Running title:
Transcriptome of maize seed

One-sentence summary:
Transcriptomic analysis of maize seed at 9 DAP will aid understanding how maize embryo and endosperm are differentially regulated in early development stage.

Correspondence:
Chunyi Zhang
Biotechnology Research Institute
Chinese Academy of Agricultural Sciences
12 ZhongGuanCun NanDaJie, Haidian District
Beijing 100081
People’s Republic of China
Tel.: 86-10-82106403
Fax: 86-10-82106138
E-mail: zhangchunyi@caas.cn

Research Category:
Genome Analysis
The differential transcription network between embryo and endosperm in the early developing maize seed

Xiaoduo Lu¹, ³, †, Dijun Chen², ⁴, ⁵, †, Defeng Shu¹, Zhao Zhang², Weixuan Wang³, ⁶, Christian Klukas⁵, Ling-ling Chen⁴, Yunliu Fan³, ⁶, Ming Chen², *, Chunyi Zhang³, ⁶ *

¹Department of Life Sciences, Qilu Normal University, Jinan 250200, China
²Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, P. R. China
³Department of Crop Genomics & Genetic Improvement, Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China
⁴State Key Laboratory of Crop Genetic Improvement, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, P. R. China
⁵Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben (IPK), 06446 Gatersleben, Germany
⁶National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI), Beijing 100081, P. R. China

†These two authors contributed equally to this work.
This work was supported by the National Basic Research Program of China (grant number 2013CB 127003 to C. Z.).

*Corresponding authors:

Ming Chen; email: mchen@zju.edu.cn

Chunyi Zhang; email: zhangchunyi@caas.cn; Fax: 86-10-82106138
ABSTRACT

Transcriptome analysis of early developing maize seed was conducted using Illumina sequencing. We mapped 11,074,508 and 11,495,788 paired-end reads from 9 DAP (days after pollination) endosperm and embryo, respectively, to define gene structure and alternative splicing events as well as transcriptional regulators of gene expression to quantify transcript abundance in both embryo and endosperm. We identified a large number of novel transcribed regions (NTRs) that did not fall within maize annotated regions and many of the NTRs were tissue-specifically expressed. We found that 50.7% (8,556/16,878) of multi-exonic genes were alternatively spliced, and some transcript isoforms were specifically expressed either in endosperm or in embryo, respectively. In addition, a total of 46 trans-splicing events, with nine intra-chromosomal and 37 inter-chromosomal events, were found in our dataset. Many metabolic activities were specifically assigned to endosperm and embryo, respectively, such as starch biosynthesis in endosperm and lipid biosynthesis in embryo. Finally, a number of transcription factors and imprinting genes were found to be specifically expressed in embryo or endosperm. This dataset will aid understanding how embryo/endosperm development in maize is differentially regulated.
Maize seeds are one of the most important crop materials which provide resource for food, feed, biofuel and raw material for processing. Maize seed development initiates from double fertilization that two of the pollen sperms fuse with egg cell and central cell to produce embryo and endosperm, respectively (Randolph, 1936; Chaudhury et al., 2001). The main function of endosperm is to provide nutrient for the developing embryo and germinating embryo.

After fertilization the zygote undergoes an asymmetric division into a small apical and a large basal cell giving rise to the embryo proper and the suspensor, respectively. The radial symmetry of the proembryo is shifted to a bilateral symmetry at the transition stage which is characterized by the protoderm formation. The shoot apical meristem (SAM) and the root apical meristem (RAM) can be distinguished at the onset of the coleoptilar stage, and then the position of the future coleoptile is marked by a small protuberance. The mature embryo is composed of the embryo axis, which is formed by the plumule with five or six short internodes and leaf primordia and primary root surrounded by the coleoptile and the coleorhiza, respectively, and the scutellum (Randolph, 1936; Abbe and Stein, 1954; Diboll, 1968; Lämmeren, 1986; Vernoud et al., 2005).

Maize endosperm follows the nuclear-type endosperm development where the fertilized central cell undergoes several rounds of synchronous division in the absence of cell wall formation and cytokinesis to produce a syncytium (Lopes and Larkins, 1993; Olsen, 2004). Cellularization allows formation of the internuclear radial microtubule systems and open-ended alveolation from the periphery of the endosperm toward the central vacuole (Olsen et al., 1999; Olsen, 2004). Four major cell types of the maize endosperm are differentiated: transfer cells, aleurone cells, starchy endosperm cells, and embryo-surrounding region (ESR) cells. During maize endosperm development, there are three different types of cell cycles (Kowles and Phillips, 1988): 1) cytokinetic mitosis results in a syncytium; 2) mitosis occurs after cellularization and lasts in the central endosperm but continues in the aleurone and subaleurone layers, and cell division pattern occurs and stop in wave-like pattern from the basal to the central region of the endosperm; and 3) endoreduplication, the reiterated rounds of DNA replication without chromatin condensation, sister chromatid segregation, or cytokinesis, results in endopolyploidy cells. Embryo and endosperm are both seed
compartments exhibiting dramatic difference from multiple aspects. Maize endosperm is not a transient tissue like Arabidopsis which is absorbed at later stage of seed development. Gene expression is a key event to determine embryo and endosperm development. Some genes have been functionally characterized in maize seed development. For example, Crinkly4 (CR4) and dek1 (defective kernel 1) contribute to the aleurone differentiation during maize endosperm development (Jin et al., 2000). BETL1 (basal endosperm transfer layer) gene, which is detectable at 12 DAP, is a key marker for transfer cell differentiation (Hueros et al., 1995; Hueros et al., 1999; Hueros et al., 1999). The outer cell layer (OCL) gene family which encode putative HD-ZP IV transcription factors play a role in the protoderm formation and maintenance (Ingram et al., 1999; Ingram et al., 2000). The lipid transfer protein2 (ltp2) gene is specifically expressed in the abaxial protoderm of the scutellum and coleoptile (Sossountzov et al., 1991). Kn1 contributes to the formation of embryo SAM (Smith et al., 1995; Kerstetter et al., 1997). However, a complete gene expression profile of embryo and endosperm development in maize is still lacking.

The transcriptome is the overall set of transcribed regions of the genome. Transcriptomic analysis of the embryo and endosperm is essential for understanding the developmental process of these two tissues. Recently, development of the next-generation high-throughput DNA sequencing technologies has provided a robust tool for mapping and quantifying transcriptome, named RNA sequencing (RNA-seq). RNA-seq data are highly reproducible, with few systematic discrepancies among technical replicates (Marioni et al., 2008). RNA-seq technology has been applied to uncovering the entire transcriptome and identifying alternative splicing (AS) and novel transcribed regions (NTRs) as well as chimeric transcripts produced by trans-splicing in human, yeast, mouse, Arabidopsis and rice (Cloonan et al., 2008; Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Pan et al., 2008; Sultan et al., 2008; Wang et al., 2008; Wilhelm et al., 2008; Filichkin et al., 2010; Lu et al., 2010; Zhang et al., 2010).

So far, transcriptomic analyses have been carried out in maize with pericycle cells of the primary root (Dembinsky et al., 2007), endosperm (Prioul et al., 2008), seedling (Fu et al., 2010), leaf bundle sheath (Li et al., 2010), kernel and leaf meristem (Kakumanu et al., 2012), and shoot apical meristem (Takacs et al., 2012). However, to our knowledge, the genome-wide transcriptional profile of embryo and endosperm by RNA-Seq technology in maize has not yet been performed. To investigate the transcriptional network that governs the seed development in maize, we conducted RNA sequencing using 9-DAP embryo and endosperm,
respectively, to profile gene expression. We explored these data with informatics tools to
depict the landscape of transcriptional network as well as major metabolic activities that are
associated with embryo/endosperm development in maize.

RESULTS

Overview of the transcriptome in 9-DAP maize seeds
To obtain an overview of the transcription profile in early developing maize seeds, we
performed high-throughput RNA-seq, utilizing paired-end Illumina sequencing technology,
on poly(A)-enriched RNAs isolated from 9-DAP endosperm and embryo tissues
(Supplemental Fig. S1). Because maize seed development is highly sensitive to
environmental conditions, we presented the structural details of the embryo and endosperm
samples harvested for RNA-seq analysis (Supplemental Fig. S2). The endosperm already
completed differentiation, with the aleurone and transfer cell as well as starchy endosperm
cells formed (Supplemental Fig. S2A, B). The embryo was around 1.3 mm in length and
characterized with the emerging leaf primordia (Supplemental Fig. S2C). To further
characterize developmental stage of the maize seeds, expression profiles of storage protein
genes were specifically compiled using the RNA-seq data generated in this study
(Supplemental Table S1). Previous studies showed expression of zein genes are temporally
regulated during endosperm development (Kodrzycki et al., 1989; Woo et al., 2001). In our
RNA-seq data, two genes, 18-kD δ-zein and 19-kD α-zein, were most abundantly expressed.
Notably, 27-kD γ-zein gene, which was reported to be expressed at 10-25 DAP (Woo et al.,
2001), was not detected in our data. All these characterizations are indicative of an early
development stage of the maize seeds analyzed in this study.

Sequencing resulted in 11,074,508 and 11,495,788 paired-end reads (100-nt read length)
from endosperm and embryo, respectively. The generated reads were then aligned to the Zea
mays genome (ZmB73_RefGen_v2) (Schnable et al., 2009) using three aligners: BWA (Li
and Durbin, 2009), Bowtie (Langmead et al., 2009) and TopHat (Trapnell et al., 2009), and
the combined results were used for further analysis (see Methods). We mapped almost 87%
of the reads from both samples to the reference genome, of which nearly 86% with both ends
were correctly aligned (Table 1). As expected, most (nearly 97%) of the reads (68.1% exonic
and 28.8% junction) were mapped to annotated gene bodies, demonstrating that the majority
of the detected genes have been predicted in the maize annotation. Besides, 3.3% reads were
mapped to intergenic regions (Fig. 1A). All junction reads corresponded to 97,938 unique splicing junction sites, of which 78,592 junctions (80.2%) were identical to the annotation. The remaining 19,346 junctions (19.8%), however, were newly discovered from our RNA-seq data and have yet to be incorporated into transcript models (Fig. 1B).

The aligned reads were then subjected to Cufflinks assembly (Trapnell et al., 2010), leading to identification of 120,828 unique exons from both samples (Fig. 1C). Among these, 89,027 exons (73.7%) were consistent with boundaries of annotated exons; the remaining were newly identified ones, mostly from either intergenic (46.36%) or intragenic regions (45.56%) (Fig. 1C). The newly detected intragenic exons and junctions indicated novel transcript isoforms to be annotated or some existing gene models to be refined. However, exons and junctions detected from intergenic regions revealed the existence of some novel transcribed regions (NTRs) (see below). The number of detected exons and junctions from endosperm and embryo are also listed separately in Table 1.

To gain a global overview of transcriptome on the genome, the read distribution along each chromosome (divided into 100 windows) was constructed (Fig. 1D and Supplemental Fig. S3). The overall pattern is correlated with the gene distribution across chromosomes. However, some regions showed quite different patterns between endosperm and embryo, indicating genes annotated within these regions were differentially expressed. For example, two differentially expressed genes (GRMZM2G304745 highly expressed in endosperm and GRMZM2G080054 highly expressed in embryo) were located in such regions (Fig. 1D). The differential expression of these two genes was confirmed by QRT-PCR (Supplemental Table S9).

**Discovery of novel transcribed regions (NTRs)**

We observed that a sizeable portion of the RNA-seq reads did not fall within annotated regions of maize genome (Fig. 1). To comprehensively identify novel transcribed regions (NTRs) that are not linked to any annotated gene models, we developed a computational pipeline that integrates RNA-seq data with available annotation data and consists of several highly stringent filtering steps (Fig. 2A; see Methods): (1) more than one exon per transcript; (2) transcript length (total length of exons in base pairs) longer than 300 nt; and (3) expression level greater than 1 RPKM (reads per kilobase of exon model per million mapped reads). In addition, we required that these NTRs were at least 300 nt distant from an
annotated gene. Using these criteria, we identified 1,286 NTRs containing 2,043 multi-exon transcripts (Supplemental Data S1).

Previous studies have shown that NTRs have fewer and shorter exons, less protein-coding potentials and tissue-specific expression than annotated protein-coding genes (Bruno et al., 2010; Lu et al., 2010; Wetterbom et al., 2010; Zhang et al., 2010; Aanes et al., 2011; Graveley et al., 2011). To determine whether endospermic and embryonic NTRs have similar features, we analyzed the structure, coding potentials, and expression level of the NTRs. We found that 80.5% of the NTRs had less than five exons and only 26 transcripts (1.27%) had more than ten exons. Moreover, NTRs had fewer exons per transcript (about 3.4) than the average protein-coding genes (about 5.2) (Fig. 2B). From another point of view, we observed that the novel transcripts were on average about half of the length of protein-coding transcripts (median size of 746 bp for novel transcripts versus 1,419 bp for protein-coding transcripts) (Fig. 2C). The average expression levels (calculated as RPKM; see Methods) of NTRs were comparable to that of annotated transcripts (Fig. 2D). However, the expression pattern of NTRs was quite different between endosperm and embryo, with 45.3% (583/1286) NTRs showing > 2-fold change between these two tissues (Fig. 2E), indicating that these NTRs were tissue-preferentially expressed.

We found that a small fraction (312; 15.3%) of NTRs was supported by available EST data (greater than 80% identity and 80% coverage). Nearly 60% (1,210/2,043) of the NTRs that have open reading frames (ORFs) with the potential to encode proteins with greater than 100 amino acids (Fig. 2F) may be bona fide novel protein-coding genes. The remaining NTRs could encode small peptides, but many are likely to serve as non-coding RNAs. Besides, we observed that some (490; 24.0%) NTRs are heterochromatic, with sequence similarity (greater than 100-nt match and 80% identity) to transposable elements (TEs) (Fig. 2F), indicating presence of substantially active TEs or TE fragments in maize seed development. Although it remains to be determined whether these NTRs have any biological function, some evidence have shown that TE-related regions are widely active and have acquired specific cell functions during their evolution (Vicient et al., 2001; de Araujo et al., 2005; Ohtsu et al., 2007; Lopes et al., 2008; Picault et al., 2009), and we did observe 46 TE-associated NTRs with EST evidence on top of the support by our RNA-seq data (Fig. 2F).

An NTR was found located between GRMZM2G063253 and GRMZM2G063754. We used the RNA-seq data to infer the structure of this NTR and identified at least nine overlapping transcripts within this region (Fig. 3A). Interestingly, all these transcripts had predicted ORFs,
but lacked EST evidence, and were unambiguously divided into two groups based on junction reads, with one (containing five transcripts) being exclusively expressed in embryo and another (containing four transcripts) in endosperm (Fig. 3A). Moreover, these two groups of RNAs differed mainly in the two variant junction regions (Fig. 3A). It remains to be elucidated that whether these RNAs function as regulatory and/or as peptide-encoding RNAs in maize seed development.

To confirm the existence of the NTRs we observed in the RNA-seq data, we performed semi-quantitative RT-PCR on some of NTRs selected (Supplemental Fig. S4). Of 18 NTRs analyzed, 16 (89%) were detected either in embryo or endosperm or both, and 11 (69%) showed expression patterns consistent with the transcriptomic data. For example, \textit{NTR.g0111} was specifically expressed in the embryo (Supplemental Fig. S4).

**Alternative splicing is differentially regulated between embryo and endosperm**

Transcript alternative splicing (AS), a universal phenomenon in higher eukaryotes, is considered a key factor in increasing the diversity of transcriptome and proteome (Nilsen and Graveley, 2010; Kalsotra and Cooper, 2011). To investigate the role of AS in regulating gene expression in maize seed, we conducted surveys of transcript isoforms in the embryo and endosperm. We first estimated the number of maize multi-exonic genes with AS by calculating the fraction of genes with more than one expressed transcript divided by the number of genes with at least one splice junction in our RNA-seq data (multi-exonic genes, including NTRs). We found that 50.7% (8,556/16,878) of multi-exonic genes were alternatively spliced, a little smaller than the most recent estimate based on RNA-seq in the maize leaf transcriptome (56.4%) (Li \textit{et al.}, 2010), but larger than that in the annotation (42.6%; 12,669/29,709), further illustrating strong power of RNA-seq technology for examining splicing diversity (Wang \textit{et al.}, 2009).

We observed that some transcript isoforms were specifically expressed in either endosperm or embryo. For example, the splicing pattern of the NTR mentioned above was quite different between endosperm and embryo (Fig. 3A), and one transcript isoform (GRMZM2G146599_T03) of the gene GRMZM2G146599 was only detected in endosperm based on junction mapping analysis (Fig. 3B). Based on this observation, we systemically examined splicing differences between endosperm and embryo in the seed development. For this purpose, we carried out a computational analysis to categorize all splicing models into eight common types of AS events (Reddy, 2007; Keren \textit{et al.}, 2010) (See Methods; Fig. 3C).
We identified a total of 15,504 splicing events, involving 7,991 multi-exonic genes (Supplemental Table S2). Retained intron (RI), in which a single intron is alternatively included or spliced out of the mature mRNA via an intron-definition splicing mechanism, was the most prevalent type of AS event (30.4%; Fig. 3C), consistent with previous studies in plants (Black, 2003; Wang and Brendel, 2006; Barbazuk et al., 2008; Filichkin et al., 2010; Li et al., 2010; Lu et al., 2010; Zhang et al., 2010; Marquez et al., 2012), but in contrast to animal AS events where exon skipping (cassette exons, CE, or coordinate cassette exons, CCE) is the predominant mechanism (Sultan et al., 2008; Wang et al., 2008; Daines et al., 2011; Graveley et al., 2011). Furthermore, alternative 3’ splice sites (A3SS; 21.0%) or alternative last exons (ALE; 11.5%), and alternative 5’ splice sites (A5SS; 12.2%) or alternative first exons (AFE; 11.7%) were other types of common AS events observed in our data (Fig. 3C), in agreement with recent findings in rice (Zhang et al., 2010), Arabidopsis (Marquez et al., 2012), human (Wang et al., 2008), and Drosophila melanogaster (Daines et al., 2011; Graveley et al., 2011). In addition, a higher frequency of tissue-regulated events of each type and more junctions were detected in embryo than in endosperm (Fig. 3C, Table 1). However, it remains to be answered if these observations are simply due to numbers of mRNAs per gene detected or types of mRNAs observed. Indeed, we detected more junctions from RNA-seq data in embryo than in endosperm. Finally, by adopting a similar method (Wang et al., 2008) to assess tissue-regulated AS, we showed that 7.0% of AS events were differentially regulated in embryo and endosperm (Fisher's exact test, P-value < 0.05; Fig. 3C). Gene Ontology (GO) analysis of the genes involved in tissue-regulated AS events revealed that these genes were functionally enriched in diverse biological processes (Supplemental Table S3), indicating that transcript processing is a prevalent phenomenon in the seed development. In addition, Pearson correlation analysis revealed that NTRs showed higher level of AFE and ALE events, while protein-coding genes had higher frequency of A3SS and A5SS events. To further confirm the AS events detected in our transcriptomic data, we selected some AS events to validate with RT-PCR analysis. Among the 20 events, 15 (75%) was detected by RT-PCR.

In summary, our analysis shows that AS is an important contributor to the extensive transcriptome complexity in maize seeds and the transcript isoform abundance seems higher in embryo than in endosperm.

Trans-splicing events occur in both embryo and endosperm
Trans-splicing is one of the RNA processing mechanisms in which alternative splicing is carried out under "trans-mode" by joining exons on separate precursor transcript molecules (Nilsen and Graveley, 2010). Trans-splicing commonly occurs in unicellular organisms but much less often in higher eukaryotes. To identify trans-splicing events in the maize seed, we first used TopHat-Fusion (Kim and Salzberg, 2011) to uncover the potential fusion transcripts, and then we dug out fusion transcripts with fusion points joined by two boundaries of annotated exons from distinct gene models. We required that a candidate fusion transcript produced by trans-splicing event be supported by five spanning reads and two supporting pairs. As a result, we found a total of 46 trans-splicing events in our dataset, with nine intra-chromosomal and 37 inter-chromosomal events (Supplemental Table S4). Of the intra-chromosomal fusions, four came from neighboring genes and the other five chimaera distant genes (Fig. 4A). Although most events occurred in both embryo and endosperm, we observed that some fusion transcripts were subjected to tissue-specific regulation (Fig. 4B and 4C). Furthermore, we found that the expression levels of fusion transcripts were lower than that of their precursors, consistent with previous study in rice (Zhang et al., 2010).

We randomly chose 9 trans-spliced fusion transcripts to validate our observation in the RNA-seq data. Among them, 6 fusion transcripts were confirmed for their existence with RT-PCR analysis and sequencing. Except for TS12, which showed embryo-specific expression, the other detected trans-spliced fusion transcripts displayed expression patterns similar to the RNA-seq results (Supplemental Table S4; Supplemental Fig. S5).

**Diverse biological processes in maize seed**

RNA-seq has been proved to be a robust tool for measurement of gene expression in a manner that is more sensitive than other methods, such as traditional hybridization-based microarray technologies (Wang et al., 2009; Wilhelm and Landry, 2009). We thus used our RNA-seq data to calculate the expression levels of annotated genes, transcripts and NTRs uncovered in this study. The gene expression values were calculated as RPKM (reads per kilobase per million mapped reads) (Mortazavi et al., 2008) in each of the samples (See Methods), resulting in 68.5% (27,167) of the annotated genes with detectable expression signal (RPKM > 0; Fig. 5A; Supplemental Table S5). Although most genes were commonly expressed in both tissues, we found that the number of genes expressed in embryo was greater than in endosperm (Fig. 5A). Using RPKM > 1 as a cutoff for gene expression, we detected a total of 19,904 genes (50.2% of the annotated genes) expressed in the maize seed.
transcriptome, with 18,384 and 17,100 genes being expressed in embryo and endosperm, respectively, and with 15,580 genes (78.3%) being commonly expressed in both tissues (Fig. 5A). Furthermore, we identified 2,982 up-regulated genes (1,422 in endosperm and 1,560 in embryo, respectively) that were highly differentially expressed (with fold change > 4) between endosperm and embryo (Fig. 5B; Supplemental Table S6), representing 15.0% of the seed transcriptome of 19,904 genes identified in the 9-DAP embryo and endosperm in this study. GO enrichment analysis of the genes up-regulated in endosperm showed that many genes encode proteins functioning as nutrient reservoir activity (P < 2.7 × 10^{-23}, false discovery rate (FDR) < 2.0 × 10^{-20}) and involved in carbohydrate and storage protein metabolic process (P < 1.6 × 10^{-5}, false discovery rate (FDR) < 0.019; Fig. 5C; Supplemental Table S7). In contrast, the genes up-regulated in embryo showed enrichment of biological processes mainly acting in the nuclei, such as chromatin regulation, nucleosome organization, DNA packaging and transcriptional regulation (Fig. 5D; Supplemental Fig. S7).

We then performed an enrichment analysis, utilizing Fisher exact test (P < 0.01, FDR = 5%), to assign differentially expressed genes to functional categories with MapMan annotation (Thimm et al., 2004). Many MapMan bins were partitioned between embryo and endosperm tissues (Fig. 6A; Supplemental Table S8). For example, genes that encode enzymes for DNA synthesis/chromatin structure (e.g. histone2A, 2B, 3 and 4 and DNA replication/chromatin binding), cell organization, nucleotide metabolism, cell cycle regulation (e.g. cyclins and peptidylprolyl isomerases), DNA repair, N-metabolism, lipid (fatty acid (FA) synthesis and elongation) metabolism, amino acid metabolism, polyamine metabolism, Calvin cycle and TCA (tricarboxylic acid cycle) pathways (Supplemental Fig. S7-S12) were greatly enriched in embryo. However, genes that showed up-regulation in endosperm were mainly involved in the biological activities including major (starch synthesis and sucrose degradation) and minor carbohydrate (e.g. trehalose) metabolism, cell vesicle transport, protein metabolism (such as glycosylation, targeting and degradation), signaling pathways (G proteins, phosphinositides and leucine-rich repeat (LRR) kinases), abiotic stress and gluconeogenese/glyoxylate cycle (Supplemental Fig. S7, S13-S16).

Together, these data demonstrate that the major biochemical differences between embryo and endosperm are reflected in part by highly dynamic, coordinated and localized transitions in transcriptome abundance, and the results are consistent with the knowledge that the endosperm provides the stored nutrients to feed the embryo and the embryo develops through a series of cell divisions to pass the genetic information to its offspring.
Survey of the gene expression regulators in maize seed

Transcription factors (TFs) are important regulators for the regulation of gene expression in plant genome. To explore the accumulation of TFs in the maize seed, we examined the expression of TFs in our RNA-seq data. Of the 1,982 expressed TFs (with RPKM > 1) in the two seed tissues, 937 were differentially expressed (fold change > 2) between embryo and endosperm, and were grouped into 61 different TF families (Fig. 6B). Overall, most of the TF families showed tissue-specific expression patterns for either embryo or endosperm, respectively (Fig. 6B and Supplemental Fig. S17). Many TFs highly expressed in embryo probably have diverse functions in seed development. Among them, several regulators identified may contribute to epigenetic inheritance and reprogramming across generations (Feng et al., 2010), including genes encoding histone acetyltransferases and deacetylases, DNA methyltransferases, Bromodomain proteins and Alfin-like family. Several TF families were found to participate in cell differentiation and vascular development, such as zf-HD family (e.g. GRMZM2G328438), Myb-related (GRMZM2G158117), ARF family (GRMZM5G874163), TCP (GRMZM2G093895), and HB family (GRMZM2G017087, GRMZM2G162481 and GRMZM2G087741) as well as those implicated in hormone signalling pathways, such as B3 domain-containing factor (Stone et al., 2001), ABI3/VP1 (abscisic acid signalling), ARF and Aux/IAA (Schruff et al., 2006; Liu et al., 2007; Sreenivasulu et al., 2008; Xing et al., 2011) (Auxin signalling), GRAS (gibberellicin signalling) and ARR (cytokinin signalling) homologs. Several TF family members are likely to play a role in cell fate determination, including C2C2-YABBY (GRMZM2G167824 and GRMZM2G529859) and GeBP (e.g. GRMZM2G036966) homologs as well as in cell growth regulation and germination, including WRKY (Zhang et al., 2011) (e.g. GRMZM5G816457) and bHLH (e.g. GRMZM2G042920) homologs. On the other hand, the majority of TFs enriched in endosperm were families such as MADS, SBP, NAC, bZIP, Myb and C2C2-GATA, most of which have been implied in seed development (Sreenivasulu et al., 2008; Bemer et al., 2010; Le et al., 2010; Agarwal et al., 2011). 33 transcription factor genes were selected for confirmation of the differential expression between embryo and endosperm with real-time RT-PCR analysis. Among the 14 up-regulated genes in endosperm, 12 genes showed expression patterns consistent with RNA-seq data. All of the 11 genes up-regulated in embryo showed expression patterns more or less the same as the RNA-seq results (Supplemental Table S9).
Besides, it is notable that the relevant core regulators in small RNA pathway, such as Dicer-like (DCL), Argonaute (AGO) and members of RNA dependent RNA polymerase (RDR) gene family, showed relatively higher expression levels in embryo (Fig. 6C), indicating the potential important biological roles of this pathway for embryo development. Interestingly, we found that \textit{DCL1}, the crucial component in microRNA pathway, and \textit{RDR2} were upregulated in endosperm (Fig. 6C). Finally, we observed that most of the surveyed imprinting genes (Zhang \textit{et al}., 2011) were preferentially expressed in endosperm. As expected, 41 out of the 45 highly differentially expressed genes (with > 4 fold change) were found in endosperm (Supplemental Table S10), consistent with previous observation that gene imprinting occurs primarily in the endosperm of flowering plants (Huh \textit{et al}., 2007). These results mentioned above demonstrate that our RNA-seq data greatly reflected the major reprogramming events in the seed transcriptome for the transcriptional regulation of gene expression.

\textbf{DISCUSSION}

In this study, we performed deep transcriptomic surveys in maize seed. Using high throughput RNA sequencing technology (RNA-seq), we mapped in detail the transcriptional differences between the embryo and endosperm. Analysis of the RNA-seq data revealed a complex landscape of transcriptional network governing maize seed development.

\textbf{Embryo and endosperm are characterized with different metabolic activities}

Among the genes expressed at levels of RPKM > 1, 1520 genes were specifically expressed in endosperm and 2804 in embryo (Fig. 5A), demonstrating a more complex biological process in embryo than in endosperm. The main function of endosperm is to provide nutrients for embryo development. Four kinds of cell type can be differentiated in endosperm, including the starchy endosperm, the aleurome layer, the transfer cells and the embryo surrounding region (Olsen, 2004). The starch endosperm cells constitute the major part of endosperm and accumulate starch and storage protein (Olsen, 2004), and the genes involved in nutrient reservoir, and storage protein accumulation, carbohydrate metabolic process were found highly expressed in endosperm in this study (Fig. 5C, 5D; Supplemental Table S7, S8). The main function of embryo is to transfer the genetic information into the next generation through a series of cell division and cell differentiation. The genes involved in genetic
information transfer, such as chromatin regulation, nucleosome organization, and DNA packaging were found highly expressed in embryo, likely to maintain, at least in part, the genome fidelity (Fig. 5D; Supplemental Fig. S6 and Table S7). It’s known that the lipid is mainly stored in embryo (Barthole et al., 2012); in consistence with this, lipid biosynthesis genes were preferentially expressed in embryo (Fig. 5D).

Embryo and endosperm exhibit differential AS patterns
Gene expression regulation occurs at different levels, and transcription regulation is one of the key players. RNA-seq technology offers a powerful tool to uncover the complexity of transcriptional expression and regulation. Alternative splicing (AS), a process with exons of pre-mRNA spliced in different arrangement to produce the structurally and functionally distinct transcripts, is an essential mechanism to contribute to increasing transcriptome plasticity and proteome diversity in eukaryotes (Blencowe, 2006). It was reported that at least 42% and 48% of genes are alternatively spliced in Arabidopsis and rice, respectively (Filichkin et al., 2010; Lu et al., 2010). The splicing events function in plant development and response to environment. For example, alternative processing of the rice WAXY gene contributes to glutinous rice (Isshiki et al., 1998), and that of disease resistance gene RPS4 is dynamically regulated during resistance response (Tzafrir et al., 2007); FCA, which undergoes both polyadenylation and alternative splicing, is involved in the regulation of flowering time (Macknight et al., 2002); SR45.1 and SR45.2, the two alternatively spliced isoforms of SR45, play a major role in flower petal development and root growth, respectively (Zhang and Mount, 2009); alternative HYH transcript contributes to the increased activity of HYH-Hy5 gene pair (Sibout et al., 2006); and a C2H2-domain protein with alternatively spliced transcripts is essential for endosperm development in Arabidopsis (Lu et al., 2012). Besides, AS has also been functionally implicated in nutrient metabolism. For example, the Arabidopsis transcription factor gene IDD14 generates two splicing variants to competitively form nonfunctional heterodimers to regulate starch metabolism (Seo et al., 2011). In our RNA-seq data of maize embryo and endosperm, 50.7% of multi-exonic genes were found to undergo alternative splicing, which is higher than that in Arabidopsis and rice (Filichkin et al., 2010; Lu et al., 2010). This indicates that the splicing diversity of maize transcriptome is more complex than Arabidopsis and rice. Interestingly, 7.0% of AS events whose genes are involved in diverse biological processes were tissue-specific between embryo and endosperm (Fig. 3A, B, and C; Table 1), and more
complex AS events were observed in embryo than in endosperm (Fig. 3C; Table 1). However, it remains to be answered that whether higher transcript isoform abundance is required for embryo development than for endosperm.

**Differential regulatory mechanisms in embryo and endosperm**

In our dataset, 1982 TFs were detected in embryo and endosperm with RPKM >1, and approximately half of them (937) were differentially expressed (fold change >2) between embryo and endosperm (Fig. 6B; Supplemental Fig. S17), implicating a differential transcriptional regulation between embryo and endosperm. It was previously reported that TFs of MADS function as seed and endosperm development (Sreenivasulu *et al*., 2008; Bemer *et al*., 2010; Le *et al*., 2010; Agarwal *et al*., 2011; Kang *et al*., 2008; Hehenberger *et al*., 2012). AGL62, a MADS protein of Arabidopsis, is the direct target of FIS Polycomb group repressive complex 2 (PRC2), establishing the molecular basis for FIS PRC2-mediated endosperm cellularization (Hehenberger *et al*., 2012). In our study, preferential expression of members of the MADS TF family in maize endosperm probably also indicates an important function of these genes in maize endosperm development (Fig. 6B). TFs involved in hormone signaling such as abscisic acid signaling (ABI3/VP1), auxin signaling (ARF and Aux/IAA), gibberellin signaling (GRAS) and cytokinin signaling (ARR) were up-regulated in embryo, suggesting an important role of hormone in embryo development (Fig. 6B). Some TFs highly expressed in embryo were found to be probably involved in cell differentiation and vascular development (Fig. 6B), such as zf-HD, Myb-related, ARF family, TCP and HB family.

Several gene expression regulators highly expressed in embryo contribute to epigenetic inheritance and reprogramming across generations, including genes encoding histone acetyltransferase and deacetylase, DNA methyltransferase, Bromodomain protein and alfin-like family (Fig. 6B). Previous studies in *Arabidopsis* support the idea that transposable elements reactivated in endosperm might enhance silencing of transposable elements in embryo, and by sacrificing genomic integrity endosperm might make an epigenetic rather than genetic contribution to the progeny (Mosher and Melnyk, 2010; Hsieh *et al*., 2009). The RNA-directed DNA methylation (RdDM) pathway function as the transposable element silence (Zhang and Zhu, 2011). The genes involved in RdDM pathway, such as Dicer-like, Argonaute (Ago) and RNA-dependent RNA polymerase (RDR) gene family, were up-regulated in embryo (Fig. 6C). Indeed, in *Arabidopsis* the small RNAs generated from RdDM pathway are specifically produced from and imprinted in endosperm (Mosher *et al*., 2009).
Unlike Arabidopsis, the core regulators in small RNA pathway expressed in maize embryo implicate that the small RNAs involved in epigenetic silencing of TE element in maize are probably produced from embryo, the producing mechanism and function of small RNAs need to be further elucidated.

The results provided in this study offer an initial step toward identification of the genes which are specifically expressed in maize embryo or endosperm and shed light, at least in part, on molecular differences between embryo and endosperm at different levels including transcription, transcript splicing, and transcription regulation as well as epigenetic regulation. Of particular, it would be tempting to functionally characterize transcripts, genes, and pathways that are preferentially or specifically expressed in either of the tissues by integrating genetic, biochemical, cytological and molecular approaches to further our understanding of maize seed development.

MATERIAL AND METHODS

Plant material
The maize inbred line B73 was grown on the field in the summer of 2009 in Langfang, Hebei province, China. Ears were bagged before silk emergence. Each set of inbred kernels were generated on the same day by self-pollination. At 9 days after pollination (DAP), endosperm and embryo were isolated with tweezers and collected in 300 mM sorbitol solution with 5 mM MES (pH 5.7) from the ovules, and then were transferred into tubes, snapped frozen in liquid nitrogen and stored at -80°C until for further use. In detail, 100 mg of embryos (around 100) or endosperm were used for RNA extraction using Trizol and 10 μg of total RNA was used for cDNA synthesis.

Histological analysis of maize seed
For histological analysis, intact maize kernels or excised embryos (9 DAP) were fixed in FAA solution (5mL of formaldehyde, 5mL of acetic acid and 90mL of 50% ethanol) at 4°C for 12 h. Following fixation, materials were dehydrated in a graded ethanol series and then infiltrated and embedded in paraffin (Sigma). The embedded samples were sliced into 8 μm slices on a KD202A microtome. Sections were then counter-stained with PAS (periodic acid-Schiff's reagent, Sigma) and 1% Amido black (Sigma). Observation was conducted under an optical microscope (CX21FS1, Olympus).
RNA-seq library sequencing
All libraries were sequenced using the Illumina Genome Analyzer (GAII). We sequenced 2 lanes for maize cells, corresponding to 22.6 million reads (Table 1).

Bioinformatics analysis
To systematically characterize the transcriptome of maize seed development, we employed various publicly available tools for mapping, assembly, and quantification of transcripts, and integrated these tools with additional informatics filtering steps to enrich the results for the most robust transcriptome construction. Besides, we specifically designed a browser database using GBrowse (Stein et al., 2002) (version 1.70) to visualize all data in this study (Supplemental Fig. S1). Details of the bioinformatic analyses are provided below.

Annotation databases
The maize reference genome sequences were downloaded from the maize sequence project (B73 AGPv2, http://www.maizesequence.org/). The gene annotation information was retrieved from B73 filter gene set (release 5b).

Read alignment and assembly
RNA-seq reads from each sample were aligned to the maize reference genome using the TopHat (Trapnell et al., 2009) (version 1.3.2), a gapped aligner capable of discovery splice junctions \textit{ab initio}. TopHat finds splice junctions (without a reference annotation) using a two-step mapping process. Briefly, TopHat firstly aligned RNA-seq reads to the genome using Bowtie (Langmead et al., 2009) (without gaps) to determine a set of "coverage islands" that may represent potential exons. Using this initial mapping as well as the presence of GT-AG and GC-AG genomic splicing motifs (for which the sense strand can be reliably inferred), TopHat builted a second set of reference sequences spanning exon-exon junctions. The unmapped reads from the first alignment step were then remapped against this splice junction database to discover all the junction-spanning reads in the sample. TopHat reports aligned results from the two mapping steps in SAM format for further analysis.

Aligned reads from TopHat mapping were subjected to Cufflinks (Trapnell et al., 2010) (version 1.1.0) for \textit{ab initio} transcript assembly. Cufflinks assembled exonic and splice-junction reads into sample-specific transcriptomes using their alignment coordinates. The
exon boundaries identified by Cufflinks were further refined based on splice junction coordinates. The method was mainly based on following criteria: the start end coordinate of an exon corresponds to the end position of a splice junction, and *vice versa*. We retained all exons with both ends supported by splice junctions. For exons that have multiple 5’ or 3’ end variances, we retained the outermost coordinates.

**Discovery of splice junctions**

We prepared two databases of possible splice junctions in this study. The first one was based on the merged annotated transcripts, which led to identification of 166,552 splice junctions from annotated exons. The second database was derived from new junctions identified using TopHat based on RNA-seq data, without relying on the genome annotations. For the *ab initio* splice junction detection, we chose the parameters which allowed putting intron size between 50 bp and 500 kb. We required a minimum of 8-nt of a junction read overhang across the anchor region. Moreover, we required that a junction has at least five supported reads which map to at least two distinct positions. We produced a total of ~98,000 high confidence splice junctions supported by ~7,154,000 reads (Table 1), ~19,000 of which were novel splice junctions (Fig. 1B).

**Discovery of alternative spliced exons**

Adjacent exons of multi-exon genes were reconnected in multiple ways via the alternative splicing (AS) mechanism. Eight different types of AS events were generally recognized, including intron retention (IR), cassette exons (CE), mutually exclusive exons (MEE), coordinate cassette exons (CCE), alternative 5’ splice sites (A5SS), alternative 3’ splice sites (A3SS), alternative first exons (AFE), and alternative last exons (ALE). The AS events were identified using the similar method in previous studies (Brooks *et al.*, 2011; Graveley *et al.*, 2011). To identify exons that were differentially spliced between samples, read counts from each sample to every exclusion and inclusion isoform were calculated. For each AS event, a 2×2 table, in which reads were divided into tissue of origin (e.g., embryo versus endosperm) and read type (i.e. inclusion versus exclusion isoform), was computed. A Fisher's exact test was then used to identify differential splicing between each pair of samples and a Benjamini-Hochberg correction was performed. Those events with an adjusted *P*-value cutoff corresponding to a false discovery rate (FDR) cutoff of 5% were considered sample-specifically regulated.
Analysis of trans-splicing events

Fusion transcripts were initially discovered by TopHat-Fusion algorithm (Kim and Salzberg, 2011) with parameters "--keep-fasta-order --no-coverage-search -r 0 --mate-std-dev 80 --fusion-min-dist 100000 --fusion-anchor-length 13". We then imposed further filters to get trans-splicing events from these potential fusions: (1) fusion points were formed by the boundaries of annotated exons from distinct gene models; (2) we required five spanning reads and two supporting pairs.

Identification of novel transcripts

We proposed a bioinformatics analysis pipeline for the detection of unannotated transcripts (termed as novel transcribed regions, NTRs). The involved six main steps of the approach are outlined in Fig. 2A. Specifically, step 1 (mapping and assembly) used TopHat and Cufflinks to obtain the splice junctions and candidate exons from RNA-seq data, respectively. In step 2 (filter known annotation), the exons and junctions that overlap, or are directly adjacent (< 200 bp) to existing annotated transcripts were filtered out. The remaining exons (some representing segmental exons, dashed boxes) were merged or extended according to the supported junction (red polylines), as shown in step 3 (merged and extended exons). During step 4 (novel transcript assembly), a graph is created where nodes are merged exons (solid boxes) and orientated edges (red polylines) between two nodes represented validated junctions. The assembled transcripts were filtered further based on their size (number of exons ≥ 2 and total size of exons > 300 bp, as shown in step 5) and expression (RPKM > 1, as shown in step 6). Besides, we manually checked the refined transcript models to adjust their orientations and boundaries.

Quantification of gene expression levels

Expression levels were computed as previously described (Mortazavi et al., 2008). Briefly, the expression of a transcript in each RNA-seq sample was calculated in reads per kilobase exon model per million mapped reads (RPKM) defined as

\[
\text{RPKM} = \frac{10^9 \times C}{N \times L}
\]

where \(C\) is the number of reads (aligned reads plus junction reads) mapped to the transcript, \(L\) is the total exonic length of the transcript and \(N\) is the total number of reads mapped in the sample. Gene expression was computed at both gene and transcript levels.
For hierarchical clustering analysis, we used clustering software Cluster 3.0 (de Hoon et al., 2004) to perform complete linkage hierarchical clustering on both genes and samples, using uncentered Pearson’s correlation as distance measure. The clustering results were visualized using Java Treeview program (Saldanha, 2004).

**Statistical analysis**
All statistical analyses were performed by using the R language (http://www.r-project.org/). For gene set analysis, based on the GO term enrichment methods, agriGO (Du et al., 2010) was used to detect the significantly enriched GO terms of interested gene sets compared to the genome-wide background with an adjusted $P$-value (FDR) cutoff of 0.05.

**Real-time RT-PCR analysis**
Total RNA was isolated from 9-DAP embryo and endosperm using Trizol reagent, respectively (Invitrogen). To eliminate any residual genomic DNA, total RNA was treated with RNase-free DNase I (NEB), and 1-2 μg RNA were used to synthesize first-strand cDNAs using the RevertAid First Strand cDNA Synthesis kit (Fermentas). ACTIN primers were used to detect genomic DNA contamination. Relative quantification values for each target gene were calculated by the $2^{-\Delta\Delta C_{T}}$ method (Livak and Schmittgen, 2001) using ACTIN as an internal reference gene to compare data from different PCR runs or cDNA samples (qRT-PCR). qRT-PCR analysis provided relative changes in gene expression, with the 9-DAP endosperm normalized to a value of 1. Data were statistically analyzed using Student’ t-test. The results shown are representative of two independent experiments, and within each experiment treatments were replicated three times, unless otherwise stated. For semi-quantitative RT-PCR for transcript expression, ACTIN was used as an internal reference to equalize RNA loading into the RT-PCR reaction. After 30 cycles of amplification, PCR products were resolved on a 2% agarose gel and stained with ethidium bromide. All primers used for the RT-PCR analysis were listed in Supplemental Table S11.

**SUPPLEMENTAL MATERIAL**

**Supplemental Fig. S1.** A. Experimental steps for RNA-sequencing. B. Strategy for analyses of RNA-seq data. Genome sequencing and annotation information were collected from B73 genome sequence project (step 1). RNA-seq reads were aligned to the genome using TopHat
(Trapnell et al., 2009) (step 2). The alignments (bam-formatted files) were subjected to Cufflinks (Trapnell et al., 2010) assembly (step 3). The mapping and assembly results were used for further analysis (step 4).

**Supplemental Fig. S2.** Structure of 9-DAP maize embryo and endosperm. A, longitudinal section of a 9-DAP seed. B, magnification of part of the seed indicated with square in (A). C, a magnified embryo as in (A). Al, Aleurone. CSEn, central starchy endosperm. Em, embryo. Pe, pericarp. SAl, subaleurone layer. TC, transfer cells.

**Supplemental Fig. S3.** The distribution of RNA-seq reads (calculated as RPKM) and annotated genes (dashed red line) across maize chromosomes. See Figure 1D for legend.

**Supplemental Fig. S4.** The NTR genes validated by RT-PCR.

**Supplemental Fig. S5.** The trans-splicing transcripts validated by RT-PCR.

**Supplemental Fig. S6.** Gene Ontology (GO) enrichment analysis for the up-regulated genes from embryo. The “Molecular function” and “Cellular component” categories are shown in A and B, respectively. Corresponding to Fig. 5D.

**Supplemental Fig. S7.** Overview of cell functions. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S8.** Heatmap of the lipid synthesis pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S9.** Heatmap of the TCA metabolism pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S10.** Heatmap of the nucleotide synthesis pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S11.** Heatmap of the photosynthesis (Calvin cycle) pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S12.** Heatmap of the terpenoid metabolism pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S13.** Heatmap of the sucrose-starch metabolism pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S14.** Heatmap of the protein targeting pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

**Supplemental Fig. S15.** Heatmap of the flavonid pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.
Supplemental Fig. S16. Heatmap of the carotenoid metabolism pathway. Red, highly expressed in embryo; blue, highly expressed in endosperm.

Supplemental Fig. S17. Overview of the expression heatmap of transcriptional factors (TFs) in maize genome. The highly expressed TF families are indicated (orange triangle for embryo and violet triangle for endosperm). Red, highly expressed in embryo; blue, highly expressed in endosperm.

Supplemental Data S1. List of novel transcribed regions (NTRs) identified

Supplemental Table S1. Storage protein gene expression in embryo and endosperm.

Supplemental Table S2. Alternative splicing events

Supplemental Table S3. GO enrichment of genes involved in tissue-regulated AS events

Supplemental Table S4. Trans-splicing event analysis

Supplemental Table S5. Gene expression levels (RPKM)

Supplemental Table S6. List of up-regulated genes

Supplemental Table S7. GO enrichment analysis of differentially expressed genes

Supplemental Table S8. Functional category enrichment analysis

Supplemental Table S9. qRT-PCR validation of the differentially expressed genes

Supplemental Table S10. Differentially expressed imprinting genes

Supplemental Table S11. Primers used for RT-PCR analysis

LITERATURE CITED

Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SG, Lim AY, Hajan HS, Collas P, Bourque G, Gong Z, Korzh V, Alestrom P, Mathavan S (2011) Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res 21: 1328-1338

Abbe EC, Stein OL (1954) The Growth of the Shoot Apex in Maize: Embryogeny. American Journal of Botany 41: 285-293

Agarwal P, Kapoor S, Tyagi AK (2011) Transcription factors regulating the progression of monocot and dicot seed development. Bioessays 33: 189-202

Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res 18: 1381-1392

Barthole G, Lepiniec L, Rogowsky PM, Baud S (2012) Controlling lipid accumulation in cereal grains. Plant Science 185-186: 33-39
Bemer M, Heijmans K, Airoldi C, Davies B, Angenent GC (2010) An atlas of type I MADS box gene expression during female gametophyte and seed development in Arabidopsis. Plant Physiol 154: 287-300

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72: 291-336

Blencowe BJ (2006) Alternative Splicing: New Insights from Global Analyses. Cell 126: 37-47

Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR (2011) Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res 21: 193-202

Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, Snyder M (2010) Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. Genome Res 20: 1451-1458

Chaudhury AM, Koltonow A, Payne T, Luo M, Tucker MR, Dennis ES, Peacock WJ (2001) Control of early seed development. Annu Rev Cell Dev Biol 17: 677-699

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613-619

Daines B, Wang H, Wang L, Li Y, Han Y, Emmert D, Gelbart W, Wang X, Li W, Gibbs R, Chen R (2011) The Drosophila melanogaster transcriptome by paired-end RNA sequencing. Genome Res 21: 315-324

de Araujo PG, Rossi M, de Jesus EM, Saccaro NL, Jr., Kajihara D, Massa R, de Felix JM, Drummond RD, Falco MC, Chabregas SM, Ulian EC, Menossi M, Van Sluys MA (2005) Transcriptionally active transposable elements in recent hybrid sugarcane. Plant J 44: 707-717

de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20: 1453-1454

Dembinsky D, Woll K, Saleem M, Liu Y, Fu Y, Borsuk LA, Lamkemeyer T, Fladerer C, Madlung J, Barbazuk B, Nordheim A, Nettleton D, Schnable PS, Hochholdinger F (2007) Transcriptomic and proteomic analyses of pericycle cells of the maize primary root. Plant Physiol 145: 575-88
Diboll AG (1968) Fine Structural Development of the Megagametophyte of Zea mays Following Fertilization. American Journal of Botany 55: 797-806

Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res 38: W64-70

Feng S, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. Science 330: 622-627

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res 20: 45-58

Fu J, Thiemann A, Schrag TA, Melchinger AE, Scholten S, Frisch M (2010) Dissecting grain yield pathways and their interactions with grain dry matter content by a two-step correlation approach with maize seedling transcriptome. BMC Plant Biol 10: 63

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE (2011) The developmental transcriptome of Drosophila melanogaster. Nature 471: 473-479

Hehenberger E, Kradolfer D, Köhler C (2012) Endosperm cellularization defines an important developmental transition for embryo development. Development 139: 2031-2039

Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D (2009) Genome-wide demethylation of Arabidopsis endosperm. Science 324: 1451-1454

Hueros G, Gomez E, Cheikh N, Edwards J, Weldon M, Salamini F, Thompson RD (1999) Identification of a promoter sequence from the BETL1 gene cluster able to confer transfer-cell-specific expression in transgenic maize. Plant Physiol 121: 1143-1152

Hueros G, Royo J, Maitz M, Salamini F, Thompson RD (1999) Evidence for factors regulating transfer cell-specific expression in maize endosperm. Plant Mol Biol 41: 403-414
Hueros G, Varotto S, Salamini F, Thompson RD (1995) Molecular characterization of BET1, a gene expressed in the endosperm transfer cells of maize. Plant Cell 7: 747-757

Huh JH, Bauer MJ, Hsieh TF, Fischer R (2007) Endosperm gene imprinting and seed development. Curr Opin Genet Dev 17: 480-485

Ingram GC, Boisnard-Lorig C, Dumas C, Rogowsky PM (2000) Expression patterns of genes encoding HD-ZipIV homeo domain proteins define specific domains in maize embryos and meristems. Plant J 22: 401-414

Ingram GC, Magnard JL, Vergne P, Dumas C, Rogowsky PM (1999) ZmOCL1, an HDGL2 family homeobox gene, is expressed in the outer cell layer throughout maize development. Plant Mol Biol 40: 343-354

Isshiki M, Morino K, Nakajima M, Okagaki RJ, Wessler SR, Izawa T, Shimanoto K (1998) A naturally occurring functional allele of the rice waxy locus has a GT to TT mutation at the 5' splice site of the first intron. Plant J 15: 133-138

Jin P, Guo T, Becraft PW (2000) The maize CR4 receptor-like kinase mediates a growth factor-like differentiation response. Genesis 27: 104-116

Kakumanu A, Ambavaram MM, Klumas C, Krishnan A, Batlang U, Myers E, Grene R, Pereira A (2012) Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq. Plant Physiol 160: 846-867

Kalsotra A, Cooper TA (2011) Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet 12: 715-729

Kang IH, Steffen JG, Portereiko MF, Lloyd A, Drews GN (2008) The AGL62 MADS domain protein regulates cellularization during endosperm development in Arabidopsis. Plant Cell 20: 635-647

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11: 345-355

Kerstetter RA, Laudencia-Chingcuanco D, Smith LG, Hake S (1997) Loss-of-function mutations in the maize homeobox gene, knotted1, are defective in shoot meristem maintenance. Development 124: 3045-3054

Kim D, Salzberg SL (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12: R72

Kodrzycki R, Boston RS, Larkins BA (1989) The opaque-2 mutation of maize differentially reduces zein gene transcription. The Plant Cell 1: 105-114
Kowles R, Phillips R (1988) Endosperm Development in Maize. In KWJ G.H. Bourne, M Friedlander, eds. Academic Press. International Review of Cytology 112: 97-136

Lammeren AAMv (1986) Developmental morphology and cytology of the young maize embryo (Zea mays L.). Acta botanica Neerlandica 35: 169-188

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

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ, Goldberg RB (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A 107: 8063-8070

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760

Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR, Reidel EJ, Turgeon R, Liu P, Sun Q, Nelson T, Brutnell TP (2010) The developmental dynamics of the maize leaf transcriptome. Nat Genet 42: 1060-1067

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523-536

Liu PP, Montgomery TA, Fahlgren N, Kasschau KD, Nonogaki H, Carrington JC (2007) Repression of AUXIN RESPONSE FACTOR10 by microRNA160 is critical for seed germination and post-germination stages. Plant J 52: 133-146

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25: 402-408

Lopes FR, Carazzolle MF, Pereira GA, Colombo CA, Carareto CM (2008) Transposable elements in Coffea (Gentianales: Rubiaceae) transcripts and their role in the origin of protein diversity in flowering plants. Mol Genet Genomics 279: 385-401

Lopes MA, Larkins BA (1993) Endosperm origin, development, and function. Plant Cell 5: 1383-1399

Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Huang X, Han B (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res 20: 1238-1249
Lu X, Li Y, Su Y, Liang Q, Meng H, Li S, Shen S, Fan Y, Zhang C (2012) An Arabidopsis gene encoding a C2H2-domain protein with alternatively spliced transcripts is essential for endosperm development. J Exp Bot 63: 5935-44

Macknight R, Duroux M, Laurie R, Dijkwel P, Simpson G, Dean C (2002) Functional significance of the alternative transcript processing of the Arabidopsis floral promoter FCA. The Plant Cell 14: 877-888

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res 22: 1184-1195

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628

Mosher RA, Melnyk CW (2010) SiRNAs and DNA methylation: seedy epigenetics. Trends Plant Sci 15: 204-210

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349

Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463: 457-463

Ohtsu K, Smith MB, Emrich SJ, Borsuk LA, Zhou R, Chen T, Zhang X, Timmermans MC, Beck J, Buckner B, Janick-Buckner D, Nettleton D, Scanlon MJ, Schnable PS (2007) Global gene expression analysis of the shoot apical meristem of maize (Zea mays L.). Plant J 52: 391-404

Olsen OA (2004) Nuclear endosperm development in cereals and Arabidopsis thaliana. Plant Cell 16 Suppl: S214-227

Olsen OA, Linnestad C, Nichols SE (1999) Developmental biology of the cereal endosperm. Trends Plant Sci 4: 253-257

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40: 1413-1415
Picault N, Chaparro C, Piegu B, Stenger W, Formey D, Llauro C, Descombin J, Sabot F, Lasserre E, Meynard D, Guiderdoni E, Panaud O (2009) Identification of an active LTR retrotransposon in rice. Plant J 58: 754-765

Prioul JL, Méchin V, Lessard P, Thévenot C, Grimmer M, Chateau-Joubert S, Coates S, Hartings H, Kloiber-Maitz M, Murigneux A, Sarda X, Damerval C, Edwards KJ (2008) A joint transcriptomic, proteomic and metabolic analysis of maize endosperm development and starch filling. Plant Biotechnol J 6: 855-69

Randolph L (1936) Developmental morphology of the caryopsis in maize. Journal of Agricultural Research 53: 881-916

Reddy AS (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol 58: 267-294

Saldanha AJ (2004) Java Treeview--extensible visualization of microarray data. Bioinformatics 20: 3246-3248

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Dechaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrana D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lameli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q,
Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112-1115

Schruff MC, Spielman M, Tiwari S, Adams S, Fenby N, Scott RJ (2006) The AUXIN RESPONSE FACTOR 2 gene of Arabidopsis links auxin signalling, cell division, and the size of seeds and other organs. Development 133: 251-261

Seo PJ, Kim MJ, Ryu JY, Jeong EY, Park CM (2011) Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. Nature Communications 2, doi:10.1038/ncomms1303.

Sibout R, Sukumar P, Hettiarachchi C, Holm M, Muday GK, Hardtke CS (2006) Opposite root growth phenotypes of hy5 versus hy5 hyh mutants correlate with increased constitutive auxin signaling. PLoS Genetics 2, e202.

Smith LG, Jackson D, Hake S (1995) Expression of knotted1 marks shoot meristem formation during maize embryogenesis. Developmental Genetics 16: 344-348

Sossountzov L, Ruiz-Avila L, Vignols F, Jolliot A, Arondel V, Tchang F, Grosbois M, Guerbette F, Miginiac E, Delseny M, Puigdomènèch P, Kader JC (1991) Spatial and temporal expression of a maize lipid transfer protein gene. Plant Cell 3: 923-933

Sreenivasulu N, Usadel B, Winter A, Radchuk V, Scholz U, Stein N, Weschke W, Strickert M, Close TJ, Stitt M, Graner A, Wobus U (2008) Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. Plant Physiol 146: 1738-1758

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12: 1599-1610

Stone SL, Kwong LW, Yee KM, Pelletier J, Lepiniec L, Fischer RL, Goldberg RB, Harada JJ (2001) LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. Proc Natl Acad Sci U S A 98: 11806-11811

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321: 956-960

31
Takacs EM, Li J, Du C, Ponnala L, Janick-Buckner D, Yu J, Muehlbauer GJ, Schnable PS, Timmermans MC, Sun Q, Nettleton D, Scanlon MJ (2012) Ontogeny of the maize shoot apical meristem. Plant Cell 24:3219-3234

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37: 914-939

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

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515

Tzafrir I, McElver JA, Liu CM, Yang LJ, Wu JQ, Martinez A, Patton DA, Zhang XC, Gassmann W (2007) Alternative splicing and mRNA levels of the disease resistance gene RPS4 are induced during defense responses. Plant Physiology 145: 1577-1587

Vernoud V, Hajduch M, Khaled A-S, Depège N, Rogowsky PM (2005) Maize embryogenesis. Maydica 50: 469-484

Vicient CM, Jaaskelainen MJ, Kalendar R, Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125: 1283-1292

Wang BB, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci U S A 103: 7175-7180

Wang ET, Sandberg R, Luo S, Kherbtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470-476

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63

Wetterbom A, Ameur A, Feuk L, Gyhlensten U, Cavelier L (2010) Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. Genome Biol 11: R78

Wilhelm BT, Landry JR (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48: 249-257
Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453: 1239-1243

Woo YM, Hu DW, Larkins BA, Jung R (2001) Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. The Plant Cell 13: 2297-317

Xing H, Pudake RN, Guo G, Xing G, Hu Z, Zhang Y, Sun Q, Ni Z (2011) Genome-wide identification and expression profiling of auxin response factor (ARF) gene family in maize. BMC Genomics 12: 178

Zhang CQ, Xu Y, Lu Y, Yu HX, Gu MH, Liu QQ (2011) The WRKY transcription factor OsWRKY78 regulates stem elongation and seed development in rice. Planta 234: 541-554

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

Zhang H, Zhu JK (2011) RNA-directed DNA methylation. Curr Opin Plant Biol 14: 142-147

Zhang XN, Mount SM (2009) Two alternatively spliced isoforms of the Arabidopsis SR45 protein have distinct roles during normal plant development. Plant Physiology 155: 1450-1458.

Zhang M, Zhao H, Xie S, Chen J, Xu Y, Wang K, Guan H, Hu X, Jiao Y, Song W, Lai J (2011) Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. Proc Natl Acad Sci U S A 108: 20042-20047

FIGURE LEGENDS

Fig. 1. Overview of the maize seed transcriptome of embryo and endosperm at 9 DAP. A. Pie chart indicating the proportion of RNA-seq reads assigned to maize annotated genomic features. Two samples were combined for analysis. B. Shared and unique splicing junctions from annotation and support with junction reads. C. Known and novel exons assembled from RNA-seq reads. The newly identified exons can be assigned to four categories according to their genomic locations: antisense to known genes (blue), intergenic regions (brown),
annotated intronic regions (violet) and overlapped with known exons but with distinct boundaries (exonic, green). D. The distribution of RNA-seq reads (calculated as RPKM) and annotated genes (dashed red line) across the chromosome 2. The whole chromosome was divided into 100 bins for visualization. Two tissue-specific genes (GRMZM2G304745 for endosperm and GRMZM2G080054 for embryo) are shown on the right panel. See Supplemental Fig. S3 for the distribution from other chromosomes.

**Fig. 2. Discovery and description of novel transcribed regions (NTRs).** A. Overview of the RNA-seq-based transcript construction pipeline that was employed to identify NTRs. The main six steps of the approach are numbered on the left. Reads were mapped to the maize genome using TopHat and initial transcripts were separately assembled by Cufflinks in each sample (step 1). After filtering the annotation-overlapped transcripts (step 2), the novel exons and junctions (step 3) were used to reconstruct the final transcript structures (step 4). These transcripts were further subject to size selection (step 5, multiple-exon with length greater than 300 bp) and expression filter (step 6, RPKM > 1). This leads to identifying 1,286 NTRs with 2,043 distinct transcript isoforms. Detailed methods see Methods section. B. The distribution of exon numbers of novel transcripts identified in this study. The majority have two exons. The number of average exons for NTRs and protein-coding genes (PC) is also shown. C. The length distribution of annotated genes and newly identified transcribed regions. Novel transcripts (median size 746 bp) are much smaller than annotated maize transcripts (median size 1,419 bp). D. The expression pattern of novel transcripts compared with annotated transcripts. E. Plot showing the different expression levels of the identified NTRs between embryo (orange) and endosperm (violet). For visualization, data points are sorted according to the expression levels from embryo. F. NTRs supported by ORF and EST evidences. ORFs were identified by getorf in the EMBOSS packages, and EST supported criteria: identity > 80% and coverage > 80%. See Fig. 3A for example of an NTR.

**Fig. 3. The splicing dynamics of maize seed development.** A. An example illustrating a newly identified transcribed region (Chr6: 123,740,814-123,743,530) and its developmentally regulated splicing events. The transcript models were quite different between embryo (orange) and endosperm (violet), as indicated by the two groups of variant junctions (colored shading). The involved three types of alternative splicing (AS) events (defined in C) in this region are indicated (middle box). The peak and junction reads of RNA-seq data are shown below. B.
Alternative splicing of GRMZM2G146599. Three types of AS events (defined in C) from this gene loci are listed below that RNA-seq data track. Black arrows indicate a novel junction that was not incorporated into existing gene models identified within the gene, and red arrow indicates the tissue-specific transcript GRMZM2G146599_T3 was expressed from endosperm. C. Analysis and categorization of eight common types of AS events. Diagram represents the eight different AS event types (left) and the numbers of events are shown as bar chart (right). Total events included annotated and newly identified splicing events. The numbers of events with both isoforms supported by RNA-seq data from embryo (orange bars) and endosperm (violet bars) are indicated. Events detected as tissue-regulated (Fisher’s exact test, \( P < 0.05 \)) are shown with red bars.

Fig. 4. Analyses of trans-splicing events in maize seed transcriptome. A. Schema representing the three types of transcript fusion events (inter-chromosomal, intra-chromosomal between neighbouring genes and intra-chromosomal between distant genes). The number of each type of events is shown in pie chart. B. Number of trans-splicing events identified in embryo (orange) and endosperm (violet). C. Examples showing three types of events. Top: a fusion event that occurred between two inter-chromosomal genes; middle: a fusion event that occurred between two neighbouring genes; bottom: a fusion event that occurred between two distant genes in the same chromosome. The numbers of spanning reads and paired-end reads that supported fusion points are shown in the inset table.

Fig. 5. The expression and functional maps of maize seed transcriptome. A. The expression map of maize seed. Bar chart shows the number of expressed genes between embryo (orange) and endosperm (violet) at different expression levels. The inserted venn diagram depicts the number of genes with RPKM > 1. B. The highly differentially expressed genes with at least a 4-fold change. Gene Ontology (GO) enrichment analysis for the up-regulated genes from embryo and endosperm are shown in C and D, respectively. The figure keys are shown in the box. In C, the enriched GO “Biological process” and “Molecular function” categories are shown (no “Cellular component” terms for up-regulated genes from endosperm). In D, only the “Biological process” category is shown. See Supplemental Fig. S6 for “Molecular function” and “Cellular component” categories.
Fig. 6. The developmental regulation map of maize seed transcriptome. A. Functional category (modified MapMan bins) enrichment analysis of differentially expressed genes (fold change > 2) between embryo and endosperm. Fisher exact test was used to investigate the significant level of each category for up-regulated genes from each tissue. Proteins involved in the categories of chromatin regulation and transcriptional regulation (arrowed) were listed in B in detail. Red, highly significant enrichment (diamond); gray, non-significant (circle). B. Distribution of up-regulated transcription factor families between embryo (orange) and endosperm (violet). The number of genes in each family is shown. C. The expression pattern of small RNA pathway-related regulators (including Dicer-like (DCL), Argonaute (AGO) and members of RNA-dependent RNA polymerase (RDR) gene family) in the embryo and endosperm tissues. Only expressed genes (with RPKM > 1) are shown. mop1, MEDIATOR OF PARAMUTATION1. In A and B, PS, photosynthesis; CHO, carbohydrate; OPP, oxidative phosphatase; HB, homeobox domain; zf-HD, Zinc-finger-homeodomain; histone DAase, histone deacetylase; histone ATase, histone acetyltransferases; DNA MT, DNA methyltransferases; PcG, Polycomb Group; Methyl BD, Methyl binding domain.
| Category              | Endosperm   | Embryo      |
|-----------------------|-------------|-------------|
| **RNA-seq data**      |             |             |
| Raw reads             | 11,074,508  | 11,495,788  |
| Mapped reads<sup>1</sup> | 9,601,041 (86.7%) | 10,016,167 (87.1%) |
| Mapped PE reads       | 8,237,322 (85.8%) | 8,548,737 (85.3%) |
| **Analysis**          |             |             |
| Junctions (Novel)<sup>2</sup> | 72,709 (12,898) | 78,809 (11,937) |
| Exons (Novel)<sup>3</sup> | 87,863 (17,088) | 96,478 (16,777) |
| Novel transcripts     | 1,939 (1,194) | 1,824 (1,095) |
| (NTRs/genes)          |             |             |

<sup>1</sup> Reads were aligned to the maize genome by BWA/Bowtie and Tophat.  
<sup>2</sup> Detected by Tophat.  
<sup>3</sup> Assembled by Cufflinks. PE, paired-end; NTRs, novel transcribed regions.
Fig. 1. Overview of the maize seed transcriptome of embryo and endosperm at 9 DAP. A. Pie chart indicating the proportion of RNA-seq reads assigned to maize annotated genomic features. Two samples were combined for analysis. B. Shared and unique splicing junctions from annotation and support with junction reads. C. Known and novel exons assembled from RNA-seq reads. The newly identified exons can be assigned to four categories according to their genomic locations: antisense to known genes (blue), intergenic regions (brown), annotated intronic regions (violet) and overlapped with known exons but with distinct boundaries (exonic, green). D. The distribution of RNA-seq reads (calculated as RPKM) and annotated genes (dashed red line) across the chromosome 2. The whole chromosome was divided into 100 bins for
visualization. Two tissue-specific genes (GRMZM2G304745 for endosperm and GRMZM2G080054 for embryo) are shown on the right panel. See Supplemental Fig. S3 for the distribution from other chromosomes.
Fig. 2. Discovery and description of novel transcribed regions (NTRs). A. Overview of the RNA-seq-based transcript construction pipeline that was employed to identify NTRs. The main six steps of the approach are numbered on the left. Reads were mapped to the maize genome using TopHat and initial transcripts were separately assembled by Cufflinks in each sample (step 1). After filtering the annotation-overlapped transcripts (step 2), the novel exons and junctions (step 3) were used to reconstruct the final transcript structures (step 4). These transcripts were further subject to size selection (step 5, multiple-exon with length greater than 300 bp) and expression filter (step 6, RPKM > 1). This leads to identifying 1,286 NTRs with 2,043 distinct transcript isoforms. Detailed methods see Methods section. B. The distribution of exon numbers of novel transcripts identified in this study. The majority have two exons. The number of average exons for NTRs and protein-coding genes (PC) is also shown. C. The length distribution of annotated genes and newly identified transcribed regions. Novel transcripts (median size 746 bp) are much smaller than annotated maize transcripts (median size 1,419 bp). D. The expression pattern of novel transcripts compared with annotated transcripts. E. Plot showing the different expression levels of the identified NTRs between embryo (orange) and endosperm (violet). For visualization, data points are sorted according to the expression levels from embryo. F. NTRs supported by ORF and EST evidences. ORFs were identified
by getorf in the EMBOSS packages, and EST supported criteria: identity > 80% and coverage > 80%. See Fig. 3A for example of an NTR.
Fig. 3. The splicing dynamics of maize seed development. A. An example illustrating a newly identified transcribed region (Chr6: 123,740,814-123,743,530) and its developmentally regulated splicing events. The transcript models were quite different between embryo (orange) and endosperm (violet), as indicated by the two groups of variant junctions (colored shading). The involved three types of alternative splicing (AS) events (defined in C) in this region are indicated (middle box). The peak and junction reads of RNA-seq data are shown below. B. Alternative splicing of GRMZM2G146599. Three types of AS events (defined in C) from this gene loci are listed below that RNA-seq data track. Black arrows indicate a novel junction that was not incorporated into existing gene models identified within the gene, and red arrow indicates the tissue-specific transcript GRMZM2G146599_T3 was expressed from endosperm. C. Analysis and categorization of eight common types of AS events. Diagram represents the eight different AS event types (left) and the numbers of events.
are shown as bar chart (right). Total events included annotated and newly identified splicing events. The numbers of events with both isoforms supported by RNA-seq data from embryo (orange bars) and endosperm (violet bars) are indicated. Events detected as tissue-regulated (Fisher’s exact test, $P < 0.05$) are shown with red bars.
Fig. 4. Analyses of trans-splicing events in maize seed transcriptome. A. Schema representing the three types of transcript fusion events (inter-chromosomal, intra-chromosomal between neighbouring genes and intra-chromosomal between distant genes). The number of each type of events is shown in pie chart. B. Number of trans-splicing events identified in embryo (orange) and endosperm (violet). C. Examples showing three types of events. Top: a fusion event that occurred between two inter-chromosomal genes; middle: a fusion event that occurred between two neighbouring genes; bottom: a fusion event that occurred between two distant genes in the same chromosome. The numbers of spanning reads and paired-end reads that supported fusion points are shown in inset tables.
Fig. 5. The expression and functional maps of maize seed transcriptome. A. The expression map of maize seed. Bar chart shows the number of expressed genes between embryo (orange) and endosperm (violet) at different expression levels. The inserted venn diagram depicts the number of genes with RPKM > 1. B. The highly differentially expressed genes with at least a 4-fold change. Gene Ontology (GO) enrichment analysis for the up-regulated genes from embryo and endosperm are shown in C and D, respectively. The figure keys are shown in the box. In C, the enriched GO “Biological process” and “Molecular function” categories are shown (no “Cellular component” terms for up-regulated genes from endosperm). In D, only the “Biological process” category is shown. See Supplemental Fig. S6 for “Molecular function” and “Cellular component” categories.
Fig. 6. The developmental regulation map of maize seed transcriptome. A. Functional category (modified MapMan bins) enrichment analysis of differentially expressed genes (fold change > 2) between embryo and endosperm. Fisher exact test was used to investigate the significant level of each category for up-regulated genes from each tissue. Proteins involved in the categories of chromatin regulation and transcriptional regulation (arrowed) were listed in B in detail. Red, highly significant enrichment (diamond); gray, non-significant (circle). B. Distribution of up-regulated transcription factor families between embryo (orange) and endosperm (violet). The number of genes in each family is shown. C. The expression pattern of small RNA pathway-related regulators (including Dicer-like (DCL), Argonaute (AGO) and members of RNA-dependent RNA polymerase (RDR) gene family) in the embryo and endosperm tissues. Only expressed genes (with RPKM > 1) are shown. mop1, MEDIATOR OF PARAMUTATION1. In A and B, PS, photosynthesis; CHO, carbohydrate; OPP, oxidative phosphatase; HB, homeobox domain; zf-HD, Zinc-finger-homeodomain; histone DAase, histone deacetylase; histone ATase, histone acetyltransferases; DNA MT, DNA methyltransferases; PcG, Polycomb Group; Methyl BD, Methyl binding domain.