

Genome Annotation of *Arabidopsis thaliana* Accession Altai-5

Anna Boss | 24-208-928

GitHub: <https://github.com/ann4boss/Annotation-of-Eukaryote-genomes>

1. Introduction

This report summarises the results obtained from the practical component of the course SBL.30004 Organization and Annotation of Eukaryote Genomes at the University of Fribourg (2025). The analysis focused on the *Arabidopsis thaliana* accession Altai-5.

The underlying genome assembly was generated in a previous course (473637 Genome and Transcriptome Assembly, University of Bern) using reads from the study by Lian et al. (2024)¹. The assembly, created using HiFiasm, resulted in a total length of 171,153,468 bp, composed of 958 contigs, with an NG50 of 15,087,710 bp. Altai-5 originates from the Chinese Altai Mountain range (Longitude: 88.400000, Latitude: 47.760000), an environment known for its high biodiversity.

2. Transposable Element Annotation and Classification

Transposable elements (TEs) were annotated using the EDTA pipeline² and subsequently classified into LTR clades using TEsorter³ and REXdb database⁴. The total TE burden in the Altai-5 genome was calculated to be approximately 13.3% of the entire assembly, corresponding to 22.8 Mb (Figure 1A). This relative abundance is within previously found ranges for *A. thaliana*, although it is lower than the 21% reported for the reference genome TAIR10⁵.

As depicted in Figure 1B, the most abundant TE class is LTR (Gypsy, Copia, and unknown), followed by non-TIR (helitron) and non-LTR (SINE) elements. The superfamily analysis showed that unidentified LTR elements (4.1%) were the most prevalent superfamily. They were likely not assigned to a more specific clade because of substantial sequence divergence. The Gypsy superfamily (2.7%) was the second most abundant, notably exceeding the abundance of the other major LTR superfamily, Copia (<1%). This dominance of Gypsy over Copia aligns with established

observations in *A. thaliana* TE evolution⁵. Helitron elements constitute the third most prevalent superfamily.

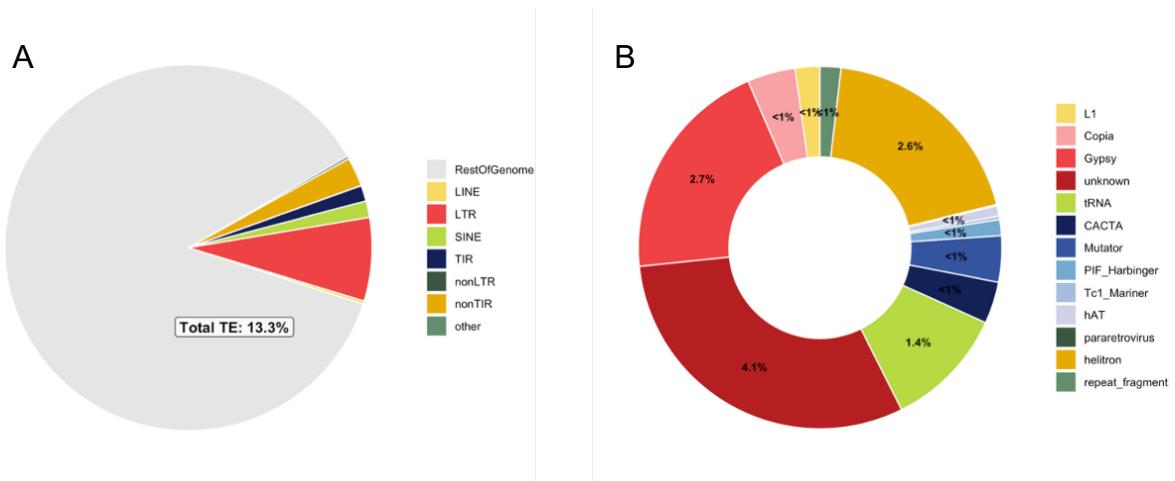


Figure 1. Transposable elements abundance

Transposable elements (TEs) were annotated using the EDTA pipeline² and classified using TEsorter³ and REXdb database⁴. (A) The total TE burden was 13.3% (22.8 Mb). (B) The figure presents the relative genomic abundance (%) for the main TE orders and their superfamilies (LTR, non-TIR, non-LTR, ect.). The TE burden is within the known range but lower than the TAIR10 reference (21%)⁵. The dominance of the LTR order is expected, and the higher prevalence of Gypsy (2.7%) over Copia (<1%) is consistent with known TE evolutionary patterns in *A. thaliana* populations⁵.

The evolutionary history and activity of TEs were assessed by estimating insertion age based on sequence divergence against consensus sequences. Since the raw divergence calculated by RepeatMasker is often inflated by hypermutation at CpG dinucleotides, the Perl script parseRM.pl⁶ was employed to yield a corrected percentage of divergence. As shown in Figure 2, Gypsy elements were active across the entire evolutionary timescale and contributed to genome size variation. They show peaks at roughly 3% and 10% divergence, which suggests periodic reactivation events. Copia elements also demonstrate sustained activity, but the contributing sequence length is substantially shorter compared to Gypsy. LINE elements are highly diverged (>20%), indicating this clade is largely extinct or silenced. Notably, the DNA/MULE-MuDR (Mutator-like elements) clade shows a resurgence in the recent periods, suggesting periodic reactivation events have occurred.

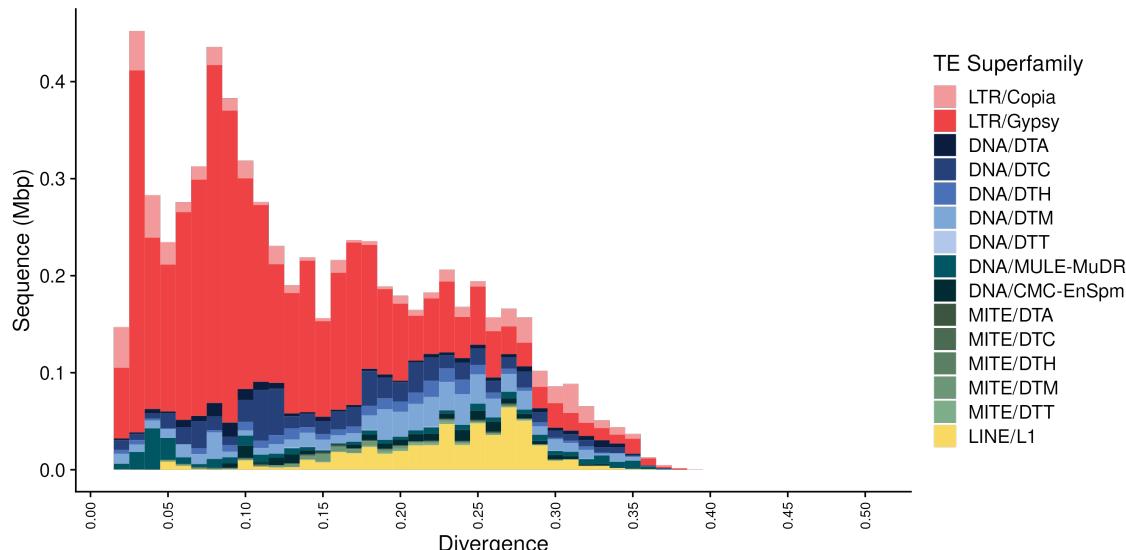


Figure 2. TE Insertion Age - Divergence Analysis

Insertion age was estimated from the raw divergence against consensus sequences, corrected for CpG hypermutation using parseRM.pl⁶. Lower divergence indicates recent activity. The plot displays the total sequence length (Y-axis) of each TE clade (e.g., Gypsy, Copia, LINE) binned by divergence percentage (X-axis). Gypsy elements are active throughout the entire history, peaking at 3% and 10% divergence. The LINE clade is highly diverged, suggesting extinction or complete silencing, while the DNA/MULE-MuDR clade shows a recent resurgence, indicating periodic re-activation.

Further analysis of Long Terminal Repeats (LTRs) retrotransposons used the identity percentage between the paired LTRs as a molecular clock to infer relative age. Figure 3 illustrates this identity analysis. Copia elements show a broad identity range, roughly 0.88 to 1.00. Gypsy elements show a similar range, roughly 0.89 to 1.00. Both clades include highly conserved copies, but Copia displays more low identity entries, which suggests greater sequence divergence or more ancient insertions in the genome. This matches the common observation that Gypsy clades often form younger, more lineage specific expansions, while Copia families tend to accumulate more divergence⁵. Athila (n=51 families), the dominant and most ancient Gypsy clade, has activity spread across the entire timeline, confirming its long-term contribution to the TE load. This clade also shows a strong peak in the 0.98 to 1.00 identity range, confirming its participation in a recent amplification event contributing to genome size expansion. This identity analysis reflects the findings seen in Figure 2.

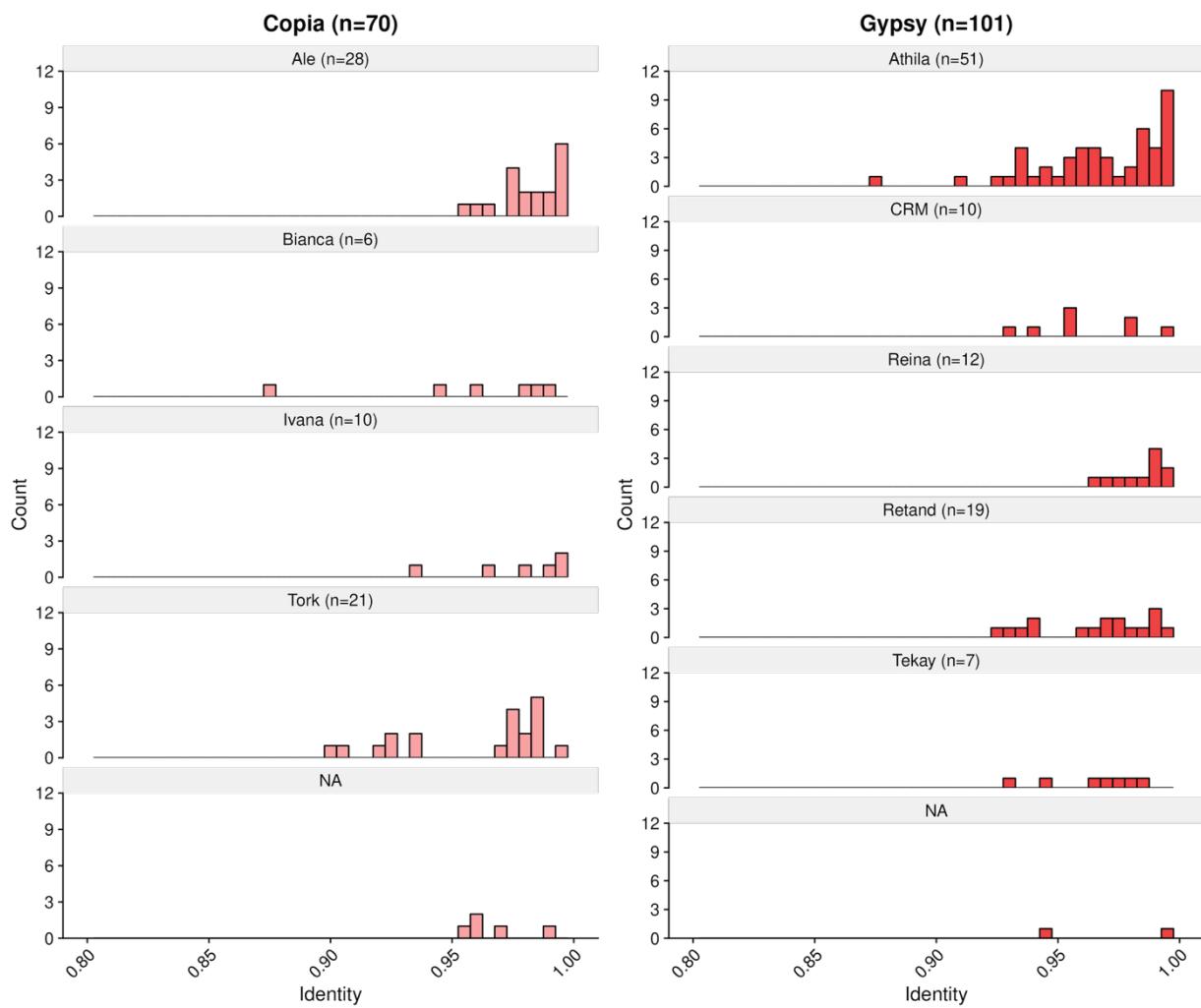


Figure 3. LTR Identity Analysis

Long Terminal Repeats (LTRs) retrotransposon activity was assessed using the identity percentage between the paired LTRs, which functions as a molecular clock. The plot compares the distribution of elements (count on the Y-axis) across identity bins (X-axis) for major LTR clades (Gypsy and Copia). Identity close to 1.00 indicates a very recent insertion event, while lower identity represents older, more diverged elements. Copia elements exhibit a wide distribution across identity bins (0.88 to 1.00), suggesting great overall sequence divergence. The Gypsy clade Athila demonstrates activity across the full timeline and a strong peak in the high-identity range (0.98–1.00), confirming its recent involvement in genome expansion, consistent with the TE age analysis shown in Figure 2.

3. Gene Annotation and Quality Control

Gene prediction was performed using the MAKER pipeline⁷, integrating RNA-seq data from the accession Sha assembled with Trinity and protein evidence using TAIR10 (TAIR10_pep_20110103_representative_gene_model)⁸. The filtered model resulted in 30,168 genes (Table 1), which is a reasonable amount compared to the reference number of 27,657 genes in TAIR10, suggesting a high degree of completeness. The

annotation appears comprehensive, with most genes functionally annotated, reasonable exon/intron statistics, and a typical proportion of monoexonic genes. Some extreme values (very short genes or extremely high exon counts) should be interpreted with caution, as they may represent artifacts, pseudogenes, or assembly errors. Additional filtering such as very short genes, might increase the quality of the annotation.

Table 1. Summary of Gene Annotation Metrics (Altai-5).

Metric	Altai-5
Number of Genes	30,168
Number of mRNA	30,168
Gene Length Mean [min, max]	2,339 [9, 199,806]
mRNA Length Mean [min, max]	2,339 [9, 199,806]
Exon Length Mean [min, max]	257 [0, 7,761]
Intron Length in CDS Mean [min, max]	171 [22, 5,603]
Monoexonic Genes	6,394 (21.19%)
Exons per Gene Mean [min, max]	5.8 [1, 7761]
Functionally Annotated Genes	25,457 (84.38%)

Key structural metrics of the Altai-5 annotation were generated and compared against the published TAIR10 reference genome. The table shows gene count, mean feature lengths, and the percentage of monoexonic genes for Altai-5. The gene count (30,168) is comparable to the TAIR10 reference (approx. 27,600), suggesting the annotation process achieved a high level of gene identification completeness. Structural metrics are consistent with typical *A. thaliana* gene architecture.

The overall quality of the annotation was assessed through the Annotation Edit Distance (AED) score. A total of 96.3% of the gene models achieved an AED score of 0.5 or lower. This indicates strong agreement between the computational predictions and the supporting biological evidence, suggesting high confidence in the predicted gene structures.

Further validation of gene content completeness was performed using BUSCO⁹ (Benchmarking Universal Single-Copy Orthologs). The analysis across three levels showed a trend of decreasing completeness (Table 2): the raw assembly exhibited near-complete detection (99.2%), confirming the underlying genome sequence is highly contiguous and robust. Transcript-level completeness was slightly lower (94.7%), reflecting minor limitations in transcript reconstruction or issues related to alternative splicing. Crucially, protein-level completeness stood at 93.3% complete, resulting in the highest rate of missing BUSCOs (5.8%). This difference confirms that

while the target genes are present in the genome, they were either missed or inaccurately delineated during the gene prediction and filtering steps of the annotation pipeline. Overall, the comprehensive BUSCO analysis demonstrated the high quality and reliability of both the genome assembly and the resultant annotation.

Table 2. Annotation Quality Control - BUSCO

Metric	Assembly	Transcripts	Proteins
Complete (C)	99.2%	94.7%	93.3%
Single-copy (S)	97.7%	90.8%	91.6%
Duplicated (D)	1.5%	3.9%	1.7%
Fragmented (F)	0.1%	0.8%	0.9%
Missing (M)	0.7%	4.5%	5.8%

Annotation quality was measured using BUSCO (Benchmarking Universal Single-Copy Orthologs)⁹. Each row shows the proportion of Benchmarking Universal Single-Copy Orthologs (BUSCOs) detected in the assembly, transcriptome, or protein predictions. “Complete (C)” indicates the fraction of expected orthologs fully present; “Single-copy (S)” represents orthologs present in one copy, while “Duplicated (D)” counts those found in multiple copies. “Fragmented (F)” refers to orthologs only partially recovered, and “Missing (M)” indicates orthologs not detected. High percentages of complete and single-copy BUSCOs reflect a high-quality, largely complete assembly and annotation, whereas low fragmented or missing scores indicate minimal gaps or errors. The high assembly completeness (99.2%) validates the genome sequence integrity. The slightly lower protein completeness (93.3%) indicates that a small percentage of core genes were missed or structurally mis-modeled during the annotation pipeline, despite being present in the underlying DNA.

4. Genome Structure and Functional Annotation

4.1. Genome Structure Analysis

A Circos plot using circlize R package¹⁰ was generated to visualize the genomic distribution of TE insertion and gene density across the assembled chromosomes. Figure 4 reveals a clear preferential insertion pattern for Gypsy elements in low gene density regions, which strongly corresponds to the inferred pericentromeric regions. These regions are under low selection pressure, thereby allowing TE accumulation. Copia and LINE elements also show increased density in these gene-sparse areas. However, they display a more dispersed insertion profile, with noticeable insertions within gene-dense regions. This broader distribution for Copia and LINE elements is likely attributed to a lack of insertion preferences, unlike Gypsy elements that are often restricted to heterochromatin sites.

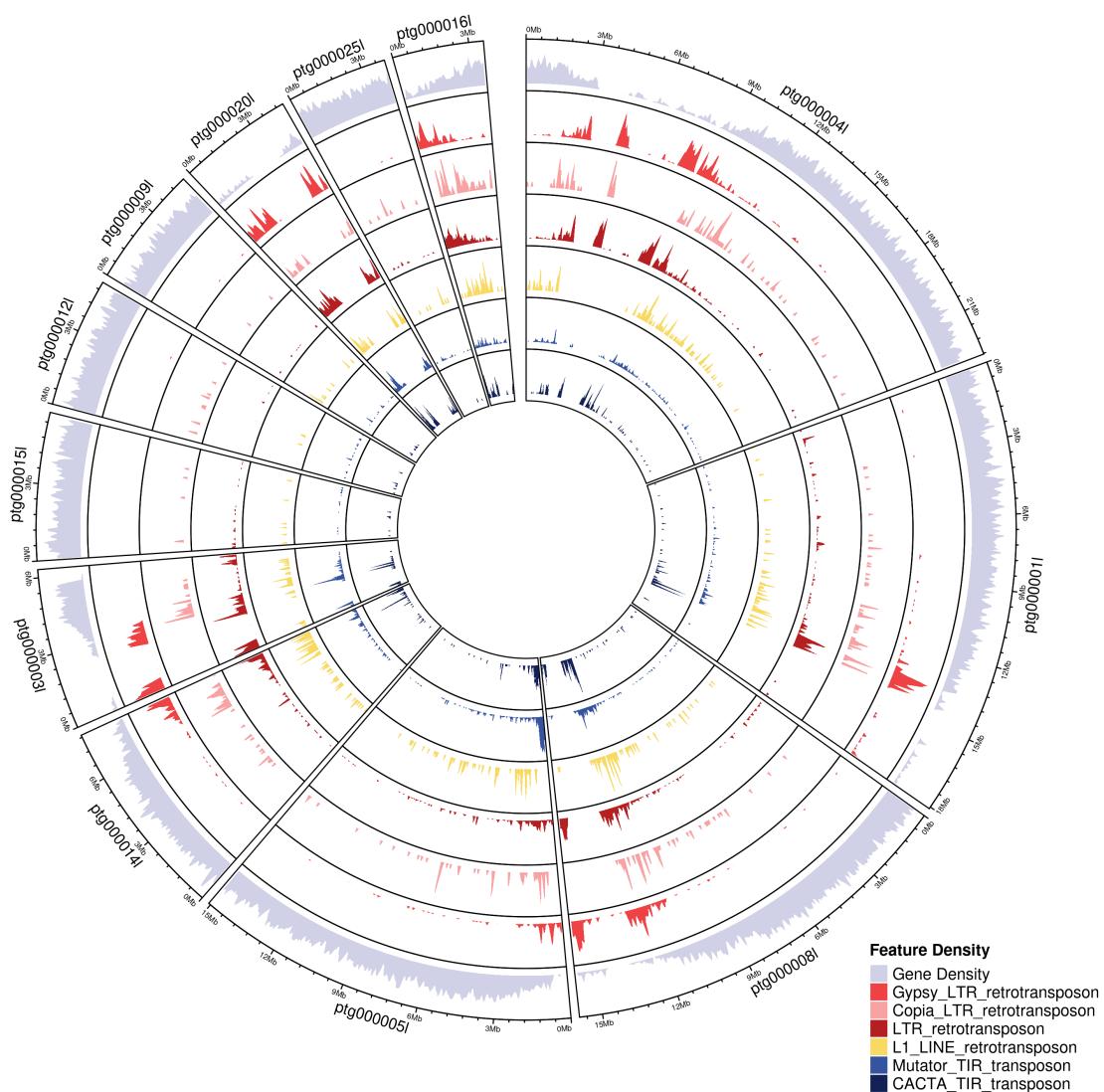


Figure 4: Circos Plot of TE Insertion and Gene Density

A Circos plot visualizes the genomic distribution of gene density and the insertion locations of the major Transposable Element superfamilies (Gypsy, Copia, LINE). The outermost ring displays contigs; inner rings show gene density and TE Superfamily Density. Gypsy elements predominantly insert into low gene density regions (inferred pericentromeric areas). Copia and LINE elements, while also concentrated in gene-sparse regions, exhibit more insertions within gene-dense areas, reflecting different target-site preferences or surviving chances and potentially lower levels of silencing.

4.2. Functional Annotation

Predicted protein sequences were subjected to homology searches using BLAST¹¹ (UniProt and TAIR10) to assess functional content. As shown in Table 3, comparison against the species-specific TAIR10 proteome yielded an excellent match rate (84.76% Well-Annotated and only 0.58% No Homology), confirming the high quality

and conservation of the predicted gene set relative to the reference species. When compared to the broader UniProt database, the majority (76.47%) were confirmed as well-characterized.

The rate of "No Homology" proteins was higher against UniProt (18.45%) compared to TAIR10 (0.58%). These 5,567 proteins without global homology were generally shorter than the functional genes (Figure 5). This suggests that they are likely not true novel accession-specific genes but rather fragmented or incomplete gene models that failed to contain sufficient conserved domain information to yield a significant BLAST hit, thus indicating an area for refinement in the gene structural prediction rather than a high number of novel genes. Based on these numbers, the annotation is high in terms of completeness (homology coverage) but requires further refinement to increase functional specificity (depth).

Table 3. Homology Analysis (TAIR10 vs. UniProt).

Classification	UniProt	TAIR10
Well-Annotated	23,069 / 76.47%	25,570 / 84.76%
Uncharacterized	1,532 / 5.07%	4,423 / 14.66%
No Homology (Novel/Error)	5,567 / 18.45%	175 / 0.58%

Predicted protein sequences were queried against the species-specific TAIR10 proteome and the global UniProt non-redundant database using BLAST. The table categorizes the predicted gene models based on the level of homology found (Well-Annotated, Uncharacterized, No Homology). The near-zero rate of "No Homology" for TAIR10 confirms high quality and conservation of the predictions. The higher rate of "No Homology" against UniProt (18.45%) suggests these are primarily fragmented gene models rather than true novel genes, as they are generally shorter than the functional set.

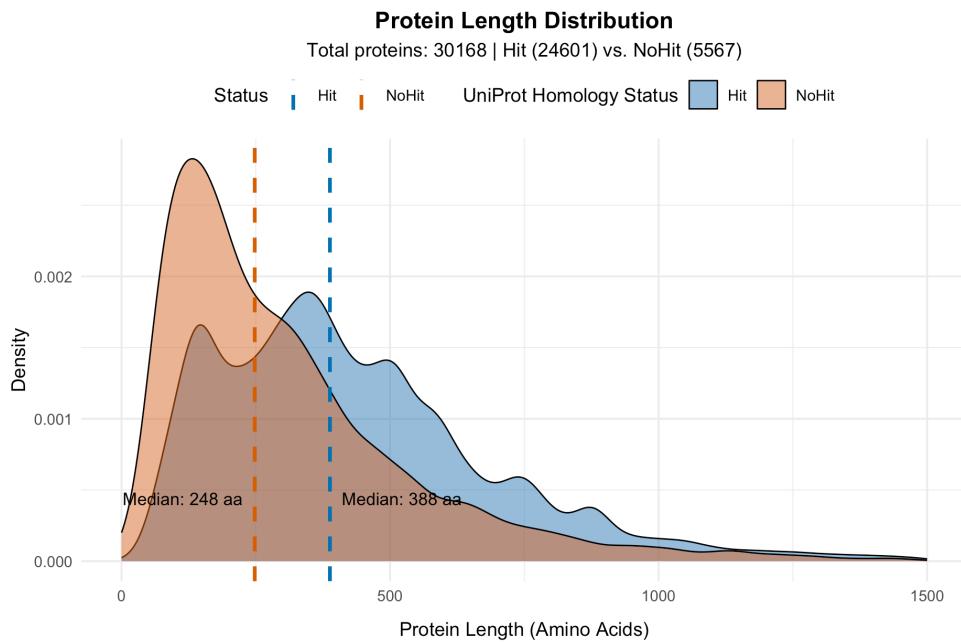


Figure 5: Protein Length Distribution

This density plot compares the length distribution of the 30,168 predicted proteins after classification via homology search against the UniProt database. Proteins were categorized as 'Hit' (24,601) if a significant homology was found, or 'NoHit' (5,567) if they returned no significant match. The X-axis represents the length of the predicted protein (in Amino Acids, aa), and the Y-axis represents the density (frequency) of proteins at that length. The dashed lines indicate the median length for each group. The 'NoHit' protein distribution (orange) is heavily skewed towards shorter lengths (Median: 248 aa) compared to the 'Hit' distribution (blue, Median: 388 aa). This significant difference suggests that the proteins lacking homology are likely incomplete or fragmented gene models, supporting the conclusion that the high 'NoHomology' count is due to structural prediction errors rather than a large population of novel accession-specific genes.

5. Comparative Genomics

The Altai-5 accession was compared to the reference TAIR10 and three additional accessions (Ice-1, Est-0, and Etna-2) from the Lian et al. (2024)¹ study using GENESPACE¹² to identify orthogroups and structural variations (SVs). The analysis (Table 4) identified a large core gene set (85.93%) which is consistent with previous comparative studies in *A. thaliana* populations¹. While 6.11% of genes are categorized as species-specific, this figure is likely inflated by the gene fragments and prediction errors identified in the functional annotation analysis, see Table 3 and Figure 5.

The GENESPACE structural comparison revealed that the Altai-5 accession's gene order and structure are highly conserved compared to the TAIR10 reference genome (only 7 SVs) and nearly identical in structure to the Ice-1 accession (0 SVs). The greatest structural divergence was observed against the Etna-2 accession (11 SVs). This gradient of structural similarity supports a closer evolutionary relationship

between Altai-5 and Ice-1, which is consistent with the phylogenetic tree established in the Lian et al.¹ study.

Table 4. GENESPACE Summary

Category	Orthogroups	Genes	Percent of total genes
Core	21,417	21,476	85.93%
Accessory	6,032	1,989	7.96%
Species-specific	5,911	1,527	6.11%

GENESPACE was used to compare the Altai-5 predicted gene set against four other *A. thaliana* accessions (TAIR10, Ice-1, Est-0, and Etna-2) to categorize genes into Orthogroups. The table displays the count of Orthogroups, the total gene count, and the percentage of the total gene set categorized as Core, Accessory, or Species-specific. The large Core gene set (85.93%) is expected for an intraspecies comparison.

An Altai-5 self-comparison of the GENESPACE output confirmed the evolutionary history of the *A. thaliana* genome. The Hits-to-Anchors ratio was 1.75 (47,609 hits / 27,128 anchors), significantly exceeding the expected 1.0 ratio for a diploid organism. This high value confirms the large-scale retention of paralogous gene copies resulting from the ancient whole genome duplication events known in the Brassicaceae lineage. The comparison also identified 721 highly fragmented syntenic blocks, which is consistent with the mapping of numerous duplicated regions that have been rearranged over evolutionary time. Finally, the detection of zero SVs confirmed the high internal consistency and non-chimeric quality of the Altai-5 assembly.

6. References

1. Lian Q, Huettel B, Walkemeier B, et al. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Nat Genet.* 2024;56(5):982-991. doi:10.1038/s41588-024-01715-9
2. Ou S, Su W, Liao Y, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 2019;20(1):275. doi:10.1186/s13059-019-1905-y
3. Zhang RG, Li GY, Wang XL, et al. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res.* 2022;9:uhac017. doi:10.1093/hr/uhac017
4. Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob DNA.* 2019;10(1):1. doi:10.1186/s13100-018-0144-1
5. Quesneville H. Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. *Mob DNA.* 2020;11(1):28. doi:10.1186/s13100-020-00223-x
6. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci U S A.* 2017;114(8):E1460-E1469. doi:10.1073/pnas.1616702114
7. Campbell MS, Law M, Holt C, et al. MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. *Plant Physiol.* 2014;164(2):513-524. doi:10.1104/pp.113.230144
8. Reiser L, Bakker E, Subramaniam S, et al. The *Arabidopsis* Information Resource in 2024. *Genetics.* 2024;227(1):iyae027. doi:10.1093/genetics/iyae027
9. Tegenfeldt F, Kuznetsov D, Manni M, Berkeley M, Zdobnov EM, Kriventseva EV. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res.* 2025;53(D1):D516-D522. doi:10.1093/nar/gkae987
10. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. *Bioinformatics.* 2014;30(19):2811-2812. doi:10.1093/bioinformatics/btu393
11. Basic Local Alignment Search Tool (BLAST) | Learn Science at Scitable. Accessed November 24, 2025. <https://www.nature.com/scitable/topicpage/basic-local-alignment-search-tool-blast-29096/>
12. Lovell JT, Sreedasyam A, Schranz ME, et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. Weigel D, ed. *eLife.* 2022;11:e78526. doi:10.7554/eLife.78526