BIOINFORMATICS

Vol. 00 no. 00 2013
Pages 1–8

Systematic Inference of Highways of Horizontal Gene Transfer in Prokaryotes

Mukul S. Bansal [1,3], Guy Banay [1], Timothy J. Harlow [2], J. Peter Gogarten [2] and Ron Shamir [1*]

[1]The Blavatnik School of Computer Science, Tel-Aviv University, Israel. [2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, USA. [3]Currently at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Horizontal gene transfer (HGT) plays a crucial role in the evolution of prokaryotic species. Typically, no more than a few genes are horizontally transferred between any two species. However, several studies identified pairs of species (or linages) between which many different genes were horizontally transferred. Such a pair is said to be linked by a *highway of gene sharing*. Inferring such highways is crucial to understanding the evolution of prokaryotes and for inferring past symbiotic and ecological associations among different species.

Results: We present a new improved method for systematically detecting highways of gene sharing. As we demonstrate using a variety of simulated datasets, our method is highly accurate and efficient, and robust to noise and high rates of HGT. We further validate our method by applying it to a published dataset of over 22,000 gene trees from 144 prokaryotic species. Our method makes it practical, for the first time, to perform accurate highway analysis quickly and easily even on large datasets with high rates of HGT.

Availability: An implementation of the method can be freely downloaded from: `http://acgt.cs.tau.ac.il/hide`.

Contact: rshamir@tau.ac.il

1 INTRODUCTION

Horizontal gene transfer (*HGT*, also called lateral gene transfer) is an evolutionary process in which genes are transferred between two organisms that do not have an ancestor-descendant relationship. HGT is known to be rampant among prokaryotes, and plays an important role in their evolution and survival. An important problem in understanding microbial evolution is to infer the HGT events (i.e., the donor and recipient species of each HGT) that occurred during the evolution of a set of species. This problem is generally solved in a comparative-genomics framework by employing a parsimony criterion, based on the observation that the evolutionary history of horizontally transferred genes does not agree with the evolutionary history of the corresponding set of species. (This is illustrated in Figure S1 in the supplement.). More formally, given a gene tree and a species tree, the *HGT inference problem* is to find the minimum number of HGT events that can explain the incongruence of the gene tree with the species tree. The HGT inference problem is known to be NP-hard under most formulations (Hallett and Lagergren, 2001; Bordewich and Semple, 2005; Hickey *et al.*, 2008) and, along with some of its variants, has been extensively studied (Hallett and Lagergren, 2001; Boc and Makarenkov, 2003; Nakhleh *et al.*, 2004; Beiko *et al.*, 2005; Nakhleh *et al.*, 2005; Than *et al.*, 2007; Jin *et al.*, 2009; Boc *et al.*, 2010; Hill *et al.*, 2010).

In general, one expects only a few genes to have been horizontally transferred between any given pair of species. However, it has been observed that some pairs of species are connected by a multitude of HGT events. This can happen, for example, when a pair of species co-inhabit the same ecological niche over a long period of time (Boussau *et al.*, 2008; Zhaxybayeva *et al.*, 2009b), due to syntrophic or other close symbiotic relationships between partners (Martin and Muller, 1998; von Dohlen *et al.*, 2001; Overmann and Schubert, 2002; Lake, 2009), or because only the acquisition of a complete pathway can provide a selective advantage to the recipient (Lawrence and Roth, 1996; Igarashi *et al.*, 2001). Such pairs of species are said to be connected by a *highway of gene sharing* (Beiko *et al.*, 2005).[1] These highways point towards major events in evolutionary history; well corroborated examples are the uptake of endosymbionts into the eukaryotic host, and the many genes transferred from the symbiont to the host's nuclear genome (Gary, 1993). Recent proposals for evolutionary events that may be reflected in highways are the role of Chlamydiae in establishing the primary plastid in the Archaeplastida (red and green algae, plants and glaucocystophytes) (Huang and Gogarten, 2007; Becker *et al.*, 2008; Moustafa *et al.*, 2008), the evolution of double membrane bacteria through an endosymbiosis between clostridia and actinobacteria (Lake, 2009), and the high rate of transfer between marine Synecchococcus and Prochlorococcus (Zhaxybayeva *et al.*, 2006, 2009a).

In this work, we introduce a fast and accurate method for highway inference. Given a rooted species tree, any two species (nodes) in it that are not related by an ancestor-descendant relationship define a *horizontal edge* connecting those two nodes. Any HGT event must

*to whom correspondence should be addressed

[1] We point out that Beiko et al. used the term highways in a slightly different way than we do: For Beiko et al., highways could exist not only between two species but also between two clades of the species tree, i.e., their highways are formed by a collection of many individual HGTs spanning two clades.

take place along a horizontal edge in one of its two directions. A horizontal edge along which an unusually large number of HGT events have taken place will be called a *highway of gene sharing* or simply a *highway*. While it is hard to give a precise threshold that would designate an edge as a highway, any horizontal edge that contains a markedly higher number of HGTs than would be expected by chance for that dataset can be viewed as a highway. In our simulation studies we defined highways to be those horizontal edges that affect at least 5% of the genes with a history of HGT. Beiko *et al.* (2005) were the first to study the problem of highway inference (but see also Kunin *et al.* 2005). They employed a straightforward approach: Given a set of gene trees and a trusted species tree, they computed the history of HGT events for each gene tree separately and then combined the results across all gene trees. The horizontal edges that are inferred in the HGT scenarios for a significant fraction of the gene trees are the postulated highways. Thus, their solution relies on inferring the individual HGT events by solving the HGT inference problem. This is problematic for several reasons: First, the HGT inference problem itself is NP-hard under most formulations, and thus, difficult to solve exactly. Second, there are often multiple (and sometimes exponentially many) optimal solutions to the HGT inference problem (Beiko *et al.*, 2005; Than *et al.*, 2007). And third, when the gene tree has only a subset of the taxa present in the species tree, the placement of the inferred HGTs on the species trees becomes ambiguous. Furthermore, when the rate of HGT is relatively high, the number of HGT events need not be parsimonious; i.e., the HGT inference problem, even if solved exactly and yielding only one optimal solution, may not infer the actual HGT events. HGT events can also be inferred using a more sophisticated reconciliation framework that explicitly accounts for gene duplication and loss events in addition to HGT events (Tofigh *et al.*, 2011; Doyon *et al.*, 2010). Since such a framework is able to explicitly consider duplication and loss events, it can handle gene trees that have multiple gene copies per species; in the approach of Beiko et al., as well as in our approach, gene trees are restricted to contain at most one copy of a gene per species. However, the same drawbacks (discussed above) that affect the HGT inference problem apply to the more sophisticated framework as well (Tofigh *et al.*, 2011).

Recently, we introduced a novel polynomial time algorithm to detecting highways that bypasses the need to infer individual HGT events (Bansal *et al.*, 2011). The method is based on the observation that highways, by definition, affect the topologies of many gene trees. Thus, the idea is to combine the phylogenetic signals for HGT events from all the gene trees and use the combined signal to infer the highways. This is achieved by employing quartet decomposition: The method decomposes each gene tree into its constituent set of quartet trees and combines the quartet trees from all the gene trees to obtain a single weighted set of quartet trees. The combined set of quartet trees is then analyzed against the given species tree to infer the highways. The intuition is that quartet trees that disagree with the species tree may indicate HGT events and thus the collective evidence from all quartet trees could pinpoint possible highways.

In this work, we propose an alternative method based on quartet decomposition to detect highways which greatly improves upon the accuracy, noise-tolerance, and applicability of the method of Bansal *et al.* (2011). Our method differs from the approach of Bansal *et al.* (2011) in many important ways: (i) we analyze the quartet decomposition of each gene tree separately before combining the results across all gene trees, (ii) we propose a new way to assign scores to horizontal edges, which is sensitive to the direction of transfer of individual HGT events, and (iii) we show how to reduce the amount of noise in the computed scores in order to pinpoint highways even more accurately. As we demonstrate using extensive simulations, compared to the approach of Bansal *et al.* (2011), our new method is significantly more accurate, much more robust to high rates of HGT and noise, and is able to seamlessly handle gene trees with many missing taxa. For example, in simulated 50-taxon datasets with 1000 genes, a highway of 100 genes, and 2000 random HGTs ("noise"), the new method identifies the correct highway in 98% of the cases, compared to less than 30% for the old method. It is worth mentioning that quartets have been previously used for phylogenetic analyses in other contexts; for example, for phylogeny construction (Strimmer and von Haeseler, 1996), supertrees (Snir and Rao, 2010), or analysis of HGTs (Zhaxybayeva *et al.*, 2006).

Two methods, EEEP (Beiko *et al.*, 2005; Beiko and Hamilton, 2006) and Prunier (Abby *et al.*, 2010), exist for inferring HGT events on unrooted gene trees. These methods were not designed for inferring highways but can be used indirectly for doing so (as in (Beiko *et al.*, 2005)). Our quartet-based approach offers important advantages over approaches based on inferring HGT events: (i) We do not face the problem of dealing with multiple optimal solutions for the HGT inference problem, (ii) we can seamlessly incorporate gene trees with only a subset of taxa, which is difficult to do effectively with HGT inference methods, and (iii) our method is significantly more scalable and time-efficient than both EEEP and Prunier. In general, inferring highways based on HGT inference methods can be slow and technically complex. Our method simplifies the process of detecting and inferring highways. As our simulation study shows, our method is also very accurate.

We also applied the method to a dataset of 144 taxa and 22430 gene trees from Beiko *et al.* (2005). Our results are largely consistent with previous analyses of this dataset, and the entire computational analysis of this very large dataset took less than two days (using a single CPU). Our new method thus makes it possible to easily, quickly, and accurately infer highways even for very large datasets as well as on datasets with high rates of HGT. We have implemented our method into a freely available software package called HiDe (short for Highway Detection). HiDe is, to the best of our knowledge, the only available software package designed specifically to address the problem of inferring highways.

2 BASIC DEFINITIONS AND NOTATION

We follow the basic definitions and notation from Bansal *et al.* (2011). Given a rooted or unrooted tree T, we denote its node, edge, and leaf sets by $V(T)$, $E(T)$, and $Le(T)$ respectively. Given a rooted tree T, the root node of T is denoted by $rt(T)$, the parent of a node $v \in V(T)$ is denoted by $pa_T(v)$, its set of children by $Ch_T(v)$, and the (maximal) subtree of T rooted at v by $T(v)$. We define $\leq_T$ to be the partial order on $V(T)$ where $u \leq_T v$ if v is a node on the path between $rt(T)$ and u. Throughout this work the term tree refers to a binary tree.

Given a rooted tree T, a *horizontal edge* on T is a pair of nodes $\{u, v\}$, where $u, v \in V(T)$, such that $u, v \neq rt(T)$, $u \not\leq_T v$, $v \not\leq_T u$, and $pa_T(u) \neq pa_T(v)$. We denote by $H(T)$ the set of all horizontal edges on T. Horizontal edges represent potential HGT events; the (directed) horizontal arc (u, v) represents the HGT

event that transfers genetic material from the species represented by edge $(pa_T(u), u)$ to the species represented by edge $(pa_T(v), v)$. Thus, the (undirected) horizontal edge $\{u, v\}$ represents the HGT events (u, v) and (v, u). We define horizontal edges to be undirected because highways can be responsible for transfer of genetic material in both directions (we elaborate more on this later). Note that our approach does not use a timed reference tree and therefore allows for the detection of HGT events from extinct or unsampled lineages. Only HGT events from extinct or unsampled lineages that would lead to transfers between ancestor and descendant are excluded.

A *quartet* is a four-element subset of some leaf set and a *quartet tree* is an unrooted tree whose leaf set is a quartet. The quartet tree with leaf set $\{a, b, c, d\}$ is denoted by $ab|cd$ if the path from a to b does not intersect the path from c to d. Given a rooted or unrooted tree T, let X be a subset of $Le(T)$ and let $T[X]$ denote the minimal subtree of T having X as its leaf set. We define the *restriction* of T to X, denoted $T|X$, to be the unrooted tree obtained from $T[X]$ by suppressing all degree-two nodes (including the root, if T is rooted). We say that a quartet tree Q is *consistent* with a tree T if $Q = T|\,Le(Q)$, otherwise Q is *inconsistent* with T. Thus, for example, in Fig. S2 in the supplement, the quartet tree $ab|ce$ is consistent with the species tree and with the Gene-1 tree but inconsistent with the Gene-2 tree. Observe that, given any T and any quartet $X = \{a, b, c, d\}$ from $Le(T)$, X induces exactly one quartet tree in T, that is, the quartet tree $T|X$. Also observe that this quartet tree must have one of three possible topologies: $ab|cd$, $ac|bd$, or $ad|bc$.

3 DETECTING HIGHWAYS

We wish to detect the highways of gene sharing in the evolutionary history of a set of species $\mathbb{S}$. The input to our method is a set of unrooted gene trees $G_1, \ldots, G_k$, and a rooted species tree S on $\mathbb{S}$. The *highway detection problem* can thus be stated as follows: Given a species tree S and a collection of gene trees, find the horizontal edges on S that are most likely to correspond to highways. The idea is to infer highways by inspecting the differences in the topologies of the gene trees compared to the species tree.

As in Bansal *et al.* (2011), our solution to the highway detection problem is based on decomposing each input gene tree T into its constituent set of $\binom{|Le(T)|}{4}$ quartet trees. The basic idea is that, in general, different HGT events produce gene trees with different sets of inconsistent quartet trees. Thus, given an incongruent gene tree and the species tree, one can infer the HGT events responsible for the incongruence by looking at the set of inconsistent quartets. An example is depicted in Figure S2 in the supplement.

3.1 Description of the method

Our method uses quartet decomposition to calculate, for each gene tree, its support for the different horizontal edges on S. For each horizontal edge, the support values are then aggregated across all gene trees to obtain the total support for that horizontal edge. In contrast, the method of Bansal *et al.* (2011) first combines the quartet decomposition from each gene tree into a single weighted set of quartet trees and then computes the total support for each horizontal edge based on that single weighted set. Considering the quartet decomposition for each gene tree separately allows us to take into account the direction of transfer of any individual gene transferred along a horizontal edge, and also allows for the proper handling of gene trees having missing taxa. Formally, the method is as follows:

1: For each input gene tree G_i, for $1 \leq i \leq k$,

- **1(a):** Decompose G_i into its constituent set Φ_i of $\binom{|Le(G_i)|}{4}$ quartet trees.
- **1(b):** Remove from Φ_i all quartet trees that are consistent with S or that can be explained by a previously inferred highway.
- **1(c):** For each horizontal edge $\{u, v\} \in H(S)$, compute the normalized score $NS(\{u, v\}, \Phi_i)$. This step is explained in more detail below.

2: For each horizontal edge $\{u, v\} \in H(S)$, compute its final score, denoted $score\{u, v\}$, to be $\sum_{i=1}^{k} NS(\{u, v\}, \Phi_i)$.

3: Select the highest scoring horizontal edge as a highway.

By iterating this procedure several times, multiple highways can be found. The normalized score $NS(\{u, v\}, \Phi_i)$ captures the support of the gene tree G_i for an HGT event along the horizontal edge $\{u, v\}$, adequately corrected for the relative impact of the evidence for that edge from Φ_i. This is done as follows. We first compute the raw scores $RS((u, v), \Phi_i)$ and $RS((v, u), \Phi_i)$ of the two HGT events (u, v) and (v, u) that constitute $\{u, v\}$. The *raw score* of HGT event (u, v) with respect to Φ_i, denoted $RS((u, v), \Phi_i)$, is defined to be the number of quartet trees from Φ_i (after Step 1(b)) that can be explained by the HGT event (u, v). Thus, the raw score of an HGT event captures the number of quartet trees, from the gene tree under consideration, that support that HGT event. However, not all horizontal gene transfers would affect the same number of quartets from the gene tree under consideration. (For instance, in the example from Figure S2, the HGT event (C, E) causes four of the quartet trees in the corresponding gene tree to become inconsistent, while the HGT event (b, c) that transfers Gene-2 causes ten of the quartet trees in the gene tree to become inconsistent.). To overcome this bias, we modify the raw score of each HGT event by dividing it by a normalization factor: the maximum number of distinct quartet trees that could be explained by that HGT event. More precisely, let Ψ_i be the set of all possible quartet trees on the leaf set $Le(G_i)$. Given an HGT event (u, v), let Q_i denote the set of quartet trees in Ψ_i that can be explained by that HGT event. The normalization factor for (u, v), with respect to Φ_i, is defined to be $|Q_i|$. The normalized score of HGT event (u, v) with respect to Φ_i is denoted by $NS((u, v), \Phi_i)$. The normalized score of the horizontal edge $\{u, v\}$ with respect to Φ_i is defined as $NS(\{u, v\}, \Phi_i) = \max\{NS((u, v), \Phi_i), NS((v, u), \Phi_i)\}$. The intuition here is that we expect any given gene to have been transferred along only one of the two HGT events along that highway.

Thus, our method computes, for each horizontal edge, the sum of the normalized support scores from each input gene tree, and reports the horizontal edge with the highest total score as a highway edge. In contrast with the approach of Bansal *et al.* (2011), which normalizes the computed raw scores only after the data from all the gene trees have been combined, our method normalizes the score for each gene tree separately before combining the scores. We will refer to the original method of Bansal *et al.* (2011) as the *Global Normalization* method, and the current method as the *Per-Gene Normalization*

(PG-Norm) method. Our PG-Norm method is more sensitive to the direction of transfer of any individual gene and, furthermore, makes it possible to use gene trees with many missing taxa.

Observe that the value $NS(\{u,v\},\Phi_i)$, for any given horizontal edge $\{u,v\}$ and any gene tree G_i, must lie between 0 and 1. The higher the value of $NS(\{u,v\},\Phi_i)$, the higher the support from gene tree G_i for an HGT along $\{u,v\}$. Note, however, that a given inconsistent quartet tree could be explained by several different HGT events on S. Thus, even if the gene corresponding to gene tree G_i was not transferred along $\{u,v\}$, the value of $NS(\{u,v\},\Phi_i)$ need not be zero since (u,v) or (v,u) could explain some of the inconsistent quartet trees from other HGT events. Such ambiguity (or noise) may confound the true solution. One possible way to overcome this noise is to ignore or to decrease the weight of smaller $NS(\cdot,\cdot)$ values. To test this, we developed two variants of the regular PG-Norm method. In the first variant, we modify Step 2 of the method by setting $score\{u,v\}$ to be $\sum_{i=1}^{k}(NS(\{u,v\},\Phi_i))^p$, for some small $p \geq 1$ (in our experiments we used $p = 2, 4$). We refer to this variant as the *PG-Norm-Exp(p)* method. The regular PG-Norm method is the same as PG-Norm-Exp(1). In the second variant, we modify Step 2 of the method by including only those terms in the sum for $score\{u,v\}$ whose value is at least equal to some cutoff value p, for $0 \leq p \leq 1$. We refer to this variant as the *PG-Norm-Cutoff(p)* method. The regular PG-Norm method is the same as PG-Norm-Cutoff(0).

3.2 Complexity and running time

We rely on the algorithms of Bansal *et al.* (2011) for quartet decomposition of gene trees and for computing the raw and normalized scores of all HGT events on the species tree with respect to any given set of quartet trees. These algorithms can be directly used for implementing the PG-Norm method (and its variants) as follows. Let $G_1,\ldots,G_k$ denote the collection of input gene trees, S denote the species tree, and n denote the number of taxa in S. Consider the time complexity of Steps 1(a) to 1(c) of the PG-Norm method, for any given gene tree G_i: Based on the algorithms from Bansal *et al.* (2011), Steps 1(a) and 1(b) can be implemented to run in $O(\binom{|Le(G_i)|}{4})$ time, and Step 1(c) requires $O(\binom{|Le(G_i)|}{4} + n^2)$ time. Step 2, for any of the PG-Norm variants, requires $O(kn^2)$ time. Thus, each iteration of the PG-Norm method (or its variants) requires $O(\sum_{i=1}^{k}(\binom{|Le(G_i)|}{4} + n^2))$ time overall.

In practice, the fast algorithms for PG-Norm make it possible to analyze datasets with hundreds of taxa and thousands of gene trees. For example, we can analyze datasets with 1000 input gene trees each and having 50, 100, and 200 taxa, in less than 15 minutes, 5 hours, and 3 days respectively. In comparison, the method of Bansal *et al.* (2011) (i.e., the Global Normalization method), on those same datasets, takes about 2 minutes, 1 hour, and 13 hours respectively. The drastic improvement in the solution quality of our method thus comes at the expense of only a 5–7 fold reduction in speed. This reduction in speed occurs because our method must compute the scores for each horizontal edge separately for each gene tree, while the Global Normalization method does so only once with the combined set of quartet trees. To compare the running time, we also ran two HGT detection programs, EEEP (Beiko and Hamilton, 2006) (in its fastest setting) and the program Prunier (Abby *et al.*, 2010), on 10 randomly chosen datasets of 50 and 100 taxa each (with 1000 gene trees, 4000 noise level), and the results are given in Table S1 in the supplement. The experiments show that our method is dramatically faster and significantly more scalable than both EEEP and Prunier. All timed experiments were run on a single core of an Intel Xeon 5160 CPU running at 3 GHz with 8 GB of RAM. The memory requirements of our algorithms are also very low since they can be controlled by partitioning the set of quartet trees from any gene tree into subsets that fit into the cache memory, computing the support for each subset to each horizontal edge, and summarizing the results.

3.3 Dealing with uncertainty in gene tree topologies

Our method makes it easy to deal with uncertainty in gene tree topologies: Each gene tree may be represented by a collection of possible trees (e.g., bootstrap replicates or samples from a Bayesian posterior distribution) and only those quartet trees that are supported by at least a certain fraction (say 70%) of the trees are included in the quartet decomposition for that gene tree. This ability to deal cleanly and robustly with phylogenetic uncertainty is one of the key strengths of our method.

4 RESULTS

4.1 Performance evaluation

We used simulations to test the performance of the algorithm in various scenarios. In the basic simulation setup, each simulated dataset consisted of a random species tree on 50 taxa generated under a Yule process using the tool TreeSample (Hartmann *et al.*, 2010), and 1000 gene trees generated as follows. We randomly chose a highway on the species tree, and randomly assigned 10% of the 1000 genes as having been transferred along this highway, with equal probability for each transfer direction. Next, we simulated "noise" as additional single-gene HGT events. For each event, the horizontal edge and direction were selected randomly and independently, and the affected gene was selected at random. Selection was done with replacement, from the set of all gene trees (including those genes that were transferred on the highway). We varied the level of noise from 0 to 6000 random HGTs, in increments of 500. For each noise level, we created 50 different datasets (different species trees) and in each set computed the scores of all horizontal edges and the rank of the implanted highway among them. These simulation parameters correspond well to real data (see Section 5).

4.1.1 Comparison of PG-Norm and the Global Normalization algorithm. We first tested the ability of the different PG-Norm variants to correctly recover implanted highways in datasets generated using the basic simulation setup. For PG-Norm-Exp(p) we performed preliminary experiments with $p = 2, 4$ and observed that PG-Norm-Exp(2) greatly outperformed PG-Norm-Exp(4). For PG-Norm-Cutoff(p) we tried $p = 0.2, 0.4, 0.6, 0.8, 0.95$ and observed that PG-Norm-Cutoff(0.4) gave the best performance (see Figs. S3 and S4 in the supplement). We therefore focused on the best parameter value for each variant.

Fig. 1 shows the relative performance of the Global Normalization method of Bansal *et al.* (2011), PG-Norm, PG-Norm-Exp(2), and PG-Norm-Cutoff(0.4) on the simulated datasets. It is immediately obvious that PG-Norm and its variants dramatically outperform Global Normalization, both in terms of accuracy and tolerance to

noise; for instance, at a noise level of 2500 HGTs, Global Normalization is able to detect the implanted highway as its top-scoring edge in only 6% of the instances, while for PG-Norm and PG-Norm-Exp(2) the corresponding numbers are 88% and 100% respectively (Fig. 1(a)). Among PG-Norm variants, PG-Norm-Exp(2) shows the best performance overall (Fig. 1). Indeed, even with very high levels of noise, PG-Norm-Exp(2) is able to detect the implanted highway among its top few highest-scoring edges. We thus chose PG-Norm-Exp(2) to be our default method and all our subsequent experiments, both on simulated as well as real data, were performed using PG-Norm-Exp(2).

Note that, for very high levels of noise, our method does not always rank the implanted highway as the highest scoring (Fig 1a). This suggests that, in datasets with very high levels of HGT, it could be misleading to blindly interpret the highest scoring edge to be a true highway. Still, as our results on the top-five highest scoring edges show (Fig 1b), the true highway is usually ranked close to the top. Indeed, the average ranks of the implanted highway in the datasets with 4000 and 5000 HGTs are 1.34 and 4 respectively. This can be extremely valuable when trying to infer highways in such difficult datasets, since it can provide independent evidence to either support or refute highways inferred by applying other methods to the dataset. It also makes it possible to narrow down the list of candidate highways to just the top few highest scoring edges, which can then be investigated in more detail using other biological knowledge.

It is worth mentioning that the noise tolerance of our method improves as the number of taxa in the dataset increases. For instance, the performance of our method on datasets with 75 taxa and 7500 random HGTs is better that its performance on the 50 taxa datasets with 6000 random HGTs (results not shown). This is not unexpected, since the same number of HGT events, when applied to a larger tree, would be spread more thinly.

To assess the effects of phylogenetic reconstruction inaccuracy on its performance, we also applied our method to datasets consisting of gene trees reconstructed from simulated sequence data, and observed that the accuracy of our method was only minimally affected. Further details appear in Section S.1 in the supplement. We also compared against the program EEEP (Beiko and Hamilton, 2006) and observed that it was effective at inferring highways on these datasets in spite of over a third of the EEEP runs terminating without a valid solution. Further details appear in Section S.4.

4.1.2 Highway width. Next, we tested the impact of HGT abundance and noise on the ability to infer highways of different widths. The *width* of a highway is the fraction of genes transferred along it. Fig. 2 shows the results of our method for different highway widths and different noise levels. Once again, we see that our method is highly effective at accurately detecting large and medium sized highways (15% and 10% of genes transferred) in datasets with even high levels of noise. For example, at 10% width and a noise level of 4500 random HGTs, our method detected the implanted highways as the highest scoring edge in 70% and among the top five edges in 92% of the test cases. Performance improves even further when the implanted highways are wider (15% transferred genes). The results also demonstrate the ability of our method to detect small highways (5% of genes transferred) in datasets with fairly high levels of noise; at a noise level of 2500 random HGTs we detected the implanted small highways as the highest scoring edge in 60%, and among the top five edges in 92% of the test cases.

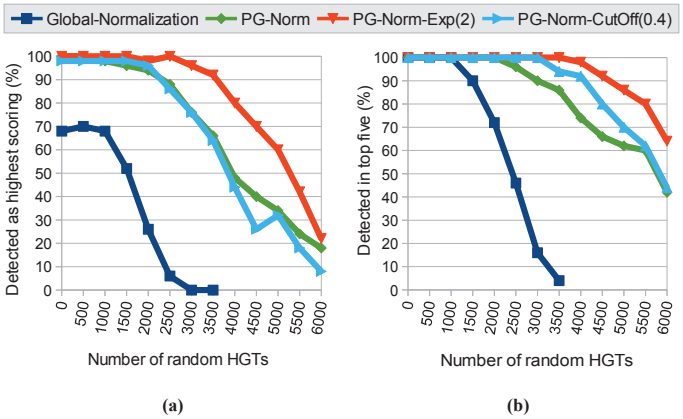

Fig. 1. Performance comparison of Global Normalization, PG-Norm, PG-Norm-Exp(2), and PG-Norm-Cutoff(0.4). (a) The percentage of times an implanted highway is detected as the highest-scoring edge in simulated datasets with varying levels of noise. (b) The percentage of times an implanted highway is detected among the top five highest-scoring edges in those same datasets. Results are based on 50 simulations for each noise level.

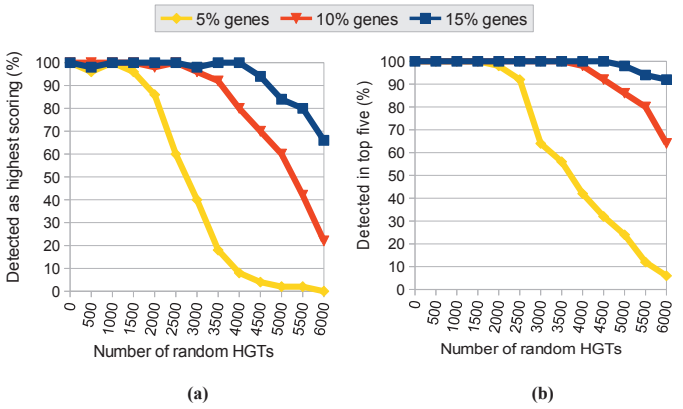

Fig. 2. Detecting implanted highways of different widths. (a) The percentage of times an implanted highway is detected as the highest-scoring edge in simulated datasets with varying levels of noise and highway widths. (b) The percentage of times an implanted highway is detected among the top five highest-scoring edges in those same simulated datasets.

4.1.3 Incomplete gene trees. In practice, one may encounter datasets where many of the gene trees do not contain genes from all the species considered in the analysis. To test the effect of such incomplete gene trees on highway inference, we modified the datasets from our basic simulation to create datasets in which the gene trees had only a fraction of their original leaf set. Specifically, we created three datasets where we restricted each gene tree to 40, 30, and 20 out of its 50 leaves respectively; and a fourth dataset where the gene tree sizes were normally distributed with mean 30 and st. dev. 7. In each case, the leaves to be removed from any gene tree were chosen randomly. Not surprisingly, performance decreases as the number of missing leaves increases. Still, as Fig. S5 in the supplement shows, our method can be effective even for datasets where most of the gene trees have a high fraction of missing leaves when the rate of HGT and/or noise is not very high. For example, at a

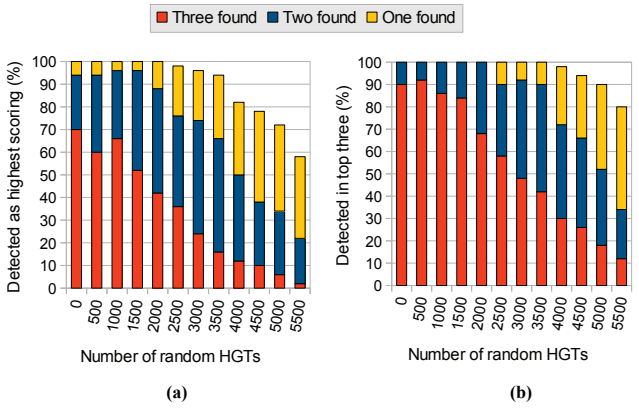

Fig. 3. Detecting multiple highways (width 10%). (a) Percentage of times three, two, or one of the implanted highways were correctly detected when only the highest scoring edge from each iteration is considered. (b) Results when the top three edges from each iteration are considered.

noise level of 2000 HGTs our method detects the implanted as the highest scoring edge in 100% and 42% of the cases in the first two datasets, respectively, and in 42% of the cases in the fourth dataset. Performance deteriorates significantly for the third dataset, in which the gene trees contain only 20 leaves each. We point out that our incomplete gene trees, with leaves pruned at random, present a greater challenge to our quartet based algorithm than incomplete gene trees with leaves drawn from a particular species tree clade.

4.1.4 Multiple highways. We went on to test the ability of our method to detect multiple highways. We used the basic simulation setup, except that we implanted three highways, each responsible for transferring randomly chosen 10% of the genes (a gene may be transferred on several highways). For each dataset we checked (i) how many of the three implanted highways were detected as the highest scoring edge in three iterations of the algorithm, and (ii) how many of the three implanted highways were detected among the top three edges from each iteration. To run the second and third iterations of the algorithm we removed the quartet trees corresponding to the highest scoring edge from each previous iteration (even if that edge was not the implanted highway). While detecting multiple highways is harder, the results, shown in Fig. 3, demonstrate the effectiveness of our method in this scenario. For instance, at a noise level of 3500 random HGTs and considering the top three edges from each iteration, we always find at least one implanted highway, find at least two implanted highways 90% of the time, and all three highways in over 40% of the cases. We also repeated the above experiment with wider 15% highways; as expected, performance improves with wider highways (see Fig. S6 in the supplement).

4.1.5 Number of highways in a dataset. Since the number of highways in a dataset is not known a priori, how does one decide if the highest scoring edge does in fact represent a highway? This can be done by studying a histogram of the edge scores reported by the method. As we demonstrate later in our analysis of the biological dataset, highways tend to be well separated from the rest of the edges in terms of their edge scores (see, e.g., Fig. S8 in the supplement). This makes it possible to infer the presence and estimate the number of highways. A similar pattern is observed in the simulated datasets, as illustrated in Fig. S7 in the supplement.

4.2 An application to biological data

We applied our method to a large biological dataset of prokaryotes studied by Beiko *et al.* (2005). The dataset contains 144 taxa and 22430 gene trees. To deal with uncertainty in gene tree topologies, we represented each of the 22430 gene trees by a collection of 100 trees sampled from a Bayesian posterior distribution. For each gene family we then considered only those quartet trees that were present in at least some predefined fraction of the 100 representatives. Thus, we only used highly supported quartet trees in our analysis. We used the 16S rRNA tree of Beiko *et al.* (2005) as our species tree.

Most of the gene trees in the dataset had very few leaves. For example, among the 22430 gene trees, only 130 had 100 taxa or more, 396 had at least 50 taxa, and 1319 had 25 or more taxa. As we observed in our experiments with simulated incomplete gene trees, it is desirable, in general, to have gene trees with as many taxa as possible in the analysis. We therefore decided to use a cutoff value for the minimum number of leaves in the gene tree. To select the values of the quartet support cutoff and the minimum gene tree size cutoff, we ran our method using all possible combinations of the two parameters. We considered values of 80%, 70%, and 60% for the quartet support cutoff and 100, 50, 25, and 0 (i.e., no cutoff) for the minimum gene tree leaf set size. Table S2 in the supplement shows the results of this analysis. In general, results were remarkably consistent across the different quartet support cutoff values. Results were also largely consistent across the gene tree leaf set size cutoff values 100, 50, and 25. With cutoff value 0, over 20000 new gene trees are introduced into the analysis, and the results change, but even then two of the three highways were the same as the ones obtained with the higher cutoff values. Based on this analysis we chose a quartet cutoff value of 80% and gene tree leaf set size cutoff of 25 for subsequent analysis.

Even with the added time complexity of performing quartet decomposition on 100 sampled trees – instead of just one – per gene family, each iteration of our method (i.e., each highway inference) required less than 6 hours on the entire 22430 gene tree dataset (using a single core on a 2.33 GHz quad-core Xeon 5410 CPU with 16 GB of RAM). We also tried to run the HGT detection method EEEP (Beiko and Hamilton, 2006) used by Beiko *et al.* (2005), on just the 130 gene trees that had at least 100 taxa (with each gene tree being represented by just one bootstrap replicate), and using the fastest version of the heuristic search (we used the strict reference tree ratchet (Beiko and Hamilton, 2006)). The analysis still required over 75 hours and only 66 of the 130 runs terminated successfully, with the remaining runs crashing due to excessive memory requirements.

The presence of highways in a dataset can be inferred from the distribution of the reported horizontal edge scores. A histogram of the horizontal edge scores reported by the first iteration of the method on our biological dataset is given in Fig. S8 in the supplement. A vast majority of the horizontal edges receive a very low edge score (99.87% of the edges have a score lower than 20 and only 6 out of the 37,418 candidate horizontal edges have a score larger than 50), and the high scoring edges are well-separated from the remaining edges in terms of their scores (the highest scoring edge has a score of 83.3). By constructing histograms of the horizontal

edge scores for successive iterations of the algorithm, we inferred that the dataset consists of at least five highways. Figure 4 shows these top five inferred highways. Most of the highways identified by our approach correspond to ones also discussed in Beiko *et al.* (2005). Similar to these authors, we find most of the transfers inside the Gamma proteobacteria, e.g., Beiko et al. identified over 175 transfers between an ancestor of *Yersinia pestis* and a common ancestor of *Escherichia coli* plus *Salmonella* (fourth highway in Fig. 4). Beiko et al. comment that the number of transfers identified between close relatives is an underestimation because transfers between sister taxa do not generate a phylogenetic signal that can be detected using bipartitions (used in Beiko *et al.* 2005) or quartets. However, while many of the within species transfers between different *E. coli* and *Shigella* strains may correspond to actual gene flow within the species (Dykhuizen and Green, 1991), we think that some of the identified transfers within the gamma proteobacteria might also be due to uncertainty in the reference tree. For example, genes from the insect endosymbionts (Wigglesworthia and Buchnera) have many shared sequence characteristics that might have originated through convergent evolution and might not reflect shared ancestry. The transfer that repositions Buchnera and Wigglesworthia as a deeper branch (first highway in Fig. 4) likely is due to phylogenetic uncertainty and may not represent an actual highway of gene transfer. The same phylogenetic conflict was observed by Lerat *et al.* (2003): The tree after transfer corresponds to the maximum likelihood (ML) tree for concatenated proteins reported in Lerat *et al.* (2003); and our reference tree corresponds to the 16S ML tree in Lerat *et al.* (2003). Lerat *et al.* (2003) found that the difference between these two topologies was not significant for most protein families when analyzed using the SH-test (Shimodaira and Hasegawa, 1999).

Since all of the top highway candidates reported by our method were within the gamma proteobacteria, we also looked for highway candidates from outside of the gamma proteobacteria. The top scoring highways of gene sharing identified outside the gamma proteobacteria are shown in Figure S9 in the supplement, and include a highway between Synechococcus and ProchlorococcusMIT. The reference tree considers Prochlorococcus as monophyletic, a grouping that is supported by the many shared derived characteristics of Prochlorococcus, including the use of chlorophyll b as an antenna pigment in the light harvesting machinery, and the absence of phycobilisomes (see Zhaxybayeva *et al.* 2006 for extended discussion). This highway was also identified by Beiko *et al.* (2005) and appears to be due to continued gene transfer between marine Synecchococcus and low light adapted Prochloroccus ssp. (Zhaxybayeva *et al.*, 2009a), likely mediated by cyanophage (Zeidner *et al.*, 2005).

Fig. 4. Results on the dataset of Beiko et al. The top five highways,along with their ranks, computed by the method are marked in red (bold edges). The reported scores for these top five highways are 83.3, 52.5, 42.9, 35.1, and 24.3. Since the top five highways are each within the gamma proteobacteria, the figure focuses on only that portion of the phylogeny (we refer the reader to Figure S9 in the supplement for a figure showing the full phylogeny). The tree was drawn using Dendroscope (Huson *et al.*, 2007)

5 DISCUSSION

In this work, we presented a new and improved method for systematically detecting highways of gene sharing. As shown in the simulation study, our method is highly accurate, and robust to noise and high rates of HGT. Our analysis of the 144 taxa, 22430 gene tree dataset of Beiko *et al.* (2005) demonstrates the utility of our method in identifying potential highways, even in large datasets.

Our simulation parameters correspond well to real data. In the dataset of Beiko *et al.* (2005), only about 5000 HGTs, including several highways of size 100–500, are reported on the entire dataset of roughly 20,000 gene trees. Our experiments show (see Section S.2 in the supplement) that by distributing the same number of HGTs across more gene trees the performance of the method is unaffected and may even slightly improve. Our simulation scenarios have a comparable or smaller highway size (50–150 HGTs), a comparable or higher noise level (1000–6000 random HGTs), and less gene trees (1000). As the number of taxa that we use is smaller (50 vs 144), the random HGTs are less spread in our case, which again makes our scenario harder as we observed in Section 4.1.1. Hence, the simulations present a scenario that is as difficult or harder than the real data in each of the parameters. The numbers of genes in our highways correspond well to other highways characterized in the literature, e.g., in Thermotogales (Zhaxybayeva *et al.*, 2009b) and in Aquificales (Boussau *et al.*, 2008).

Highways represent major shifts from the pattern of vertical inheritance and their detection is crucial for correctly representing the evolutionary history of prokaryotes. Indeed, a logical first-step towards building a comprehensive phylogenetic network for the prokaryotes (Kunin *et al.*, 2005; Williams *et al.*, 2011) is to start with the ribosomal tree as backbone and augment it with highways. By making it easy to quickly and accurately detect such highways, our method represents an important step towards building the prokaryotic net of life.

As with the method of Bansal *et al.* (2011), our new method is based on quartet decomposition. However, we employ quartet decomposition in a different way and, as a result of this and

other methodological improvements, our method achieves far better results and greatly improves upon its accuracy, noise-tolerance, and applicability. Still, our method has several limitations. It makes use of the parsimony principle, and if a dataset contains two highways that are closely related to one another then the method may only detect one of them (since many of the inconsistent quartet trees from one highway may also support the other highway). Formulating the highway detection problem in a probabilistic framework could help improve the accuracy of highway detection. Also, a statistical analysis of highway and HGT score distributions could help provide more quantifiable significance of the computed highways, which we currently lack. Our method is unable to detect HGTs between two sister species (i.e., two species that have the same parent), a limitation shared by all existing phylogenetic methods for studying HGT (but see Tofigh 2009). Our analysis of the dataset of Beiko et al. suggests that, as with all previous methods, our method is vulnerable to errors in the species tree, and that such errors can manifest themselves as highways in the analysis. Extending the method to deal cleanly with uncertainty in the species tree topology would help to pinpoint highways more accurately. It would also be interesting to study how convergent gene losses in unrelated taxa (e.g., species that have independently converged to parasitic lifestyles) affect our method. Finally, it may help to distinguish between directed and undirected highways of gene sharing (see Section S.3 in the supplement).

Acknowledgements This study was supported in part by the Israel Science Foundation (Grant 802/08) and by the Raymond and Beverly Sackler Chair in Bioinformatics. MSB was supported in part by a postdoctoral fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. JPG was supported in part by National Science Foundation grant DEB 0830024 and a fellowship from the Fulbright Program.

REFERENCES

Abby, S., Tannier, E., Gouy, M., and Daubin, V. (2010). Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinf.*, **11**(1), 324.

Bansal, M. S., Banay, G., Gogarten, J. P., and Shamir, R. (2011). Detecting highways of horizontal gene transfer. *Journal of Computational Biology*, **18**, 1087–1114.

Becker, B., Hoef-Emden, K., and Melkonian, M. (2008). Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evol. Biol.*, **8**(1), 203.

Beiko, R. and Hamilton, N. (2006). Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.*, **6**(1), 15.

Beiko, R. G., Harlow, T. J., and Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *P. Natl. Acad. Sci. USA*, **102**(40), 14332–14337.

Boc, A. and Makarenkov, V. (2003). New efficient algorithm for detection of horizontal gene transfer events. In G. Benson and R. D. M. Page, editors, *WABI*, volume 2812 of *Lecture Notes in Computer Science*, pages 190–201. Springer.

Boc, A., Philippe, H., and Makarenkov, V. (2010). Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst. Biol.*, **59**(2), 195–211.

Bordewich, M. and Semple, C. (2005). On the computational complexity of the rooted subtree prune and regraft distance. *Annals of combinatorics*, **8**(4), 409–423.

Boussau, B., Gueguen, L., and Gouy, M. (2008). Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of bacteria. *BMC Evol. Biol.*, **8**(1), 272.

Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllosi, G. J., Ranwez, V., and Berry, V. (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In *RECOMB-CG*, pages 93–108.

Dykhuizen, D. E. and Green, L. (1991). Recombination in escherichia coli and the definition of biological species. *J. Bacteriol.*, **173**(22), 7257–7268.

Gary, M. W. (1993). Origin and evolution of organelle genomes. *Curr Opin Genet Dev*, **3**, 884–890.

Hallett, M. T. and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In *RECOMB*, pages 149–156.

Hartmann, K., Wong, D., and Stadler, T. (2010). Sampling trees from evolutionary models. *Syst. Biol.*, **59**(4), 465–476.

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

Hill, T., Nordstrom, K., Thollesson, M., Safstrom, T., Vernersson, A., Fredriksson, R., and Schioth, H. (2010). Sprit: Identifying horizontal gene transfer in rooted phylogenetic trees. *BMC Evol. Biol.*, **10**(1), 42.

Huang, J. and Gogarten, J. (2007). Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biology*, **8**(6), R99.

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, **8**(1), 460.

Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K., and Nagashima, K. V. (2001). Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *Journal of Molecular Evolution*, **52**, 333–341.

Jin, G., Nakhleh, L., Snir, S., and Tuller, T. (2009). Parsimony score of phylogenetic networks: Hardness results and a linear-time heuristic. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **6**(3), 495–505.

Kunin, V., Goldovsky, L., Darzentas, N., and Ouzounis, C. A. (2005). The net of life: Reconstructing the microbial phylogenetic network. *Genome Res.*, **15**(7), 954–959.

Lake, J. A. (2009). Evidence for an early prokaryotic endosymbiosis. *Nature*, **460**, 967–971.

Lawrence, J. G. and Roth, J. R. (1996). Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**(4), 1843–1860.

Lerat, E., Daubin, V., and Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: The case of the γ-proteobacteria. *PLoS Biol.*, **1**(1), e19.

Martin, W. and Muller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature*, **392**(6671), 37–41.

Moustafa, A., Reyes-Prieto, A., and Bhattacharya, D. (2008). Chlamydiae has contributed at least 55 genes to plantae with predominantly plastid functions. *PLoS ONE*, **3**(5), e2205.

Nakhleh, L., Warnow, T., and Linder, C. R. (2004). Reconstructing reticulate evolution in species: theory and practice. In P. E. Bourne and D. Gusfield, editors, *RECOMB*, pages 337–346. ACM.

Nakhleh, L., Ruths, D. A., and Wang, L.-S. (2005). RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. In L. Wang, editor, *COCOON*, volume 3595 of *Lecture Notes in Computer Science*, pages 84–93. Springer.

Overmann, J. and Schubert, K. (2002). Phototrophic consortia: model systems for symbiotic interrelations between prokaryotes. *Arch. Microbio.*, **177**, 201–208.

Shimodaira, H. and Hasegawa, M. (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol. Biol. Evol.*, **16**(8), 1114.

Snir, S. and Rao, S. (2010). Quartets maxcut: A divide and conquer quartets algorithm. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **7**(4), 704 –718.

Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, **13**(7), 964.

Than, C., Ruths, D. A., Innan, H., and Nakhleh, L. (2007). Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. *Journal of Computational Biology*, **14**(4), 517–535.

Tofigh, A. (2009). *Using Trees to Capture Reticulate Evolution : Lateral Gene Transfers and Cancer Progression*. Ph.D. thesis, KTH Royal Institute of Technology.

Tofigh, A., Hallett, M. T., and Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **8**(2), 517–535.

von Dohlen, C. D., Kohler, S., Alsop, S. T., and McManus, W. R. (2001). Mealybug β-proteobacterial endosymbionts contain γ-proteobacterial symbionts. *Nature*, **412**, 433–436.

Williams, D., Fournier, G., Lapierre, P., Swithers, K., Green, A., Andam, C., and Gogarten, J. P. (2011). A rooted net of life. *Biology Direct*, **6**(1), 45.

Zeidner, G., Bielawski, J. P., Shmoish, M., Scanlan, D. J., Sabehi, G., and Beja, O. (2005). Potential photosynthesis gene recombination between *prochlorococcus* and *synechococcus* via viral intermediates. *Environ. Microbiol.*, **7**(10), 1505–1513.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., and Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Res.*, **16**(9), 1099–1108.

Zhaxybayeva, O., Doolittle, W. F., Papke, R. T., and Gogarten, J. P. (2009a). Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. *Genome Biol Evol*, **1**, 325–339.

Zhaxybayeva, O., Swithers, K. S., Lapierre, P., Fournier, G. P., Bickhart, D. M., DeBoy, R. T., Nelson, K. E., Nesb, C. L., Doolittle, W. F., Gogarten, J. P., and Noll, K. M. (2009b). On the chimeric nature, thermophilic origin, and phylogenetic placement of the thermotogales. *P. Natl. Acad. Sci. USA*, **106**(14), 5865–5870.