Regular manuscript

Running head: SPR Supertrees

Supertrees based on the subtree prune-and-regraft distance

Christopher Whidden, Norbert Zeh and Robert G. Beiko†

Faculty of Computer Science, Dalhousie University

6050 University Avenue, PO Box 15000

Halifax, Nova Scotia, Canada B3H 4R2

†Corresponding author.

Tel: +1 902 494 8043

Fax: +1 902 492 1517

Email: beiko@cs.dal.ca

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18

Abstract.

Supertree methods reconcile a set of phylogenetic trees into a single structure that is often interpreted as a branching history of species. A key challenge is combining conflicting evolutionary histories that are due to artifacts of phylogenetic reconstruction and phenomena such as lateral gene transfer (LGT). Although they often work well in practice, existing supertree approaches use optimality criteria that do not reflect underlying processes, have known biases and may be unduly influenced by LGT. We present the first method to construct supertrees by using the subtree prune-and-regraft (SPR) distance as an optimality criterion. Although calculating the rooted SPR distance between a pair of trees is NP-hard, our new maximum agreement forest-based methods can reconcile trees with hundreds of taxa and > 50 transfers in fractions of a second, which enables repeated calculations during the course of an iterative search. Our approach can accommodate trees in which uncertain relationships have been collapsed to multifurcating nodes. Using a series of simulated benchmark datasets, we show that SPR supertrees are more similar to correct species histories under plausible rates of LGT than supertrees based on parsimony or Robinson-Foulds distance criteria. We successfully constructed an SPR supertree from a phylogenomic dataset of 40,631 gene trees that covered 244 genomes representing several major bacterial phyla. Our SPR-based approach also allowed direct inference of highways of gene transfer between bacterial classes and genera; a small number of these highways connect genera in different phyla and can highlight specific genes implicated in long-distance LGT.

Keywords: subtree prune-and-regraft, supertrees, phylogenomics, prokaryotic phylogeny, matrix representation with parsimony, lateral gene transfer, Robinson-Foulds

An organism's genome, typically comprising many thousands of genes, provides a detailed record of its past. While sets of homologous genes from a set of genomes can provide evidence about organismal relationships, individual gene trees covering these genomes may be influenced by processes including paralogy and gene loss, lineage sorting and lateral gene transfer (LGT) (Maddison and Knowles, 2006; Galtier and Daubin, 2008). One approach to reconcile trees that differ due to these processes and to artifacts of phylogenetic inference is to construct a single tree that aims to reflect the relationships in a set of gene trees. Supertree methods generate a single tree, which may serve as a hypothesis of organismal descent or relatedness, by optimizing a similarity criterion. Supertrees have been used to represent large-scale phylogenies including the first phylogeny of nearly all extant mammals (Bininda-Emonds et al. 2007), the first family-level phylogeny of flowering plants (Davies et al. 2004), and the first species-level phylogeny of non-avian dinosaurs (Lloyd et al. 2008). They have also been used to study the extent of LGT in prokaryotes (Beiko et al. 2005) and to disentangle the origin of eukaryotic genomes (Pisani et al. 2007). One key advantage of supertree methods is that they can take as input sets of gene trees sampled from overlapping but non-identical sets of taxa, in contrast with consensus tree approaches, which require that all input trees contain exactly the same set of leaves. Simulations have shown that supertrees are more reliable in the presence of a moderate amount of misleading LGT than the supermatrix approach which is based on concatenated alignments of many gene sequences (Lapierre et al. 2012).

Many optimality criteria have been proposed for supertree construction. Matrix representation with parsimony (MRP) (Ragan 1992; Baum 1992) was among the earliest methods proposed and remains the most commonly used, but detailed work with MRP has raised several concerns with the approach. MRP converts input trees into a binary character matrix and solves the parsimony problem on this matrix. Although the parsimony problem is NP-hard, fast hill-climbing heuristics in PAUP* or TNT allow MRP to be applied to large datasets (Goloboff 1999; Swofford 2002; Roshan et al. 2004). MRP is very effective in practice, quickly constructing supertrees of competitive quality in every tested metric (Bininda-Emonds et al. 2001; Eulenstein et al. 2004; Chen et al. 2006). However, it is not clear why the MRP approach performs so well and it may generate relationships that do not belong to any of the source trees (Pisani and Wilkinson 2002), has problems resulting from unequal representation of taxa (Bininda-Emonds et al. 2002), and may include relationships contradicted by the majority of source trees (Goloboff 2005). Other developed supertree criteria include consensus supertrees (Adams 1972), majority-rule supertrees (Cotton and Wilkinson 2007), Quartet supertrees (Piaggio-Talice et al. 2004) and Triplet supertrees (Lin et al. 2009). However, like MRP, other supertree building methods that are not based on symmetric tree-to-tree similarity measures may be unduly influenced by the shapes of the input trees (Wilkinson et al. 2005).

Bansal et al. (2010) recently proposed Robinson-Foulds (RF) supertrees, which aim to minimize the total RF distance (Robinson and Foulds 1981) between the supertree and the set of input trees. The RF measure captures the number of bipartitions in one tree that do not exist in another, so the RF supertree approach aims to maintain as much phylogenetic information from the input trees as possible. Fast hill-climbing heuristics make computing rooted RF supertrees feasible from binary input trees and others have begun to extend this to unrooted trees with local search heuristics (Chaudhary et al. 2012). While RF appears to be a good criterion for supertrees, it may not be suitable for datasets with substantial amounts of LGT: a single "long-distance" LGT event between distant taxonomic relatives will result in many discordant bipartitions and a high RF distance. If many organisms participate in long-distance LGT, then "phylogenetic compromise" trees (Beiko et al. 2008) may emerge which reflect neither the correct species relationships, nor the dominant pathways of gene sharing. The requirement that all input trees be binary is also potentially limiting, as many relationships in trees inferred from sequence data are unsupported by statistics such as the bootstrap, and should be collapsed into multifurcations.

Another well-studied criterion for expressing differences between trees is the subtree prune-and-regraft (SPR) distance (Hein et al. 1996). The SPR operation involves splitting a pendant subtree from the rest of the tree, and reattaching it at a different location, with the rooting of the subtree preserved. Since SPR operations allow the pruned subtree to be reattached anywhere, they can accommodate long-distance transfers in a single step; such a transfer would increase the SPR distance by only 1, whereas the RF distance could be drastically increased. The SPR distance is the minimum number of such operations required to reconcile two trees. The relationship between an SPR operation and the topological consequences of an LGT event (Beiko and Hamilton 2006) makes SPR a natural criterion for assessing a supertree whose constituent trees contain a large number of LGT events. Given its relationship with the RF distance, the SPR criterion may also be suitable for datasets where a phenomenon other than LGT is the principal confounding factor. To date, no SPR-based supertree approach has been developed, in part because computing the SPR distance between two phylogenetic trees is NP-hard (Bordewich and Semple 2005; Hickey et al. 2008).

Combining two recent advances makes SPR supertrees feasible. First, using the equivalence between Maximum Agreement Forests (MAFs) and rooted SPR distance (Hein et al. 1996; Bordewich and Semple 2005), Whidden and Zeh (2009) and Whidden et al. (2010; 2013a) developed an algorithm with running time $O(2.42^k n)$. The resulting implementation was orders of magnitude faster than any previous algorithm and is able to compute SPR distances of up to 46 on trees with 144 prokaryotic taxa, and 99 on synthetic 1000-leaf trees, in less than 5 hours. We have extended this algorithm with several enhancements that we believe improve the running time to $O(2^k n)$ for binary input trees, and allow the inclusion of input trees in which uncertain relationships have been collapsed into multifurcating nodes. Second, Linz and Semple (2011) developed a cluster reduction technique which can reduce the computation of an MAF into several subproblems, yielding an exponential reduction of the running time in practice. The approach taken by Linz and Semple is similar to the cluster reduction rule of Baroni et al. (2005) for computing the hybridization distance but requires more care in choosing which maximum agreement forest to take for each subproblem to build the complete MAF. We have also reduced the time required to compute a cluster reduction to linear from the originally published $O(n^3)$. Neither refinement alone is fast enough to compute the thousands of SPR distances required to build an SPR-based supertree on interesting numbers of taxa. However, by combining the cluster reduction with our improved MAF-based approach we obtain dramatic improvements in running time, processing tree pairs that previously required 1-5 hours to reconcile in one second or less, thus enabling the many SPR distance computations needed to iteratively construct a supertree.

Our heuristic approach uses a greedy hill-climbing strategy to build an initial supertree, then refines this supertree using iterative global SPR rearrangements. We use a bipartition-based heuristic to identify and ignore proposed rearrangements that violate relationships that are well-supported in many trees, greatly reducing the number of rearrangements that need to be evaluated. These algorithms are implemented in the SPR Supertree software version 1.2.0, which is available at http://kiwi.cs.dal.ca/Software/SPRSupertrees. The software is freely available, open source and licensed under the GNU GPL version 3. Here we describe the steps in our approach, and demonstrate the speedups achieved using the algorithmic refinements described above. Our experiments using simulated datasets with LGT show that the SPR approach is more accurate than RF and, for some realistic rates and regimes of LGT, MRP as well. Comparisons based on the eukaryotic datasets used by Bansal et al. (2010) for benchmarking show that the SPR approach yields supertrees with lower total SPR distances to the input trees than either RF or MRP, and with slightly higher RF and parsimony scores. To demonstrate the application of the SPR supertree approach on a dataset in which considerable LGT is expected, we also used a phylogenomic data set of 244 bacteria covering 393,876 genes in 40,631 orthologous sets to analyze preferential transfer of genes between bacterial lineages. We were able to reconstruct a highly plausible supertree, and with the SPR approach we identified putative highways of gene sharing. Interestingly, preference for alternative hypotheses of the relatedness between bacterial phyla depended on the choice of gene tree rootings, suggesting that unrooted supertree methods may be ignoring plausible hypotheses.

METHODS

Calculating the Subtree Prune-and-Regraft Distance Between a Pair of Rooted Trees

We can compute the SPR distance between a pair of rooted trees quickly in practice, despite the NP-hardness of the problem (Bordewich and Semple 2005), using our efficient fixed-parameter bounded search tree algorithm in combination with our linear-time formulation of Linz and Semple’s cluster reduction (Linz and Semple 2011) to solve the equivalent Maximum Agreement Forest (MAF) problem. The MAF problem is a static version of the SPR distance problem that is easier to manipulate and analyze. An agreement forest of two trees is a forest on the same label set that can be created by cutting (deleting) edges from either tree. Bordewich and

Semple (2005) showed that a maximum agreement forest—an agreement forest that requires the fewest edge cuts—requires exactly as many edge cuts as the SPR distance between the trees. Indeed, each edge cut represents a transfer and the proposed series of transfers can be quickly inferred from the MAF (Fig. 1). Our algorithms, like most recent work on the SPR distance, compute such MAFs.

Our published MAF algorithm (Whidden et al. 2010; Whidden et al. 2013) operates in a bottom-up fashion in the first tree, denoted T_1, and reduces the second tree to a forest, denoted F_2. During the algorithm we identify subtrees that are identical in T_1 and F_2 and, in particular, pairs of such trees that are siblings in T_1 (sibling pairs). If any identical subtree is a component of F_2 we cut its corresponding parent edge in T_1. If any sibling pair in T_1 is also a sibling pair of F_2 we note that their parent nodes are identical in T_1 and F_2. If neither of these two situations applies, we identify at most three possible edge cutting scenarios and explore each recursively. We explore each scenario in turn, thus using very little memory, and use our 3-approximation algorithm (which operates similarly but simply cuts all three possible edges so that its running time scales linearly and may return at most 3 times the correct distance) to avoid exploring scenarios that are guaranteed to not return an optimal MAF.

We have enhanced our MAF algorithm to prioritize non-branching edge cut scenarios and ignore duplicate search branches through *edge protection*. First, we examine each sibling pair to select a sibling pair with only one edge cutting scenario, if any exist. This limits the exponential explosion of our search when possible. Second, we *protect* edges that have been cut in previously rejected scenarios. If we have two scenarios that cut edges e_1 and e_2, respectively, and the e_1 scenario fails to find an MAF, then the e_2 scenario will not find an MAF by cutting e_1 so we *protect* e_1 to indicate this and ignore any scenario that would cut e_1. This prevents us from exploring duplicate edge sets and increases the chance of finding a non-branching edge cut scenario. When no non-branching sibling pairs remain, we select a sibling pair with a protected member, if possible, to capitalize on this effect. For further details see Appendix I.

We have also extended our MAF algorithm to allow for reconciliation of multifurcating gene trees with the reference supertree (see Appendix I). For such gene trees we define the *soft* SPR distance (Whidden et al. 2013b; Linz and Semple 2008) to be the minimum number of SPR operations required to transform the reference tree into some binary resolution of the gene tree. This definition accounts for the general assumption that multifurcations imply uncertainty rather than simultaneous speciation. Our algorithm proceeds similarly to the binary case (as the reference tree, required to be T_1, is binary) with modifications to our considered edge scenarios that allow the resolution of multiple siblings and cutting the resulting edge.

The cluster reduction of Linz and Semple (2011) splits the input trees into smaller subproblems that can be solved iteratively (but not independently). As our algorithms' running times scale exponentially with the computed distance, this reduction has an enormous impact in practice. Two subtrees of the input trees on the same leaf sets represent a cluster. A cluster MAF with its root edge removed (representing a transfer prior to the LCA of the leaf set) is guaranteed to be part of some complete MAF of the two trees, if any such cluster MAF exists. Alternatively, if every MAF of the cluster must maintain its root edge, every cluster MAF will be part of a complete MAF. We thus modified our search strategy to prefer MAFs with their root edge removed in order to accommodate this reduction. In addition, we removed the complicated weighting scheme of the original cluster reduction method and improved the time required to compute such a cluster reduction to linear in the size of the trees from the cubic scaling reported by Linz and Semple (see Appendix II).

Recently, Chen and Wang proposed a separate improvement to our previous SPR distance algorithm for binary trees called UltraNet (Chen and Wang 2013). We do not compare our algorithms with UltraNet in detail as UltraNet requires binary trees and failed to find the correct SPR distance in 30 of our tests. However, our improved algorithm for the SPR distance even without the cluster reduction was significantly faster than UltraNet and our previous algorithm with clustering outperformed UltraNet on 65 of our tests.

Supertree Construction

We attempt to find the minimal SPR supertree for a given set of gene trees, that is, the binary rooted tree on the union of the label sets of the gene trees with the minimal cumulative SPR distance to the gene trees (hereafter, simply minimal SPR distance). When the leaf set of the (partially constructed) supertree differs from that of a gene tree, we ignore unique taxa when computing this distance. If no starting tree is provided to initiate the search, we construct an initial SPR supertree through stepwise addition of taxa and then use global SPR rearrangements to optimize the tree. To construct the initial tree, we begin with the four most common taxa in the input trees and select the tree shape on these four taxa with minimal SPR distance to the projected input trees. We then successively add taxa to the supertree, in decreasing order according to the frequency of occurrence in the gene trees. Each taxon is added in the location that minimizes the SPR distance. When determining this location, we only compute the SPR distance to gene trees containing the new taxon, as the SPR distance between the supertree and other gene trees is unchanged. Once we have constructed an initial SPR supertree (or, alternatively, are supplied an initial tree by the user) we begin the SPR rearrangement phase. For a prespecified number of iterations, we look at the O(n^2) trees that can be obtained from the current supertree of n leaves by one SPR operation and select from these the tree with minimal SPR distance. Many of these SPR rearrangements will be obviously flawed, so we incorporate a bipartition clustering approach to ignore such rearrangements. Any bipartition of the supertree that is supported by at least half of the gene trees containing two or more taxa from each of the two sets induced by the bipartition is considered "fixed", and SPR rearrangements that disrupt it are disallowed. This greatly decreases the number of considered bipartitions with little effect on the accuracy of the tree search.

Our methods were developed for rooted gene trees, but we provide three options to accommodate the unrooted gene trees that are typically produced by maximum-likelihood and Bayesian phylogenetic approaches. Our first method is to compute the minimal SPR distance between the supertree and any rooting of each gene tree using an exhaustive search of all possible rootings. Second, given a rooted (partial) supertree and unrooted gene tree we use each bipartition of the gene tree to try and identify the root bipartition of the supertree. We root the gene tree at the bipartition that best matches the supertree root bipartition according to the balanced accuracy score, an average of the similarities between each matching side of the bipartitions. Suppose that the supertree root bipartition splits the taxa into two groups A and B and a gene tree bipartition splits the taxa into two groups C and D. Then the balanced accuracy of the C|D bipartition as compared to the A|B bipartition is the larger of $((|A| \cap |C|) / 2(|A| + |C|)) + ((|B| \cap |D|) / 2(|B| + |D|)$ or $((|A| \cap |D|) / 2(|A| + |D|)) + ((|B| \cap |C|) / 2(|B| + |C|)$, depending on whether A and C or B and D are more closely matched. Third, we can root the gene trees at a set of outgroup taxa, throwing away trees where this outgroup is not monophyletic. We then build a supertree of this reduced tree set and can then, if desired, root the remainder of the trees using our balanced accuracy approach to build a final supertree.

Comparative Evaluation and Data Sets

We evaluated the performance of our SPR supertree algorithm against two other approaches: the widely used matrix representation with parsimony (MRP) approach of Baum (1992) and Ragan (1992) and the recently published Robinson-Foulds (RF) supertree algorithm (Bansal et al. 2010). Since the RF supertree approach is also based on topological distances between trees, it is an appropriate comparator for our SPR-based method. To construct MRP supertrees we used the Clann 3.2.2 (Creevey and McInerney 2005) software package to generate matrices for a PAUP* version 4.0b10 (Swofford 2003) parsimony search using 25 iterations of SPR rearrangements (to match the SPR and RF approaches). RF supertrees were constructed using version 1.8.4 of the software described by Bansal et al. (2010) which uses 25 iterations of SPR rearrangements interleaved with partial data ratchet iterations. The three methods were compared in terms of their running time on various datasets as well as their accuracy, either against the known phylogeny in the case of simulated data sets or the three supertree criteria when empirical data sets were used.

We built simulated data sets to evaluate the accuracy of SPR, MRP and RF on gene trees generated from a completely known species history. EvolSimulator (Beiko and Charlebois 2007) version 2.2 was used to generate 15 replicated speciation and extinction histories in populations limited to 25 extant genomes. 10,000 simulation iterations were run in all cases. For each of the 15 distinct histories, multiple runs were carried out in which the rate of LGT was varied between 0 (no LGT) and 2.5 events per iteration in increments of 0.1. We also simulated two different LGT regimes: random, in which transfers between any donor/recipient pair were equally probable; and divergence-biased, where donor/recipient exchanges were more likely between closely related genomes (i.e., genomes that share a recent common ancestor), with no LGT at all between genomes that diverged > 5000 generations in the past. The ancestral genome in each simulation (i.e., iteration 1) had 150 genes, and lineages could gain and lose genes to a minimum of 100 and a maximum of 200. A full list of parameter settings can be found in the sample configuration file (see online Supplemental Material). The resulting gene trees were used to infer supertrees under the SPR, MRP and RF criteria: supertree accuracy was evaluated based on dissimilarity with the known species tree, and the total distance between the supertree and all gene trees.

We also compared the three methods using published eukaryotic supertree datasets of marsupials (Cardillo et al. 2006), seabirds (Martyn and Page 2002), placental mammals (Beck et al. 2006) and papilionoid legumes (Wojciechowski et al. 2000) obtained from http://www.cs.utexas.edu/~phylo/datasets/supertrees.html. These datasets cover between 121-558 taxa in 7-726 trees and were used to compare the supertree methods according to their respective supertree optimization criteria, as was done by Bansal et al. (2010).

Finally, we constructed a 244-taxon bacterial SPR supertree from a 40,631-tree subset of the 159,905 unrooted multifurcating prokaryotic phylogenetic trees from Beiko (2011), compared it with an MRP supertree and used the SPR supertree to infer “highways of gene sharing”, that is, frequently implied pathways of LGT among major bacterial lineages. From the 1179 taxa in the original dataset, we randomly selected 15 Alphaproteobacteria, Betaproteobacteria and Deltaproteobacteria, 14 Epsilonproteobacteria, 13 Gammaproteobacteria, 40 Bacilli, 34 Clostridia, 74 Actinobacteria, 2 Deferribacteres, 11 Thermotogae, 7 Aquificae, 2 Nitrospira and 2 Synergistetes for a total of 244 taxa (listed in online Supplemental Table 1) covering a subset of well-sampled and sparsely sampled classes of bacteria and restricted the 159,905 trees to this

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18

subset. We then collapsed all branches with a bootstrap support value of less than 0.8 and discarded all star trees and trees with fewer than 4 taxa. After this procedure, 40,631 trees remained. In total, there were 393,876 leaves in the trees for an average of 9.7 taxa per tree. To construct a supertree from the set of unrooted gene trees, we used our rooting method described above with the Aquificae as outgroup. We first constructed an initial guiding supertree from the 40 largest gene trees with a monophyletic Aquificae group (Griffiths and Gupta 2004). This required 13 global rearrangement iterations and 87 CPU hours to converge on a local minimum. The remaining trees were then rooted using our balanced accuracy approach, and we constructed our SPR supertree from this data set using the guiding supertree as a base, which required 16 iterations to converge and 1198 CPU hours.

Once the final supertree was obtained, LGT events were inferred using MAF comparisons between our SPR supertree and the gene trees. We computed a single MAF for each gene tree and determined the equivalent sequence of implied LGT events in less than one minute. Transfers where both the putative donor and recipient were contained within two distinct genera were counted, and the results visualized as a heatmap and LGT affinity graph constructed using Cytoscape 2.8.3 (Smoot et al. 2012). We ignored directionality as it is often possible to identify partners but not the direction of transfer (Beiko and Ragan 2008). Heatmap values were scaled such that each row had a mean of 0 and standard deviation of 1 and relationships with fewer than 5% of the maximum transfer events for a row or only a single transfer event were filtered out. Two genera were connected by an edge if the number of inferred LGT events between them exceeded 5% of the total number of homologous genes common to at least one member of both genera.

All supertrees constructed from empirical data, as well as the input bacterial trees we used, are available online as Supplemental Material.

RESULTS

Bacterial SPR Supertree and Large-Scale Analysis of LGT

We first present our supertree of 244 bacterial taxa that was constructed from 40,631 unrooted input gene trees using our two-stage outgroup procedure. The taxa selected for our bacterial supertree analysis were chosen to examine several interesting phylogenetic questions in the Bacteria. For example, there are two competing hypotheses for the placement of the Aquificae. Informational genes such as 16S small subunit ribosomal RNA suggest that the Aquificae are deep-branching and either external to or sister with the Thermotogae but the majority of other proteins suggest that the Aquificae are sister to the Epsilonproteobacteria (or other groups such as the Deltaproteobacteria) and not the Thermotogae (Boussau et al. 2008). It has been suggested that the Aquificae may be closely related to the Epsilonproteobacteria with either LGT or a thermophilic G+C bias and long-branch attraction responsible for the observed affinity for Thermotogae (Griffiths and Gupta 2004). Informational proteins are thought to be transferred infrequently, so it has been more recently suggested that there have been large amounts of lateral gene transfer between the Aquificae and Epsilonproteobacteria (Boussau et al. 2008). Our dataset also includes members of many other groups implicated in LGT, including the Deltaproteobacteria and Clostridia: both of these groups show evidence of frequent LGT with other lineages (Dagan et al. 2010; He et al. 2010; Beiko 2011). Other genera frequently associated with high LGT rates including *Pseudomonas* and *Burkholderia* are also included. Finally, several lineages such as Deferribacteres and Synergistetes with relatively few sequenced representatives and an uncertain phylogenetic position (Jumas-Bilak et al. 2009) were included to assess their placements in the SPR supertree.

Figure 2 shows our SPR supertree of the 244-taxon bacterial dataset. The SPR supertree largely recovered the major bacterial classes as monophyletic groups with several notable exceptions. The Deltaproteobacteria are separated from the other Proteobacteria by the Actinobacteria. The Deltaproteobacteria are also split into a group containing the Myxobacteria and *Candidatus* “Nitrospira defluvii”, and a group containing all other orders of the class. Although assigned to phylum Nitrospirae, *Ca. N. defluvii* has strong affinities to other phylogenetic groups, with deltaproteobacterial genomes constituting seven of the 15 most frequently observed phylogenetic partners. This is an interesting link as *Sorangium cellulosum* has the largest known bacterial genome (Schneiker et al. 2007) and both *Candidatus Nitrospira defluvii* and *Anaeromyxobacter dehalogenans* are gram-negative nitrite reducers. Further, it has been suggested that *Ca. N. defluvii* evolved from microaerophilic or even anaerobic ancestors (Lucker et al. 2010) and *Anaeromyxobacter dehalogenans* exhibits aerobic and anaerobic growth (Sanford et al. 2002). Two other proteobacteria are separated from their classes: *Bdellovibrio bacteriovorus*, a Deltaproteobacterium that parasitizes other gram-negative bacteria (Stolp and Starr) and appears to have acquired genes from the protebacterial cells it parasitises (Gophna et al. 2006), and *Candidatus Hodgkinia cicadicola*, an alphaproteobacterial cicada symbiont with the smallest known genome (McCutcheon et al. 2009), form a pairing that is sister to the Epsilonproteobacteria.

Among other phylogenetic groups, *Veillonella parvula* and *Acidaminococcus fermentans*, initially assigned to class Clostridia, are sister to the Bacilli. *Veillonellaceae* and *Acidaminococcaceae* have a peculiar cell wall composition which stains Gram-negative, unlike most Firmicutes, and have been suggested to belong to a class Negativicutes, separate from the Bacilli and Clostridia, by Marchandin et al. (2010). *Coprothermobacter proteolyticus* groups with the Thermotagae rather than the Clostridia. *C. proteolyticus* was assigned to class Clostridia using small subunit ribosomal RNA (Rainey and Stackebrandt 1992) but phylogenomic analysis (Beiko 2011; Yutin et al. 2012) and newer phylogenetic trees built from many more samples of small subunit ribosomal RNA agree with a closer relationship between *C. proteolyticus* and Thermotogae (Munoz et al. 2011). With Aquificae as the outgroup, the next-deepest branches in the bacterial tree are *Thermodesulfovibrio yellowstonii*, the other member of phylum Nitrospirae, and the Deferribacteres, followed by Thermotogae. The Synergistetes are sister to the Firmicutes in this tree.

We then inferred LGT events between these bacteria by computing a single MAF for each gene tree and determining the equivalent sequence of implied LGT events. This entire analysis of the 40,631 gene trees required less than one minute using our refined MAF algorithms. Transfer events with source and endpoints both in a monophyletic subtree of the same genus or different genera were identified to focus on relatively recent transfers. Directionality was ignored as it is often possible to identify partners but not the direction of transfer (Beiko and Ragan 2008). Figure 3a shows the results of this analysis. Clustering based on the strength of their LGT affinities still groups most genera by class and phylum, and the majority of inferred LGT events occur within clusters of taxonomically related genera. However, there are also many linkages between genera of distinct phyla and clusters of genera with distinct classes and phyla. Online Supplemental Figure 1 shows a heatmap of the relative LGT trends between classes.

A genus-level LGT affinity graph (Fig. 3b) between genera was used to further explore these relationships and identify paths of gene sharing between distinct lineages. Genera were connected by edges representing transfer events exceeding 5% of their total number of shared homologous genes. As in Figure 3a, the majority of inferred LGT events connect members of the same class or phylum. Yet many linkages connect different classes and phyla such that all of the genera but two, *Ehrlichia* and *Wolbachia*, are connected. The large and diverse genus *Clostridium*, in particular, connects Actinobacteria, Thermotogae, four of the five classes of Proteobacteria, *Thermoanaerovibrio* (phylum Synergistetes), and has many strong connections with Bacilli and other Clostridia (online Supp. Fig. 2). Family Coriobacteriaceae, comprising *Slackia*, *Eggerthella*, and *Cryptobacterium*, had linkages with the other Actinobacterial genera *Corynebacterium* and *Bifidobacterium* but was also connected to the Firmicute genera *Clostridium, Eubacterium, and Streptococcus.* There are numerous pathways of gene sharing between actinobacterial genera such as *Acidimicrobium*, *Corynebacterium* and *Mycobacterium* on the one hand, and proteobacterial genera such as *Helicobacter*, *Sorangium*, *Xanthomonas* and *Mesorhizobium* on the other. A single path between *Nitratiruptor* and *Persephonella* connects the Epsilonproteobacteria with the Aquificae. Many connections are observed between the different classes of Proteobacteria, highlighting the numerous LGT events that occur between distinct lineages of phylum Proteobacteria. The connectedness of higher taxonomic groups is supported by the class-level affinity graph (online Supp. Fig. 3), in which each class is connected to 3.92 other classes on average, with the Actinobacteria connected to a total of ten.

Validation of Efficiency and Accuracy

We next demonstrate the improved performance of our MAF algorithms with a single SPR distance analysis of our 244-taxon bacterial supertree as compared to each of the 40,631 gene trees. Figure 4 shows the mean running time for tree comparisons with a given SPR distance on a

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18

log scale. Our improved algorithms reduced the time required for individual calculations from 5 hours to a maximum of 0.8 seconds on the initial set of binary gene trees. Both the cluster reduction and our improved algorithms are necessary to achieve these running times. Our algorithm requires slightly more time to compare the supertree with multifurcating trees for a given SPR distance but this is balanced by the reduction in SPR distance caused by collapsing unsupported bipartitions; clustered comparisons required at most 0.76 seconds. As mentioned previously, a full LGT analysis now requires just 34 seconds on a single CPU. Without our new algorithms, such an analysis would be limited to binary trees and require more than 65 hours.

Validation with Simulated Datasets

We next compared the ability of SPR, RF, and MRP based supertrees to recover the species tree in a series of simulated datasets. EvolSimulator (Beiko and Charlebois 2007) was used to evolve sets of genomes under a model of lineage duplication and extinction, with each lineage capable of gene duplication, gene loss, and LGT. Varying the rate of LGT in different sets of replicated simulations allowed us to explore the effectiveness of SPR, RF and MRP at relatively low or high levels of LGT. We also simulated two regimes of LGT: random LGT, which can interfere with the recovery of correct branching patterns, and divergence-biased LGT, which can actually reinforce the true tree due to preferential sharing between close relatives (Beiko et al. 2008).

Simulated LGT rates varied between 0 (no LGT) and 2.5 events per iteration (see Methods for details). To give context to our LGT rate simulation parameter, we computed the mean ratio of SPR distance to number of leaves in the simulated trees, to similar values inferred for the 244-taxon SPR supertree (Fig. 5). The inferred frequency of LGT in our empirical data equated to a simulated random LGT rate between 0.1 and 0.2 and a simulated divergence-biased LGT rate between 0.3 and 0.4. Since the bacterial supertree has 244 leaves rather than 25, we also restricted our bacterial supertree and gene trees to 25 randomly sampled subsets of 25 leaves and computed this ratio. We found these subsampled supertrees corresponded to lower simulated rates of LGT. This suggests that our simulations with lower rates of LGT are biologically plausible; also, since the distribution of LGT events is non-uniform across bacterial lineages (Kunin et al. 2005; Beiko et al. 2005; Thiergart et al. 2012) the higher rates are likely to be relevant to the inference of some relationships in the supertree.

Having established the relevance of our simulated rates of LGT, we then assessed the ability of different supertree algorithms to recover the correct organismal history based on analysis of the gene trees. Figure 6 shows the mean SPR difference between the simulated species histories and the RF supertree, SPR supertree, SPR supertree seeded with an MRP starting tree, and SPR supertree seeded with the correct species tree. SPR supertrees were significantly more similar to the simulated species tree than RF supertrees for the LGT rates seen in our bacterial dataset and higher ($p < 0.05$ for random LGT rates of 0.2-1.4 and divergence-biased LGT rates of 0.7,0.8 and 1.0 with a 2-tailed paired student's t-test; $p < 0.01$ for random LGT rates of 0.2-0.7, 0.9, 1.3, 1.4; the overall results were significant with $p < 10^{-5}$ for both types of LGT). Seeding the SPR supertree search with an MRP tree did not substantially change these results. Seeding the SPR supertree search with the correct tree does not substantially change the results for divergence-biased LGT or plausible rates of random LGT. We see that the SPR supertree and the simulated species tree diverge as the random LGT rate increases, even when seeded with the species tree. These results suggest that datasets with substantially higher rates of LGT than our bacterial data would require a better search strategy or a network-based analysis rather than a supertree.

Figure 7 compares the accuracy of SPR and MRP supertrees. As MRP constructs unrooted supertrees, the error is measured here as the minimum SPR distance between the simulated species history and any rooting of the inferred supertrees. The upper panels of Figure 7 show the mean supertree error between the simulated species histories and the MRP supertree, SPR supertree, SPR supertree seeded with an MRP starting tree, and SPR supertree seeded with the correct species tree. The SPR supertrees were significantly more similar to the simulated species history than the MRP trees under biologically plausible rates of LGT ($p < 0.01$ for random LGT rates of 0.3-0.5 with a two-tailed paired student's t-test; the divergence-biased results were not significantly different for individual rates other than 0.6 and 1.0 due to the small supertree error but were significantly better overall with $p < 0.001$). At higher simulated rates of LGT the accuracy of SPR supertrees matches that of the MRP trees. We observed that this occurs when the accuracy of the SPR supertree and the SPR supertree seeded with the correct tree diverge, suggesting that a better search strategy may improve these results. We also examined the accuracy of RF supertrees with this unrooted measure and found similar results to the unrooted comparison, that is, SPR supertrees and MRP supertrees were both significantly more similar to the simulated species tree than the RF supertrees (online Supp. Fig. 4). The lower panels of Figure 7 show the mean supertree error between the simulated species histories and the MRP supertree and SPR supertrees using our balanced accuracy based simple unrooted comparison without and with an MRP seed tree. The accuracy of our SPR supertrees when the gene tree roots are unknown matches that of the MRP trees for plausible rates of LGT but the performance of our SPR supertrees declines with increasing rates. Using an MRP seed tree prevented this decline which suggests that our initial tree construction step is not well suited to gene trees with unknown roots. Developing an improved method for building starting trees from unrooted gene trees could improve these results.

Comparison with MRP and RF Supertrees on Eukaryotic Datasets

Bansal et al. (2010) validated their RF supertree approach on a series of eukaryotic datasets that varied substantially in the number of input trees and total number of taxa. We compared the accuracy of each supertree method on these datasets as measured by their ability to minimize the three supertree criteria of SPR distance, RF distance, and parsimony score to the gene trees. In addition to the three basic methods, we tested a variant of SPR supertrees that uses the RF distance as a secondary optimization criterion to break ties when multiple supertrees have the same SPR distance, and tested the SPR and RF supertree methods when the MRP supertree was used as the initial tree. As MRP supertrees are unrooted, we computed the RF and SPR distances for each rooting of the MRP supertree and show the minimum value. For these tests each supertree method was run with its default parameters to match the comparisons of Bansal et al (2010) so we used the SPR and RF methods with 25 iterations of SPR rearrangements and the MRP method with 10 iterations of TBR rearrangements. Due to excessive running times (> 3 days) for the MRP method on the marsupial and legume datasets we disabled the 'multrees' option on these runs which would otherwise retain multiple trees per iteration.

The performance of each approach according to all optimality criteria is shown in Table 1. Each supertree method was best at minimizing its respective optimization measure, suggesting that each method has merit and a well-balanced analysis should either include a justification for the choice of method (e.g. the presence of LGT for the SPR distance) or consider multiple optimization criteria. The MRP method required the least amount of time and the SPR method the most. However, the SPR method converged rapidly in 3, 1, 5 and 3 iterations on the marsupial, seabird, placental mammal, and legume datasets respectively and thus produced an optimal result in only a fraction of the reported time. Seeding the search with the MRP tree greatly reduced the time required by the SPR method and reduced the resulting parsimony scores at the expense of increasing the SPR distance. Starting with the MRP tree reduced the time required by the RF method and found supertrees with better RF and MRP scores on the marsupial and placental mammal datasets but increased RF and MRP scores on the legume dataset. Using the RF distance as a tie-breaker with the SPR method found lower SPR distances, RF distances and parsimony scores in a shorter period of time over the basic method and avoided an issue with the seabird dataset where many supertrees have the same SPR distance but poor RF distances and parsimony scores. These results suggest that blended methods have merit even when only considering a single optimization criterion. In particular, the SPR distance with RF distance as a tie-breaker should be used when nontrivial amounts of lateral gene transfer are expected.

Comparison of SPR and MRP Supertrees of 244 Bacterial Genomes

To contrast with the SPR supertree described above and examine the influence of tree rootings, we constructed an MRP supertree from the 244-taxon bacterial dataset using 25 iterations of an SPR rearrangement search and compared it to our SPR supertree (Fig. 8). The MRP supertree does not recover the same arrangement of hyperthermophiles as the SPR supertree; notably, it places the Epsilonproteobacteria in close proximity to the Aquificae. If we place the root somewhat arbitrarily between the Firmicutes and all other Bacteria, the MRP supertree like the SPR supertree places the Thermotogae and *C. proteolyticus* as sisters, although this pairing is sister to the Synergistetes and not the Deferribacteres in the MRP supertree. The two Nitrospirae are again split, with *Nitrospira* sister to the Deltaproteobacteria and *Thermodesulfovibrio* with the Aquficae and Deferribacteres. As with the SPR supertree, the Deltaproteobacteria are separated from the other Proteobacteria.

The rooted nature of MAFs allowed the evaluation of our chosen rooting and alternative rootings on inferring phylogenetic relationships from this dataset. We have already described the MRP supertree rooted to separate the Firmicutes from the other taxa (MRP), the SPR supertree constructed from the 40 largest trees with a monophyletic Aquificae group (40-Aquificae) and the SPR supertree constructed using the SPR-Aquificae supertree (SPR-Aquificae). Three more supertrees were constructed to test the influence of starting topology and rooting. The first was an SPR supertree seeded with the MRP supertree (SPR-MRP). We then rooted the gene trees with both the MRP supertree and SPR-Aquificae tree using our balanced accuracy measure and constructed an SPR supertree from these two sets of rooted gene trees (SPR-MRP-Rooting and SPR-Aquificae-Rooting, respectively).

These six supertrees were compared to the two sets of rooted gene trees (see Table 2). The three MRP-rooted supertrees had a much smaller aggregate SPR distance (nearly 11% smaller) to the MRP-rooted gene trees than the Aquificae-rooted supertrees but the three Aquificae-rooted supertrees had a much smaller SPR distance (more than 8% smaller) to the Aquificae-rooted gene trees than the three MRP-rooted supertrees. Thus, it is impossible to determine which supertree is more similar to the gene trees without choosing a specific rooting of the gene trees.

The four SPR supertrees constructed from the full bacterial dataset were compared by measuring their pairwise SPR distances (see Table 3). The two Aquificae-rooted supertrees differed by only 10 SPRs, despite the fact that one was constructed from the 40-Aquificae tree and the other was constructed with our usual greedy addition procedure and no *a priori* information other than the gene tree roots. Even more telling, the two MRP-rooted supertrees were essentially identical, differing by only 2 SPRs. The SPR-MRP-Rooting supertree also differed from the MRP supertree by only 2 SPRs, so we were able to essentially recover the MRP supertree just by biasing the gene tree roots. This suggests that MRP infers relationships that are consistent with certain gene tree roots despite not implicitly assuming any rooting. As these relationships are also inconsistent with plausible alternative roots, it may be that unrooted supertree methods such as MRP are insufficient to distinguish between controversial evolutionary hypotheses such as the placement of the Aquificae.

Discussion

Large phylogenies are being built from multiple sequence datasets to reconstruct the histories of many groups of living organisms, and supertrees offer the means to carry this out in a rigorous fashion. The known limitations of widely used approaches such as MRP have motivated the development of new strategies, such as the use of Robinson-Foulds distance as an alternative optimality criterion. Although RF is frequently used to assess the dissimilarity of phylogenetic trees, it is not based on a specific phylogenetic process and can be heavily influenced by shifts in the position of single taxa. A single LGT event will influence the RF distance (and parsimony score) in proportion to the number of branches in the path between the donor and recipient lineages, and many LGT events are likely to confound RF-based supertree inference. The SPR distance is an alternative optimality criterion that is particularly well-suited to analyzing phylogenomic data where LGT or other reticulate evolutionary processes are expected to play an important role in generating phylogenetic discordance. Each SPR operation is equivalent to an LGT event, and the degree of separation between donor and recipient in the tree does not influence the SPR score. The SPR distance may thus avoid some of the “phylogenetic compromises” of other supertree methods.

Using simulations, we verified that SPR supertrees were significantly more similar to the known species history than RF supertrees given biologically plausible rates of simulated LGT. The effect was more pronounced for random LGT, which produces more "long-distance" transfers, than for divergence-biased LGT. The improved performance of SPR with random LGT events suggests that penalizing phylogenetic discordance in a manner that is insensitive to the number of impacted bipartitions may be preferable to the alternative RF criterion. However, in the future this assertion should be tested under a wider range of scenarios, with larger trees and different types of phylogenetic discordance modelled. SPR also outperformed MRP in a narrower, but still biologically relevant, range of LGT rates. However, the advantage of SPR disappeared when the gene tree roots were unknown, demonstrating that the obligately rooted SPR approach is influenced by alternative rootings of the reference and gene trees. We also verified that each of the three supertree methods excel at minimizing their respective supertree criteria on a eukaryotic dataset. Combining multiple supertree criteria, such as using the RF distance to break ties in an SPR supertree approach, yielded better results than any method did alone. This finding suggests that combinations of criteria that consider different types of phylogenetic discordance may provide even greater accuracy.

Although the history of bacteria may be better represented with a phylogenetic network than a single tree, the supertree we inferred offers a useful backdrop for the inference of highways of gene sharing. As shown in Figures 2 and 8, both SPR and MRP recovered a majority of bacterial classes as monophyletic groups, regardless of the choice of rooting. Many of the topological differences between the SPR and MRP supertrees are minor, including subtle shifts in the position of taxa such as *Nitrospira defluvii* and the Negativicutes. One point of substantial difference between the two trees related to the controversial placement of Aquificae and the Epsilonproteobacteria: MRP, being unrooted, placed these two groups adjacent to one another, corresponding to a sister relationship under the reasonable assumption that the root of the supertree is placed somewhere outside of this pairing. When the SPR supertree was rooted to reflect the MRP tree topology in the manner described above, the two supertrees were nearly identical; however, if Aquificae were treated as the outgroup then the SPR supertree produced a topology that placed other groups with many thermophiles, such as Thermotogae, as early branches. These results suggest that unrooted supertree criteria such as MRP provide hypotheses that are consistent with certain rootings despite not implicitly assuming any rooting. Furthermore, the Aquificae SPR supertree was much more similar to the Aquificae rooted gene trees than the MRP supertree, but the MRP supertree was much more similar to the MRP-rooted trees. It was thus impossible to distinguish between these two hypotheses of Aquificae placement; either could be plausible given knowledge of the correct gene tree roots. This is a practical example of the fundamental limits of unrooted supertree methods identified by Steel et al. (2000).

Using the tree in Figure 2 as a basis for LGT inference, we searched for highways of LGT between classes and genera. Not surprisingly, connections were more frequently associated with specific lineages such as *Clostridium* and interactions between the Proteobacteria and other phyla varied considerably. In addition, larger gene trees (those shared by many taxa) required proportionately more transfers to explain, including ribosomal proteins. Such biased LGT could muddy or completely obscure the vertical evolutionary signal. Our improved SPR algorithm allowed the entire set of >40,000 trees to be reconciled with the supertree in less than one minute: a similar analysis could have been carried out using any rooted reference tree, regardless of what method was used to construct this tree. The rapid inference of LGT highways raises the possibility of using information about lateral connections to construct phylogenetic networks with reticulations explicitly based on major directions of LGT (MacLeod et al. 2005; Nakhleh et al. 2005; Beiko and Hamilton 2006).

The scaling of runtimes with the number and size of trees is a central concern in phylogenomics. The analysis of Beiko et al. (2005) required over 20,000 CPU hours to reconcile 22,432 gene trees with a 144-taxon supertree, and the largest trees could not be reconciled at all due to limitations of the breadth-first search of EEEP (Beiko and Hamilton, 2006). Alternative methods of inferring highways of LGT have been proposed based on quartets (Bansal et al. 2013), but such methods are limited to finding the most obvious highways and required on the order of two days to analyze the same dataset of 22,432 gene trees. Repeated applications of SPR distances in large phylogenomic data sets were heretofore not feasible due to the complexity of the algorithm, but our efficient new methods for computing the SPR distance made the computation of these supertrees feasible even for hundreds of taxa and tens of thousands of gene trees. Of particular importance is the adaptation of the clustering strategy of Linz and Semple (2011) to subdivide the construction of a MAF for a given pair of trees. Clustering yields no improvement in theoretical runtime, because there is no guarantee that >1 cluster will be identified between a pair of trees. However, our results clearly demonstrate that clustering is effective in practice, because LGT connections are not random and consistent partitioning can usually be identified and used as the basis for subdivision. We are optimistic that our approach will be applicable to much larger phylogenomic data sets with thousands of taxa, for two reasons: first, our fixed-parameter algorithm scales exponentially with the *distance* between a pair of trees and not their *size*; and second, as the timing results of Figure 5 suggest, clustering increases the speed of the algorithm and reduces the rate of increase of running times with increasing SPR distance. With only a small number of exceptions, all trees with SPR distance < 60 were resolved in less than one second, with the time of MAF construction dominated by the single cluster with the largest distance. We expect that most large trees will have a cluster size distribution similar to that of the trees we tested here; consequently the size of the largest cluster and the corresponding computational burden may increase only slightly. This hypothesis remains to be tested on larger phylogenomic data sets.

Our methods could be expanded and refined in several ways. As we identified in our results, our current supertree search method could potentially be improved with a better strategy for constructing the initial guide tree such as SuperFine (Swenson et al. 2012), methods for avoiding local optima such as ratchet searches, or using prior knowledge to constrain the supertree search (Wehe et al. 2012). An RF supertree method has been recently proposed for multi-labelled gene trees (Chaudhary et al. 2013); extending our SPR distance algorithms to accept such trees would enable their inclusion in SPR supertrees. The rooting problem remains to be resolved. While in many cases rooting can be performed using an appropriate outgroup taxon, the bacterial case considered here lacks an obvious outgroup: the Archaea could be used to root the Bacteria and vice versa, but many gene trees have shown evidence of interdomain LGT and rooting between domains may be invalid or even impossible. Finally, our approach considers only the history of observed genes, and does not attempt to account for processes such as gene duplication and loss. Methods of reconciling multiple evolutionary processes such as duplications, losses, transfers and incompatible lineage sorting (ILS) show a great deal of promise (Bansal et al. 2012; Szöllosi et al. 2012), but are currently limited to smaller datasets (Stolzer et al. 2012).

FUNDING

This work was funded by the Natural Sciences and Engineering Research Council of Canada via grants to N.Z. and R.G.B., and a graduate fellowship to C.W. C.W. was also supported by the Killam Trusts. N.Z. and R.G.B. acknowledge the support of the Canada Foundation for Innovation and the Canada Research Chairs Program.

ACKNOWLEDGMENTS

We thank Joel Navarrete for his help in evaluating variants of our supertree approach.

REFERENCES

Adams EN. 1972. Consensus techniques and the comparison of taxonomic trees. Syst Biol 21(4):390-7.

Bansal MS, Alm EJ, Kellis M. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics 28(12):i283-91.

Bansal MS, Burleigh JG, Eulenstein O, Fernández-Baca D. 2010. Robinson-Foulds supertrees. Algorithm Mol Biol 5(1):18.

Bansal MS, Banay G, Harlow TJ, Gogarten JP, Shamir R. 2013. Systematic inference of highways of horizontal gene transfer in prokaryotes. Bioinformatics 29(5):571-9.

Baroni M, Grünewald S, Moulton V, Semple C. 2005. Bounding the number of hybridisation events for a consistent evolutionary history. J Math Biol 51(2):171-82.

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41(1):3-10.

Beck RM, Bininda-Emonds OR, Cardillo M, Liu FR, Purvis A. 2006. A higher-level MRP supertree of placental mammals. BMC Evolutionary Biology 6(1):93.

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

Beiko RG. 2011. Telling the whole story in a 10,000-genome world. Biol Direct 6:34.

Beiko RG and Charlebois RL. 2007. A simulation test bed for hypotheses of genome evolution. Bioinformatics 23(7):825-31.

Beiko RG, Doolittle WF, Charlebois RL. 2008. The impact of reticulate evolution on genome phylogeny. Syst Biol 57(6):844-56.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A 102(40):14332-7.

Bender MA and Farach-Colton M. 2000. The LCA problem revisited. In: LATIN 2000: Theoretical informatics. Springer. 88-94.

Bininda-Emonds OR and Sanderson MJ. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst Biol 50(4):565-79.

Bininda-Emonds OR, Gittleman JL, Steel MA. 2002. The (super) tree of life: Procedures, problems, and prospects. Annu Rev Ecol Syst 33:265-89.

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. Nature 446:507-12.

Bordewich M and Semple C. 2005. On the computational complexity of the rooted subtree prune and regraft distance. Ann Comb 8(4):409-23.

Boussau B, Guéguen L, 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.

Cardillo M, Bininda-Emonds R, Boakes E, Purvis A. 2004. A species-level phylogenetic supertree of marsupials. J Zool 264(1):11-31.

Chaudhary R, Burleigh J, Fernández-Baca D. 2012. Fast local search for unrooted robinson-foulds supertrees. IEEE ACM T Comput Bi 9(4):1004-13.

Chaudhary R, Burleigh J, Fernández-Baca D. 2013. Inferring Species Trees from Incongruent Multi-Copy Gene Trees Using the Robinson-Foulds Distance. arXiv preprint http://arxiv.org/abs/1210.2665.

Chen D, Eulenstein O, Fernández-Baca D, Burleigh JG. 2006. Improved heuristics for minimum-flip supertree construction. Evolutionary Bioinformatics Online 2:347.

Chen Z and Wang L. 2013. An ultrafast tool for minimum reticulate networks. J Comput Biol 20(1):38-41.

Cotton JA and Wilkinson M. 2007. Majority-rule supertrees. Syst Biol 56(3):445-52.

Creevey CJ and McInerney JO. 2005. Clann: Investigating phylogenetic information through supertree analyses. Bioinformatics 21(3):390-2.

Dagan T, Roettger M, Bryant D, Martin W. 2010. Genome networks root the tree of life between prokaryotic domains. Genome Biol Evol 2:379-392.

Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. Proc Natl Acad Sci U S A 101(7):1904-9.

Eulenstein O, Chen D, Burleigh JG, Fernández-Baca D, Sanderson MJ. 2004. Performance of flip supertree construction with a heuristic algorithm. Syst Biol 53(2):299-308.

Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Phil Trans R Soc B 363(1512): 4023-4029.

Goloboff PA. 2005. Minority rule supertrees? MRP, compatibility, and minimum flip may display the least frequent groups. Cladistics 21(3):282-94.

Goloboff PA. 1999. Analyzing large data sets in reasonable times: Solutions for composite optima. Cladistics 15(4):415-28.

Gophna U, Charlebois RL, Doolittle WF. 2006. Ancient lateral gene transfer in the evolution of bdellovibrio bacteriovorus. Trends Microbiol 14(2):64-9.

Griffiths E and Gupta RS. 2004. Signature sequences in diverse proteins provide evidence for the late divergence of the order aquificales. Int Microbiol 7(1):41-52.

He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HM, Quail MA, Rance R, Brooks K, Churcher C, Harris D, Bentley SD, Burrows C, Clark L, Corton C, Murray V, Rose G, Thurston S, van Tonder A, Walker D, Wren BW, Dougan G, Parkhill J. 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci USA 107: 7527-32.

Hein J, Jiang T, Wang L, Zhang K. 1996. On the complexity of comparing evolutionary trees. Discrete Appl Math 71(1):153-69.

Jumas-Bilak E, Roudière L, Marchandin H. 2009. Description of ‘Synergistetes’ phyl. nov. and emended description of the phylum ‘Deferribacteres’ and of the family syntrophomonadaceae, phylum ‘Firmicutes’. Int J Syst Evol Micr 59(5):1028-35.

Kennedy M, Page RD, Prum R. 2002. Seabird supertrees: Combining partial estimates of procellariiform phylogeny. Auk 119(1):88-108.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res 15: 954-9.

Lapierre P, Lasek-Nesselquist E, Gogarten JP. 2012. The impact of HGT on phylogenomic reconstruction methods. Brief Bioinform. [epub ahead of print – PubMed ID 22908214]

Lin HT, Burleigh JG, Eulenstein O. 2009. Triplet supertree heuristics for the tree of life. BMC Bioinformatics 10(Suppl 1):S8.

Linz S and Semple C. 2011. A cluster reduction for computing the subtree distance between phylogenies. Ann Comb 15(3):465-84.

Lloyd GT, Davis KE, Pisani D, Tarver JE, Ruta M, Sakamoto M, Hone DW, Jennings R, Benton MJ. 2008. Dinosaurs and the cretaceous terrestrial revolution. P Roy Soc B-Biol Sci 275(1650):2483-90.

Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B, Rattei T, Damsté JSS, Spieck E, Le Paslier D. 2010. A nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. P Natl Acad Sci USA 107(30):13479-84.

MacLeod D, Charlebois RL, Doolittle WF, 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.

Marchandin H, Teyssier C, Campos J, Jean-Pierre H, Roger F, Gay B, Carlier J, Jumas-Bilak E. 2010. *Negativicoccus succinicivorans* gen. nov., sp. nov., isolated from human clinical samples, emended description of the family Veillonellaceae and description of Negativicutes classis nov., Selenomonadales ord. nov. and Acidaminococcaceae fam. nov. in the bacterial phylum Firmicutes. Int J Syst Evol Microbiol 60(6):1271-9.

Maddison WP, Knowles LL. 2006. Inferring Phylogeny Despite Incomplete Lineage Sorting. Syst Biol 55(1): 21-30.

McCutcheon JP, McDonald BR, Moran NA. 2009. Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont. PLoS Genetics 5(7):e1000565.

Munoz R, Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer K, Glöckner FO, Rosselló-Móra R. 2011. Release LTPs104 of the all-species living tree. Syst Appl Microbiol 34(3):169.

Nakhleh L, Ruths D, Wang LS. 2005. RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer. Lect Notes Comput Sci 3595:84–93.

Piaggio-Talice R, Burleigh JG, Eulenstein O. 2004. Quartet supertrees. Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life 3:173-91.

Pisani D and Wilkinson M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. Syst Biol 51(1):151-5.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol 24(8):1752-60.

Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol 1(1):53-8.

Rainery FA and Stackebrandt E. 1993. Transfer of the type species of the genus *Thermobacteroides* to the genus *Thermoanaerobacter* as *Thermoanaerobacter acetoethylicus* (Ben-bassat and Zeikus 1981) comb. nov., description of *Coprothermobacter* gen. nov., and reclassification of *Thermobacteroides proteolyticus* as *Coprothermobacter proteolyticus* (Ollivier et al. 1985) comb. nov. Int J Syst Bacteriol 43(4):857-9.

Robinson D and Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci 53(1):131-47.

Roshan Usman W., Warnow Tandy, Moret Bernard ME and Williams Tiffani L. 2004. Rec-I-DCM3: A fast algorithmic technique for reconstructing phylogenetic trees. In Proc IEEE Comput Syst Bioinform Conf. 98 p.

Sanford RA, Cole JR, Tiedje JM. 2002. Characterization and description of *Anaeromyxobacter dehalogenans* gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium. Appl Environ Microbiol 68(2):893-900.

Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, Bartels D, Bekel T, Beyer S, Bode E. 2007. Complete genome sequence of the myxobacterium *Sorangium cellulosum*. Nat Biotechnol 25(11):1281-9.

Smoot ME, Ono K, Ruscheinski J, Wang P, Ideker T. 2011. Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics 27(3):431-2.

Steel M and Böcker S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. Syst Biol 49(2):363-8.

Stolp H and Starr M. 1963. *Bdellovibrio bacteriovorus* gen. et sp. n., a predatory, ectoparasitic, and bacteriolytic microorganism. Antonie Van Leeuwenhoek 29(1):217-48.

Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28(18):i409-15.

Swenson MS, Suri R, Linder CR, Warnow T. 2012. SuperFine: Fast and accurate supertree estimation. Syst Biol 61(2):214-27.

Swofford DL. 2003. PAUP*: Phylogenetic analysis using parsimony, version 4.0 b10.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol 4(4):466-85

Szöllosi GJ, Boussau B, Abby SS, Tannier E, Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc Natl Acad Sci U S A. 109(43): 17513-8.

Wehe A, Burleigh JG, Eulenstein O. 2012. Algorithms for knowledge-enhanced supertrees. LNCS 7292:263-274.

Whidden C and Zeh N. 2009. A unifying view on approximation and FPT of agreement forests. LNCS 5724:390-402.

Whidden C, Beiko RG, Zeh N. 2010. Fast FPT algorithms for computing rooted agreement forests: Theory and experiments. In: Experimental algorithms. Springer. p. 141-153.

Whidden C, Beiko RG, Zeh N. 2013. Fixed-Paramater Algorithms for Maximum Agreement Forests. SICOMP.

Whidden C, Beiko RG, Zeh N. 2013. Fixed-Parameter and Approximation Algorithms for Maximum Agreement Forests of Multifurcating Trees. Submitted.

Whidden C and Zeh N. 2013. Computing the SPR Distance of Binary Rooted Trees in O(2^kn) Time. In preparation.

Wilkinson M, Cotton JA, Creevey C, Eulenstein O, Harris SR, Lapointe F, Levasseur C, Mcinerney JO, Pisani D, Thorley JL. 2005. The shape of supertrees to come: Tree shape related properties of fourteen supertree methods. Syst Biol 54(3):419-31.

Wojciechowski MF, Sanderson MJ, Steele KP, Liston A. 2000. Molecular phylogeny of the “temperate herbaceous tribes” of papilionoid legumes: A supertree approach. Adv Legum Syst 9:277-98.

Yutin N, Puigbò P, Koonin EV, Wolf YI. 2012. Phylogenomics of prokaryotic ribosomal proteins. PLoS One 7(5):e36972.

Table 1: Experimental results comparing the performance of the SPR supertree method to RF and MRP supertree methods. Six analyses are shown: The SPR supertree method starting from an SPR greedy addition tree (SPR) or MRP supertree (SPR-MRP), the SPR supertree method breaking ties with the RF distance using a greedy addition tree (SPR-RF-TIES), the RF supertree method starting from random addition sequence trees (RF-Ratchet) or MRP supertree (RF-MRP), and MRP with TBR global rearrangements (MRP-TBR). The best optimization criteria or running times for a dataset are shown in bold.

Data Set	Supertree Method	SPR Distance	RF-Distance	Parsimony Score	Time (s)
Marsupial (267 taxa; 158 trees)	SPR	382	1604	2203	1097.79
	SPR-RF-TIES	**373**	1536	2149	767.01
	SPR-MRP	380	1534	2126	219.64
	RF-Ratchet	394	1520	2145	2150.30
	RF-MRP	379	**1502**	2116	2044.07
	MRP-TBR	379	1514	**2112**	**20.52**
Sea Birds (121 taxa; 7 trees)	SPR	**17**	109	235	31.15
	SPR-RF-TIES	**17**	63	208	29.44
	SPR-MRP	**17**	**61**	**208**	2.04
	RF-Ratchet	**17**	**61**	**208**	10.43
	RF-MRP	**17**	**61**	**208**	9.16
	MRP-TBR	**17**	**61**	**208**	**1.03**
Placental Mammals (116 taxa; 726 trees)	SPR	1715	5908	8946	5561.84
	SPR-RF-TIES	**1713**	5902	8934	5040.03
	SPR-MRP	**1713**	5876	8921	1819.08
	RF-Ratchet	1790	5738	8827	801.92
	RF-MRP	1780	**5692**	8810	659.32
	MRP-TBR	1783	5702	**8809**	**34.27**
Legumes	SPR	108	651	1175	21130.08

(558 taxa; 19 trees)	SPR-RF-TIES	**92**	471	1037	12376.00
	SPR-MRP	110	511	903	276.49
	RF-Ratchet	117	**401**	1102	1349.56
	RF-MRP	130	429	1068	1558.60
	MRP-TBR	140	519	**891**	**579.76**

Table 2: Aggregate SPR distance to supertrees constructed from different rootings of the bacterial protein trees. Six different construction methods were compared: The MRP supertree (MRP), the SPR supertree constructed from the 40 largest trees with a monophyletic Aquificae group (40-Aquificae), the SPR supertrees constructed using the MRP supertree (SPR-MRP) or SPR-Aquificae supertree (SPR-Aquificae), and the SPR supertrees constructed by only rooting the gene trees using the MRP supertree (SPR-MRP-Rooting) or SPR-Aquificae tree (SPR-Aquificae-Rooting) and building a greedy addition supertree. Each supertree was compared to the MRP rooted gene trees or SPR-Aquificae rooted gene trees with the SPR distance.

MRP rooted gene trees	
	SPR Distance
SPR-MRP-Rooting	52867
SPR-MRP	52896
MRP	52896
SPR-Aquificae-Rooting	58539
SPR-Aquificae	59561
40-Aquificae	60611

SPR-Aquificae rooted gene trees	
	SPR Distance
SPR-Aquificae-Rooting	53534
SPR-Aquificae	54488
40-Aquificae	55570
SPR-MRP-Rooting	58023
SPR-MRP	58057
MRP	58057

Table 3: Dissimilarity of supertrees constructed from the same rooting of bacterial protein trees. We compared the minimal SPR distance between any rooting of the SPR supertree constructed from the 40 largest trees with a monophyletic Aquificae group (40-Aquificae), the SPR supertrees constructed using the MRP supertree (SPR-MRP) or SPR-Aquificae supertree (SPR-Aquificae), and the SPR supertrees constructed by only rooting the gene trees using the MRP supertree (SPR-MRP-Rooting) or SPR-Aquificae tree (SPR-Aquificae-Rooting) and building a greedy addition supertree.

	SPR-Aquificae	SPR-Aquificae-Rooting	SPR-MRP	SPR-MRP-Rooting
SPR-Aquificae	0	10	34	33
SPR-Aquificae-Rooting	10	0	27	25
SPR-MRP	34	27	0	2
SPR-MRP-Rooting	33	25	2	0

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18

FIGURE LEGENDS

Figure 1: The equivalence between the SPR distance and MAF size. (a) The species tree S and gene tree G differ only in the placement of the grey subtree. The roots of these trees are denoted by ρ. (b) The MAF of S and G is produced by cutting the dotted edge in both trees. (c) Each component of an MAF other than the component containing ρ represents an SPR move. A single SPR move transforms S into G by moving the grey subtree in S to its position in G. (d) Each SPR move models an LGT event in the reverse direction. From the MAF of S and G we infer that a transfer of gene G has occurred from an ancestor of taxon 1 to an ancestor of taxon 4.

Figure 2. SPR supertree constructed using Aquificae as outgroup. Genera such as *Mycobacterium* with multiple representatives are shown as collapsed subtrees for brevity. Colours indicate the classes of bacteria.

Figure 3. Inferred LGT events between 135 distinct bacterial genera. (a) An LGT heatmap. The coloured side bars indicate class using the colour mapping of Figure 2. The row and column genus order is the same. The number of transfers is shown in a white-yellow-red colour scale with darker colours indicating a higher proportion of transfer events. Colour intensity is relative to the largest number of transfers in a row. Relationships with fewer than 5% of the maximum transfer events for a row or only a single transfer event were filtered out. (b) Each node of the LGT affinity graph represents a bacterial genus, colored by class and scaled relative to the number of genomes representing that genus (1-15). Two genera are connected by an edge if the number of inferred LGT events between them exceeds 5% of the number of homologous genes

common to both genomes. The shade of an edge is proportional to this ratio of LGT events to common genome size; black edges indicate relationships with at least as many LGT events as the size of their common genome. The thickness of an edge scales relative to the actual number of inferred transfers (between 2 and 370) with thicker edges indicating more transfers. The graph is shown with a spring-loaded layout.

Figure 4: Mean time required to compare gene trees with a given SPR distance from an SPR supertree of a 244-genome dataset. The time axis is on a log scale as the time required increases exponentially with the SPR distance. The left panel compares our previous ($2.42^k n$) and new ($2^k n$) algorithms, with (C) and without clustering, on the set of binary trees. The right panel compares our new algorithm with and without clustering on the set of trees with unsupported bipartitions collapsed. Note that collapsing bipartitions reduces the SPR distance.

Figure 5: A comparison of our LGT rate simulation parameter to the bacterial dataset. Supertrees of empirical data have the same mean SPR distance to leaf ratio (within 95% confidence intervals) as our simulations with a random LGT rate less than 0.2 and a divergence-biased LGT rate less than 0.4.

Figure 6: A comparison of the mean supertree error (as measured by the SPR distance) of RF supertrees (RF) to SPR supertrees using the default parameters (SPR), seeded with an MRP starting tree (SPR-MRP), or seeded with the correct tree (SPR-C).

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18

Figure 7: A comparison of the accuracy of SPR and MPR supertrees with known or unknown gene tree roots. The upper panels compare the mean supertree error (as measured by the minimal SPR distance to any rooting of a supertree) when the gene trees are correctly rooted. We compared MRP supertrees (MRP) to SPR supertrees using the default parameters (SPR), seeded with an MRP starting tree (SPR-MRP), or seeded with the correct tree (SPR-C). The lower panels compare the mean error of the MRP supertree to SPR supertrees when the gene tree roots are unknown, using our balanced accuracy based simple unrooted comparison without and with an MRP seed tree (SPR-SU and SPR-MRP-SU, respectively).

Figure 8: Comparison of SPR and MRP supertrees of 244 bacterial genomes. The SPR supertree on the left was constructed with the Aquificae as outgroup while the MRP supertree on the right is unrooted and places the Aquificae as neighbours of the Epsilonproteobacteria. Both figures show the largest monophyletic group of each class as a collapsed subtree and all members of a given class with the same color.

Supplemental Figure 1. Inferred LGT events between 13 bacterial classes. (a) LGT heatmap. The colour side bars indicate class. The row and column order is the same. The number of transfers is shown in a white-yellow-red colour scale with darker colours indicating a higher proportion of transfer events. Colour intensity is relative to the largest number of transfers in a row. Relationships with fewer than 5% of the maximum transfer events for a row or only a single transfer event were filtered out. (b) LGT affinity graph of the bacterial classes. Each node of the graph represents a bacterial class scaled relative to the number of represented taxa (2-75). Two genera are connected by an edge if the number of inferred LGT events between them exceeds 5%

of their shared genes. The shade of an edge is proportional to this ratio of LGT events to shared genes; black edges indicate relationships with at least as many LGT events as shared genes. The thickness of an edge scales relative to the actual number of inferred transfers (30-1414) with thicker edges indicating more transfers.

Supplemental Figure 2: The LGT affinity neighbourhood of genus *Clostridium*. Each node of the graph represents a bacterial genus coloured by class and scaled relative to the number of represented taxa (1-13). Two genera are connected by an edge if the number of inferred LGT events between them exceeds 5% of their shared genes. The shade of an edge is proportional to this ratio of LGT events to shared genes; black edges indicate relationships with at least as many LGT events as shared genes. The thickness of an edge scales relative to the actual number of inferred transfers (2-125) with thicker edges indicating more transfers.

Supplemental Figure 3: A comparison of the accuracy of SPR, RF and MPR supertrees as measured by the minimal SPR distance between simulated species histories and any rooting of the supertree under varying rates of random or divergence-biased simulated LGT events.

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18

Supplemental Material

Appendix 1: Fast MAF Algorithm

In this appendix we discuss the efficiency and practicality improvements of our new MAF algorithm. We first introduce our previous algorithm (Whidden et al. 2010; Whidden et al. 2013a) whose running time is bounded by $O(2.42^k n)$ for two binary trees with n leaves and an SPR distance of k. We then introduce our novel concept of "protecting" edges during the search for an MAF. This "edge protection" scheme allows us to avoid exploring the same edge cutting scenarios multiple times and greatly speeds up the search for an MAF, as we demonstrated in Figure 5. In a forthcoming paper (Whidden and Zeh 2013) we give the full details of this algorithm and prove that its running time is bounded by $O(2^k n)$. Finally, we explain how we extended our algorithm to compute MAFs of a binary and multifurcating tree and thereby account for uncertainty in the gene trees input to our supertree method. In a recently submitted manuscript (Whidden et al. 2013b) we gave the full details of this algorithm as applied to two multifurcating trees and proved that its running time remains bounded by $O(2.42^k n)$. However, by requiring that one tree be binary and applying edge protection our new MAF algorithm requires roughly the same time in practice to compute an MAF regardless of whether the other tree is multifurcating, as we demonstrated in Figure 5.

Previous MAF Algorithm.—Our previous MAF algorithm (Whidden et al. 2010; Whidden et al. 2013a) takes two binary trees T_1 and T_2 as input along with a parameter k and returns an agreement forest with at most k+1 components (and thus k edge cuts) if and only if such an agreement forest exists. To find an MAF, we run this algorithm with increasing values of k from 0 until an agreement forest is found. Since the running time of the algorithm scales exponentially with k, this entire procedure only takes a small constant factor more time than the invocation that

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18

finds the MAF. Our algorithm proceeds in a bottom-up fashion from the leaves of T_1. T_1 remains a tree through this procedure but T_2 may become a forest, denoted F_2. We maintain a set of *sibling pairs*, sibling subtrees (a,c) in T_1 such that identical subtrees a and c exist in F_2. The algorithm examines each such sibling pair in turn and applies one of three cases:

(1) If a and c are also siblings in F_2, then the subtree rooted at their parent is identical in T_1 and F_2 and so becomes a candidate for membership in a sibling pair,

(2) If a or c is a component of F_2 then it must be cut off in T_1,

(3) We identify at most 3 sets of edges in F_2 such that cutting one of these edge sets will lead to an MAF and try each edge set recursively in turn.

Case (3), which defines multiple edge sets to consider for cutting, requires detailed explanation. Assume that a is the deeper subtree of F_2, if a and c are in the same component, and let b be the sibling of a in F_2. If a and c are in separate components of F_2 then cutting off a or c will lead to an MAF. If a and c are in the same component but only one subtree, b, is on the path between them then cutting off b will always lead to an MAF. Otherwise, cutting off a, c, or simultaneously cutting off all of the subtrees between a and c in F_2 will lead to an MAF. Note that this last case is the worst case of our algorithm as it splits our computation into three branches cutting one, one, or at least two edges respectively. We previously showed in Whidden et al. (2010) that this last case bounds our running time of $O(2.42^k n)$ with a recurrence relation analysis.

New MAF Algorithm.—Our improved algorithm introduces the concept of *edge protection* to alleviate the bottleneck of the 3-way branching case of our previous MAF algorithm. Observe that if some MAF can be found by a recursive invocation of this case that cuts off subtree a in F_2 then an MAF will be found by this invocation. Thus, we can assume that cutting off subtree a does not lead to an MAF in the recursive invocation that cuts off subtree c, or we would have already found it. We *protect* edge *a* in this search branch to denote this and ignore any recursive invocations that cut a protected edge. By ignoring these search paths we reduce the running time of the algorithm to $O(2^k n)$. The proof of this bound is highly technical, as it relies on showing that this edge protection either forces our best case, cutting subtree *b* without branching, or avoids enough search branches to achieve this bound and requires some additional boundary cases. In a forthcoming paper (Whidden and Zeh 2013) we will provide the full details of our algorithm and prove this bound.

We have also developed a theory for MAFs of multifurcating trees to incorporate uncertainty in gene trees. In a recently submitted manuscript (Whidden et al. 2013b) we developed a general MAF algorithm for two multifurcating trees. This algorithm is based on our $O(2.42^k n)$ algorithm for binary trees and achieves the same running time but is significantly more complicated and requires many more cases. For the purposes of constructing SPR supertrees, however, we only need to allow that the gene trees be multifurcating; the supertree is binary. By requiring that T_1 be binary in our MAF algorithm these extra cases disappear and we can use the same overall algorithm structure but with the ability to resolve multifurcations as well as cut edges. Our MAF algorithm when T_2 is multifurcating still examines each sibling pair in turn and applies one of three cases:

(1) if *a* and *c* are also siblings in F_2, then either the subtree rooted at their parent is identical in T_1 and F_2 and so becomes a candidate for membership in a sibling pair or we resolve the multifurcation of their parent in F_2 to separate them so that this occurs.

(2) If *a* or *c* is a component of F_2 then it must be cut off in T_1.

(3) We identify at most 3 sets of edges in F_2 such that cutting one of these edge sets will lead to an MAF and try each edge set recursively in turn.

We again assume that *a* is the deeper subtree of F_2, if *a* and *c* are in the same component. Since F_2 is multifurcating, *a* may now have multiple siblings and we represent them collectively by *B* which we call a pendant subtree. If *a* and *c* are in separate components of F_2 then cutting off *a* or *c* will again lead to an MAF. If *a* and *c* are in the same component separated only by *B* then either cutting off *c* or resolving *B* separately from *a* and cutting the introduced edge will lead to an MAF. Otherwise, cutting off *a*, *c*, or resolving and cutting off all pendant subtrees of the path from *a* to *c* in F_2 will lead to an MAF. We further apply edge protection to this last case as in our improved binary algorithm. Note that this procedure is essentially identical to our prior binary algorithm with the exception that our previous best case, where we could bring *a* and *c* together in F_2 with a single cut now requires us to branch into two possibilities. Fortunately, cutting off *c* is never necessary when *a*'s parent is binary, that is, *B* is a single node *b*, so this has a negligible running time impact in practice, as we demonstrated in Figure 5. This does, however, preclude the argument we used to prove that edge protection reduces the running time of the binary MAF algorithm to $O(2^k n)$ so the running time of our MAF algorithm when one tree is multifurcating remains $O(2.42^k n)$ in the worst case.

APPENDIX 2: LINEAR-TIME CLUSTER REDUCTION

In this appendix we explain how to accelerate the computation of MAFs (and, thus, the SPR distance) using the Cluster Reduction of Linz and Semple (2011). This reduction partitions the input trees into pairs of subtrees, or clusters, that can be solved iteratively and reassembled into a full solution. The time required to solve these clusters with our MAF algorithms scales exponentially with the maximum number of components in an MAF of any cluster rather than the full MAF of the trees so this strategy greatly accelerates the recovery of MAFs in practice.

The Cluster Reduction as originally formulated is only suitable to compute an MAF variant, weighted MAFs, that cannot be computed with our algorithms. We first extend the Cluster Reduction to apply to ordinary MAFs and then show how to identify clusters in linear time, greatly improving on the previous cubic time algorithm.

Linz and Semple defined a *cluster* of two trees T_1 and T_2 to be a pair of subtrees T_1^e and T_2^f, for appropriate edges e in T_1 and f in T_2 such that both trees have the same set of labelled leaves. A *cluster sequence* of T_1 and T_2 is a sequence of tree pairs T = (T_1^1, T_2^1), (T_1^2, T_2^2), …, (T_1^t, T_2^t), (T_1^ρ, T_2^ρ) defined inductively as follows: if t = 0, then $T_1^\rho = T_1$ and $T_2^\rho = T_2$. If t > 0 then (T_1^1, T_2^1) is a cluster of T_1 and T_2 with at least two taxa, the roots of T_1^1 and T_2^1 are labelled with a new label ρ_1, and (T_1^2, T_2^2), …, (T_1^t, T_2^t), (T_1^ρ, T_2^ρ) is a cluster sequence of the two trees obtained from T_1 and T_2 by replacing the subtrees T_1^1 and T_2^2 with a single labelled leaf a_1. This is illustrated in Figure A1. Clearly, ρ is the root of T_1^{ρ} and T_2^{ρ}. An agreement forest F of T is the disjoint union of forests $F = F_1 \cup F_2 \cup \dots \cup F_t \cup F_\rho$, where F_i is an agreement forest of T_1^i and T_2^i, for all i in {1, 2, …, t, ρ}. The weight of F is defined to be $w(F) = |F| - |\{(p_i, a_i)$: p_i and a_i are singletons in $F\}| - t$, where |F| denotes the number of trees in F. We say that F is an MAF of T if it has minimum weight among all agreement forests of T. The key result proved by Linz and Semple is that the weight of an MAF of any cluster sequence is exactly the number of components in an MAF of the original trees. They also provided a divide-and-conquer approach for computing an MAF of T: Process the clusters in order, for each i computing an agreement forest F_i of T^i and T_2^i. If $F = F_1 \cup F_2 \cup \dots \cup F_{i-1}$ is the union of forests computed so far (for i=ρ, let i-1=t), then F_i is computed to be an agreement forest of T^i and T^i that minimizes $w(F_i) = |F_i| - |\{(\rho_j, a_j)$: ρ_j is a singleton in F and a_j is a singleton in $F_i\}|$. This weight corrects for the fact that we have cut the same edge twice; ρ_j and a_j are nodes introduced by the cluster reduction to represent the intersection of two clusters so the edge below ρ_j and above a_j are the same edge. Thus, for $i \neq p$, we choose F_i to be an agreement forest of T_1^i and T_2^i that minimizes this weight and such that p_i is a singleton, if possible, to capitalize on this correction. The final forest defined in this way is an MAF of T.

We used the key observation of the cluster reduction, that it is best to cut the root edge of each cluster when possible, to modify this procedure to compute unweighted MAFs. We first compute the cluster sequence as above. We then apply a modification (described below) of our MAF algorithm that returns an MAF of the current cluster such that it has the root edge cut if and only if any MAF of the current cluster i has an isolated ρ_i. If the root edge, below ρ_i, was cut in this MAF then we separate the two clusters by simply cutting the edge above a_i in its corresponding cluster and then removing a_1 and ρ_1 completely to avoid counting this double cut. If the root edge is not cut then we reattach the two clusters by cutting this root edge, removing ρ_1, and then replacing a_1 with the subtree formerly rooted by ρ_i (thereby removing this subtree from the agreement forest of the current cluster). We apply this procedure iteratively to the cluster sequence and then take the union of these forests as our MAF. We have removed each ρ_i and a_i so this is an unweighted MAF. To see that this is indeed an MAF, observe that we apply the same procedure as Linz and Semple for each cluster other than our treatment of ρ_i and a_i. If ρ_i is not isolated in a given cluster, then we remove one component from our forest by replacing a_i, whereas the weighted algorithm applies a weight of -1 (from the –t factor) to compensate. If ρ_i is isolated in a given cluster then we remove ρ_i (equivalent to the -1 weight) and remove a_i (equivalent to the singleton portion of the weight calculation, this reduces the weight by 1 if p_i and a_i are singletons in some weighted MAF). Thus, our computed forest has exactly as many components as the weight of some weighted MAF and is indeed an MAF.

We now explain how we modified our MAF algorithm to prefer MAFs with isolated roots. Recall that each recursive step of our algorithm identifies at most three edge sets to cut from the intermediate forests and tries each edge set in turn. If more than one of these edge set choices lead to an MAF then our algorithm arbitrarily chooses one of them. We simply modified our algorithm to instead select between these at most three MAFs by preferring MAFs with their root edge cut. Since our algorithm does not find all MAFs of the two trees, it is not immediately obvious that this change is sufficient to find one MAF where the root edge is cut if such an MAF exists. However, the correctness proof of our previous MAF algorithm (Whidden et al. 2013a) and our forthcoming correctness proofs start with an arbitrary agreement forest F and construct an agreement forest F' from F that has no more components than F and such that our algorithms find F'. If we choose F to be an agreement forest where ρ_i is a singleton, then this construction ensures that F' also contain ρ_i as a singleton. In other words, if there exists an MAF that has ρ_i as a singleton, our algorithms find one such MAF.

Finally, we developed a linear-time algorithm for computing a cluster sequence, greatly improving on the naïve cubic algorithm. Let n be the number of leaves in T_1 and T_2. The cubic algorithm compares each of the subtrees of T_1, starting at the leaves, to each subtree of T_2 and appends each found cluster to the cluster sequence. There are O(n) subtrees in each tree and it takes O(n) time to compare two leaf sets so this procedure requires $O(n^3)$ time. We improve on this by using least common ancestors (LCAs). The LCA of two or more nodes in a tree is their common ancestor furthest from the root. Let s_1 be a subtree of T_1 with leaf set L_1 and s_2 be a subtree of T_2 with leaf set L_2. Observe that these subtrees have the same leaf set if and only if the LCA of L_1 in T_2 is s_2 and the LCA of L_2 in T_2 is s_1. Efficient least common ancestor (LCA) query structures exist (e.g., Bender and Farach-Colton 2000) that can be built in O(n) time and that allow for constant time LCA queries of two nodes. We use such a structure to compute a mapping M of T_1 subtrees to the LCAs of their leaf sets in T_2. First, for each leaf x in T_1, we set M(x) to the corresponding leaf x of T_2. Then, for any node n of T_1 with children c_1 and c_2 such that the mapping $M(c_1)$ and $M(c_2)$ have been defined, we compute $M(n) = LCA(M(c_1), M(c_2))$. We apply this procedure again with T_1 and T_2 reversed to compute the mapping M^{-1} of T_2 subtrees to the LCAs of their leaf sets in T_1. Finally, for each subtree s_1 of T_1 in a bottom-up postorder traversal we check if s_1 is a cluster by checking if $M^{-1}(M(s_1)) = s_1$ and, if so, appending s_1 and $M(s_1)$ to the cluster sequence.

Supplemental Table 1: List of 244 bacterial genomes included in this work.

Class	Taxon
Actinobacteria	*Acidimicrobium ferrooxidans* DSM 10331
	Acidothermus cellulolyticus 11B
	Amycolatopsis mediterranei U32
	Arcanobacterium haemolyticum DSM 20595
	Arthrobacter aurescens TC1
	Arthrobacter sp. FB24
	Beutenbergia cavernae DSM 12333
	Bifidobacterium adolescentis ATCC 15703
	Bifidobacterium animalis subsp. lactis AD011
	Bifidobacterium animalis subsp. lactis Bl-04
	Bifidobacterium longum NCC2705
	Bifidobacterium longum subsp. infantis ATCC 15697
	Bifidobacterium longum subsp. longum JDM301
	Catenulispora acidiphila DSM 44928
	Cellulomonas flavigena DSM 20109
	Clavibacter michiganensis subsp. michiganensis NCPPB 382
	Corynebacterium aurimucosum ATCC 700975
	Corynebacterium efficiens YS-314
	Corynebacterium glutamicum ATCC 13032 DSM 20300
	Corynebacterium glutamicum R
	Corynebacterium jeikeium K411
	Corynebacterium kroppenstedtii DSM 44385
	Corynebacterium pseudotuberculosis FRC41
	Corynebacterium urealyticum DSM 7109
	Cryptobacterium curtum DSM 15641
	Eggerthella lenta DSM 2243
	Frankia alni ACN14a
	Gardnerella vaginalis 409-05
	Geodermatophilus obscurus DSM 43160
	Gordonia bronchialis DSM 43247
	Jonesia denitrificans DSM 20603
	Kribbella flavida DSM 17836

Kytococcus sedentarius DSM 20547

Leifsonia xyli subsp. xyli str. CTCB07

Mobiluncus curtisii ATCC 43063

Mycobacterium avium 104

Mycobacterium avium subsp. paratuberculosis K-10

Mycobacterium bovis AF2122/97

Mycobacterium bovis BCG str. Pasteur 1173P2

Mycobacterium bovis BCG str. Tokyo 172

Mycobacterium gilvum PYR-GCK

Mycobacterium leprae Br4923

Mycobacterium leprae TN

Mycobacterium marinum M

Mycobacterium smegmatis str. MC2 155

Mycobacterium sp. KMS

Mycobacterium tuberculosis F11

Mycobacterium tuberculosis H37Ra

Mycobacterium tuberculosis H37Rv

Mycobacterium vanbaalenii PYR-1

Nakamurella multipartita DSM 44233

Nocardia farcinica IFM 10152

Nocardioides sp. JS614

Propionibacterium acnes KPA171202

Propionibacterium freudenreichii subsp. shermanii CIRM-BIA1

Rhodococcus erythropolis PR4

Rhodococcus jostii RHA1

Rothia mucilaginosa DY-18

Salinispora arenicola CNS-205

Salinispora tropica CNB-440

Sanguibacter keddieii DSM 10542

Segniliparus rotundus DSM 44985

Slackia heliotrinireducens DSM 20476

Stackebrandtia nassauensis DSM 44728

Streptomyces avermitilis MA-4680

Streptomyces griseus subsp. griseus NBRC 13350

Streptomyces scabiei 87.22

	Streptosporangium roseum DSM 43021
	Thermobifida fusca YX
	Thermobispora bispora DSM 43833
	Thermomonospora curvata DSM 43183
	Tropheryma whipplei TW08/27
	Tsukamurella paurometabola DSM 20162
	Xylanimonas cellulosilytica DSM 15894
Alphaproteobacteria	*Bradyrhizobium* sp. BTAi1
	Candidatus Hodgkinia cicadicola Dsem
	Candidatus Pelagibacter ubique HTCC1062
	Ehrlichia canis str. Jake
	Ehrlichia chaffeensis str. Arkansas
	Erythrobacter litoralis HTCC2594
	Gluconacetobacter diazotrophicus PAl 5
	Mesorhizobium loti MAFF303099
	Ochrobactrum anthropi ATCC 49188
	Parvularcula bermudensis HTCC2503
	Rickettsia akari str. Hartford
	Rickettsia canadensis str. McKiel
	Rickettsia peacockii str. Rustic
	Rickettsia rickettsii str. Sheila Smith
	Wolbachia endosymbiont of *Culex quinquefasciatus* Pel
Aquificae	*Aquifex aeolicus* VF5
	Hydrogenobacter thermophilus TK-6
	Hydrogenobaculum sp. Y04AAS1
	Persephonella marina EX-H1
	Sulfurihydrogenibium azorense Az-Fu1
	Sulfurihydrogenibium sp. YO3AOP1
	Thermocrinis albus DSM 14484
Bacilli	*Bacillus anthracis* str. Sterne
	Bacillus cereus 03BB102
	Bacillus cereus AH187
	Bacillus cereus ATCC 10987
	Bacillus cereus G9842
	Bacillus cereus Q1

	Bacillus clausii KSM-K16
	Bacillus thuringiensis BMB171
	Bacillus thuringiensis str. Al Hakam
	Enterococcus faecalis V583
	Exiguobacterium sibiricum 255-15
	Exiguobacterium sp. AT1b
	Geobacillus sp. WCH70
	Lactobacillus acidophilus NCFM
	Lactobacillus casei ATCC 334
	Lactobacillus casei str. Zhang
	Lactobacillus crispatus ST1
	Lactobacillus reuteri JCM 1112
	Lactobacillus rhamnosus Lc 705
	Lactobacillus salivarius UCC118
	Lactococcus lactis subsp. cremoris MG1363
	Leuconostoc kimchii IMSNU 11154
	Listeria monocytogenes HCC23
	Listeria monocytogenes serotype 4b str. CLIP 80459
	Listeria monocytogenes serotype 4b str. F2365
	Staphylococcus aureus RF122
	Staphylococcus carnosus subsp. carnosus TM300
	Staphylococcus lugdunensis HKU09-01
	Streptococcus gordonii str. Challis substr. CH1
	Streptococcus mitis B6
	Streptococcus mutans NN2025
	Streptococcus pneumoniae 670-6B
	Streptococcus pneumoniae JJA
	Streptococcus pyogenes MGAS10270
	Streptococcus pyogenes MGAS10394
	Streptococcus pyogenes MGAS10750
	Streptococcus pyogenes NZ131
	Streptococcus pyogenes str. Manfredo
	Streptococcus suis 98HAH33
	Streptococcus thermophilus LMD-9
Betaproteobacteria	*Azoarcus* sp. BH72

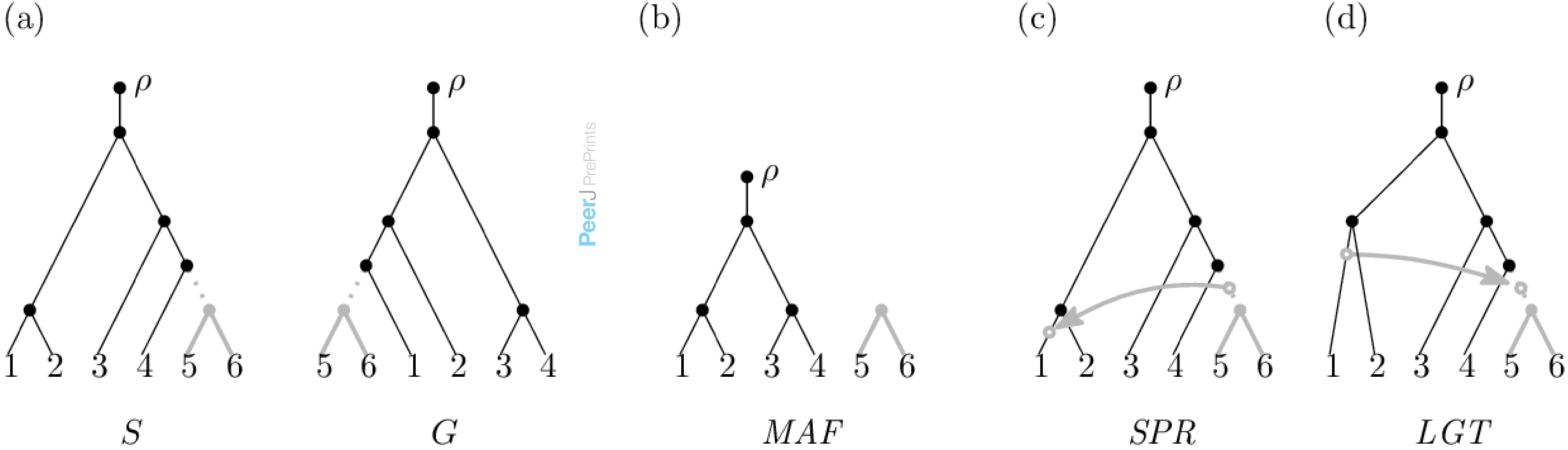
	Bordetella parapertussis 12822
	Burkholderia ambifaria MC40-6
	Burkholderia sp. 383
	Burkholderia vietnamiensis G4
	Candidatus Accumulibacter phosphatis clade IIA str. UW-1
	Gallionella capsiferriformans ES-2
	Methylibium petroleiphilum PM1
	Methylobacillus flagellatus KT
	Methylotenera mobilis JLW8
	Methylotenera sp. 301
	Nitrosomonas europaea ATCC 19718
	Ralstonia pickettii 12D
	Ralstonia solanacearum CFBP2957
	Thiobacillus denitrificans ATCC 25259
Clostridia	*Acetohalobium arabaticum* DSM 5501
	Acidaminococcus fermentans DSM 20731
	Ammonifex degensii KC4
	Caldicellulosiruptor obsidiansis OB47
	Caldicellulosiruptor saccharolyticus DSM 8903
	Caldicelulosiruptor becscii DSM 6725
	Clostridiales genomosp. BVAB3 str. UPII9-5
	Clostridium acetobutylicum ATCC 824
	Clostridium botulinum A str. ATCC 19397
	Clostridium botulinum B str. Eklund 17B
	Clostridium botulinum Ba4 str. 657
	Clostridium botulinum E3 str. Alaska E43
	Clostridium cellulovorans 743B
	Clostridium difficile CD196
	Clostridium kluyveri DSM 555
	Clostridium kluyveri NBRC 12016
	Clostridium perfringens ATCC 13124
	Clostridium perfringens SM101
	Clostridium tetani E88
	Clostridium thermocellum ATCC 27405
	Coprothermobacter proteolyticus DSM 5265

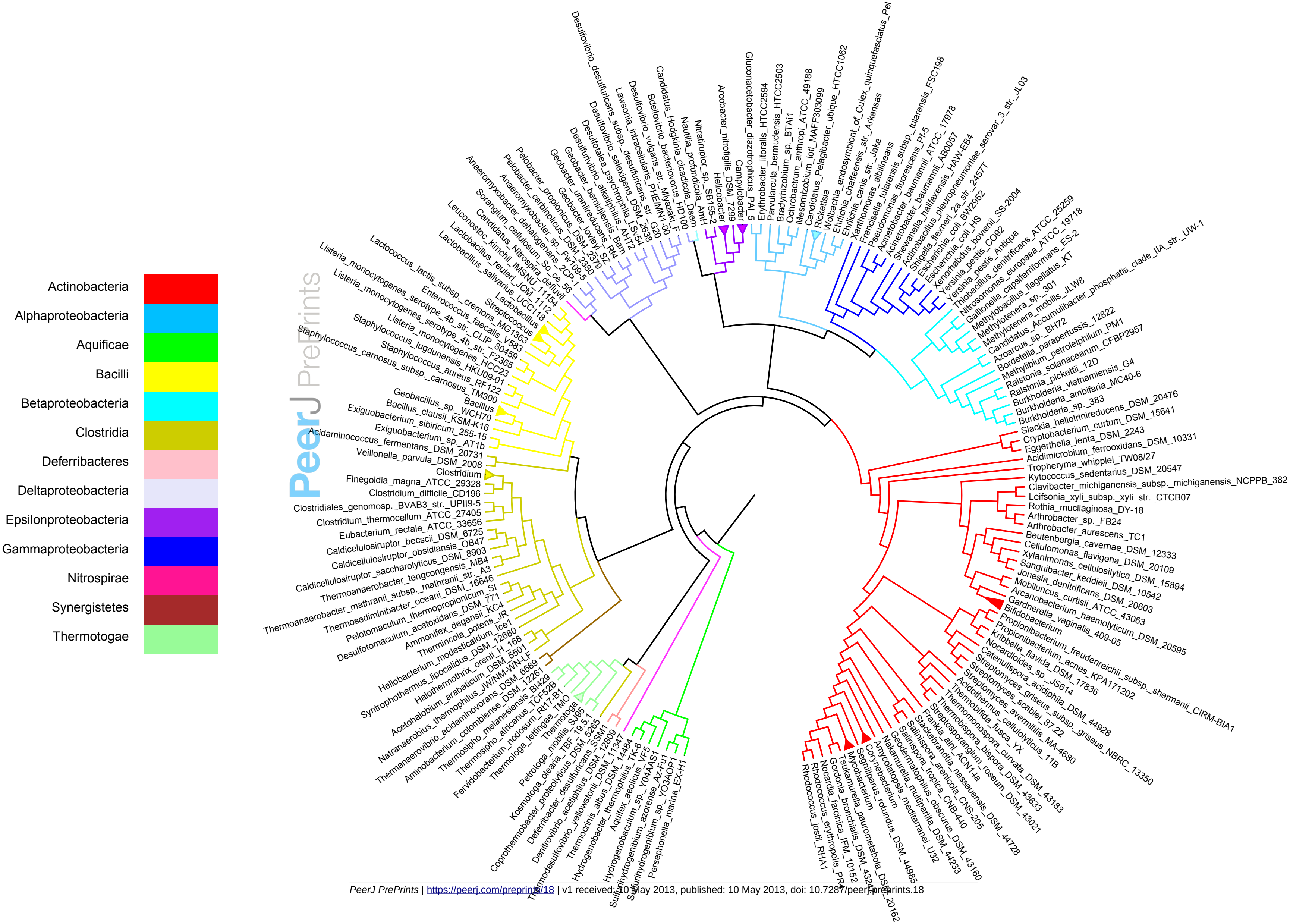
	Desulfotomaculum acetoxidans DSM 771
	Eubacterium rectale ATCC 33656
	Finegoldia magna ATCC 29328
	Halothermothrix orenii H 168
	Heliobacterium modesticaldum Ice1
	Natranaerobius thermophilus JW/NM-WN-LF
	Pelotomaculum thermopropionicum SI
	Syntrophothermus lipocalidus DSM 12680
	Thermincola potens JR
	Thermoanaerobacter mathranii subsp. mathranii str. A3
	Thermoanaerobacter tengcongensis MB4
	Thermosediminibacter oceani DSM 16646
	Veillonella parvula DSM 2008
Deferribacteres	*Deferribacter desulfuricans* SSM1
	Denitrovibrio acetiphilus DSM 12809
Deltaproteobacteria	*Anaeromyxobacter dehalogenans* 2CP-1
	Anaeromyxobacter sp. Fw109-5
	Bdellovibrio bacteriovorus HD100
	Desulfotalea psychrophila LSv54
	Desulfovibrio desulfuricans subsp. desulfuricans str. G20
	Desulfovibrio salexigens DSM 2638
	Desulfovibrio vulgaris str. Miyazaki F
	Desulfurivibrio alkaliphilus AHT2
	Geobacter bemidjiensis Bem
	Geobacter lovleyi SZ
	Geobacter uraniireducens Rf4
	Lawsonia intracellularis PHE/MN1-00
	Pelobacter carbinolicus DSM 2380
	Pelobacter propionicus DSM 2379
	Sorangium cellulosum So ce 56
Epsilonproteobacteria	*Arcobacter nitrofigilis* DSM 7299
	Campylobacter concisus 13826
	Campylobacter jejuni subsp. doylei 269.97
	Campylobacter jejuni subsp. jejuni 81116
	Campylobacter jejuni subsp. jejuni NCTC 11168

	Helicobacter acinonychis str. Sheeba
	Helicobacter hepaticus ATCC 51449
	Helicobacter mustelae 12198
	Helicobacter pylori B38
	Helicobacter pylori HPAG1
	Helicobacter pylori J99
	Helicobacter pylori Shi470
	Nautilia profundicola AmH
	Nitratiruptor sp. SB155-2
Gammaproteobacteria	*Acinetobacter baumannii* AB0057
	Acinetobacter baumannii ATCC 17978
	Actinobacillus pleuropneumoniae serovar 3 str. JL03
	Escherichia coli BW2952
	Escherichia coli HS
	Francisella tularensis subsp. tularensis FSC198
	Pseudomonas fluorescens Pf-5
	Shewanella halifaxensis HAW-EB4
	Shigella flexneri 2a str. 2457T
	Xanthomonas albilineans
	Xenorhabdus bovienii SS-2004
	Yersinia pestis Antiqua
	Yersinia pestis CO92
Nitrospirae	*Candidatus* Nitrospira defluvii
	Thermodesulfovibrio yellowstonii DSM 11347
Synergistetes	*Aminobacterium colombiense* DSM 12261
	Thermanaerovibrio acidaminovorans DSM 6589
Thermotogae	*Fervidobacterium nodosum* Rt17-B1
	Kosmotoga olearia TBF 19.5.1
	Petrotoga mobilis SJ95
	Thermosipho africanus TCF52B
	Thermosipho melanesiensis BI429
	Thermotoga lettingae TMO
	Thermotoga maritima MSB8
	Thermotoga naphthophila RKU-10
	Thermotoga neapolitana DSM 4359

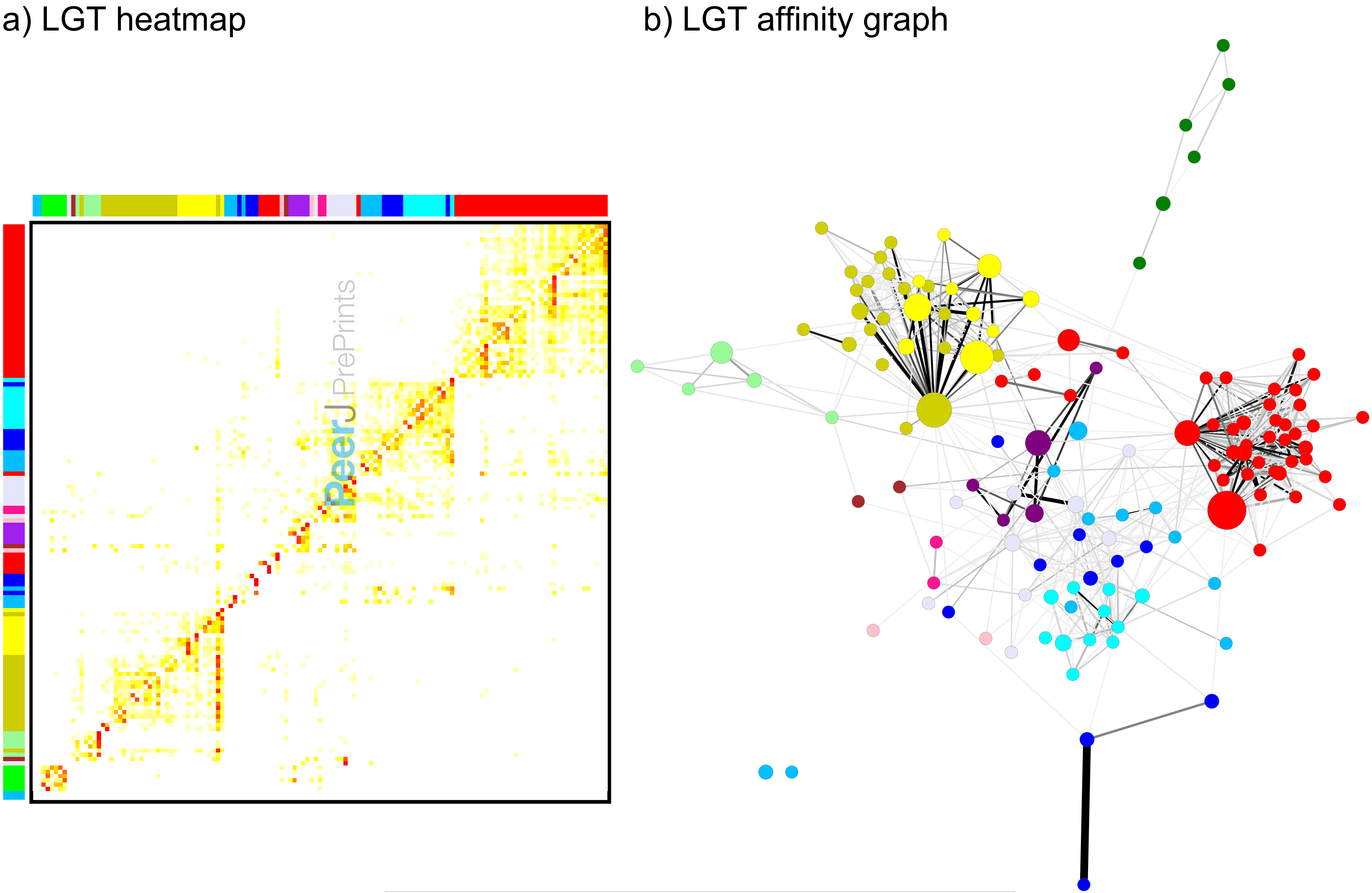
Thermotoga petrophila RKU-1

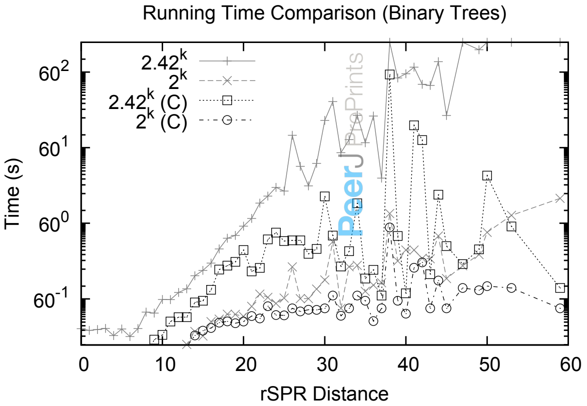
Thermotoga sp. RQ2

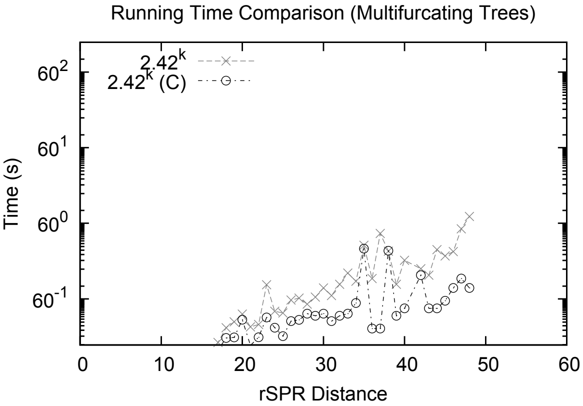

(a)
ρ
1 2 3 4 5 6
S
ρ
5 6 1 2 3 4
G
(b)
ρ
1 2 3 4 5 6
MAF
(c)
ρ
1 2 3 4 5 6
SPR
(d)
ρ
1 2 3 4 5 6
LGT


Actinobacteria 🟥
Alphaproteobacteria 🟦
Aquificae 🟩
Bacilli 🟨
Betaproteobacteria 🟦
Clostridia 🟨
Deferribacteres 🟥
Deltaproteobacteria ⬜
Epsilonproteobacteria 🟪
Gammaproteobacteria 🟦
Nitrospirae 🟥
Synergistetes 🟫
Thermotogae 🟩
Lactobacillus_salivarius_UCC118
Lactobacillus_reuteri_JCM_1112
Leuconostoc_kimchii_IMSNU_11154
Candidatus_Nitrospira_defluvii
Sorangium_cellulosum_So_ce_56
Anaeromyxobacter_dehalogenans_2CP-1
Anaeromyxobacter_sp._Fw109-5
Pelobacter_carbinolicus_DSM_2380
Pelobacter_propionicus_DSM_2379
Geobacter_lovleyi_SZ
Geobacter_uraniireducens_Rf4
Geobacter_bemidjiensis_Bem
Desulfurivibrio_alkaliphilus_AHT2
Desulfotalea_psychrophila_LSv54
Desulfovibrio_salexigens_DSM_2638
Desulfovibrio_desulfuricans_subsp._desulfuricans_str._G20
Lawsonia_intracellularis_PHE/MN1-00
Desulfovibrio_vulgaris_str._Miyazaki_F
Bdellovibrio_bacteriovorus_HD100
Candidatus_Hodgkinia_cicadicola_Dsem
Nautilia_profundicola_AmH
Nitratiruptor_sp._SB155-2
Arcobacter_nitrofigilis_DSM_7299
Helicobacter
Campylobacter
Gluconacetobacter_diazotrophicus_PAl_5
Erythrobacter_litoralis_HTCC2594
Parvularcula_bermudensis_HTCC2503
Bradyrhizobium_sp._BTAi1
Ochrobactrum_anthropi_ATCC_49188
Mesorhizobium_loti_MAFF303099
Candidatus_Pelagibacter_ubique_HTCC1062
Rickettsia
Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel
Ehrlichia_chaffeensis_str._Arkansas
Ehrlichia_canis_str._Jake
Xanthomonas_albilineans
Francisella_tularensis_subsp._tularensis_FSC198
Pseudomonas_fluorescens_Pf-5
Acinetobacter_baumannii_ATCC_17978
Acinetobacter_baumannii_AB0057
Shewanella_halifaxensis_HAW-EB4
Actinobacillus_pleuropneumoniae_serovar_3_str._JL03
Shigella_flexneri_2a_str._2457T
Escherichia_coli_BW2952
Escherichia_coli_HS
Xenorhabdus_bovienii_SS-2004
Yersinia_pestis_CO92
Yersinia_pestis_Antiqua
Thiobacillus_denitrificans_ATCC_25259
Nitrosomonas_europaea_ATCC_19718
Gallionella_capsiferriformans_ES-2
Methylobacillus_flagellatus_KT
Methylotenera_sp._301
Methylotenera_mobilis_JLW8
Candidatus_Accumulibacter_phosphatis_clade_IIA_str._UW-1
Azoarcus_sp._BH72
Bordetella_parapertussis_12822
Methylibium_petroleiphilum_PM1
Ralstonia_solanacearum_CFBP2957
Ralstonia_pickettii_12D
Burkholderia_vietnamiensis_G4
Burkholderia_ambifaria_MC40-6
Burkholderia_sp._383
Slackia_heliotrinireducens_DSM_20476
Cryptobacterium_curtum_DSM_15641
Eggerthella_lenta_DSM_2243
Acidimicrobium_ferrooxidans_DSM_10331
Tropheryma_whipplei_TW08/27
Kytococcus_sedentarius_DSM_20547
Clavibacter_michiganensis_subsp._michiganensis_NCPPB_382
Leifsonia_xyli_subsp._xyli_str._CTCB07
Rothia_mucilaginosa_DY-18
Arthrobacter_sp._FB24
Arthrobacter_aurescens_TC1
Beutenbergia_cavernae_DSM_12333
Cellulomonas_flavigena_DSM_20109
Xylanimonas_cellulosilytica_DSM_15894
Sanguibacter_keddieii_DSM_10542
Jonesia_denitrificans_DSM_20603
Mobiluncus_curtisii_ATCC_43063
Arcanobacterium_haemolyticum_DSM_20595
Gardnerella_vaginalis_409-05
Bifidobacterium
Propionibacterium_freudenreichii_subsp._shermanii_CIRM-BIA1
Propionibacterium_acnes_KPA171202
Kribbella_flavida_DSM_17836
Nocardioides_sp._JS614
Catenulispora_acidiphila_DSM_44928
Streptomyces_griseus_subsp._griseus_NBRC_13350
Streptomyces_scabiei_87.22
Streptomyces_avermitilis_MA-4680
Acidothermus_cellulolyticus_11B
Thermobifida_fusca_YX
Thermomonospora_curvata_DSM_43183
Thermobispora_bispora_DSM_43833
Streptosporangium_roseum_DSM_43021
Frankia_alni_ACN14a
Stackebrandtia_nassauensis_DSM_44728
Salinispora_arenicola_CNS-205
Salinispora_tropica_CNB-440
Geodermatophilus_obscurus_DSM_43160
Nakamurella_multipartita_DSM_44233
Amycolatopsis_mediterranei_U32
Corynebacterium
Segniliparus_rotundus_DSM_44985
Mycobacterium
Tsukamurella_paurometabola_DSM_20162
Gordonia_bronchialis_DSM_43247
Nocardia_farcinica_IFM_10152
Rhodococcus_erythropolis_PR4
Rhodococcus_jostii_RHA1
Persephonella_marina_EX-H1
Sulfurihydrogenibium_sp._YO3AOP1
Sulfurihydrogenibium_azorense_Az-Fu1
Hydrogenobaculum_sp._Y04AAS1
Aquifex_aeolicus_VF5
Hydrogenobacter_thermophilus_TK-6
Thermocrinis_albus_DSM_14484
Thermodesulfovibrio_yellowstonii_DSM_11347
Denitrovibrio_acetiphilus_DSM_12809
Deferribacter_desulfuricans_SSM1
Coprothermobacter_proteolyticus_DSM_5265
Kosmotoga_olearia_TBF_19.5.1
Petrotoga_mobilis_SJ95
Thermotoga
Thermotoga_lettingae_TMO
Fervidobacterium_nodosum_Rt17-B1
Thermosipho_africanus_TCF52B
Thermosipho_melanesiensis_BI429
Aminobacterium_colombiense_DSM_12261
Thermanaerovibrio_acidaminovorans_DSM_6589
Natranaerobius_thermophilus_JW/NM-WN-LF
Acetohalobium_arabaticum_DSM_5501
Halothermothrix_orenii_H_168
Syntrophothermus_lipocalidus_DSM_12680
Heliobacterium_modesticaldum_Ice1
Thermincola_potens_JR
Ammonifex_degensii_KC4
Desulfotomaculum_acetoxidans_DSM_771
Pelotomaculum_thermopropionicum_SI
Thermosediminibacter_oceani_DSM_16646
Thermoanaerobacter_mathranii_subsp._mathranii_str._A3
Thermoanaerobacter_tengcongensis_MB4
Caldicellulosiruptor_saccharolyticus_DSM_8903
Caldicellulosiruptor_obsidiansis_OB47
Caldicelulosiruptor_becscii_DSM_6725
Eubacterium_rectale_ATCC_33656
Clostridium_thermocellum_ATCC_27405
Clostridiales_genomosp._BVAB3_str._UPII9-5
Clostridium_difficile_CD196
Finegoldia_magna_ATCC_29328
Clostridium
Veillonella_parvula_DSM_2008
Acidaminococcus_fermentans_DSM_20731
Exiguobacterium_sp._AT1b
Exiguobacterium_sibiricum_255-15
Bacillus_clausii_KSM-K16
Geobacillus_sp._WCH70
Bacillus
Staphylococcus_carnosus_subsp._carnosus_TM300
Staphylococcus_aureus_RF122
Staphylococcus_lugdunensis_HKU09-01
Listeria_monocytogenes_HCC23
Listeria_monocytogenes_serotype_4b_str._F2365
Listeria_monocytogenes_serotype_4b_str._CLIP_80459
Enterococcus_faecalis_V583
Lactococcus_lactis_subsp._cremoris_MG1363
Streptococcus
Lactobacillus

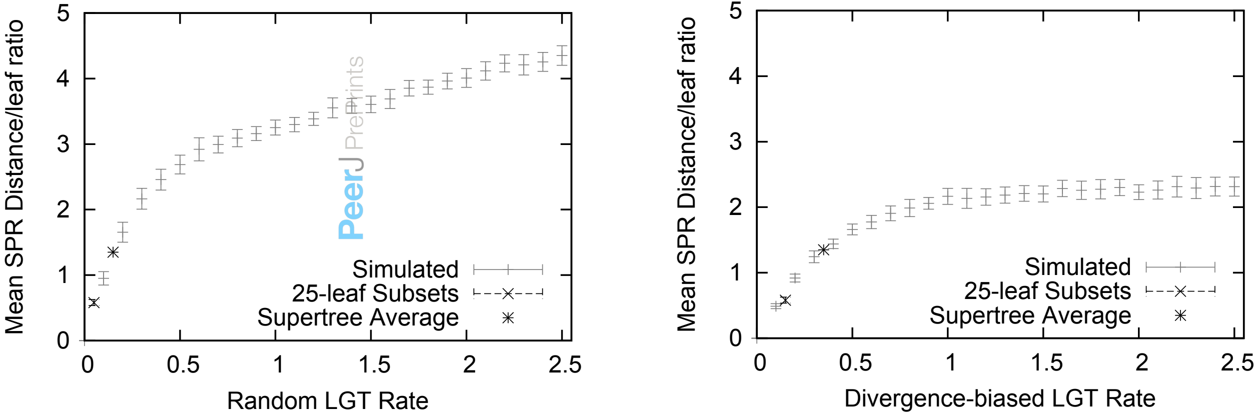
PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18


a) LGT heatmap
b) LGT affinity graph

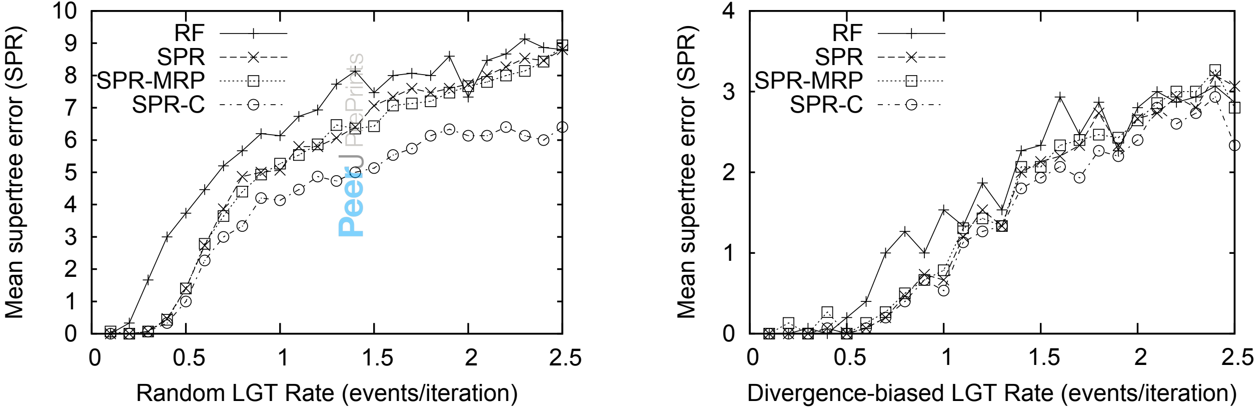

Running Time Comparison (Binary Trees)


Running Time Comparison (Multifurcating Trees)

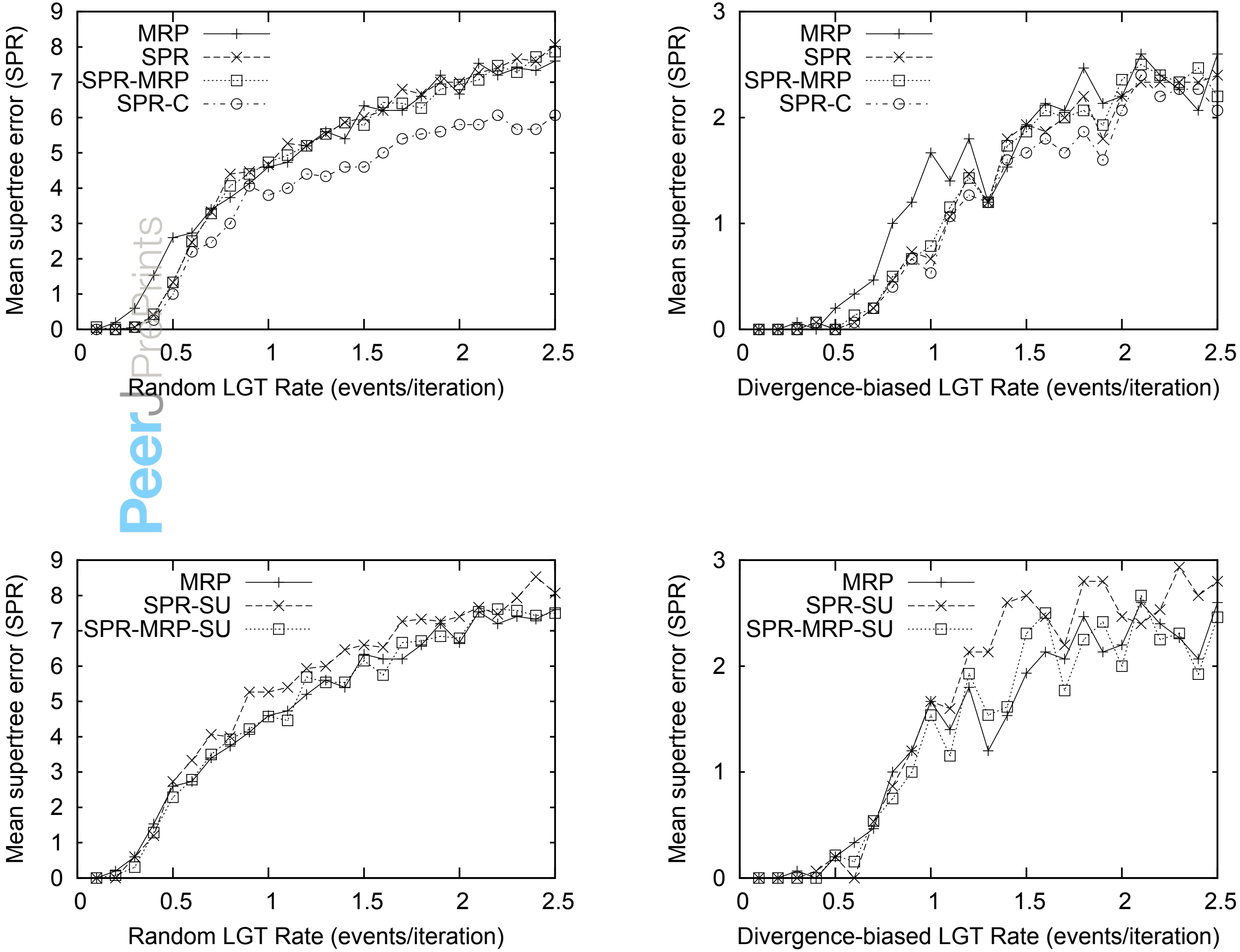
PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18



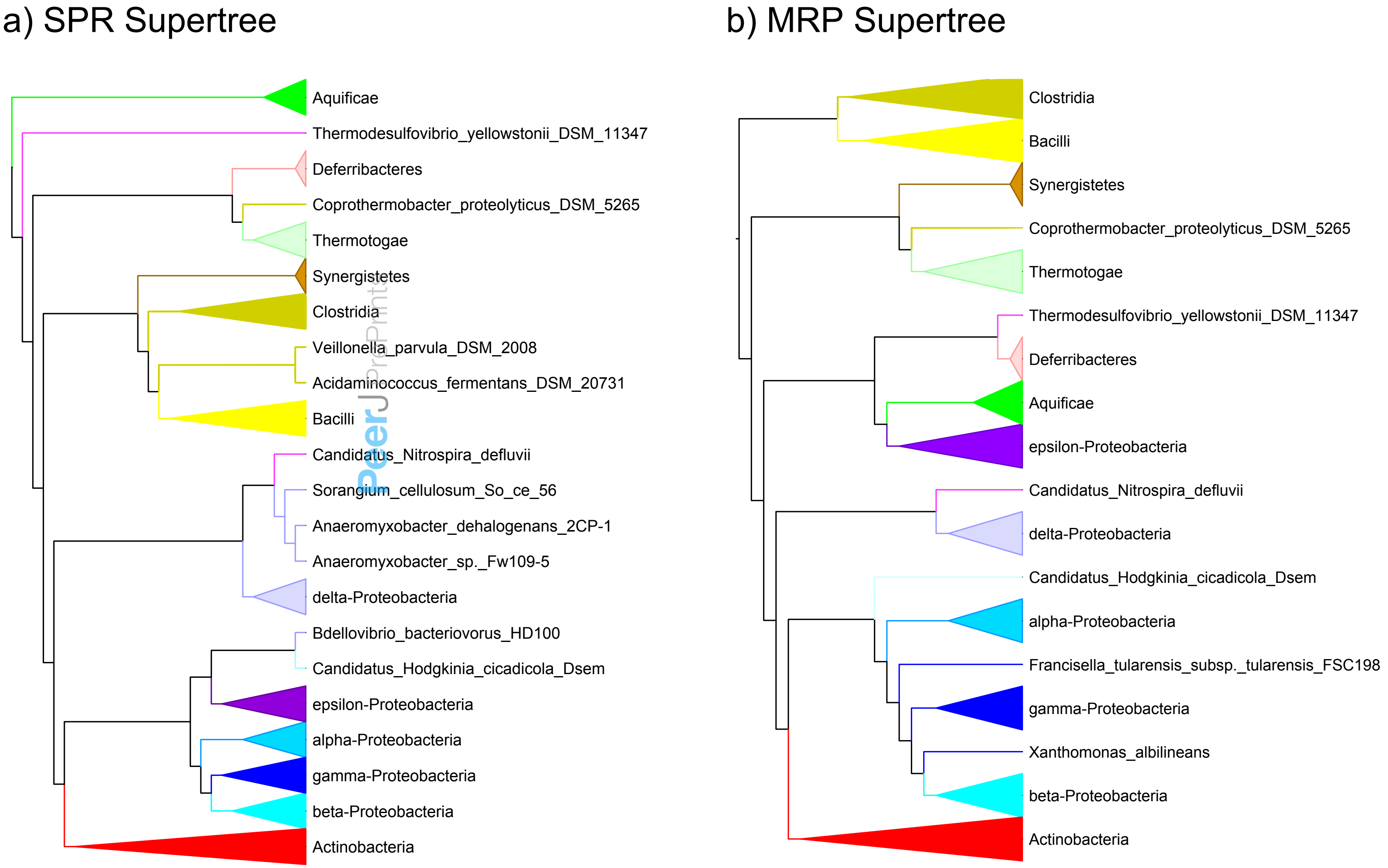
PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18



PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18



PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18



a) SPR Supertree

Aquificae
Thermodesulfovibrio_yellowstonii_DSM_11347
Deferribacteres
Coprothermobacter_proteolyticus_DSM_5265
Thermotogae
Synergistetes
Clostridia
Veillonella_parvula_DSM_2008
Acidaminococcus_fermentans_DSM_20731
Bacilli
Candidatus_Nitrospira_defluvii
Sorangium_cellulosum_So_ce_56
Anaeromyxobacter_dehalogenans_2CP-1
Anaeromyxobacter_sp._Fw109-5
delta-Proteobacteria
Bdellovibrio_bacteriovorus_HD100
Candidatus_Hodgkinia_cicadicola_Dsem
epsilon-Proteobacteria
alpha-Proteobacteria
gamma-Proteobacteria
beta-Proteobacteria
Actinobacteria

b) MRP Supertree

Clostridia
Bacilli
Synergistetes
Coprothermobacter_proteolyticus_DSM_5265
Thermotogae
Thermodesulfovibrio_yellowstonii_DSM_11347
Deferribacteres
Aquificae
epsilon-Proteobacteria
Candidatus_Nitrospira_defluvii
delta-Proteobacteria
Candidatus_Hodgkinia_cicadicola_Dsem
alpha-Proteobacteria
Francisella_tularensis_subsp._tularensis_FSC198
gamma-Proteobacteria
Xanthomonas_albilineans
beta-Proteobacteria
Actinobacteria

PeerJ PrePrints | https://peerj.com/preprints/18 | v1 received: 10 May 2013, published: 10 May 2013, doi: 10.7287/peerj.preprints.18