PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

Subgenome-aware analyses reveal the genomic consequences of ancient allopolyploid hybridizations throughout the cotton family

Pengchuan Sun[a], Zhiqiang Lu[b], Zhenyue Wang[c], Shang Wang[a], Kexin Zhao[a], Dong Mei[a], Jiao Yang[c], Yongzhi Yang[a,c,1], Susanne S. Renner[d,1], and Jianquan Liu[a,c,1]

Edited by Douglas Soltis, University of Florida, Gainesville, FL; received August 12, 2023; accepted February 27, 2024

Malvaceae comprise some 4,225 species in 243 genera and nine subfamilies and include economically important species, such as cacao, cotton, durian, and jute, with cotton an important model system for studying the domestication of polyploids. Here, we use chromosome-level genome assemblies from representatives of five or six subfamilies (depending on the placement of *Ochroma*) to differentiate coexisting subgenomes and their evolution during the family's deep history. The results reveal that the allohexaploid Helicteroideae partially derive from an allotetraploid Sterculioideae and also form a component of the allodecaploid Bombacoideae and Malvoideae. The ancestral Malvaceae karyotype consists of 11 protochromosomes. Four subfamilies share a unique reciprocal chromosome translocation, and two other subfamilies share a chromosome fusion. DNA alignments of single-copy nuclear genes do not yield the same relationships as inferred from chromosome structural traits, probably because of genes originating from different ancestral subgenomes. These results illustrate how chromosome-structural data can unravel the evolutionary history of groups with ancient hybrid genomes.

allopolyploidy | chromosomal fusion | genome duplications | protochromosome | karyotype evolution

Significance

Hybridization of divergent species can result in viable offspring if it is accompanied by genome doubling, allowing proper meiotic pairing and isolating the produced lineage from its parents. The parental genomes may coexist within the same nucleus for millions of years, but this has been challenging to show because of genomic rearrangements and loss or retention of genes. In this study, we use a approach to decipher the sequence of ancient genome duplications in the cotton family, using seven chromosome-level genomes of representative species from most of the nine subfamilies. Our results reveal that an allohexaploid clade (ranked as a subfamily) partially derives from a different allotetraploid clade and itself contributed genomes to a younger allodecaploid clade.

Allotetraploid speciation begins with the hybridization of divergent species, resulting in polyploid offspring carrying the genomes of both progenitors. The produced tetraploid is instantly genetically isolated from its diploid parents and "might therefore be expected to pursue an independent course of evolution" (1–3). The two homoeologous genomes can persist within the same nucleus for some time, and the initial stages of their evolution can be tractable in recent allotetraploids, as is the case in *Tragopogon* (4, 5). The long-term evolution and diversification of allopolyploid lineages is more challenging to infer because of mutations, genomic rearrangements, and the differential retention or loss of genes in the different subgenomes (6). Reconstructing ancient hybridizations with genome doubling requires distinguishing between orthologs of different ancestral origins in order to disentangle their individual evolutionary history, which will differ from the species tree.

Despite tangled gene trees, reconstructing the sequence of evolutionary events is possible through the comparison of multiple chromosome-level genomes, which can provide orthology criteria from shared numbers of genes and their chromosomal positions (7–12). In addition, rare large chromosome structural variations can provide independent evidence for constructing phylogenetic relationships (13–15). Comparative genomic studies on the evolution of karyotypes have been especially successful in the Brassicaceae and Poaceae (16–19).

Malvaceae, a globally distributed clade of some 4,225 species in 243 genera (20, accessed 7 July 2023), provides the world with essential natural fibers, timber, medicine, and other products, including cacao, cotton, durian, and jute. Modern classifications (21, 22) recognize nine subfamilies, namely Bombacoideae (17 genera, 164 species, mostly in the New World), Brownlowioideae (8 genera, 68 species, mostly in the Old World), Byttnerioideae (26 genera, 650 species, pantropical), Dombeyoideae (20 genera, 375 species, mostly in the Old World), Grewioideae (25 genera, 770 species, worldwide), Helicteroideae (8 to 12 genera, 95 species, mostly in Australasia), Malvoideae (78 genera, 1,800 species, worldwide), Sterculioideae (12 genera, 430 species, mostly in the Old World), and Tilioideae (3 genera, 50 species, mostly in the Northern hemisphere). Analyses of plastid, nuclear, and mitochondrial genes for up to 187 of the 243 genera and of entire plastid genomes for 35 representative species have suggested that these subfamilies form two clusters, the Byttneriina clade (Byttnerioideae and Grewioideae) and the Malvadendrina clade (Bombacoideae, Brownlowioideae, Dombeyoideae, Helicteroideae, Malvoideae, Sterculioideae, and Tiliodeae) (21–23). A phylogeny obtained from 1,214 nuclear genes (single-copy and homologous) for nine species representing eight of the nine subfamilies, however, lacked strong statistical support (24).

Author contributions: P.S., Y.Y., S.S.R., and J.L. designed research; P.S. and Y.Y. performed research; Z.L. and Y.Y. contributed new reagents/analytic tools; P.S., Z.W., S.W., K.Z., D.M., and J.Y. analyzed data; and P.S., Y.Y., S.S.R., and J.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yangyz@lzu.edu.cn, srenner@wustl.edu, or liujq@nwipb.ac.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2313921121/-/DCSupplemental.

Published April 3, 2024.

Species or lineages of the Malvaceae show signs of several whole-genome multiplications (WGMs), but how many have occurred in exactly which groups remains unclear (24–27). For example, whether *Gossypium* (Malvoideae) and *Durio* (Helicteroideae) share a direct polyploid ancestor is unresolved, with *Gossypium* experiencing a decaploidization after divergence from *Theobroma* (Byttnerioideae), while *Durio* apparently underwent another hexaploidization (26). Similarly, how many ancient polyploidy events are shared by Bombacoideae and Malvoideae remains uncertain, although minimally one shared WGD has been inferred (24).

In this study, we aimed to decipher the sequence of genome duplications deep in the history of Malvaceae, keeping in mind possible past hybridizations among species, which might have resulted in allopolyploidy and the prolonged coexistence of parental genomes. To achieve this goal, we used seven well-assembled genomes of representative species from five or six of the nine subfamilies, namely balsa (*Ochroma pyramidale*), placed variously in Bombacoideae (22, 28) or Malvoideae (21, 29), cacao (*Theobroma cacao*) and *Herrania umbratica* from Byttnerioideae, cotton (*Gossypium arboreum*) from Malvoideae, durian (*Durio zibethinus*) from Helicteroideae, *Heritiera littoralis* from Sterculioideae, and white jute (*Corchorus capsularis*) from Grewioideae (*SI Appendix*, Table S1). For balsa, we generated a de novo chromosome-level assembly (a scaffold-level assembly became available after completion of our study) (30). We assigned subgenomes for each species and extracted collinear genes for phylogenetic analyses. Our results reveal several ancient hybridizations, accompanied by genome doubling, and their long-term effects during the evolution of Malvaceae, a clade that is at least 60 My old as documented by an ample fossil record (31).

Results

A Chromosome-Level Genome Assembly for Balsa. We assembled a de novo balsa (*O. pyramidale*, Bombacoideae, voucher specimen Lu Zhiqiang-2019-Xishuangbana-01, *SI Appendix*, Fig. S1) genome using Oxford Nanopore Technologies (ONT) and Hi-C approaches. The assembled genome comprises 1.79 Gb (contig N50 of 26.34 Mb) of which 98.38% could be anchored onto 44 pseudo-chromosomes, consistent with reported haploid numbers (https://ccdb.tau.ac.il/search/) (*SI Appendix*, Fig. S2 and Tables S2–S5). We predicted 53,515 protein-coding genes, and the BUSCO genome completeness score was 99.61% (*SI Appendix*, Table S6).

Genomic collinearity revealed evidence of WGMs, with syntenic depth ratios of 5:1 when comparing balsa and cotton with cacao (Byttnerioideae), suggesting that the former two may share a decaploidization event (Fig. 1*A*). However, the distribution of synonymous substitutions per synonymous site (*Ks*) for balsa and cotton shows peaks for WGM events at 0.19 and 0.57, respectively, while the *Ks* value for their orthologs at 0.36 falls between these two values (Fig. 1*B*). This does not clearly decide between either a shared or two unique decaploidization(s) in the two species; different evolutionary rates might account for the significant discrepancy in *Ks* values among them (27, 32, 33).

If balsa and cotton indeed shared an ancestral decaploidization, one would expect more consistency in their orthologous subgenomes than in their paralogous subgenomes because orthologous subgenomes should retain more syntenic gene pairs and have higher sequence similarity scores (estimated from the BLAST results). This expectation was met (Fig. 1*C*). By using cacao as a reference and Chr10 as an example, we were able to readily identify five homeologous blocks in balsa and cotton (Fig. 1 *A* and *C*). Analysis of these five blocks reveals high consistency in their normalized retention (Fig. 1*C*, the red values within the five highlighted squares are around 0.7, while none of the other red values exceed 0.3; see *Methods*), and the sequencing depth relationship among them shows a 1:1 ratio, which supports that balsa and cotton uniquely share fivefold WGMs.

We then extracted collinear genes from the five homoeologous blocks and constructed gene trees, rooted on cacao Chr10. The coalescent phylogenetic tree revealed five paired clusters (Fig. 1*D* and *SI Appendix*, Fig. S3; see *Methods*). Therefore, all lines of evidence indicate that the most-recent common ancestor of balsa and cotton underwent a fivefold WGMs, referred to as a decaploid event, following the core eudicot triploidy event.

Long-Term Effects of Ancient Allopolyploidy Events and Evolutionary Relationships among Malvaceae Subgenomes. We next explored WGMs in the Bombacoideae, Byttnerioideae, Grewioideae, Helicteroideae, and Sterculioideae. Our analysis included balsa, cotton, durian, *H. littoralis* (Sterculioideae), *H. umbratica* (Byttnerioideae), and white jute (*C. capsularis*, Grewioideae) with *Carica papaya* (Brassicales) as an outgroup. Using cacao Chr10 as reference, we could clearly identify syntenic depth ratios of 3:1, 2:1, 1:1, 1:1, and 1:1 between *H. littoralis*, durian, white jute, *H. umbratica* and Papaya (Fig. 2*A* and *SI Appendix*, Fig. S4). Based on the complementarity and completeness of these synteny blocks, we phased them into distinct subgenomes (Fig. 2*A*). To clarify the historic relationships among these subgenomes, we extracted collinear genes from these synteny blocks and constructed gene trees, rooted on papaya (Fig. 2*B* and *SI Appendix*, Fig. S3). Phylogenies built from 900 collinear genes of subgenomes with at least four collinear genes resulted in 900 topologies from which we obtained a consensus tree. This tree showed that the subgenomes have independent histories (Fig. 2*B*). The coexisting subgenomes within the Malvadendrina clade form five clusters (labeled S_1 to S_5 in Fig. 2*B*), implying that they originated after the common ancestor of this clade had begun diversifying. That subgenomes S_1 and S_2 fall into different clades suggests a potential allopolyploid origin for the tetraploid Sterculioideae (which harbor subgenomes S_1, S_2). Subgenome S_3 forms a clade with subgenomes S_4 and S_5, rather than with S_1 and S_2, supporting the allohexaploid origin of the Helicteroideae (haboring S_1, S_2, and S_3). In other words, the hexaploid Helicteroideae derive partially from some tetraploid Sterculioideae that hybridized with another diploid (S_3). Subgenomes S_4 and S_5 also fail to coalesce and instead are nested within S_2 and S_3, suggesting that they have an allotetraploid ancestor. Put differently, the decaploid Bombacoideae/Malvoideae (harboring subgenomes S_1, S_2, S_3, S_4, and S_5) derive from some allohexaploid that hybridized with another allotetraploid (harboring subgenomes S_4, S_5). Taken together, these findings suggest a shared whole-genome duplication (WGD) event among *Heritiera litoralis* (Sterculioideae), durian (Helicteroideae), balsa (Bombacoideae), and cotton (Malvoideae). Additionally, data indicate a threefold WGMs shared by balsa (Bombacoideae), cotton (Malvoideae), and durian (Helicteroideae).

The Ancestral Karyotype of the Malvaceae and Relationships Inferred from Collinear Genes Compared to Standard DNA Alignments. To reconstruct protochromosomes, we followed the workflow of Sun et al. (13). We first identified shared chromosome-like synteny blocks (CLSBs) as protochromosomes, which resulted in an ancestral Malvaceae karyotype (AMK) of 11 protochromosomes. Ten of them are retained as independent chromosomes in at least one of the seven species, namely Chrs 1, 2, 3, 5, 6, 7, 8, 9, 10 of cacao (*T. cacao*, Byttnerioideae) and Chr27 of balsa (*SI Appendix*, Figs. S5 and S6). Protochromosome

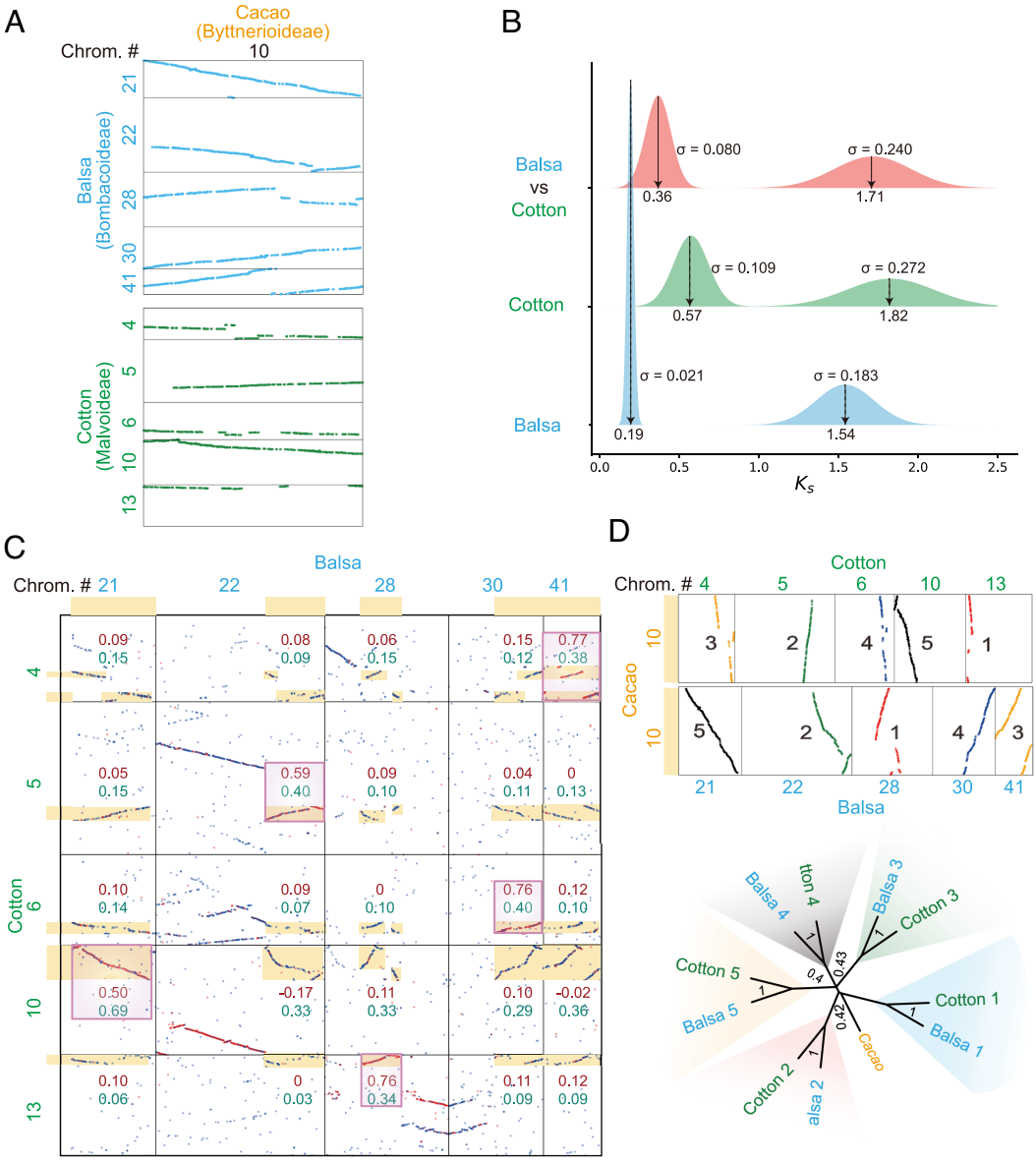

Fig. 1. Evolutionary relationship between Bombacoideae (balsa) and Malvoideae (cotton) inferred from syntenic blocks and subgenomes. (*A*) Syntenic depth ratios with cacao (Byttnerioideae) as a reference. The horizontal and vertical axes represent the chromosome numbers in all dot plots. (*B*) The *Ks* distributions of syntenic blocks in balsa and cotton. (*C*) Local dot plot between balsa and cotton showing homology and retention. Beige shading is used for syntenic blocks between subgenomes. Values in red represent normalized homology, and values in green normalized retention. (*D*) A phylogeny of the coexisting subgenomes based on 900 collinear genes from homologous blocks to cacao Chr10 supports that balsa and cotton share fivefold WGMs. Numbers 1 to 5 refer to the different subgenomes.

AMK8 fused with AMK9 by end-to-end joining (EEJ) to produce chromosome Chr4 of cacao (*SI Appendix*, Fig. S5).

To further test subgenome relationships, we conducted phylogenetic analyses using the collinear genes on each of the 11 protochromosomes separately, which yielded 11 trees with consistent topologies except for the position of white jute (Grewioideae) relative to the other subfamilies (Fig. 3*A* and *SI Appendix*, Figs. S6–S10). The relationships among the five ancestral subgenomes in these 11 phylogenetic trees agreed with the relationships obtained from the 900 collinear genes when we used Chr10 of cacao (which corresponds to the ancestral Malvaceae chromosome AMK3) as the reference (Fig. 2*B*). These results suggest that the subfamilies within the Malvadendrina clade studied here (four out of seven) are all allopolyploid and combine different ancestral subgenomes. Allopolyploidization occurred repeatedly, with the ancestral subgenomes S_1 and S_2 first combined in the ancestor of allotetraploid Sterculioideae, then further combined with S_3 to produce allohexaploid Helicteroideae, and finally combined with the last two subgenomes (S_4 and S_5) to produce the allodecaploid ancestor of Bombacoideae and Malvoideae. The relationship of white jute (Grewioideae) as sister to all other subgenomes or instead as sister to the Byttnerioideae (represented by cacao and *H. umbratica*) was the only node that differed among the 11 topologies (*SI Appendix*, Fig. S7), and only genes from chromosomes AMK1 and AMK3 yielded the latter topology, supporting the Byttneriina clade.

To further test our results, we used alignments of 1,366 single- or low-copy nuclear sequences and entire plastid genomes to infer taxon relationships (Fig. 3*B*). The nuclear and plastid topologies contradicted each other as also found in earlier studies, albeit with different taxon and gene sampling (24, 34–36). In addition, we conducted phylogenetic network analyses, using the single- or low-copy gene sets, which failed to reveal the repeated ancient hybridizations reflected by the subgenome trees analyzed with the WGDI toolkit (*SI Appendix*, Fig. S11).

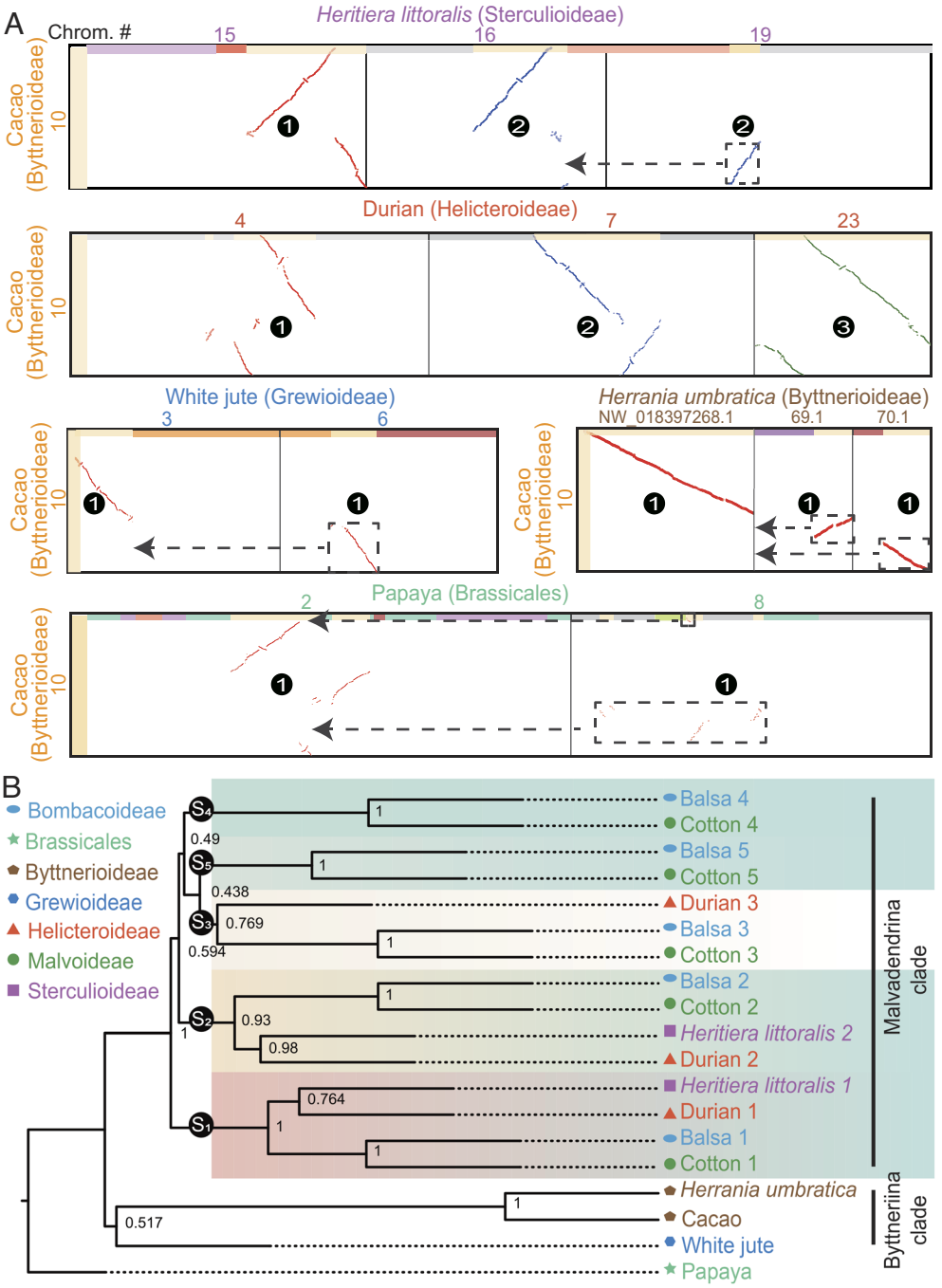

Fig. 2. Evolutionary relationship between Bombacoideae (balsa), Sterculioideae (*Hertiera littoralis*), and Helicteroideae (durian) inferred from syntenic blocks and phylogenetic analyses of collinear genes. (*A*) Syntenic depth ratios observed between cacao and various species. Distinct subgenomes derived from phased synteny blocks, as denoted by dashed boxes, arrows, and numbers. (*B*) A phylogeny of the coexisting subgenomes based on 900 collinear genes from homologous blocks to cacao Chr10 and rooted on papaya (*C. papaya*, Brassicales). Numbers 1 to 5 refer to the different subgenomes. Pie charts at nodes represent quartet frequencies and numbers next to them posterior probabilities.

Relationships among the Subgenomes Inferred from Conserved Chromosomal Changes. Fig. 4 shows the chromosomal fusions or breakpoints that have to be assumed to derive the haploid chromosome numbers of extant species from the 11 Malvaceae protochromosomes (denoted AMK1 through 11; previous section). It also shows the ancestral Malvadendrina (AMDK) and Byttneriina (ABK) clade karyotypes (*SI Appendix*, Figs. S12–S14). Our approach for inferring shared chromosomal fusions from collinearity is illustrated in Fig. 5*A*, using Chrs 13 and 32 of balsa as an example. Collinearity comparisons revealed that Malvaceae protochromosomes AMK8 and AMK11 evolved into Malvadendrina AMDK10 and AMDK11 through reciprocal translocation (RTA) (Fig. 5*A*) and that balsa (Bombacoideae), cotton (Malvoideae), durian (Helicteroideae), and *H. littoralis* (Sterculioideae) have the identical RTA (Fig. 5*B* and *SI Appendix*, Figs. S15–S18). This uniquely shared RTA supports the monophyly of the Malvadendrina clade (with four of its seven subfamilies sampled here). Another phylogenetically informative chromosomal fusion occurred between AMDK8 and AMDK9, which underwent EEJ (Fig. 5*C* and *SI Appendix*, Figs. S5 and S15–S18). This EEJ event uniquely shared by Byttnerioideae and Grewioideae supports the monophyly of the Byttneriina clade (represented by both its subfamilies).

Another chromosomal fusion involves Malvadendrina protochromosomes AMDK3 and AMDK6, which underwent a nested fusion

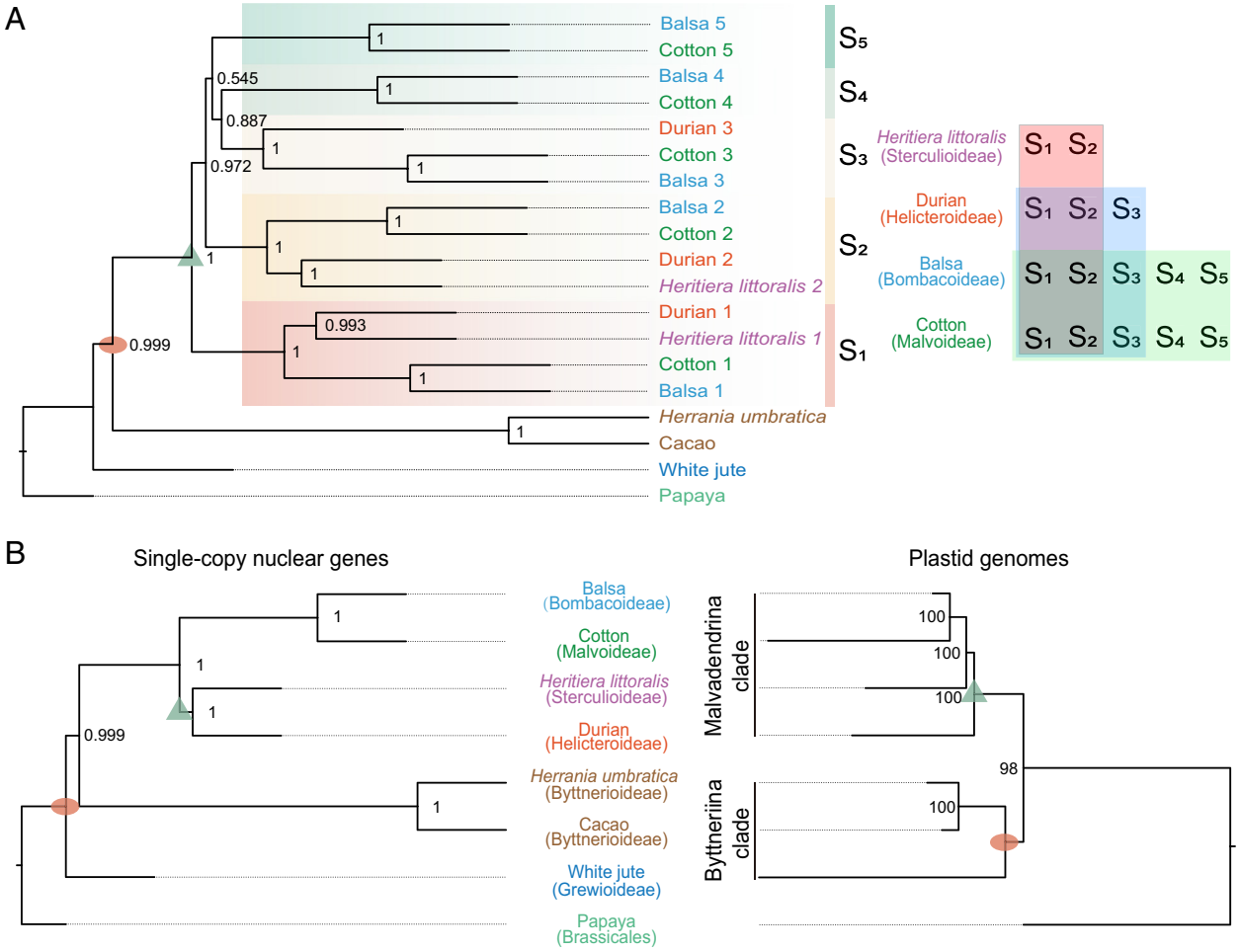

Fig. 3. Phylogenetic relationships and nucleocytoplasmic discordance among representatives of six of the nine Malvaceae subfamilies. (*A*) The consensus topology of the coexisting subgenomes (S_1 to S_5) from the 11 separate phylogenies obtained from 18,094 collinear genes extracted from the 11 Malvaceae protochromosomes, rooted on papaya (*C. papaya*, Brassicales) (*SI Appendix*, Fig. S7). Numbers at nodes represent posterior probabilities. *H. littoralis* contains subgenomes S_1 and S_2, durian contains S_1, S_2, and S_3, while balsa and cotton contain S_1 to S_5. (*B*) Species trees constructed using 1,366 aligned nuclear sequences (*Left*) and entire plastid genomes (*Right*). Conflicting nodes are marked with ellipses and triangles.

to produce another intact chromosome that is retained as Chr28 of balsa (Figs. 4 and 6*A* and *SI Appendix*, Figs. S15–S18). Malvadendrina protochromosomes AMDK8 and AMDK9 instead experienced a reciprocal translocation among two chromosomes, which are retained as Chr 10 and Chr18 of *H. littoralis* (*SI Appendix*, Fig. S15). Similarly, AMDK1 and AMDK4 underwent a reciprocal translocation and are retained as Chr33 and Chr43 of balsa (Figs. 4 and 6*A* and *SI Appendix*, Fig. S17). Malvadendrina protochromosomes AMDKs 1, 3, 4, 6, 8, and 9 are retained as independent chromosomes, indicating that the three fusion events just described occurred after the diversification of this lineage (*SI Appendix*, Figs. S15–S18).

The history of Malvadendrina protochromosome AMDK6 is complex, since it also underwent a reciprocal translocation with AMDK2; both are retained as Chr9 and Chr44 of balsa (Figs. 4 and 6*B*). This AMDK6+ AMDK2 translocation is shared by Helicteroideae, Bombacoideae, and Malvoideae but not by Sterculioideae (Figs. 4 and 6*C* and *SI Appendix*, Fig. S15), suggesting that it may have originated from another ancestral subgenome, S3 (*SI Appendix*, Fig. S19). Moreover, another AMDK2+6 reciprocal translocation and subsequent fusion with other protochromosomes gave rise to Chrs 1 and 30 of balsa and is shared by Bombacoideae and Malvoideae (Fig. 6*C* and *SI Appendix*, Fig. S17*D*). The two AMDK2+6 fusion events have similar fusion positions (*SI Appendix*, Fig. S17*C*), but the fusion positions lack support from clear synteny blocks (Fig. 6*B*). This is perhaps due to gene loss after polyploidization, resulting in sparse synteny blocks and consequently hindering the determination of shared fusion positions. Furthermore, in both subfamilies, protochromosome AMDK6 fused with AMDK1 via end-to-end joining to produce an integrated chromosome, which is retained as Chr37 of balsa (Figs. 4 and 6*C*).

The subgenome relationships inferred from the described chromosome dynamics (Fig. 4) agree well with the topology of the coalescence trees constructed based on 900 or 18,094 collinear genes (Figs. 2*B* and 3*A* and *SI Appendix*, Fig. S7), but not with topologies obtained from 1,366 aligned nuclear genes or entire plastid genomes (Fig. 3*B*).

Besides using *C. papaya* (Brassicales) as an outgroup, we also tested the effects of including *Stellera chamaejasme* (Thymelaeaceae) and *Vatica rassak* (Dipterocarpaceae) as additional outgroups (*SI Appendix*, Table S1). This revealed that *S. chamaejasme* and *V. rassak* each underwent an independent polyploidization event, with their two subgenomes clustering together (*SI Appendix*, Fig. S20). We then compared the *S. chamaejasme* and *V. rassak* genomes with cacao's Chr4 and balsa's Chrs 13 and 32, as these chromosomes contain important fusion positions (Fig. 5). The results indicated that the three genomes (Papaya, *S. chamaejasme*, *V. rassak*) do not share fusion events with the Malvadendrina and Byttneriina clades (*SI Appendix*, Fig. S21). In other words,

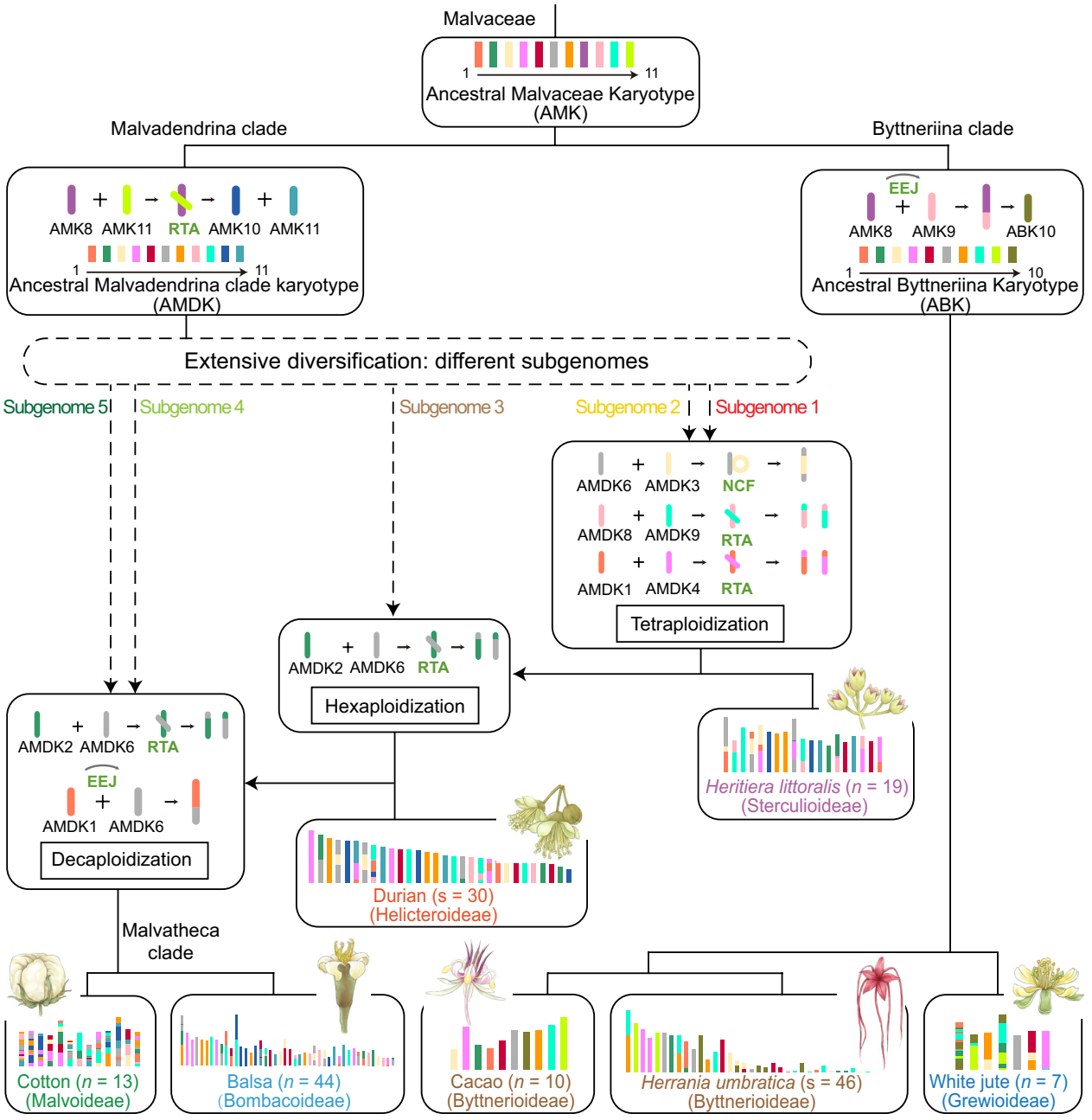

Fig. 4. Evolutionary relationships among representatives of six of the nine subfamilies of the Malvaceae inferred from shared polyploidizations and chromosome fusions. Protochromosomes are color-coded, with the same colors used in all figures. Tetraploidization (two subgenomes), hexaploidization (three subgenomes), and decaploidization (five subgenomes) are marked. The dotted lines indicate relationships not addressed in this study. "*n*" refers to observed haploid chromosome numbers, while "s" refers to genome scaffolds. AMK: Ancestral Malvaceae karyotype; ABK: Ancestral Byttneriina clade karyotype; AMDK: Ancestral Malvadendrina clade karyotype. RTA: reciprocal translocation; EEJ: end-to-end joining; NCF: nested chromosome fusion.

Thymelaeaceae and Dipterocarpaceae do not directly share subgenomes with the Malvadendrina and Byttneriina clades.

Discussion

Although previous studies have inferred complex polyploidization histories among groups of Malvaceae (24–27, 37), hybridizations, accompanied by genome doubling, during the deep history of the family were so far largely unknown. By adding a chromosome-level genome assembly of balsa, we increased the sampling of deep lineages in the Malvaceae, although our analysis lacks representatives of the Brownlowioideae, Dombeyoideae, and Tilioideae, all three belonging to the Malvadendrina clade (21–23). Because we included both subfamilies of the Byttneriina clade, the sister clade to Malvadendrina, our sampling spans the root of Malvaceae and therefore the karyotype evolution through the family's 60 My of evolution (31).

Our results reveal that the Helicteroideae are an allohexaploid lineage that derives from the hybridization between an allotetraploid Sterculioideae species and an unknown diploid species. A descendent of this allohexaploid lineage in turn became one of the parents in another hybridization that gave rise to the allodecaploid ancestor of the sister groups Bombacoideae and Malvoideae. The diversification of the Malvadendrina subfamily cluster, today comprising some 1,000 species, was accompanied by multiple irreversible chromosomal fusion events (Fig. 4). As more chromosome-level genomes become available, especially of Brownlowioideae, Dombeyoideae, and Tilioideae, the subfamily assignments of the so-far unknown

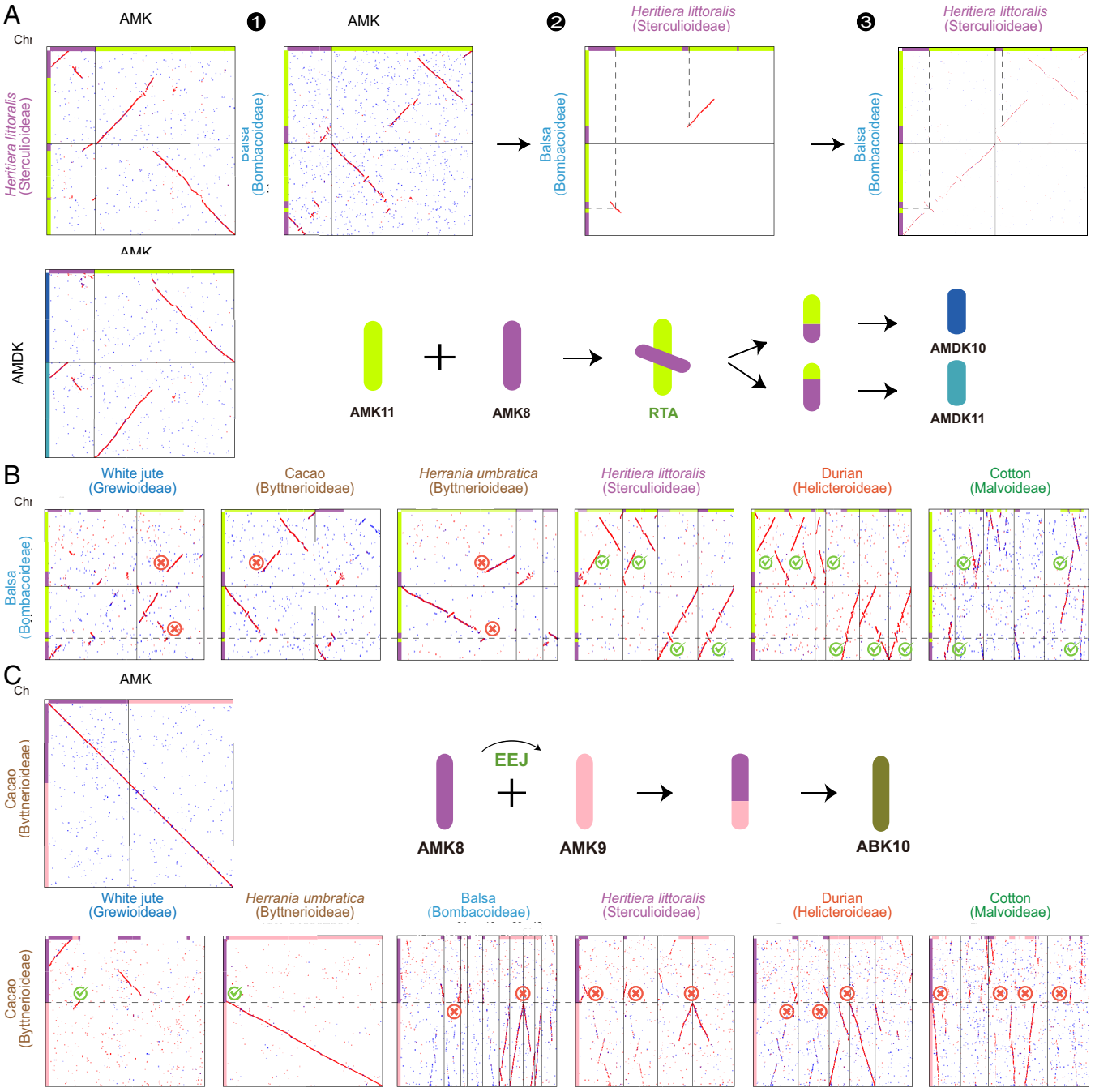

Fig. 5. Workflow for inferring evolutionary relationships from uniquely shared chromosome fusions. (*A*) Schematic of our approach for determining fusion positions. (1) Mapping color-coded protochromosomes to modern genomes based on collinearity. (2) Identification of syntenic blocks spanning different colors and their corresponding positions within the syntenic blocks of modern genomes. These points are the positions of the protochromosomes fusion event. (3) From the collinearity of all regions corresponding to the original chromosomes involved, the specific type of the protochromosomes fusion can be inferred. AMK: Ancestral Malvaceae karyotype; AMDK: Ancestral Malvadendrina clade karyotype. RTA: reciprocal translocation. (*B*) A fusion uniquely shared in the Malvadendrina clade (*H. littoralis*, durian, balsa, and cotton). A region is marked by a check mark (✓) if it is collinear to a syntenic block in a protochromosome and by a red circle with an "✗" if it is broken. (*C*) A fusion uniquely shared in the Byttneriina clade (white jute, cacao, *H. umbratica*). ABK: Ancestral Byttneriina clade karyotype; EEJ: end-to-end joining.

parents may eventually be recovered (*SI Appendix*, Fig. S22). It also may be feasible to detect whether particular fusion events preceded or followed allopolyploidization.

The subgenome relationships obtained from 900 or 18,094 collinear genes (Figs. 2*B* and 3*A*) do not agree with the topology obtained from DNA alignments of 1,366 single- or at least low-copy nuclear genes nor that obtained from plastid genomes (Fig. 3*B*; note that we did not reuse the data of Conover et al. (24), who used transcriptomes (plus the transcriptome of *Dombeya*) to extract nuclear genes, but instead relied on the de novo genome assemblies). These discordances may be due to noise introduced by single-copy genes that stem from different ancestral subgenomes, although this explanation cannot hold for the plastid genome topology, which may instead reflect instances of plastid capture or differing maternal parentages in hybridizations. We posit that chromosome structural changes, namely reciprocal translocations, nested fusions, and end-to-end fusions, provide strong evidence for subgenome relationships (because chromosome structural changes may be less prone to homoplasy than alignments

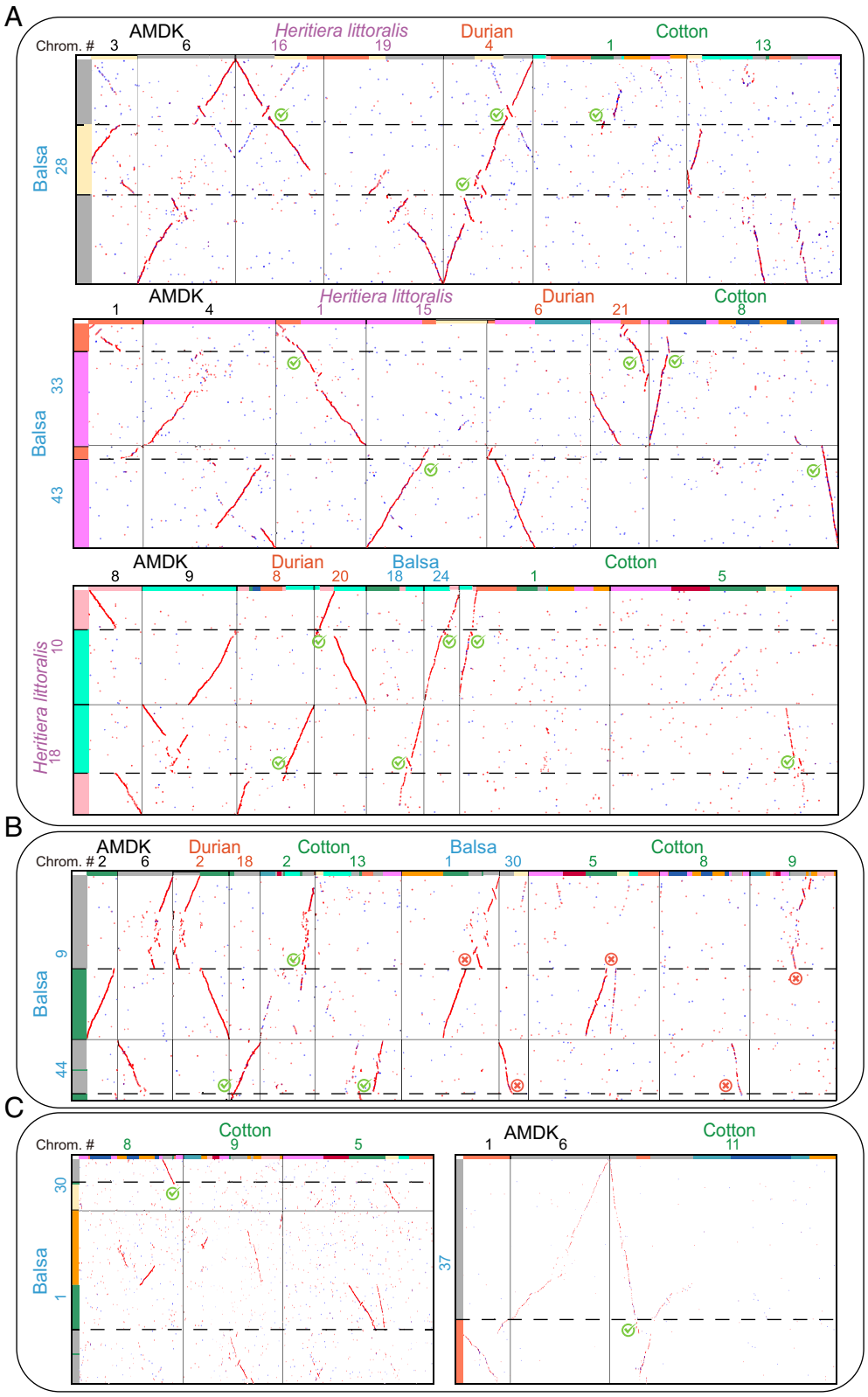

Fig. 6. Shared fusion events in protochromosomes support the successive hybridization-cum-genome-doubling events in the Malvadendrina clade. (*A*) The tetraploid ancestor of balsa, cotton, durian, and *H. littoralis* underwent three chromosomal fusion events. A region is marked by a check mark (✓) if it is collinear to a syntenic block in a protochromosome and by a red circle with an "✗" if it is broken. (*B*) Balsa and cotton had two AMDK2+6 fusion events (*SI Appendix*, Fig. S17), only one of which was inherited from their common ancestor with durian. (*C*) The decaploid ancestor of balsa and cotton underwent two other chromosomal fusion events.

of possibly paralogous gene copies). Although shared polyploidy events are currently acknowledged as evidence of evolutionary relationships, we anticipate that shared chromosome structural changes will also emerge as independent of phylogenetic relationships.

The frequency of the three types of chromosome fusions is striking (*SI Appendix*, Figs. S5*B*, S15*D*, S16*D*, and S17*D*). Thus, for balsa, with $2n = 88$, we inferred four reciprocal translocations, one nested fusion, and 10 end-to-end fusions (*SI Appendix*, Fig. S17*D*). Cotton

($2n = 26$) retains no independent Malvaceae or Malvadendrina protochromosomes, implying especially drastic reshuffling (*SI Appendix*, Fig. S18) as also inferred in earlier studies (27, 38). The inferred ancestral haploid chromosome numbers of $n = 11$ for the family and $n = 10$ for the Byttneriina clade (Fig. 4) reject hypothesized ancestral haploid numbers of $n = 6$, 7, or 8 (39).

In summary, using subgenome-aware analyses and our paleogenomic approach (13), we established reticulate relationships among the Malvaceae subgenomes, with long-term effects of allopolyploidy over perhaps 60 My of evolution (31). Similar speciation events involving hybridization, with the offspring becoming viable and setting off onto their own evolutionary trajectories through genome duplication as predicted by Stebbins (1), likely continue today, as exemplified by the repeated allopolyploid events within the cotton lineage (27, 38). Ancient reticulations during plant evolution have long been suspected but can be inferred with confidence only now, from comparisons of chromosome-level genome assemblies (12, 40). Our case study of the Malvaceae may stimulate similar paleogenomic studies for other tribes, subfamilies, families, or even orders with suspected ancient allopolyploid origins, such as the Ericales (41) and Juglandaceae (11).

Methods

Normalized Homology Score and Normalized Retention in Dot Plot. We used the WGDI toolkit (13) with the parameter "-d" to plot homologous gene dot plots within and between genomes. Each dot plot represents homologous gene pairs as dots, with red dots labeled 1, blue dots labeled 0, and gray dots labeled −1. Normalized homology scores refer to the average score assigned to gene pairs of a synteny block. Normalized retention refers to the proportion of syntenic gene pairs of a block compared to the total number of genes in the corresponding chromosomal region.

Polyploidization and Phylogenetic Analysis. We inferred the number of genome duplications by using the syntenic depth ratios of mostly red dots in homologous gene dot plots (Fig. 2 and *SI Appendix*, Fig. S4). In the synteny analyses, collinear genes were identified with the parameter "-icl" of WGDI (13), both within each genome and between genomes, and dot plots were used to calculate the syntenic ratios between different species to confirm the polyploidy level of each species. Frequencies of synonymous substitutions per synonymous site (*Ks* values) between collinear genes were estimated using the Nei-Gojobori approach as implemented in PAML (42) through WGDI with the parameter "-ks." The *Ks* values of gene pairs in synteny blocks are displayed in dot plots through WGDI with the "-bk" parameter. The median *Ks* values of each block were selected to perform *Ks* peak fitting by WGDI with the parameter "-pf."

We performed phylogenetic analysis using collinear genes. The workflow is shown in *SI Appendix*, Fig. S3, and the detailed process was as follows: First, we filtered the synteny blocks retained after lineage-specific polyploidization events through WGDI with the parameter "-c." Next, we mapped the ancestral karyotype protochromosomes onto current species based on collinearity through WGDI with the parameter "-km" and divided them into different subgenomes according to their ploidizations. Third, we recorded the subgenome regions on the chromosomes and used WGDI with the parameters "-pc" and "-a" to obtain the hierarchical gene lists. Then, the hierarchical gene lists were used to infer maximum likelihood (ML) trees using IQ-TREE (43) through WGDI with the parameter "-at." Finally, we used ASTRAL-III v.5.7.8 (44) with the parameter "-t 16" and ASTRAL-Pro 2 (45) with the parameter "-u 3" to construct the coalescent tree and estimate branch support.

We used SonicParanoid2 (46) to identify single- or low-copy nuclear sequences and GetOrganelle (47) to assemble the plastid genome for balsa. The estimation of phylogenetic networks based on the single- or low-copy nuclear sequences was conducted using the InferNetwork_MP_Allopp method from PhyloNet package (48, 49). The entire plastid genomes of other species were downloaded from NCBI (*SI Appendix*, Table S9).

Karyotypic Evolution Starting from the AMK to the Extant Sampled Species, and Construction of Subgenome Relationships. We compared all chromosomes of the sampled Malvaceae with the 11 inferred AMK protochromosomes and determined the distribution of these ancestral chromosomes in current genomes using the WGDI toolkit with the parameter "-km." Then, we compared permutations and combinations of different protochromosome in silico color-coded stretches between extant genomes using WGDI with the parameter "-sf" to rapidly identify chromosomal fusions, which were used for evolutionary inference based on shared types of fusions (or fissions) and their position on the chromosomes. Finally, we inferred the phylogenetic relationships between species by examining in which species the fusion positions in the dot plots were shared.

Data, Materials, and Software Availability. All raw sequence reads used in this study have been uploaded to the Genome Warehouse in BIG Data Centre under the BioProject accession number PRJCA016829 (50). Genome annotations, karyotypes, and phylogenetic data have been deposited in GitHub (https://github.com/SunPengChuan/Ancestral_Malvaceae_Karyotype). All other data are included in the manuscript and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank Yelin Huang (Sun Yat-sen University, CHN) for the newly annotated *H. littoralis* genome, and Liang Liu (University of Georgia, USA), Bin Tian (Southwest Forestry University, CHN), and Xiyin Wang (North China University of Science and Technology, CHN) for comments on the phylogenetic analysis. This work was supported equally by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31000000) and the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502), and further by the National Key Research and Development Program of China (2017YFC0505203), the National Natural Science Foundation of China (grant numbers 31590821, 91731301, and 31561123001), Fundamental Research Funds for the Central Universities (YJ201936, 2020SCUNL20, SCU2019D013 and 2020SCUNL207, SCU2022D003, and lzujbky-2022-ey07), Young Talent Development Project of State Key Laboratory of Herbage Improvement and Grassland Agro-ecosystems (No. 2021+02), and National High-Level Talents Special Support Plans.

Author affiliations: [a]Key Laboratory for Bio-Resources and Eco-Environment, Sichuan Zoige Alpine Wetland Ecosystem National Observation and Research Station, College of Life Sciences, Sichuan University, Chengdu 610065, China; [b]Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan 666303, China; [c]State Key Laboratory of Herbage Improvement and Grassland Agro-Ecosystem, College of Ecology, Lanzhou University, Lanzhou 730000, China; and [d]Department of Biology, Washington University, Saint Louis, MO 63105

1. G. L. Stebbins, The significance of polyploidy in plant evolution. *Am. Nat.* **74**, 54–66 (1940).
2. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
3. Y. Van de Peer, T. L. Ashman, P. S. Soltis, D. E. Soltis, Polyploidy: An evolutionary and ecological force in stressful times. *Plant Cell* **33**, 11–26 (2021).
4. D. E. Soltis *et al.*, Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): Cytogenetic, genomic and genetic comparisons. *Biol. J. Linn. Soc. Lond.* **82**, 485–501 (2004).
5. P. S. Soltis, D. E. Soltis, The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* **60**, 561–588 (2009).
6. K. H. Wolfe, Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**, 333–341 (2001).
7. D. Sankoff, J. H. Nadeau, Conserved synteny as a measure of genomic distance. *Discret. Appl. Math.* **71**, 247–257 (1996).
8. B. Snel, P. Bork, M. A. Huynen, Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110 (1999).
9. A. S. Chanderbali *et al.*, *Buxus* and *Tetracentron* genomes help resolve eudicot genome history. *Nat. Commun.* **13**, 643 (2022).
10. Z. Wang *et al.*, A high-quality *Buxus austro-yunnanensis* (Buxales) genome provides new insights into karyotype evolution in early eudicots. *BMC Biol.* **20**, 1–17 (2022).
11. Y.-M. Ding *et al.*, Genome structure-based Juglandaceae phylogenies contradict alignment-based phylogenies and substitution rates vary with DNA repair genes. *Nat. Commun.* **14**, 617 (2023).
12. R.-G. Zhang *et al.*, Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes. *Nat. Commun.* **14**, 2204 (2023).
13. P. Sun *et al.*, WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).

14. O. Simakov *et al.*, Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* **8**, eabi5884 (2022).
15. D. T. Schultz *et al.*, Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
16. T. Mandakova, M. A. Lysak, Chromosomal phylogeny and karyotype evolution in x = 7 crucifer species (Brassicaceae). *Plant Cell* **20**, 2559–2570 (2008).
17. F. Murat *et al.*, Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**, 1545–1557 (2010).
18. F. Murat, A. Armero, C. Pont, C. Klopp, J. Salse, Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* **49**, 490–496 (2017).
19. J. Salse, Deciphering the evolutionary interplay between subgenomes following polyploidy: A paleogenomics approach in grasses. *Amer. J. Bot.* **103**, 1167–1174 (2016).
20. P. F. Stevens, (2001 onwards). Angiosperm Phylogeny Website. Version 14, July 2017 [and more or less continuously updated since], https://www.mobot.org/MOBOT/research/APweb/. Accessed 7 July 2023.
21. W. S. Alverson, B. A. Whitlock, R. Nyffeler, C. Bayer, D. A. Baum, Phylogeny of the core Malvales: Evidence from *ndhF* sequence data. *Am. J. Bot.* **86**, 1474–1486 (1999).
22. C. Bayer *et al.*, 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 (1999).
23. T. Cvetković *et al.*, Phylogenomics resolves deep subfamilial relationships in Malvaceae s.l. *G3* **11**, jkab136 (2021).
24. J. L. Conover *et al.*, A Malvaceae mystery: A mallow maelstrom of genome multiplications and maybe misleading methods? *J. Integr. Plant Biol.* **61**, 12–31 (2019).
25. B. Teh *et al.*, The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat. Genet.* **49**, 1633–1641 (2017).
26. J. Wang *et al.*, Recursive paleohexaploidization shaped the durian genome. *Plant Physiol.* **179**, 209–219 (2019).
27. Z. J. Chen *et al.*, Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533 (2020).
28. R. Hernandez-Gutierrez, S. Magallon, The timing of Malvales evolution: Incorporating its extensive fossil record to inform about lineage diversification. *Mol. Phylogenet. Evol.* **140**, 106606 (2019).
29. J. G. Carvalho-Sobrinho *et al.*, Revisiting the phylogeny of Bombacoideae (Malvaceae): Novel relationships, morphologically cohesive clades, and a new tribal classification based on multilocus phylogenetic analyses. *Mol. Phylogenet. Evol.* **101**, 56–74 (2016).
30. S. K. Sahu *et al.*, Chromosome-scale genomes of commercial timber trees (*Ochroma pyramidale*, *Mesua ferrea*, and *Tectona grandis*). *Sci. Data* **10**, 512 (2023).
31. M. Carvalho, F. Herrera, C. Jaramillo, S. Wing, R. Callejas, Paleocene Malvaceae from northern South America and their biogeographical implications. *Am. J. Bot.* **98**, 1337–1355 (2011).
32. D. A. Baum *et al.*, Phylogenetic relationships of Malvatheca (Bombacoideae and Malvoideae; Malvaceae sensu lato) as inferred from plastid DNA sequences. *Am. J. Bot.* **91**, 1863–1871 (2004).
33. F. Meng *et al.*, Cotton duplicated genes produced by polyploidy show significantly elevated and unbalanced evolutionary rates, overwhelmingly perturbing gene tree topology. *Front. Genet.* **11**, 239 (2020).
34. M. S. Islam *et al.*, Comparative genomics of two jute species and insight into fibre biogenesis. *Nat. Plants* **3**, 1–7 (2017).
35. Y. Gao *et al.*, De novo genome assembly of the red silk cotton tree (*Bombax ceiba*). *GigaScience* **7**, 1–7 (2018).
36. H. Hu, P. Sun, Y. Yang, J. Ma, J. Liu, Genome-scale angiosperm phylogenies based on nuclear, plastome, and mitochondrial datasets. *J. Integr. Plant Biol.* **00**, 1–12 (2023).
37. K. L. Adams, R. Cronn, R. Percifield, J. F. Wendel, Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl Acad. Sci. U.S.A.* **100**, 4649–4654 (2003).
38. A. H. Paterson *et al.*, Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
39. A. Carta, G. Bedini, L. Peruzzi, A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytol.* **228**, 1097–1106 (2020).
40. T. Mitros *et al.*, Genome biology of the paleotetraploid perennial biomass crop *Miscanthus*. *Nat. Commun.* **11**, 5442 (2020).
41. S. Nie *et al.*, Potential allopolyploid origin of Ericales revealed with gene-tree reconciliation. *Front. Plant Sci.* **13**, 1006904 (2022).
42. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
43. B. Q. Minh *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. and Evol.* **37**, 1530–1534 (2020).
44. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 15–30 (2018).
45. C. Zhang, S. Mirarab, ASTRAL-Pro 2: Ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* **38**, 4949–4950 (2022).
46. S. Cosentino, W. Iwasaki, SonicParanoid2: Fast, accurate, and comprehensive orthology inference with machine learning and language models. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.05.14.540736 (Accessed 14 May 2023).
47. J. Jin *et al.*, GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 1–31 (2020).
48. Z. Yan, Z. Cao, Y. Liu, H. A. Ogilvie, L. Nakhleh, Maximum parsimony inference of phylogenetic networks in the presence of polyploid complexes. *Syst. Biol.* **71**, 706–720 (2022).
49. C. Than, D. Ruths, L. Nakhleh, PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinform.* **9**, 1–16 (2008).
50. P. Sun *et al.*, Subgenome-aware analyses reveal the genomic consequences of ancient allopolyploid hybridizations throughout the cotton family (March 2024), https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA016829. Accessed 9 March 2024.