Integrated Syntenic and Phylogenomic Analyses Reveal an
Ancient Genome Duplication in Monocots

Yuannian Jiao,a,1 Jingping Li,a,b,1 Haibao Tang,c,d,1 and Andrew H. Patersona,2

1 Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602
2 Institute of Bioinformatics, University of Georgia, Athens, Georgia 30602
3 Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian Province 350002, China
4 J. Craig Venter Institute, Rockville, Maryland 20850

ORCID IDs: 0000-0002-8987-2782 (Y.J.); 0000-0002-3460-8570 (H.T.)

Unraveling widespread polyploidy events throughout plant evolution is a necessity for inferring the impacts of whole-genome duplication (WGD) on speciation, functional innovations, and to guide identification of true orthologs in divergent taxa. Here, we employed an integrated syntenic and phylogenomic analyses to reveal an ancient WGD that shaped the genomes of all commelind monocots, including grasses, bromeliads, bananas (Musa acuminata), ginger, palms, and other plants of fundamental, agricultural, and/or horticultural interest. First, comprehensive phylogenomic analyses revealed 1421 putative gene families that retained ancient duplication shared by Musa (Zingiberales) and grass (Poales) genomes, indicating an ancient WGD in monocots. Intergenic syteny blocks of Musa and Oryza were investigated, and 30 blocks were shown to be duplicated before Musa-Oryza divergence an estimated 120 to 150 million years ago. Synten comparisons of four monocot (rice [Oryza sativa], sorghum [Sorghum bicolor], banana, and oil palm [Elaeis guineensis]) and two eudicot (grape [Vitis vinifera] and sacred lotus [Nelumbo nucifera]) genomes also support this additional WGD in monocots, herein called Tau (τ). Integrating synteny and phylogenomic comparisons achieves better resolution of ancient polyploidy events than either approach individually, a principle that is exemplified in the disambiguation of a WGD series of rho (ρ)-sigma (σ)-tau (τ) in the grass lineages that echoes the alpha (α)-beta (β)-gamma (γ) series previously revealed in the Arabidopsis thaliana lineage.

INTRODUCTION

Whole-genome duplication (WGD), or polyploidy, is a primary source of raw genetic material for evolutionary novelties (Ohno, 1970; Lynch and Conery, 2000; Adams and Wendel, 2005; Soltis et al., 2009). Although a straight change in ploidy is generally expected to be deleterious and often lead to an evolutionary dead end (Otto and Whitton, 2000; Mayrose et al., 2011; Arrigo and Barker, 2012), many modern organisms have descended from paleopolyploidized ancestors, including plants (Grant et al., 2000; Vision et al., 2000; Bowers et al., 2003; Blanc and Wolfe, 2004; Jaillon et al., 2007; Ming et al., 2008; Tang et al., 2008a; Barker et al., 2009; Fawcett et al., 2009; Paterson et al., 2009; International Brachypodium Initiative, 2010; Jiao et al., 2011; Wang et al., 2011; D’Hont et al., 2012; Amborella Genome Project, 2013; Singh et al., 2013). However, although several paleopolyploidy events have been intensively studied in eudicots, including the pan-core eudicot γ event (Jaillon et al., 2007; Tang et al., 2008b; Jiao et al., 2012; Ming et al., 2013), most polyploidies in monocots have remained elusive due to the availability of fewer sequenced lineages.

The first WGD event discovered in the ancestral lineage leading to modern cereals (Poaceae family) was named “p,” first estimated to have occurred ~70 million years ago (MYA) based on molecular data (Paterson et al., 2004; Wang et al., 2005), and possibly being ~30 million years older considering recent fossil discoveries (Prasad et al., 2005). The main Poaceae clades (BEP and PACCAD) evolved as separate lineages for about two-thirds of the time since p (Paterson et al., 2004, 2009). Nonetheless, in those family members that have not experienced further genome duplications and subsequent diploidization, such as rice (Oryza sativa) and sorghum (Sorghum bicolor), synten conservatism spans large genomic regions covering hundreds to thousands of orthologous genes (Paterson et al., 2009). The rice and sorghum genomes also share 97 to 98% of post-p gene loss in orthologous regions (Paterson et al., 2009), typical of genome comparisons among other cereals (Schnable et al., 2009; International Brachypodium Initiative, 2010; Bennetzen et al., 2012) and supporting the deduction that the radiation of the major cereal lineages occurred ~20 million years or more after p (Paterson et al., 2009). The banana (Musa acuminata) genome has undergone three successive rounds of WGD since its divergence with grasses...
The oil palm (Elaeis guineensis) and date palm (Phoenix dactylifera) genomes share one recent WGD (Al-Mssallem et al., 2013; Singh et al., 2013). Previous genomic analysis in the banana and oil palm provided no indications of additional more ancient WGD before the divergence of Poales and Zingiberales (D’Hont et al., 2012; Singh et al., 2013).

Nested within homoeologous regions formed in well-characterized genome duplications are sometimes found additional paralogs from more ancient paleopolyploidy events that have been obscured by subsequent reduplications and rediploidizations. For example, the event called “α” in our earlier study (Tang et al., 2010) was suspected to comprise multiple events that could not be resolved due to lack of evidence from multiple monocot lineages needed for disambiguation.

Here, by comparing the genomes of the monocots rice, sorghum (Poales), oil palm (Arecales), and banana (Zingiberales), and eudicot lineages needed for disambiguation.

**RESULTS**

**Phylogenomic Analysis**

To investigate ancient WGDs, 15 monocot genomes were used to construct putative gene families or orthologs (see Methods). Of the 39,427 total gene families, 11,503 containing at least one monocot and one eudicot sequence were used to construct maximum likelihood trees. Using a gene duplication scoring strategy previously described (Jiao et al., 2011), shared duplications (Figure 2) were identified from constructed gene family phylogenies after monocot-eudicot divergence but before grass-banana divergence (1450 duplications in 1421 gene families with bootstrap support [BS] ≥50%; 552 duplications in 540 families with BS ≥80%). Most gene families retaining high levels of shared duplicates are statistically enriched in Gene Ontology categories corresponding to kinase, transferase, transporter, transcription regulator, and transcription factor (Supplemental Data Set 1), consistent with the general postgenome duplication survivorship patterns (Maere et al., 2005; Paterson et al., 2006; Freeling, 2009).

Orthogroup gene trees show excesses of internal nodes potentially corresponding to WGD events. For example, WRKY transcription factors (orthogroup 1297) survived ancient duplication shared by grasses, palms, and Musa after divergence from eudicots (Supplemental Figure 1). Additional younger lineage-specific duplication events within the sampled grasses, palms, and Musa were largely consistent with published views of WGD timing (ρ and σ in grasses, three WGDs in Musa, and one in a palm progenitor) (Tang et al., 2010; D’Hont et al., 2012). Another exemplary gene family is the C3HC4-type RING zinc finger transcription factor family (orthogroup 1312), which showed almost maximal retention across many lineages, surviving two nested duplications in grass lineages, one duplication shared by palms, two nested duplications in Musa lineages, and the gamma duplication in eudicots (Supplemental Figure 2).

**Multiple Synteny Alignments of Musa Using Oryza as Reference**

We performed a careful analysis of Oryza-Musa synteny to pinpoint phylogenetic signals of deep-time monocot duplication events, identifying 1021 syntenic blocks using MCScanX (Wang et al., 2012), including 9371 gene pairs. Syntenic regions of Musa were then aligned against Oryza chromosomes, with 22,223 Oryza genes in regions which matched one to eight different Musa regions (hereinafter denoted as homologous clusters) and 513 Oryza genes in regions which matched 9–15 Musa regions. The redundancy ratio of 1 Oryza:8 Musa is consistent with three genome doublings in the Musa genome, while redundancy of 1 to 9–15 suggests at least one additional WGD.

Duplicated blocks were further dated by mapping the syntenic anchors onto phylogenetic trees (see Methods). Supplemental Figure 3A shows a homologous cluster of eight Musa regions aligned to Oryza chromosome 5. Most anchor genes on these eight regions were determined phylogenetically as having been duplicated after the Musa-Oryza divergence (Supplemental Figure 3B). In total, we found 47 such homologous clusters (1 Oryza to 4+ Musa) duplicated after Musa-Oryza divergence, supporting three previously identified Musa WGDs. We also found 30 homologous clusters that were generated by a more ancient duplication, before Musa-Oryza divergence. For example, one Oryza region (Chr8: 3.8 to 10.5 Mb) matched up to 11 regions in Musa (Figure 3A). The Musa region groups A (Chr2: 17.1 to 17.3 Mb and Chr6: 32.9 to 33.0 Mb) and B (Chr5: 28.4 to 28.8 Mb, Chr6: 9.9 to 10.5 Mb and 32.1 to 32.6 Mb, Chr7: 12.0 to 12.5 Mb, Chrom: 25.8 to 26.0 Mb, Chr9: 2.4 to 3.0 Mb, Chr10: 7.2 to 7.4 Mb, 21.9 to 22.2 Mb and 27.7 to 28.0 Mb, Chr11: 23.6 to 24.2 Mb) were duplicated after the divergence of Musa and Oryza, respectively, while the duplication of the ancestral blocks of A and B occurred before the divergence of Musa and Oryza (Figure 3B).

**Overview of Synteny Conservation among Six Representative Genomes**

Genome alignments of four monocot (rice, sorghum, banana, and oil palm) and two eudicot (grape and sacred lotus) species (Supplemental Table 1) revealed clear evidence of ρ and σ. In total, 97.66% (38,136) of rice genes, 98.70% (34,046) of sorghum genes, 96.50% (35,272) of banana genes, 86.10% (27,746) of oil palm genes, 95.76% (25,228) of grape genes, and 92.87% (24,782) of sacred lotus genes were covered in the aligned regions.

Comparison between lotus and oil palm genomes showed a mode of 2-to-4 multiplicity ratio between orthologs (Table 1),
with the “2” indicative of the lotus lambda (λ) duplication (Ming et al., 2013) and the “4” indicating two paleotetraploidy events in the oil palm lineage that include τ and palm-specific “p” (Figure 4A) (Singh et al., 2013). The p WGD in oil palm and the ρ and p WGDs in rice after their separation were reflected by the 2-to-4 multiplicity ratio. Collectively, the ploidy ratios across these pairwise comparisons indicate one WGD (σ) in the lineage leading to grasses before ρ and an additional ancestral event (τ) preceding the Poaceae-Arecaceae divergence (Figure 4). These events imply a total of 8 × paleo-multiplicity in the rice genome after monocot-dicot divergence, as confirmed by both lotus-rice and grape-rice comparisons (Table 1).

With successive WGDs followed by loss of most duplicated genes, ancestral gene orders become progressively more fragmented and discernible synteny blocks become smaller (Table 1). The number of anchor genes tends to show a negative correlation with species divergence (Table 1), indicating that reciprocal gene loss and/or gene transposition is a persistent process on a large evolutionary scale. Nonetheless, differential retention of ancestral loci in early Poales branches has slowed down in the cereal genomes after the most recent WGD, ρ. About 4.68% of banana and 6.85% of oil palm genes have no orthologs in the other studied monocots but do have grape and lotus (eudicot) orthologs. This percentage is smaller in rice (3.00%) and sorghum (3.51%). However, the majority (~80%) of ancestral loci that had been differentially retained in the lineage leading to cereals (since their divergence from other Commelinids 120 to ~83 MYA; The Angiosperm Phylogeny Group, 2009) are still syntenic in present-day rice and sorghum genomes. The remaining ~20% have been differentially lost or transposed since the divergence of the two species 80 to ~50 MYA (The Angiosperm Phylogeny Group, 2009).

**Circumscribing the σ Duplication**

Our grape-rice genome comparison showed that there are more paralogous regions in rice than those produced in the pan-grass ρ, indicating more ancient paleopolyploidizations designated as

---

**Figure 1.** Schematic Diagram Detailing an Integrated Synteny and Phylogenomic Approach for Circumscribing WGD Events on a Species Phylogeny.

(A) Flowchart of the methodology.
(B) Rationale of ordering speciation and WGD events in the history of two lineages. Chromosome segments are represented by gray bars, and genes are represented by colored bricks.
(C) Illustration of homoeologous genes (anchors) mapping onto the phylogeny of a homologous gene family in the phylogenomic approach used to date these homologous regions.
(Tang et al., 2010). Such nested WGDs are often best revealed by bottom-up approaches (Bowers et al., 2003), attempting to reverse the changes resulting from more recent events (in this case ρ) superimposed on them. Using such bottom-up reconstruction (see Methods), a total of 146 ρ (palm WGD) blocks covering 71.74% of the oil palm genome were merged into pre-ρ ancestral gene orders. Similarly, 140 σ (most recent grass WGD) blocks covering 70.63% of the rice genome were merged into pre-σ gene orders. Direct comparison between the ancestral orders resulted in 209 synteny blocks covering 96.34% of the pre-σ order and 72.77% of the pre-ρ order, each pre-ρ region aligning with up to two paralogous pre-σ regions, as exemplified in Figure 4B. This clearly revealed that the σ event occurred before the pan-Poaceae ρ event (~70 MYA; Paterson et al., 2004) but after the Poaceae-Arecaceae split (120 to ~83 MYA; The Angiosperm Phylogeny Group, 2009).

As the second largest monocot family and fifth largest angiosperm family, the Poaceae is rich in morphological and ecological diversity. Much of the underlying genetic diversity may have resulted from σ and ρ WGDs. The events may also have contributed to some grass-specific characters. Genomes from basal Poales species such as pineapple (Ananas comosus) could be useful to further narrow the dating of ρ and σ.

Unraveling the Precommelinid τ Duplication Event in Early Monocots

To simplify the study of τ in the deep monocot lineages, we compared oil palm and sacred lotus, each having experienced only one subsequent genome duplication. We reconstructed the oil palm pre-ρ and the sacred lotus pre-λ gene orders as described above for σ, covering 71.74 and 74.05% of the respective genomes. Alignment of the reconstructed genomes resulted in 203 ancestral synteny blocks covering 69.44% of pre-λ order and...
73.53% of pre-p order. One pre-λ region generally matched up to two pre-p regions (Figure 4B), indicating that an oil palm ancestor experienced another more ancient paleotetraploidy event before p but after it diverged with the eudicots. Syntenic regions with more than two homoeologous copies retained are distributed across all 16 oil palm chromosomes, confirming that this event was genome-wide. Since we have shown above that oil palm did not share σ, this new paleotetraploidy is inferred to be τ (Figure 4A).

Using the eudicot outgroup genomes of grape or lotus, which only had one paleopolyploid in their own lineages, up to four orthologous regions can be identified in oil palm, indicating two WGDs, τ and ρ, and up to eight orthologous regions can be identified in rice, indicating three WGDs, τ, σ, and ρ (Figure 5). However, fast evolutionary rates and three WGDs in rice have severely increased paleopolyploid loss, with only 10,877 paralogous copies of pre-τ loci retained (excluding tandem duplications) among 39,049 rice genes. In contrast, oil palm evolved at a much slower rate (about one half that of grasses) and had one less WGD. About 16,092 pre-τ paralogs are retained (excluding tandem duplications) among 32,225 oil palm genes. In both genomes, recurring WGDs severely eroded signals of paleo-paralogy from τ, with pre-τ loci retained at 1.4 copies on average in rice (versus 8 at full retention), and 1.5 copies on average in oil palm (versus 4 at full retention).

**Simultaneous Circumscription of Paleopolyploidy Events in Multiple Lineages**

To integrate many of the individual syntenic analyses described above, a streamlined procedure (see Methods) was applied to the rice, sorghum, banana, oil palm, grape, and sacred lotus genomes. The six genomes were compared in a pairwise manner and segmented at boundaries of syntenic blocks, forming sets of putative ancestral regions (PARs; clusters of homoeologous regions with high anchor density) as described (Tang et al., 2010). By integrating synteny mapping information from 1217 groups of PARs from 15 pairs of genome comparisons, paleopolyploidy levels were inferred on each branch of the given species phylogeny (Supplemental Figure 4). Two alternative topologies of the genome were inferred on each branch of the given species phylogeny (Supplemental Figure 4). Two alternative topologies of the species tree were used to re

**DISCUSSION**

Integrating synteny and phylogenomic comparisons achieves better resolution of ancient polyploidy events than either approach individually, a principle that is exemplified in the disambiguation of a WGD series of $\rho - \alpha - \tau$ in the grass lineages that echoes the $\alpha - \beta - \gamma$ series previously revealed in the Arabidopsis lineage.

Using this integrated approach, we were able to reveal a WGD, τ, in the early evolution of monocots, also confirming nine previously identified WGDs (Figure 4). Such additional WGD was hinted at previously (Tang et al., 2010) but not discernible due to lack of genomic data from non-grass monocots. We infer τ to have occurred near the origin of monocots, before the origin of commeliniids (which spans all sequenced monocot genomes thus far) but after the monocot-eudicot split. After this shared ancestral tetraploidy, grass lineages experienced another two tetracentroids. Oil palm experienced one tetracentropy, likely shared with other palms (Singh et al., 2013). Banana experienced three additional duplications (D’Hont et al., 2012), making it among the more duplicated monocot genomes known, a paleo-dotriacontaploid (2n=32x) like maize (Zea mays; Schnable et al., 2009).

Our results clearly showed that τ is shared by all 11 sequenced monocot genomes, i.e., all commeliniids. Commeliniids originated $\sim 120$ to $\sim 100$ MYA, not long after the origin of monocots $\sim 150$ to $\sim 130$ MYA (Hedges et al., 2006; The Angiosperm Phylogeny Group, 2009), and comprise $\sim 18,800$ Old World and New World monocot species that include many economically important plants, such as palm, banana, ginger, cereals, and grasses. The estimated timing of τ is very close to the monocot origin, making it analogous to the γ event foreshadowing the initial radiation of core eudicots (Figure 4A). Additional basal monocot genome sequences are needed to resolve the exact timing of τ.

**Integrating Synteny and Phylogenomic Approaches for Systematic Genome Comparison**

Repeated polyploidization and subsequent diploidization through fractionation and structural rearrangement are genome-wide mutational forces and a striking feature in plants, presumably contributing to their marvelous genetic, phenotypic, and ecological
diversity (Bowers et al., 2003; Blanc and Wolfe, 2004; Adams and Wendel, 2005; Freeling and Thomas, 2006; Doyle et al., 2008; Fawcett et al., 2009; Freeling, 2009; Schnable et al., 2009; Solis et al., 2009; Van de Peer et al., 2009; Paterson et al., 2010; D’Hont et al., 2012; Ming et al., 2013). Nonetheless, conserved genetic content and order are observed to different degrees in closely and distantly related plant genomes (Bonierbale et al., 1988; Kowalski et al., 1994; Paterson et al., 1996; Ku et al., 2000; Tang et al., 2008a; Tang et al., 2010). However, plant genome comparisons differ qualitatively from those that are suitable for mammals and most other taxa and require methods that reflect and represent the multiway homoeologous relationships resulting from paleopolyploidies.

Early studies of WGD using whole-genome sequences were phylogenetic, based either on synonymous substitution rate ($K_s$) distribution between duplicated gene pairs (Lynch and Conery, 2000; Blanc and Wolfe, 2004) or topology of gene family trees (Bowers et al., 2003). Age grouping of paralog $K_s$ values can be calculated without information about gene position and is the only method for dating WGDs using transcriptome data alone but risks blending of gene families with different modes of $K_s$ divergence and for ancient WGDs base substitution often approaches saturation and sequence-based methods become less useful. An early phylogenetic approach for dating WGD based on discriminating duplication-first (external) and speciation-first (internal) tree topologies (Bowers et al., 2003) was constrained by limited EST sampling, data for few taxa, and lineage evolutionary rate variation (Tang et al., 2008b), although later improved with more rigorous taxon sampling in the phylogenomic dating approach (Jiao et al., 2011).

While the phylogenomic approach utilizes temporal signals, the synteny approach identifies WGDs from spatial signals, i.e., conservation of gene position and order along the chromosomes (Vandepoele et al., 2002; Haas et al., 2004; Tang et al., 2008a). Synteny conservation provides independent and reliable phylogenetic signals unaffected by DNA substitution rate variation (Tang et al., 2008a). In synteny-based WGD dating, coalescence of duplicated regions can be traced before or after speciation depending on the patterns of one-to-one correspondence versus one-to-multiple (or multiple-to-multiple) correspondence (Figure 1B).

Integration of temporal and spatial evidence (Figure 1) is by far the most accurate approach for the circumscription of WGD events. It also naturally connects studies of genome structure and molecular evolution. With the fast growth of genomic data, systematic comparisons of divergent taxa aided by WGD-aware ancestral reconstruction promises better delineation and interpretation of plant genome evolution by providing an effective framework for multiway genomic alignments, with the unique
advantage of tolerating long evolutionary distance and extensive genome rearrangement.

The Value of Ancestral Gene Order in Genome Comparisons

Inferred or approximate ancestral gene order often provides the best reference to thread genomic alignments in taxa with WGDs (Bowers et al., 2003; Kellis et al., 2004; Zheng et al., 2013). First, it compensates for post-WGD gene loss, increasing the proportion of aligned genes among homoeologous regions. Second, it expands synteny blocks as lineage specific breakpoints are removed. Therefore, ancestral reconstruction reverses the effects of more recent WGDs and better reveals the interleaving pattern of gene loss (as illustrated in Figure 1; Kellis et al., 2004), greatly facilitating recovery of full syntenic mapping among homoeologous regions. When a WGD-free outgroup genome is not available,
as is the case in many plant clades, ancestral gene order can be approximated by in silico reconstruction. Reconstructed ancestral genomes, while not necessarily identical to the true ancestral genome, are important references in effective genome alignments. Recurring polyploidization and diploidization, ubiquitous in plant evolution, greatly obscure the network of homology in plant genome comparisons. More than 20 ancestral and independent paleopolyploidies have been identified in ~50 sequenced angiosperms, affecting 100% of lineages. Consequently, many angiosperm genomes contain populations of homoeologous regions of different depth resulting from combination of different levels of polyploidization and postpolyploidy gene loss (examples from the six studied genomes are shown in Supplemental Figure 5). Inter-genomic comparisons are complicated as well. Arabidopsis thaliana and rice, for example, have 829 synteny blocks averaging 43 genes, resulting from six WGDs, three in each lineage. A comparison between diploid Brassica rapa and banana genomes would involve as many as 52 copies involving eight WGD events (3x2x2x3=36 in Brassica and 2x2x2x2=16 in banana). It is beyond current technology to directly find and align all the homoeologous regions in such genomes. Instead, ancestral reconstruction based on an established framework of historical WGDs is able to effectively recover the grand hierarchy of homoeology.

Variation of Lineage Nucleotide Evolutionary Rates and Estimated Ages of \( \sigma \) and \( \tau \)

Nucleotide evolutionary rates can vary greatly among sites, gene families, nuclear and organellar genomes (Wolfe et al., 1987; Zhang et al., 2002; Mower et al., 2007; Gaut et al., 2011), and lineages with different life history traits and population characters (Smith and Donoghue, 2008; Gaut et al., 2011). Shared paleopolyploidy events provide inherent reference points to calibrate evolutionary rates among affected lineages. Homoeologs bearing synteny information are also intrinsically more accurate phylogenetic markers for multicopy gene families. By applying such reasoning, the Vitis lineage nucleotide substitution rate is estimated to be less than half that of Arabidopsis (Tang et al., 2008b), and the Nelumbo lineage rate is 30% slower than Vitis (Ming et al., 2013). The synonymous site substitution rate between paralogs formed in the shared \( \tau \) WGD (Supplemental Figure 6) is \(-1.7\) times larger in rice than oil palm. This is much less than the \(-5\) fold difference between grasses (faster) and palms estimated from chloroplast ribulose-bisphosphate carboxylase large subunit (rbcL) genes (Gaut et al., 2011) and \(-4\) fold difference estimated from a combination of rDNA, chloroplast, and mitochondrial genes (Smith and Donoghue, 2008). These differences in rate estimates emphasize the heterogeneous nature of molecular evolution in plant genomes. Reliable WGD and homoeology identification is clearly essential for detailed evolutionary analysis on a genomewide scale.

The pan-grass \( p \) event was estimated to be \(-70\) million years old (Paterson et al., 2004) based on descendant paralogs in rice having a Ks distribution of mode \(-0.36\) and estimated rice lineage rate of 6.5e-9 substitutions per synonymous site per year. Using the same rate, the age of \( \sigma \) (paralogous Ks mode \(-1.65\)) was estimated to be \(-75\) MYA. The palm WGD \( p \) event was estimated to be \(-75\) MYA (D’Hont et al., 2012; Singh et al., 2013), with descendant paralogs in oil palm having a Ks distribution of mode \(-0.36\), giving an estimated oil palm lineage rate of \(-2.4e-9\). Paralogs of \( \tau \) have Ks distributions of mode \(-1.13\) in oil palm and \(-1.87\) in rice. Taking the average rate of rice and oil palm lineages (4.45e-9) to be the approximate substitution rate on their MRCA lineage, the age of \( \tau \) can be estimated to be \(-73\) MYA before the Arecaceae-Poaceae split using Ks distribution of oil palm paralogs or \(-64\) MYA using rice Ks distribution. Again, we note that recent fossil discoveries (Prasad et al., 2005) have a host of implications for estimated evolutionary rates and associated dates of these and many other key events in monocot evolution, a topic that we are exploring comprehensively under separate cover. Taking into consideration the uncertainties in molecular evolutionary rate estimation, and availability of paleontological records, \( \tau \) likely occurred in primitive monocot branches. Due to a lack of fully sequenced early-diverging monocot genomes and limitation of resolution in current sequence-based dating methods, further refinement of the exact timing of \( \sigma \) and \( \tau \) awaits future studies.

While disambiguation of \( p = \sigma = \tau \) was manual, this approach to integrating synten and phylogenomic comparisons and demonstration of its value to resolve otherwise cryptic paleopolyploidy events has provided results useful to validate efforts to automate the approach, another topic that we plan to report on comprehensively elsewhere.

METHODS

Data Source

We selected 15 taxa to represent all of the major land plant lineages for which genome sequence data are available, including four rosid genomes (Arabidopsis thaliana, Populus trichocarpa, Theobroma cacao, and Vitis vinifera), two asterids (Solanium lycopersicum and Solanum tuberosum), one basal eudicot (Nelumbo nucifera), six monocots (Oryza sativa, Brachypodium distachyon, Sorghum bicolor, Elaeis guineensis, Phoenix dactylifera, and Musa acuminata), one lycophyte (Selaginella moellendorffii), and one moss (Physcomitrella patens). Genome data were downloaded from either Phytozone (version 6) or respective project websites.

Global Gene Family Phylogeny

The OrthoMCL method (Li et al., 2003) was used to construct a complete set of protein coding genes into orthogroups. For each ortholog, amino acid alignments were generated with MAFFT (Katoh et al., 2005) using default parameters. Corresponding DNA sequences were then forced onto the amino acid alignments using custom Perl scripts. DNA alignments were trimmed using trimAl (Capella-Gutiérrez et al., 2009) using the heuristic “automated1” method. Then, maximum likelihood phylogenetic trees were constructed using RAxML version 7.2.1 (Stamatakis, 2006) with the GTR gamma model and 100 bootstrap replicates. In total, we constructed 11,503 phylogenetic trees containing monocot and at least one outgroup genes. Due to the classification stringency of OrthoMCL, two anchor genes on synteny blocks could be classified into different orthogroups, and these two orthogroups were combined to run phylogeny for dating the duplication of anchors.

Phylogenomic Dating of Synteny Blocks

Duplication time of synteny blocks was dated by carefully mapping homologous anchor genes onto phylogenetic trees. We considered gene duplications only when the BS value equal or greater than 50%. If
homologous anchor genes are clustered in more than one orthogroup, we combined them and recomputed the phylogenetic trees for dating. After all anchor genes duplication dated, we scored the synteny blocks duplication by following criteria: (1) at least two anchor genes for each region could be dated phylogenetically with BS ≥50%, and (2) the timing of the anchor gene duplications must be mostly consistent.

**Circumscribing Paleopolyploidy Events in Multiple Lineages Based on Synteny Patterns**

For each pair of the six input genomes (Supplemental Table 1), representative coding sequences from each gene locus were aligned by LASTZ (Harris, 2007) with default parameters. Weak matches with C-score (Putnam et al., 2008) <0.5 (indicating that their similarity falls below 50% of the best matched pair between the two genomes) were filtered out. Remaining matches within 30 to ~40 Manhattan distance units (40 for grape-oil palm, grape-rice, grape-sorghum, and rice-sorghum comparisons and 30 for the remaining comparisons) were clustered into the same synteny block, with blocks having at least five anchors retained. Chromosomes (or scaffolds) in each genome were segmented at major alignment breakpoints, resulting in smaller genomic regions with simple, continuous synteny patterns. Segmented regions between each pair of genomes were hierarchically clustered based on similarity of homolog density and filtered based on their statistical significance (P < 1e-30) by modeling the probability of homologous matches with Poisson distribution as described previously. Each set of homologous region is defined as a PAR (Tang et al., 2010). We produced a total of 1217 high quality PARs across the studied genomes in this study.

Homologous regions inside each PAR were further refined at their boundaries to trim off unaligned proportions and then analyzed to infer genomic multiplicity using an evolutionary model based on pairwise synteny distance between the clustered regions. Synteny distance is calculated between pairs of aligned syntentic regions similar to the p-distance in nucleotide alignment (gene order alignment in our case), which was then used to infer their phylogenetic relationship using the UPGMA method. From the resulting tree topology, the copy number of paralogous segments in each genome was counted and used to deduce the multiplicity of genomic regions in each individual lineage since their divergence. The pairwise differences of WGD multiplicity were then combined to infer the number of WGDs on each branch of the given species phylogeny by applying the Fitch-Margoliash algorithm (Fitch and Margoliash, 1967). The Fitch-Margoliash algorithm applies a least-squares method to estimate per branch length (in our case, equivalent to the genome multiplicity) based on all pairwise distances between the genomes in comparison. The procedure was pipelined by an in-house Python script with manual inspections at various intermediate stages.

**Inferring Ancestral Gene Order Prior to Polyploidy Event**

In WGDs, ancestral regions were duplicated and then diverged by sequence and structural mutations. Each descendant homoeologous region contains a fraction of mutated ancestral gene content. Once synteny blocks are identified, the ancestral content can be reconstructed by coalescence of all the descendant regions contained in them. While ancestral reconstruction can be performed at different resolutions to facilitate comparing regions with different extents of divergence, reconstruction of ancestral gene order recovers the alignability and syntenic mapping of homoeologous regions across highly diverged genomes, like those under this study. In our approach, first, multiple alignment of homoeologous regions was computed by the top-down algorithm implemented in MCscan (Tang et al., 2008b). Local alignment inconsistencies, mostly resulted from tandem gene duplications and micro-inversions, were linearized by dynamic programming maximizing anchor gene matching. The aligned homoeologous regions were then merged into a single preduplication segment by interpolation, following methods in previous studies (Bowers et al., 2003; Aury et al., 2006). This results in approximated ancestral gene order and syntenic mappings.

**Synonymous Substitution (Ks) Calculation**

For each pair of homologous genes, we aligned their protein sequences using ClustalW2 (Larkin et al., 2007) and converted the protein alignment to DNA alignment using PAL2NAL (Suyama et al., 2006). Some homologous genes could not produce reliable alignment for various reasons and were discarded from further analysis. Ks values were calculated using the Nei-Gojobori algorithm (Nei and Gojobori, 1986) implemented in the PAML package (Yang, 1997). Ks values for gene pairs with average GC3 > 75% are considered unreliable and discarded (Tang et al., 2010). Ks values >3.0 indicate saturated substitutions at synonymous sites and those gene pairs were excluded from further analysis.

**Supplemental Data**

The following materials are available in the online version of this article.

**Supplemental Figure 1.** Exemplar Maximum Likelihood Phylogeny of WRKY Gene Family

**Supplemental Figure 2.** Exemplar Maximum Likelihood Phylogeny of Zinc Finger Gene Family.

**Supplemental Figure 3.** Example Alignment Showing Eight Homologous Musa Regions Duplicated after the Divergence of Musa and Oryza.

**Supplemental Figure 4.** WGD Multiplicity Level Estimation on Each Branch of Given Species Phylogeny.

**Supplemental Figure 5.** Homoeolog Depth (Paleopolyploidy Level) of Genomic Regions in the Studied Taxa.

**Supplemental Figure 6.** Distributions of Synonymous Substitution Rates (Ks) among Groups of Orthologous or Paralogous Gene Pairs.

**Supplemental Table 1.** Sources and Basic Information about Angiosperm Genomes Used in Synteny Comparison.

**Supplemental Data Set 1.** Significant Enrichment of GO-SLIM Term for the Gene Families with Ancient Tau Duplication Measured by Hypergeometric Test Followed by FDR Multitest Adjustment Method.

**ACKNOWLEDGMENTS**

We appreciate funding from the National Science Foundation (DBI 0849896, MCB 0821096, and MCB 1021718).

**AUTHOR CONTRIBUTIONS**

Y.J., J.L., H.T., and A.H.P. designed the research. Y.J. and J.L. performed analyses. H.T. contributed analytic tools. Y.J., J.L., H.T., and A.H.P. wrote the article.

Received May 13, 2014; revised July 3, 2014; accepted July 10, 2014; published July 31, 2014.

**REFERENCES**


