DNA Research **12**, 417–427 (2005) doi:10.1093/dnares/dsi019

Genome-wide Investigation of Intron Length Polymorphisms and Their Potential as Molecular Markers in Rice (*Oryza sativa* L.)

Xusheng Wang,† Xiangqian Zhao,† Jun Zhu, and Weiren Wu*

Institute of Bioinformatics, Huajiachi Campus, Zhejiang University, Hangzhou, 310029, P. R. China

(Received 6 October 2005; published online 23 February 2006)

Abstract

Intron length polymorphisms (ILPs) have been used as genetic markers in some studies. However, a systematic investigation and large-scale exploitation of ILP markers has not been reported. In this study, we performed a genome-wide search of ILPs between two subspecies (*indica* and *japonica*) in rice using the draft genomic sequences of cultivars 93-11 (*indica*) and Nipponbare (*japonica*) and 32 127 full-length cDNA sequences of Nipponbare obtained from public databases. We identified 13 308 putative ILPs. Based on these putative ILPs, we developed 5811 candidate ILP markers via electronic-PCR with primers designed in flanking exons. We further conducted experiment to verify the candidate ILP markers. Out of 215 candidate ILP markers tested on 93-11, Nipponbare and their hybrid, we successfully exploited 173 codominant ILP markers. Further analyses on 10 rice accessions showed that these ILP markers were widely applicable and most (71.1%) exhibited subspecies specificity. This feature suggests that ILPs would be useful for the studies of genome evolution and inter-subspecies heterosis and for cross-subspecies marker-assisted selection in rice. In addition, by testing 51 pairs of the ILP primers on five *Gramineae* plants and three dicot plants, we found another desirable characteristic of rice ILP markers that they have high transferability to other plants.

Key words: rice (*Oryza sativa* L.); intron length polymorphism (ILP); molecular marker; genome

1. Introduction

Introns are non-coding sequences in a gene that are transcribed but spliced out of the precursor mRNA.[1,2] Introns are widespread and abundant in eukaryotic genomes.[3,4] For example, introns constitute ∼11 and 24% of the fruit fly[5] and human[6] genomes, respectively. Generally speaking, introns have little functional significance, although some introns may influence the level of gene expression.[7] Therefore, introns are more variable than coding sequences.

Variations or polymorphisms in DNA sequences can be exploited as genetic markers (usually called molecular markers), which are very useful tools for genetic research (e.g. construction of genetic maps, mapping of genes or quantitative trait loci) and breeding (e.g. marker-assisted selection). Botstein et al.[8] first utilized restriction fragment length polymorphisms (RFLPs) as genetic markers to construct human genetic map. Since then, many new molecular markers have been developed, such as random amplified polymorphic DNA (RAPD),[9] amplified fragment length polymorphism (AFLP),[10] microsatellite or simple sequence repeat (SSR),[11,12] sequence-related amplified polymorphism (SRAP)[13] and single-nucleotide polymorphism (SNP).[14]

Intron polymorphisms can also be exploited as genetic markers. They have been successfully utilized in population genetics surveys[15–17] and gene mapping.[18] There could be various polymorphisms in introns, but intron length polymorphism (ILP) is the most easily recognizable type. It can be conveniently detected by the PCR. To amplify introns by PCR, primers can be designed in flanking exons. This approach is called exon-primed intron-crossing PCR (EPIC-PCR).[19] The advantage of EPIC-PCR is that exon sequences are relatively more conservative and therefore the primers designed in exons may have more extensive applications than those designed in non-coding sequences. Using this approach,

Communicated by Satoshi Tabata

* To whom correspondence should be addressed. Tel/Fax. +86-571-86971910, E-mail: wuwr@zju.edu.cn

† These authors contributed equally to this work.

Bierne et al.[20] developed several ILP markers in penaeid shrimps. To date, however, studies of exploiting intron polymorphism markers have been restricted to a few genes. No efforts on genome-wide exploitation of intron polymorphism markers have been reported.

In rice (*Oryza sativa*), draft genome sequences of two cultivars, 93-11[21] and Nipponbare,[22] representing *indica* and *japonica* subspecies, respectively, and a set of over 28 000 full-length cDNA sequences from Nipponbare[23] have been released. Moreover, complete sequences of chromosomes 1, 4 and 10 and, more recently, the whole genome of Nipponbare have been published.[24–27] In addition, a tentative assembly of all chromosomes of rice has been released (The Institute of Genomic Research, TIGR; http://www.tigr.org). These data provide us with an opportunity to systematically search for DNA polymorphisms between the two subspecies and to exploit DNA markers on a large scale in rice. Recently, Shen et al.[28] and Feltus et al.[29] independently conducted genome-wide investigations of SNPs and InDels or single-base InDels using the same set of released genome drafts of 93-11 and Nipponbare. However, it is astonishing that while Shen et al.[28] identified 1 703 176 SNPs and 479 406 InDels, among which there were 277 858 single-base InDels according to our count from their online database, Feltus et al.[29] only identified 384 341 SNPs and 24 557 single-base InDels—the numbers of SNPs and single-base InDels identified by Shen et al.[28] are 4.4 and 11.3 times greater than those identified by Feltus et al.[29] respectively. The great inconsistency between the two studies suggests that there might be many mistakes in the reported SNPs and InDels.

ILPs are caused by InDels, but many of them cannot be simply considered equivalent to the generally defined InDels because an ILP may contain several (instead of only one) InDels. In addition, even if we take all ILPs as InDels, they are at least a special subset of InDels exhibiting polymorphisms in the non-coding regions of genes. This may make ILPs possess special characteristics and usefulness. In the work described here, we performed a genome-wide search for ILPs and a large-scale exploitation of candidate ILP markers via electronic EPIC-PCR based on the released genomic and cDNA sequence data in rice. We also developed a set of ILP markers selected from the candidates and investigated their characteristics by experiment. Moreover, we established a web-accessible database for rice ILP markers.

2. Materials and Methods

2.1. Data sources of rice genomic and cDNA sequences

We downloaded genomic sequence data of two rice cultivars, Nipponbare (ssp. *japonica*) and 93-11 (ssp. *indica*), released by the International Rice Genome Sequencing Project (IRGSP) and Beijing Genomics Institute (BGI), from web sites http://www.tigr.org/ and http://rise.genomics.org.cn/rice/link/download.jsp, respectively. In addition, we downloaded 32 127 full-length cDNA sequences of Nipponbare from web site http://cdna01.dna.affrc.go.jp/cDNA/.

2.2. Search of ILPs

We developed a pipeline using Perl script to search ILPs between Nipponbare and 93-11 and exploit candidate ILP markers. The initial step was to identify the most likely positions of the available cDNA sequences in rice genome by aligning the cDNA sequences of Nipponbare with the genomic sequences of Nipponbare and 93-11 using BLASTN.[30] We used a high *E*-value ($= 10^{-20}$) for the BLASTN to remove paralogues. Then we used the program SIM4[31] to align each cDNA with its corresponding BAC clone from Nipponbare and Scaffold from 93-11 to examine the gene structure (number of introns and positions of splice sites) and putative ILPs between the two cultivars in the gene. Although there could be several possible types of ILPs, we restricted our ILP search to those genes that showed the same structure (i.e. same number of introns with same positions of splice sites) in the two cultivars because ILPs in those genes could potentially be exploited as PCR-based codominant markers.

2.3. Exploitation of candidate ILP markers by electronic EPIC-PCR

To exploit ILP markers from the putative ILPs identified by BLASTN and SIM4, we designed PCR primers based on the cDNA sequences (of Nipponbare) corresponding to the flanking exons using ePrime3 (http://www.hgmp.mrc.ac.uk/).[32] For convenience, we used a 200 bp cDNA sequence with 100 bp on each side of the target intron for the primer design for each ILP. A simulation study conducted beforehand had shown that the cDNA length could guarantee 93% success in primer design. Then we tested the designed primers by electronic PCR (e-PCR)[33] on the genomic sequences of Nipponbare and 93-11, respectively. To increase the quality and usability of the *in silico* exploited ILP markers, we required exact matches between primers and templates and set a 1100 bp margin on the product size for the e-PCR. We took a putative ILP locus as a candidate ILP marker when it was successfully and uniquely detected by the e-PCR and named it with the abbreviation RI (for Rice ILP) followed by a unique number (e.g. RI03281).

2.4. Verification and evaluation of ILP markers by experiment

We selected 215 candidate ILP markers for the experimental evaluation. According to the RGP's rice genetic map,[34] the selected candidate ILP markers were approximately evenly distributed in rice genome with an average

of ∼8 cM between adjacent ILP loci. While we directly adopted the PCR primers designed in flanking exons for most of the selected candidate ILP markers, we also redesigned primers in introns for some with smaller intron length difference (ILD) but larger intron size in order that polymorphisms could be detected via electrophoresis. All primers were synthesized by either Shanghai Sangon Biological Engineering & Technology Company or Shanghai BioAsia Biotechnology Company.

We tested the synthesized primers at first using Nipponbare, 93-11 and their F_1. We extracted genomic DNAs from young leaves using CTAB method[35] with modification. We conducted PCR in a 15 μl reaction mixture containing 50 ng template DNA, 0.5 μM of each primer, 200 μM of each dNTP, 1.5 mM $MgCl_2$, 0.1% Triton X-100 and 1 U *Taq* polymerase and 1.5 μl of 10× PCR buffer. We first tested all primer pairs with a touchdown PCR (Td-PCR)[36] program: 5 min at 94°C; 10 cycles of 30 s at 94°C, 30 s at 59°C minus 0.3°C/cycle, 1 min at 72°C; 20 cycles of 30 s at 94°C, 30 s at 56°C, 1 min at 72°C; and 5 min at 72°C for final extension. For primer pairs that did not generate good amplification results, we adjusted the initial annealing temperature (59°C) to 60 or 57°C. The purpose of using Td-PCR was to increase the specificity of amplification, but some primer pairs only required a routine PCR program: 5 min at 94°C; 35 cycles of 30 s at 94°C, 30 s at 54°C, and 1 min at 72°C; and 5 min at 72°C for final extension. For most primers, we used 6% non-denaturing PAGE (250 V, 2 h) for separating PCR products and silver staining for visualizing DNA bands following Xu et al.[37] with modification. We also used 2% agarose gel for some primers.

For primers generating correct PCR products as expected, we further tested them using 10 rice varieties including Nipponbare and 93-11 (Table 1), which were kindly provided by the China National Rice Research Institute (CNRRI). Based on the PCR data, we evaluated the allelic diversity of each ILP marker using the polymorphism information content (PIC) value defined as $PIC_i = 1 - \sum_{j=1}^{n} p_{ij}^2$, where p_{ij} is the frequency of the jth pattern for the ith marker.[38]

We also employed some of the ILP primers to perform PCR in other plants including five *Gramineae* plants (wheat, barley, maize, sorghum and bamboo) and three dicot plants (rape, cotton and tobacco), using either a Td-PCR program (with an initial annealing temperature of 55°C) or a routine PCR program (with an annealing temperature of 52°C).

Table 1. Rice accessions used for testing ILP markers

Accession	Type	Origin
93-11	*Indica*	China
Guangluai-4	*indica*	China
Xieqingzao	*indica*	China
IR64	*Indica*[a]	Philippines
Kyeema	*Indica*[a]	Australia
Nipponbare	*japonica*	Japan
Xiushui-11	*japonica*	China
Koshihikari	*japonica*	Japan
Katy	*Japonica*[a]	USA
Merim	*javanica*	Indonesia

[a]Might not be typical *indica* or *japonica* according to its pedigree.

3. Results and discussion

3.1. Number, distribution and density of ILPs in rice

By aligning 32 127 full-length cDNA sequences from Nipponbare with the genomic sequences of Nipponbare and 93-11 using BLASTN and SIM4, we found 120 489 and 108 312 introns in the two cultivars, respectively, and identified 13 308 putative ILPs between the two cultivars. All the cDNAs were localized to the BAC clones of Nipponbare as expected, but 1279 (3.98%) cDNAs did not align to the scaffolds of 93-11 with an *E*-value below the BLAST criterion of 10^{-20}. That could be the reason that fewer introns were found in 93-11. By referring to the TIGR psuedomolecule assembly of rice, we have plotted the density distribution curve of ILPs in rice genome (Fig. 1). Due to discrepancies between GenBank and TIGR in the assembly of some BAC clones, 32 ILPs could not be assigned to chromosomes and 302 ILPs could not be located, although their chromosomes were known. Therefore, we did not count these ILPs when plotting the density distribution curves, but still took those ILPs into account when calculating the total number and overall density of ILPs on each chromosome (Fig. 1). Most (>100) of the non-located ILPs were in chromosome 10.

It is obvious from Fig. 1 that the ILP density fluctuates dramatically along the genome and varies among chromosomes, ranging from 23.04 per Mb (chromosome 12) to 45.31 per Mb (chromosome 2) with an average of 32.52 per Mb; or from 6.25 per cM (chromosome 9) to 11.45 per cM (chromosome 2) with an average of 8.48 per cM. ILPs are clearly not randomly distributed in rice genome. In addition, the number of ILPs on each chromosome also varies greatly, ranging from 584 (chromosome 9) to 1848 (chromosome 3) with an average of 1106 (Fig. 1). The ILPs on chromosomes 1, 2 and 3 together constitute ∼40% of the total number.

Rice genome is estimated to contain 46 022–55 615 genes.[21] In this study, we have found 13 308 ILPs between Nipponbare and 93-11 based on 32 127 full-length cDNA sequences, suggesting that there are 0.414 ILPs per cDNA on average. If we approximately consider a cDNA as a gene, then we can deduce that the total number of ILPs between the two cultivars would be 19 064 to 23 037 according to the estimate obtained in this study. It should

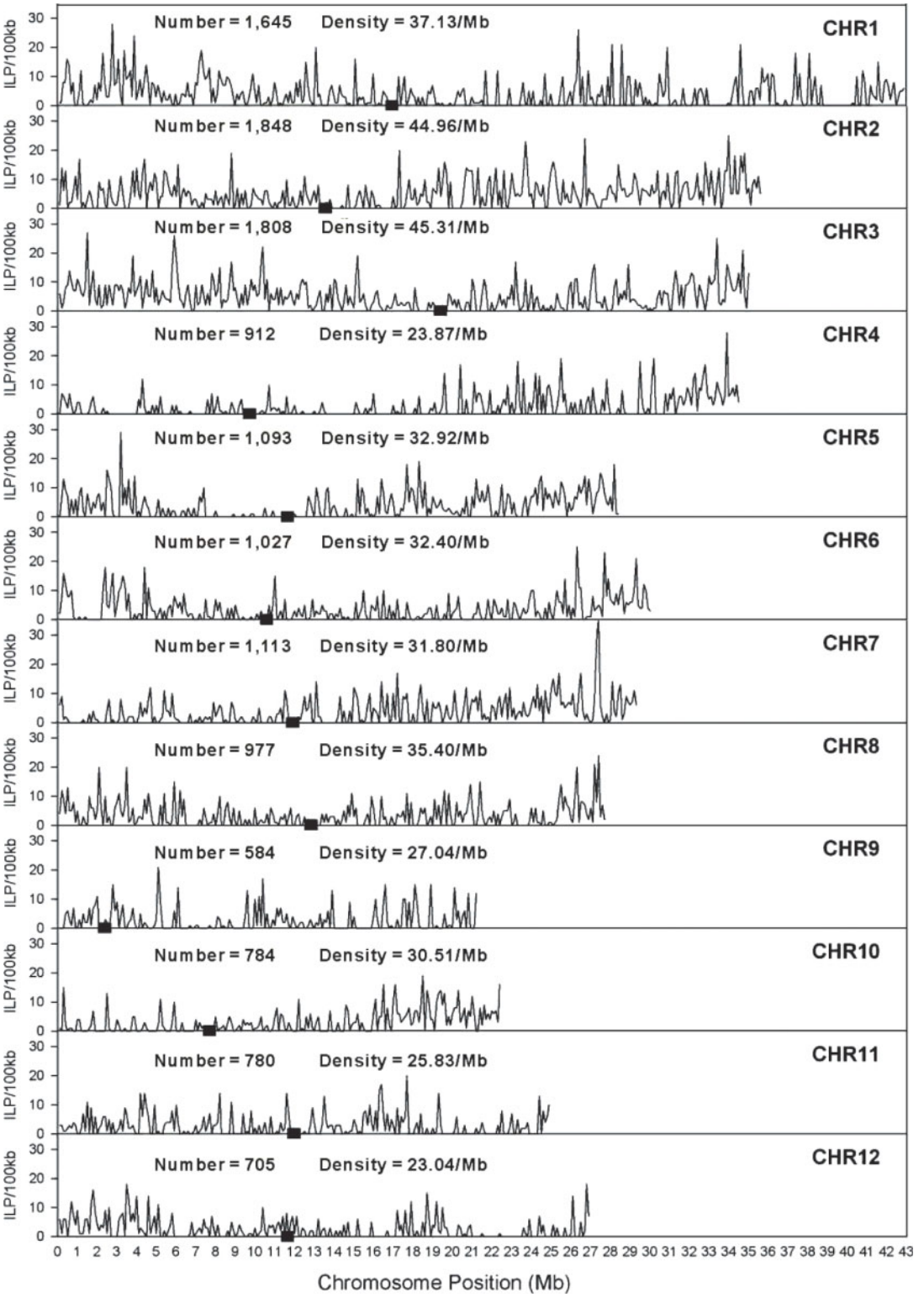

Figure 1. Density distribution of ILPs across rice genome. A gray rectangle on each x-axis indicates the position of centromere. The total number and overall density (number per Mb) of ILPs on each chromosome are also presented.

be emphasized that we restricted the ILP search to those genes having the same structure (number and positions of introns) in both Nipponbare and 93-11. Therefore, there were some genes not taken into account. In fact, we have mentioned above that 1279 (3.98%) cDNAs did not hit the scaffolds of 93-11 in the BLAST with an E-value below 10^{-20}. These genes might either be specific to *japonica* or have large variation between *indica* and *japonica* during

evolution.[39] For the former case, all introns in the genes could be taken as ILPs with null alleles in *indica* and therefore could be potentially exploited as dominant ILP markers. For the latter case, ILPs could also exist. We thus see that ILPs are very rich in the rice genome and should be a huge resource of molecular markers.

3.2. Candidate ILP markers

Using primers designed in flanking exons, we successfully obtained e-PCR products from 10 572 (79.4%) and 7742 (58.2%) putative ILP loci in Nipponbare and 93-11, respectively. Although we designed the primers based on the cDNAs from Nipponbare, we failed to acquire e-PCR products from ∼1/5 putative ILP loci in Nipponbare, probably due to the several constraint conditions set for the primer design and e-PCR (see Materials and Methods). There were a higher proportion of putative ILPs in 93-11 not detected by e-PCR because mismatches might occur between some primers and the genomic sequence of 93-11. Although perfect match between primer and template was required in the e-PCR, there still were 1009 primer pairs detecting multiple BAC clones located on different chromosomes and appearing to have multiple copies in Nipponbare. Similarly, 880 primer pairs detected multiple scaffolds and showed multiple copies in 93-11. A typical example was the primer pair designed in cDNA AK064639, which detected a total of 106 occurrences on *japonica* genome, indicating that the primer pair might be designed in the conserved sequences of a big gene family. As multiple-copy is not desirable for molecular markers, we discarded these primer pairs. It is noted that the 10 572 ILP loci detected by e-PCR in Nipponbare were located only on 2405 (∼61%) BAC clones, leaving 1526 clones without ILP hits. The result also reflects the nonrandom distribution of ILPs in rice genome.

By combining the e-PCR results in Nipponbare and 93-11, we obtained 5811 candidate single-copy ILP markers. The number of candidate ILP markers on each chromosome ranged from 130 (chromosome 9) to 990 (chromosome 2) with an average of 484; and the density on each chromosome ranged from 4.91 per Mb (chromosome 4) to 24.81 per Mb (chromosome 2) with an average of 13.06 per Mb, or from 1.39 per cM (chromosome 9) to 6.27 per cM (chromosome 2) with an average of 3.42 per cM (Fig. 2). Based on the TIGR psuedomolecule assembly of rice, we have constructed a physical map of the 5811 candidate ILP markers and a comparative map of 2275 RFLP and 2740 SSR markers (Fig. 2). The map shows that the candidate ILP markers are not evenly distributed. This seems to be consistent with the distribution of ILPs (Fig. 1). However, it is surprising that some genomic regions (particularly in the long arms of chromosomes 1, 4, 9 and 12) are nearly devoid of candidate ILP markers although the putative ILPs in these regions are not rare (Fig. 2). By examining the distribution of e-PCR hits in the whole genome, we found that these regions appeared to have high frequencies of No-EPCR-Hit-in-93-11 (NEH9) (Fig. 3). This could explain why there were so few candidate ILP markers obtained in these regions. The major reason of NEH9 for a putative ILP locus could be that mismatches occurred between the primers designed based on the cDNA of Nipponbare and the genomic sequence of 93-11. The high frequencies of mismatches in these regions imply that these regions might have a higher level of genetic variation between the two subspecies. To develop ILP markers in these regions, we need to identify conserved sequences of each gene for designing PCR primers.

The length difference between allelic introns (referred to as intron length difference, ILD) in the candidate ILP marker loci appeared to follow an exponential distribution with a mean value of 11.42 bp (Fig. 4). Most (72.6%) of the ILDs were <5 bp; 23.5% fell between 5 and 50 bp; and very few (3.9%) were >50 bp. Generally speaking, the larger the ILD is, the easier the detection will be. Therefore, the candidate ILP markers with ILDs $\geq$ 5 bp should be preferentially considered in practical studies. However, the candidate ILP markers with small ILDs (<5 bp) should not be ignored too, because they are in the majority.

Since ILPs are detected with specific PCR primers, they usually can be used as STS (sequence-tagging site) markers like SSRs. By examining the intron sequences of the 5811 candidate ILP markers, we found that only 208 (3.58%) of the ILPs were due to SSR variation, of which TA was the most frequent motif (19.2%), followed by GA (14.4%). This means that there is only a very small overlap between ILPs and SSRs in rice. In addition, we have seen that ILPs exist in many gaps in the physical map of RFLP and SSR markers (Fig. 2). Therefore, ILPs are a new source of STS markers different from SSRs and can complement SSR and RFLP markers. In principle, every ILP is potentially a genetic marker as long as a suitable detection method is available.

3.3. ILP markers exploited by experiment

In order to detect ILPs by non-denaturing PAGE or agarose gel electrophoresis, we only chose candidate ILP markers with ILD $\geq$ 3 bp for the experiment. Of the 215 candidate ILP markers tested, 173 (80.47%) yielded stable and clear PCR products as expected in both Nipponbare and 93-11, and appeared to be codominant in the F_1. Besides, six (2.79%) candidate ILP markers yielded the expected PCR products in either of the parents and appeared to be dominant in the F_1. The number of ILP markers on each chromosome is shown in Table 2.

To increase PCR specificity in the amplification of ILP loci, we adopted a Td-PCR program. Most (138/173) of the ILP markers obtained could be amplified well by the

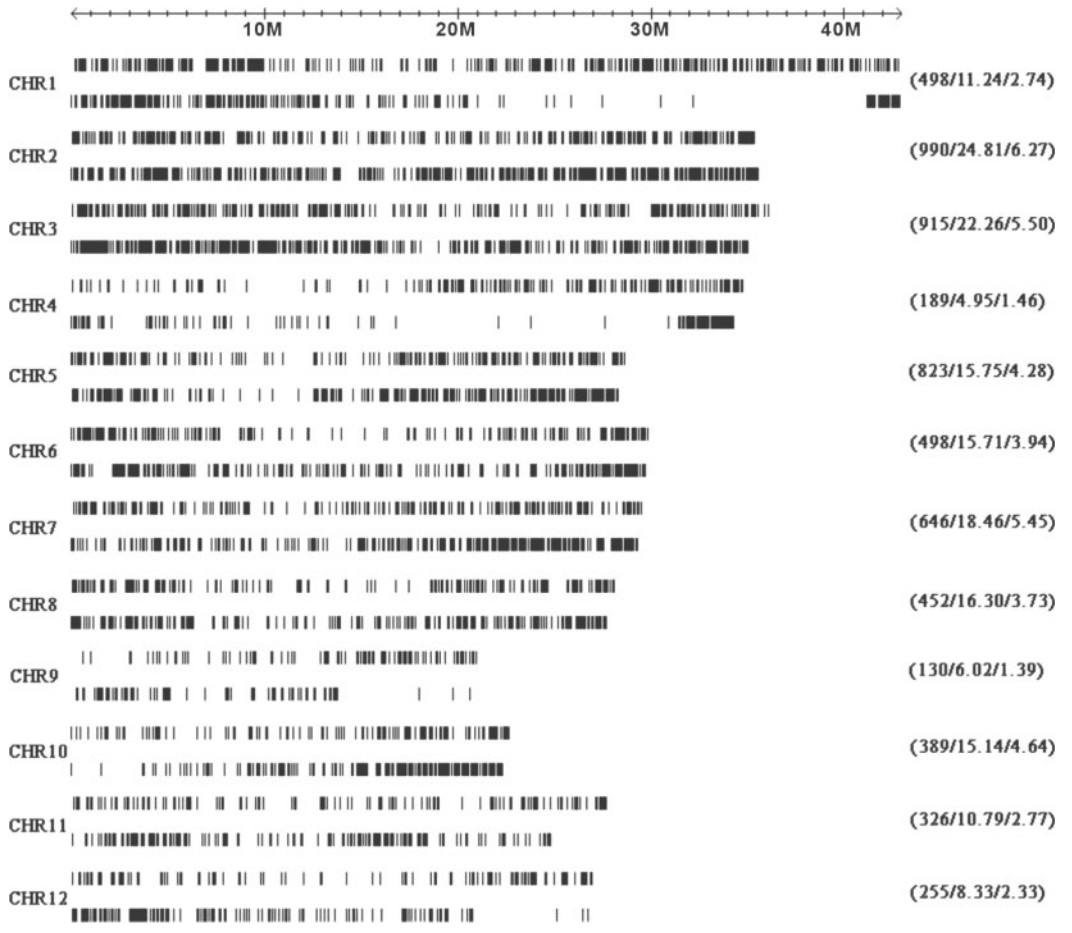

Figure 2. Physical map of 5811 candidate ILP markers and comparative map of 2275 RFLP and 2740 SSR markers. Each vertical short bar indicates the position of a candidate ILP marker. The three numbers in the brackets on the right of the map of each chromosome are the total number, number per Mb and number per cM of candidate ILP markers.

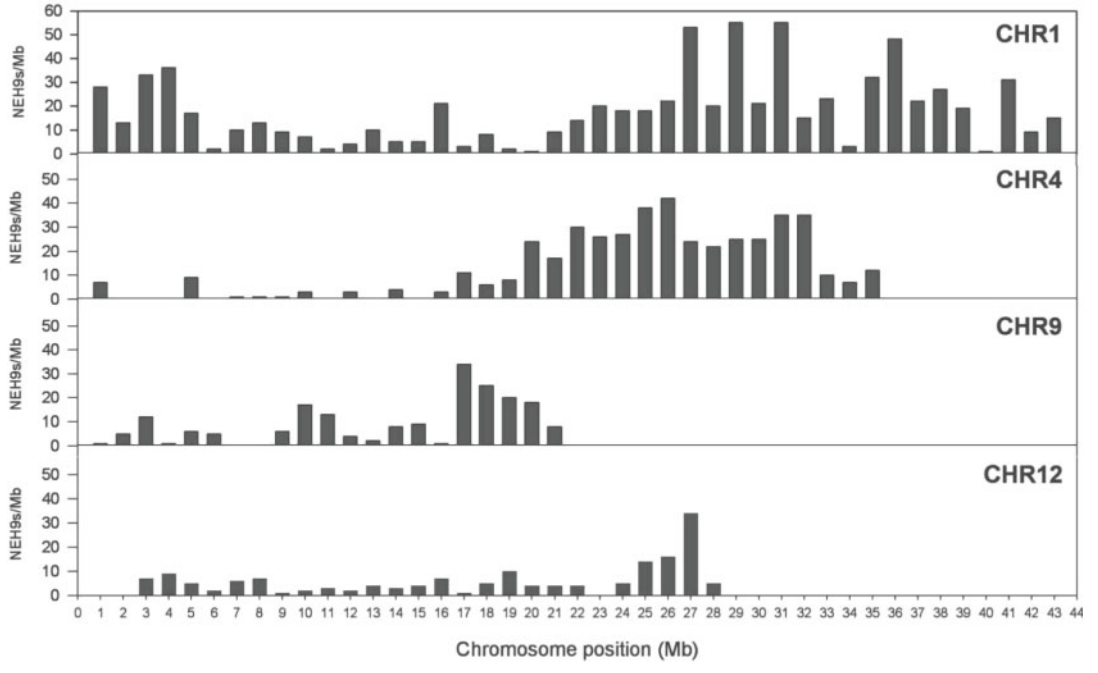

Figure 3. Density distribution of NEH9s on chromosomes 1, 4, 9 and 12. The horizontal axis shows the pseudomolecule position (Mb); the vertical axis shows the number of NEH9s per Mb.

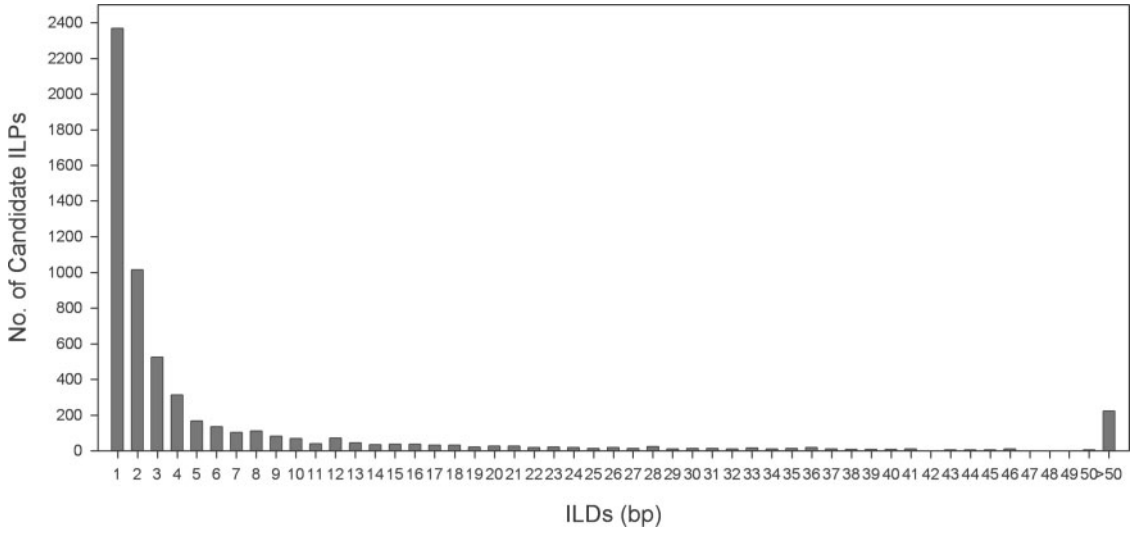

Figure 4. The number distribution of intron length difference in 5811 candidate ILP markers.

Table 2. Number of ILP markers obtained by experiment

chromosome	Based on EPIC-PCR			Based on WIN-PCR		
	Analyzed	Detected	%	Analyzed	Detected	%
1	14	10	71.43	6	5	83.33
2	20	15	75.00	6	6	100.00
3	16	14	87.50	9	7	77.78
4	5	5	100.00	4	4	100.00
5	14	13	92.86	5	4	80.00
6	12	10	83.33	10	6	60.00
7	13	11	84.62	5	4	80.00
8	15	12	80.00	8	6	75.00
9	5	5	100.00	3	3	100.00
10	10	8	80.00	6	5	83.33
11	12	11	91.67	3	3	100.00
12	10	8	80.00	4	4	100.00
Total	146	122	83.56	69	57	82.61

Td-PCR with the default initial annealing temperature (59°C). Some other (25 or 8) markers required a lower (57°C) or higher (60°C) initial annealing temperature. Eight markers could be well amplified at a constant annealing temperature (54°C).

We further tested the 173 codominant markers on 10 rice varieties. All the markers were successfully detected in those varieties (Fig. 5), suggesting that the markers exploited are widely applicable. Based on the resolution capacity of the non-denaturing PAGE or agarose gel used, it appeared that most of the markers only possessed 2 (i.e. Nipponbare's and 93-11's) alleles among the 10 varieties. Only nine markers appeared to have multiple alleles, of which five were attributed to SSR variation. The PIC values of the markers varied from 0.18 to 0.66 with an average of 0.451. The results indicate that the

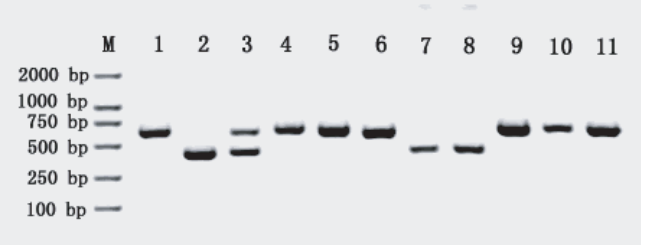

Figure 5. PCR products of ILP marker RI01015 in rice accessions separated by electrophoresis on 2% agarose gel. Lanes from left to right: M = DNA molecular weight marker; 1 = 93-11; 2 = Nipponbare; 3 = 9311/Nipponbare F_1; 4 = Guangluai-4; 5 = Xieqingzao; 6 = IR64; 7 = Koshihikari; 8 = Xiushui-11; 9 = Merim; 10 = Katy; 11 = Kyeema.

polymorphism level of ILP marker is not high in general. However, a higher estimate of the polymorphism level of ILP markers could probably be obtained if methods of DNA fragment analysis with higher resolution capacity (e.g. denaturing PAGE, usually used for SSR analysis or DNA sequencing) were adopted.

3.4. WIN-PCR for ILP detection

In the present study, we adopted electronic EPIC-PCR to screen for candidate ILP markers. However, because the ILDs of >70% candidate ILP markers were <5 bp (Fig. 3) and the average size of the EPIC-PCR products was ∼580 bp, most of the candidate ILP markers would be difficult to detect by real EPIC-PCR. To solve this problem, a possible way is to design PCR primers within introns so as to obtain smaller PCR products. For this purpose, we examined nucleotide substitutions (SNPs) in introns between Nipponbare and 93-11 by sequence comparison *in silico*. We identified 17 374 putative nucleotide substitutions in the introns of the 5811 candidate

Table 3. Single-nucleotide substitutions between 93-11 and Nipponbare in the intron sequences of 5811 candidate ILP markers

Type of substitution	Number	Proportion (%)
G/C	1757	10.11
A/T	3032	17.45
A/C	2180	12.55
G/T	2145	12.35
A/G	4117	23.70
T/C	4143	23.84
Total	17374	100

ILP markers, approximately equivalent to 6 SNPs/kb (Table 3). Although the estimate of SNP frequency in intron sequences between the two subspecies of rice is much larger than that of whole genome average (1.06 SNPs/kb),[29] it is still not very high. Hence, it should be suitable to design PCR primers within introns to detect cross-subspecies ILPs in rice. To distinguish this technique, we call this approach Within INtron PCR (WIN-PCR). In this study, we have obtained 57 ILP markers from 69 candidate ILP markers by WIN-PCR. The success rate (57/69 = 82.61%) is very similar to that of EPIC-PCR (122/146 = 83.56%) (Table 2). The results indicate that WIN-PCR could be as efficient as EPIC-PCR. In principle, WIN-PCR would permit the detection of single-nucleotide ILPs.

3.5. Subspecies specificity and intra-subspecies diversity of ILPs

Among the 10 accessions used for ILP analysis in the present study (Table 1), 93-11, Guangluai-4 and Xieqingzao could be taken as typical *indica* cultivars and Nipponbare, Xiushui-11 and Koshihikari as typical *japonica* cultivars according to their origins. Based on these typical accessions, 123 (71.1%) out of the 173 ILP markers tested showed subspecies-specific genotypes (i.e. different between but the same within subspecies). The result suggests that ILPs in rice have apparent subspecies specificity. This feature could be useful for analyzing the genetic compositions of rice cultivars. A typical example comes from the accession Kyeema. The accession was categorized as an *indica* cultivar based on morphological characters.[40] However, we have found that it is more likely to be a *japonica* cultivar because out of the 123 subspecies-specific ILP markers assayed on it, only 21 (17.1%) exhibited *indica* genotype, while 100 (81.3%) showed *japonica* genotype. This is consistent with its pedigree. In fact, Kyeema was derived from a triple cross involving one *indica* (Della) and two *japonica* (Pelde and Kulu) parents. We can expect that the offspring of the triple cross (Pelde//Della/Kulu) would contain 25% *indica* and 75% *japonica* genetic components on average. We see that the proportions in Kyeema's genome estimated by ILP markers are close to the expected values. Another example worthy of note is the accession Katy, a suggested *japonica* cultivar derived from a complicated cross Bonnet73/CI9722// Starbonnet/Tetep///Lebonnet, where Tetep is a typical *indica* parent. Our study has found that out of the 123 subspecies-specific ILP markers, 22 (17.9%) showed *indica* genotype in Katy. Therefore, Katy cannot be a typical *japonica* cultivar. In addition, 9 ILP markers showed heterozygous genotypes in Katy, suggesting that the accession might not be a pure line. A more interesting result we obtained concerns the *javanica* variety Merim. In this accession, 94 (or 76.4%) out of the 123 subspecies-specific ILP markers showed *japonica* genotype. This confirms that *javanica* belongs to *japonica*.[41]

Although ILPs in rice have strong subspecies specificity, they still exist within subspecies. Based on the three typical *indica* and three typical *japonica* cultivars mentioned above, we found that of the 173 ILP markers tested, 44 (25.4%) showed polymorphisms among *indica* cultivars and 10 (5.8%) showed polymorphisms among *japonica* cultivars. The result indicates that *indica* rice has much higher genetic diversity than *japonica* rice. This is consistent with previous studies based on RAPD[42] and RFLP[43] markers. The finding implies that *indica* rice might have evolved earlier than *japonica* rice.

3.6. Transferability of ILP markers to other plants

We randomly selected 51 pairs of rice ILP primers to perform EPIC-PCR in other 8 plants (see Materials and Methods) and found that 24 (47.1%) pairs yielded desirable results with 1–5 clear and stable bands in all the plants; 31 (60.8%) pairs worked well in all the 5 *Gramineae* plants; only 6 (11.8%) pairs did not generate any PCR products. In cotton alone, 36 (70.6%) pairs of the primers yielded clear and stable PCR products and 11 (30.6%) of them revealed polymorphisms between two cultivated cotton species, *Gossypium barbadense* L. and *Gossypium hirsutum* L. The results suggest that a high proportion of rice ILP markers are transferable to other plants.

To examine whether the PCR products in other plants were really specifically amplified or homologous to the target genes in rice, we randomly isolated electrophoretic bands produced by primer pair RI02862 from wheat, maize and cotton, respectively (Fig. 6) and sequenced them (by Shanghai Sangon Biological Engineering & Technology Company). Multiple alignment of the sequences of wheat, maize and cotton together with those of rice using computer programs ClustalX[44] and GeneDoc[45] showed that they were really from homologous genes as expected: the exon regions (two sides) were well conserved and the intron region (middle) were highly varied among the plants (Fig. 7). This suggests that

most, if not all, of the clear and stable PCR products obtained in other plants must be resulted from specific amplification.

For comparison, we also applied 32 pairs of rice SSR primers to the 9 plants. We found that although all the SSR primers could generate PCR products in all the plants, the electrophoretic bands produced by each pair of primers were generally quite many (at least 10) and unstable, suggesting that most, if not all, of the bands are produced by unspecific amplification as seen in the randomly primed markers such as RAPD.[9] Therefore, it appears that most of the rice SSR markers are not transferable to other plants, but can only serve as unspecific PCR markers in other plants.

3.7. Advantages of ILP markers

ILP is a new type of molecular marker, which has not been reported extensively. We have seen that ILPs are abundant between the two cultivated subspecies in rice.

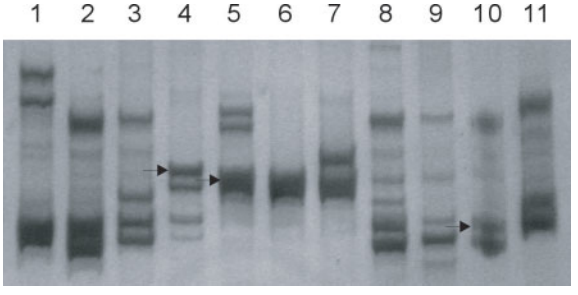

Figure 6. PCR products of ILP marker RI02862 in rice and other plants separated by electrophoresis on 6% non-denaturing PAGE. Lanes from left to right: 1 = *japonica* rice (Nipponbare); 2 = *indica* rice (93-11); 3 = barley; 4 = wheat; 5 = maize; 6 = sorghum; 7 = bamboo; 8 = rape; 9 = tobacco; 10 = cotton (*Gossypium hirsutum* L.); 11 = cotton (*Gossypium barbadense* L.).

ILP has many similar advantages to SSR including specific (being a STS marker), codominant (providing complete information of genotypes), neutral (no phenotypic effect), convenient (detectable by PCR) and reliable (result stable). In addition, ILP has a special advantage, namely, it directly reflects variation within genes. Therefore, the genetic maps constructed with ILP markers would be more valuable for genetic studies because they are similar to conventional maps consisting of morphological markers. Moreover, ILP marker would be more useful for marker-assisted breeding because it allows us to trace a gene directly as long as an ILP can be found in the gene.

We have seen that ILPs have significant subspecies specificity in rice. This characteristic could be useful for genetic study and breeding in rice. Apart from the use for analyzing genetic compositions of rice cultivars discussed above, it might be useful for the studies of genome evolution and inter-subspecies heterosis and for cross-subspecies marker-assisted breeding.

In addition, we have seen the high transferability of rice ILP markers to other plants. This characteristic would make ILP markers very useful for (i) construction of molecular marker maps in other plants that have weaker genetic research basis; (ii) genome comparison among plants; (iii) gene mapping with the help of synteny or collinearity between model plants and other plants; (iv) research of phylogenetic relationships among different species, genera, families or even higher taxonomic ranks in plants and (v) marker-assisted breeding in other plants.

3.8. ILP database

We have established a database (http://ibi.zju.edu.cn/ILPs/index.htm) for depositing the information of the

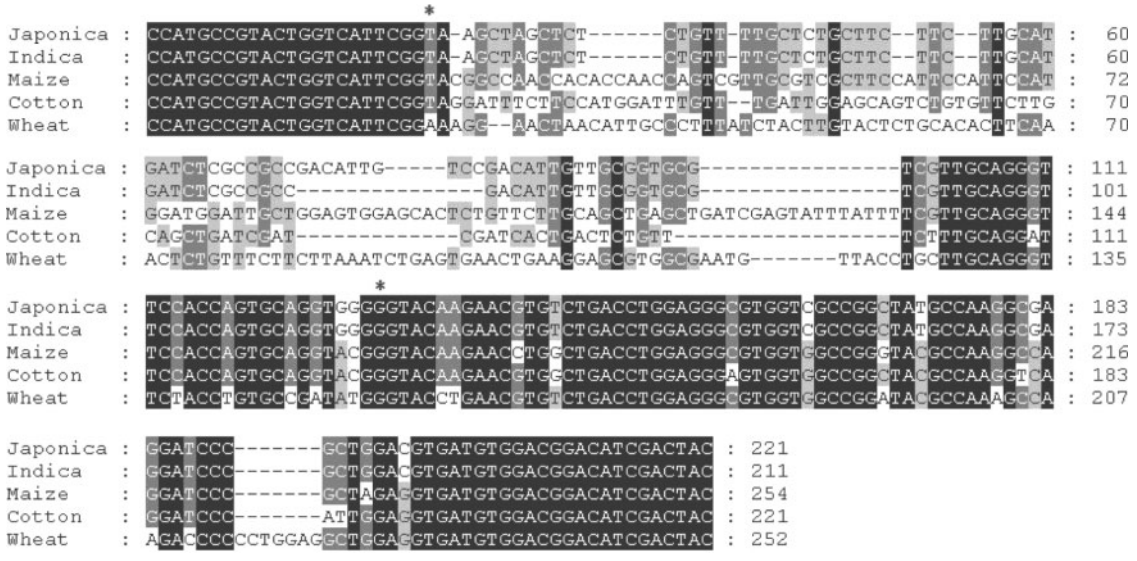

Figure 7. Multiple alignment of sequences amplified from maize, wheat and cotton (*Gossypium hirsutum* L.) by primer pair RI02862 and available target sequences from *japonica* rice (Nipponbare) and *indica* rice (93-11).

5811 candidate ILP markers, including ILP name, cDNA name, *japonica* BAC clone accession number, *japonica* BAC clone name, marker start position in *japonica* (bp), maker end position in *japonica* (bp), marker length in *japonica* (bp), indica scaffold name, marker start position in *indica* (bp), marker end position in *indica* (bp), marker length in *indica* (bp), length difference between *japonica* and *indica* (bp), position in RFLP map (cM), forward primer and reverse primer. For convenience, the ILP information is searched by key words, such as ILP name, BAC clone and scaffold. In addition, the 173 codominant ILP markers obtained have been submitted to GenBank (accession nos.: BV209990–BV210161, BV210393).

Acknowledgments: This work was funded by the National High-Tech Research and Development Program of China (project: 2003AA207160 & 2002AA234031) and by IBM Shared University Research (SUR) program. The authors thank Dr Adrian Cutler from Plant Biotechnology Institute, National Research Council of Canada for helpful suggestions on the manuscript.

References

1. Berget, S. M., Moore, C., and Sharp, P. A. 1977, Spliced segments at the 5′ terminus of adenovirus 2 late mRNA, *Proc. Natl Acad. Sci. USA*, **74**, 3171–3175.
2. Chow, L., Gilinas, R., Broker, T., and Roberts, R. 1977, An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA, *Cell*, **12**, 1–8.
3. Hawkins, J. D. 1988, A survey on intron and exon lengths, *Nucleic Acids Res.*, **16**, 9893–9908.
4. Deutsch, M. and Long, M. 1999, Intron–exon structures of eukaryotic model organisms, *Nucleic Acids Res.*, **27**, 3219–3228.
5. Adams, M. D., Celniker, S. E., Holt, R. A., et al. 2000, The genome sequence of *Drosophila melanogaster*, *Science*, **287**, 2185–2195.
6. Venter, J. C., Adams, M. C., Myers, E. W., et al. 2001, The sequence of the human genome, *Science*, **291**, 1304–1351.
7. Wang, D. G., Fan, J. B., Siao, C. J., et al. 1998, Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, *Science*, **280**, 1077–1082.
8. Botstein, D., White, R. L., Skolnick, M., et al. 1980, Construction of a genetic linkage map in man using restriction fragment length polymorphisms, *Am. J. Hum. Genet.*, **32**, 314–331.
9. Williams, J. G. K., Kubelik, A. R., Livak, K. J., et al. 1990, DNA polymorphism amplified by arbitrary primers are useful as genetic markers, *Nucleic Acids Res.*, **18**, 6531–6535.
10. Vos, P., Hogers, R., Bleeker, M., et al. 1995, AFLP: a new technique for DNA fingerprinting, *Nucleic Acid Res.*, **23**, 4407–4414.
11. Becker, J. and Heun, M. 1995, Barley microsatellites: allele variation and mapping, *Plant Mol. Biol.*, **27**, 835–845.
12. Becker, J., Vos, P., Kuiper, M., et al. 1995, Combined mapping of AFLP and RFLP markers in barley, *Mol. Gen. Genet.*, **249**, 65–73.
13. Li, G. and Quiros, C. F. 2001, Sequence-related amplified polymorphism (SRAP) a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in Brassica, *Theor. Appl. Genet.*, **103**, 455–461.
14. Kruglyak, L. and Nickerson, D. A. 2001, Variation is the spice of life, *Nat. Genet.*, **27**, 234–236.
15. Lessa, E. P. 1992, Rapid survey of DNA sequence variation in natural populations, *Mol. Biol. Evol.*, **9**, 323–330.
16. Côrte-Real, H. B. S. M., Dixon, D. R., and Holland, P. W. H. 1994, Intron-targeted PCR: a new approach to survey neutral DNA polymorphism in bivalve populations, *Mar. Biol.*, **120**, 407–413.
17. Daguin, C. and Borsa, P. 1999, Genetic characterisation of *Mytilus galloprovincialis* Lmk. in North West Africa using nuclear DNA markers, *J Exp. Mar. Biol. Ecol.*, **235**, 55–65.
18. Wydner, K. S., Sechler, J. L., Boyd, C. D., et al. 1994, Use of an intron length polymorphism to localize the tropoelastin gene to chromosome 5 in a region of linkage conservation with human chromosome 7, *Genomics*, **23**, 125–131.
19. Palumbi, S. R. 1995, Nucleic acids II: the polymerase chain reaction, In: Hillis, D. and Moritz, C. (eds) *Molecular Systematics*. 2nd edn. Sinauer Associates Inc., Sunderland, MA, pp. 205–247.
20. Bierne, N., Lehnert, S. A., Bédier, E., et al. 2000, Screening for intron-length polymorphisms in penaeid shrimps using exon-primed intron-crossing (EPIC)-PCR, *Mol. Ecol.*, **9**, 233–235.
21. Yu, J., Hu, S., Wang, J., et al. 2002, A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), *Science*, **296**, 79–92.
22. Goff, S. A., Ricke, D., Lan, T. H., et al. 2002, A draft sequence of the rice genomes (*Oryza sativa* L. ssp. *japonica*), *Science*, **296**, 92–100.
23. Kikuchi, S., Satoh, K., Nagata, T., et al. 2003, Collection, mapping, and annotation of over 28 000 cDNA clones from japonica rice, *Science*, **301**, 376–379.
24. Sasaki, T., Matsumoto, T., Yamamoto, K., et al. 2002, The genome sequence and structure of rice chromosome 1, *Nature*, **420**, 312–316.
25. Feng, Q., Zhang, Y., Hao, P., et al. 2002, Sequence and analysis of rice chromosome 4, *Nature*, **420**, 316–320.
26. The Rice Chromosome 10 Sequencing Consortium (2003), In-depth view of structure, activity, and evolution of rice chromosome 10. *Science*, **300**, 1566–1569.
27. International Rice Genome Sequencing Project (2005), The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
28. Shen, Y. J., Jiang, H., Jin, J. P., et al. 2004, Development of genome-wide DNA polymorphism database for map-based cloning of rice genes, *Plant physiol.*, **135**, 1198–1205.
29. Feltus, F. A., Wan, J., Schulze, S. R., et al. 2004, An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments, *Genome Res.*, **14**, 1812–1819.
30. Altschul, S., Madden, T., Schaffer, A., et al. 1997, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids. Res.*, **25**, 3389–3402.
31. Florea, L., Hartzell, G., Zhang, Z., et al. 1998, A computer program for aligning a cDNA sequence with a genomic DNA sequence, *Genome Res.*, **8**, 967–974.

32. Rice, P., Longden, I., and Bleasby, A. 2000, EMBOSS: The European Molecular Biology Open Software Suite, *Trends Genet.*, **16**, 276–277.
33. Schuler, G. D. 1997, Sequence mapping by electronic PCR, *Genome Res.*, **7**, 541–550.
34. Harushima, Y., Yano, M., Shomura, A., et al. 1998, A high-density rice genetic linkage map with 2275 markers using a single F_2 population, *Genetics*, **148**, 479–494.
35. Murray, M. G. and Thompson, W. F. 1980, Rapid isolation of high-molecular- weight plant DNA, *Nucleic Acids. Res.*, **8**, 4321–4325.
36. Don, R. H., Cox, P. T., Wainwright, B. J., et al. 1991, 'Touchdown' PCR to circumvent spurious priming during gene amplification, *Nucleic Acids. Res.*, **19**, 4008.
37. Xu, S., Tao, Y., Yang, Z., et al. 2002, A simple and rapid method used for silver staining and gel preservation, *Hereditas (Beijing)*, **24**, 335–336.
38. Anderson, J. A., Churchill, G. A., Autrique, J. E., et al. 1993, Optimizing parental selection for genetic linkage maps, *Genome*, **36**, 181–186.
39. Han, B. and Xue, Y. 2003, Genome-wide intraspecific DNA-sequence variations in rice, *Curr. Opin. Plant Biol.*, **6**, 134–138.
40. Wolfgang, S., Marc, H. E., and Peter, M. C. 2002, Semidwarf (*sd-1*), "green revolution" rice, contains a defective gibberellin 20-oxidase gene, *Proc. Natl Acad. Sci. USA*, **99**, 9043–9048.
41. Glaszmann, J. C. 1987, Isozymes and classification of Asian rice varieties, *Theor. Appl. Genet.*, **74**, 21–30.
42. Machill, D. J. 1995, Classifying *japonica* rice cultivars with RAPD markers, *Crop Sci.*, **35**, 889–894.
43. Lu, B. R., Zheng, K. L., Qian, H. R., et al. 2002, Genetic differentiation of wild relatives of rice as assessed by RFLP analysis, *Theor. Appl. Genet.*, **106**, 101–106.
44. Thompson, J. D., Gibson, T. J., Plewniak, F., et al. 1997, The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res.*, **24**, 4876–4882.
45. Nicholas, K. B. and Nicholas, H. B. 1997, GeneDoc: a tool for editing and annotating multiple sequence alignments, Distributed by the authors (http://www.psc.edu/biomed/genedoc).