Comparative Evolution and Transcription Activity of Transposable Elements in Indica and Japonica Rice

Zhiguo Wu (wu.zhiguo@whu.edu.cn)
Wuhan University College of Life Sciences https://orcid.org/0000-0002-3558-4478
Yanhua Wu
Wuhan University
Xi Wang
Wuhan University
Yongqiang Zheng
Wuhan University
Jie Zou
Wuhan University
Yongzhuo Guan
Wuhan University
Yang Li
Wuhan University
Yan Yang
Wuhan University
Gai Huang
Wuhan University

Research article

Keywords: transposable element, rice genome, evolution, Gaussian distribution, TE activity

Posted Date: November 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-101515/v1

Abstract

Background

Transposable elements (TEs) are able to diversify plant gene expression and function, sequentially promote plant variety and evolution. However, there is lack of efficient approach to investigate the evolution behavior and transcription activity of TEs in plants. Here we developed a pipeline Matrix-TE to comprehensively evaluate the super-families, differentiation and transcription activity of LTR/TEs in *Indica* and *Japonica* rice, the two considerable important and closely related monocots.

Results

Six LTR/TE super-families were identified by Matrix-TE in both *Indica* and *Japonica* rice genomes, in which the *OS-type1* and *OS-type2* super-families were unclassified. *Indica* rice specific TE peak P-*Gypsy* and *Japonica* rice specific TE peak P-*Copia* were observed separately. Then the two peaks were analyzed by Gaussian Probability Density Function (GPDF) fit. Significant TE transcription activities were observed in *Indica* and *Japonica* rice plants after stress treatments. Particularly, hot, cold and salt stresses induced the high expression level of LTR/TEs in rice plants.

Conclusions

We developed the approach Matrix-TE on the basis of BLASTN and GPDF algorithms, and applied it to comprehensively and quantitatively investigate LTR/TE types and contents in the close subspecies *Indica* and *Japonica* rice genomes. The individual TE burst events P-*Copia* and P-*Gypsy* were observed in *Japonica* and *Indica* rice, separately. RNA-seq and RT-PCR methods indicated that LTR/TE transcripts were induced by hot, cold and high salt stress conditions. The optimized Matrix-TE approach and procedures probably could be used in other plant species with big genomes like wheat and maize.

Background

TEs have been extensively studied for decades, and are extremely important for monocotylous and dicotyledonous plants, as TEs cause plant gene expression and function changes by insert mutations and regulatory element introduction (1). With recent advances in genomics and bioinformatics, it becomes reality to systematically evaluate the role of TEs for plant genome evolution (2, 3). The long terminal repeat transposable elements (LTR/TEs) are retrotransposons which make up the majority of total TEs in most plant genomes (4). Owing to the huge copy number accumulation of LTR/TEs in plant genomes, comprehensive interpreting of the insertion time point and insertion quantity would help to determine the biological role of LTR/TEs for plant evolution history (3).

The genomes of the most widely cultivated *Oryza sativa* subspecies *Indica* and *Japonica* were constructed and updated in the past two decades, genomic variations and wild rice genomes were sequentially studied to reveal the origin and evolution of genus *Oryza* (5–11). The most abundant

Ty1/Copia and *Ty3/Gypsy* were class I retrotransposons in rice genomes and they went through very low nature selection constraints after the insertion in the genome. Then they diverged from the ancestral sequences and become pretty polymorphism over million years of evolution (6). Single nucleotide polymorphisms (SNPs), structural variations and pan-genome constructing represent rice species genetic diversity, further providing new perspectives on rice diversity and evolutionary history (7). The *Oryza* species span ~ 15 million years of evolution, and the divergence time of cultivated *Indica* and *Japonica* rice was estimated at 0.55 million years ago (Mya) (8). Comparative annotation and phylogenetic analysis of LTR-RTs in domesticated and wild rice developed by Joshua C. Stein et. al. revealed several lineage-specific TE burst events within 2.5 million years (8). Molecular phylogenetic studies have been extensively carried out to investigate the relationships of *Indica* and *Japonica* rice, and the results indicated that they were originated independently (9). A well assembled *Indica* genome displayed significantly more solo-LTRs sequences compared with *Japonica*, indicating the two subspecies had experienced independent amplification or loss of LTRs after their divergence (10). It is essential to understand the LTR/TEs differences and similarities for the comprehensive interpreting of rice genome origin and evolution (11).

LTR/TEs content are abundant in monocotylous plants, particularly in wheat and maize as *Ty3/Gypsy* and *Ty1/Copia* super families constitute more than half of the entire genome sequences (12–16). Large quantity of complete and truncated LTR/TEs were identified in wheat D and A subgenomes, intact and solo LTR sequences were applied to the estimation of TE insertion time. Sequential bursts of TEs followed by silencing over the past 3 million years were established by this approach (12, 13). The comparison of the TEs in the maize B73, W22 and Mo17 genomes revealed high level of variations even in very close subspecies (14, 15, 16). TE mediated gene silencing in maize indicated the epigenetic regulation activities in development (17).

Most TEs of various plant species have accumulated huge number of mutations and truncation events in the evolution time course, however a number of TE families remain active and might generate new insertions in some genomes (18–23). Non-autonomous DNA transposons MITEs have been identified in both *Indica* and *Japonica* rice, and they are amplified during stress activations (19, 23). The repetitive nature and variable mutations make TEs challenging to analyze in genomic studies. RNA-seq based procedures have been developed to screen active L1 elements in animal and plant cells (21, 22). Analysis of 3000 rice genomes demonstrated that *mPing* TEs are likely active in the past century (24). A software named TRACK-POSON were developed to detect the insertion polymorphisms of TEs in rice genomes (25). Here we developed the approach Matrix-TE to systematically and quantitatively study the complete and truncated TEs in very close subspecies *Indica* and *Japonica* rice genomes, based on BLASTN and GPDF algorithms (Fig. 1). All LTR/TE super families were scanned and restricted in one super matrix by their ORFs and sequence identities, then classified and phylogenetic clustered. The peaks of TEs were analyzed by Gaussian Probability Density Function, and the nucleotide substitution ratio *Ks* of the subspecies individual TE peaks were calculated. Finally, stress induced expression of TEs in *Indica* and *Japonica* rice were cloned and quantified.

Results

LTR/TEs have been shown to make up bulk of many monocotylous plant genomes, and they are in general randomly inserted across the genome (26). This TE insertion mechanism makes a big trouble because the earlier inserted TEs could be fragmented or truncated by the later insertions at stochastic sites, as a matter of fact huge number of TE fragments are observed in plant genomes (3,26,27). It is rational to compare the nucleotide changes between the two LTR sequences of one complete TE to estimate the insertion time point in the evolution, while it is a big challenge to define the truncated TE pieces without intact LTR sequences (28). Considering the huge number of TEs caused by sequential burst events and the stochastic nucleotide substitutions under low nature selection stress, we developed the pipeline Matrix-TE with GPDF analysis to successfully synthesized the intact and truncated TEs and calculated the *Ks* of TE burst events in rice genomes (Figure 1).

Matrix-TE approach optimized and synthesized the scripts in LTR_finder, EMBOSSY, BioEdit, MEGA and BLASTN packages, to integrate the most abundant TE super families in one super matrix, and quantitatively investigate the individual TE content.

TE matrix and clusters

Rice genome and TE statistics. LTR/TEs ORF length and clusters were presented in Figure 2 (A,B,C,D). The two cultivated *Oryza sativa* rice genome sizes, gene size and total TE size were generally similar for 93-11 and Nip accessions (29). The number of annotated full length LTR/TE, ORFs over 1500bp and ORFs in super matrix were more in *Indica* than *Japonica* rice genomes (Figure 2. E). 2010 of 93-11 and 1520 of Nip ORFs which containing the majority of TE clusters were presented in the super matrix. In which the length of individual TE cluster were the same (Figure 2. A,B), and the identity of individual cluster were greater than 95% (Figure 2. C,D).

Then the TE clusters were extracted and annotated, 93-11 and Nip genomes contained 8 and 6 clusters separately (Figure 2. F). They shared the same 6 TE clusters with the annotation *OS-type1* (C1), *OS-type2* (C2), *OS-typeRT* (C3), *OS-typePHA* (C4), *Ty1/Copia* (C5) and *Ty3/Gypsy* (C6). In which the *OS-type1* (C1) and *OS-type2* (C2) TEs were unclassified, and Cluster 6 (C6), Cluster 7 (C7) and Cluster 8 (C8) in 93-11 genome were 3 subfamilies of *Ty3/Gypsy* and they were combined as C6C7C8 in the following analysis.

Phylogenetic trees and reference sequences of TE clusters. The 6 TE clusters of 93-11 and Nip genomes were constructed phylogenetic trees separately (Figure 3. A,B). *ref-C1*, *ref-C2*, *ref-C3*, and *ref-C4* and *ref-C5* on top branch of the trees were selected as the reference TE sequences for both 93-11 and Nip. *ref-C6C7C8* and *ref-C6* on top branch of the trees were selected as the reference TE sequences for 93-11 and Nip, separately. And the sequences of all the reference TE sequences were presented in Supplementary Table S1.

Genome scanning by ref-LTR/TE and TE peak differences. 93-11 and Nip genomes were scanned with the reference TE sequences extracted from 93-11 and Nip accessions, separately. All the TE super families including intact and truncated sequences were classified and combined according to their annotations (Figure 4. A). The histogram showed that *OS-type1*, *OS-type2*, *OS-typeRT* and *OS-typePHA* TE super families were almost the same contents between 93-11 and Nip genomes. While the contents of *Ty1/Copia* and *Ty3/Gypsy* super families were significantly different between 93-11 and Nip genomes (Figure 4. B,C). All the *Ty1/Copia* and *Ty3/Gypsy* sequences including intact and truncated TE were normalized to distribution curves according to their identities with reference TE sequences. And the individual peaks P-*Copia* and P-*Gypsy* were observed in Nip and 93-11 genomes, separately (Figure 4. D,E). P-*Copia* peak was quite sharp in Nip, while weak in 93-11 genome. P-*Gypsy* peak was strong in 93-11, but hardly observed in Nip genome.

Stochastic SNP distribution on TE and GPDF analysis

SNP distribution across TE. Although the nucleotide substitution rate is usually low in plants, rice genome accumulated abundant SNPs in the genome sequences, especially in TE sequences as they experienced very low nature selection stress in million years of evolution (3,30,31,32). SNP distribution was observed across *Ty1/Copia* and *Ty3/Gypsy* coding sequences in both 93-11 and Nip genomes, without conserved regions. And the SNP level was little higher for *Ty3/Gypsy* gene body than *Ty1/Copia* gene body (Figure 5. A,B).

GPDF fit of P-Copia and P-Gypsy. On the bases of stochastic nucleotide substation analysis of TEs, the individual identity distribution peaks P-*Copia* and P-*Gypsy* were extracted from Figure 4, and the peak curves were fitted well by the mathematic model Gaussian Probability Density Function, with high R square values (Figure 5. C,D). The average nucleotide substitution ratio (*Ks*) of peaks P-*Copia* and P-*Gypsy* were calculated as 2.58σ (3). By GPDF analysis, the *Ks* of P-*Copia* was 0.0043, while the *Ks* of P-*Gypsy* was 0.018.

Rice LTR/TE transcription activities

RNA-seq data mapping. Each of the total RNA for mild condition, hot, cold and salt stress treated samples was sequenced 20 million paired end reads, and the sequencing data were mapped on rice genomes. For both 93-11 and Nip genomes, majority over 95% reads were mapped in gene coding regions, while few reads less than 0.5% were mapped in LTR/TE regions (Table 1). Further indicating the low transcription activities of LTR/TEs in rice species (33).

Table 1
mapping rate of the RNA-seq data for mild condition, hot, cold and salt stress treated rice samples

	Reads in genes (% total reads)				Reads in TE ORFs (% total reads)			
	mild	42℃	5℃	NaCl	mild	42℃	5℃	NaCl
Indica (93-11)	95.75	95.16	95.11	95.51	0.18	0.20	0.20	0.15
Japonica (Nip)	97.57	97.35	97.41	97.05	0.18	0.42	0.26	0.27

Full length LTR/TE cloning from cDNA. 13 full length LTR/TE sequences were cloned from cDNA in 93-11 rice, while 20 full length LTR/TE sequences were cloned from cDNA in Nip rice, under stress treatment conditions. The agarose gel results presented the full length cDNA clones of LTR/TEs, generally the size of transcribed TEs were distributed from 1 kb to 5 kb under variable stress conditions (Figure 6). The length and sequence details of the LTR/TE clones were presented in Supplementary Table S2 and Table S3.

Expression level of rice LTR/TEs under stress treatments. The classified and annotated TE clones were applied to quantification. The RT domain (reverse transcriptase) sequences were amplified in the RT-PCR (2). And the quantified TE expression levels were shown in Figure 7. With the TE expressions under mild conditions as control, significant TE expressions induced by salt and hot stresses were observed in 93-11 rice plants, while hot and cold stressed induced TE expressions were observed in Nip rice plants.

Discussion

Advances in sequencing technology have been extensively applied to resolve the complex plant genomic regions accurately (34). Particularly the long sequencing reads generated by single-molecule technology, together with accurate short read sequencing technologies have dramatically improved representations of TE repeats in plant genomes (35). The huge number copies and non-conserved nucleotide substitution sites nature of TEs make them rough to be systematically and quantitatively studied. We established the approach Matrix-TE based on BLASTN and GPDF algorithms to comprehensively evaluate TE burst events and decline behaviors in plant genomes. Actually we have applied the primary method in cotton genome evolution analysis this year (3), and now it has been improved and optimized thoroughly. For Matrix-TE approach, it has overcome the issues including full length and fragmented piece TEs, and the stochastic nucleotide substitution sites in the coding regions. By using GPDF to fit TE identity distribution curves, it is suitable to apply Matrix-TE in big plant genome analysis.

Six TE super families were identified for Indica and Japonica rice genomes in the super matrix, indicating the six types of TEs have underwent burst events in the rice genome evolution (Fig. 2). However, no obvious TE burst event was detected in Arabidopsis genome by this method, it might be ascribed to the much smaller Arabidopsis genome size (Supplementary Figure S1) (36). As it has been described that the genome size expansion of Malvales plants was highly correlated with TE contents (3). We used the ORFs

of full length TEs to generate the super matrix, and extracted the clusters with high identities (95%) to figure out ref-LTR/TEs. This provided a much efficient strategy to do TE analysis, because any truncated or fragmented TE sequence pieces with diverse SNPs could be scanned out and collected by BLASTN algorithm for the subsequent steps.

The *Ty1/Copia* and *Ty3/Gypsy* distribution curves showed significantly differences between *Indica* and *Japonica* genomes, while the other four types of TEs were almost the same distributed in the two genomes (Fig. 4. D,E and Supplementary Figure S2. A,B,C,D). It is quite interesting that individual TE peaks P-*Copia* and P-*Gypsy* were observed for *Japonica* and *Indica* rice genome separately. And it is strongly suggested that more *Ty1/Copia* TE insertions in *Japonica* genome, while more *Ty3/Gypsy* TE insertions in *Indica* genome (Fig. 4. A,B,C).

Considering the stochastic nucleotide substitutions across TE gene body and low TE transcription activities in rice gnomes, we attempted to resolve the TE distribution curves by Gaussian Probability Density Function (Fig. 5. A,B and Table 1). The intense issue was the normalizing of intact and truncated TE pieces quantitatively, therefore we generated the raw data points of TE identities with 30 bp units, as the optimized setting of short DNA sequences for BLASTN was 30 bp in practice (37). Then we found the symmetrical single peak of TE unit identity distribution curve was fitted by GPDF perfectly, indicating GPDF was an appropriate model for stochastic events (Fig. 5. C,D).

The transcription activity of LTR/TEs was hardly observed in plants, mainly caused by the dilution of RNA-seq mapping intensity by numerous TE repeat copies (2). However, some approaches were developed to rescue the TE transcript profiling (38). We assumed the LTR/TEs with full length ORFs which encoded functional proteins were potentially active (2), the RNA-seq data and TE cDNA clones of rice samples support this hypothesis (Table 1 and Fig. 6). Then the stress induced LTR/TEs transcription were quantified by the RT domain sequences amplification (39). Particularly the high induced expression levels of LTR/TE were observed in hot stress treated rice plants (Fig. 7. A,B). However, the potential biological functions of these induced LTR/TE transcripts were expected in the future.

Conclusions

Generally, we developed the approach Matrix-TE on the basis of BLASTN and GPDF algorithms, and applied it to comprehensively and quantitatively investigate LTR/TE types and contents in the close subspecies *Indica* and *Japonica* rice genomes. The individual TE burst events P-*Copia* and P-*Gypsy* were observed in *Japonica* and *Indica* rice, separately. RNA-seq and RT-PCR methods indicated that LTR/TE transcripts were induced by hot, cold and high salt stress conditions. The optimized Matrix-TE approach and procedures probably could be used in other plant species with big genomes like wheat and maize.

Methods

TE Matrix generation and GPDF analysis

Indica and Japonica genomes.

The two cultivated Oryza Sativa Indica (93-11) and Japonica (Nip) genomes were applied for the analysis pipeline (Figure 1). The genome assembly ID were GCA_003865215.1 for Indica (93-11) and GCA_003865235.1 for Japonica (Nip), separately. As the two updated genomes were assembled by PacBio with long reads sequencing technologies (29). Then the full length LTR/TEs of 93-11 and Nip genomes were annotated by LTR_finder software with default parameters (3).

TE ORFs matrix generation.

The ORFs were extracted from the previous full length LTR/TEs by Getorf Script in EMBOSS package. All the ORFs were sorted by length from longest to shortest, and the ORFs with length less than 1500 bp were discarded. Then the sequence identity matrix was calculated by BioEdit software with default parameters (2,3).

ORF clustering and phylogenetic analysis.

The LTR/TE ORFs in the previous matrix with sequence identity over 95% were extracted and grouped by their sequence length and homology. The molecular phylogenetic trees of the ORF groups were generated by MEGA software with low homologous sequences as the root. The ORF sequences on the top branch of the phylogenetic trees were determined as the reference TE sequences (ref-TE-seq) in the following steps (3).

Genome scanning and GPDF analysis.

The Indica (93-11) and Japonica (Nip) genomes were scanned by the reference TE sequences with BLASTN software and the following parameters:

$ blastn -db Ricedb -query ref-TE-seq.fa -out ref-TE-seq-blast-Ricedb -evalue 0.00001 -word_size 11 -gapopen 5 -gapextend 2 -penalty -2 -reward 1 -culling_limit 0 -outfmt 7

The blast result sequences containing complete and truncated TE sequences with identity relative to ref-TE-seq, were fragmented into 30 bp units and the sequence identity of each unit was a data point in the following step (3). All the data points for the individual TE super family were used to generate an identity distribution curve. Then the peaks of the curve were fitted by Gaussian Probability Density Function and the average nucleotide substitution ratio Ks of each TE peak were defined as 2.58 standard deviations (σ).

TE burst events and genome evolution.

Single nucleotide polymorphism distribution across Ty1/Copia and Ty3/Gypsy gene body were calculated for 93-11 and Nip genomes, and the SNP densities were labeled along with TE gene bodies. The individual TE burst peaks were compared between Indica and Japonica rice genomes, and the different TE burst events between them might be correlated with rice subspecies differentiation (29,30,40).

Rice stress conditions and TE transcription activity

Rice plants growth and stress conditions.

The 93-11 and Nip rice were planted and stress treated as the conditions described previously (41,42). Generally, rice seedlings were cultivated on culture medium in incubator for 14 days, with cycles of 14 h of light and 10 h of dark at 28 ℃. Then the rice plants were grouped and group I were used as control (mild conditions), group II were treated at 42 ℃ for 12 h, group III were treated at 5 ℃ for 5 days, group IV were treated under 200 mM NaCl for 5 days. And the leafs of four plant groups were collected for RNA extraction (42).

Rice RNA sequencing and mapping.

The rice RNA was extracted by using FastPure Plant Total RNA Isolation kit (Vazyme Biotech) with the instructions (2). The RNA quality was examined by 1% agarose gel, and the cDNA was prepared by HiScript III 1st Strand cDNA Synthesis kit (Vazyme Biotech). The total RNA samples were applied to RNA-seq by Illumina HiSeq 3000 platforms. The RNA-seq clean data were mapped to the full length ORF sequences of Indica and Japonica TEs separately with Hisat2 software, and the transcribed TEs were screened by HOMER packages (2).

Rice transcribed TE cloning and RT-PCR quantification.

The full length TEs supported by RNA-seq data were cloned by using Phanta Max Super-Fidelity DNA Polymerase (Vazme Biotech), and validated by 1% agarose gel analysis (Figure 6). The TE transcription activities of rice under mild conditions, hot, cold and salt stresses were quantified by RT-PCR with UBQ5 as reference gene (Figure 7) (43). The primers used for cloning and RT-PCR were presented in the Supplementary Table S4 and Table S5, separately.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Not applicable

Competing interests

All authors declare no competing interests.

Funding

This work was supported by the Natural Science Foundation of China (No. 21602162, No. 31690091), National Science and Technology Major Project (No. 2016ZX08005003-001).

Authors' contributions

Z. Wu conceived the project idea, and designed the experimental work. Z. Wu, Y. Yang and G. Huang conceived the bioinformatics experiments and carried out all data analysis. Y. Wu, X. Wang, Y. Zheng, J. Zou, Y. Guan and Y. Li carried out the experimental work. Z. Wu wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Prof. Jie Zhao, Prof. Jun Hu and Prof. Xin Hou at College of Life Sciences, Wuhan University for providing the plant seeds of *Indica* and *Japonica* rice. We thank the Jie Zhao lab members for insightful discussion and technical support. And we used the supercomputing server in the Institute for Advanced Studies, Wuhan University for the genome data manipulation.

References

1. Lisch D. How important are transposons for plant evolution? Nature Rev Genet. 2013;14: 49-61.
2. Lin J, Cai Y, Huang G, Yang Y, Li Y, Wang K, et al. Analysis of the chromatin binding affinity of retrotransposases reveals novel roles in diploid and tetraploid cotton. J Integr Plant Biol. 2019;61:32-44.

3. Huang G, Wu Z, Percy RG, Bai MZ, Li Y, Frelichowski JE, et al. Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboretum and Gossypium hirsutum provide insights into cotton A-genome evolution. Nat Genet. 2020;52:516-24.
4. Möller M, Stukenbrock EH. Evolution and genome architecture in fungal plant pathogens. Nature Rev Micro. 2017;15:756-71.
5. Goff SA, Ricke D, Lan T, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (Oryza sativa L, ssp, japonica). Science. 2002;296:92-100.
6. Yu J, Hu S, Wang J, Wong G, Li S, Liu B, et al. A draft sequence of the rice genome (Oryza sativa L, ssp, indica). Science. 2002;296:79-92.
7. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in 3010 diverse accessions of Asian cultivated rice. Nature. 2018;557:43-9.
8. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation turnover and innovation across the genus Oryza. Nat Genet. 2018;50:285-296.
9. Huang X, Kurata N, Wei X, Wang Z, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012;490:497-501.
10. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, et al. Sequencing and de novo assembly of a near complete indica rice genome. Nat Commun. 2017;doi:10,1038/ncomms15324.
11. Sun J, Ma D, Tang L, Zhao M, Zhang G, Wang W, et al. Population Genomic Analysis and De Novo Assembly Reveal the Origin of Weedy Rice as an Evolutionary Game. Mol Plant. 2019;12:632–47.
12. Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, et al. Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature. 2017;551:498–
13. Ling HQ, Ma B, Shi X, Liu H, Dong L, Sun H, et al. Genome sequence of the progenitor of wheat A subgenome Triticum urartu. Nature. 2018;557:424-28.
14. Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet. 2018;50:1282-88.
15. Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, et al. European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet. 2020;doi:10,1038/s41588-020-0671-9.
16. Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. 2018;50:1289-95.
17. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature. 2004;430:471–76.
18. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. Nat Rev Genet. 2018;19:688–704.
19. Tang Y, Ma X, Zhao S, Xue W, Zheng X, Sun H, et al. Identification of an active miniature inverted-repeat transposable element mJing in rice. Plant J. 2019;98:639–53.

20. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet. 2018;50:278-84.
21. Deininger P, Morales ME, White TB, Baddoo M, Hedges DJ, Servant G, et al. A comprehensive approach to expression of L1 loci. Nucleic Acids Res. 2017;45:e31.
22. ElBaidouri M, Kim KD, Abernathy B, Arikit S, Maumus F, Panaud O, et al. A new approach for annotation of transposable elements using small RNA mapping. Nucleic Acids Res. 2015;43:e84.
23. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. Nature. 2003;421:163–7.
24. Chen J, Lu L, Benjamin J, Diaz S, Hancock CN, Stajich JE, et al. Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. Nat Commun. 2019;10:doi:10,1038/s41467-019-08451-3.
25. Carpentier MC, Manfroi E, Wei FJ, Wu HP, Lasserre E, Llauro C, et al. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nat Commun. 2019;10:doi:10,1038/s41467-018-07974-5.
26. Meyers BC, Tingey SV, Morgante M. Abundance distribution and transcriptional activity of repetitive elements in the maize genome. Genome Res. 2001;11:1660–76.
27. Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nature Rev Genet. 2017;18:292–308.
28. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20:43–5.
29. Zhang Q, Liang Z, Cui X, Ji C, Li Y, Zhang P, et al. N6-Methyladenine DNA Methylation in Japonica and Indica Rice Genomes and Its Association with Gene Expression Plant Development and Stress Responses. Mol plant. 2018;11:1492–508.
30. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA. 2004;101:12404–10.
31. Biémont C, Vieira C. Genetics: junk DNA as an evolutionary force. Nature. 2006;443:521–4.
32. Grahn RA, Rinehart TA, Cantrell MA, Wichman HA. Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. Cytogenet Genome Res. 2005;110:407–15.
33. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020;doi:10,1038/s41576-020-0251-y.
34. Bevan MW, Uauy C, Wulff BB, Zhou J, Krasileva K, Clark MD. Genomicinnovation for crop improvement. Nature. 2017;543:346–54.
35. Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii a progenitor of bread wheat with the MaSuRCA mega-reads algorithm. Genome Res. 2017;27:787–92.
36. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.

37. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36:doi:10,1093/nar/gkn201.
38. Picault N, Chaparro C, Piegu B, Stenger W, Formey D, Llauro C, et al. Identification of an active LTR retrotransposon in rice. Plant J. 2009;58:754–65.
39. Xiong Y, Ckbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 1990;9:3353–62.
40. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nature Rev Genet. 2017;18:71–86.
41. Singh AK, Kumar R, Tripathi AK, Gupta BK, Pareek A, Singla-Pareek SL. Genome-wide investigation and expression analysis of Sodium/Calcium exchanger gene family in rice and Arabidopsis. Rice. 2015;8:doi:10,1186/s12284-015-0054-5.
42. Chen K, Guo T, Li XM, Zhang YM, Yang YB, Ye WW, et al. Translational Regulation of Plant Response to High Temperature by a Dual-Function tRNAHis Guanylyltransferase in Rice. Mol plant. 2019;12:1123–42.
43. Jain M, Nijhawan A, Tyagi AK, Khurana JP. Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. Biochem Biophys Res Commun. 2006;345:646–51.

Figures

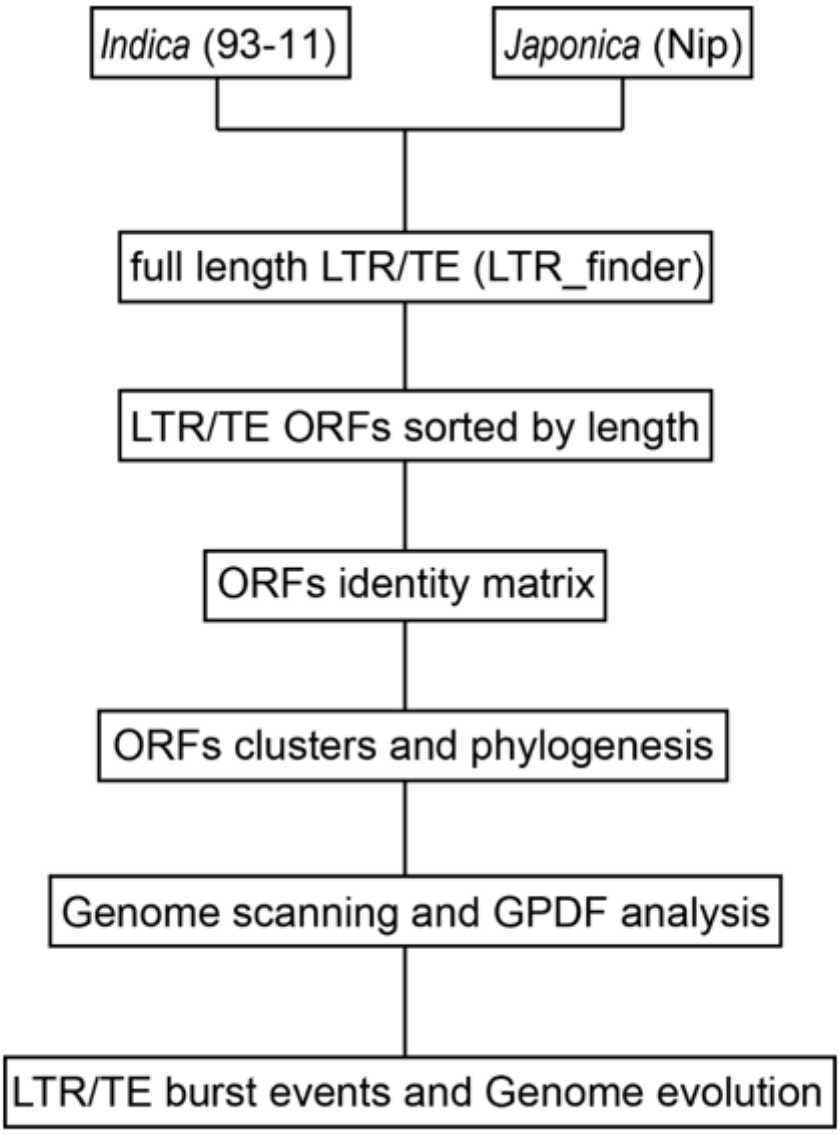
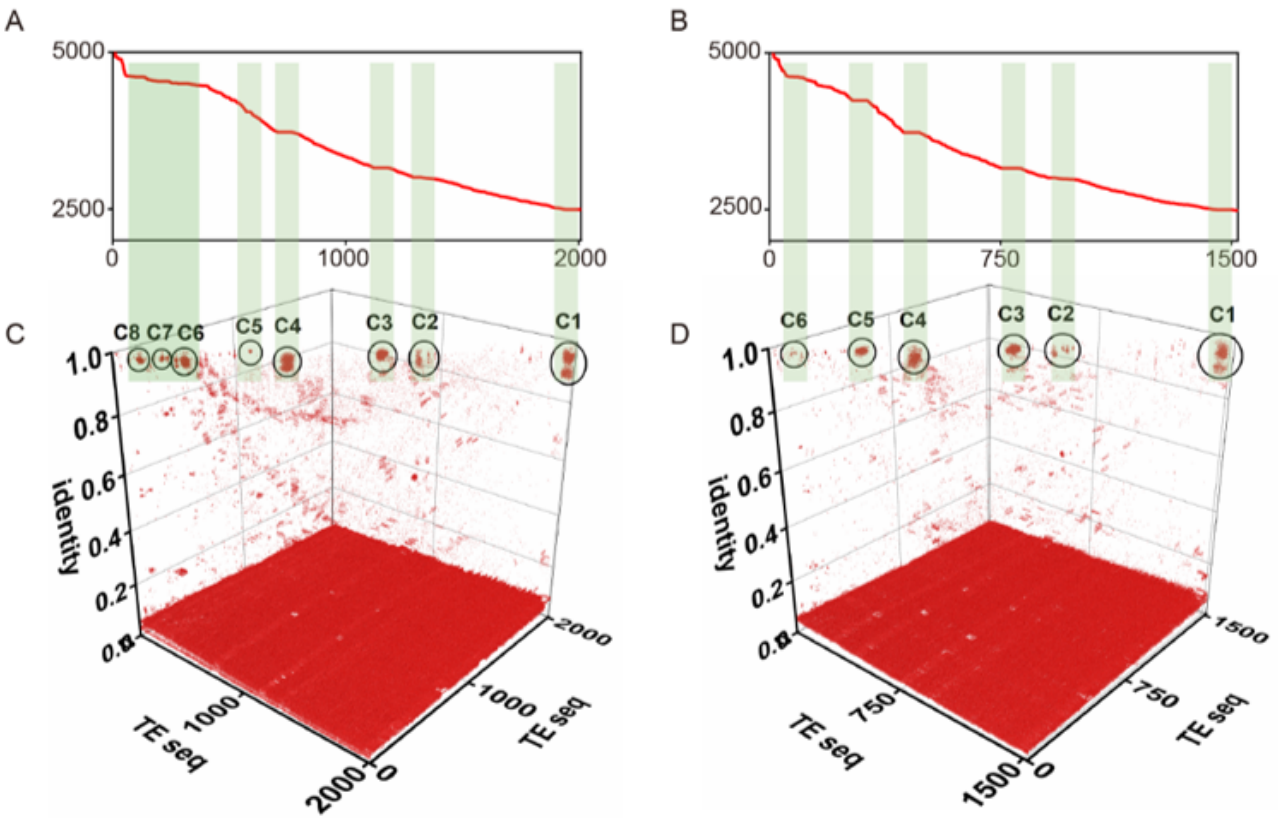


Figure 1

the Matrix-TE approach pipeline. The genomes of two close subspecies Indica (93-11) and Japonica (Nip) were used as input data. Then the scripts in LTR_finder, EMBOSSY, BioEdit, MEGA and BLASTN packages were sequentially applied to generate sets of raw data points for GPDF analysis with optimized parameters.



E

Species	*Oryza Sativa* *Indica* (93-11)	*Oryza Sativa* *Japonica* (Nip)
Genome size (Mb)	395	380
Gene size (%)	29.5	29.4
TE size (%)	46.2	44.1
Full length LTR/TE	5,146	4,744
ORFs over 1500 bp	4,870	4,201
ORFs in matrix	2,010	1,520

F

	LTR/TE type	
Species	*Indica* (93-11)	*Japonica* (Nip)
Cluster 1 (C1)	*OS-type1*	*OS-type1*
Cluster 2 (C2)	*OS-type2*	*OS-type2*
Cluster 3 (C3)	*OS-typeRT*	*OS-typeRT*
Cluster 4 (C4)	*OS-typePHA*	*OS-typePHA*
Cluster 5 (C5)	*Ty1/Copia*	*Ty1/Copia*
Cluster 6 (C6)	*Ty3/Gypsy*	*Ty3/Gypsy*
Cluster 7 (C7)	*Ty3/Gypsy*	
Cluster 8 (C8)	*Ty3/Gypsy*	

Figure 2

93-11 and Nip genome statistics and TE matrix and clusters. A,B, sorted length of ORFs from longest to shortest; C,D, TE matrixes in three dimensional format, and TE clusters with identity over 95% were circle labeled in 93-11 and Nip genomes, separately. E, genes, TEs and TE ORFs parameters of the two genomes. F, types and annotations of TE clusters for the two genomes, 8 clusters and 6 types in 93-11 and 6 clusters and 6 types in Nip genomes.

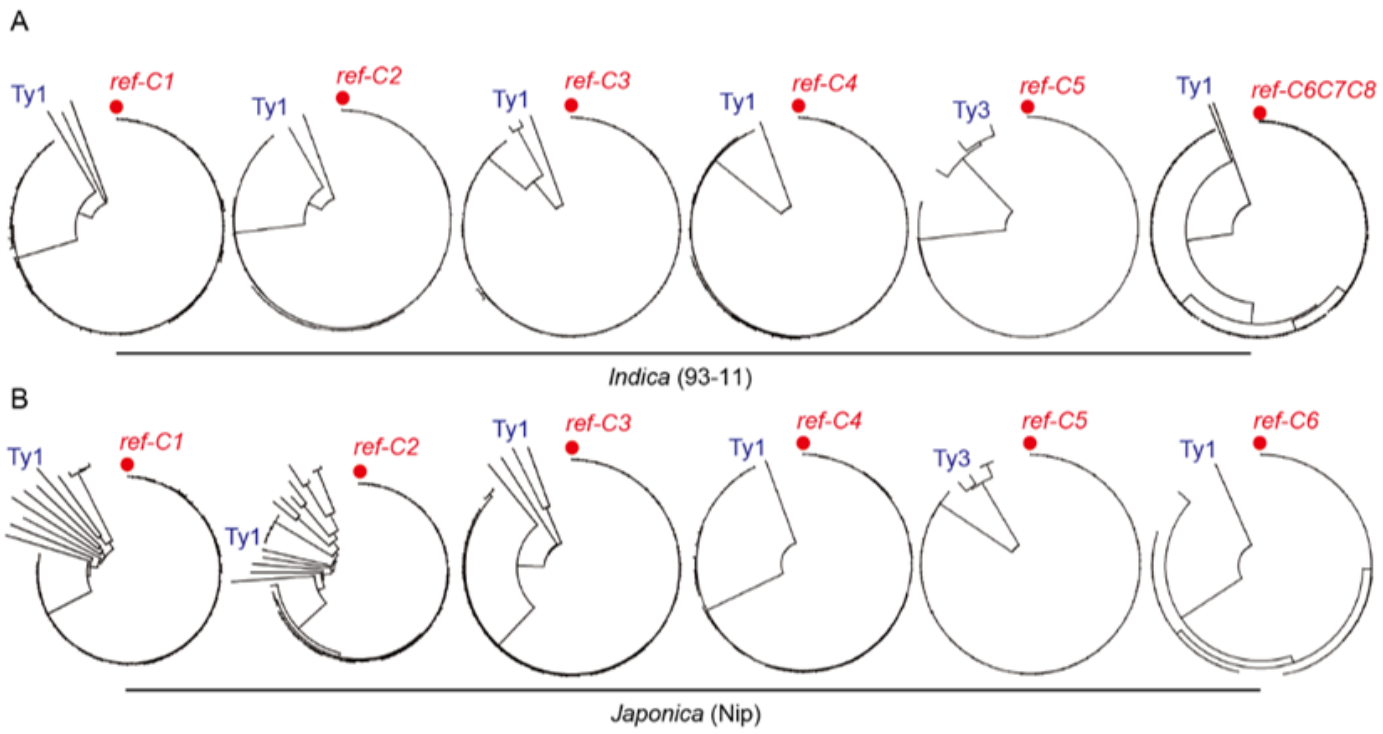


Figure 3

Phylogenetic analysis of 93-11 and Nip TE clusters and TE reference sequences. A, phylogenetic trees of TE clusters in 93-11 genomes. B, phylogenetic trees of TE clusters in Nip genomes. The sequences on root of the trees were labeled by blue, and the reference TE sequences on top branches were labeled by red.

A

Species	Genome scanning by *ref-LTR/TE*	
	Indica (93-11, bp)	*Japonica* (Nip, bp)
ref-C1 (93-11)	1,089,090	1,267,500
ref-C2 (93-11)	1,049,370	976,710
ref-C3 (93-11)	1,123,020	1,052,760
ref-C4 (93-11)	1,825,410	2,041,620
ref-C5 (93-11)	87,420	215,970
ref-C6C7C8 (93-11)	7,086,180	3,176,190
ref-C1 (Nip)	1,088,280	1,262,910
ref-C2 (Nip)	1,028,330	964,500
ref-C3 (Nip)	1,113,300	1,053,210
ref-C4 (Nip)	1,848,360	2,062,140
ref-C5 (Nip)	1,180,770	1,519,560
ref-C6 (Nip)	6,111,660	3,217,560

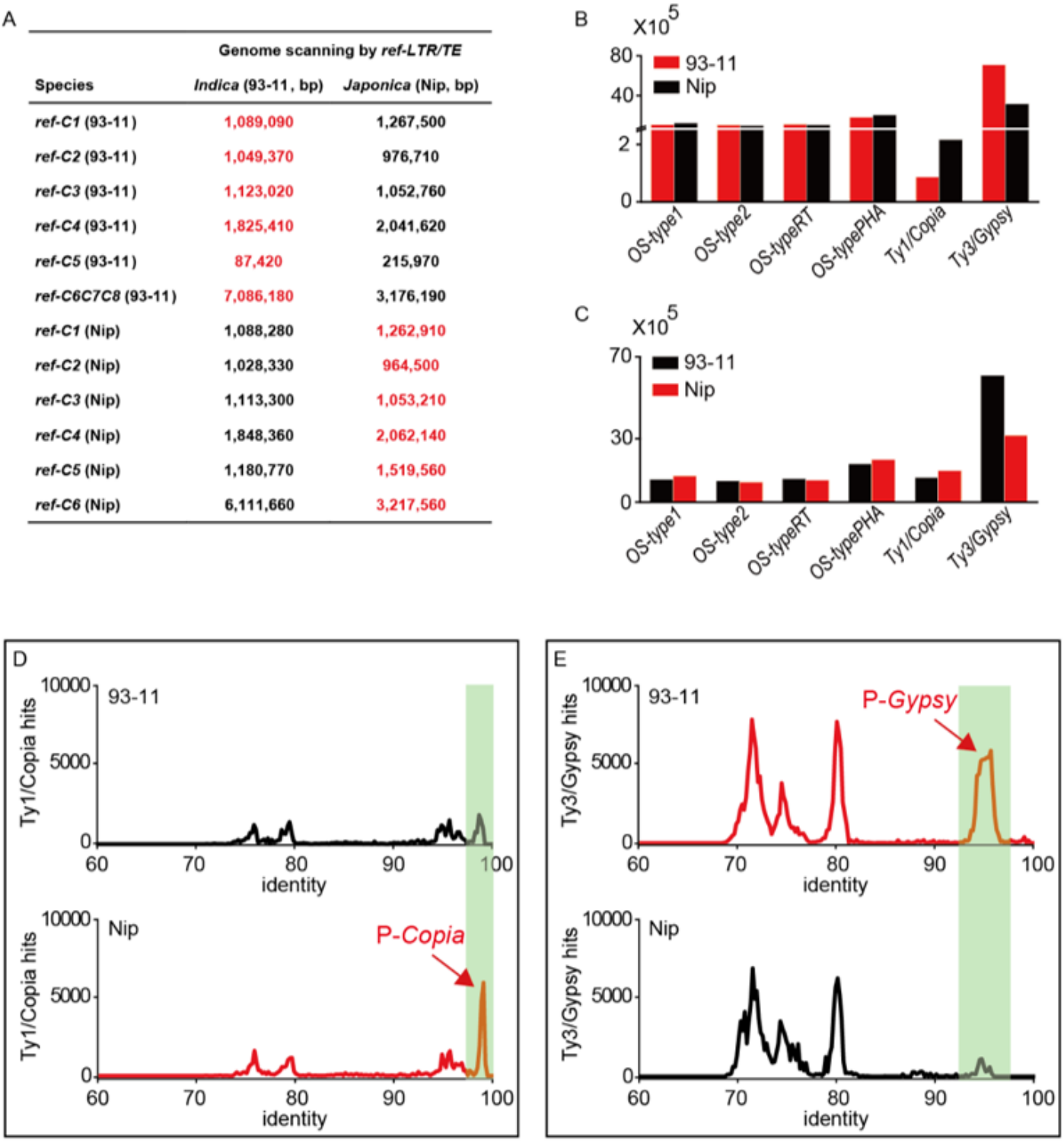


Figure 4

93-11 and Nip whole genome BLASTN scanning with reference TE sequences. A,B,C total sequence length and histogram statistics for the blast results in the two genomes. D,E, comparisons of normalized TE identity distribution curves between 93-11 and Nip genomes. Individual P-Copia and P-Gypsy peaks were observed in Nip and 93-11 genomes separately, and were labeled by red.

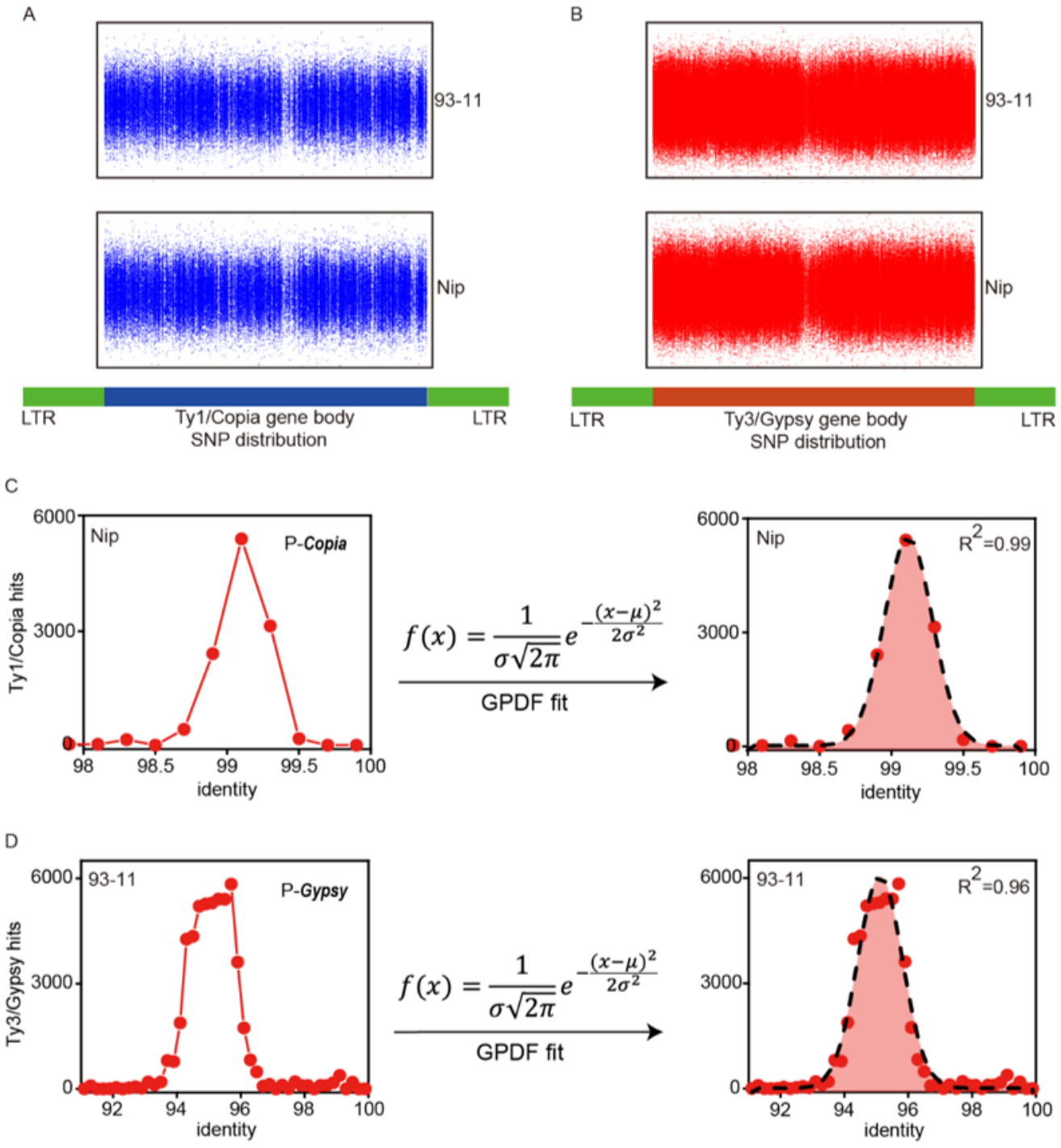


Figure 5

SNP distributions across Ty1/Copia and Ty3/Gypsy gene bodies, and GPDF fit of P-Copia and P-Gypsy peaks. A, blue SNP distribution points of Ty1/Copia in 93-11 and Nip genomes. B, red SNP distribution points of Ty3/Gypsy in 93-11 and Nip genomes. C, P-Copia peak was fitted by GPDF with R square value 0.99. D, P-Gypsy peak was fitted by GPDF with R square value 0.96. And the average nucleotide substitution ratios were defined as 2.58σ for the two peaks.

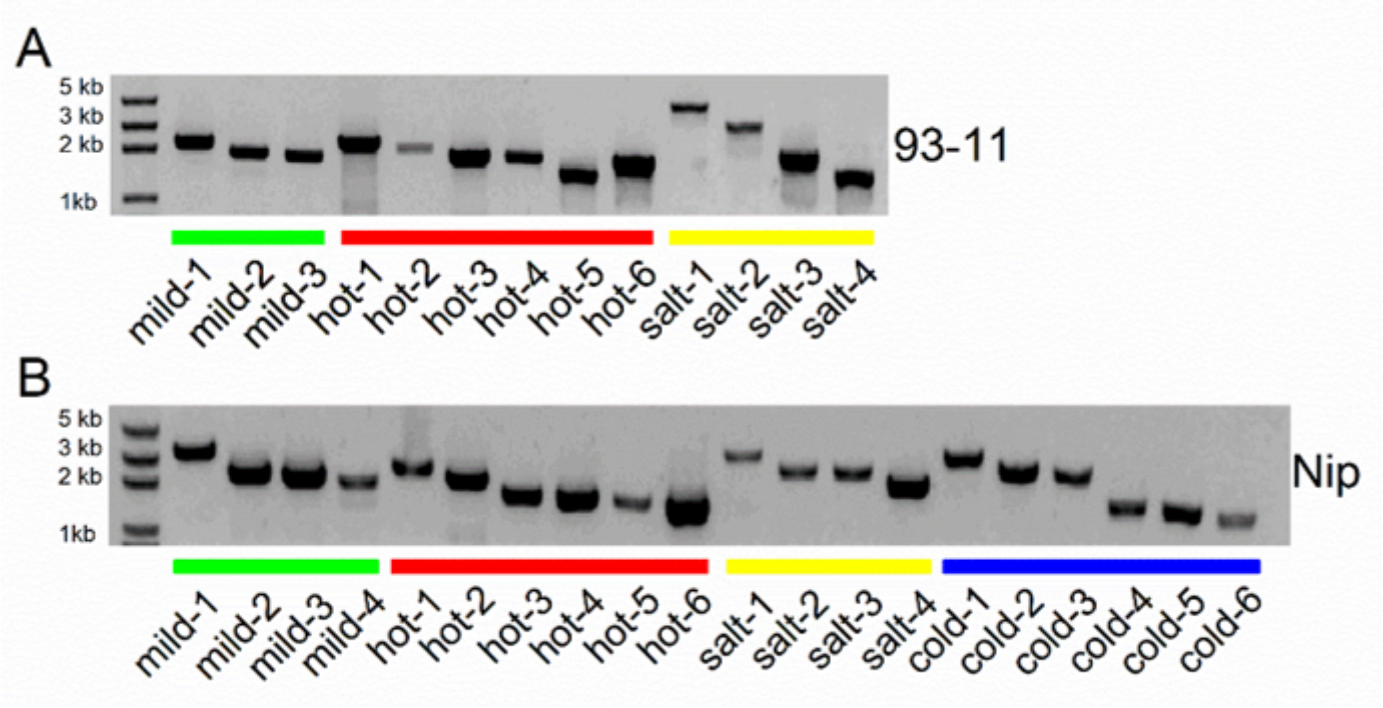


Figure 6

gel results of full length c-DNA clones from stress induced TE expression in 93-11 and Nip rice. A, 3 clones under mild conditions (green), 6 clones under hot stress (red) and 4 clones under salt stress (yellow) treated 93-11 rice plants. B, 4 clones under mild conditions (green), 6 clones under hot stress (red), 4 clones under salt stress (yellow) and 6 clones under cold stress (blue) treated Nip rice plants.

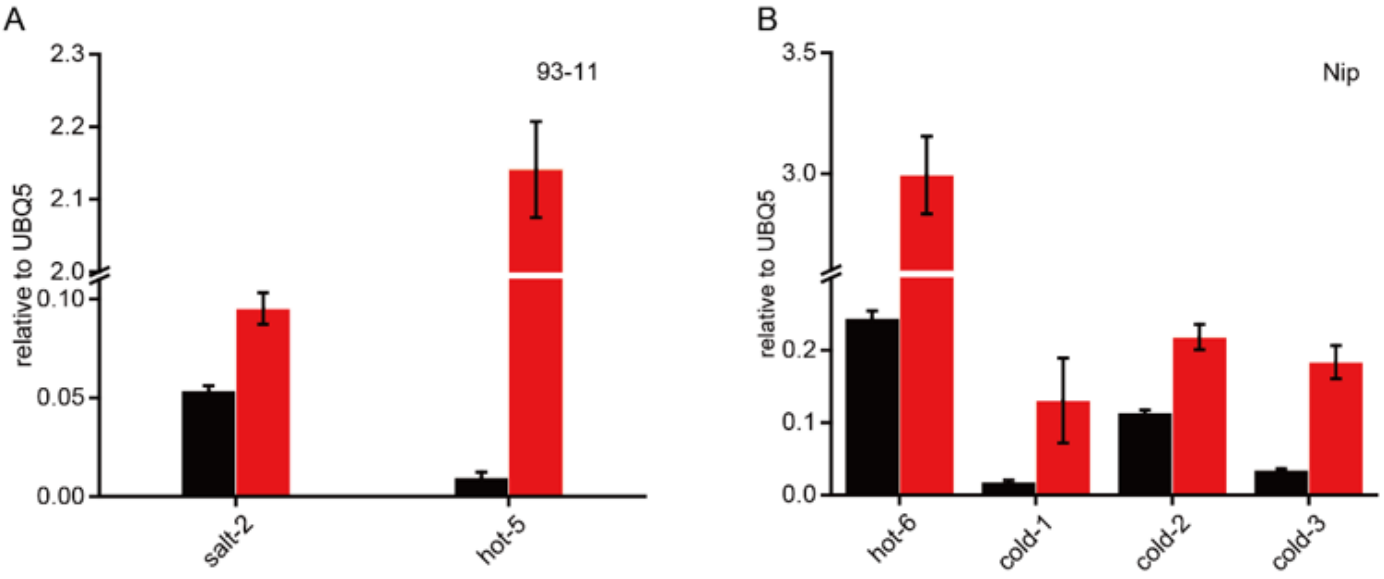


Figure 7

stress induced expression of LTR/TEs. A, salt-2 and hot-5 TE expressions were induced by NaCl and 42 °C stress treatments in 93-11 rice plants. B, hot-6, cold-1, cold-2 and cold-3 TE expressions were induced

by 42 ℃ and 5 ℃ treatments in Nip rice plants. The TE expression levels of mild conditions were labeled by black histograms, and the induced TE expression levels were labeled by red histograms.

Supplementary Files

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

- Additionalfiles.docx
- PBIOSI.docx