A super-pangenome framework of the genus Glycine unveils polyploid evolution and life-strategy transition

Yongbin Zhuang
Shandong Agricultural University
Xutong Wang
Purdue University https://orcid.org/0000-0002-4625-7659
Xianchong Li
Shandong Agricultural University
Junmei Hu
Shandong Agricultural University
Lichuan Fan
Shandong Agricultural University
Scott Jackson
University of Georgia
Jeffrey Doyle
Cornell University
Xian Sheng Zhang
Shandong Agricultural University
Dajian Zhang
Shandong Agricultural University
Jianxin Ma (maj@purdue.edu)
Purdue University https://orcid.org/0000-0002-1474-812X

Article

Keywords: Genus Glycine, life strategy transition, phylogenomics, polyploid genome evolution, super-pangenome

Posted Date: May 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-548382/v1

Version of Record: A version of this preprint was published at Nature Plants on March 14th, 2022. See the published version at https://doi.org/10.1038/s41477-022-01102-4.

Abstract

Polyploidy and life strategy transitions between annuality and perenniality often occur in flowering plants. However, the evolutionary propensities of polyploids and genetic bases of such transitions remain elusive. We assembled plantum genomes of representative perennial species across the genus Glycine including five diploids and a young allopolyploid, and constructed a Glycine super-pangenome framework by integrating 26 annual soybean genomes. The perennials exhibited greater genome stability than the annuals, with few centromere repeats abundant in the latter. Biased subgenome fractionation has occurred in the allopolyploid, primarily by accumulation of small deletions in gene clusters through illegitimate recombination, which was associated with preexisting local genomic differentiation. A gene annotated to modulate vegetative to reproductive phase transition was identified to have undergone adaptive evolution underlying the perenniality-annuality transition. Our study provides mechanistic insights into polyploid genome evolution and lays a foundation for unleashing genetic potential from the perennial gene pool for soybean improvement.

Introduction

The *Glycine* genus is composed of two subgenera that diverged ~5 million years ago (MYA) and differ in life history strategy: annuals in subgenus *Soja*, which includes only soybean (*Glycine max*) and its wild progenitor, *Glycine soja*, from which it was domesticated in East Asia about 6,000-9,000 years ago[1]; and perennials in subgenus *Glycine*, comprising ~30 species all found in diverse habitats in Australia[2] (**Fig. 1a**). Due to the economic importance of soybean as the world's most widely grown oil and protein seed crop, immense efforts have been made in sequencing both cultivated and wild accessions of the annual soybeans[3-5]. However, this primary gene pool possesses a remarkably low level of genetic diversity, which has become a key factor limiting the crop's yield potential and environmental resilience[6]. Perennial *Glycine* species represent an extended gene pool for improvement of the annual crop for traits such as large numbers of seeds per pod, resistance to cyst nematode and fungal pathogens, and tolerance to drought and salt stresses[2]. The genomes of the functionally diploid *Glycine* species all show evidence of a whole genome duplication (WGD) shared with many legume species that occurred ~65 MYA[7] and a second unique WGD that occurred ~13 MYA[2], and are thus referred to as paleopolyploids. Intriguingly, within the last 350,000 years there has been a burst of independent allopolyploidy events in the perennial subgenus, with at least eight different allopolyploids formed from various combinations of eight different diploid genomes[8]. Therefore, the genus *Glycine* presents an excellent system to understand polyploid genome evolution as well as the life-strategy transition between perenniality and annuality.

Results

Sequencing and genome annotation. We obtained high quality-standard genome sequences of five representative perennial diploid species across the *Glycine* genus, *Glycine falcata* (FF), *Glycine stenophita* (BB), *Glycine cyrtoloba* (CC), *Glycine syndetika* (AA), and *Glycine tomentella* D3 (DD), and a perennial

allopolyploid, *Glycine dolichocarpa* (AADD, dubbed $A^tA^tD^tD^t$ to distinguish it from the ancestral AA and DD genomes). These genomes were *de novo* assembled through a combination of PacBio single molecule real-time (SMRT) sequencing, Illumina sequencing, and chromatin conformation capture Hi-C technologies (Methods), and further corrected/improved through integrating our previously generated paired bacterial artificial chromosome (BAC) end sequences (BESs) from the same set of accessions (soybase.org) (**Supplementary Table 1**). Of the BESs uniquely mapped to their respective genomes, 99.0~99.7% were anchored in pairs in expected orientations and ranges of physical distances. Approximately 98.6%-99.6% of the genomes were assembled into 20 (diploids) or 40 (allopolyploid) chromosomes, with average contigs N50 ranging from 2.2 to 6.8 megabase pairs (Mbs) and average scaffold N50 ranging from 49 to 71 Mb in size (**Supplementary Table 2**), comparing favorably with those of the reference genomes of the annual soybeans[3, 9]. Completeness of the genome assemblies was assessed by BUSCO[10] and CEGMA[11] (**Supplementary Tables 3, 4**). The assembled diploid genomes range in size from 941 to 1,374 megabase pairs (Mb), with 55,376-58,312 annotated protein-coding genes (**Supplementary Table 5**). The assembled allopolyploid genome is 1,948 Mb, harboring 107,346 annotated protein-coding genes (**Supplementary Table 5**). Approximately 40%-62% of these genomes are composed of transposable elements (TEs) (**Fig. 2a and Supplementary Table 5**).

Phylogeny and karyotype stability. A phylogeny of these perennials and the annual species was constructed with 281 randomly selected single-copy orthologous genes using common bean (*Phaseolus vulgaris*), which diverged from the *Glycine* lineage ~19 MYA[12], as an outgroup. The evolutionary relationships, as reflected by the phylogenetic tree (**Fig. 1a**), are consistent with previous reports based on a limited number of genes, although the dates for the split of the perennial lineage from the annual species and the speciation of individual perennials were estimated to be ~1-2 MY earlier than the previous estimates with a different method[13] (**Fig. 1a**).

The perennial diploids exhibited a high level of chromosomal conservation (**Fig. 1b and Supplementary Fig. 1a**). Only 183 non-redundant genomic rearrangements including inversions, translocations, and insertions/deletions (InDels) of >50 kb in size were identified by pairwise comparisons with the annual soybeans using common bean as an outgroup (**Supplementary Table 6**). Perennial-specific and species-specific rearrangements were also identified (**Fig. 1c and Supplementary Fig. 1a**). Comparison among multiple species/genomes enabled validation of many genomic rearrangements and defining the relative timing and nature of those events, such as the occurrence of the inversion of a ~3.4-Mb fragment involving 398 genes in the annual species (**Supplementary Fig. 1b**) and the reorganization of chromosome 8 of the D/D^t genome by a combination of inversion, deletion, and translocation events, which resulted in a reduction of >20Mb of genomic sequence as compared with chromosome 8 of the A/A^t genome (**Fig. 1c**). Intriguingly, the ancestral telomeric region adjacent to the site where these rearrangements occurred has been retained as one of the telomeric regions of the re-arranged chromosome in the D/D^t genome (**Fig. 1c and Supplementary Fig. 1c**). The gene density along this chromosome has been reshaped primarily by accumulation of TEs in the re-structured pericentromeric regions and reduction of TE sequences in the re-structured chromosomal arm (**Fig. 1c**). Overall, fewer

genomic rearrangements have occurred in the perennials than in the annual species as revealed by comparison with common bean (**Fig. 1b, d**).

TEs, centromeric repeats, and rapid intergenic differentiation. Among all categories of TEs, long terminal repeat-retrotransposons (LTR-RTs) are most abundant, accounting for 79.4-90.6% of annotated TEs, or 31.6-49.9% of the sequenced perennial genomes (**Extended Data 1; Supplementary Table 5**). Individual LTR-RT families exhibited distinct spectra of amplification among the species as reflected by their relative abundance and estimated insertion times (**Fig. 2a, b, Supplementary Fig. 2b-e and Supplementary Tables 7, 8**). For example, the largest *gypsy*- and *copia*-LTR-RT families in the 1,374-Mb F genome, comprise 317.4 Mb (23.1%) of the genome. By contrast, few LTR-RTs belonging to these two families were seen in the annual genomes. Given that 98.2% of intact LTR-RTs were estimated to be amplified in the last 5 MYA (**Supplementary Tables 7, 8**), it is apparent that LTR-RT amplification was largely responsible for genome variation. On the other hand, the pace and degree of LTR-RT DNA loss are also striking. Of the 905 intact LTR-RTs in the perennials estimated to be amplified ~7 MYA, and thus considered to be inserted prior to the split of the perennial and annual lineages, none were found to have intact orthologs in the annuals, and fewer than a quarter of these elements had detectable remnants at putatively orthologous sites in the annuals, suggesting a loss of at least 94.3% DNA from the “original” copies. In addition, a large amount of LTR-RT DNA has been removed by formation of solo-LTRs through unequal recombination (**Supplementary Fig. 2a**). Besides LTR-RTs, Mutator and Helitrons are the two major TE superfamilies contributing to the perennial genome size variation and differentiation of intergenic regions (**Fig. 2a**).

Centromeres in most plant species are composed of long arrays of centromeric satellite repeats (CSRs), which are often interrupted by centromere-enriched retrotransposons (CRs). In the annual soybean, two subfamilies of CSRs, G^m-Cent1 and G^m-Cent2, were previously identified and found to mark two distinct subsets of the 20 chromosomes by fluorescence *in-situ* hybridization[14] and interpreted as evidence for the paleo-allopolyploid origin of soybean. Intriguingly, few copies of these CSRs were detected in the perennial genomes. Compared with the annual *G. max* (G^m) genome, which harbors ~38 Mb of CSRs, the F genome contains ~7 Mb, of CSRs, while the B, C, A, and D genomes only contain 75, 5, 0.4, and 12 kb of the CRS-homologous sequences, respectively (**Fig. 2c**). The G^m-Cent1 and G^m-Cent2 were estimated to have diverged ~8 MYA, and CRS homologs in the F genome (dubbed G^f-Cent) are slightly more similar to G^m-Cent2 than G^m-Cent1 (**Fig. 2d and Supplementary Fig, 3, 4a**). Based on the relative frequencies of physical adjacency of G^m-Cent1/G^m-Cent2 and specific retrotransposon sequences detected with the Illumina sequences, we identified Gmr17 (enriched in the G^m-Cent1 sequences) and Gmr01 (enriched in the G^m-Cent2 sequences) (**Fig. 2e-g**), as putative CR families in the annual genomes, whereas no similar retrotransposon families were found in the perennial genomes (**Fig. 2e, f, h and Supplementary Fig. 4b-d**).

Super-pangenome framework and evolutionary architecture. The representative perennial *Glycine* genomes provide a foundation for constructing a super-pangenome[15] framework of this genus. Rather than defining presence/absence of individual gene families that would not reflect the gain/loss of individual genes, we identified and compared orthologous genes among the sequenced perennial species.

A total of 109,827 non-redundant genes were annotated in the five perennial diploids, of which, 31,936 (29%) are shared by all the five perennials as orthologs and referred to as perennial core genes (**Fig. 3a and Supplementary Table 9**). By contrast, a total of 129,006 non-redundant genes were annotated in the 26 *G. soja*/*G. max* accessions[5], 31,564 (24.5%) are shared by all these annual accessions and referred to as annual core genes (**Supplementary Table 10 and Supplementary Fig. 5a**). Of the 31,936 perennial core genes, 17,922 (56.2%) overlap with the annual core genes, 8,704 (27.2%) overlap with the annual non-core genes, and 5,310 (16.6%) were perennial specific core genes (**Fig. 3b**). Of the 77,891 perennial non-core genes, 7,022 (9.0%) overlap with the annual core gene set, 6,745 (8.7%) overlap with the annual non-core genes, and 64,124 (82.3%) were identified as perennial specific non-core genes (**Fig. 3b**). The shared orthologs between any two of the five perennial species after 5.7-3.8 MY independent evolution account for 77.4-83.5% of all non-redundant genes annotated in the compared species (**Fig. 3a and Supplementary Table 9**), whereas the shared orthologs between a *G. max* and a *G. soja* accession that diverged ~0.25 MYA make up 84.5% of all non-redundant genes in the compared two accessions (**Supplementary Table 10**). These observations suggest a much higher rate of non-core gene formation in the annuals than found in the perennials.

Overall, the 17,922 core genes shared between the perennials and annuals exhibited lower rates of synonymous substitution (Ks) and non-synonymous substitution (Ka), and stronger intensities of purifying selection (w) than the 6,745 non-core genes shared between the perennials and annuals in both lineages, but neither the core genes nor the non-core genes showed differences in Ks, Ka and w between the two lineages (**Fig. 3c-e**). Among these shared genes, the duplicates showed lower rates of Ka and stronger intensities of purifying selection than the singletons in both lineages, but no differences in Ks, Ka and w were detected between the two lineages (**Fig. 3f-h**).

The duplicates were more conservative between the annuals and the perennials than the singletons, as 83.74% (23,671 in the D genome), 86.27% (21,537 in the F genome) of the duplicates were shared with the *G. max* genome compared to 46.80% (6,386 in D) and 50.85% (6,036 in F) for the singletons (**Supplementary Fig. 5b, c**). In addition, a higher ratio of duplicates to singletons was observed in the core gene sets (4.7:1) than in the non-core gene set in the perennial diploids (1:1.6) (**Supplementary Fig. 6 and Supplementary Tables 11, 12**). Interestingly, a higher ratio of duplicates to singletons was observed in the annual genome than in the perennial genomes, and the rapid emergence of non-core genes in the annuals appear to be partly caused by the rapid turnover of singletons (**Supplementary Fig. 5d**). These observations suggest the two life strategies have had distinguishable effects on gene evolution during the course of the diploidization process after the split of the annual and perennial lineages.

Adaptive evolution of flowering networks underlying perenniality. Little is known about the genetic basis of the life strategy transitions between perenniality and annuality in flowering plants. Nevertheless, it has been suggested that regulatory networks controlling flowering are involved, as the outcomes of adaptive response to particular environments. Mutagenesis analysis revealed that *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 15* (*SPL15*), the ortholog of *Arabidopsis thaliana FLOWERING LOCUS C* (*FLC*), in the perennial crucifer *Arabisalpina* contributes to perenniality through reducing flowering duration to

facilitate a return to vegetative development, preventing floral transition of some branches for polycarpic growth, and conferring flowering response to winter temperatures that restrict flowering to spring[16, 17]. In an attempt to pinpoint candidate genes underlying the perenniality-annuality transition in *Glycine*, we selected 174 genes shared by these perennials with seven wild annual soybean accessions that are orthologous/homologous to the *A. thaliana* genes controlling flowering[18] (**Supplementary Table 13**), and determined whether they have undergone adaptive evolution during the divergence of the two lineages (see Methods). Of the 174 genes examined, two genes, which are orthologous to the *Arabidopsis PLANT HOMOLOGOUS TO PARAFIBROMIN* (*PHP*) and *APETALA2* (*AP2*), were found to have experienced adaptive evolution (**Fig. 4a and Supplementary Table 14**). However, only a *PHP* ortholog in *Glycine* showed purifying selections within the annual and perennial groups (**Supplementary Fig. 7**). In *Arabidopsis*, the loss-of-function mutant *php* resulted in mis-regulation of *FLC* and activate *SPL15* and *FT* expression, conditioned accelerated phase transition from vegetative growth to flowering (**Fig. 4d and Supplementary Table 14**). Structural predictions from the consensus sequences of the PHP protein in the annuals and perennials showed noticeable differences (**Fig. 4b**). The comparion of Ka/Ks also indicates subgenera specific adaptation (**Fig. 4c**). Together, these observations suggest that the adaptive evolution of the *PHP* ortholog in *Glycine* may be partly responsible for the perennial-annual life strategy transition, although additional genes may also be involved.

Biased subgenome fractionation in the recent allopolyploid. We first evaluated the collinearity of genes among the A, D, A^t and D^t genomes/subgenomes, focusing on rearrangements each involving DNA larger than 50 kb. Compared with the more diverse perennial genomes such as F and B, which produced 7.5 rearrangements per MY, and B and C, which produced 6.8 rearrangements per MY, the A and D genomes exhibited a higher rate of rearrangements (13.5 per MY), with more rearrangements in D than A (**Supplementary Table 6**). A and A^t, and D and D^t, are highly conserved, with only 10 transpositions between A and A^t and six transpositions between D and D^t identified (**Fig. 5a**). Of the 10 transpositions between A and A^t, three occurred in A involving 113 genes, and seven occurred in A^t involving 314 genes, Of the six transpositions between D and D^t, three occurred in D involving 123 genes, and three occurred in D^t involving 146 genes. It remains unclear if the transpositions detected in the A^t and D^t subgenomes occurred before or after the allopolyploidization event. Nevertheless, 23 small inter-subgenomic transpositions involving a total of 45 genes were identified as post-allopolyploidization events (**Fig. 5b**).

Comparison among the A, D, A^t and D^t genomes revealed loss of 7,351 genes from the A^t and D^t subgenomes of the allopolyploid, including 3,242 from A^t and 4,109 from D^t (**Fig. 5c, Supplementary Fig. 8a and Supplementary table 15**). Based on their homoeologs present in the other subgenome, 60.4% of the lost genes were deduced to be singletons in their respective subgenomes prior to their losses, 39.6% were deduced to be members of duplicated gene pairs generated by the ~13-MYA WGD event prior to their losses (**Fig. 5d**). More singletons were lost in D^t than in A^t, whereas no difference in the number of losses of duplicates was observed between the two subgenomes (**Fig. 5e**). Only 0.14% of the ~13-MYAWGD-derived homoeologs lost both copies from either the A^t or D^t subgenomes (**Fig. 5e**). Based on

the orthologs of the 7,351 genes lost in A^t and D^t that are present in the A and D genomes, we found that 55.4% and 76.2% of the non-core genes present in both A and D were lost from the A^t and D^t subgenomes respectively, whereas 39.8% and 49.6% of the core genes present in both A and D were lost from the A^t and D^t subgenomes. It is apparent that non-core genes had a higher tendency of loss than the core genes in both subgenomes (**Supplementary Table 16**). In addition, the A-orthologs of the genes lost from A^t have experienced an overall lower stringency of purifying selection and exhibited an overall lower level of expression than their orthologs in the D genome, while the D-genome orthologs of the genes lost from D^t have undergone an overall lower stringency of purifying selection and exhibited an overall lower level of expression than their orthologs in the A genome (**Fig. 5f, g**). Thus, the A^t subgenome was likely dominant over the D^t subgenome upon the formation of the allopolyploid. The biased fractionation between the A^t and D^t subgenomes by gene losses were likely pre-destined by diverged genomic features of their diploid progenitors such as the levels of expression (**Supplementary Fig. 8b**), the stringencies of purifying selection (**Supplementary Fig. 8c**), the distribution of TE distal to the promofter regions of adjacent genes that is associated with levels of gene expression (**Supplementary Fig. 8d**). We found that gene losses in A^t and D^t tended to occur in clusters **Fig. 5h**, as exemplified (**Supplementary Fig. 8e**), whose orthologs in A and D generally showed a pattern of co-expression (**Fig. 5i**; r^2=0.155, P=0.012). By contrast, such a pattern of co-expression was not seen between the A- or D-orthologs of a single lost gene in A^t or D^t and its adjacent gene (**Fig. 5j**; r^2=0.014, P=0.508).

Illegitimate recombination as a key mechanism for genomic fractionation. The loss of a gene, as described above, was defined when it was annotated in A, D, and only one of the two subgenomes (A^t and D^t). After careful manual inspection of individual genes, we found that only 13.1% of the 7,351 gene losses are complete absence of the gene sequences, whereas 45.3% and 6.3% of the losses were pseudogenized genes caused by small InDels and point mutations, respectively, and 4.4% were genes interrupted by TE insertions (**Fig. 5k and Supplementary Fig. 9**). The remaining 30.9% of gene losses were defined as “losses” as they were not predicted as genes Nevertheless, this category of “losses” in the A^t and D^t subgenomes were proportional to the ratio of total lost genes in respective subgenomes and were not excluded from our analysis.

In an attempt to shed light on the mechanism(s) that gave rise to the small deletions resulting in pseudogenization that was largely responsible for subgenome fractionation, we examined 2,850 small deletions with clearly defined boundaries in 2,315 genes in the allopolyploid. We found that 31.2% of these deletions were flanked by short repeats of 2-18bp, as exemplified in **Fig. 5l**, a hallmark of illegitimate recombination[19], suggesting that illegitimate recombination that gradually deleted genic sequences was a key mechanism for subgenome fractionation.

It has been predicted that allopolyploidization by interspecific hybridization triggers ‘genomic shock’ that would lead to widespread activation of TEs[20]. If this happened in the recently synthesized *Glycine* allopolyploid, as a consequence, a large number of identical or nearly identical TEs, particularly LTR-RTs

between the A^t and D^t subgenomes would be observed. We therefore extracted the reverse transcriptase (RT) sequences from 1,202 *copia* type and 3,070 *gypsy* type intact LTR-RTs which inserted into the genome within last 350,000 years to construct their phylogenetic relationships, respectively. Out of the 4,272 LTR-RTs examined, only 38 of LTR-RTs showed a high level of sequence identity (e.g., ³99%) between the A^t and D^t subgenomes (**Supplementary Fig. 10, 11**). By contrast, none of the LTR-RTs from A and D genomes showed such a level of sequence identity (**Supplementary Fig. 12, 13**). These observations suggest that the allopolyploidization event did not lead to massive amplification of LTR-RTs.

Discussion

We have generated chromosome-level genome assemblies of six representative *Glycine* perennial species, constructed a super-pangenome framework of the *Glycine* genus, which includes 109,827 non-redundant protein coding genes from the perennials, of which ~70% were absent in the annual soybean pan-genome, representing a huge repertoire of genetic potential for improvement of the annual crop. In addition to the genomic data, this study unveils the propensities and consequences of polyploid genome evolution, genetic determinants of the life-history strategy transition, and the causes and mechanisms for subgenome fractionation.

Our analyses revealed several diverged genomic features in the perennials compared with the annual soybeans, including a lower rate of genomic rearrangements, a lower ratio of solo-LTRs to intact LTR-RTs, and a slower pace of non-core gene formation. Genomic rearrangements are generally formed by repeat-mediated recombination events[21], while solo-LTRs are the products of unequal intraelement recombination; thus, the differences in rates of rearrangements and solo-LTR formation likely reflect different rates of unequal recombination. As unequal recombination often correlates positively with allelic recombination[22], such differences are likely to be associated with distinct generation times of the perennials and annual species. The slower pace of formation of non-core genes in the perennials may also be associated with the reduced generation times[23], which would reduce recombination-induced mutations that could lead to pseudogenization and gradual degradation of genic sequences. Intriguingly, the substitution rates as reflected by Ks did not show significant differences between the perennials and annuals. One possible explanation is the potential effect of recombination on natural selection. Alternatively, it may be explained by the relatively short time frame for independent evolution of the annual and perennial lineages relative to the much longer divergence from common bean, which was used as a reference for calculation of evolutionary rates.

Plant centromeres typically harbor CSRs as integral part for maintaining centromere functions[24]. Some CSRs differ drastically between close relatives, such as sibling *Oryza* species[25], and some are highly conserved even between species diverged for millions of years, such as maize and *Oryza sativa*[26]. Given such a short time of divergence between the annual and perennial lineages, the lack of typical CSR sequences in the perennials is a surprising observation. Similarly, however, centromere satellites identified

in the annual *Brachypodium distachyon* were not detected in a few examined perennial species of the *Brachypodium* genus[27]. It remains unclear whether the CSRs were lost in these perennials or were born in the annuals, and whether they were associated with the life strategy transition.

Biased subgenome fractionation has been observed in several paleopolyploids[28, 29]. However, because of the extinction of their diploid progenitor species and extensive inter-subgenomic exchanges and reshuffling, the subgenomic origins of the hundreds of duplicated segments retained in those paleopolyploids could not be determined. As such, the “subgenomes” in each of those paleopolyploids were simply defined by sorting the duplicated segments into two distinct groups according to the degrees of regional fractionation[28, 29], which were unlikely to be representative of their ancestral diploid genomes. The availability of the A and D genomes allowed precise definition of the two subgenomes, identification of the nature of many subgenome fractionation events, and characterization of genomic features of the “lost” genes based on their existing orthologs in the ancestral diploids. Our analyses indicate that subgenome fractionation was primarily driven by local genomic features such as genes’ levels of expression and TE distribution, duplication status, levels of purifying selection, and local genetic recombination. Given that few TEs have been newly generated since the formation of the allopolyploid, but >5,000 genes have been completely or partially deleted, plus >2,000 genes that are not expressed, transcriptome modification, rather than TE proliferation, is likely the main consequence of the polyploidy-triggered “genome shock”.

Methods

Methods and any associated references are available in the online version of the paper.

Data availability. Raw sequences generated during this study are deposited in the public repository of National Center for Biotechnology Information under accession number PRJNA503746. The annotated assemblies were deposited in the European Nucleotide Archive under accession number PRJEB44023.

Note: any Supplementary Information and Source Data files are available in the online version of the paper.

Declarations

Acknowledgements

The work was mainly supported by Supported by the Taishan Scholars Program of Shandong Province (tsqn201812036), the Agricultural Variety Improvement Project of Shandong Province (2019LZGC003), and Program for Scientific Research Innovation Team of Young Scholar in Colleges and Universities of Shandong Province, China, (2020KJF008) to DZ, YZ, and XSZ, and partially supported by the National Science Foundation Plant Genome Research Program, USA, to SAJ, JJD, and JM (IOS-0822258).

Author contributions

Y.Z., D.Z., and J.M. conceived and designed the research, Y.Z., X.W., and D.Z. performed analysis, Y.Z., X.W., S.A.J, J.J.D., X.S.Z., D.Z, and J.M. interpreted the data, Y.Z., D.Z. and J.M. drafted the manuscript, and J.J.D, S.A.J., and J.M. revised the manuscript.

Competing interests

The author declares no competing financial interests.

References

1. Sedivy, E. J., Wu, F. Q. & Hanzawa, Y. Soybean domestication: the origin, genetic architecture and molecular bases. *New. Phytol.* **214,** 539-553 (2017).
2. Sherman-Broyles, S., Bombarely, A., Grimwood, J., Schmutz, J. & Doyle, J. Complete plastome sequences from *Glycine syndetika* and six additional perennial wild relatives of soybean. *G3 (Bethesda)* **4,** 2023-2033 (2014).
3. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463,** 178-183 (2010).
4. Li, Y. H. et al. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32,** 1045-1052 (2014).
5. Liu, Y. C. et al., Pan-genome of wild and cultivated soybeans. *Cell* **182,** 162-176 (2020).
6. Hyten, D. L. et al. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA.* **103,** 16666-16671 (2006).
7. Koenen, E. J. M. et al. Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytol.* **225,** 1355-1369 (2020).
8. Bombarely, A., Coate, J. E. & Doyle, J.J. Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. *Peer J.* **2,** e391 (2014).
9. Xie, M. et al. A reference-grade wild soybean genome. *Nat. Commun.* **10,** 1216 (2019).
10. Simão, F.A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210-2 (2015).
11. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061-1067 (2007).
12. Lavin, M., Herendeen, P. S. & Wojciechowski M. F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **54,** 575-594 (2005).
13. Sherman-Broyles, S. et al. The wild side of a major crop: soybean's perennial cousins from down under. *Am. J. Bot.* **101,** 1651-1665 (2014).
14. Gill, N. et al. Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol.* **151,** 1167-1174 (2009).

15. 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).
16. Wang, R. H. et al. *PEP1* regulates perennial flowering in *Arabis alpina*. *Nature* **459,** 423-427 (2009).
17. Hyun, Y. et al. A regulatory circuit conferring varied flowering response to cold in annual and perennial plants. *Science* **363,** 409-412 (2019).
18. Bluemel, M., Dally, N. & Jung, C. Flowering time regulation in crops-what did we learn from Arabidopsis?. *Curr. Opin. Biotechnol.* **32,** 121-129 (2015).
19. Devos, K. M., Brown, J. K. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12,** 1075-1079 (2002).
20. McClintock, B. The significances of responses of the genome to challenge. *Science* **226,** 792-801 (1984).
21. Bzymek, M. & Lovett, S.T. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. USA.* **98,** 8319-8325 (2001).
22. Gaut, B. S. et al. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.* **8,** 77-84 (2007).
23. Primack, R. B. Reproductive effort in annual and perennial species of Plantago (Plantaginaceae). *Am. Nat*. **114,** 51-62 (1979).
24. Talbert, P. B. & Henikoff, S. What makes a centromere?. *Exp Cell Res*. **389,** 111895 (2020).
25. Lee, H. R. et al. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. *Proc. Natl. Acad. Sci. USA.* **102,** 11793-11798 (2005).
26. Cheng, Z. K. et al. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14,** 1691-1704 (2002).
27. Li, Y. J. et al. Centromeric DNA characterization in the model grass *Brachypodium distachyon* provides insights on the evolution of the genus. *Plant J*. **93,** 1088-1101 (2018).
28. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. USA.* **108,** 4069-4074 (2011).
29. Zhao, Meixia, et al. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell* **29,** 2974-2994 (2017).

Figures

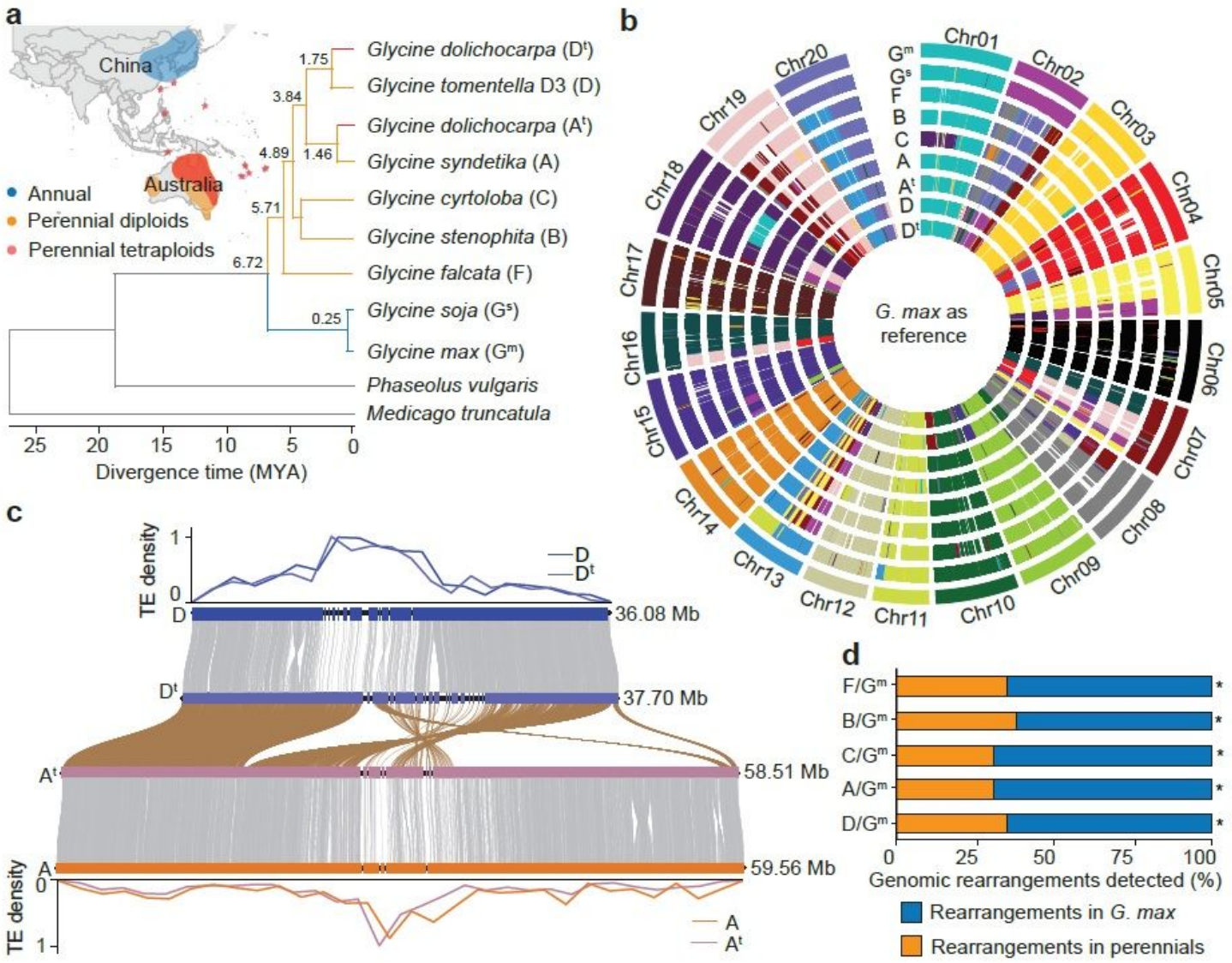

Figure 1

Geographical distribution, phylogeny, genomic synteny and rearrangements of annual and perennial Glycine species. a, Geographical distribution and phylogenetic tree of the annual (in green) and perennial (in orange/red) Glycine species. Medicago truncatula and Phaseolus vulgaris were used as outgroup species to generate a rooted tree. The divergence times of species or higher taxa are labelled. b, Synteny plot showing genomic collinearity among annual and perennial Glycine species using G. max as a reference. c, A genomic rearrangement of chromosome 8 between A/At and D/Dt. The TE density along each chromosome is plotted above and below the D and A genomes, respectively. d, Frequencies of genomic rearrangement s in G. max and perennial Glycine species evaluated using P. vulgaris as a reference.

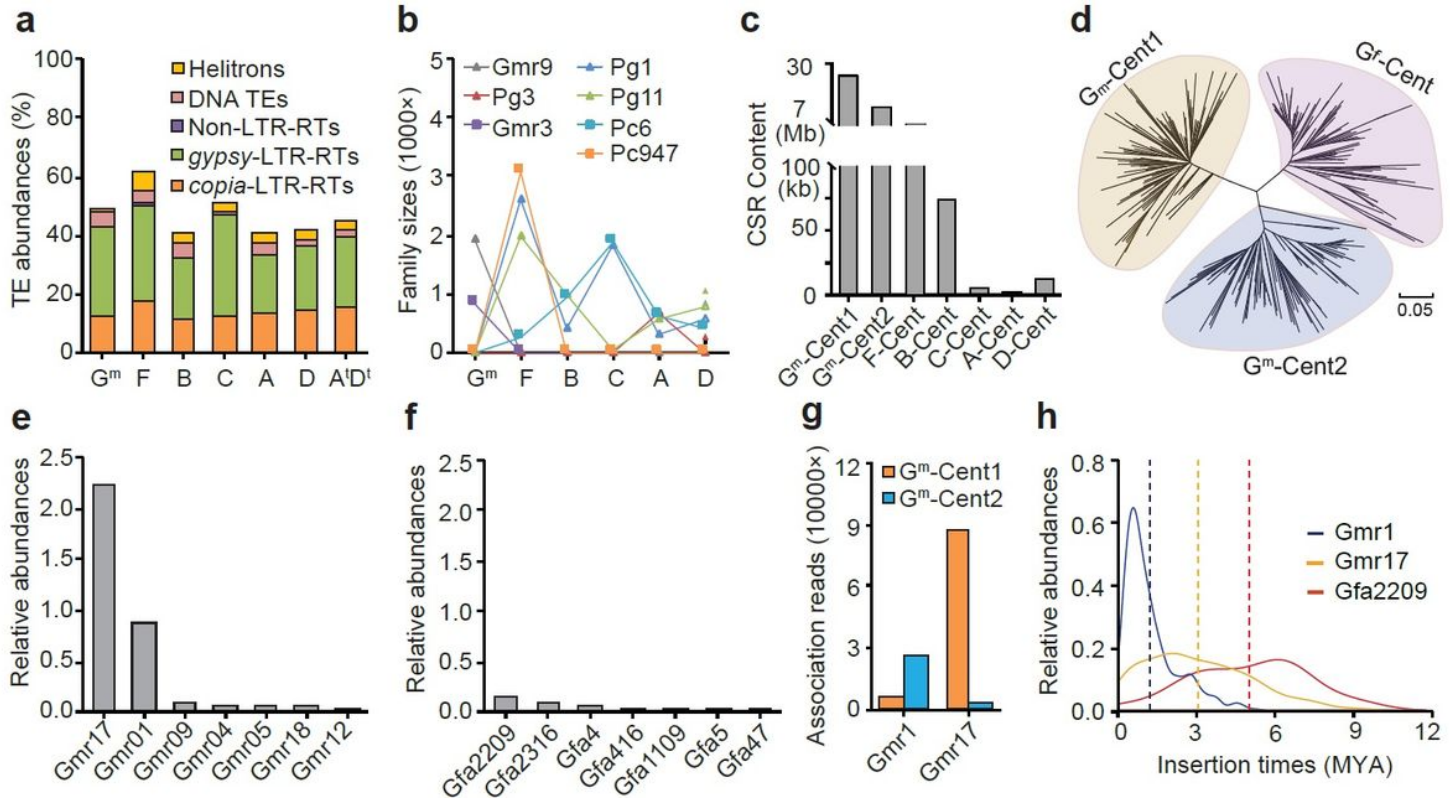

Figure 2

Analysis of repetitive sequences in G. max and perennial Glycine species. a, Composition of TE abundances in Glycine species. b, Dot plot showing the largest copia- and gypsy-LTR-RT families in G. max and the perennial Glycine species. c, Abundance of Gm-Cent1 and Gm-Cent2 in G. max and Gm-Cent in perennial Glycine species. d, Phylogeny of Gm-Cent1, Gm-Cent2 and Gf-Cent. e, f, Relative abundance of the top seven LTR-RT families showing the highest association indexes with Gm-Cent repeats in G. max (e) and the perennial species G. falcata (f). g, Specificity of association between two CR families and two CSRs in G. max. h, Estimation of the insertion times of putative CR families in G. max and G. falcata.

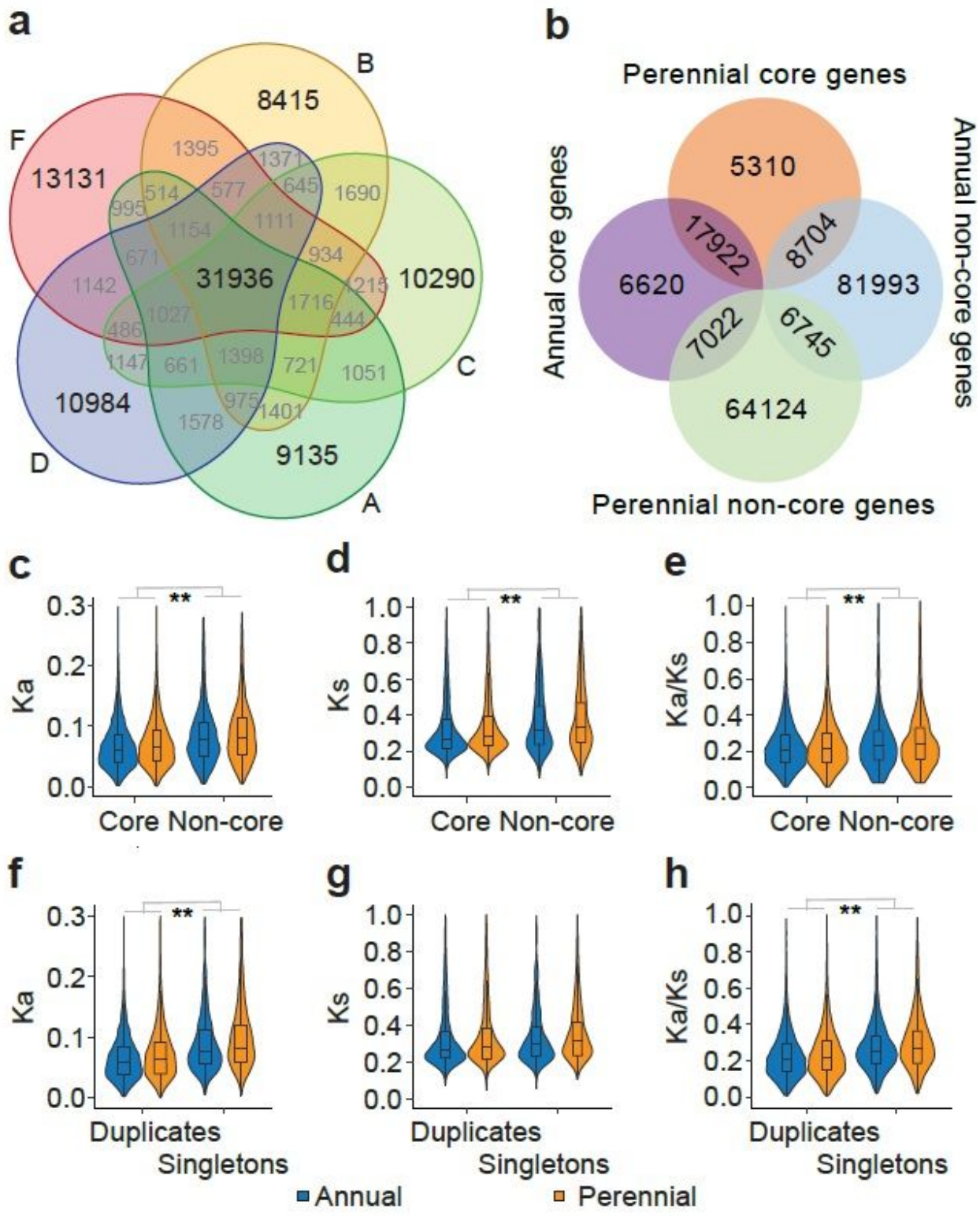

Figure 3

Comparative analysis of protein-coding genes in annual and perennial Glycine species. a, Venn diagram showing the numbers of protein-coding genes shared by all the perennial Glycine species (core genes) and not shared by all the perennials species (non-core genes). b, Venn diagram showing the numbers of core genes and non-core genes shared by annual and perennial Glycine lineages. c-h, Comparisons of Ka, Ks, and Ka/Ks of the core genes vs. non-core genes (c, d, e) or of singletons vs. duplicates (f, g, h) in the

annual and perennial species. The data for the five perennial species were combined. Significance was tested by student t-test; ** p<0.01.

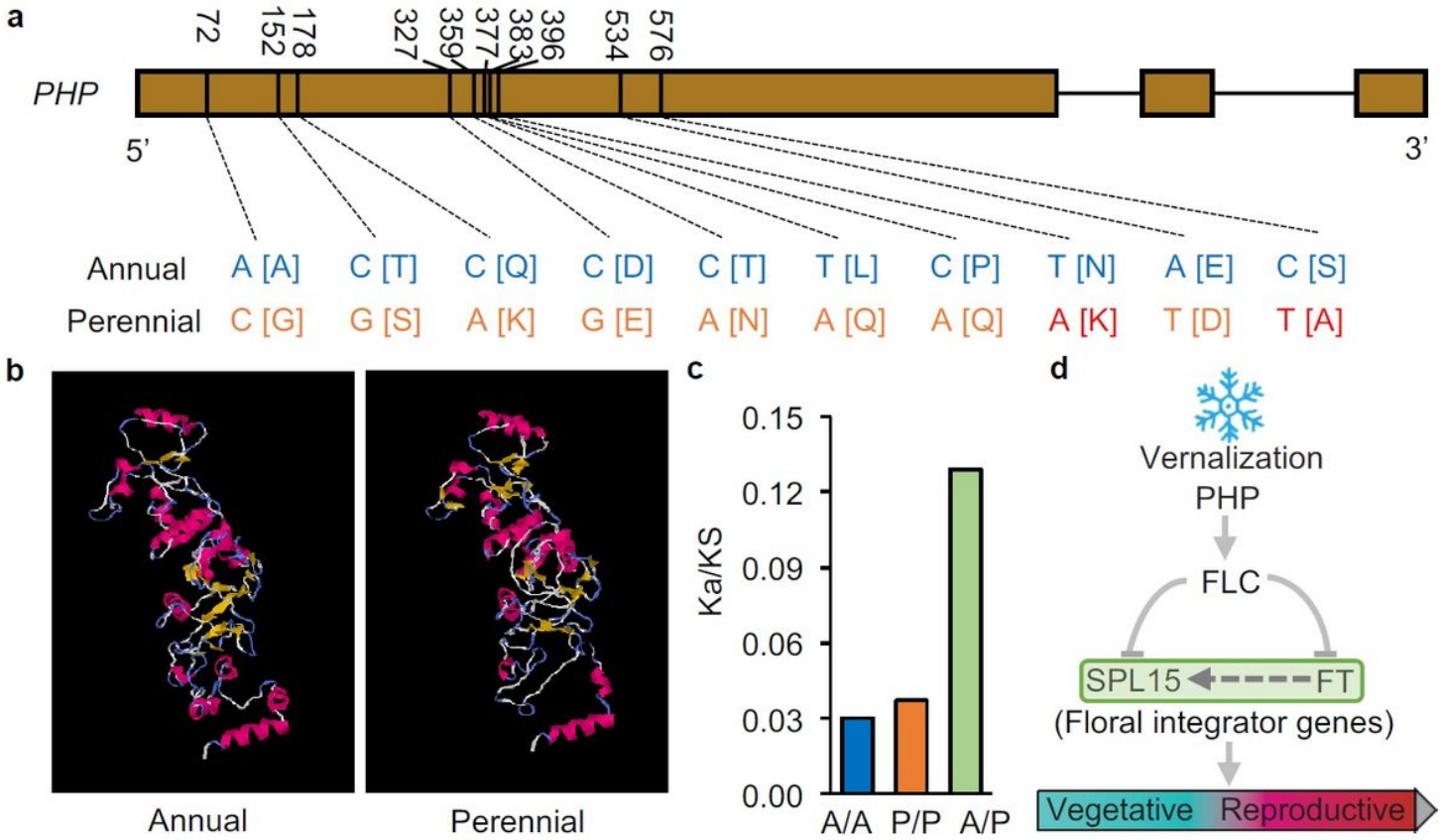

Figure 4

Structural variation and adaptive evolution of the PHP ortholog in the genus Clycine. a, Sequence comparison showing subgenera divergence. b, Predicted protein structure of the PHP orthologs representing the annual and perennial subgenera. c, Ka/Ks comparison within and between subgenera. d, Model of genetic pathway mediated by the PHP orthologs modulating the life strategy transition in Glycine according to the knowledge from Arabidopsis.

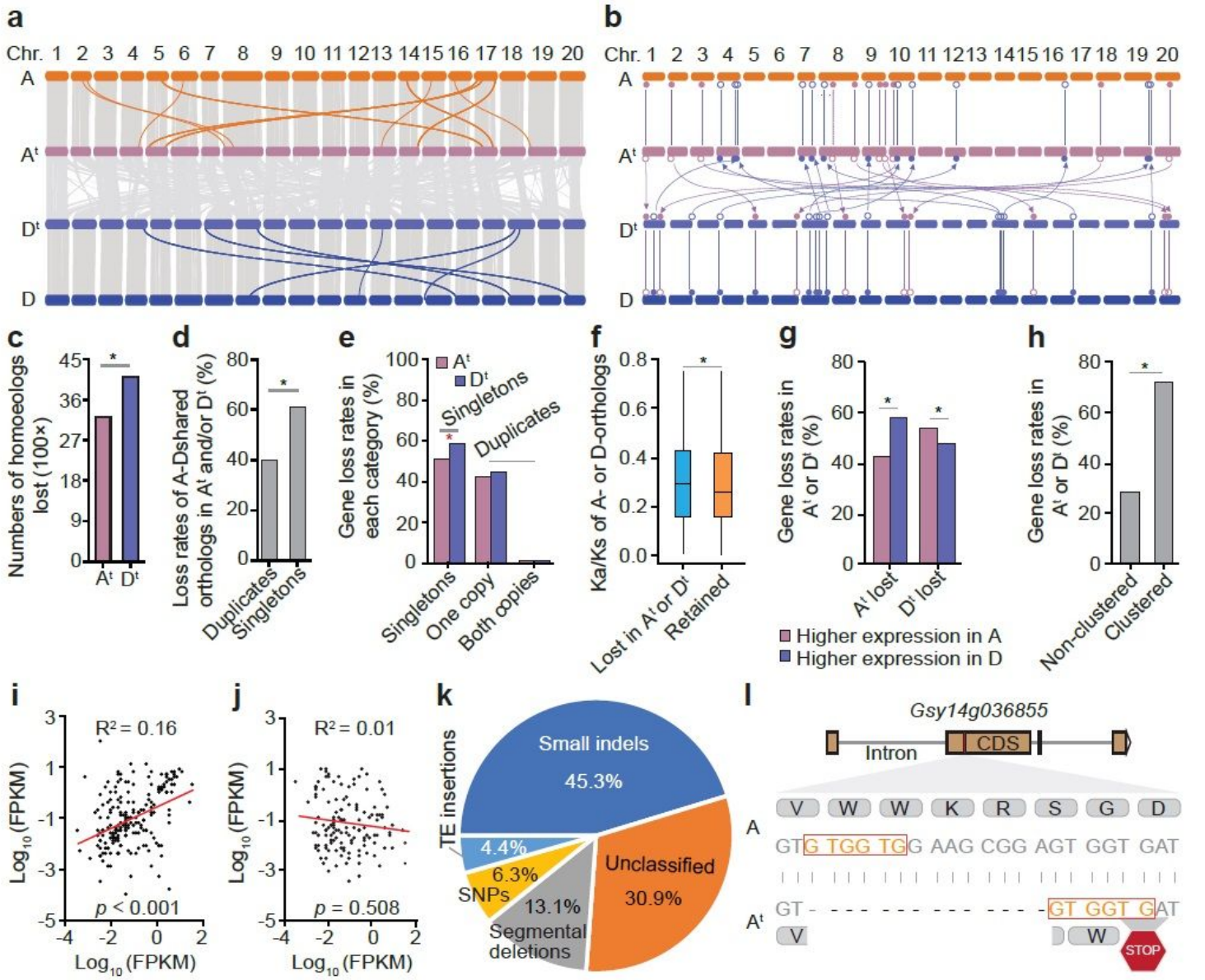

Figure 5

Subgenome differentiation and biased fractionation in the recent allopolyploid. a, Genomic collinearity and rearrangements between the allopolyploid and its two ancestral diploids illustrated based on orthologous/homoeologous genes. Curved orange and blue lines indicate segmental transpositions between A and At and between D and Dt, respectively. b, Subgenomic genic transpositions between At and Dt. Solid dots and hollow dots indicate current and original positions of transposed genes. Arrows indicate the paths of transpositions. c, losses of homoeologous genes from the At and Dt subgenomes. d, Losses of singletons and duplicated genes derived from the ~13-MYA WGD in the recent allopolyploid. e, Biased losses of singletons and unbiased losses of duplicated genes derived from the ~13-MYA WGD between At and Dt. f, Ka/Ks of A- and D-orthologs of genes lost from At and Dt versus Ka/Ks of their orthologs in D and A genomes, respectively. g, Negative correlations between expression levels of A- and D-orthologs and rates of losses of their corresponding genes from At and Dt, respectively. h, Tendency of gene losses within clusters. i, Co-expression between two adjacent genes in A or D whose counterparts in

At or Dt both were lost. j, Lack of co-expression between two adjacent genes in A or D, with only one of the two counterparts in At or Dt was lost. k, Different categories of gene losses and pseudogenizations in At and Dt. l, Exemplification of a small deletion by illegitimate recombination in the At subgenome that resulted in a premature stop codon.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ZhuangMethods.docx
- ZhuangExtendedData1.xlsx
- ZhuangSupplementaryFigures.pptx
- ZhuangSupplementaryTables.xlsx