PanTE: A Comprehensive Framework for Transposable Element Discovery in Graph-based Pangenomes

Preprint · February 2025

DOI: 10.21203/rs.3.rs-5867196/v1

12 authors, including:

Yiwen Wang
Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences
22 PUBLICATIONS **501** CITATIONS
SEE PROFILE

Cao Shuo
Huazhong Agricultural University
35 PUBLICATIONS **527** CITATIONS
SEE PROFILE

Zhenya Liu
Chinese Academy of Agricultural Sciences
18 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Hua Xiao
Chinese Academy of Agricultural Sciences
35 PUBLICATIONS **321** CITATIONS
SEE PROFILE

PanTE: A Comprehensive Framework for Transposable Element Discovery in Graph-based Pangenomes

Yiwen Wang

yiwen.wang@unimelb.edu.au

Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne
https://orcid.org/0000-0002-7067-9093

Shuo Cao

Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center

Zhenya Liu

Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

Yuting Liu

Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Zhongqi Liu

College of Plant Protection, Hunan Agricultural University

Wenqi Ma

Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

Jianzhong Lu

Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

Hua Xiao

Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Jinfeng Chen

Institute of Zoology https://orcid.org/0000-0002-5628-6322

Shujun Ou

Department of Molecular Genetics, Ohio State University

Erik Garrison

University of Tennessee Health Science Center https://orcid.org/0000-0003-3821-631X

Yongfeng Zhou

Chinese Academy of Agricultural Sciences, Shenzhen https://orcid.org/0000-0003-0780-2973

Brief Communication

Keywords:

Posted Date: February 5th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5867196/v1

PanTE: A Comprehensive Framework for Transposable Element Discovery in Graph-based Pangenomes

Shuo Cao[1, 2, †], Zhenya Liu[1, †], Yuting Liu[1], Zhongqi Liu[1], Wenqi Ma[1], Jianzhong Lu[1], Hua Xiao[1], Jinfeng Chen[3], Shujun Ou[4], Erik Garrison[2], Yongfeng Zhou[1, 5, *], Yiwen Wang[1, 6, *]

[1]National Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China.

[2]Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA.

[3]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China.

[4]Department of Molecular Genetics, Ohio State University, Columbus, OH, USA.

[5]National Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101, China.

[6]Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC, Australia.

[†]These authors contributed equally: Shuo Cao and Zhenya Liu.

*Corresponding author: Yiwen Wang (yiwen.wang@unimelb.edu.au) and Yongfeng Zhou (zhouyongfeng@caas.cn).

Abstract

Transposable element (TE) annotation is crucial for understanding genetics, genomics and evolution, yet current methods struggle to identify TEs in graph-based pangenomes. We developed a framework PanTE to construct accurate and representative TE libraries for both single genomes and graph pangenomes. PanTE is the first of its kind capable of being directly applied to graph-based pangenomes to build population-level TE libraries. By partially reimplementing RepeatModeler2 and integrating key innovations, including graph pangenome disassembly, alignment-free LTR structure detection, a machine learning-based classifier and efficiency-boosting strategies, PanTE outperformed RepeatModeler2 by efficiently handling large genomes, detecting high-abundance TEs and LTR-retrotransposons, and providing robust TE classification with superior computational efficiency. Compared to EDTA, it annotated ~ 26% more TEs in the grapevine genome and achieved up to 13 times faster runtimes in the wheat genome. PanTE represents a significant advancement in population-wide TE discovery, making it particularly valuable for pangenomic studies.

Keywords: Transposable element discovery, Graph-based pangenomes, Population-level TE libraries

Background

Transposable Elements (TEs) represent a substantial source of genetic variation and are key drivers of evolutionary processes due to their capacity to jump within the genome [1]. TEs can lead to gene inactivation or the generation of novel genes through disruption of coding regions [2], while also contributing to the regulation of gene expression and the maintenance of genome stability by modulating the three-dimensional organization of the genome [3]. Historically, the detection of TEs has been challenging because of their repetitive nature and substantial length. In the era of short-read sequencing, the technologies were limited in resolving the full extent of repetitive sequences, thus impeded the accurate assembly of TEs within the genome. Subsequently, the advent of long-read sequencing and the Telomere-to-Telomere (T2T) assembly approach have largely improved genome quality [4–6]. As a result, the primary challenge in TE research has now shifted towards the computational identification and annotation of these elements.

The genome-wide annotation of TEs involves their identification, classification and mapping across the entire genome to elucidate their locations and functions. The first two steps, collectively referred to as TE discovery, are implemented as: 1) detecting and constructing TE consensus sequences, and 2) classifying these sequences into distinct TE families. A widely used method, RepeatModeler2, relies heavily on k-mer and alignment-based methods, which are computationally intensive [7]. Additionally, the sampling size of RepeatModeler2 is restricted to 40 Kbp per time, with a maximum genome coverage of 363 Mbp through multiple samplings and repeated iterations [7]. This sampling constraint hinders the identification of long TE sequences, and the overall coverage is insufficient for

comprehensive TE scanning in large genomes, such as those of wheat and maize. Another prominent approach, Extensive de-novo TE Annotator (EDTA), generates a non-redundant TE library by filtering TEs identified through various programs that primarily rely on detecting contextual structures [8]. Consequently, its capacity to capture novel TEs with previously uncharacterized structures is inherently limited. While EDTA incorporates RepeatModeler2 to target TEs based on their repetitive nature, this step is applied only to unmasked regions after structure-based TE detection. If novel TEs are partially misclassified as known structures during earlier steps, the corresponding regions are masked, preventing RepeatModeler2 from processing them and reducing the likelihood of their identification. Thus, the effectiveness of EDTA in detecting novel TEs depends heavily on the accuracy of its initial structure-based detection steps.

In addition, due to rapid evolutionary dynamics, TEs exhibit substantial variability in sequences and structures across species [9] and even among different cultivars [10,11]. Consequently, TE discovery based on a single genome is inherently biased, underscoring the necessity of constructing population-level TE libraries. However, it is posing considerable challenges for population-level TE discovery, as there remains a notable lack of tools capable of detecting TEs in a graph-based pangenome, which integrates all genetic variations from individuals within a population [12–14]. A pragmatic approach involves filtering and combining individually constructed TE libraries into a pan-library, as exemplified by the panEDTA tool [15]. Nevertheless, this method may overlook or filter out low-frequency TEs, leading to false negatives. This omission can adversely impact subsequent genome annotation at both individual and population levels.

Results

To bridge this gap, we propose PanTE, a comprehensive framework specifically designed for the identification and classification of TEs by leveraging their intrinsic structural characteristics and repetitive nature. This framework enables the generation of population-level TE libraries, applicable to both individual genomes and graph-based pangenomes generated from the Minigraph-cactus [16] and PGGB [17] frameworks.

For single genomes (FASTA files), our framework performs a standard pre-processing step including data cleaning, sequence renaming and indexing. In contrast, for pangenomes (GFA files), the pre-processing step involves core segment selection and extension, which are deliberately designed to disassemble graph-based pangenomes while preserving sequence diversity and avoiding redundancy (Fig. 1a, 2h). The resulting sequences are then processed through two principal algorithms for TE detection: 1/ traditional detection based on k-mer and alignment using RepeatScout [18] and RECON [19], reimplemented from the RepeatModeler2 architecture, and 2/ signature-based detection using Look4LTRs [20]. Traditional algorithms detect TEs by leveraging their repetitive nature. We adopted the algorithm of RepeatScout to identify the youngest and most abundant TE families using k-mer and multiple alignments, while we adapted RECON to detect older TE families via cross-alignment. To complement these traditional approaches, PanTE also integrates the algorithm of Look4LTRs, which is specifically designed to detect Long Terminal Repeat (LTR) structures, an essential component of LTR-retrotransposons, which are the most abundant type of TEs in plant genomes, comprising up to 74.19% in species such as maize [6,21]. Look4LTRs depends on k-mer matching and a graph-based algorithm, making it

nearly alignment free and computationally efficient. The combination of these two types of detection algorithms in our framework targets the general structural and repetitive properties of TEs simultaneously, allowing for the identification of novel TEs without known specific structural signatures.

To generate high-quality TE libraries with high accuracy, complete genome coverage, and efficient process, we implemented a masking and iteration strategy. For each iteration of the RECON unit (see Methods section “TE detection”), the TE sequences identified by RepeatScout and Look4LTRs algorithms are pre-masked in the genome samples which serve as the input for RECON unit. Newly detected TEs in each iteration are similarly masked before subsequent iterations. Compared to RepeatModeler2, PanTE is featured with superior capability, with a maximum genome coverage of up to 5.6 Gbp per iteration and a sampling size of up to 200 Kbp per time. Through iterative processing, PanTE can efficiently handle large genomes exceeding 10 Gbp in size. In addition, we introduced a refinement approach in PanTE to filter out satellite DNAs and reduce false discoveries. This approach also constructs TE consensus sequences, which summarize the most common characteristics of the corresponding TE classes but do not exist as physical entities in nature. These computationally constructed sequences from different detection units are further filtered and aggregated to enhance representative capacity, reduce redundancy, and ensure greater accuracy in downstream classification (see Methods section “TE detection”).

To classify consensus sequences into TE families, PanTE integrates two types of algorithms: 1/ homology-based and 2/ machine learning-based. While homology-based

algorithms provide high accuracy in classification, they are constrained by the limited number of TE entries available in the reference databases. This limitation is further exacerbated by the fact that most reference databases are protein-based and therefore incapable of classifying TEs that lack coding regions. Machine learning algorithms therefore complement this by identifying a substantial number of TE sequences that remain unclassified by homology-based methods. The integration of these two types of algorithms maximizes both the scope and accuracy of classification, ensuring a comprehensive and reliable categorization of TE sequences.

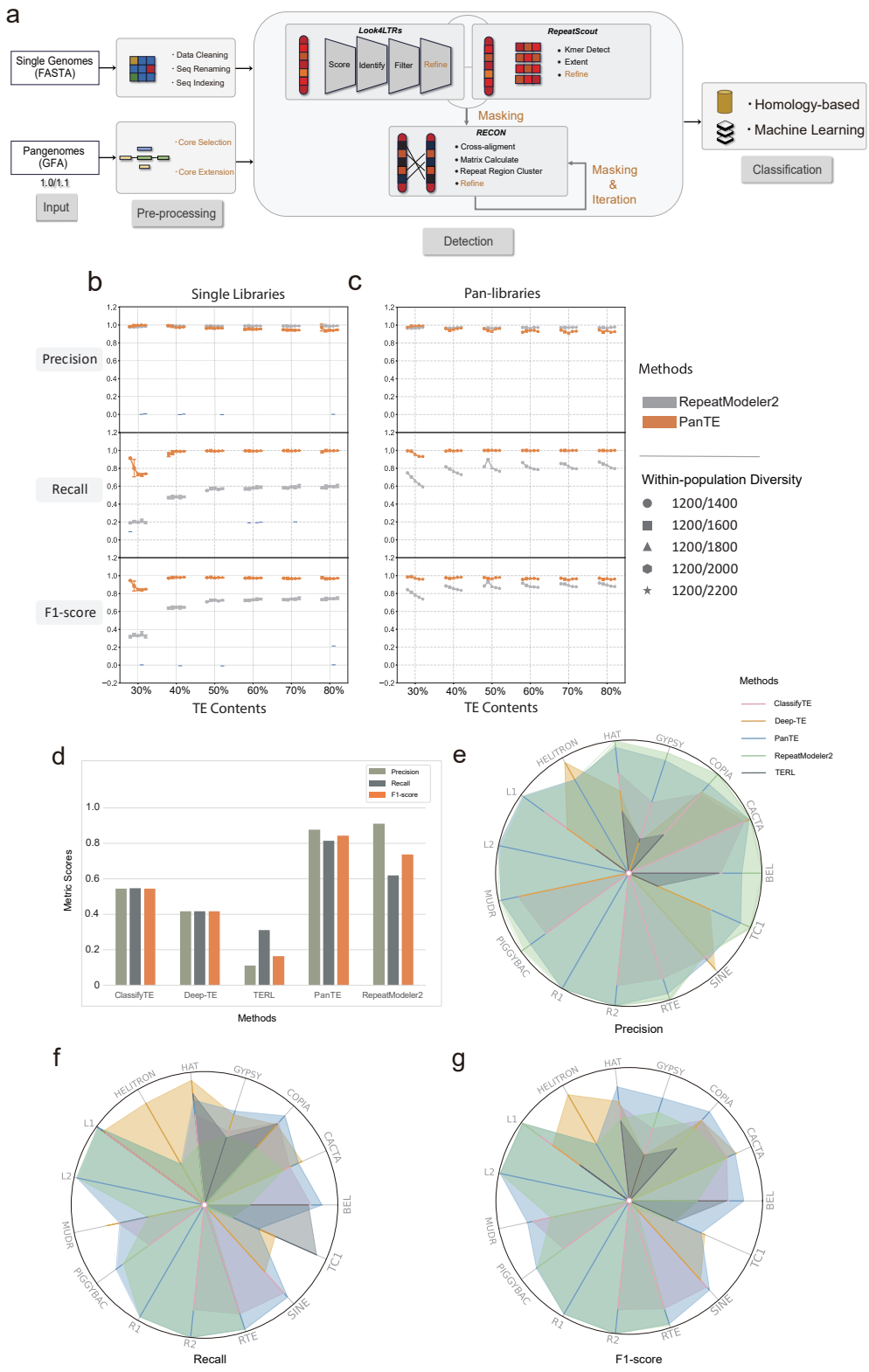
a
Single Genomes (FASTA)
Pangenomes (GFA)
1.0/1.1
Input
· Data Cleaning
· Seq Renaming
· Seq Indexing
· Core Selection
· Core Extension
Pre-processing
Look4LTRs
Score
Identify
Filter
Refine
RepeatScout
• Kmer Detect
• Extent
• Refine
Masking
RECON
• Cross-aligment
• Matrix Calculate
• Repeat Region Cluster
• Refine
Masking & Iteration
Detection
· Homology-based
· Machine Learning
Classification
b
Single Libraries
c
Pan-libraries
Precision
Recall
F1-score
TE Contents
Methods
RepeatModeler2
PanTE
Within-population Diversity
1200/1400
1200/1600
1200/1800
1200/2000
1200/2200
d
Metric Scores
Precision
Recall
F1-score
ClassifyTE
Deep-TE
TERL
PanTE
RepeatModeler2
Methods
e
Precision
f
Recall
g
F1-score
HAT
GYPSY
COPIA
CACTA
BEL
TC1
SINE
RTE
R2
R1
PIGGYBAC
MUDR
L2
L1
HELITRON

Fig. 1|Overview of PanTE and benchmarking against other methods on simulations. a, PanTE framework. The steps highlighted in brown are newly designed in this framework. **b-c**, Evaluation metrics for TE detection by RepeatModeler2 and PanTE on single TE libraries and pan-libraries integrated from individual libraires within each population. Each color represents a different method, and each shape represents a different level of within-population diversity. **d**, Evaluation metrics for TE classification globally. Different colors represent precision, recall and F1-score. **e-g**, Precision, Recall and F1-score for TE classification by ClassifyTE, Deep-TE, PanTE, RepeatModeler2 (RepeatClassifier), and TERL across 15 ground-truth families. Different colors represent different classification methods, and each node represents a TE family.

We benchmarked PanTE against RepeatModeler2 using simulated data and against both EDTA and RepeatModeler2 using real genomes with curated TE libraries from various species (see Methods). For TE classification, PanTE was evaluated alongside RepeatModeler2 [7] (RepeatClassifier), ClassifyTE [22], Deep-TE [23] and TERL [24], though this comparison was conducted exclusively on simulated data.

To evaluate TE detectability across methods, we simulated genomes embedded with known TE sequences from RepBase [25] (see Methods section "Simulations for TE detection"). EDTA relies on contextual structures of TEs, such as Target Site Duplications (TSDs) [26], which are challenging to simulate, and was therefore not included in our comparisons. When benchmarking PanTE to RepeatModeler2, for both single TE libraries derived from individual genomes and pan-libraries generated through population-level aggregation, PanTE outperformed RepeatModeler2 in recall and F1 score while

maintaining comparable precision (Fig. 1b and c, Supplementary Table 1). In single libraries, when TE content was as low as 30% and within-population diversity exceeded 1200/1600, indicating the presence of very low-frequency TEs, the recall of PanTE decreased dramatically. However, even in these challenging scenarios, PanTE still achieved substantially higher recall than RepeatModeler2. For RepeatModeler2, we observed high recall when TE content was substantial (in single libraries) and within-population diversity was low (in pan-libraries), suggesting its limitation in detecting low-frequency TEs. While PanTE shares some of these limitations, it demonstrates a greater capacity for detecting low-frequency TEs compared to RepeatModeler2.

When comparing TE classification performance across methods, PanTE thoroughly outperformed the others, achieving the highest overall Recall, F1 score, and a competitive Precision against RepeatModeler2 (RepeatClassifier; Fig. 1d, Supplementary Table 2). While Deep-TE demonstrated superior performance in certain TE families, such as Helitron and CACTA, its performance was suboptimal in other families. For comprehensive performance across all 15 TE families, as displayed by the areas in radar plots (Fig. 1e-g, Supplementary Table 3), PanTE surpassed the other methods with higher recall and F1 scores. RepeatModeler2, which classified a considerable number of sequences as “unknown” or into families outside the 15 ground-truth families, achieved slightly higher precision than PanTE but at the cost of lower recall.

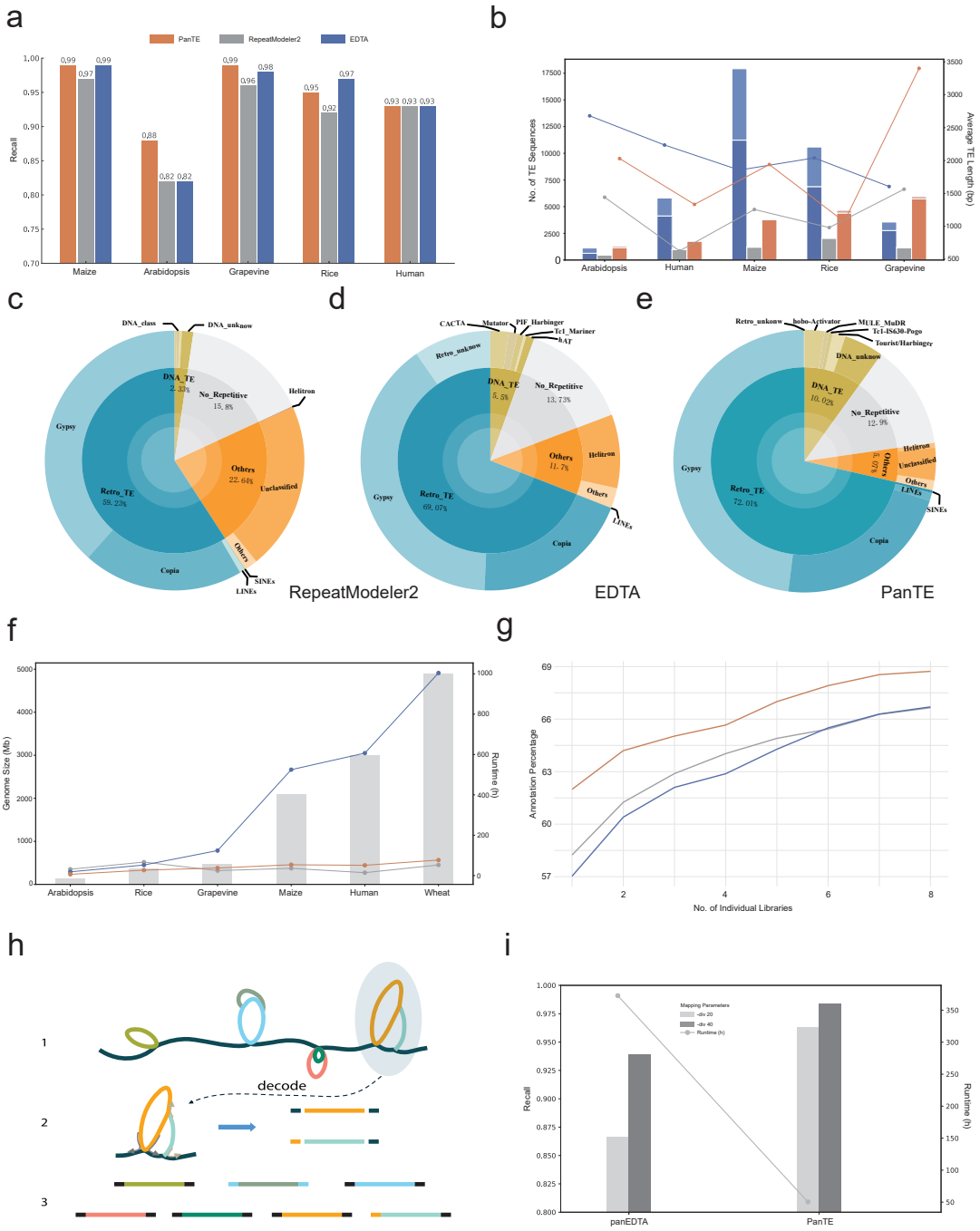
a
PanTE
RepeatModeler2
EDTA
Recall
Maize
Arabidopsis
Grapevine
Rice
Human
b
No. of TE Sequences
Average TE Length (bp)
Arabidopsis
Human
Maize
Rice
Grapevine
c
RepeatModeler2
d
EDTA
e
PanTE
f
Genome Size (Mb)
Runtime (h)
Arabidopsis
Rice
Grapevine
Maize
Human
Wheat
g
Annotation Percentage
No. of Individual Libraries
h
decode
i
Recall
Runtime (h)
panEDTA
PanTE

Fig. 2|Benchmarking PanTE with EDTA and RepeatModeler2 on real genomes. a, Recall for genome annotation using TEs identified through EDTA, RepeatModeler2 and PanTE. Each color represents a different method. **b**, TE libraries for various species generated by different methods. The bars represent the number of TE sequences identified, with the light-colored portion indicating the number of low-frequency TE sequences (<10 repeats in the genome). The points indicate the average TE length in each TE library. **c-e**, Annotation percentage of various TE families in the maize genome (GCA_022117705.1) using TEs identified through different methods. **f**, Runtime of each method across different genomes. The bars represent the genome size of each species, while the points, displayed in different colors, indicate the runtime for TE library construction by different methods. **g**, Annotation percentage of the grapevine genome (GCF_030704535.1) by integrating multiple individual TE libraries. **h**, Graph-based pangenomes disassembly, which involves: 1) core segment selection, 2) core segment extension, and 3) generation of multiple sequences. **i**, Recall for rice genome (GCF_034140825.1) annotation using pan-libraries of TEs constructed by panEDTA and PanTE. Each color represents a different mapping parameter: -div 20 or -div 40. The points indicate the runtime for TE library construction. All runtime values were calculated on an Intel 6230R processor with 384 GB RAM and 12 threads.

Given the limitation of our simulations in assessing the efficiency of EDTA, we compared PanTE with EDTA and RepeatModeler2 using real genomes with curated TE reference libraries from various species, including *Vitis vinifera* (grapevine), *Zea mays* (maize), *Oryza sativa* (rice), *Arabidopsis thaliana* (Arabidopsis), and *Homo sapiens* (human). Additionally, to compare the runtime for TE library construction, an additional species, *Triticum aestivum* (wheat) was included.

For different species, we first generated individual TE libraries and compared them to their corresponding curated reference libraries (Supplementary Table 4). The performance of each method was evaluated based on recall rather than precision, as the curated libraries may not fully encompass all true TEs. We found that the individual libraries generated by PanTE exhibited lower recall in rice, Arabidopsis and maize compared to EDTA, but outperformed EDTA in human and grapevine, while consistently surpassing RepeatModeler2 in all species except for human (Supplementary Table 4). To further assess the quality of these generated libraries, we mapped these identified TEs to their corresponding genomes and compared the actual proportion of the genomes correctly annotated by the different libraries using recall as the evaluation metric (Fig. 2a, Supplementary Table 5). PanTE either outperformed or was competitive with the other methods across all species, with the exception of rice, where EDTA achieved the best performance. Therefore, EDTA demonstrated superior performance in library comparison compared to genome annotation. This discrepancy may arise from EDTA's focus on detecting TEs through contextual structures rather than their frequencies, potentially leading to the omission of certain high-frequency TEs. Additionally, we compared genome annotation by each TE family across methods (Supplementary Table 6). For LTR-retrotransposons, PanTE effectively targeted the LTR structures and leveraged the repetitive nature of TEs, achieving a higher or competitive recall compared to the other methods, particularly for families such as Gypsy and Copia (Supplementary Table 6). PanTE also excelled in identifying relatively high-abundance TE families without LTR structures, such as LINE-1 in most species. However, in Arabidopsis, where LINE-1 elements are highly truncated and less active due to adaptation to a compact genome [27],

PanTE did not achieve a strong performance. Notably, for MuDR, a TE family characterized by known TSDs [28], PanTE obtained superior recall compared to EDTA in the grapevine genome, relying on its high frequency rather than contextual structures (Supplementary Table 6).

In other aspects, the individual TE libraries generated by EDTA contained the highest proportion of low-frequency TEs (<10 repeats in the genome), whereas RepeatModeler2 detected very few (Fig. 2b, Supplementary Table 7). EDTA tended to identify more TE families with longer sequences compared to the other two methods, likely sacrificing precision in favor of higher recall in both library comparison and genome annotation. In contrast, RepeatModeler2 detected fewer and shorter TE sequences (Fig. 2b), and annotated a larger proportion of the exemplar maize genome (GCA_022117705.1) as "No_Repetitive" (not TEs, Fig. 2c), likely prioritizing precision over recall. This inference is further supported by our simulation results, where the performance of RepeatModeler2 is much better in precision rather than recall (Fig. 1b,c). Our PanTE achieved a balance between these extremes (Fig. 2e). Despite EDTA generating a larger library, indicating it detected a wider variety of TEs than the other two methods, the proportion of the exemplar genome it annotated was not higher than that annotated by PanTE (Fig. 2d,e). In the grapevine genome, PanTE annotated ~ 26% more TEs compared to EDTA (Supplementary Fig. 1). This indicates that the representativeness of TE consensus sequences identified by EDTA may be lower than that of PanTE. As shown in Fig. 2d (also see Supplementary Table 6), EDTA excelled at detecting TEs with specific contextual structures but without high abundances, such as Helitron elements, which rely on these structures to support their

rolling-circle mechanism for replication [29]. However, the efficiency of EDTA may be compromised if the structures are disrupted or nested, as can occur through evolutionary processes.

In addition, we calculated the overlap of identified TEs between each pair of methods (Supplementary Table 8). PanTE displayed greater overlap with either EDTA and RepeatModeler2 than the overlap between EDTA and RepeatModeler2 themselves, indicating broader coverage across both methods. Regarding runtime (Fig. 2f, Supplementary Table 9), we found that the performance of RepeatModeler2 might be largely influenced by genetic complexity with respect to gene density and TE activity, rather than genome size. For instance, RepeatModeler2 processed large genomes such as the human genome, relatively quickly, but was slower with smaller yet more complex genomes, such as rice and maize. In contrast, the runtimes of PanTE and EDTA were strongly influenced by genome size. Notably, PanTE demonstrated dramatically faster processing speeds than EDTA across species with varying genome sizes, particularly for large genomes such as wheat, where its speed was up to 13 times faster than EDTA (Supplementary Table 9). The efficient performance of PanTE is attributed to its practical sampling strategy using masking and iteration, allowing it to process large genomes within a more moderate timeframe.

For grapevine genomes [5,30–32], we also constructed individual TE libraries for each of the eight varieties using different methods and progressively aggregated them into pan-libraries. These pan-libraries were then used for genome annotation. We observed that the

proportion of the genome (GCF_030704535.1 [5]) annotated as TEs increased as more individual TE libraries were incorporated (Fig. 2g, Supplementary Table 10). This finding highlights the importance of establishing a pan-library of TEs for a population. Furthermore, PanTE consistently outperformed the other two methods when the same number of individual libraries were considered, proving the superior efficacy of PanTE in constructing comprehensive TE libraries.

Unlike EDTA and RepeatModeler2, our framework PanTE can be directly applied to graph-based pangenomes (Fig. 2h). Among the methods compared, only EDTA includes an extended tool (panEDTA) for combining individual TE libraries from single genomes into a pan-library. Therefore, we compared PanTE with panEDTA in terms of runtime for TE library construction and recall for genome annotation using the TE library generated from a rice graph-based pangenome. The graph genome was constructed using seven individual rice genomes through the PGGB framework. PanTE directly utilized the pangenome as input, while panEDTA used the seven individual genomes (Supplementary Table 11). In terms of runtime, PanTE was over seven times faster than panEDTA, largely widening the performance gap observed in single genome analysis, where PanTE was less than twice as fast (Fig. 2i, Supplementary Table 9). This efficiency difference is expected to increase as more individual genomes are incorporated. Regarding recall in rice genome annotation (GCF_034140825.1), PanTE demonstrated substantial improvement. For the TE library built based on the single genome, PanTE achieved a recall of 95% using the lenient mapping setting on the genome (RepeatMasker: -div 40), slightly lower than EDTA which was 97% (Fig. 2a). However, with the pan-library, PanTE performed better than

when using its single genome TE library, even with the more stringent setting (-div 20), and outperformed panEDTA (Fig. 2i, Supplementary Table 12). Notably, the pan-library generated by panEDTA contained fewer sequences than the TE library from a single genome, decreasing from 10,596 to 3,648 due to its stringent filtering threshold. In contrast, PanTE detected a higher number of TEs, increasing from 4655 to 13375. This suggests that panEDTA may fail to capture some low-frequency TEs, particularly those present in only a few individuals. These findings also emphasize the necessity of direct TE detection on graph-based pangenomes.

Conclusions

In summary, these results demonstrate that PanTE, which identifies TEs based on their general structural characteristics and repetitive nature, is highly effective at detecting a wide variety of TEs without being constrained by specific TE structures. PanTE proficiently handles large genomes and graph pangenomes with superior performance, offering significant computational efficiency. Its ability to reliably detect high-abundance TEs and LTR-retrotransposons surpasses that of currently available methods. For TEs with specific contextual structures but without high abundances, EDTA retains an advantage, but at the cost of computational resources. However, this insufficiency of PanTE can be mitigated by using pangenomes, which have the potential to increase the TE frequency. Importantly, the application of PanTE to graph pangenomes enables comprehensive TE detection across entire populations, ensuring broader coverage of genetic diversity and the identification of population-specific TEs, making it particularly valuable for studying genetics, genomics and evolution of TEs.

Methods

Framework of PanTE

PanTE is a robust computational framework designed for the discovery of TEs in both single genomes and graph-based pangenomes. The framework is divided into three main steps: pre-processing, TE detection and TE classification.

Pre-processing

PanTE accepts input files in either FASTA or GFA format, where FASTA files correspond to single genomes and GFA files represent pangenomes. In the case of FASTA format, all non-standard nucleotide bases, i.e., those other than A, T, C, G are replaced with "N" to ensure uniformity across the dataset (data cleaning). Next, sequence identifiers (IDs) are standardized to maintain consistency (sequence renaming). Finally, sequences are indexed by matching sequence IDs with their corresponding chromosome lengths (sequence indexing), a crucial step that facilitates downstream analyses. For pangenomes, PanTE processes GFA format files generated by software such as PGGB, Minigraph-Cactus and Minigraph. In this case, the pre-processing involves the random selection of core genome fragments ranging in length from 50 bp to 40,000 bp. These core segments are subsequently extended symmetrically on both sides until the terminal ends of the sequences are reached. Only the extended segments with a total length exceeding 40,000 bp, and with less than 40% overlap in length compared to previous retained sequences, are preserved and indexed for further analysis.

TE detection

The second step of the PanTE framework involves the adaption of three widely used algorithms for the detection of TEs – Look4LTRs， RepeatScout and RECON. Each algorithm targets different aspects of TE structure and sequence similarity, enhancing the overall detection accuracy.

Look4LTRs is specifically designed to detect LTR structures, a characteristic feature of LTR-retrotransposons. This algorithm is graph-based and leverages machine learning techniques to efficiently identify LTR structures genome-wide. By focusing on structural detection rather than sequence alignment, Look4LTRs significantly improves computational speed compared to previous LTR detection methods. It is particularly effective in detecting recently generated or structurally intact TEs, including nested elements. Look4LTRs is capable of processing whole genomes at large scales and reports LTR positions with a confidence level greater than 0.95.

RepeatScout identifies repetitive sequences by detecting high-frequency k-mers, which are considered as dispersed repetitive seeds representing short regions of homology. The algorithm performs multiple alignments and iterative extension around the aligned seeds to build complete repetitive sequences. For the detection process, RepeatScout requires an input of genome fragments approx. 400 Mbp in length, which are randomly sampled from genomes typically larger than 400 Mbp. The output consists of representative sequences of potential TEs.

After the initial detection by Look4LTRs and RepeatScout, the candidate sequences are subjected to filtering using Tandem Repeat Finder (TRF) [33] to remove tandem repeats. To further refine the results and reduce false positives, the candidate sequences are aligned back to the reference genome using RMblast, allowing for the identification of all possible TE sequences. These sequences are then clustered using Ninja [34]. For each cluster, a consensus sequence is constructed using multiple sequence alignment through MAFFT [35].

RECON is a more sophisticated algorithm that employs cross-alignment of input sequences to generate TE representative sequences based on sequence similarity. This algorithm is particularly sensitive to the detection of older TE sequences, but it is computationally intensive. To address the computational burden, especially for large genomes, the TE detection process with RECON unit is divided into iterations. For genomes smaller than 500 Mbp, detection is performed in a single iteration with multiple samplings to achieve full coverage, typically 40 Kbp at length for each sampling. For genomes larger than 500 Mbp, detection is performed in multiple iterations, with progressively increasing genome coverage for each iteration. For example, in genomes exceeding 10 Gbp in size, five iterations are performed with expected genome coverages of 240 Mbp, 480 Mbp, 720 Mbp, 1400 Mbp and 5600 Mbp, respectively. In each iteration, the genome is sampled randomly with fragments of 200 Kbp in length until the required coverage is reached. To further improve computational efficiency, genome samples (e.g., 240 Mbp for a 10 Gbp genome) are pre-masked using RepeatMasker [36] based on the candidate TE consensus sequences

generated from Look4LTRs and RepeatScout units. After each iteration, the detected TE sequences are refined, as described for the previous units, to generate consensus sequences. These new consensus sequences are then masked prior to subsequent iterations.

Finally, the candidate TE consensus sequences obtained from all detection units are merged using the 80-80 rule [37], which considers two sequences as members of the same family if they share at least 80% sequence identity over 80% of their length. Sequences meeting this criteria are considered redundant and are merged using cd-hit-est [38].

TE classification

In the final step, the detected TE consensus sequences, constituting a comprehensive TE library, are classified into families using two complementary algorithms: RepeatClassifier and ClassifyTE. RepeatClassifier, a homology-based algorithm, compares the TE consensus sequences against established databases such as the Repeat Protein Database, Dfam and RepBase. To minimize false positives, RepeatClassier includes a quality control step that filters out sequences failing to meet a specified alignment score (e.g., match score > 250), labeling these sequences as “Unknown”. For sequences that do not match any entries in the reference databases, ClassifyTE, a deep-learning algorithm, is employed to classify unknown TEs into existing classes using a stacking-based machine learning algorithm [22]. This unit ensures that all detected TEs are classified into their respective families to the greatest extent possible, thereby minimizing the occurrence of "Unknown" labels.

Simulation strategies

Simulations for TE detection

To simulate single genomes for TE detection, we generated a total of 150 genomes equally distributed across five populations, each representing different levels of TE diversity. Within each population, individuals were simulated across six levels of gradient-increasing TE content, with five replicates for each level. Specifically, we began by constructing five ancestral TE repositories for the five populations, achieved by randomly sampling 1600, 1800, 2000, and 2200 TE sequences from the RepBase database. Subsequently, five distinct TE libraries were established by randomly sampling 1200 sequences from each ancestral repository without replacement, thereby simulating varying degrees of within-population diversity of TEs.

To simulate TE-free genomes and insert TE sequences into these genomes, we developed a workflow termed simuTE, which consists of two components: simu-fa and simu-ins. The simu-fa component is designed to simulate TE-free genomes and is executed using the command -L 150 Mbp (to specify genome length) and -S 3 (to specify the nucleotide expectation ratio). This ratio indicates the relative probability of different nucleotides versus the same nucleotide occurring at successive positions in the sequence, with higher ratios reducing the likelihood of consecutive identical nucleotides. For example, in this study, the nucleotide expectation ratio was 3, meaning the probability of each of the four nucleotides (A, T, C, G) occurring at the initial position was set to 0.25. For subsequent positions, the probability of the same nucleotide as the previous position was adjusted to 0.1, while the probability for alternative nucleotides was set to 0.3.

The second component, simu-ins, is designed to randomly insert predefined TE sequences into a given genome. In this study, to simulate genomes with varying TE contents, the previously generated TE libraries were duplicated 20, 30, 40, 50, 60 and 70 times to achieve TE content in the genome ranging from 30% to 80%, corresponding to the TE content levels observed in natural genomes. This process was executed using the command -i lib.fa (to specify the TE library), -g genome.fa (to specify the input genome), and -x 20 (to specify the number of duplications for the TE library).

Since the TE sequences were inserted randomly into the simulated genomes, the resulting TE locations were unpredictable, leading to considerable genomic heterogeneity. As a result, constructing a coherent pangenome from these simulations proved impractical, limiting the scope of simulations for pangenome analysis.

Simulations for TE classification

Regarding the simulations for TE classification, we randomly selected 100 sequences from each of 15 prevalent TE families, including hAT, Gypsy, Copia, CACTA, Bel, TC1, SINE, RTE, R2, R1, Piggybac, MuDR, LINE-2, LINE-1, and Helitron, from the RepBase database. These sequences were used to test and validate the ability of PanTE to classify TEs.

Real Genomes

The real reference genomes analyzed in this study include the most recent releases of complete genome assemblies for *Arabidopsis thaliana* (Arabidopsis, GCF_000001735.4), *Homo sapiens* (human, GCF_009914755.1), *Oryza sativa* (rice, GCF_034140825.1), *Zea mays* (maize, GCA_022117705.1), *Vitis vinifera* (grapevine, GCF_030704535.1) and *Triticum aestivum* (wheat, GCA_910594105.1). Among these species, the manually curated TE reference library for the rice single genome was sourced from the work of Ou *et al* [8]. For Arabidopsis, human, maize, and grapevine, we constructed the corresponding curated TE libraries using the available TE representative sequences from Repbase, which is widely recognized as a reference standard for TE annotation [25]. The TE annotated genomes were then achieved using RepeatMasker by applying the command: RepeatMasker -lib lib.fa (to specify the input TE library) -e rmblast -lcambig -dir ones -pa 3 -s -gff -dvi 40 genome.fa (to specify the genome to be annotated). The parameters used in the command line were optimized based on recommendations provided in the RepeatMasker documentation (https://www.repeatmasker.org/webrepeatmaskerhelp.html).

In addition to the reference genomes, we included seven additional genome assemblies of grapevine variates, and six additional genome assemblies of rice variates, which were downloaded from the respective databases (see Supplementary Table 11). The seven rice genomes, including the previously used reference genome, were further assembled into a graph-based pangenome using PGGB under default parameters, providing a comprehensive dataset for pangenome analysis.

Benchmarking and Assessment

Simulations for TE detection

PanTE was benchmarked against RepeatModeler2 for TE detection. To evaluate the performance of each method, we applied blastn to compare the sequences from the simulated reference TE library with those generated by each detection method. Due to the inherent complexity of TE insertions and the differing detection preferences of each method, the identified TE sequences were not expected to match exactly across methods. For instance, multiple TEs may form nested structures resulting in chimeras, particularly in simulations with high TE contents, where TE insertions occur frequently. Such TE chimeras may be detected either entirely as a single entity or partially as one or more entities. To account for these complexities, we implemented specific alignment criteria for the calculation of accuracy metrics, including precision and recall, to better reflect the detection efficiency of each method.

For precision, detected sequences were considered true positives if at least 80% of their length aligned to reference sequences (not necessarily a single reference) with a minimum of 80% sequence similarity. The precision metric was calculated as:

Precision = number of detected sequences meeting the 80-80 criteria / total number of detected sequences.

For recall, reference sequences were considered accurately detected if at least 80% of their length aligned to the detected sequences with a minimum of 80% similarity. The recall metric was calculated as:

Recall = number of reference sequences meeting the 80-80 criteria / total number of reference sequences.

The overall performance was summarized using the F1 score, calculated as:

F1 score = 2 * precision * recall/ (precision + recall).

The pan-library of TEs for each population was generated by aggregating TE libraries from multiple single genomes using cd-hit-est following the 80-80 rule.

Simulations for TE classification

For the evaluation of TE classification, we benchmarked PanTE against four existing methods using their default parameters: ClassifyTE, Deep-TE, RepeatClassifier (as part of RepeatModeler2), and TERL. Accuracy metrics were calculated both collectively across all TE families and individually for each family.

For the evaluation across all families, precision was defined as the proportion of TEs correctly classified into their respective families, calculated as:

Precision = number of TEs correctly classified into the right family/total number of TEs that were not classified as “unknown”.

Recall represented the proportion of TEs from the reference library that were correctly identified, calculated as:

Recall = number of TEs correctly classified into the right family/ total number of sequences in the reference TE library

When evaluating each TE family separately, precision was defined as the proportion of TEs classified into a given family that were accurately classified, calculated as:

Precision (each family) = number of TEs correctly classified into a specific family /total number of TEs classified into that family. Recall, in this context, referred to the proportion of TEs in a given family from the reference library that were accurately classified, calculated as:

Recall (each family) = number of TEs correctly classified into a specific family/total number of sequences in that family within the reference TE library

The F1 score was then calculated as:

F1 score = 2 * (precision * recall)/ (precision + recall).

This approach enabled a comprehensive assessment of the classification accuracy of each method, providing insights into their effectiveness across various TE families.

Real genomes

For the real reference genomes of various species including Arabidopsis, human, rice, maize, and grapevine, we constructed TE libraries for each genome using three methods: EDTA, RepeatModeler2 and PanTE. Wheat was included solely to compare the runtime for TE library construction, given its large genome size but absence of a curated TE library.

For each generated TE library, we calculated both the total number of TE consensus sequences and their average length. These TE libraries were then compared with their corresponding curated libraries using recall as defined in the section "*Simulations for TE detection*". In addition, the TE libraries generated by the different methods, along with the curated libraries from Repbase, were used to annotate the corresponding genomes. For each genome, the proportion annotated as different classes of TEs was calculated and visualized. TE sequences that matched fewer than 10 locations in the genome were categorized as low-frequency sequences. The performance of each method was further evaluated by calculating recall, both collectively across all TE families and individually for each family, to quantify the proportion of the genome correctly annotated with TEs. Specifically, recall was defined as:

Recall = overlap between the regions annotated with the curated TE library and those annotated with the generated library/ total region annotated with the curated TE library.

For the eight grapevine genomes, individual TE libraries were constructed using EDTA, RepeatModeler2, and PanTE. These libraries generated from each genome were subsequently merged into a pan-library using panEDTA.sh for EDTA or cd-hit-est for RepeatModeler2 and PanTE according to the 80-80 rule. The high-quality reference genome (GCF_030704535.1) was then annotated based on the pan-libraries of TEs using RepeatMasker, and the proportion of the genome annotated as TEs was calculated.

For the seven rice genomes and their corresponding pangenome, panEDTA constructed seven individual TE libraries, which were subsequently merged into a pan-library using the parameter -sensitive 1, while PanTE directly constructed a TE-library for the pangenome. RepeatMasker, configured with two mapping settings (lenient: -div 40; stringent: -div 20) was then used to annotate the rice reference genome (GCF_034140825.1) using both the generated TE libraries and a manually curated TE library. The recall for the genome annotation was subsequently calculated.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The real genome datasets supporting the conclusions of this article are available in the NCBI database (*Arabidopsis thaliana*: GCF_000001735.4, *Homo sapiens*: GCF_009914755.1, *Oryza sativa*: GCF_034140825.1, *Zea mays*: GCA_022117705.1, *Vitis vinifera*: GCF_030704535.1 and *Triticum aestivum*: GCA_910594105.1). The simulated data used in this paper are available from the corresponding author upon reasonable request. The genome assemblies of rice variates were obtained from the sources listed in **Supplementary Table 11**. The computational tool developed in this paper is publicly available at: https://github.com/unavailable-2374/Pan_TE.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Key Research and Development Program of China [2023YFD2200700] to Yiwen Wang and the Science Fund Program for Distinguished Young Scholars of the National Natural Science Foundation of China (Overseas) to Yongfeng Zhou.

Authors' contributions

Y.W. and Y.Z. designed the project and supervised S.C. and Z.L. S.C. developed and implemented the framework. S.C. and Z.L. benchmarked the methods. S.C., Y.W. and Z.L. wrote and edited the manuscript. Y.L., Z.L., W.M., J.L., H.X., J.C., S.O. and E.G. contributed to manuscript preparation and provided comments on the manuscript. All authors read and approved the paper.

Acknowledgements

Not applicable.

References

1. Wells JN, Feschotte C. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet. 2020;54:539–61.

2. Wessler SR. Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci. 2006;103:17600–1.

607 3. Raviram R, Rocha PP, Luo VM, Swanzey E, Miraldi ER, Chuong EB, et al. Analysis of
608 3D genomic interactions identifies candidate host genes that transposable elements
609 potentially regulate. Genome Biol. 2018;19:216.

610 4. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete
611 sequence of a human genome. Science. 2022;376:44–53.

612 5. Shi X, Cao S, Wang X, Huang S, Wang Y, Liu Z, et al. The complete reference genome
613 for grapevine (Vitis vinifera L.) genetics and breeding. Hortic Res. 2023;10:uhad061.

614 6. Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, et al. A complete telomere-to-telomere
615 assembly of the maize genome. Nat Genet. 2023;55:1221–31.

616 7. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2
617 for automated genomic discovery of transposable element families. Proc Natl Acad Sci.
618 2020;117:9451–7.

619 8. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking
620 transposable element annotation methods for creation of a streamlined, comprehensive
621 pipeline. Genome Biol. 2019;20:275.

622 9. Kidwell MG, Lisch D. Transposable elements as sources of variation in animals
623 and plants. Proc Natl Acad Sci. 1997;94:7704–11.

624 10. Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, et al. The population
625 genetics of structural variants in grapevine domestication. Nat Plants. 2019;5:965–79.

626 11. Kou Y, Liao Y, Toivainen T, Lv Y, Tian X, Emerson JJ, et al. Evolutionary Genomics

627 of Structural Variation in Asian Rice (Oryza sativa) Domestication. Mol Biol Evol.
628 2020;37:3507–24.

629 12. Yoo D, Rhie A, Hebbar P, Antonacci F, Logsdon GA, Solar SJ, et al. Complete
630 sequencing of ape genomes. bioRxiv. 2024;2024.07.31.605654.

631 13. Liu Z, Wang N, Su Y, Long Q, Peng Y, Shangguan L, et al. Grapevine pangenome
632 facilitates trait genetics and genomic breeding. Nat Genet. 2024;56:2804–14.

633 14. Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing
634 heritability and empowers tomato breeding. Nature. 2022;606:527–34.

635 15. Ou S, Scheben A, Collins T, Qiu Y, Seetharam AS, Menard CC, et al. Differences in
636 activity and stability drive transposable element variation in tropical and temperate maize.
637 Genome Res. 2024;34:1140–53.

638 16. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph
639 construction from genome alignments with Minigraph-Cactus. Nat Biotechnol.
640 2024;42:663–73.

641 17. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building
642 pangenome graphs. Nat Methods. 2024;21:2008–12.

643 18. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large
644 genomes. Bioinforma Oxf Engl. 2005;21 Suppl 1:i351-358.

645 19. Bao Z, Eddy SR. Automated De Novo Identification of Repeat Sequence Families in
646 Sequenced Genomes. Genome Res. 2002;12:1269–76.

20. Garza AB, Lerat E, Girgis HZ. Look4LTRs: a Long terminal repeat retrotransposon detection tool capable of cross species studies and discovering recently nested repeats. Mob DNA. 2024;15:8.

21. Heringer P, Benson CW, Ou S. Navigating the Maze of Maize Genomics: the Impact of Transposable Elements and Tandem Repeats. Cold Spring Harb Protoc. 2024;

22. Panta M, Mishra A, Hoque MT, Atallah J. ClassifyTE: a stacking-based prediction of hierarchical classification of transposable elements. Bioinformatics. 2021;37:2529–36.

23. Yan H, Bombarely A, Li S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics. 2020;36:4269–75.

24. da Cruz MHP, Domingues DS, Saito PTM, Paschoal AR, Bugatti PH. TERL: classification of transposable elements by convolutional neural networks. Brief Bioinform. 2021;22:bbaa185.

25. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.

26. Tan S, Ma H, Wang J, Wang M, Wang M, Yin H, et al. DNA transposons mediate duplications via transposition-independent and -dependent mechanisms in metazoans. Nat Commun. 2021;12:4280.

27. Quesneville H. Twenty years of transposable element analysis in the Arabidopsis thaliana genome. Mob DNA. 2020;11:28.

28. Dupeyron M, Singh KS, Bass C, Hayward A. Evolution of Mutator transposable

667 elements across eukaryotic diversity. Mob DNA. 2019;10:12.

668 29. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci.
669 2001;98:8714–9.

670 30. Wang X, Liu Z, Zhang F, Xiao H, Cao S, Xue H, et al. Integrative genomics reveals the
671 polygenic basis of seedlessness in grapevine. Curr Biol. 2024;34:3763-3777.e5.

672 31. Zhang T, Peng W, Xiao H, Cao S, Chen Z, Su X, et al. Population genomics highlights
673 structural variations in local adaptation to saline coastal environments in woolly grape. J
674 Integr Plant Biol. 2024;66:1408–26.

675 32. Zhang K, Du M, Zhang H, Zhang X, Cao S, Wang X, et al. The haplotype-resolved
676 T2T genome of teinturier cultivar Yan73 reveals the genetic basis of anthocyanin
677 biosynthesis in grapes. Hortic Res. 2023;10:uhad205.

678 33. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids
679 Res. 1999;27:573–80.

680 34. Wheeler TJ. Large-Scale Neighbor-Joining with NINJA. In: Salzberg SL, Warnow T,
681 editors. Algorithms Bioinforma. Berlin, Heidelberg: Springer; 2009. p. 375–89.

682 35. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of
683 multiple sequence alignment. Nucleic Acids Res. 2005;33:511–8.

684 36. Tempel S. Using and Understanding RepeatMasker. In: Bigot Y, editor. Mob Genet
685 Elem Protoc Genomic Appl [Internet]. Totowa, NJ: Humana Press; 2012 [cited 2025 Jan
686 3]. p. 29–51. Available from: https://doi.org/10.1007/978-1-61779-603-6_2

687 37. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified
688 classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.

689 38. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
690 sequencing data. Bioinforma Oxf Engl. 2012;28:3150–2.

691

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytables.xlsx
- SupplementaryMaterial.docx