METHOD

TransposonUltimate: software for transposon classification, annotation and detection

Kevin Riehl¹, Cristian Riccio¹², Eric A. Miska¹²³ and Martin Hemberg²⁴*

Abstract

Motivation: 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. Moreover, many methods are difficult to install, poorly documented, and difficult to reproduce.

Results: 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. We benchmark the classification module on the large TransposonDB covering over 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 twenty wild type strains of Caenorhabditis elegans.

Availability: Running software and source code available on https://github.com/DerKevinRiehl/TransposonClassifierRFSB. Databases, assemblies, annotations and further findings can be downloaded from https://cellgeni.cog.sanger.ac.uk/browser.html?shared=transposonultimate.

Keywords: Transposable elements; Transposon classification; Transposon annotation; Transposon detection

1 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 (Fig. 1A), and each class is associated with a set of characteristics, e.g. proteins relevant for transposition and structural features (Fig. 1 B). During transposition, transposable elements (TEs) can leave structural patterns both at the insertion and the deletion site [1, 2, 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 called transposonpathies such as subtypes of haemophilia, immunodeficiency, cancer and Alzheimer’s disease [10, 11, 12]. Also, TEs are popular for genetic engineering purposes as they allow for direct insertion of their genetic cargo into a target genome [13, 14, 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 are error prone, not robust, mostly rely on prior knowledge of transposon sequences, 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 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, 28, 29], evolutionary algorithms [30, 21] and genetic algorithms [31, 32, 33, 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 [43, 2, 1, 41]. 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 [41, 39], 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 characterised 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-read 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 (Fig. 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.

2 Materials and methods

2.1 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, 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* (Fig. 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-PSB [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/) [1], 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/SPTEdbrowse.php?species=ptr&name=Populus_trichocarpa) and TrepDB [62] (http://botserv2.uzh.ch/keldata/trep-db/downloadFiles.html).

To create the database, the taxonomies were unified, duplicates were dropped and several filter rules were applied (Table S1). Filtering included the removal of sequences with no label, the exclusion of fragments, contigs, satellites and RNA sequences. Moreover, only sequences with a length greater than 100bp 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 (Table S2 + S3). Moreover, the database is balanced and covers sufficient examples for all taxonomic nodes (Table S4).

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 (Table S5). RPSTBLASTN (v2.10.1) was used to annotate the conserved domain models at an e-value of 5.0 as it performed best in terms of classification performance (Fig. S1 A-B). In addition, two model structures were explored. The binary structure employs binary classifiers for each node (= transposon class) of the taxonomy. After inference of the binary classifiers, the taxonomic class can be determined by choosing the most probable node at each stage. 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 comprehensive 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

[1] We use version 23.08 that was the last publicly available version before the paywall was introduced.
yields the highest classification performance (Fig. S2). Based on these results, we propose a random forest classifier with a selective training strategy on a binary model structure, \textit{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 \cite{65}. The second perspective evaluates for different taxonomic levels and the third perspective captures the classification performance of single classes. We benchmark RFSB against TERL \cite{29}, TopDown \cite{24}, NLLCPN \cite{27}, HC\_LGA \cite{33} and HC\_GA \cite{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.

2.2 Transposon annotation module, \textit{reasonaTE}

Given an assembled genome, the goal of the annotation module is to find all transposon occurrences and their locations. Our \textit{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 (\url{http://www.repeatmasker.org/}), RepeatModeler (\url{http://www.repeatmasker.org/RepeatModeler/}), LTRharvest \cite{66} (\url{https://www.zbh.unihamburg.de/forschung/gi/software/ltrharvest.html}) and TIRvish \cite{67} (\url{http://genometools.org/tools/gt_tirvish.html}) are available as Conda packages. Moreover, we created Conda packages for SINE-Finder \cite{68} (\url{http://www.plantcell.org/content/suppl/2011/08/29/tpc.111.088682.DC1/Supplemental_Data_Set_1-sine_finder.txt}), SINE-Scan \cite{69} (\url{https://github.com/maohlzj/SINE_Scan}), HelitronScanner \cite{42} (\url{https://sourceforge.net/projects/helitronsscanner/files/}), MUSTv2 \cite{70} (\url{http://www.healthinformaticslab.org/supp/resources.php}), MiteFinderII \cite{71} (\url{https://github.com/jhu99/miteFinder}) and MITE-Tracker \cite{72} (\url{https://github.com/INTABiotechMJ/MITE-Tracker}) to make them accessible and to facilitate their installation. Also, we include the output files of LTRpred \cite{73} (\url{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) \cite{74, 75} and BLASTN (v2.10.1). For the annotation of transposon-characteristic proteins, we have created a Conda packaged version of TransposonPSI (\url{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 \textbf{F2}). As an application, here we annotate the genome \textit{MSU7} of \textit{Oryza sativa subspecies japonica} (\url{http://rice.plantbiology.msu.edu/index.shtml}).
2.3 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 (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 finally 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 both PacBio and OxfordNanopore data.

3 Results

3.1 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 (Fig. 4 A, Table S7). TE Learner [20] has the lowest reported performance, while the other methods have similar F1 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 TransposonDB which includes RepBase and PGSB (Fig. 4 B). 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 (Fig. 4 C). 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 longer k-mer features contribute the most to the classification performance, while protein domains have a smaller share amongst the most contributing features (Fig. 4 D). This motivated the exploration of longer k-mer features, but we did not find any significant increase of the performance when using 5-mers (Fig. 4 E).

3.2 The ensemble strategy reasonaTE finds more transposons and reduces bias

Next, we evaluated the ability of our reasonaTE pipeline to identify TEs in the genomes of three different species (Fig. 5 A-B, File F4). The TE content of almost 21% for Caenorhabditis 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 Oryza sativa subs. japonica is very close to the mean of other reports [81, 82, 83, 84, 85, 86, 87, 88, 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 Caenorhabditis elegans. Interestingly, the relative transposon class frequency reveals clear differences across species (Fig. 5 C-D). Similarly, the length distributions (Fig. 5 E-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] and [93], the share of Helitrons amounts to almost 2% of the Caenorhabditis elegans genome. Moreover, the majority of the transposons are TIR DNA transposons, as reported by [94, 95, 79]. Contrary to previous studies [80, 96, 97], we mainly find hAT, CMC and Novosib transposons to be present in the Caenorhabditis 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 Caenorhabditis elegans (Figure S3), consistent with the higher densities observed for other variants [79, 98, 99, 100, 101, 80].

The comparison of different annotation tools reveals that reasonaTE provides more unbiased results (Figure S4) as none of the other methods find 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 RepeatModeller were confirmed as transposons using our approach. Moreover, the transposon characteristic protein annotations by TransposonPSI and the 1,000 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 (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 different classes as well.

### 3.3 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 *Caenorhabditis elegans*. Each strain was compared to the two reference genomes *VC2010* and *CB4856* (Fig. 6 A, Table S8, File F5). As expected, the newly sequenced genomes of 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 (Fig. 6 B) 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* (Fig. 6 D, 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 (Fig. 6 E-F, Table S9). The strains *QX1211* and *ECA36* have the largest differences based on transposon data before [80].

### 4 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 benchmarks 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 Caenorhabditis 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 (https://vertebrategenomesproject.org/) genomes 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.

5 Conclusions
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.

Competing interests
The authors declare that they have no competing interests.

Author’s contributions
The study was conceived and designed by KR, CR, EAM and MH. The code was written by KR, and the analyses were carried out by KR and CR. The work was supervised by EAM and MH. KR and MH wrote the manuscript with input from EAM and CR.

Acknowledgements
We would like to thank Sarah Buddle, Simone Procaccia, Fu Xiang Quah and Alexandra Dallaire for assistance with testing and debugging the software.
Funding
This work was supported by Cancer Research UK (C13474/A18583, C6946/A14492) and the Wellcome Trust (219475/Z/19/Z, 092096/Z/10/Z) to EAM. 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 details
1 Gurdan Institute, University of Cambridge, CB2 1QN Cambridge, UK. 2 Wellcome Sanger Institute, Wellcome Genome Campus, CB10 1SA Hinxton, UK. 3 Department of Genetics, University of Cambridge, Downing Street, CB2 3EH Cambridge, UK. 4 Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women’s Hospital, 75 Francis Street, MA 02215 Boston, MA, USA.

References
1. Lerat, E.: Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104(6), 520–533 (2010)
2. Saha, S., Bridge, S., Magbanua, Z.V., Peterson, D.G.: Computational approaches and tools used in identification of dispersed repetitive dna sequences. Tropical Plant Biology 1(1), 85–96 (2008)
3. Wicker, T., Sabot, F., Hua-Van, A., Benetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H.: A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8(12), 973–986 (2007). doi:10.1038/nrg2165. Number: 12 Publisher: Nature Publishing Group. Accessed 2020-09-16
4. Kazazian, H.H.: Mobile elements: drivers of genome evolution. Science (New York, N.Y.) 303(5664), 1626–1632 (2004). doi:10.1126/science.1089670
5. Levin, H.L., Moran, J.V.: Dynamic interactions between transposable elements and their hosts. Nature Reviews Genetics 12(9), 615–627 (2011). doi:10.1038/nrg3030
6. Teixeira, F.K., Okusawa, C., Malone, C.D., Cudoux, R., Rio, D.C., Lehmann, R.: piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. Nature 552(7684), 268–272 (2017). doi:10.1038/nature25018. Number: 7684 Publisher: Nature Publishing Group. Accessed 2020-09-14
7. Goerner-Potvin, P., Bourque, G.: Computational tools to unmask transposable elements. Nature Reviews Genetics 19(11), 688–704 (2018). doi:10.1038/s41576-018-0050-x Number: 11 Publisher: Nature Publishing Group. Accessed 2020-08-29
8. Biémont, C., Vieira, C.: Junk DNA as an evolutionary force. Nature 443(7111), 521–524 (2006). doi:10.1038/443521a. Number: 7111 Publisher: Nature Publishing Group. Accessed 2020-08-24
9. Emera, D., Wagner, G.P.: Transposable element recruitments in the mammalian placenta: impacts and mechanisms. Briefings in Functional Genomics 11(4), 267–276 (2012). doi:10.1093/bfgp/elq013. Accessed 2020-08-21
10. Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., Antonarakis, S.E.: Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332(6160), 164–166 (1988). doi:10.1038/332164a0. Number: 6160 Publisher: Nature Publishing Group. Accessed 2020-08-21
11. Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., Nakamura, Y.: Disruption of the APC gene by a retrotransposon insertion of L1 sequence in a colon cancer. Cancer Research 52(3), 643–645 (1992)
12. Sun, W., Samimi, H., Gabel, M., Zare, H., Frost, B.: Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. Nature Neuroscience 21(8), 1038–1048 (2018). doi:10.1038/s41593-018-0194-1. Accessed 2020-08-21
13. Vilen, H., Aalto, J.-M., Kassinen, A., Paulin, L., Savilahti, H.: A Direct Transposon Insertion Tool for Modification and Functional Analysis of Viral Genomes. Journal of Virology 77(11), 123–134 (2003). doi:10.1128/JVI.77.11.123-134.2003. Accessed 2020-09-14
14. Vizváryová, M., Valková, D.: Transposons - the useful genetic tools. Biologia - Section Cellular and Molecular Biology 59, 309–318 (2004)
15. Ivics, Z., Li, M.A., Mátés, L., Boeke, J.D., Bradley, A., Izsák, Z.: Transposon-mediated Genome Manipulations in Vertebrates. Nature methods 6(6), 415–422 (2009). doi:10.1038/nmeth.1332. Accessed 2020-09-14
16. Girgis, H.Z.: Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16(1), 227 (2015). doi:10.1186/s12859-015-0654-5. Accessed 2020-09-05
17. Gilly, A., Etchevery, M., Madeou, M.A., Guy, J., Quadran, L., Alberti, A., Martin, A., Heitkam, T., Engelen, S., Labadie, K., Le Pen, J., Wincker, P., Colot, V., Aury, J.-M.: TE-Tracker: systematic identification of transposition events through whole-genome resequencing. BMC Bioinformatics 15(1), 377 (2014). doi:10.1186/s12859-014-0377-z. Accessed 2020-08-21
18. Abrusán, G., Grundmann, N., DeMester, L., Makalowski, W.: TEClass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25(10), 1329–1330 (2009). doi:10.1093/bioinformatics/btp084. Publisher: Oxford Academic. Accessed 2020-09-07
19. Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., Quesneville, H.: PASTEC: An Automatic Transposable Element Classification Tool. PLOS ONE 9(5), 91929 (2014). doi:10.1371/journal.pone.0091929. Publisher: Public Library of Science. Accessed 2020-09-16
20. Schietgat, L., Vens, C., Cerri, R., Fischer, C.N., Costa, E., Ramon, J., Carareto, C.M.A., Blockeel, H.: A machine learning based framework to identify and classify long terminal repeat retrotransposons. PLOS Computational Biology 14(4), 1000697 (2018). doi:10.1371/journal.pcbi.1000697. Publisher: Public Library of Science. Accessed 2020-09-07
21. Kamath, U., Jong, K.D., Shehu, A.: Effective Automated Feature Construction and Selection for Classification of Biological Sequences. PLOS ONE 9(7), 99982 (2014). doi:10.1371/journal.pone.0099982. Publisher: Public Library of Science. Accessed 2020-09-08
22. Arango-López, J., Orozco-Arias, S., Salazar, J.A., Guyot, R.: Application of Data Mining Algorithms to Classify Biological Data: The Coffea canephora Genome Case. In: Solano, A., Ordóñez, H. (eds.) Advances in Computing. Communications in Computer and Information Science, pp. 156–170. Springer, Cham (2017). doi:10.1007/978-3-319-66562-7_2

23. Nakano, F.K., Martiello Mastelini, S., Barbón, S., Cerri, R.: Stacking Methods for Hierarchical Classification. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 289–296 (2017). doi:10.1109/ICMLA.2017.0-145

24. Nakano, F.K., Pinto, W.J., Pappa, G.L., Cerri, R.: Top-down strategies for hierarchical classification of transposable elements with neural networks. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2539–2546 (2017). doi:10.1109/IJCNN.2017.7966165. ISBN: 2161-4407

25. Loureiro, T., Camacho, R., Vieira, J., Fonseca, N.A.: Improving the Detection of Transposable Elements Using Machine Learning. In: Mahamad, M.S., Nanni, L., Rocha, M.P., Fdez-Riverola, F. (eds.) 7th International Conference on Practical Applications of Computational Biology & Bioinformatics. Advances in Intelligent Systems and Computing, pp. 85–91. Springer, Heidelberg (2013). doi:10.1007/978-3-319-00578-2_2

26. Loureiro, T., Camacho, R., Vieira, J., Fonseca, N.A.: Improving the performance of Transposable Elements detection tools. Journal of Integrative Bioinformatics 10(3), 40–50 (2013). doi:10.1515/jib-2013-231.

Publisher: De Gruyter Section: Journal of Integrative Bioinformatics. Accessed 2020-09-07

27. Nakano, F.K., Mastelini, S.M., Barbón, S., Cerri, R.: Improving Hierarchical Classification of Transposable Elements using Deep Learning. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2018). doi:10.1109/IJCNN.2018.8489461. ISBN: 2161-4407

28. da Cruz, M.H.P., Saito, P.T.M., Paschoal, A.R., Bugatti, P.H.: Classification of Transposable Elements by Convolutional Neural Networks. In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurauda, J.M. (eds.) Artificial Intelligence and Soft Computing. Lecture Notes in Computer Science, pp. 157–168. Springer, Cham (2019). doi:10.1007/978-3-030-20915-5_5

29. Xiong, M.H.P., Domínguez, D.S., Saito, P.T.M., Paschoal, A.R., Bugatti, P.H.: TELR: Classification of Transposable Elements by Convolutional Neural Networks. bioRxiv, 2020–0325009935 (2020). doi:10.1101/2020.03.25.009935. Publisher: Cold Spring Harbor Laboratory Section: New Results. Accessed 2020-09-16

30. Ashlock, W., Datta, S.: Distinguishing Endogenous Retroviral LTRs from SINE Elements Using Features Extracted from Evolved Side Effect Machines. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6(6), 1676–1689 (2012). doi:10.1109/TCBB.2012.116. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics

31. Pereira, G.T., Santos, B.Z., Cerri, R.: A Genetic Algorithm for Transposable Elements Hierarchical Classification Rule Induction. In: 2018 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8 (2018). doi:10.1109/CEC.2018.8477642

32. Pereira, G.T., Cerri, R.: Hierarchical and Non-Hierarchical Classification of Transposable Elements with a Genetic Algorithm. Journal of Information and Data Management 9(2), 163–163 (2018). Number: 2. Accessed 2020-09-16

33. Pereira, G.T., Gabriel, P.H., Cerri, R.: Hierarchical classification of transposable elements with a weighted genetic algorithm. In: IEPI Conference on Artificial Intelligence, pp. 737–749 (2019). Springer

34. Peschotte, C., Keswani, U., Ranganathan, N., Guibots, M.L., Levine, D.: Exploring Repetitive DNA Elements Using Deep Learning. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2018). doi:10.1109/IJCNN.2018.8477642

35. Pereira, G.T., Santos, B.Z., Cerri, R.: A Genetic Algorithm for Transposable Elements Hierarchical Classification Rule Induction. In: 2018 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8 (2018). doi:10.1109/CEC.2018.8477642

36. Pereira, G.T., Gabirel, P.H., Cerri, R.: Hierarchical classification of transposable elements with a weighted genetic algorithm. In: IEPI Conference on Artificial Intelligence, pp. 737–749 (2019). Springer

37. Flutre, T., Permal, E., Quesneville, H.: Transposable element annotation in completely sequenced eukaryote genomes. Scientific reports 6, 205–220 (2016). doi:10.1038/srep20532. Publisher: Nature</ref>

38. Feschotte, C., Pritham, E.J.: DNA Transposons and the Evolution of Eukaryotic Genomes. Annual Review of Genomics and Evolution 11(1), 331–368 (2010). doi:10.1146/annurev.genet.40.110405.090448. Accessed 2020-08-21

39. Flutre, T., Permal, E., Quesneville, H.: Transposable element annotation in completely sequenced eukaryote genomes. In: Plant Transposable Elements, pp. 17–39. Springer, Heidelberg (2012)

40. Raghupathy, R., You, F.M., Cloutier, S.: Arguments for standardizing transposable element annotation in plant genomes. Trends in plant science 18(7), 367–376 (2013)

41. Arensburger, P., Piégu, B., Bigot, Y.: 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 genetic elements 6, 10263–10268 (2014)

42. Xiong, W., He, L., Lai, J., Dooner, H.K., Du, C.: HelitronScanner uncovers a large overlooked cache of helitron transposons in many plant genomes. Proceedings of the National Academy of Sciences 111(28), 10263–10268 (2014)

43. Bergman, C.M., Quesneville, H.: Discovering and detecting transposable elements in genome sequences. Briefings in bioinformatics 8(6), 382–392 (2007)

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.: Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome biology 20(1), 1–18 (2019)

45. Ye, C., Ji, G., Liang, C.: detcetmite: a novel approach to detect miniature inverted repeat transposable elements in genomes. Scientific reports 6, 19688 (2016)

46. Rho, M., Tang, H.: Msgecsan-non-ltr: computational identification and classification of autonomous non-ltr
retrotransposons in eukaryotic genomes. Nucleic acids research 37(21), 143–143 (2009)

47. Han, Y., Wessler, S.R.: Mite-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic acids research 38(22), 199–199 (2010)

48. Buisine, N., Quesneville, H., Colot, V.: Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics 91(5), 467–475 (2008)

49. Bao, W., Kojima, K.K., Kohany, O.: Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1), 11 (2015). doi:10.1186/s13100-015-0041-9. Accessed 2020-08-21

50. Alkan, C., Coe, B.P., Eichler, E.E.: Genome structural variation discovery and genotyping. Nature Reviews Genetics 12(5), 363–376 (2011). doi:10.1038/nrg2858. Number: 5 Publisher: Nature Publishing Group. Accessed 2020-09-05

51. Ewing, A.D.: Transposable element detection from whole genome sequence data. Mobile DNA 6(1), 24 (2015). doi:10.1186/s13100-015-0035-3. Accessed 2020-09-05

52. Diodero, E., Filele, J.: Lorte: Detecting transposon-induced genomic variants using low coverage pacbio long read sequences. Mobile DNA 8(1), 1–6 (2017)

53. Yi, F., Ling, J., Xiao, Y., Zhang, H., Ouyang, F., Wang, J.: ConTEdb: a comprehensive database of transposable elements in conifers. Database 2018 (2018). doi:10.1093/database/bay131. Publisher: Oxford Academic. Accessed 2020-09-15

54. Li, S.-F., Zhang, G.-J., Zhang, X.-J., Yuan, J.-H., Deng, C.-L., Gu, L.-F., Gao, W.-J.: DPTEdb, an integrative database of transposable elements in dioecious plants. Database 2016 (2016). doi:10.1093/database/baw078. Publisher: Oxford Academic. Accessed 2020-09-15

55. Nussbaumer, T., Martin, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H., Spannagl, M.: MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Research 41(D1), 1144–1151 (2013). doi:10.1093/nar/gkt1153. Publisher: Oxford Academic. Accessed 2020-09-15

56. Ma, B., Li, T., Xiang, Z., He, N.: MntEdb, a collective resource for mulberry transposable elements. Database 2015 (2015). doi:10.1093/database/bav004. Publisher: Oxford Academic. Accessed 2020-09-15

57. Chen, J., Hu, Q., Zhang, Y., Lu, C., Kuang, H.: P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Research 42(D1), 1176–1181 (2014). doi:10.1093/nar/gkt1000. Publisher: Oxford Academic. Accessed 2020-09-15

58. Bao, W., Kojima, K.K., Kohany, O.: Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1), 11 (2015). doi:10.1186/s13100-015-0041-9. Accessed 2020-09-15

59. Capetti, D., Zhang, J., El Baidouri, M., Gao, D., Wang, J., Barghini, E., Cosu, R.M., Angelova, A., Maldonado L., C.E., Roffler, S., Ohyanagi, H., Wicker, T., Fan, C., Zuccolo, A., Chen, M., Costa de Oliveira, A., Han, B., Henry, R., Hsing, Y.-i., Kurata, N., Wang, W., Jackson, S.A., Panaud, O., Wing, R.A.: RITE database: a resource database for genus-wide rice genomics and evolutionary biology. BMC Genomics 16(1), 538 (2015). doi:10.1186/s12864-015-1762-3. Accessed 2020-09-15

60. Du, J., Grant, D., Tian, Z., Nelson, R.T., Zhu, L., Shoemaker, R.C., Ma, J.: SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11(1), 113 (2010). doi:10.1186/1471-2164-11-113. Accessed 2020-09-15

61. Yi, F., Jia, Z., Xiao, Y., Ma, W., Wang, J.: SPTEdb: a database for transposable elements in salicaceous plants. Database 2018 (2018). doi:10.1093/database/bay024. Publisher: Oxford Academic. Accessed 2020-09-15

62. Wicker, T., Matthews, D.E., Keller, B.: TREP: a database for Triticeae repetitive elements. Trends in Plant Science 7(12), 561–562 (2002). doi:10.1016/S1360-1385(02)02372-5. Publisher: Elsevier. Accessed 2020-09-15

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.: Cdd/sparcle: the conserved domain database in 2020. Nucleic acids research 48(D1), 265–268 (2020)

64. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830 (2011)

65. Kiritchenko, S., Matwin, S., Nock, R., Famili, A.F.: Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization. In: Lamontagne, L., Marchand, M. (eds.) Advances in Artificial Intelligence. Lecture Notes in Computer Science, pp. 395–406. Springer, Berlin, Heidelberg (2006). doi:10.1007/11766247_4

66. Ellinghama, D., Kurtz, S., Willhoet, U.: Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. BMC bioinformatics 9(1), 18 (2008)

67. Gremme, G., Steinbiss, S., Kurtz, S.: Genometools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10(3), 645–656 (2013)

68. Wenke, T., Döbel, T., Sörensen, T.R., Junghans, H., Weisshaar, B., Schmidt, T.: Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. The Plant Cell 23(9), 3117–3128 (2011)

69. Mao, H., Wang, H.: Sine_scan: an efficient tool to discover short interspersed nuclear elements (sines) in large-scale genomic datasets. Bioinformatics 33(5), 743–745 (2017)

70. Ge, R., Mai, G., Zhang, R., Wu, X., Wu, Q., Zhou, F.: Mustv2: an improved de novo detection program for recently active miniature inverted repeat transposable elements (mites). Journal of integrative bioinformatics 14(3) (2017)

71. Hu, J., Zheng, Y., Shang, X.: Mitefinder: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. BMC medical genetics 11(5), 51–59 (2018)

72. Crescente, J.M., Zavaldo, D., Helguera, M., Vanzetti, L.S.: Mite tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC bioinformatics 19(1), 348 (2018)

73. Drost, H.-G.: Ltrpred: de novo annotation of intact retrotransposons. Journal of Open Source Software
5(50), 2170 (2020)

74. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 20(13), 1658–1659 (2004)

75. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23), 3150–3152 (2012)

76. Mao, T., Kobayashi, Y., Kameoka, H., Okuma, N., Takeda, N., Yamaguchi, K., Bino, T., Shigenobu, S., Kawaguchi, M.: Evidence of non-tandemly repeated rdnas and their intragenomic heterogeneity in rhizophagus irregularis. Communications biology 1(1), 1–13 (2018)

77. Ricico, C.e.a.: Super cool paper from cristian, check it out. Nature 1(1), 1–1000 (2021)

78. Sedgezeka, F.J., Reschneider, P., Smolka, M., Fang, H., Nattestad, M., van Haeseler, A., Schatz, M.C.: Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods 15(6), 461–468 (2018). doi:10.1038/s41592-018-0001-7. Accessed 2020-08-21

79. Bessereau, J.-L.: Transposons in c. elegans. WormBook, 1 (2006)

80. Laricchia, K., Zdraljevic, S., Cook, D., Andersen, E.: Natural variation in the distribution and abundance of transposable elements across the caenorhabditis elegans species. Molecular biology and evolution 34(9), 2187–2202 (2017)

81. Huang, X., Lu, G., Zhao, Q., Liu, X., Han, B.: Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. Plant physiology 148(1), 25–40 (2008)

82. Xu, Z., Ramakrishna, W.: Retrotransposon insertion polymorphisms in six rice genes and their evolutionary history. Gene 412(1-2), 50–58 (2008)

83. Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., Wessler, S.R.: Tuned for transposition: molecular determinants underlying the hyperactivity of a stowaway mite. science 325(5946), 1391–1394 (2009)

84. Panaud, O., Vitte, C., Hivert, J., Muzlak, S., Talag, J., Brar, D., Sarr, A.: Characterization of transposable elements in the genome of rice (oryza sativa L.) using representational difference analysis (rda). Molecular Genetics and Genomics 268(1), 113–121 (2002)

85. Mao, L., Wood, T.C., Yu, Y., Budiman, M.A., Tomkins, J., Wood, S.-s., Sasinowski, M., Presting, G., Frisch, D., Goff, S., et al.: Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Research 10(7), 982–990 (2000)

86. McCarthy, E.M., Liu, J., Lizio, G., McDonald, J.F.: Long terminal repeat retrotransposons of oryza sativa. Genome biology 3(10), 1–11 (2002)

87. Oki, N., Yano, K., Okumoto, Y., Tsukiyama, T., Teraishi, M., Tanisaka, T.: A genome-wide view of miniature inverted-repeat transposable elements (mites) in rice, oryza sativa ssp. japonica. Genes & genetic systems 83(4), 321–329 (2008)

88. Beaudet, D., Hainaut, M., Drula, E., et al.: Comparative genomics of rhizophagus irregularis, r. cerebriforme, r. diaphanus and gigaspora rosea highlights specific genetic features in glomeromycotina. New Phytologist 222(3), 1584–1598 (2019)

89. Feschotte, C., Wessler, S.R.: Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proceedings of the National Academy of Sciences 98(16), 8923–8924 (2001)

90. Garrigues, J.M., Tsu, B.V., Daugherty, M.D., Pasquinelli, A.E.: Diversification of the caenorhabditis heat shock response by helitron transposable elements. Elife 8, 51139 (2019)

91. Kapitonov, V.V., Jurka, J.: Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences 98(15), 8714–8719 (2001)

92. Sijen, T., Plasterk, R.H.: Transposon silencing in the caenorhabditis elegans germ line by natural rnai. Nature 444, 687–692 (2001)

93. Kapitonov, V.V., Jurka, J.: Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences 98(15), 8714–8719 (2001)

94. Eide, D., Anderson, P.: Transposition of tci in the nematode caenorhabditis elegans. Proceedings of the National Academy of Sciences 82(6), 1756–1760 (1985)

95. Plasterk, R.H., Izsvák, Z., Ivics, Z.: Resident aliens: the tc1/mariner superfamily of transposable elements. Annual review of genetics 33, 412 (2002)

96. Garrigues, J.M., Tsu, B.V., Daugherty, M.D., Pasquinelli, A.E.: Diversification of the caenorhabditis heat shock response by helitron transposable elements. Elife 8, 51139 (2019)

97. Waterston, R.: Genome sequence of the nematode c. elegans: a platform for investigating biology. the c. elegans sequencing consortium. Science 282(5403), 2056–2064 (1998)

98. Waterston, R.: Genome sequence of the nematode c. elegans: a platform for investigating biology. the c. elegans sequencing consortium. Science 282(5403), 2056–2064 (1998)

99. Rockman, M.V., Kruglyak, L.: Recombinational landscape and population genomics of caenorhabditis elegans. PLoS Genet 5(3), 1000419 (2009)

100. Rockman, M.V., Skrovanek, S.S., Kruglyak, L.: Selection at linked sites shapes heritable phenotypic variation in c. elegans. Science 330(6002), 372–376 (2010)

101. Andersen, E.C., Gerke, J.P., Shapiro, J.A., Crissman, J.R., Ghosh, R., Bloom, J.S., Félix, M.-A., Kruglyak, L.: Chromosome-scale selective sweeps shape caenorhabditis elegans genomic diversity. Nature genetics 44(3), 285 (2012)

102. Fuentes, R.R., Chebotar, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A., et al.: Structural variants in 3000 rice genomes. Genome research 29(5), 870–880 (2019)

103. Huang, C.R.L., Burns, K.H., Boeke, J.D.: Active transposition in genomes. Annual review of genetics 46, 651–675 (2012)

104. Nattestad, M., Schatz, M.C.: Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics 32(19), 3021–3023 (2016)
105. Sherman, R.M., Salzberg, S.L.: Pan-genomics in the human genome era. Nature Reviews Genetics 21(4), 243–254 (2020)
106. Kapitonov, V.V., Jurka, J.: A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Reviews Genetics, 2 (2008). doi:10.1038/nrg2165-c1
107. Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., Barbe, V., Mangenot, S., Alberti, A., Wincker, P., et al.: Organization and evolution of transposable elements along the bread wheat chromosome 3b. Genome biology 15(12), 546 (2014)
108. Kohany, O., Gentles, A.J., Hankus, L., Jurka, J.: Annotation, submission and screening of repetitive elements in repbase: Repbasesubmitter and censor. BMC bioinformatics 7(1), 1–7 (2006)
109. Guo, R., Li, Y.-R., He, S., Ou-Yang, L., Sun, Y., Zhu, Z.: Replong: de novo repeat identification using long read sequencing data. Bioinformatics 34(7), 1099–1107 (2018)
110. Lee, H., Lee, M., Mohammed Ismail, W., Rho, M., Fox, G.C., Oh, S., Tang, H.: Mgescan: a galaxy-based system for identifying retrotransposons in genomes. Bioinformatics 32(16), 2502–2504 (2016)
111. Xu, Z., Wang, H.: Ltr_finder: an efficient tool for the prediction of full-length ltr retrotransposons. Nucleic acids research 35(suppl_2), 265–268 (2007)
112. Valencia, J.D., Girgis, H.Z.: Ltrdetector: a tool-suite for detecting long terminal repeat retrotransposons de-novo. BMC genomics 20(1), 450 (2019)
113. Steinbiss, S., Willhoeft, U., Gremme, G., Kurtz, S.: Fine-grained annotation and classification of de novo predicted ltr retrotransposons. Nucleic acids research 37(21), 7002–7013 (2009)
Figures
Tables

**Table 1 Overview of common transposon annotation tools.** 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.

| Name          | Approach | No. | Struc. | Simil. | Class I | Class II |
|---------------|----------|-----|--------|--------|---------|----------|
|               |          |     |        |        | LTR     | LINE     |
|               |          |     |        |        | SINE    | TIR      |
|               |          |     |        |        | HEL     | MITE     |
| RepeatMasker  | x        | x   | x      | x      | x       | x        |
| RepeatModeler | x        |     | x      | x      | x       | x        |
| CLARI_TE [107]| x        | x   | x      | x      | x       | x        |
| TESeeker [41] | x        | x   | x      | x      | x       | x        |
Additional Files
File F1: TransposonDB.fasta
File F2: NCBICDD1000_Proteins.txt
File F3: Classification_FeatureImportanceAnalysis.csv
File F4: GFF3 files in "PaperSupplements/Annotation/..."
File F5: GFF3 files in "PaperSupplements/Detection/..."
File F6: Detection_SVDistribution.csv
File F7: Detection_PipelineData.csv
File F8: Detection_ClassDistribution.csv
Supplements
Table S1 Classification: TransposonDB filter rule application. This table shows the number of remaining sequences in the constituents databases of TransposonDB after the application of the different filter rules.

| database    | before filtering | #F0 | #F1_1 | #F1_2 | #F1_3 | #F1_4 | #F1 | #F2 | #F3 | #F4 |
|-------------|------------------|-----|-------|-------|-------|-------|-----|-----|-----|-----|
| ConTEdb     | 322,705          | 322,705 | 322,702 | 322,702 | 322,702 | 322,702 | 322,702 | 317,109 | 317,995 | 312,402 |
| DPTEdb      | 31,340           | 31,340 | 31,326 | 31,326 | 31,326 | 31,326 | 31,326 | 31,285 | 31,285 | 31,285 |
| PCSB        | 61,730           | 61,730 | 61,729 | 61,729 | 61,529 | 60,813 | 55,616 | 60,757 | 61,397 | 54,488 |
| MnTEDb      | 5,925            | 5,925 | 5,912 | 5,912 | 5,912 | 5,912 | 5,912 | 5,902 | 5,902 | 5,902 |
| PMITEdb     | 2,449,127        | 2,449,127 | 357,134 | 313,211 | 313,211 | 313,211 | 313,211 | 298,777 | 302,529 | 288,104 |
| RepBase     | 56,403           | 56,403 | 53,880 | 51,674 | 51,674 | 51,648 | 51,674 | 51,455 | 37,796 | 37,563 |
| RiTE        | 265,549          | 265,549 | 264,022 | 242,216 | 156,195 | 241,847 | 240,509 | 154,232 | 152,746 | 152,575 |
| Soyetedb    | 38,664           | 38,664 | 38,603 | 38,603 | 38,603 | 38,603 | 38,603 | 38,519 | 38,601 | 38,517 |
| SPTEDdb     | 18,413           | 18,413 | 18,408 | 18,408 | 18,408 | 18,408 | 18,408 | 18,402 | 18,402 | 18,402 |
| TrepDB      | 4,162            | 4,162 | 3,910 | 1,874 | 3,910 | 3,910 | 1,870 | 3,669 | 3,869 | 1,822 |
| TransposonDB|                 |       |       |       |       |       |       |       |       | 891,051 |

This table shows the number of remaining sequences in the constituents databases of TransposonDB after the application of the different filter rules.
### (A) Proposed transposon classification taxonomy

| Class | Taxonomy / hierarchy | Notes on constituents |
|-------|----------------------|-----------------------|
| 1     | Class I (Retrotransposons) | Copia, Gypsy, Bel-Pao, Retrovirus, ERV |
| 1/1   | LTR                  | Copia |
| 1/1/1 |                     | Gypsy |
| 1/1/2 |                     | ERV   |
| 1/2   | Non-LTR             | DIRs (VIPER, Ngaro), LINEs, SINEs |
| 1/2/1 | LINE                | R2, RTE, Jockey, L1, I, Randl, Penelope, DRE… |
| 1/2/2 | SINE                | tRNA, 7SL, 5S |

### (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) | |
| Poly(A) tail | x | x |

| Protein Features | Class I | Class II |
|------------------|---------|----------|
| Helicase | x | (x) |
| Capsid protein (GAG) | x | (x) |
| RPA-like (RAPl) 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) | |
| RnaseH (RH) | |

Figure 1 Transposon taxonomy and transposon structure. (A) The taxonomy used in this study is based on multiple classification schemes [49], [36], [106], [3] 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.
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.

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.
Figure 4 Evaluation of the RFSB classifier. (A) Benchmark of different transposon classifiers by reported numbers in publications. (B) The performance of selected, reproducible classifiers applied on RepBase+PGSB database using the taxonomy in Fig. 1 A. 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 tenfold cross validation.

| databases       | ConTEdb | DPTEdb | mipsREdat-PGSB | PfamTEdb | PMITEdb | RepBase23.08 | RIPE | Soytedb | SPTEDdb | TrepDB |
|-----------------|---------|--------|----------------|----------|---------|---------------|------|----------|---------|--------|
| Eukaryota       | 312,334 | 29,452 | 52,691         | 3,154    | 282,500 | 36,681        | 129,911| 36,429   | 6,733   | 1,068  |
| Animalia        | 0       | 0      | 0              | 0        | 0       | 21,854        | 0    | 0        | 0       | 20     |
| Chromista       | 0       | 0      | 0              | 0        | 0       | 256           | 988  | 0        | 0       | 74     |
| Fungi           | 0       | 0      | 0              | 0        | 0       | 0             | 1,961| 0        | 0       | 354    |
| Plantae         | 312,334 | 29,452 | 52,691         | 3,154    | 282,244 | 11,721        | 129,911| 36,429   | 6,733   | 620    |
| Protozoa        | 0       | 0      | 0              | 0        | 0       | 150           | 0    | 0        | 0       | 0      |
| UnicellularFlagellate | 0     | 0      | 0              | 0        | 0       | 7             | 0    | 0        | 0       | 0      |
| Prokaryota      | 0       | 0      | 0              | 0        | 0       | 54            | 0    | 0        | 0       | 6      |
| Virus           | 0       | 0      | 0              | 0        | 0       | 31            | 0    | 0        | 0       | 0      |
| **Total**       | 312,334 | 29,452 | 52,691         | 3,154    | 282,500 | 36,766        | 129,911| 36,429   | 6,733   | 1,074  |

Table S2 Classification: TransposonDB sequences across biological domains. This table shows the number of sequences in TransposonDB by the source database and biological domains.
Figure 5 reason TE results for three species. The colors used in this figure represent *Caenorhabditis elegans* (red), *Oryza sativa subsp. 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% to 75% percentiles, including the orange bar at the 50% percentile. The length of whiskers amounts to 150% of the interquartile range.
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].
Figure S1 Classification: feature design experiments. All performance measures reported as average across a tenfold cross validation (in %) on RepBase+PGSB dataset for a classifier distinguishing class I and II transposons. (A) Performance of standard classifiers using the binary protein feature for different e-thresholds. This feature is either zero or one depending on whether the protein domain is detected for given e-threshold by BLASTN. (B) Performance of standard classifiers using the numerical protein features for different e-thresholds. This feature represents the number of times the protein domain is detected for given e-threshold by BLASTN. (C) Combinations of the protein features with relative k-mer frequency for a random forest classifier.
Figure S2 Classification: structure, strategy and model design experiments. All performance measures reported as average across a tenfold cross validation (in %) on RepBase+PGSB from an overall perspective. Panel (A) exhibits the performance of standard classifiers for different model structure and training strategy combinations. Panel (B) shows the performance of SVM classifier model for different kernel functions. Panel (C) summarises the performances of standard classifiers and best SVM classifier for the selective binary (Sel-Bin) combination. Panel (D) compares the proposed “RFSB” (Random Forest Selective Binary) approach with existing benchmark classifiers. In addition, RFSB trained on TransposonDB is reported as well.
Figure S3 Annotation: transposon annotation density plots. This diagram shows the density of transposon annotations across the chromosomes of the different *Caenorhabditis elegans* and the *Oryza sativa subs. japonica* genomes.
Figure S4 Annotation: tool intersection heat map. This heat map presents the intersection of annotations from different tools. The content of cell in row i and column j represents the number of intersecting basepairs of annotations from tools i and j, divided by the total number of basepairs in annotations of tool row i. Numbers are reported in percent, as average across the three reference genomes VC2010, N2 and CB4856. Repeat annotations are the combination of annotated repeats by RepeatMasker and RepeatModeler. Protein annotations are the combination of annotated transposon characteristic proteins by NCBICDD1000 and TransposonPSI. The colour of a cell represents its value. Darker colours represent values closer to 100%, while lighter colours represent values closer to 0%.
Figure S5 Annotation: tool class heat map. This heat map presents the share of annotated transposons by different transposon classes. The content of a cell represents the number of basepairs of a specific tool’s annotations related to a specific class divided by the total genomes length. Numbers are reported in percent, as average across the three reference genomes VC2010, N2 and CB4856. The colour of a cell represents its value. The darker the colours, the more a tool was able to capture most of the transposons that the ensemble (reasonaTE) found for this specific class.
### Table S3: Classification: TransposonDB sequences across eukaryotic kingdoms.

This table shows the number of sequences in TransposonDB by the source database and biological kingdoms.

| Kingdoms          | ConTeDb | DPTEdb | mipsREdat-PGSB | MinTEdb | PMITEdb | RepBase-23.08 | RITE | Soytedb | SPTEDd | TrepDB |
|-------------------|---------|--------|----------------|---------|---------|----------------|------|---------|---------|--------|
| Animalia          | 0       | 0      | 0              | 0       | 0       | 21,854         | 0    | 0       | 0       | 0      |
| Annelida          | 0       | 0      | 0              | 0       | 0       | 287            | 0    | 0       | 0       | 0      |
| Anthropoda        | 0       | 0      | 0              | 0       | 0       | 8,030          | 0    | 0       | 0       | 0      |
| Antropoda         | 0       | 0      | 0              | 0       | 0       | 1              | 0    | 0       | 0       | 0      |
| Ascomycota        | 0       | 0      | 0              | 0       | 0       | 4              | 0    | 0       | 0       | 0      |
| Bilateria         | 0       | 0      | 0              | 0       | 0       | 1              | 0    | 0       | 0       | 0      |
| Brachiopoda       | 0       | 0      | 0              | 0       | 0       | 1              | 0    | 0       | 0       | 0      |
| Chordata          | 0       | 0      | 0              | 0       | 0       | 10,182         | 0    | 0       | 0       | 0      |
| Cnidaria          | 0       | 0      | 0              | 0       | 0       | 1,228          | 0    | 0       | 0       | 0      |
| Ctenophora        | 0       | 0      | 0              | 0       | 0       | 31             | 0    | 0       | 0       | 0      |
| Deuterostomia     | 0       | 0      | 0              | 0       | 0       | 26             | 0    | 0       | 0       | 0      |
| Echinodermata     | 0       | 0      | 0              | 0       | 0       | 221            | 0    | 0       | 0       | 0      |
| Hemichordata      | 0       | 0      | 0              | 0       | 0       | 37             | 0    | 0       | 0       | 0      |
| Mollusca          | 0       | 0      | 0              | 0       | 0       | 749            | 0    | 0       | 0       | 0      |
| Nematoda          | 0       | 0      | 0              | 0       | 0       | 574            | 0    | 0       | 0       | 0      |
| Placozoa          | 0       | 0      | 0              | 0       | 0       | 8              | 0    | 0       | 0       | 0      |
| Platyhelminthes   | 0       | 0      | 0              | 0       | 0       | 343            | 0    | 0       | 0       | 0      |
| Porifera          | 0       | 0      | 0              | 0       | 0       | 7              | 0    | 0       | 0       | 0      |
| Priapulida        | 0       | 0      | 0              | 0       | 0       | 1              | 0    | 0       | 0       | 0      |
| Rotifera          | 0       | 0      | 0              | 0       | 0       | 123            | 0    | 0       | 0       | 0      |
| Chromista         | 0       | 0      | 0              | 0       | 0       | 256            | 988  | 0       | 0       | 74     |
| Ciliophora        | 0       | 0      | 0              | 0       | 0       | 11             | 0    | 0       | 0       | 0      |
| Ciliophora        | 0       | 0      | 0              | 0       | 0       | 3              | 0    | 0       | 0       | 0      |
| Haptophyta        | 0       | 0      | 0              | 0       | 0       | 15             | 0    | 0       | 0       | 0      |
| Myzozoa           | 0       | 0      | 0              | 0       | 0       | 104            | 0    | 0       | 0       | 7      |
| Ochrophyta        | 0       | 0      | 0              | 0       | 0       | 92             | 0    | 0       | 0       | 0      |
| Oomycota          | 0       | 0      | 0              | 0       | 0       | 256            | 760  | 0       | 0       | 40     |
| Orchophyta        | 0       | 0      | 0              | 0       | 0       | 3              | 0    | 0       | 0       | 0      |
| Fungi             | 0       | 0      | 0              | 0       | 0       | 1,961          | 0    | 0       | 0       | 354    |
| Ascomycota        | 0       | 0      | 0              | 0       | 0       | 749            | 0    | 0       | 0       | 354    |
| Basidiomycota     | 0       | 0      | 0              | 0       | 0       | 1,111          | 0    | 0       | 0       | 0      |
| Blastocladiomycota| 0       | 0      | 0              | 0       | 0       | 14             | 0    | 0       | 0       | 0      |
| Chytridiomycota   | 0       | 0      | 0              | 0       | 0       | 15             | 0    | 0       | 0       | 0      |
| Mucoromyceta      | 0       | 0      | 0              | 0       | 0       | 9              | 0    | 0       | 0       | 0      |
| Mucoromycota      | 0       | 0      | 0              | 0       | 0       | 71             | 0    | 0       | 0       | 0      |
| Plantae           | 312,334 | 29,452 | 52,691         | 3,154   | 282,244 | 11,721         | 129,911 | 36,429  | 6,733  | 620    |
| Angiosperms       | 0       | 29,452 | 48,583         | 3,154   | 257,898 | 9,997          | 129,911 | 36,429  | 6,733  | 618    |
| Bryophyta         | 0       | 1,060  | 0              | 0       | 45      | 0              | 0       | 0       | 0       | 2      |
| Chlorophyta       | 0       | 0      | 0              | 105     | 0       | 106            | 0      | 0       | 0       | 0      |
| Rhodophyta        | 0       | 0      | 0              | 0       | 0       | 595            | 0      | 0       | 0       | 0      |
| Tracheophyta      | 312,334 | 0      | 3,047          | 0       | 24,241  | 978            | 0      | 0       | 0       | 0      |
| Protozoa          | 0       | 0      | 0              | 0       | 0       | 150            | 0      | 0       | 0       | 0      |
| Amoebozoa         | 0       | 0      | 0              | 0       | 0       | 78             | 0      | 0       | 0       | 0      |
| Eozona            | 0       | 0      | 0              | 0       | 0       | 2              | 0      | 0       | 0       | 0      |
| Euglenozoa        | 0       | 0      | 0              | 0       | 0       | 14             | 0      | 0       | 0       | 0      |
| Metamonada        | 0       | 0      | 0              | 0       | 0       | 42             | 0      | 0       | 0       | 0      |
| Perozolozoa       | 0       | 0      | 0              | 0       | 0       | 14             | 0      | 0       | 0       | 0      |

Table S3: Classification: TransposonDB sequences across eukaryotic kingdoms. This table shows the number of sequences in TransposonDB by the source database and biological kingdoms.
### Table S4: Classification: TransposonDB sequences across transposon classes.

This table shows the number of sequences in TransposonDB by the source database and transposon classes.

| database | TransposonDB | ConTEdb | DPTEdb | mipsREdat-PGSB | MnTEdb | PMITEdb | RepBase23.08 | RITE | Soyetedb | SPTEDdb | TrepDB |
|----------|--------------|---------|--------|----------------|---------|----------|--------------|------|----------|---------|--------|
| 1        | 412,975      | 252,236 | 23,832 | 49,853         | 1,280   | 0        | 25,908       | 23,508 | 30,538   | 4,972   | 848    |
| 1.1      | 360,618      | 210,831 | 22,343 | 48,186         | 1,267   | 0        | 21,542       | 20,670 | 30,404   | 4,769   | 606    |
| 1.1.1    | 122,715      | 81,329  | 5,010  | 11,059         | 600     | 0        | 6,423        | 4,277  | 12,482   | 1,316   | 219    |
| 1.1.2    | 144,221      | 70,605  | 9,694  | 18,497         | 430     | 0        | 9,218        | 14,490 | 17,922   | 3,026   | 339    |
| 1.1.3    | 8,118        | 4,087   | 651    | 0              | 0       | 0        | 3,276        | 56    | 0        | 48      | 0      |
| 1.2      | 51,943       | 41,405  | 1,489  | 1,253          | 13      | 0        | 4,366        | 2,838  | 134      | 203     | 242    |
| 1.2.1    | 47,328       | 40,679  | 1,440  | 931            | 13      | 0        | 3,418        | 272    | 134      | 199     | 242    |
| 1.2.2    | 3,640        | 0       | 0      | 322            | 0       | 0        | 753          | 2,565  | 0        | 0       | 0      |
| 2        | 478,070      | 60,098  | 5,620  | 2,838          | 1,874   | 0        | 282,500      | 10,859 | 106,403  | 5,891   | 1,761  |
| 2.1      | 397,819      | 131     | 468    | 1,416          | 1,752   | 0        | 282,500      | 7,791  | 97,710   | 5,809   | 59     |
| 2.1.1    | 92,563       | 5       | 41     | 261            | 0       | 0        | 86,195       | 2,078  | 2,301    | 1,645   | 7      |
| 2.1.2    | 94,929       | 5       | 144    | 228            | 135     | 0        | 918          | 93,494 | 0        | 5       | 0      |
| 2.1.3    | 19,335       | 25      | 108    | 77             | 1,084   | 0        | 15,076       | 2,376  | 487      | 65      | 22     |
| 2.1.4    | 35,981       | 0       | 0      | 0              | 33,537  | 0        | 13           | 2,370  | 0        | 61      |       |
| 2.1.5    | 145,791      | 4       | 93     | 139            | 285     | 0        | 142,679      | 755    | 131      | 1,664   | 8      |
| 2.1.6    | 7,552        | 0       | 0      | 711            | 0       | 4,996    | 467          | 1,280  | 65       | 0       | 33     |
| 2.2      | 56,786       | 47,199  | 4,652  | 14             | 4       | 0        | 665          | 2,458  | 82       | 1,669   | 43     |
| 2.3      | 20,129       | 12,767  | 500    | 475            | 118     | 0        | 2            | 6,235  | 0        | 32      | 0      |
| **Total**| **891,045**  | **312,334** | **29,452** | **52,691** | **3,154** | **282,500** | **36,767** | **129,911** | **36,429** | **6,733** | **1,074** |

Table S4: Classification: TransposonDB sequences across transposon classes. This table shows the number of sequences in TransposonDB by the source database and transposon classes.
| Constituents                      | NCBI CDD ID                      |
|----------------------------------|----------------------------------|
| Aspartic proteinase              | cd00303, cd05481, cd05484        |
| Apurinic endonuclease            | tigr00587                        |
| Integrase (core domain)          | pfam0665, cog3335, pfam13358, pfam01359 |
| GAG pre-integrase               | pfam13976                        |
| Ribosomal-processing proteinase  | cd16332, prk14553                |
| Cysteine proteinase              | tigr01586                        |
| Peptidase (Prp)                  | pfam04327, pfam04231, cog4636, pfam05685, cog2356, pfam01844, pfam05551, pfam07510, pfam13391, pfam13392, pfam13395, pfam14414, prk15137, cd00719, smart00478, cog0648, prk01060, smart00518, prk02308, pfam04493, pfam08459 |
| Endonuclease                     | pfam03732, pfam16297              |
| Helicase                         | prk13767, tigr04121, pfam06733, pfam00270, pfam00271, pfam04851, pfam05970, pfam14617, pfam13307 |
| DNA polymerase                   | cd08637, cd08638, cd08639, cd08640, cd08641, cd08642, cd08643 |
| RNAse H                          | cd06266, cd09272, cd09273, cd09274, cd09275, cd09276, cd09279, prk06863, prk06751, prk06752, prk08182, prk06293, prk06461, prk06958, prk07274, prk10053, smart00976, cog0629, cog2965, cog3111, prk05733, cog4085, cog3390, prk06341, prk09010, pfam00436, tigr00621, pfam02765, pfam04057, pfam08646, pfam16686, pfam09104, pfam09103, pfam08661, pfam16900, pfam13742 |
| Replication protein A            | pfam13966, pfam07727, pfam00078, pfam11474, tigr04416, pfam13655, pfam17984, pfam17919, pfam17917, pfam13456, pfam06817, pfam06815, cog3344, cd03715, cd03714, cd03487, cd01709, cd01699, cd01651, cd01648, cd01647, cd01646, cd01644, cd05471, nf033179, pfam13006, pfam14706, pfam02281, pfam13701, pfam13007, pfam13005, pfam04986, pfam03050, pfam01610, pfam01609, pfam01548, pfam01526, pfam18759, pfam18758, pfam17906, pfam13751, pfam13612, cd01187, cd01186, pfam11427, pfam02371, pfam01797, pfam1373, pfam13586, pfam13359, pfam13308, pfam13613 |
| Reverse transcriptase            | pfam01196, cd01195, cd01194, cd01192, cd01191, cd01184, cd01188, tigr02224, prk02436, prk00283, cd00796, cd00397, cd00799, cd06094, pfam03564, pfam05380, pfam05585, pfam08284, pfam13975, pfam14223, pfam14244, pfam03184 |

Table S5 Classification: selection of protein domains: This table lists the selected NCBI CDD PSSM model IDs considered for the protein features used in the classification module.
Table S6 Annotation: case study genomes. This diagram shows source, species, sequencing technology, number of sequences, length in bp, strain name and location for the 25 case study genomes.
### Table S7 Classification: benchmark of classifiers

All performance measures reported as average across 10 folds (in %) are supplemented by the standard deviations in brackets (in %). Bold numbers mark the best performance amongst different classifiers within same category. Panel (A) outlines performance measures of the benchmark algorithms reported in their publications (meaning these results were gathered from different datasets and taxonomies, depending on the specific publication). Panel (B) outlines performance measures of several benchmark algorithms to the proposed “RFSB” classifier methodology. All results were calculated based on the same dataset RepBase+PGSB and the same, proposed taxonomy. The measures are reported for taxonomic levels and overall perspective. In addition, the proposed “RFSB” classifier is applied to TransposonDB and reported in the most right column.

**Panel (A) Benchmark of classifiers, performances reported in publications of classifiers (different databases and taxonomies)**

|                | Random Even | Random Gaussian | HC_GA | HC_LGA | NLLCPN | TERL | TopDown | RFSB | (TransposonDB) | RFSB |
|----------------|-------------|-----------------|-------|--------|--------|------|---------|------|----------------|------|
| **Overall**    |             |                 |       |        |        |      |         |      |                |      |
| ACC            | 47.58%      | 47.39%          | 83.00%| 84.00% | 90.00% | 85.80%| 83.00%  | 91.34%| 93.75%         | 93.28%|
| (0.27%)        | (0.38%)     | (0.26%)         |       |        |        |      |         |      |                |      |
| F1             | 31.03%      | 30.44%          | 77.91%| 74.02% | 70.84% | 7.11% | 24.43%  | 91.34%| 89.66%         | 86.45%|
| (0.32%)        | (0.26%)     | (0.26%)         |       |        |        |      |         |      |                |      |
| MCC            | -11.14%     | -11.71%         | 65.99%| 59.45% | 57.52% | -54.45%| 24.43%  | 91.34%| 84.68%         | 85.44%|
| (0.54%)        | (0.63%)     | (0.63%)         |       |        |        |      |         |      |                |      |
| **Level 1**    |             |                 |       |        |        |      |         |      |                |      |
| ACC            | 79.23%      | 79.52%          | 92.86%| 91.78% | 92.32% | 82.18%| 77.99%  | 96.48%| 96.11%         | 96.11%|
| (0.09%)        | (0.07%)     | (0.06%)         |       |        |        |      |         |      |                |      |
| F1             | 29.52%      | 29.14%          | 75.00%| 71.17% | 70.10% | 0.10% | 23.05%  | 91.34%| 89.66%         | 87.74%|
| (0.3%)         | (0.23%)     | (0.23%)         |       |        |        |      |         |      |                |      |
| MCC            | 17.35%      | 17.18%          | 70.92%| 66.44% | 66.70% | -7.36%| 10.31%  | 96.48%| 85.44%         | 85.44%|
| (0.36%)        | (0.26%)     | (0.26%)         |       |        |        |      |         |      |                |      |
| **Level 2**    |             |                 |       |        |        |      |         |      |                |      |
| ACC            | 82.60%      | 82.91%          | 93.63%| 92.72% | 93.26% | 89.21%| 82.33%  | 96.99%| 96.67%         | 96.67%|
| (0.09%)        | (0.05%)     | (0.05%)         |       |        |        |      |         |      |                |      |
| F1             | 17.57%      | 17.13%          | 68.51%| 63.70% | 60.18% | 0.00% | 12.89%  | 96.99%| 95.30%         | 85.66%|
| (0.38%)        | (0.23%)     | (0.23%)         |       |        |        |      |         |      |                |      |
| MCC            | 7.85%       | 7.62%           | 65.12%| 59.89% | 59.21% | -1.01%| 3.21%   | 84.65%| 83.81%         | 83.81%|
| (0.43%)        | (0.24%)     | (0.24%)         |       |        |        |      |         |      |                |      |
| **Level 3**    |             |                 |       |        |        |      |         |      |                |      |
| ACC            | 86.30%      | 86.57%          | 92.87%| 91.69% | 90.91% | 90.90%| 86.96%  | 97.33%| 97.76%         | 97.76%|
| (0.05%)        | (0.04%)     | (0.04%)         |       |        |        |      |         |      |                |      |
| F1             | 9.63%       | 9.32%           | 49.93%| 38.67% | 0.00%  | 0.00% | 7.96%   | 84.95%| 87.51%         | 87.51%|
| (0.41%)        | (0.34%)     | (0.34%)         |       |        |        |      |         |      |                |      |
| MCC            | 2.59%       | 2.56%           | 48.61%| 37.48% | 0.00%  | -0.21%| 2.28%   | 83.51%| 86.29%         | 86.29%|
| (0.42%)        | (0.34%)     | (0.34%)         |       |        |        |      |         |      |                |      |
Table S8 Detection: Number of observed transposition events. This table shows the number of observed transposition events for different probe reference genome combinations, alignment and structural variant calling tools.

| strain        | PBSV (pbmm2) | Sniffles (ngmrl) | PBSV (pbmm2) | Sniffles (ngmrl) |
|---------------|--------------|------------------|--------------|------------------|
| SX3383 (QX1791) | 665          | 848              | 295          | 289              |
| SX3399 (GXW1)  | 384          | 546              | 278          | 27               |
| SX3356 (QX1211) | 862          | 831              | 371          | 297              |
| SX3351 (DL200) | 354          | 390              | 260          | 237              |
| SX3368 (AB1)   | 262          | 288              | 284          | 233              |
| SX3287 (PS2025) | 521          | 571              | 257          | 210              |
| SX3304 (EG4725) | 528          | 417              | 268          | 185              |
| SX3306 (ECA36) | 927          | 837              | 372          | 299              |
| SX3254 (N2)    | 13           | 10               | 246          | 212              |
| SX3292 (CB4856) | 670          | 744              | 2            | 28               |
| SX3302 (JU394) | 208          | 252              | 273          | 234              |
| SX3303 (JT11398) | 275         | 309              | 262          | 220              |
| SX3298 (JU775) | 494          | 509              | 271          | 55               |
| SX3295 (LKC34) | 378          | 336              | 261          | 203              |
| SX3367 (DL238) | 671          | 861              | 279          | 287              |
| SX3305 (ECA248) | 454          | 411              | 271          | 225              |
| SX3296 (MY23)  | 555          | 472              | 279          | 187              |
| SX3286 (EG4946) | 277          | 272              | 270          | 214              |
| SX3350 (PB306) | 458          | 540              | 284          | 258              |
| SX3341 (JU751) | 431          | 424              | 263          | 192              |
Table S9 Detection: Genetic distance and number of transposition events found. This table shows the number of observed transposition event candidates, the length of their mask in bp, and the phylogenetic distance of probe and reference genome.

| Strain | Total BP VC2010 CB4856 | Total BP VC2010 CB4856 | Total BP VC2010 CB4856 | Total BP VC2010 CB4856 | Phylogenetic distance VC2010 CB4856 |
|--------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------------------|
| AB1    | 288 233 2528558 864269  | 262 284 450579 907524   | 0.023819 0.094459       |
| CB4856  | 744 28 3508318 86028    | 670 2 1350133 11185     | 0.096336 0.07201       |
| DL200  | 390 237 1260330 918529   | 354 260 737484 613060   | 0.078775 0.144549       |
| DL238  | 861 287 5340946 1120948  | 671 279 1635629 701238  | 0.100631 0.077201       |
| ECA248 | 411 225 1674096 828618   | 454 271 861328 978863   | 0.193356 0.185248       |
| ECA36  | 837 299 5547803 932049   | 927 372 2241605 779615   | 0.133636 0.104196       |
| EG4725 | 417 185 1116717 976099   | 528 268 1077970 542523   | 0.063796 0.104196       |
| EG4946 | 272 214 948485 1126127   | 277 270 839962 797286   | 0.036709 0.093291       |
| GXW1   | 546 27 1990299 42255     | 384 278 784283 62929    | 0.036727 0.096107       |
| JT11398| 309 220 86683 1137900    | 275 262 557395 664228   | 0.027652 0.093426       |
| JU394  | 252 234 1413635 120411   | 208 273 747156 795337   | 0.014571 0.109029       |
| JU751  | 424 192 402712 880873    | 431 263 971156 842471   | 0.043152 0.09734        |
| JU775  | 509 55 2324057 198557    | 494 271 1175322 589865  | 0.115199 0.165997       |
| LKC34  | 336 203 935153 763459    | 378 261 909664 710945   | 0.037956 0.088754       |
| MY23   | 472 187 2079515 480242   | 555 279 1227010 500555  | 0.08266 0.139242       |
| MY72   | 10 212 39344 576541      | 13 246 20414 520726     | 0 0.096336              |
| PB306  | 540 258 2226567 1042300  | 458 284 854083 735233   | 0.056832 0.113414       |
| PS205  | 571 210 3381679 856256   | 521 257 1018228 738095  | 0.071363 0.127945       |
| QX1211 | 831 297 3556466 936120   | 862 371 1895254 723441  | 0.194891 0.186783       |
| QX1791 | 848 289 5571466 1232941  | 665 295 1521737 860000  | 0.105931 0.097823       |