MCHelper automatically curates transposable element libraries across species

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

[1] Institute of Evolutionary Biology, CSIC, UPF, Barcelona, Spain.

[2] Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

*Corresponding author: josefa.gonzalez@csic.es

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

Competing Interest Statement: The authors declare no competing interest.

Keywords: genome annotation, consensus sequence extension, false positive filtering, structural checking, transposable element classification

Abstract

The number of species with high quality genome sequences continues to increase, in part due to 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 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 transposable element sequences is still technically challenging. Several de novo transposable element 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 two de novo transposable element identification tools, RepeatModeler2 and REPET, in fruit fly, rice, and zebrafish, we show a substantial improvement in the quality of the transposable element libraries and genome annotations. MCHelper libraries are less redundant, with up to 54% reduction in the number of consensus sequences, have up to 11.4% fewer false positive sequences, and also have up to ~45% fewer “unclassified/unknown” transposable element consensus sequences. Genome-wide transposable element annotations were also improved, including larger unfragmented insertions. MCHelper is a fast, easy to install, and easy to use tool and is available at https://github.com/GonzalezLab/MCHelper.

Significant statement

The number and quality of genome sequences available is rapidly increasing across the tree of life. Transposable elements, DNA fragments with the ability to move within the genome often generating copies of themselves in the process, constitute significant proportions of the total genome content across species. Although transposable elements are a major source of genome variation, they are often ignored as it is still challenging to accurately annotate them. We have developed MCHelper, an automatic tool to curate transposable element libraries that allows for high quality, fast, and replicable transposable elements annotations.

Introduction

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 (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 (Payer and Burns 2019, De Cecco et al 2019). Thus, 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 (Flynn et al., 2020), REPET (Flutre et al., 2011), and EDTA (Ou et al., 2019) (for a more comprehensive list visit: https://tehub.org/; Elliot et al 2021). However, libraries produced by these tools lead to low quality and incomplete TE annotations (Goubert et al., 2022; Rodriguez & Makałowski, 2022). This is mainly due to the high diversity of TE sequences present in genomes that precludes 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 friendly-user manuals have been recently published (Baril et al., 2022; Goubert et al., 2022; Jamilloux et al., 2016; Platt et al., 2016; J. M. Storer et al., 2021, 2022). Additionally, some specific tools to aid in the manual curation have also been produced (Baril et al., 2022; Goubert et al., 2022; Orozco-Arias, et al., 2021; 2022). Still, manual curation of TE libraries is time-consuming 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., 2022). 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 higher quality 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 two *de novo* TE tools, RepeatModeler2 and REPET, which differ in their output formats, and in three species, which differ in TE content and genome size: the fruit fly *D. melanogaster,* rice (*O. sativa*) and zebrafish (*D. rerio*).

RESULTS

MCHelper integrates protocols to automatically curate transposable element libraries

MCHelper is a computational tool that integrates transposable element (TE) manual curation protocols developed by TE experts to improve the completeness and accuracy of TE libraries generated by automatic *de novo* tools (Baril et al., 2022; Goubert et al., 2022; Jamilloux et al., 2016; Ou & Jiang, 2018; Platt et al., 2016; J. M. Storer et al., 2021, 2022). MCHelper is a flexible tool designed to accept libraries generated by *de novo* TE identification tools that produce a fasta file output, such as RepeatModeler2 (RM2; Flynn et al., 2020) or EDTA (Ou et al., 2019), as well as the output files produced by the *TEdenovo* pipeline from REPET (Flutre et al., 2011; Figure 1). 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: (i) reduces consensus redundancy, once at the beginning and once at the end of the pipeline; (ii) extends the consensus sequences; (iii) identifies and filters false positive consensus; (iv) identifies previously known TEs based on homology; (v) performs structural checks on consensus sequences initially labeled by *de novo* TE identification tools as “classified” TEs; and (vi) assigns a classification to consensus sequences initially labeled as “unclassified/unknown” TEs (Figure 1 and Figure S1).

MCHelper offers three distinct levels of automation: fully-automatic, semi-automatic, and fully-manual, thus enabling users to tailor the software's functionality to their curation requirements and time availability (Figure 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 Material and 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. MCHelper adds the suffix "_inc" to “incomplete” sequences for easy identification by the user. In the semi-automatic 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 from three species

To test the MCHelper tool, we ran the fully-automatic mode for three species that differ in genome size and TE content and that have available reference TE libraries: the fruit fly (*Drosophila melanogaster*), rice (*Oryza sativa*), and zebrafish (*Danio rerio*). For each species we ran MCHelper with the raw TE libraries obtained from two different *de novo* tools, RepeatModeler2 (RM2; Flynn et al., 2020) and REPET (Flutre et al., 2011), and then compared the MCHelper libraries with the available reference libraries: BDGP (Kaminker et al., 2002) and MCTE libraries (Rech et al., 2022) for *D. melanogaster*, the manual curated library provided by Ou et al. (2019), named hereafter as "EDTA library" for *O. sativa*, and the Dfam (J. Storer et al., 2021) and Repbase (Jurka et al., 2005) libraries for *D. rerio*.

MCHelper reduced the total number of families (consensus sequences) in the raw libraries produced by RM2 and REPET in the three species analyzed. Using the RM2 output, MCHelper reduced the number of families by 54.2%, 32.2%, and 34.7% in *D. melanogaster*, *O. sativa*, and *D. rerio*, respectively (Figure 2A and Table S1A). Additionally, MCHelper was able to remove 4%, 11.4%, and 5.7% of false positives in these three species (Figure S2 and Table S1D). Using REPET's output, MCHelper reduced the number of families by 57.7% in *D. melanogaster*, 51.4% in *O. sativa* and 48.7% in *D. rerio*, and removed 3.7%, 6.7%, and 5% of false positives, respectively. 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 six times more perfect families than the raw libraries (Figure 2B and Table S1B). 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 Materials and Methods). We also found that the MCHelper automatically curated libraries contained fewer missing families for all genomes and *de novo* tools except for the REPET libraries of *D. melanogaster* and *O.sativa*, for which the number of missing families was very similar (*D. melanogaster*: 32 REPET vs 31 MCHelper; *O. sativa*: 1,484 REPET vs 1,483 MCHelper; Figure 2B).

Finally, MCHelper consistently generated longer consensus sequences than the raw libraries, thus exhibiting lengths more similar to those reported in the reference libraries (Figure 1C and

Table S1C) across all major orders of TEs (Helitron, LINE, LTR, and TIR). 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 (Figure 1) (Baril et al., 2022; Goubert et al., 2022; J. M. Storer et al., 2021). However, as noted by Baril et al. (2022), longer consensus sequences do not always equate to a better representation of the TE family. To avoid over-extensions and fuzzy terminations, MCHelper incorporates an approach that detects when the consensus sequence reaches either end (Figure S1A). We observed that the percentage of TEs that do not contain their corresponding terminal repeats, or polyA tail sequences decreased in MCHelper libraries in comparison with RM2's raw libraries up to 20.5% and with REPET's raw libraries up to 11.6% (Table S2).

MCHelper recovers up to ~45% of the TEs originally labeled as "unclassified/unknown"

Another common issue with raw libraries is the high number of unclassified TEs, especially in non-model organisms due to the lack of TE consensus sequences for these species in the available TE databases. To recover a classification for 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) (Fig S1C; see Material and Methods). Using this approach, MCHelper was able to assign a TE class for up to 44.9% of the TEs initially labeled as "unclassified/unknown" TEs (Figure 3A and Table S3A). The majority of TEs in the three genomes analyzed were re-classified in the homology step, with the REPET output showing the bigger improvement compared to RM2 (Figure 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 three genomes analyzed (Figure 3B and Table S3B).

Genome-wide TE annotations with MCHelper libraries are more accurate

Genome-wide TE annotations performed with raw TE libraries often lead to overestimation of the number of TE copies, mostly because a single copy is annotated as several fragmented copies (Jamilloux et al., 2016). In addition, raw libraries may have chimeric consensus sequences leading to the annotation of a single TE copy by multiple consensus sequences (M. Rodriguez & 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 RM2 library, and more similar to the number of copies annotated with the reference library (Figure 4A and Table S4A): 20.57%, 16.97% and 11.20% less copies in *D. melanogaster*, *O. sativa* and *D. rerio*, respectively. Also, a 5.74% and 6.50% decrease in the number of copies was found in the *D. melanogaster* and *D. rerio* annotations performed with the MCHelper using REPET's outputs. While in *O. sativa* a 2.21% increase in the number of MCHelper copies was found when compared with the raw library, the number of MCHelper copies is more similar to the number obtained with the reference 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 three species when comparing with the raw RM2 annotations (11.84%, 20.53% and 19.8% fewer short copies in *D. melanogaster*, *O. sativa* and *D. rerio*, respectively). When comparing MCHelper with REPET, the number of short annotated copies was also reduced in *D. rerio* (22.15%) whereas there were similar numbers in *D. melanogaster* (1,799 Raw vs 1,800 MCHelper) and in *O. sativa* (12,237 Raw vs 12,317 MCHelper; Figure 4B and Table S4B). Thus, we checked whether the short insertions annotated with the MCHelper library in *D. melanogaster* were present in the reference annotation, suggesting they are real insertions, or whether they were enriched for false positive annotations. We found that 44.27% of these short copies were also annotated as TE insertions (overlapping of 100%) or as a part of a TE in the reference annotation. On the other hand, 226 (12.55%) short copies annotated with the MCHelper library were enriched for simple sequence repeats (SSR>60% of the annotation length), suggesting that these short annotations were indeed false positives. The remaining 43.18% copies may also be false positives as some *de novo* tools are not able to accurately annotate short copies (Rech et al., 2022).

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 (Figure 4C and Table S4C). The MCHelper library led to a decrease of

overlapping annotations of 30.4% (RM2) and 19.9% (REPET) in *D. melanogaster*, of 14.4% (RM2) and 7% (REPET) in *O. sativa*, and of 12.0% (RM2) and 16.5% (REPET) in *D. rerio*. 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. To further investigate this, we manually inspected the 100 consensus sequences of the REPET library that annotated the largest number of overlapping copies (47%) in the *O. sativa* genome. We found that 39 of these consensus sequences were indeed chimeric, while the other 61 were non-chimeric (Table S4D). Finally, we plot the number of overlapping annotations *vs* the length of the 100 consensus sequences, and we found that the consensus sequences >8 kb accounted for 56.1% of the chimeric consensus sequences while they only contained 1.9% of the non-chimeric consensus, suggesting that consensus sequences >8kb are enriched for being chimeric (Figure S3 and Table S4D).

Annotations made by MCHelper libraries are longer and more contiguous than 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., 2016). 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., 2016).

Our results showed that MCHelper libraries produced longer annotations than the raw library across tools and genomes (Figure 5 and Supplementary Data 1). The NTE50 values were smaller when annotating the genome with the MCHelper library compared with the raw library: 18.8% (RM2) and 9.1% (REPET) smaller in *D. melanogaster*, 27.4% (RM2) and 9.87% (REPET) smaller in *O. sativa*, and 30.4% (RM2) and 12.10% (REPET) smaller in *D. rerio* (Figure 5 and Table 1). As shown in Table 1, and as expected if the quality of MCHelper TE annotations is higher than the annotation performed with the raw library, the LTE50 values increased by 1.4x

(RM2) and 1.2x (REPET) in *D. melanogaster*, by 1.5x (RM2) and 1.2x (REPET) in *O. sativa*, and by 1.4x (RM2) and 1.1x (REPET) in *D. rerio* when comparing MCHelper libraries to raw libraries.

Finally, we also examined the proportions of each TE order in the annotations produced by the raw, MCHelper, and reference libraries. In all cases, the proportions of TE orders were higher in annotations performed with the MCHelper libraries than in annotations with raw libraries (Figure 6 and Table S5). 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 (Figure 4A) (31.1% REPET vs 37.5% EDTA library). In *D. rerio*, we also found that the raw library from REPET annotated lower TE proportions compared with MCHelper library and the reference library (34.4%, 44.7% and 54.4%, respectively).
Overall, MCHelper libraries produce longer and less fragmented annotations than raw libraries in all tested genomes and across *de novo* tools. Importantly, and as expected, although the number of consensus sequences in the MCHelper libraries was smaller compared with the raw libraries (Figure 2A), the proportions of TE orders annotated with the MCHelper libraries were closer to the ones produced with the reference libraries (Figure 6).

MCHelper curates TE libraries in hours 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 ten times for each raw library of the three species produced by the two *de novo* tools used in this study, on a server with 48 CPUs. We measured the execution time for ten distinct processes included in the MCHelper tool (see Material and Methods). On average, MCHelper took between 1.42 (*D. melanogaster* - RM2) and 29.91 (*D. rerio* - RM2) 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.

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 species, by integrating and automating TE curation protocols (Figure 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 et al 2022, Storer et al 2022). Based on several benchmarking metrics, we showed that MCHelper's automatically curated libraries for *D. melanogaster*, *O. sativa* and *D, rerio* are more accurate and complete compared with raw libraries, and thus more similar to high-quality manually curated (reference) libraries (Figure 2). Moreover, 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 45% of these sequences (Figure 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 2020).

MCHelper is a fast tool which can be executed in parallel in multiple CPUs taking advantage of currently available supercomputing resources (Table 2). MCHelper allows to curate raw TE libraries in hours with the Consensus extension module being the more time-consuming step. 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 run time, 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 2021; 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 2022). 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 2021; Storer et al 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 (Figure 1 and Figure 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 semi-automatic 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 towards understanding the dynamics of these important genome components but also the biology of genomes given their important roles in function, structure and evolution.

Material and methods

MCHelper workflow

First, MCHelper conducts pre-processing analysis, ensuring file completeness and consistency with Wicker's classification nomenclature (Wicker et al., 2007). For the raw libraries of RM2 (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 Supplementary Material and Methods), as well as satellites and RNAs and creates a file with the list of excluded sequences and the reason of exclusion. For the raw libraries from REPET, MCHelper checks that all the necessary paths and files exist (see github for more details). After checking that all the required files are correct, MCHelper will start the steps to do the curation (Figure 1).

Redundancy reduction. MCHelper reduces redundancy by clustering all consensus sequences using CD-HIT v.4.8.1 (Li & Godzik, 2006) 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; Baril et al., 2022; Platt et al., 2016; J. M. Storer et al., 2021). 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 multiple sequence alignment generated (see Supplementary Material and Methods). All consensus sequences are extended until the sequence reaches both TE ends or until the end of the iterations (see Supplementary Material and Methods).

False positive filtering. Consensus sequences with (i) very few full-length fragments (FLF) in the genome (by default one FLF but this parameter can be defined by the user; Jamilloux et al., 2016); (ii) having homology with multi-copy genes or rRNAs; and (iii) containing simple sequence repeats (SSR) (by default 60% also adjustable by the user) are considered false positive consensus sequences and filtered out (see Supplementary Materials and Methods; J. M. Storer et al., 2021, 2022).

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 (Wicker et al., 2007) are saved in the final curated library (see Supplementary Materials and Methods). Those that do not fulfill the 80-80-80 rule are divided into classified and unclassified sequences and follow two distinct paths (Figure 1).

Structural Checking. For classified sequences, MCHelper looks for structural information: presence of protein domains, terminal repeats (TRs), and polyA tails, to check if the original classification of the consensus sequence matches the expected TE features (see Supplementary Materials and Methods; Goubert et al., 2022). In the MCHelper semi-automatic 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 Supplementary Materials and 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 Supplementary Materials and Methods).

Manual inspection. MCHelper allows the user to manually visualize useful information to decide if a consensus sequence should be kept, removed or re-classified. All the graphical information is generated by TE+Aid v.1.0 (Goubert et al., 2022) except the multiple sequence alignment 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 Supplementary Materials and Methods).

Genome assemblies used

We evaluated the performance of the MCHelper tool on three different species with manually curated TE libraries of high quality: *D. melanogaster*, *O. sativa*, and *D. rerio*. These species were selected based on their varying genome sizes (180 Mb (Ashburner et al., 2005; Berlin et al., 2015), 466 Mb (Yu et al., 2002; Kawahara et al., 2013), and 1.7 Gb (Howe et al., 2013), 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) and 59.5% in *D. rerio* (Chang et al., 2022).

Source of the raw libraries

To obtain the raw TE libraries, we utilized two de novo tools, namely RepeatModeler2 (RM2; Flynn et al., 2020) 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). Regarding

the REPET libraries, we used the same assemblies that were used for the RM2 library generation: GenBank assemblies GCA_002050065.1 for *D. melanogaster*, and GCA_000002035.4 for *D. rerio*, and the RGAP assembly version 7 for *O. sativa* (available at https://phytozome-next.jgi.doe.gov/info/Osativa_v7_0; Ouyang et al., 2007). 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). Furthermore, the REPET group kindly provided us with the output generated with the *O. sativa* genome available in the RepetDB database (Amselem et al., 2019). 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 2016).

Source of the reference libraries

As the reference libraries, we used the Berkeley Drosophila Genome Project (BDGP) dataset (Kaminker wt al., 2002) and the Manual Curated TE library (MCTE, Rech, et al., 2022) for *D. melanogaster*, the library published by Ou et al. (2019) for *O. sativa* (referred to as "EDTA library" in this study), and the manual curated TE models available in Dfam release 3.7 (Storer et al., 2021) and in Repbase version 20181026 for *D. rerio*. For the EDTA, 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 was executed with default parameters (to see customizable parameters see 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*, the viriplantae set (viriplantae_odb10), and for *D. rerio*, the actinopterygii set (actinopterygii_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.

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., 2016; Storer et al., 2021; Storer et al., 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 not present 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.

To measure accuracy, we used the number of false positive sequences as calculated in the False Positive filtering step of MCHelper (see Supplementary Materials and Methods). Briefly, we calculated the number of consensus sequences with >60% of their sequence length corresponding to single sequence repeats (SSR), 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 three 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. 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/) with parameters "-d -1 -c 9,9 -o count,collapse" to merge the overlapping copies and 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 2016). 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 at https://github.com/GonzalezLab/MCHelper/tree/main/benchmarking_scripts.

Short annotations analysis

To check if the short copies annotated by the MCHelper library obtained from REPET in the *D. melanogaster* genome were real insertions, we compared their coordinates with the coordinates of the annotation made by using the reference library. To do this, we used bedtools intersect with the following parameters: “-f 1 -wa”. Short copies that overlapped between the two annotations were considered as real insertions and the rest were analyzed with the TRF program (parameters: 2 3 5 80 10 20 15 -h -d) to look for simple repeats.

Manual inspection of the TE consensus that annotated overlapping copies

For the MCHelper library produced from the REPET library of *O. sativa*, we conducted a manual inspection to assess the quality of the consensus sequences that annotated overlapping copies. We employed TE-Aid (Goubert et al., 2022) to examine the structural characteristics, genomic coverage, and self-dot plot of each TE model. By scrutinizing this information, we were

able to identify and label consensus sequences as "chimeric" if they exhibited evidence of combining two or more TE sequences in the same structure. We manually reviewed multiple hits with domains from different TE orders, unusual patterns in the coverage plot, divergent patterns in comparison to the consensus, and those that contained only a few full-length fragments.

Run-time 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 ten times to see the variability of the execution times and we calculated the average and the standard deviation of the ten executions. Besides measuring total time, we calculated the execution time of ten subroutines: pre-processing, 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.

Acknowledgements

This project is funded by by MCIN/AEI/ 10.13039/501100011033 (grant number PID2020-115874GB-I00) awarded to J.G. P.S. is funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956229. R.D. acknowledges funding from Wellcome grant 207492.

References

Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., Maumus, F., Letellier, T., Luyten, I., Pommier, C., & others. (2019). RepetDB: a unified resource for transposable element references. *Mobile DNA*, *10*, 1-8.

Ashburner, M., Hawley, R., & Golic, K. (2005). *Drosophila. A Laboratory Handbook. Cold Spring Harbor Lab*.

Bailly-Bechet, M., Haudry, A., & Lerat, E. (2014). “One code to find them all”: A perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, *5*, 1-15.

Baril, T., Imrie, R. M., & Hayward, A. (2022). *Earl Grey: A fully automated user-friendly transposable element annotation and analysis pipeline*.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic acids research*, *27*(2), 573-580.

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, *33*(6), 623-630.

Casacuberta, E., & González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular ecology*, *22*(6), 1503-1517.

Chang, N.-C., Rovira, Q., Wells, J., Feschotte, C., & Vaquerizas, J. M. (2022). Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Research*, *32*(7), 1408-1423.

Consortium, D. T. of L. P. (2022). Sequence locally, think globally: The Darwin tree of life project. *Proceedings of the National Academy of Sciences*, *119*(4), e2115642118.

De Cecco, M., Ito, T., Petrashen, A. P., Elias, A. E., Skvir, N. J., Criscione, S. W., Caligiana, A., Brocculi, G., Adney, E. M., Boeke, J. D., & others. (2019). L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, *566*(7742), 73-78.

Dunn-Fletcher, C. E., Muglia, L. M., Pavlicev, M., Wolf, G., Sun, M.-A., Hu, Y.-C., Huffman, E., Tumukuntala, S., Thiele, K., Mukherjee, A., & others. (2018). Anthropoid primate–specific retroviral element THE1B controls expression of CRH in placenta and alters gestation length. *PLoS biology*, *16*(9), e2006337.

Elliott, T. A., Heitkam, T., Hubley, R., Quesneville, H., Suh, A., & Wheeler, T. J. (2021). TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mobile DNA*, *12*(1), 1-5.

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS one*, *6*(1), e16526.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451-9457.

Gilbert, C., Peccoud, J., & Cordaux, R. (2020). Transposable elements and the evolution of insects. *Annual Review of Entomology*, *66*, 355-372.

Goubert, C., Craig, R. J., Bilat, A. F., Peona, V., Vogan, A. A., & Protasio, A. V. (2022). A beginner's guide to manual curation of transposable elements. *Mobile DNA*, *13*(1), 1-19.

Hayward, A., & Gilbert, C. (2022). Transposable elements. *Current Biology*, *32*(17), R904-R909.

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., & others. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498-503.

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H. A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H., & others. (2016). Discovery of an active RAG transposon illuminates the origins of V (D) J recombination. *Cell*, *166*(1), 102-114.

Jamilloux, V., Daron, J., Choulet, F., & Quesneville, H. (2016). De novo annotation of transposable elements: Tackling the fat genome issue. *Proceedings of the IEEE*, *105*(3), 474-481.

Jebb, D., Huang, Z., Pippel, M., Hughes, G. M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L. S., Skirmuntt, E. C., Katzourakis, A., & others. (2020). Six reference-quality genomes reveal evolution of bat adaptations. *Nature*, *583*(7817), 578-584.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, *110*(1-4), 462-467.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., & others. (2002). The transposable elements of the Drosophila melanogaster euchromatin: A genomics perspective. *Genome biology*, *3*(12), 1-20.

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., & others. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, *6*, 1-10.

Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., & others. (2022). The earth BioGenome project 2020: Starting the clock. En *Proceedings of the National Academy of Sciences* (Vol. 119, Número 4, p. e2115635118). National Acad Sciences.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658-1659.

Mérel, V., Boulesteix, M., Fablet, M., & Vieira, C. (2020). Transposable elements in Drosophila. *Mobile DNA*, *11*, 1-20.

Orozco-Arias, S., Candamil-Cortes, M. S., Jaimes, P. A., Valencia-Castrillon, E., Tabares-Soto, R., Isaza, G., & Guyot, R. (2022). Automatic curation of LTR retrotransposon libraries from plant genomes through machine learning. *Journal of Integrative Bioinformatics*, *19*(3), 20210036.

Orozco-Arias, S., Candamil-Cortés, M. S., Valencia-Castrillón, E., Jaimes, P. A., Orozco, N. T., 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. *2021 IEEE 2nd International Congress of Biomedical Engineering and Bioengineering (CI-IB&BI)*, 1-4.

Osmanski, A. B., Paulat, N. S., Korstian, J., Grimshaw, J. R., Halsey, M., Sullivan, K. A., Moreno-Santillán, D. D., Crookshanks, C., Roberts, J., Garcia, C., & others. (2023). Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science*, *380*(6643), eabn1430.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., & others. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology*, *20*(1), 1-18.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., & others. (2007). The TIGR rice genome annotation resource: Improvements and new features. *Nucleic acids research*, *35*(suppl_1), D883-D887.

Pastuzyn, E. D., Day, C. E., Kearns, R. B., Kyrke-Smith, M., Taibi, A. V., McCormick, J., Yoder, N., Belnap, D. M., Erlendsson, S., Morado, D. R., & others. (2018). The neuronal gene arc encodes a repurposed retrotransposon gag protein that mediates intercellular RNA transfer. *Cell*, *172*(1), 275-288.

Payer, L. M., & Burns, K. H. (2019). Transposable elements in human genetic disease. *Nature Reviews Genetics*, *20*(12), 760-772.

Petersen, M., Armisén, D., Gibbs, R. A., 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 Ecology and Evolution*, *19*(1), 1-15.

Platt, R. N., Blanco-Berdugo, L., & Ray, D. A. (2016). Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome biology and evolution*, *8*(2), 403-410.

Rech, G. E., 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. *Nature Communications*, *13*(1), 1948.

Rodriguez, M., & Makalowski, W. (2022). Software evaluation for de novo detection of transposons. *Mobile DNA*, *13*(1), 1-14.

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, *19*(6), 329-346.

Smit, A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0. 2013–2015*.

Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2016). Evolution and diversity of transposable elements in vertebrate genomes. *Genome biology and evolution*, *9*(1), 161-177.

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, *12*(1), 1-14.

Storer, J. M., Hubley, R., Rosen, J., & Smit, A. F. (2021). Curation guidelines for de novo generated transposable element families. *Current protocols*, *1*(6), e154.

Storer, J. M., Hubley, R., Rosen, J., & Smit, A. F. (2022). Methodologies for the de novo discovery of transposable element families. *Genes*, *13*(4), 709.

Tumescheit, C., Firth, A. E., & Brown, K. (2022). CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ*, *10*, e12983.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., & others. (2007). A unified classification system for eukaryotic transposable elements. *Nature reviews genetics*, *8*(12), 973-982.

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., & others. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. Indica). *science*, *296*(5565), 79-92.

Figures and Tables

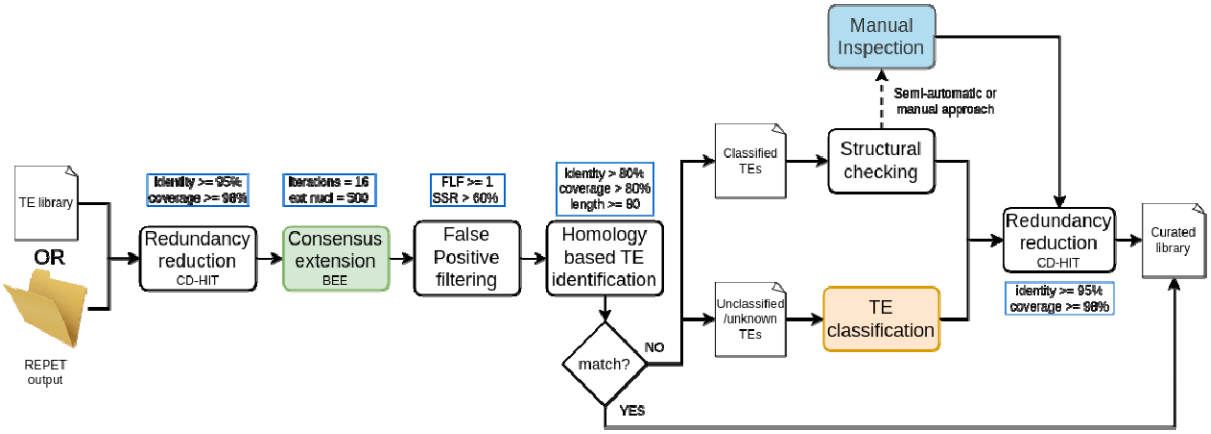

Figure 1. MCHelper's tool workflow
Inputs, main steps, and default parameters of the MCHelper tool. MCHelper accepts as input a single *fasta* file containing consensus sequences produced by a *de novo* TE identification tool or a folder containing files produced by the *TEdenovo* pipeline from REPET (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 Figure S1.

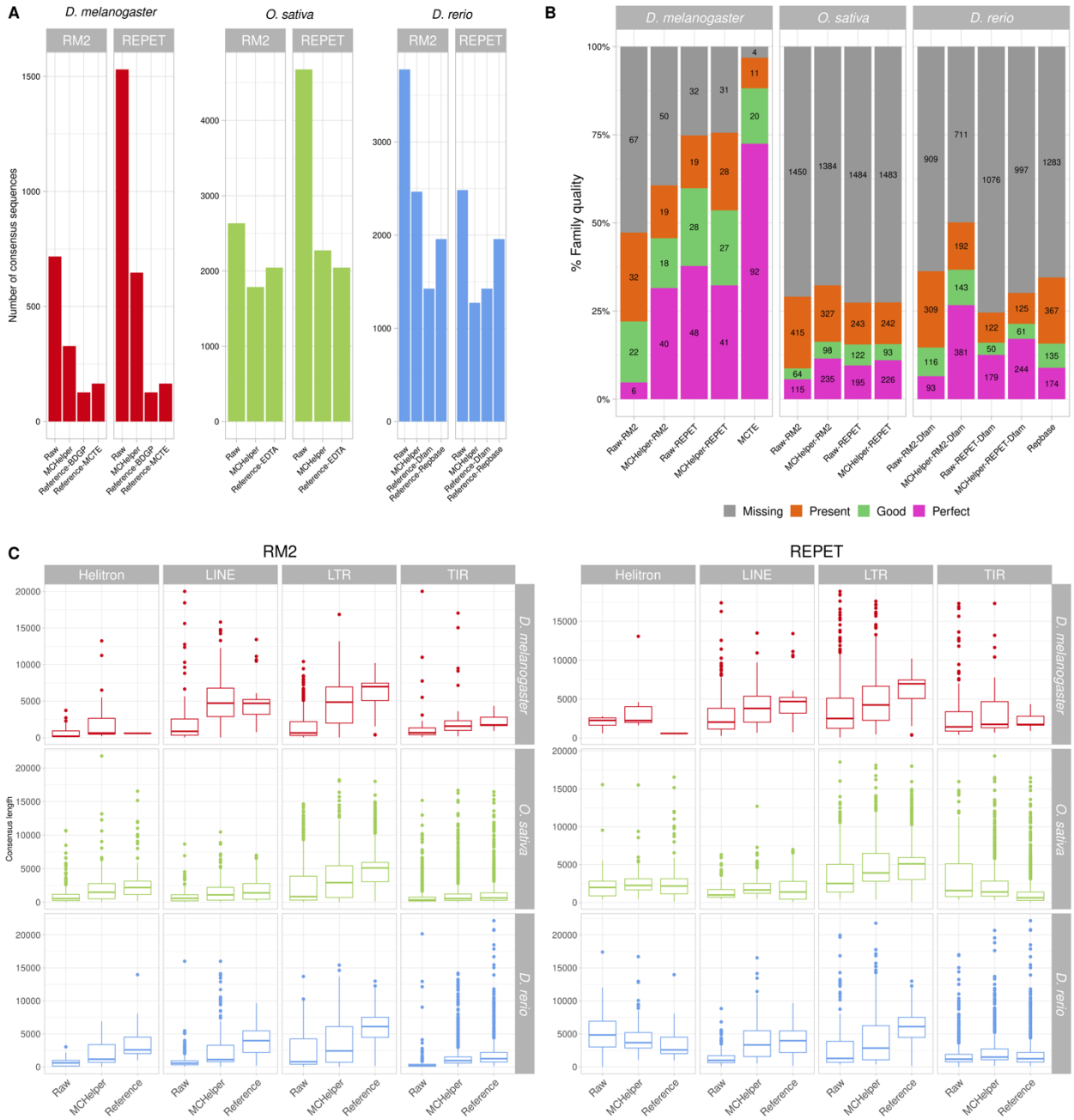

Figure 2. Comparison of MCHelper libraries with raw and reference libraries in the three genomes analyzed

A) Number of consensus sequences contained in the raw library, the MCHelper library and the reference libraries of the three species analyzed. Reference libraries used are: BDGP (Kaminker et al., 2002) and MCTE (Rech et al., 2022) for *D. melanogaster*, EDTA (Ou et al., 2019) for *O. sativa*, and Dfam (J. Storer et al., 2021) and Repbase (Jurka et al., 2005) for *D. rerio.* B) Number of consensus sequences classified as Perfect, Good, Present and Missing families following the methodology proposed by Flynn et al. (2020). C) Consensus length distribution of each TE order in the three libraries.

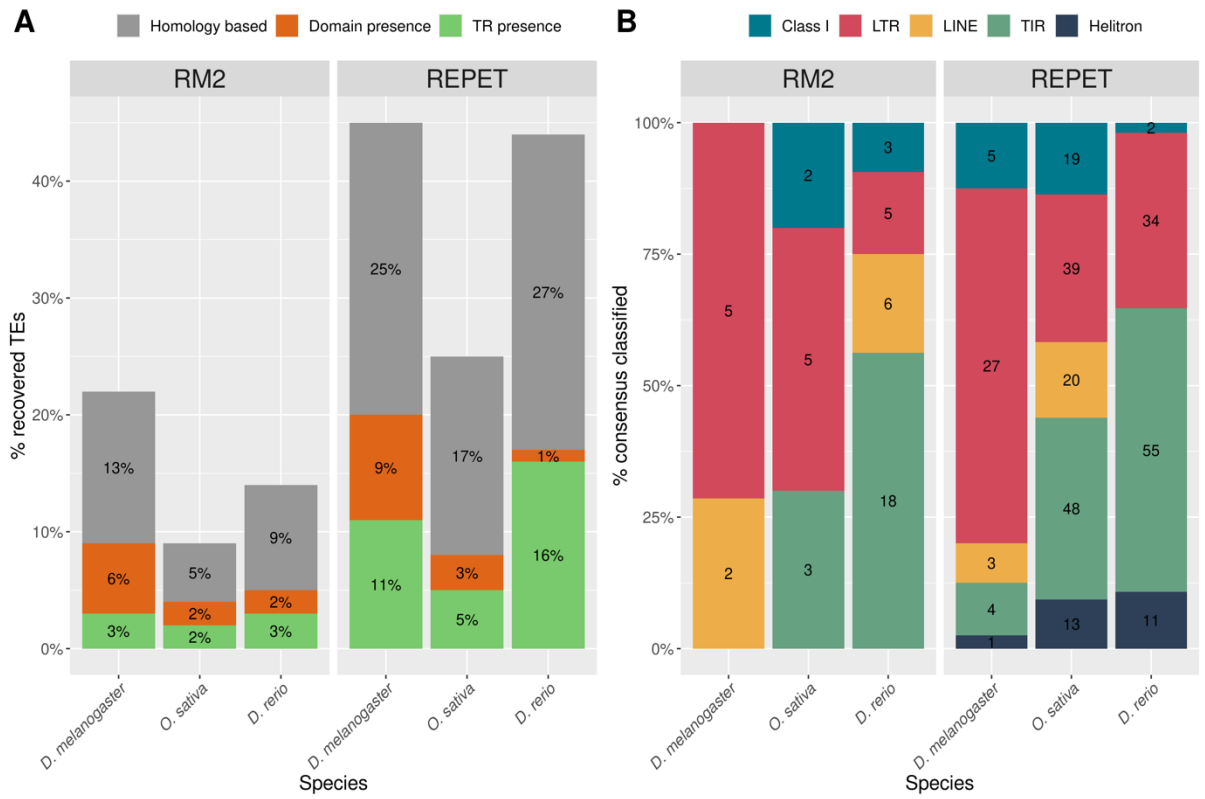

Figure 3: "Unclassified/unknown" consensus sequences in the raw libraries that are classified by MCHelper

A) Proportion of TEs originally labeled as "unclassified/unknown", which were recovered by the three-step approach of MCHelper. B) Number of recovered TE sequences by TE order/class. Class I indicates all the elements that contain one or more retrotransposon-like domains, but it was not possible to assign a deeper classification.

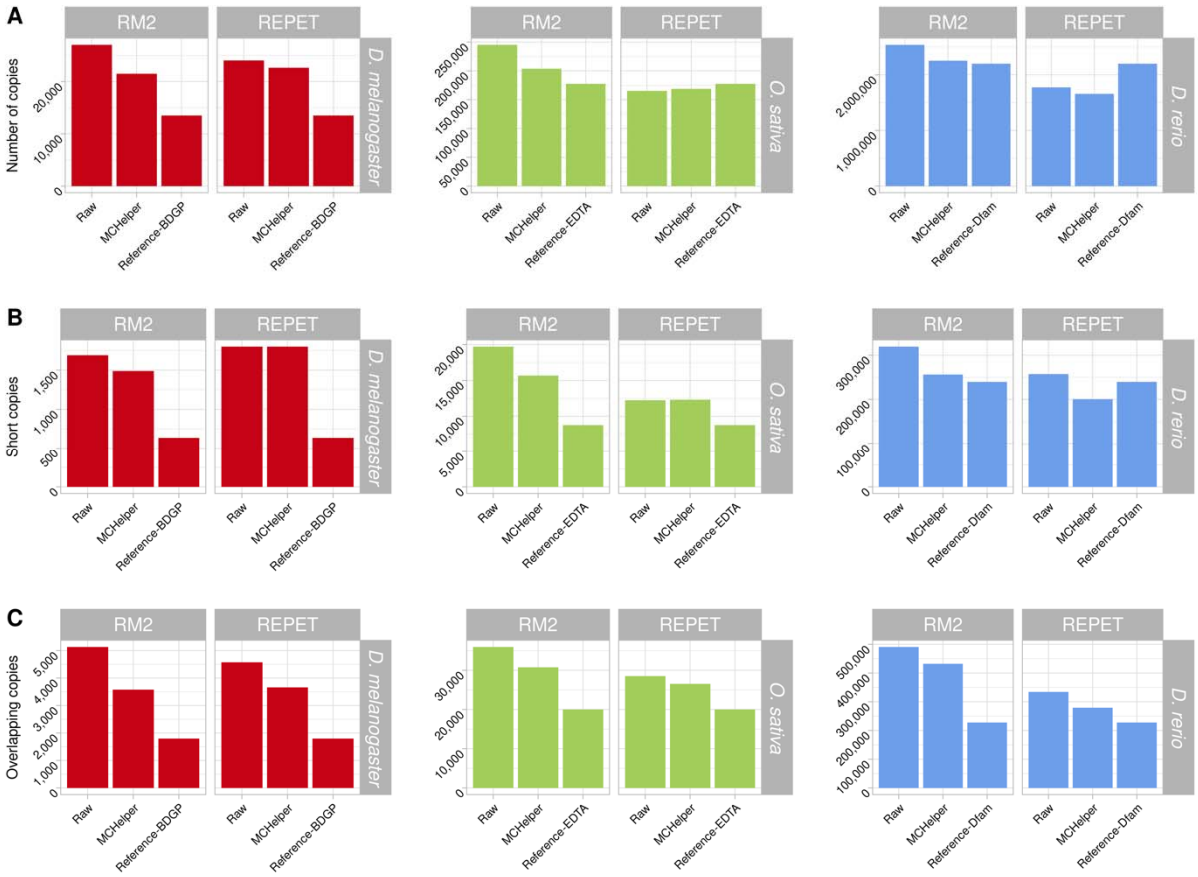

Figure 4. Genome-wide number and type of TE annotations for each of the three libraries analyzed in the three species

The BDGP, EDTA and Dfam libraries were used as reference for *D. melanogaster*, *O. sativa and D. rerio* respectively. A) Number of copies annotated by each library. B) Number of copies categorized as short (<100bp). C) Number of copies categorized as overlapping (annotated by two or more consensus sequences).

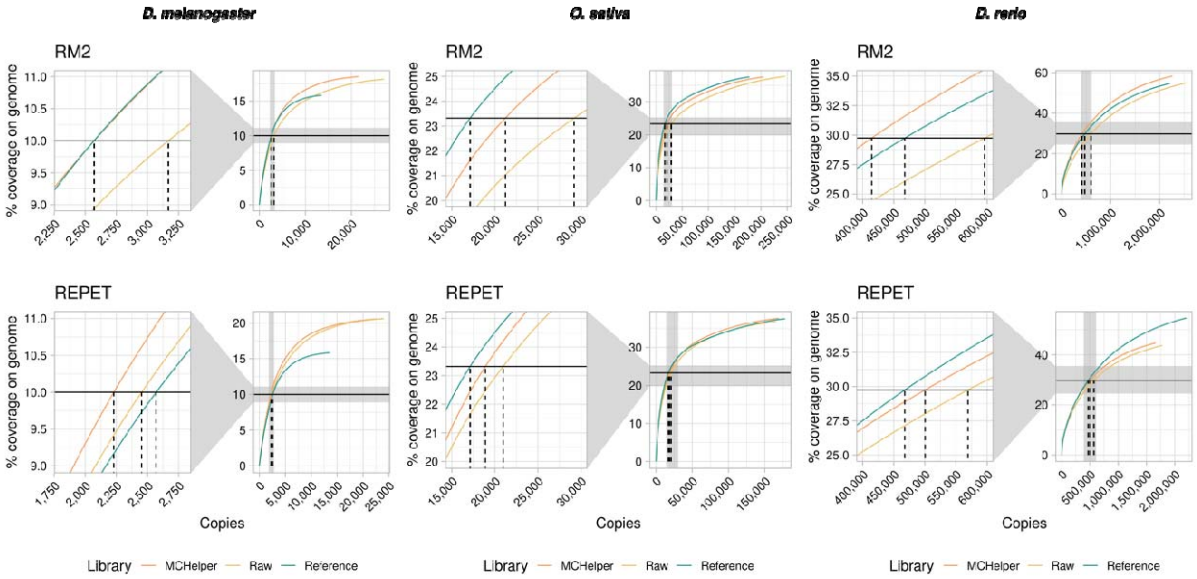

Figure 5. Cumulative length coverage plot
The black horizontal line represents half of the mobilome and it was used to calculate the LTE50 values (Table 1). The dash vertical line indicates the NTE50 values. For each graph, a zoom capture is provided on the left to show when the cumulative length plot of each library crosses the black horizontal line (NTE50 values). Mobilome proportions are 20% in *D. melanogaster* (Goubert et al., 2015; Hill 2019; Mérel et al., 2020), 46.64% in *O. sativa* (Ou et al., 2019) and 59.5% in *D. rerio* (Chang et al., 2022). NTE50 and LTE50 values are shown in Table 1.

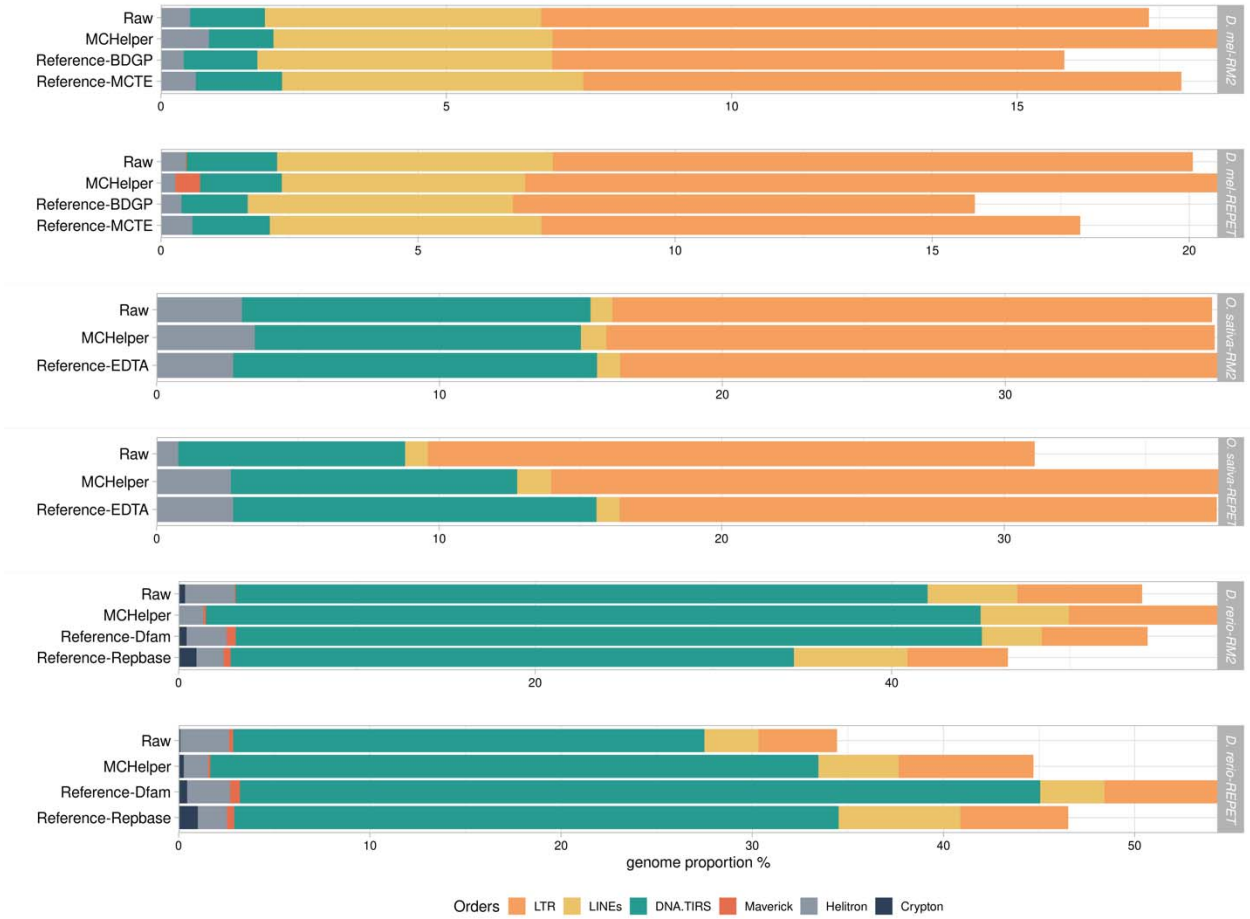

Figure 6. Genomic proportions of TE orders annotated by the raw, MCHelper and reference libraries

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.

Table 1. NTE50 and LTE50 values of each library

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., 2016).

	***D. melanogaster* (RM2)**		***D. melanogaster* (REPET)**		***O. sativa* (RM2)**		***O. sativa* (REPET)**		***D. rerio* (RM2)**		***D. rerio* (REPET)**	
Library	**NTE50**	**LTE50**	**NTE50**	**LTE50**	**NTE50**	**LTE50**	**NTE50**	**LTE50**	**NTE50**	**LTE50**	**NTE50**	**LTE50**
Raw	3167	2188	2453	3429	29221	918	20958	1497	596048	374	569157	332
MCHelper	2571	3007	2229	4034	21216	1400	18888	1773	414669	508	500255	362
Reference	2570	3158	2570	3158	17119	1800	17119	1800	467769	433	467769	433

Table 2. MCHelper average runtimes for the three species analyzed
MCHelper was executed 10 times for each raw library (RM2 and REPET) and the average and standard variation (STD) was calculated. Total runtimes are given in hours.

Species	**RM2**		**REPET**	
	Total runtime (h)	**Extension step runtime (%)**	**Total runtime (h)**	**Extension step runtime (%)**
D. melanogaster	1.42 ± 0.004	76.87%	3.46 ± 0.004	81.67%
O. sativa	10.82 ± 0.050	83.64%	16.55 ± 0.084	79.76%
D. rerio	29.91 ± 0.094	95.91%	27.92 ± 0.60	70.12%