New Resources

New lncRNA annotation reveals extensive functional divergence of the transcriptome in maize

Running title: Comprehensive analysis of maize lncRNAs

Linqian Han, Zhenna Mu, Zi Luo, Qingchun Pan and Lin Li*

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

* Correspondence: Lin Li (hzaulilin@mail.hzau.edu.cn)

Edited by: Jin-Sheng Lai, China Agricultural University, China

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/jipb.12708]

This article is protected by copyright. All rights reserved.
Received: July 5, 2018; Accepted: August 6, 2018
Abstract Long non-coding RNAs (lncRNAs), whose sequences are ~200 bp or longer and unlikely to encode proteins, may play an important role in eukaryotic gene regulation. Although the latest maize (*Zea Mays* L.) reference genome provides an essential genomic resource, genome-wide annotations of maize lncRNAs have not been updated (Jiao et al. 2017). Here, we report on a large transcriptomic dataset collected from 749 RNA sequencing experiments across different tissues and stages of the maize reference inbred B73 line and 60 from its wild relative teosinte. We identified 18,165 high-confidence lncRNAs in maize, of which 6,873 are conserved between maize and teosinte. We uncovered distinct genomic characteristics of conserved lncRNAs, non-conserved lncRNAs, and protein-coding transcripts. Intriguingly, Shannon entropy analysis showed that conserved lncRNAs are likely to be expressed similarly to protein-coding transcripts. Co-expression network analysis revealed significant variation in the degree of co-expression. Furthermore, selection analysis indicated that conserved lncRNAs are more likely than non-conserved lncRNAs to be located in regions subject to recent selection, indicating evolutionary differentiation. Our results provide the latest genome-wide annotation and analysis of maize lncRNAs and uncover potential functional divergence between protein-coding, conserved lncRNA, and non-conserved lncRNA genes, demonstrating the high complexity of the maize transcriptome.
INTRODUCTION

With the sequencing of many genomes, including that of humans, eukaryote genomes have been discovered to be complex beyond expectations (Djebali et al. 2012). In addition to protein-coding genes, significant fractions of the genomes of humans, animals, and plants undergo transcription, demonstrating the pervasive nature of this process (David et al. 2006; Dinger et al. 2009; Berretta et al. 2009; Djebali et al. 2012). This has broadened the understanding of so-called “junk” sequences, many of which encode long non-coding RNAs (lncRNAs). These lncRNAs are transcripts of more than 200 bp in length that likely function as non-coding molecules. As with mRNAs, the majority of lncRNAs are transcribed by RNA polymerase II (Ulitsky et al. 2013). Previous studies have shown that lncRNAs can also be synthesized by RNA polymerases IV and V, especially in plants (Wierzbicki et al. 2009).

Early studies on lncRNAs considered them to be transcriptional noise (Makunin et al. 2014; Kern et al. 2015). However, over half a million lncRNAs have been identified across hundreds of species (Bhatia et al. 2017), some of which have been reported to function as microRNA mimics, epigenetic complex decoys, and precursors of small RNAs (Ponting et al. 2009; Zhu et al. 2012; Kung et al. 2013; Marín-Béjar et al. 2015; Wu et al. 2016). Clearly, lncRNAs have multiple functions and likely play important roles in evolution and speciation.

Maize is a worldwide-grown grain crop and is important for food, animal feed, and bio-energy production. Although the maize reference genome was released in 2009 and updated in 2017, the genome-wide annotation of lncRNAs is still ongoing (Schnable et al. 2009; Jiao et al. 2017). A computational pipeline for lncRNA identification uncovered hundreds of lncRNAs in maize (Boerner et al. 2012), and a later study identified and characterized some 2,000 high-confidence lncRNAs (Li et al. 2014), suggesting that maize lncRNAs are widespread. Further, 664 maize lncRNAs were shown to be responsive to drought stress (Zhang et al. 2014). Additionally, ribosomal RNA depletion and deep total RNA sequencing on maize B73 leaves, at the V7 stage under conditions of nitrogen deficiency and sufficiency, identified a large number of nitrogen-
responsive intergenic and intron-derived IncRNAs (Lv et al. 2016). Increasing numbers of functional maize IncRNAs have been reported. However, since the release of the latest maize reference genome, a new comprehensive annotation of IncRNAs has been needed.

Maize was domesticated from ancient teosinte about 10,000 years ago (Hufford et al. 2012). Although this is a short speciation period, the morphology of maize plants has diverged dramatically from that of teosinte. It has been proposed that IncRNAs are likely to be intermediates in the formation of new genes (Hezroni et al. 2015). The genome-wide identification and characterization of conserved IncRNAs and non-conserved IncRNAs between maize and teosinte would provide insight into the molecular mechanisms underlying maize speciation from the ancient teosinte. However, the evolutionary roles of conserved and non-conserved IncRNAs during maize domestication and improvement are largely unknown.

In this study, we collected a large transcriptome dataset from 749 RNA-Seq experiments on the maize reference inbred line B73, for the annotation of IncRNAs. We identified 18,165 maize IncRNAs, providing a near-complete landscape of the maize IncRNAs known thus far. Comparative transcriptomic analysis identified conserved and non-conserved IncRNAs, as well as genomic and expression features distinct from protein-coding genes. Co-expression network and co-localization with selective genomic regions demonstrated that protein-coding genes and conserved and non-conserved RNA genes exhibited extensive evolutionary divergence. Our study provides a nearly complete landscape of IncRNAs, along with an analysis of the evolutionary divergence of IncRNAs in maize. The substantial number of IncRNAs and the dramatic functional differentiation of protein-coding transcripts, conserved IncRNAs, and non-conserved IncRNAs describe an extremely complex maize transcriptome.
RESULTS

Extreme complexity of the maize transcriptome

To establish a transcriptome landscape of lncRNAs in maize, we collected a transcriptome dataset consisting of data from 749 RNA sequencing (RNA-Seq) experiments (420 single-end datasets and 329 pair-end datasets; Table S1 and Table S2) on the maize reference inbred line B73, encompassing 17 different tissues and developmental stages (Figure 1A). A genome-guided transcriptome assembly strategy was used to identify transcribed genomic loci and transcription isoforms. A substantial number (86,683) of genomic loci were detected as being expressed. The number of genomic loci detected ranged from 46,712 in the spikelet to 52,980 in the anther. Accordingly, over 424,000 transcript isoforms were detected across different tissues and stages in maize (Figure 1B). Notably, anthers, where the largest number of genomic loci were transcribed, did not express the highest number of transcript isoforms per expressed genomic locus. Rather, the maize kernel expressed the most transcript isoforms overall. Interestingly, these transcribed genomic loci covered a large proportion (60.3%) of the genome, indicating that there is a prevalence of transcription in the maize genome, similar to what has been shown in other eukaryotes (David et al. 2006; Dinger et al. 2009; Berretta et al. 2009; Djebali et al. 2012). These results confirmed the very high complexity of the maize transcriptome, which has also been previously reported (Wang et al. 2016).

A series of genome-wide lncRNA identification and quality-control steps were implemented to ensure the inclusion of only high-confidence lncRNAs for further analysis. To date, this is the largest set (18,165) of maize lncRNAs that has been assembled (Dataset S1; Dataset S2). Compared with genome-wide annotation by the maize genome sequencing project (Schnable et al. 2009; Jiao et al. 2017), the transcriptome landscape annotated by our genome-guided assembly strategy uncovered six other types of maize high-confidence lncRNAs (Figure 1C; Table S3). The newest set of maize lncRNAs covers 92.1% of previously annotated lncRNAs (Li et al. 2014; Wang...
et al. 2016). Moreover, 15,315 novel maize lncRNAs were detected in the maize reference inbred line B73.

To determine the evolutionary roles of lncRNAs, we conducted transcriptomic analyses in teosinte so as to establish the teosinte transcriptome landscape and perform genome-wide identification of teosinte lncRNAs. A total of 60 RNA-Seq datasets were collected from teosinte (Table S4) and we uncovered 334,600 teosinte transcript isoforms. A genome-wide lncRNA scan identified 6,873 lncRNAs in the teosinte transcriptome, providing a reference transcriptome for the detection of conserved and non-conserved lncRNAs in maize.

Transcriptomic comparison between maize and teosinte identified not only a large number of conserved protein-coding transcripts, but also a substantial number of conserved and non-conserved lncRNAs (Figure 1D). As expected, the proportion of lncRNAs conserved between maize and teosinte is much higher than that between maize, sorghum (*Sorghum bicolor* L.), and rice (*Oryza Sativa* L.), confirming the close relationship between maize and its relative teosinte. The proportion of conserved and non-conserved lncRNAs varies dramatically across different classes of lncRNAs, relative to annotated transcription loci in the reference genome (Figure 1E; Table S5). It is worth noting that lncRNAs of class j, which are derived from intronic regions of protein-coding genes, are more likely to be conserved between species, while intergenic lncRNAs (class code u) are the least conserved. These results indicated that lncRNAs may have played a role in maize domestication.

**Distinct genomic and expression features of conserved and non-conserved lncRNAs compared with protein-coding transcripts**

All of the maize transcripts were next classified into three groups, according to protein-coding potential and evolutionary conservation: i.e., protein-coding transcripts, conserved lncRNAs, and non-conserved lncRNAs. As expected, all three types of transcripts are distributed along the entire maize genome (Figure 2A). However, protein-coding transcripts are more likely to be distributed at the ends of chromosomes and are depleted
in the pericentromeric regions, whereas IncRNAs are more evenly distributed, especially non-conserved IncRNAs, which are also present in the centromeres (Figure 2A,B).

Previous studies showed that most IncRNAs are derived from repetitive genomic regions (Kapusta et al. 2013), which explains the even distribution of IncRNAs along the chromosomes. Additionally, the majority of IncRNAs have a single exon and are less than 2.5 Kb in length, significantly different from the protein-coding transcripts (Figure 2C,D). The annotation of maize reference genome V4 accompanied by single-molecule real-time (SMRT) sequencing on the transcriptome could detect longer IncRNAs with more exons as compared with routine mRNA-Seq, confirming the advantage of the SMRT technique (Wang et al. 2016). However, the IncRNA loci detected by SMRT are otherwise largely equivalent to those detected by routine mRNA-Seq. Moreover, we quantified the expression levels of all transcripts across different tissues and stages of B73. Based on the ratio of transcripts detected in a specific tissue to total transcripts, most protein-coding transcripts were expressed in the majority of tissues and stages, although at different levels, whereas conserved IncRNAs showed less tissue specificity and non-conserved IncRNAs were the least likely to be expressed constitutively (Figure 2E).

Interestingly, most non-conserved IncRNAs were detected in anthers, the maize reproductive organ. The Shannon entropy value, ranging from zero for expression in a single tissue to log2(number of tissues) for uniform expression in all tissues considered, provides a robust statistic for detecting the expression patterns of transcripts (Schug et al. 2015). Comparison of Shannon entropy values showed a significant expression variation across different tissues and stages between protein-coding transcripts, conserved IncRNAs, and non-conserved IncRNAs (Figure 2F). Protein-coding transcripts had the highest Shannon entropy values, indicating a strong likelihood of being constitutively expressed, whereas non-conserved IncRNAs had the lowest Shannon entropy values, suggesting the most tissue and/or stage-specific expression pattern. These dramatic genomic and expression variations between protein-coding transcripts and IncRNAs indicate potential functional differentiation of the transcriptome in maize.
Functional differentiation of conserved lncRNAs, non-conserved lncRNAs, and protein-coding genes

To detect the potential functional roles of lncRNAs, we constructed a comprehensive co-expression network, including protein-coding, conserved lncRNA genes, and non-conserved lncRNA genes simultaneously. The network was constructed using normalized expression levels of all identified lncRNAs (18,165) and protein-coding RNAs (133,836) annotated by the latest maize reference genome. It contained 750 co-expression modules with 15,460 co-expressed genes (Figure 3A), which is comparable to co-expression networks previously observed in maize (Schaefer et al. 2014; Li et al. 2016).

Degree of co-expression is the number of genes/transcripts that are significantly associated with the target gene/transcript, and weighted degree is the degree of co-expression adjusted based on the Pearson correlation coefficient. The higher the degree, the more functionally important the co-expression network node (Zhu et al. 2007). A comparison of co-expression between protein-coding transcripts, conserved lncRNAs, and non-conserved lncRNAs revealed that protein-coding genes had higher degrees and weighted degrees of co-expression than conserved lncRNAs overall, as expected. However, unexpectedly, both the weighted and the non-weighted degrees of co-expression were significantly higher for non-conserved transcripts than for conserved transcripts (Figure 3B, C). This may be due to either the greater importance of non-conserved lncRNAs for maize speciation or the tissue- and stage-specific expression patterns for a large number of non-conserved lncRNAs in the reproductive organ. Interestingly, a certain number of conserved and non-conserved lncRNAs had very high co-expression degrees and weighted degrees. Variation in the degree of co-expression is likely to be related to the functional divergence of the maize transcriptome.

To further detect functional divergence in the co-expression network, we analyzed the proportion of different classes of transcripts in each module. Interestingly, some co-expression modules were dominated by a specific class of transcripts (Figure 3D). For example, one module consisted of a single protein-coding gene co-expressed
with many non-conserved lncRNA genes (Figure 3E), whereas some modules had one lncRNA gene associated with many protein-coding genes (Figure 3F).

Additionally, several modules demonstrated complex co-expression relationships between conserved lncRNAs, non-conserved lncRNAs, and protein-coding transcripts (Figures S1, S2, S3). Specifically, we identified three co-expression modules that contained only non-conserved lncRNAs and protein-coding RNAs. These protein-coding genes that were associated with non-conserved lncRNAs showed strong functional enrichment in transferase activity and nutrient reservoir activity (Figure S4). Moreover, hub genes, potential super-regulators, were identified in a few co-expression modules (Figure S5). The existence of such classes of dominant co-expression modules and the existence of potential super-regulatory factors between protein-coding and non-coding genes are suggestive of the complex regulatory relationships in the transcriptome.

Co-localization with selective regions during domestication and breeding suggests evolutionary differentiation of the maize transcriptome

Maize was domesticated from ancient teosinte and improved to satisfy human needs (Doebley 2006; Hufford et al. 2012). During the process of domestication and improvement, a large fraction of the maize genome was subjected to selection (Hufford et al. 2012). Genes in the genomic regions affected by selection are more likely to be domestication and improvement targets, and are likely to be evolutionarily important in maize. As expected, co-localization of selective signals with protein-coding and lncRNA genes showed an enrichment of protein-coding genes, along with a depletion of lncRNA genes, in the selected genomic regions (Figure 4A). Interestingly, more than twice as many conserved lncRNA genes than non-conserved lncRNA genes are located in the selected genomic regions, suggesting the greater evolutionary importance of conserved lncRNAs. Additionally, some conserved lncRNA genes located in the selected genomic regions are hub genes associated with many other protein-coding genes in the co-expression network (Figure 4B). These protein-coding genes show related function
according to the Gene Ontology (GO) analysis (Figure 4C). These results imply that the maize transcriptome exhibits evolutionary differentiation.

**DISCUSSION**

Maize (*Zea mays* L.), one of the most important crops worldwide, has been extensively studied. However, genome-wide and systematic annotation, characterization, and genetic mapping of maize lncRNAs are needed to complement the current updated genome annotation, facilitate QTL cloning and functional genomics studies, and enable more comprehensive expression regulation research. Here, we integrated all available transcriptomic datasets to identify a comprehensive set of lncRNAs, providing a new annotation resource of the maize genome.

All of the RNA-Seq data we analyzed are publicly available and might have been collected from different individuals of the B73 inbred line. It has been reported that there is about 2.3% genomic variation within the B73 line (Liang et al. 2016). However, the genomic variation among B73 samples is largely due to single-nucleotide polymorphisms (SNPs) and, thus, less likely to alter the expression of lncRNAs. Therefore, our genome-wide annotation of lncRNAs in the B73 genome provides a near-complete landscape of lncRNAs in maize (Dataset S3; Dataset S4).

Maize was domesticated from teosinte (Hufford et al. 2012). As the wild relative of maize, teosinte provides a perfect counterpart for the detection of selection signals that operated during maize domestication (Hufford et al. 2012). Notably, however, the teosinte dataset is not as comprehensive as that for maize, and the teosinte used for our study is not likely to be equivalent to the actual maize ancestors. Thus, some of the non-conserved lncRNAs that we identified as being specific to maize, in our data, might also be uncovered in a more extensive analysis of the teosinte transcriptome. However, the identification of conserved lncRNAs should be robust.

Most interestingly, lncRNAs were expressed in a tissue- and stage-specific manner overall, but non-conserved lncRNAs were more likely to be detected in reproductive organs, such as anthers, than conserved lncRNAs. This may imply that
conserved lncRNAs have functional divergence from non-conserved lncRNAs. During the reproductive process in plants, the heterochromatin compartment is reorganized, the genome-wide histone modification landscape is altered, and nucleosome composition is remodeled, resulting in reprogramming at the physiological, cytological, and transcriptome levels (She and Baroux 2014; Baroux and Autran 2015). Further transcriptomic reprogramming may lead to random transcription of non-coding genomic regions, mostly non-conserved lncRNAs. It is worth noting that a substantial number of non-conserved lncRNAs were also detected in non-reproductive organs and co-expressed with tens of protein-coding genes, potentially having functions related to the divergence of maize morphology from that of teosinte.

As in earlier studies (Moreno-Risueno et al. 2010; Movahedi et al. 2011; Mutwil et al. 2011), we established that a substantial number of co-expression modules contain genes with related functional GO enrichments, suggestive of regulatory relationships and functions of the co-expression modules. The co-expression divergence identified in our study implies the functional divergence of protein-coding, conserved lncRNA, and non-conserved lncRNA genes. Conserved lncRNA genes were also more likely to be located in the domestication and selection regions of maize than non-conserved lncRNA genes. These conserved lncRNA genes would have been subject to functional alteration during the speciation of maize.

Non-conserved maize-specific lncRNAs may stem from the divergence point of maize from ancient teosinte, less than 10,000 years ago (Hufford et al. 2012). These non-conserved lncRNA genes could have been intermediates in the formation of new protein-coding genes, thereby playing vital roles in the speciation process (Hezroni et al. 2015). We did identify several co-expression modules in which non-conserved lncRNAs are associated with tens of protein-coding genes. The associated genes are significantly enriched in GO groups related to transferase activity and nutrient reservoir activity (Figure S4). Transferase activity is a molecular function of catalysis involving the transfer of a chemical group, such as a methyl group, glycosyl group, acyl group, phosphorus-containing group, or other group (http://www.geneontology.org/). These non-conserved lncRNAs might have altered the expression levels of genes encoding proteins
with transferase activity as part of the methylation landscape switch during maize speciation (Zhang et al. 2018). Our study provides a genome-wide glimpse into the functional divergence of lncRNAs in maize.

MATERIALS AND METHODS

RNA-Seq data collection

Transcriptome datasets from the maize (Zea Mays L.) reference inbred line B73 and teosinte were collected from the Sequence Read Archive of the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) website. These collected datasets were generated by 1,027 RNA-Seq experiments on different tissues and stages of the maize reference inbred line B73 (Table S1). After Trimomatic quality control and TopHat mapping filtering, a total of 420 single-end and 329 paired-end RNA-Seq datasets were used to identify long non-coding RNAs of B73 (Table S2) (Bolger et al. 2014; Trapnell et al. 2009). Additionally, 52 single-end and 8 paired-end RNA-Seq datasets from teosinte were collected and employed for the identification of teosinte transcripts (Table S3).

Genome-wide identification of lncRNAs in maize

For the RNA-Seq data, all sequence reads from each experiment were aligned to the B73 reference genome (v4.32) using TopHat with parameters “-G Zea_mays.AGPv4.32.gff3 --max-multihits 1”. Next, Cufflinks was employed to assemble the transcriptome for each experiment (Mortazavi et al. 2008). To reduce transcriptional noise, only those transcript isoforms which were detected in two or more experiments were retained for further analyses. Lastly, we used LncRNA_Finder.pl to identify long non-coding RNAs through a series of filters to eliminate potential protein-coding RNAs (Li et al. 2014).

Classification of maize conserved and non-conserved lncRNAs

Trinity was used to de novo assemble the teosinte transcriptome with parameters “--seqType fq --JM 225G --CPU 15” (Grabherr et al. 2011). Transcript isoforms of teosinte were aligned against long non-coding RNAs of maize using blastn with the parameters “-
If sequence similarity between two species was greater than 50%, we defined these IncRNA transcripts as conserved IncRNAs, and otherwise, as non-conserved IncRNAs.

**Construction and analysis of a co-expression network for protein-coding genes, conserved IncRNA genes, and non-conserved IncRNA genes**

Transcripts per million (TPM) was adopted as the proxy for normalized expression level using RSEM with the parameters: "--bowtie2-k 20 --bowtie2-mismatch-rate 0.03 --estimate-rspd --append-names" (Li et al. 2011). A matrix of the maize transcriptome from 156 single-end RNA-Seq experiments was used for the construction of a maize co-expression network (152,001 annotated maize transcripts, which included 6,873 conserved IncRNAs, 11,292 non-conserved IncRNAs, and 133,836 protein-coding RNAs across 749 different tissues and stages). To reduce the bias on Pearson correlation coefficients (PCCs) caused by highly expressed transcripts, the TPM value was transformed by an inverse hyperbolic sine function. Co-expression networks were constructed by calculating PCCs between all pairs of transcripts (152,001 × 152,001).

MR (mutual rank) values were then employed for the final co-expression relationship identification through the following formula (Obayashi and Kinoshita, 2009; Jennifer et al. 2017):

\[ \text{mutual rank} = \sqrt{\text{rank}(a \rightarrow b) \times \text{rank}(b \rightarrow a)} \]

Any edge with PCC <0.3 or edge weight = $e^{-(\text{MR} - 1)/5} < 0.01$ was excluded. Remaining isoforms were clustered using the mcl (https://github.com/JohannesBuchner/mcl) clustering algorithm. Finally, 750 modules with at least 10 isoforms were obtained. The software Gephi (https://gephi.org/) was used for visualization and feature extraction of the co-expression networks.

**Evolutionary analysis by co-localization with selected genomic regions during maize domestication and improvement**
We downloaded the supplementary data regarding genetic regions associated with domestication and improvement (Hufford et al. 2012). According to the class codes of the transcriptome relative to the annotated maize reference genome, conserved lncRNAs, non-conserved lncRNAs, and protein-coding RNAs were recorded as the gene accessions. Then, we evaluated the proportion of genes located in different domestication and improvement genomic regions as follows:

$$\text{con}_{\text{dom}}\% = \frac{\text{number located in domestication region}}{\text{gene number of all conserved lncRNAs}} \times 100\%,$$

where \( \text{con}_{\text{dom}}\% \) represents percentage of conserved genes in a given domesticated region.

Comparisons between protein-coding genes, conserved lncRNA genes, and non-conserved lncRNA genes were performed using 1,000 simulations for enrichment analysis in the domestication and improvement regions.

**ACKNOWLEDGEMENTS**

We thank all of the contributors in the maize research community for the generation of RNA-Seq data for the reference maize inbred line B73. This research was supported by the National Key Research and Development Program of China (2016YFD0100404) and the Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (2662016PY096).

**AUTHOR CONTRIBUTIONS**

L.L. designed and supervised this study. L.H., Z.L. and Q.P. performed the data analysis. L.L. and L.H. prepared the manuscript, which was read and approved by all authors.
REFERENCES


Figures

Figure 1. Transcriptome landscape in maize

(A) Flowchart of the bioinformatics pipeline for the identification of lncRNAs in maize. (B) Numbers of genes and transcript isoforms identified across different tissues and stages. (C) Different classes of lncRNAs in terms of physical relationships between putative lncRNA genes and protein-coding genes. (D) Proportion of conserved (con) and non-conserved (non) lncRNAs of maize based on separate alignments against teosinte (Teo), sorghum (Sbi), and rice transcripts. **, difference at P < 0.01 significance level. (E) Number of conserved and non-conserved maize lncRNAs across different lncRNA classes.

Figure 2. Distinct genomic and expression features of maize lncRNAs

(A) Genome-wide distribution of lncRNA and protein-coding genes. (B) Close-up view of the distribution of lncRNA and protein-coding genes on chromosome 1. (C) Numbers of exons in conserved lncRNA, non-conserved lncRNA, and protein-coding transcripts. v4_non and v4_con represent non-conserved and conserved lncRNAs, respectively, annotated in the fourth version of maize reference genome (Jiao et al. 2017). (D) Lengths of different types of transcripts. (E) Percentages of non-conserved lncRNA, conserved lncRNA, and protein-coding transcripts detected across different tissues and stages. (F) Shannon entropy of expression levels for different types of transcripts. Pro represents protein-coding RNAs; non and con represent non-conserved and conserved lncRNAs identified in our study, respectively.
Figure 3. Co-expression network shows potential functional divergence of maize transcriptome

(A) Gene-to-gene co-expression networks, including over 700 modules. (B) and (C) Differentiation of degree (B) and weighted degree (C) for protein-coding (pro), conserved (con) lncRNA, and non-conserved (non) lncRNA genes. **, difference at $P < 0.01$ significance level. The horizontal line within the box plots is the median value of the number of co-expression degrees. The ‘dots’ at the end of the boxplot represent outliers, which have values larger than the upper inner fence of the box plot. (D) Proportion of protein-coding, conserved lncRNA, and non-conserved lncRNA genes for all co-expression modules. (E) Example of a co-expression module with a high proportion of non-conserved RNA genes. (F) Example of a co-expression module with a high proportion of protein-coding genes.

Figure 4. Co-localization of selective signals demonstrates that lncRNAs are less subject to selection during maize domestication and improvement than protein-coding genes

(A) Percentage of protein-coding (pro), conserved (con), and non-conserved (non) RNA genes located in the domestication and improvement genomic regions. **, difference at $P < 0.01$ significance level. (B) Example of a conserved lncRNA gene located in the domestication and improvement genomic region, which is also a hub gene in co-expression networks. (C) Significant Gene Ontology enrichment for the genes co-expressed with the conserved lncRNA gene located in the domestication and improvement genomic region.
SUPPORTING INFORMATION

Dataset S1. Conserved lncRNAs in fasta format.

Dataset S2. Non-conserved lncRNAs in fasta format.

Dataset S3. GTF file of conserved lncRNAs.

Dataset S4. GTF file of non-conserved lncRNAs.

Figure S1. Co-expression modules of all conserved lncRNAs (green) and non-conserved lncRNAs (blue).

Figure S2. Co-expression modules of all conserved lncRNAs (green) and protein-coding RNAs (red).

Figure S3. Co-expression modules of all non-conserved lncRNAs (blue) and protein-coding RNAs (red).

Figure S4. Gene Ontology analysis of three co-expression modules which only contain non-conserved lncRNAs (blue) and protein-coding RNAs (red) by using AGRIGO (http://systemsbiology.cau.edu.cn/agriGOv2/).

Figure S5. Co-expression modules with obvious hub genes for all conserved lncRNAs (green), non-conserved lncRNAs (blue), and protein-coding RNAs (red).

Table S1. Datasets collected in this study. A total of 1,027 RNA-Seq experiments on different tissues and stages of the maize reference inbred line B73 were collected and used.

Table S2. A total of 420 single-end and 329 paired-end RNA-Seq datasets were used to identify long non-coding RNAs of B73.

Table S3. Summary of different transcripts in B73.

Table S4. Datasets used in this study. A total of 52 single-end and 8 paired-end RNA-Seq datasets of teosinte were collected and employed for the identification of teosinte transcripts.

Table S5. The number of genes, isoforms, conserved lncRNAs, non-conserved lncRNAs, and protein-coding RNAs across 156 different tissues/stages of maize.
Figure 1

Figure 2
Figure 3

Figure 4