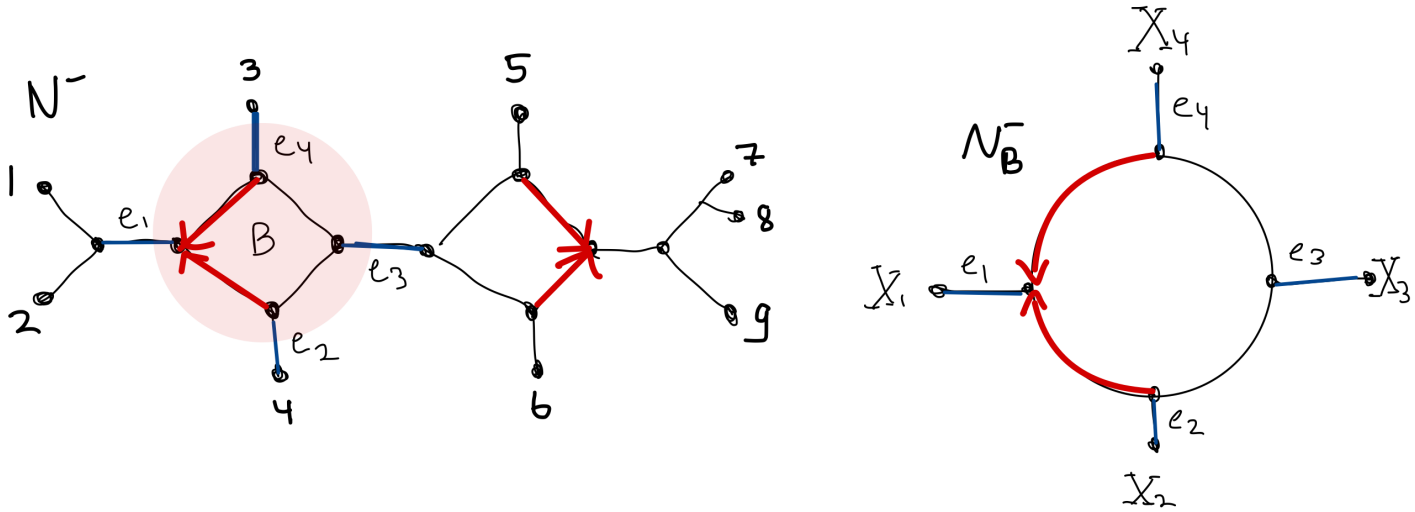
Let B be a blob of N^-. Let $e_1, \ldots, e_m$ denote the cut edges incident with B yielding a partition $X_1 \sqcup \ldots \sqcup X_m$ of X. Then, the **generalized sunlet network N_B^- induced by** B is obtained from N^- as follows: Delete all nodes and edges not contained in or incident with B. Then label the degree-1 node of each cut edge e_i incident with B by X_i.





Here, $X_1 = \{1, 2\}$, $X_2 = \{4\}$, $X_3 = \{5, 6, 7, 8, 9\}$, and $X_4 = \{3\}$.

Group distances for a blob from the generalized sunlet



- For any choice of $\boldsymbol{\rho}$, we have a distance $d_{\boldsymbol{\rho}}$ on the labels $\{X_i\}$ on N_B^- using the form of the induced quartet network on 4 of these X_i.
- This induced quartet network must have the same form as the full network's induced quartet network on taxa x_i chosen from each X_i.
- However, to estimate this distance from data, we must allow for the fact that different choices of $x_i \in X_i$ may in fact support different *inferred quartet topologies*.

Group distances around a blob from quartets

Suppose $X = X_1 \sqcup \ldots \sqcup X_m$ is a partition of the taxa induced by an m-blob and $\mathcal{Q}$ is a collection of quartet networks on X such that for all distinct i, j, k, l and $x_i \in X_i$, $x_j \in X_j$, $x_k \in X_k$, $x_l \in X_l$, $\mathcal{Q}$ contains a level-1 quartet network $Q_{x_i x_j x_k x_l}$ on x_i, x_j, x_k, x_l. Then, the **parametric group distance** between $X_i \neq X_j$ for $\mathcal{Q}$ is

$$d_{\boldsymbol{\rho}}^{\mathcal{Q}}(X_i, X_j) = 2 \sum_{k,l \neq i,j} \frac{1}{|X_i||X_j||X_k||X_l|} \sum_{\substack{x \in X_i, y \in X_j, \\ z \in X_k, w \in X_l}} \rho_{xy}(Q_{xyzw}) + 2m - 4,$$

Group distances around a blob from quartets

Suppose $X = X_1 \sqcup \ldots \sqcup X_m$ is a partition of the taxa induced by an m-blob and $\mathcal{Q}$ is a collection of quartet networks on X such that for all distinct i, j, k, l and $x_i \in X_i$, $x_j \in X_j$, $x_k \in X_k$, $x_l \in X_l$, $\mathcal{Q}$ contains a level-1 quartet network $Q_{x_i x_j x_k x_l}$ on x_i, x_j, x_k, x_l. Then, the **parametric group distance** between $X_i \neq X_j$ for $\mathcal{Q}$ is

$$d_{\boldsymbol{\rho}}^{\mathcal{Q}}(X_i, X_j) = 2 \sum_{k,l \neq i,j} \frac{1}{|X_i||X_j||X_k||X_l|} \sum_{\substack{x \in X_i, y \in X_j, \\ z \in X_k, w \in X_l}} \rho_{xy}(Q_{xyzw}) + 2m - 4,$$

Proposition (Allman, Baños, Rhodes, W (2024+))

For a generalized sunlet N_B^- on $\{X_i\}$ induced from a level-1 network N^- on X, let $\mathcal{Q}$ denote the set of induced quartet networks Q_{xyzw} for all choice of $x, y, z, w \in X$ from four distinct X_i. Then

$$d_{\boldsymbol{\rho}}^{N_B^-} = d_{\boldsymbol{\rho}}^{\mathcal{Q}}.$$

Cycle resolution

Algorithm ResolveCycle

Input: Unrooted tree of blobs T on X with a designated m-multifurcation representing a blob B; a collection $\mathcal{Q}$ of level-1 quartet topologies for each set of 4 taxa drawn from different taxon groups around the blob; $\boldsymbol{\rho}$.
Output: A circular order of the taxon groups for B and, if $m > 4$, a designation of the hybrid group.

1. Compute $d_{\boldsymbol{\rho}}^{\mathcal{Q}}$ and $d_{\boldsymbol{\rho}}^{C}$ for an m-sunlet C.
2. For each circular order of the X_i, and if $m > 4$, a designated hybrid group, compute the ordinary least-squares residual r between $d_{\boldsymbol{\rho}}^{\mathcal{Q}}$ and the reordered $d_{\boldsymbol{\rho}}^{C}$.
3. Return the circular order and, if $m > 4$, the hybrid group giving the minimal r.

Cycle resolution

Algorithm ResolveCycle

Input: Unrooted tree of blobs T on X with a designated m-multifurcation representing a blob B; a collection $\mathcal{Q}$ of level-1 quartet topologies for each set of 4 taxa drawn from different taxon groups around the blob; $\boldsymbol{\rho}$.
Output: A circular order of the taxon groups for B and, if $m > 4$, a designation of the hybrid group.

1. Compute $d_{\boldsymbol{\rho}}^{\mathcal{Q}}$ and $d_{\boldsymbol{\rho}}^{C}$ for an m-sunlet C.
2. For each circular order of the X_i, and if $m > 4$, a designated hybrid group, compute the ordinary least-squares residual r between $d_{\boldsymbol{\rho}}^{\mathcal{Q}}$ and the reordered $d_{\boldsymbol{\rho}}^{C}$.
3. Return the circular order and, if $m > 4$, the hybrid group giving the minimal r.

Step 2 is potentially limiting computationally. For larger cycles, we use a heuristic.

Tree of blobs resolution

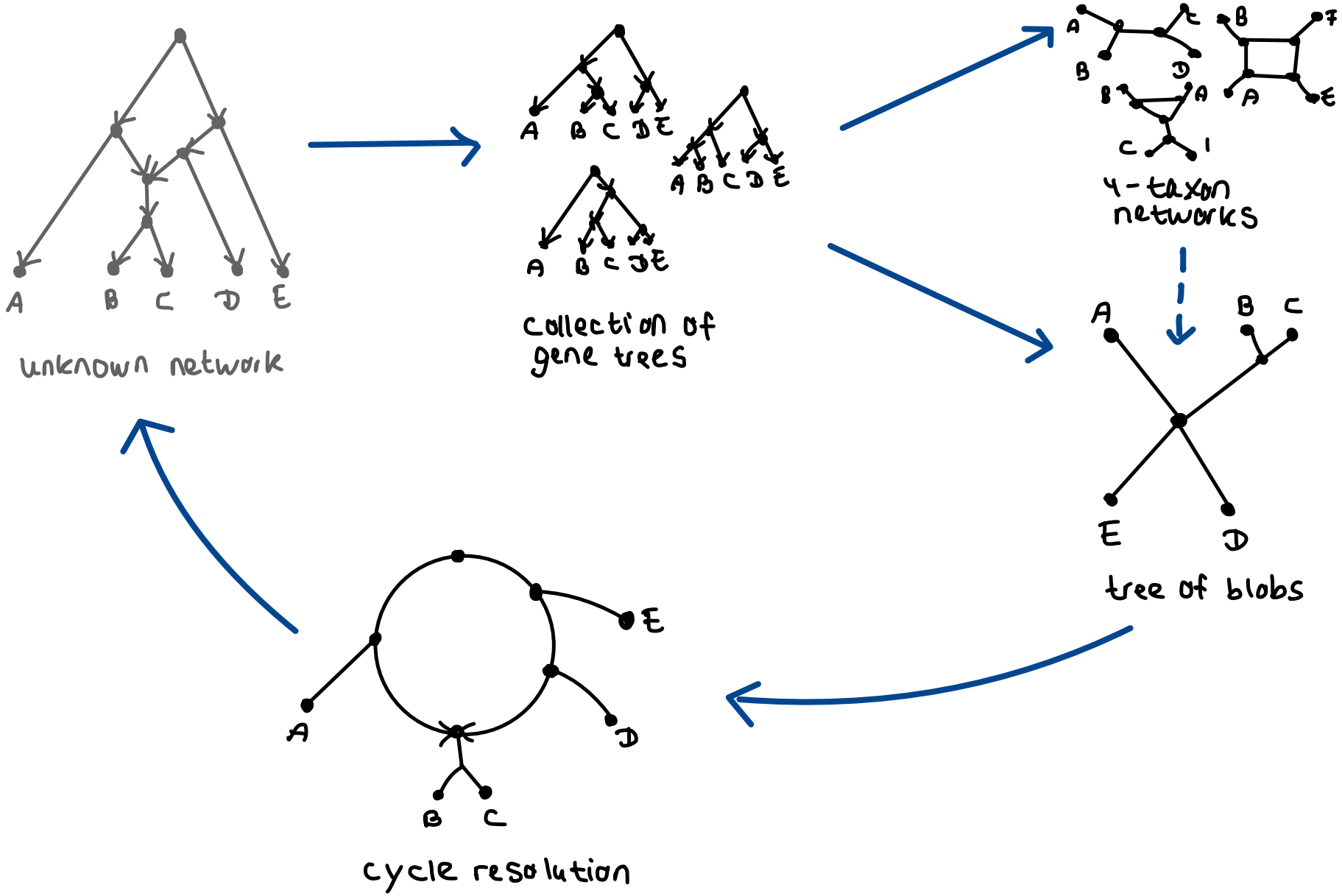
Algorithm Tree of Blobs Resolution

Input: Unrooted tree of blobs T on X; a collection $\mathcal{Q}$ of level-1 quartet topologies for each set of 4 taxa drawn from different taxon groups around the blob; $\boldsymbol{\rho}$.

Output: A binary level-1 network resolving the tree of blobs, or FAIL

1. Resolve each blob on T;
2. Determine if the designated hybrid nodes are compatible (in the sense of permitting a rooting of the network);
3. If so, return this level-1 network; otherwise return FAIL.

Big picture

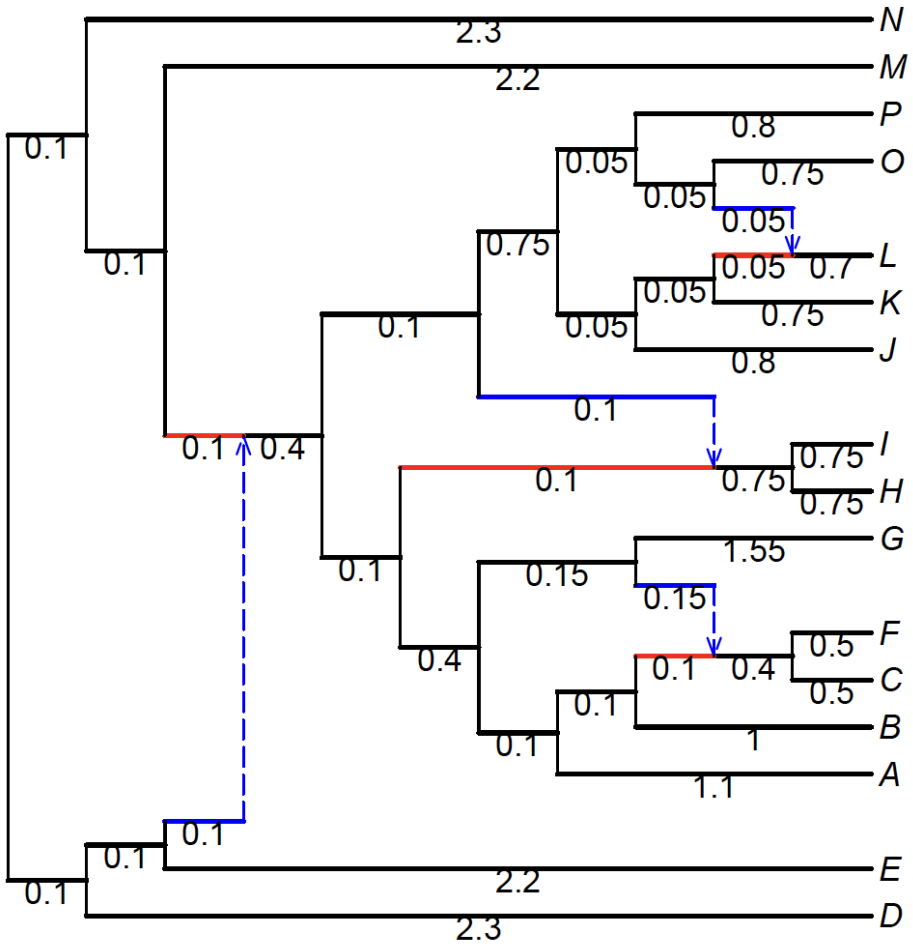

A B C D E
unknown network
A B C D E
A B C D E
A B C D E
collection of gene trees
4-taxon networks
A B C
E D
tree of blobs
E
A
D
B C
cycle resolution

Implementation

Our divide-and-conquer approach for resolving a tree of blobs into a level-1 network will be added to the R package `MSCquartets`. It will be possible to

- Resolve individual cycles,
- Combine cycle resolutions for different cycles,
- Resolve the full tree of blobs.

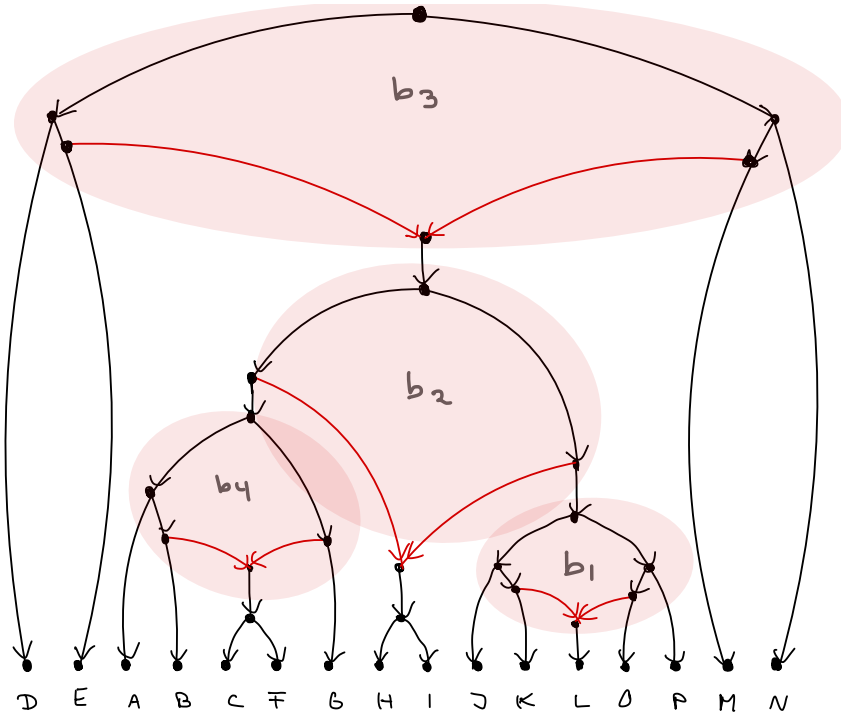
Simulation study

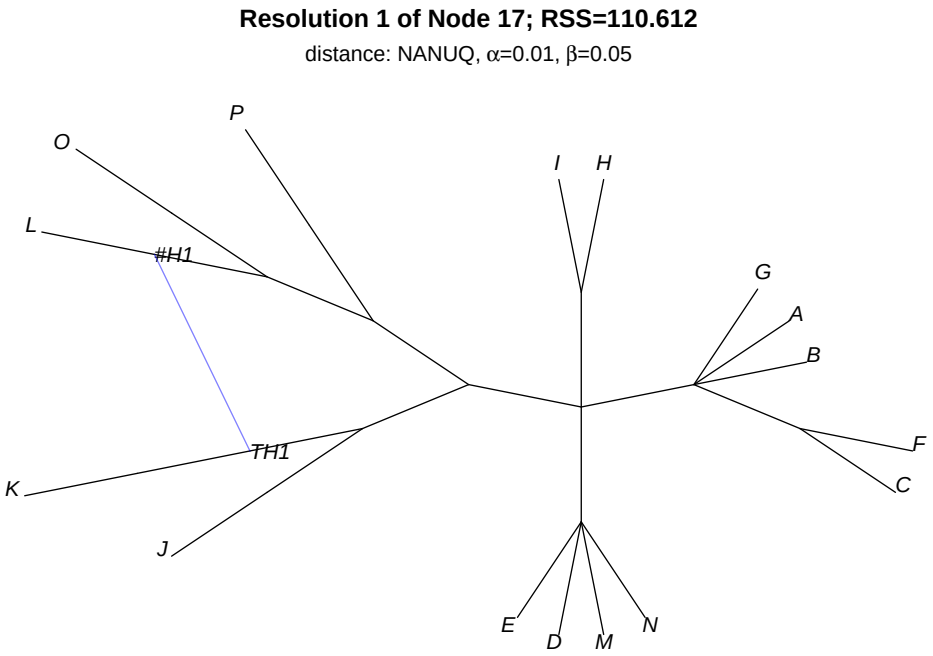


- Simulated m gene trees with `PhyloCoalSimulations` (Fogg et al., 2023)
  $m = 300, 500, 1000, 10000$
- Varied amount of ILS by scaling network branch lengths
  $k = 1.0, 1.5, 2.0, 4.0$

Simulation study – Results

$k = 1.0,\ m = 10,000,\ \alpha = 0.01,\ \beta = 0.05$

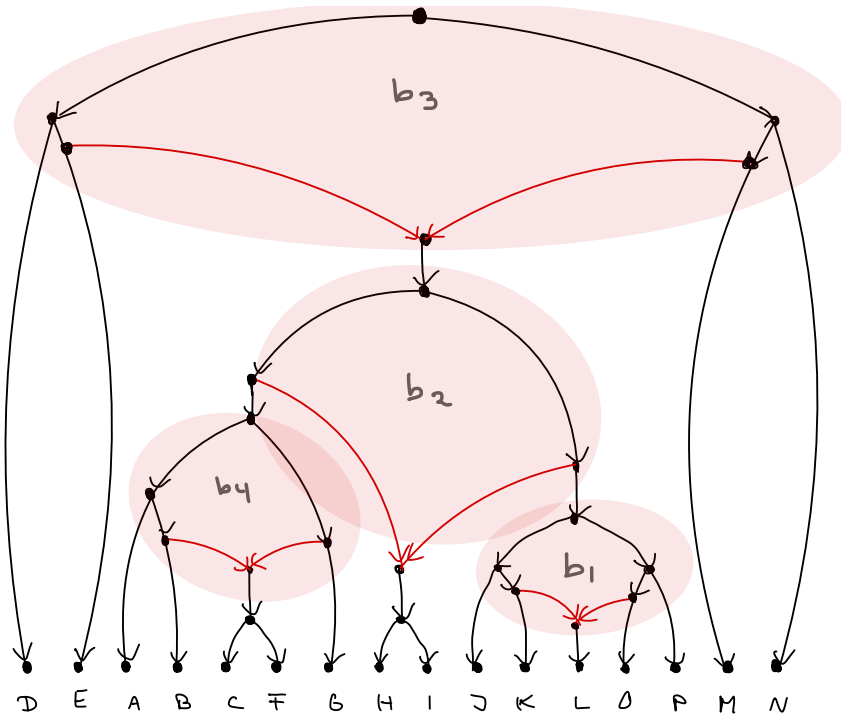


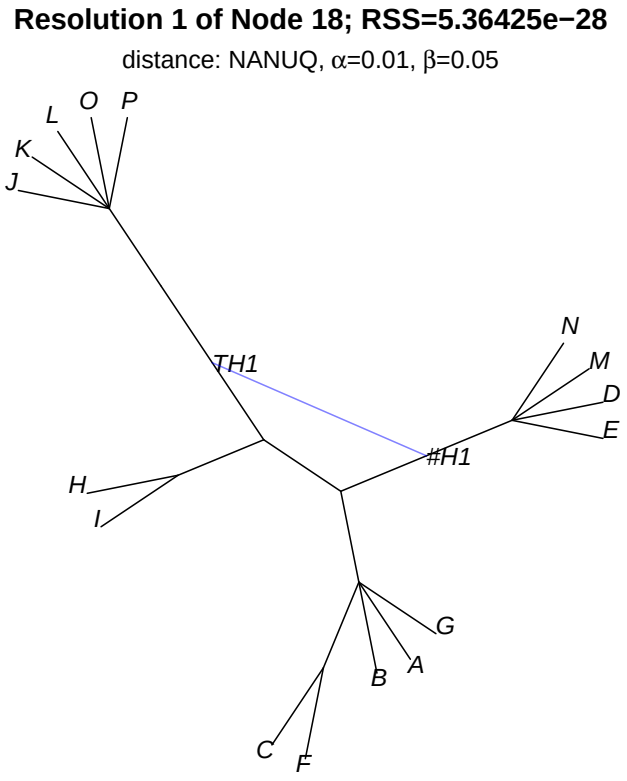


Network should be semidirected; Rooting is arbitrary.

Simulation study – Results

$k = 1.0,\ m = 10,000,\ \alpha = 0.01,\ \beta = 0.05$

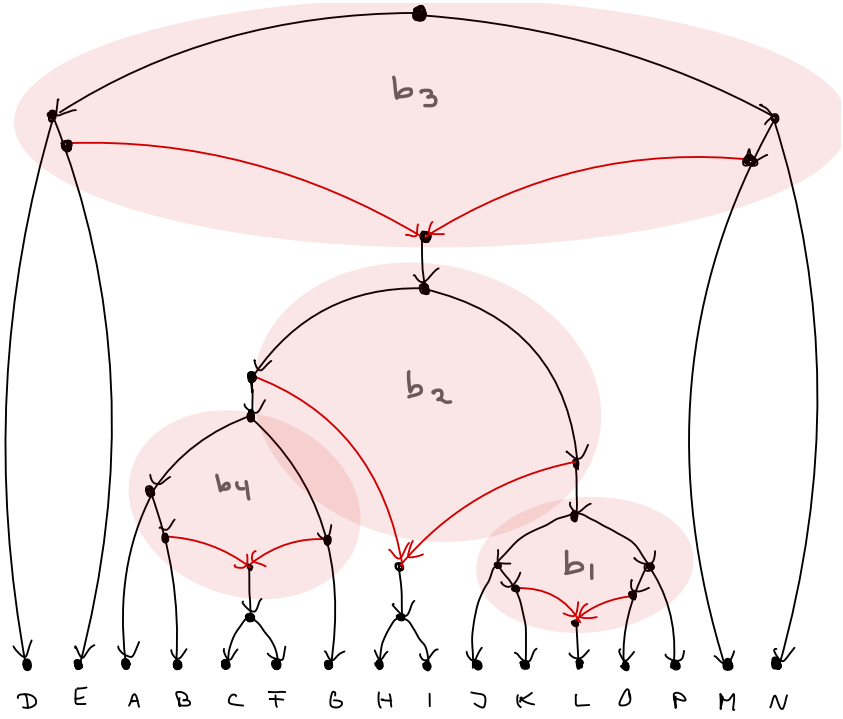


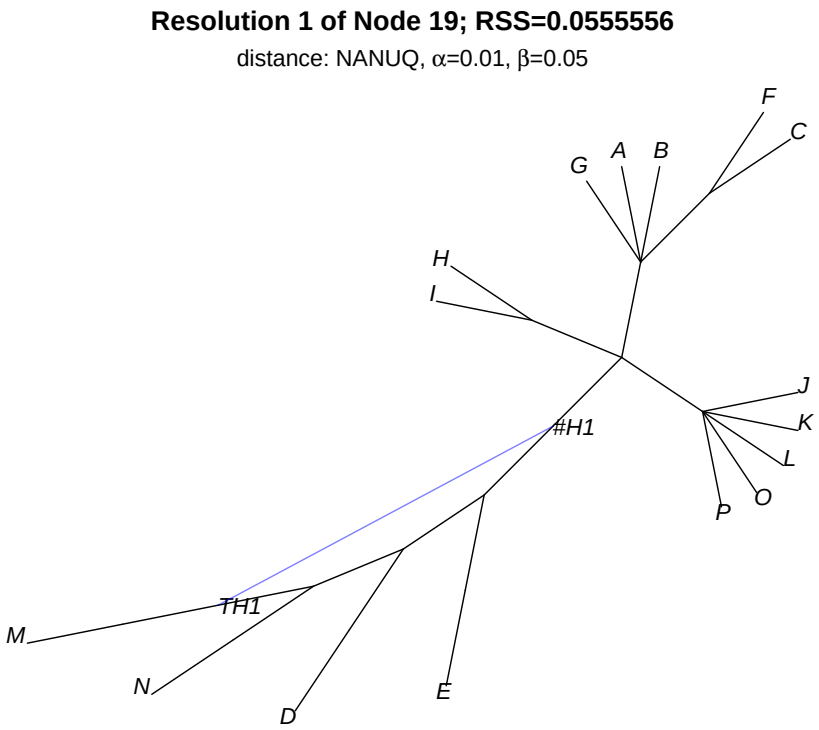


Network should be semidirected; Rooting is arbitrary.
Hybrid node on 4−cycle is not identifiable.

Simulation study – Results

$k = 1.0,\ m = 10,000,\ \alpha = 0.01,\ \beta = 0.05$

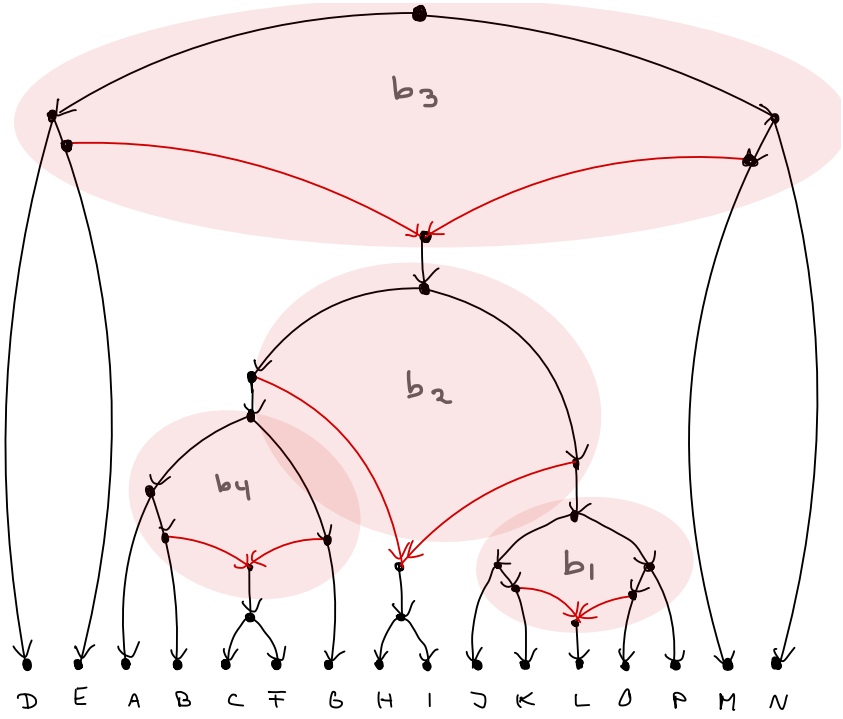


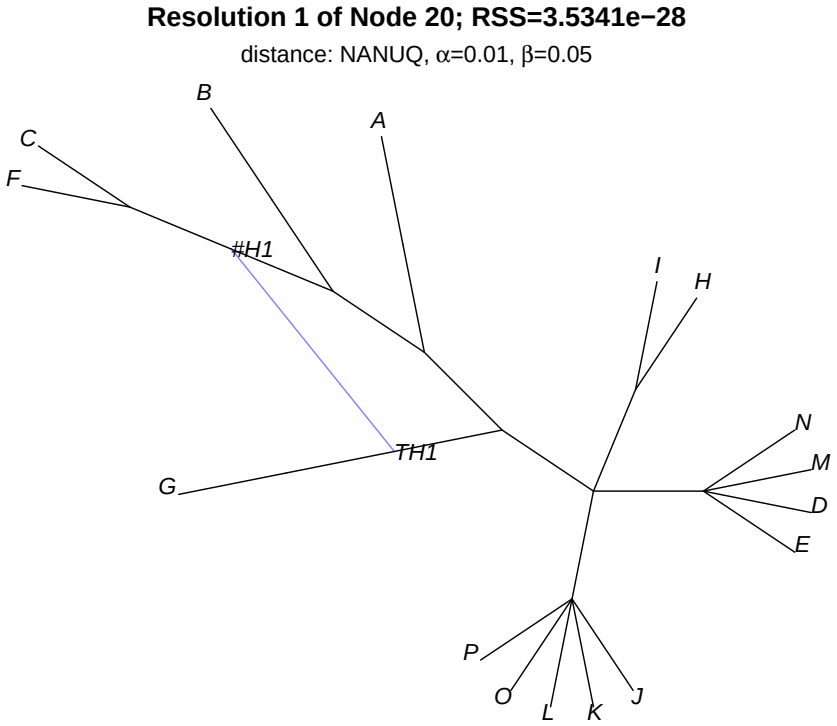


Network should be semidirected; Rooting is arbitrary.

Simulation study – Results

$k = 1.0,\ m = 10,000,\ \alpha = 0.01,\ \beta = 0.05$

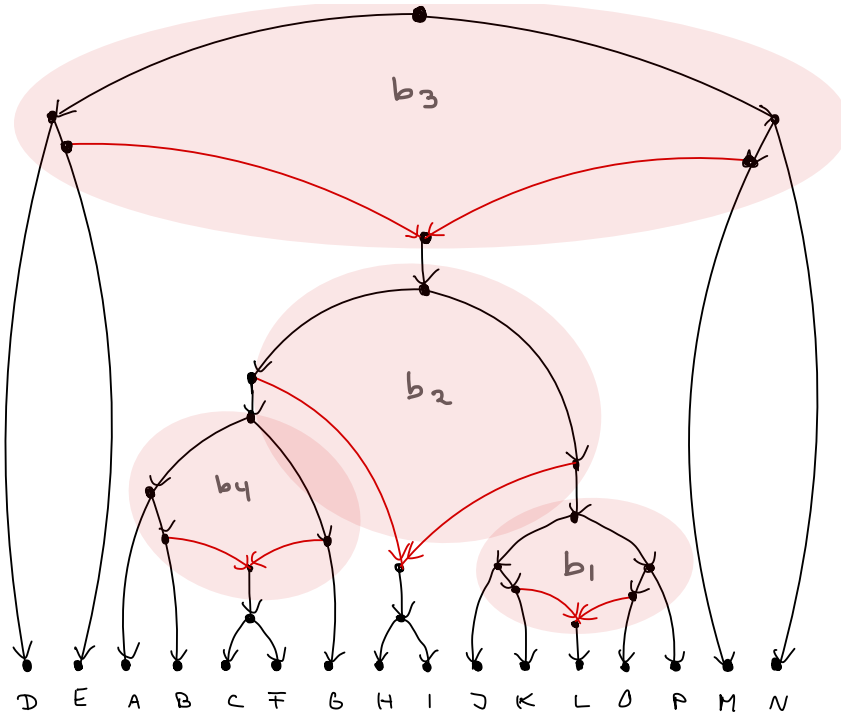




Network should be semidirected; Rooting is arbitrary.

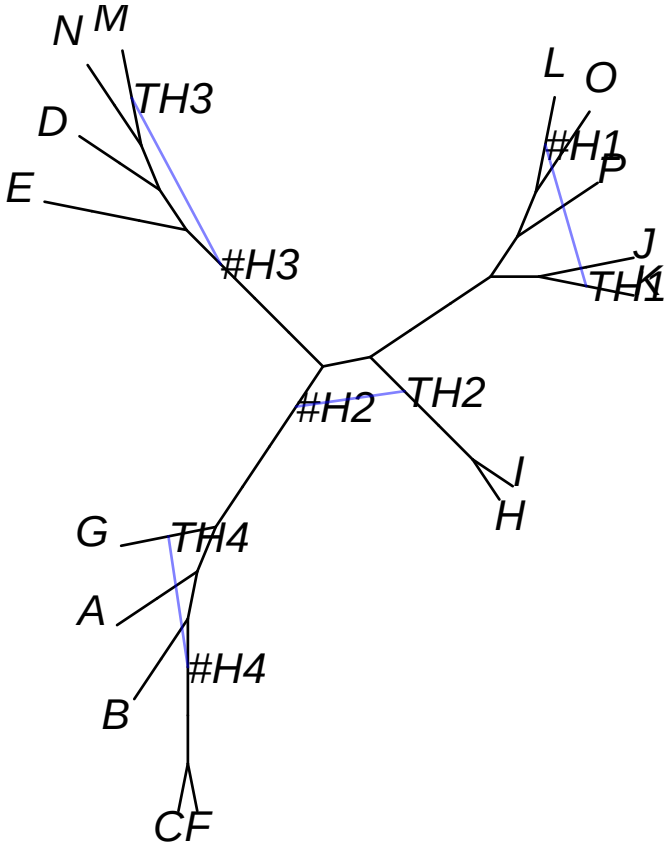
Simulation study – Results

$k = 1.0,\ m = 10,000,\ \alpha = 0.01,\ \beta = 0.05$



Inferred Level−1 Network

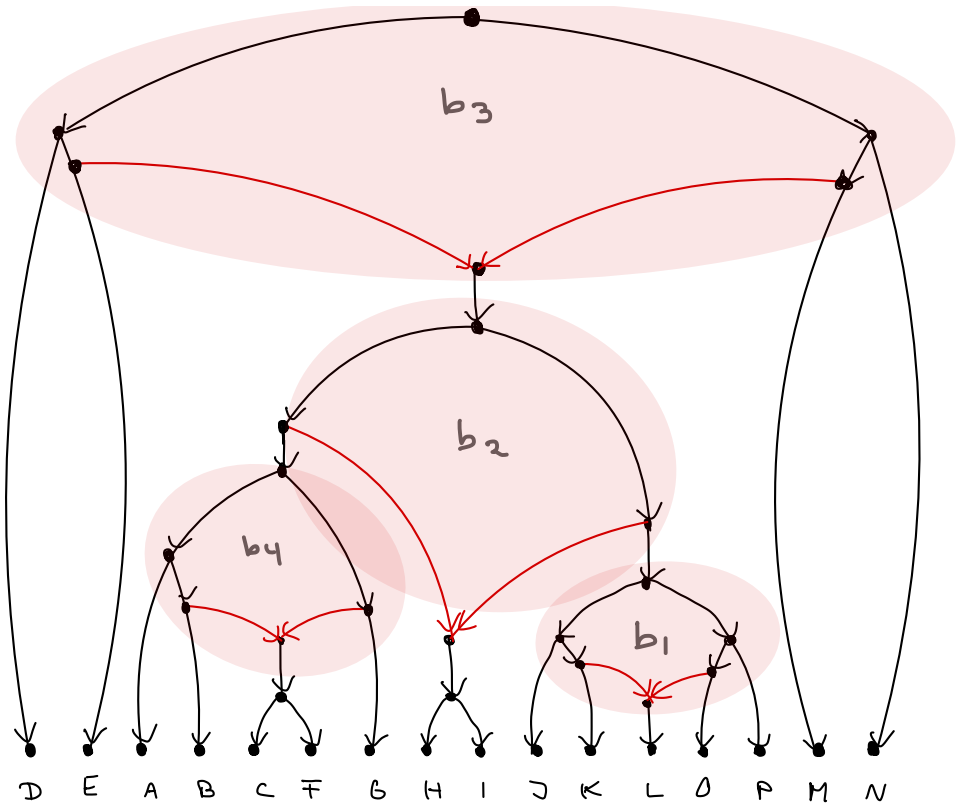
distance: NANUQ, α=0.01, β=0.05

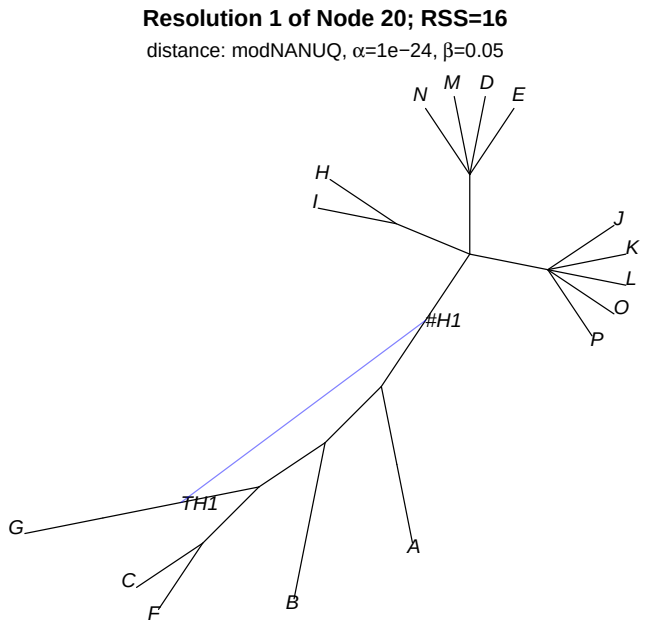


Simulation study – Results

$k = 1.0,\ m = 10,000,\ \alpha = 1e-24,\ \beta = 0.05$

Blobs b_1, b_2, b_3 as before, but tie for b_4:

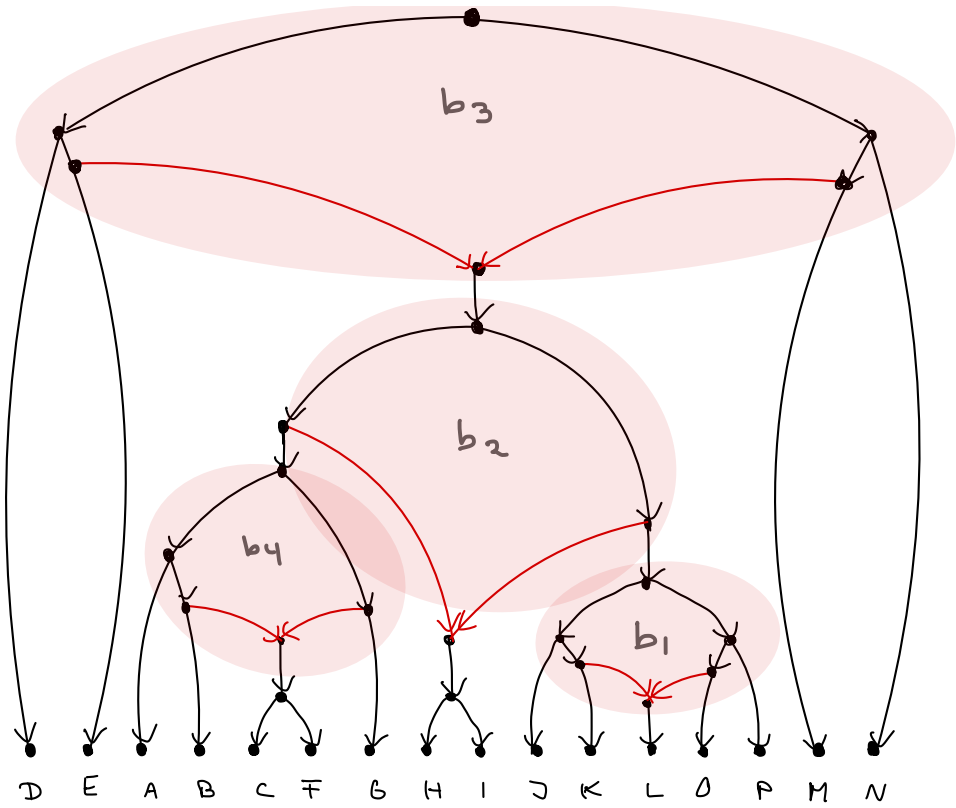


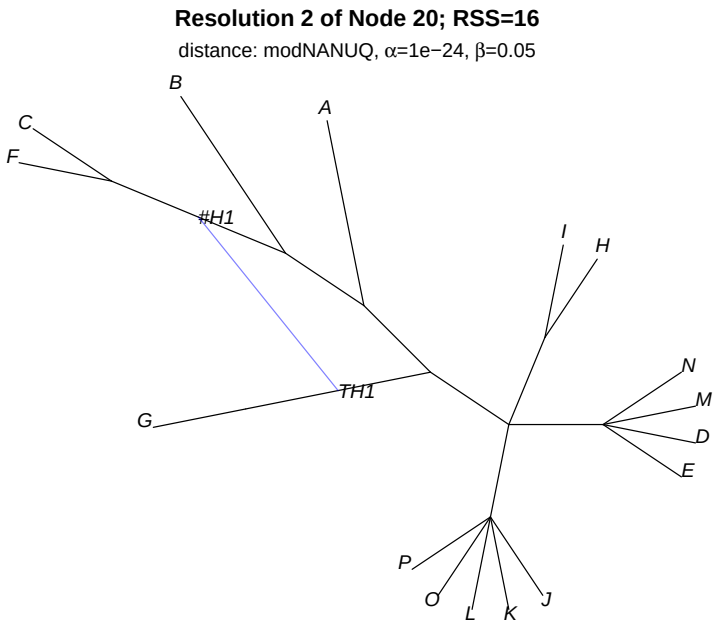


Simulation study – Results

$k = 1.0,\ m = 10,000,\ \alpha = 1e-24,\ \beta = 0.05$

Blobs b_1, b_2, b_3 as before, but tie for b_4:





Network should be semidirected; Rooting is arbitrary.

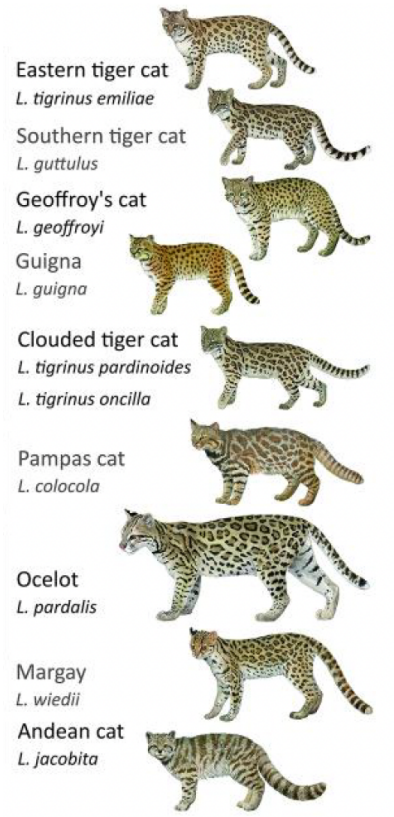
Simulation study – Results

We will present more extensive results in a forthcoming preprint. Overall observations so far:

- Given the true tree of blobs, the approach works very well across a variety of tested settings and model networks.
- Sample size and branch lengths matter.
- Given the gene trees, the approach is very fast.

Leopardus

Lescroart et al. (2023)
Extensive Phylogenomic Discordance and the Complex Evolutionary History of the Neotropical Cat Genus *Leopardus*



- 16 taxa
- 23,136 gene trees

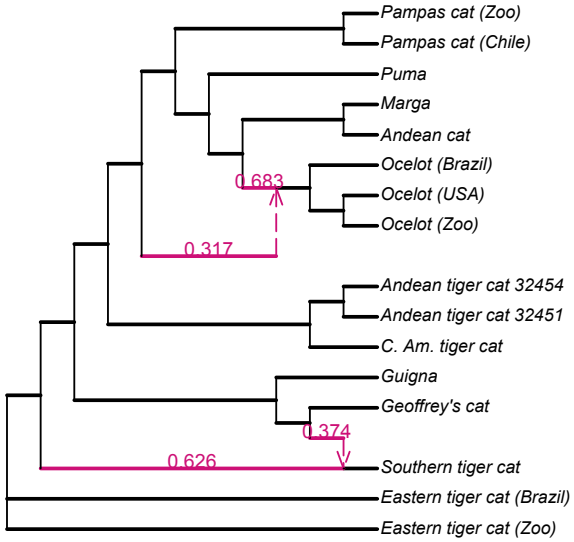
Leopardus

- Running TINNiK (with $\alpha = 5e - 29$ and $\beta = 0.95$), we obtain a tree of blobs with 2 multifurcations.
- We obtain a unique resolution for one cycle, and a 5-tie for the other cycle.
- Combining the cycle resolutions into 5 "candidate" networks, we use functionality of `PhyloNetworks` to optimize parameters under pseudolikelihood.

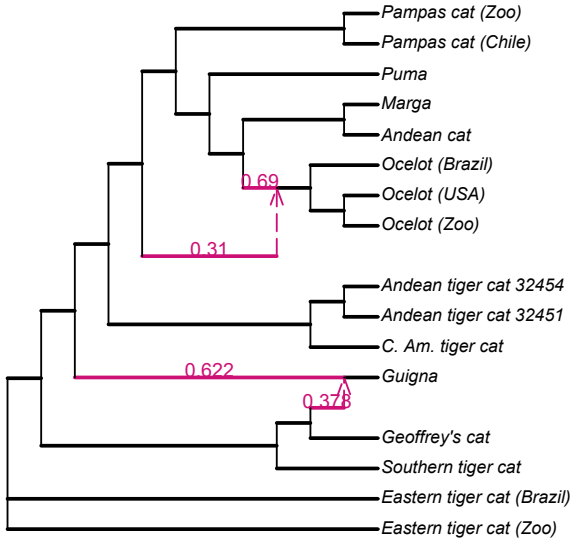
$\Rightarrow$ This is very quick ($\sim$ 30 seconds for finding candidate networks; 5 mins per candidate for optimizing parameters)!

Solis-Lemus C, Bastide P, Ane, C (2017): PhyloNetworks: a package for phylogenetic networks. Mol Biol Evol 34(12):3292–3298.

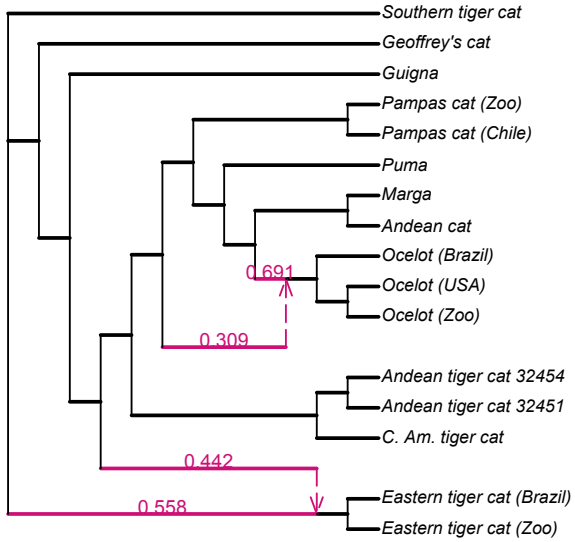
Leopardus



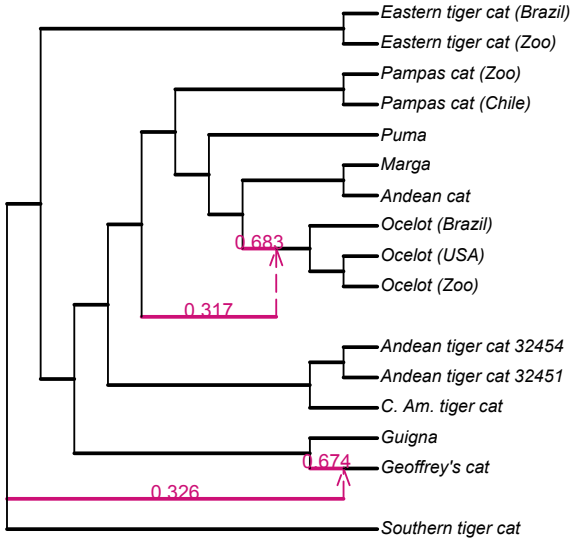
254.0



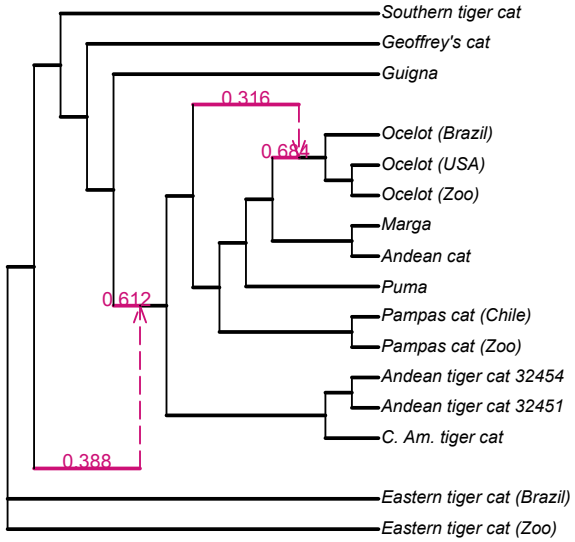
1026.2



406.0



240.2

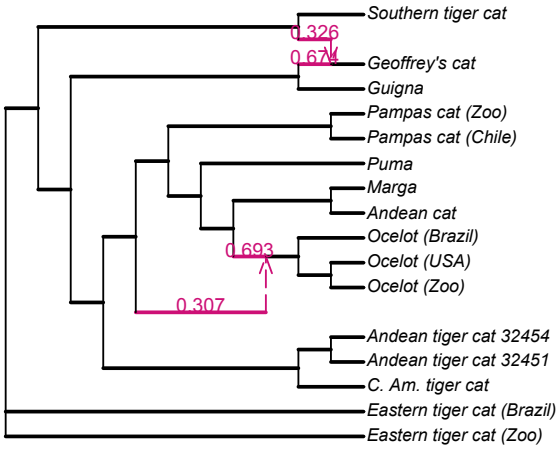


490.3

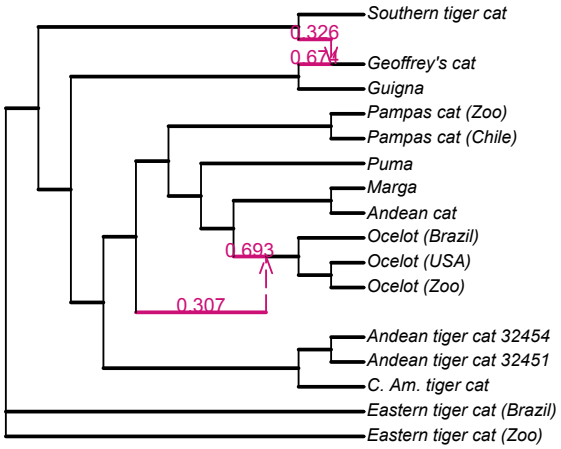
Leopardus

- Comparison with `SNaQ` (Solís-Lemus and Ané, 2016) with $h_{\max} = 2$.
- Runtime $\sim$ 9 hours (default settings).

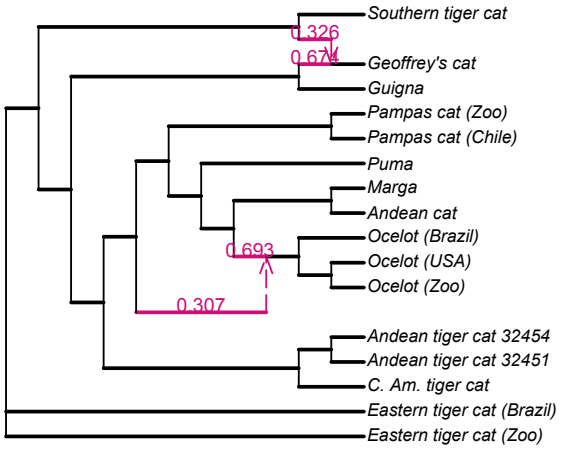
Leopardus



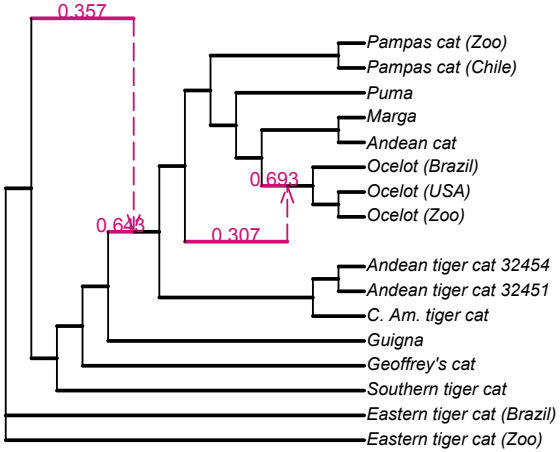
234.6



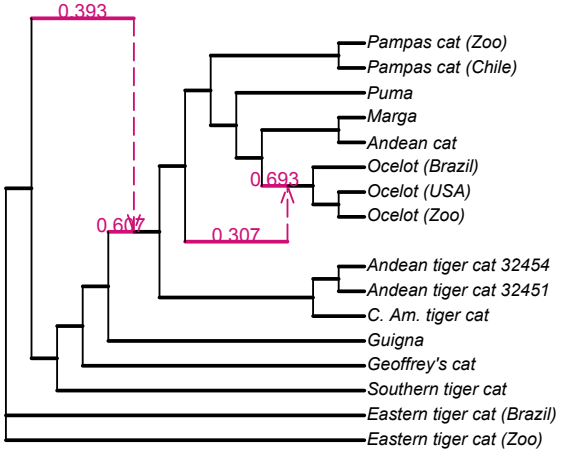
234.8



234.8



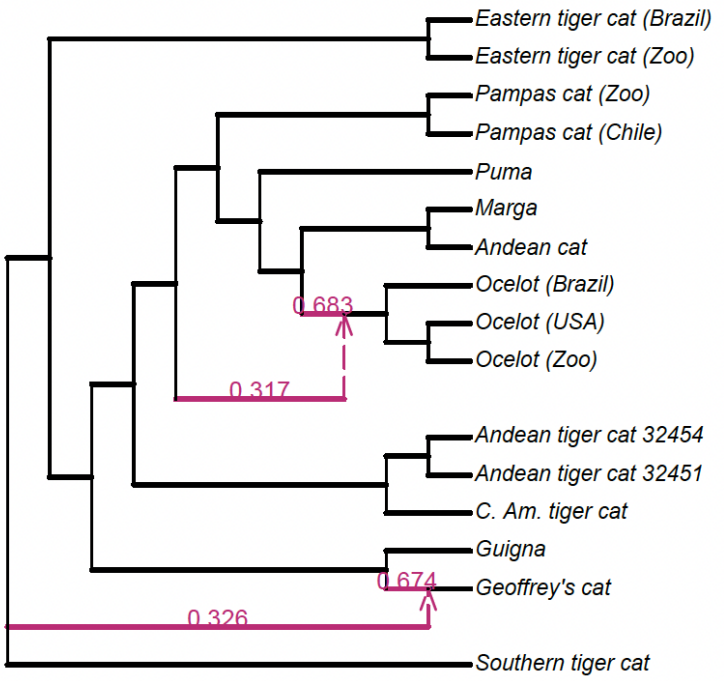
364.7



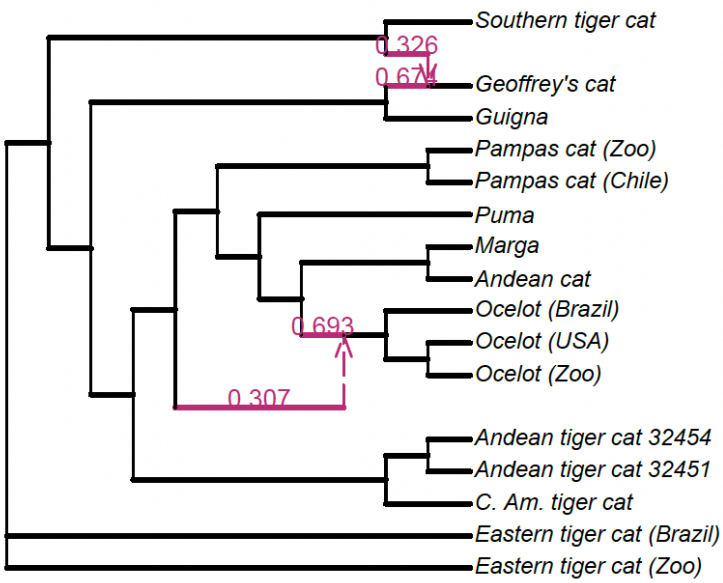
486.1

Leopardus

Our best-scoring candidate network agrees with the best-scoring network found by SNaQ (in terms of topology):



240.2

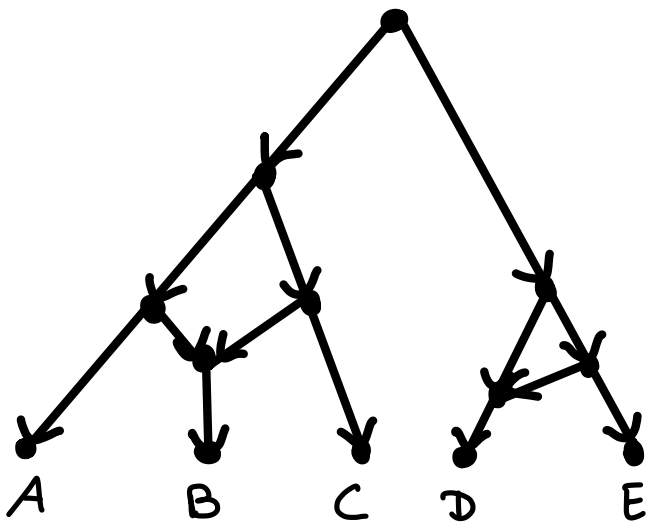


234.6

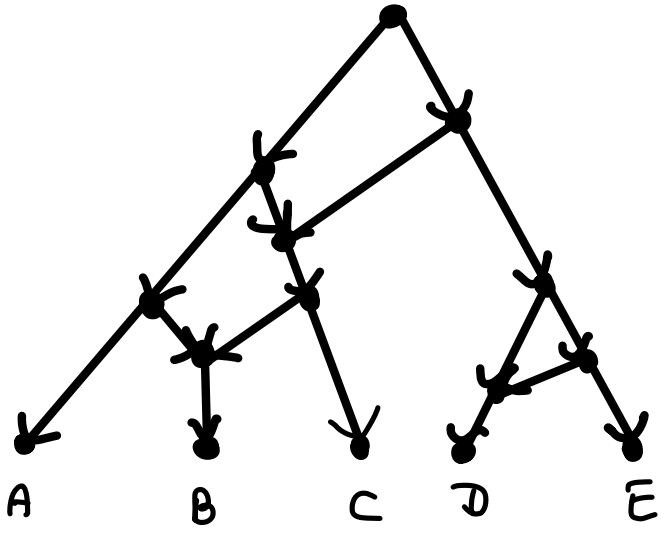
Future directions

Networks of higher level

- Our theoretical results for identifying the circular order and hybrid taxon on an m-sunlet from the (modified) NANUQ distance are for level-1 networks.
- Future work: Extend the approach to networks of higher level (if possible).



level-1

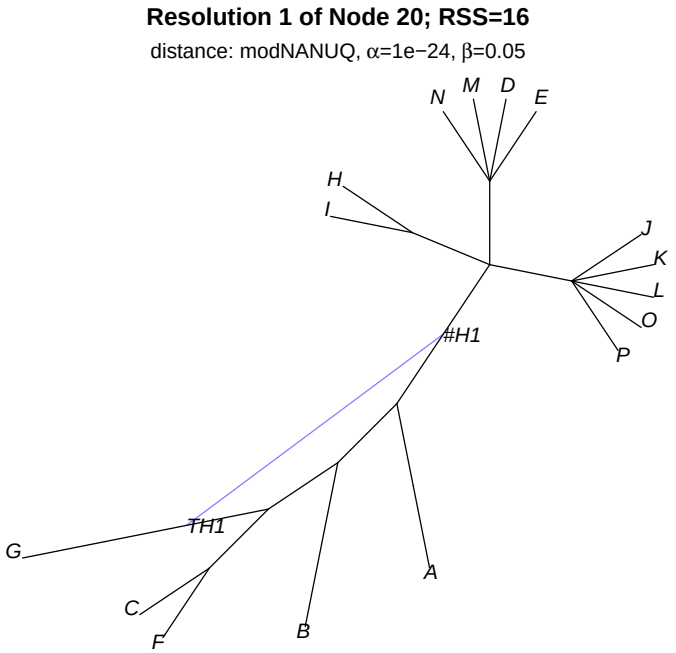


not level-1

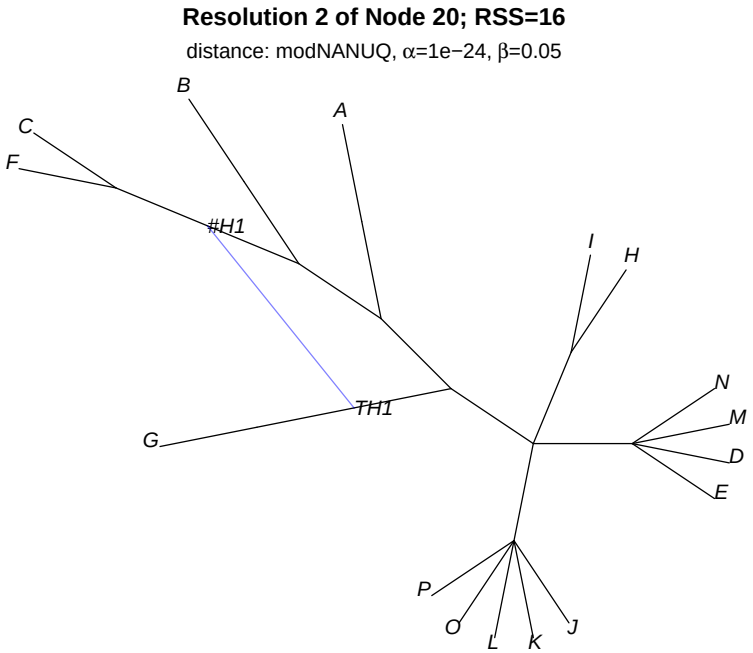
Future directions

Visualizing consensus among cycle resolutions

Sometimes different cycle resolutions are tied. How do we visualize their consensus?

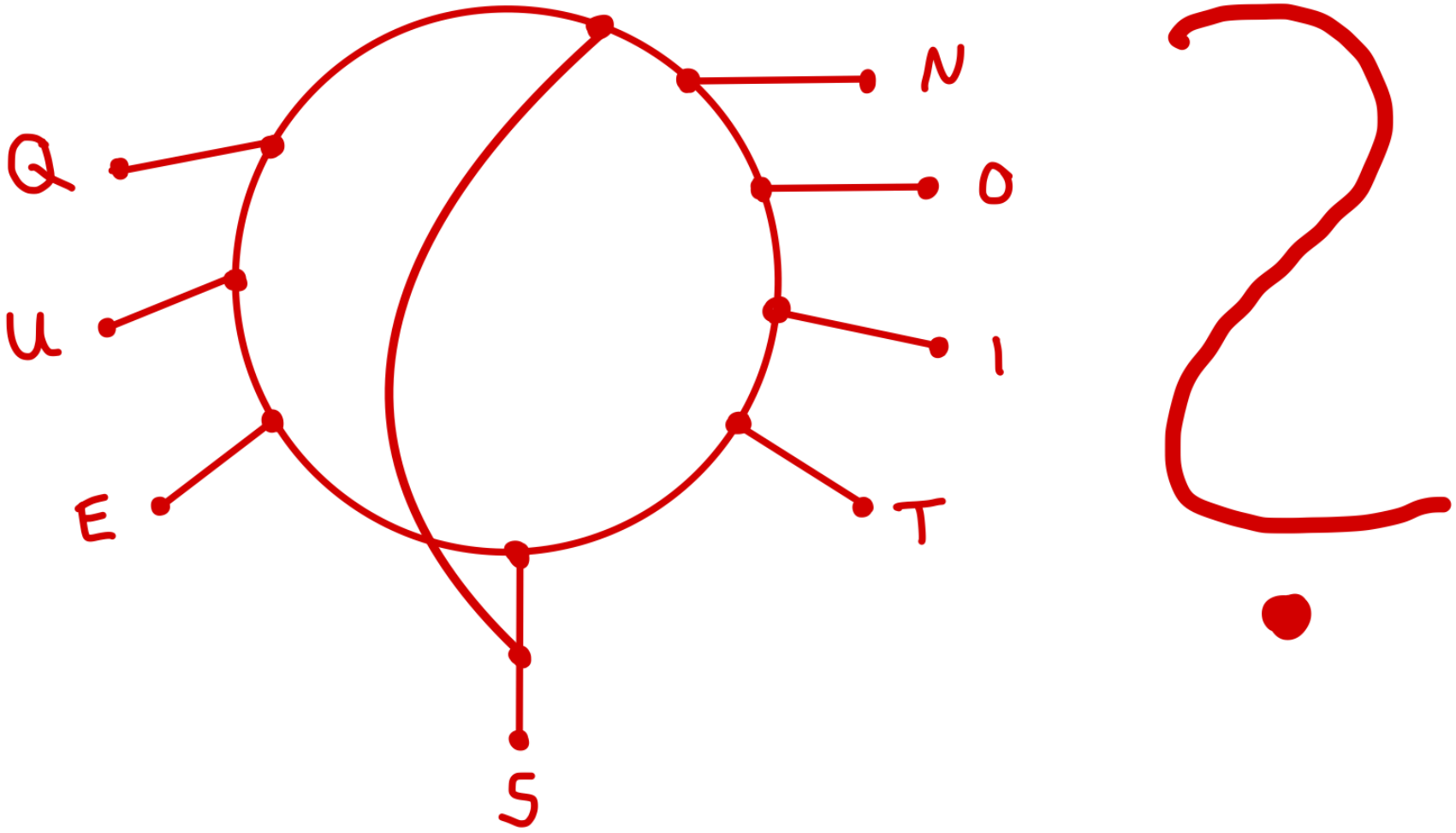


Network should be semidirected; Rooting is arbitrary.



Network should be semidirected; Rooting is arbitrary.

Thank you!



Bibliography I

- Allman, Elizabeth S., Hector Baños, and John A. Rhodes (2019). "NANUQ: a method for inferring species networks from gene trees under the coalescent model". In: *Algorithms for Molecular Biology* 14.1.
- Allman, Elizabeth S, Jonathan D Mitchell, and John A Rhodes (2021). "Gene Tree Discord, Simplex Plots, and Statistical Tests under the Coalescent". In: *Systematic Biology* 71.4, 929–942.
- Allman, Elizabeth S. et al. (2022). "The tree of blobs of a species network: identifiability under the coalescent". In: *Journal of Mathematical Biology* 86.1.
- Allman, E.S. et al. (2024). "TINNiK: Inference of the Tree of Blobs of a Species Network Under the Coalescent". In: *bioarxiv*.
- Baños, Hector (2018). "Identifying Species Network Features from Gene Tree Quartets Under the Coalescent Model". In: *Bulletin of Mathematical Biology* 81.2, 494–534.

Bibliography II

- Fogg, John, Elizabeth S Allman, and Cécile Ané (2023). "PhyloCoalSimulations: A Simulator for Network Multispecies Coalescent Models, Including a New Extension for the Inheritance of Gene Flow". In: *Systematic Biology* 72.5, 1171–1179.
- Lescroart, Jonas et al. (2023). "Extensive Phylogenomic Discordance and the Complex Evolutionary History of the Neotropical Cat Genus Leopardus". In: *Molecular Biology and Evolution* 40.12.
- Rhodes, John A. (2020). "Topological Metrizations of Trees, and New Quartet Methods of Tree Inference". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.6, 2107–2118.
- Solís-Lemus, Claudia and Cécile Ané (2016). "Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting". In: *PLOS Genetics* 12.3, e1005896.