The impact of HGT on phylogenomic reconstruction methods

Pascal Lapierre, Erica Lasek-Nesselquist and Johann Peter Gogarten

Abstract

Supermatrix and supertree analyses are frequently used to more accurately recover vertical evolutionary history but debate still exists over which method provides greater reliability. Traditional methods that resolve relationships among organisms from single genes are often unreliable because of the frequent lack of strong phylogenetic signal and the presence of systematic artifacts. Methods developed to reconstruct organismal history from multiple genes can be divided into supermatrix and supertree approaches. A supermatrix analysis consists of the concatenation of multiple genes into a single, possibly partitioned alignment, from which phylogenies are reconstructed using a variety of approaches. Supertrees build consensus trees from the topological information contained within individual gene trees. Both methods are now widely used and have been demonstrated to solve previously ambiguous or unresolved phylogenies with high statistical support. However, the amount of misleading signal needed to induce erroneous phylogenies for both strategies is still unknown. Using genome simulations, we test the accuracy of supertree and supermatrix approaches in recovering the true organismal phylogeny under increased amounts of horizontally transferred genes and changes in substitution rates. Our results show that overall, supermatrix approaches are preferable when a low amount of gene transfer is suspected to be present in the dataset, while supertrees have greater reliability in the presence of a moderate amount of misleading gene transfers. In the face of very high or very low substitution rates without horizontal gene transfers, supermatrix approaches outperform supertrees as individual gene trees remain unresolved and additional sequences contribute to a congruent phylogenetic signal.

Keywords: supertree; supermatrix; quartets; concatenation; horizontal gene transfer; phylogenomics

INTRODUCTION

Disentangling organismal descent from confounding factors such as horizontal gene transfers (HGT) remains an essential objective in the study of evolution. Genomes, especially in case of single-cell organisms, are mosaics, with different parts having different histories. Over short time intervals, vertical inheritance from ancestor to descendant dominates over HGT, and an organismal line of descent may be defined as a ‘plurality consensus’ of gene histories [1]. However, discerning organismal lineages in the tangle of gene trees remains a challenge [2–5]. In addition, over long time intervals, HGT between preferred partners can create patterns that may overwhelm the signal due to shared ancestry [6–9]. Lumping genes into a single dataset for analyses can improve resolution when the phylogenetic signal per gene is low, but the presence of HGT might yield phylogenetic trees that reflect neither the plurality of gene histories nor organismal history [10–13]. It is frequently considered safer to restrict the analysis to only a few genes that show no obvious conflict but these analyses often lack resolution due to most genes being excluded and a tree depicting the evolution of...
a single cellular component can hardly be considered sufficient to describe organismal phylogeny [14]. Supertree approaches, in addition to using sequence data, can also include morphological characters such as the presence–absence of genes, indels or shared genome rearrangements to resolve relationships among organisms [15,16]. Phylogenomic methods rely on the premise that it is possible to recover the organismal phylogeny by capturing a plurality signal of gene phylogenies while minimizing conflicting signals introduced by horizontally transferred genes, unrecognized paralogy, or shared biases, such as similar GC composition [10–12]. The level of misleading information needed to tilt the balance toward erroneous reconstruction is unknown and likely differs depending on the strategy used.

Simulation of sequence evolution using parameters estimated from real data is a powerful but underutilized strategy to test phylogenetic (phylogenies based on single genes) and phylogenomic approaches (phylogenies based on multiple genes) [17–20]. Previous simulations have shown that supermatrix methods are better in recovering the true phylogeny, especially for the resolution of branches with few substitutions, where single gene phylogenies result in unresolved branches [21,22]. However, in the presence of incomplete lineage sorting, both methods were found to be statically unreliable in the recovery of the true species tree [23,24]. Shared sequence biases (such as similar nucleotide and amino acid composition), non-homogenous sequence composition and long branch attraction artifacts caused by unequal rates of evolution among taxa can also lead to well-supported topologies that differ from the true species history [11,25]. While supertrees might be preferable to single gene datasets, debate still exists as to which phylogenomic method—supermatrices or supertrees—outperforms the other [26,27]. In previous simulations, supermatrix methods outperformed supertree methods, particularly in cases where there was poor overlap in taxonomic representation among gene trees [28–30]. However, no simulation study addressed the impact of HGT on supertree versus supermatrix performance. Supertree methods rely on the information contained within individual gene tree topologies to build a consensus tree. Many supertree approaches perform parsimony reconstruction on a matrix representing splits contained in the individual gene trees [25,31]. Some methods extract a plurality consensus from the decomposition of gene trees into embedded quartets [20,32,33]. The loss of information when supertree analyses summarize characters into trees and the inability of supertrees to detect hidden signal that is amplified by multiple genes might serve as weaknesses to the method [26,27]. Supertree approaches can also mislead in their ability to accurately recover the expected tree depending on the shapes and sizes of the individual gene trees used for the reconstruction [34,35]. However, the fact that supermatrix approaches incorporate all character information present in the dataset paradoxically serves as a strength and a weakness: phylogenetic resolution improves if concatenation amplifies a consistent but weak signal across genes but has the potential to recover a phylogeny that does not reflect the history of any of its constituent genes if random noise overwhelms the underlying signal [36]. Here, we explore the performance of supermatrix and supertree methods in the presence of two confounding factors to the resolution of organismal phylogenies: HGT and substitution rate variation. Specifically, using genome simulations, we tested the ability of supermatrix and supertree approaches to recover the original tree topology from sequence data when increasing levels of HGT or changes in substitution rates were included in the simulations. Our results suggest that overall supermatrix approaches are better in recovering the species history in the presence of slow or rapidly evolving sequences, while supertree approaches are more robust in the presence of increasing amounts of HGT. Since it is often difficult to determine the amount of HGT present in a real dataset, we suggest using both methods and investigating the conflicts when performing phylogenomic analyses.

**METHODS**

**Genome simulation and tree calculations**

All simulations were performed as depicted in Figure 1. Genome sequences were generated using EvolSimulator [17]. Each simulated genome contained a total of 100 genes ranging from 50 to 600 amino acids. Genes were concatenated into a single sequence from each simulated genome for the supermatrix tree calculations, or gene trees were calculated individually from each gene for the supertree approaches. EvolSimulator does not simulate insertions and deletions. The simulated sequences therefore did not contain any gaps or indels and did not
need to be aligned prior to the phylogenetic reconstruction step. In the analysis of real-world data, sequence alignment is a serious complication and prone to introduce additional artifact. These will be worthwhile to explore in future analyses; however, their consideration is beyond the scope of this study. Neighbor Joining (NJ) Trees were calculated with Kimura correction in Clustalw version 2.0.12 [37–39]. Maximum likelihood trees were calculated using FastTree version 2.1.2 [40] with the following parameters as recommended by the FastTree manual for improved accuracy: JTT model, optimized nearest neighbor interchanges (slownni), optimized all five branches at each NNI in two rounds (mlacc 2), four subtree-prune-regraft moves (spr 4) and estimated gamma under 20 categories. Five supertree reconstruction methods were tested on our simulations. The quartet-based method decomposed individual phylogenies into sets of quartets, and the frequency of each quartet reconstructed a supertree using Wilson’s inconsistency minimizing approach [41]. Matrix Representation using Parsimony (MRP) can be used to calculate a supertree from the original gene trees. MRP encodes each branch or split in each individual tree as a column of binary characters and uses parsimony to reconstruct the resulting supertree from the character matrix [31]. Robinson–Foulds (RF) distance supertrees attempt to find the optimal supertree with the least total RF distances (the dissimilarity between two trees expressed as the number of differences in bipartitions) between the supertree and the input trees. Similarly, the most similar supertree (dfit) method finds the optimal supertree with a scoring matrix that compares each source trees against every possible supertree. Finally, maximum split fit (sfit) methods identify the best supertree by looking at the number of shared splits between the input trees and all possible supertrees. Embedded quartets as well as the resulting plurality trees (supertrees)

Figure 1: General flowchart for the performed genome simulations and phylogenetic reconstruction methods. The simulations were run with increasing amount of proposed HGTs between genomes in EvolSimulator [17]. The accuracy of each reconstruction method was measured using the number of times the original phylogeny was recovered in 100 simulations and calculating RF distances [45] between true tree and reconstructed phylogeny. The reconstruction methods tested consist of gene concatenations, embedded quartets, RF distances, MRP and Clann Maximum Split Fit (sfit) and Most Similar Supertree (dfit) methods. The same flowchart was used with Evolver to test the effect of substitution rate variation on phylogenomic reconstruction.
Horizontal Gene transfer simulations

EvolSimulator is a software package that creates genome simulations mimicking evolution with more than 70 adjustable parameters. These simulations include lineage extinctions and transfers from lineages that later go extinct to lineages that survive until the end of the simulation. For these simulations, HGTs do not occur with the same probability between organisms. Rather, transfers are biased toward partners sharing the same habitats or genomes with similar G+C content. While we do not simulate highways of gene sharing directly, highways of gene sharing emerge as a consequence of the imposed biases [46]. HGT plays an important evolutionary role in all domains of life, best studied in single organisms [4,13,47]. Our parameter choices for genome simulations have levels of HGTs and imposed biases in relation to sequence divergence more reflective of bacterial evolution than multicellular animals or plants. The parameters for the simulation were empirically determined by a series of trials. The chosen parameter set resulted in a species tree with well-defined clades with a good representation of the different tree reconstruction methods in recovering the original tree topology and compared using RF distances between the recovered and true tree [45].

Assessing the effects of filtering for supermatrices

We examined the effects of filtering our data on the percent recovery of the true phylogeny by removing genes from each run that showed incongruent/inconsistent signal with the consensus phylogeny (the phylogeny generated from all 100 genes). In some instances, filtering genes might decrease the overall phylogenetic signal while in other cases filtering might remove genes that recover a grossly incongruent topology to the majority signal. Paup* [44] for the MRPP supermatrix calculations. The simulations were repeated 100 times to measure the reproducibility of the different tree reconstruction methods in recovering the original tree topology and compared using RF distances between the recovered and true tree [45].

for the uncertainty, we explore a wide range of transfer rates using the same species tree [49,50].

We ran EvolSimulator for 1000 iterations using the default parameters except for the following: preferred minimum and maximum number of lineages of 15 and 25, respectively; speciation/extinction rate of 0.25; 50 habitat spaces; genome size of 100 genes; gene size in amino acids of 50–600; mutation drift factor of 0.1; minimum and maximum mutation rates of 0.1 and 0.999 and initial mutation rates of 0.999. Parameters related to HGT were set to a random HGT probability of 0.1, divergence HGT probability of 0.2, gene complement HGT probability of 0.0, G+C HGT probability of 0.2, habitats HGT probability of 0.5, ortholog replacement probability of 1, HGT receptivity drift of 0.01, HGT minimum receptivity of 0.0, HGT maximum receptivity of 1.0 and initial HGT receptivity of 0.5. The resulting simulation yields a species tree with 22 extant genomes (Figure 2). To retain the same genome evolutionary history between runs but with different gene histories, all simulations were done using the same genome seed number \((z = 388767242)\). The number of HGTs introduced throughout the simulations was determined by varying the mean number of proposed HGTs per iteration in the control file from 0 to 100 in increments of 10. Evolver from the PAML package [51] simulated sequences with the EvolSimulator tree as a reference tree to assess the success in recovering the true species topology when faced with lower or higher substitutions rates in the sequences. This simulation sequentially increased or decreased the branch lengths of the reference tree by a factor of two without introducing any HGT. Each of the simulated genomes was composed of 100 genes that were each 300 amino acids long.
which allows the inclusion of majority rule bipartitions even if they are recovered fewer than 50% of the time. Then, a subroutine from Spruce v.1.1 (http://www.cs.utexas.edu/~phylo/software/spruce/) compared the number of shared bipartitions between the consensus tree and each gene tree. Supermatrices were created by removing gene trees that shared less than a certain number of shared bipartitions with its
consensus. We removed trees that shared fewer than 4, 7 and 15 bipartitions (from a total of 19 possible shared bipartitions). Under the most stringent condition, we removed all trees that conflicted with the consensus (i.e. did not match the consensus exactly).

RESULTS
As we increased the number of proposed HGTs per iteration in the simulation, only a fraction of the proposed HGTs actually affected genes of the extant species in the simulation as determined from the EvolSimulator output files (Figure 3). At the highest number of proposed HGTs, each gene on average transferred about 15 times with the number of transfers affecting each gene ranging from as few as five transfers up to a maximum of 32 transfers along the simulation. In presence of gene transfers, the supermatrix (gene concatenation) approach was found to be the least accurate method with the exception of supertrees reconstructed with the Clann dfit model (Figure 4). This pattern was true for both the actual number of correct trees recovered and the average RF distance between the tree reconstructed and the original tree topology (except for when no HGTs where proposed in the simulation). RF distances provide a measurement of topological dissimilarity between two trees, calculated by the number of bipartitions that differ in the compared trees. RF distances can be very sensitive to small changes between trees, which might inflate how poorly the various tree reconstruction methods perform when subject to small amounts of HGT. Calculating the percentage of trees recovered with the original topology leads to the same problem: the strict requirement for reconstructed trees to match the true tree exactly discounts that some trees with small changes might still be of overall good quality. While more accurate tree distance methods exists, i.e. SPR [52], geodesic tree distance [53], the computation power needed to calculate the distances using these methods increases dramatically as more disagreement is present between two trees (in our case increasing amount of HGTs), rendering them impractical for our simulations. Supertrees constructed from quartets embedded in maximum likelihood gene phylogenies were the most reliable with a 95% success rate in recovering the original tree topology for up to 20 proposed HGT events per iteration (Figure 4). The MRP reconstruction method was second best with a success rate >89% for up to 20 proposed HGT events per iteration, but matched or outperformed the quartet-based method for overall tree reconstruction accuracy as observed with the RF distances (Figure 4). Quartet and MRP supertrees tolerate the introduction of more HGT than all other methods before dropping precipitously in their success at recovering the true tree. In contrast, the supermatrix approach declines immediately and drastically in success rate for NJ and maximum-likelihood trees, suggesting that this strategy tolerates only minimal HGT rates before becoming

![Figure 3: Average number of accepted transfers per gene in the extant species. This figure represents the average number of transfers affecting each gene as the number of proposed transfer increases throughout the simulation. The error bars represent the standard deviation from the average. The gray zone represents the actual minimum and maximum number of transfers found in the simulation for all the genes.](https://academic.oup.com/bib/article-abstract/15/1/79/186338)
unreliable. In summary, supertree approaches were superior in recovering the true tree topology in the presence of HGT and have better overall tree resolution as reflected by lower RF scores. Maximum likelihood tree reconstruction always gave better reconstruction results for both supertree and supertrees.

We also tested the efficiency of the described supertree and supermatrix methods in the presence of little phylogenetic information, either due to highly conserved sequences (short branches) or high substitution rates (long branches or sequences saturated with substitutions). Our result revealed that sequence concatenation is superior to supertree approaches for both cases but particularly for highly conserved sequences. For example, we found that for sequences with low substitution rates the supermatrix approach had at least six to eight times the power of resolution as the supertree approach (Figure 5). At low substitution rates, supertrees calculated from NJ trees with Kimura corrected distances were less accurate in solving genome histories than the supermatrix method but both methods had comparable results for sequences with fast substitution rates (as reflected by large RF distances; Figure 5). When trees were calculated with maximum likelihood in the presence of substitution rates near saturation, supermatrices fared better as reflected by the absolute percent true tree recovery and RF distance plots (Figure 5). The quartet analysis recovered the true tree fewer times than all other supertree methods at lower sequence substitution rates for NJ and maximum likelihood trees but all supertree methods performed equally well at high substitution rates (according to the ‘percent recovery of original topology’ criterion; Figure 5).

Figure 4: Effect of HGT on phylogenomic tree reconstruction methods. As the number of proposed HGT events per iteration increases between simulations, all supertree approaches with the exception of the dfit method quickly outperformed gene concatenations for both NJ and maximum likelihood tree calculations. Supertrees using quartets embedded in maximum likelihood gene phylogenies were found to be the most reliable method with a success rate >95% in recovering the original tree topology for up to 20 proposed HGT events per iteration. The MRP reconstruction method was second best with a success rate >89% for up to 20 proposed HGT events per iteration but matched or outperformed the quartet based method on the overall tree reconstruction accuracy as observed with the RF distances. RF distances provide a measure for topological dissimilarity between two trees. Maximum likelihood-based tree reconstruction outperforms phylogenies calculated via NJ from Kimura-corrected distances.
A typical procedure before concatenation involves removing genes from the dataset that are incongruent with the majority in order to improve the strength of the phylogenetic signal [54]. However, our filtering strategy provided no improvement to the number of supermatrices that recovered the true phylogeny. Even when no HGT is present, the average number of shared bipartitions between the consensus tree and genealogies is 15/19 (Table 1). This number drops consistently as the level of HGT increases and only reaches a meager average of 3.8 shared bipartitions for the highest level of HGT (HGT100 dataset; Table 1). Thus, removing any gene from the dataset that does not recover the same tree as the consensus is too stringent—eliminating 97–100% of the trees—and essentially provides no information for further phylogenetic reconstruction (Table 1). We examined other cutoff values for a balance between accepting enough genes to provide adequate phylogenetic information and including a certain amount of HGT that would not destroy the central tendency. While the ‘15 plus’ cutoff value is still too stringent (except for HGT0), which reflects the loss of too much data, continuing to decrease the stringency (‘7 plus’ and ‘4 plus’) provides no benefit for any dataset with HGT (i.e. the percentage of supermatrices that recover the correct topology is either worse or remains the same as the unfiltered data; Table 1).

DISCUSSION
Genomes of microorganisms are dynamic and individual genes can have different origins and
evolutionary histories [4,9,55]. Genes can be duplicated, transferred or lost, and they can experience very different substitution rates [56,57]. Vertical inheritance is more frequent than horizontal transfer over short periods of time, and the majority of genes traveling together through vertical inheritance may be used to define the line of descent over short distances [1,58]. However, in the long run no single gene may have traveled through this lineage from beginning to end, and highways of gene sharing between divergent organisms may overwhelm the signal due to shared organismal ancestry [8,9]. Supertree and supermatrix methods are two popular approaches that aim to resolve organismal phylogeny with greater accuracy as compared to single gene phylogenies by pooling the phylogenetic signal common to many genes. Our analyses show that supertree and supermatrix procedures each have intrinsic advantages and disadvantages with respect to horizontal gene transfer and changes in substitution rates.

Supermatrix methods evaluate informative sites lumped together from multiple genes. Our results demonstrate that this approach provides better resolution when the overall phylogenetic signal in the sequences is very sparse. This can be explained by the fact that trees reconstructed from gene concatenations tend to reflect the average evolutionary signal present in the complete dataset. For this reason, gene concatenation should be used when the user knows that the data analyzed contain genes that are less likely to have been affected by gene transfers. Housekeeping genes and genes that form large interacting complexes, such as ribosomal proteins, often are considered to be infrequently transferred [22,59]. Additionally, when evolutionary rates were either very high or very low (Figure 5), supermatrices outperformed supertrees. However, at the extremes of substitution rates in our simulations, each gene contains very little phylogenetic information [60], leading to over-fitting [61] and unresolved gene trees in the analysis of individual gene families; whereas the increased non-conflicting signal obtained by concatenating genes helps the supermatrix strategy (particularly for maximum likelihood) to converge on the right answer. Supermatrices may also function reasonably well in cases where most transfers occur between closely related organisms and do not alter the phylogeny in the larger context. For example, a gene transferred between two species of the same genus might obscure the phylogenetic history of organisms within this genus but it will not impact inferred relationships between the genus and other bacterial clades. Only transfers between divergent organisms erode away phylogenetic signal while transfers biased in frequency toward close relatives tend to reinforce the phylogenetic signal due to shared ancestry [49,58]. To minimize the problem posed by HGT for supermatrix tree reconstruction, a filter can be used to find and eliminate sequences that either diverged too much from the rest of the

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Avg. No. of biparts</th>
<th>% of gene trees rejected</th>
<th>% recovery of true phylogeny</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>19</td>
</tr>
<tr>
<td>HGT0</td>
<td>15 ± 2.0</td>
<td>0</td>
<td>97</td>
</tr>
<tr>
<td>HGT10</td>
<td>12 ± 3.5</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>HGT20</td>
<td>10 ± 3.5</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>HGT30</td>
<td>8.0 ± 3.5</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>HGT40</td>
<td>7.0 ± 3.1</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>HGT50</td>
<td>6.0 ± 3.0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>HGT60</td>
<td>5.6 ± 2.75</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>HGT70</td>
<td>5.0 ± 2.5</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>HGT80</td>
<td>4.6 ± 2.5</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>HGT90</td>
<td>4.0 ± 2.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>HGT100</td>
<td>3.8 ± 2.2</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

Summary statistics for filtering applied to each dataset. Avg. # bipartitions, average number of bipartitions shared between gene trees and their consensus (±SD); % of gene trees rejected, percent of maximum-likelihood gene trees rejected from each dataset at different cut-off values; % recovery of true phylogeny, the percent of supermatrices after filtering that yielded the correct topology; All, the entire dataset (100 genes) without filtering (used for comparison to filtering results); 19, the most stringent filtering value that accepted only gene trees with identical topologies to the true phylogeny; 15 plus, filtering value that accepted gene trees that shared 15 or more bipartitions with the true phylogeny; 7 plus, filtering value that accepted gene trees with 7 or more shared bipartitions to the true phylogeny; 4 plus, filtering value that accepted gene trees with four or more shared bipartitions.
data or for which individual tree phylogenies show conflicting signals with the consensus. However, implementing the pre-filtering step to our simulations did not improve the percentage of supermatrices that recovered the true phylogeny, except when no HGT was present. In certain instances, this was due to the fact that the filtering parameters were too stringent—eliminating most of the genes from the dataset—but relaxed stringency generated filtered supermatrices with essentially equivalent recovery rates as their unfiltered counterparts. Examining the types of incongruencies between the genealogies of each dataset (HGT0, 10, 20, etc.) and the true phylogeny provides clues to the inefficacy of this filtering method. All topological differences between HGT0 genealogies and the true phylogeny stem from branching differences among closely related taxa; major clades and subclades remain intact and most genealogies show only one to two conflicts with the true phylogeny. Thus, HGT0 supermatrices that include the maximum number of genes (100), despite minor incongruencies fare decently; the whole is greater than the sum of individual parts. This pattern generally holds for the HGT10 dataset as well, with transfers among major clades being rare within genealogies. However, by the time HGT reaches level 50, deeper nodes are not recovered and both recent and more ancient HGT impact genealogies—making the recovery of a consistent signal with concatenation virtually impossible. Our filtering method also displays weakness due to certain flaws including not taking bootstrap support into consideration (due to computational limitations). A better filtering method, such as a likelihood ratio test to identify genes that generate significantly different trees, might improve the ability of supermatrices to recover the true phylogeny [62]. Yet, our filtering method suggests that the supermatrix approach is relatively robust to small amounts of HGT (depending on perspective) because eliminating genes with the most egregious signal still recovers the true phylogeny at the same rate as when these genes were simply kept in the analysis.

In the presence of HGT, supertree reconstruction provides a more robust alternative to supermatrices with quartet decomposition being the most accurate of the methods we tested. Quartet decomposition disassembles gene trees into embedded quartets [41] and constructs supertrees only from plurality/majority quartets, which will remove phylogenetic information from HGT. In other words, this approach decomposes trees into simple parts to find the majority signal linking each branch individually without being affected by the overall average signal. If the individual gene families contain a clear phylogenetic signal, the quartet-based supertree approach reliably recovers the phylogeny of the majority of individual genes [22]. In contrast, MRP calculates supertrees directly on all of the gene trees and does not filter out the phylogenetic signals due to HGT prior to supertree assembly. Therefore, MRP is expected to be more susceptible to misleading signal created through HGT, which we observe in our results. Sfit (split fit) and dfit (most similar supertree) methods assemble supertrees via the number of shared splits or a distance matrix between the proposed supertree and the input trees to find the best tree and are more similar to MRP than quartet methods. Other probabilistic approaches that do not include a filtering step probably fall somewhere in between these two major methods, but further studies are needed to evaluate the sensitivity of the different supertree approaches toward HGT. Regardless of the specific method, supertree approaches all rely on individual tree reconstructions and any lack of resolution from the individual trees could pose serious problems in the reconstruction of the consensus tree just as conflicting phylogenetic information from HGT will pose serious problems for supermatrices. Thus, depending on the dataset, either the supertree or supermatrix method might have the performance advantage. Realistic datasets (especially for the prokaryotic realm) contain HGT and sequences with little phylogenetically useful information. Tree reconstruction methods must negotiate a path that adequately accounts for both of these confounding factors. We suggest performing both supertree and supermatrix analyses and filtering the data to minimize the impact of HGT and optimize resolution due to low phylogenetic informativeness for very recent and very deep divergences.

---

**Key points**

- Supermatrices and supertrees are two methods that are frequently used to resolve phylogenies with information from multiple genes.
- However, the reliability of both methods in the presence of poor phylogenetic signal or horizontal gene transfer remains unknown.
- Our data based on genome simulation indicate that supermatrix methods performed better when the signal contained in the data is highly diluted while supertrees are better in recovering the true tree topology in presence of HGT.
SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

This work was supported by the National Science Foundation Assembling the Tree of Life Program (DEB 0830024).

References

37. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific


