Honors Thesis: Fast branch length estimation under the multi-species coalescent model using triplets

Michael Suehle
Department of Computer Science
University of Maryland, College Park

(Advisor: Erin Molloy)

December 20, 2022

1 Introduction

Species trees model how different species evolved over time. They are used in biological research for studying the evolution of traits and how different species are related. Gene trees model how genes (regions of the genome) evolved over time. They can be estimated from genomic data and then used to estimate a species tree (topology and branch lengths) under the multi-species coalescent (MSC) model (Rannala and Yang, 2003). Popular methods for estimating species trees under this model involve counting triplets (3-leaf rooted subtrees) or quartets (4-leaf unrooted subtrees) in the gene trees. This thesis presents a new triplet-based method for estimating branch lengths, which is similar to the fast quartet-based method ASTRAL (Sayyari and Mirarab, 2016). The remainder of this thesis is structured as follows. In Section 2, I define background and terminology. In Section 3, I present Trips, a new triplet-based branch length estimation algorithm that runs in $O(n^2k)$ time, where n is the number of species and k is the number of gene trees. In Section 4, I describe how I evaluated Trips using simulated data sets, and in Section 5, I present the results. In Section 6, I conclude with a discussion of future work.

2 Background

2.1 Terminology

A phylogenetic tree T is a tree that models the evolutionary history of the leaves, typically labeled by a set S of species. A tree T is said to be on species set S if there is a bijection between the leaves of T and the labels in S. For simplicity, I refer to leaf vertices by their labels and use $L(T)$ to denote the set of leaves/labels. Rooted phylogenetic trees are directed acyclic graphs that have a designated root vertex with no parent. Leaf vertices have no children, and internal vertices have both a parent and children. Terminal branches connect leaves to their parents; all other branches are internal branches. A rooted tree is binary if every internal vertex has exactly two children and every vertex, aside from the root, has one parent (note: all trees are binary in this thesis). An ancestor of vertex v is any vertex u on the directed path from the root to v in T (conversely, v is a descendant of u). The lowest common ancestor (LCA) of a set of vertices is the ancestor of all vertices that is farthest from the root.

A rooted tree can be made into an unrooted tree by suppressing its root vertex and undirecting its edges. An induced subtree on label set $R \subseteq S$, denoted $T|_R$, is the resulting tree from removing all leaves with labels that are not in R from T (note: vertices with degree two must be suppressed, after removing the leaves). A triplet is a rooted phylogenetic tree with three leaves; there are three possible topologies (Fig. 1a). A quartet is an unrooted phylogenetic tree with four leaves; there are also three possible topologies (Fig. 1b).

2.2 Incomplete Lineage Sorting (ILS) and Multi-Species Coalescent (MSC) Model

The MSC model generates gene trees from a species tree. The leaves of species trees are populations and the internal nodes are events where populations split. The branches of the species tree are in coalescent units (CUs). Coalescent units are defined as: $\tau = \frac{t}{2N_e}$ where t is the number of generations on the branch and N_e is the effective population size on the branch. The leaves of gene trees are DNA sequences (labeled by the species whose genomes they came from), and the internal nodes are coalescent events, indicating the DNA sequences were trace back to a common ancestor. Incomplete lineage sorting (ILS) occurs when gene tree topologies differ from the species tree topology due to population modeled by the MSC. Species tree branches with shorter lengths cause more ILS, so branch lengths are important for biological research.

Probability of Triplets under MSC Model. Given a rooted species tree with three leaves and branch lengths in coalescent units, the probability of the three possible gene tree topologies $Q_1 = ((A,B),C)$, $Q_2 = ((B,C),A)$, and $Q_3 = ((A,C),B)$ can be computed. If the rooted species tree has the same topology as Q_1, then the probabilities are

$$P(Q_1) = 1 - \frac{2}{3}e^{-d} \tag{1}$$

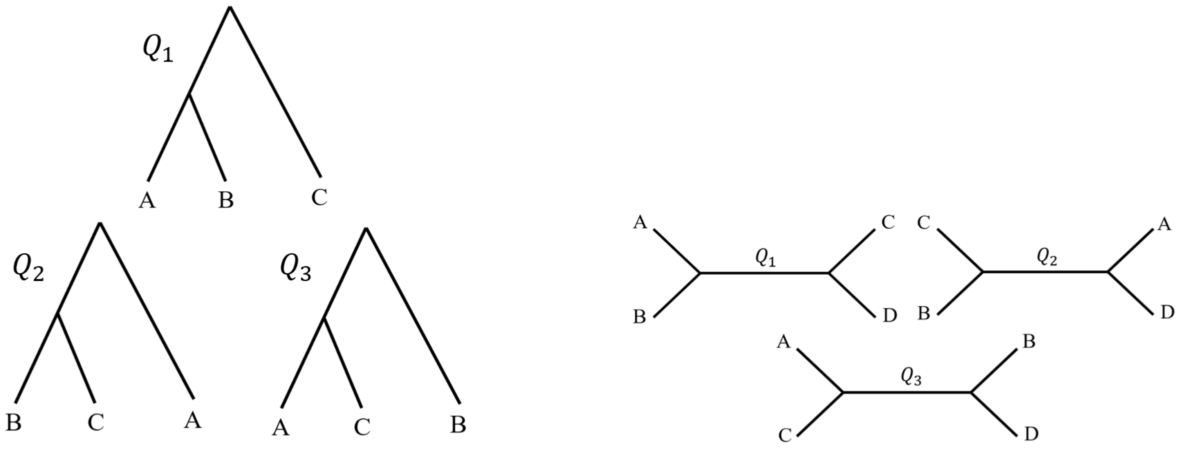
$$P(Q_2) = P(Q_3) = \frac{1}{3}e^{-d} \tag{2}$$

where d is the length of the internal branch in coalescent units (CUs) in the species tree. This result is well-known for triplets (Degnan and Rosenberg, 2006). Given a rooted species tree with three leaves and a collection of rooted gene trees, the length of the internal branch in the species tree can be estimated as

$$\hat{d} = -\ln \frac{3}{2}\left(1 - \frac{t}{k}\right) \tag{3}$$

where t is the number of triplets with the same topology as the species tree and k is the number of gene trees. The same relationships hold for quartets $Q_1 = A,B|C,D$, $Q_2 =$

$A, D|B, C$, and $Q_3 = A, C|B, D$, although in this case the species tree and gene trees are unrooted (Allman et al., 2011).



(a) Three Possible Triplet Topologies

(b) Three Possible Quartet Topologies

Figure 1

2.3 Branch length estimation in ASTRAL

I now describe the branch length estimation technique in ASTRAL (Sayyari and Mirarab, 2016), which takes an unrooted species tree and a set of unrooted gene trees as input. Their algorithm extends the idea above for estimating branch lengths for the case where there are more than four species in the tree. In equation 3, the probability of the quartet that agrees with the species tree is estimated by computing the frequency of that quartet in the input gene trees. If there are more than four species, then the probability could be estimated using different subsets of four species and taking the average.

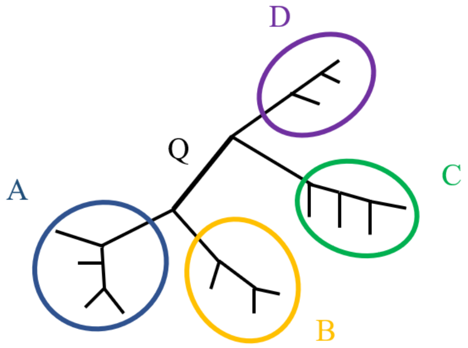


Figure 2: Example oriented quadripartition $A, B|C, D$ induced by branch Q

Definition 2.1. *A branch $Q = (u, v)$ in an unrooted phylogenetic tree T induces an* ***oriented quadripartition*** *$A, B|C, D$ as follows. If you delete edge (u, v) but not its endpoints,* *you get two subtrees: one rooted at vertex u and one rooted at vertex v. Then, let A be the set of leaves that are descendents of the left child or the right child of u, let B be the set of leaves that are descendents of the other child of u, let C be the set of leaves that are descendents of the left child or the right child of v, and let D be the set of leaves that are descendents of the other child of v (note: we could also swap u and v to get an induced oriented quadripartition). The subset of leaves A, B, C, and D are referred to as* ***clusters*** *(note: they are pairwise disjoint and their union is $L(T)$).*

Definition 2.2. *A quartet $w, x|y, z$ is* ***consistent*** *with branch Q (with* ***oriented quadripartition*** *$A, B|C, D$) if either*

1. *$w \in A$, $x \in B$, $y \in C$, and $z \in D$,*
2. *$w \in B$, $x \in A$, $y \in C$, and $z \in D$,*
3. *$w \in A$, $x \in B$, $y \in D$, and $z \in C$,*
4. *$w \in B$, $x \in A$, $y \in D$, and $z \in C$,*
5. *$w \in C$, $x \in D$, $y \in A$, and $z \in B$,*
6. *$w \in D$, $x \in C$, $y \in A$, and $z \in B$,*
7. *$w \in C$, $x \in D$, $y \in B$, and $z \in A$, or*
8. *$w \in D$, $x \in C$, $y \in B$, and $z \in A$.*

Now suppose that Q_1 is a branch (with orientated quadripartition $A, B|C, D$) in the input species tree, and let Q_2 and Q_3 denote the alternative branches $A, D|B, C$ and $A, C|B, D$, respectively. Let z_i denote the number of induced quartets in the input gene trees that are consistent with Q_i. ASTRAL implements a fast algorithm to compute z_1, z_2, z_3 and uses these quantities to estimate the length of Q_1 using the equation below.

$$\hat{d} = -\ln \frac{3}{2}\left(1 - \frac{z_1}{\sum_{i=1}^{3} z_i}\right) \tag{4}$$

2.4 Branch length estimation with MP-EST

I now describe the branch length estimation technique in MP-EST (Liu et al., 2010), which takes an rooted species tree and a set of rooted gene trees as input. MP-EST begins by counting the triplets induced by the input gene trees and initializing the branch lengths in the input species tree to 0.1 CUs for internal branches and and 1 CU for terminal branches. Then, it computes the pseudo-loglikelihood for species tree with current branch lengths given triplets. The pseudo-loglikelihood function is computed as:

$$\sum_{j=1}^{\binom{n}{3}} \sum_{k=1}^{3} x_{j,k} \log p_{j,k} \tag{5}$$

where $p_{j,k}$ is the probability of observing triplet on species set j with the topology k given the model species tree and $x_{j,k}$ is the frequency of triplet on species set j with topology k in the input gene trees. To estimate branch lengths on a fixed tree, the code seems to randomly perturb the branch lengths, keeping the tree (with branch lengths) that increases the likelihood until either max rounds or convergence (see `https://github.com/lliu1871/mp-est/blob/master/src/mpest.c#L1034`).

3 Trips: Extending ASTRAL's approach to Triplets

I now extend the algorithm of Sayyari and Mirarab, 2016 to rooted trees and triplets. To begin, I define what it means for triplets to be consistent with a branch.

Definition 3.1. *Each internal edge $Q = u \mapsto v$ in a rooted phylogenetic tree induces a* ***oriented tripartition*** *$A, B|C$, where A is the set of leaves that are descendants of the left or right child of v, B is the set of leaves that are descendants of the other child of v, and C is the set of leaves that are descendants of the sibling of v. The subsets of leaves A, B, and C are referred to as* ***clusters*** *(note: they are pairwise disjoint).*

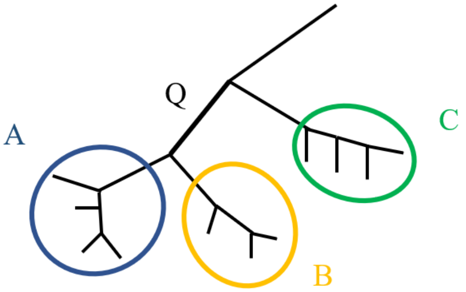


Figure 3: Example oriented tripartition $A, B|C$ induced by Q

Definition 3.2. *A triplet $((x, y), z)$ is* ***consistent*** *with branch Q (with oriented tripartition $A, B|C$) if either $x \in A$, $y \in B$, and $z \in C$ or $x \in B$, $y \in A$, and $z \in C$.*

Now I provide an efficient algorithm to count triplets in a gene tree that are consistent with a given branch in the species tree.

Algorithm 1 TripletCounting$(t, Q = A, B|C)$

```
**Input:** Phylogenetic tree $t$ and oriented tripartition $A, B|C$ where $A \cup B \cup C \subseteq L(t)$
**Output:** Number of triplets in $t$ that are consistent with $A, B|C$
$V \leftarrow$ structure indexed by tree vertices and cluster indices ($V[u][i]$ denotes the number of leaves that are descendants of vertex $u$ that are in cluster $i$)
$Trips \leftarrow$ structure indexed by tree vertices ($Trips[u]$ denotes the number of triplets in $t$ with LCA $u$ that are consistent with $A, B|C$)
$Total \leftarrow 0$
for $u$ in PostOrderTraversal of $t$ do
    if $u$ is a leaf then
        $V[u] \leftarrow (A[u], B[u], C[u])$
    else
        $Trips[u] \leftarrow V[u.left.left][0] * V[u.left.right][1] * V[u.right][2] +$
                  $V[u.left.left][1] * V[u.left.right][0] * V[u.right][2] +$
                  $V[u.left][2] * V[u.right.left][0] * V[u.right.right][1] +$
                  $V[u.left][2] * V[u.right.left][1] * V[u.right.right][0] +$
                  $V[u.left.left][0] * V[u.left.left][1] * V[u.right][2] +$
                  $V[u.left.right][0] * V[u.left.right][1] * V[u.right][2] +$
                  $V[u.left][2] * V[u.right.left][0] * V[u.right.left][1] +$
                  $V[u.left][2] * V[u.right.right][0] * V[u.right.right][1]$
        $Total \leftarrow Total + Trips[u]$
        $V[u] \leftarrow V[u.left] + V[u.right]$
    end if
end for
return $Total$
```

Note: $V[u][i]$ returns 0 if node u does not exist.
Note: $A[u]$ returns 1 if leaf u is in cluster A and 0 otherwise (same for clusters B and C).

Lemma 3.3. *The Triplet Counting Algorithm correctly computes quantity $Trips[u]$, which denotes the number of triplets in the input gene tree t with LCA u that are consistent with the input oriented tripartition $A, B|C$.*

Proof. By definitions 3.1 and 3.2, $Trips[u]$ is the number of ways to pick one leaf from A (call a), one leaf from B (call b), and one leaf from C (call c) so that the LCA of a, b, c is u and the LCA of a, b is a descendant of u. This breaks down into two cases: the number of ways to pick a, b so that they are descendants of the left child of u (denoted $u.left$) and c so that it is a descendant of the right child of u (denoted $u.right$)—and vice versa. The number of triplets in the first case can be computed as $V[u.right][C]*V[u.left][A]*V[u.left][B]$, and the number of triplets in the second case can computed as $V[u.left][C] * V[u.right][A] * V[u.right][B]$ (recall that $V[u][i]$ denotes the number of leaves that are descendants of vertex u that are in cluster i and that $V[u][i]$ is 0 if vertex u does not exist). In my algorithm, I break this down into 8 subcases (Fig. 4).

1. Leaf a is a descendant of $u.left.left$, leaf b is a descendent of $u.left.right$, and leaf c is a descendent of $u.right$. The number of triplets that satisfy this case is $V[u.left.left][A]* V[u.left.right][B] * V[u.right][C]$.

2. Leaf b is a descendant of $u.left.left$, a is a descendant of $u.left.right$, and c is a descendent of $u.right$. The number of triplets that satisfy this case is $V[u.left.left][B]* V[u.left.right][A] * V[u.right][C]$.

3. Leaf c is a descendant of $u.left$, a is a descendant of $u.right.left$, and b is a descendant of $u.right.right$. The number of triplets that satisfy this case is $V[u.left][C] * V[u.right.left][A] * V[u.right.right][B]$.

4. Leaf c is a descendant of $u.left$, b is a descendant of $u.right.left$, and a is a descendant of $u.right.right$. The number of triplets that satisfy this case is $V[u.left][C] * V[u.right.left][B] * V[u.right.right][A]$.

5. Leaves a and b are descendants of $u.left.left$, and c is a descendant of $u.right$. The number of triplets that satisfy this case is $V[u.left.left][A] * V[u.left.left][B] * V[u.right][C]$.

6. Leaves a and b are descendants of $u.left.right$, and c is a descendant of $u.right$. The number of triplets that satisfy this case is $V[u.left.right][A] * V[u.left.right][B] * V[u.right][C]$.

7. Leaves a and b are descendants of $u.right.left$, and c is a descendant of $u.left$. The number of triplets that satisfy this case is $V[u.right.left][A] * V[u.right.left][B] * V[u.right][C]$.

8. Leaves a and b are descendants of $u.right.right$, and c is a descendant of $u.left$. The number of triplets that satisfy this case is $V[u.right.right][A] * V[u.right.right][B] * V[u.right][C]$.

The values for these 8 cases are added together, giving us the total number of triplets in t with LCA u that are consistent with $A, B|C$. □

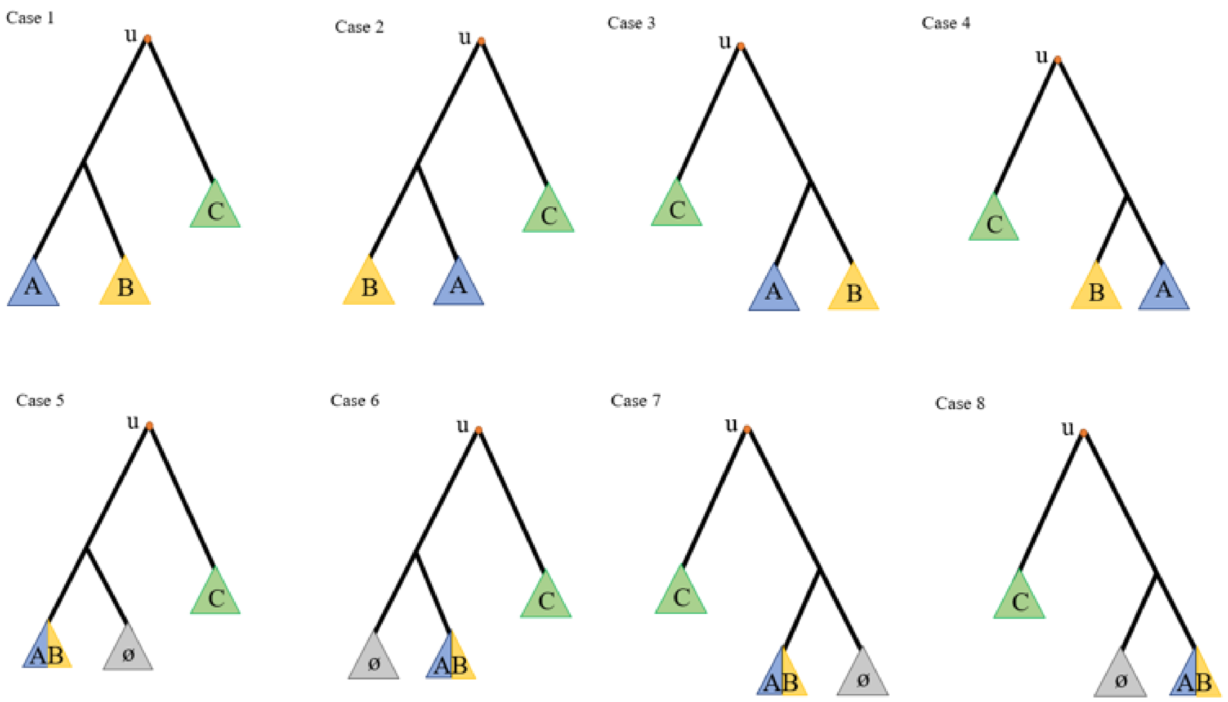


Figure 4: Cases for counting the number of induced triplets in a gene tree with LCA u that are consistent with the oriented tripartition $A, B|C$

Theorem 3.4. *The Triplet Counting Algorithm returns the number of triplets in the input gene tree t that are consistent with the input oriented tripartition $A, B|C$.*

Proof. The Triplet Counting Algorithm computes the number of triplets in t with LCA u that are consistent with $A, B|C$, summing this value across all vertices in t. This is correct because of Lemma 3.3 and because any triplet on leaves $x, y, z \in L(t)$ has a single LCA in t and thus is counted only once. □

Lemma 3.5. *The Triplet Counting Algorithm has time complexity $O(n)$, where n is the number of leaves (species) in the gene tree.*

Proof. At each vertex u of the gene tree t, $Trips[u]$ and $V[u]$ can be calculated in constant time because they can be calculated using information obtained at previous vertices (i.e., child and grand child vertices) in the post-order traversal. Therefore, the Triplet Counting Algorithm has time complexity $O(n)$. □

Now suppose that Q_1 is a branch (with orientated tripartition $A, B|C$) in the input species tree, and let Q_2 and Q_3 denote the alternative branches $A, B|C$ and $A, C|B$, respectively. The Triplet Counting Algorithm can be used to get the triplet counts for Q_1, Q_2, and Q_3 (call z_1, z_2, z_3) so that branch length of Q_1 can be estimated using Equation 4 in a similar manner to ASTRAL.

Algorithm 2 TripsBranchLengthEstimation($T, \mathcal{P}$)

```
    Input: A rooted species tree topology $T$ and a set $\mathcal{P}$ of rooted gene trees
    Output: The rooted species tree $T$ with estimated branch lengths (in CUs)
$L \leftarrow$ structure indexed by tree vertices ($L[v]$ denotes the list of leaves that are descendents of vertex $v$)
for $v$ in PostOrderTraverasl of $T$ do
    if $v$ is the root then
        do nothing
    else if $v$ is a leaf then
        $L[v] \leftarrow List(v)$
    else
        $z1 \leftarrow 0$
        $z2 \leftarrow 0$
        $z3 \leftarrow 0$
        for $g$ in $\mathcal{P}$ do
            $A \leftarrow L[v.left]$
            $B \leftarrow L[v.right]$
            $C \leftarrow L[v.sibling]$
            if $A, B$, and $C$ are not empty then
                $z1 \leftarrow z1 + TripletCounting(g, (A, B|C))$
                $z2 \leftarrow z2 + TripletCounting(g, (A, C|B))$
                $z3 \leftarrow z3 + TripletCounting(g, (B, C|A))$
            end if
        end for
        $z \leftarrow z1 + z2 + z3$
        if $z$ is 0 then
            $d \leftarrow 0$
        else
            $d \leftarrow -ln\frac{3}{2}(1 - \frac{z1}{z})$
        end if
        $v.setBranchLength(d)$
        $L[v] \leftarrow L[v.left] + L[v.right]$
    end if
end for
return $T$
```

Theorem 3.6. *The Trips Branch Length Estimation Algorithm has time complexity $O(n^2k)$, where n is the number of species and k is the number of gene trees.*

Proof. The Trips Branch Length Estimation Algorithm applies the Triplet Counting Algorithm on each of the k gene trees in $\mathcal{P}$ to calculate triplet counts for estimating each of the $O(n)$ internal branches of the species tree T. By Lemma 3.5, this results in an overall runtime of $O(n^2k)$. $\square$

4 Experimental Study

All computational experiments were performed on the CBCB cluster. The experimental protocol is described below, and the scripts used to run these experiments are available on Github (`https://github.com/shwaylay/triplet-brlen-study`).

4.1 Data Sets

Simulated data sets. Methods were evaluated on data sets taken from the study by Mahbub et al. (2021). Gene trees were simulated using MSC model species trees estimated in prior studies. Specifically,

- the *mammalian simulated* data sets were generated using an MSC model species tree estimated from 37 mammalian species and 447 loci (Song et al., 2012) and
- the *avian simulated* data sets were generated using an MSC model species tree estimated from 48 avian species and 14,446 loci (Jarvis et al., 2014).

DNA sequences were then simulated down each of the gene trees. The simulation protocol varied the following parameters: species tree height by scaling factor (which varies the level of ILS), number of gene trees, and sequence length (bp), which varies the level of gene tree estimation error (GTEE). They also have 20 replicates for each of the combinations of parameters.

4.2 Methods Evaluated

We evaluated our branch length estimation method, Trips, in comparison to the techniques implemented within MP-EST (Liu et al., 2010) and ASTRAL (Sayyari and Mirarab, 2016).

Trips. My method takes in three arguments, a species tree file, a gene trees file, and an output path. Tree files should be in the newick string format. You can find it here (`https://github.com/shwaylay/triplet-brlen-study/blob/main/tools/compute_triplets.py`) and run it with the following command:

```
python compute_triplets.py -s <species tree file> \
                           -g <genetrees file> \
                           -o <output path>
```

MP-EST. MP-EST v2.1 was downloaded from Github (`https://github.com/lliu1871/mp-est/`) on November 9, 2022 (commit: f85d76f) and built using gcc v4.8.5. I created a python script `https://github.com/shwaylay/triplet-brlen-study/blob/main/tools/generate_control_file.py`, which takes in a species tree file and a gene trees file and generates a control file to estimate branch lengths in species tree with MP-EST.

ASTRAL. ASTRAL v5.7.8 (henceforth called ASTRAL-III) was downloaded from Github (`https://github.com/smirarab/ASTRAL`) on November 9, 2022 (commit: 3114c92). I ran ASTRAL-III using the following command:

```
java –Xmx36G \
    –D"<astral lib directory>" \
    –jar <astral jar file> \
    –q <input species tree> \
    –t2 \
    –i <input gene trees> \
    –o <output species tree with branch lengths> \
    &> <output log file>
```

4.3 Evaluation Metrics

I evaluated Astral, MP-EST, and my method Trips by comparing their average absolute and percent errors for each branch estimation over 20 replicates. Let d^* denote the true length and let $\hat{d}$ denote the estimated length. Absolute error is calculated as

$$absolute\_error(d^*, \hat{d}) = |d^* - \hat{d}| \tag{6}$$

and percent error is calculated as

$$percent\_error = \frac{absolute\_error(d^*, \hat{d})}{d^*}. \tag{7}$$

When comparing the performance of methods under different model conditions, I also calculated error as $\hat{d} - d^*$, to get a sense as to whether methods over or under estimate. A negative error signifies an underestimate whereas a positive error signifies an overestimate.

5 Results and Discussion

Overall trends. In terms of accuracy, my method Trips preformed very similarly to ASTRAL and MP-EST. The error comparison plots (Figs. 5 and 6) all show similar trends between the three methods. All three methods tend to have low absolute error and high variation in percent error for short true branches, relatively moderate absolute and percent error for true branch lengths between 2 and 6 CUs, and a large amount of variation for true branch lengths greater than 6 CUs.

These trends motivated me to split up the data into four categories: data with true branch lengths < 0.25, $[0.25, 0.5)$, $[0.5, 2)$, and $[2, 6)$. With the split up data, I looked at (1) how varying species tree height (and thus ILS), (2) varying number of estimated gene trees, and (3) varying sequence length (and thus GTEE) impacts the performance of each method. Again, all three methods have similar error trends. Some notable trends are below.

Experiment #1: Varying species tree height (ILS level). For the avian data, varying species tree height made very little impact. All three scaling factors (0.5X, 1X, and 2X) yielded very similar errors. The errors are, however, slightly better for 2X scaling (Fig. 7).

For the mammalian data, estimates tended to improve as species tree height increased. Parameter scale2d (or 0.5X) yielded the worst errors, being nearly all underestimates. Errors for parameter noscale (or 1X) were closer to zero, but still mostly underestimates. Errors for parameter scale2u (or 2X) were the best, mostly being centered around 0. Estimates for parameter scale2u that are in the $[2, 6)$ category were mostly overestimates however (Fig. 8).

Experiment #2: Varying numbers of genes. For the avian data, varying the number of genes did not make much of a difference. All three methods also performed with MP-EST having more overestimate outliers than the other methods (Fig. 9).

For the mammalian data, varying the number of genes also did not make much of a difference, but there is a slight increase in underestimates as the number of genes increases. It is also worth noting that the spread of errors for MP-EST and Trips tends to become narrower as the number of genes increases, but the spread of ASTRAL tends to be fairly consistent (Fig. 10).

Experiment #3: Varying numbers of sequence length (GTEE level). For the avian data, errors when using true gene trees were all centered around zero. It is worth noting that MP-EST tends to have more outliers than ASTRAL and Trips. Branch lengths computed from estimated gene trees (sequence length 500) tend to become worse underestimates as the true branch length increases (Fig. 11).

Similarly, for the mammalian data, the estimates using true gene trees tend to be the best. Aside from the true gene trees, the estimated gene trees with varying sequence length produce similar results. All methods behave similarly, with MP-EST having a slightly higher variance and number of outliers than ASTRAL and Trips (Fig. 12).

6 Conclusions

In this thesis, I develop a new method, Trips, for estimating branch lengths in a species tree under the MSC model. I compared Trips to existing methods ASTRAL and MP-EST with datasets that differed in species tree height, number of genes, and bp sequence length. There were no major differences in error between ASTRAL, MP-EST, and Trips based on my preliminary results. Varying species tree height, and thus ILS level, did not make a large difference for the avian dataset, but did make a difference for the mammalian data, with estimations improving as height increases. Varying the number of genes and varying the sequence length did not make much of a difference for both datasets, aside from true gene trees producing the best estimates.

There are many ways this study can be expanded upon. Running statistical tests on our results would provide more insight in whether there is any significant difference between the methods' performances. Comparing the run-time of each method would also be a good next step for this study since the three methods had similar accuracy. Trips would need to be optimized in C++ before comparing run-time because its current implementation is in Python which is a notably slower language. Comparing the impact of missing data would be another good test. I also want to implement another method that involves treating branch lengths as a linear system.

Overall, it is interesting to see how my method, Trips, compares to the commonly used methods, ASTRAL and MP-EST, and there are still many ways this study can be improved upon.

7 Results Figures

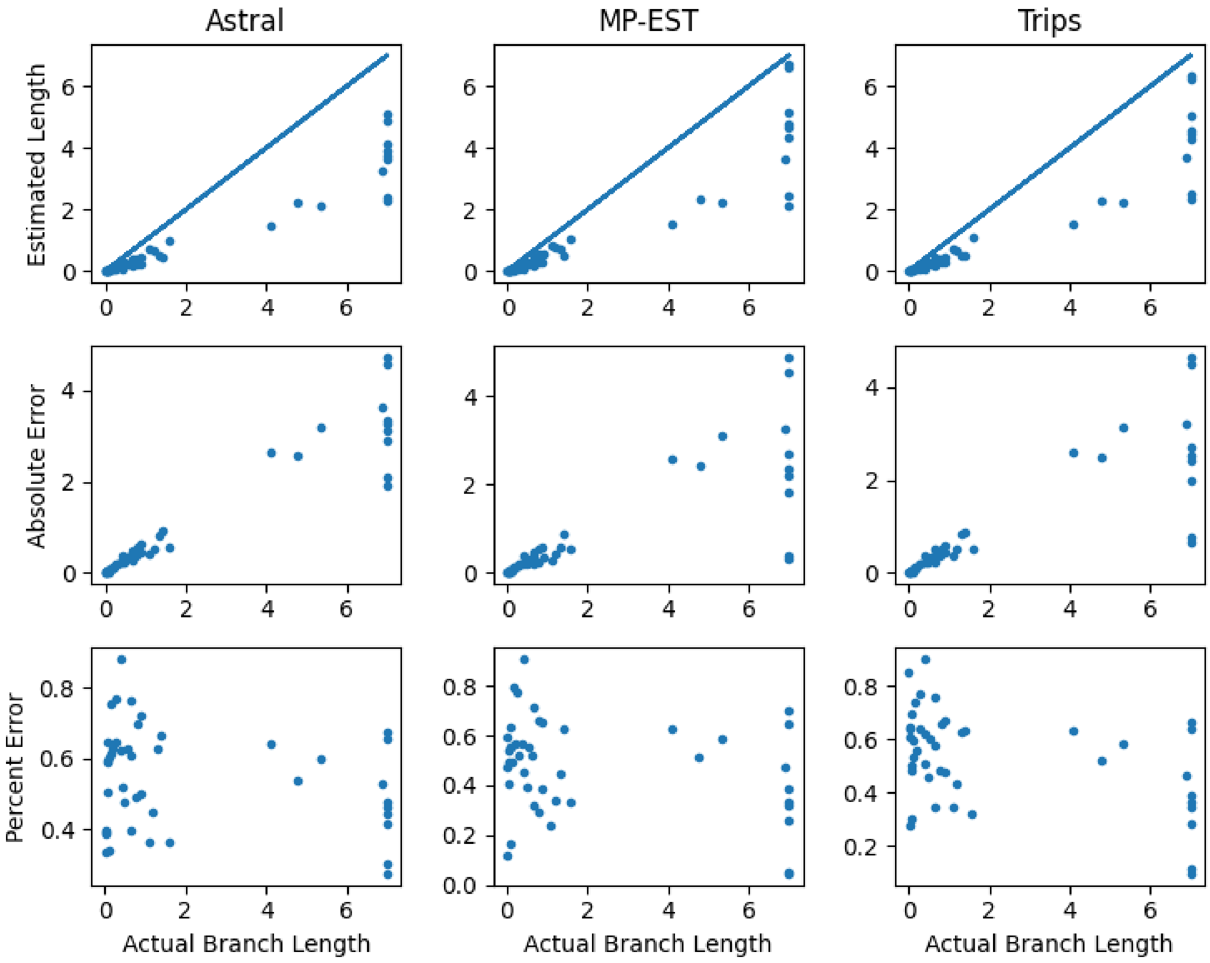


Figure 5: **Performance Comparison for ASTRAL, MP-EST, and Trips on Avian data (1X scaling, 1000 estimated gene trees, and 500 bp sequence length).** The first row shows the estimated length vs the actual length, the second row shows the absolute error and the third row shows the percent error. The lengths are in coalescent units. Each column is for one of the three methods: ASTRAL, MP-EST, and my method (called Trips).

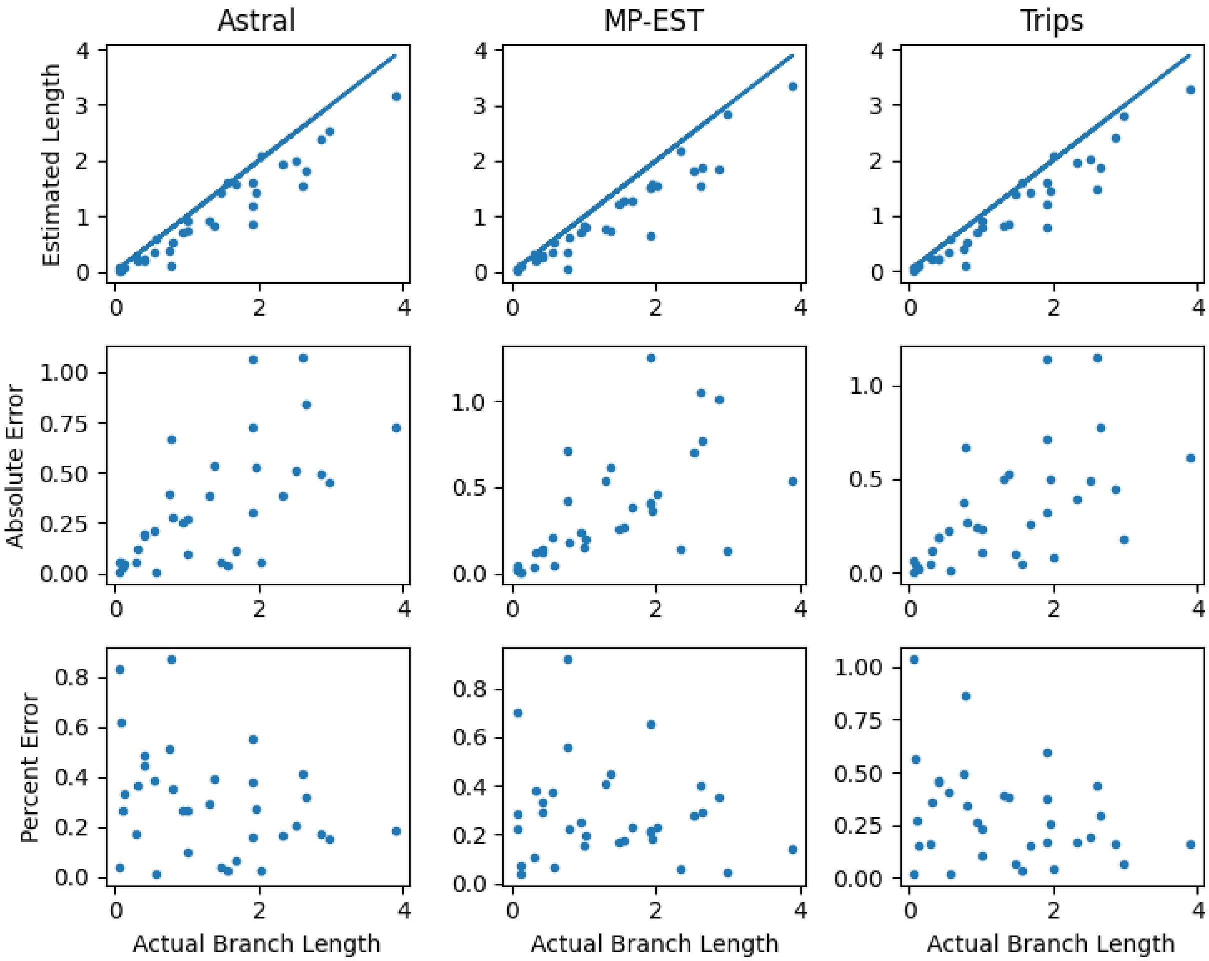


Figure 6: **Performance Comparison for ASTRAL, MP-EST, and Trips on Mammalian data (1X scaling, 200 estimated gene trees, and 500 bp sequence length).** The first row shows the estimated length vs the actual length, the second row shows the absolute error and the third row shows the percent error. The lengths are in coalescent units. Each column is for one of the three methods: ASTRAL, MP-EST, and my method (called Trips).

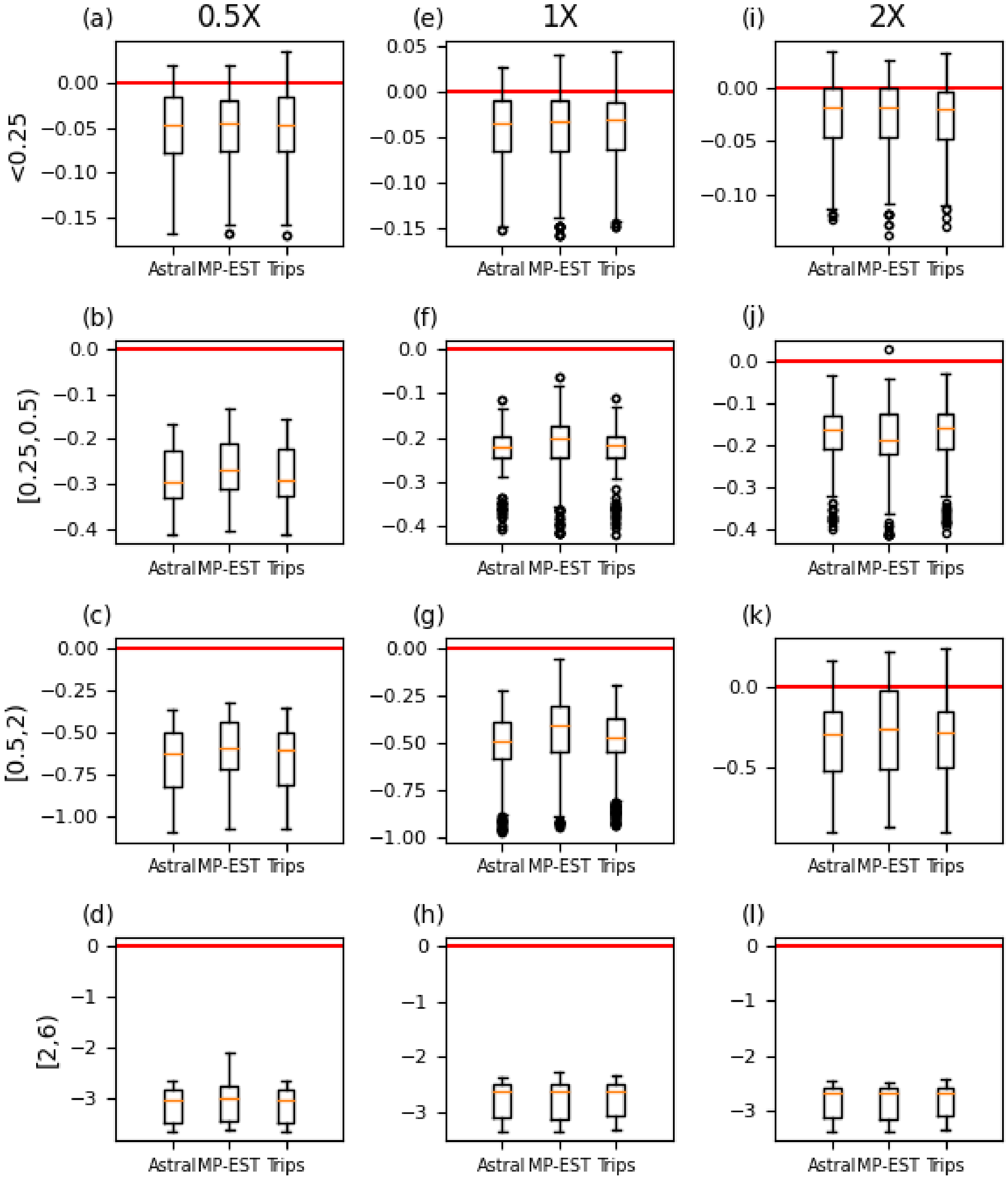


Figure 7: **Impact of varying species tree height (and thus ILS) on avian data with 1000 estimated gene trees and 500 bp sequence length.** Each subfigure shows branch length estimation error (in coalescent units) for the thee methods. The subplots in the same row show binning the true branch lengths. The subplots in the same column have the same model condition based on species tree height scaling factor.

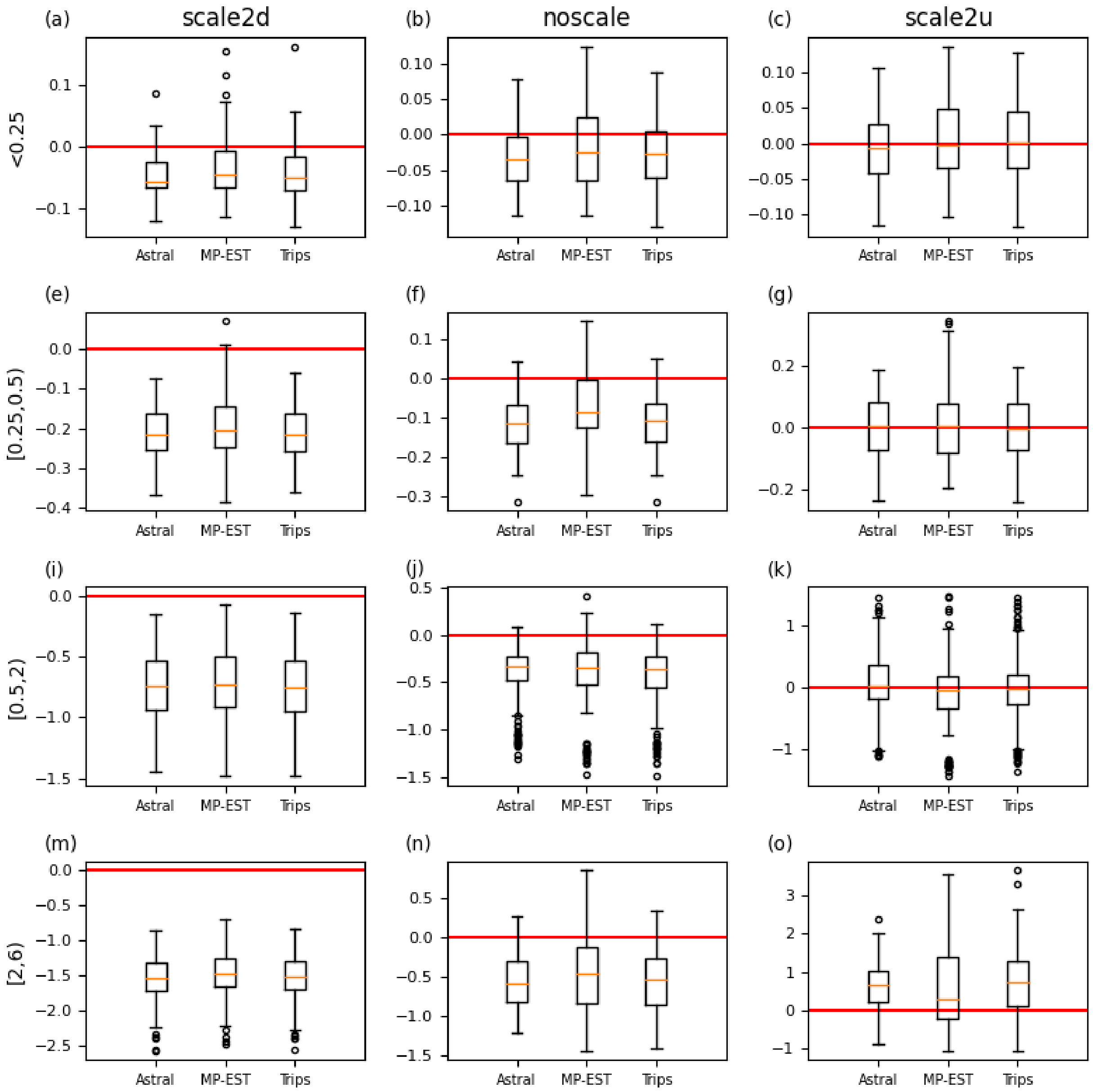


Figure 8: **Impact of varying species tree height (and thus ILS) on Mammalian data with 200 estimated gene trees - 500 bp sequence length.** Each subfigure shows branch length estimation error (in coalescent units) for the thee methods. The subplots in the same row show binning the true branch lengths. The subplots in the same column have the same model condition based on species tree height scaling factor.

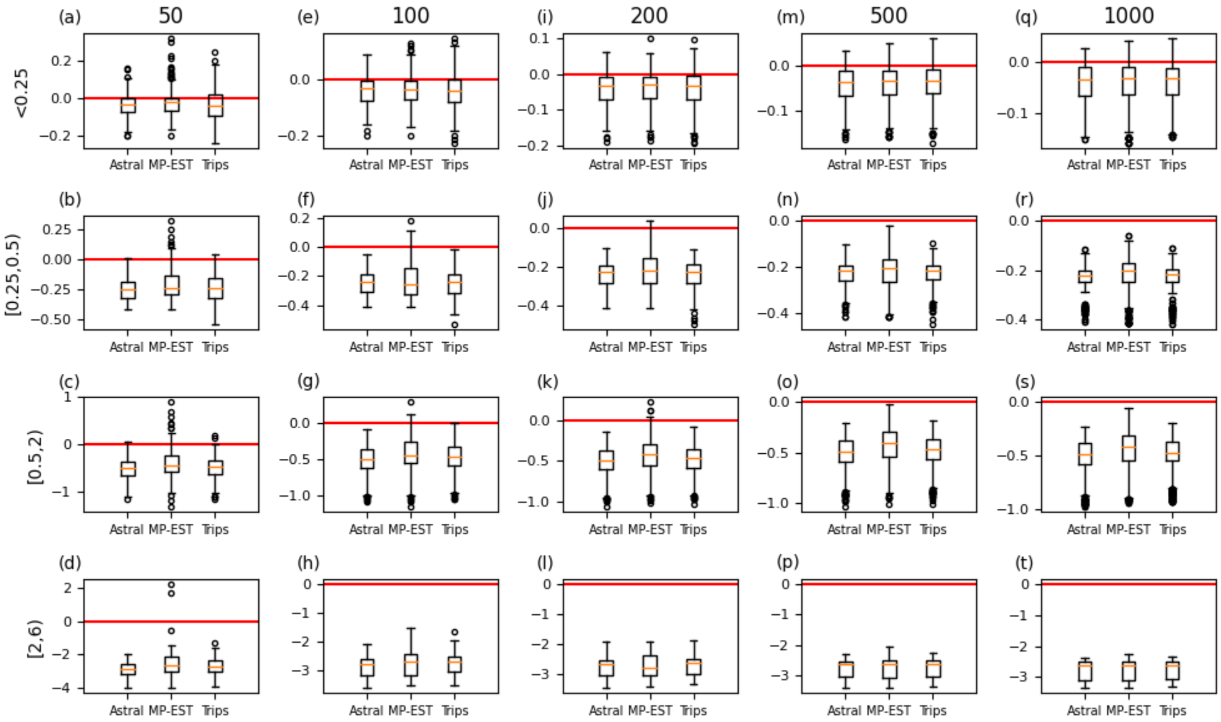


Figure 9: **Impact of varying number of genes on avian data with 1X scaling and estimated gene trees (500 bp sequence length).** Each subfigure shows branch length estimation error (in coalescent units) for the thee methods. The subplots in the same row show binning the true branch lengths. The subplots in the same column have the same model condition based on number of gene trees.

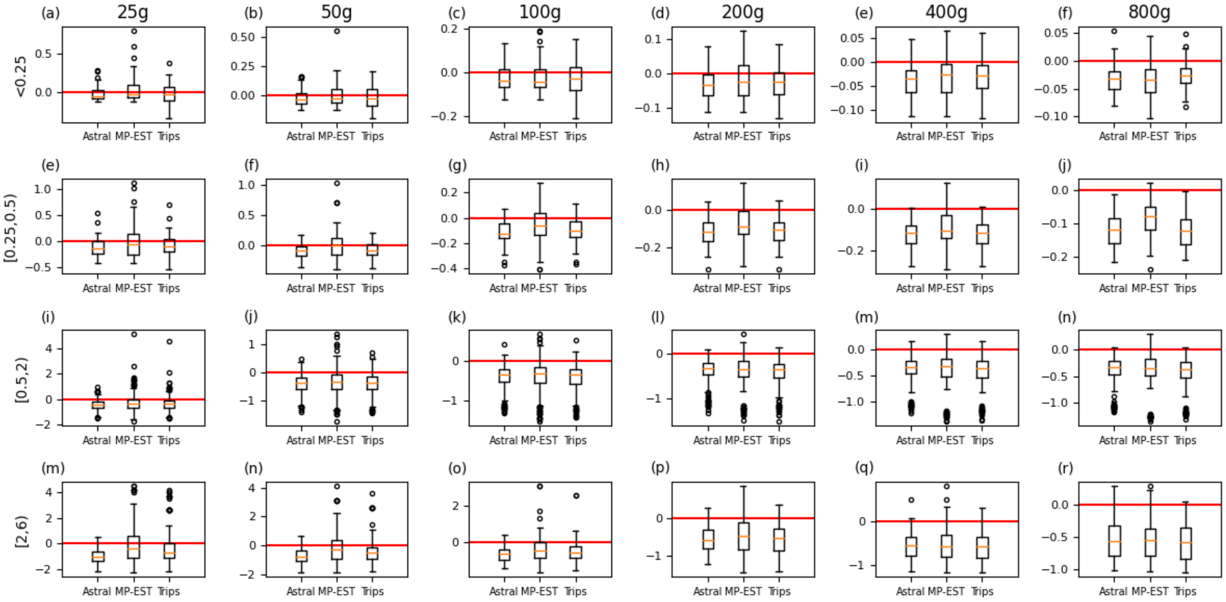


Figure 10: **Impact of varying number of genes on mammalian data with no scaling and estimated gene trees (500 bp sequence length).** Each subfigure shows branch length estimation error (in coalescent units) for the thee methods. The subplots in the same row show binning the true branch lengths. The subplots in the same column have the same model condition based on number of gene trees.

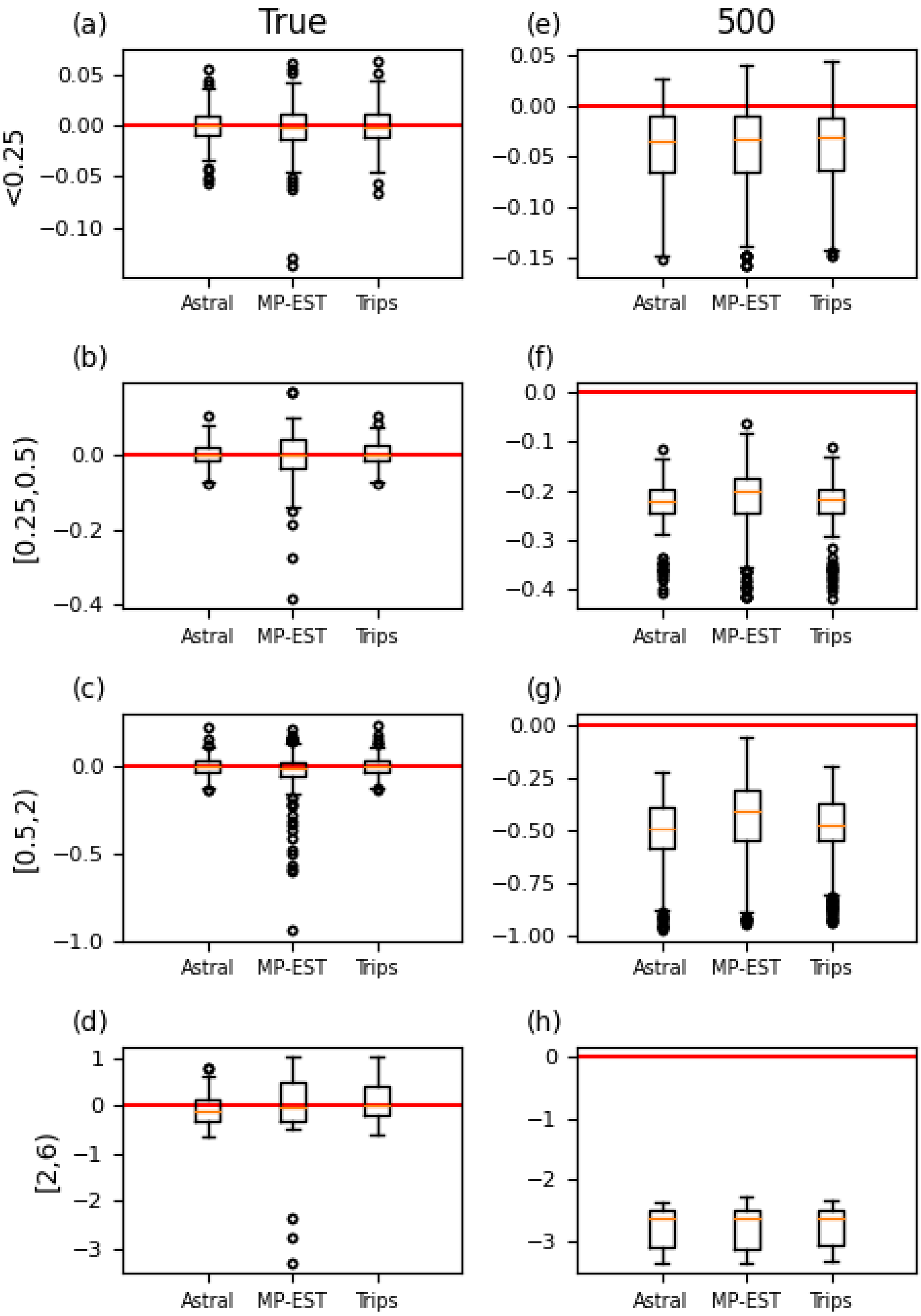


Figure 11: **Impact of gene tree estimation error on avian data with 1X scaling and 1000 gene trees** Each subfigure shows branch length estimation error (in coalescent units) for the thee methods. The subplots in the same row show binning the true branch lengths. The subplots in the same column have the same model condition based on true or estimated (sequence length 500 bp) gene trees.

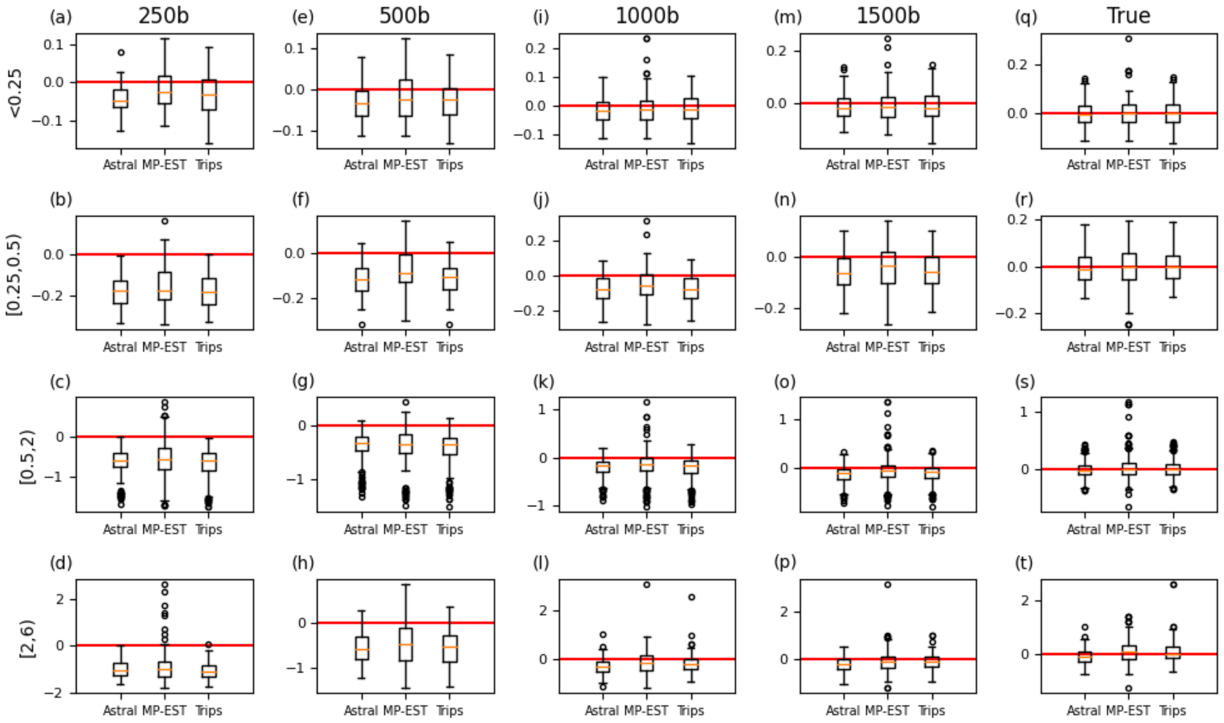


Figure 12: **Impact of gene tree estimation error on mammalian data with noscaling and 200g gene trees** Each subfigure shows branch length estimation error (in coalescent units) for the thee methods. The subplots in the same row show binning the true branch lengths. The subplots in the same column have the same model condition based on true or estimated (varying sequence length) gene trees.

References

Allman, E. S., J. H. Degnan, and J. A. Rhodes. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. Journal of Mathematical Biology 62:833–862.

Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLOS Genetics 2:1–7.

Jarvis, E. D., S. Mirarab, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Liu, L., L. Yu, and S. V. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology 10:302.

Mahbub, M., Z. Wahab, R. Reaz, M. S. Rahman, and M. S. Bayzid. 2021. wQFM: highly accurate genome-scale species tree estimation from weighted quartets. Bioinformatics Btab428.

Rannala, B. and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

Sayyari, E. and S. Mirarab. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Molecular Biology and Evolution 33:1654–1668.

Song, S., L. Liu, S. V. Edwards, and S. Wu. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. PNAS 109:14942–14947.