BRIEF REPORT

An Atlas of Plant Transposable Elements [version 1; peer review: 1 approved]

Daniel Longhi Fernandes Pedro¹, Tharcisio Soares Amorim¹, Alessandro Varani², Romain Guyot³, Douglas Silva Domingues⁴¹,⁵, Alexandre Rossi Paschoal¹

¹Department of Computer Science; Bioinformatics and Pattern Recognition Group, Graduation Program in Bioinformatics, Federal University of Technology - Paraná (UTFPR), Cornélio Procópio, Paraná, 86300000, Brazil
²Departament of Agricultural and Environmental Biotechnology, School of Agricultural and Veterinary Sciences, São Paulo State University (UNESP), Jaboticabal, São Paulo, 14884-900, Brazil
³Institut de Recherche pour le Développement, IRD, University of Montpellier, Montpellier, France
⁴Department of Electronics and Automatization, Universidad Autónoma de Manizales, Manizales, Colombia
⁵Group of Genomics and Transcriptomes in Plants, Institute of Biosciences, São Paulo State University (UNESP), Rio Claro, São Paulo, 13506-900, Brazil

Abstract

Advances in genomic sequencing have recently offered vast opportunities for biological exploration, unraveling the evolution and improving our understanding of Earth biodiversity. Due to distinct plant species characteristics in terms of genome size, ploidy and heterozygosity, transposable elements (TEs) are common characteristics of many genomes. TEs are ubiquitous and dispersed repetitive DNA sequences that frequently impact the evolution and composition of the genome, mainly due to their redundancy and rearrangements. For this study, we provided an atlas of TE data by employing an easy-to-use portal (APTE website). To our knowledge, this is the most extensive and standardized analysis of TEs in plant genomes. We evaluated 67 plant genomes assembled at chromosome scale, recovering a total of 49,802,023 TE records, representing a total of 47,992,091,043 (~47.62%) base pairs (bp) of the total genomic space. We observed that new types of TEs were identified and annotated compared to other data repositories. By establishing a standardized catalog of TE annotation on 67 genomes, new hypotheses, exploration of TE data and their influences on the genomes may allow a better understanding of their function and processes. All original code and an example of how we developed the TE annotation strategy is available on GitHub (Extended data).

Keywords

mobile elements, atlas, large-scale, genome-wide, standardized, plants

Open Peer Review

Reviewer Status

Invited Reviewers

1

version 1
24 Nov 2021

1. Gökhan Karakülah, Dokuz Eylül University, Izmir, Turkey

Any reports and responses or comments on the article can be found at the end of the article.
Introduction

The growing number of sequenced plant genomes is providing unprecedented opportunities for biological studies, evolution, and growing of many algal and Viridiplantae species. We estimate more than 13k plant genomes have been released (NCBI), revealing that plant genomes are faintly explored. High diversity in terms of ploidy, heterozygosity, and genome size, probably due to a dynamic set of old and recent bursts of transposable elements (TEs), are common hallmarks of many plant genomes.\textsuperscript{1} TEs can comprise between 32\% to 56\% (\textit{Utricularia} genomes),\textsuperscript{2,3} to up to 90\% in many plant genomes,\textsuperscript{4–6} e.g., maize\textsuperscript{7} and wheat.\textsuperscript{5,8,9}

TEs can be organized into two main classes. Each class is hierarchically organized into orders, superfamilies, families, and subfamilies. This terminology is primarily associated with the type of their transposition mechanisms.\textsuperscript{3,10} They are classified into: (i) retrotransposons (Class I), which are propagated by a “copy-and-paste” mobility mechanism and are the most redundant TE class in plant genomes; and (ii) DNA transposons (Class II), which are known for the “cut-and-paste” mechanism that allows them to move to a completely different position. Moreover, both Classes may contain autonomous members\textsuperscript{10,11} for which the transposition mechanism depends on an autonomous and cognate type of TE.\textsuperscript{11,12}

However, despite their significance, in-depth identification and analysis of TEs content in the sequenced plant genomes remains barely explored.\textsuperscript{13,14} The lack of concise data available may prevent the enrichment of \textit{in silico}, functional genomics research and compromises the appearance of new strategies to investigate TEs. Recently, many computational models and entire wet-lab efforts have increasingly been helping to understand these sequences.\textsuperscript{15–18} For example, \textit{Ensembl Plants}\textsuperscript{19} provides high-quality, primary genomic information for 67 plant (in the broad sense, including green plants, green and red algae) genomes, assembled near or at chromosome scale; however, mobile sequences are poorly systematized and have a humble coverage.

These observations prompted us to standardize tools and methods aiming to improve TE detection, annotation and standardization. In this work, we developed a new method for systematic annotation of plant TEs, using the 67 genomes available at Ensembl Plants assembled at chromosome scale as a starting point. Our identification was standardized, applying the same methodolo\textsuperscript{ies to all genomes and delivering a concise Atlas of TEs annotation in plant genomes. We also provided an updated analysis of non-coding RNAs (ncRNAs) overlapping TEs. This annotation is accessible on the Atlas website for exploration and download, which might be relevant to any type of research involving mobile sequences.

Methods

Data source

All genomes (Supplementary Material 1, \textit{Extended data}) were downloaded from the Ensembl Plants\textsuperscript{19} database, version 41 (57 genomes) and 45 (plus 10 new genomes).

Annotation of transposable elements

We used similarity-based methods and \textit{de novo} techniques to build a collection of putative transposable elements, based on the SPTEdb pipeline.\textsuperscript{21} We refined, extended and increased steps in order to produce a novel annotation (Figure 1). Our reformulated steps (details in Supplementary Material 2, \textit{Extended data}) guarantee a comprehensive knowledgebase of these TEs.

RepeatScout was performed separately; the output was unified in a library to be labeled by \texttt{PASTECclassifier}\textsuperscript{32} and later combined into a final annotation. To automate the pipeline, an in-house framework in Perl language was developed for each software output to be uniformized, described in steps 1 to 4. A main script in Bash starts the process of automatization using Perl scripts. All steps were supervised by researchers, carefully checked, and the output was manually verified at each step for each genome. Records classified as low complexity, simple repeat and other nomenclature not related to Class I or Class II TEs were discarded.

Due to the extensive genome sizes of \textit{Triticum aestivum} (14.5 Gb), \textit{Triticum dicoccum} (10.4 Gb) and \textit{Triticum turgidum} (10.4 Gb), we adapted our pipeline for their analysis, based on the approach of Jamilloux \textit{et al.}\textsuperscript{24} For these species, we applied our pipeline on chromosome 1 (which is the longest pseudomolecule), as the large genomes were eventually duplicated into new copies, increasing the number of these same repeats in the genome, and did not significantly impact discoveries related to new or different TEs families.\textsuperscript{21}
To test the reliability of our TE annotation pipeline, we scored sequences that had duplicated annotation in the same loci (Figure 2). We developed a statistical metric (labeled as TE-Score, shown in each record as the ninth column for each genome annotation file) that identify and ponder sequences types that have been identified by the programs. The TE-Score is a metric \(0 \leq \text{TE-Score} \leq 1\) that is given by:

\[
\text{TE-Score} = \frac{Q_{IP}}{Q_{P}}
\]

Where:
- \(Q_{IP}\): The total number of identical calls made by all programs.
- \(Q_{P}\): The total number of unique calls made by all programs.

**Figure 1. Steps in transposable elements identification.** Dataset: Genome assemblies were downloaded from Ensembl Plants. Identification: 1A) RepeatScout was used to search for putative repetitive sequences and further classification by PASTEClassifier, resulting in a library. 1B) RepeatModeler was also used to find a consensus of TEs sequences. 2) RepeatMasker was run with Repbase library and libraries from RepeatModeler and RepeatScout. 3) For Class II - Subclass 2 TEs, we also used HelitronScanner and MITE-Hunter. 4) In order to find LTR and Non-LTR retrotransposons, we used LTR_retriever and MGEScan-non-LTR, respectively. Filter: A cut-off filter was applied to remove low complexities, simple repeats and other nomenclatures that were not classified into TEs. Annotation: In result of the pipeline, we have a Transposable Element annotation for each genome analyzed.

**Figure 2. The TE Score: the average amount of sequence identification made by programs in all genomes.**

**TE evidence score**

To test the reliability of our TE annotation pipeline, we scored sequences that had duplicated annotation in the same loci (Figure 2). We developed a statistical metric (labeled as TE-Score, shown in each record as the ninth column for each genome annotation file) that identify and ponder sequences types that have been identified by the programs. The TE-Score is a metric \(0 \leq \text{TE-Score} \leq 1\) that is given by:

\[
\text{TE-Score} = \frac{Q_{IP}}{Q_{P}}
\]
where QIP = Quantity of identification by program and QP = Quantity of programs. To illustrate an average of the amount sequences annotation by programs, see Figure 2.

**Correlation analysis**

To test for correlations between genome size and transposable elements percentage by genome in base pairs, we first normalized using log10, and then we applied the Pearson Correlation Coefficient in SPSS version 25.

**Web implementation**

APTE is hosted at the Universidade Tecnológica Federal do Paraná (Cornélio Procópio, PR, Brazil). It uses Debian 11 as operating system, Apache 2 as web server, PHP 5.6 as web programming language. We also used Zend Framework 2, which implements model, view, controller (MVC), a methodology for web development that can be expanded for any future additional functionality. On the front-end, we used HyperText Markup Language 5 (HTML5), Cascading Style Sheet 3 (CSS3) and JavaScript to perform dynamic functions that provide user-friendly navigation. A built-in genome browser (JBrowse, version 1.14.1) is available to visualize and download the data as well.

**Computational resources**

To run the pipeline described in Figure 1, we used three platforms: (i) to run RepeatModeler, RepeatScout, RepeatMasker, LTR_retriever, MITE-Hunter and HelitronScanner; (ii) to perform MGEScan-non-LTR and PASTEClassifier; and (iii) to unify and filter outputs to the main annotation. The hardware utilized were (i) Xeon E7540 2.00 GHz 256GB memory, Xeon E5-2620v3 2.40 GHz 64GB memory, 2x Intel i7-3820 3.60GHz 32GB memory and Intel i7-3820 3.60GHz 64GB memory, (ii) Intel i7-3820 3.60GHz 64GB memory, and (iii) Intel Core 2 Duo 2.4 GHz 8GB memory, a total of 30 physical cores and 456 GB of memory. In order to present a scale of time elapsed to measure, filter and standardize the results, we estimate that for the *A. thaliana* genome, the time needed to get the final annotation was ~18 hours, using all resources mentioned, including post-processing scripts (detailed on our website).

**Results and discussion**

**Overview of TE portion**

We retrieved a total of 49,802,023 TE records from 67 plant genomes, representing a total of 47,992,091,043 (~47.62%) base pairs (bp) of the total genomic space. This information is distributed in ~57.36% (28,565,034) TEs organized into

![Figure 3. Class, order and superfamilies identified among the 67 plant genomes used in this study.](image-url)
class, order and superfamily. In addition, ~42.64% (21,236,989) elements could not be assigned to any type of known TE and they were labelled as unknown. They likely represent chimeric and/or partial elements for which we were not able to perform the full classification. For known TEs, we identified that ~62.85% were retrotransposons, and ~37.15% were DNA transposons. All assigned classifications of TEs identified along the 67 genomes are shown in Figure 3. The distribution of TEs in the analyzed genomes are somewhat similar (Figure 4), especially in genomes that have a shorter phylogenetic distance (e.g., *Oryza* spp, *Triticum* spp). However, even close-related genomes exhibit uneven TEs distribution (e.g., *Arabidopsis* spp). Two main hypotheses might explain the variation of TE content: (a) different evolutionary stories, since the two major genome duplication events are shared by all seed plants (*epsilon*) and flowering plants (*gamma*), followed by the lineage-specific duplication events,20 and (b) specific pressures to maintain, expand and purge TEs in each lineage.

We have noted that our approach permitted better TE annotations in genomes assembled at chromosome scale, and we also observed that the amount of TEs is generally related to the genome size, since larger genomes have higher occurrences of TEs (Figure 5). However, for incompletely and draft-assembled genomes, it tends to decrease the number of TEs, once the assembly into small parts (scaffolds or contigs) may impact the genome assembly quality, collapsing repeated contigs (mostly TE-derived) and interfering with the proper identification of these TEs.

![Figure 4. Overview of Class I and Class II composition of TEs in each genome organized in a phylogenetic tree.](image-url)
TE database comparison

To compare the results of the identification performed and to ensure the reliability (details in Supplementary Material 1, Extended data) of our approach, we used SPTEdb\textsuperscript{21} annotation data of the genome \textit{Populus trichocarpa} (black cottonwood), which is explored in Table 1. The second comparison of TE annotations was performed for the \textit{Glycine max} (soybean) genome, in which we used SoyTEdb\textsuperscript{22} to compare the amount vs. type of TEs, shown in Table 1. A third comparison used data from GrTEdb\textsuperscript{23} to explore the amount of TEs in \textit{Gossypium raimondii} (cotton), available in Table 1.

Figure 5. Correlation between genome size and TE content. On the left, the bar chart in blue, the genome size (in Gb), and, in green, the transposable elements distribution in analyzed genomes (in percentage). On the right, we normalized, in base pair, genome size and TE using log(10) and then we correlated (Pearson) the genome size by transposable elements. \(r\) and \(p\)-value are shown in the top-left of each chart. A) Using all the 67 annotated genomes; B) For all genomes with recent WGD (Whole Genome Duplication) events, blue circles; C) Excluding genomes that experienced recent WGD, red circles.
Table 1. Transposable elements amount annotated in *Populus trichopoda*, *Glycine max* and *Gossypium raimondii* by class, order and superfamily compared to other databases which have their own annotation.

<table>
<thead>
<tr>
<th>Class</th>
<th>Order</th>
<th>Superfamily</th>
<th>Quantity</th>
<th>Our work</th>
<th>SPTedb</th>
<th>Glycine max</th>
<th>SoyTEdb</th>
<th>Gossypium raimondii</th>
<th>GrTEdb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td></td>
<td>LTR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cassandra</td>
<td>17</td>
<td>-</td>
<td>1,126</td>
<td>-</td>
<td>157</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caulimovirus</td>
<td>469</td>
<td>6</td>
<td>574</td>
<td>-</td>
<td>1,479</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copia</td>
<td>28,259</td>
<td>1,557</td>
<td>60,399</td>
<td>13,318</td>
<td>50,543</td>
<td>2,929</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gypsy</td>
<td>57,505</td>
<td>5,587</td>
<td>92,396</td>
<td>19,052</td>
<td>161,281</td>
<td>10,368</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LARD</td>
<td>142</td>
<td></td>
<td>6,804</td>
<td>-</td>
<td>2,832</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TRIM</td>
<td>1,164</td>
<td>-</td>
<td>3,698</td>
<td>-</td>
<td>2,239</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pao</td>
<td>-</td>
<td>140</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unknown</td>
<td>38,146</td>
<td>-</td>
<td>22,978</td>
<td>-</td>
<td>42,814</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LINE</td>
<td>CR1</td>
<td>84</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRE</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>I</td>
<td>173</td>
<td>-</td>
<td>168</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L1</td>
<td>3,472</td>
<td>87</td>
<td>11,897</td>
<td>-</td>
<td>13,018</td>
<td>299</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L2</td>
<td>695</td>
<td>-</td>
<td>177</td>
<td>-</td>
<td>215</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RTE</td>
<td>362</td>
<td>-</td>
<td>3,927</td>
<td>-</td>
<td>78</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEX</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unknown</td>
<td>4,484</td>
<td>-</td>
<td>6,119</td>
<td>182</td>
<td>3,572</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LINE</td>
<td>rRNA</td>
<td>174</td>
<td>-</td>
<td>193</td>
<td>-</td>
<td>509</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>tRNA</td>
<td>18,775</td>
<td>-</td>
<td>2,784</td>
<td>-</td>
<td>850</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>snRNA</td>
<td>-</td>
<td>-</td>
<td>64</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unknown</td>
<td>12,922</td>
<td>-</td>
<td>4,152</td>
<td>-</td>
<td>4,485</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NGARO</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLE</td>
<td>-</td>
<td>-</td>
<td>14</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DIRS</td>
<td>-</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Class II</td>
<td></td>
<td>TIR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CMC-EnSpm</td>
<td>7,031</td>
<td>3</td>
<td>26,395</td>
<td>-</td>
<td>1,547</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crypton</td>
<td>-</td>
<td>-</td>
<td>535</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dada</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EnSpm-CACTA</td>
<td>3,590</td>
<td>-</td>
<td>4,596</td>
<td>65</td>
<td>1,419</td>
<td>275</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harbinger</td>
<td>584</td>
<td>-</td>
<td>1,182</td>
<td>-</td>
<td>1,036</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hAT</td>
<td>3,383</td>
<td>17</td>
<td>3,348</td>
<td>65</td>
<td>4,837</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hAT-Ac</td>
<td>3,301</td>
<td>-</td>
<td>9,676</td>
<td>-</td>
<td>22,333</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hAT-Charlie</td>
<td>-</td>
<td>-</td>
<td>180</td>
<td>-</td>
<td>134</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hAT-Tag1</td>
<td>11,022</td>
<td>4,612</td>
<td>4,650</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hAT-Tip100</td>
<td>935</td>
<td>-</td>
<td>2,805</td>
<td>-</td>
<td>7,733</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maverick</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>115</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MuDR</td>
<td>2,246</td>
<td>-</td>
<td>15,920</td>
<td>2,373</td>
<td>9,933</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MuLE-MuDR</td>
<td>2,712</td>
<td>6</td>
<td>27,566</td>
<td>-</td>
<td>11,207</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Novosib</td>
<td>74</td>
<td>-</td>
<td>577</td>
<td>-</td>
<td>55</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PIF-Harbinger</td>
<td>6,430</td>
<td>3</td>
<td>8,157</td>
<td>90</td>
<td>4,342</td>
<td>435</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

Our analysis brought an exhaustive, systematic and comprehensive genome identification in plant genomes, using seven programs to annotate TEs in plant genomes. In both TE classes, several orders and superfamilies were found ubiquitously in all genomes. Additionally, 21,236,989 out of 49,802,023 mapped TE sequences could not be classified into any of the nomenclatures known for TEs, and were labeled as “Unknown” in GFF3, a standard file format for gene annotation.

For plant species whose TE complement may be quite well-annotated, i.e., *Arabidopsis thaliana*, we yielded an increased number of identified TEs. In species with less curated annotation in Ensembl, we were able to deliver a more detailed identification of TEs. For example, in three particular genomes, i.e., *Populus trichocarpa* (black cottonwood), *Glycine max* (soybean) and *Gossypium raimondii* (cotton), we increased the TE identification levels by 2,295%, 900% and 2,643%, respectively. We observe that for several other genomes, new types of TEs were identified and annotated; this ensures that our pipeline delivers not only the same TE identification, but also new ones, making the annotation process possible to use for any species.

In this study, we contributed to expand the knowledge on TEs, by providing a large-scale, organized and standardized TE Atlas. We integrated all annotations to make it available to download in each genome separately from the Atlas of Plant Transposable Elements (APTE) website. An example how our pipeline works using the *A. thaliana* genome, software dependencies, and in-house scripts developed, which can be downloaded, used and changed freely, are available from https://github.com/alerpaschoal/apte_pipeline/.

Data availability

Underlying data

All data underlying the Plant TE Atlas is available in the portal http://apte.cp.utfpr.edu.br/.

Extended data


This project contains the following extended data:

- SuppMat_1.xlsx (the genome assembly reference access from Ensembl Plants species used)
- SuppMat_2.docx (a brief transposable elements annotation steps used in this work)

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).
References


32. PASTECclassifier. Reference Source

Gökhan Karakülah
Izmir Biomedicine and Genome Center, Dokuz Eylül University, Izmir, Turkey

In this work, Pedro and colleagues present a novel transposable elements database for plants using publicly available genome sequences from the Ensembl Plant database. The methodology was clearly described and seems reproducible. The main output of the current work is a publicly available database named APTE and its interfaces are elegantly designed. All transposon annotations are available without any restriction and the database is easy-to-use.

I believe this unique and valuable database has potential to help researchers in both computational and wet lab analyses of mobile elements.

Is the work clearly and accurately presented and does it cite the current literature?  
Yes

Is the study design appropriate and is the work technically sound?  
Yes

Are sufficient details of methods and analysis provided to allow replication by others?  
Yes

If applicable, is the statistical analysis and its interpretation appropriate?  
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?  
Yes

Are the conclusions drawn adequately supported by the results?  
Yes

Competing Interests: No competing interests were disclosed.
Reviewer Expertise: Bioinformatics, Genomics, Mobile elements, Repetitive DNA, Non-coding RNAs

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com