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

1 indica Oryza sativa indica japonica

Received: 9 Februar 2007 / Accepted: 13 April 2007 / Published online: 24 Ma 2007
Springer Science+Business Media B.V. 2007

Relativel fe indica rice full-length cDNAs ere available to aid in the annotation of rice genes. The data presented here described the sequencing and anal sis of 10,096 full-length cDNAs from *Oryza sativa* subspecies *indica* Guangluai 4. Of them, 9,029 matched rice genomic sequences in publicl -available databases, and 1,200 ere identified as ne rice genes. Comparison ith the kno ledge-based Or a Molecular Biological Enc clopedia *japonica* cDNA collection indicated that 3,316 (41.6%) of the 7,965 *indica-japonica* cDNA pairs sho ed no distinct variations at protein level (2,117 *indica-japonica* cDNA pairs sho ed full identical and 1,199 *indica-japonica* cDNA pairs sho ed no frame shift). Moreover, 3,645 (45.8%) of the *indica-japonica* pairs sho ed substantial differences at the protein level due to single nucleotide pol morphisms (SNPs), insertions or deletions, and sequence-segment variations bet een *indica* and *japonica* subspecies. Further e perimental verifications using PCR screening and quantitative reverse transcriptional PCR revealed unique transcripts for *indica* subspecies. Comparative anal sis also sho ed that most of rice genes ere evolved under purif ing selection. These variations might distinguish the phenot pic changes of the t o cultivated rice subspecies *indica* and *japonica*. Anal sis of these cDNAs e tends kno n rice genes and identifies ne ones in rice.

Comparative anal sis · Full-length cDNA · *Indica* and *japonica* rice · *Oryza sativa* · Transcriptome

Rice is a major crop that feeds about half the orld's population. Rice is also a model plant for molecular biological and genomic research because of its relativel small genome si e, transformabilit and completion of genome sequencing. There is a ell-established divergence bet een the t o major Asian cultivated rice (*Oryza sativa* L.) subspecies, *indica* and *japonica*, but ner levels of genetic structure are suggested b the breeding histor (Panaud et al. 2002; Garris et al. 2005). *Indica* and *japonica* rice diverged about appro imatel 0.2 0.44 million ears ago (Ma and Bennet en 2004; Vitte et al. 2004). *Indica* and *japonica* rice had a pol ph letic origin. *Indicas* ere

Xiaohui Liu, Tingting Lu, Shuliang Yu and Ying Li contributed equall to this ork.

The online version of this article (doi:10.1007/s11103-007-9174-7) contains supplementar material, hich is available to authori ed 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 Ph siolog and Ecolog , Shanghai Institutes for Biological Sciences, Chinese Academ 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 & Biotechnolog , Shanghai Jiaotong Universit , Shanghai, China

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

Z. Sun · Q. Qian
The State Ke Laborator of Rice Biolog , China Rice Research Institute, Chinese Academ of Agricultural Sciences, Hang hou, China

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 verification and the cDNA clones will be distributed upon request.

[illegible]

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′-TAGGCCTTCCAGGCCAGTCGAGACGACGTGA-3′ and 5′-NNNTCGCGTCGTCTCGACTGGCCTGGAAGGCCTA-3′ (N = G, A, T, C), respectively. First-strand cDNA was synthesized by superscript II RNase H- reverse transcriptase (Invitrogen) with oligo dT20VN carrying a XhoI site (5′-AGCTAATCGGTCTCCTCGAGGCCAAGCTGGCC(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′-NNNNNNVVVVVV-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′-GTAGTACGGGTCTCGAATTCGGTAGGCCTTCCAGGCCAGTCGAG-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.

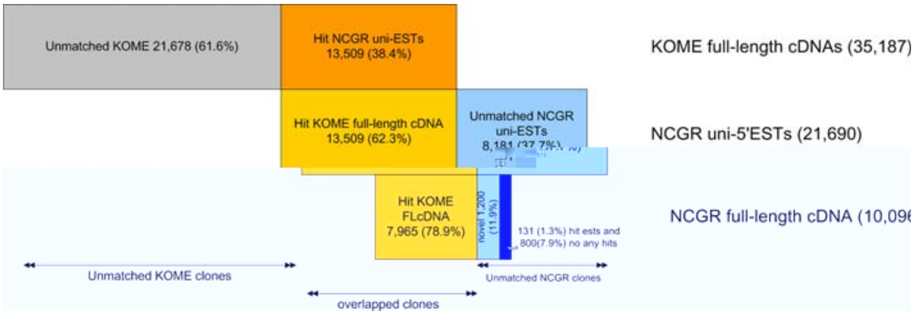
Double-stranded cDNA was digested with EcoRI and XhoI, and cDNA fragments of >2 kb, 1–2 kb, 0.5–1 kb and <0.5 kb were subsequently size-fractioned through an agarose gel electrophoresis. cDNAs were then cloned directionally into the EcoRI and XhoI sites of vector pBluescript SK+ (Strategene) and transformed 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 chromatograph and PvuII digestion to roots, and panicles using an advanced 5 Cap-Tagging method. Initiall , e completed 180,000 single-pass sequencing reactions on the selected clones from the normali ed libraries. A total of 149,857 clones comprised of at


Unmatched KOME 21,678 (61.6%)
Hit NCGR uni-ESTs 13,509 (38.4%)
KOME full-length cDNAs (35,187)
Hit KOME full-length cDNA 13,509 (62.3%)
Unmatched NCGR uni-ESTs 8,181 (37.7%)
NCGR uni-5'ESTs (21,690)
Hit KOME FLcDNA 7,965 (78.9%)
novel 1,200 (11.9%)
131 (1.3%) hit ests and 800(7.9%) no any hits
NCGR full-length cDNA (10,09
Unmatched KOME clones
Unmatched NCGR clones
overlapped clones

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 bet een the KOME and the NCGR cDNA collection ere indicated in the orange and ello bo es respectivel . The number of novel *indica* cDNAs ere indicated in the blue and light blue bo es.

content of most of the 3 UTR sequences onl ranged from 25% to 55% (Fig. 2A). Similar results ere obtained for the anal sis 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 idespread in both rice and *Arabidopsis* and these species share man common features (Campbell et al. 2006). Mapping of the full-length cDNAs to rice genome sho ed 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 ere identified in both *indica* and *japonica* subspecies. The other 447 TUs sho ed no AS events in *japonica*. Assigned functions of the 93 TUs bet een *indica* and *japonica* ere assessed using searching against the PFAM protein famil database (Ap eiler et al. 2001; Bateman et al. 2004; http:// .sanger.ac.uk/Soft- are/Pfam/). The results sho ed that 53 TUs had similarit ith 45 PFAM protein families (*P*-score belo 1e-10) (Supplementar Table S2).

Antisense RNAs are pairs of transcripts that are transcribed bi-directionall from an overlapping genome region. Among the 7,965 *indica* cDNAs that matched KOME cDNAs, 179 ere identified to have *japonica* cDNA hits on the opposing strand, and therefore these cDNAs ere annotated as anti-sense sequences. Additionall , e found 34 pairs of internal anti-sense transcripts in the NCGR *indica* cDNAs. T ent -three of the 34 pairs ere found to be pairs of internal anti-sense transcripts in the KOME *japonica* cDNAs and thus conserved in the t o rice subspecies (Supplementar Table S3).

Graphics sho ing the distributions of GC contents in 5 UTRs (blue), 3 UTRs (black) and ORFs (red) of the NCGR *indica* ( ) and KOME *japonica* ( ) 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 (*K*a) and synonymous (*K*s) changes (41). Generally, the ratio of *K*a/*K*s provides a measure of evolutionary constrains (*K*a/*K*s = 1 neutral evolution, *K*a/*K*s > 1 positive selection, and *K*a/*K*s < 1 negative selection), while *K*s represents the age of divergence between two homologous sequences. Percentages of the calculated *K*a, *K*s and *K*a/*K*s were shown in Table 2 and Fig. 3. Most of the rice genes have evolved under purifying and neutral selections. However, 136 genes showed *K*a/*K*s > 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 *K*a/*K*s > 1 in all protein categories was 17.2%. However, relatively high percentages of the proteins with *K*a/*K*s > 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.

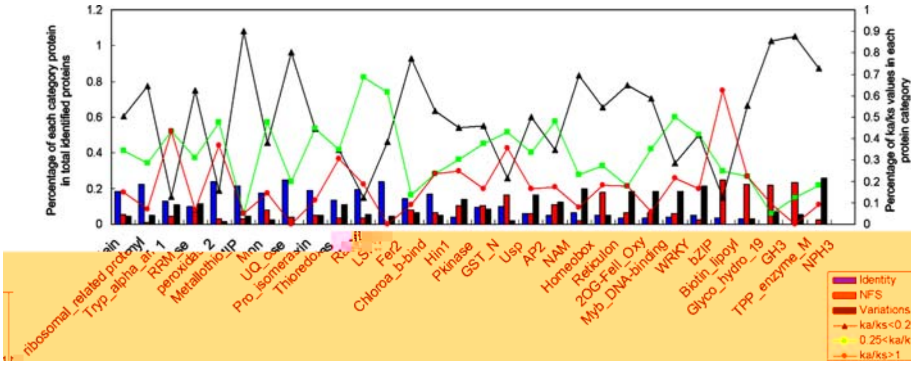
In addition, we searched 1,200 novel NCGR cDNAs against the PFAM protein database. The results showed that only 8.5% (102) of the 1,200 novel cDNAs matched proteins in PFAM database (p-score below 1e-10). As mentioned above, 39.9% (2,776) of the NCGR-KOME cDNA pairs matched PFAM proteins. Obviously,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



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 *K*a/*K*s > 1, *K*a/*K*s < 0.25 and 0.25 < *K*a/*K*s < 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*.

[illegible]

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 major QTL for grain length, is identified to encode a *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 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.

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).

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 abilit 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, Ha ashi aki Y (2001) Computer-based methods for the mouse full-length cDNA enc clopedia: real-time sequence clustering for construction of a non redundant cDNA librar . Genome Res 11:281 289

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

Kuhl JC, Cheung F, Yuan Q, Martin W, Ze die Y, McCallum J, Catanach A, Rutherford P, Sink KC, Jenderek M, Prince JP, To n CD, Have MJ (2004) Aunique set of 11,008 onion e pressed sequence tags reveals e pressed sequence and genomic differences bet een the monocot orders and Asparagales and Poales. Plant Cell 16:114 125

Le in B (2000) Genes VII. O ford Universit Press, O ford

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

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

Ma J, Bennet en JL (2004) Rapid recent gro th and divergence of rice nucleargenomes. Proc Natl Acad Sci USA 101:12404 12410

Nishi ama T, Fujita T, Shin-I T, Seki M, Nishide H, Uchi ama I, Kami a A, Carninci P, Ha ashi aki Y, Shino aki K, Kohara Y, Hasebe M (2003) Comparativegenomics of *Physcomiyrella patens* gametoph tic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. Proc Natl Acad Sci USA 100:8007 8012

Oden ald WF, Rasband W, Ku in A, Brod T (2005) EVOPRINTER, a multigenomic comparative tool for rapid identification of functionall important DNA. Proc Natl Acad Sci USA 102:14700 14705

Osato N, Itoh M, Konno H, Kondo S, Shibata K, Carninci P, Shiraki T, Shinaga a A, Araka a T, Kikuchi S, Sato K, Ka ai J, Ha ashi aki Y (2002) A computer-based method of selecting clones for a full-length cDNA project: simultaneouscollection of negligibl redundant and variant cDNAs. Genome Res 12:1127 1134

Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Su uki K, Ka ai J, Carninci P, Ohtomo Y, Murakami K, Matsubara K, Kikuchi S, Ha ashi aki Y (2003) Antisense transcripts ith rice full-length cDNAs. Genome Biol 5:R5

Ota T, Su uki Y, Nishika a T, Otsuki T, Sugi ama T, Irie R, Wakamatsu A, Ha ashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Oba ashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto J, Saito K, Ka ai Y, Isono Y,Nakamura Y, Nagahari K, Murakami K, Yasuda T, I a anagi T, Wagatsuma M,Shiratori A, Sudo H, Hosoiri T, Kaku Y, Kodaira H, Kondo H, Suga ara M, Takahashi M, Kanda K, Yokoi T, Furu a T, Kikka a E, Omura Y, Abe K, Kamihara K, Katsuta N, Sato K, Tanika a M, Yama aki M, Ninomi a K, Ishibashi T, Yamashita H, Muraka a 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, Togi a S, Komai F, Hara R, Takeuchi K, Arita M, Imose N, Musashino K, Yuuki H, Oshima A, Sasaki N, Aotsuka S, Yoshika a Y, Matsuna a H, Ichihara T, Shiohata N, Sano S, Mori a S, Momi ama H, Satoh N, Takami S, Terashima Y, Su uki O, Nakaga a S, Senoh A, Mi oguchi H, Goto Y, Shimi u F, Wakebe H, Hishigaki H, Watanabe T, Sugi ama A, Takemoto M, Ka akami B, Yama aki M, Watanabe K, Kumagai A, Itakura S, Fuku umi Y, Fujimori Y, Komi ama M, Tashiro H, Tanigami A, Fuji ara T, Ono T, Yamada K, Fujii Y, O aki K, Hirao M, Ohmori Y, Ka abata A, Hikiji T, Kobatake N, Inagaki H, Ikema Y, Okamoto S, Okitani R, Ka akami T, Noguchi S, Itoh T, Shigeta K, Senba T, Matsumura K, Nakajima Y, Mi uno T, Morinaga M, Sasaki M, Togashi T, O ama M, Hata H, Watanabe M, Komatsu T, Mi ushima-Sugano J, Satoh T, Shirai Y, Takahashi Y, Nakaga a 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 characteri ation of 21,243 full-length human cDNAs. Nat Genet 36:40 45

Panaud O, Vitte C, Hivert J, Mu lak S, Talag J, Brar D, Sarr A (2002) Characteri ation of transposable elements in the genome of rice (*Oryza sativa* L.) using representational difference anal sis (RDA). Mol Gen Genomics 268:113 121

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

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karam cheva S, Lee Y, White J, Cheung F, Parvi i B, Tsai J, Quackenbush J (2003) TIGR gene indices clustering tools (TGICL): a soft are s stem for fast clustering of large EST datasets. Bioinformatics 19:651 652

Rensink WA, Buell CR (2004) Arabidopsis to rice. Appl ing kno ledge from a eed to enhance our understanding of a crop species. Plant Ph siol 135:622 629

Rice P, Longden I, Bleasb A (2000) EMBOSS: the European molecular biolog open soft are suite. Trends Genet 16:276 277

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Kata ose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, Antonio BA, Kanamori H, Hosoka a S, Masuka a M, Arika a K, Chiden Y, Ha ashi M, Okamoto M, Ando T, Aoki H, Arita K, Hamada M, Harada C, Hijishita S, Honda M, Ichika a Y, Idonuma A, Iijima M, Ikeda M, Ikeno M, Ito S, Ito T, Ito Y, Ito Y, I abuchi A, Kami a K, Karasa a W, Katagiri S, Kikuta A, Koba ashi N, Kono I, Machita K, Maehara T, Mi uno H, Mi uba ashi 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, Shimoka a T, Shomura A, Song J, Taka aki Y, Terasa a K, Tsuji K, Waki K, Yamagata H, Yamane H, Yoshiki S, Yoshihara R, Yuka a K, Zhong H, I ama 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, Kami a A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Aki ama K, Oono Y, Muramatsu M, Ha ashi aki Y, Ka ai J, Carninci P, Itoh M, Ishii Y, Araka a T, Shibata K, Shinaga a A, Shino aki K (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. Science 296:141 145

Soga a K, Li Y, Zhang J, Liu G, Yao H (2003) Genealogical anal sis of resistance to the hitebacked planthipper *Sogatella furcifera* in Chinese *japonica* rice Chunjiang 06. Chinese J Rice Sci 17:67 72

Soltis DE, Soltis PS (2003) The role of ph logenetics in comparative genetics. Plant Ph siol 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

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

Su uki Y, Taira H, Tsunoda T, Mi ushima-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