Method

MCHelper automatically curates transposable element libraries across eukaryotic species

Simon Orozco-Arias,[1] Pío Sierra,[2] Richard Durbin,[2] and Josefa González[1,3]

[1] *Institute of Evolutionary Biology, CSIC, UPF, 08003 Barcelona, Spain;* [2] *Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom;* [3] *Institut Botànic de Barcelona (IBB), CSIC-CMCNB, 08038 Barcelona, Spain*

The number of species with high-quality genome sequences continues to increase, in part due to the scaling up of multiple large-scale biodiversity sequencing projects. While the need to annotate genic sequences in these genomes is widely acknowledged, the parallel need to annotate transposable element (TE) sequences that have been shown to alter genome architecture, rewire gene regulatory networks, and contribute to the evolution of host traits is becoming ever more evident. However, accurate genome-wide annotation of TE sequences is still technically challenging. Several de novo TE identification tools are now available, but manual curation of the libraries produced by these tools is needed to generate high-quality genome annotations. Manual curation is time-consuming, and thus impractical for large-scale genomic studies, and lacks reproducibility. In this work, we present the Manual Curator Helper tool MCHelper, which automates the TE library curation process. By leveraging MCHelper's fully automated mode with the outputs from three de novo TE identification tools, RepeatModeler2, EDTA, and REPET, in the fruit fly, rice, hooded crow, zebrafish, maize, and human, we show a substantial improvement in the quality of the TE libraries and genome annotations. MCHelper libraries are less redundant, with up to 65% reduction in the number of consensus sequences, have up to 11.4% fewer false positive sequences, and up to ∼48% fewer "unclassified/unknown" TE consensus sequences. Genome-wide TE annotations are also improved, including larger unfragmented insertions. Moreover, MCHelper is an easy-to-install and easy-to-use tool.

[Supplemental material is available for this article.]

After two decades of sequencing projects, reference genomes for thousands of eukaryotic species are already available and many more are currently being sequenced as part of large-scale coordinated initiatives (Darwin Tree of Life Project Consortium 2022; Lewin et al. 2022). Thanks to long-read sequencing technologies, our ability to identify genomic variants has now expanded from the well-characterized single-nucleotide polymorphisms and short indels, to structural variants, which have been shown to contribute more diversity at the nucleotide level than any other type of genetic variant (Sedlazeck et al. 2018). Accurate annotation of transposable elements (TEs), a major source of structural variants, is still challenging and so far, has only been performed in a few selected groups of species (Jebb et al. 2020; Osmanski et al. 2023). However, TEs are not only present in virtually all genomes studied to date, but in some groups of species, such as mammals and plants, they are also the largest genome component. Moreover, TEs have been shown to be major contributors to the organization, rearrangement, and regulation of genomes across species (Casacuberta and González 2013; Hayward and Gilbert 2022). For example, TEs have been recruited to perform diverse biological functions in adaptive immunity, placental development, memory formation, and brain development (Huang et al. 2016; Dunn-Fletcher et al. 2018; Pastuzyn et al. 2018), while they have also been associated with autoimmune and neurological diseases, cancer, and aging (De Cecco et al. 2019; Payer and Burns 2019). Thus, the lack of accurate annotations of TE sequences results in extensive undiscovered genetic variation with potentially important implications for genome function, structure, and evolution.

While attempts at including TE diversity in the analysis of genome sequences are now common, they are often restricted to the use of homology-based methods, which are based on searching for known elements present in databases of the genome of interest, or that of a closely related species. However, the accuracy of TE identification using only homology-based methods decreases as the phylogenetic distance between the species being annotated and the species where the known elements were described increases (Platt et al. 2016). Combining homology-based and de novo TE identification in the species of interest is needed to accurately annotate TEs in genome sequences (Platt et al. 2016; Sotero-Caio et al. 2017). De novo methodologies use the repetitive nature and the structural characteristics of TE sequences to generate genome-specific libraries. Currently, there are several automatic tools available that allow de novo discovery and annotation of TE insertions, e.g., RepeatModeler2 (RM2) (Flynn et al. 2020), REPET (Flutre et al. 2011), and EDTA (Ou et al. 2019) (for a more comprehensive list visit: https://tehub.org/) (Elliott et al. 2021). However, libraries produced by these tools lead to low-quality and incomplete TE annotations (Goubert et al. 2022; Rodriguez and Makalowski 2022). This is mainly due to the high diversity of TE sequences present in genomes that preclude the generation of high-quality libraries unless manual curation of the raw libraries produced is carried out. Due to the increasing interest in TE biology, several user-friendly manuals have been recently published (Platt et al. 2016; Jamilloux et al. 2017; Storer et al. 2021b, 2022; Goubert et al. 2022; Baril et al. 2024). Additionally, some specific tools to aid in the manual curation have also been produced (Orozco-Arias et al. 2021, 2022; Goubert et al. 2022; Baril et al. 2024). Still, manual curation of TE libraries is time-consuming

Corresponding author: josefa.gonzalez@csic.es

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278821.123. Freely available online through the *Genome Research* Open Access option.

and requires acquiring knowledge on the biology of TEs present in the genomes of interest. As an example, the recent annotation of TEs in 248 mammal genome assemblies required between 1 and 19 rounds of detailed curation (Osmanski et al. 2023). Another important limitation of manual curation is the lack of reproducibility, since the curator takes decisions based on the available information but also based on his/her own expertise and experience (Baril et al. 2024). This lack of reproducibility is especially relevant for comparative genomic approaches that require consistent and accurate annotations of a large number of genomes, unlikely to be done by a single researcher.

In this work, we introduce the Manual Curator Helper, MCHelper, a tool that automates all the steps required to curate TE libraries, enabling non-TE experts to generate high-quality eukaryotic TE libraries, as well as helping experienced TE curators to do so more efficiently. MCHelper is based on published manual curation protocols developed by experts and allows for faster and reproducible TE annotations. We tested MCHelper on libraries generated by three de novo TE tools, RM2, EDTA and REPET, which differ in their output formats, and in six species, which differ in TE content and genome size: the fruit fly (*Drosophila melanogaster*), rice (*Oryza sativa*), the hooded crow (*Corvus cornix*), zebrafish (*Danio rerio*), maize (*Zea mays*), and human (*Homo sapiens*).

Results

MCHelper integrates protocols to automatically curate transposable element libraries

MCHelper is a computational tool that integrates TE manual curation protocols developed by TE experts to improve the completeness and accuracy of TE libraries generated by automatic de novo tools (Platt et al. 2016; Jamilloux et al. 2017; Ou and Jiang 2018; Storer et al. 2021b, 2022; Goubert et al. 2022; Baril et al. 2024). MCHelper is a flexible tool designed to accept libraries generated by de novo TE identification tools that produce a FASTA file output, such as RM2 (Flynn et al. 2020) or EDTA (Ou et al. 2019), as well as the output files produced by the TEdenovo pipeline from REPET (Fig. 1; Flutre et al. 2011). MCHelper was designed to reduce the most common problems present in libraries generated with automatic tools: redundant, fragmented, false positive, and unclassified/unknown TE sequences. Briefly, MCHelper: (1) reduces consensus redundancy, once at the beginning and once at the end of the pipeline, using either the CD-HIT (Li and Godzik 2006) or the MeShClust V3 (Girgis 2022) clustering algorithms; (2) extends the consensus sequences; (3) identifies and filters false positive consensus sequences, i.e., sequences wrongly classified as TEs (see Methods); (4) identifies previously known TEs based on homology; (5) performs structural checks on consensus sequences initially labeled by de novo TE identification tools as “classified” TEs; and (6) assigns a classification to consensus sequences initially labeled as “unclassified/unknown” TEs (Fig. 1; Supplemental Fig. S1).

MCHelper offers three distinct levels of automation: fully automatic, semiautomatic, and fully manual, thus enabling users to tailor the software’s functionality to their curation requirements and time availability (Fig. 1). In the fully automatic mode, consensus sequences that pass the structural check (complete) and those that fail (incomplete) are both kept in the curated library (see Methods). “Incomplete” sequences are kept as they could correspond to older families for which a full-length copy is no longer present in the genome analyzed. MCHelper adds the suffix “_inc” to “incomplete” sequences for easy identification by the user. In the semiautomatic mode, MCHelper automatically retains only “complete” sequences and provides graphical information, such as TE-Aid plots (Goubert et al. 2022), multiple sequence alignment (MSA) plots, and information on TE length and copy number, to facilitate the user’s manual inspection of the “incomplete” sequences. The user is then responsible for deciding whether to keep, remove, or reclassify the “incomplete” sequences. In the fully manual mode, MCHelper provides all the information needed to manually inspect all the consensus sequences.

MCHelper improves the number, quality, and length of consensus sequences in transposable element libraries across species

To test the MCHelper tool, we ran the fully automatic mode on six species that differ in genome size and TE content and that have available reference TE libraries: the fruit fly (*D. melanogaster*), rice (*O. sativa*), the hooded crow (*C. cornix*), zebrafish (*D. rerio*), maize (*Z. mays*), and human (*H. sapiens*). For each species, we ran MCHelper with the raw TE libraries obtained from three different de novo tools, RM2 (Flynn et al. 2020), EDTA (Ou et al. 2019),

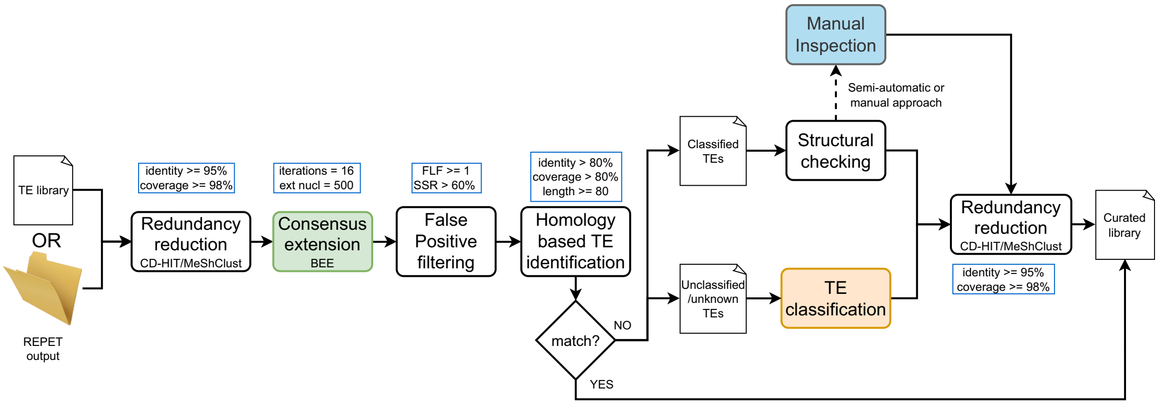

Figure 1. MCHelper’s tool workflow. Inputs, main steps, and default parameters of the MCHelper tool. MCHelper accepts as input a single FASTA file containing consensus sequences produced by a de novo TE identification tool or a folder containing files produced by the TEdenovo pipeline from REPET (see the MCHelper’s GitHub for details). Colored boxes correspond to modules that can also be executed independently. Workflows of these modules are provided in Supplemental Figure S1.

and REPET (Flutre et al. 2011), with the exception of the human T2T genome where a REPET library could not be obtained. We compared the performance of the two clustering algorithms implemented in MCHelper, CD-HIT, and MeShClust, and we found that the proportion of overlapping clusters between the algorithms is high and depends on the genome analyzed (Supplemental Fig. S2).

We then compared the MCHelper libraries (generated using CD-HIT as the clustering algorithm) with the available reference libraries: Berkeley Drosophila Genome Project (BDGP) (Kaminker et al. 2002) and Manual Curated TE (MCTE) libraries (Rech et al. 2022) for *D. melanogaster*, the manually curated "standard library" provided by Ou et al. (2019), for *O. sativa*, the MClibrary (Weissensteiner et al. 2020) for *C. cornix*, Dfam (Storer et al. 2021a) and Repbase (Jurka et al. 2005) for *D. rerio*, MTEC curated by Ou et al. (https://github.com/oushujun/MTEC) for *Z. mays*, and the Dfam library for *H. sapiens*.

MCHelper reduced the total number of consensus sequences in the raw libraries produced by RM2, EDTA, and REPET in the six species analyzed. Using the RM2 output, MCHelper reduced the number of consensus sequences between 34.3% in *O. sativa* and 54.8%, in *D. melanogaster* (Fig. 2A; Supplemental Table S1A). Additionally, MCHelper was able to remove between 3.1% (in *D. melanogaster*) and 11.4% (in *O. sativa*) of false positives in these six species (Supplemental Fig. S3; Supplemental Table S1B). Using EDTA's output, MCHelper reduced the number of families between 40.7% (in *C. cornix*) and 65.4% (in *H. sapiens*), and removed 2.4% (in *C. cornix*) and 3.9% (in *D. rerio*) of false positives. Using REPET's output, MCHelper reduced the number of families

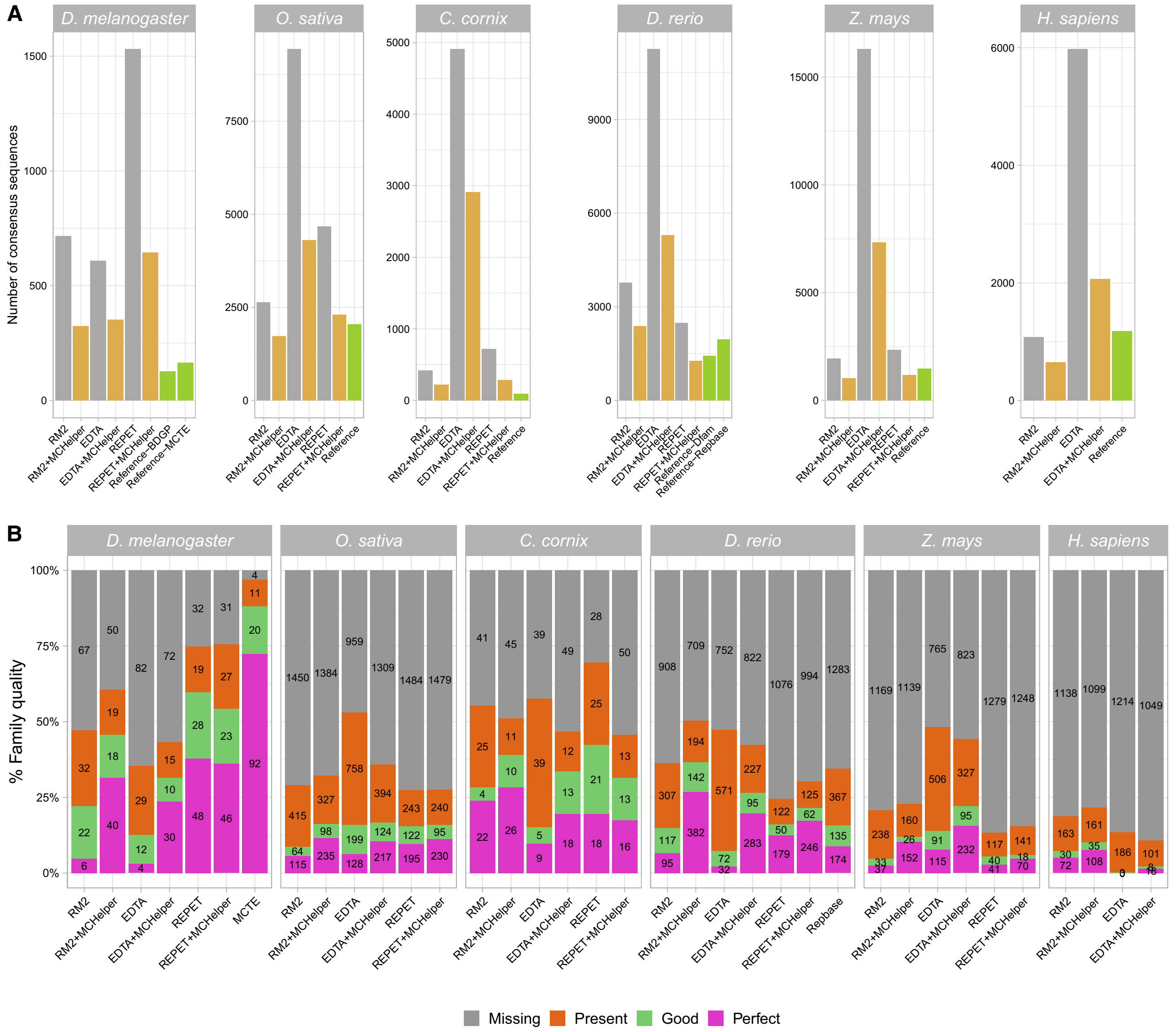

Figure 2. Comparison of MCHelper libraries with raw and reference libraries in the six genomes analyzed. (*A*) Number of consensus sequences contained in the raw library, the MCHelper library, and the reference libraries of the six species analyzed. Reference libraries used are: BDGP (Kaminker et al. 2002) and MCTE (Rech et al. 2022) for *D. melanogaster*, standard library (Ou et al. 2019) for *O. sativa*, MClibrary (Weissensteiner et al. 2020) for *C. cornix*, Dfam (Storer et al. 2021a) and Repbase (Jurka et al. 2005) for *D. rerio*, MTEC curated by Ou et al. (https://github.com/oushujun/MTEC) for *Z. mays*, and Dfam for *H. sapiens*. (*B*) Number of consensus sequences classified as perfect, good, present, and missing families following the methodology proposed by Flynn et al. (2020). Note that for *D. melanogaster* and *D. rerio*, the MCTE and the Repbase libraries are also compared to the BDGP and Dfam libraries, respectively.

between 48.9% (in *D. rerio*) and 60.6% (in *C. cornix*), and removed between 2.8% (in *Z. mays*) and 6.7% (in *O. sativa*) of false positives. Thus overall, MCHelper libraries are more similar to the reference libraries than to the raw libraries, consistent with the findings of Storer et al. (2022), who noted that raw de novo libraries are typically twice as large as necessary.

MCHelper libraries contained up to 8.8 times more perfect families than the raw libraries (Fig. 2B; Supplemental Table S1C). Family quality was defined according to the methodology described in Flynn et al. (2020), which classifies each family contained in a reference library as perfect, good, present, or missing in the comparison library (see Methods). We also found that the majority of the MCHelper automatically curated libraries contained fewer or very similar numbers of missing families for all genomes and de novo tools (Fig. 2B).

Finally, MCHelper consistently generated longer consensus sequences than the raw libraries, thus exhibiting lengths more similar to those reported in the reference libraries (Supplemental Fig. S4; Supplemental Table S1D) across almost all major orders of TEs: LINE, SINE, LTR, and MITEs. We observed that for all the species, except for *C. cornix*, MCHelper also generated, in average, longer consensus sequences for terminal inverted repeat (TIR) elements in the raw libraries produced by 2/3 of the programs tested (2/2 in *H. sapiens*). In the case of Helitrons, MCHelper generated longer consensus sequences in the raw libraries produced by 1/3 of the programs tested in four species (*C. cornix*, *D. rerio*, *Z. mays*, and *H. sapiens*) and in 2/3 in the other two (*D. melanogaster* and *O. sativa*). This observation reinforces that raw libraries often contain fragmented TEs instead of complete consensus sequences, underscoring the importance of the consensus extension step in the curation process (Fig. 1; see Methods; Storer et al. 2021b; Goubert et al. 2022; Baril et al. 2024). However, as noted by Baril et al. (2024), longer consensus sequences do not always equate to a better representation of the TE family. To avoid overextensions and fuzzy terminations, MCHelper incorporates an approach that detects when the consensus sequence reaches either end (Supplemental Fig. S1A). We observed that the percentage of TEs that do not contain their corresponding terminal repeats (TRs), or poly (A) tail sequences decreased in MCHelper libraries in comparison with RM2's raw libraries up to 38% (in *Z. mays*), with EDTA's raw libraries up to 13% (in *D. melanogaster*), and with REPET's raw libraries up to 12% (in *Z. mays*) (Supplemental Table S2). Additionally, we tested how often chimeric TE sequences can be found in the raw libraries and in the MCHelper libraries. As a proxy to identify chimeric TE sequences, we did a BLASTN search between each library and the MCHelper's internal TE database (see Supplemental Methods) to look for those consensus sequences with hits with >80% identity and >50% coverage with consensus sequences from two or more orders. We found that the MCHelper libraries contained less chimeric TEs than the raw libraries in all species, except for *H. sapiens* in which the MCHelper library contains one chimeric sequence while the raw EDTA library does not contain any (Supplemental Table S3).

Finally, we compared the MCHelper tool with SENMAP, a tool specifically designed to manually curate full-length long terminal repeat (LTR) consensus sequences (Orozco-Arias et al. 2021). While SENMAP works better in plant genomes compared with animal genomes, as expected since SENMAP was trained with plant data, MCHelper identifies a higher number of LTR elements across tools as well as a higher number of perfect families (Supplemental Table S4A). If we focus on complete LTRs, while SENMAP identifies more LTRs from *D. melanogaster* REPET libraries compared with MCHelper, the number of perfect families identified by MCHelper is higher (Supplemental Table S4B).

MCHelper recovers up to ∼48% of the TEs originally labeled as "unclassified/unknown"

Another common issue with raw libraries is the high number of unclassified TEs, especially in nonmodel organisms due to the lack of TE consensus sequences for these species in the available TE databases. To assign a classification to previously unclassified TEs, MCHelper follows a three-step process: homology search, coding domain presence (e.g., transposase and reverse transcriptase), and terminal repeat presence (LTRs and TIRs) (Supplemental Fig. S1C; see Methods). Using this approach, MCHelper was able to assign a TE class for up to 48.24% of the TEs initially labeled as "unclassified/unknown" TEs (Fig. 3A; Supplemental Table S5A). The majority of TEs in the six genomes analyzed were reclassified in the homology step, with the REPET output showing the bigger improvement compared to RM2, with the exception of *Z. mays* (Fig. 3A). Across species and de novo TE identification tools, TEs from the LTR and TIR orders were the most commonly recovered sequences, consistent with the relative abundance of these orders in the genomes analyzed, with the exception of *H. sapiens* where LINE was more common (Fig. 3B; Supplemental Table S5B).

We also used TEClass2 (Bickmann et al. 2023), TERL (da Cruz et al. 2021), and DeepTE (Yan et al. 2020) to classify the "unclassified/unknown" TEs (see Methods) to compare the performance against MCHelper and we found that although the recovering proportion was smaller in MCHelper compared with the other tools, the MCHelper precision was generally higher (Supplemental Table S6).

Genome-wide TE annotations with MCHelper libraries are more accurate

Genome-wide TE annotations performed with raw TE libraries often lead to an overestimation of the number of TE copies, mostly because a single copy is annotated as several fragmented copies (Jamilloux et al. 2017). In addition, raw libraries may have chimeric consensus sequences leading to the annotation of a single TE copy by multiple consensus sequences (Rodriguez and Makalowski 2022). To evaluate the quality of the annotations obtained with the different TE libraries (raw, MCHelper, and reference), we analyzed the total number of copies annotated, the number of short copies annotated (<100 bp), and the number of overlapping annotations (copies annotated with more than one consensus).

We found that the number of TE copies annotated with the MCHelper library was smaller than the number of copies annotated with the raw libraries for the majority of tools and species analyzed (Fig. 4A; Supplemental Table S7A). The reduction in number of copies ranged between 9.9% (in *D. rerio*) and 37.8% (in *Z. mays*) compared with the RM2 libraries, between 24.5% (in *H. sapiens*) and 48.8% (in *Z. mays*) compared with EDTA libraries and between 0.2% (in *O. sativa*) and 16.5% (in *Z. mays*) compared with REPET libraries. The only exceptions were the *C. corvix* genome, in which the number of annotated copies by MCHelper libraries was slightly bigger, and the annotation of the human genome with the raw RM2 library.

The number of short copies (<100 bp) was also reduced in MCHelper annotations, getting closer to the number of short copies in reference annotations in the six species when comparing with the raw RM2 annotations (between 10.8% in *D. melanogaster* and 39% in *Z. mays* fewer short copies) (Fig. 4B; Supplemental

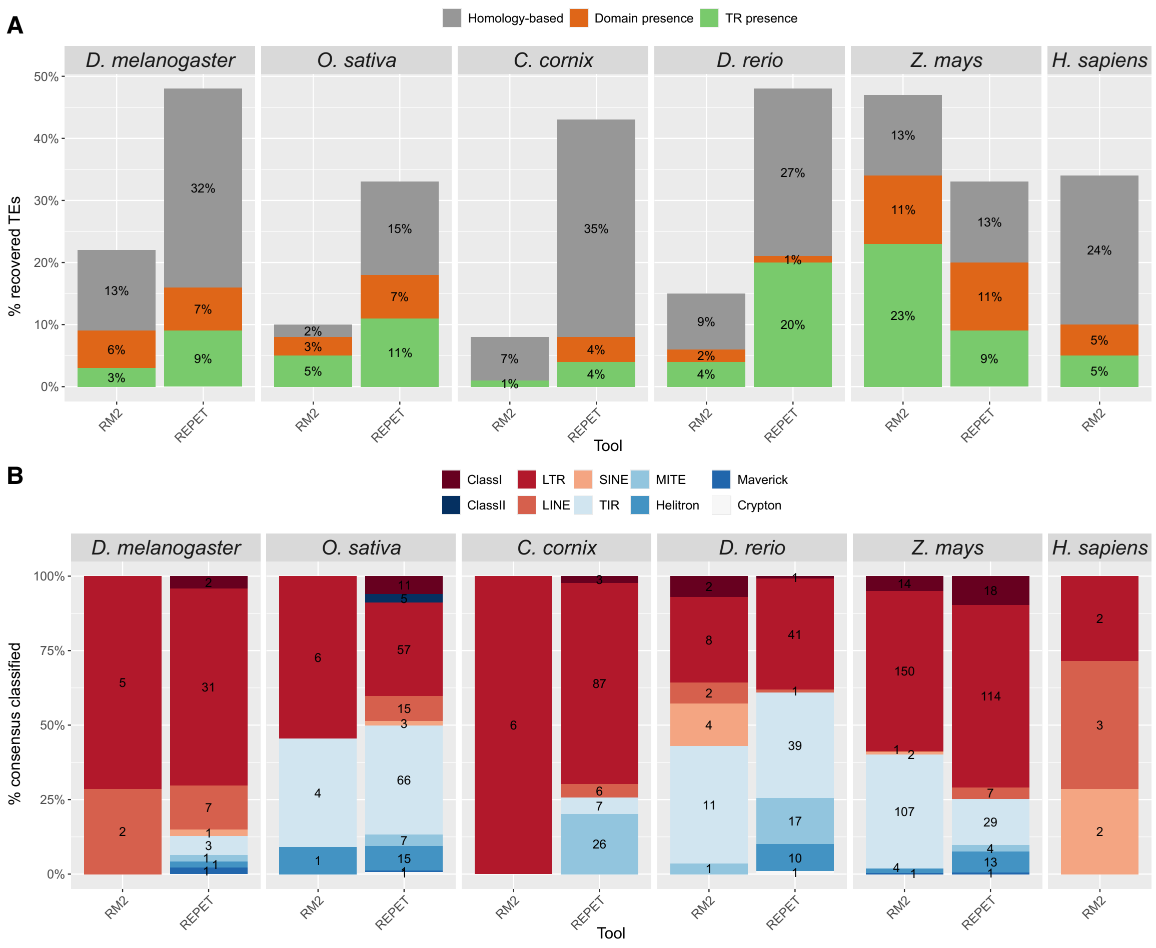

Figure 3. "Unclassified/unknown" consensus sequences in the raw libraries that are classified by MCHelper. (*A*) The proportion of TEs originally labeled as "unclassified/unknown," which were recovered by the three-step approach of MCHelper. (TR) terminal repeat. (*B*) Number of recovered TE sequences by TE order/class. Class I and Class II indicate all the elements that contain one or more retrotransposon-like domains in the first case or DNA-like domains in the latter case, but it was not possible to assign a deeper classification. EDTA is not included in this figure because the software does not report any "unknown/unclassified" elements.

Table S7B). When comparing MCHelper with EDTA and REPET, the number of short annotated copies was also reduced between 11.2% (in *C. cornix*) and 67.7% (in *D. melanogaster*) for EDTA outputs and between 2.4% (in *O. sativa*) and 20.9% (in *Z. mays*) in REPET outputs. The only exception found was *C. cornix* with the REPET + MCHelper, where the number of short copies increased by 14.5%.

Finally, MCHelper reduces the number of overlapping annotations, i.e., those annotated using more than one consensus sequence, again resulting in annotations that were more similar to the reference ones (Fig. 4C; Supplemental Table S7C). The MCHelper library led to a decrease of overlapping annotations of 29.4% (RM2), 44% (EDTA), and 19.2% (REPET) in *D. melanogaster*, of 7.8% (RM2), 24.4% (EDTA), and 17.6% (REPET) in *O. sativa*, of 47.6% (RM2) and 9.9% (EDTA) in *C. cornix*, of 11.7.% (RM2), 61.8% (EDTA), and 14.7% (REPET) in *D. rerio*, of 31% (RM2), 31.9% (EDTA) and 9.7% (REPET) in *Z. mays*, and of 25.4% (RM2) and 52.1% (EDTA) in *H. sapiens*. The only exception was found in *C. cornix* where MCHelper annotated 17.9% more overlapping copies compared with REPET. Note that the number of overlapping copies might be an overestimate as the raw libraries, and thus the MCHelper libraries, are only classified to the superfamily level. Thus copies that are detected as annotated with two consensus sequences might actually be annotated with redundant consensus from the same family or subfamily.

Annotations made by MCHelper libraries are longer and more contiguous than those of raw libraries

To assess the improvements in the genome-wide TE annotation based on the MCHelper libraries compared with the annotation based on the raw libraries, we used as a proxy the NTE50 and the LTE50 metrics, and the proportion of each TE order in the genome annotations. Briefly, the NTE50 is the number of the largest copies needed to annotate 50% of the mobilome. Thus, lower NTE50 values indicate that annotated copies are longer and thus that the annotation includes large unfragmented TE insertions (Jamilloux et al. 2017). The LTE50 is the length such that 50% of the mobilome is annotated by copies longer than that. So, higher LTE50 indicates better annotations (Jamilloux et al. 2017).

Our results showed that MCHelper libraries produced longer annotations than the raw library for the majority of tools and

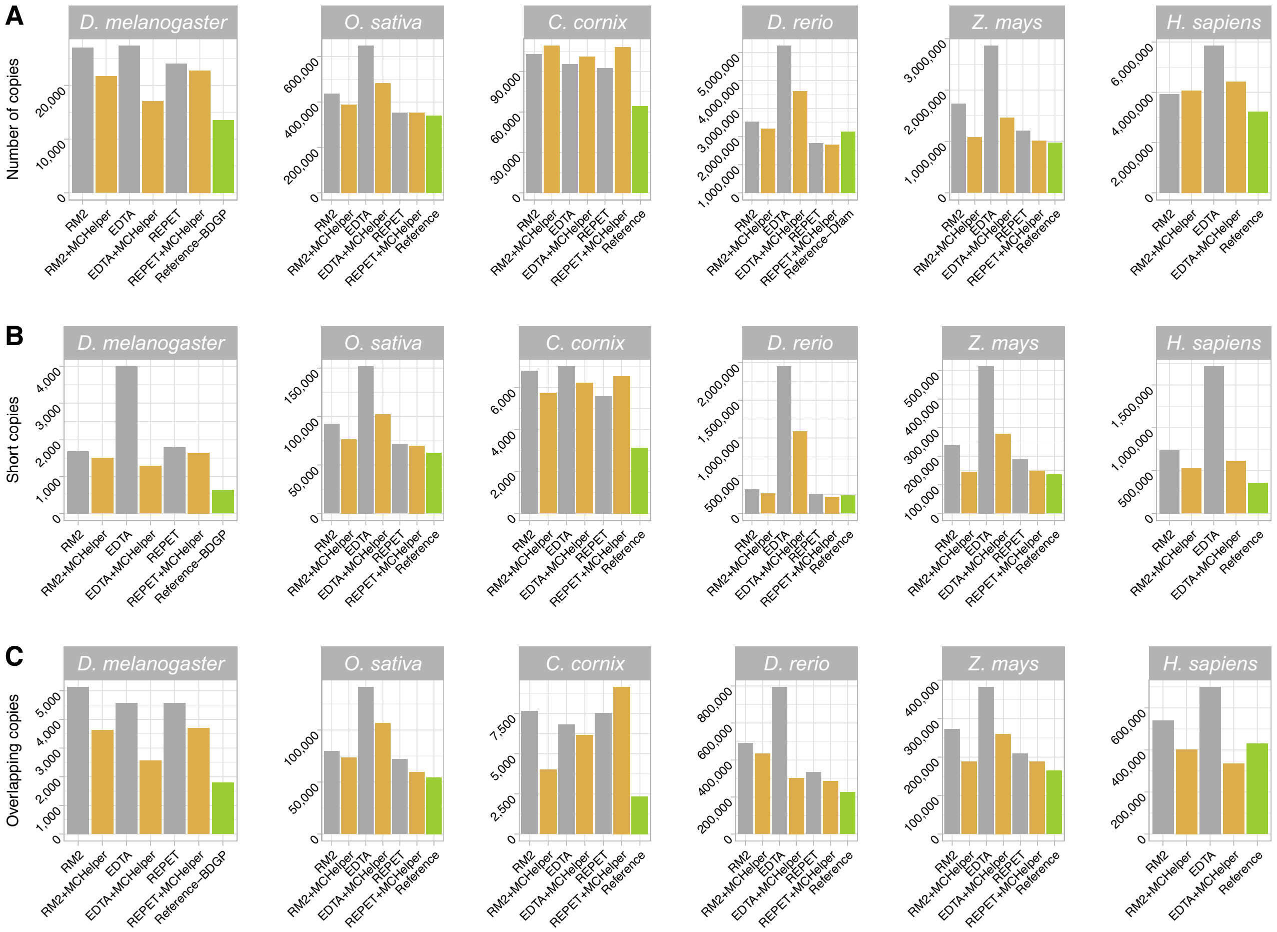

Figure 4. Genome-wide number and type of TE annotations for each of the three libraries analyzed in the six species. The BDGP, EDTA, MClibrary, Dfam-zebrafish, MTEC, and Dfam-Human libraries were used as reference for *D. melanogaster*, *O. sativa*, *C. cornix*, *D. rerio*, *Z. mays*, and *H. sapiens*, respectively. (*A*) Number of copies annotated by each library. (*B*) Number of copies categorized as short (<100 bp). (*C*) Number of copies categorized as overlapping (annotated by two or more consensus sequences).

genomes (Fig. 5). The NTE50 values were smaller when annotating the genome with the MCHelper library compared with the raw library with only three exceptions: EDTA NTE50 values were smaller in *D. melanogaster* and *C. cornix*, and REPET NTE50 values were smaller in *C. cornix* (Fig. 5; Table 1). Similarly, and as expected if the quality of MCHelper TE annotations is higher than the annotation performed with the raw library, the LTE50 values were higher when the genome was annotated with MCHelper libraries except for the annotation with EDTA libraries in three of the six species analyzed: *D. melanogaster*, *O. sativa*, and *C. cornix*.

Finally, we also examined the proportions of each TE order in the annotations produced by the raw, MCHelper, and reference libraries. EDTA raw libraries consistently annotated a higher proportion of TE orders compared both to MCHelper and raw libraries across genomes (Fig. 6; Supplemental Table S8). However, in *Z. mays* the number of nucleotides annotated as TEs was equal to the number of nucleotides in the assembly, suggesting that annotations with this library overestimates the number of TEs. For RM2 and REPET, the proportions of TE orders were higher in annotations performed with the MCHelper libraries compared with annotations performed with raw libraries except for REPET in *Z. mays* (Fig. 6; Supplemental Table S8). In *D. melanogaster*, the proportion of TE orders obtained with the MCHelper library was even higher than the one obtained with the reference-BDGP library and more similar to the MCTE library, suggesting that the MCHelper produces a high-quality annotation. In *O. sativa*, the TE order proportions were very similar when comparing the MCHelper and the raw RM2 library, while the proportions were smaller for the raw REPET library, consistent with the lower number of copies detected by REPET in this species (Fig. 4A) (39.2% REPET vs. 48.1% standard library). In *C. corvix*, the manually curated library annotated the lowest proportion of TEs, and only from two orders, LTR and LINE. This result suggests that although the *C. corvix* manually curated library is of high quality, it is probably incomplete. In *D. rerio*, we also found that the raw library from REPET annotated lower TE proportions compared with the MCHelper library and the reference library (34.7%, 45.7%, and 54.4%, respectively). In *Z. mays*, the MCHelper library processed from REPET annotated virtually the same proportion as the reference (85% REPET + MCHelper vs. 84.7% reference).

Overall, MCHelper libraries produce longer and less fragmented annotations than raw libraries for the majority of tested genomes and de novo tools. Importantly, and as expected, although the number of consensus sequences in the MCHelper libraries was smaller compared with the raw libraries (Fig. 2A), the proportions of TE orders annotated with the MCHelper

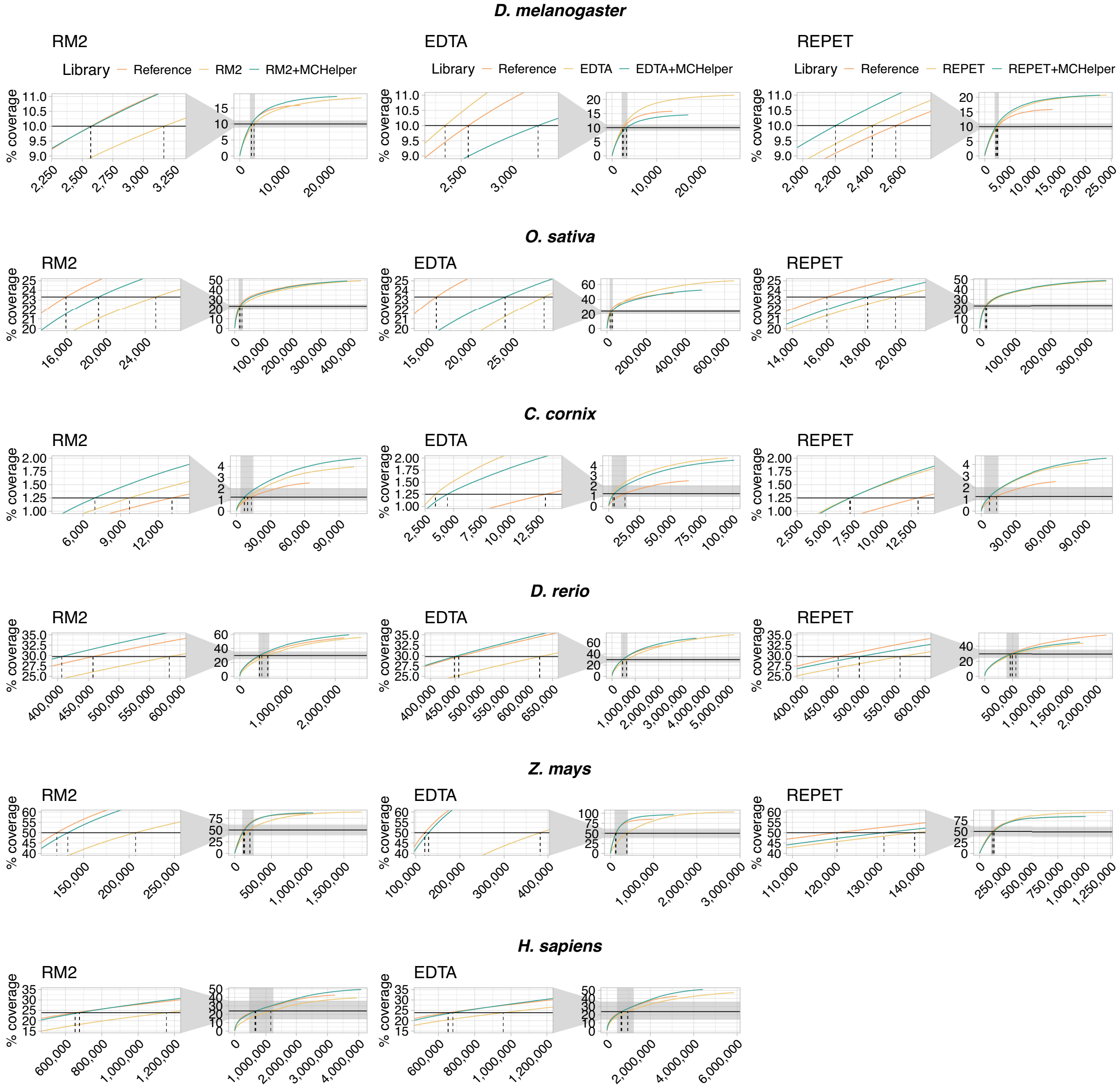

Figure 5. Cumulative length coverage plots for TE annotations in the six species analyzed. The black horizontal line represents half of the mobilome and it was used to calculate the LTE50 values (Table 1). The dashed vertical line indicates the NTE50 values. For each graph, a zoom capture is provided on the *left* to show when the cumulative length plot of each library crosses the black horizontal line (NTE50 values). Mobilome proportions are 20% in *D. melanogaster* (Mérel et al. 2020), 46.64% in *O. sativa* (Ou et al. 2019), 2.5% in *C. cornix*, 59.5% in *D. rerio* (Chang et al. 2022), and 47.68% in *H. sapiens* (Hoyt et al. 2022). In *Z. mays*, we used 50% of the full genome space to calculate the NTE50 and LTE50 as recommended by Jamilloux et al. (2017). NTE50 and LTE50 values are shown in Table 1.

libraries were closer to the ones produced with the reference libraries (Fig. 6).

MCHelper curates TE libraries in hours for small genomes and it is easy to install and run

Besides improving the completeness and accuracy of TE libraries, another main goal of MCHelper was to reduce the required time to curate TE sequences. To test the MCHelper runtimes, we executed MCHelper 10 times for the raw library of the four species with genomes <2 Gb produced by the three de novo tools used in this study, on a server with 48 CPUs. We measured the execution time for 10 distinct processes included in the MCHelper tool (see Methods). On average, MCHelper took between 1.10 (*D. melanogaster*—EDTA) and 41.95 (*D. rerio*—EDTA) hours (Table 2). In all cases, the extension step was the most time-consuming: up to 95.91% of the total MCHelper execution time for the *D. rerio* when using the RM2 library. For the two T2T genomes analyzed in this work, the running times of MCHelper ranged between 10,526 and 3342 h using 100 cores.

Table 1. NTE50 and LTE50 values of each library

	D. melanogaster		*O. sativa*		*C. cornix*		*D. rerio*		*Z. mays*		*H. sapiens*	
Library	NTE50	LTE50	NTE50	LTE50	NTE50	LTE50	NTE50	LTE50	NTE50	LTE50	NTE50	LTE50
RM2	3167	2188	25,075	1296	9699	769	582,490	378	207,065	2564	1,155,606	327
RM2 + MCHelper	2577	3007	19,235	1824	6953	1077	406,526	515	132,536	5046	676,997	464
EDTA	2342	3925	28,674	1584	3405	2164	623,608	340	382,400	1565	957,177	321
EDTA + MCHelper	3259	1762	24,031	1411	4453	1371	448,658	468	129,175	5328	680,093	464
REPET	2426	3514	19,675	1754	7108	979	559,160	337	138,840	4897	–	–
REPET + MCHelper	2201	4082	18,144	2032	7150	1051	492,326	364	131,590	5074	–	–
Reference	2570	3158	15,900	2279	13,088	569	457,644	439	120,514	5453	653,254	412

The NTE50 is calculated as the number of the largest copies needed to annotate 50% of the mobilome, while the LTE50 is defined as the length of the shortest TE annotation that annotates 50% of the mobilome (Jamilloux et al. 2017).

MCHelper is an easy tool to set up, as all the dependencies are available in a YML file that allows them all to be installed in an Anaconda environment using a single command. Also, MCHelper is easy to use because it can be run using a single Python command line with only three inputs: the raw library, the genome assembly, and the BUSCO gene set. As such, this tool can be easily integrated into any automatic pipeline to annotate TEs.

Discussion

We have shown that MCHelper increases TE libraries accuracy and completeness, producing longer and unfragmented TE genome annotations across eukaryotic species, by integrating and automating TE curation protocols (Fig. 1). Libraries generated by de novo TE identification tools (raw libraries) overestimate the number of TE copies as they contain TE consensus sequences that are redundant, fragmented, and false positive sequences (Rodriguez and Makalowski 2022; Storer et al. 2022). Based on several benchmarking metrics, we showed that MCHelper's automatically curated libraries for *D. melanogaster*, *O. sativa*, *C. cornix*, *D. rerio*, *Z. mays*, and *H. sapiens* are more accurate and complete compared with raw libraries, and thus more similar to high-quality manually curated (reference) libraries (Fig. 2). Moreover, the MCHelper tool incorporates a module specifically designed to classify TE sequences initially labeled as "unknown/unclassified" by de novo TE identification tools, which allows to recover classifications for up to 48% of these sequences (Fig. 3). This is highly relevant as the percentage of unclassified elements in raw libraries can be as high as 93% leading to very incomplete and inaccurate TE annotations (Petersen et al. 2019; Gilbert et al. 2021).

MCHelper can be executed in parallel in multiple CPUs taking advantage of currently available supercomputing resources. MCHelper allows to curate raw TE libraries in hours for small genomes, with the consensus extension module being the more time-consuming step (Table 2). Thus, the number of iterations chosen to run the consensus extension module will affect the total time needed to run MCHelper. Therefore, we recommend the users to test the appropriate value of this parameter to obtain the best possible results in the shortest possible runtime, especially when analyzing a large number of genomes.

Most previous TE curation automation efforts consisted of scripts dealing with individual limitations of the raw libraries that needed to be executed independently by the user for each one of the consensus sequences (Storer et al. 2021b; Goubert et al. 2022). Recently, a pipeline to de novo annotate TEs included multiple steps to deal with some specific limitations of raw libraries, in particular the presence of redundant and fragmented consensus (Baril et al. 2024). Machine learning approaches have also been implemented to automatically curate particular TE orders in plants (Orozco-Arias et al. 2021, 2022). Thus, previous attempts at automation were partial and still required substantial time investment by the researcher. On the other hand, MCHelper implements multiple manual curation protocols developed by TE experts that tackle the most common limitations of de novo TE identification tools across TE orders and species (Platt et al. 2016; Jamilloux et al. 2017; Storer et al. 2021b, 2022; Goubert et al. 2022).

Besides the fully automatic mode that allows the user to improve library accuracy and completeness in a time-effective manner, MCHelper also provides the user with the possibility of performing manual inspection of the generated libraries through a dedicated module (Fig. 1; Supplemental Fig. S1B). This module provides the user with information to detect and remove low-quality TE consensus sequences to further increase the overall quality of the TE libraries (Goubert et al. 2022; Tumescheit et al. 2022). We envision that the automatic mode of MCHelper will be highly useful for research projects that need to annotate TEs in a large number of genomes, while the semiautomatic or fully manual modes should be implemented to generate TE libraries for groups of species for which there are no reference libraries available yet.

MCHelper, in combination with one of the many de novo TE identification tools already available, should be instrumental in substantially increasing the number of genomes with accurate TE annotations, which is currently reduced to a very small fraction of the eukaryotic diversity available. Accurate TE annotations in the increasing number of eukaryotic genomes available are the first step toward understanding the dynamics of these important genome components but also the biology of genomes given their important roles in function, structure, and evolution.

Methods

MCHelper workflow

First, MCHelper conducts preprocessing analysis, ensuring file completeness and consistency with Wicker's classification nomenclature (Wicker et al. 2007). For the raw libraries of RM2 and

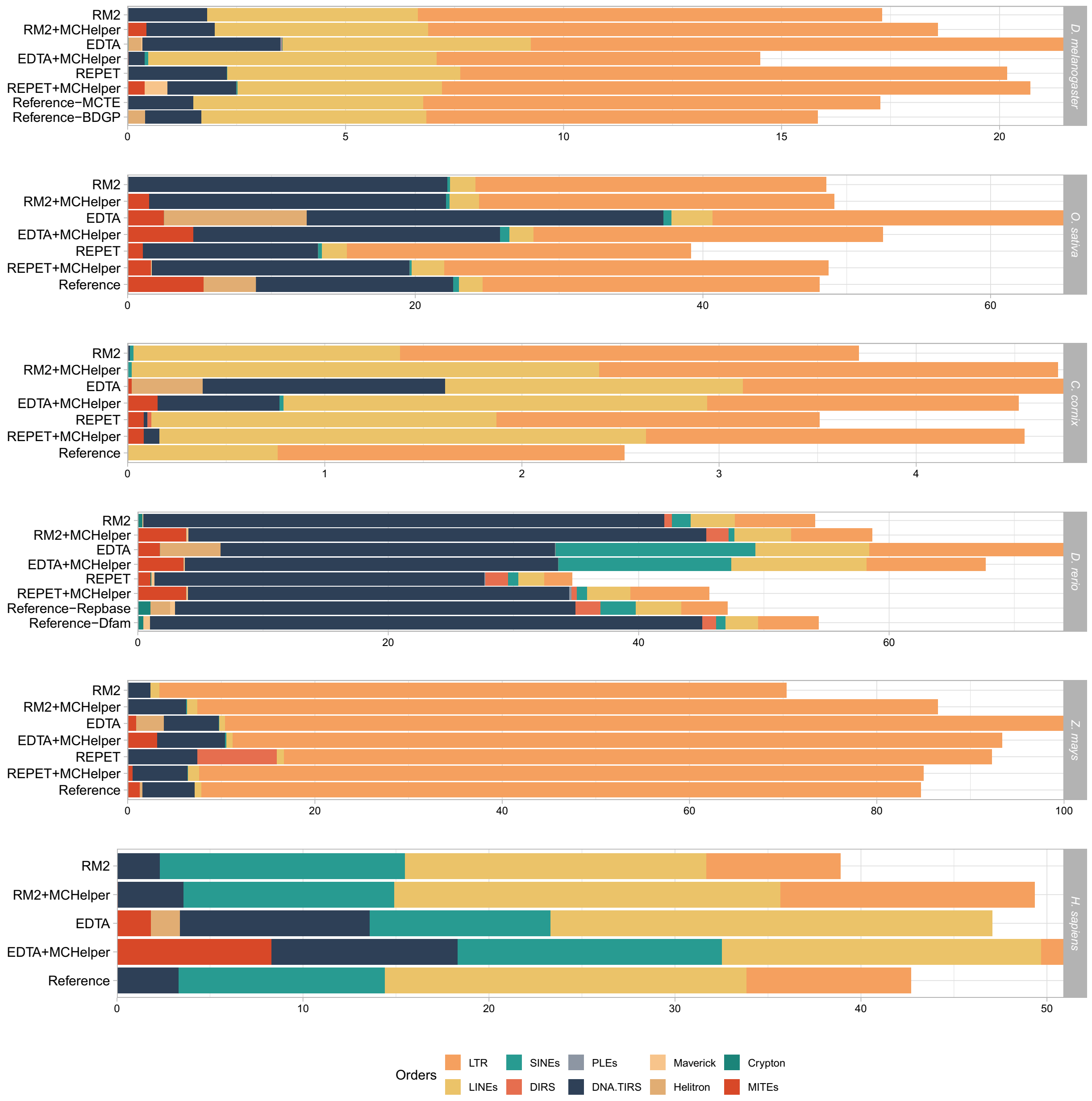

Figure 6. Genomic proportions of TE orders annotated by the raw, MCHelper, and reference libraries. Note that the proportion of each genome labeled as "unclassified/unknown" TEs is not shown to ease the comparison of TE order proportions across libraries. Annotation with the EDTA library clearly overestimates the amount of TEs in *Z. mays* which could be due to an incomplete defragmentation of TE copies.

Table 2. MCHelper average runtimes for the four species analyzed

	RM2		EDTA		REPET	
Species	Total runtime (h)	Extension step runtime (%)	Total runtime (h)	Extension step runtime (%)	Total runtime (h)	Extension step runtime (%)
D. melanogaster	1.42 ± 0.004	76.87	1.10 ± 0.025	62.02	3.46 ± 0.004	81.67
O. sativa	10.82 ± 0.050	83.64	10.71 ± 0.052	82.28	16.55 ± 0.084	79.76
C. cornix	5.98 ± 1.067	72.34	7,52 ± 0.033	63.48	12.33 ± 0.439	54.85
D. rerio	29.91 ± 0.094	95.91	41.95 ± 0.1056	72.92	27.92 ± 0.60	70.12

MCHelper was executed 10 times for each raw library (RM2 EDTA and REPET) and the average and standard deviation (SD) were calculated. Total runtimes are given in hours.

EDTA (or any other FASTA library), MCHelper replaces DNA order names with TIR, and "unknown" with "unclassified." Moreover, it excludes sequences with classifications not contained in the MCHelper nomenclature list (see Supplemental Methods), as well as satellites and RNAs, and creates a file with the list of excluded sequences and the reason of exclusion.

Redundancy reduction

MCHelper reduces redundancy by clustering all consensus sequences using either CD-HIT v.4.8.1 (Li and Godzik 2006) or MeShClust V3 (Girgis 2022), depending on the user decision, with an identity threshold of 95% and coverage of 98% as suggested by Flutre et al. (2011).

Consensus extension

Consensus sequences are processed using *n* iterations of BLAST, Extract, and Extend (BEE) rounds (16 by default) (Platt et al. 2016; Storer et al. 2021b; Baril et al. 2024). MCHelper uses NCBI-BLAST v.2.10.1 with parameter "-evalue 1×10^{-20}" based on Goubert et al. (2022). After each BEE round, consensus families are split into subfamilies, if needed, by applying a clustering approach based on the Kimura two-parameter distance on the MSA generated (see Supplemental Methods). All consensus sequences are extended until the sequence reaches both TE ends or until the end of the iterations (see Supplemental Methods).

False positive filtering

Consensus sequences with (1) very few full-length fragments (FLFs) in the genome (by default one FLF but this parameter can be defined by the user) (Jamilloux et al. 2017); (2) having homology with multicopy genes or rRNAs; and (3) containing simple sequence repeats (SSRs) (by default 60% also adjustable by the user) are considered false positive consensus sequences and filtered out (see Supplemental Methods; Storer et al. 2021b, 2022). For genes and rRNAs searching, MCHelper uses hmmscan from HMMER (Eddy 2011) v.3.3.2 with the following parameters "-E 10 --noali" and the rest by default.

Homology-based TE identification

Once false positives are filtered out, the remaining sequences are searched against a curated TE database, and those with an 80-80-80 hit (at least 80 bp long with 80% sequence identity over 80% of their length) (Wicker et al. 2007) are saved in the final curated library (see Supplemental Methods). Those that do not fulfill the 80-80-80 rule are divided into classified and unclassified sequences and follow two distinct paths (Fig. 1). MCHelper uses NCBI-BLAST v.2.10.1 to perform this search.

Structural checking

For classified sequences, MCHelper looks for structural information: the presence of protein domains, TRs, and poly(A) tails, to check if the original classification of the consensus sequence matches the expected TE features (see Supplemental Methods; Goubert et al. 2022). In the MCHelper semiautomatic mode, those sequences that present the expected features based on their classification are kept, and the others are manually inspected in the Manual Inspection Module (see Supplemental Methods).

TE classification

For "unclassified" elements, MCHelper tries to infer the correct classification using three steps: homology (using 70-70-70 hits), coding domain presence (e.g., transposase and reverse transcriptase), and terminal repeats (LTRs and TIRs) (see Supplemental Methods).

Manual inspection

MCHelper allows the user to manually visualize useful information to decide if a consensus sequence should be kept, removed, or reclassified. All the graphical information is generated by TE+Aid v.1.0 (Goubert et al. 2022) except the MSA plot that is generated by CIAlign v.1.0.18 (Tumescheit et al. 2022). Finally, structural information needed for manual inspection is provided through the terminal (see Supplemental Methods).

Genome assemblies used

We evaluated the performance of the MCHelper tool on six different species with high-quality manually curated TE libraries: *D. melanogaster*, *O. sativa*, *C. cornix*, *D. rerio*, *Z. mays*, and *H. sapiens*. The assemblies used were GenBank assemblies GCA_002050065.1 for *D. melanogaster*, GCF_000738735.6 for *C. cornix*, and GCA_000002035.4 for *D. rerio*, the RGAP assembly version 7 for *O. sativa* (available at https://phytozome-next.jgi.doe.gov/info/Osativa_v7_0) (Ouyang et al. 2007), and the T2T assemblies GCA_022117705.1 for *Z. mays*, and GCF_009914755 for *H. sapiens*. These species were selected based on their varying genome sizes: *D. melanogaster*: 180 Mb (Ashburner et al. 2005; Berlin et al. 2015); *O. sativa*: 466 Mb (Yu et al. 2002; Kawahara et al. 2013); *C. corvix*: 1.1 Gb (Weissensteiner et al. 2020); *D. rerio*: 1.7 Gb (Howe et al. 2013); *Z. mays*: 2.3 Gb (Schnable et al. 2009); and *H. sapiens*: 3.3 Gb (Nurk et al. 2022), respectively, and differences in TE content: ∼20% in *D. melanogaster* (Mérel et al. 2020; Sicat et al. 2022), 46.64% in *O. sativa* (Ou et al. 2019), 59.5% in *D. rerio* (Chang et al. 2022), 85% in *Z. mays* (Schnable et al. 2009), and 47.68% in *H. sapiens* (Hoyt et al. 2022).

Source of the raw libraries

To obtain the raw TE libraries, we utilized three de novo tools: RM2 (Flynn et al. 2020), EDTA (Ou et al. 2019), and TEdenovo from REPET (Flutre et al. 2011). For the RM2 raw libraries, we employed those available in its repository (https://github.com/jmf422/TE_annotation/tree/master/benchmark_libraries/RM2) for *D. melanogaster*, *O. sativa*, and *D. rerio*. For *C. cornix*, *Z. mays*, and *H. sapiens* the raw libraries were generated using the TEtools Docker image available at https://github.com/Dfam-consortium/TETools, with the assemblies mentioned in the previous section. With respect to the EDTA libraries, we generated them for the six species using the EDTA pipeline v.2.0 with --anno 0 and the other parameters by default and the same assemblies that were used for the RM2 library generation. Regarding the REPET libraries, we used the same assemblies that were used for the two previous tools. For the *D. melanogaster* genome, we ran the TEdenovo pipeline using default parameters and followed the recommended steps in the user's guideline (https://urgi.versailles.inra.fr/Tools/REPET/TEdenovo-tuto). For the *D. rerio* genome, we created a subset of 300 Mb using the PreProcess.py script available with the REPET package and then we ran TEdenovo pipeline using default parameters as for the *D. melanogaster* genome (Jamilloux et al. 2017). The REPET group kindly provided us with the output generated with the *O. sativa* genome available in the RepetDB database (Amselem et al. 2019) and with the libraries for *C. cornix* and *Z. mays*.

Source of the reference libraries

As the reference libraries, we used the BDGP data set (Kaminker et al. 2002) and the MCTE library (Rech et al. 2022) for *D. melanogaster*, the library published by Ou et al. (2019) for *O. sativa* (referred to as "standard library" by the authors), the MClibrary available in Weissensteiner et al. (2020) for *C. cornix*, the manual curated TE models available in Dfam release 3.7 (Storer et al. 2021a) and in Repbase version 20181026 for *D. rerio*, MTEC curated by Ou et al. (https://github.com/oushujun/MTEC) for *Z. mays*, and curated TE models in Dfam 3.7 for *H. sapiens*. For the rice "standard library," Dfam and Repbase libraries, we unified the LTR sequences with their corresponding internal part before performing further analysis, using in-house scripts.

Generation of the MCHelper libraries

MCHelper v.1.7.0 was executed with default parameters (customizable parameters can be found at GitHub: https://github.com/GonzalezLab/MCHelper). In the false positive filtering step, MCHelper requires a gene set to detect homology between the consensus sequences and multicopy gene families. We used the following BUSCO gene sets: for *D. melanogaster*, we utilized the Diptera set (diptera_odb10), for *O. sativa* and *Z. mays*, the Viriplantae set (viriplantae_odb10), for *C. cornix*, the aves set (aves_odb10), for *D. rerio*, the Actinopterygii set (actinopterygii_odb10), and for *H. sapiens*, the mammalian set (mammalia_odb10). To accommodate MCHelper's expectation of a single file with all the hidden Markov models (HMMs), we concatenated all the available HMM files into a single file for each species. Note that all the libraries generated in this work are available at our institutional repository (https://digital.csic.es/handle/10261/362092) and as Supplemental Data files S1–S6. The genome assemblies and the genome annotations are also available at our institutional repository (https://digital.csic.es/handle/10261/362092).

Benchmarking metrics used

At the library level, we used different metrics to benchmark the quality of the consensus models in the raw, MCHelper, and reference libraries used. We divided those metrics into two distinct categories: completeness and accuracy, based on the criteria of a "high-quality library" (Jamilloux et al. 2017; Storer et al. 2021b, 2022; Goubert et al. 2022). Metrics used to measure completeness were: the total number of consensus sequences, family quality evaluation, and TE order length distribution. The family quality evaluation consists of defining perfect, good, present, and missing families, as follows: Perfect families are those that have a match with one sequence in the analyzed library with >95% nucleotide similarity and >95% length coverage. Good families are those that have multiple overlapping matches in the analyzed library with >95% nucleotide similarity and >95% coverage. Present families are the same as good families, but with >80% similarity and >80% coverage. Below those thresholds, the families are considered missing (Flynn et al. 2020; Rodriguez and Makalowski 2022). We used the script developed by Flynn et al. (2020) and available at https://github.com/jmf422/TE_annotation/ to classify the consensus sequences.

To measure accuracy, we used the number of false positive sequences as calculated in the false positive filtering step of MCHelper (see Supplemental Methods). Briefly, we calculated the number of consensus sequences with >60% of their sequence length corresponding to single sequence repeats (SSRs), and the number of consensus sequences with hits with BUSCO genes and rRNAs.

To benchmark the quality of the annotations, we ran RepeatMasker v.4.1.2-p1 (Smit et al. 2015) on each of the six genomes with the raw, reference and MCHelper libraries using the -lib parameter, and the following additional parameters: -gff -nolow -no_is -norna. We then used the OneCodeToFindThemAll script (Bailly-Bechet et al. 2014) in each genome annotation to defragment the copies, and we used the "--strict" parameter for *D. melanogaster*, *D. rerio*, and *C. cornix*. The output of this script was used to estimate the benchmarking metrics. We used the following groups of metrics: quantity of copies, lengths of copies, and genomic proportions occupied by the annotated copies. The metrics used for the quantity of copies were: the total number of annotations, the number of short annotations (<100 pb), and the number of overlapping annotations (i.e., copies annotated by two or more TE consensus sequences). To define overlapping copies, we considered those that contained any bases annotated with more than one consensus sequence. We used BEDTools merge from the BEDTools suite version 2.30 (https://bedtools.readthedocs.io/en/latest/) (Quinlan and Hall 2010) with parameters "-d -1 -c 9,9 -o count,collapse" to merge the overlapping copies, and we counted the number of merged copies that came from more than one consensus sequence. To benchmark the length of copies, we used the NTE50 and the LTE50. To calculate these two metrics, we first ordered the copy lengths from longest to shortest and then calculated a cumulative sum of the copy lengths until we reached 50% of the total TE content. The NTE50 is analogous to the N50 metric, used in assembly quality checks, and corresponds to the number of the largest copies needed to annotate 50% of the mobilome. Lower NTE50 values are indicative of longer annotations suggesting that the annotation has captured large unfragmented TE insertions. LTE50 corresponds to the length of the TE annotation at the point in the ordering that 50% of the mobilome is annotated. Higher LTE50 indicates better annotations (Jamilloux et al. 2017). Note that for *Z. mays*, NT50 and LTE50 were calculated for 50% of the full genome space instead of for 50% of the mobilome. Finally, we also calculated the TE order genomic proportions as the percentages of nucleotides that were annotated with TE consensus sequences for each one of the orders analyzed. We obtained the genomic proportions directly from the outputs of the *OneCodeToFindThemAll* script, which sums up all the bases corresponding to each TE order divided by the total assembly size. All scripts used to calculate the metrics described above are available in Supplemental Code File and at https://github.com/GonzalezLab/MCHelper/tree/main/benchmarking_scripts.

Comparisons with classification tools

To compare the MCHelper efficiency when classifying unknown TEs, we processed the "unclassified/unknown" TEs from the RM2 libraries from four of the species analyzed in this study (*D. melanogaster*, *O. sativa*, *C. cornix*, and *D. rerio*), and then we followed two different approaches to try to assign a classification. First, we manually inspected all the consensus sequences to infer the classification based on the observed structure. Second, we used cross_match with a substitution matrix specially trained for noncoding sequences available in RepeatMasker (named 25p41g.matrix) and against the Repbase library version 20181026 to infer the classification based on homology. Then, we created a subset of the sequences which we inferred the classification based on either structure or homology and we took as ground true the classification inferred. Using this subset as input, we executed TEClass2 (Bickmann et al. 2023), TERL (da Cruz et al. 2021), and DeepTE (Yan et al. 2020) to predict the classification of the sequences. TEClass2 was run using the web GUI (https://www.bioinformatics.uni-muenster.de/tools/teclass2/index.pl?), TERL was executed with the parameter "-m Models/DS3/" to indicate the model to be used and the rest of parameters by default. Finally, DeepTE was run using the Metazoans_model for *D. melanogaster*, Plans_model for *O. sativa*, and Others_model for *C. cornix* and *D. rerio*. Finally, we calculated three metrics to evaluate the performance of the three tools together with MCHelper: Recovering proportion, overlapping proportion, and precision. We defined the recovering proportion as the proportion of the sequences in the subset recovered by each tool, without taking into account the classification predicted. The overlapping proportion was defined as the proportion of the sequences in the subset that was recovered by each tool and for which the classification was correctly predicted at the order level. And the precision was defined as the proportion of the recovered sequences by each tool that was correctly classified.

Comparison with SENMAP

To compare the performance of MCHelper with another curation algorithm, we ran the SENMAP neural network through the script available at: https://github.com/simonorozcoarias/SENMAP. SENMAP was trained only with plant LTRs and its objective is to detect whether an element is complete or not. For this purpose, we used an animal (*D. melanogaster*) and a plant (*O. sativa*) and calculated the number of elements in the final library of both SENMAP (filtering out those that the network detects as incomplete) and MCHelper. In addition, we calculated how many elements were classified as perfect, good, present, and missing families.

Runtime tests

To test the time required to curate TE libraries, we executed MCHelper in the fully automated mode, using 16 extension iterations extending 500 bases at each iteration, 48 CPUs, and the default values of other MCHelper parameters. We ran the software 10 times to see the variability of the execution times and we calculated the average and the standard deviation of the 10 executions. Besides measuring total time, we calculated the execution time of 10 subroutines: preprocessing, extension (all the iterations), FLF filtering, false positive filtering, TE feature calculation, classification assignment for the classified consensus sequences, FLF filtering, TE feature calculation, classification assignment by homology for the unclassified consensus sequences, and order inferred from structural features also for the unclassified consensus sequences. All the executions were performed in a dedicated computing node with 64 CPUs and 240 GB in RAM.

Software availability

All the in-house scripts used in this paper as well as the MCHelper source code is available in Supplemental Code File and at https://github.com/GonzalezLab/MCHelper.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank the REPET team for providing the *O. sativa*, *C. cornix*, and *Z. mays* REPET libraries. This work was supported by grants PID2020-115874GB-I00 funded by MICIU/AEI/10.13039/501100011033 and PID2023-148838NB-I00 funded by MICIU/AEI/10.13039/501100011033/FEDER, EU awarded to J.G. and by grant 2021 SGR 00417 funded by Departament de Recerca i Universitats, Generalitat de Catalunya awarded to J.G. P.S. is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 956229. R.D. acknowledges funding from Wellcome grant 207492.

Author contributions: S.O.-A.: designed research, performed research, analyzed data, contributed analytical tools, and wrote the paper; P.S.: contributed analytical tools; R.D.: designed research; J.G.: designed research, analyzed data, and wrote the paper. All authors reviewed and commented on the submitted manuscript.

References

Amselem J, Cornut G, Choisne N, Alaux M, Alfama-Depauw F, Jamilloux V, Maumus F, Letellier T, Luyten I, Pommier C, et al. 2019. RepetDB: a unified resource for transposable element references. *Mob DNA* **10:** 6. doi:10.1186/s13100-019-0150-y

Ashburner M, Hawley R, Golic K. 2005. *Drosophila: a laboratory handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mob DNA* **5:** 13. doi:10.1186/1759-8753-5-13

Baril T, Galbraith J, Hayward A. 2024. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. *Mol Biol Evol* **41:** msae068. doi:10.1093/molbev/msae068

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33:** 623–630. doi:10.1038/nbt.3238

Bickmann L, Rodriguez M, Jiang X, Makalowski W. 2023. TEclass2: classification of transposable elements using Transformers. bioRxiv doi:10.1101/2023.10.13.562246

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. *Mol Ecol* **22:** 1503–1517. doi:10.1111/mec.12170

Chang N-C, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. 2022. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res* **32:** 1408–1423. doi:10.1101/gr.275655.121

da Cruz MHP, Domingues DS, Saito PTM, Paschoal AR, Bugatti PH. 2021. TERL: classification of transposable elements by convolutional neural networks. *Brief Bioinform* **22:** bbaa185. doi:10.1093/bib/bbaa185

Darwin Tree of Life Project Consortium. 2022. Sequence locally, think globally: the Darwin Tree of Life Project. *Proc Natl Acad Sci* **119:** e2115642118. doi:10.1073/pnas.2115642118

De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, Caligiana A, Brocculi G, Adney EM, Boeke JD, et al. 2019. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **566:** 73–78. doi:10.1038/s41586-018-0784-9

Dunn-Fletcher CE, Muglia LM, Pavlicev M, Wolf G, Sun M-A, Hu Y-C, Huffman E, Tumukuntala S, Thiele K, Mukherjee A, et al. 2018. Anthropoid primate-specific retroviral element THE1B controls expression of *CRH* in placenta and alters gestation length. *PLoS Biol* **16:** e2006337. doi:10.1371/journal.pbio.2006337

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* **7:** e1002195. doi:10.1371/journal.pcbi.1002195

Elliott TA, Heitkam T, Hubley R, Quesneville H, Suh A, Wheeler TJ. 2021. TE hub: a community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mob DNA* **12:** 16. doi:10.1186/s13100-021-00244-0

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6:** e16526. doi:10.1371/journal.pone.0016526

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci* **117:** 9451–9457. doi:10.1073/pnas.1921046117

Gilbert C, Peccoud J, Cordaux R. 2021. Transposable elements and the evolution of insects. *Annu Rev Entomol* **66:** 355–372. doi:10.1146/annurev-ento-070720-074650

Girgis HZ. 2022. MeShClust v3.0: high-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. *BMC Genomics* **23:** 423. doi:10.1186/s12864-022-08619-0

Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. 2022. A beginner's guide to manual curation of transposable elements. *Mob DNA* **13:** 7. doi:10.1186/s13100-021-00259-7

Hayward A, Gilbert C. 2022. Transposable elements. *Curr Biol* **32:** R904–R909. doi:10.1016/j.cub.2022.07.044

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496:** 498–503. doi:10.1038/nature12111

Hoyt SJ, Storer JM, Hartley GA, Grady PG, Gershman A, de Lima LG, Limouse C, Halabian R, Wojenski L, Rodriguez M, et al. 2022. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* **376:** eabk3112. doi:10.1126/science.abk3112

Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H, et al. 2016. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell* **166:** 102–114. doi:10.1016/j.cell.2016.05.032

Jamilloux V, Daron J, Choulet F, Quesneville H. 2017. De novo annotation of transposable elements: tackling the fat genome issue. *Proc IEEE* **105:** 474–481. doi:10.1109/JPROC.2016.2590833

Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler S, Jermiin LS, Skirmuntt EC, Katzourakis A, et al. 2020. Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583:** 578–584. doi:10.1038/s41586-020-2486-3

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462–467. doi:10.1159/000084979

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol* **3:** research0084.1. doi:10.1186/gb-2002-3-12-research0084

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6:** 4. doi:10.1186/1939-8433-6-4

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. 2022. The Earth BioGenome Project 2020: starting the clock. *Proc Natl Acad Sci* **119:** e2115635118. doi:10.1073/pnas.2115635118

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658–1659. doi:10.1093/bioinformatics/btl158

Mérel V, Boulesteix M, Fablet M, Vieira C. 2020. Transposable elements in *Drosophila*. *Mob DNA* **11:** 23. doi:10.1186/s13100-020-00213-z

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Orozco-Arias S, Candamil-Cortés MS, Valencia-Castrillón E, Jaimes PA, Orozco NT, Arias-Mendoza M, Tabares-Soto R, Guyot R, Isaza G. 2021. SENMAP: a convolutional neural network architecture for curation of LTR-RT libraries from plant genomes. In *2021 IEEE 2nd International Congress of Biomedical Engineering and Bioengineering (CI-IB&BI)*, pp. 1–4. IEEE, Bogotá, Colombia.

Orozco-Arias S, Candamil-Cortes MS, Jaimes PA, Valencia-Castrillon E, Tabares-Soto R, Guyot R, Isaza G. 2022. Deep neural network to curate LTR retrotransposon libraries from plant genomes. In *Practical Applications of Computational Biology & Bioinformatics, 15th International Conference (PACBB 2021)*, pp. 85–94. Springer International Publishing, Salamanca, Spain.

Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KA, Moreno-Santillán DD, Crookshanks C, Roberts J, Garcia C, et al. 2023. Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science* **380:** eabn1430. doi:10.1126/science.abn1430

Ou S, Jiang N. 2018. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176:** 1410–1422. doi:10.1104/pp.17.01310

Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20:** 275. doi:10.1186/s13059-019-1905-y

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al. 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* **35(**Database issue**):** D883–D887. doi:10.1093/nar/gkl976

Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, et al. 2018. The neuronal gene arc encodes a repurposed retrotransposon gag protein that mediates intercellular RNA transfer. *Cell* **172:** 275–288.e18. doi:10.1016/j.cell.2017.12.024

Payer LM, Burns KH. 2019. Transposable elements in human genetic disease. *Nat Rev Genet* **20:** 760–772. doi:10.1038/s41576-019-0165-8

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecol Evol* **19:** 11. doi:10.1186/s12862-018-1324-9

Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biol Evol* **8:** 403–410. doi:10.1093/gbe/evw009

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V, Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. *Nat Commun* **13:** 1948. doi:10.1038/s41467-022-29518-8

Rodriguez M, Makalowski W. 2022. Software evaluation for de novo detection of transposons. *Mob DNA* **13:** 14. doi:10.1186/s13100-022-00266-2

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326:** 1112–1115. doi:10.1126/science.1178534

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* **19:** 329–346. doi:10.1038/s41576-018-0003-4

Sicat JPA, Visendi P, Sewe SO, Bouvaine S, Seal SE. 2022. Characterization of transposable elements within the *Bemisia tabaci* species complex. *Mob DNA* **13:** 12. doi:10.1186/s13100-022-00270-6

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. http://repeatmasker.org.

Sotero-Caio CG, Platt RN, Suh A, Ray DA. 2017. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol Evol* **9:** 161–177. doi:10.1093/gbe/evw264

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021a. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12:** 2. doi:10.1186/s13100-020-00230-y

Storer JM, Hubley R, Rosen J, Smit AF. 2021b. Curation guidelines for de novo generated transposable element families. *Curr Protoc* **1:** e154. doi:10.1002/cpz1.154

Storer JM, Hubley R, Rosen J, Smit AF. 2022. Methodologies for the de novo discovery of transposable element families. *Genes (Basel)* **13:** 709. doi:10.3390/genes13040709

Tumescheit C, Firth AE, Brown K. 2022. CIAlign: a highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ* **10:** e12983. doi:10.7717/peerj.12983

Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al. 2020. Discovery and population genomics of structural variation in a songbird genus. *Nat Commun* **11:** 3403. doi:10.1038/s41467-020-17195-4

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973–982. doi:10.1038/nrg2165

Yan H, Bombarely A, Li S. 2020. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* **36:** 4269–4275. doi:10.1093/bioinformatics/btaa519

Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Indica*). *Science* **296:** 79–92. doi:10.1126/science.1068037

Received December 5, 2023; accepted in revised form September 18, 2024.