



ICERM
November 2018



Elizabeth S. Allman

Reconstructing Hybridization Networks

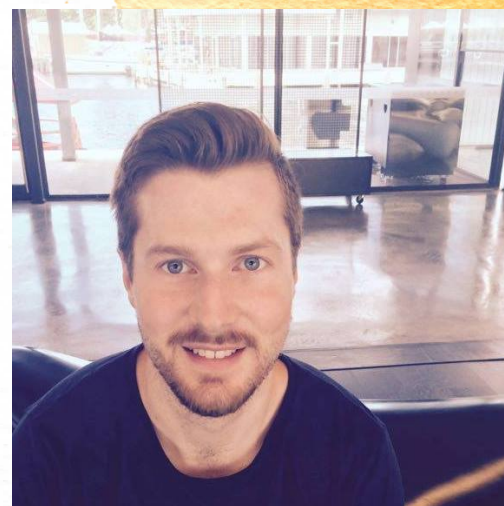
John Rhodes



Hector Baños-Cervantes



Jonathan Mitchell



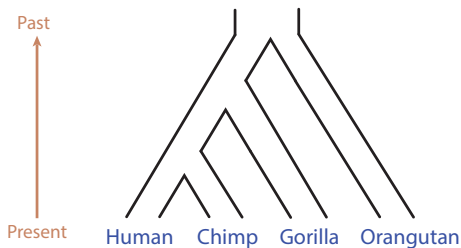
Joint work with these Nanooks

and nourished by NIH.



Inferring species trees

Phylogenetics is the branch of biology concerned with inferring evolutionary relationships between *taxa* (species).

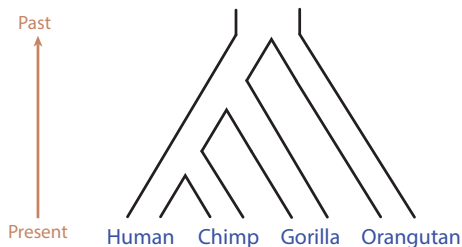


The data

For inference of a species tree, data input might be

- aligned molecular sequences
- collections of gene trees
- ...

A species tree inference method might lead to

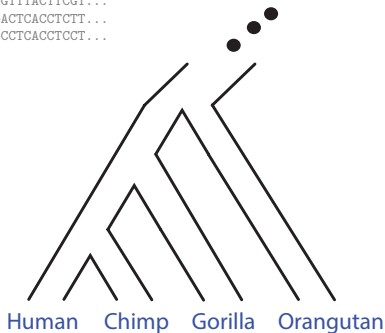


Eg. Primate mitochondrial DNA sequences, HindIII

Hayasaka, K., T. Gojobori, and S. Horai. MBE (1988) 5:626-644.

Gorilla	AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCACGGACTTACATCAT...
Orangutan	AAGCTTCACCGGCGCAACCAACCCTCATGATTGCCCATGGACTCACATCCT...
Human	AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...
Chimpanzee	AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gibbon	AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCCT...
Crab-eat_Mac	AAGCTTCTCCGGGCGCAACCAACCCTTATAATCGCCCACGGGCTCACCTCCT...
Lemur	AAGCTTCATAGGAGCAACCATTTCTAATAATCGCACATGGCCTTACATCAT...
Barbary_Mac	AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCCT...
Japanese_Mac	AAGCTTTTCCGGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCCT...
Squirrel_Mon	AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGT...
Rhesus_Mac	AAGCTTTTCTGGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCCT...
Tarsier	AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCT...

Inference scheme



Eg. Rokas et al. gene phylogenies

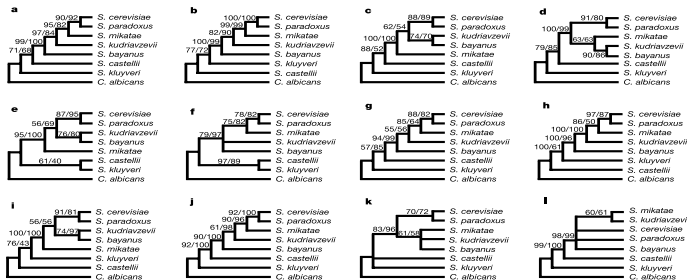
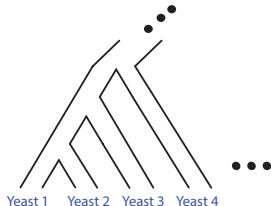


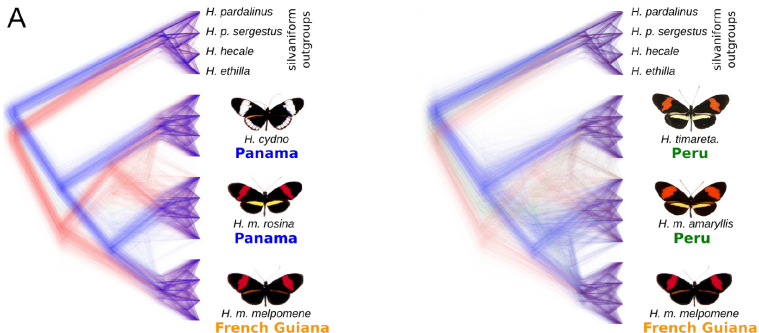
Figure 1 Single-gene data sets generate multiple, robustly supported alternative topologies. Representative alternative trees recovered from analyses of nucleotide data of 106 selected single genes and six commonly used genes are shown. The trees are the 50% majority-rule consensus trees from the genes YBL091C (a), YDL031W (b),

YER005W (c), YGL001C (d), YNL155W (e) and YOL097C (f), as well as those from the commonly used genes actin (g), hsp70 (h), β -tubulin (i), RNA polymerase II (j) elongation factor 1- α (k) and 18S rDNA (l). Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides).

Inference scheme
→

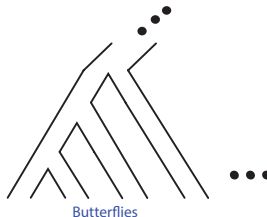


Eg. Martin et al. *Heliconius* butterflies



Genome Research 23:1817-1828 (2013).

Inference scheme



Gene trees vs. species trees

As these datasets make clear

**Gene trees and species trees are NOT the same,
and often disagree.**

- ▶ **Gene tree:** tree that represents the evolutionary history for a particular gene
 - ▶ Estimated using genetic data (e.g. DNA sequence alignments)
- ▶ **Species tree:** tree that represents the sequence of speciation events that gave rise to the observed collection of species
 - ▶ Genes and gene tree data are only indirectly informative about the species tree.

Sources of conflict

There are many reasons gene trees may differ from species trees

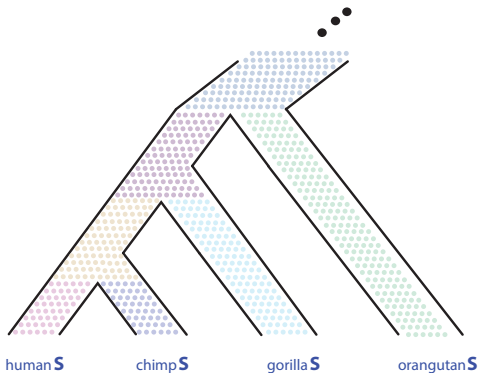
- ▶ lateral gene transfer
(e.g. viral insertion of genetic material into host genome)
- ▶ hybridization — **species network**
(interbreeding between distinct species to produce hybrid population that shares genetic contributions from both parental organisms)
- ▶ effects from population genetics — **incomplete lineage sorting**

Gene tree discord gives information about the species tree or network.

Goal: use sample of gene trees to infer species tree or network

Multispecies Coalescent Model

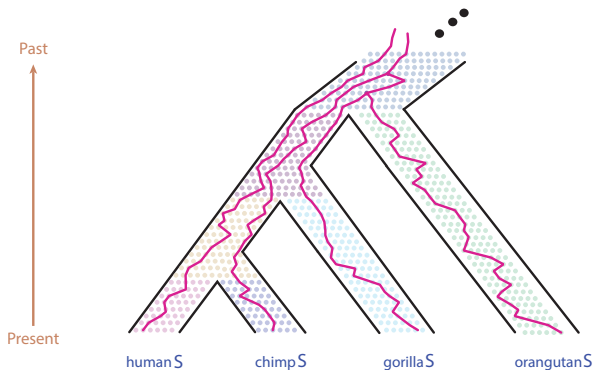
Models Incomplete Lineage Sorting (ILS) in **populations** of individuals, and extant and ancient populations are genetically diverse.



Multispecies Coalescent Model

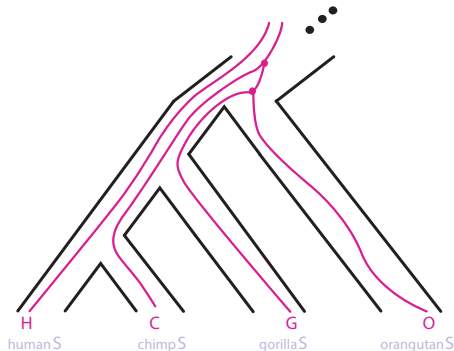
MSC models formation of gene trees in species trees as a *stochastic process*.

Under the MSC, choose one **lineage** per species and trace backwards in time the genetic history of these lineages. (discretization)



Multispecies Coalescent Model

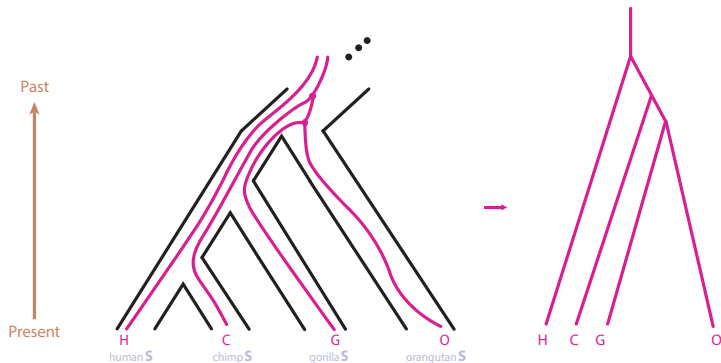
The MSC describes the formation of **gene trees** within species trees. (cont.)



Gene tree $(H, (C, (G, O)))$ beginning to form in species tree.

Multispecies Coalescent Model

The MSC describes the formation of **gene trees** within species trees.

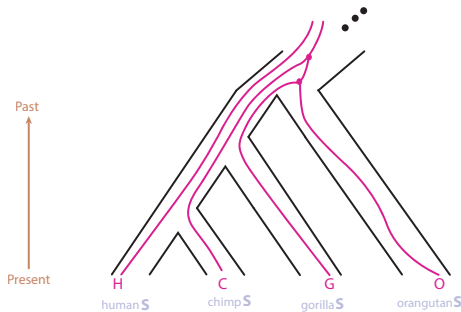


Gene tree $(H, (C, (G, O)))$ forming in species tree.

Multispecies Coalescent Model

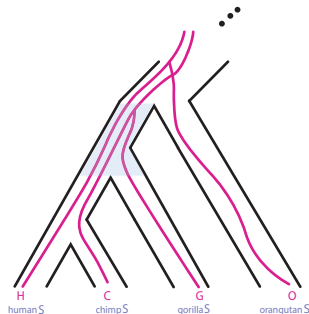
The **Multispecies Coalescent** models ILS.

- ▶ Viewing time backward (present \rightarrow past), lineages within a population coalesce, one pair at a time.
- ▶ The species tree constrains which lineages may coalesce at any given time.



Multispecies Coalescent model

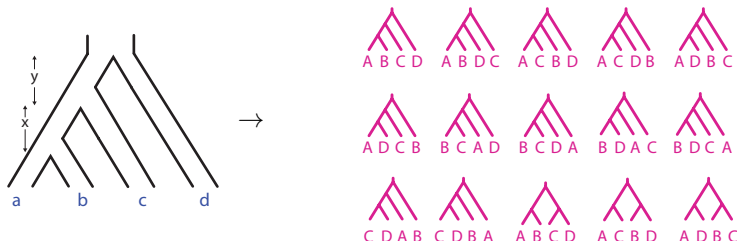
- ▶ Coalescent events occur in ‘populations,’ or internal branches of species tree. Internal branch lengths and population sizes are parameters of the MSC.



- ▶ If k lineages enter a population from below, then
 - ▶ coalescent events follow a Poisson process with rate $\binom{k}{2}$; waiting time T_k for the first event is exponential, $T_k \sim \exp(\binom{k}{2})$.
 - ▶ When an event occurs, every pair of lineages is equally likely to coalesce, with probability $\binom{k}{2}^{-1}$

Multispecies Coalescent model

The **MSC** model gives probability of rooted, metric gene trees.

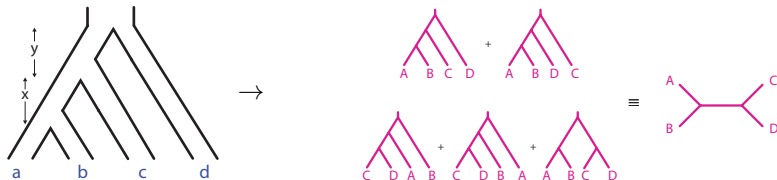


Complicated model due to multiple coalescent events in populations.

By integrating over gene tree branch lengths, the MSC can give the distribution of *topological gene trees*, both rooted and unrooted.

Multispecies Coalescent model

The **MSC** model gives probability of rooted, metric gene trees.



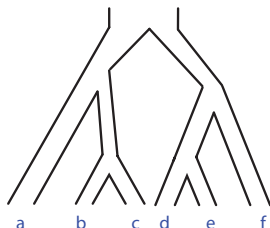
Complicated model due to multiple coalescent events in populations.

By integrating over gene tree branch lengths, the MSC can give the distribution of *topological gene trees*, both rooted and unrooted.

Multispecies Coalescent model

Summary:

The **MSC** model gives probability of rooted, metric gene trees.



By integrating over gene tree branch lengths and/or summing appropriately, the MSC can give the distribution of



- ▶ rooted topological gene trees
- ▶ unrooted topological gene trees
- ▶ ...
- ▶ quartet trees

Parameters:

Species tree topology

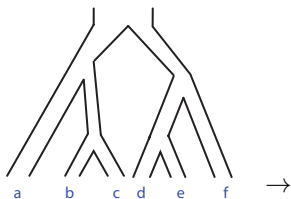
Internal branch lengths

Population sizes

Gene tree quartets

From the MSC, one can compute **concordance factors** of gene tree **quartets**:

The probabilities that a gene tree display any of three unrooted quartet trees.



1. Under the MSC, compute the probability of each (un)rooted topological gene tree on the set of taxa $X = \{a, b, c, d, e, f\}$.

Any gene tree G has positive probability $p_G = P(G \mid \sigma)$ under the MSC.

The distribution $\{p_G\} = \{P(G \mid \sigma)\}$ is polynomial in exponentials of the branch lengths.

After a change of variable, $X = \exp(-\ell)$ this gives rise to a parameterized variety.

Parameters σ :

Species tree topology

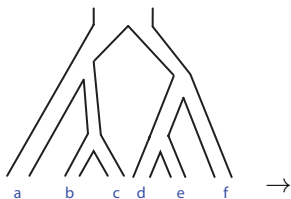
Internal branch lengths

Population sizes

Gene tree quartets

From the MSC, one can compute **concordance factors** of gene tree **quartets**:

The probabilities that a gene tree display any of three unrooted quartet trees.



1. Under the MSC, compute the probability of each (un)rooted topological gene tree on the set of taxa $X = \{a, b, c, d, e, f\}$.

Any gene tree G has positive probability $p_G = P(G \mid \sigma)$ under the MSC.

The distribution $\{p_G\} = \{P(G \mid \sigma)\}$ is polynomial in exponentials of the branch lengths.

After a change of variable, $X = \exp(-\ell)$ this gives rise to a parameterized variety.

Parameters σ :

Species tree topology

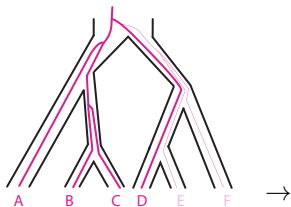
Internal branch lengths

Population sizes

Gene tree quartets

From the MSC, one can compute **concordance factors** of gene tree **quartets**:

The probabilities that a gene tree display any of three unrooted quartet trees.



2. For a given gene tree G and a 4-taxon subset of X , G displays one quartet tree.

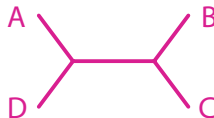
For example, the gene tree matching the species tree displays the **quartet** $AD \mid BC$.

Parameters σ :

Species tree topology

Internal branch lengths

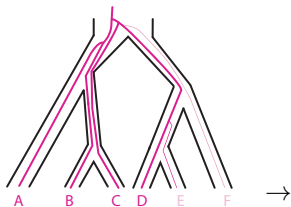
Population sizes



Gene tree quartets

From the MSC, one can compute **concordance factors** of gene tree **quartets**:

The probabilities that a gene tree display any of three unrooted quartet trees.



2. For a given gene tree G and a 4-taxon subset of X , G displays one quartet tree.

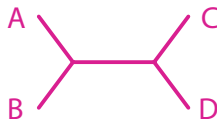
A different gene tree arising on σ displays the **quartet $AB \mid CD$** .

Parameters σ :

Species tree topology

Internal branch lengths

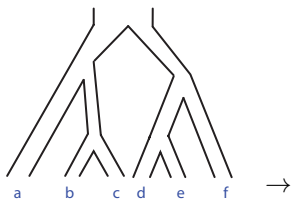
Population sizes



Gene tree quartets

From the MSC, one can compute **concordance factors** of gene tree **quartets**:

The probabilities that a gene tree display any of three unrooted quartet trees.



Parameters σ :

Species tree topology

Internal branch lengths

Population sizes

3. Given σ and any 4-taxon subset of X , under the MSC one can compute the probabilities of each of the 4 resolved quartet topologies being displayed by a gene tree.



$$P(AB \mid CD)$$



$$P(AC \mid BD)$$



$$P(AD \mid BC)$$

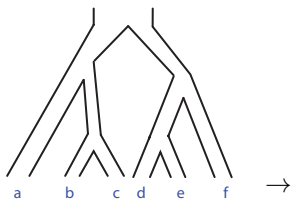
These are **concordance factors**.

Thm: CFs do not depend on the root location.

Gene tree quartets

From the MSC, one can compute **concordance factors** of gene tree **quartets**:

The probabilities that a gene tree display any of three unrooted quartet trees.



- Given σ and any 4-taxon subset of X , under the MSC one can compute the probabilities of each of the 4 resolved quartet topologies being displayed by a gene tree.



$$P(AB \mid CD)$$

$$P(AC \mid BD)$$

$$P(AD \mid BC)$$

Parameters σ :

Species tree topology

Internal branch lengths

Population sizes

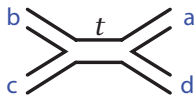
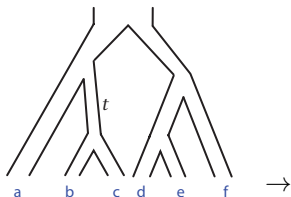
These are **concordance factors**.

Thm: CFs do not depend on the root location.

Gene tree quartets

From the MSC, one can compute **concordance factors** of gene tree **quartets**:

4. Moreover, with t denoting the branch length indicated, the concordance factors are given by the formulas:



Parameters σ :

Species tree topology

Internal branch lengths

Population sizes



$$P(AD \mid BC)$$

$$= 1 - \frac{2}{3}e^{-t}$$



$$P(AB \mid CD) = P(AC \mid BD)$$

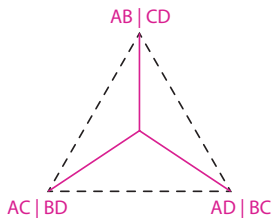
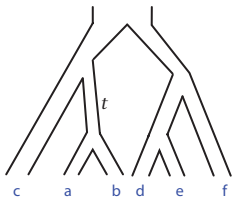
$$= \frac{1}{3}e^{-t}$$

Quartet concordance factors

Under the MSC, the quartet concordance factors are

$$CF_{abcd} = (P_{AB|CD}, P_{AC|BD}, P_{AD|BC}),$$

and can be displayed in the 2-simplex.

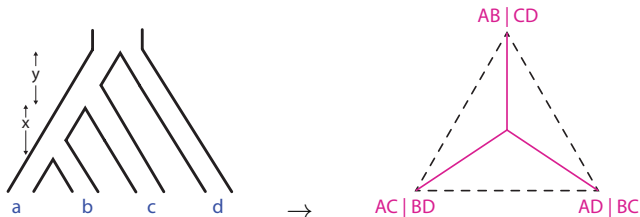


The largest concordance factor is for the quartet matching the unrooted species tree topology. The other two concordance factors are equal.

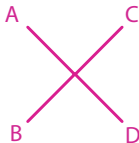
$$P_{AB|CD} = 1 - \frac{2}{3}e^{-t} \geq \frac{1}{3}, \quad P_{AC|BD} = P_{AD|BC} = \frac{1}{3}e^{-t} \leq \frac{1}{3}.$$

Quartet concordance factors

For a quartet species tree, the lines in the simplex denote the model space.

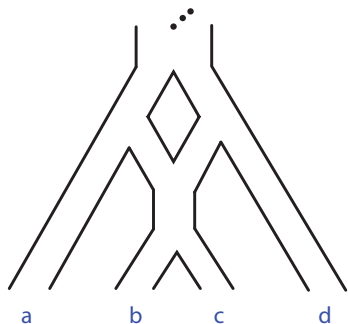


The model has a singularity at $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ which corresponds to the star tree.



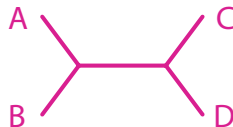
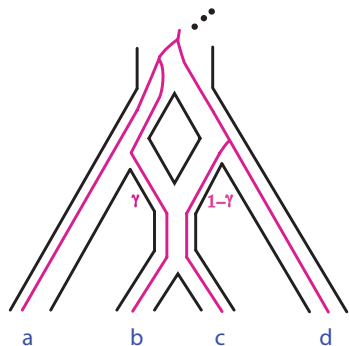
Network Multispecies Coalescent model

These ideas can be extended to **species networks**.



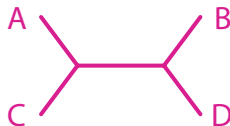
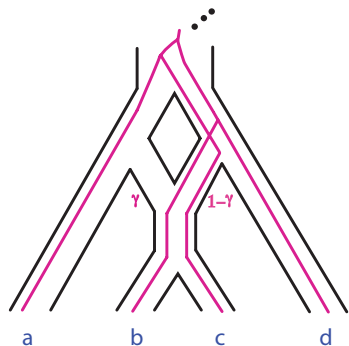
Network Multispecies Coalescent model

These ideas can be extended to **species networks**.



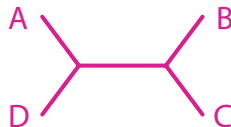
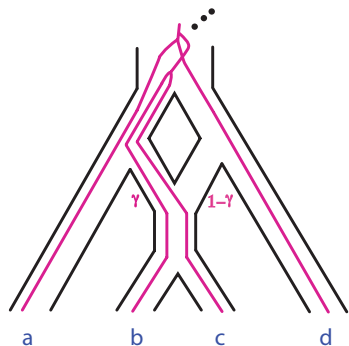
Network Multispecies Coalescent model

These ideas can be extended to **species networks**.



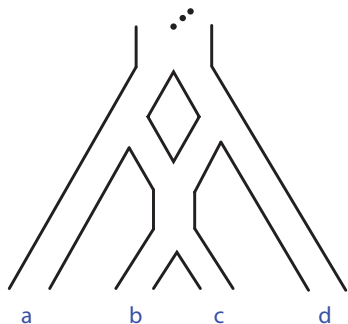
Network Multispecies Coalescent model

These ideas can be extended to **species networks**.



Network Multispecies Coalescent model

These ideas can be extended to **species networks**.



And triplets of concordance factors can be computed under the Network multi-species coalescent (NMSC).

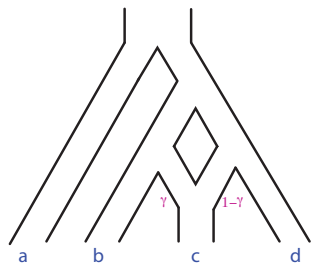
$$CF_{abcd} = (P_{AB|CD}, P_{AC|BD}, P_{AD|BC}).$$

N.B. CF_{abcd} is an ordered triplet.

Theorem: (Solís-Lemus/Ané 2016, Baños)
CFs are independent of root location on networks.

Concordance factors under the NMSC

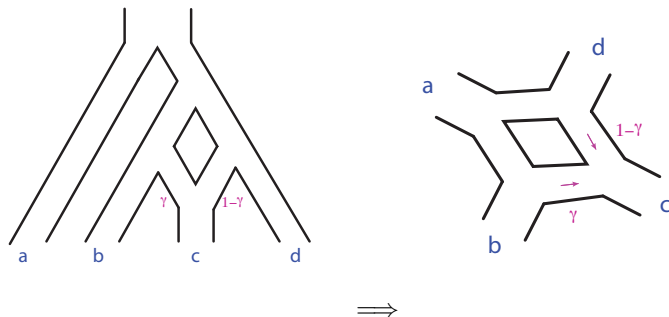
Given: rooted, metric, level-1 species network on X with cycles of size $k \geq 4$.



Unroot it, keep direction only on hybrid edges to obtain \mathcal{N}^- .

Concordance factors under the NMSC

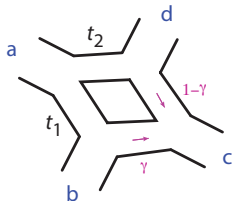
Given: rooted, metric, level-1 species network on X with cycles of size $k \geq 4$.



Unroot it, keep direction only on hybrid edges to obtain \mathcal{N}^- .

Concordance factors under the NMSC

Given: \mathcal{N}^- (or any rooted version of it)



Then for \mathcal{N}^- and with $T_i = e^{-t_i}$,
the concordance factors CF_{abcd} are
(in order):

$$P_{AB|CD} = (1 - \gamma) \left(1 - \frac{2}{3} T_2 \right) + \gamma \left(\frac{1}{3} T_1 \right)$$

$$P_{AC|BD} = (1 - \gamma) \left(\frac{1}{3} T_2 \right) + \gamma \left(\frac{1}{3} T_1 \right)$$

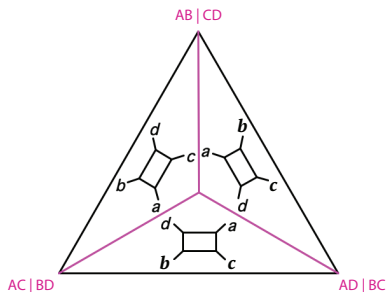
$$P_{AD|BC} = (1 - \gamma) \left(\frac{1}{3} T_2 \right) + \gamma \left(1 - \frac{2}{3} T_1 \right)$$

and similar polynomial formulas describe CFs for other networks.

Concordance factors under the NMSC

Moreover,

CFs derived under the NMSC model further partition the simplex.



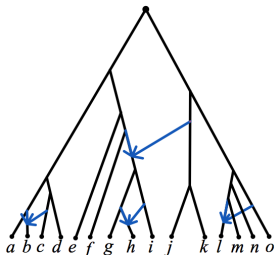
Proposition: (Solís-Lemus/Ané 2016, Baños 2018) Concordance factors CF_{abcd} under the NMSC generically identify topological 4-cycles in 4-taxon \mathcal{N}^- .

Concordance factors under the NMSC

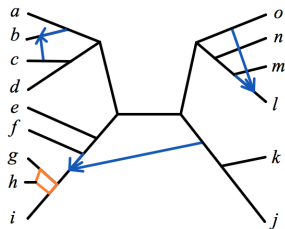
Theorem: (Baños) Let \mathcal{N} be a rooted metric level-1 network on X , $|X| \geq 4$, with cycles of size $k \geq 4$ in \mathcal{N}^- . Let $CF_{\mathcal{N}}$ be the collection of concordance factors for each 4-element subset of X ,

$$CF_{\mathcal{N}} = \{CF_{abcd} \mid a, b, c, d \text{ distinct elements of } X\}.$$

Then the unrooted topological network \mathcal{N}^- is identifiable from $CF_{\mathcal{N}}$. Moreover, for any k -cycle in \mathcal{N}^- with $k > 4$, the hybrid edges are also identifiable.



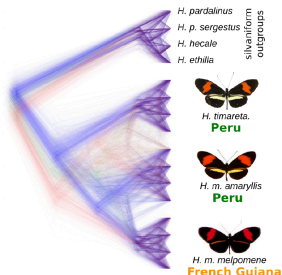
Original species network parameter.



Generically identifiable topological species network \mathcal{N}^- .

Concordance factors estimated from data

Given a large multilocus data set on taxa X , concordance factors can be (relatively) quickly estimated. $\mathcal{O}(N^4)$ quartets on each of m loci.

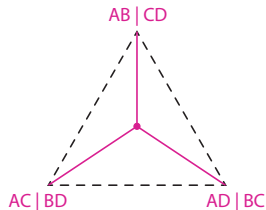


For the Peruvian butterfly dataset (2476 loci, 16 individuals), there are 1820 quartets.

Q	\widehat{CF}_Q	guess
amar.48, eth.67, tim.57, tim.86	(2244, 118, 114)	tree-like ?
amar.160, amar.216, amar.293, eth.67	(801, 889, 786)	star-like ?
amar.160, hec.273, melp.9317, tim.313	(165, 1741, 570)	????

Hypothesis tests

Testing whether gene tree quartets might arise from a **species TREE** or a **species STAR-TREE** can be formalized with statistical hypothesis tests, using the *geometry* of the MSC model. Under the MSC, quartet gene trees are binary and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is a singularity of the model.



Species *TREE* test

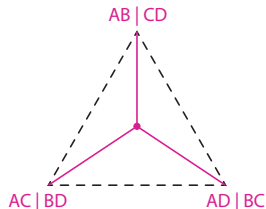
H_0 : A species tree generated gene tree quartets under the MSC.

H_1 : No species tree and/or MSC alone is insufficient.

J. Mitchell developed such tests in R and they will be available as part of the **MSCquartet** R package. (depend on sample size, proximity to singularity, etc.)

Hypothesis tests

Testing whether gene tree quartets might arise from a **species TREE** or a **species STAR-TREE** can be formalized with statistical hypothesis tests, using the *geometry* of the MSC model. Under the MSC, quartet gene trees are binary and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is a singularity of the model.



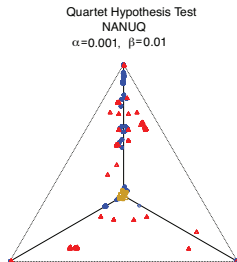
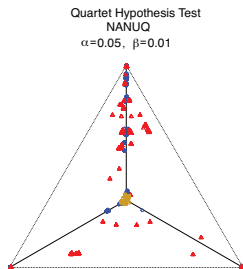
Species *STAR-TREE* test

H_0 : A STAR species tree generated gene tree quartets under the MSC.

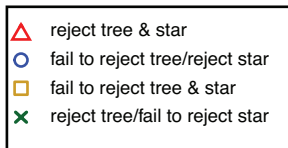
H_1 : *WIP*: Binary species tree gen. gt quartets under MSC

J. Mitchell developed such tests in R and they will be available as part of the **MSCquartet** R package. (depend on sample size, proximity to singularity, etc.)

Hypothesis tests on Peruvian butterflies



KEY:

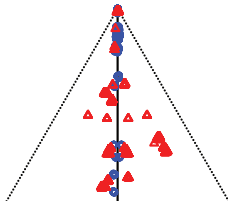


α = significance level of 'species-tree' test.

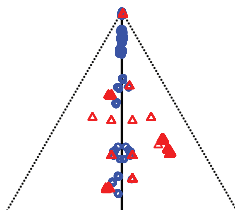
β = significance level of 'species-star-tree' test.

Hypothesis tests on Peruvian butterflies

$\alpha=0.05, \beta=0.01$



$\alpha=0.001, \beta=0.01$



KEY: With smaller significance level α , more quartets are accepted as tree-like.

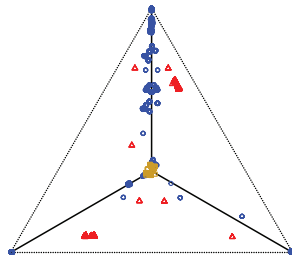
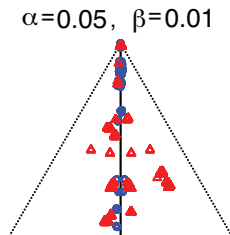
- | | |
|--|---------------------------------|
| | reject tree & star |
| | fail to reject tree/reject star |
| | fail to reject tree & star |
| | reject tree/fail to reject star |

α = significance level of 'species-tree' test.





β = significance level of 'species-star-tree' test.

Hypothesis tests on Peruvian butterflies

Quartet Hypothesis Test, NANUQ, $\alpha=1e-14$, $\beta=0.01$



KEY: With smaller significance level α , more quartets are accepted as tree-like.

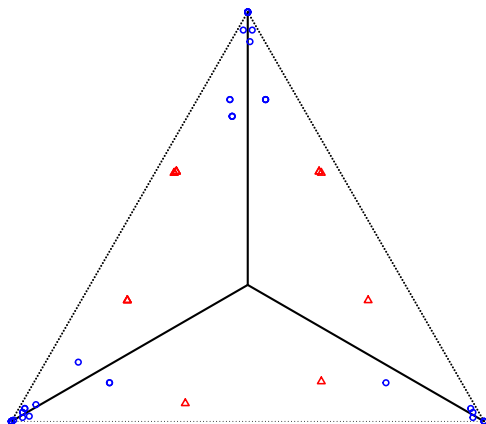
-  reject tree & star
-  fail to reject tree/reject star
-  fail to reject tree & star
-  reject tree/fail to reject star

α = significance level of 'species-tree' test.

β = significance level of 'species-star-tree' test.

Hypothesis tests on Yeast data

Quartet Hypothesis Test, NANUQ, $\alpha=0.05$, $\beta=0.01$



Shows evidence for non-species-tree-like evolution under the MSC.

Reconstructing hybridization networks

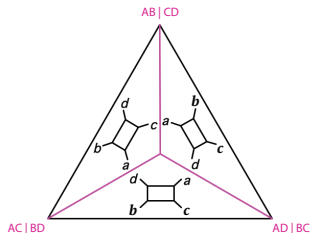
The NANUQ algorithm for inference of topological species tree networks.



Nanuq = polar bear (Inupiaq Eskimo)

NANUQ =

Network inference **A**lgorithm via **N**eighbour-net **U**sing **Q**uartet distance



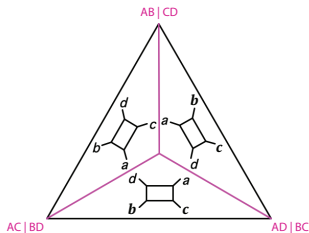
Input:

A collection of unrooted topological gene trees on a taxon set X .

A hypothesis testing level $0 < \alpha < 1$.

Steps:

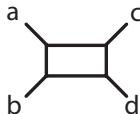
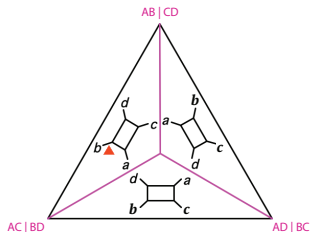
1. For each subset Q of 4 taxa, determine the empirical quartet frequencies \widehat{CF}_Q across the gene trees for each of the 3 resolved topologies.



2. Apply a statistical hypothesis test to each \widehat{CF} with level α with null hypothesis H_0 : the quartet is species-tree-like.

If the null hypothesis is rejected, use the values of the \widehat{CF} to determine a topological quartet 4-cycle network for the 4 taxa.

If it is accepted, use the value to determine a quartet tree topology.

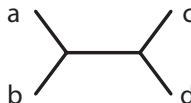
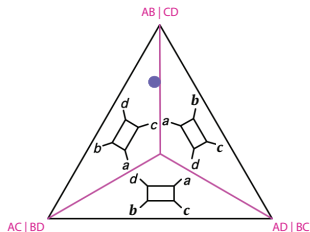


Choose this quartet 4-cycle species network topology.

2. Apply a statistical hypothesis test to each \widehat{CF} with level α with null hypothesis H_0 : the quartet is species-tree-like.

If the null hypothesis is rejected, use the values of the \widehat{CF} to determine a topological quartet 4-cycle network for the 4 taxa.

If it is accepted, use the value to determine a quartet tree topology.



Accept this quartet species tree topology for the data point.

2. Apply a statistical hypothesis test to each \widehat{CF} with level α with null hypothesis H_0 : the quartet is species-tree-like.

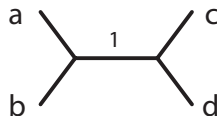
If the null hypothesis is rejected, use the values of the \widehat{CF} to determine a topological quartet 4-cycle network for the 4 taxa.

If it is accepted, use the value to determine a quartet tree topology.

3. Use the quartet networks/trees from the previous step to construct a network quartet distance between taxa. Based on (Rhodes 2018).

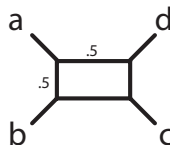
Idea: To get the pairwise distance $d(a, b)$ between taxa a and b , for every quartet separating these species, sum the weights using the rules:

Case 1: tree relates a, b, x, y



Tree contributes 1 to $d(a, d)$, $d(a, c)$, and 0 to $d(a, b)$.

Case 2: 4-cycle network relates a, b, x, y



4-cycle network contributes $\frac{1}{2}$ to $d(a, b)$, $d(a, d)$, and 1 to $d(a, c)$.

4. Use the NeighborNet Algorithm (Bryant et al. 2007) to determine a weighted circular split system approximating the quartet distance.
5. Use the Circular Network Algorithm of (Dress et al. 2004) to determine a splits graph for the circular system.

If this is confusing to you, the upshot is that steps 4 and 5 take pairwise distances estimated from data and construct a network known as a *splits network*. Examples to follow.

4. Use the NeighborNet Algorithm (Bryant et al. 2007) to determine a weighted circular split system approximating the quartet distance.
5. Use the Circular Network Algorithm of (Dress et al. 2004) to determine a splits graph for the circular system.

If this is confusing to you, the upshot is that steps 4 and 5 take pairwise distances estimated from data and construct a network known as a *splits network*. Examples to follow.

Output:

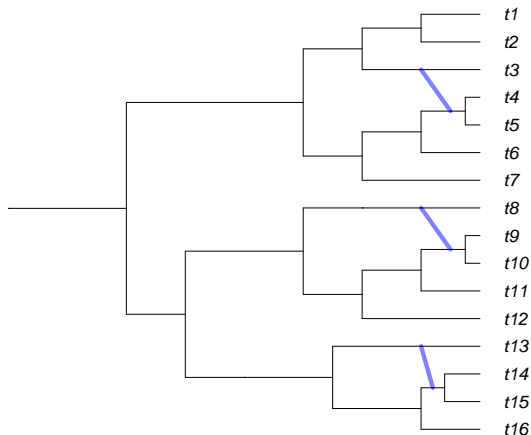
A splits graph on the induced topological network \mathcal{N}^-
(which needs some interpretation).

Theorem: (ABR)

Under the NMSC model, for generic numerical parameters, the NANUQ species tree estimator is statistically consistent for inferring an unrooted topological network \mathcal{N}^- associated to a rooted metric species network \mathcal{N}^+ .

NANUQ splits graph

How does NANUQ work on simulated data?



Model species network (16 taxa, 4-, 5-, and 6-cycle)

$\{t_{14}, t_{15}\}$, $\{t_9, t_{10}\}$, $\{t_4, t_5\}$

NANUQ splits graph

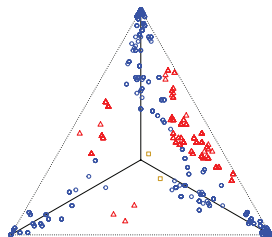
Input parameter: Model species network

1. Generate $n = 100, 1000$ gene trees under the NMSC. (hybrid-Lambda)
2. Use NANUQ to choose topology of 4-taxon subsets from $\{t_1, \dots, t_{16}\}$
3. Use NANUQ to compute the network quartet distance to compute pairwise distances between t_i, t_j .
4. Use the NeighborNet algorithm (Bryant et al.) to determine a weighted circular split system approximating the quartet distance.
5. Use the Circular Network Algorithm of (Dress et al.) to determine a splits graph for the circular system.
6. Plot resulting network in SplitsTree (Huson)

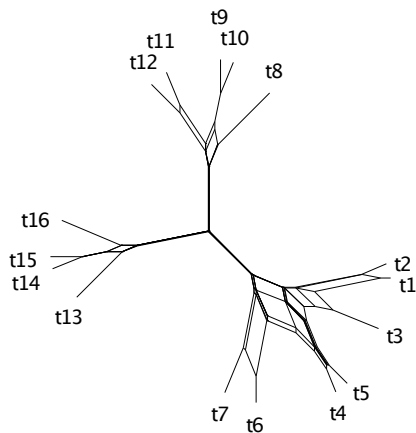
NANUQ splits graph

$n = 100$ gene trees, $\alpha = .001$, $\beta = .01$

Quartet Hypothesis Test, NANUQ, $\alpha=0.001$, $\beta=0.01$



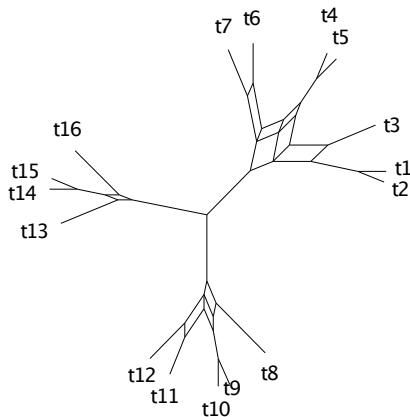
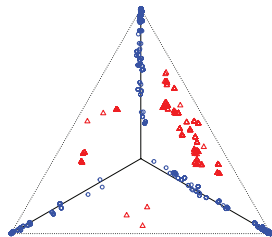
10.0



NANUQ splits graph

$n = 1000$ gene trees, $\alpha = .001$, $\beta = .01$

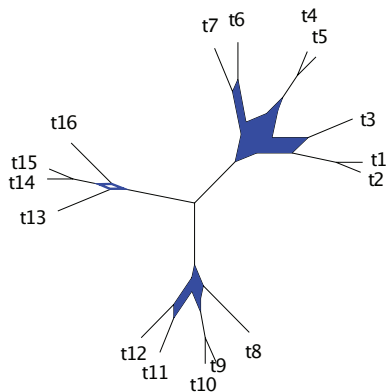
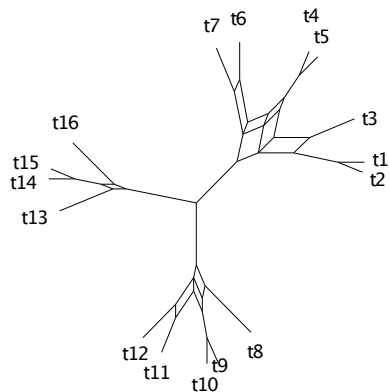
Quartet Hypothesis Test, NANUQ, $\alpha=0.001$, $\beta=0.01$



The silhouette of the cycles is showing.....

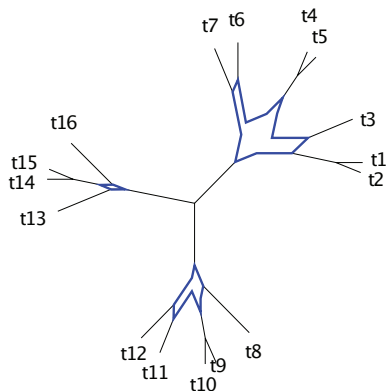
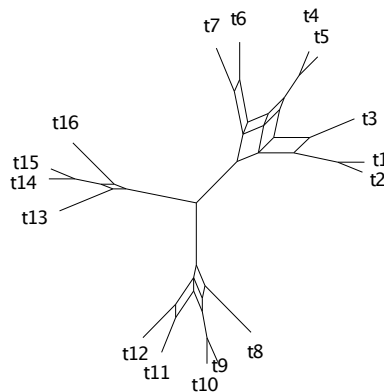
NANUQ splits graph

Splits graphs (L) were designed to show conflicting splits signal in data, not cycles (R). Replace blob with frontier cycle to recover parameter.



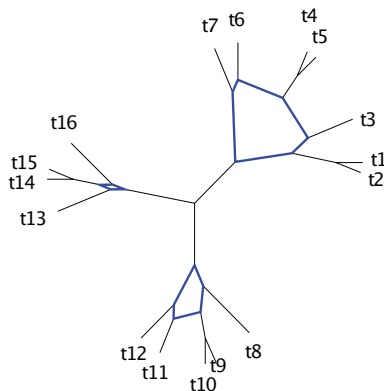
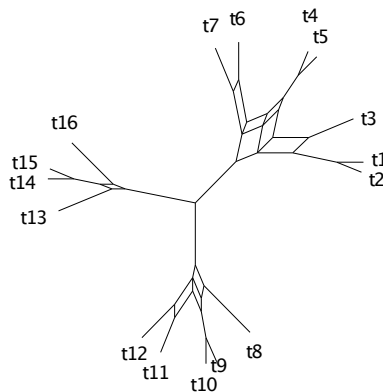
NANUQ splits graph

Splits graphs (L) were designed to show conflicting splits signal in data, not cycles (R). Replace blob with frontier cycle to recover parameter.



NANUQ splits graph

Splits graphs (L) were designed to show conflicting splits signal in data, not cycles (R). Replace blob with frontier cycle to recover parameter.



Theorem: (ABR)

As the number n of gene trees $\rightarrow \infty$, the network quartet distance exactly fits a circular split system, such that each blob in its splits graph corresponds to a cycle in the original network.

For cycles of size > 4 , this blob has features indicating the hybrid direction.

While splits graphs were introduced to show conflict in data, this theorem shows that they are consistent under the NMSC (in the sense of the theorem).

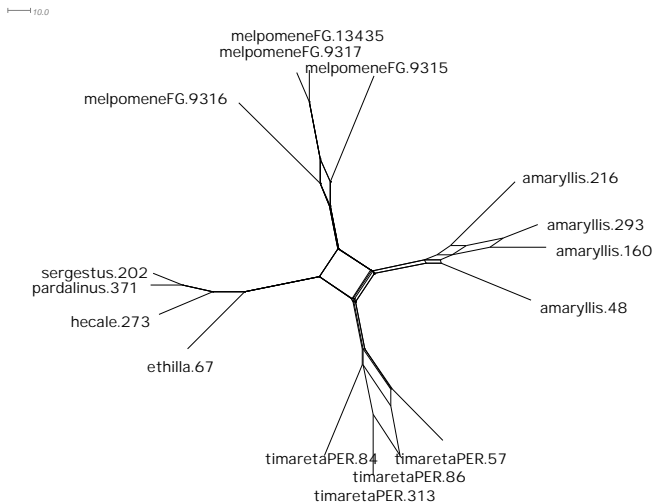
Theorem: (ABR)

As the number n of gene trees $\rightarrow \infty$, the network quartet distance exactly fits a circular split system, such that each blob in its splits graph corresponds to a cycle in the original network.

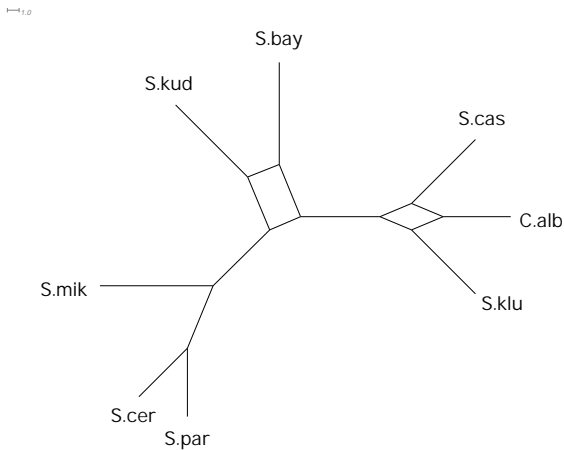
For cycles of size > 4 , this blob has features indicating the hybrid direction.

While splits graphs were introduced to show conflict in data, this theorem shows that they are consistent under the NMSC (in the sense of the theorem).

NANUQ splits graph for the butterfly data



and for the yeast data



Thank you!



H. Baños, J. Mitchell, E. Allman, and J. Rhodes. “MSCquartet” R package including NANUQ and Hypothesis Tests. in preparation.

J. Mitchell, E. Allman, and J. Rhodes. “Hypothesis testing near singularities and boundaries.” 2018. submitted.

H. Baños, E. Allman, and J. Rhodes. “A new method of inferring hybridization networks from gene trees.” in preparation.

H. Banos. “Identifying species network features from gene tree quartets.” *Bulletin of Mathematical Biology*, 2018. to appear.

J. A. Rhodes. “Topological metrizations of trees, and new quartet methods of tree inference.” 2018. submitted.

Bibliography

D. Bryant, V. Moulton, and A. Spillner. "Consistency of the neighbor-net algorithm." *Algorithms for Molecular Biology*, 2:8, 2007.

A. Dress and D. Huson. "Constructing splits graphs." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(3):109–115, July 2004.

D. Huson. "Splitstree—a program for analyzing and visualizing evolutionary data." *Bioinformatics*, 14:68–73, 1998.

D. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks*. Cambridge University Press, Cambridge, 2010.

C. Solís-Lemus and C. Ané. "Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting." *PLoS Genetics*, 12(3), 2016.