On Partial Gene Transfer and its Impact on Gene Tree Reconstruction

Sumaira Zaman[1] and Mukul S. Bansal[1,2]

[1] Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269, USA
[2] Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269, USA
sumaira.zaman@uconn.edu, mukul.bansal@uconn.edu

Abstract. Horizontal transfer of genetic material between different organisms is one of the most important evolutionary processes in microbial evolution. Such horizontal transfer events can result in the transfer of genomic fragments containing multiple complete genes, complete single genes, or partial genes. However, partial gene transfer (PGT) remains poorly understood and generally underappreciated. Indeed, existing phylogenetic approaches for studying microbial evolution and horizontal gene transfer largely ignore PGT, leading to potential biases and errors in downstream inferences.
In this work, we (i) perform a systematic study of the impact of PGT on the ability to correctly reconstruct the evolutionary histories of gene families (i.e., gene trees) and (ii) propose a simple, yet effective approach, called *trippd*, to detect if a given gene family has been affected by PGT. Our analysis, using simulated and real biological datasets, reveals many interesting insights related to when and how PGT affects gene tree reconstruction, demonstrates the utility of trippd, and sheds light on the importance of detecting and accounting for PGT when studying microbial evolution.

1 Introduction

Horizontal gene transfer (HGT) is known to play an important role in microbial evolution and many different computational techniques have been developed to infer HGTs; see, e.g., [29] for a review. While most methods for inferring and studying HGT view single genes as the "unit" of HGT, it is well known that multiple genes can be transferred in a single transfer event [5, 10, 14, 24] and that many transfers result in the transfer of only partial genes (i.e., fraction of a gene) [3, 6, 7, 34, 36]. Partial gene transfer (PGT), in particular, remains poorly understood and existing phylogenetic approaches for studying microbial evolution and horizontal gene transfer largely ignore PGT. Such PGTs can occur not only when the transferred genomic fragments themselves are small but also when boundaries of larger genomic fragments containing one or more complete genes overlap flanking genes. Moreover, integration of new genetic material into a genome often occurs through homologous recombination in flanking regions [25].

While many approaches have been developed for studying recombination in genomes, e.g., [11, 19–21, 35], such methods have been observed to have high false-positive rates for breakpoint detection [2], decreasing their utility for PGT detection. To our knowledge, the two approaches most directly applicable to the problem of detecting PGTs within gene families are T-REX [3] and PhyML-Multi [4]. T-REX [3] uses a sliding window technique and infers PGT by constructing window trees and comparing them to a known species tree to infer possible transfer events. However, T-REX assumes all discordance is due to replacing transfer (or homologous recombination) and cannot be directly applied to gene families with a history of gene duplication or additive transfer. PhyML-Multi [4] uses a more sophisticated HMM based approach and can partition the given gene family alignment into a user-specified number of partitions with distinct evolutionary histories. Notably, PhyML-Multi does not rely on a known species tree or on inference of actual transfer events, both of which are known to be error-prone [1, 12, 18], and can therefore be directly applied to any gene family alignment to detect possible PGT. However, the utility of PhyML-Multi for PGT detection has not been sufficiently explored and its effectiveness for this problem has not been previously studied. Furthermore, the impact of PGT on gene tree reconstruction itself remains poorly understood and generally underappreciated. Previously, Posada and Crandall [26] systematically evaluated the impact of recombination on phylogeny inference. However, that work did not focus directly on PGTs and used small, 8-taxon trees with only a single recombination event per tree.

In this work, we advance the study of PGT and gene family evolution by (i) performing a systematic assessment of the impact of PGT on gene tree reconstruction, (ii) evaluating the ability of PhyML-Multi to accurately detect PGTs, and (iii) proposing a conceptually simple and easy-to-use approach, called *trippd*, based on alignment tri-partitioning, to identify gene families affected by non-negligible PGT. Among many interesting findings, we demonstrate that PGTs can significantly impact gene tree reconstruction and identify the scenarios under which PGTs may or may not significantly affect gene tree reconstruction accuracy; despite considerable conceptual and methodological differences, some of these findings are also consistent with previous results from [26]. Our evaluation of PhyML-Multi as the basis for PGT detection reveals that such an approach has a very high false-positive rate of PGT detection. At the same time, our experimental analysis shows how our new approach, trippd, can help address this limitation of PhyML-Multi, achieving a false-negative rate comparable to that of the PhyML-Multi based approach while having a negligible false-positive rate. An application of trippd to two biological datasets demonstrates the prevalence of PGT in real gene families.

Overall, this work sheds fresh light on the importance of detecting PGTs and accounting for them in microbial evolutionary analyses, reveals new insights into when and how gene tree reconstruction is impacted by PGT, and proposes a simple approach that can help end-users easily identify gene families affected by sufficient PGT to impact gene tree reconstruction. Scripts im-

plementing trippd, along with all simulated datasets, are freely available from `https://github.com/suz11001/Tripartition`.

2 Materials and Methods

We use an extensive simulation study to assess the impact of PGT on gene tree reconstruction accuracy and to evaluate the effectiveness of the PhyML-Multi based approach and of our proposed PGT detection approach trippd.

2.1 Simulated datasets

We used the phylogenetic simulation framework SaGePhy [17] to generate a large collection of simulated datasets consisting of gene families affected by PGT. Each gene family is represented by a gene family alignment, where the alignment is composed of a *genic-region*, consisting of sequences evolved down a gene tree, and a *PGT-region* consisting of sequences evolved down the same gene tree but with a certain rate of replacing transfer (homologous recombination). In other words, each gene family alignment represents two or more distinct evolutionary histories, appended together, with one representing the evolution of the gene tree and the other(s) representing the evolutionary history of a gene sequence region (or locus) affected by PGT. The resulting gene family datasets represent a wide range of evolutionary conditions, with varying gene lengths, PGT-region to genic-region ratios (i.e, fraction of gene sequence affected by PGTs), rates of PGT, sequence evolution rate, etc. We divide these datasets into three broad categories: *baseline* datasets, *multi-PGT* datasets, and *PGT-location* datasets. We describe the construction of these datasets below:

Baseline datasets. Our baseline collection consists of 14 distinct datasets, each representing a distinct combination of evolutionary parameter settings and consisting of 100 gene families generated under the corresponding parameter settings. To generate each dataset, we first simulated 100 species trees with 100 leaves each using a birth-death process and then simulated a gene tree inside each species tree using specific rates of gene duplication, replacing HGT, additive HGT, and gene loss. (The exact parameter values used for simulating species trees and gene trees, along with all simulated data, are freely available from the GitHub page linked above.) This yields 100 gene trees per dataset, and this same set of 100 gene trees was used for simulating the 100 gene families in each dataset. These gene trees had between between 20 and 236 leaves, with an average of 98.2, and each gene tree was of height 1. To simulate PGTs within each gene tree, we used SagePhy to simulate 3 different "subgene" trees, each with a different rate of PGT, within each of the 100 gene trees. Each subgene tree represents a history of PGT via homologous recombination within the corresponding gene tree. Specifically, each subgene tree was evolved down the gene tree under a certain rate of replacing subgene transfer and no other events. We used replacing transfer rates of 0.2, 0.4, and 0.6 (per unit branch length) to simulate low, medium, and high rates of partial gene transfer, resulting in 3 subgene trees per

gene tree, each with the same height and number of leaves as the corresponding gene tree. These three resulting sets of subgene trees correspond to, on average, 2.03, 3.87, and 5.55 PGTs per gene family, respectively.

The resulting set of 100 gene trees and 300 subgene trees was then used to simulate sequences under different evolutionary scenarios, resulting in the 14 baseline simulated gene family datasets. For these datasets, only one PGT-region is included in each gene family alignment and this PGT-region is always appended at the end of the genic-region. To generate the 14 baseline datasets, we considered the three PGT evolution rates (0.2, 0.4, and 0.6) as discussed above and, in addition, varied the following sequence-related parameters: (i) total sequence length (500, 1000, and 2000nt; for reference, the average prokaryotic gene length is roughly 1000nt [15]), (ii) substitution rates (0.1, 0.5, 1, 2, and 5 substitutions per site per unit branch length, capturing a wide range of evolutionary distances from closely related to distantly related), and (iii) fraction of sequence length represented by PGT-region (10%, 20%, 30%, 40%, 50% and 60%). We created one dataset with default parameter values of 0.4 for PGT evolution rate, 1000nt for sequence length, 0.5 for substitution rate, and 40% for fraction of sequence length represented by PGT-region. To study the impact of different parameters on gene tree reconstruction and PGT detection, we generated additional datasets by varying one parameter value at a time and keeping other parameters at their default values. This resulted in $2+2+4+5=13$ additional datasets, yielding a total of 14 baseline datasets. All sequences were generated using Seq-Gen [27] under the GTR model with gamma distributed rates and default settings for other Seq-Gen parameters.

Multi-PGT datasets. To assess how gene tree reconstruction is impacted by the presence of *multiple* PGT-regions within the same gene family, we created 4 additional datasets, each containing 2 PGT-regions. Specifically, we used default values for PGT evolution rate, total sequence length, and substitution rates, but varied the fraction of sequence length represented by PGT-regions as well as the specific fractions corresponding to each of the two PGT-regions. The 5 Multi-PGT datasets correspond to the following splits of PGT-region length between the two PGT-regions: {20%, 20%}, {30%, 10%}, {40%, 20%}, {60%, 10%}.

PGT-location datasets. To further assess the impact of PGT-region location within gene family alignments, we created 3 additional datasets corresponding to offsets of 34 base pairs (bps), 84 base pairs, and 134 base pairs from the end of the sequence alignment. These datasets otherwise use default parameter settings for all parameters. This small number of PGT-location datasets is sufficient to assess the impact of PGT-region location (Section 4.2).

2.2 Biological datasets

To assess the prevalence of PGTs in real microbial gene families, we used samples from two large published biological datasets: a dataset consisting of over 4700 gene families from 100 broadly sampled species (11 eukaryotic, 12 archaeal, and 67 bacterial) [8], and a dataset of 8,277 gene families from 103 *Aeromonas*

strains [14, 28]. The first dataset represents a scenario where, due to the great evolutionary divergence between included taxa, we do not expect to see much PGT. In contrast, the second dataset represents a scenario where the taxa under consideration are closely related and so a high prevalence of PGT is expected due to the ease of homologous recombination.

For each dataset, we first filtered the collection of gene families present in each original dataset by removing all gene families that had fewer than 40 genes or alignments shorter than 150 amino acids or 450nt. After applying this filtering we were left with 823 gene families for the 100-taxon dataset and 3,357 gene families for the 103-taxon *Aeromonas* dataset. Since the *Aeromonas* dataset is quite large, we randomly sampled 500 gene families from the remaining 3,357. During subsequent analysis of the resulting datasets, we found that some gene families had very large gaps (greater than one-third of the total alignment length) in the alignment of one or more sequences. We therefore removed all gene families with such large gaps, leaving us with 784 and 466 gene families for the 100-taxon dataset and 103-taxon *Aeromonas* dataset, respectively.

2.3 Gene tree construction and comparison

To study the impact of PGTs on gene tree reconstruction accuracy, we compared the topologies of the three main tree types for each gene family in each dataset: The *true gene tree* for that gene family (as simulated using SaGePhy), the *pre-PGT gene tree* reconstructed using the genic-region of the corresponding sequence alignment, and the *post-PGT gene tree* reconstructed using the full sequence alignment (appended genic- and PGT-regions). A pre-PGT gene tree represents the best tree we could reasonably reconstruct given only the sequence alignment and knowledge of the presence of PGTs in that gene family. A post-PGT gene tree represents the tree we would reconstruct if we were unaware of the presence of PGTs in that gene family.

All pre-PGT and post-PGT gene trees were reconstructed using RAxML v8.2.11 [32] (with 100 rapid bootstrap samples (-f a -N 100) and under the GTRCAT model). Divergence between any pair of (unrooted) gene tree topologies was measured using Robinson-Fould's distance [30]. Specifically, we count the number of splits present in only one of the two trees being compared. We refer to the resulting number as the RF-score and use $RF(T_1, T_2)$ to denote the RF-score between trees T_1 and T_2. Note that the RF-score counts unique splits of *both* trees (i.e., we do not divide the computed score by 2).

2.4 Using PhyML-Multi to detect PGTs

PhyML-Multi [4] is an existing state-of-the-art approach designed to identify plausible recombination breakpoints within a given sequence alignment and to reconstruct phylogenetic trees for each identified recombination-free region. Next, we briefly describe how PhyML-Multi can be used to detect PGTs. Our new approach, trippd, is introduced later in Section 3.

For our analysis, we used the more rigorous HMM-based implementation of PhyML-Multi and used suitable parameter settings expected to maximize inference accuracy. Specifically, we specified the number of expected partitions/trees to be 2 (which is the correct expected number for all baseline datasets), used the TN93 model of evolution (the closest one to GTR, since GTR is not available within PhyML-Multi), used 4 rate categories, allowed PhyML-Multi to estimate the transition/transversion ratio, proportion of invariable sites, and gamma shape parameter, and used BIONJ to build starting trees (instead of providing user-specified starting trees).

The output from PhyML-Multi includes breakpoints for the number of specified partitions along with the PhyML maximum likelihood (ML) tree corresponding to each partition. For a fair comparison with trippd, we ignored the output PhyML trees and instead used the breakpoints/partitions identified by PhyML-Multi to generate the corresponding RAxML tree for each partition using the same RAxML parameter settings as described above.

Note that PhyML-Multi will always find the specified number of partitions (and trees) for the given sequence alignment, even if no homologous recombination has occurred. Thus, PhyML-Multi cannot be directly used to determine if PGT has occurred. We therefore use a simple histogram intersection test to determine if any phylogenetic differences for the sequence partitions identified by PhyML-Multi may, in fact, be due to PGT. We describe this test below. A similar test is also used as part of trippd.

Histogram intersection test for PGT presence and absence. Given the two partitions of a gene family alignment output by PhyML-Multi, we employ a simple classification procedure to determine if any inferred phylogenetic differences between the two partitions are likely to have been caused by PGT. As part of this test, we compute 100 bootstrap replicates for each of the two partitions using RAxML (under GTRCAT, as above). Let A and B denote the two partitions and $\{A_1, ..., A_{100}\}$ and $\{B_1, ..., B_{100}\}$ denote the corresponding bootstrap replicate trees, respectively. We also compute a maximum likelihood tree (using RAxML) for the full, unpartitioned sequence alignment for that gene family. Let $\mathcal{R}$ denote this maximum likelihood tree.

We then compute the RF-scores between each bootstrap replicate A_i and $\mathcal{R}$, and between each bootstrap replicate B_i and $\mathcal{R}$, for each $i \in \{1, \ldots, 100\}$. This generates two discrete distributions of 100 RF-scores for the two partitions. The classification is based on the histogram intersection of these two distributions: If the intersection is less than a certain threshold, fixed at 50% in our experiments, then the phylogenetic difference between partitions A and B is assumed to be due to PGT, and otherwise assumed to be due to inference uncertainty or other random effects. The key idea is that if both partitions are a result of the same evolutionary process, i.e., no PGT, then the differences between the bootstrap trees for each partition and the overall ML tree should be similar for the two partitions. An illustration appears in Figure 1A.

3 trippd: tri-partition based PGT detection

In our experiments, we found that PhyML-Multi showed a high false positive rate for identifying gene families affected by PGT (Section 4.2). We therefore devised a simple, proof-of-concept approach that, in our experiments, nearly matches the accuracy of PhyML-Multi in correctly detecting PGT (i.e., has low false negative rate) while also achieving a very low false positive rate. Our new approach, called trippd (short for tri-partition based PGT detection, and pronounced "tripped") has three key features: (i) unlike PhyML-Multi, it does not rely on breakpoint detection and is therefore robust to errors in detecting the breakpoints/partitions correctly, (ii) it does not require any advance knowledge of the number of partitions or PGT-regions, and (iii) it leverages insights from our experimental evaluation of the impact of PGTs on gene tree reconstruction and is especially designed to classify gene family alignments as those having *sufficient* or *insufficient* PGT to impact gene tree reconstruction. trippd is illustrated in Figure 1B, and a step-by-step description of trippd follows:

Alignment tri-partitioning. The given gene family alignment is partitioned into three equal (or roughly equal) parts, each consisting of one-third of the sites in the alignment. We refer to these partitions as window-1, window-2, and window-3.

ML window tree reconstruction. Use RAxML to compute a maximum likelihood tree for each of the three windows.

Identifying most similar and most dissimilar pairs of windows. Compute the RF-score between each pair of ML window trees. Identify the pairs with smallest RF-score, denoted ww_{min}, and largest RF-score, denoted ww_{max}. Note that if $ww_{min} = ww_{max}$ then subsequent steps need not be executed and PGT is assumed to be absent.

Bootstrap replicates for each window. Compute several (100 in our experiments) bootstrap replicates for each of the three windows using RAxML. Denote these as $\{w_1^i, ..., w_b^i\}$ for window-i, where $1 \leq i \leq 3$ and b denotes the number of bootstrap replicates per window.

Computing distributions of RF-scores. Given the bootstrap replicates for any two windows i and j, define $D(i, j)$, to be the distribution of RF-scores $RF(w_k^i, w_k^j)$, where $k \in \{1, \ldots, b\}$. Compute $D(ww_{min})$ and $D(ww_{max})$.

Histogram intersection test. Apply a simple test (similar to the one for PhyML-Multi described in Section 2.4) to determine if differences between $D(ww_{min})$ and $D(ww_{max})$ are likely due to PGT or not. Specifically, compute the histogram intersection of $D(ww_{min})$ and $D(ww_{max})$ and check if the intersection is no more than a certain threshold, fixed at 50% in our experiments. An intersection smaller than or equal to the threshold percentage indicates presence of PGT. Intersection greater than the threshold indicates lack of significant PGT.

Our choice of using only three static windows in trippd is based on several observations and considerations: For example, we find in our experiments (see

Results) that a PGT-region that spans less than a third of the total sequence does not have measureable impact on gene tree reconstruction. At the same time, any PGT-region longer than a third of the total sequence length would overlap significantly with at least one of the three static windows, impacting at least one of the window trees. Having three windows, rather than just two, also allows for multiple pairwise window comparisons. The three-window approach is also relatively robust to the size of the PGT-region, allowing for the PGT-region to dominate and the genic-region to be relatively short, as long as the genic-region still makes up a majority of at least one of the three windows. Finally, this approach is also relatively robust to the presence of multiple PGT-regions, as long as there are at least two windows in which either the genic-region or one of the PGT-regions constitutes the majority of the sequence.

Selecting histogram intersection test threshold. We used a simple simulation study to determine a reasonable (not optimized) threshold for the histogram intersection test. Specifically, we used the baseline dataset with default parameter values to measure false-positive and false-negative inference at thresholds of 0% (i.e., complete separation between the two distributions), 25% and 50%. At the very strict threshold of 0%, we observed a very high false negative rate of about 0.5 and no false positives. At 25% the false negative rate improved only slightly. At 50% we observed a large reduction in the false negative rate, while still observing no false positives. We therefore fixed the threshold at 50%. We did not further optimize this threshold to maintain robustness to varying evolutionary conditions. An evaluation of its robustness appears in Section 4.

4 Results

4.1 Impact of PGT on gene tree reconstruction accuracy

We first assessed the impact of PGT on gene tree reconstruction using the baseline and multi-PGT simulated datasets described earlier. Recall that the 14 baseline datasets encompass a wide range of evolutionary scenarios, allowing for an assessment of the impact of PGT rate, total sequence length, sequence evolution (substitution) rate, and ratio of PGT- to genic-region length. In addition, the 4 multi-PGT datasets make it possible to assess the impact of multiple distinct PGT-regions within the same gene family alignment.

For each gene family within each dataset, we reconstruct two gene trees by applying RAxML to the simulated sequence data: A *pre-PGT* gene tree reconstructed using only the geneic-region of the sequence, and a *post-PGT* gene tree reconstructed using the entire sequence alignment (consisting of both the genic- and PGT-regions). Thus, a *pre-PGT* gene tree represents the best tree we can reasonably reconstruct if all PGTs were correct detected and accounted for, while a *post-PGT* gene tree represents the gene tree we would normally reconstruct if we do not account for possible PGT. The results of our analysis are shown in Figure 2, where we plot the average RF-scores between each pre-PGT gene tree and true (simulated) gene tree and between each post-PGT gene tree and the true gene tree, for each dataset. We describe these results below.

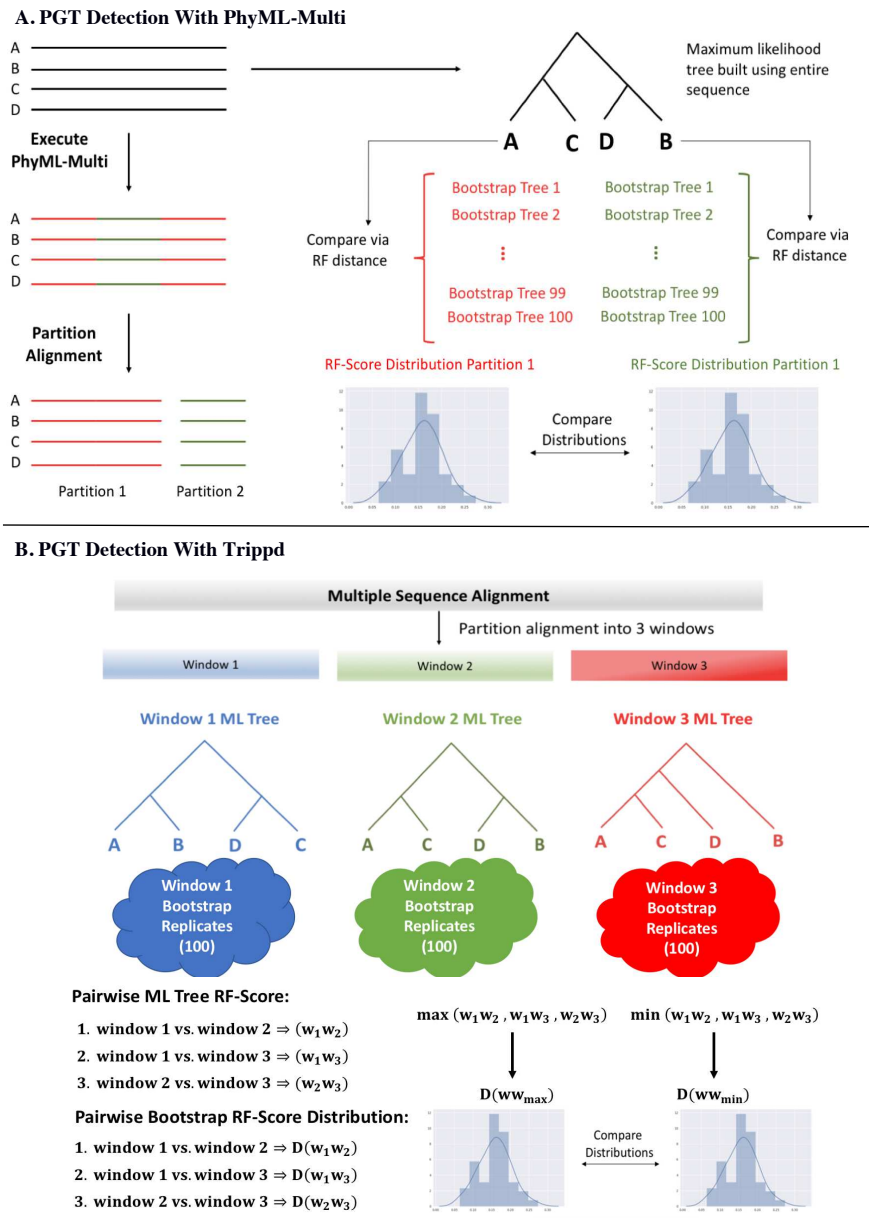

Fig. 1. Overview of PhyML-Multi and trippd for PGT detection. Both approaches start with a given multiple sequence alignment for the gene family. (A) The PhyML-Multi based approach works by using PhyML-Multi to partition the alignment into two regions, using RAxML to compute multiple bootstrap replicates for the two regions, comparing the resulting trees to the maximum likelihood (ML) tree for the entire sequence alignment, and using a simple histogram intersection test to determine if the resulting distributions of RF-scores suggest different evolutionary histories for the two regions. (B) tripped executes the step-by-step approach described in Section 3.

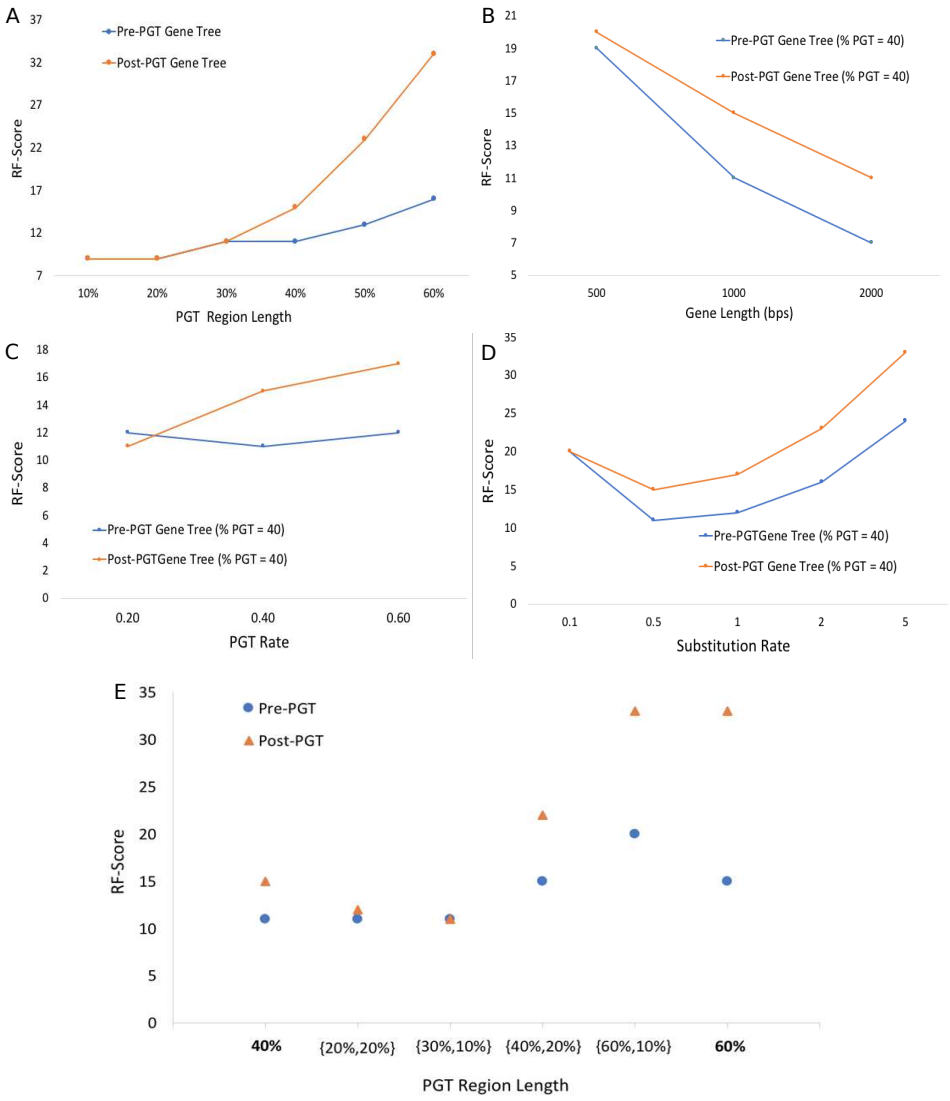

Fig. 2. Impact of PGT on gene tree reconstruction accuracy. The plots show the impact of various evolutionary parameters on the reconstruction error of pre-PGT (blue) and post-PGT (orange) gene trees. (A) shows the impact of PGT region length (as percentage of total gene length), (B) of total gene length, (C) of PGT rate, (D) of the rate of sequence evolution, and (E) of having multiple PGT regions of different lengths within a single gene family. Plots (A)-(D) are based on baseline datasets and plot (E) is based on multi-PGT datasets. For Plot (E), the first and last columns (40% and 60%) show results for the corresponding baseline (single-PGT) datasets for reference. Reconstruction error is measured in terms of RF-score by comparing reconstructed pre- and post-PGT gene trees against corresponding true gene trees. All results are averaged over the 100 gene families in the corresponding dataset.

Impact of PGT-region to genic-region ratio. As expected, PGT-region to genic-region ratio has a direct and drastic impact on gene tree reconstruction. However, to our surprise, we found that gene tree reconstruction was only impacted once the length of the PGT-region exceeds 30% of total sequence length. This is shown in Figure 2(A), where a difference between pre-PGT and post-PGT gene tree accuracy is observed only when the PGT-region represents at least 40% of total sequence length. The figure also shows how post-PGT reconstruction accuracy rapidly degrades as the relative length of the PGT region increases. We note that this observation remains robust to changes in other parameter values, showing no or minimal impact at 30% length and a clear impact at 40% length consistently across all baseline datasets (results not shown). This finding is consistent with previous results of Posada and Crandall [26] who found that phylogenetic reconstruction was not affected if the recombining region was small compared to the length of the non-recombining region. We also point out the slight upward trend in pre-PGT error rate; this occurs because genic-region length decreases as PGT-region length increases, reducing the amount of information available for pre-PGT gene tree reconstruction.

Impact of total gene length. As Figure 2(B) shows, increasing the total gene length, while keeping other parameters at their default values, reduces both pre-PGT and post-PGT error rates. The post-PGT gene tree also remains considerable less accurate compared to the pre-PGT gene tree, except at the smallest gene length setting where both trees show high error-rate.

Impact of PGT rate. As Figure 2(C) shows, increasing the PGT rate (i.e., more transfer events in the subgene tree) leads to increased inference error in the post-PGT gene tree. As expected, the accuracy of pre-PGT gene trees remains unaffected (except for small random fluctuations). Interestingly, we found that the post-PGT gene tree was more accurate than the pre-PGT gene tree for the smallest setting of PGT rate (which corresponds to 2.03 PGTs per gene family, on average). This is because, when PGT events are rare, the benefit of using the full (longer) sequence alignment may outweigh the benefit of discarding PGT regions and using the resulting shorter sequence alignment.

Impact of sequence evolution rate. The impact of sequence evolution rate is similar to that of total gene length, affecting both pre-PGT and post-PGT gene trees similarly. This is shown in Figure 2(D) where both pre- and post-PGT gene trees are either simultaneously worsened or simultaneously improved as substitution rate changes. Somewhat surprisingly, we found that the error rates of both pre-PGT and post-PGT gene trees were nearly identical for the smallest setting of substitution rate. This is likely because at low evolutionary rates there may not be sufficient information in the sequence alignment to confidently reconstruct either type of gene tree. As the figure shows, and as expected, error rates also start to increase at higher substitution rates.

Impact of multiple PGT-regions. Figure 2(E) shows pre-PGT and post-PGT gene tree reconstruction accuracies for the four multi-PGT datasets. Unsurprisingly, the accuracy of pre-PGT gene trees increases with increasing length of the genic-region. However, careful analysis of post-PGT gene tree error rates reveals an

important, unexpected insight: We find that the error-rate of the post-PGT gene trees is impacted not by the total length of PGT-regions, but rather by the length of the single longest PGT-region. For instance, as the figure shows, post-PGT error rates for the {20%, 20%} and {30%, 10%} multi-PGT datasets are the same as their pre-PGT error rates, and much lower than the corresponding baseline dataset post-PGT error rate for PGT-region length 40%, despite the total PGT-region length being 40% in both these multi-PGT datasets. Likewise, the post-PGT error-rate for the {40%, 20%} multi-PGT dataset is much lower than for the corresponding baseline dataset with PGT-region length 60%.

A key insight from the above results is that when PGT regions are small (say less than a third of the total sequence length) or when PGTs occur very rarely, and even if multiple such PGT-regions appear within the same gene family, it may be beneficial to use the full gene family sequence alignment for gene tree reconstruction. At the same time, these results clearly demonstrate the significant adverse impact of longer and frequent PGTs on gene tree reconstruction.

We note that the results above show results averaged across all 100 gene families in the corresponding baseline dataset, even though not all 100 gene families in each dataset may have PGTs. Given the randomness inherent in any simulation framework, we found that, in datasets with the default PGT-rate of 0.4, 75 out of the 100 gene families had at least one PGT. These numbers were 71 and 88, out of 100, for datasets with PGT rates of 0.2 and 0.6, respectively. The results shown in Figure 2 are only minimally impacted even when limited only to gene families with at least one PGT (detailed results not shown). We also point out that post-PGT gene tree reconstruction accuracy does not depend on the "location" of the PGT-region within the sequence alignment since the gene tree reconstruction methods assume each site evolves independently. We therefore did not separately evaluate reconstruction accuracy on PGT-location datasets.

4.2 PGT detection accuracy

We used the baseline dataset with default parameter values (i.e., with 0.4 PGT evolution rate, 1000nt total sequence length, 0.5 substitution rate, and 40% PGT-region to genic-region ratio) to assess the ability of the PhyML-Mutli based approach and trippd to correctly detect the presence or absence PGTs. Since baseline datasets have the PGT-region appended at the end of the alignment, we also used the PGT-location datasets to further assesss the impact (if any) of PGT location within the sequence alignment. We also simulated additional datasets without any PGTs to further assess the false-positive rate of PGT detection for these methods.

Detecting PGTs using the PhyML-Multi approach. Recall that the default baseline dataset consists of 75 gene families with at least one PGT and 25 gene famlies without any PGT. We found that the PhyML-Multi based approach, using a histogram intersection test threshold of 50%, was correctly able to classify 63 of the 75 gene families as having PGT. However, PhyML-Multi also incorrectly classified 11 of the 25 gene families without any PGTs as having

PGT. For additional false positive testing, we ran PhyML-Multi on an additional dataset of 100 gene families with no PGTs and found that the method incorrectly detected PGTs in 65 out of the 100 gene families. Thus, the PhyML-Multi based approach shows a false negative rate of 0.16 (12/75) and a false positive rate of about 0.5 (more precisely, 0.44 (11/25) for the baseline dataset and 0.65 for the additional simulated dataset). Importantly, we found that these results are robust to the specific histogram intersection test threshold used and Table 1 shows the clear tradeoff between false-positive and false-negative rates of this approach as the threshold is decreased or increased.

Observe that the accuracy of his PhyML-Multi-based approach depends on PhyML-Multi's ability to correctly identify PGT boundaries/breakpoint(s). We found that, out of the 75 baseline dataset gene families with PGTs, PhyML-Multi was able to correctly detect the breakpoint to within 5 basepairs for 54 gene families. Thus, the breakpoint could not be accurately detected for 28% of the gene families.

Table 1. PGT detection accuracy using the PhyML-Multi based approach and trippd. False-positive and false-negative rates for both methods are shown when applied to the default baseline dataset and to the additional simulated dataset of 100 gene families with no PGTs. Results are shown for three different histogram intersection test thresholds, where the default threshold is 50%.

PhyML-Multi Based Approach

Baseline dataset	Threshold = 40%	Threshold = 50%	Threshold = 60%
False Positive Rate	0.36 (9/25)	0.44 (11/25)	0.52 (13/25)
False Negative Rate	0.24 (18/75)	0.16 (12/75)	0.13 (10/75)
No PGT dataset	Threshold = 40%	Threshold = 50%	Threshold = 60%
False Positive Rate	0.57	0.65	0.70
False Negative Rate	N/A	N/A	N/A

trippd

Baseline dataset	Threshold = 40%	Threshold = 50%	Threshold = 60%
False Positive Rate	0 (0/25)	0 (0/25)	0.16 (4/25)
False Negative Rate	0.27 (20/75)	0.2 (15/75)	0.11 (8/75)
No PGT dataset	Threshold = 40%	Threshold = 50%	Threshold = 60%
False Positive Rate	0	0.02	0.09
False Negative Rate	N/A	N/A	N/A

Detecting PGTs using trippd. As the lower half of Table 1 shows, an application of trippd to the same datasets shows much better PGT detection accuracy. In particular, we find that tripped has a drastically lower false positive rate and a comparable false negative rate as compared to PhyML-Multi. For instance, at the 50% histogram intersection test threshold, we found that tripped had a false positive rate of 0 on the baseline dataset and just 0.02 on the additional simulated dataset with no PGTs, compared to 0.44 and 0.65, respectively, for the PhyML-Multi approach. The false negative rate was also a relatively low 0.2,

which is roughly comparable to the 0.16 false negative rate for the PhyML-Multi approach. In fact, at a threshold of 60% both false positive and false negative rates of tripped are lower than those for PhyML-Multi.

To assess the impact of PGT-region length (default 40%) on the detection accuracy of trippd, we applied it to the baseline datasets with PGT lengths of 20%, 30%, 50% and 60% of the total gene length. We found that tripped was able to correctly classify 56, 58, 48, and 59 gene families, respectively, out of 75, as having PGTs. This corresponds to false-negative rates of 0.25, 0.22, 0.36, and 0.21, respectively; only slightly higher than for the default baseline dataset. Importantly, false-positive rates remained extremely low at 0.04, 0.07, 0, and 0, respectively.

We also assessed the impact of substitution (sequence evolution) rate (default 0.5) on trippd. Since sequences that are more similar are expected to undergo homologous recombination more easily, we applied trippd to the baseline dataset with a much smaller substitution rate of 0.1 and observed false-negative and false-positive rates of 0.29 and 0.16, respectively. Crucially, the increased false-positive rate is still much lower than the false-positive rate for PhyML-Multi. We also applied trippd to the baseline dataset with a very high substitution of 5. As expected, performance degrades substantially and the false-negative rate increases to 0.63. This is not surprising since the error-rates of the trees constructed for each of the three window are likely to be very high under this setting. Notably, the false-positive rate still remains very low, at 0, for this setting.

Interestingly, we observed that there were 11 gene families with PGT (out of the 75) that were consistently incorrectly classified as not having PGT, regardless of PGT-region length. We discovered that these 11 gene families had only a single PGT event. Thus, most of the gene families for which trippd fails to detect the presence of PGT are those in which only a very small amount of PGT has occurred. Furthermore, we found that among these 11 gene families, 8 had a lower post-PGT RF-score than pre-PGT RF-score. This indicates that for many of the gene families where tripped fails to correctly detect the presence of PGTs, it may, in fact, be beneficial to use the entire gene sequence alignment for gene tree reconstruction.

Impact of multiple PGT regions. We also assessed the impact of the presence of multiple PGT regions on the detection accuracy of trippd. Since trippd relies on phylogenetic discordance between pairs of windows, we chose the most challenging of all multi-PGT datasets, {40%, 20%}, for our evaluation. This specific dataset is particularly challenging for trippd since it has the property that each of the three windows largely represent three different evolutionary histories; window 1 consists entirely of the genic sequence, window 2 consists almost entirely of the first PGT region, and window 3 consists mostly of the second PGT region. On this dataset, 14 of the 100 gene families did not have any PGTs. Using our default histogram intersection test threshold of 50%, we found that 49 of the 86 gene families with PGT were correctly classified as having PGTs and only one of the 14 gene families without PGT was classified as having PGT. This corre-

sponds to a false positive rate of 0.07 and a false negative rate of 0.43. Thus, as expected, the PGT detection accuracy of trippd suffers when multiple PGTs occur in such a way that all three windows largely represent different evolutionary histories. However, such instances are expected to be relatively rare in practice.

Impact of PGT location. Finally, we used the three PGT-location datasets to assess the impact of PGT location on trippd. These results are shown in Figure 3. We find that as the evolutionary histories of window 2 and window 3 become more similar, the false negative rate of trippd increases. Specifically, from the baseline false negative rate of 0.2 (on the default baseline dataset using a threshold of 50%), the rate increases to 0.24 for the dataset with 34 bps offset, 0.35 for the dataset with 84 bps offset, and 0.35 for the 134 bps offset. The false-positive rate is also affected but remains relative low for all settings, with a high of 0.07 for the 84 bps offset dataset. Note that, since the three windows are treated identically by trippd, just these three PGT-location datasets cover all relevant cases. These results show that the PGT detection accuracy of trippd can be affected, though not drastically, if the PGT-region does not appear towards the beginning or end of a gene. However, since horizontal gene transfer often occurs through homologous recombination in flanking regions [25], PGTs may be more likely to occur at the beginnings or ends of genes.

Fig. 3. Impact of PGT location on trippd. The PGT detection accuracy of tripped, in terms of false positve and false negative rates, is reported for various locations of the PGT region within the gene sequence alignment. The first scenario describes our default baseline case where the PGT region is 40% of the gene length and occurs at the end of the gene. For this baseline case, the PGT region (orange) falls into two windows, 334 bps of the PGT-region is in the last window and the remaining 66 bps is in the middle window. The remaining three scenarios correspond to the three PGT-location datasets with offsets of 34, 84, and 134 bps, respectively.

4.3 Application to biological datasets

To assess the impact of trippd in practice, we applied it to the two biological datasets previously described. For the 784 gene families of the 100-taxon broadly sampled dataset, we observed that 62 (7.5%) were identified as having PGT. It is not surprising to see only a small number of gene families with detectable PGTs for this dataset since its species are broadly sampled from the entire tree of life and are therefore very distantly related to each other.

On the 466 gene family *Aeromonas* dataset, trippd identified 151 (32.4%) of the gene families as having PGT. This much higher percentage, compared with the 100-taxon broadly sampled dataset, is expected since the taxa in the *Aeromonas* dataset are much more closely related; thus, homologous recombination is expected to be both abundant and more easily detectable (due to relative recency) in this dataset.

Recall that trippd shows a very low false-positive rate of PGT detection. Thus, our results on these biological datasets indicate that PGTs, particularly those that are capable of affecting gene tree reconstruction, occur frequently in real biological datasets. trippd can easily help identify such cases for further analysis or filtering. Note, however, that these results about PGT prevalence are preliminary and should therefore be interpreted with caution.

5 Discussion and conclusion

In this work, we used a simulation study to assess the impact of partial gene transfer on gene tree reconstruction and presented a simple computational approach, trippd, based on alignment tri-partitioning to detect the presence of PGTs in gene family alignments. Our study of the impact of PGT reveals several important insights: We find that there can be significant adverse impacts of PGT on gene tree reconstruction accuracy. In such cases, it can be helpful to identify and remove the PGT region(s) from the alignment and reconstruct the gene tree on the reduced alignment. However, our results also show that if PGT regions are small (no more than a third of the total sequence length), or if only a very small number of PGTs have occurred, then gene tree reconstruction is unlikely to be impacted and it is likely beneficial to use the full gene family sequence alignment for gene tree reconstruction. We also find that multiple small PGTs do not significantly impact gene tree reconstruction accuracy and that adverse impacts depend on the length of the longest PGT-region. Our experiments with using PhyML-Multi to detect PGTs show that such an approach is effective at detecting PGTs, showing low false negative rate, but that it also has a very high false positive rate. Furthermore, we find a clear tradeoff between false positive rate and false negative rate for such an approach. The new approach, tripped, attempts to address this limitation and we demonstrate how tripped matches the false-negative rate of the PhyML-Multi based approach while having a negligible false-positive rate. Having a low false-positive rate is important for any effective PGT detection method since incorrect detection of PGTs can inflate or overestimate the impact of PGT in a dataset and lead to corrective

measures (such as using only an identified "non-PGT" region of the alignment) that ultimately lower the accuracy of reconstructed gene trees.

We view trippd as a preliminary, proof-of-concept approach, and it has several important limitations worth addressing. Most importantly, trippd can only *detect* the presence of PGT and not *identify* actual PGT regions. It may be possible to combine the strengths of recombination/breakpoint detection approaches such as PhyML-Multi and of tripped to both detect and identify PGT regions with high accuracy. Furthermore, it would be helpful to not only identify the different regions of an alignment but also to identify which region represents the underlying genic region and which represent PGT regions. The accuracy of trippd is also somewhat sensitive to PGT length, PGT location, and substitution rates, and methodological refinements could help address this limitation.

Several aspects of our simulation study can also be improved. In particular, our current study assumes that the same region of the underlying gene sequence undergoes repeated homologous recombination. A more reasonable model would be to allow each homologous recombination event to independently affect any region of the recipient gene. Likewise, it may help to appropriately model when homologous recombination between two gene sequences can occur (e.g., based on sequence similarity).

It is also possible that species-tree-aware approaches for gene tree reconstruction [1, 9, 13, 16, 22, 23, 31, 33] are more robust to the presence of PGTs and the impact of PGT on such approaches is worth investigating further. Finally, while our preliminary experimental analysis indicates that methods used to study genomic recombination, such as those implemented in RDP4 [21], have high false-positive rates of PGT detection (results not shown), it may be useful to evaluate the utility of such methods for PGT detection and identification more systematically.

Funding: This work was supported in part by NSF award IIS 1553421 to MSB.

References

1. Bansal, M.S., Wu, Y.C., Alm, E.J., Kellis, M.: Improved gene tree error correction in the presence of horizontal gene transfer. Bioinformatics **31**(8), 1211–1218 (2015). https://doi.org/10.1093/bioinformatics/btu806
2. Bay, R.A., Bielawski, J.P.: Recombination detection under evolutionary scenarios relevant to functional divergence. J Mol Evol p. 273–286 (2011)
3. Boc, A., Makarenkov, V.: Towards an accurate identification of mosaic genes and partial horizontal gene transfers. Nucleic Acids Res. **39**(21), e144 (Nov 2011)
4. Boussau, B., Guéguen, L., Gouy, M.: A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. Evol. Bioinform. Online **5**, 67–79 (Jun 2009)
5. Brinkmann, H., Göker, M., Koblízek, M., Wagner-Döbler, I., Petersen, J.: Horizontal operon transfer, plasmids, and the evolution of photosynthesis in Rhodobacteraceae. The ISME Journal **12**, 1994 – 2010 (2018)
6. Chan, C.X., Beiko, R.G., Darling, A.E., Ragan, M.A.: Lateral transfer of genes and gene fragments in prokaryotes. Genome Biology and Evolution **1**, 429–438 (2009). https://doi.org/10.1093/gbe/evp044

7. Chan, C.X., Darling, A.E., Beiko, R.G., Ragan, M.A.: Are protein domains modules of lateral genetic transfer? PLoS ONE **4**(2), e4524 (02 2009). https://doi.org/10.1371/journal.pone.0004524
8. David, L.A., Alm, E.J.: Rapid evolutionary innovation during an archaean genetic expansion. Nature **469**(7328), 93–96 (Jan 2011)
9. David, L.A., Alm, E.J.: Rapid evolutionary innovation during an archaean genetic expansion. Nature **469**, 93–96 (2011)
10. Dunning, L.T., Olofsson, J.K., Parisod, C., Choudhury, R.R., Moreno-Villena, J.J., Yang, Y., Dionora, J., Quick, W.P., Park, M., Bennetzen, J.L., Besnard, G., Nosil, P., Osborne, C.P., Christin, P.A.: Lateral transfers of large dna fragments spread functional genes among grasses. Proceedings of the National Academy of Sciences **116**(10), 4416–4425 (2019). https://doi.org/10.1073/pnas.1810031116
11. Etherington, G.J., Dicks, J., Roberts, I.N.: Recombination analysis tool (RAT): a program for the high-throughput detection of recombination. Bioinformatics **21**(3), 278–281 (Aug 2004)
12. Gogarten, J.P., Doolittle, W.F., Lawrence, J.G.: Prokaryotic evolution in light of gene transfer. Molecular Biology and Evolution **19**(12), 2226–2238 (2002)
13. Jacox, E., Chauve, C., Szollosi, G.J., Ponty, Y., Scornavacca, C.: eccetera: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics **32**(13), 2056 (2016). https://doi.org/10.1093/bioinformatics/btw105
14. Kloub, L., Gosselin, S., Fullmer, M., Graf, J., Gogarten, J.P., Bansal, M.S.: Systematic Detection of Large-Scale Multigene Horizontal Transfer in Prokaryotes. Molecular Biology and Evolution **38**(6), 2639—-2659 (2021). https://doi.org/10.1093/molbev/msab043
15. Koonin, E.V., Wolf, Y.I.: Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. **36**(21), 6688–6719 (2008). https://doi.org/10.1093/nar/gkn668
16. Kordi, M., Bansal, M.S.: Treesolve: Rapid error-correction of microbial gene trees. In: Martín-Vide, C., Vega-Rodríguez, M.A., Wheeler, T. (eds.) Algorithms for Computational Biology. pp. 125–139. Springer International Publishing, Cham (2020)
17. Kundu, S., Bansal, M.S.: SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution. Bioinformatics **35**(18), 3496–3498 (Feb 2019)
18. Lewis, P.O., Chen, M.H., Kuo, L., Lewis, L.A., Fucikova, K., Neupane, S., Wang, Y.B., Shi, D.: Estimating bayesian phylogenetic information content. Systematic Biology **65**(6), 1009–1023 (2016). https://doi.org/10.1093/sysbio/syw042
19. Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., Ray, S.C.: Full-length human immunodeficiency virus type 1 genomes from subtype c-infected seroconverters in india, with evidence of intersubtype recombination. J. Virol. **73**(1), 152–160 (Jan 1999)
20. Martin, D.P., Lemey, P., Posada, D.: Analysing recombination in nucleotide sequences. Molecular Ecology Resources **11**(6), 943–955 (2011). https://doi.org/https://doi.org/10.1111/j.1755-0998.2011.03026.x
21. Martin, D.P., Murrell, B., Golden, M., Khoosal, A., Muhire, B.: RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol **1**(1) (May 2015)
22. Morel, B., Kozlov, A.M., Stamatakis, A., Szollosi, G.J.: GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. Molecular Biology and Evolution **37**(9), 2763–2774 (2020). https://doi.org/10.1093/molbev/msaa141

23. Nguyen, T.H., Doyon, J.P., Pointet, S., Chifolleau, A.M.A., Ranwez, V., Berry, V.: Accounting for gene tree uncertainties improves gene trees and reconciliation inference. In: Raphael, B.J., Tang, J. (eds.) WABI. LNCS, vol. 7534, pp. 123–134. Springer (2012)
24. Petersen, J., Wagner-Dobler, I.: Plasmid transfer in the ocean – a case study from the roseobacter group. Frontiers in Microbiology **8**, 1350 (2017). https://doi.org/10.3389/fmicb.2017.01350
25. Polz, M.F., Alm, E.J., Hanage, W.P.: Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics **29**(3), 170 – 175 (2013)
26. Posada, D., Crandall, K.: The effect of recombination on the accuracy of phylogeny estimation. Journal of Molecular Evolution **54**, 396–402 (2002)
27. Rambaut, A., Grass, N.C.: Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics **13**(3), 235–238 (06 1997). https://doi.org/10.1093/bioinformatics/13.3.235
28. Rangel, L.T., Marden, J., Colston, S., Setubal, J.C., Graf, J., Gogarten, J.P.: Identification and characterization of putative Aeromonas spp. T3SS effectors. PLOS ONE **14**(6), 1–20 (06 2019). https://doi.org/10.1371/journal.pone.0214035
29. Ravenhall, M., Škunca, N., Lassalle, F., Dessimoz, C.: Inferring horizontal gene transfer. PLOS Computational Biology **11**(5), 1–16 (05 2015). https://doi.org/10.1371/journal.pcbi.1004095
30. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Math. Biosci. **53**(1), 131–147 (Feb 1981)
31. Sjostrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B., Lagergren, J.: A bayesian method for analyzing lateral gene transfer. Systematic Biology **63**(3), 409–420 (2014). https://doi.org/10.1093/sysbio/syu007
32. Stamatakis, A.: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9), 1312–1313 (May 2014)
33. Szollosi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E., Daubin, V.: Efficient exploration of the space of reconciled gene trees. Systematic Biology **62**(6), 901–912 (2013)
34. Tuomanen, E.I., Hollingshead, S.K., Becker, R., Briles, D.E.: Diversity of PspA: Mosaic Genes and Evidence for Past Recombination in *Streptococcus pneumoniae*. Infection and Immunity **68**(10), 5889–5900 (2000). https://doi.org/10.1128/IAI.68.10.5889-5900.2000
35. Weiller, G.F.: Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol. Biol. Evol. **15**(3), 326–335 (Mar 1998)
36. Zhaxybayeva, O., Lapierre, P., Gogarten, J.P.: Genome mosaicism and organismal lineages. Trends in Genetics **20**(5), 254–260 (2004)