nature communications

Article

https://doi.org/10.1038/s41467-025-61676-3

The subordinate role of pseudogenization to recombinative deletion following polyploidization in angiosperms

Received: 15 May 2024

Accepted: 26 June 2025

Published online: 09 July 2025

Check for updates

Ewout Crombez[1,2] ✉, Yves Van de Peer[1,2,3,4] ✉ & Zhen Li[1,2] ✉

Extensive gene loss is a hallmark of rediploidization following polyploidization, but its molecular basis remains unclear: whether it occurs primarily through pseudogenization or DNA deletion. Here, we examine pseudogenization in collinear segments from ancient whole-genome multiplications (WGMs) across 12 angiosperms. Although total pseudogenes are abundant, we find far fewer WGM-derived pseudogenes than expected if pseudogenization and DNA deletion contribute equally to gene loss. Simulations of neutrally evolving pseudogenes indicate that, if DNA deletion is absent, pseudogenes should be detectable for far longer than observed in the paleo-polyploid genomes, suggesting gene loss driven by DNA deletion. Analyses of three neo-autopolyploid genomes confirm this pattern: among substantial gene loss, DNA deletions occur on average 1.5 times more frequently than pseudogenization. Our findings imply that gene loss post-polyploidization primarily takes place via DNA deletion, enabled by a genomic environment with an elevated recombination rate created by WGMs. In contrast, small-scale duplications yield scattered duplicated genes, which appear less exposed to deletion and hence result in a high number of pseudogenes. This model is further reinforced by an enrichment of WGM-derived pseudogenes in high recombination regions. Moreover, some pseudogenes may govern a function, as indicated by non-neutral K_a/K_s ratios and overlap with lncRNAs.

Polyploidization, the duplication or multiplication of the total set of chromosomes, is a prominent evolutionary phenomenon that has been observed across the entire plant kingdom, especially in angiosperms[1–3]. After formation, polyploidy is often followed by drastic changes in the genome, such as genomic rearrangements and loss events, which may eventually lead to a return to a diploid state[4]. This process, referred to as rediploidization or diploidization, significantly contributes to the evolutionary importance of polyploidy, influencing speciation, diversity, and evolutionary innovation[5–8]. Interestingly, a typical contemporary angiosperm genome has undergone one or more rounds of whole genome duplication or multiplication (WGM), each time followed by rediploidization[9,10]. This suggests that rediploidization plays a pivotal role in the long-term evolutionary trajectory following polyploidization. Namely, rediploidization has been postulated to mitigate the drawbacks associated with polyploidy, such as meiotic instability, gene dosage imbalance, and increased nutrient demands, while concurrently generating variation, e.g., in gene copies and genomic rearrangements, upon which selection can operate[5,11]. As such, rediploidization can facilitate speciation through the differential loss of duplicate genes and distinct genomic rearrangements[12].

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [2]Center for Plant Systems Biology, VIB, Ghent, Belgium. [3]Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa. [4]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, China. ✉e-mail: ewout.crombez@ugent.be; yves.vandepeer@psb.ugent.be; zhen.li@psb.ugent.be

One of the key features of post-WGM genome evolution and rediploidization is extensive gene loss. To this end, the fraction of retained WGM-derived duplicate genes appears to rapidly decrease over time, exhibiting an apparent exponential distribution[4,13,14]. This rapid, pervasive loss of genes can be attributed to the genome-wide redundancy for all genes after WGM. As such, it has been hypothesized that one copy of the duplicated gene pair may get lost without negatively impacting the fitness of an organism, as there is still another functional copy of the gene present in the genome[13]. Indeed, after gene and genome duplication, gene loss or non-functionalization is considered as the most probable evolutionary path[13,15–17]. Interestingly, despite this redundancy, duplicate genes are not simply lost at random. Some duplicate genes involved in specific functional categories are preferentially retained after WGM[4]. Several models have been put forward to explain this biased retention of duplicate pairs[15,18]. For example, according to the dosage balance hypothesis, duplicate genes whose proteins participate in a pathway and/or multi-protein complex are preferentially retained after polyploidization due to dosage constraints, specifically through selection against the disruption of the stoichiometric balance of that pathway or multi-protein complex[19]. This hypothesis can explain the functional bias of retained duplicate genes towards transcription regulation, signaling, and developmental processes[20–22].

Although duplicate gene retention has been extensively studied, a major gap in our understanding of rediploidization pertains to the processes by which genes are lost. As a result, little is known about the mechanistic underpinnings of gene loss following WGM. In principle, gene loss can occur through two notable mechanisms, DNA deletion and pseudogenization. DNA deletion refers to the physical removal of a segment of DNA from the genome. The deleted DNA segment may encompass one or multiple genes. Hereby, genetic processes leading to DNA deletion are mainly related to recombination or mobilization of a transposable element[23]. Illegitimate recombination, i.e., the recombination of two non-homologous sequences, has been suggested as the most prominent mechanism for gene loss for angiosperms[4,24–27]. Pseudogenization, on the other hand, refers to the process whereby a gene acquires deleterious mutations leading to the non-functionality of that gene, followed by further degradation through mutation. Pseudogenes are typically defined as genomic sequences that resemble functional protein-coding genes but have lost their function[28]. In this study, we define a pseudogene more broadly as any noncoding genomic sequence (i.e., in intergenic regions and introns) that shares sequence similarity with a functional gene but carries deleterious mutations. In animals and ferns, pseudogenization has long been considered the most prominent mechanism for gene loss[29–31]. Although originally considered rarer in angiosperms, several studies have indicated that pseudogenization may not be as limited as previously thought[32–37]. Moreover, Xie et al. and Mascagni et al. investigated pseudogenes that originated from WGMs and pinpointed a number ranging from hundreds to thousands WGM-derived pseudogenes in seven and five paleo-polyploid species, respectively, suggesting pseudogenization plays a role in gene loss after WGM in angiosperms[32,36].

The two mechanisms of gene loss, DNA deletion and pseudogenization, may have very different biological implications. Importantly, DNA deletion results in the complete removal of the sequence from the genome, whereas after pseudogenization, the sequence is still retained. As such, a pseudogene sequence may still have the potential to regain its original functional role through the reversal of deleterious mutations if the selective environment shifts[38], a capability that is not possible after DNA deletion. Furthermore, pseudogenes may even acquire a novel protein-coding function, e.g., as a truncated protein[28], or play a significant role in the origin of regulatory elements[28,36,39,40]. Indeed, several studies have put pseudogenes forward as an important source of long noncoding RNAs (lncRNAs), which govern regulatory functions[36,41–43]. Gene loss through pseudogenization may thus be evolutionary appealing as it is reversible and may also lead to functional novelty and diversity. A growing body of research indicates that at least some pseudogenes exhibit some form of functionality[28,34,39,44–46].

Here, we investigate the relative importance of pseudogenization in comparison to recombinative DNA deletion and analyze its temporal evolution during the process of rediploidization. To this end, we identify pseudogenes that trace back to WGMs in 12 paleo-polyploid plant genomes. All these paleo-polyploid plant species have undergone at least one WGM at various timings in their evolution, but are currently functionally diploid, allowing us to understand gene loss in the process of rediploidization. Furthermore, to obtain a more thorough comprehension of gene loss in a shorter timeframe following WGM, we investigate pseudogenization in three neo-autopolyploid genomes that have not undergone complete rediploidization yet. Although there have been previous studies that identified WGM-derived pseudogenes[32,36], our study specifically focuses on the relationship between pseudogenes and the process of rediploidization. Our results indicate that a large fraction of duplicate genes appear to rapidly get lost through DNA deletion following WGM, most likely via recombination-related processes. Nonetheless, WGM-derived pseudogenes are identifiable, albeit in much lower numbers than expected given their rediploidization histories. Moreover, at least some of these pseudogenes have acquired a functional role as lncRNAs.

Results and discussion

Identification of pseudogenes in paleo-polyploid species

We investigated the extent of pseudogenization in 12 paleo-polyploid plant species that underwent WGMs at different times in their evolutionary past (Fig. 1a and Supplementary Table 1). The PseudoPipe pipeline[47] was utilized to identify putative pseudogenes. Pseudogenes were identified as genomic sequences that reside in the noncoding regions of a genome, encompassing all areas outside the coding exons (i.e., intergenic regions and introns), and have sequence similarity with a functional gene but contain deleterious mutations. In total, we obtain a highly variable total number of putative pseudogenes for different species, ranging from 6989 in *Arabidopsis thaliana* to 236,487 in *Zea mays* (Supplementary Fig. 1). These putative pseudogenes were classified into three classes in accordance to PseudoPipe: (1) "retro-transposed (or processed) pseudogenes" that originated by retrotransposition of a reverse transcribed mRNA, (2) "duplicated pseudogenes" that originated from the duplication of a gene, and (3) "fragmented pseudogenes", encompassing pseudogenes characterized by a highly fragmented structure, making it challenging to attribute them to a specific origin (Methods).

In addition to these generic classes, we further defined a class of "WGM-derived pseudogenes", encompassing our focal interest, for pseudogenes initially classified as "duplicated" or "fragmented" that were located in intergenic regions. This classification is based on their presence in collinear homoeologous (i.e., derived from a WGD or hybridization event) segments, which are defined as two or more regions within a genome containing several paralogous genes or pseudogenes that largely maintain their order. As such, a "duplicated" or "fragmented" pseudogene that forms a homologous pair with a functional paralogous gene, is assumed to be "WGM-derived". Duplicated pseudogenes that are not WGM-derived are then considered to be derived from small-scale duplication (SSD). The WGM-derived pseudogenes are consistently the smallest class, ranging from 61 in *Sorghum bicolor* to 1761 in *Glycine max* (Fig. 1b and Supplementary Fig. 1). Most WGM-derived pseudogenes came from the fragmented rather than the duplicated class, suggesting that pseudogenes of this class generally have a fragmented structure. We compared our results with previous studies, revealing variability in the identification of pseudogenes that can be attributed to the absence of a precise

https://doi.org/10.1038/s41467-025-61676-3

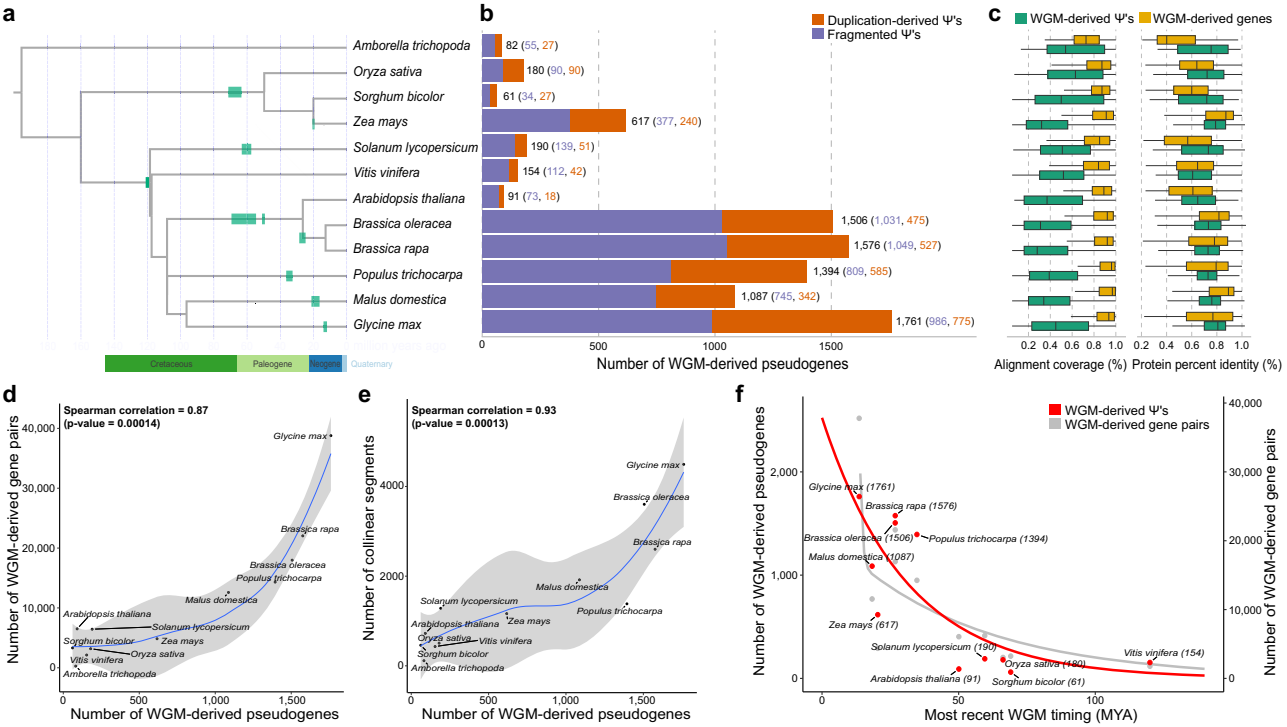

Fig. 1 | Characteristics of identified pseudogenes derived from whole genome multiplications (WGMs) in twelve paleo-polyploid plant species. a A time tree of the selected species obtained with TimeTree 5[80]. Whole Genome Multiplications (WGMs), as identified and dated by Vanneste et al.[50], are highlighted on the tree as green rectangles. The divergence time between *S. bicolor* and *Z. mays* was taken from Vanneste et al.[50] as the divergence time from TimeTree does not correspond with the lineage-specific WGD in *Z. mays*, i.e., not shared with *S. bicolor*. **b** Number of identified WGM-derived pseudogenes per species with duplicated pseudogenes in orange and fragmented pseudogenes in purple. **c** Boxplots of alignment coverage and protein percent identity of WGM-derived pseudogenes (green; *A. trichopoda*: $n = 82$, *A. thaliana*: $n = 91$, *B. oleracea*: $n = 1506$, *B. rapa*: $n = 1576$, *G. max*: $n = 1761$, *M. domestica*: $n = 1087$, *O. sativa*: $n = 180$, *P. trichocarpa*: $n = 1394$, *S. lycopersicum*: $n = 190$, *S. bicolor*: $n = 61$, *V. vinifera*: $n = 154$, and *Z. mays*: $n = 617$) and WGM-derived duplicate genes (yellow; *A. trichopoda*: $n = 306$, *A. thaliana*: $n = 6318$, *B. oleracea*: $n = 17{,}461$, *B. rapa*: $n = 22{,}149$, *G. max*: $n = 38{,}579$, *M. domestica*: $n = 11{,}828$, *O. sativa*: $n = 3179$, *P. trichocarpa*: $n = 14{,}336$, *S. lycopersicum*: $n = 6495$, *S. bicolor*: $n = 3406$, *V. vinifera*: $n = 2078$, and *Z. mays*: $n = 4493$) compared to their functional paralogous genes. The center, upper, and lower bound of the boxplots correspond, respectively, to the median, 25th and 75th percentiles. The bounds of the whiskers of the boxplots extend to 1.5 times the inter-quartile range starting from the bounds of the boxplots. **d** Correlation between number of WGM-derived pseudogenes and number of WGM-derived gene pairs. **e** Correlation between number of WGM-derived pseudogenes and number of collinear segments. The blue lines in d and e represent the smoothed conditional mean, with the grey shaded area around the regression line delineating the 95% confidence interval. Two-sided correlation tests were conducted to evaluate statistical significance. **f** Number of WGM-derived pseudogenes (red) and WGM-derived gene pairs (gray) correlate with the timing of the most recent WGM event. Timing of the most recent WGM is expressed in million years ago (MYA). The red exponential curve is the best fit for the WGM-derived pseudogenes. The gray double-exponential curve is the best fit for the WGM-derived gene pairs (Methods). Underlying data files on Zenodo[107].

computational definition for pseudogenes (Supplementary Note 1 and Supplementary Fig. 2).

The total number of pseudogenes exhibits a moderately significant correlation with genome size (Spearman's $\rho = 0.57$, p-value = 0.003) and the number of retrotransposons in the genome (Spearman's $\rho = 0.59$, p-value = 0.02). Moreover, the number of fragmented pseudogenes also correlates significantly with the number of retrotransposons in the genome (Spearman's $\rho = 0.77$, p-value = 0.03). Together with the large number of retro-transposed pseudogenes, which were very often larger than the number of duplicated (SSD- and WGM-derived) pseudogenes (Supplementary Fig. 1), the significant correlations suggest that a large fraction of the pseudogenes originated from retro-transposition events. Surprisingly, the number of retro-transposed pseudogenes itself does not correlate with the number of retrotransposons in the genome (Spearman's $\rho = 0.36$, p-value = 0.18). A potential explanation for the lack of correlation may be that retro-transposed pseudogenes were identified using stringent cut-off criteria, resulting in many of them being categorized as fragmented pseudogenes. More correlations between different categories of pseudogenes and various genomic features are available in Supplementary Table 2.

In all species analyzed, fragmented pseudogenes formed the most abundant class, suggesting that a large proportion of pseudogenes have experienced extensive fragmentation. This may be the result of recombinational processes that lead to rearrangements and deletions in the genome. Fragmentation is also reflected in the alignment coverage of pseudogenes against their functional paralogs, which is consistently lower for WGM-derived pseudogenes than the alignment coverage observed between pairs of paralogous genes within collinear segments, i.e., WGM-derived duplicate genes or anchor pairs (Fig. 1c, Supplementary Note 2, and Supplementary Fig. 3). The protein percent identity between WGM-derived pseudogenes and their functional paralogous gene is relatively high when compared to the percent identity between WGM-derived duplicate genes, especially for species with an ancient WGM (Fig. 1c, Supplementary Note 2, and Supplementary Fig. 3). However, only looking at protein percent identity gives an incomplete picture as it overlooks the alignment coverage. Multiplying the protein percent identity by the alignment coverage provides a metric for the protein percent identity, including non-aligned parts, i.e., global protein percent identity. This combined metric tends to be higher for WGM-derived duplicate genes than for WGM-derived pseudogenes (Supplementary Fig. 4).

Although the total number of pseudogenes does not significantly correlates with the timing of the most recent WGM in a species (Spearman's $\rho = -0.43$, p-value = 0.20), the number of WGM-derived pseudogenes is strongly negatively correlated with the timing of the most recent WGM (Spearman's $\rho = -0.79$, p-value = 0.01, Fig. 1f). The more recent the WGM event, the more WGM-derived pseudogenes are identified, suggesting that pseudogenes are lost over time, either through further sequence decay, or DNA deletion. In concordance, we find significant enrichments of various meiosis and RNA/DNA-related Gene Ontology (GO) terms in WGM-derived pseudogenes for paleo-polyploid species with a WGM <35 million years ago (MYA), while for paleo-polyploid species with a WGM more than 35 MYA, we observed no significant enrichments (Supplementary Note 3, Supplementary Tables 3, 4). We would indeed expect that WGM-derived pseudogenes are enriched in functional categories associated with meiosis and RNA/DNA-related processes, as these are underrepresented in retained duplicate genes[21,22,48]. The lack of enrichment observed in ancient paleo-polyploids (WGM > 35 MYA) suggests that these pseudogenes have disappeared over time.

The number of WGM-derived pseudogenes is also significantly positively correlated with the total number of genes in the genome (Spearman's $\rho = 0.83$, p-value = 0.001), the number of WGM-derived duplicate genes (Spearman's $\rho = 0.87$, p-value = 0.00014, Fig. 1d), and the number of collinear segments (Spearman's $\rho = 0.93$, p-value = 0.00013, Fig. 1e). These strong correlations are likely a reflection of the method that was used to identify WGM-derived pseudogenes. The number of WGM-derived pseudogenes could be impacted by several factors, a concern that we discuss in the following sections.

Why so few WGM-derived pseudogenes?

Considering the rediploidization histories of the species and introducing a null hypothesis that pseudogenization and DNA deletion contribute equally to gene loss after WGM, we would expect to observe a comparable number of WGM-derived pseudogenes as DNA deletions to reflect this balanced contribution. Due to the nature of DNA deletion, here we could only infer the occurrence of gene loss via DNA deletion post-WGM, when no detectable WGM-derived duplicate or pseudogene remains in the genome. Although the total number of identified pseudogenes is large, often surpassing the total number of functional genes within the given species, the number of WGM-derived pseudogenes is relatively limited compared to the total number of lost genes after WGM (Supplementary Table 5). For example, in *A. thaliana*, which underwent its most recent WGM about 50 MYA, only about 21% of the total number of genes are retained as gene anchor pairs, suggesting that about 79% were lost or translocated. However, we only find 91 WGM-derived pseudogenes in *A. thaliana*, suggesting that the large majority of duplicate genes were eventually lost through DNA deletion. Species with a more recent WGM, such as *G. max*, which underwent its most recent WGM less than 13.6 MYA, tend to have more WGM-derived pseudogenes (1761 pseudogenes), but the number is still modest compared to the gene number (52,872 genes) and the estimated 48% of WGM-derived duplicate genes that were lost or translocated after the WGM. Therefore, a substantially lower number of WGM-derived pseudogenes, compared to inferred DNA deletions, suggests that DNA deletion contributes disproportionately to gene loss following WGM.

It should be noted that, here, we cannot know whether this DNA deletion occurred in a direct (one-step) process in which a functional gene gets deleted, or an indirect (two-step) process in which pseudogenization has preceded DNA deletion (Supplementary Fig. 5). Since both scenarios would leave no detectable duplicate or pseudogene, they are indistinguishable in our analyses and are therefore collectively classified as gene losses by DNA deletion.

To gain deeper insight into the respective loss trajectories of duplicated gene copies and pseudogenes after WGM, we fitted two competing models separately to WGM-derived duplicates and WGM-derived pseudogenes over time[49]. The first, exponential model, consists of one phase with a constant loss rate whereby pseudogenes or genes get lost passively (i.e., under no selective constraint), and independently (i.e., not as part of a coordinated event affecting multiple genes at once). This corresponds to a single-exponential decay curve. The second, double-exponential model, consists of an initial rapid loss phase, possibly involving the simultaneous removal of adjacent genes or co-regulated genes, followed by a slower loss phase. This can be described by a double-exponential decay curve, which starts with a steep drop and then flattens out[49]. The initial rapid loss phase is thought to reflect the relaxed purifying selection due to genome-wide redundancy of genes immediately following a WGM. WGM-derived pseudogene loss consistently fits the single-exponential model best ($r^2 = 0.65$; red curve in Fig. 1f). In contrast, WGM-derived gene loss of a duplicate copy fits the double-exponential model better ($r^2 = 0.8$ vs 0.56; grey curve in Fig. 1f). Both model comparisons are supported by the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) comparisons (Supplementary Tables 6, 7). These results remain consistent when using the timing of WGMs represented by the peaks observed in distributions of synonymous substitutions per synonymous site (K_s) for all duplicate genes from Vanneste et al.[50] (Supplementary Fig. 6, Supplementary Tables 6, 7).

While merely fitting a curve does not guarantee the correctness of the underlying assumed model, the notable differences in fits between WGM-derived pseudogenes and WGM-derived gene pairs are intriguing. If the fits reflect the essence of the underlying models on the loss processes, our findings indicate that duplicated genes undergo an initial rapid loss phase after WGM events, followed by a phase of gene loss with a slower rate. The pseudogenes, in contrast, show a consistent loss rate, indicating a different loss dynamic. Combined with the scarcity of WGM-derived pseudogenes, this indicates that pseudogenization could not solely account for the rapid gene loss right after WGM events, and one-step DNA deletion, without a pseudogene intermediate, is likely relevant. Although the double-exponential model does not consider the mechanisms by which the rapid loss happens, our results here would suggest that DNA deletion dominates the gene loss processes after WGMs over pseudogenization. These observations may indicate that pseudogenization only has limited importance for gene loss compared to DNA deletion, especially right after WGMs, in angiosperms, as suggested before[26,27]. However, before jumping to this conclusion, we further explore whether the life span of pseudogenes and the degradation of collinear segments may lead to potential biases and complicate the interpretation of WGM-derived pseudogene numbers.

Simulating the evolution of neutrally evolving pseudogenes indicates a longer expected identifiability span than observed in paleo-polyploids

One potential bias which may have contributed to the low number of WGM-derived pseudogenes in the studied paleo-polyploids is that a pseudogene may accumulate small-scale mutations, rendering it undetectable due to falling below the sequence similarity thresholds for pseudogene identification. To interrogate this, we adopted the classical gene duplication model whereby gene duplicates are under relaxed functional constraint and free to accumulate mutations[13,17,51]. Once a deleterious mutation, e.g., a premature stop codon or frameshift, occurs, the gene becomes a pseudogene and continues to evolve without selective constraint[52]. As such, the more ancient the WGM, the more likely it is that its associated WGM-derived pseudogenes have degraded beyond detectability. We intentionally do not include large-scale DNA deletion as part of the simulation, instead solely focusing on small-scale mutations, i.e., substitutions and short indels, which drive pseudogenization.

https://doi.org/10.1038/s41467-025-61676-3

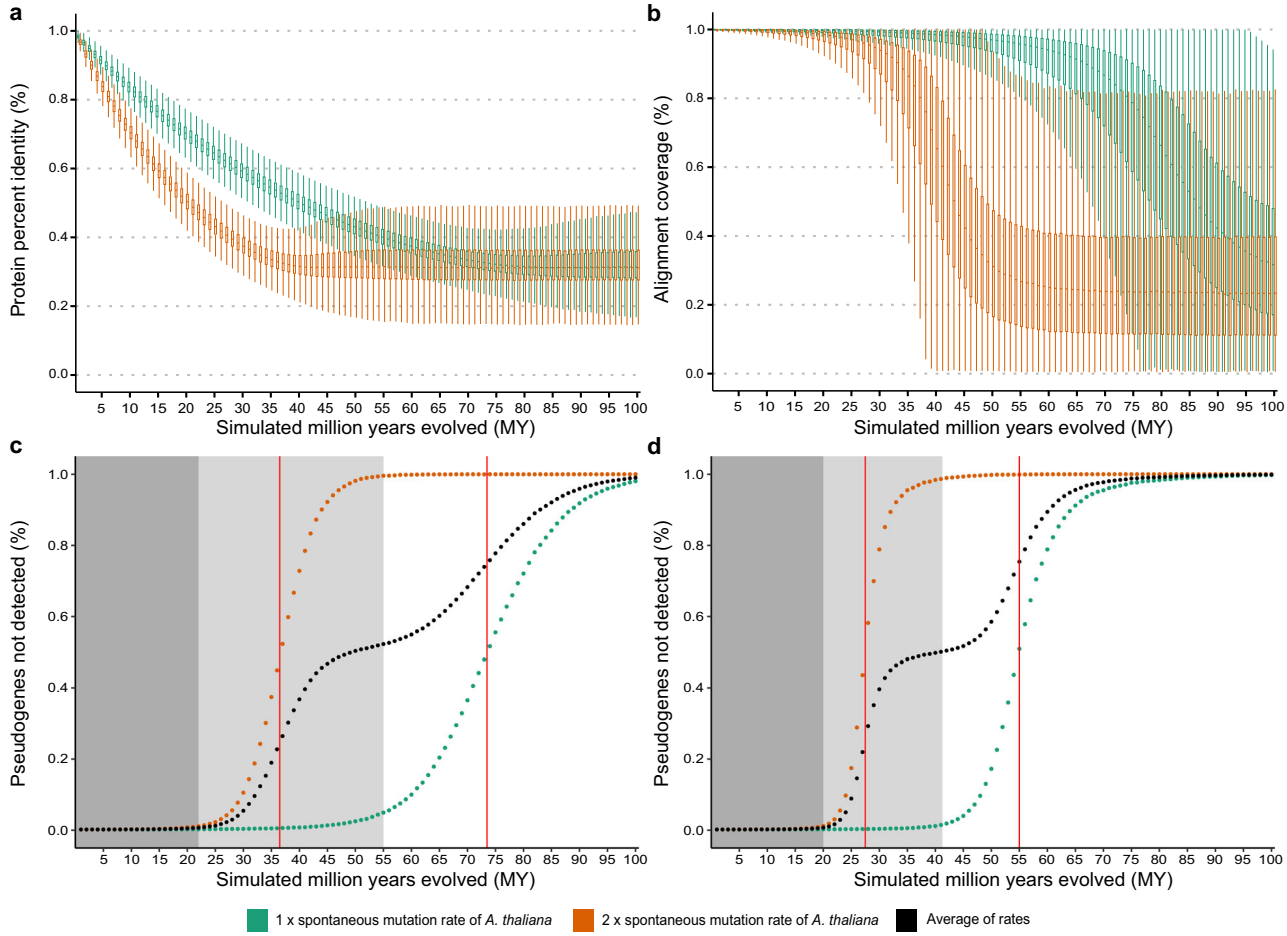

Fig. 2 | Simulated evolution of pseudogenes. Sequences were evolved following once (green) and twice (orange) the spontaneous mutation rate of *Arabidopsis thaliana*[53] ($n = 27{,}651$ coding DNA sequences). **a** Simulated evolution of protein percent identity between evolved and ancestral proteins over 100 million years (MY) represented by boxplots per million years. **b** Simulated evolution of alignment coverage between evolved and ancestral proteins over 100 MY represented by boxplots per million years. The center, upper and lower bound of the boxplots in a and b correspond, respectively, to the median, 25th and 75th percentiles. The bounds of the whiskers of the boxplots extend to 1.5 times the inter-quartile range starting from the bounds of the boxplots. **c** Percentage of the evolved pseudogenes that are not detectable over 100 MY (per million years) with the cut-offs used for fragmented pseudogenes (E-value $\leq 10^{-5}$, alignment coverage ≥ 0.05 and protein percent identity ≥ 0.2). **d** Percentage of the evolved pseudogenes that are not detectable over 100 MY (per million years) with the cut-offs used for duplicated pseudogenes (E-value $\leq 10^{-5}$, alignment coverage ≥ 0.05 and protein percent identity ≥ 0.4). The average percentage of the two spontaneous mutation rates is indicated in black. The dark gray region represents the period in which essentially all pseudogenes are still detectable. The light grey region represents the period after which half of the pseudogenes are detectable following the average of the spontaneous mutation rates. The two red lines represent the time at which half of the pseudogenes are detectable following, respectively, twice and once the spontaneous mutation rate of *Arabidopsis thaliana*. Underlying data files on Zenodo[107].

Starting with all coding sequences (CDS) of *A. thaliana*, we duplicated them to simulate a WGM event and then evolved one copy for at least 100 million years using a spontaneous substitution rate of 7×10^{-9} substitutions per site per year and 1- to 3-bp insertion and deletion rates of 0.3×10^{-9} and 0.6×10^{-9} per site per year, respectively[53]. Per one-million-year interval, we translated and aligned the evolved sequence to the ancestral one and record pairwise alignment statistics, most notably protein percent identity, alignment coverage, and E-value. Importantly, for pseudogene identification in the paleo-polyploids, we compared pseudogenes to their functional paralogs rather than their ancestral copies, which also accumulated mutations, albeit under different selection pressures. To address this, we additionally ran a second set of simulations using double the spontaneous substitution and indel rates to mimic both duplicate genes evolving neutrally. As such, these two simulations provide lower ($1 \times$ spontaneous mutation rates) and upper ($2 \times$ spontaneous mutation rates) bounds of sequence divergence, with the actual sequence divergence between a pseudogene and its functional paralog likely falling in between these two (Methods).

By tracking the protein percent identity and coverage of the simulated protein relative to its ancestral form over time, our simulation reveals alignment plateaus for simulations with both mutation rates. For protein percent identity, the plateau is reached at a median of around 30% protein percent identity after 40 to 80 MY for the simulations using 1× and 2× the spontaneous mutation rates, respectively (Fig. 2a). Similarly, alignment coverage also reaches a plateau at a median of about 23% after 120 MY and 60 MY using respectively 1× and 2× the spontaneous mutation rates (Fig. 2b and Supplementary Fig. 7). In addition, the variation of both alignment statistics increases through time.

By further exploring the plateaus in these alignment statistics, we found that they can be attributed to the pairwise alignment algorithm, which is inherently designed to seek similarities between two sequences to align most of the sequences. This, hence, underscores the importance of alignment significance, most often expressed as E-value, i.e., the number of expected alignments that would score at least as high as the observed alignment just by random chance[54]. As expected, the E-value increases over time, indicating that the obtained

alignments become less significant over time (Supplementary Fig. 8). Next, we employed the thresholds utilized for duplicated (protein percent identity ≥ 40%, alignment coverage ≥ 5% and *E*-value < 10^{-5}) and fragmented pseudogenes (protein percent identity ≥ 20%, alignment coverage ≥ 5% and *E*-value < 10^{-5}) for the paleo-polyploids to track the fraction of unidentifiable simulated pseudogenes over time. Adhering to the criteria established for fragmented pseudogenes, half of the pseudogenes should remain detectable on average 55 MY, ranging between 37 and 73 MY using the upper and lower boundaries, respectively (Fig. 2c). Similarly, when applying the more stringent criteria of duplicated pseudogenes, half of the simulated pseudogenes are expected to remain detectable on average 42 MY, ranging between 28 and 55 MY (Fig. 2d).

Therefore, considering that ~79% of the duplicate genes are lost after the most recent WGM timing of *A. thaliana* (50 MYA) (Supplementary Table 5), assuming that pseudogenization has contributed equally as deletion events (50%), and taking the average percentage of pseudogenes that are detectable in the simulations at the most recent WGM timing of *A. thaliana* (50% and 41% using the cut-offs of respectively the fragmented and duplicated pseudogenes), we expect to see a proportion between 16 and 20% of the number of genes prior to the WGM that are detectable as pseudogenes. Moreover, considering that ~21% of the WGM-derived duplicate genes are retained, we would thus expect to see a proportion between 13 and 16% of the current number of genes—3595–4425 lost genes—that are detectable as pseudogenes, which is much larger than what we observe (91 WGM-derived pseudogenes).

Similarly, for *Populus trichocarpa*, considering that ~58% of the duplicate genes were lost after its most recent WGM (34.7 MYA) (Supplementary Table 5), the same null hypothesis and that nearly all pseudogenes should remain identifiable following its most recent WGM timing, as it has a much slower spontaneous mutation rate of 1.33×10^{-10} per site per year[55] than *A. thaliana* (Supplementary Fig. 9), we would expect to see a proportion of about 20% of the current number of genes—6940 lost genes—that are detectable as pseudogenes. This is again much higher than the 1394 WGM-derived pseudogenes we detect. Moreover, these expected proportions are also much higher than the WGM-derived pseudogene numbers found by Xie et al.[36] and Mascagni et al.[32]. Although the above calculations are based on the strong assumption that all genes that have no collinear paralog have lost its paralog from the genome, these calculations give us a more robust indication that the number of identified WGM-derived pseudogenes is, indeed, lower than expected, and could not be attributed to the loss of detectability due to small-scale mutations. In addition, following the spontaneous mutation rates of *A. thaliana*, in the absence of DNA deletion, almost all simulated WGM-derived pseudogenes are identifiable after 20 to 45 MY (Fig. 2c, d), indicating that pseudogenes where a WGM occurred less than 20 MYA, e.g., *G. max*, *M. domestica*, and *Z. mays*, should be readily identifiable.

We acknowledge that two factors may affect our simulations. First, only using the spontaneous mutation rates of *A. thaliana* and *P. trichocarpa* may not accurately reflect reality for other species. Nevertheless, the two selected species are among the species with the highest and lowest spontaneous mutation rates among angiosperms measured so far, respectively[56]. The obtained results may thus represent a minimal and maximal expected identifiability time span for pseudogenes, with pseudogenes in other species experiencing an expected identifiability time span in between these two species.

Second, assuming immediate loss of selective constraint after duplication is of course not realistic in our simulation, reflected by our observation that both WGM-derived pseudogenes from *A. thaliana* and *P. trichocarpa* tend to have higher protein percent identity than the simulated results but comparable alignment coverage. Indeed, our simulations using the spontaneous mutation rate of *A. thaliana* revealed that frameshifts and premature stop codons could rapidly accumulate within the first 5 MY, indicating that pseudogenization can happen easily in the absence of selection (Supplementary Figs. 10, 11). In reality, before the acquisition of a deleterious mutation, the duplicated genes may initially be still under some selective constraint, resulting in a greater protein percent identity than anticipated under a scenario of complete neutral evolution. That said, a delayed process of pseudogenization by selection constraints would only prolong the detectability of pseudogenes, hence leading to a higher number of pseudogenes than our simulation. As such, by considering both factors, we would expect many more WGM-derived pseudogenes than the ones observed in all the studied paleo-polyploids.

Since the substantially lower numbers of WGM-derived pseudogenes diverge from our simulation expectation, we must consider other explanations beyond the loss of detectability resulting from the accumulation of small-scale mutations in neutrally evolving pseudogenes. The most likely explanation is loss through DNA deletion, a mutational process excluded from the simulations. The simulation results do not clarify whether this occurred in a direct (one-step) process, where a functional gene is deleted, or an indirect (two-step) process, where pseudogenization occurs before deletion (Supplementary Fig. 5). However, previous model fitting suggests a relevant role for the one-step process, especially immediately after the WGM (Fig. 1f). In the next section, we further investigate the mechanism of rapid loss and the respective importance of pseudogenization and DNA deletion in neo-polyploids.

Gene loss in neo-autopolyploid genomes confirms bias toward DNA deletion

Another potential bias which may limit WGM-derived pseudogene identification is the ability to detect collinear segments. After WGM, intragenomic collinearity typically diminishes over time because of genome rearrangements and gene loss itself[57]. This decline of collinearity through time is supported by a significant negative correlation between the timing of the most recent WGM and both the number of collinear segments (Spearman's $\rho = -0.84$, p-value = 0.022) and the number of WGM-derived gene pairs (Spearman's $\rho = -0.77$, p-value = 0.033). As such, the more ancient a WGM, the more challenging it becomes to find collinear segments, and subsequently to identify WGM-derived pseudogenes therein. Moreover, the number of WGM-derived pseudogenes itself significantly correlates with both the number of collinear segments and number of anchor pairs (Fig. 1d, e), indicating that the ability to detect collinear segments surely influences the identification of WGM-derived pseudogenes. Nonetheless, for species that underwent a WGM very recently and are still polyploid, i.e., neo-polyploids, we would expect to observe extensive collinearity. As such, if the loss of collinearity is the main reason for the low number of WGM-derived pseudogenes and gene loss equally occurs through pseudogenization and DNA deletion, we would expect to identify both loss categories comparably in neo-polyploid species.

To this end, we investigate three high-quality autopolyploid genomes: (1) the auto-octoploid *Saccharum spontaneum* AP85-441 (sugarcane) genome, which underwent two rounds of WGDs less than 0.8 MYA[58,59], (2) the auto-tetraploid *Solanum tuberosum* Otava (potato) genome, which underwent a WGD in 1981[60], and (3) the auto-tetraploid *Actinidia arguta* (kiwiberry) genome, which underwent a WGD about 3.13 MYA[61]. The auto-octoploid sugarcane has been sequenced to a sub-genome level with four sets of unphased sub-genomes, while the two tetraploids are sequenced to a haplotype level with four sets of phased haplotypes. Crucially, we focused on autopolyploids, but not allopolyploids, to reduce the potential inclusion of gene loss events that occurred prior to the WGM event during the period in which parental genomes evolved independently. This concern is most problematic for allopolyploids, since genetic diversity is generally greater between species than within a single species[62,63].

https://doi.org/10.1038/s41467-025-61676-3

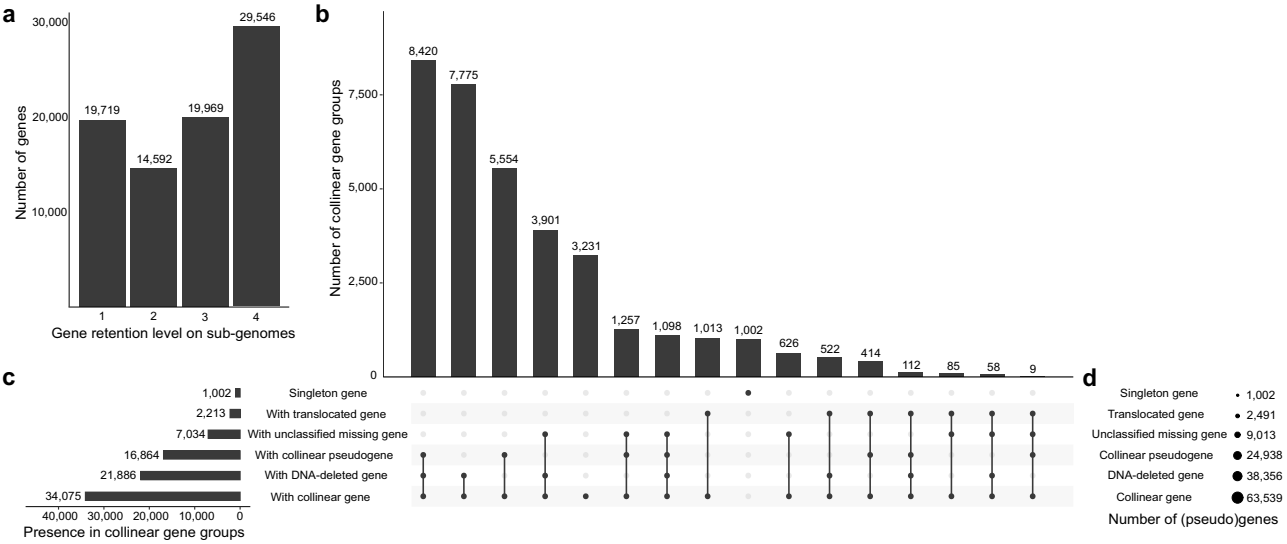

Fig. 3 | Gene retention and loss in the auto-octoploid *Saccharum spontaneum* AP85-441 (sugarcane) genome. a Gene retention level across the four sub-genomes of sugarcane. Gene retention level is expressed as the number of sub-genomes on which a gene has a collinear paralog, with a value of one indicating that it is found in only one sub-genome with no collinear paralogs. **b** Upset plot illustrating the combinations of various labels for evolutionary scenarios across collinear gene groups in sugarcane. A collinear gene group can contain the following labels: (1) With collinear gene, (2) With DNA-deleted gene, i.e., a gene lost through DNA deletion, (3) With collinear pseudogene, (4) With unclassified missing gene, (5) With translocated gene, or (6) Singleton gene, i.e., a single gene with no collinearity to any other sub-genome (Methods). **c** The number of collinear gene groups with certain labels. **d** The number of (pseudo)genes in the sugarcane genome with certain labels. Underlying data files on Zenodo[107].

Overall, the gene collinearity of the three species is relatively well maintained between sub-genomes/haplotypes, with a limited extent of inversions, translocations, and deletions (Supplementary Figs. 12–14). For sugarcane, in total, 14,592 (17.4%), 19,969 (23.8%), and 29,546 (35.2%) of the 83,826 genes had paralogs in two, three, and four sub-genomes, leaving 19,719 (23.5%) genes that were only found in one sub-genome (Fig. 3a). For the tetraploid potato and kiwiberry genome, gene retention was higher, with respectively 77.2% and 80.2% of the genes found in all four haplotypes (Supplementary Figs. 15, 16). This difference in gene retention between sugarcane and the two tetraploids can be attributed to their difference in genome sequencing levels (i.e., sub-genome vs haplotype) and ploidy level (i.e., octaploid vs tetraploid). Although less pronounced in potato and kiwiberry, the above numbers clearly indicate that extensive gene loss has already been ongoing in the three autopolyploid genomes. To better understand which genes are preferentially retained, we examined functional enrichment patterns associated with different levels of paralog retention (Supplementary Note 4, Supplementary Data 1, 2).

To investigate gene loss across the sub-genomes of sugarcane and the haplotypes in potato and kiwiberry, we first grouped paralogous genes retained in a collinear segment into so-called “collinear gene groups”, each consisting of all paralogs that are collinear across sub-genomes or haplotypes. In total, we identified 35,077, 38,320, and 35,445 nonredundant collinear gene groups in the sugarcane, potato, and kiwiberry genome, respectively (Supplementary Data 3). Each collinear gene group was annotated based on the presence or absence of genes across sub-genomes/haplotypes and labeled with the most likely evolutionary scenarios: “With collinear gene” (default), “With translocated gene”, “With collinear pseudogene”, and “With DNA-deleted gene” (Supplementary Fig. 17 and Methods). DNA-deleted genes are defined as those with no detectable collinear or translocated duplicate or collinear pseudogene, consistent with physical removal from the genome. Groups with unresolved scenarios were labeled as “With unclassified missing gene”, and singleton genes with no collinear paralogs were labeled as “Singleton gene”. A collinear gene group may have experienced different evolutionary scenarios on different sub-genomes/haplotypes, and thus may receive multiple labels.

Among 35,077 collinear gene groups in the sugarcane genome, we found that 6% (2213 groups) have at least one translocated gene. This is considerably lower than the 48% (16,864 groups) having at least one collinear pseudogene and 62% (21,886 groups) having at least one DNA-deleted gene (Fig. 3c). Similarly, the most frequently occurring collinear gene groups include at least one pseudogene and one deleted gene (8420 groups), followed by the collinear gene group with only DNA-deleted genes (7775 groups) and the one with only collinear pseudogenes (5554 groups) (Fig. 3b). Therefore, our results show that gene loss instead of gene translocation plays a significant role among the missing collinear genes in the auto-octoploid sugarcane genome. Moreover, the identified number of collinear pseudogenes (24,938 pseudogenes) and DNA-deleted genes (38,356 DNA-deleted genes) shows again that gene loss through DNA deletion is the most frequent gene loss mechanism, with an ~1.5 times higher frequency than pseudogenization (Fig. 3d).

Applying the same labeling strategy, the potato and kiwiberry genomes also showed that gene loss through deletion occurs more often than pseudogenization, with a ratio of 1.8 (30,600 DNA-deleted genes/16,767 pseudogenes) in potato and 1.3 (30,352 DNA-deleted genes/23,869 pseudogenes) in kiwiberry (Supplementary Figs. 18, 19). Different from the sugarcane genome, the two haplotype-resolved genomes of potato and kiwiberry show higher collinear gene retention as reflected by not only the collinear alignment (Supplementary Figs. 20–22) but also the most frequent collinear gene group having only collinear genes and no loss or translocation (Supplementary Figs. 18, 19).

To further corroborate that putative deleted genes were indeed lost via DNA deletions, we checked if neighboring paralogs of the corresponding retained gene also tend to get lost, as DNA deletion may not necessarily only affect one gene but may affect a group of neighboring genes. In sugarcane, genes with corresponding deleted genes are clustered together with an average of 1.90 adjacent genes with corresponding deleted genes, indicating these genes are lost together, likely through DNA deletion. In contrast, for randomly selected genes an average of 0.48 adjacent genes with corresponding deleted genes is found (Supplementary Fig. 23). Similarly, for potato

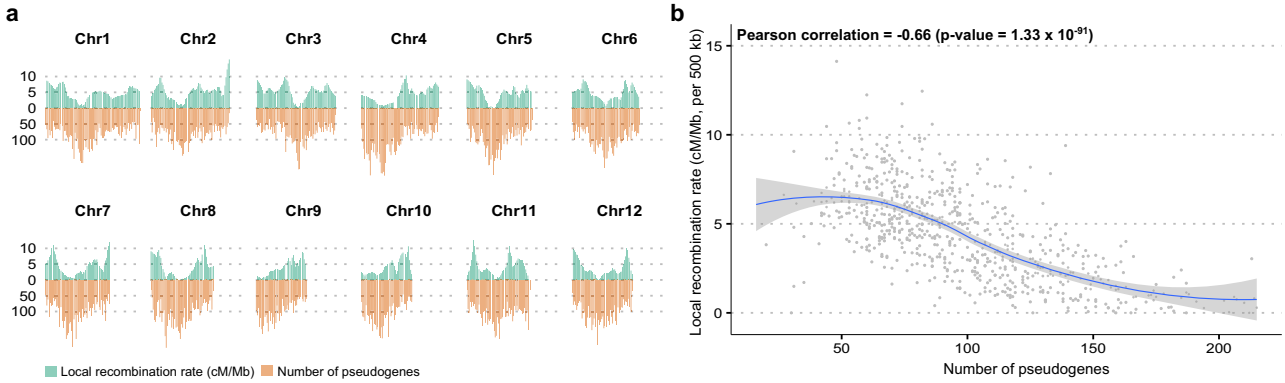

Fig. 4 | Relationship between recombination rate and pseudogene number in *Oryza sativa*. **a** Distribution of local recombination rate (green), expressed in centimorgan per megabase (cM/Mb), and pseudogenes (orange) across the different chromosomes of *O. sativa*. **b** Number of pseudogenes in function of local recombination rate in *O. sativa* (cM/Mb). The blue line represents the smoothed conditional mean, with the gray shaded area around the regression line delineating the 95% confidence interval. Two-sided correlation tests were conducted to evaluate statistical significance. The average recombination rate and the number of pseudogenes were calculated per window of 500 kb. Underlying data files on Zenodo[107].

and kiwiberry, we find a significant number of adjacent genes with corresponding deleted genes, with 0.38 (compared to 0.16 in random) in kiwiberry and 4.58 (compared to 0.19 in random) in potato (Supplementary Figs. 24, 25). The varied number of adjacent genes may suggest that the average size of deletions differs across species.

Together, although the degradation of collinear segments and the lifespan of pseudogenes influence pseudogene identification, our findings demonstrate that DNA deletion, rather than pseudogenization, primarily drives gene loss following WGMs in both paleo- and neo-polyploid angiosperms. Since neo-polyploid genomes already exhibit a bias towards DNA deletion, a direct one-step gene deletion, without a pseudogene intermediate, seems to be a likely and frequent mechanism. That said, the relative contribution of DNA deletion versus pseudogenization may differ across plant lineages. For example, in ferns, pseudogenization likely plays a more significant role in gene loss, while DNA deletion is far less prevalent[4,64]. Moreover, although pseudogenes could eventually be eliminated through DNA deletion, those that persist may still provide an evolutionary spandrel that contributes to novel biological functions[34,39,40,45].

WGM-derived pseudogenes, unlike other pseudogenes, are enriched in regions of high recombination

Because most pseudogenes are not under purifying selection, they are likely removed by recombination in regions of high recombination[65]. This leads to the expectation of a negative correlation between local recombination rate and pseudogene abundance. A previous study has confirmed this pattern in *G. max*[36]. Here, we repeated and extended such analysis for *A. thaliana*, *G. max*, *O. sativa*, and *Z. mays*. We find a significant negative Pearson correlation between recombination rate and total pseudogene number in *A. thaliana* ($r = -0.35$, $p = 4.18 \times 10^{-8}$), *G. max* ($r = -0.58$, $p = 1.4 \times 10^{-107}$), and *O. sativa* ($r = -0.66$, $p = 1.33 \times 10^{-91}$) (Fig. 4, Supplementary Figs. 26, 27, and Supplementary Table 8). In contrast, *Z. mays* shows no significant correlation ($r = -0.05$, p-value = 0.12, Supplementary Fig. 28), likely due to its unusually uniform distribution of pseudogenes. This pattern is probably linked to a recent burst of LTR retrotransposons in *Z. mays*[66], which also resulted in a uniform distribution of LTR retrotransposons (Supplementary Fig. 29) and led to the creation of retro-transposed pseudogenes along the genome. Across all four species, a significant positive Pearson correlation exists between pseudogene number and LTR retrotransposon number, indicating that a substantial portion of pseudogenes arose through retro-transposition (Supplementary Table 9). In *Z. mays*, this has obscured the expected correlation between recombination and total pseudogenes, as recombination may not have removed them (yet).

In contrast to other pseudogene classes, we found a significant positive Pearson correlation between recombination rate and WGM-derived pseudogene number in *G. max* ($r = 0.45$, $p = 5.54 \times 10^{-59}$), *O. sativa* ($r = 0.24$, $p = 1.78 \times 10^{-11}$) and *Z. mays* ($r = 0.13$, $p = 4.23 \times 10^{-6}$) (Supplementary Table 8). *A. thaliana* was the exception, likely due to the sparse distribution of WGM-derived pseudogenes across its compact genome. The occurrence of WGM-derived pseudogenes in regions of high recombination may help explain its overall scarcity. At first glance, this pattern may seem surprising given that high recombination regions are also hotspots for DNA deletion. However, consistent with the nature of WGMs, WGM-derived pseudogenes arise in regions of high gene density that also experience high recombination. In contrast, other classes of pseudogenes, such as those originating from retrotransposons, are often found in regions of low recombination, where they are more likely to escape recombinative DNA deletion and persist. This difference in genomic context helps account for the contrasting abundance of different pseudogene types. Moreover, considering that the persistent WGM-derived pseudogenes are retained in collinear segments alongside functional genes, which also have a positive correlation with recombination[67], our results would suggest that these WGM-derived pseudogenes may be under stronger selective constraints compared to other pseudogene classes.

Based on the above findings and the scarcity of WGM-derived pseudogenes, we propose a model, where WGM creates a transient genomic environment with an elevated recombination rate, as has been observed in both allo- and auto-polyploids[68–70]. This elevated recombination rate leads to the accelerated removal of redundant duplicated genes, especially in the early phase following WGM. While the exact cause of this elevated recombination rate remains unclear, the phenomenon is specific to WGM, because SSDs, in contrast, generate a proportionally high number of pseudogenes. SSDs produce scattered duplicated genes, which are likely less exposed to recombination, limiting their removal via recombinative DNA deletion and thereby allowing the accumulation of pseudogenes. This distinction highlights the unique consequences of WGM on genome evolution. Once the genome re-establishes a diploid-like state, the elevated recombination rate may decline. Interestingly, even in polyploid genomes, there is evidence of selection acting to reduce recombination. Indeed, natural polyploid populations of *A. arenosa* have evolved lower recombination rates through selection on meiosis-related genes[71,72]. Similarly, in *Brassica napus*, the loss of *MSH4* duplicates has been shown to reduce nonhomologous crossovers and enhance meiotic stability[73]. These observations suggest selective pressure

https://doi.org/10.1038/s41467-025-61676-3

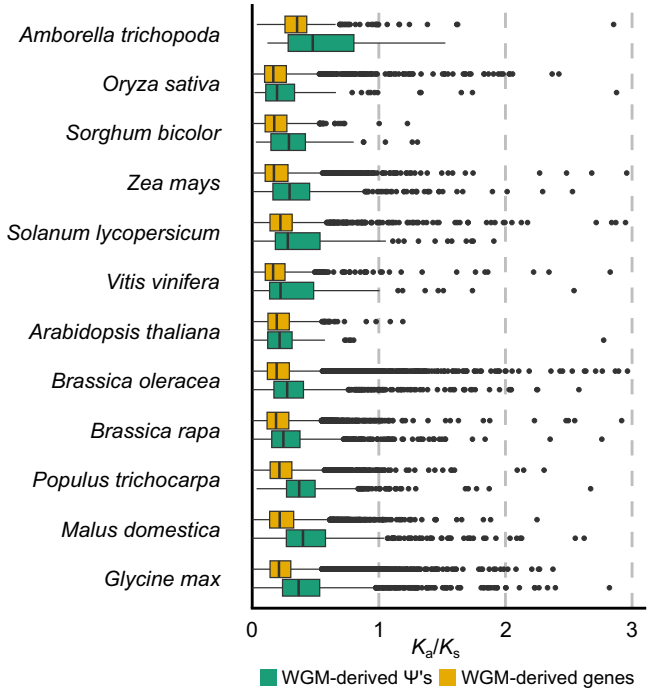

Fig. 5 | Ratio of non-synonymous substitutions per non-synonymous site (K_a) to synonymous substitutions per synonymous site (K_s) of whole genome multiplication (WGM)-derived gene pairs and pseudogenes. To obtain the K_a/K_s ratio for pseudogenes, they were aligned to their functional paralogous gene. Boxplots were generated for WGM-derived pseudogenes (green; *A. trichopoda*: $n = 72$, *A. thaliana*: $n = 84$, *B. oleracea*: $n = 1466$, *B. rapa*: $n = 1516$, *G. max*: $n = 1704$, *M. domestica*: $n = 1015$, *O. sativa*: $n = 166$, *P. trichocarpa*: $n = 1364$, *S. lycopersicum*: $n = 157$, *S. bicolor*: $n = 53$, *V. vinifera*: $n = 130$, and *Z. mays*: $n = 574$) and for WGM-derived duplicate genes (yellow; *A. trichopoda*: $n = 273$, *A. thaliana*: $n = 5356$, *B. oleracea*: $n = 16{,}698$, *B. rapa*: $n = 20{,}766$, *G. max*: $n = 34{,}564$, *M. domestica*: $n = 11{,}080$, *O. sativa*: $n = 2267$, *P. trichocarpa*: $n = 12{,}906$, *S. lycopersicum*: $n = 5030$, *S. bicolor*: $n = 2359$, *V. vinifera*: $n = 1850$, and *Z. mays*: $n = 3925$). The center, upper and lower bound of the boxplots correspond, respectively, to the median, 25th and 75th percentiles. The bounds of the whiskers of the boxplots extend to 1.5 times the inter-quartile range starting from the bounds of the boxplots. Data beyond the end of the whiskers are referred to as outlying points and are shown individually. Underlying data files on Zenodo[107].

against the initially elevated recombination rate induced by WGM and confirm its transient state.

Pseudogenes exhibit a non-neutral K_a/K_s ratio

To assess the selective constraints on WGM-derived pseudogenes, we estimated the ratio of non-synonymous substitutions per non-synonymous site (K_a) to synonymous substitutions per synonymous site (K_s). In theory, pseudogenes evolving neutrally should have a K_a/K_s ratio close to 1, while functional genes are typically under purifying selection and thus have a K_a/K_s ratio lower than 1. Whereas WGM-derived pseudogenes do exhibit a higher K_a/K_s ratio than that of functional paralogous genes, their K_a/K_s ratios remain below 1, suggesting some level of selective constraint or bias (Fig. 5 and Supplementary Fig. 30). This deviation from the neutral expectation ($K_a/K_s \approx 1$) has also been observed in previous studies, e.g., in rice and barley[32,33,35]. Several factors may explain this: (1) synonymous substitutions continue to accumulate in the functional paralogous gene, inflating the denominator of the K_a/K_s ratio; and (2) prior to pseudogenization, the ancestral duplicate gene may have undergone purifying selection. The latter may be particularly relevant for WGM-derived pseudogenes, which originate from duplicated genes with the ancestral regulatory context and hence may retain functionality for some time following the WGM.

Nonetheless, it is possible that at least some WGM-derived pseudogenes have adopted a novel function and are under selective constraints. By definition, pseudogenes are expected to have no functionality because they lost their coding potential through detrimental mutations. However, non-functionality is not evaluated during the bioinformatic identification of pseudogenes, as it necessitates labor-intensive experimental verification. In fact, several studies have demonstrated that some pseudogenes are still transcribed, with a subset playing biological roles[34,39,40,45]. For example, pseudogenes have been proposed as an important source of lncRNAs, which govern regulatory functions[36,41–43].

To explore whether pseudogenes may contribute to regulatory functions via lncRNAs, we assessed the overlap between our predicted pseudogenes and annotated lncRNAs from the PLncDB v2.0[74] in *A. thaliana* and *S. lycopersicum*. We defined an overlap when at least half of a pseudogene and lncRNA sequence coincided. As such, we find that 297 and 178 of the pseudogenes overlap with lncRNAs in *A. thaliana* and *S. lycopersicum*, respectively (Supplementary Table 10). Using a shuffled empirical null model (Methods), these overlaps were statistically significant (p-value = 9.99×10^{-4} and 0.049, respectively, Supplementary Fig. 31). Moreover, in *S. lycopersicum*, filtering the fragmented pseudogenes strengthened this signal (with 106 overlaps, p-value = 9.99×10^{-4}, Supplementary Fig. 32), while the fragmented pseudogenes alone show no significance (with 72 overlaps, p-value = 0.99, Supplementary Fig. 33).

Although the absolute numbers of WGM-derived pseudogenes overlapping with lncRNAs were low, with 12 in *A. thaliana* and two in *S. lycopersicum*, these were still statistically significant when compared to the shuffled empirical null distribution (Supplementary Fig. 34). The low overlapping numbers are mainly due to the small overall number of WGM-derived pseudogenes in the two species. For SSD-derived pseudogenes, *A. thaliana* and *S. lycopersicum* have 57 and 75 statistically significant pseudogene-lncRNA overlaps, respectively (Supplementary Fig. 35), while retro-transposed pseudogenes do not show significant overlaps in either species (Supplementary Fig. 36). Together, these results suggest that pseudogenes, in particular the ones originating from (WGM and SSD) duplications, may serve as a source of lncRNAs with potential regulatory roles.

In this study, we investigated the role of pseudogenization during rediploidization following WGM across 12 paleo-polyploid and three neo-polyploid angiosperm genomes with varying timings of WGMs. We identified pseudogenes arising from WGMs and found that their numbers were consistently lower than expected under a null hypothesis assuming equal contributions from pseudogenization and DNA deletion to gene loss after WGM. After accounting for potential identification biases that could cause a low number of pseudogenes, our analyses reject the null hypothesis but embrace a subordinate role of pseudogenization to DNA deletion following WGMs in the studied angiosperms. Based on this, we propose a model in which WGM creates a transient genomic environment with an elevated recombination rate, leading to an increase in DNA deletion. Notably, despite the limited number of WGM-derived pseudogenes, the persisting ones may still govern functional potential, as indicated by their non-neutral K_a/K_s ratios and significant overlaps with lncRNAs.

Methods

Data retrieval

We retrieved 12 species that differed in the occurrence and timing of WGMs: *Amborella trichopoda*, *Arabidopsis thaliana*, *Brassica oleracea*, *Brassica rapa*, *Glycine max*, *Malus domestica*, *Oryza sativa*, *Populus trichocarpa*, *Solanum lycopersicum*, *Sorghum bicolor*, *Vitis vinifera*, and *Zea mays*. The genomes of the analyzed paleo-polyploid species were retrieved from PLAZA v5.0[75]. The timings of the WGMs were obtained from Vanneste et al.[50] for all species except *Vitis vinifera*, which was obtained from Jiao et al.[76] (Supplementary Table 1). For *Amborella*

trichopoda, we refrained from providing an estimated timing due to the ongoing controversy regarding whether a WGD occurred before the divergence of angiosperms, with *A. trichopoda* being the earliest extant diverging lineage[77–79]. A species tree depicting the relationships between the paleo-polyploid species was generated using TimeTree 5[80]. Timings of WGMs were annotated on the tree using shinyWGD (https://github.com/li081766/shinyWGD).

The assembly and annotation of the haploid auto-octoploid *Saccharum spontaneum* AP85-441 genome[59], and two haplotype-resolved auto-tetraploid genomes of *Actinidia arguta*[61] and *Solanum tuberosum* Otava[60] are available on NCBI (GCA_003544955.1), Genome Warehouse (accession number GWHBJWW00000000), and SpudDB (Otava v1 assembly), respectively. Timings of WGMs were summarized in Supplementary Table 1.

Pseudogene identification in genomes of paleo-polyploid species

First, repeats were identified in the genomes of the paleo-polyploid species and masked using RepeatMasker v4.1.1[81]. Next, putative pseudogenes were identified using the publicly available PseudoPipe pipeline[47]. Briefly, PseudoPipe attempts to identify pseudogenes by searching for homologous sequences of functional genes in the non-coding part of the genome. The proteome of a species is searched against the repeat- and exon-masked genome of that species using TBLASTN[82] with an *E*-value cut-off of 1×10^{-4} and further processed (see ref. 47 for further details). Pseudogenes are classified as (1) "retrotransposed (or processed)" if they lack introns and possess a poly-A tail, (2) "duplicated" if they comprise multiple exons and lack a poly-A tail, and (3) "fragmented" if they show sequence similarity to a gene but fail to meet the criteria for processed and duplicated pseudogenes, or fall below the thresholds of the other two classes. Both processed and duplicated pseudogenes must exhibit a minimum of 40% protein identity with the corresponding functional paralogous protein and an *E*-value lower than 1×10^{-10}. Moreover, processed pseudogenes are specifically required to align more than 70% with the functional protein. After running PseudoPipe, some further post-processing was performed. Putative pseudogenes with more than 30 bps overlap with an exon were removed and pseudogenes were filtered based on some minimal criteria: they must align with at least 5% of the functional protein, have a minimum of 20% protein identity, and an *E*-value smaller or equal to 1×10^{-5}. Furthermore, in cases where pseudogenes exhibited an overlap of more than 30 bps, only the best pseudogene hit was preserved, determined on the following order of priority: *E*-value, alignment coverage, and protein percent identity. Subsequently, i-ADHoRe v3.0[83] was executed to identify regions within the genome with similar gene content and order, i.e., collinear segments. As input for i-ADHoRe, we extracted gene-pseudogene pairs from the duplicated and fragmented classes of PseudoPipe, filtering out intronic pseudogenes as these do not likely originate from a WGM. In addition, gene pairs were extracted from an all-vs-all BLASTP of BLAST + v2.16.0 suite[82], whereby the following criteria were met: a maximum *E*-value of 1×10^{-5}, an alignment coverage of at least 30%, and a difference in length no more than the length of the smallest protein. In order to remain consistent and avoid biases due to parameters, we ran i-ADHoRe using the following parameters for all species: (1) A minimum number of three anchorpairs, i.e., pairs of putative homologous genes and/or pseudogenes in a collinear segment, (2) a maximum gap size that should exist between two points of 35, (3) a maximum cluster gap size of 40, (4) maximum number of gaps in alignment of 40, (5) a probability cut-off of 0.01, (6) a *q*-value cut-off of 0.75, and (7) disabled level 2 only. Pseudogenes that form homologous pairs with their functional paralogous gene within collinear segments were considered to be WGM-derived. Using this approach, we were able to identify pseudogenes both in recent and ancient collinear segments for the different species. Although species that experienced multiple recent WGMs may also show collinear segments from older WGMs, this does not significantly affect our findings. We evaluated this by constructing K_s-age distributions using the median K_s values of paralogs retained in collinear segments where pseudogenes were found. This analysis revealed that the vast majority of identified collinear segments indeed stemmed from the most recent WGM (Supplementary Fig. 37). Collinear segments were visually inspected using GenoPlotR v0.8.11[84] in R.

Fitting single- and double-exponential functions to WGM-derived numbers

To explore the evolution through time of the number of WGM-derived pseudogenes and gene pairs, these quantities were graphed against the timing of the most recent WGM, utilizing both absolute (as million years ago) and relative (as K_s) timing of the most recent WGM in each species. Subsequently, we fitted a single- and double-exponential function to these datasets using the lmfit library v1.3.0[85] in Python. The single-exponential function corresponds to the simple passive or random loss model whereby genes or pseudogenes are lost randomly because there is no selection to retain them[86]. The double-exponential function approximates the two-phase model described by Inoue et al.[49]. This model comprises an initial rapid phase, characterized by the simultaneous loss of multiple genes, followed by a second phase with a lower rate of loss. Regression fits were evaluated and compared through r^2, AIC, BIC, and root-mean-square error statistics. Furthermore, we conducted a regression analysis excluding the two *Brassica* species and *Zea mays* to confirm that our results were not biased by species that underwent multiple recent WGMs. We observed the same patterns as when these species were included, indicating that our overall findings are unlikely to be significantly affected by multiple ancient WGMs (Supplementary Fig. 38).

Simulation of pseudogene evolution

To understand the expected evolution of genes without selective constraints, we conducted simulations. These simulations represent the situation where redundant genes are immediately released from selective constraints and can freely accumulate mutations at the spontaneous substitution rate after a WGM. These simulations were performed using the phylogenetic sequence simulator AliSim of IQ-TREE v2.2.2.6[87]. For this, we used the spontaneous substitution rate of 7×10^{-9} substitutions per site per generation of 1 year and the rate of 1- to 3-bp deletions as 0.6×10^{-9} per site per year and the rate of 1- to 3-bp insertions as 0.3×10^{-9} per site per year estimated for *Arabidopsis thaliana*[53]. In addition, we explored multiples of this substitution rate, e.g., twice (×2), ten times (×10), and one tenth (1/10), along with the estimated spontaneous substitution rate of *Populus trichocarpa* as 1.33×10^{-10} per site per year[55]. Employing 1× the spontaneous mutation rate represents the case whereby a pseudogene evolves neutrally, while the functional gene remains identical to the ancestral gene. As the functional gene will also accumulate mutations, albeit under selection, employing 1× the spontaneous mutation rate provides a lower boundary of expected divergence. Employing 2× the spontaneous mutation rate leads to the pseudogene evolving twice as fast from the ancestral sequence compared to the substitution rate. This is equivalent to having two sequences that evolve at the spontaneous mutation rate, and thus both genes evolving as neutral pseudogenes. Consequently, this serves as an upper boundary of expected divergence. The evolution of all current CDS sequences of *A. thaliana* was simulated for 100 million years in steps of 1 million year. Each step consisted of a simulation along a simple two-branch tree, using a GTR model, taking into account the %GC of *A. thaliana* (36%) and following a Lavalette distribution (with parameter $a = 0.5$ and max = 3) for both insertion and deletion size. The used branch length represents the expected amount of substitutions per site. For example, a branch length of 0.007 was used to represent the spontaneous substitutions per site per million years for *A. thaliana*. For each step, the obtained

simulated sequence of the previous simulation was used as input. After each simulation step, the simulated sequence was aligned to the original translated CDS sequence using tfasty from the FASTA v36.3.8 d suite[88]. This is the aligner that was also used within PseudoPipe for realignment. By using the same aligner, we ensure that the observed differences between reality and simulation are not due to a different alignment method. For the obtained alignment, summary statistics were calculated: percent identity between the evolved and ancestral protein, percentage of the ancestral functional protein that is aligned by the evolved protein, *E*-value of the alignment between the evolved and ancestral protein, number of frameshifts in the simulated DNA sequence, and the number of stop codons in the simulated DNA sequence. Furthermore, we applied the cut-offs that were used for paleo-polyploids (see above) to infer the detectability of pseudogenes over time.

Pseudogene identification in the neo-polyploid genomes

For the neo-polyploid genomes, collinear segments were inferred between sub-genomes using i-ADHoRe v3.0[83], with gene pairs extracted from an all-vs-all BLASTP run as input (*E*-value < 1×10^{-10}, alignment coverage ≥ 30%, and protein percent identity ≥ 40%). We utilized the same parameters for i-ADHoRe v3.0 as for the paleo-polyploids (see above), except that only collinear segments of level 2 were inferred. A riparian plot was generated to show the syntenic relationships of genes using GENESPACE v1.2.3[89].

To evaluate the different evolutionary trajectories of genes across sub-genomes or haplotypes, we adopted a strategy with a labeling system as outlined in Supplementary Fig. 17. We grouped paralogous collinear genes into so-called "collinear gene groups". A collinear gene group consists of all paralogs that are collinear to each other, across sub-genomes or haplotypes. If no collinear paralogs were found for a gene, the gene was placed in its own separate group. We also refer to these groups as "collinear gene groups" because they were most likely collinear in all sub-genomes/haplotypes in the ancestor, but have undergone extensive gene loss. Each collinear gene group was annotated based on the presence of genes across the sub-genomes/haplotypes, and the most likely evolutionary scenarios explaining collinear gene absence. If a gene was missing in one or more sub-genomes but collinearity was maintained with other sub-genomes, we considered two possible causes: translocation or gene loss (through pseudogenization or DNA deletion).

First, all collinear gene groups were labeled as "With collinear gene" by default, while groups consisting of only one gene which had no collinearity with any other sub-genome were labeled as "Singleton gene". For example, in the sugarcane genome, about 3% (1002) of the collinear gene groups consisted of a single gene with no paralogs or collinear correspondence to any other gene.

Second, per collinear gene group, we searched for translocated genes. For paralogs to be classified as translocated genes, they must (1) have at least 60% protein percent identity and at least 70% alignment coverage to a gene in the group, outside of a collinear segment, and (2) be present in a corresponding translocated region identified by SyRI v1.6.3[90], using default parameters. The results of SyRI were plotted using plotsr v1.1.0[91]. Translocated genes were inferred both within and among collinear gene groups. Collinear gene groups for which all genes were identified as translocated genes corresponding to other groups, were considered as redundant and removed from further analyses. Collinear gene groups with at least one translocation event were labeled as "With translocated gene".

Third, if a group had at least one sub-genome with no corresponding collinear paralogous gene, we further searched for collinear pseudogenes. If the sub-genome with the missing gene did not maintain any collinearity with other sub-genomes in the group, we labeled such groups as "With unclassified missing gene". For example, in the sugarcane genome, about 20% (7034) of the collinear gene groups had at least one missing gene that could not be explained by one of the evolutionary scenarios. Collinear pseudogenes were identified as sequences similar to a gene from a collinear gene group within the noncoding regions of collinear segments, using exonerate v2.4.0[92] with the protein2genome model and a maximum intron size of 6000. Hereby, genic regions were masked using maskfasta of the BEDtools v2.31.1 suite[93]. Hits that match the same gene were merged if they were within 6000 bp, and only the hit with the longest length was retained. Next, the inferred hits were re-aligned with the corresponding gene using exonerate with the affine:local model. Hits with at least 20% protein percent identity and at least 20% alignment coverage were retained. MACSE v2.06[94] was then used to align each pseudogene hit to its functional paralogous gene to calculate summary statistics of the alignment (number of frameshift mutations, number of premature stop codons, the absence of a start or stop codon, alignment coverage, and protein percent identity). If a putative pseudogene retained its full functional potential (i.e., no frameshift mutations, no premature stop codons, and the presence of both start and stop codons) and exhibited more than 95% alignment coverage and protein identity with its functional paralog, it was considered an unannotated gene and added as a collinear paralogous gene in the group. Otherwise, such collinear groups were labeled as "With collinear pseudogene".

Fourth, if no paralogous genes and/or pseudogenes were found in a sub-genome and the genes in the collinear group were part of a collinear segment between the sub-genome that had a missing paralog, we concluded that the gene had likely been deleted from the genome by DNA deletion. The group was then labeled "With DNA-deleted gene". Please note that a collinear gene group may have experienced multiple evolutionary scenarios hence being assigned with multiple labels. To summarize the labels for collinear groups and genes, we generated barplots and upset plots using ggplot2 v3.5.2[95] and UpsetR v1.4.0[96] packages in R.

Finally, we checked whether DNA-deleted genes are clustered together more than expected by chance. For each gene with an inferred DNA-deleted paralog, we determined how many adjacent genes also had inferred DNA-deleted paralogs. An average was then calculated and compared to an empirical distribution, generated by randomly selecting an equal number of genes that had an inferred DNA-deleted paralog and counting how many adjacent genes had DNA-deleted paralogs. This process was repeated 1000 times to generate an empirical null distribution.

Link between pseudogenization and recombination rate

To explore the relationship between the distributions of pseudogenes and recombination rate across the genome, we obtained genetic maps and their corresponding physical maps for *A. thaliana*[97], *G. max*[98], *O. sativa*[99], and *Z. mays*[100]. Subsequently, we estimated local recombination rates using the MareyMap R package v1.3.7[101]. A similar approach was taken as Brazier and Glémin[67]. Hereby, a two-degree polynomial loess regression was fitted to the marey map (i.e., map showing the relationship between genetic and physical distances) to obtain recombination rate estimates, expressed in centimorgan per megabase. Hereby, the span was set to 0.2 and negative estimates were set to 0. Subsequently, the average recombination rate of the markers and number of pseudogenes were calculated per bin of 500 kb across the genome. Using these binned estimates, we evaluated the correlation between pseudogene number and recombination rate.

Calculation of evolutionary distances: K_a, K_s, and K_a/K_s

To get insight into pseudogene evolution, we calculated the number of non-synonymous substitutions per non-synonymous site (K_a), the number of synonymous substitutions per synonymous site (K_s) and the ratio between the two (K_a/K_s) for all pseudogenes. To do this, pseudogene nucleotide sequences were aligned to their functional

paralogous gene using MACSE v2.06[94]. MACSE can accommodate pseudogenes by recognizing them as less reliable sequences, assigning them a reduced penalty for frameshifts and stop codons in comparison to functional genes. Next, the alignments were converted to AXT format, and evolutionary distances were calculated using KaKs_Calculator v2.0[102] through model averaging (MA). Moreover, K_a, K_s, and K_a/K_s were also calculated for WGM-derived gene pairs.

Link between pseudogenes and lncRNAs

To investigate the relation between pseudogenes and lncRNAs, we assessed whether there was overlap between our inferred pseudogenes and lncRNAs in the genome of *Arabidopsis thaliana* and *Solanum lycopersicum* using the intersect function of the BEDtools v2.31.1 suite[93] within Python with the pybedtools v0.10.1 library[103]. For an overlap to be acknowledged, a minimum of 50% of the pseudogene and lncRNA needed to be contained in the overlap. The coordinates of the lncRNAs were obtained from the PLncDB database v2.0[74]. To evaluate whether the observed number of overlaps was significantly higher than expected by chance, we generated an empirical null distribution by randomly rearranging lncRNAs on the genome using the shuffle function of the pybedtools library[103], and determining the number of overlapping pseudogenes with these randomly shuffled lncRNAs. This process was carried out for 1000 iterations. Subsequently, *p*-values were computed for the observed counts by determining the fraction of iterations that exhibit an equal or more extreme value, plus one, compared to our observed number divided by the total number of iterations plus one. The assessment of overlap was conducted for all pseudogenes, as well as separately for each class of pseudogenes.

Functional assessment of gene loss

To assess whether groups of pseudogenes or lost genes were enriched in specific functional categories, GO enrichment analyses were performed. GO annotations were obtained from PLAZA v5.0[75] for the paleo-polyploid species. For the neo-polyploid genomes, GO terms were obtained from the closest related genome available on PLAZA and linked to the respective genome by identifying the best BLASTP hit for each gene, ensuring an *E*-value below 1×10^{-5}.

GO enrichment analyses were performed using the clusterProfiler v4.0 R package[104]. Hereby, all GO-annotated genes of that analysis were used as background. Pseudogenes were assumed to have the same functional category as their functional paralogous genes. GO enrichment analyses were performed for WGM-derived pseudogenes (1) in each species, (2) in all paleo-polyploid species combined, and (3) further stratified based on species that underwent its most recent WGM less than 35 MYA and those that underwent its most recent WGM more than 35 MYA.

For the neo-polyploid genomes, GO enrichment analyses were performed for pseudogenes. In addition, for sugarcane, GO enrichment analyses were performed for genes that were retained at different levels on the sub-genomes, i.e., retained on all four, three, two or one sub-genome(s). For potato and kiwiberry, it was not possible to focus on sub-genomes as they were sequenced to a haplotype level. Therefore, we focused exclusively on genes that were fully retained or reduced to a single copy.

Finally, for the neo-polyploid genomes, it was inferred whether genes were part of protein–protein interaction (PPI) networks (both functional associations and physical links) by annotating the proteomes of the neo-polyploid species using STRING 12.0[105]. We assessed whether genes that were retained at different levels were enriched in such PPI genes by performing hypergeometrical tests using the phyper function of the stats v4.5.0 R package[106].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The genome data of the paleo-polyploid species used in this study are available in the PLAZA v5.0 database[75] under accession codes atr (*Amborella trichopoda*), ath (*Arabidopsis thaliana*), bol (*Brassica oleracea*), bra (*Brassica rapa*), gma (*Glycine max*), mdo (*Malus domestica*), osa (*Oryza sativa*), ptr (*Populus trichocarpa*), sly (*Solanum lycopersicum*), sbi (*Sorghum bicolor*), vvi (*Vitis vinifera*), and zma (*Zea mays*) [https://bioinformatics.psb.ugent.be/plaza.dev/_dev_instances/feedback/download/download]. The *Saccharum spontaneum* AP85-441 genome data used in this study are available at NCBI under accession code GCA_003544955.1 [https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_003544955.1/]. The *Actinidia arguta* genome data used in this study are available at Genome Warehouse under accession code GWHBJWW00000000 [https://ngdc.cncb.ac.cn/gwh/Assembly/83526/show]. The *Solanum tuberosum* Otava genome data used in this study are available at SpudDB under accession code Otava v1 [https://spuddb.uga.edu/otava_potato_download.shtml]. Data underlying figures are provided on Zenodo (https://doi.org/10.5281/zenodo.15552828)[107]. Source file 1 in Zenodo contains the list of identified pseudogenes in twelve paleo-polyploid species.

Code availability

The code used in this study is available on GitHub (https://github.com/ewoutcrombez/PseudogeneProject_scripts) and Zenodo (https://doi.org/10.5281/zenodo.15528393)[108].

References

1. Qiao, X., Zhang, S. & Paterson, A. H. Pervasive genome duplications across the plant tree of life and their links to major evolutionary innovations and transitions. *Comput. Struct. Biotechnol. J.* **20**, 3248–3256 (2022).
2. Wood, T. E. et al. The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. USA* **106**, 13875–13879 (2009).
3. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
4. Li, Z. et al. Patterns and processes of diploidization in land plants. *Annu. Rev.* **72**, 387–410 (2021).
5. Dodsworth, S., Chase, M. W. & Leitch, A. R. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Bot. J. Linn. Soc.* **180**, 1–5 (2016).
6. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141 (2005).
7. Soltis, P. S., Marchant, D. B., Van de Peer, Y. & Soltis, D. E. Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* **35**, 119–125 (2015).
8. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
9. Wendel, J. F. The wondrous cycles of polyploidy in plants. *Am. J. Bot.* **102**, 1753–1756 (2015).
10. Mandáková, T. & Lysak, M. A. Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* **42**, 55–65 (2018).
11. Doyle, J. J. & Coate, J. E. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *Int. J. Plant Sci.* **180**, 1–52 (2019).
12. Hufton, A. L. & Panopoulou, G. Polyploidy and genome restructuring: a variety of outcomes. *Curr. Opin. Genet. Dev.* **19**, 600–606 (2009).
13. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science*. **290**, 1151–1155 (2000).
14. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).
15. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev.* **11**, 97–108 (2010).

16. Kimura, M. & King, J. L. Fixation of a deleterious allele at one of two 'duplicate' loci by mutation pressure and random drift. *Proc. Natl. Acad. Sci. USA* **76**, 2858–2861 (1979).
17. Ohno, S. *Evolution by Gene Duplication*. *Evolution by Gene Duplication* (Springer, 1970).
18. Hahn, M. W. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* **100**, 605–617 (2009).
19. Conant, G. C., Birchler, J. A. & Pires, J. C. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* **19**, 91–98 (2014).
20. Li, Z. et al. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28**, 326–344 (2016).
21. Tasdighian, S. et al. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *Plant Cell* **29**, 2766–2785 (2017).
22. Maere, S. et al. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 5454–5459 (2005).
23. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
24. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* https://doi.org/10.1101/gr.132102 (2002).
25. Freeling, M., Scanlon, M. J. & Fowler, J. F. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.* **35**, 110–118 (2015).
26. Freeling, M. et al. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr. Opin. Plant Biol.* **15**, 131–139 (2012).
27. Woodhouse, M. R. et al. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol* **8**, e1000409 (2010).
28. Cheetham, S., Faulkner, G. & Dinger, M. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat. Rev. Genet.* **21**, 191–201 (2020).
29. Schrider, D. R., Costello, J. C. & Hahn, M. W. All human-specific gene losses are present in the genome as pseudogenes. *J. Comput. Biol.* **16**, 1419–1427 (2009).
30. Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
31. Zhong, Y. et al. Genomic insights into genetic diploidization in the homosporous fern *Adiantum nelumboides*. *Genome Biol. Evol.* **14**, evac127 (2022).
32. Mascagni, F., Usai, G., Cavallini, A. & Porceddu, A. Structural characterization and duplication modes of pseudogenes in plants. *Nat. Sci. Rep.* **11**, 5292 (2021).
33. Prade, V. M. et al. The pseudogenes of barley. *Plant J.* **93**, 502–514 (2018).
34. Zou, C. et al. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.* **151**, 3–15 (2009).
35. Thibaud-Nissen, F., Ouyang, S. & Buell, C. R. Identification and characterization of pseudogenes in the rice gene complement. *BMC Genom.* **10**, 317 (2009).
36. Xie, J. et al. Evolutionary origins of pseudogenes and their association with regulatory sequences in plants. *Plant Cell* **31**, 563–578 (2019).
37. Hoopes, G. et al. Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Mol. Plant* **15**, 520–536 (2022).
38. Olson, M. V. When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18 (1999).
39. Li, W., Yang, W. & Wang, X.-J. Pseudogenes: pseudo or real functional elements? *J. Genet. Genom.* **40**, 171–177 (2013).
40. Guo, X., Zhang, Z., Gerstein, M. B. & Zheng, D. Small RNAs originated from pseudogenes: cis- or trans-acting? *PLoS Comput. Biol.* **5**, e1000449 (2009).
41. Milligan, M. J. & Lipovich, L. Pseudogene-derived lncRNAs: Emerging regulators of gene expression. *Front. Genet.* **4**, 117941 (2015).
42. Wen, Z. Y., Kang, Y. J., Ke, L., Yang, D. C. & Gao, G. Genome-wide identification of gene loss events suggests loss relics as a potential source of functional lncRNAs in humans. *Mol. Biol. Evol.* **40**, msad103 (2023).
43. Liu, W. H., Tsai, Z. T. Y. & Tsai, H. K. Comparative genomic analyses highlight the contribution of pseudogenized protein-coding genes to human lincRNAs. *BMC Genom.* **18**, 786 (2017).
44. Pink, R. C. & Carter, D. R. F. Pseudogenes as regulators of biological function. *Essays Biochem.* **54**, 103–112 (2013).
45. Zheng, D. & Gerstein, M. B. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.* **23**, 219–224(2007).
46. Esfeld, K. et al. Pseudogenization and Resurrection of a speciation gene. *Curr. Biol.* **28**, 3776–3786.e7 (2018).
47. Zhang, Z. et al. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–1439 (2006).
48. Blanc, G. & Wolfe, K. H. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**, 1679–1691 (2004).
49. Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K. & Nishida, M. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc. Natl. Acad. Sci. USA* **112**, 14918–14923 (2015).
50. 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.* **24**, 1334–1347 (2014).
51. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
52. Li, W. H., Gojobori, T. & Nei, M. Pseudogenes as a paradigm of neutral evolution. *Nature* **292**, 237–239 (1981).
53. Ossowski, S. et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
54. Kerfeld, C. A. & Scott, K. M. Using BLAST to teach "*E*-value-tionary" concepts. *PLoS Biol.* **9**, e1001014 (2011).
55. Hofmeister, B. T. et al. A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial *Populus trichocarpa*. *Genome Biol.* **21**, 1–27 (2020).
56. De La Torre, A. R., Li, Z., Van De Peer, Y. & Ingvarsson, P. K. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.* **34**, 1363–1377 (2017).
57. Van de Peer, Y. Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* **5**, 752–763 (2004).
58. Zhang, J. et al. Recent polyploidization events in three Saccharum founding species. *Plant Biotechnol. J.* **17**, 264–274 (2019).
59. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum L*. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0237-2 (2018).
60. Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).
61. Lu, X. M. et al. Genome assembly of autotetraploid *Actinidia arguta* highlights adaptive evolution and enables dissection of important economic traits. *Plant Commun.* **5**, 100856 (2024).

62. Michael, T. P. et al. Triploidy is prominent in the duckweed Lemna minor complex. *bioRxiv* 2025.02.18.638736 https://doi.org/10.1101/2025.02.18.638736 (2025).
63. Khan, A. W. et al. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant Sci.* **25**, 148–158 (2020).
64. Nakazato, T., Barker, M. S., Rieseberg, L. H. & Gastony, G. J. Evolution of the nuclear genome of ferns and lycophytes. in *Biology and Evolution of Ferns and Lycophytes* 175–198 https://doi.org/10.1017/CBO9780511541827.008 (Cambridge University Press, 2008).
65. Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J. & Anderson, L. K. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.* **8**, 77–84 (2007).
66. Stitzer, M. C., Anderson, S. N., Springer, N. M. & RossIbarra, J. The genomic ecosystem of transposable elements in maize. *PLoS Genet.* **17**, e1009768 (2021).
67. Brazier, T. & Glémin, S. Diversity and determinants of recombination landscapes in flowering plants. *PLoS Genet.* **18**, e1010141 (2022).
68. Leflon, M. et al. Crossovers get a boost in *Brassica Allotriploid* and Allotetraploid hybrids. *Plant Cell* **22**, 2253–2264 (2010).
69. Desai, A., Chee, P. W., Rong, J., May, O. L. & Paterson, A. H. Chromosome structural changes in diploid and tetraploid A genomes of Gossypium. *Genome* **49**, 336–345 (2006).
70. Pecinka, A., Fang, W., Rehmsmeier, M., Levy, A. A. & Mittelsten Scheid, O. Polyploidization increases meiotic recombination frequency in Arabidopsis. *BMC Biol.* **9**, 1–7 (2011).
71. Lloyd, A. & Bomblies, K. Meiosis in autopolyploid and allopolyploid Arabidopsis. *Curr. Opin. Plant Biol.* **30**, 116–122 (2016).
72. Yant, L. et al. Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Curr. Biol.* **23**, 2151–2156 (2013).
73. Gonzalo, A. et al. Reducing MSH4 copy number prevents meiotic crossovers between non-homologous chromosomes in Brassica napus. *Nat. Commun.* **10**, 1–9 (2019).
74. Jin, J. et al. PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. *Nucleic Acids Res.* **49**, D1489–D1495 (2021).
75. Van Bel, M. et al. PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Res.* **50**, D1468–D1474 (2022).
76. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, 1–14 (2012).
77. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
78. Ruprecht, C. et al. Revisiting ancestral polyploidy in plants. *Sci. Adv.* **3**, e1603195 (2017).
79. Zwaenepoel, A. & Van de Peer, Y. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol. Biol. Evol.* **36**, 1384–1404 (2019).
80. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).
81. Chen, N. Using repeatmasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **5**, 4.10.1–4.10.14 (2004).
82. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
83. Proost, S. et al. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11–e11 (2012).
84. Guy, L., Kultima, J. R., Andersson, S. G. E. & Quackenbush, J. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
85. Newville, M. et al. Lmfit: non-linear least-square minimization and curve-fitting for Python (Astrophysics Source Code Library 2016).
86. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345 (2006).
87. Ly-Trong, N., Naser-Khdour, S., Lanfear, R. & Minh, B. Q. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. *Mol. Biol. Evol.* **39**, msac092 (2022).
88. Pearson, W. R. Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinform.* **53**, 3.9.1–3.9.25 (2016).
89. Lovell, J. T. et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11**, e78526 (2022).
90. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 1–13 (2019).
91. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
92. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 1–11 (2005).
93. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
94. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**, 2582–2584 (2018).
95. Wickham, H. *ggplot2: Elegant Graphics For Data Analysis* (Springer, 2009).
96. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
97. Serin, E. A. R. et al. Construction of a high-density genetic map from RNA-Seq data for an Arabidopsis bay-0 × Shahdara RIL population. *Front. Genet.* **8**, 310192 (2017).
98. Grant, D. & Nelson, R. T. SoyBase: a comprehensive database for soybean genetic and genomic data 193–211. in (eds Nguyen, H. & Bhattacharyya, M.) *The Soybean Genome. Compendium of Plant Genomes* (Springer, 2017).
99. De Leon, T. B., Linscombe, S. & Subudhi, P. K. Molecular dissection of seedling salinity tolerance in rice (Oryza sativa L.) using a high-density GBS-based SNP linkage map. *Rice* **9**, 1–22 (2016).
100. Kianian, P. M. A. et al. High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nat. Commun.* **9**, 1–10 (2018).
101. Rezvoy, C., Charif, D., Guéguen, L. & Marais, G. A. B. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* **23**, 2188–2189 (2007).
102. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteom. Bioinform.* **8**, 77–80 (2010).
103. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
104. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* https://doi.org/10.1016/j.xinn.2021.100141 (2021).
105. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
106. R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org (2025).

107. Crombez, E., Van de Peer, Y. & Li, Z. The subordinate role of pseudogenization to recombinative deletion following polyploidization in angiosperms. *Zenodo (Source Data)* https://doi.org/10.5281/zenodo.15552828 (2025).
108. Crombez, E., Van de Peer, Y. & Li, Z. The subordinate role of pseudogenization to recombinative Deletion following polyploidization in angiosperms. *Zenodo (ewoutcrombez/PseudogeneProject_scripts: Release for Paper Publication)* https://doi.org/10.5281/zenodo.15528393 (2025).

Acknowledgments

This work was supported by the European Research Council under the European Union's Horizon 2020 Research and Innovation program (No. 833522) and by Ghent University (Methusalem funding, BOF.-MET.2021.0005.01) (to Y.V.d.P.). Z.L. acknowledges funding from the Junior Research Project of FWO (G0ADO25N) and the Special Research Grant from Ghent University (BOF.BAF.2024.0889.01).

Author contributions

E.C., Z.L., and Y.V.d.P. conceived and managed the project. E.C. conducted the genomic analyses of this study. E.C. and Z.L. wrote the manuscript with the support of Y.V.d.P. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-61676-3.

Correspondence and requests for materials should be addressed to Ewout Crombez, Yves Van de Peer or Zhen Li.

Peer review information *Nature Communications* thanks Shifeng Cheng, Jeffrey Doyle and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.