bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Polynomial-Time Statistical Estimation of Species Trees under Gene Duplication and Loss

Brandon Legried[1], Erin K. Molloy[2[0000−0001−5553−3312]], Tandy Warnow[2[0000−0001−7717−3514]], and Sébastien Roch[1[0000−0002−7608−8550]]

[1] University of Wisconsin-Madison, Madison, WI, USA
{roch,blegried}@math.wisc.edu
[2] University of Illinois at Urbana-Champaign, Urbana, IL, USA
{warnow,emolloy2}@illinois.edu

Abstract. Phylogenomics—the estimation of species trees from multi-locus datasets—is a common step in many biological studies. However, this estimation is challenged by the fact that genes can evolve under processes, including incomplete lineage sorting (ILS) and gene duplication and loss (GDL), that make their trees different from the species tree. In this paper, we address the challenge of estimating the species tree under GDL. We show that species trees are *identifiable* under a standard stochastic model for GDL, and that the polynomial-time algorithm ASTRAL-multi, a recent development in the ASTRAL suite of methods, is *statistically consistent* under this GDL model. We also provide a simulation study evaluating ASTRAL-multi for species tree estimation under GDL. All scripts and datasets used in this study are available on the Illinois Data Bank: https://doi.org/10.13012/B2IDB-2626814_V1.

Keywords: Species trees · gene duplication and loss · identifiability · statistical consistency · estimation · ASTRAL.

1 Introduction

Phylogeny estimation is a statistically and computationally complex estimation problem, due to heterogeneity across the genome resulting from processes such as incomplete lineage sorting (ILS), gene duplication and loss (GDL), rearrangements, gene flow, horizontal gene transfer, introgression, etc. [23]. Much is known about the problem of estimating species trees in the presence of ILS, as modelled by the Multi-Species Coalescent (MSC) [19,37]. For example, because the most probable unrooted tree for every four species is the species tree on those species [1], the unrooted species tree topology is identifiable under the MSC from its gene tree distribution, and quartet-based species tree estimation methods that operate by combining gene trees (such as BUCKy-pop [20] and ASTRAL [25,27,44]) are statistically consistent estimators of the unrooted species tree topology (i.e., as the number of sampled genes increases, almost surely the tree returned by these methods will be the true species tree). It is also known that concatenation (whether partitioned or unpartitioned) is not statistically consistent, and can

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

even be positively misleading (i.e., converge to the wrong tree as the number of loci increases) [34,32]. In general, establishing whether a method is statistically consistent or not is important for understanding its performance guarantees.

Yet, correspondingly little has been established about species tree estimation in the presence of GDL. For example, although likelihood-based approaches for species tree estimation have been developed (e.g., PHYLDOG [7]), they have not been established to be statistically consistent. Key to understanding the performance of species tree estimation under GDL is whether the species tree topology itself is identifiable from the distribution it defines on the gene trees it generates. However, since gene trees can have multiple copies of each species when gene duplication occurs, this question can be formulated as: "Is the species tree identifiable from the distribution on MUL-trees?", where a MUL-tree is a tree with potentially multiple copies of each species.

In this paper, we prove that unrooted species tree topologies are identifiable from the distribution implied on MUL-trees (Section 3) under the simple GDL model of [2]. Furthermore, we prove that the polynomial-time method ASTRAL-multi [29], a recent variant of ASTRAL designed to enable analyses of datasets with multiple individuals per species, is statistically consistent under this model (Section 3). We then present an experimental study evaluating ASTRAL-multi on 16-taxon datasets simulated under the DLCoal model (a unified model of GDL and ILS) [30]; the results of this study show that when given a sufficiently large number of genes, ASTRAL-multi is competitive with other methods (e.g., DupTree [4], MulRF [9], and ASTRID-multi [39], the implementation of ASTRID for multi-allele datasets) that also estimate species trees from MUL-trees (Section 4). We conclude with remarks about future work and implications for large-scale species tree estimation (Section 5).

2 Species tree estimation from gene families

Our input is a collection $\mathcal{T}$ of gene trees representing the inferred evolutionary histories of gene families. In the presence of gene duplication and loss events, such gene trees may be multi-labeled trees (MUL-trees), meaning that the same species label may be assigned to several gene copies. Our goal is to reconstruct a species tree T over the corresponding set S of species.

ASTRAL. We provide theoretical guarantees and empirically validate an approach based on ASTRAL [25] in its variant for multiple alleles [29], which we refer to as ASTRAL-multi. Following [12], the input consists of unrooted MUL-trees $\mathcal{T}$ from all gene families, where copies of a gene in a species are treated as multiple alleles within the species.

ASTRAL-multi proceeds as follows. Let S be the set of n species and let R be the set of m individuals. The input are the gene trees $\mathcal{T} = \{t_i\}_{i=1}^{k}$, where t_i is labeled by individuals $R_i \subseteq R$. For any (unrooted) species tree $\widetilde{T}$ labeled by S, an extended species tree $\widetilde{T}_{ext}$ labeled by R is built by adding to each leaf of $\widetilde{T}$ all individuals corresponding to that species as a polytomy. The quartet score

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of $\widetilde{T}$ with respect to $\mathcal{T}$ is then

$$Q_k(\widetilde{T}) = \sum_{i=1}^{k} \sum_{\mathcal{J}=\{a,b,c,d\}\subseteq R_i} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, t_i^{\mathcal{J}}), \tag{1}$$

where $\mathbf{1}(T_1, T_2)$ is the indicator that T_1 and T_2 agree and $T_1^{\mathcal{J}}$ is the restriction of T_1 to individuals $\mathcal{J}$. Run in its *exact* version (i.e., an unrooted species tree that maximizes the quartet score), ASTRAL-multi is guaranteed to find an optimal solution, but can use exponential time. The *default* mode, which runs in polynomial time, uses dynamic programming to solve a constrained version of the problem, requiring that the output tree draw its bipartitions from a set Σ of bipartitions that ASTRAL computes on the input, where Σ by construction includes all the bipartitions on S that occur in any gene tree in $\mathcal{T}$.

3 Theoretical results

In this section, we provide theoretical guarantees for the reconstruction algorithm discussed in Section 2. Specifically, we establish statistical consistency under a standard model of GDL [2]. First we show that the species tree is identifiable.

3.1 Gene duplication and loss model

We assume in this section that gene tree heterogeneity is due exclusively to GDL (and so no ILS) and that the true gene trees are known. That is, there is no gene tree estimation error (GTEE).

Birth-death process of gene duplication and loss The rooted n-species tree $T = (V, E)$ has vertices V and directed edges E with lengths (in time units) η that depend on the edge. For ease of presentation, we assume that there is a single copy of each gene at the root of T and that the rates of duplication λ and loss μ are fixed throughout T (although our proofs do not use these assumptions). Each gene tree is generated by a top-down birth-death process within the species tree. That is, on each edge, each gene copy independently duplicates at exponential rate λ and is lost at exponential rate μ; at speciation events, each gene copy bifurcates and proceeds similarly in the descendant edges. Each duplication is indicated in the gene tree by a bifurcation. The resulting gene tree is then pruned of lost copies to give the observed unrooted gene tree t_i. The gene trees $\{t_i\}_{i=1}^{k}$ are assumed independent and identically distributed. See more details in [2].

3.2 Identifiability of the species tree under the GDL model

We first show that the unrooted species tree is identifiable from the distribution of MUL-trees $\mathcal{T}$ under the GDL model over T. That is, that two distinct unrooted species trees necessarily produce different gene tree distributions.

We begin with a quick proof sketch. The idea is to show that, for each 4-tuple of species $\mathcal{Q} = \{A, B, C, D\}$, the corresponding species quartet topology can be

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

identified by taking an independent uniform random gene copy in each species in $\mathcal{Q}$ and showing that the quartet topology consistent with the species tree is most likely to result in the gene tree restricted to these copies. It should be noted that the proof is not as straightforward as it is under the multispecies coalescent [1], as we explain next. Assume the species tree restricted to $\mathcal{Q}$ is $((A, B), (C, D))$, let R be the most recent common ancestor of $\mathcal{Q}$ in T, and let a, b, c, d be random gene copies in A, B, C, D respectively.

- When all ancestral copies of a, b, c, d in R are distinct, by symmetry *all quartet topologies are equally likely.* The *ancestral copy of x in R* is the vertex of the gene tree that is ancestral to x and corresponds to a speciation event at node R of the species tree.
- When the ancestors of a and b (or c and d) in R are the same, the *species quartet topology results.*
- *However,* there are further cases. For example, if the ancestors of a and c in R coincide while being distinct from those of b and d, then the resulting quartet topology *differs* from that of the species tree.

Hence, one must carefully account for all possible cases to establish that the species quartet topology is indeed likeliest, which we do next. Our argument relies primarily on the symmetries (i.e., exchangeability) of the process.

Theorem 1 (Identifiability). *Let T be a species tree with $n \geq 4$ leaves. Then T, without its root, is identifiable from the distribution of MUL-trees $\mathcal{T}$ under the GDL model over T.*

Proof. It is known that the unrooted topology of a species tree is defined by its set of quartet trees [3]. Let $\mathcal{Q} = \{A, B, C, D\}$ be four distinct species in T and let $T^{\mathcal{Q}}$ be the species tree restricted to $\mathcal{Q}$. Assume without loss of generality that the corresponding unrooted quartet topology is $AB|CD$. Let t be a MUL-tree generated under the GDL model over T and let $t^{\mathcal{Q}}$ be its restriction to the gene copies from species in $\mathcal{Q}$. Conditioning on having at least one gene copy in the species $\mathcal{Q}$, independently pick a uniformly random gene copy a, b, c, d in species A, B, C, D respectively and let q be the corresponding quartet topology under $t^{\mathcal{Q}}$. We show that the most likely outcome is $q = ab|cd$. There are two cases: $T^{\mathcal{Q}}$ is 1) balanced or 2) a caterpillar.

In case 1), let R be the most recent common ancestor of $\mathcal{Q}$ in T and let I be the number of gene copies exiting (forward in time) R. By the law of total probability, $\mathbf{P}'[q = ab|cd] = \mathbf{E}'[\mathbf{P}'_I[q = ab|cd]]$, where the primes indicate that we are conditioning on having at least one gene copy in each species in $\mathcal{Q}$ and the subscript I indicates conditioning on I. So it suffices to prove

$$\mathbf{P}'_I[q = ab|cd] > \max\left\{\mathbf{P}'_I[q = ac|bd], \mathbf{P}'_I[q = ad|bc]\right\}, \tag{2}$$

almost surely. Let $i_x \in \{1, \ldots, I\}$ be the ancestral lineage of $x \in \{a, b, c, d\}$ in R. Then

$$\begin{aligned}\mathbf{P}'_I[q = ab|cd] = \mathbf{P}'_I[i_a = i_b] + \mathbf{P}'_I[i_c = i_d] - \mathbf{P}'_I[i_a = i_b, i_c = i_d] \\ + \mathbf{P}'_I[q = ab|cd \text{ and } i_a, i_b, i_c, i_d \text{ all distinct}].\end{aligned} \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

On the other hand,

$$\begin{aligned}\mathbf{P}'_I[q = ac|bd] \leq \mathbf{P}'_I[i_b \neq i_a = i_c \neq i_d] + \mathbf{P}'_I[i_a \neq i_b = i_d \neq i_c] \\ + \mathbf{P}'_I[q = ac|bd \text{ and } i_a, i_b, i_c, i_d \text{ all distinct}],\end{aligned} \tag{4}$$

and similarly for $\mathbf{P}'_I[q = ac|bd]$, where note that we double-counted the case $i_a = i_c \neq i_d = i_b$ to simplify the expression. By symmetry of the GDL process above R (which holds under $\mathbf{P}'_I$), the last term on the RHS of (3) and (4) are the same. The same holds for the first two terms on the RHS of (4) this time by the independence and exchangeability of the pairs (i_a, i_b) and (i_c, i_d) under $\mathbf{P}'_I$, which further implies

$$\begin{aligned}&\mathbf{P}'_I[q = ab|cd] - \mathbf{P}'_I[q = ac|bd] \\ &\geq \mathbf{P}'_I[i_a = i_b] + \mathbf{P}'_I[i_c = i_d] - \mathbf{P}'_I[i_a = i_b, i_c = i_d] - 2\mathbf{P}'_I[i_b \neq i_a = i_c \neq i_d] \\ &= x + y - xy - 2(1-x)(1-y)\mathbf{P}'_I[i_a = i_c \,|\, i_a \neq i_b, i_c \neq i_d] \\ &= x + y - xy - 2(1-x)(1-y)\mathbf{P}'_I[i_a = i_c] \\ &= x + y - xy - 2(1-x)(1-y)\frac{1}{I} \equiv h(x, y).\end{aligned}$$

where $x = \mathbf{P}'_I[i_a = i_b]$ and $y = \mathbf{P}'_I[i_c = i_d]$.

For fixed y, $h(x, y)$ is linear in x and $h(1, y) = 1$. So $h(\cdot, y)$ achieves its minimum at the smallest value allowed for x. The same holds for y. Intuitively, i_a and i_b are "positively correlated" so $x \geq 1/I$. We prove this formally next.

Lemma 1. *Almost surely, $x, y \geq 1/I$.*

Proof. For $j \in \{1, \ldots, I\}$, let N_j be the number of gene copies at the most recent common ancestor R' of A and B that descend from copy j in R. Upon conditioning on $(N_j)_j$, the choice of a and b is independent, with i_a and i_b being picked proportionally to the corresponding N_j's (i.e., the gene copies in R' are equally likely to have given rise to a). By the law of total probability and the fact that the quadratic mean is greater than the arithmetic mean,

$$\mathbf{P}'_I[i_a = i_b] = \mathbf{E}'_I[\mathbf{P}'_I[i_a = i_b \,|\, (N_j)_j]] = \mathbf{E}'_I\left[\frac{\sum_{j=1}^I N_j^2}{\left(\sum_{j=1}^I N_j\right)^2}\right] \geq \frac{1}{I},$$

and similarly for $\mathbf{P}'_I[i_c = i_d]$. □

Returning to the proof of the theorem, evaluating h at $x, y = 1/I$ gives

$$h(1/I, 1/I) = 2\frac{1}{I} - \frac{1}{I^2} - 2\frac{(I-1)^2}{I^3} = \frac{2I^2 - I}{I^3} - \frac{2I^2 - 4I + 2}{I^3} = \frac{3I - 2}{I^3} > 0.$$

That establishes (2) in case 1), which implies

$$\mathbf{P}'[q = ab|cd] > \max\left\{\mathbf{P}'[q = ac|bd], \mathbf{P}'[q = ad|bc]\right\}, \tag{5}$$

as desired. The proof in case (2) can be found in the appendix. □

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

As a direct consequence of our identifiability proof, it is straightforward to establish the statistical consistency of the following pipeline, which we refer to as ASTRAL/ONE (see also [12]): for each gene tree t_i, pick in each species a random gene copy (if possible) and run ASTRAL on the resulting set of modified gene trees $\tilde{t}_i$. The proof can be found in the appendix.

Theorem 2 (Statistical Consistency: ASTRAL/ONE). *ASTRAL/ONE is statistically consistent under the GDL model. That is, as the number of input gene trees tends toward infinity, the output of ASTRAL/ONE converges to T almost surely, when run in exact mode or in its default constrained version.*

3.3 Statistical consistency of ASTRAL-multi under GDL

The following consistency result is not a direct consequence of our identifiability result, although the ideas used are similar.

Theorem 3 (Statistical Consistency: ASTRAL-multi). *ASTRAL-multi, where copies of a gene in a species are treated as multiple alleles within the species, is statistically consistent under the GDL model. That is, as the number of input gene trees tends toward infinity, the output of ASTRAL-multi converges to T almost surely, when run in exact mode or in its default constrained version.*

Proof. First, we show that ASTRAL-multi is consistent when run in exact mode. The input are the gene trees $\mathcal{T} = \{t_i\}_{i=1}^k$ with t_i labelled by individuals (i.e., gene copies) $R_i \subseteq R$. Then the quartet score of $\widetilde{T}$ with respect to $\mathcal{T}$ is given by (1). For any 4-tuple of gene copies $\mathcal{J} = \{a, b, c, d\}$, we define $m(\mathcal{J})$ to be the corresponding set of species. It was proved in [29] that those $\mathcal{J}$'s with fewer than 4 species contribute equally to all species tree topologies. As a result, it suffices to work with a modified quartet score

$$\widetilde{Q}_k(\widetilde{T}) = \sum_{i=1}^{k} \sum_{\substack{\mathcal{J}=\{a,b,c,d\}\subseteq R_i \\ |m(\mathcal{J})|=4}} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, t_i^{\mathcal{J}}).$$

By independence of the gene trees (and non-negativity), $\widetilde{Q}_k(\widetilde{T})/k$ converges almost surely to its expectation simultaneously for all unrooted species tree topologies over S.

The expectation can be simplified as

$$\begin{aligned}\mathbf{E}\left[\frac{1}{k}\widetilde{Q}_k(\widetilde{T})\right] &= \mathbf{E}\left[\sum_{\substack{\mathcal{J}=\{a,b,c,d\}\subseteq R_1 \\ |m(\mathcal{J})|=4}} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, t_1^{\mathcal{J}})\right] \\ &= \sum_{\mathcal{Q}=\{A,B,C,D\}} \mathbf{E}\left[\sum_{\mathcal{J}\subseteq R_1 : m(\mathcal{J})=\mathcal{Q}} \mathbf{1}(\widetilde{T}_{ext}^{\mathcal{J}}, t_1^{\mathcal{J}})\right]. \qquad (6)\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Let $\mathcal{N}^{\mathcal{Q}}_{AB|CD}$ (respectively $\mathcal{N}^{\mathcal{Q}}_{AC|BD}, \mathcal{N}^{\mathcal{Q}}_{AD|BC}$) be the number of choices consisting of one gene copy in t_1 from each species in $\mathcal{Q}$ whose corresponding restriction $t_1^{\mathcal{Q}}$ agrees with $AB|CD$ (respectively $AC|BD$, $AD|BC$). Then each summand in (6) may be written as $\mathbf{E}[\mathcal{N}^{\mathcal{Q}}_{\widetilde{T}^{\mathcal{Q}}}]$. We establish below that this last expression is maximized at the true species tree $T^{\mathcal{Q}}$, that is,

$$\mathbf{E}[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] > \max\left\{\mathbf{E}[\mathcal{N}^{\mathcal{Q}}_{AC|BD}], \mathbf{E}[\mathcal{N}^{\mathcal{Q}}_{AD|BC}]\right\}, \tag{7}$$

when (without loss of generality) $T^{\mathcal{Q}} = AB|CD$. From (6) and the law of large numbers, it will then follow that almost surely the quartet score is eventually maximized by the true species tree as $k \to +\infty$.

It remains to establish (7). Fix $\mathcal{Q} = \{A, B, C, D\}$ a set of four distinct species in T. Assume that the corresponding unrooted quartet topology in T is $AB|CD$. Let t_1 be a MUL-tree generated under the GDL model over T. Again, there are two cases: $T^{\mathcal{Q}}$ is 1) balanced or 2) a caterpillar.

In case 1), let R be the most recent common ancestor of $\mathcal{Q}$ in T and let I be the number of gene copies exiting (forward in time) R. For $j \in \{1, \ldots, I\}$, let $\mathcal{A}_j$ be the number of gene copies in A descending from j in R, and similarly define $\mathcal{B}_j$, $\mathcal{C}_j$ and $\mathcal{D}_j$. By the law of total probability, $\mathbf{E}[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] = \mathbf{E}[\mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}]]$. We show that, almost surely,

$$\mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] > \max\left\{\mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AC|BD}], \mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AD|BC}]\right\}, \tag{8}$$

which implies (7). By symmetry, we have $X^{=} \equiv \mathbf{E}_I[\mathcal{A}_j\mathcal{B}_j] = \mathbf{E}_I[\mathcal{A}_1\mathcal{B}_1]$, $Y^{=} \equiv \mathbf{E}_I[\mathcal{C}_j\mathcal{D}_j] = \mathbf{E}_I[\mathcal{C}_1\mathcal{D}_1]$, $X^{\neq} \equiv \mathbf{E}_I[\mathcal{A}_j\mathcal{B}_k] = \mathbf{E}_I[\mathcal{A}_1]\mathbf{E}_I[\mathcal{B}_1]$ as well as $Y^{\neq} \equiv \mathbf{E}_I[\mathcal{C}_j\mathcal{D}_k] = \mathbf{E}_I[\mathcal{C}_1]\mathbf{E}_I[\mathcal{D}_1]$ for all j, k with $j \neq k$. Hence, the expected number of pairs consisting of a single gene copy from A and B is $X = IX^{=} + I(I-1)X^{\neq}$. Arguing similarly to (3) and (4),

$$\begin{aligned}
&\mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] - \mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] \\
&\quad \geq (IX^{=})Y + X(IY^{=}) - (IX^{=})(IY^{=}) - I(I-1)X^{\neq}[2(I-1)Y^{\neq}] \\
&\quad = XY\left[x + y - xy - 2(1-x)(1-y)\frac{1}{I}\right],
\end{aligned}$$

where here we define $x = \frac{IX^{=}}{X}, y = \frac{IY^{=}}{Y}$. Following the argument in the proof of Theorem 1, to establish (8) it suffices to show that almost surely, $x, y \geq 1/I$. That is implied by the following positive correlation result.

Lemma 2. *Almost surely,* $X^{=} \geq X^{\neq}$.

Indeed, we then have: $x = \frac{IX^{=}}{IX^{=}+I(I-1)X^{\neq}} \geq \frac{IX^{=}}{IX^{=}+I(I-1)X^{=}} = \frac{1}{I}$.

Proof (Lemma 2). For $j \in \{1, \ldots, I\}$, let N_j be the number of gene copies at the divergence of the most recent common ancestor of A and B that are descending from j in R. Then, for $j \in \{1, \ldots, I\}$, since $\mathcal{A}_j$ and $\mathcal{B}_j$ are conditionally independent given $(N_j)_j$ under $\mathbf{E}_I$, it follows that

$$X^{=} = \mathbf{E}_I[\mathbf{E}_I[\mathcal{A}_j\mathcal{B}_j \mid (N_j)_j]] = \mathbf{E}_I[(N_j\alpha)(N_j\beta)] = \alpha\beta\mathbf{E}_I[N_j^2],$$

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where α (respectively β) is the expected number of gene copies in A (respectively B) descending from a single gene copy in the most recent common ancestor of A and B under $\mathbf{E}_I$. Similarly, for $j \neq k \in \{1, \ldots, I\}$,

$$X^{\neq} = \mathbf{E}_I[\mathbf{E}_I[\mathcal{A}_j \mathcal{B}_k \mid (N_j)_j]] = \mathbf{E}_I[(N_j\alpha)(N_k\beta)] = \alpha\beta\mathbf{E}_I[N_j N_k] \leq \alpha\beta\mathbf{E}_I[N_j^2],$$

by Cauchy-Schwarz and $\mathbf{E}_I[N_j^2] = \mathbf{E}_I[N_k^2]$. □

We establish (8) in case (2) in the appendix. Thus, ASTRAL-multi is statistically consistent when run in exact mode, because it is guaranteed to return the optimal tree, and that is realized by the species tree. To see why the default version of ASTRAL-multi is also statistically consistent, note that the true species tree will appear as one of the input gene trees, almost surely, as the number of MUL-trees sampled tends to infinity. For instance, the probability of observing no duplications or losses is strictly positive. Furthermore, when this happens, the true species tree bipartitions are all contained in the constraint set Σ used by the default version. Hence, as the number of sampled MUL-trees increases, almost surely ASTRAL-multi will return the true species tree topology. □

4 Experiments

We performed a simulation study to evaluate ASTRAL-multi and other species tree estimation methods on 16-taxon datasets with model conditions characterized by three GDL rates, five levels of gene tree estimation error (GTEE), and four numbers of genes. Due to space constraints, we briefly describe the study here and provide details sufficient to reproduce the study on bioRxiv: https://doi.org/10.1101/821439. In addition, all scripts and datasets used in this study are available on the Illinois Data Bank: https://doi.org/10.13012/B2IDB-2626814V1.

Our simulation protocol uses parameters estimated from the 16-taxon fungal dataset studied in [12,30]. First, we used the species tree and other parameters estimated from the fungal dataset to simulate gene trees under the DLCoal [30] model with three GDL rates (the lowest rate 1×10^{-10} reflects the GDL rate estimated from the fungal dataset, so that the two higher rates reflect more challenging model conditions). Specifically, for each GDL rate, we simulated 10 replicate datasets (each with 1000 model gene trees that deviated from the strict molecular clock) using SimPhy [24]. Although we simulated gene trees under a unified model of GDL and ILS, there was effectively no ILS in our simulated datasets (Table 1 on bioRxiv: https://doi.org/10.1101/821439). Second, for each model gene tree, we used INDELible [15] to simulate a multiple sequence alignment under the GTR+GAMMA model with parameters based on the fungal dataset. Third, we ran RAxML [35] to estimate a gene tree under the GTR+GAMMA model from each gene alignment. By varying the length of each gene alignment, four model conditions were created with 23% to 65% mean GTEE, as measured by the normalized Robinson-Foulds (RF) distance [31] between true and estimated gene trees, averaged across all gene trees. Fourth, we ran species tree estimation methods given varying numbers of gene trees as

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

input. Finally, we evaluated species tree error as the normalized RF distance between true and estimated species trees.

In our first experiment, we explored ASTRAL-multi on both true and estimated gene trees (Fig. 1). ASTRAL-multi was very accurate on true gene trees; even with just 25 true gene trees, the average species tree error was less than 1% for the two lower GDL rates and was less than 6% for the highest GDL rate (5×10^{-10}). As expected, species tree error increased with the GDL rate, increased with the GTEE level, and decreased with the number of genes.

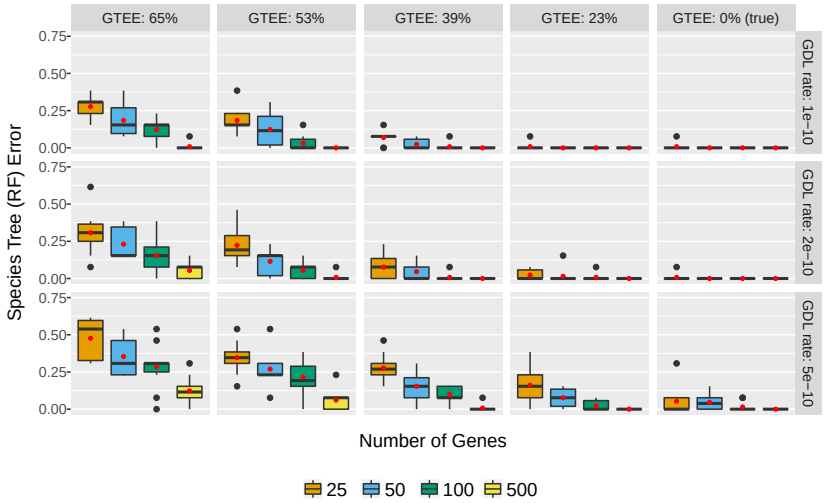

Fig. 1. ASTRAL-multi on true and estimated gene trees generated from the fungal species tree (16 taxa) under a GDL model using three different rates (subplot rows). Estimated gene trees had four different levels of gene tree estimation error (GTEE), by varying the sequence length (subplot columns). We report the average Robinson-Foulds (RF) error rate between the true and estimated species trees. There are 10 replicate datasets per model condition. Red dots indicate means, and bars indicated medians.

In our second experiment, we compared ASTRAL-multi to four other species tree methods (DupTree [42], MulRF [9], STAG [14], and ASTRID-multi, which is ASTRID [40] run under the multi-allele setting) that take gene trees as input. Figure 2 shows species tree error for model conditions with mean GTEE of 53%. As expected, the error increased for all methods with the GDL rate and GTEE level, and decreased with the number of genes. Differences between methods depended on the model condition. When given 500 genes, all five methods were competitive (with a slight disadvantage to STAG); a similar trend was observed when methods were given 100 genes provided that the GDL rate was one of the two lower rates. When given 50 genes, ASTRAL-multi, MulRF, and ASTRID-multi were the best methods for the two lower GDL rates. On the remaining

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

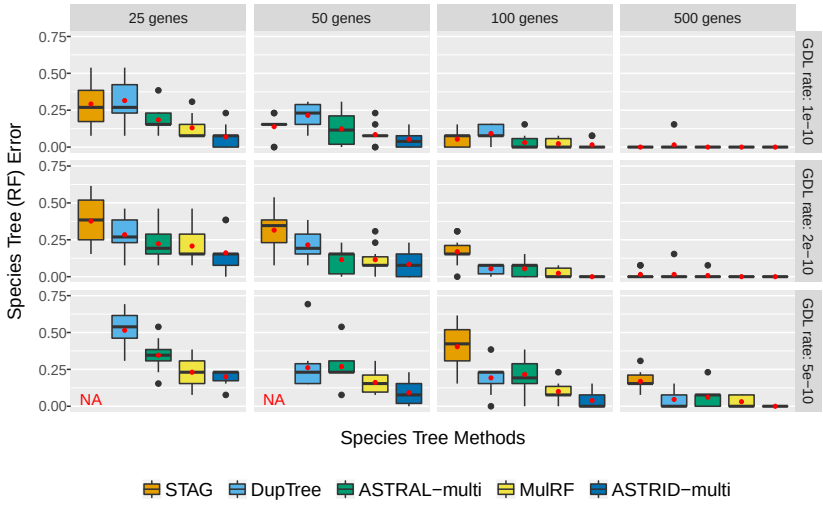

Fig. 2. Average RF tree error rates of species tree methods on estimated gene trees (mean GTEE: 53%) generated from the fungal 16-taxon species tree using three different GDL rates (subplot rows) and different numbers of genes (subplot columns). STAG failed to run on some replicate datasets for model conditions indicated by "NA", because none of the input gene trees included at least one copy of every species.

model conditions, ASTRID-multi was the best method. Finally, STAG was unable to run on some datasets when the GDL rate was high and the number of genes was low; this result was due to STAG failing when none of the input gene trees included at least one copy of every species. Results for other GTEE levels are provided Table 2 on bioRxiv: https://doi.org/10.1101/821439, and show similar trends.

5 Discussion and Conclusion

This study establishes the identifiability of unrooted species trees under the simple model of GDL from [2] and that ASTRAL-multi is statistically consistent under this model. In our simulation study, ASTRAL-multi was accurate under challenging model conditions, characterized by high GDL rates and high GTEE, provided that a sufficiently large number of genes is given as input. When the number of genes was smaller, ASTRID-multi often had an advantage over ASTRAL-multi and the other methods.

The results of this study can be compared to the previous study by Chaudhary *et al.* [8], who also evaluated species tree estimation methods under model conditions with GDL. They found that MulRF and gene tree parsimony methods had better accuracy than NJst [22] (a method that is similar to ASTRID). Their study has an advantage over our study in that it explored larger datasets (up to

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

500 species); however, all genes in their study evolved under a strict molecular clock, and they did not evaluate ASTRAL-multi.

Our study is the first study to evaluate ASTRAL-multi on *estimated* gene trees, and we also explore model conditions with varying levels of GTEE. Evaluating methods under conditions with moderate to high GTEE is critical, as estimated gene trees from four recent studies [6,17,18,36] all had mean bootstrap support values below 50% (see Table 1 in [28]), suggesting high GTEE.

Our study is limited to one underlying species tree topology with 16 species. Previous studies [41] have shown that MulRF (which uses a heuristic search strategy to find solutions to its NP-hard optimization problem) is much slower than ASTRAL on large datasets, suggesting that ASTRAL-multi may dominate MulRF as the number of species increases. Hence, future studies should investigate ASTRAL-multi and other methods under a broader range of conditions, including larger numbers of species. Future research should also consider empirical performance and statistical consistency under different causes of gene tree heterogeneity.

We note with interest that the proof that ASTRAL-multi is statistically consistent is based on the fact that the most probable unrooted gene tree on four leaves (according to two ways of defining it) under the GDL model is the true species tree (equivalently, there is no anomaly zone for the GDL model for unrooted four-leaf trees). This coincides with the reason ASTRAL is statistically consistent under the MSC as well as under a model for random HGT [33,10]. Furthermore, previous studies have shown that ASTRAL has good accuracy in simulation studies where both ILS and HGT are present [11]. Hence ASTRAL, which was originally designed for species tree estimation in the presence of ILS, has good accuracy and theoretical guarantees under different sources of gene tree heterogeneity.

We also note the surprising accuracy of DupTree, MulRF, and ASTRID-multi, methods that, like ASTRAL-multi, are not based on likelihood under a GDL model. Therefore, DynaDup [26,5] is also of potential interest, as it is similar to DupTree in seeking a tree that minimizes the duploss score (though the score is modified to reflect true biological loss), but has the potential to scale to larger datasets via its use of dynamic programming to solve the optimization problem in polynomial time within a constrained search space. In addition, future research should explore these methods compared to more computationally intensive methods such as InferNetwork_ML and InferNetwork_MPL (maximum likelihood and maximum pseudo-likelihood methods in PhyloNet [38,43]) restricted so that they produce trees rather than reticulate phylogenies, or PHYLDOG [7], a likehood-based method for co-estimating gene trees and the species tree under a GDL model.

6 Acknowledgments

This study was supported in part by NSF grants CCF-1535977 and 1513629 (to TW) and by the Ira and Debra Cohen Graduate Fellowship in Computer Sci-

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ence (to EKM). SR was supported by NSF grants DMS-1614242, CCF-1740707 (TRIPODS), and DMS-1902892, as well as a Simons Fellowship and a Vilas Associates Award. BL was supported by NSF grant DMS-1614242 (to SR). All computational analyses were performed on the Illinois Campus Cluster and the Blue Waters supercomputer, computing resources that are operated and financially supported by UIUC in conjunction with the National Center for Supercomputing Applications. Blue Waters is supported by the NSF (grants OCI-0725070 and ACI-1238993) and the state of Illinois.

References

1. Allman, E.S., Degnan, J.H., Rhodes, J.A.: Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. Journal of Mathematical Biology **62**(6), 833–862 (2011). https://doi.org/10.1007/s00285-010-0355-7
2. Arvestad, L., Lagergren, J., Sennblad, B.: The gene evolution model and computing its associated probabilities. Journal of the ACM **56**(2), 7 (2009). https://doi.org/10.1145/1502793.1502796
3. Bandelt, H.J., Dress, A.: Reconstructing the shape of a tree from observed dissimilarity data. Advances in Applied Mathematics **7**(3), 309–343 (1986). https://doi.org/10.1016/0196-8858(86)90038-2
4. Bansal, M.S., Burleigh, J.G., Eulenstein, O., Fernández-Baca, D.: Robinson-foulds supertrees. Algorithms for Molecular Biology **5**(1), 18 (2010). https://doi.org/10.1186/1748-7188-5-18
5. Bayzid, M.S., Warnow, T.: Gene tree parsimony for incomplete gene trees: addressing true biological loss. Algorithms for Molecular Biology **13**(1), 1 (2018). https://doi.org/10.1186/s13015-017-0120-1
6. Blom, M.P.K., Bragg, J.G., Potter, S., Moritz, C.: Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards. Syst. Biol. **66**(3), 352–366 (2017). https://doi.org/10.1093/sysbio/syw089
7. Boussau, B., Szöllősi, G.J., Duret, L., Gouy, M., Tannier, E., Daubin, V.: Genome-scale coestimation of species and gene trees. Genome Research **23**(2), 323–330 (2013). https://doi.org/10.1101/gr.141978.112
8. Chaudhary, R., Boussau, B., Burleigh, J.G., Fernández-Baca, D.: Assessing approaches for inferring species trees from multi-copy genes. Systematic Biology **64**(2), 325–339 (2015). https://doi.org/10.1093/sysbio/syu128
9. Chaudhary, R., Fernández-Baca, D., Burleigh, J.G.: MulRF: a software package for phylogenetic analysis using multi-copy gene trees. Bioinformatics **31**(3), 432–433 (2014). https://doi.org/10.1093/bioinformatics/btu648
10. Daskalakis, C., Roch, S.: Species trees from gene trees despite a high rate of lateral genetic transfer: A tight bound (extended abstract). In: Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 1621–1630 (2016). https://doi.org/10.1137/1.9781611974331.ch110
11. Davidson, R., Vachaspati, P., Mirarab, S., Warnow, T.: Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. BMC Genomics **16**(10), S1 (2015). https://doi.org/10.1186/1471-2164-16-S10-S1

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

12. Du, P., Hahn, M.W., Nakhleh, L.: Species tree inference under the multispecies coalescent on data with paralogs is accurate. bioRxiv (2019). https://doi.org/10.1101/498378
13. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research **32**(5), 1792–1797 (2004). https://doi.org/10.1093/nar/gkh340
14. Emms, D., Kelly, S.: STAG: Species tree inference from all genes. bioRxiv (2018). https://doi.org/10.1101/267914
15. Fletcher, W., Yang, Z.: INDELible: A Flexible Simulator of Biological Sequence Evolution. Molecular Biology and Evolution **26**(8), 1879–1888 (2009). https://doi.org/10.1093/molbev/msp098
16. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology **59**(3), 307–321 (2010). https://doi.org/10.1093/sysbio/syq010
17. Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T.: Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). Molecular Biology and Evolution **33**(4), 1110–1125 (2016). https://doi.org/10.1093/molbev/msv347
18. Jarvis, E.D., Mirarab, S., et al.: Whole-genome analyses resolve early branches in the tree of life of modern birds. Science **346**(6215), 1320–1331 (2014). https://doi.org/10.1126/science.1253451
19. Kingman, J.F.C.: The coalescent. Stochastic processes and their applications **13**(3), 235–248 (1982). https://doi.org/10.1016/0304-4149(82)90011-4
20. Larget, B.R., Kotha, S.K., Dewey, C.N., Ané, C.: BUCKy: Gene Tree/Species Tree Reconciliation with Bayesian Concordance Analysis. Bioinformatics **26**(22), 2910–2911 (2010). https://doi.org/10.1093/bioinformatics/btq539
21. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Molecular Biology and Evolution **32**(10), 2798–2800 (2015). https://doi.org/10.1093/molbev/msv150
22. Liu, L., Yu, L.: Estimating Species Trees from Unrooted Gene Trees. Systematic Biology **60**(5), 661–667 (2011). https://doi.org/10.1093/sysbio/syr027
23. Maddison, W.: Gene Trees in Species Trees. Systematic Biology **46**(3), 523–536 (1997). https://doi.org/10.1093/sysbio/46.3.523
24. Mallo, D., De Oliveira Martins, L., Posada, D.: SimPhy: Phylogenomic simulation of gene, locus, and species trees. Systematic Biology **65**(2), 334–344 (2016). https://doi.org/10.1093/sysbio/syv082
25. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T.: ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics **30**(17), i541–i548 (2014). https://doi.org/10.1093/bioinformatics/btu462
26. Mirarab, S.: DynaDup Github Repository: A software package for species tree estimation from rooted gene trees under gene duplication and loss, https://github.com/smirarab/DynaDup, last accessed October 3, 2019
27. Mirarab, S., Warnow, T.: ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics **31**(12), i44–i52 (2015). https://doi.org/10.1093/bioinformatics/btv234
28. Molloy, E.K., Warnow, T.: To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. Systematic Biology **67**(2), 285–303 (2018). https://doi.org/10.1093/sysbio/syx077

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

29. Rabiee, M., Sayyari, E., Mirarab, S.: Multi-allele species reconstruction using ASTRAL. Molecular Phylogenetics and Evolution **130**, 286–296 (2019). https://doi.org/10.1016/j.ympev.2018.10.033
30. Rasmussen, M.D., Kellis, M.: Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research **22**(4), 755–765 (2012). https://doi.org/10.1101/gr.123901.111
31. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1), 131–147 (1981). https://doi.org/10.1016/0025-5564(81)90043-2
32. Roch, S., Nute, M., Warnow, T.: Long-branch attraction in species tree estimation: Inconsistency of partitioned likelihood and topology-based summary methods. Systematic Biology **68**(2), 281–297 (2018). https://doi.org/10.1093/sysbio/syy061
33. Roch, S., Snir, S.: Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. Journal of Computational Biology **20**(2), 93–112 (2013). https://doi.org/10.1089/cmb.2012.0234
34. Roch, S., Steel, M.: Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theoretical Population Biology **100**, 56–62 (2015). https://doi.org/10.1016/j.tpb.2014.12.005
35. Stamatakis, A.: RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics **30**(9), 1312–1313 (2014). https://doi.org/10.1093/bioinformatics/btu033
36. Streicher, J.W., Schulte, II, J.A., Wiens, J.J.: How Should Genes and Taxa be Sampled for Phylogenomic Analyses with Missing Data? An Empirical Study in Iguanian Lizards. Systematic Biology **65**(1), 128–145 (2016). https://doi.org/10.1093/sysbio/syv058
37. Takahata, N.: Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics **122**(4), 957–966 (1989)
38. Than, C., Ruths, D., Nakhleh, L.: PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics **9**(1), 322 (2008). https://doi.org/10.1186/1471-2105-9-322
39. Vachaspati, P., Warnow, T.: ASTRID: Accurate Species Trees from Internode Distances. BMC Genomics **16**(10), S3 (2015), 10.1186/1471-2164-16-S10-S3
40. Vachaspati, P., Warnow, T.: ASTRID: Accurate Species TRees from Internode Distances. BMC Genomics **16**(10), S3 (2015). https://doi.org/10.1186/1471-2164-16-S10-S3
41. Vachaspati, P., Warnow, T.: FastRFS: fast and accurate Robinson-Foulds Supertrees using constrained exact optimization. Bioinformatics **33**(5), 631–639 (2016). https://doi.org/10.1093/bioinformatics/btw600
42. Wehe, A., Bansal, M.S., Burleigh, J.G., Eulenstein, O.: DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics **24**(13), 1540–1541 (2008). https://doi.org/10.1093/bioinformatics/btn230
43. Wen, D., Yu, Y., Zhu, J., Nakhleh, L.: Inferring phylogenetic networks using PhyloNet. Systematic Biology **67**(4), 735–740 (2018). https://doi.org/10.1093/sysbio/syy015
44. Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S.: ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics **19**(6), 153 (2018). https://doi.org/10.1186/s12859-018-2129-y

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A Additional proofs

A.1 Proof of Theorem 1: case (2)

In case (2), assume that $T^{\mathcal{Q}} = (((A, B), C), D)$, let R be the most recent common ancestor of A, B, C (but not D) in $T^{\mathcal{Q}}$, and let I be the number of gene copies exiting R. As in case 1), it suffices to prove (2) almost surely. Let $i_x \in \{1, ..., I\}$ be the ancestral lineage of $x \in \{a, b, c\}$ in R. Then

$$\mathbf{P}'_I[q = ab|cd] = \mathbf{P}'_I[i_a = i_b] + \mathbf{P}'_I[q = ab|cd \text{ and } i_a, i_b, i_c \text{ all distinct}]. \tag{9}$$

On the other hand,

$$\mathbf{P}'_I[q = ac|bd] = \mathbf{P}'_I[i_b \neq i_a = i_c] + \mathbf{P}'_I[q = ac|bd \text{ and } i_a, i_b, i_c \text{ all distinct}], \tag{10}$$

with a similar result for $\mathbf{P}'_I[q = ad|bc]$. By symmetry again, the last term on the RHS of (9) and (10) are the same. This implies

$$\begin{aligned}\mathbf{P}'_I[q = ab|cd] - \mathbf{P}'_I[q = ac|bd] &= \mathbf{P}'_I[i_a = i_b] - \mathbf{P}'_I[i_b \neq i_a = i_c] \\ = x - (1-x)\mathbf{P}'_I[i_a = i_c|i_a \neq i_b] &= x - (1-x)\frac{1}{I} \equiv g(x),\end{aligned}$$

where $x = \mathbf{P}'_I[i_a = i_b]$. This function g attains its minimum value at the smallest possible of x, which by Lemma 1 is $x = 1/I$. Evaluating at $x = 1/I$ gives

$$g(1/I) = \frac{1}{I} - \frac{1}{I} + \frac{1}{I^2} = \frac{1}{I^2} > 0,$$

which establishes (2) in case (2).

A.2 Proof of Theorem 2

First, we prove consistency for the exact version of ASTRAL. The input to the ASTRAL/ONE pipeline is the collection of gene trees $\mathcal{T} = \{t_i\}_{i=1}^k$, where t_i is labeled by individuals (i.e., gene copies) $R_i \subseteq R$. For each species and each gene tree t_i, we pick a uniform random gene copy, producing a new gene tree $\tilde{t}_i$. Recall that the quartet score of $\widetilde{T}$ with respect to $\widetilde{\mathcal{T}} = \{\tilde{t}_i\}_{i=1}^k$ is then

$$Q_k(\widetilde{T}) = \sum_{i=1}^{k} \sum_{\mathcal{J}=\{a,b,c,d\}\subseteq R_i} \mathbf{1}(\widetilde{T}^{\mathcal{J}}_{ext}, \tilde{t}_i^{\mathcal{J}}).$$

We note that the score only depends on the unrooted topology of $\widetilde{T}$. Under the GDL model, by independence of the gene trees (and non-negativity), $Q_k(\widetilde{T})/k$ converges almost surely to its expectation simultaneously for all unrooted species tree topologies over S.

For a species $A \in S$ and gene tree $\tilde{t}_i$, let A_i be the gene copy in A on $\tilde{t}_i$ if it exists and let $\mathcal{E}_i^A$ be the event that it exists. For a 4-tuple of species

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\mathcal{Q} = \{A, B, C, D\}$, let $\mathcal{Q}_i = \{A_i, B_i, C_i, D_i\}$ and $\mathcal{E}_i^{\mathcal{Q}} = \mathcal{E}_i^A \cap \mathcal{E}_i^B \cap \mathcal{E}_i^C \cap \mathcal{E}_i^D$. The expectation can then be written as

$$\mathbf{E}\left[\frac{1}{k} Q_k(\widetilde{T})\right] = \sum_{\mathcal{Q}=\{A,B,C,D\}} \mathbf{E}\left[\mathbf{1}(\widetilde{T}_{ext}^{\mathcal{Q}_1}, \tilde{t}_1^{\mathcal{Q}_1}) \,\middle|\, \mathcal{E}_1^{\mathcal{Q}}\right] \mathbf{P}[\mathcal{E}_1^{\mathcal{Q}}], \tag{11}$$

as, on the event $(\mathcal{E}_1^{\mathcal{Q}})^c$, there is no contribution from $\mathcal{Q}$ in the sum over the first sample.

Based on the proof of Theorem 1, a different way to write $\mathbf{E}[\mathbf{1}(\widetilde{T}_{ext}^{\mathcal{Q}_1}, \tilde{t}_1^{\mathcal{Q}_1}) \,|\, \mathcal{E}_1^{\mathcal{Q}}]$ is in terms of the original gene tree t_1. Let a, b, c, d be random gene copies on t_1 in A, B, C, D respectively. Then if q is the topology of t_1 restricted to a, b, c, d,

$$\mathbf{E}\left[\mathbf{1}(\widetilde{T}_{ext}^{\mathcal{Q}_1}, \tilde{t}_1^{\mathcal{Q}_1}) \,\middle|\, \mathcal{E}_1^{\mathcal{Q}}\right] = \mathbf{P}'[q = \widetilde{T}^{\mathcal{Q}}].$$

From (5), we know that this expression is maximized (strictly) at the true species tree $\mathbf{P}'[q = T^{\mathcal{Q}}]$. Hence, together with (11) and the law of large numbers, almost surely the quartet score is eventually maximized by the true species tree as $k \to +\infty$. This completes the proof for the exact version.

The default version is statistically consistent for the same reason as in the proof of Theorem 3. As the number of MUL-trees sampled tends to infinity, the true species tree will appear as one of the input gene trees almost surely. So ASTRAL returns the true species tree topology almost surely as the number of sampled MUL-trees increases.

A.3 Proof of Theorem 3: case (2)

In case (2), assume that $T^{\mathcal{Q}} = (((A, B), C), D)$ and let R be the most recent common ancestor of A, B, C (but not D) in T. We want to establish (8) in this case. For $i = 1, 2, 3$, let $\mathcal{N}_{AB|CD}^{\mathcal{Q},\{i\}}$ (respectively $\mathcal{N}_{AC|BD}^{\mathcal{Q},\{i\}}$) be the number of choices consisting of one gene copy from each species in $\mathcal{Q}$ whose corresponding restriction on $t^{\mathcal{Q}}$ agrees with $AB|CD$ (respectively $AC|BD$) and where, in addition, copies of A, B, C descend from i distinct lineages in R. We make five observations:

- Contributions to $\mathcal{N}_{AB|CD}^{\mathcal{Q},\{2\}}$ necessarily come from copies in A and B descending from the same lineage in R, together with a copy in C descending from a distinct lineage and any copy in D. Similarly for $\mathcal{N}_{AC|BD}^{\mathcal{Q},\{2\}}$
- Moreover $\mathcal{N}_{AC|BD}^{\mathcal{Q},\{1\}} = 0$ almost surely, as in that case the corresponding copies from A and B coalesce (backwards in time) below R.
- Arguing as in the proof of Theorem 1, by symmetry we have the equality $\mathbf{E}_I[\mathcal{N}_{AB|CD}^{\mathcal{Q},\{3\}}] = \mathbf{E}_I[\mathcal{N}_{AC|BD}^{\mathcal{Q},\{3\}}]$.
- For $j \in \{1, \ldots, I\}$, let $\mathcal{A}_j$ be the number of gene copies in A descending from j in R, and similarly define $\mathcal{B}_j$, $\mathcal{C}_j$. Let $\mathcal{D}$ be the number of gene copies in D. Then, under the conditional probability $\mathbf{P}_I$, $\mathcal{D}$ is independent of $(\mathcal{A}_j, \mathcal{B}_j, \mathcal{C}_j)_{j=1}^I$. Moreover, under $\mathbf{P}_I$, $(\mathcal{C}_j)_{j=1}^I$ is independent of $(\mathcal{A}_j, \mathcal{B}_j)_{j=1}^I$.

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

– Similarly to case 1), by symmetry we have $X^{=} \equiv \mathbf{E}_I[\mathcal{A}_{j_1}\mathcal{B}_{j_1}] = \mathbf{E}_I[\mathcal{A}_1\mathcal{B}_1]$, $X^{\neq} \equiv \mathbf{E}_I[\mathcal{A}_{j_1}\mathcal{B}_{k_1}] = \mathbf{E}_I[\mathcal{A}_1]\mathbf{E}_I[\mathcal{B}_1]$ for all j_1, k_1 with $j_1 \neq k_1$. Define also $X = IX^{=} + I(I-1)X^{\neq}$, $Y \equiv \mathbf{E}_I[\mathcal{C}_1]$ and $Z \equiv \mathbf{E}_I[\mathcal{D}]$.

Putting these observations together, we obtain

$$\begin{aligned}
&\mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AB|CD}] - \mathbf{E}_I[\mathcal{N}^{\mathcal{Q}}_{AC|BD}] \\
&\quad = \mathbf{E}_I[\mathcal{N}^{\mathcal{Q},\{1\}}_{AB|CD}] + \mathbf{E}_I[\mathcal{N}^{\mathcal{Q},\{2\}}_{AB|CD}] - \mathbf{E}_I[\mathcal{N}^{\mathcal{Q},\{2\}}_{AC|BD}] \\
&\quad = IX^{=}YZ + I(I-1)X^{=}YZ - I(I-1)X^{\neq}YZ \\
&\quad > 0,
\end{aligned}$$

where we used Lemma 2 on the last line.

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

B Details of Simulation Study

All scripts and datasets used in this study are available on the Illinois Data Bank: https://doi.org/10.13012/B2IDB-2626814_V1.

B.1 SimPhy Simulation

Our simulation protocol is based on the 16-taxon fungal dataset from Rasmussen and Kellis [30]. First, we used the species tree estimated by Rasmussen and Kellis [30] (download: http://compbio.mit.edu/dlcoal/pub/config/fungi.stree) as the model species tree, modifying the branch lengths to be ultrametric and in generations (assuming 10 generations per year). Below is the Newick string for the resulting model species tree (height: 1,800,000,337.5 generations).

(((((((scer : 70617600.0 , spar : 70617600.0) : 49996800.0 , smik : 120614400.0): 59706000.0 , sbay : 180320400.0) : 526823100.0 , cgla : 707143500.0) : 72206550.0 , scas : 779350050.0) : 231815475.0 , ((agos : 785532600.0 , klac : 785532600.0) : 104349600.0 , kwal : 889882200.0) : 121283325.0) : 788834812.5 , (((calb : 412758000.0 , ctro : 412758000.0) : 296329500.0 , (cpar:523231200.0 , lelo : 523231200.0) : 185856300.0) : 311495850.0 , ((cgui : 756158400.0 , dhan : 756158400.0) : 140068800.0 , clus : 896227200.0) : 124356150.0) : 779416987.5);

We then simulated gene trees from this model species tree under the DLCoal model [30] with three different GDL rates using SimPhy Version 1.0.2 with the command:

```
simphy-1.0.2-mac64 -rs $nreps -rl F:ngens -rg 1 -s $stree \
    -si F:1 -sp F:$psize -su F:$mrate -sg F:10 -lb F:$drate \
    -ld F:lb -hg LN:1.5,1 -o <output directory> -ot 0 -om 1 \
    -od 1 -op 1 -oc 1 -ol 1 -v 3 -cs 293745 &> <log file>
```

where `$nreps` is the number of replicate datasets (10), `$ngens` is the number of genes trees per dataset (1000), `$stree` is the model species tree (Newick string above), `$psize` is the effective population size (1×10^7), `$mrate` is the tree-wide substitution rate (4×10^{-10} substitutions per generation per site), `$drate` is the duplication rate (either 1×10^{-10}, 2×10^{-10}, or 5×10^{-10} duplication events per generation), and the loss rate always equals the duplication rate.

These parameters (with the GDL rate of 1×10^{-10}) are similar to those estimated from the fungal dataset by Rasmussen and Kellis [30] and are the same parameters used in the simulation study by Du *et al.* [12]. Unlike in the simulation study by Du *et al.* [12], we did not enable gene conversion and allowed gene trees to deviate from the strict molecular clock by using the gene-by-lineage-specific rate heterogeneity modifiers (`-hg`). This means that a gamma distribution was defined for each gene tree by drawing α from a log-normal distribution with a location of 1.5 and a scale of 1 (same parameters as [44]), and then each branch in a gene tree was multiplied by a value drawn the gamma distribution corresponding to that gene tree.

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In summary, there were three model conditions, characterized by the three GDL rates; each of these model conditions had 10 replicate datasets, and each of these replicate datasets had 1000 gene trees. For each model gene tree, we simulated a multiple sequence alignment (1000 base pairs) using INDELible Version 1.03 with GTR+GAMMA model parameters drawn from distributions; specifically, GTR base frequencies (A, C, G, T) were drawn from Dirchlet(113.48869, 69.02545, 78.66144, 99.83793), GTR substitution rates (AC, AG, AT, CG, CT, GT) were drawn from Dirchlet(12.776722, 20.869581, 5.647810 9.863668, 30.679899, 3.199725), and α was drawn from LogNormal(-0.470703916, 0.348667224), where the first parameter is the meanlog and the second parameter is the sdlog.

These distributions were based on the fungal dataset from Rasmussen and Kellis [30] (download: http://compbio.mit.edu/dlcoal/pub/data/real-fungi.tar.gz), which included a multiple sequence alignment estimated using MUSCLE [13] and a maximum likelihood tree estimated using PhyML [16] for each of the 5,351 genes. We estimated GTR+GAMMA model parameters using RAxML Version 8.2.12 with the command:

```
raxmlHPC-SSE3 -m GTRGAMMA -f e -t <PhyML gene tree file> \
    -s <MUSCLE alignment file> -n <output name>
```

We then fit distributions to the parameters estimated from alignments with at least 500 distinct alignment patterns and at most 25% gaps.

Table 1. SimPhy Simulation Summary. All values shown below are averages ($\pm$ standard deviations) across the 10 replicate datasets for each of the three GDL rates. We quantify the level of ILS as the normalized RF distance between each true locus tree and its respective true gene tree (which are on the same leaf set), averaged across all 1000 locus/gene trees. Because these values are all less than 1%, *there is effectively no ILS in these datasets*. We also report the number of taxa per gene tree as well as the number of leaves per gene tree, both averaged across all 1000 gene trees. Because the gene duplication and gene loss rates were equal, the number of leaves per locus/gene tree was close to the number of leaves in the species tree. As the GDL rate increased, the number of species per locus/gene tree decreased, and thus, even though locus/gene trees had the same number of leaves on average, these leaves were labeled by fewer species as the GDL rate increased.

GDL Rate	ILS Level	Mean # of taxa per gene tree	Mean # of leaves per gene tree
1×10^{-10}	0.20% $\pm$ 0.03%	13.61 $\pm$ 0.09	16.09 $\pm$ 0.14
2×10^{-10}	0.37% $\pm$ 0.07%	12.02 $\pm$ 0.12	16.36 $\pm$ 0.14
5×10^{-10}	0.68% $\pm$ 0.14%	9.20 $\pm$ 0.10	17.42 $\pm$ 0.25

B.2 Gene Tree Estimation

On gene trees with four or more species, we estimated gene trees using RAxML Version 8.2.12 with the command:

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


```
raxmlHPC-SSE3 -m GTRGAMMA -p <random seed> -n <output name> \
    -s <alignment file>
```

We truncated sequences to the first 25, 50, 100, and 250 base pairs in order to produce datasets with varying levels of gene trees estimation error (GTEE). Sequence lengths of 25, 50, 100, and 250 resulted in mean GTEE of 65%, 53%, 39%, 23% respectively. Mean GTEE was measured as the normalized RF distance between true and estimated gene trees, averaged across all gene trees.

B.3 Species Tree Estimation

We estimated species trees using either the first 25, 50, 100, or 500 (true or estimated) gene trees. ASTRAL Version 5.6.3 was run with the command:

```
java -Xms2000M -Xmx20000M -jar astral.5.6.3.jar -i <gene tree file> \
    -a <name map file> -o <output file> &> <log file>
```

ASTRID Version 2.2.1 was run with the command:

```
./ASTRID -u -s -i <gene tree file> -a <name map file> \
    -o <output file> &> <log file>
```

DupTree (download: http://genome.cs.iastate.edu/CBL/DupTree/linux-i386.tar.gz) was run with the command:

```
./duptree -i <gene tree file> -o <output file> &> <log file>
```

MulRF Version 2.1 was run with the command:

```
./MulRFSupertreeLin -i <gene tree file> -o <output file> &> <log file>
```

STAG (download: https://github.com/davidemms/STAG) was run with the command:

```
python stag.py <name map file> <gene tree folder> &> <log file>
```

We ran STAG with FastME Version 2.1.5 [21]. Importantly, STAG only uses gene trees that include at least one copy of every species. When the level of GDL was high (i.e., 5×10^{-10}), STAG failed to run on 3/10 replicates with 25 genes and 2/10 replicates on 50 genes, because none of the input gene trees included at least one copy of every species; we do not show results using STAG for those model conditions.

C Additional Results of Simulation Study

bioRxiv preprint doi: https://doi.org/10.1101/821439; this version posted January 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2. For each model condition, species tree error (mean ± standard deviation across 10 replicate datasets) is shown for five different methods. Species tree error was measured as the normalized RF distance between true and estimated species trees.

# of genes	ASTRAL-multi	MulRF	DupTree	STAG	ASTRID
GDL Rate: 1×10^{-10}, *Mean GTEE: 65%*					
25	0.28 ± 0.06	0.22 ± 0.10	0.49 ± 0.19	0.39 ± 0.11	0.21 ± 0.07
50	0.18 ± 0.10	0.12 ± 0.07	0.45 ± 0.11	0.24 ± 0.07	0.14 ± 0.08
100	0.12 ± 0.06	0.06 ± 0.06	0.25 ± 0.11	0.15 ± 0.10	0.05 ± 0.05
500	0.01 ± 0.02	0.00 ± 0.00	0.08 ± 0.10	0.02 ± 0.05	0.00 ± 0.00
GDL Rate: 1×10^{-10}, *Mean GTEE: 39%*					
25	0.07 ± 0.04	0.03 ± 0.04	0.11 ± 0.12	0.13 ± 0.08	0.06 ± 0.06
50	0.02 ± 0.04	0.02 ± 0.04	0.09 ± 0.10	0.06 ± 0.05	0.03 ± 0.04
100	0.01 ± 0.02	0.01 ± 0.02	0.02 ± 0.05	0.02 ± 0.04	0.01 ± 0.02
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GDL Rate: 1×10^{-10}, *Mean GTEE: 23%*					
25	0.01 ± 0.02	0.01 ± 0.02	0.02 ± 0.03	0.06 ± 0.06	0.00 ± 0.00
50	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.02	0.01 ± 0.02	0.00 ± 0.00
100	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.02	0.00 ± 0.00
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GDL Rate: 1×10^{-10}, *Mean GTEE: 0%*					
25	0.01 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
50	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
100	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GDL Rate: 2×10^{-10}, *Mean GTEE: 65%*					
25	0.31 ± 0.14	0.28 ± 0.10	0.53 ± 0.19	0.56 ± 0.15	0.14 ± 0.10
50	0.23 ± 0.10	0.21 ± 0.11	0.39 ± 0.16	0.45 ± 0.12	0.08 ± 0.07
100	0.15 ± 0.12	0.11 ± 0.09	0.33 ± 0.16	0.26 ± 0.10	0.05 ± 0.08
500	0.05 ± 0.05	0.00 ± 0.00	0.13 ± 0.09	0.05 ± 0.06	0.01 ± 0.02
GDL Rate: 2×10^{-10}, *Mean GTEE: 39%*					
25	0.08 ± 0.08	0.12 ± 0.09	0.11 ± 0.05	0.29 ± 0.15	0.08 ± 0.07
50	0.05 ± 0.06	0.03 ± 0.05	0.05 ± 0.04	0.15 ± 0.14	0.02 ± 0.05
100	0.01 ± 0.02	0.00 ± 0.00	0.03 ± 0.05	0.09 ± 0.08	0.01 ± 0.02
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GDL Rate: 2×10^{-10}, *Mean GTEE: 23%*					
25	0.02 ± 0.04	0.02 ± 0.04	0.04 ± 0.05	0.11 ± 0.08	0.02 ± 0.03
50	0.02 ± 0.05	0.00 ± 0.00	0.01 ± 0.02	0.06 ± 0.05	0.00 ± 0.00
100	0.01 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.02	0.00 ± 0.00
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GDL Rate: 2×10^{-10}, *Mean GTEE: 0%*					
25	0.01 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
50	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
100	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
GDL Rate: 5×10^{-10}, *Mean GTEE: 65%*					
25	0.48 ± 0.13	0.36 ± 0.12	0.65 ± 0.09	$nan \pm nan$	0.34 ± 0.12
50	0.35 ± 0.13	0.30 ± 0.10	0.48 ± 0.24	$nan \pm nan$	0.19 ± 0.15
100	0.28 ± 0.15	0.18 ± 0.08	0.36 ± 0.23	0.52 ± 0.11	0.09 ± 0.08
500	0.12 ± 0.09	0.04 ± 0.04	0.20 ± 0.10	0.22 ± 0.10	0.00 ± 0.00
GDL Rate: 5×10^{-10}, *Mean GTEE: 39%*					
25	0.28 ± 0.09	0.25 ± 0.08	0.16 ± 0.12	$nan \pm nan$	0.14 ± 0.07
50	0.15 ± 0.10	0.09 ± 0.07	0.06 ± 0.07	$nan \pm nan$	0.05 ± 0.06
100	0.10 ± 0.05	0.05 ± 0.05	0.03 ± 0.05	0.28 ± 0.10	0.02 ± 0.03
500	0.01 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.06 ± 0.07	0.00 ± 0.00
GDL Rate: 5×10^{-10}, *Mean GTEE: 23%*					
25	0.16 ± 0.12	0.05 ± 0.04	0.05 ± 0.09	$nan \pm nan$	0.02 ± 0.03
50	0.08 ± 0.06	0.05 ± 0.05	0.01 ± 0.02	$nan \pm nan$	0.01 ± 0.02
100	0.02 ± 0.04	0.03 ± 0.05	0.00 ± 0.00	0.17 ± 0.11	0.00 ± 0.00
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.03	0.00 ± 0.00
GDL Rate: 5×10^{-10}, *Mean GTEE: 0%*					
25	0.05 ± 0.09	0.00 ± 0.00	0.00 ± 0.00	$nan \pm nan$	0.01 ± 0.02
50	0.05 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	$nan \pm nan$	0.00 ± 0.00
100	0.02 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.05 ± 0.08	0.00 ± 0.00
500	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00