Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome

Guojie Zhang, Guangwu Guo, Xueda Hu, et al.

*Genome Res.*, published online March 19, 2010
Access the most recent version at doi:10.1101/gr.100677.109
Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome

Guojie Zhang,1,2,5 Guangwu Guo,1,5 Xueda Hu,1,5 Yong Zhang,1,5 Qiye Li,1 Ruiqiang Li,1,3 Ruhong Zhuang,1 Zhike Lu,1 Zengquan He,1 Xiaodong Fang,1 Li Chen,1 Wei Tian,1 Yong Tao,4 Karsten Kristiansen,3 Xiuqing Zhang,1 Songgang Li,1 Huanming Yang,1 Jian Wang,1,6 and Jun Wang1,3

1Beijing Genomics Institute at Shenzhen, Shenzhen 518000, China; 2CAS-Max Planck Junior Research Group, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China; 3Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark; 4Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

Understanding the dynamics of eukaryotic transcriptome is essential for studying the complexity of transcriptional regulation and its impact on phenotype. However, comprehensive studies of transcriptomes at single base resolution are rare, even for modern organisms, and lacking for rice. Here, we present the first transcriptome atlas for eight organs of cultivated rice. Using high-throughput paired-end RNA-seq, we unambiguously detected transcripts expressing at an extremely low level, as well as a substantial number of novel transcripts, exons, and untranslated regions. An analysis of alternative splicing in the rice transcriptome revealed that alternative cis-splicing occurred in ~33% of all rice genes. This is far more than previously reported. In addition, we also identified 234 putative chimeric transcripts that seem to be produced by trans-splicing, indicating that transcript fusion events are more common than expected. In-depth analysis revealed a multitude of fusion transcripts that might be by-products of alternative splicing. Validation and chimeric transcript structural analysis provided evidence that some of these transcripts are likely to be functional in the cell. Taken together, our data provide extensive evidence that transcriptional regulation in rice is vastly more complex than previously believed.

[Supplemental material is available online at http://www.genome.org. RNA-seq and small RNA data, and digital gene expression sequence reads from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession nos. GSE16631 and GSE16507, respectively.]

Over the past several decades, rice has been intensively studied for its agricultural importance. Besides providing more than one-fifth of the calories consumed worldwide (Smith and Pluciennik 1995), rice is also an excellent model system for monocots. There are a growing number of essential resources available for its analysis—in particular the complete genome sequence for both cultivated rice strains, Oryza sativa subsp. indica and subsp. japonica (Goff et al. 2002; Yu et al. 2002). With the availability of complete genome sequences and a variety of other tools for rice genome analysis, extensive functional genomics work is underway to identify and determine the activity of all the functional elements in the rice genome (for review, see Jung et al. 2008). The success of this research is dependent on the availability of deep and detailed rice transcriptome data.

The transcriptome represents a comprehensive set of transcribed regions throughout the genome. Understanding the dynamics of the transcriptome is essential for unveiling functional elements of the genome and interpreting phenotypic variation produced by combinations of genotypic and environmental factors. Recent studies have shown that massively parallel sequencing technology is more sensitive at detecting lowly expressed transcripts compared to traditional SAGE (serial analysis of gene expression) and microarray hybridizations (Cloonan et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Sultan et al. 2008; Wang et al. 2008; Wilhelm et al. 2008). The application of massively parallel sequencing in transcript profiling indicates a far greater complexity of the eukaryotic transcriptome than previously believed, owing to the presence of extensive alternative splicing and the rapidly increasing number of newly identified transcripts (Lister et al. 2008; Nagalakshmi et al. 2008; Sultan et al. 2008; Wang et al. 2008, 2009; Wilhelm et al. 2008; Hillier et al. 2009; Filichkin et al. 2010). The complicated nature of the transcriptome poses challenges for obtaining accurate estimates of its size and determining changes in expression activity at different times in different organs.

Here, we present transcriptome profiles with nucleotide resolution representing different stages and organs of cultivated rice. We used deep RNA sequencing (RNA-seq), which allowed us to rapidly identify and analyze the vast majority of the transcriptomes in a cost-effective way. It also allows us to analyze their extensive level of alternative splicing or those lowly expressed regions. Our analysis uncovered numerous new transcripts that are expressed at very low levels, potentially functional noncoding RNAs, newly identified exons, and untranslated regions (UTRs). In this context, we also developed a new method to determine the optimal sequencing depth for such detailed analyses, and this method can be generalized for transcriptome analyses of any genome.

We also found that a far greater amount of alternative splicing occurs than previously shown, in addition to significantly...
improving current genome annotation. Detailed data on these alternative transcripts provide information about splice site junctions and underlying splicing mechanisms. Most intriguingly, we identified a large number of chimeric transcripts. These fusion events add yet another level of complexity to transcriptional regulation in plants.

Results

Rice transcriptome obtained for eight organs

To obtain a global view of the rice transcriptome and gene activity at nucleotide resolution, we performed high-throughput RNA-seq, using Illumina sequencing technology, on poly(A)-enriched RNAs from eight different rice organs: callus, seedling shoot, seedling root, leaf at the tillering stage, leaf at the flowering stage, booting panicle, flowering panicle, and filling panicle. To minimize the likelihood of systematic biases in transcriptome sampling, multiple cDNA libraries were prepared and data were generated from six paired-end libraries with insert sizes ranging from 100 to 500 base pairs (bp). We conducted in-depth sequencing by paired-end RNA-seq on two of the eight organs, callus and booting panicle, and carried out single-end RNA-seq on all eight organs.

In total, we acquired more than 410 million paired-end reads of 35–75 bp in length for callus and booting panicle (218 and 199 million, respectively). In addition, we obtained >5-million single-end reads from each of the eight organs (Supplemental Table S1). The total length of the reads was over 28 gigabases (Gb), representing about 67-fold of the rice genome size. We aligned all these short reads onto the reference Oryza sativa subsp. indica genome, and found that about 73% of the reads can be uniquely mapped to the genome, including 11% junction reads that spanned splice junction sites (see Methods; Supplemental Table S2). Our deep sequencing of the rice transcriptome covered 99.7% (32,959) of the available rice full-length cDNA data (Kikuchi et al. 2003), and resulted in having ~38.1% of the rice genome covered by at least one read.

The use of deep-sequencing RNA-seq technology provided an unprecedented chance to address the fundamental questions of how to estimate the transcriptome size, and how many transcripts should be sequenced for a given transcriptome. To address these questions, we developed a new statistical method using a combination of the digital gene expression (DGE) profile method (’t Hoen et al. 2008) and the Lander–Waterman model (Lander and Waterman 1988) to determine the fraction of transcriptome coverage that one could obtain at a particular sequencing depth. At first, DGE profiling was performed to evaluate the abundance of all transcripts in a transcriptome by considering the real number and the level of different transcript isoforms (Supplemental Table S3). We proposed that the location of sequenced reads on the transcriptome followed the discrete uniform distribution. Then, we introduced the Lander–Waterman model to deduce the growth curve of percentage of transcriptome coverage with increasing sequencing depth (see Methods). To evaluate the applicability of this model on transcriptome sampling, we used the transcriptome data of deep-sequenced booting panicle sample to demonstrate how the model fits the experimental data. Assuming the uniform distribution of reads, we estimated the size of covered transcriptome by summing up the number of covered nucleotides in each transcript copy (see Methods). We found that coverage at different sequencing depths estimated by experimental data fit well with the simulation in the Lander–Waterman model (Supplemental Fig. S1), suggesting the suitability of this model in transcriptome analysis. Using this method, we were able to estimate that 5 Gb of sequencing reads is sufficient to cover 90% of the genes that are expressed at an average tag density of 1 per million (Supplemental Fig. S1).

Rice transcriptome analysis reveals an extensive number of novel transcripts

The total number of nontransposable element (non-TE)-related genes of indica rice is currently estimated to be 38,130, of which 21,071 have been empirically validated by the presence of RNA transcripts, including ESTs and full-length cDNA sequences (Ohyanagi et al. 2006). We quantified the expression level of the genes in our data by counting the number of reads per kilobase per million mapped reads (RPKM). This calculation normalized the read density measurement, and therefore, could be used in comparisons within and between different organ samples. About 27,200 genes showed expression with 95% confidence in our sampled organs, represented at ≥0.78 RPKM (Supplemental Fig. S2).

Extensive read mapping and clustering revealed 38,650 transcript units (TUs), larger than 150 bp in size and supported by at least four reads per base. We compared our data with TUs detected by tilling array analysis (Li et al. 2007), and found that more than 60% of the transcript regions identified from tilling array were present in our TUs, and more than 72% of the TUs defined by RNA-seq were not detected by tilling array analysis (Fig. 1). In comparison with the tilling array results and known gene models, we detected an additional set of 7232 novel TUs that had not been previously identified (Supplemental Table S4). Of the novel TUs, 1133 were located in transposable element (TE) regions, indicating a substantial activity of TEs in rice. Among the 6099 non-TE novel TUs, we found about 59% with lengths shorter than 500 bp (Fig. 2C), but 3769 contain multiple exons. Based on the presence of an ORF larger than 50 amino acids within the transcripts, we found that about 2707 (44.4%) of the non-TE novel TUs are likely to represent protein-coding genes (Fig. 2A), and 1087 of these contain known protein domains (Supplemental Table S5).

To investigate whether the identified noncoding RNAs were potentially functional (e.g., microRNA precursors), we carried out deep sequencing of small RNAs (20–30 bp) derived from the same organ as used above (see Methods). Systematic analysis that integrated a robust computational approach and exhaustive sequencing...
Alternative splicing in plants

Alternative splicing (AS) plays a major role in the generation of proteomic and functional complexity in higher organisms (Black 2003; Matlin et al. 2005; Blencowe 2006; Reddy 2007). To explore potential AS events, we carried out computational analyses to determine all theoretical splicing junctions and then identified sequence reads that mapped to these regions (see Methods). We found that up to 9042 rice genes had undergone AS, displaying 23,800 AS events, representing the primary seven known types of AS models (Fig. 3; Supplemental Table S10; Black 2003; Matlin et al. 2005). In our data, we found that 78% of all indica genes contained two or more exons, and 42.4% of these produced two or more AS isoforms. Therefore, in total, about 33% of all rice genes are alternatively spliced. We also found that about 58% of the rice AS-related genes undergo multiple AS events producing a variety of transcripts from a single gene (Supplemental Fig. S4), illustrating the extremely high complexity of transcriptome regulation. Expression analysis showed that 59% of the AS events were organ-specific, indicating a strong association of AS events with organ-specific regulation and a major role for AS in the functional complexity of plants (Table 1). Analysis of Gene Ontology (GO) enrichment for AS genes, using all rice genes as background, revealed that AS gene functions were enriched in many different biological processes, indicating the prevalent influence of AS in rice gene expression (Supplemental Table S11).

Compared with previously characterized AS in japonica, based on full-length cDNA and ESTs analyses (Wang and Brendel 2006), our present ratio of alternative-spliced genes is much higher (see Methods). As for 4793 AS genes annotated by ESTs/cDNAs that have homology found in indica, 3290 of them (68.6%) were included in our list (Fig. 4). About 57.1% of the rice AS genes annotated by our RNA-seq had not been detected previously. We also searched for close homologs of indica rice AS genes in Arabidopsis by reciprocal BLAST best hits (see Methods), and found that 1694 AS genes in rice have close homologs in Arabidopsis that also are alternatively spliced. Indeed, the same AS type is conserved in 1023 homologous pairs.

Gene structure analyses of AS genes were performed to determine the exact position of the splicing events within a gene transcript and to assess the likelihood that the resulting transcripts are functional. We found that about 76% of all AS events occurred in protein-coding regions. In 27.5% of these AS events the coding frame was retained (Supplemental Table 12), whereas 48.5% caused frame shifts or produced premature termination codons (PTC) that can destabilize transcripts by nonsense-mediated mRNA decay (NMD) (Leeds et al. 1991; Reddy 2007). The remaining 24% of AS events occurred in UTRs. Although they do not affect the coding frame, they create a variety of UTRs that may play a role in gene regulation.

Given that the identification and refinement of an extensive number of UTR boundaries could be useful for interpreting cis-and trans-regulatory factors that modify gene expression, we investigated the presence of one important translational control signal in the UTR: A small open reading frame (uORF) upstream of the start codon (Kawaguchi and Bailey-Serres 2002). We found that 5837 genes (20.4%) contained uORFs (see Methods; Supplemental Table S9), indicating this translational regulation mechanism may be prevalent in rice. Further analyses of these novel and extended UTRs should greatly aid research directed at identifying all known regulatory elements in the rice genome.
We carried out an analysis of the rice AS events to determine which of the main seven AS mechanisms (Black 2003; Matlin et al. 2005) is most common in rice. As previously reported by others (Ner-Gaon et al. 2004; Wang and Brendel 2006), we found that intron retention, in which a single intron is alternatively included or spliced, is the primary type of AS. This is in contrast to human and yeast AS events where exon skipping is the most prevalent mechanism (Sultan et al. 2008; Wang et al. 2008). Intron retention occurred in 47% of all AS events in rice, while exon skipping only constituted 25%. Based on previous analyses in plants, and now in rice, intron retention seems to be a specific and common AS feature in plants—rather than being technical noise from spurious sequences. The high frequency of intron retention in rice corroborates with previous findings from cDNA/ESTs alignment studies in plants (Ner-Gaon et al. 2004; Wang and Brendel 2006), which were experimentally confirmed in Arabidopsis (intron retention at 50% or more) (for data on the remaining AS events, see Table 1; Ner-Gaon et al. 2004).

Our analysis also shows that the average size of retained-introns is about 183 bp, which is significantly shorter than that of rice introns in general (470 bp, \( P < 2.2 \times 10^{-16} \); Supplemental Fig. S5), indicating that intron retention may be related to intron size. This supports the hypothesis that organisms that typically have small introns use an intron-definition splicing mechanism, which results in intron retention; whereas other organisms that have large introns use an exon definition mechanism, which results in exon skipping (Nakai and Sakamoto 1994). Thus, the differences in AS frequency and most common AS-type between plants and animals may reflect underlying differences in pre-mRNA splicing regulation.

### Chimeric transcripts produced by trans-splicing

Transcription fusion events, potentially resulting in fusion proteins from different genes, have been reported to be quite common in humans (Akiva et al. 2006; Parra et al. 2006), but these events have only been sporadically demonstrated in plants (Koller et al. 1987; Kawasaki et al. 1999). We examined our sequence data for the presence of chimeric transcripts (CT) that would be a hybrid of transcripts from two different genes, in rice callus and booting panicle organs for which we had paired-end sequences. We used a stringent criterion, which required each CT to be supported by five or more independent paired-end reads that had different start positions, to determine the genome-wide level of such transcript fusion events. We then carried out step-by-step filtering to eliminate potentially artificial fused RNAs (see Methods).

It has been reported that the standard Illumina paired-end sequencing protocol can produce chimeric paired reads (Quail et al. 2008); therefore, some of the chimeric transcripts may be artificially produced by the experimental manipulation. Nevertheless, some mechanisms have been reported to produce real fusion transcripts in the cell, including trans-splicing (Horiuchi and Aigaki 2006; Kapranov et al. 2007), read-through (Akiva et al. 2006; Parra et al. 2006), and trans-splicing (Horiuchi and Aigaki 2006; Kapranov et al. 2007).

### Table 1. The number of each type of AS event and the number of organ-specific AS events.

<table>
<thead>
<tr>
<th>AS events</th>
<th>ES</th>
<th>IR</th>
<th>A5SS</th>
<th>A3SS</th>
<th>MXE</th>
<th>AFE</th>
<th>ALE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total organ</td>
<td>6048(25.4%)</td>
<td>11,297(47.5%)</td>
<td>1871(7.9%)</td>
<td>3483(14.6%)</td>
<td>567(2.4%)</td>
<td>258(1.1%)</td>
<td>276(1.2%)</td>
</tr>
<tr>
<td>Organ specific</td>
<td>3296(26.2%)</td>
<td>5800(46.0%)</td>
<td>1057(8.4%)</td>
<td>1637(13.0%)</td>
<td>306(2.4%)</td>
<td>240(1.9%)</td>
<td>262(2.1%)</td>
</tr>
</tbody>
</table>

ES: Exon-skipping; IR: intron retention; A5SS: alternative 5’ spliced site; A3SS: alternative 3’ spliced site; MXE: mutually exclusion exon; AFE: alternative first exon; ALE: alternative last exon.
2006), and transcriptional slippage (Li et al. 2009). Here, we focused on the trans-splicing model, which is similar to cis-splicing, as both require canonical splice sites (GU-AG). In contrast to cis-splicing, which occurs within a single transcript, the splicing factors for trans-splicing join the 5’ splice donor of one pre-mRNA to the 3’ splice acceptor of a pre-mRNA from a different gene (Horiuchi and Aigaki 2006). To identify fusion transcripts that resulted from trans-splicing, we enumerated all possible exon–exon combinations from fusion pairs to detect junction reads. We required that a candidate chimeric transcript predicted to be produced by trans-splicing be supported by multiple chimeric short reads that spanned the junction of the two genes with at least five bases that were a perfect match at each site. To exclude false fusion transcripts that might be produced by genome rearrangement, we also required that the precursors of trans-splicing should be represented by the transcript units. Using this strategy, we identified a total of 234 rice fusion events likely to be produced by trans-splicing. Of these, 173 were interchromosomal, where the two precursor RNAs came from different chromosomes. The remaining 61 were intrachromosomal, with 25 chimeras that occurred between neighboring genes and 36 chimeras between distant genes (Fig. 5A–C; Supplemental Table S13).

To validate our trans-spliced fusion transcripts, we randomly chose 25 candidates for RT-PCR confirmation. In all, 17 (68%) of the 25 fusions could be experimentally validated, indicating the authenticity of these chimeric RNAs (Fig. 5A–C; Supplemental Table S14). Because many of the fusion transcripts are present only at a very low level, we believe that our validation rate is an underestimate; thus, the percentage of fusion transcripts in the transcriptome is likely to be higher.

Our analysis revealed several characters of trans-splicing: (1) Expression analysis showed that the fusion transcripts were expressed at a lower level than its precursors (Supplemental Fig. S6). (2) We found that over 92% of fusion pairs were generated from precursors where at least one in the pair also engages in alternative cis-splicing. This ratio is significantly higher than the ratio expected by chance, and it is consistent with previous reports on trans-splicing in animals (Chatterjee and Fisher 2000; Dorn et al. 2001; Finta and Zaphiropoulos 2002; Horiuchi et al. 2003). Thus, it is possible that either alternative cis-splicing mechanisms are involved in and facilitate the formation of trans-spliced transcripts or that trans-splicing is a by-product of alternative cis-splicing (Maniatis and Tasic 2002). (3) We found a significant positive correlation between the efficiency of trans-splicing and that of cis-splicing ($r = 0.68$; Supplemental Fig. S7). (4) The exonic splice enhancer was found in many fused exons, suggesting it may play as important a role in regulation of trans-splicing as in cis-splicing (Supplemental Fig. S8).

(5) We found that 58 (25%) of the analyzed CTs contained intact open reading frames (ORFs) that integrated amino acids derived from separate proteins. One can expect that many trans-splicing events may be the “splicing-noise” or by-product of cis-splicing.
(Maniatis and Tasic 2002). However, some may also represent a novel form of alternative splicing with an expression level that is similar to that of normal alternative cis-splicing (for a more detailed analysis, see Supplemental material).

**Discussion**

Here, we have presented the most complete transcriptome data sets currently available along with additional analyses to investigate mechanisms involved in genome evolution and to assess the complexity of transcriptional regulation. In this study, we combined RNA-seq, which allows rapid and efficient deep sequencing, with paired-end methods that allow comprehensive detection of read position and linkage. This offers an unprecedentedly powerful tool for interrogation of complex transcriptomes and it has several distinct advantages for detecting very rare RNA isoforms. Our transcriptome data profoundly improve existing gene annotation in a variety of ways, including evidence for numerous novel transcripts, novel exons and 5′/3′ UTRs, and extension of known 5′/3′ UTRs.

To assess the level of complexity in the rice transcriptome, we investigated the frequency of the different forms for alternative splicing. We identified highly complex patterns of AS in 9042 rice genes. We further identified 12,598 AS events that were organ-specific, which is similar in percent to what has been observed in humans (Wang et al. 2008). Overall, our data indicate that 33% of all rice genes undergo alternative splicing. This number is substantially higher than that previously reported for rice (21.2%) and Arabidopsis (21.8%) (Wang and Brendel 2006). The amount of AS in rice, however, still remains significantly lower than that predicted in higher animals, especially in humans, where AS is estimated to occur in 86% of all genes (Wang et al. 2008).

One of the most intriguing findings is that the rice transcriptome was found to contain a large number of chimeric transcripts, indicating that transcript fusion is much more common than previously believed. The chimeric fusion events further expand the complexity of the rice transcriptome. Detailed sequence analyses of the CTs indicate that trans-splicing is likely to be an important mechanism, forming fusion transcripts. We also find that many of these fusion transcripts contain open reading frames bringing together specific protein domains from different genes, which may result in proteins with novel functional interactions. Although an assessment of the true functionality of fusion transcripts in rice will require additional analyses, the potential for post-transcriptional combination of sequences from different genes in rice provides new means to produce more complex exon shuffling (Gilbert 1978).

Taken together, our work provides extensive new knowledge of the rice transcriptome, and the use of this resource has enabled us to examine mechanisms for transcript diversification that aid in understanding transcriptome complexity and should augment future research in many agricultural areas.

**Methods**

**Organ collection and RNA isolation**

Oryza sativa L. ssp. indica cv. 9311 was used in all experiments. Eight organs from different developmental stages were collected from this strain of rice, including callus, root at seedling stage of 14 d, shoot at seedling stage of 14 d, flag leaves at tillering stage, flag leaves at flowering stage, panicle at booting stage, panicle at flowering stage, and panicle at filling stage. Total RNA was isolated with TRIzol (Invitrogen) from each sample according to the manufacturer's instructions. It was treated with RNase-free DNase I for 30 min at 37°C (New England BioLabs) to remove residual DNA.

**cDNA library preparation and sequencing**

Beads with oligo(dT) were used to isolate poly(A) mRNA. First-strand cDNA was synthesized using random hexamer-primer and reverse transcriptase (Invitrogen). The second-strand cDNA was synthesized using RNase H (Invitrogen) and DNA polymerase I (New England BioLabs). Then the cDNA libraries were prepared according to Illumina’s protocols. In total, we constructed one individual single-end cDNA library each for eight organs, and sequenced them on the Illumina GA platform for 35 cycles. Six additional paired-end cDNA libraries were constructed from RNAs of callus and booting panicle with insert-sizes ranging from 170 to 500 bp (Supplemental Table 1). The paired-end libraries were sequenced for 44–75 bp.

**Small RNA library preparation**

Small RNA sized at 18–30 nt was gel-purified from total RNA for each organ. 5′ and 3′ Illumina RNA adapter were ligated to the small RNA molecules. The adapter-ligated small RNAs were then reverse transcribed and amplified for 15 cycles with PCR. Then eight small RNA libraries were constructed and sequenced according to the Illumina GA platform sequencing protocols.

**mRNA tag library preparation and Illumina sequencing**

The mRNAs from booting panicle and callus were separately isolated with oligo(dT) ligated beads, and were then reverse transcribed into double-stranded cDNA. The ds cDNAs were digested by the restriction enzymes NlaIII and MmeI. After ligation with sequencing adapters, the molecules were amplified by PCR. The two mRNA tag libraries were sequenced on the Illumina GA platform.

**Mapping short reads to the rice genome and annotated gene**

The indica rice genome and annotated gene set (The Beijing Gene Finder, http://bgf.genomics.org.cn/) were downloaded from the Rice Information System (http://rice.genomics.org.cn/rice/index2.jsp), and 37,029 japonica rice full-length cDNAs were collected from the KOME (knowledge-based oryza molecular biological encyclopedia) site (http://cdna01.dna.affrc.go.jp/cDNA/, release 2007/09/14). The cDNAs were aligned to the indica rice genome and those with identities higher than 80% were retained for further analysis. A nonredundant gene set of indica rice was created by merging the sequences of BGF predicted gene and cDNAs and then removing the smaller one if two transcripts have at least 100 bp overlapping.

After removing reads containing sequencing adapters and reads of low quality (reads containing Ns > 5), we aligned reads to the indica genome using SOAP (Li et al. 2008) allowing up to two mismatches. Reads that failed to be mapped were progressively trimmed off one base from the 3′-end and mapped to genome again until a match was found. For paired-ends reads, we set insert size between paired reads at 1 bp–10 kb to allow reads spanning introns of different sizes. A similar strategy was used to align reads to the nonredundant gene set, but the insert length range was restricted to 1 kb for paired-end reads mapping.

**Illumina DGE tag annotation**

We obtained ~8.9 and 8.1 M Illumina tags for RNA samples of booting panicle and callus, respectively, with equal sample
amounts used in paired-end RNA-seq. We filtered out low quality tags (containing Ns) and adaptor sequences. The DGE tags, which consist of the CATG restriction enzyme digested site and an additional 17 bp from each transcript, were aligned onto the transcriptome including known genes and novel transcripts (see paragraph below). Ambiguous tags with multihits were excluded. The expression abundance of transcripts was measured by the number of tags mapped.

**Statistical model of transcriptome sampling**

We developed a statistical model by combining DGE and the Lander–Waterman theory (Lander and Waterman 1988; Port et al. 1995) to address the problems of how to estimate the transcriptome size and how to estimate the fraction of transcriptome coverage with different depths of sequencing.

First, we estimated the transcriptome size by summing up the length of all genes multiplied by the corresponding expression level of each gene measured by DGE:

\[
S_t = \sum_{i=1}^{n} E_i \times L_i, 
\]

where \(S_t\) represent the size of the transcriptome, \(E\) denotes the sequence abundance determined by DGE, \(L\) means the length of the gene, and \(n\) is the gene number. Because the \(E\)-value depends on the expression levels of different transcript isoforms, it can represent the copy number of all transcripts and \(S_t\) can be considered as an estimate of the size of all the annotated transcripts that are expressed. Here we only considered the annotated genes, because some of the novel transcripts may be transcriptional noise.

The real sequencing depth of the transcriptome can be estimated by the following formula:

\[
D_t = \frac{S_t}{D_t},
\]

where \(D_t\) is the sequencing depth of the annotated transcriptome, and \(S_t\) is the total size of sampling reads.

Under the assumption that reads within mRNA follow a discrete uniform distribution, the fraction of transcriptome coverage, \(C_t\), can be estimated by

\[
C_t = \frac{\sum_{i=1}^{n} E_{o_{ij}}}{S_t},
\]

In Equation 3, \(o_{ij}\) is defined as

\[
o_{ij} = \begin{cases} 1, & \frac{C_{ij}}{E_i} \geq 1, \\ \frac{C_{ij}}{E_i} < 1, & \end{cases}
\]

where \(C_{ij}\) represents the number of reads covered at the position \(j\) of gene \(i\) in a sequencing depth. \(E_i\) denotes the copy number of gene \(i\), and \(o_{ij}\) is the covered coefficient factor of nucleotide \(j\) of gene \(i\). If the number of reads covered on nucleotide \(j\) is higher than the copy number of gene \(i\), this suggests that this nucleotide is covered in each copy; thus, the coefficient factor \(o_{ij}\) is equal to 1. However, if the \(C_{ij} < E_i\), assuming the reads are randomly distributed, the coefficient factor is equal to the ratio of \(C_{ij}\) to \(E_i\).

Applying Equations 2–4, the fraction of transcriptome coverage for certain depths of sequencing can be determined. We used this method to survey the transcriptome of booting panicle using a data set of cDNA genes and short reads mapping to cDNAs. The relationship between coverage and depth of a transcriptome is shown in Supplemental Fig. S1a.

Theoretically, it can be assumed that the reads fit a discrete, uniform distribution. Therefore, we applied the Lander and Waterman (LW) theory to calculate the coverage of a sequenced transcriptome. According to the LW theory, the fraction of transcriptome coverage based on a certain sequencing depth, \(d\), can be implied by the following formula:

\[
C_{L,W} = 1 + e^{-2d\theta}(d\theta - 1)/2 - e^{-d\theta} - d\theta - d\theta^2, 
\]

where

\[
\theta = \frac{\text{overlap between reads}}{\text{reads length}}, \sigma = 1 - \theta,
\]

in which the minimum value of \(\theta\) is 0 as the minimum overlap between two reads was equal to 0. We observed that the distribution of transcriptome coverage relative to the sequencing depth accessed by the LW theory fits fairly well with the real coverage distribution (see Supplemental Fig. S1a).

We can also estimate the percentage of represented cDNA genes under different sequencing depths by considering the \(E_i = 1\) in Equations 1 and 3 (see Supplemental Fig. S1b).

**Normalization of gene expression level s based on RNA-seq**

The gene expression levels based on RNA-seq was measured as numbers of reads per kilobase of exon region in a gene per million mapped reads (RPKM) (Mortazavi et al. 2008). Only reads from single-end libraries of eight organs were used to calculate the RPKM value, in order to avoid bias of assessment for booting panicle and callus for which additional paired-end reads were obtained. The cutoff value for determining the background expression level was the 95% confidence limit of the RPKM for all genes in eight organs (Supplemental Fig. S2).

**Organ specificity of gene expression**

Organ specificity of gene expression in rice was measured as the organ-specific index \(\tau\) value developed by Yanai et al. (2005) with minor modification:

\[
\tau = \frac{\sum_{i=1}^{n} \left(1 - \frac{S_{i,j}/S_{i,j,\text{max}}}{n - 1}\right)}{n}
\]

In the formula above, \(n\) is the number of rice organs surveyed; \(S_{i,j}\) is the RPKM value of \(i\) gene in \(j\) organ, and \(S_{i,j,\text{max}}\) is the highest RPKM of gene \(i\) in the \(n\) organs. We set the \(S\) value as zero for genes expressed at a level lower than background.

**Detection of transcript units and discovery of novel transcripts**

A contiguous expression region with each base supported by at least four reads and with a size larger than 35 bp was considered as a transcriptionally active region (TAR). The TARs that were joined by at least one set of paired-end reads were connected into a transcript unit (TU). The TUs with length <150 bp or average expression of <10 reads per base were excluded in the further analysis. In total, we detected 38,650 TUs in our sequenced samples. To discover novel transcribed regions, we compared our TUs with both an annotated gene model and TARs detected by tiling array (Li et al. 2007). Finally, 7232 TUs that were not overlapping with genomics transcripts in the above two data sets were defined as novel TUs. Some of these TUs are produced by transposons or retrotransposons. We blasted these TUs against the repetitive DNA sequences downloaded from Repbase Update (http://www.girinst.org/repbase/update/index.html, release 14.02) to remove TE. TUs consisting of <50% of repetitive sequences were then screened by Augustus (Stanke et al. 2006) using gene model parameters based on training with known maize genes to predict putative protein-coding genes.
Identification of miRNA in rice

The small RNA reads produced by the Illumina 1G Genome Analyzer were subjected to several filtering processes: (1) Filter out low quality reads; (2) trim three prime adaptor sequence by a modified dynamic programming algorithm; (3) remove adaptor contaminations formed by adaptor ligation; and (4) retain only short trimmed reads of sizes from 18 to 30 nt. Finally, 4,866,269 clean small RNA reads representing 2,682,621 distinct reads remained.

To annotate and classify small RNAs into different categories and identify novel miRNA candidates, we performed the following analyses: (1) We excluded small RNA reads that might be from known noncoding RNAs by comparing with known noncoding RNAs (only consider rRNA, tRNA, snRNA, and snoRNA) deposited in the Rfam database (http://www.sanger.ac.uk/Software/Rfam/, release 9.0); (2) small RNAs aligned to repetitive regions were filtered out; and (3) small RNA reads assigned to exonic regions were also excluded. After excluding small RNA reads falling within the above categories (Supplemental Table S6), the rest were subjected to “MIREAP” (https://sourceforge.net/projects/mireap/), which identifies miRNA candidates according to the canonical hairpin structure and sequencing data. The identified miRNAs that were absent in the miRBase database (http://www.microrna.sanger.ac.uk/sequences/, release 13.0) were regarded as novel miRNAs.

UTR analysis based on RNA-seq data

To identify the UTRs of annotated genes, we searched the boundary of the transcribed region in both ends of genes. The upstream open reading frames (uORF) were screened in 5’ UTRs with computational methods. We predicted the uORF upstream of the start codon and required a potential uORF should be 10–50 bp.

Determination of alternative splicing using RNA-seq

To identify all potential splice sites, we searched for putative donor sites (GT on the forward strand or AC on the reverse strand) and acceptor sites (AG on the forward strand and CT on the reverse strand) inside intron regions. We then enumerated all possible pairs of donor sites and acceptor sites and detected 1,709,842 potential splice junctions.

All reads that could not be matched to the genome were aligned onto the splice junctions to find junction reads. We require a junction site to be supported by at least two unambiguously mapped reads with nonrepetitive match position within the splice junction and also having a minimum of five bases on both sides of the junction. In total 124,242 junction sites were identified.

Adjacent exons were joined into multiexon genes via the spliced junctions. We empirically classified the AS events into seven different types according to the structures of exons. These seven types include exon skipping (ES), intron retention (IR), alternative 5’ splice site (ASSS), alternative 3’ splice site (A3SS), mutually exclusive exons (MXE), alternative first exons (AFE), and alternative last exons (ALE) as described by Wang et al. (2008).

Gene enrichment analysis

GO enrichment was performed using Cytoscape software (http://www.cytoscape.org, version 2.5.2) with Bingo plugin (http://www.psb.ugent.be/cdb/papers/BINGO/, version 2.3). Hypergeometric test with Benjamini & Hochberg false discovery rate (FDR) were performed using the default parameters to adjust the P-value.

Comparison of AS genes in indica, japonica rice, and Arabidopsis

The annotated AS proteins in japonica rice and Arabidopsis were aligned against all indica rice proteins, and vice versa, by using BLAST (Altschul et al. 1997) with an E-value of 1 × 10^-20 as the cutoff. The reciprocal uniquely best hits were considered as closely homologous pairs. If both homologous genes show the same AS type, they were considered as conserved AS. The genes undergoing AS of AFE and ALE, undefined in japonica rice and Arabidopsis, were excluded for this analysis.

Identification of chimeric transcripts

Connections between two separate genes sequences supported by more than five nonredundant paired-end RNA-seq reads were considered as candidate fusion events. In total, 2,772,516 paired-end reads spanned two separate genes, defining 2540 candidate fusions events. To avoid false-positives, we removed 349 neighboring transcript pairs that are likely to be produced by alternative splicing. Some of the neighboring transcripts represent a shared chromosomal region, and reads aligned to overlapping regions may also create artificial fusions. Therefore, we removed 82 neighboring pairs that have overlapping regions. Additionally, since gene pairs that contain highly similar sequences may cause ambiguous alignments, we required that candidate gene pairs be supported by more than five paired-end reads aligned to unique regions. Using these stringent criteria, we found 1356 highly reliable chimeric fusions in rice.

One possible mechanism that can lead to chimeric transcripts is trans-splicing, in which the splicing factors join the 5’ splice donor of one pre-mRNA to the 3’ splice acceptor of another pre-mRNA. We enumerated all theoretical splice junctions between any two fusion pairs. Short reads that cannot be mapped to the rice genome, rice cDNA, or junctions of alternative cis-splicing were aligned to the potential trans-splicing junctions. Finally, we detected 28,446 chimeric reads spanning the trans-splicing junctions, representing 234 fusion genes that clearly seem to be created by a trans-splicing mechanism.

Real-time PCR validation for trans-splicing

To validate fusion transcripts that seem to be produced by trans-splicing, we performed qPCR using Power SYBR Green Mastermix (AB) in an Applied Biosystems Step One Plus Real Time PCR System. The RNAs from callus and panicle used in the RNA-seq study were reverse transcribed into cDNAs. The fusion specific and precursor specific oligonucleotide primers are listed in Supplemental Table S14. The ACTB gene was used as a control in these experiments. The qPCR was performed using the following program: 95°C for 3 min; 32 cycles of 95°C for 15 sec, melting temperature for 15 sec, and 72°C for 45 sec; and 72°C for 5 min.

Acknowledgments

This project is supported by the Chinese Academy of Science (KSCX2-YW-023, GJHZ0701-6), the National Natural Science Foundation of China (30725008), and the Chinese 973 program (2007CB815701, 2007CB815703, 2007CB815705). This project is also funded by the Shenzhen Municipal Government and the Yantian District local government of Shenzhen, the Danish Platform for Integrative Biology, and the Ole Ramer grant from the Danish National Science Research Council (272-07-0196). Laurie Goodman edited the manuscript.
References


Received September 15, 2009; accepted in revised form February 2, 2010.