Assembly and annotation of Kar-1 *Arabidopsis thaliana* accession: Zooming in on Lian et al. (2024)

Abstract

The aim of this project was to assemble and annotate the Kar-1 accession of *Arabidopsis thaliana*, based on the study by Lian et al. ("A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range," 2024), and to replicate their results. Three genome assemblies were generated using FLye, LJA, and Hifiasm. These assemblies were of good quality, as assessed by QUAST and BUSCO results, with a slight preference for the Flye assembly, which showed the highest similarity in length to the assembly from Lian et al. However, for the annotation, the LJA assembly was used. Annotation of transposable elements (TEs) revealed several differences, particularly in the identification of Gypsy and LTR elements, likely due to assembly differences. Insertion events were found to be relatively recent, with TEs mostly located in centromeric regions, as expected. However, Mutators exhibited unusual behaviour, possibly due to assembly errors. Overall, the annotation quality was good, as indicated by BUSCO and OMArk results. The number of genes identified with a high-quality BLAST hit was lower than in the reference paper. In contrast, the number of orthogroups and genes within orthogroups was higher, likely due to the comparison with only two other genomes, including the St-0 accession, which may represent a mix of two accessions, skewing the results. The high number of contigs made it difficult to identify potential translocations or inversions, but no contigs contained material from multiple chromosomes. Overall, the results of this project reproduced those of Lian et al. (2024).

Introduction

Arabidopsis thaliana is a common model organism used in plant biology from the same family as the mustards, the family Brassicaceae. It was the first ever plant genome to be sequenced, in the year 2000 (Feldmann and Goff, 2014) and it has been used for many experimentations, making it like "lab rat" but for plant biology. Its key advantages are its small genome, approximately 135 Mbp, distributed across only five chromosomes, its haploid nature, and its rapid life cycle (Ensembl Plants, 2024). Additionally, its wide global distribution makes it an excellent candidate for studying the impact of environmental conditions on the genome structure.

Those characteristics made it perfect to learn how to assemble and annotate a genome, which is the goal of this report, and the two lectures associated to it, namely Genome and Transcriptome assembly & Organization and annotation of eukaryote genomes. To do that, *A. thaliana* sequencing was needed. By chance, on April 18, 2024, a paper was published, called "A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range" from Lian et al. (2024). This paper, referred to as the reference paper, analyses the genome structure of 69 accessions of *A. thaliana*. Its primary goals were to determine whether the genome structure is conserved across these accessions and overall to assess *A. thaliana* genome diversity. As the title indicates, the structure is indeed conserved, with counter-selection acting against chromosomal arm rearrangements but they also found structural variation in the centromeric regions explaining most of the variation in genome size between the accessions (Lian et al., 2024).

To conduct their study, Lian et al. assembled and annotated the genomes of 69 accessions, which are mostly from the Northern Hemisphere. In the report you are currently reading, one of those accessions, Kar-1, was used for assembly and annotation. The objective was to reproduce as well as possible the results presented in the reference paper. For that, starting from the raw sequencing reads, three assemblies were created with three different assemblers. Two of these assemblies were later annotated—one by me and one by E. Diethelm. Additionally, other *A. thaliana* accessions from the reference paper, such as St-0, were analysed by my colleagues whose work was used for comparison during this project.

Overall, the focus of this project was to learn how to assemble and annotate a genome, as well as understanding the outputs and evaluate their quality.

Materials and Methods

Data

The dataset used in this project is composed PacBio HiFi sequencing data from *Arabidopsis thaliana* (Kar-1 accession), collected in Kyrgyzstan, as described in the study by Lian et al. (2024). Additionally, RNA-seq data from *A. thaliana* (Sha accession), collected in Shadara, Tajikistan, was obtained from the study by Jiao et al. (2020). The RNA-seq dataset was generated using a combination of PacBio sequencing and Illumina whole-genome shotgun sequencing.

To assess the quality of the raw sequencing reads, FastQC v0.12.1 (Andrews, 2010) was used. The Kar-1 assession exhibited satisfactory quality, but the Sha assession required trimming, which was performed using fastp v0.23.4 (Chen et al., 2023) with default parameters. For the Kar-1 dataset, fastp was additionally run with trimming disabled to obtain detailed statistics on the reads.

Assembly

To do some estimation on the genome before the assembly, a k-mer analysis was made using Jellyfish v2.3.0 (Marçais & Kingsford, 2011) with default parameters to count the k-mers and create a histogram which was visualized using GenomeScope2.0 (Ranallo-Benavidez et al., 2020).

For the genome assembly, three assemblers were applied to the Kar-1 assession: Flye v2.9.5 (Kolmogorov et al., 2019), Hifiasm v0.19.8 (Cheng et al., 2021) and LJA v0.2 (Bankevich et al., 2021), all with default parameters. In the case of the Sha assession, transcriptome assembly was performed using Trinity v2.15.1 (Grabherr et al., 2011), also with default parameters.

The quality of the genome assemblies was assed using multiple tools. QUAST v5.2.0 (Gurevich et al., 2013) was used to obtain different metrics such as NG50 and total assembly length. Merqury v1.3 (Rhie et al., 2020) was used to calculate k-mer based statistics and BUSCO v5.4.2 (Manni et al. 2021) assessed the completeness of the assemblies. For the genome assemblies, BUSCO was run in genome mode with the "Brassicales" lineage dataset, whereas transcriptome mode was used for the Sha transcriptome assembly.

The three genome assemblies were aligned to the *A. thaliana* TAIR10 reference genome (The Arabidopsis Information Resource, 2011) using NUCmer from MUMmer4 (Marçais et al., 2018). The alignments were visualized using mummerplot, also from MUMmer4.

Annotation

Genome annotation was performed on the Kar-1 assembly generated by LJA. This process involved three main steps: the annotation and classification of transposable elements (TEs), gene annotation, and orthology based gene annotation quality control and comparative genomics.

The annotation of TEs was conducted using EDTA v2.2 (Ou et al., 2019) with the anno option enabled to ensure annotation of both intact and fragmented TEs. The coding sequences were included to minimize misclassification. TE annotations were visualized using R v4.4.2 (R Core Team, 2024) and RStudio v2024.09.1 (RStudio team, 2024), where the distribution of Long Terminal Repeats (LTRs) across clades, the percentage of the genome occupied by TE superfamilies, and genome-wide TE distributions were plotted. For the latter, the R package circlize (Gu et al., 2014) was employed, and scaffold lengths were calculated using SAMtools v1.13 (Danecek et al., 2021). TE classification was refined using TEsorter v1.3.0 (Zhang et al., 2022) on Copia and Gypsy sequences extracted with SeqKit2

(Shen et al., 2024). Insertion ages of TEs were estimated by parsing RepeatMasker outputs using the Perl script parseRM.pl (available at https://github.com/4ureliek/Parsing-RepeatMasker-Outputs), followed by visualization in R. A phylogenetic analysis of TEs was also performed using Clustal-Omega v1.2.4 (Sievers et al., 2011), FastTree v2.1.11 (Price et al., 2010) and iTOL (Letunic and Bork, 2024).

Gene annotation was made using the MAKER pipeline v3.01.03 (Cantarel et al., 2008), with parameters configured to infer gene predictions from ESTs and protein homology. The alternative splicing option was also enabled. Protein sequences were functionally annotated using InterProScan v5.70 (Jones et al., 2014) with the Pfam database, and annotations were filtered based on Annotation Edit Distance (AED < 0.5). The quality of annotations was further validated with BUSCO v5.4.2, and functional validation was performed through sequence homology searches against the Uniprot viridiplantae database using BLAST v2.15.0 (NCBI, 2023).

Orthology-based quality control of gene annotations was conducted using OMArk v0.3.0 (Nevers et al., 2024). Gene models were refined using Miniprot v0.13 (Li, 2023), which employed Hierarchical Orthologous Groups (HOGs). Comparative genomics analyses were performed using GENESPACE (Lovell et al., 2022), incorporating annotations from the St-0 accession provided by Léo Wütschert. Some outputs were visualized in R.

Results

Assembly

Basic reads statistics

For Kar-1 (DNA), the length of the reads is between 52 and 48019 bases, with a total of 6'155'100'885 bases, based on FastQC and fastp results. This number correspond to an estimated coverage depth of 45x, assuming the genome size of *A. thaliana* is 135 Mb (Ensembl Plants, 2024). The FastQC report indicates a satisfying quality for the reads.

For Sha (RNA), the length of the reads is 101bp. The FastQC report indicates inadequate quality in the raw reads. After the filtering with fastp, 4536518 bases were removed, which improved the quality of the dataset.

K-mer analysis

Figure 1 shows GenomeScope profile of Kar-1. The estimated coverage depth is around 32x, and the genome size is approximated to 126Mb. It also shows a low error rate (0.192%) and a low level of heterozygosity (between 0.08% and 0.11%).

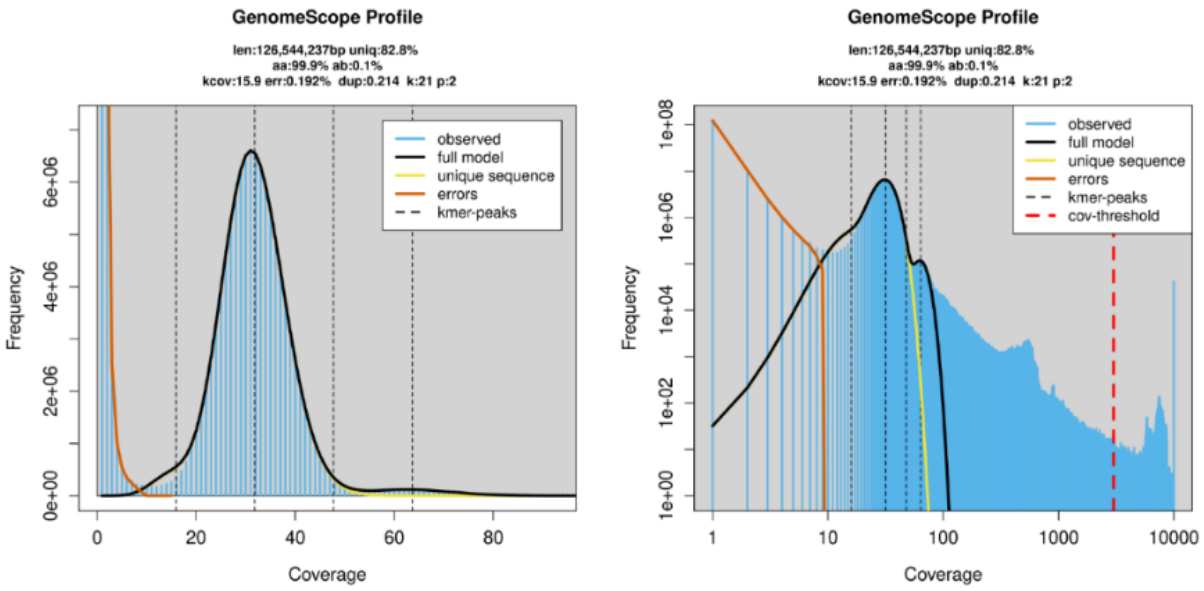

Figure 1: GenomeScope Profile for Kar-1 accession.

Assemblies' evaluation

The results from QUAST, as shown in Table 1 and Figure 2, indicate differences between the three assemblers in terms of the number of contigs, length of the assemblies, and metrics such as N50 and NG50.

Flye created the lowest number of contigs (148), while Hifiasm and LJA produced significantly more (490 and 500 respectively). Hifiasm assembled the longest assembly (155Mbp) which is around 20Mbp longer than the one assembled by Flye (134Mbp) and 15Mbp longer than the one assembly by LJA (140Mbp). The highest values for N50, NG50 and N90 are obtained by LJA, however Hifiasm has a higher NG90. The number of missassemblies is similar for Hifiasm and LJA, but lower for Flye.

In comparison, the assembly from the reference paper had 57 contigs, a length of 135061189bp and a N50 of 10478321. This assembly was made using different assemblers and merging the obtained assemblies together.

Table 1: Summary of QUAST result on the three genome assemblies.

	Flye	Hifiasm	LJA
Number of contigs	148	490	500
Total length (bp)	134 060 777	155 497 067	140 215 039
N50	5 870 153	9 242 573	10 481 239
NG50	6 461 146	9 899 701	10 979 033
N90	1 074 843	68 840	1 689 671
NG90	2 410 800	3 174 890	2 754 934
Number of missassemblies	3 626	4 839	4 937

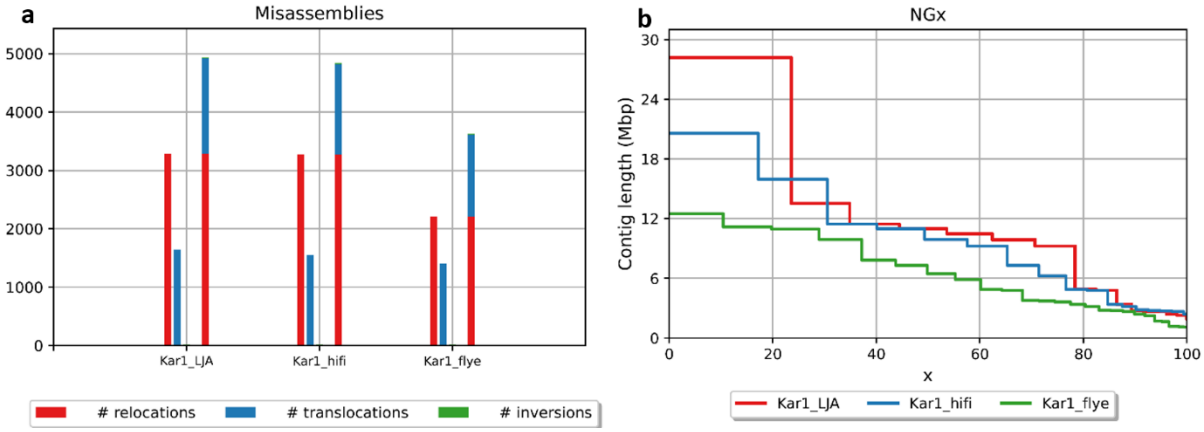

*Figure 2: QUAST results. **a**) Misassemblies in the genome assemblies, **b**) NGx of the three genome assemblies*

The Merqury results, Figure 3, show similar profile for the three assemblies, namely a peak around 35, representing the coverage depth in the assemblies, a low duplicity level, and read-only k-mers only present at low k-mer multiplicity, indicating sequencing error.

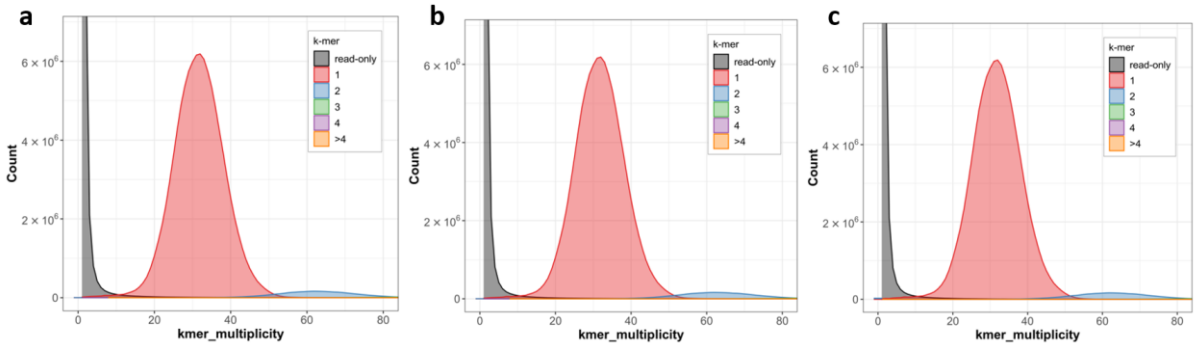

*Figure 3: Merqury results for Flye (**a**), Hifiasm (**b**) and LJA (**c**) genome assemblies.*

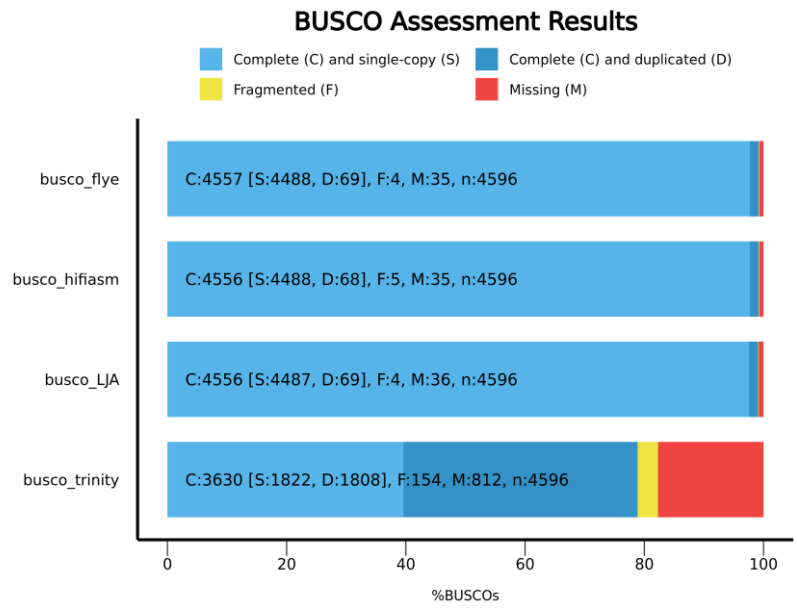

Figure 4: BUSCO results for the three genome assemblies and the transcriptome assembly.

The BUSCO results, shown in Figure 4, indicate similar completeness across the three genome assemblies. They all have high percentage of complete BUSCO (~99%), with only few differences, like the number of missing BUSCOs being higher for LJA (36) than for Flye and Hifiasm (35).

However, the transcriptome assembly (Trinity) shows significantly lower quality with a higher proportion of duplicated and missing BUSCOs, but also more fragmented BUSCOs.

Genomes comparison

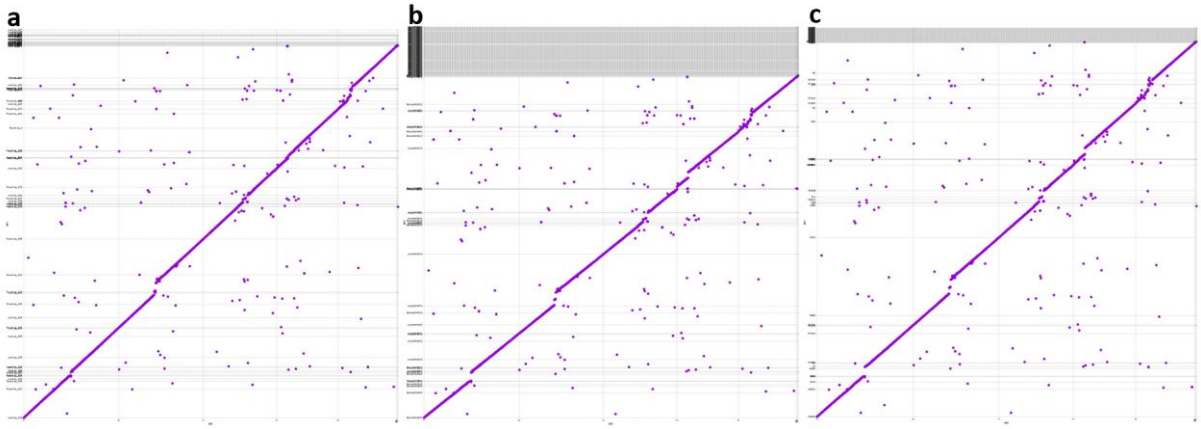

*Figure 5: Comparison of the assembled genomes from Flye (**a**), Hifiasm (**b**) and LJA (**c**) against the* Arabidopsis thaliana *reference. The reference is on the x axis.*

The three genome assemblies show good alignment with the reference genome across all chromosomes (Figure 5). The “jumps” in alignment are observed in every alignment around the middle of the chromosomes, which correspond probably to the centromeric regions. At the end of the Hifiasm assembly, there is something that is not present in the reference genome, showed by the grey bloc on top of Figure 5b.

Annotation

Transposable Elements

Based on EDTA summary, 15.31% of the genome is composed of repetitive elements. In comparison, in the reference paper this number is 16.22%. Figure 6 shows the differences in superfamilies of transposable elements (TEs) between the annotation made here (purple) and the one from the reference paper (orange). The biggest differences between both annotations are present in unknown long terminal repeats (LTR) and Gypsy (LTR).

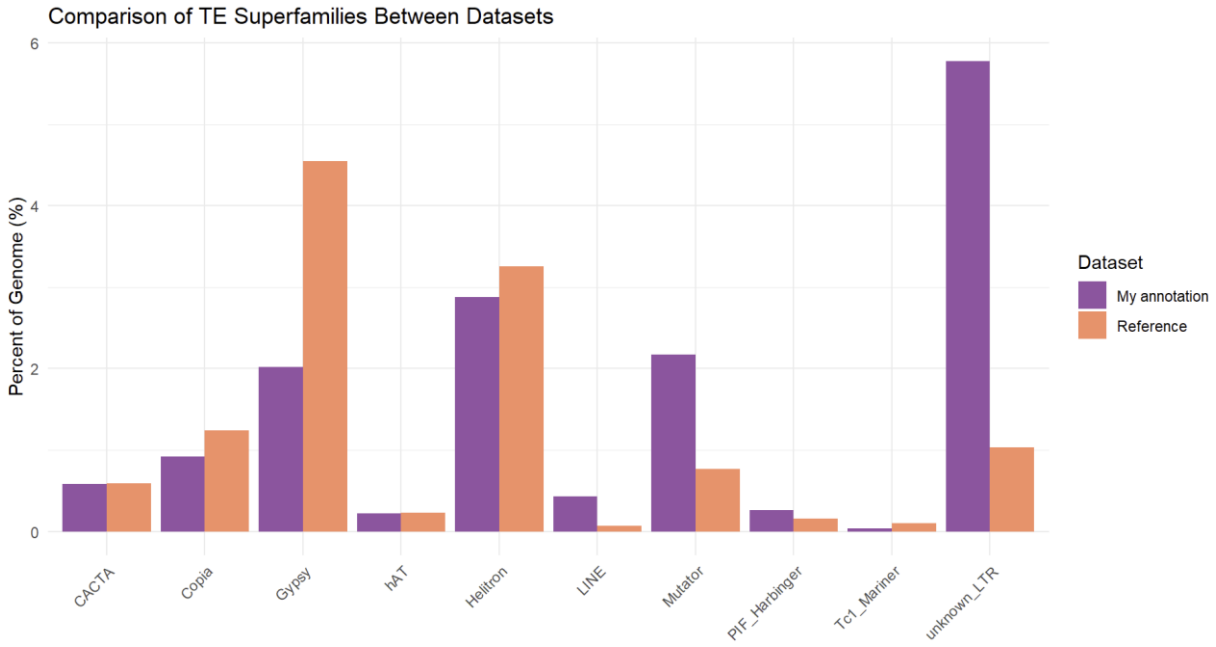

Figure 6: Comparison of TE superfamilies between the annotation made here (purple) and the annotation of the reference paper (orange) on Kar-1 accession

Figure 7 shows the identity percentage for some selected clades of Copia and Gypsy (Long Terminal Repeats (LTRs) retrotransposons (LTR-RTs)). Most of the count have a high percent identity (>95%), But few counts, for example in Reina, are present at lower percent identity. The clades not shown all had few counts but a high percent identity.

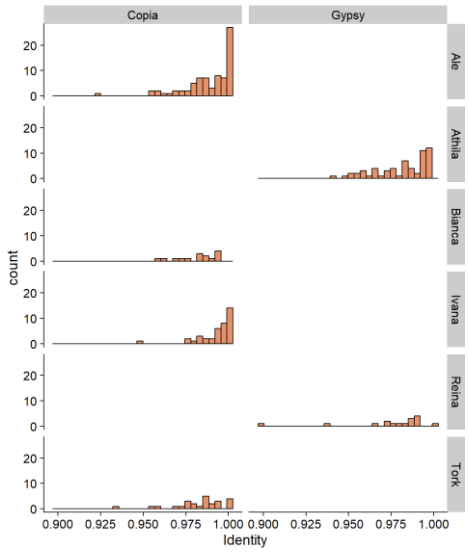

Figure 7: Number of LTR-RTs in chosen clades with their corresponding percent identity.

Figure 8 shows the distribution of some TE superfamilies across the genome. CRM and Athila clades are also present to indicate the centromeric regions. Gypsy are mostly present near those centromeric regions, in comparison with Copia that are present almost everywhere along the contigs. Mutator are almost not present expect for a wide peak in contig chr454. At this position, not other TEs are present.

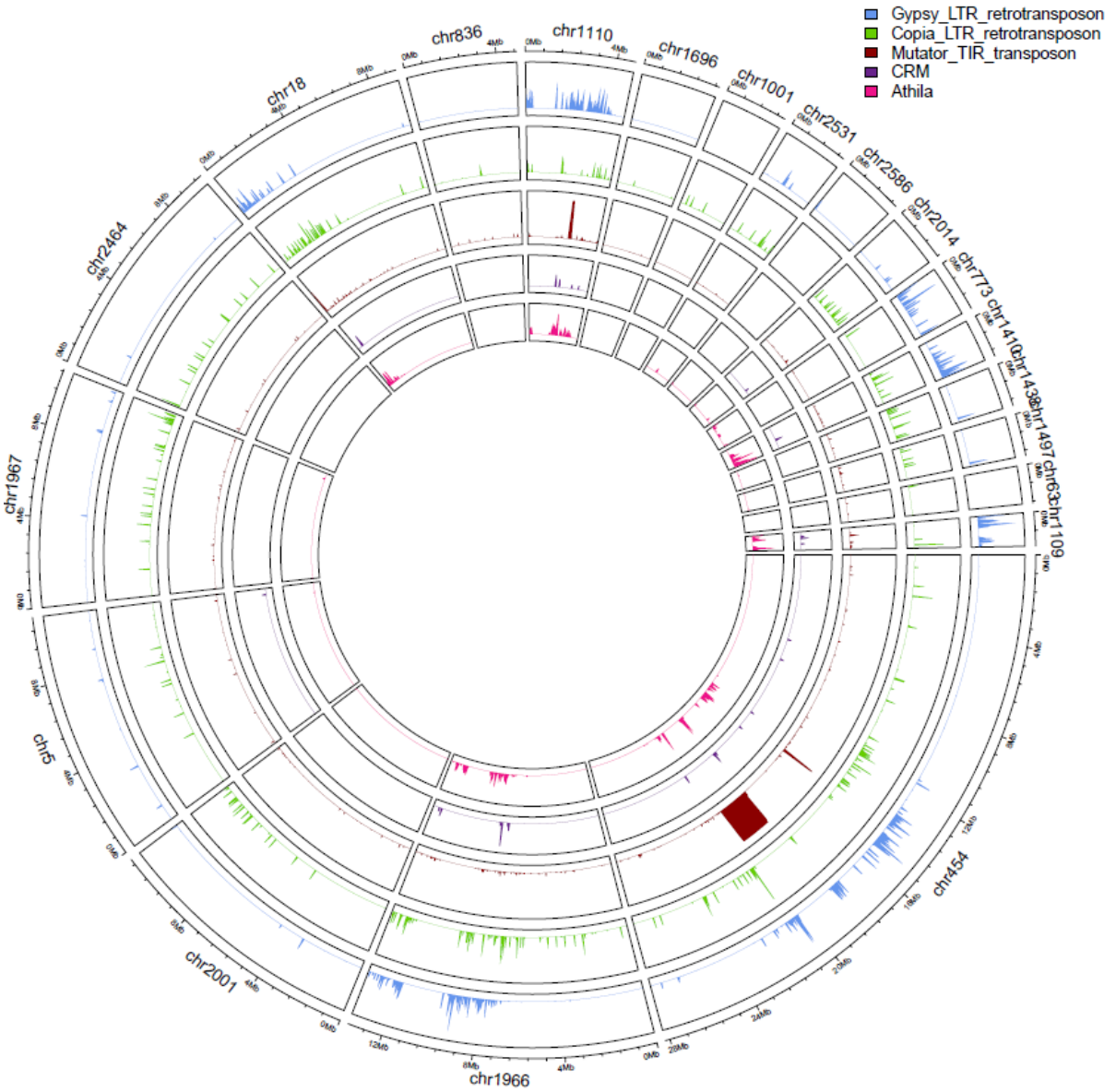

Figure 8: Distribution of three superfamilies, Gypsy, Copia and Mutator and two clades, CRM and Athila, across the genome.

The estimation of insertion Age for TEs is presented in Figure 9. Most of the activity is present at a small distance, indicating recent activity. MITE/DTM have the most recent activity. LTR/Gypsy has also a peak, a bit further. LTR/Copia seem to have more activity in the past, then stopped during Gypsy activity and more recently is being more active. For the other TE superfamilies, the activity seems to be constant.

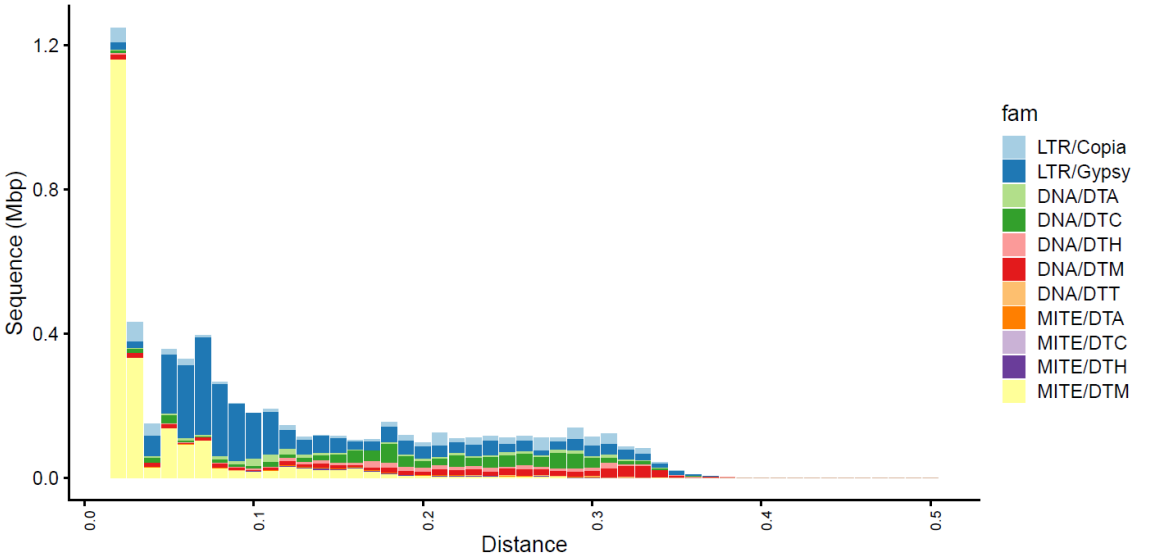

Figure 9: Estimation of Insertion Age of TE superfamilies.

Genome Annotation

The number of genes found before filtering is 28 821 and 28 206 after filtering. 23 244 of those genes have blast hits. In the reference paper, 26720 genes have high quality blast hits.

Figure 10a shows BUSCO results of the longest protein sequences and longest transcript sequences produced during the annotation. In both case a high fraction of BUSCO orthologs are fully present in the annotation (completeness) and single copy, indicating a high-quality annotation. There are also duplication and missing BUSCO orthologs, but it is less than 7% in both cases.

Figure 10b shows OMArk results. Like BUSCO, it shows that most of the genes are present in single copy, but there are still some duplicated and missing genes (top of the bar plot). The other half of the bar plot shows that most of the proteome consist of taxonomically consistent genes, with some being partially mapped or fragmented. Around 7% are genes with no detected homology.

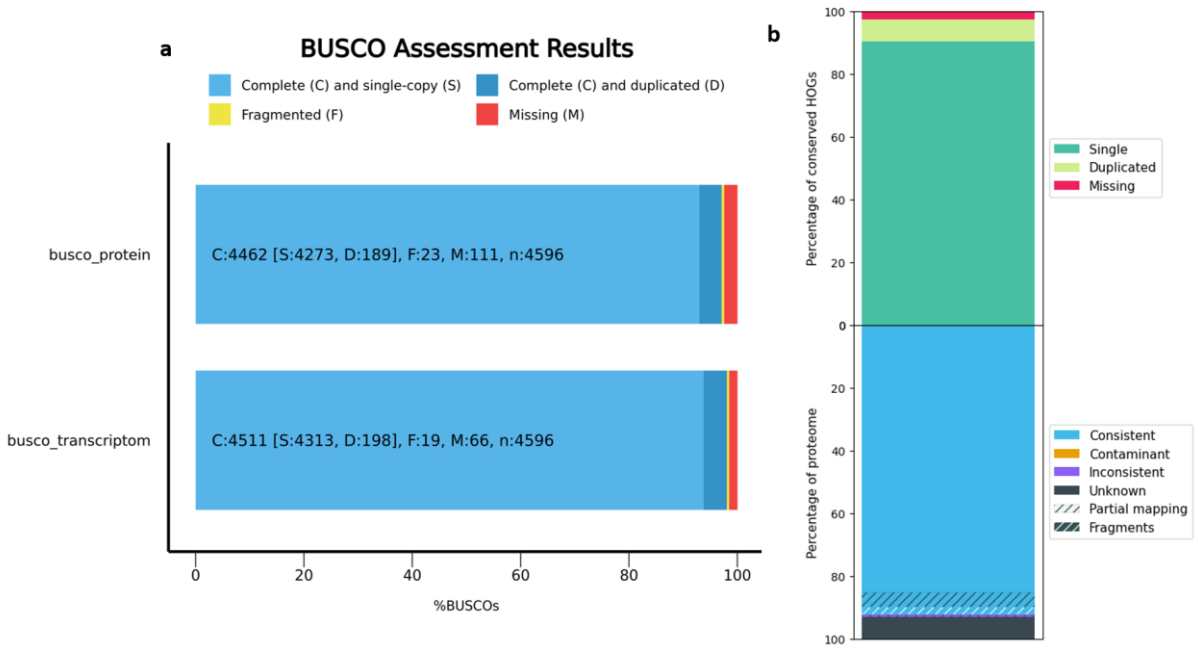

*Figure 10: Evaluation of the quality of the annotation using BUSCO (**a**) and OMArk (**b**)*

Figures 11 & 12 show a comparison between three genomes, Kar-1, St-0 and the reference genome TAIR10. In Figure 11, the orthogroups shared and unique between the three accessions are presented. In terms of genes in core orthogroups, 25304 are shared between the three genomes and 118 are specific to Kar-1. In the reference paper, Kar-1 as 24070 genes were shared with the other assessions, and 57 were unique to KAR-1.

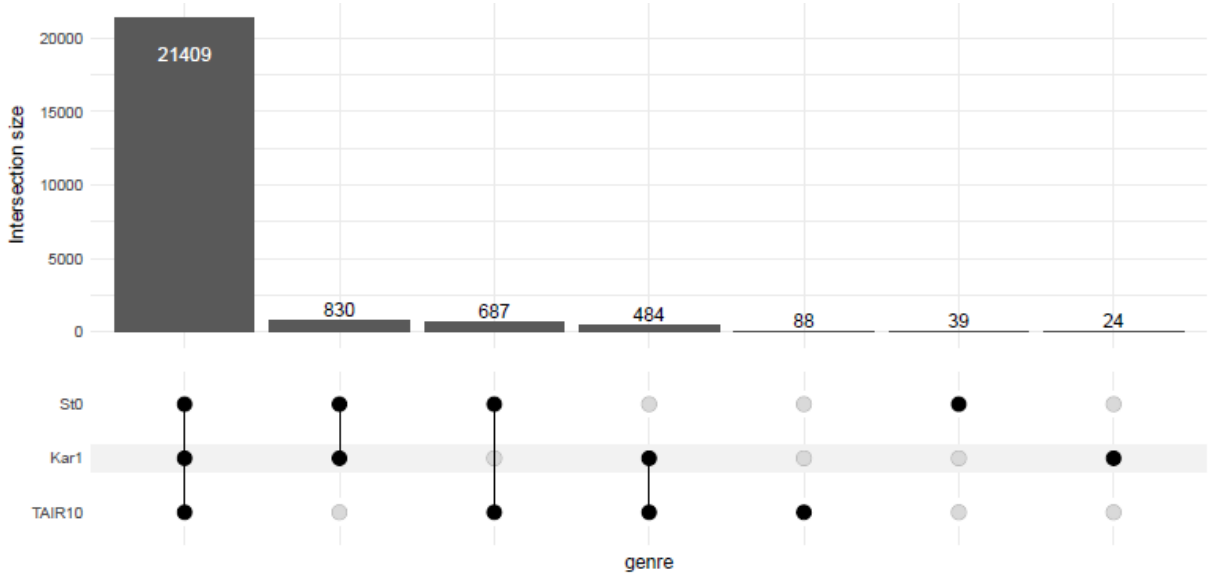

Figure 11: OrthoFinder results, showing the number of orthogroups shared or unique between Kar-1, St-0, and the reference genome TAIR10.

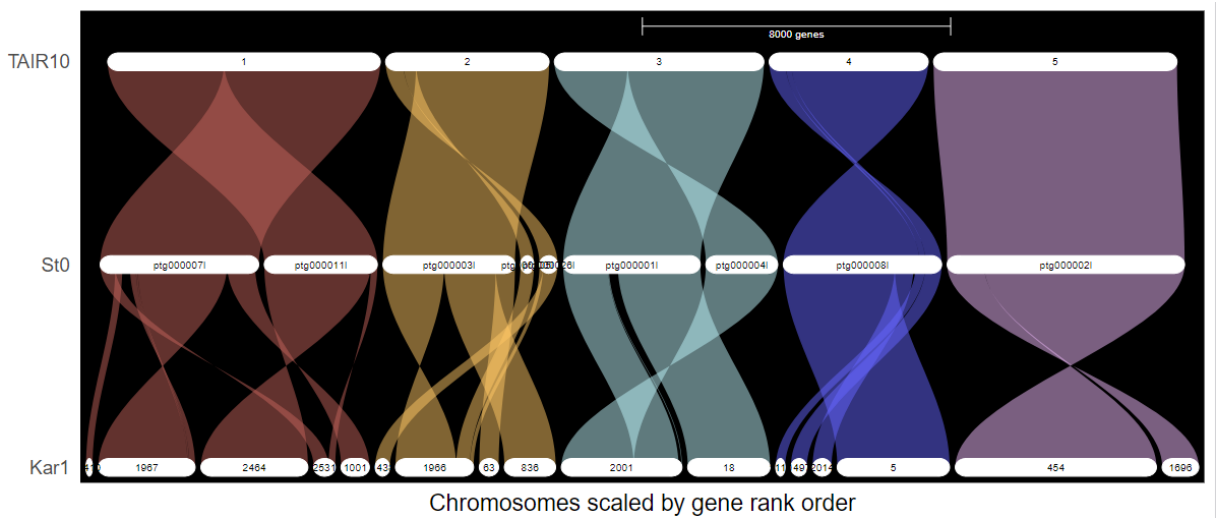

Figure 12: Riparian plot representing syntenic relationship between Kar-1, St-0, and the reference genome TAIR10.

Figure 12 shows syntenic relationships between TAIR10 (reference), St-0 and Kar-1. It seems to indicate a lot of movement between the three genomes. Many contigs since to be inversed or not in the correct order, compared to the reference genome. Such rearrangements are not present in the reference paper.

Discussion

Assembly

The three genome assemblies are of excellent quality. This was expected knowing that we had long and accurate reads (PacBio Hifi reads) and a good coverage (from Genomscope around 32).

However, the assemblers provided different assemblies with the same dataset. The QUAST results show it. For example, the length of the assemblies varies. This is because the three assemblers use different method. LJA uses Multiplex de Bruijn Graphs (Bankiev et al., 2020), Flye Repeat Graphs (Kolmogorov et al., 2019) and Hifiasm phased assembly graphs (Cheng et al., 2021). To evaluate the assemblies and determine which one is the best, three categories are important: the contiguity, the completeness, and the correctness.

The contiguity corresponds to the number of fragments the assembly consist of and how long they are. To evaluate that, the metrics NG50 and NG90 are used. They represent the sequence length of the shortest contig at 50% and 90% respectively of the estimated genome size. For NG50, LJA as the higher value, but for NG90, it is Hifiasm. On Figure 2b, it is clear that LJA has few longer contigs, but after approximatively NG25 is similar to Hifiasm. So, for the contiguity, LJA is probably the best assembly followed closely by Hifiasm and finally Flye.

The completeness is evaluated using the BUSCO score. It corresponds to the fraction of the genome that is assembled. In Figure 4, the BUSCO score for the three genome assemblies is really similar, with a complete single copy of more than 99% for all of them. Since LJA has one more missing BUSCO than the two others, for the completeness, Hifiasm and Flye are the best ones.

Finally, the correctness is the number of errors an assembly contains. It can be seen by looking at QUAST results, on Figure 2a. Flye as slightly less missassemblies than the two others. However, this metric depends on the reference genome and its quality. Since the reference used here is quite old (14 years old), it might not be too reliable. So for this one Flye is better, but should be treated with caution.

If we look at the genome comparison (Figure 5), they are all good with the centromeric region clearly visible, but the Hifiasm assembly exhibits something that is not present in the reference genome at the end of the assembly (the grey part). LJA and Flye assemblies also have something that is not present in the reference at the end of the assembly, but it is way thicker for Hifiasm. It means that this assembler does something at the end of the assembly that other do not do. It is hard to tell if it is something missing from the reference or added by the assembler. Since it is less significantly present in the other assemblies, it will be treated as a malus for Hifiasm assembly.

Based on all that, the best assembly is probably the Flye one, followed by the LJA and finally the Hifiasm one, but overall, they were all good. The assembly from the paper seems to be closer to the Flye one, at least for the length. This might be because in the paper, they used three different assemblers, including Hifiasm and Flye and merged them based on the best of them, so if the Flye was also their best, it might explain the similarity in length.

Regarding the transcriptome, the BUSCO score (Figure 4) is not as high as for the genome assemblies, showing a considerable number of duplicated and missing BUSCOs. This outcome is actually expected, as transcriptome sequencing captures only the genes that are actively expressed at the time of sampling. Therefore, missing, and duplicated BUSCOs are normal since not all genes are expressed under all conditions, and some genes might be expressed multiple times.

Annotation

The annotation results presented below were made using the LJA assembly. The Flye assembly was annotated by Etienne Diethelm.

Regarding the annotation of transposable elements (TEs), it is important to note that the high proportion of helitrons is likely inaccurate because EDTA is known to falsely identify them. A comparison with the reference paper (Figure 6) reveals that the annotation made here detected significantly more unknown LTRs but fewer Gypsy and Copia elements. Since both Gypsy and Copia belong to the LTR category, it is possible that EDTA identified them as LTRs but failed to classify them further. Additionally, the annotation made here shows a high fraction of Mutator elements, which will be addressed later.

From Figure 7, we can see that the insertions of most of the LTR is recent, because the percentage identity is high. Only Reina and Ale seem to have ancient insertion events. Those results are consistent with the result from Figure 9 that shows that many Gypsy and Copia insertion event are recent, but they also have older insertion events in a smaller proportion.

Figure 8 highlights the presence of TEs near the centromeres, which aligns with expectations. The low recombination rate in centromeric regions allows TEs to persist indefinitely once inserted. On this same plot, there is this wide and high peak of Mutators which is likely responsible for the higher fraction of Mutators compared to the reference paper. Such pattern is not expected. Indeed, it would mean that a high proportion of a contig is only composed of Mutators. This anomaly might indicate an error during the assembly.

The number of genes identified after filtering is slightly lower than the number reported in the reference paper, which might be caused by differences in assembly and annotation methods. Overall, the quality of the annotation was satisfying, according to Figure 10. It is important to note that the annotation was performed using the transcriptome for the Sha accession, rather than the Kar-1 one. It might explain why certain genes are missing, as Sha and Kar-1 likely do not share identical gene sets.

The number of genes within core orthogroups is approximately one thousand higher than reported in the reference paper. Initially, this was thought to be due to the genetic proximity of the two accessions. However, after verification, St-0 belongs to the European cluster, while Kar-1 is from the Asian cluster. Another possible explanation is that the St-0 accession may not only represent St-0. Its observed heterozygosity and unusually high gene count suggest it could be a fusion of two accessions (cf. Léo Wütschert results). Additionally, since we compared only three genomes, it is expected that the gene count would be higher than in the reference paper, which analysed 69 genomes. This explanation also accounts for why the number of private genes for Kar-1 is higher compared to the reference paper.

Figure 12 reveals significant movement between the three genomes, but this movement always follows contig boundaries. Importantly, there is no evidence of translocations where a single contig contains material from two different chromosomes, the movements might be only assembly artefacts. If the riparian plot were untangled, the result would likely show five distinct colour blocks, supporting the conclusion of the reference paper that there are no major structural rearrangements. Less and longer contigs would be needed to confirm that.

To conclude this report, the tree assemblies and genome annotation of *A. thaliana* were of high quality, allowing me to achieve results that align with those from the reference paper.

References

Andrew, S. “Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data.” Accessed December 5, 2024. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

“Arabidopsis_thaliana - Ensembl Genomes 60,” October 2024. https://plants.ensembl.org/Arabidopsis_thaliana/Info/Index.

Bankevich, Anton, Andrey Bzikadze, Mikhail Kolmogorov, Dmitry Antipov, and Pavel A. Pevzner. “LJA: Assembling Long and Accurate Reads Using Multiplex de Bruijn Graphs.” *bioRxiv*, January 1, 2021, 2020.12.10.420448. https://doi.org/10.1101/2020.12.10.420448.

Cantarel, Brandi L., Ian Korf, Sofia M. C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. “MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes.” *Genome Research* 18, no. 1 (January 1, 2008): 188–96. https://doi.org/10.1101/gr.6743907.

Chen, Shifu. “Ultrafast One-Pass FASTQ Data Preprocessing, Quality Control, and Deduplication Using Fastp.” *iMeta* 2, no. 2 (2023): e107. https://doi.org/10.1002/imt2.107.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. “Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm.” *Nature Methods* 18, no. 2 (February 2021): 170–75. https://doi.org/10.1038/s41592-020-01056-5.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10, no. 2 (February 1, 2021): giab008. https://doi.org/10.1093/gigascience/giab008.

Feldmann, Kenneth A., and Stephen A. Goff. “The First Plant Genome Sequence—*Arabidopsis Thaliana*.” In *Advances in Botanical Research*, edited by Andrew H. Paterson, 69:91–117. Genomes of Herbaceous Land Plants. Academic Press, 2014. https://doi.org/10.1016/B978-0-12-417163-3.00004-4.

“Genome Assembly - Arabidopsis Community - Confluence.” Accessed January 3, 2025. https://phoenixbioinformatics.atlassian.net/wiki/spaces/COM/pages/42216252/Genome+Assembly.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. “Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data.” *Nature Biotechnology* 29, no. 7 (May 15, 2011): 644–52. https://doi.org/10.1038/nbt.1883.

Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. “Circlize Implements and Enhances Circular Visualization in R.” *Bioinformatics* 30, no. 19 (October 1, 2014): 2811–12. https://doi.org/10.1093/bioinformatics/btu393.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics (Oxford, England)* 29, no. 8 (April 15, 2013): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Jiao, Wen-Biao, and Korbinian Schneeberger. “Chromosome-Level Assemblies of Multiple Arabidopsis Genomes Reveal Hotspots of Rearrangements with Altered Evolutionary Dynamics.” *Nature Communications* 11, no. 1 (February 20, 2020): 989. https://doi.org/10.1038/s41467-020-14779-y.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30, no. 9 (May 1, 2014): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. “Assembly of Long, Error-Prone Reads Using Repeat Graphs.” *Nature Biotechnology* 37, no. 5 (May 2019): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Letunic, Ivica, and Peer Bork. “Interactive Tree of Life (iTOL) v6: Recent Updates to the Phylogenetic Tree Display and Annotation Tool.” *Nucleic Acids Research* 52, no. W1 (July 5, 2024): W78–82. https://doi.org/10.1093/nar/gkae268.

Li, Heng. “Protein-to-Genome Alignment with Miniprot.” *Bioinformatics (Oxford, England)* 39, no. 1 (January 1, 2023): btad014. https://doi.org/10.1093/bioinformatics/btad014.

Lian, Qichao, Bruno Huettel, Birgit Walkemeier, Baptiste Mayjonade, Céline Lopez-Roques, Lisa Gil, Fabrice Roux, Korbinian Schneeberger, and Raphael Mercier. “A Pan-Genome of 69 Arabidopsis Thaliana Accessions Reveals a Conserved Genome Structure throughout the Global Species Range.” *Nature Genetics* 56, no. 5 (May 2024): 982–91. https://doi.org/10.1038/s41588-024-01715-9.

Lovell, John T, Avinash Sreedasyam, M Eric Schranz, Melissa Wilson, Joseph W Carlson, Alex Harkess, David Emms, David M Goodstein, and Jeremy Schmutz. “GENESPACE Tracks Regions of Interest and Gene Copy Number Variation across Multiple Genomes.” Edited by Detlef Weigel. *eLife* 11 (September 9, 2022): e78526. https://doi.org/10.7554/eLife.78526.

Madeira, Fábio, Nandana Madhusoodanan, Joonheung Lee, Alberto Eusebi, Ania Niewielska, Adrian R N Tivey, Rodrigo Lopez, and Sarah Butcher. “The EMBL-EBI Job Dispatcher Sequence Analysis Tools Framework in 2024.” *Nucleic Acids Research* 52, no. W1 (July 1, 2024): W521–25. https://doi.org/10.1093/nar/gkae241.

Manni, Mosè, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular Biology and Evolution* 38, no. 10 (October 1, 2021): 4647–54. https://doi.org/10.1093/molbev/msab199.

Marçais, Guillaume, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. “MUMmer4: A Fast and Versatile Genome Alignment System.” *PLOS Computational Biology* 14, no. 1 (January 26, 2018): e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Marçais, Guillaume, and Carl Kingsford. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of *k* -Mers.” *Bioinformatics* 27, no. 6 (March 15, 2011): 764–70. https://doi.org/10.1093/bioinformatics/btr011.

Nevers, Yannis, Alex Warwick Vesztrocy, Victor Rossier, Clément-Marie Train, Adrian Altenhoff, Christophe Dessimoz, and Natasha M. Glover. “Quality Assessment of Gene Repertoire Annotations with OMArk.” *Nature Biotechnology*, February 21, 2024, 1–10. https://doi.org/10.1038/s41587-024-02147-w.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. “Benchmarking Transposable Element Annotation Methods for Creation of a

Streamlined, Comprehensive Pipeline." *Genome Biology* 20, no. 1 (December 16, 2019): 275. https://doi.org/10.1186/s13059-019-1905-y.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." *PLOS ONE* 5, no. 3 (March 10, 2010): e9490. https://doi.org/10.1371/journal.pone.0009490.

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Ranallo-Benavidez, T. Rhyker, Kamil S. Jaron, and Michael C. Schatz. "GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes." *Nature Communications* 11, no. 1 (March 18, 2020): 1432. https://doi.org/10.1038/s41467-020-14998-3.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21, no. 1 (September 14, 2020): 245. https://doi.org/10.1186/s13059-020-02134-9.

RStudio Team (2024). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. URL: https://posit.co/.

Shen, Wei, Botond Sipos, and Liuyang Zhao. "SeqKit2: A Swiss Army Knife for Sequence and Alignment Processing." *iMeta* 3, no. 3 (2024): e191. https://doi.org/10.1002/imt2.191.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (October 11, 2011): 539. https://doi.org/10.1038/msb.2011.75.

Staff, NCBI. "Faster and Focused Searches with BLAST+ 2.15.0." NCBI Insights, November 21, 2023. https://ncbiinsights.ncbi.nlm.nih.gov/2023/11/21/faster-focused-searches-blast-2-15/.

The Arabidopsis Genome Initiative. "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana." *Nature* 408, no. 6814 (December 2000): 796–815. https://doi.org/10.1038/35048692.

"The Bioperl Toolkit: Perl Modules for the Life Sciences." Accessed December 5, 2024. https://www.researchgate.net/publication/11091711_The_Bioperl_Toolkit_Perl_Modules_for_the_Life_Sciences.

"Viridiplantae | Taxonomy | UniProt." Accessed December 5, 2024. https://www.uniprot.org/taxonomy/33090.

Zhang, Ren-Gang, Guang-Yuan Li, Xiao-Ling Wang, Jacques Dainat, Zhao-Xuan Wang, Shujun Ou, and Yongpeng Ma. "TEsorter: An Accurate and Fast Method to Classify LTR-Retrotransposons in Plant Genomes." *Horticulture Research* 9 (January 5, 2022): uhac017. https://doi.org/10.1093/hr/uhac017.

Link to GitHub

https://github.com/lieselty/assembly-annotation-course

Appendix

Phylogenetic trees of Copia and Gypsy superfamilies.

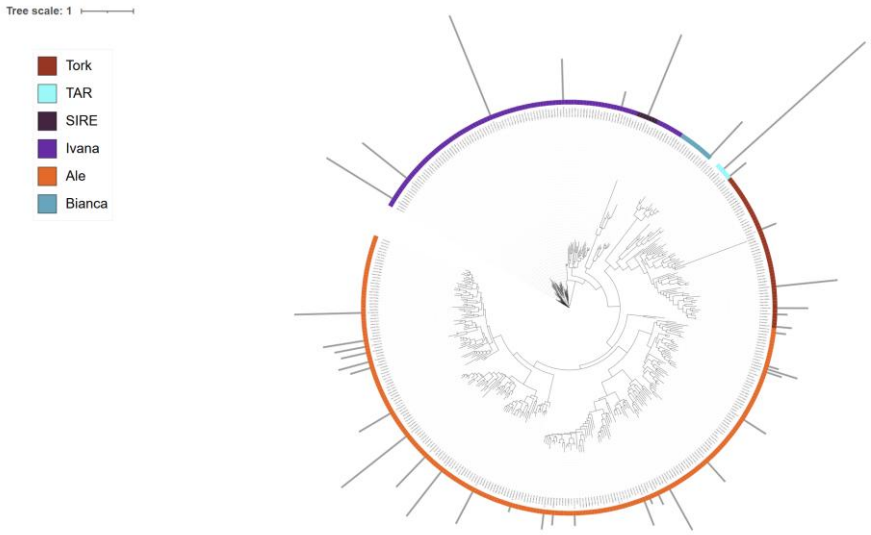

Figure 13: Copia's phylogenetic tree.

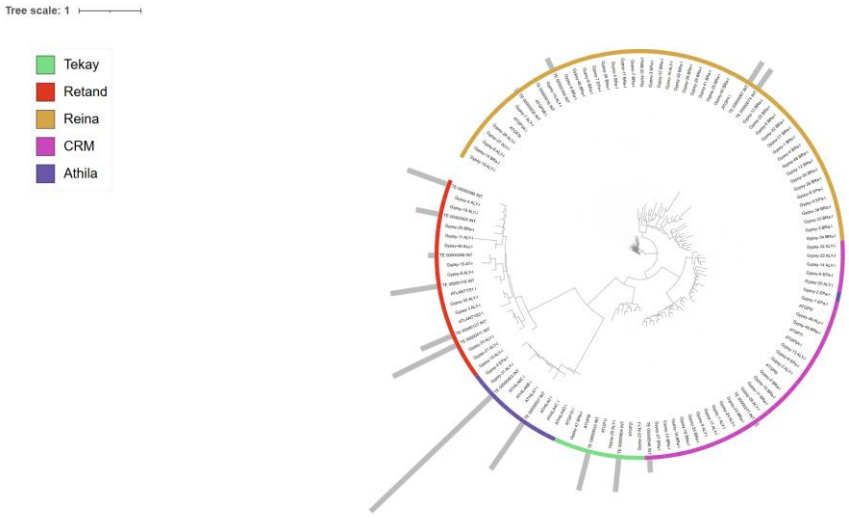

Figure 14: Gypsy's phylogenetic tree.

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 used the following AI technologies: Introduction, Material and Methods & Discussion – Chat GPT-4o – to improve writing (word choice and syntax).

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.