Botanical Studies (2007) 48: 243-253.

Detection of SNPs between Tainung 67 and Nipponbare rice cultivars

Ai-Ling HOUR[1,a], Yao-Cheng LIN[1,a], Pei-Fang LI[1,2], Teh-Yuan CHOW[1,3], Wei-Fu LU[1,4], Fu-Jin WEI[1], and Yue-Ie C. HSING[1,*]

[1]*Institute of Plant and Microbial Biology, Academia Sinica, Taipei 11529, Taiwan*

(Received September 28, 2006; Accepted December 18, 2006)

ABSTRACT. Single nucleotide polymorphisms (SNPs) are known as the most detectable variations among related genomes. We estimated the SNPs between Tainung 67 (TNG67), an elite cultivar of rice (*Oryza sativa*) in Taiwan, and Nipponbare, the cultivar used for rice genome sequencing by the international consortium. More than 6,000 expressed sequence tag (EST) sequences from developing panicles of TNG67 were compared with the annotated gene sequences of Nipponbare. The estimated SNP rate is about 0.3% to 0.4% between the two cultivars, with most of the insertions or deletions (indels) occurring on the 5′ or 3′ untranslated regions (UTRs). The rate of transition substitutions on the 3′ UTR and the third codon positions is higher than that of transversions but lower on 5′ UTR and first codon positions. The synonymous (Ks) and non-synonymous (Ka) substitution distances are also calculated, and most of the Ka/Ks ratios are less than 1. Because the SNP density is higher than that of other traditional markers, detection of SNPs in this report with subsequent development of markers will allow genetic mapping and positional cloning between TNG67 and Nipponbare.

Keywords: Expressed sequence tag **(**EST); Nipponbare; RAP-DB; Rice; Single-nucleotide polymorphisms (SNP); TNG67.

INTRODUCTION

Rice is the most important staple food for half the world's population. The increase in global rice production in recent years is no longer keeping pace with the growth in consumption. Rice production in the next few decades will face even greater challenges with a larger and more affluent population, with greater demands for higher production and better-quality rice. However, future enhancement of global rice production faces the difficulties of reduced arable land, water, and labor while maintaining a sustainable agriculture system. Thus, great demands are put on biotechnology to improve rice production.

Better understanding of the rice genome will facilitate research on rice, which in turn speeds up the development of rice biotechnology methods. The highly accurate map-based genomic sequence of a japonica cultivar, Nipponbare, was decoded in 2005 by the International Rice Genome Sequencing Project (IRGSP) (IRGSP, 2005). Shortly before this sequence became available to the public, a whole genome shotgun sequence of the rice indica cultivar, 93-11, a parent of super hybrid rice, was released (Yu et al., 2002). Polymorphisms between the japonica and indica cultivars over the whole genomic region were analyzed with use of the above two genomic sequences (Feltus et al., 2004; Shen et al., 2004).

Single-nucleotide polymorphisms (SNPs) are the most abundant sequence variations among closely related genomes. They may be used as genetic markers because they are detectable on a large scale, and they exist in high density throughout the genome. Besides, they are generally more stable than microsatellite markers. Many studies have investigated the SNP distribution in human, mouse, *Arabidopsis*, maize, and other model organisms. For instance, previous research found a rate of one SNP per 242 or 348 sites from human expressed sequence tag (EST) data (Cargill et al., 1999; Halushka et al., 1999). When the International Human Genome Sequencing Consortium announced the available draft sequence of the human genome, the International SNP Map Working Group also reported that SNPs occur every 1,000-2,000 bases, on average, in a comparison of chromosomes from several human beings (Sachidanandam et al., 2001). With the

[a]Equal contributors.

[2]Present address: Department of Biotechnology, Fooyin University, Kaohsiung 831, Taiwan.

[3]Present address: Institute of Medical Biotechnology, Central Taiwan University of Science and Technology, Taichung 40601, Taiwan.

[4]Present address: Department of Biotechnology and Bioinformatics, Asia University, Taichung 41354, Taiwan.

*Corresponding author: E-mail: bohsing@gate.sinica.edu.tw; Tel: 886-2-7892496; Fax: 886-2-7827954.

progress of the International HapMap Project, more than 10 million SNPs have been characterized, many of them associated with diseases, although the allele frequencies are low (Kruglyak and Nickerson, 2001; Carlson et al., 2003; The International Hap Map Consortium, 2003; Thorisson and Stein, 2003).

The differences between japonica and indica rice have been assessed from their relatively distinct rice genomic sequences. About 1.7 to 3.7 SNPs and 0.1 indels per kb were found in a comparison of Nipponbare and 93-11 (Feltus et al., 2004; Shen et al., 2004). Using EST or PCR-sequencing methods, the SNP frequencies of Kasalath (an indica land race), Koshihikari (a japonica cultivar in Japan), W1943 (a wild accession of *Oryza rufipongon*), Guang-Lu-Ai 4 (an indica cultivar from China) and Senshou (an upland rice in Taiwan) were also analyzed for SNPs (Nasu et al., 2002; Monna et al., 2006). Recently, an Oryza SNP project initiated by the international rice community plans to compare 21 rice genomes with each other using a high throughput strategy (McNally et al., 2006). The discovery of genome-wide SNPs will provide a useful tool for selection in rice breeding and new genetic knowledge of evolution and domestication.

Tainung 67 (TNG67), an elite japonica cultivar in Taiwan, has been a leading commercial variety for several decades (Huang, 1979). It is photoperiod insensitive, and possesses many good agronomic characteristics. Currently, it is a popular genetic stock in breeding programs and scientific research in Taiwan. In the present work, we used EST sequences of TNG67 to explore the polymorphism between this cultivar adapted in Taiwan and another japonica cultivar (Nipponbare) adapted in Japan.

MATERIALS AND METHODS

Construction of cDNA libraries and the subsequent processing

The 0.5 mm and 6-8 cm developing panicles of rice (*Oryza sativa* L. cv Tainung No. 67) were used to construct a 5MPR and 6CPR cDNA library, respectively. Tissues were harvested, immediately frozen in liquid nitrogen, and stored at -80°C. Total RNA was extracted for cDNA library construction using Lambda ZAPII (Stratagene). About 4,000 clones from each library were randomly selected for sequencing. Raw sequencing data were evaluated by the base calling program PHRED (Ewing and Green, 1998; Ewing et al., 1998). Vector-derived sequences were screened by CROSS_MATCH (P. Green, http://www.phrap.org/phredphrapconsed.html) according the vector database. The EST reads were quality trimmed by the PHRED quality score where Phred score < 20 within 10 consecutive bases. Reads that comprised over 100 bp of contiguous satisfying quality were kept as successful sequence for later analysis and were submitted to the GenBank dbEST database with the accession numbers EG709278 - EG713055 and EG713056-EG715450 for 5MPR and 6CPR, respectively.

Assembly of EST contigs and data download

Since ESTs generally represent only partial cDNA sequences, the assembly of overlapping ESTs into putative unique transcript contigs constitutes the first step for all EST analyses. Two cDNA libraries were constructed from developing panicles of 5 mm and 6 cm long, which resulted in 3,787 and 2,354 EST sequences or a total length of 1,473,449 and 871,401 bps for the two libraries, respectively. The 6,112 sequences in these two EST libraries were clustered by the TGICL program (Pertea et al., 2003), which resulted in 659 assembled contigs and 3834 singletons (Figure 1).

The japonica variety Nipponbare has been sequenced completely (IRGSP, 2005). The most updated annotations of cDNA and genomic sequences from The Rice Annotation Project Database (RAP-DB) (Ohyanagi et al., 2006) were downloaded for similarity searches.

The EST sequences of other rice varieties reported previously (Feltus et al., 2004; Katagiri et al., 2004; Shen et al., 2004; Monna et al., 2006) were also downloaded from GeneBank for comparison study.

Polymorphism level estimation

Similarity searches were performed with BLASTN program (Altschul et al., 1997). Both of the assembled contigs and singletons described above were aligned against RAP-DB, with a cut-off E-value of 1E-50. The resulting high-scoring segment pairs (HSPs) with a minimum similarity of 95% and minimum coverage of 50% were then screened (Table 1).

The divergence levels based on the HSPs between TNG67 and Nipponbare were estimated on ESTs, contigs, and singletons. The polymorphisms were divided into substitutions and indels. The polymorphisms on the 5' and 3' UTRs and three codon sites were calculated separately.

Synonymous and non-synonymous substitutions ratio

The synonymous (Ka) and non-synonymous (Ks) substitution distances between contig sequence pairs were estimated simultaneously by use of nucleotide and amino acid sequences. The estimation was implemented by use of the Diverge program of the GCG package (Wisconsin Package). The methods and parameters used were Li's model (Li, 1993) and Kimura's two-parameter substitution model (Kimura, 1980). The Ka/Ks ratios were calculated for sequence pairs for which both Ka and Ks values were larger than zero.

RESULTS

We estimated the polymorphism proportions between TNG67 and Nipponbare cultivars by using panicle ESTs and annotated cDNA sequences. The distributions of SNPs or indels on different gene coding regions were also assessed. Furthermore, the Ka/Ks ratios were calculated.

Table 1. Summary of aligned and screened polymorphisms from the BLAST results of TNG67 ESTs, contigs and singletons against the RAP1 database. All data are percentages. Both the original and screened data were used.

	Aligned HSPs	Screened HSPs*
EST		
Sequence numbers	87.07	95.25
Length	80.56	91.29
Mismatch	0.42	0.31
InDel	0.10	0.10
Polymorphism	0.53	0.41
Contig		
Sequence numbers	94.69	96.15
Length	85.54	90.40
Mismatch	0.31	0.22
InDel	0.10	0.10
Polymorphism	0.41	0.31
Singleton		
Sequence numbers	82.37	94.14
Length	75.69	90.50
Mismatch	0.48	0.34
InDel	0.11	0.10
Polymorphism	0.58	0.44

*The criterion for screening was at least 50% length alignment and greater than 95% identity.

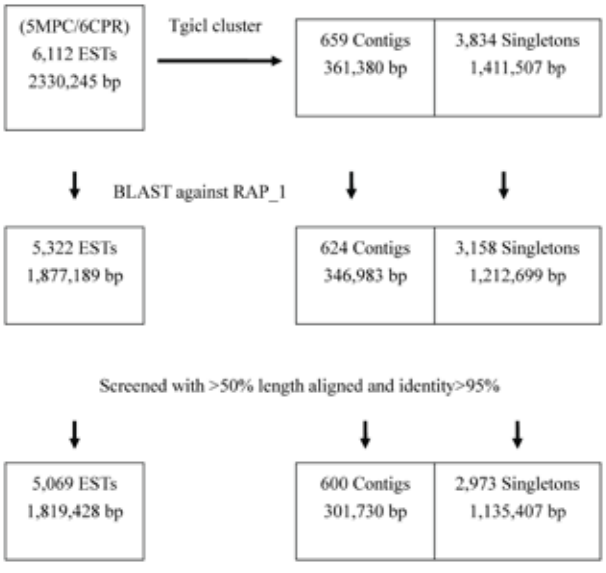

Figure 1. Flow chart of the clustering of ESTs from panicles of TNG67 and detection of SNPs. Total sequence numbers and lengths are indicated.

Length and hit frequencies of EST sequences

Redundant ESTs were assembled into putative unique transcript contigs before further analysis. These contig sequences are longer than single-pass EST sequences, which allows for construction of nearly full-length cDNAs. As calculated from data shown in Figure 1, the mean length of 6,112 EST sequences was 381 bps before clustering. After clustering, the mean length for the 659 clustered contigs and 3,834 singletons were 548 and 348 bps, respectively. Both the EST and singleton sequences showed a platykurtic distribution and were left skewed while the clustered contigs were leptokurtic distributed and right skewed.

Table 1 lists the percentages of sequences and lengths of TNG67 EST sequences aligned with RAP-DB sequences by use of the BLASTN program. Both the original and the screened HSPs were used. The homologs of assembled contigs found a match for 95% of the sequences in RAP-DB, with about 85% of lengths aligned. This ratio was much higher than with the other two kinds of unassembled sequences, ESTs or singletons, and indicates increased sequence quality after clustering.

The HSPs with more than 50% length alignment and higher than 95% identity were screened. Table 1 shows the similar proportions of HSPs screened from ESTs, assembled contigs and other singletons. Only the screened contig sequences were used for further analysis.

Polymorphisms were about 0.3% to 0.5%

With the HSPs received from BLASTN analysis between TNG67 ESTs and RAP-DB sequences, the proportion of polymorphisms, including substitutions and gaps for ESTs, contigs and singletons, were calculated separately. The proportion of total polymorphisms was about 0.4% to 0.6% between aligned HSPs. The values were reduced to 0.3% to 0.4% with screened alignments (Table 1). The indel proportion was about 0.1% with or without screening. However, the mismatched proportions were reduced from 0.4% to 0.3% after screening. The polymorphisms observed from the assembled contigs, including 0.3% mismatch and 0.1% indels, with different lengths from 1 bp to 11 bps (Table 2), were investigated in

Table 2. Frequency distribution of continuous SNPs and indels detected from assembled contigs of TNG67.

Length	SNP	Indel
1	523	181
2	44	20
3	2	7
4	1	3
5	0	4
6	0	1
6~10	0	6
Total	570	222

detail to uncover the distribution on different positions and types of substitution.

Distribution of indels and gaps in different regions

The indels and transition/transversion substitutions on different codon positions or 5′ and 3′ UTRs were calculated for the screened contig sequences. Table 3 and Figure 2 illustrate that the highest polymorphism portion was found on the 3′ UTR (0.53%), with decreasing polymorphisms on the 5′ UTR (0.41%) and the third codon (0.30%). Of the three codon positions, the indels occurred similarly, but the substitutions were distributed differently; the transition proportions on the third codon were double those of the other two codons. The indel and substitution proportions of the first codon were slightly larger than those of the second codon with the indel proportion in the first codon being the highest among the three positions. Only on the 5′ UTR was the proportion of indels higher than that of substitutions and exactly the reverse on the 3′ UTR and coding regions.

Most Ka/Ks ratios were less than 1

The nucleotide substitutions on protein coding sites are defined as synonymous or non-synonymous substitutions according to whether the corresponding amino acids were changed. The ratio of Ka (number of non-synonymous changes per nonsynonymous site) and Ks (number of synonymous changes per synonymous site) reflects the selective constraint on a gene.

The synonymous and non-synonymous substitution distances were calculated with use of the selected homologous sequence pairs. Of the 154 selected contigs, 133 had Ka/Ks ratios less than 1 (Figures 3 and 4), which indicates that the selection constraint of these genes was conserved. The annotations of the genes with Ka larger than Ks are listed in Table 4. Interestingly, most of the Ka and Ks values of these genes are quite low, which indicates a high similarity between homologous pairs.

Table 3. Total length (base pairs) of TNG67 panicle EST contigs and their polymorphism between homologous pairs on comparison with Nipponbare genome annotation genes.

	Total length	Indel	Transition	Transversion
5′ UTR	22,488	48	19	25
3′ UTR	79,951	155	144	121
Codon	199,028	124	163	147
1st	66,318	48	32	46
2nd	66,338	37	38	37
3rd	66,372	39	93	64

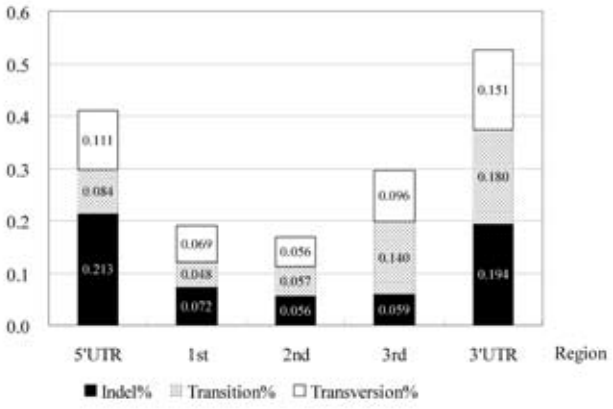

Figure 2. Polymorphism frequencies between TNG67 and Nipponbare sequences. Three types of differences, including indels, transitions and transversions, on different genomic regions are indicated.

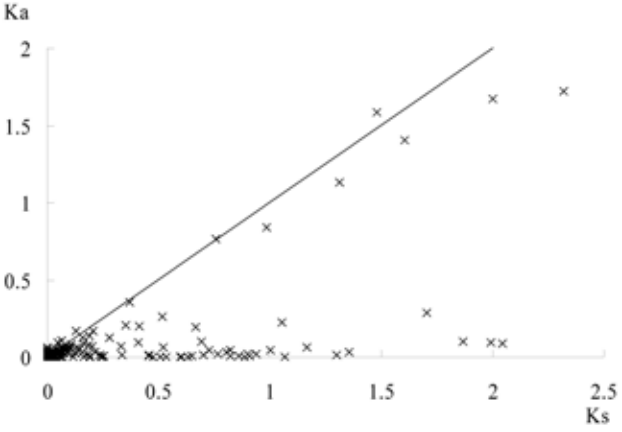

Figure 3. Relation between Ka and Ks from 154 homologous matches between TNG67 and Nipponbare sequences. The solid diagonal line represents Ka=Ks. Most of the observed points are Ka<Ks.

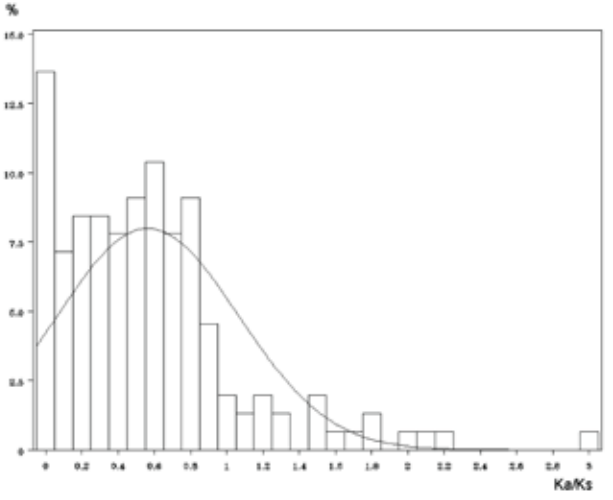

Figure 4. Frequency distribution of Ka/Ks ratio for 154 homologous pairs between TNG67 and Nipponbare sequences. The fitted normal distribution is shown as a curved line.

Table 4. Genes with Ka/Ks >1, annotated by homologous RAP-DB sequences.

EST Contig #	Locus name	Sequence name	Ks	Ka	Annotation descriptions
CL8Contig1	AK100159	-	1.481	1.584	Glyceraldehyde-3-phosphate dehydrogenase (Gpc) protein
CL386Contig1	AK103374	Os07g0466300	0.758	0.762	Rurm1 protein
CL340Contig1	AK102983	Os09g0484300	0.127	0.166	HECT domain containing protein
CL641Contig1	Os03g0750300	Os03g0750300	0.064	0.103	Hypothetical protein
CL574Contig1	AY224520	Os06g0643300	0.046	0.092	Superal1
CL639Contig1	AK103440	Os07g0191000	0.039	0.051	Inositol monophosphatase family protein
CL340Contig1	AK102983	Os09g0484300	0.027	0.032	HECT domain containing protein
CL194Contig1	AK065975	Os05g0498700	0.029	0.03	Gda-1 protein
CL487Contig1	AK074018	Os05g0113900	0.025	0.029	Histone H2A
CL315Contig1	AK103501	Os03g0192700	0.017	0.025	Myo-inositol-1-phosphate synthase
CL343Contig1	AK061681	Os05g0553000	0.011	0.017	ATP synthase beta chain, mitochondrial precursor (EC 3.6.3.14)
CL484Contig1	AB117994	Os09g0326900	0.008	0.017	Translation initiation factor IF5 domain containing protein
CL507Contig1	AK059049	Os08g0242700	0.009	0.015	Hypothetical protein.
CL512Contig1	AK120568	Os03g0836500	0.008	0.014	Conserved hypothetical protein
CL391Contig1	AY332470	Os03g0794500	0.021	0.012	Glutamate dehydrogenase (EC 1.4.1.3) (GDH)
CL391Contig1	AY332470	Os03g0794500	0.011	0.012	Glutamate dehydrogenase (EC 1.4.1.3) (GDH)
CL410Contig1	Os06g0723900	Os06g0723900	0.009	0.011	(not hit)
CL288Contig1	AK106095	Os01g0502400	0.005	0.011	2OG-Fe(II) oxygenase domain containing protein
CL255Contig1	AK061254	Os02g0317400	0.005	0.009	Clathrin adaptor complex, small chain family protein
CL55Contig1	X81691	Os11g0220800	0.004	0.006	60S ribosomal protein L10 (QM protein homolog)
CL290Contig1	AK065538	Os01g0703600	0.002	0.006	Mu1 adaptin

DISCUSSION

The detection of SNPs between two rice cultivars may be approached in several ways. Because the whole genome sequence of Nipponbare and the EST sequences of TNG67 are available, *in silico* SNP detection by bioinformatics is a relatively low-cost method.

Using the EST data generated in our laboratory, we searched more than 6,000 ESTs from panicles and detected a 0.3% polymorphism between TNG67 and Nipponbare. The SNPs or indels were distributed over UTRs and coding regions. The resulting data gave useful information for future research involving the use of TNG67 as an experiment material and Nipponbare genomic sequence as a database resource.

Frequency of polymorphisms in TNG67 panicle ESTs

Agronomic traits and physiological characteristics differ genetically among rice subspecies or cultivars, and molecular polymorphisms have been reported recently (McNally et al., 2006). This integrated knowledge enables analysis of traits undergoing selection in the course of domestication. In the present work, we compared TNG67 and Nipponbare sequences and found about three SNPs per kb, or 0.3%. This rate is intermediate of two previously published rates for Nipponbare compared with 93-11, a japonica and an indica variety: 0.17% from Feltus et al. (2004) and 0.71% from Shen et al. (2004). However, the rate is higher than that found among japonica cultivars by PCR analysis (Nasu et al., 2002; Monna et al., 2006). We downloaded EST sequences of other rice cultivars for analysis as described in Materials and Methods, and the results are listed in Supplementary Table 1; previous results of different approaches are listed in Supplementary Table 2. The estimation of SNP frequency may differ because of the calculation criteria of each method or the biased bases of ratios. For example, Monna et al. (2006) selected only 490 amplicons from 1,117 sites for which effective data were obtained in all cultivars used. The frequency of SNPs detected among japonica cultivars by PCR was 0.03% to 0.05%, much lower than our result by EST sequencing and *in silico* estimation.

Information on the pedigree of TNG67 varieties was retrieved from the Taiwan Rice Information System constructed by the Taiwan Agricultural Research Institute (http://tris.tari.gov.tw:8080) and redrawn as Figure 5. Since this variety includes many indica varieties in pedigree, the TNG67 genome may diverge from that of traditional japonica cultivars from Japan such as Nipponbare. Breeders' selection preferences may also affect the genetic distances between cultivars. The rice breeding program in Taiwan usually involves TNG67 in the pedigree, but the commonly available database and information are constructed with Nipponbare. Therefore, SNP analysis is essential for the study of genetic mapping and phenotype-genotype association.

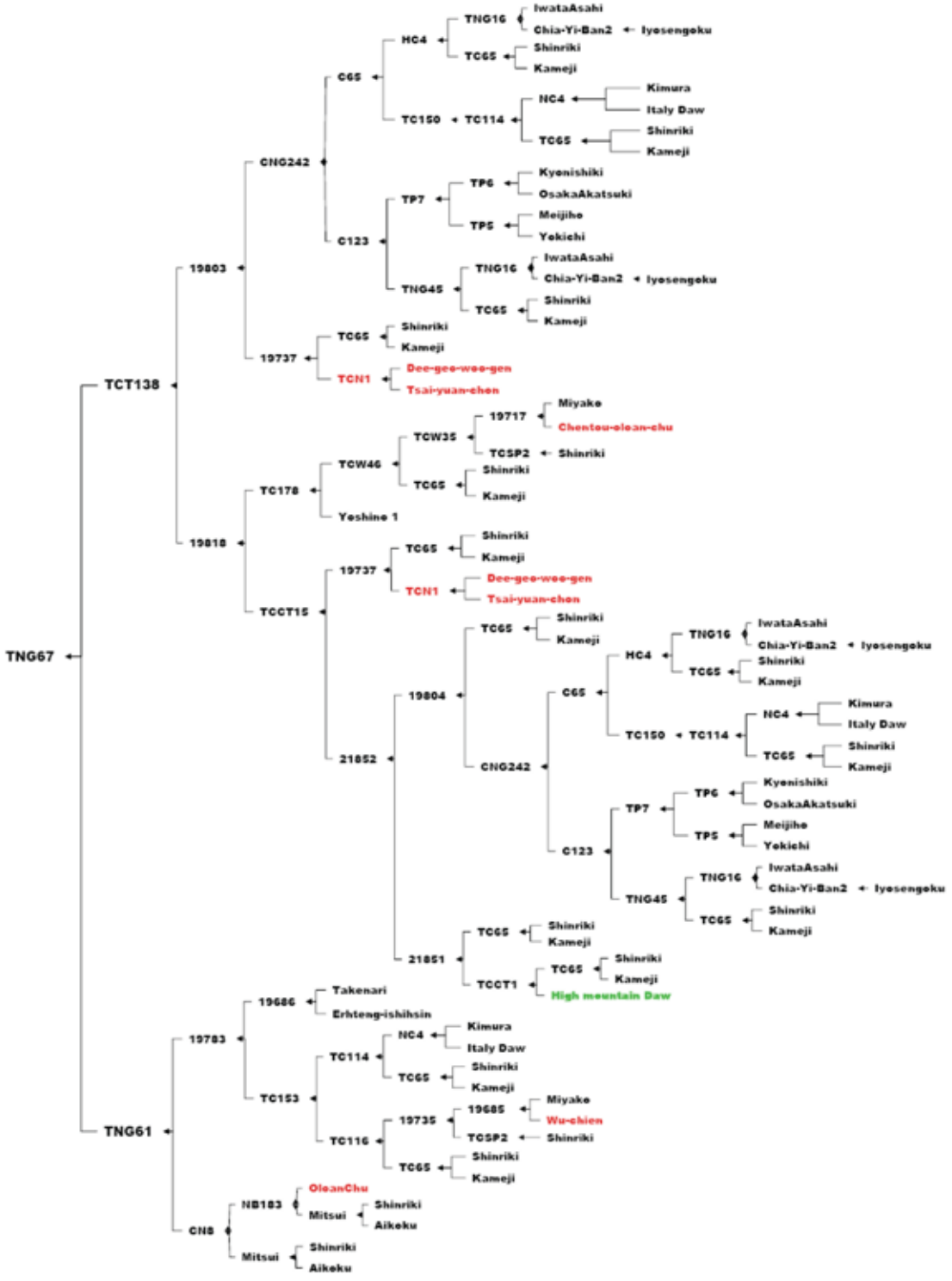

Figure 5. The pedigree of TNG67. Black indicates japonica varieties, green indica varieties, and red javanica varieties.

For example, TNG67 was used as research material in serial analysis of gene expression (SAGE) (Su et al., 2005), mutant generation involving sodium azide (Jeng et al. 2003), T-DNA insertion mutants (Hsing et al., 2007), and AFFYMETRIX array analysis. If the annotation of flanking sequences or expression genes should be based on the genomic sequences and cDNA databases of Nipponbare, the polymorphism between TNG67 and Nipponbare must be considered. Thus, for SAGE, the probability of sequence polymorphisms is approximately 1% for the 4-bp restriction sites and approximately 3% for the 10-bp tag sequences. An annotation system allowing mismatches in such scales should be implemented.

Confirmation of homologous alignment by screening

The sequencing of ESTs was not confirmed by duplicated processes; thus, errors in sequences cannot be overlooked (Pertea et al., 2003). The sequence of panicle ESTs have been estimated by autoSNP (Barker et al., 2003) with the assembled contigs. Only 22 of 659 contigs have SNPs, and most of them occurred on GC-rich or simple-repeated regions. Since we are using consensus sequences of contigs, the sequencing errors are very low and should be neglected. The clustering of EST sequences to longer and higher coverage contigs reduce the error mentioned above. Genome sequencing and comparative genomic studies have documented gene duplications in the rice genome (Goff et al., 2002; Yu et al., 2002; IRGSP, 2005; Yu et al., 2005). In the current study, the screening of HSPs with thresholds of 95% similarity and 50% coverage was a strategy intended to reduce the possibility of non-orthologous alignments.

From the results illustrated in Table 1, the total polymorphism was reduced from 0.41% to 0.31% after clustering, which implies that sequencing error may account for approximately 0.1% of polymorphism. Similar results can be estimated from other cultivars (Supplementary Table 1). With the described screening criteria, confirmation of homologous alignments can be processed automatically and objectively.

The coverage rates of alignments were reduced by 10% when we screened the homologous alignments with similarity higher than 95% and when more than 50% of the length of sequences was aligned. The criteria of screening are empirical and usually result in robust assessments. Thus, in subsequent analyses, including SNP distribution in different coding positions and Ka/Ks ratio, we used the screened contigs as query sequences for correctness.

Distribution of polymorphism

The polymorphisms observed in the current study did not just occur on one base pair; instead, about 10% of polymorphisms occurred on more than 1-bp lengths (Table 2).Under the cut-off E-value of 1E-50 and screening criteria, the total probability of indels was lower than that of SNPs (Figure 2), but the probability of indels longer than 1 bp was higher than that of SNPs. The mean total polymorphism frequency was 0.5% for the 3' UTR region, 0.4% for the 5' UTR region, and 0.3% for the three codons of the coding regions. The average polymorphism on the third codon was higher than on the other two, which is consistent with theoretical knowledge of selection constraint (Nei, 2005). The indels occurred more frequently in UTR regions than in coding regions because of possible serious alternation consequences in the latter. The frequency of transition and transversion substitutions was almost the same in the first and second codons but differed in the third codon.

Implication from evolutionary distances

The calculation of Ks and Ka substitution distances gives aspects of functional constraint for corresponding genes. In the current study, the divergence of most aligned genes was low and Ka/Ks ratios were usually below 1; thus, the gene function is conserved between these two cultivars with high polymorphism. Table 4 lists 21 genes with Ka>Ks. Most of them have small absolute values of Ka and Ks because the sequences' coding regions are consistent even under unconstrained function. A possible cause was the bias introduced by low Ks (Xing and Lee, 2006).

Conclusion

SNPs, the basic units of genomic diversity, have high density even among closely related genomes, which makes them preferred markers for fine mapping and understanding evolutionary dynamics. These SNP sites can then be used in designing primers and marker-assisted selection. SNPs detected from panicle EST libraries may be especially helpful in investigating the development and function of panicles for the elite local cultivar Tainung 67.

Acknowledgments. This work was supported by the Thematic Project, Academia Sinica, Taipei, Taiwan. The authors would like to thank Anthony H.C. Huang for his critical review of this manuscript, Dr. Ming-Hsing Lai for his helpful comments and discussion, and Ms. Laura Heraty for English editing.

LITERATURE CITED

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

Barker, G., J. Batley, H. O'Sullivan, K.J. Edwards, and D. Edwards. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics **19:** 421-422.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington,

R. Lipshutz, G.Q. Daley, and E.S. Lander. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231-238.

Carlson, C.S., M.A. Eberle, M.J. Rieder, J.D. Smith, L. Kruglyak, and D.A. Nickerson. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat. Genet. **33:** 518-521.

Feltus, F.A., J. Wan, S.R. Schulze, J.C. Estill, N. Jiang, and A.H. Paterson. 2004. An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res. **14:** 1812-1819.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175-185.

Ewing, E. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:** 186-194.

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

Halushka, M.K., J.B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239-247.

Hsing, Y.I., C. G. Chern, M.J. Fan, P.C. Lu, K.T. Chen, S.F. Lo, P.K. Sun, S.L. Ho, K.W. Lee, Y.C. Wang, R. Ko, W.L. Huang, J.L. Chen, C.I. Chung, Y.C. Lin, A.L. Hour, Y.W. Wang, Y.C. Chang, M.W. Tsai, Y.S. Lin, Y.C. Chen, S. Chen, H.M. Ten, C.P. Li, C.K. Wey, C.S. Tseng, M.H. Lai, S.C. Huang, L.J. Chen, and S.M. Yu. 2007. A rice gene activation/knockout mutant resource for functional genomics. Plant Mol. Biol. **63:** 351-364.

Huang, C.S. 1979. Development of rice variety Tainung 67. J. Agri. Res. China **28:** 57-66.

IRGSP. 2005. The map-based sequence of the rice genome. Nature **436:** 793-800.

Jeng, T.L., C.S. Wang, T.H. Tseng, C.L. Chen, and J.M. Sung. 2003. Starch biosynthesizing enzymes in developing grains of rice cultivar Tainung 67 and its sodium azide-induced rice mutant. Field Crops Res. **84:** 261-269.

Katagiri, S., J. Wu, Y. Ito, W. Karasawa, M. Shibata, H. Kanomori, Y. Katayose, N. Namiki, T. Matsumoto, and T. Sasaki. 2004. End sequencing and chromosomal *in silico* mapping of BAC clone derived from an *indica* rice cultivar, Kasalath. Breed. Sci. **54:** 273-279.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111-120.

Kruglyak, L. and D.A. Nickerson. 2001. Variation is the spice of life. Nat. Genet. **27:** 234-236.

Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96-99.

McNally, K.L., R. Bruskiewich, D. Mackill, C.R. Buell, J.E. Leach, and H. Leung. 2006. Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. Plant Physiol. **141:** 26-31.

Monna, L., R. Ohta, H. Masuda, A. Koike, and Y. Minobe. 2006. Genome-wide searching of single-nucleotide polymorphisms among eight distantly and closely related rice cultivars (*Oryza sativa* L.) and a wild accession (*Oryza rufipogon* Griff.). DNA Res. **13:** 43-51.

Nasu, S., J. Suzuki, R. Ohta, K. Hasegawa, R. Yui, N. Kitazawa, L. Monna, and Y. Minobe. 2002. Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNP markers. DNA Res. **9:** 163-171.

Nei, M. 2005. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. **22:** 2318-2342.

Ohyanagi, H., T. Tanaka, H. Sakai, Y. Shigemoto, K. Yamaguchi, T. Habara, Y. Fujii, B. A. Antonio, Y. Nagamura, T. Imanishi, K. Ikeo, T. Itoh, T. Gojobori, and T. Sasaki. 2006. The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. *japonica* genome information. Nucleic Acids Res. **34:** D741-744.

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

Sachidanandam, R., D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, D.L. Willey, S.E. Hunt, C.G. Cole, P.C. Coggill, C.M. Rice, Z. Ning, J. Rogers, D.R. Bentley, P.Y. Kwok, E.R. Mardis, R.T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R.H. Waterston, J.D. McPherson, B. Gilman, S. Schaffner, W.J. Van Etten, D. Reich, J. Higgins, M.J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M.C. Zody, L. Linton, E.S. Lander, and D. Altshuler. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928-933.

Shen, Y.J., H. Jiang, J.P. Jin, Z.B. Zhang, B. Xi, Y.Y. He, G. Wang, C. Wang, L. Qian, X. Li, Q.B. Yu, H.J. Liu, D.H. Chen, J.H. Gao, H. Huang, T.L. Shi, and Z.N. Yang. 2004. Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. **135:** 1198-1205.

Su, C.L., C.I. Chung, Y.C. Lin, P.C. Lu, F.J. Wei, Y.I.C. Hsing, and A.L. Hour. 2005. Statistics analysis of rice SAGE data. J. Genet. Mol. Biol. **16:** 248-260.

The International Hap Map Consortium. 2003. The International HapMap Project. Nature **426:** 789-796.

Thorisson, G.A. and L. D. Stein. 2003. The SNP Consortium website: past, present and future. Nucleic Acids Res. **31:** 124-127.

Wisconsin Package Version 10.3. Genetics Computer Group (GCG), Madison, WI.

Xing, Y. and C. Lee. 2006. Can RNA selection pressure distort the measurement of Ka/Ks? Gene **370:** 1-5.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, and H. Yang. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79-92.

Yu, J., J. Wang, W. Lin, S. Li, H. Li, J. Zhou, P. Ni, W. Dong, S. Hu, C. Zeng, J. Zhang, Y. Zhang, R. Li, Z. Xu, S. Li, X. Li, H. Zheng, L. Cong, L. Lin, J. Yin, J. Geng, G. Li, J. Shi, J. Liu, H. Lv, J. Li, J. Wang, Y. Deng, L. Ran, X. Shi, X. Wang, Q. Wu, C. Li, X. Ren, J. Wang, X. Wang, D. Li, D. Liu, X. Zhang, Z. Ji, W. Zhao, Y. Sun, Z. Zhang, J. Bao, Y. Han, L. Dong, J. Ji, P. Chen, S. Wu, J. Liu, Y. Xiao, D. Bu, J. Tan, L. Yang, C. Ye, J. Zhang, J. Xu, Y. Zhou, Y. Yu, B. Zhang, S. Zhuang, H. Wei, B. Liu, M. Lei, H. Yu, Y. Li, H. Xu, S. Wei, X. He, L. Fang, Z. Zhang, Y. Zhang, X. Huang, Z. Su, W. Tong, J. Li, Z. Tong, S. Li, J. Ye, L. Wang, L. Fang, T. Lei, C. Chen, H. Chen, Z. Xu, H. Li, H. Huang, F. Zhang, H. Xu, N. Li, C. Zhao, S. Li, L. Dong, Y. Huang, L. Li, Y. Xi, Q. Qi, W. Li, B. Zhang, W. Hu, et al. 2005. The genomes of *Oryza sativa*: a history of duplications. PLoS Biol. **3:** e38.

檢測水稻台農 67 與日本晴品種間之單核苷酸多型性

侯藹玲[1]　林耀正[1]　李佩芳[1,2]　周德源[1,3]　呂威甫[1,4]　魏甫錦[1]　邢禹依[1]

[1] 中央研究院植物暨微生物研究所

[2] 目前地址：輔英科技大學生物技術系

[3] 目前地址：中臺科技大學醫學生物科技研究所

[4] 目前地址：亞洲大學生物科技與生物資訊學系

單核苷酸多型性（Single Nucleotide Polymorphisms, SNPs）被視為基因體內數量最多的變異型態。我們採用台灣優良品種台農 67 幼穗發育期的 6 千筆以上基因表現標籤（Expressed Sequence Tag, EST）為材料，估算台農 67 與國際水稻基因體計畫所用品種日本晴（Nipponbare）之間的多型性。估計得到台農 67 與日本晴間的單核苷酸多型性頻度約為 0.3-0.4%，而大部分的核苷酸增減則發生在基因不轉譯的5端及 3 端。在鹼基置換部分，3 端不轉譯區域及第三密碼子位置的異類鹼基互換較同類者多，而 5 端不轉譯區域及第一密碼子位置的情況相反。同時也計算同類置換與非同意置換計算演化距離。由於單核苷酸多型性通常比一般傳統分子標記之密度較高，對於像台農 67 及日本晴這樣的近源品種，偵測其單核苷酸多型性並依此設計分子標記將有助於進一步對遺傳圖譜或基因定位研究探討。

關鍵詞：基因表現標籤；日本晴；水稻註解計畫資料庫；水稻；單核苷酸多型性；台農 67。

Supplementary Table 1. Polymorphism percentages, including mismatches and gaps, between the Nipponbare annotated cDNA sequences, RAP-DB sequences, and EST sequences of other cultivars. The similarity search was performed by using BLASTN with E-value at 1E-50 and screened alignments of at least 95% identity and 50% coverage.

Original		Contig				Singleton			
subsp.	cultivar	coverage	mm	gap	pol	coverage	mm	gap	pol
hybrid	LYP9	92.765	0.675	0.146	0.821	87.276	0.913	0.224	1.137
	PA64s	89.982	0.540	0.124	0.663	84.265	0.833	0.223	1.056
indica	9311	92.736	0.733	0.164	0.898	85.295	1.036	0.205	1.241
	IR36	88.448	0.766	0.067	0.833	83.883	1.173	0.086	1.259
	IR64	80.875	1.719	0.162	1.881	64.812	2.224	0.234	2.458
	Milyang23	90.899	1.660	0.387	2.047	76.821	2.595	0.752	3.347
	Minghui63	73.666	1.216	0.305	1.522	67.576	1.342	0.471	1.813
	Nagina22	83.904	0.965	0.210	1.174	68.110	3.104	0.787	3.891
	Pokkali	81.958	1.241	0.193	1.434	73.218	1.666	0.239	1.905
japonoca	Donganbyeo	79.565	0.809	0.309	1.118	79.565	0.809	0.309	1.118
	Koshihikari	93.352	0.582	0.179	0.761	86.488	0.715	0.276	0.991
	Nackdong	88.185	0.631	0.094	0.725	93.835	1.672	0.105	1.777
	PI560247	94.405	0.543	0.046	0.589	92.722	0.756	0.050	0.806
Screened		Contig				Singleton			
subsp.	cultivar	coverage %	mm %	gap %	pol %	coverage %	mm %	gap %	pol %
hybrid	LYP9	94.858	0.478	0.143	0.621	93.868	0.597	0.205	0.802
	PA64s	93.768	0.379	0.120	0.500	92.376	0.471	0.195	0.666
indica	9311	95.816	0.564	0.157	0.720	94.948	0.699	0.192	0.891
	IR36	93.012	0.528	0.060	0.588	92.542	0.674	0.069	0.743
	IR64	91.372	1.052	0.144	1.196	90.460	1.256	0.207	1.463
	Milyang23	91.234	1.228	0.368	1.596	88.617	1.692	0.545	2.237
	Minghui63	76.390	0.812	0.288	1.100	76.762	0.843	0.356	1.200
	Nagina22	92.282	0.666	0.201	0.867	91.457	0.939	0.406	1.345
	Pokkali	86.120	0.914	0.184	1.098	86.369	1.005	0.225	1.230
japonoca	Donganbyeo	83.271	0.714	0.312	1.026	84.779	1.085	0.419	1.505
	Koshihikari	95.272	0.218	0.170	0.388	94.969	0.328	0.218	0.546
	Nackdong	91.828	0.274	0.089	0.363	93.031	0.305	0.104	0.409
	PI560247	96.796	0.310	0.041	0.350	96.023	0.351	0.040	0.391

Supplementary Table 2. Polymorphism between cultivars from the present work and previous studies.

Comparison type	Cultivars compared			SNP %	Indel %	Polymorphism %	Data sources	Ref.
hybrid/japonica	LYP9	vs.	Nipponbare	0.478	0.143	0.621	EST / RAP-DB	a
	PA64s	vs.	Nipponbare	0.379	0.120	0.500	EST / RAP-DB	a
indica/japonica	9311	vs.	Nipponbare	0.564	0.157	0.720	EST / RAP-DB	a
	9311	vs.	Nipponbare	0.710	0.200	0.910	BGI / IRGSP	3
	9311	vs.	Nipponbare	0.170	0.011	0.181	BGI / IRGSP	1
	GLA4	vs.	Nipponbare	0.630			PCR amplicons	4
	GLA4	vs.	Koshihikari	0.637			PCR amplicons	4
	IR36	vs.	Nipponbare	0.528	0.060	0.588	EST / RAP-DB	a
	IR64	vs.	Nipponbare	1.052	0.144	1.196	EST / RAP-DB	a
	Kasalath	vs.	Nipponbare	0.733			PCR amplicons	4
	Kasalath	vs.	Koshihikari	0.744			PCR amplicons	4
	Kasalath	vs.	Nipponbare	0.710	0.123	0.833	BES / IRGSP	2
	Milyang23	vs.	Nipponbare	1.228	0.368	1.596	EST / RAP-DB	a
	Minghui63	vs.	Nipponbare	0.812	0.288	1.100	EST / RAP-DB	a
	Nagina22	vs.	Nipponbare	0.666	0.201	0.867	EST / RAP-DB	a
	Pokkali	vs.	Nipponbare	0.914	0.184	1.098	EST / RAP-DB	a
japonoca/japonoca	Donganbyeo	vs.	Nipponbare	0.714	0.312	1.026	EST / RAP-DB	a
	Koshihikari	vs.	Nipponbare	0.218	0.170	0.388	EST / RAP-DB	a
	Koshihikari	vs.	Nipponbare	0.038			PCR amplicons	4
	Nackdong	vs.	Nipponbare	0.274	0.089	0.363	EST / RAP-DB	a
	PI560247	vs.	Nipponbare	0.310	0.041	0.350	EST / RAP-DB	a
	TNG67	vs.	Nipponbare	0.216	0.095	0.311	EST / RAP-DB	a
indica/indica	GLA4	vs.	Kasalath	0.547			PCR amplicons	4

[a]Present study.
[1]Feltus et al., 2004; [2] Katagiri et. al., 2004; [3] Shen et. al., 2004; [4] Monna et. al., 2006.