**ONLINE METHODS**

**Plant materials and genome sequencing.** Five diploid perennial *Glycine* species [*G. falcata* (FF), *G. stenophita* (BB), *G. cyrtoloba* (CC), *G. syndetika* (AA), *G. tomentella* D3 (DD)] and one recent allotetraploid species [*G. dolichocarpa* (*G. syndetika* × *G. tomentella* D3)], were obtained from Cornell University. Young leaves were collected three weeks after planting in the greenhouse and used for DNA extraction using the cetyltrimethylammonium bromide (CTAB) method30. For Illumina short-read sequencing, DNA was physically sheared to a size of ~500 bp and sequenced with an Illumina HiSeqX platform to generate 150-bp paired-end reads at a mean of ~100× coverage. High-molecular-weight DNA (>20 kb) libraries were created following Pacbio’s standard protocol and sequenced using a Pacbio Sequel platform at ~60× coverage (Novogene, Tianjin, China). Hi-C libraries were created and sequenced at Phase Genomics at a minimum of ~100× coverage (Phase Genomic, Seattle, USA).

**Genome assembly and construction of pseudomolecule chromosomes.** Quality trimming of raw Illumina reads removed adaptor sequences and low-quality bases (*Q* < 20), as well as contamination with mitochondrial, chloroplast and PhiX reads, using Trimmomatic (2:30:10)31. Long reads generated by Pacbio were self-corrected using its built-in pipeline. Error-corrected reads were assembled using FALCON v.0.3.032. Multiple assembly runs were tested with adjusted parameters and the best assembly was selected based on contig N50. The acquired draft genome then underwent two rounds of error correction using Quiver33 and PILON34 to obtain contig-level assembly. The draft assembly was further scaffolded using Hi-C reads. Briefly, Hi-C reads were aligned to the draft genome using BWA-MEM35 and processed using SAMBLASTER and SAMTOOLS36 (bwa mem -5SP [assembly.fasta] [fwd\_hic.fastq] [rev\_hic.fastq] | samblaster | samtools view -S -h -b -F 2316). Quality control of the alignments was then performed using the Phase Genomics open-source Hi-C alignment QC tool, and scaffolding was carried out with the Phase Genomics Proximo Hi-C genome scaffolding platform to obtain chromosome-level assembly.

**Error correction and assembly evaluation.** Paired bacterial artificial chromosome (BAC) end sequences (BESs) generated from the SoyMapII project (https://soybase.org/soymap2/) were used to check and correct the orientation of scaffolded contigs. First, paired reads were mapped independently to the assembled genome using BLAST+37, and at least one of the paired reads that mapped uniquely to the assembly was kept for downstream analysis. Screening was then performed for paired reads that spanned two contigs. The order and orientation of contigs were manually adjusted on the basis of the expected insert size (~150 kb) and the orientation of mapped reads. The final statistics for each corrected assembly was generated with QUAST38, and the completeness of the assemblies was evaluated using BUSCO39 and CEGMA11.

**Annotation of repetitive sequences.** Repetitive elements were identified through a combination of *de novo* and structure-based approaches. The *de novo* prediction of LTR-RTs was performed with LTR harvest40 and LTR\_finder41. Identified LTR-RT candidates were further screened for the integrity of their LTR PPT and PBS signatures. A retrotransposon database was compiled from the intact LTR-RTs together with known retrotransposons from *G. max* and sequences downloaded from Repbase42 to form a confident retrotransposon database, and this was used to mask the genome for identification of nested LTR-RTs. The whole process was repeated until no new predictions could be made. *Gypsy*- and *copia*-LTR-RTs were distinguished according to their sequence similarity and the order of integrase, reverse transcriptase and RNase H sequences. Identified LTR-RTs were further classified into different families based on the 80-80-80 rules suggested by Wicker et al. (2007) using SiLiX43, 44. Non-LTR-RT SINEs were identified using SINE\_scan with default parameters45. Helitrons were identified with HelitronScanner46. A similar pipeline was employed to identify LINEs and DNA TEs, such as Mutator, hAT, CACTA, Mariner, PIF-Harbiger and Ping elements. Briefly, Enzymes specific to each type of TE were compiled and used to perform local BLASTx searches against assembled genomes. Upstream and downstream sequences were then extracted on the basis of the expected length of each type of transposon, and the patterns of target site duplications (TSDs) and terminal inverted repeats (TIRs) were screened for signatures distinguishing each type of transposon. The identified candidate sequences of each type of transposon were aligned, manually checked and clustered into families. Families that contained two or more members were kept. The insertion times of LTR-RTs were calculated as described by Du *et al.* (2010)47. Briefly, the Jukes-Cantor distance (K) of two LTRs was calculated in MEGAX48 and the insertion time (*T*) was estimated as *T* = *K*/2*r,* where a substitution rate (*r*) of 1.3 × 10-8 per synonymous sites per year was used49.

**Identification of centromeric repeats and associated retrotransposons.** To identify centromeric repeats in perennial *Glycine* species, we performed a genome-wide search for all tandem repeats using Tandem Repeats Finder50. The identified repeats were then clustered into families at 80% similarity using SiLiX44. The abundance of each family was estimated by *k*-mer analysis using Illumina short-read sequences with Jellyfish51 and was compared with known centromeric repeat sequences from *G. max* with BLAST+37. The distribution of the top five most abundant repeats along each chromosome was manually checked to assess whether they possessed the features of centromere repeats. A neighbor-joining tree was created for identified centromeric repeats on the basis of random sampling of 200 repeats from each family, using MEGAX48. Divergence among three types of identified centromeric repeats was calculated based on pair-wise comparisons of 200 random sample repeats from each type using the formula above. Centromere-related retrotransposons (CRs) were identified by mapping paired-end short reads to a database created from the LTR repeats and centromeric repeats, using BLASTn37. LTR repeats that were linked by paired-end reads to centromeric repeats within a reasonable distance (~500 bp) were considered as candidate CRs. The degree of association of each CR was calculated using the formula: Index = the number of mapped reads/(sequencing depth × family size).

**Annotation of protein-coding genes.** RNA-seq libraries were created using a mixture of leaf, root and flower tissue for each perennial *Glycine* species. Quality-trimmed reads were assembled *de novo* using Trinity, and a reference-based assembly was generated using the HISAT2 and StringTie pipeline52. The *de novo* and reference-based assemblies were combined, and the identical transcripts were removed using CD-HIT53. Fragments per kilo base per million mapped reads (FPKM) values generated by StringTie were used to estimated transcript abundances. Annotation of the genome was performed using the MAKER-P pipeline54. Species-specific TE libraries as well as the repeat database from *G. max* and Repbase were used to mask the genome by RepeatMasker55. Assembled transcripts were used as evidence for expressed sequence tags, and proteins downloaded from the NCBI protein database for *G. max* (av2) and other flowering plant species56 were used as evidence for proteins. To obtain confident annotations, we first performed gene model training with the ab initio gene prediction tools used by MAKER-P, including SNAP57 and AUGUSTUS58, and then conducted two rounds of MAKER annotation to obtain the final annotation file.

**Phylogenetic and** **genome collinearity analysis**. To construct phylogenetic tree and estimate the divergence time of annual and perennial *Glycine* species, proteins from *M. truncatula*, *P. vulgaris* (common bean), *G. max* (accession William82), *G. soja* (accession W05) and selected perennial *Glycine* species were classified into gene families using OrthoFinder54 (using an e-value cutoff of 1e−10 and >50% match). 281 single copy genes shared by all species were randomly selected to construct phylogenetic tree using neighbour-joining method with 100 bootstraps in MEGAX48. Divergence time estimation was performed using the RealTime-ML method within MEGAX48. The divergence time among *Glycine* species was calibrated using the time of divergence between *P. vulgaris* and *G. max*12. Genome synteny as compared among 26 recent reported annual *Glycine* accessions5, William 82, and our perennial *Glycine* species. Amino acid sequences from each pair of species selected were compared using all-against-all BLASTP (e < 1e−10; similarity >95% and coverage >80). Protein pairs belonging to the reciprocal best hits were extracted and fed into MCScanX59 to identify syntenic blocks between each pair of species. Except for *P. vulgaris*, for which each protein was expected to have two orthologous pairs in *Glycine* species, proteins were classified into 1:1 pairs (putative orthologs) based on the symmetrical best hit within syntenic blocks that had the lowest *P* value between the two genomes. Other homologous genes within each family were considered to be co-orthologs or in-paralogs.

**Gene classification.** Perennial core orthologs were defined as those genes shared by all five diploid perennial *Glycine* species for which a 1:1 orthologous pair could be identified within syntenic regions. The remaining genes were considered to be non-core genes. Genes were further categorized based on their duplication modes with the DupGen\_finder pipeline60 (<https://github.com/qiao-xin/DupGen_finder)> using *P. vulgaris* as the outgroup reference. All-to-all BLASTp was performed between each studied *Glycine* species and common bean (*e* value 1e-10; similarity >70%; coverage >70%; the top five matches were kept if more than five hits met the pre-set requirements). The same parameters were applied for all-against-all self-genome BLASTp comparisons, except that the similarity was set >80%.

**Comparison of Ka, Ks and Ka/Ks among annual and perennial *Glycine* species.** The Ka, Ks and Ka/Ks were calculated for the annual and perennial *Glycine* groups with the KaKs\_Calculator package using the YN method61. *G. soja* accessions W01, W02, and W03 were selected as representative species for annual *Glycine* species. For each orthologous gene pair, the Ka, Ks and Ka/Ks were calculated based on its orthologous gene identified in common bean as a reference. Coding sequences and their corresponding amino-acid sequences were aligned using MUSCLE62 and transformed into codon-aligned format with PAL2NAL63, low-quality alignments (similarity <85, coverage <85) were removed, and the resulting alignment files were manually checked and fed into KaKs\_Calculator. For the comparison between core genes and non-core genes, only genes shared by annual and perennial *Glycine* groups were selected. The evolutionary rate and selective pressure between annual and perennial *Glycine* species was tested using a Student’s *t*-Test at a significance level of *p* ≤ 0.05.

**Comparative analysis of genes as singleton and duplicates.** For calculated the percentage of duplicated genes shared between each perennial species and *G. max,* we only focus on duplicated genes generated by whole genome duplication (WGS), and duplicates shared with *G. max* was defined as both copies (homoeologs) are retained in both species based on the 1:1 orthologous genes pair identified above. The singletons shared between each perennial species and *G. max* was also calculated based on the 1:1 orthologous pair identified. For the comparison of Ka, Ks and Ka/Ks between the annual and the perennial groups as WGS duplicated genes and singletons, only core genes shared by annual and perennial groups were selected.

**Detection of genome rearrangements.** Genome rearrangements among selected *Glycine* species were initially identified using the SyRI pipeline64 with default parameters. Briefly, each pair of genomes was aligned using the NUCMer utility65 (--maxmatch -c 500 -b 500 -l 100) and filtered with a delta-filter module (-m -i 90 -l 100). The resulting files were fed to SyRI to detect variations in genome structure. Identified genome rearrangements were further filtered on the basis of the genome synteny file generated above. Only rearrangements occurred within syntenic blocks were kept for further analysis. The comparison of genome stability among annual and perennial species was calculated using *P. vulgaris* as a reference. The variation within shared syntenic regions among *P. vulgaris* and each pair of genomes of *Glycine* species was compared. Structural rearrangements were considered to have occurred in species that differed from *P. vulgaris*.

**Flowering-time control and genes related to perenniality.** Genes that control flowering time in *Arabidopsis* were used as queries to perform BLASTp against annotated proteins of each annual and perennial *Glycine* species. Genes with a significant hit (*e*-value < 1e-10) were extracted for downstream analysis. A neighbour-joining tree was created for each reference gene from *Arabidopsis* and its candidate orthologous gene in each *Glycine* species. The constructed phylogenetic tree was manually checked, and genes that clustered within the same clade as the *Arabidopsis* gene were considered to be bona fide orthologous genes in each *Glycine* species. For each pair of species studied, the Ka, Ks and Ka/Ks were calculated for orthologous gene pairs using the method described above. Identification of genes showing adaptive evolution between annual and perennial *Glycine* groups was performed according to the approach described by McDonald and Kreitman66, seven *G. soja* accessions67 were used as the representatives of annual *Glycine* species, and the divergence was corrected by Jukes and Cantor68 (<http://mkt.uab.es/mkt>). SNPs that distinguished annual and perennial *Glycine* species were extracted and their genotypes in annual *Glycine* populations were analysed in 302 soybean resequencing datasets69 to identify SNPs related to annual/perennial speciation. The impact of identified sequence variations on protein structure was predicted with I-TASSER70.

**Subgenome exchange in the allopolyploid.** Structural variations between each subgenome (AtAt and DtDt) and its ancestor genomes, *G. tomentella* D3 (AA) and *G. syndetika* (DD), were identified using SyRI as described above. Genetic exchanges between the two subgenomes in *G. dolichocarpa* were identified by mapping Illumina paired-end short reads of the two diploid contributors *G. tomentella* D3 and *G. syndetika* to the assembled tetraploid genome (AtAt/ DtDt). The mapping density along the chromosomes was calculated as Density (*d*) = the number of reads from A/the number of reads from D using a 10-kb window and was compared with syntenic regions of in a simulated tetraploid genome (AA/DD) to ensure that no bias was caused by an unequal sequencing coverage. When *d* was significantly <1 on the At genome, this indicated a translocation from Dt to At, and vice versa.

**Mechanism of gene fractionation and deletion in the allopolyploid.** The 1:1 orthologous pairs shared by *G. tomentella* D3 (A) and *G. syndetika* (D) were extracted and the presence/absence of an orthologous pair in the At and Dt genomes was examined. A subset of genes that had lost their orthologous pair either in At or Dt but not in both genomes was then selected to study the mechanism of gene fractionation and deletion. A Chi-squared Test was performed to test whether the number of A/At lost genes differed significantly from the number of D/Dt lost genes (*P* ≤ 0.05) genome-wide. To study local effect on gene deletion, 200 fragments (100 kb) were random sampled from each subgenome, and the number of genes lost in At and Dt of each orthologous pair was counted. Genes in each orthologous pair were also categorized as singletons or duplicates, on the basis of their duplication status in the A and D genomes. For the orthologous pair that one gene was annotated as singleton and another was annotated as duplicates, the percentage of lost genes in each category was calculated and tested according to their deviation from a 1:1 ratio, using a Chi-squared Test. The Ka/Ks values and transcript abundance were also measured for the lost and conserved orthologous gene pairs, based on their counterparts from the A and D genomes. To test the effect of TE insertion on gene expression, the distance of the nearest TE inserted into the upstream region of a gene was identified using bedtools71 (closest -id -D a) and the correlation was compared for the orthologous pairs between A and D genomes. The distance for lost genes along a chromosome was calculated according to the number of genes that separated them. If this number was smaller than 5, the two lost genes were considered to be in a cluster; otherwise, the gene was defined as an independently lost gene. To test whether lost genes had similar local expression levels, the correlation between the gene expression of the two nearest lost genes within the same cluster was calculated and compared with the correlation between the expression of the lost gene and its nearest conserved gene. For genes without an annotated orthologous pair in the tetraploid genome, genomic DNA from either the A or D genome was used as a query to perform BLASTn against the syntenic regions of the tetraploid subgenome to study the mechanism of fractionation/gene deletion (*e*-value 1e-10; word\_size 30; -qcov\_hsp\_perc 0.3). If genomic DNA could be identified, a local alignment of each exon in the lost gene was performed using BLAST+ and was analysed to examine the presence of mutations, indels and transposon insertions. This helps to identify the mechanism that regulates gene deletion/loss, flanking sequences of each deletion identified was further manually checked for the presence of Illegitimaterecombination.

**Investigation of TE activation following the formation of allopolyploid.** To study whether the formation of allopolyploid by interspecific hybridization triggers ‘genomic shock’, which would lead to the activation and transposition of TEs between subgenomes, the reverse transcriptase of LTR-RTs that estimated to be amplified less than 350,000 years were extracted from A, D, At and Dt genome. A neighbour-joining tree with 200 bootstrap replications was constructed for *gypsy* and *copia*-type LTRs using MEGA X48 for At/Dt and A/D genomes,respectively. The sequence divergence of transposons located on different subgenomes within each clade was calculated using the same general approach. TEs transposed between subgenomes were identified as those located on different subgenomes but placed in the same clade.

30. Murray, M. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8,** 4321-4326 (1980).

31. Bolger, A. M., Lohse, M. & Usadel. B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114-2120 (2014).

32. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods.***13,** 1050-1054 (2016).

33. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods.* **10,** 563-569 (2013).

34. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One.* **9,** e112963 (2014).

35. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078-2079 (2009).

36. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997v2 (2013).

37. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421-429 (2009).

38. Gurevich, A. et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072-1075 (2013).

39. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962,** 227-245 (2019).

40. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9,** 18 (2008).

41. Ou, S. & Jiang, N. LTR\_FINDER\_parallel: parallelization of LTR\_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DN*A **10,** 48 (2019).

42. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462-467 (2005)

43. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8,** 973-982 (2007).

44. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12,** 116-124 (2011).

45. Mao, H. & Wang, H. "SINE\_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets." Bioinformatics **33**(5), 743-745 (2017).

46. Xiong, W. W, et al. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl Acad. Sci. USA.* **111,** 10263-10268 (2014).

47. Du, J. C. et al. Evolutionary conservation, diversity and specificity of LTR retrotransposons in flowering plants: Insights from genome-wide analysis and multi-specific comparison. *Plant J.* **63,** 584-598 (2010).

48. Kumar, S. et al. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.***35,** 1547-1549 (2018).

49. Ma, J. & Jackson, S.A. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* **16,** 251-259 (2006).

50. Benson, G. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.***27,** 573-580 (1999).

51. Marcais, G. & C. Kingsford. Jellyfish: A fast k-mer counter. Tutorialis e Manuais **1,** 1-8 (2012).

52. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protocols* **11,** 1650–1667 (2016).

53. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics***22,** 1658–1659 (2006).

54. Campbell, M. S. et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164,** 513–524 (2014).

55. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Prot. Bioinform.* **25,** 4.10.1–4.10.14 (2009).

56. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** D61–D65 (2007).

57. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).

58. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** (suppl. 2). ii215–ii225 (2003)

59. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40,** e49 (2012).

60. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20,** 38 (2019).

61. Zhang, Z. et al. KaKs\_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4,** 259–263 (2006).

62. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

63. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.***34,** W609–W612 (2006).

64. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20,** 277 (2019).

65. Marcais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14,** e1005944 (2018).

66. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652-654 (1991).

67. 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).

68. Egea, R., Casillas, S. & Barbadilla, A. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.* **36,** W157-W162 (2008).

69. Zhou, Z. et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33,** 408–414 (2015).

70. Yang, J. et al. The I-TASSER suite: protein structure and function prediction. *Nat. Methods*. **12,** 7-8 (2014).

71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).