Molecular Plant
Correspondence

CellPress
Partner Journal

The Chromosome-Level Reference Genome of Tea Tree Unveils Recent Bursts of Non-autonomous LTR Retrotransposons in Driving Genome Size Evolution

Dear Editor,

The tea tree *Camellia sinensis*, a member of the genus *Camellia* in the Theaceae family, includes two major cultivated varieties, *C. sinensis* var. *assamica* (*CSA*; Assam type) and *C. sinensis* var. *sinensis* (*CSS*; Chinese type) (Ming and Bartholomew, 2007). Due to the high economic importance of the tea tree, considerable efforts have been made to explore genetic basis of the biosynthesis of natural metabolites that determine health benefits and diverse tea flavors (Shi et al., 2011; Li et al., 2015; Xia et al., 2017; Liu et al., 2019). We first *de novo* sequenced and assembled the highly heterozygous genome sequence of *C. sinensis* var. *assamica* cv. *Yunkang-10* (*CSA-YK10*) (Xia et al., 2017). Subsequently, a draft genome of *C. sinensis* var. *sinensis* cv. *Shuchazao* (*CSS-SCZ*) was reported, which was generated by using the same sequencing platform and then filling gaps with PacBio long reads (Wei et al., 2018). However, it remains a great challenge to obtain a high-quality genome assembly of the tea tree, because short Illumina reads and even hybrid assembly strategies have intrinsic difficulties in generating the chromosome-level, high-quality *de novo* assembly of complex large plant genome harboring highly heterozygous and repetitive DNA sequences due to its self-incompatibility.

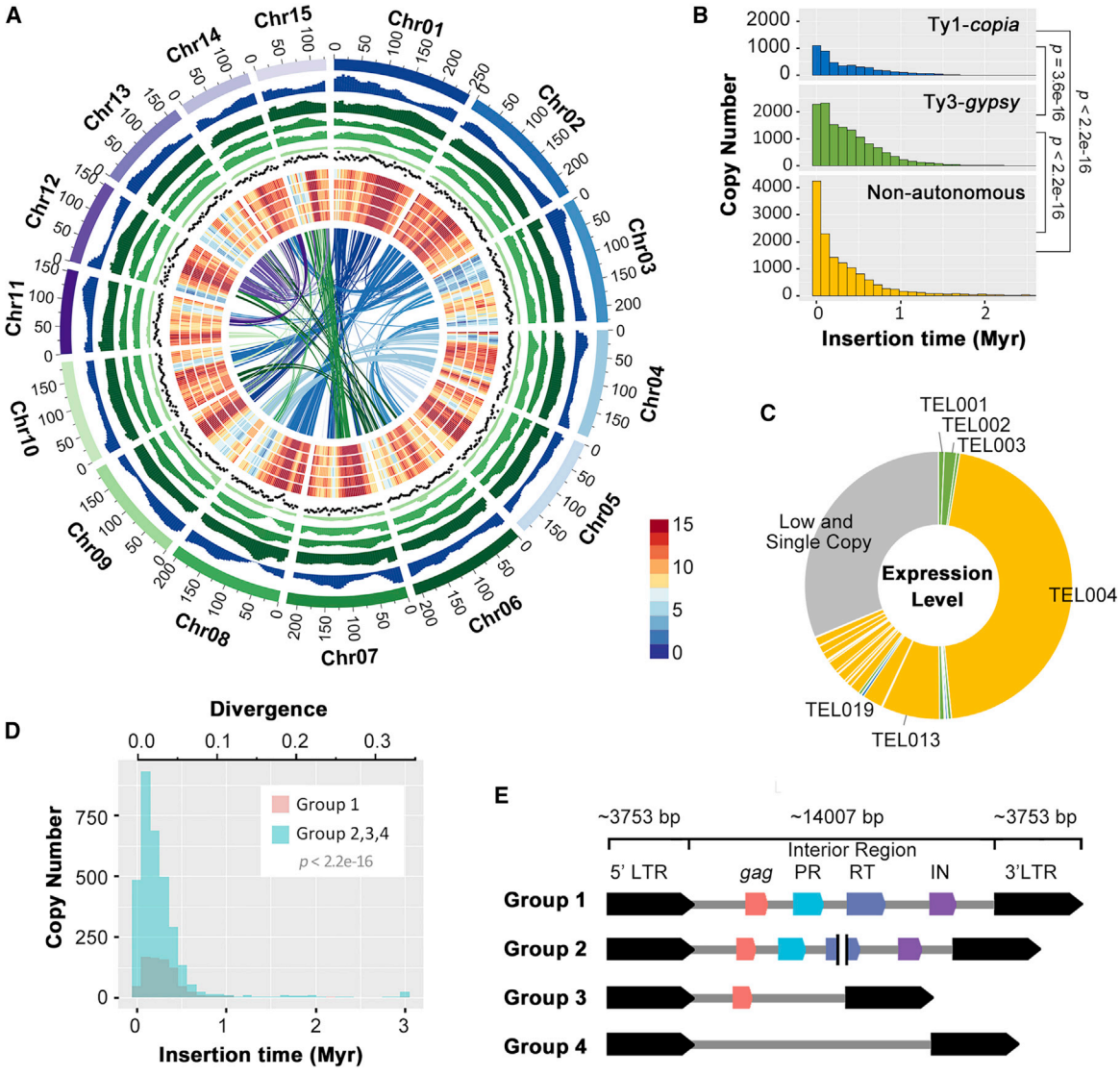
To further investigate the tea tree genome size evolution, we generated the high-quality chromosome-length reference genome of *C. sinensis* var. *sinensis* by using long-read single-molecule real-time (SMRT) (~417.95 Gb, ~127.66-fold coverage) and Hi-C (909 454 810 Hi-C reads) sequencing technologies (Supplemental Tables 1–3). We first employed the Illumina short-read technology with paired-end libraries on the HiSeq X Ten sequencing platform to screen 12 representative tea tree cultivars. We then selected the commercial variety (*CSS-BY*) for long-read genome sequencing due to its relatively low heterozygosity (~1.22%). We estimated that the genome size of *CSS-BY* is ~3.25 Gb using 17-*mer* analysis (Supplemental Figure 1 and Supplemental Table 1). We obtained a final assembly of ~2.92 Gb, accounting for ~89.85% of the estimated genome size, ~2.86-Gb (~97.87%) of which was anchored into 15 pseudochromosomes (Figure 1A and 1F; Supplemental Figure 2; Supplemental Tables 4–6). The assembly comprised 13 006 contigs with a contig N50 length of ~625.11 kb, ~9.32-times longer than the previously reported genome assembly of *C. sinensis* var. *sinensis* cv. *Shuchazao* (*CSS-SCZ*) (~67.07 kb) (Wei et al., 2018) (Figure 1F). The assembly was composed of 4153 scaffolds with a scaffold N50 length of ~195.68 Mb, ~140.78-times longer than the previously reported genome assembly of *C. sinensis* var. *sinensis* cv. *Shuchazao* (*CSS-SCZ*) (~1.39 Mb) (Wei et al., 2018) (Figure 1F). The lengths of 15 chromosomes of the *CSS-BY* genome ranged from ~253 Mbp (Chr01) to ~128 Mbp (Chr15) with an average size of ~190 Mbp (Figure 1A and Supplemental Table 6).

We predicted a total of 40 812 protein-coding genes (Figure 1F), of which 34 722 (85.08%) were supported by transcriptome-based evidence (Supplemental Tables 12–14). The average gene length and exon number were 6263 bp and 5.2 per gene, which are much higher than those in *CSS-SCZ* with 4053 bp and 3.3 per gene, respectively (Figure 1F and Supplemental Figure 4). The annotation of noncoding RNA genes yielded 659 transfer RNA, 2845 ribosomal RNA, 471 small nucleolar RNA, 207 small nuclear RNA, and 139 microRNA genes (Supplemental Table 15). We performed comprehensive analyses of the *CSS-BY* and *CSS-SCZ* genome assemblies (Figure 1F; Supplemental Table 17; Supplemental Figures 3B, 3C, and 10–12). We observed that, compared with the previously reported *CSS-SCZ* genome assembly (Wei et al., 2018), the SMRT sequencing and assembly strategy has produced a *CSS-BY* assembly of superior contiguity containing accurate long-range information.

Using this new genome assembly of *CCS*we annotate functionally important gene families such as those involved in the biosynthesis of secondary metabolites. We attempted all 23 gene families encoding enzymes potentially involved in catalyzing reactions of the flavonoid, theanine, and caffeine pathways (Supplemental Tables 18–21; Supplemental Figures 13–16) and produced more reliable annotations of almost all gene families as compared with the previously reported *CSS-SCZ* genome assembly (Wei et al., 2018).

The long reads generated by SMRT technology also allow to characterize almost all transposable elements (TEs) in *CSS-BY* genome(Figure 1A and Supplemental Figure 3A). Ty3-*gypsy* LTR retrotransposon elements dominate the genome with ~34.11% (~996.15 Mb) of the assembled sequence length, ~7.11-fold larger than Ty1-*copia* LTR retrotransposon families (~140.11 Mb; ~4.80%) and ~2.03-fold larger than non-autonomous LTR retrotransposon families (~490.84 Mb; ~16.81%) (Supplemental Table 10 and Supplemental Figure 3A). We classified all full-length LTR retrotransposons into 8844 families, of which the top 111 families with more than

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, CAS.



F

Assembly	*CSS-BY*	*CSS-SCZ*	Annotation	*CSS-BY*	*CSS-SCZ*
Estimated genome size (Gb)	3.25	2.98	No. of protein-coding genes	40,812	33,932
Total length of scaffolds (Gb)	2.92	3.14	Average gene length (bp)	6,263	4,053
Coverage of the assembly	89.85%	105.37%	Average exon length (bp)	251	259
Scaffold number	4,153	14,051	Average exon per gene	5.2	3.3
N50 of scaffolds (Mb)	195.68	1.39	Mean intron length (bp)	1,168	1,408
N50 of contigs (Kb)	625.11	67.07	Masked repeat length (Mb)	2,165	2,008
GC content of the genome	38.24%	37.84%	Percentage of repeats	74.13%	64.78%

Figure 1. The Genome Features of *C. sinensis* var. *sinensis* cv. *Biyun*.

(A) Circular representation of the 15 pseudochromosomes. From outside to inside, the density of genes, the distribution of TEs, Ty3-*gypsy* LTR retrotransposons, Ty1-*copia* LTR retrotransposons, DNA TEs, the density of single sequence repeats, the relative transcript levels for young leaf, tender shoot, flower bud, fruit, and stem, and genomic synteny are shown.

(B) Insertion times of Ty1-*copia* (blue), Ty3-*gypsy* (green), and non-autonomous (yellow) LTR retrotransposons. The insertion times for LTR retrotransposons were calculated by the formula $T = K/2r$, where T is insertion time, r is synonymous mutations/site/Myr, and K is the divergence between the two LTRs. A substitution rate of 5.62×10^{-9} per site per year was used to calculate the insertion times.

(C) Expression levels calculated by transcripts read count of LTR retrotransposon families. All transcripts from five tissues were collected using HISAT2 and StringTie to classify the LTR retrotransposons related transcripts into different LTR families by BLAST. Reads number of each LTR retrotransposon family were then counted by HTSeq.

(D) Insertion times of LTR retrotransposons. The distribution differences in (B) and (D) were evaluated by Wilcoxon rank-sum test.

(E) Structural features of the four groups of the top *TEL001* retrotransposon family.

(F) Global statistics for the assembly and annotation of the two *Camellia sinensis* var. *sinensis* genome assemblies.

10 copies contained 75% full-length LTR retrotransposons and occupied 36.47% of the genome (Supplemental Table 22). A total of 13 172 Ty3-*gypsy* and 4630 Ty1-*copia* retrotransposon sequences were extracted to construct phylogenetic trees (Supplemental Figure 17), yielding 11 lineages, consistent with previous results (Vitte et al., 2007; Wicker and Keller, 2007; Llorens et al., 2009; Hřibová et al., 2010). The repetitive nature of tea tree genome is determined by a handful of LTR retrotransposon families with extremely high copy numbers; for example, the amplification of *Tat* (~671.13 Mb; ~22.98%) and *Tekay* (~303.84 Mb; ~10.41%) of Ty3-*gypsy* has largely contributed to the large size of tea tree genome (Supplemental Figures 3B, 17B, and 20). Notably, incessant bursts of the *Tat* lineage predominantly came from eight of the top 12 families, resulting in ~50% of full-length LTR retrotransposons that accounted for ~29.65% of this genome assembly (Supplemental Table 22). The largest family *TEL001*, for instance, contains 4062 full-length LTR retrotransposons with the longest average length of 18 204 bp, contributing most to the genome size (~18.27%) (Supplemental Figure 3B and 3E; Supplemental Table 22). Meanwhile, *Ale*, *TAR*, *GMR*, *Maximus Angela*, and *Ivana* of Ty1-*copia* retain full-length LTR retrotransposons, suggesting that Ty1-*copia* has experienced a long and slow amplification history (Supplemental Figures 17A and 20).

The continuous genome assembly provides new insights into evolutionary dynamics of LTR retrotransposons in the tea tree genome. The analyses of retrotransposon sequences from the two major tea tree variety genomes, *CSS-BY* and *CSA-YK10*, showed that they may have experienced a similar evolutionary history, except that considerably large numbers of retrotransposons (e.g., *Tat* and *Tekay* lineages) were detected in the SMRT-based *CSS-BY* genome assembly (Supplemental Figure 17C and 17D). In sharp contrast to the failure to assemble the recently generated retrotransposons from the previously reported *CSS-SCZ* genome assembly (Supplemental Figures 18 and 19), the resulting 32 367 full-length LTR retrotransposons account for nearly 18.5% of the assembled sequence length, allowing us to further date the very recent evolutionary history of LTR retrotransposons in the *CSS-BY* genome (Supplemental Table 11 and Supplemental Table 22). The expansion of Ty3-*gypsy* retrotransposon families makes the genome currently predominate (Supplemental Figure 3D and 3E), such as *Tat* members of Ty3-*gypsy*, which have rapidly amplified during the last 1 million years (Myr) before declining rapidly (Figure 1D; Supplemental Figures 3E, 20, and 21; Supplemental Table 22). Surprisingly, the *Tekay* lineage of Ty3-*gypsy* (e.g., *TEL005*, *TEL021*, and *TEL022*) and non-autonomous LTR retrotransposon families (e.g., *TEL004* and *TEL010*) (Supplemental Figures 20 and 21; Supplemental Table 22) were found to be predominant in recently affecting the dynamic genome size variation (Supplemental Figure 3E). It is of great interest to observe recent insertions of many single-copy LTR retrotransposon families, although the retrotransposon abundance is expectedly governed by recent activities of multi-copy LTR retrotransposons.

The degree to which non-autonomous LTR retrotransposons impede the proliferation of autonomous retroelements has critical evolutionary impacts on the genome size (Zhang and Gao, 2017). We found a rapid and recent propagation of more than 4000 non-autonomous elements (Figure 1B). Of them, some were derived from autonomous Ty3-*gypsy* or Ty1-*copia* families that have slowly lost internal protein-coding genes. However, it is problematic to determine the counterpart autonomous families for others. *TEL001* was selected as an exemplar to show that partial and/or complete loss of internal protein-coding genes has resulted in a quick increase of incomplete autonomous and/or non-autonomous retroelements that have far exceeded autonomous ancestral elements within the last 1 Myr. Based on structural features of *TEL001*, 4062 full-length LTR retrotransposons were classified into the four groups (Figure 1E and Supplemental Figure 23). Group 1 contains 451 copies with complete sequences of *gag* and *pol* (protease [PR], reverse transcriptase [RT], and integrase [IN]) genes; group 2 comprises 352 copies with the loss of at least one of the *gag*, PR, RT, and IN domains; group 3 had 1063 copies with only the *gag* domain; and group 4 includes 2196 non-autonomous copies without any internal *gag* and *pol* genes (Figure 1E). Due to the dominance of the non-autonomous retroelements, the proportion of effective retrotransposition-related source proteins declines dramatically and insertion rates of the entire *TEL001* family largely decreases most recently (Figure 1D). In addition, there are many non-autonomous families, such as *TEL004*, which is a very young family that has undergone a large number of recent insertions (Supplemental Figure 21). There are also many low-copy and single-copy non-autonomous families reproduced most recently, together making the recent inserted non-autonomous elements far exceed Ty3-*gypsy* or Ty1-*copia* copies (Figure 1B; Supplemental Figures 22 and 24). We then assessed expression levels of all types of LTR retrotransposons using Illumina RNA-sequencing data from the five tissues (Figure 1C; Supplemental Tables 12 and 23). We detected ~16.70% (~7586) of all expressed transcripts and ~10.38% of all mapped reads, on average, for five tissues that are associated with LTR retrotransposons (Supplemental Table 24). About 63.59% of Illumina reads mapped to multi-copy non-autonomous LTR retrotransposon families (e.g., *TEL004*, ~45.88%; *TEL013*, ~7.03%; *TEL019*, ~2.45%) exhibit notably high levels of gene expression than Ty1-*copia* and particularly Ty3-*gypsy* families in multi-copy families (Figure 1C and Supplemental Table 23). Proteins (including *gag*, PR, RT, and IN domains in *pol*) necessary for the retrotransposition were further annotated. Surprisingly, ~94.23% of the expressed LTR retrotransposon-related transcripts are not related to encoding *gag* and *pol* genes, and only 5.77% of the retrotransposon-related transcripts mapped to at least one of the aforementioned genes (Supplemental Table 24). Our findings thus offer more evidence that recently increased non-autonomous LTR retrotransposons with high expression levels may limit the efficiency by reducing the supply of enzymes needed for a successful retrotransposition (Zhang and Gao, 2017).

In conclusion, we have generated a highly continuous and accurate tea tree genome assembly for *CSS-BY* by using SMRT technology combined with Hi-C. This chromosome-level genome assembly of the tea tree is powerful in identifying all types of long LTR retrotransposons and characterizing the abundance of retrotransposon diversity, allowing to resolve the nature of the repetitive landscape of such a large genome. The evolutionary history of very recently augmented LTR retrotransposon

families could also be tracked genome-wide by dating bursts of non-autonomous LTR retrotransposons and measuring their interaction with autonomous LTR retrotransposons that drove the evolution of genome size. Such a high-quality reference genome as the tea tree would be valuable to the broad tea research community, enabling researchers to not only accurately obtain functionally significant gene families but also determine agronomically important traits relevant to the improvement of tea quality and production.

ACCESSION NUMBERS

Raw PacBio and Illumina sequencing reads of *CSS-BY* have been deposited in the National Genomics Data Center under accession number PRJCA002071. Genome assembly, gene prediction, gene functional annotations, and transcriptomic data may be accessed via the web site at www.plantkingdomgdb.com/CSS-BY/.

SUPPLEMENTAL INFORMATION

Supplementary Information is available at *Molecular Plant Online*.

FUNDING

This study was supported by a startup grant from the South China Agricultural University and Yunnan Innovation Team Project (to L.-Z.G.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

L.-Z.G. designed and managed the project; C.S., Y.T., H.N., Y.-L.L., X.-L.Y., and X.-H.W. collected materials; C.S., C.L., C.-F.W., and X.-X.L. prepared and purified DNA and RNA samples; K.L. performed the genome assembly; Q.-J.Z., W.L., H.N., Y.Z., D.Z., L.-Y.F., J.-A.H., W.-K.J., and Z.-Y.D. performed genome annotation and subsequent data analyses; L.-Z.G. and Q.-J.Z. wrote the manuscript; L.-Z.G., Q.-J. Z., Z.-H.L., X.-C.Z., and E.E.E. revised the manuscript.

ACKNOWLEDGMENTS

We appreciate the anonymous reviewers for their comments on the manuscript. The authors thank T. Brown for assistance in editing the manuscript. No conflict of interest declared.

Received: October 29, 2019
Revised: December 26, 2019
Accepted: April 25, 2020
Published: April 27, 2020

Qun-Jie Zhang[1,11], Wei Li[1,11], Kui Li[2,3,11], Hong Nan[4,5,11], Cong Shi[4,5,11], Yun Zhang[4], Zhang-Yan Dai[6], Yang-Lei Lin[4,5], Xiao-Lan Yang[4,5], Yan Tong[4], Dan Zhang[1], Cui Lu[1], Li-Ying Feng[1], Chen-Feng Wang[1], Xiao-Xin Liu[1], Jian-An Huang[7], Wen-Kai Jiang[2], Xing-Hua Wang[8], Xing-Cai Zhang[9], Evan E. Eichler[10], Zhong-Hua Liu[7,] and Li-Zhi Gao[1,4,*]*

[1]Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou 510642, China
[2]Novogene Bioinformatics Institute, Building 301, Zone A10 Jiuxianqiao North Road, Chaoyang District, Beijing 100083, China
[3]School of Life Sciences, Nanjing University, Nanjing 210023, China
[4]Plant Germplasm and Genomics Center, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650204, China
[5]University of the Chinese Academy of Sciences, Beijing 100039, China
[6]Agro-biological Gene Research Center, Guangdong Academy of Agricultural Sciences, Guangzhou, 510640, China
[7]Hunan Agricultural University, Changsha 410128, China
[8]Yunnan Pu'er Tea Tree Breeding Station, No. 212 Zhenxing Avenue, Simao District, Pu Er, Yunnan 665099, China
[9]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[10]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
[11]These authors contributed equally to this article.
***Correspondence: Zhong-Hua Liu (larkin-liu@163.com), Li-Zhi Gao (lgaogenomics@163.com)**
https://doi.org/10.1016/j.molp.2020.04.009

REFERENCES

Hřibová, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J., and Doležel, J. (2010). Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biol. **10**:204.

Li, C.-F., Zhu, Y., Yu, Y., Zhao, Q.-Y., Wang, S.-J., Wang, X.-C., Yao, M.-Z., Luo, D., Li, X., and Chen, L. (2015). Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (*Camellia sinensis*). BMC Genomics **16**:560.

Liu, Z., Gao, L., Chen, Z., Zeng, X., Huang, J.a., Gong, Y., Li, Q., Liu, S., Lin, Y., Cai, S., et al. (2019). Leading progress on genomics, health benefits and utilization of tea resources in China. Nature **566**:7742.

Llorens, C., Munoz-Pomer, A., Bernad, L., Botella, H., and Moya, A. (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol. Direct **4**:41.

Ming, T., and Bartholomew, B. (2007). Theaceae. In Flora of China, Z. Wu, P. Raven, and D. Hong, eds. (Beijing and St. Louis: Science Press and Missouri Botanical Garden), pp. 367–412.

Shi, C.-Y., Yang, H., Wei, C.-L., Yu, O., Zhang, Z.-Z., Jiang, C.-J., Sun, J., Li, Y.-Y., Chen, Q., and Xia, T. (2011). Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics **12**:131.

Vitte, C., Panaud, O., and Quesneville, H. (2007). LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics **8**:218.

Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., Xia, E., Lu, Y., Tai, Y., and She, G. (2018). Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. Proc. Natl. Acad. Sci. U S A **115**:E4151–E4158.

Wicker, T., and Keller, B. (2007). Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res. **17**:1072–1081.

Xia, E.-H., Zhang, H.-B., Sheng, J., Li, K., Zhang, Q.-J., Kim, C., Zhang, Y., Liu, Y., Zhu, T., and Li, W. (2017). The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Mol. Plant **10**:866–877.

Zhang, Q.-J., and Gao, L.-Z. (2017). Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA-genome *Oryza* species. G3 (Bethesda) **7**:1875–1885.