Journal of Animal & Plant Sciences, 32(2): 2022, Page: 421-429
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2022.2.043

PRELIMINARY ANALYSIS OF NUCLEAR TRANSCRIPTOME REVEALED THE EXTENSIVE NUCLEAR-ENRICHED SUBCELLULAR DISTRIBUTION PATTERN OF TRANSPOSABLE ELEMENT-DERIVED TRANSCRIPTS IN RICE

L. Zeng[1], X. Tan[2], S. Hu[2,3,*] and Y. Luo[2,3,*]

[1] Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing, China.; [2] State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China; [3] University of Chinese Academy of Sciences, Beijing, China.

* Corresponding author's email: husn@im.ac.cn (Songnian Hu), sanmaolyf@sina.com (Yingfeng Luo)

ABSTRACT

The complexity and its functional consequence of nuclear transcriptome have come to be documented in mammal cells. However, little is known about plant nuclear transcriptome. In this study, we explored the complexity of rice nuclear transcriptome via the profiling of whole protoplasts and isolated nucleus using rRNA-depleted RNA-seq strategy. The expression characteristics of the long non-protein coding RNAs, transposable element-related genes and conventional protein coding genes were investigated. These analyses found that the proportion of expressed TE-related genes and reads mapped to these genes in nucleus were four time of that in protoplasts. In particular, no obvious nuclear-enriched expression pattern of lncRNAs or conventional protein coding genes was observed. Therefore, the extensive transcription of TE-related genes was the major contribution of the larger diversity of nuclear RNA species. Further muti-tissues comparative transcriptome analysis indicated that the transcription of TE-related genes, especially for nuclear-specific expressed TE-related genes defined in this study, was likely to be underestimated by conventional bulk RNA-seq. Furthermore, Gene Ontology functional enrichment analysis indicated that some nuclear abundant lncRNAs were co-expressed with nuclear-located RNA binding proteins. Together, the global features of rice nuclear transcriptome will prompt the functional characterization of the nuclear-expressed TE-related genes and lncRNAs.

Key words: Nuclear transcriptome, Rice, RNA-seq, long non-protein coding RNAs, transposable element-related genes

Published first online November 25, 2021 Published final March 15, 2022.

INTRODUCTION

The transcriptional landscape of nucleus extended the transcriptome complexity of mammal cells. The nucleus expressed a higher proportion of novel transcripts stemmed from intergenic regions (Grindberg, *et al*. 2013; Mondal, *et al*. 2010). Besides of novel transcripts, quite a lot of long non-protein coding RNAs (lncRNAs, with length ≥ 200 nt) enriched in nucleus and involved in important nuclear activities (Bergmann, *et al*. 2015; Lee, *et al*. 2014). Most of nuclear-enriched lncRNAs were evolutionary conserved both in sequence level and expression pattern, suggesting their important roles in nuclear biological processes (Derrien, *et al*. 2012). In addition to lncRNAs, comparative analysis of nuclear and cytoplasmic transcriptomes of mammal cell lines found that the complexity of nuclear retrotransposon-derived transcripts was substantial larger than that in cytoplasms. Subsequent functional study documented the nuclear transcription of retrotransposons played a regulatory role in pluripotency maintenance for mammalian stem cells (Fort, *et al*. 2014). In brief, subcellular transcriptome profiling has been established as an important perspective for functional characterization of lncRNAs and TE-related genes in mammal cells (Derrien, *et al*. 2012; Fort, *et al*. 2014; Mondal, *et al*. 2010). Although plant genomes are rich of lncRNAs and transposable elements (TEs), little is known about plant nuclear transcriptomes at present.

Rice is the stable food for one half of the world's population and a model organism for plant research. Thousands of lncRNAs were identified in rice and other plants (Li, *et al*. 2014; Liu, *et al*. 2012; Zhang, *et al*. 2010; Zhang, *et al*. 2014), but few of them have been functional studied (Liu, *et al*. 2015; Zhang, *et al*. 2014). Additionally, TEs are important genomic components and accounts for > 35% of rice genome (International Rice Genome Sequencing 2005). Rice genomic annotation revealed that about a quarter of open reading frames (ORFs) were highly similar to rice repeat sequences (Ouyang and Buell 2004) or contained TE-related protein domains (Yuan, *et al*. 2005). Therefore, these ORFs were defined as transposable element-related genes (TE-related genes) (Jiao and Deng 2007). Multiple studies revealed that TE-related genes were conditionally transcribed and their dynamic transcriptions were associated with stress responses in rice and other plants (Jiao and Deng 2007; Liang, *et al*. 2021; Vicient 2010;

Wang, *et al*. 2017). From this point, rice is an ideal plant model to study lncRNAs and TE-related genes from the perspective of subcellular transcriptome profiling. For now, the subcellular location information has been revealed in rice transcriptome studies (Reynoso, *et al*. 2018; Yuan, *et al*. 2018). However, the comprehensive analysis of rice nuclear transcriptome, especially from the perspective of TE-related genes, is still lacking.

Here, the rRNA-depleted RNA-seq technology was used to study the transcriptome of rice nucleus isolated from the protoplasts of rice seedlings. The analyses were focused on the subcellular expression characteristics of TE-related genes, lncRNAs and conventional protein-coding genes from universally established rice gene models in the form of comparison between nucleus and protoplasts transcriptomes. This study provided an overall understanding of rice nuclear transcriptome.

MATERIALS AND METHODS

Protoplasts preparation: Rice (Oryza sativa L.) cultivar Nipponbare was incubated on MS medium at 28 °C with a photoperiod of 16 hours light (150 µmol m-2 s-1)/ 8 hours dark. The 14 days seedlings were used for protoplasts isolation. Green rice sheath was cut into 0.5 mm stripes with sharp razors. The stripes were incubated in the enzyme solution (0.6 M mannitol, 10 mM MES, 1.5% cellulase RS, 0.75% macerozyme, 0.1% BSA, 3.4 mM CaCl2 dihydrate, 5 mM β-mercaptoethanol and 50µg/ml carbenicillin) and applied vacuum infiltration for 30 minutes (15 Hg) in the dark. After 5 hours digestion with shaking (60-80 rpm), an equal volume of W5 (154 mM NaCl, 125 mM CaCl2, 5 mM KCl, 2 mM MES) solution was added and gently mixed, then the protoplasts solution was filtered through a 35 µm nylon meshes screen. The protoplasts pellets were collected by centrifugation at 150 g for 5 minutes. The protoplasts were washed two times with W5 solution (Zhang, *et al*. 2011b). All solutions were prepared with DEPC water.

Nucleus isolation: The freshly isolated protoplasts were suspended in nuclei isolation buffer (NIB, 0.25 M sucrose, 15 mM PH 6.8 PIPES , 5 mM MgCl2, 60 mM KCl, 15mM NaCl, 1 mM CaCl2, 0.9% Triton X-100, 1 mM PMSF, 2 ug/ml pepstatin A, 2 ug/ml aprotinin) and were smashed on ice by constantly shaking for 15 minutes. The pellets were collected at 4°C by centrifugation at 150 g for 5 minutes. The nuclei in pellets were suspended in NIB buffer, and were layered on 10 ml 2 M sucrose by centrifugation at 6000 g for 15 minutes at 4°C. Nucleus were collected and washed with NIB buffer for three times (Tan, *et al*. 2007). The purity and integrity of nucleus were checked using fluorescence microscopy (Leica TCS SP2) after DAPI (4',6-Diamidine-2'-phenylindole dihydrochloride) staining.

RNA-seq library construction: Total RNA of nucleus and whole protoplasts were separately extracted by TRIzol reagent (Tiangen), and treated with RNase free DNase (Takara) to digest genomic DNA. Ribominus kit (NEB) was used to remove rRNAs according to manu's protocol. The strand-specific libraries for RNA-seq were constructed using NEXTflexTM Rapid Illumina Directional RNA-Seq Library Prep Kit (Bioo). The RNA-seq libraries were constructed and sequenced using HiSeq2000 (Illumina) with 2x100 bp module.

qRT-PCR verification: For quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) analysis, cDNAs was prepared from 2µg of total RNA isolated from nucleus and protoplasts by Quantscript RT Kit (Tiangen) with random primers, respectively. qRT-PCR was conducted by a CFX96 Real-Time PCR detection system (Bio-Rad). The primers (Table S1) of U1 (AP007257) and U6 (AK059727) were designed online (https://primer3.ut.ee/). Three technical replicates were used for each sample.

Gene models integration and RNA-seq data analysis: To better annotate RNA-seq datasets, Cuffmerge module embedded in software Cufflinks (Ghosh and Chan 2016) was used to integrate two well-established rice gene models from the Rice Genome Annotation Project (RGAP http://rice.plantbiology.msu.edu) and the Rice Annotation Project (RAP, https://rapdb.dna.affrc.go.jp/).

The raw RNA-seq reads were processed to remove adapter sequences and terminal low quality bases (Zhang, *et al*. 2011). These >30 nt qualified reads were mapped to rice chloroplast genome, mitochondria genome and rRNA genes by bowtie2 (Langmead and Salzberg 2012) and mapped reads were removed. Lastly, the clean reads were mapped to the rice genome guided by integrated gene models using GSNAP (Wu, *et al*. 2016). The genomic distribution of mapped reads in exonic, intronic and intergenic regions were calculated by custom Perl scripts. Cufflinks (Ghosh and Chan 2016) was used to compute the gene expression level. And the gene with Reads Per Kilobase per Million mapped reads (RPKM) value larger than 0 was considered as expressed gene. All genes in integrated gene models were divided into four groups based on subcellular expression characteristics in this study, (a) nucleus-specific expressed gene group, includes the genes only detected in nuclear transcriptome, (b) protoplasts-specific expressed gene group, includes the genes only detected in protoplasts transcriptome, (c) nucleus/protoplasts expressed gene group, includes the genes detected in nuclear and protoplasts transcriptomes, (d) undetected gene group, includes the genes not detected in nuclear or protoplasts transcriptomes from this study.

TE-related genes identification and analysis: The repeat types of TE-related gene were based on the protein

domain or the annotation of the MSU Oryza Repeat Database (http://rice.plantbiology.msu.edu/). The value of "Gene expression bias from nucleus to protoplasts" for each nucleus/protoplasts expressed gene was calculated by the formula "(RPKM nucleus-RPKM protoplasts)/(RPKM nucleus+RPKM protoplasts)".

LncRNAs identification: To identify lncRNAs in rice genome, the nucleotide sequences of each transcript were downloaded from corresponding RGAP and RAP websites and computed for protein coding potential with Coding Potential Calculator (CPC, http://cpc.gao-lab.org/programs/run_cpc.jsp) (Kong, *et al*. 2007). The transcripts with CPC value smaller than 0 were considered as putative lncRNAs.

Gene expression breadth and gene co-expression analysis: Raw RNA-seq data from eleven rice tissues (shoot, root, leave, anther, inflorescence, embryo, endosperm, pistil, seed, callus and seedling) were downloaded from RGAP website (http://rice.plantbiology.msu.edu/expression.shtml) and published datasets (Reynoso, *et al*. 2018; Yuan, *et al*. 2018; Zhang, *et al*. 2014). Reads pre-processing, mapping and gene expression level calculation were using the same pipeline as the datasets produced in this study. The Pearson Correlation Coefficient for each combination of lncRNAs and protein-coding genes was computed using custom Perl scripts. Protein coding genes with Pearson Correlation Coefficient larger than 0.8 or smaller than -0.8 for twenty nuclear abundant lncRNAs were selected for Gene Ontology enrichment analysis using agriGO (http://bioinfo.cau.edu.cn/agriGO/) (Tian, *et al*. 2017).

RESULTS

Rice nucleus isolation, RNA-seq and quality confirmation: The protoplasts and nucleus were isolated from the seedlings of 14-day-old rice (*Oryza sativa L.*) cultivar Nipponbare (Fig 1A). After confirming the purity of nucleus using fluorescence microscopy (Fig 1B), the nucleus and protoplasts were immediately used for total RNA isolation. The relative concentration of two small nuclear RNAs (U1 and U6) in nucleus and protoplasts were measured by qRT-PCR using comparable amount of total RNA. The results of two biological replicates indicated that the concentration of U1 in nucleus was 8 times higher than protoplasts, while U6 in the nucleus was 2 times higher than protoplasts (Fig 1C). Lastly, the rRNA depleted RNA of nucleus and protoplasts were converted to double-stranded cDNAs for RNA-seq, respectively.

In order to facilitate the comparison of rice nucleus and protoplasts transcriptomes, the gene models from two well-established rice gene annotation projects (RGAP and RAP) were combined by removing redundancy. The integrated gene models consisted of 68,082 genes, including 15,848 TE-related genes, 7,124 putative lncRNAs and 45,110 conventional protein coding genes (Table 1). Then, the high-quality clean reads were aligned to the latest rice genome (TIGR 7.0) guided by the integrated gene models. Initial analysis indicated the biological replicates of nucleus and protoplasts exhibited high consistence in gene expression level (Fig S1) and genomic location distribution of mapped reads (Fig S2). Therefore, the RNA-seq data from two biological replicates were combined for following analysis, resulting 17.7 and 24.3 million high-quality reads for nucleus and protoplasts transcriptomes, respectively. Averagely, reads mapped to intronic regions accounted for 14.2% of total mapped reads in nucleus (exonic region 59.1%, intergenic region 26.7%), which was larger than that in protoplasts (intronic region 11.3%, exonic region 82.2% and intergenic region 6.5%, Fig 1D). The genomic location analysis result was consistent with two relevant publications (Reynoso, *et al*. 2018; Yuan, *et al*. 2018) (Table S2 and Fig S3) that nucleus harbored a larger proportion of primary transcripts from intronic and intergenic regions. Importantly, these results together illustrated the good performances of the nuclei isolation and RNA-seq experimental workflow used in this study (Fig 1A), and thereby providing a solid basis for the subsequent feature analysis of rice nuclear transcriptome.

The general features of rice nuclear transcriptome: Totally, the number of expressed genes detected in nucleus was 27,900, while that in protoplasts was 24,228. The number of genes simultaneously detected in nucleus and protoplasts was 16,398. Therefore, the number of nucleus-specific expressed genes (11,502) was larger than protoplasts-specific expressed genes (7,830) (Table 1, Fig 2A and Table S3). Additional gene expression level distribution analysis indicated that the nucleus expressed a smaller number of highly expressed genes and a larger amount of moderately expressed genes (Fig 2B and Table S4). For instance, the number of genes with RPKM $\geq$ 100 detected in nucleus was 374 (harboring 28.9% of total mapped reads in nucleus), while that in protoplasts was 562 (harboring 33.1% of total mapped reads in protoplasts). Therefore, in nucleus, the deficiency of highly expressed transcripts facilitated the detection of more transcripts, especially for the moderately expressed genes (Fig 2C, Table S4 and S5).

Combining the finding that more reads transcribed from intergenic regions in nucleus (Fig 1D), these results indicated that the rice nucleus expressed not only larger number of genes compared with protoplasts, but also novel transcripts in addition to known gene models. Taken together, the general features of rice nuclear transcriptome were similar with that of mammal cells (Grindberg, *et al*. 2013), suggesting the conservative

subcellular distribution pattern of transcripts in eukaryotic cells.

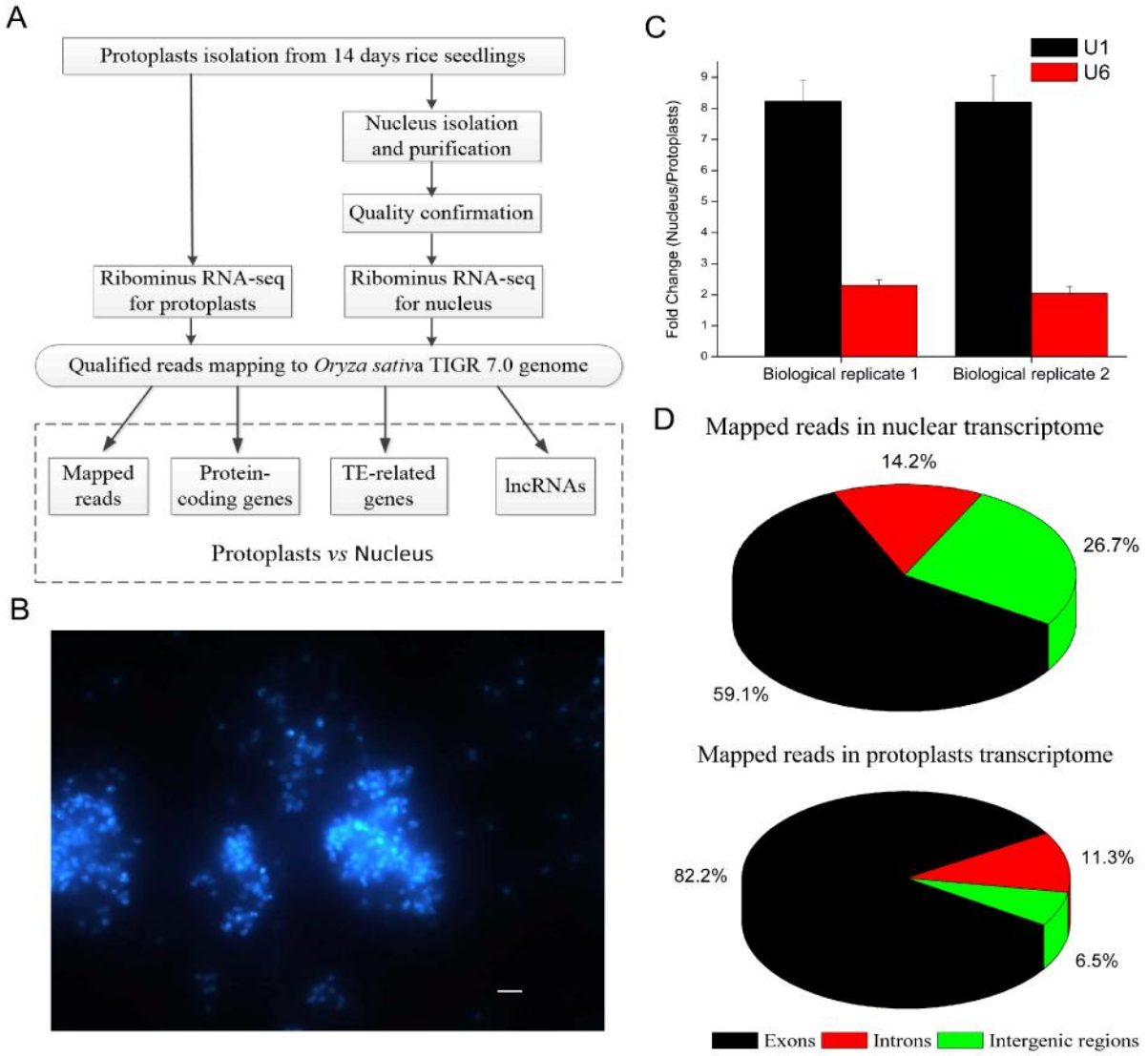

Fig 1. The experimental and analysis workflows for rice nuclear transcriptome. (A) Flowchart showing a summary of the nucleus isolation, RNA-seq and bioinformatics analysis. (B) The image of DAPI-stained nuclei by fluorescence microscopy. The length of white bar is 20μm. (C) The qRT-PCR of small nuclear RNA (U1 and U6) for two biological replicates. (D) Pie chart showing the genomic location distribution of the mapped reads in nucleus and protoplasts transcriptomes.

Nucleus-specific (%)[a] means the number of nucleus-specific expressed genes, and the number in parentheses is the percentage of reads derived from these genes among total reads mapped to genetic regions in nucleus transcriptome; **Nucleus/protoplasts (%/%)[b]** means the number of genes both detected in nucleus and protoplasts transcriptome. The first and second number in parentheses are the percentage of reads from nucleus/protoplasts expressed genes among total reads mapped to genetic regions in nucleus and protoplasts transcriptomes, respectively; **Protoplasts-specific (%)[c]** means the number of protoplasts-specific expressed genes, and the number in parentheses it the percentage of reads derived from these genes among total reads mapped to genetic regions in protoplasts transcriptome; **Undetected[d]** was the genes not detected to be expressed in nucleus or protoplasts transcriptomes.

Table 1. The expression features of genes in rice nucleus and protoplasts transcriptomes.

Gene groups	# Genes	Nucleus-specific (%)[a]	Nucleus/ protoplasts (%/%)[b]	Protoplasts specific (%)[c]	Undetected[d]
Noncoding (high)	3,546	288 (0.21)	923 (3.61/4.53)	487 (0.26)	1,848
Noncoding (weak)	3,578	402 (0.33)	976 (4.42/6.17)	458 (0.26)	1,742
TE-related genes	15,848	4,796 (5.31)	1,372 (5.11/1.97)	760 (0.36)	8,920
Protein coding genes	45,110	6,016 (5.51)	13,127 (75.49/82.56)	6,125 (3.87)	19,842
All genes	68,082	11,502 (11.36)	16,398 (88.64/95.25)	7,830 (4.75)	32,352

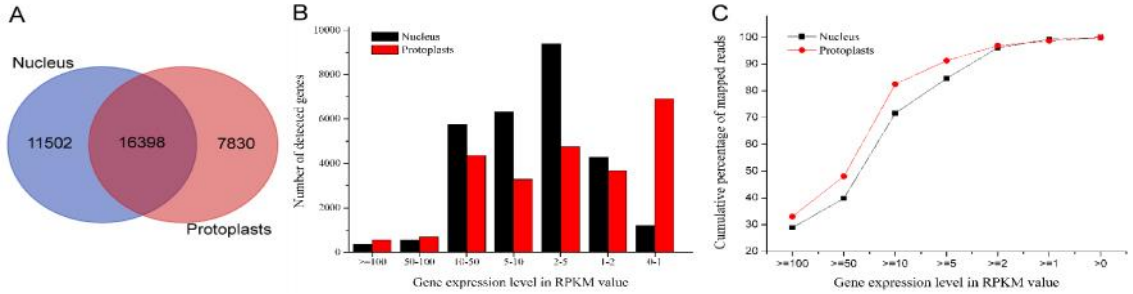

Fig 2. The general features of rice nuclear transcriptome. (A) The Venn diagram representing the number of genes detected in nucleus and protoplasts transcriptomes. (B) The number of expressed genes in nucleus and protoplasts transcriptomes within different gene expression level ranges. (C) Cumulative percentage of the mapped reads in nucleus and protoplasts transcriptomes.

The expression characteristics of TE-related genes: It was found that the subcellular location information was useful to reveal the functional clues of retrotransposon-derived transcripts in mammal cells (Fort, *et al*. 2014). The TEs account for about 35% of rice genome (International Rice Genome Sequencing Project 2005) and quite a proportion TE-related genes were expressed in different tissues (Jiao and Deng 2007; Wang, *et al*. 2017). Therefore, rice TE-related genes were identified and their expression characteristics were extensive compared between nucleus and protoplasts transcriptomes.

There are 15,848 TE-related genes annotated in rice genome, including 12,017 retrotransposon-related genes (Class Ⅰ) and 3,831 transposon-related genes (Class Ⅱ). And these two classes could be further classed into several subclasses (Fig 3A and Table S6). Generally, 43.7% of TE-related genes (6,928/15,848) were detected to be expressed in rice seedlings, which was smaller than that of conventional protein-coding genes (56.0%, 25,268/45,110, Chi-square test, $p < 0.01$). In detail, the number of nuclear expressed genes and protoplasts expressed genes were 6,168 and 2,132, accounting for 32.2% (6,168/27,900) and 9.1% (21,32/24,228) of total expressed genes in nucleus and protoplasts, respectively. And importantly, the tendency with a larger proportion of TE-derived transcripts in nucleus than that in protoplasts was always observed irrespective of sequencing depth (Fig S4A and Table S7). The public rice nuclear transcriptomes presented comparable subcellular pattern (Table S8). Moreover, the proportion of nuclear-specific expressed genes among total TE-related genes was significantly greater (32.2%, 4,796/15,848) than that among conventional protein-coding genes (13.3%, 6,061/45,110, Chi-square test, $p < 0.01$). Almost all of these subclasses from both retrotransposons-related genes and transposon-related genes followed this rule (Fig 3A and Table S6).

The overall expression abundance of TE-related genes was also different between nucleus and protoplasts. The proportion of reads mapped to TE-related genes in nucleus transcriptome was consistently larger than that in protoplasts transcriptome, irrespective of sequencing depth (Fig S4B and Table S7). Those two relevant public rice nuclear transcriptomes from roots (Reynoso, *et al*. 2018) and leaves (Yuan, *et al*. 2018) also followed this rule (Table S8). More detailed, in nucleus, reads derived from nucleus-specific and nucleus/protoplasts expressed TE-related genes accounted for 5.3% and 5.1% of total reads mapped to genetic regions, while in protoplasts, reads derived from protoplasts-specific and nucleus/protoplasts expressed TE-related genes accounted only for 0.4% and 1.9% of total reads mapped to genetic regions (Table 1). Consistently, within the group of nucleus/protoplasts expressed TE-related genes, the expression level of both retrotransposon-related genes and transposon-related genes in nucleus were higher than that in protoplasts, exhibiting extensive nuclear-enriched subcellular distribution pattern (Fig 3BC and Fig S5).

Collectively, these analyses revealed a greater diversity of nuclear-located TE-derived transcripts in rice, indicating the conservative subcellular distribution pattern of TE-related transcripts among rice seedlings and mammal cells. Furthermore, to explore the impact of nuclear-enriched subcellular distribution pattern on the

transcriptional profiling of TE-related genes, public transcriptomes from eleven rice representative tissues (Zhang, *et al*. 2014) were used to explore the expression breadth of four TE-related gene groups classed by the subcellular expression characteristics. The nucleus/protoplasts expressed TE-related gene group have the greatest proportion of genes detected in all eleven tissues and was the gene group with the greatest expression breadth, following by the protoplasts-specific expressed gene group (Fig 3D). Notably, the nuclear-specific expressed gene group, together with the undetected gene group, had larger than 30% of genes not detected to be expressed in any eleven tissues. Both retrotransposon-related genes and transposon-related genes followed this rules (Fig S6). These findings suggested that the transcription of TE-related genes, especially for the genes within nuclear-specific expressed TE-related gene group defined in this study, was prone to be underestimated using conventional bulk RNA-seq strategy.

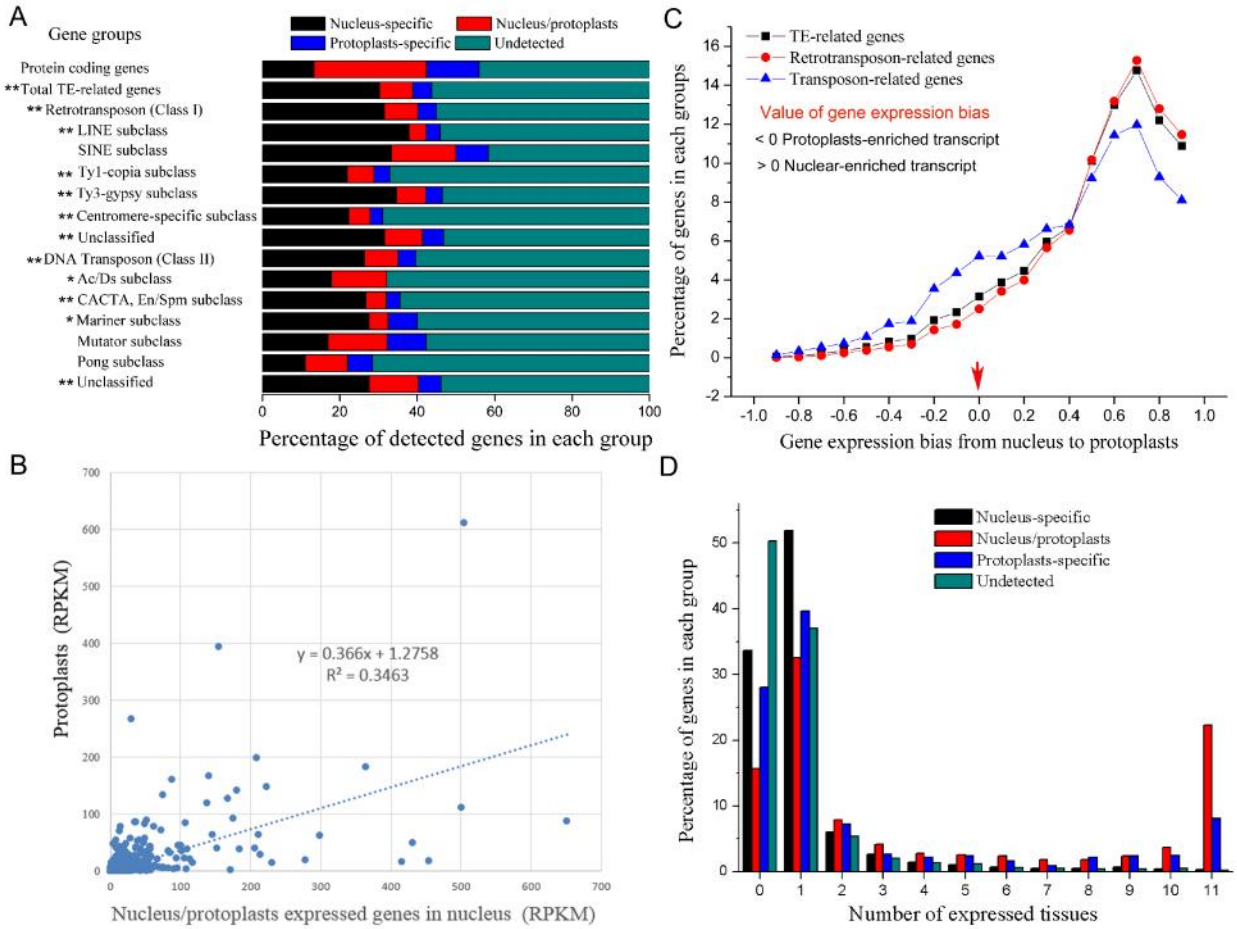

Fig 3. The expression characteristics of TE-related genes. (A) The proportion distribution of four gene groups classified based on the subcellular expression characteristics of TE-related genes and conventional protein coding genes. The repeat types of TE-related gene were downloaded from RGAP annotation (http://rice.plantbiology.msu.edu/). The classes/subclasses with significant different number of the nucleus and protoplasts specific expressed genes were marked (Chi-square test, ** means $p < 0.01$, * means $p < 0.05$). (B) The expression level of nucleus/protoplasts expressed TE-related genes in nucleus and protoplasts. (C) The distribution of gene expression bias from nucleus to protoplasts for the nucleus/protoplasts expressed TE-related genes. The gene with value (> 0) was nuclear-enriched expressed transcripts, while the gene with value (< 0) was protoplasts-enriched expressed transcripts. The gene with value (1) is the nuclear-specific expressed gene, while the gene with value (-1) is the protoplasts-specific expressed gene. (D) The expression breadth of four gene groups classified based on the subcellular expression characteristics.

The expression characteristics and functional clues of lncRNAs: There are 7,124 putative lncRNAs in integrated gene models of rice genome, including 3,546 high-confidence lncRNAs (CPC ≤-1) and 3,578 weak-

confidence lncRNAs (CPC > -1 & < 0). Totally, the subcellular distribution pattern of detected lncRNAs was similar to conventional protein coding genes with comparable proportion of nuclear-specific expressed genes and protoplasts-specific expressed genes (Fig 4A and Table 1). Moreover, the proportion of reads transcribed from high-confidance lncRNAs in nucleus was 3.8%, while that in protoplasts was 4.8%. Meanwhile, the proportion of reads derived from weak-confidence lncRNAs in nucleus was 4.8%, which was slightly lower than that in protoplasts (6.4%) (Table 1).

When it comes to the expression level of the lncRNAs simultaneously expressed in nucleus and protoplasts, both high-confidence and weak-confidence lncRNAs exhibited obvious different pattern as TE-related genes, which have a great higher proportion of nuclear-enriched transcripts (Fig 4B). Moreover, compared to conventional protein coding genes (Fig S7), both high-confidence and weak-confidence lncRNAs, irrespective within which gene groups classified based on subcellular expression characteristics, tended to be expressed with narrower expression breadth and with a higher proportion of non-expressed genes (Fig S8). This finding was in line with feature of previous detected lncRNAs in multiple plants (Li, *et al*. 2014; Liu, *et al*. 2012; Zhang, *et al*. 2014).

Lastly, twenty high-confidence lncRNAs with the highest expression level in nucleus were selected and characterized to illustrate their expression pattern among eleven representative tissues. Briefly, some of these lncRNAs had universal expression pattern or expressed with high abundance (Fig S9). Gene co-expression analysis for these nuclear abundant lncRNAs identified 921 positive co-expressed genes (Pearson Coefficient values ≥ 0.8) and 171 negative co-expressed genes (Pearson Coefficient values ≤ -0.8). Ge Ontology enrichment analysis indicated that positive co-expressed genes were enriched in "RNA binding" (GO: 0003723, p=0.00037, FDR=0.027) and "nuclear part" (GO: 0044428, p=0.00019, FDR=0.012) (Fig S10). Meanwhile, no significant enriched GO terms was found for the negative co-expressed genes of these nuclear abundant lncRNAs. These results provide functional clues of nuclear-expressed lncRNAs in rice seedlings.

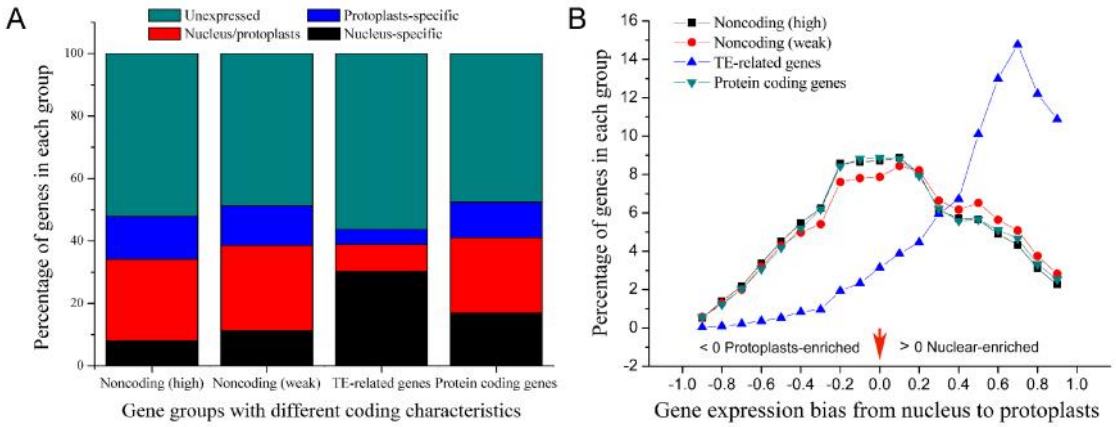

Fig 4. The expression characteristics of lncRNAs. (A) The proportion distribution of four gene groups classified based on subcellular expression characteristics with different coding characteristics. (B) The proportion distribution of gene expression bias from nucleus to protoplasts for nucleus/protoplasts expressed genes in four gene groups with different coding characteristics.

DISCUSSION

This study revealed that genes were expressed with greater diversity in rice nucleus and more genes can be detected. Meanwhile, rice nucleus were prevalent in intergenic transcription. These findings suggested that nucleus might be valuable transcriptome resources for improving existing gene models and for discovering novel genes/transcripts in rice.

It was notably that the TE-derived transcripts exhibited uneven subcellular distribution pattern, which was characterized by a great many of nuclear-specific and nuclear-enriched TE-derived transcripts. Although the enrichment of TE-related genes in nuclear transcriptome seems to be a common phenomenon in rice and mammal cells, the prevalence of nuclear-enriched TE-derived transcripts in plants and the molecular mechanism behind this observation require further study to verify. Two post-transcriptional regulation mechanisms may cooperatively explain the over-presentation of nuclear TE-derived transcripts. The first one is the siRNAs-mediated transcripts degradation. Taking that siRNAs were found in both nuclear and cytoplasmic components and TE-derived transcripts were the major targets of siRNAs (Castel and Martienssen 2013), we speculated that siRNA-mediated degradation may take place in nucleus and cytoplasm with different efficiency, thus resulting the

higher concentration of TE-derived transcripts in nucleus. The second one is the selective RNA nuclear export, which may also led to the nuclear enrichment of TE-derived transcripts. Preliminary study indicated that subcellular location of a member of retrotransposon was associated with the variation of transcript structure in Barley, which was belonged to the same grass family as rice. A Barley retrotransposon named BARE had two isoforms, one was cytoplasmic isoform with capped and polyadenylated structures, and the other was nuclear retention isoform with neither of capped or polyadenylated structure (Chang, *et al*. 2013). Therefore, the comprehensive subcellular comparison of full transcriptomes (small RNAs, long and full RNAs, long and truncated RNAs) from transcript abundance and structure levels could distinguish the contributions of above two mechanisms.

What is the biological function of extensive nuclear-enriched subcellular distribution pattern of TE-devised transcripts? These nucleus-retained TE RNA isoforms may affect the transcription of nearby genes, which was revealed by the prior study on mammalian stem cell lines (Fort, *et al*. 2014). Additionally, the lower proportion of TE-derived transcripts in the cytoplasm may reduce the burden of cellular translation machine and thus benefit for organisms. Undoubtedly, any progresses from relevant high-throughput sequencing and/or certain functional studies would enhance the understanding of the post-transcriptional regulation mechanism of TE-related genes and its biological consequence in plants.

Conclusions: In summary, this study exhibited an attractive feature of rice nuclear transcriptome. Two observations are worth paying much more attention. Firstly, more intergenic transcripts and TE-derived transcripts were detected in rice nucleus. Secondly, some nuclear abundant lncRNAs may play regulatory roles along with nuclear-located RNA binding proteins. It was the first step to uncover the complexity of rice nuclear transcriptome. Importantly, the experimental and analysis approaches developed in this study would be applicable to other plants.

Acknowledgments: This work was supported by National Key Research and Development Program of China (2018YFD0901503), National Natural Science Foundation of China (Grant No.31771473 and 31000561), and the Youth Innovation Promotion Association of Chinese Academy of Science (2017140).

REFERENCES

Bergmann, J. H., J. Li, M. A. Eckersley-Maslin, F. Rigo, S. M. Freier and D. L. Spector (2015). Regulation of the ESC transcriptome by nuclear long noncoding RNAs. Genome Res. 25(9): 1336-1346.

Castel, S. E. and R. A. Martienssen (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat. Rev. Genet. 14(2): 100-112.

Chang, W., M. Jaaskelainen, S. P. Li and A. H. Schulman (2013). BARE retrotransposons are translated and replicated via distinct RNA pools. PLoS One. 8(8): e72270.

Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas , C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow and R. Guigo (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22(9): 1775-1789.

Fort, A., K. Hashimoto, D. Yamada, M. Salimullah, C. A. Keya, A. Saxena, A. Bonetti, I. Voineagu, N. Bertin, A. Kratz, Y. Noro, C. H. Wong, M. D. Hoon, R. Andersson, A. Sandelin, H. Suzuki, C. L. Wei, H. Koseki, F. Consortium, Y. Hasegawa, A. R. Forrest and P. Carninci (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat. Genet. 46(6): 558-566.

Ghosh, S. and C. K. Chan (2016). Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol. Biol. 1374: 339-361.

Grindberg, R. V., J. L. Yee-Greenbaum, M. J. McConnell, M. Novotny, A. L. O'Shaughnessy, G. M. Lambert, M. J. Arauzo-Bravo, J. Lee, M. Fishman, G. E. Robbins, X. Lin, P. Venepally, J. H. Badger, D. W. Galbraith, F. H. Gage. and R. S. Lasken (2013). RNA-sequencing from single nuclei. Proc. Natl. Acad. Sci. U. S. A. 110(49): 19802-19807.

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. Nature. 436(7052): 793-800.

Jiao, Y. and X. W. Deng (2007). A genome-wide transcriptional activity survey of rice transposable element-related genes. Genome Biol. 8(2): R28.

Kong, L., Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei and G. Gao (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 35(W): W345-349.

Langmead, B. and S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9(4): 357-359.

Lee, J. H., E. R. Daugharthy, J. Scheiman, R. Kalhor, J. L. Yang, T. C. Ferrante, R. Terry, S. S. Jeanty, C. Li, R. Amamoto, D. T. Peters, B. M. Turczyk, A. H. Marblestone, S. A. Inverso, A. Bernard, P. Mali, X. Rios, J. Aach and G. M. Church (2014). Highly multiplexed subcellular RNA sequencing in situ. Science. 343(6177): 1360-1363.

Li, L., S. R. Eichten, R. Shimizu, K. Petsch, C. T. Yeh, W. Wu, A. M. Chettoor, S. A. Givan, R. A. Cole, J. E. Fowler, M. M. Evans, M. J. Scanlon, Y. Yu, P. S. Schnable, M. C. Timmermans, N. M. Springer and G. J. Muehlbauer (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. 15(12): R40.

Liang, Z., S. N. Anderson, J. M. Noshay, P. A. Crisp, T. A. Enders and N. M. Springer (2021). Genetic and epigenetic variation in transposable element expression responses to abiotic stress in maize. Plant Physiol. 186(1): 420-433.

Liu, J., C. Jung, J. Xu, H. Wang, S. Deng, L. Bernad, C. Arenas-Huertero and N. H. Chua (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell. 24(11): 4333-4345.

Liu, X., L. Hao, D. Li, L. Zhu and S. Hu (2015). Long non-coding RNAs and their biological roles in plants. Genomics Proteomics Bioinformatics 13: 137-147.

Mondal, T., M. Rasmussen, G. K. Pandey, A. Isaksson and C. Kanduri (2010). Characterization of the RNA content of chromatin. Genome Res. 20(7): 899-907.

Ouyang, S. and C. R. Buell (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 32(D): D360-363.

Reynoso, M.A., C. C. Pauluzzi, K. Kajala, S. Cabanlit, J. Velasco, J. Bazin, R. Deal, N. R. Sinha, S. M. Brady and J. Bailey-Serres (2018). Nuclear Transcriptomes at High Resolution Using Retooled INTACT. Plant Physiol. 176(1): 270-281.

Tan, F., G. Li, B. R. Chitteti and Z. Peng (2007). Proteome and phosphoproteome analysis of chromatin associated proteins in rice (Oryza sativa). Proteomics. 7(24): 4511-4527.

Tian, T., Y. Liu, H. Yan, Q. You, X. Yi, Z. Du, W. Xu and Z. Su (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 45(W1): W122-W129.

Vicient, C. M. (2010). Transcriptional activity of transposable elements in maize. BMC Genomics. 11: 601.

Wang, D., Z. Qu, L. Yang, Q. Zhang, Z. H. Liu, T. Do, D. L. Adelson, Z. Y. Wang, I. Searle and J. K. Zhu (2017). Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants. Plant J. 90(1): 133-146.

Wu, T. D., J. Reeder, M. Lawrence, G. Becker and M. J. Brauer (2016). GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. Methods Mol. Biol. 1418: 283-334.

Yuan, J., J. Li, Y. Yang, C. Tan, Y. Zhu, L. Hu, Y. Qi and Z. J. Lu (2018). Stress-responsive regulation of long non-coding RNA polyadenylation in Oryza sativa. Plant J. 93(5): 814-827.

Yuan, Q., S. Ouyang, A. Wang, W. Zhu, R. Maiti, H. Lin, J. Hamilton, B. Haas, R. Sultana, F. Cheung, J. Wortman and C. R. Buell (2005). The institute for genomic research Osa1 rice genome annotation database. Plant Physiol. 138(1): 18-26.

Zhang, G., G. Guo, X. Hu, Y. Zhang, Q. Li, R. Li, R. Zhuang, Z. Lu, Z. He, X. Fang, L. Chen, W. Tian, Y. Tao, K. Kristiansen, X. Zhang, S. Li, H. Yang., J. Wang and Jun Wang (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res. 20(5): 646-654.

Zhang, T., Y. Luo, K. Liu, L. Pan, B. Zhang, Y. Yu and S. Hu (2011). BIGpre: a quality assessment package for next-generation sequencing data. Genomics Proteomics Bioinformatics. 9(6): 238-244.

Zhang, Y., J. Su, S. Duan, Y. Ao, J. Dai, J. Liu, P. Wang, Y. Li, B. Liu, D. Feng, J. Wang and H. Wang (2011). A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. Plant Methods. 7(1): 30.

Zhang, Y. C., J. Y. Liao, Z. Y. Li, Y. Yu, J. P. Zhang, Q. F. Li, L. H. Qu, W. S. Shu and Y. Q. Chen (2014). Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol. 15(12): 512.