Comparative Genomics:
Glycine max

Prepared by group 5:

Aya Ben TAGHALINE

Joelle ASSY

Phuc Khanh Nhi NGUYEN

Rayane ADAM

Yazid HOBLOS

M2 GENIOMHE

Supervisor: Prof. Carene RIZZON

Université Évry Paris-Saclay

2025 - 2026

Table of Contents

Introduction

The objective of this project is to investigate the evolution of duplicated genes and transposable elements in the Glycine max genome. Using a well-defined protocol below, we aim to identify duplicated genes, classify them according to their duplication mechanisms, and analyze their genomic organization and functional properties. In parallel, we seek to characterize the abundance, distribution, and genomic context of transposable elements and assess their potential role in shaping genome structure and gene duplication patterns.

Ks represents the number of synonymous substitutions per synonymous site between two homologous coding sequences and is commonly used as a molecular clock proxy because synonymous mutations are generally assumed to be selectively neutral and therefore accumulate at an approximately constant rate over evolutionary time. If gene duplications and deletions occur randomly, the genome-wide age distribution of duplicated gene pairs is expected to show an L-shaped Ks distribution, characterized by a high proportion of young duplicates with low Ks values and a rapid decrease in older duplicates due to gene loss over time. In contrast, large-scale duplication events such as whole-genome duplications (WGD) occurring within a time range generate a secondary Ks peak, whose position reflects the relative timing of the duplication event. However, *large Ks values are associated with increasing estimation error, and Ks-based analyses are therefore typically restricted to lower Ks ranges to balance accuracy and statistical power.* Under these assumptions, Ks distributions provide an effective framework to distinguish ancient WGD and paleopolyploidy (Eckardt, 2004).

Glycine max (soybean) is one of the most important crops, it has been a spotlight of genomic research in plant evolution due to its complex evolutionary history.It's a diploid with paleopolypoid origins, having undergone 2 whole genome duplication (WGD) events approximately 59 and 13 million years ago. This has resulted in a highly duplicated genome, with about 75% of paralogous genes (Shmutz et al., 2010). As a polypolyploid, it had undergone WGD in the distant evolutionary past, but over millions of years its genome has been diploidized. This is why Glycine max has been studied to see how ancient WGDs shaped the plant's genome today, understand diploidization and gene loss and retention bias as well as the functional bias of duplicated genes, in addition to transposable elements' dynamism and interactions in the genome.

Gene duplication is a major evolutionary mechanism that generates new genetic material and allows functional diversification, while duplicated genes can be retained or lost depending on selective pressure. Transposable elements are also major components of eukaryotic genomes and contribute to genomic evolution by shaping the genomic environment surrounding genes through their dynamism. Previous studies have shown that different classes of transposable elements accumulate differently around duplicated and singleton genes, indicating that the local TE landscape is associated with gene retention and essentiality. This highlights the importance of considering transposable elements when studying the evolutionary dynamics of duplicated genes (Correa et al., 2021).

Our biological questions that we aim to tackle are:

WGD Events:

- What duplication events are revealed by the Ks distribution? Do they align with previous studies on soybean?

Duplicates Analysis:

- Can the orientation patterns of TAGs provide insights into the evolutionary processes generating tandem duplications in Glycine max?
- Do duplicates exhibit any particular structural patterns for their organization in the genome?

Functional Analysis:

- How does gene duplication type influence gene function, specifically comparing singletons, tandemly arrayed genes (TAGs), and non-TAG duplicated genes.
- How does gene family size affect gene function, comparing small versus large gene families.

  ⇒ Together, these questions aim to understand how genome organization and duplication contribute to functional specialization in Glycine max.

Transposable Elements Analysis:

- How are TEs characterized within the *Glycine max* genome in terms of their abundance, chromosomal coverage, and physical length across various classes and superfamilies?

- How are TEs spatially arranged across the 20 chromosomes, and which superfamilies show a preference for 'self-nesting' or acting as hosts for other elements?
- Does the spatial proximity between TEs and genes vary depending on whether a gene is a singleton, a TAG, or a non-TAG duplicate?

Materials and Methods

Data

Glycine max has a 974.4Mb long genome with 89598 isoforms belonging to 56679 coding genes (EnsemblPlants, v2.0). With 2n=40, these genes are distributed across 20 chromosomes.

For our duplicated genes analysis, data consisted of a peptides fasta file comprising all coding genes (including isoforms, which will be filtered later on). At a later stage for Ks computation, the CDS fasta (nucleotide) file was also retrieved from EnsemblPlants (v2.0).

Functional annotations were based on GO Slim terms retrieved through the PANTHER database (version 19.0) and were used for downstream functional enrichment analyses.

Transposable elements of the same genome's version were retrieved from APTEdb (apte.cp.utfpr.edu.br).

Tools and algorithms

- Ncbi-blast+: command line
- MCL: command line
- mafft: command line
- PAL2NAL: command line
- PAML: command line
- MCScanX: command line
- GenomicRanges: R package
- ggplot2: R package
- matplotlib: python library
- Panther: webserver

Pipelines

Pipeline 1

`pipeline_1/`: directory containing all scripts for this part of the pipeline

This project implements a modular bash pipeline to identify protein families in Glycine max using sequence similarity and graph-based clustering:

- (1) Protein sequences were first downloaded from Ensembl Plants and made sure **not to include mitochondrial and chloroplast genes** (Mt and Pt are not part of evolutionary history (endosymbiosis theory) ⇒ wont consider duplicated genes from these 2 organelles) and processed to extract sequence metadata, after which only the longest isoform per gene was retained to reduce redundancy.
- (2) An all-vs-all BLASTP was performed on the filtered protein set to detect pairwise similarities.
- (3) Query and subject coverage (qcov, scov) were then computed using protein lengths and added to the BLAST output to ensure alignments covered a sufficient fraction of both sequences.
- (4) BLAST hits were filtered using user-defined default thresholds (at the beginning of the project), with default parameters of >30% sequence identity, >50% query coverage, and >50% subject coverage, and optional filtering on e-value (e.g. 1e-5 or 1e-10) or bit score if user wants (not used here).
- (5) The filtered similarities were converted into a weighted protein similarity network, using the BLAST bit score as the edge weight.
- (6) Finally, proteins were clustered into families using the Markov Clustering Algorithm (MCL), with key parameters including the inflation value (I = 2.0) and removal of self-loops and singleton clusters.

Overall, the pipeline is based on what we have performed in the TD in class, but for the project we tried to make it fully reproducible and parameterizable, and varying identity, coverage, and e-value thresholds revealed that increasing stringency leads to lower duplication ratios and smaller protein families, which we tried to explore in depth next. The pipeline scripts can be found in folder `pipeline_1/`.

We tried 24 different filtrations based on identity, coverage and evalue, at this stage to really pick the thresholds that we will work with for the rest of the project. At the end we fixed in defining 2 datasets (more on that in results section):

- Low stringency on id 30%, cov 50% and evalue 1e-10
- High stringency on id 50%, cov 70% and evalue 1e-10

Pipeline 2

`pipeline_2/`: directory containing all scripts for this part of the pipeline.

Pipeline 2 was designed to estimate pairwise Ks values between duplicated genes within protein families identified in Pipeline 1.

The pipeline follows a protein-guided codon alignment strategy and consists of the following steps:

- (1) Prepare data: all unique gene pairs were generated within each protein family. Only the longest isoform per gene was retained to avoid redundancy. For each gene pair, a protein FASTA file and a corresponding CDS FASTA file were created.
- (2) Protein alignment: protein sequences were aligned pairwise using MAFFT.
- (3) CDS alignment: protein alignments were converted into codon-based CDS alignments using PAL2NAL, ensuring correct reading frames.
- (4) Control file generation: a control file was automatically generated for each gene pair using a template for PAML yn00.
- (5) Ks estimation: pairwise dN, dS, and Ks values were estimated using PAML yn00.

All steps were organized by gene family, with start/end parameters enabling iterative development, debugging, and partial reruns on selected subsets of families. Pipeline 2 was therefore first validated on a small subset of families from the low-stringency dataset before being executed on the full dataset (30% identity, 50% coverage, e-value = 1e-10). Under this configuration, the two largest gene families contained 1,308 and 707 genes, corresponding to approximately 855,878 and 249,171 pairwise comparisons, respectively. Running the full pipeline on these families required almost 10 days for the largest family and about 7 days for the second-largest family. The main computational bottlenecks were protein alignment and PAML yn00 execution, both scaling directly with the number of gene pairs.

These observations showed that applying the pipeline to all families using the initial configuration was not computationally tractable. This motivated the need for pipeline optimization. After optimization, the remaining parts of the pipeline were executed using the optimized code.

TAGs

Identification

The spacer-based approach we used to identify tandemly arrayed genes (TAGs) was based on several works like Lallemand et al. (2020), also mentioned in this review (Shoja & Zhang, 2006). In the review (Lallemand et al., 2020). Tandemly arrayed genes are defined as genes that are physically close on the chromosome and share high sequence similarity. As we have dealt previously with the definition of "sequence similarity" to be considered duplicated, what's left to define is "physically close". The gene spacer strategy revolves around setting a max spacer number, i.e. threshold of intervening genes, to consider two genes as tandemly duplicated. Usually this spacer number ranges between 0 (a perfect TAG cluster with no intervening genes) to 10, with 1, 5 and 10 being common choices. To choose ours, we refer to Shoja & Zhang (2006) who tried the 11 different spacer numbers from 0 to 10 on 3 different genomes (human, mouse, rat) and observed the increase in the number of TAGs detected with increasing spacer number.

We tried the same approach on our data, running the detection of TAGs with spacer numbers from 0 to 10 and plotting the results. We performed this through writing the script `script/TAGs_compute.sh` which is built on another R script `scripts/TAGs_detect.R` to detect TAGs with a given spacer number on 0:10 range (plotting in

`analysis/duplicated_genes/plot_TAGs_distribution.R`). We were able to do that using the ***GenomicRanges*** package in R (needs hours to run).

Orientation analysis

The orientation of TAG gene pairs can provide insights into their evolutionary history and functional relationships. To assess whether the observed orientation of TAG gene pairs deviates from random expectation, we applied a chi-squared test. TAGs were classified into three categories based on relative orientation: tandem (→ → or ← ←), convergent (→ ←), and divergent (← →). For each dataset, the observed number of TAGs in each orientation was compared to the expected counts under the assumption of equal probability across the three categories. The chi-squared statistic was calculated and a significant p-value indicates that TAG orientations are not randomly distributed, with a bias toward direct orientation reflecting the influence of unequal crossing-over as a major mechanism generating tandem duplicates.

CODE: the scripts we wrote for the orientation analysis can be found in `scripts/TAGs_pairs_orientation.R`, we used chi square test on both datasets,

Functional Analysis

Two independent datasets were analysed: a low-stringency dataset and a high-stringency dataset. Both datasets were processed using the same functional annotation workflow. As similar enrichment trends were observed across the two datasets, only results from the high-stringency dataset are presented for clarity.

Functional enrichment analyses were performed using the PANTHER Overrepresentation Test, with all *Glycine max* genes available in the PANTHER database (55,853 genes) used as the reference set. GO Slim annotations were used for Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) categories. However, CC terms were generally broad and showed limited differences between gene groups; therefore, they were not included in the interpretation.

Functional enrichment was assessed using Fisher's exact test to determine whether the proportion of genes associated with a given GO term in a query set was significantly higher than expected by chance, using the full *Glycine max* genome as the reference. For each GO term, a contingency table was constructed comparing the number of genes annotated and not annotated to the term within the query set against the corresponding counts in the reference genome outside the query set, allowing accurate estimation of enrichment. P-values were corrected for multiple testing using the False Discovery Rate (FDR) method. For each gene group, the top ten GO terms ranked by FDR were selected for further analysis.

Two complementary functional annotation strategies were applied. The first examined functional enrichment based on gene family size, testing whether small and large gene families differ in their functional roles. Gene families were classified as small (2–5 genes) or large (>5 genes). Several threshold values were tested, yielding consistent results, and PANTHER overrepresentation analysis was applied to each group. The second strategy focused on duplication type, comparing singleton genes, tandemly arrayed genes (TAGs), and non-TAG

duplicated genes as defined by our protocol. TAGs were identified using the gene spacer method with a maximum of one intervening gene, while non-TAGs consisted of duplicated genes not arranged in tandem. Singleton genes were defined as non-duplicated. All gene lists were analysed using the same PANTHER workflow.

Functional enrichment results were summarized using dot plots. In these plots, the x-axis represents the GeneRatio, defined as the proportion of query genes associated with a given GO term relative to the total number of reference genes annotated to that term. Dot size corresponds to the number of enriched genes in the query set, while dot colour reflects statistical significance expressed as −log10 of the FDR-adjusted p-value.

Transposable Elements

The classification of TEs followed the hierarchical structure provided by the APTE database, categorized by Class, Order, and Superfamily. The initial abundance analysis was performed at the order level to determine the distribution of Class I (retrotransposons) and Class II (DNA transposons). This was followed by a more detailed superfamily analysis where results were presented with full classification info, using panels organized by Class and grouping the data by Order to maintain the hierarchical context.

To understand the physical characteristics of these elements, a length distribution analysis was performed. The data was processed by separating TE orders into their respective Classes to visualize the spread of individual element lengths across the genome.

The coverage analysis was then conducted to map the chromosomal landscape of the repeatome. To determine the most informative visualization, several bin sizes were tested during the processing phase, including 10 kb, 50 kb, 100 kb, 500 kb, and 1 Mb windows. While these various scales were explored, the results presented here utilize 500 kb non-overlapping windows to provide the best balance of genomic detail and clarity. To calculate these values, the GenomicRanges R package was utilized; the algorithm identifies the intersection of TE coordinates within each genomic bin to sum the total occupancy. These results were then visualized as stacked bar plots across the 20 chromosomes.

To handle the structural complexity of the soybean genome, the analysis moved to the superfamily level to account for the way these elements are often fragmented or stacked on top of one another. To find the true genomic coverage for each superfamily, the GenomicRanges package in R was used to "reduce" the data, an algorithm that merges all overlapping or touching coordinates of the same superfamily into a single physical footprint so that the same piece of DNA isn't counted twice. From there, a Nesting Index was calculated by taking the simple sum of all annotated lengths for a superfamily and dividing it by its own merged "true" footprint. This index measures "self-nesting"; a score of 1.0 means the elements of that family usually sit alone, while higher scores identify families that frequently jump into members of their own group.

To identify relationships between different families, a pairwise nesting matrix was constructed for the top 20 most abundant superfamilies to ensure results remained clear and computationally efficient. The script utilized a combination of the `findOverlaps` and `pintersect` functions from GenomicRanges. The algorithm treated each superfamily as a Host (the container) and every other superfamily as a Guest (the nested element). For every possible pair,

`findOverlaps` identified the indices where a Guest and Host shared genomic space. A critical step was taken for self-nesting pairs (e.g., Gypsy vs Gypsy), where the script specifically ignored "self-hits" to ensure an element wasn't being counted as nested inside itself. Once the overlaps were identified, `pintersect` was used to calculate the exact number of base pairs shared between the Guest and Host coordinates. To make the data comparable across different groups, this overlap was normalized by dividing the shared base pairs by the total genomic length of the Guest superfamily. The final matrix was visualized as a heatmap using ggplot2, with a custom color scale, ranging from white to deep purple, to highlight which Guest families have the highest percentage of their sequence physically located inside the boundaries of specific Host superfamilies.

To investigate the genomic environment of different gene classes, the spatial proximity between TEs and genes categorized by duplication history was calculated. This process was performed independently for the two distinct classification datasets: High Stringency and Low Stringency. Genes were partitioned into three mutually exclusive groups: Singletons, TAGs, and Non-TAG Duplicates. The spatial relationship was defined using an edge-to-edge distance calculation. This method measures the shortest physical gap (in base pairs) between the nearest boundary of a gene (Start or End) and the nearest boundary of a TE. Under this logic, a distance of 0 bp was assigned to any gene where a TE is either immediately adjacent to the gene boundary or physically overlaps with the gene body, such as TEs located within intronic sequences. Distances were calculated using the GenomicRanges package in R and transformed using a log1p scale (log(x+1)) for visualization. Statistical significance was determined using a Pairwise Wilcoxon Rank Sum Test with Bonferroni correction.

RESULTS

Basic Statistics

Taking into consideration only longest isoforms, the duplication rate in *Glycine max* is quite high as can be seen in the bar plot, where >75% of genes in all chromosomes are paralogs. This can also be seen through a dotplot across the whole genome (20 chromosomes, in dotplot gm stands for *Glycine max*'s chromosome). Each dot represents a pair of **paralogous genes identified based on sequence similarity**, with their genomic positions plotted on the corresponding *Glycine max* chromosomes. The diagonal and off-diagonal blocks observed in the dotplot reflect extensive duplications of a variety of types, where some are proximal and others are segmental and dispersed. Highlighting the high duplication content of *Glycine max.*

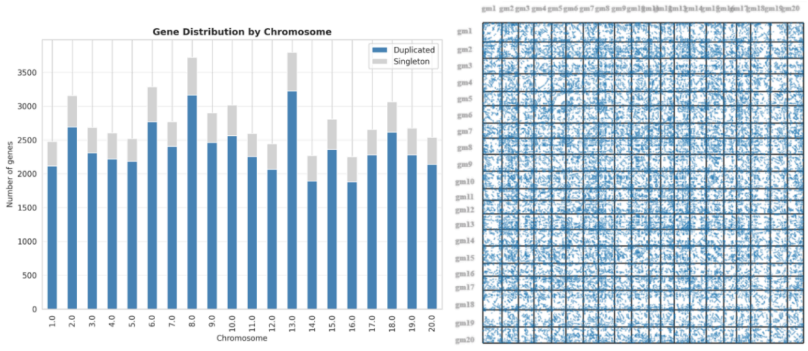


Figure 1. Left is a barplot showing distribution of duplicated and singleton genes across the 20 chromosomes, right is a dotplot of sequence similarity across 20 chromosomes

Gene families

Starting from blast results, we identify duplicated genes in the genome. First we used default thresholds to run `blastP` used in our pipeline, we tried performing the full analysis on a low stringency dataset (30% id, 50% cov) to check initial results knowing that we can use a more stringent dataset later on based on filtration thresholds. The dataset had a 99% duplication rate which led to a pretty exhaustive search space to work with in pipeline 2 in particular, requiring really high computational power to run. To be more exact:

30% identity, 50% coverage without e-value filter ⇒ 99% of genes are duplicated genes (89598 isoforms filtered to 56680 genes, 56144 out of them are duplicated genes)

From a large search space of the filtration thresholds, we got preliminary statistics on gene families to judge better what to go for, for the rest of the analysis.

Running pipeline on different thresholds to see how results change, by comparing number of duplicated genes in each file in `output/clusters/` containing gene families detected with different thresholds.

Table 1. 24 different runs of the pipeline generating 24 different sets of gene families based on different filtrations (id, cov, eval), reporting the %duplication and largest family size

id	coverage	evalue	percentage duplication (%)	largest family size
30	50	1e-10	91.79	1308
30	50	1e-5	92.65	1311
30	60	1e-10	91.33	1306
30	60	1e-5	92.14	1309
30	70	1e-10	90.18	1302
30	70	1e-5	90.90	1304
40	60	1e-10	90.65	841
40	60	1e-5	91.41	842
40	70	1e-10	89.49	838
40	70	1e-5	90.15	840
40	80	1e-10	87.77	834
40	80	1e-5	88.36	835
50	70	1e-10	88.13	349
50	70	1e-5	88.71	357
50	80	1e-10	86.45	348
50	80	1e-5	86.96	355
50	90	1e-10	84.72	347
50	90	1e-5	85.14	354
60	70	1e-10	85.77	210
60	70	1e-5	86.21	214
60	80	1e-10	84.12	208
60	80	1e-5	84.50	212
70	90	1e-10	78.16	177

We chose 2 datasets to work with for the rest of the analysis:

- **Low stringency**: 30% id, 50% coverage, 1e-10 e-value ⇒ 91.8% duplication rate
- **High stringency**: 50% id, 70% coverage, 1e-10 e-value ⇒ 88.1% duplication rate

Main results for each can be found in files suffixed with `_low` and `_high` respectively in subdirectories of the `output/` folder.

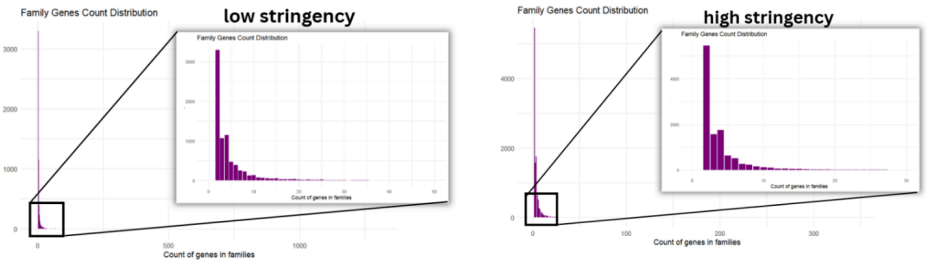


Figure 2. Family gene count distribution for the low and high datasets respectively

We notice the high stringency dataset has a largest family of size 349 genes, almost 4x smaller than the low stringency one: 1308 genes. As for the number of small families, we can see the

bar in the high stringency dataset hitting 4000s while it hits the 3000s in the low dataset. This is expected since the high stringency dataset is splitting the large families into many smaller ones due to higher filtration thresholds, this way the size of the largest family decreases and the number of small families increase.

Ks Distribution & WGD Events

We executed pipeline 2 to produce the pairwise synonymous substitution estimates (Ks) from BLAST-derived homologs (BLAST prefilters: ≥30% identity, ≥50% coverage). These Ks results were then filtered to retain reliable values for duplication-age analysis. Primary filters removed non-numeric or missing Ks, non-positive values (Ks ≤ 0) and saturated/high values (commonly Ks > 2.0) that confound peak detection. Larger Ks values (e.g., >0.75) are associated with increasingly large error due to substitutional saturation and the limitations of synonymous site estimation (Li, 1997). To minimize this error while retaining a reasonably sized data set, we adopted a Ks cutoff of 2.0, following the approach of Blanc and Wolfe (2004). This threshold ensures that the retained gene pairs represent reliable duplication events, while reducing the confounding effects of highly diverged or saturated pairs. Additional considered filters included minimum aligned length, minimum percent identity on the alignment used for Ks, and removal of problematic back-translations (frameshifts/stop codons). At each step we retained both the filtered and the full archived datasets.

Filtered Ks values are then visualized as histograms and smoothed density curves to reveal duplication-age peaks. When converting Ks to approximate ages, we use the molecular clock rate $\lambda = 6.1\times10^{-9}$ substitutions/site/year to compute age with the formula: Age (MYA) = Ks / (2 × λ) / 1e6. This lambda rate is reported for soybean and used in multiple publications in the literature (Duan et al., 2023; Miura et al., 2008).

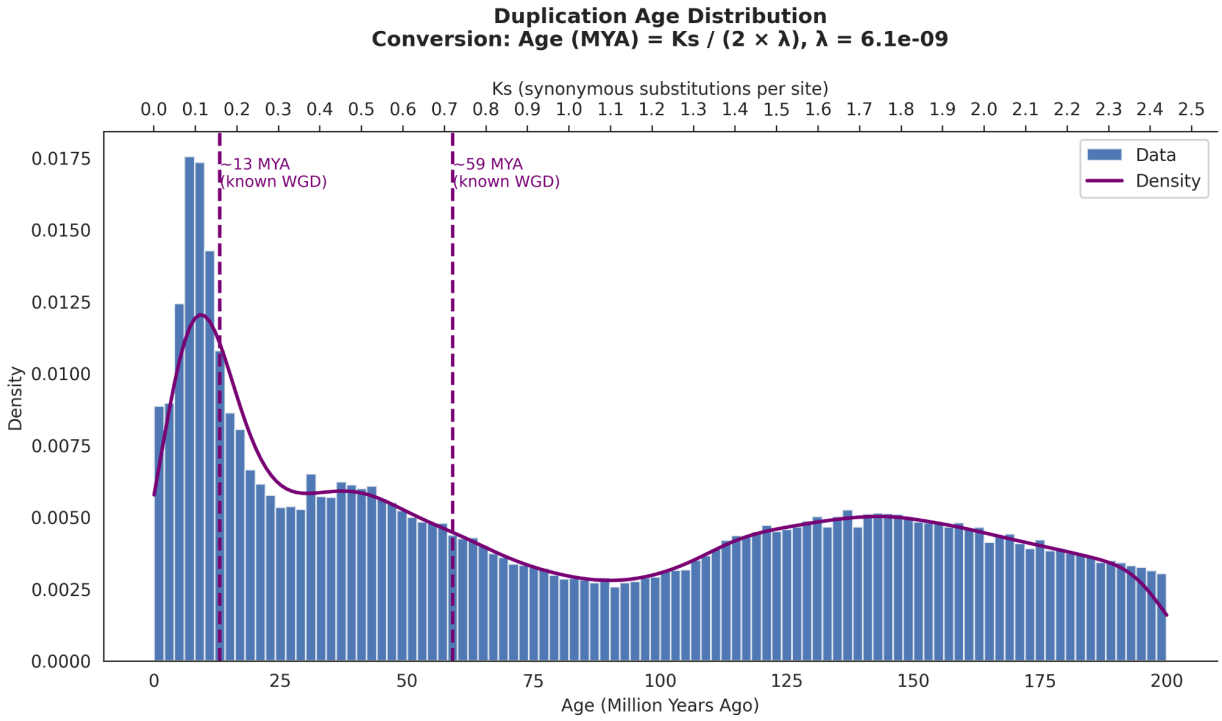


Figure 3. Ks distribution of all BLAST-derived homologous gene pairs in soybean, showing major duplication events/peaks. Both Ks and age scales are used concurrently, with the conversion formula and adopted lambda value highlighted. The literature-based duplication events (13 and 59 Mya) are shown.

We note two peaks, one around 0.1 and another around 0.5, as highlighted in Fig 3. This is in accordance with previous publications on soybean. Yang et al. (2013) reported peaks at 0.15 and 0.42, while Rulin et al. (2013) reported them at 0.1 and 0.53. However, the corresponding ages of these two duplication events of glycine max have been reported to be 10 and 56.5 Mya (Kim et al., 2015), or 13 and 59 Mya (Schmutz et al., 2010). Accordingly, the age of the second peak does not match the ks peak value using the previously highlighted lambda value for conversion.

According to Schmutz et al. (2010), this discrepancy is explainable by the fact that the older Ks peak (reported at Ks≈0.59) was assigned to an early-legume WGD by anchoring it to fossil evidence that dates the origin of the papilionoid legumes to ~58–60 Mya, which was then used to reverse compute an effective synonymous substitution rate of ~5.17 × 10^{-9} substitutions per site per year. The authors stated: "If the older duplication is assumed to have occurred around 58 Myr ago, then the calculated rate of silent mutations extending back to the duplication would be 5.17 × 10-3".

The more recent Glycine-specific WGD was dated independently using a higher lineage-specific substitution rate, resulting in an estimated age of ~13 Mya. The 6.1×10^{-9} lambda could be used for the recent event (glycine max-specific), but another rate has to be introduced to account for the older early-legume event. In our case, we deduced a lambda ~4.1 × 10^{-9}. We explain the

discrepancy from that reported by Schmutz et al. (2010) in the fact that they limited their analysis to families of sizes 2 to 6.

Rulin et al. also reported a third highly diffuse peak ~1.5 (Schmutz et al., 2010). This is also in accordance with our findings as could be seen in the plot. This likely corresponds to an ancient eudicot whole-genome triplication, which occurred near the origin of core eudicots (~115–130 Mya).

Anchors

To further clarify our peaks more distinctively, we used MCScanX to identify collinear anchor pairs (gene pairs retained in conserved blocks across the soybean genome). MCScanX detects these anchors by scanning for regions where gene order and orientation are preserved, indicating large-scale duplication events such as WGDs. By plotting the age distribution of these anchored pairs, we observe much more distinct peaks aligning with the previous reports in the literature, as shown in Fig 4. This approach ensures that the Ks peaks are specifically associated with large-scale, collinear duplications, providing a robust link between the observed molecular signatures and historical genome duplication events.

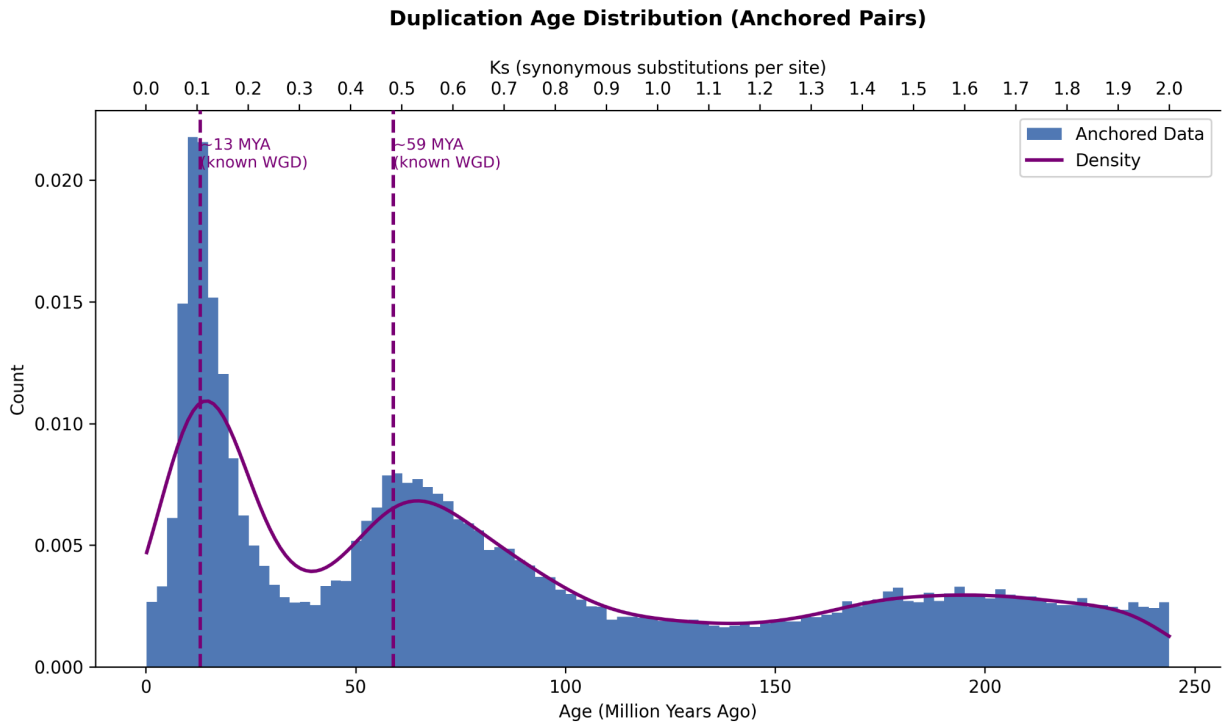


Figure 4. Ks distribution of collinear anchor pairs identified by MCScanX, highlighting distinct peaks associated with large-scale duplication events.

Filtering

Using our computed Ks values for all 2.4M pairs (at identity >30% and coverage 50%), we experimented with stricter filtering criteria to examine the effect on the Ks distribution plot. As highlighted in Fig 5, we note that only ~12.5% of pairs are retained for Ks range 0-2, with more stringent filtering primarily affecting the huge jump in the 1.5-2 region.

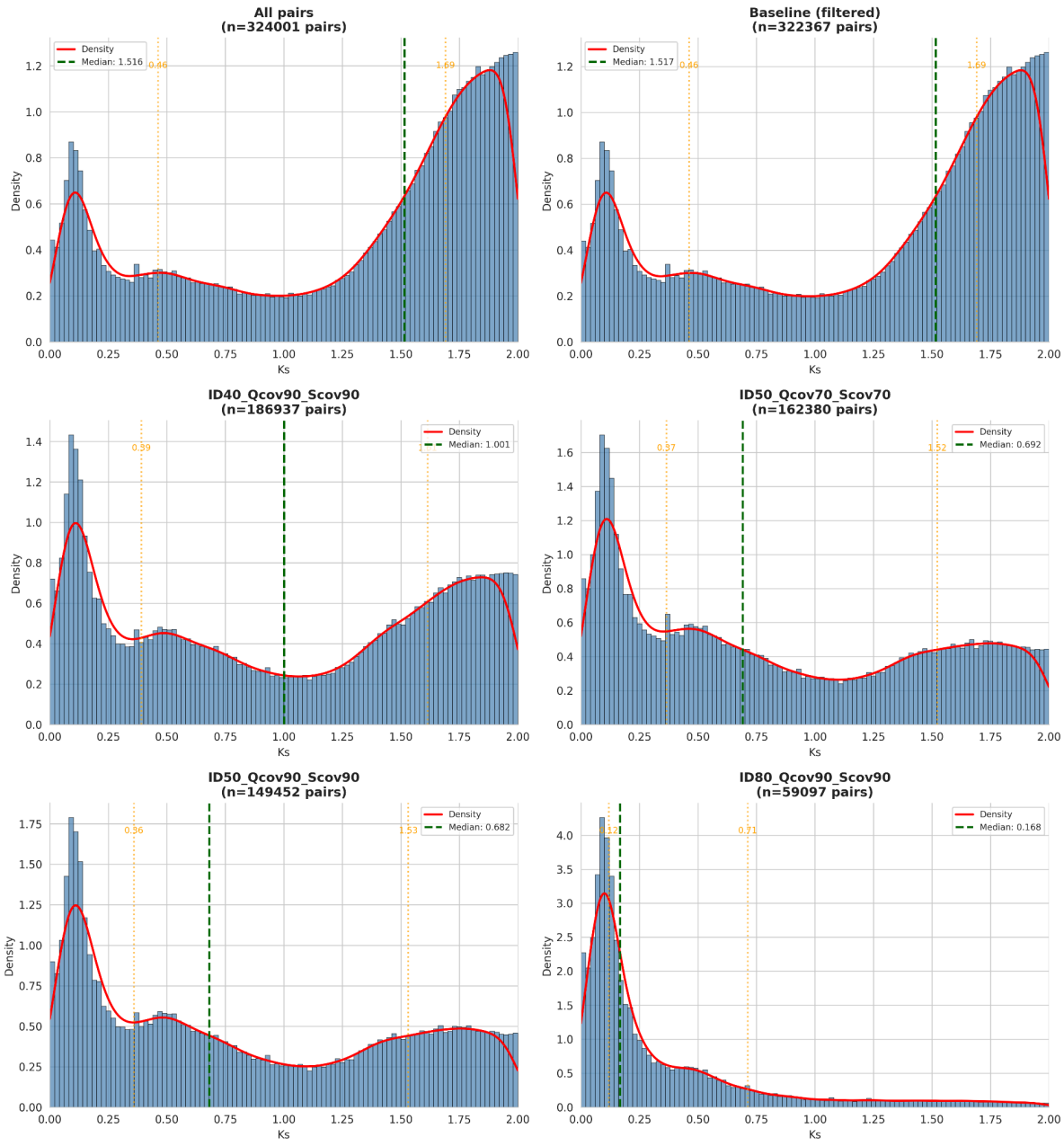


Figure 5. Comparison of Ks distributions under different filtering criteria, illustrating the effect of stricter thresholds on the retention of gene pairs.

Pipelines Automation

Beyond implementing pipeline 1 and 2 for Ks estimation, we fully automated the entire workflow to enable robust, reproducible, and scalable comparative genomics analyses. Each pipeline step—from data extraction and BLAST searches to alignment, Ks calculation, and downstream

filtering—was scripted for end-to-end execution, minimizing manual intervention and reducing error. Batch processing, parallelization, and checkpointing were integrated to efficiently handle millions of gene pairs, optimize runtime, and ensure resilience to interruptions. All parameters (e.g., identity, coverage, Ks thresholds) are configurable, allowing rapid adaptation to new species or datasets and facilitating comparative analysis across genomes.

The modular design supports easy replication and extension, with outputs and intermediate results systematically logged for transparency. This automation not only streamlines analysis but also enables efficient exploration of different filtering criteria and thresholds, ensuring that the workflow remains flexible and scalable for future studies. We validated our pipeline on additional species, demonstrating rapid and reliable replication of all analysis steps and outputs for new genomic datasets.

TAGs

Using the spacer approach, we observed a similar trend as in Shoja and Zhang (2006), with a rapid increase in the number of TAGs detected from spacer 0 to 1, then a slower one follows. Based on the same approach, we will consider spacer number 1 as our threshold to define TAGs in our data (more tables and figures in Appendix)

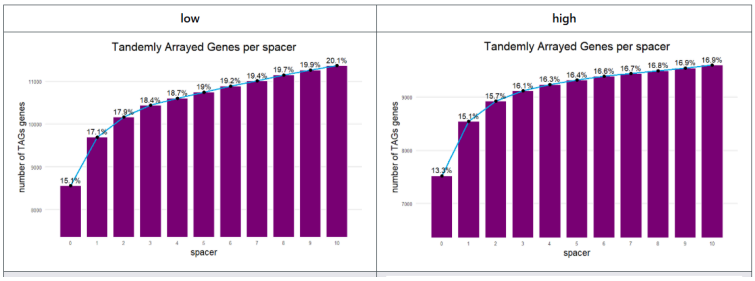


Figure 6. Number (and % out of duplicated genes) of identified TAGs using spacers from 0 to 10 for the low and high datasets (more figures in Appendix)

As for their orientation, both the low and high stringency dataset show a significant p-value for the orientation chi squared test. 82.5% of the low dataset TAGs, and 83.5% of the high one were shown to be parallel TAGs, following the same trend as in Rizzon et al. (2006) (more on that in the discussion).

	Low Stringency (count, %)	High Stringency (count, %)
Tandem	14,425 (82.5%)	12,685 (83.5%)
Convergent	1,637 (9.36%)	1,320 (8.69%)
Divergent	1,423 (8.14%)	1,192 (7.84%)
Chi-square test	X^2 = 7392.3, df = 2, p < 2.2e-16	X^2 = 6812, df = 2, p < 2.2e-16

* Orientation analysis on our identified TAGs genes with spacer=1

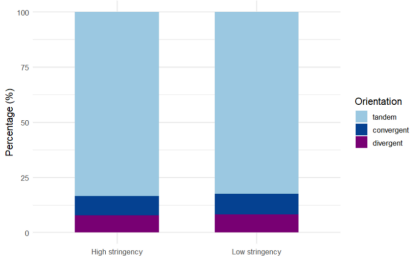


Figure 7. Left, table showing numbers of tandem convergent and divergent TAGs in low and high datasets respectively, right is a visualization of the table

Other approaches

We used multiple approaches to obtain TAGs. The venn diagram in Fig 8 highlights the overlap of predictions.

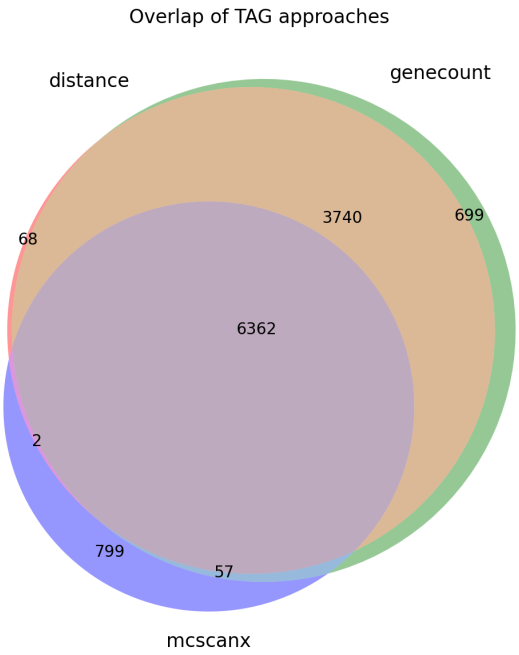


Figure 8. Venn diagram showing the overlap of tandem adjacent gene (TAG) predictions from distance-based, genecount-based, and MCScanX approaches.

The discrepancies observed between TAGs identified by the distance-based, genecount-based, and MCScanX approaches is due to their different definitions of tandem adjacency. The distance-based approach classifies TAGs as gene pairs located within a specified physical distance (≤100,000 base pairs by default) on the chromosome, regardless of the number of intervening genes. In contrast, the genecount-based method considers pairs as TAGs if they are separated by no more than 10 genes (by default), independent of the actual base pair distance. MCScanX predicts tandem adjacent genes (TAGs) using a collinearity-based approach. First, MCScanX identifies homologous gene pairs through sequence similarity (typically from BLAST results). It then scans chromosomes for collinear blocks—regions where homologous genes are arranged in the same order and orientation. Within these blocks, MCScanX applies a gap threshold (default: ≤10 intervening genes) to define whether genes are sufficiently close to be considered part of the same block. For TAG prediction specifically, MCScanX looks for homologous gene pairs that are directly adjacent (no intervening genes) on the same chromosome and belong to the same collinear block. Only pairs meeting these strict criteria—direct adjacency, shared block membership, and sequence homology—are classified as TAGs by MCScanX. This approach is more conservative than simple distance or genecount methods, as it requires both physical proximity and evidence of conserved genomic context.

The overlap in the Venn diagram represents TAGs consistently detected by all three methods, while the unique regions highlight pairs identified only by one or two approaches due to these

differing criteria. This comparison underscores the importance of method selection and parameter choice in tandem duplication analysis, as each approach captures distinct aspects of genomic organization.

Age analysis

We examined the age distribution of TAG pairs, as determined by their Ks values. The distribution reveals a concentration of younger TAGs, reflecting ongoing tandem duplication activity, with a tail of older events, as highlighted in Fig 9. This pattern supports the biological expectation that tandem duplications are a continuous process, contributing to gene family expansion and functional diversification.

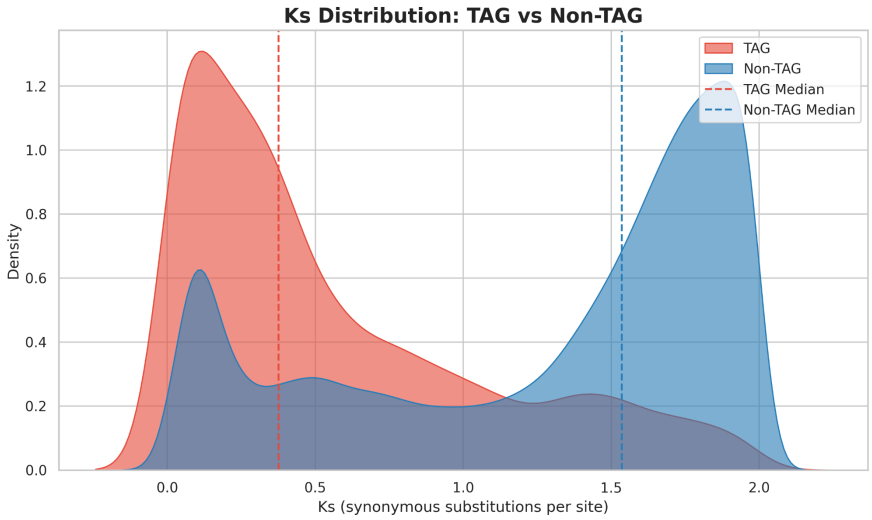


Figure 9. Ks density distributions for TAG versus non-TAG gene pairs, demonstrating the age bias of tandem duplications.

Structural Analysis

Our structural analysis of duplicated gene pairs in soybean reveals a significant enrichment near chromosome ends, in accordance with what was reported by Yang et al. (2013). As shown in the bar plot and histogram in Fig 10, a substantial proportion of both tandem (TAG) and whole-genome duplication (WGD) pairs are located within 4 Mb of the chromosome termini. Quantitatively, we observe that approximately 40% of duplicated gene pairs are found in these subtelomeric regions. This spatial bias was also examined per chromosome, as highlighted in Fig S4 for chromosome 1, with centromere region highlighted as well. The frequency of gene pairs declines with increasing distance from the ends and peaks within the first few megabases.

The per-chromosome histograms further illustrate that this enrichment is consistent across individual chromosomes, with the majority of TAG and WGD pairs clustering near the telomeric regions. The centromeric regions, highlighted in purple, show a marked depletion of duplicated pairs, supporting the notion that chromosomal ends are hotspots for gene duplication retention. These results reinforce the conclusion that the chromosomal environment, particularly proximity to telomeres, plays a significant role in the retention and distribution of duplicated genes in soybean, as previously described by Yang et al. (2013).

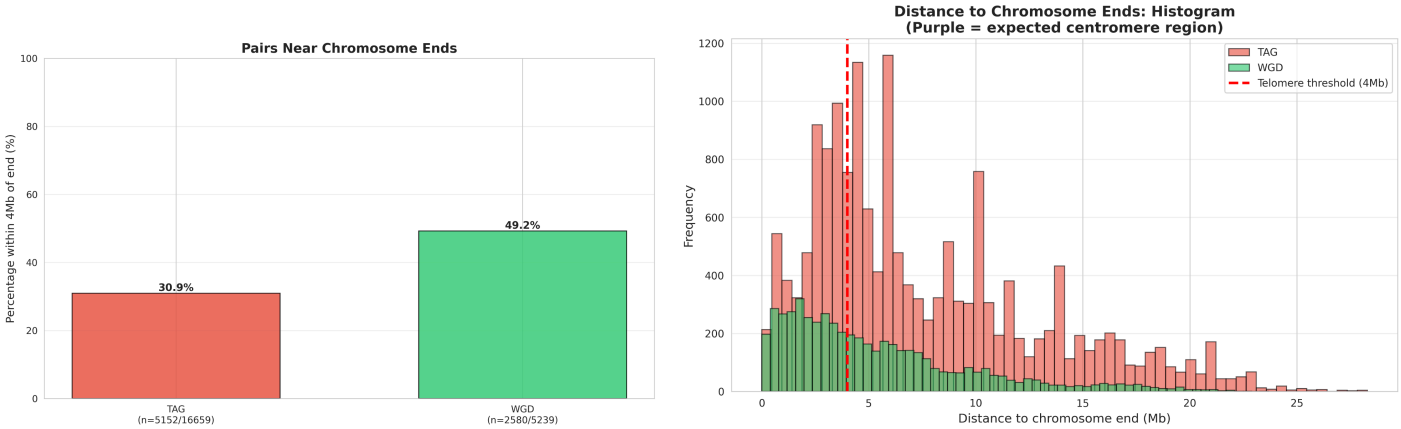


Figure 10. (Left) Percentage of TAG and WGD gene pairs located within 4 Mb of chromosome ends, indicating enrichment near telomeres; (Right) Histogram of distances to chromosome ends for TAG and WGD pairs, across all chromosomes, with telomere threshold (4 Mb) marked in red.

Functional analysis

Table 2. Number of genes per functional group in the low and high stringency datasets

	Singletons	Tags	Non Tags	Small family	Large family
High dataset	6723	8545	41411	25853	24103
Low dataset	4648	9696	42335	16748	35283

Functional Enrichment by Duplication Type

Biological Process enrichment revealed that singletons are primarily associated with genome maintenance, DNA repair, DNA/RNA modification, and core nucleic-acid metabolic processes. This suggests that singleton genes are essential for genome stability and fundamental cellular functions. TAGs were enriched in processes related to defense responses, response to stimulus, and stress-related mechanisms, indicating their role in adaptation to environmental challenges. Non-TAG duplicated genes were enriched in biosynthetic and general metabolic processes, highlighting their participation in core cellular metabolism.

Molecular Function analysis further supported these distinctions. Singleton genes showed enrichment in DNA-interacting enzymatic activities and catalytic functions related to genome maintenance, confirming their housekeeping roles. TAGs were enriched in catalytic, oxidoreductase, and glutathione transferase activities, reflecting functions in adaptive metabolism and detoxification. Non-TAGs were mainly associated with RNA and DNA binding, transcription regulation, and core catalytic functions, emphasizing their role in transcriptional regulation and essential cellular processes.

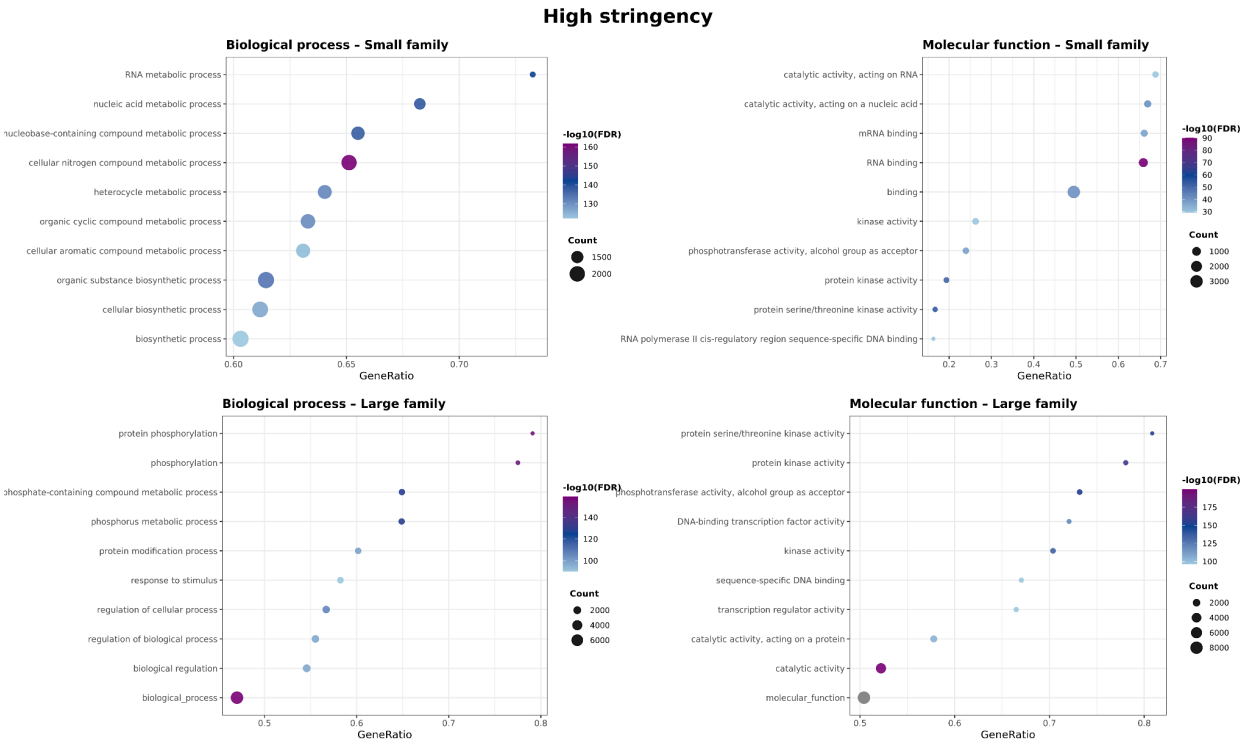


Figure 11. Functional enrichment by duplication type in Glycine max

Functional Enrichment by Gene Family Size

Small gene families (2–5 genes) were enriched in biological processes such as core metabolic and biosynthetic pathways. Molecular Function analysis revealed associations with RNA and mRNA binding, RNA-related catalytic activities, and basic enzymatic functions. These results suggest that small families primarily encode proteins involved in essential biochemical and post-transcriptional processes.

Large gene families (>5 genes) were enriched in biological processes such as protein phosphorylation, protein modification, response to stimulus, and regulation of cellular processes. Molecular Function enrichment highlighted protein kinase activity and DNA-binding transcription factor activity, indicating that these families support regulatory, signalling, and adaptive functions.

⇒ This distinction between small and large gene families aligns with previous observations that gene family expansion often contributes to functional innovation, while small families maintain essential housekeeping roles.

⇒ The two complementary analyses together show that gene function in Glycine max is closely associated with both duplication type and family size. Singletons and small families are mainly involved in essential metabolic and genome maintenance functions, whereas TAGs and large families are enriched in adaptive, regulatory, and stress-responsive functions. These results highlight how gene duplication and family expansion contribute to functional specialization in the genome.

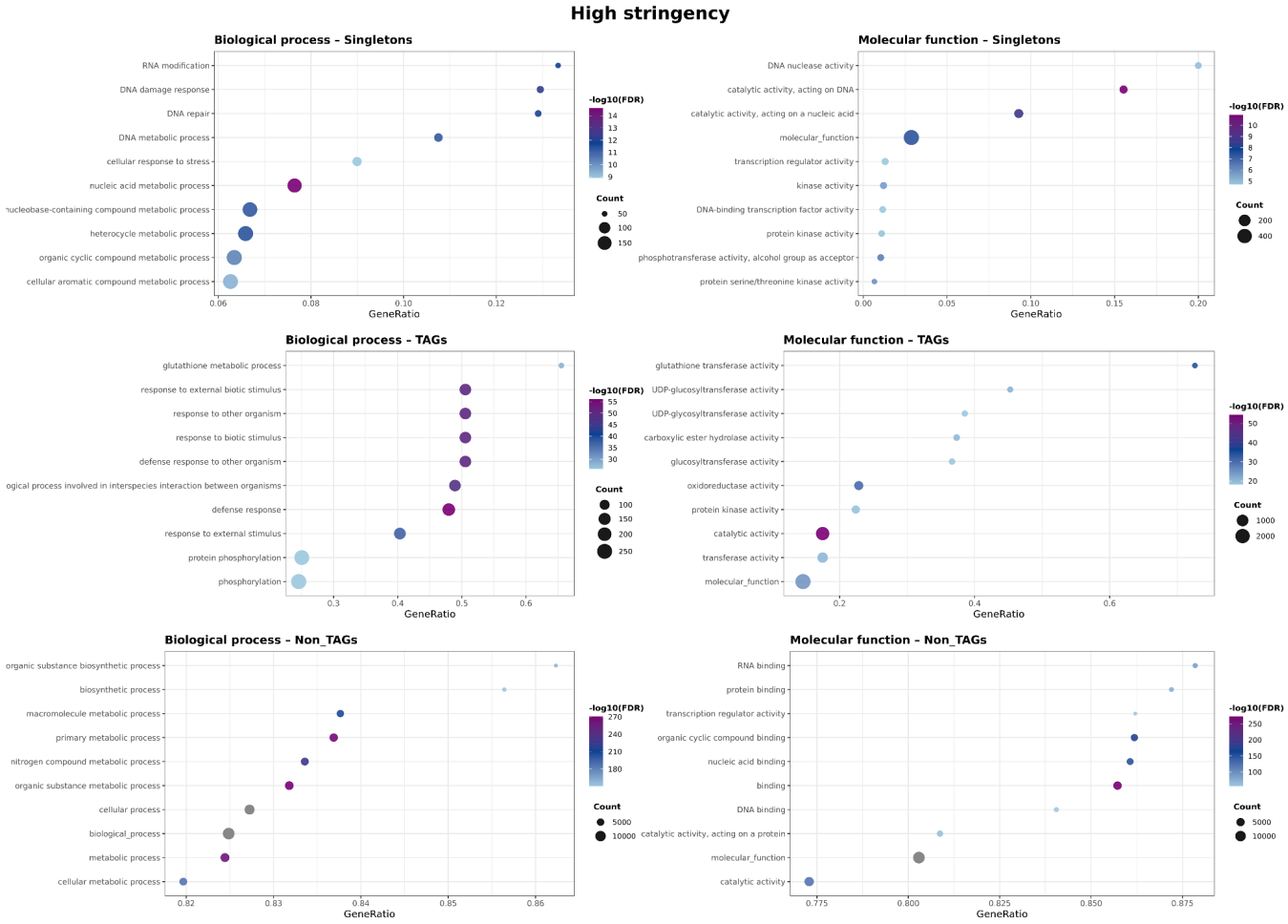


Figure 12. Functional enrichment by gene family size in Glycine max

Transposable Elements

Abundance Analysis

The numerical distribution of transposable elements (TEs) in the *G. max* genome was analyzed at multiple taxonomic levels, as presented in Figure 13. The total dataset consists of 966,883 annotated TEs. As detailed in the summary table (Figure 13A), a significant majority of the repeat landscape (64.07%) remains unclassified at the order level. Among the identified sequences, Class I (Retrotransposons) are the most abundant, representing over 22% of the total count. Figure 13A shows that the LTR order is the primary driver of this class, accounting for 187,975 elements or 19.44% of all annotations. The superfamily-level abundance in Figure 13B further illustrates that the LTR group is heavily dominated by the Gypsy superfamily, followed by Copia, with other families like LARD and TRIM contributing much smaller fractions. Class II (DNA Transposons) comprise 13.45% of the total annotations. The TIR order is the primary component of this class (11.81%), characterized by a high degree of superfamily diversity. As shown in Figure 13B, this diversity is led by the MuLE-MuDR and CMC-EnSpm superfamilies. Smaller numerical contributions to the overall repeat profile are provided by other

orders, including LINEs (2.31%), MITEs (0.97%), and Helitrons (0.61%), as visualized in the relative proportions across both panels of Figure 13B.

A.

Order	Class	Count	%
Unknown	Unknown	619,524	64.07
LTR	Class I	187,975	19.44
TIR	Class II	114,148	11.81
LINE	Class I	22,294	2.31
MITE	Class II	9,355	0.97
SINE	Class I	7,190	0.74
Helitron	Class II	5,860	0.61
Crypton	Class II	535	0.06
DIRS	Class I	2	<0.01
Total		**966,883**	**100**

B.

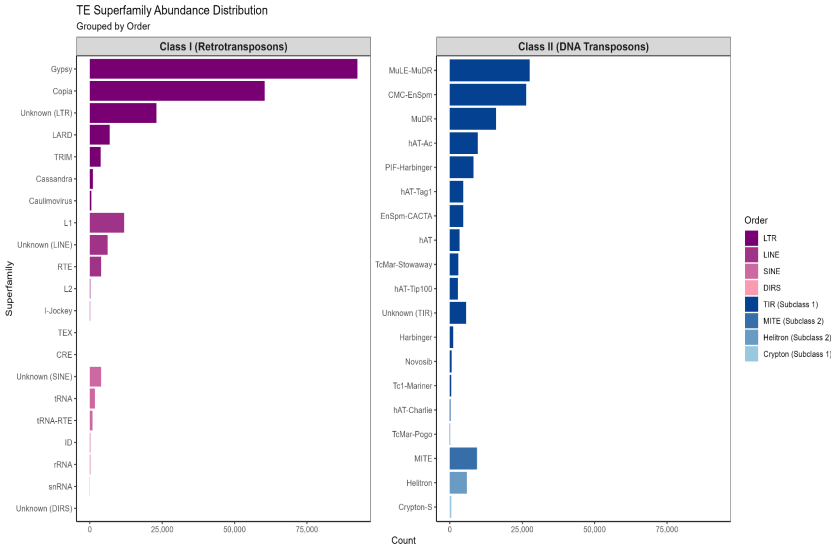


Figure 13: Taxonomic Abundance of TEs in G. max. A. Total abundance and frequency (%) of TEs categorized by Order and Class (Class I: Retrotransposons; Class II: DNA Transposons). B. Superfamily-level distribution of element counts faceted by Class and colored by taxonomic Order.

Length Distribution Analysis

A length distribution analysis was performed on all the TEs to compare the size frequency of each order across the three genomic classes shown in Figure 14. Class I (Retrotransposons): This group has the most variety in length. Elements like DIRS and SINEs show very high, sharp peaks between 100 bp and 300 bp, meaning they are mostly found as short, uniform sequences. In contrast, LTRs and LINEs have much wider, flatter shapes that go up to 10 kb or 15 kb, showing they are the main source of long, full-length elements in the genome. Class II (DNA Transposons): These are generally much shorter and stay within a smaller size range. Most orders, like TIRs and MITEs, are concentrated between 100 bp and 1 kb. While Cryptons have a clear peak at 500 bp, Helitrons are more spread out across the smaller sizes. Unknown

Elements: The unclassified TEs show a major peak around 150 bp. This curve looks very similar to the small fragments seen in the other orders, which suggests that most "Unknown" TEs are likely broken or degraded pieces of known families that are now too small to identify.

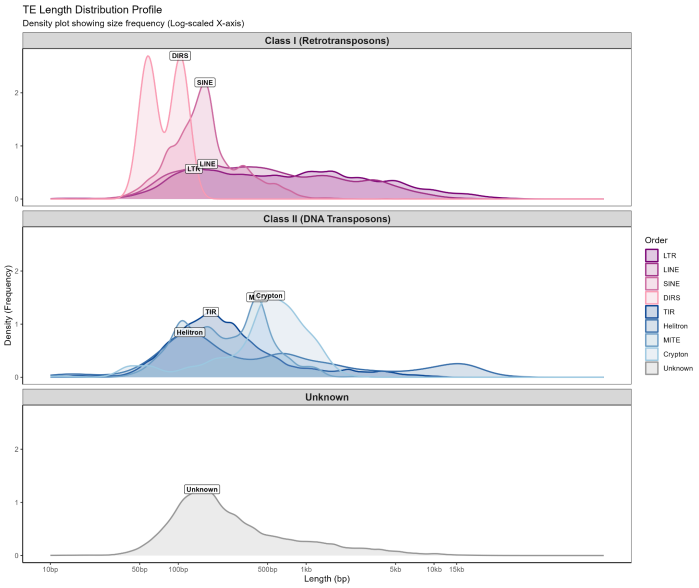


Figure 14. TE Length Distribution Profile. Density plots showing the size frequency of all TEs on a Log-scaled X-axis. Panels show Class I (Retrotransposons), Class II (DNA Transposons), and Unknown elements.

Coverage and Nesting Analysis

The spatial distribution of TEs across the 20 chromosomes is shown in Figure 15, visualizing coverage in 500 kb windows. While many chromosomes exhibit a symmetrical, bell-shaped pattern with TEs concentrated in the central pericentromeric regions, several chromosomes deviate from this standard layout. Chr 13 stands out significantly as its high-density TE clusters are heavily skewed toward one end of the chromosome rather than the center. Other chromosomes, such as Chr 8, Chr 10, and Chr 15, also show asymmetrical distributions where the main "mountain" of LTR (purple) and Unknown (gray) elements is shifted away from the middle of the chromosome.

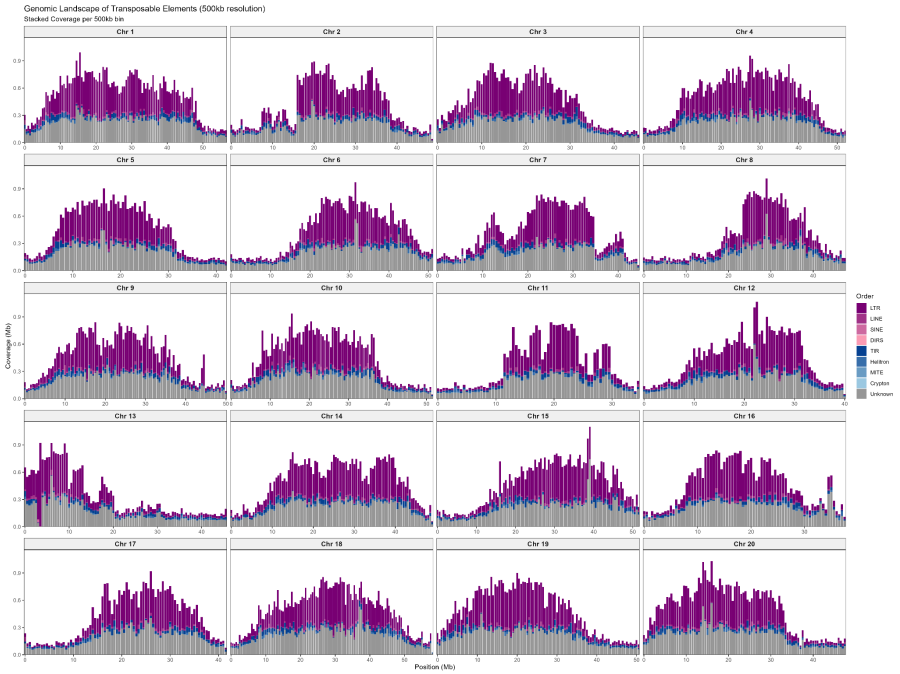


Figure 15. Chromosomal Distribution of TEs in G. max. Stacked bar plots showing the genomic coverage (Mb) of transposable elements in 500 kb windows across all 20 chromosomes. Bars are colored by order, with Class I (Retrotransposons) elements in purples, Class II (DNA Transposons) in blues, and Unknown elements in gray. The Y-axis represents the total Mb of TE sequence within each window.

After looking at how these elements are spread out across the chromosomes, Figure 16 breaks down exactly how much of the genome each specific superfamily occupies. TEs take up more than half the soybean genome at 56.149% (533 Mb). From Figure 16A, it's clear that Class I (Retro) elements are the main players here; LTR/Gypsy alone covers nearly a quarter of the genome (24.25%), followed by LTR/Copia at 8.77%. In contrast, there is no single dominant group within Class II (DNA); even the most frequent superfamily, TIR/MuDR, only reaches a maximum of 1.84% of the genome.

To get a better sense of how these sequences are structurally organized, Figure 16B uses an internal nesting index to measure "self-stacking" within each superfamily. LTR/Gypsy and LTR/Copia show the highest indices (1.12 and 1.10 respectively), which indicates that a percentage of their total genomic length is composed of elements from the same superfamily inserting into one another. While most Class II (DNA) elements stay right at the 1.0 baseline, meaning they typically exist as single, independent segments, there are notable exceptions. For example, the TIR/Unknown and TIR/MULE-MuDR groups show a bit higher indices.

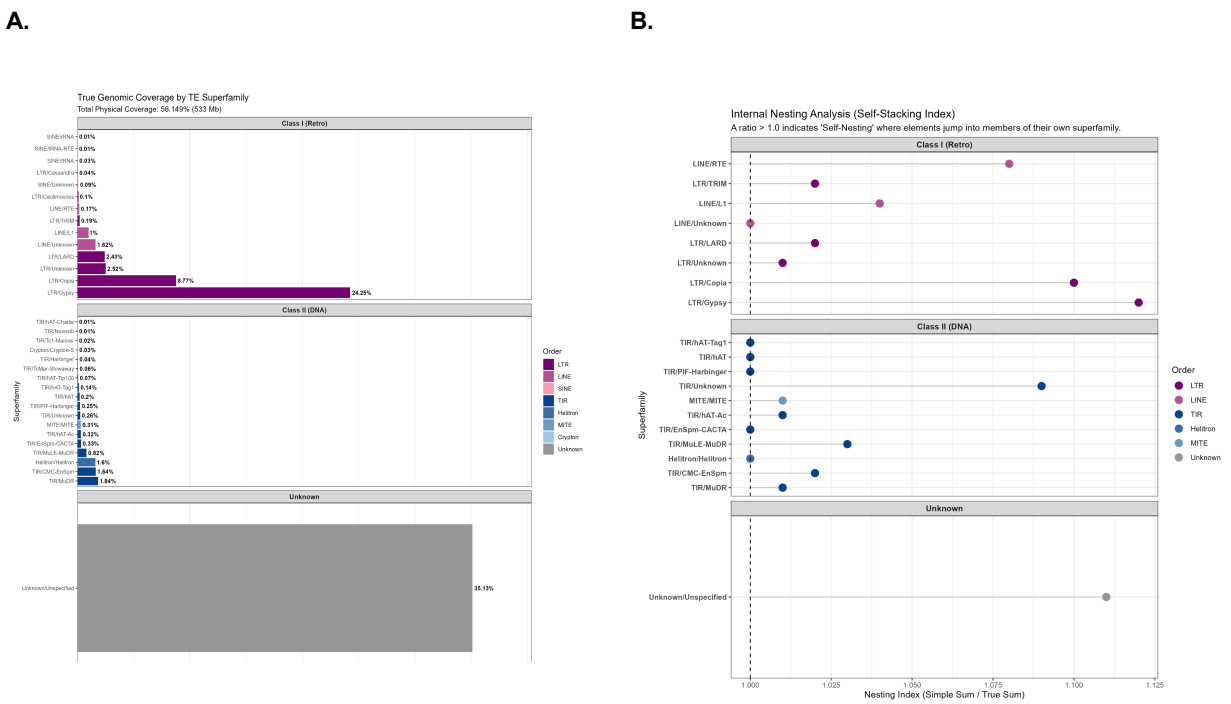


Figure 16. Genomic coverage and nesting of TEs. A. Total genomic coverage for each TE superfamily in G. max. Percentages show the actual DNA coverage after merging overlapping sequences. B. Nesting index by superfamily, calculated as the ratio of total annotations to true coverage. A ratio above 1.0 identifies "host" groups where other elements have inserted themselves.

Finally, the nesting matrix in Figure 17 maps out these relationships to show "who is inside whom". The darkest purple squares represent the highest nesting percentages and are found where LTR/Copia and LTR/Gypsy stack inside themselves, as well as where LTR/LARD and LTR/TRIM are inserted into LTR/Copia or LTR/Gypsy. While LTR/LARD and LTR/TRIM do show self-nesting, the color is lighter compared to their high preference for jumping into the larger Copia and Gypsy hosts. Additionally, Class II elements like Helitrons and TIR/Harbinger frequently use these same large LTR retrotransposons as their physical containers. Overall, these results show that the soybean genome is built in layers, with the largest retrotransposons providing the main foundation for both different TE classes and members of their own group to insert into.

Gene-TE Proximity by Duplication Type Analysis

Proximity analysis reveals that the spatial relationship between genes and transposable elements (TEs) follows a consistent pattern across both High and Low stringency datasets (Figure 18A, 18B). While all gene groups show a high frequency of physical association with TEs, Singletons exhibit a notably broader distribution. The upper quartiles and whiskers for Singletons extend significantly into the kilobase range (100 bp to >1 kb), indicating that a substantial portion of these genes occupy relatively repeat-depleted genomic regions.In contrast, the distributions for TAGs and Non-TAG Duplicates are heavily compressed at the 0 bp

baseline in both analyses. This demonstrates that duplicated gene classes are almost exclusively restricted to TE-dense environments. Statistical analysis confirms that the spatial distribution of Singletons is significantly distinct from both duplicated categories ($p < 0.0001$, Wilcoxon test), highlighting their relative isolation from repetitive elements compared to tandem and other duplicates.

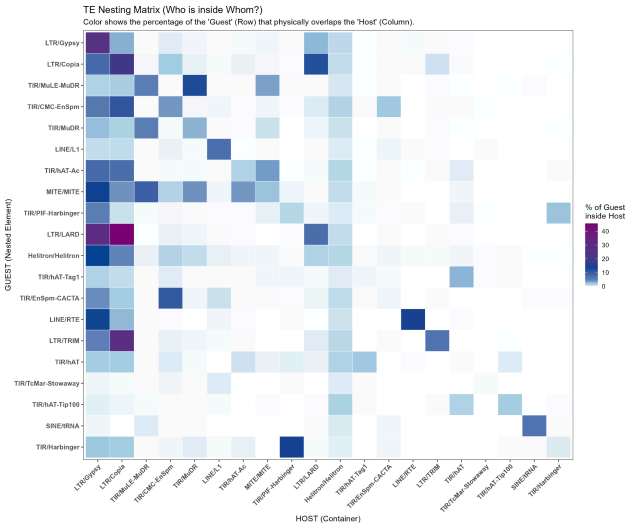


Figure 17. TE nesting relationships. Heatmap showing which TE superfamilies act as "guests" inside others. Colors represent the percentage of a Guest (row) found within the physical space of a Host (column).

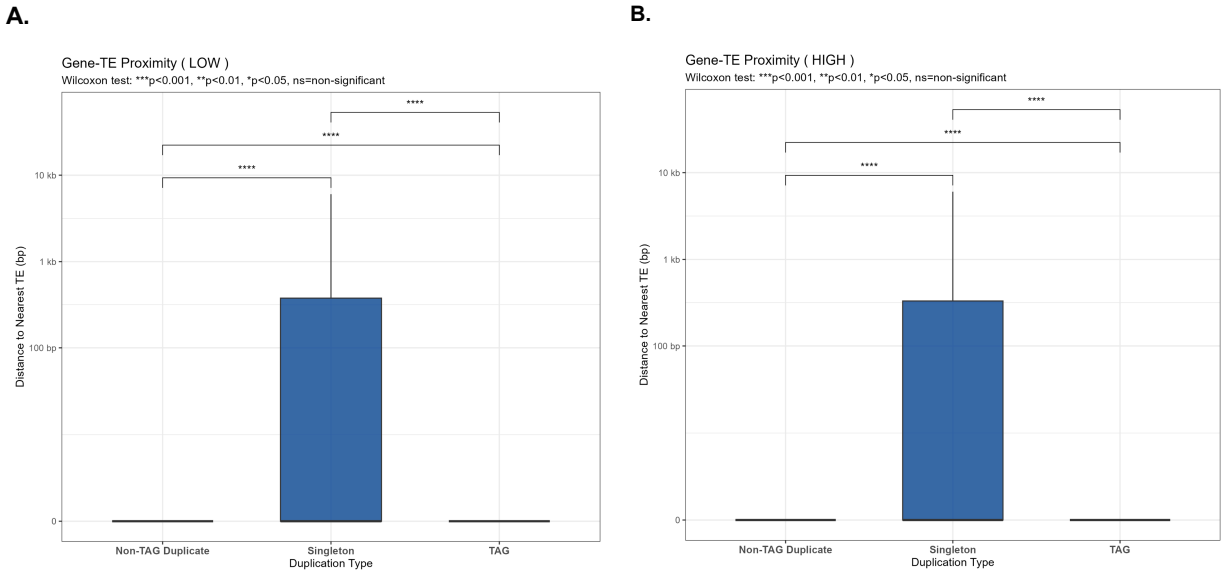


Figure 18. Spatial proximity between transposable elements and gene duplication classes in Glycine max. Boxplots represent the edge-to-edge distance (bp) to the nearest transposable element (TE) for Singletons (blue), Non-TAG Duplicates (purple), and TAGs (peach). Analysis is

shown for (A) Low Stringency and (B) High Stringency duplication classification datasets. The y-axis is log1p transformed. Distances of 0 bp indicate physical overlap or immediate adjacency between genes and TEs. Statistical significance was determined via Pairwise Wilcoxon Rank Sum Tests with Bonferroni correction.

Discussion

Gene Families

In Lallemand et al. (2020), the work was performed on rice, with a relatively higher number of initial genes (42534, closer to our number than the other studies that used the same protocol we used), they used 2 filtering strategies:

- 30% identity and 70% coverage ⇒ low stringency dataset (reduced to ~18k)
- 50% identity and 90% coverage ⇒ high stringency dataset (reduced to ~9k)

Several other studies actually tried analyzing different stringency datasets.

Our genome is much larger and has more duplicates, thus we also considered having 2 datasets, and considered to increase stringency from default ones (30% id and 50% cov). We wanted to see a bigger picture like that referenced work, that's why we tried 24 different thresholds to pick a low and high datasets, and the ones we picked (30/50 and 50/70) show duplication rates >88% in both of them: the high duplication rates indicate a rich search space for downstream analyses (pipeline 2)

The gene family sizes follow a power law distribution: **a lot of families are of small size and few of them are large.** This pattern can be seen in both the low and high stringency datasets and is actually expected.**To explain this biologically,** gene duplications happen all the time, but most duplicates are quickly lost. Only some survive long-term, either keeping the original function, splitting it (subfunctionalization), or gaining a new one (neofunctionalization). This creates many small families and few large ones.

TAGs Analysis

The spacer-based approach successfully identified TAGs in *Glycine max*, and the trend of increasing TAG detection with spacer number was consistent with prior studies. Selecting a spacer of 1 balances detection, as it allows a level of stringency as to keep into consideration only genes that are very close (max 1 gene in between), and is more loose than a perfect TAG with spacer 0 that will only consider directly neighbor genes which would miss some biological events like having a gene inserted in between.

As TAGs originate from duplicated genes that are positioned next to each other on a chromosome, unequal crossing-over (UCO) is a recombination-based process in which misaligned homologous chromosomes exchange genetic material, generating these tandem duplicates typically in direct orientation. According to Rizzon et al (2006), tandemly arrayed genes (TAGs) are predominantly arranged in direct orientation, reflecting the influence of unequal crossing-over (UCO) as a major mechanism generating TAGs. It showed that around 80% of TAGs in rice and 88% in Arabidopsis are in direct orientation, suggesting that UCO is

likely the primary driver of tandem gene duplication, although other evolutionary processes such as selection, recombination, gene gain, and loss may also play a role.

In fact, Shoja & Zhang (2006) also observed that in human, mouse, and rat genomes, the majority of TAGs are arranged in direct orientation, with percentages being 68%, 76%, and 72%, respectively. And they performed a chi-squared test to confirm the significance of these percentages compared to what would be expected by chance.

Biologically speaking, what's interesting about studying TAGs and the results we got from a >80% tandem TAGs, they show that they are going to be expressed together. And as they are tandemly expressed, they are implied together in the same work, hence same or similar function. The nearby location comes as an *advantage.*

Functional Analysis

Functional enrichment analyses revealed clear functional differences among gene categories in *Glycine max*. Singleton genes were predominantly enriched in biological processes related to DNA repair, DNA metabolic processes, and genome maintenance, indicating that single-copy genes are mainly involved in essential housekeeping functions.

Consistent patterns have been reported in other plant species. In *Arabidopsis thaliana*, single-copy genes are enriched in DNA repair and DNA replication processes (De Smet et al., 2013), supporting the idea that genes performing essential cellular roles are preferentially retained as single copies in plant genomes.

In contrast, TAGs in *Glycine max* were enriched in stress response, defense response, and stimulus-related biological processes, as well as oxidoreductase and transferase activities. Similar functional categories have been observed for TAGs in pear (Qiao et al., 2018), suggesting that tandem duplication commonly contributes to the expansion of genes involved in environmental adaptation in plants.

More broadly, plant genome studies have shown that gene family expansion is often associated with regulatory and adaptive functions, whereas genes involved in core cellular processes tend to be retained in smaller families (Panchy et al., 2016). Together, these observations indicate that the functional biases observed in *Glycine max* reflect general trends reported across plant genomes, supporting the robustness of our functional enrichment results.

Transposable Elements

The analysis shows that transposable elements (TEs) occupy 56.14% of the soybean genome. This total coverage is consistent with the original genome study, which estimated that approximately 57% of the *G. max* sequence consists of repeat-dense heterochromatin (Schmutz et al., 2010). While the physical coverage matches these historical estimates, the number of individual annotations in this dataset is significantly higher than in previous reports. By using the APTEdb atlas (Pedro et al., 2018) instead of the older SoyTEdb (Du et al., 2010), there is a nearly 900% increase in the number of identified TE annotations. This difference in count likely reflects the high sensitivity of the APTEdb pipeline, which is updated for the current genome version and is designed to capture smaller or putative sequences that might have been grouped together or excluded in earlier studies.

This sensitivity may also explain why 64.07% of the elements are labeled as "Unknown." The length distribution (Figure 14) shows that these unclassified sequences are mostly short, which could suggest they are older, broken pieces of DNA that have lost the specific features needed for a clear classification (Ma et al., 2004). However, it is also possible that a portion of these "Unknown" elements represents novel or highly divergent TE families unique to the soybean lineage that are not yet well-defined in standard classification databases.

The spatial distribution of TEs across the 20 chromosomes (Figure 15) follows a distinct pattern that varies between individual chromosomes. While most chromosomes show a "centric" accumulation of TEs in the middle, others show density peaks that are shifted significantly toward the ends. To understand these different patterns, we referred to the physical map of the soybean genome provided by Schmutz et al. (2010). Their work (Figure 1, Schmutz et al., 2010) confirms that soybean chromosomes are not all uniform; while many are metacentric, several others are submetacentric or have centromeres located off-center. This explains why we see shifted peaks in our results for chromosomes such as Chromosome 13, Chromosome 8, and Chromosome 17. Furthermore, recent studies on the soybean pan-centromere have shown that centromeres can undergo "repositioning", where the functional centromere moves to a new location on the chromosome over time (Liu et al., 2023). By comparing our plots to these established genomic maps, it is clear that the TE distribution accurately highlights the recombination-suppressed heterochromatin. These areas serve as regions where TEs accumulate, keeping them separate from the gene-rich portions of the genome (Schmutz et al., 2010).

The Nesting Index and Pairwise Matrix (Figures 16 and 17) show that certain TE families have a clear preference for where they insert. LTR/Gypsy and LTR/Copia are the most frequent "Hosts" for other elements, which is expected as they are the dominant elements in *G. max* and drive genome expansion (Fatima et al., 2024). The high frequency of elements jumping into these LTR groups supports the hypothesis that large retrotransposons serve as primary targets for new insertions. According to literature, this layered organization is thought to be a survival strategy that allows the genome to manage its high repeat content while keeping coding regions stable (Bennetzen, 2000).

The proximity patterns suggest that different gene types occupy distinct genomic environments in the soybean genome. Singletons are often located further from TEs, likely because these unique genes are essential and are maintained in stable, repeat-depleted regions to avoid disruption. In contrast, the 0 bp distance observed for TAGs and Non-TAG Duplicates (which contain dispersed and proximal copies) indicates a constant physical association with TEs. This measurement shows that TEs are either immediately adjacent to gene boundaries or residing within introns.This close association suggests a high tolerance for repetitive sequences in these regions or the result of TE-mediated duplication, where transposable elements facilitate the movement and copying of gene sequences. According to studies in soybean (Du et al., 2010), tandem and dispersed duplicates are frequently found in these repeat-rich areas. This indicates that the duplication mechanism is closely linked to the transposable element landscape, where duplicates are either born in or relocated to these high-repeat environments.

Bibliography

Eckardt, N. A. (2004). Two genomes are better than one: widespread paleopolyploidy in plants and evolutionary effects. The Plant Cell, 16(7), 1647–1649.

Correa, M., Lerat, E., Birmelé, E., Samson, F., Bouillon, B., Normand, K., & Rizzon, C. (2021). The transposable element environment of human genes differs according to their duplication status and essentiality. Genome Biology and Evolution, 13(5), evab062.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. Nature, 463, 178–183. https://doi.org/10.1038/nature08670

Lallemand, T., Leduc, M., Landès, C., Rizzon, C., & Lerat, E. (2020). An overview of duplicated gene detection methods: why the duplication mechanism has to be accounted for in their choice. Genes, 11(9), 1046.

Shoja, V., & Zhang, L. (2006). A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Molecular Biology and Evolution, 23(11), 2134–2141.

Rizzon, C., Ponger, L., & Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. PLoS Computational Biology, 2(9), e115.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., & Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proceedings of the National Academy of Sciences of the United States of America, 110(8), 2898–2903.

Qiao, X., Yin, H., Li, L., Wang, R., Wu, J., Wu, J., & Zhang, S. (2018). Different modes of gene duplication show divergent evolutionary patterns in pear (Pyrus bretschneideri). Plant Molecular Biology, 98(4–5), 337–348.

Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiology, 171(4), 2294–2316.

Duan, X., Zhang, K., Duanmu, H., & Yu, Y. (2023). The myosin family genes in soybean: genome-wide identification and expression analysis. South African Journal of Botany, 160, 338–346. https://doi.org/10.1016/j.sajb.2023.06.054

Miura, K., Toh, H., Hirakawa, H., Sugii, M., Murata, M., Nakai, K., Tashiro, K., Kuhara, S., Azuma, Y., & Shirai, M. (2008). Genome-wide analysis of Chlamydophila pneumoniae gene expression at the late stage of infection. DNA Research, 15(2), 93–102. https://doi.org/10.1093/dnares/dsn001

Kim, K. D., El Baidouri, M., Abernathy, B., Iwata-Otsubo, A., Chavarro, C., Gonzales, M., … Jackson, S. A. (2015). A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. Plant Physiology, 168(4), 1433–1447. https://doi.org/10.1104/pp.15.00408

Yang, Y., Wang, J., & Di, J. (2013). Comparative inference of duplicated genes produced by polyploidization in the soybean genome. International Journal of Genomics, 2013, Article 810403. https://doi.org/10.1155/2013/810403

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R. W., & Jackson, S. A. (2013). The fate of duplicated genes in a polyploid plant genome. The Plant Journal, 73(1), 143–153. https://doi.org/10.1111/tpj.12026

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., ... & Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. nature, 463(7278), 178-183.

Pedro DLF, Amorim TS, Varani A et al. An Atlas of Plant Transposable Elements [version 1; peer review: 2 approved]. F1000Research 2021, 10:1194 (https://doi.org/10.12688/f1000research.74524.1)

Du, J., Grant, D., Tian, Z., Nelson, R. T., Zhu, L., Shoemaker, R. C., & Ma, J. (2010). SoyTEdb: a comprehensive database of transposable elements in the soybean genome. Bmc Genomics, 11(1), 113.

Ma, J., Devos, K. M., & Bennetzen, J. L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome research, 14(5), 860-869.

Liu, Y., Yi, C., Fan, C., Liu, Q., Liu, S., Shen, L., ... & Han, F. (2023). Pan-centromere reveals widespread centromere repositioning of soybean genomes. Proceedings of the National Academy of Sciences, 120(42), e2310177120.

Fatima, A., Muhammad, G., Waheed, U., Ijaz, S., Khanum, P., & Khan, Z. (2025). Retrotransposons in Soybean (Glycine max). Plant Retrotransposons, 176-196.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant molecular biology*, *42*(1), 251-269.

Appendices

Code Availability

All the code written in this project, from pipelines to plots, can be found on this github link: https://github.com/raysas/comparative-genomics-project. Results tables, summaries and figures can also be found there, and are fully reproducible.

Pipelines Optimization

Processing over 2.4 million gene pairs required substantial workflow optimization to ensure both speed and reliability. We implemented parallelization throughout the pipelines, distributing BLAST searches, alignments, and Ks calculations across multiple CPU cores to dramatically reduce runtime. Batch processing was used to split large tasks into manageable chunks, allowing for efficient memory usage and easier error recovery. Checkpointing was integrated at key stages, so that intermediate results could be saved and the workflow could resume from the last successful step in case of interruption or failure. Additionally, all command-line steps were carefully optimized for performance, including the use of efficient file formats, streamlined data parsing, and robust error handling. These optimizations enabled us to process large genomic datasets reproducibly and efficiently, and to easily adapt the workflow for new species or parameter settings.

MCScanX

Duplicates Prediction & Methodology

To systematically classify duplicated genes, we used MCScanX (besides our own implementations), a widely adopted tool for detecting and categorizing gene duplications based on sequence similarity and chromosomal context. MCScanX first identifies homologous gene pairs through all-vs-all BLAST results, then scans the genome for collinear blocks—regions where multiple homologous genes are arranged in the same order and orientation. Based on the arrangement and proximity of these homologs, MCScanX assigns each gene pair to one of several duplication categories:

- (Tandem) duplicates: Adjacent homologous genes on the same chromosome, with no or very few intervening genes.

- (Proximal) duplicates: Homologous genes located on the same chromosome but separated by a small number of non-homologous genes.

- (Dispersed) duplicates: Homologous genes located on different chromosomes or far apart on the same chromosome, not fitting other categories.

- (Segmental/WGD) duplicates: Homologous gene pairs that are part of larger collinear blocks, typically resulting from whole-genome or large-scale segmental duplications.

This classification enables a comprehensive view of the duplication landscape, distinguishing between local (tandem/proximal) and large-scale (segmental/WGD) events, and providing a robust framework for downstream evolutionary and functional analyses. For our analyses, we validated MCScanX TAG predictions with our own spacer-based implementation (described above), and we validated WGD predictions with our identified Ks peak regions, i.e. ensured that MCScanX predicted WGD duplicates have Ks values that fall close to the identified Ks peaks.

Synteny Blocks

MCScanX also detects collinear (syntenic) blocks—conserved regions of gene order between different chromosomal segments—by chaining together homologous gene pairs that are co-linear and co-oriented. These blocks represent the remnants of ancient duplication events, such as WGDs, and are key to understanding genome evolution.

To visualize the extent and distribution of these collinear blocks, we used the MCScanX output to generate synteny plots (using SynVisio). The resulting synteny plot is highlighted in Fig 19. Its maps provide a genome-wide view of duplicated segments, revealing patterns of large-scale structural conservation and rearrangement. We matched it and found it in accordance with the equivalent synteny map provided by Roulin et al. (2013). These visualizations are instrumental in interpreting the evolutionary history of the soybean genome and in identifying regions of functional or evolutionary significance.

In addition, we examined the best-matching chromosome pairs associated with each WGD event by combining MCScanX collinear blocks with their corresponding Ks values produced by our pipeline. The results are depicted in Fig 20. This allowed us to assign duplicated chromosomal regions to specific duplication episodes and to distinguish between the Glycine-specific and early-legume WGDs. Consistent chromosome pairings and comparable block structures were observed across large portions of the genome, indicating extensive retention of WGD-derived synteny despite subsequent diploidization, gene loss, and rearrangements. This chromosome-level analysis helps localize regions that have remained evolutionarily conserved since polyploidization.

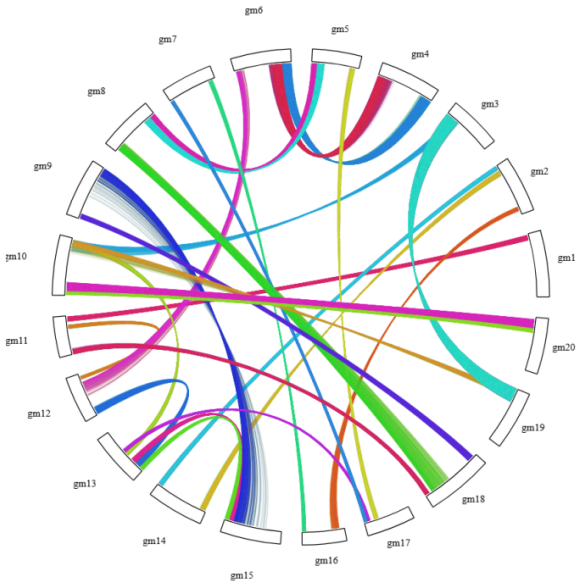


Figure 19. Synteny plot based on MCScanX predicted synteny blocks (collinearity file), produced by SynVisio.

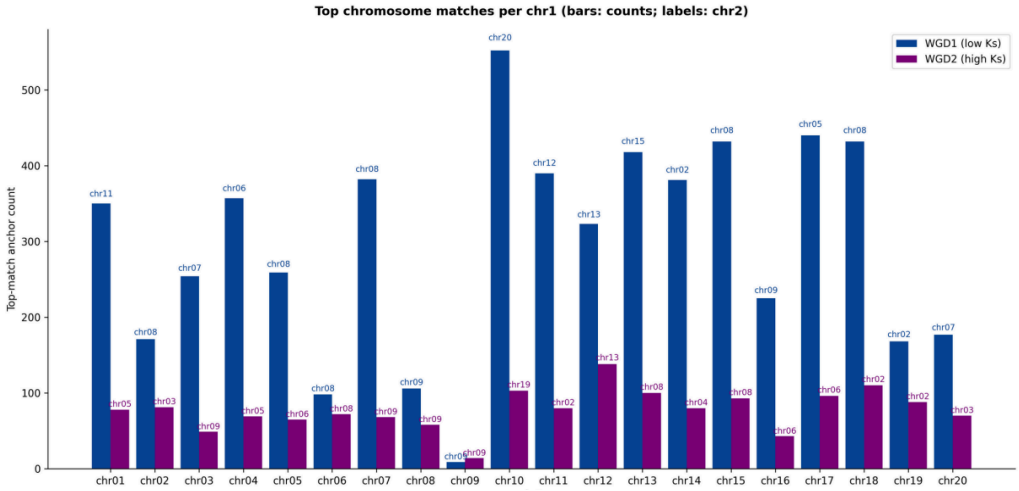


Figure 20. The best corresponding match (most overlaps) per chromosome, for each of the two WGD events.

Supplementary Figures

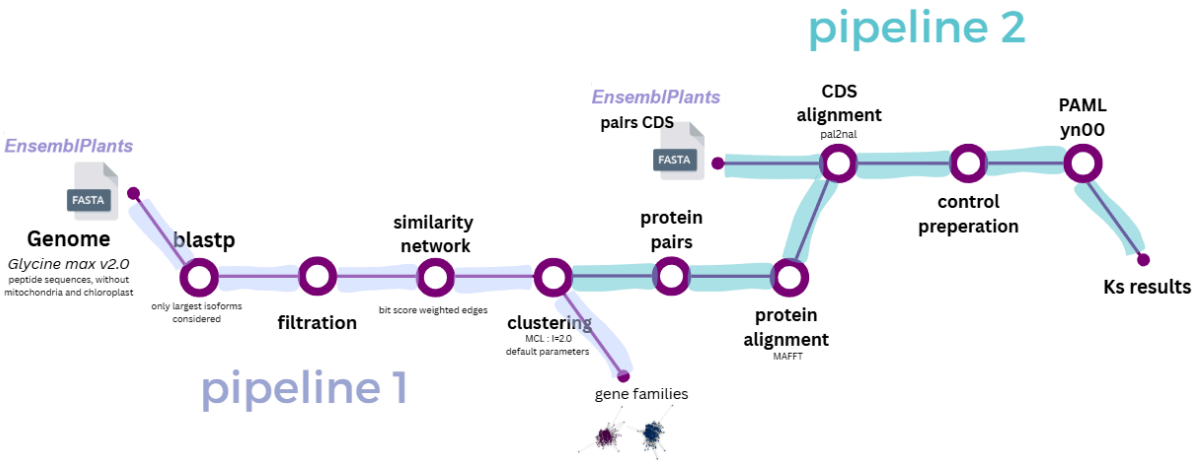


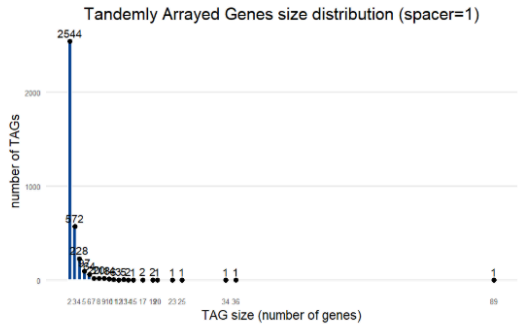
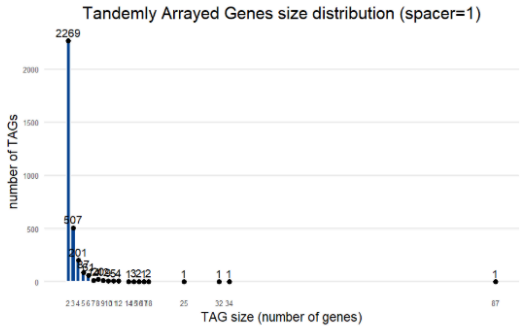
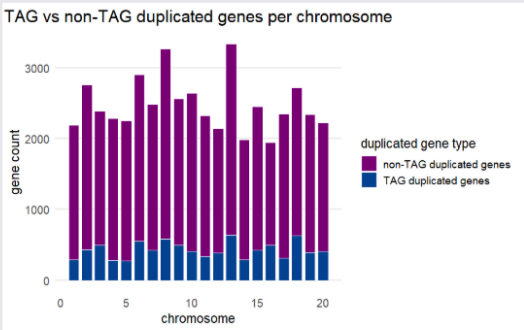
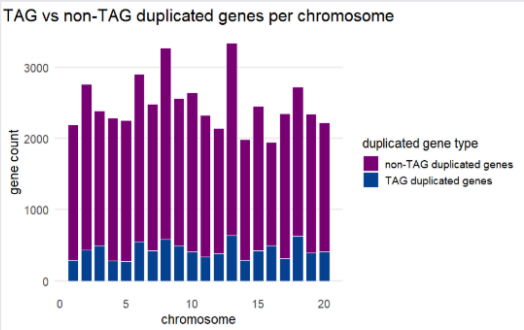
Figure S1. Diagram summarizing the pipeline built for duplicated genes protocol

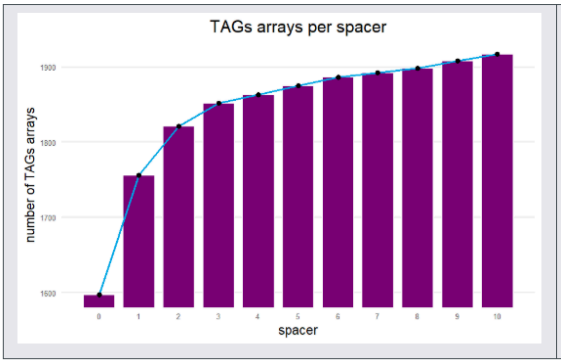
- low stringency dataset (`output/statistics/TAGs_spacers_ratios_low.tsv`)

spacer	n_TAG_genes	n_TAG_arrays	TAG_genes_percent
0	8563	1263	15.107%
1	9696	1381	17.106%
2	10164	1425	17.932%
3	10442	1447	18.423%
4	10607	1458	18.714%
5	10751	1468	18.968%
6	10894	1474	19.220%
7	11010	1480	19.425%
8	11151	1487	19.673%
9	11267	1495	19.878%
10	11370	1499	20.060%

- high stringency dataset (`output/statistics/TAGs_spacers_ratios_high.tsv`)

spacer	n_TAG_genes	n_TAG_arrays	TAG_genes_percent
0	7524	1597	13.274%
1	8545	1756	15.076%
2	8926	1821	15.748%
3	9122	1852	16.094%
4	9239	1863	16.300%
5	9321	1875	16.445%
6	9393	1886	16.572%
7	9447	1892	16.667%
8	9501	1898	16.762%
9	9551	1908	16.851%
10	9605	1917	16.946%

low	high
	
	



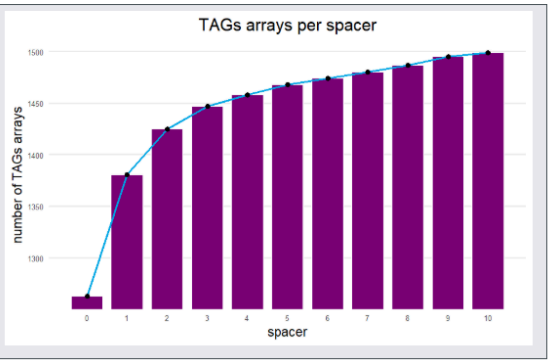


Figure S2. Tables and figures representing data on identified TAGs from duplicated genes of low and high datasets respectively, using spacers from 0 to 10 (section: TAGS identification)

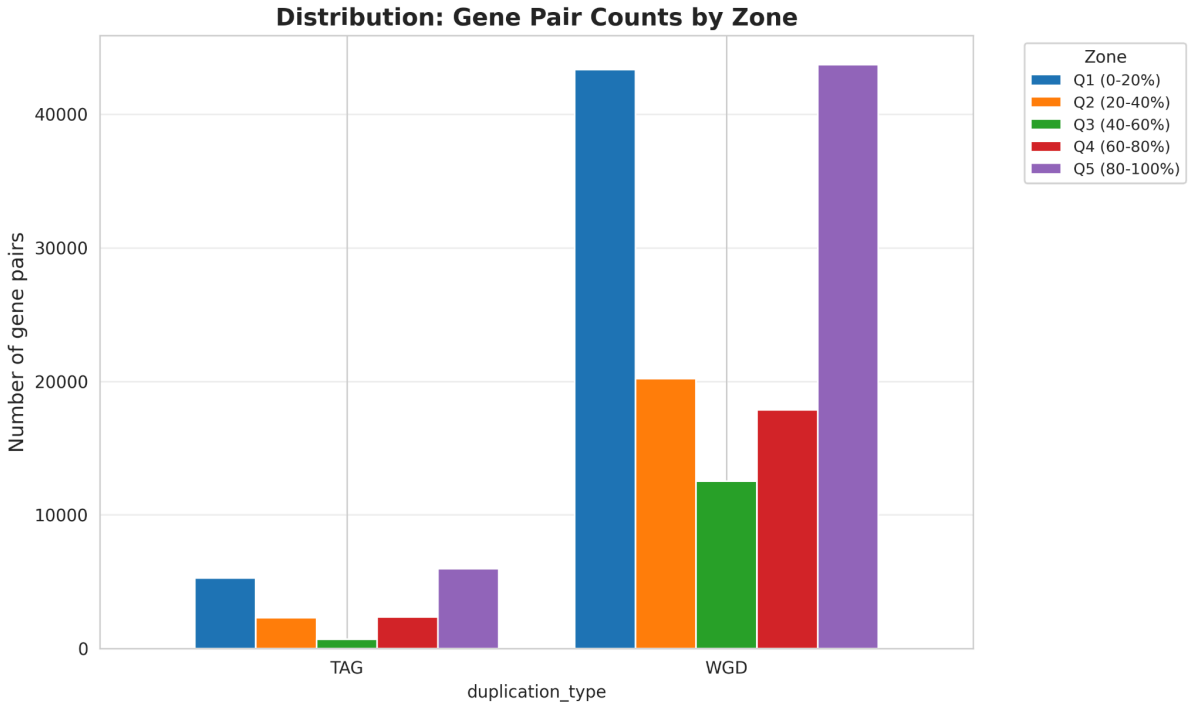


Figure S3. Distribution of gene pair counts by chromosomal zone (quintiles), for TAG and WGD pairs.

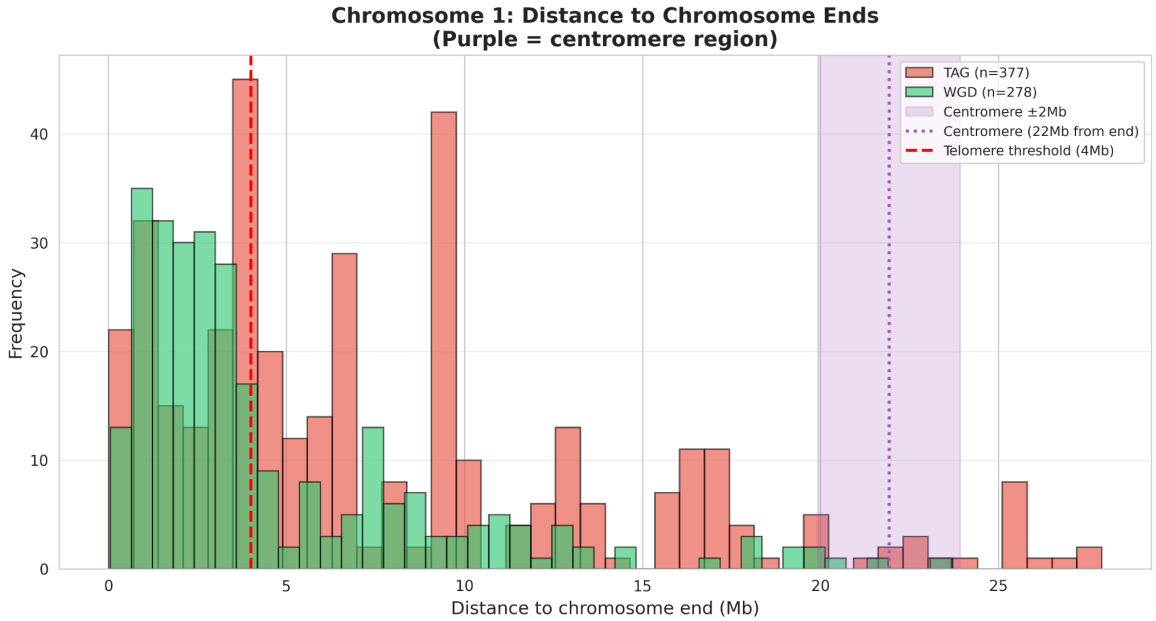


Figure S4. Per-chromosome histogram (example: Chromosome 1) of distances to chromosome ends for TAG and WGD pairs, with centromere region and telomere threshold highlighted.

Table S1. Functional enrichment by duplication type in *Glycine max*

Category	Group	GO Name	Gene Ratio	Gene Count	-Log10(FDR)
BP	Singletons	RNA modification	0.133	50	11.830
BP	Singletons	DNA damage response	0.129	54	12.207
BP	Singletons	DNA repair	0.129	52	11.914
BP	Singletons	DNA metabolic process	0.108	63	11.258
BP	Singletons	cellular response to stress	0.090	69	8.928
BP	Singletons	nucleic acid metabolic process	0.076	159	14.607
BP	Singletons	nucleobase-containing compound metabolic process	0.067	170	11.243
BP	Singletons	heterocycle metabolic process	0.066	180	11.373
BP	Singletons	organic cyclic compound metabolic	0.063	184	10.321

		process			
BP	Singletons	cellular aromatic compound metabolic process	0.063	176	9.386
BP	TAGs	glutathione metabolic process	0.655	57	28.243
BP	TAGs	defense response to other organism	0.506	137	50.277
BP	TAGs	response to biotic stimulus	0.506	137	50.180
BP	TAGs	response to other organism	0.506	137	50.056
BP	TAGs	response to external biotic stimulus	0.506	137	49.879
BP	TAGs	biological process involved in interspecies interaction between organisms	0.489	137	48.171
BP	TAGs	defense response	0.480	165	56.031
BP	TAGs	response to external stimulus	0.403	140	37.200
BP	TAGs	protein phosphorylation	0.251	247	25.914
BP	TAGs	phosphorylation	0.246	259	25.883
BP	Non TAGs	organic substance biosynthetic process	0.862	3019	157.745
BP	Non TAGs	biosynthetic process	0.856	3060	150.539
BP	Non TAGs	macromolecule metabolic process	0.838	4948	206.267
BP	Non TAGs	primary metabolic process	0.837	6280	267.103
BP	Non TAGs	nitrogen compound metabolic process	0.834	5590	225.351
BP	Non TAGs	organic substance metabolic process	0.832	6691	270.577
BP	Non TAGs	cellular process	0.827	8917	Inf
BP	Non TAGs	biological_process	0.825	12580	Inf
BP	Non TAGs	metabolic process	0.824	7030	262.498
BP	Non TAGs	cellular metabolic process	0.820	5468	185.625
MF	Singletons	DNA nuclease activity	0.200	14	4.928

MF	Singletons	catalytic activity, acting on DNA	0.155	39	10.936
MF	Singletons	catalytic activity, acting on a nucleic acid	0.093	66	8.658
MF	Singletons	molecular_function	0.029	470	7.366
MF	Singletons	transcription regulator activity	0.013	17	4.712
MF	Singletons	kinase activity	0.012	17	5.693
MF	Singletons	DNA-binding transcription factor activity	0.012	13	4.712
MF	Singletons	protein kinase activity	0.011	12	4.790
MF	Singletons	phosphotransferase activity, alcohol group as acceptor	0.010	14	6.296
MF	Singletons	protein serine/threonine kinase activity	0.007	6	5.936
MF	TAGs	glutathione transferase activity	0.726	61	34.264
MF	TAGs	UDP-glucosyltransferase activity	0.453	67	20.759
MF	TAGs	UDP-glycosyltransferase activity	0.385	74	18.209
MF	TAGs	carboxylic ester hydrolase activity	0.373	87	20.316
MF	TAGs	glucosyltransferase activity	0.367	81	18.356
MF	TAGs	oxidoreductase activity	0.228	368	30.406
MF	TAGs	protein kinase activity	0.224	244	18.827
MF	TAGs	catalytic activity	0.175	1583	54.431
MF	TAGs	transferase activity	0.175	672	20.277
MF	TAGs	molecular_function	0.146	2386	24.478
MF	Non TAGs	RNA binding	0.878	1446	85.848
MF	Non TAGs	protein binding	0.872	1342	74.699
MF	Non TAGs	transcription regulator activity	0.862	1119	56.102
MF	Non TAGs	organic cyclic compound binding	0.862	3375	176.907
MF	Non TAGs	nucleic acid binding	0.861	2941	150.959

MF	Non TAGs	binding	0.857	5248	272.583
MF	Non TAGs	DNA binding	0.841	1423	56.682
MF	Non TAGs	catalytic activity, acting on a protein	0.809	2274	62.019
MF	Non TAGs	molecular_function	0.803	13137	Inf
MF	Non TAGs	catalytic activity	0.773	6993	121.316

Table S2. Functional enrichment by gene family size in Glycine max

Category	**Family**	**GO Name**	**Gene Ratio**	**Gene Count**	**-Log10(FDR)**
BP	Small	RNA metabolic process	0.733	1112	141.767
BP	Small	nucleic acid metabolic process	0.683	1419	138.690
BP	Small	nucleobase-containing compound metabolic process	0.655	1666	137.959
BP	Small	cellular nitrogen compound metabolic process	0.651	1990	161.759
BP	Small	heterocycle metabolic process	0.640	1749	131.344
BP	Small	organic cyclic compound metabolic process	0.633	1835	130.907
BP	Small	cellular aromatic compound metabolic process	0.631	1772	124.089
BP	Small	organic substance biosynthetic process	0.614	2151	133.921
BP	Small	cellular biosynthetic process	0.612	2082	126.848
BP	Small	biosynthetic process	0.603	2155	122.272
BP	Large	protein phosphorylation	0.791	780	153.767
BP	Large	phosphorylation	0.775	817	151.474
BP	Large	phosphate-containing compound metabolic process	0.649	1240	127.910
BP	Large	phosphorus metabolic process	0.649	1255	129.167

BP	Large	protein modification process	0.602	1307	99.131
BP	Large	response to stimulus	0.582	1371	90.092
BP	Large	regulation of cellular process	0.567	1811	106.331
BP	Large	regulation of biological process	0.555	1847	97.943
BP	Large	biological regulation	0.546	2014	98.025
BP	Large	biological_process	0.470	7170	158.947
MF	Small	catalytic activity, acting on RNA	0.688	302	29.227
MF	Small	catalytic activity, acting on a nucleic acid	0.669	476	41.311
MF	Small	mRNA binding	0.661	453	37.305
MF	Small	RNA binding	0.659	1085	90.124
MF	Small	binding	0.495	3029	41.210
MF	Small	kinase activity	0.263	368	29.547
MF	Small	phosphotransferase activity, alcohol group as acceptor	0.240	321	38.034
MF	Small	protein kinase activity	0.194	211	50.939
MF	Small	protein serine/threonine kinase activity	0.167	150	53.747
MF	Small	RNA polymerase II cis-regulatory region sequence-specific DNA binding	0.163	79	29.772
MF	Large	protein serine/threonine kinase activity	0.808	726	153.333
MF	Large	protein kinase activity	0.781	850	161.777
MF	Large	phosphotransferase activity, alcohol group as acceptor	0.732	980	153.000
MF	Large	DNA-binding transcription factor activity	0.721	805	118.921
MF	Large	kinase activity	0.704	986	135.110
MF	Large	sequence-specific DNA binding	0.670	840	96.708
MF	Large	transcription regulator activity	0.665	863	96.478

MF	Large	catalytic activity, acting on a protein	0.578	1625	104.495
MF	Large	catalytic activity	0.522	4725	199.959
MF	Large	molecular_function	0.504	8250	Inf