The Plant Journal (2005) **42**, 772–780 doi: 10.1111/j.1365-313X.2005.02408.x

Features of the expressed sequences revealed by a large-scale analysis of ESTs from a normalized cDNA library of the elite *indica* rice cultivar Minghui 63

Jianwei Zhang[1], Qi Feng[2], Caoqing Jin[2], Deyun Qiu[1], Lida Zhang[1], Kabin Xie[1], Dejun Yuan[1], Bin Han[2], Qifa Zhang[1] and Shiping Wang[1,*]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China, and
[2]National Center for Gene Research, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China

Received 7 December 2004; revised 26 February 2005; accepted 7 March 2005.

*For correspondence (fax +86 27 87287092; e-mail swang@mail.hzau.edu.cn).

Summary

The *indica* subspecies of cultivated rice occupies the largest area of rice production in the world. However, a systematic analysis of cDNA sequences from the *indica* subspecies has not been performed. The aim of the present study was to collect and analyze the expressed sequence tags (ESTs) of *indica* rice on a large scale. A total of 39 208 raw sequences were generated from a normalized cDNA library prepared by use of 15 different tissues of the *indica* cultivar Minghui 63. After trimming, processing and analysis, 17 835 unique sequences were obtained, each of which presumably represents a unique gene. Of these sequences, 2663 were novel, and at least 70 were *indica* specific. Comparison of the Minghui 63 sequences with the ESTs/full-length cDNAs in GenBank revealed a large number of deletion/insertion/substitution (DIS) at both the inter- and intra-subspecific levels. The overall number of polymorphisms in the expressed sequences was higher in the inter-subspecific comparisons than in the intra-subspecific comparisons. However, the extent of DIS-based polymorphism was highly variable among different rice varieties. In total, 15 726 unique sequences, including 697 novel sequences, were assigned to regions where large numbers of quantitative trait loci (QTLs) for agronomic traits had been detected previously. These results may be useful for developing new molecular markers for genetic mapping, detecting allelic polymorphisms associated with phenotypic variations between rice varieties, and facilitating QTL cloning by providing the starting points for candidate-gene identification.

Keywords: ***indica* rice, expressed sequence tag, single nucleotide polymorphism, Indel, quantitative trait locus, mutator-like transposase.**

Introduction

Rice is a major crop that feeds about half of the human population of the world. Rice has also become a model system for genomic research in cereal plants because of its relatively small genome size and near completion of the genome sequencing. The rice genome was first estimated to have 32 000–56 000 genes, on the basis of whole-genome shotgun sequence data (Goff *et al.*, 2002; Yu *et al.*, 2002a). Further studies have indicated that the number of genes in the rice genome has been overestimated because of the existence of transposable elements or transposable element fragments, suggesting that the number of genes is <40 000 (see review by Bennetzen *et al.*, 2004). Recently, using an improved version of rice genome sequences that brings the coverage of the *indica* rice genome from the original 4.2× (Yu *et al.*, 2002a) to 6.28× and a new gene-prediction procedure that removes erroneous predictions resulting from the presence of transposable elements, Yu *et al.* (2005) reestimated that the rice genome contains at least 38 000–40 000 genes. Only a very small portion of the rice genes has been experimentally studied; currently, the process of gene identification is difficult because none of the existing gene-prediction programs can identify gene structures with satisfactory accuracy (Mathe *et al.*, 2002; Rogic *et al.*, 2002). Thus, aligning predicted genes with expressed sequence tags (ESTs) has become the most practical and reliable strategy for accurate gene prediction and annotation.

Asian cultivated rice (*Oryza sativa* L.) consists of two major groups, which are known by the subspecies names *indica* and *japonica*. Approximately 285 758 ESTs and 28 469 full-length cDNA sequences (The Rice Full-Length cDNA Consortium, 2003) of rice are currently available in GenBank (http://www.ncbi.nlm.nih.gov, release 071604). Most (63.6%) of the ESTs and all the full-length cDNAs available in GenBank are from *japonica* cultivars, while the remaining ESTs are from a number of *indica* varieties. However, large numbers of ESTs in the database are redundant. Thus, identification of new ESTs is an approach to annotation of the rice genes. Moreover, a comparison of DNA sequences, covering approximately the 2.3 Mb region, between an *indica* cultivar and a *japonica* cultivar revealed that genomic divergence occurred as differences in total number of genes, single-nucleotide polymorphisms (SNPs) and insertion/deletions (Indels) in coding regions (Feng *et al.*, 2002; Han and Xue, 2003). Therefore, a large-scale analysis of the expressed sequences from *indica* cultivars not only may lead to the identification of genes that are not present in the *japonica* genome but also will reveal differences in the type and amount of polymorphisms in the coding sequences between the two subspecies.

An important characteristic of gene regulation in eukaryotes is that large portions of the genes are expressed only in specific tissues and/or at certain growth stages. In addition, many genes are expressed only in special situations such as exposure to environmental stresses. Thus, obtaining all or most of the expressed genes in one or a few experiments is difficult. To obtain expressed sequences of *indica* rice on a large scale, with an emphasis on rarely expressed genes, we constructed a normalized cDNA library by use of 15 tissues harvested at nine developmental stages, including tissues treated with biotic and abiotic stresses, from the *indica* cultivar Minghui 63 (Chu *et al.*, 2003). Inverse Northern blot analysis showed that this cDNA library contained many rarely expressed sequences. Thus, it is a valuable source for the identification of novel ESTs. Furthermore, Minghui 63 is the restorer line for a number of planted rice hybrids that collectively have comprised >20% of the total rice production area in China during the last 2 decades. The hybrids produced with Minghui 63 have a number of desirable characteristics, including a high yield and wide adaptability. Thus, characterization of the genome of Minghui 63 at the level of gene expression will facilitate the identification of genes that control important agronomic traits.

The present study was undertaken to: (i) analyze the sequences of the cDNA clones in the normalized cDNA library described above, (ii) characterize the polymorphisms of expressed sequences at both inter- and intra-subspecific levels, and (iii) examine the relationships between the ESTs and quantitative trait loci (QTLs) associated with agronomically important traits. It is expected that the large-scale analysis of ESTs from such an *indica* cultivar will help in the annotation of the *indica* rice genome, facilitate identification of new genes, and provide data for comparative studies of the two subspecies. The results of this study may also serve as a starting point for the identification of genes of QTLs for various traits.

Results and discussion

Sequence analysis of the cDNA library

Random sequencing of >40 000 cDNA clones of the normalized cDNA library composed of clones from 15 different tissues generated 39 208 raw sequences. The sequences were trimmed by removing those that showed homology to sequences of *Escherichia coli*, *Xanthomonas axonopodis* pv. *citri*, *Xanthomonas campestris* pv. *campestris*, *Magnaporthe grisea* or vectors and those that were <100 bp. A total of 36 672 trimmed sequences were obtained, ranging in length from 101 to 1380 bp, with an average length of 583 bp. In total, 6006 of the trimmed sequences had poly(A/T) stretches.

Clustering and assembling of these sequences by use of a modified EST clustering program (Zhang *et al.*, 2003) produced 4592 contigs and 15 092 singletons. The sequences covering entire spans were taken from the contigs, and consensus sequences were used for regions with overlapping ESTs, resulting in a total of 19 684 processed sequences (GenBank accession nos CX099458–CX119141) that varied in length from 101 to 3082 bp, with an average length of 691 bp. Poly(A/T) stretches were detected in 3679 of the processed sequences. For identification, most of the processed sequences from the singletons were named with the prefix EI or BI (e.g., EI079O08 and BI120F05), and a small number of the processed sequences from the singletons was named according to origin, such as POLLN for pollen, PANIC for panicle at flowering stage, DEVEP for panicle at panicle-development stage, FLAGL for flag leaves, and THREE for whole plant at the three-leaf stage, while the processed sequences from the contigs were renamed with the prefix RECm (e.g., RECm0001, RECm0002, etc.). All the sequences are available at the web site for the Rice EST DataBase (REDB) [http://redb.ricefgchina.org (or) http://bioinformatics.hzau.edu.cn].

For better correspondence between the ESTs and the genes, the above-mentioned processed sequences were reclustered with reference to the 59 712 nucleotide sequences of the rice gene model from the TIGR web site (http://www.tigr.org/tdb/e2k1/osa1/). Although the gene model may have overestimated the number of rice genes, in accordance with the recent estimation of rice genes (Yu *et al.*, 2005), the reclustering reduced the 19 684 processed sequences to 17 835 sequences, which were referred to as unisequences because each of them presumably represents a unique gene. We designated the sequences joined by the reclustering with the prefix RECn (e.g., RECn0011) and filled in the missing sequences between the ends of the ESTs, as indicated by the gene models, with the same lengths of 'X' segments.

Of the 17 835 unisequences, 12 762 were single copies, 3396 (19%) were generated from two to three overlapping sequences, 1237 (7%) were from four to eight overlapping sequences, and 440 (<3%) were based on >8 overlapping ESTs. The redundancy of this library was approximately 51.4%. The largest cluster (RECm0300) was formed of 508 ESTs spanning 891 bp in length. It was homologous with a gene coding for a Zn-induced protein (RezA) (GenBank accession no. U46138; *E*-value = 0). The longest cluster (RECm0103) was 3082 bp in length and was formed of 225 ESTs corresponding to the rice 25S ribosomal RNA gene (GenBank accession no. M11585; *E*-value = 0). As the cDNA library used for the present study was normalized through hybridization of cDNA with saturated genomic DNA (Chu *et al.*, 2003), the frequency of each gene sequence in the normalized library reflected both the expression level in the tissues and the number of homologous regions in the genomic DNA. Both RECm0300 and RECm0103 showed high levels of expression (Figure 1). Analysis also indicated that both sequences detected multiple homologous regions in the rice genome. For example, RECm0300 detected 58 homologous regions in the rice genome, ranging from 30 to 781 bp in length, while RECm0103 detected 25 homologous regions, ranging from 30 to 1167 bp in length. Thus, the high frequencies of these ESTs in the library could be attributed both to their high levels of expression and to multiple homologous regions in the rice genome.

Analysis of the 17 835 unisequences, by use of the BLASTN program, revealed that 2663 (14.9%) of them had no match or a poor match (*E*-value > 10^{-5}) with 285 758 rice EST and 28 469 full-length cDNA (The Rice Full-Length cDNA Consortium, 2003) entries in GenBank. As the ESTs in this study were from a cDNA library constructed with tissues from nine developmental stages and challenged with biotic and abiotic stresses (Chu *et al.*, 2003), these novel ESTs may mainly represent tissue-specific genes or stress-responsive genes.

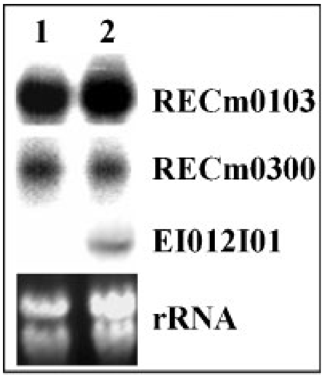

Figure 1. Expression patterns of three expressed sequence tags (ESTs) in rice cultivar Minghui 63.
The highly redundant ESTs, RECm0103 and RECm0300, in cultivar Minghui 63 (lane 1) were examined by use of cDNA clones EI073H13 and EI016A06, which carried part of the sequences of RECm0103 and RECm0300 as probes, respectively. The expression of the single-copy EST EI012I01 could only be detected in transgenic rice with the genetic background of Mudanjiang 8 (*Oryza sativa* ssp. *japonica*) (lane 2), in which EI012I01 was overexpressed.

Functional classification of the ESTs

In total, 16 954 (95.1%) of the 17 835 unisequences were classified according to the Gene Ontology (GO) database (Gene Ontology Consortium, 2001), by use of the BLASTX program with the default criterion for sequence alignment (Table 1). Of the 16 954 unisequences, 9140, 12 228 and 10 126 were classified by the GO terms 'molecular function,' 'biological process' and 'cellular component,' respectively. Using the classification of 'molecular function,' the unisequences representing 9140 genes were divided into 13 groups (Table 2); 87.3% of the classified unisequences had known molecular functions and the remaining 12.7% were in the category 'molecular function unknown.'

Of the 2663 novel unisequences, 1421 were classified by the GO term 'molecular function,' 1876 by the term 'biological process' and 1463 by the term 'cellular component,'

Table 1 Criteria for sequence alignment

Threshold	*E*-value	Alignment length (bp)	Nucleotide identity[a]
Default	10	–	–
Stringent	≤10^{-5}	≥40	≥94%

[a] Identity of overlapping regions.

Table 2 Functional classification of the unisequences by the Gene Ontology database term 'molecular function'

	No. of unisequences[a]	
Category	Total	Novel
Enzyme	2163 (23.7)	239 (16.8)
Transcription regulator	1528 (16.7)	248 (17.5)
Transporter	1296 (14.2)	192 (13.5)
Molecular function unknown	1165 (12.7)	224 (15.8)
Ligand binding or carrier	1016 (11.1)	139 (9.8)
Signal transducer	910 (10.0)	205 (14.4)
Structural molecule	449 (4.9)	84 (5.9)
Enzyme regulator	166 (1.8)	45 (3.2)
Obsolete	193 (2.1)	20 (1.4)
Translation regulator	130 (1.4)	11 (0.8)
Antioxidant	77 (0.8)	3 (0.2)
Motor	31 (0.3)	10 (0.7)
Chaperone	16 (0.2)	1 (0.1)
Total	9140 (100)	1421 (100)

[a] Numbers in parentheses are proportions (%) of the total unisequences classified according to molecular function.

involving a total of 2335 (87.7%) of the unisequences. Compared with the categories formed by the classification of total unisequences by 'molecular function,' the proportions of novel sequences in the categories 'signal transducer' and 'molecular function unknown' were increased, while the proportion in the category 'enzyme' was greatly reduced (Table 2), suggesting that the normalized library enhanced the discovery of rarely expressed new genes.

Inter- and intra-subspecific polymorphisms of the expressed sequences

Two classes of expressed sequence polymorphisms, SNPs and Indels, were analyzed among different rice varieties by comparing the consensus sequences of Minghui 63, as described above, with the rice ESTs/full-length cDNAs from GenBank. In GenBank, approximately 64% of the rice ESTs/full-length cDNAs are from 25 *japonica* varieties, and the remaining ESTs are from 22 *indica* varieties. This rich source provided an opportunity for examining polymorphisms of the gene sequences at both inter- and intra-subspecific levels. Detailed deletion/insertion/substitution (DIS) information for such comparisons is available in the Rice DIS Database of the REDB (http://redb.ricefgchina.org/cgi-bin/dis.pl).

Three analyses were performed for comparisons (Table 3). In the mEST2EST analysis, each consensus sequence of Minghui 63 was compared simultaneously with multiple homologous sequences from GenBank. In this analysis, the numbers of sequences compared with each Minghui 63 sequence ranged from two to 261 for inter-subspecific comparisons and from two to 149 for intra-subspecific comparisons. A difference was considered to be a polymorphism only when all sequences from GenBank differed from the consensus sequence of Minghui 63 at the same site. The analysis showed that, in general, DISs occurred more frequently in inter-subspecific comparisons than in intra-subspecific comparisons (Table 3). Approximately 73.3% (275) and 76.1% (105) of the Indels that were detected in inter- and intra-subspecific comparisons, respectively, were 1 bp in length. Indels of 2–4 bp in length occurred in 20.8% (78) of the inter-subspecific comparisons and in 19.6% (27) of the intra-subspecific comparisons. The longest Indels detected were 10 bp in the inter-subspecific comparisons and 8 bp in the intra-subspecific comparisons. Most of the SNPs were detected as two-alternative-base substitutions, in which a nucleotide in the Minghui 63 sequence was substituted by either of two nucleotides in all the homologous sequences. One-alternative-base substitutions, in which a nucleotide in the Minghui 63 sequence was substituted by only one other nucleotide in all the homologous sequences, was observed only in 15.2% (334) of the SNPs in the inter-subspecific comparisons and in 17.1% (139) of the SNPs in the intra-subspecific comparisons. SNPs with three-alternative-base substitutions were not observed. As each DIS was identified by comparison of the Minghui 63 sequence and multiple aligned ESTs/full-length cDNAs from GenBank, in this analysis, the polymorphisms detected should represent the major deviations between the expressed sequences of Minghui 63 and those of most of the *japonica* varieties, as well as between those of Minghui 63 and most of the other *indica* varieties.

In total, 2625 Minghui 63 sequences were used in both inter- and intra-subspecific comparisons of the mEST2EST analysis. Analysis of the DIS, using the same Minghui 63

Table 3 Polymorphisms of the expressed sequences revealed by inter- and intra-subspecific comparisons

	mEST2EST[a]		sEST2EST[b]		EST2GNM[c]	
	indica	*japonica*	*indica*	*japonica*	*indica*	*japonica*
No. of Minghui 63 sequences compared[d]	2681	4279	2586	4128	4386	4397
Total length of homologous regions (kb)	706	1668	954	1720	1831	1930
No. of DISs						
SNP	814	2197	3136	2897	4291	5659
Deletion	53	143	518	265	319	245
Insertion	85	232	648	793	809	833
Total	952	2572	4302	3955	5419	6737
Frequency of DISs (bp per DIS)						
SNP	868	759	304	594	427	341
Deletion	13,328	11,655	1842	6491	5740	7878
Insertion	8310	7184	1472	2169	2263	2317
Total	742	648	222	435	338	286

DIS, deletion/insertion/substitution; EST, expressed sequence tag; SNP, single nucleotide polymorphism.

[a] Each Minghui 63 sequence was compared with multiple rice ESTs/full-length cDNAs from GenBank. A DIS was recorded only when the Minghui 63 sequence detected a polymorphism with all the aligned homologous sequences at the same site.

[b] Each Minghui 63 sequence was compared with the most similar rice EST/full-length cDNA from GenBank.

[c] Each Minghui 63 sequence was compared with the most similar genomic sequence of either *indica* variety 93-11 or *japonica* variety Nipponbare.

[d] The consensus sequences from regions with overlapping ESTs.

sequences as in the inter- and intra-subspecific comparisons, revealed that only approximately 170 of the 1519 SNPs, seven of the 96 deletions and 33 of the 156 insertions detected in the inter-subspecific comparisons occurred in the intra-subspecific comparisons. Thus, approximately 88% (1561 of 1771) of the total DIS detected between Minghui 63 and multiple *japonica* varieties were also present between the *indica* varieties and the *japonica* varieties. These results suggest that the DIS detected in the inter-subspecific comparisons (Table 3) may largely represent the basal variation between the expressed sequences of the two subspecies.

In the second analysis (sEST2EST), each consensus sequence of Minghui 63 was compared with the most similar rice EST/full-length cDNA from GenBank (i.e., the sequence with the smallest *E*-value). Approximately 79% of the *japonica* ESTs/full-length cDNAs were from the variety Nipponbare in the pairwise inter-subspecific comparisons and 59% of the *indica* ESTs were from the varieties IR36 and IR64 in the pairwise intra-subspecific comparisons. Interestingly, DISs were detected more frequently in the intra-subspecific comparisons than in the inter-subspecific comparisons. The 1 and 2–4-bp Indels occurred in 50.2% (531) and 45.0% (476) of the inter-subspecific comparisons, respectively, whereas the respective frequencies were 74.3% (866) and 23.1% (269) in the intra-subspecific comparisons. The longest Indels detected in inter- and intra-subspecific comparisons were 12 and 15 bp, respectively.

To examine whether the DISs detected in the sEST2EST analysis might be partly due to sequence errors, as a large number of ESTs in GenBank are single-pass sequences, a third analysis, EST2GNM, was performed to compare the Minghui 63 sequences with the genomic sequences of *indica* variety 93-11 and *japonica* variety Nipponbare as the reference genomes. These genomic sequences were obtained by sequencing the genomes 4–10 times (Yu *et al*., 2002a; http://rgp.dna.affrc.go.jp/genomicdata/seqstrategy/newstrategy.html). The total frequency of DISs in the inter-subspecific comparisons of the EST2GNM analysis was even higher than that detected in the sEST2EST analysis, suggesting that EST sequence error may not be the major concern in previous comparisons. The total frequency of DISs in the intra-subspecific comparisons of the EST2GNM analysis was lower than that in the sEST2EST analysis. However, this difference could have been due to the fact that sequences from a different *indica* variety, 93-11, were used for the comparison. The 1 and 2–4-bp Indels occurred in 45.5% (490) and 50.6% (545) of the inter-subspecific comparisons, respectively, whereas the respective frequencies were 48.5% (547) and 46.7% (527) in the intra-subspecific comparisons. The longest Indels detected in the inter- and intra-subspecific comparisons were both 10 bp. As in the sEST2EST analysis, Indels were more frequently detected in the intra-subspecific comparisons than in the inter-subspecific comparisons in the EST2GNM analysis, although the later comparisons identified a higher total frequency of DISs (Table 3).

Comparison of 2.3 Mb of genomic sequences of *indica* GLA4 with 2.4 Mb of genomic sequences of *japonica* Nipponbare identified SNPs in the predicted exon regions at frequencies of 1 SNP per 379 bp for GLA4 and 1 SNP per 304 bp for Nipponbare (Han and Xue, 2003). Three SNPs and 0.22 Indel per 1000 bp were observed in the predicted coding regions, when Nipponbare and *indica* 93-11 were compared (Yu *et al*., 2005). The results of the present study show that the basal polymorphisms in expressed sequences were approximately 1 SNP per 759 bp and 1 Indel per 4448 bp, between Minghui 63 and *japonica* varieties, and 1 SNP per 868 bp and 1 Indel per 5116 bp, between Minghui 63 and other *indica* varieties. However, approximately 1 SNP per 341 bp and 1 Indel per 1790 bp were found between the expressed sequences of Minghui 63 and the corresponding genomic sequences of *japonica* variety Nipponbare, and 1 SNP per 427 bp and 1 Indel per 1623 bp were found between the expressed sequences of Minghui 63 and the corresponding genomic sequences of *indica* variety 93-11. The different frequencies of DISs detected in different studies further suggest that the extent of DIS-based polymorphisms is highly variable among different rice varieties.

Mapping of the ESTs to the rice genome

In total, 15 726 (88.2%) of the 17 835 unisequences, including 697 of the 2663 novel unisequences, were anchored in the rice genome according to their sequence homology with the genomic sequences of *japonica* variety Nipponbare. The 15 726 EST sites were distributed throughout the 12 chromosomes (Table 4). This high-density transcript map can be obtained from the REDB (http://redb.ricefgchina.org or http://bioinformatics.hzau.edu.cn) or Table S1. According to the latest rice RFLP map (http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html), the rice genome spans a total of 1530.4 cM. The present EST map covers 1526.6 cM – that is, 99.8% of the rice genome – and has a resolution of 10 EST sites per cM. This high density plus the annotation of the ESTs will greatly facilitate gene identification and isolation and the comparative study of gene evolution between subspecies. The average density of EST sites differed among the chromosomes (Table 4). Chromosome 3 had the highest density of EST sites, and chromosomes 8, 11 and 12 had the lowest density. Rice chromosomes 1, 2 and 3 have been reported to contain approximately 41% of a total of 6591 EST sites (Wu *et al*., 2002). The present results show that chromosomes 2, 3 and 5 have large numbers of EST sites, collectively accounting for 35% of the total 15 726 EST sites (Table 4). The total map length of the three chromosomes is 446.6 cM, which is 29% of the rice genome (http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html). Thus, chromosomes 2, 3 and 5

Table 4 Distribution of the unisequences in the rice genome[a]

Chromosome	No. of sites	Percentage	Density (site per cM)
1	2038	13.0	11
2	1950	12.4	12
3	2173	13.8	13
4	1227	7.8	9
5	1409	9.0	12
6	1206	7.7	10
7	1225	7.8	10
8	1029	6.5	8
9	870	5.5	9
10	858	5.4	10
11	898	5.7	8
12	843	5.4	8
Total	15 726	100	10

[a]Detailed information can be found in the Rice EST DataBase (http://redb.ricefgchina.org or http://bioinformatics.hzau.edu.cn) or in Table S1.

have a higher gene density than the other nine rice chromosomes.

Approximately 2109 (11.8%) of the 17 835 unisequences could not be mapped to the molecular linkage map by sequence homology analysis against rice genomic sequences. Among them, 106 sequences (including 25 novel ones) showed homology with *indica* genomic sequences with unknown chromosomal locations released by the Beijing Genomics Institute (http://btn.genomics.org.cn/rice/). The remaining 2003 sequences (including 1941 novel ones) showed no sequence homology with either *indica* or *japonica* genomic sequences. This non-match may be due to the incomplete sequencing of the rice genome or to sequence divergence leading to a poor match between Minghui 63 and Nipponbare, as has been previously reported for the *indica* and *japonica* subspecies (Feng *et al*., 2002; Sasaki *et al*., 2002; The Rice Chromosome 10 Sequencing Consortium, 2003; Yu *et al*., 2002a). Some of the ESTs are also likely to be specific to *indica* rice. For example, the novel sequence RECm4091, which was not found to be homologous to any of the Nipponbare sequences, showed sequence homology (99% identity) with the genomic sequence of *indica* variety 93-11 (accession no. AAAA02003735) at the position of 16 248–16 612 bp, whereas a fragment of AAAA02003735 at the position of 8148–10 314 bp was highly homologous (98% identity) to the Nipponbare genomic sequence (accession no. AP003410) located in an ungapped segment of the genomic sequence on chromosome 1. At least 70 of the Minghui 63 unisequences (including 15 novel sequences) were identified as *indica* specific by comparison with the available sequences (Table S2). Approximately 2–3% of the predicted genes have been found to be subspecies specific, on the basis of comparison with draft genomic sequences of *indica* and *japonica* varieties (Yu *et al*., 2005). More *indica*-specific ESTs are highly likely to be identified among the 2003 sequences that currently are not matched with genomic sequences in the databases, when more genomic sequences of *indica* rice become available for comparison. It may also be possible that a small portion of the unmapped sequences might have resulted from contamination by non-rice sources.

A novel type of transcriptionally active mutator-like transposase

One EST, EI045F13, detected 335 homologous sites (*E*-value range from 2e-105 to 2.4e-06) distributed across all 12 rice chromosomes. The deduced product of EI045F13 showed sequence homology with a mutator-like transposase (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov; protein database accession no. NP_922866; *E*-value = 2e-13), as analyzed by use of the BLASTX method, indicating that rice cultivar Minghui 63 contains a transcriptionally active mutator-like transposon. Two mutator-like transposable elements, OsMu4-2 and OsMu10-1, have been cloned from a *japonica* cultivar, and one of these elements is transcriptionally active (Asakura *et al*., 2002). However, EI045F13 showed no sequence similarity with either of the two transposable elements, when either nucleotide or putative encoding products were compared. EI045F13 also did not show sequence similarity with RMu1-A23, a rice genomic sequence (GenBank accession no. AB023047) with 99% identity to a rice cDNA sequence (GenBank accession no. C98506; *E*-value = 0), that is most likely a mutator-like element (Lisch *et al*., 2001). These results suggest that EI045F13 may represent a new type of mutator-like transposon. This mutator-like element appeared to be expressed rarely in rice, as only one copy of the sequence had been detected among the 39 208 sequences generated from the sequencing of >40 000 cDNA clones. In addition, analysis of EI045F13 against the 285 758 rice ESTs and 28 469 full-length cDNAs in GenBank, by use of the BLASTN program, did not find any homologous sequences. The existence of a large number of EI045F13 homologous genomic sequences in the rice genome suggests that this element was once very active during the course of evolution of the rice genome.

Association of the ESTs with QTLs for agronomically important traits

Hybrids produced with Minghui 63 as the restorer line have been widely cultivated in China. One of the best examples of the hybrids is Shanyou 63, which was released in the early 1980s, occupied approximately 6.7 million hectares during its peak period, and is still the most widely cultivated hybrid in rice production in China. A large number of QTLs related to important traits of these hybrids has been identified by use of populations derived from one of the hybrids, Shanyou 63, which is a cross between Zhenshan 97 (*O. sativa* ssp.

indica) and Minghui 63. These QTLs can be classified into seven categories, including yield and its components (Cui *et al.*, 2002b, 2003; Hua *et al.*, 2002, 2003; Li *et al.*, 2000; Xing *et al.*, 2002; Yu *et al.*, 1997), grain quality (Tan *et al.*, 1999, 2000, 2001; Xing *et al.*, 2001a), disease resistance (Chen, 2001; Chen *et al.*, 2003; Han *et al.*, 2002), growth vigor (Cui *et al.*, 2002a,c), biochemical products (Cui *et al.*, 2002a,c; Tan *et al.*, 2001), development (Xing *et al.*, 2001c; Yu *et al.*, 2002b), and morpho-physiological traits (Cui *et al.*, 2002a,b,c, 2003; Xing *et al.*, 2001b; Yu *et al.*, 2002b). Despite the extensive studies, little is known about the genes underlying the QTLs because of the complexity of the quantitative traits.

For comparison, all the QTLs mentioned above were placed onto the same framework map (Figure S1), which was constructed by use of a segregation population including 241 recombinant inbred lines developed by single-seed descent from a cross between Zhenshan 97 and Minghui 63 (Xing *et al.*, 2002). The corresponding relationship between the QTLs and the ESTs from Minghui 63 was established by assigning the 697 novel unisequences onto the framework map for QTL mapping (Figure S1). In Figure S1, the locations of these novel ESTs on the high-density EST map (Table S1) were also marked. Thus, the relationship between the QTLs and other ESTs that were only mapped on the high-density EST map can be easily determined. Almost all the QTLs detected by use of the segregation populations developed from the cross between Zhenshan 97 and Minghui 63 were co-localized with one or more of the expressed sequences (Figure S1 and Table S1).

Recent studies have shown that the candidate gene approach may provide a way to illustrate the genes underlying the QTLs for important agronomic traits (Borevitz and Chory, 2004). The ESTs co-localized with QTLs can serve as primary candidates for gene discovery. As each QTL corresponds to a segment of the chromosome in which two or more ESTs were mapped (Figure S1 and Table S1), the evidence for the candidate(s) of a QTL may be further evaluated by means of other information such as the microarray data associated with various phenotypes and the flanking sequence collections of insertion mutants. Finally, the candidate of a QTL needs to be confirmed by a complementation study such as transformation. The approach described above has been successfully used in our laboratory to identify the genes of two QTLs for disease resistance in rice (S. Wang *et al.*, unpublished data).

Conclusions

The *indica* subspecies of cultivated rice is the most widely cultivated form of rice produced worldwide. In this paper we present data on the large-scale collection, annotation and mapping of ESTs from an *indica* cultivar. This information may be useful for annotation of the *indica* rice genome, identification of *indica*-specific genes and the comparative study of the evolution between *indica* and *japonica* subspecies. The inter- and intra-subspecific expressed sequence polymorphisms may be used to develop new molecular markers for the identification, mapping and cloning of rice genes. These DISs will also facilitate the detection of allelic polymorphisms associated with phenotypic variations among different rice varieties. The establishment of the relationship between the high-density transcript map and the QTLs controlling various agronomically important traits, including yield and its components, disease resistance, grain quality, biochemical products, development, morpho-physiology and growth vigor, will provide a convenient reference source for studying the genetic basis of the QTLs. This information, in combination with other data such as gene expression profiles or flanking sequence databases of rice-mutant libraries, will facilitate the discovery of genes underlying the QTLs and the understanding of the molecular basis of quantitatively regulated plant activities.

Experimental procedures

Materials

The clones from a normalized whole-life-cycle cDNA library were used for obtaining EST sequences. This library was constructed by use of 15 tissues collected from nine developmental stages of rice cultivar Minghui 63 (*O. sativa* ssp. *indica*) and was normalized by saturation hybridization with genomic DNA (Chu *et al.*, 2003). In brief, a cDNA library was first constructed for each of the 15 tissues, which included those from calluses, sprouts, etiolated whole plants at the three-leaf stage, whole green plants at the three-leaf stage, whole plants at the five-leaf stage, flag leaves, whole plants except roots at the panicle development stage, whole plants except roots at the heading stage, panicles at the panicle development stage, panicles at the flowering stage, panicles at the grain-filling stage, stems at the flowering stage, pollens before flowering, and fungal pathogen (*M. grisea*)-inoculated leaves and bacterial pathogen (*Xanthomonas oryzae* pv. *oryzae*)-inoculated leaves. Denatured plasmids purified from the 15 cDNA libraries were mixed and hybridized with saturated genomic DNA from Minghui 63. Well-matched plasmids were then recovered from the hybridized genomic DNA to form the normalized cDNA library, referred to as a normalized whole-life-cycle cDNA library. A total of 62 000 clones were collected and stored. An analysis determined that the average insert length of this library was 1.4 kb (Chu *et al.*, 2003).

Analysis of cDNA sequences

The cDNA clones were sequenced from the 5′ ends by use of a T7 primer. First, the cDNA sequences were trimmed by comparison with sequences from *E. coli*, *X. axonopodis* pv. *citri*, *X. campestris* pv. *campestris* and vectors in GenBank (http://www.ncbi.nlm.nih.gov), and with the sequences of *M. grisea* provided by the *Magnaporthe* Sequencing Project (http://www.fungalgenomics.ncsu.edu). The BLASTN program (Altschul *et al.*, 1997) was used, with a stringent criterion for sequence alignment (Table 1), to remove contaminating sequences.

The processed sequences were obtained by clustering and assembling the ESTs using a modified ESTClustering program (Zhang *et al.*, 2003) in which the 'Phrap' program was replaced by the 'CAP3' program (Huang and Madan, 1999). The processed sequences were assembled with a stringent criterion for alignment (Table 1). To avoid the influence of poly(A/T) segments on the assembly results, all poly(A/T) segments of the ESTs were cut to 5 bp in length before assembly.

The 'getorf' program in the EMBOSS package (Rice *et al.*, 2000) was used to find the CDS of the processed sequences. The processed sequences with a CDS shorter than 90 bp or without a CDS were removed from the data. The gene numbers represented by the ESTs were determined by reclustering the processed sequences, using a stringent criterion for sequence alignment (Table 1), with reference to the 59 712 nucleotide sequences of the rice gene model containing the untranslated region but without the intron, which were downloaded from TIGR (http://www.tigr.org/tdb/e2k1/osa1/).

Analysis of SNPs and Indels

In the analysis of Indels, an insertion or deletion meant that the sequence(s) from GenBank had an insertion or deletion when compared with the sequence of rice cultivar Minghui 63 in the present study. To reduce the effects of sequencing errors, only the consensus sequences from overlapping regions of 4592 EST contigs of Minghui 63 were used for the analysis. The total length of the consensus sequences was 2077 kb.

Two strategies, EST2EST and EST2GNM, were used to analyze SNPs and Indels. Two methods, mEST2EST and sEST2EST, were used in the EST2EST strategy. In mEST2EST, rice ESTs and full-length cDNAs (The Rice Full-Length cDNA Consortium, 2003) collected from GenBank were divided into two local databases, which contained sequences from *indica* (i-dbEST) and *japonica* (j-dbEST) subspecies. The Minghui 63 sequences were used to screen for homologous sequences in i-dbEST and j-dbEST, by use of the BLASTN program (Altschul *et al.*, 1997) with a stringent criterion for sequence alignment (Table 1). If a Minghui 63 sequence identified two or more hits in the i-dbEST or j-dbEST, the Minghui 63 sequence and the homologs from the databases were aligned by use of the 'CAP3' program (Huang and Madan, 1999) with default settings for analysis of SNPs and Indels. In the sEST2EST analysis, the Minghui 63 sequence and its most homologous sequence (i.e., the sequence with the smallest *E*-value) from the i-dbEST or j-dbEST were first aligned by use of the BLASTN program with a stringent criterion (Table 1); the 'diffseq' program in the EMBOSS package (Rice *et al.*, 2000) was then used to screen the aligned sequences for SNPs and Indels.

In the EST2GNM strategy, genomic sequences of *japonica* rice (variety Nipponbare), collected from TIGR (http://www.tigr.org/tdb/e2k1/osa1/), and *indica* rice (variety 93-11), collected from the Beijing Genomics Institute (http://btn.genomics.org.cn/rice/), were divided into two local databases, j-dbGNM and i-dbGNM, respectively. The Minghui 63 sequence and its most homologous sequence (i.e., the sequence with the smallest *E*-value) from j-dbGNM or i-dbGNM were aligned by use of the BLASTN program with a stringent criterion (Table 1), and the SNPs and Indels were identified by use of the 'diffseq' program.

Mapping of ESTs

The ESTs were mapped to the rice chromosomes using the BLASTN program by searching for homologous genomic sequences with known chromosomal locations (Chen *et al.*, 2002) in the TIGR Rice Genome Database of Pseudomolecules (version 3; http://www.tigr.org/tdb/e2k1/osa1/) using a stringent criterion for sequence alignment (Table 1). When this stringent threshold was used, the two matched sequences had a sequence identity ranging from 80 to 100%, with an average of 91%, in this study. This level of sequence identity is much higher than that obtained by wet-lab DNA hybridization under conditions of low stringency, in which the hybridized DNA molecules share ≥65% sequence identity (Joseph and David, 2001). The most recent molecular linkage map (JRGP RFLP 2000) for rice, containing 3267 RFLP markers (http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html), was used as the framework map for mapping the ESTs.

Acknowledgements

This research was supported by grants from the National Special Key Program of Rice Functional Genomics, the National Program on the Development of Basic Research in China and the National Natural Science Foundation of China.

Supplementary Material

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ2408/TPJ2408sm.htm

Figure S1. Co-mapping of novel expressed sequence tags (ESTs) and quantitative trait loci (QTLs).

This map is also available online at http://redb.ricefgchina.org or http://bioinformatics.hzau.edu.cn.

Table S1 High-density map of expressed sequence tags (ESTs) of rice

This map is also available online at http://redb.ricefgchina.org (or) http://bioinformatics.hzau.edu.cn

Table S2 Unmapped, *indica* specific, novel expressed sequence tags (ESTs) showing sequence homology only with the genomic sequences of *indica* rice

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Asakura, N., Nakamura, C., Ishii, T. and Kasai, Y. (2002) A transcriptionally active maize MuDR-like transposable element in rice and its relatives. *Mol. Genet. Genomics*, **268**, 321–330.

Bennetzen, J.L., Coleman, C., Liu, R., Ma, J. and Ramakrishna, W. (2004) Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7**, 732–736.

Borevitz, J.O. and Chory, J. (2004) Genomics tolls for QTL analysis and gene discovery. *Curr. Opin. Plant Biol.* **7**, 132–136.

Chen, H. (2001) Population structure of *Pyricularia grisea* from central and southern China and comparative mapping of QTL for blast- and bacterial blight-resistance in rice and barley (in Chinese). PhD Dissertation. Huazhong Agriculture University, Wuhan, China.

Chen, M., Presting, G., Barbazuk, W.B. *et al.* (2002) An integrated physical and genetic map of the rice genome. *Plant Cell*, **14**, 537–545.

Chen, H., Wang, S., Xing, Y., Xu, C., Hayes, P.M. and Zhang, Q. (2003) Comparative analyses of genomic locations and race specificities of loci for quantitative resistance to *Pyricularia grisea* in rice and barley. *Proc. Natl Acad. Sci. USA*, **100**, 2544–2549.

Chu, Z., Peng, K., Zhang, L., Zhou, B., Wei, J. and Wang, S. (2003) Construction and characterization of a normalized whole-life-cycle cDNA library of rice. *Chin. Sci. Bull.* **48**, 229–235.

Cui, K.H., Peng, S.B., Xing, Y.Z., Xu, C.G., Yu, S.B. and Zhang, Q. (2002a) Molecular dissection of seedling-vigor and associated physiological traits in rice. *Theor. Appl. Genet.* **105**, 745–753.

Cui, K.H., Peng, S.B., Xing, Y.Z., Yu, S.B. and Xu, C.G. (2002b) Genetic analysis of the panicle traits related to yield sink size of rice. *Acta Genetica Sinica*, **29**, 144–152.

Cui, K.H., Peng, S.B., Xing, Y.Z., Yu, S.B. and Xu, C.G. (2002c) Molecular dissection of relationship between seedling characteristics and seed size in rice. *Acta Botanica Sinica*, **44**, 702–707.

Cui, K.H., Peng, S.B., Xing, Y.Z., Yu, S.B., Xu, C.G. and Zhang, Q. (2003) Molecular dissection of the genetic relationships of source, sink and transport tissue with yield traits in rice. *Theor. Appl. Genet.* **106**, 649–658.

Feng, Q., Zhang, Y., Hao, P. *et al.* (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433.

Goff, S.A., Ricke, D., Lan, T.H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.

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

Han, Y.P., Xing, Y.Z., Chen, Z.X., Gu, S.L., Pan, X.B., Chen, X.L. and Zhang, Q.F. (2002) Mapping QTLs for horizontal resistance to sheath blight in an elite rice restorer line, Minghui 63. *Acta Genetica Sinica*, **29**, 622–626.

Hua, J., Xing, Y., Xu, C., Sun, X., Yu, S. and Zhang, Q. (2002) Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. *Genetics*, **162**, 1885–1895.

Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S. and Zhang, Q. (2003) Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl Acad. Sci. USA*, **100**, 2574–2579.

Huang, X. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877.

Joseph, S. and David, W. (eds) (2001) Preparation and analysis of eukaryotic genomic DNA. In *Molecular Cloning: A Laboratory Manual*, Volume 1, 3rd edn. New York: Cold Spring Harbor Laboratory Press, pp. 6.1–6.64.

Li, J.X., Yu, S.B., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H. and Zhang, Q. (2000) Analyzing quantitative trait loci for yield using a vegetatively replicated F_2 population from a cross between the parents of an elite rice hybrid. *Theor. Appl. Genet.* **101**, 248–254.

Lisch, D.R., Freeling, M., Langham, R.J. and Choy, M.Y. (2001) Mutator transposase is widespread in the grasses. *Plant Physiol.* **125**, 1293–1303.

Mathe, C., Sagot, M.-F., Schiex, T. and Rouze, P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30**, 4103–4117.

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

Rogic, S., Francis Ouellette, B.F. and Mackworth, A.K. (2002) Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics*, **18**, 1034–1045.

Sasaki, T., Matsumoto, T., Yamamoto, K. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.

Tan, Y.F., Li, J.X., Yu, S.B., Xing, Z.Y., Xu, C.G. and Zhang, Q. (1999) The three important traits for cooking and eating quality of rice grains are controlled by a single locus in an elite rice hybrid, Shanyou 63. *Theor. Appl. Genet.* **99**, 642–648.

Tan, Y.F., Xing, Y.Z., Li, J.X., Yu, S.B., Xu, C.G. and Zhang, Q. (2000) Genetic bases of appearance quality of rice grains in Shanyou 63, an elite rice hybrid. *Theor. Appl. Genet.* **101**, 823–829.

Tan, Y.F., Sun, M., Xing, Y.Z., Hua, J.P., Sun, X.L., Zhang, Q. and Corke, H. (2001) Mapping quantitative trait loci for milling quality, protein content and color characteristics of rice using a recombinant inbred line population derived from an elite rice hybrid. *Theor. Appl. Genet.* **103**, 1037–1045.

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.

Wu, J., Maehara, T., Shimokawa, T. *et al.* (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell*, **14**, 525–535.

Xing, Y., Tan, Y., Xu, C., Hua, J. and Sun, X. (2001a) Mapping quantitative trait loci for grain appearance traits of rice using a recombinant inbred line population. *Acta Botanica Sinica*, **43**, 840–845.

Xing, Y., Xu, C., Hua, J. and Tan, Y. (2001b) Analysis of QTL × environment interaction for rice panicle characteristics. *Acta Genetica Sinica*, **28**, 439–446.

Xing, Y., Xu, C., Hua, J., Tan, Y. and Sun, X. (2001c) Mapping and isolation of quantitative trait loci controlling plant height and heading date in rice. *Acta Botanica Sinica*, **43**, 721–726.

Xing, Y., Tan, Y., Hua, J., Sun, X., Xu, C. and Zhang, Q. (2002) Characterization of the main effects, epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice. *Theor. Appl. Genet.* **105**, 248–259.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q. and Saghai Maroof, M.A. (1997) Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl Acad. Sci. USA*, **94**, 9226–9231.

Yu, J., Hu, S., Wang, J. *et al.* (2002a) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Li, X.H. and Zhang, Q. (2002b) Identification of quantitative trait loci and epistatic interactions for plant height and heading date in rice. *Theor. Appl. Genet.* **104**, 619–625.

Yu, J., Wang, J., Lin, W. *et al.* (2005) The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38.

Zhang, L.D., Yuan, D.J., Zhang, J.W., Wang, S. and Zhang, Q. (2003) A new method for EST clustering. *Acta Genetica Sinica*, **30**, 147–153.

GenBank accession numbers: CX099458–CX119141