Received: 19 March 2021 | Accepted: 12 August 2021

DOI: 10.1111/1755-0998.13489

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines

Ellen A. Bell[1] | Christopher L. Butler[1] | Claudio Oliveira[2] | Sarah Marburger[3] | Levi Yant[4] | Martin I. Taylor[1]

[1]School of Biological Sciences, University of East Anglia, Norwich, Norfolk, UK

[2]Department of Structural and Functional Biology, Institute of Biosciences/UNESP, Rua Professor Doutor Antonio Celso Wagner Zanin, Botucatu, São Paulo, Brazil

[3]Department of Cell and Developmental Biology, John Innes Centre, Norwich, Norfolk, UK

[4]Future Food Beacon of Excellence and the School of Life Sciences, University of Nottingham, Nottingham, UK

Correspondence
Ellen A. Bell and Christopher L. Butler, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, Norfolk, UK.
Emails: ellen.bell@uea.ac.uk and c.butler@uea.ac.uk

Funding information
This work was supported by the Natural Environment Research Council (NERC) and EnvEast doctoral training partnership (DTP) (grant no. NE/L002582/1) awarded to CB; the Biotechnology and Biological Science Research Council (BBSRC) (grant no. BB/R017174/1) awarded to MIT; the National Council for Scientific and Technological Development (CNPq) and Sao Paulo Research Foundation (FAPESP) under a Brazilian Science Without Borders visiting fellowship scheme awarded to MIT; the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. ERC-StG 679056 HOTSPOT) awarded to LY; the BBSRC (grant number BB/P013511/1) awarded to the John Innes Centre (LY) and the Norwich Research Park seed links fund awarded to MIT and LY

Abstract

Transposable elements (TEs) are significant genomic components which can be detected either through sequence homology against existing databases or de novo, with the latter potentially reducing the risk of underestimating TE abundance. Here, we describe the semi-automated generation of a de novo TE library using the newly developed EDTA pipeline and DeepTE classifier in a non-model teleost (*Corydoras fulleri*). Using both genomic and transcriptomic data, we assess this de novo pipeline's performance across four TE based metrics: (i) abundance, (ii) composition, (iii) fragmentation, and (iv) age distributions. We then compare the results to those found when using a curated teleost library (*Danio rerio*). We identify quantitative differences in these metrics and highlight how TE library choice can have major impacts on TE-based estimates in non-model species.

KEYWORDS
Corydoras, de novo, genomics, teleost, transcriptomics, transposon annotation

Ellen A. Bell and Christopher L. Butler contributed equally to this work and should be considered as joint first authors.

1 | INTRODUCTION

Transposable elements (TEs) are sequences of repetitive, non-coding DNA found in high abundance across the tree of life (Bourque et al., 2018; Wells & Feschotte, 2020; Wicker et al., 2007). Historically overlooked during genomic analysis and annotation, TEs are now recognised as key contributors to genome evolution and regulation, providing alternative promoters, neofunctionalisation, novel exons, and large-scale rearrangements (Bourque et al., 2018; Cowley and Oakey 2013; Hoen & Bureau, 2015). This realisation, coupled with the increased availability of genome sequences, has generated a growing need for both accessible and comprehensive TE annotation in non-model species.

TEs can be detected using either homology or de novo approaches. Homology-based approaches detect TEs through sequence comparisons against existing databases, whilst de novo approaches identify TEs through signatures such as structure or elevated copy number (Kennedy et al., 2011; Ou et al., 2019). Homology searches may lead to TE underestimates because sequence divergence can render certain TEs unrecognisable, which may be particularly common in non-model organisms where there may be large phylogenetic distances to their closest database entry (Bergman & Quesneville, 2007). Furthermore, due to their potential absence from databases, homology-based searches may bias detection away from species-specific TEs which have inserted since the common ancestor of focal species and library (Platt et al., 2016). This may be particularly true in the case of horizontally transferred TEs which are increasingly recognised to move between vertebrate genomes and may be important for long term TE persistence (Groth & Blumenstiel, 2017; Panaud, 2016; Zhang et al.,2020). Consequently, the generation of TE libraries that do not rely solely on homology-based searching is recommended (Hoen & Bureau, 2015; Platt et al., 2016). However, de novo TE libraries also have disadvantages, as they may fail to detect low-copy number elements or erroneously identify/classify TEs (Bergman & Quesneville, 2007). De novo TE libraries may therefore require a degree of manual curation, which can be both time consuming and labour intensive.

Several semi-automated pipelines for de novo library construction have been created to streamline their development, both in terms of annotation and classification. These include the Extensive de novo TE Annotator (EDTA) (Ou et al., 2019) and DeepTE (Yan et al., 2020). EDTA combines a suite of best-performing packages (LTR_FINDER, LTRharvest, LTR_retriever, Generic Repeat Finder, TIR Learner, HelitronScanner and RepeatMasker) to produce nonredundant TE libraries. EDTA also has an option to use RepeatModeler to do a final sweep for remaining unidentified TEs, thereby utilising two very powerful TE annotating tools (Ou et al., 2019). After initially performing well in rice (*Oryza sativa*, Ou et al., 2019), EDTA has subsequently been run across numerous nonvertebrate genomes, including sweet corn (*Zea mays*, Hu et al., 2021), field mustard (*Brassica rapa*, Cai et al., 2021) and sawfly (*Euura lappo*, Michell et al., 2021). DeepTE classifies TEs using machine learning, specifically by using convolutional neural networks to assign TEs to superfamily and order, with good performance in terms of accuracy and sensitivity against other similar classifiers as well in the assignment of previously unknown TEs (Yan et al., 2020). However, the impacts of EDTA's implementation on TE annotation in non-model, vertebrate genomes, particularly when combined with DeepTE, have yet to be fully explored.

In this study, we describe the use of both EDTA and DeepTE to construct a de novo library for *Corydoras fulleri*, a member of the Corydoradinae which are a species-rich subfamily of Neotropical catfishes with highly variable TE content (Alexandrou et al., 2011, Marburger et al., 2018). Teleost genomes contain the most abundant and diverse TE content of all vertebrates, including numerous horizontally integrated elements, making them interesting organisms to assess de novo pipelines (Sotero-Caio et al., 2017; Zhang et al., 2020). Here, we compare the performance of our *Corydoras*-specific TE library against the RepBase *D. rerio* library by estimating TE content within two *Corydoras* species; *Corydoras fulleri* and *Corydoras maculifer*. Specifically, our *Corydoras*-specific TE library was quantitatively assessed using estimates of four key TE-based metrics; (i) abundance, (ii) composition, (iii) fragmentation, (i.e., the likelihood that genomic TE copies have not been captured in a single contiguous manner during library creation), and (iv) sequence divergence distributions. We also use a mixture of both genomic and transcriptomic sequences to test how library type affects TE landscapes across different transposon age groups. Finally, we present this pipeline as a GitHub resource that will be applicable to a diverse range of species in the future (Figure 1, https://github.com/ellenbell/FasTE).

2 | MATERIALS AND METHODS

2.1 | Extraction and sequencing of DNA and genome assembly

The genome of *C. fulleri* was assembled using both long-read PacBio sequencing and short-read Illumina Sequencing. Genomic DNA was extracted from *C. fulleri* using MagAttract HMW DNA Kit (Qiagen) for high molecular weight PacBio sequencing and PureLink Genomic DNA mini kit (by ThermoFisher Scientific) for Illumina Hiseq. Sequencing was performed on two PacBio Sequel cells and one Hiseq lane using 300 bp paired-end reads, which was estimated to generate 60x long-read PacBio coverage and 100x Illumina Hiseq coverage. All genomic library preparation and sequencing of *C. fulleri* was performed by Novogene Co Ltd.

Genome assembly for *C. fulleri* was performed using wtdbg2 (version 2.5) to create an initial long-read assembly from PacBio data (Ruan & Li, 2019). This first pass assembly was then polished using wtdbg2-racon-pilon.pl v04 script (Schellt, 2019; https://github.com/schellt/wtdbg2-racon-pilon) which performs three iterative corrections, firstly with long-read mapping using minimap2 (version 2.17, Li, 2018) and polishing with Racon (version 1.4.15, Vaser et al., 2017)

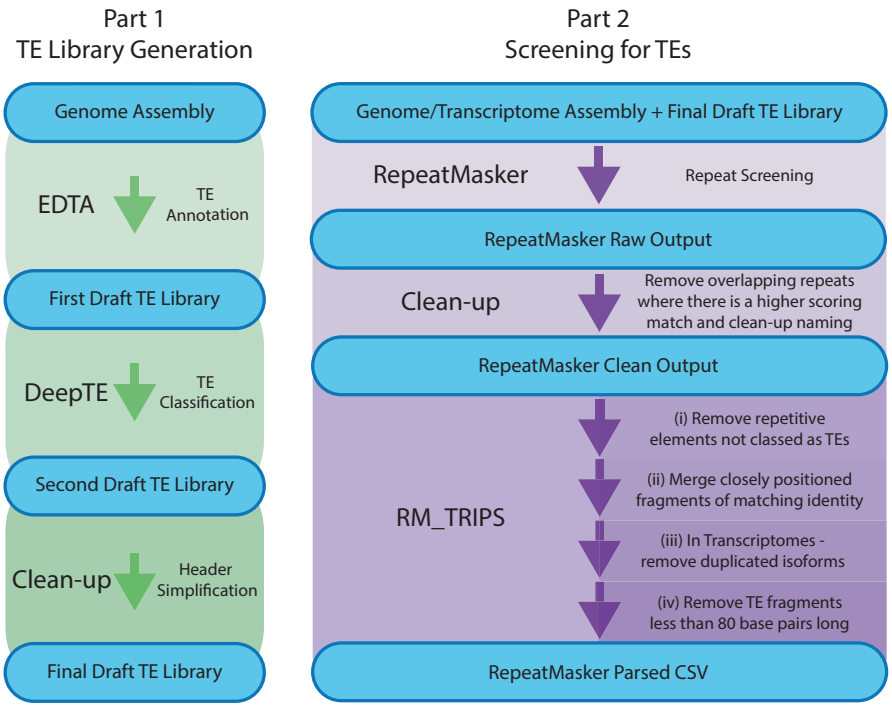

FIGURE 1 FasTE pipeline schematic. Part 1; The three major steps behind de novo TE library generation with EDTA and DeepTE. Part 2; Utilisation of RepeatMasker and de novo libraries to generate estimates of genome wide repeat abundance alongside subsequent parse steps with RM_Trips

and then with short-read mapping using bwa mem (version 0.7.17, Li, 2013), merging and sorting using Samtools (version 1.10, Li et al., 2009) and polishing with Pilon (version 1.23, Walker et al., 2014).

The genome of *C. maculifer* was assembled using short-read Illumina based sequencing. Genomic DNA was extracted from *C. maculifer* using Qiagen DNeasy Blood and Tissue extraction kit. Paired-end PCR-free libraries were produced and sequenced on a single lane of an Illumina Hiseq platform using 250 bp paired-end reads, estimated to provide 50X coverage. Twelve Nextera long mate paired (LMP) libraries were also generated and sequenced on a second lane of Illumina Hiseq using 300 bp paired-end reads from which the two libraries with the largest insert size were selected (average insert sizes 8678.2 bp and 8730.0 bp, respectively). These two libraries were then sequenced on an Illumina Hiseq platform with 250 bp paired-end reads to assist with scaffolding. All library preparation and sequencing of *C. maculifer* was performed by the Earlham Institute, Norwich. Paired-end libraries were assembled using w2rap-contigger (Clavijo et al. 2017) under default settings. LMP libraries were cleaned using NextClip (Leggett et al., 2014) and combined with contigs from paired-end assemblies using SOAPdenovo2 (Luo et al., 2012) under default settings but using a kmer size of 19 to produce scaffolds. Genome coverage for both assemblies was assessed using Quast (version 5.0.2, Gurevich et al., 2013) and completeness measured using BUSCO (version 4.1.0, Seppey et al., 2019) (Table S1).

2.2 | Extraction and sequencing of RNA and transcriptome assembly

The transcriptome of *C. maculifer* was assembled from short read Illumina based sequencing. RNA extraction (TRIzol Plus RNA Purification Kit) was conducted on somatic muscle tissue. The size selection and integrity of the extracted RNA was confirmed using an Agilent 2100 Bioanalyser (Agilent Technologies) which met internal QC standards of the sequencing provider. Transcriptomic library preparation and sequencing was performed by the Animal Biotechnology Laboratory of Esalq/Piracicaba and the cDNA library was then built using a TruSeq RNA Sample Prep kit (Illumina, Inc). The *C. maculifer* cDNA was sequenced using paired-end sequencing on an Illumina HiSeq (as part of a larger multiplexed run), generating 10.09 million paired reads. The library was then demultiplexed and cleaned using Trimmomatic (version 0.2.36, Bolger et al., 2014) and subsequently assembled using the de novo transcriptome assembler Trinity (version 2.6.9, Grabherr et al., 2013). Transcriptome quality was later assessed using TransRate (version 1.03, Smith-Unna et al., 2016) (Table S2).

2.3 | Transposable element annotation

A de novo TE library was generated from the long-read PacBio *C. fulleri* genome using the Extensive de novo TE Annotator (EDTA) (Ou et al., 2019) set to the "others" species parameter. We utilised the inbuilt RepeatModeller (Smit and Hubley, 2008) support which identifies any remaining TEs which might have been overlooked by the EDTA algorithm (--sensitive 1). Classifications within this library were refined using DeepTE using the predefined metazoan model parameter setting (-m) (Yan et al., 2020). TE identification was performed using RepeatMasker (RM; version 1.332) utilising the NCBI/RMBLAST (version 2.6.0+) search engine. This analysis was conducted either against the *Danio rerio* Repbase (26 October 2018) entry, which was also run through DeepTE (to allow for uniformity in

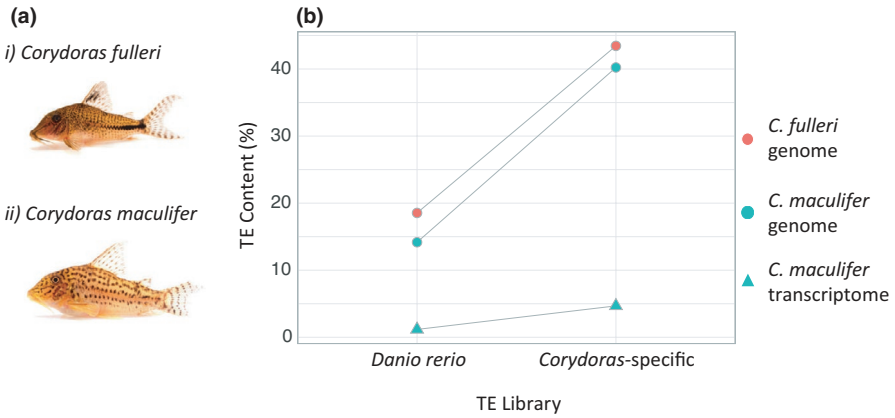

FIGURE 2 TE library type influences TE abundance. (a) the two *Corydoras* species used in this study (i) *Corydoras fulleri* and (ii) *Corydoras maculifer*. (b) Estimated TE abundance is given as percentage of total genome/transcriptome size for the C. *fulleri* genome and the *C. maculifer* genome and transcriptome

TE classification, referred to as the "*D. rerio* library" henceforth), or the *Corydoras*-specific library. RM was run under the most sensitive (-s) parameter setting in all instances. The genomic and transcriptomic RM output files were subsequently cleaned of nondistinct elements by removing overlapping repeats where a match with a higher likelihood score was available. Outputs were then parsed through a custom R script; RM_TRIPS (which is publically available at https://github.com/clbutler/RM_TRIPS). RM_TRIPS was used to (i) remove repetitive elements not classed as TEs (e.g., microsatellites, simple repeats & sRNAs), (ii) merge elements found on the same contig if they had the same name, orientation, and their combined sequence length was less than or equal to the corresponding reference sequence in the repeat library, (iii) remove merged repeats with a length less than 80 base pairs, and (iv) for transcriptomic data, if multiple identical repeats were found across different transcript isoforms, only one was retained. This was to ensure that each repeat represented a unique genomic locus.

This complete pipeline from de novo library generation through to RM output parsing has been consolidated into the annotated tool, FasTE, which is publically available at https://github.com/ellenbell/FasTE (see Figure 1).

2.4 | Comparative assessment of the performance of Corydoras-specific TE library

TE abundance estimates were calculated from parsed RM output files derived from the *Corydoras*-specific and *D. rerio* libraries. These were then standardised across both *Corydoras* species by calculating the percentage of total genome or transcriptome length (bp) represented by TEs. For compositional comparison, TEs were grouped into Helitrons, Maverick elements, DDE DNA elements, long terminal repeat retrotransposons (LTRs) (including dictyosteilium intermediate repeat sequences DIRS), long interspersed nuclear elements (LINEs), Penelope like elements (PLEs) and short interspersed nuclear elements (SINEs) (Wicker et al., 2007). These compositional comparisons were standardised across genomic and transcriptomic sequence data by scaling TE abundance by megabase (MB).

Library fragmentation was assessed firstly by visualising the cumulative abundance estimates of elements against the standardised number of TE entries within both the *Corydoras*-specific library and the *D. rerio* library. Second, we compared genomic TE lengths using the *Corydoras*-specific library against the C. *fulleri* genome and the *D. rerio* library against the *D. rerio* genome (GCF_000002035.6_GRCz11) (Howe et al., 2013).

Age distributions of TEs were compared across library types using their sequence divergence from library entry as a proxy. This made use of the RM outputs which reports the percentage of substitutions in a matching TE compared to its corresponding library hit. Age/sequence divergence distributions were generated for the four major TE classes - DNA transposons, LTR retrotransposons, SINEs and LINEs.

To investigate the potential origin of *C. maculifer* Mariner elements, we extracted every genomic copy with a matching length of >80% against its library hit, and every transcript copy where an element made up >80% of the transcript's length. We subsequently ran a BLASTn search against the RepeatMasker library, with elements potentially horizontally inherited if sequences had both (i) a best match (lowest E value) against a non-teleost species and (ii) following rationale used in (Rogers et al., 2018) a greater than 2% sequence similarity than its best teleost hit. Figures were produced using the ggplot2 package in R (Wickham, 2016).

3 | RESULTS

To assess the impact of de novo library creation using EDTA/DeepTE pipelines we generated a de novo TE library (*Corydoras*-specific) from a long-read (PacBio) *Corydoras fulleri* genome assembly and benchmarked it against the *D. rerio* RepBase entry. TE content was then assessed across two *Corydoras* species and sequence types including: (i) a *C. fulleri* genome (ii) a *C. maculifer* genome (another species of the same lineage) and (iii) a *C. maculifer* transcriptome (Figure 2a).

3.1 | Use of the Corydoras-specific TE library led to a 2–3-fold increase in TE abundance estimates

Total TE abundance estimates were higher across both species and sequence types when using the *Corydoras*-specific library. For *C.*

fulleri, estimated TE abundance more than doubled from 18.54% of the genome (755.96 hits per MB) using the *D. rerio* RepBase library to 43.45% of the genome (1499.91 hits per MB) using the *Corydoras*-specific library (Figure 2b). For the closely related species *C. maculifer*, estimated TE abundance almost tripled from 14.17% of the genome (626.87 hits per MB) using the *D. rerio* RepBase library to 40.23% of the genome (2218.25 hits per MB) using the *Corydoras*-specific library (Figure 2b). We then assessed the estimated abundance of TEs across the transcriptome of *C. maculifer*, where TE derived transcripts are expected to represent younger, potentially active, transposons (Lanciano & Cristofari, 2020). Transcriptional TE content was substantially lower than in the *C. maculifer* genome, varying between 1.17% (68.07 hits per MB) and 4.68% (263.22 hits per MB) of the transcriptome when using the *D. rerio* and *Corydoras*-specific library respectively (Figure 2b). The substantial increases associated with the use of the *Corydoras*-specific library suggests that the *D. rerio* library missed a large fraction of *Corydoras*-specific elements. We therefore investigated the total number of different TE entries within the *Corydoras*-specific and *D. rerio* RepBase libraries detected in the *C. fulleri* genome. The *Corydoras*-specific library led to an average fourfold increase in the number of different TEs detected (Figure S1). Furthermore, across all classes (and particularly for DDE DNA and LTR classes), a number of elements present in the *D. rerio* library were not detected at all within the *C. fulleri* genome (Figure S1).

3.2 | Use of the Corydoras-specific TE library led to substantial changes in estimated TE composition

Using the *Corydoras*-specific library impacted TE composition estimates across both species and sequence types, which we assessed using DeepTE assigned classification. Similar to other teleosts, DDE DNA elements (particularly Tc1 Mariner and hAT transposons) made up substantial proportions of both species genomes and transcriptomes (Figure 3). Estimated genomic TE compositions were similar across both genomes investigated, indicating a high level of intralineage TE similarity. TE annotation using the *D. rerio* library detected a similar, relatively high, proportion of SINEs within both genomes, which is in contrast to other teleost species which typically have SINE-depleted genomes (Gao et al., 2016; Shao et al., 2019). On closer inspection however, absolute SINE abundance (29 MB, 4.57% genome) was similar to that reported in the *D. rerio* genome (30.64 MB, 2.24% of genome) (Gao et al., 2016), suggesting that SINE over-representation was a consequence of (i) non-SINE elements being missed when using the *D. rerio* library and (ii) SINEs being undetected during de novo library construction and therefore poorly represented in subsequent analyses that depend on the library. Supporting this we found that the number of SINEs detected using the *D. rerio* library was largely driven by a single element (HE1 DR1, 84.52% of SINEs in *C. maculifer* and 84.12% of SINEs in *C. fulleri*) which, following confirmation using BLASTn, was absent in the

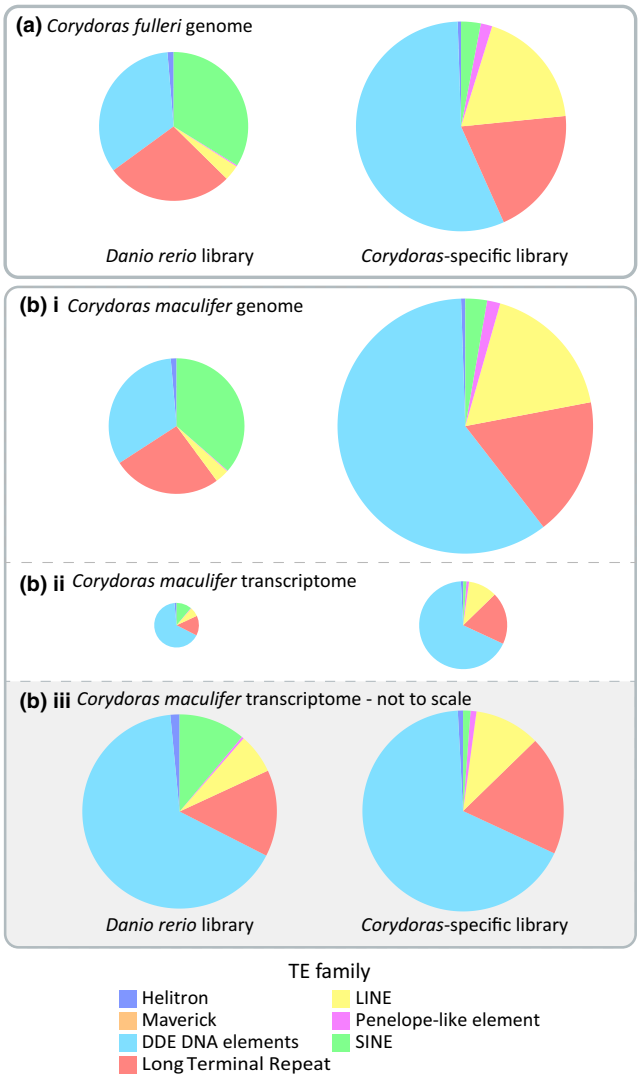

FIGURE 3 TE library type alters TE composition. Estimated TE composition are given in (a) the *C. fulleri* genome, (bi) the *Corydoras maculifer* genome and (bii) the *C. maculifer* transcriptome after using the *Danio rerio* (left) and *Corydoras*-specific (right) TE libraries. Pie charts are scaled based on TE abundance per MB in all cases apart from (biii) which, for clarity, is the unscaled *C. maculifer* transcriptome composition

de novo *Corydoras*-specific library. We also note that the choice of TE library did not generate large compositional changes within transcriptomic sequences (Figure 3(b) iii). To investigate whether any compositional bias had been introduced by DeepTE, we also ran the curated RepBase *D. rerio* TE library through DeepTE and compared its classification outputs against the original RepBase library. As expected, the RepBase curated library had a greater range of classifications than the DeepTE classified library (Figure S2). Although general classification patterns were similar, there was some bias exhibited by DeepTE towards both TIR elements (hAT and Mariner-like) and LTR elements (BEL and Copia) (Figure S2).

3.3 | The Corydoras-specific library was more fragmented when compared to a curated TE library

We assessed the degree of fragmentation in the *Corydoras*-specific library against the curated *D. rerio* library using (i) cumulative frequency of estimated individual TE abundances and (ii) TE length distributions across the *C. fulleri* genome (Figure 4). We define fragmentation as genomic TE copies which have not been captured as a single contiguous unit during library creation (Figure 5). An excess of fragmented TE library entries will push a cumulative frequency curve further to the right because many entries will be found at low abundance within the genome (singletons) (Figure 5). When standardised by total number of hits, we found little difference between the two libraries (Figure 4a) although the *Corydoras*-specific library was inflated with singletons (6.25% of library entries). When looking at the TE length distributions and benchmarking the *Corydoras*-specific library against the RepBase *D. rerio* library (run on their respective genomes) we see markedly similar patterns across all TE classes, with one anomalous peak at c. 350 bp in the LINE distributions (Figure 4b). On closer investigation, this peak consisted of a single element (TE_00002410) which, following reanalysis with BLASTx, closely matched a LTR copia element, so is probably a product of misidentification or misclassification by EDTA or DeepTE. We also calculated the average proportion of hits that map back to a single element, with lower values indicating higher degrees of fragmentation. For the *C. fulleri* genome the median number of hits that map to a single element was 0.003% (36 hits per element) and within *D. rerio* it was 0.009% (218 hits per element).

3.4 | The Corydoras-specific TE library reduces average TE age estimates

We investigated the impact the *Corydoras*-specific library had on estimated TE age distributions compared to using the *D. rerio* RepBase library. The *Corydoras*-specific library reduced average sequence divergence against corresponding library entries across all major TE classes and sequence types, suggesting a recent TE accumulation within the *Corydoras* which would have been missed if relying solely on the *D. rerio* library (Figure 6). Specifically, the use of the *Corydoras*-specific library significantly reduces the divergence estimates of each element by an average of ~4% (*D. rerio* library 19.90 ± 5.38 sd; *Corydoras*-specific library 15.60 ± 6.71 sd; Welch's $t = -519.93$, d.f = 1,173,000 $p < .001$). Finally, the proportion of elements that were very young (estimated to be <5% divergent from its corresponding library entry) was 7.52% within transcriptomic sequences and 5.54% within genomic sequences, suggesting that expressed TEs are on average younger than their genomic counterparts.

4 | DISCUSSION

Our results demonstrate how TE library choice can have major implications during TE detection and quantification. The use of the *Corydoras*-specific library led to a 2–3-fold increase in estimated TE abundance in *Corydoras* spp., meaning that ~40% of the two *Corydoras* genomes investigated consist of TEs. TE abundance is highly variable amongst teleosts, ranging between 5% in pufferfish (*Tetraodon nigroviridis*) to 56% in zebrafish (*D. rerio)* (Shao et al.,

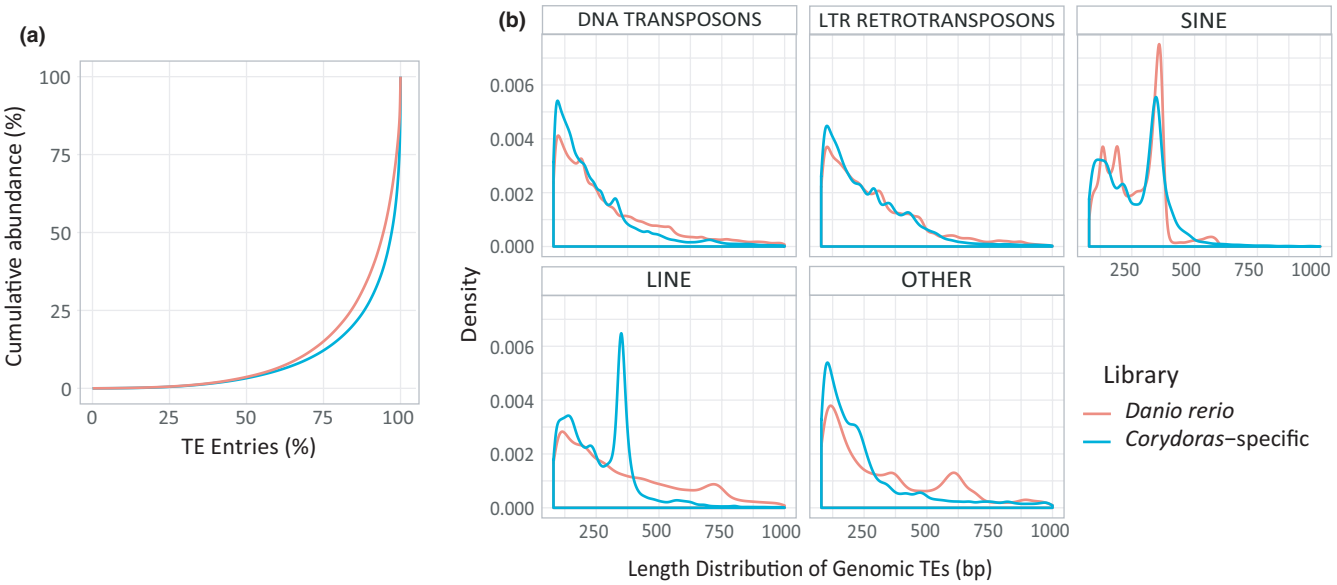

FIGURE 4 Degree of fragmentation is similar for both Corydoras-specific and *Danio rerio* TE libraries. (a) Cumulative frequency of standardised (%) estimated TE abundance using the Corydoras-specific library and the *D. rerio* library. (b) TE length distributions across the Corydoras-specific library and *D. rerio* library, when run on their respective genomes. For visual purposes the length distribution had a cut off of 1000 bp

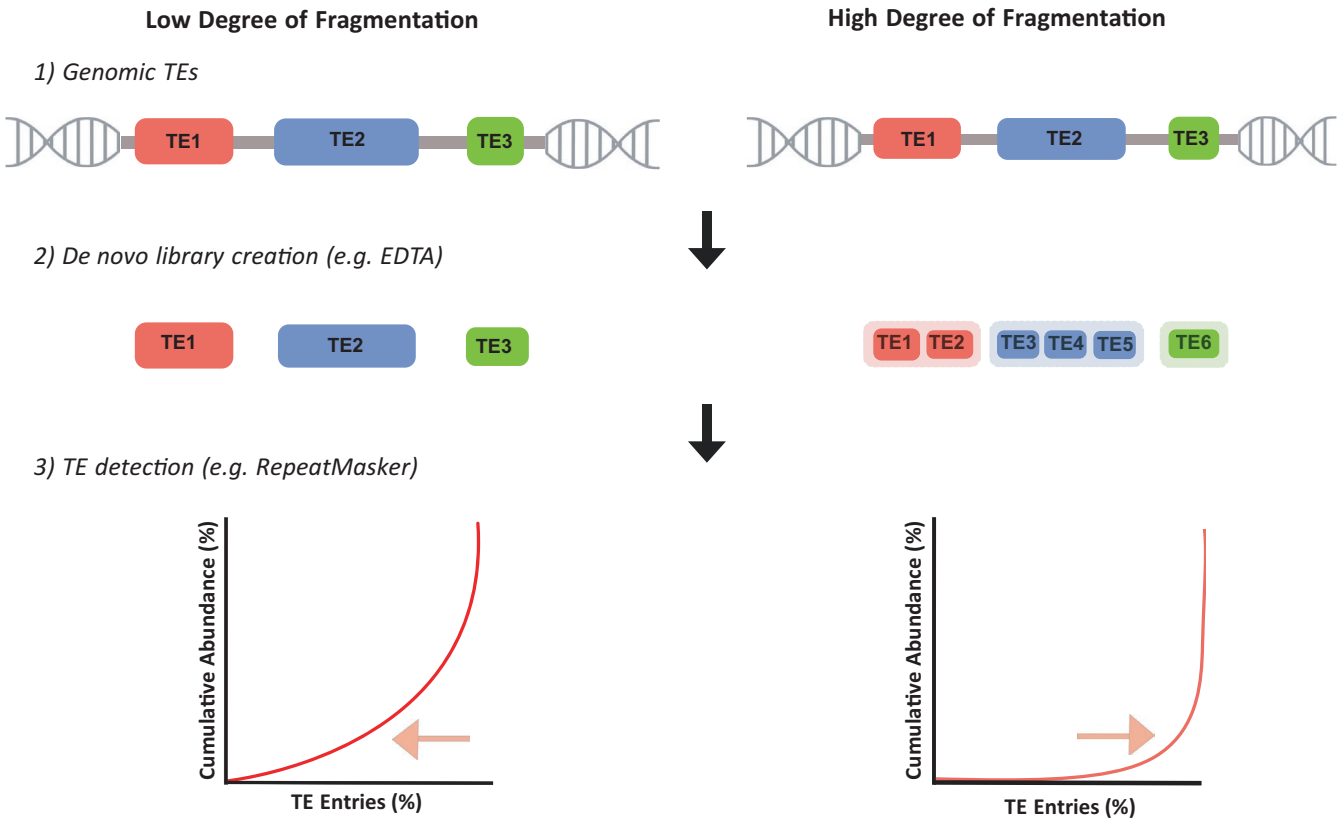

FIGURE 5 Schematic depicting TE fragmentation as a result of de novo library creation. Fragmentation during de novo library creation occurs when single TE copies are detected as multiple fragmented copies. This creates an overinflation of unique library entries, and results in a skewed cumulative frequency curve due to an excess of singletons (TEs detected once only)

2019). Use of the *Corydoras*-specific library indicates that TE abundance within these two *Corydoras* genomes (both spp. lineage 1) is comparable to other teleosts, particularly *D. rerio* (~56% of the genome) and *Oryzias latipes* (~33.7% of the genome) (Gao et al., 2016). Theoretically, an inverse relationship between homology-based identification rates and phylogenetic distance exists, in which sequence differences between the species used to develop a library and the target species may provide an obstacle for accurate TE detection. A previous comparison across 40 mammalian genomes demonstrated that TE detection rates exhibit a 'threshold limit', in which TE abundance underestimates are largely avoided until a phylogenetic distance greater than ~90MY is reached, above which homology-based searching may detect as few as 20% of total TEs (Platt et al., 2016). It is therefore no surprise that *Corydoras* TE content was probably underestimated when assessed using the *D. rerio* library given that these species are separated by ~150 million years of evolution (Chen et al., 2013).

In addition, estimated transcriptomic TE abundance was approximately an order of magnitude lower than genomic content, which probably reflects the fact that: (i) TEs may largely be located within non-coding regions of the genome, (ii) many TEs found within *Corydoras* genomes may be degraded and no longer possess the ability to be transposed, or (iii) epigenetic silencing mechanisms (such as CpG methylation and histone modifications) may prevent TE expression (Slotkin & Martienssen, 2007). It is also worth noting that this study used RNA-seq data originating from somatic muscle tissue. TE expression is likely to vary between different tissue types, theoretically evolving to be most active in the germline and comparatively silent in the soma (Haig, 2016).

TE composition estimates within the *Corydoras* were found to be similar to that of other teleosts, with DDE DNA transposons being the most abundant TE class, largely driven by a high abundance of Tc1-Mariner and hAT elements (Shao et al., 2019). Due to their "blurry promoters" Mariner elements appear to have a particular propensity for horizontal transfer across the vertebrate kingdom (Zhang et al., 2020). Despite the suggestion that homology-based methods may miss horizontally transferred elements, a BLASTn search against genomic *C. maculifer* Mariner elements (see methods for full details) demonstrated that the percentage that have a best hit against a non-teleost species differed very little between library type (5.58% for the *Corydoras*-specific and 4.22% for the *D. rerio* library). Interestingly, it appears that the percentage of expressed Mariner elements with a best hit against a non-teleost species within the *C. maculifer* transcriptome was much higher than within the genome (17.14%), suggesting potential horizontally transferred elements may be more likely to be under purifying selection and retain their transposition ability (Zhang et al., 2020). The evolutionary impacts of horizontally transferred TEs are potentially wide-reaching

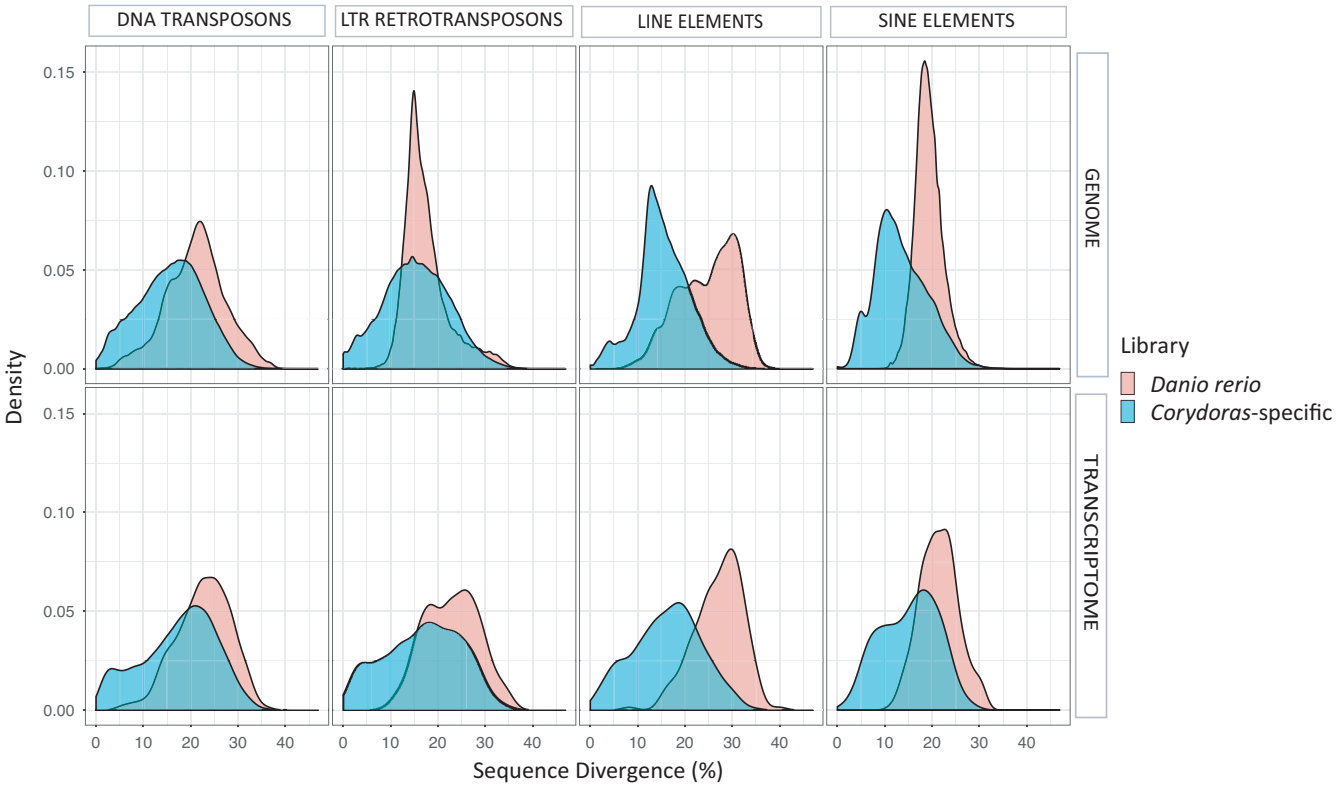

FIGURE 6 TE library type alters TE age distributions. Density plots were used to highlight sequence divergence distributions identified using the *Corydoras*-specific library and the *Danio rerio* library. Plots are faceted by the four main TE classes (DNA transposons, LINE elements, LTR Retrotransposons and SINE elements) and sequence type (genome vs transcriptome). All plots are based on *C. maculifer* sequence data

(see Schaack et al., 2010), and thus their accurate annotation is important. More conservative testing across a wider range of elements would be required to fully investigate the role that library type has on the detection of horizontally transferred TEs, and is an important avenue to explore in the future.

The use of the *D. rerio* TE library led to skewed estimates of TE compositions. In particular, homology-based searching inflated the relative proportion of genomic SINE elements to a level equivalent to DDE DNA transposons, which was unexpected given other teleost species contain particularly SINE-depleted genomes (Gao et al., 2016; Shao et al., 2019) and potentially caused by a predisposition for homology-based searching against the detection of certain TE classes (e.g., DNA transposons). Furthermore, the majority of SINEs found by homology searching were represented by a single SINE element, which was not present in the *Corydoras*-specific library, suggesting a failure to comprehensively detect SINE elements during de novo library creation. This finding was, in part, expected: SINE elements have a propensity to be missed during de novo library creation because of high sequence variation and lack of terminal repeats and supports a prediction made by Ou et al. (2019). Such compositional differences were not observed within transcriptomic sequences, possibly because expressed TEs (which are often younger) tend to have higher levels of sequence similarity (Lanciano & Cristofari, 2020). Over-reliance on homology-based searching may lead to similar inaccuracies during TE abundance and compositional estimates, particularly when working with organisms that are phylogenetically distant from a model organism in which a curated TE library exists.

The substantial increase in individual elements detected in the *Corydoras*-specific library compared to the *D. rerio* library raises the possibility of false discovery and/or fragmentation, whereby libraries contain multiple fragmented entries representing different regions of a single contiguous element (Flynn et al., 2020). Both false discovery and library fragmentation are common pitfalls associated with de novo pipelines (Flynn et al., 2020; Ou et al., 2019,). Without full manual curation, false discovery rate is difficult to assess; however, performance analysis of EDTA within the model rice *Oryza sativa* indicated that EDTA exhibits an overall false-positive rate of ~15% which, even in the unlikely absence of false-negatives, would not explain the degree of estimated TE abundance increase we observed in the *Corydoras* (Ou et al., 2019). We assessed fragmentation by (i) measuring the cumulative percentage abundance of TEs that mapped back to a single element, and (ii) plotting distributions of masked genomic TE lengths across each TE class. The cumulative curve created from the *Corydoras*-specific library suggested that it is likely to be more fragmented than the manually curated RefBase *D. rerio* library. However, when looking at the number of TEs that mapped back to a single element, the differences between the *Corydoras*-specific and *D. rerio* libraries were within the same order of magnitude, and at a similar level to the variation observed

when comparing fragmentation levels between de novo generated libraries produced by different pipelines (Flynn et al., 2020). When comparing distributions of TE length between the *Corydoras*-specific and *D. rerio* libraries, the results were remarkably similar with a single anomalous peak within the LINE class, which on further investigation proved to be evidence of misidentification/misclassification by EDTA and DeepTE. We have provided a framework for comparative analysis against a model species library which may be an efficient way to check de novo libraries for deviations, either in false discovery, fragmentation or missing elements.

When assessing estimated TE age distributions associated with both library types, we found that the *Corydoras*-specific library led to a reduction in average sequence divergence across multiple TE classes, suggestive of a more recent TE accumulation within the *Corydoras*. This mirrors findings in mammals and insects where the use of species-specific TE libraries also reduced relative TE ages (Platt et al., 2016). Analysis of other teleost genomes suggest that TE age distributions can vary considerably among species. Three waves of TE accumulation have been proposed in the evolutionary history of cichlid fish for example, whereas a single, recent insertion peak was identified in the piranha genome (*Pygocentrus nattereri)* (Brawand et al., 2014; Schartl et al., 2019). Furthermore, the variation in estimated sequence divergence associated with the *Corydoras*-specific library was larger than the *D. rerio* library, suggesting that the *Corydoras*-specific library was able to detect TEs of a greater age range. Taken together these findings further support the notion that the majority of TEs detected using the *D. rerio* library are probably those inherited from the common ancestor of the *Corydoras* and *D. rerio*. Additionally, we found that elements found within transcriptomic data had a greater probability of being less divergent than their genomic counterparts, further supporting the hypothesis that expressed TEs tend to be both younger and more intact, with a greater potential for active transposition.

5 | CONCLUSION AND OUTLOOK

To conclude, we have combined two recent bioinformatic pipelines (EDTA & DeepTE) to generate a novel semi-automated de novo TE library for a non-model group of teleosts (*Corydoras*). We assessed the performance of this *Corydoras*-specific library against a distantly related but highly curated TE library (*D. rerio*). Across both species and sequence types, the use of the *Corydoras*-specific TE library increased estimated transposon abundance between 2–3x and altered TE composition estimates. We stress that future work on non-model organisms will probably encounter substantial TE underestimates/classification biases if researchers are to rely heavily on homology-based TE detection. Furthermore, we demonstrate that TEs missed by homology methods are likely to be species-specific, and thus elements of most interest if the focal aim of a study is assessing lineage specific TE impacts. Furthermore, use of the de novo library reduced the estimated average sequence divergence/age distributions. This is likely to have important implications for researchers particularly interested in identifying elements that have recently proliferated within a lineage. Many of these TE-based traits varied across sequence type, with expressed TEs being estimated at lower abundance and exhibiting a younger average age. Finally, we provided an assessment of potential fragmentation associated with EDTA generated TE libraries, and whilst the *Corydoras*-specific library exhibited some fragmentation, it is within the same order of magnitude as the manually curated *D. rerio* library. A set of relatively small alterations highlighted by our fragmentation analysis could improve the *Corydoras*-specific library further. By providing both (i) a quantitative assessment of how library choice can influence numerous important TE-based metrics, and (ii) a stepwise pipeline (https://github.com/ellenbell/FasTE) for replication, we hope this study can provide a useful resource for all TE-based researchers, and particularly those who may be new to the field.

ACKNOWLEDGEMENTS

We would like to thank Dr Alexander Suh and Professor Tracey Chapman for their valuable insights on the analysis and compilation of this manuscript and to Dr Riviane Garcez da Silva for her help in RNA extraction. The genomic assemblies presented in this paper were produced on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

CONFLICT OF INTEREST

None.

AUTHOR CONTRIBUTIONS

The study was conceived and developed by Ellen A. Bell, Christopher L. Butler and Martin I. Taylor. Sample collection was completed by Martin I. Taylor and Claudio Oliveira. Genomic data for *Corydoras fulleri* was extracted, assembled and analysed by Ellen A. Bell. Genomic data for *Corydoras maculifer* was extracted and analysed by Ellen A. Bell and assembled by Sarah Marburger. Transcriptomic sequence data was extracted and sequenced by Claudio Oliveira, assembled by Ellen A. Bell and Christopher L. Butler and further analysed by Christopher L. Butler. The de novo repeat library was constructed by Ellen A. Bell. Repeat-Masker output files were parsed using a custom script written and developed by Christopher L. Butler. This manuscript was written in equal parts by Ellen A. Bell and Christopher L. Butler with contributions from Martin I. Taylor, Claudio Oliveira, Sarah Marburger and Levi Yant.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareabledata necessary to reproduce the reported results. The data is available at https://doi.org/10.5061/dryad.m0cfxpp3h, https://github.com/ellenbell/, https://github.com/clbutler/RM_TRIPS and https://doi.org/10.5061/dryad.m0cfxpp3h.

DATA AVAILABILITY STATEMENT

Genomic and transcriptomic assemblies have been made publicly available from Genbank. The *Corydoras fulleri* genome assembly has been deposited under the BioProject Assession No: PRJNA706371 and the *Corydoras maculifer* genome assembly has been deposited under the BioProject Accession No: PRJNA706379. The Transcriptome Shotgun Assembly (TSA) for *Corydoras maculifer* has been deposited at under the Accession no: GJAY00000000. The associated Sequence Read Archive (SRA) data has been deposited under the BioProject Accession No: PRJNA706424.

Repeat libraries in fasta file format and unparsed Repeat Masker output files have been submitted to Dryad (https://doi.org/10.5061/dryad.m0cfxpp3h). The full pipeline for de novo library generation and utilisation is available on GitHub: https://github.com/ellenbell/FasTE Scripts for the analyses associated with this manuscript have either been published on GitHub: https://github.com/clbutler/RM_TRIPS or are available from Dryad (https://doi.org/10.5061/dryad.m0cfxpp3h).

ORCID

Ellen A. Bell https://orcid.org/0000-0001-7427-6331
Christopher L. Butler https://orcid.org/0000-0001-6337-6652
Claudio Oliveira https://orcid.org/0000-0002-7010-8880
Levi Yant https://orcid.org/0000-0003-3442-0217
Martin I. Taylor https://orcid.org/0000-0002-3858-0712

REFERENCES

Alexandrou, M. A., Oliveira, C., Maillard, M., McGill, R. A. R., Newton, J., Creer, S., & Taylor, M. I. (2011). Competition and phylogeny determine community structure in Müllerian co-mimics. *Nature*, *469*, 84–89. https://doi.org/10.1038/nature09660

Bergman, C. M., & Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, *8*(6), 382–392. https://doi.org/10.1093/bib/bbm048

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1577-z

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A. Y., Lim, Z. W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., … Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*(7518), 375–381. https://doi.org/10.1038/nature13726

Cai, X. U., Chang, L., Zhang, T., Chen, H., Zhang, L., Lin, R., Liang, J., Wu, J., Freeling, M., & Wang, X. (2021). Impacts of allopolyploidization and structural variation on intraspecific diversification in *Brassica rapa*. *Genome Biology*, *22*(166), 1–24. https://doi.org/10.1186/s13059-021-02383-2

Chen, W. J., Lavoué, S., & Mayden, R. L. (2013). Evolutionary origin and early biogeography of otophysan fishes (ostariophysi: teleostei). *Evolution*, *67*(8), 2218–2239. https://doi.org/10.1111/evo.12104

Clavijo, B. J., Garcia Accinelli, G., Wright, J., Heavens, D., Barr, K., Yanes, L., & Di-Palma, F. (2017). W2RAP: A pipeline for high quality, robust assemblies of large complex genomes from short read data, *bioRxiv*, 1–12. doi: https://doi.org/10.1101/110999

Cowley, M., & Oakey, R. J. (2013). Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLoS Genetics*, *9*(1), 1–7. https://doi.org/10.1371/journal.pgen.1003234

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. https://doi.org/10.1073/pnas.1921046117

Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., & Song, C. (2016). The contribution of transposable elements to size variations between four teleost genomes. *Mobile DNA*, *7*(4), 1–16. https://doi.org/10.1186/s13100-016-0059-7

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., & Regev, A. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Groth, S. B., & Blumenstiel, J. P. (2017). Horizontal transfer can drive a greater transposable element load in large populations. *Journal of Heredity*, *108*(1), 36–44. https://doi.org/10.1093/jhered/esw050

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haig, D. (2016). Transposable element: Self-seekers of the germline, team-players of the soma. *BioEssays*, *38*, 1158–1166. https://doi.org/10.1002/bies.201600125

Hoen, D. R., & Bureau, T. E. (2015). Discovery of novel genes derived from transposable elements using integrative genomic analysis. *Molecular Biology and Evolution*, *32*(6), 1487–1506. https://doi.org/10.1093/molbev/msv042

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. https://doi.org/10.1038/nature12111

Hu, Y., Colantonio, V., Müller, B. S. F., Leach, K. A., Nanni, A., Finegan, C., Wang, B. O., Baseggio, M., Newton, C. J., Juhl, E. M., Hislop, L., Gonzalez, J. M., Rios, E. F., Hannah, L. C., Swarts, K., Gore, M. A., Hennen-Bierwagen, T. A., Myers, A. M., Settles, A. M., Tracy, W. F., & Resende, M. F. R. (2021). Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. *Nature Communications*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-21380-4

Kennedy, R. C., Unger, M. F., Christley, S., Collins, F. H., & Madey, G. R. (2011). An automated homology-based approach for identifying transposable elements. *BMC Bioinformatics*, *12*, 1–10. https://doi.org/10.1186/1471-2105-12-130

Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nature Reviews Genetics*, *21*(12), 721–736. https://doi.org/10.1038/s41576-020-0251-y

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., & Caccamo, M. (2014). Next clip: An analysis and read preparation tool for nextera long mate pair libraries. *Bioinformatics*, *30*(4), 566–568. https://doi.org/10.1093/bioinformatics/btt702

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., & Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(18), 1–18. https://doi.org/10.1186/s13742-015-0069-2

Marburger, S., Alexandrou, M. A., Taggart, J. B., Creer, S., Carvalho, G., Oliveira, C., & Taylor, M. I. (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. *Proceedings of the Royal Society B*, *285*, 1–10. https://doi.org/10.1098/rspb.2017.2732

Michell, C., Wutke, S., Aranda, M., & Nyman, T. (2021). Genomes of the willow-galling sawflies *Euura lappo* and *Eupontania aestiva* (Hymenoptera: Tenthredinidae): a resource for the research on ecological speciation, adaptation and gall induction. *G3*, *11*(5), 1–7. https://doi.org/10.1093/g3journal/jkab094

Ou, S., Su, W., Liao, Y. I., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, *20*(1), 1–18. https://doi.org/10.1186/s13059-019-1905-y

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. https://doi.org/10.1093/gbe/evw009

Rogers, R. L., Zhou, L., Chu, C., Márquez, R., Corl, A., Linderoth, T., Freeborn, L., MacManes, M. D., Xiong, Z., Zheng, J., Guo, C., Xun, X. U., Kronforst, M. R., Summers, K., Wu, Y., Yang, H., Richards-Zawacki, C. L., Zhang, G., & Nielsen, R. (2018). Genomic takeover by transposable elements in the strawberry poison frog. *Molecular Biology and Evolution*, *35*(12), 2913–2927. https://doi.org/10.1093/molbev/msy185

Ruan, J., & Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *bioRxiv*. doi: https://doi.org/10.1101/530972

Schaack, S., Gilbert, C., & Feschotte, F. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology and Evolution*, *25*(9), 537–546. https://doi.org/10.1016/j.tree.2010.06.001

Schartl, M., Kneitz, S., Volkoff, H., Adolfi, M., Schmidt, C., Fischer, P., Minx, P., Tomlinson, C., Meyer, A., & Warren, W. C. (2019). The piranha genome provides molecular insight associated to its unique feeding behavior. *Genome Biology and Evolution*, *11*(8), 2099–2106. https://doi.org/10.1093/gbe/evz139

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In Kollmar, M. (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227–245). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_14

Shao, F., Han, M., & Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-51888-1

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*(4), 272–285. https://doi.org/10.1038/nrg2072

Smit, A. F., & Hubley, R. RepeatModeler Open-1.0. Available from: http://www.repeatmasker.org

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*(8), 1134–1144. https://doi.org/10.1101/gr.196469.115

Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biology and Evolution*, *9*(1), 161–177. https://doi.org/10.1093/gbe/evw264

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, *9*(11), 1–14. https://doi.org/10.1371/journal.pone.0112963

Wells, J. N., & Feschotte, C. (2020). A field guide to transposable elements. *Annual Review of Genetics*, *54*, 1–23. https://doi.org/10.1002/9781118500156.ch2

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982. https://doi.org/10.1038/nrg2165

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Yan, H., Bombarely, A., & Li, S. (2020). DeepTE: A computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics*, *36*(15), 4269–4275. https://doi.org/10.1093/bioinformatics/btaa519

Zhang, H. H., Peccoud, J., Xu, M. R. X., Zhang, X. G., & Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-15149-4

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Bell, E. A., Butler, C. L., Oliveira, C., Marburger, S., Yant, L., & Taylor, M. I. (2021). Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Molecular Ecology Resources*, 00, 1–11. https://doi.org/10.1111/1755-0998.13489