Plant Mol Biol (2007) 65:403–415
DOI 10.1007/s11103-007-9174-7

A collection of 10,096 *indica* rice full-length cDNAs reveals highly expressed sequence divergence between *Oryza sativa indica* and *japonica* subspecies

Xiaohui Liu · Tingting Lu · Shuliang Yu · Ying Li · Yuchen Huang · Tao Huang · Lei Zhang · Jingjie Zhu · Qiang Zhao · Danlin Fan · Jie Mu · Yingying Shangguan · Qi Feng · Jianping Guan · Kai Ying · Yu Zhang · Zhixin Lin · Zongxiu Sun · Qian Qian · Yuping Lu · Bin Han

Received: 9 February 2007 / Accepted: 13 April 2007 / Published online: 24 May 2007

Abstract Relatively few *indica* rice full-length cDNAs were available to aid in the annotation of rice genes. The data presented here described the sequencing and analysis of 10,096 full-length cDNAs from *Oryza sativa* subspecies *indica* Guangluai 4. Of them, 9,029 matched rice genomic sequences in publicly-available databases, and 1,200 were identified as new rice genes. Comparison with the knowledge-based Oryza Molecular Biological Encyclopedia *japonica* cDNA collection indicated that 3,316 (41.6%) of the 7,965 *indica-japonica* cDNA pairs showed no distinct variations at protein level (2,117 *indica-japonica* cDNA pairs showed fully identical and 1,199 *indica-japonica* cDNA pairs showed no frame shift). Moreover, 3,645 (45.8%) of the *indica-japonica* pairs showed substantial differences at the protein level due to single nucleotide polymorphisms (SNPs), insertions or deletions, and sequence-segment variations between *indica* and *japonica* subspecies. Further experimental verifications using PCR screening and quantitative reverse transcriptional PCR revealed unique transcripts for *indica* subspecies. Comparative analysis also showed that most of rice genes were evolved under purifying selection. These variations might distinguish the phenotypic changes of the two cultivated rice subspecies *indica* and *japonica*. Analysis of these cDNAs extends known rice genes and identifies new ones in rice.

Keywords Comparative analysis · Full-length cDNA · *Indica* and *japonica* rice · *Oryza sativa* · Transcriptome

Xiaohui Liu, Tingting Lu, Shuliang Yu and Ying Li contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11103-007-9174-7) contains supplementary material, which is available to authorized users.

X. Liu · T. Lu · S. Yu · Y. Li · Y. Huang · T. Huang · L. Zhang · J. Zhu · Q. Zhao · D. Fan · J. Mu · Y. Shangguan · Q. Feng · J. Guan · K. Ying · Y. Zhang · Y. Lu · B. Han (✉)
National Center for Gene Research & Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China
e-mail: bhan@ncgr.ac.cn

T. Lu · Y. Li · Y. Huang · J. Zhu · Q. Zhao · Q. Feng · Z. Lin
College of Life Science & Biotechnology, Shanghai Jiaotong University, Shanghai, China

S. Yu
School of Life Sciences, Fudan University, Shanghai, China

Z. Sun · Q. Qian
The State Key Laboratory of Rice Biology, China Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou, China

Introduction

Rice is a major crop that feeds about half the world's population. Rice is also a model plant for molecular biological and genomic research because of its relatively small genome size, transformability and completion of genome sequencing. There is a well-established divergence between the two major Asian cultivated rice (*Oryza sativa* L.) subspecies, *indica* and *japonica*, but finer levels of genetic structure are suggested by the breeding history (Panaud et al. 2002; Garris et al. 2005). *Indica* and *japonica* rice diverged about approximately 0.2 ~ 0.44 million years ago (Ma and Bennetzen 2004; Vitte et al. 2004). *Indica* and *japonica* rice had a polyphyletic origin. *Indicas* were

probably domesticated in the foothills of Himalayas in Eastern India and *japonicas* somewhere in South China (Khush 1997). The *indica* subspecies of rice is the most widely cultivated subspecies in China, India and most of the rest of Asia, while the *japonica* rice subspecies is favored in Japan and other countries with temperate climates.

The entire rice genome sequence of *Oryza sativa* ssp. *japonica* Nipponbare, which is a typical *japonica* inbred variety, was completed by the International Rice Genome Sequencing Project (IRGSP) using a map-based sequencing strategy (Feng et al. 2002; Sasaki et al. 2002; The Rice Chromosome 10 Sequencing Consortium 2003; International Rice Genome Sequencing Project 2005). The draft genome sequences of the *japonica* Nipponbare and *indica* variety 93-11 have also been made available using a whole-genome shotgun sequencing approach (Goff et al. 2002; Yu et al. 2002, 2005). Overall synteny at the genome-wide level was reported previously using intra-specific genomic sequence comparisons (Feng et al. 2002; Han and Xue 2003; Ma and Bennetzen 2004; Yu et al. 2005), while comparison of the *indica* rice genome sequence with the *japonica* data provided insights into rice genetic diversity (Bennetzen 2002).

Full-length complementary DNA (cDNA) clones are important, not only for gene annotation and the determination of transcriptional start sites, but also for functional analyses (Suzuki et al. 2001; Wang and Brendel 2006). The methods for preferential cloning of cDNA that corresponds to full-length mRNAs with 5′-end-proximal cap structures (Kristiansen and Pandey 2002) have been developed and used in large-scale analyses of transcripts from human (Suzuki et al. 2002; Ota et al. 2004), mouse (Konno et al. 2001; The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium 2001; Osato et al. 2002; Carninci 2003), fruit fly (Stapleton et al. 2002), *Arabidopsis thaliana* (Seki et al. 2002), and rice (The Rice Full-Length cDNA Consortium 2003; Osato et al. 2003). Genomic comparisons of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy (Town et al. 2006). Similarly, Expressed Sequence Tag (EST) and cDNA sequences are also used for comparative genome analysis. The moss *Physcomitrella patens* transcriptome has been compared with the *A.thaliana* genome, revealing a large number of putative transcripts with high levels of similarity to vascular plant genes (Nishiyama et al. 2003). A unique set of 11,008 onion ESTs was used to assess the genomic differences between the *Asparagales* and *Poales* (Kuhl et al. 2004).

Computational annotation of the rice genome has been reported (Yuan et al. 2003) and collections of cDNAs and ESTs have provided valuable information toward our understanding of gene structure and genome coding capacity (Wu et al. 2002; The Rice Full-Length cDNA Consortium 2003; Rensink and Buell 2004; Zhang et al. 2005). Expression profiling of the rice genome using DNA microarrays has provided information on the coding potential and expression patterns of the chromosome 4 and the entire genome (Jiao et al. 2005; Li et al. 2006). Although 1,211,078 rice ESTs (http://www.ncbi.nlm.nih.gov/dbEST/dbEST) are presented in publicly-available databases, a large number of them is redundant. The Rice Full-Length cDNA Consortium has collected 28,469 unique full-length cDNA sequences from the *japonica* variety Nipponbare and provided a detailed description of the rice transcriptome (The Rice Full-Length cDNA Consortium 2003). The total number of rice full-length cDNA of publicly available KOME database is about 32,127 (Osato et al. 2003). These cDNAs provide the complete coding region of the encoded protein and complete 5′, 3′ untranslated regions (UTRs) that define the boundary of transcriptional unit and provide a functional resource for biological function verification. As part of the National Rice Functional Genomics Project in China, collection of 17,835 unique ESTs and 10,828 putative full-length cDNAs from *indica* variety Minghui 63 have been achieved (Xie et al. 2005; Zhang et al. 2005). Overall, the cDNA resources of the publicly available databases are still incomplete as it has been estimated that there are 37,500 ~ 43,000 genes predicted in the rice genome (International Rice Genome Sequencing Project 2005; Paterson et al. 2005). Comparative analysis of *indica* and *japonica* genomes at the expression level is likely to reveal some details of intra-specific variations as sequence polymorphisms in coding regions might influence the expression of genes and thus result in the phenotypic variations (Windsor and Mitchell-Olds 2006). In addition, gene structure as predicted by *ab initio* gene finders is never 100% accurate. Thus, a whole-genome full-length cDNA collection is a powerful resource for accurate gene prediction and comparative transcriptome analysis.

In this study, we constructed five full-length cDNA libraries from different *indica* rice tissues and various developmental stages including various stress treatments. In total, we collected and sequenced 10,096 unique full-length cDNA clones from a typical *indica* variety Guangluai 4 (Lin and Min 1991). Through transcriptome comparative analysis, we found that the overall number of polymorphisms, including insertions or deletions, single nucleotide polymorphisms (SNPs) and sequence polymorphisms between *indica* and *japonica* subspecies was very high. Further experimental verifications from quantitative reverse transcriptional (RT) PCR and Southern blotting hybridization analyses revealed unique transcripts in *indica* and *japonica* rice. These full-length cDNA clones also serve as a resource for further protein function

verification and the cDNA clones will be distributed upon request.

Materials and methods

Plant materials

Five cDNA libraries of *Oryza sativa* ssp. *indica* Guangluai 4 were constructed from five different tissues or at various developmental stages: (1) Two-day germinated shoots and roots were collected when roots reached 1–2 cm long; (2) Rice seedlings were grown in plant growth chamber with a cycle of 16 h light/8 h dark at 30°C. Seedling shoots and roots were harvested 2 weeks after germination; (3) Panicles were harvested from rice grown in paddy fields; (4) Two-week seedlings treated individually with various stresses, such as high-salinity (100 mM NaCl, treated for 20 min, 3, 12, 24, 48 h, 3 days and recovered for 72 h), dehydration (15% PEG-4000, treated for the same time duration as high-salinity), cold (6°C for 1, 12, 24, 48 h, 3 days and recovered for 72 h), heat (45°C, for the same time duration as cold), or immersion under water (for 1, 12, 24, 48 h, 3, 5 days) were harvested, and equimolar amounts of poly(A+) mRNA from the five tissues under stress treatments were combined for synthesis of cDNA.

Genomic DNA of the three *indica* (Guangluai 4, 93-11 and Nanjing 11) and five *japonica* (Nipponbare, Lansheng, Zhonghua 11, Xiushui 4 and Chunjiang 6) varieties were prepared from two-week rice seedling shoots. Classification and genealogy of Guangluai 4, Nanjing 11 and Xiushui 4 varieties were described by Lin and Min (1991). The *japonica* Chunjiang 6 variety was described by Sogawa et al. (2003).

Construction of full-length cDNA libraries

We utilized the Cap-Tagging method from the Oligo-Capping technique for full-length cDNA library construction (Suzuki et al. 2001). The 5′Cap-Tagging method utilizes the 5′Cap capture technique through the combined treatments of calf intestinal phosphatase (CIP) and tobacco acid pyrophosphatase (TAP) so that only the full-length cDNA was targeted for library construction (as outlined in Supplementary Fig. S1). Normalization and subtraction procedures were applied to reduce the frequency of highly expressed mRNA in the library and to enhance the discovery of new cDNAs (Carninci et al. 2000). Subtraction procedures were based on hybridization of the single-stranded DNA with RNA drivers from previously sequenced cDNAs or DNA primers designed from already known *japonica* cDNAs.

Total RNA was extracted with Trizol, and mRNA was purified with oligotex mRNA kit (Qiagen). mRNA was treated with CIP to remove the phosphate from truncated mRNA while the 5′capped full-length mRNA was not affected. Dephosphated mRNA was ligated with the first adapter (blocking tag) to block phosphate terminus residue of mRNA. The top strand sequence of the blocking tag is 5′-GGAATGATCCAG-3′ and bottom strand sequence is 5′-NNNCTGGATCATTCC-3′ (N=G, A, T, C). After purification, mRNA was treated at 37°C for 1 h with 50 units TAP (Epicentre) to remove the 5′cap from a full-length mRNA. De-capped mRNA was ligated to the second adapter (cap tag). The top and low strand adapter sequences are 5′-TAGGCCTTCCAGGCCAGTCGAGAC GACGTGA-3′ and 5′-NNNTCGCGTCGTCTCGACTGG CCTGGAAGGCCTA-3′ (N = G, A, T, C), respectively. First-strand cDNA was synthesized by superscript II RNase H- reverse transcriptase (Invitrogen) with oligo dT20VN carriying a XhoI site (5′-AGCTAATCGGTCTCCTCGAG GCCAAGCTGGCC(T)20VN-3′) (V = G, A, C; N = G, A, T, C).

Enrichment of full-length cDNA was utilized by biotin-labelling and magnetic porous glass (MPG)-streptavidin (CPG) sorting. Biotin-dCTP and random primer 5′-NNNN NNVVVVV-3′ (V = A, G, C; N = G, A, T, C) were added to the reverse transcription for additional 1-hour incubation at 42°C. Then, partial cDNA incorporated with biotin-dCTP was removed by MPG-streptavidin beads. Second-strand cDNA was synthesized with primer carrying a EcoRI site (5′-GTAGTACGGGTCTCGAATTCGGTAGG CCTTCCAGGCCAGTCGAG-3′) using cycling conditions of denature at 95°C for 2 min; 10 cycles of 45°C, 1 min, 55°C, 1 min, 68°C, 10 min and a final extension at 68°C for 10 min.

Double-stranded cDNA was digested with EcoRI and XhoI, and cDNA fragments of >2 kb, 1–2 kb, 0.5–1 kb and <0.5 kb were subsequently size-fractioned through an agarose gel electrophoresis. cDNAs were then cloned directionally into the EcoRI and XhoI sites of vector pBluescript SK+ (Strategene) and transformed into *E.coli* DH10B competent cells (Invitrogen).

Subtraction of full-length cDNA libraries

Two sources of subtraction drivers were utilized for cDNA library subtraction to increase novel cDNA discovery. One was the in vitro transcribed RNA originated from 5′-end sequenced 20,000 clones. The other was the 6,000 primers designed from 6,000 *japonica* Nipponbare non-redundant cDNA 3′-UTR sequences.

The in vitro transcribed driver cRNA was produced from 20,000-pooled plasmid DNAs. The tester single-stranded DNA (ssDNA) (+) was prepared from entire cDNA library with the help of the M13 helper phage. ssDNA was then enriched by means of hydroxyapatite

(HAP) column chromatography and PvuII digestion to eliminate double stranded plasmid DNA.

Hybridization was performed in 15-μl volume hybridization buffer (50% deionized formamide, 250 mM NaCl, 0.1% SDS and 5 mM EDTA) with 5-μg driver cRNA and 50-ng ssDNA (+) at 30°C for 70 h. After hybridization, non-hybrid ssDNA (+) was separated from hybrid ssDNA (+)-cRNA through HAP column chromatography and converted into dsDNA for transformation. The clones of subtracted libraries were sequenced at 5′ends to verify the subtraction efficiency.

Six thousand primers (sense strand, 30 bases long) were designed from 6,000 *japonica* Nipponbare non-redundant cDNA 3′-UTR sequence. They were pooled together and used as subtraction drivers. Tester ssDNA (–) was prepared from plasmid pool by the combined action of Gene II and Exo III (GeneTrapper kit, Invitrogen). The 6,000 primers were annealed with tester ssDNA and biotin labeled through primer extension. Biotin labeled ssDNA (–) was removed by MPG-streptavidin beads. Non-labeled ssDNA (–) was converted into double-stranded DNA, transformed into *E.coli* DH10B competent cells and sequenced at 5′ends to verify the subtraction efficiency.

DNA sequencing

DNA sequencing was carried out on ABI3730 sequencers. The clones were sequenced from both ends by the dideoxy chain termination method using BigDye Terminator Cycle sequencing V2.0 Ready Reaction (Applied Biosystems). The sequences were assembled using the PHRED and PHRAP programs (Ewing and Green 1998). Manual editing was utilized to validate the accuracy of the assembly.

Strategy of four sequencing phases of full-length cDNAs

One hundred and eighty thousand cDNA clones were sequenced at 5′ends and clustered into 21,690 non-redundant groups using the Lucy (Chou and Holmes 2001), TGICL (Pertea et al. 2003) and CAP3 programs (Huang and Madan 1999). One representative clone from each group was completely sequenced through four phases as following. Phase I: 5′end sequencing. Phase II: 3′end sequencing. Phase III: primer walking to obtain the full-length sequence. Phase IV: cDNA two strands were sequenced from both directions to get the solid sequence.

Quantitative RT-PCR

Gene expression variation between *indica* Guangluai 4 and *japonica* Nipponbare was verified by quantitative RT-PCR. First-strand cDNA was synthesized by superscript III RNase H^- reverse transcriptase (Invitrogen) with random 9 mers. Quantitative PCR was performed on the Applied Biosystems 7,500 real time PCR System using SYBR green (TaKaRa) as the reporter dye. The PCR thermal cycle conditions were as following: denature at 95°C for 1 min and 40 cycles for 95°C, 15 s; 60°C, 1 min. Then, the dissociation-curve analysis was performed to detect nonspecific products.

Annotation and comparative sequence analysis

Similarity searches against other cDNA and genomic sequences and proteins in the publicly-available databases (NCBI GenBank nt DB, est-other DB, rice *japonica* genomic sequence http://www.rgp.dna.affrc.go.jp/IRGSP/) were performed using BLAST program with the thresholds of *E*-value less than 1e-30 and 1e-10, respectively. We aligned cDNAs to genomic sequences by using ''est2genome'' program from EMBOSS package. Open reading frames (ORF) were determined by using ''getorf'' program of EMBOSS package (Rice et al. 2000). To calculate *K*a, *K*s values, a program of ''PAL2NAL'' (version: v11) was used (Suyama et al. 2006). The conserved sequences, polymorphic segments, indels, and SNPs in the collinear regions of *indica* and *japonica* were identified by using ''diffseq'' program of EMBOSS package. Multiple-sequence alignments were done using the ClustalW (version 1.81) program.

Accession numbers for submitted data in EMBL database

For *Oryza sativa* ssp. *indica* cv. Guangluai 4 full-length cDNAs (10,096) CT827960-CT834770, CT836522-CT836598, CT837477-CT837976, CT834771-CT836521, CT827880-CT827943, CT836599-CT837476 For *Oryza sativa* ssp. *indica* cv. Guangluai 4 unique-ESTs (21, 686): CT842009-CT842118, CT842120-CT842121, CT842123-CT842130, CT842132-CT842147, CT842149-CT842158, CT842160-CT842178, CT842180-CT842181, CT842183-CT842515, CT842516, CT842518-CT842528, CT842530-CT842531, CT842533-CT842540, CT842542-CT842595, CT842597-CT854325, CT854327-CT857270, CT857272-CT859335, CT859337-CT863709.

Results

Cloning and sequencing of full-length cDNAs of *O. sativa* ssp. *indica* Guangluai 4

Five full-length enriched cDNA libraries of *O. sativa* ssp. *indica* variety Guangluai 4 were constructed from tissues of 2-day germinated shoots and roots, seedling shoots and

roots, and panicles using an advanced 5′Cap-Tagging method. Initially, we completed 180,000 single-pass sequencing reactions on the selected clones from the normalized libraries. A total of 149,857 clones comprised of at least 100 contiguous nucleotides with a Phred score above 20 were obtained after eliminating low-quality sequences and vector trimming (Chou and Holmes 2001). These sequences were assembled into 21,690 unique ESTs (uni-ESTs) using the TGICL (Pertea et al. 2003) and CAP3 Sequence Assembly program (Huang and Madan 1999). We further selected potential full-length cDNA clones for completely sequencing through 3′end sequencing and primer-walking. In total, 10,096 full-length cDNAs were identified with an average length of 1,042 base pairs (bp); the longest clone sequence was 3,173 bp. We estimated that the average accuracy of the cDNA sequences was 99.9%. All sequences have been submitted to public EMBL sequence database with the accession numbers as described in the Materials and Methods. The sequences are also available at the National Center for Gene Research (NCGR) website (http://www.ncgr.ac.cn/cDNAs/). Here, we refer to these data resources of *indica* cDNAs and ESTs as the NCGR cDNA database.

We analyzed the proportion of full-length cDNA clones that contained mRNA start sites. As transcription usually starts at a purine base (Lewin 2000; Nishiyama 2003), we detected the base composition of first and subsequent nucleotides at the 5′ ends of the 21,690 ESTs. The first nucleotide showed a strong bias for purine (78% G or A; Table 1), whereas the 2–6th nucleotides showed much lower percentage of purine (42–51%). In the 5′end sequence data, the T/C ratio at the first position (0.68) differed from that at the second (1.45) and subsequent (0.8–0.98) positions (the third to the sixth ones) (Table 1).

Characterization of *indica* full-length cDNAs

To determine whether the cDNAs span the entire coding sequence, we used the BLASTN and BLASTX programs for comparing our sequences with the Knowledge-based Oryza Molecular Biological Encyclopedia (KOME, http://www.cdna01.dna.affrc.go.jp/cDNA) cDNA collection, the NCBI protein database, and the rice *japonica* (http://www.rgp.dna.affrc.go.jp/IRGSP/) and *indica* (http://www.rise.genomics.org.cn/rice/index2.jsp) genomic sequences. Of the 21,690 uni-ESTs, 13,509 matched cDNAs in the KOME database, 13,940 matched publicly-available rice ESTs, and 16,504 matched rice genomic sequences leaving 2,564 novel rice uni-ESTs. The remaining 5,286 uni-ESTs failed to match rice sequences in publicly-available databases (Fig. 1).

Among 10,096 full-length cDNAs, 7,965 (78.9%) matched KOME cDNAs (with ≥80% identity and at least 100-bp continuous hits). The remaining 2,131 *indica* cDNAs failed to match the KOME entries, but 1,200 of them matched the rice *indica* and *japonica* genomic sequences in publicly-available databases (with ≥80% identity and ≥60% length coverage), and these sequences were therefore identified to be novel rice full-length cDNAs. Among the remaining 931 cDNAs, 131 matched EST entries in the NCBI dbest, and 800 had no hits in the publicly-available databases (Fig. 1).

Of the 10,096 *indica* full-length cDNAs, 9,029 (89.4%) were mapped onto rice genome, and these cDNAs were distributed throughout the 12 rice chromosomes. To further analyze the *indica* and *japonica* genomic sequence hits, we found that 12 *indica* full-length cDNAs failed to map onto the *japonica* Nipponbare genome, while they could be fully matched KOME cDNAs, rice ESTs and the *indica* 93-11 genomic sequences. These clones were therefore deduced to be located in the gap of the current *japonica* Nipponbare genomic sequences (Supplementary Table S1). Our data also showed that 58 *indica* full-length cDNAs specifically mapped to the *indica* 93-11 genomic sequences, indicating that some of the 58 *indica* cDNAs might be *indica*-specific genes or they may map to the gaps in the *japonica* genome sequence.

We analyzed guanine plus cytosine (GC) composition of 5′- and 3′- untranslated regions (UTRs) and open reading frames (ORFs) of *indica* cDNAs (Fig. 2). Mean GC values for 5′UTRs (54.4%) and ORFs (56.1%) were equal, whereas mean GC content of 3′UTRs was very low (38.5%). The GC content of most of the 5′UTR and ORF sequences ranged from 35% to 75%; however, the GC

Table 1 The purine and T/C contents of the first six nucleotides of the NCGR cDNAs

	A	T	C	G	Total	AG%	T/C
1	12,235	1,920	2,825	4,514	21,494	**0.77924**	0.67965
2	6,659	7,282	5,019	2,587	21,547	0.42911	**1.45089**
3	6,344	4,838	6,030	4,342	21,554	0.49578	0.80232
4	6,873	5,142	5,230	4,307	21,552	0.51875	0.98317
5	6,336	5,528	5,878	3,797	21,539	0.47045	0.94046
6	5,662	5,800	6,226	3,854	21,542	0.44174	0.93158

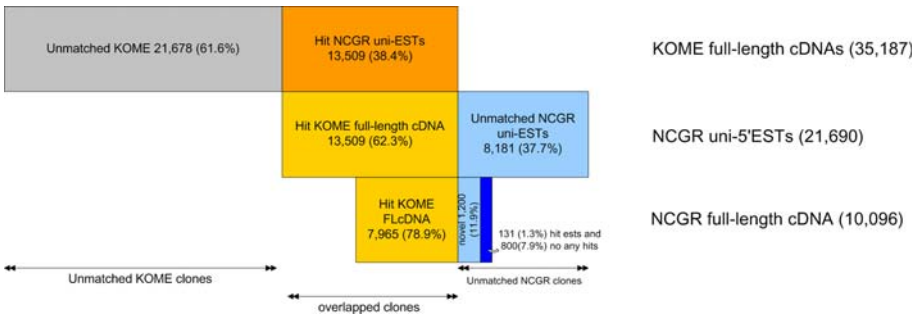

Fig. 1 Comparison of NCGR (*Indica* Guangluai 4) and KOME (*Japonica* Nipponbare) cDNA sequences. The total numbers of the KOME full-length cDNAs (35,187) and the NCGR full-length cDNAs (10,096) and the NCGR uni-ESTs (21,690) are indicated. The total numbers of overlapped cDNA clones between the KOME and the NCGR cDNA collection were indicated in the orange and yellow boxes respectively. The number of novel *indica* cDNAs were indicated in the blue and light blue boxes.

content of most of the 3′UTR sequences only ranged from 25% to 55% (Fig. 2A). Similar results were obtained for the analysis of 35,187 KOME *japonica* full-length cDNAs. The GC content of most 5′UTR and ORF sequences in *japonica* ranged from 35% to 75% and the GC content of 3′UTRs in *japonica* ranged from 25% to 55% (Fig. 2B).

Alternative splicing (AS) and antisense transcripts

Alternative splicing is widespread in both rice and *Arabidopsis* and these species share many common features (Campbell et al. 2006). Mapping of the full-length cDNAs to rice genome showed that 9,029 cDNAs represented 7,372 transcription units (TUs) in the rice genome. We identified 1,382 *indica* alternative splicing transcripts corresponding to 540 TUs. The conserved AS events corresponding to 93 TUs were identified in both *indica* and *japonica* subspecies. The other 447 TUs showed no AS events in *japonica*. Assigned functions of the 93 TUs between *indica* and *japonica* were assessed using searching against the PFAM protein family database (Apweiler et al. 2001; Bateman et al. 2004; http://www.sanger.ac.uk/Software/Pfam/). The results showed that 53 TUs had similarity with 45 PFAM protein families (*P*-score below 1e-10) (Supplementary Table S2).

Antisense RNAs are pairs of transcripts that are transcripted bi-directionally from an overlapping genome region. Among the 7,965 *indica* cDNAs that matched KOME cDNAs, 179 were identified to have *japonica* cDNA hits on the opposing strand, and therefore these cDNAs were annotated as anti-sense sequences. Additionally, we found 34 pairs of internal anti-sense transcripts in the NCGR *indica* cDNAs. Twenty-three of the 34 pairs were found to be pairs of internal anti-sense transcripts in the KOME *japonica* cDNAs and thus conserved in the two rice subspecies (Supplementary Table S3).

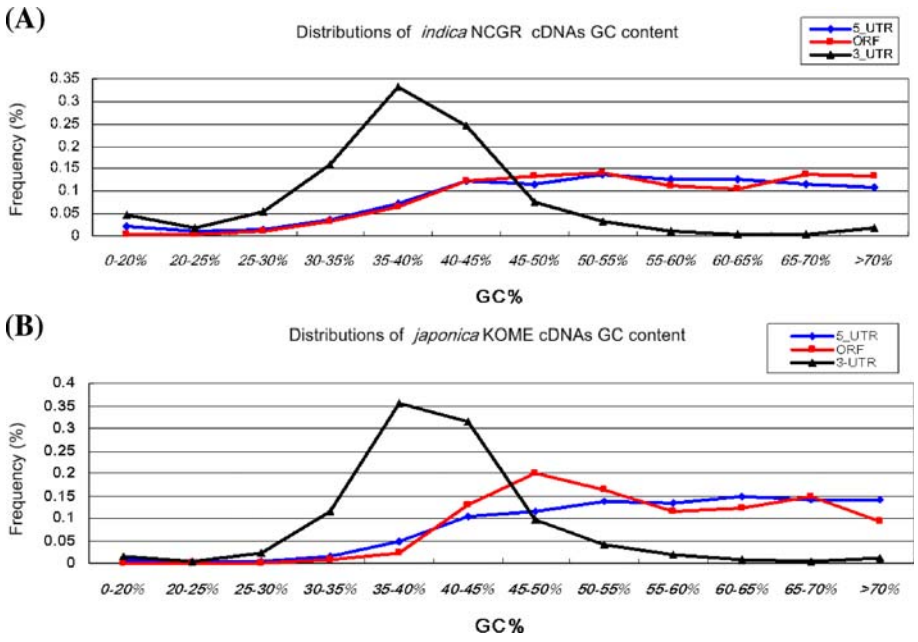

Fig. 2 Graphics showing the distributions of GC contents in 5′UTRs (blue), 3′UTRs (black) and ORFs (red) of the NCGR *indica* (**A**) and KOME *japonica* (**B**) full-length cDNAs

Transcriptome comparison between *indica* and *japonica*

We extracted the ORF of each full-length cDNA sequence using ''getorf'' program. Among 7,965 NCGR-KOME cDNA homologue pairs, 7,918 were predicted to have ORFs. Comparison of these *indica* and *japonica* ORFs revealed that 3,316 (41.6%) had no distinct variations at protein levels. Among these, 2,117 (26.6%) *indica-japonica* pairs were identical at protein level (designated as Identity protein), and 1,199 (15.1%) pairs were highly conserved with more than 96% identity at protein level (designated as Non-Frame Shift (NFS) proteins). Additionally, 3,645 (45.8%) NCGR-KOME cDNA pairs showed variations at protein level due to SNPs, insertions and deletions, non-homologous sequences and alternative splicing (designated as Variations).

We searched the NCGR-KOME cDNA pairs against the PFAM database. Overall, we found that 2,776 (39.9%) of these cDNAs showed similarity with 1,143 PFAM protein families (*P*-score below 1e-10). Of them, 789 NCGR-KOME pairs were classified into 30 major FPAM families after excluding the ''Domain of unknown Function (DUF)'' and ''Uncharacterized Protein Family (UPF)'' PFAM families (Fig. 3 and Table 2). Furthermore, we calculated these 789 pairs with the rate of non-synonymous (*Ka*) and synonymous (*Ks*) changes (41). Generally, the ratio of *Ka*/*Ks* provides a measure of evolutionary constrains (*Ka*/*Ks* = 1 neutral evolution, *Ka*/*Ks* > 1 positive selection, and *Ka*/*Ks* < 1 negative selection), while *Ks* represents the age of divergence between two homologous sequences. Percentages of the calculated *Ka*, *Ks* and *Ka*/*Ks* were shown in Table 2 and Fig. 3. Most of the rice genes have evolved under purifying and neutral selections. However, 136 genes showed *Ka*/*Ks* > 1, indicating that these *indica* and *japonica* genes were diverged under positive selections. Some proteins were highly diverged between the two subspecies. The average rate of the percentage of the protein with *Ka*/*Ks* > 1 in all protein categories was 17.2%. However, relatively high percentages of the proteins with *Ka*/*Ks* > 1 were found in some protein categories of ''Biotin_lipoyl'' (62.5%), ''RRM_1'' (43.6%) and ''Metallothio_2'' (36.8%) (Table 2). In contrast, other proteins seemed highly conserved between *indica* and *japonica*, which included ribosomal, ''peroxidase'', ''Tryp_alpha_amyl'', ''MIP'' and ''GH3'', as higher proportions of these identical proteins were found in each category.

In addition, we searched 1,200 novel NCGR cDNAs against the PFAM protein database. The results showed that only 8.5% (102) of the 1,200 novel cDNAs matched proteins in PFAM database (p-score below 1e-10). As mentioned above, 39.9% (2,776) of the NCGR-KOME cDNA pairs matched PFAM proteins. Obviously, the novel *indica* cDNAs identified in this study showed significant higher percentage of unknown functions.

Comparative analysis of the chromosome 4 cDNAs from two subspecies

As a total of 22.1-Mb chromosome 4 of *indica* Guangluai 4 has been sequenced (in publicly-available databases), we used the 23.2-Mb chromosome 4 collinear sequence of the *japonica* Nipponbare to compare exon-intron organization between *indica* and *japonica* (Table 3). Among 10,096 *indica* full-length cDNAs, 523 were mapped onto the *indica* 22.1-Mb region. Only five of them were not mapped on the 23.2-Mb *japonica* collinear region. We selected 361 NCGR-KOME collinear cDNAs for identifying exon-intron organizations in the two subspecies. We aligned the NCGR *indica* and KOME *japonica* cDNAs on the GLA4 and Nipponbare chromosome 4 collinear regions, respectively. Table 3 showed that mean exon sizes between *indica* Guangluai 4 (301 bp) and *japonica* Nipponbare (307 bp) were similar, but mean intron sizes between GLA4 (415 bp) and Nipponbare (461 bp) were different. These results were slightly different from the previous studies (Han and Xue 2003; International Rice Genome Sequencing Project 2005). Introns can be classified into phases 0, 1 and 2 depending on their position relative to the reading frame of the gene. Intron may interrupt the reading frame of a gene between two consecutive codons (phase 0 introns), between the first and second nucleotide of a codon (phase 1 introns), or between the second and the third nucleotide (phase 2 introns). In order to detect whether intron-phase variation existed between the two subspecies, we compared 56 identical NCGR-KOME transcripts referring to their corresponding chromosome 4 genomic sequences. The result showed no intron-phase variations observed. We scanned the unspliced mRNA (Supplementary Table S4). Thirty-two pairs of 361 NCGR-KOME cDNAs were found to be single exon in both subspecies. However, eight pairs of them were found to be single exon in *indica* but multiple exons in *japonica*, and four pairs of them were found to be single exon in *japonica* but multiple exons in *indica*.

Real time PCR analysis of the subspecific expressions

As described above, 12 *indica* cDNAs (assigned as Type I cDNAs) were assumed to be located in the gaps of the current *japonica* Nipponbare genome sequence, and 58 *indica* full-length cDNAs (assigned as Type II) were only aligned to *indica* 93-11 genomic sequences. Expression analysis of type I and type II cDNAs were carried out by real time RT-PCR. The results were shown in Fig. 4 and

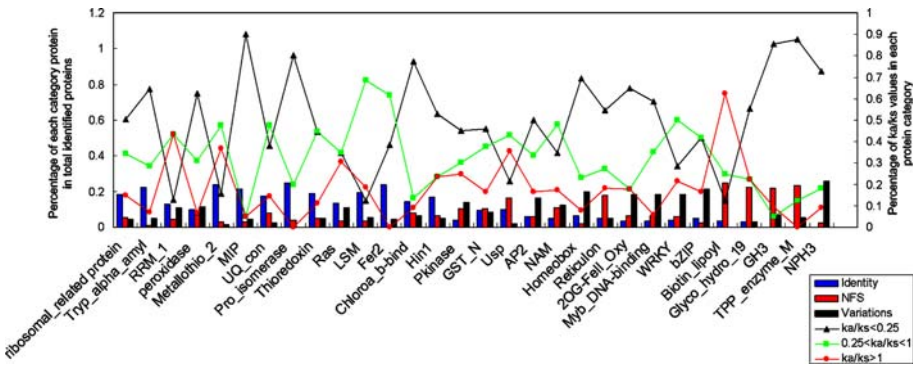

Fig. 3 Comparison of the percentages of conservations and variations at each InterPro classified NCGR-KOME homologous protein category. A total of 30 major categories were shown. Blue represents the percentage of identical protein (Identity). Red represents the percentage of Non-Frame Shift proteins (NFS). Green represents the percentage of variation protein resulted from SNPs, insertions/deletions and non-homologous sequences (Variations). Within each protein category, the percentages of the proteins with $Ka/Ks > 1$, $Ka/Ks < 0.25$ and $0.25 < Ka/Ks < 1$ were indicated by the orange, brown, and light green curves, respectively

Supplementary Table S5. Six of the Type I cDNAs were randomly selected for real time RT-PCR verification. All of these genes were expressed in both *indica* Guangluai 4 and *japonica* Nipponbare. Twenty-two type II cDNAs were only expressed in *indica* Guangluai 4, not in *japonica* Nipponbare (Supplementary Table S5) and may be *indica*-specific genes. We further detected whether the Type-II transcripts are present in the genomes of other *indica* and *japonica* varieties using PCR. The specific primers were designed for screening the type II genes in three *indica* varieties (Guangluai 4, 93-11 and Nanjing11) and four *japonica* varieties (Nipponbare, Lansheng, Zhonghua 11 and Chunjiang). Twenty-seven of the Type II genes were only detected in the *indica* varieties, indicating they are unique genome to *indica* varieties. The results were shown in Fig. 5. Further evidence was obtained from the real time RT-PCR analysis. Among the 58 Type II cDNAs, 27 appeared to be expressed only in *indica*.

Discussion

The domesticated Asian rice *Oryza sativa indica* subspecies represents the largest amount of rice production in the world. Although a collection of *indica* rice ESTs has been performed, large-scale *indica* rice full-length cDNA collection has not been available in public databases. In this study, we collected and completely sequenced 10,096 full-length cDNA clones and identified 21,690 *indica* uni-ESTs from *Oryza sativa* ssp. *indica* cv. Guangluai 4 to aid in the annotation of rice *indica* genes. This *indica* cDNA resource increased the number of publicly available rice expressed sequences and provided a platform for genome-wide comparison of two subspecies both in gene structure and further biological function verification.

We collected *indica* EST or mRNA sequences using a 5′Cap-Tagging approach to randomly select cDNA clones. This approach for rapid collecting of most transcript sequences from a novel genome was highly efficient. Other approaches such as ORFeome which is relying on large-scale PCR amplification of specific cDNAs followed by sequencing of the amplifications have been used to amplify cDNAs (Guigo et al. 2003; Wei et al. 2005). This method needs reasonably accurate gene predictions to use for PCR primer design. It will be much efficient through significant improvements in *de novo* gene prediction and optimizing and automating both the informatics and wet lab components of large-scale RT-PCR (Brent 2005).

Comparative genomics provides a powerful tool to study gene structure and the evolution of gene function and regulation (Soltis and Soltis 2003; Castelli et al. 2004; Katari et al. 2005; Odenwald et al. 2005). A recent study of exploring the plant transcriptome through phylogenetic profiling provides strong evidence for the existence of at least 33,700 genes in rice (Vandepoele and Van de Peer 2005). Among 7,965 *indica-japonica* (NCGR-KOME) homologue pairs, 3,316 (41.6%) showed no distinct variations at the protein level between *indica* and *japonica* subspecies, but 3,645 (45.8%) of the *indica-japonica* pairs showed large differences at protein level because of SNPs, insertions or deletions, and sequence-segment variations between *indica* and *japonica* subspecies. These variations might distinguish the phenotypic changes of the two cultivated rice subspecies, *indica* and *japonica*. The evidence for supporting this hypothesis was obtained from a recent cloning of the GS3 gene in rice (Fan et al. 2006). Rice grain size is a highly important quality trait. The long and slender grain is generally characteristics for *indica* rice, and short and round grain is for *japonica* rice. A recent report showed that the GS3 gene, which is controlling a

Table 2 PFAM categories

Pfam Acc.	Description	Identity	NFS	Variations	*Ka*/*Ks* < 0.25	0.25<*Ka*/*Ks* < 1	*Ka*/*Ks* > 1
null	Ribosomal protein	120	36	30	94 (50.5%)	64	28
PF00234	Tryp_alpha_amyl	22	1	5	18 (64.3%)	8	2
PF00076	RRM_1	18	6	15	5	17	17 (43.6%)
PF00141	Peroxidase	16	11	18	28 (62.2%)	14	3
PF01439	Metallothio_2	16	2	1	3	9	7 (36.8%)
PF00230	MIP	15	2	3	18 (90%)	1	1
PF00179	UQ_con	13	6	2	8	10	3
PF00160	Pro_isomerase	13	2	0	12 (80%)	3	0
PF00085	Thioredoxin	12	3	3	8	8	2
PF00071	Ras	11	3	9	8	8	7
PF01423	LSM	11	2	3	2	11 (68.8%)	3
PF00111	Fer2	11	0	2	5	8 (61.5%)	0
PF00504	Chloroa_b-bind	11	6	5	17 (77.3%)	3	2
PF07320	Hin1	10	4	3	9 (52.9%)	4	4
PF00069	Pkinase	10	27	36	33	22	18
PF02798	GST_N	8	9	7	11	9	4
PF00582	Usp	5	8	1	3	6	5
PF00847	AP2	5	5	14	12	8	4
PF02365	NAM	4	9	10	8	11	4
PF00046	Homeobox	3	1	9	9 (69.2%)	3	1
PF02453	Reticulon	2	7	2	6 (54.5%)	3	2
PF03171	2OG-FeII_Oxy	2	4	11	11 (64.7%)	3	3
PF00249	Myb_DNA-binding	2	4	11	10 (58.8%)	6	1
PF03106	WRKY	2	3	9	4	7 (50%)	3
PF00170	bZIP	2	1	9	5	5	2
PF00364	Biotin_lipoyl	1	7	0	1	2	5 (62.5%)
PF00182	Glyco_hydro_19	1	7	1	5	2	2
PF03321	GH3	0	16	5	18 (85.7%)	1	2
PF00205	TPP_enzyme_M	0	13	3	14 (87.5%)	2	0
PF03000	NPH3	0	1	10	8	2	1

Conservations and variations of NCGR-KOME homologous proteins were classified into 30 major categories of PFAM protein families. In each category, numbers of identical protein (Identity), >=96% identity with no frame shift (NFS) or Variations from insertion/deletion and non-homologous sequences were shown, numbers of the proteins with *Ka*/*Ks* > 1, 0.25< *Ka*/*Ks* < 1 and *Ka*/*Ks* < 1 were also shown in each column

major QTL for grain length, is identified to encode a putative transmembrane protein. Comparative sequencing analysis identified a nonsense mutation, shared among all the large-grain varieties (mostly are *indica* varieties) tested in comparison with the small grain varieties (mostly are *japonica* varieties), in the second exon of the putative GS3 gene. This mutation causes a 178-aa truncation in the C-terminus of the predicted protein, suggesting that GS3 may function as a negative regulator for grain size.

The large number of rice sense-antisense transcript pairs has been identified (Osato et al. 2003). We identified 179 rice antisense transcript pairs through comparative analysis of *indica* and *japonica* transcripts. Thirty-four sense-antisense transcript pairs were identified in the NCGR *indica* internal sequences, and 23 of them were also identified in the KOME *japonica* internal sequences. Clearly, conserved antisense transcripts in *indica* and *japonica* must play some roles in gene regulation in rice. Alternative spicing enables one gene to specify several proteins. We identified 1,382 *indica* alternative splicing transcripts corresponding to 540 predicted genes. Ninety-three genes of them were identified to have conserved AS events in both *indica* and *japonica* subspecies, but only 53 genes matched proteins in 45 PFAM protein families. The function of rice antisense transcripts and alternative splicing events should be carried out by experimental studies. In addition, collection of no-polyA transcripts from rice should be carried out for transcriptome analyses in rice.

Calculation of the *Ka* and *Ks* values showed that some gene families were highly diverged between *indica* and

Table 3 Characteristics of exon-intron organization in *indica* Guangluai 4 and *japonica* Nipponbare varieties

	Indica (cv. Guangluai 4)	*Japonica* (cv. Nipponbare)
cDNA	361	361
No. of exon	1,288	1,548
Mean No. of exon	3.6	4.3
Mean exon size(bp)	301 (281)[a]	307 (306)[a] (254)[b]
Mean 1st exon size (bp)	434	476
Mean last exon size (bp)	578	658
No. of intron	927	1,187
Mean intron size (bp)	415 (320)[a]	461 (316)[a] (413)[b]

The analyzed cDNAs were NCGR-KOME homology pairs to have been mapped on the *indica* and *japonica* chromosome 4. The data from the previous studies were indicated by *a* (Han and Xue 2003) and *b* (IRGSP 2005) individually

japonica. Of 789 *indica-japonica* gene pairs, 136 genes (*Ka*/*Ks* > 1) showed significant divergence between *indica* and *japonica*. These genes might be evolved under positive selection. We estimated that about 26.6% of the rice genes were identically conserved in the two *indica* and one *japonica* varieties, revealing that there were about 11,400 identity genes in total between *indica* and *japonica* subspecies. The large amount of the identity genes between the two subspecies indicated that *indica* and *japonica* were very closely related subspecies, and were not diverged for very long.

The expressions of novel *indica* cDNAs were detected by real time RT-PCR analysis. Our results indicated that Type I of the novel cDNAs seemed to be expressed in both *indica* and *japonica* variety, but their expression varies among two subspecies. However, the Type II genes were specifically expressed in *indica* Guangluai 4, indicating that these might be *indica* specific genes. Most of the Type II genes were only detected in *indica* varieties (Fig. 4). These results were in good accordance with the BLAST searching results against rice genomic sequences. In addition, 6 *indica* cDNAs that had no matches in Nipponbare genome were expressed in both Guangluai 4 and Nipponbare varieties, revealing their existence in Nipponbare genome sequences. As known that 95% of Nipponbare genome sequences had been finished, but still about 5% of Nipponbare genome sequences were uncompleted. These

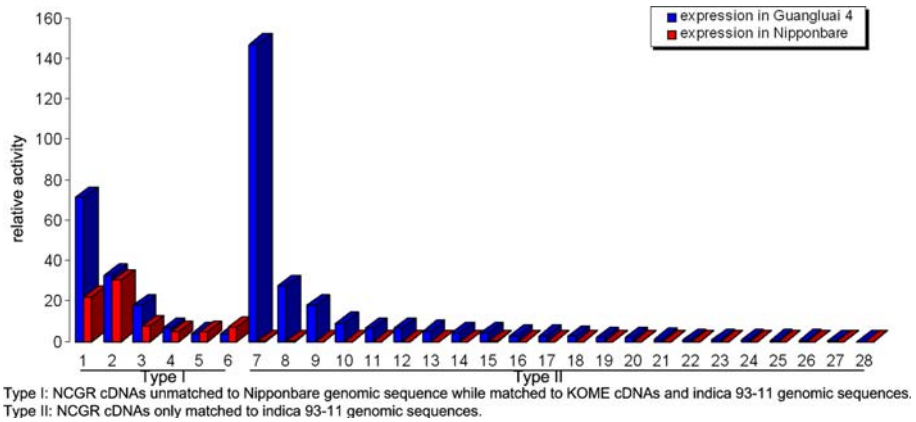

Fig. 4 Expression analysis of Type I and Type II novel cDNAs in *indica* Guangluai 4 and *japonica* Nipponbare. Type I represent the NCGR *indica* cDNAs that unmatched to *japonica* Nipponbare genomic sequence, but matched to the KOME cDNAs and *indica* 93-11 genomic sequences. Type II represent the NCGR *indica* cDNAs that only matched to *indica* 93-11 genomic sequences. Red represents cDNA expression level in *japonica* Nipponbare. Blue represents cDNA expression level in *indica* Guangluai 4. The relative quantification was calculated against the reference gene *actin 1*

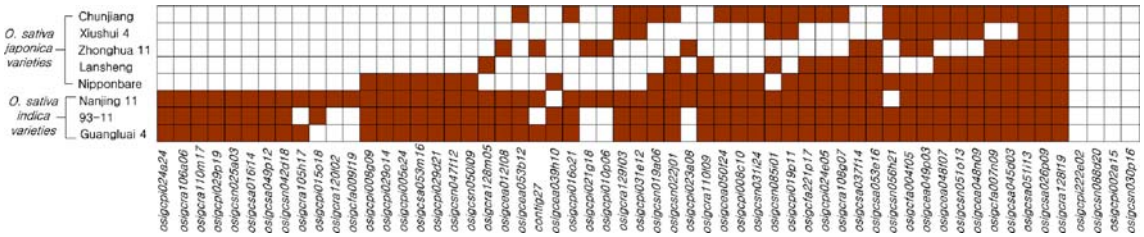

Fig. 5 Detection of the 58 Type II genes in 3 *indica* and 5 *japonica* varieties using PCR. The Type II cDNA clones and the varieties were indicated. The solid and empty boxes represent positive and negative results, respectively

genes were believed to be located in the Nipponbare sequence gaps, and could be used as probes for identifying the genomic bacterial artificial chromosomes (BACs) to fill the rice genome sequencing gaps. We identified a number of *indica* specific transcripts through PCR and real time RT-PCR analysis. Among the 58 Type II cDNAs, 27 seemed to be *indica* specific, indicating the proportion of the *indica* specific ones in 9,029 cDNAs was 3%. We would then estimate that there were about 130 *indica* specific transcripts in the 43,000 rice genes.

So, large-scale comparative analysis of *indica* and *japonica* full-length cDNAs showed gene expression variations that might lead to the discovery of molecular mechanism for phenotypic difference between two subspecies and will make impact on rice molecular breeding. Comprehensive analysis of the genomes, transcriptomes and proteomes of the rice *indica* and *japonica* subspecies will lead to a better understanding of the intra-specific divergence and functions of rice genes.

Acknowledgements We thank Robin Buell for critical comments on the manuscript. This research was supported by the grants from the Ministry of Science and Technology of China (The China Rice Functional Genomics Programs, grant no. 2002AA2Z1003/2006AA10A102), the Chinese Academy of Sciences (grant no. 038019315), and the National Natural Science Foundatation (grant no. 30325014).

References

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P,Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29:37–41

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2004) The Pfam protein familiesdatabase. Nucleic Acids Res 32:D138–D141

Bennetzen J (2002) Opening the door to comparative plant biology. Science 296:60–63

Brent MR (2005) Genome annotation past, present and future: how to define an ORFat each locus. Genome Res 15:1777–1786

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. BMC Genomics 7:327 doi: 10.1186/1471-2164-7-327

Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. Genome Res 10:1617–1630

Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D, Bono H, Kondo S, Sugahara Y, Saito R, Osato N, Fukuda S, Sato K, Watahiki A, Hirozane-Kishikawa T, Nakamura M, Shibata Y, Yasunishi A, Kikuchi N, Yoshiki A, Kusakabe M, Gustincich S, Beisel K, Pavan W, Aidinis V, Nakagawara A, Held WA, Iwata H, Kono T, Nakauchi H, Lyons P, Wells C, Hume DA, Fagiolini M, Hensch TK, Brinkmeier M, Camper S, Hirota J, Mombaerts P, Muramatsu M, Okazaki Y, Kawai J, Hayashizaki Y (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res 13:1273–1289

Castelli V, Aury JM, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quetier F, Scarpelli C, Schachter V, Temple G, Caboche M, Weissenbach J, Salanoubat M (2004) Whole genome sequence comparisons and ''full-length'' cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. Genome Res 14:406–413

Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. Bioinformatics 17:1093–1104

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Fan C, Xing Y, Mao H, Lu T, Han B, Xu C, Zhang Q (2006) *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice,encodes a putative transmembrane protein. Theor Appl Genet 112:1164–1171

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X, Jia P, Zhang Y, Zhao Q, Ying K, Yu S, Tang Y, Weng Q, Zhang L, Lu Y, Mu J, Lu Y, Zhang LS, Yu Z, Fan D, Liu X, Lu T, Li C, Wu Y, Sun T, Lei H, Li T, Hu H, Guan J, Wu M, Zhang R, Zhou B, Chen Z, Chen L, Jin Z, Wang R, Yin H, Cai Z, Ren S, Lv G, Gu W, Zhu G, Tu Y, Jia J, Zhang Y, Chen J, Kang H, Chen X, Shao C, Sun Y, Hu Q, Zhang X, Zhang W, Wang L, Ding C, Sheng H, Gu J, Chen S, Ni L, Zhu F, Chen W, Lan L, Lai Y, Cheng Z, Gu M, Jiang J, Li J, Hong G, Xue Y, Han B (2002) Sequence and analysis of rice chromosome 4. Nature 420:316–320

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence the rice genome *Oryza sativa* L. ssp. *japonica*. Science 296:92–100

Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, Antonarakis SE, Brent MR (2003) Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. Proc Natl Acad Sci USA 100:1140–1145

Han B, Xue Y (2003) Genome-wide intraspecific DNA-sequence variations in rice. Curr. Opin Plant Biol 6:134–138

Huang X, Madan A (1999) A DNA sequence assembly program. Genome Res9:868–877

International Rice Genome Sequencing Project (IRGSP) (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jiao Y, Jia P, Wang X, Su N, Yu S, Zhang D, Ma L, Feng Q, Jin Z, Li L, Xue Y, Cheng Z, Zhao H, Han B, Deng XW (2005) A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. Plant Cell 17:1641–1657

Katari MS, Balija V, Wilson RK, Martienssen RA, McCombie WR (2005) Comparing low coverage random shotgun sequence data from *Brassica oleracea* and *Oryza sativa* genome sequence for

their ability to add to the annotation of *Arabidopsis thaliana*. Genome Res 15:496–504

Khush GS (1997) Origin, dispersal, cultivation and variation of rice. Plant Mol. Biol 35:25–34

Konno H, Fukunishi Y, Shibata K, Itoh M, Carninci P, Sugahara Y, Hayashizaki Y (2001) Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a non redundant cDNA library. Genome Res 11:281–289

Kristiansen TZ, Pandey A (2002) Resources for full-length cDNAs. Trends Biochem Sci 27:266–267

Kuhl JC, Cheung F, Yuan Q, Martin W, Zewdie Y, McCallum J, Catanach A, Rutherford P, Sink KC, Jenderek M, Prince JP, Town CD, Havey MJ (2004) Aunique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders and Asparagales and Poales. Plant Cell 16:114–125

Lewin B (2000) Genes VII. Oxford University Press, Oxford

Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J, Deng XW (2006) Genome-wide transcription analyses in rice using tiling microarrays. Nat Genetics 38:124–129

Lin SC, Min SK (1991) Rice varieties and their genealogy in China. Shanghai Scientific and Technical Publishers, Shanghai

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nucleargenomes. Proc Natl Acad Sci USA 101:12404–12410

Nishiyama T, Fujita T, Shin-I T, Seki M, Nishide H, Uchiyama I, Kamiya A, Carninci P, Hayashizaki Y, Shinozaki K, Kohara Y, Hasebe M (2003) Comparativegenomics of *Physcomiyrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. Proc Natl Acad Sci USA 100:8007–8012

Odenwald WF, Rasband W, Kuzin A, Brody T (2005) EVOPRINTER, a multigenomic comparative tool for rapid identification of functionally important DNA. Proc Natl Acad Sci USA 102:14700–14705

Osato N, Itoh M, Konno H, Kondo S, Shibata K, Carninci P, Shiraki T, Shinagawa A, Arakawa T, Kikuchi S, Sato K, Kawai J, Hayashizaki Y (2002) A computer-based method of selecting clones for a full-length cDNA project: simultaneouscollection of negligibly redundant and variant cDNAs. Genome Res 12:1127–1134

Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K, Matsubara K, Kikuchi S, Hayashizaki Y (2003) Antisense transcripts with rice full-length cDNAs. Genome Biol 5:R5

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto J, Saito K, Kawai Y, Isono Y,Nakamura Y, Nagahari K, Murakami K, Yasuda T, Iwayanagi T, Wagatsuma M,Shiratori A, Sudo H, Hosoiri T, Kaku Y, Kodaira H, Kondo H, Sugawara M, Takahashi M, Kanda K, Yokoi T, Furuya T, Kikkawa E, Omura Y, Abe K, Kamihara K, Katsuta N, Sato K, Tanikawa M, Yamazaki M, Ninomiya K, Ishibashi T, Yamashita H, Murakawa K, Fujimori K, Tanai H, Kimata M, Watanabe M, Hiraoka S, Chiba Y, Ishida S, Ono Y, Takiguchi S, Watanabe S, Yosida M, Hotuta T, Kusano J, Kanehori K, Takahashi-Fujii A, Hara H, Tanase TO, Nomura Y, Togiya S, Komai F, Hara R, Takeuchi K, Arita M, Imose N, Musashino K, Yuuki H, Oshima A, Sasaki N, Aotsuka S, Yoshikawa Y, Matsunawa H, Ichihara T, Shiohata N, Sano S, Moriya S, Momiyama H, Satoh N, Takami S, Terashima Y, Suzuki O, Nakagawa S, Senoh A, Mizoguchi H, Goto Y, Shimizu F, Wakebe H, Hishigaki H, Watanabe T, Sugiyama A, Takemoto M, Kawakami B, Yamazaki M, Watanabe K, Kumagai A, Itakura S, Fukuzumi Y, Fujimori Y, Komiyama M, Tashiro H, Tanigami A, Fujiwara T, Ono T, Yamada K, Fujii Y, Ozaki K, Hirao M, Ohmori Y, Kawabata A, Hikiji T, Kobatake N, Inagaki H, Ikema Y, Okamoto S, Okitani R, Kawakami T, Noguchi S, Itoh T, Shigeta K, Senba T, Matsumura K, Nakajima Y, Mizuno T, Morinaga M, Sasaki M, Togashi T, Oyama M, Hata H, Watanabe M, Komatsu T, Mizushima-Sugano J, Satoh T, Shirai Y, Takahashi Y, Nakagawa K, Okumura K, Nagase T, Nomura N, Kikuchi H, Masuho Y, Yamashita R, Nakai K, Yada T, Nakamura Y, Ohara O, Isogai T, Sugano S (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36:40–45

Panaud O, Vitte C, Hivert J, Muzlak S, Talag J, Brar D, Sarr A (2002) Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using representational difference analysis (RDA). Mol Gen Genomics 268:113–121

Paterson AH, Freeling M, Sasaki T (2005) Grains of knowledge: genomics of model cereals. Genome Res 15:1643–1650

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–652

Rensink WA, Buell CR (2004) Arabidopsis to rice. Applying knowledge from a weed to enhance our understanding of a crop species. Plant Physiol 135:622–629

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, Antonio BA, Kanamori H, Hosokawa S, Masukawa M, Arikawa K, Chiden Y, Hayashi M, Okamoto M, Ando T, Aoki H, Arita K, Hamada M, Harada C, Hijishita S, Honda M, Ichikawa Y, Idonuma A, Iijima M, Ikeda M, Ikeno M, Ito S, Ito T, Ito Y, Ito Y, Iwabuchi A, Kamiya K, Karasawa W, Katagiri S, Kikuta A, Kobayashi N, Kono I, Machita K, Maehara T, Mizuno H, Mizubayashi T, Mukai Y, Nagasaki H, Nakashima M, Nakama Y, Nakamichi Y, Nakamura M, Namiki N, Negishi M, Ohta I, Ono N, Saji S, Sakai K, Shibata M, Shimokawa T, Shomura A, Song J, Takazaki Y, Terasawa K, Tsuji K, Waki K, Yamagata H, Yamane H, Yoshiki S, Yoshihara R, Yukawa K, Zhong H, Iwama H, Endo T, Ito H, Hahn JH, Kim HI, Eun MY, Yano M, Jiang J, Gojobori T (2002) The genome sequence and structure of rice chromosome 1. Nature 420:312–316

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. Science 296:141–145

Sogawa K, Li Y, Zhang J, Liu G, Yao H (2003) Genealogical analysis of resistance to the whitebacked planthipper *Sogatella furcifera* in Chinese *japonica* rice Chunjiang 06. Chinese J Rice Sci 17:67–72

Soltis DE, Soltis PS (2003) The role of phylogenetics in comparative genetics. Plant Physiol 132:1790–1800

Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, Wan K, Rubin GM, Celniker SE (2002) A Drosophila full-length cDNA resource. Genome Biol 312:research0080.1–0080.8

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–W612

Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, Okubo K, Sakaki Y,

Nakamura Y, Suyama A, Sugano S (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep 2:388–393

Suzuki Y, Yamashita R, Nakai K, Sugano S (2002) DBTSS: database of human transcriptional start sites and full-length cDNAs. Nucleic Acids Res 30:328–331

The Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. Science 300:1566–1569

The Rice Full-Length cDNA Consortium (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. Science 301:376–379

The RIKEN genome exploration research group phase II team, the FANTOM consortium (2001) Functional annotation of a full-length mouse cDNA collection. Nature 409:685–690

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ, Vigouroux M, Trick M, Bancroft I (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell 18:1348–1359

Vandepoele K, Van de Peer Y (2005) Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol 137:31–42

Vitte C, Ishii T, Lamy F, Brar D, Panaud O (2004) Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). Mol Gen Genomics 272:504–511

Wang BB, Brendel V (2006) Genome wide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA 103:7175–7180

Wei C, Lamesch P, Arumugam M, Rosenberg J, Hu P, Vidal M, Brent MR (2005) Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. Genome Res 15:577–582

Windsor AJ, Mitchell-Olds T (2006) Comparative genomics as a tool for gene discovery. Curr Opin Biotec 17:1–7

Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J, Fujii F, Shomura A, Ando T, Kono I, Waki K, Yamamoto K, Yano M, Matsumoto T, Sasaki T (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell 14:525–535

Xie K, Zhang J, Xiang Y, Feng Q, Han B, Chu Z, Wang S, Zhang Q, Xiong L (2005) Isolation and annotation of 10828 putative full length cDNAs from indica rice. Sci China Ser C Life Sci 48:445–451

Yuan Q, Ouyang S, Liu J, Suh B, Cheung F, Sultana R, Lee D, Quackenbush J, Buell CR (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res 31:229–233

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome *Oryza sativa* L. ssp. *indica*. Science 296:92–100

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J, Zhang Y, Li R, Xu Z, Li S, Li X, Zheng H, Cong L, Lin L, Yin J, Geng J, Li G, Shi J, Liu J, Lv H, Li J, Wang J, Deng Y, Ran L, Shi X, Wang X, Wu Q, Li C, Ren X, Wang J, Wang X, Li D, Liu D, Zhang X, Ji Z, Zhao W, Sun Y, Zhang Z, Bao J, Han Y, Dong L, Ji J, Chen P, Wu S, Liu J, Xiao Y, Bu D, Tan J, Yang L, Ye C, Zhang J, Xu J, Zhou Y, Yu Y, Zhang B, Zhuang S, Wei H, Liu B, Lei M, Yu H, Li Y, Xu H, Wei S, He X, Fang L, Zhang Z, Zhang Y, Huang X, Su Z, Tong W, Li J,Tong Z, Li S, Ye J, Wang L, Fang L, Lei T, Chen C, Chen H, Xu Z, Li H, Huang H, Zhang F, Xu H, Li N, Zhao C, Li S, Dong L, Huang Y, Li L, Xi Y, Qi Q, Li W, Zhang B, Hu W, Zhang Y, Tian X, Jiao Y, Liang X, Jin J, Gao L, Zheng W, Hao B, Liu S, Wang W, Yuan L, Cao M, McDermott J, Samudrala R, Wang J, Wong GK, Yang H (2005) The genomes of *Oryza sativa*: a history of duplications. PloS Biology 32:e38

Zhang J, Feng Q, Jin C, Qiu D, Zhang L, Xie K, Yuan D, Han B, Zhang Q, Wang S (2005) Features of the expressed sequences revealed by a large-sale analysis of ESTs from a normalized cDNA library of the elite *indica* rice cultivar Minghui 63. Plant J 42:772–780