*Laura Fernández Cerro | University of Bern*

# Whole genome assembly and annotation of *Arabidopsis Thaliana*

## Abstract

*Arabidopsis thaliana* is a model plant extensively used for research. Its genome has been sequenced and assembled multiple times. Nonetheless, Arabidopsis thaliana is broadly distributed across different geographic regions and has adapted to diverse environments. Whole-genome assemblies and annotations from different geographic distributions can help study the evolution and adaptation of this species, particularly by identifying regions where genomic collinearity is conserved or regions with high genetic diversity. High-quality long-read sequencing approaches, such as PacBio HiFi sequencing, significantly improve genome assembly, enabling the accurate reconstruction and annotation of repetitive elements. Interestingly, transposable elements and centromeric regions are key areas of study, as *A. thaliana* exhibits higher variability in centromeric regions compared to chromosomal arms, where large genome rearrangements are rare. In this project, in the first step, the genome of an European accession, Est-0, has been assembled using three different tools (Flye, LJA and Hifiasm) to compare and select the best outcome. In a second step, the best assembly was annotated, focusing on transposable elements and the gene annotation. Additionally, a preliminary explorative comparative genomic analysis was conducted between Est-0, the reference genome (TAIR10), and the accession Stw-0.

## 1. Introduction

*Arabidopsis thaliana* (*A. Thaliana*) is a crucial and widely used model organism in plant research due to its genetic simplicity (small genome size, 5 chromosomes, genetic tractability) and rapid life cycle. Its adaptability to diverse environments and geographic locations has made it an interesting specie to study different evolutionary dynamics across its genome [1].

The first complete assembly of a plant genome was *A. Thaliana* (Col-0), considered to be the reference sequence. This approach was based on BACs sequenced with Sanger technology [2]. Later, other *Arabidopsis* genomes have been largely studied based on short-reads sequencing or reference-guided assembly, where the identification of genomic rearrangements was limited [3 , 4 ,  5]. In contrast, the arrival of third generation sequencing and long-reads, such as PacBio technology with HIFI reads allows to chromosome-level assemblies with lower error rates. The longer reads allow to span repetitive regions, reducing fragmentation and improving contiguity. Large repetitive sequences, such as transposable elements (TE), centromeric arrays or large structural variants (inversions, translocations or duplications) are often miss-assembled in the short-read sequencing strategies, whereas the long-read sequencing approach handles better this challenge [6].

Different assembly strategies applied over the same reads would lead to different results. The choice of the algorithm often relies on the coverage and the longitude of the reads. In the case of high-fidelity (HiFi) reads, three of the most common used tools are LJA, Hifiasm and Flye. LJA utilizes a De Bruijn graph approach optimized for long reads. It employs a multiplex De Bruijn graph to efficiently resolve repetitive regions and produce highly contiguous assemblies, making it suitable for genomes with complex repeat content [7]. Hifiasm, on the other hand, leverages phased assembly graphs to generate haplotype-resolved assemblies, which are valuable for analysing heterozygous regions and structural variants, but sometimes result in more fragmented assemblies compared to other tools [8]. Flye uses a repeat graph-based method specifically designed for long-read sequencing data, and it is remarkable for handling repetitive sequences and producing highly contiguous assemblies [9].

High-quality assemblies are crucial for accurate annotation, as the reliability of annotations is directly dependent on the quality of the underlying assembly. Comparative analyses of *A. thaliana* whole-genome *de novo* assemblies and their corresponding annotations across different geographical accessions are essential for understanding genome evolution and synteny conservation [10].

The gene annotation strategy can combine intrinsic and extrinsic evidence. The *ab initio* methods, rely on features within the genome sequence itself, such as sequence composition, open reading frames (ORFs), gene structure signals (e.g., start and stop codons), and splice site motifs, to predict genes. The extrinsic approach can integrate evidence from the well-characterized proteins, transcriptome information derived from RNA-sequencing data or expressed sequence tags (ESTs), and genome evidence, using comparative genomics to transfer annotations from closely related reference genomes [11].

Interestingly, the annotation and study of transposable elements can be crucial, since it has been shown that they can drive evolution by controlling gene expression through mutagenic transposition [12]. The classification of TEs can be divided into retrotransposon (Class I), which are known for its replication via RNA and cDNA intermediates, and DNA transposons (Class II), which replicate through a DNA intermediate. Overall, the transposon frequency correlates with the genome size, since in eukaryotes is expected that the higher content in repetitive elements results in a larger genome size [13]. It is estimated that TEs occupy the 75% of the maize genome, 40% of the human genome, and 20% of *A. Thaliana* genome. The distribution of TEs and TE classes is not homogeneous between genomes neither across different regions of the same genome, there is variation in abundance [14] .In *Arabidopsis*, the Gypsy and Copia families are the predominant LTR retrotransposons, and LINE elements are the predominant non-LTR retrotransposons [15].

In a recent study from Lian et al. [16], the genomes different accessions of *A. thaliana* from different geographic locations classifies in groups as Europe, Madeira, Asia and Africa were assembled in chromosome-level and further annotated. It was found that the genomic collinearity is highly conserved among geographically and genetically distant accessions, indicating a quasi-fixed karyotype. Along chromosomal arms large rearrangements are rare; however, the centromeres account for the divergences and structural variations.

In this project, the PacBio HIFI reads from the European accession Est-0 and short Illumina reads from the RNA-sequencing of the transcriptome of *A. Thaliana* are assembled [16], [17]. The main objective is to perform the whole-genome assembly of the accession Est-0 and its consequent annotation regarding genes and transposable elements. The whole-transcriptome assembly was used as evidence and guidance for the gene annotation. Three different assembly strategies - Hifiasm, LJA and Flye - were performed over the whole-genome assembly in order to identify the best-performing approach. Beyond the assembly and annotation, this study further explored the relationship between the TE content and distribution across the genome, as well as its dynamics and impact on the evolution of the Est-0 accession.

## 2. Materials and Methods

### 2.1. Raw reads and quality control

The sequencing data of the accession Est-0 come from a previous studies [16], [17], concretely the whole genome reads were sequenced with PacBio HiFi and the whole transcriptome was sequenced with Illumina RNA-seq.

The quality of the reads was assessed with FastQC (version 0.11.9) [18]. In order to improve the quality of the short Illumina reads, they were filtered and trimmed using Fastp (version 0.23.2) [19].

In addition, the *k-mer* counting was performed over the genome reads with Jellyfish (version 2.2.6) [20] in order to estimate the depth of coverage, the number of *k-mers* found without sequencing errors, the genome size and the percentage of heterozygosity. Concretely, canonical *k-mers* were used to ensure that the forward and reverse complement of a *k-mer* are treated as identical, since the DNA sequences are double-stranded.

### 2.2. Genome and transcriptome assembly

The software used for the whole transcriptome assembly was Trinity (version 2.15.1) [21]. For the whole genome assembly three different tools were used: Flye (version 2.9.5) [9] , Hifiasm (version 0.19.8)[8] and LJA (version 0.2) [7] to compare their performance and choose the best assembly achieved for annotation. Due to the high quality of the reads, the polishing step was not performed.

### 2.3. Assembly evaluation

The contiguity, completeness and correctness of the three whole genome assemblies was evaluated using BUSCO (version 5.4.2)[22], QUAST (version 5.0.2) [23] and merqury (version 1.3) [24]. The whole transcriptome assembly was also evaluated with BUSCO (version) [22] .

The mode (transcriptome or protein) and the lineage (brassicales_odb10) was specified to run BUSCO. In the case of protein mode, the input file contained only the MAKER-produced longest protein sequences to ensure that each coding gene is represented by its most complete sequence, since shorter or alternative protein isoforms might lead to duplicated BUSCO hits, distorting the results. QUAST was performed with reference genome (TAIR10) and without reference genome.

### 2.4. Genome assemblies' comparison

Pair-wise comparisons were performed between the assembled genomes from flye, hifiasm and LJA and the *Arabidopsis thaliana* reference genome using MUMmer (version 3.0) [25].

### 2.5. Annotation of TEs

Only the Flye genome assembly, chosen as the best, regarding the continuity, completeness and correctness was further annotated. EDTA (Extensive de novo TE annotator, version 1.9.6) pipeline was used in order to annotate both intact and fragmented transposable elements (TEs), create a non-redundant TE library and classify the TEs into superfamilies.

Subsequently, the TE classification was refined using TEsorter (version 1.3.0) based on homology-based detection in order to study the TE clade abundance.

### 2.5.1. TE Age Estimation

RepeatMasker (version 3.01.03) [26] allowed to compute the percentage of diversity of TE copies compared to the references in the TE library, the age of TE insertions was estimated to study the evolutionary dynamics of TEs. The insertion time (T) was calculated as $T = \frac{K}{2r}$, where K is the sequence divergence, and r is the substitution rate. It was assumed that the rate of substitutions per synonymous site per year is $8.22 \cdot 10^{-9}$ according to previous estimations [27].

### 2.5.2. Phylogenetic analysis of TEs

Reverse transcriptase (RT) protein sequences from Copia and Gypsy LTR retrotransposon superfamilies, Ty1 and Ty3, respectively, were extracted, aligned, and used to infer approximately-maximum-likelihood phylogenetic trees. Tools such as Clustal Omega (version 1.2.4) [28] and FastTree (version 2.1.11) [29] were employed, and the resulting trees were visualized and annotated using iTOL (version 7) [30].

## 2.6. Annotation of genes

The annotation of the genes was performed using MAKER pipeline (version 3.01.03) [11], since it integrates several sources of evidence; ab initio prediction models, RNA-Seq data and protein homology. Particularly, the transcriptome assembled in the previous steps was included as EST evidence.

Additional filtering and refinement were made over the gene annotations. InterProScan (version 5.70-102) [31] was used to annotate the protein sequences with functional domains, according to Pfam database [32]. Moreover, the gene models were evaluated based on the Annotation Edit Distance (AED) values, to filter high confidence models genes, with AED ≤ 0.5. Finally, only the mRNA and proteins sequences of high-quality gene models were extracted and retained, to ensure the accuracy and reliability for downstream steps.

## 2.7. Gene annotation evaluation

To evaluate the quality of the gene annotations, BUSCO (version 5.4.2) [22] was used to assess completeness by identifying conserved single-copy orthologs, with results indicating the proportion of complete, duplicated, fragmented, or missing genes. Functional annotation quality was further validated by aligning protein sequences against the UniProt database [33] to identify homologs with known functions. Additionally, the protein sequences were evaluated with OMArk (version 0.3.0) [34] according to the completeness, consistency of protein-coding genes based their homologs and possible contaminations from other species. LUCA database from OMA Browser website was used [35]. For further improvement of the gene annotation, the sequences of conserved Hierarchical Orthologous Groups (HOGs) for which the gene models are fragmented or missing were mapped against the genome assembly with Miniprot (version 0.13) [36] and visualized with JBrowse 2 [37], in order to identify and correct discrepancies.

## 2.8. Comparative genomics

To explore genomic differences between the *Arabidopsis thaliana* reference genome (TAIR10) and the genome assemblies of the accessions Est-0 and Stw-0, GENESPACE (version 1.4) [38] and OrthoFinder (version 2.5) [39] were used. First, gene annotation data for each accession, including BED files for gene coordinates and FASTA files for peptide sequences, were prepared. The 20 longest contigs were selected based on assembly statistics.

OrthoFinder was employed to identify orthogroups shared between the accessions and the reference genome. The results were visualized through summary plots displaying the distribution of shared orthogroups and riparian plots, which showed the syntenic blocks and structural rearrangements.

## 3. Results

### 3.1. Quality control of the reads

As expected, the HiFi-PacBio reads, 319,355 sequences in total, showed overall good quality, in terms of per base sequence quality, per sequence quality scores and sequence duplication levels. However, the short Illumina paired end reads exhibit worse quality, specifically in the ending positions of the reads. In addition, adapter sequences were detected as overrepresented sequences. After trimming and applying the quality filter, 40,704 M reads were retained (roughly the 90%) and 4,536 M of reads (nearly the 10%) were discarded due to low quality.

The *k-mers* counting yielded an estimated genome size of 164,494,408 bp, and low percentages of heterozygosity (0.001%) and read error rate (0.147%).

### 3.2. Genome and transcriptome assemblies

BUSCO estimates the completeness and redundancy based on universal single-copy orthologs. As summarized in Figure 1, the BUSCO results are classified in complete BUSCOs, e.g. number of BUSCOs that are present in the assembly, that could be a single copy or duplicated, the fragmented BUSCOs, and the missing BUSCOs, that are the orthologs that are not found in the assembly. Comparing the transcriptome assembly to the genome assemblies, it is noticeable that there is a higher proportion of duplicated, fragmented and missing BUSCOs, what aligns with the expectations due to isoforms, alternative splicing, and the dynamic nature of the expression of genes.

When comparing the whole genome assemblies (Flye, Hifiasm and LJA), all of them demonstrate high percentages of completeness, both Flye and Hifiasm showed a 99.2% of completeness, whereas Hifiasm completeness was slightly lower, 96.5%. In terms of duplicated and fragmented BUSCOs the three assemblies exhibit similar and successful results, approximately 70 duplicated BUSCOs and 5 fragmented BUSCOs each one. However, the Hifiasm assembly contains 156 missing BUSCOs, in contrast to Fly and LJA, which each contain only 32 missing BUSCOs.

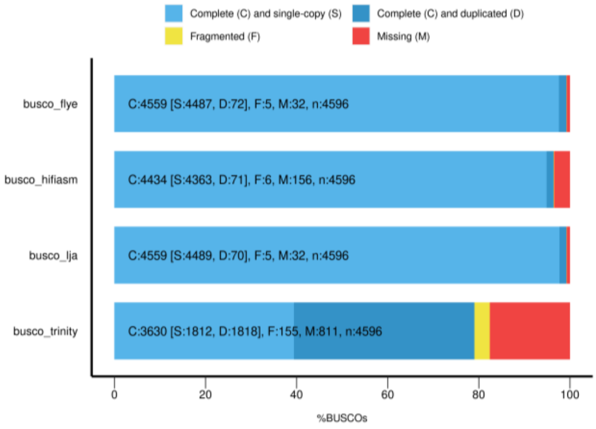
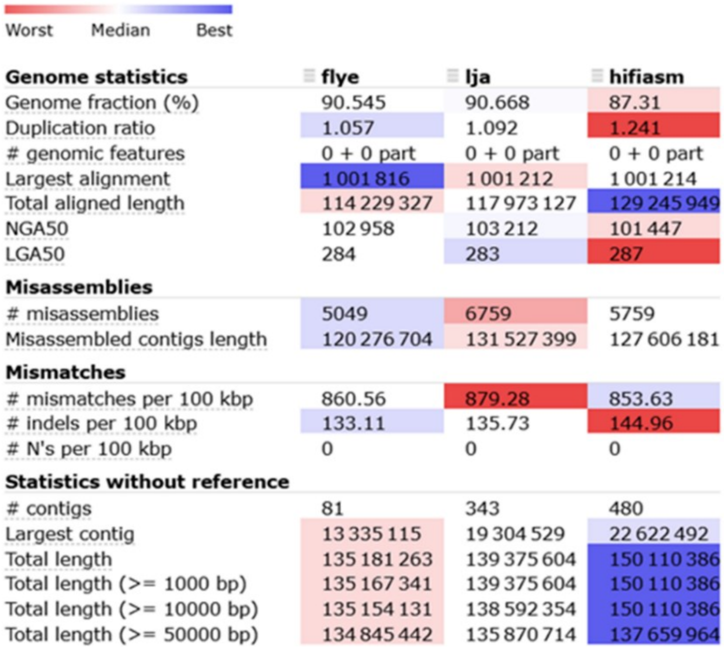


**Figure 1 .** BUSCO assessment results of the whole genomes assemblies using Fly, Hifiasm or LJA software and whole transcriptome assembly using Trinity software.

Based on QUAST metrics (Figure 2), Flye and LJA demonstrate higher genome fractions (90.5% and 90.7%) compared to Hifiasm (87.3%), indicating that the Hifiasm assembly, with 450 contigs, is more fragmented than the LJA and Flye assemblies, with 343 and 84 contigs, respectively. In addition, Flye assembly is considered to be slightly better than LJA and Hifiam in terms of contiguity and correctness, due to lower number of miss-assemblies and its largest alignment. However, the NGA50 of LJA assembly is slightly larger than the one of the Flye assembly. It is remarkable that the interpretation of this metrics depends on the quality of the reference genome, which is considered as high-quality in this case.



Worst Median Best

| **Genome statistics** | **flye** | **lja** | **hifiasm** |
|---|---|---|---|
| Genome fraction (%) | 90.545 | 90.668 | 87.31 |
| Duplication ratio | 1.057 | 1.092 | 1.241 |
| # genomic features | 0 + 0 part | 0 + 0 part | 0 + 0 part |
| Largest alignment | 1 001 816 | 1 001 212 | 1 001 214 |
| Total aligned length | 114 229 327 | 117 973 127 | 129 245 949 |
| NGA50 | 102 958 | 103 212 | 101 447 |
| LGA50 | 284 | 283 | 287 |
| **Misassemblies** | | | |
| # misassemblies | 5049 | 6759 | 5759 |
| Misassembled contigs length | 120 276 704 | 131 527 399 | 127 606 181 |
| **Mismatches** | | | |
| # mismatches per 100 kbp | 860.56 | 879.28 | 853.63 |
| # indels per 100 kbp | 133.11 | 135.73 | 144.96 |
| # N's per 100 kbp | 0 | 0 | 0 |
| **Statistics without reference** | | | |
| # contigs | 81 | 343 | 480 |
| Largest contig | 13 335 115 | 19 304 529 | 22 622 492 |
| Total length | 135 181 263 | 139 375 604 | 150 110 386 |
| Total length (>= 1000 bp) | 135 167 341 | 139 375 604 | 150 110 386 |
| Total length (>= 10000 bp) | 135 154 131 | 138 592 354 | 150 110 386 |
| Total length (>= 50000 bp) | 134 845 442 | 135 870 714 | 137 659 964 |

**Figure 2**. Summary of assessment over the three assemblies (Flye, LJA and Hifiasm) from QUAST.

Merqury results further support the findings from QUAST and BUSCO (Supplementary 1). Regarding the spectra-cn plots, the hifiasm assembly shows a slighter higher peak of duplicates sequence. The error rate is similar for the three genome assemblies, assuming that the *k-mers* only found in assembly are bp errors. The three assemblies show an expected sequencing coverage of 24, that coincides with the peak of homozygous *k-mers*.

The Flye assembly was chosen as the best one, since its balance between contiguity, correctness and completeness. The downstream steps were only performed over this genome assembly.

### 3.3. Genome annotation

The Table 1 summarizes the comparison between the Flye assembly and its annotation achieved in this study to the assembly presented in a previous study [16] based on Quickmerge.

**Table 1.** Comparison of the assembly and annotation metrics of the Est-0 Flye assembly of this study and Quickmerge assembly from *[16]*.

| | Flye assembly Est-0 accession | Quickmerge Est-0 accession |
|---|---|---|
| **Size of assembly** | 135,182,495 bp | 136,122,196 bp |
| **Number of contigs** | 84 | 41 |
| **N50** | 6,923,766 | 11,478,368 |
| **Number of TEs** | 31,382 | - |
| **Base pairs covered by TEs** | 19,489,436 (14.42 %) | 22,171,713.96 (16.29 %) |
| **Number of genes** | 27,711 | - |
| **Number of filtered genes** | 27,193 | - |
| **Genes with Blast hits** | 21,528 | 26,778* (High quality with blastn hit) |
| **Genes without Blast hits** | 5,665 | - |
| **Number of mRNAs** | 29,880 | - |
| **Number of monoexogenic genes** | 9,423 | - |
| **Genes in core orthogroups** | 12,146 | - |
| **Genes in Accession specific orthogroups** | 52 | - |

### 3.3.1. Annotation and dynamics of TEs

As shown in Figure 3A, the most abundant superfamilies are Gypsy, Helitron and Copia. However, the Helitron annotation with EDTA software is not reliable without a further step of curation, due to its complex structure and less conservation. It is observable that the TE distribution of TEs across superfamilies is similar between different accessions (Altai_5, Est_0, Ms_0 and Stw_0).

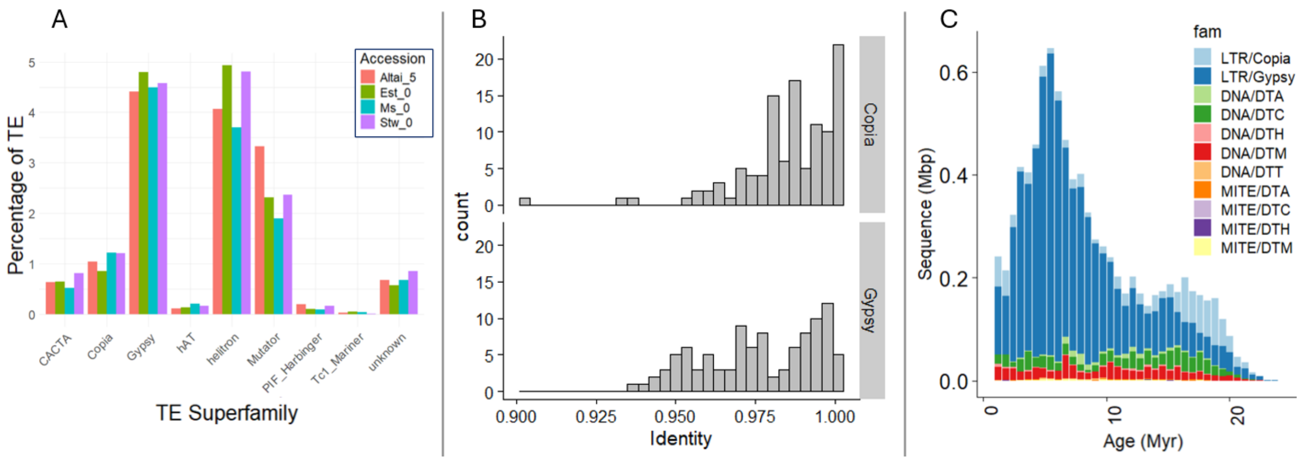


**Figure 3**. A) TE content from EDTA output across different accessions (Altai_5, Est_0, Ms_0 and Stw_0), B) Percentage of identity between long terminal repeats (LTR) of retrotransposons in Copia and Gypsy superfamilies, C) Insertion time (millions of years) of TEs based on sequence divergence.

The total number of TEs after filtering regarding TEs features (long terminal repeats, repeat regions and target site duplications) are 225, where 111 TEs belong to Copia superfamily and 101 to Gypsy superfamily. In a subclassification by clades, the most abundant ones are Ale (45 TEs), Athila (31 TEs) and Retand (28 TEs).

Based on the mutations accumulated over the two identical long terminal repeats (LTR), the percentage of identity can be computed, giving insight about the time. It is assumed that mutations accumulate over time, decreasing the percentage of identity. The Figure 3B shows that the TEs belonging to the superfamilies Copia and Gypsy mainly have high percent identity (99-100%), meaning that they are relatively young insertions. In contrast, old insertions (low percent identity 70-90%) are not found. Furthermore, the TE insertions were dated according to the divergence from the consensus sequence for *Brassicaceae* species. In the Figure 3C, both Copia and Gypsy elements exhibit a recent peak of activity around 5-10 million years ago (Myr). However, there is a gradual decline in activity beyond 10 Myr that is extended until 20 Myr.

In addition, the TE annotations across the scaffolds of the genome assembly were visualized (Supplementary 2), where specially the TEs from the clades CRM and Athila, can be used as markers of the centromeric and pericentromeric regions of the chromosomes.

### 3.3.2. Annotation of genes

As summarized in Table 1, the Flye assembly annotation resulted in a total of 27,711 genes, of which 27,193 were filtered as high confidence. In comparison, the Quickmerge annotation reported 26,778 high-quality genes with blastn hits, indicating similar levels of gene annotation quality. However, the Flye assembly annotation contains more genes lacking blast hits (5,665), potentially representing novel or less-characterized sequences. Moreover, in the Flye assembly 29,880 mRNAs were identified, including 9,423 monoexonic genes, with an average of 5.15 exons per gene (Table 2). Gene lengths presented a median length of 1,782 bp and an average length of 2,287 bp, which accounts for the variability in gene structure. Exon lengths are also variable, with a median of 132 bp and an average of 246.89 bp. Intron lengths are much shorter, with an average of 171.61 bp.

Although both annotations manifest similar results in gene annotation, the quality of the Flye assembly annotation could be improved through further refinement or manual curation, as the extremely small minimum sizes, such as 6 bp for gene and mRNA lengths and 1 bp for exon and intron lengths, may indicate potential annotation errors or artifacts (Table 2).

**Table 2.** Statistics about the gene length, mRNA length, exon length, intro length (in bp) and exons per gene after filtering the gene annotation.

| | Minimum | Maximum | Median | Mean |
|---|---|---|---|---|
| **Gene length** | 6 | 187516 | 1782 | 2287.17 |
| **mRNA length** | 6 | 187516 | 1854 | 2348 |
| **Exon length** | 1 | 7761 | 132 | 246.89 |
| **Intron length** | 1 | 20 | 105 | 171.61 |
| **Exons per gene** | 1 | 960 | 3 | 5.15 |

## 3.4. Comparison of genomes

The genome assemblies developed with the three different software were visualized and compared against the reference genome using Mummer plots. Additionally, the LJA assembly was compared against the Flye assembly (Figure 4). None of the three genome assemblies show major rearrangements compared to the reference genome, indicating as a quality control an overall good performance of the three software. Notably, as previously mentioned, the Hifiasm assembly is less contiguous than the LJA and Flye assemblies, as it contains a greater number of contigs. Figure 4D highlights that the LJA and Flye assemblies are largely concordant, as evidenced by the clear diagonal observed in the plot.

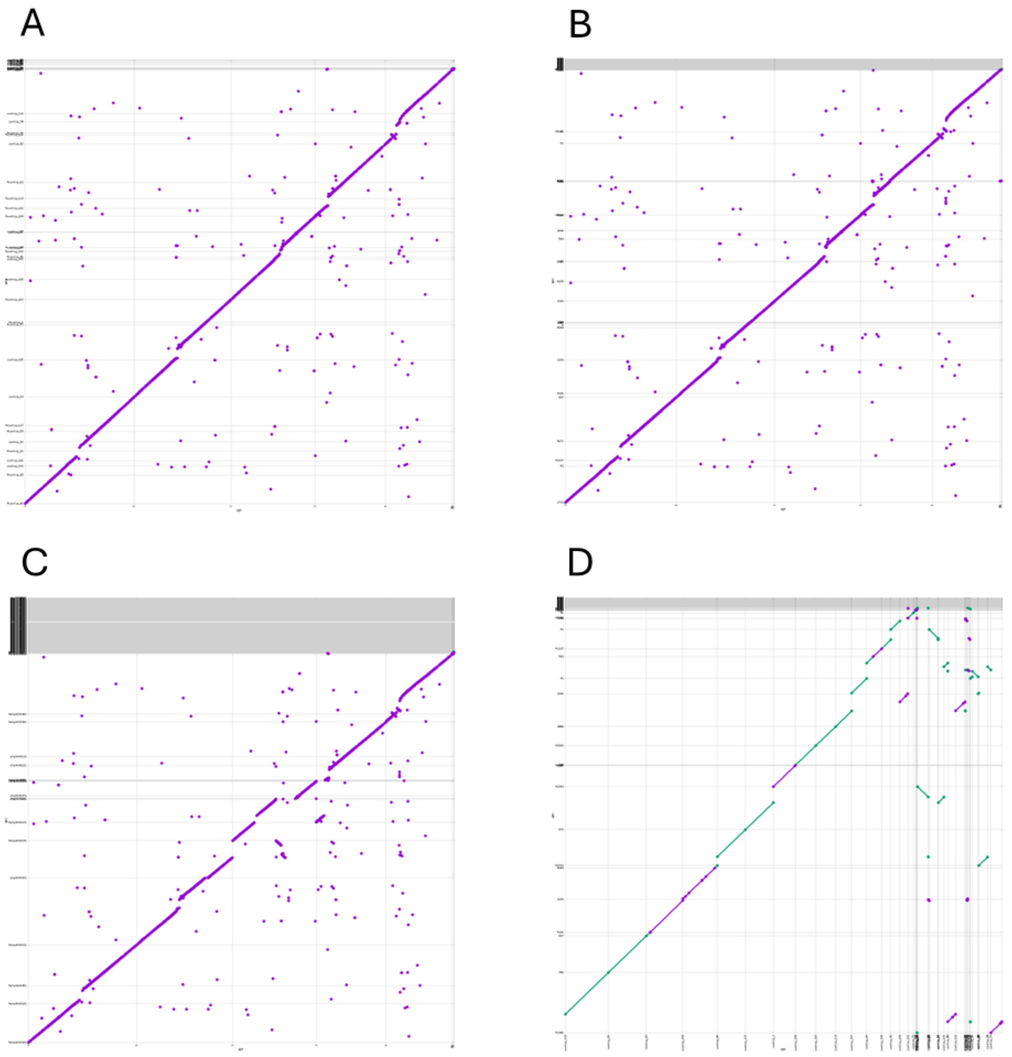


**Figure 4.** Mummer plots of pairwise genome assembly comparisons: A) Flye assembly vs reference genome, B) LJA assembly vs reference genome, C) Hifiasm assembly vs reference genome, D) LJA assembly (green) vs Flye assembly (purple).

The orthogroup analysis shows a substantial overlap of orthogroups between the reference genome (TAIR10) and the Est-0 and Stw-0 accessions, with 9.407 orthogroups shared among all three genomes, indicating strong conservation across *A. thaliana* accessions (Supplementary 3). Furthermore, the riparian plot (Figure 5) remarks this conservation across the three genomes, where the syntenic blocks (contigs) are phased by the reference genome. Concretely, the Est-0 accession is more contiguous with TAIR10, than Stw-0, exemplified by the fragmentation of the chromosome 4 in more than 5 different contigs. It could be due to the fact that this chromosome 4 suffered an inversion in Stw-0 accession, as reported in [16], and has not been assembled properly.

It is observed that the chromosome 2 is inverted in the two accessions compared to the reference genome, which can be explained by the close geographic location of both accessions, inferring that is part of the evolution of these European accessions.

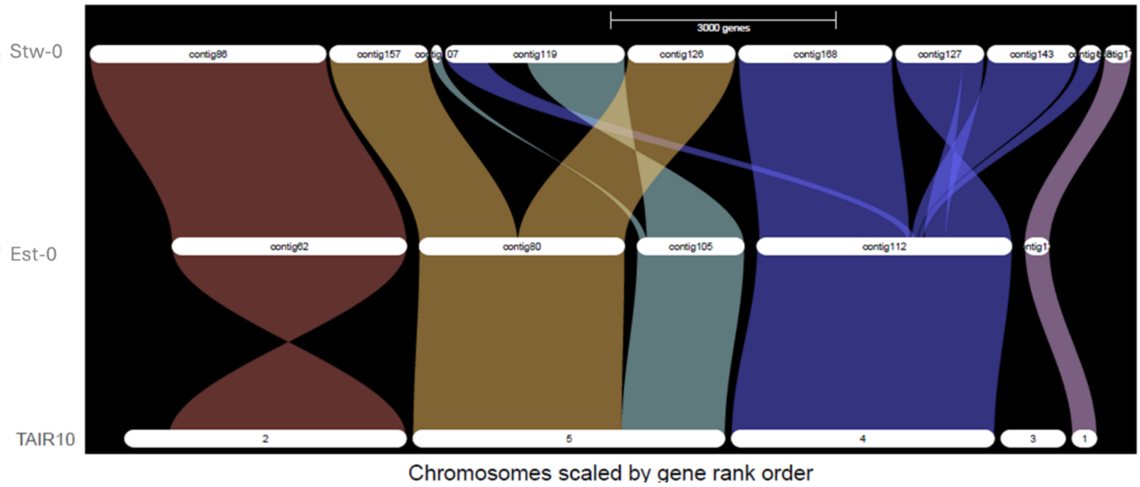


**Figure 5**. Riparian plot of the accessions Est-0 and Stw-0 phased to the reference genome (TAIR10). Each colour represents one synteny block.

## 4. Discussion

In comparing genome size estimation, is important to consider the discrepancies between different methods, like flow cytometry, *k-mer* analysis, and assembly-based methods. From previous studies [4] ,it is known that genome size estimations based on flow cytometry suffer from overestimation, ranging the *A. Thaliana* genome from 161 Mb to 184 Mb. Other estimates derived *k-mer* analysis defines the range from 138 Mb to 175 Mb [40]. In the case of the Est-0 genome, the *k-mer* analysis generates an estimation of 164 Mbp, within the expected range but larger than the estimation from the Flye assembly, whose length was approximately 135 Mbp. In addition, the *k-mer* analysis gives insight into the assembly, since in our case the estimations of low percentage of heterozygosity (0.001%), of low read error rate (0.147%) and expected coverage x24.

The comparison of the three assembly strategies: LJA, Hifiasm and Flye, resulted in the choice of Flye as the one that outperformed for this accession of *A. Thaliana*. However, it should be considered that there is no single best assembler, and it depends on the quality and coverage of the quality input data, and noticeably, in the heterozygosity and genome complexity of the specie. The *k-mer* analysis over the Est-0 accession revealed low percentages of heterozygosity, what leads to a more reliable assembly, consequently, facilitates the annotation step and leads to a higher-quality annotation.

Furthermore, the assembly evaluation should be based on the combination of the three metrics: contiguity, completeness and correctness. Often the best assembly is selected based on a higher NG50, instead of on the trade-off between contiguity and correctness. Aggressive assemblers will create longer contigs at the cost of correctness. The validation of the assembly is challenging when the reference genome is not available, as shown in Figure 2, the QUAST results without reference suggest that Hifiasm assembly would be better than LJA or Flye assembly.

An alternative strategy of assembly is to merge the three assemblies in one using quickmerge, as exposed in [16], with Canu, Flye and Hifiasm software. It relies on the assumption that different strategies would be better assembling different regions of the genome, however it increases the uncertainty. As it is shown in Table 1, in this case the merging strategy outperforms to the single use of Flye in terms of contiguity, achieving higher N50 and fewer number of contigs. This fact is reflected in the percentage of TE, it is expected a lower number of base-pairs covered by TEs in the Flye genome assembly than in the quickmerge genome assembly, since it is more fragmented and less TEs will be annotated.

Comparing the TE content across different accessions, there are not significant differences, pointing out that the *A. Thaliana* genome can be considered as a model, that over time has largely managed to eliminate transposon dynamics over time, maintaining conserved synteny blocks and overall genome stability. Despite its quasi-fixed karyotype as noted in [16], the adaptation of *A. Thaliana* to different environments should be reflected, and it has been shown that specially near the centromeres, the diversity increases, and large and abundant rearrangement occur, in contrast to the chromosome arms. In this study over the accession Est-0, the distribution of TEs across the genome (at scaffold level) support this hypothesis. There are contigs enriched in TE content, such as the contig 69, the contig 74, the contig 105, the contig 110 and the contig 126, whereas there are others contigs like the contig 1, the contig 61, the contig 62, the contig 80 or the contig 112 that exhibit low density in TEs (Supplementary 2). In this sense, the high density of TEs can be seen as an engine of diversity. It has been previously studied that CRM elements and Athila elements are clades that usually target centromeric and pericentromeric regions of the chromosome [41], [42]. Notably, the contigs 69, 74 and 110 where the clades CRM and Athila are present, indicating potential centromeric regions of the genome, there is an overlap with the presence of other TEs from the Gypsy, Copia, and CACTA superfamilies. It could suggest that these regions corresponding potentially to centromeres, may accumulate more diversity and higher density in TEs compared to the other regions with low density in TEs, that could be attributed to the chromosomal arms. Interestingly, the recent peaks detected in the TE age analysis came mainly from the Copia and Gypsy superfamilies, and with further study could be relevant if they correlate with centromeric regions, again supporting the hypothesis of the centromeres evolving and leading to different centromeric haplotypes. Furthermore, the riparian plots that report synteny conservation include the contigs 1, 62, 80 and 112 that showed low density in TEs (Supplementary 2).

In perspective to the future, further genomic investigation is needed to confirm that hypothesis, such as selecting and annotating carefully the contigs that presented high density in TEs, improving the assembly to reach the chromosomal level or further functional analysis to understand the impact of the possible rearrangements in the centromeric regions. In addition, studying the *A. Thaliana* from a pan-genome perspective with accession from all over the world would reveal more insights about the correlation between the genomic diversity and the phenotypic adaptation.

## References

[1] D. W. Meinke, J. M. Cherry, C. Dean, S. D. Rounsley, and M. Koornneef, 'Arabidopsis thaliana: A Model Plant for Genome Analysis', *Science*, vol. 282, no. 5389, pp. 662–682, Oct. 1998, doi: 10.1126/science.282.5389.662.

[2] The Arabidopsis Genome Initiative, 'Analysis of the genome sequence of the flowering plant Arabidopsis thaliana', *Nature*, vol. 408, no. 6814, pp. 796–815, Dec. 2000, doi: 10.1038/35048692.

[3] J. Cao *et al.*, 'Whole-genome sequencing of multiple Arabidopsis thaliana populations', *Nat. Genet.*, vol. 43, no. 10, pp. 956–963, Oct. 2011, doi: 10.1038/ng.911.

[4] Q. Long *et al.*, 'Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden', *Nat. Genet.*, vol. 45, no. 8, pp. 884–890, Aug. 2013, doi: 10.1038/ng.2678.

[5] K. Schneeberger *et al.*, 'Reference-guided assembly of four diverse Arabidopsis thaliana genomes', *Proc. Natl. Acad. Sci.*, vol. 108, no. 25, pp. 10249–10254, Jun. 2011, doi: 10.1073/pnas.1107739108.

[6] T. Hon *et al.*, 'Highly accurate long-read HiFi sequencing data for five complex genomes', *Sci. Data*, vol. 7, no. 1, p. 399, Nov. 2020, doi: 10.1038/s41597-020-00743-4.

[7] A. Bankevich, A. Bzikadze, M. Kolmogorov, D. Antipov, and P. A. Pevzner, 'LJA: Assembling Long and Accurate Reads Using Multiplex de Bruijn Graphs', *bioRxiv*, p. 2020.12.10.420448, Jan. 2021, doi: 10.1101/2020.12.10.420448.

[8] H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, and H. Li, 'Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm', *Nat. Methods*, vol. 18, no. 2, pp. 170–175, Feb. 2021, doi: 10.1038/s41592-020-01056-5.

[9] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, 'Assembly of long, error-prone reads using repeat graphs', *Nat. Biotechnol.*, vol. 37, no. 5, pp. 540–546, May 2019, doi: 10.1038/s41587-019-0072-8.

[10] M. Goel, H. Sun, W.-B. Jiao, and K. Schneeberger, 'SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies', *Genome Biol.*, vol. 20, no. 1, p. 277, Dec. 2019, doi: 10.1186/s13059-019-1911-0.

[11] C. Holt and M. Yandell, 'MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects', *BMC Bioinformatics*, vol. 12, no. 1, p. 491, Dec. 2011, doi: 10.1186/1471-2105-12-491.

[12] E. B. Chuong, N. C. Elde, and C. Feschotte, 'Regulatory activities of transposable elements: from conflicts to benefits', *Nat. Rev. Genet.*, vol. 18, no. 2, pp. 71–86, Feb. 2017, doi: 10.1038/nrg.2016.139.

[13] M. (viaf)118816478 Lynch, *The origins of genome architecture*. Sunderland (Mass.) : Sinauer associates, 2007. [Online]. Available: http://lib.ugent.be/catalog/rug01:001266846

[14] R. K. Slotkin and R. Martienssen, 'Transposable elements and the epigenetic regulation of the genome', *Nat. Rev. Genet.*, vol. 8, no. 4, pp. 272–285, Apr. 2007, doi: 10.1038/nrg2072.

[15] N. Buisine, H. Quesneville, and V. Colot, 'Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets', *Genomics*, vol. 91, no. 5, pp. 467–475, May 2008, doi: 10.1016/j.ygeno.2008.01.005.

[16] Q. Lian *et al.*, 'A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range', *Nat. Genet.*, vol. 56, no. 5, pp. 982–991, May 2024, doi: 10.1038/s41588-024-01715-9.

[17] W.-B. Jiao and K. Schneeberger, 'Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics', *Nat. Commun.*, vol. 11, no. 1, p. 989, Feb. 2020, doi: 10.1038/s41467-020-14779-y.

[18] Simon Andrews, *FastQC: a quality control tool for high throughput sequence data.* (2010). [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[19] S. Chen, 'Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp', *iMeta*, vol. 2, no. 2, p. e107, May 2023, doi: 10.1002/imt2.107.

[20] T. R. Ranallo-Benavidez, K. S. Jaron, and M. C. Schatz, 'GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes', *Nat. Commun.*, vol. 11, no. 1, p. 1432, Mar. 2020, doi: 10.1038/s41467-020-14998-3.

[21] M. G. Grabherr *et al.*, 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, Jul. 2011, doi: 10.1038/nbt.1883.

[22] M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov, 'BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes', *Mol. Biol. Evol.*, vol. 38, no. 10, pp. 4647–4654, Oct. 2021, doi: 10.1093/molbev/msab199.

[23] A. Mikheenko, V. Saveliev, P. Hirsch, and A. Gurevich, 'WebQUAST: online evaluation of genome assemblies', *Nucleic Acids Res.*, vol. 51, no. W1, pp. W601–W606, Jul. 2023, doi: 10.1093/nar/gkad406.

[24] A. Rhie, B. P. Walenz, S. Koren, and A. M. Phillippy, 'Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies', *Genome Biol.*, vol. 21, no. 1, p. 245, Sep. 2020, doi: 10.1186/s13059-020-02134-9.

[25] S. Kurtz *et al.*, 'Versatile and open software for comparing large genomes', *Genome Biol.*, vol. 5, no. 2, p. R12, Jan. 2004, doi: 10.1186/gb-2004-5-2-r12.

[26] Smit, AFA, Hubley, R & Green, P., *RepeatMasker Open-4.0.* (2015 2013). [Online]. Available: http://www.repeatmasker.org

[27] S. Kagale *et al.*, 'The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure', *Nat. Commun.*, vol. 5, no. 1, p. 3706, Apr. 2014, doi: 10.1038/ncomms4706.

[28] F. Sievers *et al.*, 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Mol. Syst. Biol.*, vol. 7, no. 1, p. 539, Jan. 2011, doi: 10.1038/msb.2011.75.

[29] M. N. Price, P. S. Dehal, and A. P. Arkin, 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments', *PLoS ONE*, vol. 5, no. 3, p. e9490, Mar. 2010, doi: 10.1371/journal.pone.0009490.

[30] I. Letunic and P. Bork, 'Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool', *Nucleic Acids Res.*, vol. 52, no. W1, pp. W78–W82, Jul. 2024, doi: 10.1093/nar/gkae268.

[31] P. Jones *et al.*, 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, May 2014, doi: 10.1093/bioinformatics/btu031.

[32] J. Mistry *et al.*, 'Pfam: The protein families database in 2021', *Nucleic Acids Res.*, vol. 49, no. D1, pp. D412–D419, Jan. 2021, doi: 10.1093/nar/gkaa913.

[33] The UniProt Consortium, 'UniProt: the Universal Protein Knowledgebase in 2025', *Nucleic Acids Res.*, p. gkae1010, Nov. 2024, doi: 10.1093/nar/gkae1010.

[34] Y. Nevers *et al.*, 'Quality assessment of gene repertoire annotations with OMArk', *Nat. Biotechnol.*, vol. 43, no. 1, pp. 124–133, Jan. 2025, doi: 10.1038/s41587-024-02147-w.

[35] A. M. Altenhoff *et al.*, 'OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem', *Nucleic Acids Res.*, vol. 52, no. D1, pp. D513–D521, Jan. 2024, doi: 10.1093/nar/gkad1020.

[36] H. Li, 'Protein-to-genome alignment with miniprot', *Bioinformatics*, vol. 39, no. 1, p. btad014, Jan. 2023, doi: 10.1093/bioinformatics/btad014.

[37] C. Diesh *et al.*, 'JBrowse 2: a modular genome browser with views of synteny and structural variation', *Genome Biol.*, vol. 24, no. 1, p. 74, Apr. 2023, doi: 10.1186/s13059-023-02914-z.

[38] J. T. Lovell *et al.*, 'GENESPACE tracks regions of interest and gene copy number variation across multiple genomes', *eLife*, vol. 11, p. e78526, Sep. 2022, doi: 10.7554/eLife.78526.

[39] D. M. Emms and S. Kelly, 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biol.*, vol. 20, no. 1, p. 238, Nov. 2019, doi: 10.1186/s13059-019-1832-y.

[40] H. Sun, J. Ding, M. Piednoël, and K. Schneeberger, 'findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies', *Bioinformatics*, vol. 34, no. 4, pp. 550–557, Feb. 2018, doi: 10.1093/bioinformatics/btx637.

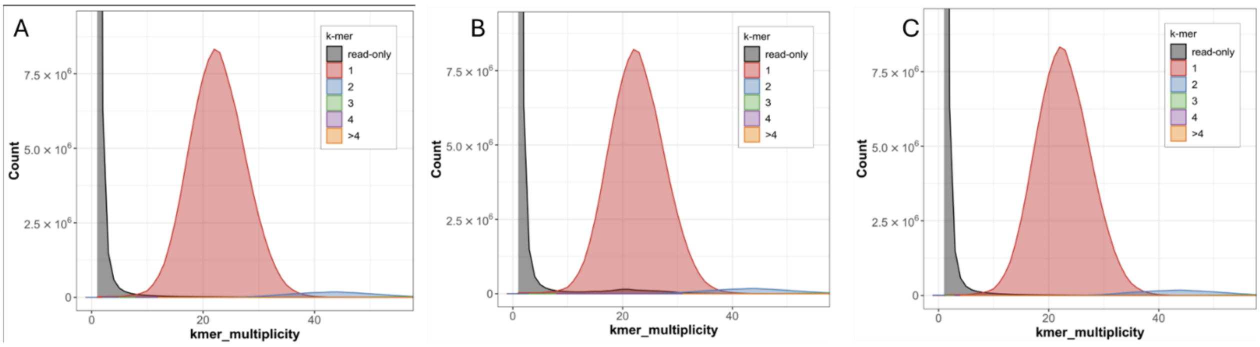
[41] A. Miura, M. Kato, K. Watanabe, A. Kawabe, H. Kotani, and T. Kakutani, 'Genomic localization of endogenous mobile CACTA family transposons in natural variants of Arabidopsis thaliana', *Mol. Genet. Genomics*, vol. 270, no. 6, pp. 524–532, Jan. 2004, doi: 10.1007/s00438-003-0943-y.

[42] V. Pereira, 'Insertion bias and purifying selection of retrotransposons in the Arabidopsis thalianagenome', *Genome Biol.*, vol. 5, no. 10, p. R79, Sep. 2004, doi: 10.1186/gb-2004-5-10-r79.

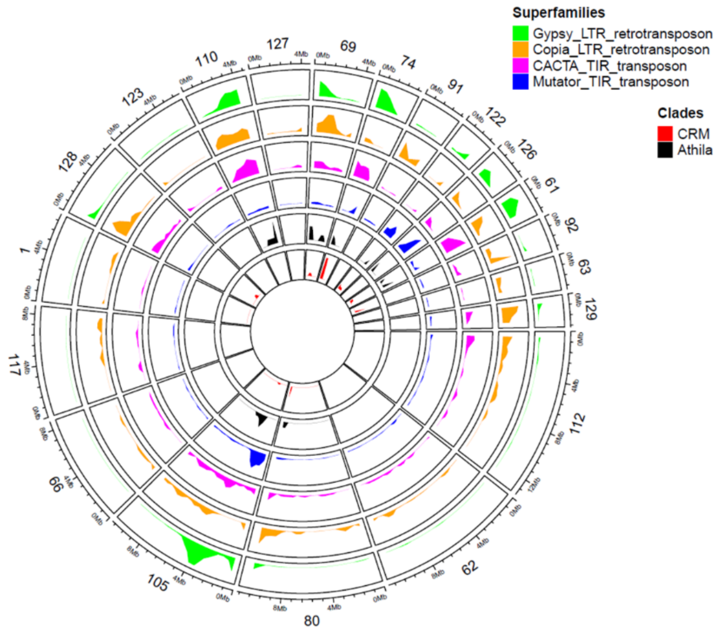
## Code link

https://github.com/lfercer/Assembly_Annotation_course

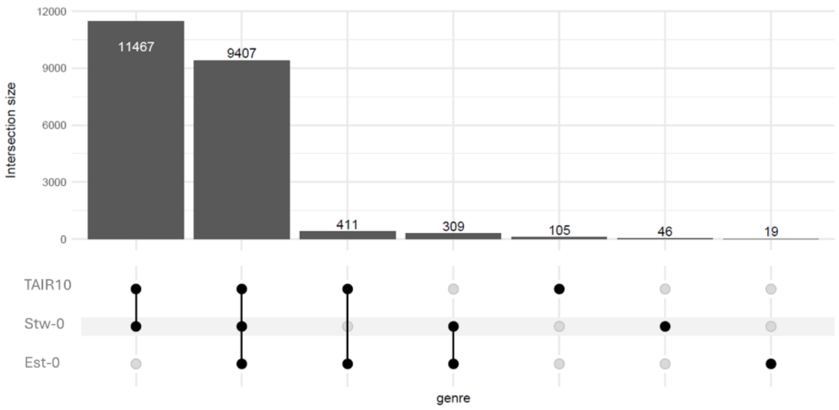
## Supplementary material



***Supplementary 1****. Merqury outputs performed over: A) Fly assembly, B) Hifiasm assembly, C) LJA assembly. The error rate is similar for the three genome assemblies, assuming that the k-mers only found in assembly are bp errors. The three assemblies present an expected sequencing coverage of 24, that coincides with the peak of homozygous k-mers.*



**Supplementary 2**. *Distribution of TEs across the contigs, in concrete the superfamilies: Gypsy, Copia, CACTA and Mutator, and the clades CRM and Athila. There are contigs enriched in TE content, such as the contig 110, the contig 69, the contig 74 and the contig 105, whereas others seem to have low density of TEs, like the contig 1, 62 or 117.*



***Supplementary 3.*** *UpSet plot representing the intersection sizes of orthogroups annotated in three different genome assemblies: TAIR10, Est-0, and Stw-0. The bar plot indicates the number of orthogroups in each intersection.*

**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 I have used the following AI technologies:

- Abstract - ChatGPT- Check the grammatical syntax and errors.
- Introduction – ChatGPT - Check the grammatical syntax and errors.
- Coding – ChatGPT – Debugging errors

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.