Assembly and Annotation of Arabidopsis thaliana Accession Lu-1

MAALOUF Andrew

University of Bern, MSc in Bioinformatics and Computational Biology, 23-117-583

KEYWORDS

Arabidopsis Thaliana
PacBio HIFI Reads
Illumina Short Reads
Assembly
Quality Evaluation
Annotation
Genome
Proteome
Transcriptome
Transposable Elements

ABSTRACT

Arabidopsis thaliana, a widely used model organism, offers insights into genome evolution and structural variation. This report investigates the genomic features of the Lu-1 accession, a dataset excluded from the recent study by Lian et al. (2024) due to heterozygosity concerns. Using PacBio HiFi sequencing, the genome was assembled with three tools: LJA, Flye, and Hifiasm, selecting Hifiasm for downstream analyses based on different metrics and its ability to produce a non-redundant, streamlined assembly for annotation. Transposable elements (TEs), constituting 12.86% of the Lu-1 genome, were annotated with EDTA. LTR retrotransposons, particularly the Gypsy and Copia superfamilies, dominated, with Gypsy elements comprising 2.07% of the genome. By estimating the insertion times, we found that most of the TEs across the genome expanded within the last 3 or 17 million years: this indicates either an ancient insertion from a distant ancestor or a highly conserved TE preserved through evolutionary time. The annotation pipeline using MAKER resulted in the identification of 38,431 protein-coding genes which decreased to 35,572 after filtering for high-confidence annotations (AED $\leq$ 0.5), which is higher than the 27,416 genes annotated in the reference sequence. Comparative analyses with GENESPACE identified 19,744 orthogroups shared across some accessions, while structural variations between different accessions and TAIR10 were visualized through synteny and riparian plots. Despite the limitations posed by heterozygosity in Lu-1, this report has assembled and annotated the Lu-1 genome, leveraging the strengths of Hifiasm to preserve genetic diversity and heterozygosity. Several additional enhancements to improve the current analysis, especially the structural contiguity and incomplete annotation, are presented in the discussion.

INTRODUCTION

Arabidopsis thaliana, a part of the brassicales family, has played a role in basic biological discoveries[1,2] and continues to be adopted as a model for plant research[3]. It is mainly self-fertilizing, with an out-crossing rate of about 3%[4]. Its diploid nature and small genome (~135 Mb) make A. thaliana ideal for sequencing, genetic research, and computational analysis[5,6].

The difference in environmental conditions in regions such as Europe, Africa, North America, and Japan has led to variants of A. thaliana known as ecotypes, represented through different accessions[7,8]. Evolution is driven by mutation and selection, with random genetic drift contributing to diversification. Mutations can range from small point changes in nucleotides to larger alterations like insertions or deletions, known as structural variants. These structural variations are often linked to the repetitive nature and functional roles of transposable elements (TEs) which are abundant in many eukaryotic genomes, with nearly half of the human genome[9] and about 85% of the genome of wheat consisting of TEs[10]. These elements play a key role in genome evolution during adaptation, such as influencing recombination hotspots[11], or plant architecture[12]. In response to ongoing environmental pressures, A. thaliana may

increase genetic recombination, generating more diversity to better adapt and survive, much like the idea behind the Red Queen hypothesis, where species must keep evolving just to stay competitive.

TEs mainly consist of 2 major classes: Class I elements, known as retrotransposons, follow a "copy and paste" mechanism for their transposition[13], which directly contributes to an increased genome size. In fact, active movements of retrotransposons have been considered the main driver of gene obesity[14]. Retrotransposons are further divided into long terminal repeats (LTR) retrotransposons, as well as non-LTRs depending on structural features such as the target site duplication[15]. Moreover, Copia and Gypsy are 2 main superfamilies of LTR retrotransposons, differing in their terminal sequences[16]. On the other hand, Class II elements, known as DNA transposons, follow a "cut and paste" mechanism for their transposition[17].

TEs, while abundant and impactful, are often underexplored outside of model organisms. Their high copy numbers and complex nesting patterns create significant challenges for genome assembly. Although short-read sequencing efficiently assembles gene-rich regions, repetitive regions like TEs remain fragmented, and incorrect mapping of reads will result in a loss of genetic information. In contrast, long-read technologies such as PacBio and scaffolding methods like Hi-C have markedly improved the assembly of these regions, enabling thus better annotation and insights into genomic features. In fact, long-read sequencing was able to properly identify novel genetic information and structural variation that was missed when using short-read sequencing[18].

Consequently, recent research has utilized long-read sequencing technologies to study the structural variation and genetic diversity of Arabidopsis thaliana. Multiple studies have focused on resequencing diverse accessions to better understand the natural variability within the species[19].

In this course, each student analyzed one of 69 accessions of A. thaliana that had been re-sequenced using long-read sequencing, particularly using PacBio HiFi technology[19]. The workflow involves genome assembly, selecting the best assembler based on performance metrics, followed by annotation and quality assessment. My assigned accession, Lu-1, was excluded from the publication by Lian et al. (2024) due to indications of heterozygosity. As a result, while my analysis followed the same rigorous methodology, the significance of my findings was limited compared to other accessions in the dataset.

METHODS

The TAIR10 genome served as reference for Arabidopsis thaliana. Furthermore, transcriptome data was used to provide insights into transcript structures and isoform diversity. RNA-Seq reads from the Sha accession[20] were assembled and then used to validate gene annotations and improve functional predictions.

Reads and Quality Check

The raw data includes whole-genome PacBio HiFi reads for accession Lu-1[19] and whole-transcriptome Illumina RNA-seq reads for accession Sha[20]. Read quality was assessed using FastQC (v0.12.1)[21]. Fastp (v0.23.2)[22] was then applied to the paired-end Illumina RNA-seq reads with default settings, removing reads with a quality score below 15. For the PacBio HiFi reads, Fastp was run without filtering to calculate the total number of bases. The estimated depth of coverage was calculated using the following formula:

$$Depth\ of\ Coverage = \frac{mean\ read\ length\ \times number\ of\ reads}{total\ genome\ size\ (bp)}$$

The genome size of Lu-1 was estimated using k-mer counting with Jellyfish (v2.3.0)[23] and the percentage of heterozygosity was analyzed using

GenomeScope (v2.0)[24], both with default settings. Canonical k-mers were used to reduce redundancy, enhancing computational efficiency during genome size and heterozygosity estimation.

Genome Assembly

For the Lu-1 accession, the initial de novo assembly was performed with three different assembly tools: LJA (v0.2)[25], Flye (v2.9.2)[26], and Hifiasm[27]. For the latter, the output *bp.p_ctg* was selected for further assembly and downstream analysis.

Transcriptome Assembly

For transcriptome assembly of the Sha accession, Trinity (v2.15.1)[28] was used to reconstruct transcripts from short-read RNA-Seq data.

Assembly Evaluation

The quality of the genome and transcriptome assemblies was assessed using multiple tools. BUSCO (Benchmarking Universal Single-Copy Orthologs) (v5.7.1)[29] was run using the lineage dataset *brassicales_odb10* and the correct mode to investigate the presence of essential genes in the assemblies. QUAST (v5.2.0)[30] was employed with the Arabidopsis thaliana TAIR10 reference and its genomic features to evaluate the eukaryote genome assemblies by analyzing key metrics. Merqury (v1.3)[31] compared k-mers of length 19 derived from the assemblies against those from unassembled PacBio Hifi reads. This step provided insights into the quality (QV), error rate, and completeness of the assemblies. Moreover, Nucmer (mummer4_gnuplot.sif)[32] was used to align each of the genome assemblies against the A. thaliana reference genome. Dotplots were then generated with Mummerplot[33] to visualize structural differences and similarities. Additionally, basic assembly statistics were found using gfastats (gfastats_1.3.7.sif)[34].

For the following section, the Hifiasm assembly was used. Reasons for that are included in the discussion.

Transposable Element Annotation Using EDTA

To structurally annotate transposable elements (TEs) in the A. thaliana accession Lu-1 genome, the Extensive De-Novo TE Annotator (EDTA) tool[35] was used while ensuring all steps of TE annotation and providing the Arabidopsis thaliana coding sequence file *'TAIR10_cds_20110103_representative_gene_model_updated'* to avoid misclassification of gene sequences as TEs. To analyze intact LTR-RTs dynamics, the percent identity values were extracted, and the LTR-RT sequences were classified into clades using a homology-based tool TEsorter (v1.3)[36] and the rexdb-plant database.

Visualizing and Comparing TE Annotations from EDTA

The TE annotations summary file, which provides details on the base pair and genome percentage occupied by each TE superfamily, was examined. The most abundant TE superfamily was identified, and the percentage of TE content was compared against other accessions and other plant genomes. Moreover, the distribution of TEs was visualized using the circlize package[37] in R (v4.1.0) while selecting the top 10 longest scaffolds as pseudo-chromosomes for visualization. The scaffold lengths were obtained using samtools (v1.13)[39] faidx command on the assembly FASTA file. The distribution of some superfamilies such as Gypsy and Copia retrotransposons as well as some clades was plotted.

Refining TE Classification with TEsorter

Focusing on Class I LTR-RTs, TEsorter (v1.3) was employed to refine the classification on clade-level, including the Gypsy and Copia superfamilies. The non-redundant TE library FASTA file was used to extract sequences belonging to the Gypsy and Copia superfamilies separately. After correlating the different TE clades with their genome-wide abundance, the results of TE clades were compared against other accessions.

TE Age Estimation

The insertion age of TEs was estimated by measuring the sequence divergence of individual TEs from their consensus sequences. The RepeatMasker alignment output was parsed using the *parseRM.pl* script from BioPerl (v1.7.8)[40] to calculate the corrected percentage of divergence for each TE sequence. This step takes into consideration the high mutation rates at CpG sites. Next, the divergence of TEs was plotted. The insertion time was calculated using the formula $T = \frac{K}{2r}$ where K is the sequence divergence and $r = 8.22 \times 10^{-9}$ is the synonymous substitution rate per site per year[41].

Phylogenetic analysis of TEs

To construct relationships among TEs and identify families that proliferated from common ancestral sequences, protein-coding sequences for reverse transcriptase (RT) domains from the rexdb-plant were used to perform phylogenetic analysis of the Copia and Gypsy superfamilies, which differ in the order of their protein-coding domains. Additionally, TEsorter was run with the Brassicaceae TE database using the rexdb-plant database to refine the classification of the RT sequences. Sequences were then aligned separately for each superfamily using Clustal Omega (v1.2.4)[42] to generate protein alignment in FASTA format, a file needed to infer phylogenetic tree using FastTree (v2.1.11)[43]. The resulting trees were visualized using iTOL[44].

Homology-Based Genome Annotation with MAKER

To obtain high-quality genome annotation, the MAKER[45] pipeline was used while combining 3 methods: ab initio predictions (using Augustus with the pre-trained A. thaliana model), RNA-Seq evidence (using the transcriptome assembly of the Sha accession), and protein homology (using the A. thaliana genome and the UniProt database). Additionally, protein sequences were annotated using InterProScan (v5.67-99.0)[46] with the Pfam database to identify functional domains and assign Gene Ontology terms. The quality of the gene models was assessed using Annotation Edit Distance (AED). High-confidence gene models were filtered by keeping annotations with AED score less or equal to 0.5. Filtered gene models were processed to extract the corresponding protein and transcript sequences and create the final annotation set for downstream analyses.

Quality Assessment of Gene Annotations

The quality and completeness of annotations were evaluated with BUSCO (v5.7.1) using the *brassicales_odb10* database with default parameters, run separately on the protein and transcript sequences generated by MAKER. For that, the longest isoforms were extracted as the representative gene sequence.

To estimate protein-coding set completeness, homology was tested by aligning the protein sequences against the UniProt Viridiplantae Reviewed Database of functionally validated proteins using blastp (v2.15.0)[47] with a significance threshold set at 1e-10.

Orthology-Based Gene Annotation Quality Check Using OMArk

The quality of protein-coding gene sets was further evaluated using OMArk (v0.3.0)[48] which leverages the OMA orthology datasets to identify Hierarchical Orthologous Groups (HOGs). Additionally, a way to improve the annotation was explored using fragmented or missing gene models to retrieve sequences for conserved HOGs which can then be mapped to the genome using MiniProt[49].

Comparative Genomics Using GENESPACE and OrthoFinder

The GENESPACE (v1.2.3)[50] R package was employed to identify orthogroups and orthologues between different accessions, Lu-1, Kar-1, Altai-5, and the reference genome. Multiple plots were then created to visualize a summary of the orthogroup distribution. Furthermore, synteny between accessions was

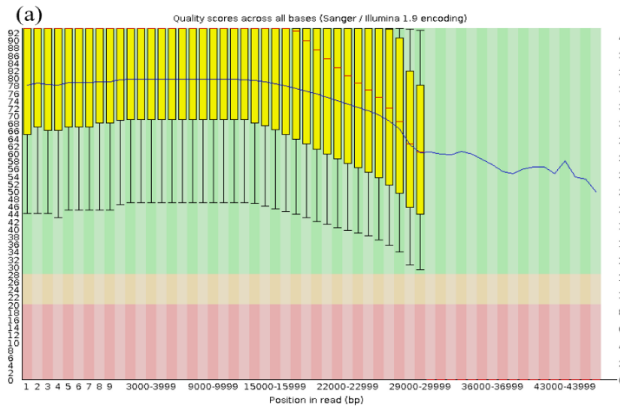
visualized using dotplots, and structural rearrangements between accessions were checked using riparian plots.

RESULTS

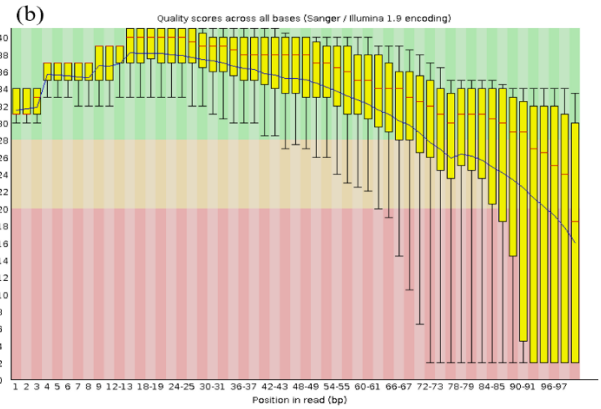
Reads and Quality Check

Read lengths were different across datasets, with PacBio HiFi reads ranging from 62 to 46,056 bp, while Illumina RNA-Seq reads were consistently 101 bp. The GC content was 46% for RNA-seq reads and 36% for PacBio reads, consistent with expected values for A. thaliana genomes. Looking at the quality per base graph (fig.1-[a,b]), filtering was deemed necessary using Fastp to improve Illumina RNA-Seq reads quality. A total of 4.54 million reads were filtered, improving the percentage of high-quality bases(fig.1-c). After trimming, Q20 and Q30 bases increased from 88.92% to 93.15% and from 78.55% to 83.06%, respectively.

(a)



(b)



(c)

	Before Filtering	After Filtering
Total Reads	45.241360 M	40.704842 M
Q20 Bases (%)	88.924755	93.146078
Q30 Bases (%)	78.550219	83.056931
GC Content (%)	46.334435	45.970670

(d)

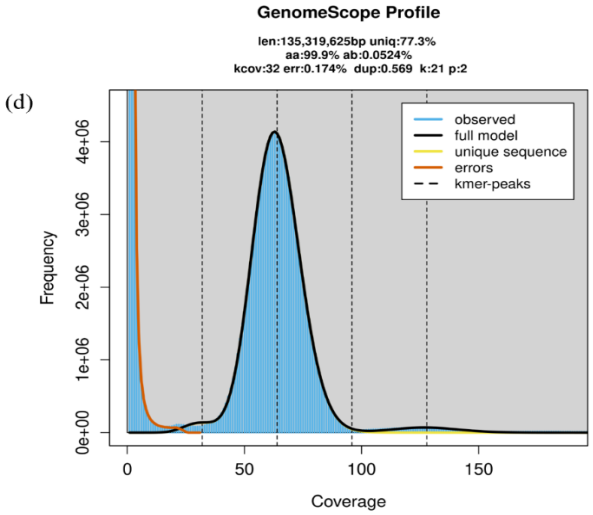


Figure 1 (a) Quality score distribution per base for the Lu-1 genome, (b) quality score distribution per base for Sha transcriptome R1, (c) fastp quality control statistics before and after filtering for the transcriptome, and (d) k-mer frequency distribution generated by Jellyfish.

With a mean read length of 17,365, 570,354 reads, and a total genome size of 135,000,000 bp, the depth of coverage was calculated at 73.36x which ensured sufficient coverage of the Arabidopsis genome. Using Jellyfish and GenomeScope, the haploid genome size was estimated to be around 135.3 Mbp, matching the expected size for Arabidopsis thaliana. The heterozygosity was calculated to be 0.0524%. However, as noted in the introduction, Lu-1 accession exhibiting signs of heterozygosity was observed in GenomeScope k-mer histogram showing a small peak at half the sequencing depth of the main homozygous peak. This indicates alternate alleles contributing to the reads.

BUSCO Results

All three assemblers achieved very high completeness (>99%), indicating assemblies successfully capturing essential genes. Flye and LJA exhibited similar duplication rates (~6%), while Hifiasm showed a significantly lower duplication rate (1%) and thus a higher single-copy BUSCO genes (fig.2-a). The transcriptome assembly's BUSCO completeness (78.7%) is lower than that of the genome assemblies (fig.2-a), which is expected given that transcriptomes are inherently more complex due to alternative splicing and dynamic gene expression. A significant portion of BUSCOs(39.4%) were found to be duplicated, potentially reflecting multiple isoforms of transcripts or assembly errors.

Genome Assembly Statistics (QUAST Results)

All three assemblies achieved a genome fraction of ~91%, indicating good reference genome coverage. The final assembly sizes ranged from 138 to 155.9Mb with contig N50 sizes of 11.4 – 14.5Mb. LJA had the highest number of contigs, Hifiasm had the highest number of misassemblies, while Flye exhibited the fewest contigs and missassemblies. Stats are summarized in figure 2-b.

Merqury Analysis

Flye demonstrated the highest consensus QV (66.4) while LJA achieved the highest completeness (99.39%) (fig.2-b). When comparing the k-mer plots (fig.2-c), Flye k-mer plot shows 2 peaks, red and blue, at the same main sequencing depth indicating that Flye retained both copies of certain regions in the assembly. These peaks likely reflect uncollapsed heterozygous regions or duplicated sequences. On the other hand, Hifiasm and LJA k-mer plots show 1 peak only at the main sequencing depth suggesting possible collapse of haplotypes and duplicates into a single consensus sequence.

Dotplots and Structural Comparisons

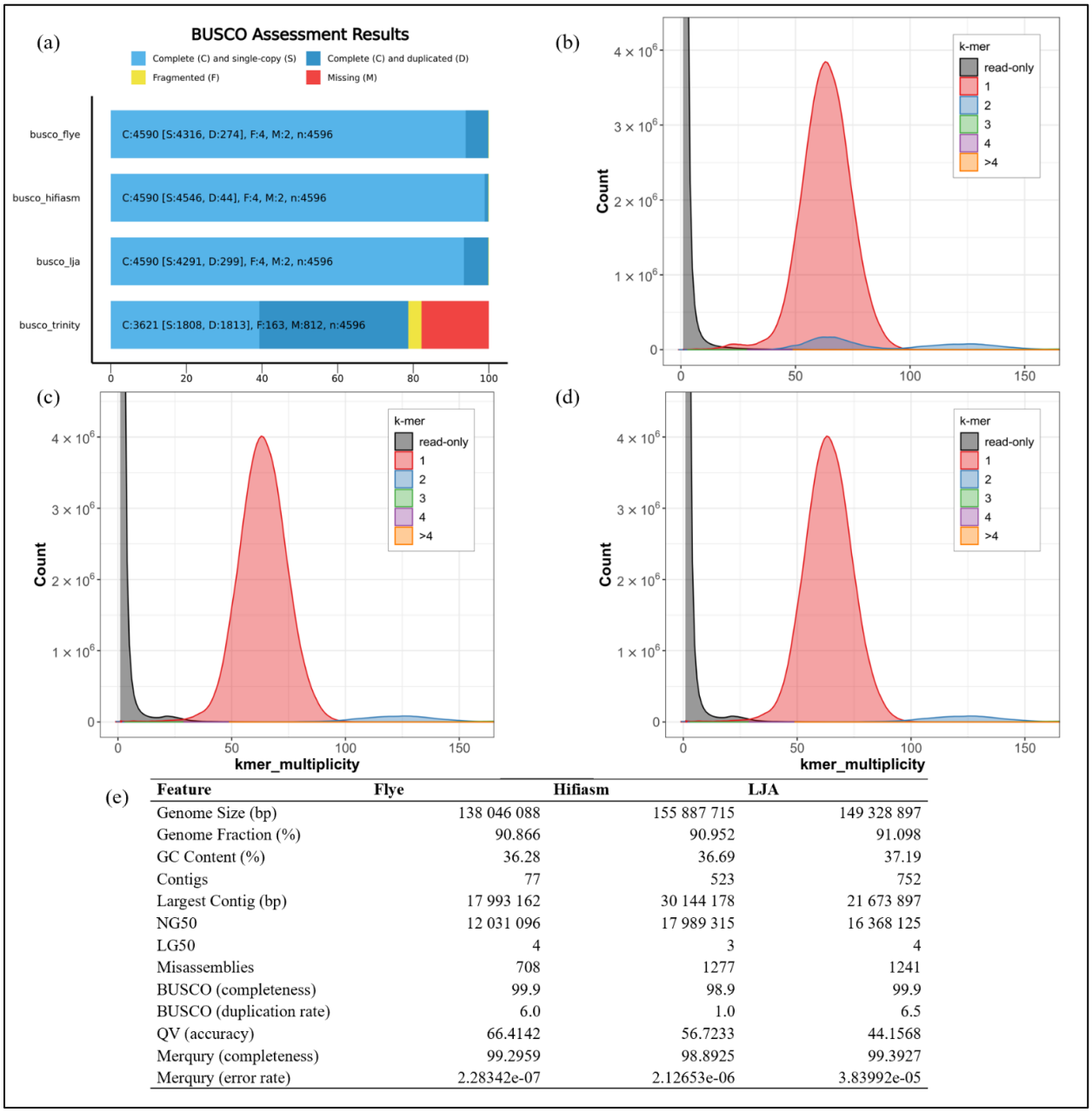
In the Flye and LJA assemblies, some regions are aligning multiple times to the same location in the reference genome indicating the retention of duplicated sequences. In contrast, Hifiasm assembly does not show duplicated alignments in the dotplot, indicating that the assembler has likely collapsed duplicate sequences and handled repeats more accurately (fig.3).

TE Annotation Using EDTA

The transposable element (TE) annotation of the Arabidopsis thaliana Lu-1 genome revealed a total of 596 intact full-length LTR retrotransposons (fig.4-a). Among these, 126 belonged to the Copia superfamily, 160 to Gypsy, and the remaining 310 were classified as unknown. Looking at their percent identity, the high values suggest that they are young insertions that have not yet accumulated significant mutations. This supports the idea that these elements are part of a recent wave of transposition activity, with minimal evolutionary divergence from the parental sequence. Copia superfamily show higher identity compared to Gypsy suggesting more active transposon activity for Copia. The difference in the number of full length LTR-RTs between clades is shown in figure 4-b.

Analysis of clade-level classification showed a dominance of the unknown category, with Athila, Ale, Tekay, and Ivana clades also contributing

significantly (fig.4-c). Most of the elements exhibited very high sequence identity, ranging between 99% and 100%, indicating that they represent young insertions. There was no significant representation of older TEs with lower sequence identity, suggesting either a recent amplification wave or efficient genome mechanisms to eliminate ancient TEs. Athila, Ale, and Ivana showed significant activity, likely contributing to recent transposon expansion.



(e)

Feature	Flye	Hifiasm	LJA
Genome Size (bp)	138 046 088	155 887 715	149 328 897
Genome Fraction (%)	90.866	90.952	91.098
GC Content (%)	36.28	36.69	37.19
Contigs	77	523	752
Largest Contig (bp)	17 993 162	30 144 178	21 673 897
NG50	12 031 096	17 989 315	16 368 125
LG50	4	3	4
Misassemblies	708	1277	1241
BUSCO (completeness)	99.9	98.9	99.9
BUSCO (duplication rate)	6.0	1.0	6.5
QV (accuracy)	66.4142	56.7233	44.1568
Merqury (completeness)	99.2959	98.8925	99.3927
Merqury (error rate)	2.28342e-07	2.12653e-06	3.83992e-05

Figure 2 (a) BUSCO completeness plot for genome and transcriptome assemblies, (b) Merqury k-mer frequency plot for the Flye assembly, (c) Merqury k-mer frequency plot for the Hifiasm assembly, (d) Merqury k-mer frequency plot for the LJA assembly, and (e) table summarizing genome assembly statistics.

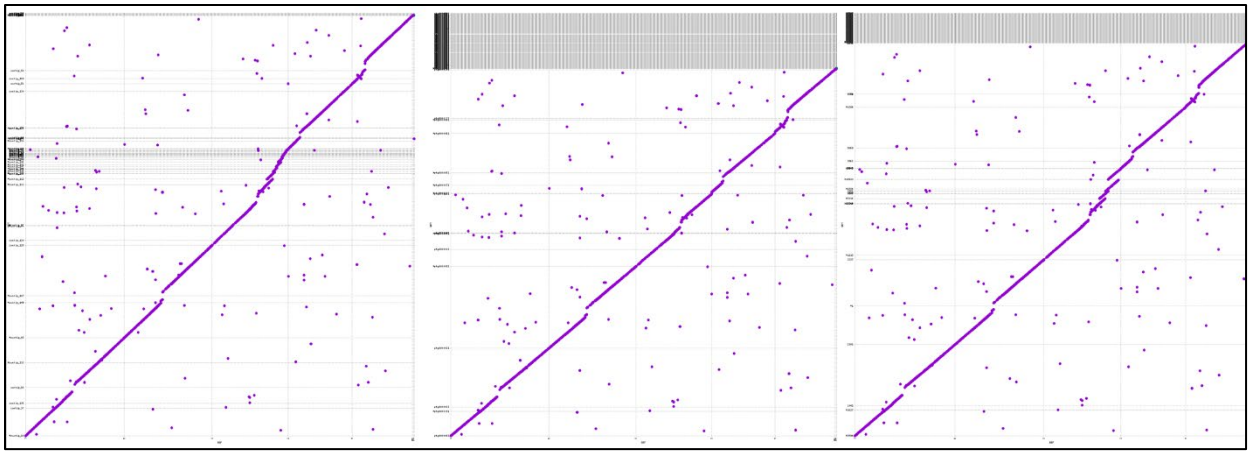


Figure 3 Dotplots comparing genome assemblies: Flye (left), Hifiasm (center), and LJA (right).

In all TE families, 23,082 TEs were annotated which accounted for approximately 12.86% of the genome, masking over 20 million base pairs out of the total genome size with LTRs accounting for the majority of TEs (60.42% of all TEs). The Gypsy retrotransposons alone masked over 2% of the genome (fig.4-a). Among all TE categories, LTRs and helitrons were the two most abundant categories across genomes.

The TE content for other accessions ranged from 15 to 16%, which is higher than the Lu-1 accession possibly due to the latter's large genome size compared to other accessions. When comparing TE content among plant genomes, maize has a much higher TE content at about 77%, largely due to retrotransposons. Similarly, wheat is among the most TE-dense, with over 80% of its genome composed of TEs, contributing to its large size. Meanwhile, rice lies in the middle with TEs representing around 35%–40% of its genome[60].

Visualization of TE density along the top longest scaffolds revealed high TE density regions shown as overlapping peaks particularly for the Gypsy and Copia superfamilies (fig.5-a). Additionally, the TE clade abundance comparison between my accession and other Arabidopsis accessions (Kar-1 and St-0) was compared (fig.5-b). It is worth mentioning that the accession St-0 had shown signs of heterozygosity or even possible concatenation of raw reads coming from 2 different accessions.

TE Age Estimation

The high sequence identity observed in most elements confirmed their recent origins, with minimal divergence from the ancestral sequences. By estimating the insertion times for the 2 main peaks, we found that most of the TEs across the genome expanded within the last 3 or 17 million years, though numerous originated more recently, suggesting ancient TE insertions from a distant common ancestor or a TE with high sequence conservation maintained across evolutionary time, possibly through horizontal transfer or exceptional preservation (fig.6).

Gene Models and Functional Annotation

The annotation pipeline using MAKER resulted in the identification of 38,431 protein-coding genes which decreased to 35,572 after filtering for high-confidence annotations, which is higher than the 27,416 genes annotated in the reference sequence.

Assessment of Annotation Quality

The analysis revealed for the proteins and transcripts sets respectively 96.7% and 97.8% complete BUSCOs, of which respectively 1.5% and 4% were duplicated (fig.7). Furthermore, out of 35572 proteins in the filtered annotation set, 27,084 showed hits with BLAST.

(a)

Order/Superfamily	Percentage of Genome Size (%)	Number of Elements (Intact Elements)
LTR/Copia	0.77	821 (126)
LTR/Gypsy	2.07	2273 (160)
LTR/Unknown	4.93	5822 (310)
LINE	0.43	1202
Helitron	3.19	8895
Total	12.86	23082

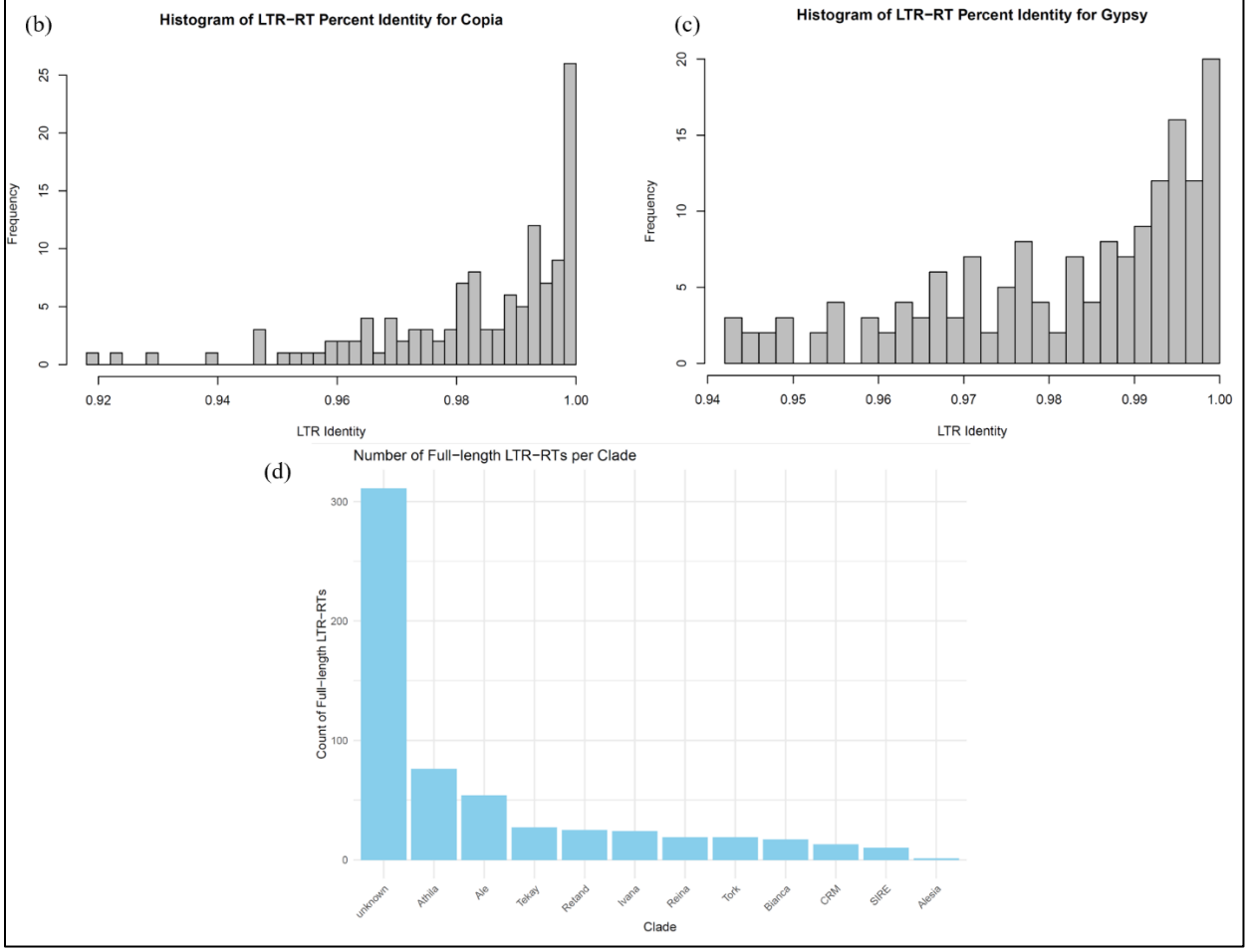


Figure 4 (a) Table summarizing TE content and the number of intact elements, (b) histogram showing percent identity of full-length LTR-RT Copia elements, (c) histogram showing percent identity of full-length LTR-RT Gypsy elements, and (d) bar chart depicting the number of full-length LTR-RTs per clade.

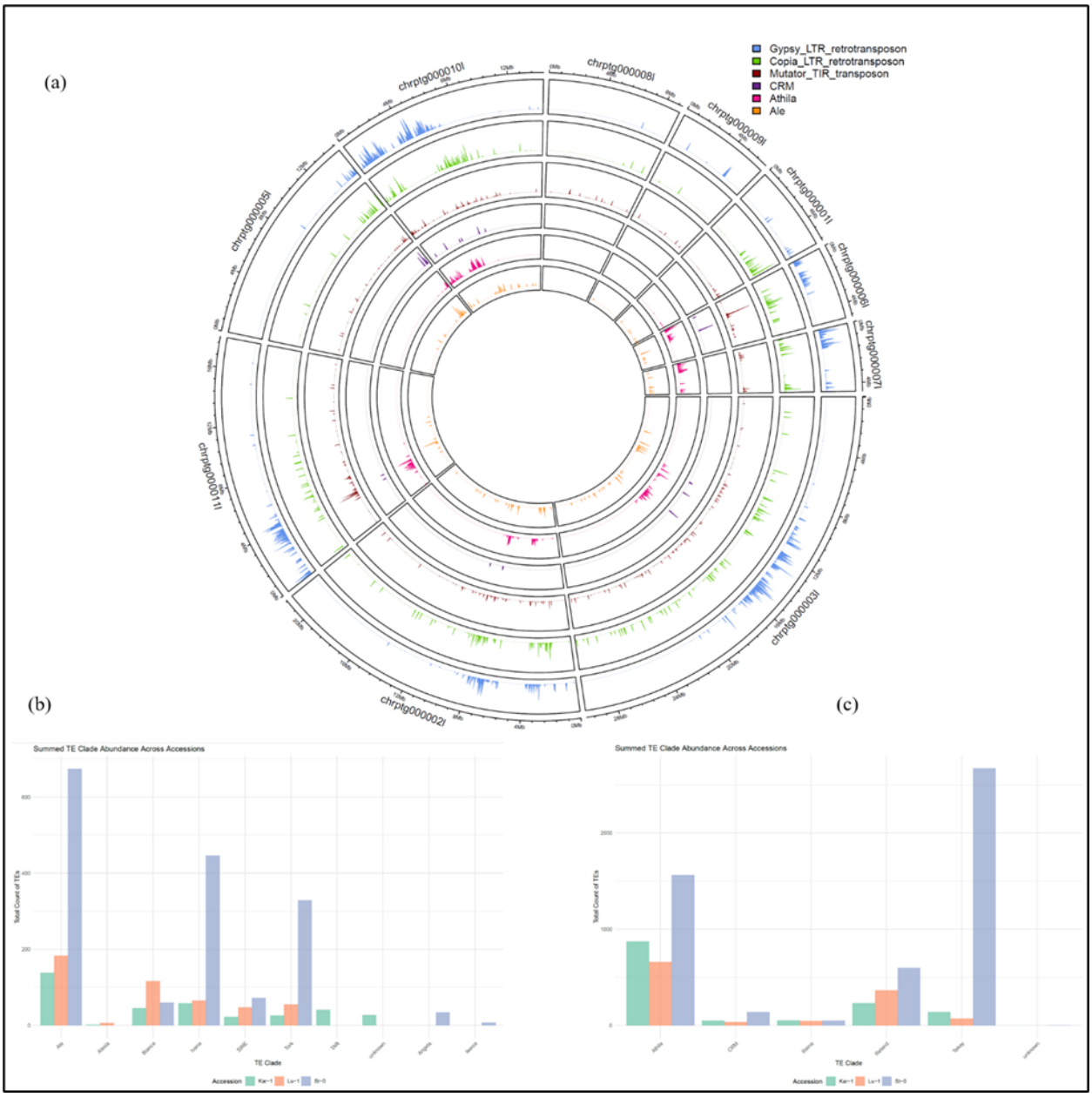


Figure 5 (a) TE density along the top longest scaffolds, (b) comparison of Copia clades abundance between different accessions accession, and (c) comparison of Gypsy clades abundance between different accessions.

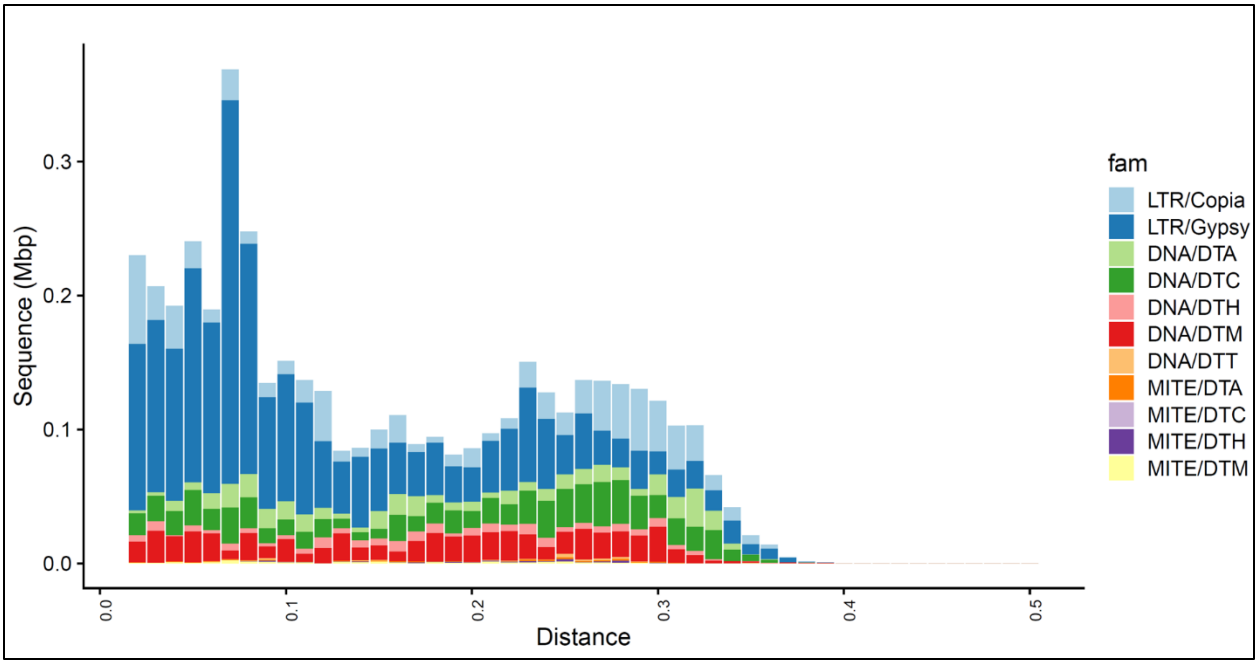


Figure 6 TE age distribution showing recent expansion events.

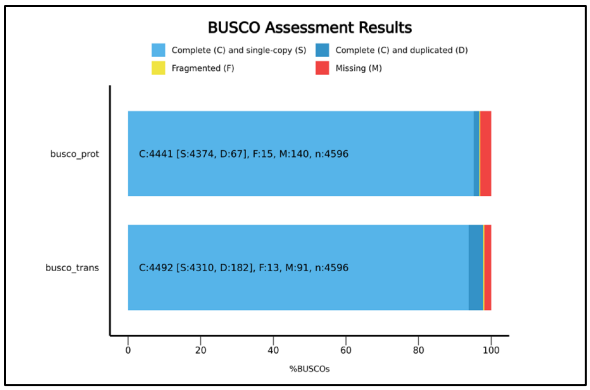


Figure 7 BUSCO analysis evaluating the quality and completeness of the proteome and transcriptome annotations.

Orthology-Based Quality Assessment of Gene Annotations

A total of 17996 HOGs were identified. Additionally, the analysis revealed 91.7% of the annotated proteins had orthologs, indicating strong conservation within the dataset.

Comparative Genomics and Orthogroup Distribution

Comparative genomic analysis using GENESPACE identified 27,730 genes with 98.6% of them assigned to orthogroups. 19,744 orthogroups were core across all 3 accessions and the reference genome. In contrast, 251 genes (0.9% of all genes) were unique to Lu-1. Orthologue distributions were summarized in visualizations in figure 8.

Synteny and Structural Rearrangements

Synteny between the same accessions and the reference genome was visualized using dotplots, which showed syntenic blocks indicating conserved regions of the genome. Structural rearrangements, including inversions, translocations, and duplications, were observed in regions, as highlighted in riparian plots (fig.9).

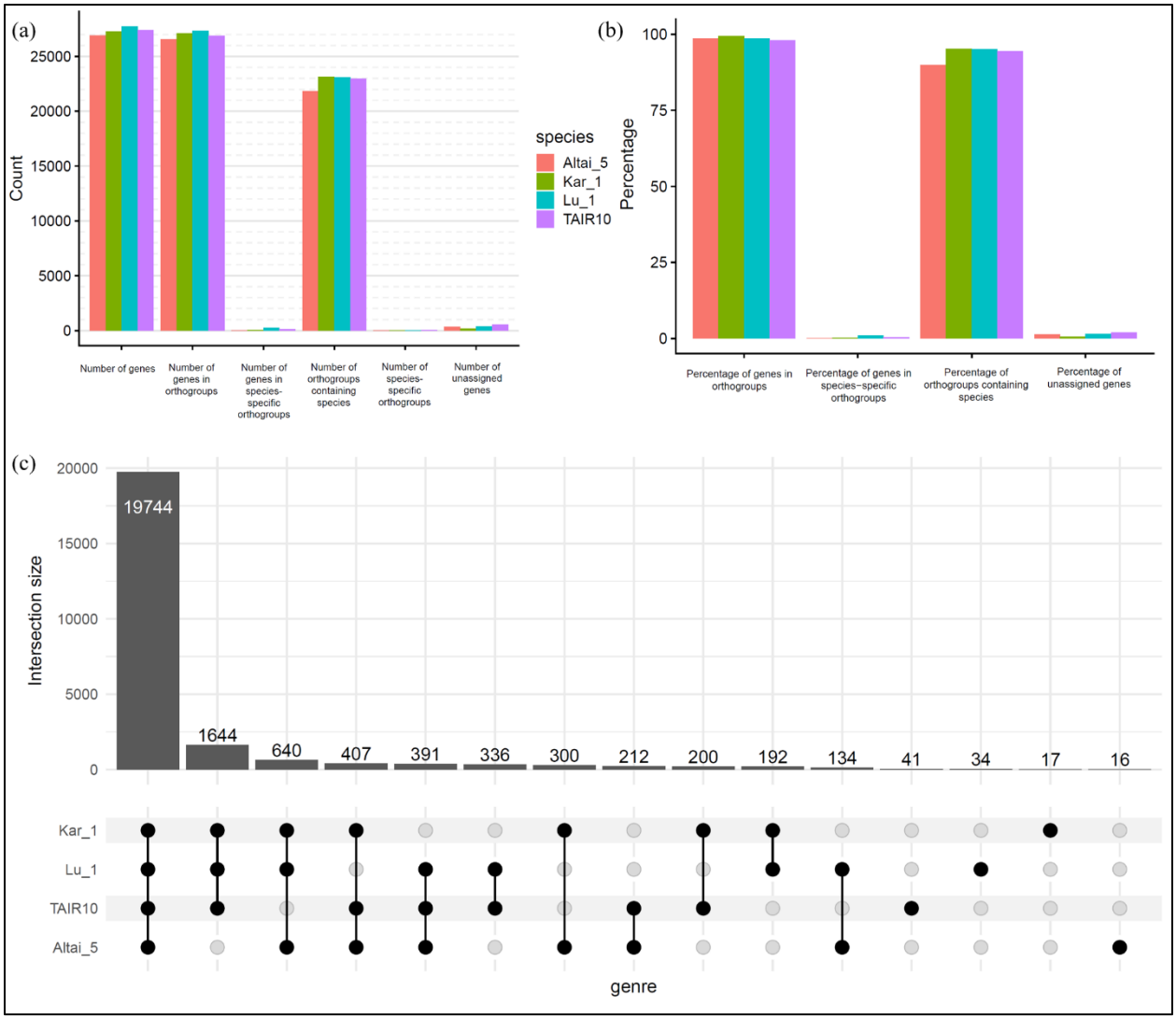


Figure 8 (a) Orthogroup distribution across the reference genome (TAIR10), Lu-1, Altai-5, and Kar-1 accessions, (b) percentage of shared orthogroups among the four genomes, and (c) plot of one-to-one orthogroups highlighting direct gene correspondences.

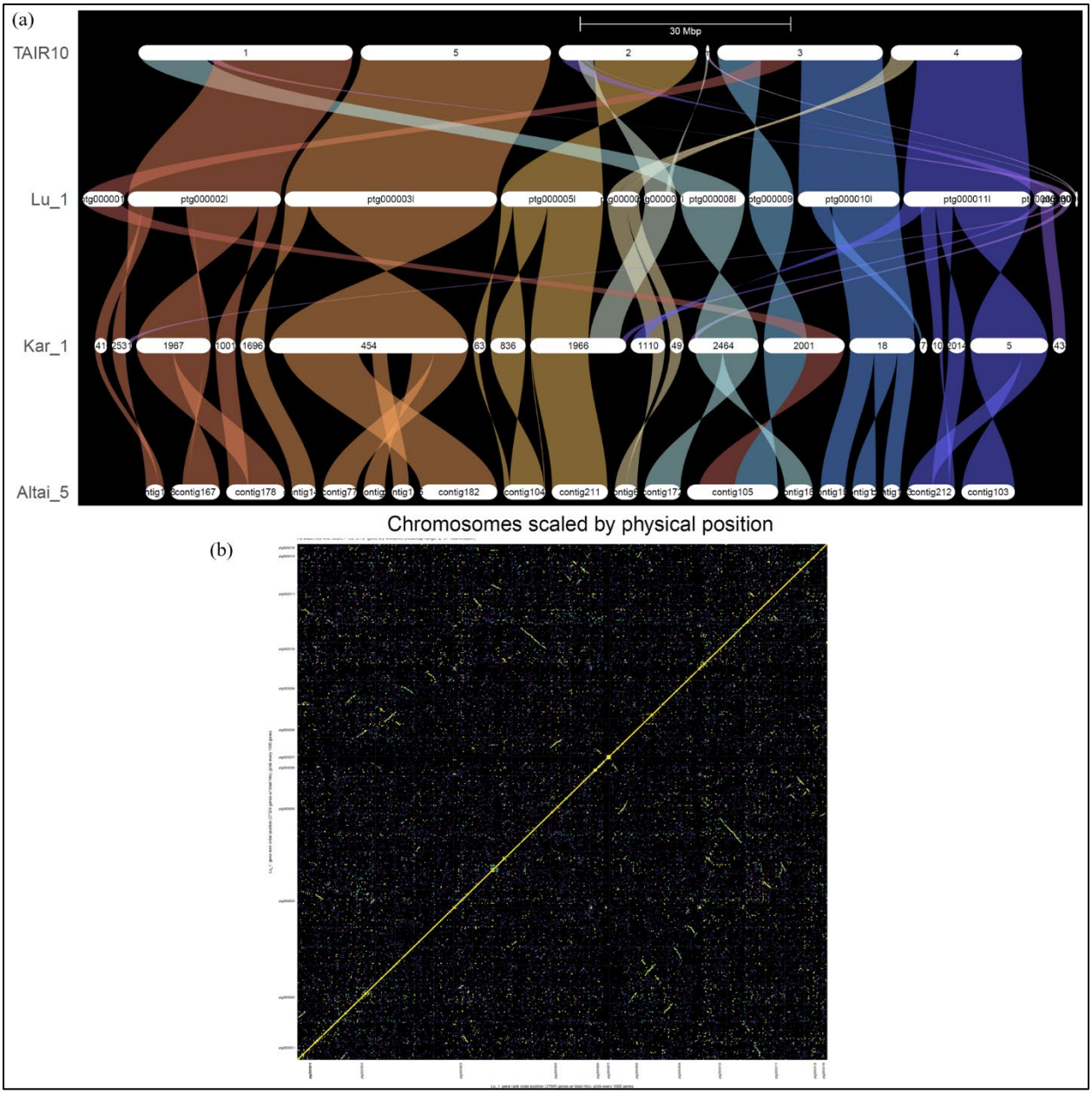


Figure 9 (a) Riparian plot displaying chromosomes scaled by physical position for TAIR10, Lu-1, Kar-1, and Altai-5, and (b) raw hit alignments within the Lu-1 genome (Lu_1_vs_Lu_1.rawHits).

DISCUSSION

Choosing the Best Assembler for Lu-1

Choosing the best assembler for accession Lu-1 depends on different factors such as the quality of the assembly, the complexity of the genome, and the specific objectives of the research. It is important to note that no single assembler consistently outperforms others across all metrics. However, each tool has shown its unique advantages which emphasize its suitability for different research applications. Considering the indications of heterozygosity in my accession mentioned in the paper, this aspect will play a critical role in determining the most appropriate assembler for this report. For that, it is also important to note that reference-based metrics used to evaluate real read assemblies rely on comparisons with an assembly rather than the original genome from which the reads were sequenced.

Comparative Analysis of Assembly Metrics and Insights from K-mer Analysis

Analyzing BUSCO results reveals that Hifiasm, with its low duplication rate of 1%, appears to have effectively minimized over-assembly and maintained haplotype structures more accurately. In contrast, Flye and LJA, with their duplication rate of 6% and 6.5% respectively, show a slightly higher tendency toward duplication. Nonetheless, all 3 assemblers achieved very high completeness indicating that the assemblies are largely comprehensive. QUAST results reveal that they have a genome fraction close to 91%. LJA and Hifiasm produced remarkably more contigs then Flye, which could be indicative of higher fragmentation. Although Hifiasm exhibited the highest NG50, a metric that does not account for assembly errors or errors in the reference genome, it still showed a higher number of misassemblies, mainly due to relocations. The large number of misassemblies suggests that many result from differences between maternal and paternal alleles, even for an inbreeding species, since the Lu-1 accession shows traces of heterozygosity. Furthermore, metrics like quality value (QV) scores and k-mer-based completeness are crucial to assess genome assembly quality. The highest achieved completeness by LJA can likely be attributed to their k-mer-based assembly strategy, particularly the use of the Multiplex de Bruijn Graph methodology. However, Flye had the highest QV score and the lowest error rate out of the 3 assemblers. Analyzing the k-mer plots generated by Merqury, the largest peak in the middle corresponds to the main sequencing coverage depth of the genome for single-copy homozygous regions. The presence of a second peak with the double-copy portion at that same location could indicate heterozygosity that was not successfully collapsed. In the case of residual heterozygosity in the assembly, heterozygous regions are misassembled as 2 separate haplotypes; the reads from each of these regions would have the same k-mer multiplicity as single-copy regions. Consequently, Flye plot shows that it has retained alternative alleles, however the 2 other assemblers could have been able to collapse most haplotypes into a single consensus sequence. To check whether this can be supported, the dotplots for Flye and LJA assemblies show that some regions are aligning multiple times to the same location in the reference genome; these duplicates may explain the small peak of the double-copy portion at the main sequencing coverage depth in the Flye k-mer plot. On the other hand, the Hifiasm assembly does not show duplicated alignments when aligned against the reference genome. Taking all this into account, Flye and LJA produce shorter assemblies by collapsing heterozygous regions, which may result in a biologically incomplete representation caused by the loss or merging of genes; this will reduce the annotation accuracy. In contrast, Hifiasm, resulting in the longest assembly, retains both haplotypes and thus preserves heterozygosity; this offers better foundation for annotation.

Selecting an assembler often involves balancing various strengths and limitations to match the specific goals of an assembly project. Since this

accession was excluded from the study by Lian et al. (2024)[19], the ideal assembler is the one that has clearly demonstrated its ability to effectively address this particular challenge. Since a non-redundant, streamlined assembly is needed for annotation, the Hifiasm assembly will be picked for the rest of the report.

Limitations of the Current Assembly Approach

In this report, long PacBio Hifi reads were successfully used to assemble Arabidopsis Thaliana Lu-1 genome. The Hifiasm assembly yielded a total of 523 contigs, which is significantly larger than the expected 5 chromosomes under perfect conditions. This huge difference reveals limitations in the current approach and its failure to achieve a chromosome-level assembly. In other words, using 1 assembler as well as 1 type of reads was insufficient for achieving the level of contiguity required for resolving entire chromosomes. Introducing Illumina short reads for polishing and error correction[51] in addition to genetic maps for structural scaffolding, or even using multiple assemblers can improve the assembly[52,53,54]. Furthermore, using Hi-C data instead of PacBio Hifi reads could be another solution enabling the assembly of telomere-to-telomere sequences and thus reducing the number of unplaced contigs significantly and improving contiguity. This will directly overcome the challenge of resolving large-scale structural relationships between contigs.

Gene Annotation and Orthogroup Analysis

A total of 35,572 filtered gene models were predicted using the described methods, a number significantly higher than in other accessions, where the expected number of genes ranges from 27,246 to 28,989. Of these, 19,721 genes are reported in the paper as being conserved across all accessions, meaning every accession should have at least this core set of genes. In this report, the initial filtering step was based on AED scores below 0.5. When comparing orthogroups between Lu-1 and other accessions, specifically Kar-1 (from a nearby region in Europe) and Altai-5 (from a distant region in Central Asia), 19,744 genes were found to be shared among all these accessions and the reference genome TAIR10. This consistency supports the findings reported in the paper. There were 27730 annotated genes in the Lu-1 accession, and 1.4% of them were unassigned to orthogroups. Focusing on one-to-one orthogroups, Lu-1 and Kar-1 share 192 orthogroups, while Lu-1 and Altai-5 share 134. This difference reflects their geographical proximity, as accessions from closer regions are expected to share a greater degree of genetic similarity compared to those from more distant regions. Moreover, the higher number of private genes in Lu-1 (0.9% of all genes compared to the expected 0.4%) could be a direct consequence of heterozygosity, or it could reflect its unique evolutionary trajectory. Private genes, resulting from gene duplication and horizontal gene transfer, have been shown to play roles in biotic and abiotic stress responses, including pathogen resistance and drought tolerance[55,56].

When further filtering based on Pfam domain presence was applied, the number of predicted genes dropped to 28012. While this step prioritizes functional annotations, it came with the cost of a significantly reduced BUSCO completeness (80%): some removed genes lack Pfam domains but still encode conserved orthologous sequences captured by BUSCO analysis.

Enhancements for Future Analysis

Given the lack of confidence in the generated results because of heterozygosity, several additional enhancements to improve the current analysis will be discussed for the remainder of the report. First, the clustering inconsistencies observed in the phylogenetic trees indicate problems in resolving relationships within each clade. A better approach would involve using maximum likelihood methods with bootstrapping, such as those implemented in RAxML[57] to provide statistical support. Second, structural variations can be explored; they can have a huge impact on gene function,

regulation, and genome architecture. Third, integrating transcriptome profiles under different conditions could help characterize the functional significance of annotated genes and TEs. TEs are known to contribute to gene regulation through epigenetic modification or promoter hijacking. Studying differential gene expression across various environmental conditions can provide insights into how TEs modulate gene expression in response to stimuli. Additionally, differential gene expression studies could help validate the functional relevance of private genes in Lu-1[58,59]. Fourth, both the study in question as well as this report used RNA transcriptome data from only a single accession to annotate all genomes which causes some limitations. Using a single transcriptome as a universal reference introduces biases and may overlook accession-specific genes, alternative splicing events, and regulatory elements, leading to an incomplete annotation, especially for genes expressed under specific environmental conditions. Though more costly, generating transcriptomic data for each accession would allow to capture the gene expression and splicing diversity. Fifth, to further investigate the functional significance of the private genes, enrichment analyses can identify pathways enriched in private genes and thus provide insights into adaptive traits in Lu-1.

CONCLUSION

In conclusion, this report has assembled and annotated the Lu-1 genome, leveraging the strengths of Hifiasm to preserve genetic diversity and heterozygosity. Despite some limitations, such as unresolved structural contiguity and incomplete annotation, the assembly provides a robust foundation for further functional studies. Future efforts to integrate additional data types and assembly strategies could further enhance genomic insights into the Lu-1 accession.

SUPPLEMENTARY MATERIALS

All scripts can be found in the GitHub repository: https://github.com/andrew-maalouf/assembly-annotation-course. Additional plots are included in the appendix.

REFERENCES

[1] Jones, A. M. E., Whiteman, S. A., Serazetdinova, L., Sanders, D., Rathjen, J., Peck, S. C., & Maathuis, F. J. M. (2008). Identification of novel proteins and phosphorylation sites in a tonoplast-enriched membrane fraction of Arabidopsis thaliana. Proteomics, 8(17), 3536–3547. https://doi.org/10.1002/pmic.200701104

[2] Jones, J. D. G., & Dangl, J. L. (2006). The plant immune system. Nature, 444(7117), 323–329. https://doi.org/10.1038/nature05286

[3] Provart, N. J., Alonso, J., Assmann, S. M., Bergmann, D., Brady, S. M., & et al. (2016). 50 years of Arabidopsis research: Highlights and future directions. New Phytologist, 209(3), 921–944. https://doi.org/10.1111/nph.13687

[4] Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., ... & Bergelson, J. (2010). The scale of population structure in Arabidopsis thaliana. PLoS Genetics, 6(2), e1000843. https://doi.org/10.1371/journal.pgen.1000843

[5] Hays, J. B. (2002). Arabidopsis thaliana, a versatile model system for study of eukaryotic genome-maintenance functions. DNA Repair, 1(6), 579–600. https://doi.org/10.1016/S1568-7864(02)00015-0

[6] Hou, X., et al. (2022). Chromosome-scale assembly and annotation of eight Arabidopsis genomes. Genome Biology, 23(1), 241. https://doi.org/10.1186/s13059-022-02797-0

[7] Koornneef, M., et al. (2004). Natural variation and natural selection in Arabidopsis thaliana. Trends in Plant Science, 9(5), 219–226. https://doi.org/10.1016/j.tplants.2004.03.004

[8] The 1001 Genomes Consortium (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell, 166(2), 481–491. https://doi.org/10.1016/j.cell.2016.05.063

[9] Mills, R.E., Bennett, E.A., Iskow, R.C., & Devine, S.E. (2007). Which transposable elements are active in the human genome? Trends in Genetics, 23(5), 183–191. https://doi.org/10.1016/j.tig.2007.02.001

[10] International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators, Appels, R., Eversole, K., Feuillet, C., Keller, B., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 361(6403), eaar7191. https://doi.org/10.1126/science.aar7191

[11] Marand, A.P., Zhao, H., Zhang, W., Zeng, Z., Fang, C., & Jiang, J. (2019). Historical meiotic crossover hotspots fueled patterns of evolutionary divergence in rice. Plant Cell, 31(3), 645-662. https://doi.org/10.1105/tpc.18.00751

[12] Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nature Genetics, 43(12), 1160-1163. https://doi.org/10.1038/ng.983

[13] Kumar, A., & Bennetzen, J. L. (1999). Plant retrotransposons. Annual Review of Genetics, 33, 479-532. https://doi.org/10.1146/annurev.genet.33.1.479

[14] Hays, D. B. (1997). Repeat landscape analysis. The Plant Cell, 9(9), 1509-1519. https://doi.org/10.1105/tpc.9.9.1509

[15] Eickbush, T. H., & Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Research, 134(2), 221-234. https://doi.org/10.1016/j.virusres.2007.11.012

[16] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., ... & Pasy, A. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12), 973-982. https://doi.org/10.1038/nrg2165

[17] Kunze, R., Saedler, H., & Lönnig, W. E. (1997). Plant transposable elements. Advances in Botanical Research, 27, 331-470.

[18] Jaeglé, B., Soto-Jiménez, L. M., Burns, R., et al. (2023). Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity. Genome Biology, 24(1), 44. https://doi.org/10.1186/s13059-023-02875-3

[19] Lian, Q., et al. (2024). A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. Nature Genetics, 56(5), 982–991. https://doi.org/10.1038/s41588-024-01715-9

[20] Jiao WB, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nature Communications. 2020;11:1–10. Available from: http://dx.doi.org/10.1038/s41467-020-14779-y

[21] Andrews, S. (2010). FastQC: A Quality Control tool for High Throughput Sequence Data. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[22] Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

[23] Ranallo-Benavidez, T. R., Harkess, A. E., & Ganko, E. W. (2020). PBJelly: a tool for improving PacBio assembly using short-read data. bioRxiv. https://doi.org/10.1101/2020.01.27.921538

[24] Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., ... & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, 33(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

[25] Bankevich, A., Bzikadze, A., Kolmogorov, M., Antipov, D., & Pevzner, P. A. (2022). Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. Nature Biotechnology, 40(4), 491–499. https://doi.org/10.1038/s41587-022-01220-6.

[26] Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Flye: A novel and efficient assembler for single-molecule sequencing. Bioinformatics, 35(9), 1797–1806. https://doi.org/10.1093/bioinformatics/bty102

[27] Cheng, H., Liang, H., & Liu, Z. (2021). Hifiasm: A fast and accurate long-read assembly method for large genomes. Nature Methods, 18(2), 168–175. https://doi.org/10.1038/s41592-020-00963-x

[28] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., & Amit, I. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. https://doi.org/10.1038/nbt.1883

[29] Manni, M., Berkeley, M. R., & Seppey, M. (2021). BUSCO: Assessing genome assembly and annotation quality with a comprehensive set of single-copy orthologs. Bioinformatics, 37(1), 200–208. https://doi.org/10.1093/bioinformatics/btaa787

[30] Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

[31] Rhie, A., McCarthy, S. A., Pugach, I., Targaryen, S., & Schmitz, M. (2020). Merqury: An evaluation tool for genome assemblies. Nature Methods, 17(6), 629-633. https://doi.org/10.1038/s41592-020-0792-4

[32] Kurtz, S., Phillippy, A. M., Delcher, A. L., Smoot, M., Shumway, M., & Antonescu, C. (2004). Versatile and efficient sequence alignment with Mauve. Bioinformatics, 20(12), 3081-3089. https://doi.org/10.1093/bioinformatics/bth427

[33] Delcher, A. L., Salzberg, S. L., & Phillippy, A. M. (2003). Using MUMmer to identify similar regions in large sequence sets. Current Protocols in Bioinformatics, 1(1), 10. https://doi.org/10.1002/0471250953.bi1010s00

[34] Formenti, G., Abueg, L., Brajuka, A., Brajuka, N., Gallardo-Alba, C., Giani, A., Fedrigo, O., & Jarvis, E. D. (2022). Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. Bioinformatics, 38(17), 4214–4216. https://doi.org/10.1093/bioinformatics/btac460

[35] Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology, 20(1), 275. https://doi.org/10.1186/s13059-019-1905-y

[36] Zhang, R.-G., Li, G.-L., Wang, X.-L., et al. (2022). TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Horticulture Research, 9, uhac017. https://doi.org/10.1093/hr/uhac017

[37] Gu, Z., Eils, R., & Schlesner, M. (2014). circlize: An R package for circular visualization. Bioinformatics, 30(19), 2811-2812. https://doi.org/10.1093/bioinformatics/btu393

[38] R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org

[39] Li H, Handsaker B, Wysoker A, et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

[40] Stajich JE, Block D, Boulez K, et al. (2002). BioPerl: Perl modules for the life sciences. Genome Research, 12(10), 1603–1608. https://doi.org/10.1101/gr.361502

[41] Kagale, S., & Rozwadowski, K. (2014). Bioinformatics tools for plant biology and genomics. Journal of Integrative Plant Biology, 56(1), 2-16. https://doi.org/10.1111/jipb.12106

[42] Sievers, F., & Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. Multiple Sequence Alignment Methods, 105-116. https://doi.org/10.1007/978-1-4939-0372-0_7

[43] Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution, 26(7), 1641-1650. https://doi.org/10.1093/molbev/msq053

[44] Letunic, I., & Bork, P. (2016). Interactive Tree Of Life (iTOL) v3: An online tool for the display and annotation of phylogenetic trees. Nucleic Acids Research, 44(W1), W242-W245. https://doi.org/10.1093/nar/gkw290

[45] Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., ... & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research, 18(1), 188-196. https://doi.org/10.1101/gr.6743907

[46] Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236-1240. https://doi.org/10.1093/bioinformatics/btu031

[47] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

[48] Nevers, Y., Glover, N., Dessimoz, C., & Boussau, B. (2024). Quality assessment of gene repertoire annotations with OMArk. Nature Biotechnology, 42(4), 497–505. https://doi.org/10.1038/s41587-024-02147-w

[49] Li, H. (2023). Protein-to-genome alignment with miniprot. Bioinformatics, 39(1), btad014. https://doi.org/10.1093/bioinformatics/btad014

[50] Schmutz, J., Lovell, J., & USDOE. (2021). GENESPACE R Package (GENESPACE) v1.0. OSTI.GOV. https://doi.org/10.11578/dc.20210610.3

[51] Laumann, A. E., Stanke, M., & Zerbino, D. R. (2021). Genome polishing with Illumina short reads. Bioinformatics, 37(3), 458-460. https://doi.org/10.1093/bioinformatics/btaa716.

[52] Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., & Drake, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6), 563-569. https://doi.org/10.1038/nmeth.2474

[53] Michael, T. P., & VanBuren, R. (2020). Building near-complete plant genomes with single-molecule sequencing technologies. Nature Plants, 6(5), 4-11. https://doi.org/10.1038/s41582-020-0330-0

[54] Jiao, Y., Peluso, P., Shi, J., & et al. (2017). Improved maize reference genome with single-molecule technologies. Nature, 546(7659), 524-527. https://doi.org/10.1038/nature22971

[55] Van de Peer, Y., Mizrachi, E., & Marchal, K. (2009). The evolutionary significance of polyploidy. Nature Reviews Genetics, 10(5), 375-385. https://doi.org/10.1038/nrg2576

[56] Panchy, N., Lehti-Shiu, M. D., & Shiu, S. H. (2016). Evolution of gene duplication in plants. Plant Physiology, 171(4), 2294-2314. https://doi.org/10.1104/pp.16.00873

[57] Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

[58] Anders, S., Pyl, P. T., & Huber, W. (2013). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

[59] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

[60] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2013). Classification, identification, and nomenclature of transposable elements in plants: A revised proposal. Genome Biology and Evolution, 5(3), 494–504. https://doi.org/10.1093/gbe/evt141

APPENDIX

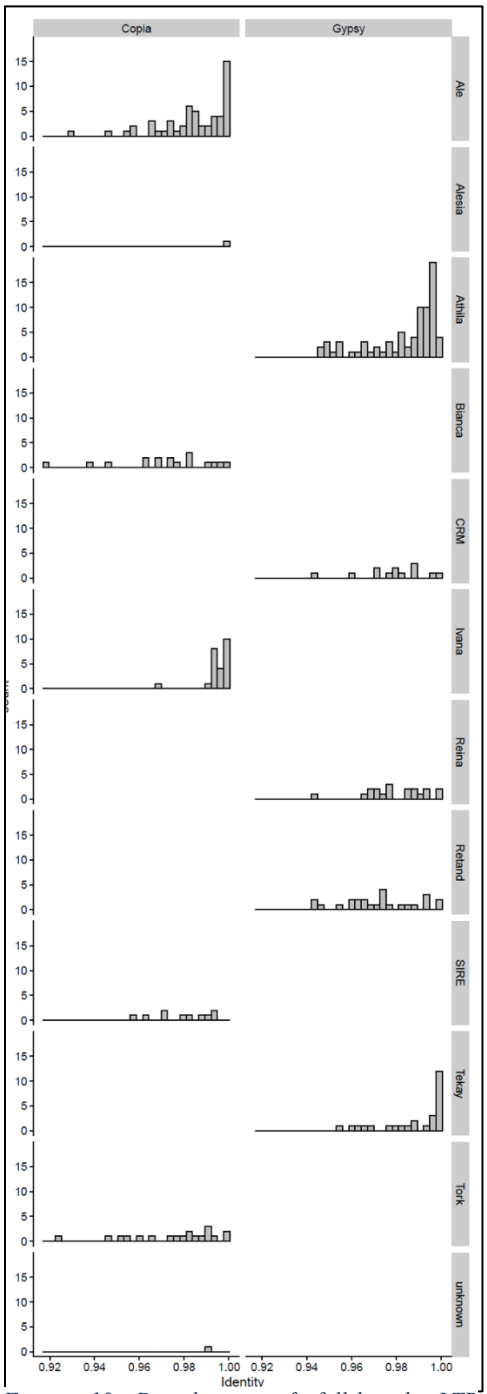


Figure 10 Distribution of full-length LTR retrotransposons (LTR-RTs) across all clades. The x-axis represents percent identity, while the y-axis shows the count of LTR-RTs. Distinct patterns are observed for superfamilies such as Gypsy and Copia, as well as specific clades, highlighting variability in sequence conservation and abundance.

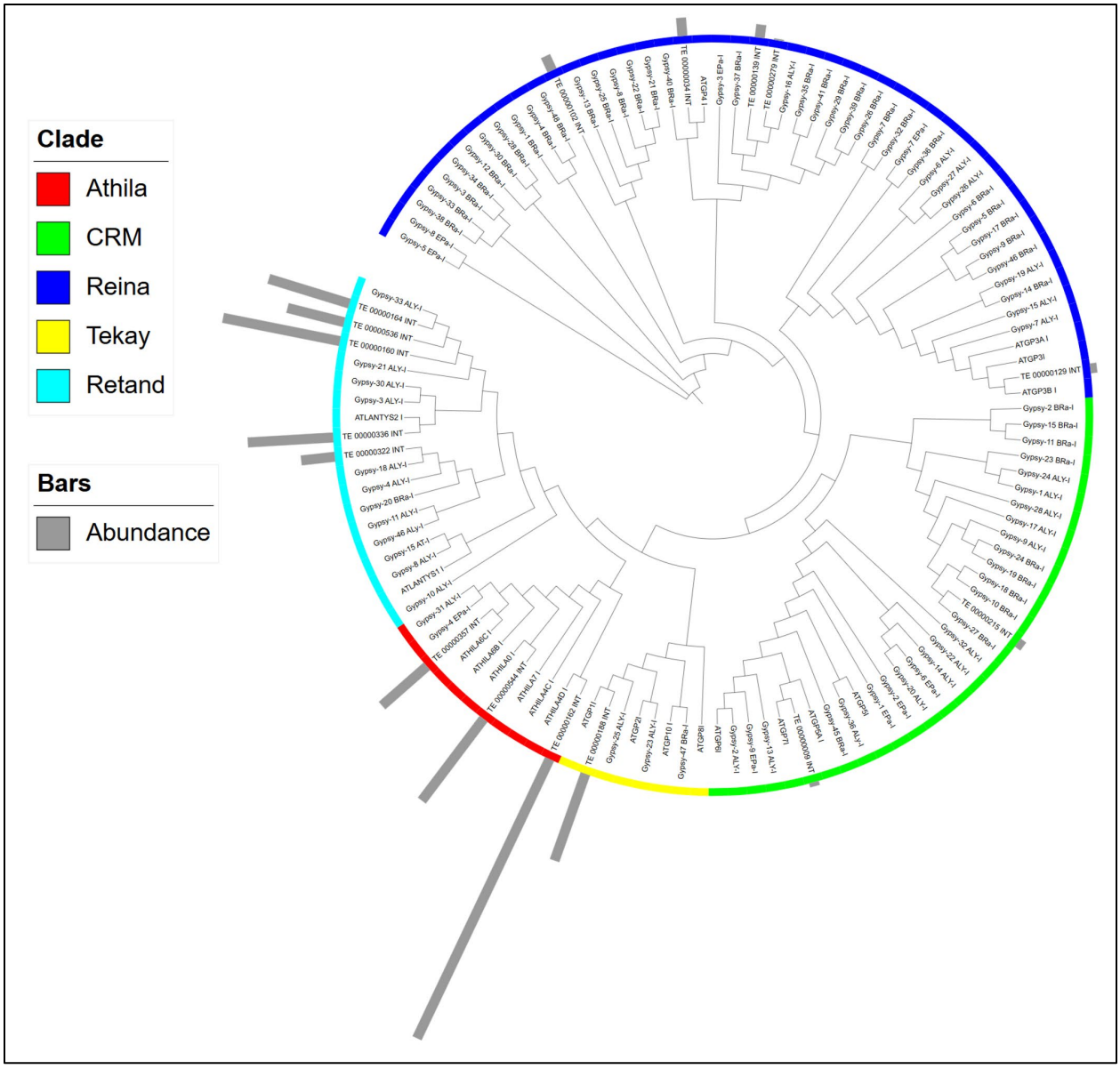


Figure 11 Phylogenetic analysis of transposable elements (TEs) from the Gypsy superfamily.

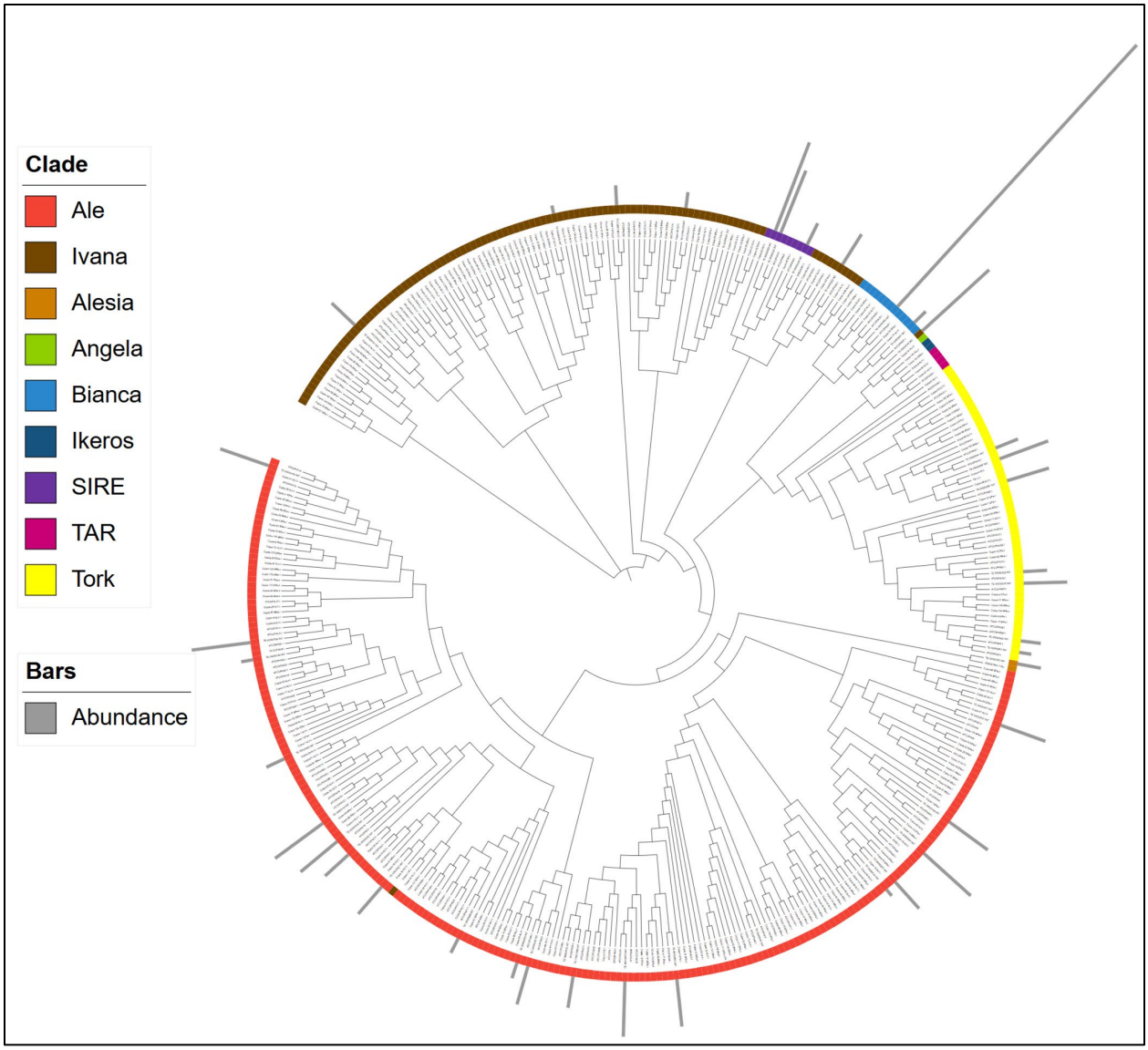


Figure 12 Phylogenetic analysis of transposable elements (TEs) from the Copia superfamily.

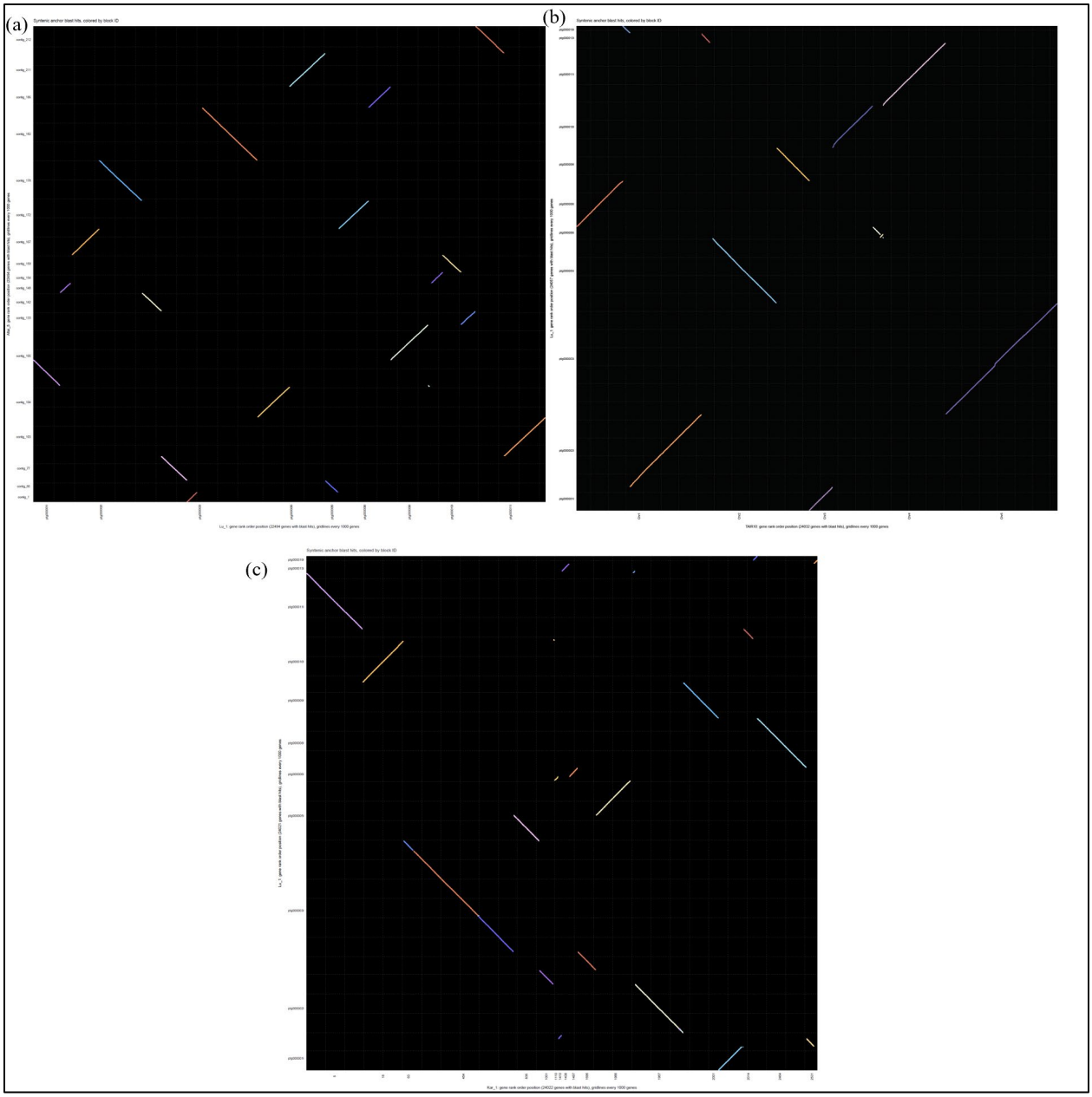


Figure 13 Dotplots showing syntenic hits among Arabidopsis thaliana accessions: (a) Syntenic relationships between Lu-1 and Altai-5, (b) Syntenic relationships between TAIR10 and Lu-1, and (c) Syntenic relationships between Kar-1 and Lu-1. Each dot represents a conserved syntenic block between the compared genomes, visualizing genomic rearrangements and conservation patterns.

Declaration

I hereby declare that I have written this report independently and have not used any sources other than those indicated. I have marked as such all passages, including illustrations, which have been taken literally or analogously from sources. I am aware that otherwise the lecturer responsible may assign an unsatisfactory grade for the work, even retrospectively.

I declare that for this work ...

❑ have not used any AI technologies.

☒ have used the following AI technologies:

I acknowledge the use of ChatGPT for refining the sentence structure and enhancing the clarity of the English in this report.

After using these AI services, I have checked the work and take full responsibility for the content of the submitted work. I am aware that in case of unreflected use of these services, the generated text may be considered as plagiarism.