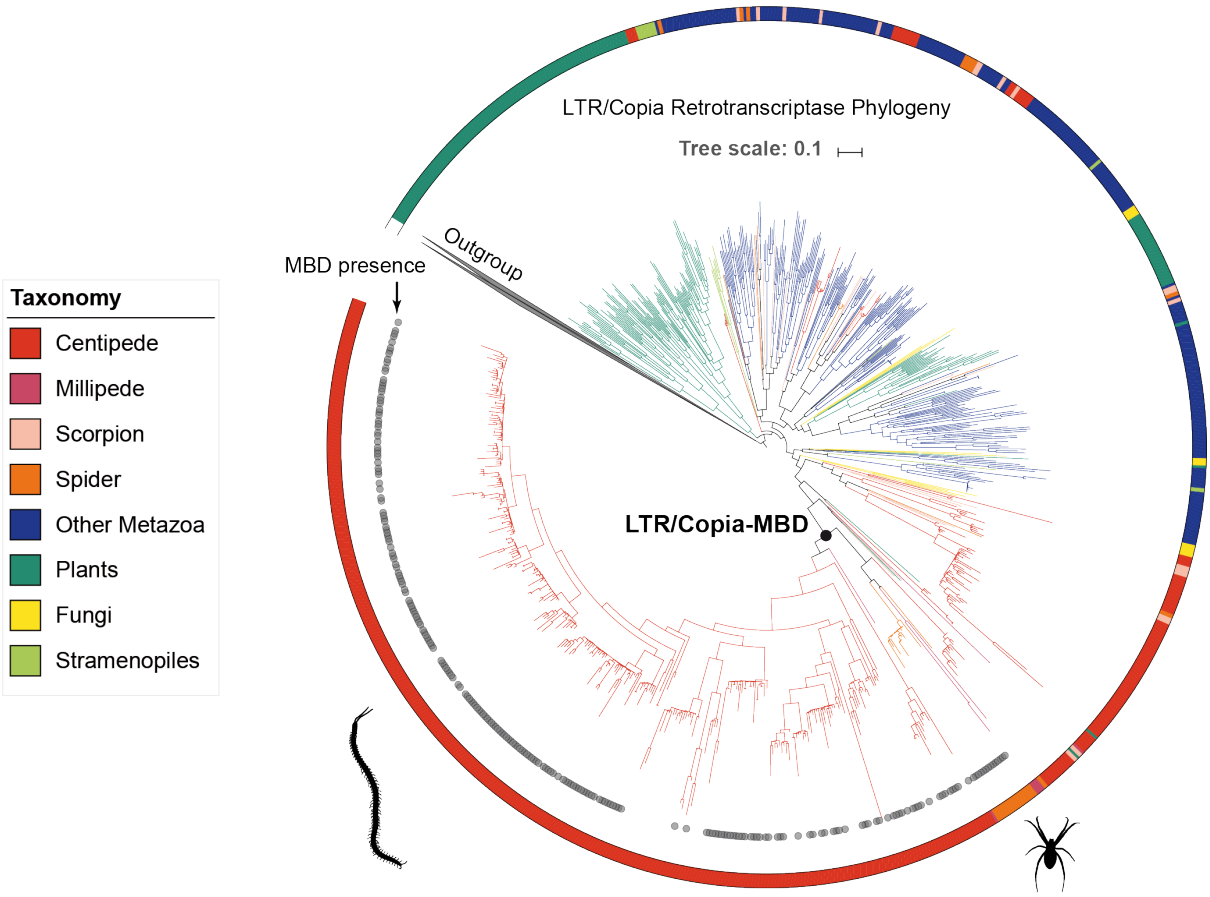
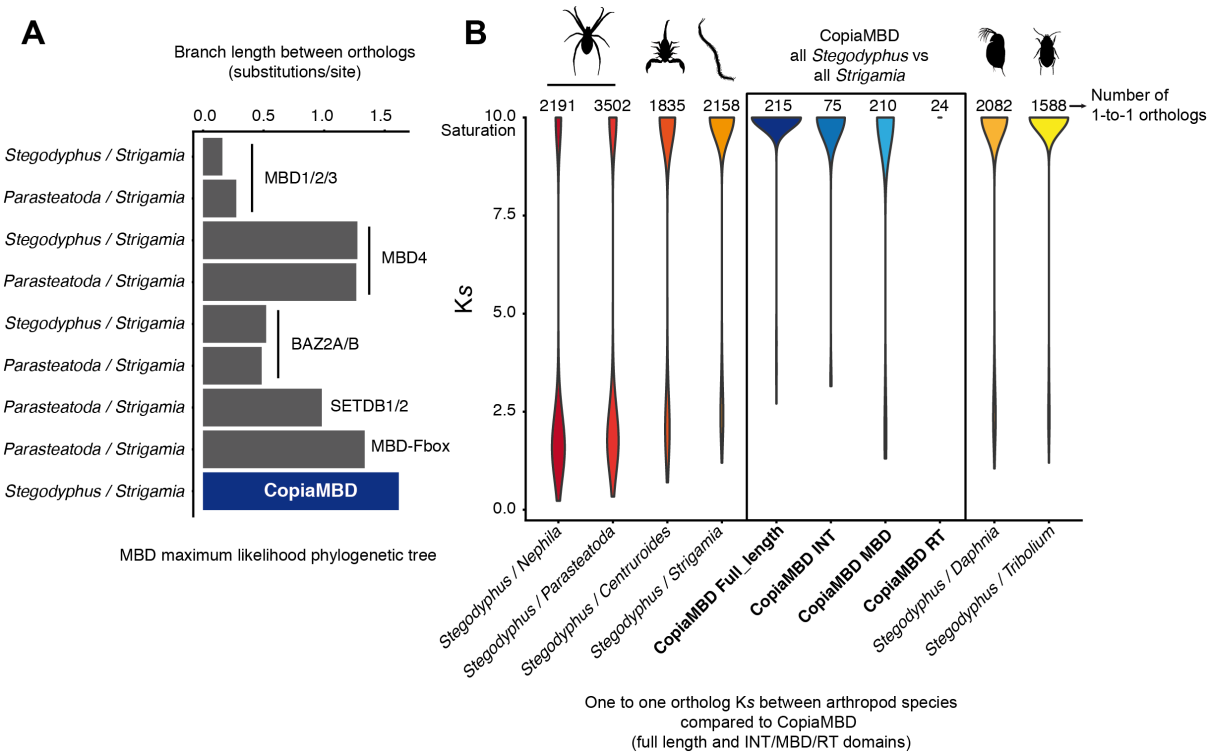
Supplemental figures for:

Capture of a functionally active Methyl-CpG Binding Domain by an arthropod retrotransposon family

Alex de Mendoza, Jahnvi Pflueger, Ryan Lister

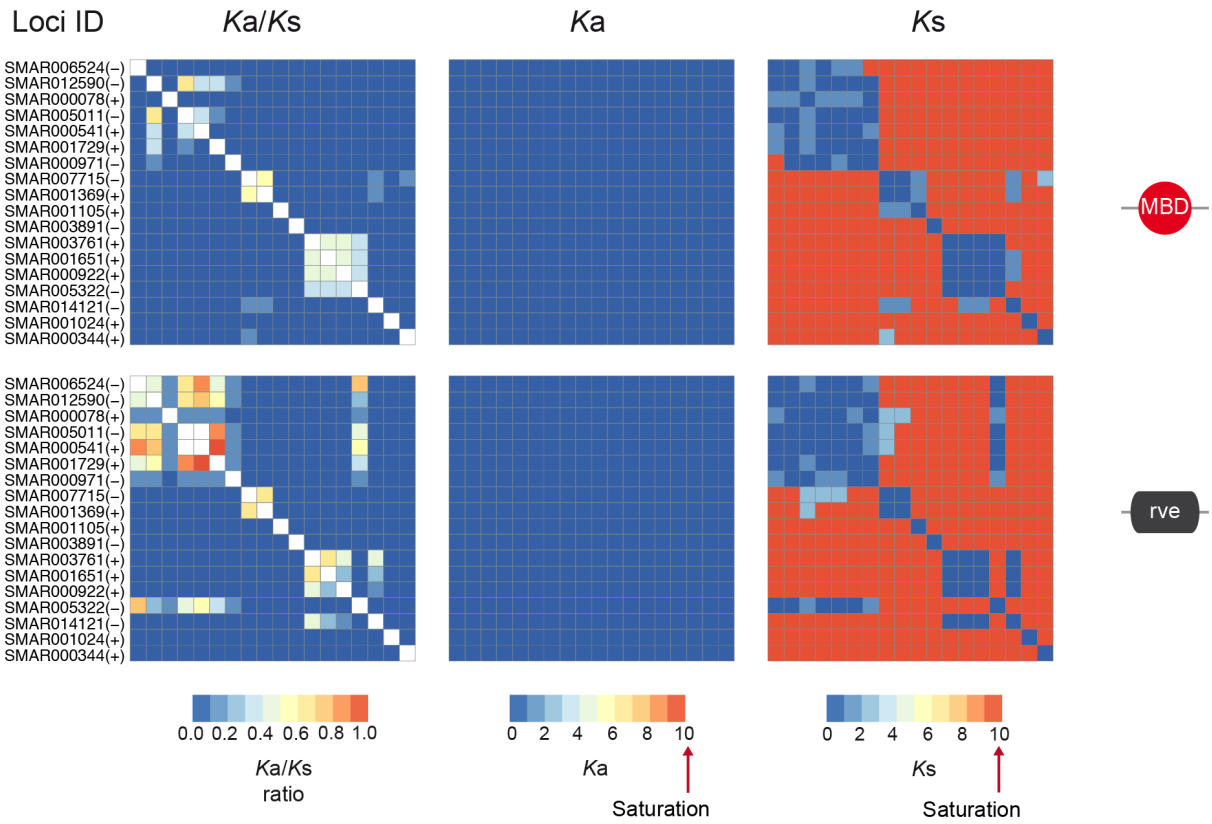


Supplemental Figure 1. The MBD domain was acquired once in the history of Copia retrotransposons. Maximum likelihood phylogenetic tree of Copia Retrotransposons based on the reverse transcriptase (RVT_2) domain (not including the MBD). Outer circle shows taxonomic affiliation of each sequence as color coded in the legend. An outgroup of non-Copia LTR reverse transcriptases have been used to root the tree. Grey dots indicate which reverse transcriptase sequences co-occur with an MBD domain.

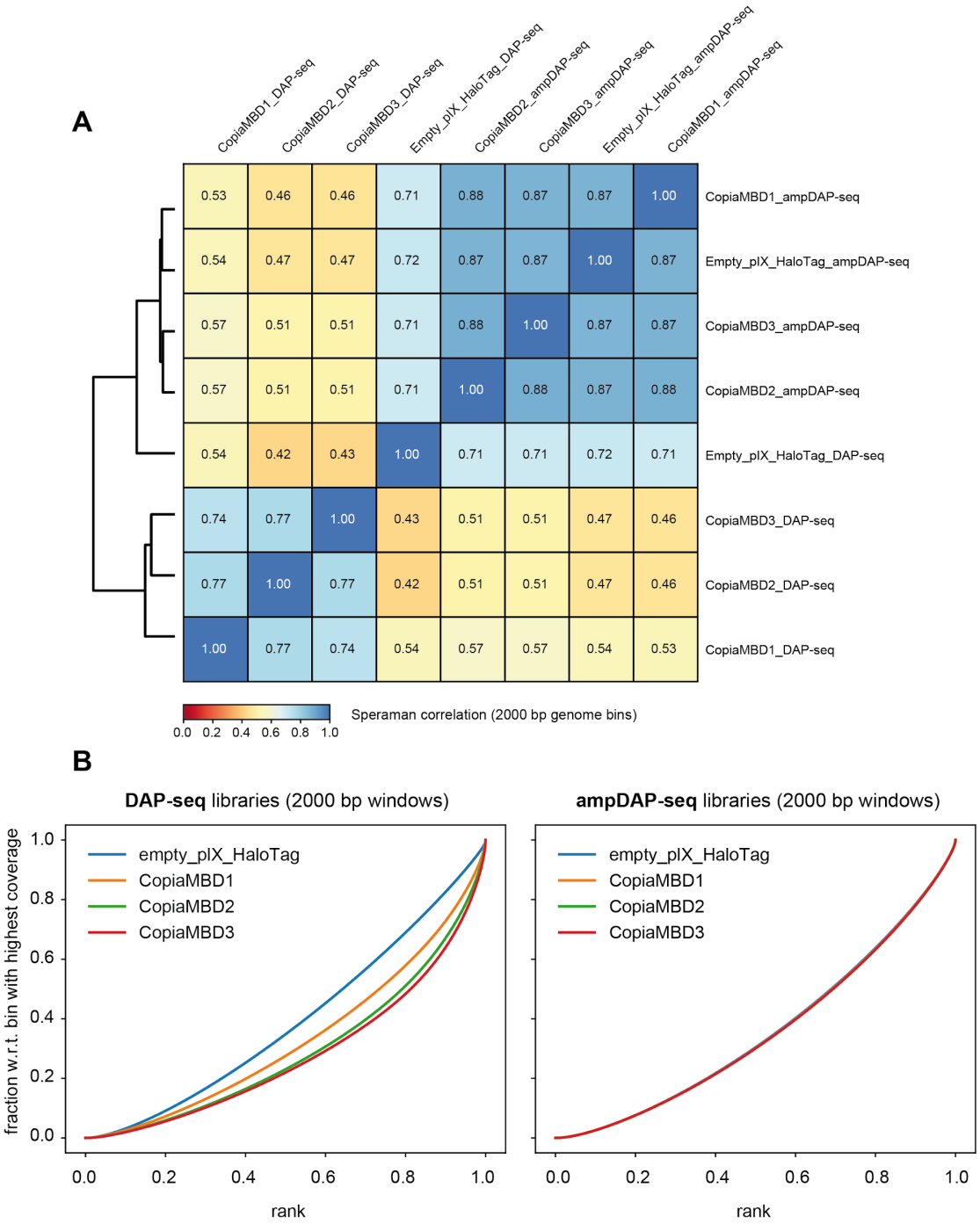


Supplemental Figure 2. CopiaMBDs do not show evidence for a recent horizontal transfer event. (A) Branch length distances from a maximum likelihood phylogeny between spider and *Strigamia* conserved MBD encoding families, as well as between CopiaMBD *Stegodyphus* and *Strigamia* clades. The sequences from the spider *Parasteatoda* have been introduced to show reproducibility of branch lengths between *Stegodyphus* and *Strigamia* sequences. (B) Distribution of synonymous substitution (K_s) rates across conserved one to one BUSCO orthologs between *Stegodyphus* and other arthropod genomes as well as between *Stegodyphus* and *Strigamia* CopiaMBD sequences.

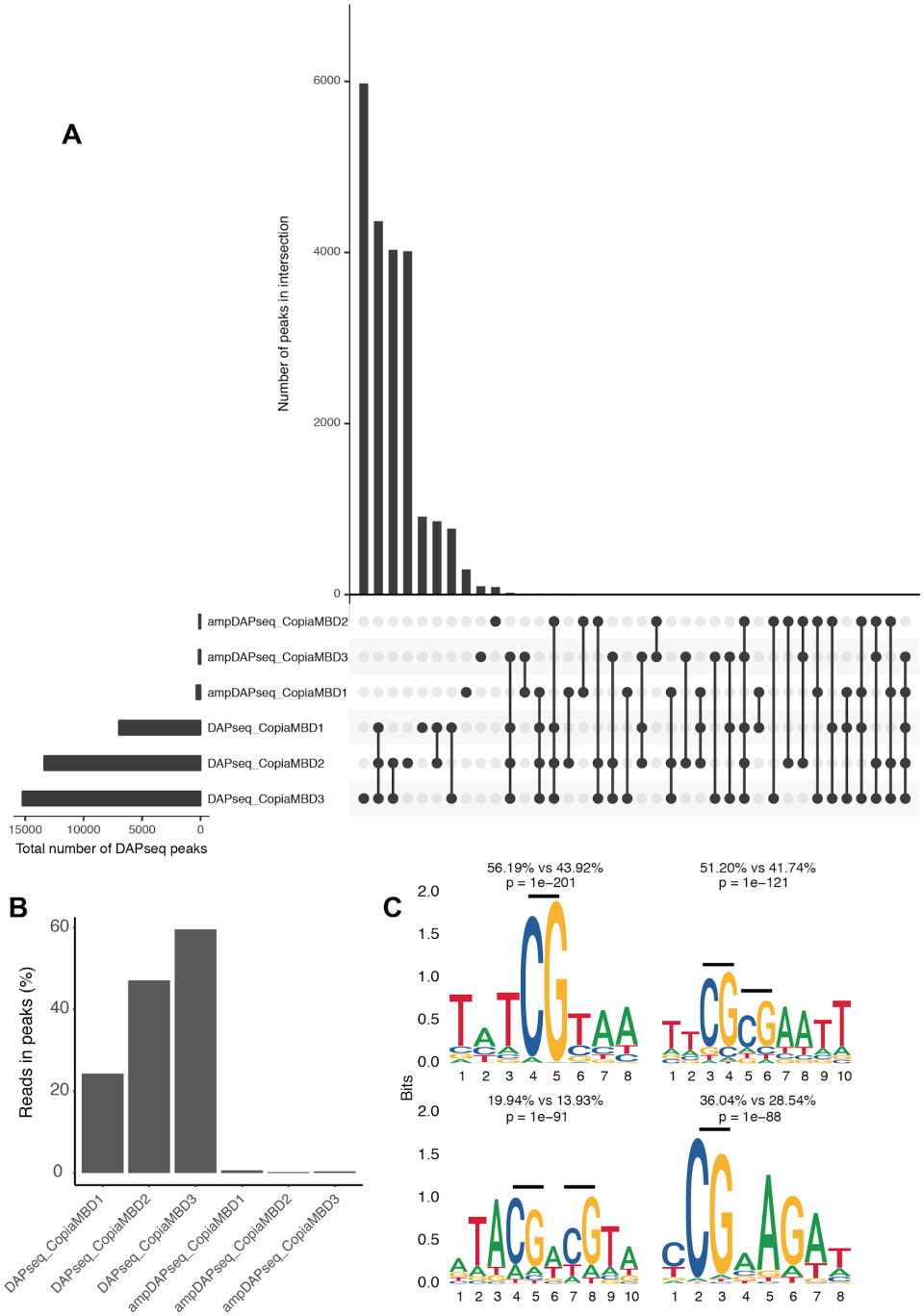
Strigamia **CopiaMBD**



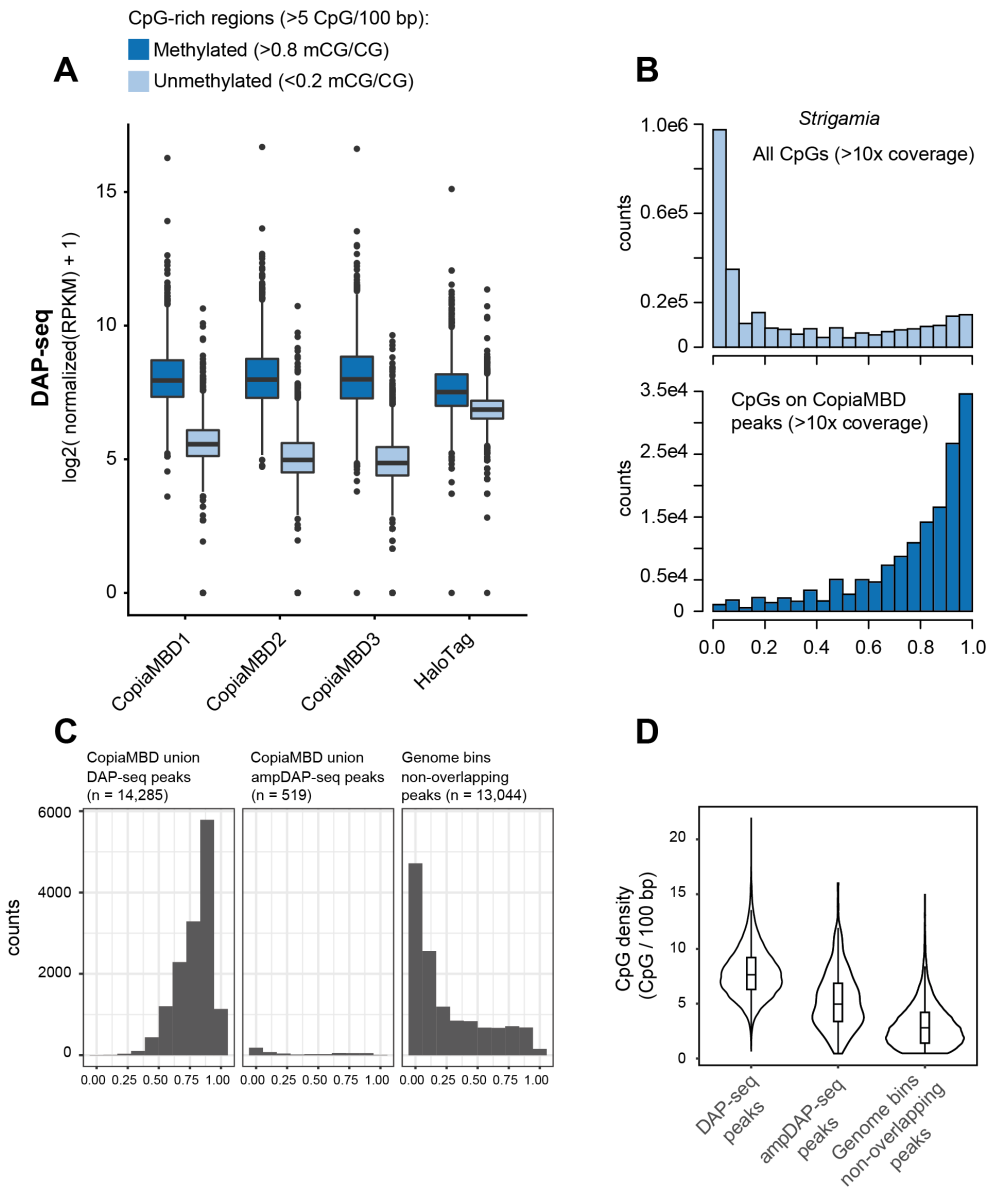
Supplemental Figure 3. Heatmaps showing all-versus-all pair-wise codon alignments between MBD and Integrase (rve) domains belonging to *Strigamia* Copia-MBD retrotransposons. Ordering of the sequences from MBD and Integrase domains belonging to the same retrotransposons is matched across heatmaps. K_s and K_a values close to 10 represent saturated substitution rates. Self-comparisons of K_a/K_s ratios are shown in white across the diagonal. Neutral selection is estimated as $K_a/K_s \sim 1$.



Supplemental Figure 4. CopiaMBD DAP-seq samples show consistent enrichments over background. (A) Spearman correlation distance matrix between DAP-seq and ampDAP-seq samples based on genomic coverage on 2000 base pair genome bins. (B) Fingerprint plot profiles showing cumulative read coverages for each sample on reads overlapping 2000 base pair windows (bins). Empty pIX HaloTag DAP-seq and ampDAP-seq samples are the background signal. Both plots have been computed using deepTools2.

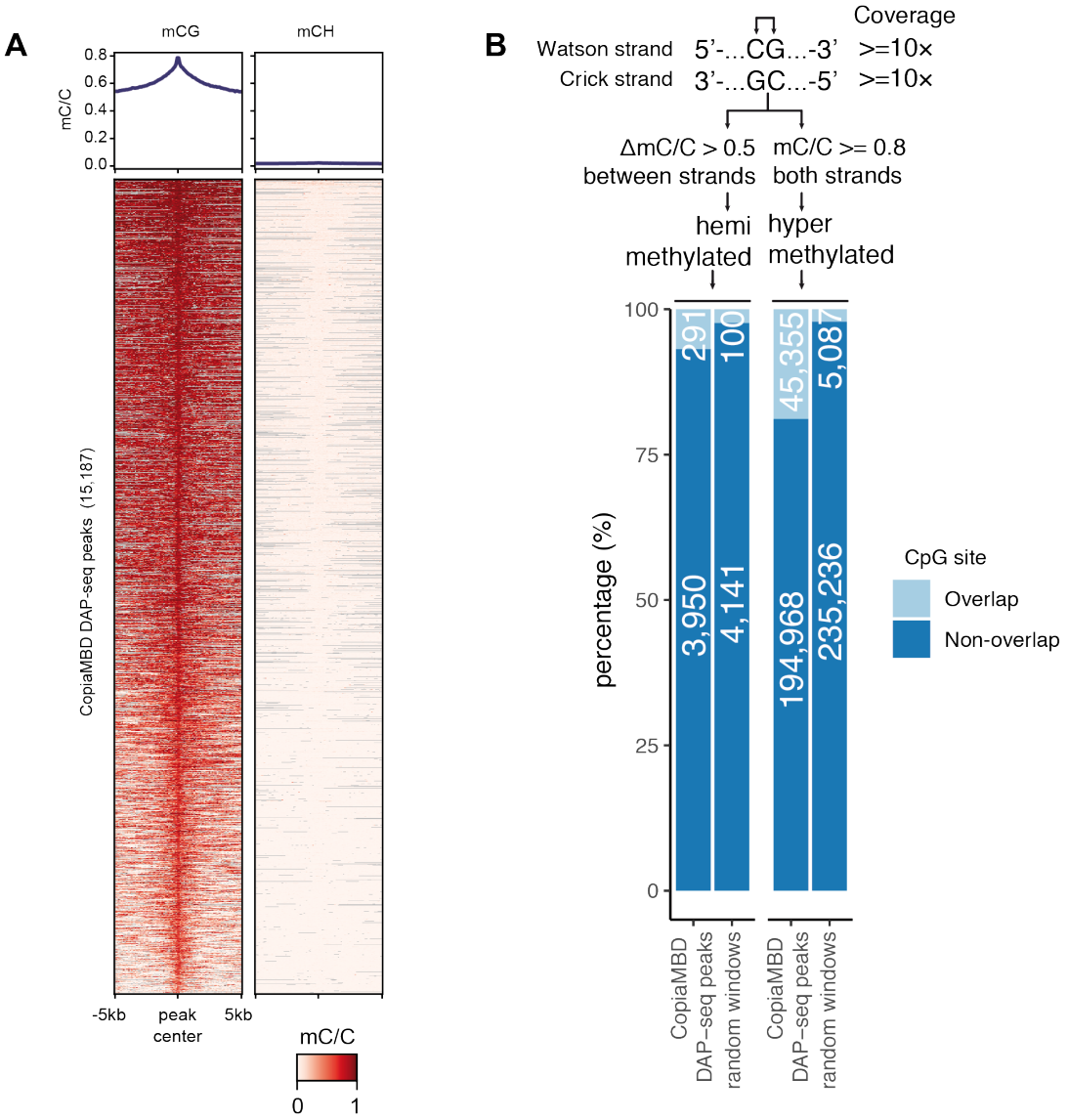


Supplemental Figure 5. CopiaMBD MBDs mostly bind to natively methylated CpGs. (A) Upset plot showing the intersection of peaks (as defined by macs2) for all DAP-seq and ampDAP-seq samples for three phylogenetically distant CopiaMBD domains (CopiaMBD1, CopiaMBD2 and CopiaMBD3). On the left is the total number of peaks per sample. (B) Fraction of total number of reads located in peaks for each DAP-seq and ampDAP-seq sample. (C) Top 4 motifs from a motif search on the intersection of peaks from the three DAP-seq CopiaMBD samples.

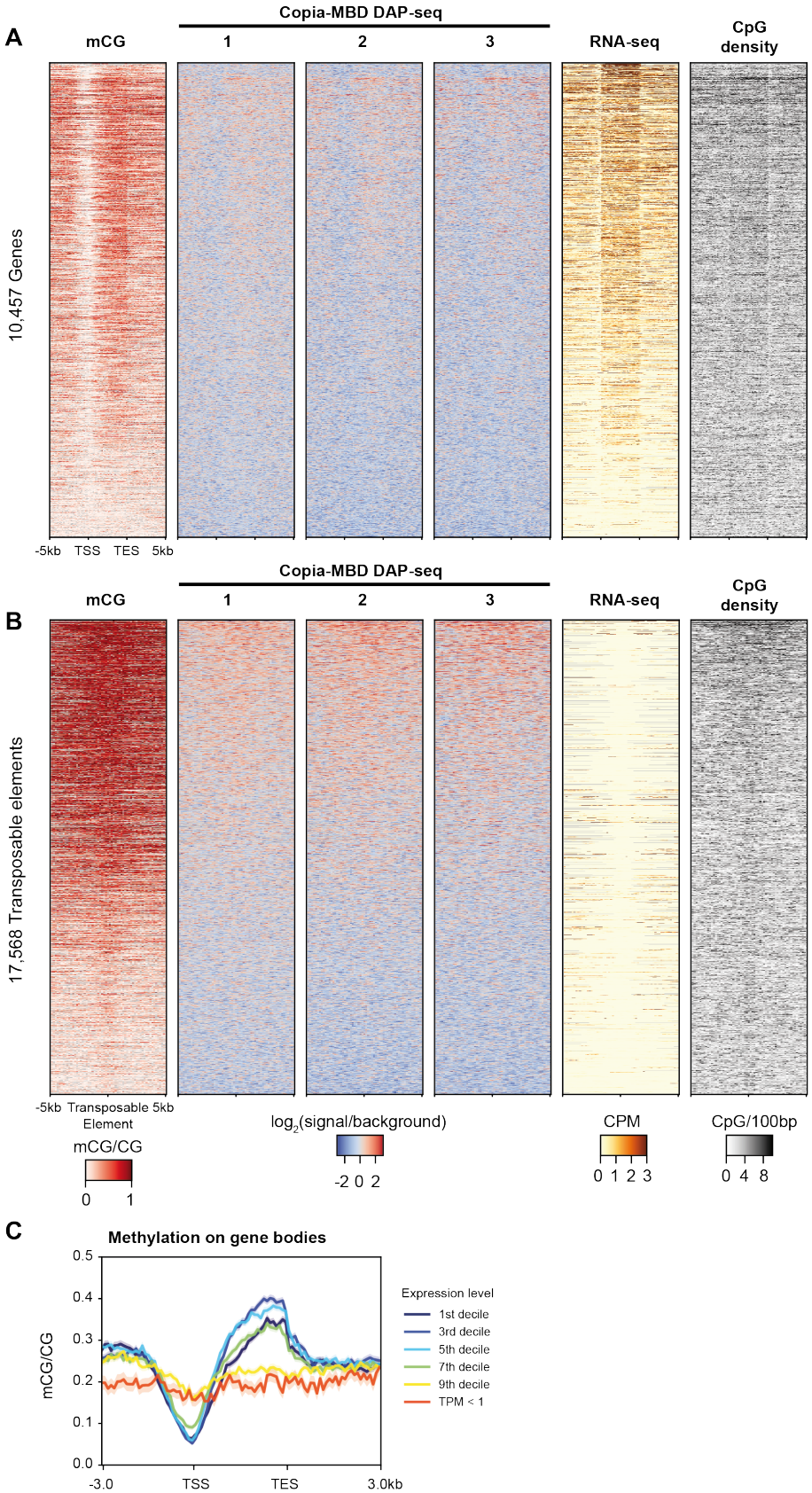


Supplemental Figure 6. CopiaMBD DAP-seq is enriched in highly methylated DNA. (A) Distribution of coverage values for DAP-seq samples on Strigamia CpG rich regions that have high and low methylated levels as assessed by WGBS. HaloTag is the background. RPKM values have been calculated using EdgeR and further normalised using the formula for Transcripts Per Million (TPM). (B) Distribution of methylation levels on CpGs across the genome and those CpGs found in the union of CopiaMBD DAP-seq peaks. (C) Distribution of aggregated methylation levels on CopiaMBD DAP-seq peaks, ampDAP-seq peaks and a set of random genome bins with the same width but non-overlapping CopiaMBD DAP-seq peaks. All regions have been filtered for a minimum WGBS coverage >4× and for

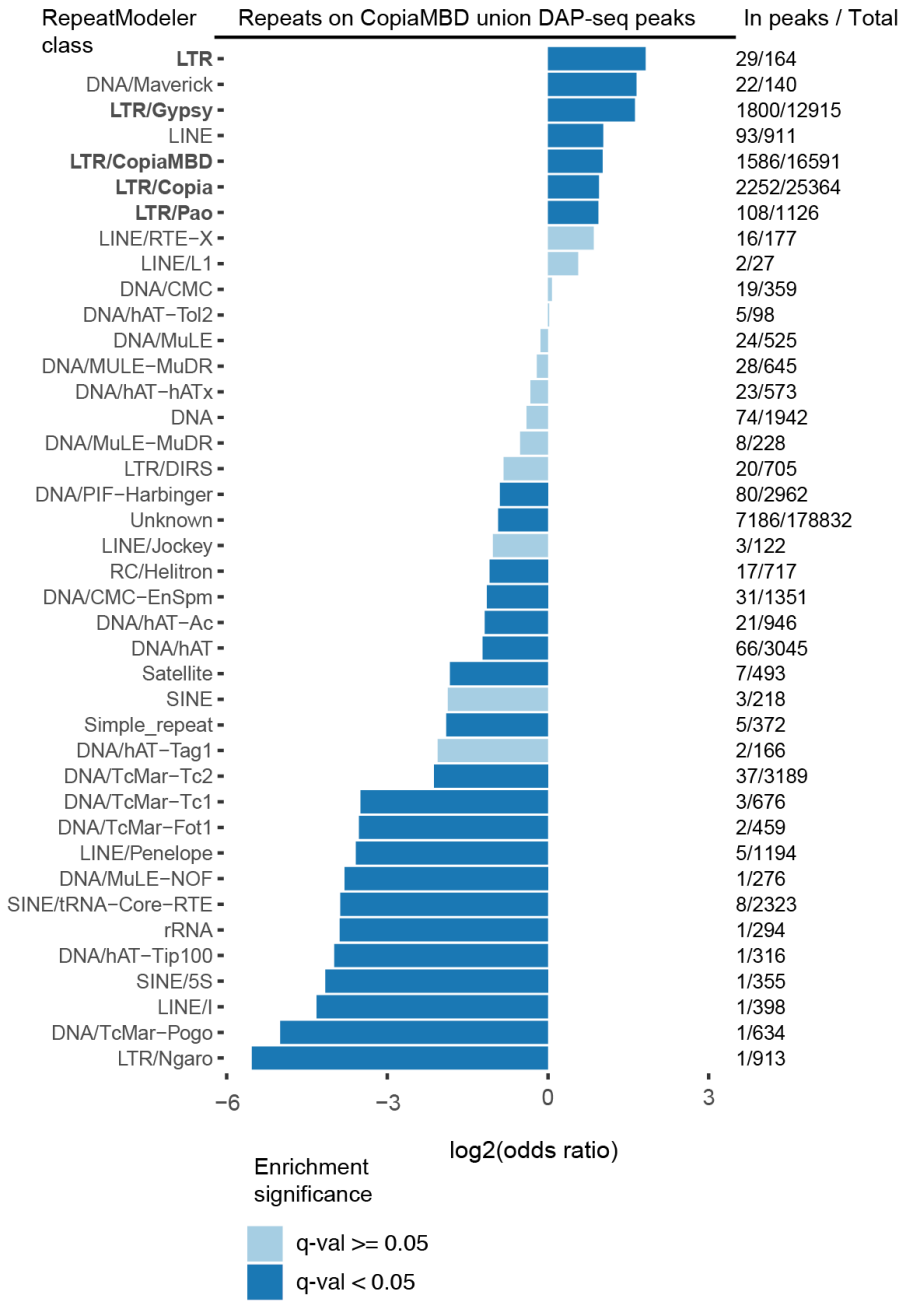
encompassing at least 1 CpG. (D) Distribution of CpG densities on the same set of bins shown in panel (C).



Supplemental Figure 7. CopiaMBD DAP-seq peaks are not enriched in non-CG methylation or hemi-methylated CpGs. (A) Heatmap showing enrichment levels of CpG methylation and non-CG methylation (CH, where H = A,T,C) on the union of CopiaMBD DAP-seq peaks. (B) Proportion of CopiaDAP-seq peak overlapping highly methylated CpGs and hemi-methylated CpGs compared to random genomic windows. CpG were required to have at least 10× coverage in each strand, and hemi-methylated positions were defined as showing a difference in methylation level higher than 0.5. Values shown on random windows represent the average overalp between the subset of genomic random windows and the CpG positions on 100 permutations.


A
mCG
Copia-MBD DAP-seq
1
2
3
RNA-seq
CpG density
10,457 Genes
-5kb TSS TES 5kb
B
mCG
Copia-MBD DAP-seq
1
2
3
RNA-seq
CpG density
17,568 Transposable elements
-5kb Transposable Element 5kb
mCG/CG
0 1
log2(signal/background)
-2 0 2
CPM
0 1 2 3
CpG/100bp
0 4 8
C
Methylation on gene bodies
mCG/CG
0.5
0.4
0.3
0.2
0.1
0.0
-3.0
TSS
TES
3.0kb
Expression level
1st decile
3rd decile
5th decile
7th decile
9th decile
TPM < 1

Supplemental Figure 8. Global profiles of CopiaMBD MBDs on genes and transposable elements. (A) Heatmap showing enrichment levels of CpG methylation, CopiaMBD DAP-seq samples, transcription and CpG density on *Strigamia* gene models filtered for mean WGBS coverage >4 and length >1000 bp. TSS Transcription Start Site, TES Transcription End Site. (B) Heatmap showing enrichment levels of CpG methylation, CopiaMBD DAP-seq samples, transcription, and CpG density on *Strigamia* transposable elements filtered for mean WGBS coverage >4. Heatmap has been centered in the middle of the transposable element. CPM Counts Per Million, RNA-seq from SRR1267275. (C) Profile of mean methylation levels on gene bodies grouped by expression level (deciles). Thick line depicts mean values and background shade depicts standard error per decile.



Supplemental Figure 9. CopiaMBD DAP-seq peaks are enriched in LTR retrotransposons. Enrichment ratios of RepeatModeler annotated repeats coloured according to significance. Two sided Fisher's exact test used to compute odds ratio enrichments and p-values, and p-values were corrected for multiple testing using Benjamini and Hochberg adjustment in R.