American Journal of Botany

RESEARCH ARTICLE

INVITED SPECIAL ARTICLE
For the Special Issue: Using and Navigating the Plant Tree of Life

Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta)

Matthew B. Parks[1,*], Teofil Nakov[2,*], Elizabeth C. Ruck[2], Norman J. Wickett[1], and Andrew J. Alverson[2,3]

Manuscript received 25 August 2017; revision accepted 18 December 2017.

[1] Daniel F. and Ada L. Rice Plant Conservation Science Center, Chicago Botanic Garden, Glencoe, IL 60022, USA

[2] Department of Biological Sciences, University of Arkansas, 1 University of Arkansas, SCEN 601, Fayetteville, AR 72701, USA

[3] Author for correspondence (e-mail: aja@uark.edu)

*These authors contributed equally to this work.

Citation: Parks M. B., T. Nakov, E. C. Ruck, N. J. Wickett, and A. J. Alverson. 2018. Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *American Journal of Botany* 105(3): 1–18.

doi:10.1002/ajb2.1056

PREMISE OF THE STUDY: Diatoms are one of the most species-rich lineages of microbial eukaryotes. Similarities in clade age, species richness, and primary productivity motivate comparisons to angiosperms, whose genomes have been inordinately shaped by whole-genome duplication (WGD). WGDs have been linked to speciation, increased rates of lineage diversification, and identified as a principal driver of angiosperm evolution. We synthesized a large but scattered body of evidence that suggests polyploidy may be common in diatoms as well.

METHODS: We used gene counts, gene trees, and distributions of synonymous divergence to carry out a phylogenomic analysis of WGD across a diverse set of 37 diatom species.

KEY RESULTS: Several methods identified WGDs of varying age across diatoms. Determining the occurrence, exact number, and placement of events was greatly impacted by uncertainty in gene trees. WGDs inferred from synonymous divergence of paralogs varied depending on how redundancy in transcriptomes was assessed, gene families were assembled, and synonymous distances (Ks) were calculated. Our results highlighted a need for systematic evaluation of key methodological aspects of Ks-based approaches to WGD inference. Gene tree reconciliations supported allopolyploidy as the predominant mode of polyploid formation, with strong evidence for ancient allopolyploid events in the thalassiosiroid and pennate diatom clades.

CONCLUSIONS: Our results suggest that WGD has played a major role in the evolution of diatom genomes. We outline challenges in reconstructing paleopolyploid events in diatoms that, together with these results, offer a framework for understanding the impact of genome duplication in a group that likely harbors substantial genomic diversity.

KEY WORDS diatoms; gene tree; genome duplication; paleopolyploidy; polyploidy; synonymous divergence.

Duplicated genes are a hallmark of eukaryotic genomes. For example, some two thirds of the genes in *Arabidopsis* are present in more than one copy (Ambrosino et al., 2016), a proportion that is typical of most plant genomes (Panchy et al., 2016). Duplicated genes can provide raw materials for evolutionary innovation, thereby representing an important source of novel traits in lineages spanning the eukaryotic tree of life (Ohno, 1970). In flowering plants, for example, gene duplications have been linked to changes in a diverse set of traits, including floral pigmentation and structure, flowering time, disease and herbivore resistance, fruit characteristics, and stress response (Soltis et al., 2014; Panchy et al., 2016; Soltis and Soltis, 2016). Gene duplication can occur across multiple scales, from small tandem duplications affecting one or a few genes to, most dramatically, doubling of the entire genome (whole-genome duplication [WGD] or polyploidy) (Flagel and Wendel, 2009; Panchy et al., 2016; Van de Peer et al., 2017).

The evolutionary history of angiosperms is marked by ancient polyploidy events, such that a majority of the duplicated genes in *Arabidopsis*, for example, can be traced to a series of at least four separate WGDs dating back to the origin of flowering plants (Bowers et al., 2003; Jiao et al., 2011). In addition to providing a source of novel and potentially adaptive traits, gene and genome duplications can also serve as mechanisms of speciation (Winge, 1917; Lynch and Force, 2000). Whole-genome duplications, in particular, frequently coincide with speciation events in flowering plants (Otto and Whitton, 2000; Wood et al., 2009; Zhan et al., 2016). An association between WGDs and increased diversification rate is also emerging in angiosperms (Otto and Whitton, 2000; Soltis et al., 2009; Tank et al., 2015; but see Wood et al., 2009; Mayrose et al., 2011; Schranz et al., 2012; Kellogg, 2016), highlighting WGD as a potentially important driver of species diversification. Polyploidy has been an important source of genetic novelty in other species-rich lineages as well, including vertebrates (Ohno, 1970; Dehal and Boore, 2005) and fungi (Wolfe and Shields, 1997; Albertin and Marullo, 2012), though it is unclear whether WGD has significantly impacted species diversification in these groups (Santini et al., 2009; Glasauer and Neuhauss, 2014; Laurent et al., 2017). With longstanding genetic model systems and a wealth of genomic data, these groups represent some of the most intensively studied eukaryotes. Growing genomic resources for equally diverse but historically understudied groups have made it possible to begin exploring whether WGD has played a similarly important role in non-model lineages.

With diversity estimates in the tens to hundreds of thousands of species (Guiry, 2012; Mann and Vanormelingen, 2013), a prominent role in the global cycling of carbon and oxygen (Field et al., 1998), a critical position at the base of their native food webs, and a crown age of roughly 200 Myr (Sorhannus, 2007), diatoms are in many respects the angiosperms of the sea. They exhibit many layers of diversity beyond their species richness, including a broad range of ecological niches, life history strategies, and most famously in the diverse patterns and ornamentations of their silicified cell walls (Fig. 1; Round et al., 1990). Very little is known, however, about the primary sources of genetic change underlying the origins and evolutionary shifts in these traits. Many independent lines of direct and indirect evidence collected over decades suggest that WGD may be common in diatoms. For example, although karyotypes are available for very few species, chromosome counts range from $2n$ = 8–130 among raphid pennate species alone (Kociolek and Stoermer, 1989). Flow cytometric measurements have shown substantial variation in genome size, with estimates spanning more than three orders of magnitude among the few dozen species that have been surveyed (Connolly et al., 2008; von Dassow et al., 2008). Within species, a recent genome doubling distinguishes natural populations of the polar centric species, *Ditylum brightwellii* (Koester et al., 2010), and WGDs apparently can occur in strains maintained in long-term cell culture as well (von Dassow et al., 2008). Finally, and perhaps most compellingly, simultaneous fusions of three or four gametes, leading to the formation of autopolyploid auxospores (i.e., zygotes), have been directly observed in several raphid pennate diatoms, including *Cocconeis* (Geitler, 1927), *Craticula* (Mann and Stickle, 1991), *Dickea* (Mann, 1994), *Achnanthes* (Chepurnov and Roschin, 1995), and *Seminavis* (Chepurnov et al., 2002). The latter set of observations, in particular, led to the prediction that polyploidy might be an important driver of speciation in diatoms (Mann, 1994, 1999b). Finally, there is some evidence for polyploidy in non-diatom stramenopiles, the broader lineage to which diatoms belong (Coyer et al., 2006; Ioos et al., 2006). In light of this relatively large body of evidence, the most surprising discovery might be lack of a genomic signature for paleopolyploidy in diatoms.

We compiled new and previously sequenced genomic and transcriptomic data for 37 phylogenetically diverse diatom species to estimate, for the first time, the extent to which diatom genomes have been shaped, if at all, by WGD events. Gene counts, gene trees, and patterns of synonymous sequence divergence (Ks) between gene duplicates identified numerous putatively allopolyploid-driven WGDs across the phylogeny dating as far as back as 200 Myr ago (Ma). We discuss possible modes of polyploid formation in diatoms and outline research directions that will help shed light on the mechanisms and evolutionary consequences of WGD in diatoms.

MATERIALS AND METHODS

Taxon sampling

We sampled 37 diatom species that spanned the known breadth of extant phylogenetic diversity, the bolidophyte *Triparma pacifica*, and two additional pelagophyte outgroups (Appendix S1, see the Supplemental Data with this article).

Transcriptome sequencing and assembly

For newly sequenced transcriptomes, diatom cultures were grown in L1 marine medium (Guillard, 1975) or COMBO freshwater medium (Kilham et al., 1998) at 22°C on a 12 h light/12 h dark cycle. We extracted total RNA from exponentially growing cultures using the Qiagen (Hilden, Germany) RNeasy kit and prepared indexed sequencing libraries with the Illumina TruSeq RNA Sample Preparation Kit v2. Multiplexed libraries were sequenced with the Illumina HiSeq 2000 or HiSeq 4000 platforms. Sequencing reads were deposited in the Sequence Read Archive database maintained by the National Center for Biotechnology Information (NCBI) under BioSample accessions SAMN07688919–SAMN07688929 (Appendix S1).

We filtered and assembled sequencing reads following guidelines outlined in the Oyster River Protocol (MacManes, 2015). Briefly, raw reads were corrected with Rcorrector (Song and Florea, 2015) and quality-trimmed with Trimmomatic (ver. 0.32) (Bolger et al., 2014). The corrected and trimmed reads were filtered for common laboratory vectors and diatom rRNA genes with Bowtie2 (ver. 2.2.3) (Langmead and Salzberg, 2012). Overlapping forward and reverse reads were then merged with BBMerge (ver. 8.8) (Bushnell et al., 2017), and merged and unmerged reads were assembled with Trinity (ver. 2.2.0) (Grabherr et al., 2011). Assembled transcripts were translated into amino acid sequences using TransDecoder (ver. 2.0.1) (https://transdecoder.github.io/), with translation predictions enabled by searches of the longest identified open reading frames to the Swiss-Prot and Pfam (Finn et al., 2015) databases using NCBI-BLASTP (ver. 2.3.0+) (Camacho et al., 2009) and HMMER (Eddy, 2011), respectively. Assembly quality was measured by TransRate scoring (ver. 1.01) (Smith-Unna et al., 2016) and recovery of conserved eukaryotic orthologs present in the BUSCO database (Simão et al., 2015).

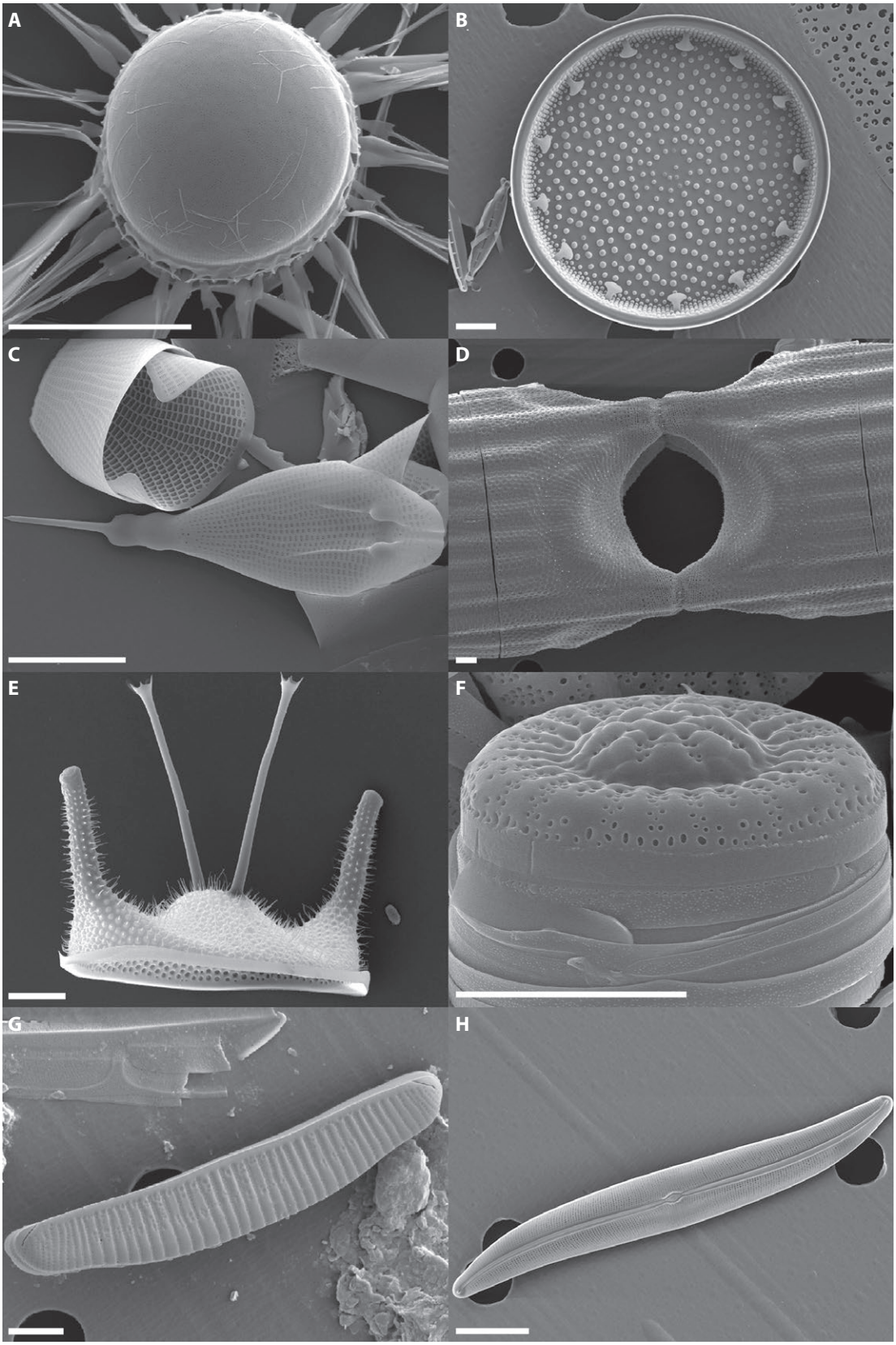

FIGURE 1. Scanning electron micrographs of the siliceous cell walls of select diatom taxa: (A) *Corethron hystrix*, (B) *Actinocyclus* sp., (C) *Rhizosolenia fallax*, (D) *Lampriscus shadboltianus*, (E) *Odontella longicruris*, (F) *Discostella stelligera*, (G) *Eunotia* sp., and (H) *Gyrosigma* sp. Scale bars = 5 μm.

Orthology/paralogy-based transcriptome clustering

We used CD-HIT (-c 0.99 -n 5) (Fu et al., 2012) to remove redundant isoform transcripts from the full set of amino acid sequences for each species. We used NCBI-BLASTP (e-value ≤ 10^{-5} and max-target sequences = 100) to search the resulting nonredundant transcriptome of each species against a database of all 40 (nonredundant) transcriptomes and used this output to identify putative orthologous clusters with MCL (ver. 12-135) (Van Dongen, 2001; Enright et al., 2002; Van Dongen and Abreu-Goodger, 2012), using an e-value cutoff of 10^{-30} and an inflation value of 1.4. Any MCL clusters with fewer than four taxa were excluded from subsequent analyses.

Species-tree reconstructions

We used the "phylogenomic_dataset_construction" pipeline of Yang and Smith (2014) to build and prune ortholog trees for species-tree reconstruction. As part of this pipeline, we aligned sequences with MAFFT (ver. 7.309) (Katoh and Standley, 2013) and reconstructed gene and ortholog trees with RAxML (ver. 8.2.9) (Stamatakis, 2014) using the PROTCATWAG model and 100 rapid bootstrap pseudoreplicates per alignment. As part of the pruning pipeline that selects a single representative transcript (per taxon, per orthologous cluster) for phylogenetic analyses, alignments were trimmed to (1) include only sites with column occupancy ≥0.1, (2) remove terminal branches longer than two branch-length units or 10 times longer than the sister branch, (3) remove subclades subtended by branches longer than two branch-length units, and (4) prune sister tips belonging to the same taxon to include only the tip with the largest number of unambiguous characters in the trimmed alignment. We used Yang and Smith's (2014) RT strategy to create final ortholog alignments with a single representative transcript per sample, with the two pelagophyte samples specified as outgroup taxa and all diatom samples plus *Triparma pacifica* specified as the ingroup, allowing the final set of gene trees to be rooted with a nondiatom outgroup. We used SumTrees (Sukumaran and Holder, 2010) to collapse nodes on the final ortholog trees with less than 33% bootstrap support.

For species tree reconstructions, we filtered ortholog alignments and trees to include only those alignments with 100% taxon occupancy and alignment columns with less than 20% missing data and/or gap characters. We then reconstructed species trees using summary-coalescent and concatenation-based approaches. We used ASTRAL (ver. 4.10.8) for summary-coalescent species-tree reconstruction, with topology and support estimated with local posterior probabilities (Sayyari and Mirarab, 2016) and multilocus bootstrapping (Seo, 2008). We refer to these as ASTRAL and ASTRAL-mlbs, respectively. For the concatenation-based analysis, we used ProtTest (ver. 3.4.2) with the AICc selection criterion to determine the best-fitting model of protein evolution for each ortholog alignment (Guindon et al., 2010; Darriba et al., 2011). We used AMAS (Borowiec, 2016) to concatenate the alignments, and we used IQ-TREE with ultrafast bootstrapping and SH-aLRT testing (1000 replicates each) to infer the species tree (Guindon et al., 2010; Minh et al., 2013; Chernomor et al., 2016). We recovered relatively high levels of gene tree discordance and low levels of gene tree support across gene trees, under which conditions concatenation-based methods may outperform summary-coalescent methods (Mirarab and Warnow, 2015). As a result, we used the concatenation-based tree as the reference species tree for all subsequent analyses. Importantly, the topologies of the concatenated and summary-coalescent trees were nearly identical. Gene tree support was summarized with PhyParts (analysis=fullconcon) (Smith et al., 2015) and a companion script, phypartspiecharts.py (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts), with gene tree concordance estimated against the IQ-TREE species tree based on a 33% bootstrap support threshold.

The IQ-TREE species tree was time-calibrated with TreePL (Smith and O'Meara, 2012) and 10 fossil-derived calibration points (Appendix S2). The minimum and maximum bounds were calculated following Norris et al. (2015), except the calibration for the most recent common ancestor (MRCA) of diatoms and Parmales was constrained to a maximum age of 250 Ma. The optimal rate-smoothing parameter for TreePL was estimated using random-subsample-and-replace cross-validation with a range of tested values on a log scale between 10^{5} and 10^{-5}.

Overall approach to identification of paleopolyploidy events

Identifying WGDs from transcriptome data relies on temporal or phylogenetic signal, rather than spatial syntenic signal, and so may be impacted by historical variation in molecular evolutionary rates and saturation artifacts (McKain et al., 2016). Nevertheless, several complementary methods are now available that together can provide increased confidence in transcriptome-based WGD inferences in the absence of synteny information. These approaches are broadly divided into three categories: (1) paralog divergence (i.e., Ks-based) methods (Lynch and Conery, 2000; Blanc and Wolfe, 2004), (2) gene tree/species tree reconciliation methods (Durand et al., 2006; Jiao et al., 2011; Thomas et al., 2017), and (3) gene count methods (Rabier et al., 2014). Each of the three approaches provides incrementally more rigorous and specific tests of WGD: (1) the Ks analyses provide semiquantitative evidence for the presence of synchronously duplicated genes, (2) gene tree reconciliation pipelines identify specific branches on the species tree with elevated numbers of gene duplications and losses (Durand et al., 2006; Yang et al., 2015), and one of the reconciliation approaches used here allows for specific tests about the mechanism of inferred WGD events (auto- vs. allopolyploidy) (Thomas et al., 2017), and (3) a gene count method for detecting and locating WGD events independent of both Ks and gene tree information. As described in the following sections, we applied each of these methods to one or more sets of orthologous clusters and their corresponding gene trees (Fig. 2).

Synonymous divergence (Ks) of paralogs

We looked for evidence of WGDs in diatom and outgroup taxa using traditional approaches based on pairwise divergence between paralogs at synonymous sites (Ks) (Lynch and Conery, 2000; Blanc and Wolfe, 2004). Methods for identifying secondary Ks peaks vary in several key parameters (e.g., clustering criteria for paralogs and codon substitution model), and the behaviors of different Ks pipelines have not been systematically evaluated, so we used several different Ks pipelines and settings. We restricted these analyses to a set of relatively conserved genes, based on a BLASTP search (e-value ≤ 10^{-10}) of each transcriptome against a database of complete proteomes from 17 protist species, including

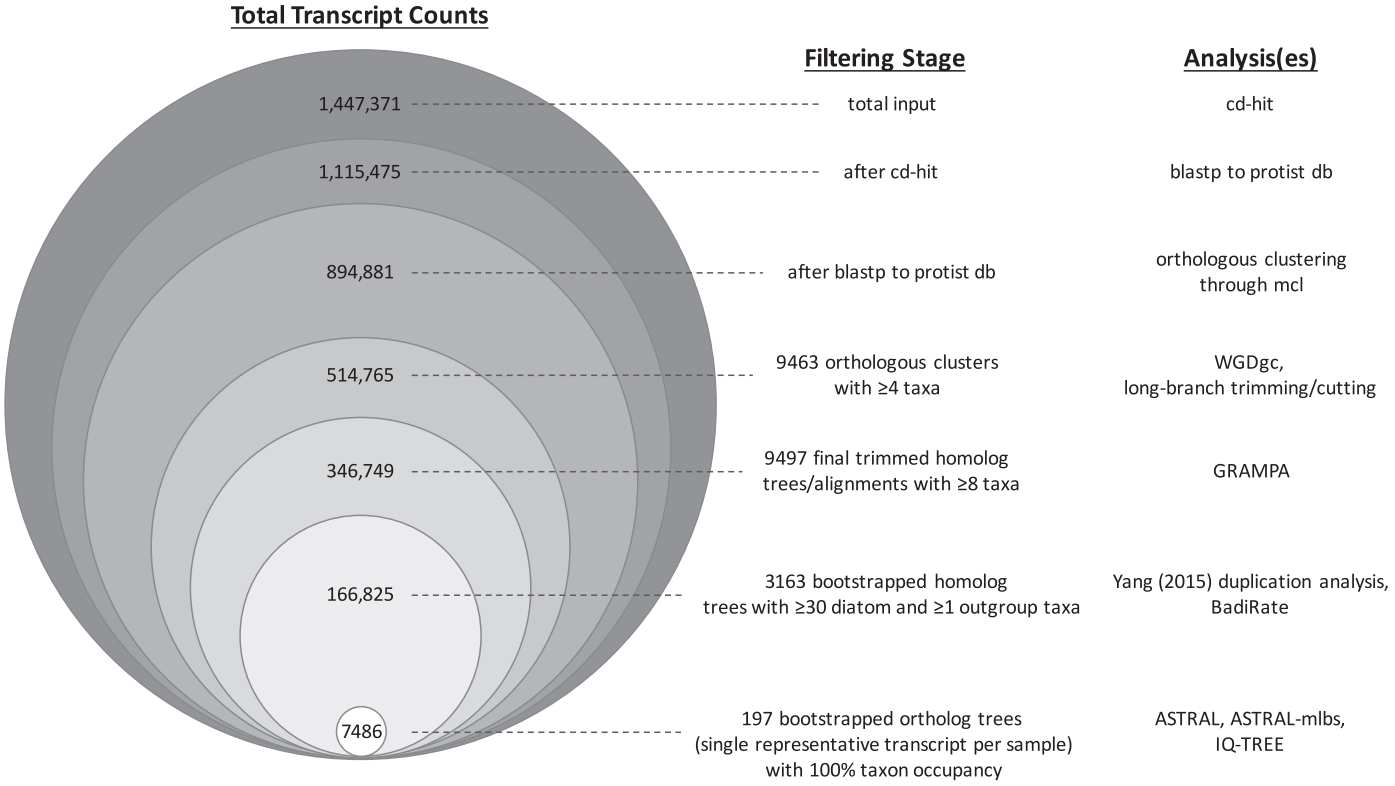

FIGURE 2. Data set sizes at critical stages of analysis. The area of each circle is proportional to the total transcript count at that stage of analysis. Total transcript counts represent all assembled transcripts (transcriptomes) and predicted genes (genomes) available from all taxa at a given stage of analysis.

the diatoms *Cyclotella nana* (formerly *Thalassiosira pseudonana*; heretofore *Cyclotella*) and *Phaeodactylum tricornutum*. The first approach followed Johnson et al. (2016), with initial filtering of each gene set to remove highly similar sequences (e.g., isoforms or very recent duplicates) using CD-HIT-EST (-c 0.98 -aS 0.90). Remaining proteins were then clustered for each species with CD-HIT (-c 0.40 -aL 0.75 -n 2), aligned with MAFFT, and back-translated by forcing nucleotide sequences to protein alignments using Pal2Nal (Suyama et al., 2006) with gap regions and internal stop codons removed. For each pair of paralogous nucleotide sequences in the CD-HIT clusters, Ks was calculated using the KaKs_Calculator (Zhang et al., 2006) under both the YN (Yang and Nielsen, 1998) and GY (Goldman and Yang, 1994) codon substitution models, hereafter referred to as JYN and JGY. We also estimated Ks distributions using the FASTKs pipeline with default settings (McKain et al., 2016). In the FASTKs pipeline, translated transcriptomes are searched against themselves with BLASTP to identify pairs of putative paralogs, which are then filtered by alignment length and percentage identity, then re-aligned and back-translated before calculating Ks. It is important to note here that the Trinity transcriptome assembler makes a distinction between closely related paralogous genes vs. isoforms of the same gene (Grabherr et al., 2011). As a result, transcript assemblies are hierarchically organized according to assembly read clusters, which are comprised of "genes" and gene "isoforms". In some cases, isoforms of the same Trinity gene might represent very recently diverged paralogs, and some Ks pipelines are "Trinity-agnostic", relying on alternative filtering strategies to distinguish paralogs and isoforms (Jiao et al., 2011; Johnson et al., 2016). Due to this ambiguity, Ks distributions were determined using the FastKs pipeline both before and after removing BLASTN self-hits at the "gene" level for the Trinity assemblies (i.e., BLASTN hits between two Trinity isoforms of the same Trinity gene). These analyses are hereafter referred to as MBA (McKain BLAST All) and MGC (McKain gene-collapsed), respectively. For both pipelines, we tested for multiple normal distributions in the Ks distributions using the R package MClust (Fraley et al., 2012), with the best fit model chosen using the Bayesian information criterion (BIC). This method applies finite mixture modeling through a semiparametric, model-based approach to estimate the probability distribution for a set of values, including for one-dimensional data sets (such as Ks estimates). We considered MClust-identified peaks that met all of the following criteria as providing strong Ks-based evidence for WGD: (1) total count of paralogous pairs within a Ks analysis ≥ 200, (2) value of Ks peak ≥ 0.05 and ≤ 2.0, and (3) ≥ 20% of all paralog pairs used in an analysis residing within an MClust-identified peak.

Gene tree reconciliation against the species tree

We used gene tree/species tree reconciliation to identify branches on the species tree with concentrations of gene duplications or losses. We filtered the RAxML gene trees to include only those with at least 30 diatoms and one outgroup, resulting in a final set of 3163 trees for this analysis (Fig. 2). Bootstrap support values across these trees were generally low, with only about 30% of nodes across all trees supported by bootstrap values >50%. This result prompted us to explore how sensitive our inferences of WGD were

to uncertainty (i.e., lack of bootstrap support) in the gene trees. First, we filtered trees based on mean bootstrap support across all nodes, retaining only those trees with mean bootstrap support ≥40% or ≥50%. Second, we performed bootstrap-based rearrangement of the homolog trees at 40%, 50%, 60%, and 70% bootstrap cutoffs. These rearrangements minimized the reconciliation cost by rearranging gene trees around poorly supported nodes. These filtering and rearranging steps resulted in a total of seven sets of homolog trees: the original set of 3163 trees, 1054 trees with mean bootstrap ≥40%, 301 trees with mean bootstrap ≥50%, and four sets of the same size as the original set but rearranged at 40%, 50%, 60%, and 70% bootstrap thresholds. Each set of gene trees was rerooted and reconciled against the species tree with Notung (ver. 2.9) (Durand et al., 2006; Darby et al., 2017). We ran Notung's phylogenomic pipeline to estimate the number of gains and losses in each gene tree and total counts of duplication and loss per node. In addition to a priori filtering or rearranging of homolog trees, we also applied the approach used by Yang et al. (2015) to the original set of 3163 homolog alignments. This pipeline maps rooted clades of gene trees, which include orthologs and paralogs, to a species tree to determine the proportion of gene families with one or more duplications at each node, taking into account confidence in gene tree topologies as measured by average bootstrap support across a subclade.

We repeated these analyses using gene trees pruned to include only pennate diatoms (Fig. 3) or only Thalassiosirales (Fig. 3), two clades of particular interest because of their high degree of species richness. We analyzed these clades separately to identify potential WGDs specific to these lineages while reducing computational time by working with smaller data sets and eliminating the signal of duplications that preceded the origin of these lineages. For each homolog tree, we extracted the complete focal clade if it was monophyletic or, if the entire focal clade was not monophyletic, we extracted all subclades composed of at least four terminal taxa. For each of these extracted tree sets, we performed the filtering, rearrangement, and reconciliation steps described above. These additional manipulations resulted in 14 additional sets of homolog trees.

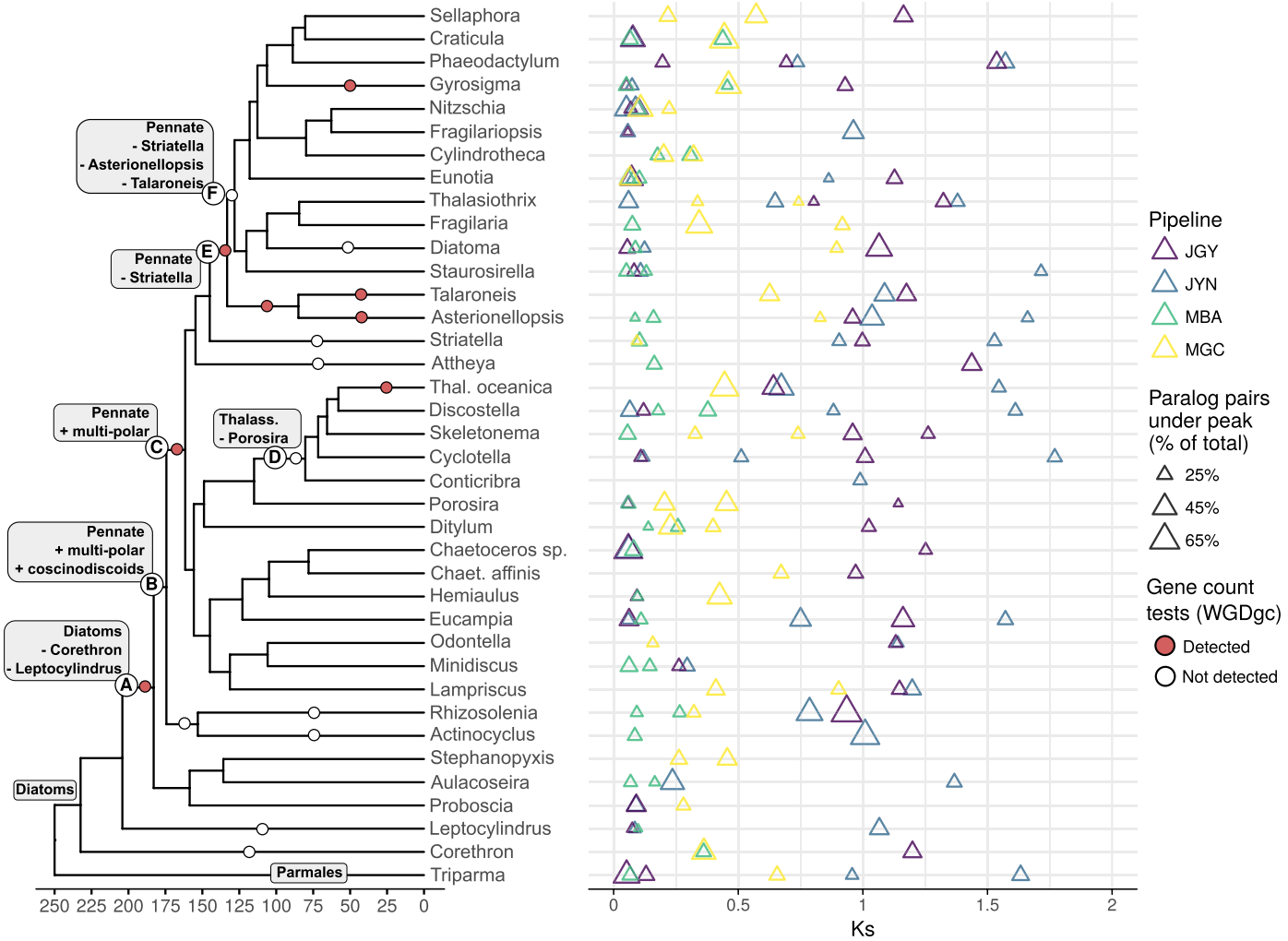

FIGURE 3. Time-calibrated species tree of 37 diatoms and the outgroup *Triparma* (*Bolidomonas*) *pacifica* (Parmales) reconstructed from a concatenated alignment of 197 pruned orthologous groups with 100% taxon occupancy and a single representative transcript per sample. Nodes relevant to downstream analyses are labeled A–F. Statistically significant secondary peaks in Ks distributions are indicated for four different Ks pipelines (see Methods). Triangle size is proportional to the fraction of paralog pairs in a Ks analysis that fall within a given Ks peak. Detailed Ks results are given in Appendices S3 and S4.

Gene tree reconciliation against multiply labeled trees

We used the software package GRAMPA (Thomas et al., 2017) to specifically test for the mechanism of WGD formation at focal nodes highlighted by the Yang and Notung pipelines. GRAMPA compares the reconciliation scores of multiply labeled (MUL) trees that correspond to allo- or autopolyploid scenarios against the singly labeled species tree. Cases in which the MUL tree—a topology in which a taxon or clade appears twice as the result of a duplication—had a better reconciliation score than the species tree were considered supportive of a WGD event. By default, GRAMPA performs least-common ancestor (LCA) reconciliation of all gene trees against both the species tree and all possible MUL trees, and reports the number of duplications and losses as well as their sum (the reconciliation score). Overly complex gene trees, which might take a prohibitively long time to reconcile, are filtered out based on a maximum allowed number of polyploid groups, which we left to its default value (the group cap setting, default = 8).

We performed two types of GRAMPA searches: (1) scoring all possible MUL trees, and (2) scoring only the MUL trees relevant for nodes flagged by the Notung and Yang pipelines as having large concentrations of gene duplications. Both approaches tested all relevant arrangements for the two parents of a putative allopolyploid event, including the same parent for autopolyploid events, and compared their reconciliation scores to the reconciliation score of the singly labeled species tree. As with the Notung analyses, we ran separate GRAMPA searches at the level of all diatoms, within Thalassiosirales, and within pennate diatoms, testing the robustness of inferred WGDs to bootstrap support by following the filtering and rearrangement strategies detailed above.

Gene count analyses

We used gene count data derived from the 3163 bootstrapped trees (3.1K) and a broader set of 9497 homolog trees with at least eight diatoms (9.5K) (Fig. 2) to test a number of WGDs supported by Ks or gene tree reconciliation analyses with the R package WGDgc (Rabier et al., 2014). The computational demands of these analyses restricted the number of branches we were able to test, so we focused on 11 terminal and seven internal branches, chosen because they represent groups of longstanding interest (e.g., Thalassiosirales) or to test whether significant Ks peaks in closely related species represented multiple independent or single shared events. Initial tests used the entire species phylogeny and required an orthologous cluster to include *Triparma pacifica* and at least one ingroup species, thereby removing orthologous clusters unique to diatoms. With this strategy, most of the putative WGD events identified through Ks analyses were not detectable, likely due to excessively stringent filtering to meet the above criterion. Similar results have been observed in other studies that use gene count data, and one common solution is to focus analyses on subtrees that maximize the amount of data relevant to testing a particular WGD hypothesis (Tiley et al., 2016). To increase the pool of orthologous clusters for detection of WGDs while keeping computational memory and time reasonable, we created reduced gene count data sets and pruned accordingly the time-calibrated chronogram to include only those taxa relevant to a specific WGD hypothesis. For example, when testing for a putative Ks-inferred WGD in *Gyrosigma*, we pruned the species tree down to include raphid pennates only (Fig. 3). The final data sets represented orthologous clusters with representation in the outgroup and at least one species of the ingroup. WGDgc analyses were run with the root prior set to the mean number of copies per cluster in each of the data sets and with the option "oneInBothClades" that reflected our filtering strategy. The putative WGD events were assumed to have occurred at the midpoint of branches leading to the focal node. Hypotheses were tested using likelihood ratio tests against a null model of no WGDs (Rabier et al., 2014; Tiley et al., 2016).

RESULTS

Transcriptome assemblies

We assembled transcriptomes for 34 diatom taxa and one outgroup (*Triparma pacifica*) using paired-end RNA-seq read pools that ranged in size from 21.3 to 424 million reads. Trinity assemblies ranged in size from 13,578 to 61,091 genes and 16,145 to 70,488 transcripts (including isoforms). BUSCO recovery in assembled transcriptomes averaged 70 ± 8% for combined complete and fragmented orthologs. Gene counts for protein sets from the five genome sequences ranged from 10,402 to 27,137 genes, with a corresponding average BUSCO recovery of 83 ± 6%. Sample information and assembly details are available in Appendix S1.

Homology and orthology inference

A total of 9463 orthologous clusters containing at least four taxa were circumscribed with MCL (Fig. 2). These clusters were pruned and subdivided into a total of 9497 ortholog alignments and corresponding phylogenetic trees, each with at least eight taxa. These alignments were then filtered based on various taxon-occupancy thresholds to create data subsets for further analyses (Fig. 2).

Species tree reconstruction

A total of 197 ortholog alignments with a single representative transcript per sample and with 100% taxon occupancy were recovered (Fig. 2), representing a combined alignment length of 58,294 amino acids. Coalescent-summary (ASTRAL, ASTRAL-mlbs) and concatenation-based (IQ-TREE) inference methods recovered well-supported species trees with identical branching orders, with the exception of the polar centric diatom *Ditylum brightwellii* (Fig. 3, Appendix S3), which has been difficult to place with phylogenomic data (Parks et al., 2018). Similar to previous findings (Parks et al., 2018), relationships among the major multipolar centric clades were the most difficult to resolve, with deep splits supported by few or no gene trees (Appendix S3).

Synonymous divergence (Ks) between paralogs

Ks-based age distributions of gene duplicates revealed strongly supported secondary Ks peaks, typically interpreted as evidence of paleopolyploid events, in all diatom species in at least one of the four Ks analyses (Fig. 3, Appendices S4 and S5). The numbers, sizes, and placements (ages) of secondary Ks peaks varied considerably by gene family clustering algorithm, treatment of isoforms in the transcriptome assemblies, and model of sequence evolution for calculating synonymous distances (Fig. 3). Strongly supported peaks in the JYN and JGY analyses had significantly higher Ks values than in the MBA and MGC analyses, and MGC analyses identified peaks

TABLE 1. Summary of strongly supported secondary peaks in Ks-based analyses of whole-genome duplication in diatoms. The average numbers of peaks per sample were not significantly different in paired *t*-tests at $p < 0.05$ ($0.138 < |t| < 1.357$). Significant differences in the average Ks peak value per species are indicated by contrasting superscript letters based on standard *t*-tests at $p < 0.05$. Degrees of freedom for average Ks value/sample *t*-tests: JYN = 44; JGY = 42; MCA = 34; MCG = 33; *t*-values for significant differences were $2.528 < |t| < 6.063$; *t*-value for JYN vs. JGY = 0.501.

Ks analysis	Total number of peaks	Average peaks/ sample (SD)	Average Ks value/ sample (SD)
JYN	49	1.216 (0.976)	0.728[a] (0.594)
JGY	47	1.162 (0.727)	0.668[a] (0.518)
MBA	37	0.946 (0.815)	0.152[b] (0.113)
MCG	36	0.919 (0.795)	0.425[c] (0.241)

with significantly higher mean Ks values than in the MBA analyses (Table 1, Fig. 3).

Gene tree reconciliation (Yang and Notung pipelines)

Notung and Yang pipeline reconciliation with the original set of 3163 unfiltered and unrearranged gene trees identified six branches along the backbone of the species tree with high concentrations of gene duplications (40–70% of gene families; Figs. 3 and 4, branches A–F). Average bootstrap support of these gene trees was relatively low, so a potentially large fraction of these duplications may have been inferred from poorly supported trees or nodes. To assess the sensitivity of these results to bootstrap support, we repeated the Notung reconciliation by filtering for trees with at least 40% or 50% average bootstrap support or by rearranging the original set of gene trees at bootstrap thresholds ranging from 40–70%. Both of these strategies reduced the number of inferred gene duplications, with rearrangements having a much stronger effect (Fig. 4). Gene-tree filtering generally resulted in a 10–20% drop in the percentage of inferred duplications, though the percentage of duplicated genes did not change with bootstrap-based filtering for branches C and E (Fig. 4). For rearrangements, the percentage of duplications decreased more or less linearly with increasing bootstrap threshold, ultimately resulting in a 60–80% drop in inferred gene duplications with a rearrangement threshold of 70% (Fig. 4). Rearrangement results indicated that the large duplication fractions at focal nodes were influenced to a large degree by gene trees with poor support; however, filtering for homolog alignments that contained more phylogenetic signal and produced better supported trees (mean bootstrap ≥40% or ≥50%) still retained substantial signal for large-scale duplication events at these nodes. Overall, even with the most conservative strategy (rearrangements with 70% bootstrap threshold), we found that ≥19% of gene families experienced duplications at the top three nodes (A, D, and C), suggesting that signal for synchronous duplications at these branches could still be detected despite the generally poor bootstrap support within gene trees (Figs. 3 and 4, branches A, C, and D; Appendix S6, see the Supplemental Data with this article). Results from the Yang pipeline with a bootstrap cutoff of 40% corroborated these findings (Fig. 4).

Gene tree reconciliation (GRAMPA)

The Notung and Yang pipelines identified four nodes with large concentrations of duplications, consistent with ancient WGD events: (1) the MRCA of all diatoms excluding *Corethron hystrix* and *Leptocylindrus danicus* (Fig. 3, branch A), (2) the MRCA of pennate+multipolar centric diatoms (Fig. 3, branch C), (3) the MRCA of Thalassiosirales excluding *Porosira pseudodenticulata* (Fig. 3, branch D), and (4) the MRCA of all pennate diatoms excluding *Striatella unipunctata* (Fig. 3, branch E) (Fig. 4). We used GRAMPA to further investigate these putative WGDs with the original, filtered, and rearranged sets of gene trees. As GRAMPA reconciles gene trees against multiply labeled trees representing different scenarios for the parental lineages of a WGD event, these analyses provided additional tests of whether our tree sets showed signs of polyploidy and, if so, whether the more likely reconstruction pointed to allo- or autopolyploid ancestry.

Whole-genome duplication at deep internodes (Figs. 3 and 4, branch A/C)

The Notung/Yang reconciliations against the singly labeled species tree highlighted concentrations of families with duplicated genes at the MRCA of all diatoms excluding *Leptocylindrus* and *Corethron* (Figs. 3 and 4, branch A) as well as the MRCA of pennate+multipolar diatoms (Figs. 3 and 4, branch C). In unconstrained GRAMPA searches, MUL trees involving these nodes were better than the species tree (Fig. 5), though the optimal topology implied a different and more complex scenario. Namely, in unconstrained GRAMPA searches, the lowest-scoring MUL tree pinpointed the WGD to an intermediate branch (Figs. 3 and 4, branch B). The WGD was inferred to have been an allopolyploid event involving an unsampled or extinct lineage sister to branch B (Fig. 3) and an unsampled or extinct lineage sister to all diatoms. Searches using sets of trees with mean bootstrap support ≥40% ($N = 1173$) or ≥50% ($N = 348$) gave similar results, with all three of these scenarios (WGD at branches A, B or C; Figs. 3 and 4) among the top five MUL trees, and in each case, the second parental lineage was an unsampled or extinct lineage sister to all diatoms (Fig. 5). In addition to these possibilities, a MUL tree involving the clade of multipolar diatoms excluding *Attheya*, with *Attheya* as the second parental lineage of an allopolyploid event, was also among the best-supported scenarios. However, in unconstrained GRAMPA searches using the sets of trees rearranged at bootstrap thresholds as low as 40%, the singly labeled species tree (i.e., no inferred WGDs) received the lowest reconciliation score. Overall, our analyses of unaltered and filtered sets of gene trees supported at least one deep allopolyploid event, with an uncertain location between branches A, B, and C or possibly at the MRCA of multipolar diatoms excluding *Attheya* (Figs. 3–5; Appendix S6).

Whole-genome duplication within Thalassiosirales (Figs. 3 and 4, branch D/D′)

Multiple lines of evidence supported one or more WGDs in Thalassiosirales, specifically involving a clade composed of all sampled members of Thalassiosirales except *Porosira pseudodenticulata* (heretofore *Porosira*) (Fig. 6, branch D), a less inclusive clade consisting of *Skeletonema marinoi* (heretofore *Skeletonema*), *Discostella pseudostelligera* (heretofore *Discostella*), and *Thalassiosira oceanica* (heretofore *Thalassiosira,* not to be confused with *Thalassiosira pseudonana*, referred to here as *Cyclotella*, or *Thalassiosira weissflogii*, referred to here as *Conticribra*) (Fig. 6, branch D′), and *Thalassiosira* alone (Fig. 6). We tested these putative events with GRAMPA using extracted Thalassiosirales subtrees with four or

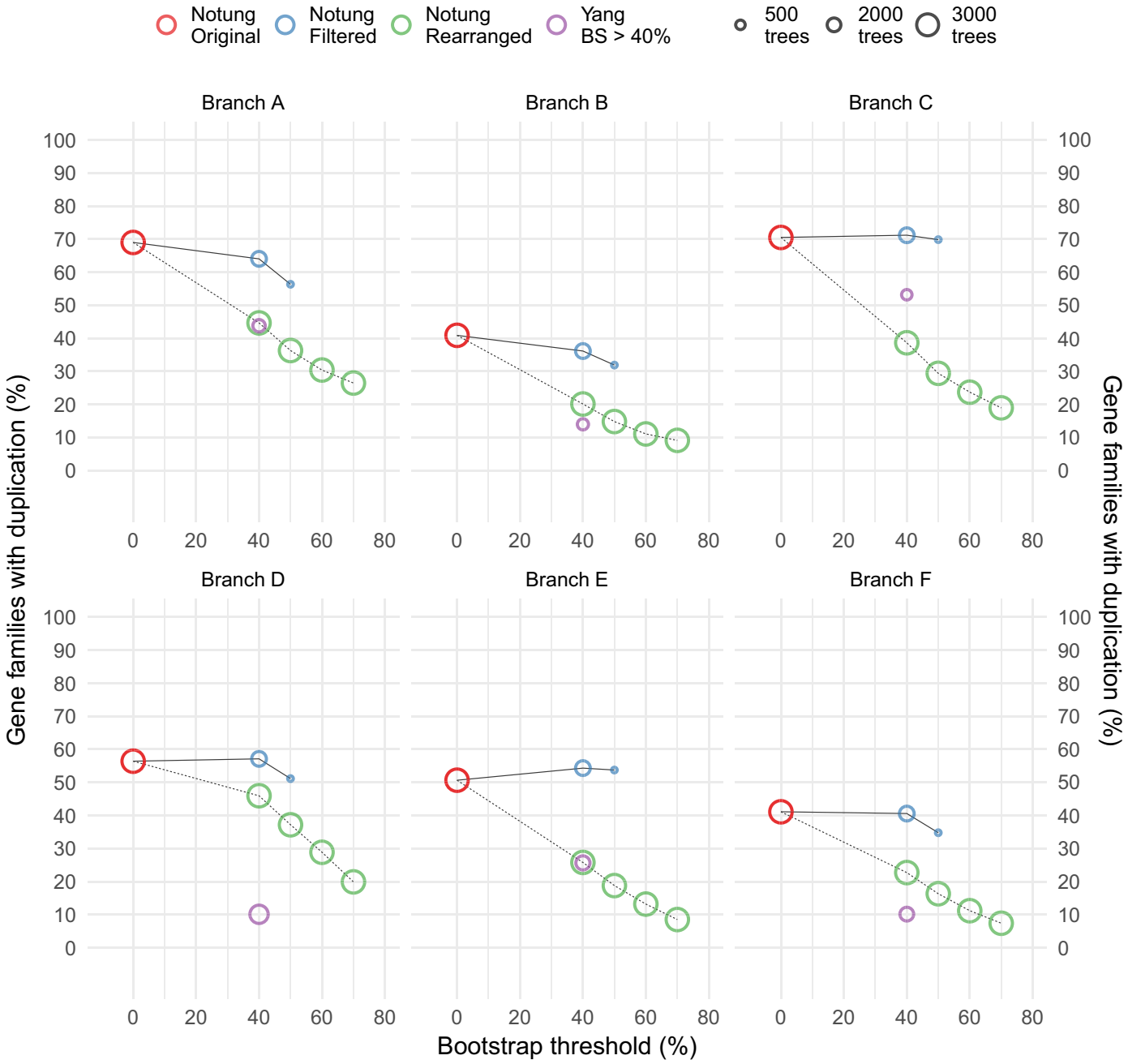

FIGURE 4. Trends in gene duplication and loss across six select branches of the species tree (see Fig. 3 for key to branch names). Percentages of gene duplication and loss were reconstructed with different sets of homolog trees, different reconciliation pipelines, and at different bootstrap thresholds for filtering and rearrangement.

more species ($N = 3259$). An unconstrained GRAMPA search found that MUL trees defined by the branch representing Thalassiosirales excluding *Porosira* (Figs. 3 and 4, branch D) were never better than the species tree, irrespective of the set of trees used and whether or not the trees were rearranged. The Notung and Yang pipelines, by contrast, found the highest concentration of duplications within Thalassiosirales at this node (Fig. 4). Instead, the strongest support for WGD within the Thalassiosirales was for a clade comprising *Skeletonema*, *Discostella*, and *Thalassiosira* (Figs. 2 and 4, branch D′). The best MUL tree corresponded to an allopolyploidy event between an extinct or unsampled lineage represented by the MRCA of these three species and an extinct or unsampled lineage represented by the MRCA of all Thalassiosirales excluding *Porosira* (Fig. 5). This event was robust to bootstrap support, being detectable with trees filtered up to mean bootstrap ≥70% ($N = 745$) and with bootstrap-based rearrangement up to a threshold of 50%. In a GRAMPA run with the unfiltered and unrearranged gene trees, seven of the 10 MUL trees better than the species tree included lineages nested within the (*Skeletonema*, (*Thalassiosira*, *Discostella*)) clade, including one MUL tree for each of *Thalassiosira* and

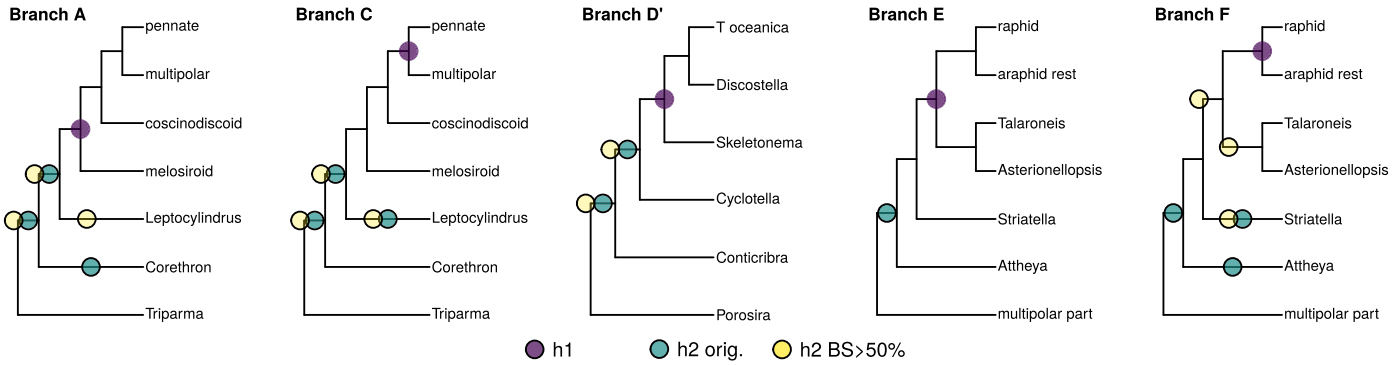

FIGURE 5. GRAMPA reconciliation for several putative whole-genome duplications and at different bootstrap filtering thresholds. Branch names correspond to lineages where the Notung and Yang pipelines detected a high concentration of duplications and match the notations used in Figs. 3 and 4. For each branch, we highlight the first parental lineage of the allopolyploid event (h1, purple circles) and the top three possible positions for attachment of the second parental lineage (h2, green and yellow circles). Green circles denote the results from the analysis of the unaltered trees ($N = 3163$), whereas yellow circles denote the analysis of trees with mean bootstrap support ≥50% ($N = 348$). The species tree was simplified to include only those branches relevant to each event. MUL trees representing autopolyploid events always scored worse than the allopolyploid alternatives and the singly labeled species tree (no polyploidy).

Discostella independently and for these two species combined. In all cases, the second parental lineage was represented by the MRCA of Thalassiosirales except *Porosira*, pointing to a discrepancy between the age of the polyploid lineage and the concentration of duplications mapped onto the singly labeled species tree (Figs. 3–6). Similar findings in yeast have been interpreted as strong support for ancient hybridization and allopolyploidy (Marcet-Houben and Gabaldon, 2015; Thomas et al., 2017). The discrepancy between the relative age of the reconciliation-inferred duplication peak (Fig. 6, branch D) and the polyploid clade (Fig. 6, branch D′) is therefore likely due to the earlier divergence of the hybridization-derived homeologs present in the genome of the polyploid lineage, which trace back to the earlier branch D, compared to the age of the polyploid lineage itself, which traces back to the younger branch D′ (Figs. 3 and 6).

Whole-genome duplication within the pennate clade (Figs. 3 and 4, branch E/F)

Within the clade of pennate diatoms (Fig. 3), Notung and Yang reconciliations against the singly labeled species tree detected the highest concentrations of duplications at the nodes representing the MRCA of all pennates except *Striatella* (Fig. 4, branch E) and the MRCA of all pennates except *Striatella*, *Asterionellopsis*, and *Talaroneis* (Fig. 4, branch F). GRAMPA searches to further assess support for these putative WGDs used extracted subtrees of pennate diatoms ($N = 2156$) that were either unaltered, filtered, or rearranged. In unconstrained GRAMPA searches, MUL trees corresponding to these nodes were weakly supported (ranked 6th and 7th out of seven MUL trees better than the species tree), and instead, the best reconciliation scores were obtained for smaller clades nested within the clades descendant from branches E or F (Figs. 3 and 6). With the full set of unaltered subtrees, the best MUL tree involved the clade (*Phaeodactylum*, (*Sellaphora*, *Craticula*)), whereas sets of trees filtered at bootstrap cutoffs of 40% ($N = 1206$), 50% ($N = 798$), and 70% ($N = 187$) supported a MUL tree formed by the clade of all raphid pennate diatoms except *Eunotia* (Fig. 3). Using the rearranged trees, GRAMPA found no MUL trees better than the species tree. By constraining our searches to MUL trees involving branch F (Fig. 3), for which GRAMPA's filter retained and reconciled larger sets of gene trees, we were able to detect an allopolyploid event in which the second parental lineage was an extinct or unsampled lineage represented by either the MRCA of all pennates except *Striatella*, or the MRCA of all pennate diatoms, though these reconstructions varied across sets of trees and were not detected with rearranged trees (Fig. 5; Appendix S6).

Gene count analyses

Analyses based on gene counts were designed to test 18 putative WGD events inferred either from analyses of Ks divergence (10 terminal and two internal branches; Fig. 3) or gene tree reconciliation (six internal branches) (Fig. 3). Each putative event was tested independently through comparison to a null, non-WGD model. We performed these analyses using gene counts based on the 3.1K and 9.5K data sets and, with three exceptions described below, recovered the same set of results for both analyses. We detected WGDs in eight of 18 tested branches (Fig. 6), with relatively low rates of homolog retention, whereby the retention rate (q) is defined as the probability of retaining the WGD-derived copy of a gene (Rabier et al., 2014). Retention of two WGD-derived homologs following a duplication event was generally less than 2%, although four WGDs had retention rates between 3% and nearly 15%. All eight tests that returned support for WGD with retention rates >0 were significantly better than their corresponding no-WGD null models (likelihood ratio tests, df = 1, χ^2 $p \leq 0.001$ for all tests; Appendix S7).

Within pennate diatoms, gene count analyses corroborated the Ks-inferred WGDs in *Gyrosigma* ($q = 14.8\%$), *Asterionellopsis* ($q = 7.0\%$), and *Talaroneis* ($q = 1.4\%$) (Fig. 6). There was also signal for WGD along the branch leading to the MRCA of *Asterionellopsis* and *Talaroneis* ($q = 3.6\%$), suggesting that the Ks peaks observed in these two taxa might represent a single shared WGD (Figs. 3 and 5). Finally, in agreement with gene tree reconciliations, there was also signal for WGD along the branch leading to the MRCA of all pennate diatoms excluding *Striatella* (Fig. 3, branch B; $q = 1.1\%$) (Fig. 6). The latter event was not detected with gene counts from the 3.1K data set, which instead detected signal for WGD in *Attheya* alone ($q = 0.3\%$). WGD along the branch leading to *Attheya* was not observed in the analysis of counts calculated from the 9.5K data set.

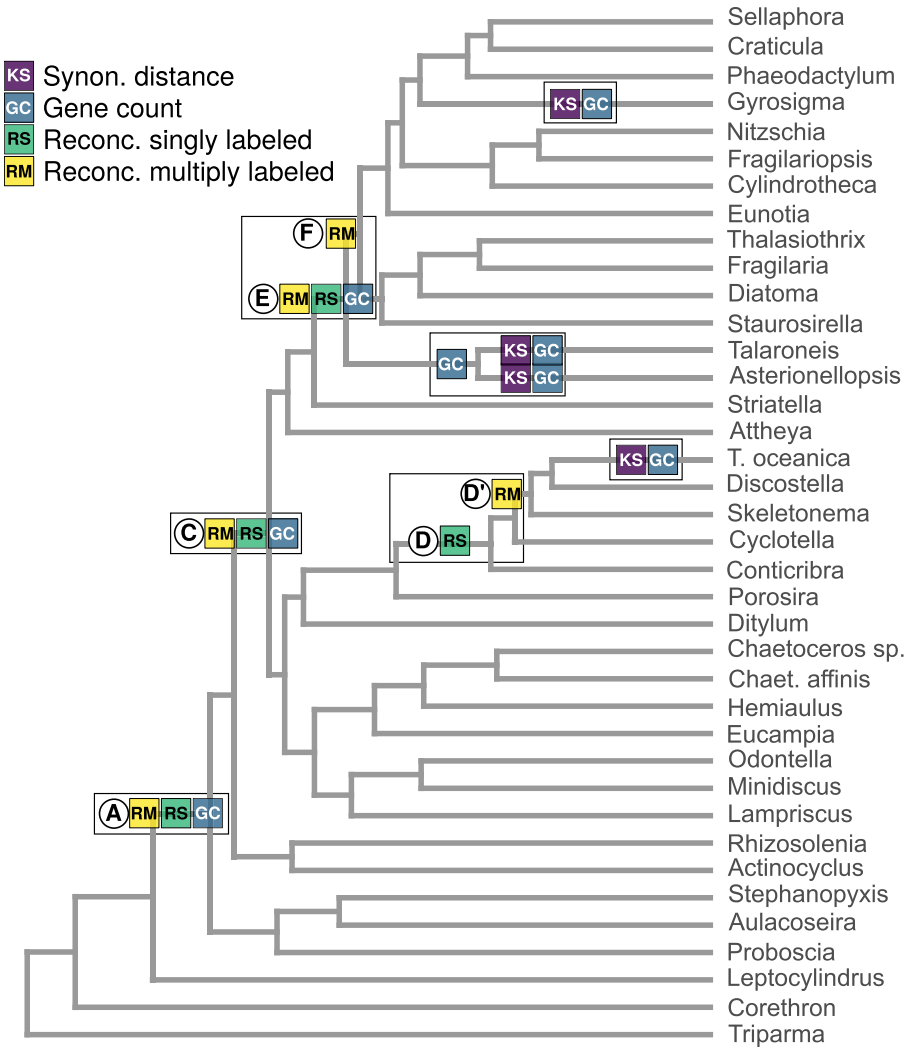

FIGURE 6. Summary of whole-genome duplications (WGDs) across diatoms. Purple: WGDs supported by Ks-based age distributions of duplicated genes (KS). Blue: WGDs detected with gene count data (GC). Green: WGDs inferred by reconciliation of gene trees against the singly labeled species tree (RS) (includes both Yang and Notung results; for details, see Appendix S6). Yellow: WGDs inferred by reconciliation of gene trees against multiply labeled trees (RM). Branches discussed in the text are labeled A–F, as shown here and in Fig. 3. Within Thalassiosirales, branch D' denotes a GRAMPA-inferred allopolyploid clade (RM) that corresponds to the duplication peak inferred from the Notung and Yang analyses (branch D). Within pennate diatoms, GRAMPA-inferred events are added to both branches E and F to reflect uncertainty in the placement of the WGD.

We used WGDgc to test for polyploid events at several nodes across the centric and multipolar centric diatoms. We tested two distinct hypotheses within Thalassiosirales: a prominent, Ks-inferred WGD in *Thalassiosira* and the GRAMPA-inferred WGD at the MRCA of *Thalassiosira*, *Skeletonema*, and *Discostella*. We detected signal for an event within the *Thalassiosira* lineage (q = 1.2%) but found no evidence for the older WGD (Fig. 6). Despite distinct Ks peaks, we did not detect WGD events in *Actinocyclus*, *Rhizosolenia*, or their MRCA, nor did we find support for the WGD events implied by secondary Ks peaks in *Corethron* and *Leptocylindrus*. Finally, we tested for two events supported by both reconciliation and GRAMPA results at the MRCA of pennate+multipolar diatoms (Figs. 3 and 5, branch C) and the MRCA of all diatoms excluding *Corethron* and *Leptocylindrus* (Figs. 3 and 5, branch A). The gene count analysis detected WGDs on both branches, including a WGD with retention rate q = 1.7% along the branch leading to pennate+multipolar diatoms using both data sets, as well as an event on the branch leading to the MRCA of all diatoms excluding *Corethron* and *Leptocylindrus* using the 9.5K data set (q = 6.4%) (Fig. 6).

DISCUSSION

Substantial variation in genome size and chromosome number, a high rate of genome size evolution, and direct observations of polyploidization in cell cultures together suggest that WGD may have been common throughout the course of diatom evolution (Mann, 1994, 1999b; Oliver et al., 2007; von Dassow et al., 2008; Koester et al., 2010; Whittaker et al., 2012). Our survey of 37 diatom genomes and transcriptomes provided strong support for this hypothesis, identifying seven WGDs supported by multiple lines of evidence and many more suggested by synonymous divergence of paralogs (Figs. 3 and 6). Our coarse taxon sampling precluded precise pinpointing of the timing of these events, with two strongly supported events assigned to terminal branches that represent ca. 60–100 Myr of evolutionary history (Fig. 6). Nevertheless, despite coarse taxon sampling and the general challenges of working with a group of non-model organisms, our analyses point to an extensive history of WGD in diatoms.

Combined genomic evidence for whole-genome duplication in diatoms

We applied three strategies to characterize the history of WGD across diatoms: (1) traditional Ks-based age distributions of duplicated genes, (2) phylogenetically based reconciliation methods to identify nodes on the species tree with concentrations of gene duplications and to construct specific tests for allopolyploidy, and (3) gene count methods, which provide conservative inferences of WGD that are agnostic to information in the gene sequences and gene trees. Although each of these approaches suffers some drawbacks, we considered a putative WGD as strongly supported when results from two or more of these three very different approaches were in agreement (Fig. 6).

The Ks distributions showed extensive signal for WGD in our analyses, usually identifying multiple significant Ks peaks in every species in our analysis (Fig. 3). Although Ks distributions are useful for initial explorations of duplication signal in a genome, several general limitations of this approach led us to rely much more heavily on gene tree reconciliations and gene count analyses. For

example, it can be challenging, even with statistical tests, to discern real duplication peaks in a Ks distribution. In some cases, peaks are identified essentially by eye (Blanc and Wolfe, 2004; Fawcett et al., 2009; Tang et al., 2010; Cannon et al., 2015), which easily becomes an exercise in reading tea leaves. Although several statistical approaches have been adopted to identify discrete shifts or peaks in Ks distributions (Schlueter et al., 2004; Cui et al., 2006; Vanneste et al., 2015), it still can be challenging to distinguish WGD-derived peaks from stochastic variation in the background rate of gene duplication. In particular, component-selection strategies such as BIC are prone to overfitting (Naik et al., 2007; Vekemans et al., 2012), though complementary statistical tests may ameliorate some of these issues (Barker et al., 2008; Vekemans et al., 2012). Ks-based inferences of ancient duplications can also be confounded by saturation at synonymous sites (Vanneste et al., 2013), which should be more pronounced in lineages with high substitution rates, such as diatoms (Bowler et al., 2008). In these cases, Ks-based age distributions will saturate sooner, erasing the signature of ancient WGDs and potentially creating false signal for more recent WGDs (Vanneste et al., 2013). On average, 45% of the paralog pairs in a given diatom species had Ks values that are considered out of range (>2) for drawing Ks-based inferences of WGD (Vanneste et al., 2013).

Finally, Ks distributions are generally calculated from just one of many different available software packages that vary considerably in how synonymous distances are calculated and gene families are clustered. To the best of our knowledge, this study is one of the first to demonstrate just how widely Ks distributions—and the biological inferences made from them—can vary based on these two fundamental but largely unexplored aspects of Ks-based inference of WGD. We found that BLAST-based clustering (MBA and MGC) identified younger WGD events, whereas CD-HIT clustering (JYN and JGY) tended to identify older ones (Fig. 3), likely as a result of creating larger "gene family" clusters with more divergent sequences. Of course, parameters of the two clustering algorithms could be adjusted to more closely align the size and composition of the resulting clusters, but this still does not address how the parameters should be set for Ks analyses. Differences among the various pipelines used here were evident both in broadscale trends across species as well as within individual genomes (Fig. 3). For example, a single strongly supported Ks peak was identified in *Attheya septentrionalis* by JGY, MBA, and MGC analyses, but the mean Ks values indicated that different analyses either were identifying different events or were assigning vastly different age estimates to the same event, with mean Ks values of 1.46 vs. 0.16/0.18 for JGY and MBA/MGC, respectively (Fig. 3). It was not clear how best to resolve these and many more such disparities (e.g., *Cylindrotheca* J vs. M pipelines, *Triparma* J vs. M and JGY vs. JYN pipelines).

Gene-tree reconciliation methods allow for more precise timing of WGDs (Durand et al., 2006; Jiao et al., 2011), naturally accommodate uncertainty in the data, and even allow for specific tests of auto- vs. allopolyploidy (Thomas et al., 2017). The power of these approaches is limited, however, by the quality of the gene trees, which were generally poorly supported in our data sets. For the gene trees in our 3.1K data set, the overall distribution of bootstrap values across all nodes and trees was relatively low (median bootstrap support = 29). A total of 68% and 80% of the nodes across gene trees had bootstrap support lower than 50% and 70%, respectively. Bootstrap-based rearrangement of gene trees to minimize the numbers of inferred duplications and losses is a common strategy for guarding against false inferences of WGD from poorly supported gene trees (Durand et al., 2006; Inoue et al., 2015; Thomas et al., 2017). All of our gene trees, and a majority of nodes within our trees, had the potential to be rearranged. Deciding on a bootstrap threshold on which to base our inferences, therefore, depended on our confidence in the correct reconstruction of the gene trees. The inference of WGDs ideally should be based on strongly supported nodes that, when reconciled against the species tree, identify duplication events (Hahn, 2007). However, the amount of data necessary to obtain strong support for nodes depends on tree shape and the distribution of internal branch lengths (Alfaro et al., 2003; Hahn, 2007; Philippe et al., 2011). Short internal branches can require substantial amounts of data to obtain strong bootstrap support, but the number of phylogenetically informative characters within an individual gene tree is clearly limited. Simulation studies have shown that even correct nodes can receive low bootstrap values under a variety of conditions (Alfaro et al., 2003). These considerations highlight the difficulties in determining an empirical cutoff for what should be considered an accurate bipartition and, by extension, the bootstrap threshold for gene tree rearrangement for reconciliation analyses.

Empirical and simulation studies have shown that gene count analyses provide another powerful means of identifying WGDs (Rabier et al., 2014; Tiley et al., 2016). The methods are conservative, however, and may fail to discern nested WGDs or WGDs followed by high rates of gene loss (Hahn, 2007; Tiley et al., 2016). In this regard, it is difficult to determine the number and sequence of events underlying the Ks peaks in, for example, *Talaroneis* and *Asterionellopsis*, which are sister taxa in our species tree (Figs. 3 and 6). Although our gene count analyses did not test for multiple duplication events in these or any other terminal taxa, gene count and Notung analyses highlighted very low overall rates of duplicate gene retention (alternatively, high rates of gene loss) across diatoms. These low retention rates likely limited the power of WGDgc to identify WGDs in our dataset and suggest, more broadly, that high rates of molecular and genome evolution in diatoms might rapidly mask the signal of historic duplications and lead to underestimation of the number of duplication events. On the other hand, high rates of gene loss following duplication, coupled with the ignorance of WGDgc to gene tree topology and potential asymmetric gene loss, together increase confidence in WGDs inferred from gene count data. Denser taxon sampling should increase the chance of detecting duplications and gene losses in descendent lineages and, as a result, possibly provide gene count support for some of the putative terminal duplications identified in the Ks analyses.

Finally, to better contextualize the support for the WGD events identified here, WGDs identified in lineages with many more genomic resources (e.g., flowering plants [Tang et al., 2010; Amborella Genome Project, 2013; Jiao et al., 2014] and fungi [Kellis et al., 2004]) are similarly supported by multiple lines of evidence, including synteny analyses of sequenced genomes. Lack of collinearity in two distantly related diatoms (*C. nana* and *P. tricornutum*) suggested that the limited duplication signal in those species was due to small segmental, rather than whole genome, duplications (Bowler et al., 2008). In our Ks analyses, we only recovered strong WGD peaks for these two species with the JYN and JGY pipelines (Fig. 3). Moreover, the total number of duplicated genes in these genomes was low, with a maximum of just 404 paralog pairs identified out of a total of >10,000 genes in both cases (Appendix S5). Lack of historic WGD signal in these species could, therefore, be due to low retention rates of duplicated genes in these two compact (<30 Mb)

genomes. Nevertheless, a fuller understanding of paleopolyploidy in diatoms will benefit from an increased number and diversity of high-quality genome sequences (Kellogg, 2016), though high rates of nucleotide substitution, gene loss, and genome rearrangement will likely prove to be persistent challenges.

Paleopolyploidy in diatoms

We identified two deep WGDs occurring roughly 200 and 170 Ma (Fig. 6, branches A and C) that resulted in ancient polyploid ancestry for the vast majority of diatom diversity. Although both events were supported by multiple approaches (Fig. 6), limitations stemming from our coarse taxon sampling made it difficult to pinpoint the timing of these events. For example, the WGDs inferred at branches A and C were supported by gene tree reconciliation (Fig. 4) and gene count data (Appendix S7), and reconciliation against multiply labeled trees clearly supported an allopolyploid mode of origin for both events. In both cases, the second parental lineage of the allopolyploid event was an extinct or unsampled lineage vaguely identified as the MRCA of all diatoms or all diatoms excluding *Corethron* (Fig. 5). Although it is possible (or perhaps likely) that these ancestors are extinct, the fact that only two branches separate the older of these events and the diatom stem lineage leaves open the possibility that our sampling is too coarse for a precise determination of the lineages involved in this allopolyploid event.

Whole-genome duplication in Thalassiosirales

Thalassiosirales is among the most common and abundant diatom lineages in the plankton of both marine and freshwaters. It is also a long-established, genome-enabled model system for studies of diatom physiology, morphology, and ecology (Guillard and Ryther, 1962; Armbrust et al., 2004; Poulsen and Kroger, 2004; Alverson et al., 2007), and previous studies have shown substantial variation in cellular DNA content among species (von Dassow et al., 2008; Whittaker et al., 2012). The discovery of ancient hybridization and allopolyploidy in this group further establishes it as an excellent system for understanding genome-scale evolutionary processes in diatoms.

The signal for polyploidy within Thalassiosirales was the strongest recovered in our analyses, being detectable even after applying a relatively stringent (given our gene trees) 50% bootstrap rearrangement threshold. Gene-tree reconciliation supported an allopolyploid event involving the clade comprised of *Thalassiosira*, *Skeletonema*, and *Discostella* (Fig. 6, branch D′). Uncertainty in the species tree, however, makes it difficult to accurately circumscribe this event. Although most nodes within the Thalassiosirales species tree were well supported, gene tree discordance was especially high for splits within the putatively polyploid subtree, (*Skeletonema*, (*Thalassiosira*, *Discostella*)) (Appendix S3). Interestingly, the two nodes immediately predating this clade (the MRCA of Thalassiosirales minus *Porosira* and the MRCA of Thalassiosirales) had many more concordant gene trees, suggesting that the high levels of discordance in the (*Skeletonema*, (*Thalassiosira*, *Discostella*)) clade may reflect, at least in part, conflict resulting from past hybridization. Densely sampled phylogenies of Thalassiosirales inferred from ribosomal RNA and chloroplast genes have produced an alternative topology, ((*Thalassiosira*, *Skeletonema*), *Discostella*), with *Thalassiosira* sister to *Skeletonema* (Alverson et al., 2007). If this relationship is correct, the strong secondary Ks peak in *Thalassiosira* and the heavily tailed Ks distribution in *Skeletonema* (Appendix S4) raise the possibility that the polyploid lineage comprises these two lineages only, with the inclusion of *Discostella* representing an artifact of sparse taxon sampling and uncertainty in the species tree.

Finally, secondary peaks were evident in Ks-based age distributions of duplicated genes from both *Thalassiosira* and *Skeletonema*, and although less pronounced, age distributions of *C. nana* and *Contricribra* paralogs also had heavy right tails (Fig. 3; Appendix S4). Additional taxon sampling will show whether the Ks data from these species are indicative of a more extensive history of paleopolyploidy across Thalassiosirales, a hypothesis that has some support from the *Thalassiosira* data. Here, gene tree reconciliation supported one deep WGD event, gene count data supported a *Thalassiosira*-specific WGD, and the Ks distribution featured several strongly supported secondary peaks (Fig. 3; Appendix S5). As a result, we cannot rule out the possibility that the *Thalassiosira* genome carries signal from two different WGD events within Thalassiosirales—its genome, therefore, representing the product of as many as four serial paleopolyploidy events (Fig. 6) dating back to the common ancestor of diatoms.

Whole-genome duplication in pennate diatoms

The transition from radial to axial cell wall symmetry and from oogamous to isogamous sexual reproduction were landmark events in diatom evolution (Round et al., 1990), circumscribing a clade whose species diversity vastly outnumbers all remaining diatoms (Guiry and Guiry, 2017) and, as a result, motivating great interest in identifying the underlying drivers of their diversification (Nakov et al. 2018). We found multiple lines of evidence supporting as many as six independent WGD events within pennate diatoms (Figs. 3 and 6). Three of these events were supported by at least two of the three strategies, whereas the others were suggested by Ks distributions (Fig. 3). The best-supported events included (1) a deep split within the pennates, circumscribing nearly the entirety of the clade (Fig. 6, branch E or F), (2) a deep split within araphid pennates (Fig. 6, the MRCA of *Asterionellopsis* and *Talaroneis*), and (3) a terminal branch representing the highly diverse navicuoloid diatoms, with a stem age of >100 Myr (Fig. 6, *Gyrosigma*). Placements of these events suggest that the majority of pennate diatoms share an ancient WGD, followed by multiple rounds of additional, nested polyploidizations that have affected several subclades of pennate diatoms (Figs. 3 and 6). Note also that these pennate-specific events might have occurred in addition to at least two earlier WGDs (Fig. 4, branches A and C), analogous to the complex serial polyploid ancestry of numerous angiosperm lineages (Bowers et al., 2003; Jiao et al., 2011).

As with WGDs in other parts of the tree, a degree of uncertainty exists with regard to the deep, nearly pennate-wide WGDs (Fig. 6, branches E and F). Specifically, reconciliation against the species tree and gene count analyses indicated that the most likely placement of this event was at the branch representative of the MRCA of all pennate diatoms excluding *Striatella* (Figs. 4 and 6, branch E). Reconciliation against MUL trees was more equivocal, however, supporting either this branch or the next branch up the backbone as the likely placement of the duplication (Figs. 5 and 6, branch F). As before, we were unable to determine whether this uncertainty was a byproduct of our exemplar sampling, i.e., the lineages relevant for pinpointing the placement of this event might be missing from our data set. Alternatively, uncertainty in the species tree might be carried over into the placement of this WGD. More densely sampled

phylogenies based on conventional phylogenetic markers place *Striatella* (along with *Asterionellopsis* and *Talaroneis*) within one of two or three clades that comprise the paraphyletic araphid pennate diatoms (Theriot et al., 2015). Phylogenomic analyses with fewer species but more markers placed *Striatella* as sister to all other pennate diatoms (Fig. 3 and Appendix S3; see also Parks et al. [2018]). These competing hypotheses have important implications for our ability to infer the location and timing of this WGD and further highlight this part of the tree as a primary target for additional genomic sampling.

Finally, with respect to hybridization and polyploidy, raphid pennate diatoms have received far more attention than any other group of diatoms (see introduction and *Mechanisms of polyploid formation in diatoms* section below). There is direct evidence for autopolyploid formation within raphid pennates in vitro (Mann, 1994; Chepurnov and Roschin, 1995) and strong genetic evidence for natural hybrids in the few species that have been examined (Casteleyn et al., 2009; Tanaka et al., 2015). Their unique suite of traits, species richness, availability of established and emerging genetic models, and extensive body of research on their reproductive biology establish raphid pennates as the premier lineage for uncovering the mechanisms and evolutionary consequences of polyploidy in diatoms.

Mechanisms of polyploid formation in diatoms

Although auto- and allopolyploids are equally abundant in angiosperms (Barker et al., 2016), the modes of polyploid formation in diatoms are much more poorly understood. Our results suggest that allopolyploidy may be especially common in diatoms. There is some genomic support for this hypothesis in the highly heterozygous genome of the raphid pennate diatom, *Fistulifera solaris*, which appears to be an allodiploid (Tanaka et al., 2015). The relatively distant parental lineages in allopolyploid events supported in our GRAMPA analyses appear to contrast with the low hybrid viability (Vanormelingen et al., 2008; Casteleyn et al., 2009; Amato and Orsini, 2015) and high levels of reproductive isolation (Amato et al., 2007; Vanormelingen et al., 2008) seen at low taxonomic levels in the raphid pennate genera *Pseudo-nitzschia* and *Eunotia*, though levels of intraspecific genetic divergence in the latter are comparable to intergeneric divergence levels in angiosperms (Baldwin et al., 1995). Incongruence between these results should be further explored at multiple scales, using a combination of laboratory experiments and comparative genomics. Comparative genomic analyses involving the broader stramenopile lineage, including taxa capable of interfamilial hybridization (Liptack and Druehl, 2000), may provide further insight into the evolution and maintenance of reproductive barriers and hybrid viability. Considering the time- and labor-intensive nature of experimental reproductive studies of diatoms (Chepurnov et al., 2004, 2008, 2012; Mann et al., 2004), evidence for hybridization and introgression in diatoms may be more efficiently pursued using genomic data (Mallet, 2005), emphasizing the need for more intensive genome sequencing projects at lower phylogenetic scales. Candidates for such studies include *Ditylum brightwellii* (Koester et al., 2010), *Sellaphora* (Mann et al., 2004; Evans et al., 2008), *Seminavis* (Moeys et al., 2016), *Pseudo-nitzschia* (Casteleyn et al., 2009; Basu et al., 2017), and *Cocconeis* (Geitler, 1927, 1973). Given the strong evidence for ancient hybridizations uncovered by our analyses, including in the pennate diatoms, it will also be important to determine the specificity of sex pheromone systems used by pennate diatoms for mate attraction (Sato et al., 2011; Gillard et al., 2013; Moeys et al., 2016). Finally, although our analyses highlighted allopolyploidy as a potentially important mode of WGD in diatoms, it is important to note that autopolyploidy may be shown to be equally, if not more, common with increased sampling. Autopolyploid formation has been directly observed in vitro for several different species of raphid pennate diatoms (Geitler, 1927; Mann and Stickle, 1991; Mann, 1994; Chepurnov and Roschin, 1995; Chepurnov et al., 2002).

A number of observed meiotic anomalies suggest that diatom polyploids could form in a variety of ways. First, meiotic nonreduction, which is thought to be the predominant mode of polyploid formation in plants (Thompson and Lumaret, 1992; Ramsey and Schemske, 1998), likely occurs in diatoms as well. Although the rate of meiotic nonreduction in diatoms is unknown, Mann (1994) observed that failed cleavage in gametangia of the raphid pennate diatom, *Dickea ulvacea*, led to the formation of "double gametes" that produced a dikaryotic, triploid-like zygote following fusion with a reduced gamete. "Centric" diatoms may also be prone to cleavage failure (von Dassow et al., 2008). Second, although polyspermy is thought to occur rarely in plants (Ramsey and Schemske, 1998), triploid and tetraploid zygotes have been produced from simultaneous gamete fusions in culture studies of several raphid pennate diatoms (Geitler, 1927; Mann and Stickle, 1991; Mann, 1994; Chepurnov and Roschin, 1995; Chepurnov et al., 2002), suggesting that this may be a principal pathway to polyploidization in diatoms. These studies have found mixed populations of co-occurring haploid, triploid, and tetraploid zygotes following one or two rounds of crossing in culture, suggesting that two-step, "triploid-bridge" routes to stable polyploidy may be more common in diatoms than other groups (Ramsey and Schemske, 1998). These hypotheses further underscore the value of the experimental reproductive studies in diatoms that initially led to these discoveries. Extending these studies to include longer-term tracking of in vitro polyploids will help clarify the long-term viability and reproductive dynamics of vegetative haploids, triploids, and tetraploids, thereby distinguishing culturing anomalies from observations that hint at the natural frequencies and mechanisms of polyploid formation in diatoms.

CONCLUSIONS

The phylogenomic results presented here provide strong initial support for a history of paleopolyploidy in diatoms that, with increased taxonomic sampling, will likely prove to be more extensive than what was uncovered with our coarse sampling. Although WGD may be common in diatoms, its roles in speciation, lineage diversification, trait and life history evolution, and habitat shifts remain unknown. Establishing these associations, and further establishing causal links between WGD and any potential evolutionary consequences, are notoriously challenging problems, even with the benefit of data sets much larger than those available for diatoms (Kellogg, 2016; Panchy et al., 2016). As with all species-rich and ecologically diverse groups, however, establishing these links, if they do indeed exist, represents an important direction in evolutionary research of diatoms.

Extending our sampling to more fully capture the broad ecological diversity of diatoms across environmental gradients in temperature, pH, and salinity will help show how consequential WGDs have, or have not, been in the ecological and evolutionary diversification of diatoms. A larger comparative framework and a more

precise reconstruction of the pattern and timing of paleopolyploid events—coupled with laboratory experiments—will show whether physiological shifts, either in short-term stress responses or in major habitat transitions, have been facilitated by genetic novelties introduced by gene or genome duplication. Compared to their diploid progenitors, for example, polyploid *Arabidopsis* have increased tolerance to salinity (Chao et al., 2013), which is one of the principal ecological divides in diatoms and other microbial eukaryotes (Round and Sims, 1980; Mann, 1999a; Logares et al., 2009).

Finally, our results highlight several important gaps in our understanding of diatom genomes. For example, for a group of this size and diversity, very few karyotypes and genome size estimates are available. The few data points that we do have, however, point to a level of genomic diversity and complexity that is proportional to their many other, and much better known, layers of morphological and ecological diversity (Kociolek and Stoermer, 1989; Connolly et al., 2008). As genomic data for diatoms continue to accumulate, a coordinated effort to establish a reference genome data set that better captures their phylogenetic and ecological diversity, similar to the current call for angiosperms (Galbraith et al., 2011), is necessary to fully understand the evolution of genome size, structure, and ploidy in diatoms. Although genome size data are few, cell size data are available for every described diatom species and so could help guide these efforts (Connolly et al., 2008). Finally, although diatoms are generally assumed to be diploid, very little is known about natural variation in ploidy levels. Although very few species have been surveyed for intraspecific variation in genome size and ploidy, the data that are available (Geitler, 1973; Koester et al., 2010) suggest that polyploidy may be an important driver of speciation in diatoms.

ACKNOWLEDGEMENTS

The authors thank the organizers of this special issue for the opportunity to contribute this manuscript. The authors also thank Ya Yang, the Associate Editor, and two anonymous reviewers for comments on earlier versions of the manuscript. Matt Ashworth generously provided several of the scanning electron micrographs shown in Fig. 1, and Gregg Thomas kindly helped with GRAMPA analyses and interpretation. The authors thank David Chafin, Jeff Pummill, and Pawel Wolinski for providing computational support through the Arkansas High Performance Computing Center (AHPCC), and the Chicago Botanic Garden for hosting and support of the *Treubia* and *Fabronia* computing clusters. This work was supported by the National Science Foundation (NSF) (Grant No. DEB-1353131 to A.J.A. and DEB-1353152 to N.J.W.), multiple awards from the Arkansas Biosciences Institute to A.J.A., and by a grant from the Simons Foundation (403249, A.J.A.). This research used computational resources available through the AHPCC, which were funded through multiple NSF grants and/or the Arkansas Economic Development Commission, and resources available at the Chicago Botanic Garden, which were funded by NSF (DEB-1239992 and DEB-1342873 to N.J.W.).

DATA ACCESSIBILITY

New RNA-seq data used in this study have been deposited in the National Center for Biotechnology Information's Sequence Read Archive (SRA) database under accessions SAMN07688919–SAMN07688929.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

LITERATURE CITED

Albertin, W., and P. Marullo. 2012. Polyploidy in fungi: Evolution after whole-genome duplication. *Proceedings of the Royal Society, B, Biological Sciences* 279: 2497–2509.

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution* 20: 255–266.

Alverson, A. J., R. K. Jansen, and E. C. Theriot. 2007. Bridging the Rubicon: Phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. *Molecular Phylogenetics and Evolution* 45: 193–210.

Amato, A., W. H. C. F. Kooistra, J. H. L. Ghiron, D. G. Mann, T. Pröschold, and M. Montresor. 2007. Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* 158: 193–207.

Amato, A., and L. Orsini. 2015. Rare interspecific breeding in *Pseudo-nitzschia* (Bacillariophyceae). *Phytotaxa* 217: 145–154.

Amborella Genome Project. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* 342: 1241089.

Ambrosino, L., H. Bostan, P. di Salle, M. Sangiovanni, A. Vigilante, and M. L. Chiusano. 2016. pATsi: Paralogs and singleton genes from *Arabidopsis thaliana*. *Evolutionary Bioinformatics Online* 12: 1–7.

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. Zhou, et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* 306: 79–86.

Baldwin, B. G., M. J. Sanderson, J. M. Porter, M. F. Wojciechowski, C. S. Campbell, and M. J. Donoghue. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82: 247–277.

Barker, M. S., N. Arrigo, A. E. Baniaga, Z. Li, and D. A. Levin. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 210: 391–398.

Barker, M. S., N. C. Kane, M. Matvienko, A. Kozik, R. W. Michelmore, S. J. Knapp, and L. H. Rieseberg. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

Basu, S., S. Patil, D. Mapleson, M. T. Russo, L. Vitale, C. Fevola, F. Maumus, et al. 2017. Finding a partner in the ocean: Molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytologist* 215: 140–156.

Blanc, G., and K. H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16: 1667–1678.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* 30(15): 2114–2120.

Borowiec, M. L. 2016. AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4: e1660.

Bowers, J. E., B. A. Chapman, J. K. Rong, and A. H. Paterson. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari, A. Kuo, U. Maheswari, et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456: 239–244.

Bushnell, B., J. Rood, and E. Singer. 2017. BBMerge—Accurate paired shotgun read merging via overlap. *PloS One* 12: e0185056.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.

Cannon, S. B., M. R. McKain, A. Harkess, M. N. Nelson, S. Dash, M. K. Deyholos, Y. Peng, et al. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution* 32: 193–210.

Casteleyn, G., N. G. Adams, P. Vanormelingen, A. E. Debeer, K. Sabbe, and W. Vyverman. 2009. Natural hybrids in the marine diatom *Pseudo-nitzschia pungens* (Bacillariophyceae): Genetic and morphological evidence. *Protist* 160: 343–354.

Chao, D. Y., B. Dilkes, H. B. Luo, A. Douglas, E. Yakubova, B. Lahner, and D. E. Salt. 2013. Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science* 341: 658–659.

Chepurnov, V., P. Chaerle, K. Vanhoutte, and D. Mann. 2012. How to breed diatoms: Examination of two species with contrasting reproductive biology. *In* R. Gordon and J. Seckbach [eds.], The science of algal fuels, cellular origin, life in extreme habitats and astrobiology, 323–340. Springer, Dordrecht, Netherlands.

Chepurnov, V. A., D. G. Mann, P. von Dassow, P. Vanormelingen, J. Gillard, D. Inze, K. Sabbe, and W. Vyverman. 2008. In search of new tractable diatoms for experimental biology. *BioEssays* 30: 692–702.

Chepurnov, V. A., D. G. Mann, K. Sabbe, and W. Vyverman. 2004. Experimental studies on sexual reproduction in diatoms. *International Review of Cytology* 237: 91–154.

Chepurnov, V. A., D. G. Mann, W. Vyverman, K. Sabbe, and D. B. Danielidis. 2002. Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology* 38: 1004–1019.

Chepurnov, V. A., and A. M. Roschin. 1995. Inbreeding influence on sexual reproduction of *Achnanthes longipes* Ag. (Bacillariophyta). *Diatom Research* 10: 21–29.

Chernomor, O., A. von Haeseler, and B. Q. Minh. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology* 65: 997–1008.

Connolly, J. A., M. J. Oliver, J. M. Beaulieu, C. A. Knight, L. Tomanek, and M. A. Moline. 2008. Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). *Journal of Phycology* 44: 124–131.

Coyer, J. A., G. Hoarau, G. A. Pearson, E. A. Serrão, W. T. Stam, and J. L. Olsen. 2006. Convergent adaptation to a marginal habitat by homoploid hybrids and polyploid ecads in the seaweed genus *Fucus*. *Biology Letters* 2: 405.

Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738–749.

Darby, C. A., M. Stolzer, P. J. Ropp, D. Barker, and D. Durand. 2017. Xenolog classification. *Bioinformatics* 33: 640–649.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2011. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* 27: 1164–1165.

von Dassow, P., T. W. Petersen, V. A. Chepurnov, and E. V. Armbrust. 2008. Inter- and intraspecific relationships between nuclear DNA content and cell size in selected members of the centric diatom genus *Thalassiosira* (Bacillariophyceae). *Journal of Phycology* 44: 335–349.

Dehal, P., and J. L. Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* 3: 1700–1708.

Durand, D., B. V. Halldorsson, and B. Vernot. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology* 13: 320–335.

Eddy, S. R. 2011. Accelerated profile HMM searches. *PLoS Computational Biology* 7: e1002195.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30: 1575–1584.

Evans, K. M., A. H. Wortley, G. E. Simpson, V. A. Chepurnov, and D. G. Mann. 2008. A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta). *Journal of Phycology* 44: 215–231.

Fawcett, J. A., S. Maere, and Y. Van de Peer. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Sciences, USA* 106: 5737–5742.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281: 237–240.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, et al. 2015. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research* 44: D279–D285.

Flagel, L. E., and J. F. Wendel. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* 183: 557–564.

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca. 2012. mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. University of Washington, Seattle, WA, USA.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150–3152.

Galbraith, D. W., J. L. Bennetzen, E. A. Kellogg, J. C. Pires, and P. S. Soltis. 2011. The genomes of all angiosperms: A call for a coordinated global census. *Journal of Botany* 2011: 1–10.

Geitler, L. 1927. Somatische teilung, reduktionsteilung, copulation und parthenogenese bei *Cocconeis placentula*. *Archiv für Protistenkunde* 59: 506–549.

Geitler, L. 1973. Auxosporenbildung und systematik bei pennaten diatomeen und die cytologie von *Cocconeis-Sippen*. *Österreichische Botanische Zeitschrift* 122: 299–321.

Gillard, J., J. Frenkel, V. Devos, K. Sabbe, C. Paul, M. Rempt, D. Inze, et al. 2013. Metabolomics enables the structure elucidation of a diatom sex pheromone. *Angewandte Chemie-International Edition* 52: 854–857.

Glasauer, S. M. K., and S. C. F. Neuhauss. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics* 289: 1045–1060.

Goldman, N., and Z. H. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11: 725–736.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology* 29: 644–652.

Guillard, R. R. L. 1975. Culture of phytoplankton for feeding marine invertebrates. *In* Culture of marine invertebrate animals, 29–60. Springer, Boston, MA, USA.

Guillard, R. R. L., and J. H. Ryther. 1962. Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt and *Detonula confervacea* (Cleve) Gran. *Canadian Journal of Microbiology* 8: 229–239.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* 59: 307–321.

Guiry, M. D., and G. M. Guiry. 2017. AlgaeBase. Available at http://www.algaebase.org.

Guiry, M. D. 2012. How many species of algae are there? *Journal of Phycology* 48: 1057–1063.

Hahn, M. W. 2007. Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biology* 8: R141.

Inoue, J., Y. Sato, R. Sinclair, K. Tsukamoto, and M. Nishida. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences, USA* 112: 14918–14923.

Ioos, R., A. Andrieux, B. Marcais, and P. Frey. 2006. Genetic characterization of the natural hybrid species *Phytophthora alni* as inferred from nuclear and mitochondrial DNA analyses. *Fungal Genetics and Biology* 43: 511–529.

Jiao, Y., J. Li, H. Tang, and A. H. Paterson. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* 26: 2792–2802.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Johnson, M. G., C. Malley, B. Goffinet, A. J. Shaw, and N. J. Wickett. 2016. A phylotranscriptomic analysis of gene family expansion and evolution in the

largest order of pleurocarpous mosses (Hypnales, Bryophyta). *Molecular Phylogenetics and Evolution* 98: 29–40.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428: 617–624.

Kellogg, E. A. 2016. Has the connection between polyploidy and diversification actually been tested? *Current Opinion in Plant Biology* 30: 25–32.

Kilham, S. S., D. A. Kreeger, S. G. Lynn, C. E. Goulden, and L. Herrera. 1998. COMBO: A defined freshwater culture medium for algae and zooplankton. *Hydrobiologia* 377: 147–159.

Kociolek, J. P., and E. F. Stoermer. 1989. Chromosome numbers in diatoms: A review. *Diatom Research* 4: 47–54.

Koester, J. A., J. E. Swalwell, P. von Dassow, and E. V. Armbrust. 2010. Genome size differentiates co-occurring populations of the planktonic diatom *Ditylum brightwellii* (Bacillariophyta). *BMC Evolutionary Biology* 10: 1.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.

Laurent, S., N. Salamin, and M. Robinson-Rechavi. 2017. No evidence for the radiation time lag model after whole genome duplications in Teleostei. *PloS One* 12: e0176384.

Liptack, M. K., and L. D. Druehl. 2000. Molecular evidence for an interfamilial laminarialean cross. *European Journal of Phycology* 35: 135–142.

Logares, R., J. Brate, S. Bertilsson, J. L. Clasen, K. Shalchian-Tabrizi, and K. Rengefors. 2009. Infrequent marine-freshwater transitions in the microbial world. *Trends in Microbiology* 17: 414–422.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155.

Lynch, M., and A. G. Force. 2000. The origin of interspecific genomic incompatibility via gene duplication. *The American Naturalist* 156: 590–605.

MacManes, M.D. 2015. An opinionated guide to the proper care and feeding of your transcriptome. *bioRxiv* 10.1101/035642.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution* 20: 229–237.

Mann, D. G. 1994. Auxospore formation, reproductive plasticity and cell structure in *Navicula ulvacea* and the resurrection of the genus *Dickieia* (Bacillariophyta). *European Journal of Phycology* 29: 141–157.

Mann, D.G. 1999a. Crossing the Rubicon: The effectiveness of the marine/freshwater interface as a barrier to the migration of diatom germplasm. *In* S. Mayama, M. Idei, and I. Koizumi [eds.], Proceedings of the 14th International Diatom Symposium, 1–21. Koeltz Scientific Books, Koenigstein. Germany.

Mann, D. G. 1999b. The species concept in diatoms. *Phycologia* 38: 437–495.

Mann, D. G., S. M. McDonald, M. M. Bayer, S. J. M. Droop, V. A. Chepurnov, R. E. Loke, A. Ciobanu, and J. M. H. Du Buf. 2004. The *Sellaphora pupula* species complex (Bacillariophyceae): Morphometric analysis, ultrastructure and mating data provide evidence for five new species. *Phycologia* 43: 459–482.

Mann, D. G., and A. J. Stickle. 1991. The genus *Craticula*. *Diatom Research* 6: 79–107.

Mann, D. G., and P. Vanormelingen. 2013. An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology* 60: 414–420.

Marcet-Houben, M., and T. Gabaldon. 2015. Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology* 13: e1002220.

Mayrose, I., S. H. Zhan, C. J. Rothfels, K. Magnuson-Ford, M. S. Barker, L. H. Rieseberg, and S. P. Otto. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333: 1257.

McKain, M. R., H. Tang, J. R. McNeal, S. Ayyampalayam, J. I. Davis, C. W. Depamphilis, T. J. Givnish, et al. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biology and Evolution* 8: 1150–1164.

Minh, B. Q., M. A. Nguyen, and A. von Haeseler. 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* 30: 1188–1195.

Mirarab, S., and T. Warnow. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31: i44–i52.

Moeys, S., J. Frenkel, C. Lembke, J. T. F. Gillard, V. Devos, K. Van den Berge, B. Bouillon, et al. 2016. A sex-inducing pheromone triggers cell cycle arrest and mate attraction in the diatom *Seminavis robusta*. *Scientific Reports* 6: 19252.

Naik, P. A., P. Shi, and C.-L. Tsai. 2007. Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association* 102: 244–254.

Nakov, T., J.M. Beaulieu, and A.J. Alverson. 2018. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytologist*: in press.

Norris, R.W., C.L. Strope, D.M. McCandlish, and A. Stoltzfus. 2015. Bayesian priors for tree calibration: Evaluating two new approaches based on fossil intervals. *bioRxiv* 014340.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, NY, NY, USA.

Oliver, M. J., D. Petrov, D. Ackerly, P. Falkowski, and O. M. Schofield. 2007. The mode and tempo of genome size evolution in eukaryotes. *Genome Research* 17: 594–601.

Otto, S. P., and J. Whitton. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* 34: 401–437.

Panchy, N., M. Lehti-Shiu, and S. H. Shiu. 2016. Evolution of gene duplication in plants. *Plant Physiology* 171: 2294–2316.

Parks, M., N.J. Wickett, and A.J. Alverson. 2018. Signal, uncertainty, and conflict in phylogenomic data for a diverse lineage of microbial eukaryotes (diatoms, Bacillariophyta). *Molecular Biology and Evolution* 35: 80–93.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology* 9: e1000602.

Poulsen, N., and N. Kroger. 2004. Silica morphogenesis by alternative processing of silaffins in the diatom *Thalassiosira pseudonana*. *Journal of Biological Chemistry* 279: 42993–42999.

Rabier, C. E., T. Ta, and C. Ane. 2014. Detecting and locating whole genome duplications on a phylogeny: A probabilistic approach. *Molecular Biology and Evolution* 31: 750–762.

Ramsey, J., and D. W. Schemske. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics* 29: 467–501.

Round, F. E., R. M. Crawford, and D. G. Mann. 1990. The diatoms: Biology & morphology of the genera. Cambridge University Press, Cambridge, UK.

Round, F. E., and P. A. Sims. 1980. The distribution of diatom genera in marine and freshwater environments and some evolutionary considerations. *In* R. Ross [ed.], Proceedings of the Sixth Symposium on Recent and Fossil Diatoms, 301–320. Otto Koeltz Science Publishers, Hirschberg, Germany.

Santini, F., L. J. Harmon, G. Carnevale, and M. E. Alfaro. 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology* 9: 194.

Sato, S., G. Beakes, M. Idei, T. Nagumo, and D. G. Mann. 2011. Novel sex cells and evidence for sex pheromones in diatoms. *PloS One* 6: e26923.

Sayyari, E., and S. Mirarab. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* 33: 1654–1668.

Schlueter, J. A., P. Dixon, C. Granger, D. Grant, L. Clark, J. J. Doyle, and R. C. Shoemaker. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868–876.

Schranz, M. E., S. Mohammadin, and P. P. Edger. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Current Opinion in Plant Biology* 15: 147–153.

Seo, T. K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution* 25: 960–971.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.

Smith, S. A., and B. C. O'Meara. 2012. treePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28: 2689–2690.

Smith-Unna, R., C. Boursnell, R. Patro, J. Hibberd, and S. Kelly. 2016. TransRate: Reference free quality assessment of de novo transcriptome assemblies. *Genome Research* 26(8): 1134–1144.

Soltis, D. E., V. A. Albert, J. Leebens-Mack, C. D. Bell, A. H. Paterson, C. Zheng, D. Sankoff, et al. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

Soltis, P. S., X. Liu, D. B. Marchant, C. J. Visger, and D. E. Soltis. 2014. Polyploidy and novelty: Gottlieb's legacy. *Philosophical Transactions of the Royal Society, B, Biological Sciences* 369: 20130351.

Soltis, P. S., and D. E. Soltis. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Current Opinion in Plant Biology* 30: 159–165.

Song, L., and L. Florea. 2015. Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* 4: 48.

Sorhannus, U. 2007. A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Marine Micropaleontology* 65: 1–12.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Sukumaran, J., and M. T. Holder. 2010. DendroPy: A Python library for phylogenetic computing. *Bioinformatics* 26: 1569–1571.

Suyama, M., D. Torrents, and P. Bork. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34: W609–W612.

Tanaka, T., Y. Maeda, A. Veluchamy, M. Tanaka, H. Abida, E. Maréchal, C. Bowler, et al. 2015. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *Plant Cell* 27: 162.

Tang, H., J. E. Bowers, X. Wang, and A. H. Paterson. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences, USA* 107: 472–477.

Tank, D. C., J. M. Eastman, M. W. Pennell, P. S. Soltis, D. E. Soltis, C. E. Hinchliff, J. W. Brown, et al. 2015. Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. *New Phytologist* 207: 454–467.

Theriot, E. C., M. P. Ashworth, T. Nakov, E. Ruck, and R. K. Jansen. 2015. Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Molecular Phylogenetics and Evolution* 89: 28–36.

Thomas, G. W. C., S. H. Ather, and M. W. Hahn. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Systematic Biology* 66: 1007–1018.

Thompson, J. D., and R. Lumaret. 1992. The evolutionary dynamics of polyploid plants: origins, establishment and persistence. *Trends in Ecology & Evolution* 7: 302–307.

Tiley, G. P., C. Ane, and J. G. Burleigh. 2016. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biology and Evolution* 8: 1023–1037.

Van de Peer, Y., E. Mizrachi, and K. Marchal. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411–424.

Van Dongen, S., and C. Abreu-Goodger. 2012. Using MCL to extract clusters from networks. *Bacterial Molecular Networks: Methods and Protocols* 281–295.

Van Dongen, S.M. 2001. Graph clustering by flow simulation. Ph.D. dissertation, University of Utrecht, Utrecht, Netherlands.

Vanneste, K., L. Sterck, A. A. Myburg, Y. Van de Peer, and E. Mizrachi. 2015. Horsetails are ancient polyploids: Evidence from *Equisetum giganteum*. *Plant Cell* 27: 1567–1578.

Vanneste, K., Y. Van de Peer, and S. Maere. 2013. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* 30: 177–190.

Vanormelingen, P., V. A. Chepurnov, D. G. Mann, K. Sabbe, and W. Vyverman. 2008. Genetic divergence and reproductive barriers among morphologically heterogeneous sympatric clones of *Eunotia bilunaris* sensu lato (Bacillariophyta). *Protist* 159: 73–90.

Vekemans, D., S. Proost, K. Vanneste, H. Coenen, T. Viaene, P. Ruelens, S. Maere, et al. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: Significance for MADS-box gene and species diversification. *Molecular Biology and Evolution* 29: 3793–3806.

Whittaker, K. A., D. R. Rignanese, R. J. Olson, and T. A. Rynearson. 2012. Molecular subdivision of the marine diatom *Thalassiosira rotula* in relation to geographic distribution, genome size, and physiology. *BMC Evolutionary Biology* 12: 209.

Winge, Ö. 1917. The chromosomes. Their numbers and general importance. *Comptes Rendus des Travaux du Laboratoire Carlsberg* 13: 131–175.

Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708–713.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* 106: 13875–13879.

Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. K. S. Wong, E. J. Carpenter, Y. Zhang, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* 32: 2001–2014.

Yang, Y., and S. A. Smith. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* 31: 3081–3092.

Yang, Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* 46: 409–418.

Zhang, Z., J. Li, X.-Q. Zhao, J. Wang, G. K.-S. Wong, and J. Yu. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics* 4: 259–263.

Zhan, S. H., M. Drori, E. E. Goldberg, S. P. Otto, and I. Mayrose. 2016. Phylogenetic evidence for cladogenetic polyploidization in land plants. *American Journal of Botany* 103: 1252–1258.