11-26-2018

A Malvaceae mystery: A mallow maelstrom of genome multiplications and maybe misleading methods?

Justin L. Conover
Iowa State University, jconover@iastate.edu

Nisa Karimi
University of Wisconsin-Madison

Noah Stenz
University of Wisconsin-Madison

Cécile Ané
University of Wisconsin-Madison

Corrinne E. Grover
Iowa State University, corrinne@iastate.edu

See next page for additional authors

A Malvaceae mystery: A mallow maelstrom of genome multiplications and maybe misleading methods?

Abstract

Previous research suggests that Gossypium has undergone a 5- to 6-fold multiplication following its divergence from Theobroma. However, the number of events, or where they occurred in the Malvaceae phylogeny remains unknown. We analyzed transcriptomic and genomic data from representatives of eight of the nine Malvaceae subfamilies. Phylogenetic analysis of nuclear data placed Dombeya (Dombeyoideae) as sister to the rest of Malvadendrina clade, but the plastid DNA tree strongly supported Durio (Helicteroideae) in this position. Intraspecific Ks plots indicated that all sampled taxa, except Theobroma (Byttnerioideae), Corchorus (Grewioideae), and Dombeya (Dombeyoideae), have experienced whole genome multiplications (WGMs). Quartet analysis suggested WGMs were shared by Malvoideae-Bombacoideae and Sterculioideae-Tilioideae, but did not resolve whether these are shared with each other or Helicteroideae (Durio). Gene tree reconciliation and Bayesian concordance analysis suggested a complex history. Alternative hypotheses are suggested, each involving two independent autotetraploid and one allopolyploid event. They differ in that one entails an allopolyploid origin for the Durio lineage, whereas the other invokes an allopolyploid origin for Malvoideae-Bombacoideae. We highlight the need for more genomic information in the Malvaceae and improved methods to resolve complex evolutionary histories that may include allopolyploidy, incomplete lineage sorting, and variable rates of gene and genome evolution.

Disciplines

Ecology and Evolutionary Biology | Genetics and Genomics | Plant Breeding and Genetics

Authors

Justin L. Conover, Nisa Karimi, Noah Stenz, Cécile Ané, Corrinne E. Grover, C. Skema, Jennifer A. Tate, Kirsten Wolff, Samuel A. Logan, Jonathan F. Wendel, and David A. Baum

Research Article

A Malvaceae mystery: A mallow maelstrom of genome multiplications and maybe misleading methods?

Running Title: WGMs in the Malvaceae

Justin L. Conover[1†], Nisa Karimi[2†], Noah Stenz[2,3], Cécile Ané[2,4], Corrinne E. Grover[1], C. Skema[5], Jennifer A. Tate[6], Kirsten Wolff[7], Samuel A. Logan[7], Jonathan F. Wendel[1] and David A. Baum[2,8]*

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, 50011, USA

[2]Department of Botany, University of Wisconsin Madison, WI, 53706, USA

[3]Current address: 7421 Allemande Way, Chattanooga, TN, 37421, USA

[4]Department of Statistics, University of Wisconsin Madison, WI, 53706, USA

[5]Morris Arboretum of the University of Pennsylvania, 100 E. Northwestern Avenue, Philadelphia, PA, 19118, USA

[6]Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand

[7]School of Natural and Environmental Sciences, Newcastle University, NE1 7RU, UK

[8]Wisconsin Institute for Discovery, 330 N Orchard St, Madison, WI 53715, USA

†These authors contributed equally to this work.
*Correspondence: dbaum@wisc.edu

Special Issue: Polyploidy and Evolution
Edited by:Yuxian Zhu, Wuhan University, China

Received: August 23, 2018; Accepted: November 18, 2018

Abstract

Previous research suggests that *Gossypium* has undergone a 5- to 6-fold multiplication following its divergence from *Theobroma*. However, the number of events, or where they occurred in the Malvaceae phylogeny remains unknown. We analyzed transcriptomic and genomic data from representatives of eight of the nine Malvaceae subfamilies. Phylogenetic analysis of nuclear data placed *Dombeya* (Dombeyoideae) as sister to the rest of Malvadendrina clade, but the plastid DNA tree strongly supported *Durio* (Helicteroideae) in this position. Intraspecific Ks plots indicated that all sampled taxa, except *Theobroma* (Byttnerioideae), *Corchorus* (Grewioideae), and *Dombeya* (Dombeyoideae), have experienced whole genome multiplications (WGMs). Quartet analysis suggested WGMs were shared by Malvoideae-Bombacoideae and Sterculioideae-Tilioideae, but did not resolve whether these are shared with each other or Helicteroideae (*Durio*). Gene tree reconciliation and Bayesian concordance analysis suggested a complex history. Alternative hypotheses are suggested, each involving two independent autotetraploid and one allopolyploid event. They differ in that one entails an allopolyploid origin for the *Durio* lineage, whereas the other invokes an allopolyploid origin for Malvoideae-Bombacoideae. We highlight the need for more genomic information in the Malvaceae and improved methods to resolve complex evolutionary histories that may include allopolyploidy, incomplete lineage sorting, and variable rates of gene and genome evolution.

INTRODUCTION

An exciting revelation from the genomics revolution is the large number of ancient whole genome multiplication (WGM) events that punctuate angiosperm evolution, with many plant lineages experiencing multiple WGMs over their histories (Adams and Wendel 2005; Cui et al. 2006; Van de Peer et al. 2009; Wood et al. 2009; Jiao et al. 2011; Bowers et al. 2013; Soltis et al. 2015; Van de Peer et al. 2017). Polyploidy is frequently considered a source of evolutionary innovation (Stebbins 1947; Grant 1971) leading to increased species diversification and phenotypic innovation (Soltis and Soltis 2016). Thus, WGMs just prior to the radiation of such families as the Asteraceae, Brassicaceae, Fabaceae, Poaceae, and Solanaceae have been assigned a causal role in those clade's species-richness (Paterson et al. 2004; Soltis et al. 2009; Cannon et al. 2015).

Despite the profound importance of WGMs in plant evolution, the identification and characterization of past events is often challenging. In particular, mechanisms such as rapid fractionation and rediploidization tend to remove evidence of the WGMs that precipitated them (Wendel 2015; Cheng et al. 2018). Maize, for example, was classically considered a pure diploid, but genomic research has since reclassified maize as a stabilized paleo-allotetraploid that rapidly jettisoned much of its duplicated genome (Gaut and Doebley 1997; Schnable et al. 2011; Zhao et al. 2017).

Multiple methods have been developed to infer ancient WGMs, each leveraging different, though sometimes overlapping, lines of evidence. Early inferences were based on replicated intragenomic synteny, such as seen in maize (Wendel et al.1986; Helentjaris et al.1988; Gaut and Doebley 1997) and *Arabidopsis* (Vision et al. 2000). Soon after, synonymous substitution "peaks" in EST distance data among paralogs within a genome were used to infer WGMs (Schlueter et al. 2004). Since then, methods including gene tree/species tree reconciliation (Ness et al. 2011), collinearity of full genome sequences (Tang et al. 2010), and statistical methods based on gene counts (Rabier et al. 2013) have been combined to uncover many WGM events, some of them very old.

Genomic evidence for WGM in the Malvaceae was first described in the publication of the *Gossypium raimondii* genome, where collinearity data supported an ancient 5- to 6-fold WGM sometime after its divergence from chocolate (*Theobroma cacao*), estimated at approximately 90 million years ago (Paterson et al. 2012). Subsequent analyses (using various lines of evidence) have attempted to infer whether this event consisted of two successive whole genome duplication/triplication events, or a single deca- or dodecaploidization event (Wang et al. 2016). In addition, studies have investigated if any other Malvaceae taxa show evidence of the same event(s), namely *Durio zibethinus* (Teh et al. 2017) and *Firmiana danxiaensis* (Chen et al. 2015)*.* The draft genome of *D. zibethinus* (Teh et al. 2017) suggested a shared WGM history with *G. raimondii*; however, while the 28 chromosomes of *D. zibethinus* show remarkable (hexaploid, 3x) synteny with respect to the 10 chromosomes of *T. cacao*, this inference is insufficient to explain the decaploid (5x) or dodecaploid (6x) multiplication history of *G. raimondii* relative to *Theobroma*.

In this study, we used comparative genomics to further explore WGM(s) in the Malvaceae. The Malvaceae includes nine subfamilies (Bombacoideae, Brownlowioideae, Byttnerioideae, Dombeyoideae Grewioideae, Helicteroideae, Malvoideae, Sterculioideae, and Tilioideae), several of which were at one time treated at the family rank (Alverson et al.1999; Bayer et al.1999). Phylogenetic analyses have consistently supported a sister relationship between Bombacoideae and Malvoideae, in a clade called Malvatheca (Baum et al. 1998; Alverson et al. 1999; Bayer et al.1999; Baum et al. 2004). Likewise, all subfamilies except Grewioideae and Byttnerioideae, the latter containing *Theobroma*, have been grouped into a large clade, Malvadendrina, which is supported by a unique and unreversed 21-bp deletion in the plastid-encoded *ndhF* gene (Alverson et al. 1999). Relationships between Malvatheca and the other five subfamilies of Malvadendrina remain uncertain (Nyffeler et al. 2005).

Here, we employed transcriptome and genome sequences for representatives of eight subfamilies (all but Brownlowioideae) to investigate the phylogeny and timing of WGM events. We analyzed distributions of synonymous divergence (Ks) to detect evidence of ancient WGMs, and inferred orthologous nuclear gene trees to test alternative possibilities for the timing and placement of WGMs within the family. Collectively, our analyses suggest a complicated history and considerable residual uncertainty. This study highlights the challenges of inferring WGM events in the face of reticulation and other phenomena that impact species-tree inference. Nonetheless, we propose alternative hypotheses for the history of the Malvaceae and suggest how these might ultimately be distinguished through additional sampling.

RESULTS

Novel transcriptomic data sets

To complement already published genome and transcriptome data (see Methods and Supplemental Table 1 for summary of taxa sampled and references), we generated transcriptomes for *Tilia cordata* (Tilioideae), *Adansonia digitata* (Bombacoideae)*,* and *Bombax ceiba* (Bombacoideae), although the *Bombax* transcriptome was superseded by the recent publication of a genome sequence for *Bombax ceiba* (Gao et al. 2018). Raw data for these three transcriptomes are available on NCBI Sequence Read Archive, under accession PRJNA493960.

After quality filtering, 141,394, 109,236, and 168,390 scaffolds were assembled *de novo* for *Adansonia*, *Bombax*, and *Tilia*, respectively. The combined total number of bases was 71,972,881 for *Adansonia* with a mean contig length of 577 bp*,* 50,737,486 bases for *Bombax* with a mean length of 509 bp, and 66,261,905 bases for *Tilia* with a mean length of 449 bp. After Transdecoder filtering, the number of gene models recovered for *Adansonia*, *Bombax* and *Tilia* were 54,054, 39,027, and 51,336, respectively (Table S1). BUSCO recovery results for all assembled transcriptomes ranged from 17.6% complete BUSCO genes present (*Dombeya*) to 81.5% (*Corchorus*), whereas the proportion of missing BUSCO genes ranged from 9.7%

(*Corchorus*) to 57.2% (*Dombeya*) (Table S1). For the sake of readability, we will generally refer to each taxon by generic name.

Malvaceae phylogeny inferred by plastid genes

For eight taxa, genomic DNA-based plastomes were available, either in published genome sequences or from nuclear targeted sequence capture data sets (*Adansonia* and *Bombax*). For three taxa (*Dombeya*, *Heritieria*, and *Corchorus*), DNA sequences were not available, so we instead mapped RNAseq reads to the published *Theobroma* plastid genome sequence. We aligned 67 plastid-encoded genes that were not located in the inverted repeat regions and were present in all taxa sampled in this study (see Methods). The tree was rooted with the published plastid sequence of *Arabidopsis thaliana.*

As genes derived from RNAseq may be affected by RNA editing, we were concerned that mixing transcriptomic and genomic data could create artefactual synapomorphies. To investigate this possibility, we analyzed three different datasets: genomic sequences only (Figure 1A), genomic sequences plus *Dombeya* (Figure 1B), or all sequences combined (Figure 1C). The first two datasets yielded trees where all internal nodes had high bootstrap support (>75), whereas the dataset including all samples had a similar topology as the other two, but markedly lower bootstrap values on internal nodes, perhaps indicative of RNA-editing-mediated attraction. After rooting between *Theobroma* (Byttnerioideae) and *Corchorus* (Grewioideae) and the remaining Malvaceae taxa, all three phylogenies placed *Durio* as sister to the remaining lineages of Malvadendrina: *Dombeya* (Dombeyoideae), *Tilia* (Tilioideae), *Firmiana* (Sterculioideae), *Heritiera* (Sterculioideae), and Malvatheca (*Gossypium*+*Hibiscus*, Malvoideae and *Adansonia*+*Bombax*, Bombacoideae)*.*

A clade composed *of Tilia* (Tilioideae) and *Firmiana*+*Heritiera* (Sterculioideae) was sister to the well-supported Malvatheca clade. Of note, the internodes near the base of Malvadendrina are extremely short, suggesting a rapid radiation of the six subfamilies and the potential for a large amount of incongruence due to incomplete lineage sorting. Combined with other source of phylogenetic noise, such as homoplasy, gene conversion, and orthology-paralogy issues, these short branches help to explain prior difficulties in resolving subfamilial relationships.

To minimize dependence on transcriptomic datasets, and because the topology of the three trees did not change between datasets, we pruned *Corchorus* and *Heritiera* and used the resulting tree for downstream analyses. This pruning is further justified because the phylogenetic placement of *Corchorus* (Grewioideae, outside of Malvadendrina; Baum et al.

1998; Alverson et al. 1999; Bayer et al.1999) and *Heritiera* (in the same subfamily as *Firmiana* [Sterculioideae]; Wilkie et al. 2006) are uncontroversial.

Ks histograms identify subfamilies with a history of WGM

If a species were to have undergone a WGM event in its past, a large proportion of the paralogs would originate from this event. Therefore, when the synonymous rates of divergence (Ks) between paralogs is plotted against frequency, a WGM event may be evident as a “peak” in this distribution. The position of this peak along the Ks axis provides a proxy for divergence time. In autopolyploids this divergence time coincides with an estimate of when autopolyploidy occurred. In allopolyploidy, this peak coincides with the time of divergence of the progenitor diploids that subsequently hybridized.

Ks distributions of all pairwise combinations of paralogs in a single genome/transcriptome showed no evidence of a WGM in *Theobroma*, *Corchorus*, or *Dombeya,* whereas all other species showed one or two supplemental peaks (Figure 2). Six taxa each had a single peak, with Ks medians of ~0.2 (*Tilia*), ~0.3 (*Durio*, *Firmiana*, *Heritiera*) or ~0.4 (*Bombax, Adansonia*), whereas the two members of Malvoideae (*Gossypium* and *Hibiscus*)*,* each had two peaks at ~0.1/0.5 and ~0.3/0.6, respectively. We interpret the younger peak in *Gossypium* as part of the background gene duplication rate, most likely due to tandemly duplicated genes that retain high sequence similarity to one another, via gene conversion (Panchy et al. 2016; Train et al. 2017). The younger peak in *Hibiscus,* in contrast, is likely the result of a recent WGM unique to this lineage, as previous work suggests this was an independent event with respect to *Gossypium* (Kim et al. 2016; Teh et al. 2017). Furthermore, we established *Hibiscus* to have twice as many gene models as other taxa sampled (Table S1).

Allowing for differences in rates of molecular evolution across lineages, which can be considerable among Malvaceae subfamilies (Baum et al. 2004), the pattern could be suggestive of a single shared WGM in all Malvadendrina subfamilies except Dombeyoideae (Kim et al. 2016; Teh et al. 2017). The peaks of these distributions closely coincide with Ks divergences associated with the divergence of *Theobroma, Durio, Gossypium, and Bombax* (see Methods), indicating that the WGM event(s) we detected occurred early in the group’s radiation. However, the subfamilies that have experienced a WGM in Malvaceae are not monophyletic on the plastid tree, which places *Durio* rather than *Dombeya* as sister to the rest of Malvdendrina. This indicates either that WGM events occurred independently in *Durio* and other Malvadendrina, or there was deep allopolyploidy, or that the plastid DNA has a discordant evolutionary history (e.g., due to reticulation or incomplete lineage sorting).

Identification of shared whole genome multiplications using gene quartets

To help assess whether the Ks peaks represent the same WGM event shared between pairs of species, we created gene trees consisting of two paralogs with a Ks distance of 0.2-0.6 from each species (i.e. four genes total), tabulated whether inferred tree topologies were consistent with either a shared duplication (i.e., conspecific paralogs not sister to each other) or separate duplications (i.e., conspecific paralogs sister to each other).

The frequency of trees for each taxon pair suggestive of separate duplication events varied from well below to well above the 33% expected for random trees (Table 1). Evidence for shared duplications between *Hibiscus* and any other taxon is the least supported. This is likely because *Hibiscus* has experienced an independent WGM since its divergence from *Gossypium* (Kim et al. 2016) such that paralogs from this recent WGM event obscure evidence of earlier events shared with other taxa. With this exception, the gene topologies consistently support shared duplications for *Gossypium* and *Adansonia-Bombax* (Bombacoideae-Malvoideae) and for *Firmiana+Heritiera+Tilia* (Sterculioideae-Tilioideae). The signal for a duplication shared between these two groups and *Durio* (Helicteroideae) is equivocal, with the number of genes supporting separate or shared duplications being close to the null expectation.

Multi-labeled trees (MUL-trees) reveal hybridization and allopolyploidy events

We used GRAMPA (Gregg et al. 2017) to place WGM event(s) on the Malvaceae phylogeny, inferred from both the nuclear and plastid genes, to determine whether these were allo- or autopolyploidy events. GRAMPA uses a parsimony approach to reconcile multiply-labeled gene trees (which include paralogs) with various species trees. The method searches for multiply-labeled trees that minimizes the sum of the implied number of gene duplications and gene losses; i.e., the reconciliation score. The MUL-trees derived when using the plastid phylogeny as the species tree had lower reconciliation scores and were used instead of the nuclear topology for downstream analysis (although each yielded similar conclusions).

Using the 12,426 gene trees extracted from either the transcriptomic or genomic datasets, GRAMPA identified 53 MUL-trees (File S1) that had a lower reconciliation score, across all input gene trees, than did the single-labeled species tree (Figure 3). The twenty topologies with the lowest reconciliation scores suggested that an allopolyploidization event generated either a Malvatheca+Sterculioideae+Tilioideae clade (six MUL-trees), Malvatheca alone (seven MUL-trees), or Malvoideae alone (seven MUL-trees). For each of these possible allopolyploidization scenarios there is conflict regarding the parents involved.

One possible confounding factor in these analyses is the high rate of implied "deletion," which includes a lack of expression in the species represented only by transcriptomic data. Therefore, we repeated the analyses with only those taxa that had published genome sequences (*Theobroma*, *Gossypium*, *Bombax*, and *Durio*), plus our *de novo* transcriptome assembly of *Dombeya* and *Firmiana*. To ensure proper rooting of the gene trees, we also included the published genome sequence of *Carica papaya* L. (*Brassicacee*), which is not reported to have undergone any independent WGM events (Ming et al. 2008).

To be conservative, we restricted our dataset to only include those 1,032 gene trees in which all internal nodes had at least 90% bootstrap support (see Materials and Methods). GRAMPA identified seven MUL-tree topologies that fit the data better than a single-labeled species tree matching the optimal plastid tree (File S2). The three MUL-tree topologies with the lowest reconciliation scores suggest that Malvatheca (*Gossypium* and *Bombax*) arose via an allopolyploidization event between an ancestor of Malvatheca and either an extinct lineage sister to Malvadendrina, or to Malvadendrina minus *Durio*, or to *Durio* (Figure 3). The fourth lowest scoring tree topology suggested that the clade consisting of Malvatheca + *Firmiana* was formed, via an allopolyploidy event between an ancestor of the clade and a member of the *Durio* lineage.

The remaining three topologies suggested a hybrid origin of clades including *Dombeya*, which based on Ks plots, does not have evidence of WGM in its evolutionary history. We attribute these topologies to an artifact of GRAMPA, possibly due to the very short internodes in the species tree and many other possible causes of gene tree incongruence, including erroneous splitting of isoforms and/or paralogs during transcriptome assembly, incomplete lineage sorting (ILS), and homoplasy.

Species tree inference: Bayesian concordance analysis and maximum likelihood inference of nuclear genes

The short internal branches of the species tree, possibly combined with a history of allopolyploidy, should lead us to expect a large amount of phylogenetic incongruence among single-copy gene trees. One way around this is to use genes that are represented by a single gene in each species (presumably because they are lost rapidly after each WGM). However, because only two orthogroups were identified once, and only once in all taxa, we focused on a smaller taxon set, which included the published genomic data of four genera (*Theobroma*, *Gossypium*, *Bombax*, and *Durio*) as well as transcriptomic data from *Dombeya*. We included *Dombeya*, even though it does not contain published genomic information, because it is the lone sampled Malvadendrina taxon whose Ks plot does not show evidence of a WGM. *Theobroma* was treated as the outgroup.

We identified 1,214 singleton groups (genes present as only one copy in each taxon sampled). We conducted Bayesian concordance analysis using BUCKy (Ané et al. 2007; Larget et al. 2010), which, taking account of uncertainty in individual gene trees, estimates the optimal population ("species") tree under the assumption that all discordance is due to incomplete lineage sorting (ILS). The resulting primary concordance (Baum 2007) and population trees had the same topology (Figure 4A). There is strong support for a Malvatheca clade composed of Malvoideae (as represented by *Gossypium*) and Bombacoideae (represented by *Bombax*) with a concordance factor (CF) of 0.935, with a 95% credibility interval [CI] of 0.921-0.948, and a branch length of 2.615 coalescent units. The other internal branch is short (0.067 coalescent units) and supports a Malvatheca + *Durio* clade with CF = 0.365 [0.336-0.395].

Interestingly, the two other possible resolutions of this node are not at equal frequency as expected if all discordance were due to ILS: the *Durio* + *Dombeya* (CF = 0.342 [0.315-0.370]) resolution does not have a significantly lower CF than the optimal resolution, whereas Malvatheca + *Dombeya* (CF = 0.263 [0.236, 0.291]) does. The observation that one of three possible resolutions of a node has a significantly lower CF than the other two cannot readily be explained by ILS and instead suggests hybridization (Rodriquez et al. 2009; Ané 2010). Specifically, this result could be explained by a population tree with the topology (*Theobroma,* (*Dombeya*, (*Durio*, Malvatheca))) with subsequent gene flow between the *Durio* and *Dombeya* lineages, perhaps involving allopolyploid hybridization between the *Dombeya* and ancestral *Durio* lineages to generate an allopolyploid *Durio* lineage. This hypothesis, and why it did not emerge from GRAMPA analysis is discussed further below.

To further characterize the possible scenarios of hybridization suggested by BUCKy and GRAMPA, we identified 381 orthogroups that are present in the genomes of *Gossypium*, *Theobroma*, *Durio*, and *Bombax* plus the transcriptome of *Dombeya*, and have every node supported by a maximum likelihood bootstrap support >80%. We then used a custom R script to recursively combine sister groups from the same species (treating Malvatheca as a single species). Of the 381 trees, 265 reduced down to a tree with one tip per taxon, and of these, 113 trees had the rooted topology (*Dombeya*, (*Durio*, Malvatheca)), 95 had (Malvatheca, (*Durio*, *Dombeya*)), and 57 had (*Durio*, (*Dombeya*, Malvatheca)). This distribution is significantly different from 1/3 proportions (p-value=9.584e-05), although the frequency of the first two is not significantly different. Thus, despite using a different (though overlapping) gene tree set than was analyzed in BUCKy, we see a very similar pattern: significantly more trees supporting either *Dombeya* or Malvatheca as sister to the rest of Malvadendrina over a *Durio*-sister topology.

Modeling orthologous gene family sizes

Analyses of gene trees provides rich information for a subset of gene families, but a majority of families are ignored because they yield inadequately resolved gene trees or because the interpretation of gene trees is confounded by a history of gene duplication and gene loss. Therefore, as a complement to gene topology-based methods, we conducted analysis of gene family size, which can also provide information on WGM events (Rabier et al. 2013; Parks et al. 2018).

Using the four published genome sequences of *Gossypium, Theobroma, Durio,* and *Bombax*, we used OrthoFinder to identify 19,621 gene families, with an average number of genes/family/genome ranging from 1.13 to 1.58 (Table S2, column A). Excluding the 1,368 gene families that are not represented by at least one gene in all taxa, increases the number of genes/family, especially in *Theobroma* (Table S2, column B). Further excluding the 68 clusters that had 31 or more genes in total (the largest cluster had 94 genes) had only a modest effect on the average gene family size in each species (Table S2, column C).

The R-program WGDgc (Rabier et al. 2013) was used to analyze the resulting gene count data and test alternative models for the distribution of WGMs over the three terminal and two internal branches of the rooted four-taxon tree used. While the tree topology is not in doubt, relative branch-lengths could significantly affect the conclusions because WGDgc explicitly models rates of gene duplication and loss. We used two alternative sets of branch lengths: (1) the branch lengths in substitutions/site as estimated from concatenated analysis of the singleton genes, and (2) branch length in substitutions per unit time, as inferred by imposing ultrametricity using penalized likelihood.

When raw branch lengths are used, the best model supports three separate WGM events (Figure 5A): one whole genome duplication (WGD) shared by *Gossypium* + *Bombax*, one whole genome triplication (WGT) specific to *Bombax*, and one WGT specific to *Durio*. When the tree is rendered ultrametric the best model suggests four separate events (Figure 5B): a WGT+WGD specific to *Gossypium*, a WGT specific to *Bombax*, and a WGT specific to *Durio*. Hypotheses with shared events are disfavored (see Tables S3, 4 for AIC differentials), but the best such scenario has one WGT event shared among the ingroup taxa (*Durio, Bombax* and *Gossypium*) and another event (WGD) specific to *Bombax*. Detailed results for likelihood, Alkaike Information

Criterion (AIC), and Bayesian Information Criterion (BIC) of all models can be found in Tables S2, 3.

DISCUSSION

Since the discovery that the genome of *Gossypium raimondii* has an evolutionary history that includes WGMs (Paterson et al. 2012), it has been unclear whether WGMs detected in *Gossypium* are independent or shared with other clades of Malvaceae. Here, we attempted to ascertain the number and phylogenetic placements of WGM events, using a combination of tools, including nuclear and plastid phylogenetic reconstructions, Ks plots, quartet tree inferences, a reconciliation method (GRAMPA), Bayesian concordance analysis, and a method based on gene copy number variation. We anticipated integrating these multiple approaches to reconcile our results with those of previously published studies. However, there was no agreement among approaches, leaving considerable uncertainty as to the history of genome evolution in the Malvaceae. We discuss some of the better-supported alternative hypotheses and note some methodological issues that complicate WGM inference, including ILS, asymmetrical gene fractionation, varying rates of evolution, and gene conversion (Panchy et al. 2016; Train et al. 2017).

Nuclear-plastid discordance

To investigate the phylogenetic placement of such WGM events, we first sought to infer a phylogeny of the Malvaceae using multiple plastid gene sequences, as previously explored by La Duke and Doebley (1995) and Nyffeler et al. (2005), among others. We compared this to the trees inferred by 1214 nuclear genes for a core set of taxa. The nuclear and plastid data yielded similar trees supporting Malvatheca (Malvoideae + Bombacoideae) sister to Sterculioideae + Tilioideae (represented by *Firmiana* + *Tilia,* respectively). They differed in having either *Dombeya* (Dombeyoideae) or *Durio* (Helicteroideae) as strongly supported sister group to the rest of the Malvadendrina for the nuclear and plastid data, respectively.

One possible explanation of this discordance is that our plastome tree does not track the dominant evolutionary history. Due to the lower effective population size of plastid DNA, however, plastomes should be less prone to ILS. Nonetheless ILS remains a possibility, as do other sources of phylogenetic conflict, such as errors in tree estimation. For example, it is possible that, despite high bootstrap support, the short and ancient internodes of the plastid tree were incorrectly resolved due to artifacts such as long-branch attraction (Bergsten 2005; Qu et al. 2017). That said, it is reasonable to assert that the plastid tree is the best current estimated of at least the maternal side of what appears to be a complex history of reticulation and genome multiplication.

Bayesian concordance analysis of singleton nuclear orthologs documented extensive phylogenetic incongruence, much of it presumably due to ILS along short branches near the base of Malvadendrina. These short branches, in fact, are expected under the scenario of a rapid radiation, which the plastid tree also implies. Although BUCKy does not directly return a credibility interval (CI) on a population tree topology as a whole, by looking at the CIs of conflicting concordance factors it is possible to evaluate the implied population tree. In this case credibility interval analysis on a small subset of taxa with genome sequences (The outgroup *Theobroma* and ingroup taxa *Durio, Bombax*, and *Gossypium*), plus transcriptome-derived data from *Dombeya*, allowed rejection of a topology in which *Durio* and *Dombeya* formed a clade, but could not distinguish between the alternative topologies in which either *Durio* or *Dombeya* is sister to Malvatheca (*Bombax* plus *Gossypium*). This pattern is most easily explained by a history of reticulation (discussed below). It should be borne in mind, however, that WGM followed by different rates of gene loss has the potential to cause erroneous inferences of orthology, which could potentially lead to systematic errors in population tree inference.

Ks plots and quartet analyses imply shared WGMs

The release of a draft genome sequence of *Hibiscus syriacus* (Kim et al. 2016) indicated two independent WGD events since its divergence from *G. raimondii*. We detected two broad peaks in the Ks plot for *H. cannabinus*, which has less than half as many chromosomes as *H. syriacus* (CCDB online: http://ccdb.tau.ac.il/), leading us to infer that at least the older Ks peak is a WGM shared with *Gossypium*. Likewise, phylogenetic inference, quartet analysis, and Ks plots suggest that Bombacoideae (*Bombax* and *Adansonia*) shares at least one WGM with *Gossypium*. However, while the two other subfamilies closely related to Malvatheca, namely Tilioideae and Sterculioideae, have experienced at least one WGM, our results are equivocal as to whether this was shared with Malvatheca (i.e., Malvoideae + Bombacoideae).

The Ks approach identified a Ks peak in most species, corroborating prior analyses (Paterson et al. 2012; Chen et al. 2015; Kim et al. 2016; Teh et al. 2017; Gao et al. 2018). While the modes of these distributions differ between species, the full distributions overlap markedly among species, and also overlap with speciation estimates. These results are compatible with a shared WGM event early in the radiation of Malvadendrina, but multiple separate events cannot be ruled out. Furthermore inference from Ks plots alone should be treated with caution. As a case in point, there are both known and suspected recent WGM events in our phylogeny that are not visible in the Ks plots. For example, *Adansonia digitata* is a recent tetraploid (Baum & Oginuma 1994), but no recent peak is observed. One possibility is that this was an autotetraploidy event that occurred recently enough that tetrasomic inheritance is still in effect, which could prevent paralogs diverging enough to result in two separate gene models during transcriptome assembly (Scott et al. 2016).

There is also a suspected WGM in the lineage leading to *Dombeya burgessiae* based on chromosome counts (2n = 46, 54; Seyani 1991) that are more than double that reported for some other Dombeyoideae (e.g., Corchoropsis with n=10; Tang 1992), but this is not visible in the Ks plot except, perhaps, in the relatively high frequency of paralogs with small Ks values (0.0-0.1). We note that the transcriptome assembly for *Dombeya* was the lowest quality assembly of our sampled taxa, highlighting that there should remain a degree of caution in interpreting results based solely on transcriptomic data. Additional high-quality genome sequences, especially for *Dombeya*, would help overcoming the limitations of transcriptome-based analyses.

Modeling WGMs by gene counts

Whereas Ks plots, quartet analyses, and GRAMPA analyses suggest some sharing of WGM events among subfamilies, gene cluster size analyses generally supported separate WGM events. We suspect that this is a failure of the WGDgc method resulting from errors during clustering and orthology detection, which could lead to violations of model assumptions. The model of gene size evolution assumes a constant rate of gene turnover (gene loss and small scale gene duplications) within and among lineages. The model seems sensitive to this assumption, as evidenced by the different results on trees with different branch lengths. The different lineages might have different rates of gene deletion and the assembled genomes also differ in quality, which could inflate implied deletion rates in less well validated genomes. As a case in point, the preference for more WGM events in the *Bombax* lineage than in the *Gossypium* lineage might be caused by a smaller number of genes per family in *Gossypium* than in any other ingroup species.

Another assumption of the gene count model is that, after each WGM, fractionation occurs rapidly compared to the rate of speciation. Here, we found evidence of rapid diversification between subfamilies, resulting in short internal branch lengths. Consequently, fractionation might have spanned one or more rapid speciation events, violating the model assumption. This might explain the prevalence of non-shared WGM events in WGDgc results. Thus, pending availability of more high-quality genome assemblies, we consider gene family size to provide insufficient information for resolving WGM events in Malvaceae.

Hypotheses to reconcile the conflicting signals from the different data sets

The foregoing synopsis explains why there might be reasons to doubt inference coming from Ks plots or gene-count based analysis. However, the discrepancy between the analysis of singly-labelled gene trees (BUCKy and gene tree counting) and multiple-labelled gene trees (GRAMPA) is less easily discounted. Specifically, as discussed below and illustrated in Figure 6, these two approaches suggest somewhat different hypotheses to account for the patterns observed in our data, both involving shared WGMs and allopolyploidy.

Hypothesis one

The first hypothesis to explain these data is that soon after the initial split separating *Dombeya* (Dombeyoideae) from the remaining Malvadendrina lineages, the Malvadendrina ingroup lineage underwent autopolyploidization to generate a tetraploid. The *Durio* lineage then arose from an allopolyploid event involving a diploid ancestor of *Dombeya* and a tetraploid member of the Malvadendrina ingroup, yielding the hexaploid genome of *Durio* (Figure 6A). Such an allopolyploidization event is plausible if the internodes were short, as illustrated by cotton, for example, where modern A and D genomes are on different continents, last shared a common ancestor 5+ MYA, have about 2.2% divergence in genic regions (Page et al. 2013), and yet still can form hybrids. To explain the decaploid or dodecaploid genome of *Gossypium* (Wang et al. 2016) we would hypothesize a subsequent auto-triplication in the Malvoideae (i.e., unique to cotton and *Hibiscus* among the species sampled), or deep in the Malvatheca (i.e., shared with Bombacoideae).

This hypothesis provides an explanation for why all Malvadendrina except *Dombeya* (Dombeyoideae) have a peak in Ks plots. Additionally, collinearity analysis shows that at least some duplicate chromosomal arms are shared between *Durio* and *Gossypium* (Teh et al. 2017), which supports such a shared WGM. This hybridization model is also consistent with the BUCKy results, since it explains why 1) a plurality of genes support a *Durio*+Malvatheca clade, which is the expected pattern for *Durio* genes from the tetraploid Malvatheca parent, 2) a significant subset of genes support a *Durio*+*Dombeya* clade, which is the expected pattern for *Durio* genes from the diploid dombeyoid parent, and 3) there are many fewer genes showing a *Dombeya*+Malvatheca clade, which can only arise by ILS. On the other hand, this hypothesis would predict a plastome phylogeny in which either *Dombeya* or *Dombeya*+*Durio* was sister to the remainder of Malvadendrina, rather than the observed *Durio*-sister topology. It should be borne in mind, however, that the plastid DNA tree is inferred from a single locus, which is subject to ILS, and might, therefore, not match the population tree.

Hypothesis two

A second hypothesis proposes an allopolyploid origin for Malvatheca (see Figure 6B). The initial split within Malvadendrina was between Dombeyoideae-Sterculoideae-Tilioideae and Helicteroideae-Malvatheca. Thereafter, the Helicteroideae-Malvatheca lineage experienced a WGT and, after Sterculoideae-Tilioideae diverged from the Dombeyoideae, Sterculoideae-Tilioideae experienced an independent WGD. Finally, Malvatheca was formed via allopolyploidization between the hexaploid ancestor of Helicteroideae and the tetraploid ancestor of Sterculoideae+Tilioideae to form a decaploid Malvatheca lineage. If the Sterculoideae+Tilioideae lineage was the maternal donor in this allopolyploid hypothesis, then the expected plastid tree would match the inferred topology in which *Durio* alone is sister to the remainder of Malvadendrina.

This hypothesis is one of the scenarios supported by GRAMPA, using both transcriptomic and genomic data, which identified the Malvatheca clade as having an allopolyploid origin. It readily explains the Ks peaks in all of Malvadendrina except *Dombeya*; the observation that *Durio* has undergone an ancient hexaploidy event after its divergence from *Theobroma*; that *Gossypium* has undergone an ancient 5- or 6-fold multiplication; and that at least one of the paleopolyploidy event(s) in cotton must have been an allopolyploid event that resulted in biased fractionation (Renny-Byfield et al. 2015). Additionally, the pattern of WGMs is consistent with our plastid DNA tree. While this hypothesis is contradicted by the BUCKy results of a Dombeya-Malvatheca clade being significantly rarer than other resolutions, this could be attributed to noise or an artifact, for example caused by using a poor quality *de novo* transcriptome assembly for *Dombeya*.

CONCLUSIONS

Here we have illustrated the application of a number of approaches to dissect the complexities involved in phylogenetic inference for taxa that have experienced multiple polyploidization and hybridization events. Our analyses demonstrate that even when large-scale genomic and transcriptomic datasets are analyzed, with the most current methodologies, teasing apart the true history of WGM events is a challenging endeavor, and that large ploidy increases magnify these difficulties.

Even in sampling genomic-scale data for eight of the nine subfamilies in the Malvaceae, it is still uncertain as to whether cotton has undergone a 5- or 6-fold multiplication (Paterson et al. 2012; Wang et al. 2016), and whether this same history is shared by other subfamilies in the Malvaceae. Nonetheless, our data suggest two alternative hypotheses, which are quite different despite both invoking two autotetraploidy events and one allopolyploidy event.

There is reason to be hopeful that future work can determine which of these is correct. A higher quality transcriptome or genome assembly for *Dombeya* could resolve whether it has a history of WGM. Furthermore, genome sequences of additional exemplar species are likely to be able to distinguish our main competing models, one of which suggests independent WGMs in Malvatheca and Sterculioideae/Tilioideae and also that *Durio* is an allopolyploid, whereas the second model suggests that Malvatheca, Sterculioideae, and Tilioideae share a WGM and that members of Malvatheca are allopolyploid. Such work would not only help clarify the history of WGM in the Malvaceae, but would also help the community better understand causes of disagreement among methods used to study genome evolution.

MATERIALS AND METHODS

Sources of transcriptomic/genomic data

We used published genome sequences for *Gossypium raimondii* L. (Malvoideae; Paterson et al. 2012), *Theobroma cacao* L. (Byttnerioideae; Motamayor et al. 2013), and *Durio zibethinus* L. (Helicteroideae; Teh et al. 2017). For the *Dombeya burgessiae* Gerrard ex Harv. & Sonder (Dombeyoideae) transcriptome, RNA from pooled leaves and floral parts was sequenced as part of the 1KP project (Johnson et al. 2012; Matasci et al. 2014; Wickett et al. 2014; Xie et al. 2014).

RNA-Seq data were downloaded from the NCBI Sequence Read Archive for *Durio zibethinus* (Helicteroideae; SRR6040092; Teh et al. 2017), *Firmiana danxiaensis* H.H.Hsue & H.S.Kiu (Sterculioideae; PRJNA274165; Chen et al. 2015) and *Heritiera littoralis* Aiton (Sterculioideae; SRR5138318; Dassanayake et al. 2009), *Hibiscus cannabinus* L. (Malvoideae; SRR2089299; Li et al. 2016)*,* and *Corchorus capsularis* L. (Grewoideae; SRR2089352; Zhang et al. 2015). Table S1 summarizes transcriptome assembly statistics.

Transcriptomes from *Adansonia digitata* L. and *Bombax ceiba* L. were generated at the University of Wisconsin - Madison. RNAs were extracted from greenhouse-grown leaves of

Adansonia digitata and Bombax ceiba using an optimized CTAB extraction protocol based on (Chang et al. 1993) and the Qiagen RNeasy Mini Kit (Cat #74106). Illumina TruSeq RNA libraries were prepared at the University of Wisconsin- Madison Biotechnology Center (Madison, WI). DNAs were purified using Agencourt AMPure XP beads (Beckman Coulter, USA). Quality and quantity of the finished libraries were assessed using an Agilent DNA1000 chip (Agilent Technologies, Inc., Santa Clara, California, USA) and Qubit® dsDNA HS Assay Kit (Invitrogen, Thermo Fisher Scientific, USA), respectively. Libraries were standardized to 2 nM. Cluster generation was performed using standard Cluster Kits and the Illumina Cluster Station (Illumina Inc., San Diego, CA, USA). Paired-end, 100bp sequencing was performed, using standard SBS chemistry on an Illumina HiSeq2000 sequencer. The *Tilia cordata* Mill*.* transcriptome (Tilioideae) was derived from leaf RNA prepared at Newcastle University, United Kingdom, and sequenced at the W.M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign, USA. RNAseq libraries were prepared with the Illumina TruSeq Stranded RNAseq Sample Prep kit (Illumina Inc., San Diego, CA), with 250 nt target insert size, quantified by qPCR and, sequenced on one lane for 101 cycles from each end of the fragments on a HiSeq2000 using a TruSeq Stranded RNA Sample Preparation Kit (Illumina Inc., San Diego, CA). Fastq files were generated and demultiplexed with Casava 1.8.2 (Illumina Inc., San Diego, CA). These three transcriptomes generated are available on NCBI Sequence Read Archive accession PRJNA493960.

Raw reads were quality trimmed and removed of adapter sequences using TrimGalore (Krueger 2015) and assembled with SOAP-Denovo-Trans (Luo et al. 2012) with Kmer size of 51. Open reading frames were identified with Transdecoder (Haas et al. 2013), and only those ORFs with sequence similarity to those of the published *Gossypium raimondii* (Paterson et al. 2012) reference genome annotation were kept for further analysis. Quality of transcriptome assembly was assessed using BUSCO v2 and the Embryophyta odb9 database (Simão et al. 2015).

These data were complemented by published genome annotations for *G. raimondii* (Malvoideae) and *Theobroma cacao* (Byttneroideae), obtained from Phytozome version 9.1.

Phylogenetic analysis of plastid DNA

Published plastid DNA sequences were downloaded from NCBI (Figure S2), for all taxa in this study, including *Arabidopsis thaliana* as an outgroup, except for the following taxa: *Adansonia*, *Bombax*, *Corchorus*, *Dombeya*, and *Heritiera.* Data for *A. digitata* and *B. ceiba* were generated by mapping nuclear targeted sequence capture reads (Karimi et al. unpublished) to the published plastid genome of *Theobroma* as described below. For the remaining three taxa for which no published plastid DNA sequences were available for the loci of interest (*Heritiera, Corchorus, and Dombeya*), we utilized our transcriptomic data. Trimmed RNAseq reads from these remaining taxa were mapped to the *Theobroma* published plastid genome using HISAT version 2.0.4 (Kim et al. 2015). Alignment (bam) files were converted into fasta files using *bam2consensus* (Page et al. 2014). Genes which were present in all taxa in single copy (i.e., not encoded in the IR regions of the plastid genome) were retained, and any gap sites were excluded to decrease phylogenetic errors caused by missing data. Nucleotide sequences were aligned with MAFFT (Katoh et al. 2002) using FFT-NS-2.

We performed phylogenetic analyses on three concatenated data sets: 1) genomic-derived sequences only (excluding taxa from RNA-Seq data, i.e., no *Dombeya*, *Heritiera*, or *Corchorus*), 2) genomic-derived sequences plus *Dombeya* (Dombeyoideae), 3) all sequences combined. We compared topologies from these three data sets given that transcriptome data are subject to RNA editing while samples represented by genomic data are not; thus the inclusion of more than one sample with RNA editing could potentially alter tree topology and bootstrap support due to the introduction of artificial homoplasy (Bowe and dePamphilis1996). For the concatenated datasets, maximum likelihood phylogenetic inference was performed with RAxML version 8.2.10 (Stamatakis 2014) using the GTR + Γ model (as determined by jModelTest2; Darriba et al. 2012) and 100 bootstrap replicates.

Ks distributions

We used full predicted gene regions for those taxa that have published genome sequences (*G. raimondii*, and *Theobroma cacao*) and full coding region predictions from Transdecoder (Haas et al. 2013) for all remaining taxa. We clustered the predicted protein sequences of all open reading frames (ORFs) from the *de novo* assembled transcriptomes (Table S1) with the published protein sequences of *G. raimondii* and *T. cacao* using OrthoFinder version 2.1.2 (Emms & Kelly 2015) to create orthogroups, groups of homologs (i.e., groups containing both orthologs and paralogs) defined by inclusion of the outgroup *Theobroma*. For all orthogroups that contained more than one gene for a given species (i.e., potential paralogs), we used every possible pairwise combination of genes and calculated the synonymous site (Ks) difference of these genes using the codeml package of PAML (Yang 1997). We then created species-specific distributions of Ks values to look for peaks that may indicate a WGM event.

Gene quartets

If a gene duplication event that gives rise to paralogs occurs after a speciation event (i.e., the gene duplication event was not shared between taxa of interest), then we expect conspecific paralogs to be sister in the resulting gene tree. Alternatively, if the duplication event occurred prior to speciation, we expect one paralog from each of the species to be sister to one another in the resulting gene tree. By comparing the proportion of all gene quartets that share the latter topology, we can assess the likelihood that two species share the same WGM event. Each species was compared to itself with BLAT (Kent 2002), to find protein pairs in the Ks range of 0.2-0.6, with a minimum of 300bp in their alignment. ProteinOrtho (Lechner et al. 2011) was then used to cluster proteins across two species of interest, using one representative sequence from each pair obtained in the previous step. ProteinOrtho is similar to OrthoMCL or OrthoFinder, but uses "spectral" partitioning to decompose large clusters into smaller clusters. The one-to-one clusters were retained, and the second protein in each pair was merged back to form 4-protein clusters, each with 2 proteins from each of the 2 species. Each cluster was aligned with MUSCLE, and alignments were restricted to homologous sites with BLASTn. Finally, a maximum likelihood tree and its bootstrap support were obtained for each quartet with RAxML version 8.2.10 (Stamatakis 2014), run with the GTR + Γ model and 100 bootstrap replicates. A given quartet was only kept if its tree (with a single internal edge) had a bootstrap >70%. Quartet trees were then summarized across all quartets, for each given pair of species. The code for this pipeline is available at https://github.com/cecileane/wgd-analysis.

Bayesian concordance analysis of nuclear genes

To infer phylogenies based on nuclear genes, we used Orthofinder version 2.1.2 (Emms and Kelly 2015) on the genomes of *Bombax, Gossypium, Durio,* and *Theobroma*, and the transcriptome of *Dombeya* to identify single homologous genes for analysis. Bayesian phylogenetic analysis was performed independently on each gene using MrBayes version 3.2.2 (Ronquist and Huelsenbeck 2003) with 2 million generations, 3 chains with a swap rate of 0.45, and 10% discarded as burn-in. The resulting posterior probabilities were used for Bayesian concordance analysis implemented in BUCKy version 1.4.4 (Ané et al. 2007; Larget et al. 2010), with an alpha of 1 and one million generations.

Classifying MUL gene tree topologies

We extracted protein-sequence orthogroups using OrthoFinder version 2.1.2 (Emms and Kelly 2015) from genomic datasets of *Bombax*, *Durio, Gossypium*, and *Theobroma,* and our *de novo* assembly of *Dombeya*. Sequences were aligned with MAFFT (Katoh et al. 2002) and gene trees were estimated using the PROTGAMMAWAG model of substitution in RAxML (Stamatakis 2014) with 100 bootstrap replicates. Gene trees were filtered to include only those with >80% bootstrap support on every branch. Trees were rooted on *Theobroma* and we used custom R scripts to classify the topologies (see Results).

Multiple-Labeled (MUL) gene tree analysis and hybridization testing

Using the same orthogroups used to construct the Ks plots, we aligned protein sequence data and obtained gene trees using OrthoFinder version 2.1.2 (Emms and Kelly 2015), which is based on the DendroBlast (Kelly and Maini 2013) algorithm. The resulting 21,757 non-trivial gene trees were rooted (-O flag) using Urec version 1.02 (Górecki and Tiuryn 2006) with the plastid DNA tree defined as the most probable species tree. Because gene trees often contain multiple sets of orthologs (each derived from a single gene at the root of the species tree), there may be more than one correct rooting. Urec minimizes the number of gene gains and gene loss events to reconcile each gene tree with the given species tree. These rooted gene trees were then input into GRAMPA (Gregg et al. 2017) with the maximum number of multiply labeled taxa in each gene tree set to 8 (default), and no restrictions as to which nodes in the species tree (if any) were involved in allopolyploidy.

For additional analyses using only published genomic datasets and our *de novo* assembly of *Dombeya*, protein sequences were clustered using OrthoFinder version 2.1.2 (Emms and Kelly 2015), and sequence alignments for each orthogroup were constructed using MAFFT (Katoh et al. 2002). Gene trees were inferred using the PROTGAMMAWAG model of substitution in RAxML (Stamatakis 2014) with 100 bootstrap replicates. Only those gene trees in which all internal nodes had at least 50% bootstrap support were retained for further analyses. Gene trees were rooted with Urec version 1.02 (Górecki and Tiuryn 2006) using the species tree topology generated from our plastid sequences (see Results), and GRAMPA was run with the maximum number of multiply labeled taxa in each gene tree set to 18. No restrictions were set on which lineages were involved in allopolyploidy.

Modeling orthologous gene family sizes

To further test the hypothesis of a single or separate WGM events, we implemented a statistical framework for using counts of orthologous genes between species to comparing hypotheses for the placement of WGD or WGT events on a phylogenetic tree. Counts of orthologs per orthogroup were obtained from OrthoFinder version 2.1.2 (Emms and Kelly 2015) for the published genomes (*Gossypium, Theobroma, Bombax,* and *Durio*). We used concatenation of single-copy nuclear-encoded proteins whose gene trees matched the expected species tree topology to infer a phylogeny in RAxML (Stamatakis 2014) using the PROTGAMMAWAG model of substitution with branch lengths in substitutions/site. This phylogeny was also rendered ultrametric with penalized likelihood (PL) using Chronos (Sanderson et al. 2002; Kim and Sanderson 2008) in the R package APE (Paradis et al. 2004; Paradis 2011; Popescu et al. 2012), with default parameters of a correlated clock and lambda=1 (as in Paradis 2013). On the non-calibrated tree, to make birth rate comparable with that on the calibrated tree, branch lengths were rescaled to have an average distance of 1 between the root and the tips. Analyses described below were run both on this rescaled maximum likelihood tree and on the PL-calibrated tree.

Modeling gene family sizes was implemented in the R package WGDgc as described in Rabier et al. (2013). Using the phylogenetic tree inferred above, gene family size was modelled using equal background (small-scale) gene duplication and gene loss rates, with the same rate per unit branch rate assumed for the entire tree. Shared duplication or triplication events were modeled by doubling or tripling counts in each orthogroup, followed by immediate fractionation and loss of each duplicate with some probability, 1 minus the retention rate, with each event having an independent retention rate. Parameters were estimated jointly by maximum likelihood and models were compared using the Bayesian Information Criterion (BIC). The estimated parameters include the background duplication/loss rate, one retention rate for each WGM, the timing of each event, and the average number of genes per family at the root of the tree (with the number of genes at the root assumed to follow a geometric distribution, translated to start at 1). The likelihood calculation accounts for some sources of ascertainment bias. In particular, we corrected for the fact that we cannot observe families that went extinct or are unsampled (zero genes in all the sampled taxa) and that OrthoFinder excluded singletons, i.e. families with a single gene. To correct for these missing data, means were calculated excluding gene families that went extinct either in *Theobroma*, or in all of the other three taxa.

Given our four-taxon tree, we tested a total of 89 models accounting for possible phylogenetic placements of WGD/WGT events. Tables S3 and 4 present results for each scenario, but briefly the models tested were: 1) No events, 2) one WGD or one WGT event occurring on five possible tree edges resulting in total of 10 models, 3) either two WGD or one WGD + one WGT, each event placed on any of the 5 possible edges, resulting in 40 possible models, 4) three events with one specific to Durio and two events (two WGD or one WGD + one WGT) in Malvatheca (3 edges), accounting for 30 models, and 5) four events, but none of them shared between taxa, accounting for 8 models. Our constraint on four-event scenarios were to have two in *Gossypium* (two WGD or one WGD + one WGT), one in *Bombax*, and one in *Durio*. The likelihood of shared WDG/WDT models was compared by BIC.

ACKNOWLEDGEMENTS

Funding was provided through the National Science Foundation (award 1354268 to DAB, CA, CEG, and JFW) and Cotton Incorporated (to JFW). The authors thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility for providing RNA-Seq library preparation and transcriptome sequencing services, and the Iowa State University Research IT Unit (http://researchit.las.iastate.edu/) for computational support.

Photograph credits for Figure 2 are as follows: Corrinne Grover for *Gossypium raimondii*, C. Skema for *Dombeya burgessiae,* Samuel Logan for *Tilia cordata,* Nisa Karimi for *Adansonia digitata*, and under the Creative Commons Attribution-Share Alike 3.0 license from Wikimedia Commons: *Bombax ceiba* by Earth100, *Firmiana colorata* by Delonix*, Durio* by Mokkie, and *Theobroma cacao* by Domste*.*

AUTHOR CONTRIBUTIONS

D.A.B. conceived of the study; C.E.G., N.K., S.A.L., C.S., N.S., J.T., and K.W. generated data; J.L.C., D.A.B., N.K., and C.A. performed most of the data analysis; J.L.C., N.K., C.A., C.S., D.A.B., C.E.G., and J.F.W. wrote the manuscript;all authors edited the manuscript and approved of its content.

REFERENCES

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. **Curr Opin Plant Biol** 8: 135-141

Alverson WS, Whitlock BA, Nyffeler R, Bayer C, Baum DA (1999) Phylogeny of the core Malvales: Evidence from ndhF sequence data. **Am J Bot** 86: 1474–1486

Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. **Mol Biol Evol** 24: 412–426

Ané C (2010) Reconstructing concordance trees and testing the coalescent model from genome-wide data sets. In: Knowles L and Kubatko L, eds. *Estimating Species Trees: Practical and Theoretical Aspects*, Wiley-Blackwell

Baum DA (2007) Concordance trees, concordance factors, and the exploration of reticulate genealogy. **Taxon** 56: 417-426

Baum DA, Alverson WS, Nyffeler R (1998) A durian by any other name: Taxonomy and nomenclature of the core malvales. **Harv Pap Bot** 3: 315–330

Baum DA, Oginuma K (1994) A review of chromosome numbers in Bombacaceae with new counts for Adansonia. **Taxon** 43: 11-20

Baum DA, Smith SD, Yen A, Alverson WS, Nyffeler R, Whitlock BA, Oldham RL (2004) Phylogenetic relationships of Malvatheca (Bombacoideae and Malvoideae; Malvaceae sensu lato) as inferred from plastid DNA sequences. **Am J Bot** 91: 1863–1871

Bayer C, Fay MF, De Bruijn AY, Savolainen V, Morton CM, Kubitzki K, Alverson WS, Chase MW (1999) Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: A combined analysis of plastid atpB and rbcL DNA sequences. **Bot J Linn Soc** 129: 267–303

Bergsten J (2005) A review of long-branch attraction. **Cladistics** 21: 163-193

Bowe LM, dePamphilis CW (1996) Effects of RNA editing and gene processing on phylogenetic reconstruction. **Mol Biol Evol** 13: 1159-1166

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. **Nature** 422: 433–438

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN Jr, Rolf M, Kutchan T, Tan X, Chen C, Zhang Y, Carpenter E, Wong GK, Doyle JJ, Leebens-Mack J (2015) Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. **Mol Biol Evol** 32: 193–210

Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. **Plant Mol Biol Rep** 11: 113–116

Chen SF, Li MW, Jing HJ, Zhou RC, Yang GL, Wu W, Fan Q, Liao WB (2015) *De novo* transcriptome assembly in *Firmiana danxiaensis*, a tree species endemic to the Danxia landform ZJ Liu, ed. **PLoS ONE** 10: e0139373

Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X (2018) Gene retention, fractionation and subgenome differences in polyploid plants. **Nat Plants** 4: 258-268

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA (2006) Widespread genome duplications throughout the history of flowering plants. **Genome Res** 16: 738-749

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: More models, new heuristics and parallel computing. **Nat Methods** 9: 772

Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM (2009) Shedding light on an extremophile lifestyle through transcriptomics. **New Phytol** 183: 764-775

Emms DM, Kelly S (2015) OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biol** 16: 157

Gao Y, Wang H, Liu C, Chu H, Dai D, Song S, Yu L, Han L, Fu Y, Tian B, Tang L (2018) *De novo* genome assembly of the red silk cotton tree (*Bombax ceiba*). **Gigascience** 7: giy051

Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. **Proc Natl Acad Sci** 94: 6809-6814

Górecki P, Tiuryn J (2006) URec: A system for unrooted reconciliation. **Bioinformatics** 23: 511-512

Grant V (1971) Plant speciation, 2nd edition. New York, New York, USA (Vol. 194).

Gregg WCT, Ather SH, Hahn MW (2017) Gene-tree reconciliation with MUL-Trees to resolve polyploidy events. **Syst Biol** 66: 1007–1018

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. **Nat Protoc** 8: 1494–1512

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

Helentjaris T, Weber D, Wright S (1988) Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. **Genetics** 118: 353-363

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE (2011) Ancestral polyploidy in seed plants and angiosperms. **Nature** 473: 97-100

Johnson MT, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, Chase MW, Clarke ND, Covshoff S, Edger PP, Goh F (2012) Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. **PLoS ONE** 7: e50226

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Res** 30: 3059–3066

Kelly S, Maini PK (2013) DendroBLAST: Approximate phylogenetic trees in the absence of multiple sequence alignments. **PLoS ONE** 8: e58537

Kent WJ (2002) BLAT—the BLAST-like alignment tool. **Genome Res** 12: 656-664

Kim YM, Kim S, Koo N, Shin AY, Yeom SI, Seo E, Park SJ, Kang WH, Kim MS, Park J, Jang I (2016) Genome analysis of Hibiscus syriacus provides insights of polyploidization and indeterminate flowering in woody plants. **DNA Res** 24: 71-80

Kim D, Langmead B, Salzberg SL (2015) HISAT: A fast spliced aligner with low memory requirements. **Nat Methods** 12: 357-360

Kim J, Sanderson MJ (2008) Penalized likelihood phylogenetic inference: Bridging the parsimony-likelihood gap. **Syst Biol** 57: 665-674

Krueger F (2015) Trim Galore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. Available: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

La Duke JC, Doebley J (1995) A chloroplast DNA based phylogeny of the Malvaceae. **Syst Bot** 20: 259-271

Larget BR, Kotha SK, Dewey CN, Ané C (2010) BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. **Bioinformatics** 26: 2910-2911

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ (2011) Proteinortho: Detection of (co-) orthologs in large-scale analysis. **BMC Bioinformatics** 12: 124

Li H, Li D, Chen A, Tang H, Li J, Huang S (2016) Characterization of the kenaf (*Hibiscus cannabinus*) global transcriptome using illumina paired-end sequencing and development of EST-SSR markers. **PLoS ONE** 11: e0150548

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J (2012) SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. **Gigascience** 1: 18

Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, Burleigh JG (2014) Data access for the 1,000 plants (1KP) project. **Gigascience** 3: 17

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw J, Senin P, Wang W, Ly B, Lewis K, Salzberg S, Feng L, Jones M, Skelton R, Murray J, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull R, Michael T, Wall K, Rice D, Albert H, Wang M, Zhu Y, Schatz M, Nagarajan N, Acob R, Guan P, Blas A, Wai C, Ackerman C, Ren Y, Liu C, Wang J, Wang J, Na J, Shakirov E, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers J, Gschwend A, Delcher A, Singh R, Suzuki J, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Pérez R, Torres M, Feltus

F, Porter B, Li Y, Burroughs A, Luo M, Liu L, Christopher D, Mount S, Moore P, Sugimura T, Jiang J, Schuler M, Friedman V, Mitchell-Olds T, Shippen D, dePamphilis C, Palmer J, Freeling M, Paterson A, Gonsalves D, Wang L, Alam M (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* L.). **Nature** 452: 991

Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone III D, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, Saski C (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. **Genome Biol** 14: r53

Ness RW, Graham SW, Barrett SC, (2011) Reconciling gene and genome duplication events: Using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae. **Mol Biol Evol** 28: 3009-3018

Nyffeler R, Bayer C, Alverson WS, Yen A, Whitlock BA, Chase MW, Baum DA (2005) Phylogenetic analysis of the Malvadendrina clade (Malvaceae sl) based on plastid DNA sequences. **Org Divers Evol** 5: 109-123

Page JT, Huynh MD, Liechty ZS, Grupp K, Stelly DM, Hulse AM, Ashrafi H, Van Deynze A, Wendel JF Udall JA (2013) Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. **G3: Genes Genom Genet** 3: 1809-1818

Page JT, Liechty ZS, Huynh MD, Udall JA (2014) BamBam: Genome sequence analysis tools for biologists. **BMC Res Notes** 7: 829

Panchy N, Lehti-Shiu M, Shiu SH (2016) Evolution of gene duplication in plants. **Plant Physiol** 171: 2294–316

Paradis E (2011) *Analysis of Phylogenetics and Evolution with R*. Springer Science & Business Media.

Paradis E (2013) Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. **Mol Phylogenet Evol** 67: 436–444

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. **Bioinformatics** 20: 289-290

Parks MB, Nakov T, Ruck EC, Wickett NJ, Alverson AJ (2018) Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). **A J Bot** 105: 330-347

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. **Proc Natl Acad Sci** 101: 9903-9908

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffman LV, Hovav R, Jones DC, Lemke C, Mansoor S, ur Rahman M, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly DA, Tripplett BA, Van Deynze A, Vaslin MF, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KF, Peterson DG, Roksar DS, Wang X, Schmutz J (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. **Nature** 492: 423

Popescu AA, Huber KT, Paradis E (2012) ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. **Bioinformatics** 28: 1536-1537

Qu XJ, Jin JJ, Chaw SM, Li DZ, Yi TS (2017) Multiple measures could alleviate long-branch attraction in phylogenomic reconstruction of Cupressoideae (Cupressaceae). **Sci Rep** 7: 41005

Rabier CE, Ta T, Ané C (2013) Detecting and locating whole genome duplications on a phylogeny: A probabilistic approach. **Mol Biol Evol** 31: 750-762

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF (2015) Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. **Mol Biol Evol** 32: 1063-1071

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. **Bioinformatics** 19:1 572-1574

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. **Mol Biol Evol** 19: 101-109

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. **Genome** 47: 868-876

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. **Proc Natl Acad Sci** 108: 4069-4074

Scott AD, Stenz NW, Ingvarsson PK, Baum DA (2016) Whole genome duplication in coast redwood (Sequoia sempervirens) and its implications for explaining the rarity of polyploidy in conifers. **New Phytol** 211: 186-193

Seyani JH (1991) The genus *Dombeya* (Sterculiaceae) in continental Africa. **Opera Bot Belg** 2: 32

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics** 31: 3210-3212

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, de Pamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm diversification. **Am J Bot** 96: 336-348

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE (2015) Polyploidy and genome evolution in plants. **Curr Opin Genet Dev** 35: 119-125

Soltis PS, Soltis DE (2016) Ancient WGD events as drivers of key innovations in angiosperms. **Curr Opin Plant Biol** 30: 159-165

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

Stebbins GL (1947) Types of polyploids: Their classification and significance. **Adv Genet** 1: 403–429

Tang Y (1992) The systematic position of Corchoropsis Sieb. & Zucc. Cathaya 4: 131–150

Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. **Proc Natl Acad Sci** 107: 472-477

Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, Soh PS (2017) The draft genome of tropical fruit durian (*Durio zibethinus*). **Nat Genet** 49:1633-1641

Train CM, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C (2017) Orthologous Matrix (OMA) algorithm 2.0: More robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. **Bioinformatics** 33: i75–i82

Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. **Nat Rev Genet** 10: 725

Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. **Nat Rev Genet** 18: 411

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in *Arabidopsis*. **Science** 290: 2114-2117

Wang X, Guo H, Wang J, Lei T, Liu T, Wang Z, Li Y, Lee TH, Li J, Tang H, Jin D (2016) Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. **New Phytol** 209: 1252-1263

Wendel JF (2015) The wondrous cycles of polyploidy in plants. **Am J Bot** 102: 1753-1756

Wendel JF, Stuber CW, Edwards MD, Goodman MM (1986) Duplicated chromosome segments in maize (*Zea mays* L.): Further evidence from hexokinase isozymes. **Theor Appl Genet** 72: 178-185

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. **Proc Natl Acad Sci** 111: E4859-E4868

Wilkie P, Clark A, Pennington RT, Cheek M, Bayer C, Wilcock CC (2006) Phylogenetic relationships within the subfamily Sterculioideae (Malvaceae/Sterculiaceae-Sterculieae) using the chloroplast gene ndhF. **Syst Bot** 31: 160-170

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH (2009) The frequency of polyploid speciation in vascular plants. **Proc Natl Acad Sci** 106: 13875-13879

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. **Bioinformatics** 30: 1660-1666

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. **Bioinformatics** 13: 555-556

Zhang L, Ming R, Zhang J, Tao A, Fang P, Qi J (2015) De novo transcriptome sequence and identification of major bast-related genes involved in cellulose biosynthesis in jute (*Corchorus capsularis* L.). **BMC Genomics** 16: 1062

Zhang L, Wan X, Xu J, Lin L, Qi J (2015) De novo assembly of kenaf (*Hibiscus cannabinus*) transcriptome using Illumina sequencing for gene discovery and marker identification. **Mol Breeding** 35: 192

Zhao M, Zhang B, Lisch D, Ma J (2017) Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. **Plant Cell** 29: 2974-2994

Supporting Information

Table S1. Taxa sampled in this study, subfamily, and genomic or transcriptomic, source, and transcriptome assembly statistics, including BUSCO

Table S2. Average gene family size in each taxon

Table S3. Gene count analyses from WGDgc. Models, their likelihoods and parameter estimates for the starting tree with branch lengths proportional to their raw values in substitutions/site

Table S4. Gene count analyses from WGDgc. Models, their likelihoods and parameter estimates for the starting tree with branch lengths calibrated with Penalized Likelihood

Figure legends:

Figure 1. Maximum Likelihood trees based on the concatenated data sets of 67 plastid genes

Trees were rooted with *Arabidopsis thaliana*. Unlabeled branches have 100% bootstrap support. (**A**) Phylogeny inferred from genomic data only. (**B**) Phylogeny excluding taxa for which we only had RNAseq data, except for *Dombeya*. (**C**) All taxa sampled.

Figure 2. Malvaceae plastid phylogeny and Ks frequency histograms

(**A**) Malvaceae phylogeny based on 67 plastid genes, as in Figure 1B. Internodes without labels have 100% bootstrap support. Dotted lines indicate placement of taxa not included in Figure 1. (**B**) Ks frequency histograms generated for each transcriptome analyzed. Vertical lines on the Ks plots indicate the median Ks of that species relative to either *Bombax* (red), *Dombeya* (orange), *Theobroma* (yellow), *Durio* (green), or *Gossypium* (blue).

Figure 3. Polyploidy events inferred by GRAMPA

GRAMPA suggests multiple alternative polyploidy events based on reconciliation of gene trees and multiply-labeled species trees (MUL-trees). Clades that are the result of allopolyploidy are indicated by a triangle. Because the phylogeny used is based on (presumably) maternally-inherited plastomes, triangles also represent that maternal progenitor in the allopolyploidy event. Several potential paternal progenitors are indicated by color-coded branches. Triangles with more than one color indicate alternative hypotheses of the paternal progenitor branch. (**A**) Using transcriptomic data for all but *Gossypium* and *Theobroma*, 53 possible polyploidy scenarios had reconciliation scores better than the single-labeled species tree. The top 20 are summarized. (**B**) Analysis of genomic data from *Gossypium*, *Theobroma*, *Durio*, *Bombax*, and *Carica papaya*, with the top four of the seven polyploid scenarios shown (the three not shown all suggested *Dombeya* has an allopolyploid origin, which contradicts the Ks plots).

Figure 4. Phylogeny inferred by Bayesian concordance analysis of 1214 nuclear genes and maximum likelihood inference of concatenation of genes

(**A**) Population tree inferred by Bayesian concordance analysis with BUCKy from 1214 singleton nuclear genes. Concordance factors are printed above the branches, branch lengths in coalescence units below. (**B**) Maximum likelihood phylogeny inferred from the concatenated 1214 nuclear genes by RAxML. All branches have 100% bootstrap support.

Figure 5. WGDgc results for gene count data and alternative models for the distribution of WGMs

(**A**) Branch lengths proportional to their raw values in substitutions/site. (**B**) Branch lengths calibrated with Penalized Likelihood. On each tree, the best model (according to BIC) is shown. Numbers by WGM symbols are the estimated retention probability of each extra copy. λ is the estimated gene turnover (duplication & loss) rate. Numbers in parentheses, next to taxon names, are the mean number of genes/family among the families that were used for analysis.

Figure 6. Hypotheses presented to explain the data

(**A**) Hypothesis one, which is consistent with Bayesian concordance analysis, ML phylogeny of concatenated nuclear genes, Ks histograms, and gene tree topology testing, suggests that *Durio* is an allopolyploid. (**B**) Hypothesis two, which is consistent with Ks histograms, plastid phylogeny, and GRAMPA, suggests that the Malvatheca clade is allopolyploid.

Table 1. Quartet-based gene tree topology test

For any two species that contained a Ks peak, we extracted quartets (pairs of paralogs) under the peak and constructed gene trees to test if the tree topology is consistent with a shared gene duplication event. Below diagonal: number of genes that support separate gene duplication event + genes suggesting shared duplications. Above diagonal: percentage of quartets (pairs of paralogs) yielding a topology consistent with a separate duplication. Values well-below or well-above the expected 33% for random sampling are colored: <15% = red; >45% = blue.

	Hibiscus	*Gossypium*	*Adansonia*	*Bombax*	*Durio*	*Firmiana*	*Heritiera*	*Tilia*
Hibiscus	---	36.92%	64.69%	59.87%	90.81%	84.41%	89.80%	93.42%
Gossypium	223+381	---	6.00%	7.43%	49.33%	48.35%	50.88%	56.25%
Adansonia	240+131	21+329	---	3.82%	20.71%	33.06%	36.05%	40.35%
Bombax	273+183	35+436	24+604	---	22.07%	27.70%	22.95%	31.43%
Durio	168+17	37+38	29+111	32+113	---	11.86%	7.37%	7.69%
Firmiana	157+29	44+47	40+81	59+154	14+104	---	3.78%	13.89%
Heritiera	88+10	29+28	31+55	28+94	7+88	9+229	---	7.27%
Tilia	71+5	18+14	23+34	22+48	5+60	10+62	4+51	---

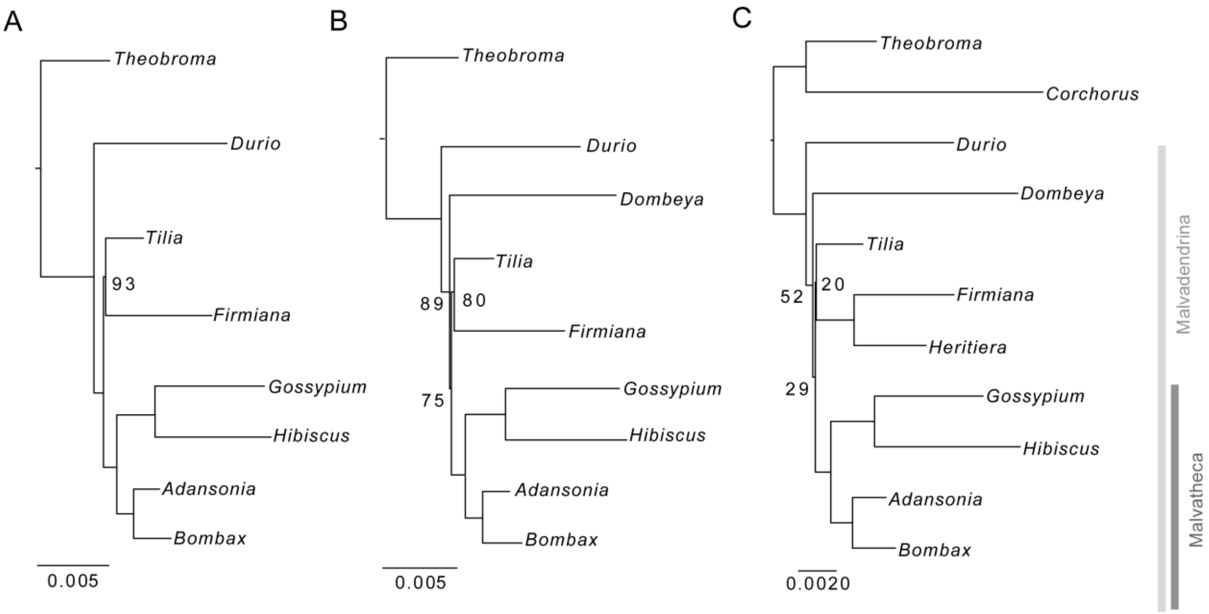
A
Theobroma
Durio
Tilia
93
Firmiana
Gossypium
Hibiscus
Adansonia
Bombax
0.005
B
Theobroma
Durio
Dombeya
Tilia
89
80
Firmiana
75
Gossypium
Hibiscus
Adansonia
Bombax
0.005
C
Theobroma
Corchorus
Durio
Dombeya
Tilia
52
20
Firmiana
Heritiera
29
Gossypium
Hibiscus
Adansonia
Bombax
0.0020
Malvadendrina
Malvatheca

Figure 1

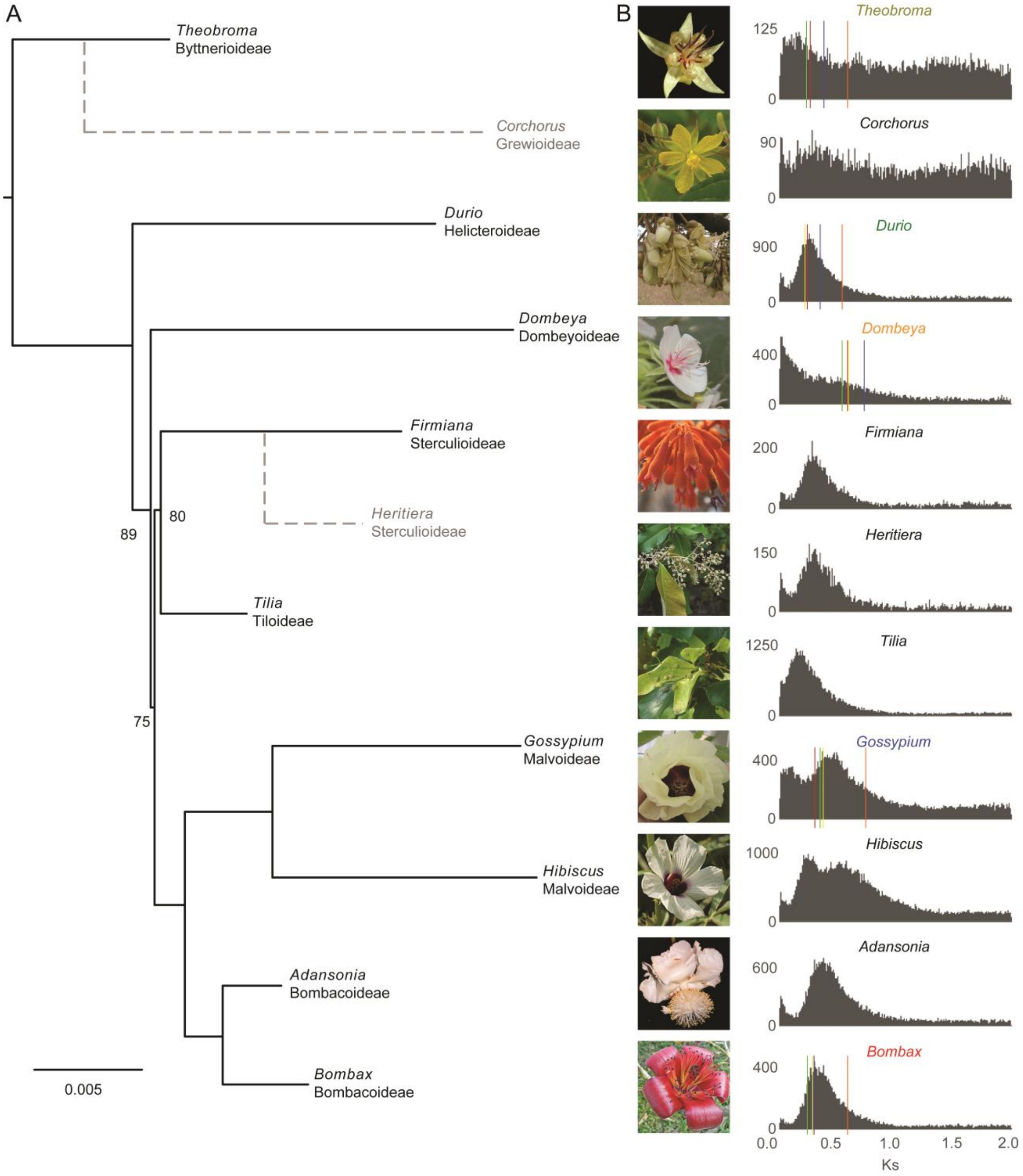
A
Theobroma
Byttnerioideae
Corchorus
Grewioideae
Durio
Helicteroideae
Dombeya
Dombeyoideae
Firmiana
Sterculioideae
80
89
Heritiera
Sterculioideae
Tilia
Tiloideae
75
Gossypium
Malvoideae
Hibiscus
Malvoideae
Adansonia
Bombacoideae
Bombax
Bombacoideae
0.005
B
Theobroma
125
0
Corchorus
90
0
Durio
900
0
Dombeya
400
0
Firmiana
200
0
Heritiera
150
0
Tilia
1250
0
Gossypium
400
0
Hibiscus
1000
0
Adansonia
600
0
Bombax
400
0
0.0
0.5
1.0
1.5
2.0
Ks

Figure 2

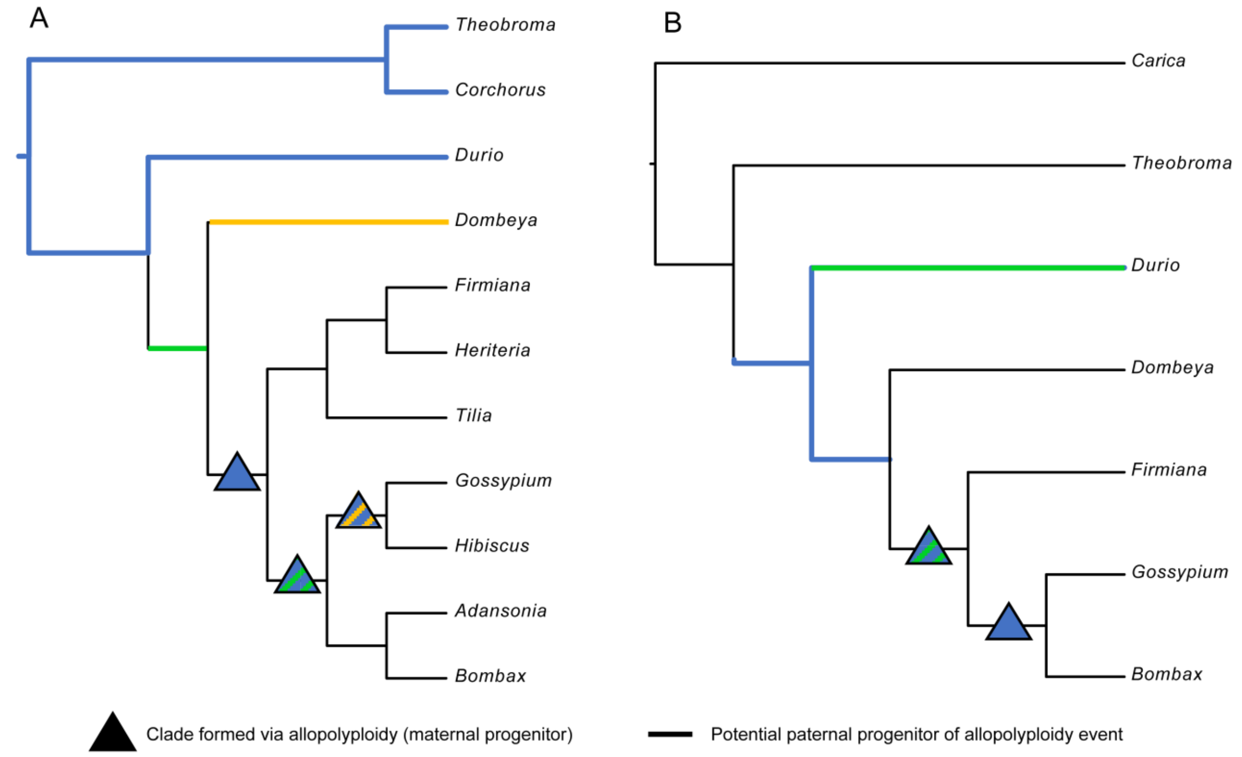

Figure 3

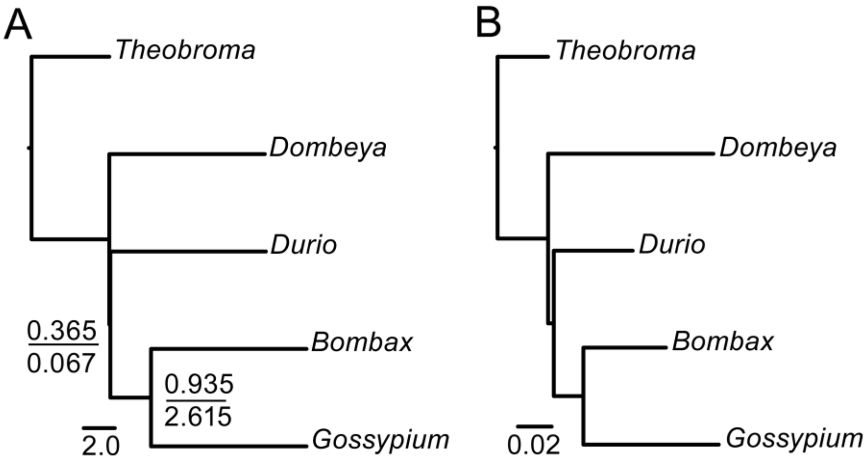

Figure 4

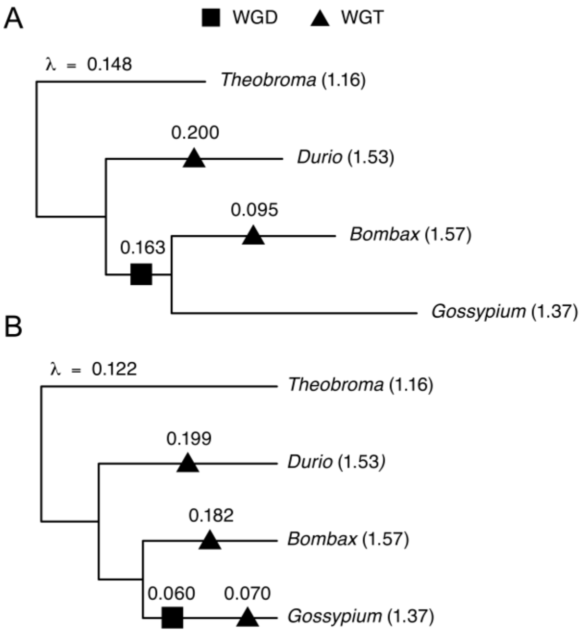

Figure 5

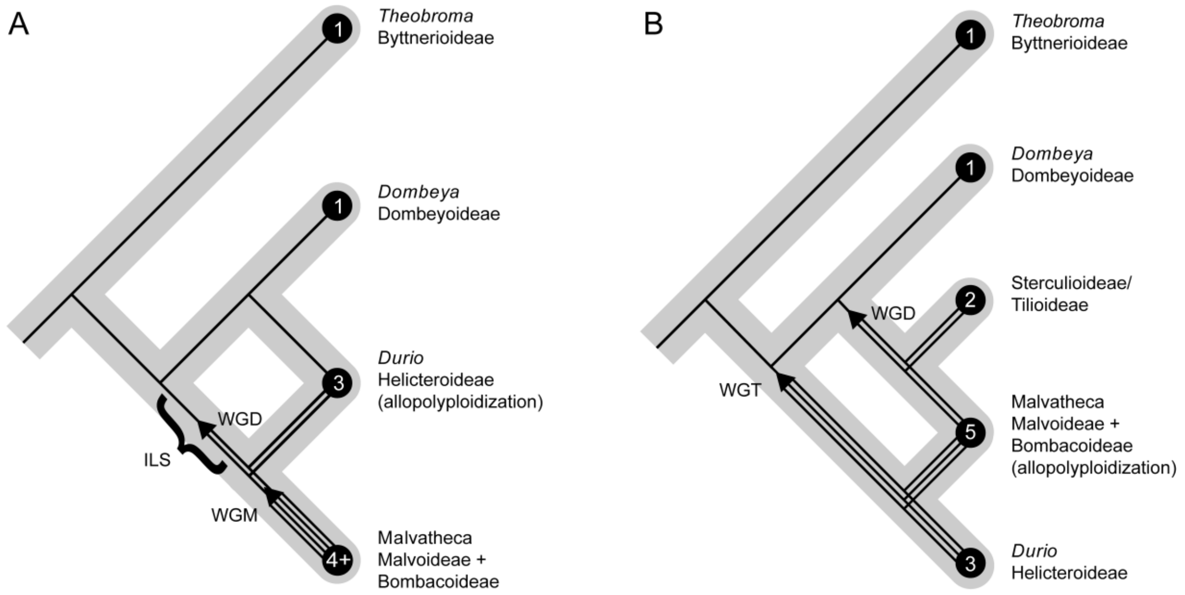

Figure 6