TransposonUltimate: software for transposon classification, annotation and detection

Kevin Riehl[1], Cristian Riccio[1,2], Eric A. Miska[1,2,3] and Martin Hemberg[2,4,*]

[1]Gurdon Institute, University of Cambridge, Cambridge CB2 1QN, UK, [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK, [3]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK and [4]Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02215, USA

Received June 09, 2021; Revised February 09, 2022; Editorial Decision February 14, 2022; Accepted February 14, 2022

ABSTRACT

Most genomes harbor a large number of transposons, and they play an important role in evolution and gene regulation. They are also of interest to clinicians as they are involved in several diseases, including cancer and neurodegeneration. Although several methods for transposon identification are available, they are often highly specialised towards specific tasks or classes of transposons, and they lack common standards such as a unified taxonomy scheme and output file format. We present TransposonUltimate, a powerful bundle of three modules for transposon classification, annotation, and detection of transposition events. TransposonUltimate comes as a Conda package under the GPL-3.0 licence, is well documented and it is easy to install through https://github.com/DerKevinRiehl/TransposonUltimate. We benchmark the classification module on the large *TransposonDB* covering 891,051 sequences to demonstrate that it outperforms the currently best existing solutions. The annotation and detection modules combine sixteen existing softwares, and we illustrate its use by annotating *Caenorhabditis elegans*, *Rhizophagus irregularis* and *Oryza sativa subs. japonica* genomes. Finally, we use the detection module to discover 29 554 transposition events in the genomes of 20 wild type strains of *C. elegans*. Databases, assemblies, annotations and further findings can be downloaded from (https://doi.org/10.5281/zenodo.5518085).

INTRODUCTION

Transposons are evolutionary ancient mobile genetic elements that can move via copy&paste and cut&paste transposition mechanisms. They can be classified within a taxonomic scheme (Figure 1A), and each class is associated with a set of characteristics, e.g. proteins relevant for transposition and structural features (Figure 1B). During transposition, transposable elements (TEs) can leave structural patterns both at the insertion and the deletion site (1–3). Autonomous transposons encode the tools necessary for transposition events, e.g. genes producing transposase, integrase and other enzymes (3), while non-autonomous transposons depend on proteins encoded elsewhere (4). As the insertion of a transposon can be detrimental, many species have developed repression mechanisms, e.g. TE promoter methylation (5) and piRNAs (6). Even though transposition events occur rarely (7), in many organisms large sections of DNA consist of either transposons or their transposition-incompetent descendants that have accumulated mutations over time (4). It is estimated that transposons make up a large share of the genome in many species; 45% in humans, 20% in fruit flies, 40% in mice, 77% in frogs and 85% in maize (8).

Studying TEs is highly relevant for understanding evolutionary processes (9), developmental biology, gene regulation, and many diseases are suspected to be related to transposon activity such as subtypes of haemophilia, immunodeficiency, cancer and Alzheimer's disease (10–12). Also, TEs are popular for genetic engineering purposes as they allow for direct insertion of their genetic cargo into a target genome (13–15). However, the repetitive nature of transposons and their descendants is a challenge for their analysis and discovery, in particular when using short-read sequencing technologies (7). Long-read technologies facilitate studies of transposons and their functional consequences, but they also require novel computational tools. Although various approaches for identifying transposons have been proposed recently (16), current tools do not provide the flexibility to combine, filter and order annotated elements on a unifying scale, and are often limited to a family of transposons or a group of species (17).

Here, we present a bundle of tools addressing three different tasks related to transposon identification: classification, annotation and detection. The goal of classification is to determine which taxonomic class a given transposon

*To whom correspondence should be addressed. Tel: +1 857 307 1422; Fax: +1 617 525 5566; Email: mhemberg@bwh.harvard.edu

Proposed transposon classification taxonomy

Class	Taxonomy / hierarchy	Notes on constituents
1	Class I (Retrotransposons)	
1/1	LTR	Copia, Gypsy, Bel-Pao, Retrovirus, ERV
1/1/1	Copia	Copia
1/1/2	Gypsy	Gypsy
1/1/3	ERV	ERV
1/2	Non-LTR	DIRs (VIPER, Ngaro), LINEs, SINEs
1/2/1	LINE	R2, RTE, Jockey, L1, I, Randl, Penelope, DRE…[1]
1/2/2	SINE	tRNA, 7SL, 5S
2	Class II (DNA transposons)	TIR, Crypton, Helitron, Maverick / Polinton, MITEs
2/1	TIR	Tc1-Mariner, hAT, Mutator, Merlin…[2]
2/1/1	Tc1-Mariner	Tc1-Mariner
2/1/2	hAT	hAT
2/1/3	CMC	CMC
2/1/4	Sola	Sola
2/1/5	Zator	Zator
2/1/6	Novosib	Novosib
2/2	Helitron	Helitron
2/3	MITEs	Tourist, Stowaway

[1] R2 (CRE, R4, Hero, NeSL, R2), RTE (RTETP, Proto2, RTEX, RTE), Jockey (Rex1, CR1, L2, L2A, L2B, Daphne, Crack), L1 (Proto1, Tx1), I (Ingi, Nimb, Tad1, Loa, R1), Randl, Penelope, DRE

[2] Tc1-Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF-Harbinger, CACTA / ENSPM / Chapaev, MuLE / MUDR, CMC, Sola, Ginger, Academ, Dada, Kolobok, Zator, Novosib

B Transposon structure overview

Structural Features	Class I			Class II		
	LTR	LINE	SINE	TIR	HEL	MITE
Target site duplication (TSD)	x	x	x	x		x
Terminal inverted repeat (TIR)				x		x
Long terminal repeat (LTR)	x					
Primer binding site (PBS)	x					
Polypurine tract (PPT)	x					
Begin A-TC					x	
End CTRR-T					x	
Open reading frames (ORF)	x	x		x		x
Palindromic sequence (hairpin loop)					(x)	
Poly(A) tail		x	x			
Protein Features						
Helicase					(x)	
Capsid protein (GAG)	x					
RPA-like (RAPI) replication protein					(x)	
Envelope (ENV)	(x)					
Transposase				x		
Endonuclease		(x)				
Nucleic acid binding protein (NABP)		x				
Aspartic proteinase (AP)	x					
Apurinic endonuclease (AE)		(x)	(x)			
Pol gene (pol)	x					
Protease (PR)	x					
Integrase (INT)	x					
Reverse transcriptase (RT)	x	(x)				
RnaseH (RH)	x					

Figure 1. Transposon taxonomy and transposon structure. (**A**) The taxonomy used in this study is based on multiple classification schemes (3,36,49,106) and the taxonomies used by the transposon databases. (**B**) Autonomous, transposition competent transposons have characteristic structural and protein features depending on their class. The proteins are necessary for the transposons to move via class-specific transposition mechanisms. The x mark which structural and protein features are characteristic to different transposon classes and sub classes for complete, autonomous transposons. The (x) mark features that are not required but if present are indicative.

sequence belongs to. The annotation task consists of scanning a genome sequence to mark all transposons. Finally, the detection task involves the comparison of two genomes to identify structural variants arising from the insertion of TEs.

Existing transposon classifiers are difficult to compare directly since they vary in their approach, which features and taxonomies they use, how they evaluate predictions, and which databases are used for training. Applications of SVMs (18), hidden Markov models (19), random forests (20), Gaussian naive Bayes (21), decision trees (22), stacking (23,24), boosting (25,26), neural networks (27–29), evolutionary algorithms (21,30) and genetic algorithms (31–34) can be found in the literature. Most methods use sequence features, such as the k-mer frequency, the occurrence of structural (35) and protein features (18) for classification. Besides, another approach is to classify TEs using the similarity to known transposons based on a sequence library (36).

The annotation of transposons in nucleotide sequences is challenging due to the presence of transposition-incompetent TEs that have been mutated, truncated, degraded, fragmented and dismembered due to nesting (37). Annotation is further complicated by a lack of standards (38) and disagreement on definition, taxonomy and terminology (39,40). Since transposons do not adhere to a universal structure (41), many researchers have employed class-specific approaches (42). Moreover, most of the software employed for transposon annotation was originally designed for gene annotation, neglecting the peculiarities of transposons (39). Existing transposon annotation methods (Table 1) can be assigned to one or more approaches (1,2,41,43). The *de novo* approach finds transposons by identifying repetitive sequences. It is effective in discovering previously unknown transposons with high prevalence (41), but it is computationally costly (39,41), unable to find degraded transposons (41), and risks misidentifying repetitive DNA or high copy number genes as transposons (44,45). The structure-based approach (also called motif-based (42) or signature-based approach (2)) is based on knowledge of the structure of transposons and annotates by finding combinations of characteristic patterns (38,46). This approach enables the discovery of transposition-incompetent transposons thanks to their unique structural properties (41). However, these approaches are often characterized by high false discovery rates (37,44) and they miss transposons with weak signatures (37). The similarity-based approach (also called library-based approach (2)) employs a library of known transposons together with BLAST(-like) tools. The high accuracy (41) and short runtimes (44,47) of this approach come at the cost of its inability to find unrelated transposons (41,47) and the dependency on quality and exhaustiveness of the library (38,44,48). Moreover, the current version of the most widely used database RepBase (49) is behind a paywall and the related tools RepeatMasker and RepeatModeler are not transparent with regards to how transposons were curated and consensus sequences were generated (39).

Previous efforts to detect transposition events by comparing two genomes have been based on the analysis of the depth of coverage, discordant and split read pairs (50,51). However, both the task of detecting structural variants (SVs) and annotating TEs are very challenging when using short reads (7). Recently, long-reads technologies have become more widely available, but to the best of our knowledge the only existing method that can take advantage of them for TE detection is LoRTE (52). Although results indicate that LoRTE performs well even on low coverage reads, it is limited to PacBio data and insertion and deletion SVs only.

Here, we present TransposonUltimate, a set of tools for the identification of transposons, consisting of three modules for accurate classification, annotation in nucleotide sequences and detection of transposition events (Figure 2). Our new classifier is benchmarked against existing softwares, and we use the annotation module to analyse the genomes of three different species. Finally, the detection module is employed to identify transposition events in 20 high quality genomes from *Caenorhabditis elegans* wild isolates that were assembled using a combination of long- and short-read technologies.

MATERIALS AND METHODS

Transposon classification module, RFSB

Given a nucleotide sequence that is considered to be a transposon, the goal is to determine the class of a transposon according to a given taxonomy. This task is a hierarchical classification problem, meaning the classifier needs to identify multiple classes that stand in a relationship described by a taxonomic hierarchy. The design of the classification module includes several aspects; choosing a transposon database for training and testing, feature selection, model structure, training strategy, model implementation, evaluation and benchmarking.

The classifiers considered here are supervised learning algorithms, and consequently their performance is limited by the data used for training. Previous studies used small transposon sequence databases, each with different taxonomic schemes, which does not allow for a direct comparison. Therefore, we created *TransposonDB* (Figure 3, File F1), a large collection of transposon sequences that consists of ten databases: ConTEdb (53) (http://genedenovoweb.ticp.net:81/conTEdb/index.php), DPTEdb (54) (http://genedenovoweb.ticp.net:81/DPTEdb/browse.php?species=cpa&name=Carica_papaya_L.), mipsREdat-PGSB (55) (https://pgsb.helmholtz-muenchen.de/plant/recat/index.jsp), MnTEdb (56) (http://genedenovoweb.ticp.net:81/MnTEdb1/), PMITEdb (57) (http://pmite.hzau.edu.cn/download_mite/), RepBase (58) (https://www.girinst.org/repbase/, we use version 23.08 that was the last publicly available version before the paywall was introduced), RiTE (59) (https://www.genome.arizona.edu/cgi-bin/rite/index.cgi), Soyetedb (60) (https://www.soybase.org/soytedb/#bulk), SPTEDdb (61) (http://genedenovoweb.ticp.net:81/SPTEdb/browse.php?species=ptr&name=Populus_trichocarpa) and TrepDB (62) (http://botserv2.uzh.ch/kelldata/trep-db/downloadFiles.html). To create the database, the taxonomies were unified, duplicates were dropped and several filter rules were applied (Supplementary Table S1). Filtering

Table 1. Overview of common transposon annotation tools

		Approach			Class I			Class II		
Name		Novo.	Struc.	Simil.	LTR	LINE	SINE	TIR	HEL	MITE
RepeatMasker		x		x	x	x	x	x	x	x
RepeatModeler		x			x	x	x	x	x	x
CLARI_TE	(107)	x	x	x	x	x	x	x	x	x
TESeeker	(41)			x	x	x	x	x	x	x
PILER	(40)	x			x	x	x	x	x	x
Censor	(108)	x			x	x	x	x	x	x
RepLong	(109)	x			x	x	x	x	x	x
EDTA	(44)	x	x	x	x	x	x	x	x	x
MGEScan	(110)	x	x	x	x	x	x			
LTR_Finder	(111)		x		x					
LtrDetector	(112)		x		x					
LTRpred	(73)	x	x	x	x					
LTRharvest	(66)	x	x	x	x					
LTRdigest	(113)		x		x					
SINE-Finder	(68)	x	x				x			
SINE-Scan	(69)	x	x				x			
TIRvish	(67)		x					x		
HelitronScanner	(42)		x						x	
MUSTv2	(70)		x							x
MiteFinderII	(71)		x							x
MITE-Tracker	(72)		x							x
detectMITE	(45)		x							x
MITE-Hunter	(47)		x							x

The most commonly used tools such as RepeatMasker and RepeatModeler cover a variety of transposons, while others focus on certain classes only. The tools use one or more of the *de novo*, structural and similarity-based transposon annotation approaches.

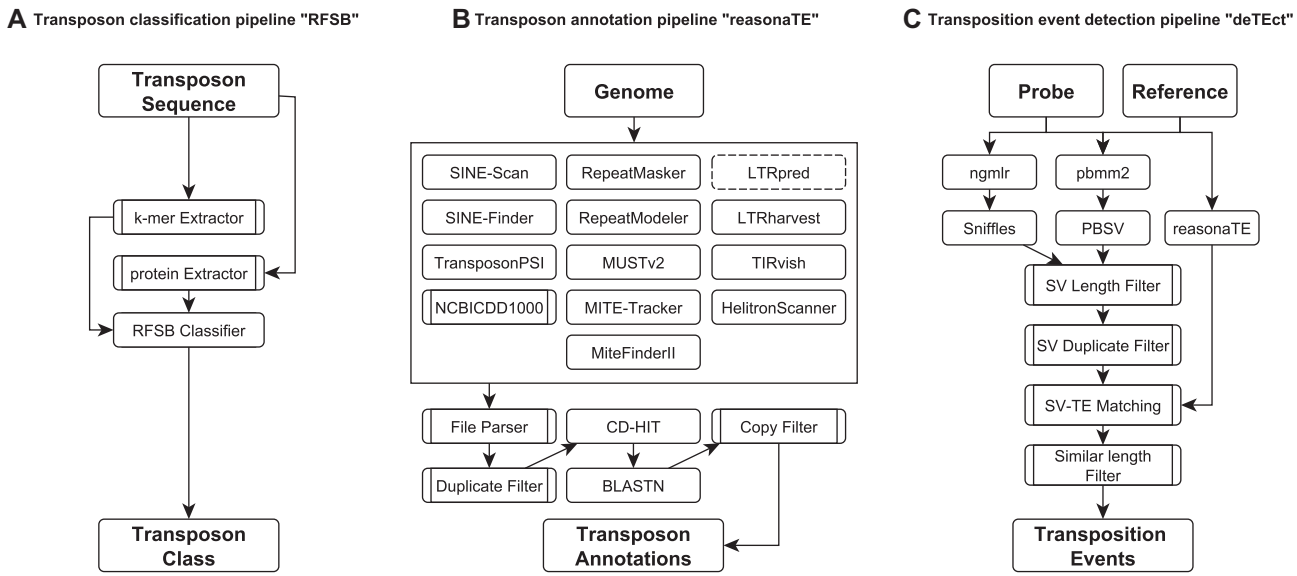

Figure 2. Three pipelines of the TransposonUltimate framework. (**A**) Given the nucleotide sequence of a transposon, relative *k*-mer frequencies (for $k = 2, 3, 4$) and binary protein features are extracted. These features are used by the random forest selective binary classifier (RFSB) to infer the transposon's class. (**B**) Published transposon and protein annotation tools are applied to a given genome. Resulting annotations are filtered, merged and clustered using CD-HIT. Then, BLASTN is used to find additional full-length copies. (**C**) Sequencing reads obtained using a long-read technology from a probe genome are aligned onto a reference genome using ngmlr and pbmm2. Next, the alignments are used to discover structural variants. After filtering the structural variants, they are matched to the transposon annotations to detect transposition events.

included the removal of sequences with no label, the exclusion of fragments, contigs, satellites and RNA sequences. Moreover, only sequences with a length >100 bp and those including at least once each of the letters 'A','C','G' and 'T' were kept. To the best of our knowledge, this is the largest database of transposon sequences available. Since TransposonDB covers all relevant Eukaryotic kingdoms, it allows for the training and evaluation of a robust, cross-species hierarchical classification model (Supplementary Tables S2 and S3). Moreover, the database is balanced and covers sufficient examples for all taxonomic nodes (Supplementary Table S4). However, TransposonDB is still likely to be biased as most of the TEs are from plant genomes.

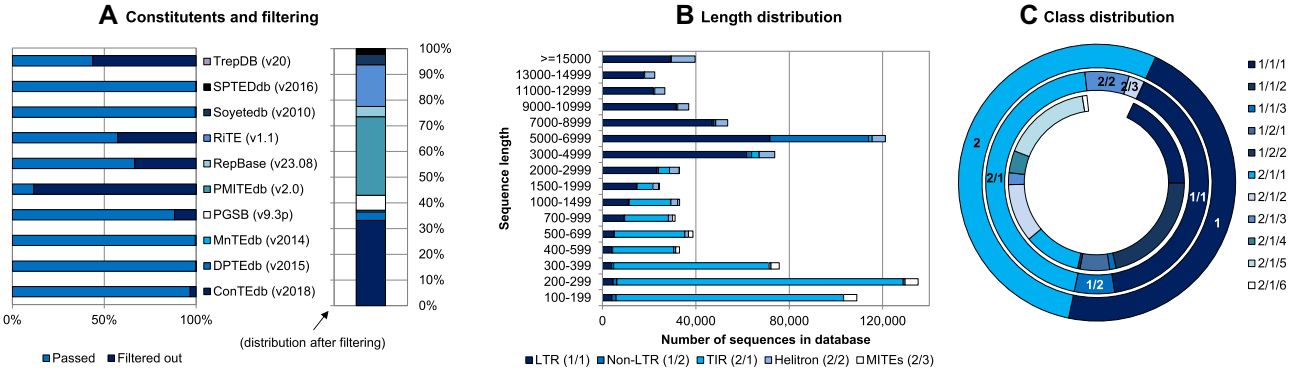

Figure 3. Summary statistics for the TransposonDB. (**A**) Ten publicly available transposon databases were filtered and combined. Sequences with no (valid) class label, fragments, contigs, satellites, RNA, shorter than 100 bp were filtered out. Moreover, duplicates were dropped when merging. Taxonomic schemes by different databases were unified. (**B**) The length distribution of sequences in the databases reveals that most DNA transposons are shorter than 500 bp, while most retrotransposons are longer than 3,000 bp. However, Helitrons are significantly longer than other DNA transposons. (**C**) TransposonDB is balanced in terms of class occurrence, although ERV (1/1/3), SINE (1/2/2) and Novosib (2/1/6) transposons occur rarely.

We selected the combination of relative *k*-mer frequencies and binary protein features for our classifier. Relative *k*-mer frequencies represent the number of occurrences of a *k*-mer within a sequence divided by the number of times it would appear if the sequence consisted of this *k*-mer only. Protein features are binary, indicating the presence of a certain protein domain in the sequence. The feature vector consists of *k*-mer frequencies ($k = 2, 3, 4$) and 169 selected domains from NCBI CDD (63) covering class-specific transposons (Supplementary Table S5). We used the 169 domains as query sequences for RPSTBLASTN (v2.10.1) to annotate the conserved domain models at an *e*-value of 5.0 as it performed best in terms of classification performance (Supplementary Figure S1A, B). In addition, two model structures were explored. The binary structure employs binary classifiers for each node (= transposon class) of the taxonomy, and it assigns a probability for each sequence to belong to the node in question. For each internal node, the child is chosen as the node associated with the highest probability. The multilabel structure employs a multilabel classifier for each parent node of the taxonomy with $n + 1$ classes representing the taxonomic child classes and -1 (return scenario). After inference, the taxonomic class can be determined by choosing the most probable child node at each stage or to return to a higher level and then choose the second most probable child node at that stage. Moreover, we explored two training strategies. The *all* training strategy trains each classification node with the whole training set, while the *selective* training strategy trains each classification child node with a training set that was activated by the parent node. All training strategies, model structures and feature generation were implemented in Python (v3.6.9). Models implementing random forests, AdaBoost, logistic regression, SVM and Naive Bayes from the machine learning package scikit-learn (v0.23) (64) were explored. Random forest consistently yields the highest classification performance (Supplementary Figure S2). Based on these results, we propose a **r**andom **f**orest classifier with a **s**elective training strategy on a **b**inary model structure, *RFSB*.

Previous transposon classification studies use different performance measures, taxonomies, training and testing sets, making it hard to compare them. To evaluate the performance, we consider three perspectives. The first perspective is based on hierarchical precision and recall, meaning it considers the whole taxonomy, as proposed in (65). The second perspective evaluates for different taxonomic levels and the third perspective captures the classification performance of single classes. We benchmark RFSB againts TERL (29), TopDown (24), NLLCPN (27), HC_LGA (33) and HC_GA (31), as their published code allowed for reproduction. To ensure a fair comparison, source codes were partially modified to allow the training and evaluation of these models on the taxonomy used in our work and TransposonDB and can be found on Github https://github.com/DerKevinRiehl/transposon_classifier_rfsb/blob/main/benchmark/ClassifierCode.rar.

Transposon annotation module, reasonaTE

Given an assembled genome, the goal of the annotation module is to find all transposon occurrences and their locations. Our *reasonaTE* pipeline produces rich annotations, including transposon mask regions (union of all annotated base pairs) as well as transposon annotations, classification, structural and protein features. This is achieved by combining the advantages of thirteen published transposon annotation tools covering different annotation approaches and transposon classes: RepeatMasker v2.0.1 (http://www.repeatmasker.org/), RepeatModeler v4.1.1 (http://www.repeatmasker.org/RepeatModeler/), LTRharvest (66) (https://www.zbh.uni-hamburg.de/forschung/gi/software/ltrharvest.html) and TIRvish (67) (http://genometools.org/tools/gt_tirvish.html) are available as Conda packages. Moreover, we created Conda packages for SINE-Finder (68) (http://www.plantcell.org/content/suppl/2011/08/29/tpc.111.088682.DC1/Supplemental_Data_Set_1-sine_finder.txt), SINE-Scan (69) (https://github.com/maohlzj/SINE_Scan), HelitronScanner (42) (https://sourceforge.net/projects/helitronscanner/files/), MUSTv2 (70) (http://www.healthinformaticslab.org/supp/resources.php), MiteFinderII (71) (https://github.com/jhu99/miteFinder) and MITE-Tracker (72) (https://github.com/INTABiotechMJ/MITE-Tracker)

to make them accessible and to facilitate their installation. Also, we include the output files of LTRpred (73) (https://hajkd.github.io/LTRpred/articles/Introduction.html) into the pipeline, as this tool provides high quality annotations, but is available as a Docker image only. As the tools have different output formats, we developed a parser module to convert all outputs to GFF3 format.

After running the annotation tools, additional copies of the identified transposons are searched using the clustering tool CD-HIT (v4.8.1) (74,75) at an identity threshold of 0.9 and BLASTN (v2.10.1) at an e-value of 0.1. If not mentioned further, we used the standard settings for all other parameters of these tools. For the annotation of transposon-characteristic proteins, we have created a Conda packaged version of TransposonPSI (http://transposonpsi.sourceforge.net/), and we also use the protein domains from NCBI CDD for this task. Using TransposonDB, NCBI CDD and RPSTBLASTN, we selected the 1,000 most frequently occurring protein domains that are characteristic to transposons (File F2). As an application, here we annotate the genome *MSU7* of *Oryza sativa* subspecies *japonica* (http://rice.plantbiology.msu.edu/index.shtml), the genome *DAOM197198* of *Rhizophagus irregularis* (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJDB4945) (76), three reference genomes *VC2010* (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB28388), *N2* (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA13758), *CB4856* (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA275000) and 20 novel wild type strains (77) of *Caenorhabiditis elegans* (Supplementary Table S6).

Transposition event detection module, deTEct

Given an assembled reference genome and sequenced probe genome reads, the goal is to identify transposition events that are manifested as structural variants. This requires both a list of SVs and annotation of TEs as inputs. We employ the structural variant caller Sniffles on ngmlr (78) alignments and PBSV (https://github.com/PacificBiosciences/pbsv) structural variant caller on pbmm2 alignments of PacBio reads (https://github.com/PacificBiosciences/pbmm2). Moreover, the TE annotations are generated using the proposed reasonaTE pipeline mentioned before.

SVs are filtered twice. First, variants shorter than 50 bp or longer than 1% of the genome length were excluded. Second, duplicate structural variants of the same type are merged. Consecutively, the remaining variants and TE annotations are matched and reported if their length corresponds to each other. Transposon annotations were matched to structural variants if they intersected for at least 10% and their length was similar by a threshold of 50%. We chose to do so as structural variant callers and transposon annotators have an uncertainty regarding exact locations. We therefore consider a similar length more important than a high overlap. The proposed deTEct pipeline is applicable to long-read sequencing technologies, and it has been tested with PacBio data. It has not been tested for short reads and thus we advice against using the pipeline for this type of data.

RESULTS

RFSB outperforms other transposon classifiers

We benchmarked our RFSB method against other transposon classifiers, and the results show that it has the highest sensitivity and specificity (Figure 4A, Supplementary Table S7). TE Learner (20) has the lowest reported performance, while the other methods have similar *F*1 scores. However, this comparison is based on reported numbers from different studies with different evaluation schemes, taxonomies and datasets for training and testing. For a more fair comparison some of the tools were applied to the subset of TranspsonDB which includes RepBase and PGSB (Figure 4B). The comparison of the results reveals large discrepancies. Surprisingly, TERL and TopDown have a performance which is worse than random guessing, and closer inspection of the outputs from NLLCPN reveals that it has learned a constant distribution rather than a relationship between sequences and classes.

A detailed analysis of the classification performance of RFSB across different taxonomic levels and classes reveals a small decrease in performance when considering deeper taxonomic levels (Figure 4C). Underrepresented classes, e.g. Helitrons and MITEs, perform worse, and the results are consistent for both F1 and MCC scores. Moreover, for some classes the performance of RFSB on the large, cross-species TransposonDB is better than for the more homogeneous subset of RepBase and PGSB, which suggests that it is robust, generalisable, and applicable to different species. An inspection of the most informative features (File F3) shows that long *k*-mer features contribute the most to the classification performance, while protein domains have a smaller share amongst the most contributing features (Figure 4D). This motivated the exploration of longer *k*-mer features, but we did not find any significant increase of the performance when using 5-mers (Figure 4E). We also evaluated the runtime (All computations were executed on the cluster CB-GPU1 of the Gurdon institute (OS Ubuntu v18.04.4 LTS). The cluster consists of 80 Intel(R) Xeon(R) Gold 6148 CPUs (2.40 GHz), 315 GB CPU-RAM, two GeForce RTX 2080 GPUs (each 60T RTX-OPS) and 16 GB GPU-RAM.) of the different classifiers, and the results show that the superior classification performance of RFSB comes at a cost of it taking almost twice as long to run as the other methods (Supplementary Tables S10 and S11).

The ensemble strategy reasonaTE finds more transposons

Next, we evaluated the ability of our reasonaTE pipeline to identify TEs in the genomes of three different species (Figure 5A, B, File F4). The TE content of almost 21% for *C. elegans* is higher than previously reported values of 12% (8), 17% (79) and 12−16% (80). However, as these studies used methods that were biased towards finding specific classes of transposons, it is to be expected that our ensemble strategy finds more TEs. By contrast, the prediction of 33% for *O. sativa* ssp. *japonica* is very close to the mean of other reports (81–89). The content of 23% in *Rhizophagus irregularis* is close to a previous estimate of 27% (90). The low variation of transposon content across different strains becomes obvious for the cluster of *C. elegans*. Interestingly,

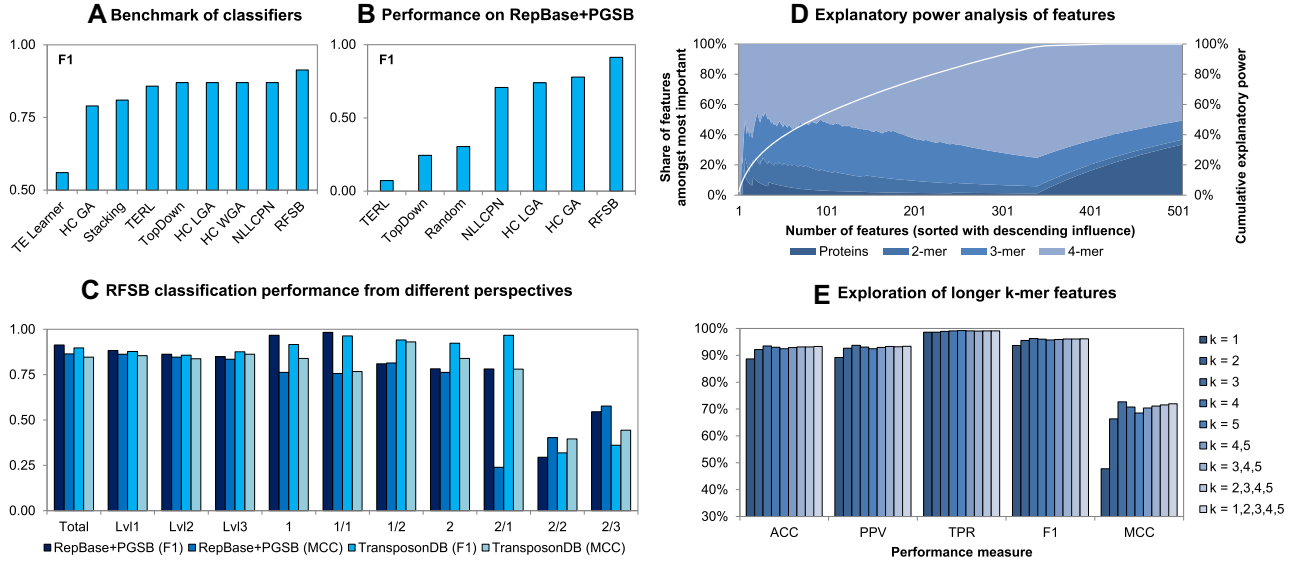

Figure 4. Evaluation of the RFSB classifier. (**A**) Benchmark of different transposon classifiers by reported numbers in publications. (**B**) The performance of selected, reproducable classifiers applied on RepBase+PGSB database using the taxonomy in Figure 1A. Reported numbers represent performances from a total perspective. (**C**) RFSB classification performances from total, taxonomic level and class perspective. (**D**) Analysis of each feature's contribution to classifier's explanatory power. The white line shows the cumulative explanatory power. (**E**) Analysis of different *k*-mer features in combination with protein features for a binary classifier differentiating between class 1 and 2 transposons. All values presented were calculated as average across a 10-fold cross validation.

the relative transposon class frequency reveals clear differences across species (Figure 5C, D). Similarly, the length distributions (Figure 5E–G) exhibit substantial differences between transposons of the same class found in different species. Helitrons in particular vary in length as was observed before (91).

In concordance with (92,93), the share of Helitrons amounts to almost 2% of the *C. elegans* genome. Moreover, the majority of the transposons are TIR DNA transposons, as reported by (79,94,95). Contrary to previous studies (80,96,97), we mainly find hAT, CMC and Novosib transposons to be present in the *C. elegans* genome rather than Tc1-Mariner transposons. Our findings for the rice genome are consistent with previous findings. The high frequency of Gypsy (class 1/1/2) compared to other LTR (class 1/1) and non-LTR (class 1/2) was reported in *Oryza sativa subs. japonica* (87). Moreover, the small share of MITEs, up to 2%, is similar to the previously reported share of 4% (89). A previous study (44) found that class 1 transposons have a larger share (25%) than class 2 transposons (20%) and the frequencies for the subclass level (LTR 23.5% and non-LTR 2%, TIR 17.5% and Helitrons 3.6%) match our findings. Inspection of the annotation density across the chromosomes revealed a characteristic concentration at the arms for *C. elegans* (Supplementary Figure S3), consistent with the higher densities observed for other variants (79,80,98–101).

The comparison of different annotation tools reveals that reasonaTE finds more TEs (Supplementary Figure S4) as none of the other methods finds more than 31.8% of the TEs reported by reasonaTE. In addition, the analysis shows that around 40% of the repetitive elements found by RepeatMasker and RepeatModeler were confirmed as transposons using our approach. Moreover, the transposon characteristic protein annotations by TransposonPSI and the 1000 most frequently occurring proteins from NCBI CDD intersect significantly with reasonaTE's transposon annotations. The analysis also reveals large overlap between some tools, e.g. MUSTv2 & MITE-Tracker, LTRpred & LTRharvest and SINE-Finder with all other tools.

Closer inspection of the class composition of the TEs found for *Caenorhabditis elegans* confirms the advantages of the ensemble technique of reasonaTE (Supplementary Figure S5). None of the tools is able to find the same share of TEs on its own as the ensemble. Moreover, we find that tools that were designed to identify a specific transposon class annotate TEs from other classes as well.

The runtime of reasonaTE depends on many different factors, including the number of TEs, their distribution, and the size of the genome. In general, the runtime is proportional to the size of the genome, but as we have only examined three different organisms we cannot extrapolate how runtimes will scale. For the investigated genomes in this study, the annotation tasks took around 6 days in the given cluster environment setup. RepeatMasker and RepeatModeler made up the largest share of runtime, the actual post-processing of reasonaTE did not exceed 10% of the total runtime.

29 554 transposition event candidates were observed analyzing 20 wild type strains of *Caenorhabditis elegans* using deTEct

Finally, we applied the deTEct pipeline to 20 whole genome assemblies of wild type strains of the nematode *C. elegans*. Each strain was compared to the two reference genomes *VC2010* and *CB4856* (Figure 6A, Supplementary Table S8, File F5). As expected, the newly sequenced genomes of

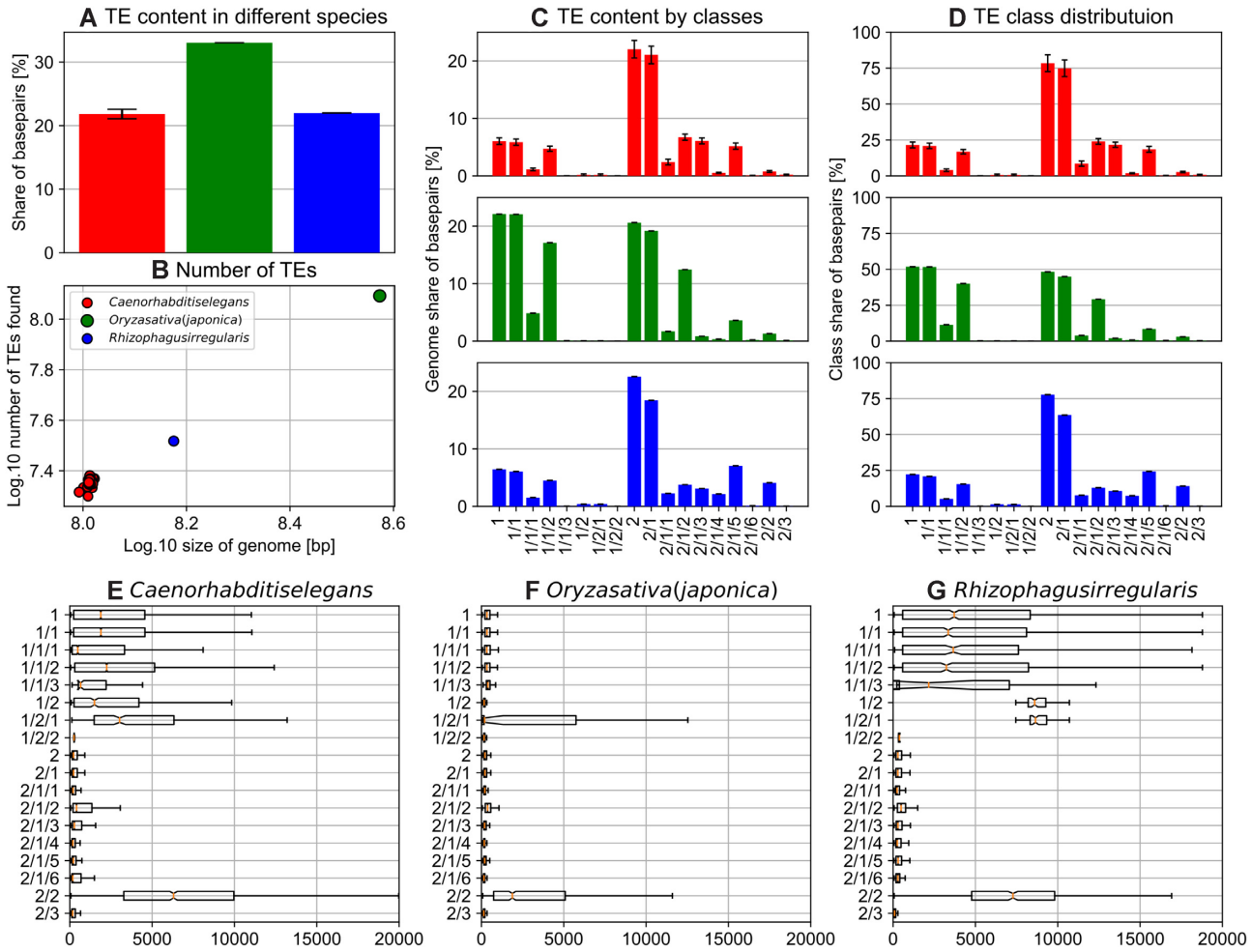

Figure 5. reasonaTE results for three species. The colors used in this figure represent *Caenorhabditis elegans* (red), *Oryza sativa subs. japonica* (green) and *Rhizophagus irregularis* (blue). (**A**) The average TE content of different species. The TE content is calculated as ratio of the sum of all basepairs part of the transposon region mask and the total genome size. The whiskers represent standard deviations. (**B**) The dot size represents the TE content as reported in the first panel, and the figure shows a linear relationship between genome size and the total number of transposons found. (**C**) Average TE content by transposon classes. The values were calculated by dividing the sum of the lengths of all transposons of a specific class by the total genome length. The whiskers represent the standard deviation. (**D**) The class distribution across all TEs based on the number of elements. (E–G) The transposon length distribution by classes for the three species. The boxes cover 25–75% percentiles, including the orange bar at the 50% percentile. The length of whiskers amounts to 150% of the interquartile range.

these two strains have almost no transposition events when compared to their reference. Closer inspection of the transposon and transposition event densities reveals that the putative transposition events are primarily located at the ends of the chromosomes (Figure 6B) as reported by (79). From the initial list of SVs, 3.97% were identified as transposition events. However, the list included numerous duplicates or very short variants that were subsequently filtered out. Consequently, we find that after filtering, 7.37% of all SVs are caused by transposition events.

Most of the transposition events were observed due to deletions (60%) while insertions, duplications and inversions cause the remaining variation (File F6 + F7). One difficulty in interpreting these proportions stems from the known biases of sequencing data (102) which make insertions hard to detect. This results in an elevated number of observations of cut transpositions (deletions), but fewer paste transpositions (insertions). Nonetheless, we find certain classes of transposons to be especially active in the comparisons of probe and reference genomes, such as Helitrons and SINEs relative to *VC2010*, and LINEs and Novosib when compared to *CB4856* (Figure 6D, File F8). The activity of Helitrons was observed previously (92,93). Helitrons were implicated in the divergence of GPCR genes and heat shock elements. Moreover, they are considered to play an important role in evolution (42). Comparing the two major classes, we conclude that the biggest contribution stems from DNA transposons (82% for VC2010 comparisons and 95% for CB4856 comparisons), similar to the findings in (103).

Moreover, we observe a linear relationship between the number of transposition events found and the phylogenetic distance of the given strains (Figure 6E-F). This result can be observed consistently for PacBio data (Supplementary Table S9). The strains *QX1211* and *ECA36* have the largest differences based on transposon data before (80). Although

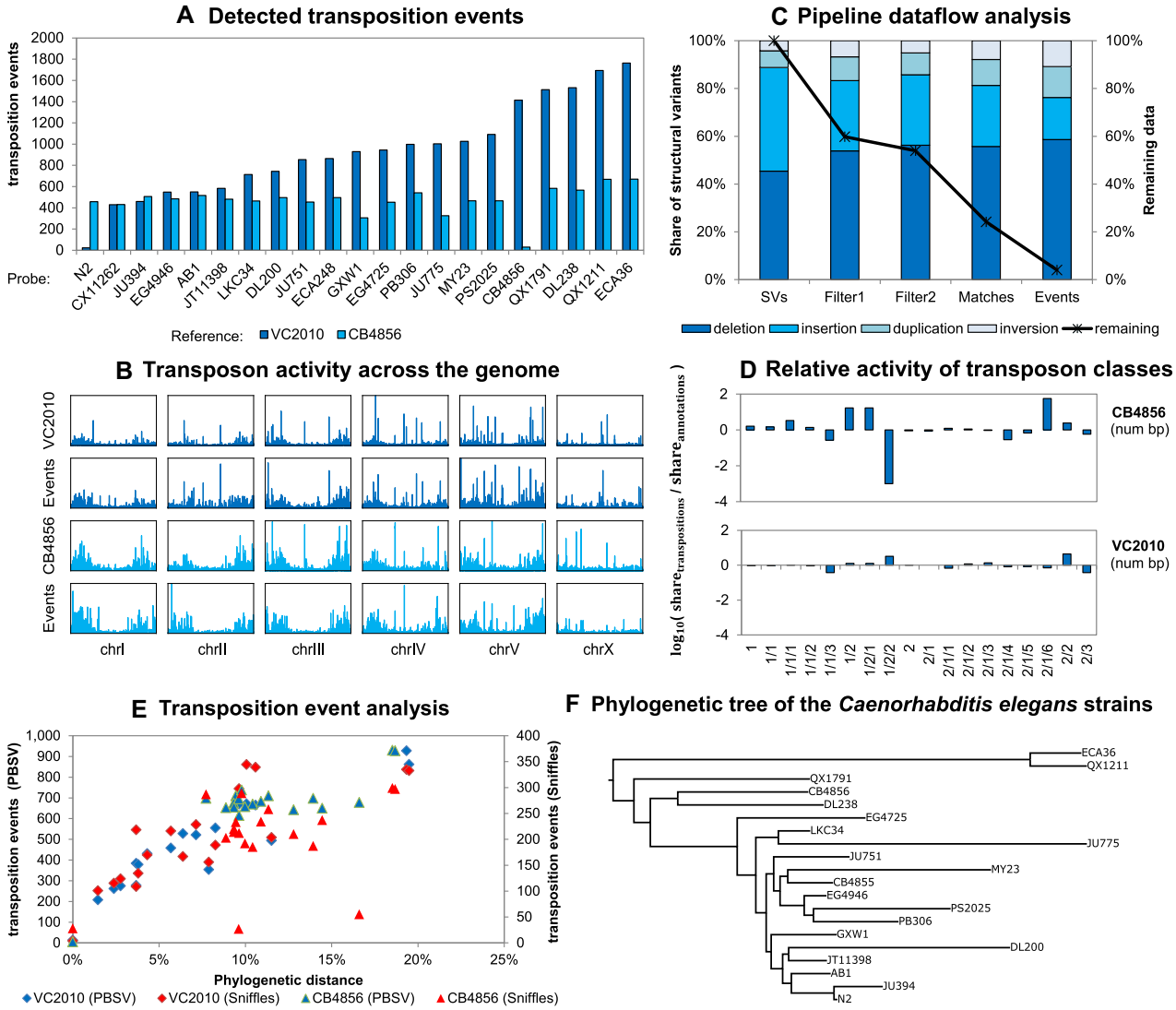

Figure 6. deTEct results and discovered transposition events. (**A**) Results show the number of detected transposition event candidates by probe strain for both reference genomes *VC2010* and *CB4856*. (**B**) The transposon activity in the *Caenorhabditis elegans* genome by chromosomes. The first row shows the density of transposon annotations in *VC2010*. The second row shows the density of transposition events. The following two rows represent results for *CB4856*. For all autosomal chromosomes we identify a characteristic pattern of transposon activity at the ends of chromosomes. (**C**) Dataflow analysis of the pipeline. The diagram shows the share of different structural variant categories at each stage of the pipeline (left y-axis). Deletions make up the largest share of transposition events. Additionally, the share of remaining data is outlined (right y-axis). Approximately 4% of all structural variants initially found are finally identified as transposition events. (**D**) Helitrons and SINEs are more active relative to *VC2010*, while Novosib are especially active relative to *CB4856*. Relative activity is calculated by the share of a class' basepairs appearing in transposition events divided by its share of a the classes basepairs in the transposon annotation. (**E**) A linear relationship between phylogenetic distance and the number of observed transposition events becomes obvious for the *Caenorhabditis elegans* strains for both SV callers PBSV and Sniffles. Phylogenetic distance is calculated as sum of distances in the phylogenetic tree to the last common ancestor. (**F**) The phylogenetic tree of the *Caenorhabditis elegans* strains. The branch lengths are proportional to the number of polymorphisms that differentiate each pair. Tree based on data from (101).

the identification of SVs and TEs are computationally demanding tasks, the identification of transposition events using deTEct takes only a few seconds to run.

DISCUSSION

Here, we present TransposonUltimate, a bundle of three modules for transposon classification, annotation and transposition event detection. Moreover, we present TransposonDB, a database containing more than 891 051 transposon sequences from a wide range of species. Our benchmark shows that the classification module RFSB outperforms existing methods. Although *RFSB* has a very high accuracy, we believe that performance could be improved by developing species specific classifiers. It would also be helpful to explore new feature representations that strongly correlate to phylogenetic distance metrics.

The annotation module combines existing annotation approaches using an ensemble strategy, and this ensures a less biased outcome than existing methods that tend to favor certain TE classes. The annotation module could be extended by the search for fragmented copies of annotated

transposons connected with filters to avoid false positives. Application to three different species revealed that TEs from the same family vary drastically in length. Thus, an important question for future research is to determine to what extent such differences reflect hitherto uncharacterized families, and to what extent the differences correspond to overall sequence divergence.

The detection module enables the identification of transposition events through structural variants in genomes profiled using long-read sequencing technologies. Application of the *deTEct* pipeline to 20 wild type strains of *C. elegans* suggests that transposon events are responsible for 7.37% of structural variants. Although previous studies have argued that transposons are a major driver of structural variation (102), our results suggest that at least for wild isolates of *Caenorhabditis elegans* this is not the case. As additional high quality assemblies become available, it will be interesting to further explore this important question. Moreover, the development of localisation algorithms of target and donor sites of transposons seems a promising add-on for the detection module. Besides, structural variants gathered from whole genome comparison using anchor filtering (104) could be included and compared.

As long-read technologies are becoming more widely used and the number of sequenced genomes rises quickly, there is an urgent need for methods to identify and annotate TEs which correspond to plurality and in some cases a majority of genome sequences. In particular, as more human (105) and other vertebrate genomes (https://vertebrategenomesproject.org/) are profiled using these technologies, TransposonUltimate will be a valuable tool to improve our understanding of the impact of TEs on both traits and diseases.

CONCLUSION

Our TransposonUltimate bundle of software tools provides a powerful and user-friendly means of analyzing TEs. In addition to providing highly accurate classifications, our analysis also provides insights as to what features are most informative for predicting TE class. Our ensemble approach to annotation is more unbiased than existing methods that tend to focus on one or a few classes. Finally, our transposition event detection module can take advantage of long-read technologies to identify to what extent TEs underlie SVs.

DATA AVAILABILITY

Databases, assemblies, annotations and further findings can be downloaded from https://cellgeni.cog.sanger.ac.uk/browser.html?shared=transposonultimate. TransposonDB is available at Zenodo with DOI 10.5281/zenodo.5518085 . Source codes, Conda package, installation manual and further documentation and further instructions can be found on https://github.com/DerKevinRiehl/TransposonUltimate.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Sarah Buddle, Simone Procaccia, Fu Xiang Quah and Alexandra Dallaire for assistance with testing and debugging the software. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author contributions: The study was conceived and designed by K.R., C.R., E.A.M. and M.H. The code was written by K.R., and the analyses were carried out by K.R. and C.R. The work was supervised by E.A.M. and M.H. K.R. and M.H. wrote the manuscript with input from E.A.M. and C.R.

FUNDING

Cancer Research UK [C13474/A18583, C6946/A14492]; Wellcome Trust [219475/Z/19/Z, 092096/Z/10/Z to E.A.M.]. Funding for open access charge: Wellcome Trust.
Conflict of interest statement. None declared.

REFERENCES

1. Lerat,E. (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, **104**, 520–533.
2. Saha,S., Bridges,S., Magbanua,Z.V. and Peterson,D.G. (2008) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Trop. Plant Biol.*, **1**, 85–96.
3. Wicker,T., Sabot,F., Hua-Van,A., Bennetzen,J.L., Capy,P., Chalhoub,B., Flavell,A., Leroy,P., Morgante,M., Panaud,O. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.
4. Kazazian,H.H. (2004) Mobile elements: drivers of genome evolution. *Science (New York, NY)*, **303**, 1626–1632.
5. Levin,H.L. and Moran,J.V. (2011) Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.*, **12**, 615–627.
6. Teixeira,F.K., Okuniewska,M., Malone,C.D., Coux,R.-X., Rio,D.C. and Lehmann,R. (2017) piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. *Nature*, **552**, 268–272.
7. Goerner-Potvin,P. and Bourque,G. (2018) Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, **19**, 688–704.
8. Biémont,C. and Vieira,C. (2006) Junk DNA as an evolutionary force. *Nature*, **443**, 521–524.
9. Emera,D. and Wagner,G.P. (2012) Transposable element recruitments in the mammalian placenta: impacts and mechanisms. *Brief. Funct. Genom.*, **11**, 267–276.
10. Kazazian,H.H., Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G. and Antonarakis,S.E. (1988) Haemophilia A resulting from de novo insertion of L 1 sequences represents a novel mechanism for mutation in man. *Nature*, **332**, 164–166.
11. Miki,Y., Nishisho,I., Horii,A., Miyoshi,Y., Utsunomiya,J., Kinzler,K.W., Vogelstein,B. and Nakamura,Y. (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.*, **52**, 643–645.
12. Sun,W., Samimi,H., Gamez,M., Zare,H. and Frost,B. (2018) Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. *Nat. Neurosci.*, **21**, 1038–1048.
13. Vilen,H., Aalto,J.-M., Kassinen,A., Paulin,L. and Savilahti,H. (2003) A direct transposon insertion tool for modification and functional analysis of viral genomes. *J. Virol.*, **77**, 123–134.
14. Vizváryová,M. and Valková,D. (2004) Transposons - the useful genetic tools. *Biologia*, **59**, 309–318.
15. Ivics,Z., Li,M.A., Mátés,L., Boeke,J.D., Bradley,A. and Izsvák,Z. (2009) Transposon-mediated genome manipulations in vertebrates. *Nat. Methods*, **6**, 415–422.
16. Girgis,H.Z. (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics*, **16**, 227.

17. Gilly,A., Etcheverry,M., Madoui,M.-A., Guy,J., Quadrana,L., Alberti,A., Martin,A., Heitkam,T., Engelen,S., Labadie,K. *et al.* (2014) TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics*, **15**, 377.
18. Abrusán,G., Grundmann,N., DeMester,L. and Makalowski,W. (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, **25**, 1329–1330.
19. Hoede,C., Arnoux,S., Moisset,M., Chaumier,T., Inizan,O., Jamilloux,V. and Quesneville,H. (2014) PASTEC: an automatic transposable element classification tool. *PLOS ONE*, **9**, e91929.
20. Schietgat,L., Vens,C., Cerri,R., Fischer,C.N., Costa,E., Ramon,J., Carareto,C. M.A. and Blockeel,H. (2018) A machine learning based framework to identify and classify long terminal repeat retrotransposons. *PLoS Comput. Biol.*, **14**, e1006097.
21. Kamath,U., Jong,K.D. and Shehu,A. (2014) Effective automated feature construction and selection for classification of biological sequences. *PLoS ONE*, **9**, e99982.
22. Arango-López,J., Orozco-Arias,S., Salazar,J.A. and Guyot,R. (2017) Application of data mining algorithms to classify biological data: the coffea canephora genome case. In: Solano,A. and Ordoñez,H. (eds). *Advances in Computing*. Springer International Publishing Communications in Computer and Information Science, Cham, pp. 156–170.
23. Nakano,F.k., Martiello Mastelini,S., Barbon,S. and Cerri,R. (2017) Stacking methods for hierarchical classification. In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. pp. 289–296.
24. Nakano,F.K., Pinto,W.J., Pappa,G.L. and Cerri,R. (2017) Top-down strategies for hierarchical classification of transposable elements with neural networks. In: *2017 International Joint Conference on Neural Networks (IJCNN)*, ISSN: 2161-4407, pp. 2539–2546.
25. Loureiro,T., Camacho,R., Vieira,J. and Fonseca,N.A. (2013) Boosting the detection of transposable elements using machine learning. In: Mohamad,M.S., Nanni,L., Rocha,M.P. and Fdez-Riverola,F. (eds). *7th International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer International Publishing Advances in Intelligent Systems and Computing, Heidelberg, pp. 85–91.
26. Loureiro,T., Camacho,R., Vieira,J. and Fonseca,N.A. (2013) Improving the performance of transposable elements detection tools. *J. Integr. Bioinformatics*, **10**, 40–50.
27. Nakano,F.K., Mastelini,S.M., Barbon,S. and Cerri,R. (2018) Improving hierarchical classification of transposable elements using deep neural networks. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. ISSN: 2161-4407, pp. 1–8.
28. da Cruz,M.H.P., Saito,P.T.M., Paschoal,A.R. and Bugatti,P.H. (2019) Classification of transposable elements by convolutional neural networks. In: Rutkowski, L., Scherer,R., Korytkowski,M., Pedrycz,W., Tadeusiewicz,R. and Zurada,J.M. (eds). *Artificial Intelligence and Soft Computing*. Springer International Publishing Lecture Notes in Computer Science, Cham, pp. 157–168.
29. Cruz,M. H. P.D., Domingues,D.S., Saito,P. T.M., Paschoal,A.R. and Bugatti,P.H. (2021). TERL: classification of transposable elements by convolutional neural networks. *Briefings in bioinformatics* , **22** , bbaa185.
30. Ashlock,W. and Datta,S. (2012) Distinguishing endogenous retroviral LTRs from SINE elements using features extracted from evolved side effect machines. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **9**, 1676–1689
31. Pereira,G.T., Santos,B.Z. and Cerri,R. (2018) A genetic algorithm for transposable elements hierarchical classification rule induction. In: *2018 IEEE Congress on Evolutionary Computation (CEC)*. pp. 1–8.
32. Pereira,G.T. and Cerri,R. (2018) Hierarchical and non-hierarchical classification of transposable elements with a genetic algorithm. *J. Inform. Data Manage.*, **9**, 163–163.
33. Pereira,G.T., Gabriel,P. H.R. and Cerri,R. (2019) A lexicographic genetic algorithm for hierarchical classification rule induction. In: *Proceedings of the Genetic and Evolutionary Computation Conference New York*. Association for Computing Machinery GECCO '19, NY, pp. 846–854.
34. Pereira,G.T., Gabriel,P.H. and Cerri,R. (2019) Hierarchical classification of transposable elements with a weighted genetic algorithm. In: *EPIA Conference on Artificial Intelligence*. Springer, pp. 737–749.
35. Feschotte,C., Keswani,U., Ranganathan,N., Guibotsy,M.L. and Levine,D. (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol. Evol.*, **1**, 205–220.
36. Feschotte,C. and Pritham,E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Ann. Rev. Genet.*, **41**, 331–368.
37. Flutre,T., Permal,E. and Quesneville,H. (2012) Transposable Element Annotation in Completely Sequenced Eukaryote Genomes. In: Grandbastien,M.A. and Casacuberta,J. (eds). *Plant Transposable Elements. Topics in Current Genetics*. Springer, Berlin, Heidelberg, Vol. **24**, pp. 17–39.
38. Ragupathy,R., You,F.M. and Cloutier,S. (2013) Arguments for standardizing transposable element annotation in plant genomes. *Trends Plant Sci.*, **18**, 367–376.
39. Arensburger,P., Piégu,B. and Bigot,Y. (2016) The future of transposable element annotation and their classification in the light of functional genomics-what we can learn from the fables of Jean de la Fontaine? *Mobile Genet. Elem.*, **6**, e1256852.
40. Edgar,R.C. and Myers,E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics*, **21**, i152–i158.
41. Kennedy,R.C., Unger,M.F., Christley,S., Collins,F.H. and Madey,G.R. (2011) An automated homology-based approach for identifying transposable elements. *BMC Bioinformatics*, **12**, 130.
42. Xiong,W., He,L., Lai,J., Dooner,H.K. and Du,C. (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Nat. Acad. Sci. U.S.A.*, **111**, 10263–10268.
43. Bergman,C.M. and Quesneville,H. (2007) Discovering and detecting transposable elements in genome sequences. *Brief. bioinform.*, **8**, 382–392.
44. Ou,S., Su,W., Liao,Y., Chougule,K., Agda,J.R., Hellinga,A.J., Lugo,C. S.B., Elliott,T.A., Ware,D., Peterson,T. *et al.* (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, **20**, 275.
45. Ye,C., Ji,G. and Liang,C. (2016) detectMITE: a novel approach to detect miniature inverted repeat transposable elements in genomes. *Sci. Rep.-UK*, **6**, 19688.
46. Rho,M. and Tang,H. (2009) MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res.*, **37**, e143.
47. Han,Y. and Wessler,S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.*, **38**, e199.
48. Buisine,N., Quesneville,H. and Colot,V. (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, **91**, 467–475.
49. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.
50. Alkan,C., Coe,B.P. and Eichler,E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.
51. Ewing,A.D. (2015) Transposable element detection from whole genome sequence data. *Mobile DNA*, **6**, 24.
52. Disdero,E. and Filée,J. (2017) LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mobile DNA*, **8**, 5.
53. Yi,F., Ling,J., Xiao,Y., Zhang,H., Ouyang,F. and Wang,J. (2018) ConTEdb: a comprehensive database of transposable elements in conifers. *Database*, **2018**, bay131.
54. Li,S.-F., Zhang,G.-J., Zhang,X.-J., Yuan,J.-H., Deng,C.-L., Gu,L.-F. and Gao,W.-J. (2016) DPTEdb, an integrative database of transposable elements in dioecious plants. *Database*, **2016**, baw078.
55. Nussbaumer,T., Martis,M.M., Roessner,S.K., Pfeifer,M., Bader,K.C., Sharma,S., Gundlach,H. and Spannagl,M. (2013) MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.*, **41**, D1144–D1151.
56. Ma,B., Li,T., Xiang,Z. and He,N. (2015) MnTEdb, a collective resource for mulberry transposable elements. *Database*, **2015**, bav004.

57. Chen,J., Hu,Q., Zhang,Y., Lu,C. and Kuang,H. (2014) P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.*, **42**, D1176–D1181.
58. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.
59. Copetti,D., Zhang,J., El Baidouri,M., Gao,D., Wang,J., Barghini,E., Cossu,R.M., Angelova,A., Maldonado,L. C.E., Roffler,S. *et al.* (2015) RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics*, **16**, 538.
60. Du,J., Grant,D., Tian,Z., Nelson,R.T., Zhu,L., Shoemaker,R.C. and Ma,J. (2010) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. *BMC Genomics*, **11**, 113.
61. Yi,F., Jia,Z., Xiao,Y., Ma,W. and Wang,J. (2018) SPTEdb: a database for transposable elements in salicaceous plants. *Database*, **2018**, bay024.
62. Wicker,T., Matthews,D.E. and Keller,B. (2002) TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.*, **7**, 561–562.
63. Lu,S., Wang,J., Chitsaz,F., Derbyshire,M.K., Geer,R.C., Gonzales,N.R., Gwadz,M., Hurwitz,D.I., Marchler,G.H., Song,J.S. *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.*, **48**, D265–D268.
64. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
65. Kiritchenko,S., Matwin,S., Nock,R. and Famili,A.F. (2006) Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization. In: Lamontagne,L. and Marchand,M. (eds). *Advances in Artificial Intelligence*. Springer Lecture Notes in Computer Science, Berlin, Heidelberg, pp. 395–406.
66. Ellinghaus,D., Kurtz,S. and Willhoeft,U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.
67. Gremme,G., Steinbiss,S. and Kurtz,S. (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **10**, 645–656.
68. Wenke,T., Döbel,T., Sörensen,T.R., Junghans,H., Weisshaar,B. and Schmidt,T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117–3128.
69. Mao,H. and Wang,H. (2017) SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. *Bioinformatics*, **33**, 743–745.
70. Ge,R., Mai,G., Zhang,R., Wu,X., Wu,Q. and Zhou,F. (2017) MUSTv2: an improved de novo detection program for recently active miniature inverted repeat transposable elements (MITEs). *J. Int. Bioinform.*, **14**, 20170029.
71. Hu,J., Zheng,Y. and Shang,X. (2018) MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. *BMC Med. Genom.*, **11**, 51–59.
72. Crescente,J.M., Zavallo,D., Helguera,M. and Vanzetti,L.S. (2018) MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics*, **19**, 348.
73. Drost,H.-G. (2020) LTRpred: _de novo_ annotation of intact retrotransposons. *J. Open Source Softw.*, **5**, 2170.
74. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
75. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
76. Maeda,T., Kobayashi,Y., Kameoka,H., Okuma,N., Takeda,N., Yamaguchi,K., Bino,T., Shigenobu,S. and Kawaguchi,M. (2018) Evidence of non-tandemly repeated rDNAs and their intragenomic heterogeneity in *Rhizophagus irregularis*. *Commun. Biol.*, **1**, 87.
77. Ricico,C.E.A. (2021) Super cool paper from Cristian, check it out. *Nature*, **1**, 1–1000.
78. Sedlazeck,F.J., Rescheneder,P., Smolka,M., Fang,H., Nattestad,M., von Haeseler,A. and Schatz,M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, **15**, 461–468.
79. Bessereau,J.-L. (2006) Transposons in *C. elegans*. *WormBook*, 1–13.
80. Laricchia,K., Zdraljevic,S., Cook,D. and Andersen,E. (2017) Natural variation in the distribution and abundance of transposable elements across the Caenorhabditis elegans species. *Mole. Biol. Evol.*, **34**, 2187–2202.
81. Huang,X., Lu,G., Zhao,Q., Liu,X. and Han,B. (2008) Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physio.*, **148**, 25–40.
82. Jiang,N., Bao,Z., Zhang,X., Hirochika,H., Eddy,S.R., McCouch,S.R. and Wessler,S.R. (2003) An active DNA transposon family in rice. *Nature*, **421**, 163–167.
83. Picault,N., Chaparro,C., Piegu,B., Stenger,W., Formey,D., Llauro,C., Descombin,J., Sabot,F., Lasserre,E., Meynard,D. *et al.* (2009) Identification of an active LTR retrotransposon in rice. *Plant J.*, **58**, 754–765.
84. Xu,Z. and Ramakrishna,W. (2008) Retrotransposon insertion polymorphisms in six rice genes and their evolutionary history. *Gene*, **412**, 50–58.
85. Yang,G., Nagel,D.H., Feschotte,C., Hancock,C.N. and Wessler,S.R. (2009) Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *science*, **325**, 1391–1394.
86. Panaud,O., Vitte,C., Hivert,J., Muzlak,S., Talag,J., Brar,D. and Sarr,A. (2002) Characterization of transposable elements in the genome of rice (Oryza sativa L.) using representational difference analysis (RDA). *Mol. Genet. Genom.*, **268**, 113–121.
87. Mao,L., Wood,T.C., Yu,Y., Budiman,M.A., Tomkins,J., Woo,S.-S., Sasinowski,M., Presting,G., Frisch,D., Goff,S. *et al.* (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. *Genome Res.*, **10**, 982–990.
88. McCarthy,E.M., Liu,J., Lizhi,G. and McDonald,J.F. (2002) Long terminal repeat retrotransposons of Oryza sativa. *Genome Biol.*, **3**, research0053.1–research0053.11.
89. Oki,N., Yano,K., Okumoto,Y., Tsukiyama,T., Teraishi,M. and Tanisaka,T. (2008) A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica. *Genes Genet. Syst.*, **83**, 321–329.
90. Morin,E., Miyauchi,S., San Clemente,H., Chen,E.C., Pelin,A., de la Providencia,I., Ndikumana,S., Beaudet,D., Hainaut,M., Drula,E. *et al.* (2019) Comparative genomics of Rhizophagus irregularis, R. cerebriforme, R. diaphanus and Gigaspora rosea highlights specific genetic features in Glomeromycotina. *New Phytol.*, **222**, 1584–1598.
91. Feschotte,C. and Wessler,S.R. (2001) Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 8923–8924.
92. Garrigues,J.M., Tsu,B.V., Daugherty,M.D. and Pasquinelli,A.E. (2019) Diversification of the Caenorhabditis heat shock response by helitron transposable elements. *Elife*, **8**, e51139.
93. Kapitonov,V.V. and Jurka,J. (2001) Rolling-circle transposons in eukaryotes. *Proc. Nat. Acad. Sci. U.S.A.*, **98**, 8714–8719.
94. Sijen,T. and Plasterk,R.H. (2003) Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature*, **426**, 310–314.
95. Waterston,R. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* sequencing consortium. *Science*, **282**, 2012–2018.
96. Eide,D. and Anderson,P. (1985) Transposition of Tc1 in the nematode *Caenorhabditis elegans*. *Proc. Nat. Acad. Sci. U.S.A.*, **82**, 1756–1760.
97. Plasterk,R.H., Izsvák,Z. and Ivics,Z. (1999) Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet.*, **15**, 326–332.
98. Cutter,A.D. and Payseur,B.A. (2003) Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol. Biol. Evol.*, **20**, 665–673.
99. Rockman,M.V. and Kruglyak,L. (2009) Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet.*, **5**, e1000419.
100. Rockman,M.V., Skrovanek,S.S. and Kruglyak,L. (2010) Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science*, **330**, 372–376.
101. Andersen,E.C., Gerke,J.P., Shapiro,J.A., Crissman,J.R., Ghosh,R., Bloom,J.S., Félix,M.-A. and Kruglyak,L. (2012) Chromosome-scale

selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.*, **44**, 285.
102. Fuentes,R.R., Chebotarov,D., Duitama,J., Smith,S., De la Hoz,J.F., Mohiyuddin,M., Wing,R.A., McNally,K.L., Tatarinova,T., Grigoriev,A. *et al.* (2019) Structural variants in 3000 rice genomes. *Genome Res.*, **29**, 870–880.
103. Huang,C.R.L., Burns,K.H. and Boeke,J.D. (2012) Active transposition in genomes. *Ann. Rev. Gen.*, **46**, 651–675.
104. Nattestad,M. and Schatz,M.C. (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*, **32**, 3021–3023.
105. Sherman,R.M. and Salzberg,S.L. (2020) Pan-genomics in the human genome era. *Nat. Rev. Genet.*, **21**, 243–254.
106. Kapitonov,V.V. and Jurka,J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, **9**, 411–412.
107. Daron,J., Glover,N., Pingault,L., Theil,S., Jamilloux,V., Paux,E., Barbe,V., Mangenot,S., Alberti,A., Wincker,P. *et al.* (2014) Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol.*, **15**, 546.
108. Kohany,O., Gentles,A.J., Hankus,L. and Jurka,J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, **7**, 474.
109. Guo,R., Li,Y.-R., He,S., Ou-Yang,L., Sun,Y. and Zhu,Z. (2018) RepLong: de novo repeat identification using long read sequencing data. *Bioinformatics*, **34**, 1099–1107.
110. Lee,H., Lee,M., Mohammed Ismail,W., Rho,M., Fox,G.C., Oh,S. and Tang,H. (2016) MGEScan: a Galaxy-based system for identifying retrotransposons in genomes. *Bioinformatics*, **32**, 2502–2504.
111. Xu,Z. and Wang,H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*, **35**, W265–W268.
112. Valencia,J.D. and Girgis,H.Z. (2019) LtrDetector: a tool-suite for detecting long terminal repeat retrotransposons de-novo. *BMC Genomics*, **20**, 450.
113. Steinbiss,S., Willhoeft,U., Gremme,G. and Kurtz,S. (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.*, **37**, 7002–7013.