Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval

Liming Cai¹, Zhenxiang Xi¹-², André M. Amorim³, M. Sugumaran⁴, Joshua S. Rest⁵, Liang Liu⁶ and Charles C. Davis¹

¹Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Avenue, Cambridge, MA 02138, USA; ²Key Laboratory of Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610064, China; ³Departamento de Ciências Biológicas, Universidade Estadual de Santa Cruz, Ilhéus, 45.662-900, Bahia, Brazil; ⁴Rimba Ilmu Botanic Garden, Institute of Biological Sciences, University of Malaya, 50603, Kuala Lumpur, Malaysia; ⁵Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA; ⁶Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

Author for correspondence:
Charles C. Davis
Tel: +1 6174960515
Email: cdavis@oeb.harvard.edu

Received: 4 March 2018
Accepted: 21 June 2018

doi: 10.1111/nph.15357

Key words: climatic upheaval, flowering plants, genome evolution, global change, phylogenomics, speciation.

Summary

• Whole-genome duplications (WGDs) are widespread and prevalent in vascular plants and frequently coincide with major episodes of global and climatic upheaval, including the mass extinction at the Cretaceous–Tertiary boundary (c. 65 Ma) and during more recent periods of global aridification in the Miocene (c. 10–5 Ma). Here, we explore WGDs in the diverse flowering plant clade Malpighiales.
• Using transcriptomes and complete genomes from 42 species, we applied a multipronged phylogenomic pipeline to identify, locate, and determine the age of WGDs in Malpighiales using three means of inference: distributions of synonymous substitutions per synonymous site (Ks) among paralogs, phylogenomic (gene tree) reconciliation, and a likelihood-based gene-count method.
• We conservatively identify 22 ancient WGDs, widely distributed across Malpighiales sub-clades. Importantly, these events are clustered around the Eocene–Paleocene transition (c. 54 Ma), during which time the planet was warmer and wetter than any period in the Cenozoic.
• These results establish that the Eocene Climatic Optimum likely represents a previously unrecognized period of prolific WGDs in plants, and lends further support to the hypothesis that polyploidization promotes adaptation and enhances plant survival during episodes of global change, especially for tropical organisms like Malpighiales, which have tight thermal tolerances.

Introduction

Whole-genome duplication (WGD), or polyploidy, is an important evolutionary force that has shaped plant evolution. It has long been appreciated that the formation of recent polyploids in vascular plants is common (Stebbins, 1947; Barker et al., 2016), and mounting evidence suggests that ancient polyploids are more frequent than once thought. WGD has been identified in ferns using three means of inference: distributions of synonymous substitutions per synonymous site (Ks) among paralogs, phylogenomic (gene tree) reconciliation, and a likelihood-based gene-count method.

Whole-genome duplication (WGD), or polyploidy, is an important evolutionary force that has shaped plant evolution. It has long been appreciated that the formation of recent polyploids in vascular plants is common (Stebbins, 1947; Barker et al., 2016). Whole-genome duplications (WGD) are widespread and prevalent in vascular plants and frequently coincide with major episodes of global and climatic upheaval, including the mass extinction at the Cretaceous–Tertiary boundary (c. 65 Ma) and during more recent periods of global aridification in the Miocene (c. 10–5 Ma). Here, we explore WGDs in the diverse flowering plant clade Malpighiales.
In summary, it is established that WGDs are a prominent feature of vascular plant evolution, yet they remain underexplored in many clades. Here, we investigate WGDs in the large and diverse angiosperm order Malpighiales, which contains more than 16,000 mostly tropical species with tremendous morphological and ecological diversity. Members of this clade also include numerous economically important crops, such as rubber, cassava, and flax. The Malpighiales have long been recognized as one of the most difficult clades to resolve in the flowering plant tree of life (Davis et al., 2005; Wurdack & Davis, 2009), which has been attributed in part to its rapid radiation in the mid-Cretaceous (Davis et al., 2005; Xi et al., 2012b). However, recent efforts utilizing phylogenomic approaches have greatly increased our understanding of deep-level relationships in the order (Xi et al., 2012b). Additionally, the clade includes numerous species that have previously been targeted for genomic investigation of WGDs. Eight genomes are currently available for interrogation: *Hevea brasiliensis* (rubber), *Manihot esculenta* (cassava), *Linum usitatissimum* (flax), *Ricinus communis* (castor bean), *Jatropha curcas* (Barbados nut), *Populus trichocarpa* (black cottonwood), *Salix suchouensis* (shrub willow), and *Salix purpurea* (purple willow). Notably, three WGDs have been identified in Malpighiales using these data: in the common ancestor of *Populus* and *Salix* (35–65 Ma; Tuskan et al., 2006; Dai et al., 2014); in the common ancestor of *Manihot* and *Hevea* (35–47 Ma; Bredeson et al., 2016); and more recently in *L. usitatissimum* (5–9 Ma; Z. Wang et al., 2012). In addition, studies using transcriptomes have identified an older WGD shared by all blue-flowered *Linum* species, including *L. usitatissimum*, at 20–40 Ma (Sveinsson et al., 2014). More complicated polyploidy histories involving multiple rounds of WGDs and hybridization have also been reported using chromosome count data in the genera *Passiflora* (Mayrose et al., 2009) and *Viola* (Marcussen et al., 2012, 2014). In short, given the apparent propensity of WGDs in Malpighiales and the existing complete sequence data available for the order, this clade is an ideal study system for investigating the frequency and timing of WGDs.

**Materials and Methods**

**Taxon sampling and transcriptome sequencing**

We collected genomic and transcriptomic data for 36 species representing 21 families of Malpighiales, spanning all major clades (*sensu* Wurdack and Davis (Wurdack & Davis, 2009; Chase et al., 2016) (Supporting Information Tables S1–S3). In addition, three closely related outgroups from Celastraceae (Celastrales), Elaeocarpaceae (Oxalidales), and Oxalidaceae (Oxalidales), plus three more distantly related outgroups (*Cucumis sativus* (Eurosid I), *Theobroma cacao* (Eurosid II), and *Vitis vinifera* (basal Rosidae)) were used for rooting (Chase et al., 2016). We sequenced transcriptomes of 15 Malpighiales species following the protocol described by Xi et al. (2012a). Total RNA from leaf tissue was extracted using the RNAqueous and Plant RNA Isolation Aid kits (Ambion Inc.), and treated with the TURBO DNA-free kit (Ambion, Inc.) at 37°C for 4 h to remove residual DNA. The complementary DNA library was synthesized from total RNA following the protocols of Novaes et al. (2008). Illumina paired-end libraries were prepared for complementary DNA following the protocols of Bentley et al. (2008). Each library was sequenced in a single lane of a Genome Analyzer II instrument (Illumina, Inc.) with paired-end 150 bp read lengths. We additionally included 13 annotated transcriptomes from the OneKP project (Table S2) to complete our taxon sampling for Malpighiales (Matasci et al., 2014). These sequences were obtained following the protocol outlined by Wickett et al. (2014). Finally, we also obtained whole-genome sequence data from eight published genomes of Malpighiales plus three outgroup species: *H. brasiliensis* (Willd. ex A.Juss.) Müll.Arg., *J. cacao* L., *L. usitatissimum* L., *M. esculenta* Crantz, *P. trichocarpa* Torr. & A.Gray ex Hook., *R. communis* L., *S. purpurea* L., *S. suchouensis* W.C.Cheng ex G.H.Zhu, *C. sativus* L., *T. cacao* L., *V. vinifera* L. (Table S3).

**Transcriptome assembly**

Raw sequencing reads were first corrected for errors using *RCORRECTOR* (Song & Florea, 2015). Reads marked as ‘unfixable’, generally constituting regions of low complexity, were discarded. Sequencing and PCR adapters were identified and trimmed using *TRIMGLORE* v.0.4.4.2 (Krueger, 2015). We examined the quality of trimmed reads using *FASTQC* v.0.11.5 (Anders, 2010) and then assembled the reads using *TRINITY* v.2.1.1 (Grabherr et al., 2011). We used the longest isoform from each Trinity assembly and further reduced the redundancy generated from sequencing error or alternative splicing by performing a similarity-based clustering (~c 0.99 -n 10, threshold following Yang et al., 2015) via *CD-HIT-EST* v.4.6.4 (Li & Godzik, 2006). The completeness of our assemblies was assessed by comparison against the single-copy orthologs plant database, BUSCO (Simão et al., 2015; Fig. S1). Coding regions of each putative transcript were predicted following the transdecoder workflow (Haas & Papanicolaou, 2012). Finally, to control for transcriptome quality in our subsequent assessments of WGD, we reanalyzed our combined transcriptome and complete genome data following the methods described later with only high-quality transcriptomes. Here, transcriptomes with more than 40% (382/956) missing BUSCOs were removed, including *Bhesa paniculata*, *Flacourtia jangomas*, *Galearia maingayi*, *Isonanthus reticulata*, *Podostemum ceratophyllum*, *Rinorea anguifera*, and *Tristellateia australisae*.

**Gene family clustering and orthology inference**

To assign sequences into homologous gene families, we used an integrated method that takes into account sequence similarity and species phylogeny. We first constructed whole genome/transcriptome homology scans using *PROTEINORTHO* v.5.13 (Lechner et al., 2011) with default parameter settings. This program extends the reciprocal best *BLAST* hit method and is computationally efficient. Clusters were searched to identify gene families containing at least 22 (> 60%) in-group species. This resulted in
8465 candidate homolog clusters. This similarity-based homology search can sometimes be erroneous due to deep paralogs, misassembly, or frame shifts (Yang & Smith, 2014). To reduce such errors in orthology inference, we further applied a tree-based method to sort genes into orthology groups (Yang & Smith, 2014). This method does not rely on a known species tree, but rather iteratively searches for the subtree with the highest number of in-group taxa to assign as orthology groups (Fig. S2). Here, we first aligned the protein sequences of each homolog cluster with MAPFT v.7.2.299 (Katoh & Standley, 2013) using the local alignment algorithm (-localpair -maxiterate 1000). The resulting protein alignments were converted into the corresponding codon alignments using PAL2NAL v.14 (Suyama et al., 2006). A gene family tree of each codon alignment was then reconstructed using RAxML v.8.1.5 (Stamatakis, 2014b) with 10 random starting points. To sort homologs into orthologous groups, we first pruned exceptionally long and short branches within each gene family tree because we suspected such branches to be incorrect homologs, sequencing errors, or transcript isoforms. In addition, short branches may cause problems of overfitting in the subsequent penalized likelihood dating method (Sanderson, 2004). Along these lines, branches that were 10 times longer than the ‘5% trimmed mean branch length’, or shorter than an absolute value of $1 \times 10^{-15}$, were pruned. The ‘5% trimmed mean branch length’ was defined as the mean branch length after discarding the lowest and highest 5% of the branch length distribution in each gene family. Orthology was then inferred based on this pruned gene family tree using the ‘rooted tree’ method (‘prune_paralogs_RT.py’) following Yang & Smith (2014). The resulting 5113 orthology clusters were realigned as amino acids using the method already described. Finally, the back-translated nucleotide alignments were used for subsequent phylogenetic analysis.

Phylogeny reconstruction and molecular dating

To infer the phylogeny of each orthology group, we first removed sites containing >80% gaps using TRIMAL v.1.4.15 (Capella-Gutiérrez et al., 2009). We then applied RAxML v.8.1.5 to reconstruct maximum likelihood (ML) trees under the GTR + F model with 20 random starting points. We chose the GTR + F model because it accommodates rate heterogeneity among sites, whereas the other available GTR model in RAxML, the GTRCAT model, is less appropriate owing to our small taxon sampling size (Stamatakis, 2014a). We then filtered each gene tree to eliminate exceptionally long and short branches using the method outlined earlier. The remaining gene accessions were realigned and a final round of ML tree inference was conducted. Statistical confidence of each gene tree was assessed by performing 100 bootstrap (BS) replicates with branch length (-N 100 -k).

In addition to determining the phylogeny of orthology groups, we also estimated the divergence time of each orthologous tree for the purpose of dating WGDs. We applied the penalized likelihood method for its high efficiency when dealing with large data sets including thousands of genes. By contrast, Bayesian divergence time estimation is computationally intensive and difficult to implement with a data set of this size. We estimated molecular divergence times for each ML gene tree as well as the BS trees with penalized likelihood as implemented in r8s v.1.7 (Sanderson, 2003). The following four calibration points were applied to each tree: the root age was fixed at 109 Ma, representing the approximate age of the crown group divergence in Malpighiales (Xi et al., 2012b); the minimum age of crown group clusios (including Calophyllum macrocarpum, Clusia rosea, P. ceratophyllum, Garcinia oblongifolia, Hypericum perforatum, and Mammea americana) was set to be 89 Ma, representing the oldest known fossil in Malpighiales, Palaeoclusia chevalieri (Ruhfel et al., 2013); and two additional secondary minimum age constraints from Xi et al. (2012b) to constrain stem group Euphorbiaceae (97 Ma) and Salicaceae (69 Ma). For each clade, the age constraint was placed on the most recent common ancestor of all gene accessions forming a monophyletic group for that clade. The optimal smoothing parameter for each gene tree was determined within the range of parameter space ($1 \times 10^{-4.5}, 1 \times 10^{4.5}$) by cross-validation (Sanderson, 2003). Trees were subsequently dated under the assumption of a relaxed molecular clock by applying a semiparametric penalized likelihood approach using a truncated Newton optimization algorithm in r8s.

Species tree estimation

We inferred a single reference species tree for our analysis of WGD applying a summary coalescent method as implemented in ASTRA v.4.10.5 (Mirarab & Warnow, 2015). As input gene trees for our species tree analysis we utilized all 5113 gene trees derived from each orthology cluster described earlier. Before species tree inference, we applied an additional branch trimming process to remove duplicated taxa from individual gene trees following Yang & Smith (2014). At each node where two descendant clades contain overlapping taxa, the branch with the smaller number of taxa was pruned. These pruned gene trees were subsequently used in ASTRA for species tree estimation. We additionally processed the 100 BS trees for each gene using the same pruning method described earlier. These BS trees are used in ASTRA analyses to conduct BS replicates for species tree estimation. Molecular divergence time estimates were subsequently inferred for the species tree using the penalized likelihood method described earlier using a concatenated sequence matrix derived from all 40 genes containing at least 34 in-group taxa.

Overview of the methods applied to infer WGD

We utilized three lines of inference described in the following three subsections to identify, locate, and determine the age of WGDs in Malpighiales. Each of these methods has been commonly applied in angiosperms and elsewhere, and includes distribution of synonymous substitutions per synonymous site ($K$) among paralogs (Cui et al., 2006; Barker et al., 2008), phylogenomic (gene tree) reconciliation (Jiao et al., 2011; Li et al.,...
Ks-based method for WGD identification

This method identifies a proliferation of duplicated genes from WGDs under the assumption that synonymous substitutions between duplicate genes accrue at a relatively constant rate. Here, each species was subjected to a reciprocal BLAST search to identify putative paralogous gene pairs in their protein coding sequences. Paralogous pairs were identified as sequences that demonstrated 40% sequence similarity over at least 300 bp from a contiguous MEGABLAST search (Zhang et al., 2000; Barker et al., 2008). Each paralogous protein sequence pair was aligned using MAFFT (Katoh & Standley, 2013) and then back-translated to their coding sequences using PAL2NAL (Suyama et al., 2006). All sites containing gaps were removed from the alignment. Ks values for each duplicate pair were calculated using the ML method implemented in codeml of the PAML package (Yang, 1997) under the F3 x 4 model (Goldman & Yang, 1994). To infer WGDs from the Ks distribution, we employed the one sample Kolmogorov–Smirnov goodness-of-fit test followed by 100 BS resampling to assess statistical confidence (Cui et al., 2006). A significant P value (<0.05) rejected the null hypothesis of a birth–death process of gene duplication, thus supporting evidence of WGD. Because peaks produced by paleopolyploidy are expected to be approximately Gaussian (Blanc & Wolfe, 2004; Schluter et al., 2004), we applied the expectation–maximization (EM) algorithm to fit mixtures of Gaussian distributions to our data using the normalmixEM() function in the R package mixtools (Benaglia et al., 2009). Estimated mean peak values for each taxon are reported in Figs S3, S4. Alternative splicing can confound the signal of WGD using the Ks method (Barker et al., 2008). To alleviate this concern, sites containing gaps were removed from paralog alignments; thus, transcript isoforms generated from alternative splicing will receive a Ks value of zero. All pairs with a Ks value of less than 0.001, which would include these transcript isoforms as well as recent tandem duplications, were discarded and not considered in the Ks distribution. Finally, to further explore the sensitivity of misassemblies and tandem gene duplications on WGD identification, we conducted the same Ks analysis but applied a more stringent 0.95 threshold in the CD-HIT-EST analysis (~c 0.95 - r 10).

Placing and dating WGDs using phylogenetic reconciliation and molecular divergence time estimation

We applied a phylogenetic approach to identify more precise placements of WGDs and to determine the approximate age of these events. We first reconciled each orthology tree to the species tree under the duplication–transfer–loss model in Notung v2.9 (Chen et al., 2000). Here, total numbers of gene duplications inferred from well-supported gene tree nodes (>70 BS) are summarized onto the species tree. For each branch in the species tree, we calculated the percentage of genes duplicated along that branch (total number of inferred gene duplications along the branch/total number of genes containing at least one descended copy on that branch (i.e., from both single and duplicated gene copies)). The range of duplicated genes along branches where WGD was inferred was 10.0–84.3% (Table S4). Our threshold percentage for identifying a WGD was 10.0%, which is the lowest percentage for a terminal WGD as determined by our Ks analysis (in genus Bhesa). Terminal WGDs are likely to exhibit a higher percentage of retained duplicated genes, and thus represent a more conservative filter for WGD assessment. This threshold is also well above the percentage identified from fully sequenced genomes and transcriptomes that do not exhibit recent WGDs. These fully sequenced genomes and transcriptomes instead show maximally only 1.2–2.2% of duplicated genes where tandem duplications, not WGDs, have been inferred (calculated from Chrysobalanus and Jatropha, Table S4). Together, this represents a reasonable approach for estimating WGDs in Malpighiales.

Following our phylogenetic localization of WGDs, we applied a customized R script to extract the divergence times from our r8s analyses to summarize the age of gene duplications along each branch of the species tree. Only ages of nodes supported with >70 BS were used. The inferred distribution of divergence times was fitted to a mixture of Gaussian models using the R package mixtools as described earlier to estimate mean age of each WGD. In several cases, the best model inferred two WGDs along the same branch. These cases were independently supported by our phylogenetic analysis but with smaller percentages of duplications, presumably due to gene loss or missing data (Table S4). Estimated mean peak values for each taxon are reported in Table S4 and Fig. S5. Confidence intervals for mean age of each WGD were estimated using the 100 bootstrapped trees for each gene as outlined earlier (Table S4). To further explore the impact of root age on divergence time estimation using this penalized likelihood method, we fixed the root age to be 125 Ma, which is the maximum age of crown group eudicots based on the fossil record (Hughes & McDougall, 1990; Doyle & Hotton, 1991; Magallón et al., 2015). Moreover, this age is over 20 Ma older than the optimal age estimates for crown group Malpighiales (103.6 Ma) as inferred by Magallón et al. (2015). We applied the same set of minimum age constraints described earlier to conduct the divergence time estimation (Table S4).

Gene-count-based method for confirmation of WGDs

To further assess statistical confidence for the 24 putative WGDs inferred from the Ks method and using the aforementioned phylogenetic approach, we applied the ML method to test the number and placement of WGDs using gene count data (Rabier et al., 2014). This method estimates the likelihood of prespecified putative WGDs on a phylogeny using a gene count matrix summarized across all orthologous genes. It is advantageous because it suffers less from false-positive rates due to tandem duplication and assembly error (Rabier et al., 2014), which could create an artificial signal of polyploidization. We first tested the utility of this method by examining three independent WGDs previously identified from fully sequenced genomes using synteny analyses.
These three WGDs were identified in *Populus* and *Salix* (Hanley et al., 2006; Tuskan et al., 2006), in *Hevea* and *Manihot* (Bredeson et al., 2016; Tang et al., 2016), and in *Linum* (Z. Wang et al., 2012). Filtering of the gene count matrix to avoid missing data is critical for this method, which may otherwise lead to biased estimates (Rabier et al., 2014). To accomplish this, our dated species trees were first pruned to contain only species with fully sampled genomes, including these five species, plus *Jatropha* and *Ricinus* (where WGDs have not been previously detected) and *Vitis* (as outgroup). Next, a gene count matrix was summarized for all species across all ortholog trees (Fig. S6). We further conditioned the data matrix for all gene families to contain one or more gene copies descended from the branch along which the WGD was tested. Along these lines, gene families that were missing in the taxa affected by WGD were removed in each test. The conditional likelihoods were subsequently estimated for models with and without the WGD of interest using a prior geometric mean of 1.5 (Tiley et al., 2016). After convergence of the likelihood scores for all runs, we performed a series of likelihood ratio tests (LRTs) to determine the significance of individual WGDs. All three previously identified WGDs were successfully identified with confidence (LRT statistic \( > 9.55 \), probability of type I error \( \ll 0.001 \)). In addition, *Ricinus* and *Jatropha* showed no evidence of WGD (Table S4), suggesting a low false-positive rate of this method.

We tested the remaining WGDs by sequentially adding species associated with each WGD to the seven-species phylogeny. In each case, we added all of the species from which a single WGD to be tested were descended and tested for a WGD along the added branch of interest. We generated the conditioned gene count matrix containing varying numbers of gene families as described earlier (Table S4).

### Clustering of WGDs in time

To assess whether WGDs were clustered in time, we tested for the optimum number of clusters in age distribution using the finite Gaussian mixture modeling in R package *mclust* (Fraley & Raftery, 2002). The EM algorithm was used for mixture estimation and the Bayesian information criterion (BIC) was used for comprehensive clustering. In order to assess the confidence of the clustering, we conducted the same analysis on WGD age distributions derived from 100 BS replicates. We conducted this Gaussian mixture modeling for age estimations derived from r8s analyses with fixed root age at 109 Ma as well as age estimations with fixed root age at 125 Ma. To further explore the clustering time of the more ancient WGDs, we excluded WGDs younger than 20 Ma and conducted the same Gaussian mixture modeling analysis.

### Synteny-based assessment of WGD in *Linum*

Syntenic analysis serves as the gold standard for inferring WGDs, but it is only amenable to the mostly completely assembled genomes. Our phylogenetic and \( K_s \) approach identified two WGDs in *Linum*. Here, subsampled gene families containing three or four gene copies of *Linum* consistent with two duplications. The most closely related paralogous gene pairs of *Linum* were expected to arise from the most recent WGD, while their relationship with the remaining copy(ies) arose from the more ancient WGD (Fig. S7). We then mapped these paralogs onto the genome using MCSAnX_h (Y. Wang et al., 2012) and visualized the result using RCircos (Zhang et al., 2013).

### Results

Our final data set included 36 in-group taxa derived from eight genomes and 28 transcriptomes (15 newly acquired for this study) plus six outgroup species (Tables S1–S3). Summary statistics of the transcriptome assemblies are reported in Table S5. The taxon sampling includes 21 traditionally recognized families in the order, thus representing the broad outline of phylodiversity within Malpighiales (*sensu* Angiosperm Phylogeny Group IV; Wurdack & Davis, 2009; Chase et al., 2016).

Our analyses identified a total of 22–24 WGDs broadly distributed across the Malpighiales phylogeny (Fig. 1). In nearly all cases these events are corroborated by all three methods. Our \( K_s \) analysis identified WGD in 22 species regardless of the sequence similarity threshold (Figs 1, S3, S4 for the 0.99 \( CD-HIT-EST \) threshold; Fig S8 for the 0.95 \( CD-HIT-EST \) threshold); our phylogenomic reconciliation analysis identified 24 WGDs (Table S4); and 22 WGDs were verified with the likelihood-based gene-count method (LRT statistic \( > 9.55 \); Table S4). Moreover, these results are robust to data quality and phylogenetic uncertainty (by applying our BS resampling procedure described earlier).

We additionally analyzed a reduced data set containing only completely sequenced genomes and high-quality transcriptomes to further verify these results. Even with this more conservative data set, we still identified 22 WGDs using all three methods (Table S6). Three of the WGDs we identify validate those from previous studies in the common ancestor of *Populus* (Tuskan et al., 2006) and *Salix* (Sterck et al., 2005; Hanley et al., 2006), in the common ancestor of *Manihot* (Bredeson et al., 2016) and *Hevea* (Tang et al., 2016), and in *Linum* (Z. Wang et al., 2012). In addition, recent analyses have inferred two independent WGDs in *Linum* using \( K_s \) distributions (Z. Wang et al., 2012; Sveinsson et al., 2014). These events, however, have not previously been corroborated simultaneously due to the limited power of this method. Our reconciliation and gene-count method, by contrast, identifies these two duplications in the lineage leading to *Linum*. Furthermore, we verified these two independent WGDs using a phylogenetically guided synteny analysis (Jiao et al., 2014) based on 879 gene families containing three or four gene copies in *Linum* (Fig. S7). Here, we identified 15 syntenic regions across large scaffolds reflecting the four-parted paralogous relationship created by two independent WGDs in the *Linum* lineage.

Our divergence time estimates indicated that the timing of WGDs range broadly from 3.2–85.0 Ma (Fig 2; Table S4). Surprisingly, however, these events are not randomly distributed in time. Instead, our Gaussian mixture model (Fraley...
indicates that the inferred ages of WGDs display a bimodal distribution (BIC = 225.05, compared with BIC = 227.30 for the univariate normal distribution) with peaks at the Eocene-Paleocene (mean age 53.89 Ma) and late Miocene (mean age 7.39 Ma). Our assessment using BS replicates provides confident statistical support for this interpretation: 98% of the replicates support a bimodal distribution of WGDs with the mean age of each cluster ranging from 6.06 to 13.78 Ma and from 52.26 to 57.79 Ma. In our case, an overwhelming number of WGDs (n = 19) occur during the older, early Eocene time period, vs only five during the more recent late Miocene period. When we excluded WGDs with mean age younger than 20 Ma, the 19 WGDs were estimated to cluster at 54.21 Ma according to the Gaussian mixture model analysis. The ages of these WGDs are also robust to alternative age constraints. When we fixed the root age of each gene tree at 125 Ma, age estimations of most WGDs were 0–9.9 Ma older than previous estimations except for the recent WGD in Endospermum, which is 0.3 Ma younger (Table S4). However, regardless of these older ages, our Gaussian mixture model still supported a clustering age of 56.8 Ma (54.81–60.42 Ma in the 100 BS replicates), which remains closer to the Eocene-Paleocene transition than to the KT boundary.
Discussion

Massive WGDs in Malpighiales

The WGDs we identified are broadly distributed across 18 branches of the Malpighiales phylogeny. All of these WGDs are inferred in the crown groups of major clades where phylogenetic uncertainty is minimum. As a result, although the backbone of the phylogeny in Malpighiales is still challenging to resolve (species tree with BS support Fig. S9), the phylogenetic placements of these WGDs are clear. Moreover, we identified the same set of WGDs in our full data set compared with the more stringent, filtered data set (cf. Bhesa, which we excluded in the more conservative analysis). This result indicates that low-quality transcriptomes do not have a major impact on our assessment of WGD in Malpighiales.

Interestingly, these WGDs are commonly associated with the most diverse clades in the order, including in the clusioids, ochnoids, euphorbioids, phyllanthoids, violets and passion flowers. Additionally, we note that some species-poor clades show no evidence of duplication (e.g., Malesherbia, Rinorea, and Elatinaeae, among others; Fig. 1), despite having species-rich sister clades. This lends tentative support to suggestions that WGDs may fuel species diversification (Tank et al., 2015), possibly via the establishment of reproductive barriers (Husband et al., 2016). However, other studies have concluded that, although polyploidization is important to cladogenesis in plants, it likely does not enhance species diversification rates (Wood et al., 2009; Estep et al., 2014). We cannot adequately address this question with existing data (Kellogg, 2016); but regardless, our analyses set the stage for establishing the finer scale taxon sampling necessary to pinpoint these events to clarify the association between WGDs and the tempo of diversification in Malpighiales. Namely, do WGDs precede prolific diversification of Malpighiales subclades?

One WGD requires more detailed exploration. In our phylogenetic reconciliation and gene-count analysis, a WGD is inferred to predate the common ancestor of *Populus, Salix,* and
Flacourtia (the $K_c$ analysis is inconclusive for Flacourtia; Figs 1, S4). Chromosome count data, however, do not support such an early WGD. Instead, the chromosome number of Populus and Salix are approximately twice that of Flacourtia ($2n = 38$, vs $2n = 20$ or 22 in Flacourtia; IPCC Chromosome Reports; http://www.tropicos.org/Project/IPCN), suggesting that this WGD event likely occurred more recently and is thus restricted to the clades, including Asteraceae (Barker et al., 2008; Huang et al., 2016), Poaceae (Estep et al., 2014), and Caryophyllales (Yang et al., 2015, 2017), also show histories characterized by prolific WGDs. Collectively, these results suggest that Malpighiales and other previously examined clades represent a more pervasive pattern of WGDs characteristic of possibly hundreds of major angiosperm clades (and extending to other vascular plant clades, including gymnosperms and ferns; Wood et al., 2009; Li et al., 2015). Further investigation is required to address this question more broadly, but this seems a plausible hypothesis in light of recent findings.

Timing of WGDs coincides with events of climate upheaval

WGDs in Malpighiales show a bimodal distribution through time, which is supported by 98% of our BS replicates. The majority of the WGDs were identified to cluster around the Eocene–Paleocene transition (mean age 53.89 Ma), during which time the planet was warmer and wetter than any other period in the Cenozoic. The age of this older peak is estimated to range between 52.26 and 57.79 Ma based on our BS replicates, which falls entirely within the prolonged warming trend from the late Paleocene through the early Eocene (58–50 Ma, Zachos et al., 2008). Moreover, the late Paleocene–early Eocene age of this peak is also robust when we excluded WGDs younger than 20 Ma (mean age 54.21 Ma) or apply an older root age of 125 Ma consistent with the oldest eudicots (mean age 56.8 Ma). Previously, WGD has been hypothesized to buffer plants through episodes of major global and climatic upheavals (Fawcett et al., 2009). Studies have identified WGDs associated with the earlier KT boundary (c. 65 Ma) when a large meteor impacted off the Yucatán Peninsula disrupting the global climate, precipitating a major reorganization of the terrestrial biota (Fawcett et al., 2009; Vanneste et al., 2014). Similarly, the late Miocene–early Pliocene (c. 10–5 Ma) has been implicated as another period of climatic instability when WGDs were pervasive (Estep et al., 2014). The expansion of Cs grassland as a result of widespread global aridification (Gerling et al., 1997; Arakaki et al., 2011) has, in particular, been inferred to be correlated with numerous polyploidizations in grasses, which are among the most important members of these arid and cooler habitats that bear their namesake; that is, grassland and steppe biomes (Estep et al., 2014).

By contrast, relatively little is known about WGDs during the Eocene. This is surprising, because the Eocene–Paleocene transition is associated with an extended and prolonged period of intense warming. Most notably, the Paleocene–Eocene Thermal Maximum (56 Ma) and the subsequent Eocene Climatic Optimum (49 Ma) constitute the warmest and most humid period during the Cenozoic. During this time, mean global temperatures increased by 5–10°C due to massive release of C-depleted carbon (Pagani et al., 2006; Zeebe et al., 2009). This dramatic climate upheaval is thought to have stimulated profound reshuffling of the terrestrial biome, spurring plant migrations, extensive species turnover, and accelerated species diversification in numerous plant and animal clades (Clay & Gingerich, 1998; Wilf, 2000; Bowen et al., 2002; Wing et al., 2005; McKenna & Farrell, 2006; Ramírez et al., 2007; Schuettbelz & Pryor, 2009; Jaramillo et al., 2010). Our results establish a record of at least 19 WGDs during this period, suggesting a role in adaptation during Paleocene–Eocene warming. We hypothesize that this may be common for predominantly tropical groups, like Malpighiales (Davis et al., 2005), which are likely much more impacted by warming given the relatively tight thermal tolerances exhibited by many such groups (Janzen, 1967; Tewksbury et al., 2008; Wright et al., 2009).

What may have stimulated interactions that facilitated increased polyploid formation during the warmer, wetter period of the Eocene beyond the generally increased rates of angiosperm diversification during this window of time? One hypothesis, coined the ‘neutral’ process (Van de Peer et al., 2017), posits that these sorts of upheavals increase the formation of unreduced gametes and therefore result in an excess of polyploids. It is widely appreciated that external stimuli, temperature in particular, has a pronounced effect on unreduced gamete formation (De Storme & Geelen, 2014). In this case, it appears that both high and low temperatures can promote unreduced gametes in various taxa, as demonstrated in Arabidopsis (De Storme et al., 2012) and
some roses (Pécix et al., 2011) respectively. In addition, a recent study in Arabidopsis thaliana demonstrated that supernumerary sperm fusion could generate viable polyploid offspring that exhibited vegetative vigor (Nakel et al., 2017). Thus, the extensive early Eocene warming and even the Miocene aridification might have significantly increased unreduced gametes, perhaps contributing to enhanced polyploid formation. This is supported by evidence of increased levels of unreduced gametes in gymnosperms and lycophytes during comparable upheavals, including during the Triassic–Jurassic (Kürschner et al., 2013) and Permian–Triassic (Visscher et al., 2004; Foster & Afonin, 2005) transitions.

Although such polyploidizations might have initially arisen more neutrally, it is also possible that polyploids were adapted for survival in these changing landscapes (the ‘adaptive’ processes; Van de Peer et al., 2017). Of course, both neutral and adaptive processes can contribute simultaneously to the formation and establishment of polyploids. Polyploids are often viewed as evolutionary dead ends owing to their small population sizes, relatively restricted distributions, high extinction risks, and seemingly sparser representation in the deep angiosperm phylogeny (Stebbins, 1970; Comai, 2005; Mayrose et al., 2011). However, these may not apply under less stable environments, such as during major climatic upheavals, when polyploids may outperform their diploid progenitors (Van de Peer et al., 2017). It has been hypothesized that such genomic novelty and epigenetic repatterning may result in phenotypic variability, including variants that confer selective advantages in stressful conditions (Wendel, 2000; Comai, 2005; Madlung, 2013; Van de Peer et al., 2017). In particular, the advantages of especially allopolyploids include altered gene expression leading to hybrid vigor and increased genetic variation (Comai, 2005; Lynch, 2007). Along these lines, polyploids have been reported to be more frequent at higher latitudes and in xeric environments, which may mimic such upheavals (Löve & Löve, 1949; Löve, 1953; Hanelt, 1966; Brochmann et al., 2004). Polyploids also occur with greater frequency among invasive plants, which commonly become established on disturbed grounds (Prentis et al., 2008; te Beest et al., 2011). In support of this argument, Brochmann et al. (2004) reported an unexpected overabundance of recently formed polyploids in newly deglaciated areas in the Arctic. Additionally, megaflora fossils from Wyoming, USA, indicate that the dynamics of plant community assembly after dramatic warming during the Paleocene–Eocene Thermal Maximum is very similar to late and postglacial florals (Wing et al., 2005), suggesting that observations for the Arctic may represent similar responses to warmer and wetter periods during the Eocene transition. Regardless of these competing ideas, the striking propensity and clustered distribution of WGDs in time strengthens the hypothesis that polyploidization may be an important means of lineage persistence during episodes of major global change.

Acknowledgements

We thank S. Edwards, G. Giribet, E. Kellogg, A. Knoll, and members of the Davis laboratory for technical assistance and valuable discussions. D. Soltis provided access to the Malpighiales, Celastrales, and Oxalidales sampling from the 1KP project and read a final version of the draft in October 2017. Funding for this study came from US National Science Foundation (NSF) Assembling the Tree of Life grants DEB-0622764, DEB-1120243, and DEB-0544039 (to C.C.D.) and from Harvard University. J.S.R. acknowledges financial support from US National Institutes of Health/National Institute of General Medical Sciences grant R01 GM108904. A.M.A. acknowledges financial support from Brazil National Council for Scientific and Technological Development (CNPq) PROTAX Malpighiales grant 440543/2015-0 and Research Productivity Fellowship grant 310717/2015-9.

Author contributions

L.C. and C.C.D. designed the research; L.C. analyzed the data; Z.X. collected the data; L.C. and C.C.D. wrote the initial paper draft with input from A.M.A., M.S., J.S.R. and L.L.

References


**Supporting Information**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Fig. S1** BUSCO assessment of transcriptome and genomic completeness of 42 sampled species.

**Fig. S2** Bioinformatic pipeline depicting transcriptome assembly, homology and orthology inferences, and three methods for WGD identification, placement, and dating.

**Fig. S3** Histograms of the synonymous substitutions per synonymous site (*K*) of duplicated gene pairs among 22 Malpighiales taxa showing WGDs.

**Fig. S4** Histograms of the synonymous substitutions per synonymous site (*K*) of duplicated gene pairs among 11 Malpighiales taxa where WGD is absent (or inconclusive in the case of three taxa).

**Fig. S5** Histograms depicting divergence time estimations inferred using penalized likelihood for 22 species that exhibited WGDs (summarized in Fig. S3).

**Fig. S6** Gene count matrix for 5113 ortholog groups of 36 Malpighiales species.

**Fig. S7** Phylogeny-guided syntenic analyses demonstrate successive WGDs in *Linum*.

**Fig. S8** Histograms of the synonymous substitutions per synonymous site (*K*) of duplicated gene pairs among 22 Malpighiales taxa showing WGDs.

**Fig. S9** Cladogram of 50% majority-rule bootstrap tree of Malpighiales inferred from 5113 gene trees using ASTRAL.

**Table S1** Voucher and GenBank information for 15 species in Malpighiales used for *de novo* transcriptome assembly.

**Table S2** Taxa sampled from the OneKP data set and the corresponding OneKP library ID.

**Table S3** Taxa sampled with complete genomes and corresponding reference.

**Table S4** Whole genome duplications (WGDs) in Malpighiales identified with complete taxon sampling, including their phylogenetic placement, percentage of gene families supporting gene duplication, log likelihood, and age estimations of each WGD.

**Table S5** Summary statistics for the transcriptome assembly in Malpighiales.

**Table S6** Whole genome duplications (WGDs) in Malpighiales identified with more conservative taxon sampling, including their phylogenetic placement, percentage of gene families supporting gene duplication, log likelihood, and age estimations of each WGD.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.