Lower Statistical Support with Larger Data Sets: Insights from the Ochrophyta Radiation

Arnaud Di Franco [1], Denis Baurain [2], Gernot Glöckner,[†,3] Michael Melkonian,[4] and Hervé Philippe*[,1,5]

[1]Station d'Ecologie Théorique et Expérimentale de Moulis, UMR CNRS 5321, Moulis, France
[2]InBioS–PhytoSYSTEMS, Unité de Phylogénomique des Eucaryotes, Université de Liège, Liège, Belgium
[3]Institut für Biochemie I, Medizinische Fakultät, Universität zu Köln, Köln, Germany
[4]Max Planck Institute for Plant Breeding Research, Integrative Bioinformatics, Cologne, Germany
[5]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC, Canada
[†]Deceased.
***Corresponding author:** E-mail: herve.philippe@sete.cnrs.fr.
Associate editor: Tal Pupko

Abstract

It is commonly assumed that increasing the number of characters has the potential to resolve evolutionary radiations. Here, we studied photosynthetic stramenopiles (Ochrophyta) using alignments of heterogeneous origin mitochondrion, plastid, and nucleus. Surprisingly while statistical support for the relationships between the six major Ochrophyta lineages increases when comparing the mitochondrion (6,762 sites) and plastid (21,692 sites) trees, it decreases in the nuclear (209,105 sites) tree. Statistical support is not simply related to the data set size but also to the quantity of phylogenetic signal available at each position and our ability to extract it. Here, we show that this ability for current phylogenetic methods is limited, because conflicting results were obtained when varying taxon sampling. Even though the use of a better fitting model improved signal extraction and reduced the observed conflicts, the plastid data set provided higher statistical support for the ochrophyte radiation than the larger nucleus data set. We propose that the higher support observed in the plastid tree is due to an acceleration of the evolutionary rate in one short deep internal branch, implying that more phylogenetic signal per position is available to resolve the Ochrophyta radiation in the plastid than in the nuclear data set. Our work therefore suggests that, in order to resolve radiations, beyond the obvious use of data sets with more positions, we need to continue developing models of sequence evolution that better extract the phylogenetic signal and design methods to search for genes/characters that contain more signal specifically for short internal branches.

Key words: **phylogenomics, radiations, phylogenetic signal, model of sequence evolution, long branch attraction.**

Introduction

Resolving ancient radiations remains difficult at the age of genomics. They combine three main limitations encountered during phylogenetic inference. First, they are characterized by short internal branches, meaning a scarce genuine (historical) phylogenetic signal, which resides in the rare substitutions accumulated during a short amount of time (i.e., very few synapomorphies). Second, their ancient nature makes the sites subject to substitutional saturation (multiple substitutions at the same site). As a result, an ancient synapomorphy can easily be masked by subsequent substitutions, possibly leading the tree reconstruction method to interpret it as a convergence. The misinterpretation of site history, if not random (i.e., biased), creates a nonphylogenetic signal that conflicts with the genuine phylogenetic signal. Third, all loci possibly do not share the same evolutionary history, because of incomplete lineage sorting (ILS) or, even more problematically, hybridization (Maddison 1997). To sum up, the phylogenetic signal left behind by ancient radiations is both scarce and difficult to extract (Whitfield and Lockhart 2007).

The use of large amounts of data in phylogenomics initially generated great hope to resolve such radiations (Gee 2003). However, the failure of phylogenomics in several important cases suggests that more data may not be sufficient (see Philippe et al. 2011) and might be due to limitations of currently available tree reconstruction methods. To explain the issue, let us consider a data set D containing n sites. The genuine phylogenetic signal contained in D for a given branch B (i.e., the total number of synapomorphies of the data set supporting this branch) is directly proportional to n and $\lambda_B(D)$, the expected number of substitutions per site in branch B for data set D. Yet, this last value is actually unknown, as we only have access to the apparent phylogenetic signal (Baurain and Philippe 2010) inferred by a specific method. Tree reconstruction methods, owing to their limitations, generate spurious phylogenetic signals (that we will collectively embrace under the umbrella term

Article

"nonphylogenetic signal") that compete with the genuine signal. For instance, when a strong bias favors an alternative branching order (e.g., a long branch attraction [LBA] between two unrelated fast-evolving taxa), the nonphylogenetic signal may overcome the genuine signal, resulting in an apparent phylogenetic signal in favor of an alternative, incorrect, branch. This nonphylogenetic signal depends not only on the properties of the radiation (e.g., age) and the data set (e.g., global rate of evolution or taxon sampling), but also ultimately on how correctly the model of evolution infers the substitution history at each position. In cases where all loci do not share the same history, the nonphylogenetic signal might be further increased by other model violations, for example, when using a concatenated model in the presence of ILS. Overall, the failure to resolve most radiations is due to the fact that the phylogenetic signal ($n^*\lambda_B(D)$) is too small with respect to the amount of nonphylogenetic signal generated by model violations. In phylogenomics, n tends to be maximal, leading improvements on $n^*\lambda_B(D)$ to become asymptotically smaller. This is especially true when considering the finite collection of orthologous sequences relevant to the issue at hand. Although alternative approaches based on other types of characters may exist (e.g., retrotransposon insertions or intron positions), they will not be considered here. Thus, the hope of supermatrix-based phylogenomics to resolve radiations mainly hinges on reducing the nonphylogenetic signal.

Because phylogenetic inference should be viewed as a statistical problem (Felsenstein 1983), it requires the formalization of an explicit model. The nonphylogenetic signal generated by probabilistic tree reconstruction methods ultimately depends on the validity of model assumptions (Simion et al. 2020). In other words, model violations increase the nonphylogenetic signal. Two main types of violations exist: 1) violations of the model of sequence evolution and 2) violations of the model of gene evolution. The first type of violation is unavoidable, as we fail to fully apprehend the complexity of sequence evolution, and it affects all tree reconstruction methods. The second violation is due to the fact that, because of ILS, gene duplication or loss, horizontal gene transfer (HGT) or hybridization (i.e., gene flux between closely related species), single-gene trees can differ from the species tree (Maddison 1997), which is not taken into account by the concatenation approach. In theory, gene duplication and HGT are not present, given that only orthologs should be included in the supermatrix. However, ILS can affect orthologs and is expected to be all the more frequent in phylogenies with short durations between speciation events. The effect of these model violations can be studied through the comparison of trees inferred by models fitting data to a different degree or the variation of taxon sampling.

To study the impact of the nonphylogenetic signal on the power of the supermatrix approach to resolve ancient radiations, we chose the diversification of Ochrophyta (i.e., photosynthetic Stramenopiles). Stramenopiles (also known as heterokonts) is a eukaryotic clade composed mostly (but not only, e.g., kelps) of unicellular species, and is closely related to Alveolata and Rhizaria, the three clades forming the supergroup SAR (Burki et al. 2007). Inside Stramenopiles, Ochrophyta is a monophyletic group of photosynthetic organisms that appeared around 500 Ma (Brown and Sorhannus 2010). The diversity of this clade is large, ranging from the picoplanktonic *Nannochloropsis* (Eustigmatophyceae) to ecologically important diatoms (Bacillariophyta) and multicellular brown algae (Phaeophyceae). As photosynthetic eukaryotes, they harbor three genomic compartments, a nucleus (nu), a mitochondrion (mt), and a plastid (cp), the latter inherited from a red algal endosymbiont (Archibald 2015).

The diversification of the major ochrophyte lineages seems to have occurred relatively rapidly, as demonstrated by the corresponding short internal branches (Yang et al. 2012; Derelle et al. 2016). The apparent phylogenetic signal for these branches (B) is expected to vary across compartments. First, the three genomes have a quite different size ($n_{mt} < n_{cp} \ll n_{nu}$), suggesting that we could expect their respective apparent phylogenetic signal to be proportional to the size of each data set (if we assume similar branch lengths across compartments and a negligible role of the nonphylogenetic signal). Second, they have evolved under very different mutation/selection pressures (e.g., different G + C content, presence of recombination in the nucleus but likely not in the organelles), leading to different mean substitution rates and different amounts of nonphylogenetic signal per site, due to differences in types and levels of model violations. Although these values vary across the three genomes, it is difficult to predict whether these variations are major. Moreover, although complex red algae are now thought to have repeatedly exchanged plastids laterally, there is no evidence that it was the case within Ochrophyta (Dorrell et al. 2017; Sibbald and Archibald 2020). Therefore, the latter constitute an interesting case study to evaluate the relative importance of n and of the nonphylogenetic/phylogenetic signal ratio in our ability to resolve ancient radiations. In particular, it might help phylogeneticists to determine whether they should increase n or rather work on reducing model violations.

For this study, we sequenced mitochondrial and plastid genomes from five ochrophyte species belonging to Chrysophyceae, Dictyochophyceae, Pinguiophyceae, and Synurophyceae. From these new data, we built three supermatrices, one for each genomic compartment, all representing most of the major ochrophyte clades. Each data set was carefully constructed, so as to maximize matrix size and completeness, while minimizing erroneous inclusion of nonorthologous genes, contaminated sequences and sequencing errors. With these three largest stramenopiles supermatrices to date, we studied how model violations affect phylogenetic inference. We first observed incongruent topologies between the three genomes for deep ochrophyte relationships, along with surprisingly lower bootstrap support (BS) for the largest data set when using the conventional model LG4X (Le et al. 2012). We then studied the impact of model violations by varying taxon sampling and by using an alternative model of sequence evolution. Finally, we explored ways to resolve the deep ochrophyte radiation.

Table 1. Data Set Composition Summary.

Genomic Compartment	Number of Species	Number of Amino Acid Positions	Number of Genes	Missing Data (%)
Mitochondrion	64	6,762	32	5.84
Plastid	63	21,692	99	4.12
Nucleus	124	209,105	797	25.56

Results and Discussion

Recovery of the Major Ochrophyte Clades

We carefully assembled three supermatrices from mitochondrial, plastid and nuclear genome sequences, containing 6,672, 21,692 and 209,105 amino acid positions, respectively (table 1). They included species from eleven major clades of Ochrophyta: Bacillariophyta, Bolidophyceae, Chrysophyceae, Dictyochophyceae, Eustigmatophyceae, Pelagophyceae, Phaeophyceae, Pinguiophyceae, Raphidophyceae, Synurophyceae, and Xanthophyceae. The nuclear data set also included *Synchroma pusilla*, a species of Synchromophyceae, but some clades (Chrysomeridoephyceae, Phaeothamniophyceae, Schizocladiophyceae, Aurearenophyceae, Phaeosacciophyceae, and Chrysoparadoxophyceae) were absent. Mitochondrial, plastid, and nuclear phylogenies inferred using the LG4X model (fig. 1A–C) retrieved the monophyly of all major clades with maximal BS except Chrysophyceae (see supplementary figs. 1–3, Supplementary Material online). Chrysophyceae came out as a monophyletic group in the mitochondrion and plastid data sets (BS = 56% and 88%, respectively), but were represented by two species only. In the nuclear phylogeny, which includes 12 chrysophycean species, they were paraphyletic, with Synurophyceae nested within Chrysophyceae, in agreement with previous studies (Yang et al. 2012). Monophyly of Synurophyceae + Chrysophyceae (SC clade) was always maximally supported. In the nuclear data set, their grouping with Synchromophyceae (SSC clade) was highly supported, as in Yang et al. (2012) and Derelle et al. (2016). This SSC clade was also recovered in a plastid phylogeny built with partial *Synchroma* sequences obtained from RNAseq data (Keeling et al. 2014) (data not shown). Consequently, in the following, we consider the SSC clade as one of the ten major ochrophyte clades contained in our analyses. The PX clade (Phaeophyceae and Xanthophyceae) (Kai et al. 2008) was recovered using all three data sets, as well as their sister relationship with Raphidophyceae (PXR clade) (Graf et al. 2020), but with limited support in the mitochondrial data set (BS = 92% and 71%, respectively). Two other previously reported relationships (Yang et al. 2012; Derelle et al. 2016; Han et al. 2018) were highly supported—monophyly of Pelagophyceae and Dictyochophyceae (PD clade) and monophyly of Bolidophyceae and Bacillariophyta (BB clade)—but again mitochondrial support was not maximal for the PD clade (BS = 85%). Inside Bacillariophyta, in our nuclear tree, Coscinodiscophyceae were paraphyletic at the base, followed by a monophyletic group composed of Mediophyceae and Bacillariophyceae, as in Parks et al. (2018). Overall, our results were thus in excellent agreement with existing knowledge.

In sharp contrast with the concatenated approach, single-gene phylogenies could only recover the monophyly of the seven well-established, clades (Ochrophyta, BB, Eustigmatophyceae, PD, Pinguiophyceae, PXR, and SSC). Whatever the model of sequence evolution used (LG + G, LG4X, or CAT + G), their monophylies were found on average approximately 46% across the 797 nuclear genes, and only 15, 15 and 17 genes recovered all seven clades, respectively. This result is expected given the age of the ochrophyte radiation (~500 Ma) and the limited size of the genes (~260 positions), which favor systematic and stochastic errors, respectively. However, this prevents us from studying the impact of ILS on the resolution of Ochrophyta radiation. On one hand, coalescent-based methods that jointly infer gene and species trees (e.g., *BEAST) (Heled and Drummond 2010) are still not accessible to phylogenomics because of their computational burden. On the other hand, the proxies to this joint inference (e.g., ASTRAL) (Mirarab and Warnow 2015; Zhang et al. 2018) are too sensitive to single-gene tree estimation errors to be considered accurate (Gatesy and Springer 2014). Consequently, we cannot study the impact of ILS on the amount of nonphylogenetic signal in the case of the ancient ochrophyte radiation and will focus on the effect of violations of the model of sequence evolution on concatenated data.

Incongruence between Compartments for Deep Ochrophyte Relationships

Although the monophyly of the ten major clades was consistently recovered across the three data sets, the basal phylogeny of Ochrophyta showed incongruent relationships depending on which data set was used. Although the plastid tree strongly grouped Eustigmatophyceae with the SSC clade (BS = 100), nuclear and mitochondrial trees separated them, and instead supported the grouping of Pinguiophyceae with SSC on one side (BS = 68% and 74%, respectively) and the grouping of Eustigmatophyceae with PXR on the other side (BS = 81% and 22%, respectively). Our plastid phylogeny (fig. 1*B*) was in agreement with the work of Yang et al. (2012), which was based on nuclear SSU rRNA and four plastid-encoded genes, suggesting that their inferences were dominated by the plastid signal. Comparison with more recently published plastid (Ševčíková et al. 2015), mitochondrial (Ševčíková et al. 2016), and nuclear trees (Derelle et al. 2016) was more difficult, as those data sets lacked some major clades: Bolidophyceae, Dictyochophyceae, and Pinguiophyceae (Ševčíková et al. 2015); Bolidophyceae, Dictyochophyceae, Pinguiophyceae, and Xanthophyceae (Ševčíková et al. 2016); Eustigmatophyceae and Pinguiophyceae (Derelle et al. 2016). Still, the plastid tree of Ševčíková et al. (2015) was congruent with our plastid tree. However, the mitochondrial tree of Ševčíková et al. (2016) did

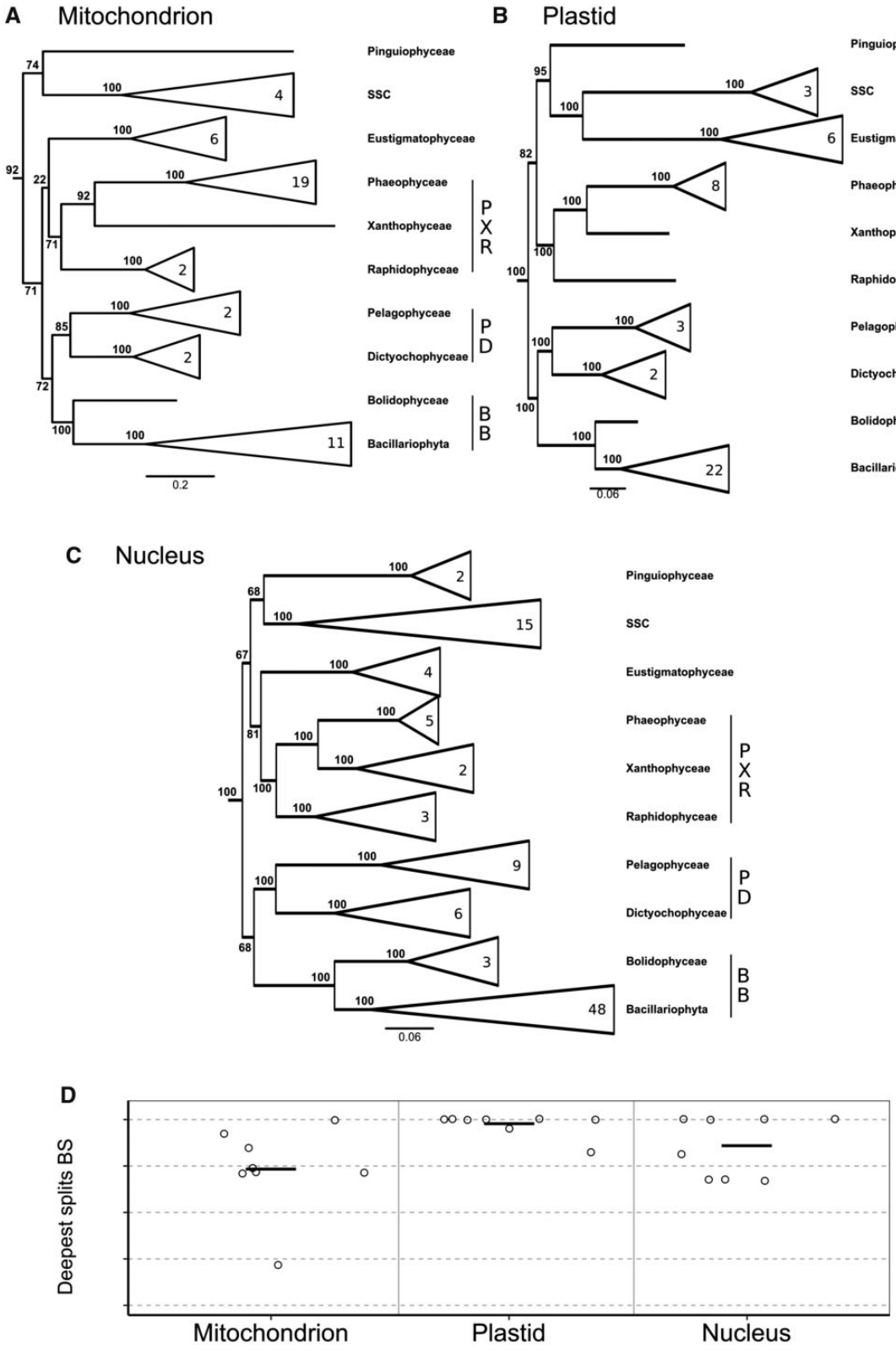

FIG. 1. Collapsed trees inferred under the LG4X model from the three different compartments. Statistical support was computed via 100 fast bootstrap replicates in RAxML. The ten major clades of Ochrophyta were collapsed when more than two species were present. SSC stands for the monophyly of Synchromophyceae, Synurophyceae, and Chrysophyceae, which is only represented by species of Synurophyceae and Chrysophyceae clades in the mitochondrial and plastid data sets. (*A*) Mitochondrial data set—outgroup of 15 species not shown; (*B*) plastid data set—outgroup of 15 species not shown; (*C*) nuclear data set—outgroup of 27 species not shown; (*D*) BS for eight internal splits from the three data sets; average values are indicated by a line.

not recover the monophyly of the PXR clade, contrary to our mitochondrial tree (fig. 1A), but both trees recovered a basal position for Chrysophyceae. In our nuclear tree (fig. 1C), we observed the dichotomy between Diatomista (BB + PD) and Chrysista (PXR + SSC), first proposed in Derelle et al. (2016). Overall, although phylogenies based on the three genomic compartments yielded incongruent deep ochrophyte relationships (fig. 1A–C), they were each in good agreement with previously published trees based on the same compartment (Yang et al. 2012; Ševčíková et al. 2015; Derelle et al. 2016; Ševčíková et al. 2016).

Unexpectedly, statistical support for the eight deep splits that connect the ten major lineages, displayed a surprising pattern with respect to the number of positions (fig. 1D). The average BS for these eight splits increased from 73% in the mitochondrion tree (6,762 positions) to 97% in the plastid tree (21,692 positions), disregarding the fact that these two trees differed for basal relationships. With approximately three times more positions than the mitochondrial data set, the plastid data set thus confirmed the expectation that the apparent phylogenetic signal (as measured by BS) increases with data set size. In sharp contrast, the nuclear data set (209,105 positions), which is approximately ten times larger than the plastid data set (~7.5 times larger if taking into account missing data; table 1), did not follow that expectation, with an average BS of 86%. The deep ochrophyte phylogeny inferred from the three compartments therefore showed not only incongruent relationships but also unexpected statistical supports.

Comparison of the Apparent Phylogenetic Signal across the Three Genomes When Controlling for the Number of Sites and the Number of OTUs

Despite $n_{cp} \ll n_{nu}$, there was more apparent signal in the plastid than in the nuclear data set for the deep branches connecting the major clades as shown by BS values (fig. 1D). However, differences in taxon sampling (64/63 species in the mitochondrial/plastid data sets vs. 124 in the nuclear data set) can affect the amount of nonphylogenetic signal, making our comparison of the three data sets difficult to interpret. To cancel out the impact of taxon sampling, we reduced each data set to a common set of 23 species (22 for the mitochondrion). The phylogenies, inferred with the same LG4X model as above (supplementary fig. 4, Supplementary Material online) were virtually identical to those inferred with more species (supplementary figs. 1–3, Supplementary Material online). Yet, we observed a slight decrease in the apparent phylogenetic signal when reducing the number of species: The average BS for the eight deep splits went down from 73% to 64% for the mitochondrion, from 97% to 93% for the plastid, and from 86% to 69% for the nucleus. This is in agreement with the widely recognized idea that the use of a large number of species improves phylogenetic accuracy, hence reducing the nonphylogenetic signal (Zwickl and Hillis 2002). The higher apparent phylogenetic signal in the plastid versus nuclear compartment was thus still observed, in spite of controlling for taxon sampling.

To better characterize the apparent phylogenetic signal of the three compartments, we used the variable length bootstrap (VLB), or partial bootstrap, approach with the set of common species (Lecointre et al. 1994; Springer et al. 2001; Baurain et al. 2010). Usually, VLB analyses are used to define the number of sites needed to reach a predefined level of apparent phylogenetic signal (e.g., BS = 95%), in order to compare the resolving power of different data sets (Springer et al. 2001; Baurain et al. 2010). Here, they allowed us to study the variation in apparent phylogenetic signal between the three compartments without being affected by the different sizes of the data sets.

Interestingly, VLBs revealed that the apparent phylogenetic signal of most splits was highly similar for the nucleus, the plastid and, to a lesser extent, the mitochondrion (fig. 2). For the monophyly of the major clades (fig. 2A–F), VLB curves always reached 100% BS below 1,000 positions. For the five higher-level groupings (BB, PD, PX, PXR, and BB + PD) that were easily recovered with the three genomes (fig. 1A–C), the curves displayed similar increasing trends between compartments (fig. 2G–K). The mitochondrial data set required more sites to reach a given BS, which could be due to an increased nonphylogenetic signal related to the high rate of evolution in this compartment (Neiman and Taylor 2009). Yet, nucleus and plastid curves were virtually identical, sometimes the plastid increasing slightly faster (fig. 2H) or slower (fig. 2I) than the nucleus. In sharp contrast, support for the monophyly of E + SSC (fig. 2L), as well as their subsequent grouping with Pinguiophyceae and PXR (fig. 2M), rose much faster and higher in the plastid data set than in the two other compartments. None of the bipartitions conflicting with E + SSC (fig. 2N–P) showed the same rapid increase in the mitochondrion or the nucleus, showing that a strong apparent phylogenetic signal only exists in the plastid data set for positioning these taxa.

Hypotheses to Explain the Conflict between Plastid and Nucleus

When controlling for the number of species and the number of sites, the apparent phylogenetic signal in plastid and nuclear compartments was almost identical for most splits, suggesting that the amount of nonphylogenetic signal per site was virtually identical across most splits. The higher average support observed over the eight deep splits with the plastid data set (fig. 1D) was therefore due to a few splits (e.g., E + SSC and E + Ping + PXR + SSC). The comparison of branch lengths in figure 1 allowed us to formulate two hypotheses accounting for this inequality. In the plastid tree (fig. 1B), Eustigmatophyceae and SSC are connected by a long internal branch, which is approximately three times longer than the internal branch basal to Pinguiophyceae and SSC in the nucleus tree (fig. 1C) and evolved much faster than the other clades. The first hypothesis is that E + SSC is correct and that this grouping benefits from a genuinely high value of λ_{E+SSC}(cp). In contrast, the second hypothesis is that E + SSC is incorrect and that their grouping is the result of a LBA artifact. Note that this is not the LBA artifact originally described by

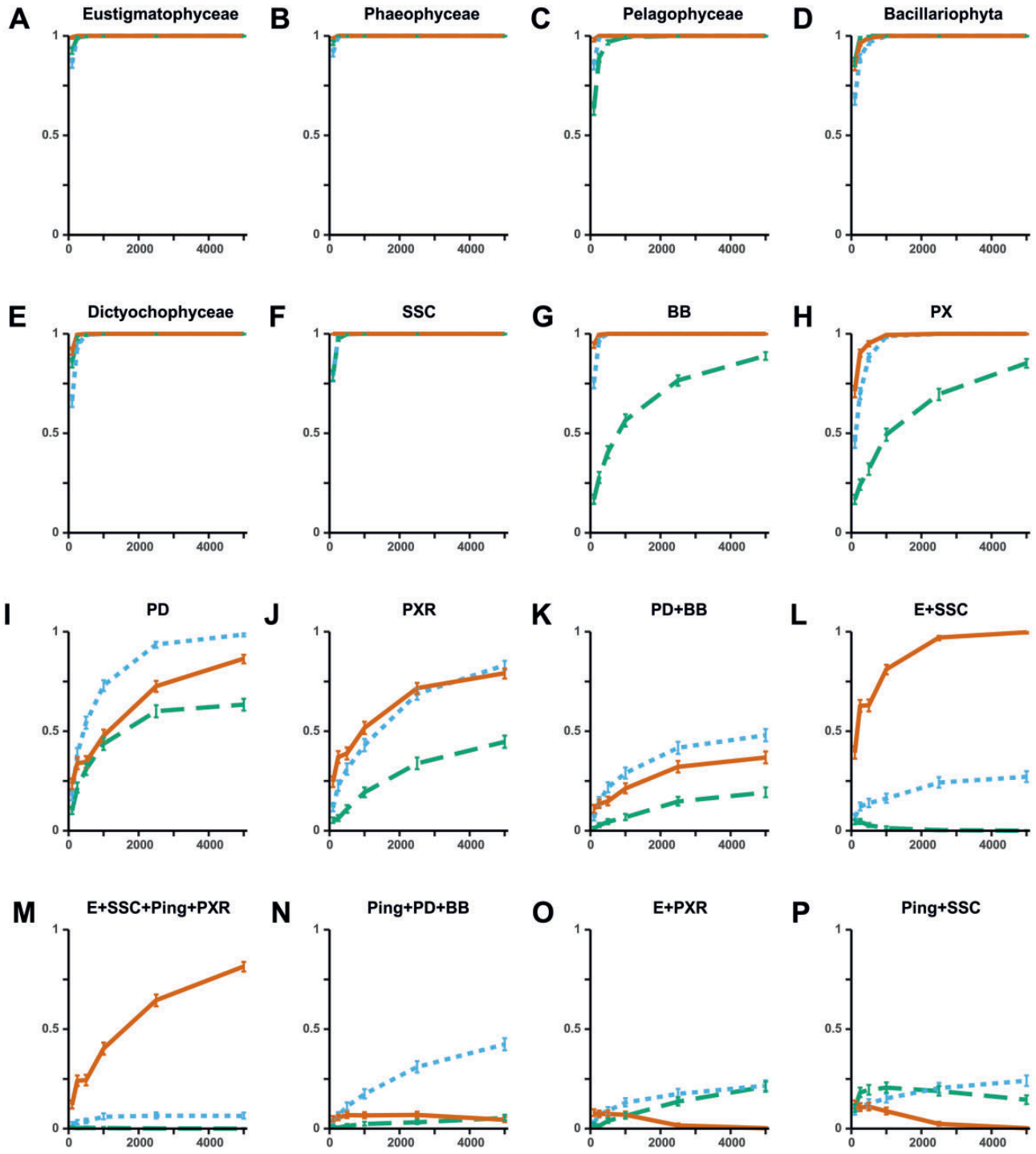

FIG. 2. VLB results for a set of groupings in the three compartments under the LG4X model. *X*-axes represent the number of sites used to infer phylogeny, whereas *Y*-axes represent the BS observed for the grouping of interest. Line colors represent the compartments: nucleus (dotted blue), plastid (solid orange), and mitochondrion (dashed green). Error bars represent Binomial proportion confidence interval for a 95% confidence level (Wilson score interval). (SSC) Synchromophyceae, Synurophyceae, and Chrysophyceae; (BB) Bolidophyceae and Bacillariophyta; (PX) Phaeophyceae and Xanthophyceae; (PD) Pelagophyceae and Dictyochophyceae; (PXR) Phaeophyceae, Xanthophyceae, and Raphidophyceae; (E) Eustigmatophyceae; and (Ping) Pinguiophyceae.

Felsenstein (1978) in the case of maximum parsimony, because probabilistic methods used here do take branch lengths into account. Nevertheless, fast-evolving lineages not only evolve faster but also evolve differently from other lineages, being more subject to heterotachy (e.g., differences in the sets of sites free to vary [Lockhart et al. 2006; Germot and Philippe 1999]) and/or heteropecilly (different substitution processes at play) (Roure and Philippe 2011), which violate the stationarity assumption made by almost all models. For the sake of simplicity, in

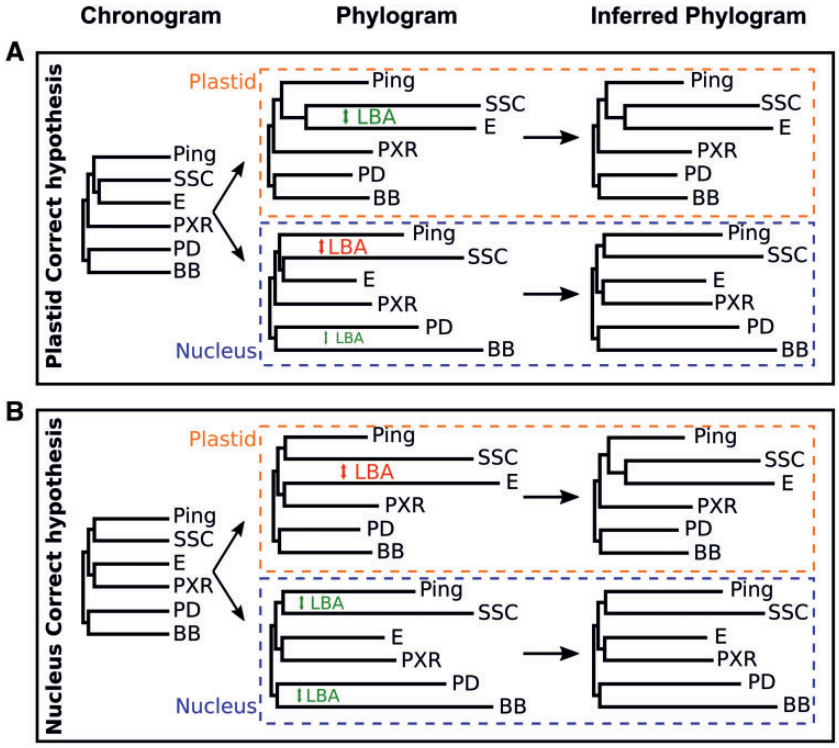

FIG. 3. Two hypotheses to explain the incongruence between plastid and nucleus phylogenies. The first column shows the chronogram assumed to be correct in each hypothesis (top: PC; bottom: NC), whereas the second column shows the corresponding phylograms (with branch lengths further accounting for the evolutionary rate of each lineage) and the third column the phylograms expected to be inferred when using a model unable to deal with LBA artifacts. Phylograms are shown for each compartment (top: plastid; bottom: nucleus). Depending on the true topology and the respective evolutionary rates of the lineages, LBA can either reinforce (green) or overwhelm (red) the phylogenetic (historical) signal, which results in incongruent apparent signals.

what follows, we will present our results in terms of LBA, without explaining anew that LBA in a probabilistic setting is due to model violations.

For illustrative purposes, let us make a simplifying assumption: Either the plastid or the nucleus tree is fully representative of the true phylogeny (fig. 3). In both cases, the durations between speciation events at the base of Ochrophyta are very short (on the left of fig. 3). In the Plastid-Correct (PC) hypothesis (fig. 3A), the long branch length of E + SSC is genuine in the plastid data set (due to an increased evolutionary rate), hence a large λ_{E+SSC}(cp), increasing the amount of phylogenetic signal per position. In addition, as both E and SSC evolved faster, LBA generates a nongenuine signal for this grouping, which leads to the correct topology being inferred with very high BS. In contrast, in the nucleus tree, Pinguiophyceae and SSC evolved faster than E and PXR. Therefore, LBA between Pinguiophyceae and SSC creates a nonphylogenetic signal in favor of the erroneous groupings P + SSC and E + PXR. In the Nucleus-Correct (NC) hypothesis (fig. 3B), E and SSC evolved much faster than the other ochrophytes in the plastid data set, generating a strong LBA artifact that exceeds the genuine P + SSC signal. The nucleus tree was easier to infer, because the fast-evolving taxa (P and SSC on one hand and PD and BB on the other) are in this case sister groups, so LBA reinforces the genuine phylogenetic signal. Because both hypotheses imply an erroneous branching due to a high amount of nonphylogenetic signal, distinguishing between them requires estimating whether the unavoidable model violations are sufficient to generate erroneous trees with the plastid or the nucleus data sets. Here, we applied two commonly used approaches against the LBA artifact (i.e., to reveal the effect of model violations): varying taxon sampling (to favor or disfavor LBA) and using different models of sequence evolution (more or less sensitive to the aforementioned model violations).

Evidence for the Presence of Model Violations

First, we evaluated the impact of major variations of the taxon sampling. The rationale was to reveal a possible artifact of the tree reconstruction method (i.e., model violations sufficiently important to make the nonphylogenetic signal stronger than the genuine phylogenetic signal) through the discovery of incongruence between phylogenies inferred from different subsets of species. We investigated two strategies: 1) use of only a distant outgroup (to increase LBA by creating a long unbroken branch) and 2) independent removal of highly supported ochrophyte lineages. We selected the six clades that were strongly supported by the three data sets: Eustigmatophyceae (E), Pinguiophyceae (Ping), SSC clade (represented by SC in the plastid), PXR clade, PD clade, and BB clade. Because the phylogenetic signal is more accurately extracted with many taxa, we focus on the analyses with complete (albeit different, see above) taxon sampling. Analyses with the common set of 23 species returned

Table 2. Bootstrap Support of High-Level Ochrophyte Clades with Varying Taxon Sampling under the LG4X Model.

	Plastid								Nucleus							
Groupings	**All**	**Out**	**E**	**SSC**	**Ping**	**PXR**	**PD**	**BB**	**All**	**Out**	**E**	**SSC**	**Ping**	**PXR**	**PD**	**BB**
BB + E + PD + PXR			.			.	.	.		78	.			.	.	.
BB + E + PD + PXR + SSC			.			.	.	.		100	.			.	.	.
BB + PD	100	100	98	90	100	99	.	.	68	100	97		100		.	.
BB + PD + Ping					.		.	.				95	.	88	.	.
BB + PD + Ping + PXR		95			.	.	.	.					.	.	.	.
BB + PD + PXR		98			56	.	.	.						.	.	.
BB + Ping					.			.				95	.	88	83	.
E + Ping			.	*67*	.						.		.			
E + Ping + PXR			.	*71*	.	.					.		.	.		
E + Ping + PXR + SSC	82		.	.	.	.	77	53	67		.	.	.	.		100
E + Ping + SSC	95		.	.	.	91	80	66			.	.	.			
E + PXR			.			.			81	82	.	96		.	91	100
E + PXR + SSC			.	.		.					.	.	*83*	.	83	
E + SSC	100	100	.	.	100	100	100	100			.	.	56	91		
Ping + PXR + SSC			*87*	.	.	.					*96*	.	.	.		
Ping + SSC			*95*	.	.				68		97	.	.			100

Rows correspond to the observed high-level groupings and columns to major clades that were left out from the taxon sampling (all means that all species were considered). Dots (.) indicate groupings not testable with the corresponding taxon sampling of the column, italics indicate groupings that are compatible, but not directly comparable, to the corresponding grouping formed when all the species are considered, underline indicates groupings that are not observed when all the species are considered. Abbreviations are as in figure 2, and Out means use of a distant outgroup (i.e., removal of the close outgroup).

comparable results, but with weaker BS values (supplementary table 1, Supplementary Material online). All groupings of the six major clades observed through the 14 taxon sampling variations (2 compartments * 1 + 6 taxon samplings) are reported in table 2.

For the plastid, only two taxon sampling variations produced an incongruent topology, the use of a distant outgroup and the removal of Pinguiophyceae (three incompatibilities, BS shown in underline in table 2). Both resulted in the same likely artifactual topological move: the attraction of the fast-evolving E + SSC group by the outgroup. Attractions were explained by the presence of a longer unbroken branch, either the distant outgroup or the branch of E + SSC in the absence of their slowly evolving sister-group Pinguiophyceae. Taxon sampling variations of the plastid data set revealed that model violations of LG4X sometimes produced LBA artifacts. This suggested that the nonphylogenetic signal generated by this combination of model and data set did not often dominate the phylogenetic signal, suggesting that the hypothesis NC may be correct. Yet, it is important to notice that the grouping E + SSC was always observed.

Variations of the taxon sampling in the nucleus data set showed more incompatibilities with the tree inferred from all species (10, in underline in table 2, corresponding to six alternative groupings) than in the plastid (only three). The incompatibilities were more complicated to understand, as the six clades evolved at a more homogeneous rate in the former than in the latter (fig. 1*B* and *C*). Pinguiophyceae appeared to be the most unstable clade: They emerged as the sister-group to the remaining ochrophytes (BS = 100%) when using the distant outgroup, as the sister-group to BB (BS = 95%) when removing SSC, and as the sister-group to SSC (BS = 100) when removing BB. Pinguiophyceae were only represented by two closely related species (*Phaeomonas* and *Pinguiococcus*), leaving a long unbroken branch at their base (fig. 1C). As BB and SSC are the fastest evolving clades in the nuclear data set, the placement of Pinguiophyceae can be explained by a LBA with the longest branch available in each of the three cases, that is, the outgroup, BB, and SSC, respectively. The limited support for Ping + SSC (BS = 68%) when using the complete data set could then result from an average among these contradictory attractions, all the more so that the competing bipartition (32%) is BB + Ping. The removal of PXR, a relatively slowly evolving clade, had the most dramatic effect, all the deep relationships becoming incongruent. It may have allowed the grouping of E + SSC (BS = 91%), hence reducing the attraction between SSC and Pinguiophyceae, the latter being then attracted by BB (BS = 88%). Interestingly, E + SSC was also recovered through the removal of Pinguiophyceae (BS = 56%). Altogether, these results suggest the presence of a high amount of nonphylogenetic signal under the LG4X model and/or a limited genuine phylogenetic signal in the nuclear data set, thereby supporting hypothesis PC (fig. 3A).

Although taxon sampling variations of the nuclear data set argued in favor of hypothesis PC, as E + PXR and Ping + SSC groupings failed to be robustly recovered, the plastid data set also showed incongruence that may instead argue in favor of hypothesis NC. The higher number of incompatibilities observed with the nucleus than the plastid (10 vs. 3) indicates an amount of nonphylogenetic signal that is lower in the plastid and/or an amount of genuine phylogenetic signal that is lower in the nucleus data set, which yields a less reliable tree. For instance, whereas plastid taxon samplings consistently recovered two high-level clades (E + SSC and BB + PD), the nuclear data set failed to recover any such clade consistently. However, the main result of taxon sampling variations was the evidence for a major impact of model violations with LG4X, especially for the nuclear compartment. These observations are in agreement with the sensitivity to

LBA of models that do not fully incorporate the heterogeneity of the substitution process across sites (Lartillot et al. 2007; Philippe et al. 2011, 2019; Simion et al. 2017). They further suggest that neither the PC nor the NC hypothesis is correct and that we need to use a better model to get insights into the ochrophyte radiation.

Impact of Using a Better Fitting Model of Sequence Evolution

We tested two site-heterogeneous models that have been shown to be less sensitive to LBA (Lartillot et al. 2007): The CAT + Γ model implemented in the Bayesian framework (Lartillot and Philippe 2004) and the C20 + LG + Γ model, an empirical version of the CAT + Γ model implemented in the maximum-likelihood format (Le et al. 2008). We first compared LG4X with C20 + LG + Γ using ModelFinder (Kalyaanamoorthy et al. 2017) from IQ-TREE (Nguyen et al. 2015), which showed C20 + LG + Γ to be better than LG4X with both the plastid and the nuclear data sets (supplementary table 2, Supplementary Material online). Second, cross-validation demonstrated CAT + Γ to have a better fit than C20 + LG + Γ for both data sets (plastid: likelihood difference between CAT + Γ and C20 + LG + Γ of 370 ± 152; nucleus: 488 ± 141). Consequently, the combination of these two tests showed CAT + Γ to have a better fit than LG4X and C20 + LG + Γ for our data sets. Even if CAT + Γ is computationally very demanding (e.g., Philippe et al. 2019), we were able to compute CAT + Γ BS for the complete plastid data set given its moderate size (63× 21,692). In contrast, this model could not be used on the much larger nuclear data set (124×209,105). To make the analysis tractable, we resorted to a gene jackknife approach instead (Delsuc et al. 2008). We chose to generate data sets of approximately 50,000 positions and to run 50 replicates. In the case of LG4X, we verified that the jackknife supports (JS) were comparable with BS, despite being based on approximately four times less positions: As expected, JS values were lower than BS values (supplementary table 3, Supplementary Material online). Yet, and more importantly, the same groupings were recovered in all but one case (the position of SSC when using a distant outgroup). Therefore, JS is a reasonable proxy for BS to evaluate the effect of taxon sampling on the nucleus-based phylogeny inferred with the better fitting CAT + Γ model.

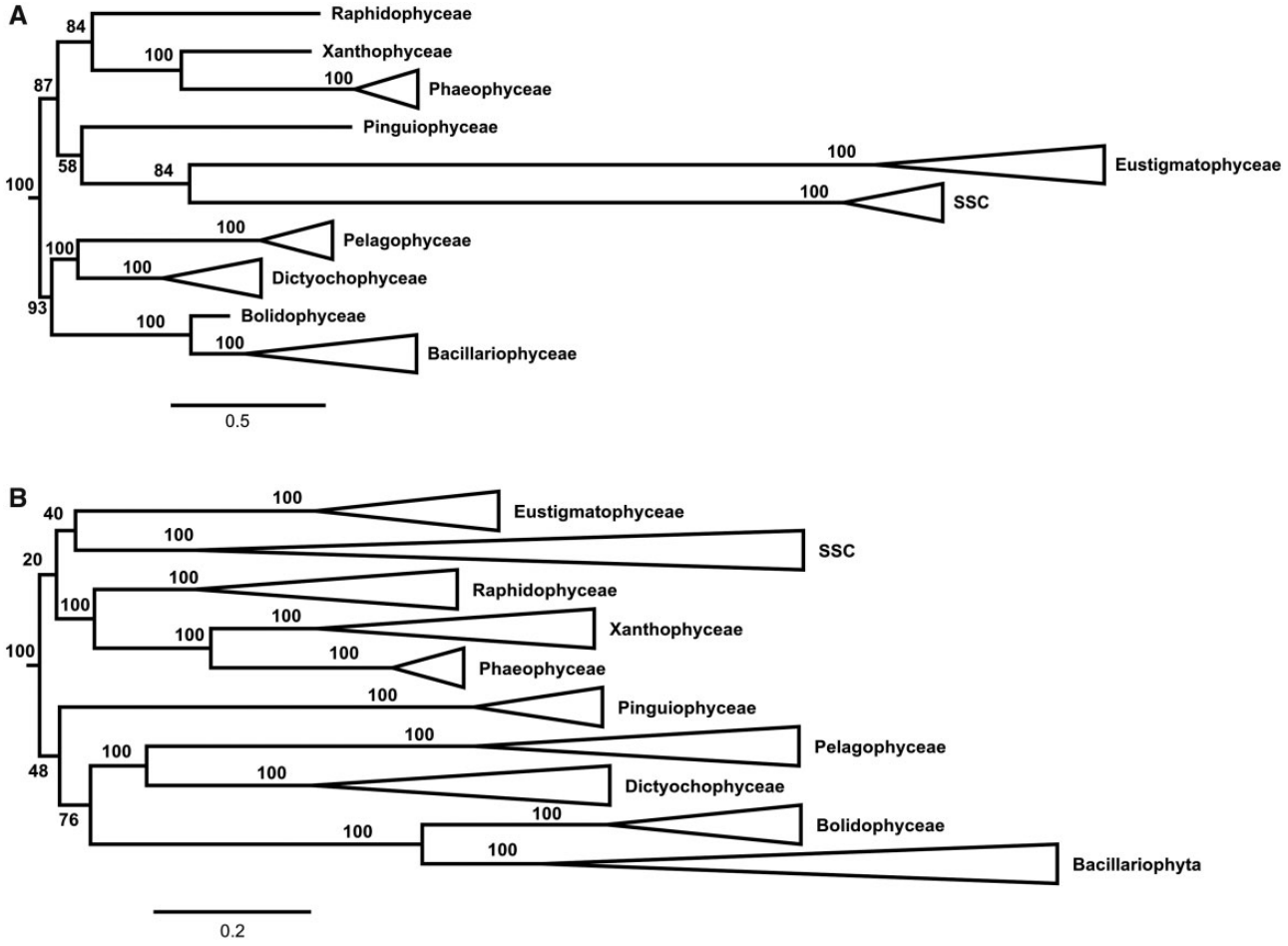

FIG. 4. Phylogenetic trees inferred using PhyloBayes-MPI under the CAT + Γ4 model. The ten major clades of Ochrophyta have been collapsed when more than two species were present. Statistical support values are displayed above their relative splits. (*A*) Plastid data set (63 species and 21,692 positions). Statistical support based on 100 nonparametric bootstrap replicates. (*B*) Nuclear data set (124 species and 209,105 positions). Statistical support based on 50 gene jackknife replicates of about 50,000 positions. Interestingly, the three most frequent alternative groupings for the nucleus, Ping + SSC (40%), E + PXR (20%), and E + Ping + PXR + SSC (20%), were all recovered with LG4X (fig. 1C).

The plastid tree inferred using the CAT + Γ model (fig. 4A) had the same topology as the LG4X tree, but with lower statistical support, especially for E + SSC (BS = 84% vs. 100%) and the position of its sister-group Pinguiophyceae (BS = 58% vs. 95%). Lower support for E + SSC can be explained by the fact that the CAT + Γ model is more suspicious when it has to group two long branches (Eustigmatophyceae and SSC) together. In other words, it assumes more shared amino acids to be due to convergence than LG4X, the very reason for its reduced sensitivity to LBA (Lartillot et al. 2007). In contrast, the topology inferred from the nucleus supermatrix using CAT + Γ (fig. 4B) was different from that inferred with LG4X (fig. 1C): Only the monophyly of BB + PD (JS = 76%) was common among the high-level relationships observed between the six major clades. In the CAT + Γ tree, SSC was sister of Eustigmatophyceae (E + SSC; JS = 54%) instead of Pinguiophyceae, whereas the latter group was sister of BB + PD (JS = 56%). Finally, PXR was weakly grouped with BB + PD + Ping (JS = 32%). Overall, the use of a better fitting model, which likely reduces the nonphylogenetic signal, allows the common recovery of the relationship between Eustigmatophyceae and SSC (E + SSC) by both the plastid and the nucleus data sets, a relationship key to distinguish between the two hypotheses explaining the conflicts observed between the two compartments when using the LG4X model (fig. 3).

To confirm that using a better fitting model decreases the nonphylogenetic signal, and thus reduces incongruence, we performed the same variations of taxon sampling as above, but using the CAT + Γ model. The results (table 3) revealed less incompatibilities within each compartment (1 vs. 3 for the plastid and 4 vs. 10 for the nucleus) and better congruence between the two compartments. In particular, CAT + Γ recovered BB + PD and E + SSC in all analyses of the two compartments. The position of Pinguiophyceae and PXR remained unstable, displaying various sister relationships to one of the two previous clades with limited support. However, the nucleus supported the relationship between Pinguiophyceae and BB + PD in all taxon sampling experiments, except when Eustigmatophyceae were removed, in which case Pinguiophyceae were sisters to fast-evolving SSC, hence possibly a LBA artifact. In contrast, the plastid data set has never recovered BB + PD + Ping. Despite the use of fewer sites (50,000) than LG4X (i.e., increased stochastic error), CAT + Γ thus turned out to be more robust to taxon sampling variations, thereby demonstrating its success in reducing the amount of nonphylogenetic signal.

We then estimated the performance of the computationally more efficient, but more poorly fitting, site-heterogeneous model C20 + LG + Γ. As expected from its intermediate fit between LG4X and CAT, the impact of taxon sampling (supplementary table 4, Supplementary Material online) was intermediate: four incongruences for the plastid and three for the nucleus with C20 + LG + Γ (to compare with 3 and 10 for LG4X and one and four for CAT). Importantly, only one grouping, BB + PD, was consistently recovered in all experiments. Although E + SSC was always recovered by the plastid data set, the nucleus data set found it in only one case (after removal of PXR) and generally grouped Eustigmatophyceae with PXR and SSC with Pinguiophyceae. Albeit less sensitive to taxon sampling, C20 + LG + Γ did not improve the congruence between the plastid and the nucleus, suggesting that its model violations remained serious. The poor performance of C20 + LG + Γ could be due to the limited number of categories used to handle across-site heterogeneities (20), because the CAT + Γ model, which infers the optimal number of categories from the data (Lartillot and Philippe 2004), recovered hundreds of categories (data not shown). We therefore performed the same analysis with C60 + LG + Γ. As expected, C60 + LG + Γ had a better fit than C20 + LG + Γ (data not shown) and produced similar topologies but with a slightly increased sensitivity to taxon

Table 3. Support of High-Level Ochrophyte Clades with Varying Taxon Sampling under the CAT + Γ4 Model.

	Plastid (Bootstrap)								Nucleus (Jackknife)							
Groupings	**All**	**Out**	**E**	**SSC**	**Ping**	**PXR**	**PD**	**BB**	**All**	**Out**	**E**	**SSC**	**Ping**	**PXR**	**PD**	**BB**
BB + PD	93	95	88	90	94	83	.	.	92	82	93	79	98	91	.	.
BB + PD + Ping					.		.	.	56	30		81	.	77	.	.
BB + PD + Ping + PXR									32				.	.	.	.
BB + PD + PXR						.	.	.			27		50	.	.	.
BB + Ping					.			.					.		52	.
BB + Ping + PXR					.	.		.					.	.	*46*	.
E + Ping + PXR			.	*90*	.	.					.		.	.		
E + Ping + PXR + SSC	87	89	.	.	.	.	89	89				.	.	.		
E + Ping + SSC	57	28	.	.	.	83	59	64			.	.	.			
E + PXR			.	80		.					.	52		.		
E + PXR + SSC			.	.	*86*	.				28	.	.		.		44
E + SSC	84	55	.	.	90	100	85	98	54	32	.	.	90	79	54	56
PD + Ping													.		.	*54*
Ping + PXR + SSC			*96*	.	.	.						.	.	.		
Ping + SSC			*86*	.	.						65	.	.			

Rows correspond to the observed high-level groupings and columns to major clades that were left out from the taxon sampling (all means that all species were considered). Dots (.) indicate groupings not testable with the corresponding taxon sampling of the column, italics indicate groupings that are compatible, but not directly comparable, to the corresponding grouping formed when all the species are considered, underline indicates groupings that are not observed when all the species are considered. Abbreviations are as in figure 2, and Out means use of a distant outgroup (i.e., removal of the close outgroup).

sampling (supplementary table 5, Supplementary Material online): four incongruences for the plastid and five for the nucleus (to compare with four and three). This confirms that improvement of model fit does not always improve topological accuracy (Spielman 2020; Yang 1997) and suggests that robustness to taxon sampling is a useful complementary approach. The heterogeneities of the functional constraints across sites in phylogenomic data sets are probably too important for being handled by only tens of categories. As a result, although C20/C60 + LG + Γ are the best fitting models under ML and can be used with the complete data set, the CAT + Γ model appeared as the most suitable to accurately address the difficult question of the Ochrophyta radiation.

Reducing the nonphylogenetic signal favored one facet of hypothesis PC (fig. 3A), that is, the grouping of Eustigmatophyceae and SSC is correct and more highly supported in the plastid than in the nucleus compartment because of an acceleration of the substitution rate (and thus of λ_{E+SSC}(cp)) in the branch at the base of E + SSC in plastid loci. First, E + SSC was always recovered by the two compartments for 16 (2 ×8) different taxon sampling variations. Second, this relationship is also supported by a common split of the plastid-encoded gene *clpC*, which is involved in the protein degradation pathway mediated by the *ClpP* protease (Ševčíková et al. 2015). Third, we observed five common losses of plastid genes in these two clades (ATP synthase CF1 delta subunit, hypothetical protein *Ycf39*/Isoflavone reductase, PSI reaction center subunit XII, hypothetical protein *Ycf35*, and cytochrome b6-f complex subunit 6/*petL*; data not shown), although we cannot exclude convergence because the plastid genome is reduced in both cases. Overall, our experiments proved that decreasing the nonphylogenetic signal through selection of a better fitting model reduced incongruence and increased robustness to taxon sampling variations. Although only two (BB + PD and E + SSC) high-level relationships out of four were consistently recovered in the case of Ochrophyta, the use of the CAT + Γ model demonstrated the key importance of adequately handling the nonphylogenetic signal when trying to resolve ancient radiations.

Using Branch Length Heterogeneity to Tackle Radiations

Hypothesis PC (fig. 3A), which is corroborated by several lines of evidence, postulates an acceleration of the evolutionary rate in the internal branch connecting Eustigmatophyceae and SSC in the plastid compartment (i.e., high value of λ_{E+SSC}(cp)). Although we were lucky that the plastid compartment was enriched with such markers, it is possible that markers displaying a similar acceleration in this branch (hence containing a large amount of phylogenetic signal) are also present in the nucleus. If it is indeed the case, finding markers with a relatively long internal branch (i.e., with a high value of λ) would be helpful. Obviously, looking for such genes is difficult because it requires being able to accurately infer the value of λ. However, testing the potential of such an approach is possible by assuming the knowledge of the correct phylogeny. More precisely, we can estimate branch lengths for each gene, constrained to a candidate topology, and select those displaying the longest (or shortest) length for the branch of interest. Finally, we can infer a phylogeny using a concatenation of the resulting set of markers and compare it to the phylogeny obtained without such a selection to study the effect of filtering the data set by the signal of interest.

We applied this protocol, using the LG4X model, to the nucleus data set by selecting the 200 genes with the longest internal branch at the base of E + SSC, yielding a supermatrix of 47,386 positions ($LONG_{nu}$) and, as a negative control, the 200 genes with the shortest internal branch, yielding a supermatrix of 39,867 positions ($SHORT_{nu}$). Not surprisingly, the phylogeny inferred from $SHORT_{nu}$ with the LG4X model (supplementary fig. 5A, Supplementary Material online) did not recover E + SSC, but strongly grouped Pinguiophyceae and SSC (BS = 96%) and Eustigmatophyceae and PXR (BS = 100%), in agreement with the topology observed with the full data set (fig. 1C). In the absence of a strong genuine phylogenetic signal (for E + SSC), the misleading nonphylogenetic signal dominated, and the support for two erroneous groupings (P + SSC and E + PXR) increased, despite the use of a much smaller data set (BS rose from 68/81 to 96/100, respectively). Note that a zero branch length might also be due to the fact that a locus has a different history (e.g., due to hybridization or ILS), amounting to the branch being nonexistent. When the time separating two nodes is very short, the probability to observe at least one substitution in the corresponding branch is proportional to the size of the genes. We therefore expect the genes having a very short branch length to be shorter in terms of positions than the ones with a long branch. This prediction is fulfilled (200 vs. 238 positions on average, Mann–Whitney test *P* value of 0.017), suggesting that rate variation and short gene length rather than different history are the main causes of the observed short branch lengths.

In contrast, the phylogeny inferred from $LONG_{nu}$ with LG4X (supplementary fig. 5B, Supplementary Material online) strongly supported E + SSC (BS = 100%) with the same complete taxon sampling. This suggests that the genuine phylogenetic signal was now stronger than the nonphylogenetic signal created by the serious violations affecting this model, thereby leading to a strong apparent signal in favor of E + SSC. As the full data set did not support E + SSC, under the assumption that the nonphylogenetic signal per site generated by the use of LG4X is the same for all genes, the nonphylogenetic signal produced over 209,105 positions is probably stronger than the corresponding signal in the 47,386 positions of the $LONG_{nu}$ set of genes. This protocol cannot be used to resolve radiations, because it assumes the species phylogeny to be known, but it can be used to reveal the contradictory attractions present in a large data set (here SSC attracted either by Pinguiophyceae or Eustigmatophyceae), these attractions stemming either from model violations or from the genuine (historical) signal. More importantly, it validates the idea of looking for innovative methods to detect genes with a high signal for internal splits of a species phylogeny, disregarding the global topologies of the gene trees. Such approaches could be another

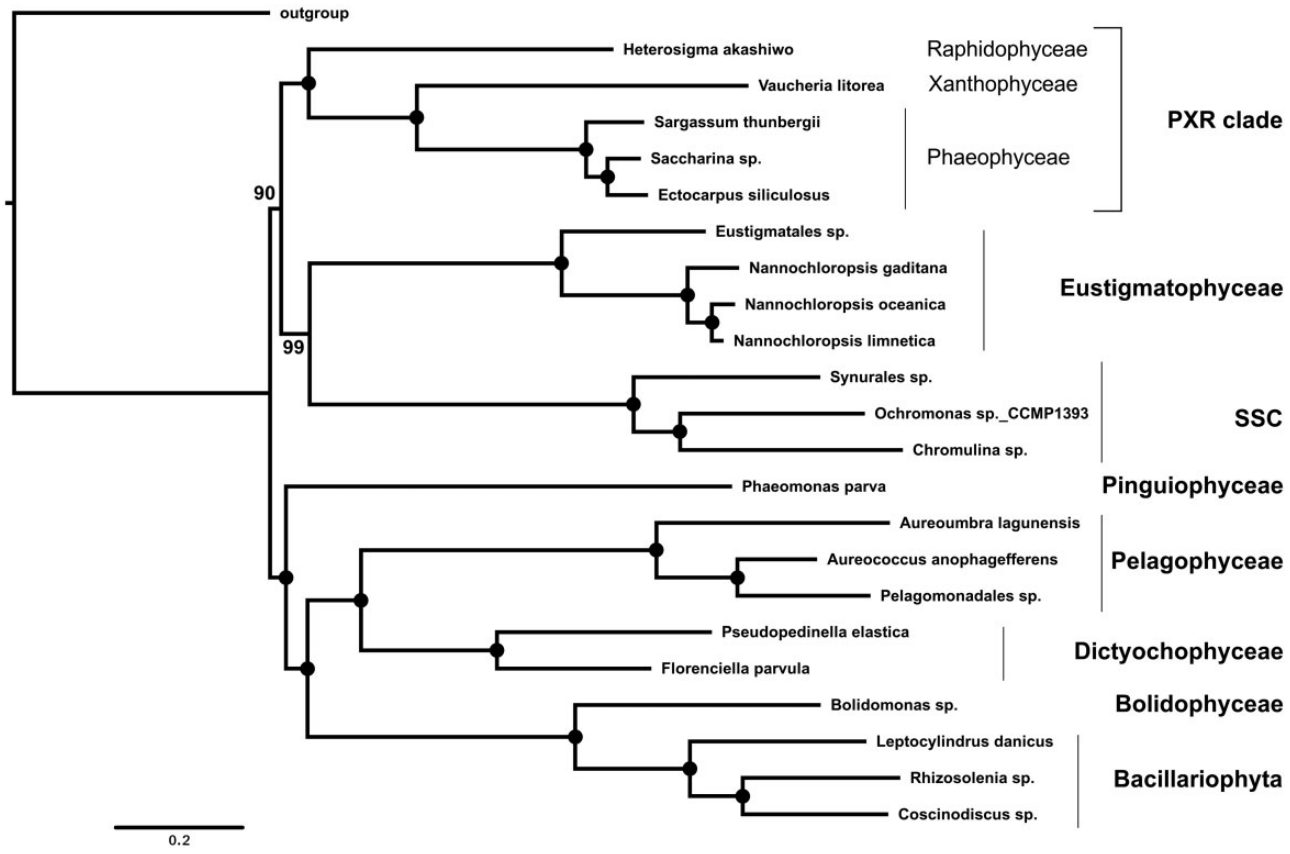

FIG. 5. Consensus phylogenetic tree of the fusion (nu + cp) data set, inferred from 100 jackknife replicates (~80,000 positions) under the CAT + Γ4 model using PhyloBayes-MPI. Statistical support corresponds to JS, with black circles meaning 100% JS. Species named sp. correspond to chimeras between the corresponding species of the plastid and nuclear data set presented in supplementary table 5, Supplementary Material online.

avenue to alleviate the impact of model violations when trying to resolve radiations, without designing ever-more complex evolutionary models.

Toward Resolving the Ochrophyta Phylogeny

Our analysis showed that the resolution of the deep ochrophyte relationships was extremely difficult, because of short internal branches and serious model violations. Interestingly, the small plastid data set appeared to contain a relatively large amount of phylogenetic signal, in particular because of its high value of λ_{E+SSC}(cp). Because the CAT + Γ model did not show evidence of a strong nonphylogenetic signal for the nucleus or the plastid, it should be interesting to combine the high number of positions of the nucleus and the high λ of the plastid to increase the apparent phylogenetic signal of the ochrophyte radiation. Indeed, by combining the plastid and nuclear data sets, we should lengthen at least one of the difficult branches, that is, λ_{E+SSC}(nu + cp) > λ_{E+SSC}(nu), thus making the problem easier to resolve for any method of phylogenetic inference. However, there are potential drawbacks to this approach, such as the fact that combining those data sets would reduce the taxon sampling down to 23 common species, along with the potential introduction of additional model violations (in particular branch length heterogeneity across compartments) (Kolaczkowski and Thornton 2004). Nevertheless, reducing the number of species allowed us to use more sites (80,000) with the best fitting model (CAT).

Such a combined phylogeny inferred with the nu + cp supermatrix using the CAT + Γ model (fig. 5) showed a much higher support for deep ochrophyte relationships: BB + PD (JS = 100%), BB + PD + Ping (JS = 100%), E + SSC (JS = 99%) and E + PXR + SSC (JS = 90%). Interestingly, the nu + cp phylogeny (fig. 5) is different from both the plastid (fig. 4A) and the nucleus (fig. 4B) trees. However, a higher statistical support is not necessarily incontrovertible evidence for a given grouping, as the inference method might be inconsistent. Therefore, we applied the same taxon sampling variations as above (i.e., the use of a distant outgroup and the removal of each major ochrophyte clade) to the nu + cp supermatrix. Interestingly, all seven variations returned trees fully compatible with the phylogeny of figure 5 (table 4). In contrast, the use of a more poorly fitting model (LG4X) on the same supermatrix yielded a lower support and displayed sensitivity to taxon sampling (supplementary table 6, Supplementary Material online), thereby confirming the key role of the model of sequence evolution in the accurate resolution of short internal branches. Even under difficult phylogenetic inference conditions (small number of taxa and residual violations affecting the CAT + Γ

Table 4. Jackknife Support of High-Level Ochrophyte Clades of the Fusion (nu + cp) Data Set with Varying Taxon Sampling under the CAT + Γ4 Model.

	Gene Jackknife of 80,000 Sites							
Groupings	**All**	**Distant**	**E**	**SSC**	**Ping**	**PXR**	**PD**	**BB**
BB + PD	**100**	**100**	**98**	**99**	**100**	**99**	.	.
BB + PD + Ping	**100**	**96**	**88**	**99**	.	**79**	.	.
BB + PD + Ping + SSC				.	.		.	.
E + PXR			.	*100*		.		
E + PXR + SSC	**90**	**98**	.	.	**96**	.	**96**	**86**
E + SSC	**99**	**99**	.	.	**100**	**100**	**100**	**98**
PD + Ping					.		*99*	.
PD + Ping					.		.	*100*
PXR + SSC			*70*	.		.		

Rows correspond to the observed high-level groupings and columns to major clades that were left out from the taxon sampling (all means that all species were considered). Dots (.) indicate groupings not testable with the corresponding taxon sampling of the column, italics indicate groupings that are compatible, but not directly comparable, to the corresponding grouping formed when all the species are considered, underline indicates groupings that are not observed when all the species are considered. Abbreviations are as in figure 2, and Out means use of a distant outgroup (i.e., removal of the close outgroup).

model), the robustness to taxon sampling variations argued for the nu + cp phylogeny (fig. 5) to be a credible working hypothesis for the deep ochrophyte relationships.

Conclusion

A common belief is that increasing the number of positions (n) has the potential to resolve evolutionary radiations. Our work confirms that this is a necessary condition (fig. 2; supplementary table 7, Supplementary Material online), but that heterogeneity of branch length (λ) across loci and model violations cannot be neglected. In particular, the nonphylogenetic signal is a major limiting factor, because our models are necessarily oversimplified with respect to the complexity of biological evolution. The accumulation of data, not of positions but of species, is certainly useful, as the use of more taxa generally helps in the extraction of the phylogenetic signal. Yet, this approach has some serious limitations: 1) some branches are unavoidably unbroken because of extinction, 2) some (rogue) species decrease extraction accuracy, and 3) the resulting increase in computational time limits us to the use of the simplest models. Studies are thus needed to evaluate what are the best compromises between the number of species and the complexity of models to optimize the reduction of the nonphylogenetic signal.

The reduction of model violations achieved when dropping LG4X in favor of the CAT + Γ model allowed us to reduce the incongruence revealed by taxon sampling variations and improve the resolution of the ochrophyte radiation, especially for the nucleus data set. However, the CAT + Γ model is still far from perfect. For instance, it does not take into account the genetic code to weigh amino acid substitutions (see Rodrigue et al. 2010), and it assumes that the evolutionary process is the same all over the phylogeny (e.g., ignoring compositional biases, heterotachy, or heteropecilly). These simplifications are bound to result in model violations that could lead to an incorrect phylogeny. The improvement of models of sequence evolution, both in terms of fit and of computational efficiency, should thus be a priority to resolve ancient radiations. For recent radiations, the impact of these model violations is expected to be more limited (fewer multiple substitutions at the same position), and it is key to address another kind of model violation (not studied in our work), the presence of interspecies gene flux (hybridization) and ILS, using coalescent methods such as *BEAST (Heled and Drummond 2010). However, when a non-negligible fraction of gene trees is different from the species tree, the interest in resolving the radiation is limited because hemiplasy is so frequent that the species tree is no longer useful to study the evolution of characters and organisms (Hahn and Nakhleh 2016).

In addition to the number of positions and the reduction of model violations, the strength of the apparent phylogenetic signal is also dependent on branch length. The length of a given branch is variable across loci, for example, being longer at a locus that underwent reduced purifying selection or directional selection. In the case of the plastid data set, we were lucky to have had a large number of loci that underwent a substitution rate acceleration in the E + SSC basal branch. This acceleration likely explains the observation that the small plastid data set (21,692 positions) is able to strongly recover the monophyly of this clade otherwise very difficult to resolve, whereas the large nuclear data set (209,105 positions) cannot. The difference was more pronounced with the LG4X model than with the CAT + Γ model, probably because the long branches of Eustigmatophyceae and SSC were further artifactually attracted. This "lucky" rate acceleration suggests a new approach to resolve ancient radiations: searching for loci having accelerated in the short internal branches of interest, so as to facilitate extraction of a signal that is scarce for other, more regular, loci.

Finally, combining the nuclear and plastid data sets, along with the use of the CAT + Γ model, helped us to simultaneously increase n and λ and decrease model violations, leading to a well-supported tree, robust to taxon sampling variations. Given the difficulties to resolve the ochrophyte radiation, this phylogeny needs to be confirmed with a richer taxon sampling and/or with a better model. It nevertheless constitutes a working hypothesis to understand in which order the remarkably diverse phenotypes of Ochrophyta emerged, from the picoplanktonic *Nannochloropsis* to the silica frustule-bearing diatoms and to giant marine kelps.

Materials and Methods

Cultures, Organelle Genome Sequencing, and Assembling

Cultures of *Chromulina chionophila* (CCAP 909/9), *Pseudopedinella elastica* (SAG B43.88), *Synura petersenii* (CCAC 0052), *Phaeomonas parva* (CCMP 2877), and *Florenciella parvula* (RCC 446) were obtained from their respective algal culture collections (CCAP: https://www.ccap.ac.uk/; SAG: http://www.uni-goettingen.de/en/culture±collection±of±algae±%28sag%29/184982.html; CCAC: https://www.uni-due.de/biology/ccac/; CCMP: https://ncma.bigelow.org/; RCC: http://roscoff-culture-collection.org/). Algae were grown in the culture media recommended by the collections in aerated 1-l Erlenmeyer flasks at 15 °C and 20 µmol photons/m^2/s in a 14:10 h L/D cycle. They were harvested by centrifugation and, after grinding in liquid nitrogen, total DNA was extracted using either the NucleoSpin Plant II Midi Kit (Macherey-Nagel, Düren, Germany) or a modified CTAB protocol (Rogers and Bendich 1985; see Supplementary Material online).

Sequencing and Assembly of Organelle Genomes

DNA samples were converted to Illumina sequencing libraries according to the manufacturer's protocols and sequenced in paired end mode (150 bases sequencing length). The resulting reads were assembled using ABySS (Simpson et al. 2009). Organellar contigs were extracted using gene sequences from the respective *Ectocarpus* genomes as queries in BLAST searches. Gaps were closed with GapFiller (Nadalin et al. 2012) and annotation was carried out with the sequin tool from NCBI.

Creation of Phylogenomic Data Sets

For each compartment, we assembled the data sets following a semiautomatic protocol similar to the one described in our previous phylogenomic studies (Irisarri et al. 2017; Simion et al. 2017). In summary (see figure 1 of Simion et al. for an overview and https://github.com/psimion/SuppData_Metazoa_2017/blob/master/utilities_src.tgz for software availability), we used protein annotations obtained from genomic data to define orthologous groups with OrthoFinder version 1.4 (Emms and Kelly 2015). Sequence similarity matrices were computed with BLAST for mitochondrial and plastid data sets and with USEARCH (Edgar 2010) for the nuclear data set (e-value threshold = 1e−5) before being divided with the MCL algorithm using the default inflation value (1.5). We filtered the resulting orthogroups for minimal taxonomic representation before validating their orthology relationships. Then we improved their taxon sampling by adding species from transcriptomic and genomic data using 42 (https://metacpan.org/dist/Bio-MUST-Apps-FortyTwo). Detailed description for each compartment, as well as on the computational treatments undertaken to remove paralogous and xenologous sequences from the multiple sequence alignments, can be found in Supplementary Material online. In particular, we used the Branch Length Correlation method (Simion et al. 2017) to detect and remove outlier genes. Finally, our analyses focused on the three data sets summarized in table 1 and available at https://doi.org/10.6084/m9.figshare.7680395.v2.

Phylogenetic Inferences

All supermatrices used in our analyses were concatenated using SCaFoS (Roure et al. 2007). We inferred phylogenetic trees using RAxML version 8.2 (Stamatakis 2014) with the LG4X mixture model (Le et al. 2012) using 100 fast bootstrap replicates. Inferences under the C20 + LG + Γ and C60 + LG + Γ models were carried out using IQ-TREE 1.6.8 (Nguyen et al. 2015). Inferences under the CAT + Γ mixture model (Lartillot and Philippe 2004) were carried out using PhyloBayes-MPI version 1.8 (Lartillot et al. 2013), either on bootstrap replicates for mitochondrial and plastid data sets or on gene jackknife replicates for the nucleus data set. Preliminary analyses demonstrated that convergence was not reachable for a data set of 124 species and 209,105 amino acid positions with the current implementation of PhyloBayes-MPI. Following Delsuc et al. (2008), we thus used a gene jackknife approach and generated replicates of approximately 50,000 or approximately 80,000 positions with a custom script. Convergence assessment and consensus tree construction were performed as in Simion et al. (2017).

VLB Analyses

We reduced each data set to an ingroup taxon sampling of 22 comparable species (21 for the mitochondrion as one out of four Pelagophyceae species was missing), that is, identical or closely related (supplementary table 8, Supplementary Material online). For the outgroup, we used *Guillardia theta* for the plastid and *Phytophthora sojae* for the mitochondrion and *Phytophthora parasitica* for the nucleus. We used distinct species to have a similar branch length leading to the outgroup in each compartment, whereas using the same species (e.g., *G. theta*) would have generated a much longer branch in the mitochondrion/nucleus than in the plastid. Out of the three resulting supermatrices, we drew 1,000 VLB replicates of different sizes (100, 250, 500, 1,000, 1,500, 2,000, and 2,500 sites) and 100 replicates of 5,000 sites using seqboot from the PHYLIP package (Felsenstein 1989). The best tree was obtained for each VLB replicate with RAxML under the LG4X mixture model. Finally, we retrieved the bootstrap proportion of each bipartition for each matrix length with the program consense from the PHYLIP package, and further analyzed them using a custom R script.

Model Comparison

AIC, AICc, and BIC between LG4X and GTR + Γ4 models were computed using ModelFinder (Kalyaanamoorthy et al. 2017) from IQ-TREE version 1.6.8 (Nguyen et al. 2015), with the constrained topology previously obtained under the LG4X model with RAxML. Crossvalidations between GTR + Γ4 and CAT + Γ4 were carried out using PhyloBayes version 4.1. For both plastid and nuclear data sets, ten training data sets of 10,000 positions were used, and likelihoods were computed on ten test data sets of 2,000 positions.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Zehra Çebi for growing up algal strains and for extraction of DNA, and Paul Simion and Rik Verdonck for critical reading of the manuscript. Sequencing was carried out by the Cologne Center for Genomics (CCG). Computations were performed on the supercomputers Mp2 and Ms2 from the Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI), and the Fonds de recherche du Québec—Nature et technologies (FRQ-NT). Computational resources were also provided by the Consortium des Équipements de Calcul Intensif (CÉCI) funded by the F.R.S.-FNRS (2.5020.11), and through two research grants to DB: B2/191/P2/BCCM GEN-ERA (Belgian Science Policy Office—BELSPO) and CDR J.0008.20 (F.R.S.-FNRS). This work was supported by the TULIP Laboratory of Excellence (ANR-10-LABX-41).

Data Availability

The newly sequenced organelle genomes and their corresponding annotations are available online in the NCBI databases with accession numbers ranging from MK546602 to MK546611. The alignments used in this study as well as the resulting phylogenetic trees are available on figshare at the following address: https://doi.org/10.6084/m9.figshare.7680395.v2.

References

Archibald JM. 2015. Endosymbiosis and eukaryotic cell evolution. *Curr Biol*. 25(19):R911–R921.

Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol Biol Evol*. 27(7):1698–1709.

Baurain D, Philippe H. 2010. Current approaches to phylogenomic reconstruction. In: Caetano-Anollés G, editor. Evolutionary genomics and systems biology. Hoboken (NJ): John Wiley & Sons, Inc. p. 17–41.

Brown JW, Sorhannus U. 2010. A molecular genetic timescale for the diversification of autotrophic Stramenopiles (Ochrophyta): substantive underestimation of putative fossil ages. *PLoS One* 5(9):e12759.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* 2(8):e790.

Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H. 2008. Additional molecular support for the new chordate phylogeny. *Genesis* 46(11):592–604.

Derelle R, López-García P, Timpano H, Moreira D. 2016. A phylogenomic framework to study the diversity and evolution of Stramenopiles (=Heterokonts). *Mol Biol Evol*. 33(11):2890–2898.

Dorrell RG, Gile G, McCallum G, Méheust R, Bapteste EP, Klinger CM, Brillet-Guéguen L, Freeman KD, Richter DJ, Bowler C. 2017. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *ELife* 6:e23717. doi:10.7554/eLife.23717.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 16(1):1–14.

Felsenstein J. 1983. Parsimony in systematics: biological and statistical issues. *Annu Rev Ecol Syst*. 14(1):313–333.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*. 27(4):401.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package - v3.2. *Cladistics* 5:164–166. doi:10.1111/j.1096-0031.1989.tb00562.x

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogenet Evol*. 80(1):231–266.

Gee H. 2003. Ending incongruence. *Nature* 425(6960):782–782.

Germot A, Philippe H. 1999. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. *J Eukaryot Microbiol*. 46(2):116–124.

Graf L, Yang EC, Han KY, Küpper FC, Benes KM, Oyadomari JK, Herbert RJH, Verbruggen H, Wetherbee R, Andersen RA, et al. 2020. Multigene phylogeny, morphological observation and re-examination of the literature lead to the description of the Phaeosacciophyceae classis nova and four new species of the Heterokontophyta SI clade. *Protist* 171(6):125781.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. *Evolution* 70(1):7–17.

Han KY, Graf L, Reyes CP, Melkonian B, Andersen RA, Yoon HS, Melkonian M. 2018. A re-investigation of *Sarcinochrysis marina* (Sarcinochrysidales, Pelagophyceae) from its type locality and the descriptions of *Arachnochrysis*, *Pelagospilus*, *Sargassococcus* and *Sungminbooa genera* nov. *Protist* 169(1):79–106.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*. 27(3):570–580.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J-Y, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, et al. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol*. 1(9):1370–1378.

Kai A, Yoshii Y, Nakayama T, Inouye I. 2008. Aurearenophyceae classis nova, a new class of Heterokontophyta based on a new marine unicellular alga *Aurearena cruciata* gen. et sp. nov. inhabiting sandy beaches. *Protist* 159(3):435–457.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14(6):587–589.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*. 12(6):e1001889.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431(7011):980–984.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. 7(Suppl 1):S4.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21(6):1095–1109.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol*. 62(4):611–615.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol*. 29(10):2921–2936.

Le SQ, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24(20):2317–2323.

Lecointre G, Philippe H, Vân Lê HL, Le Guyader H. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Mol Phylogenet Evol*. 3(4):292–309.

Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T. 2006. Heterotachy and tree building: a case study with plastids and Eubacteria. *Mol Biol Evol*. 23(1):40–45.
Maddison WP. 1997. Gene trees in species trees. *Syst Biol*. 46(3):523–536.
Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31(12):i44–i52.
Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13(S14):S8.
Neiman M, Taylor DR. 2009. The causes of mutation accumulation in mitochondrial genomes. *Proc Biol Sci*. 276(1660):1201–1209.
Nguyen L-TT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.
Parks MB, Wickett NJ, Alverson AJ. 2018. Signal, uncertainty, and conflict in phylogenomic data for a diverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *Mol Biol Evol*. 35(1):80–93.
Philippe H, Brinkmann H, Lavrov DV, Timothy Littlewood DJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9(3):e1000602.
Philippe H, Poustka AJ, Chiodin M, Hoff KJ, Dessimoz C, Tomiczek B, Schiffer PH, Müller S, Domman D, Horn M, et al. 2019. Mitigating anticipated effects of systematic errors supports sister-group relationship between Xenacoelomorpha and Ambulacraria. *Curr Biol*. 29(11):1818–1826.e6.
Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*. 107(10):4629–4634.
Rogers SO, Bendich AJ. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol Biol*. 5(2):69–76.
Roure B, Philippe H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evol Biol*. 11(1):17.
Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol*. 7(S1):1–12.
Ševčíková T, Horák A, Klimeš V, Zbránková V, Demir-Hilton E, Sudek S, Jenkins J, Schmutz J, Přibyl P, Fousek J, et al. 2015. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Sci Rep*. 5(March):10134–10112.
Ševčíková T, Klimeš V, Zbránková V, Strnad H, Hroudová M, Vlček Č, Eliáš M. 2016. A comparative analysis of mitochondrial genomes in Eustigmatophyte algae. *Genome Biol Evol*. 8(3):705–722.
Sibbald SJ, Archibald JM. 2020. Genomic insights into plastid evolution. *Genome Biol Evol*. 12(7):978–990.
Simion P, Delsuc F, Philippe H. 2020. To what extent current limits of phylogenomics can be overcome? Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher, Authors open access book. p. 2.1:1–2.1:34.
Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol*. 27(7):958–967.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res*. 19(6):1117–1123.
Spielman SJ. 2020. Relative model fit does not predict topological accuracy in single-gene protein phylogenetics. *Mol Biol Evol*. 37(7):2110–2123.
Springer MS, DeBry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. *Mol Biol Evol*. 18(2):132–143.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. *Trends Ecol Evol*. 22(5):258–265.
Yang EC, Boo GH, Kim HJ, Cho SM, Boo SM, Andersen RA, Yoon HS. 2012. Supermatrix data highlight the phylogenetic relationships of photosynthetic stramenopiles. *Protist* 163(2):217–231.
Yang Z. 1997. How often do wrong models produce better phylogenies? *Mol Biol Evol*. 14(1):105–108.
Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19(Suppl 6):153.
Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol*. 51(4):588–598.