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

Tobias Baril[1,2,3]*, James Galbraith[1,4], Alex Hayward[1,5]*

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

[2] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Rue Émile-Argand 11, 2000 Neuchâtel, Switzerland

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

[4] ORCID: 0000-0002-1871-2108

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

* Corresponding Author information:

Tobias Baril: tobias.baril@unine.ch

Alex Hayward: alex.hayward@exeter.ac.uk

ABSTRACT

Transposable elements (TEs) are major components of eukaryotic genomes and are implicated in a range of evolutionary processes. Yet, TE annotation and characterisation remains challenging, particularly for non-specialists, since existing pipelines are typically complicated to install, run, and extract data from. Current methods of automated TE annotation are also subject to issues that reduce overall quality, particularly: (i) fragmented and overlapping TE annotations, leading to erroneous estimates of TE count and coverage; (ii) repeat models represented by short sections of total TE length, with poor capture of 5' and 3' ends. To address these issues, we present Earl Grey, a fully automated TE annotation pipeline designed for user-friendly curation and annotation of TEs in eukaryotic genome assemblies. Using nine simulated genomes and an annotation of *Drosophila melanogaster*, we show that Earl Grey outperforms current widely-used TE annotation methodologies in ameliorating the issues mentioned above, whilst scoring highly in benchmarking for TE annotation and classification, and being robust across genomic contexts. 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 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, with chromosome-level resources 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, with their own distinct evolutionary trajectories (McClintock 1956). TEs are major components of genome content across the eukaryotic tree of life, and are a major reservoir of genetic variation (Wells and Feschotte 2020). As such, TEs are implicated in a range of evolutionary processes involved in 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 et al. 2017; Cosby et al. 2019).

Tools for annotating the location and identity of TEs within a 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 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 many 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 at least 80% identity, across at least 80bp and at least 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 by eye by a human investigator, which involves identifying hallmarks of TE boundaries, such as target site duplications and long-terminal repeats, extending TE alignments as necessary, or trimming to remove poorly aligned regions unlikely to represent TE sequence. 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. Manual curation requires expert knowledge of the structure of diverse TE types, which is a specialist skill. It also requires significant time investment and is not a feasible option for general application given the rate that new genomic data are being produced. 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

104 patterns, and human error. Consequently, manual curation is generally only applicable to
105 studies focusing on single genomes, or small set of genomes.
106 To overcome the limitations of manual TE curation and facilitate large-scale
107 comparative and population-level studies, a diversity of automated TE annotation tools have
108 been developed (Goerner-Potvin and Bourque 2018). However, many of these tools are
109 species- or element-specific and need a level of technological expertise to install and
110 configure, often requiring multiple dependencies and databases, which can be difficult to
111 locate and apply due to dead weblinks, a lack of updates, and retirement of
112 codebases/dependencies. Compounding these problems is a lack of focus on user-
113 friendliness in the usage processes, which frequently includes issues relating to data
114 compatibility with downstream tools, necessitating rounds of data reformatting or cleaning.
115 Overall, this can act as a significant barrier to discourage interested biologists from
116 performing research on TEs.
117 Automated TE annotation remains a computationally challenging process. This is due
118 to the great diversity and complexity of TE sequences, the huge number of TEs that likely
119 remain uncharacterised, and the gradual post-integration erosion of TE sequences due to
120 host mutational processes, which decreases the ability to recognise and identify them.
121 Consequently, automated TE annotation tools often struggle to accurately identify TE
122 sequences. Methodologies that generate *de novo* TE consensus sequences often produce
123 TE consensus libraries with high levels of redundancy, where a single TE is represented by
124 multiple TE consensus models, and these models often perform poorly in their designation of
125 TE boundaries (Rodriguez and Makałowski 2022). Following consensus generation, TE
126 annotation can also lead to challenges that affect both *de novo* and library-based methods.
127 Fragmented TE annotations are often inferred, where a single TE is represented by multiple
128 separate annotations instead of a single continuous annotation. In addition, TE annotations
129 often contain overlaps, where a single base pair is annotated with multiple TE identities.
130 Such issues can result in distorted TE counts and coverage estimates, with knock-on
131 implications for downstream analyses. Many of these issues are particularly acute when

132 considering non-model organisms that are distantly related to model reference species,
133 where considerable effort has been expended to provide high quality TE curation.
134 The release of new genome assemblies has outpaced detailed efforts to characterise
135 their associated TEs. Furthermore, this situation will accelerate with the progress of
136 massive-scale genome sequencing efforts such as the Darwin Tree of Life Project and the
137 Earth Biogenome Project (Lewin et al. 2018), which seek to provide high quality
138 chromosomal-level genome assemblies for all eukaryotic life in the British Isles and planet
139 Earth respectively. Such projects offer huge opportunities to expand our understanding of TE
140 biology. However, the scale of genomic resources available also greatly increases the need
141 for rapid, robust, and user-friendly automated TE annotation and analysis approaches, so
142 that genomes are accurately annotated for TEs, and public TE databases are not
143 overwhelmed with poor quality sequences uploaded from substandard annotations.
144 Here, we present Earl Grey, a fully automated TE annotation pipeline combining
145 widely-used library-based and *de novo* TE annotation tools, with TE consensus and
146 annotation refinements, aimed at generating high-quality TE libraries, annotations, and
147 analyses for eukaryotic genome assemblies.
148

149 **IMPLEMENTATION**

150 To make Earl Grey as user-friendly as possible, we have provided several solutions to suit
151 the needs of most researchers. Earl Grey is available from a Github repository
152 (https://github.com/TobyBaril/EarlGrey), as a package in Bioconda (Grüning et al. 2018)
153 (recommended installation), as a preconfigured Docker/Singularity container
154 (https://biocontainers.pro/tools/earlgrey), and in a web browser via gitpod (https://gitpod.io).
155 It can be run on Linux distributions, such as Ubuntu, and installed on a local system or HPCs
156 supporting Conda, Docker, or Singularity. Earl Grey is parallelised and makes use of multiple
157 CPU threads to reduce runtime.

Earl Grey runs in a configured environment to avoid conflicts between tool versions and to streamline the installation procedure. Users who do not have RepeatMasker (Smit et al. 2013) and RepeatModeler2 (Flynn et al. 2020) installed, and who do not wish to do this, can install Earl Grey using the Bioconda package or via interactive containers with Docker and Singularity, which will install and configure dependencies automatically and provide a virtual machine in which to perform all Earl Grey analyses. For installation directly from GitHub, it is necessary to have RepeatMasker and RepeatModeler2 pre-installed. Instructions for the installation of dependencies are provided for users who do not currently have them installed. In an effort to maintain Earl Grey as open-source, Earl Grey has been tested, and is recommended to be used, with the curated subset of the Dfam database of repetitive DNA elements (Hubley et al. 2016) (tested with all versions from release 3.6 onwards. Earlier versions are not compatible due to the transition of RepeatMasker libraries to the H5 format). If users have access to RepBase RepeatMasker edition libraries (Jurka et al. 2005; Kapitonov and Jurka 2008), they can also choose to configure RepeatMasker with these in addition to Dfam.

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 3 required options, and 6 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/`'.

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). Relative file paths can be used.	Y
-r	RepeatMasker species search term used for the optional initial mask of known elements, in string (“Arthropoda”) or NCBI taxonomy ID (6656) format.	N
-l	User-provided custom TE library in FASTA format for the optional initial mask of known elements.	N
-t	Number of threads (default: 1)	N
-c	Should the *de-novo* consensus sequences be clustered to reduced library redundancy? (yes|no, Default: no)	N
-m	Do you want to remove all potentially spurious TE annotations <100bp? (yes|no, Default: no)	N
-d	Generate a softmasked genome after	N

	annotation? (yes|no, Default: no)	
-i	Number of iterations to run the BLAST, Extract, Extend, Trim process (Default: 10)	N
-f	Number of flanking bases to add in each round of the BLAST, Extract, Extend process (Default: 1000)	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). 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 '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. If required by the user, known repeats are identified and masked using RepeatMasker and a user-specified subset of the TE consensus libraries (ie Dfam and/or RepBase depending on RepeatMasker configuration), or a user-supplied custom FASTA file of consensus sequences. This is a non-default option, as in most cases it is expected that

203 annotation is being performed on previously unscreened species and a new search required,
204 rather than grouping TE families with previously described elements. A sensitive search is
205 performed ignoring small RNA genes '`-s -norna`', and a hard masked version of the input
206 genome is produced with nucleotides within known TEs replaced with 'N'.

207 3. The input genome, or masked genome if step 2 is performed, is analysed with
208 RepeatModeler2 for *de novo* TE identification (Flynn et al. 2020). The optional LTR
209 identification step included as part of RepeatModeler2 is not used, as we implement a
210 separate LTR curation step later in the Earl Grey pipeline during the RepeatCraft stage, as
211 this is a requirement for RepeatCraft (see step 12). RepeatModeler2 outputs a library of *de*
212 *novo* consensus sequences using the following naming convention: "rnd-n_family-
213 n#TE_Classification" (e.g rnd-1_family-256#LINE/R2-Hero), where rnd stands for the
214 RepeatModeler analysis round in which an element was identified (the default of 6 rounds
215 are run in Earl Grey. If annotating very large genomes, manually increasing RepeatModeler2
216 sampling rounds is recommended).

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

224 5. To generate maximum-length *de novo* TE consensus sequences, the *de novo* TE
225 library undergoes an automated implementation of the "BLAST, Extract, Extend" (BEE)
226 process described by Platt et al. (2016). In Earl Grey, this process is termed "BLAST,
227 Extract, Align, Trim" (BEAT), and is adapted from TEstrainer
228 (https://github.com/jamesdgalbraith/TEstrainer). BEAT is iteratively performed on each
229 consensus sequence until either the sequence cannot be considerably improved (based on
230 increases in sequence length), or, to prevent infinite loops, the maximum number of

specified iterations is complete. The default parameters used throughout BEAT (including extending flanks by 1000bp, a maximum of 10 iterations, and the use of 20 sequences for consensus construction) were selected through extensive manual testing. However, all parameters can be changed by editing the 'TEstrainer_for_EarlGrey.sh' script. The BEAT process proceeds as follows: (i) Each iteration begins with a search for tandem repeats within the consensus sequence using Tandem Repeat Finder (TRF) (Benson 1999). If TRF identifies >90% of the consensus sequence as tandem repeats, the consensus is classified as a tandem repeat and not curated further. If 50% to 90% of the consensus sequence is identified as tandem repeats, the longest section not identified as tandem repeats is extracted and treated as the final consensus sequence, and the sequence is not curated further. If TRF finds <50% of the consensus sequence to be tandem repeats, curation proceeds. (ii) Genome-wide copies of the repeat are identified using BLAST+ (Zhang et al. 2000) (`-task dc-megablast`). The top 20 matching sequences (defined as having >70% pairwise identity and >50% coverage of the query consensus sequence), are selected and both flanks of each of the 20 sequences is extended by 1,000bp. If at any point of the iterative process filtering reduces the number of genome-wide matches to below three, the iteration's starting consensus sequence is considered as high quality as possible, and further iterative curation ceases. To ensure the selected sequences improve on the original RepeatModeler sequence, the extended sequences are aligned to the initial RepeatModeler consensus sequence using BLAST (`-task dc-megablast`), and any which do not align with >70% pairwise identity and >50% coverage are discarded. Manual testing found that, without this step, it is possible for the final consensus sequence to occasionally represent a more abundant repetitive sequence that neighbours a copy of the initial RepeatModeler consensus sequence. (iii) The extended sequences are pairwise aligned to each other using BLAST (`-task dc-megablast`). Using these alignments, the flanking regions of the extended sequences are trimmed to only include sequence which aligns to at least one other extended sequence. A multiple sequence alignment (MSA) of the trimmed, extended sequences and

258 the starting consensus sequence is constructed using MAFFT (Katoh and Standley 2013).
259 Any columns of the resultant multiple sequence alignment (MSA) which contain only a single
260 nucleotide are removed from the MSA, followed by removal of the starting consensus
261 sequence, so that only the new extended and trimmed sequences remain. This first MSA is
262 realigned with MAFFT and a new consensus sequence is constructed using majority rule.
263 Manual testing found fewer ambiguous positions in consensus sequences constructed from
264 realigned MSAs compared to a consensus sequences constructed from the first MSA. This
265 new consensus sequence is aligned to the previous iteration's consensus sequence using
266 BLAST (`-task dc-megablast`) to verify if the sequence has been improved. Any new
267 sequence either shorter than, or having less than 80% coverage of, the previous consensus
268 sequence is treated as a reduction in quality, and the previous iteration's consensus
269 sequence is treated as the final consensus sequence. New consensus sequences longer
270 than the starting sequence by <50% of the flank extension size are treated as complete and
271 not curated further. New consensus sequences longer than the previous consensus
272 sequence by >50% of the flank extension size are submitted for a further round of curation.
273 (iv) Following the iterative BEAT curation process, TRF (Benson 1999), MREPS (Kolpakov
274 et al. 2003) and SA-SSR (Pickett et al. 2016) are used to identify potential satellite repeats in
275 the BEAT curated consensus sequence library. Any consensus sequences identified as
276 >50% tandem repeats are considered likely to be simple or satellite repeats. Within this set,
277 any sequences identified as being >90% a contiguous tandem repeat with a period of 200bp
278 or more are classified as macrosatellites, and the consensus is trimmed to a single period of
279 the repeat.
280 6. If clustering is specified by the user, the set of *de novo* consensus sequences are
281 clustered using cd-hit-est (Li and Godzik 2006; Fu et al. 2012) with parameters satisfying the
282 TE family definition of Wicker et al. (2007), implemented as described by Goubert et al.
283 (2022), i.e. '`-d 0 -aS 0.8 -c 0.8 -G 0 -g 1 -b 500 -r 1`'. This step reduces the
284 presence of redundant TE consensus sequences, such as sequences on opposite DNA
285 strands or consensus sequences formed from older fragments of a recognisable TE, which

286 are not recognised as the same element in RepeatModeler. However, it should be used with
287 caution. For example, it is possible that one TE could transpose into another, creating a new
288 transposition-competent chimeric TE, which may or may not continue to mobilise as a unit.
289 Clustering would result in the TE library containing a single consensus for the longest of
290 similar consensus sequences. Consequently, the individual independently mobilising TEs
291 would be annotated as smaller fragments of the large chimeric TE only, and not as separate
292 elements. Therefore, the default behaviour is to avoid clustering, but this can be activated at
293 the discretion of the user.

294 7. If an initial RepeatMasker step was performed, the curated *de novo* consensus
295 sequences are combined with the known TE library subset used during the initial
296 RepeatMasker step to produce a combined TE library. Otherwise, the *de novo* library is
297 used.

298 8. The *de novo* TE library (or optionally, the combined library, produced in step 7) is
299 used to annotate TEs in the input genome using RepeatMasker, with sensitive search
300 settings (which takes longer, but improves TE detection) and ignoring small RNA genes '`-s`
301 `-norna`'.

302 9. The input genome is analysed with LTR_Finder (version 1.07) (Xu and Wang
303 2007), using the LTR_Finder parallel wrapper (Ou and Jiang 2019), to identify full-length
304 LTR elements, which is required for the TE annotation defragmentation performed in step
305 10.

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

310 11. Following defragmentation, Earl Grey removes any overlapping TE annotations
311 using a custom R script employing GenomicRanges (Lawrence et al. 2013), which assigns
312 half of the length of overlapping segments to each of the two implicated TEs. For example, in

313 the case of a 10bp overlap between two elements, 5bp will be assigned to the first TE, and
314 5bp will be assigned to the second, thereby ensuring each TE is only annotated and
315 attributed to a single TE.

316 12. Finally, to decrease the incidence of spurious hits unlikely to be true TE
317 sequences, all TE annotations less than 100bp in length can be removed before the final set
318 of annotated TEs are quantified. This is at the discretion of the user and can be activated
319 with the flag '`-m yes`'.

320 13. Summary figures are generated. Earl Grey automatically produces the following
321 summary figures providing a general overview of TEs in the input genome (Figure 2): (i) A
322 pie chart illustrating the proportion of the genome assembly annotated with the main TE
323 classifications and non-TE sequence; (ii) A repeat landscape plot, which illustrates the
324 genetic distance between each identified TE and its respective consensus sequence
325 (calculated using the '`calcDivergenceFromAlign.pl`' utility of RepeatMasker), and is
326 broadly indicative of patterns of TE activity (i.e. recently active TE copies are assumed to
327 have low levels of genetic distance to their respective family consensus). Consistent colour
328 keys are used for the pie chart and repeat landscape to facilitate comparison. At the
329 discretion of the user, a softmasked version of the input genome is generated, with
330 annotated TE sequence replaced with lower-case letters in the resulting FASTA file.

331 14. Upon completion, the main results are saved within the summary files directory,
332 which contains: (i) TE annotation coordinates in both GFF3 and Bed format; (ii) The TE
333 library used for the final annotation; (iii) High-level summary tables quantifying the main TE
334 classifications, and a family-level summary to show the most abundant TE families; (iv)
335 Summary figures.

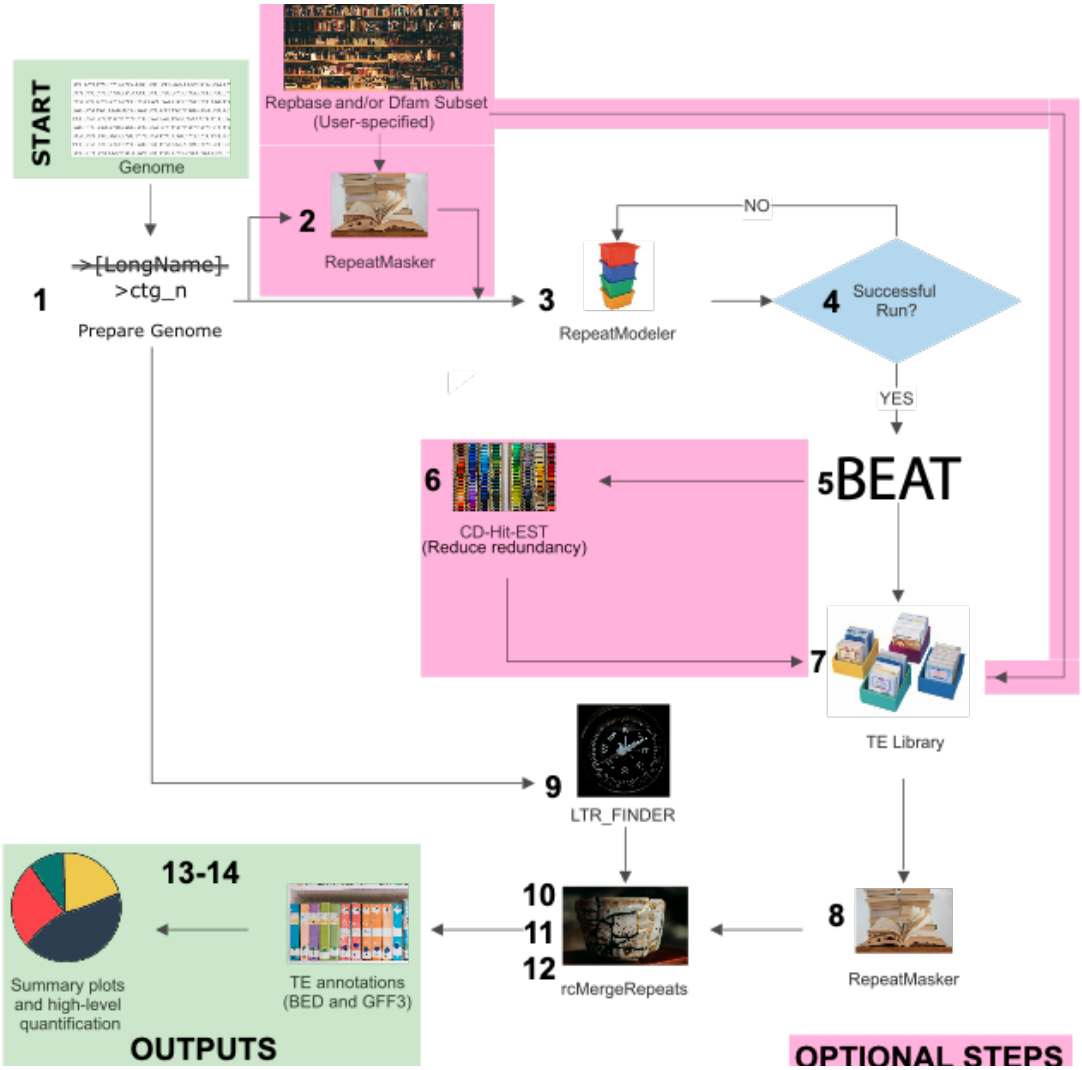


Figure 1. Flow chart illustrating the Earl Grey workflow. Processing is fully automated from start to finish. Stage numbers correspond to steps in the main text.

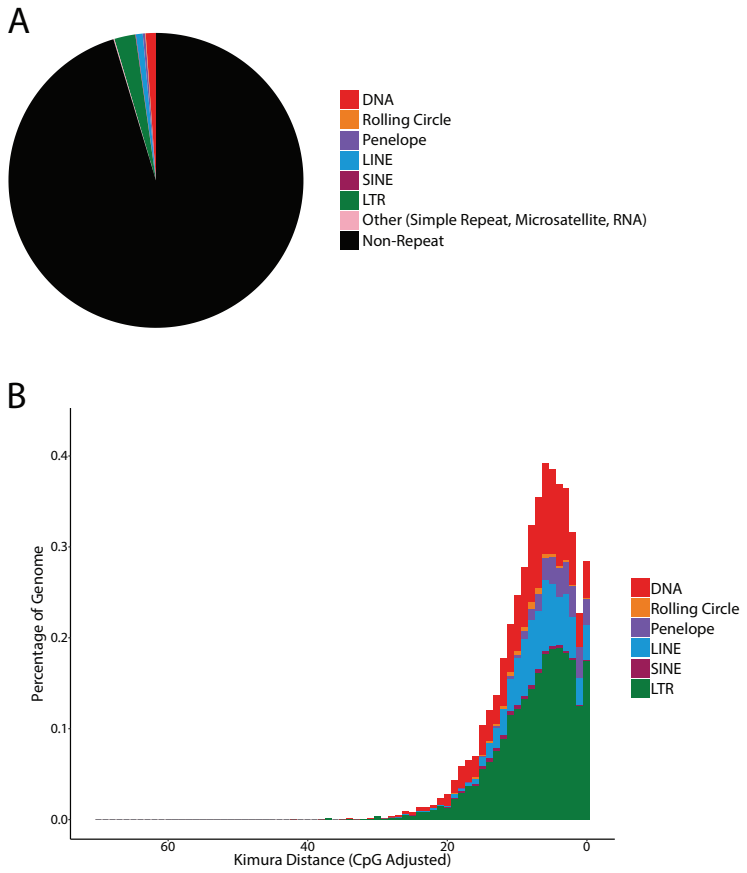


Figure 2. Example summary figures produced by Earl Grey using the results from the annotation of an 840Mb simulated genome with 58% GC content. **A**. Summary pie chart showing the proportion of the genome annotated with the main TE classifications. **B**. Repeat landscape plot, where the X axis indicates divergence from consensus in Kimura distance, and the Y axis indicates the percentage of the genome annotated as TE for each level of divergence. Landscape plots 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.

347 # METHODS

348 To assess the performance of Earl Grey, we compared it to two widely used existing
349 automated methods: (i) Extensive *de novo* TE Annotator (EDTA) (Ou et al. 2019); (ii)
350 RepeatModeler2 (Flynn et al. 2020) for *de novo* TE consensus generation, followed by
351 RepeatMasker (Smit et al. 2013) for genome assembly annotation. To benchmark Earl Grey
352 against these software, we simulated nine genome assemblies where the coordinates and
353 divergence of all TE copies was known, using scripts from Rodriguez and Makałowski (2022)
354 (https://github.com/IOB-Muenster/denovoTE-eval). We generated the nine simulated
355 genomes in a framework that facilitated consideration of the impacts of both total genome
356 size and extremes of GC content. Genomes were generated with initial sizes (before artificial
357 TE insertion) of 250Mb, 400Mb, and 800Mb, with GC content of 21%, 42% or 58%
358 (Additional File 1). Low and high GC contents were chosen to reflect extremes observed in
359 real genomes. For example, bacterial GC content can vary from <25% to <75% (Hershberg
360 2016). Fungi also display large variability in GC content, with the rumen fungus
361 *Neocallimastix californiae* having a very low GC content (18.2%) (Peng et al. 2021), whilst
362 the endophytic fungus *Falciphora oryzae* has a high GC content (56.5%) (Xu et al. 2014).
363 Considering sequenced avian, mammalian, and reptilian genomes, GC content tends to lie
364 between 40 and 50% (Bohlin and Pettersson 2019). Therefore, we selected three levels of
365 GC content to reflect TE annotation of species across the tree of life. We inserted 11,883 TE
366 sequences from 30 TE families sourced from Dfam (v3.7) (Hubley et al. 2016; Storer et al.
367 2021) into each of the simulated genome assemblies, including cut, nested, and diverged
368 copies (up to 30% divergence from consensus). Representatives from a variety of TE
369 classifications were selected, including non-autonomous elements such as MITEs (Full
370 configuration datasets for each genome are provided in Additional File 1). TE copy number
371 was determined by generating random numbers following a normal distribution, with the
372 lowest TE copy number being 5 and the highest being 732. This generated a distribution
373 close to what we might expect in a 'real' genome assembly, where we anticipate that few TE

374 families would be found at low copy number, the majority at intermediate copy number, and
375 few at very high copy number (Figure 3a). For each simulated genome, the copy number,
376 divergence percentage, nesting percentage, and cutting percentage of each TE family were
377 randomised.

378 The simulated genomes were annotated with EDTA, RepeatModeler2, and Earl Grey
379 using the default options for all tools. For EDTA, the species flag was set to 'others', as
380 defined in the documentation for analyses of species other than rice or maize. Using scripts
381 developed and described in Rodriguez and Makałowski (2022), TE annotation results for
382 each methodology ('test annotations') were compared to the 'reference annotation'
383 coordinates, detailing the exact position and identity of each TE sequence in the simulated
384 genome, to create a confusion matrix from which the Matthews Correlation Coefficient
385 (MCC) was calculated. An MCC score of +1 arises if all annotations are correct, a score of 0
386 suggests that test annotations are no better than random guesses, and a score of -1
387 indicates that all annotations are wrong. Nucleotides were classified based on their
388 agreement between the reference and test annotations as follows: nucleotides found in both
389 reference and test annotations were designated 'true positive' (TP), nucleotides absent in
390 both annotations were designated 'true negative' (TN), nucleotides found only in the test
391 annotation were designated 'false positive' (FP), and nucleotides found only in the reference
392 annotation were designated 'false negative' (FN) (Figure 3b) (Rodriguez and Makałowski
393 2022). TE classifications were compared between the reference and the test annotations
394 using BEDTools intersect and subsequent analyses in R using Rstudio and the tidyverse
395 and ape packages (Paradis et al. 2006; Racine 2013; Team 2013; Wickham et al. 2019).

396 Following initial benchmarking, the genome assembly of the fruit fly, *Drosophila*
397 *melanogaster* (Release 6.52) (Strelets et al. 2014) was annotated to test Earl Grey's
398 performance in a real genomic context.

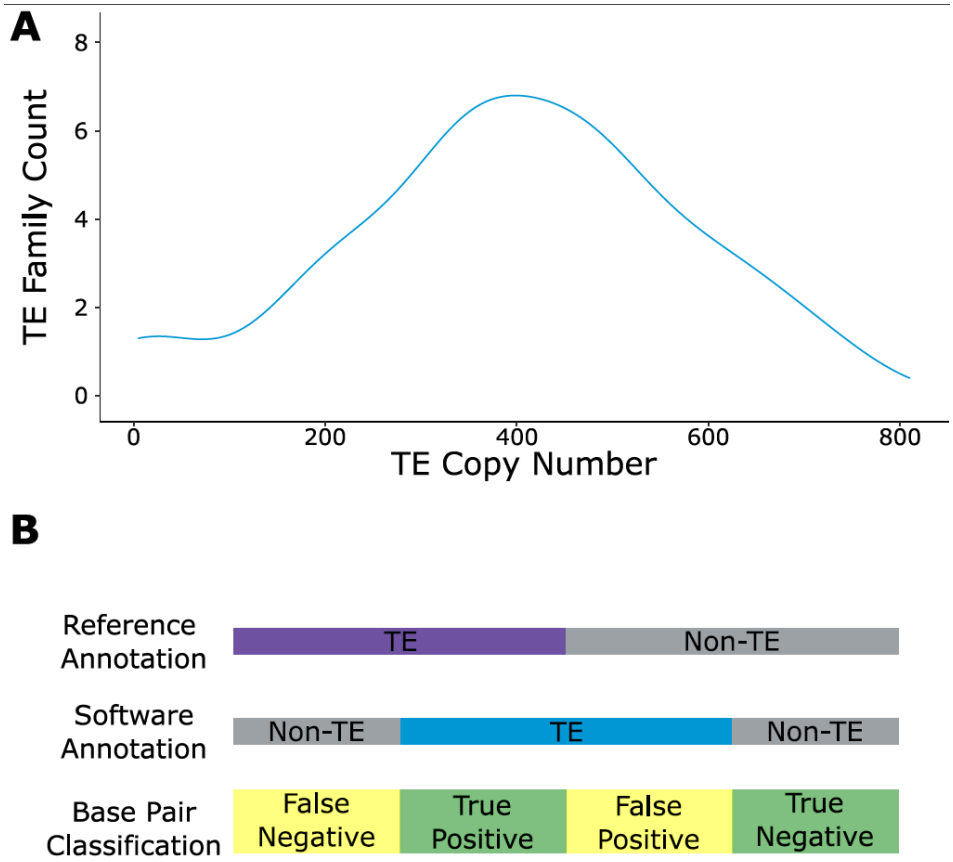
399 Overlapping annotations were filtered using a custom R script and GFF files were
400 compared before and after filtering in R using Rstudio and the tidyverse and ape packages

401 (Paradis et al. 2006; Racine 2013; Team 2013; Wickham et al. 2019). Shared and unique TE

402 annotations were identified using BEDTools intersect (`-wao`) (Quinlan and Hall 2010).

403

404



405

406

407 Figure 3. **A**. Distribution showing the number of TE families across copy numbers in the

408 simulated genomes. **B**. Schematic indicating the classification of nucleotides annotated by

409 each method in comparison to their true identity.

410

411

412

413

414

415 # RESULTS

416 ## Simulated Datasets: Identification of TEs

417 The performance of Earl Grey was compared to the following widely-used TE annotation
418 methods: (i) Extensive *de novo* TE Annotator (EDTA) (Ou et al. 2019), (ii) RepeatModeler2
419 (Flynn et al. 2020) for *de novo* TE consensus generation followed by RepeatMasker (Smit et
420 al. 2013) for genome assembly annotation. To assess the relative performance of Earl Grey,
421 we generated nine simulated genomes of varying size and GC content, containing TE
422 insertions extracted from Dfam (see methods). The annotations generated by each pipeline
423 were compared to the real coordinates of each TE insertion in the simulated 'reference'
424 genome to create confusion matrices from which the Matthews Correlation Coefficient
425 (MCC) was calculated. A score between +1 and -1 is calculated, where +1 indicates a
426 perfect annotation, -1 indicates a totally wrong annotation, and 0 indicates that the
427 annotation is as good as a random guess. Raw annotation files are provided in additional file
428 2.

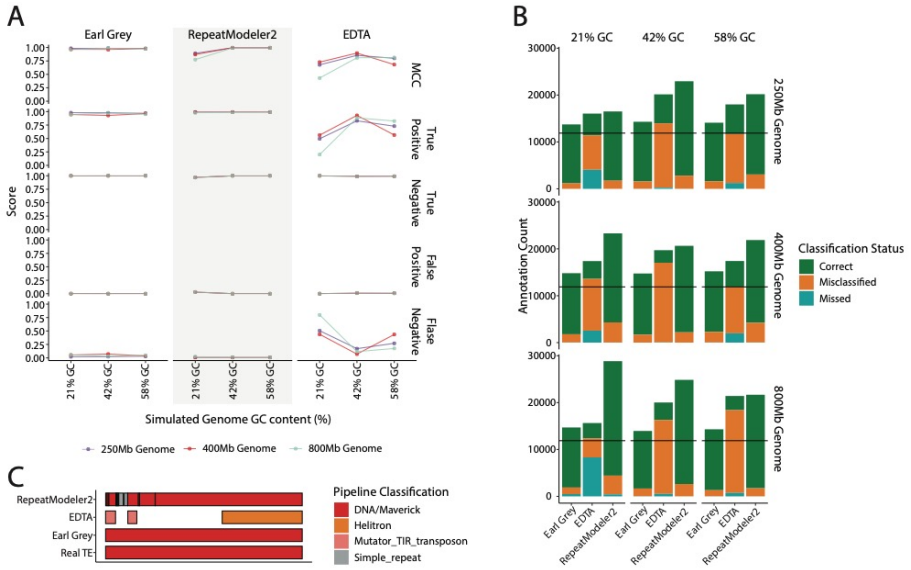
429 In the simulated genomes, Earl Grey and RepeatModeler2 outperform EDTA, with
430 MCC scores averaging 0.97 and 0.94, respectively (Figure 4). EDTA scored the lowest with
431 an average MCC of 0.67 and has the highest rates of false positive and false negative
432 annotations (Table S1; Additional File 3). Comparing Earl Grey and RepeatModeler2, the
433 difference in average MCC scores comes from Earl Grey's robust TE annotation in GC-poor
434 (or conversely AT-rich) genomes, where Earl Grey averaged an MCC score of 0.98, whilst
435 RepeatModeler2 averaged 0.84. Considering mid to high GC content, Earl Grey and
436 RepeatModeler2 both score highly (Figure 4, Table S1, Additional File 3).

437 Importantly, MCC scores do not consider whether the annotated TE nucleotides were
438 classified correctly. Rather than just identifying that nucleotides are derived from TE
439 sequence as opposed to host genome sequence, it is also important that TEs are correctly
440 classified as the correct TE type. Therefore, we quantified the number of correct,
441 misclassified, and missing TE annotations for each method. Earl Grey had the highest

442 average correct classification (88.5%) whilst RepeatModeler2 averaged 86.6% (Figure 4,
443 Table S2, Additional File 3). EDTA struggled to correctly classify TEs, with a successful
444 classification average of just 23.9% (Figure 4b, Table S2, Additional File 3). Annotations
445 were much more fragmented when using RepeatModeler2, demonstrated by elevated TE
446 counts compared to the number of actual TE insertions, which arises through a single TE
447 being annotated as multiple separate fragments (Figures 4b & 4c).

448 Overall, Earl Grey performs very well when annotating TEs, with very low false
449 positive and false negative rates and annotations that closely match the real TE loci, leading
450 to very high MCC scores and high correct classification rates. When assessing the small
451 number of TE insertions missed by Earl Grey (an average of 0.47% of real annotations), we
452 find no systematic bias in their classification.

453



454

455

456 Figure 4. **A**. MCC and nucleotide classification rates for each method compared using the
457 simulated genomes. Colours show genome size as indicated in the key. **B**. Number of TE
458 annotations identified by each method with each annotation labelled as classified correctly,
459 misclassified, or missing when compared to the reference TE annotations. Colours in the key
460 indicate classification status for each TE. The black line indicates the real number of TE

461 insertions. **C.** Schematic demonstrating how the number of TEs annotated can be higher
462 than the real number of TE insertions, as a single TE might be annotated as multiple
463 separate fragments.
464

465 **Simulated Datasets: TE Consensus Libraries**

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

469 30 real TE families were inserted into the nine simulated genomes. Earl Grey and
470 RepeatModeler2 performed comparably, generating an average of 57.8 and 57.6 consensus
471 sequences, respectively. Earl Grey generated between 49 and 64 consensus sequences,
472 whilst Repeatmodeler2 generated between 47 and 64. EDTA generated the highest number
473 of consensus sequences, with a greatly inflated average of 1,303. EDTA also showed a
474 large range in TE consensus generation, identifying between 21 and 2,948 distinct TE
475 consensus sequences (Table S3, Additional File 3). On average, EDTA identified 796 DNA
476 elements, 169 rolling circle elements, and 338 LTR elements per simulated genome. EDTA
477 did not detect any LINEs, SINEs, or *Penelope*-like elements. Given the high number of
478 consensus sequences generated by EDTA, considering only 30 real TE families were
479 inserted into the simulated genome, we combined the EDTA TE libraries and annotated their
480 consensus sequences with RepeatMasker to interrogate these further. Of the initial 11,727
481 consensus sequences, 4,903 were annotated with homology to known TEs from Dfam
482 (curated version 3.7) and RepBase (release 20181026). Of these, 3,333 were classified
483 correctly (3,036 LTR, 274 DNA, and 23 rolling circle elements), whilst 1,570 were
484 misclassified, including several that should be classified as non-LTR retroelements (Table
485 S4, Additional File 3). The remaining 6,824 TE consensus sequences generated by EDTA
486 share no similarity to known TEs in Dfam and RepBase, and so we cannot confirm that they
487 represent real TEs. The exclusion of tools to identify non-LTR retroelements is

488 acknowledged in the original EDTA paper: "*Particularly, there is no structure-based program*
489 *available for the identification of LINEs. The EDTA package may therefore miss a number of*
490 *elements in, for instance, vertebrate genomes that contain many SINEs and LINEs.*". The
491 authors suggest the use of RepeatModeler following EDTA annotation. However, this
492 suggestion is not repeated in the current GitHub repository
493 (https://github.com/oushujun/EDTA) or software documentation and may result in
494 researchers missing this suggestion and assuming that EDTA is a complete TE annotation
495 pipeline suitable for analysing diverse genome assemblies.
496 A key aim of Earl Grey is to generate longer consensus sequences, but without
497 extending them past real TE boundaries. In all simulated genomes, Earl Grey does not
498 generate TE consensus sequences that are significantly different in length from the real
499 reference TE sequences (Figure 5a, Table S5, Additional File 3). Further, when considering
500 all TE consensus sequences, Earl Grey generates significantly longer consensi than both
501 RepeatModeler2 and EDTA (Figure 5a, Table S5, Additional File 3), suggesting that these
502 pipelines underestimate TE consensus length, and thus can miss considerable proportions
503 of true TE sequence within a genome. Considering individual classifications, Earl Grey
504 generates significantly longer LTR consensus sequences than RepeatModeler2 irrespective
505 of genome size or TE content, whilst generating significantly longer LINE (400Mb 21% and
506 42% GC; 800Mb 42% GC) and DNA consensus sequences (250Mb 42% GC; 800Mb 42%
507 GC) in certain genomic contexts (Figure 5a, Table S5, Additional File 3). LTR consensus
508 sequences generated by RepeatModeler2 were significantly shorter than the real reference
509 sequences in all simulated genome assemblies except for the 800MB genome with 21% GC
510 content, in which the LINE consensus sequences were significantly shorter (Figure 5a, Table
511 S5, Additional File 3). When compared to EDTA, we find that Earl Grey generates
512 significantly longer LTR consensus sequences (400Mb 21% and 42% GC; 800Mb 21% and
513 58% GC) (Figure 5a, Table S5, Additional File 3). EDTA LTR consensus sequences were
514 significantly shorter than the real reference sequences in larger simulated genomes (400Mb
515 21% and 42% GC; 800Mb 21% and 58% GC). EDTA generated significantly longer DNA

consensus sequences than RepeatModeler2 in the 800Mb simulated genome with 42% GC content (Figure 5a, Table S5, Additional File 3). However, this result should be taken with caution given the high levels of misclassification observed with EDTA. The generation of longer consensus sequences in Earl Grey can be attributed to the automated implementation of the iterative “BLAST, Extract, Align, Trim” (BEAT) process, that seeks to generate maximum-length consensus sequences from the initial *de novo* TE consensus sequences identified by RepeatModeler2, whilst ensuring consensus sequences do not overextend true TE boundaries.

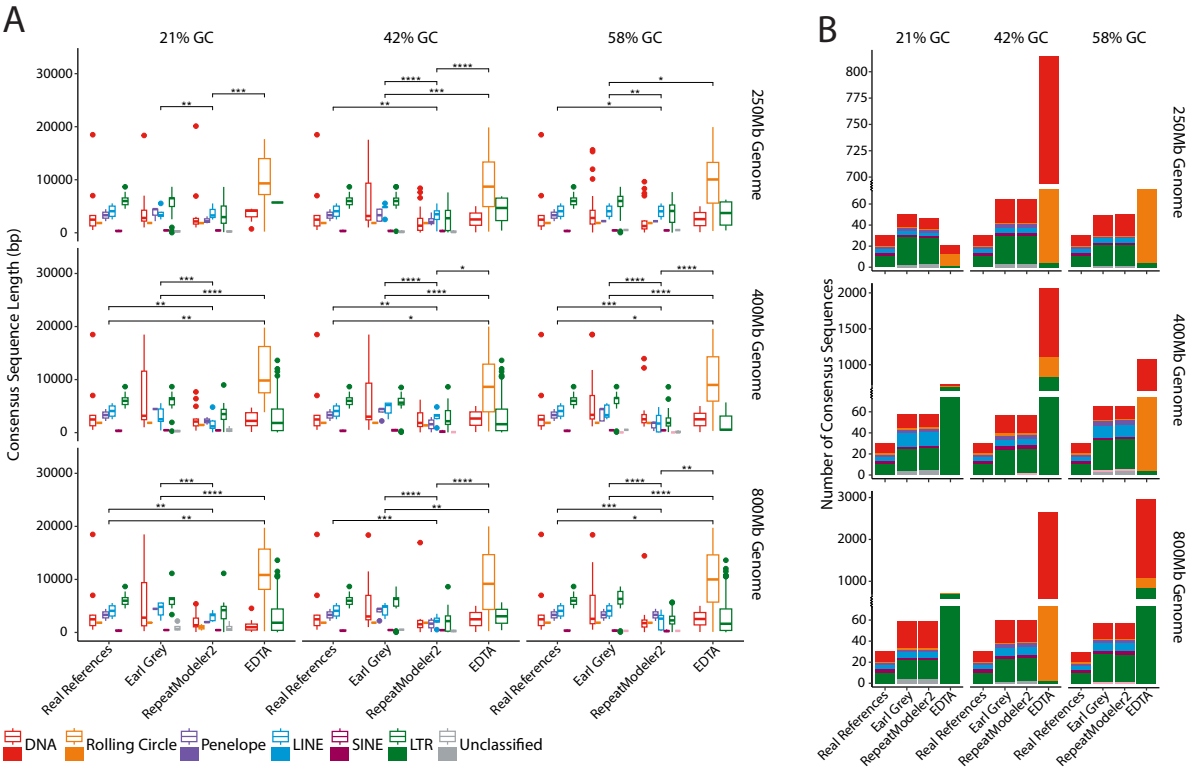


Figure 5. **A**. TE consensus sequence lengths for each of the main TE classifications. Colours show TE classifications determined by each respective methodology, as indicated in the key. Asterisks indicate levels of significance for comparisons of TE consensus length using Wilcoxon Rank-Sum Tests (*: p≤0.05; **: p≤0.01; ***: p≤0.001; **** p≤0.0001). Only significant comparisons are shown. **B**. Number of consensus sequences generated by each methodology, along with the number of real reference sequences. EDTA consensus sequence counts required axis scale adjustment due to the very high consensus sequence number generated. Colours show major TE classifications, as indicated in the key.

533 **Simulated Datasets: TE Annotations**

534 As discussed above, improving the accuracy of TE consensus sequence can have a
535 considerable impact on consensus length, and consequently the amount of correctly
536 annotated TE sequence within a genome. Additionally, Earl Grey was developed to improve
537 several other aspects of TE annotation. Firstly, Earl Grey aims to address current issues with
538 overlapping TE annotations. It is important to address overlapping annotations to prevent
539 inflation of TE count and coverage estimates, as it is not physically possible for a single
540 locus to belong to multiple TEs. When interrogating final annotation outputs, we find
541 overlapping annotations in the outputs of EDTA and RepeatMasker (used to annotate TEs
542 generated with RepeatModeler2), which inflate TE count and coverage estimates. The
543 greatest amount of overlapping annotated TE sequence was found when using EDTA,
544 where removing overlaps reduced overall TE coverage by an average of 7.8% (ranging from
545 2.4% to 12.7%) (Figure 6a, Table S6, Additional File 3) and TE count by an average of 4.8%
546 (ranging from 1.9% to 8.3%) (Table S6, Additional File 3).

547 Secondly, Earl Grey aims to address current issues with the fragmentation of TE
548 annotations, where a single TE can be annotated as multiple fragments, for example due to
549 degradation of the TE sequence. This can inflate estimates of TE copy number and it can
550 also complicate investigations into TE association with host gene features. To address this,
551 Earl Grey includes a post-annotation process employing RepeatCraft (Wong and Simakov
552 2018) to merge annotations that are likely to belong to the same TE insertion. When
553 examining annotation fragmentation, we find that all methods produce annotations that are
554 more fragmented than Earl Grey, indicated by a mean number of annotations per Earl Grey
555 annotation of >1 for all TE classifications (Figure 6b). Levels of fragmentation were
556 comparable between EDTA and RepeatMasker (using the RepeatModeler2 consensus
557 library), with an average of 1.33 and 1.42 annotations per Earl Grey annotation, respectively
558 (Figure 6b).

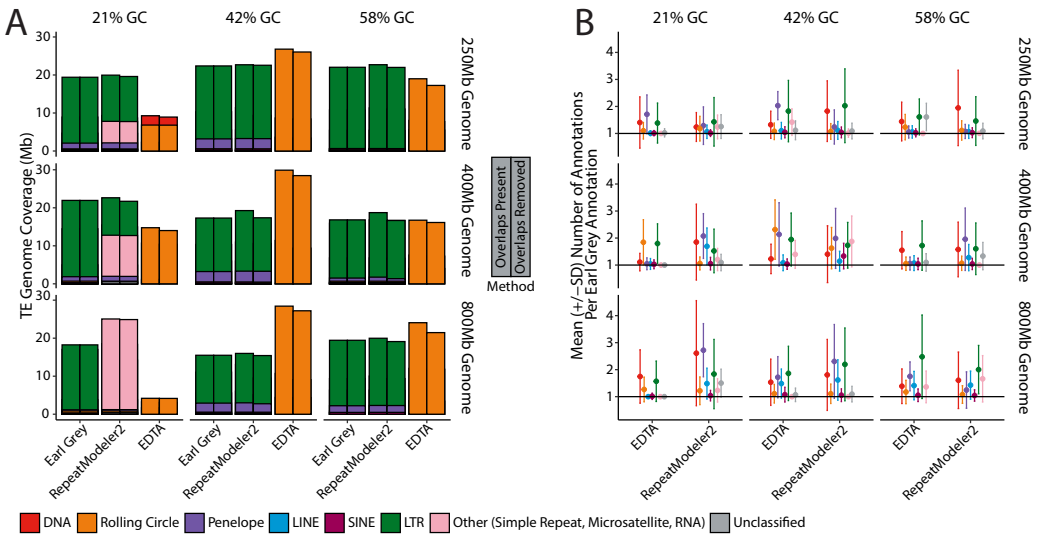


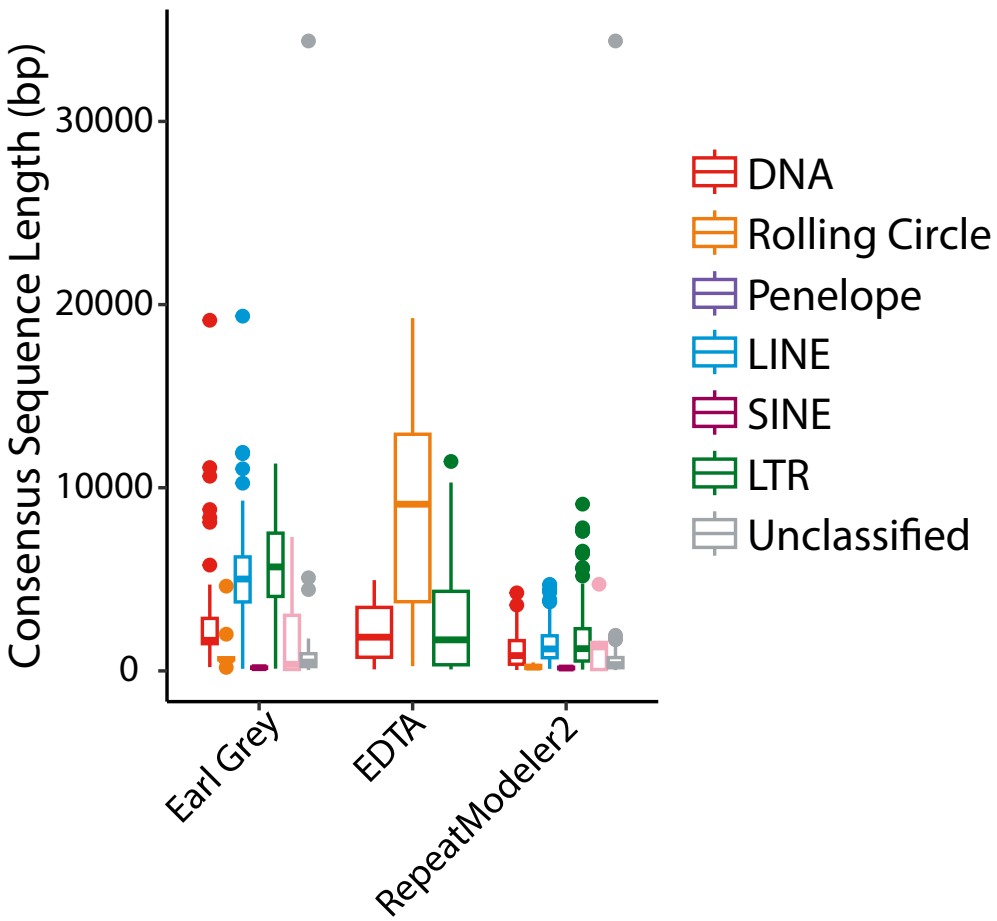
Figure 6. **A**. Number of TEs annotated in the simulated genomes before and after overlapping annotations have been removed. For each method, left hand bars show TE coverage including overlapping annotations, whilst right hand bars indicate TE coverage following removal of overlaps. TE classifications are indicated in the key **B**. Number of annotations per Earl Grey annotation for each simulated genome, demonstrating the level of fragmentation in TE annotation. The 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.

A test using a real genome assembly: TE Consensus Library quality

Following benchmarking of Earl Grey using simulated genomes, we annotated the genome of *Drosophila melanogaster* (Release 6.52) (Strelets et al. 2014) to compare Earl Grey to other software in a real genome assembly. TE annotations from each pipeline are provided in Additional File 4.

Earl Grey and RepeatModeler2 each generated 388 consensus sequences. However, there were less unclassified consensus sequences when using Earl Grey (Earl Grey: 88% classified; RepeatModeler2: 81% classified) (Table S7, Additional File 3). Meanwhile, EDTA generated 583 consensus sequences (Table S7, Additional File 3) (Figure 7). With the exception of rolling circle elements, SINEs, and unclassified elements, Earl Grey

also generated significantly longer TE consensus sequences (DNA: Kruskal-Wallis, $\chi^2_2 = 21.3$, $p < 0.01$; LTR: Kruskal-Wallis, $\chi^2_2 = 204.0$, $p < 0.01$; LINE: Wilcoxon Rank Sum, $W = 5578.5$, $p < 0.01$; Unclassified: Wilcoxon Rank Sum, $W = 1919.5$, $p < 0.01$;) (Figure 8). This can be attributed to the automated implementation of the iterative "BLAST, Extract, Align, Trim" process in Earl Grey 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, $\chi^2_2 = 48.1$, $p < 0.01$) should be treated with caution, due to the very high levels of misclassification observed when using EDTA in the simulated genomes, with these consensus sequences likely belonging to a variety of TE classifications. No significant difference in consensus length between Earl Grey and RepeatModeler2 was found for SINEs and unclassified elements (SINEs: Wilcoxon Rank Sum, $W = 2.5$, $p = 1$; Unclassified: Wilcoxon Rank Sum, $W = 1919.5$, $p = 0.15$).



592

593 Figure 7. Length of TE consensus sequences split by TE classification and consensus

594 generation methodology. EDTA results should be taken with caution given widespread

595 instances of TE misclassification identified in the simulated genomes. Colours represent TE

596 classifications, as indicated in the key.

A test using a real genome assembly: TE Annotation quality

When comparing TE annotation results, we find higher TE proportions annotated when using Earl Grey in comparison to EDTA (Figure 8, Table S8; Additional File 3). However, RepeatMasker used with the RepeatModeler2 *de novo* TE library led to higher proportions of the genome being annotated as TE than Earl Grey in *D. melanogaster* (Figure 8, Table S8; Additional File 3).

The highest abundance of unclassified elements were annotated when using RepeatModeler2. This is due to RepeatModeler2 annotating putative TE sequences that do not share sufficient similarity with the TEs in Dfam and RepBase to enable their classification using the RepeatClassifier module. This is surprising given the abundance of knowledge regarding the TE content of *D. melanogaster*, which has been the focus of TE research for at least 43 years (Green 1980; Mérel et al. 2020), and for which substantial TE resources exist. Earl Grey also makes use of the RepeatClassifier module, but it does so with longer and more refined consensus sequences, which likely reveals homology to known elements to enable their classification.

When interrogating final annotation outputs, we find numerous overlapping annotations in the outputs of EDTA and RepeatMasker (using the RepeatModeler2 library), which inflate TE count and coverage estimates (Figure 8, Table S8, Additional File 3). For EDTA, removing these overlaps reduces total annotation coverage by 22%, whilst annotation coverage is reduced by 9.5% for RepeatMasker (using the RepeatModeler2 library) (Figure 8).

Annotations produced with RepeatMasker (using the RepeatModeler2 library) 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). Specifically, each Earl Grey Annotation was represented by an average of 2.14 annotations when using EDTA, and 1.98 annotations when using RepeatMasker (with the RepeatModeler2 library). For all TE classifications, the

level of fragmentation was reduced when using Earl Grey, demonstrating Earl Grey's ability to defragment repeat annotations in a real genomic context.

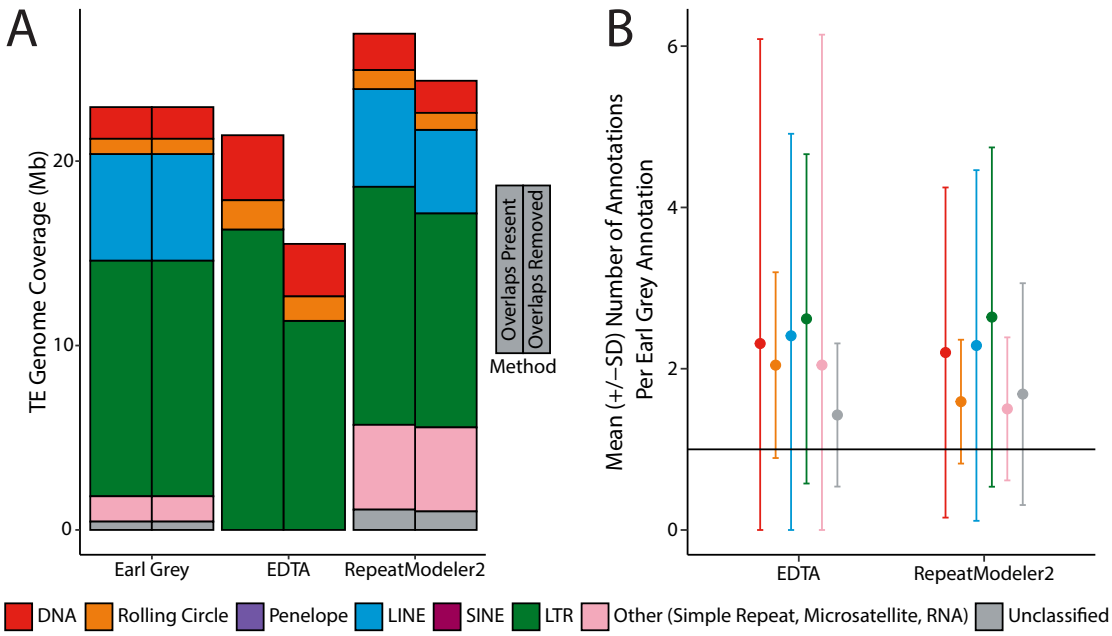


Figure 8. TE classifications are indicated by the colour in the key. **A**. TE coverage expressed as percentage of total genome assembly annotated using each methodology. Coverage before and after removing overlapping annotations is shown, with TE types coloured as indicated in the key. **B**. Mean number of TE annotations per Earl Grey annotation at shared loci. Error bars indicate 1 standard deviation (SD) from the mean (circles). The black line indicates a perfect match where one annotation matches one annotation using Earl Grey.

633 **DISCUSSION**

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

642 Benchmarking of Earl Grey shows favourable improvement in TE annotation, with an
643 MCC very close to a perfect score of +1. In addition to being robust in terms of TE
644 consensus generation and subsequent TE annotation, Earl Grey will benefit researchers
645 requiring TE annotation in an "all-in-one" automated package requiring no extra analysis
646 tools or steps. Although not as accurate as Earl Grey, RepeatModeler2 scored favourably in
647 benchmarking, but it requires a separate RepeatMasker run following TE library generation
648 to annotate a genome. In addition, Earl Grey is more robust than existing methods when
649 analysing AT-rich genomes, making it suitable for annotation of genome assemblies
650 representing diverse species from across the eukaryotic tree of life.

651 Furthermore, Earl Grey also provides extra polishing steps to further optimise TE
652 annotation results. Through the implementation of the automated "BLAST, Extract, Align,
653 Trim" process, Earl Grey succeeds in producing longer TE consensus sequences without
654 extending significantly beyond true element boundaries (Figure 7b). In the context of TE
655 annotation, the longer TE consensus sequences generated by Earl Grey are beneficial in
656 comparison to the shorter ones produced by the other software, as finding longer matches
657 between individual putative TE copies in a genome brings us closer to confident
658 identification of TE boundaries and more accurate estimations of genome-wide TE content,
659 which remains a significant challenge within the TE field, and of direct utility for various

research applications. 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, finding longer matches among individual putative TE copies in a genome, without extending to generate consensus sequences longer than real insertions, presents a step forward in our ability to accurately identify TEs in genome assemblies.

As previously discussed by Rodriguez and Makałowski (2022), redundancy in the models generated by existing *de novo* TE identification tools is an issue during annotation. We have addressed this in Earl Grey through the inclusion of optional library redundancy reduction steps. By clustering after the BEAT process, only TE sequences from different putative TE families are included in the final TE consensus library. However, this remains an optional step in the Earl Grey pipeline and this should only be used with caution, with manual evaluation of redundancies.

When considering the extent of overlap in annotations produced by other methods, Earl Grey demonstrates a significant improvement over existing pipelines by 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.

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, which can be built upon, to meet the needs of the research community. To this end, we have identified several areas of development to be considered for future implementation.

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

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

703 The advancement of genome sequencing technology has been accompanied by a
704 corresponding decrease in sequencing costs. Consequently, the vast majority of genome
705 assemblies being released today are chromosome-level. For example, the Darwin Tree of
706 Life project is aiming to release ~70,000 chromosome-level genome assemblies
707 (https://www.darwintreeoflife.org/). With the release of chromosome-level assemblies comes
708 distinct opportunities to further characterise TE landscapes. To add to the current summary
709 plots generated by Earl Grey, it would be informative to automatically generate karyoplots to
710 show the chromosomal distribution of the main TE classifications across the genome
711 assembly (e.g. see examples in (Li et al. 2020; Baril and Hayward 2022)). As well as
712 providing a visual aid, these can help to identify areas of interest for further interrogation. In
713 addition to visual additions, there are opportunities to present the chromosomal distribution
714 of TEs within a quantitative framework. An additional module for Earl could identify

715 statistically significant hotspots and coldspots for TE insertion, where TEs are found in
716 higher, or lower, densities than expected if TEs were evenly distributed across the host
717 genome (Baril and Hayward 2022). To build on this and enable comparisons across higher
718 taxonomic levels, this could include a metric for “evenness of spread” (Baril and Hayward
719 2022). This will assess the distribution of TEs within a given context and provide an estimate
720 of how even the distribution of TE sequences is in a cross comparable manner.
721 Overall, we have presented Earl Grey, a new fully-automated TE annotation pipeline
722 for the annotation of genome assemblies. We have shown that Earl Grey improves upon
723 current widely-used TE annotation methodologies in terms of consensus generation,
724 removing overlapping annotations, and reducing fragmentation of annotations. Earl Grey will
725 help to facilitate large-scale comparative genomic studies, as well as being user-friendly and
726 producing outputs in common formats compatible with downstream analyses. There remains
727 scope for further improvements of Earl Grey, especially via incorporating suggestions and
728 feedback from the needs of the research community. Finally, Earl Grey is an open-source
729 project hosted on GitHub. Our aim is for the TE community to request and contribute new
730 features and improvements to the Earl Grey project, so that it is a community-led effort to
731 improve TE annotation.
732

733 AVAILABILITY AND REQUIREMENTS

734 Project Name: Earl Grey
735 Project Home Page: https://github.com/TobyBaril/EarlGrey
736 Operating Systems: Linux-based Systems (e.g Ubuntu), MacOS-based systems (OSX), or a
737 web browser (for the gitpod implementation)
738 Programming Language: Pipeline including software coded in Python, R, Bash, and Perl
739 Other Requirements: Anaconda3, Docker (Optional), Singularity (Optional)
740 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). AH was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/N020146/1), which also supported JG.

Authors Contributions

TB developed Earl Grey, performed analysis, produced the figures, and drafted the manuscript. JG developed the 'BLAST, Extract, Align, Trim' 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 thank Dirk-Jan van Workum for assisting with the conda recipe for Earl Grey, users of early iterations of the software for constructive feedback and suggestions, and the reviewers for their constructive comments.

For the purpose of open access, the author has applied a 'Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Figure Legends

Figure 1. Flow chart illustrating the Earl Grey workflow. Processing is fully automated from start to finish. Stage numbers correspond to steps in the main text.

Figure 2. Example summary figures produced by Earl Grey using the results from the annotation of an 840Mb simulated genome with 58% GC content. **A**. Summary pie chart showing the proportion of the genome annotated with the main TE classifications. **B**. Repeat landscape plot, where the X axis indicates divergence from consensus in Kimura distance, and the Y axis indicates the percentage of the genome annotated as TE for each level of divergence. Landscape plots 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 simulated genomes. **B**. Schematic indicating the classification of nucleotides annotated by each method in comparison to their true identity.

Figure 4. **A**. MCC and nucleotide classification rates for each method compared using the simulated genomes. Colours show genome size as indicated in the key. **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. **C.** Schematic demonstrating how 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 show TE classifications determined by each respective methodology, as indicated in

813 the key. Asterisks indicate levels of significance for comparisons of TE consensus length
814 using Wilcoxon Rank-Sum Tests (*: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$; **** $p \leq 0.0001$). Only
815 significant comparisons are shown. **B**. Number of consensus sequences generated by each
816 methodology, along with the number of real reference sequences. EDTA consensus
817 sequence counts required axis scale adjustment due to the very high consensus sequence
818 number generated. Colours show major TE classifications, as indicated in the key.
819
820 Figure 6. **A**. Number of TEs annotated in the simulated genomes before and after
821 overlapping annotations have been removed. For each method, left hand bars show TE
822 coverage including overlapping annotations, whilst right hand bars indicate TE coverage
823 following removal of overlaps. TE classifications are indicated in the key **B**. Number of
824 annotations per Earl Grey annotation for each simulated genome, demonstrating the level of
825 fragmentation in TE annotation. The black line at 1 indicates the number of annotations
826 expected if annotations are no more fragmented than Earl Grey. TE classifications, as
827 identified by Earl Grey, are indicated in the key.
828
829 Figure 7. Length of TE consensus sequences split by TE classification and consensus
830 generation methodology. EDTA results should be taken with caution given widespread
831 instances of TE misclassification identified in the simulated genomes. Colours represent TE
832 classifications, as indicated in the key.
833
834 Figure 8. TE classifications are indicated by the colour in the key. **A**. TE coverage expressed
835 as percentage of total genome assembly annotated using each methodology. Coverage
836 before and after removing overlapping annotations is shown, with TE types coloured as
837 indicated in the key. **B**. Mean number of TE annotations per Earl Grey annotation at shared
838 loci. Error bars indicate 1 standard deviation (SD) from the mean (circles). The black line
839 indicates a perfect match where one annotation matches one annotation using Earl Grey.

Additional Files

Additional File 1: Tar archive containing configuration files and TE families to generate all nine simulated genomes.

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 genomes.

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

Additional File 4: Tar archive containing TE annotations for *D. melanogaster* annotated with each software, in GFF format.

REFERENCES

Baril T, Hayward A. 2022. Migrators within migrators: exploring transposable element dynamics in the monarch butterfly, Danaus plexippus. *Mob. DNA* [Internet] 13:5. Available from: http://dx.doi.org/10.1186/s13100-022-00263-5

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* [Internet] 27:573–580. Available from: http://dx.doi.org/10.1093/nar/27.2.573

Bohlin J, Pettersson JH-O. 2019. Evolution of genomic base composition: From single cell microbes to multicellular animals. *Comput. Struct. Biotechnol. J.* [Internet] 17:362–370. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6429543/

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biol.* [Internet] 19:199. Available from: http://dx.doi.org/10.1186/s13059-018-1577-z

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics* [Internet] 48. Available from: http://dx.doi.org/10.1002/0471250953.bi0411s48

Chung H, Bogwitz MR, McCart C, Andrianopoulos A, Ffrench-Constant RH, Batterham P, Daborn PJ. 2007. Cis-regulatory elements in the accord retrotransposon result in tissue-specific expression of the Drosophila melanogaster insecticide resistance gene Cyp6g1. *Genetics* [Internet] 175:1071–1077. Available from:

http://dx.doi.org/10.1534/genetics.106.066597

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* [Internet] 18:71–86. Available from: http://dx.doi.org/10.1038/nrg.2016.139

Cosby RL, Chang N-C, Feschotte C. 2019. Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev.* [Internet] 33:1098–1116. Available from: http://genesdev.cshlp.org/content/33/17-18/1098.full

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* [Internet] 117:9451–9457. Available from: https://www.pnas.org/content/117/17/9451

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152.

Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. *Nat. Rev. Genet.* [Internet] 19:688–704. Available from: http://dx.doi.org/10.1038/s41576-018-0050-x

Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. 2022. Correction: A beginner's guide to manual curation of transposable elements. *Mob. DNA* [Internet] 13:15. Available from: http://dx.doi.org/10.1186/s13100-022-00272-4

Green MM. 1980. Transposable elements in Drosophila and other Diptera. *Annu. Rev. Genet.* [Internet] 14:109–120. Available from: http://dx.doi.org/10.1146/annurev.ge.14.120180.000545

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J, Bioconda Team. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* [Internet] 15:475–476. Available from: http://dx.doi.org/10.1038/s41592-018-0046-7

Hershberg R. 2016. Codon Usage and Translational Selection. In: Encyclopedia of Evolutionary Biology. Elsevier. p. 293–298. Available from: http://dx.doi.org/10.1016/b978-0-12-800049-6.00178-5

Hof AEV t., Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ, van't Hof AE, et al. 2016. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* [Internet] 534:102–105. Available from: http://dx.doi.org/10.1038/nature17951

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* [Internet] 44:D81–D89. Available from: http://dx.doi.org/10.1093/nar/gkv1272

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

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* [Internet] 9:411–412. Available from: http://dx.doi.org/10.1038/nrg2165-c1

916 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
917 Improvements in performance and usability. *Mol. Biol. Evol.* [Internet] 30:772–780.
918 Available from: http://dx.doi.org/10.1093/molbev/mst010

919 Kollmar M. 2019. Gene Prediction: Methods and Protocols. Humana Press Available from:
920 https://books.google.com/books/about/Gene_Prediction.html?hl=&id=iEkZvwEACAA
921 J

922 Kolpakov R, Bana G, Kucherov G. 2003. mreps: Efficient and flexible detection of tandem
923 repeats in DNA. *Nucleic Acids Res.* [Internet] 31:3672–3678. Available from:
924 http://dx.doi.org/10.1093/nar/gkg617

925 Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey
926 VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput.*
927 *Biol.* [Internet] 9:e1003118. Available from:
928 http://dx.doi.org/10.1371/journal.pcbi.1003118

929 Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R,
930 Edwards SV, Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project:
931 Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* [Internet]
932 115:4325–4333. Available from: http://dx.doi.org/10.1073/pnas.1720115115

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

935 Li Y, Nong W, Baril T, Yip HY, Swale T, Hayward A, Ferrier DEK, Hui JHL. 2020.
936 Reconstruction of ancient homeobox gene linkages inferred from a new high-quality
937 assembly of the Hong Kong oyster (Magallana hongkongensis) genome. *BMC*
938 *Genomics* 21:1–17.

939 McClintock B. 1956. “Controlling Elements and the Gene.” *Cold Spring Harb. Symp. Quant.*
940 *Biol.* 21:197–216.

941 Mérel V, Boulesteix M, Fablet M, Vieira C. 2020. Transposable elements in Drosophila. *Mob.*
942 *DNA* [Internet] 11:23. Available from: http://dx.doi.org/10.1186/s13100-020-00213-z

943 Mombaerts P. 1999. Seven-transmembrane proteins as odorant and chemosensory
944 receptors. *Science* [Internet] 286:707–711. Available from:
945 http://dx.doi.org/10.1126/science.286.5440.707

946 Ou S, Jiang N. 2019. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid
947 identification of long terminal repeat retrotransposons. *BioRxiv*:2–6.

948 Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D,
949 Peterson T, et al. 2019. Benchmarking transposable element annotation methods for
950 creation of a streamlined, comprehensive pipeline. *Genome Biol.* [Internet] 20:1–45.
951 Available from: http://dx.doi.org/10.1186/s13059-019-1905-y

952 Paradis E, Strimmer K, Claude J, Jobb G, Opgen-Rhein R, Dutheil J, Noel Y, Bolker B,
953 Lemon J. 2006. ape: Analyses of Phylogenetics and Evolution. *R package version*
954 [Internet] 1. Available from: http://ape-package.ird.fr/ep/diapo_LaReunion_2009.pdf

955 Peng X, Wilken SE, Lankiewicz TS, Gilmore SP, Brown JL, Henske JK, Swift CL, Salamov
956 A, Barry K, Grigoriev IV, et al. 2021. Genomic and functional analyses of fungal and
957 bacterial consortia that enable lignocellulose breakdown in goat gut microbiomes.
958 *Nat. Microbiol.* [Internet] 6:499–511. Available from: http://dx.doi.org/10.1038/s41564-

959 020-00861-0

960 Pickett BD, Karlinsey SM, Penrod CE, Cormier MJ, Ebbert MTW, Shiozawa DK, Whipple CJ,
961 Ridge PG. 2016. SA-SSR: a suffix array-based algorithm for exhaustive and efficient
962 SSR discovery in large genetic sequences: Table 1. *Bioinformatics* [Internet]
963 32:2707–2709. Available from: http://dx.doi.org/10.1093/bioinformatics/btw298

964 Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is
965 vital when analyzing new genome assemblies. *Genome Biol. Evol.* [Internet] 8:403–
966 410. Available from: http://dx.doi.org/10.1093/gbe/evw009

967 Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic
968 features. *Bioinformatics* [Internet] 26:841–842. Available from:
969 http://dx.doi.org/10.1093/bioinformatics/btq033

970 Racine JS. 2013. RSTUDIO: A PLATFORM-INDEPENDENT IDE FOR R AND SWEAVE. *J.*
971 *Appl. Econometrics* [Internet] 27:167–172. Available from:
972 http://dx.doi.org/10.1002/jae

973 Rodriguez M, Makałowski W. 2022. Software evaluation for de novo detection of
974 transposons. *Mob. DNA* [Internet] 13:14. Available from:
975 http://dx.doi.org/10.1186/s13100-022-00266-2

976 Smit AFA, Hubley RR, Green PR. 2013. RepeatMasker Open-4.0. *http://repeatmasker.org*.

977 Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of
978 transposable element families, sequence models, and genome annotations. *Mob.*
979 *DNA* [Internet] 12:2. Available from: http://dx.doi.org/10.1186/s13100-020-00230-y

980 Strelets VB, Schroeder AJ, Thurmond J, Goodman JL, Gelbart WM, dos Santos G, Emmert
981 DB, Crosby MA, the FlyBase Consortium. 2014. FlyBase: introduction of the
982 Drosophila melanogaster Release 6 reference genome assembly and large-scale
983 migration of genome annotations. *Nucleic Acids Res.* [Internet] 43:D690–D697.
984 Available from: http://dx.doi.org/10.1093/nar/gku1099

985 Team RC. 2013. R: A language and environment for statistical computing.

986 Wells JN, Feschotte C. 2020. A Field Guide to Eukaryotic Transposable Elements. *Annu.*
987 *Rev. Genet.* [Internet] 54:539–561. Available from: http://dx.doi.org/10.1146/annurev-
988 genet-040620-022145

989 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,
990 Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic
991 transposable elements. *Nat. Rev. Genet.* [Internet] 8:973–982. Available from:
992 http://dx.doi.org/10.1038/nrg2165

993 Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes
994 A, Henry L, Hester J. 2019. Welcome to the Tidyverse. *Journal of Open Source*
995 *Software* 4:1686.

996 Wong WY, Simakov O. 2018. RepeatCraft: a meta-pipeline for repetitive element de-
997 fragmentation and annotation. *Bioinformatics* [Internet] 35:1051–1052. Available
998 from: https://doi.org/10.1093/bioinformatics/bty745

999 Xu X-H, Su Z-Z, Wang C, Kubicek CP, Feng X-X, Mao L-J, Wang J-Y, Chen C, Lin F-C,
1000 Zhang C-L. 2014. The rice endophyte Harpophora oryzae genome reveals evolution

1001 from a pathogen to a mutualistic endophyte. *Sci. Rep.* [Internet] 4:5783. Available
1002 from: http://dx.doi.org/10.1038/srep05783

1003 Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR
1004 retrotransposons. *Nucleic Acids Res.* [Internet] 35:W265–W268. Available from:
1005 https://www.ncbi.nlm.nih.gov/pubmed/17485477

1006 Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA
1007 sequences. *J. Comput. Biol.* [Internet] 7:203–214. Available from:
1008 http://dx.doi.org/10.1089/10665270050081478

1009