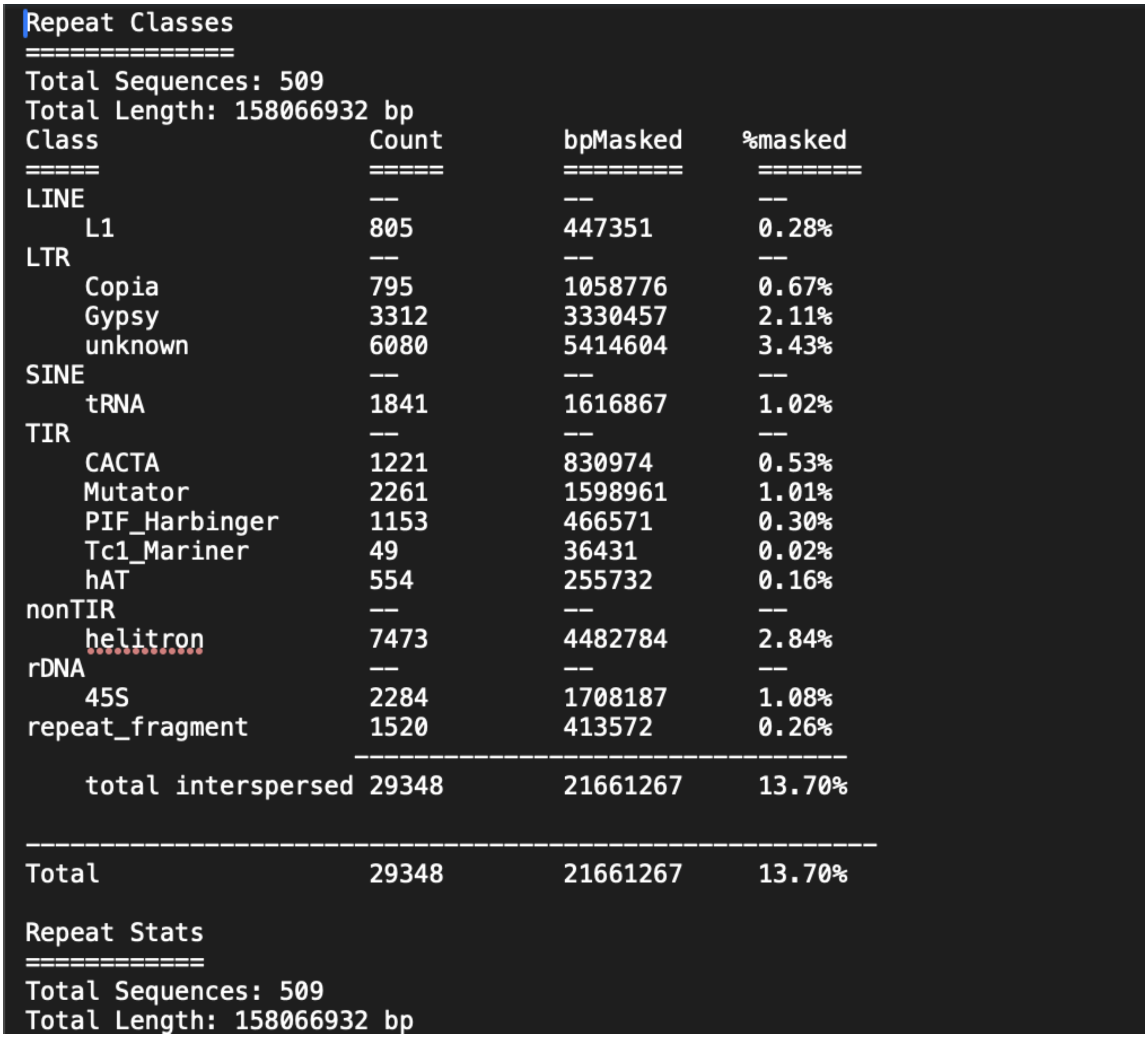
Organization and Annotation of Eukaryote Genomes

Yuwei Liu Student ID: 24-100-430
Accession: Kas-1
GitHub: `https://github.com/Saul-Gooodman/Organization-and-annotation-of-eukaryote-genomes`

November 19, 2025



```
Repeat Classes
==============
Total Sequences: 509
Total Length: 158066932 bp
Class                  Count        bpMasked    %masked
=====                  =====        ========     =======
LINE                   --           --           --
    L1                 805          447351       0.28%
LTR                    --           --           --
    Copia              795          1058776      0.67%
    Gypsy              3312         3330457      2.11%
    unknown            6080         5414604      3.43%
SINE                   --           --           --
    tRNA               1841         1616867      1.02%
TIR                    --           --           --
    CACTA              1221         830974       0.53%
    Mutator            2261         1598961      1.01%
    PIF_Harbinger      1153         466571       0.30%
    Tc1_Mariner        49           36431        0.02%
    hAT                554          255732       0.16%
nonTIR                 --           --           --
    helitron           7473         4482784      2.84%
rDNA                   --           --           --
    45S                2284         1708187      1.08%
repeat_fragment        1520         413572       0.26%
                      ---------------------------------
    total interspersed 29348        21661267     13.70%

---------------------------------------------------------
Total                  29348        21661267     13.70%

Repeat Stats
============
Total Sequences: 509
Total Length: 158066932 bp
```

Figure 1: Transposable element (TE) class composition and genome coverage in the Kas-1 genome based on EDTA annotation. Each bar represents the fraction of the assembly covered by a given TE class; the relative bar heights indicate which classes dominate the repeat landscape. Kas-1 shows a moderate repeat load dominated by LTR retrotransposons, which is typical for an *Arabidopsis thaliana* accession.

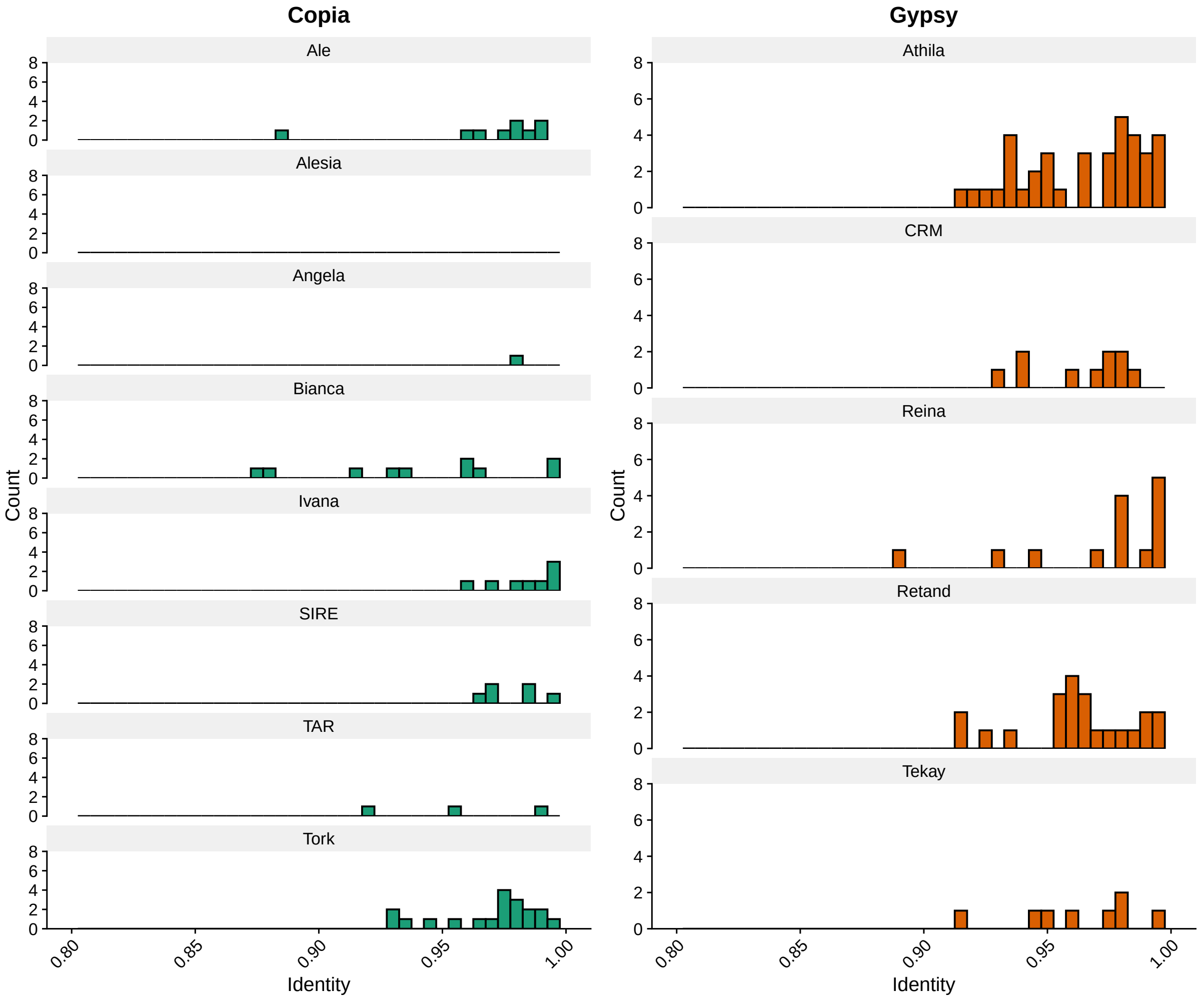


Figure 2: Percent-identity distributions between the two LTRs of full-length Copia and Gypsy elements in Kas-1. The x-axis shows LTR–LTR identity (younger insertions to the right), and the y-axis shows the number of elements; peaks indicate periods of intense activity. Kas-1 exhibits strong peaks at high identity values for both Copia and Gypsy families, consistent with relatively recent LTR retrotransposon bursts.

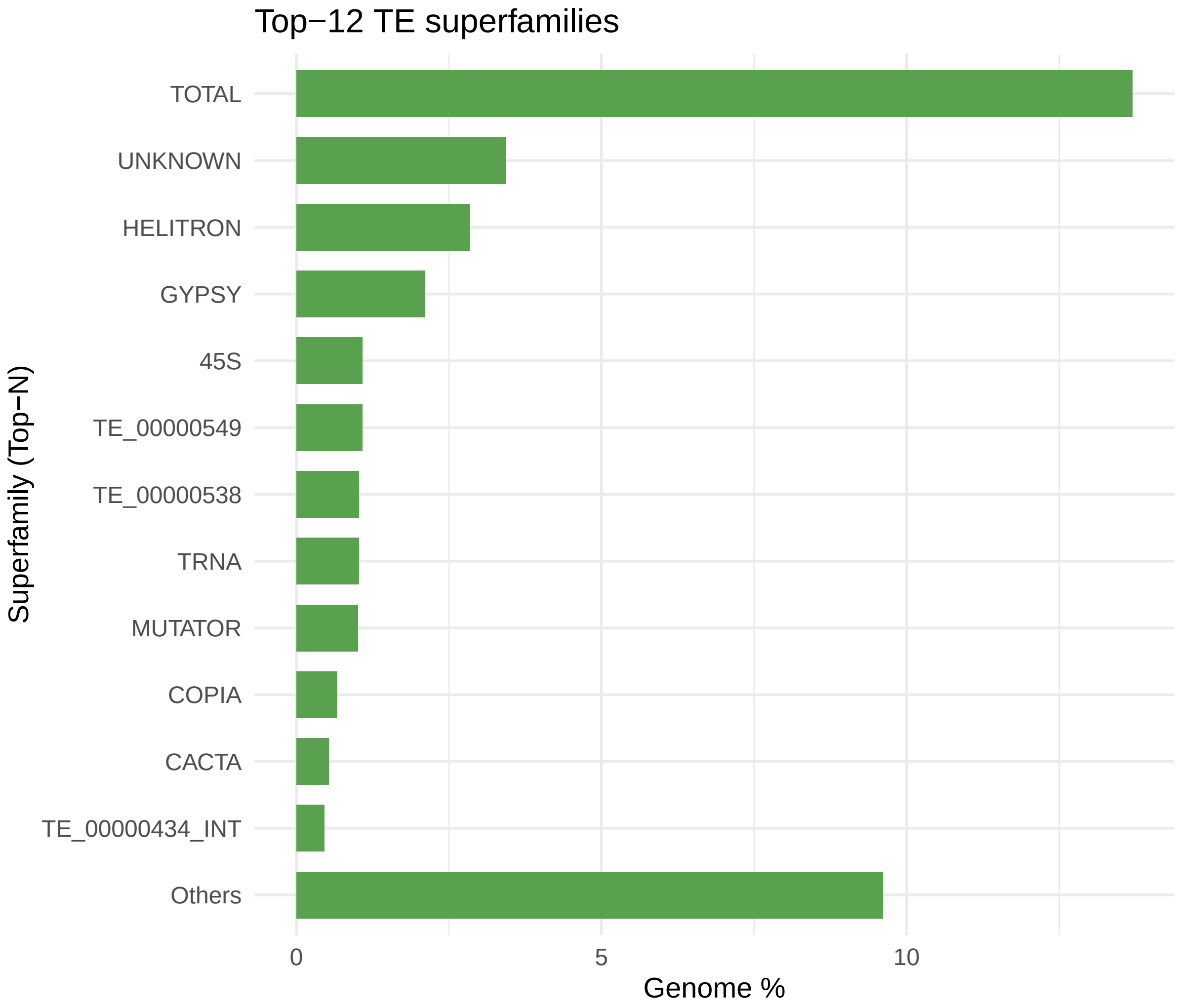


Figure 3: Percent genome coverage of the most abundant TE superfamilies in the Kas-1 assembly. Each bar corresponds to the fraction of total base pairs contributed by one superfamily, allowing direct comparison of their relative abundance. Gypsy and Copia LTR retrotransposons contribute most of the repetitive content, indicating that LTR-RTs are the main drivers of genome expansion in Kas-1.

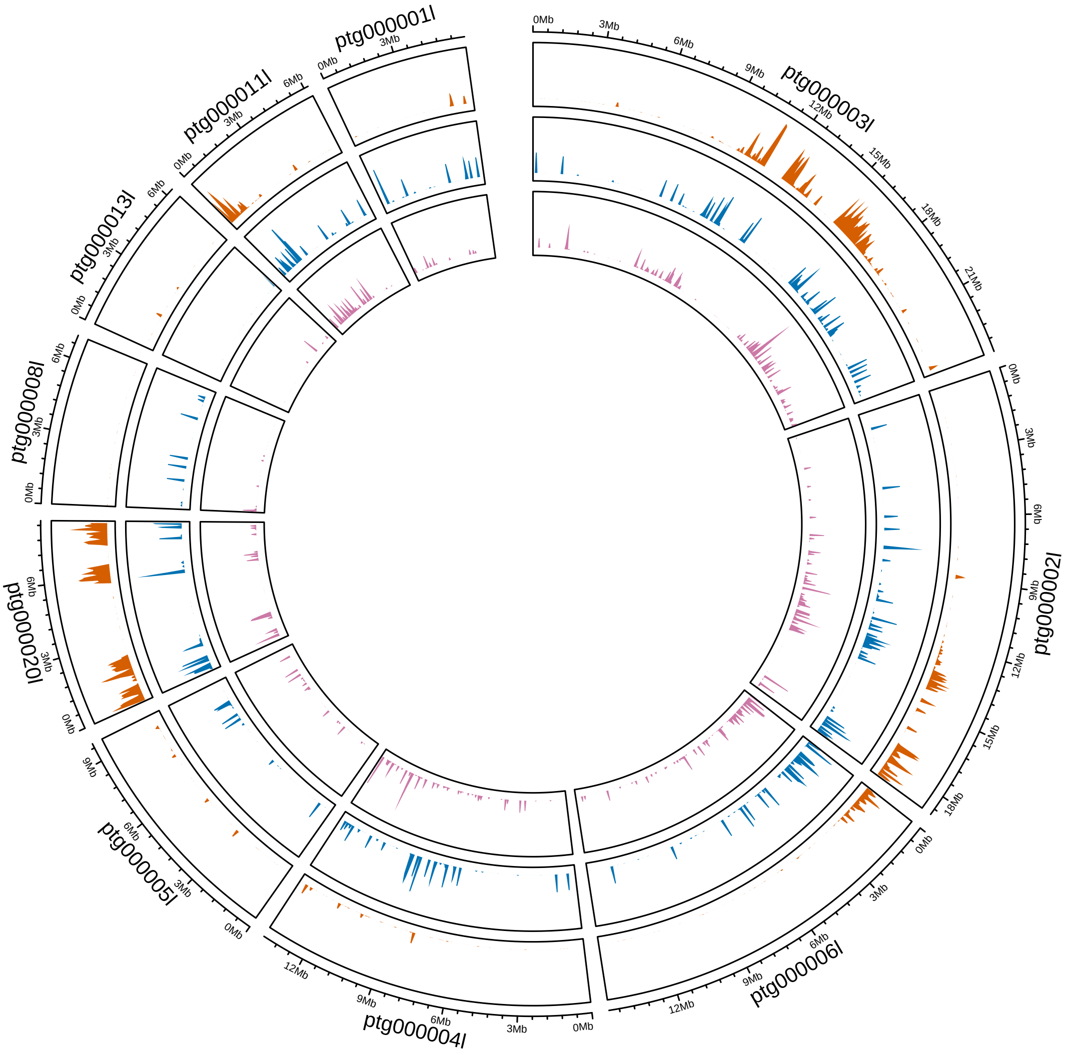


Figure 4: Genome-wide TE density across the longest Kas-1 scaffolds in 100 kb windows. The x-axis represents genomic position, and the y-axis represents TE density; peaks mark local TE-rich regions, while troughs indicate TE-poor segments. Kas-1 shows highly uneven TE distribution, with several clear TE hotspots that may correspond to structurally dynamic or low-recombination regions.

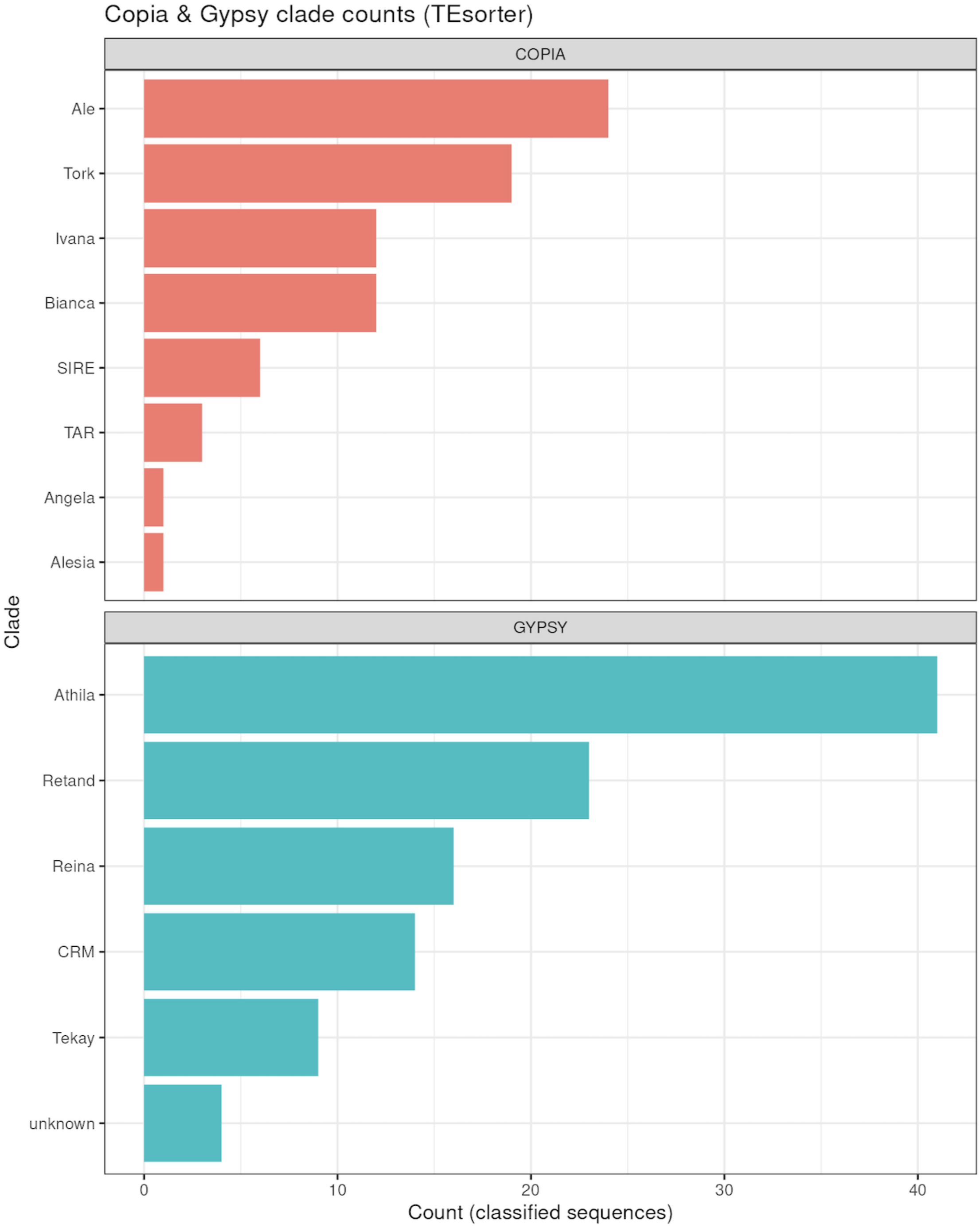


Figure 5: Clade-level classification of Copia and Gypsy LTR retrotransposons in Kas-1 using TEsorter. Bars show the number of elements assigned to each Copia or Gypsy clade, allowing comparison of which lineages are expanded. Kas-1 is enriched for Ale/Tork (Copia) and Athila (Gypsy) clades, reflecting a TE composition similar to other *A. thaliana* accessions.

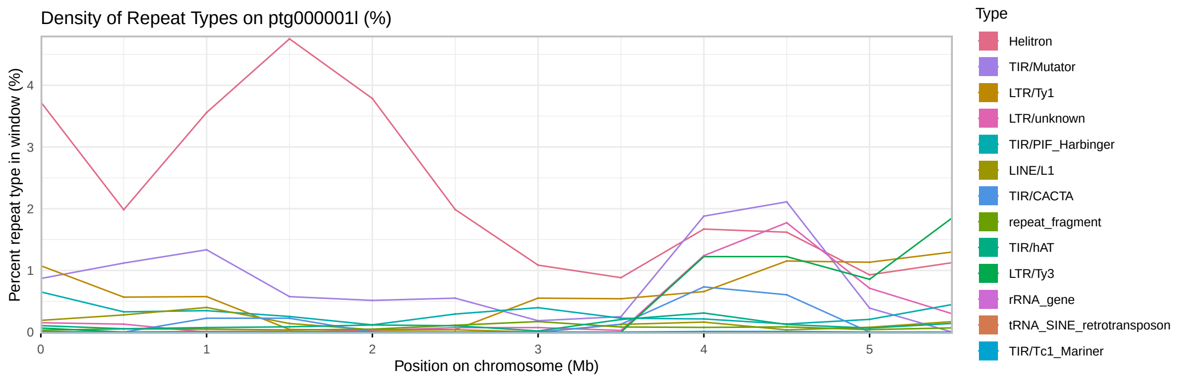


Figure 6: TE density tracks for major TE superfamilies across the longest Kas-1 scaffolds. Each track shows how the local density of a given TE group changes along the chromosomes, with peaks marking regions of enrichment. Different superfamilies have distinct spatial patterns, indicating that TE accumulation and removal have been heterogeneous across the Kas-1 genome.

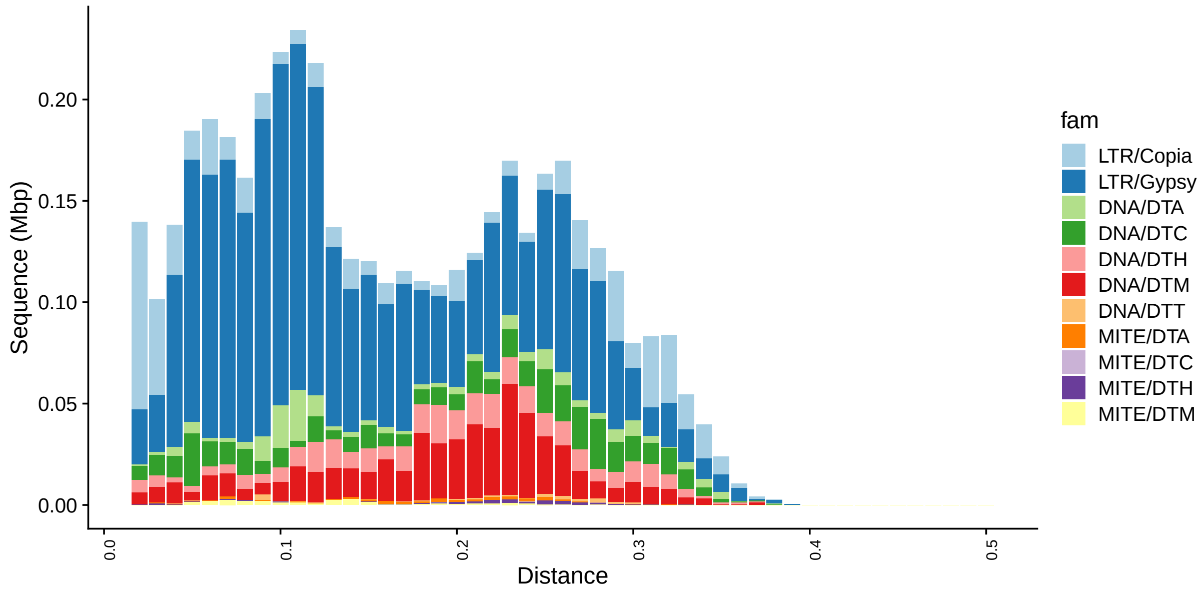


Figure 7: Distribution of sequence divergence among TE families in the Kas-1 genome. The x-axis shows Kimura divergence (as a proxy for age), and the y-axis shows counts per divergence bin; younger insertions cluster at low divergence, older ones at higher values. Kas-1 displays both low- and high-divergence peaks, suggesting multiple waves of TE activity over its evolutionary history.

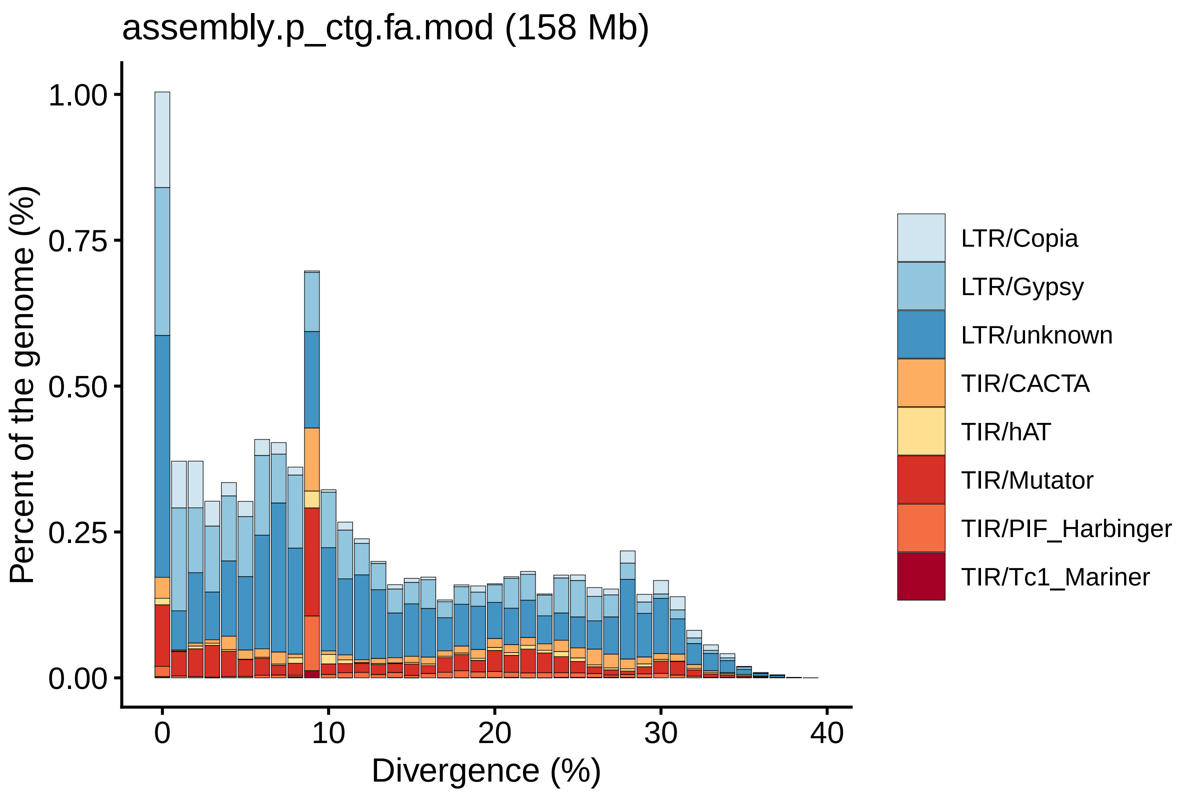
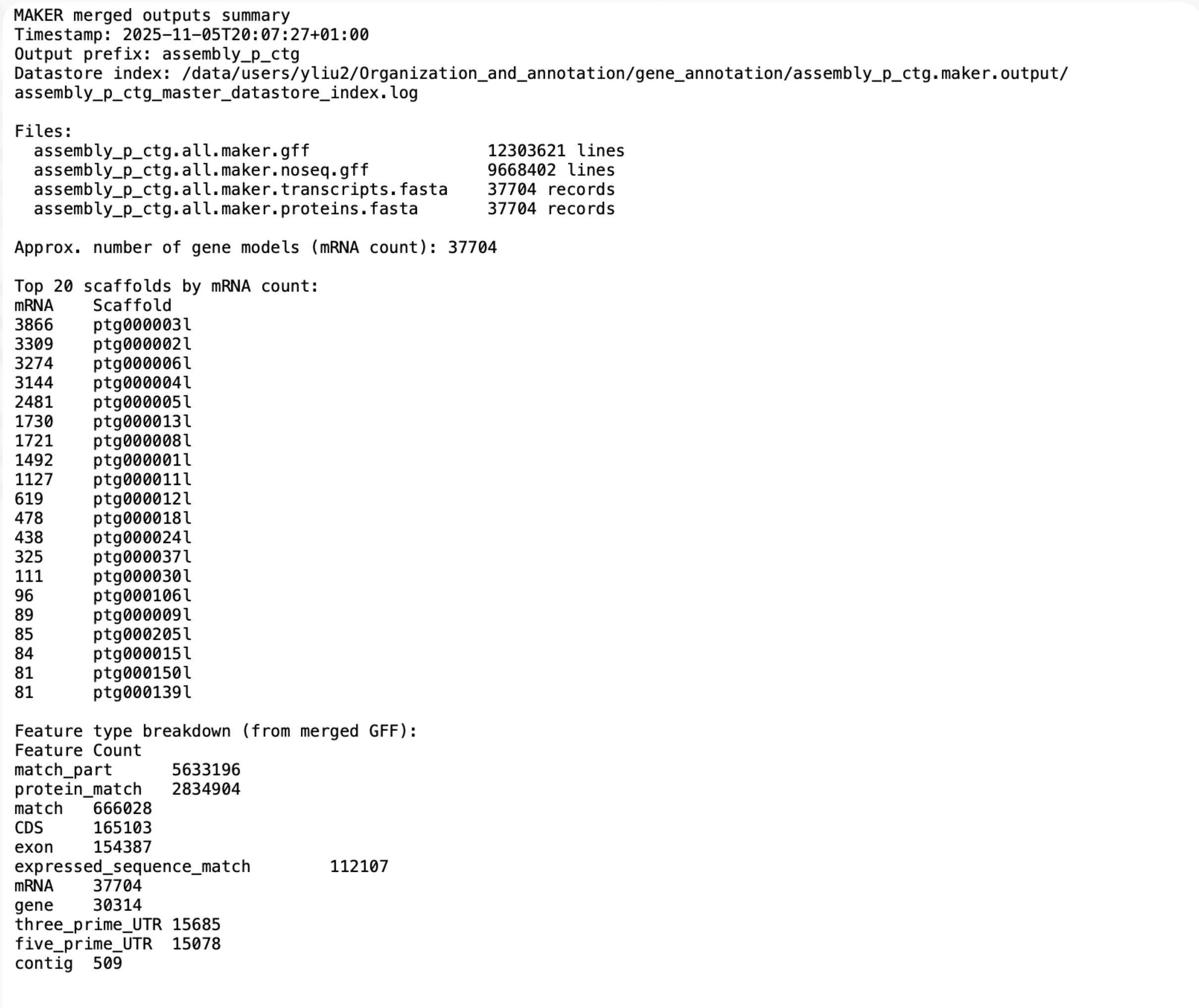
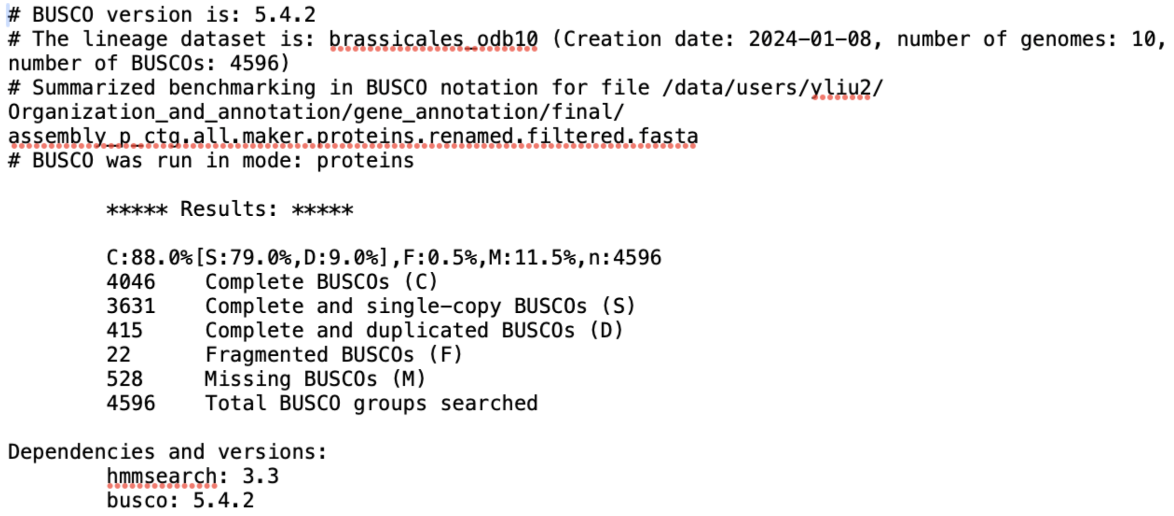


Figure 8: TE landscape plot summarizing divergence-based age distributions of the main TE superfamilies. For each superfamily, the curve height reflects the cumulative amount of sequence at a given divergence level, so peak positions and shapes reveal the timing and intensity of TE bursts. In Kas-1, LTR retrotransposons show pronounced peaks at relatively low divergence, indicating a strong recent expansion on top of an older TE background.



```
MAKER merged outputs summary
Timestamp: 2025-11-05T20:07:27+01:00
Output prefix: assembly_p_ctg
Datastore index: /data/users/yliu2/Organization_and_annotation/gene_annotation/assembly_p_ctg.maker.output/
assembly_p_ctg_master_datastore_index.log

Files:
  assembly_p_ctg.all.maker.gff                  12303621 lines
  assembly_p_ctg.all.maker.noseq.gff            9668402 lines
  assembly_p_ctg.all.maker.transcripts.fasta    37704 records
  assembly_p_ctg.all.maker.proteins.fasta       37704 records

Approx. number of gene models (mRNA count): 37704

Top 20 scaffolds by mRNA count:
mRNA    Scaffold
3866    ptg000003l
3309    ptg000002l
3274    ptg000006l
3144    ptg000004l
2481    ptg000005l
1730    ptg000013l
1721    ptg000008l
1492    ptg000001l
1127    ptg000011l
619     ptg000012l
478     ptg000018l
438     ptg000024l
325     ptg000037l
111     ptg000030l
96      ptg000106l
89      ptg000009l
85      ptg000205l
84      ptg000015l
81      ptg000150l
81      ptg000139l

Feature type breakdown (from merged GFF):
Feature Count
match_part      5633196
protein_match   2834904
match   666028
CDS     165103
exon    154387
expressed_sequence_match        112107
mRNA    37704
gene    30314
three_prime_UTR 15685
five_prime_UTR  15078
contig  509
```

Figure 9: Summary of MAKER gene annotation results for the Kas-1 genome. The figure reports counts of genes, mRNAs, exons, and other features, as well as the proportion of models supported by RNA-seq and protein homology. Kas-1 has just over 30,000 predicted genes backed by multiple lines of evidence, indicating a robust and biologically plausible gene set.



```
# BUSCO version is: 5.4.2
# The lineage dataset is: brassicales_odb10 (Creation date: 2024-01-08, number of genomes: 10,
number of BUSCOs: 4596)
# Summarized benchmarking in BUSCO notation for file /data/users/yliu2/
Organization_and_annotation/gene_annotation/final/
assembly_p_ctg.all.maker.proteins.renamed.filtered.fasta
# BUSCO was run in mode: proteins

        ***** Results: *****

        C:88.0%[S:79.0%,D:9.0%],F:0.5%,M:11.5%,n:4596
        4046    Complete BUSCOs (C)
        3631    Complete and single-copy BUSCOs (S)
        415     Complete and duplicated BUSCOs (D)
        22      Fragmented BUSCOs (F)
        528     Missing BUSCOs (M)
        4596    Total BUSCO groups searched

Dependencies and versions:
        hmmsearch: 3.3
        busco: 5.4.2
```

Figure 10: BUSCO short summary report for assessing the completeness of the Kas-1 gene annotation. Bars summarize the fractions of complete, duplicated, fragmented, and missing Brassicales single-copy orthologs, providing a standardized quality metric. The high proportion of complete BUSCOs confirms that the Kas-1 assembly and annotation capture most expected conserved genes.

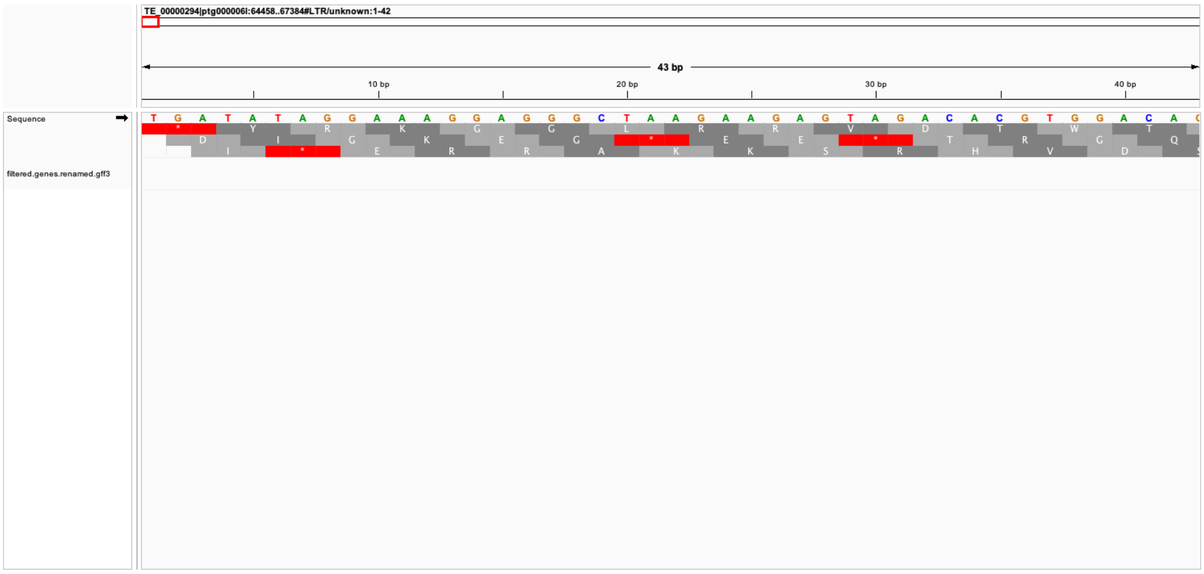
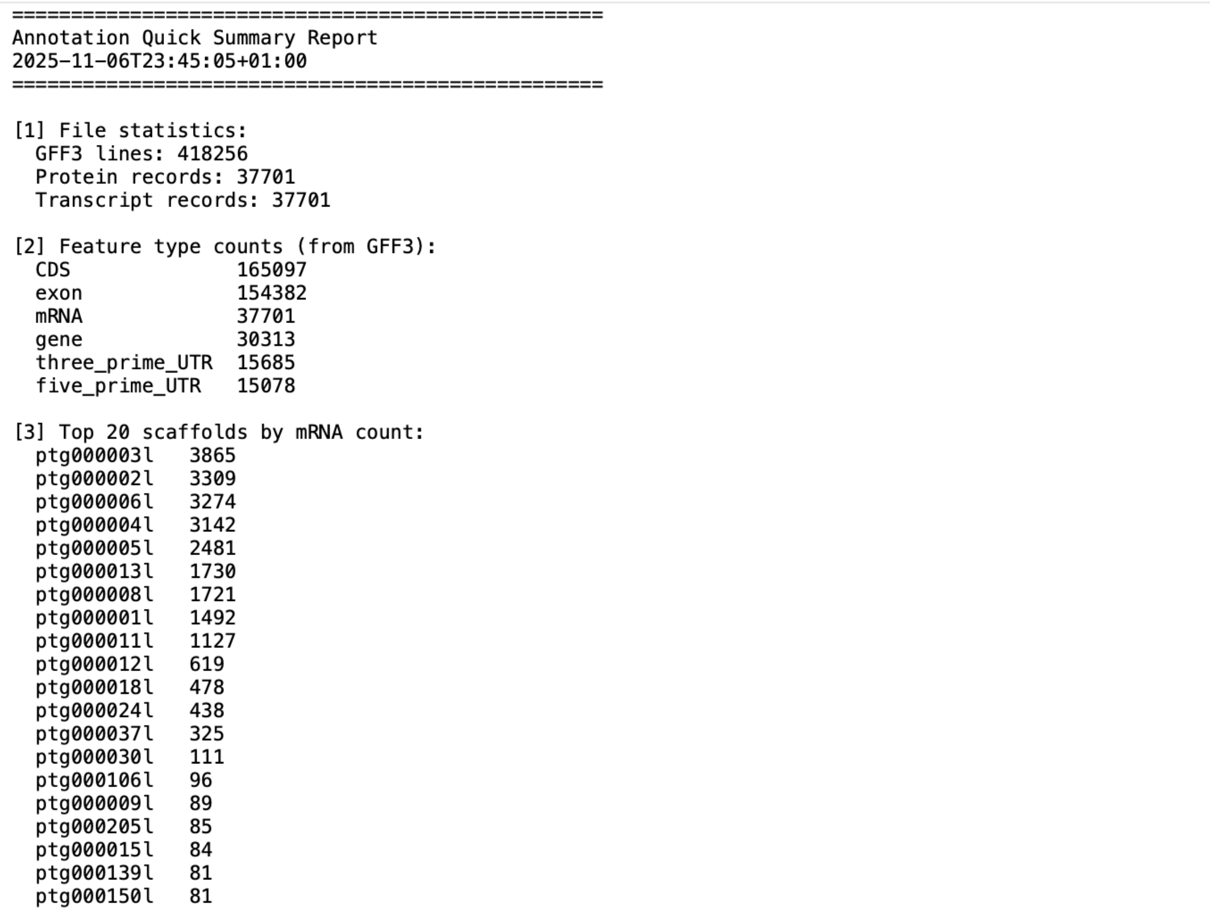

TE_00000294|ptg000006l:64458..67384#LTR/unknown:1-42
43 bp
10 bp
20 bp
30 bp
40 bp
Sequence
T G A T A T A G G A A A G G A G G G C T A A G A A G A G T A G A C A C G T G G A C A
filtered.genes.renamed.gff3

Figure 11: IGV snapshot showing Kas-1 gene models and nearby TE annotation at a representative locus. Tracks display the genomic sequence, coding exons, introns, and TE features, allowing visual inspection of gene structure and the proximity of repeats. This example illustrates how TE insertions can occur near genes in Kas-1, potentially affecting local regulation or structural stability.



```
==================================================
Annotation Quick Summary Report
2025-11-06T23:45:05+01:00
==================================================

[1] File statistics:
  GFF3 lines: 418256
  Protein records: 37701
  Transcript records: 37701

[2] Feature type counts (from GFF3):
  CDS              165097
  exon             154382
  mRNA             37701
  gene             30313
  three_prime_UTR  15685
  five_prime_UTR   15078

[3] Top 20 scaffolds by mRNA count:
  ptg000003l   3865
  ptg000002l   3309
  ptg000006l   3274
  ptg000004l   3142
  ptg000005l   2481
  ptg000013l   1730
  ptg000008l   1721
  ptg000001l   1492
  ptg000011l   1127
  ptg000012l   619
  ptg000018l   478
  ptg000024l   438
  ptg000037l   325
  ptg000030l   111
  ptg000106l   96
  ptg000009l   89
  ptg000205l   85
  ptg000015l   84
  ptg000139l   81
  ptg000150l   81
```

Figure 12: Annotation quick summary report for the MAKER gene set in the Kas-1 assembly. The figure aggregates counts of genes, mRNAs, CDSs, and exons across the assembly and highlights scaffold-level distributions. Kas-1 shows over 30,000 genes with several gene-rich scaffolds, consistent with a high-quality and relatively compact *Arabidopsis* genome.

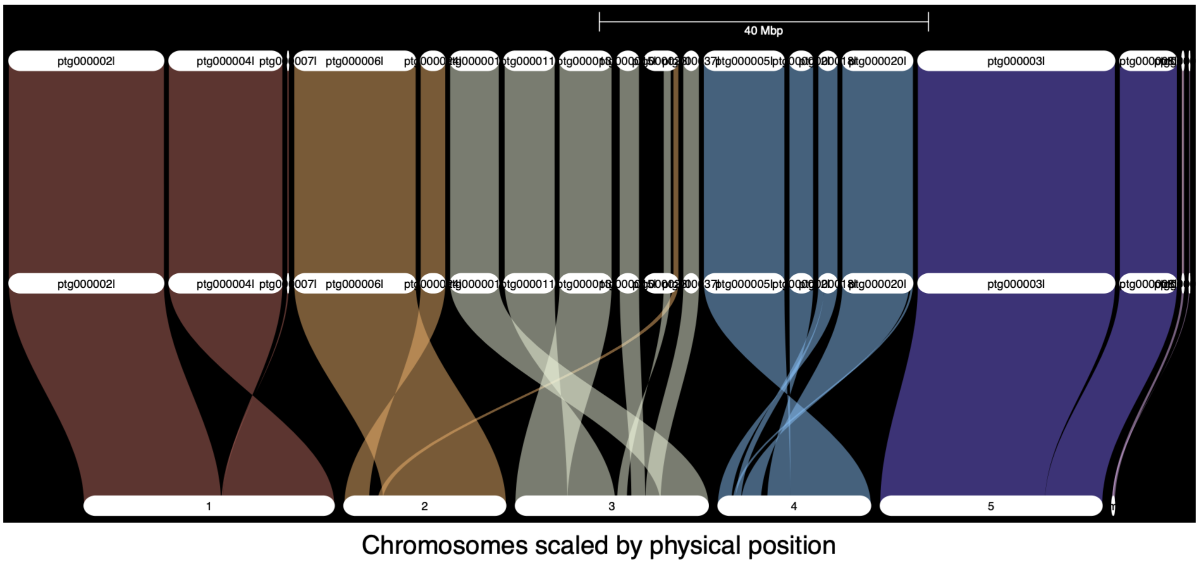
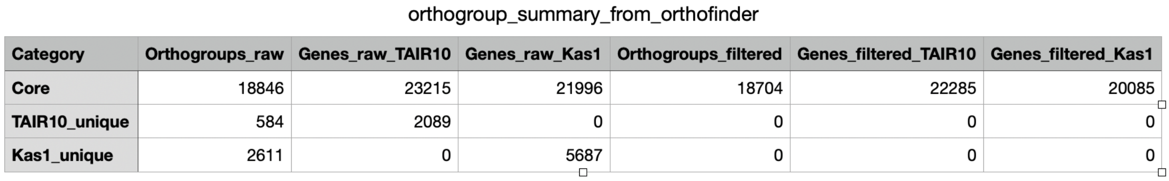


Figure 13: Riparian synteny plot comparing Kas-1, Kas-1.norm, and the *Arabidopsis thaliana* TAIR10 reference. Each polygon represents a chromosome or scaffold, and colored ribbons link syntenic blocks across genomes, so continuous ribbons indicate conserved gene order. Kas-1 shows largely collinear synteny with TAIR10, supporting high assembly accuracy and a largely conserved chromosome structure with limited large-scale rearrangements.



orthogroup_summary_from_orthofinder

Category	Orthogroups_raw	Genes_raw_TAIR10	Genes_raw_Kas1	Orthogroups_filtered	Genes_filtered_TAIR10	Genes_filtered_Kas1
Core	18846	23215	21996	18704	22285	20085
TAIR10_unique	584	2089	0	0	0	0
Kas1_unique	2611	0	5687	0	0	0

Figure 14: Orthogroup sharing between TAIR10 and Kas-1 based on OrthoFinder results. Bars distinguish orthogroups shared between both genomes from those specific to Kas-1, allowing a quick readout of core versus accession-specific content. Kas-1 shares 20,085 orthogroups with TAIR10 but also carries 5,687 accession-specific orthogroups, indicating substantial lineage-specific gene content variation.