Text S3. Repeat annotation protocol following the advanced repeat library construction tutorial of MAKER.

To mask genomes before gene annotation, repetitive element identification and classification were performed following the advanced repeat library construction tutorial of MAKER ver. 2.31.10 (Holt and Yandell 2011). Miniature inverted transposable elements (MITEs), as well as other small (< 2Kb) class 2 nonautonomous transposable elements (TEs), were first searched with MITE-hunter (Han and Wessler 2010) using parameters by default. Then, candidate elements with long terminal repeats (LTRs) were searched with LTRharvest (Han and Wessler 2010; Ellinghaus, Kurtz, and Willhoeft 2008) in two stages consisting of detecting recent and old divergent sequences. Accordingly, we collected recent LTR elements (≥99% similarity) with terminal repeats of 100-6000 bp long, size of the entire element ranging from 1.5 kb to 25 kb, TSD (target site duplication) of 5 bp flanking the element, and within 10 of its end. In addition, LTRdigest was applied to find elements with PPT (polypurine tract) or PBS (primer binding site) using eukaryotic tRNAs sequences obtained from the Genomic tRNA database (Chan and Lowe 2016). To ensure the correct boundary of the LTR, at least 50% of the PPT of PBS sequences had to be located in internal regions and with a maximum distance of 20 bp from the LTR. Moreover, we also performed further filtering of candidate elements to reduce the chance of detecting false positives: 1) candidate elements with more than 50 unknown bases (Ns) were removed; 2) pairwise alignments of candidate elements with their 50 bp flanking sequences were made using MUSCLE ver. 3.8.31 (Edgar 2004) and, if flank sequences were also alignable (half of the total nucleotides identical with at least 60% identity excluding gaps), the candidate element was excluded; 3) LTR sequences from candidate elements retained after the steps above were used to mask the putative internal regions of other elements and, if LTR or MITE sequences were detected in internal regions, remove those elements nested with other insertions; 4) a transposase database available at the RepeatMasker ver. 3.3.0 repository (Smit, Hubley & Green at [http://www.repeatmasker.org/]), was used to also search DNA

transposons in internal regions and remove elements with significant matches (*e-value* = $1x10^{-10}$).

Redundancy in LTR elements was reduced by selecting only representative sequences. To this end, several rounds of BLASTN were employed in all-vs-all searches of LTR elements. The most representative element in each round was chosen using the greatest number of matches obtained considering 80% identity and 90% coverage cut-off thresholds, and retained in the final repeat library. Then, this representative sequence and its matches were excluded from the set of elements in the next round of BLASTN searches, leading to the generation of a second representative sequence. This selection process was repeated until no additional sequences could be found, and all representative elements were combined to form the set of recent LTR elements at 99% identity (LTR99). In the second stage, we repeated the LTR identification to recover old LTR retrotransposons by applying a 85% similarity cutoff (LTR85). The detected LTR85 sequences were masked with the representative LTR99 sequences and excluded if the match had 80% identity and 90% coverage.

As a way to create species-specific repeat libraries, we employed a de novo approach implemented in RepeatModeler ver. 1.0.11 [https://github.com/Dfam-consortium/RepeatModeler] using parameters by default. The genome was masked with the MITEs, LTR99, and LTR85 libraries to detect additional repeat sequences that were classified as ‘known’ or ‘unknown’. The new unknown elements were searched against the transposase database using BLASTX and reclassified as ‘known’ if significant matches (*e-value* = $1x10^{-10}$) to a transposon superfamily were found. All repeat sequences collected at this stage were compared to a *Drosophila* protein database downloaded from FlyBase release FB2019_01 (Thurmond et al. 2019). Elements with significant hits to genes were removed with ProtExcluder ver. 1.2 (Thurmond et al. 2019; Campbell et al. 2014). After excluding the possible gene fragments, we combined MITEs, LTR99, LTR85, RepeatModeler known, and unknown repeat sequences to generate species-specific libraries.