# Organization and Annotation of Eukaryote Genomes

Yuwei Liu  Student ID: 24-100-430

November 15, 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 class composition and genome coverage in the Kas-1 genome.

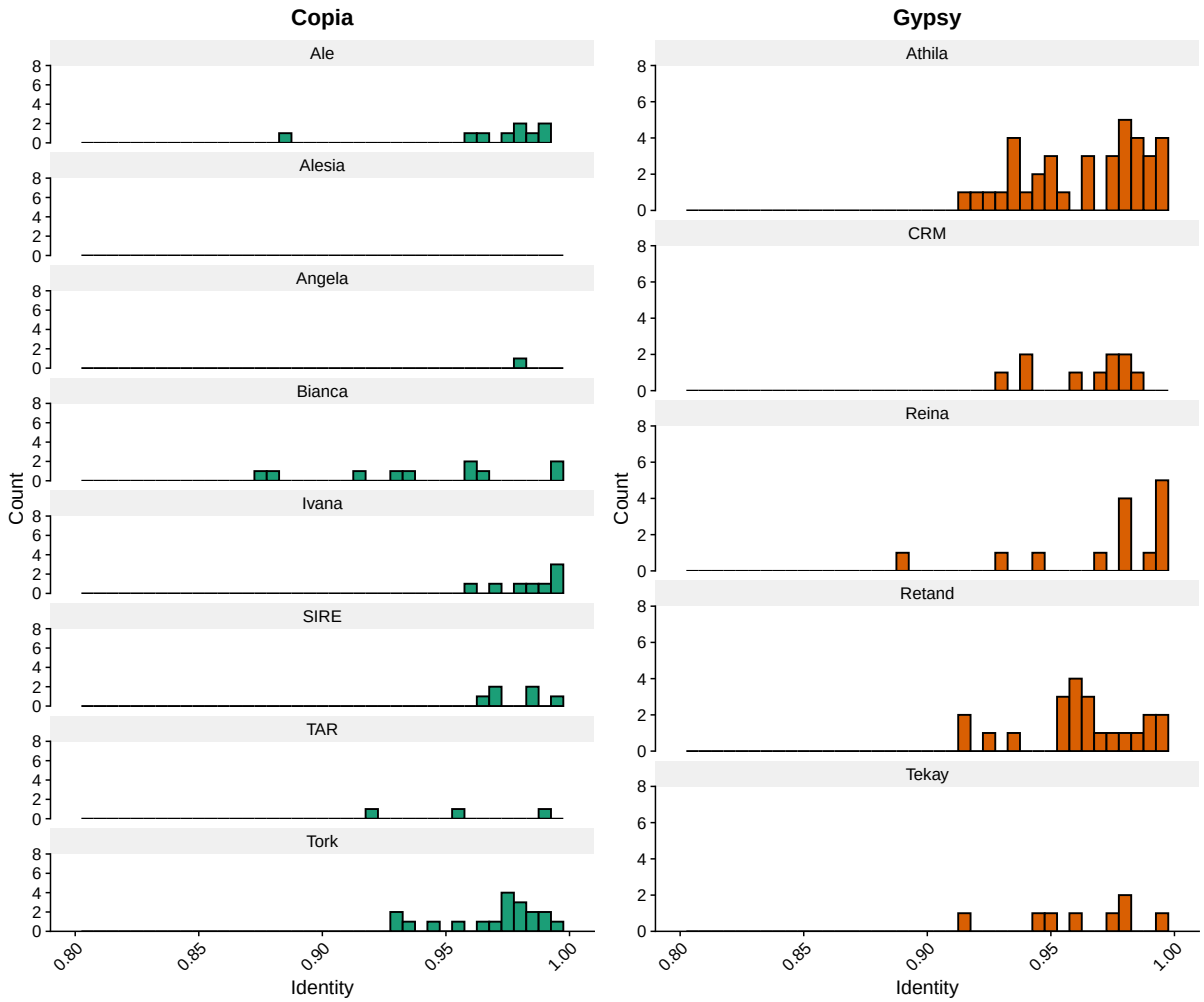


Figure 2: Percent-identity distribution of full-length Copia and Gypsy LTR retrotransposons in Kas-1.

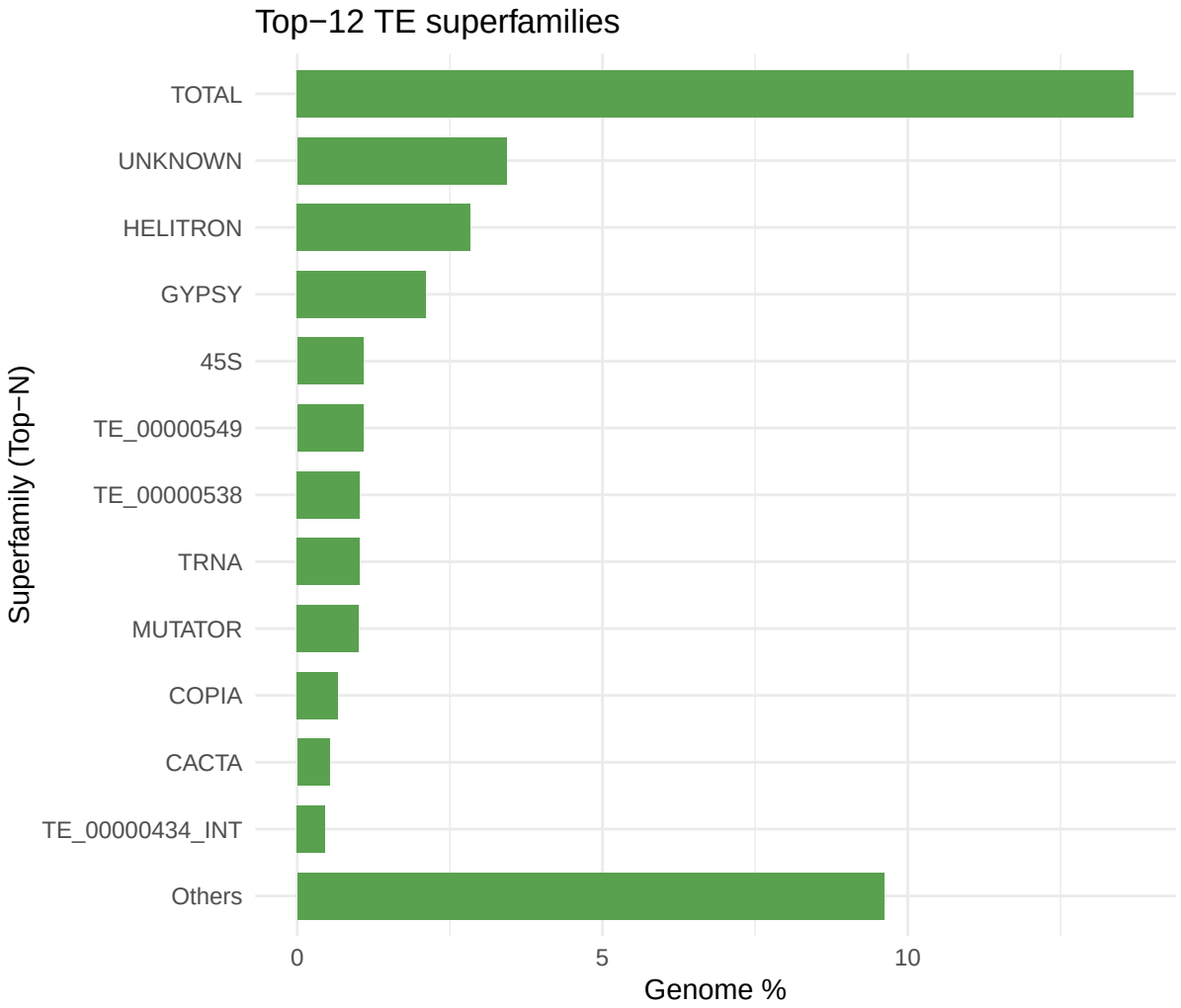


Figure 3: Percent genome coverage of the most abundant TE superfamilies in the Kas-1 assembly.

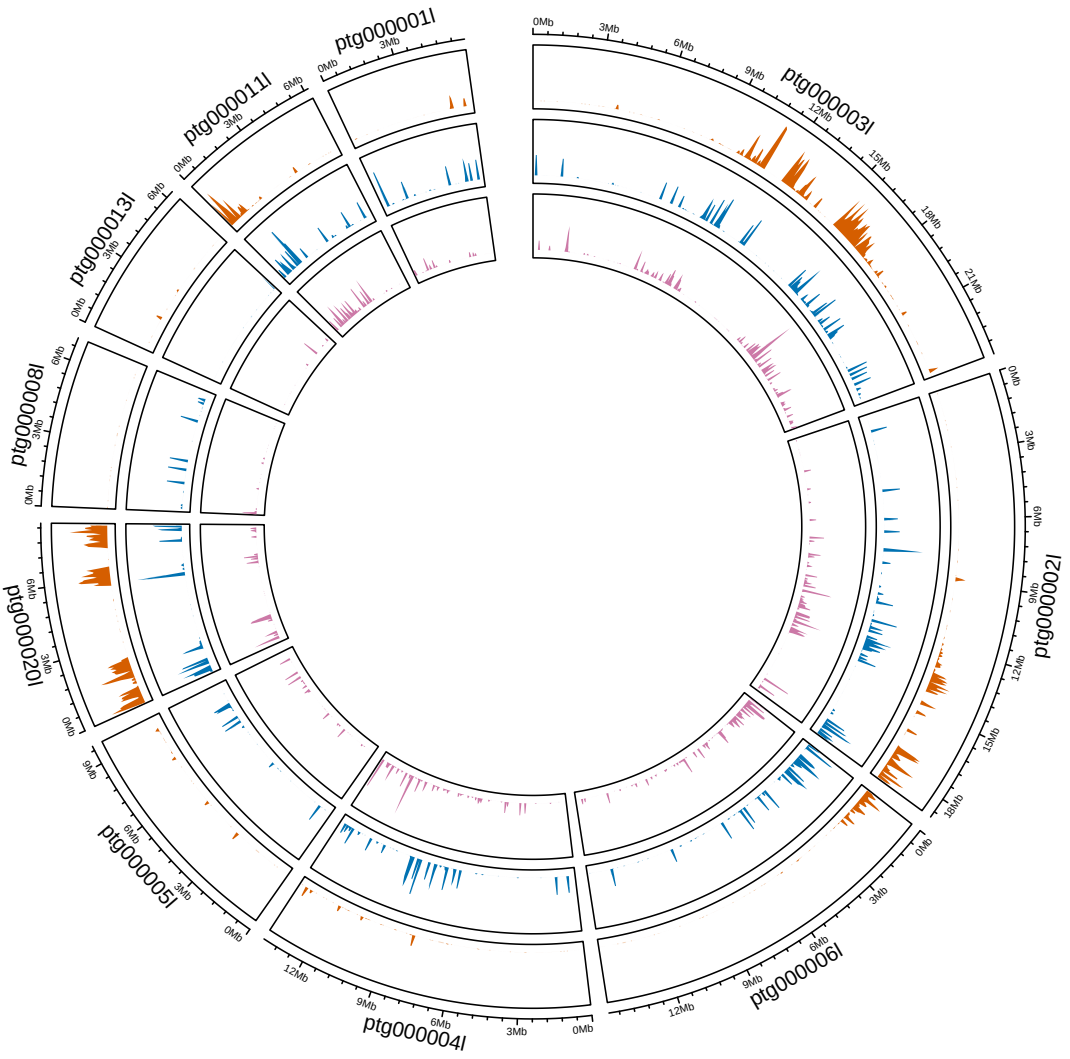


Figure 4: Genome-wide TE density across the longest Kas-1 scaffolds in 100 kb windows.

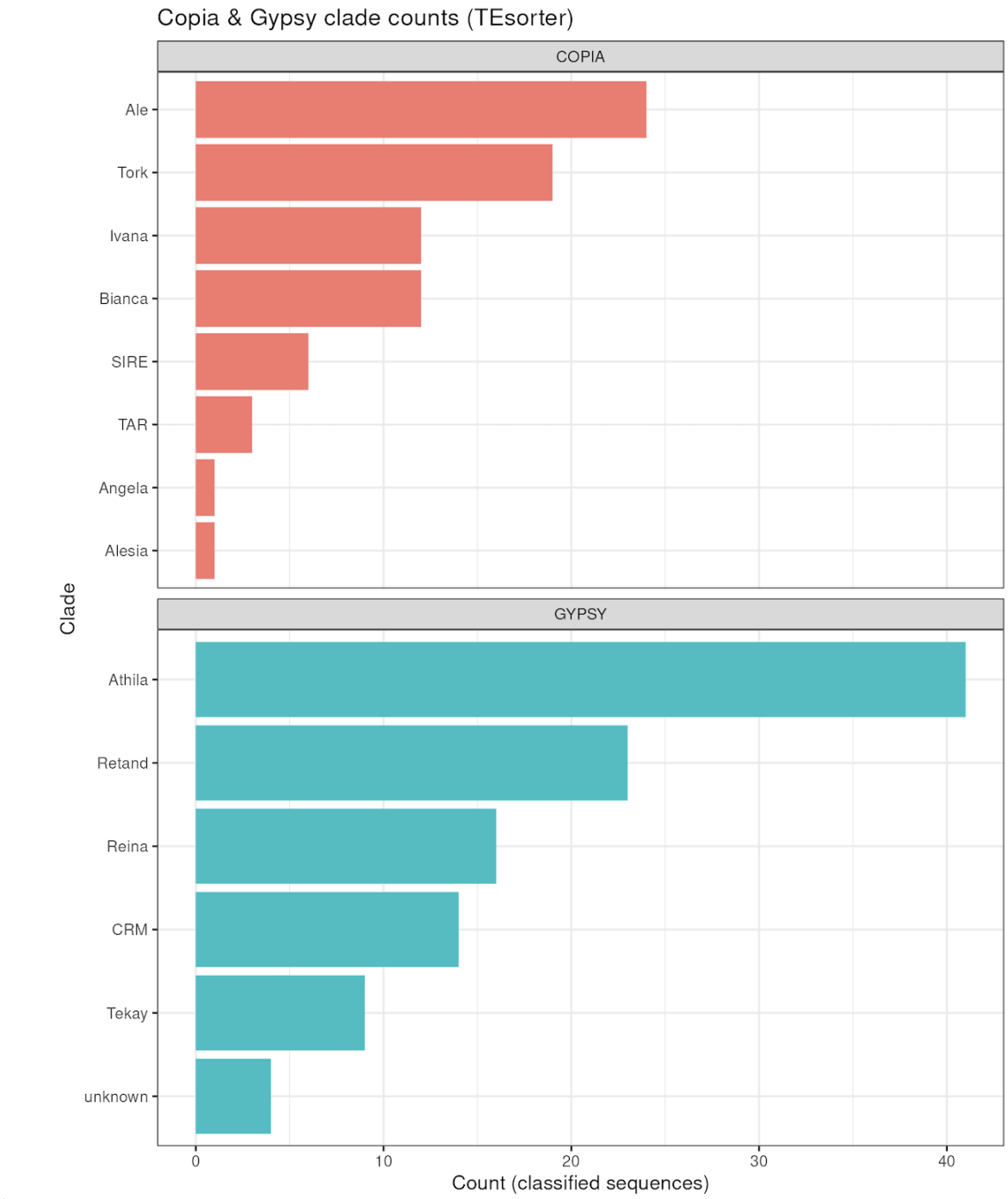


Figure 5: Clade-level classification of Copia and Gypsy LTR retrotransposons in Kas-1 using TEsorter.

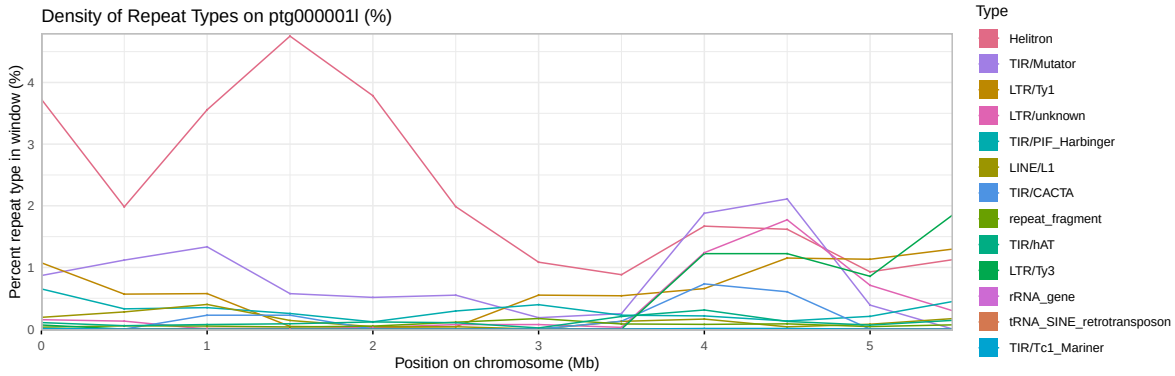


Figure 6: TE density tracks for major TE superfamilies across the longest Kas-1 scaffolds.

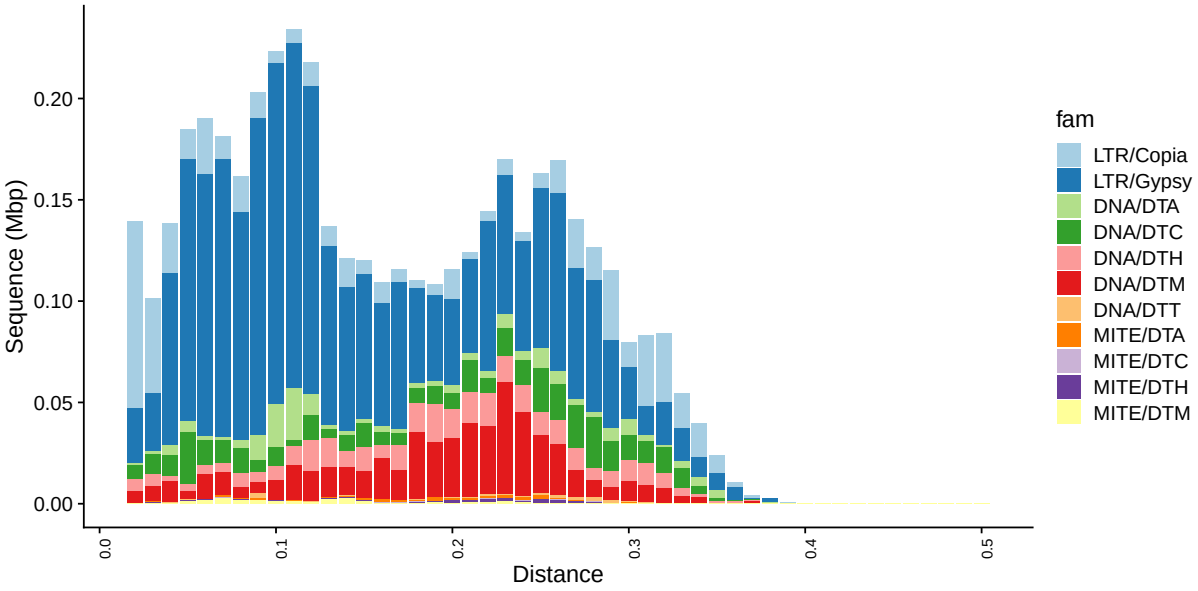


Figure 7: Distribution of sequence divergence among TE families in the Kas-1 genome.

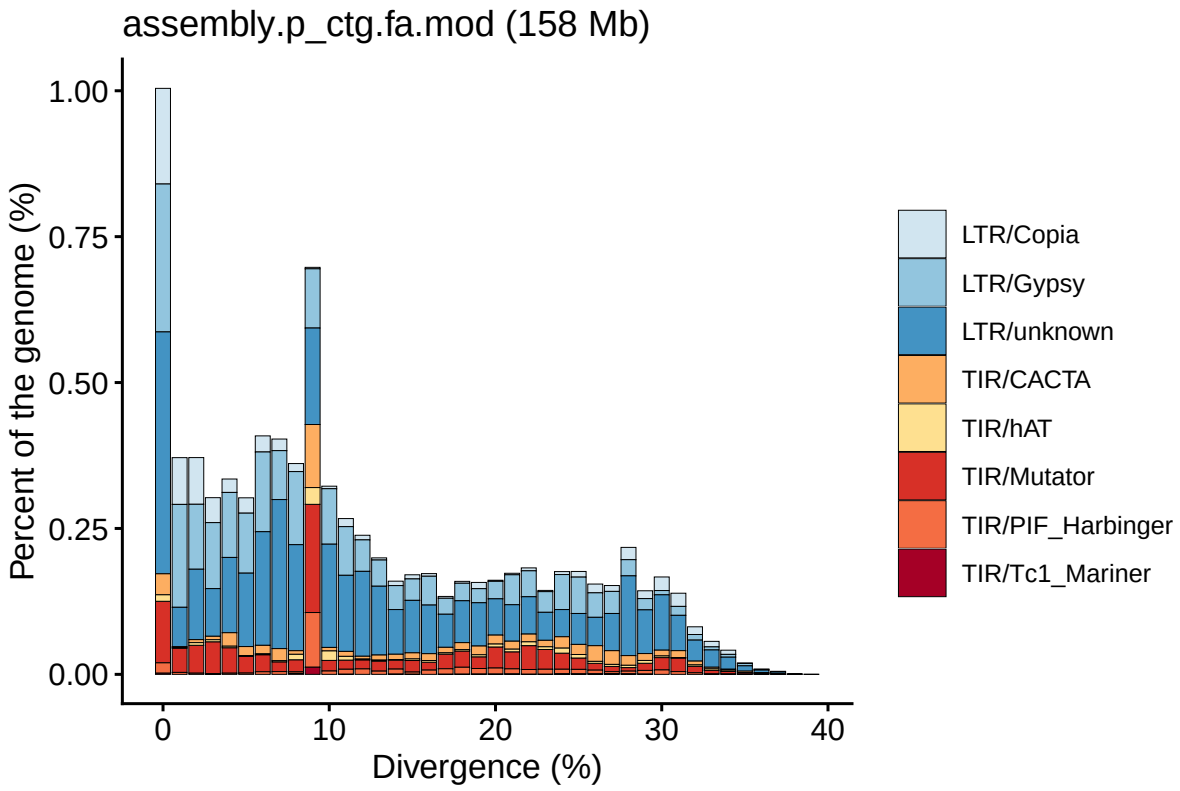


Figure 8: TE landscape plot showing divergence-based age distribution of major TE superfamilies.

```
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.

```
# 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 completeness of the Kas-1 gene annotation.


Figure 11: IGV snapshot showing Kas-1 gene models and nearby TE annotation at a representative locus.

```
==================================================
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.

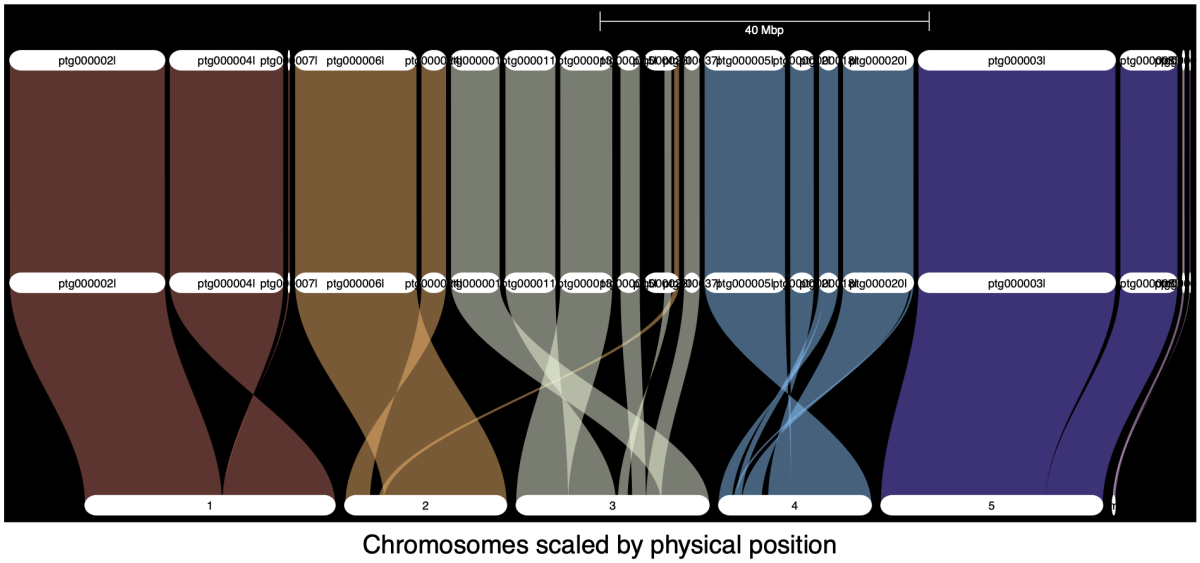


Figure 13: Riparian synteny plot comparing Kas-1, Kas-1.norm, and the Arabidopsis thaliana TAIR10 reference.

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.