Isolation and annotation of 10828 putative full length cDNAs from *indica* rice

XIE Kabin[1], ZHANG Jianwei[1], XIANG Yong[1], FENG Qi[2], HAN Bin[2], CHU Zhaohui[1], WANG Shiping[1], ZHANG Qifa[1] & XIONG Lizhong[1]

1. National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China;
2. National Center for Gene Research, Chinese Academy of Sciences, Shanghai 200233, China
Correspondence should be addressed to Xiong Lizhong (email: lizhongx@mail.hzau.edu.cn)

Received November 8, 2004

Abstract We reported the isolation and identification of 10828 putative full-length cDNAs (FL-cDNA) from an *indica* rice cultivar, Minghui 63, with the long-term goal to isolate all full-length cDNAs from *indica* genome. Comparison with the databases showed that 780 of them are new rice cDNAs with no match in *japonica* cDNA database. Totally, 9078 of the FL-cDNAs contained predicted ORFs matching with *japonica* FL-cDNAs and 6543 could find homologous proteins with complete ORFs. 53% of the matched FL-cDNAs isolated in this study had longer 5′UTR than *japonica* FL-cDNAs. *In silico* mapping showed that 9776 (90.28%) of the FL-cDNAs had matched genomic sequences in the *japonica* genome and 10046 (92.78%) had matched genomic sequences in the *indica* genome. The average nucleotide sequence identity between the two subspecies is 99.2%. A majority of FL-cDNAs (90%) could be classified with GO (gene ontology) terms based on homology proteins. More than 60% of the new cDNAs isolated in this study had no homology to the known proteins. This set of FL-cDNAs should be useful for functional genomics and proteomics studies.

Keywords: ***Oryza sativa* L.*, indica,* functional genomics.**

DOI: 10.1360/062004-90

Rice has become a model plant for genomic studies of monocot species, because of its relative small genome size (430 Mb), high synteny with other important crop species such as maize, barley and wheat, the release of draft sequences of both *indica*[1] and *japonica*[2] genomes, and the near completion of the map-based sequencing of rice genome by the International Rice Genome Sequencing Project. Currently, more than 340 Mb of non-overlapping genomic sequences including completely sequenced chromosomes 1[3], 2, 4[4], 7, 8 and 10[5], covering 85.6% of rice genome, have been released to public databases. With completion of rice genome sequencing, functional annotation of rice genome has become a new challenge to biologists. Although high throughput analysis of gene expression can be done with the aid of DNA chip technology, information from millions of expressed sequence tags (ESTs) in database as well as bioinformatics tools, complete and accurate annotation of functional genes rely largely on the information of full-length cDNAs (FL-cDNAs) since eukaryotic genes often have differential splicing of introns and unpredictable transcription starting sites, making gene prediction less accurate. Full-length cDNA, as a copy

of mRNA, can provide not only the structure information of the gene but also the complete sequence for functional studies. Therefore, a full set of full-length cDNAs from rice is invaluable in functional annotation of the rice genome.

Large scale isolation and annotation of FL-cDNAs has been reported in a few model species including 60770 FL-cDNAs from *Mus musculus*[6,7], 14034 FL-cDNAs from *Arabidopsis thaliana*[8], 15000 FL-cDNAs from *Mus musculus* and *Homo sapiens*[9], 10910 FL-cDNAs from *Drosophila melanogaster*[10,11] and 21143 FL-cDNAs from *Homo sapiens*[12]. In addition, FL-cDNA projects have been initiated in various species such as *Danio rerio*, *Xenopus laevis* and *Ciona intestinalis* (http://www.ncbi.nlm.nih.gov/genome /flcdna/). Meanwhile, three Japanese groups, NIAS, FAIS and RIKEN, jointly reported a set of more than 28000 ORF-contained putative full-length cDNAs from *japonica* rice as a part of their KOME (Knowledge-based *Oryza* Molecular Biological Encyclopedia) project[13] and the number of FL-cDNAs has recently reached 32127 (http://cdna01.dna.affrc.go.jp/cDNA/) that has covered more than half of the predicted genes in *japonica* genome. Nevertheless, large scale of FL-cDNAs isolation has not been reported in the economically more important subspecies, *Oryza sativa* L. ssp. *indica*, that is the major cultivated type in China.

As a part of the National Rice Functional Genomics Project in China (RFGC), we report here 10828 putative FL-cDNAs isolated from three cDNA libraries including a normalized full-life-cycle cDNA library[14] of Minghui 63, one of the most important *indica* rice in China and the proposed material by RFGC initiative.

1 Materials and methods

1.1 Construction of cDNA libraries and sequencing

The *indica* rice Minghui 63 used in this study is the restorer line for a number of elite rice hybrids including Shanyou 63 that has been the most widely cultivated hybrid in China during the last two decades. Total RNA and mRNA extraction was performed following the manufacture instructions of Trizol^{TM} reagent (Invitrogen) and Dynal beads oligo d(T)$_{25}$ (Dynal, Norway) respectively. The normalized full-life-cycle cDNA library[14] was constructed following the protocol of Plasmid System for cDNA Synthesis and Plasmid Cloning^{TM} (Invitrogen). Full-length-enriched cDNA libraries of pollen and young panicles (less than 1 cm in length) were constructed according to the instruction of GeneRacer kit (Invitrogen). A total of 62208 cDNA clones were sequenced with single-pass from the 5′-ends using BigDye^{TM} Terminator Cycle Sequencing Ready Reaction kit. Raw sequences were collected using Phred software[15,16]. Vector or other junk sequences were removed by CROSS MATCH software (http://www.phrap.org/).

1.2 Sequence analysis and identification of FL-cDNA

For an efficient sequence analysis to identify FL-cDNA, a sequence analysis system (http://redb.croplab.org/) was established based on Linux platform, bioinformatics tools (such as Phred/Phrap/Copnsed and BLAST) and local rice database including updated nucleotide and protein sequences from public databases. Homology search of rice cDNAs was performed by software BLAST2.2.9[17] and cDNAs were clustered by using software EST-Clustering[18] and CAP3[19]. The longest cDNA in each cluster was picked out for further analysis to identify FL-cDNA (Fig. 1).

Two methods as illustrated in Fig. 1 were used to determine putative FL-cDNAs. The first one is based on homology proteins. All unique cDNAs from Minghui 63 were searched against updated protein databases (PDB, SwissProt, PIR, PRF and predicted protein sequences) by BLASTX, and significant (with a threshold of Score value more than 80) homologies were checked at the N-termini of protein sequences. The sequence was considered to be of FL-cDNA if the cDNA was predicted to encode peptide with complete N-terminus as referenced to the homologous sequence. The second method is based on direct sequence comparison of cDNA from Minghui 63 against *japonica*

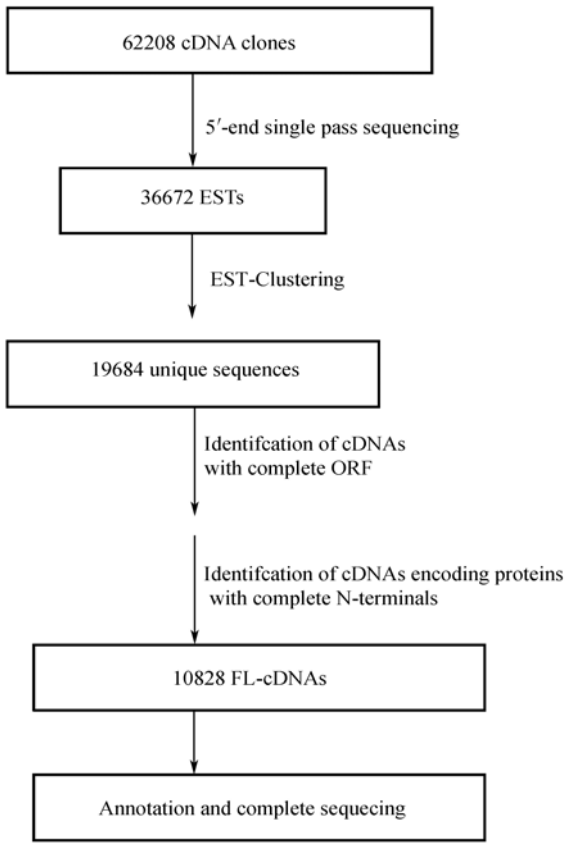

Fig. 1. Flow chart of sequence analysis.

FL-cDNAs. Putative FL-cDNA was claimed if the sequence has covered the 5′-end of the ORF (opening reading frame) of the *japonica* homologue (with threshold of E-value less than 1E-5). Putative FL-cDNAs identified by these two methods were sequenced from their 3′-end. FL-cDNAs with complete 3′-end were then completely sequenced.

FL-cDNAs were compared against rice genomic sequences by BLASTN (with threshold of E-value less than 9E-30 and sequence identity more than 98%) for determining chromosomal locations (*in silico* mapping). Putative functions of FL-cDNAs from Minghui 63 were deduced from their homologues based on various analyses: (i) BLASTN against FL-cDNAs in KOME database and *OSGI* (*Oryza sativa* Gene Index) from TIGR database; (ii) BLASTX against protein databases; and (iii) GO[20,21] classification based on terms of homologous genes or proteins.

2 Results

2.1 cDNA libraries and sequencing

Three cDNA libraries were constructed including the normalized full-life-cycle cDNA library[14] and two full-length-enriched libraries from pollen and young panicle respectively. The average insert size was about 1.4 kb. A total of 62208 clones were sequenced with single-pass from 5′-end, resulting in 36672 cDNAs each with accurate rice sequence longer than 200 bp. Poor sequences (less than 200 bp or inaccurate sequences) were not included in further analysis. The average length of sequences was 582 bp with accuracy of 99.81% based on comparison of redundant sequences. By using EST-Clustering and CAP3, 19684 unique (non-redundant) cDNAs were identified with 4592 clusters each containing two or more clones and 15092 clusters each containing a single clone.

2.2 Identification of FL-cDNAs

Comparison of 19684 unique cDNAs from Minghui 63 with KOME cDNAs identified 18377 cDNAs showing significant homology with the average sequence identity of 97.14%. A total of 9078 cDNAs covered the 5′-end of predicted ORFs of *japonica* cDNAs. Among them, 53% cDNAs had longer 5′-UTR (5′-end untranslated region) than the *japonica* homologues, whereas 46.8% cDNAs had shorter 5′-UTR than the *japonica* counterparts (Fig. 2).

By BLASTX analysis, a total of 17504 unique cDNAs from Minghui 63 found significant homologous protein sequences. Among them, 6543 cDNAs were predicted to encode peptides containing complete N-termini of the homologous proteins. The average identity was 70.49% on the basis of amino acids. Of the 6543 cDNAs, 4792 were also identified to be FL-cDNAs by comparison with *japonica* FL-cDNAs. Combining the results of the two methods, a total of 10828 cDNAs from Minghui 63 were considered to be putative FL-cDNAs (Fig. 3).

By scanning all putative FL-cDNAs for start codon (ATG), 10604 cDNAs were identified with start codon at 5′-end and 82.4% of them had the featured start codon sequence ATGG. Based on the predicted start

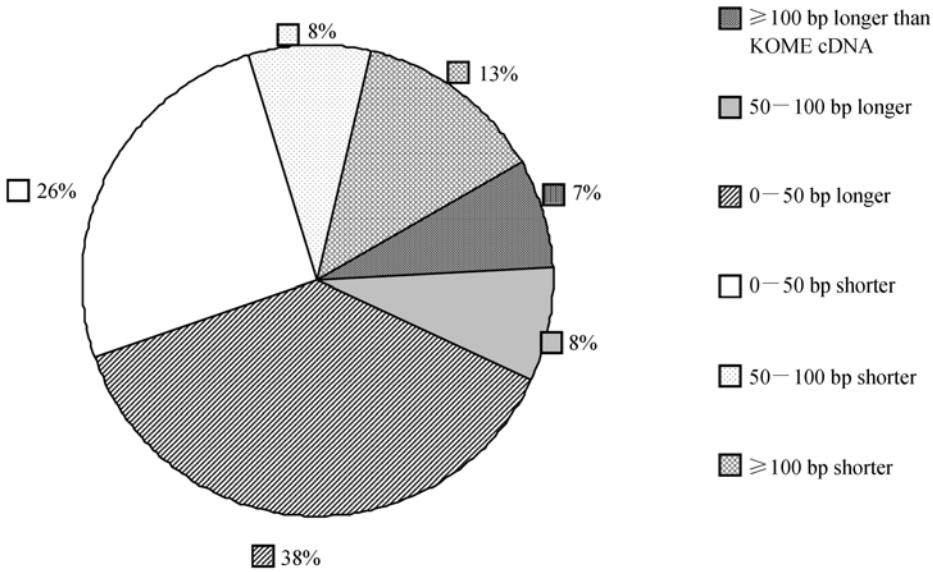

Fig. 2. Comparison of FL-cDNAs from Minghui 63 with *japonica* counterparts for the length of 5′-UTR.

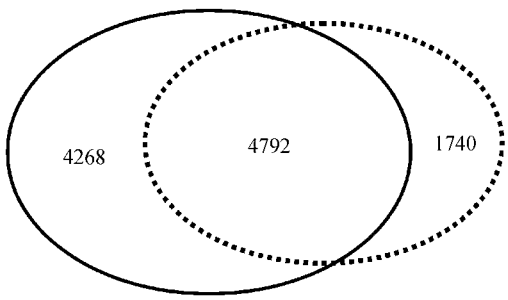

Fig. 3. 10828 FL-cDNAs from Minghui 63. 6432 (dotted oval circle) were identified based on homology proteins with complete N-terminals; 9078 (solid oval circle) were identified based on comparison with the ORFs of japonica FL-cDNAs.

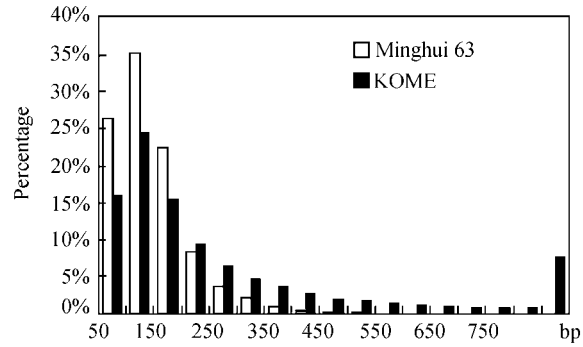

Fig. 4. Distribution of the 5′-UTR length of FL-cDNAs from Minghui 63 and *japonica* FL-cDNAs.

codon, the length of 5′-UTR was compared between *japonica* and *indica* cDNAs. The result showed that the distribution of 5′-UTR length was very similar between the two subspecies, with 85% of the 5′-UTR in the range of 100—300 bp (Fig. 4).

2.3 Genomic location of FL-cDNAs from Minghui 63

By sequence alignment of the FL-cDNAs from Minghui 63 against genomic sequences with a threshold E-value of 9E-30 and sequence identity more than 98%, 90.28% (9,776) and 92.78% (10,046) of FL-cDNAs were *in silico* mapped to the *japonica* and *indica* genomes, respectively. A minor portion of cDNAs without matches in genomic sequences may be partly due to the incomplete sequencing and /or release of rice genomic at the time of data analysis. A total of 8216 FL-cDNAs from Minghui 63 and their *japonica* counterparts were mapped to the same genomic regions (BAC sequences used). The average sequence identity of these cDNAs between *japonica* and *indica* was 99.2% with single nucleotide polymorphism (SNP) about 0.26%. By BLASTN analysis of the FL-cDNAs against all rice cDNAs and ESTs in public databases

including GenBank, KOME and OsGI, 780 FL-cDNAs (7.2%) from Minghui 63 had no match, which may represent novel transcription units (TU) or genes.

2.4 Annotation of FL-cDNAs

Functional prediction of the FL-cDNAs from Minghui 63 was based on their homologies (at a threshold E-value of 10E-5) with sequences in various genomic databases including TAIR (The *Arabidopsis* Information Resources), SGD (*Saccharomyces* Genomic Database), MGI (Mouse Genome Informatics), UNIPROT, KOME and TIGR Gene Index. The FL-cDNA genes were classified using GO (Gene ontology) into three major groups (Table 1): biological processes (3186), cellular components (4288) and diverse molecular functions (8956). More than 40% of the cDNA genes were assigned with two or more GO terms. Based on GO classification, about 52.4% of the FL-cDNAs were classified in the category of unknown function.

The putative functions of the 780 new FL-cDNAs from Minghui 63 were further analyzed based on MIPS (Munich Information on Protein Sequences, http://mips.gsf.de/proj/thal/db/tables/tables_func_frame.html) classification system. These cDNA genes were classified into 10 categories including a predominant (60.9%) group with unknown function (Table 2). Genes involved in transcription regulation (12.3%), cellular communication (6.9%), cell rescue and defense (6.5%) and development (4.7%) had much larger proportions than other five categories including cellular organization (3.3%), metabolism (1.5%), energy (0.5%), transport facilitation (1.2%), and protein fate (1.0%).

Table 1 Gene ontology classification of FL-cDNAs from Minghui 63

GO term	GO ID	No. cDNAs[a)]	Percentage
A. Biological process			
Biological process unknown	4	1983	62.26%
Development	7275	989	31.05%
Physiological process	7582	166	5.21%
Regulation of biological process	50789	47	1.48%
Total		3185	100%
B. Cellular component			
Cellular component unknown	8372	2118	49.39%
Cell	5623	1918	44.73%
Extracellular	5576	195	4.55%
Unlocalized	5941	55	1.28%
Virion	19012	2	0.05%
Total		4288	100%
C. Molecular function			
Binding	5488	2959	33.04%
Catalytic activity	3824	2423	27.05%
Molecular function unknown	5554	1868	20.86%
Transporter activity	5215	829	9.26%
Triplet codon-amino acid adaptor activity	30533	307	3.43%
Transcription regulator activity	30528	189	2.11%
Translation regulator activity	45182	120	1.34%
Signal transducer activity	4871	79	0.88%
Nutrient reservoir activity	45735	73	0.82%
Enzyme regulator activity	30234	56	0.63%
Motor activity	3774	33	0.37%
Antioxidant activity	16209	20	0.22%
Total		8956	100%

a) Some cDNAs have more than one GO annotation.

Table 2 Classification of 780 novel FL-cDNAs from Minghui 63

Category[a)]	Number of FL-cDNAs	Percentage (%)
Metabolism	12	1.5
Cellular organization	26	3.3
Transport facilitation	17	2.2
Protein fate	8	1.0
Development	37	4.7
Cell rescue and defense	51	6.5
Energy	4	0.5
Transcriptional regulation	96	12.3
Cellular communication	54	6.9
Unknown	475	60.9
Total	780	100.0

a) According to MIPS (http://mips.gsf.de/proj/thal/db/tables/tables_func_frame.html) classification system.

3 Discussion

The FL-cDNAs isolated in this study were mainly from a normalized full-life-cycle cDNA library[14] of *indica* rice Minghui 63. The original purpose for constructing such a library was to identify new expressed sequences and the library construction was essentially based on the conventional adapter-orientated method in which both complete and incomplete cDNAs were cloned. Compared with the reported full-length-enriched method such as Oligo-capping[22] and Cap-trapper[23,24], the percentage of FL-cDNAs in this library (approximately 40%) was not so high. Full-length-enriched libraries of pollen and young panicle were also constructed in this study based on a modified Oligo-capping method with the Gateway™ recombination technology (Invitrogen) incorporated. The proportion of FL-cDNAs in these libraries can be as high as more than 90%, however, the redundancy of abundant genes was 3—5 fold higher than cDNA library constructed by the conventional methods, which may result from the PCR amplification involved in this method (data not shown). An efficient normalization method combined with full-length- enriched cDNA synthesis may help to accelerate isolating more novel full-length cDNAs in rice.

The FL-cDNAs isolated in this study contained complete ORFs and predicted 5′-UTR sequences. However, it is hard to determine which one is a biologically complete cDNA considering the complexity of eukaryotic mRNA processing. Experiments such as RACE (rapid amplification of cDNA end) and primer extension can be done to determine a biologically complete cDNA. However, such experiments can hardly be performed in a high throughput manner at present. To date, thousands of FL-cDNAs (including *japonica* FL-cDNAs[13]) in database are actually cDNAs with complete ORFs. We used the same criteria (complete ORF) to determine full-length-like cDNAs in this study. It should be pointed out that some of these FL-cDNAs are already biologically complete full-length cDNAs. For example, by comparison of the FL-cDNAs from Minghui 63 with experimentally verified rice full-length cDNAs in GenBank, 92 of the 146 matches showed almost identical length at the 5′-end with length difference less than 5 bp. Moreover, complete ORF of cDNAs provide basis not only for prediction of protein sequences and upstream regulatory sequences but also for multiple reverse genetics approaches (such as overexpression and ectopic expression) and *in vitro* experiments to reveal gene function. Therefore, FL-cDNAs with complete ORF from whole genome is one of the most important platforms for functional genomics studies.

Discovering the transcriptome difference between *indica* and *japonica*, two subspecies of cultivated rice, has both theoretical and practical significance. Currently more than 30000 FL-cDNAs have been identified in *japonica* rice Nipponbare, whereas FL-cDNAs from *indica* rice is very limited. Although only 10828 FL-cDNAs were isolated in this study, these cDNAs represented a reasonable *indica* collection for transcriptome comparison between two subspecies. Considerable high level of SNP (0.26%) was detected within the transcribed region of the two genomes. In-

terestingly, 3.4% of cDNAs (among 8216 FL-cDNAs compared) was detected with differential splicing between *indica* and *japonica*. Different transcripts from same gene by differential splicing, as an important regulatory means of gene expression[25,26], may contribute to the differentiation of rice subspecies.

Among the 10828 FL-cDNAs from Minghui 63, 1700 had no match in *japonica* FL-cDNA database (high homologies but with different chromosomal locations were considered different genes) and 780 had no match of rice EST in databases. It is likely that some of the genes corresponding to these cDNAs may also be expressed in the *japonica* genome but their FL-cDNAs have not been isolated yet, since FL-cDNAs for about one third of predicted genes have not been obtained. Considering the smaller collection of *indica* FL-cDNAs in this study, however, the relatively large proportion of novel cDNAs may suggest that some cDNAs were transcribed in one subspecies but not in another, and *vice versa*. Further isolation and characterization of these genes may help to reveal the molecular basis of differentiation between these two subspecies.

Acknowledgements This work was supported by National Science and Technology Special Key Project —— Functional Genomics and Biological Chip and National Natural Science Foundation of China (Grant No. 30321005).

References

1. Yu, J., Hu, S., Wang, J. et al., A draft sequence of the rice genome (*Oryza sativa* L. *ssp. indica*), Science, 2002, 296: 79—92.[DOI]
2. Goff, S. A., Ricke, D., Lan, T. H. et al., A draft sequence of the rice genome (*Oryza sativa* L. *ssp. japonica*), Science, 2002, 296: 92—100.[DOI]
3. Sasaki, T., Matsumoto, T., Yamamoto, K., et al., The genome sequence and structure of rice chromosome 1, Nature, 2002, 420: 312—316.[DOI]
4. Feng, Q., Zhang, Y., Hao, P. et al., Sequence and analysis of rice chromosome 4, Nature, 2002, 420: 316—320.[DOI]
5. Rice Chromosome 10 Sequencing Consortium, In-depth view of structure, activity, and evolution of rice chromosome 10, Science, 2003, 300: 1566—1569.[DOI]
6. Okazaki, Y., Furuno, M., Kasukawa, T. et al., Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, Nature, 2002, 420: 563—573.[DOI]
7. Kawai, J., Shinagawa, A., Shibata, K. et al., Functional annotation of a full-length mouse cDNA collection, Nature, 2001, 409: 685v690.
8. Seki, M., Narusaka, M., Kamiya, A. et al., Functional annotation of a full-length *Arabidopsis* cDNA collection, Science, 2002, 296: 141—145.[DOI]
9. Strausberg, R. L., Feingold, E. A., Grouse, L. H. et al., Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences, Proc. Natl. Acad. Sci. USA, 2002, 99: 16899—16903.[DOI]
10. Stapleton, M., Carlson, J., Brokstein, P. et al., A Drosophila full-length cDNA resource, Genome Biol., 2002, 3: RESEARCH0080.
11. Stapleton, M., Liao, G., Brokstein, P. et al. The *Drosophila* gene collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes, Genome Res, 2002, 12: 1294—1300.[DOI]
12. Ota, T., Suzuki, Y., Nishikawa, T. et al., Complete sequencing and characterization of 21,243 full-length human cDNAs, Nat. Genet., 2004, 36: 40—45.[DOI]
13. Kikuchi, S., Satoh, K., Nagata, T. et al., Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice, Science, 2003, 301: 376—379.[DOI]
14. Chu, Z. H., Peng, K. M., Zhang, L. D. et al., Construction and characterization of a normalized whole-life-cycle cDNA library of rice, Chinese Science Bulletin, 2003, 48: 229—235.
15. Ewing, B., Hillier, L., Wendl, M. C. et al., Base-calling of automated sequencer traces using phred (I): Accuracy assessment, Genome Res., 1998, 8(3): 175—185.
16. Ewing, B., Green P., Base-calling of automated sequencer traces using phred (II): Error probabilities, Genome Res., 1998, 8: 186—194.
17. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res., 1997, 25: 3389—3402.[DOI]
18. Zhang, L. D., Yuan, D. J., Zhang, J. W. et al., A new method for EST clustering, Yi Chuan Xue Bao, 2003, 30: 147—153.
19. Huang, X., Madan, A., CAP3: A DNA sequence assembly program, Genome Res., 1999, 9: 868—877.[DOI]
20. Ashburner, M., Ball, C. A., Blake, J. A. et al., Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet., 2000, 25: 25—29.[DOI]
21. Harris, M. A., Clark, J., Ireland, A. et al., The Gene Ontology (GO) database and informatics resource, Nucleic Acids Res., 2004, 32 Database issue: D258—261.
22. Maruyama, K., Sugano, S., Oligo-capping: A simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, Gene, 1994, 138: 171—174.[DOI]
23. Carninci, P., Kvam, C., Kitamura, A. et al., High-efficiency full-length cDNA cloning by biotinylated CAP trapper, Genomics, 1996, 37: 327—336.[DOI]
24. Carninci, P., Shibata, Y., Hayatsu, N. et al., Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes, Genome Res., 2000, 10: 1617—1630.[DOI]
25. Cartegni, L., Chew, S. L., Krainer, A. R., Listening to silence and understanding nonsense: Exonic mutations that affect splicing, Nat. Rev. Genet., 2002, 3: 285—298.[DOI]
26. Taneri, B., Snyder, B., Novoradovsky, A. et al., Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific, Genome Biol., 2004, 5: R75.[DOI]