Available online at www.sciencedirect.com
ScienceDirect
Genomics 91 (2008) 186–194
GENOMICS
www.elsevier.com/locate/ygeno

Case study for identification of potentially indel-caused alternative expression isoforms in the rice subspecies *japonica* and *indica* by integrative genome analysis ☆

Fengxia Liu [a,1], Wenying Xu [b,*,1], Lubin Tan [c], Yongbiao Xue [b], Chuanqing Sun [c], Zhen Su [a,*]

[a] *State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100094, China*
[b] *Laboratory of Molecular and Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences and National Center for Plant Gene Research, Beijing 100080, China*
[c] *Department of Plant Genetics and Breeding and State Key Laboratory of Agrobiotechnology, China Agricultural University, Beijing 100094, China*

Received 25 July 2007; accepted 3 October 2007
Available online 26 November 2007

Abstract

Alternative splicing (AS) is one of the most significant components of the functional complexity of the eukaryote genome, increasing protein diversity, creating isoforms, and affecting mRNA stability. Recently, whole genome sequences and large microarray data sets have become available, making data integration feasible and allowing the study of the possible regulatory mechanism of AS in rice (*Oryza sativa*) by erecting and testing hypotheses before doing bench studies. We have developed a new strategy and have identified 215 rice genes with alternative expression isoforms related to insertion and deletion (indel) between subspecies *indica* and subspecies *japonica*. We did a case study for alternative expression isoforms of the rice peroxidase gene LOC_Os06g48030 to investigate possible mechanisms by which indels caused alternative splicing between the *indica* and the *japonica* varieties by mining of array data together with validation by RT-PCR and genome sequencing analysis. Multiple poly(A) signals were detected in the specific indel region for LOC_Os06g48030. We present a new methodology to promote more discoveries of potentially indel-caused AS genes in rice, which may serve as the foundation for research into the regulatory mechanism of alternative expression isoforms between subspecies.

Keywords: InDel; Alternative splicing; Indica; Japonica; Polyadenylation signal

Alternative splicing (AS) is an important and common feature in eukaryotic gene expression [1–11]. There are different types of AS, such as exon skipping, intron retention, alternative donor site, alternative acceptor site, alternative terminus, etc. Since Walter Gilbert proposed AS phenomena in eukaryotes in 1978 [12], more and more genes with different expression isoforms have been reported, especially from high-throughput sequencing genome and expressed sequence tag (EST) data [13–22], and microarray technologies, such as exon junction array and tiling array, have been used to detect alternative isoforms [23]. To date, on the basis of genome-wide analysis, about 60% of human genes are considered to be AS [3]. In 2006, Wang and Brendel reported 22% of *Arabidopsis* genes and 21.2% of rice genes with AS isoforms identified by comparison between genomic sequences and EST/cDNA sequences [24].

With the expanding number of alternative isoforms expected with the increased availability of EST and cDNA data, it is necessary to investigate populations of AS transcripts and study the possible mechanisms underlying the generation of AS isoform diversity. Conventionally, AS is thought to occur within the same species under different environmental and/or development conditions. But insertion and deletion (indel) of subspecies may contribute to the generation of AS isoform diversity during evolution, leading to expanded populations of AS transcripts in rice and other organisms. It has been reported that indel of a genome sequence may lead to AS isoforms. In 1992, Kenneth R. Luehrsen and Virginia Walbot added a non-intron sequence to

☆ Sequence data from this article have been deposited with the GenBank Data Library under Accession Nos. EF990900 and EF990901.

* Corresponding authors.
E-mail addresses: wyxu@genetics.ac.cn (W. Xu), zhensu@cau.edu.cn (Z. Su).

[1] These authors contributed equally to this work.

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.10.001

two maize introns and used a transient expression assay to explore the impact of inserted sequences on splicing [25]. They reported that transposable element insertion into or near introns can cause AS events. Two other groups reported that mobile retrotransposons can induce AS of the host gene upon insertion [26,27].

Rice (*Oryza sativa*) is the staple food for almost half of the world population, and it is a model organism for studies of crop plants. The entire rice genome determined by high-quality sequencing is freely available [28–33]. A genome-wide comparative analysis was conducted for DNA sequences of two major cultivated rice subspecies, *O. sativa* L. ssp. *indica* and *O. sativa* L. ssp. *japonica* [34,35]. The variations affect gene structures and may cause intraspecific phenotypic adaptation [36]. The availability of public microarray data makes it feasible to use microarray data mining and a comparative genomics approach for identifying rice AS possibly due to indel. The rice microarray database provides a powerful tool with which to identify different rice gene expression patterns, predict possible gene functions, and analyze genotyping by data mining. There are two sets of rice tissue/organ-specific microarray data available in the GEO data sets, GSE7951 and GSE6893 (http://www.ncbi.nlm.nih.gov/geo/). These two microarray data sets were compiled for different research purposes but they used the same platform, the Affymetrix GeneChip rice genome array (GPL2025). GSE6893 was generated by Dr. Jitendra P. Khurana's laboratory in India and the array samples from *indica* variety IR64 were used for identifying the genes expressed differentially during various stages of reproductive development [37]. GSE7951 was generated by Dr. Yongbiao Xue's laboratory in China and the array samples from *japonica* variety Nipponbare were used for genome-wide gene expression profiling in rice stigma [38]. These data sets provide a good opportunity for global comparison of gene expression levels of AS genes between the two rice subspecies *japonica* and *indica.*

We developed a new strategy to do data mining through Affymetrix microarray data for predicted AS genes with multiple probe sets that were differentially expressed in *indica* and *japonica* varieties, and searched for possible indel regions between *japonica* variety Nipponbare and *indica* variety 93-11 contigs. Furthermore, to identify possible mechanisms of indel-caused AS transcripts between *indica* and *japonica* varieties, we conducted a case study for alternative expression isoforms of

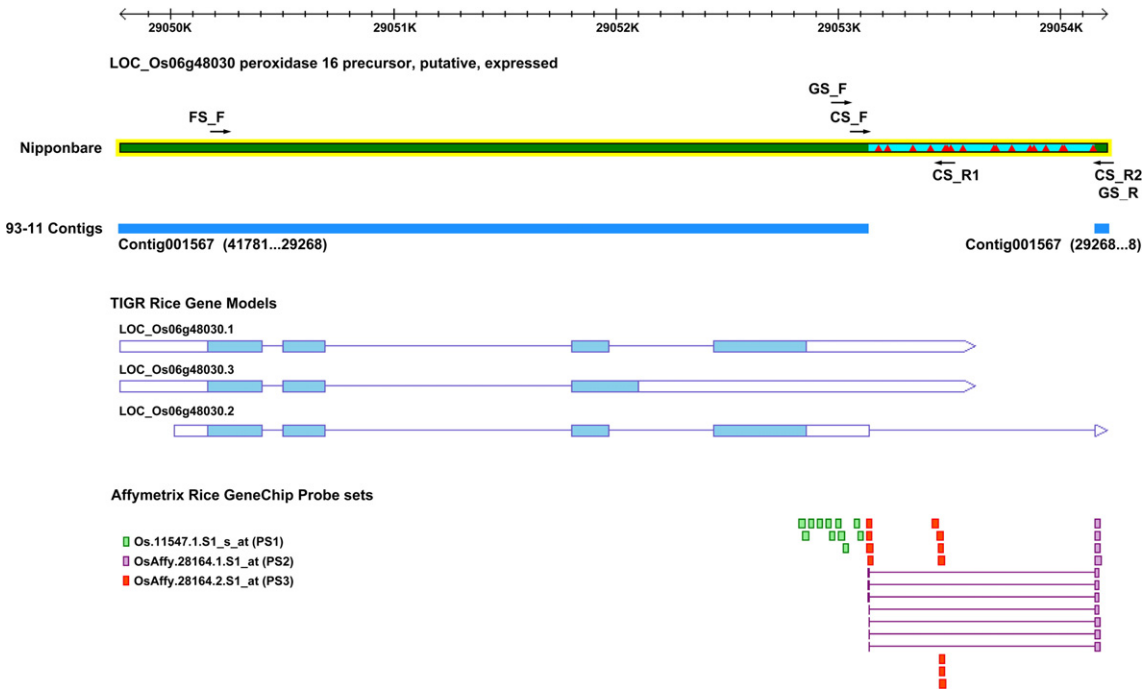


Fig. 1. A scheme for gaining rice genomic information in the region of LOC_Os06g48030. Rice peroxidase 16 precursor, LOC_Os06g48030, is located on rice chromosome 6, from 29,049,770 to 29,054,208 bp in the pseudomolecule. The position of the region is shown in the topmost track. The green bar with a yellow outline represents the genome region of LOC_Os06g48030 in *japonica* (Nipponbare), the light blue line with red triangles indicates the indel region between the *japonica* and the *indica* genomes; the red triangles show the locations of poly(A) signals in the indel region. The blue bars represent the contigs of the *indica* (93-11) genome; contig 001567 has a gap in the genome region of LOC_Os06g48030 compared with the *japonica* (Nipponbare) genome. The three tracks with blue and white boxes represent the models of gene LOC_Os06g48030. There are three alternative expression isoforms predicted by the TIGR Web site, LOC_Os06g48030.1, LOC_Os06g48030.2, and LOC_Os06g48030.3, and each track indicates the structure of one isoform. The boxes represent the exons and the blue bars represent the coding region. The small colored boxes represent the positions of the probes in the three probe sets in the Affymetrix GeneChip rice whole genome. Each color indicates one probe set: green indicates Os.11547.1.S1_s_at, named PS1; purple indicates OsAffy.28164.1.S1_at, named PS2; red indicates OsAffy.28164.2.S1_at, named PS3. The purple lines connecting purple boxes represent the location of the probe in the exon junction. The arrows beside the *japonica* (Nipponbare) genome region represent the positions of primers designed for RT-PCR experiments. The primer pair CS_F and CS_R1 is for RT-PCR of expression corresponding to PS3, with a 305-bp product. The primer pair CS_F and CS_R2 is for RT-PCR of expression corresponding to PS2, with a 101-bp product. The primer pair FS_F and CS_R2 is for RT-PCR of full-length cDNA of LOC_Os06g48030.2. The primer pair GS_F and GS_R is for PCR of the indel region between the *japonica* and the *indica* genomes, with a 1231-bp product for *japonica* and a 213-bp product for *indica.*

the rice peroxidase gene LOC_Os06g48030. Peroxidases are known to respond to leaf senescence [39] and have a role in increasing a plant's defense against pathogens [40]. Genomic analysis indicated that there is a large indel region between Nipponbare and BGI 93-11 contigs for LOC_Os06g48030, whose two probe sets (OsAffx.28164.1.S1_at and OsAffx.28164.2.S1_at) are located in/near the indel region. Further studies using comparative genomics, RT-PCR validation, and genotyping analysis revealed that AS isoforms of LOC_Os06g48030 associate with indel between the *indica* and the *japonica* varieties. This is the first study to use microarray data mining and a comparative genomics approach for identifying alternative splicing possibly due to indel between rice subspecies. We describe a new methodology to predict indel-related AS genes in rice and other plant species globally. This new methodology will promote more discoveries in potentially indel-related AS genes in rice and provide a foundation for research into the regulatory mechanisms of alternative expression of isoforms between subspecies.

Results

Identification of alternative expression isoforms possibly caused by indel within genes, such as LOC_Os06g48030, through genome analysis and microarray comparison of transcript profiles between the rice subspecies japonica and indica

In TIGR release version 5 for rice pseudomolecules, 6497 rice AS genes with 10,431 additional gene models were curated on the basis of the rice EST and full-length cDNA sequences (http://www.tigr.org/tdb/e2k1/osa1/expression/alt_spliced.info.shtml). The rice genome browser (http://www.tigr.org/tigr-scripts/osa1_web/gbrowse/rice/) shows that some of those genes have indel variations between contigs from Nipponbare versus 93-11, leading us to ask whether there is alternative splicing possibly due to indel between these rice subspecies. We took a systematic approach to identifying transcript variants between the two rice subspecies by mining microarray data generated from the hybridization of various tissue RNAs from *japonica* variety Nipponbare (GSE7951) and *indica* variety IR64 (GSE6893). We mapped the 6498 predicted rice AS genes to about 7504 probe sets of the Affymetrix GeneChip rice whole genome, which contains more than 2086 predicted AS genes with multiple probe sets.

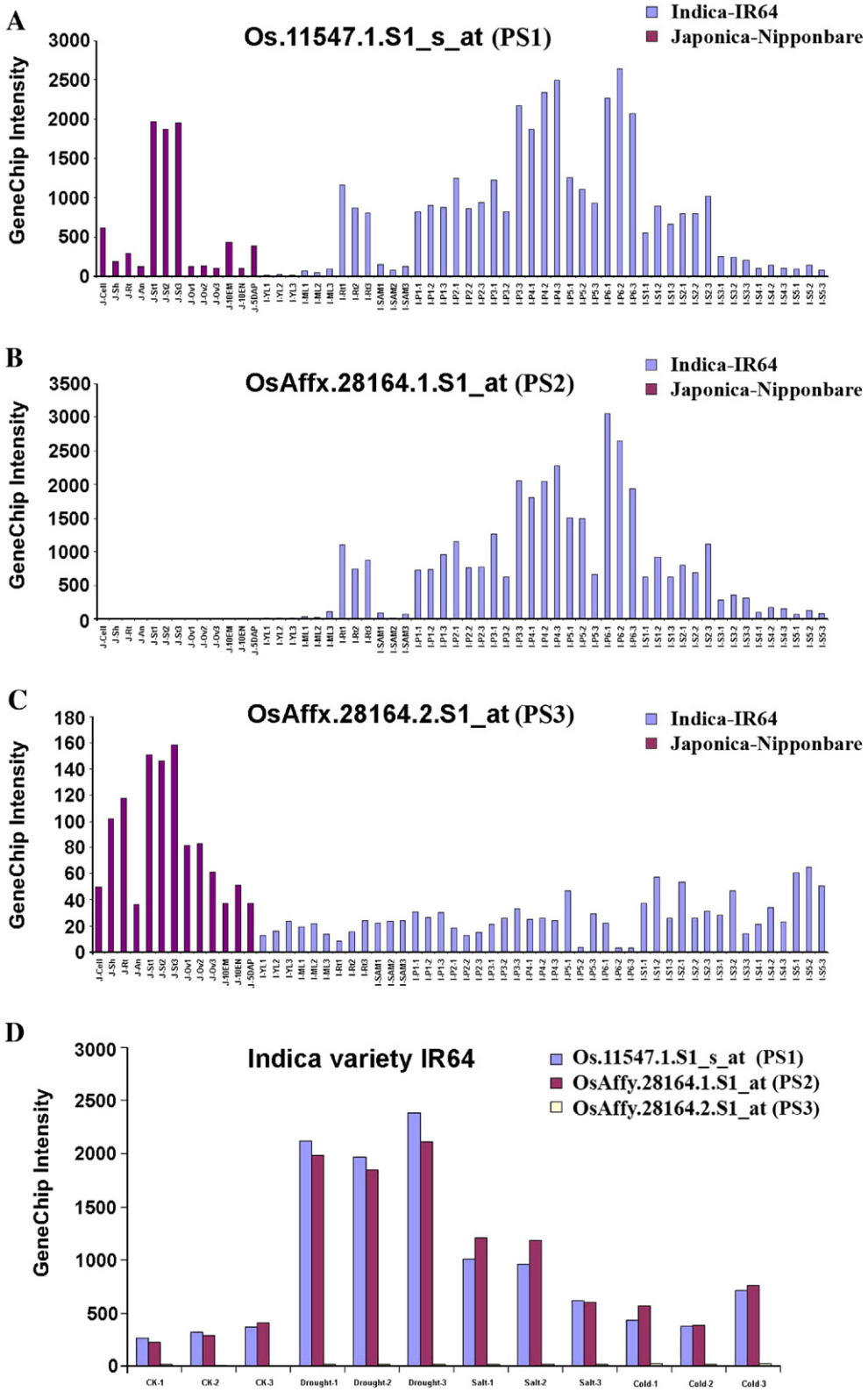
Although these two data sets do not have exactly the same number of tissue/organ types, there are enough common tissue/organ types, such as root, leaf, seed, and flower, for a sound comparison. Furthermore, the overall expression levels of presence/absence in the two cultivars provide relevant data for the AS analysis. To compare the array data from GSE7951 and GSE6893, we rescaled the data from GSE7951 and set the mean target intensity of each array to 100 using Affymetrix GCOS software, and we used the *Z*-score transformation normalization method to compare expression levels from the two microarray data sets. In total, we found 215 candidate genes through investigation of predicted AS genes with transcript variants and indel between *japonica* and *indica* subspecies of rice (Supplemental Table 1). Some of them have probe sets located exactly in the indel regions between contigs from Nipponbare versus 93-11 and match to different AS isoforms, such as LOC_Os06g48030, LOC_Os04g49757, and LOC_Os01g49529.

To identify possible mechanisms whereby indel caused AS transcripts between the *indica* and the *japonica* varieties, we used the rice peroxidase gene LOC_Os06g48030 as an example for a case study. Fig. 1 shows the scheme for the genome analysis of LOC_Os06g48030, including comparison of genomic regions in *japonica* and *indica* with highlighted indel and poly(A) signals and a map of three different isoforms that include predicted introns and exons. In the 3′ end of LOC_Os06g48030 there is a large gap in *indica* variety 93-11. Further analysis of this indel region indicates the presence of multiple poly(A) signals. Three gene models predicted by TIGR for LOC_Os06g48030 are shown in Fig. 1: LOC_Os06g48030.1 (isoform 1), LOC_Os06g48030.2 (isoform 2), and LOC_Os06g48030.3 (isoform 3).

Fig. 1 also indicates the locations of three probe sets for LOC_Os06g48030, including Os.11547.1.S1_s_at (PS1), OsAffx.28164.1.S1_at (PS2), and OsAffx.28164.2.S1_at (PS3). PS3 locates mainly in the indel region of LOC_Os06g48030. PS1 locates in all three isoforms. PS2 expands the end sequence region of isoform 2. PS3 locates in the end of both isoforms 1 and 3, four probes of PS3 partially hit isoform 2, and seven other probes are completely outside the isoform 2 region. Fig. 2 shows the results of the comparison for three probe sets of LOC_Os06g48030 in different tissues between IR64 and Nipponbare. All the tissue expression data from GSE7951 and GSE6893 for each probe set are given in one histogram, with the dark bar (left-hand side) for *japonica* variety Nipponbare and the gray bar (right-hand side) for the *indica* variety IR64. Both Nipponbare and IR64 bars are shown in PS1 (Fig. 2A). Interestingly, the dark bar is prominent in PS3 (Fig. 2C) but almost invisible in PS2 (Fig. 2B), while the gray bar (IR64) has a relatively low level in PS3 (Fig. 2C) but significant expression in PS2 (Fig. 2B). As shown in Figs. 2A, B, and C, different LOC_Os06g48030 isoforms are expressed differentially in tissues/organs from *indica* variety IR64 and *japonica* variety

Fig. 2. Comparison of GeneChip expression data between IR64 (*indica*) and Nipponbare (*japonica*) for three probe sets of LOC_Os06g48030 in different tissues. For comparison, the mean target intensity of each array was arbitrarily set to 100. The dark bar (left-hand side) stands for the expression levels of the Nipponbare variety, including the following tissues: triplicate for stigma (J-St) and ovary (J-Ov), suspension cell (J-Cell), shoot (J-Sh), root (J-Rt), anther (J-An), 10-day embryo after pollination (J-10EM), 10-day endosperm after pollination (J-10EN), and 5-day seed after pollination (J-5DAP). The gray bar (right-hand side) stands for the expression levels from the IR64 variety, including the following tissues with triplicate experiments: 7-day seedling root (I-Rt), mature leaf (I-ML), Y leaf (I-YL), SAM (I-SAM), young inflorescence (P1, 0–3 cm (I-P1); P2, 3–5 cm (I-P2); P3, 5–10 cm (I-P3); P4, 10–15 cm (I-P4); P5, 15–22 cm (I-P5); and P6, 22–30 cm (I-P6)), and seed (S1, 0–2 dap (I-S1); S2, 3–4 dap (I-S2); S3, 5–10 dap (I-S3); S4, 11–20 dap (I-S4); S5, 21–29 dap (I-S5)). (A) Expression levels of the probe set Os.11547.1.S1_s_at (PS1) from IR64 and Nipponbare. (B) Expression levels of probe set OsAffy.28164.1.S1_at (PS2) from IR64 and Nipponbare. (C) Expression levels of OsAffy.28164.2.S1_at (PS3) from IR64 and Nipponbare. (D) Expression levels of three probe sets under stress treatment. Raw data from the GEO database: GSE7951 (Nipponbare), GSE6893 (IR64), and GSE6901 (IR64). SAM, shoot apical meristem; dap, days after pollination.

Nipponbare, especially in root and reproductive tissue such as stigma and panicles. In addition, the *indica* variety IR64 stress treatment array data show that PS1 and PS2 were up-regulated under drought stress and induced slightly under salt and cold stress; PS3 was almost completely absent, irrespective of control or stress conditions (Fig. 2D). Therefore, the different expression patterns for probe sets of the rice peroxidase gene LOC_Os06g48030 indicate that cDNA sequence variance and alternative expression isoforms exist between the *japonica* and *indica* varieties.



RT-PCR validation for alternative expression isoforms of LOC_Os06g48030

To validate further the alternative expression patterns in different probe sets of the peroxidase gene LOC_Os06g48030, we carried out RT-PCR analysis of gene expression under cold stress using the specific primer pairs shown in Fig. 1: CS_F and CS_R1 (validating PS3) and CS_F and CS_R2 (validating PS2). As shown in Fig. 3A, the CS_1 primer set revealed one band (about 300 bp) in Nipponbare, but none in 93-11; and the CS_2 primer set reproduced the product (about 100 bp) in 93-11 and not in Nipponbare. Fig. 3A shows that the RT-PCR expression patterns were similar to those of the probe sets PS3 and PS2 in the microarray.

On the basis of the array and RT-PCR results, we undertook further analysis of the gene structure of LOC_Os06g48030. Due to limited information about the full-length cDNA sequence of LOC_Os06g48030 in *indica,* an additional primer pair FS_F and CS_R2 was designed (Fig. 1). The PCR product was reproduced by primer set FS_F and CS_R2 in *indica* variety 93-11 but not in *japonica* variety Nipponbare (Fig. 3B). We cloned the cDNA with 1347 bp and the sequencing results have been submitted to NCBI and are given in the supplemental data. The cloned *indica* cDNA sequence was aligned with full-length cDNA sequences published by the Japanese group and given in Fig. 3C. Variation between *indica* and *japonica* is located in the 3′ end of LOC_Os06g48030. Alternative isoform 2 may exist only in *indica* varieties, and isoforms 1 and 3 exist in *japonica* varieties.

Genotyping of LOC_Os06g48030 between indica and japonica varieties

To investigate the potential cause of the alternative expression patterns for the two probe sets of LOC_Os06g48030 (PS2 and PS3) in *japonica* and *indica* varieties, we conducted further genotyping analysis to establish whether any indel could be detected between *japonica* and *indica* varieties.

In Fig. 1, an indel was analyzed in the genome regions between Nipponbare (*japonica*) and 93-11 (*indica*), and there is large gap in the 3′ end of LOC_Os06g48030 in the *indica* variety 93-11. The primer pair GS_F and GS_R was designed for genome sequence analysis (Fig. 1). PCR was conducted using genomic DNAs isolated from five *indica* varieties (93-11, IR24, 03A-11, 03A-9, and Zhongyou13) and five *japonica* varieties (Nipponbare, Hua1, 746, Yunfeng7, and Xiangjingnuo). Fig. 4 shows that the size of the PCR products from *japonica* varieties is about 1200 bp, significantly larger than those from *indica* varieties (about 200 bp), suggesting there is a deletion in 93-11. To confirm the presence of indels in the gene LOC_Os06g48030 between *indica* and *japonica* subspecies, we cloned the PCR products for sequencing, which showed that

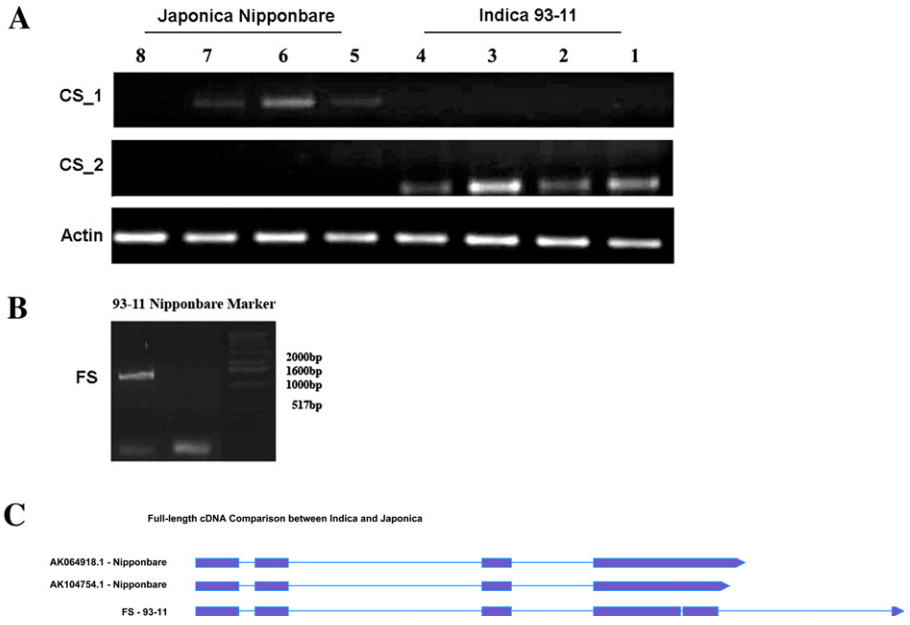


Fig. 3. RT-PCR validation for different probe sets and full-length cDNA comparison for LOC_Os06g48030 between *indica* and *japonica* varieties. (A) RT-PCR results under low temperature (4°C) in 93-11 (*indica* variety) and Nipponbare (*japonica* variety) for CS_1 (PCR product using CS_F and CS_R1) and CS_2 (PCR product using CS_F and CS_R2). Actin was used as a control. The RT-PCR samples are lanes 1, 9311—0 h; 2, 9311—12 h; 3, 9311—24 h; 4, 9311—48 h; 5, Nipponbare—0 h; 6, Nipponbare—12 h; 7, Nipponbare—24 h; 8, Nipponbare—48 h. The CS_1 primer set could amplify and detect one band (about 300 bp) in Nipponbare—0 h, Nipponbare—12 h, and Nipponbare—24 h, but not in 93-11. The CS_2 primer set could reproduce the product (about 100 bp) in 93-11, but not in Nipponbare. (B) RT-PCR result for FS (PCR product using primers FS_F and CS_R2). In the *indica* variety, the PCR product could be amplified with the FS primer set and produced cDNA of 93-11 and revealed one band. In the *japonica* variety, PCR could not be amplified with the FS primer set and there is no band in Nipponbare. (C) Full-length cDNA sequence comparison for LOC_Os06g48030 between *indica* and *japonica.* The RT-PCR product FS cloned and sequenced was compared to full-length cDNA sequences published by the Japanese group.

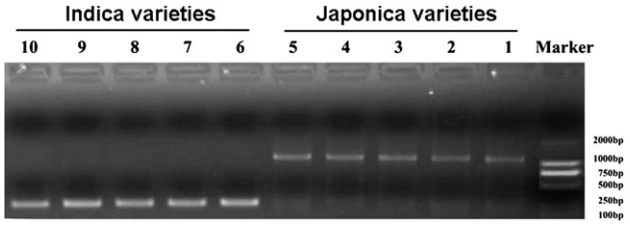


Fig. 4. The analysis of the indel in LOC_Os06g48030 between the *indica* and the *japonica* varieties. PCR analysis using genomic DNA of the *indica* and *japonica* varieties is presented. Lanes 1–5, the PCR product using the genomic DNA of *japonica* varieties Nipponbare, Hua1, 746, Yunfeng7, and Xiangjingnuo, respectively; lanes 6–10, the PCR product using the genomic DNA of the *indica* varieties 93-11, IR24, 03A-11, 03A-9, and Zhongyou13, respectively. DL2000 (Invitrogen) was used as a marker for the size of PCR product.

there is a 1017-bp fragment deleted from 93-11 (the sequencing and BLASTN results are given as supplemental data). We have submitted this specific region to NCBI.

Discussion

Alternative splicing is one of the most significant components of the complexity of eukaryote genomes, increasing protein diversity, creating a few isoforms, and affecting mRNA stability. Unlike conventional thought that AS happens in the same species under different development conditions or environmental stresses, we propose another hypothesis, that indels of subspecies may contribute to generating AS isoform diversity during evolution, leading to expanded populations of AS transcripts in rice and other organisms. Recently, whole genome sequences and large microarray data sets have become available, giving us the opportunity to undertake data integration and study the possible regulatory mechanism of rice AS by erecting and testing hypotheses before doing bench studies. We developed a new strategy, through genome-wide investigation of indels and microarray-based comparison of transcript profiles between *japonica* and *indica* subspecies of rice, and identified 215 rice candidate genes. In this study, we use the peroxidase gene LOC_Os06g48030 as an example to study the possible regulatory mechanism of alternative expression of isoforms between subspecies.

Microarray probe set PS1 of LOC_Os06g48030 hits in three isoforms and was expressed in both IR64 and Nipponbare; PS3 hits isoforms 1 and 3 and showed significantly lower expression in IR64, while PS2 hits isoform 2 and is expressed only in IR64 (Figs. 1 and 2).

During microarray analysis, we asked why probe set PS3 still has some expression in the *indica* variety shown in Fig. 2C. On the basis of sequence analysis, although there are 157 peroxidase genes in the rice genome, there should be no cross-hybridization between the PS3 probe sequence and any other rice peroxidase gene. We further investigated all 11 probes for PS3 and found that 4 probes of PS3 partly hit isoform 2 (Fig. 1); the other 7 probes of PS3 have multiple hits in other genes, although they are not peroxidases. This might be the reason the array intensity of PS3 shows slight expression in the *indica* variety and may be caused by cross-hybridization. Our RT-PCR analysis confirmed the array results (Fig. 3). The indel regions between *indica* variety 93-11 and *japonica* variety Nipponbare were identified by the TIGR genome browser and confirmed by our genotyping analysis using PCR amplification and genome sequencing for genome DNA samples from *indica* and *japonica* varieties (Fig. 4). The gene structure of LOC_Os06g48030 in *indica* and *japonica* was analyzed by comparing our cloned *indica* full-length cDNA with the published *japonica* cDNA sequence. It is likely that isoform 2 came from *indica* varieties, and isoforms 1 and 3 may have come from *japonica* varieties (Fig. 1).

The biological function of alternative splicing isoforms of LOC_Os06g48030 remains an enigma. The microarray data provide clues suggesting that LOC_Os06g48030 has an important role in plant development and stress tolerance. All three probe sets of LOC_Os06g48030 are highly expressed in root and reproductive tissues (stigma and panicle) (Fig. 2): PS1 is expressed significantly in the stigma of Nipponbare and in the

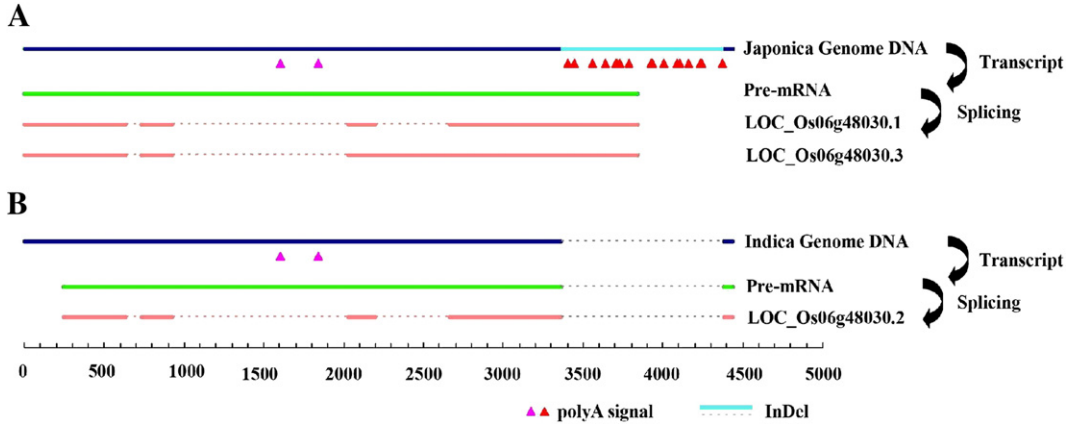


Fig. 5. A possible model for indel-caused alternative expression isoforms in the *indica* and *japonica* varieties. The model describes the transcription process from genomic DNA to mRNA in the *indica* and *japonica* varieties. (A) In the *japonica* varieties there is a 1017-bp region that contains multiple poly(A) signals (the light blue bar with red triangles). The transcription terminated in this region, and then the pre-mRNA was converted into two isoforms, LOC_Os06g48030.1 and LOC_Os06g48030.3, through an mRNA splicing process. (B) In the *indica* varieties, without the 1017-bp in genomic DNA (gray broken line), the transcription went farther and the pre-mRNA was converted into isoform LOC_Os06g48030.2, which contains a small exon in the 3′ end.

panicle of IR64, PS2 is expressed preferentially in the panicle of IR64, and PS3 is expressed preferentially in the stigma of Nipponbare. The IR64 stress treatment array data show that PS1 and PS2 are up-regulated under drought stress and induced slightly under salt and cold stress (Fig. 2D). There may be cross talk between drought stress and pollination for the peroxidase gene LOC_Os06g48030.

It is very puzzling, however, that LOC_Os06g48030 in *japonica* variety Nipponbare does not have isoform 2 and has a larger cDNA sequence compared to that of *indica* varieties. Here, we propose a potential model (Fig. 5) in which the alternative expression of different isoforms of LOC_Os06g48030 may be due to an indel(s) between *indica* and *japonica* varieties. In our model, the transcription process from genomic DNA to mRNA may be affected by the indel between *indica* and *japonica* varieties. In *japonica* varieties (Fig. 5A), there is a 1017-bp insertion region that contains multiple poly(A) signals; the analysis of poly(A) signals for the insertion sequence is given as supplemental material and this sequence has been submitted to NCBI. When the transcription terminates in this region, the pre-mRNA converts into isoforms 1 and 3 through an mRNA splicing process. This AS model is supported by two Nipponbare full-length cDNA sequences available in the KOME database (AK064918.1 and AK104754.1) (Fig. 3C). It is known that the 3′ ends of mRNAs terminate with a poly(A) tail, and poly(A) signals such as the AAUAAA motif have a very important role during posttranscriptional modification directed by sequence features present in the 3′ untranslated region [42]. In agreement with the literature [43,44], our results indicate that the mechanism for AS forms 1 and 3 may be related to both alternative splicing and alternative polyadenylation, due to the existence of poly(A) signals in the *japonica* insertion region. For *indica* varieties (Fig. 5B), without the 1017-bp region in the genomic DNA, the transcription extends farther and the pre-mRNA converts into isoform 2, which contains the small exon piece in its 3′ end. We cloned full-length cDNA from *indica* variety 93-11 and confirmed our predicted AS (Fig. 3C). Very interestingly, the indel region is conserved in both *japonica* and *indic*a varieties (Fig. 4). Here, we propose that isoform 2 of LOC_Os06g48030 exists in *indica* and may be the original form in evolution, and the 1017-bp DNA fragment with multiple poly(A) signals may have been inserted into *japonica* later, leading to isoforms 1 and 3. The multiple poly(A) signals in the indel region may be the key reason for generating the different isoforms of LOC_Os06g48030 between *indica* and *japonica* varieties. We found this sequence only in the rice genome and with no hit in any other species from the NCBI sequence databases. Further study of the origin of this sequence and its possible molecular function during rice evolution may be necessary.

In addition to LOC_Os06g48030, isoform 2 of LOC_Os04g49757 (Supplemental Fig. 1) and isoform 4 of LOC_Os01g49529 (Supplemental Fig. 2) were identified with transcript variants between *indica* and *japonica* varieties using the same approach. On the basis of the genome sequence analysis, it seems that the AS isoforms may be due to an indel(s) between rice subspecies, but the mechanisms may be completely different. In addition, the Affymetrix whole genome array was not designed specifically for studying indel-based alternative splicing, so there is the possibility that the probe sets were not matched perfectly in the isoform regions and in the indel regions. Further bench work is needed for studying the possible mechanisms.

This study provides a new approach to identifying AS between subspecies. This is the first application of microarray data mining and comparative genomics for identifying AS possibly due to indel between the rice subspecies *indica* and *japonica*. It provides a foundation for the development of a new microarray chip designed specifically for identification of AS isoforms between rice subspecies due to indels in a high-throughput fashion. In addition, AS candidate genes such as LOC_Os06g48030 could be new markers for identifying *indica* and *japonica* varieties (Fig. 4), which will have significant implications in future rice breeding. This new methodology will allow more discoveries in potentially indel-caused AS transcripts in rice and in other species. Our research will be very useful for the identification of the regulatory mechanism underlying alternative expression of isoforms between subspecies.

Materials and methods

Plant materials

DNA isolation

Fresh leaves from various cultivars (*indica* cultivars 93-11, IR24, 03A-11, 03A-9, and Zhongyou13; *japonica* cultivars Nipponbare, Hua1, 746, Yunfeng7, and Xiangjingnuo) were harvested from rice plants grown under natural conditions.

RNA isolation

Seeds of two rice cultivars (93-11 and Nipponbare) were surface-sterilized in 5% (w/v) sodium hypochlorite for 20 min and then washed in distilled water three or four times. The seeds were placed onto water-saturated Whatman paper for 1 day at 37°C to allow germination. The seedlings were transferred to a greenhouse (28°C day/25°C night, 12 h light/12 h dark, and 83% relative humidity). About 1 week after germination, the temperature was changed to 4–5°C, and budburst and root tissues were harvested after 12, 24, and 48 h cold treatment, frozen in liquid nitrogen, and stored at −80°C. Control plants were harvested at the same time.

DNA extraction and PCR analysis

Fresh leaves were collected and ground in liquid nitrogen. DNA was extracted from the ground tissues by the CTAB method [41]. Genome region primer sets (GS_F, 5′-CATGTTCACAAGTCCACCGCGC-3′; GS_R, 5′-CAAAAAG-GAATGGCATATGTATGGGA-3′) were designed according to the genome sequence of Nipponbare. A 25-μl reaction mixture was composed of 30 ng of total DNA, 10 mM Tris–HCl (pH 9.0), 50 mM $MgCl_2$, 0.1% (v/v) Triton X-100, 2 μM each primer, and 1 unit of Taq DNA polymerase (Promega). Amplification for the initial denaturing step was for 3 min at 94°C, followed by 35 cycles of 1 min at 94°C, 1 min at 58°C, 2 min at 72°C, with a final extension for 10 min at 72°C. The PCR product was separated by electrophoresis in a 1.2% (w/v) agarose gel.

RNA isolation and RT-PCR

All seedling samples from varieties 93-11 and Nipponbare were homogenized in liquid nitrogen before isolation of the RNA. Total RNA was isolated using TRIzol reagent (Invitrogen, CA, USA) and purified using Qiagen RNeasy columns (Qiagen, Hilden, Germany). Reverse transcription was performed using Moloney murine leukemia virus (M-MLV; Invitrogen). We heated 10-μl samples containing 2 μg of total RNA and 20 pmol of random hexamers (Invitrogen) at 70°C for 2 min to denature the RNA, and then chilled the samples on ice for 2 min. We added reaction buffer and M-MLV to a total volume of

20 μl containing 500 μM dNTPs, 50 mM Tris–HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$, 5 mM dithiothreitol, 200 units of M-MLV, and 20 pmol random hexamers. The samples were then heated at 42°C for 1.5 h. The cDNA samples were diluted to 8 ng/μl. The specific primer pairs CS_1 (CS_F and CS_R1), CS_2 (CS_F and CS_R2), and FS (FS_F and CS_R2) were designed by Primer3, and the primer pairs were CS_F, 5′-CCTACCCATGTGATATGATAGAAGG-3′; CS_R1, 5′-AGACGAGTCTAGAGTTCATATAGG-3′; CS_R2, 5′-CAAAAAGGAATGGCATATGTATGGGA-3′; and FS_F, 5′-ATGGGGCAGAGGAGGAGGTC-3′.

The amplification of actin was used as an internal control to normalize all data (ActinF, 5′-TATGGTCAAGGCTGGGTTCG-3′; ActinR, 5′-CCATGCTCGATGGGGTACTT-3′).

Array data reanalysis

We downloaded the CEL files of each experiment in the three microarray data sets (GSE7951 generated by the Chinese group and GSE6893 and GSE6901 generated by the Indian group) from the GEO Web site (http://www.ncbi.nlm.nih.gov/geo/). There are 70 chip data (13 from GSE7951, 45 from GSE6893, and 12 from GSE6901). All CEL files were reprocessed by Affymetrix GCOS software to produce the CHP file, and the target mean value was rescaled as 100 for each chip.

To map the probe set ID to the locus ID in the rice genome, the consensus sequence of each probe set was compared by BLAST (Basic Local Alignment and Search Tool) against the newest release of TIGR rice genome, version 5. The cutoff *E*-value was set as 1×10^{-20}. Within the 57,195 designed probe sets in the Affymetrix rice genome array, there are 52,697 probe sets mapped to rice genes in TIGR rice pseudomolecules.

Z-score transformation was used to identify the differential expression features between the *indica* and the *japonica* cultivars. The *Z* scores were calculated by taking the difference between the average expression level of *japonica* tissues (μ_j) and the average expression level of *indica* tissues (μ_i) divided by the standard deviation (SD_i) of the expression levels of *indica* tissues (for $Z_{j\_in\_i}$) or by the standard deviation (SD_j) of the expression levels of *japonica* tissues (for $Z_{i\_in\_j}$) using the following equations:

$$Z_{j\_in\_i} = (\mu_j - \mu_i)/SD_i,$$

$$Z_{i\_in\_j} = (\mu_j - \mu_i)/SD_j.$$

The *p* value was calculated on the basis of the *Z* score, with $p \le 0.05$ set as the level of statistical significance.

For gene LOC_Os06g48030, we define isoform 1 as LOC_Os06g48030.1, isoform 2 as LOC_Os06g48030.2, and isoform 3 as LOC_Os06g48030.3; for three probe sets of LOC_Os06g48030, we define PS1 for probe set Os.11547.1.S1_s_at, PS2 for OsAffx.28164.1.S1_at, and PS3 for OsAffx.28164.2.S1_at.

Sequence analysis

The sequencing results were assembled using SeqMan from the DNAstar package. Sequence alignment was done by MegAlign from the DNAStar package and bl2seq from NCBI. The polyadenylation signal was identified by the PLACE database (http://www.dna.affrc.go.jp/PLACE/).

Acknowledgments

This work was supported by grants from the Ministry of Science and Technology of China (2006CB100105) and the China Agriculture University. We thank Dr. Huaqin Helen Pan for helpful discussions and comments on the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.10.001.

References

[1] C.W. Smith, J. Valcarcel, Alternative pre-mRNA splicing: the logic of combinatorial control, Trends Biochem. Sci. 25 (2000) 381–388.
[2] A.C. Goldstrohm, A.L. Greenleaf, M.A. Garcia-Blanco, Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing, Gene 277 (2001) 31–47.
[3] B. Modrek, C. Lee, A genomic view of alternative splicing, Nat. Genet. 30 (2002) 13–19.
[4] D. Brett, H. Pospisil, J. Valcarcel, J. Reich, P. Bork, Alternative splicing and genome complexity, Nat. Genet. 30 (2002) 29–30.
[5] T. Maniatis, R. Reed, An extensive network of coupling among gene expression machines, Nature 416 (2002) 499–506.
[6] J. Leipzig, P. Pevzner, S. Heber, The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome, Nucleic Acids Res. 32 (2004) 3977–3983.
[7] L.F. Lareau, R.E. Green, R.S. Bhatnagar, S.E. Brenner, The evolving roles of alternative splicing, Curr. Opin. Struck. Biol. 14 (2004) 273–282.
[8] Y. Xing, C. Lee, Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes, Nat. Rev. 7 (2006) 499–509.
[9] H.Y. Kim, V.N. Gladyshev, Alternative first exon splicing regulates subcellular distribution of methionine sulfoxide reductases, BMC Mol. Biol. 7 (2006) 11.
[10] E. Kim, A. Magen, G. Ast, Different levels of alternative splicing among eukaryotes, Nucleic Acids Res. 35 (2007) 125–131.
[11] A.S. Reddy, Alternative splicing of pre-messenger RNAs in plants in the genomic era, Annu. Rev. Plant Biol. 58 (2007) 267–294.
[12] W. Gilbert, Why genes in pieces? Nature 271 (1978) 501.
[13] Z. Kan, E.C. Rouchka, W.R. Gish, D.J. States, Gene structure prediction and alternative splicing analysis using genomically aligned ESTs, Genome Res. 11 (2001) 889–900.
[14] Y. Zhou, C. Zhou, L. Ye, J. Dong, H. Xu, L. Cai, L. Zhang, L. Wei, Database and analyses of known alternatively spliced genes in plants, Genomics 82 (2003) 584–595.
[15] H. Kim, R. Klein, J. Majewski, J. Ott, Estimating rates of alternative splicing in mammals and invertebrates, Nat. Genet. 36 (2004) 915–916 (author reply 916-917).
[16] H. Itoh, T. Washio, M. Tomita, Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes, RNA (New York) 10 (2004) 1005–1018.
[17] S. Gupta, D. Zink, B. Korn, M. Vingron, S.A. Haas, Genome wide identification and classification of alternative splicing based on EST data, Bioinformatics (Oxford) 20 (2004) 2579–2585.
[18] P.K. Shah, L.J. Jensen, S. Boue, P. Bork, Extraction of transcript diversity from scientific literature, PLoS Comput. Biol. 1 (2005) e10.
[19] S. Stamm, J.J. Riethoven, V. Le Texier, C. Gopalakrishnan, V. Kumanduri, Y. Tang, N.L. Barbosa-Morais, T.A. Thanaraj, ASD: a bioinformatics resource on alternative splicing, Nucleic Acids Res. 34 (2006) D46–D55.
[20] V. Le Texier, J.J. Riethoven, V. Kumanduri, C. Gopalakrishnan, F. Lopez, D. Gautheret, T.A. Thanaraj, AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation, BMC Bioinform. 7 (2006) 169.
[21] F.C. Chen, S.S. Wang, S.M. Chaw, Y.T. Huang, T.J. Chuang, Plant gene and alternatively spliced variant annotator: a plant genome annotation pipeline for rice gene and alternatively spliced variant identification with cross-species expressed sequence tag conservation from seven plant species, Plant Physiol. 143 (2007) 1086–1095.
[22] H. Ner-Gaon, N. Leviatan, E. Rubin, R. Fluhr, Comparative cross-species alternative splicing in plants, Plant Physiol. 144 (2007) 1632–1641.
[23] J.M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P.M. Loerch, C.D. Armour, R. Santos, E.E. Schadt, R. Stoughton, D.D. Shoemaker, Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays, Science 302 (2003) 2141–2144.
[24] B.B. Wang, V. Brendel, Genomewide comparative analysis of alternative splicing in plants, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 7175–7180.

[25] K.R. Luehrsen, V. Walbot, Insertion of non-intron sequence into maize introns interferes with splicing, Nucleic Acids Res. 20 (1992) 5181–5187.
[26] M.J. Varagona, M. Purugganan, S.R. Wessler, Alternative splicing induced by insertion of retrotransposons into the maize waxy gene, Plant Cell 4 (1992) 811–820.
[27] A.S. Leprinc, M.A. Grandbastien, M. Christian, Retrotransposons of the Tnt1B family are mobile in Nicotiana plumbaginifolia and can induce alternative splicing of the host gene upon insertion, Plant Mol. Biol. 47 (2001) 533–541.
[28] G.F. Barry, The use of the Monsanto draft rice genome sequence in research, Plant Physiol. 125 (2001) 1164–1165.
[29] J. Yu, S. Hu, J. Wang, G.K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science 296 (2002) 79–92.
[30] S.A. Goff, D. Ricke, T.H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*), Science 296 (2002) 92–100.
[31] T. Sasaki, T. Matsumoto, K. Yamamoto, K. Sakata, T. Baba, Y. Katayose, J. Wu, Y. Niimura, Z. Cheng, Y. Nagamura, et al., The genome sequence and structure of rice chromosome 1, Nature 420 (2002) 312–316.
[32] Q. Feng, Y. Zhang, P. Hao, S. Wang, G. Fu, Y. Huang, Y. Li, J. Zhu, Y. Liu, X. Hu, et al., Sequence and analysis of rice chromosome 4, Nature 420 (2002) 316–320.
[33] I. Bancroft, Insights into cereal genomes from two draft genome sequences of rice, Genome Biol. 3 (2002) (REVIEWS1015).
[34] B. Han, Y. Xue, Genome-wide intraspecific DNA-sequence variations in rice, Curr. Opin. Plant Biol. 6 (2003) 134–138.
[35] Y.J. Shen, H. Jiang, J.P. Jin, Z.B. Zhang, B. Xi, Y.Y. He, G. Wang, C. Wang, L. Qian, X. Li, et al., Development of genome-wide DNA polymorphism database for map-based cloning of rice genes, Plant Physiol. 135 (2004) 1198–1205.
[36] M. Morgante, Plant genome organisation and diversity: the year of the junk! Curr. Opin. Biotechnol. 17 (2006) 168–173.
[37] M. Jain, A. Nijhawan, R. Arora, P. Agarwal, S. Ray, P. Sharma, S. Kapoor, A.K. Tyagi, J.P. Khurana, F-box proteins in rice: genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress, Plant Physiol. 143 (2007) 1467–1483.
[38] M. Li, W. Xu, W. Yang, Z. Kong, Y. Xue, Genome-wide gene expression profiling reveals conserved and novel molecular functions of the stigma in rice (*Oryza sativa* L.), Plant Physiol. 144 (2007) 1797–1812.
[39] M. Kar, D. Mishra, Catalase, peroxidase, and polyphenoloxidase activities during rice leaf senescence, Plant Physiol. 57 (1976) 315–319.
[40] J.M. Chittoor, J.E. Leach, F.F. White, Differential induction of a peroxidase gene family during infection of rice by Xanthomonas oryzae pv. oryzae, Mol. Plant-Microb. Interact. 10 (1997) 861–871.
[41] O.S. Rogers, A.J. Bendich, Extraction of DNA from plant tissue, Plant Mol. Biol. Manual A6 (1998) 1–10.
[42] A. Hajarnavis, I. Korf, R. Durbin, A probabilistic model of 3′ end formation in Caenorhabditis elegans, Nucleic Acids Res. 32 (2004) 3392–3399.
[43] G. Yeung, L.M. Choi, L.C. Chao, N.J. Park, D. Liu, A. Jamil, Martinson, Poly(A)-driven and poly(A)-assisted termination: two different modes of poly(A)-dependent transcription termination, Mol. Cell. Biol. 18 (1998) 276–289.
[44] S.J. Kim, H.G. Martinson, Poly(A)-dependent transcription termination: continued communication of the poly(A) signal with the polymerase is required long after extrusion in vivo, J. Biol. Chem. 278 (2003) 41691–41701.