RESEARCH ARTICLE

Large-Scale Gene Relocations following an Ancient Genome Triplication Associated with the Diversification of Core Eudicots

Yupeng Wang[1], Stephen P. Ficklin[2], Xiyin Wang[1], F. Alex Feltus[3], Andrew H. Paterson[1]*

1 Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia, United States of America, **2** Department of Horticulture, Washington State University, Pullman, Washington, United States of America, **3** Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina, United States of America

* paterson@uga.edu

OPEN ACCESS

Citation: Wang Y, Ficklin SP, Wang X, Feltus FA, Paterson AH (2016) Large-Scale Gene Relocations following an Ancient Genome Triplication Associated with the Diversification of Core Eudicots. PLoS ONE 11(5): e0155637. doi:10.1371/journal.pone.0155637

Editor: Lewis Lukens, University of Guelph, CANADA

Received: June 22, 2015

Accepted: May 2, 2016

Published: May 19, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: AHP appreciates funding from the National Science Foundation (NSF: DBI 0849896, MCB 0821096, MCB 1021718). This study was supported in part by resources and technical expertise from Georgia Advanced Computing Resource Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Different modes of gene duplication including whole-genome duplication (WGD), and tandem, proximal and dispersed duplications are widespread in angiosperm genomes. Small-scale, stochastic gene relocations and transposed gene duplications are widely accepted to be the primary mechanisms for the creation of dispersed duplicates. However, here we show that most surviving ancient dispersed duplicates in core eudicots originated from large-scale gene relocations within a narrow window of time following a genome triplication (γ) event that occurred in the stem lineage of core eudicots. We name these surviving ancient dispersed duplicates as relocated γ duplicates. In *Arabidopsis thaliana*, relocated γ, WGD and single-gene duplicates have distinct features with regard to gene functions, essentiality, and protein interactions. Relative to γ duplicates, relocated γ duplicates have higher non-synonymous substitution rates, but comparable levels of expression and regulation divergence. Thus, relocated γ duplicates should be distinguished from WGD and single-gene duplicates for evolutionary investigations. Our results suggest large-scale gene relocations following the γ event were associated with the diversification of core eudicots.

Introduction

Gene duplication by different mechanisms is a primary raw material for the origin and evolution of new genes, as well as generating functional novelty and specialization [1]. The angiosperms (flowering plants) are an outstanding model in which to investigate the modes and consequences of gene duplication. Large-scale gene duplications such as whole-genome duplications (WGDs) have been recurring in angiosperm evolution [2–5]. *Arabidopsis*, a model angiosperm, experienced two WGDs since its divergence from other members of the Brassicales clade (αand β), and a more ancient triplication (γ) shared with all core eudicots [3, 6, 7]. Single-gene duplications in angiosperms are also widespread [8–10].

Single-gene duplications have been subdivided into local and dispersed duplications [8]. Local duplications may occur by tandem duplication (consecutive in the genome and presumed to arise through unequal crossing over) [8] and proximal duplication (near one another but separated by a few genes, thought to occur by localized transposon activities) [11–13]. Dispersed duplicates are neither adjacent to each other in the genome nor within homologous chromosomal segments [11, 12, 14]. Transposon-induced single-gene transpositions or transposed gene duplications, via either DNA or RNA-based mechanisms, have been suggested to account for the widespread existence of dispersed duplicates [8, 15–19].

Following WGDs, genomes have preferentially retained genes encoding transcription factors [20, 21]. Moreover, gene fates are often correlated across multiple WGD events [22]. For example, in *Arabidopsis*, γ duplicates are more often retained in duplicate for both β and α events [22]. In contrast to WGDs, genomes have preferentially retained local duplications of genes encoding membrane proteins or functioning in stress response [23]. In *Populus trichocarpa*, whole-genome and local duplicates were found to differ in gene functions, protein lengths and expression patterns [24]. In addition, the age distribution of the duplicates derived from a large-scale event exhibits a peak in ancient times due to dramatic increase in the number of duplicated genes, while single-gene duplicates show an *L* shaped age distribution due to a high and constant rate of gene loss [25, 26].

However, WGD and local duplicates comprise only part of the whole set of duplicated genes. A substantial part of the duplicated gene population is dispersed duplicates [8, 9]. To date, only a few things are known regarding the preservation and evolution of dispersed duplicates. Recent dispersed duplicates, often referred to as transposed duplicates, are often associated with flanking repeats, and certain classes of genes that tend to form local duplications are more likely to have transposed than other gene classes [27]. Certain *Arabidopsis* gene families that are prone to mobility have transposed in different epochs throughout the rosids [10].

WGD events have long been thought to play a major role in plant speciation [28, 29]. Recently, bioinformatics analysis of 41 plant genomes suggested that polyploidy extensively occurred around the Cretaceous–Paleogene (K–Pg) extinction event about 66 million years ago (Mya) [30]. However, following the γ genome triplication around 125 Mya, core eudicots diverged in a narrow window of time during which few WGD events occurred [6, 7, 31]. Thus, WGD itself may not explain the rapid diversification of core eudicots. In this work, we study the evolutionary origins of dispersed duplicates and explore relationships between dispersed duplicates and the diversity of core eudicots.

Results

A substantial proportion of the survivalyduplicates were relocated in core eudicots

We identified WGD, local and dispersed duplicates (see Methods) in 17 sequenced core eudicot genomes, including *Ricinus communis*, *Populus trichocarpa*, *Lotus japonicus*, *Medicago truncatula*, *Glycine max*, *Cajanus cajan*, *Cucumis sativus*, *Malus x domestica*, *Fragaria vesca*, *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Brassica rapa*, *Carica papaya*, *Theobroma cacao*, *Vitis vinifera*, *Solanum tuberosum* and *Solanum lycopersicum*. Synonymous substitution rates (Ks) were initially used to estimate the relative ages of gene duplications. Large-scale gene duplication events such as WGDs often result in Ks peaks because of the short periods during which large numbers of duplicated genes were created [25]. In contrast, single-gene duplications such as local and transposed duplications typically show Ks distributions of *L* shape, because gene duplications and subsequent deletions are random and at relatively constant rates during evolution [26]. We compared Ks distributions among WGD, local and dispersed duplicates in these 17 core eudicot genomes (Fig 1). In all these genomes, the Ks distributions of local duplicates show an *L* shape, and the Ks distributions of WGD duplicates exhibit peaks depending

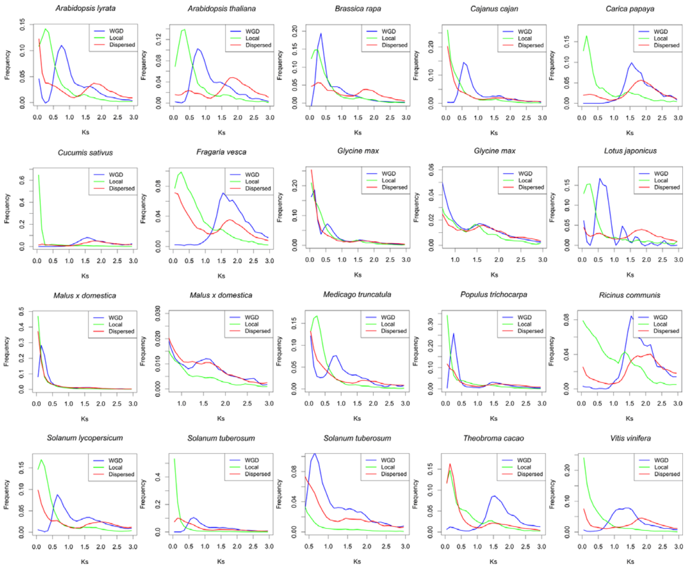

Fig 1. Comparison of Ks distributions among WGD, local and dispersed duplicates in the investigated core eudicot genomes. For *Glycine max*, *Solanum tuberosum* and *Malus x domestica* whose genomes are large and recent duplicates are rampant, a second plot is provided to show the right tails of Ks distributions.

doi:10.1371/journal.pone.0155637.g001

on the epochs of WGD events. However, in most investigated core eudicot genomes, dispersed duplicates show a secondary Ks peak (in the few others whose genome are big and recent duplications are rampant they show a thicker Ks distribution than local duplicates) at relatively large Ks levels (1.5<Ks<2.2). Dispersed duplicates, if all of them were created by single-gene events, should show an L-shaped Ks distribution. The secondary Ks peaks/thickness at high Ks levels for dispersed duplicates in the investigated core eudicot genomes suggest that the ages of dispersed duplicates are a mixture of two distributions–"*L* shaped + a secondary peak". We hypothesized that in a typical core eudicot genome, part of the dispersed duplicates originated from single-gene duplication events that lead to the *L* shaped distribution, while there was also a burst of dispersed duplicates created by ancient large-scale gene duplication events. Since the levels of the secondary Ks peaks are very similar among the 17 investigated core eudicot genomes, the burst of dispersed duplicates was likely to begin in the common ancestor of core eudicots. It is known that a genome triplication event (γ) occurred in the stem lineage of core eudicots [7]. It is highly possible that most of the ancient dispersed duplicates in these investigated core eudicot genomes originated from the γ event.

To examine whether there was a burst of dispersed duplication in the early evolution of core eudicots, we investigated the detailed origins of duplicated genes in *Arabidopsis thaliana*, since the genome of *Arabidopsis thaliana* has been extensively studied. Using the procedure described in Methods, we classified *Arabidopsis thaliana* duplicated genes into 5156α, 2142β, 802γ, 3602 tandem, 1197 proximal, and 9345 dispersed duplicates. We used gene colinearity conservation between *Arabidopsis thaliana* and outgroups (phylogeny shown in Fig 2A) to estimate the epochs (ages) of *Arabidopsis thaliana* dispersed duplicates [32, 33], i.e. to examine between which speciation events each dispersed duplicate was created. For example, if an *Arabidopsis thaliana* duplicated gene shows colinearity conservation with *Populus trichocarpa*, but not with *Vitis vinifera* or more distant outgroups, this duplicated gene is estimated to be created during the epoch between *Arabidopsis-Populus* and *Arabidopsis-Vitis* divergence. Using such logic, the ages of *Arabidopsis thaliana* dispersed duplicates were assigned to eight epochs, including after *Arabidopsis thaliana-Arabidopsis lyrata* divergence (<5 Mya), between

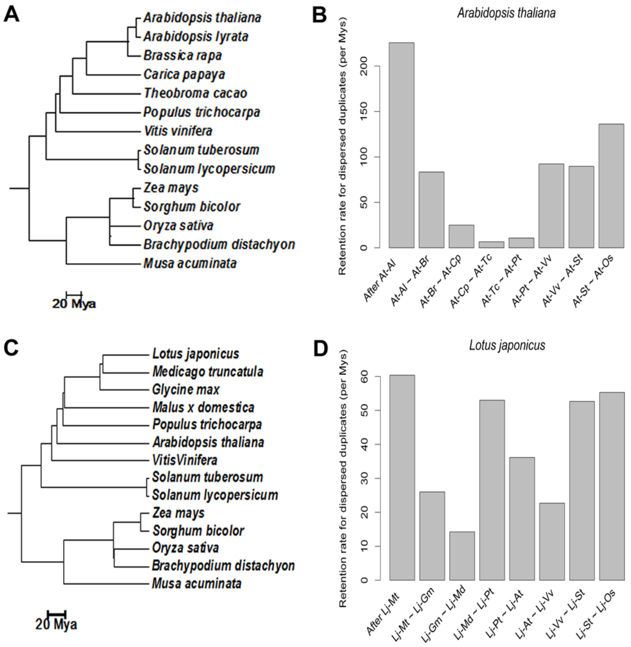

Fig 2. Different retention rates of dispersed duplicates during the evolution of core eudicots. (A) Phylogenetic relationships between *Arabidopsis thaliana* and its outgroups used for estimating the epochs (ages) of dispersed duplicates according to colinearity conservation. (B) Retention rates of *Arabidopsis thaliana* dispersed duplicates in different epochs. Abbreviations: At: *Arabidopsis thaliana*; Al: *Arabidopsis lyrata*; Br: *Brassica rapa*; Cp: *Carica papaya*; Tc: *Theobroma cacao*; Pt: *Populus trichocarpa*; Vv: *Vitis vinifera*; St: *Solanum tuberosum*; Os: *Oryza sativa*. Dispersed duplicates created after At-Br divergence were named transposed duplicates, while those created between At-Pt and At-St divergence were named relocated γ duplicates. (C) Phylogenetic relationships between *Lotus japonicus* and its outgroups used for estimating the epochs (ages) of dispersed duplicates according to colinearity conservation. (D) Retention rates of *Lotus japonicus* dispersed duplicates in different epochs. Abbreviations: Lj: *Lotus japonicas*; Mt: *Medicago truncatula*; Gm: *Glycine max*; Md: *Malus x domestica*; Pt: *Populus trichocarpa*; At: *Arabidopsis thaliana*; Vv: *Vitis vinifera*; St: *Solanum tuberosum*; Os: *Oryza sativa*.

doi:10.1371/journal.pone.0155637.g002

Arabidopsis thaliana-Arabidopsis lyrata and *Arabidopsis-Brassica* divergence (5~16 Mya), between *Arabidopsis-Brassica* and *Arabidopsis-Carica* divergence (16~72 Mya), between *Arabidopsis-Carica* and *Arabidopsis-Theobroma* divergence (72~90 Mya), between *Arabidopsis-Theobroma* and *Arabidopsis-Populus* divergence (90~107 Mya), between *Arabidopsis-Populus* and *Arabidopsis-Vitis* divergence (107~113 Mya), between *Arabidopsis-Vitis* and *Arabidopsis-Solanum* divergence (113~125 Mya), and between *Arabidopsis-Solanum* and *Arabidopsis-Oryza* divergence (125~148 Mya). For each epoch, we computed a retention rate for dispersed duplicates, which was the number of retained dispersed duplicates divided by the epoch length (in millions of years). We compared retention rates of dispersed duplicates across different epochs (Fig 2B), and found that there was indeed a burst of dispersed duplicates between *Arabidopsis-Populus* and *Arabidopsis-Oryza* divergence (107~148 Mya), i.e., in the early evolutionary period of core eudicots. To test whether this observation is unique to *Arabidopsis thaliana*, we also made plots for *Lotus japonicus*, which also showed a high retention rate for dispersed duplicates between *Lotus-Vitis* and *Lotus-Oryza* divergence (113~148 Mya, Fig 2C and 2D). Further, we observed very similar patterns of retention for the dispersed duplicates in *Arabidopsis lyrata* and *Medicago truncatula* (S1 and S2 Figs), based on the above phylogenies (Fig 2A and 2C). These observations support the hypothesis that dispersed gene duplications in the initial tens of millions of years following the γ event were much more frequent than before and thereafter in core eudicots. Since single-gene duplications tend to have steady deletion rates, the most plausible explanation of this burst of gene dispersal is that extensive chromosome rearrangements following the γ event relocated most γ duplicates to new positions.

We named the 4763 *Arabidopsis thaliana* dispersed duplicates generated between *Arabidopsis-Populus* and *Arabidopsis-Solanum* divergence (two epochs) as 'relocated γ duplicates', and reserve "γ duplicates" to denote genes remaining at the colinear positions within γ colinear blocks. Because the γ event occurred prior to *Arabidopsis-Solanum* divergence, among the dispersed duplicates generated between *Arabidopsis-Solanum* and *Arabidopsis-Oryza* divergence there might be a possibility that some were created prior to the γ event. Thus, we excluded the dispersed duplicates generated during this epoch from relocated γ duplicates. The evolutionary origins of all duplicated genes in *Arabidopsis thaliana* are listed in S1 Table. Relocated γ duplicates tend to have smaller Ks than γ duplicates (*P* = 0.047, *t*-test), precluding the possibility that most of the relocated γ duplicates were created by the ancestral duplication events of seed plants and angiosperms [5]. Additionally, we named the 2047 *Arabidopsis thaliana* dispersed duplicates created after *Arabidopsis-Brassica* divergence as "transposed duplicates". The number of γ duplicates in *Arabidopsis thaliana* is only 802. Thus, a large proportion (85.6%) of the duplicates created by the γ event appear to have been relocated in *Arabidopsis*. It might be possible that recent WGD events have erased most signals of the γ event in *Arabidopsis*. In the lineages of *Carica Papaya* and *Theobroma cacao* whose genomes have not experienced recent WGD events, we estimated 80.2% and 47.8% of the γ duplicates were relocated respectively (see Methods). Thus, it is reasonable to conclude that a substantial proportion of the retained γ duplicates were relocated in core eudicots.

Relocated γ duplicates are functionally distinct from WGD and single-gene duplicates in *Arabidopsis thaliana*

We investigated the functional features of relocated γ duplicates in *Arabidopsis thaliana* by comparison withα, β, γ, tandem, proximal, transposed and relocated γ duplicates. We used the Gene Ontology (GO) terms to represent the biological functions of *Arabidopsis thaliana* genes. As described in Methods, the functional profile of each origin of duplicates was denoted by its fold enrichment for the 29 biggest "biological process" GO terms (see details in S2 Table). We clustered the functional profiles of different origins of duplicated genes using a heat map (Fig 3A).

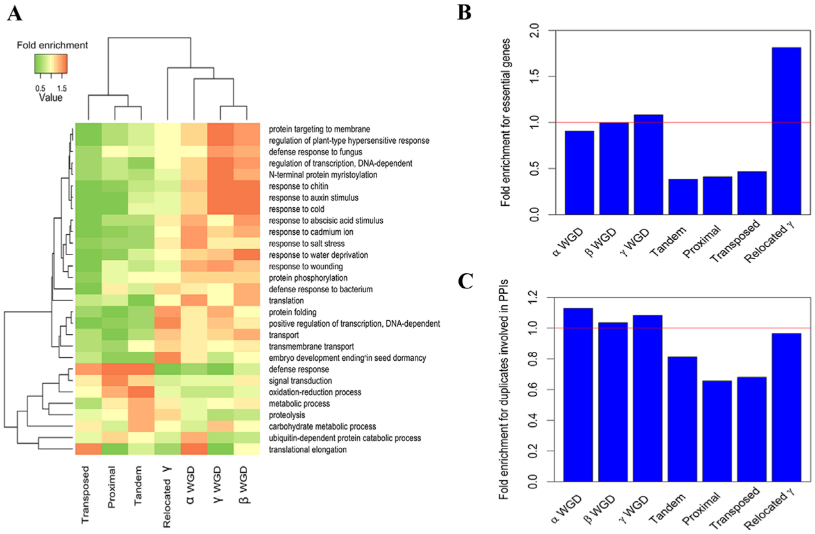

Fig 3. Functional comparison of different origins of duplicates. (A) Clustering of functional profiles. (B) Comparison of fold enrichment for essential genes. (C) Comparison of fold enrichment for the genes involved in protein-protein interactions (PPIs).

doi:10.1371/journal.pone.0155637.g003

Relocated γ duplicates were clustered with WGD duplicates at the second level of the dendrogram, while single-gene duplicates formed another cluster at the second level. Thus, relocated γ duplicates slightly resemble WGD duplicates, but are more different from single-gene duplicates in terms of biological functions.

Essential genes cause loss-of-function mutant phenotypes in single-gene knock-out experiments. We analyzed 2400 genes with a loss-of-function mutant phenotype in *Arabidopsis thaliana* [34]. Fold enrichment of essential genes in different origins of duplicates was compared (Fig 3B). Relocated γ duplicates are significantly enriched for essential genes ($P<2.2\times10^{-16}$), α duplicates are only slightly enriched for essential genes ($P = 0.04$), β and γ duplicates are not distinguishable from random probabilities of being essential genes ($P>0.1$), while single-gene duplicates are depleted for essential genes ($P<3.0\times10^{-9}$). This observation suggests that relocated γ duplicates tend to perform important biological functions, which are not likely to be compensated by other genes.

The gene balance hypothesis suggests that any surviving genome has retained an optimal balance of gene products that bind with one another to form protein complexes, or are involved in multiple steps of biological pathways [35]. In terms of network connectivity, more connected gene products should be more essential as phenotype is more likely to change if dosage imbalance happens [8]. According to this hypothesis, WGD duplicates should be more frequently involved in protein-protein interactions (PPIs) than single-gene duplications. Using a large-scale PPI data [36], we compared the enrichment of duplicates involved in PPIs among different origins of duplicates (Fig 3C), which shows that α duplicates are more likely to be involved in PPIs ($P = 0.64\times10^{-4}$); single-gene duplicates including tandem, proximal and transposed duplicates are less likely to be involved in PPIs ($P<0.005$); and β, γ and relocated γ duplicates have close-to-average probabilities of involvement in PPIs ($P>0.1$). In partial summary, the comparisons of biological functions, essentiality and PPIs among different origins of duplicated genes in *Arabidopsis thaliana* suggest that in terms of functional landscape, relocated γ duplicates are distinct from WGD and single-gene duplicates.

Evolutionary significance of gene relocations following the γ event in *Arabidopsis*

We investigated the evolutionary significance of gene relocations following the γ event in *Arabidopsis thaliana*. We first examined if gene relocations following the γ event increased coding sequence divergence. We used non-synonymous substitution rates (Ka) and Ka/Ks to indicate coding sequence divergence and selection pressure respectively. Comparison of Ka among γ, relocated γ and transposed duplicates shows that relocated γ duplicates tend to code more divergent proteins than γ duplicates (Fig 4A). Comparison of Ka/Ks shows that relocated γ duplicates are under more relaxed purifying selection than γ duplicates (Fig 4B), suggesting that relocated γ duplicates are less affected by the gene balance constraints which often impose strong purifying selection on duplicated genes within tens of millions of years following WGDs [8]. However, higher Ka/Ks values for transposed duplicates may be explained by the theory that most transposed duplicates are accumulating degenerative mutations and will be ultimately pseudogenized, although admittedly there is a small chance of transposed duplicates being preserved for long times due to relaxed selective constraints and neofunctionalization [33]. Further, we investigated whether gene relocations following the γ event increased gene expression and regulation divergence. Expression divergence between γ, relocated γ and transposed duplicates is plotted in Fig 4C, showing that expression divergence is in general comparable among these three categories of duplicates ($P>0.1$, *t*-test). Furthermore, gene regulation divergence (denoted by the difference of transcription factor binding sites in their promoter

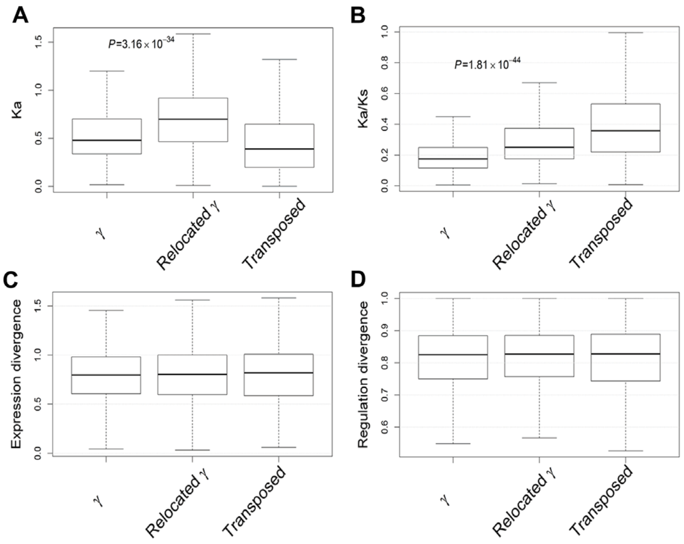

Fig 4. Evolutionary significance of relocated γ duplicates. (A) Comparison of Ks among γ, relocated γ and transposed duplicates. (B) Comparison of Ka/Ks among γ, relocated γ and transposed duplicates. (C) Comparison of expression divergence among γ, relocated γ and transposed duplicates. (D) Comparison of gene regulation divergence among γ, relocated γ and transposed duplicates.

doi:10.1371/journal.pone.0155637.g004

regions) between γ, relocated γ and transposed duplicates is plotted in Fig 4D, showing that gene regulation divergence is also comparable among the three categories of duplicates ($P>0.1$, *t*-test). These observations suggest that the primary evolutionary impact of gene relocations following the γ event was to make duplicated genes less affected by purifying selection, perhaps rendering them freer to acquire amino acid changes and to evolve novel and critical biological functions.

Discussion

The γ event was suggested to have occurred at the base of the core eudicots. Following the γ event, core eudicot lineages rapidly radiated, giving rise to nearly 75% of angiosperm species [6]. The large-scale gene relocations following the γ event were thus associated with the diversification of core eudicots. We observed Ks peaks around 2.0 for the dispersed duplicates in the investigated core eudicots. Due to extensive lineage-specific genome and gene duplications which are meanwhile mixed with extensive gene losses, the original pairs of ancient dispersed duplicates are largely lost in the extant genomes. However, future studies can investigate pairs of ancient dispersed duplicates which have been preserved in all core eudicot genomes and their contributions to the survival of core eudicots.

Large-scale, extensive gene relocations could provide a huge number of recombinations of genetic materials for the genomes to rapidly evolve. The number of relocated γ duplicates (4763) is almost 6 times the number of γ duplicates (802) in *Arabidopsis thaliana*. In contrast, the number of transposed duplicates (2047) is significantly smaller than that of α duplicates (5156). Thus, in *Arabidopsis thaliana*, most of the duplicates originating from the γ event were relocated shortly, while a small proportion of the duplicates originating from the α event have been relocated. In *Carica* and *Theobroma*, >47% of the γ duplicates were found to have relocated. The relocated γ duplicates we identified in *Arabidopsis thaliana* all show colinearity with *Populus trichocarpa* and/or *Vitis vinifera*, suggesting that they did not relocate again within the recent 100 million years. Thus, large-scale, extensive gene relocations may have been a singular feature of the γ event. In addition to fractionation, small-scale and mostly deleterious gene relocations are common across different WGD events [8]. A possible distinct mechanism for the gene relocations following the γ event could be that the triplicated chromosomes were fragile and underwent many rearrangements, which were then gradually fused into bigger chromosomes [37].

Methods

Genome annotations

Gene locations, coding sequences and colinear gene pairs for eudicot genomes including *Ricinus communis*, *Populus trichocarpa*, *Lotus japonicus*, *Medicago truncatula*, *Glycine max*, *Cajanus cajan*, *Cucumis sativus*, *Malus x domestica*, *Fragaria vesca*, *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Brassica rapa*, *Carica papaya*, *Theobroma cacao*, *Vitis vinifera*, *Solanum tuberosum*, *Solanum lycopersicum*, and outgroup monocot genomes including *Oryza sativa*, *Brachypodium distachyon*, *Sorghum bicolor*, *Zea mays* and *Musa acuminata* were obtained from the Plant Genome Duplication Database (http://chibba.agtec.uga.edu/duplication/) [38]. Divergence time between species was retrieved from Time Tree [39].

Ka and Ks computation

To generate a coding sequence alignment for a pair of duplicated genes, their protein sequences were first aligned using Clustalw [40] with default parameters. Then, the protein alignment was converted to coding sequence alignment using the "Bio::Align::Utilities" module of the BioPerl package (http://www.bioperl.org/). Ka and Ks were calculated using the Yang & Nielsen method [41] via the "Bio::Tools::Run::Phylo::PAML::Yn00" module in the BioPerl package.

Identification of WGD, local and dispersed duplicates

In each investigated eudicot genome, the population of potential gene duplications was identified using BLASTP [42] with the following parameter: top five non-self matches and $E<10^{-10}$. MCScanX [43] was applied to each BLASTP output. WGD duplicate pairs were the pairs of colinear genes generated by MCScanX. Potential local duplicate pairs were obtained from the ". tandem" files generated by MCScanX (within 10 annotated genes of each other). To reduce redundancy in potential local duplicate pairs, for each local duplicate, only the local duplicate pair with the smallest Ks was kept for analysis. Potential dispersed duplicate pairs were derived by excluding WGD duplicate pairs and potential local duplicate pairs from the population of potential gene duplications. To reduce redundancy in potential dispersed duplicate pairs, for each dispersed duplicate, only the dispersed duplicate pairs with smallest Ks was kept for analysis.

Specially in *Arabidopsis thaliana*, WGD duplicate pairs created by α, β and γ events were initially obtained from a previous study [3]. Then, α WGD duplicates were updated according to another study [44], to exclude tandemly-duplicated WGD duplicates which were shown to have very similar evolutionary patterns with local duplicates [45]. Tandem and proximal duplicate pairs were obtained from another previous study [33]. Potential dispersed duplicate pairs were then derived by excluding WGD duplicate pairs, tandem and proximal duplicate pairs from the population of potential gene duplications (generated by BLASTP as previously noted). To reduce redundancy in potential dispersed duplicate pairs, for each dispersed duplicate, only the dispersed duplicate pairs with smallest Ks was kept for analysis.

Estimation of relocated γ duplicates in *Carica* and *Theobroma*

Phylogeny between *Carica* (or *Theobroma*) and outgroups is shown in Fig 2A. Gamma duplicates in *Carica* (or *Theobroma*) were initially the intra-genome colinear genes generated by MCScanX since no lineage-specific WGDs have occurred. We then excluded the duplicates without any colineary homologs in outgroups from the γ duplicates, i.e. to exclude recent segmental duplicates. Gene colinearity conservation between *Carica* (or *Theobroma*) and outgroups was used to estimate the epochs of *Carica* (or *Theobroma*) dispersed duplicates. Relocated γ duplicates in *Carica* (or *Theobroma*) were the dispersed duplicates which show colinearity with *Populus*, *Vitis* and/or *Solanum*, but do not show colinearity with monocots or *Musa*. We got 996 γ duplicates and 4024 relocated γ duplicates in *Carica*, and 4267 γ duplicates and 3906 relocated γ duplicates in *Theobroma*.

Enrichment analysis

Enrichment analysis was performed by Fisher's exact test.

Gene function analysis

Gene functions in *Arabidopsis thaliana* were denoted by Gene Ontology (GO) terms, available from TAIR [46]. We selected the 29 GO terms belonging to "biological process" and with >300 *Arabidopsis thaliana* genes for further analysis. For each origin of duplicates, the fold enrichment for each selected GO term was the fraction of the duplicates of this origin having the GO term divided by the faction of the pooled duplicates having the GO term. Then, for each origin of duplicates, its functional profile was denoted by a vector consisting of 29 ratios, which correspond to the fold enrichment for the 29 selected GO terms. Functional similarity between different origins of duplicates was denoted by the Pearson's correlation coefficient (*r*) of their functional profiles. To make the heat map of the functional relationships among different origins of duplicates, average linkage hierarchical clustering with distance = 1- *r* was employed.

Phenotypic data

The phenotypic effects of 5,360 *Arabidopsis* mutant genes (by single locus knockout) were obtained from a published study [34], of which 1,742 showed phenotypic changes, whereas 3,618 did not.

Protein-protein interaction data

We downloaded a dataset of genome-wide protein-protein interactions (*Arabidopsis* Interactome version 1 "main screen", i.e. AI1-main) from a previous study, involving 5,664 binary interactions between 2661 proteins [36]. Note that AI1-main was derived from a population of 8,595 *Arabidopsis* genes. So when we compared the proportions of genes involved in PPIs among different gene groups, the genome background consisted of only these 8,595 genes.

Gene expression and regulation analysis

Processed microarray data measured by the Affymetrix *Arabidopsis* ATH1 Genome Array (GPL198) were obtained from previous studies [12, 47]. The expression divergence between two genes was measured by 1-*r*, where *r* is the Pearson's correlation coefficient between their expression profiles.

The promoter region of each *Arabidopsis thaliana* gene was defined as the DNA sequence between 600bp upstream and 200bp downstream of its transcription start site. A total of 155 position weight matrices (PWMs) for plant genomes were retrieved from the TRANSFAC [48] and JASPAR [49] databases. Based on these PWMs, FIMO [50], a DNA motif search tool, was executed to detect the matched PWMs for the promoter regions of each *Arabidopsis thaliana* gene. The gene regulation similarity between duplicated genes was defined as the fraction of their shared PWMs in their promoter regions. The regulation divergence between duplicated genes is computed by:

$$\text{Gene regulation divergence} = 1 - \frac{\{\text{PWMs in duplicate 1}\} \cap \{\text{PWMs in duplicate 2}\}}{\{\text{PWMs in duplicate 1}\} \cup \{\text{PWMs in duplicate 2}\}}.$$

Supporting Information

S1 Fig. Retention rates of *Arabidopsis lyrata* dispersed duplicates in different epochs.
(PNG)

S2 Fig. Retention rates of *Medicago truncatula* dispersed duplicates in different epochs.
(PNG)

S1 Table. *Arabidopsis thaliana* duplicated genes and their evolutionary origins.
(XLSX)

S2 Table. Functional enrichment analysis of *Arabidopsis thaliana* duplicated genes of different evolutionary origins.
(XLSX)

Acknowledgments

We thank Barry Marler for IT support.

Author Contributions

Conceived and designed the experiments: YW FAF AHP. Performed the experiments: YW. Analyzed the data: YW AHP. Contributed reagents/materials/analysis tools: YW SPF XW. Wrote the paper: YW AHP.

References

1. Ohno S. Evolution by gene duplication. New York: Springer Verlag; 1970.
2. Paterson AH, Freeling M, Tang H, Wang X. Insights from the comparison of plant genome sequences. Annu Rev Plant Biol. 2010; 61:349–72. Epub 2010/05/06. doi: 10.1146/annurev-arplant-042809-112235 PMID: 20441528.
3. Bowers JE, Chapman BA, Rong JK, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 2003; 422(6930):433–8. doi: 10.1038/Nature01521 PMID: ISI:000181801200045.
4. Tang H, Bowers JE, Wang X, Paterson AH. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci U S A. 2010; 107(1):472–7. Epub 2009/12/08. 0908007107 [pii] doi: 10.1073/pnas.0908007107 PMID: 19966307; PubMed Central PMCID: PMC2806719.
5. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011; 473(7345):97–100. Epub 2011/04/12. nature09916 [pii] doi: 10.1038/nature09916 PMID: 21478875.
6. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, et al. A genome triplication associated with early diversification of the core eudicots. Genome Biol. 2012; 13(1):R3. Epub 2012/01/28. gb-2012-13-1-r3 [pii] doi: 10.1186/gb-2012-13-1-r3 PMID: 22280555; PubMed Central PMCID: PMC3334584.

7. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome research. 2008; 18(12):1944–54. doi: 10.1101/gr.080978.108 PMID: ISI:000261398900010.

8. Freeling M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 2009; 60:433–53. Epub 2009/07/07. doi: 10.1146/annurev.arplant.043008.092122 PMID: 19575588.

9. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. Genome research. 2008; 18(12):1924–37. doi: 10.1101/gr.081026.108 PMID: ISI:000261398900008.

10. Woodhouse MR, Tang H, Freeling M. Different gene families in *Arabidopsis thaliana* transposed in different epochs and at different frequencies throughout the rosids. Plant Cell. 2011; 23(12):4241–53. Epub 2011/12/20. tpc.111.093567 [pii] doi: 10.1105/tpc.111.093567 PMID: 22180627; PubMed Central PMCID: PMC3269863.

11. Wang Y, Wang X, Paterson AH. Genome and gene duplications and gene expression divergence: a view from plants. Ann N Y Acad Sci. 2012; 1256(1):1–14. Epub 2012/01/20. doi: 10.1111/j.1749-6632.2011.06384.x PMID: 22257007.

12. Wang Y, Wang X, Tang H, Tan X, Ficklin SP, Feltus FA, et al. Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. PLoS One. 2011; 6(12):e28150. Epub 2011/12/14. doi: 10.1371/journal.pone.0028150 PONE-D-11-17549 [pii]. PMID: 22164235; PubMed Central PMCID: PMC3229532.

13. Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, et al. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res. 1998; 8(5):479–92. Epub 1998/05/20. PMID: 9582192.

14. Ganko EW, Meyers BC, Vision TJ. Divergence in expression between duplicated genes in *Arabidopsis*. Mol Biol Evol. 2007; 24(10):2298–309. doi: 10.1093/molbev/msm158 PMID: ISI:000250437000015.

15. Yang L, Bennetzen JL. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci U S A. 2009; 106(47):19922–7. Epub 2009/11/21. 0908008106 [pii] doi: 10.1073/pnas.0908008106 PMID: 19926865; PubMed Central PMCID: PMC2785268.

16. Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004; 431(7008):569–73. doi: 10.1038/Nature02953 PMID: ISI:000224156700047.

17. Cusack BP, Wolfe KH. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol. 2007; 24(3):679–86. Epub 2006/12/21. msl199 [pii] doi: 10.1093/molbev/msl199 PMID: 17179139.

18. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009; 457(7229):551–6. Epub 2009/02/04. doi: 10.1038/nature07723 PMID: 19189423.

19. Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009; 10(1):19–31. Epub 2008/11/26. nrg2487 [pii] doi: 10.1038/nrg2487 PMID: 19030023.

20. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell. 2004; 16(7):1679–91. Epub 2004/06/23. doi: 10.1105/tpc.021410 [pii]. PMID: 15208398; PubMed Central PMCID: PMC514153.

21. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 2005; 102(15):5454–9. Epub 2005/04/01. 0501102102 [pii] doi: 10.1073/pnas.0501102102 PMID: 15800040; PubMed Central PMCID: PMC556253.

22. Chapman BA, Bowers JE, Feltus FA, Paterson AH. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci U S A. 2006; 103(8):2730–5. doi: 10.1073/pnas.0507782103 PMID: ISI:000235554900046.

23. Rizzon C, Ponger L, Gaut BS. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol. 2006; 2(9):e115. Epub 2006/09/05. 06-PLCB-RA-0127R2 [pii] doi: 10.1371/journal.pcbi.0020115 PMID: 16948529; PubMed Central PMCID: PMC1557586.

24. Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, et al. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. Genome research. 2012; 22(1):95–105. Epub 2011/10/07. gr.125146.111 [pii] doi: 10.1101/gr.125146.111 PMID: 21974993; PubMed Central PMCID: PMC3246211.

25. Blanc G, Wolfe KH. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 2004; 16(7):1667–78. Epub 2004/06/23. doi: 10.1105/tpc.021345 [pii]. PMID: 15208399; PubMed Central PMCID: PMC514152.

26. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000; 290 (5494):1151–5. Epub 2000/11/10. 8976 [pii]. PMID: 11073452.

27. Woodhouse MR, Pedersen B, Freeling M. Transposed genes in *Arabidopsis* are often associated with flanking repeats. PLoS Genet. 2010; 6(5):e1000949. ARTN e1000949 doi: 10.1371/journal.pgen.1000949 PMID: ISI:000278557300019.

28. Grant V. Plant speciation. New York: Columbia Univ Press; 1981.

29. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci U S A. 2009; 106(33):13875–9. doi: 10.1073/pnas.0811575106 PMID: 19667210; PubMed Central PMCID: PMC2728988.

30. Vanneste K, Baele G, Maere S, Van de Peer Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome research. 2014; 24(8):1334–47. doi: 10.1101/gr.168997.113 PMID: 24835588; PubMed Central PMCID: PMC4120086.

31. Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, Ruelens P, et al. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. Mol Biol Evol. 2012; 29(12):3793–806. doi: 10.1093/molbev/mss183 PMID: 22821009.

32. Wang Y, Li J, Paterson AH. MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics. 2013; 29(11):1458–60. doi: 10.1093/bioinformatics/btt150 PMID: 23539305.

33. Wang Y, Tan X, Paterson AH. Different patterns of gene structure divergence following gene duplication in Arabidopsis. BMC Genomics. 2013; 14:652. doi: 10.1186/1471-2164-14-652 PMID: 24063813; PubMed Central PMCID: PMC3848917.

34. Hanada K, Kuromori T, Myouga F, Toyoda T, Li WH, Shinozaki K. Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. Genome Biol Evol. 2009; 1:409–14. Epub 2009/01/01. doi: 10.1093/gbe/evp043 PMID: 20333209; PubMed Central PMCID: PMC2817435.

35. Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell. 2007; 19(2):395–402. Epub 2007/02/13. tpc.106.049338 [pii] doi: 10.1105/tpc.106.049338 PMID: 17293565; PubMed Central PMCID: PMC1867330.

36. *Arabidopsis* Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. Science. 2011; 333(6042):601–7. Epub 2011/07/30. 333/6042/601 [pii] doi: 10.1126/science.1203877 PMID: 21798944; PubMed Central PMCID: PMC3170756.

37. Wang X, Jin D, Wang Z, Guo H, Zhang L, Wang L, et al. Telomere-centric genome repatterning determines recurring chromosome number reductions during the evolution of eukaryotes. The New phytologist. 2015; 205(1):378–89. doi: 10.1111/nph.12985 PMID: 25138576.

38. Lee TH, Tang H, Wang X, Paterson AH. PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res. 2013; 41(Database issue):D1152–8. doi: 10.1093/nar/gks1104 PMID: 23180799; PubMed Central PMCID: PMC3531184.

39. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006; 22(23):2971–2. doi: 10.1093/bioinformatics/btl505 PMID: 17021158.

40. Thompson JD, Higgins DG, Gibson TJ. Clustal-W—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22(22):4673–80. PMID: ISI:A1994PU19900018.

41. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000; 17(1):32–43. Epub 2000/02/10. PMID: 10666704.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. doi: 10.1016/S0022-2836(05)80360-2 PMID: 2231712.

43. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012; 40(7):e49. doi: 10.1093/nar/gkr1293 PMID: 22217600; PubMed Central PMCID: PMC3326336.

44. Thomas BC, Pedersen B, Freeling M. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome research. 2006; 16(7):934–46. Epub 2006/06/09. gr.4708406 [pii] doi: 10.1101/gr.4708406 PMID: 16760422; PubMed Central PMCID: PMC1484460.

45. Wang Y. Locally duplicated ohnologs evolve faster than nonlocally duplicated ohnologs in Arabidopsis and rice. Genome Biol Evol. 2013; 5(2):362–9. doi: 10.1093/gbe/evt016 PMID: 23362157; PubMed Central PMCID: PMC3590777.

46. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 2008; 36 (Database issue):D1009–14. Epub 2007/11/08. gkm965 [pii] doi: 10.1093/nar/gkm965 PMID: 17986450; PubMed Central PMCID: PMC2238962.

47. Spangler JB, Subramaniam S, Freeling M, Feltus FA. Evidence of function for conserved noncoding sequences in *Arabidopsis thaliana*. The New phytologist. 2012; 193(1):241–52. Epub 2011/10/01. doi: 10.1111/j.1469-8137.2011.03916.x PMID: 21955124.

48. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34(Database issue):D108–10. doi: 10.1093/nar/gkj143 PMID: 16381825; PubMed Central PMCID: PMC1347505.

49. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010; 38(Database issue):D105–10. doi: 10.1093/nar/gkp950 PMID: 19906716; PubMed Central PMCID: PMC2808906.

50. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27(7):1017–8. doi: 10.1093/bioinformatics/btr064 PMID: 21330290; PubMed Central PMCID: PMC3065696.