PlantRep: a resource of plant repeats

Xizhi Luo
Chinese Academy of Agricultural Sciences Agricultural Genomes Institute at Shenzhen
Shiyu Chen
Chinese Academy of Agricultural Sciences Agricultural Genomes Institute at Shenzhen
Yu Zhang (zhangyu07@caas.cn)
Chinese Academy of Agricultural Sciences Agricultural Genomes Institute at Shenzhen
https://orcid.org/0000-0001-6547-6243

Research Article

Keywords: database of repeat sequences, evolution of plant genome, transposable element

Posted Date: October 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-498874/v2

PlantRep: a resource of plant repeats

Xizhi Luo[1], Shiyu Chen[1], Yu Zhang[1,2,*]

Key Message

We re-annotated repeats of 459 plant genomes and released a new datatabase: PlantRep (http://www.plantrep.cn/). PlantRep sheds lights of repeat evolution and provides fundamental data for deep exploration of genome.

Key words: database of repeat sequences, evolution of plant genome, transposable element.

[1] Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China

[2] School of Agriculture, Sun Yat-sen University, Shenzhen 518107, China

Email: zhangyu07@caas.cn

Introduction

With the rapid release of genome, exploring the comparative genomics of repeats enabled us to elucidate how repeat sequences originated and amplified in different plant lineages. However, large-scale evolution analysis in plants takes a considerable amount of time and computing resources, hence the comparative genomics and evolution analysis of transposons have only been conducted in a few selected plants[1-3] and the current reference repeat databases only contain a few model plants[4-6]. Besides, the repeat annotations carried out independently using heterogeneous pipelines cannot be used directly for comparative studies. To cope with this issue, we utilized a uniformed pipeline to re-annotate repeats for 459 plant genomes and compared the repeat sequences among the plant groups, including the composition, family diversity, genomic distribution and evolutionary rate. The results provide a resource for the analysis and study of the repeat sequences in different lineages of plants.

Result and Discussion

We re-annotated repeats from 459 released plant genomes and generated 45.72 Gb seed alignments and 601,731 consensus sequences of repeats from *de novo* repeat annotation of each plant genome. The repeat libraries are available in our database PlantRep. Combined with the reference-based annotation, 206.04 Gb of 396,041,410 repeats were identified and categorized. We released this consensus repeat annotations (PlantRep) for the plant community as an updated resource for the future data-mining studies. Repeats in the PlantRep database were categorized with repeat types adapted from the existing eukaryotic transposable element classes and the Dfam database[6,7] (Supplementary Table 1c). Retrotransposon includes Long Terminal Repeat (LTR), Inverted Long Terminal Repeat (DIRS), Penelope (PLE), Long Interspersed Nuclear Element (LINE), and Short LINE-dependent Retroposons (SINE). DNA transposons were categories into Terminal Inverted Repeat (TIR), Circular dsDNA Intermediate (CirdsDNA), DNA Polymerase (DP), and Circular ssDNA Intermediate (Rolling Circle, RC). Besides, low complexity, satellite and simple repeat were also included.

To examine the diversity of repeat families, the 459 plant species were divided into 15 clades based on their phylogeny: algae, bryophyte, lycophytes, fern, gymnosperms, ANA (early angiosperms), magnolids, monocots, base eudicots, super rosids, fabids, malvid, super asterids, lamiids, and campanulids[8] (Supplementary Table 1b; Supplementary Fig. 2). The abundance and diversity of each repeat type in each clade were characterized.

We examined the distribution of different types of repeat sequences in each clade (Supplementary table 1c). The average percentage of repeats within the genome across all species was 45.49%. The top five abundant types of repeats are LTR, TIR, LINE, Simple repeat and Rolling Circle, accounting for 21.66%, 5.44%, 2.25%, 1.54%, and 0.59% of the plant genome on average (Fig 1a; Supplementary table 2), respectively. Plants from different lineages display distinct proportion of repeat types. In general, the proportion of repeats increased from algae, bryophyte, lycophytes, fern to gymnosperms (Supplementary Fig. S2-S15). LTRs, as the largest family of plant repeats, might be the major contributor for the increase of total TE population. For algae, the proportions of simple repeats and LINEs exceeded LTRs, which were the highest in all the lineages, suggesting a possible mechanism of controlling the LTR amplification in algae. Ferns showed a higher proportion of SINE transposons. In ANA, TIRs accounted for a prominent proportion of repeats. The diverse compositions of the repeat types among plant species could provide a source for their unique genome evolution trajectory.

To trace back the evolutionary history of plant transposons, we analyzed the presence/absence of types of repeats (Fig 1b; Supplementary table 3; Supplementary Fig. S16, S17). We found that most of the repeat families existed in algae, indicating the common ancestor of algae and land plants had already evolved the fundamental transposon layout of modern green plants (Fig 1b).

The diversity of transposon nucleotide can reflect the evolution rate of transposon to some extent, we investigate the nucleotide diversity of transposons for each plant linage (Fig 1c; Supplementary table 4). In plants, the main nucleotide diversity peaks of the LTR concentrates is at 16%, and the TIR is about 20%, but the LINE concentrates is around 25%, indicating that LINE lack recently replication activity compared to LTR and TIR. The result to some extent explains the genome different between plant and animal: the proportion of LINE in the plant genome is usually

less than 5%, but in some animal, LINE is the main repeat content of genome. As the main contributor to plant genome, LTRs were selected to investigate the evolutionary history of amplification (Fig 1c; Supplementary table 4). The main peak of algae is 0.2–0.23, which is smaller than the peaks of all of the other species, implying that algae carried ancient LTR groups. LTR amplified more recently in bryophytes compared with algae. The nucleotide diversity of LTR in ferns and lycophytes showed the lowest diversity across all the land plants. The amplification of LTR in the genome of gymnosperms leads to a large genome, and the results show that the degree of divergence of LTR is high, indicating that it has no recent activity, which is consistent with the results of Norway spruce. The nucleotide diversity of ANA and magnoliids is similar as that of gymnosperms. Monocots and base eudicots displayed a more recent amplification. For dicots, the nucleotide diversity was broad, indicating several rounds of amplification of LTRs. According to the divergence of LTRs, we also estimated the amplification time of LTRs along the plant, with the main detected amplicon that can be traced back to 1-4 Mya ago (Supplementary fig. S18). This indicates that the LTRs amplify independently in each linage, playing important roles in the evolution of genome size and environmental adaptability.

To elucidate the contribution of repeats to genes, we calculate the frequencies of repeat sequences at different sites around genes. We found that LTRs, LINEs, SINEs, and DNA transposons display decrease of frequency from 1 kb upstream to 1kb downstream of transcriptional start site genes (Supplementary Fig S19; Supplementary table 5), indicating that plants tend to suppress transposon insertions around gene transcription start sites (TSS), the transposons located near the gene might impact the expression and function of gene. The common feature of LTRs and LINEs near the gene is that there is an inflection point which falls sharply from 2 kb upstream of the TSS to the lowest frequency at 1 kb within genes TSS (Supplementary Fig S19-S21). It then continues to rise within the gene to a distribution of 10k. The frequency of 10kb within gene is close to or even higher than the frequency at 10kb upstream of the TSS. Similarly, the frequency of SINE and DNA transposons from 1kb upstream of the TSS to 1kb within the gene decreases, in gene internal frequency rises; but then the frequency shows a downward or stable trend within gene (Supplementary Fig S19, S22), which is different from LTRs and LINEs. The result implies different transposon families might adopt specific integration strategies and occupy different “niches” of

genome[9]. Unlike TEs, the frequency of simple repeats near TSS is opposite. Simple repeats increase from 4.5 kb upstream of TTS to 0.5 kb within the gene where it reaches the highest frequency (Supplementary Fig S19) , which is similar to the results of before study[10]. Therefore, one can speculate that the high frequency of simple repeats around the gene provides a certain fault tolerance rate for the stability of gene transcription.

In summary, we re-annotated repeats of 459 plant species and characterized the abundance, presence/absence and nucleotide diversity of the repeat types for 15 plant taxonomic groups. The frequency of repeats along the gene models showed unique patterns for different repeat types. Our work supplies a new resource for the future study of repeat sequences and will be helpful to plant genome structure annotation.

Declarations

Funding

This work was supported by the National Natural Science Foundation of China (32070250), the Natural Science Foundation of Guangdong Province (2020A1515011030) and the open research project of "Cross-Cooperative Team" of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the website: http://www.plantrep.cn/.

Code availability

The code used during the current study available from the corresponding author on reasonable request.

Ethics approval and

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Authors' contributions

X.L. and Y.Z. designed the study. X.L. and Y.Z. wrote the manuscript. S.C. reviewed and edited the manuscript. X.L. analyzed data. Y.Z. oversaw the study.

Acknowledgements

We thank Sanwen Huang and Xinyan Zhang for providing suggestions, Shujun Ou for discussion on intact LTR insert time calculation. Computational support was provided by the Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture and Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. This work was supported by the National Natural Science Foundation of China (32070250), the Natural Science Foundation of Guangdong Province (2020A1515011030) and the open research project of “Cross-Cooperative Team” of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences.

Reference

1 El Baidouri, M. & Panaud, O. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome biology and evolution* **5**, 954-965, doi:10.1093/gbe/evt025 (2013).

2 Elliott, T. A. & Gregory, T. R. Do larger genomes contain more diverse transposable elements? *Bmc Evol Biol* **15**, 69, doi:10.1186/s12862-015-0339-8 (2015).

3 Schaper, E. & Anisimova, M. The evolution and function of protein tandem repeats in plants. *New Phytol* **206**, 397-410, doi:10.1111/nph.13184 (2015).

4 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467, doi:10.1159/000084979 (2005).

5 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11, doi:10.1186/s13100-015-0041-9 (2015).

6 Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, D81-89, doi:10.1093/nar/gkv1272 (2016).

7 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982, doi:10.1038/nrg2165 (2007).

8 Group, T. A. P. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**, 1-20, doi:10.1111/boj.12385 (2016).

9 Zhang, X., Zhao, M., McCarty, D. R. & Lisch, D. Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. *Nucleic Acids Res*, doi:10.1093/nar/gkaa370 (2020).

10 Huda, A., Mariño-Ramírez, L., Landsman, D. & Jordan, I. K. Repetitive DNA elements, nucleosome binding and human gene expression. *Gene* **436**, 12-22, doi:10.1016/j.gene.2009.01.013 (2009).

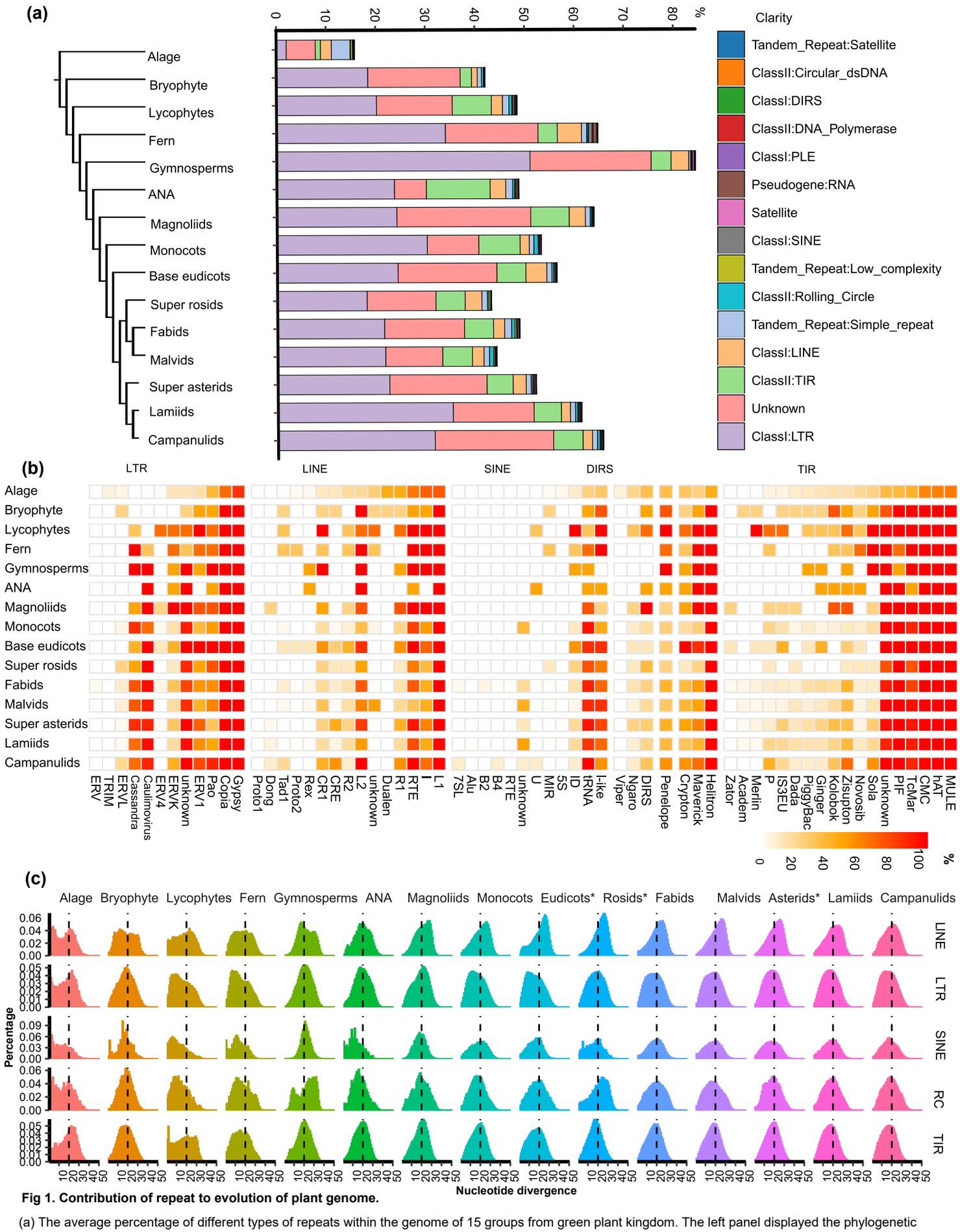

Fig 1. Contribution of repeat to evolution of plant genome.

(a) The average percentage of different types of repeats within the genome of 15 groups from green plant kingdom. The left panel displayed the phylogenetic trees of the plant lineages. (b) The percentage of species carrying certain repeat family in each group. (c) Nucleotide diversity of transposons within each plant lineages. The x-axial label means the nucleotide diversity percentage of repeat. The vertical dashed line represents the divergence rate of 20%. Eudicots*, base eudicots; Rosids*, super rosids; Asterids*, super asterids.

Supplementary Files

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

- PlantReparesourceofplantrepeatsSupplementaryFigure.pdf
- PlantReparesourceofplantrepeatsSupplementaryMaterial.pdf
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xls
- SupplementaryTable5.xls