Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline

Tobias Baril[1,2]*, Ryan M. Imrie[1,3], Alex Hayward[1,4]

[1] University of Exeter, Penryn Campus, Cornwall, TR10 9FE, UK

[2] ORCID: 0000-0002-5936-7531

[3] ORCID: 0000-0002-5055-2888

[4] ORCID: 0000-0001-7413-718X

* Corresponding Author information:

Tobias Baril: tb529@exeter.ac.uk

ABSTRACT

Background

Transposable elements (TEs) are found in nearly all eukaryotic genomes and are implicated in a range of evolutionary processes. Despite considerable research attention on TEs, their annotation and characterisation remain challenging, particularly for non-specialists. Current methods of automated TE annotation are subject to several issues that can reduce their overall quality: (i) fragmented and overlapping TE annotations may lead to erroneous estimates of TE count and coverage; (ii) repeat models may represent small proportions of their total length, where 5’ and 3’ regions are poorly captured; (iii) resultant libraries may

contain redundancy, with the same TE family represented more than once. Existing pipelines can also be challenging to install, run, and extract data from. To address these issues, we present Earl Grey: a fully automated transposable element annotation pipeline designed for the user-friendly curation and annotation of TEs in eukaryotic genome assemblies.

Results

Using a simulated genome, three model genome assemblies, and three non-model genome assemblies, Earl Grey outperforms current widely used TE annotation methodologies in ameliorating the issues mentioned above by producing longer TE consensus sequences in non-redundant TE libraries, which are then used to produce less fragmented TE annotations without the presence of overlaps. Earl Grey scores highly in benchmarking for TE annotation (MCC: 0.99) and classification (97% correctly classified) in comparison to existing software.

Conclusions

Earl Grey provides a comprehensive and fully automated TE annotation toolkit that provides researchers with paper-ready summary figures and outputs in standard formats compatible with other bioinformatics tools. Earl Grey has a modular format, with great scope for the inclusion of additional modules focussed on further quality control aspects and tailored analyses in future releases.

KEYWORDS

Transposon; Repeat; Mobile genetic element; Genome; Software; Consensus.

BACKGROUND

Over recent decades, massive advances in genome sequencing technologies have accompanied significant decreases in sequencing costs. This has led to a huge increase in

the availability of genome assemblies for lineages across the eukaryotic tree of life. Further, the advent of long-read sequencing technologies has resulted in a wealth of high quality and highly contiguous genome assemblies, and chromosome-level resources are now becoming common for both model and non-model organisms. With the accumulation of genomic resources, it has become clear that large proportions of eukaryotic genomes are composed of transposable elements (TEs) (Bourque *et al.*, 2018). TEs are DNA sequences that are capable of moving from one location to another in a host genome (McClintock, 1956). TEs are almost ubiquitous across the eukaryotic tree of life and have been acknowledged as a potential reservoir of genetic variation (Wells and Feschotte, 2020). As such, TEs are implicated in a range of evolutionary processes through the generation of genomic novelty via processes including alteration of coding sequence, chromosomal rearrangements and modification of gene regulatory networks (Chung *et al.*, 2007; Hof *et al.*, 2016; Chuong, Elde and Feschotte, 2017; Cosby, Chang and Feschotte, 2019).

Tools for annotating the location and identity of TEs within the genome are an integral aspect of most genomics projects, even where there is little ultimate research focus on TE biology. A key example is the assembly of a new genome, where TE annotation is performed so that TEs do not interfere with gene prediction pipelines (Campbell *et al.*, 2014; Kollmar, 2019). This involves screening for TEs present in the genome and masking them so they are filtered out during subsequent steps, typically by using annotations that replace their nucleotide sequences with either lowercase letters (softmasking), or another symbol such as ‘N’ or ‘X’ (hardmasking). Given that TE annotation is typically carried out prior to gene prediction, the quality of TE annotation performed can influence the resultant quality of host gene annotation. For example, many TEs contain protein-coding regions that could be incorrectly annotated as host genes (Bourque *et al.*, 2018), while repetitive host gene families can be mistaken for TEs.

TEs are inherently repetitive in nature, and TE annotation approaches firstly seek to identify ‘families’ of elements composed of closely related TE copies considered to originate from a single initial integration in the genome of interest. TE families are typically defined by

thresholds, such that two TE sequences are considered to belong to the same family if they align with 80% identity, across at least 80bp and 80% of their length (Wicker *et al.*, 2007). TE families are commonly represented by a single consensus sequence, constructed by aligning multiple copies sampled from the genome that fall within the defined sequence threshold. Collectively, the set of resultant TE models for a given species is referred to as a 'TE library'. TE families can be identified using 'library-based' approaches, which search the genome for similarity to elements already present in TE databases. Alternatively, '*de novo*' analyses use either existing knowledge of TE structure or their repetitive nature to detect TEs within a genome.

The gold standard for TE annotation is manual curation, whereby a *de novo* TE prediction program is run on a genome assembly to identify putative TE sequences before individual TE copies are extracted and a multiple alignment is generated for each TE 'family' (Goubert *et al.*, 2022). These alignments are then curated manually, which involves trimming poorly aligned regions, and identifying hallmarks of TE boundaries, such as target site duplications and long-terminal repeats. The strength of manual curation lies in the individual care taken to annotate each TE model. However, the inherent difficulties associated with manual curation methods are not easily surmountable. Firstly, manual curation requires a significant time investment and expert knowledge of the structure of diverse TE types. The nature of manual curation also leads to reproducibility issues, due to variability introduced when trimming sequence alignments by eye, especially when considering differences in experience, the ability to spot patterns, and human error. Consequently, manual curation is most effective in studies focusing on single species and is generally unviable for large-scale comparative studies due to the levels of human resources required, and the potential influence of curator-based variability.

To overcome the limitations of manual TE curation and facilitate large-scale comparative and population-level studies, a diversity of automated TE annotation tools have been developed (Goerner-Potvin and Bourque, 2018). However, many of these tools are species- or element-specific and require a level of technological expertise to install and

configure, often requiring multiple dependencies and databases, some of which are difficult to locate and apply due to dead weblinks, a lack of updates, and retirement of codebases/dependencies. Compounding these problems is a lack of focus on user-friendliness in the usage processes, which frequently includes issues relating to data compatibility with downstream tools, necessitating rounds of data reformatting or cleaning. Overall, this can act as a significant barrier to discourage interested biologists from performing research on TEs.

Automated TE annotation remains a computationally challenging process. This is due to the great diversity and complexity of TE sequences, the huge number of TEs that remain uncharacterised, and the gradual post-integration erosion of TE sequences due to host mutational processes, which decreases the ability to recognise and identify them. Consequently, automated TE annotation tools often struggle to accurately identify TE sequences. Methodologies that generate *de novo* TE consensus sequences often produce TE consensus libraries with high levels of redundancy, where a single TE is represented by multiple TE consensus models, and these models often perform poorly in their designation of TE boundaries (Rodriguez and Makałowski, 2022). Following consensus generation, TE annotation can also lead to challenges that affect both *de novo* and library-based methods. Fragmented TE annotations are often inferred, where a single TE is represented by multiple separate annotations instead of a single continuous annotation. In addition, TE annotations often contain overlaps, where a single base pair is annotated with multiple TE identities. Such issues can result in distorted TE counts and coverage estimates, with knock-on implications for downstream analyses. Many of these issues are particularly acute when considering non-model organisms that are distantly related to model reference species, where considerable effort has been expended to provide high quality TE curation.

The release of new genome assemblies has now outpaced efforts to characterise their associated TEs. Furthermore, this situation will accelerate with the initiation of massive-scale genome sequencing efforts such as the Darwin Tree of Life Project (Blaxter *et al.*, 2022) and the Earth Biogenome Project (Lewin et al. 2018), which seek to provide high

quality chromosomal-level genome assemblies for all eukaryotic life in the British Isles and planet Earth respectively. Such projects offer huge opportunities to expand our understanding of TE biology. However, the scale of genomic resources available also greatly increases the need for rapid, robust, and user-friendly automated TE annotation and analysis approaches.

Here, we present Earl Grey, a fully automated TE annotation pipeline combining widely used library-based and *de novo* TE annotation tools, with TE consensus and annotation refinements, aimed at generating high-quality TE libraries, annotations, and analyses for eukaryotic genome assemblies.

IMPLEMENTATION

Earl Grey is available from a Github repository (https://github.com/TobyBaril/EarlGrey) and can be run on Linux distributions, such as Ubuntu. It can be installed on a local system or a HPC. Earl Grey is parallelised and makes use of multiple CPU threads to reduce runtime. Given that storage input/output speed can be a limiting factor when annotating TEs, analysing individual genome assemblies back-to-back (instead of running multiple genome analyses at once) can be most time efficient.

Earl Grey runs in Conda or Docker environments to avoid conflicts between tool versions and to streamline the installation procedure. Users who do not have RepeatMasker (Smit, Hubley and Green, 2013) and RepeatModeler2 (Flynn *et al.*, 2020) installed, and who do not wish to do this, can install Earl Grey in an interactive Docker container, which will install and configure all dependencies automatically and provide a virtual machine in which to perform all Earl Grey analyses. For a Conda installation (which uses the Anaconda Distribution Platform), it is necessary to have RepeatMasker and RepeatModeler2 pre-installed. This is because the Conda implementations of these packages are difficult to configure with the expanded Dfam and RepBase libraries, and our aim is to minimise the difficulty of configuration for the user. Instructions for the installation of these dependencies

are provided for users who do not currently have them installed. As a minimum requirement, RepeatMasker must be configured to use the Dfam database of repetitive DNA elements (Hubley *et al.*, 2016) (tested with all versions from release 3.2 onwards). If users have access to RepBase RepeatMasker edition libraries (Jurka *et al.*, 2005; Kapitonov and Jurka, 2008), they are encouraged to configure RepeatMasker with these in addition to Dfam. A configuration script provided with Earl Grey will automatically check for dependencies and configure the Conda environment and related tools to enable the user to run Earl Grey.

Once installed and configured, Earl Grey will run on a given input genome assembly in FASTA format with a single command. Prior to running Earl Grey, the "earlGrey" conda environment must be activated '`conda activate earlGrey`'. Once the conda environment is active, Earl Grey can be called with the command '`earlGrey`'. There are four required options, and 5 optional parameters (Table 1). For example, a run on the *Homo sapiens* genome, with the genome located in the current directory, could be started with the minimum parameters: '`earlGrey -g homoSapiens.fasta -s homoSapiens -o ./homoSapiens_outputs/ -r 9606`'.

Command Flag	Description	Required?
-g	Path to genome FASTA file	Y
-s	Name for results files (cannot contain spaces)	Y
-o	Path to output directory (. represents current directory).	Y
-r	RepeatMasker species search term used for the initial mask of known elements, in string ("Arthropoda") or NCBI taxonomy ID (6656) format.	Y

-t	Number of threads (default: 1)	N
-l	Path to FASTA file containing a custom library for initial mask of known elements (e.g if the user has some previous manually-curated elements to use)	N
-i	Number of iterations to run the BLAST, Extract, Extend process (Default: 5)	N
-f	Number of flanking bases to add in each round of the BLAST, Extract, Extend process (Default: 1000)	N
-d	Maximum distance between two TEs to define clusters	N

Table 1. Parameters for Earl Grey.

Earl Grey runs through a multi-step TE curation and annotation pipeline to annotate a given genome assembly with all intermediate results saved in their respective directories (Figure 1, Table 2). Logs are printed to '`stdout`' (the console) and saved to a log file in the Earl Grey output directory. The steps involved in the Earl Grey TE annotation procedure are outlined below:

1. The first step of Earl Grey prepares the input genome for analysis. Some tools used in the pipeline are sensitive to long header names. To prevent associated issues, header names are stored in a dictionary and replaced with generic headers using the naming convention 'ctg_n', where 'ctg' is a short generic name for each contig, and 'n' is a unique integer for each entry in the FASTA file. Ambiguous nucleotide IUPAC codes are replaced with "N" due to incompatibility with some tools, including the search engines used by RepeatMasker. The original input genome file is backed up and compressed, with the prepared input genome version saved under the same file name appended with the extension '.prep'.

2. Known repeats are identified and masked using RepeatMasker and a user-specified subset of the TE consensus libraries (i.e. RepBase and/or Dfam depending on RepeatMasker configuration). A sensitive search is performed ignoring low-complexity repeats and small RNA genes '`-s -nolow -norna`', and a hard masked version of the input genome is produced with nucleotides within known TEs replaced with 'N'.

3. The hard masked version of the genome is subsequently analysed with RepeatModeler2 for *de novo* TE identification. The optional LTR identification step included as part of RepeatModeler2 is not used, as we implement a separate LTR curation step later in the Earl Grey pipeline during the RepeatCraft stage, as this is a requirement for RepeatCraft (see step 12). RepeatModeler2 outputs a library of *de novo* consensus sequences using the following naming convention: "rnd-n_family-n#TE_Classification" (e.g rnd-1_family-256#LINE/R2-Hero).

4. The success of the RepeatModeler2 run is verified as failures can occur when annotating certain genome assemblies. For example, when annotating a genome where enough unsampled nucleotides remain to initiate a new round of RepeatModeler2, but where there are not enough unsampled long sequences for the additional round to run successfully, this leads to a program failure (e.g. https://github.com/Dfam-consortium/RepeatModeler/issues/118). If this occurs, Earl Grey will automatically restart the RepeatModeler2 run with a reduced maximum stage number to ensure it runs successfully.

5. To reduce redundancy, the set of *de novo* consensus sequences are clustered using cd-hit-est (Li and Godzik, 2006; Fu *et al.*, 2012) with parameters satisfying the TE family definition of Wicker et al. (2007), implemented as described by Goubert et al. (2022) '`-d 0 -aS 0.8 -c 0.8 -G 0 -g 1 -b 500 -r 1`'. This step is required due to RepeatModeler2 generating multiple redundant consensus sequences for a single TE family.

6. To generate maximum-length *de novo* TE consensus sequences, the *de novo* TE library is processed with an automated implementation of the "BLAST, Extract, Extend" (BEE) process described by Platt et al. (2016). Each iteration consists of three main steps: (i) BLAST. A BLASTn (Camacho *et al.*, 2009) search is performed with a *de novo* TE

majority rule consensus sequence (generated by RepeatModeler2) as the query, and the input genome as the subject, to obtain up to the top 20 hits for each TE consensus sequence. Up to the top 20 hits are taken to ensure only the highest quality TE sequences of the TE family are used to generate new consensus sequences, as including more hits introduced a higher number of ambiguous positions in new consensus sequences during testing. Top hits are selected based on BLASTn scoring parameters: bitscore, percentage identity, and length of match, where hits with the highest bitscores (indicating statistical significance of an alignment between the query TE sequence and subject genome assembly), highest percentage identities (indicating the proportion of identical positions shared in the query TE sequence and subject genome assembly), and longest lengths of match are selected. (ii) Extract. Genomic sequences corresponding to the selected BLASTn hits are extracted from the input genome. (iii) Extend. The extracted hits are extended by adding 1kb of flanking sequence to both the 5' and 3' ends (other flank lengths can be defined by the user). The resultant output files each contain sequences corresponding to a single TE family sequence with extended flanks. Each set of TE sequences is aligned using MAFFT (version 7.487) (Katoh and Standley, 2013) with the '`--auto`' flag to generate a multiple sequence alignment of the genomic TE sequences with their extended flanks. Resultant alignments are trimmed with trimAl (version 1.4.rev22) (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009) to retain high quality positions in the alignment ('`-gt 0.6`'). Updated consensus sequences are generated with EMBOSS (version 6.6.0) (Rice, Longden and Bleasby, 2000) cons ('`-plurality 3`'), resulting in new TE consensus sequences with extended ends.

7. The updated consensus sequence is then used as the initial query in a new BEE process, until five iterations have been performed, but up to ten iterations can be specified by the user.

8. Following the BEE process, redundancy is again removed by clustering sequences likely to represent the same family using cd-hit-est with the same parameters applied in the previous clustering step. It is important to include this post-BEE clustering step to reduce TE

library redundancy following the generation of extended TE consensus sequences, as the process of generating extended TE consensus sequences can reveal homology between library sequences.

9. The resulting TE consensus sequences then undergo a TE sequence reclassification step, where similarity between the *de novo* consensus sequences and known TEs is identified using a RepeatMasker run. If sufficient similarity is not detected, the repeat family is designated as 'Unclassified'.

10. The curated *de novo* consensus sequences are combined with the known TE library subset used during the initial RepeatMasker step to produce a combined TE library.

11. This combined TE library is used to annotate TEs in the input genome using RepeatMasker applying a conservative score threshold of 400 ('`-cutoff 400`') to exclude poor matches unlikely to be true TE sequences.

12. The input genome is then analysed with LTR_Finder (version 1.07) (Xu and Wang, 2007), using the LTR_Finder parallel wrapper (Ou and Jiang, 2019), to identify full-length LTR elements.

13. The TE annotations from the final RepeatMasker run are defragmented and combined with the LTR_Finder results using the loose merge '`-loose`' process in RepeatCraft (Wong and Simakov, 2018) (https://github.com/niccw/repeatcraftp), which produces a modified GFF file containing the refined TE annotations.

14. Following defragmentation, Earl Grey removes overlapping TE annotations using a custom R script employing GenomicRanges (Lawrence *et al.*, 2013), which ignores strand information and retains the longest TE of overlapping pairs.

15. Finally, to decrease the incidence of spurious hits unlikely to be true TE sequences, all TE annotations less than 100bp in length are removed before the final set of annotated TEs are quantified.

16. Summary figures are generated. Earl Grey automatically produces figures providing a general overview of TEs in the input genome (Figure 2): (i) A pie chart illustrating the proportions of the genome assembly annotated with the main TE classifications and non-

TE sequence; (ii) A repeat landscape plot, which illustrates the genetic distance between each identified TE and their respective consensus sequences (calculated using '`calcDivergenceFromAlign.pl`' utility of RepeatMasker), broadly indicative of patterns of TE activity (recently active TE copies are assumed to have low levels of genetic distance to their respective family consensus). Consistent colour keys are used for the pie chart and repeat landscape to facilitate comparability.

17. The final TE annotation file is analysed with bedtools cluster to identify clusters of TEs. By default, TEs are considered to be part of the same cluster if they are separated by <200bp, although this can be user-defined.

18. Upon completion, main results are stored within the summary files directory. This directory contains: (i) TE annotation coordinates in both GFF3 and Bed format; (ii) The combined TE library used for the final RepeatMasker annotation; (iii) The *de novo* TE library containing all *de novo* TE consensus sequences processed using the BEE methodology; (iv) Summary figures.

Path to Directory	**Contents**
outputDirectory/{speciesName}EarlGrey	All Earl Grey outputs. The EarlGrey directory is created within the directory designated as the output directory by the user, and where {speciesName} is set with the -s flag.
All directories described below are found within {speciesName}EarlGrey/	
/{species}_RepeatMasker	Results from the initial RepeatMasker run.

/{species}_Database	Hard masked input genome database required for RepeatModeler2.
/{species}_RepeatModeler	Results of the RepeatModeler2 run, including the raw and clustered *de novo* TE consensus families.
/{species}_BLASTN	Results of the first BLASTn search to identify copies of *de novo* TEs in the input genome.
/{species}_ExtractAlign	Results of the iterative BEE process, including the improved *de novo* TE consensus library with reduced redundancy.
/{species}_Curated_Library	Subset of known TE sequences used in initial RepeatMasker run, improved *de novo* TE consensus library, and a file of the two combined. All in FASTA format.
/{species}_Masked_de_novo_Repeats	Data for the reclassification of TEs in the *de novo* library.
/{species}_RepeatMasker_Against_Custom_Library	Results of the final RepeatMasker run using the combined known and *de novo* repeat libraries.
/{species}_RepeatLandscape	.divsum file generated by

	RepeatMasker. Used to calculate percentage divergence from consensus for production of repeat landscape plot.
/{species}_mergedRepeats	Results of TE defragmentation process.
/{species}_summaryFiles	Main results of Earl Grey: TE loci in input genome (GFF3 and bed), combined TE consensus library, processed *de novo* TE consensus library, TE proportion pie chart, and TE landscape plots.
/{species}_clusTErs	Results of TE clustering step

Table 2. Directories created by Earl Grey, and the outputs generated in each.

METHODS

To assess the performance of Earl Grey, we compared its performance to three existing widely-used methods: (i) RepeatMasker (version 4.1.2) with RepBase (Release 20181026) and Dfam (Release 3.4); (ii) RepeatModeler2 (Flynn *et al.*, 2020); (iii) Extensive *de novo* TE Annotator (EDTA) (Ou *et al.*, 2019). To benchmark Earl Grey against these software, we simulated a genome assembly where the coordinates and divergence of all TE copies was known, using scripts from Rodriguez and Makałowski (2022). (https://github.com/IOB-Muenster/denovoTE-eval). We generated a simulated genome with an initial size of 400Mb with 42% GC content. We inserted 11,883 TE sequences from 30 TE families sourced from Dfam into the simulated genome assembly, including nested and

diverged copies (up to 30% divergence from consensus). Representatives from a variety of TE classifications were selected, including non-autonomous elements such as MITEs. TE copy number was determined by generating random numbers following a normal distribution, with the lowest TE copy number being 5 and the highest being 782. This generated a distribution close to what we might expect in a 'real' genome assembly, where we anticipate that few TE families would be found at low copy number, the majority at intermediate copy number, and few at very high copy number (Figure 3a). Following TE insertion, the simulated genome had a size of 440Mb. Configuration files for the simulated genome are provided in Additional File 1.

The simulated genome was annotated with EDTA, RepeatMasker, RepeatModeler2, and Earl Grey (without the initial RepeatMasker step, to effectively treat all TE insertions as novel, as the sequences used to simulate the genome are from the libraries RepeatMasker is configured to use, which would lead to all TE sequences being masked in the initial RepeatMasker stage). Using scripts developed and described in Rodriguez and Makałowski (2022), TE annotation results for each methodology ('test annotations') were compared to the 'reference annotation' coordinates, detailing the exact position and identity of each TE sequence in the simulated genome, to create a confusion matrix from which the Matthews Correlation Coefficient (MCC) was calculated. An MCC score of +1 arises if all annotations are correct, a score of 0 suggests that test annotations are no better than random guesses, and a score of -1 indicates that all annotations are wrong. Bases were classified based on their agreement between the reference and test annotations as follows: bases found in both reference and test annotations were designated 'true positive' (TP), bases absent in both annotations were designated 'true negative' (TN), bases found only in the test annotation were designated 'false positive' (FP), and bases found only in the reference annotation were designated 'false negative' (FN) (Figure 3b) (Rodriguez and Makałowski, 2022). TE classifications were compared between the reference and the test annotations using bedtools intersect and subsequent analyses in R using Rstudio and the tidyverse and ape packages (Paradis *et al.*, 2006; Racine, 2013; Team, 2013; Wickham *et al.*, 2019).

Following initial benchmarking, the genome assemblies of three model and three non-model organisms were also compared: human, *Homo sapiens* (GRCh38.p13, GCF_000001405.39); fruit fly, *Drosophila melanogaster* (Release 6 plus ISO1 MT, GCF_000001215.4); thale cress, *Arabidopsis thaliana* (TAIR10.1, GCF_000001735.4); rough-barked apple tree, *Angophora floribunda* (GCA_014182895.1); malachite beetle, *Malachius bipustulatus* (GCA_910589415.1); Indian cobra, *Naja naja* (GCA_009733165.1). These sets were selected to assess the performance of Earl Grey for species with well-defined (model) and poorly defined (non-model) pre-existing TE libraries.

Earl Grey, RepeatModeler2, RepeatMasker and EDTA were all tested using default parameters. For Earl Grey and RepeatMasker (using Dfam and RepBase), the known repeat library terms used for each genome were: *H. sapiens* ('Homo sapiens'), *D. melanogaster* ('Drosophila'), *A. thaliana* ('Arabidopsis'), *A. floribunda* ("plantae"), *M. bipustulatus* ("coleoptera"), *N. naja* ("squamata"). For EDTA, the species flag was set to 'others', as defined in the documentation for analysis of species other than rice or maize.

For each *de novo* TE Library, redundancy was determined by clustering consensus sequences using the parameters described by Goubert *et al.* (2022) based on the family definition of Wicker *et al.* (2007). Overlapping annotations were filtered using a custom R script and GFF files were compared before and after filtering in R using Rstudio and the tidyverse and ape packages (Paradis *et al.*, 2006; Racine, 2013; Team, 2013; Wickham *et al.*, 2019). Shared and unique TE annotations were identified using bedtools intersect (-wao) (Quinlan and Hall, 2010).

RESULTS

Simulated Datasets: Identification of TEs

The performance of Earl Grey was compared to the following widely-used TE annotation methods: (i) RepeatMasker (version 4.1.2) with RepBase (Release 20181026) and Dfam (Release 3.4), (ii) RepeatModeler2 (Flynn *et al.*, 2020), and (iii) Extensive *de novo*

TE Annotator (EDTA) (Ou *et al.*, 2019). To assess the relative performance of Earl Grey, we generated a simulated genome of 440Mb containing TE insertions from Dfam (see methods). The annotations generated by each software were compared to the real coordinates of each TE insertion in the simulated genome to create confusion matrices from which the Matthews Correlation Coefficient (MCC) was calculated. A score between +1 and -1 is calculated, where +1 indicates a perfect annotation, -1 indicates a totally wrong annotation, and 0 indicates that the annotation is as good as a random guess. Raw annotation files are provided in additional file 2.

In the simulated genome, Earl Grey outperforms EDTA and RepeatMasker, whilst performing similarly to RepeatModeler2, with an MCC score of 0.99 (Figure 4). EDTA scored the lowest and has the highest rates of false positive and false negative annotations (Table S1; Additional File 3). RepeatMasker and RepeatModeler2 also perform well, with MCC scores of 0.97 and 0.99, respectively, showing these methods to be effective in annotating TEs.

Rather than just annotating the correct bases, it is also important that TEs are correctly classified. Therefore, we quantified the number of correct, misclassified, and missing TE annotations when using each method. RepeatMasker was excluded here as it does not generate and classify novel sequences but makes use of existing ones. Both Earl Grey and RepeatModeler2 performed equally, with high rates of correctly classified TEs (Earl Grey: 97%, RepeatModeler2: 97%), however, annotations were much more fragmented when using RepeatModeler2, demonstrated by elevated TE counts compared to the number of actual TE insertions, which arises through a single TE being annotated as multiple separate fragments (Figure 4b). EDTA struggled to correctly classify TEs with a successful classification rate of just 16% (Figure 4b). When investigating this further, EDTA appears to be annotating the majority of TEs as rolling circle elements, with 2,377 DNA elements, 1,568 LINEs, 2,957 LTR elements, 1,047 PLEs, and 418 SINEs all misclassified instead as rolling circle elements (Figure 4, Table S2; Additional File 3).

Overall, Earl Grey performs very well when annotating TEs, with very low false positive and false negative rates and annotations that closely match the real TE loci, leading to a very high MCC score and high correct classification rate. When assessing the small number of TE insertions missed by Earl Grey (n=205, 1.6%), we find no systematic bias in their classification. However, TEs that are currently missed occur adjacent to other TE insertions with overlapping annotation coordinates. These elements are currently removed due to the removal of overlapping annotations in the final results parsing in Earl Grey.

Simulated Datasets: TE Consensus Libraries

Given Earl Grey's high performance in correctly annotating TEs, we next examined how Earl Grey's TE consensus libraries compare to other software that generate *de novo* TE consensus sequences (namely EDTA and RepeatModeler2).

30 real TE families were inserted into the simulated genome. Earl Grey generated the lowest number of consensus sequences with a total of 33, Repeatmodeler2 generated 69 consensus sequences, whilst EDTA generated the highest number of consensus sequences, totalling 2,071 (Table S3, Additional File 3), with the majority of these classified as DNA elements (n=954) and LTR elements (n=829). Given the high number of consensus sequences generated by EDTA, considering only 30 real TE families were inserted into the simulated genome, we annotated the EDTA TE library with RepeatMasker to interrogate these further. Of the initial 2,071 consensus sequences, 1,223 were annotated with homology to known TEs from Dfam and RepBase. Of these, 889 were classified correctly (824 LTR, 58 DNA, and 7 rolling circle elements), whilst 334 were misclassified, including several that should be classified as non-LTR retroelements (Table S4, Additional File 3). The remaining 848 TE consensus sequences generated by EDTA share no similarity to known TEs in Dfam and RepBase, and so we cannot confirm that they represent real TEs. We acknowledge the lack of TE consensus sequences for LINEs, SINEs, *Penelope*-like elements (PLEs), and unclassified elements when using EDTA. The exclusion of tools to

identify non-LTR retroelements is acknowledged in the original EDTA paper: "*Particularly, there is no structure-based program available for the identification of LINEs. The EDTA package may therefore miss a number of elements in, for instance, vertebrate genomes that contain many SINEs and LINEs.*". The authors suggest the use of RepeatModeler following EDTA annotation. However, this suggestion is not repeated in the current GitHub repository (https://github.com/oushujun/EDTA) or software documentation, which may result in researchers missing this suggestion and assuming that EDTA is a complete TE annotation software suitable for analysing diverse genome assemblies.

Compared to EDTA and RepeatModeler2, Earl Grey generated significantly longer TE consensus sequences for DNA elements (Kruskal-Wallis, $\chi^2_2 = 14.51$, $p < 0.01$), LTR elements (Kruskal-Wallis, $\chi^2_2 = 22.91$, $p < 0.01$), and LINEs (Wilcoxon Rank Sum, $W = 35$, $p < 0.01$), whilst TE consensus sequences were not significantly different among software for rolling circle elements (Kruskal-Wallis, $\chi^2_2 = 5.31$, $p > 0.05$), PLEs (Wilcoxon Rank Sum, $W = 8$, $p > 0.05$), and SINEs (Wilcoxon Rank Sum, $W = 8$, $p > 0.05$) (Figure 5a). The generation of longer consensus sequences can be attributed to the automated implementation of the iterative "BLAST, Extract, Extend" process, that seeks to generate maximum-length consensus sequences from the initial *de novo* TE consensus sequences identified by RepeatModeler2.

Considering TE annotation, a key issue we aimed to address with Earl Grey was the redundancy that often occurs in *de novo* TE consensus libraries generated by other software. To assess redundancy levels, the *de novo* TE libraries generated by each method were each clustered to the family definition as described by Wicker *et al.* (2007), using CD-HIT-EST (-d 0 -aS 0.8 -c 0.8 -G 0 -g 1 -b 500 -r 1). *De novo* TE library redundancy was 0% for all analyses when using Earl Grey, due to the inclusion of redundancy reduction steps designed to maintain the longest representative copy of any TE family with multiple models. The highest levels of redundancy were observed when using RepeatModeler2, with 19% of TE consensus models found to be redundant (Figure 5b, Table S3; Additional File 3). This finding is consistent with those of Rodriguez and Makalowski (2022), suggesting that

redundancy in *de novo* TE libraries using RepeatModeler2 is a common occurrence. Meanwhile, 1% of TE consensus models were found to be redundant when using EDTA, which is particularly surprising given the high number of TE consensus models generated. For both RepeatModeler2 and EDTA, the highest number of redundant sequences was found for LTRs (RepeatModeler2: 25 reduced to 20; EDTA: 829 reduced to 806) (Figure 5b, Table S3; Additional File 3).

Simulated Datasets: TE Annotations

Earl Grey was developed not only to improve the generation of TE libraries, but to improve final TE annotations. Firstly, Earl Grey aims to address current issues with overlapping TE annotations. It is important to address these overlapping annotations to prevent inflation of TE count and coverage estimates, as it is not physically possible for a single locus to belong to multiple TEs. When interrogating final annotation outputs, we find overlapping annotations in the outputs of EDTA and RepeatMasker (used with either known repeats or *de novo* TEs from RepeatModeler2) which inflate TE count and coverage estimates. The most overlaps were found when using EDTA, where removing overlaps reduced overall TE coverage by 5% and TE count by 11% (Figure 6).

Secondly, Earl Grey aims to address current issues with the fragmentation of TE annotations, where a single TE can be annotated as multiple fragments due to degradation of the TE sequence. Again, this can inflate estimates of TE copy number and can also complicate investigations into TE association with host gene features. To address this, Earl Grey includes a post-annotation process employing RepeatCraft (Wong and Simakov, 2018) to merge annotations that are likely to belong to the same TE insertion. When examining annotation fragmentation, we find that all methods produce annotations that are more fragmented than Earl Grey, indicated by a mean number of annotations per Earl Grey annotation of >1 for all TE classifications (Figure 6c). Levels of fragmentation were comparable between EDTA, RepeatMasker, and RepeatModeler2.

Real Genome Assemblies: TE Consensus Sequences

Following benchmarking of Earl Grey using a simulated genome, we annotated three model and three non-model genome assemblies to compare Earl Grey to other software in real genome assemblies. TE annotations for each software are provided in Additional File 4.

Earl Grey generated the lowest number of consensus sequences for all main TE classifications. For vertebrate genomes and *A. thaliana*, EDTA generated the most consensus sequences, whilst RepeatModeler2 generated the most for invertebrates and *A. floribunda* (Figure 7). With the exception of rolling circle elements, Earl Grey also generated significantly longer TE consensus sequences (DNA: Kruskal-Wallis, $\chi^2_2 = 810.88$, $p < 0.01$; LTR: Kruskal-Wallis, $\chi^2_2 = 387.9$, $p < 0.01$; PLE: Wilcoxon Rank Sum, $W = 1649$, $p < 0.01$; LINE: Wilcoxon Rank Sum, $W = 458344$, $p < 0.01$; SINE: Wilcoxon Rank Sum, $W = 3303$, $p < 0.01$; Unclassified: Wilcoxon Rank Sum, $W = 11799556$, $p < 0.01$;) (Figure 8). This can be attributed to the automated implementation of the iterative “BLAST, Extract, Extend” process that works to generate maximum-length consensus sequences from initial *de novo* TE consensus sequences identified by RepeatModeler2. The significantly longer rolling circle consensus sequences generated by EDTA (Kruskal-Wallis, $\chi^2_2 = 271.66$, $p < 0.01$) should be treated with caution, due to the very high levels of misclassification observed when using EDTA in the simulated genome, with these consensus sequences likely belonging to a variety of TE classifications.

For RepeatModeler2, the TE classification with the most redundant sequences was SINEs, whilst the most redundant sequences generated by EDTA were LTRs. The highest levels of redundancy were found when using RepeatModeler2, with 59% of SINE models in *N. naja* found to represent the same TE family (Figure 7, Table S5; Additional File 3).

The highest levels of redundancy were seen in *N. naja* (Figure 7). As a non-model vertebrate with a large genome, this is perhaps unsurprising as one would expect large proportions of this genome to be composed of TEs, given correlations between genome size

and TE content in eukaryotes (Chénais *et al.*, 2012). The TEs in this genome are likely complex, leading to numerous models being generated for the same TE family.

Real Genome Assemblies: TE Consensus Sequences

When comparing the annotation results, we find higher proportions of all genome assemblies annotated as TE when using Earl Grey in comparison to EDTA (Figure 8, Table S6; Additional File 3). However, RepeatMasker used with Dfam and RepBase led to higher proportions of genomes annotated as TE for the three model species and *N. naja*, whilst Earl Grey annotated higher proportions in *M. bipustulatus* and *A. floribunda*. RepeatModeler2 annotated a higher proportion of *M. bipustulatus* than Earl Grey, but a lower proportion of all other genome assemblies (Figure 8, Table S6; Additional File 3).

The highest abundance of unclassified elements were identified when using Earl Grey or RepeatModeler2. This is due to RepeatModeler2 identifying putative TE sequences that do not share sufficient similarity with the TEs in Dfam and RepBase to enable their classification using the RepeatClassifier module. As such, we find higher levels of unclassified TEs in non-model species compared to model species (Figure 8, Table S6; Additional File 3). This is unsurprising given the lack of knowledge regarding the TE content of non-model organisms for which the closest libraries of known TEs are from very distant species. These species are unlikely to share a similar TE landscape as distantly related models, as this can vary considerably, even within a single genus (Wells and Feschotte, 2020; Baril and Hayward, 2022). This is demonstrated when comparing the results of *de novo* methods to the known repeat RepeatMasker run, where *M. bipustulatus* is estimated to have a TE content between 31.86% and 56.57% using methods with *de novo* TE detection, whilst RepeatMasker only annotated 3.41% of the genome assembly (Table S6, Additional File 3). This massive difference highlights the need for *de novo* TE annotation methods to be applied for non-model genome assemblies, as a lack of reference sequences in Dfam for closely related species leads to large proportions of putative TE sequences being missed. In

this case, the Coleoptera library in Dfam and RepBase consists of TEs from *Tribolium*, *Tenebrio*, *Palorus*, *Nicrophorus, Hippodamia, Dendroctonus, Cicindela, and Agrilus* beetle species, the closest of which is separated from *M. bipustulatus* by ~203Mya (Kumar *et al.*, 2017). Given the evolutionary time separating these species, TE sequences from the current Coleoptera library cannot be used to accurately identify all TEs in *M. bipustulatus* based on sequence similarity alone. Therefore, the use of library-based methodologies in the absence of *de novo* methods may lead to massive underestimation of total TE content due to sampling gaps in TE databases where non-model species are vastly underrepresented.

When interrogating final annotation outputs, we find numerous overlapping annotations in the outputs of EDTA and RepeatMasker (when used with either the Repbase and Dfam libraries, and the RepeatModeler2 library), which inflate TE count and coverage estimates (Figure 8). Using EDTA, removing these overlaps reduces annotation coverage by 28.9% on average across the 6 species, whilst annotation coverage is reduced by 8.8% for RepeatMasker and 13.5% for RepeatModeler2 (Figure 8).

Annotations produced with RepeatMasker, RepeatModeler2, and EDTA are more fragmented than Earl Grey annotations. This is demonstrated by each Earl Grey annotation, (of those that are shared), overlapping with multiple annotation loci generated by the other methodologies (Figure 8). When considering EDTA annotations, the mean number of annotations per Earl Grey annotation varied between 1.64 in *A. thaliana* and 2.67 in *N. naja* (Figure 8), whilst RepeatMasker annotations varied between 1.41 (*M. bipustulatus*) and 2.23 (*N. naja*), and RepeatModeler2 annotations varied between 1.84 (*H. sapiens*) and 2.42 (*M. bipustulatus*). In all cases, the level of fragmentation was reduced when using Earl Grey, demonstrating Earl Grey's ability to defragment repeat annotations regardless of whether the genome assembly is for a model or non-model organism.

DISCUSSION

Here, we have introduced Earl Grey, a fully automated transposable element annotation and analysis pipeline for repeat identification in genome assemblies of diverse organisms. Earl Grey provides various benefits over other pipelines employed for TE annotation. Specifically, Earl Grey was designed to increase TE consensus sequence length and reduce TE consensus library redundancy, resolve spurious overlapping and fragmented annotations, and provide users with results in standard formats for compatibility with downstream analyses. In addition, “paper-ready” summary figures are produced to provide researchers with a high-level overview of the TE landscape and activity profile for any given genome assembly.

Benchmarking of Earl Grey shows a favourable improvement in TE annotation, with an MCC very close to a perfect score of +1. In addition to being robust in terms of TE consensus generation and subsequent TE annotation, Earl Grey will benefit researchers requiring TE annotation in an “all-in-one” automated package requiring no extra analysis tools or steps. Whilst RepeatModeler2 scored highly in benchmarking, it still requires a separate RepeatMasker run following TE library generation to annotate a genome, whereas Earl Grey can be run uninterrupted from start to finish to perform all required steps in addition to providing informative summaries and figures.

Furthermore, Earl Grey also provides extra polishing steps to further optimise TE annotation results. Through the implementation of the automated “BLAST, Extract, Extend” process, Earl Grey succeeds in producing longer TE consensus sequences (Figure 7b). We acknowledge that some TE consensus sequences generated by Earl Grey are longer than the original TE sequences inserted into the simulated genome and that some of these contain ‘fuzzy ends’ of ambiguous bases that extend beyond the true TE boundaries. In the context of TE annotation, the longer TE consensus sequences generated by Earl Grey are beneficial in comparison to the shorter ones produced by the other software. Specifically, when the consensus sequences are used with RepeatMasker to annotate TEs, the modified BLAST algorithm will identify a match to the real TE, including the boundaries which are found in the longer TE consensus sequence generated by Earl Grey, whilst the algorithm will

not find matches to ambiguous fuzzy ends. Meanwhile, there is little chance of the TE boundary sequence being annotated if it is not found in the shorter TE consensus sequences generated by other software. Therefore, whilst longer TE consensus sequences cannot always be defined as ‘better’ than shorter ones, finding longer matches between individual putative TE copies in a genome brings us closer to confident identification of TE boundaries, which remains a significant challenge within the TE field. Currently, the stringent parameters used in Earl Grey during consensus generation aim to ensure a conservative approach is taken to reduce incidences of including spurious sequences at the ends of each TE consensus sequence. We will continue to develop and refine the methodologies applied to reduce the fuzzy ends on some consensus sequences and better define TE boundaries in consensus curation.

As previously discussed by Rodriguez and Makałowski (2022), redundancy in the models generated by existing *de novo* TE identification tools is an issue. We have addressed this in Earl Grey through the inclusion of library redundancy reduction steps to remove redundancy in TE consensus libraries. By clustering before and after the BEE process, it is ensured that only TE sequences from different putative TE families are included in the final TE consensus library.

When considering the extent of overlap in annotations produced by other methods, Earl Grey demonstrates a significant improvement over existing pipelines by totally eliminating overlapping TE annotations. This reduces the risk of inflated TE counts whilst remaining accurate in estimating TE coverage in genome assemblies. Furthermore, the removal of overlapping annotations ensures that each base pair of the genome assembly is only annotated as a single TE, so that results remain biologically plausible.

CONCLUSIONS AND FUTURE DEVELOPMENTS

A key aim of Earl Grey was to produce a user-friendly TE annotation pipeline that can facilitate large-scale comparative studies through a fully automated process. This first

release of Earl Grey provides an improved starting point for automated TE curation and annotation, and we will continue to develop, improve, and expand Earl Grey to meet the needs of the research community. To this end, we have identified several areas of development to be considered for future implementation.

Whilst efforts have been made to improve current automated curation methodologies, we acknowledge that for the foreseeable future, the gold standard will remain manual curation. However, this is also possible following an Earl Grey analysis, which can be used to accelerate the process, as demonstrated previously for an in-depth annotation of the Monarch butterfly genome (Baril and Hayward, 2022).

A major challenge in identifying *de novo* TEs in genome assemblies concerns the annotation of non-TE sequences. Many *de novo* methods, including those employed by Earl Grey, work by identifying sequences found in multiple copies in the genome. As TE annotation is often performed prior to gene annotation, care should be taken that multicopy genes are not incorrectly designated as TE sequences. For example, this could occur when annotating TEs in animals, where olfactory receptor genes are known to be frequently duplicated (Mombaerts, 1999). To overcome this, gene annotations or models can be used to retain genomic loci known to contain these genes of interest prior to TE annotation, although this can only be done if such models exist, and prior knowledge is held. To address this, we plan to develop a module that will identify multicopy genes and prevent them ending up in TE consensus libraries.

The advancement of genome sequencing technology has been accompanied by a corresponding decrease in sequencing costs. Consequently, the vast majority of genome assemblies being released today are of chromosomal level. For example, the Darwin Tree of Life project is aiming to release ~70,000 chromosomal level genome assemblies (https://www.darwintreeoflife.org/). With the release of chromosomal assemblies come distinct opportunities to further characterise TE landscapes. To add to the current summary plots generated by Earl Grey, we plan to release an update that will automatically generate karyoplots to show the chromosomal distribution of the main TE classifications across the

genome assembly (e.g Li *et al.* (2020); Baril and Hayward (2022)). As well as providing a visual aid, these can help to identify areas of interest for further interrogation. In addition to visual additions, there are opportunities to present the chromosomal distribution of TEs within a quantitative framework. An additional module for Earl Grey is planned that will identify statistically significant hotspots and coldspots for TE insertion, where TEs are found in higher, or lower, densities than expected if TEs were evenly distributed across the host genome (Baril and Hayward, 2022). To build on this and enable comparisons to be drawn across higher taxonomic levels, we will apply a metric for 'evenness of spread' (Baril and Hayward, 2022). This will assess the distribution of TEs within a given context and provide an estimate of how even the distribution of TE sequences is in a cross comparable manner.

A major challenge in TE annotation remains the characterisation of TE boundaries. Simulated genomes can prove a powerful tool to progress towards near-perfect TE annotation, as different pipelines can be benchmarked against genomes in which the exact TE boundaries are known. Further, there are additional opportunities to refine simulated genomes by, for example, including diverse TE sequences and host genes to better represent real genome assemblies. The tools provided by Rodriguez and Makałowski (2022) are a powerful resource for the development of TE annotation tools, as the subjectivity associated with evaluating true TE loci is removed (given that a simulated genome is populated with known elements). We suggest that benchmarking with simulated genomes should become standard practice for assessing the performance of new tools, where aspects including TE classification, base pair annotation accuracy, annotation fragmentation, and overlapping annotations can be assessed against perfect reference TE annotations. It is an open question to what extent the simulated dataset that we apply represents an optimum for such comparisons. We encourage further discussion regarding choice of simulated genome size, diversity of TEs included, TE copy number profiles, and the extent of TE decomposition modelled, towards developing a standardised simulated dataset against which tools can be compared formally and quantitatively within a standardised framework.

Overall, we have presented Earl Grey, a new fully automated TE annotation pipeline for the annotation of genome assemblies. We have shown that Earl Grey outperforms current widely used TE annotation methodologies in terms of consensus generation, reduction of TE library redundancy, removing overlapping annotations, and reducing fragmentation of annotations. Earl Grey will help to facilitate large scale comparative studies whilst maintaining reproducibility, as well as being user-friendly and producing outputs in common formats compatible with downstream analyses. We plan to continue improving on Earl Grey, incorporating suggestions and feedback from the research community. Finally, Earl Grey is an open-source project hosted on GitHub. Our aim is for the TE community to request and contribute new features and improvements to the Earl Grey project, so that it is a community-led effort to improve TE annotation.

AVAILABILITY AND REQUIREMENTS

Project Name: Earl Grey

Project Home Page: https://github.com/TobyBaril/EarlGrey

Operating Systems: Linux-based Systems (e.g Ubuntu)

Programming Language: Pipeline including software coded in Python, R, Bash, and Perl

Other Requirements: Anaconda3, Docker (Optional)

License: Open Software License v 2.1

Restrictions for Non-Academic Users: None. Some dependencies may require licences for use by non-academic users.

ABBREVIATIONS

TE: Transposable element

Non-LTR: Non-long terminal repeat

LTR: Long terminal repeat

LINEs: Long INterspersed Elements

SINEs: Short INterspersed Elements

PLEs: *Penelope*-like Elements

DECLARATIONS

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of Data and Materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing Interests

The authors declare that they have no competing interests.

Funding

TB was supported by a studentship from the Biotechnology and Biological Sciences Research Council-funded South West Biosciences Doctoral Training Partnership (BB/M009122/1). RMI was supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (109356/Z/15/Z). AH was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/N020146/1).

Authors Contributions

TB developed Earl Grey, performed analysis, produced the figures, and drafted the manuscript. RMI assisted in the development of scripts for Earl Grey. AH conceived and coordinated the study and participated in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Dr James Galbraith and Ryan Biscocho for constructive comments on early drafts of this manuscript.

Figure Legends

Figure 1. Flow chart illustrating the Earl Grey workflow. Processing is fully automated from start to finish.

Figure 2. Example summary figures produced by Earl Grey using the results from the annotation of a 440Mb simulated genome. **A**. Summary pie chart showing the proportion of the genome annotated by the main TE classifications. **B**. Repeat landscape plot. X axis shows the divergence from consensus (Kimura distance). Y axis shows the percentage of the genome annotated as TE for each level of divergence. This plot can be used to provide an estimate of TE activity within a given genome assembly, with recent activity indicated by a low distance from consensus.

Figure 3. **A**. Distribution showing the number of TE families across copy numbers in the 440Mb simulated genome. **B**. Schematic demonstrating the classification of base pairs annotated by each method in comparison to their true identity.

Figure 4. **A**. MCC and base pair classification rates for each method compared using the simulated genome. Category for each bar is indicated in the key. False positive and false negative rates shown on adjusted scale for clarity. **B**. Number of TE annotations identified by each method with each annotation labelled as classified correctly, misclassified, or missing when compared to the reference TE annotations. Colours in the key indicate classification status for each TE. The black line indicates the real number of TE insertions. The Number of TEs annotated by each software can be higher than the real number of TE insertions, as a single TE might be annotated as multiple separate fragments.

Figure 5. **A**. TE consensus sequence lengths for each of the main TE classifications. Colours indicate software used to generate consensus sequences, as indicated in the key. Black boxes indicate length of reference TE sequences from Dfam that were inserted into the simulated genome. Asterisks indicate families where Earl Grey consensus sequences are significantly longer than those generated by other software. **B**. Number of consensus sequences for EDTA and RepeatModeler2 before and after clustering redundant sequences. Grey box highlights EDTA consensus sequence count, which required axis scale adjustment due to the very high consensus sequence number generated. Colours indicate main TE classifications, as indicated in the key.

Figure 6. **A**. Number of TEs annotated in the 440Mb genome before and after overlapping annotations have been removed. **B**. Coverage of TEs annotated in the 440Mb genome before and after overlapping annotations have been removed. **C**. Number of annotations per Earl Grey annotation, demonstrating the level of fragmentation in annotation. Black line at 1 indicates the number of annotations expected if annotations are no more fragmented than Earl Grey. TE classifications, as identified by Earl Grey, are indicated in the key.

Figure 7. **A**. Number of TE consensus sequences before and after clustering consensus sequences using the Wicker definition for a TE family (80-80-80 rule) to identify redundant TE sequences. Colours represent main TE classifications as indicated in the key. **B**. Length of TE consensus sequences split by TE classification and consensus generation methodology. Asterisk (*) highlights results to take with caution given EDTA misclassification. Colours represent annotation methodology, as indicated in the key.

Figure 8. TE classifications are indicated by the colour in the key. **A**. TE coverage expressed as percentage of total genome assembly annotated using each methodology for each species. For EDTA, RepeatMasker, and RepeatModeler2, the coverage before and after removing overlapping annotations are shown. **B**. Mean number of TE annotations per Earl Grey annotation at shared loci. Error bars indicate 1 standard deviation (SD) from the mean.

Additional Files

Additional File 1: Tar archive containing configuration files to generate the 440Mb simulated genome.

Additional File 2: Tar archive containing reference TE coordinates in GFF format and GFF annotation files for each software used to annotate TEs in the simulated genome.

Additional File 3: Excel file containing all supplementary tables with contents page.

Additional File 4: Tar archive containing TE annotations for each real genome assembly annotated with each software, in GFF format.

REFERENCES

Baril, T. and Hayward, A. (2022) ‘Migrators within migrators: exploring transposable element dynamics in the monarch butterfly, Danaus plexippus’, *Mobile DNA*, 13(1), p. 5. doi:10.1186/s13100-022-00263-5.

Blaxter, M. *et al.* (2022) ‘Sequence locally, think globally: The Darwin Tree of Life Project’, *Proceedings of the National Academy of Sciences*, 119(4), p. e2115642118. doi:10.1073/pnas.2115642118.

Bourque, G. *et al.* (2018) ‘Ten things you should know about transposable elements’, *Genome biology*, 19(1), p. 199. doi:10.1186/s13059-018-1577-z.

Camacho, C. *et al.* (2009) ‘BLAST+: Architecture and applications’, *BMC bioinformatics*, 10, pp. 1–9. doi:10.1186/1471-2105-10-421.

Campbell, M.S. *et al.* (2014) ‘Genome Annotation and Curation Using MAKER and MAKER-P’, *Current Protocols in Bioinformatics*. doi:10.1002/0471250953.bi0411s48.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) ‘trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses’, *Bioinformatics* , 25(15), pp. 1972–1973. doi:10.1093/bioinformatics/btp348.

Chénais, B. *et al.* (2012) ‘The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments’, *Gene*, 509(1), pp. 7–15. doi:10.1016/j.gene.2012.07.042.

Chung, H. *et al.* (2007) ‘Cis-regulatory elements in the accord retrotransposon result in tissue-specific expression of the Drosophila melanogaster insecticide resistance gene Cyp6g1’, *Genetics*, 175(3), pp. 1071–1077. doi:10.1534/genetics.106.066597.

Chuong, E.B., Elde, N.C. and Feschotte, C. (2017) ‘Regulatory activities of transposable elements: From conflicts to benefits’, *Nature reviews. Genetics*, 18(2), pp. 71–86. doi:10.1038/nrg.2016.139.

Cosby, R.L., Chang, N.-C. and Feschotte, C. (2019) ‘Host–transposon interactions: conflict, cooperation, and cooption’, *Genes & development*, 33(17-18), pp. 1098–1116. doi:10.1101/GAD.327312.119.

Flynn, J.M. *et al.* (2020) ‘RepeatModeler2 for automated genomic discovery of transposable element families’, *Proceedings of the National Academy of Sciences*, 117(17), pp. 9451–9457. doi:10.1073/PNAS.1921046117.

Fu, L. *et al.* (2012) ‘CD-HIT: accelerated for clustering the next-generation sequencing data’, *Bioinformatics* , 28(23), pp. 3150–3152.

Goerner-Potvin, P. and Bourque, G. (2018) ‘Computational tools to unmask transposable elements’, *Nature reviews. Genetics*, 19(11), pp. 688–704. doi:10.1038/s41576-018-0050-x.

Goubert, C. *et al.* (2022) ‘A beginner’s guide to manual curation of transposable elements’, *Mobile DNA*, pp. 1–98.

Hof, A.E.V. t. *et al.* (2016) ‘The industrial melanism mutation in British peppered moths is a transposable element’, *Nature*, 534(7605), pp. 102–105. doi:10.1038/nature17951.

Hubley, R. *et al.* (2016) ‘The Dfam database of repetitive DNA families’, *Nucleic acids research*, 44(D1), pp. D81–D89. doi:10.1093/nar/gkv1272.

Jurka, J. *et al.* (2005) ‘Repbase Update, a database of eukaryotic repetitive elements’, *Cytogenetic and genome research*, 110(1-4), pp. 462–467. doi:10.1159/000084979.

Kapitonov, V.V. and Jurka, J. (2008) ‘A universal classification of eukaryotic transposable elements implemented in Repbase’, *Nature reviews. Genetics*, 9(5), pp. 411–412. doi:10.1038/nrg2165-c1.

Katoh, K. and Standley, D.M. (2013) ‘MAFFT multiple sequence alignment software version 7: Improvements in performance and usability’, *Molecular biology and evolution*, 30(4), pp. 772–780. doi:10.1093/molbev/mst010.

Kollmar, M. (2019) *Gene Prediction: Methods and Protocols*. Humana Press. Available at: https://books.google.com/books/about/Gene_Prediction.html?hl=&id=iEkZvwEACAAJ.

Kumar, S. *et al.* (2017) 'TimeTree: A Resource for Timelines, Timetrees, and Divergence Times', *Molecular biology and evolution*, 34(7), pp. 1812–1819. doi:10.1093/molbev/msx116.

Lawrence, M. *et al.* (2013) 'Software for computing and annotating genomic ranges', *PLoS computational biology*, 9(8), p. e1003118. doi:10.1371/journal.pcbi.1003118.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. "Earth BioGenome Project: Sequencing Life for the Future of Life." *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–33.

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

Li, Y. *et al.* (2020) 'Reconstruction of ancient homeobox gene linkages inferred from a new high-quality assembly of the Hong Kong oyster (Magallana hongkongensis) genome', *BMC genomics*, 21(1), pp. 1–17.

McClintock, B. (1956) '"Controlling Elements and the Gene"', *Cold Spring Harbor symposia on quantitative biology*, 21, pp. 197–216.

Mombaerts, P. (1999) 'Seven-transmembrane proteins as odorant and chemosensory receptors', *Science*, 286(5440), pp. 707–711. doi:10.1126/science.286.5440.707.

Ou, S. *et al.* (2019) 'Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline', *Genome biology*, 20(1), pp. 1–45. doi:10.1186/s13059-019-1905-y.

Ou, S. and Jiang, N. (2019) 'LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons', *BioRxiv*, pp. 2–6.

Paradis, E. *et al.* (2006) ‘ape: Analyses of Phylogenetics and Evolution’, *R package version*, 1(4). Available at: http://ape-package.ird.fr/ep/diapo_LaReunion_2009.pdf.

Platt, R.N., Blanco-Berdugo, L. and Ray, D.A. (2016) ‘Accurate transposable element annotation is vital when analyzing new genome assemblies’, *Genome biology and evolution*, 8(2), pp. 403–410. doi:10.1093/gbe/evw009.

Quinlan, A.R. and Hall, I.M. (2010) ‘BEDTools: A flexible suite of utilities for comparing genomic features’, *Bioinformatics* , 26(6), pp. 841–842. doi:10.1093/bioinformatics/btq033.

Racine, J.S. (2013) ‘RSTUDIO: A PLATFORM-INDEPENDENT IDE FOR R AND SWEAVE’, *Journal of Applied Econometrics*, 27, pp. 167–172. doi:10.1002/jae.

Rice, P., Longden, L. and Bleasby, A. (2000) ‘EMBOSS: The European Molecular Biology Open Software Suite’, *Trends in genetics: TIG*, 16(6), pp. 276–277. doi:10.1016/S0168-9525(00)02024-2.

Rodriguez, M. and Makałowski, W. (2022) ‘Software evaluation for de novo detection of transposons’, *Mobile DNA*, 13(1), p. 14. doi:10.1186/s13100-022-00266-2.

Smit, A.F.A., Hubley, R.R. and Green, P.R. (2013) ‘RepeatMasker Open-4.0’, *http://repeatmasker.org* [Preprint].

Team, R.C. (2013) ‘R: A language and environment for statistical computing’.

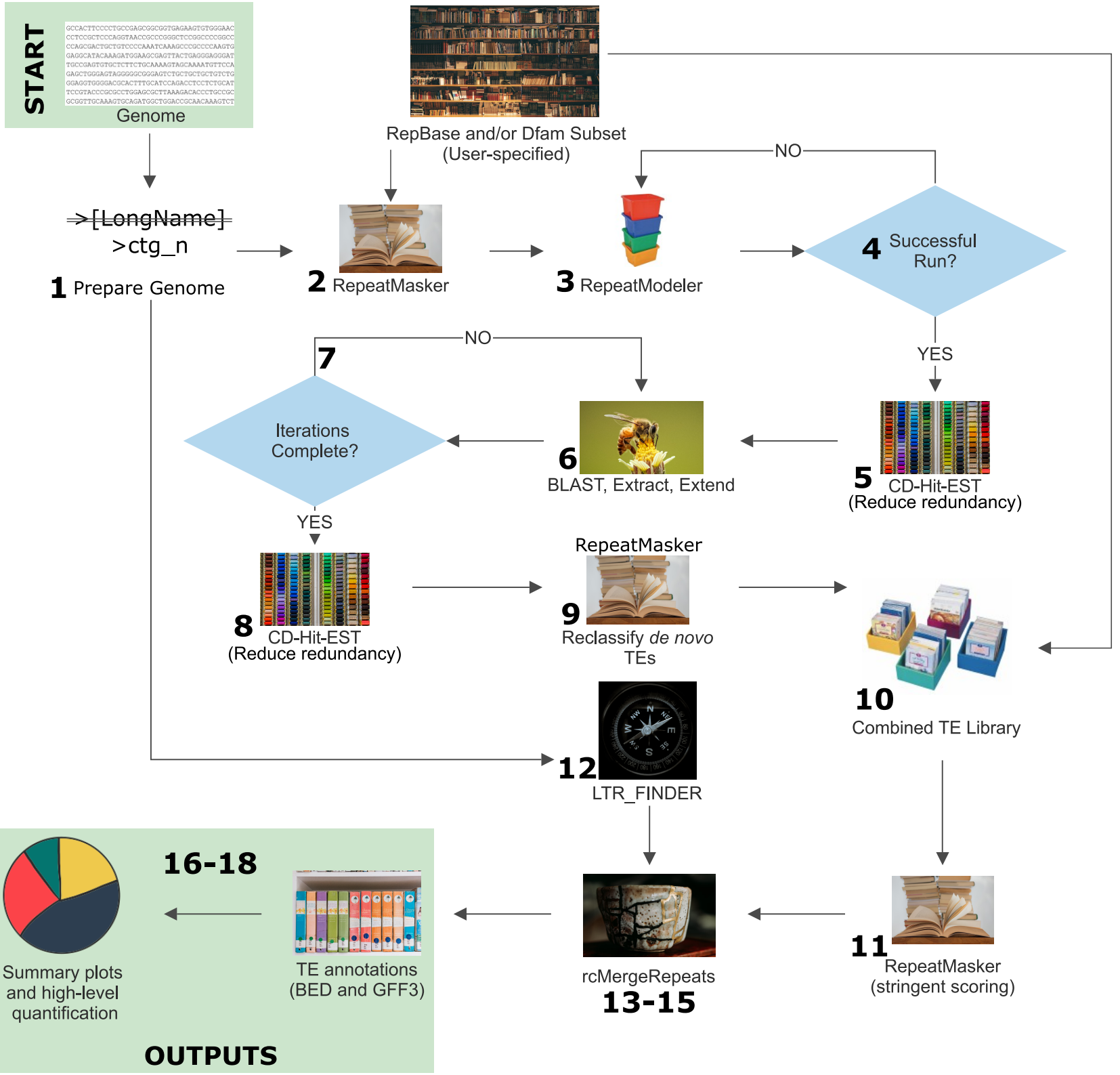
Wells, J.N. and Feschotte, C. (2020) ‘A Field Guide to Eukaryotic Transposable Elements’, *Annual review of genetics*, 54, pp. 539–561. doi:10.1146/annurev-genet-040620-022145.

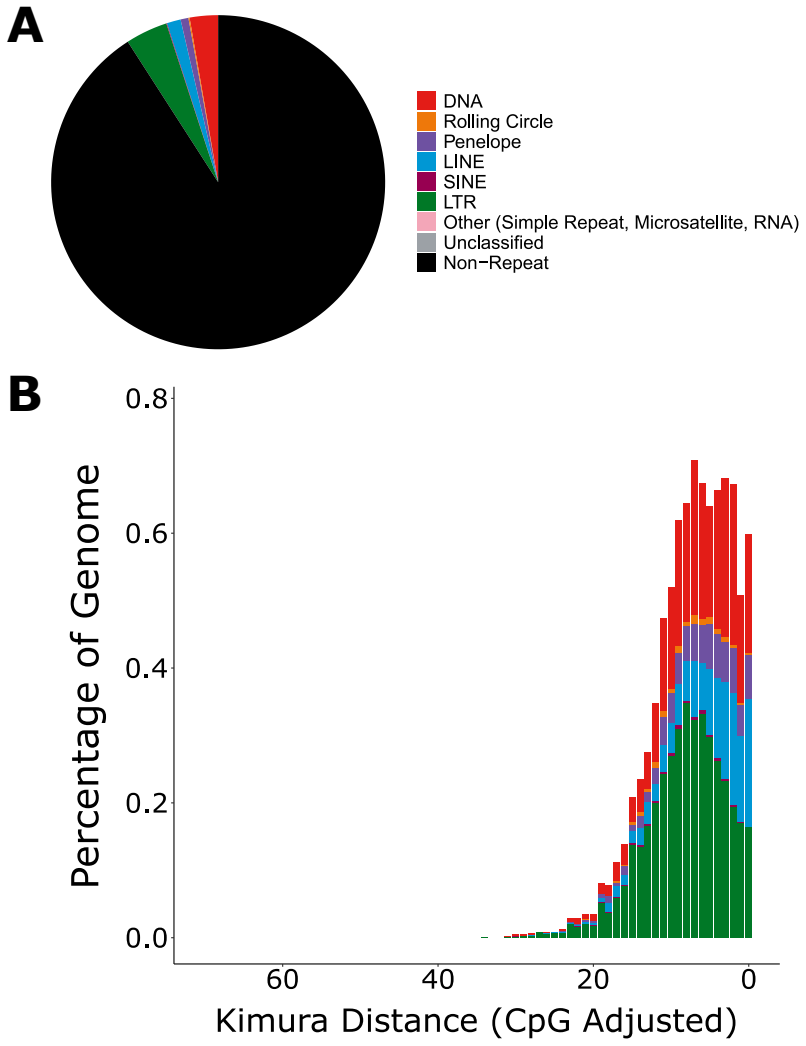
Wicker, T. *et al.* (2007) ‘A unified classification system for eukaryotic transposable elements’, *Nature reviews. Genetics*, 8(12), pp. 973–982. doi:10.1038/nrg2165.

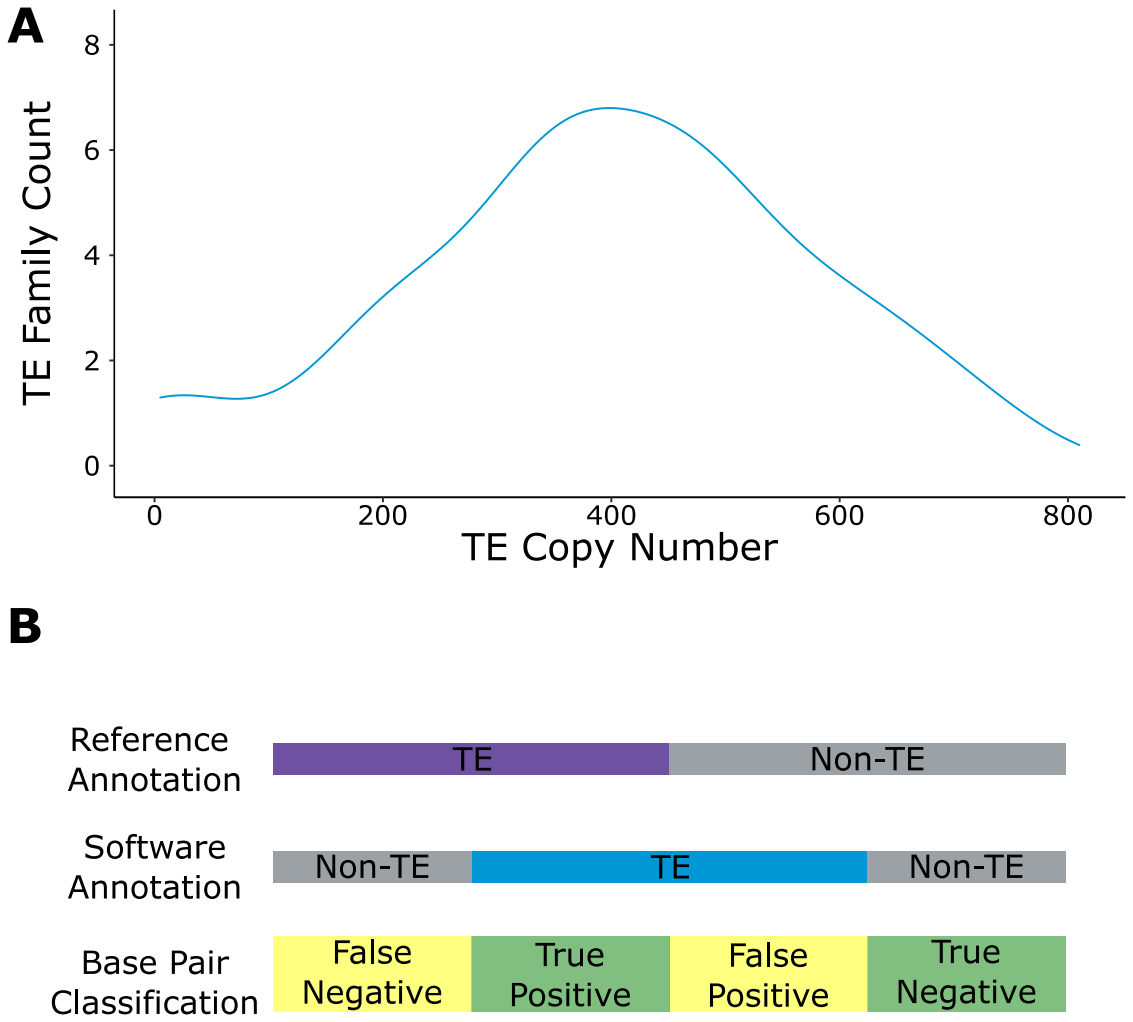
Wickham, H. *et al.* (2019) ‘Welcome to the Tidyverse’, *Journal of Open Source Software*, 4(43), p. 1686.

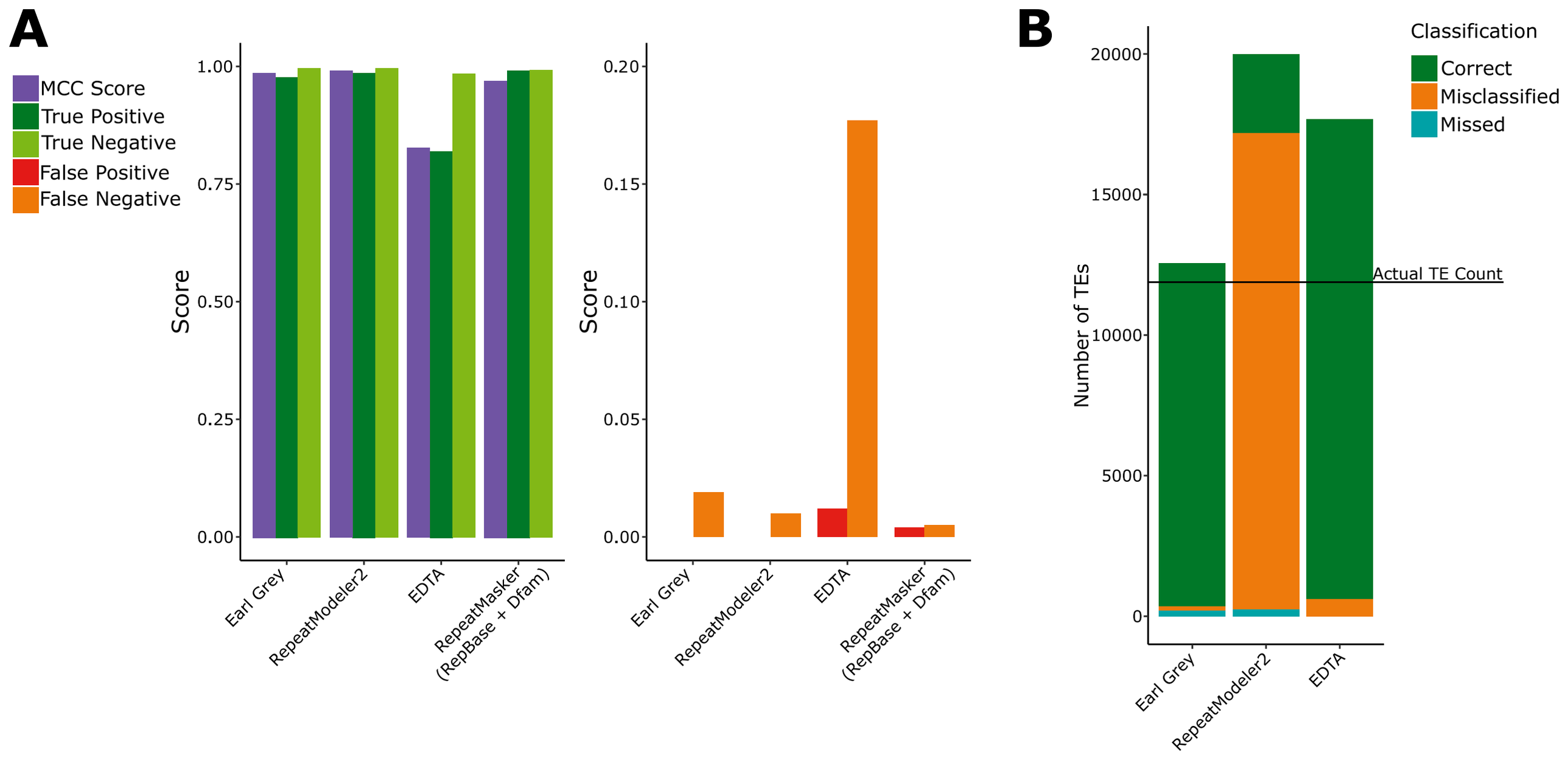
Wong, W.Y. and Simakov, O. (2018) ‘RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation’, *Bioinformatics* , 35(6), pp. 1051–1052. doi:10.1093/bioinformatics/bty745.

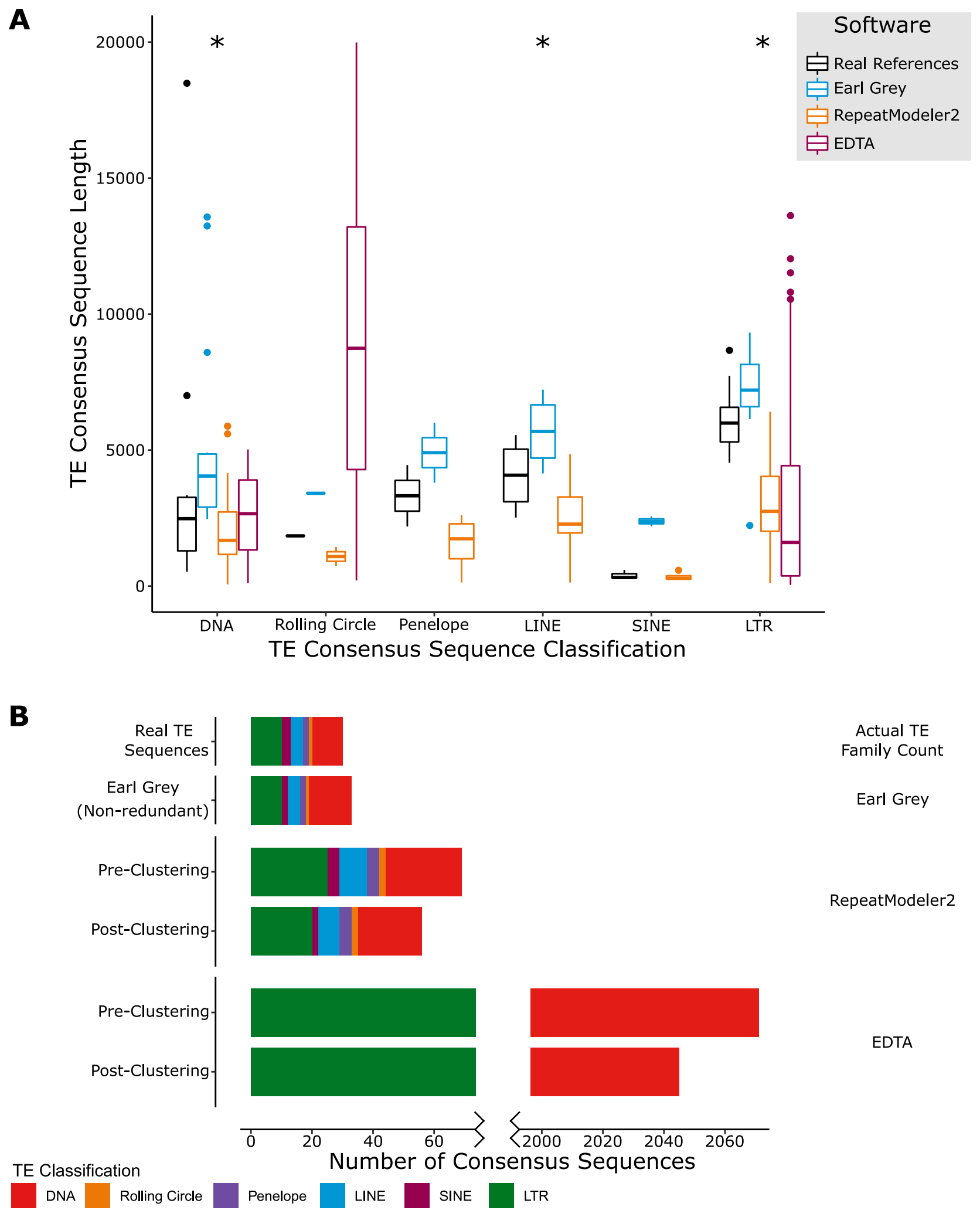
Xu, Z. and Wang, H. (2007) ‘LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons’, *Nucleic acids research*, 35(Web Server issue), pp. W265–W268. doi:10.1093/nar/gkm286.

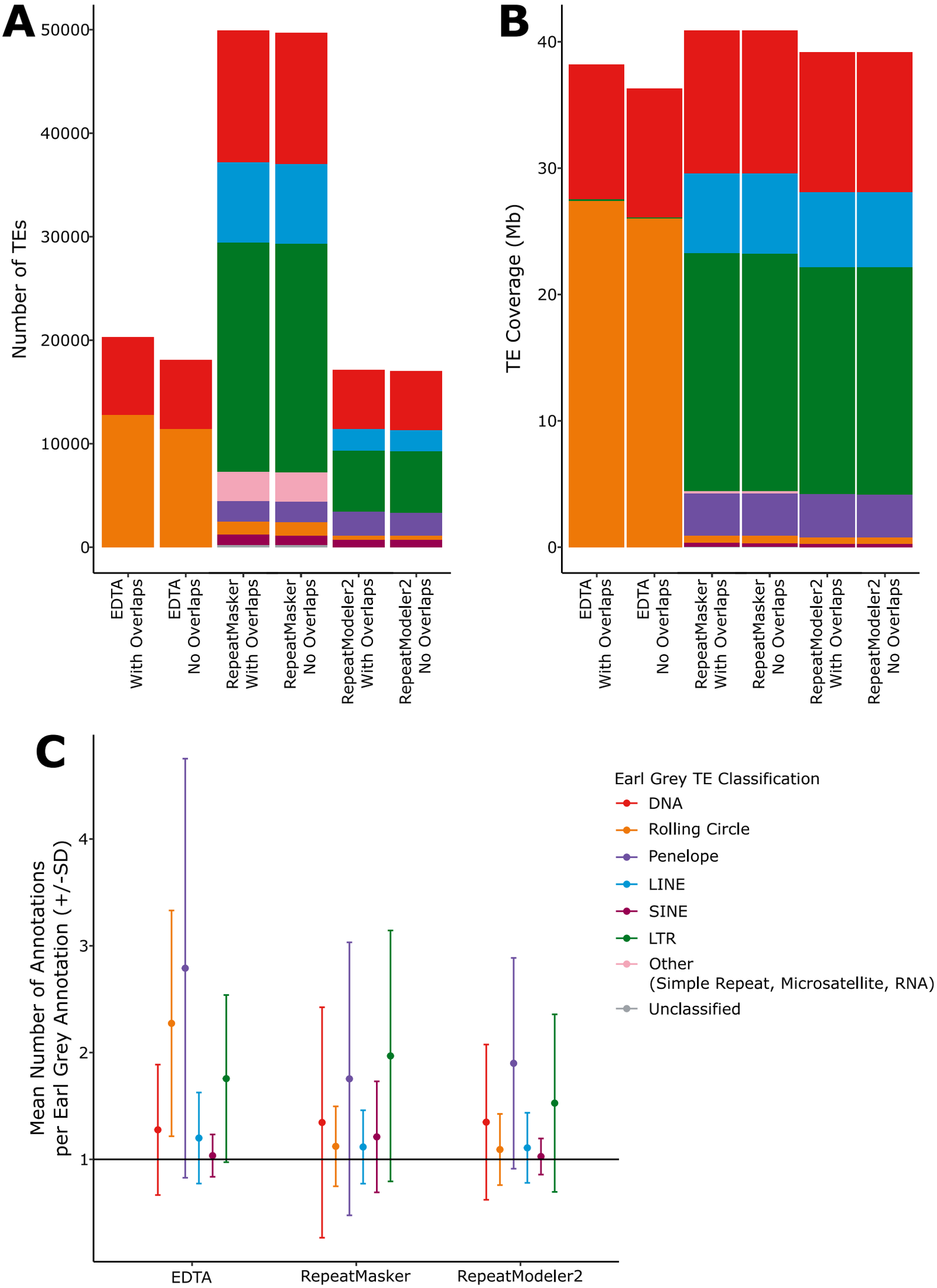

START
Genome
RepBase and/or Dfam Subset
(User-specified)
NO
~~>[LongName]~~
>ctg_n
1 Prepare Genome
2 RepeatMasker
3 RepeatModeler
4 Successful Run?
YES
7
NO
Iterations Complete?
6
BLAST, Extract, Extend
5 CD-Hit-EST
(Reduce redundancy)
YES
8 CD-Hit-EST
(Reduce redundancy)
RepeatMasker
9
Reclassify *de novo* TEs
10
Combined TE Library
12
LTR_FINDER
rcMergeRepeats
13-15
11
RepeatMasker
(stringent scoring)
16-18
TE annotations
(BED and GFF3)
Summary plots
and high-level
quantification
OUTPUTS

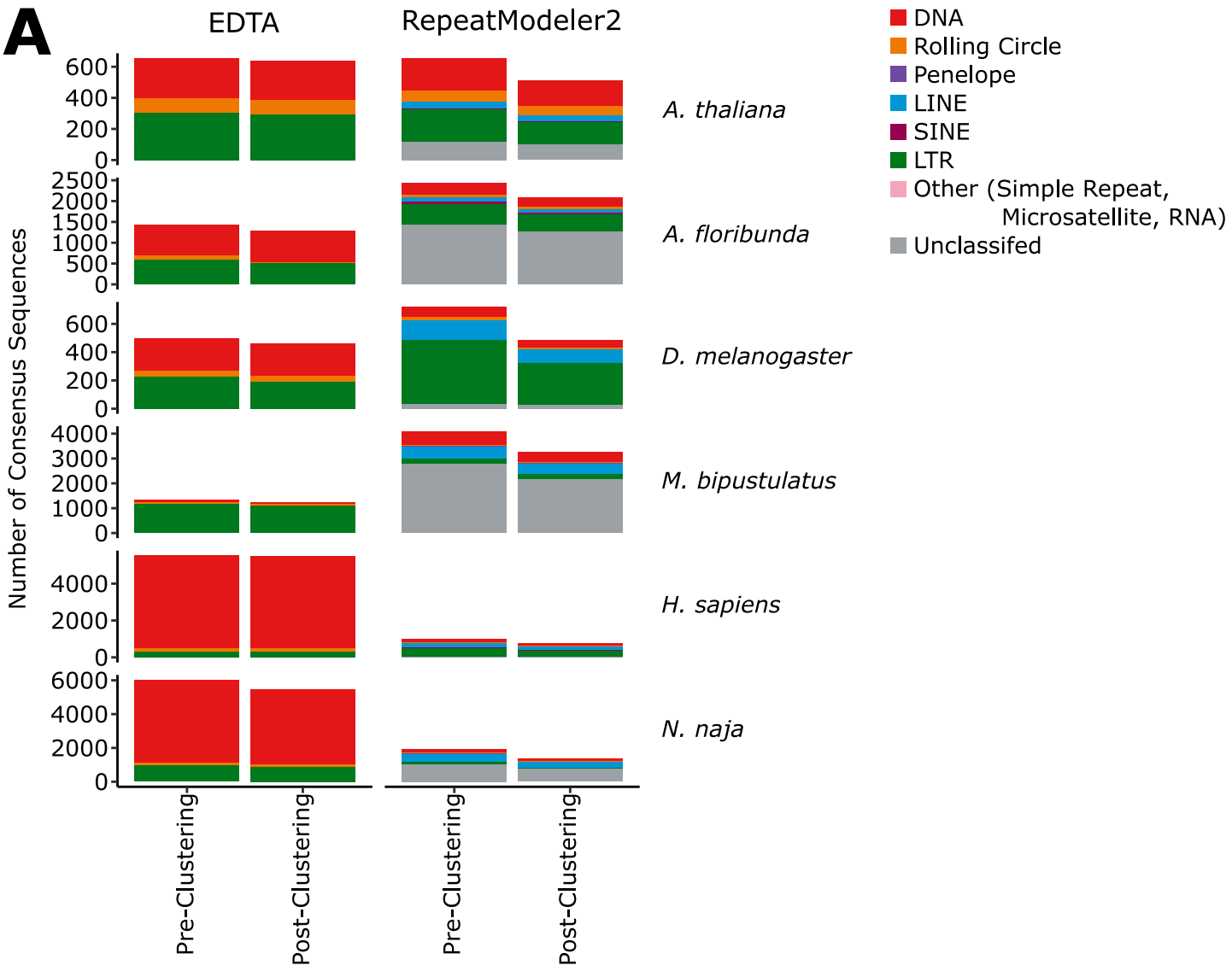

A
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other (Simple Repeat, Microsatellite, RNA)
Unclassified
Non−Repeat
B
Percentage of Genome
0.8
0.6
0.4
0.2
0.0
60
40
20
0
Kimura Distance (CpG Adjusted)

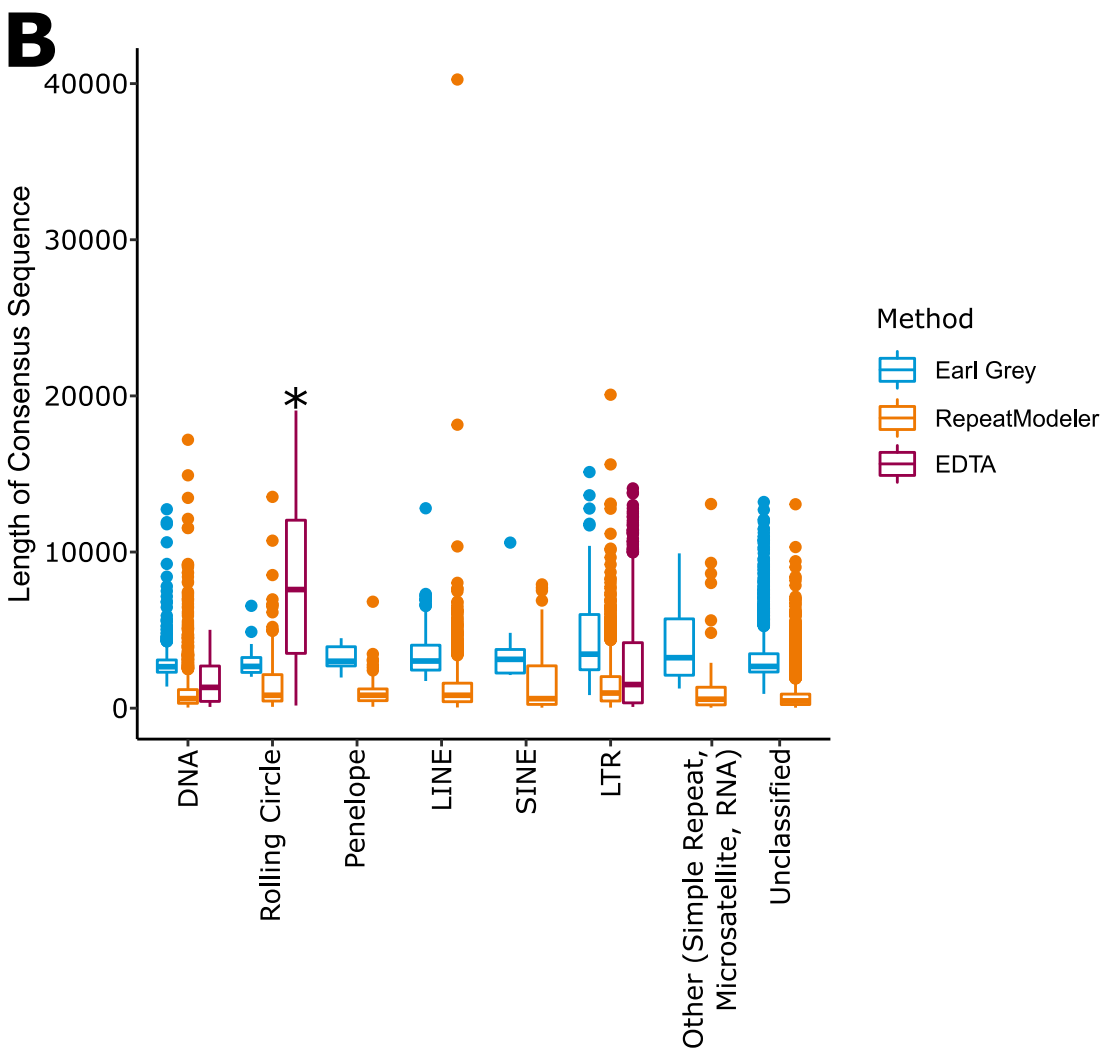

A
TE Family Count
0
2
4
6
8
0
200
400
600
800
TE Copy Number
B
Reference Annotation
TE
Non-TE
Software Annotation
Non-TE
TE
Non-TE
Base Pair Classification
False Negative
True Positive
False Positive
True Negative

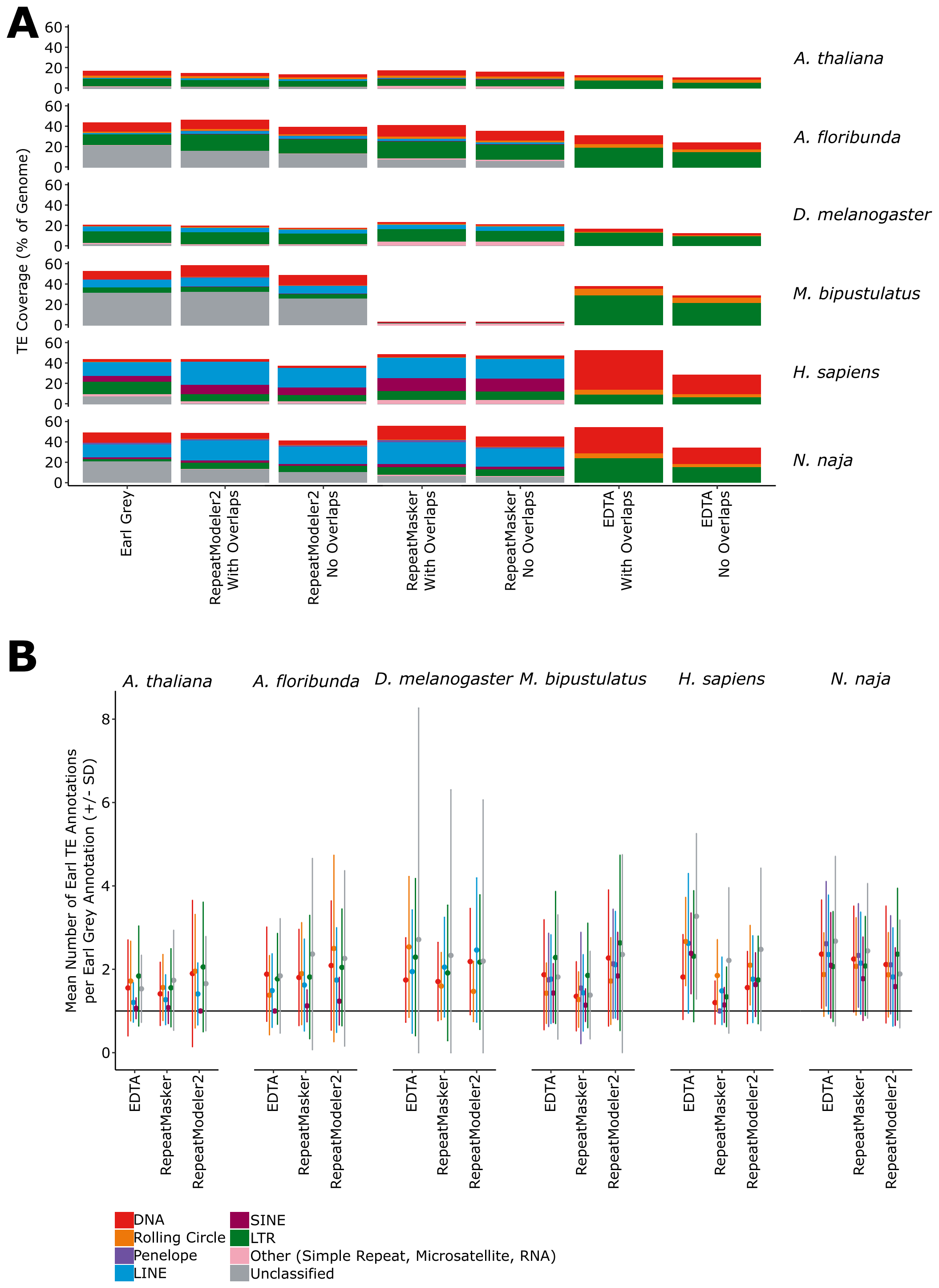

A
MCC Score
True Positive
True Negative
False Positive
False Negative
Score
1.00
0.75
0.50
0.25
0.00
Earl Grey
RepeatModeler2
EDTA
RepeatMasker (RepBase + Dfam)
Score
0.20
0.15
0.10
0.05
0.00
Earl Grey
RepeatModeler2
EDTA
RepeatMasker (RepBase + Dfam)
B
Classification
Correct
Misclassified
Missed
Number of TEs
20000
15000
10000
5000
0
Actual TE Count
Earl Grey
RepeatModeler2
EDTA


A
Software
Real References
Earl Grey
RepeatModeler2
EDTA
TE Consensus Sequence Length
20000
15000
10000
5000
0
*
*
*
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
TE Consensus Sequence Classification
B
Real TE Sequences
Earl Grey (Non-redundant)
Pre-Clustering
Post-Clustering
Pre-Clustering
Post-Clustering
Actual TE Family Count
Earl Grey
RepeatModeler2
EDTA
0
20
40
60
2000
2020
2040
2060
Number of Consensus Sequences
TE Classification
DNA
Rolling Circle
Penelope
LINE
SINE
LTR


A
Number of TEs
50000
40000
30000
20000
10000
0
EDTA With Overlaps
EDTA No Overlaps
RepeatMasker With Overlaps
RepeatMasker No Overlaps
RepeatModeler2 With Overlaps
RepeatModeler2 No Overlaps
B
TE Coverage (Mb)
40
30
20
10
0
EDTA With Overlaps
EDTA No Overlaps
RepeatMasker With Overlaps
RepeatMasker No Overlaps
RepeatModeler2 With Overlaps
RepeatModeler2 No Overlaps
C
Mean Number of Annotations per Earl Grey Annotation (+/-SD)
4
3
2
1
EDTA
RepeatMasker
RepeatModeler2
Earl Grey TE Classification
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other
(Simple Repeat, Microsatellite, RNA)
Unclassified


A
EDTA
RepeatModeler2
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other (Simple Repeat, Microsatellite, RNA)
Unclassifed
A. thaliana
A. floribunda
D. melanogaster
M. bipustulatus
H. sapiens
N. naja
Number of Consensus Sequences
Pre-Clustering
Post-Clustering
Pre-Clustering
Post-Clustering


B
Length of Consensus Sequence
Method
Earl Grey
RepeatModeler
EDTA
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other (Simple Repeat, Microsatellite, RNA)
Unclassified


A
TE Coverage (% of Genome)
A. thaliana
A. floribunda
D. melanogaster
M. bipustulatus
H. sapiens
N. naja
Earl Grey
RepeatModeler2 With Overlaps
RepeatModeler2 No Overlaps
RepeatMasker With Overlaps
RepeatMasker No Overlaps
EDTA With Overlaps
EDTA No Overlaps
B
Mean Number of Earl TE Annotations per Earl Grey Annotation (+/- SD)
EDTA
RepeatMasker
RepeatModeler2
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other (Simple Repeat, Microsatellite, RNA)
Unclassified