DR PENG TIAN (Orcid ID : 0000-0002-0473-5712)

DR RUI XIA (Orcid ID : 0000-0003-2409-1181)

PROF. BLAKE MEYERS (Orcid ID : 0000-0003-3436-6097)

Article type : Regular Manuscript

# Evolution and diversification of reproductive phased small interfering RNAs in *Oryza* species

Peng Tian[1, 2], Xuemei Zhang[1, 2], Rui Xia[3], Yang Liu[1, 2], Meijiao Wang[1, 2], Bo Li[1], Tieyan Liu[1], Jinfeng Shi[1], Rod A. Wing[4], Blake C. Meyers[5, 6, *], Mingsheng Chen[1, 2, *]

[1] State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

[2] University of Chinese Academy of Sciences, Beijing 100039, China

[3] State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Horticulture, South China Agricultural University, Guangzhou 510642, China

[4] Arizona Genomics Institute, BIO5 Institute and School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

[5] Division of Plant Sciences, 52 Agriculture Lab, University of Missouri, Columbia, MO 65211, USA

[6] Donald Danforth Plant Science Center, 975 North Warson Road, St Louis, MO 63132, USA

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/nph.17035

*Correspondence:

Blake C. Meyers

Telephone +1 314 587 1422

E-mail: BMeyers@danforthcenter.org

Mingsheng Chen

Telephone +86-10-64806540

E-mail: mschen@genetics.ac.cn

Received: *21 July 2020*

Accepted: *12 October 2020*

ORCID IDs:

Peng Tian: http://orcid.org/0000-0002-0473-5712

Xuemei zhang: https://orcid.org/0000-0002-1554-3388

Rui Xia: http://orcid.org/0000-0003-2409-1181

Yang Liu: http://orcid.org/0000-0003-3003-1089

Meijiao Wang: https://orcid.org/0000-0001-7568-4433

Bo Li: https://orcid.org/0000-0002-6994-9535

Tieyan Liu: https://orcid.org/0000-0001-8928-8052

Jinfeng shi: https://orcid.org/0000-0002-0208-4266

Rod A. Wing: https://orcid.org/0000-0001-6633-6226

Blake C. Meyers: http://orcid.org/0000-0003-3436-6097

Mingsheng Chen: https://orcid.org/0000-0001-7757-2777

**Summary**

- In grasses, two types of phased, small interfering RNAs (phasiRNAs) are expressed largely in young, developing anthers. They are 21 or 24 nucleotides (nt) in length, and are triggered by miR2118 or miR2275, respectively. However, most of their functions and activities are not fully understood.
- We performed comparative genomic analysis of their source loci (*PHAS*) in five *Oryza* genomes, and combined this with analysis of high-throughput sRNA and degradome datasets. In total, we identified 8216 21-*PHAS* and 626 24-*PHAS* loci. Local tandem and segmental duplications mainly contributed to the expansion and supercluster distribution of the 21-*PHAS* loci. Despite their relatively conserved genomic positions, *PHAS* sequences diverged rapidly, except for the miR2118/2275 target sites, which were under strong selection for conservation.
- We found 21-nt phasiRNAs with a 5′-terminal uridine (U) demonstrated *cis* cleavage at *PHAS* precursors, and these *cis*-acting sites were also variable among close species. miR2118 could trigger phasiRNA production from antisense transcript of its own, and the derived phasiRNAs might reversely regulate miR2118 precursors.
- We hypothesized that successful initiation of phasiRNA biogenesis is conservatively maintained, while the phasiRNA products diverged quickly and are not individually conserved. In particular, phasiRNA production is under the control of multiple reciprocal regulation mechanisms.

**Keywords**: *Oryza*, phasiRNA, male reproductive, miR2118, diversification, *cis* activity

## Introduction

Plant small RNAs (sRNA), including microRNAs (miRNAs) and small interfering RNAs (siRNAs), are short, processed transcripts with sizes ranging from 20- to 24-nt. They are incorporated into Argonaute (AGO) proteins and play important functions in multiple biological processes, such as development, phase transitions, biotic and abiotic stress responses, and transposon control (Brodersen *et al.*, 2008; Axtell, 2013; Castel & Martienssen, 2013). Both miRNAs and siRNAs are categorized into several subclasses based on their size, origin, and functional pathways. Among classes of siRNAs, the phased, small interfering RNAs (phasiRNAs) are recognized by their mode of

generation (Fei *et al.*, 2013; Komiya, 2017). The precursors of phasiRNAs, transcribed by RNA polymerase Ⅱ (Pol Ⅱ), are either protein-coding or long, noncoding RNAs. A trigger miRNA, typically 22-nt in length, initiates the processing by cleavage at a complementary site, i.e. the miRNA target site. One of the cleavage products, usually the 3′ fragment, is used as a template for synthesis of the double strand RNA (dsRNA) by RNA-DEPENDENT RNA POLYMERASE 6 (RDR6). Subsequently, the dsRNA is successively sliced by DICER-LIKE 4 (DCL4) in most land plants, or by DCL5 in grasses and some non-grasses monocots (Kakrana *et al.*, 2018), into 21- or 24-nt siRNAs, i.e. phasiRNAs. The phasiRNA-producing loci are named as *PHAS* loci (Fei *et al.*, 2013).

Functional studies of phasiRNAs are exemplified by studies of eight *trans*-acting siRNA (tasiRNA) genes in *Arabidopsis*, including *TAS1a/b/c*, *TAS2*, *TAS3a/b/c* and *TAS4*, from which the derived tasiRNAs regulate downstream genes in *trans* (Vazquez *et al.*, 2004; Axtell *et al.*, 2006). However, phasiRNAs originating from protein-coding transcripts could target genes of the same family, both in *cis* and *trans*, which is regarded as an efficient cascade regulation of large gene families, such as the pentatricopeptide repeat (*PPR*), *MYB*, *NAC*, *F-box* and nucleotide binding site leucine-rich repeat (*NLR*) families (Rajagopalan *et al.*, 2006; Howell *et al.*, 2007; Zhai *et al.*, 2011; Xia *et al.*, 2015; Sosa-Valencia *et al.*, 2017). *NLR*s, the disease resistance genes in plants, were reported to be the major phasiRNA-producing loci in many eudicots. Several miRNAs were able to induce *NLR*s into phasiRNA processing (Zhai *et al.*, 2011). One of them, the miR482/2118 family, is relatively conserved in flowering plants and targeted at the P-loop encoding region in *NLRs*; miRNA diversification is driven by the *NLR* diversity (Zhang *et al.*, 2016).

In grasses, two types of phasiRNA are abundantly expressed in developing anthers: the 21-nt phasiRNAs triggered by miR2118, and the 24-nt by miR2275. In contrast to the protein-coding transcripts derived phasiRNAs in eudicots, the precursors of phasiRNAs in grasses are mostly long noncoding RNAs (lncRNAs) (Johnson *et al.*, 2009; Song *et al.*, 2012a; Komiya *et al.*, 2014). Recent research in maize revealed that the 21- and 24-nt phasiRNAs are independently and spatiotemporally regulated in the process of anther development, with the 21-nt premeiotic phasiRNAs dependent on epidermal cell differentiation and the 24-nt meiotic phasiRNAs dependent on tapetal cell differentiation (Zhai *et al.*, 2015). Similar expression pattern was also found in rice (Fei *et al.*, 2016; Tamim *et al.*, 2018), indicating different roles of the two types of phasiRNAs in male organ development. Several mutations in the phasiRNA pathway are known to result in failure of male fertility in rice, including defects in the key components

RDR6 (Song *et al.*, 2012b) and DCL4 (Liu *et al.*, 2007). MEIOSIS ARRESTED AT LEPTOTENE 1 (MEL1) is the only AGO protein that is confirmed to preferentially interact with 21-nt phasiRNAs with a 5′-terminal cytosine (C) (Komiya *et al.*, 2014). The *mel1* loss of function mutant showed abnormal tapetum and aberrant pollen mother cells (Nonomura *et al.*, 2007). Photoperiod-sensitive male sterility (PSMS), which initiated two-line hybrid rice breeding and contributes tremendously to rice production in China, is a phenomenon in which particular rice strains show male sterility or fertility, under long or short day conditions, respectively. Two controlling loci, *pms1* and *pms3*, encode lncRNAs *PMS1T* and *LDMAR*, respectively (Ding *et al.*, 2012; Fan *et al.*, 2016). Both of them are targets of miR2118 and produce 21-nt phasiRNAs (Fan *et al.*, 2016; Tamim *et al.*, 2018). These studies highlight the role of phasiRNAs in male reproductive development, with some loci controlling important agronomical traits. Recently, Tamim *et al.* (2018) found that 21-nt reproductive phasiRNAs can direct *cis* cleavage of their own precursor or bottom-strand transcripts and yield tertiary sRNAs in both rice and maize, indicating that cleavage activity is directed by some 21-nt phasiRNAs (Tamim *et al.*, 2018). However, how these reproductive phasiRNAs evolve and whether there are functionally conserved phasiRNAs remains uninvestigated.

The genus *Oryza* is a model system for comparative and evolutionary studies. With the advances in the International *Oryza* Map Alignment Project (I-OMAP), at least 13 species in this genus have completed whole genome sequences (Stein *et al.*, 2018). In the present study, we used five *Oryza* species to investigate the reproductive phasiRNAs, including *Oryza sativa*, *Oryza rufipogon*, and *Oryza glaberrima* with the AA genome type, *Oryza punctata* with the BB genome type, and *Oryza brachyantha* with the FF genome type (Chen *et al.*, 2013). These five species have clear phylogenetic relationships, with an evolutionary gradient of separation ranging from ten thousand years to 15 million years (MYA) (Stein *et al.*, 2018). Thus, we could analyze the conservation patterns and evolutionary changes of phasiRNAs and *PHAS* loci within this model system. In addition, using high-depth sRNA and degradome data, we could explore the potential function of phasiRNAs using several computational approaches.

## Materials and Methods

### Plant materials

*O. sativa* (Nipponbare), *O. rufipogon* (W1943), *O. glaberrima* (IRGC96717), *O. punctata* (IRGC105690) and *O.*

*brachyantha* (IRGC101232) were grown in an experimental field in Beijing, China. To economically and efficiently identify *PHAS* loci, we used mixed young panicles with lengths including 1-2, 2-4, 6-8 and 8-10 cm in each species.

**High-throughput sRNA and degradome sequencing**

Total RNA was isolated using the Trizol reagent (Invitrogen, Carlsbad, CA, USA). sRNA library preparation was as described previously (Zhao *et al.*, 2013). The degradome library construction was performed according to a previous study (Ma *et al.*, 2010). All libraries were sequenced on an Illumina HiSeq-2000 instrument at the Beijing Genomics Institute (BGI), Shenzhen, China.

Mapping of sRNA and degradome reads was performed using Bowtie (V1.1.0) (Langmead *et al.*, 2009). Any read with more than 20 perfect hits to the genome was excluded from further analysis. Abundances of short reads in each library were normalized to "TP10M" (transcripts per 10 million). The genome sequence and annotation of *O. sativa* referred to RGAP 7 (http://rice.plantbiology.msu.edu/), the other four species referred to I-OMAP (Stein *et al.*, 2018).

***PHAS* loci and trigger miRNA target site identification**

*PHAS* loci were identified as described previously (Xia *et al.*, 2013). In addition, regions overlapping with tRNAs, rRNAs and snoRNAs were filtered out. To identify the trigger miRNA target sites, (1) we used MEME (V4.10.0) (Bailey *et al.*, 2009) to predict the 22-nt recognition motif of miR2118 and miR2275 in the whole genome, respectively; (2) identified the cleavage signals at the 10 - 11 position relative to the 3′ site of the recognition motif on the opposite strand using the degradome data (≥ 3 degradome reads at cleavage site); (3) for each *PHAS*, we searched for these corresponding motifs with cleavage signals at both sides (100 bp extension); and (4) *PHAS* loci with more than two recognition sites were manually divided into separate *PHAS* loci.

**Comparison of *PHAS* between species**

Orthologous genes between *O. sativa* and other *Oryza* species were identified using SynOrths (V1.0) (Cheng *et al.*, 2012). Synteny blocks were determined with two adjacent orthologous genes as anchors. We performed homology searches of *O. sativa PHAS* sequences in the other four genomes, or conversely, searching for the *PHAS* sequences of the four species in the *O. sativa* genome. If *PHAS* loci from two species were inside syntenic blocks, they were considered to have syntenic positions. If their sequences were also homologous (BLASTN, E-value < 1e-8 and with

coverage ≥ 1/3 of the query sequence), they were regarded as possible orthologs.

**Substitution rate analysis**

Homologous sequences were aligned using MUSCLE (V3.8.31) program (Edgar, 2004) using default parameters. Nucleotide sequence divergence (K) of noncoding sequence were estimated using the baseml modules in the PAML (V4.6) software (Yang, 2007).

**Sequence data**

Sequence data from this article can be found in the NCBI Sequence Read Archive database under accession number PRJNA504938.

## Results

### Identification and characterization of *PHAS* loci

To identify the *PHAS* loci in *Oryza*, we performed high-throughput sRNA sequencing (sRNA-seq) of young panicles from *O. sativa*, *O. rufipogon*, *O. glaberrima*, *O. punctata* and *O. brachyantha*, separately. Degradome sequencing (degradome-seq) or parallel analysis of RNA ends (PARE) (German *et al.*, 2008) was carried out using the same samples to analyze miRNA and siRNA targets. In addition, two additional replicates of the sRNA-seq and degradome-seq were prepared for *O. sativa* (Table S1). Based on the sRNA-seq mapping results and *PHAS* analysis, a total of 8216 21-*PHAS* and 626 24-*PHAS* loci were detected in the five genomes (Fig. 1A and Table S2). miR2118 target sites were detected in 91% of the 21-*PHAS* loci, and miR2275 target sites in 76% of the 24-*PHAS* loci, respectively. The degradome reads were strand-specific and consistent with the original transcripts. Thus, we could determine the transcriptional direction of the phasiRNA primary precursors, i.e. the single-stranded, Pol Ⅱ RNA transcripts (only named as phasiRNA precursors subsequently), according to the reads indicating cleavage at the target sites. From the sRNA mapping results, 21- and 24-nt sRNAs were the most abundant in quantity (Fig. 1B), and the 24-nt showed the greatest sequence diversity (Fig. 1C). These sRNAs were derived from diverse regions of the genomes (Fig. 1D), and more than 92% corresponded to 21-nt in 21-*PHAS* or 24-nt in 24-*PHAS*. Approximately 90% of the 21-nt sRNAs in 21-*PHAS* were “in phase” to the cycles of processing (Fig. 1E), while roughly 80% of 24-nt in

24-*PHAS* were "in phase" (Fig. 1F). These results support the accuracy of the data for the *PHAS* and miRNA target sites, thus facilitating subsequent analyses.

Based on the regions producing sRNAs, the size of the *PHAS* loci ranged from 100 to 2605 bp, with a mean size of ~400 bp. About 87% are distributed in the intergenic regions, while a few overlap with annotated genes (~7%), or transposable elements (TE, ~6%), at either the sense or the antisense strand (Fig. S1A, B). We analyzed the GC composition at the 21-*PHAS* and 24-*PHAS* regions using a sliding-window approach. We found that the GC composition within the 21-*PHAS* loci was significantly higher than that in the flanking regions (P = 2.84e-11), and also higher than that in the 24-*PHAS* loci (P = 2.702e-16) while the GC composition at 24-*PHAS* loci was relatively lower than their flanking regions (P = 7.046e-09). Such GC patterns were consistent in all five genomes (Figs S1C-G). The difference was more significant at the miRNA target sites, which is consistent with the high GC content of miR2118 compared with that of miR2275 (Figs S1H, I). These results suggest a quite different GC composition between 21-*PHAS* and 24-*PHAS* loci.

We found miR2118 predicted to target multiple *NLR* genes in *Oryza* species. Consistent with reports in eudicots (Zhai *et al.*, 2011), target sites were inside P-loop encoding regions (Figs S2A, B), suggesting an evolutionarily conserved regulation of miR2118 at *NLRs*. However, most *NLR* targets did not produce obvious phasiRNAs (Figs S2A, B), and the number of *NLR* targets was less. In rice, over 500 *NLRs* are encoded in the genome (Luo *et al.*, 2012), at least 228 of them expressed (> 1 FPKM) in anther (based on the analysis of rice RNA-seq data from anthers, DRR016141), from which the miR2118 recognition motif was found in 85 genes. Only 18 of these were identified as targets with obvious miR2118-induced cleavage, and five of them showed weak phasiRNA signals after the cleavage sites, indicating a relatively minimal interaction of miR2118 with *NLR* transcripts in *Oryza*.

**Local tandem and segment duplications contributed to the expansion and supercluster distribution of 21-*PHAS***

The 21-*PHAS* loci were found to form superclusters, as previously reported (Johnson *et al.*, 2009). We further found that 21-*PHAS* loci in the same superclusters tended to maintain a consistent precursor direction. For example, the largest cluster in rice (on chromosome (Chr) 12 with position from 21,716,057 bp to 22,124,703 bp) contains one hundred and eleven 21-*PHAS*, 104 of which have the same direction on the genomic top or "plus" strand, suggesting

that these 21-*PHAS* precursors were consistently transcribed from the same strand. In addition, the large quantity and supercluster distribution of 21-*PHAS* loci suggested a localized gene expansion in *Oryza*. To investigate the expansion pattern, we analyzed DNA duplication by homologous searching of *PHAS* sequences over the repeat-masked genome using BLASTN, and compared the position of *PHAS* loci across the duplicate segments (E-value < 1e-8, and a homologous region ≥ 1/3 of the query *PHAS*).

Overall, most *PHAS* sequences are unique or low-copy, with over 70% of them found as single copy in *Oryza* genomes (Fig. S3A). For duplications that contributed to the expansion of *PHAS* loci, 83% ~ 100% of these duplicate segments were adjacent intrachromosomal duplications (Fig. 2A), that is, localized duplications with a median distance less than 8 kb. The directions of the paired, duplicated *PHAS* were mostly consistent (77% ~ 96%). This was found in both 21-*PHAS* and 24-*PHAS*, indicating the expansion of *PHAS* loci was primarily induced by localized tandem or segmental duplications. Only a few intrachromosomal or interchromosomal segment duplications contributed to the expansion of *PHAS* loci (Figs S3B, C). For instance, a detailed analysis of the duplication status of 21-*PHAS* loci in a supercluster on *O. sativa* Chr 4 revealed that all sixty-six 21-*PHAS* loci were transcribed from the plus strand, and tandem duplications were frequent in this region (Fig. 2B). The sequences of OS_PHAS1311, 1313, 1315, 1317 and 1319 were all homologous, demonstrating high levels of duplication of some 21-*PHAS* loci. One duplication event might include, and thus lead to the duplication of several adjacent 21-*PHAS* loci (Fig. 2C). Local duplications, especially tandem duplications, would lead to closely distributed and consistent precursor directions of *PHAS* loci, yielding the superclusters that we observed.

**Position but not sequence conservation of *PHAS* loci between more divergent species**

To investigate the sequence conservation of these *PHAS* loci, we constructed syntenic blocks between *O. sativa* and other *Oryza* species with orthologous genes as borders. Then, we performed bidirectional, homology searches of *PHAS* sequences between *O. sativa* and other *Oryza* species to identify potentially homologous *PHAS* pairs (Figs 3A, B). The proportion of *PHAS* loci with homologs decreased quickly as the evolutionary distance increased. Within the less-diverged AA group (<1 MYA) (Stein *et al.*, 2018), 70 to 90% of *PHAS* loci maintained homology. However, less than 40% of *PHAS* loci in *O. sativa* were homologous to those in *O. punctata* (6.76 MYA). Furthermore, only 5% of

*PHAS* loci in *O. brachyantha* (15 MYA) were homologous in *O. sativa*. Over a wider evolutionary distance, the proportion of sequences with homology in syntenic regions also decreased in CDS, UTR and intron of protein-coding genes (Fig. 3C). Nevertheless, the ratio was lowest in *PHAS* loci and their flanking regions, even lower than introns, which were assumed to be neutrally selected (after removing of the two splice site adjacent ends). These results indicated rapid divergence at *PHAS* regions.

Although the sequences of *PHAS* loci changed quickly, their genomic distributions were mostly consistent among the five species (Figs 3D, S4). Further investigation of these *PHAS* regions showed that the majority of *PHAS* loci were embedded in common syntenic blocks. For the two most divergent species, we found that ~85% (1986) of 21-*PHAS* loci in *O. sativa* and ~ 90% (892) in *O. brachyantha* were present in 106 common syntenic blocks. The number of 21-*PHAS* loci in the syntenic blocks were highly correlated (Pearson's product-moment correlation $r = 0.70$, $P < 2.2e\text{-}16$), suggesting that the number of *PHAS* loci were comparable between syntenic positions. Furthermore, the number of 21-*PHAS* loci on the two strands were also highly correlated (plus strand: $r = 0.80$, $P < 2.2e\text{-}16$, minus strand: $r = 0.62$, $P < 2.2e\text{-}16$), suggesting the precursor directions of these 21-*PHAS* were also mostly consistent in the syntenic regions.

To illustrate the properties of their genomic distribution, we examined a syntenic region of the 21-*PHAS* supercluster on Chr 6, in which 27, 27, 25, 28, and eighteen 21-*PHAS* loci were identified in the five species, respectively (from top to bottom, Fig. 3E). Of the twenty-seven 21-*PHAS* loci in *O. sativa*, all had orthologs in *O. rufipogon*; 22 had orthologs, three were segments (partial ortholog) and one was a homolog (outside the syntenic region) in *O. glaberrima*; 15 were orthologs, six were segments and two were homologs in *O. punctata*; and only three were segments in *O. brachyantha*, further supporting the rapid diversification of *PHAS* sequences. Most of these loci were found on the minus strand in the five species, suggesting the transcription directions of these *PHAS* precursors were conservatively maintained. However, in *O. punctata*, an inversion event was found to change the direction of at least 10 *PHAS* loci, suggesting that directional changes may result from specific evolutionary events (Fig. 3E). Finally, because of their low numbers and sparse distribution, the positional conservation pattern was weak at 24-*PHAS* regions.

**miR2118 and miR2275 target-site biased selection**

Reproductive phasiRNA processing is initiated by miR2118 or miR2275 targeting with high base complementarity at target sites. The mature sequences of the two miRNAs were mostly conserved among *Oryza* species (Table S3). Correspondingly, the target sites should be under purifying selection to match the trigger miRNAs. To test this hypothesis, we calculated the nucleotide substitution rates (K) of *PHAS* loci, the 500 bp flanking sequences on both sides, the miRNA target sites (separately), and we compared these results with the K of neutrally-selected introns from protein-coding genes that adjacent to *PHAS* loci (intron_a). To evaluate the conservation level of *PHAS* and adjacent region, we compared the K of intron_a with randomly selected introns (intron_r), and also compared the K of *PHAS* with intergenic lncRNAs (lincRNAs) that didn't produce phasiRNAs. We found that the K values of *PHAS* and their flanking regions were significantly higher than those of introns. Notably, the K values of intron_a were also significantly higher than intron_r (Figs 4A, B), further indicating that *PHAS* sequences were in regions experiencing high rates of variation. As expected, the K values of miR2118 and miR2275 target sites were significantly lower than introns (Figs 4A, B), consistent with purifying selection at target sites. The K values of 21-*PHAS* loci were relatively lower than in the flanking sequences (Fig. 4A), indicating weak selection at the 21-nt phasiRNA producing regions. In addition, we found no significant difference of the K between *PHAS* and non-*PHAS* lincRNAs, suggesting the sequence conservation level of *PHAS* were comparable with other lincRNAs.

To further assess the pattern of variation at *PHAS* regions, we analyzed the average variation at *PHAS* loci and their flanking regions using the distribution density of population SNPs and indels from the 3000 rice genomes project (Alexandrov *et al.*, 2015). In addition, the occupation ratio of TEs, a major cause of genome variation, was also used to measure the variation pattern (Fig. 4C). We found an unbalanced pattern of variation, especially in the 21-*PHAS* regions, in which sequences close to the miR2118 target site tended to be more conserved than those distal, with the lowest density of variation at the target site, indicating relatively strong selection at target site sequences. Only a few 24-*PHAS* loci were available for the analysis; therefore, we could not determine the pattern of variation at 24-*PHAS* loci with strong support. However, the miR2275 target sites also showed more conservation based on the indel and TE distribution patterns (Fig. 4C). Furthermore, we found all three variation types occurred more frequently in 24-*PHAS* loci than those in 21-*PHAS* (Fig. 4C), indicating a higher mutation rate at 24-*PHAS* loci.

We conducted sequence analysis of two rice 21-*PHAS* loci known to function in PSMS. In addition to *PMS1T* at

the *pms1* site (Fan *et al.*, 2016) (Fig. S5A), *LDMAR* from *pms3* is also the target of miR2118 and produces 21-nt phasiRNAs (Ding *et al.*, 2012) (Fig. S5B). The two 21-*PHAS* loci from *LDMAR* and *PMS1T* corresponded to OS_PHAS750 and OS_PHAS1809, respectively. However, the regions of the two 21-*PHAS* loci were located inside the two lncRNAs (Figs S5A, B top panel). Based on the phasiRNA processing mechanism, this suggests that only a portion of the 3′ fragments after the miR2118 cleavage site of the lncRNAs were used for phasiRNA generation. The orthologous loci of *LDMAR* were found in the AA group and *O. punctata*, while the orthologs of *PMS1T* were only present in the AA group, suggesting the recent emergence of *pms1* and *pms3*. Consistent to the target site biased conservation, sequences at the miR2118 target sites in the two *PHAS* loci were more conserved, with a low SNP frequency in both (Figs S5A, B middle panel). Crucially, it was reported that the causal variants of PSMS in the two loci were SNP mutations in the second cycle after the miR2118 target sites (Figs S5A, B bottom panel) (Ding *et al.*, 2012; Fan *et al.*, 2016). These two studies showed that sequences surrounding the SNPs, which contained the miR2118 target sites, are critical for the function of the two lncRNAs. *LDMAR* and *PMS1T* without the target site were both non-functional, probably because of failure to generate phasiRNAs (Ding *et al.*, 2012; Fan *et al.*, 2016).

**21-nt phasiRNAs with 5′-terminal U tend to induce *cis*-cleavage**

Recently, Tamim et al. (2018) reported *cis*-cleavage function by some 21-nt reproductive phasiRNAs that target their own precursor or bottom strand transcripts. However, the rapid variation of *PHAS* sequences might lead to numerous divergent phasiRNAs in different species. The kinds of 21-nt phasiRNAs that can direct the target cleavage need to be investigated.

To identify functional (cleavage-directing) 21-nt phasiRNAs and their targets, we analyzed three high-throughput degradome datasets (Table S1) in *O. sativa*. We selected 21-nt phasiRNAs with at least 5 TP10M for target prediction at a genome-wide scale. In brief, we first mapped these degradome reads to the *O. sativa* genome, and selected the cleavage signals with sharp peaks using a described method (Zhai *et al.*, 2015). Next, we extracted the sequence around the cleavage sites (we named this sequence the “cleavage-tag”). Then, we predicted the complementarity and scored the 21-nt phasiRNAs with these cleavage-tags using TargetFinder (Fahlgren & Carrington, 2010). Lastly, we selected those with best complementary (i.e. lowest penalty score) to the cleavage-tag and slicing site exactly 10 bp

after the 5′ terminus of the 21-nt phasiRNAs. Based on this analysis, we identified 2303, 1237, and 1557 potentially functional 21-nt phasiRNA-target pairs in the three datasets (Fig. S6A and Table S4). However, many of these 21-nt phasiRNA-target relationships were sample-specific, with only 553 shared across the three replicates (Table S5), accounting for 24.0%, 44.7% and 35.5% in each replicate. For comparison, we also analyzed rice miRNA targets using the same degradome datasets under the same criteria. A total of 1201 consistent miRNA-target pairs were identified, accounting for 84.4%, 87.3% and 85.5% in each replicate (Fig. S6B). This included many previously reported interacting partners, like miR156 to *OsSPL*s (Xie *et al.*, 2012), miR160 to *OsARFs* (Huang *et al.*, 2016), miR2118 and miR2275 at corresponding target sites in *PHAS* loci, supporting the robustness of our methods. It is probable that the reason for the relatively low reproducibility for most 21-nt phasiRNA-target interactions is weak or unstable cleavage, indicated by much lower quantity of degradome reads at those sites only detected in one library (V1) than those replicated in two (V2) or three (V3) libraries (Figs S6C, D). It may also be that the transcripts targeted by reproductive phasiRNAs are expressed at low levels and yield few degradome reads.

The 553 shared 21-nt phasiRNAs-targets are more likely to be functional than the sample-specific pairs. To determine whether there are common characteristics for these 21-nt phasiRNAs (a total of 540) and their target sites, we analyzed the nucleotide composition of the 21-nt sequences and the genomic features of the target sites, separately. We found that the 5′-terminal nucleotide of these 21-nt phasiRNAs was heavily biased toward uridine (U), accounting for over 92% of the instances, while the other positions showed no obvious bias (Fig. 5A). Consistent with the *cis*-activity (Tamim *et al.*, 2018), most of the target sites were inside 21-*PHAS* regions (Fig. 5B). The *cis* regulation included the *cis*-targeting of bottom strand phasiRNAs at their precursors derived from RNA Pol Ⅱ (*PHAS*_top), which accounted for ~65% of the targets. Another kind of *cis* regulation was top-strand phasiRNAs targeting at RDR6-derived bottom strand RNA (*PHAS*_bottom), as described by Tamim et al. (2018), which account for ~20%. We also discovered some *trans* regulations, including 21-nt phasiRNAs from one locus targeting the RNAs from another *PHAS* (4.7%), and a few 21-nt phasiRNAs targeting annotated genes (6.5%, many with unknown function and no enrichment at any GO term), TEs (0.9%) and unannotated intergenic regions (3.1%) (Table S5). To analyze whether functional 21-nt phasiRNAs tend to be present in specific phase cycles, we calculated the frequency of occurrence of these 540 21-nt phasiRNAs occurred in the first 25 cycles. We found that functional 21-nt phasiRNAs could be from several cycles, but with a relatively high ratio from the cycles close to the miR2118 target sites (Fig.

5C), consistent with the view that sequences adjacent to the target sites tend to be more conserved, i.e. more conservation at functional 21-nt phasiRNAs. Exceptionally, the proportion of functional 21-nt phasiRNAs from the first cycle (C1) was much lower. This might be consistent with the former report that 21-nt phasiRNAs from C1, which contained half of the miR2118 target site, was highly biased with low abundance and stability and might be non-functional (Tamim *et al.*, 2018). In brief, these results showed that 5′ U 21-nt phasiRNAs have a high potential to direct *cis*-cleavage, perhaps reflecting loading into AGO1, which preferentially interacts with 5′ U 21-nt siRNA to induce cleavage (Mi *et al.*, 2008), rather than the more typical AGO5 loading for 5′ C 21-nt phasiRNAs.

**_Cis_-acting 21-nt phasiRNAs show low conservation**

To identify evolutionarily conserved and functional (cleavage-directing) 21-nt phasiRNAs, we analyzed the targets of 21-nt phasiRNAs in the other four *Oryza* species using the degradome datasets (Table S1), in comparison with that in *O. sativa*. Consistent with the rapid divergence of *PHAS* sequences, the number of homologous 21-nt phasiRNAs (≤ 4 nt difference) from orthologous *PHAS* loci also decreased quickly with increasing evolutionary distance, and were rarely detected between *O. sativa* and *O. brachyantha* (Table S6). However, the abundance of homologous phasiRNAs between *O. sativa* and other *Oryza* species was highly correlated (Fig. S7A), suggesting conservation of expression for homologous phasiRNAs, to a certain degree. Using the same method as mentioned above, we found 504, 976, 572 and 1035 potentially functional 21-nt phasiRNA-target pairs in *O. rufipogon*, *O. glaberrima*, *O. punctata* and *O. brachyantha*, separately (Table S4). Consistently, 42% to 75% of these functional 21-nt phasiRNAs were 5′ U (Fig. S7B), and 60% to 78% of the targets were under *cis*-regulation in the four species (Fig. S7C). For the 540 functional 21-nt phasiRNAs in *O. sativa*, 57% and 55% of them have homologs in less diverged *O. rufipogon* and *O. glaberrima*, respectively. Their targets were also mostly similar or shared (Table S5). We further confirmed that these functional phasiRNAs tended to be from relatively conserved loci in *O. rufipogon* and *O. glaberrima*, based on Fisher's exact test (Table S6). However, most of these consistent targets were *cis*-regulated, and we identified few consistencies at *trans*-regulated sites.

To illustrate the *cis*-cleavage activity of 21-nt phasiRNAs, two relatively conserved 21-*PHAS* loci were analyzed. In C6 of OS_PHAS768 (Fig. 5D), *cis*-cleavage was found in *O. sativa* and the homologous site in *O. glaberrima*, and

*O. punctata*, with obvious cleavage at the $10^{th}/11^{th}$ position relative to the 21-nt phasiRNA from the bottom strand. The sequence of *O. rufipogon* at C6 is consistent with that of *O. sativa* (Fig. S8A). The lack of *cis*-cleavage in *O. rufipogon* might be the weak cleavage not detected in this degradome library. In OS_PHAS1660, multiple *cis*-acting 21-nt phasiRNAs were identified (Fig. 5E). Particularly, *cis*-cleavage at C4, C8 and C9 was supported in the AA group but not in *O. punctata*. Two indel mutations between the miR2118 target sites and the functional sites lead to the shift of phase register and resulting different 21-nt phasiRNA products (Fig. S8B). The 5′ nucleotides of the functional phasiRNAs (bottom strand) in the three cycles were U in AA group, while they were A, G, and G in *O. punctata* (Fig. S8B, Table S5), which are less likely to be loaded by AGO1 than the 5′ U type. However, there were other *cis*-cleavage sites (C3, C5) in *O. punctata* (Fig. 5E), suggesting *cis*-regulation were conserved in this 21-*PHAS* but with species-specific sites between AA and *O. punctata*. From this case, we also realized that indels with length that were not the multiple of 21 were easy to change the phase cycles and derived phasiRNAs at *PHAS* loci.

**miR2118 induces its natural antisense transcripts into phasiRNA processing**

The precursors of miR2118 distribute as clusters on Chr 4 and Chr 11 in the *Oryza* AA and BB groups, and primarily on Chr 4 in *O. brachyantha* (Fig. 6A). In *O. sativa*, the lncRNA AK068680 was found to be encoded on the antisense strand of the miR2118 cluster at the Chr 4 loci (Fig. 6B), with miR2118 precursors mostly in the intron of AK068680, which were also shown in the work of Song et al., 2012a. According to published lncRNAs in *O. sativa* (Zhang et al., 2014), XLOC_035472 (equal to AK068680) and XLOC_012293, are transcribed from the antisense stand of miR2118 clusters on Chr 4 and Chr 11, respectively (Fig. S9). These findings indicated that both strands were transcribed at the miR2118 loci, i.e. there were natural antisense transcripts (NATs) for miR2118 in *O. sativa*. Interestingly, 21-*PHAS* loci were also found in this region, with their Pol Ⅱ precursor direction consistent with the lncRNAs (NATs, Fig. 6B). Particularly, the miR2118 target sites of these 21-*PHAS* were only on the antisense strand of mature miR2118 (Fig. 6C). This indicated that miR2118 directed *cis*-cleavage of its own NATs or preRNAs of the NATs, as most 21-*PHAS* were within the intron or partially overlap with the exon of the NATs, and triggered the processing of 21-nt phasiRNAs (Fig. 6D), with the perfect target site on the opposite strand of mature miR2118. In *O. sativa*, *cis*-cleavage was found in two-thirds (20/30) of the mature miR2118 sites on the opposite strand, and at least nine 21-*PHAS* loci

were produced in this manner. Similar results were also found in the other four *Oryza* species at the miR2118 cluster regions (Fig. 6A), suggesting a conserved *cis*-regulation between miR2118 and its NATs in *Oryza*. Based on the *cis*-activity of 21-nt phasiRNAs, the derived 21-nt phasiRNAs from miR2118 NATs might have the potential to target the precursor of miR2118 or the NATs. Perhaps due to the low abundance of 21-nt phasiRNAs or weak cleavage signals in the panicle tissue that did not pass the threshold we set above, we observed a few 21-nt phasiRNAs that direct cleavage of miR2118 precursors in *cis*.

## Discussion

Reproductive phasiRNAs in grasses were mysterious for their specific expression, large varieties and important but mostly unknown function in anthers development. Here, we comprehensively investigated the genomic, evolutionary and functional properties of reproductive phasiRNAs and *PHAS* loci in *Oryza*. We noticed that the position but not sequence conservation of *PHAS* loci is quite similar to other non-*PHAS* lncRNAs (Mohammadin *et al.*, 2015; Wang *et al.*, 2015; Deng *et al.*, 2018). However, *PHAS* loci, especially the 21-*PHAS*, are distinct for their trigger miRNA target-site biased selection and supercluster distribution patterns (Johnson *et al.*, 2009). We found miR2118 could also target *NLRs* in *Oryza* but producing no or sparse phasiRNAs, which differs from many eudicots, in which *NLRs* are frequently found to be loci generating miR482/2118-triggered phasiRNAs (Zhai *et al.*, 2011). It was reported that intron splicing, occurring in most protein-coding RNAs, might prevent siRNA generation in RDR6 mediated gene silencing (Christie *et al.*, 2011). Research in *Arabidopsis* revealed that phasiRNA biosynthesis from noncoding *TAS* transcripts occurred, or at least initiated, on membrane-bound polymers at rough endoplasmic reticulum (ER), a place suggested to inhibit the entry of mRNAs into phasiRNA biosynthesis (Li *et al.*, 2016). These studies indicate that lncRNAs other than mRNAs are preferentially processed into phasiRNAs. Thus, we proposed that the large numbers of reproductive-specific *PHAS* lncRNAs in rice (possibly over 1000) (Komiya *et al.*, 2014) may strongly compete with *NLRs* for miR2118 targeting and the phasiRNA processing machinery, lead to minimal regulation of miR2118 to *NLRs*. Besides, the different expression pattern of miR2118 between eudicots and grasses might be another cause of the targets difference. In eudicots, miR482/2118 is expressed in both vegetative and reproductive tissues (Zhai *et al.*, 2011), while miR2118 in grasses is expressed mainly in male reproductive tissues

(Song *et al.*, 2012a), in which reproductive lncRNAs are abundant.

The discovery that 5′ U 21-nt phasiRNAs tend to direct *cis*-cleavage is expected. The functions of siRNAs are AGO-dependent (Fang & Qi, 2016). AGO1 in plants preferentially binds to 5′ U 21-nt miRNAs or siRNAs and induces target cleavage (Mi *et al.*, 2008). Recently, AGO1d and AGO1b were described as strong candidates to interact with 21-nt phasiRNAs in rice, based on coordinated expression of these AGOs with 21-nt phasiRNAs (Fei *et al.*, 2016; Araki *et al.*, 2020), possibly they bind the 5′ U type. Though we proposed the *cis*-cleavage of 5′ U 21-nt phasiRNAs, the possibility of some other functions of 21-nt phasiRNAs with 5′ C, A or G should not be excluded. Especially, MEL1- or AGO5c- loaded 5′ C 21-nt phasiRNAs are most abundant (Komiya *et al.*, 2014; Patel *et al.*, 2018), suggesting an important function for some 5′ C 21-nt phasiRNAs. Recent studies in maize have proposed a putative function of phasiRNAs in *cis* DNA methylation (Dukowic-Schulze *et al.*, 2016). Based on the analysis of public MethylC-seq datasets in rice, we also observed that CHG and CHH (H = A, C, or T) DNA methylation in *PHAS* regions, especially the body of 24-*PHAS*, at reproductive panicles (Li *et al.*, 2012) was significantly higher than that at vegetative leaf tissues (Zemach *et al.*, 2010) (Fig. S10). C*is*-activities including *cis*-cleavage and possibly *cis*-methylation might serve as negative feedback regulation in phasiRNAs processing (Patel *et al.*, 2018). On the one hand, *cis*-cleavage by bottom-strand, 21-nt phasiRNAs may directly reduce the quantity of *PHAS* precursors. On the other hand, increased DNA methylation at *PHAS* regions may also influence the transcription of the precursors. *Cis*-activities of phasiRNAs were consistent with their genomic property. As most *PHAS* loci were found to be unique, with few copies, and the derived phasiRNAs lacked complementarity to other sites, phasiRNAs might target mainly in *cis* (Zhai *et al.*, 2015; Patel *et al.*, 2018). In addition, we found that the *cis*-acting 21-nt phasiRNAs tend to be from conserved phase cycles in closely-related species, suggesting that this feedback regulation might be advantageous in the phasiRNA pathway. Furthermore, *cis* feedback regulation might be easily maintained during phasiRNA evolution. As bottom-strand phasiRNAs are always complementary to the *PHAS* precursors with no mismatch, no matter how the *PHAS* sequence diversified, there might be new *cis*-acting sites occurred randomly. Although we describe *cis*-activity of phasiRNAs, their *trans* functions may well be equally or more important, given that there is a tremendously diverse set of phasiRNAs active at reproductive stages that could have innumerable targets in *trans*. However, no matter whether they are *cis*- or *trans*-acting, the phasiRNA sequences appear almost random in their variation, and are both species-specific and poorly conserved.

It is unclear how top-strand 21-nt phasiRNAs could direct the cleavage of RDR6-derived bottom-strand RNA, as these are theoretically well-annealed in double-stranded RNA, while AGO-loaded siRNAs typically cleave single-stranded RNA. One possibility is that both strands of these *PHAS* region are transcribed, i.e. there are NATs for the *PHAS* precursors (just like miR2118 and its NATs), and the top-strand 21-nt phasiRNAs might target the NATs, resulting in a 12 nt shift of cleavage site relative to the miR2118 target site. With respect to previously published lncRNA results in *O. sativa* (Zhang *et al.*, 2014), we found 18 lncRNAs overlapping the 21-*PHAS* loci investigated in our study, and 12 of them were on the opposite strand of the 21-*PHAS* precursor, suggesting some *PHAS* NATs exist and are potential *cis*-targets of top strand 21-nt phasiRNAs.

We observed a potentially complex reciprocal regulation mechanism between miR2118 and its NATs, and the derived 21-nt phasiRNAs have the potential to target miR2118 precursors. However, it is unknown how and when these NATs arose during grass evolution. In addition, how miR2118 is activated in the reproductive stage, and how to maintain the balance of miR2118, NATs and derived 21-nt phasiRNAs are important questions, the answers to which might increase our understanding of the function of reproductive phasiRNAs. In the study of PSMS, it was still undetermined how the SNPs in the second phasiRNA cycle lead to the phenotypic changes (i.e. male sterility). However, the two studies have indicated the functional importance of sequences close to the miR2118 target sites (Ding *et al.*, 2012; Fan *et al.*, 2016). Moreover, these observations plus the trigger miRNA target-sites biased conservation pattern prompted us to consider whether editing the miR2118/2275 target sites of other potential *PHAS* loci, or *LDMAR* and *PMS1T* orthologs in other rice varieties, might produce additional valuable mutant materials, which would facilitate rice breeding efforts.

**Acknowledgements**

This work in the Chen lab is supported by the National Natural Science Foundation of China (NSFC award no. 31571309 and no. 31771409) and the State Key Laboratory of Plant Genomics. Work in the Meyers lab is supported by US NSF award no. 1754097. We thank three anonymous referees and the handling editor for very helpful comments. We also thank Joanna Friesner for useful comments and edits to the manuscript. The authors declare no conflicts of interest.

**Author contributions**

M.C., B.C.M. and P.T. designed the experiments. P.T., X.Z., Y.L., M.W., B.L., T.L. and J.S conducted the experiments, P.T., X.Z., R.X., M.C. and B.C.M. analyzed the data. R.A.W. provided *Oryza* genome assemblies. P.T., M.C. and B.C.M. wrote the article. R.X. and B.L. edited the article. All authors read and approved the final manuscript.

**References**


**Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, Ulat Victor J, Chebotarov D, Zhang G, Li Z, et al. 2015.** SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res* **43**: D1023-D1027.

**Araki S, Le NT, Koizumi K, Villar-Briones A, Nonomura K-I, Endo M, Inoue H, Saze H, Komiya R. 2020.** miR2118-dependent U-rich phasiRNA production in rice anther wall development. *Nat Commun* **11**: 3115.

**Axtell MJ. 2013.** Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol* **64**: 137-159.

**Axtell MJ, Jan C, Rajagopalan R, Bartel DP. 2006.** A two-hit trigger for siRNA biogenesis in plants. *Cell* **127**: 565-577.

**Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009.** MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

**Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, Dunoyer P, Yamamoto YY, Sieburth L, Voinnet O. 2008.** Widespread translational inhibition by plant miRNAs and siRNAs. *Science* **320**: 1185-1190.

**Castel SE, Martienssen RA. 2013.** RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat Rev Genet* **14**: 100-112.

**Chen J, Huang Q, Gao D, Wang J, Lang Y, Liu T, Li B, Bai Z, Luis Goicoechea J, Liang C, et al. 2013.** Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nat Commun* **4**: 1595.

**Cheng F, Wu J, Fang L, Wang X. 2012.** Syntenic gene analysis between *Brassica rapa* and other *Brassicaceae* species. *Front Plant Sci* **3**: 198.

**Christie M, Croft LJ, Carroll BJ. 2011.** Intron splicing suppresses RNA silencing in Arabidopsis. *Plant J* **68**: 159-167.

**Deng P, Liu S, Nie X, Weining S, Wu L. 2018.** Conservation analysis of long non-coding RNAs in plants. *SCI CHINA LIFE SCI* **61**: 190-198.

**Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, Xu C, Li X, Xiao J, Zhang Q. 2012.** A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proc Natl Acad Sci USA* **109**: 2654-2659.

**Dukowic-Schulze S, Sundararajan A, Ramaraj T, Kianian S, Pawlowski WP, Mudge J, Chen CB. 2016.** Novel meiotic miRNAs and indications for a role of phasiRNAs in meiosis. *Front Plant Sci* **7**: 762.

**Edgar RC. 2004.** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

**Fahlgren N, Carrington JC. 2010.** miRNA Target Prediction in Plants. *Methods Mol Biol* **592**: 51-57.

**Fan Y, Yang J, Mathioni SM, Yu J, Shen J, Yang X, Wang L, Zhang Q, Cai Z, Xu C, et al. 2016.** *PMS1T*, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. *Proc Natl Acad Sci USA* **113**: 15144-15149.

**Fang X, Qi Y. 2016.** RNAi in plants: an Argonaute-centered view. *Plant Cell* **28**: 272-285.

**Fei Q, Xia R, Meyers BC. 2013.** Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell* **25**: 2400-2415.

**Fei Q, Yang L, Liang W, Zhang D, Meyers BC. 2016.** Dynamic changes of small RNAs in rice spikelet development reveal specialized reproductive phasiRNA pathways. *J Exp Bot* **67**: 6037-6049.

**German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, Kannan V, Rymarquis LA, Nobuta K,**

**German R, et al. 2008.** Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol* **26**: 941-946.

**Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, Kasschau KD, Carrington JC. 2007.** Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell* **19**: 926-942.

**Huang J, Li Z, Zhao D. 2016.** Deregulation of the OsmiR160 Target Gene *OsARF18* Causes Growth and Developmental Defects with an Alteration of Auxin Signaling in Rice. *Sci Rep* **6**: 29938.

**Johnson C, Kasprzewska A, Tennessen K, Fernandes J, Nan GL, Walbot V, Sundaresan V, Vance V, Bowman LH. 2009.** Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome Res* **19**: 1429-1440.

**Kakrana A, Mathioni SM, Huang K, Hammond R, Vandivier L, Patel P, Arikit S, Shevchenko O, Harkess AE, Kingham B, et al. 2018.** Plant 24-nt reproductive phasiRNAs from intramolecular duplex mRNAs in diverse monocots. *Genome Res* **28**: 1333-1344.

**Komiya R. 2017.** Biogenesis of diverse plant phasiRNAs involves an miRNA-trigger and Dicer-processing. *J Plant Res* **130**: 17-23.

**Komiya R, Ohyanagi H, Niihama M, Watanabe T, Nakano M, Kurata N, Nonomura K. 2014.** Rice germline-specific Argonaute MEL1 protein binds to phasiRNAs generated from more than 700 lincRNAs. *Plant J* **78**: 385-397.

**Langmead B, Trapnell C, Pop M, Salzberg SL. 2009.** Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

**Li S, Le B, Ma X, Li S, You C, Yu Y, Zhang B, Liu L, Gao L, Shi T, et al. 2016.** Biogenesis of phased siRNAs on membrane-bound polysomes in Arabidopsis. *eLife* **5**: e22750.

**Li X, Zhu J, Hu F, Ge S, Ye M, Xiang H, Zhang G, Zheng X, Zhang H, Zhang S, et al. 2012.** Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene

expression. *BMC Genomics* **13**: 300.

**Liu B, Chen Z, Song X, Liu C, Cui X, Zhao X, Fang J, Xu W, Zhang H, Wang X, et al. 2007.** *Oryza sativa* dicer-like4 reveals a key role for small interfering RNA silencing in plant development. *Plant Cell* **19**: 2705-2718.

**Luo S, Zhang Y, Hu Q, Chen J, Li K, Lu C, Liu H, Wang W, Kuang H. 2012.** Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. *Plant Physiol* **159**: 197-210.

**Ma ZR, Coruh C, Axtell MJ. 2010.** *Arabidopsis lyrata* Small RNAs: Transient *MIRNA* and Small Interfering RNA Loci within the *Arabidopsis* Genus. *Plant Cell* **22**: 1090-1103.

**Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, et al. 2008.** Sorting of small RNAs into *Arabidopsis* argonaute complexes is directed by the 5' terminal nucleotide. *Cell* **133**: 116-127.

**Mohammadin S, Edger PP, Pires JC, Schranz ME. 2015.** Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. *BMC Plant Biol* **15**: 1-12.

**Nonomura K, Morohoshi A, Nakano M, Eiguchi M, Miyao A, Hirochika H, Kurata N. 2007.** A germ cell specific gene of the ARGONAUTE family is essential for the progression of premeiotic mitosis and meiosis during sporogenesis in rice. *Plant Cell* **19**: 2583-2594.

**Patel P, Mathioni S, Kakrana A, Shatkay H, Meyers BC. 2018.** Reproductive phasiRNAs in grasses are compositionally distinct from other classes of small RNAs. *New Phytol* **220**: 851-864.

**Rajagopalan R, Vaucheret H, Trejo J, Bartel DP. 2006.** A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev* **20**: 3407-3425.

**Song X, Li P, Zhai J, Zhou M, Ma L, Liu B, Jeong DH, Nakano M, Cao S, Liu C, et al. 2012a.** Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. *Plant J* **69**: 462-474.

**Song X, Wang D, Ma L, Chen Z, Li P, Cui X, Liu C, Cao S, Chu C, Tao Y, et al. 2012b.** Rice RNA-dependent RNA polymerase 6 acts in small RNA biogenesis and spikelet development. *Plant J* **71**: 378-389.

**Sosa-Valencia G, Palomar M, Covarrubias AA, Reyes JL. 2017.** The legume miR1514a modulates a NAC transcription factor transcript to trigger phasiRNA formation in response to drought. *J Exp Bot* **68**: 2013-2026.

**Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, et al. 2018.** Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat Genet* **50**: 285-296.

**Tamim S, Cai Z, Mathioni SM, Zhai J, Teng C, Zhang Q, Meyers BC. 2018.** *Cis*‐directed cleavage and nonstoichiometric abundances of 21‐nucleotide reproductive phased small interfering RNAs in grasses. *New Phytol* **220**: 865-877.

**Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crete P. 2004.** Endogenous *trans*-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. *Mol Cell* **16**: 69-79.

**Wang H, Niu Q-W, Wu H-W, Liu J, Ye J, Yu N, Chua N-H. 2015.** Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits. *Plant J* **84**: 404-416.

**Xia R, Meyers BC, Liu Z, Beers EP, Ye S, Liu Z. 2013.** MicroRNA superfamilies descended from miR390 and their roles in secondary small interfering RNA biogenesis in Eudicots. *Plant Cell* **25**: 1555-1572.

**Xia R, Ye S, Liu Z, Meyers BC, Liu Z. 2015.** Novel and recently evolved microRNA clusters regulate expansive *F-BOX* gene networks through phased small interfering RNAs in wild diploid strawberry. *Plant Physiol* **169**: 594-610.

**Xie K, Shen J, Hou X, Yao J, Li X, Xiao J, Xiong L. 2012.** Gradual increase of miR156 regulates temporal expression changes of numerous genes during leaf development in rice. *Plant Physiol* **158**: 1382-1394.

**Yang Z. 2007.** PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591.

**Zemach A, McDaniel IE, Silva P, Zilberman D. 2010.** Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**: 916-919.

**Zhai J, Jeong DH, De Paoli E, Park S, Rosen BD, Li Y, Gonzalez AJ, Yan Z, Kitto SL, Grusak MA, et al. 2011.** MicroRNAs as master regulators of the plant *NB-LRR* defense gene family via the production of phased, *trans*-acting siRNAs. *Genes Dev* **25**: 2540-2553.

**Zhai J, Zhang H, Arikit S, Huang K, Nan GL, Walbot V, Meyers BC. 2015.** Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc Natl Acad Sci USA* **112**: 3146-3151.

**Zhang Y, Xia R, Kuang H, Meyers BC. 2016.** The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them. *Mol Biol Evol* **33**: 2692-2705.

**Zhang YC, Liao JY, Li ZY, Yu Y, Zhang JP, Li QF, Qu LH, Shu WS, Chen YQ. 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**: 512.

**Zhao YP, Xu ZH, Mo QC, Zou C, Li WX, Xu YB, Xie CX. 2013.** Combined small RNA and degradome sequencing reveals novel miRNAs and their targets in response to low nitrate availability in maize. *Ann Bot* **112**: 633-642.

**Supporting Information**

**Fig. S1** *PHAS* loci were mostly in the intergenic regions and 21-*PHAS* have high GC composition than 24-*PHAS*.

**Fig. S2** miR2118 could target *NLRs* and produce no or sparse phasiRNAs in *O. sativa*.

**Fig. S3** Most 21-*PHAS* and 24-*PHAS* were unique or with low copies, and examples showing 21-*PHAS* loci expanded by segment duplication.

**Fig. S4** The syntenic distribution property of *PHAS* on the 11 chromosomes of five *Oryza* genomes.

**Fig. S5** Both *pms1* and *pms3* are 21-*PHAS* loci in *O. sativa*.

**Fig. S6** 21-nt phasiRNA-targets were less stable than miRNA-targets.

**Fig. S7** 21-nt phasiRNAs with 5As with 5with 5ble than miRN*cis*-cleavage in the four *Oryza* species.

**Fig. S8** The sequences of two 21-*PHAS* loci with *cis*-cleavage sites.

**Fig. S9** The expression level of two potential miR2118 NATs in different tissues.

**Fig. S10** DNA methylation level of *PHAS* loci was higher in reproductive yong panicles than vegetative leaves.

**Table S1** Summary of reads quantities and genome mapping qualities of sRNA and degradome data for five *Oryza* species.

**Table S2** The lists of *PHAS* loci found in the five *Oryza* genomes.

**Table S3** miR2118 and miR2275 members in the five *Oryza* genomes.

**Table S4** The lists of 21-nt phasiRNAs targets identified by degradome analysis in the five *Oryza* species.

**Table S5** The 553 functional 21-nt phasiRNA-targets found in *O. sativa*, and their orthologous 21-nt phasiRNAs and targets in the other *Oryza* species.

**Table S6** The number of total, or subset of the 540 functional 21-nt phasiRNAs, with homology between *O. sativa* and the other four *Oryza* species.

## FIGURE LEGENDS

**Fig. 1. *PHAS* loci identified from the sRNAs expressed at reproductive stage in five *Oryza* species.**

(A) The number of *PHAS* loci identified in five *Oryza* species. (B) Size distribution of total sRNAs in each sRNA library, which showed that 21- and 24-nt sRNAs were mostly abundant at this stage. (C) Size distribution of unique sRNAs in each sRNA library, which showed that 24-nt have more varieties. (D) The ratio of 21, 22 and 24 nt sRNAs in different genomic regions. “sRNA” include tRNA, rRNA and snoRNA here. “other” represents unannotated region. (E) The histograms show the ratio of 21-nt sRNAs in the defined phase cycles (In *PHAS*) or outside phase cycles (Out of *PHAS*) at 21-*PHAS* region. (F) The histograms show the ratio of 24-nt sRNAs in the defined phase cycles (In *PHAS*) or outside phase cycles (Out of *PHAS*) at 24-*PHAS* region. Species abbreviations used in the manuscript: *O. sativa* (OS), *O. rufipogon* (OR), *O. nivara* (ON), *O. barthii* (OB), *O. glaberrima* (OGB), *O. glumaepatula* (OGL), *O. meridionalis* (OM), *O. punctata* (OP), *O. brachyantha* (FF).

**Fig. 2. *PHAS* loci were expanded by local tandem duplication.**

(A) The ratio of different duplication types according to their position. “adj”: the two duplicates were adjacent or less

than 1 Mb apart on the same chromosome; "dis": the distance of the two duplicates was larger than 1 Mb on the same chromosome; "inter": interchromosomal duplication. (B) A schematic graph showing 21-*PHAS* supercluster in *Oryza sativa* on chromosome 4, ranged from 20,991,722 bp to 21,237,202 bp, in which the *PHAS* copies induced by tandem duplications (green curves) have consistent *PHAS* direction. (C) A schematic graph showing one tandem duplication event that led to the expansion of three adjacent *PHAS* loci, as highlighted by the purple frame in panel B.

**Fig. 3. *PHAS* loci show position but not sequence conservation among *Oryza* species.**

(A) The percentage of *O. sativa* 21-*PHAS* and 24-*PHAS* loci with homology in the other four *Oryza* genomes. (B) The percentage of 21-*PHAS* and 24-*PHAS* in the other four *Oryza* genomes with homology in the *O. sativa* genome. Types are: "homolog" are homologous sequences found outside the syntenic region; "segment" are homologous sequences found in the syntenic region with a length less than 1/3 of the query sequence; "ortholog" are homologous sequences in a syntenic region; and "specific" are sequences with no homology in the *O. sativa* genome. (C) The ratios of 21-*PHAS*, 24-*PHAS*, flanking regions of *PHAS*, CDS, UTR, intron of genes in *O. sativa* genomes that with orthologs in the other four *Oryza* species. The results show that both *PHAS* loci and their flanking regions have low sequence conservation than other genomic regions. Flanking regions involve " 5′ " (500 bp before the trigger site) and " 3′ " (500 bp after the terminal site of *PHAS*). Genes in *O. sativa* are selected from the adjacent regions of *PHAS* with distance less than 50 kb. (D) The distribution of 21-*PHAS* and 24-*PHAS* loci on chromosome 6 of the five *Oryza* species, coded as described in Fig. 1. The position of *PHAS* is highlighted. The red arrow points to the 21-*PHAS* clusters in Fig. 3E. (E) Detailed presentation of 21-*PHAS* clusters in the syntenic blocks of five *Oryza* genomes, with the two terminal border genes as LOC_Os06g09390 and LOC_Os06g09540 in *O. sativa*. The lines or bands between species indicate sequence homology. The blue lines show an inversion that changed the strand distribution of 10 *PHAS* loci in *O. punctata*.

**Fig. 4. *PHAS* loci experience high mutation with biased selection at trigger miRNA target sites.**

(A) Boxplots showing nucleotide substitution rates of 21-*PHAS*, flanking regions, miR2118 target sites (T), introns and non-*PHAS* lincRNAs. (B) Boxplots showing nucleotide substitution rates of 24-*PHAS*, flanking regions, miR2275 target sites (T), introns and non-*PHAS* lincRNAs. The horizontal lines in the boxes are the median value, and the two

bars were the 25th and 75th percentiles, respectively. The nucleotide substitution rate was calculated based on pairwise comparison of *Oryza sativa* and the other *Oryza* species from AA group. "a" above the bars indicates K values were significantly higher ($P < 0.01$) than that of intron_a. "b" above the bars indicates K values were significantly lower ($P < 0.01$) than that of intron_a. "intron_a" indicates an intron selected from the protein-coding genes adjacent or inside *PHAS* superclusters; "intron_r" indicates an intron from randomly selected protein-coding genes. Flanking regions involve " 5′ " (500 bp before the target site) and " 3′ " (500 bp after the terminal site of *PHAS* locus). non-*PHAS* lincRNAs were selected from the published intergenic lncRNAs (Zhang *et al.*, 2014) that have no overlap with our *PHAS* loci in *O. sativa*. (C) The three types of average variation densities at *PHAS* loci and its flanking regions. The red arrowheads point to the miRNA target sites. SNPs and indels were only analyzed in the *O. sativa* genome. Transposon data is from the repeat annotation of the five *Oryza* genomes. Only 21-*PHAS* or 24-*PHAS* regions with definite precursor direction and sequence length ≥ 200 bp were included. For each *PHAS*, up to 1 kb of sequences flanking each side were interrogated. Here, an overlapping sliding window of 5% of the sequence length, at a step of 2.5% of the sequence length, was used in both flanking sides and the *PHAS* body. The mean value at same window position was used to depict the distribution pattern.

**Fig. 5. 21-nt phasiRNAs with 5′-terminal U tend to induce *cis* cleavage.**

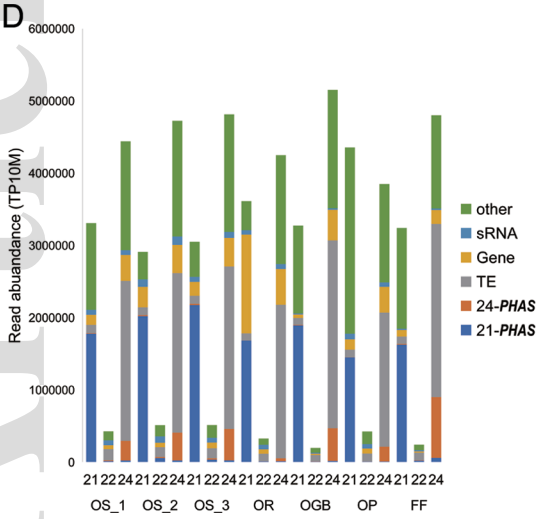
(A) Nucleotide composition of 21-nt phasiRNAs with potential cleavage-directing functions, which shows that the 5′-terminal nucleotides were mostly uridine (U). (B) The number of possible target sites in different categories. "*PHAS*_top" is the 21-nt phasiRNA precursor derived from RNA Pol Ⅱ. "*PHAS*_bottom" is consistent to the RDR6 derived bottom strand RNA relative to the phasiRNA precursor. "*PHAS*_other" mean one 21-nt phasiRNA target transcripts from another *PHAS* locus in *trans*. (C) Relative ratios of functional 21-nt phasiRNAs in different phase cycles, which shows that functional 21-nt phasiRNAs tend to be present in the cycles close to the miR2118 target site, except for the first cycle (C1). (D) The *cis*-cleavage of 21-PHAS768 precursor at C6, which was consistent in *Oryza sativa*, *O. glaberrima* and *O. punctata*. (E) Multiple *cis*-acting phasiRNAs were found in 21-PHAS1660, in which *cis*-cleavage at C4, C8 and C9 were found in the three AA genomes, but not in *O. punctata*.

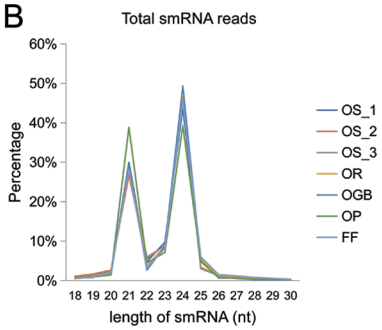
**Fig. 6. miR2118 triggers the processing of 21-nt phasiRNAs from its natural antisense transcripts**.

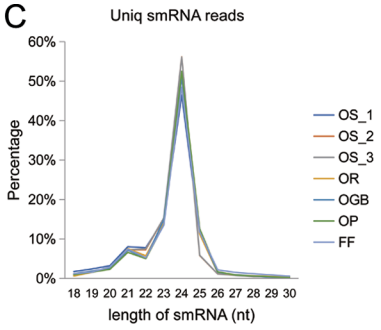
(A) The distribution of miR2118 clusters and antisense 21-*PHAS* loci in the syntenic regions at Chr 4 and Chr 11 in the five *Oryza* genomes. The green rectangle indicates tandem duplication in *O. punctata* at Chr 4. (B) The detailed distribution of genetic elements at the miR2118 cluster on *O. sativa* Chr 4. miR2118 clusters are encoded on the same strand, while the natural antisense transcripts AK068680 and at least six 21-*PHAS* loci are on the opposite strand. (C) The antisense sites of mature miR2118 are the target sites of 21-*PHAS* loci, as highlighted by the light-blue box in panel B. (D) *Cis*-cleavage at the antisense site of mature miR2118 triggers the production of 21-nt phasiRNAs. The red arrows at the sRNA track point to the 22-nt mature miR2118 sequence. The blue arrows at the degradome track point to the *cis*-cleavage reads at the antisense strand of mature miR2118.

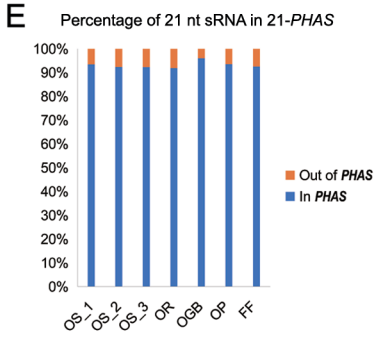
A

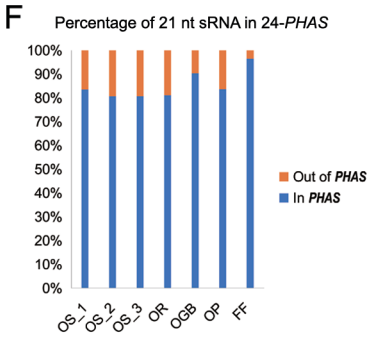
| Species | 21-*PHAS* | | 24-*PHAS* | |
|---|---|---|---|---|
| | Total | miR2118 triggered | Total | miR2275 triggered |
| *O. sativa* | 2,338 | 2,115 | 172 | 108 |
| *O. rufipogon* | 1,618 | 1,381 | 88 | 61 |
| *O. glaberrima* | 1,799 | 1,721 | 157 | 129 |
| *O. punctata* | 1,473 | 1,386 | 117 | 102 |
| *O. brachyantha* | 988 | 937 | 92 | 80 |



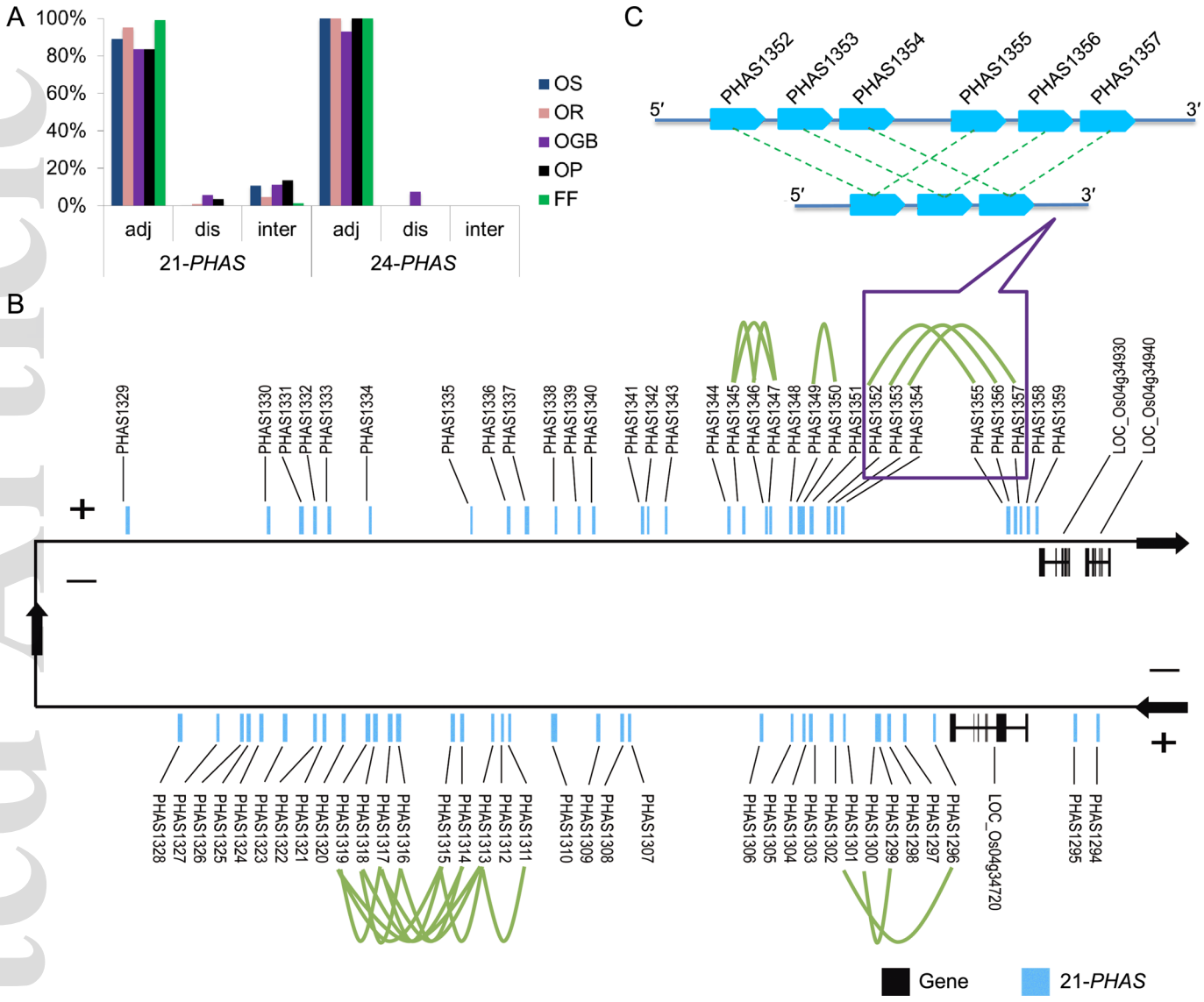



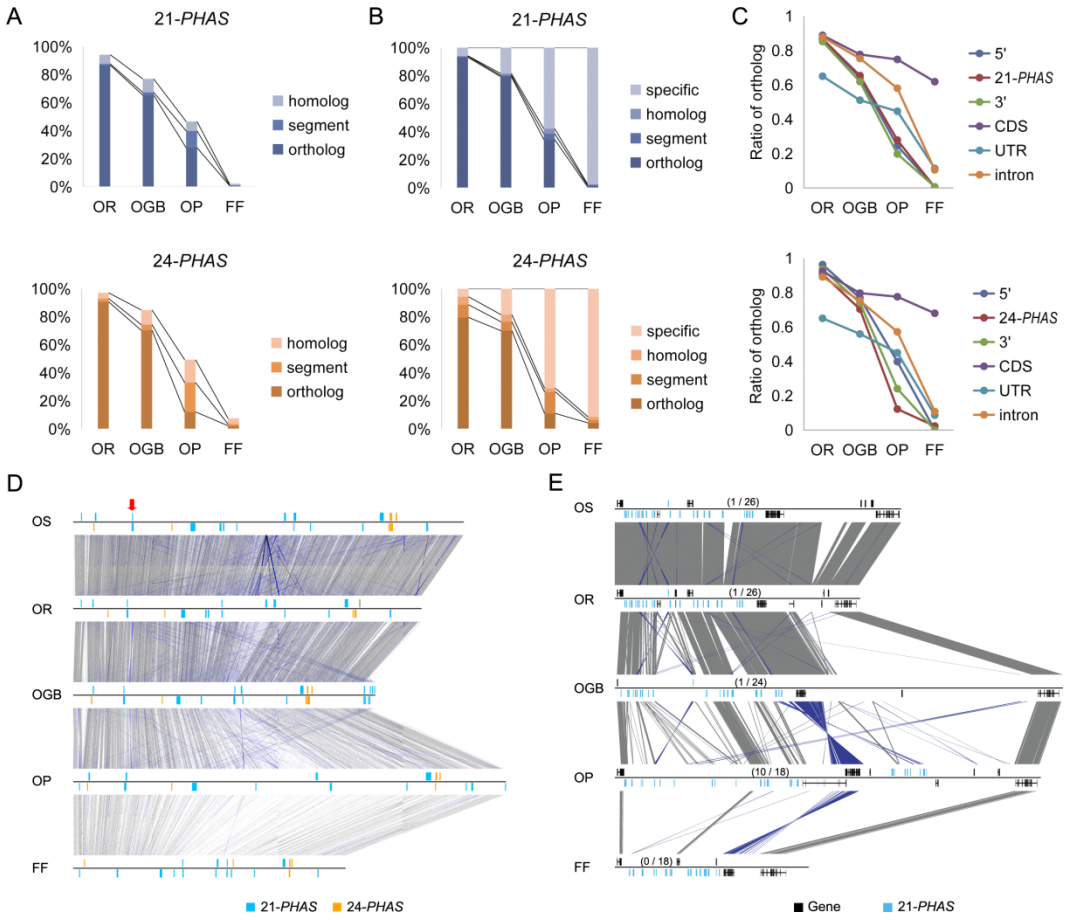



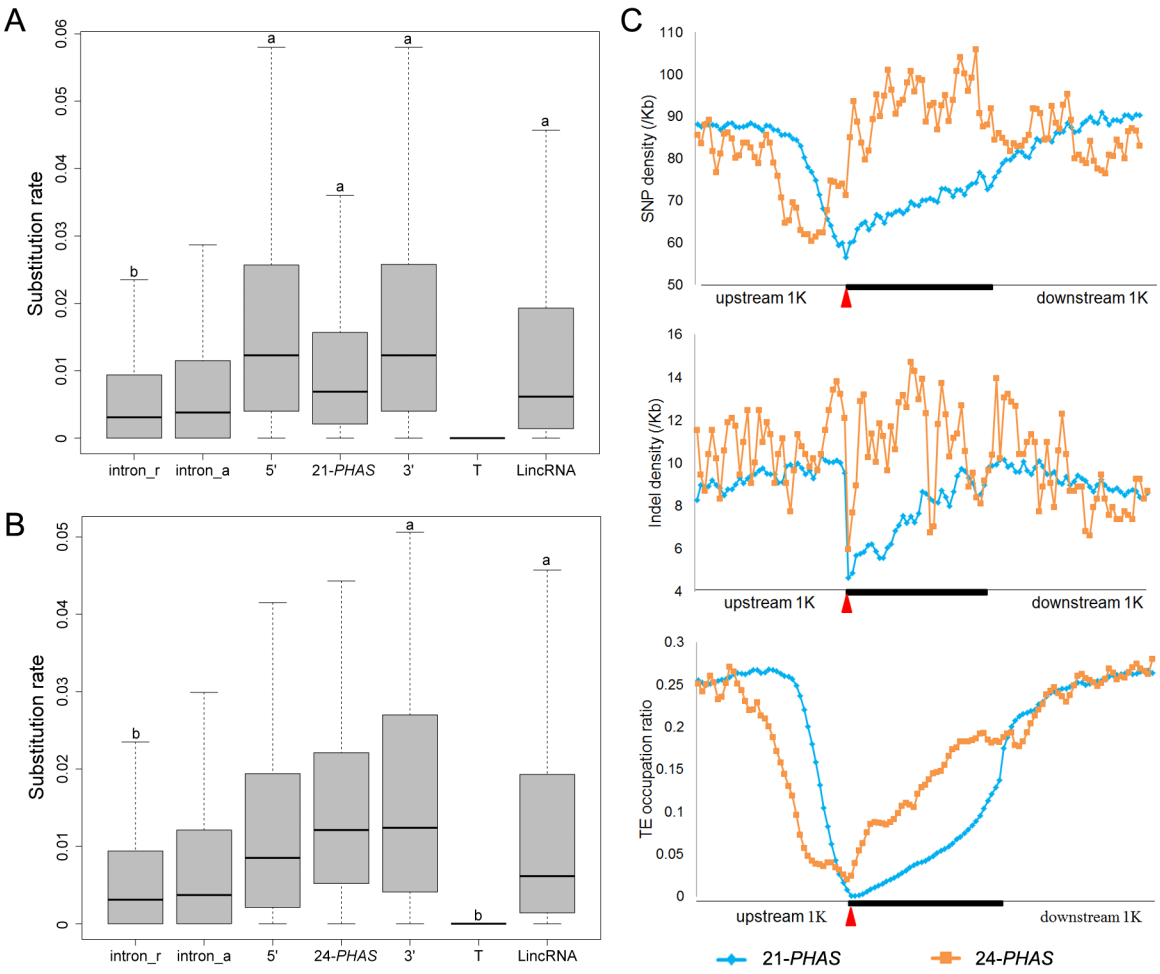

nph_17035_f1.tif



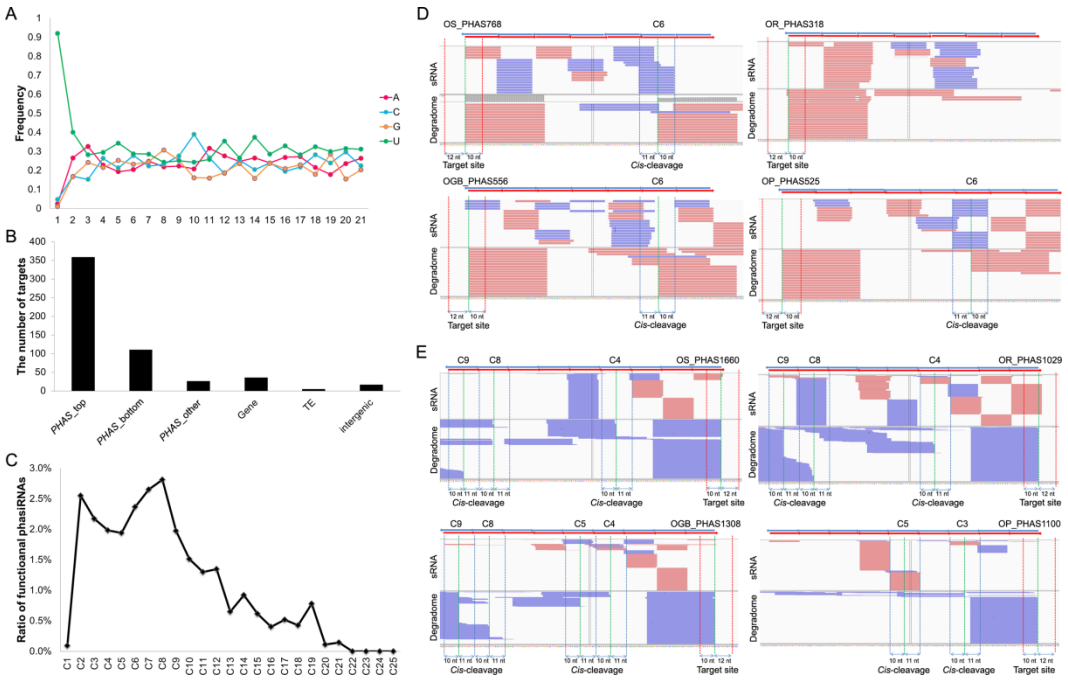

nph_17035_f2.tif



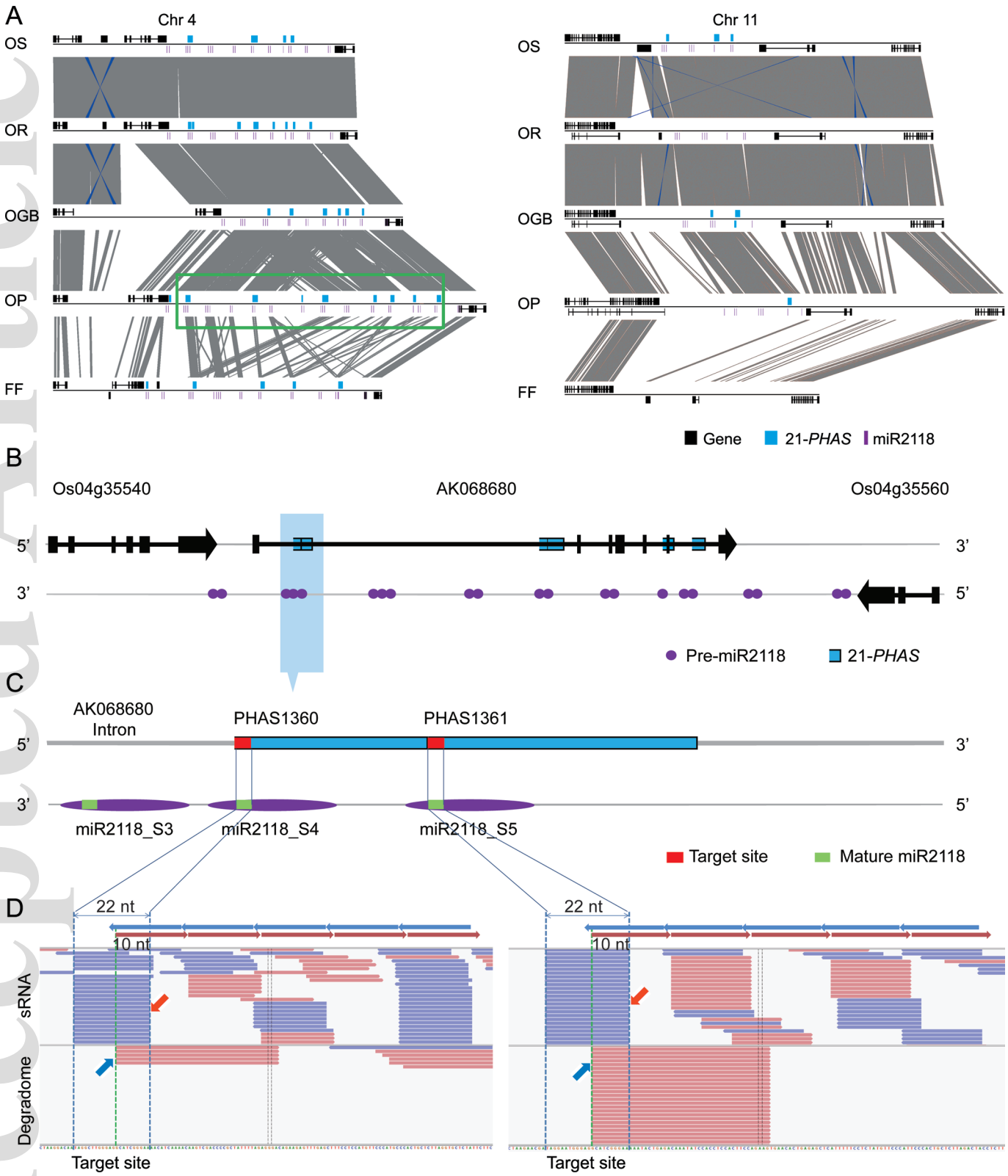

nph_17035_f3.tif




nph_17035_f4.tif




nph_17035_f5.tif




nph_17035_f6.tif