Untangling hybrid phylogenetic signals - HGT and artifacts of phylogenetic reconstruction

Running head: Hybrid phylogenetic signals

Robert G. Beiko and Mark A. Ragan

Summary

Phylogenomic methods can be used to investigate the tangled evolutionary relationships among genomes. Building "all the trees of all the genes" can potentially identify common pathways of horizontal gene transfer (HGT) among taxa at varying levels of phylogenetic depth. Phylogenetic affinities can be aggregated and merged with information about genetic linkage and biochemical function to examine hypotheses of adaptive evolution via HGT. Additionally, the use of many genetic data sets increases the power of statistical tests for phylogenetic artifacts. However, large-scale phylogenetic analyses pose several challenges, including the necessary abandonment of manual validation techniques, the need to translate inferred phylogenetic discordance into inferred HGT events, and the challenges involved in aggregating results from search-based inference methods. In this chapter we describe a tree search procedure to recover most-parsimonious pathways of HGT, and examine some of the assumptions that are made by this method.

Keywords: Phylogenetics; phylogenomics; horizontal gene transfer; subtree prune-and-regraft; bipartitions; model violation

1. Introduction: Building large sets of trees

HGT detection methods can be classified into two categories, depending on whether or not they rely on comparisons among genomes to identify homologous sequences. Homology-independent methods typically rely on the distribution of different compositional patterns such as G+C content ***(1)*** and are sometimes identified as *surrogate* ***(2)*** methods. Surrogate methods carry certain advantages: they can be applied to an entire genome, including protein-coding sequences that have a small number of orthologs in other genomes. However, coding sequences in a genome will not all be influenced to the same degree by background signals, and the amelioration process can lead to ambiguous classifications of putatively transferred genes. Different surrogate methods have been shown to generate sets of predicted acquired genes that overlap poorly ***(1,3).***

Homology-based approaches, particularly those based on phylogenetic analysis, can be applied only to sequences for which a sufficiently large number of reliable homologs can be identified. Such methods can be based on the observed 'patchiness' of a distribution of homologous or orthologous sequences ***(4,5)*** with sparse distributions of such sequences across a reference tree of organisms constituting *prima facie* evidence for HGT as opposed to multiple gene loss events. Such methods are sensitive to the sparsity and bias of taxon sampling, and the unknown intrinsic rates of gene loss versus HGT. Phylogenetic or 'phylogenomic' approaches are based on the consistency of phylogenetic signals across many genes: if all observed similarity relationships ***(6)*** or phylogenetic trees ***(7)*** are compatible with one another over a relatively unbiased sample of taxa, then

there is taken to be little evidence of HGT. In contrast with comparisons among surrogate methods, Beiko et al. found good agreement between a tree-based and a distribution-based approach to identifying phylogenetic discordance ***(8)***. However, differences in the choice of methodology, taxon sampling and data set (e.g., informational genes, ubiquitous genes, or whole genomes) have produced a wide range of HGT frequency estimates from different homology-based approaches ***(5,7-9)***.

Tree-based approaches potentially carry several advantages for HGT inference, including statistically consistent likelihood-based methods, the ability to specify models of sequence change, and the possibility of identifying donor-recipient relationships. Conversely, tree-based methods also have disadvantages, including sensitivities to model violation, computational demand, and challenges in interpreting and summarizing observed patterns of discordance.

2. Data-set generation / software

Inference of orthologs is an essential component of the tree-based approach, and there are many ways to perform this step. Protein data are typically used to compare distantly related taxa: since amino acids evolve more slowly than nucleotides and have more character states than nucleotides, they are less prone to effects of substitutional saturation. Many ortholog inference methods begin with all-*versus*-all BLAST (or a similarly defined heuristic) to identify putative homology relationships within the set of sequences. The challenge then lies in converting the resulting graph, with sequences as vertices and similarity relationships defining edges, into a set of discrete and non-overlapping groups to be analyzed separately. Methods to do this include CD-HIT ***(10)***,

hybrid Markov-single linkage clustering ***(11)***, and phylogenetic approaches such as BranchClust ***(12)***.

There are many ways to build and trim multiple sequence alignments and infer phylogenetic trees; comprehensive discussion of these is beyond the scope of this chapter, but a key challenge in phylogenomic analysis is the high degree of reliance on automation. Techniques such as manual alignment curation are neither viable nor sufficiently consistent for very large data sets. The alignment 'word-oriented' objective function WOOF ***(13)*** is an example of a method intended to replace manual curation by examining the same 'cues' a human would look for (consistent alignment of conserved residues in the case of WOOF). In addition to being the only viable approach for large-scale analysis, such methods are more rigorous, consistent in their treatment of the data, and repeatable.

There are many approaches to quantifying the dissimilarity between phylogenetic trees. The most relevant of these to the HGT inference problem is the subtree prune-and-regraft (SPR) distance ***(14)***, typically defined as the minimum number of SPR operations needed to reconcile two trees. Computing the minimum SPR distance between two unrooted trees is an NP-hard problem ***(15)***, and the most frequent approach involves a search of possible intermediates to recover the final path or paths. Different constraints can be placed on the set of allowable SPR operations to eliminate transfers that are evolutionarily impossible (such as from descendant to ancestor), and to reduce the size of the search space. Programs that implement different variants of tree comparison include LatTrans ***(16)***, HorizStory ***(17)***, Efficient Evaluation of Edit Paths (EEEP) ***(18)*** and RIATA-HGT ***(19)***. In their approaches HorizStory and EEEP are somewhat similar, but

all four methods differ in the type of trees they take as input (rooted or unrooted, strictly binary or possibly multifurcating), the evaluation of potential HGT partners, and the strategy used to restrict the size of the search space.

3. Assessing and aggregating phylogenetic evidence for HGT

This section assumes that phylogenetic trees have already been generated through an appropriate combination of methods for orthology inference, multiple sequence alignment, alignment 'trimming' and phylogenetic tree inference or calculation. The approach used in the phylogenomic project of ***(8)*** is described in stepwise fashion in Beiko and Ragan ***(20)***, but many other approaches have been taken to generate large-scale phylogenomic data ***(7, 21)***. We also assume that a *reference tree* describing a plausible scenario of vertical descent of organisms is available; ideally this tree will be rooted (see Section **3.1** below). This reference tree can represent the plurality signal from the inferred trees under consideration, or can represent a phylogenetic hypothesis that is extrinsic to the data (*e.g.* based on small-subunit ribosomal DNA, cellular ultrastructure, or other information thought to be phylogenetically informative). We refer to each member of the set of inferred trees as a *test tree*, to distinguish it from the reference tree.

3.1. Assessment of phylogenetic discordance

An efficient way to assess the degree of compatibility between a reference tree and a set of test trees, is to count incidences of reference tree features (such as bipartitions or quartets) that are topologically *congruent* or *incongruent* with individual test trees. An example of mapping concordant and discordant features is shown in **Fig. 1**.

For the mixture of trees in **Fig. 1A**, the number of bipartitions that are concordant and discordant with each internal branch of the reference tree is shown above and below the corresponding branches in **Fig. 1B**. Although this example treats all trees as being completely resolved, the total number of trees examined should also be taken into account: a high proportion of unresolved bipartitions (i.e., with low Bayesian posterior or supporting proportion of bootstrap replicates) may suggest a difficult relationship to resolve. Bipartitions can be mapped onto a tree, but the complete spectrum of *embedded* quartets (**Fig. 1C**), while more difficult to summarize, confers more information about the consistency of relationships among specific taxa in the reference tree. Quartet decomposition of a phylogenetic tree ***(22)*** is performed by pruning away the complements of all possible *n* choose 4 = n!/(4!(n-4)!) sets of taxa from a tree covering *n* taxa.

[Insert Fig. 1 here]

Phylogenetic discordance can also be assessed with *consensus network* methods such as Neighbor-Net ***(23)*** or *super-network* methods such as Z-closure ***(24)***. All of these methods can reveal cases in which the consensus signal from many trees has significant representation of specific alternative relationships, which are unlikely to be due to rare events or noisy data. Such relationships appear as reticulations in the network, with parallel edges separating groups of taxa that cluster together with significant frequency. As shown in **Fig. 2**, super-networks can display short-range (**Fig. 2B**) and long-range (**Fig. 2C**) transfers. Super-networks typically display a 'web' of reticulations when two

trees disagree due to a long-range transfer, instead of a single additional connection between donor and recipient lineages that might be expected. A single such transfer can still be identified (e.g., from the reticulations shown in **Fig. 2C**), but two or more long-range transfers may not be distinguishable if the webs they produce intersect in the network.

[Insert Fig.2 here]

3.2. Recovery of HGT pathways

If the reference tree is rooted and has accurate branch lengths that are proportional to time, then any paths of HGT could be constrained to occur only between contemporary lineages in the reference tree (but see **Note 1**). However, parsimony methods including the Matrix Representation with Parsimony (MRP) approach ***(25)*** for generating reference trees do not themselves estimate branch lengths, although an MRP tree could be fixed and branch lengths subsequently estimated from the data using a likelihood approach. If the reference tree is based on distance or likelihood analysis of e.g. the small-subunit ribosomal RNA gene or a concatenated alignment of many such genes, then branch lengths in the tree will reflect the number of substitutions per site. However, such branch lengths will be proportional to time only if the rate of sequence change has been constant through time (i.e., evolves according to a molecular clock). Since this assumption rarely holds for distantly related sets of taxa, it is customary to ignore branch lengths when inferring HGT from phylogenetic trees, and focus on differences in the branching order.

The SPR distance between two trees ***(14)*** considers only the branching order of taxa and is not sensitive to differences in branch lengths between two trees. Each SPR operation involves a donor and a recipient lineage, and the topological impact of SPR is to prune the recipient subtree and reattach it to the branch corresponding to the donor lineage. This is directly analogous to an HGT event in which a given gene is replaced by its ortholog from a different genome. We term a sequence of such operations an *edit path*, and the minimal-length edit path from the reference tree to a given test tree corresponds to minimum number of HGT operations that need to be inferred in the history of the gene whose evolution is described by the test tree. Therefore, in seeking a minimum-length edit path we are trying to recover the most-parsimonious explanation for gene evolution, in terms of the number of HGT operations that need to be mapped onto the reference tree (see **Note 2**)**.** Even if branch lengths are ignored, the reference tree imposes a partial ordering on lineages by specifying ancestor / descendant relationships, and we can prohibit donations of genetic material from an ancestral lineage to one of its descendants, or vice versa (but again see **Note 1**). This constraint on SPR operations actually produces a different distance measurement, the *hybridization distance* ***(26, 27)***.

For a set of donor lineages D_i and recipient lineages R_i, an edit path recovered by EEEP is of the form $(D_1 \rightarrow R_1)$, $(D_2 \rightarrow R_2)$, $(D_3 \rightarrow R_3)$. However, unambiguous answers where only a single edit path is returned are rare: a complete set of most-parsimonious solutions may include rearrangements of the order of edit paths, inversions of the donor and recipient lineages, and alternate paths that include other lineages in place of the ones above. Questions about the extent and nature of HGT can be posed at several levels of

detail, and the amount of information that can be recovered depends on several factors that are outlined below.

What is the length (number of SPR operations) of the shortest edit path or paths? This question, unlike the ones below, can be answered by an exact algorithm (although the computation may be limited by available time and memory). In computing the shortest distance we are making a parsimony assumption, and indeed model-based approaches to the problem have been developed in which the minimum length does not necessarily correspond to the maximum likelihood or maximum *a posteriori* HGT scenario (see **Note 2**).

Which lineages are identified as HGT partners? This question addresses which lineages in the tree are donor-recipient pairs, without a clear indication of the direction of transfer. Even if a reference / test tree reconciliation (achieved by choosing an appropriate donor/recipient pair) can be unambiguously attributed to lineages A and B, the direction of the implied HGT event may not be determinable from the data. The unambiguous recovery of donor-recipient pairs is not guaranteed, which is best illustrated via an HGT event that changes the topology of a four-taxon tree: if (A, B) and (C, D) are the sister-taxa pairs in the reference tree, then the phylogenetic effects of a transfer between B and C will be indistinguishable from one between A and D, unless meaningful branch lengths are present and exploited. In this case, the identity of transfer partners cannot be uniquely identified. The cases in which donor-recipient pairs *can* be recovered identifies 'partners'

in HGT events; even this incomplete information can be useful in constructing and testing hypotheses about HGT.

What is the direction of transfer? In some cases a donor-recipient pair *does* have an unambiguous direction of transfer (i.e., A→B is present in the reconciliation path, but not B→A), in which case the identity of the donor and recipient lineages can be inferred. 'Long-distance' transfers, where a gene from one organism is acquired by a very distant relative (e.g., interphylum transfers), leave the donor in its original grouping, but displace the recipient away from its canonical place in the reference tree. In addition to information about partnerships, this information can shed light on possible adaptive roles of the HGT event, since the ecology and metabolic capacity of the recipient genome can be compared to its closest relatives in the reference tree. A surrogate method might also be of use in identifying which genome (A or B) is the recipient, since an acquired gene that is only partially ameliorated ***(1)*** may still show unusual compositional patterns in the recipient genome.

What is the order in which these transfers have taken place? The HGT events implied by an edit path of two or more transfers can be thought of as independent if none of the donor or recipient edges is an ancestor or descendant of another. In other cases, it may be possible to assign a time ordering on successive transfers into a recipient genome if there are sufficiently many internal branches to resolve separate transfer events.

When the answers to one or both of the last two questions are ambiguous, we can merge edit paths in the solution set that differ only in the direction of transfer between donor-recipient pairs, as well as those that differ only in the ordering of successive HGT events. While the search algorithm will likely visit all of these solutions separately, considerable gains in storage efficiency and clarity of results can be achieved through this merging. For instance, a path consisting of five edits has 5! = 120 possible orderings, but if the distinction between orderings is not of interest, then these can all be represented as a single path with an explicit indication that the ordering of elements is arbitrary. EEEP uses a compressed format to represent such paths, where a single path is shown in full, followed by numerical representations of all recovered permutations of that path.

3.3. Aggregating inferred HGT events from a phylogenomic analysis

If every comparison between the reference tree and a test tree yielded a single, most-parsimonious edit path reconciliation, then aggregating the implied HGT events would simply require a summing up over all edits. However, the ambiguities outlined in **Section 3.2.3** were observed in a majority of the trees examined in ***(8)***. Therefore, non-trivial aggregation techniques need to be built. In this section we discuss three different aggregation schemes: a *greedy* approach, *weighting* of edit paths, and *refinement* techniques.

The *greedy approach* to aggregation involves choosing the single most-likely scenario from among the set of recovered paths for each test tree in turn. In the absence of models which assign different probabilities to different donor-recipient pairs, a greedy approach can be used to favour those donor-recipient pairs that are observed most

frequently in the complete set of reference tree-test tree reconciliations. If lineages A and B are proposed as potential partners in the reconciliation paths of 200 out of 1000 different test trees, and if no other pair has an equal or greater frequency, then the greedy approach favours edit paths containing this pair over all possible alternatives in each of these 200 cases. In doing this, many alternative edit paths will be eliminated from consideration. With this decision made, the next most-frequent pair from the set of all edit paths still under consideration is found, with the same selection and elimination process. This iterative procedure can be repeated until each test tree has a single remaining reconciliation path with the test tree (see **Note 3**). The greedy approach is sensitive to the order in which edit paths are chosen, and variation in rankings can lead to the recovery of different sets of edit paths. However, its robustness with respect to any data set can be tested by randomly permuting the ranking of paths many times, and comparing the sets of paths recovered from each permutation.

When aggregating donor-recipient pairs across many test tree reconciliations, the *weighting approach* assigns fractional values to donor-recipient pairs based on their frequency in each reconciliation after path permutations have been merged. Therefore, if 50% of all reconciliation paths for a given test tree contain the donor-recipient pair (A, B), then we will add 0.5 to the total observed count of that pair (see **Note 4**).

Refinement approaches aim to decrease the ambiguity of results by reducing the precision of the phylogenetic question that is asked. For instance, a given test tree may have several mutually exclusive reconciliation paths with the reference tree, with lineages A, B, C, D, or E as possible donors and F, G, H, and I as possible recipients. Although there is no unambiguous donor / recipient pair, if A, B, C, D, and E are all part of the

same grouping (e.g., the same genus or class, or clade in the reference tree), and F, G, H, and I are part of the same group (which may be the same or a different genus, class, clade, etc.), then we can say that the transfer is *obligately* between these two groups. The loss in precision is offset by the ability to address hypotheses about gene sharing partners, even if the exact donor and recipient lineages cannot be identified. A related question is how many transfers have *possibly* been observed between two groups; in this case, all of A, B, C, D, E may not be a part of the same group, but a group that contains a subset of these lineages may be implicated in the transfer event. Summaries of obligate or possible transfers can be evaluated in light of other intersecting ecological or functional hypotheses.

3.4 Testing for errors of phylogenetic inference

Many types of methodological error or bias can lead to incorrect inferences of HGT. Conflation of paralogs with orthologs can lead to inferences being made on sets of sequences whose relationships are not congruent with the organismal tree, without the need to invoke HGT ***(28)***. Lineage sorting is another phenomenon that can lead to mistaken inference of HGT ***(19, 29)***. Errors in multiple sequence alignment can lead to incorrect phylogenetic inferences based on non-homologous residues. Sources of bias that are particularly relevant to microbial genomes include violations of the assumptions of model-based phylogenetic methods.

Non-tree-like signals. Inferring a phylogenetic tree from a given dataset (e.g., gene or protein sequence) makes the assumption that the underlying data have a unique, tree-like

history. However, reticulate evolution events such as gene conversions can produce genes which contain multiple incongruent phylogenetic signals, and traditional statistical methods such as bootstrap resampling are not suited to the detection of clusters of mutually incompatible phylogenetic signals within an alignment. There is some evidence that HGT can affect fragments of genes in addition to whole genes and operons ***(30,31)***. Phylogenetic inference from such data may support one or the other of the correct signals, or may support a 'phylogenetic compromise' that reflects neither scenario. Phylogenetic recombination detection methods such as RecPars ***(32)*** and BARCE ***(33)*** can be used to identify such hybrid signals, but methods designed specifically with fragmentary HGT in mind could consider variations in sequence composition as well.

Stochastic error. Phylogenetic methods based on likelihood scores are statistically consistent, assuming the evolutionary model is correct. However, it has been demonstrated ***(34)*** that large numbers of alignment positions (relative to the length of a single gene or protein) may be necessary to yield highly accurate trees. Many 'short-distance' transfers may arise as a consequence of undetected stochastic error (but see **Note 5**), with the influence of stochastic error increasing with decreasing alignment length, and with increasing ratios of terminal to internal branch lengths ***(35)***. Supermatrix methods were developed to overcome stochastic error by concatenating many sequences from the same group of taxa. However, building a single tree from all available genes assumes that no reticulate evolution has taken place, and supermatrix methods are consequently inappropriate for the assessment of HGT. Concaterpillar ***(34)*** aims to group sequences based on their probable phylogenetic signals, and has the potential to approach

the power of supermatrix methods while identifying sets of phylogenetically discordant genes.

Compositional change along branches of the tree. Microbes have a very large range of genomic G+C composition, and amino acid usage differs both as a consequence of DNA composition ***(37)*** and environmental requirements ***(38)***. Simulations have shown that violations of assumptions such as compositional stationarity can bias otherwise consistent phylogenetic methods in favour of incorrect trees ***(39)***. Matched-site compositional tests such as Bowker's test ***(40)*** can be applied to diagnose the magnitude and significance of the problem, and the problem can be remedied either by using composition-insensitive methods such as LogDet distances ***(41)*** or by correcting for compositional differences via e.g. purine-pyrimidine recoding of nucleotides ***(42)***.

Variation in evolutionary rates. Likelihood-based phylogenetic methods are sensitive to variations in the rate of substitution unless the substitution model used is accurate ***(43)***. The best-known example of this type of sensitivity is the 'long branch attraction' artifact. The relative rate test ***(44)*** is one widely-used method to assess the degree of rate variation in the tree.

4. Conclusion: estimating the number of HGT events

In ***(8)***, we used Bayesian phylogeny and an MRP supertree to estimate the balance of vertical versus horizontal evolutionary signal in a set of 144 genomes. Based on a comparison of strongly supported (Bayesian posterior probability ≥ 0.95) bipartitions against the reference tree, we found that 86.9% of all strongly supported bipartitions were concordant. There was considerable agreement with the reference supertree, and EEEP was used to recover frequent pathways of apparent HGT among lineages. But do these results indicate that the underlying 'frequency' of HGT is 13.1%, and why does this number differ so dramatically from other estimates?

There are several reasons why our comparative approach might overestimate the true number of HGT events. Discordance may arise due to the stochastic and systematic errors outlined in Section **3.4** above, and other phenomena such as lineage sorting may contribute when internal branches are short. A single long-distance HGT event will also disrupt more than one bipartition, since the recipient taxon will be lost from its original species, genus, etc. groupings, and disrupt bipartitions that include the donor taxon as well.

The choice of reference tree will also influence the number of recovered HGT events. Using an MRP supertree ensures that some relationship will exist between the reference and test trees. Even so, the most-parsimonious supertree will not necessarily minimize the number of HGT events that need to be invoked, due to the mismatch between HGT events and disrupted bipartitions identified above. Other analyses have used different types of reference tree: for instance, Dagan and Martin ***(5)*** used a

concatenation of three (5S, 16S, and 23S) ribosomal DNA alignments across 190 species. However, ribosomal DNA is sensitive to the mutation biases that affect the rest of the genome, and can lead to artifactual grouping of organisms based on their G+C content ***(43)***. Use of an incorrect reference tree will overestimate the number of HGT events, particularly if the test data (gene trees or gene presence / absence data) are less sensitive to compositional artifacts than is the original reference tree.

Conversely, it is also probable that tree-based methods underestimate the number of HGT events that have occurred. A methodological reason for this is the need to choose a threshold of significance for inferred phylogenetic relationships. As this threshold decreases, more discordant events will appear in the dataset: for instance, while 13.1% of bipartitions were discordant at a BPP threshold of 0.95, over 23% were discordant when the threshold was reduced to 0.51. With 90% often treated as a minimum threshold of reliability for Bayesian posterior values, most of the discordant features with support in the 50-60% range are likely to be a consequence of weak phylogenetic signal or uncorrected errors.

Tree-based methods cannot detect events that do not disrupt the branching order of the recovered phylogenetic tree. The inability to detect transfers among sister taxa was mentioned above, but Ge et al. ***(9)*** extended this by eliminating a considerable amount of short-distance discordance that *was* detected in reference-test tree comparisons, which produced a low estimate of ~ 2.0%. Another way to reduce the inference of short-distance HGT events is to allow the reference tree to multifurcate, which can be done with HorizStory ***(17)***. A trifurcating reference tree node imposes no ordering on the descendant lineages A, B, and C, whereas a bifurcating tree would necessarily contain a

nested pair such as (A,(B,C)). The multifurcating reference tree will not be discordant with any grouping of these three lineages, so short-distance transfers among them will not be detected.

Using trees also overlooks the distributional evidence for HGT. Simple presence/absence analysis (phylogenetic profiles) reveals many orthologs with patchy distributions across groupings of taxa, including genes involved in aerobic respiration ***(46)***. Indeed the analysis in ***(5)*** was based on the presence and absence of orthologous genes in different lineages, and led to estimates of > 1 exchange per gene family, and > 50% overall.

These methodological differences are all important and also highlight the importance of defining a precise question prior to inferring networks of gene sharing. Given the evolutionary and ecological diversity of microbes and the evidence from many comparative analyses of HGT, it is clear that HGT frequency cannot be characterized with a single percentage that describes all organisms and all genes. This is reflected in hypotheses that reflect the role of HGT in the emergence of specific functions such as photosynthesis ***(47)*** and variable degrees of 'transferability' based on compatibility with the recipient genome and proteome, such as the Complexity Hypothesis ***(48)***. These and other considerations will have a strong influence on the type of network analysis that is performed, and the consequent interpretation of phylogenetic uncertainty and discordance.

5. Notes

1. The requirement that donor and recipient lineages be contemporary ignores the possibility that some donor lineages are not represented in the reference tree. The sampled genomes may be descendants of recipient organisms for which the corresponding donor is not available; absence from the data set may be due to a subsequent extinction event, or to incomplete sampling of extant lineages. A 'phantom donor' event may still be detectable, but the reconciling SPR operation will appear to merge the recipient lineage with one of its ancestors. Examples of this type of event are shown in Fig. 2 of MacLeod et al. ***(17)*** and Fig. 2 of Beiko and Hamilton ***(18)***. HorizStory explicitly allows such events in the reconciliation path, whereas EEEP excludes them by default but has a command-line argument that allows the relevant ancestor-descendant SPR operations to be performed.

2. The parsimony approach seeks to minimize the number of transfers in the reconciliation path, without the use of an explicit model of HGT. HGT probabilities can be modeled for specific pairs of lineages based on their G+C compatibility, shared ecological context, phylogenetic distance, or other factors ***(49,50)*** . Such weightings may be useful in breaking ties among many possible donor-recipient pairs, and may even identify cases in which longer edit paths should be favoured over shorter ones based on the probability of certain lineage pairs. It is unclear, however, how phylogenetic uncertainty and the phantom donor problem should be incorporated into models of HGT.

3. The greedy approach favors the recovery of major 'highways' of gene sharing among organisms, since donor-recipient pairs are concentrated wherever possible. A serious problem arises in the resolution of ties, when many possible pairs have the same frequency in the data set. The greedy approach could select one of these at random, but it would be worth evaluating decisions that result from many possible choices, to assess the robustness of the final set of edit paths that is returned.

4. The weighted approach shares out the inferred HGT events across all possible alternative pathways. A serious problem with this approach is its sensitivity to the completeness of taxonomic sampling in the relevant gene tree. If the gene tree has few of the taxa found in the reference tree, then many possible lineage pairs will be proposed to account for any phylogenetic discordance. Consequently, adding or subtracting taxa from the gene tree can affect the number of possible alternate pairs recovered, and can have a dramatic impact on their weighting.

5. While the number of inferred short-distance transfers is likely inflated by stochastic error, such transfers are plausible given the increased likelihood of shared vectors (viruses and plasmids) that can shuttle DNA between donor and recipient cells, and a higher propensity toward homologous recombination (although see ***(51)*** for potential selective constraints on short-range transfers of some informational genes). Tree-based methods for detecting HGT also *underestimate* the number of short-distance transfers due to their inability to detect transfers between lineages that are sisters in

the reference tree, since an SPR operation with sister taxa as donor and recipient will not modify the branching order of a tree.

6. References

1. Lawrence, J. G., Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**, 383-97.

2. Ragan, M. A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* **201**, 187-91.

3. Ragan, M. A., Harlow, T. J., Beiko, R. G. (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol* **14**, 4-8.

4. Ragan, M. A., Charlebois, R. L. (2002) Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *Int J Syst Evol Microbiol* **52**, 777-87.

5. Dagan, T., Martin, W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* **104**, 870-5.

6. Clarke, G. D. P., Beiko, R. G., Ragan, M. A., Charlebois, R. L. (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* **184**, 2072-80.

7. Lerat, E., Daubin, V., Moran, N. A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol* **1**, E19.

8. Beiko, R. G., Harlow, T. J., Ragan, M. A. (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* **102**, 14332-7.

9. Ge, F., Wang, L. S., Kim, J. (2005) The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol* **3**, E16.

10. Li, W., Jaroszewski, L., Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–3.

11. Harlow, T. J., Gogarten, J. P., Ragan, M. A. (2004) A hybrid clustering approach to recognition of protein families in 114 microbial genomes. *BMC Bioinformatics* **5**, 45.

12. Poptsova, M. S., Gogarten, J. P. (2007) BranchClust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics* **8**, 120.

13. Beiko, R. G., Chan, C.-X., Ragan, M. A. (2005) A word-oriented objective function for alignment validation. *Bioinformatics* **21**, 2230-9.

14. Allen, B. L., Steel, M. (2001) Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. Combinatorics* **5**: 1-15.

15. Hickey, G., Dehne, F., Rau-Chaplin, A., Blouin, C. (2008) SPR distance computation for unrooted trees, *Evolutionary Bioinformatics* **4**: 17-27.

16. Hallett, M., Lagergren, J. (2001) Efficient algorithms for lateral gene transfer problems. *RECOMB 2001*, p. 149-56.

17. MacLeod, D., Charlebois, R. L., Doolittle, W. F., Bapteste, E. (2005) Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol Biol* **5**, 27.

18. Beiko, R. G., Hamilton, N. (2006) Phylogenetic identification of lateral genetic transfer events. *BMC Evol Biol* **6**, 15.

19. Than, C., Ruths, D., Innan, H., Nakhleh, L. (2007) Confounding factors in HGT detection: statistical error, coalescent effects, multiple solutions. *J Comp Biol* **14**, 517-35.

20. Beiko, R. G., Ragan, M. A. (2008) Detecting lateral genetic transfer: a phylogenetic approach, in *Bioinformatics* (Keith, J. M., ed.), Humana, Totowa, NJ, in press.

21. Creevey, C. J., Fitzpatrick, D. A., Philip, G. K., Kinsella, R. J., O'Connell, M. J., Pentony, M. M., Travers, S. A., Wilkinson, M., McInerney, J. O. (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc Biol Sci* **271**, 2551-8.

22. Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., Papke, R. T. (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* **9**, 1099-1108.

23. Bryant, D., Moulton, V. (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**, 255-65.

24. Huson, D. H., Dezulian, T., Klopper, T., Steel, M. (2004) Phylogenetic super-networks from partial trees. *IEEE Trans. Comput. Biol. Bioinform.* **1**: 151-8.

25. Ragan, M. A. (1992) Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol* **1**, 53-8.

26. Nakhleh, L., Warnow, T., Linder, C. R., St. John, K. (2005) Reconstructing reticulate evolution in species - theory and practice, *J Comput Biol* **12**, 796-811.

27. Bordewich, M., Semple, C. (2007) Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Appl Math* **155**, 914-28.

28. Kurland, C. G., Canback, B., Berg, O. G. (2003) Horizontal gene transfer: A critical view. *Proc Natl Acad Sci USA* **100**, 9658-62.

29. Maddison, W. P. (1997) Gene trees in species trees. *Syst Biol* **46**, 523-46.

30. Inagaki, Y., Susko, E., Roger, A. J. (2006) Recombination between elongation factor 1α genes from distantly related archaeal lineages. *Proc Natl Acad Sci USA* **103**, 4528-33.

31. Chan, C.-X., Beiko, R. G., Ragan, M. A. (2007) A two-phase strategy for detecting recombination in nucleotide sequences. *South Africa Comp J* **38**, 20-7.

32. Hein, J. (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *J Mol Evol* **36**, 396-405.

33. Husmeier, D., McGuire, G. (2002) Detecting recombination with MCMC. *Bioinformatics* **18 Suppl 1**, S345-53.

34. Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O., Rogers, J. S. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol* **50**, 525-39.

35. Philippe, H., Delsuc, F., Brinkmann, H., Lartillot, N. (2005) Phylogenomics. *Annu Rev Ecol Evol Syst* **36**, 541-62.

36. Leigh, J. W., Susko, E., Baumgartner, M., Roger, A. J. (2008) Testing congruence in phylogenomic analysis. *Syst Biol* **57**, 104-15.

37. Singer, G. A. C., Hickey, D. A. (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**, 1581-8.

38. Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M., Nishikawa, K. (2003) Unique amino acid composition of proteins in halophilic bacteria. *J Mol Biol* **327**, 347-57.

39. Ho, S. Y., Jermiin, L. S. (2004) Tracing the decay of the historical signal in biological sequence data. *Syst Biol* **53**, 623-37.

40. Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J., Larkum, A. W. D. (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol* **53**, 638-43.

41. Lockhart, P. J., Steel, M. A., Hendy, M. D., Penny, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol* **11**, 605-12.

42. Delsuc, F., Phillips, M. J., Penny, D. (2003) Comment on "Hexapod origins: monophyletic or paraphyletic?" *Science* **301**, 1482.

43. Sullivan, J. and Swofford, D. L. (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J Mamm Evol* **4**, 77-86.

44. Wu, C. I. and Li, W. H. (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci USA* **82**, 1741–5.

45. Woese, C. R., Achenbach, L., Rouviere, P., Mandelco, L. (1991) Archaeal phylogeny: re-examination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artefacts. *Syst Appl Microbiol* **14**, 364–71.

46. E. Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E., Nesbø, C. L., Case, R. J., Doolittle, W. F. (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* **37**, 283-328.

47. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., Blankenship, R. E. (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science* **298**, 1616-20.

48. Jain, R., Rivera, M. C., Lake, J. A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**, 3801-6.

49. Galtier, N. (2007) A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst Biol* **56**, 633-42.

50. Beiko, R. G., Charlebois, R. L. (2007) A simulation test bed for hypotheses of genome evolution. *Bioinformatics* **23**, 825-31.

51. Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., Rubin, E. M. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**, 1449-52.

Fig. 1. Summarizing the relationships among a set of trees using bipartition and quartet compatibility. The trees in **1A** represent gene or 'test' trees inferred from 50 hypothetical orthologous data sets covering taxa A-F, with numbers below indicating the number of times each tree was recovered. A hypothetical reference tree (which happens to coincide with the most frequently observed topology) is shown in **1B** with concordant and discordant bipartitions mapped above and below the internal branches. For instance, the bipartition or split (ADE | BCF) is present in the first and third test trees, and is therefore present a total of 30 + 5 = 35 times in the data set. The second test tree exhibits an incompatible bipartition (ABC | DEF) which accounts for the other 15 cases. Panel **1C** shows a Lento plot of three distinct quartets (of the 15 possible sets of four taxa) onto the reference tree, with compatible topology frequencies shown in black above the 0 line and incompatible quartet frequencies aggregated below. The quartet BCEF always appears as (BC | EF) in the 50 test trees, and this topology is compatible with the reference, so a compatible proportion of 1.0 is displayed. Quartet ABDE is compatible with the reference as (AB | DE) in 90% of test trees, and the only incompatible alternative seen (AD | BE) is shown below the zero line. Finally, all three possible configurations of AFDE are observed in the set of test trees: the two incompatible versions (AD | EF) and (AE | DF) are represented together below the zero line.

Fig. 2. Super-networks constructed from two trees, covering the same set of taxa A-P. Both super-networks consist of one 'reference' tree as shown in **1A**, and one additional tree. In **1B** the second tree reflects a possible short-range transfer from B to the common ancestor of C and D (indicated with the dashed arrow in panel **A**), while the second tree

in **1C** reflects a transfer from an ancestor of taxon P to an ancestor of taxon A (dotted arrow in panel **A**). Networks were generated using the Z-closure method in SplitsTree version 4.8.