Molecular Phylogenetics and Evolution 71 (2014) 184–192

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The impact of automated filtering of BLAST-determined homologs in the phylogenetic detection of horizontal gene transfer from a transcriptome assembly

Jennifer H. Wisecaver *[,1], Jeremiah D. Hackett

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85719, USA

ARTICLE INFO

Article history:
Received 22 January 2013
Revised 9 October 2013
Accepted 25 November 2013
Available online 7 December 2013

Keywords:
Phylogenomics
Taxon sampling
Homolog selection
Horizontal gene transfer
Gene trees
Dinoflagellates

ABSTRACT

Phylomes (comprehensive sets of gene phylogenies for organisms) are built to investigate fundamental questions in genomics and evolutionary biology, such as those pertaining to the detection and characterization of horizontal gene transfer in microbes. To address these questions, phylome construction demands rigorous yet efficient phylogenetic methods. Currently, many sequence alignment and tree-building models can analyze several thousands of genes in a high-throughput manner. However, the phylogenetics is complicated by variability in sequence divergence and different taxon sampling among genes. In addition, homolog selection for automated approaches often relies on arbitrary sequence similarity thresholds that are likely inappropriate for all genes in a genome. To investigate the effects of automated homolog selection on the detection of horizontal gene transfer using phylogenomics, we constructed the phylome of a transcriptome assembly of *Alexandrium tamarense*, a microbial eukaryote with a history of horizontal and endosymbiotic gene transfer, using seven sequence similarity thresholds for selecting putative homologs to be included in phylogenetic analyses. We show that no single threshold recovered informative trees for the majority of *A. tamarense* unigenes compared to the pooled results from all pipeline iterations. As much as 29% of trees built could have misleading phylogenetic relationships that appear biased in favor of those otherwise indicative of horizontal gene transfer. Perhaps worse, nearly half of the unigenes were represented by a single tree built at just one threshold, making it difficult to assess the validity of phylogenetic relationships recovered in these cases. However, combining the results from several pipeline iterations maximizes the number of informative phylogenies. Moreover, when the same phylogenetic relationship for a given unigene is recovered in multiple pipeline iterations, conclusions regarding gene origin are more robust to methodological artifact. Using these methods, the majority of *A. tamarense* unigenes showed evolutionary relationships indicative of vertical inheritance. Nevertheless, many other unigenes revealed diverse phylogenetic associations, suggestive of possible gene transfer. This analysis suggests that caution should be used when interpreting the results from phylogenetic pipelines implementing a single similarity threshold. Our approach is a practical method to mitigate the problems associated with automated sequence selection in phylogenomics.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

The phylome is the complete set of phylogenies for every gene in an organism (Sicheritz-Pontén and Andersson, 2001). As practical tools for biologists, phylomes can be used to predict the function of uncharacterized proteins (Eisen, 1998), distinguish orthologs from paralogs (Gabaldón, 2008), and predict species trees from gene trees (Delsuc et al., 2005). Phylomes also offer insight into large-scale processes in evolution including patterns of gene duplication (Huerta-Cepas et al., 2010) and horizontal gene transfer (Peña et al., 2010). Given the advantage of incorporating an evolutionary perspective into genomic analyses (i.e., phylogenomics), the automation of phylogenetic tree building for the purpose of reconstructing these phylomes is often the first hurdle faced by many investigators.

For those who study the evolution of microbial eukaryotes, phylogenomics offers a tremendous opportunity as well as a significant challenge. Increasingly, it is clear that many microbial

Abbreviations: HGT, horizontal gene transfer; EGT, endosymbiotic gene transfer; FC, fraction conserved; GO, gene ontology; SAR, stramenopiles, alveolates, and rhizarians.

* Corresponding author.
E-mail address: jen.wisecaver@vanderbilt.edu (J.H. Wisecaver).
[1] Current address: Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA.

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.11.016

eukaryotes have mosaic genomes, with a significant number of genes non-vertically acquired (Keeling and Palmer, 2008). Gene influx in these organisms is typically divided into two main categories: the more general horizontal gene transfer (HGT), and the special case of endosymbiotic gene transfer (EGT), which includes genes transferred to host nuclear genomes during organelle endosymbiosis. Many eukaryotic lineages are impacted by both of these processes, having a complicated history of heterotrophy, mixotrophy and autotrophy (Archibald et al., 2003; Nosenko et al., 2006; Minge et al., 2010; Wisecaver and Hackett, 2010; Maruyama et al., 2011). Untangling such complex evolutionary history benefits greatly from automated phylome construction, which has been used to quantify and qualify genes acquired from these different sources and thus has shed light on broader processes in eukaryote evolution (Stiller, 2011).

Three practical challenges for automated phylogenomics in the study of microbial eukaryotes are that (1) these analyses necessarily include highly divergent organisms from across the tree of life, (2) many branches of the eukaryotic tree are unresolved, and (3) most eukaryotic lineages are still poorly sampled, with many groups represented by none or only a handful of genomes. Gene sequences that are available for non-model microbial eukaryotes often contain errors due to poor gene models, partial cDNA sequences or contamination, creating complex taxon sampling issues susceptible to long-branch attraction and phylogenetic noise (Hartmann and Vision, 2008). As a result, eukaryotic tree of life studies invest heavily on manual curation of sequences (Rodriguez-Ezpeleta et al., 2007; Parfrey et al., 2010). Unfortunately, the scale of phylogenomic analysis necessitates the use of methods that emphasize speed often at the expense of rigour. Nonetheless, approximate methods do not always result in decreased accuracy. Although branch-length estimates are reliant on best-fit phylogenetic models, tree topology is relatively robust to model choice (Kelchner and Thomas, 2007). Additionally, new alignment and tree-building methods employ faster algorithms with little degradation in accuracy (e.g., Liu et al., 2011). Perhaps the weakest step in the process currently, is the selection of homologs included in the individual gene trees.

This paper evaluates a common methodology for selecting appropriate sequences for automatic tree building and its impact on detection of HGT/EGT in microbial eukaryotes. Generally, all genes from an organism are queried against a database of representative taxa using local alignment methods. Putative homologs are then extracted based on some set of thresholds for assessing the match significance (recent examples include Peña et al., 2010; Nowack et al., 2011; Maruyama et al., 2011; Price et al., 2012; Curtis et al., 2012). Often, the top 50-1000 BLAST hits are selected and filtered using a combination of *E*-value, sequence identity, and coverage cutoffs. The total number of sequences included from a particular species or lineage can be restricted in an attempt to minimize the number of paralogs in the analysis (e.g., Chan et al., 2011). When working with closely related and well-sampled groups, such as model plants or animals, precompiled clustered orthologs are available for tree building (O'Brien et al., 2005; Penel et al., 2009). However, clustering methods require complete gene sets from a limited number of genomes (Remm et al., 2001; Li et al., 2003). These predefined ortholog clusters can be augmented with additional sequences from further taxa, but this approach requires that the phylogeny of core taxa be known (Ebersberger et al., 2009). Ambiguity in the backbone of the eukaryotic tree and poor representation of many lineages makes query-based similarity searches the most sensitive approach for investigating divergent microbial eukaryotes (Chen et al., 2007).

Whether any method of homolog selection listed in the previous paragraph is appropriate for accurate and comprehensive phylome reconstruction has not been rigorously investigated. No standard guidelines exist to automate the taxon sampling process even after two decades of debate (Nabhan and Sarkar, 2012), despite widespread knowledge that sufficient taxon sampling is one of the most important factors influencing phylogenetic accuracy (Heath et al., 2008). In many cases, taxon sampling is critical not only for phylogenetic inclusiveness, but also because including more taxa decreases the average length of terminal branches, which can ameliorate the effects of long branch attraction (Brinkmann et al., 2005). Importantly, the best BLAST hits are often not equal to the nearest neighbor once a phylogenetic model is applied, further illustrating the importance of parsing a sufficient number of homologs from the BLAST report in automated pipelines (Koski and Golding, 2001). However, limiting the number of superfluous sequences is equally important, because increased sequence divergence can decrease accuracy by introducing additional long branches and increasing the number of poorly aligned columns in an alignment. A large number of sequences in phylogenetic analyses also impose practical limitations by adding computational complexity and running time. In particular, the inclusion of unnecessary out-paralogs (i.e., homologous genes originating through gene duplication in the common ancestor of multiple species) not only complicates tree building but also confounds downstream analysis of monophyly. The lack of standard guidelines is sensible, to some degree, because genes evolve at different rates, both in terms of sequence divergence and rate of duplication (Pal et al., 2006). Accordingly, no one set of significance thresholds is sufficient for all genes. There may exist some sampling optimum for each gene, where depth of taxon coverage is maximized while minimizing the number of more divergent sequences and thereby achieve the most accurate phylogenetic tree. What remains to be seen is whether this optimum is similar for most genes in a genome and if a single set of significance thresholds can be used to extract appropriate homologs from local alignments for the majority of gene trees in a phylome.

We created an automated phylogenomic pipeline to reconstruct the phylome of the dinoflagellate *Alexandrium tamarense*. The dinoflagellates are an ideal group for investigating the impact of taxon sampling in automated phylogenetic pipelines. A substantial amount of HGT and EGT has occurred in this lineage, with genes from bacterial and algal sources prevalent in the genome (Hackett et al., 2005, 2013; Janouskovec et al., 2010; Chan et al., 2012). However, dinoflagellate phylogenetic placement among related eukaryotes is consistently resolved, well-supported, and stable in phylogenetic analyses (Parfrey et al., 2010). Predicted proteins from the recently sequenced transcriptome of *A. tamarense* were run through a custom phylogenomic pipeline seven times while altering the threshold for what was considered an acceptable match for extracting putative homologs for tree building. This study tests the sensitivity of phylome reconstruction to automated phylogenetic methods. Our results suggest that no single threshold recovers the majority of trees in the phylome. However, pooling the results from the different pipeline iterations creates a set of supported trees with phylogenetic associations that can be evaluated for evolutionary signals of interest such as horizontal gene transfer.

2. Materials and methods

Details on cell culturing and Illumina sequencing has been published elsewhere (Hackett et al., 2013). Briefly, RNA-seq data were produced for *Alexandrium tamarense* CCMP1598 using cultures grown under nutrient replete as well as nitrogen and phosphorus limiting conditions. Fragments from polyadenylated RNAs were isolated and prepared for paired-end sequencing using an mRNA-seq reagent kit from Illumina (RS-100-0801). Sequencing was

completed on an Illumina Genome Analyzer 2.0 by the Biomicro Center at MIT (Cambridge, MA). RNA-seq data from *A. tamarense* (SRA052316) were quality trimmed with the trim read module in the CLC genomics workbench (www.clcbio.com) using a quality score limit of 0.05 and removing all ambiguous nucleotides. Trimmed reads were assembled in Velvet (version 1.1.02) using the Oases extension (version 0.1.20) with tracking of short read positions enabled (Zerbino and Birney, 2008). A range of hash lengths (23, 27, 31, 35, 39, 43, 51) was used to create a final merged assembly (Schulz et al., 2012). The *A. tamarense* assembly has been deposited in DDBJ/EMBL/GenBank under the accession GAIQ01000000.

A. tamarense contigs were queried against local protein and nucleotide databases by BLAST (see 2.1. Databases). Protein sequences were predicted from the nucleotide data using top BLAST hit information and FrameDP to correct for frameshift assembly errors (Gouzy et al., 2009). Generic Gene Ontology (GO) slim terms were assigned to *A. tamarense* translated sequences in Blast2GO using default cutoff values (Conesa et al., 2005). GO slim enrichment tests were done in Blast2GO using two-tailed Fisher's Exact Test with a false discovery rate of 0.05.

2.1. Databases

Two local databases were constructed for the purpose of this paper. The protein database included NCBI's Reference Sequence (release 42) and predicted protein sequences from recently sequenced microbial eukaryotes (JGI genome portal and Ghent University's online genome annotation server BOGAS). Sequences from additional algal and protist species and strains not present in the protein database were included in a nucleotide database comprised of expressed sequence tags and next-generation transcriptome assemblies from NCBI's dbEST and TSA. Both the protein and nucleotide databases were further subdivided based on major taxonomic groups (for list of groups see the Supplementary Table S1), and each taxonomic group was individually queried using BLAST.

2.2. Phylogenetic pipeline

The phylome of *A. tamarense* was constructed using a custom phylogenetics pipeline (Fig. 1A); scripts are available from the authors upon request. Predicted amino acid sequences were first queried using BLASTP and TBLASTN against the local databases. For each BLAST result, a hit was considered significant if the *E*-value was less than $1e^{-3}$ and the bit score was greater than 60. To investigate the impact of taxon sampling on phylome reconstruction, the BLAST reports were parsed seven times using a range of fraction conserved (FC) thresholds (0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). If a hit passed the *E*-value, bitscore, and FC thresholds, the associated sequence was extracted from the database using a custom perl script. For matches to the nucleotide database, only the translations of the high-scoring segment pairs were included. To reduce the number of paralogs in the analysis, only the top hit per species was extracted.

Extracted sequences were reordered based on global similarity to the query sequence with MAFFT using the minimum linkage clustering method and rough distance measure (number of shared 6mers) (Katoh et al., 2005). After reordering, the files were reduced to include only the top 1000 sequences, and files with less than 4 sequences were eliminated. Alignments were performed with MAFFT using the auto strategy selection and the BLOSUM scoring matrix closest to the FC threshold (i.e., a lower BLOSUM matrix was used when average sequence conservation was low, and a higher BLOSUM matrix was used when average sequence conservation was high). Poorly aligned positions and sequences were removed from the alignment using REAP (Hartmann and Vision, 2008), and trimmed alignments were further refined by a second MAFFT alignment using the same parameters as above. Phylogenetic trees were inferred using FastTree assuming a JTT + CAT amino acid model of substitution and 1000 resamples (Price et al., 2009, 2010).

2.3. Identifying nearest neighbors

Trees were filtered using a perl script that eliminated trees in which only dinoflagellates were present or dinoflagellates did not form a monophyletic group. Trees that contained only one dinoflagellate species were also removed because monophyly could not be assessed and to mitigate any potential signal from contamination. For each tree, dinoflagellate nearest neighbors were identified using Phylosort, a tool for sorting phylogenetic trees by searching for a user specified grouping of interest (Moustafa and Bhattacharya, 2008). For this analysis, a nearest neighbor association is defined as a sister relationship between a clade of dinoflagellates and another group of organisms with FastTree local support of 0.75 or greater (see Fig. 1D for a list of groups analyzed). Although we required dinoflagellate monophyly, other members of the neighbor group could be present elsewhere in the tree. This approach identified the most closely related sequences to our dinoflagellate clade while allowing for HGT and paralogous sequences in other lineages. To determine all possible nearest neighbors to dinoflagellates, we iterated through phylosort using all the lineages shown in Fig. 1D and identified all trees in which each lineage formed a neighbor association with dinoflagellates. For each iteration, the trees were rerooted with an outgroup that was automatically selected from taxa outside the relationship of interest.

3. Results and discussion

The *A. tamarense* transcriptome data assembled into 142,638 contigs. After clustering the full assembly into 101,118 putative unigenes, the longest contig from each unigene set was translated and run through our phylogenetic pipeline (Fig. 1). The haploid genome of *A. tamarense* is estimated to be approximately 100 Gbp based on flow cytometry experiments (LaJeunesse et al., 2005), but the number of genes in this organism is still unknown. A regression of gene content and genome size in sequenced genomes predicted 87,688 protein-coding genes in a dinoflagellate with a genome of comparable size (Hou and Lin, 2009), which suggests that the amount of unigenes in our assembly is not unreasonable for a dinoflagellate transcriptome.

To investigate the impact of BLAST significance thresholds on taxon sampling in phylome construction, the phylogenetic pipeline was repeated seven times, each time using a different fraction of conserved amino acids threshold for identifying putatively homologous sequences. The fraction conserved (FC) score accounts for amino acid substitutions that occur frequently (given a substitution matrix) and is a more relaxed metric of sequence similarity than percent identity. In total, 40,342 contigs were represented by a phylogenetic tree at one or more FC thresholds (Table 1). The pipeline iteration using the lowest FC threshold (0.30) was the most inclusive and had the largest number of trees (40,192), whereas using the highest FC threshold (0.90) resulted in the smallest number of trees built (10,787).

Phylogenetic trees in which dinoflagellates did not form a monophyletic group, as well as trees for which dinoflagellate monophyly could not be assessed, were excluded from the analysis. This conservative yet robust approach greatly reduces the risk of culture or sequencing contamination in the analysis because the genes must be present in multiple dinoflagellate species. This

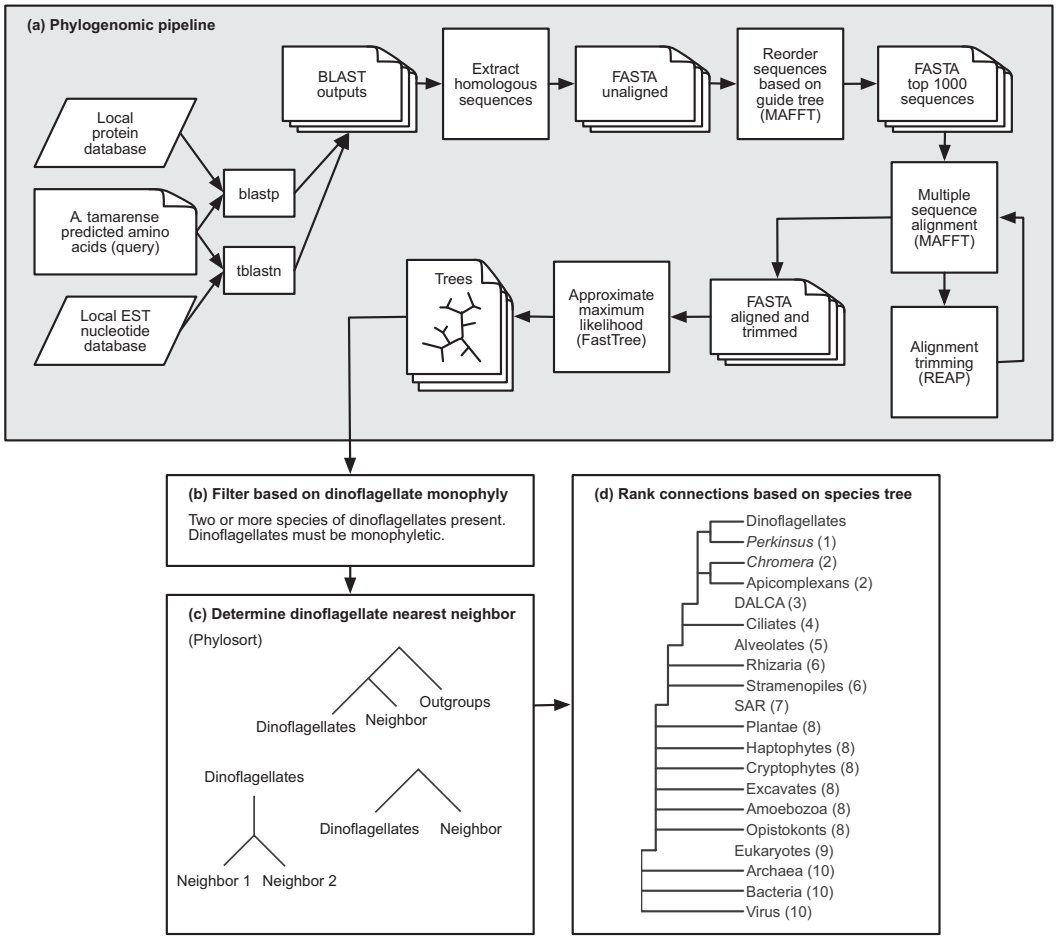

Fig. 1. Flowchart depicting the order of operations for the phylogenetic pipeline (a) and subsequent tree parsing (b–d). Only nearest neighbors to dinoflagellates with local support of 0.75 or higher were included. Because trees were unrooted, two supported neighbor associations were possible for a single tree (c). Groups were ranked (1–10) relative to their phylogenetic distance from dinoflagellates based on the species tree (d) with groups closely related to dinoflagellates given a lower number. Internal branches were drawn off center to label internal nodes in line with the leaves. DALCA = Dinoflagellate-Apicomplexan Last Common Ancestor. SAR = Stramenopile, Alveolate, Rhizaria last common ancestor.

Table 1
Results of the phylogenetic pipeline for constructing trees for transcriptome contigs from the dinoflagellate *Alexandrium tamarense*.

FC threshold	Trees built	Passed monophyly filter (as % of trees built)	Phylogenetically informative (as % of filtered trees)	Singleton trees (as % of phylogenetically informative trees)
0.30	40,192	10,789 (27%)	9749 (90%)	526 (5%)
0.40	38,697	11,063 (29%)	9980 (90%)	418 (4%)
0.50	38,176	12,717 (33%)	11,660 (92%)	2132 (18%)
0.60	36,050	13,185 (37%)	12,491 (95%)	4459 (36%)
0.70	30,475	8152 (27%)	7896 (97%)	3065 (39%)
0.80	21,494	2692 (13%)	2652 (99%)	936 (35%)
0.90	10,787	686 (6%)	682 (99%)	304 (45%)
Total[a]	40,342	27,458 (68%)	26,787 (98%)	11,840 (44%)

[a] Number of unique contigs with one or more trees built across all thresholds.

filtering step also removes trees containing dinoflagellate out-paralogs. One limitation of requiring dinoflagellate monophyly is that this analysis will not recover lineage specific HGT. We consider this tradeoff acceptable because of the difficulty in interpreting more recent HGT due to limited sequence data from dinoflagellates and potential gene donors. Depending on the FC threshold, as little as six percent of trees passed this filtering stage (Table 1). Using a lower FC threshold resulted in a larger proportion of trees failing to recover dinoflagellate monophyly, presumably due to the addition of dinoflagellate out-paralogs (Fig. 2). In contrast, at higher FC thresholds a large proportion of trees contained only dinoflagellate taxa and were thus uninformative for this analysis (Fig. 2). The 0.60 FC threshold appeared optimal with the largest proportion of trees passing the filter (37%). Of the trees that passed the filtering stage, the vast majority (90–99% depending on threshold) were considered informative. For this study, informative trees had a nearest neighbor clade to dinoflagellates that contained sequences from a single taxonomic group (as defined in Fig. 1) and FastTree local support of 0.75 or greater. Because the trees were unrooted, it was possible for dinoflagellates to have two nearest neighbors in

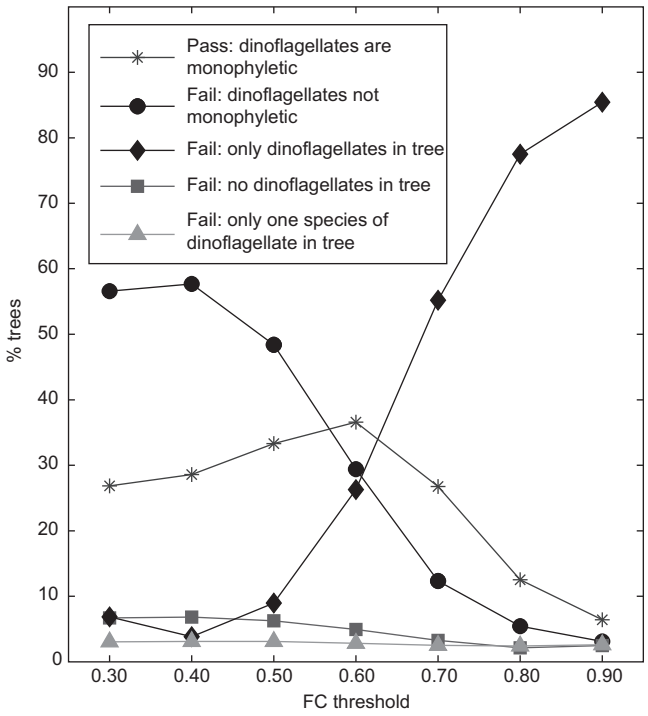

Fig. 2. Effect of FC threshold on dinoflagellate monophyly. For each threshold, the proportion of trees in which dinoflagellates formed a monophyletic group is plotted with an asterisk. All other trees failed the filtering step and were rejected. This included trees in which dinoflagellates were not monophyletic (circles), trees that only contained dinoflagellates (diamonds), trees that contained no dinoflagellates after alignment trimming (squares), and trees that contained only one dinoflagellate species (triangles).

a single tree depending on the location of the root (Fig. 1C). When possible, the neighbors were ranked using the species tree (Fig. 1D), with the neighbor sharing the more recent common ancestor with dinoflagellates being preferred. For example, if a tree showed phylogenetic associations between dinoflagellates and both apicomplexans and haptophytes, the haptophyte association was discarded, because haptophytes are the more plausible outgroup. This approach is intended to provide a conservative estimate of HGT by assuming that the majority of genes are vertically inherited and thus expected to agree with the species tree. However, if the two neighbors shared the same most recent common ancestor with dinoflagellates (e.g., haptophytes and excavates), both were included in downstream analyses.

3.1. Effects of FC threshold choice on phylome interpretation

3.1.1. Distribution of gene trees

In total, 26,788 contigs were represented by one or more gene trees in the analysis, but use of any one FC threshold failed to produce informative trees for the majority of these contigs. Use of a FC threshold of 0.60 produced the largest number of trees with at least one discernable nearest neighbor to dinoflagellates (12,491 or 47% of total contigs with trees). However, a stepwise comparison of different FC thresholds showed that pipeline iterations often yielded conflicting nearest neighbor associations to dinoflagellates (Table 2). The degree of conflict served to illustrate the importance of sequence similiarity thesholds in phylogenomic pipelines and raised the practical challenge of selecting the most appropriate nearest neighbor when different FC thresholds yielded different dinoflagellate nearest neighbors.

Two different methods were used to reconcile nearest neighbors when contigs were represented by multiple trees built using

Table 2
Stepwise comparison of FC thresholds.

FC threshold comparison	Total overlapping[a]	Disagree (%)[b]
0.3–0.4	8160	1970 (24%)
0.4–0.5	6897	2673 (39%)
0.5–0.6	5535	2889 (52%)
0.6–0.7	3681	1978 (54%)
0.7–0.8	1343	689 (51%)
0.8–0.9	306	177 (58%)

[a] Number of contigs with phylogenetically informative trees built at both thresholds.
[b] Number of contigs in which the two FC thresholds produced different dinoflagellate nearest-neighbors.

different FC thresholds. For a given contig, the consensus-based reconciliation consisted of the neighbor(s) present in the largest number of trees. In the case of a tie, all nearest neighbors supported by the largest number of trees were included in the final reconciliation. The majority of trees (85–93% depending on the FC threshold) did not contradict this reconciliation (Fig. 3). The second approach, termed the tree-based reconciliation again utilized the species tree from Fig. 1D to rank nearest neighbors and selected the most plausible given the species tree. The majority of trees (71–81% depending on the FC threshold) also agreed with this second reconciliation (Fig. 3). Fig. 3B illustrates the overlap between trees built and their agreement with either the tree-based or consensus-based reconciliation. See the Supplementary Dataset S1 for overlap results comparing all seven FC thresholds analyzed.

3.1.2. Distribution of dinoflagellate nearest neighbors

The ability of the pipeline to recover a tree for any single contig was heavily dependent on FC threshold (Fig. 4). The majority of contigs had trees recovered using two or more FC thresholds, however, a significant fraction (44%) were represented by a single tree built using only one of the seven thresholds (Table 1). In addition, the distribution of trees was not nested as FC threshold decreased. These results raised the concern that nearest neighbors could be differentially represented in pipeline iterations, and thereby affect the interpretation of the phylome. We compared neighbors present in the phylome at different FC thresholds and found that the proportions of neighbors were stable, regardless of FC threshold (Supplementary Fig. S1). In addition, we compared the distribution of Generic Gene Ontology (GO) slim terms assigned to contigs with trees built using different FC thresholds and found no significant differences between thresholds (Supplementary Fig. S2). However, highly conserved genes (i.e. Ribosome (GO: 0005840), nucleotide

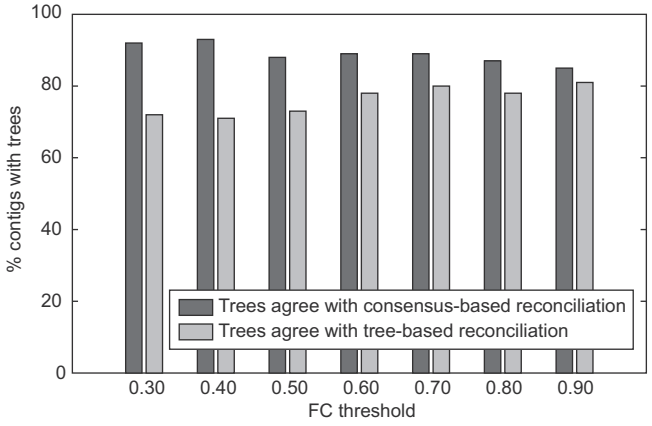

Fig. 3. Tree reconciliation. Bar graph summarizing the percentage of contigs with trees that agreed with either the tree-based or consensus-based reconciliation.

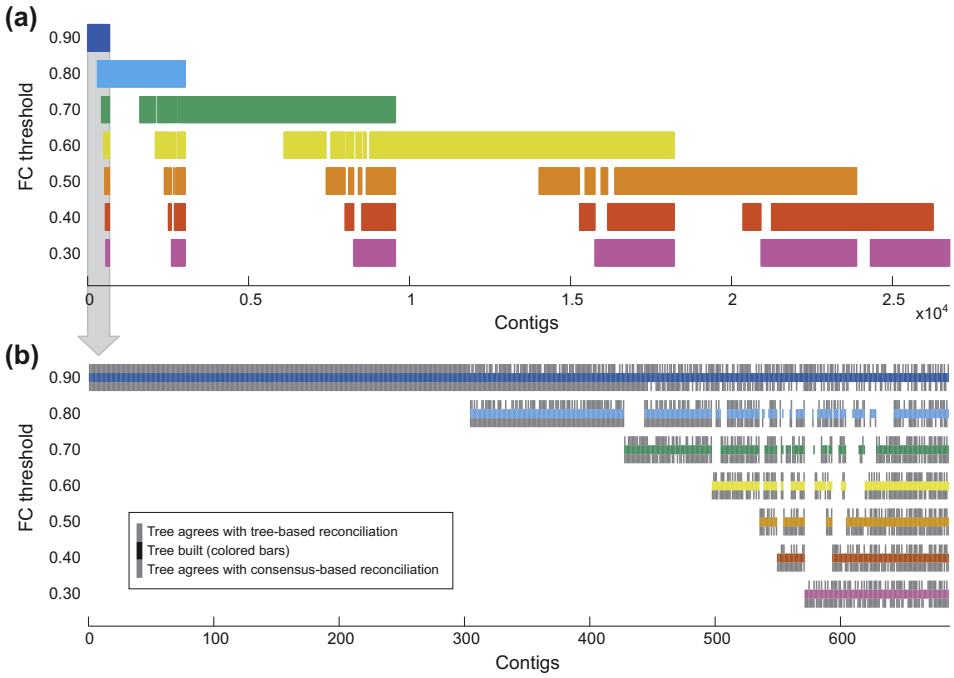

Fig. 4. Distribution of gene trees built across FC thresholds. After filtering, 27,458 *A. tamarense* contigs were represented by one or more trees. (a) Contigs are ordered horizontally and colored bars indicate a phylogenetically informative tree was built at a particular threshold. (b) Expanded view of trees built at the 0.90 FC threshold (shaded gray in part a) to illustrate the overlap between trees built and their agreement with either the tree-based or consensus-based reconciliation.

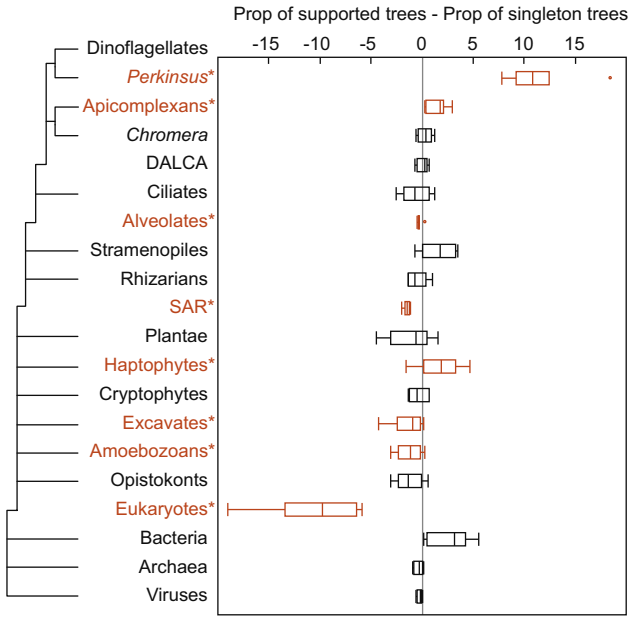

Fig. 5. Plot summarizing the differential representation of dinoflagellate nearest neighbors between supported and singleton trees. For each possible neighbor, we calculated the difference between its proportion in supported trees and its proportion in singleton trees. Box plots summarize the variation in the proportional difference across FC thresholds. Neighbors with mean proportions statistically different between supported and singleton trees are colored red and marked with an asterisk (*). Species tree from Fig. 1 is included for orientation.

binding (GO: 0000166), and translation (GO: 0006412)) made up a larger proportion of the higher FC threshold trees (0.90 and 0.80).

To determine how potentially spurious nearest neighbors affect the interpretation of phylomes, we compared the distribution of dinoflagellate nearest neighbors found in two separate tree subsets. The first subset consisted of contigs represented by one tree recovered by a single FC threshold, hereafter referred to as singleton trees. The neighbors found in singleton trees were, by definition, unable to be substantiated by agreement with trees built in other pipeline iterations. In contrast, the second subset was comprised of contigs represented by trees built at multiple FC thresholds whose neighbors agreed with both the consensus-based and tree-based reconciliations, hereafter referred to as supported trees. Five dinoflagellate neighbors (alveolates, SAR, amoebozoans, excavates, and eukaryotes) were overrepresented in singleton trees (two-tailed t test; $p = 0.013$, <0.0001, 0.044, 0.018, and 0.002 respectively), whereas dinoflagellate nearest neighbor associations to *Perkinsus*, apicomplexans, and haptophytes were overrepresented in supported trees (two-tailed t test; $p \leqslant 0.0001$, 0.003, and 0.043 respectively). The difference in proportion between supported trees and singleton trees, plotted for every possible nearest neighbor to dinoflagellates, further illustrates how these associations are differentially represented in the two subsets (Fig. 5). The neighbors *Perkinsus*, apicomplexans, stramenopiles, haptophytes, and bacteria made up a larger fraction of total neighbors in the supported trees compared to singleton trees. These results agree with what is already known regarding dinoflagellate evolutionary relationships. *Perkinsus*, apicomplexans, and stramenopiles are all related to dinoflagellates, and these associations are likely indicative of vertical gene inheritance. In addition, stramenopiles and haptophytes are sources of plastid related genes in dinoflagellates (Yoon et al., 2002; Nosenko et al., 2006; Wisecaver and Hackett, 2010; Minge et al., 2010). Lastly, bacteria-to-dinoflagellate HGT has also been described (Morse et al., 1995; Hackett et al., 2005).

In contrast, the associations overrepresented in singleton trees, particularly dinoflagellates being nearest neighbor to amoebozoans, excavates, and opisthokonts (e.g., animals and fungi), are not supported by previous, single-gene phylogenetics for dinoflagellates. Because of the large evolutionary distance between these groups and dinoflagellates, these associations could be interpreted as evidence for HGT. However, these lineages are also minimally

overlapping with dinoflagellates in terms of ecology, which makes HGT extremely doubtful. Given the lack of support for these associations across pipeline iterations, it is more likely that they are due to other sources of atypical phylogenetic placement, which include taxon sampling (e.g., Rokas et al., 2003), long branch attraction (e.g., Brinkmann et al., 2005), incomplete lineage sorting (e.g., Ebersberger et al., 2007), and differential gene loss (e.g., Qiu et al., 2012).

3.2. Pipeline implementation for analysis of recent evidence of HGT in dinoflagellates

HGT is a commonly described mechanism of gene innovation in prokaryotes, but the importance of HGT in the evolution of microbial eukaryotes is only now being recognized thanks to the increase in genome and transcriptome data for these organisms (Keeling and Palmer, 2008). The genomes of microbial eukaryotes, such as dinoflagellates, are chimeras of vertically retained genes as well as genes acquired through endosymbiosis and HGT. However, despite their mosaic genomes, the placement of dinoflagellates among related members of the SAR supergroup (e.g., apicomplexans, ciliates, rhizarians and stramenopiles) is consistently resolved and well supported in phylogenetic analyses (Reece et al., 1997; Fast et al., 2002; Burki et al., 2007; Parfrey et al., 2010), providing a strong phylogenetic framework for testing non-vertical inheritance of genes in dinoflagellates. Gene transfer has already been shown to be an important process in the evolution of distinguishing features of dinoflagellates (Morse et al., 1995; Hackett et al., 2005; Nosenko et al., 2006; Wisecaver and Hackett, 2010; Minge et al., 2010). However, despite seemingly adequate phylogenetic resolution, determining the pattern of inheritance for individual genes from these mosaic genomes can be quite challenging. For example, genes trees showing a phylogenetic relationship between dinoflagellates and other algae (e.g., stramenopiles) could be interpreted as vertically acquired from a common ancestor and lost in other alveolates (i.e., apicomplexans and ciliates). However, EGT through plastid endosymbiosis or HGT from algal prey could result in the same phylogenetic association. For this study, we focused on trees in which dinoflagellates grouped with bacteria, which are easier to interpret as being horizontally acquired compared to trees that show dinoflagellates grouping with other eukaryotes. Large amounts of *A. tamarense* contigs show a phylogenetic affinity between dinoflagellates and bacteria when using our pipeline, a signal that is indicative of HGT. The fact that this relationship is overrepresented in supported trees versus singleton trees suggests that the association is robust to variation in sequence selection criteria. As much as 17% of trees built demonstrate this relationship, which is recovered regardless of the FC threshold used for tree building (Table 3).

Genes vary in their propensity to be horizontally transferred (Nakamura et al., 2004). The complexity hypothesis predicts that the transferability of a gene can be inferred from its biological function and connectivity (Jain et al., 1999). Specifically, information genes (e.g., involved in translation and transcription) as well as genes with high connectivity (i.e., a large number of protein–protein interactions) are generally less likely to be transferred. We tested the expectations of the complexity hypothesis by evaluating the distribution of GO slim terms assigned to *A. tamarense* contigs that our pipeline suggested were horizontally acquired from bacteria in dinoflagellates. As predicted, GO slims related to information processing were underrepresented in putatively transferred genes (e.g., Ribosome GO: 0005840, $p = 3.50E{-}12$; Translation GO: 0006412, $p = 1.02E{-}10$). In contrast, operational genes involved in catalytic activity (GO: 0003824) were statistically overrepresented ($p = 1.88E{-}7$). See the Supplementary Table S2 for the full list of under/over represented GO slim terms.

Table 3
Number of gene phylogenies containing a dinoflagellate-bacteria phylogenetic association.

FC threshold	Genes of putative bacterial origin	Proportion of phylogenetically informative trees (%)
0.3	1203	12
0.4	1267	13
0.5	1440	12
0.6	1409	11
0.7	853	11
0.8	253	10
0.9	27	4
Total[a]	4013	15
Supported[b]	1281	17

[a] Number of contigs with one or more trees built across all thresholds.
[b] Number of contigs with supported dinoflagellate nearest neighbors corroborated by multiple thresholds.

The results from our pipeline support a growing body of evidence that dinoflagellate genomes are heavily impacted by HGT. Rather than the plastid-encoded, two subunit RuBisCO present in most eukaryotes, dinoflagellates have a nuclear-encoded form acquired from Proteobacteria (Morse et al., 1995; Janouskovec et al., 2010). In addition, histone-like proteins, basic nuclear proteins hypothesized to play a role in the unique chromosome structure of dinoflagellates (Chan and Wong, 2007), have also been acquired from Proteobacteria (Hackett et al., 2005). A recent analysis of dinoflagellate genes showed as much as 435 genes (17% of genes analyzed) arising through HGT from a variety of sources (Chan et al., 2012). To mitigate the effects of automated tree construction, the aforementioned study used conservative sequence selection criteria, reducing the numbers of genes analyzed, but nevertheless recovering a significant amount HGT. In our study, the corroboration of dinoflagellate nearest neighbors across multiple thresholds permitted the analysis of many more sequences and found a larger total number but comparable proportion of possible HGT in dinoflagellates, particularly from bacteria. However, these analyses should not be interpreted as quantitative measures of HGT in dinoflagellates because of the difficulty of connecting contigs from *de novo* transcriptome assemblies back to discrete genomic loci. As a result, the present study does not address possible duplications and gene family expansions in dinoflagellates that might over represent the number of transfer events. Also unknown is whether the proportion of bacterial genes can be extrapolated to the full *A. tamarense* transcriptome including the contigs that did not produce informative trees. In addition, our pipeline required dinoflagellate monophyly, which is a conservative estimate of HGT in dinoflagellates for two reasons. First, genes that were recently acquired in only some dinoflagellates would have been excluded from our analysis (e.g., genes acquired during plastid replacement, see Minge et al., 2010; Wisecaver and Hackett, 2010), potentially underestimating HGT. Second, even genes horizontally acquired in the dinoflagellate ancestor could be missed by our screen if the transferred gene functionally replaced a preexisting gene, which could lead to differential loss of the two copies in different dinoflagellate species. Regardless of these limitations, the total of number of transfers predicted by our analysis and others suggests that dinoflagellate genomes are more amenable to the incorporation of foreign DNA than many other eukaryotic lineages (Andersson, 2005; Keeling and Palmer, 2008).

4. Conclusions

Many automated phylogenetic pipelines rely on extracting information from BLAST reports. However, it is well understood that the best BLAST hit does not always reveal the phylogenetic nearest neighbor (Koski and Golding, 2001). This study illustrates

the importance of homolog selection in automated phylogenetic pipelines used for detecting cases of HGT/EGT in microbial eukaryotes. The phylogenetic associations recovered are highly dependent on the significance thresholds used to extract putative homologs. This pattern is no doubt amplified by the nature of HGT detection pipelines that require the inclusion of highly divergent sequences from across the tree of life as well as EST and transcriptome sequences from poorly sampled groups (i.e., microbial eukaryotes outside Fungi) in order to accurately predict gene donors.

Our results show that no single FC threshold recovers trees for the majority of contigs compared to the pooled results from all seven FC thresholds. Nearly half of all contigs (44%) are represented by a tree built in just one pipeline iteration (singleton trees), making it difficult to assess the validity of the nearest neighbors present for these contigs. Comparing all neighbor associations to those supported by our two reconciliation methods suggests that as much as 29% of trees could have erroneous phylogenetic relationships. In addition, it is troubling that the relative proportions of several neighbors were different when comparing singleton trees to trees whose phylogenetic relationships were well supported across multiple thresholds. In the case of dinoflagellates, some of the neighbors overrepresented in singleton trees could be misinterpreted as evidence for HGT (e.g., excavates or ameobozoans), when the association is likely due to phylogenetic artifact. Our results suggest that caution should be exercised when using just one threshold for automated phylogenetic pipelines, particularly when the results of phylome construction are used for quantifying how many gene trees support different conclusions related to gene and genome evolution. If the results of our analysis are applicable to other phylomes, as much as 29% of trees built using standard query based methods could have misleading phylogenetic relationships that are biased in favor of those otherwise indicative of HGT.

One encouraging aspect of our analyses is that the proportion of spurious neighbor associations is consistent across pipeline iterations using different FC thresholds. In addition, different pipeline iterations do not appear overly biased in individual functional GO categories that would otherwise affect our interpretation of the phylome. By pooling the results of several FC thresholds, we built trees for many more contigs than was possible using any single threshold. When the same relationship was recovered across multiple pipeline iterations, conclusions regarding patterns of gene origin, including sources of HGT, were more strongly supported. Our approach is a potential method to mitigate this key problem associated with automated sequence selection in phylogenomic pipelines for the detection and quantification of HGT.

Acknowledgments

We are grateful to Betsy Arnold, Mike Sanderson, and Matt Sullivan for reviewing the manuscript and providing helpful feedback. We thank Susan Miller for her computational help. We thank Bonnie Hurwitz and Dan DeBlasio for their help with the bioinformatics. We thank Galen Holt, Will Driscoll, and Ellen Martinson for their help with the statistical analysis. JHW was supported by the NSF IGERT Program in Comparative Genomics at the University of Arizona (DGE-0654435). This work was supported by grants from the National Science Foundation (OCE-0723498 and EF-0732440) and funding provided by the BIO5 Institute at the University of Arizona to JDH.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2013.11.016.

References

Andersson, J.O., 2005. Lateral gene transfer in eukaryotes. Cell. Mol. Life Sci. 62, 1182–1197.
Archibald, J.M., Rogers, M., Toop, M., Ishida, K., Keeling, P.J., 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc. Natl. Acad. Sci. USA 100, 7678–7683.
Brinkmann, H., Van der Giezen, M., Zhou, Y., De Raucourt, G., Philippe, H., 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54, 743–757.
Burki, F., Shalchian-Tabrizi, K., Minge, M., 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One 2, e790.
Chan, Y.-H., Wong, J.T.Y., 2007. Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCc3. Nucleic Acids Res. 35, 2573–2583.
Chan, C.X., Yang, E.C., Banerjee, T., Yoon, H.S., Martone, P.T., Estevez, J.M., Bhattacharya, D., 2011. Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. Curr. Biol. 21, 328–333.
Chan, C.X., Soares, M.B., Bonaldo, M.F., 2012. Analysis of *Alexandrium tamarense* (Dinophyceae) genes reveals the complex evolutionary history of a microbial eukaryote. J. Phycol. 48, 1130–1142.
Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2, e383.
Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.
Curtis, B.A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., Hopkins, J.F., Kuo, A., Rensing, S.A., Schmutz, J., Symeonidi, A., Elias, M., Eveleigh, R.J.M., Herman, E.K., Klute, M.J., Nakayama, T., Oborník, M., Reyes-Prieto, A., Armbrust, E.V., Aves, S.J., Beiko, R.G., Coutinho, P., Dacks, J.B., Durnford, D.G., Fast, N.M., Green, B.R., Grisdale, C.J., Hempel, F., Henrissat, B., Höppner, M.P., Ishida, K.-I., Kim, E., Kořený, L., Kroth, P.G., Liu, Y., Malik, S.-B., Maier, U.-G., McRose, D., Mock, T., Neilson, J.A.D., Onodera, N.T., Poole, A.M., Pritham, E.J., Richards, T.A., Rocap, G., Roy, S.W., Sarai, C., Schaack, S., Shirato, S., Slamovits, C.H., Spencer, D.F., Suzuki, S., Worden, A.Z., Zauner, S., Barry, K., Bell, C., Bharti, A.K., Crow, J.A., Grimwood, J., Kramer, R., Lindquist, E., Lucas, S., Salamov, A., McFadden, G.I., Lane, C.E., Keeling, P.J., Gray, M.W., Grigoriev, I.V., Archibald, J.M., 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492, 59–65.
Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.
Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M., Haeseler von, A., 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24, 2266–2276.
Ebersberger, I., Strauss, S., Haeseler von, A., 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. 9, 157.
Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8, 163–167.
Fast, N., Xue, L., Bingham, S., Keeling, P.J., 2002. Re-examining alveolate evolution using multiple protein molecular phylogenies. J. Eukaryot. Microbiol. 49, 30–37.
Gabaldón, T., 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 9, 235.
Gouzy, J., Carrere, S., Schiex, T., 2009. FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 25, 670–671.
Hackett, J.D., Scheetz, T.E., Yoon, H.S., Soares, M.B., Bonaldo, M.F., Casavant, T.L., Bhattacharya, D., 2005. Insights into a dinoflagellate genome through expressed sequence tag analysis. BMC Genom. 6, 80.
Hackett, J.D., Wisecaver, J.H., Brosnahan, M.L., Kulis, D.M., Anderson, D.M., Bhattacharya, D., Plumley, F.G., Erdner, D.L., 2013. Evolution of saxitoxin synthesis in cyanobacteria and dinoflagellates. Mol. Biol. Evol. 30, 70–78.
Hartmann, S., Vision, T.J., 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol. Biol. 8, 95.
Heath, T.A., Hedtke, S.M., Hillis, D.M., 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46, 239–257.
Hou, Y., Lin, S., 2009. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS One 4, e6978.
Huerta-Cepas, J., Marcet-Houben, M., Pignatelli, M., Moya, A., Gabaldón, T., 2010. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. Insect Mol. Biol. 19, 13–21.
Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA 96, 3801–3806.
Janouskovec, J., Horak, A., Oborník, M., Lukes, J., Keeling, P.J., 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc. Natl. Acad. Sci. USA 107, 10949–10954.
Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.
Keeling, P.J., Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9, 605–618.
Kelchner, S.A., Thomas, M.A., 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22, 87–94.
Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542.
LaJeunesse, T., Lambert, G., Andersen, R., Coffroth, M., Galbraith, D., 2005. *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. J. Phycol. 41, 880–886.

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189.
Liu, K., Linder, C.R., Warnow, T., 2011. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. PLoS One 6, e27731.
Maruyama, S., Suzaki, T., Weber, A.P.M., Archibald, J.M., Nozaki, H., 2011. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. BMC Evol. Biol. 11, 105.
Minge, M.A., Shalchian-Tabrizi, K., Tørresen, O.K., Takishita, K., Probert, I., Inagaki, Y., Klaveness, D., Jakobsen, K.S., 2010. A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate *Lepidodinium chlorophorum*. BMC Evol. Biol. 10, 191.
Morse, D., Salois, P., Markovic, P., Hastings, J., 1995. A nuclear-encoded form II RuBisCO in dinoflagellates. Science 268, 1622–1624.
Moustafa, A., Bhattacharya, D., 2008. PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*. BMC Evol. Biol. 8, 6.
Nabhan, A.R., Sarkar, I.N., 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Brief. Bioinform. 13, 122–134.
Nakamura, Y., Itoh, T., Matsuda, H., Gojobori, T., 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat. Genet. 36, 760–766.
Nosenko, T., Lidie, K.B., Van Dolah, F.M., Lindquist, E., Cheng, J., Bhattacharya, D., 2006. Chimeric plastid proteome in the Florida "red tide" dinoflagellate *Karenia brevis*. Mol. Biol. Evol. 23, 2026–2038.
Nowack, E.C.M., Vogel, H., Groth, M., Grossman, A.R., Melkonian, M., Gloeckner, G., 2011. Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. Mol. Biol. Evol. 28, 407–422.
O'Brien, K.P., Remm, M., Sonnhammer, E.L.L., 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33, D476–80.
Pal, C., Papp, B., Lercher, M., 2006. An integrated view of protein evolution. Nat. Rev. Genet. 7, 337–348.
Parfrey, L.W., Grant, J., Tekle, Y.I., Lasek-Nesselquist, E., Morrison, H.G., Sogin, M.L., Patterson, D.J., Katz, L.A., 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst. Biol. 59, 518–533.
Peña, A., Teeling, H., Huerta-Cepas, J., Santos, F., Yarza, P., Brito-Echeverría, J., Lucio, M., Schmitt-Kopplin, P., Meseguer, I., Schenowitz, C., Dossat, C., Barbe, V., Dopazo, J., Rosselló-Mora, R., Schüeler, M., Glöckner, F.O., Amann, R., Gabalón, T., Antón, J., 2010. Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting *Salinibacter ruber* strains. ISME J. 4, 882–895.
Penel, S., Arigon, A.M., Dufayard, J.F., Sertier, A.S., Daubin, V., Duret, L., Gouy, M., Perrière, G., 2009. Databases of homologous gene families for comparative genomics. BMC Bioinform. 10 (Suppl. 6), S3.
Price, M.N., Dehal, P.S., Arkin, A.P., 2009. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26, 1641–1650.
Price, M.N., Dehal, P.S., Arkin, A.P., 2010. Fasttree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.
Price, D.C., Chan, C.X., Yoon, H.S., Yang, E.C., Qiu, H., Weber, A.P.M., Schwacke, R., Gross, J., Blouin, N.A., Lane, C., Reyes-Prieto, A., Durnford, D.G., Neilson, J.A.D., Lang, B.F., Burger, G., Steiner, J.M., Löffelhardt, W., Meuser, J.E., Posewitz, M.C., Ball, S., Arias, M.C., Henrissat, B., Coutinho, P.M., Rensing, S.A., Symeonidi, A., Doddapaneni, H., Green, B.R., Rajah, V.D., Boore, J., Bhattacharya, D., 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science 335, 843–847.
Qiu, H., Yang, E.C., Bhattacharya, D., Yoon, H.S., 2012. Ancient gene paralogy may mislead inference of plastid phylogeny. Mol. Biol. Evol. 29, 3333–3343.
Reece, K., Siddall, M., Burreson, E., Graves, J., 1997. Phylogenetic analysis of *Perkinsus* based on actin gene sequences. J. Parasitol. 83, 417–423.
Remm, M., Storm, C.E., Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314, 1041–1052.
Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399.
Rokas, A., King, N., Finnerty, J., Carroll, S.B., 2003. Conflicting phylogenetic signals at the base of the metazoan tree. Evol. Dev. 5, 346–359.
Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092.
Sicheritz-Pontén, T., Andersson, S.G.E., 2001. A phylogenomic approach to microbial evolution. Nucleic Acids Res. 29, 545–552.
Stiller, J.W., 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. BMC Evol. Biol. 11, 259.
Wisecaver, J.H., Hackett, J.D., 2010. Transcriptome analysis reveals nuclear-encoded proteins for the maintenance of temporary plastids in the dinoflagellate *Dinophysis acuminata*. BMC Genom. 11, 366.
Yoon, H.S., Hackett, J.D., Bhattacharya, D., 2002. A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. Proc. Natl. Acad. Sci. USA 99, 11724–11729.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.