the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **84,** 404–416 doi: 10.1111/tpj.13018

Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits

Huan Wang[1], Qi-Wen Niu[1], Hui-Wen Wu[1,†], Jun Liu[1,‡], Jian Ye[2,§], Niu Yu[1] and Nam-Hai Chua[1,*]

[1]*Laboratory of Plant Molecular Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA, and*

[2]*Temasek Life Sciences Laboratory, 1 Research Link, National University of Singapore, Singapore City 117604, Singapore*

Received 21 April 2015; accepted 26 August 2015; published online 8 September 2015.

*For correspondence (e-mail chua@mail.rockefeller.edu).

†Present address: Temasek Life Sciences Laboratory, 1 Research Link, National University of Singapore, Singapore City 117604, Singapore.

‡Present address: National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China.

§Present address: State Key Laboratory of Plant Genomics, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China.

Accession numbers: GSE56459, GSE56460, GSE56462, GSE56463.

SUMMARY

Long non-coding RNAs (lncRNAs) have recently been found to widely exist in eukaryotes and play important roles in key biological processes. To extend our knowledge of lncRNAs in crop plants we performed both non-directional and strand-specific RNA-sequencing experiments to profile non-coding transcriptomes of various rice and maize organs at different developmental stages. Analysis of more than 3 billion reads identified 22 334 long intergenic non-coding RNAs (lincRNAs) and 6673 pairs of sense and natural antisense transcript (NAT). Many lincRNA genes were associated with epigenetic marks. Expression of rice lincRNA genes was significantly correlated with that of nearby protein-coding genes. A set of NAT genes also showed expression correlation with their sense genes. More than 200 rice lincRNA genes had homologous non-coding sequences in the maize genome. Much more lincRNA and NAT genes were derived from conserved genomic regions between the two cereals presenting positional conservation. Protein-coding genes flanking or having a sense-antisense relationship to these conserved lncRNA genes were mainly involved in development and stress responses, suggesting that the associated lncRNAs might have similar functions. Integrating previous genome-wide association studies (GWAS), we found that hundreds of lincRNAs contain trait-associated SNPs (single nucleotide polymorphisms [SNPs]) suggesting their putative contributions to developmental and agriculture traits.

Keywords: *Oryza sativa* L. ssp. *Japonica* cultivar Nipponbare, *Zea mays* L. ssp. *mays*, lincRNA, NAT, conservation, RNA-seq.

INTRODUCTION

RNA molecules that are not translated into proteins are referred as non-coding transcripts. Non-coding RNAs have recently emerged as important regulators of gene expression and have attracted increasing attention. This group of RNAs can be classified by size: (1) small non-coding RNAs, including microRNAs (miRNAs), small interfering RNAs (siRNAs), piwi-interacting RNAs (piRNAs), *trans*-acting siRNAs (ta-siRNAs) and natural antisense transcript siRNAs (NAT-siRNAs); and (2) long non-coding RNAs (lncRNAs) which are usually more than 200 nucleotides in length (Chen, 2009; Rinn and Chang, 2012). According to the positional relationship of their encoding genes to nearby protein-coding genes, lncRNAs can be further grouped into long intergenic non-coding RNAs (lincRNAs), natural antisense transcripts (NATs) and intronic RNAs (incRNAs) (Ma *et al.*, 2013).

In the past decade, thousands of lncRNAs have been identified in several eukaryotes, mainly yeast, model animal species and human. Compared with these eukaryotes, genome-wide identification of lncRNAs in plants is more recent and not as comprehensive (Ulitsky and Bartel, 2013; Zhang *et al.*, 2014). By integrating directional tiling array and RNA-sequencing (RNA-seq) data we have recently identified around 6480 lincRNAs and 37 238 NAT pairs in Arabidopsis (Liu *et al.*, 2012; Wang *et al.*, 2014). Recent analysis of maize EST and RNA-seq datasets has uncovered

many lncRNAs (Boerner and McGinnis, 2012; Li *et al.*, 2014). As these maize lncRNAs were assembled from non-directional transcriptome data and mainly with relative short sequences (<50 nt), the characterization of lncRNAs and especially for NATs, was limited with respect to transcription directions and transcript boundaries. Moreover, owing to their association with transcription activities, histone modification marks have been used to define genomic loci encoding lncRNAs in animals (Guttman *et al.*, 2009; Ulitsky *et al.*, 2011). A plant lincRNA, APOLO, was recently reported to play a role in the formation of a chromatin loop in response to auxin (Ariel *et al.*, 2014). Hence, information relating to epigenetic modification could also be used in the characterization and re-annotation of genes for non-coding transcripts in rice and maize.

Conservation analysis can advance our knowledge of lncRNAs in several aspects: (1) identification of functional sequence elements and structures; (2) potential biological functions; and (3) clues for their mode of action (Ulitsky and Bartel, 2013). Animal lincRNA sequences evolve very rapidly and are poorly conserved, and the functional sequence of lincRNAs is usually short and difficult to detect by current alignment methods. A good example is the *Miat/Gomafu/Rncr2* lincRNA which is involved in specifying cell identity in the nervous system (Sone *et al.*, 2007; Rapicavoli *et al.*, 2010). In vertebrates, all *Miat* homologs contain a short region with multiple copies of the (U) ACUAAC(C) motif which could not be uncovered by BLAST (Rapicavoli *et al.*, 2010; Tsuiji *et al.*, 2011). Conversely, some lincRNAs have conserved exon-intron structures or are located in conserved genomic regions although without any detectable sequence conservation. Likewise, X-inactive specific transcript (*Xist*) which regulates X chromosome inactivation is only conserved in its exon-intron structure (Brockdorff *et al.*, 1992; Brown *et al.*, 1992; Penny *et al.*, 1996; Nesterova *et al.*, 2001). Also, a functional zebrafish lincRNA gene is embedded in a conserved genomic locus and its nearby protein-coding gene has orthologs in both human and mouse (Ulitsky *et al.*, 2011). Because of the limited number of studies on rice and maize lncRNAs, it is still an open question to what extend rice and maize lncRNAs are conserved at the sequence or genomic position level. Clearly, a more comprehensive lncRNA catalog from these two cereal crops is needed to address these issues.

Genetic and molecular analyses have shown that an increasing number of lncRNAs are involved in the regulation of gene expression and chromatin structure. LncRNAs can regulate their targets in *cis* or *trans* through various mechanisms (Faghihi and Wahlestedt, 2009; Ulitsky and Bartel, 2013). However, the key issue of whether lncRNAs serve specific biological functions or are simply transcriptional by-products is still controversial, because most of them do not have a recognizable function (Katayama *et al.*, 2005; Schultes *et al.*, 2005; Struhl, 2007). Genome-wide association studies (GWAS) have been successful in unravelling the genetic basis of trait variation and in identifying causal loci linked to phenotypic diversity in plants (Buckler *et al.*, 2009; McMullen *et al.*, 2009; Atwell *et al.*, 2010; Brachi *et al.*, 2010; Huang *et al.*, 2010; Nemri *et al.*, 2010; Famoso *et al.*, 2011; Kump *et al.*, 2011; Poland *et al.*, 2011; Tian *et al.*, 2011; Zhao *et al.*, 2011). Moreover, single–nucleotide polymorphism (SNP) typing allowed GWAS to identify small haplotype blocks that are correlated with trait variation (Brachi *et al.*, 2011). Based on current GWAS results in animals and plants, <20% of these causal loci are assigned to coding regions (Hindorff *et al.*, 2009; Huang *et al.*, 2010; Kumar *et al.*, 2012). Those trait-associated SNPs that are embedded in intergenic regions serve as a rich resource to address potential functions of lncRNAs. Genetic variations associated with diseases have been shown to impact lincRNA expression in human (Kumar *et al.*, 2013). Also, a rice SNP has been found to regulate the expression of the corresponding lncRNA gene resulting in photoperiod-sensitive male sterility (Ding *et al.*, 2012). These examples suggest that association analysis of rice and maize lncRNAs and GWAS data may not only provide clues for potential functions of lncRNAs but also identify candidate non-coding genes important for GWAS. Understanding the genetic basis of trait variation is also critical to improving the yield and quality of these two cereal crops.

In this study, we performed both directional and non-directional high-throughput RNA-seq experiments to investigate non-coding transcriptome in various organs at different developmental stages of two important cereal crops, rice and maize. Here, we compiled a comprehensive list of more than 22 000 lincRNAs and 6673 NATs. Systematic evolutionary analysis revealed that lncRNAs displayed more positional conservation than sequence conservation. Integrating with published GWAS datasets, our results highlighted potential contributions of lncRNAs to the specification of agriculture traits of rice and maize.

RESULTS

Identification of long non-coding RNAs in rice and maize

To uncover non-coding transcripts in rice and maize, we first collected eight rice (*Oryza sativa* L. ssp. *Japonica* cultivar Nipponbare) samples of various organs at different developmental stages, including flower buds, flowers, flag leaves and roots, sampled before and after flowering, milk grains and mature seeds (Figure 1a). Two maize (*Zea mays* L. ssp. *mays*) organs, shoots and roots, were also collected. We then performed RNA-seq experiments on polyadenylated RNAs to investigate rice and maize transcriptomes. We first carried out non-directional paired-end RNA-seq (peRNA-seq) and collected 2.5 billion mate read

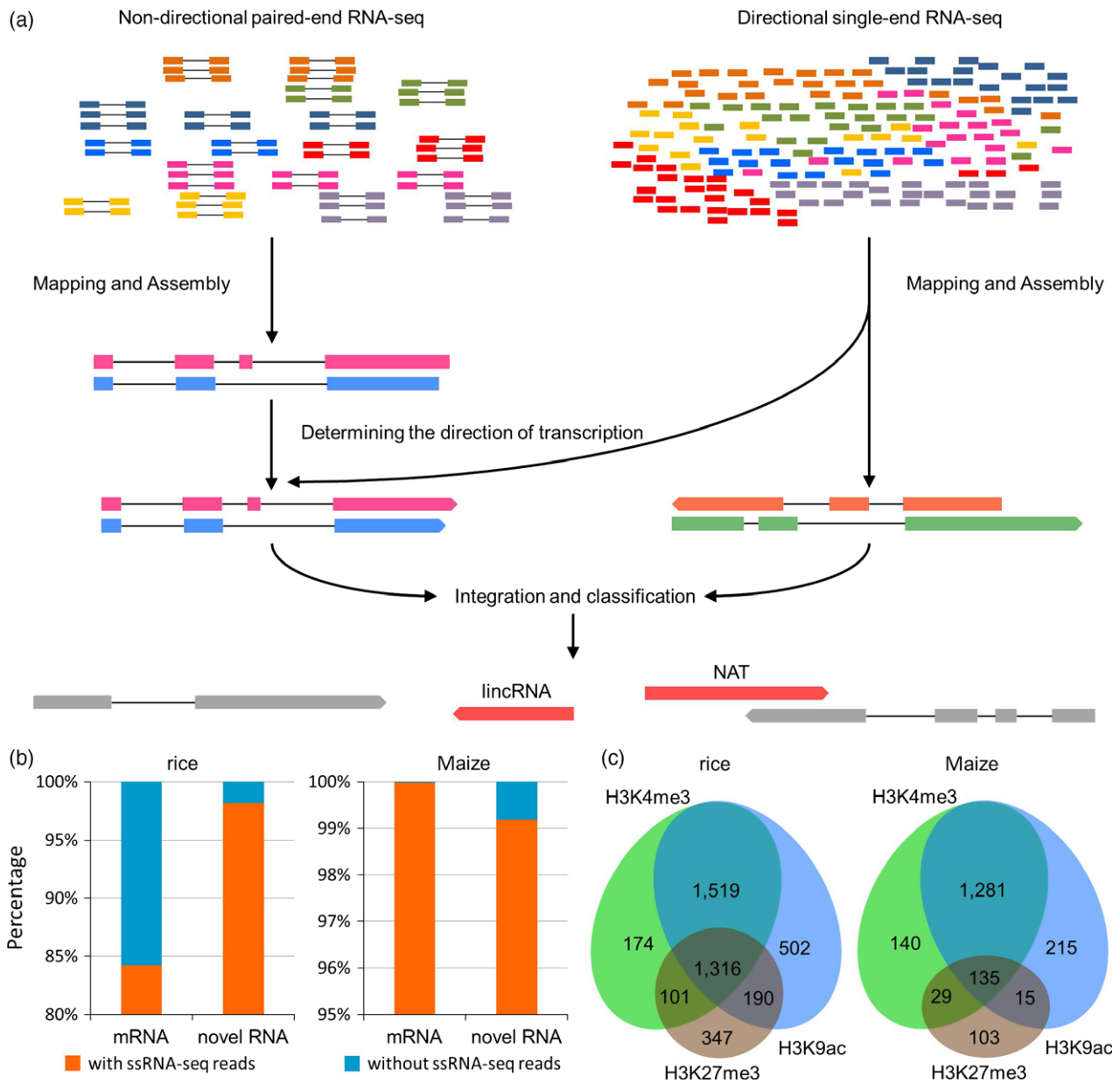

Figure 1. Identification of lncRNA genes in rice and maize and their general features.
(a) Pipeline for the identification of lncRNA genes in rice and maize genomes.
(b) Proportion of peRNA-seq assemblies aligned with or without ssRNA-seq reads.
(c) Venn diagrams of three kinds of histone modifications associated with lincRNA genes.

pairs (Table S1). In total, we identified 67 692 and 39 085 unique transcripts from peRNA-seq data in rice and maize, respectively. On average, 34.6% of these transcripts were not identical to annotated gene sequences, including protein-coding genes, pseudogenes, ribosomal RNA, tRNA, miRNA, and other known classes of ncRNAs, and these were referred to as unclassified transcripts.

Although non-directional peRNA-seq provides an accurate estimate of isoform abundance, it is not useful in defining transcription directions (Katz *et al.*, 2010). For example, it is difficult to resolve NAT from the sense transcript. Therefore, we employed strand-specific single-end RNA-seq (ssRNA-seq) protocols to investigate transcriptomes from the same set of samples and obtained more than 677 million reads (Table S1). First, we aligned ssRNA-seq reads to unclassified transcripts assembled from the peRNA-seq data set to determine their potential transcription directions. More than 99% of peRNA-seq-assembled transcripts were supported by ssRNA-seq reads and transcription direction of around 91% of them was defined

(Figures 1b and S1). Second, we assembled ssRNA-seq data independently and identified 16 090 rice and 13 211 maize directional transcripts. Third, all transcripts identified from peRNA-seq and ssRNA-seq data were integrated based on their genomic positions and transcription directions. In total, we obtained 27 065 and 22 814 unclassified transcripts with defined transcription direction as well as 577 and 709 non-directional ones in rice and maize, respectively. These transcripts were then further filtered and classified into two groups: long intergenic non-coding RNAs (lincRNAs) and long NATs.

LincRNAs in rice and maize

To identify lincRNAs, short transcripts (<200 nt) and transcripts overlapping with either strand of annotated transcripts were first removed from our transcript dataset; subsequently, their coding potentials were evaluated. Because lincRNAs are defined as transcripts with an open reading frame (ORF) region of <100 amino acids (Dinger *et al.*, 2008) and *bona fide* protein-coding genes are likely to have sequence similarities to entries in protein databases, we excluded transcripts (1) encoding more than 100 amino acids as screened by GENSCAN (Burge and Karlin, 1998); or (2) having amino acid sequence similarity to protein sequences deposited in National Center for Biotechnology Information (NCBI) nr database (E-value ≤0.001). Applying our criteria we retained 11 229 and 11 105 lincRNAs in rice and maize, respectively. The lincRNA loci were named using *lnc* gene classifier, i.e. LNC_OsNNgXXXXXX or LNC_ZmNNgXXXXXX. 'LNC' stands for long non-coding gene and 'Os' and 'Zm' refer to rice and maize, respectively. 'NN' gives the chromosome number and 'XXXXXX' is the *lnc* gene identifier.

LincRNA genes ranged in length from 200 base pair (bp) to around 30 000 bp with a mean length of 676 and 741 bp in rice and maize, respectively. Rice lincRNAs were close to flanking genes with an average distance of 871 bp, and maize lincRNAs were on average 6761-bp away from their flanking genes (Figure S2a,b). Comparing the transcription directions of lincRNAs and neighboring coding genes, we found that lincRNAs could be transcribed from the same or the opposite strand of neighboring protein-coding genes (Figure S1). Because the uneven expansion of cereal chromosome was mainly driven by transposon insertion (Bruggmann *et al.*, 2006) we examined the gene density of genomic regions encoding lincRNAs. We found that lincRNAs were embedded in genomic regions with similar gene density to annotated protein-coding genes, suggesting association between lincRNA genes and protein-coding genes (Figure S2c,d).

Many mammalian lincRNAs are associated with chromatin-modifying complexes and their expression levels could be linked to histone modification marks. For example, trimethylated histone H3 lysine 4 (H3K4me3), which is a positive mark for transcription initiation regions, was used in combination with expression data to assemble lincRNA collections (Guttman *et al.*, 2009; Khalil *et al.*, 2009). We collected published ChIP-seq data profiling three kinds of histone marks in rice and maize, including two positive marks, H3K4me3 and acetylated histone three lysine 9 (H3K9ac), and one negative mark, trimethylated histone H3 lysine 27 (H3K27me3). To generate these data, He *et al.* (He *et al.*, 2010) used shoots of four-leaf stage seedlings of rice (*Oryza sativa* L. ssp. *japonica* cultivar Nipponbare) and Wang *et al.* (Wang *et al.*, 2009) used shoots and roots of 14-day-old maize (*Zea mays* L. ssp. *mays*) B73 seedlings. The organs used in their studies were represented in our samples used for lincRNA identification. We therefore re-analyzed their data and compared their histone modification marks with loci encoding lincRNAs by genomic position. In total, around 37% rice and 17% maize lincRNA loci were associated with histone modification peaks in selected samples. We further investigated the histone mark association with protein-coding genes and transposon elements. We found the proportion of histone mark-associated lincRNA genes is lower than that of coding genes, but significantly higher than that of transposon elements (hypergeometric test, *P*-value <0.01; see details about lincRNA gene association with each kind of histone marks in Table S2). In mammals, genomic regions modified with H3K4me3 and H3K27me3, referred to as bivalent marks, are generally associated with developmental regulator genes in embryonic stem cells (Bernstein *et al.*, 2006; Mikkelsen *et al.*, 2007). Here, we found that 16–34% lincRNA genes were associated with positive marks, 27.7% rice and 14.3% maize lincRNA genes were associated with H3K4me3 mark and 31.4% rice and 14.8% maize lincRNA genes were associated with H3K9ac mark. On the other hand, 17.4% rice and 2.5% maize lincRNA genes carried the negative mark, H3K27me3. The positive marks were enriched in the up-stream and transcribed regions of lincRNA genes whereas the negative mark was mainly associated with transcribed region of lincRNA genes (Figure S3a,b). More than 1700 lincRNA genes contained both positive and negative marks (Figure 1c). We further examined expression levels of histone-mark-associated lincRNA genes in rice leaves sampled before flowering and in maize shoots. We found that positive marks (H3K4me3 and H3K9ac) were enriched in highly expressed lincRNA genes and H3K27me3 mark was associated with lincRNA genes of low expression levels (Figure S3c,d). Enrichment of histone marks was significantly correlated with lincRNA gene expression level (Mann–Whitney *U*-test, rice *P*-value = 1.524e-08, maize *P*-value = 0.001182).

Genomic properties of NATs

We defined NAT as a long RNA (≥200 nt) transcribed from the opposite strand of an annotated gene with at least a

50-nt overlapping sequence. We identified NATs from unclassified transcripts and focused on NAT pairs composed of one annotated mRNA and one non-coding RNA. We defined the annotated mRNA as the sense transcript and the non-coding RNA as the antisense transcript. In total, we found 4681 rice NAT pairs derived from 4455 unclassified genes and 3827 annotated genes. Using the same method, 1992 maize NAT pairs were identified. Around one-third of the antisense RNAs were completely covered by the sense transcripts and another one-third of the NAT pairs were transcriptionally divergent in which sense and antisense transcripts overlapped at the 5′-end (Figure S4). Gene ontology (GO) enrichment analysis showed that sense genes preferentially encoded important proteins, such as transcription factors, involved in regulating developmental transition and responses to stresses (Table S3). This result suggests potential biological function of the corresponding antisense RNAs.

Specific expression of lncRNAs and expression correlation with neighboring protein-coding genes

Mammalian lincRNAs are usually expressed at specific developmental stages or tissues with an organ preference towards brains and testis (Ravasi *et al.*, 2006; Cabili *et al.*, 2011; Ulitsky *et al.*, 2011; Pauli *et al.*, 2012). This organ-specific expression is also observed in the model dicot plant, *Arabidopsis* (Liu *et al.*, 2012). To see whether it is also true for the lincRNAs identified here, we compared expression levels of each transcript among all sequenced samples and found lincRNAs displayed more expression variation than mRNA in rice ($P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test, Figure 2a). Similar results were seen in maize (Figure S5). We found that the largest number of rice lincRNAs showed expression peaks in mature seeds followed by flower buds (Figure 2b). Using K-means clustering, we identified several groups of organ- and/or developmental stage-specific lincRNAs (Figure S6). For example, we uncovered 152 rice lincRNAs preferentially expressed at high levels in mature seeds but barely detectable in other samples. Several hundred organ-specific rice lincRNAs were identified in flowers, leaves or roots. We also found some developmental stage-specific lincRNAs, e.g. 445 lincRNAs expressed only in roots sampled before flowering but not in roots collected after flowering or in any other organs (Figure 2c). We further examined several differentially expressed lincRNAs by quantitative real-time PCR (qRT-PCR) experiments and confirmed their expression specificity using a different technical platform (Figures 2d and S7).

To investigate potential function of lincRNA in *cis,* i.e. regulating expression of nearby genes, we examined possible correlation of lincRNA expression with expression of its closest flanking protein-coding gene and found there is a correlation on a genome-wide scale. We found an expression correlation between lincRNA and its closest flanking coding gene (Pearson's product-moment correlation = 0.32) (Figure 2e). Whereas the expression correlation between protein-coding genes and their neighboring protein-coding genes was much lower. As negative controls, neither the expression of lincRNA genes and that of randomly selected genes nor the expression of genes and that of randomly selected genes showed such correlation. Moreover, we observed a strong correlation in expression between hundreds of lincRNA genes and their flanking genes (Pearson's product-moment correlation for positively correlated lincRNA and flanking gene = 0.80, *P*-value <0.001; Pearson's product-moment correlation for negatively correlated lincRNA and flanking gene = 0.78, *P*-value <0.001 Figure S8) suggesting their co-expression or *cis*-regulation.

NAT may regulate the expression of its sense transcript in either a positive or a negative way (Lavorgna *et al.*, 2004); the expression levels of sense and antisense transcripts could be positively correlated (concordant NAT pairs) or negatively correlated (discordant NAT pairs). We scanned differentially expressed (>=2-fold) NAT pairs between any two samples and found 62.0% rice and 6.3% maize NAT pairs to be organ-specific and showed either positively correlated or negatively correlated expression of sense and antisense RNAs (Figure S9). However, with respect to positively correlated NAT pairs there is the concern that some of them might be transcription by-products of sense RNAs. To be on the conservative side, we considered the 1702 rice and 41 maize negatively correlated NATs more likely to be independent transcripts and may have regulatory functions. Around 60% of negatively correlated NAT pairs in rice showed substantial expression changes between mature seeds and other organs, suggesting a potential function of negatively correlated NAT pairs in rice seeds (Figure 2f).

LncRNAs show more positional conservation than sequence conservation

To investigate possible sequence conservation of lincRNAs, we carried out whole-genome alignment between rice and maize. Around 20% of rice lincRNAs (2281 out of 11 229 lincRNAs) showed detectable sequence conservation to the maize genomic sequence; however, only 5% of them (117 out of 2281 rice lincRNA) had sequence similarity to our maize lincRNAs. Because lincRNAs are usually expressed at low levels with enrichment in specific organs, they could be uniquely detected in different experiments. To provide a more comprehensive view of lincRNA conservation in the two cereals, we integrated previously reported maize lincRNAs in our evolutionary analysis (Li *et al.*, 2014). In total, we found 264 rice lincRNAs displaying sequence conservation with maize lincRNAs and more than half of rice lincRNAs (1177 out of 2281 lincRNAs,

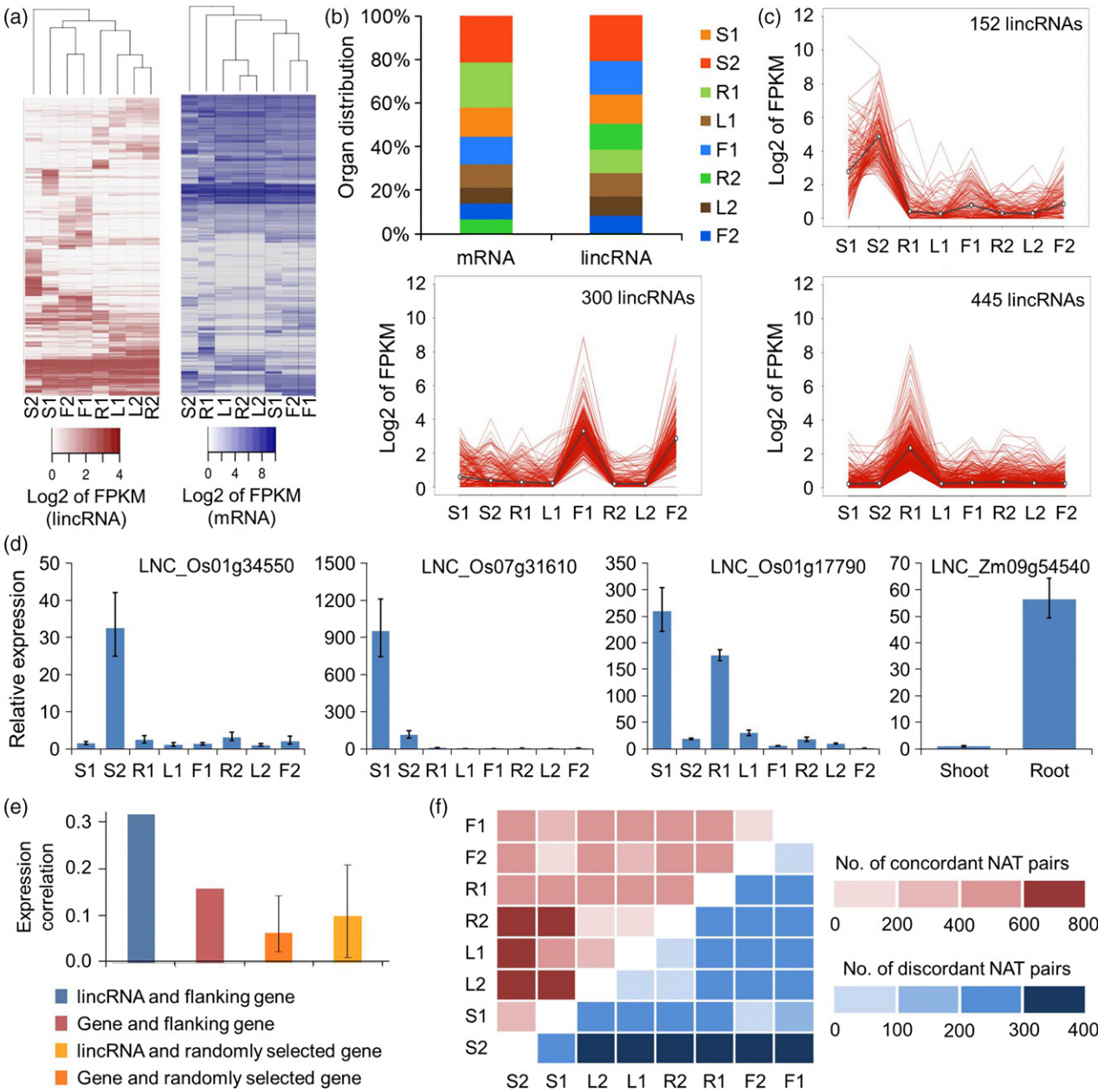

Figure 2. Temporal-spatial expression of lincRNAs and NATs in rice and maize.
(a) Heat maps of rice lincRNA and mRNA expression levels (log2 of FPKM value).
(b) Organ distribution of expression peaks of rice mRNAs and lincRNAs.
(c) Three groups of organ- and/or developmental stage-specific rice lincRNAs. *Y*-axis gives the log2 value of FPKMs for lincRNAs. The total number of lincRNAs in each group is also given.
(d) Validation of differentially expressed lincRNAs by qRT-PCR. *Y*-axis gives the relative expression levels. Error bars represent standard errors ($n = 3$).
(e) Expression correlation of rice lincRNAs and flanking protein-coding genes. *Y*-axis gives Pearson product-moment correlation coefficient. Error bars represent standard deviations ($n = 3$).
(f) Number of rice positively correlated and negatively correlated NAT pairs identified from comparison between any two samples. S1, milk grains. S2, mature seeds. R1, roots sampled before flowering. L1, leaves sampled before flowering. F1, flower buds. R2, roots sampled after flowering. L2, leaves sampled after flowering. F2, flowers.

51.6%) showing sequence conservation to maize mRNAs (Figure 3a). On the other hand, 19.0% of maize lincRNAs (2110 out of 11 105 lincRNAs) contained sequences conserved in the rice genome. Also 5% of these 2110 maize lincRNAs had lincRNA homologs in rice and a much larger percentage of them (1874 out of 2110 maize lincRNAs,

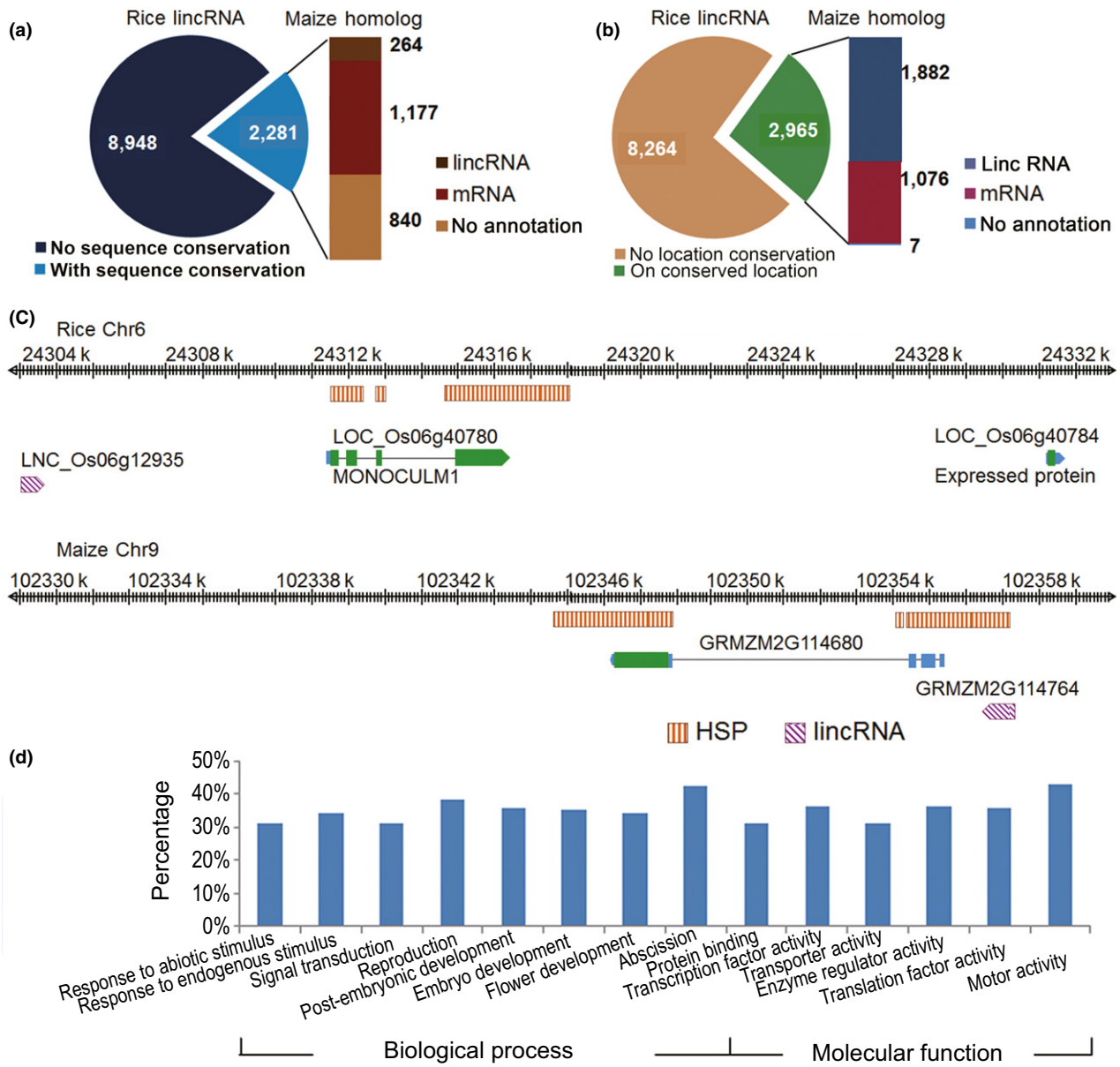

Figure 3. Sequence and positional conservation of rice lncRNAs.
(a) Sequence conservation of rice lincRNAs based on results of whole-genome alignment.
(b) Positional conservation of rice lincRNA genes based on analysis of synteny blocks.
(c) An example of a positionally conserved rice lincRNA. Gene structures are shown in solid color. lincRNA gene and high scoring pair (HSP) of synteny block are marked by diagonals and vertical lines.
(d) Functional enrichment of flanking genes of positionally conserved rice lincRNAs.

88.8%) displayed sequence similarity to mRNAs (Figure S10a).

As some lincRNAs are embedded in conserved genomic locations bioinformatics tools designed to search for sequence homology often failed to detect such lincRNA orthologs (Ulitsky and Bartel, 2013). To further uncover lincRNAs located in conserved genomic regions but with limited sequence conservation, we performed synteny search and found many more lincRNAs with positional conservation. Around 26.4% of rice (2965 out of 11 229) and 23.3% of maize lincRNAs (2589 out of 11 105) were embedded in synteny blocks (Figures 3b and S10b). Sharing positional conservation, 1882 rice lincRNAs could be traced to detectable maize lincRNAs identified here or by Li *et al.*,2014;. The number of lincRNAs with positional conservation was around seven times that the number of sequence conserved lincRNAs. Results of a chi-squared test showed that the number of lincRNA genes with

positional conservation was significantly greater than that of lincRNA genes containing homologous sequences (*P*-value <0.001). For example, we found a 30-kb synteny region containing 1 lincRNA gene and one protein-coding gene, *MONOCULM 1* (*MOC1*) gene, which is the first identified key regulator of rice tiller number (Li *et al.*, 2003) (Figure 3c). As a control, we performed similar analysis of the sequence and positional conservation of all rice and maize protein-coding genes. We found that the number of protein-coding genes sharing conserved sequences is almost two times that of protein-coding genes with positional conservation in both rice and maize, Moreover, the extent of sequence and positional conservation among lincRNA genes and protein-coding genes are significantly different (chi-squared test, *P*-value <0.001).

We found the synteny blocks carrying lincRNA homologs were usually enriched for genes encoding transcription factors; these genes are involved in several important biological processes, such as responses to stress and developmental processes (Figure 3d and Table S4). Moreover, 45.2% of rice and 48.1% of maize lincRNA genes with positional conservation were differentially expressed between leaves and roots (fold change ≥2). Around 40% of the differentially expressed lincRNA homologs showed similar organ preference in rice and maize. This conserved expression preference of lincRNA genes with positional conservation suggests that they may play conserved functions in similar biological process.

We further investigated positional conservation of NATs in the two cereals. Here, we defined a NAT pair as having position conservation if the corresponding sense gene has a homolog in the other genome and the homolog also has a NAT identified from RNA-seq data. In total, we identified 166 and 198 NAT pairs with positional conservation in rice and maize, respectively. Function enrichment analysis revealed that sense genes involved in these NAT pairs with positional conservation are preferentially involved in the regulation of developmental processes and stress responses (Figure S11). In some cases, the positional conservation of NAT pairs was even extended to the dicot *Arabidopsis.* For example, we found the NATs of miR156, miR159, miR167 and miR399 genes were detectable in both rice and *Arabidopsis.* Also, previously unidentified NATs were found derived from the opposite strand of miR169 gene in both maize and *Arabidopsis.* Moreover, expression of some NATs with positional conservation was negatively correlated with the corresponding miRNA gene expression.

Potential functions of lincRNAs in determining morphological, developmental and agronomic traits

To explore potential functions of lincRNAs in rice and maize, we collected trait-associated SNPs identified by recent GWAS in the two cereals and aligned them to genomic loci encoding lincRNAs. Zhao *et al.* (2011) evaluated the molecular basis of 34 morphological, developmental and agronomic traits and identified 234 related SNPs (Zhao *et al.*, 2011). As their work was based on an old assembled version of the rice genome (MSU.v6), we first aligned our lincRNA genes to the genome sequences used in their study and then selected the best match of each lincRNA gene for comparison with trait-associated SNPs. We uncovered five rice lincRNA genes carrying trait-associated SNPs related to leaf and seed morphology and yield components (Figure 4a). We further randomly selected a group of protein-coding genes ($n = 6000$) and examined their association with trait-associated SNPs. Compared to protein-coding genes, we did not find rice lincRNA genes to be significantly associated with trait-associated SNPs (hypergeometric test, *P*-value = 0.08). The expression of one of these lincRNA was validated by qRT-PCR experiments (Figure 4b). In maize, Chia and collaborators reported more than 6000 SNPs to be associated with leaf development traits, including leaf angle, leaf width, leaf length, Southern leaf blight and northern leaf blight (Chia *et al.*, 2012). We found 951 maize lincRNA genes with SNPs associated with leaf development traits. Compared with all protein-coding genes, we found that trait-associated SNPs were significant enriched in genomic loci encoding maize lincRNAs (hypergeometric test, *P*-value = 1.8e-214).

Moreover, we checked the expression level of maize SNP-containing lincRNAs and their neighboring protein-coding genes in the intermated B73 × Mo17 recombinant inbred line population. Li *et al.* conducted RNA-seq experiments on the shoot apices of 2-week old maize seedlings from the inbred lines B73 and Mo17, and 105 recombinant inbred lines (Li *et al.*, 2013). Analysis of their RNA-seq data showed that expression levels of 15.0% of SNP-associated lincRNAs and those of their neighboring genes are significantly correlated (Pearson correlation coefficient *P*-value <0.001).

Two rice lincRNA genes linked to trait-associated SNP were transcribed from conserved genomic loci; one was LNC_Os03 g44325 associated with seed color-related SNP (Figure 4c) and the other one was LNC_Os05 g27795 associated with leaf pubescence-related SNP. Moreover, we also found maize homologs of LNC_Os05 g27795 in the corresponding synteny blocks and these maize lincRNA genes also carried trait-associated SNPs, leaf angle-related SNPs (Figure 4d). Our results revealed that several lincRNA genes with positional conservation carry leaf-trait-associated SNPs indicating conserved functions of non-coding transcripts. LNC_Os05 g27795 was mainly expressed in vegetative organs before flowering and two maize lincRNAs, LNC_Zm08 g18080 and LNC_Zm08 g18085, specifically accumulated in shoots (Figure 4e). The potential functions of these conserved SNP-associated lincRNAs await further investigation.

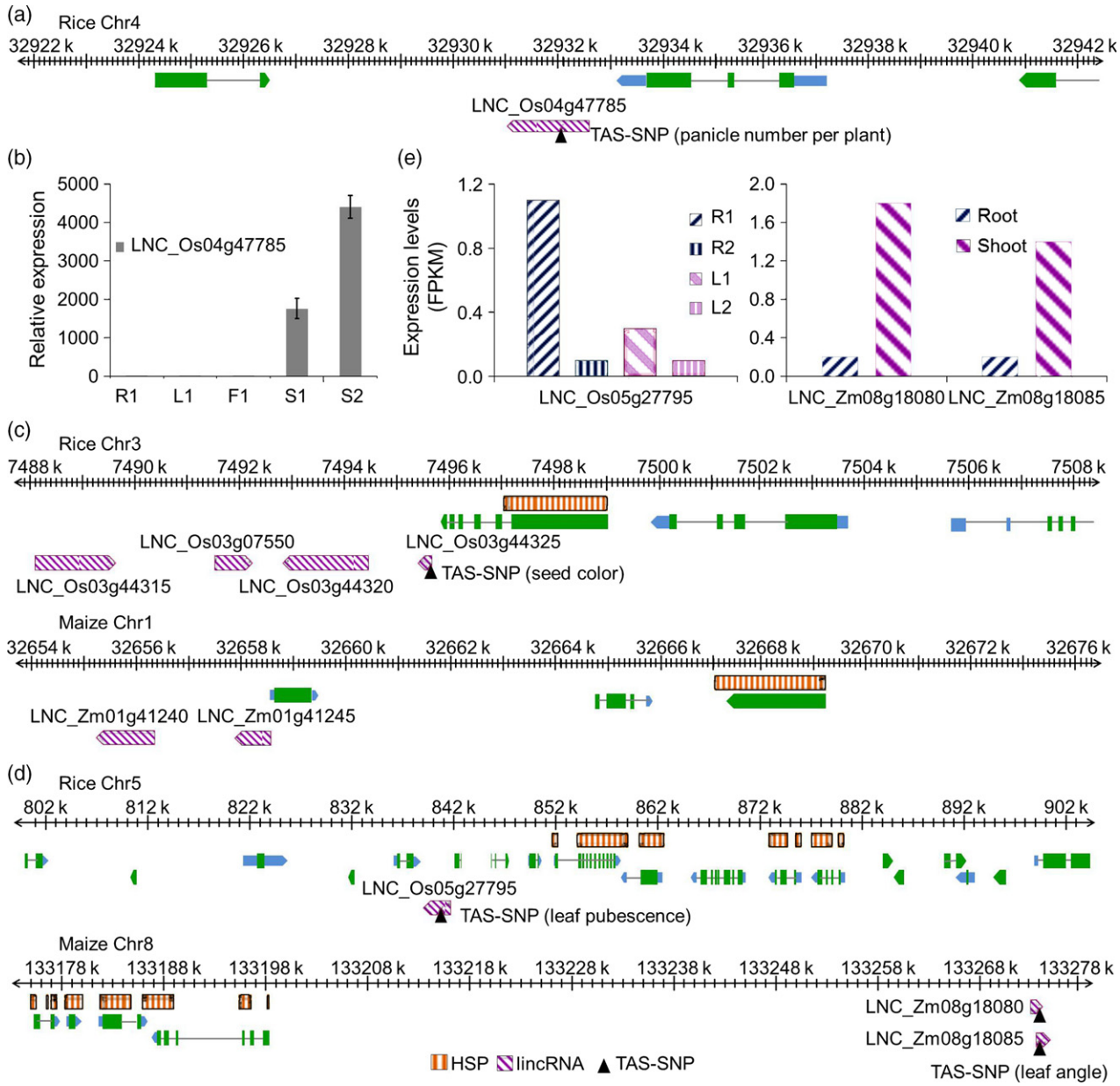

Figure 4. LincRNA genes containing agriculture trait-associated SNPs.
(a) A 20-kb flanking region of a rice lincRNA gene carrying a casual locus related to panicle number per plant.
(b) Validation of a rice lincRNA gene containing a trait-associated SNP by qRT-PCR. Error bars give standard errors, $n = 3$.
(c, d) Conserved lincRNAs associated with agronomic traits.
(e) Organ-specific expression of conserved lincRNAs with potential functions. *Y*-axis gives FPKM values of lincRNAs. Agronomic trait-associated SNPs are indicated by triangles. Gene structures are shown in solid color. lincRNA locus and HSP of synteny block are marked by diagonals and vertical lines.

DISCUSSION

To investigate the coverage of the non-directional paired-end and the directional single-end RNA-seq methods, we independently assembled and annotated transcripts from peRNA-seq and ssRNA-seq reads. We found that the majority of assembled transcripts was uncovered by the two methods; more than 90% rice and 78% maize transcripts were reproducibly detected by both methods (Figure S1). The difference between the percentage of commonly detected transcripts in rice and maize can be mainly attributed to the sequencing depth. For example, we first pooled peRNA-seq reads from all eight rice samples and assembled transcripts. Then, we merged ssRNA-seq reads from two, four, six, and eight rice samples and assembled the transcripts independently. We compared transcripts obtained from 8-sample peRNA-seq and those

from 2/4/6/8-sample ssRNA-seq reads. We found that the number of commonly detected transcripts increased with the number of ssRNA-seq reads used in transcript assembly.

To remove transcripts generated by possible transcription noise, we used an arbitrary criterion to exclude products of basal transcription before identification of lncRNAs. We required that the expression level of each transcript to be no <1 fragments per kilobase of exon per million fragments mapped (FPKM) in at least one sample. Applying this criterion, we were able to detect 51% rice and 69% maize annotated protein-coding transcripts. Since not all annotated protein-coding transcripts were detected by our experiments, it is likely that our list of reported lncRNAs is incomplete. Considering the organ-specific expression of lncRNAs in previous reports and in this work, we believed additional lncRNAs would be uncovered using specific plant organs or plants subjected to various stresses.

Li *et al.* (2014) assembled maize lncRNAs using public EST database, maize genome annotation and RNA-seq datasets from 30 different experiments (Li *et al.*, 2014). Around 33% of their high-confidence lncRNAs were also found in our maize lncRNA collection. The large number of lincRNAs unique to each study raises the issue of reliability of these putative lincRNAs. However, lncRNAs that are identified by such methods have unknown transcription direction. Because the sensitivity and coverage of different experiments are not the same, it is not surprising that there are differences in the population of detected lincRNAs. Moreover, Li and colleagues found that a large number of lncRNAs could be smRNA precursors by searching homologous sequences of smRNAs on lncRNAs. We also aligned our maize lincRNAs to the smRNAs reported by Wang *et al.*(Wang *et al.*, 2009), and followed the cutoff used by Boerner and McGinnis (Boerner and McGinnis, 2012). We found that 14.9% (1657 out of 11 105) of them carrying smRNAs. However, we could not determine whether the smRNA-associated lincRNAs are true smRNA precursors. Even lincRNAs without significant amount of associated smRNAs could still be smRNA precursors, because current maize smRNA dataset may be not comprehensive enough. Thus, we did not exclude smRNA-containing lincRNAs in our study.

In the identification of NATs, we note that the proportion of negatively correlated NAT pairs is much smaller in maize than in rice; this is mainly due to the fewer number of maize organs used in our comparative analysis. In rice, we compared expression levels of sense and antisense transcripts in any two out of eight organs. But, only two maize organs were used for the same analysis, which greatly reduced the possibility to observe negatively correlated expression. We expect more negatively correlated maize NAT pairs when more organ samples are included in the comparison.

Consistent with results of evolutionary analysis in vertebrate lincRNAs, we found little sequence conservation between rice and maize lincRNAs suggesting a rapid evolution of lncRNA sequence. As our sequence alignment was focused on large genomic regions we were unable to exclude the possibility of smaller conserved regions in lncRNAs. It will be interesting to search for short conserved elements from lncRNAs in future analysis.

Systematic synteny search showed that a large number of lncRNAs is embedded in conserved genomic regions. The positional conservation of lncRNA genes is much higher than their sequence conservation. We should emphasize that lncRNA genes with positional conservation were so defined only if they were found in both rice and maize synteny regions. For a large number of synteny regions, we found detectable lncRNA genes in only one species but not in the other and this type of lncRNA genes was not considered as having positional conservation. Due to the low abundance and organ specificity of lncRNAs, current identification of rice and maize lncRNAs is clearly not yet saturated. It is likely that the synteny regions might encode additional lncRNAs not detected by this study since we did not sample all the possible tissues/stages from both species. Accordingly, we proposed that the number of lncRNAs with positional conservation is still under estimated.

We found that a large proportion of lincRNAs could be traced to mRNAs in rice and maize, which is also observed in the analysis of zebrafish lincRNAs (Ulitsky *et al.*, 2011). Possible explanations for this phenomenon include lncRNA gene annotation errors and conversion between coding and non-coding RNA genes. There are several examples to illustrate this type of possible inter-conversion. For example, *Xist* was documented to have evolved from a protein-coding gene *Lnx3,* which is present in non-eutherian vertebrates (Duret *et al.*, 2006). New protein-coding genes can also be born from genes for non-coding transcripts and this is supported by evidence from human and other species (Cai *et al.*, 2008; Carvunis *et al.*, 2012; Xie *et al.*, 2012). Thus, our work here provide a comprehensive dataset for further investigation regarding the birth of coding and non-coding genes, which will aid in a better understanding of the evolution and function of plant lncRNAs.

GWAS has been widely used to uncover the genetic basis of phenotypic variation in both plants and animals. Such studies have helped to identify disease etiology, improve agriculture productivity and illuminate adaptive processes. However, previous research usually focused on the GWAS-identified loci in protein-coding regions, because of the apparent molecular function of SNP in affecting the amino acid sequence of the encoded protein. However, a large number of trait-associated SNPs was embedded in intergenic regions. In this study, we identified a large number of transcripts derived from the non-coding regions and uncov-

ered several agriculture trait-associated SNPs associated with lncRNA genes. Note that the association with trait-associated SNPs does not provide direct evidence of lncRNA biological function as flanking genes of these SNP-associated lncRNAs may also contribute to the trait. To address this issue, we also screen the function of genes in 20 kb flanking regions of each SNP-associated lincRNAs. However, we did not find the existence of any crucial protein-coding regulator that can help to explain the trait. Thus, we provisionally linked the trait-associated SNPs to lncRNA genes and proposed that these lncRNA genes might be true transcription units and functional elements contributing to important agriculture traits.

EXPERIMENTAL PROCEDURES

Plant materials and RNA extraction

Seeds from the cultivated rice subspecies *Oryza sativa* L. ssp. *Japonica* cultivar Nipponbare were grown in a greenhouse in Singapore under natural light conditions. Flower buds were collected before flowering and flowers were collected at the flowering day. Flag leaves and roots were collected at both the before- and after-flowering stage. The before-flowering sample was defined as a mixture of different stages in a period from panicle initiation to 1 day before flowering. The after-flowering sample was defined as a mixture of different stages after the flowering day. Milk grains and mature seeds were also collected. Maize (*Zea mays* L. ssp. *mays*) B73 seeds were germinated on wetted paper towel in plates for 2 days and then transferred to soil and grown for 2 weeks under 26°C and 16 h light and 8 h dark condition in a growth chamber at The Rockefeller University. Shoot and root tissues were separately collected. All samples were frozen in liquid nitrogen.

RNA was extracted using the Qiagen RNeasy Plant Mini kit. Total RNA was treated with Turbo DNase I (Life Technologies AM2238, http://www.thermofisher.com/) according to product specification. DNase I treated RNA was then applied again to an RNeasy spin column with 0.5 volumes of ethanol, washed and eluted according to the manufacturer's instructions.

Illumina non-directional paired-end RNA sequencing

Sequencing libraries were prepared using the TruSeq RNA Sample Preparation Kits v2, set A (RS-122-2001; Illumina Inc., http://www.illumina.com/) according to the manufacturer's instructions. The quality and size of cDNA libraries were checked using the Agilent 2200 TapeStation system (Agilent Inc., http://www.agilent.com/). The libraries were sequenced for 100 cycles (paired-end) on a Hiseq™ 2000 machine (Illumina Inc.).

Illumina stranded single-end RNA sequencing

Sequencing libraries were prepared using the Illumina Stranded mRNA sample Prep kit, set A (RS-122-2101; Illumina Inc.) according to the manufacturer's instructions. The quality and size of cDNA libraries for sequencing were checked using the Agilent 2200 TapeStation system (Agilent Inc.). The libraries were sequenced for 100 cycles on Hiseq™ 2500 (Illumina Inc.).

cDNA synthesis and qRT-PCR

DNase-treated total RNA was used for cDNA synthesis by using the Superscript III First strand synthesis system (Invitrogen, 18080-051, http://www.thermofisher.com/) following the manufacturer's instructions. real-time PCR was performed using SYBR Premix Ex *Taq* (TaKaRa) in a Bio-Rad CFX96 real-time PCR system. Data were collected and analyzed by Bio-Rad CFX96 real-time system. Levels of rice *ACTIN 3* and maize *ACTIN 1* were used for normalization. Primers are listed in Table S5.

Genomic data sources

Genome assemblies and gene annotations of rice (*Oryza sativa* L. ssp. *Japonica* cultivar Nipponbare, MSU Rice Genome Annotation Project Release 7) and maize (*Zea mays* L. ssp. *mays,* Maize Golden Path B73 RefGen_v2) were used for the identification of lncRNAs (Schnable *et al.*, 2009; Kawahara *et al.*, 2013).

Analysis of RNA-seq data

Both peRNA-seq and ssRNA-seq data were aligned to the rice and maize genome using TopHat v2.0.8 with default parameters (Trapnell *et al.*, 2009). Transcripts were assembled using Cufflinks v2.1.1 and their expression levels were evaluated by Cuffdiff v2.1.1 as described by Trapnell *et al.* (Trapnell *et al.*, 2012). Only transcripts with more than 1 FPKM in at least one sample were used in this study. Genomic positions of rice and maize lincRNAs and NATs were listed in Tables (S6-S9).

Histone modifications and lincRNAs

ChIP-seq data of H3K4me3, H3K9ac and H3K27me3 in rice and maize were downloaded from the NCBI Gene Expression Omnibus under accession number GSE19602 and GSE15286, respectively (Wang *et al.*, 2009; He *et al.*, 2010; Barrett *et al.*, 2013). Genomic positions of histone marks were updated to the version used in this study and compared to lincRNA loci as well as 500-bp up/down-stream genomic regions.

Conservation analysis of lincRNA

Whole-genome alignment between rice and maize were carried out by threaded blockset aligner, TBA, with default parameters (Blanchette *et al.*, 2004). Using the same version of genomic sequences, synteny blocks were screened by CoGe (Lyons and Freeling, 2008). We used following parameters: word size 8, gap start penalty 400, gap extend penalty 30, no chaining, score threshold 3000, mask threshold 0.

GWAS analysis of lincRNAs

We downloaded trait-associated SNPs data reported by Zhao *et al.* (2011) in rice and by Chia *et al.* (2012) in maize (Zhao *et al.*, 2011; Chia *et al.*, 2012). Genomic coordinates of rice SNPs were based on MSU.v6 genome assembly and maize SNPs were reported according to maize B73 RefGen_v1. After having downloaded the genomic sequences used in their studies we aligned lincRNAs to the corresponding genomic sequences. We then obtained genomic coordinates of lincRNA genes corresponding to the reported SNP datasets. LincRNA genes that overlapped with trait-associated SNPs were selected (Table S10).

Expression analysis of maize recombinant inbred line population. RNA-seq data were downloaded from NCBI SRA database (accession number: SRA054779) and aligned to B73 genome using TopHat. Only uniquely mapped reads were used in further analysis. Expression levels of protein-coding genes and lincRNAs were calculated and normalized by HTSeq and DESeq2 (Anders and Huber, 2010; Anders *et al.*, 2015).

ACKNOWLEDGEMENTS

We thank Connie Zhao and Bin Zhang (Genomics Resource Center, The Rockefeller University) for technical support, Li-Fang Huang for collecting maize samples and other lab members for fruitful discussion. This work was supported by grants from the Singapore Millennium Foundation, the Cooperative Research Program for Agriculture Science & Technology Development, Rural Development Administration, Republic of Korea [PJ906910] and Klein Wanzlebener Saatzucht (KWS) SAAT AG, Germany.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Transcription directions of peRNA-seq-assembled transcripts were supported by ssRNA-seq reads.

Figure S2. Genomic features of lincRNAs in rice and maize.

Figure S3. Distribution of three kinds of histone marks on the upstream, transcribed and down-stream regions of lincRNA genes and the correlation with expression levels.

Figure S4. Genomic features of NATs in rice and maize.

Figure S5. Temporal-spatial expression of lincRNAs and mRNAs in maize.

Figure S6. K-means clustering of rice lincRNAs.

Figure S7. Validation of differentially expressed lincRNAs by qRT-PCR.

Figure S8. Expression correlation of selected rice lincRNAs and their nearest genes.

Figure S9. Numbers of positively-correlated and negatively correlated NAT pairs in rice and maize.

Figure S10. Sequence and positional conservation of maize lincRNAs.

Figure S11. Function enrichment analysis of sense genes involved in positionally conserved NAT pairs.

Table S1. RNA-seq dataset.

Table S2. Histone marks association with lincRNA genes, protein-coding genes and transposon elements.

Table S3. GO enrichment analysis of sense genes involved in NAT pairs.

Table S4. GO enrichment analysis of genes surrounding conserved lincRNA genes.

Table S5. Primers used in qRT-PCR experiments.

Table S6. Rice lincRNAs.

Table S7. Rice NATs.

Table S8. Maize lincRNAs.

Table S9. Maize NATs.

Table S10. Original genomic positions of trait-associated SNPs (TAS-SNPs) and their associated lincRNAs.

REFERENCES

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.

Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Ariel, F., Jegu, T., Latrasse, D., Romero-Barrios, N., Christ, A., Benhamed, M. and Crespi, M. (2014) Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. *Mol. Cell* **55**, 383–396.

Atwell, S., Huang, Y.S., Vilhjalmsson, B.J. *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Barrett, T., Wilhite, S.E., Ledoux, P. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995.

Bernstein, B.E., Mikkelsen, T.S., Xie, X. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.

Blanchette, M., Kent, W.J., Riemer, C. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715.

Boerner, S. and McGinnis, K.M. (2012) Computational identification and functional predictions of long noncoding RNA in *Zea mays*. *PLoS ONE*, **7**, e43047.

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J. and Roux, F. (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet.* **6**, e1000940.

Brachi, B., Morris, G.P. and Borevitz, J.O. (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* **12**, 232.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S. and Rastan, S. (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, **71**, 515–526.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J. and Willard, H.F. (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**, 527–542.

Bruggmann, R., Bharti, A.K., Gundlach, H. *et al.* (2006) Uneven chromosome contraction and expansion in the maize genome. *Genome Res.* **16**, 1241–1251.

Buckler, E.S., Holland, J.B., Bradbury, P.J. *et al.* (2009) The genetic architecture of maize flowering time. *Science*, **325**, 714–718.

Burge, C.B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346–354.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927.

Cai, J., Zhao, R., Jiang, H. and Wang, W. (2008) De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*, **179**, 487–496.

Carvunis, A.R., Rolland, T., Wapinski, I. *et al.* (2012) Proto-genes and de novo gene birth. *Nature*, **487**, 370–374.

Chen, X. (2009) Small RNAs and their roles in plant development. *Annu. Rev. Cell Dev. Biol.* **25**, 21–44.

Chia, J.M., Song, C., Bradbury, P.J. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807.

Ding, J., Lu, Q., Ouyang, Y., Mao, H., Zhang, P., Yao, J., Xu, C., Li, X., Xiao, J. and 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.

Dinger, M.E., Pang, K.C., Mercer, T.R. and Mattick, J.S. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**, e1000176.

Duret, L., Chureau, C., Samain, S., Weissenbach, J. and Avner, P. (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, **312**, 1653–1655.

Faghihi, M.A. and Wahlestedt, C. (2009) Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol.* **10**, 637–643.

Famoso, A.N., Zhao, K., Clark, R.T., Tung, C.W., Wright, M.H., Bustamante, C., Kochian, L.V. and McCouch, S.R. (2011) Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. *PLoS Genet.* **7**, e1002221.

Guttman, M., Amit, I., Garber, M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

He, G., Zhu, X., Elling, A.A. *et al.* (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell* **22**, 17–33.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367.

Huang, X., Wei, X., Sang, T. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967.

Katayama, S., Tomaru, Y., Kasukawa, T. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.

Katz, Y., Wang, E.T., Airoldi, E.M. and Burge, C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015.

Kawahara, Y., de la Bastide, M., Hamilton, J.P. *et al.* (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N Y)*, **6**, 4.

Khalil, A.M., Guttman, M., Huarte, M. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106**, 11667–11672.

Kumar, V., Wijmenga, C. and Withoff, S. (2012) From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. *Semin. Immunopathol.* **34**, 567–580.

Kumar, V., Westra, H.J., Karjalainen, J. *et al.* (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.* **9**, e1003201.

Kump, K.L., Bradbury, P.J., Wisser, R.J. *et al.* (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* **43**, 163–168.

Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C.M. and Casari, G. (2004) In search of antisense. *Trends Biochem. Sci.* **29**, 88–94.

Li, X., Qian, Q., Fu, Z. *et al.* (2003) Control of tillering in rice. *Nature*, **422**, 618–621.

Li, L., Petsch, K., Shimizu, R. *et al.* (2013) Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS Genet.* **9**, e1003202.

Li, L., Eichten, S.R., Shimizu, R. *et al.* (2014) Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol.* **15**, R40.

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

Lyons, E. and Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.*, **53**, 661–673.

Ma, L., Bajic, V.B. and Zhang, Z. (2013) On the classification of long noncoding RNAs. *RNA Biol.* **10**, 925–933.

McMullen, M.D., Kresovich, S., Villeda, H.S. *et al.* (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737–740.

Mikkelsen, T.S., Ku, M., Jaffe, D.B. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Nemri, A., Atwell, S., Tarone, A.M., Huang, Y.S., Zhao, K., Studholme, D.J., Nordborg, M. and Jones, J.D. (2010) Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proc. Natl Acad. Sci. USA* **107**, 10302–10307.

Nesterova, T.B., Slobodyanyuk, S.Y., Elisaphenko, E.A., Shevchenko, A.I., Johnston, C., Pavlova, M.E., Rogozin, I.B., Kolesnikov, N.N., Brockdorff, N. and Zakian, S.M. (2001) Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.* **11**, 833–849.

Pauli, A., Valen, E., Lin, M.F. *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **22**, 577–591.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S. and Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. *Nature*, **379**, 131–137.

Poland, J.A., Bradbury, P.J., Buckler, E.S. and Nelson, R.J. (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc. Natl Acad. Sci. USA* **108**, 6893–6898.

Rapicavoli, N.A., Poth, E.M. and Blackshaw, S. (2010) The long noncoding RNA RNCR2 directs mouse retinal cell specification. *BMC Dev. Biol.* **10**, 49.

Ravasi, T., Suzuki, H., Pang, K.C. *et al.* (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16**, 11–19.

Rinn, J.L. and Chang, H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166.

Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schultes, E.A., Spasic, A., Mohanty, U. and Bartel, D.P. (2005) Compact and ordered collapse of randomly generated RNA sequences. *Nat. Struct. Mol. Biol.* **12**, 1130–1136.

Sone, M., Hayashi, T., Tarui, H., Agata, K., Takeichi, M. and Nakagawa, S. (2007) The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. *J. Cell Sci.* **120**, 2498–2506.

Struhl, K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14**, 103–105.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B. and Buckler, E.S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Tsuiji, H., Yoshimoto, R., Hasegawa, Y., Furuno, M., Yoshida, M. and Nakagawa, S. (2011) Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. *Genes Cells*, **16**, 479–490.

Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26–46.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H. and Bartel, D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**, 1537–1550.

Wang, X., Elling, A.A., Li, X., Li, N., Peng, Z., He, G., Sun, H., Qi, Y., Liu, X.S. and Deng, X.W. (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell*, **21**, 1053–1069.

Wang, H., Chung, P.J., Liu, J., Jang, I.C., Kean, M.J., Xu, J. and Chua, N.H. (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Res.* **24**, 444–453.

Xie, C., Zhang, Y.E., Chen, J.Y., Liu, C.J., Zhou, W.Z., Li, Y., Zhang, M., Zhang, R., Wei, L. and Li, C.Y. (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8**, e1002942.

Zhang, Y.C., Liao, J.Y., Li, Z.Y., Yu, Y., Zhang, J.P., Li, Q.F., Qu, L.H., Shu, W.S. and Chen, Y.Q. (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, K., Tung, C.W., Eizenga, G.C. *et al.* (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2**, 467.