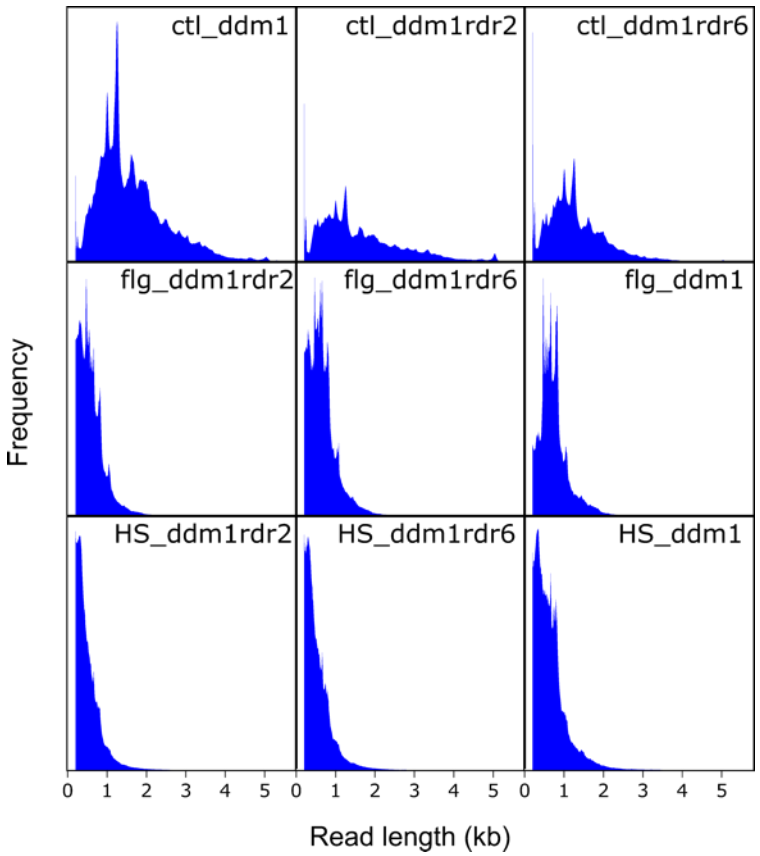
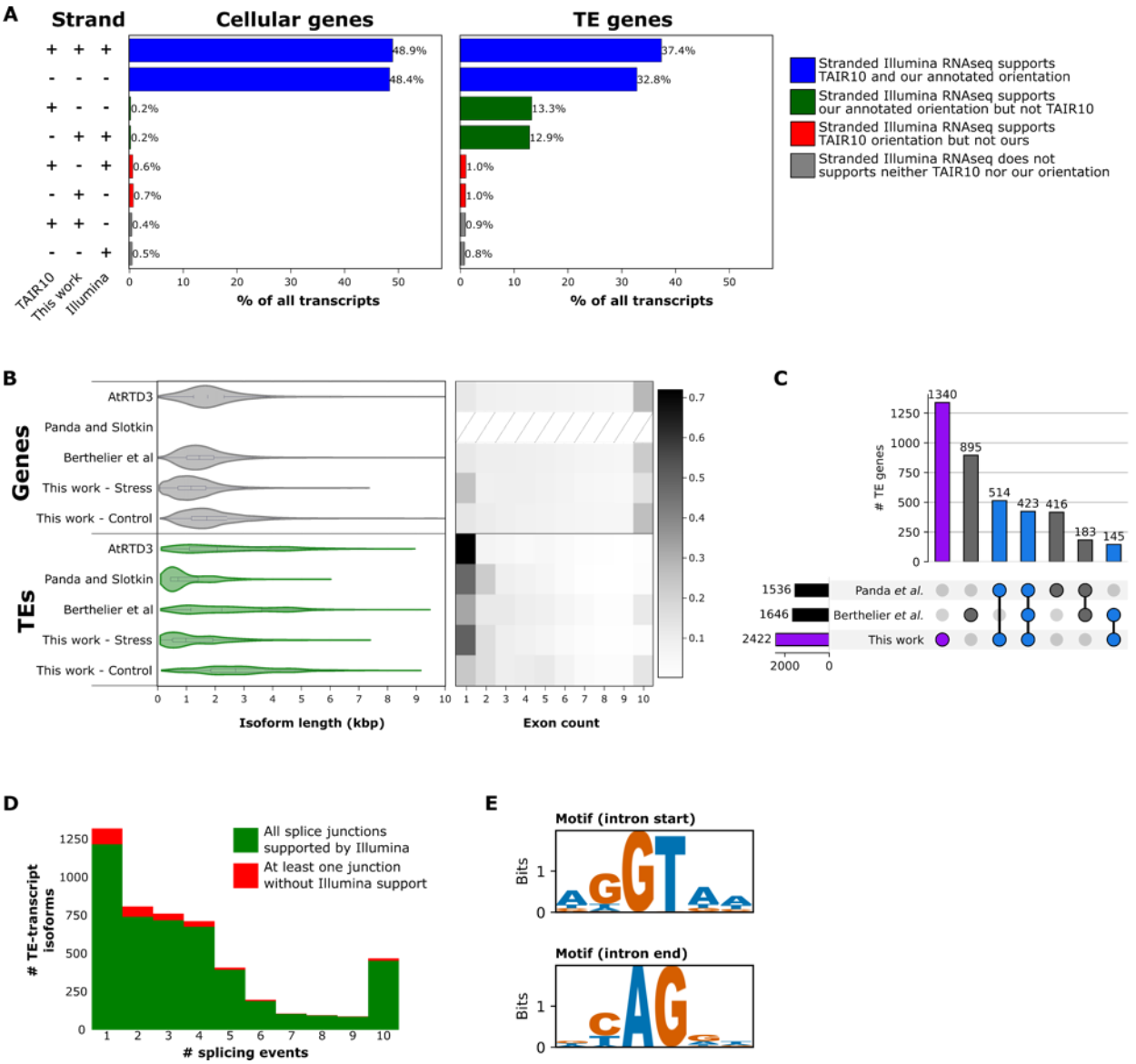
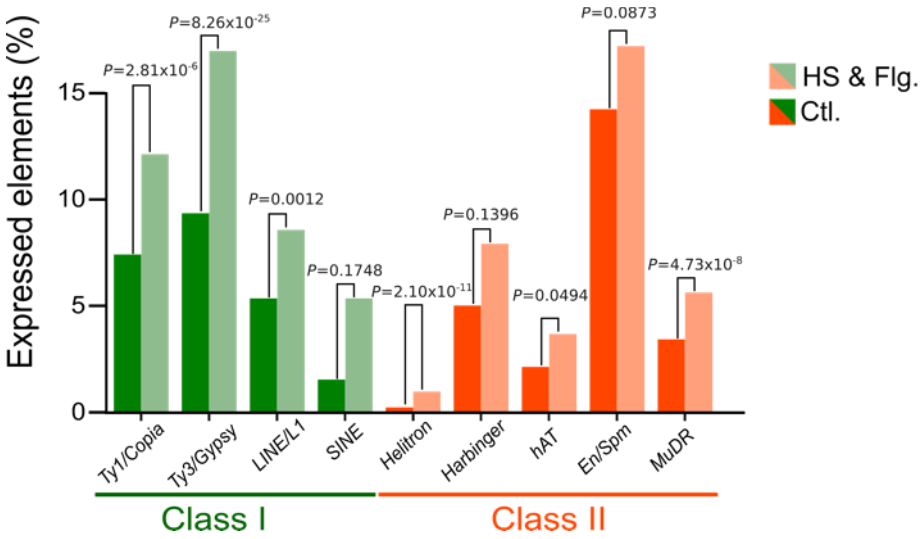


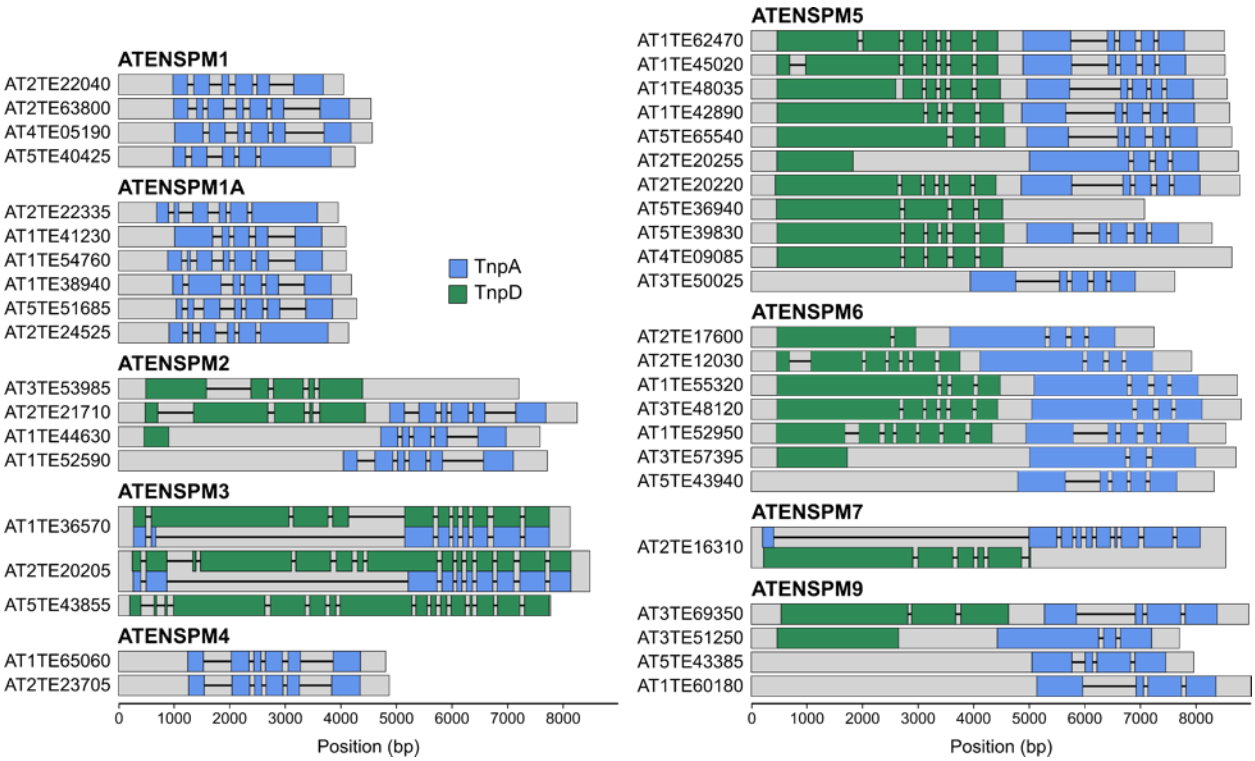
Supplementary Figure 1. ONT read length distribution, related to Figure 1. Read length distribution of the different ONT sequencing experiments.



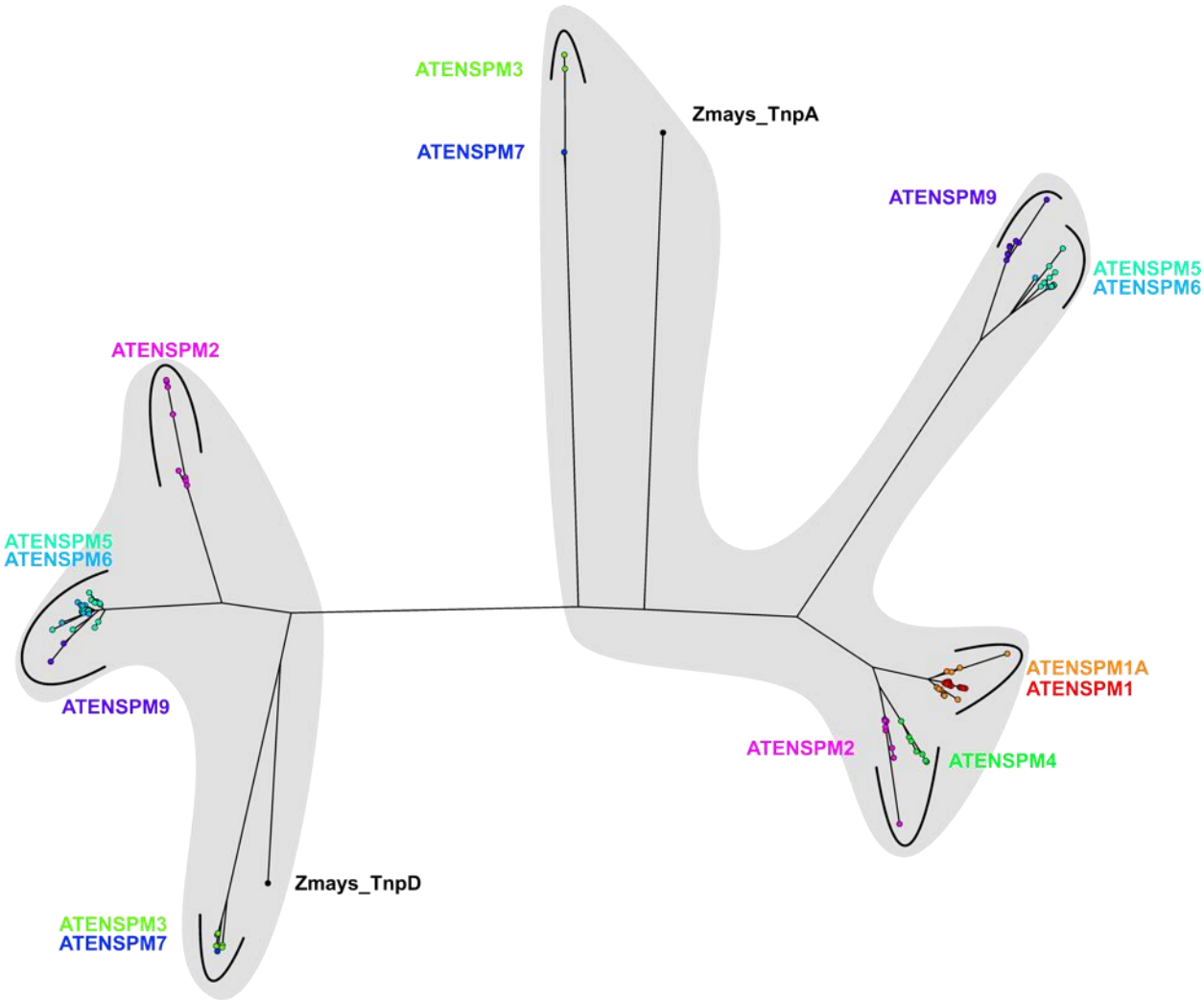
Supplementary Figure 2: TE transcript strandness and annotation lengths, related to Figure 1. **A.** Strand comparison between TAIR10 annotation, our annotation, and stranded Illumina RNA-seq data. Only transcripts with assigned strands in the three datasets are compared. **B**. Length distribution and number of exons comparing our annotation with previous annotation efforts. **C**. Upset plot depicting the number of novel and known TE genes. **D**. Number of isoforms with splicing events validated by short-read data. **E**. Motif of the start and end sequences of introns defined in our pipeline.



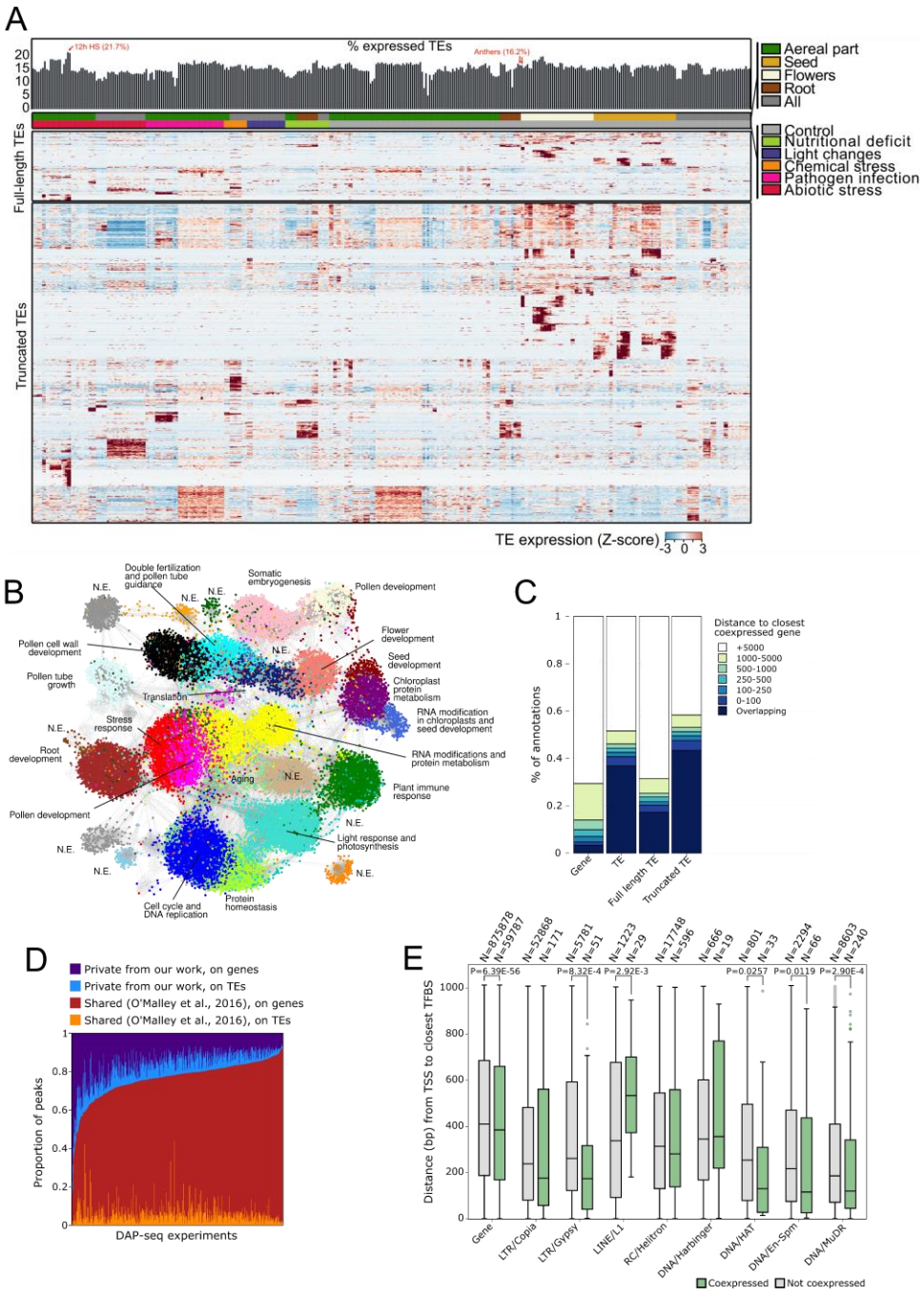
Supplementary Figure 3: TE reactivation upon stress, related to Figure 1D. Proportion of transcriptionally competent TEs per superfamily, detected on mutant plants grown either under control conditions or subject to stress. Statistical significance of the differences per superfamily was calculated using a Chi^2 test.



Supplementary Figure 4. *EN/SPM* gene architectures on the *Arabidopsis thaliana* genome, related to Figure 1. TE-genes and isoforms from all the *EN/SPM* elements characterized in this work. The identity of each transcript was assigned by aligning the proteins encoded by them with the *Zea mays* TnpA and TnpD (NCBI protein database accessions AAG17044 and AAG17043, respectively).

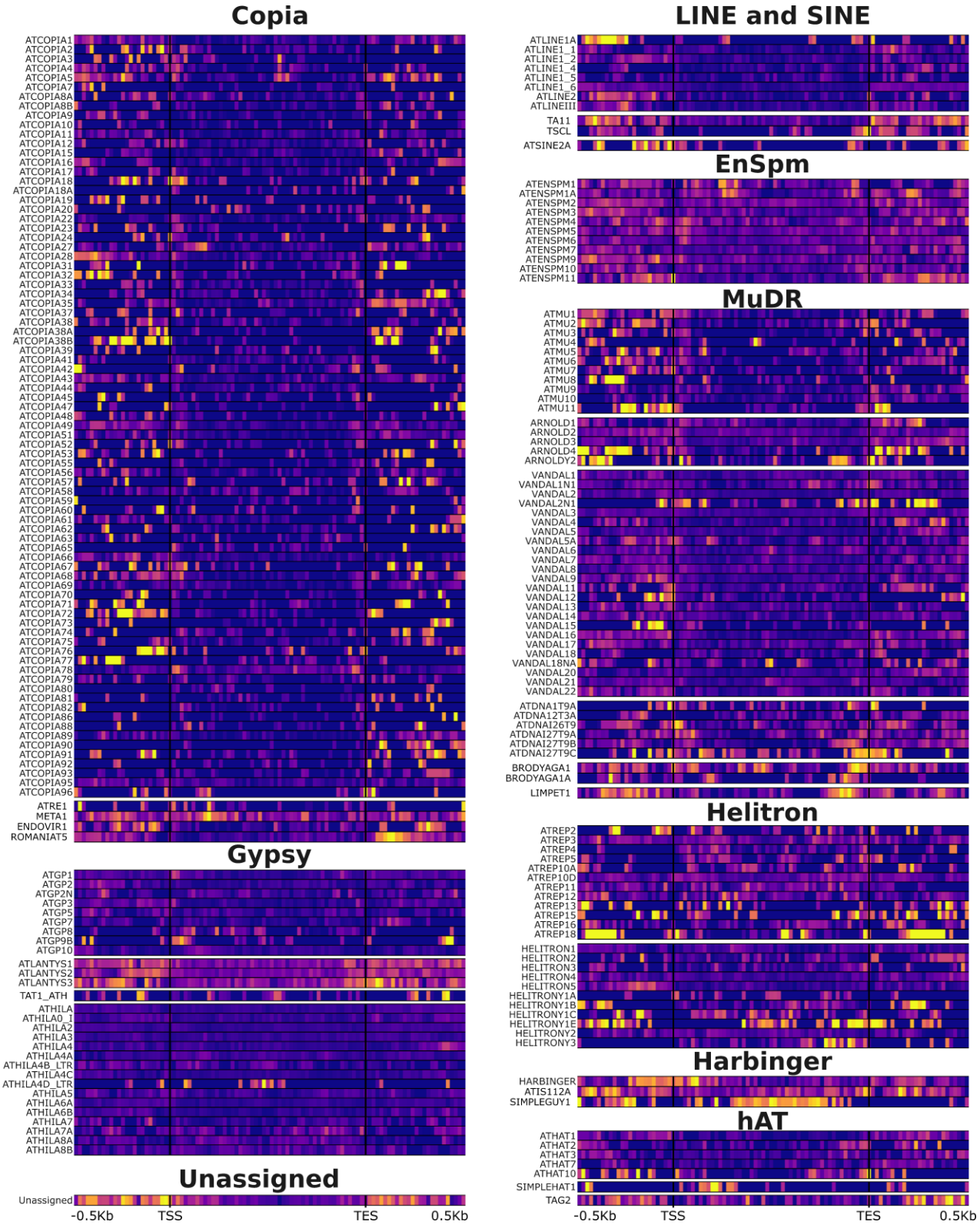


Supplementary Figure 5. Phylogenetic tree of *EN/SPM*-encoded proteins, related to Figure 1. All proteins with a length over 200 amino acids encoded by transcripts derived from *ATENSPM1 to ATENSPM9* elements are included. Two major clades are highlighted.

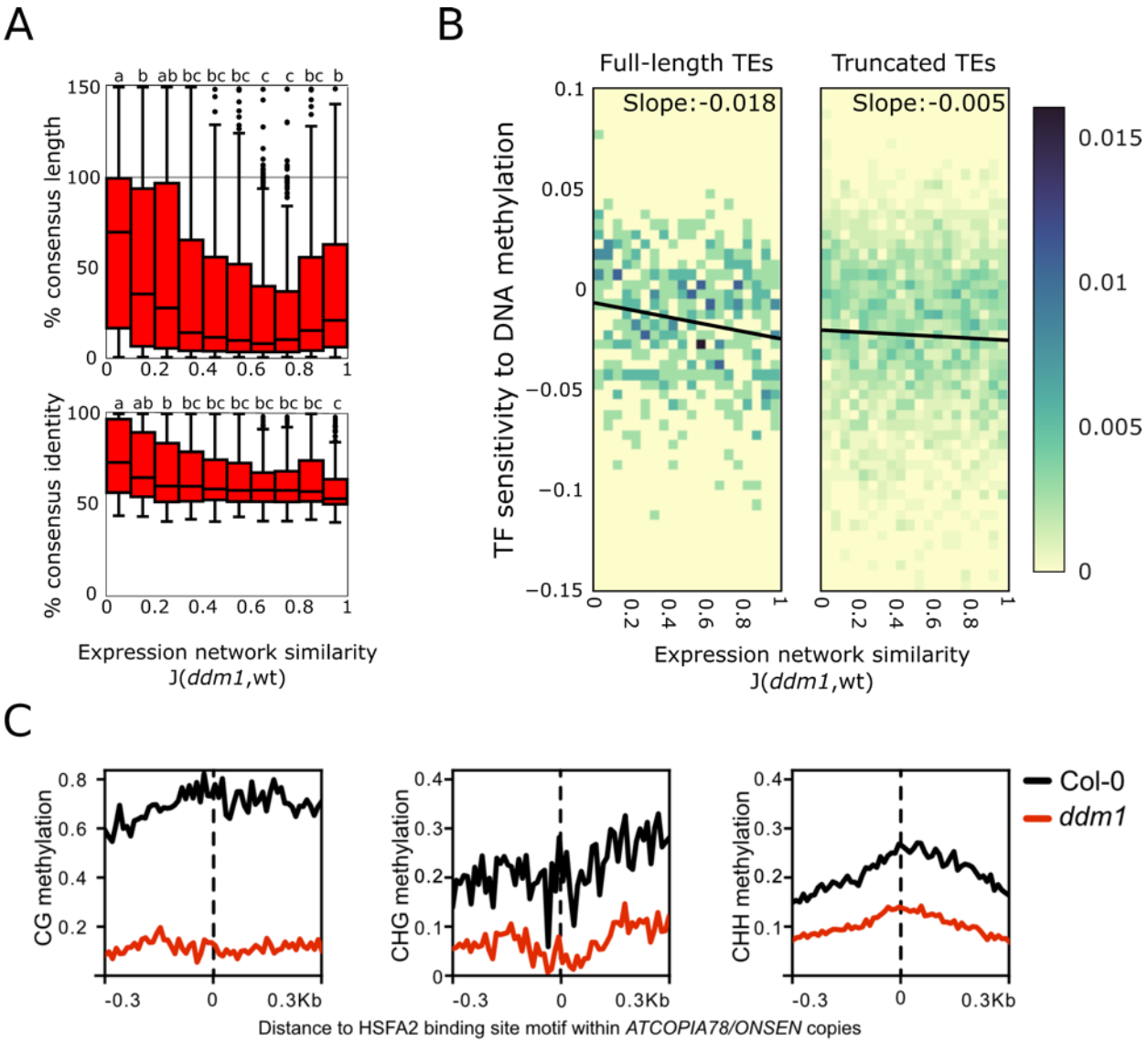


Supplementary Figure 6. TE transcriptome, related to Figure 2. A. Heatmap of normalized expression levels of TEs across 375 transcriptomic datasets from wild-type Col-0 plants. Expression for full-length and truncated TEs are shown separately. Percentage of transcriptional competent TEs expressed (TPM>5) in each dataset is shown as barplot. **B**. Coexpression clusters are shown in different colors. Genes not included in any cluster are hidden. Major GO terms associated with each cluster are shown, for the full list see Supplementary Data 3 and 4. N.E.: Non enriched in Gene Ontology terms. **C**. Distance from TE-genes and cellular genes to their closest coexpressed gene. Full length TEs were defined as those spanning over 90% of the length of their corresponding family's consensus sequence. **D**. Proportion of DAP-seq peaks shared between our analysis and that reported previously by O'Malley et al 2016. **E**. Absolute distance between each TSS and all TFBS within 1kbp from it. TF-TE-gene and TF-gene pairs in the same expression module were considered coexpressed. Statistically significant differences are marked

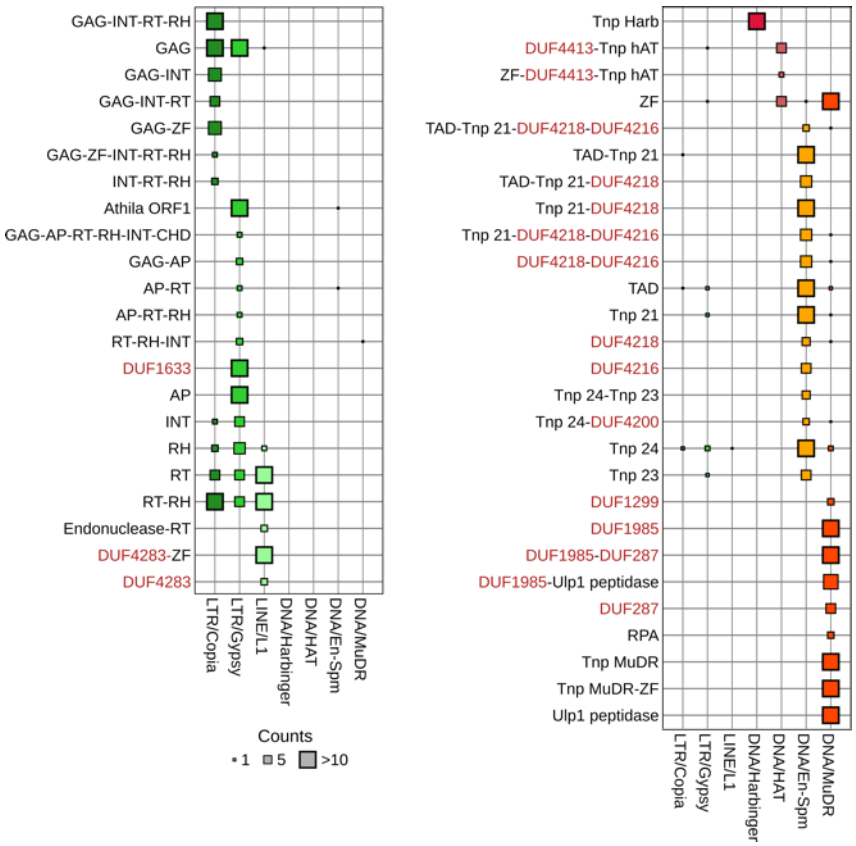
(Mann-Whitney U test).



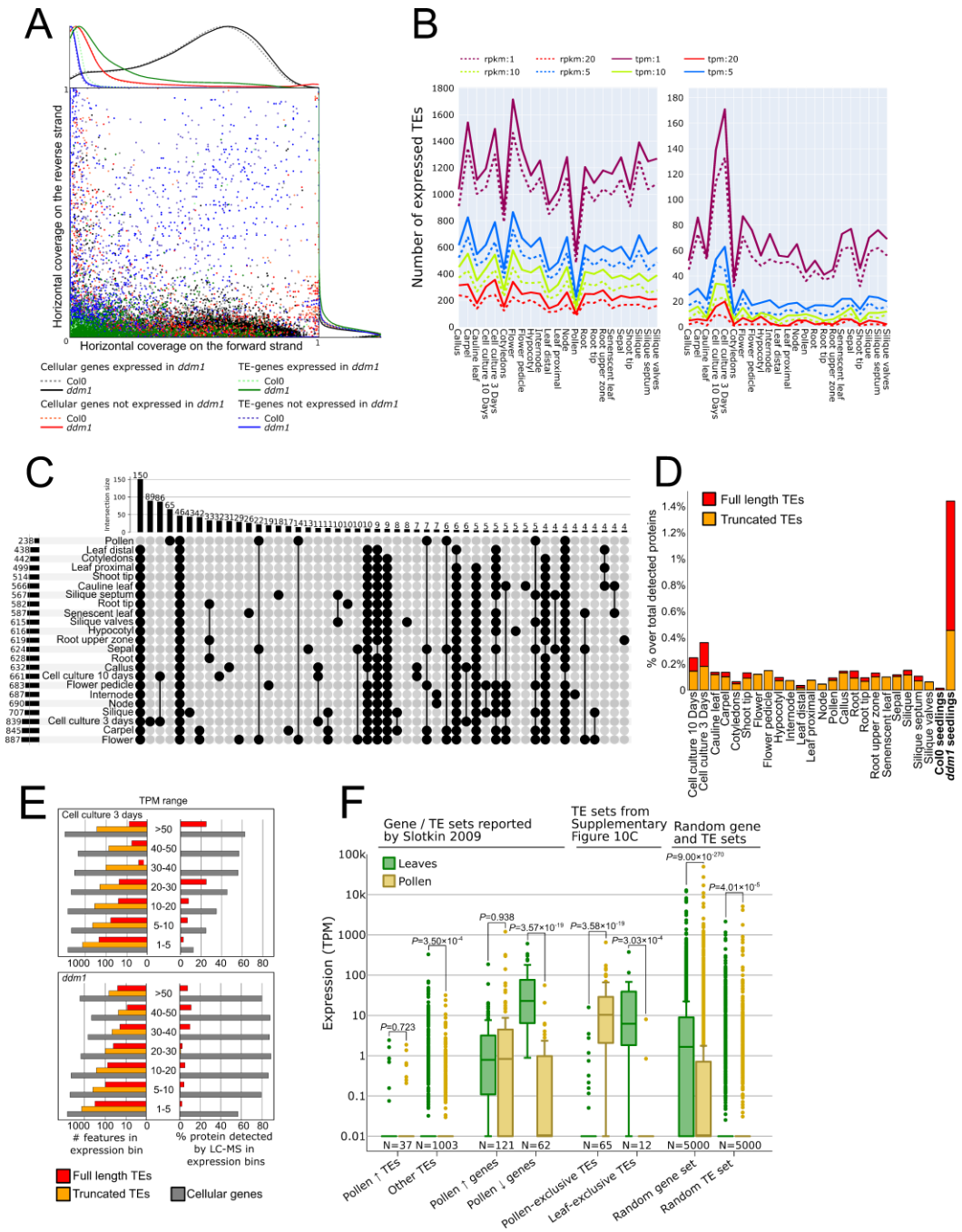
Supplementary Figure 7. Transcription factor binding site distribution. Heatmap of TFBS abundance across each TE family inside the TE boundaries and 500bp up and downstream. Only TEs with a total length over 1000bp are displayed.



Supplementary Figure 8. DNA methylation shapes the transcriptional network of TE-genes, related to Figure 2. **A**. Distribution of sizes and similarities of TEs compared to their corresponding consensus sequences across different deciles of Jaccard indexes. Significant differences were tested using pairwise Tukey HSD tests. Each box includes one decile of the TE-genes, or 327 TEs. Boxes indicate the median value, the 25th and 75th percentiles. Whiskers show the minimum and maximum points up to 1.5 IQR away from Q1 and Q3. **B**. Difference in the percentage of bases bound by TFs in naked versus native genomic DNA across different percentiles of Jaccard indexes for TE-genes within full-length or truncated TEs. Slopes are significatively different (ANCOVA interaction term > 0, $P = 0.0446$). **C**. Metaplot of DNA methylation in three contexts around HSFA2 TFBS within *ATCOPIA78* copies for wild type and *ddm1* plants

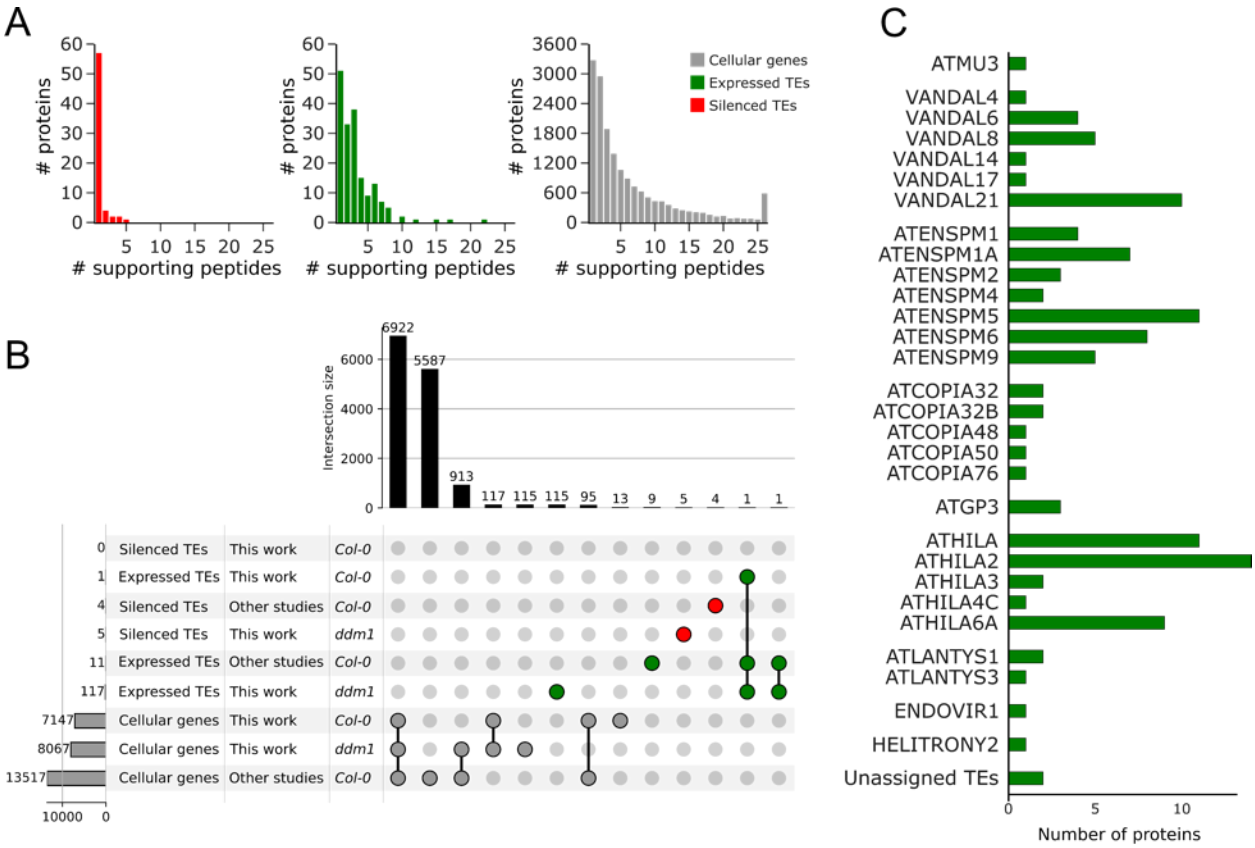


Supplementary Figure 9. Conserved domain architectures, related to Figure 3. Main conserved domain architectures for Class I (left) and Class II (right). Point size indicates the number of TE elements harboring at least one protein with a given architecture. Domains of Unknown functions (DUF) are highlighted in red.

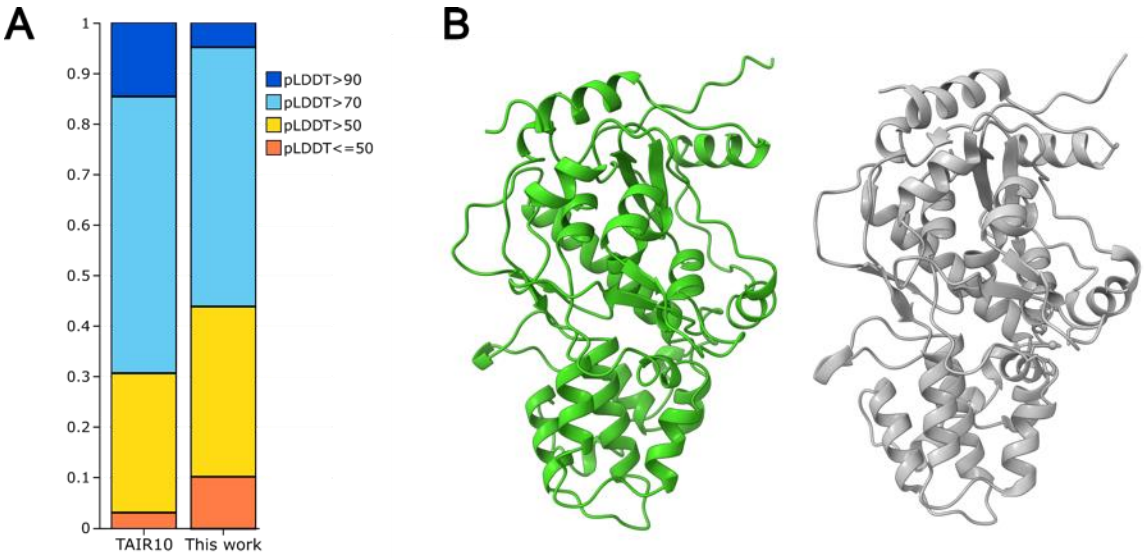


Supplementary Figure 10. TE expression and translation, related to Figure 3. A. Ribo-Seq coverage in the forward and reverse strand. Each dot represents a cellular of TE gene in either wild type or *ddm1* background. Density plots are included for each axis. **B**. Number of expressed TEs (full-length and truncated) across different tissues and organs. Various methods to quantify expression from RNA-seq data (TPM or RPKM) as well as different expression thresholds to call expressed TEs are indicated. **C**. Upset-plot of the overlap of expressed TEs (TPM> 5) across the different samples. Only the 60 largest sets are shown **D**. Percentage of proteins detected by LC-MS proteomics in different samples of wild-type plants relative to the total number of proteins detected. Proteomic data across tissues was obtained from (Mergner *et al.,* 2020). **E**. Percentage of TE- (orange and red) and cellular genes-encoded proteins (gray) detected in by LC-MS proteomics in relation to the expression levels (Transcripts per kilobase million; TPM) of the encoding transcript in Cell culture 3 days and *ddm1* samples. **F**. Average expression

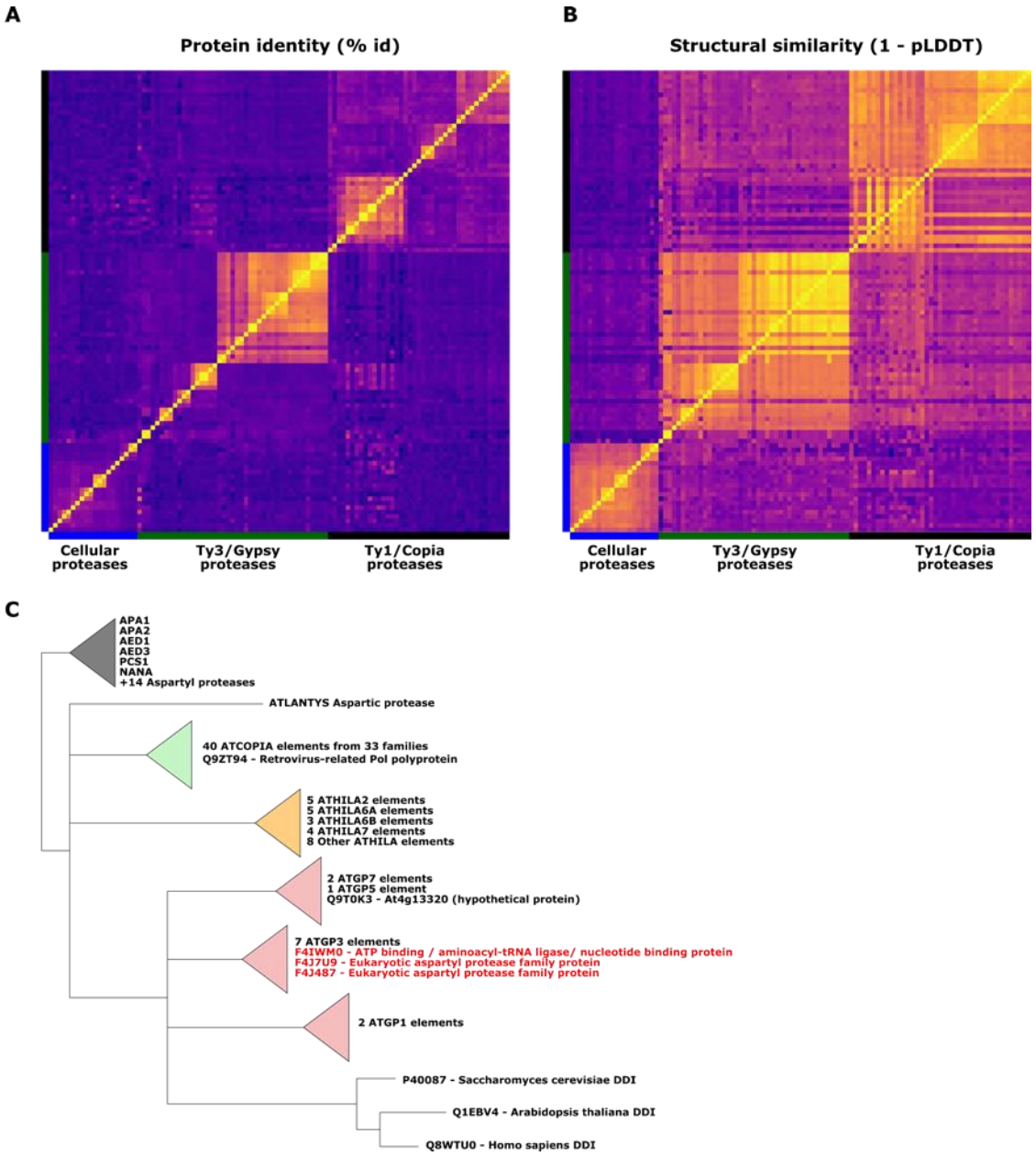
(TPM) calculated from long read cDNA sequencing of pollen and leaves on different sets of TE and cellular genes. From the TEs reported by Slotkin *et al.* 2009, only those with a reciprocal coverage of 50% with TE-genes as defined in the present work were included (99% of all TEs reported by Slotkin *et al.* 2009). TEs exclusively expressed (TPM>5) on either pollen or a set of leaf tissues (proximal leaf, distal leaf and cauline leaf), derived from the transcriptomic data generated by Mergner *et al.,* 2020 and reanalyzed in this work (Supplementary Figure 10C), as well as a set of 5,000 randomly selected genes and TEs, were also included. Mann-Whitney U tests indicate that the subset of TEs previously reported as being pollen-induced (Slotkin et al 2009) have comparable expression levels in leaves and pollen. Boxes indicate the median value, the 25th and 75th percentiles. Whiskers show the minimum and maximum points up to 1.5 IQR away from Q1 and Q3. Points beyond this range are marked as outliers.



Supplementary Figure 11. Comparison of proteomic analyses between cellular and TE proteins, related to Figure 3. A. Distribution of number of peptides supporting each protein detected by both DIA and DDA proteomics performed on either wild type or *ddm1* plants. Only proteomic datasets produced in this study were included. TE proteins derived from our functional TE annotation were classified as expressed TEs. All 6-ORF translations from the remaining TEs were classified as silenced-TE proteins. **B.** Comparison of the number of high-quality proteins (supported by at least two independent peptides) identified in this study, Mergner *et al*., 2020 and Zhang *et al*., 2019. Colors legend matches Supplementary Figure 11A. **C.** Number of high-quality proteins detected per TE family.

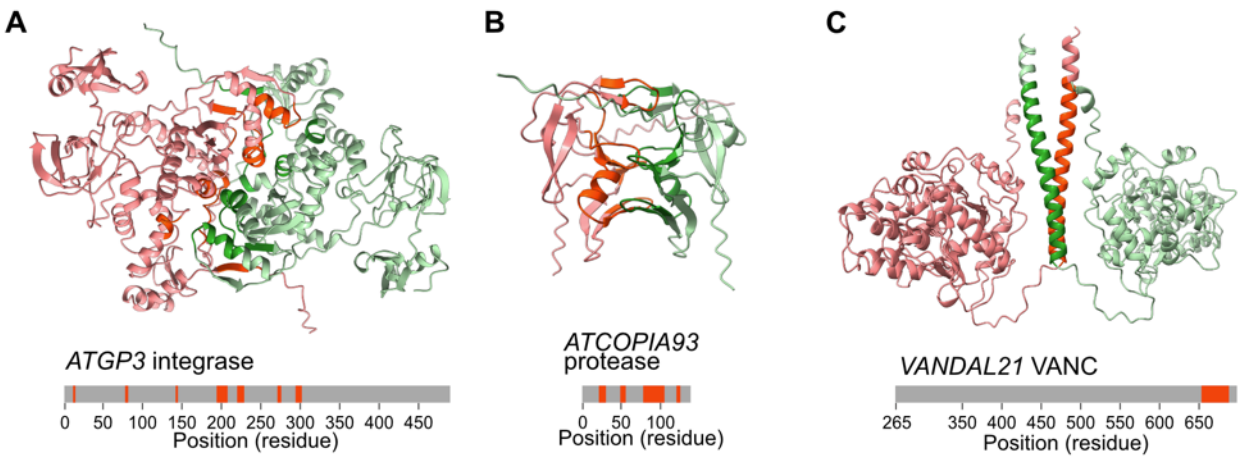


Supplementary Figure 12. pLDDT distribution, related to Figure 4. A. pLDDT distribution of modeled TE-encoded proteins and the Arabidopsis reference proteome available at AlphaFold Protein Structure Database. **B.** Predicted (left) and resolved (right) protein structure for the VANDAL21 VANC protein. Resolved protein corresponds to the RCSB PDB entry 9LNK.

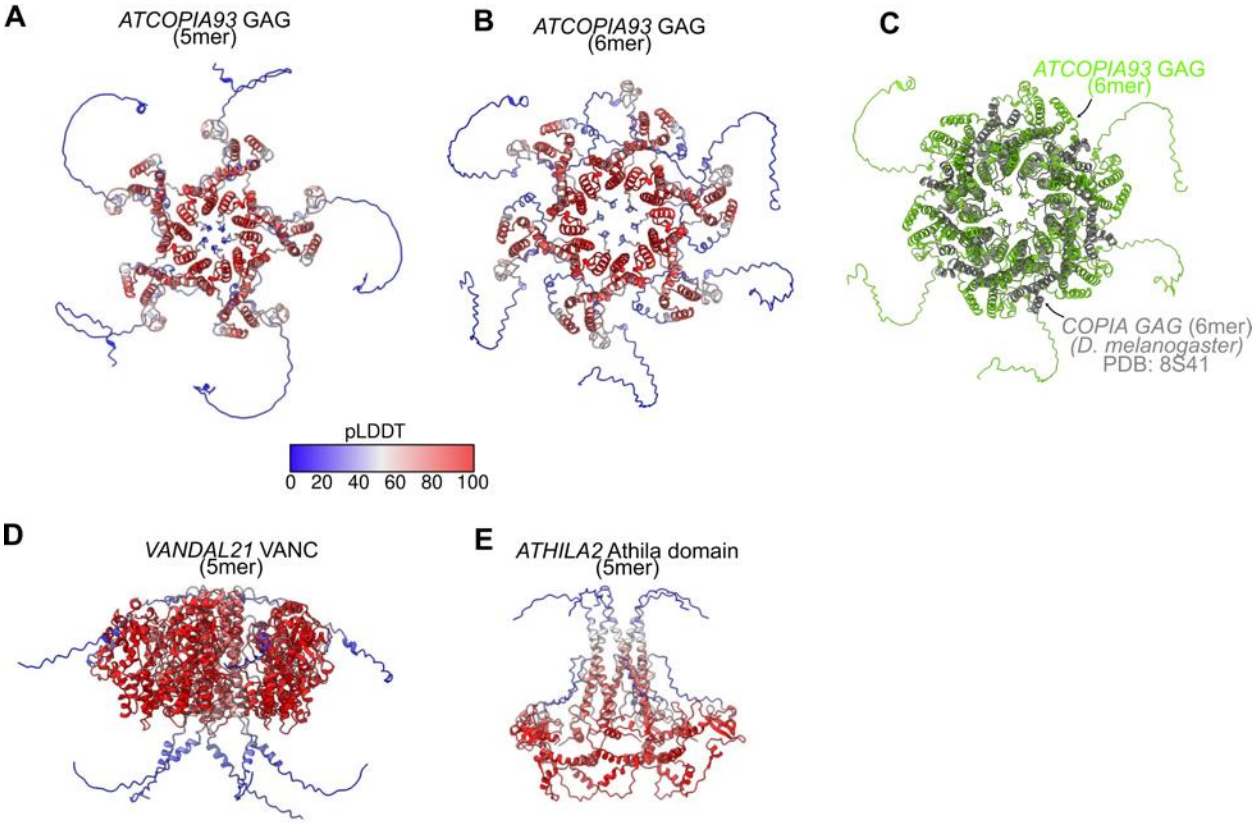


Supplementary Figure 13: Aspartic protease structural domains, related to Figure 4. A. Sequence identity measured between every pair of proteins with structural similarity to aspartic proteases, determined from neighborhood in the UMAP projection. **B**. Structural similarity measured between the same pairs of proteins. **C**. Phylogenetic tree of the proteins included in A and B, together with yeast and human DDI1 homolog. Cellular monomeric aspartic proteases are highlighted in red.

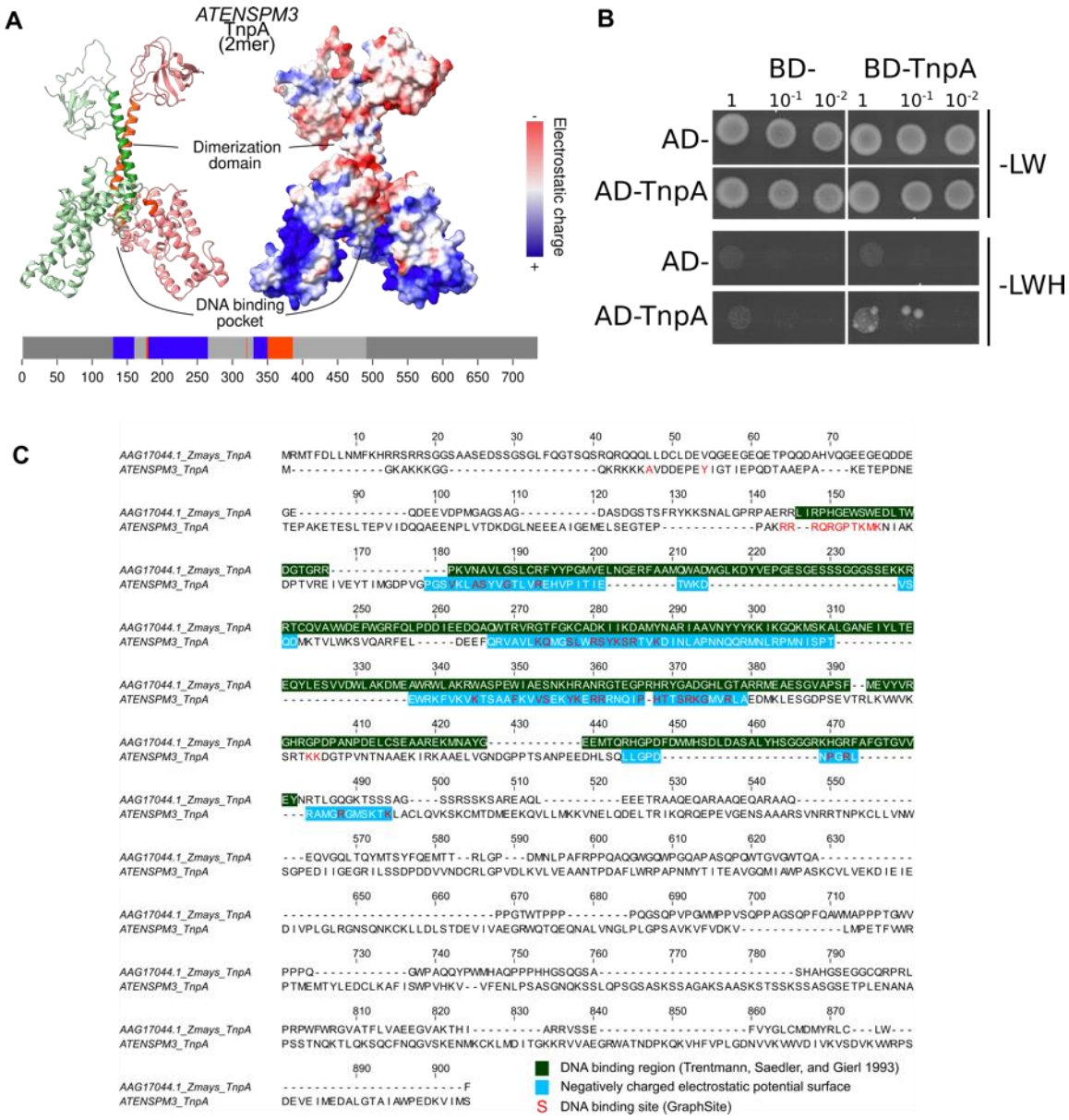
Supplementary Figure 14: TnpA conserved domain prevalence in plants, related to Figure 4. Sunburst of prevalence of conserved domains associated with TnpA main structural domain (Transposase_24, PF03004), TnpA beta barrel (Transposase_23, PF03017) and Nodulin homeobox domain (IPR039325). The concentric circles represent the phylogenetic distribution of each domain across major taxonomic levels, from inner to outer: domain, kingdom-level group, phylum, and class



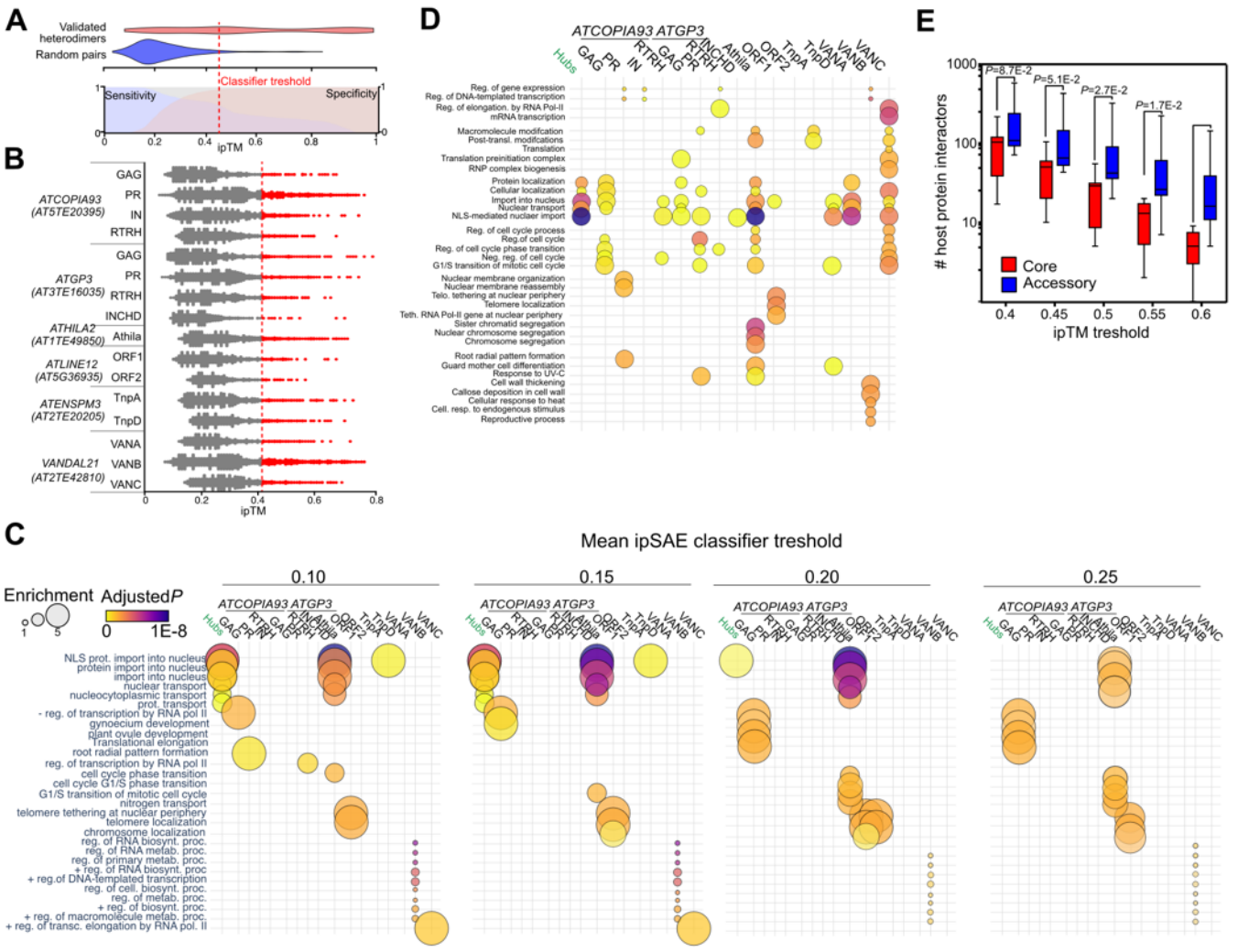
Supplementary Figure 15: TE-encoded protein predicted dimers, related to Figure 5. A. Predicted dimer conformation of *ATGP3*-encoded integrase. Interacting residues are marked in orange in the diagram below. **B**. Predicted dimer conformation of *ATCOPIA93*-encoded protease. **C**. Predicted dimer conformation of *VANDAL21*-encoded VANC protein. The disordered N-terminal region was hidden for clarity.



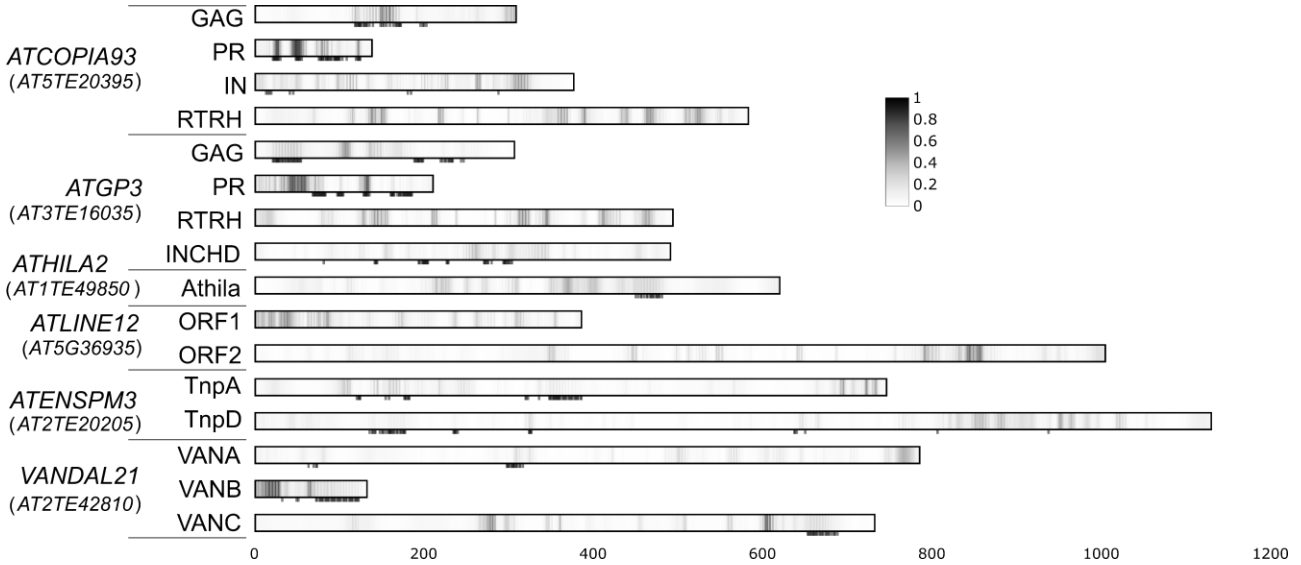
Supplementary Figure 16. Multimer structure for TE-encoded proteins, related to Figure 5. **A**. *ATCOPIA93* GAG pentamer structure. **B**. *ATCOPIA93* GAG hexamer structure. **C**. Overlay of predicted *ATCOPIA93* GAG hexamer and structurally resolved *D. melanogaster COPIA* GAG hexamer (PDB 8S41). RMSD=1.278 Å across 223 atom pairs. **D.** *VANDAL21* VANC pentamer structure. **E.** *ATHILA2* Athila-exclusive domain pentamer structure. Local confidence structure prediction (pLDDT) is shown for A, B, D, and E.



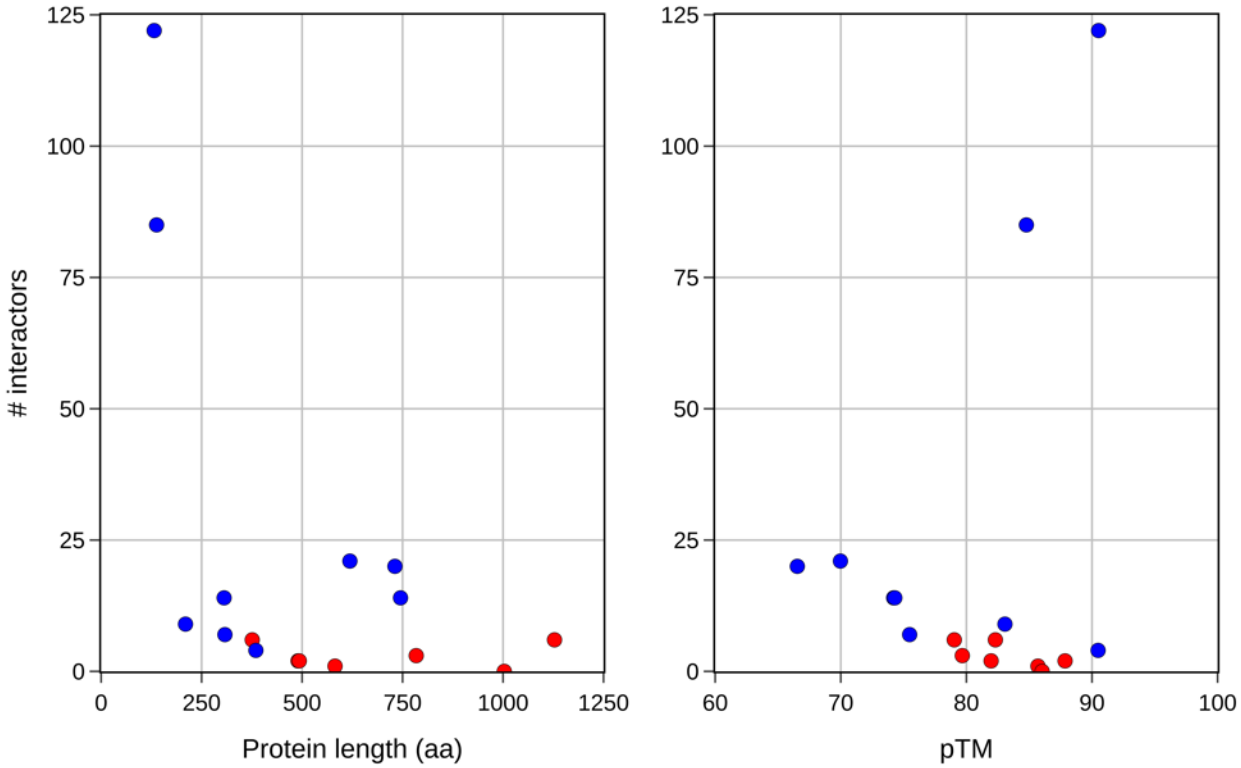
Supplementary Figure 17. Characterization of *EN/SPM3* TnpA, related to Figure 5. A. Predicted dimer conformation of *ATENSPM3*-encoded TnpA protein and its electrostatic surface. Residues mediating the dimerization or located in the predicted DNA binding pocket are shown in orange and blue, respectively. The disordered N-terminal tail and two β-barrel domains are hidden. **B**. Yeast two-hybrid assay for the homodimerization of *En/Spm3*-encoded TnpA. Activating domain (AD) and binding domain (BD) empty vectors were used to generate the negative controls. After serial dilutions (OD600nm = 5, 0.5 and 0.05, left to right) and 7 days of incubation at 30°C, growth of double transformants is shown on synthetic medium without leucine and tryptophan (-LW), or without leucine, tryptophan, and histidine (-LWH). **C**. Protein alignment of the modeled TnpA from *ATENSPM3* (AT2TE20205) and a previously reported *Zea mays* TnpA (AAG17044). DNA binding regions predicted by alternative methods are shown.



Supplementary Figure 18. TE-host PPI enrichments, related to Figure 6. A. Distribution of ipTM scores for experimentally validated versus random matched protein pairs. A threshold of 0.42, providing over 95% specificity and 61% sensitivity, was selected as the cutoff for confident PPIs. **B**. Distribution of ipTM scores obtained from AlphaFold-Multimer screens of 16 representative TE-encoded proteins against 4,164 host-encoded proteins. **C**. Gene Ontology enrichment analysis of predicted host proteins interacting with each individual TE protein, as well as host interaction hubs, detected at different ipSAE thresholds. **D.** Gene Ontology enrichment analysis of predicted host proteins interacting with each individual TE protein, as well as host interaction hubs, detected based on ipTM (ipTM threshold>0.42). **E**. Number of predicted PPIs per TE-encoded protein as a function of ipTM threshold. Boxes include one point per TE protein, with 7 core proteins and 9 accessory proteins as in Figure 6D. They display the median value, the 25th and 75th percentiles; whiskers extend to the maximum and minimum value.



Supplementary Figure 19. PPI-involved regions among TE-encoded proteins, related to Figure 6. Per residue interaction potential for 16 TE-encoded proteins. Color intensity indicates the proportion of TE-encoded and cellular heterodimer models in which each residue in the TE-encoded has at least an interacting residue in the cellular protein. Black marks below each heatmap indicate residues involved in TE-encoded protein homodimerization.



Supplementary Figure 20. Number of PPIs per TE-encoded protein against their protein length or prediction quality, related to Figure 6. Core and accessory TE proteins are marked in red and blue, respectively.