Qiao *et al. Genome Biology* (2019) 20:38
https://doi.org/10.1186/s13059-019-1650-2

Genome Biology

RESEARCH Open Access

Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants

Xin Qiao[1], Qionghou Li[1], Hao Yin[1], Kaijie Qi[1], Leiting Li[1], Runze Wang[1], Shaoling Zhang[1*]
and Andrew H. Paterson[2*]

Abstract

Background: The sharp increase of plant genome and transcriptome data provide valuable resources to investigate evolutionary consequences of gene duplication in a range of taxa, and unravel common principles underlying duplicate gene retention.

Results: We survey 141 sequenced plant genomes to elucidate consequences of gene and genome duplication, processes central to the evolution of biodiversity. We develop a pipeline named *DupGen_finder* to identify different modes of gene duplication in plants. Genes derived from whole-genome, tandem, proximal, transposed, or dispersed duplication differ in abundance, selection pressure, expression divergence, and gene conversion rate among genomes. The number of WGD-derived duplicate genes decreases exponentially with increasing age of duplication events—transposed duplication- and dispersed duplication-derived genes declined in parallel. In contrast, the frequency of tandem and proximal duplications showed no significant decrease over time, providing a continuous supply of variants available for adaptation to continuously changing environments. Moreover, tandem and proximal duplicates experienced stronger selective pressure than genes formed by other modes and evolved toward biased functional roles involved in plant self-defense. The rate of gene conversion among WGD-derived gene pairs declined over time, peaking shortly after polyploidization. To provide a platform for accessing duplicated gene pairs in different plants, we constructed the Plant Duplicate Gene Database.

Conclusions: We identify a comprehensive landscape of different modes of gene duplication across the plant kingdom by comparing 141 genomes, which provides a solid foundation for further investigation of the dynamic evolution of duplicate genes.

Keywords: Gene duplication, Evolution, Polyploidization, Gene conversion, Plant

Background

The finding that the first fully sequenced eukaryote genome, that of the budding yeast (*Saccharomyces cerevisiae*) [1], had experienced whole-genome duplication (WGD, or defined as polyploidization) [2] invigorated research into this evolutionary mechanism of central importance. The otherwise compact ciliate (*Paramecium tetraurelia*) genome (72 Mb) has nonetheless retained a high number of gene sets (40,000) after at least three successive whole-genome duplications [3–5]. The ancestral vertebrate is thought to have undergone two rounds of ancient WGD (defined as 1R and 2R) at least ~ 450 million years ago (Mya) [6–8]—about 20–30% of human genes are thought to be paralogs produced by these two WGDs, and these “ohnologs” have a strong association with human disease [7, 9]. Additional WGDs occurred in the common ancestor of teleost fish (3R, ~ 320 Mya) [10, 11] and salmonids including the rainbow trout (*Oncorhynchus mykiss*) and Atlantic salmon (*Salmo salar*) (4R) dated to ~ 80 Mya [12, 13]. The most recent genome duplication currently known in vertebrates has been uncovered in the common carp (*Cyprinus carpio*) (4R, ~ 8.2 Mya) [14, 15].

* Correspondence: slzhang@njau.edu.cn; paterson@uga.edu
[1]Centre of Pear Engineering Technology Research, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China
[2]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30605, USA

In contrast with fungi and animals, the most frequent occurrence of paleo-polyploidization has been detected in angiosperms, flowering plants. It has been suggested that one to two genome duplications preceded angiosperm diversification [16], and only one angiosperm is known that did not experience additional WGDs, *Amborella trichopoda* [17]. *Arabidopsis thaliana*, chosen to be the first fully sequenced angiosperm in part due to its apparent genomic simplicity, is, ironically, a member of the Brassicaceae family that is as yet unmatched in its propensity for genome duplication—*Brassica napus* has experienced an aggregate 72× multiplication, in five events (3 × 2 × 2 × 3 × 2) at times ranging from > 100 million to ~ 10,000 years ago [18]. A WGD series of rho (ρ)–sigma (σ)–tau (τ) in Poaceae [19] echoes the now-classic alpha (α)–beta (β)–gamma (γ) series in Brassicaceae [20]. While most plant paleopolyploidies are duplications, several are triplications [21–23] and at least one is a penta-plication [24].

Whole-genome duplication is thought to have contributed much to the evolution of morphological and physiological diversity [25, 26]. However, WGD is often followed by loss of most duplicated genes over a few million years [27] and is episodic [19, 20]. Successive WGD events are often separated by tens of millions of years, failing to provide a continuous supply of variants available for adaptation to continuously changing environments. Diploidization is thought to occur "quickly" (i.e., in the first few million years, [27]) following WGD to return to disomic inheritance, by genome modifications including chromosomal rearrangement, gene loss, gene conversion, subgenome dominance, and expression divergence between duplicate copies [28–30]. The tiny genome (82 Mb) of bladderwort (*Utricularia gibba*) which accommodates a typical number of genes for a plant but purges almost all intergenic DNA and repeat sequence exemplifies the extreme genome reduction or fractionation after multiple rounds of WGD [31].

With a diploidized state restored soon after genome duplication, what is the raw material for adaptation in taxa that have abstained from genome duplication for long time periods? Various types of single-gene duplication occur more or less continuously and have been implicated in key environmental adaptations [32, 33], but yield genes with short half-lives [27]. De novo gene evolution, for example as a result of transposable element activities [34], may often form fragmentary products of uncertain function [35]. In addition to whole-genome duplication, other modes of gene duplication are collectively deemed single-gene duplications [36–38]. Single genes can move, or be copied, from the original chromosomal position to a new position by various ways [39–41]. Tandem duplicates are closely adjacent to each other in the same chromosome, a phenomenon which is speculated to occur through unequal crossing over [36]. Proximal duplication (PD) generates gene copies that are near each other but separated by several genes (10 or fewer genes), possibly through localized transposon activities [42] or originating from ancient tandem duplicates interrupted by other genes [39]. It has been revealed that neighboring genes tend to be co-regulated, especially tandem duplicates [43], and neighboring gene pairs still show interchromosomal colocalization after their separation [44]. Moreover, tandem duplicates have been commonly found to be important for plant adaptation to rapidly changing environments [45]. The transposed duplication (TRD) generates a gene pair comprised of an ancestral and a novel locus and is presumed to arise through distantly transposed duplications occurred by DNA-based or RNA-based mechanisms [38, 46]. Dispersed duplication (DSD) happens through unpredictable and random patterns by mechanisms that remain unclear, generating two gene copies that are neither neighboring nor colinear [47]. The dispersed duplicates are prevalent in different plant genomes [48].

Herein, we exploited a pipeline incorporating syntenic and phylogenomic approaches to identify the different modes of gene duplication in 141 sequenced plant genomes. Duplicated genes were classified into five types, including whole-genome duplication, tandem duplication (TD), proximal duplication, transposed duplication, and dispersed duplication. Integrated large-scale genome and transcriptome datasets were used to investigate selection pressures, expression divergence, and gene conversion underlying duplicate gene evolution. In addition, construction of gene families using all genes from 141 plant genomes suggested 232 families most widely preserved across the plant kingdom. The results of this study lay a substantial foundation for further investigating the contributions of gene duplication to gene regulatory network evolution, epigenetic variation, morphological complexity, and adaptive evolution in plants.

Results

The landscape of gene duplication in the plant kingdom

In 141 sequenced plant genomes, we identified duplicated genes using *DupGen_finder* (freely available at https://github.com/qiao-xin/DupGen_finder) and classified them into one of the five categories (Additional file 1: Figure S1 and Additional file 2), being derived from WGD, TD, PD, TRD, and DSD. The number of duplicate gene pairs for each category in each taxon was determined (Fig. 1 and Additional file 3). The higher percentages of WGD-derived gene pairs were detected in plants experiencing more recent WGDs such as soybean (*Glycine max*, ~ 13 Mya) and flax (*Linum usitatissimum*, 3.7~6.8 Mya). Interestingly, the highest frequency of whole-genome triplication (WGT) occurred in plants belonging to

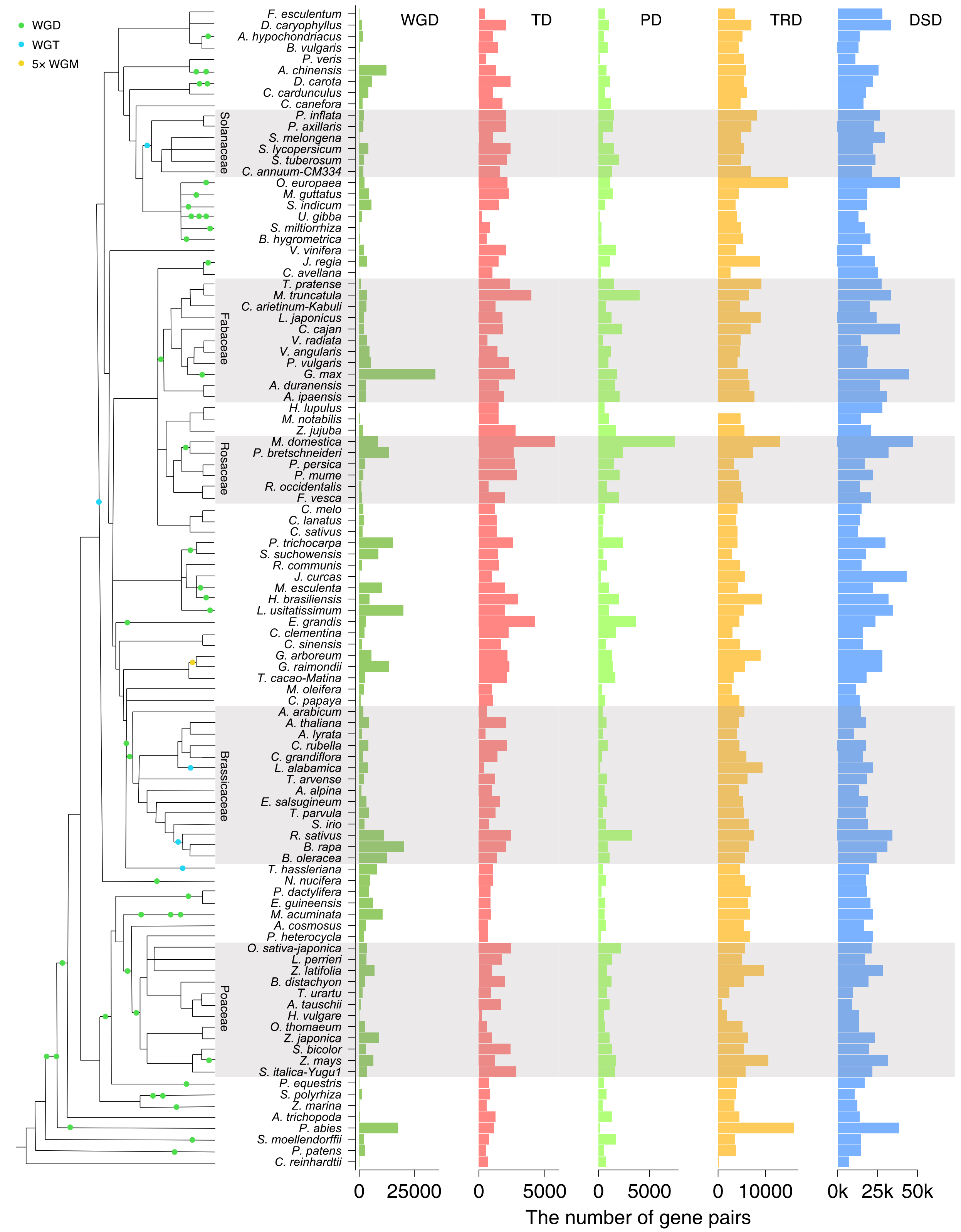

Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 The number of gene pairs derived from different modes of duplication in representative plant genomes. WGD whole-genome duplication, TD tandem duplication, PD proximal duplication, TRD transposed duplication, DSD dispersed duplication. A schematic diagram of phylogeny of different plant species [231–233] and the WGDs occurred in different branches were labeled. Branch length is not directly proportional to time scale. WGT whole-genome triplication, WGM whole-genome multiplication

Brassicaceae such as cabbage (*Brassica oleracea*), radish (*Raphanus sativus*), and *Leavenworthia alabamica*. In addition, the occurrence of genome duplication is frequent in some individual plants such as kiwifruit (*Actinidia chinensis*, two rounds of WGD), carrot (*Daucus carota*, a WGT (Dc-β) and a WGD (Dc-α)), and banana (*Musa acuminata*, three rounds of WGD). Larger percentages of WGD-derived gene pairs are still maintained in the aforementioned species although genome fractionation occurred quickly after genome duplication. To provide a platform for accessing and searching duplicated gene in 141 sequenced plants, we constructed a public database named Plant Duplicate Gene Database (PlantDGD, freely available at http://pdgd.njau.edu.cn:8080).

Identifying K_s peaks corresponding to genome duplication events of different ages in each species

The most recent and more ancient genome duplication events that affect each of the taxa were identified (Additional file 4). To identify the most recent and more ancient K_s peaks (or WGDs) in each species, we estimated the mean K_s values for the gene pairs contained in each syntenic block within a species, and in addition, the K_s distribution was fitted using Gaussian mixture models (GMM) (the code is freely available at https://github.com/qiao-xin/Scripts_for_GB).

Ranges of K_s values for estimates of individual genome duplication events (e.g., γ WGT in core eudicots) from different taxa reflect substantial divergence in evolutionary rates (clock-like rates, substitutions/synonymous site/year) in specific lineages (Fig. 2 and Additional file 4). There are 16 species which have not been influenced by recent genome duplication event but share the core eudicot γ WGT events. The K_s peaks corresponding to the γ WGT from these 16 taxa range from 1.91 to 3.64 (Fig. 2a). For example, we detected strong signal of γ WGT in grape (*Vitis vinifera*) (Fig. 2b). The K_s values corresponding to the cucurbit-common tetraploidization (CCT) range from 2.44 to 2.56. The K_s values corresponding to the Poaceae ρ WGD range from 1.98 to 2.34. For example, two K_s peaks corresponding to ρ WGD and σ/τ WGD were detected in rice (*Oryza sativa*) (Fig. 2d). The K_s values corresponding to the Fabaceae common WGD range from 1.13 to 1.66. The K_s values corresponding to the Brassicaceae α/β WGD range from 1.18 to 1.66. For example, two K_s peaks corresponding to α/β WGD and γ WGT were fitted by a GMM method in *Arabidopsis* (Fig. 2f). The K_s values corresponding to the Solanaceae common WGT range from 1.17 to 1.46. The K_s values corresponding to the cotton 5× WGM (whole-genome multiplication) range from 0.86 to 0.93. The K_s values corresponding to the Brassica common WGT range from 0.48 to 0.52. For example, three K_s peaks corresponding to Brassica WGT, α/β WGD, and γ WGT respectively, were fitted in *Brassica oleracea* (Fig. 2i). The K_s values corresponding to the Salicaceae common WGD range from 0.34 to 0.56. The K_s values corresponding to the Pomoideae WGD range from 0.27 to 0.39.

Dynamic changes in abundance of duplicated genes over time

The most recent K_s peaks were used to determine the order in which the taxa are shown in Fig. 3a (types of gene duplications). Genomes with abnormal K_s peaks were not included in Fig. 3a because fragmented assembly hindered the identification of large syntenic blocks. We detected whole-genome duplication in all plant genomes investigated except for several with highly fragmented assemblies such as Hop (*Humulus lupulus*) and European hazelnut (*Corylus avellana*). The K_s values for duplication events show a steady decline with decreasing antiquity (Fig. 3b), as expected.

Linear regression between the number of each type of duplicated gene pair and the K_s peaks from different taxa showed that the number of gene pairs derived from WGD generally declines with increasing antiquity of duplication events ($r = -0.45$, $P < 0.001$, Additional file 1: Figure S2A), although again with substantial fluctuation among taxa (Fig. 3a). Paralleling the decline in WGD-derived gene pairs with increasing antiquity is decreases in TRD-derived ($r = -0.50$, $P < 0.001$) or DSD-derived gene pairs ($r = -0.57$, $P < 0.001$) (Additional file 1: Figure S2D and E). Tandem and proximal duplicate pairs show a nominal (nonsignificant) decrease ($r = -0.11$, $P = 0.25$ and $r = -0.10$, $P = 0.28$) (Additional file 1: Figure S2B and C).

Further, the absolute number of duplicate gene pairs for each category in each taxon was converted to log10-transformed number to mitigate the effect of genome size and total gene number variation among taxa. Linear regression between the log10-transformed number of each type of duplicated gene pair and the K_s peaks from different taxa strongly supported that the number of duplicated gene pairs derived from WGD ($r = -0.70$, $P < 0.001$), TRD ($r = -0.49$, $P < 0.001$), and DSD ($r = -$

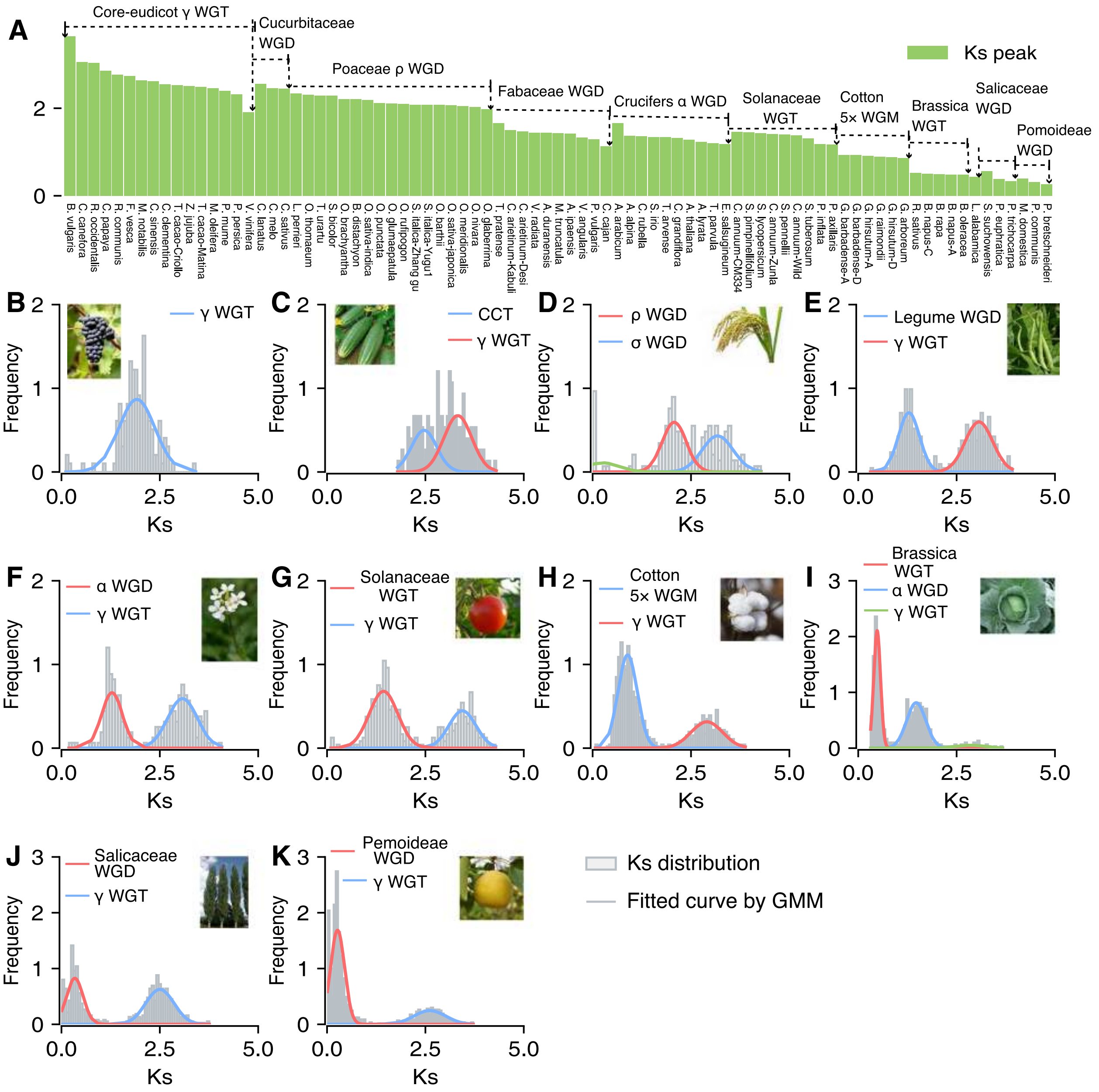

Fig. 2 Lineage-specific genome duplication events. **a** The genome duplication events identified in different lineages. A range of K_s values for estimates of individual genome duplication events from different taxa. The K_s distribution in each species was fitted using Gaussian mixture models (GMM). The K_s peaks corresponding to core eudicot γ WGT, cucurbit-common tetraploidization (CCT), Poaceae ρ WGD, Fabaceae common WGD, Brassicaceae α WGD, Solanaceae WGT, cotton 5× WGM, Brassica WGT, Salicaceae WGD, and Pomoideae WGD were respectively detected in *Vitis vinifera* (**b**), *Cucumis sativus* (**c**), *Oryza sativa* (**d**), *Phaseolus vulgaris* (**e**), *Arabidopsis thaliana* (**f**), *Solanum lycopersicum* (**g**), *Gossypium raimondii* (**h**), *Brassica oleracea* (**i**), *Populus trichocarpa* (**j**), and *Pyrus bretschneideri* (**k**)

0.61, $P < 0.001$) significantly declines with increasing antiquity of duplication events (Additional file 1: Figure S3A, D and E). However, the number of tandem and proximal duplicates showed no significant decrease over time ($r = -0.08$, $P = 0.43$ and $r = 0.02$, $P = 0.84$) and may provide a continuous supply of genes potentially useful for plant adaptation. Moreover, the exponential fit was performed between log10-transformed numbers (y axis) and K_s peaks (x axis). The number of WGD-derived pairs decreases exponentially with increasing antiquity of duplication events (Fig. 3c). The chi-squared goodness of fit test supports this observation (or null hypothesis) ($\chi2 = 2.33$, $P = 1.0$). Exponential decrease of number of duplicated genes over time was also found in TRD- and DSD-derived duplicate genes ($\chi2 = 0.47$, $P = 1.0$ and $\chi2 = 0.37$, $P = 1.0$) (Fig. 3f, g). Significant exponential decay

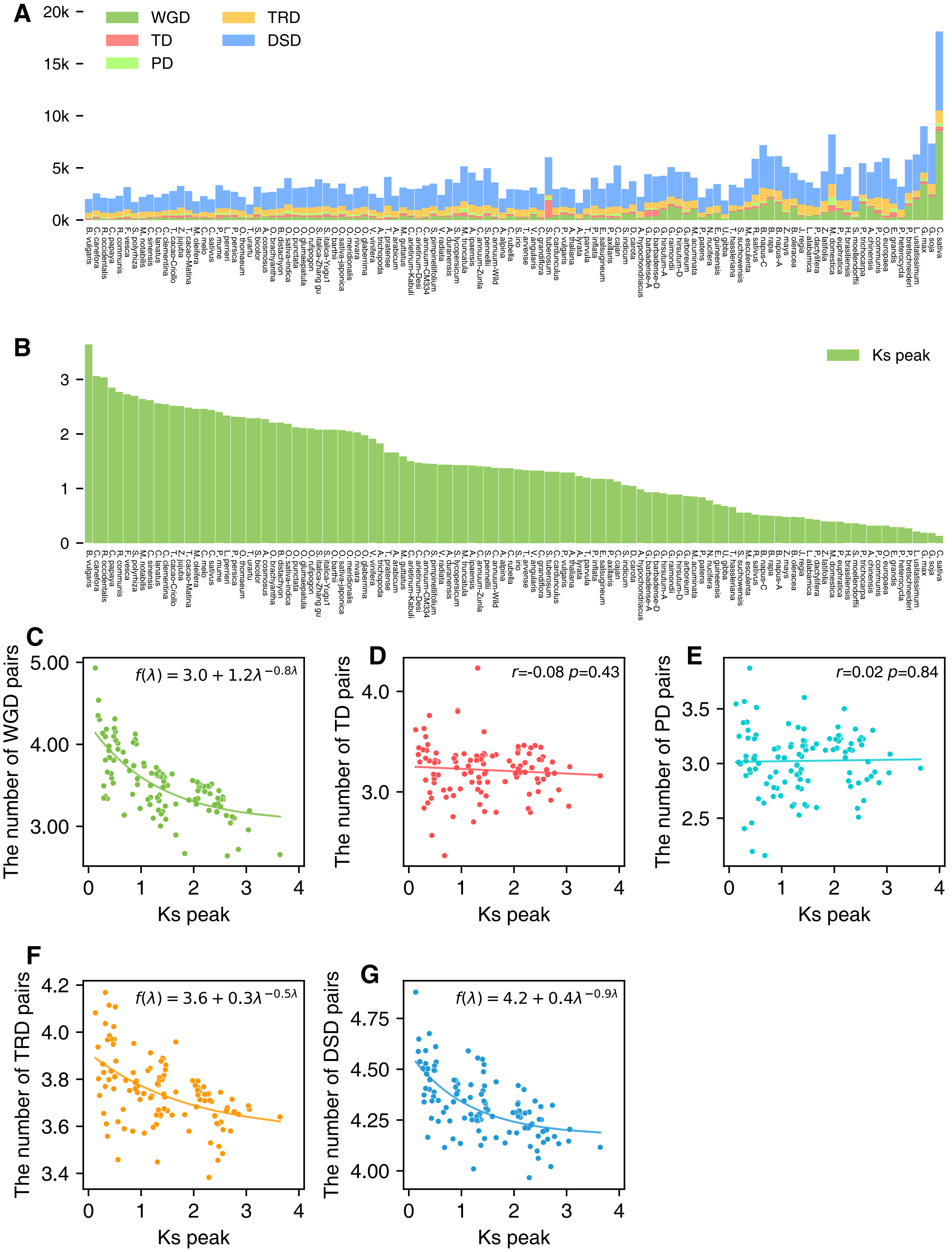

Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Changes in abundance of different modes of duplicated gene pairs over time. **a** The distribution of number of gene pairs derived from different modes of duplication in 141 plant genomes. Genomes with abnormal K_s peaks because fragmented assembly hindered the identification of large syntenic blocks were not included. **b** The fitted K_s peak corresponding to the most recent WGD for each species. **c**–**g** The relationship between the log10-transformed number of different types of gene pairs and K_s peak of WGD genes from different taxa, excluding those taxa with abnormal K_s peaks due to fragmented assembly. **c** WGD-pairs. **d** TD-pairs: transposed gene pairs. **e** PD-pairs: proximal gene pairs. **f** TRD-pairs: transposed gene pairs. **g** DSD-pairs: dispersed gene pairs. Exponential fit and linear regression analysis were performed. The exponential equation was annotated in subplots **c**, **f**, and **g**; Pearson correlation coefficient (*r*) was annotated in subplots **d** and **e**

was not found in TD- and PD-derived duplicate genes (Fig. 3d, e).

To investigate whether results of the aforementioned linear regression analyses have bias due to some individual genome duplication events being shared among different taxa, we undertook new analyses using only one that sampled each of the most recent genome duplication events (Fig. 2, noting that ancient events were unavoidably shared across species). The results from this new analysis supported that the number of duplicated gene pairs derived from WGD ($r = -0.39$, $P < 0.05$), TRD ($r = -0.46$, $P < 0.001$), and DSD ($r = -0.56$, $P < 0.001$) declines significantly with increasing antiquity of duplication events (Additional file 1: Figure S4A, D and E). The number of tandem and proximal duplicates showed no significant decrease over time ($r = -0.22$, $P = 0.17$ and $r = -0.23$, $P = 0.16$) (Additional file 1**:** Figure S4B and C). Linear regression analysis using the log10-transformed number of each type of duplicated gene pair also supported our prior observation (Additional file 1: Figure S5).

Evolutionary forces inferred to affect duplicated genes

The K_a (number of substitutions per nonsynonymous site), K_s (number of substitutions per synonymous site), and K_a/K_s values were estimated for gene pairs generated by different modes of duplication. We compared the K_a, K_s, and K_a/K_s distributions across 141 plants (Fig. 4 and Additional file 1: Figure S6 and S7). The K_a/K_s ratios among different modes of gene duplications showed a striking trend, with tandem and proximal duplications having qualitatively higher K_a/K_s ratios than other modes. The TD- and PD-derived gene pairs have relatively smaller K_s values (Additional file 1: Figure S7). This finding suggests that tandem and proximal duplications of younger age that have been preserved have experienced more rapid sequence divergence than other gene classes, although concerted evolution may also preserve homogeneity of TD or PD genes to a greater degree than genes that are not located near one another. In contrast, WGD genes are more conserved with smaller K_a/K_s ratios.

We further explored the roles of purifying selection ($K_a/K_s < 1$) and positive selection ($K_a/K_s > 1$) in the evolution of duplicated genes in seven model plants, including *Arabidopsis thaliana* (eudicots), *Oryza sativa* (monocots), *Amborella trichopoda* (angiosperm, Amborellales), *Picea abies* (Norway spruce, gymnosperms), *Selaginella moellendorffii* (Lycophytes), *Physcomitrella patens* (Bryophytes), and *Chlamydomonas reinhardtii* (Chlorophytes). The majority of duplicated genes evolve under purifying selection ($K_a/K_s < 1$) **(**Additional file 1: Table S1 and Figure S8-S14). In *Arabidopsis*, 100% WGD-, 96.5% TD-, 94.9% PD-, 99.7% TRD-, and 99.3% DSD-derived duplicate genes experienced purifying selection, while only 0.0–5.1% of duplicated genes show evidence of positive selection (Additional file 1: Figure S8). Likewise, evidence of purifying selection is found for 98.3–99.7% of duplicated genes in *O. sativa*, 91.9–97.2% in *A. trichopoda*, 86.2–98.8% in *P. abies*, 91.8–98.3% in *S. moellendorffii*, 91.4–99.7% in *P. patens*, and 95.7–98.7% in *C. reinhardtii*. Consistent with our earlier observation, tandem and proximal duplicates experienced stronger positive selection than other modes (Fig. 4 and Additional file 1: Table S1), reflected by the high percentages of gene pairs showing $K_a/K_s > 1$ in *Arabidopsis* (PD (5.1%) > TD (3.3%) > DSD (0.6%) > TRD (0.3%) > WGD (0.0%)) and other model plants. This finding suggests that tandem and proximal duplication is an important source of genetic material for evolving new functions.

Does stronger selective pressure drive the evolution of tandem and proximal duplicates toward specific biological functions? To answer this question, we performed GO enrichment analysis to investigate the functional roles of tandem and proximal genes in the model plant *A. thaliana*, given its high-quality genome annotation and extensive functional analysis. Tandem and proximal duplicates exhibited divergent functional roles although they shared several enriched GO terms involved in defense response, drug binding, endomembrane system, monooxygenase activity, oxidoreductase activity, and oxygen binding, which are critical for plant self-defense and adaptation (Additional file 5). In particular, proximal duplicates are enriched in GO terms involved in apoptotic processes, cell death, programmed cell death, immune response, and signaling receptor activity. Tandem duplicates are enriched in GO terms involved in "binding," such as tetrapyrrole binding, iron

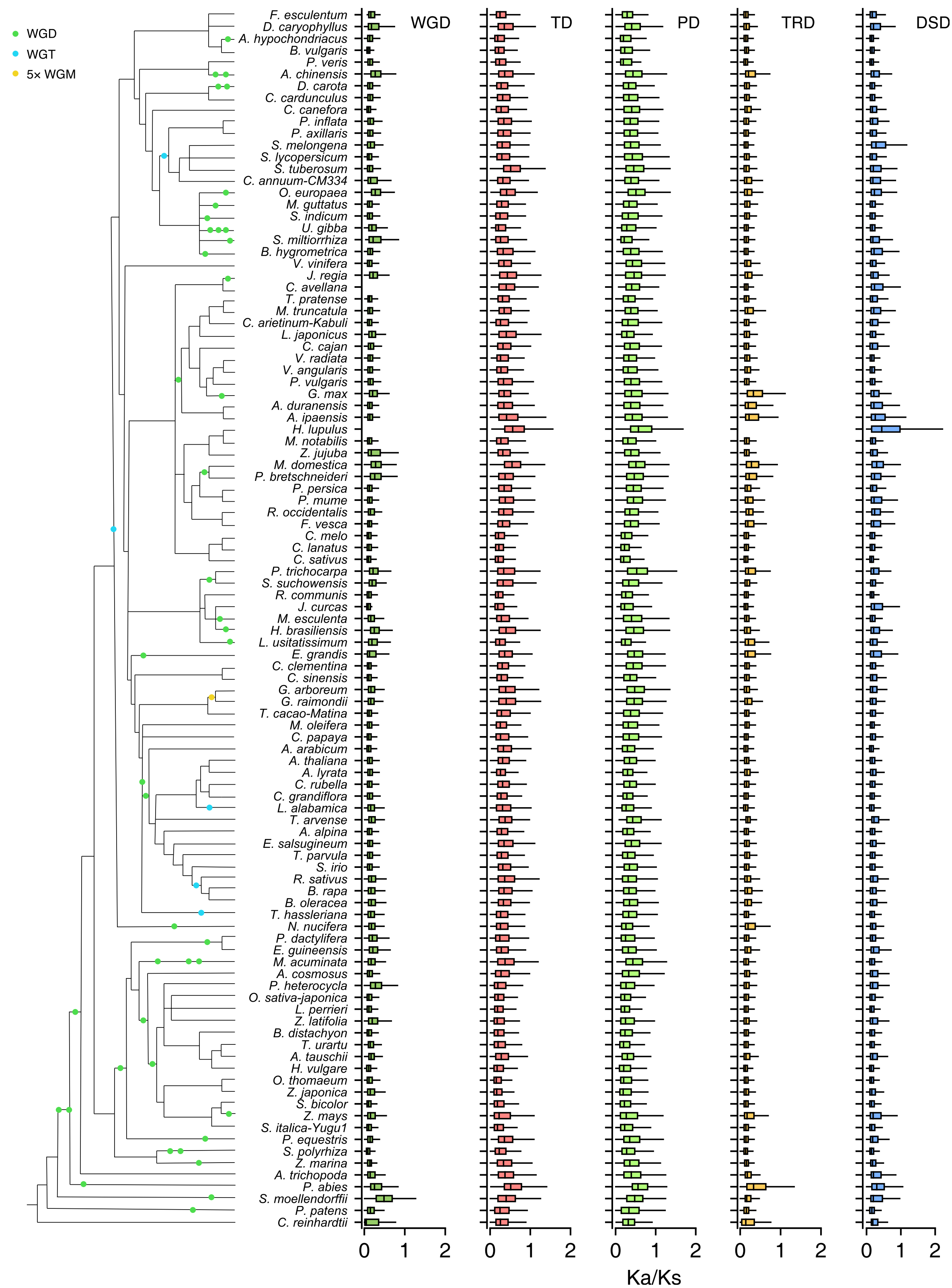

Fig. 4 The K_a/K_s ratio distributions of gene pairs derived from different modes of duplication in representative plant genomes. WGD whole-genome duplication, TD tandem duplication, PD proximal duplication, TRD transposed duplication, DSD dispersed duplication

ion binding, heme binding, and cofactor binding, and "activity" such as transferase activity, hydrolase activity, electron transfer activity, and catalytic activity (Fig. 5).

Expression divergence between duplicated genes

Large-scale RNA-seq data from different tissues, development stages, and treatments are available for a range of plant taxa (Additional file 6). Here, we investigated patterns of expression divergence between duplicated genes in eight model plants for WGD, tandem, proximal, transposed, and dispersed gene pairs. Log10-transformed TPM (transcripts per million) values were used as a proxy for expression levels. For duplicated pairs in which both gene copies are expressed in at least one tissue or condition, Pearson's correlation coefficient (*r*) was calculated between the expression profiles of the two genes, also calculating *r* for 10,000 randomly selected gene pairs for each species. The 95% quantile in the *r* value distribution for random gene pairs was taken as the significance threshold for determining that two gene copies of a duplicated pair have diverged in expression (Additional file 1: Figures S15 and S16). The results showed diverged expression profiles (Fig. 6a–h) for 87%, 66%, 80%, 81%, 84%, 66%, 85%, and 71% of WGD-derived gene pairs in *C. reinhardtii* (Chlorophytes), *P. patens* (Bryophytes), *S. moellendorffii* (Lycophytes), *P. abies* (Norway spruce, gymnosperms), *A. trichopoda* (angiosperm, Amborellales), *O. sativa* (monocots), *N. nucifera* (eudicots, Proteales), and *A. thaliana* (eudicots), respectively. Similarly, 63–85% TD-, 76–85% PD-, 73–92% TRD-, and 74–88% DSD-derived pairs showed expression divergence.

Furthermore, we investigated expression divergence between duplicated genes after genome duplication or triplication events of different ages in strategically chosen monocots and eudicots. Grasses share sigma WGD (σ, 100~120 Mya) and tau WGD (τ, 110~135 Mya) with *A. comosus* (not including the angiosperm-wide event and beyond) [49]. After divergence from the lineage of *A. comosus*, the common ancestor of grasses including *S. bicolor*, *O. sativa*, and *Z. mays* experienced rho WGD (ρ, 95~115 Mya) [50]. In addition, *Z. mays* experienced an additional species-specific event (mWGD, ~ 26 Mya) [50]. Brassicaceae share core eudicot gamma WGT events (γ, ~ 117 Mya) with *V. vinifera* [23]. After divergence with *V. vinifera*, the common ancestor of Brassicaceae including *Arabidopsis*, *B. oleracea* and *C. sativa* experienced alpha WGD (α, ~ 35 Mya) [21] and beta WGD (β, 50~60 Mya) [16, 51]. Following Brassicaceae diversification, *B. oleracea* and *C. sativa* independently experienced species-specific genome triplication events, at ~ 15.9 Mya [21] and ~ 5.41 Mya [52] respectively.

Eudicots *C. sativa*, *B. oleracea* (cabbage), and *V. vinifera* (grape) have been influenced by three different ages of whole-genome triplication, estimated to have occurred at ~ 5.41 [52], ~ 15.9 [21], and ~ 117 [23] Mya respectively. The proportion of WGD-pairs with divergent expression in these three plants increases with the time after duplication from 43 to 86% ($P < 0.001$, Fisher's exact test) (Fig. 6n), with > 50% of WGD-pairs still undifferentiated in expression ~ 5.41 My after duplication. Monocots *Z. mays* (maize), *S. bicolor* (sorghum), and *A. comosus* (pineapple) also offer stratified ages of whole-genome duplication, at ~ 26 [50], 95~115 [50], and 100~120 [49] Mya respectively. The proportion of WGD-pairs with divergent expression in these three plants increases from 50 to 77% ($P < 0.001$, Fisher's exact test) (Fig. 6l), with 50% of WGD-pairs showing undifferentiated expression after ~ 26 My.

Moreover, the model plant *Arabidopsis* alone provided three genome duplications: alpha (α), beta (β), and gamma (γ). The proportion of gene pairs with divergent expression from these three WGD events increases from 65 to 84% ($P < 0.001$, Fisher's exact test) (Fig. 6k). The cereal crop rice provided two rounds of genome duplication, rho (ρ) and sigma (σ). The proportion of gene pairs with divergent expression from these two WGD events increases from 63 to 74% ($P < 0.05$, Fisher's exact test) (Fig. 6i). These results indicated that WGD-derived gene pairs show gradually increasing expression divergence with age.

The rate of gene conversion between WGD-derived paralogs declined over time

We investigated the gene conversion rates of duplicated genes derived from Brassicaceae α WGD and Poaceae ρ WGD over evolutionary time (Fig. 7). Firstly, high-confidence α WGD-derived gene pairs from *A. thaliana* were retrieved from a previous report [20]. The gene conversion rates after divergence between the *A. thaliana* lineage and those of *Aethionema arabicum*, *Eutrema salsugineum*, *Capsella rubella*, or *Arabidopsis lyrata* were examined respectively. K_s was used as a proxy for evolutionary time. Brassicaceae α WGD has been dated to ~ 35 Mya [21]. In this study, the average of a range of K_s values for estimates of the Brassicaceae α WGD event from different Brassicaceae plants is approximately 1.3. To estimate the time of speciation, we calculated the mean K_s values for the gene pairs contained in each syntenic block between the *Arabidopsis* lineage and each outgroup species, and further, the K_s distribution was fitted using Gaussian mixture models (GMM) (Additional file 1: Figures S17 and S18). The divergence between the *Arabidopsis* lineage and *A. arabicum* occurred shortly after α WGD and dated to $K_s = 1.0$. The divergences between the *Arabidopsis* lineage and *E. salsugineum*, *C. rubella*, and *A. lyrata* were respectively dated to 0.5, 0.4, and 0.2. The number of gene conversion events among α WGD-derived duplicated gene pairs is 104 after the divergence of *Arabidopsis* and *A. arabicum*, over 50-fold higher

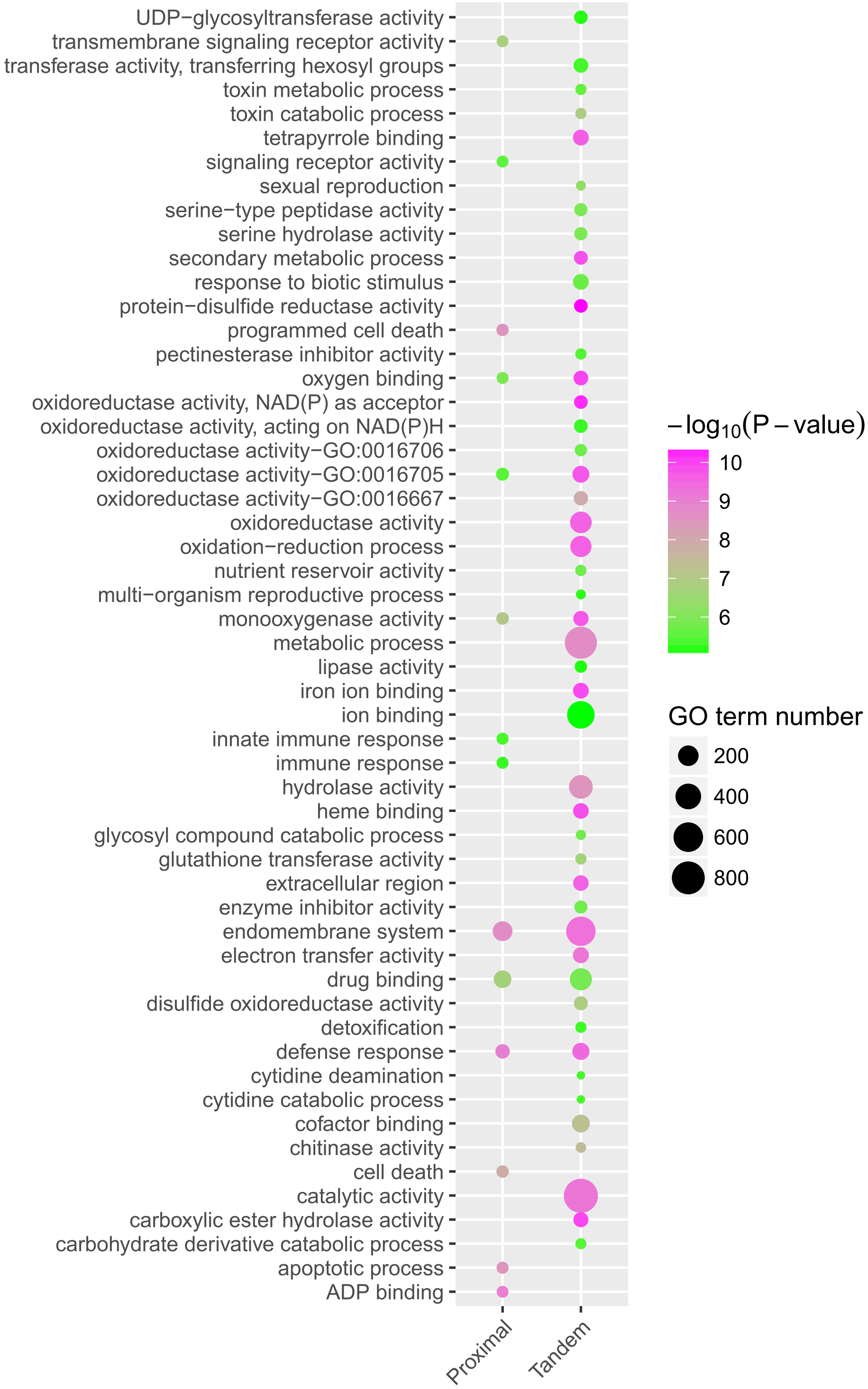

Fig. 5 Functional enrichment analysis of tandem and proximal duplicates in *Arabidopsis*. The enriched GO terms with corrected *P* value < 0.01 are presented. The color of circle represents the statistical significance of enriched GO terms. The size of the circles represents the number of occurrences of a GO term

than the number after the divergence of *Arabidopsis* and *A. lyrata* (Fig. 7a, b). This result indicated that gene conversion was extensive shortly after polyploidization and declined over time, a result that has been strongly supported by independent evidence [53]. Moreover, the gene conversion rates after divergence between *Oryza sativa* L. (ssp. japonica) lineage and

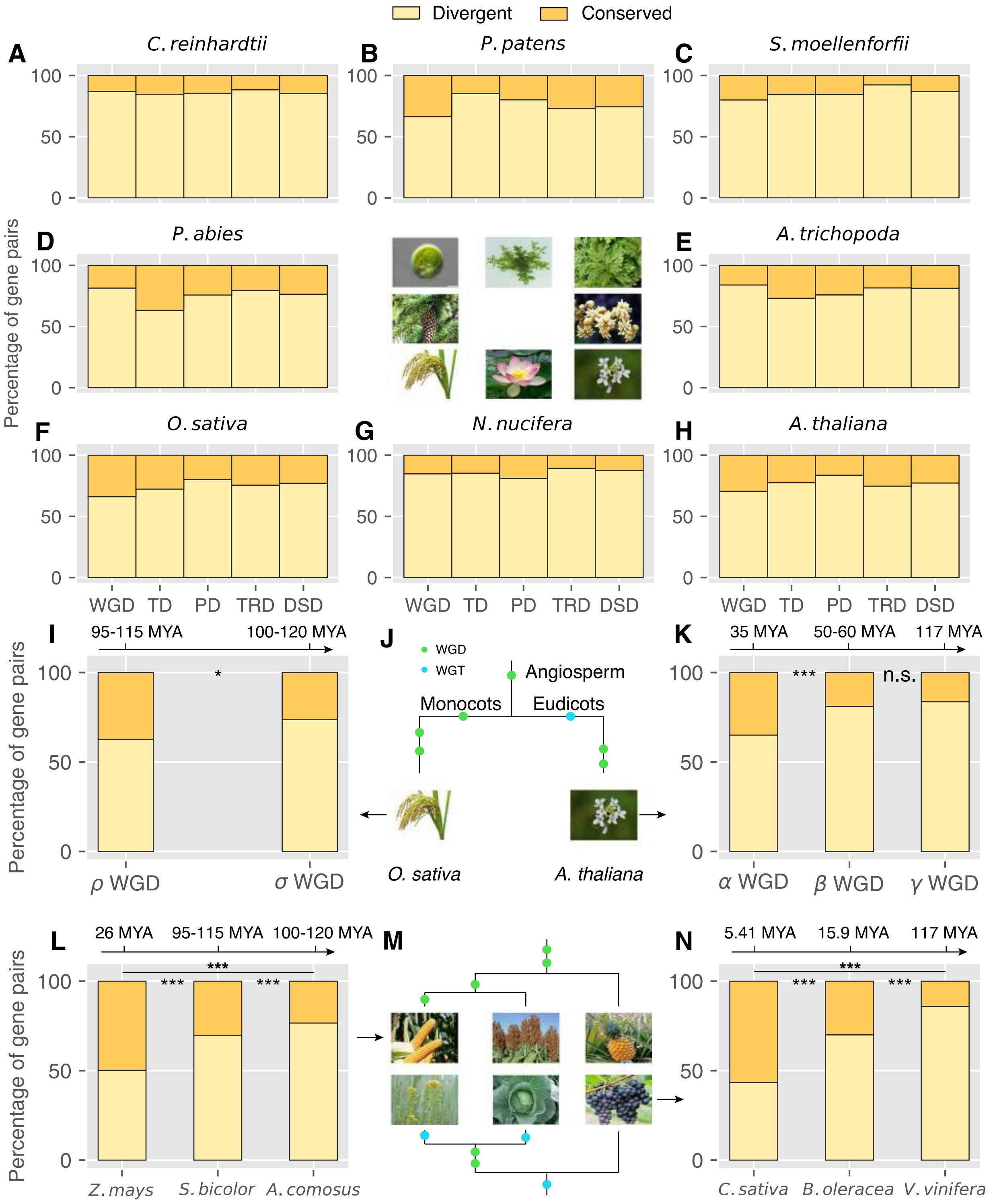

Fig. 6 Expression divergence between duplicate genes derived from WGD, tandem (TD), proximal (PD), transposed (TRD), and dispersed (DSD) duplication in **a** *C. reinhardtii*, **b** *P. patens*, **c** *S. moellendorffii*, **d** *P. abies*, **e** *A. trichopoda*, **f** *O. sativa*, **g** *N. nucifera*, and **h** *A. thaliana*. The proportion of gene pairs conserved and divergent in expression, respectively, was indicated by different colors. **i**–**k** The expression divergence between duplicate genes derived from genome duplication events of different ages in model plants *A. thaliana* (eudicot) and *O. sativa* (monocot). **l**–**n** Expression divergence between duplicate genes in eudicot and monocot plants: **l** *Zea mays*, *Sorghum bicolor*, and *Ananas comosus*; **n** *Camelina sativa*, *Brassica oleracea*, and *Vitis vinifera*. **j**, **m** The phylogeny of different species and genome duplication or triplication events occurring in different branches were labeled. Significant differences (Fisher's exact test): $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{n.s.}P > 0.05$

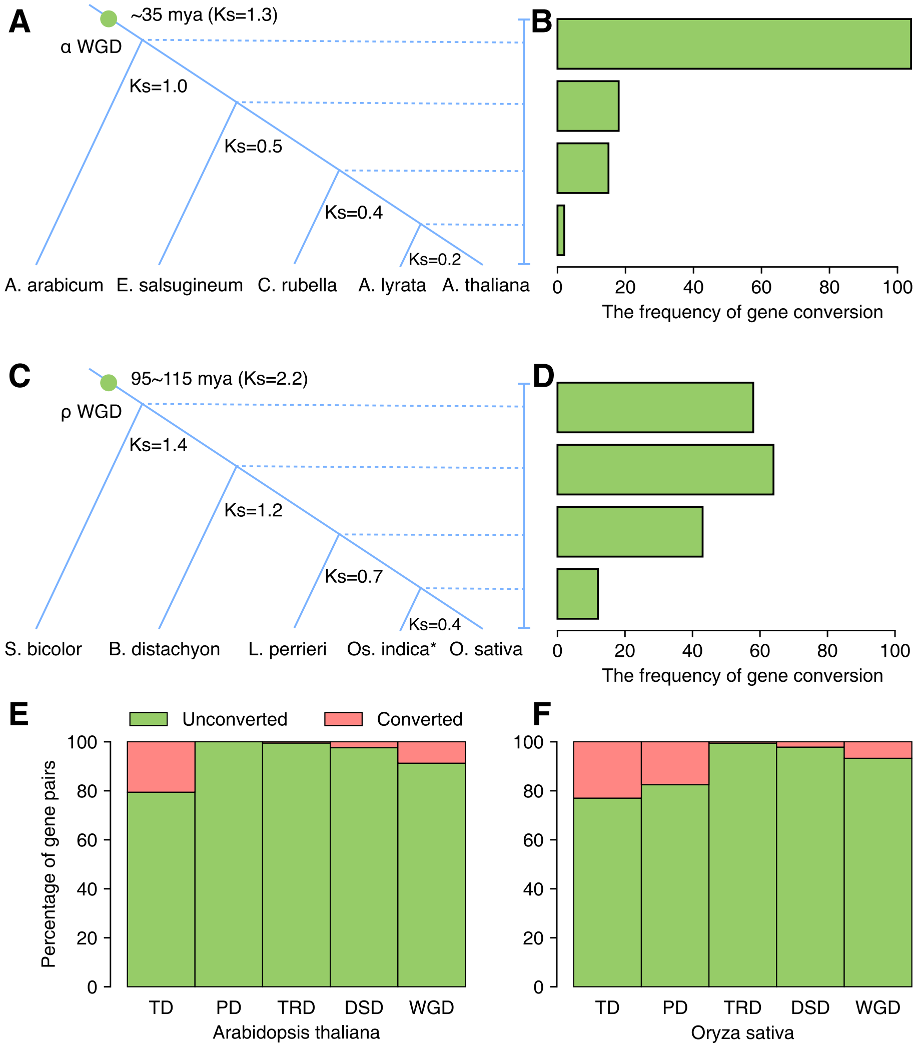

Fig. 7 Factors affecting gene conversion rates in plants. **a**, **b** Gene conversion rates after divergence between the *A. thaliana* lineage and those of *Aethionema arabicum*, *Eutrema salsugineum*, *Capsella rubella*, or *A. lyrata*, respectively. **c**, **d** Gene conversion rates after divergence between the *O. sativa* L. (ssp. japonica) lineage and those of *Sorghum bicolor*, *Brachypodium distachyon*, *Leersia perrieri*, or *O. sativa* L. (ssp. indica), respectively. **e**, **f** The frequency of gene conversion events for different modes of duplicated gene pairs in model plants *Arabidopsis* (**e**) and rice (**f**). **O. sativa* L. (ssp. indica). WGD whole-genome duplication, TD tandem duplication, PD proximal duplication, TRD transposed duplication, DSD dispersed duplication

those of *Sorghum bicolor*, *Brachypodium distachyon*, *Leersia perrieri*, or *Oryza sativa* L. (ssp. indica) were examined, using high-confidence Poaceae ρ WGD-derived gene pairs from *O. sativa*-japonica [54]. The divergence between *O. sativa*-japonica lineage and *S. bicolor*, *B. distachyon*, *L. perrieri*, or *O. sativa*-indica were respectively dated to $K_s = 1.4$, 1.2, 0.7, and 0.4. The rate of gene conversion events among ρ WGD-derived gene pairs decelerated over time compared with shortly after WGD. The number of gene conversion events is 58 after the divergence of *O. sativa*-japonica and *S. bicolor*, about fivefold higher than the number after the (much more recent) divergence of *O. sativa*-japonica and *O. sativa*-indica (Fig. 7c, d).

The proportion of tandem or proximal gene pairs experiencing gene conversion is more than that for other modes of gene duplication for model plants *Arabidopsis* and rice (Fig. 7e, f). In *Arabidopsis*, the percentage of converted TD-, PD-, TRD-, DSD-, and WGD-pairs is 20.6%, 0.0%, 0.5%, 2.4%, and 8.8% respectively. In rice, the percentage of converted TD-, PD-, TRD-, DSD-, and WGD-pairs is 23.0%, 17.5%, 0.6%, 2.2%, and 6.7% respectively. Rare gene conversion events were found in TRD-derived gene pairs, consistent with extensive sequence and expression divergence between TRD-duplicated genes.

Inferring core gene families from 141 green plant genomes

The whole-genome protein sequences of 141 green plants containing 4,921,214 genes were used to construct core gene families by using OrthoFinder [55]. Large-scale BLASTP searches were carried out for each pair of 141 species. We identified 86,831 gene families (or orthologous groups) (freely available at figShare, https://doi.org/10.6084/m9.figshare.7264667.v1), including 4,333,638 (88.1%) genes, 6266 (18,889 genes, 0.4% of all genes) species-specific families, and 232 most conserved families (Additional file 7) in which all species have at least one gene. We found no strict single-copy gene families for these 141 species, which may be due to errors in genome annotation, frequent single-gene duplication, or pseudogenization. We further identified the most-preserved, intermediate-preserved, and least-preserved gene families in 141 plants. The most-preserved plant gene families are those orthologous groups in which all species must have at least one gene. The intermediate-preserved gene families are those orthologous groups in which the absence (or missing) of orthologous genes in up to three species was allowed. The least-preserved gene families are those orthologous groups in which the absence of orthologous genes in up to five species was allowed. Functional enrichment analysis for most-preserved, intermediate-preserved, and least-preserved plant gene families using *Arabidopsis* genes as a reference revealed that these genes were collectively enriched in GO terms involved in "membrane" and "organelle" such as plasma membrane, organelle part, nucleus, membrane–bounded organelle, intracellular organelle, and cytoplasmic part (Additional file 1: Figure S19 and Additional file 8). In addition, the enriched GO terms are also collectively involved in small GTPase-mediated signal transduction, nucleosome, cytoskeletal, light-harvesting complex, ATPase activity, actin filament-based movement.

We further assigned the genes in each orthogroup into each single species and acquired the repertoire of gene families for each species (freely available on FigShare, DOI: https://doi.org/10.6084/m9.figshare.7264667). For each species, we calculated the percentage of gene families of a given size with respect to the total number of all gene families in this species, then investigated the distribution of gene family size in all plants (Fig. 8). A large percentage of small gene families (one to three members) were observed across all plants, showing a strong bias toward single-copy status. In green algae, the majority of gene families contained only one gene. For example, in *C. reinhardtii*, the proportion of single-gene families is 81.4%. The highest proportion (95.8%) of single-copy gene families was found in the marine angiosperm *Zostera marina*, forming a sharp contrast with closely related *Z. muelleri* with only 20.6% single-copy gene families. Species influenced by recent WGD or WGT, such as soybean (*Glycine max*), apple (*Malus domestica*), flax (*L. usitatissimum*), banana (*Musa acuminata*), and maize (*Z. mays*), possess more gene families of moderate number than other plants.

Discussion

Classification and comparison of the five major types of gene duplication in 141 plant genomes affected by a diverse set of whole-genome multiplications spanning more than 100 million years provides new insight into genome evolution and biological innovation. Whole-genome duplication increases all genes in a genome in a balanced manner that may favor modification of entire pathways and processes [56] and is associated with longer half-lives of the resulting gene duplicates [27]. However, it is unclear whether these advantages outweigh the relatively constant availability of new tandem and proximal duplicates that may be important for plants to adapt to dramatic environmental changes [45, 57–60]. The C4 photosynthetic pathway, thought to have been an adaptation to hot, dry environments or CO_2 deficiency [61–64] and independently appearing at least 50 times during angiosperm evolution [65, 66], includes some elements resulting from WGD and others from single-gene duplication, despite that all were in principle available from WGD in a cereal common ancestor [33]. Indeed, we found that the K_s peaks for WGD, transposed, and dispersed duplicates commonly overlapped in the same plant, suggesting that whole-genome duplication was also accompanied by extensive transposed and dispersed gene duplication, consistent with a recent study showing extensive relocation of γ duplicates shortly after the γ WGT event in core eudicots [48].

Different classes of gene duplicates showed distinct patterns of temporal and functional evolution. WGD-derived duplicates are more conserved with smaller K_a/K_s ratios than tandem and proximal duplicates, suggesting that they have experienced long-term purifying selection. Proximal and tandem duplicates preserved in modern genomes, with relatively high K_a/K_s ratios but relatively small K_s values per se, appear to experience more rapid functional divergence than other gene classes—supporting that positive selection plays an important role in the early stage of duplicate gene retention [67–69]. While concerted evolution may preserve homogeneity of tandem or proximal duplicates to a greater degree than genes that are distant from one another, this is not incompatible with rapid functional divergence [38].

Paralleling sequence divergence, expression divergence of duplicated genes gradually increases with age. Transposed duplicates preserved in modern genomes have high percentage of expression divergence in nearly all investigated species; this is consistent with both their antiquity and the nature of their evolution, with novel copies potentially being separated from *cis*-regulatory sequences at the original site and/or exposed to different ones at the new site. Environmental factors may accelerate expression divergence between duplicate genes [70], and frequent occurrence of transposed duplication may be important for plants to adapt to dramatic environmental changes [45, 57–60]. Physically linked (or

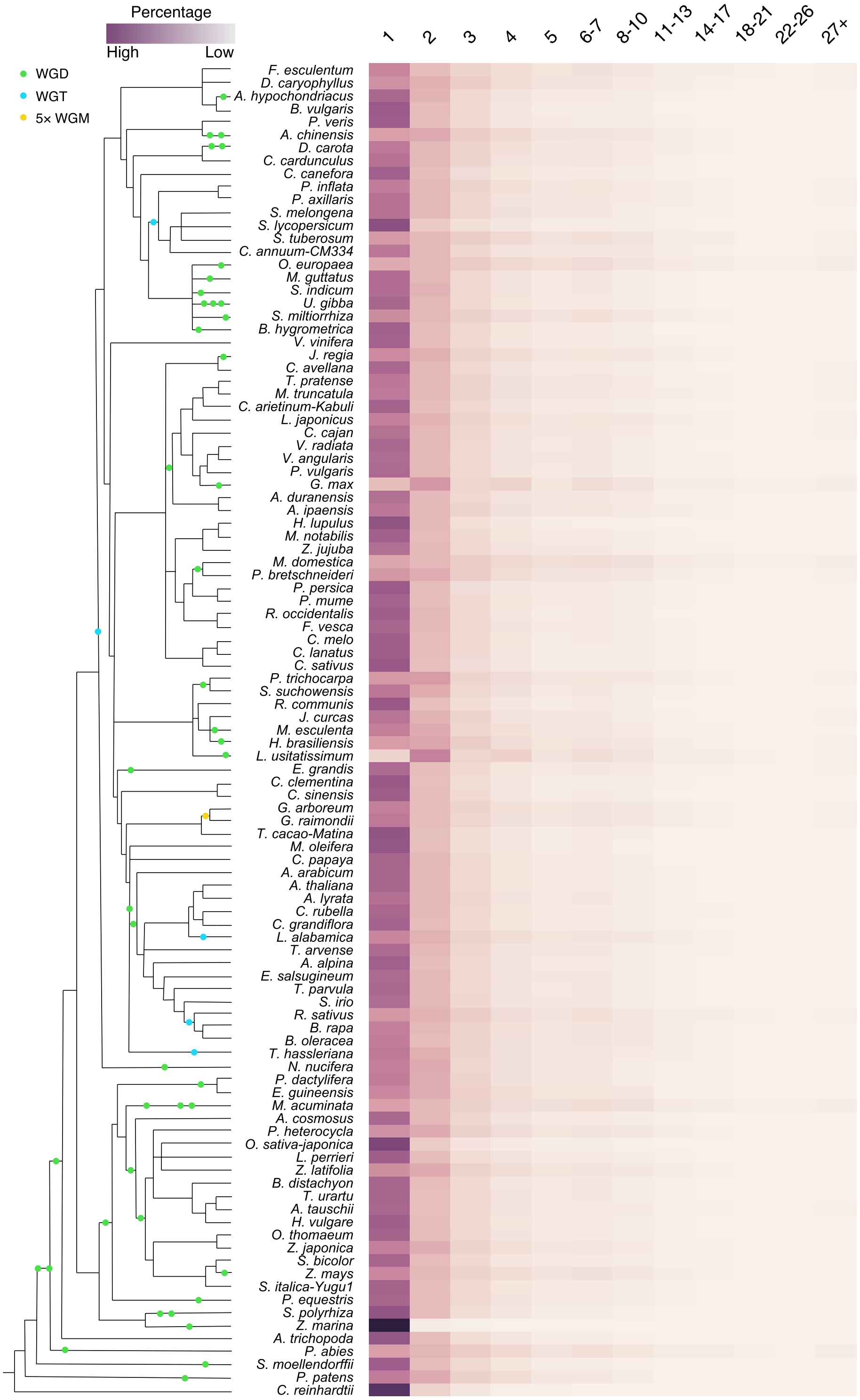

Fig. 8 The distribution of gene family sizes across 141 plants. For each species, we calculated the percentage of gene families of a given size with respect to the total number of all gene families in this species. The top labels of the *x* axis indicate the gradient of different sizes of gene family

tandem) duplications show generally less expression divergence than distant duplications, a result supported by many prior studies, e.g., [43, 44, 71–74]. Indeed, physically linked genes in the same paralogon (or syntenic block) are preferentially retained in *cis*-PPIs (protein–protein interactions) after WGD [75, 76].

Two types of subfunctionalization (SF) have been proposed [77–80]. One type of subfunctionalization takes place by complementary coding sequence changes between duplicated genes, leading to their functional divergence at the protein level, and eventually resulting in division of multiple functions of the progenitor gene. However, divergence at the biochemical level between two copies is limited even over long evolutionary times. The other type of subfunctionalization occurs by complementary loss or degenerative mutation of *cis*-regulatory elements between duplicated genes, creating inter-dependence between partially degenerated copies to maintain the full expression profiles of the ancestral gene in different tissues and/or conditions (defined as expression subfunctionalization (ESF)) [43, 78]. Many previous studies revealed that expression divergence between duplicate genes often occurred quickly after gene duplication [47, 81–84]. In this study, widespread divergence between expression profiles of duplicated genes was found in different modes of gene duplication—this can be largely explained by the expression subfunctionalization (or subfunctionalization) models, under which two duplicate genes evolved toward the partitioning of ancestral gene expression profiles in different tissues or conditions. The expression neofunctionalization (ENF) hypothesis, that one of the two gene copies gains a new *cis*-regulatory element in its promoter region and expresses in a new tissue, could also result in divergent expression profiles between duplicated genes such as some observed in this study [43, 85].

Among the earliest changes following polyploidization is gene conversion, nonreciprocal recombination between alleles or paralogous loci which homogenizes paralogous sequences or even chromosomal regions [86–89]. Gene conversion appears to occur virtually immediately in synthetic polyploid *Arachis* (peanut) [90]—indeed, abundant gene conversion after hybridization or polyploidization plays an important role in maintaining genome stability in plants and fungi [5, 18, 91, 92]. We detected relatively abundant gene conversion events in TD-, PD-, and WGD-pairs, which may be associated with their reduced expression divergence. The TRD- and DSD-pairs may have escaped the constraints induced by gene conversion. The dynamic changes of gene conversion rate found in this study, being high shortly after polyploidization and declining over time, show that prior findings over about 1 MY of cotton evolution [53] are generally applicable to a wide range of taxa and polyploidization events. The extensive gene conversion events occurring immediately after gene or genome duplication homogenize paralogs for a period of time and maintain a higher probability of functional compensation between duplicated genes, buffering the phenotypic effect caused by loss of one of two members of a duplicated pair [93–95]. Evolutionary divergence between duplicate genes may be suppressed by extensive gene conversion events during the early stage of genome duplication; however, this is not incompatible with rapid functional divergence of the TD- or PD-derived gene pair [96].

Conclusions

The sharp increase in the number of sequenced plant genomes has empowered investigation of key aspects of evolution by application of uniform techniques to taxa spanning hundreds of million years of divergence, including model and non-model, crop and non-crop, flowering and non-flowering, seed and non-seed, vascular and non-vascular, and unicellular and multicellular species. Building on many studies of individual genomes, the comprehensive landscape of different modes of gene duplication identified across the plant kingdom by virtue of the ability to compare 141 genomes provides a solid foundation for further investigating the dynamic evolution and divergence of duplicate genes and for validating evolutionary models underlying duplicate gene retention. The contributions of gene duplication to gene regulatory networks, epigenetic variation, morphological complexity, and adaptive evolution are intriguing subjects for further investigation by this approach.

Methods

Collecting genome datasets

In this study, the genome datasets of 141 plants were downloaded from multiple comprehensive databases such as Phytozome (v11), NCBI, Ensembl Plants, and many other individual genome databases. These 141 plant genomes sample diverse taxa ranging from unicellular green alga (Chlorophytes) to Bryophytes, Lycophytes, gymnosperms, and angiosperms. The detailed information of these 141 species and their data sources can be retrieved in Additional file 2. Only the transcript with the longest CDS was selected for further analysis when several transcripts were available for the same gene.

Identifying gene duplications

The different modes of gene duplication were identified using the *DupGen_finder* pipeline (https://github.com/qiao-xin/DupGen_finder). Firstly, the all-versus-all local BLASTP was performed using protein sequences ($E < 1e^{-10}$, top 5 matches and m8 format output) to search all potential homologous gene pairs within each genome. Secondly, the *MCScanX* algorithm [97] was utilized to identify the WGD-derived gene pairs. Then, we excluded these WGD-pairs from the whole set of homologous pairs (or BLASTP hits) to further determine the single-gene duplications. If the two genes in a BLASTP hit that are adjacent to each other on the same chromosome, they were defined as tandem gene pair. Proximal gene pairs were defined as non-tandem pairs separated by 10 or fewer genes on the same chromosome. To identify transposed duplications, WGD, tandem, and proximal gene pairs were deducted from the whole set of homologous gene pairs. A transposed duplicate pair was required to meet the following criteria: one gene existed in its ancestral locus (named the parent copy) and the other was located in a non-ancestral locus (transposed copy). Two types of genes can be regarded as ancestral loci: (i) intra-species colinear genes and (ii) inter-species colinear genes. The intra-species colinear genes can be obtained from WGD-derived gene pairs, which have been identified above. Inter-species colinear genes were discerned by intergenomic synteny analysis, executing *MCScanX* on inter-species BLASTP files between the target and outgroup genomes. The sacred lotus (*Nelumbo nucifera*) and *Spirodela polyrhiza* were respectively taken as outgroup for all eudicot plants and all monocot plants to identify ancestral syntenic blocks. *Amborella trichopoda* was adopted as outgroup for *N. nucifera* and *S. polyrhiza* to find ancestral syntenic blocks. Genes located in these conserved syntenic blocks were deemed to be ancestral loci. The rarity of syntenic blocks between green algae (Chlorophytes), Bryophytes, Lycophytes, and other plants hindered the identification of ancestral loci in these species by applying inter-species synteny analysis. Therefore, we constructed orthologous relationships among genes of these species with large evolutionary distances to deduce the conserved ancestral genes. To identify the ancestral loci in *P. patens* (a Bryophyte) and *S. moellendorffii* (a Lycophyte), OrthoFinder [55] and whole-genome protein sequences were used to infer orthogroups among these two species and five other species: *P. abies*, *S. polyrhiza*, *N. nucifera*, *Amborella trichopoda*, and *Arabidopsis thaliana*. Based on the above orthogroups, if a gene in *P. patens* or *S. moellendorffii* has an ortholog pair in at least two other lineages, it is considered ancient and likely to have been present in the common ancestor of land plants. Similarly, we built the orthogroups among eight green algae species to determine the ancestral loci within each green algae genome. Based on the above steps, BLASTP hits to both an ancestral and a novel locus were defined as transposed duplications. Finally, after removing WGD, tandem, proximal, and transposed duplications from the whole set of homologous gene pairs, the remaining gene pairs were classified as dispersed duplications. Noting that the same dispersed gene may have several BLASTP hits resulting in multiple gene pairs for one gene, we only considered the dispersed gene pairs with highest similarity in this situation.

For gymnosperm species, we applied an alternative method to infer gene duplications. In this study, we initially selected two reference gymnosperm species: *Picea abies* and *Pinus taeda*, both belonging to Pinaceae. However, no or few syntenic or colinear blocks could be detected within these two genomes due to the fragmented assembly; thus, we used an alternative strategy to find potential duplicate gene pairs derived from WGDs. A recent study suggested that Pinaceae lineages had experienced one ancient WGD shared with other seed plants corresponding to a K_s peak with a median K_s = 0.75 to 1.5 and one younger WGD in a Pinaceae ancestor corresponding to K_s peak with a median K_s = 0.2 to 0.4 [98]. According to the above results, we firstly selected duplicate gene pairs corresponding to these two putative WGD peaks in the K_s age distribution from all-blast-all output. Furthermore, we identified orthogroups among genes from *P. abies*, *P. taeda*, and three other Pinaceae species (*Pinus lambertiana*, *Pseudotsuga menziesii*, and *Picea glauca*) by using the OrthoFinder software [55], which utilize a novel method to infer orthogroups of protein coding genes and is suitable for orthogroup inference from incomplete genome assemblies. Based on the above two steps, if each gene of a duplicate pair from the aforementioned two K_s peaks in *P. abies* or *P. taeda* has an ortholog pair in at least two other lineages, we assumed that this duplicate pair was created by WGDs in a common Pinaceae ancestor rather than independently in each lineage. By using the same rules applied in other plants, the tandem and proximal gene pairs were identified in *P. abies* or *P. taeda*. Based on orthogroups among five gymnosperm species, if a gene in *P. abies* or *P. taeda* has an ortholog pair in at least two other lineages, it is considered ancient and likely to have been present in the common ancestor of Pinaceae species. Then, we determined the transposed gene pairs comprised of an ancestral and a novel locus after excluding the WGD, tandem, and proximal gene pairs from the population of BLASTP hits. At last, the remaining gene pairs after removing other modes of gene duplications from BLASTP hits were classified as dispersed gene pairs.

Calculating K_a, K_s, and K_a/K_s values

For each duplicate gene pair, we aligned their protein sequences using MAFFT (v7.402) [99] with the L-INS-i option and converted the protein alignment into a codon alignment using PAL2NAL [100]. Then, the resulting codon alignment was formatted into an AXT format using a custom Perl script. γ-MYN method (a modified version of the Yang–Nielsen method) [101, 102] incorporated in KaKs_Calculator 2.0 [103] was used to calculate K_a and K_s values by implementing the Tamura–Nei model [104]. The K_s values > 5.0 were excluded from further analysis due to the saturated substitutions at synonymous sites [105, 106]. The pipeline used to calculate

K_a and K_s values is freely available on GitHub (https://github.com/qiao-xin/Scripts_for_GB).

RNA-seq data and quantification

Single-end or paired-end RNA-seq reads were downloaded from NCBI SRA (https://www.ncbi.nlm.nih.gov/sra). The RNA-seq samples used in this study were documented in Additional file 6. The raw reads were filtered using Trimmomatic (version 0.36) (http://www.usadellab.org/cms/?page=trimmomatic). We filtered the raw reads according to the following procedure: (1) removing adapters (pair-end: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 and single-end: ILLUMINACLIP:TruSeq3-SE:2:30:10); (2) removing leading low quality or N bases (below quality 15) (LEADING:15); (3) removing trailing low quality or N bases (below quality 15) (TRAILING:15); (4) scanning the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15); and (5) dropping reads below 55 or 36 bases long (pair-end: MINLEN:55 and single-end: MINLEN:36). Next, the abundances of transcripts from RNA-Seq data were estimated using kallisto [107]. The reference transcripts obtained from genome annotation files were used to build kallisto indices. Then, the kallisto quantification algorithm was performed with default parameters (for single-ends, -l 200 -s 20) to process either single-end or paired-end reads, outputting the normalized count estimates and TPM (transcripts per million) values for each transcript. The TPM value was used as the measure of expression level of the genes in different tissues and conditions.

We further extracted all intergenic regions at the whole-genome level for investigated species and quantified their expression abundances using the same procedure and RNA-seq reads used for exonic regions. The medians of the distributions of TPM values for intergenic sequences in different tissues and conditions are close to 0. Therefore, we used the mean value of the medians (the 50th percentile) obtained from the TPM distributions for intergenic sequences in different tissues and conditions as the threshold of expression (Additional file 1: Figure S20 and S21).

Estimating expression divergence

Duplicated gene pairs in which both gene copies were expressed in at least one tissue or development stage were used to calculate Pearson correlation coefficients (r) between expression profiles of the two gene copies. The two genes in a random pair should have unrelated function and differential expression, so we can determine the cutoff for divergent expression by comparing distributions of r values for random gene pairs to those for duplicated gene pairs. We randomly selected 10,000 gene pairs from each species and computed r values between their expression profiles. We determined a cutoff from the distribution of r values for random gene pairs in each species and required that 95% of the r values obtained from the distribution be less than this cutoff value. The duplicated gene pairs with r less than this cutoff can be considered to have diverged in expression (Additional file 1: Figure S15 and S16).

Detecting gene conversion

The method used to detect gene conversion is as described in former studies [89, 108, 109]. Firstly, we identified homologous gene quartets, comprised of two paralogs in the species of interest and their respective orthologs in outgroup species. For *Arabidopsis*, *Aethionema arabicum*, *Eutrema salsugineum*, *Capsella rubella*, and *Arabidopsis lyrata* were used as outgroup species. For rice (*O. sativa* L. (ssp. japonica)), *Sorghum bicolor*, *Brachypodium distachyon*, *Leersia perrieri*, and *Oryza sativa* L. (ssp. indica) were used as outgroup species. The number of homologous gene quartets identified between *Arabidopsis* and the four outgroup species *A. arabicum*, *E. salsugineum*, *C. rubella*, and *A. lyrata* are 615, 1165, 1355, and 788 respectively. The number of homologous gene quartets identified between rice and the four outgroup species *S. bicolor*, *B. distachyon*, *L. perrieri*, and *O. sativa*-indica are 761, 718, 917, and 1140. To identify the gene conversion events in different modes of duplicated gene pairs, we chose *A. arabicum* and *S. bicolor* as outgroups for *Arabidopsis* and rice respectively to determine homologous gene quartets. The frequency of gene conversion events for different modes of duplicated gene pairs was determined in model plants *Arabidopsis* and rice. Then, we compared gene similarity or tree topology between homologs in quartets by estimating synonymous nucleotide substitution rates (K_s) between them. We performed a bootstrap test to evaluate the significance of putative gene conversions with 1000 repetitive samplings to produce a bootstrap frequency indicating the confidence level of the supposed conversion [89, 108]. The pipelines used to identify homologous gene quartets and detect gene conversion are available on GitHub (https://github.com/qiao-xin/Scripts_for_GB/tree/master/detect_gene_conversion).
All homologous gene quartets identified in this study have been deposited on FigShare (https://doi.org/10.6084/m9.figshare.7264667.v1).

Inferring the orthogroups of 141 green plants

The OrthoFinder [55] algorithm was utilized to construct the orthogroups for the 141 plants. It has been demonstrated that the OrthoFinder is more accurate and faster than other commonly used orthogroup inference methods such as OrthoMCL [55, 110]. To run OrthoFinder with pre-computed BLAST results, we performed all-vs-all BLASTP searches ($E < 1e^{-10}$, top 5 matches and m8 format output) for each pairwise genome comparison between species and self-genome comparisons by using protein sequences. Then, we ran OrthoFinder with default parameters using the BLASTP outputs as inputs and obtained a file containing the orthologous groups (or gene families) of genes from these 141 species. Furthermore, we assigned the genes in each orthogroup into each single species and acquired the repertoire of gene families for each species (freely available on FigShare, https://doi.org/10.6084/m9.figshare.7264667.v1). We then investigated the distribution of gene family size in all studied plants.

Gene ontology enrichment analysis

Because the members of a gene family have similar functions, we only conducted the functional enrichment analysis for the *Arabidopsis* gene sets from the most-conserved, intermediate-conserved, and least-conserved gene families (or orthogroups) in 141 plants. Firstly, we retrieved all *Arabidopsis* genes from most-preserved, intermediate-preserved, and least-preserved gene families respectively. GO annotations for the genes in *Arabidopsis* were downloaded from Phytozome11 (https://phytozome.jgi.doe.gov/pz/portal.html). Furthermore, we detected the overrepresented GO slim terms in these *Arabidopsis* genes by using the GOATOOLS package [111]. The *P* values used to evaluate significant enrichment of certain GO terms were calculated based on Fisher's exact test and corrected by an FDR test correction method (false discovery rate implementation using resampling). Finally, we used corrected *P* value < 0.01 as the threshold to determine significant overrepresentation of certain GO terms.

Additional files

Additional file 1: Supplementary Table S1 and Figure S1-21. (DOCX 5373 kb)

Additional file 2: The detailed information of 141 plant species used in this study. (XLSX 76 kb)

Additional file 3: The absolute number of different modes of duplicate gene pairs in each taxon. (XLSX 49 kb)

Additional file 4: The fitted K_s peak for WGD genes in each species. (XLSX 62 kb)

Additional file 5: The enriched GO terms for tandem and proximal duplicate genes in *Arabidopsis thaliana*. (XLSX 69 kb)

Additional file 6: The list of all RNA-Seq samples collected from different plants investigated in this study. (XLSX 105 kb)

Additional file 7: The 232 most conserved gene families in 141 plants. (XLSX 1968 kb)

Additional file 8: The enriched GO terms for the most-preserved, intermediate-preserved and least-preserved gene families in 141 plants. (XLSX 1018 kb)

Acknowledgements

We thank Barry Marler and Leiting Li for IT support. We thank Xiyin Wang for providing valuable script to detect gene conversion events. We also thank Maosong Pei and Weiwei Zeng for their help in data visualization.

Funding

This work was funded by the National Key Research and Development Program of China (2018YFD1000107), Key Program of National Natural Science Foundation of China (31830081), "Taishan Scholar" project from Shandong Province of China, the Earmarked Fund for China Agriculture Research System (CARS-28), Jiangsu Province Science and Technology Support Program (BE2018389), the US National Science Foundation to AHP and XW (MCB-1021718; 1339727), and the China Scholarship Council.

Availability of data and materials

All accession numbers and URLs for raw data used in this study are provided in Additional file 2 [112–227]. The *DupGen_finder* pipeline is freely available on GitHub (https://github.com/qiao-xin/DupGen_finder) [228]. The different modes of duplicate gene pairs identified in 141 plant genomes are available at Plant Duplicate Gene Database (PlantDGD, http://pdgd.njau.edu.cn:8080) [229]. The orthologous groups inferred by OrthoFinder using all genes from the 141 plants are available for download from FigShare (https://doi.org/10.6084/m9.figshare.7264667.v1) [230]. The custom scripts used in this study are freely available on GitHub (https://github.com/qiao-xin/Scripts_for_GB) [228].

Authors' contributions

AHP, SZ, and XQ conceived and designed the experiments. XQ performed the experiments. XQ and AHP analyzed the data. LL and RW contributed analysis tools/materials/Perl scripts. QL, HY, and KQ assisted in the data analysis. XQ, SZ, and AHP wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors are aware of the content and agree with the submission.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 18 May 2018 Accepted: 8 February 2019
Published online: 21 February 2019

References

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. Life with 6000 genes. Science. 1996; 274:546–67.
2. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 1997;387:708–13.
3. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006;444:171–8.
4. McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M. Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics. 2014;197:1417–28.
5. McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome research. 2014;24:1665–75.
6. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 2005;3:e314.
7. Nakatani Y, Takeda H, Kohara Y, Morishita S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res. 2007;17:1254–65.
8. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet. 2016;48:427–37.
9. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci. 2010; 107:9270–4.

10. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A. 2004;101:1638–43.
11. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007;447:714–9.
12. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:3657.
13. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200.
14. Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, Xu J, Zheng X, Ren L, Wang G, et al. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. Nat Genet. 2014;46:1212–9.
15. Li JT, Hou GY, Kong XF, Li CY, Zeng JM, Li HD, Xiao GB, Li XM, Sun XW. The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (*Cyprinus carpio*). Sci Rep. 2015;5:8199.
16. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473:97–100.
17. Albert VA, Barbazuk WB, Der JP, Leebens-Mack J, Ma H, Palmer JD, Rounsley S, Sankoff D, Schuster SC, Soltis DE. The *Amborella* genome and the evolution of flowering plants. Science. 2013;342:1241089.
18. Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science. 2014;345:950–3.
19. Jiao Y, Li J, Tang H, Paterson AH. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell. 2014;26:2792–802.
20. Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 2003;422:433–8.
21. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, Zhao M, Ma J, Yu J, Huang S, et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun. 2014;5:3930.
22. Tomato Genome C. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485:635–41.
23. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ. A genome triplication associated with early diversification of the core eudicots. Genome Biol. 2012;13:R3.
24. Wang X, Guo H, Wang J, Lei T, Liu T, Wang Z, Li Y, Lee TH, Li J, Tang H, et al. Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. New Phytol. 2016;209:1252–63.
25. Paterson AH, Freeling M, Tang H, Wang X. Insights from the comparison of plant genome sequences. Ann Rev Plant Biol. 2010;61:349–72.
26. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS: Polyploidy and angiosperm diversification. Am J Botany 2009, 96:336–348.
27. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290:1151–5.
28. Jiao Y, Paterson AH. Polyploidy-associated genome modifications during land plant evolution. Philos Trans R Soc Lond B Biol Sci. 2014;369:20130355.
29. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016;17:37.
30. Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in plants. Curr Opin Genet Dev. 2015;35:119–25.
31. Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Juárez MJA, Simpson J. Architecture and evolution of a minute plant genome. Nature. 2013;498:94–8.
32. Cuevas HE, Zhou C, Tang H, Khadke PP, Das SK, Lin YR, Ge Z, Clemente T, Upadhyaya HD, Hash CT, Paterson AH. The evolution of photoperiod-insensitive flowering in sorghum, a genomic model for panicoid grasses. Mol Biol Evol. 2016;33:2417–28.
33. Wang X, Gowik U, Tang H, Bowers JE, Westhoff P, Paterson AH. Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. Genome Biol. 2009;10:R68.
34. Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431:569–73.
35. Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res. 2005;15:1292–7.
36. 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.
37. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of gene duplication in plants. Plant Physiol. 2016;171:2294–316.
38. Wang Y, Wang X, Paterson AH. Genome and gene duplications and gene expression divergence: a view from plants. Ann N Y Acad Sci. 2012;1256:1–14.
39. 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 Res. 2008;18:1924–37.
40. Woodhouse MR, Tang HB, Freeling M. Different gene families in *Arabidopsis thaliana* transposed in different epochs and at different frequencies throughout the rosids. Plant Cell. 2011;23:4241–53.
41. Woodhouse MR, Pedersen B, Freeling M. Transposed genes in Arabidopsis are often associated with flanking repeats. PLoS Genet. 2010;6:e1000949.
42. X-p Z, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson AH. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res. 1998;8:479–92.
43. Lan X, Pritchard JK. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science. 2016;352:1009–13.
44. Dai Z, Xiong Y, Dai X. Neighboring genes show interchromosomal colocalization after their separation. Mol Biol Evol. 2014;31:1166–72.
45. Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 2008;148:993–1003.
46. Cusack BP, Wolfe KH. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol. 2007;24:679–86.
47. Ganko EW, Meyers BC, Vision TJ. Divergence in expression between duplicated genes in Arabidopsis. Mol Biol Evol. 2007;24:2298–309.
48. Wang Y, Ficklin SP, Wang X, Feltus FA, Paterson AH. Large-scale gene relocations following an ancient genome triplication associated with the diversification of core eudicots. PLoS One. 2016;11:e0155637.
49. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang M-L, Chen J, Biggers E. The pineapple genome and the evolution of CAM photosynthesis. Nat Genet. 2015;47:1435–42.
50. Wang X, Wang J, Jin D, Guo H, Lee TH, Liu T, Paterson AH. Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. Mol Plant. 2015;8:885–98.
51. 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 Res. 2014;24:1334–47.
52. Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, Spillane C, Robinson SJ, Links MG, Clarke C, et al. The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. Nat Commun. 2014;5:3706.
53. Guo H, Wang X, Gundlach H, Mayer KF, Peterson DG, Scheffler BE, Chee PW, Paterson AH: Extensive and biased intergenomic non-reciprocal DNA exchanges shaped a nascent polyploid genome, Gossypium (cotton). Genetics 2014:genetics. 114.166124.
54. Tang HB, Bowers JE, Wang XY, Paterson AH. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci U S A. 2010;107:472–7.
55. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
56. Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell. 2007;19:395–402.
57. Xu G, Ma H, Nei M, Kong H. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc Natl Acad Sci U S A. 2009;106:835–40.
58. Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, Yun DJ, Bressan RA, Zhu JK, Bohnert HJ, Cheeseman JM. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet. 2011;43:913–8.
59. Woodhouse M, Freeling M. Tandem duplications and gene transposition in plants. Maydica. 2009;54:463.

60. Tamate SC, Kawata M, Makino T. Contribution of nonohnologous duplicated genes to high habitat variability in mammals. Mol Biol Evol. 2014;31:1779–86.
61. Seemann JR, Sharkey TD, Wang J, Osmond CB. Environmental effects on photosynthesis, nitrogen-use efficiency, and metabolite pools in leaves of sun and shade plants. Plant Physiol. 1987;84:796–802.
62. Hattersley PG. The distribution of C3 and C4 grasses in Australia in relation to climate. Oecologia. 1983;57:113–28.
63. Ehleringer JR, Bjorkman O. A comparison of photosynthetic characteristics of Encelia species possessing glabrous and pubescent leaves. Plant Physiol. 1978;62:185–90.
64. Cerling TE, Harris JM, MacFadden BJ, Leasey MG, Quade J, Eisenmann V, Ehleringer JR. Global vegetation change throught the Miocene/Pliocene boundary. Nature. 1997;389:153–8.
65. Sage RF. The evolution of C4 photosynthesis. New Phytologist. 2004;161: 341–70.
66. Mulhaidat R, Sage RF, Dengler NG. Diversity of kranz anatomy and biochemistry in C4 eudicots. Am J Botany. 2007;94:20.
67. Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH. Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci U S A. 2006;103:2232–6.
68. Ren LL, Liu YJ, Liu HJ, Qian TT, Qi LW, Wang XR, Zeng QY. Subcellular relocalization and positive selection play key roles in the retention of duplicate genes of populus class III peroxidase family. Plant Cell. 2014;26: 2404–19.
69. Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. Evidence for the fixation of gene duplications by positive selection in Drosophila. Genome Res. 2016;26:787–98.
70. Ha M, Li WH, Chen ZJ. External factors accelerate expression divergence between duplicate genes. Trends Genet. 2007;23:162–6.
71. Sémon M, Duret L. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. 2006;23:1715–23.
72. Lou XL, Han B. Evolutionary conservation of neighbouring gene pairs in plants. Gene. 2009;437:71–9.
73. Ghanbarian AT, Hurst LD. Neighboring genes show correlated evolution in gene expression. Mol Biol Evol. 2015;32:1748–66.
74. Guang-Zhong Wang W-HC, Martin J. Lercher: coexpression of linked gene pairs persists long after their separation. Genome Biol Evol. 2011;3:565.
75. Makino T, McLysaght A. Interacting gene clusters and the evolution of the vertebrate immune system. Mol Biol Evol. 2008;25:1855–62.
76. Makino T, McLysaght A. Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant. Genome Res. 2012;22: 2427–35.
77. Hahn MW. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered. 2009;100:605–17.
78. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003; 18:292–8.
79. Force A, Lynch M, Pickett FB, Amores A, Yan Y-L, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151:1531–45.
80. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154:459–73.
81. Gu Z, Nicolae D, Lu HH, Li W-H. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. 2002;18:609–13.
82. Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldon T. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Brief Bioinform. 2011;12:442–8.
83. Makova KD, Li W-H. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. 2003;13:1638–45.
84. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell. 2004;16:1679–91.
85. Ohno S. Evolution by gene duplication. Berlin Heidelberg: Springer; 1970.
86. Wang X-Y, Paterson AH. Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. Genes. 2011;2:1–20.
87. Fawcett JA, Innan H. Neutral and non-neutral evolution of duplicated genes with gene conversion. Genes (Basel). 2011;2:191–209.
88. Hurles M. Gene duplication: the genomic trade in spare parts. PLoS Biol. 2004;2:E206.
89. Wang XY, Tang HB, Bowers JE, Feltus FA, Paterson AH. Extensive concerted evolution of rice paralogs and the road to regaining independence. Genetics. 2007;177:1753–63.
90. Chen X, Li H, Pandey MK, Yang Q, Wang X, Garg V, Li H, Chi X, Doddamani D, Hong Y, et al. Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. Proc Natl Acad Sci U S A. 2016;113:6785–90.
91. Sriswasdi S, Takashima M, Manabe R, Ohkuma M, Sugita T, Iwasaki W. Global deceleration of gene evolution following recent genome hybridizations in fungi. Genome Res. 2016;26:1081–90.
92. Yang S, Yuan Y, Wang L, Li J, Wang W, Liu H, Chen J-Q, Hurst LD, Tian D. Great majority of recombination events in Arabidopsis are gene conversion events. Proc Natl Acad Sci. 2012;109:20992–7.
93. 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.
94. Hanada K, Sawada Y, Kuromori T, Klausnitzer R, Saito K, Toyoda T, Shinozaki K, Li WH, Hirai MY. Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. Mol Biol Evol. 2011; 28:377–82.
95. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. Role of duplicate genes in genetic robustness against null mutations. Nature. 2003;421:63.
96. Wang YP, Wang XY, Tang HB, Tan X, Ficklin SP, Feltus FA, Paterson AH. Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. Plos One. 2011;6:e28150.
97. Wang YP, Tang HB, DeBarry JD, Tan X, Li JP, Wang XY, Lee TH, Jin HZ, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49.
98. Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. Early genome duplications in conifers and other seed plants. Sci Adv. 2015;1:e1501084.
99. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772–80.
100. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609–12.
101. Wang DP, Wan HL, Zhang S, Yu J. Gamma-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. Biol Direct. 2009;4:20.
102. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000;17:32–43.
103. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics. 2010;8:77–80.
104. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10:512–26.
105. Vanneste K, Van de Peer Y, Maere S. Inference of genome duplications from age distributions revisited. Mol Biol Evol. 2013;30:177–90.
106. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell. 2016;28:326–44.
107. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
108. Wang XY, Tang HB, Bowers JE, Paterson AH. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. Genome Res. 2009;19:1026–32.
109. Wang XY, Tang HB, Paterson AH. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. Plant Cell. 2011;23:27–37.
110. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
111. Klopfenstein DV, Zhang L, Pedersen BS, Ramirez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, et al. GOATOOLS: a Python library for Gene Ontology analyses. Sci Rep. 2018;8:10872.
112. Clouse JW, Adhikary D, Page JT, Ramaraj T, Deyholos MK, Udall JA, Fairbanks DJ, Jellen EN, Maughan PJ. The amaranth genome: genome, transcriptome, and physical map assembly. Plant Genome. 2016;9:1–14.
113. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43:476–81.

114. Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, et al. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408:796–815.
115. International Brachypodium I. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010;463:763–8.
116. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS, Newman LK, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet. 2013;45:831–5.
117. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature. 2008;452:991–6.
118. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 2007;318:245–51.
119. Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J, et al. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol. 2014;32:656–62.
120. Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP, et al. The draft genome of sweet orange (*Citrus sinensis*). Nat Genet. 2013;45:59–66.
121. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biol. 2012;13:R39.
122. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. The genome of the cucumber, *Cucumis sativus* L. Nat Genet. 2009;41:1275–81.
123. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al. The genome of *Eucalyptus grandis*. Nature. 2014;510:356–62.
124. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, et al. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet. 2011;43:109–16.
125. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–83.
126. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature. 2012;492:423–7.
127. Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, et al. The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J. 2012;72:461–73.
128. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, et al. The genome of the domesticated apple (*Malus x domestica* Borkh.). Nat Genet. 2010;42:833–9.
129. Wang W, Feng B, Xiao J, Xia Z, Zhou X, Li P, Zhang W, Wang Y, Moller BL, Zhang P, et al. Cassava genome from a wild ancestor to cultivated varieties. Nat Commun. 2014;5:5110.
130. Young ND, Debelle F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, et al. The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature. 2011;480:520–4.
131. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science. 2009;324:268–72.
132. Hellsten U, Wright KM, Jenkins J, Shu SQ, Yuan YW, Wessler SR, Schmutz J, Willis JH, Rokhsar DS. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proc of the Natl Acad Sci U S A. 2013;110:19478–82.
133. Matsumoto T, Wu JZ, Kanamori H, Katayose Y, Fujisawa M, Namiki N, Mizuno H, Yamamoto K, Antonio BA, Baba T, et al. The map-based sequence of the rice genome. Nature. 2005;436:793–800.
134. Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A. 2007;104:7705–10.
135. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu SQ, Song QJ, Chavarro C, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics. 2014;46:707–13.
136. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, et al. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science. 2008;319:64–9.
137. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313:1596–604.
138. International Peach Genome I, Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 2013;45:487–94.
139. Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, et al. Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol. 2010;28:951–6.
140. Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, de Pamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science. 2011;332:960–3.
141. Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, et al. Reference genome sequence of the model plant Setaria. Nat Biotechnol. 2012;30:555–61.
142. Xu X, Pan SK, Cheng SF, Zhang B, Mu DS, Ni PX, Zhang GY, Yang S, Li RQ, Wang J, et al. Genome sequence and analysis of the tuber crop potato. Nature. 2011;475:189–U194.
143. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009;457:551–6.
144. Wang W, Haberer G, Gundlach H, Glasser C, Nussbaumer T, Luo MC, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J, et al. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun. 2014;5:3311.
145. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 2013;14:r53.
146. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–U465.
147. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science. 2010;329:223–6.
148. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–5.
149. Olsen JL, Rouze P, Verhelst B, Lin YC, Bayer T, Collen J, Dattolo E, De Paoli E, Dittami S, Maumus F, et al. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. Nature. 2016;530:331–5.
150. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, et al. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nature Biotechnol. 2011;29:521–U584.
151. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng CF, Alberti A, Anthony F, Aprea G, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 2014;345:1181–4.
152. Wu J, Wang ZW, Shi ZB, Zhang S, Ming R, Zhu SL, Khan MA, Tao ST, Korban SS, Wang H, et al. The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Research. 2013;23:396–408.
153. Chagne D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, Ireland H, Fiers M, Dzierzon H, Cestaro A, Fontana P, et al. The draft genome sequence of European pear (*Pyrus communis* L. 'Bartlett'). PLoS One. 2014;9:e92644.
154. Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, Tao Y, Wang J, Yuan Z, Fan G, et al. The genome of *Prunus mume*. Nat Commun. 2012;3:1318.
155. Huang S, Ding J, Deng D, Tang W, Sun H, Liu D, Zhang L, Niu X, Zhang X, Meng M, et al. Draft genome of the kiwifruit *Actinidia chinensis*. Nat Commun. 2013;4:2640.

156. He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, Lee TH, Wang X, Cai Q, Li D, et al. Draft genome sequence of the mulberry tree *Morus notabilis*. Nat Commun. 2013;4:2445.
157. Martin G, Baurens FC, Droc G, Rouard M, Cenci A, Kilian A, Hastie A, Dolezel J, Aury JM, Alberti A, et al. Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics. 2016;17:243.
158. Liu MJ, Zhao J, Cai QL, Liu GC, Wang JR, Zhao ZH, Liu P, Dai L, Yan GJ, Wang WJ, et al. The complex jujube genome provides insights into fruit tree biology. Nature Commun. 2014;5:5315.
159. Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, et al. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res. 2011;18:65–76.
160. Nowak MD, Russo G, Schlapbach R, Huu CN, Lenhard M, Conti E. The draft genome of *Primula veris* yields insights into the molecular basis of heterostyly. Genome Biol. 2015;16:12.
161. Yagi M, Kosugi S, Hirakawa H, Ohmiya A, Tanase K, Harada T, Kishimoto K, Nakayama M, Ichimura K, Onozaki T, et al. Sequence analysis of the genome of carnation (*Dianthus caryophyllus* L.). DNA Res. 2014;21:231–41.
162. Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Chen LJ, He Y, Xu Q, Bian C, et al. The genome sequence of the orchid *Phalaenopsis equestris*. Nat Genet. 2015;47:65–72.
163. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–84.
164. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology. 2014;15:R59.
165. Natsume S, Takagi H, Shiraishi A, Murata J, Toyonaga H, Patzak J, Takagi M, Yaegashi H, Uemura A, Mitsuoka C, et al. The draft genome of hop (*Humulus lupulus*), an essence for brewing. Plant Cell Physiol. 2015;56: 428–41.
166. Wang XW, Wang HZ, Wang J, Sun RF, Wu J, Liu SY, Bai YQ, Mun JH, Bancroft I, Cheng F, et al. The genome of the mesopolyploid crop species *Brassica rapa*. Nature Genet. 2011;43:1035–U1157.
167. Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Gonzalez VM, Henaff E, Camara F, Cozzuto L, Lowy E, et al. The genome of melon (*Cucumis melo* L.). Proc Natl Acad Sci U S A. 2012;109:11872–7.
168. Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet. 2013;45:51–8.
169. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nature Genetics. 2014;46: 270–8.
170. Qin C, Yu C, Shen Y, Fang X, Chen L, Min J, Cheng J, Zhao S, Xu M, Luo Y, et al. Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc Natl Acad Sci U S A. 2014;111:5135–40.
171. Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, Yamaguchi H, Sato S, Isobe S, Tabata S, Fukuoka H. Draft genome sequence of eggplant (*Solanum melongena* L.): the representative *Solanum* species indigenous to the Old World. DNA Research. 2014;21:649–60.
172. Jeong YM, Kim N, Ahn BO, Oh M, Chung WH, Chung H, Jeong S, Lim KB, Hwang YJ, Kim GB, et al. Elucidating the triplicated ancestral genome structure of radish based on chromosome-level comparison with the Brassica genomes. Theor Appl Genet. 2016;129:1357–72.
173. Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F, Tafer H, Rupp O, Sorensen T, Stracke R, Reinhardt R, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014; 505:546.
174. Dorn KM, Fankhauser JD, Wyse DL, Marks MD. A draft genome of field pennycress (*Thlaspi arvense*) provides tools for the domestication of a new winter biofuel crop. DNA Res. 2015;22:121–31.
175. Singh R, Ong-Abdullah M, Low ETL, Manaf MAA, Rosli R, Nookiah R, Ooi LCL, Ooi SE, Chan KL, Halim MA, et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. Nature. 2013;500: 335–9.
176. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci U S A. 2006;103:11647–52.
177. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature Genet. 2013;45:891–U228.
178. Wu HJ, Zhang ZH, Wang JY, Oh DH, Dassanayake M, Liu BH, Huang QF, Sun HX, Xia R, Wu YR, et al. Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proc Natl Acad Sci U S A. 2012;109:12219–24.
179. Li FG, Fan GY, Wang KB, Sun FM, Yuan YL, Song GL, Li Q, Ma ZY, Lu CR, Zou CS, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nature Genet. 2014;46:567–72.
180. Zhang TZ, Hu Y, Jiang WK, Fang L, Guan XY, Chen JD, Zhang JB, Saski CA, Scheffler BE, Stelly DM, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nature Biotechnol. 2015;33:531–U252.
181. Yuan DJ, Tang ZH, Wang MJ, Gao WH, Tu LL, Jin X, Chen LL, He YH, Zhang L, Zhu LF, et al. The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. Scientific Reports. 2015;5:17662.
182. Willing EM, Rawat V, Mandakova T, Maumus F, James GV, Nordstrom KJV, Becker C, Warthmann N, Chica C, Szarzynska B, et al. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. Nature Plants. 2015;1:1–7.
183. Varshney RK, Chen WB, Li YP, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan GY, Whaley AM, et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nature Biotechnol. 2012;30:83–U128.
184. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B, et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol. 2013; 31:240–6.
185. Parween S, Nawaz K, Roy R, Pole AK, Suresh BV, Misra G, Jain M, Yadav G, Parida SK, Tyagi AK, et al. An advanced draft genome assembly of a desi type chickpea (*Cicer arietinum* L.). Sci Reports. 2015;5:12806.
186. Kang YJ, Kim SK, Kim MY, Lestari P, Kim KH, Ha BK, Jun TH, Hwang WJ, Lee T, Lee J, et al. Genome sequence of mungbean and insights into evolution within Vigna species. Nat Commun. 2014;5:5443.
187. De Vega JJ, Ayling S, Hegarty M, Kudrna D, Goicoechea JL, Ergon A, Rognli OA, Jones C, Swain M, Geurts R, et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Sci Reports. 2015;5:17394.
188. Sakai H, Naito K, Ogiso-Tanaka E, Takahashi Y, Iseki K, Muto C, Satou K, Teruya K, Shiroma A, Shimoji M, et al. The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. Sci Reports. 2015;5:16780.
189. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, et al. Genome structure of the legume, *Lotus japonicus*. DNA Res. 2008;15:227–39.
190. Bertioli DJ, Cannon SB, Froenicke L, Huang GD, Farmer AD, Cannon EKS, Liu X, Gao DY, Clevenger J, Dash S, et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet. 2016;48:438–46.
191. Wang LH, Yu S, Tong CB, Zhao YZ, Liu Y, Song C, Zhang YX, Zhang XD, Wang Y, Hua W, et al. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. Genome Biol. 2014;15:R39.
192. VanBuren R, Bryant D, Edger PP, Tang HB, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature. 2015;527:508–U209.
193. Peng ZH, Lu Y, Li LB, Zhao Q, Feng Q, Gao ZM, Lu HY, Hu T, Yao N, Liu KY, et al. The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). Nat Genet. 2013;45:456–61.
194. Ming R, VanBuren R, Liu YL, Yang M, Han YP, Li LT, Zhang Q, Kim MJ, Schatz MC, Campbell M, et al. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biol. 2013;14:R41.
195. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sorensen I, Lichtenstein G, et al. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. Nat Genet. 2014;46:1034–8.
196. Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Muehlbauer GJ, et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491:711–6.

197. Mayer KFX, Rogers J, Dolezel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ, Sourdille P, et al. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science. 2014;345:1251788.
198. Ling HQ, Zhao SC, Liu DC, Wang JY, Sun H, Zhang C, Fan HJ, Li D, Dong LL, Tao Y, et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature. 2013;496:87–90.
199. Jia JZ, Zhao SC, Kong XY, Li YR, Zhao GY, He WM, Appels R, Pfeifer M, Tao Y, Zhang XY, et al. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature. 2013;496:91–5.
200. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang CJ, Chougule K, Gao DY, Iwata A, Goicoechea JL, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet. 2018;50:285–96.
201. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp indica). Science. 2002;296:79–92.
202. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al. The genome of *Theobroma cacao*. Nat Genet. 2011;43:101–8.
203. Zhang GY, Liu X, Quan ZW, Cheng SF, Xu X, Pan SK, Xie M, Zeng P, Yue Z, Wang WL, et al. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. Nat Biotechnol. 2012;30:549–54.
204. Sierro N, Battey JND, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV. The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun. 2014;5:3833.
205. Jung S, Ficklin SP, Lee T, Cheng CH, Blenda A, Zheng P, Yu J, Bombarely A, Cho I, Ru S, et al. The Genome Database for Rosaceae (GDR): year 10 update. Nucleic Acids Res. 2014;42:D1237–44.
206. Tanaka H, Hirakawa H, Kosugi S, Nakayama S, Ono A, Watanabe A, Hashiguchi M, Gondo T, Ishigaki G, Muguerza M, et al. Sequencing and comparative analyses of the genomes of zoysiagrasses. DNA Res. 2016;23:171–80.
207. Martinez-Garcia PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, Paul R, Butterfield TS, Britton MT, Reagan RL, et al. The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. Plant J. 2016;87:507–32.
208. Qi X, Li MW, Xie M, Liu X, Ni M, Shao G, Song C, Kay-Yuen Yim A, Tao Y, Wong FL, et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun. 2014;5:4340.
209. Ma T, Wang JY, Zhou GK, Yue Z, Hu QJ, Chen Y, Liu BB, Qiu Q, Wang Z, Zhang J, et al. Genomic insights into salt adaptation in a desert poplar. Nat Commun. 2013;4:2797.
210. Tang CR, Yang M, Fang YJ, Luo YF, Gao SH, Xiao XH, An ZW, Zhou BH, Zhang B, Tan XY, et al. The rubber tree genome reveals new insights into rubber production and species adaptation. Nat Plants. 2016;2:16073.
211. Guo L, Qiu J, Han Z, Ye Z, Chen C, Liu C, Xin X, Ye CY, Wang YY, Xie H, et al. A host plant genome (*Zizania latifolia*) after a century-long endophyte infection. Plant J. 2015;83:600–9.
212. Scaglione D, Reyes-Chin-Wo S, Acquadro A, Froenicke L, Portis E, Beitel C, Tirone M, Mauro R, Lo Monaco A, Mauromicale G, et al. The genome sequence of the outbreeding globe artichoke constructed de novo incorporating a phase-aware low-pass sequencing strategy of F1 progeny. Sci Rep. 2016;6:19427.
213. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang JY, Bowman M, Iovene M, Sanseverino W, Cavagnaro P, et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet. 2016;48:657–66.
214. Bombarely A, Moser M, Amrad A, Bao M, Bapaume L, Barry CS, Bliek M, Boersma MR, Borghi L, Bruggmann R, et al. Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. Nat Plants. 2016;2:16074.
215. Dai XG, Hu QJ, Cai QL, Feng K, Ye N, Tuskan GA, Milne R, Chen YN, Wan ZB, Wang ZF, et al. The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res. 2014;24:1274–7.
216. Rowley ER, Fox SE, Bryant DW, Sullivan CM, Priest HD, Givan SA, Mehlenbacher SA, Mockler TC. Assembly and characterization of the European hazelnut 'Jefferson' transcriptome. Crop Sci. 2012;52:2679–2686.
217. Yasui Y, Hirakawa H, Ueno M, Matsui K, Katsube-Tanaka T, Yang SJ, Aii J, Sato S, Mori M. Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes. DNA Res. 2016;23:215–24.
218. Zhang G, Tian Y, Zhang J, Shu L, Yang S, Wang W, Sheng J, Dong Y, Chen W. Hybrid de novo genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). Gigascience. 2015;4:62.
219. Zhang J, Tian Y, Yan L, Zhang G, Wang X, Zeng Y, Zhang J, Ma X, Tan Y, Long N, et al. Genome of plant maca (*Lepidium meyenii*) illuminates genomic basis for high altitude adaptation in the central Andes. Mol Plant. 2016;9:1066–1077.
220. Yan L, Wang X, Liu H, Tian Y, Lian J, Yang R, Hao S, Wang X, Yang S, Li Q, et al. The genome of *Dendrobium officinale* illuminates the biology of the important traditional Chinese orchid herb. Mol Plant. 2015;8:922–34.
221. Tian Y, Zeng Y, Zhang J, Yang C, Yan L, Wang X, Shi C, Xie J, Dai T, Peng L, et al. High quality reference genome of drumstick tree (*Moringa oleifera* Lam.), a potential perennial crop. Sci China Life Sci. 2015;58:627–38.
222. Xiao L, Yang G, Zhang L, Yang X, Zhao S, Ji Z, Zhou Q, Hu M, Wang Y, Chen M, et al. The resurrection genome of *Boea hygrometrica*: a blueprint for survival of dehydration. Proc Natl Acad Sci U S A. 2015;112:5833–7.
223. Zhang GQ, Xu Q, Bian C, Tsai WC, Yeh CM, Liu KW, Yoshida K, Zhang LS, Chang SB, Chen F, et al. The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. Sci Reports. 2016;6:19029.
224. Cheng SF, van den Bergh E, Zeng P, Zhong X, Xu JJ, Liu X, Hofberger J, de Bruijn S, Bhide AS, Kuelahoglu C, et al. The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. Plant Cell. 2013;25:2813–30.
225. Cruz F, Julca I, Gomez-Garrido J, Loska D, Marcet-Houben M, Cano E, Galan B, Frias L, Ribeca P, Derdak S, et al. Genome sequence of the olive tree, *Olea europaea*. Gigascience. 2016;5:29.
226. Lee H, Golicz AA, Bayer PE, Jiao YN, Tang HB, Paterson AH, Sablok G, Krishnaraj RR, Chan CKK, Batley J, et al. The genome of a Southern hemisphere seagrass species (*Zostera muelleri*). Plant Physiol. 2016;172:272–83.
227. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, et al. The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell. 2010;22:2943–55.
228. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. GitHub; 2018. https://github.com/qiao-xin. Accessed 18 Feb 2019.
229. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Plant Duplicate Gene Database; 2018. http://pdgd.njau.edu.cn:8080. Accessed 18 Feb 2019.
230. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. figshare; 2018. https://doi.org/10.6084/m9.figshare.7264667.v1. Accessed 18 Feb 2019.
231. Michael TP, VanBuren R. Progress, challenges and the future of crop genomes. Curr Opin Plant Biol. 2015;24:71–81.
232. Federhen S. The NCBI taxonomy database. Nucleic Acids Res. 2012;40:D136–43.
233. Castillo AI, Nelson ADL, Haug-Baltzell AK, Lyons E: A tutorial of diverse genome analysis tools found in the CoGe web-platform using Plasmodium spp. as a model. Database 2018, 2018:bay030-bay030.