Taxon-Rich Phylogenomic Analyses Resolve the Eukaryotic Tree of Life and Reveal the Power of Subsampling by Sites

LAURA A. KATZ1,∗ AND JESSICA R. GRANT1

1Department of Biological Sciences, Smith College, Northampton, MA 01063, USA and 2Program in Organismic and Evolutionary Biology, UMass-Amherst, Amherst MA 01003, USA

∗Correspondence to be sent to: Department of Biological Sciences, Smith College, Northampton, MA 01063, USA; E-mail: lkatz@smith.edu

Received 9 March 2014; reviews returned 17 July 2014; accepted 15 December 2014

Associate Editor: Thomas Buckley

Abstract.—Most eukaryotic lineages are microbial, and many have only recently been sampled for phylogenetic studies or remain in the “dark area” of the tree of life where there are no molecular data. To assess relationships among eukaryotic lineages, we perform a taxon-rich phylogenomic analysis including 322 eukaryotes selected to maximize taxonomic diversity and up to 1554 genes chosen as vertically inherited based on their broad distribution among eukaryotes. We also include sequences from 486 bacteria and 84 archaea to assess the impact of endosymbiotic gene transfer (EGT) from plastids and to detect contamination. Overall, our analyses are consistent with other less taxon-rich estimates of the eukaryotic tree of life, and we recover strong support for five major clades: Amoebozoa, Excavata (without the genus Malachimonas), Opisthokonta, Archaeplastida, and SAR (Stramenopila, Alveolata, and Rhizaria). Our analyses also highlight the existence of “orphan” lineages, lineages that lack robust placement in the eukaryotic tree of life, and indicate the possibility of as yet undiscovered diversity. In analyses including bacteria and archaea, we find that approximately 10% of the 1554 genes, which we choose because they are found in four or five of the five major eukaryotic clades and hence may be more likely to be inherited vertically, appear to have been acquired from cyanobacteria through EGT in photosynthetic lineages. Removing these EGT genes places the green algae as sister to the glaucophytes instead of the red algae, suggesting that unknowingly including genes of plastid origin, and combining them with genes of nuclear origin, may mislead phylogenetic estimates. Finally, the large size of our data set allows comparative analyses of subsets of data; alignments built from randomly sampled sites provide greater support, particularly for deep relationships, than do equivalent-sized data sets built from randomly sampled genes. [Endosymbiotic gene transfer; eukaryotic phylogeny; microbial diversity; phylogenetics; sampling strategy; tree of life.]

The power of phylogenomics for resolving relationships is well-documented for groups shallower than eukaryotes (Delsuc et al. 2005; Philippe et al. 2005; Giribet and Edgecombe 2012; Kumar et al. 2012), and has been applied previously to questions of eukaryotic phylogenomics (Hampl et al. 2009; Burki et al. 2010; Burki et al. 2012; Torruella et al. 2012; Sierra et al. 2013) and eukaryotic origins (Cotton and McInerney 2010; Thiergart et al. 2012; Williams et al. 2012). However, given the constraints on large-scale phylogenomic analyses and data availability, the bulk of these studies focused on relatively small numbers of lineages, often in the range of 50–70 species. Yet, the past few years have seen an explosion in genome scale data (e.g., whole genomes and transcriptomes) and high-powered computer resources allowing for more data rich analyses.

Reconstructing eukaryotic phylogeny is particularly challenging given the approximately 1.8 billion year age of this clade (Knoll et al. 2006; Parfrey et al. 2011), the tremendous diversity among eukaryotes (Adl et al. 2012), and the impact of both lateral and endosymbiotic gene transfer (LGT and EGT; Andersson 2005; Keeling and Palmer 2008; Archibald 2009; Cavalier-Smith 2010; Hug et al. 2010; Nowack and Melkonian 2010; Szklarczyk and Huyten 2010). Eukaryotes are marked by tremendous heterogeneity in rates and patterns of molecular evolution, including the rapid evolution of proteins in many lineages (Dacks et al. 2002; Zufall et al. 2006; Yoon et al. 2008; Cuomo et al. 2012). Despite these difficulties, five major clades of eukaryotes have emerged: Amoebozoa, Excavata, Opisthokonta, Archaeplastida, and SAR (stramenopila, alveolate, and rhizaria; e.g., Adl et al. 2012; Katz et al. 2012).

The evolutionary history of eukaryotic genomes has also been impacted by the transfer of genes from endosymbionts (i.e., EGT; Martin et al. 1993; 2002). Eukaryotes acquired plastids by endosymbiosis of a cyanobacterium or through secondary endosymbiosis in which a heterotrophic eukaryote engulfed a red or green alga (Delwiche 1999; Cavalier-Smith 2004; Archibald 2009). The major clade Archaeplastida includes green algae, red algae, and glaucophytes, and is argued to have diverged following a single primary acquisition of chloroplasts (Adl et al. 2005). The remaining lineages of photosynthetic eukaryotes (e.g., diatoms, brown algae, euglenoids, cryptomonads, haptophytes, and dinoflagellates) are argued to have resulted from secondary, or in the case of some dinoflagellates, tertiary, or perhaps even quaternary endosymbiosis (Delwiche 1999; Archibald 2009).

Given the importance of taxon sampling for phylogenetic reconstruction (Heath et al. 2008), we aimed to capture as broad a diversity of eukaryotes as possible. For example, in studies of eukaryotes (Parfrey et al. 2010), Metazoa (Dunn et al. 2008; Hejnol et al. 2009), and vertebrates (Wiens 2005; Wiens and Tiu 2012), inclusion of taxa in key positions has been critical in stabilizing phylogenies even if these taxa introduced extensive missing data. With well-sampled phylogenomic data, it is also possible to test the idea that robust phylogenies are better obtained from analyses of randomly chosen sites when compared with data
sets of the same size constructed from unlinked genes (Cummings et al. 1995, 1999; Cummings and Meyer 2005).

Here we combine a taxon and gene rich approach, with up to 802 species (232 eukaryotes, 486 bacteria, and 84 archaea) and up to 1554 genes (~500,000 characters), to assess the evolutionary relationships among eukaryotes and the position of eukaryotes on the tree of life. Inclusion of bacterial lineages also enabled us to identify eukaryotic genes that fall sister to cyanobacteria and are hence likely the result of EGTs following acquisition of plastids. The large size of our data set allows us to assess the stability of clades using a variety of subsampling approaches.

**MATERIALS AND METHODS**

**Taxon and Gene Selection**

Given our goal of creating a taxon rich phylogenomic study, we collected data from diverse eukaryotes containing at least 100 proteins available on public databases, and curated these sequences in a custom-built pipeline as described in Grant and Katz (2014). To this end, we began with the taxa included in OrthoMCL (Li et al. 2003; Chen et al. 2006), a database of over 124,000 orthologous groups from whole genomes of 98 eukaryotes, 44 bacteria, and 16 archaea. We then added sequence data from all eukaryotic taxa in GenBank with at least 100 proteins either in the protein database or as EST projects and data from transcriptome projects from our laboratory (*Chilodonella uncinata*, *Subulatomonas tetraspora*, *Corallomyxa tenera*) or from the Marine Microbial Eukaryote Transcriptome Sequencing project (http://marinemicroeukaryotes.org/; *Trichospherium* sp. ATCC40318, *Filamoeba nolandii* ATCC50430, *Mayorella* sp., *Stromcomixa ramosa*, *Pesonella* sp. PRA-29, *Rhodomonas lens*, *Tetraselmis chuii*, *Mesodinium pulex*, *Strombidinopsis* sp., and *Tiarina fusus*).

To maintain some taxonomic evenness, we subsampled from groups that are well represented in GenBank (i.e., plants, animals, fungi, and some parasitic microbial lineages such as *Plasmodium* and *Trypanosoma*). We also removed taxa from the analyses if the data appeared to be highly contaminated; for example, the EST data from the rhizarian plant parasites *Plasmodiophora brassicae* and *Spongospora subterranea* included many sequences that fell within Archaeplastida in preliminary analyses. Other taxa (e.g., *Welwitschia mirabilis* and *Ginkgo biloba*) were removed because they had few data that remained in our matrices after data refinement was complete. In the final analyses, we also removed the microsporidian fungi as they consistently fell on very long branches and their position was unstable across analyses.

To capture putatively vertically inherited genes, we identified 1554 orthologous groups from OrthoMCL (Chen et al. 2006) that were present in at least four of five major eukaryotic clades (i.e., Opisthokonta, Archaeplastida, Amoebozoa, Excavata and SAR) as represented in OrthoMCL (http://orthomcl.org/orthomcl/; Fig. 1 and Supplementary Table S2).

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Depiction of output from phylogenomic pipeline reveals heterogeneity in data availability for the 1554 genes sampled. Each column represents a gene and the warmth (darkness) of the color represents the number of sequences captured for each gene and clade. The genes are ranked by evenness as measured by representation across both major clades and subclades, and the most evenly sampled 150 genes are marked by the bar at the bottom of figure. The last row indicates genes identified as impacted by EGT as orthologs in photosynthetic eukaryotes nest among cyanobacteria. Am = Amoebozoa; ac, Acanthamoebidae; ac, Archamoebae; da, Dactylopodida; di, Dicyoteldida; h, Filamoebia, is, incertae sedis; my, Mycetozoa; v, Vannellidae; EE = “everything else”: ap, Apusomonadidae; br, Breviatea; cr, Cryptophyta; ha, Haptophyta; ka, Katablepharidophyta; Ex = Excavata; eu, Euglenozoa; fo, Foraminifera; he, Heterolobosea; ja, Jakobida; ma, Malawimonadida; ox, Oxymonadida; pa, Parabasalia; Op = Opisthokonta; ch, Choranglidellia; tu, Eug; sc, lichytesporopora; me, Metazoa; Pr = Archaeplastida/Plantae; gl, Glaucocystophyceae; gr, Virdiplantae; rh, Rhodophyta; Sr = SAR: ap, Apicomplexa; ch, Chromerida; ci, Ciliophora; di, Dinophyceae; pe, Perkinsida; rh, Rhizaria; st, Stramenopiles.)
available from http://dx.doi.org/10.5061/dryad.db78g). We generated alignments using a pipeline that is described in detail in Grant and Katz (2014). In brief, genes are identified first by BLAST, using data from OrthoMCL as query and then pairwise comparisons of sequences are used to remove too similar (>98% identical; e.g., allelic variation) or too divergent (<70% identical; e.g., non-orthologous) data. Alignments are then refined using Guidancce (Penn et al. 2010) and RaxML ( Stamatakis 2006) is used to build single gene trees. The resulting data set had 802 taxa (232 eukaryotes, 486 bacteria, and 84 archaea) and up to 493,050 amino acids. The well-sampled marker SSU-rDNA gene was also concatenated to some gene matrices. To maximize the power of our tree reconstructions, we put considerable effort into analysis of the 150 genes that had the most evenly distributed taxa across the eukaryotic clades (e.g., Excavata, Amoebozoa, Alveolata, Rhizaria, Stramenopile, Archaeplastida, and Opisthokonta; indicated in Fig. 1).

**Phylogenetic Analyses**

For all analyses, the best fitting model for each gene was determined using ProtTest (Darriba et al. 2011). The model LG was determined to be the best fitting model for all but 8 of the 1554 genes. Of the eight, six had WAG as the best fit, one had JTT, and one had BLOSUM, but of these genes all had LG as the second-best fit with a deltaBIC < 10. Because estimates from ProtTest were very similar for +G and +G+I, and because of the computational cost of partitioning these data, we ran all amino acid matrices under the PROTGAMMALG model using default parameters as implemented in RAxML. Where SSU-rDNA nucleotide data were included, we partitioned the data to run with GTRGAMMA for nucleotide data and PROTGAMMALG for amino acids.

Phylogenies were reconstructed in one of several ways: 1) for smaller and more tractable data sets we used RAxML (Stamatakis et al. 2005; Stamatakis 2006; Berger et al. 2011) with rapid bootstrapping of 100 replicates (Stamatakis et al. 2008) followed by full maximum likelihood reconstruction; 2) for larger data sets, and for subsampling, we used ExaML (Stamatakis and Aberer 2013), a faster implementation of the RAxML algorithm that provides the topology of the most likely tree without bootstrap support (BS); 3) we also ran an analysis in PhyloBayes (PhyloBayesMPI 1.4f; Martillot et al. 2009) as implemented at CIPRES (Miller et al. 2010) using the GTR CAT model, to assess the stability of our results to changing models. Although the PhyloBayes run did not converge after approximately 52 h on a 192 core computer (up to 2750 cycles), likely due to instability among closely related taxa, deeper nodes were consistent with our other analyses.

Analyses are named in part based on their inclusion of either the 150 most evenly sampled genes or the full set of 1554 genes ("150" vs. "1554"), whether they were run without genes impacted by ("EGT" vs. "+EGT" and varying taxon inclusion to focus only on eukaryotes ("E"), or to include also Bacteria and Archaea ("A, B, E"); Table 1). We explored the impact of missing data using a range of levels (40–70%) and found that topologies were generally stable to varying levels of missing data (i.e., supported clades retained across analyses). Based on these preliminary analyses, we choose an arbitrary cut-off of more than 50% missing data (per column) for removal in all analyses reported here. As detailed in the legend for Figures 1 and 2, we also used a standard naming system with the first two letters referring to major clade (e.g., Op for Opisthokonta), the second two letters referring to subclade (e.g., me for Metazoa), and the next four letters referring to species (e.g., hsap for Homo sapiens).

**Comparing Sampling by Sites versus Genes**

Data sets of the size generated by our pipeline are too large to be analyzed by most phylogenetic software programs. This allowed us to ask whether researchers analyzing phylogenomic data are best off subsampling by genes or by randomly chosen sites. Subsampling by genes at random might enable capture of diverse genes with varying levels of functional constraint, whereas subsampling by sites may better fit the assumption of site independence inherent in most phylogenetic models. We subsampled the full data set of 1554 genes and 493,050 amino acids (again masking columns with ≥50% missing data), choosing random genes or random sites to make new matrices of varying lengths from 1000 to 45,000 characters. For each sampling strategy, we made 20 alignments of lengths 1000, 5000, 10,000, 15,000, 20,000, 25,000, 30,000, 35,000, 40,000, and 45,000. These were analyzed with ExaML and the proportion of the 20 topologies that supported monophyly of a particular clade is reported.

We assessed the impact of fast evolving sites in our analyses using TIGER (Cummins and Mclnerney 2011) with default parameters. TIGER infers the evolutionary rate of homologous sites, and groups sites with similar rates into bins. We removed sites from the two and three highest bins (i.e. fastest 14,800 and 26,017 sites, respectively) from the alignments “150” and “150-EGT” with eukaryotes only, respectively, and analyzed the resulting matrices in RAxML as described above.

**Endosymbiotic Gene Transfer**

To assess EGTs and possible contamination, we also included whole, genome data from 486 bacteria and 84 archaea, limiting ourselves to two species per genus for well-sampled genera. Prior to concatenation, we removed possible contaminating sequences identified as single eukaryotes (i.e., without eukaryotic sisters) within bacterial and/or archael clades. To search for genes with putative signal of EGT, we use python scripts that employed the tree-walking methods in p4 (Foster 2004) and identified genes with eukaryotic sequences sister to or nested within cyanobacteria. We identified 122 genes...
from the 1554 (12 of which were from the 150 most even genes) that are putative examples of EGT. We removed these genes from some analyses to assess the impact of EGT (e.g., 150-EGT and 1554-EGT).

Analyses of 150 Most Evenly Sampled Genes Yield Robust Estimate of the Eukaryotic Tree of Life

We put the bulk of our effort into analyses of the 150 genes that were most evenly sampled across major clades, representing a data set of 36,346 characters (Fig. 1 and Table 1 analysis “150”). The resulting topology supports five major clades of eukaryotes (Fig. 2)—Amoebozoa, Excavata (without the genus Malawimonas), Opisthokonta, Archaeplastida, and SAR—and is consistent with a subset of previous analyses that used fewer taxa (Hampi et al. 2009) or fewer genes (Yoon et al. 2008; Farrey et al. 2010). The placement of Malawimonas has also been unstable in other analyses (Simpson et al. 2006; Hampi et al. 2009;
Figure 2. Most likely tree from 150 most even genes, including 232 eukaryotes and 36,346 characters, reveals five major clades of eukaryotes and several "orphan" lineages (lineages lacking clear sister taxa at this time). The tree is drawn rooted between Opisthokonta and the remaining eukaryotic clades, though we recognize that the root of the eukaryotic tree of life is under debate. Subclades indicated in bold are monophyletic and include all subclades except the diatoms as the single stramenopile taxon, *Vaucheria litorea*, interrupts the monophyly of diatoms in this analysis. Notations on branches indicate full BS (asterisk) or ≥80% BS (filled circles). This analysis is entitled "150" in Table 1, and abbreviations are as in Figure 1.
Derelle and Lang 2011; Zhao et al. 2012), suggesting that this genus may not belong to the major clade Excavata but may instead represent an orphan lineage (i.e., one without clear position given current taxon sampling). In a RAxML analysis, BS for these major clades varied from 100% for Opisthokonta, to more than 80% for Amoebozoa, Archaeplastida, and Excavata (without Malawimonas), to 66% for SAR (Table 2 and Table 1 analysis “150”).

Nested relationships within major eukaryotic clades are generally concordant with previous phylogenies, with monophyly indicated by bold text in Figure 2. Monophyletic clades include animals, fungi, green algae, embryophytes, red algae, brown algae, and apicomplexa (Table 1, online Supplementary Data). The one exception is that the diatoms are not monophyletic as the single representative of the golden algae, *Vaucheria litorea*, falls on a long branch within the diatoms (Fig. 2). Analyses with varying taxon inclusion and software (i.e., RAxML, ExaML, and PhyloBayes) generate similar topologies (Table 1). Removing rapidly evolving sites from these alignments using TIGER (Cummins and McInerney 2011) yields consistent topology with lower support for most clades. We depict the topology rooted between Opisthokonta and the remaining eukaryotes (consistent with some previous studies (Stechmann and Cavalier-Smith 2011; Derelle and Lang 2011; Katz et al. 2012), though we recognize the debate around the placement of the root of the eukaryotic tree of life.

As further evidence of the robustness of analyses of the 150 most even genes, several hypotheses of relationships within major clades are supported. For example, we recover the monophyly of ascomycetes and basidiomycetes within the fungi (Hibbett et al. 2007) and the Discoba within Excavata, which contains Heterolobosea, Euglenozoa, and Jakobida (Fig. 2 and Supplementary Table S3; Hampl et al. 2009). We also assessed the controversial topology for early diverging animal lineages and found that our analyses yield a topology for Metazoa with a clade containing a sponge (*Oscarella*) and ctenophore (*Mnemiopsis*) as sister to all remaining animals (Fig. 2 and Table 1), a finding consistent with some recent phylogenomic analyses (Dunn et al. 2008; Hejnol et al. 2009).

Our taxon rich analyses also highlight the presence of several clades of “orphan” lineages, lineages that lack clear sister taxa and/or are unstable in position across our trees. The first of these clades, “Orphan clade 1” (Fig. 2), contains several photosynthetic lineages such as the small (∼6 μm) picobiliphytes, the katablepharid taxon *Roombia truncata*, and the *incertae sedis* taxon *Telenenu subtilis*. These lineages fall sister to the haptophytes and cryptophytes, two lineages previously considered to be part of the now falsified taxon “Chromalveolata” (Parfrey et al. 2010). The position of these taxa is controversial, in part because some have been placed in the “Hacrobia”, which in turn is a taxon with changing definition (Okamoto et al. 2009; Burki et al. 2012). A second group of orphans, “Orphan clade 2” (Fig. 2), includes the controversial genus *Malawimonas*, that has been argued to belong to Excavata, though it is often found outside of this major clade (Simpson et al. 2006; Hampl et al. 2009; Derelle and Lang 2011; Zhao et al. 2012). In many of our analyses, *Malawimonas* falls sister to *Collodictyon triciliatum* (Zhao et al. 2012), though the placement of these two taxa is unstable (see online Supplementary Data). The final orphan clade includes two taxa previously identified as a potential novel clade of eukaryotes, *S. tetraspora* and *Breviata anathea* (Walker et al. 2006; Katz et al. 2011; Grant et al. 2012), and *Thecamonas (Amastigomonas) tralensis*, a member of the Apusozoa (Cavalier-Smith 1997; Cavalier-Smith and Chao 2003). The position of these lineages varies by analyses (online supplementary Data).

**Analyses of 1554 Eukaryotic Genes**

We also assessed relationships using the full data set of 1554 genes, though the large size here (232 eukaryotes × 493,150 characters) made analyses challenging. For example, full RAxML analyses with bootstraps were not possible even with the high-powered computing that we accessed through CIPRES (Miller et al. 2010) and University Florida Research Computing Center. Instead, we used ExaML, despite differences in reconstructions based on these analyses as compared to RAxML; ExaML generally fails to recover monophyly of Excavata (even without *Malawimonas*) or Amoebozoa because Archamoeba go within Excavata, sister to a clade containing Fornicata and parabasalids (Table 1 analysis “1554”, online Supplementary Data). Given the growing body of evidence supporting these clades (e.g., Hampl et al. 2009; Parfrey et al. 2010; Adl et al. 2012), we suspect that lack of monophyly of Amoebozoa and Excavata (with or without the controversial *Malawimonas*) may be due to differences in implementation of ML analysis in ExaML when compared with RAxML. Beyond differences that might be attributed to software package, the 1554 analyses are generally concordant with analyses of 150 most evenly sampled genes in that we recover Opisthokonta, Archaeplastida, SAR, and many of the nested clades (Table 1).

**Analyses Including Bacterial and/or Archaeal Orthologs**

Inclusion of bacterial and archaeal orthologs enabled us to look at the impact of genes of endosymbiotic origin and the placement of eukaryotes on the tree of cellular life. We identified 122 genes where homologs in photosynthetic eukaryotes appear to be descended from the cyanobacterial ancestor of plastids (see dashes at bottom of Fig. 1, and Supplementary Table S2). These genes were identified by the placement of photosynthetic eukaryotes sister to at least 1 of the 10 cyanobacteria included in our pipeline. Removal of 12 potential EGT genes from the 150 analyses had little impact on the structure of the tree except for in a few places (compare “150 euk” to “150-EGT” in Table 1). Notably, the support for the monophyly of Archaeplastida dropped from 82%...
BS to 40% BS and within Archaeplastida green algae are sister to glaucophytes (75% BS) instead of sister to red algae (67% BS). Removal of 122 potential EGT genes from the 1554 had the same topological effect, though the size of this data set constrained us to use ExaML (Table 1). These data indicate that inclusion of just a few genes impacted by EGT can have a substantial impact on interpretation of evolutionary relationships among photosynthetic taxa.

To evaluate the placement of eukaryotes on the tree of life and to explore root of eukaryotes, we concatenated genes that met two criteria: they included bacteria and/or archaea and generated single gene trees with monophyletic eukaryotes. We restricted ourselves to the trees with monophyletic eukaryotes as a means of lessening the impact of both EGT and LGT on our analyses. Phylogenetic reconstructions of the resulting 207 genes revealed that eukaryotes emerge as a lineage of archaea, sister to species in the archaeal clade Thaumarchaeota (Fig. 3). This result is consistent with multiple analyses that challenge the validity of the three domain hypotheses (e.g., Lake et al. 1984; Tourasse and Gouy 1999; Brown et al. 2001; Williams et al. 2013), including analyses that find eukaryotes emerging from among the “TACK” (Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota) clade of archaea (Williams et al. 2012). Within eukaryotes, the root appears between Fornicata (Excavata such as Giardia lamblia), Archamoeba, and the remaining eukaryotes. However, this root may be due to systematic error such as: 1) neither Amoebozoa nor Excavata are monophyletic in these analyses; and 2) ‘long branch’ lineages such as Trichomonas vaginalis and Entamoeba spp. fall toward root of tree (Fig. 3).

More Genes or More Sites?

Subsampling phylogenomic data for sites chosen at random yields more robust phylogenies than subsampling by gene (Fig. 4). The large size of our data set enabled us to assess clade stability by subsampling characters to build matrices of varying sizes. We took two approaches here, assessing the consistency of nodes from 20 data sets of a given size created using a subsampling of either genes or randomly chosen sites from the supermatrix. In nearly all cases, nodes were more stable in analyses of randomly chosen sites than in matrices of the same size that consisted of gene segments (Fig. 4). For example, 10,000 characters were sufficient to yield monophyletic Opisthokonta in 19 of 20 trees when subsampling by site but only half of the trees generated with same-sized subsamples of genes (Fig. 4 and Supplementary Table S4). Similarly, the monophyly of SAR is more stable in subsamples by site; in data sets of 30,000 characters, SAR is recovered in 19 of 20 trees sampled by site and only 12 of 20 trees by gene (Supplementary Table S4). Our findings at a phylogenomic scale are consistent with analyses of smaller data sets. For example, the same inference on the greater power of random sites when compared with gene segments was found analyzing loci within fully linked mitochondrial genomes (Cummings et al. 1995, 1999). One explanation is that the non-independence of sites within a gene can lead to
FIGURE 4. Subsampling by sites chosen at random yields more robust phylogenetic estimates than subsampling in gene pieces, based on assessment of monophyly of clades in 20 trees produced for each sampling strategy and data set size. Analyses are divided into major clades (top panel), nested hypotheses (mid panel), and groups with morphological synapomorphies (lower panel). Warmer (darker) colors represent greater numbers of trees supporting monophyly, and numerical values are in Supplementary Table S4. Clades recovered significantly more often when sampling by site than by gene are marked with asterisk (*P < 0.05 as assessed by a two-tailed t-test implemented in R (http://www.r-project.org/)). Abbreviations are: Ap, Di, Ch, Pu* = apicomplexa, dinoflagellates, chromerids, and perkinsida and (Sp + Ct) + Me = sponges and ctenophores together, outside of the rest of the metazoon.

inaccurate phylogenetic estimates (Cummings et al. 1995, 1999), which contributed to the suggestion that sampling multiple, unlinked genes from various parts of the genome will help overcome this effect (Cummings and Meyer 2005). However, our sampling of genes chosen at random with respect to their position in genomes did not perform as well as sampling random sites from within these unlinked genes (Fig. 4). Intriguingly, the impact of the sampling effect was greatest on deeper nodes (estimated at ~1 billion years or more), suggesting the need to rethink sampling approaches for studies of early evolutionary events.

SYNTHESIS

Using a data set of up to 802 species and 1554 genes, we are able to construct a robust estimate for the eukaryotic tree of life, including identifying major clades plus several clusters of “orphan” lineages (lineages without clear sister taxa). Inclusion of bacterial and archaeal sequences also enabled us to investigate the impact of lateral gene transfer from plastids to host nuclei (EGT) on estimates of evolutionary relationships among photosynthetic eukaryotes. Finally, we demonstrate that random sampling sites provides better estimates of phylogeny when compared with data sets of the same size made from gene segments.

SUPPLEMENTARY MATERIAL
Supplementary material related to this manuscript has been deposited on Treebase (eukaryotes only concatenated alignments and trees, due to size constraints; TB2:S16260) and Dryad (supplementary data tables, plus all concatenated alignments used for analyses and their associated trees; http://dx.doi.org/10.5061/dryad.db78g).

FUNDING

This work was support by the National Science Foundation A V AT oL grant [DEB-1208741 to L.A.K.].

ACKNOWLEDGMENTS

The authors are grateful to Gordon Burleigh (University of Florida), T. Martin Embley (Newcastle University), Tom Williams (Newcastle University) for help with computer resources and conversations about the work presented here, and to Jean-David Grattepanche (Smith College) for advice on statistics. They also thank to Mark Miller and Wayne Pfeiffer at CIPRES and Charles Taylor and Oleksandr Moskalenko at the University of Florida for support in using computer clusters. The authors also thank William F. Martin (University of Düsseldorf) and two anonymous reviewers for comments on earlier versions of this manuscript. The Gordon and Betty Moore Foundation provided sequence data and assembly through the Marine Microbial Eukaryote Transcriptome project (http://marinemicroeukaryotes.org/; last accessed September 1, 2015).

REFERENCES


