Breeding Science 63: 246–254 (2013)
doi:10.1270/jsbbs.63.246

Development of INDEL markers to discriminate all genome types rapidly in the genus *Oryza*

Shinichiro Yamaki[1)], Hajime Ohyanagi[1,2)], Masanori Yamasaki[3)], Mitsugu Eiguchi[4)], Toshie Miyabayashi[4)], Takahiko Kubo[1,5)], Nori Kurata[1,5)] and Ken-Ichi Nonomura*[4,5)]

[1)] *Plant Genetics Laboratory, National Institute of Genetics*, Mishima, Shizuoka 411-8540, Japan
[2)] *Mitsubishi Space Software Co., Ltd.*, Takezono 1-6-1, Tsukuba, Ibaraki 305-0032, Japan
[3)] *Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University*, Uzurano1348, Kasai, Hyogo 675-2103, Japan
[4)] *Experimental Farm, National Institute of Genetics*, Mishima, Shizuoka 411-8540, Japan
[5)] *Department of Life Science, Graduate University for Advanced Studies (SOKENDAI)*, Mishima, Shizuoka 411-8540, Japan

The wild *Oryza* species are rich in genetic diversity and are good resources for modern breeding of rice varieties. The reliable *ex situ* conservation of various genetic resources supports both basic and applied rice research. For this purpose, we developed PCR-based and co-dominant insertion/deletion (INDEL) markers which enable the discrimination of the genome types or species in the genus *Oryza*. First, 12,107 INDEL candidate sequences were found in the BAC end sequences for 12 *Oryza* species available in public databases. Next, we designed PCR primers for INDEL-flanking sequences to match the characteristics of each INDEL, based on an assessment of their likelihood to give rise to a single or few PCR products in all 102 wild accessions, covering most *Oryza* genome types. Then, we selected 22 INDEL markers to discriminate all genome types in the genus *Oryza*. A phylogenetic tree of 102 wild accessions and two cultivars according to amplicon polymorphisms for the 22 INDEL markers corresponded well to those in previous studies, indicating that the INDEL markers developed in this study were a useful tool to improve the reliability of identification of wild *Oryza* species in the germplasm stocks.

Key Words: insertion/deletion marker, genome type, genus *Oryza*, rice, wild species, genetic resource.

Introduction

Rice (*Oryza sativa* L.) is one of the most important crops in many countries and so many breeders have made extraordinary efforts to improve yield, grain quality and other agronomical traits in cultivated rice. The completion of the rice genome sequence has promoted molecular breeding and contributed to shorten of the periods for development of new varieties. On the other hand, large scale and long-term cultivation of a few varieties have resulted in a genetic breakdown of modern varieties, for example, in terms of a resistance to biotic stresses (Kottapalli *et al.* 2010). Wild species are known to be a source of useful genes for potential use in modern rice breeding.

The genus *Oryza* is composed of 23 species, two cultivated and 21 wild (reviewed by Vaughan and Morishima 2003). On the other hand, regarding chromosome structure, *Oryza* species are classified into nine genomic types, A to J (I is absent), according to the pairing affinity of meiotic chromosomes in hybrid plants. AA wild species are closely related to cultivated rice, *O. sativa* and have been useful resources for current breeding programs (Jena 2010) because of high cross compatibility with cultivars. Non-AA species are distantly related to cultivars and seldom used for breeding, due to lower affinity of the meiotic chromosomes and consequently seed sterility of hybrids with cultivated rice. However, non-AA species contain many agronomically important genes (Jena 2010, Nonomura *et al.* 2010), so availability of genetically reliable populations of *Oryza* species would be beneficial in rice breeding.

The *ex situ* conservation of wild species are always faced with problems arising from outcrossing of species or ecotypes, mixed collections of different species in natural habitats and artificial contamination during conservation. The great efforts by many taxonomists have solved these problems and enabled us to classify and conserve the *Oryza* species reliably. However, it is still difficult for most researchers to classify the *Oryza* species based on morphological and physiological traits.

In this study, we aimed to develop new molecular markers for even inexperienced researchers to discriminate species or genome types of wild *Oryza* accessions easily by

Communicated by Duncan Vaughan
Received January 7, 2013. Accepted April 12, 2013.
*Corresponding author (e-mail: knonomur@nig.ac.jp)

polymerase chain reaction (PCR) and gel electrophoresis. Single nucleotide polymorphisms (SNPs) and insertion/deletion (INDEL) markers are the most commonly used markers in plants, because they are easy to use, PCR-based, co-dominant and relatively abundant (Pacurar 2012). Simple sequence repeat (SSR) markers are also useful, but being rapidly replaced with SNP markers, because of the characteristics of SNPs more stable and amenable to automation (reviewed by McCouch *et al.* 2010). Numerous SNP markers have been developed and used for genomic selection, genomic association and quantitative-trait-loci mapping for inbreeding populations of modern varieties with local varieties or closely-related wild species. In this study, we developed INDEL markers, because they have a merit in easy detection of polymorphims by PCR and direct gel electrophoresis, while SNPs information might be converted into CAPS (cleaved amplified polymorphic sequences) or dCAPS (derived CAPS) and restriction endonuclease cleavage is necessary for SNP detection prior to gel electrophoresis (Michaels and Amasino 1998, Neff *et al.* 1998). Here, we show that 22 INDEL markers are available to discriminate species or genome types reliably in the genus *Oryza*.

Materials and Methods

Plant materials

All the wild accessions of the genus *Oryza* were provided by the National Institute of Genetics (NIG) and the National Bioresource Project (NBRP), Japan (Nonomura *et al.* 2010) and are listed in Table 1. Of 282 wild accessions of the NIG core collection, 42 of rank 1 and 60 of rank 2 were used. The accessions were selected to cover 20 of the 21 wild *Oryza* species (Fig. 1) and four or more accessions were analyzed for each of most species. In addition, *O. sativa* cv. Nipponbare and cv. Kasalath were used as standard varieties representing ssp. *japonica* and *indica*, respectively. All plants were grown in the summer on 2005 and 2006 in a field of the NIG, Mishima, Japan. Genomic DNA was extracted from mature leaves by the CTAB method (Rogers and Bendich 1988).

Identification of INDEL markers

The bacterial artificial chromosome (BAC)-end sequences (BESs) released by the *Oryza* Map Alignment Project (OMAP) (Ammiraju *et al.* 2006, Wing *et al.* 2005) and available on public databases were used to design PCR primers. The OMAP BESs derived from 12 wild *Oryza* species (Table 2) were mapped *in silico* to the reference *O. sativa* cv. Nipponbare genome (IRGSP Build 5) (http://rgp.dna.affrc.go.jp/E/IRGSP/Build5/build5.html) by a BLAST search (Altschul *et al.* 1997). In this step, we selected the BESs that displayed high similarity to the reference sequences except for INDEL regions. For further selection, the gap size was set in the range from 51 to 2,000 base pairs (bp). The primer pair to amplify each of the BESs selected above was designed by Primer3 software (http://primer3.sourceforge.net). The sequences of the reference genome and all primer pairs were put into e-PCR software (http://www.ncbi.nlm.nih.gov/projects/e-pcr/) and the primer pairs virtually amplifying a single band were selected. Each of the pairs was further screened for giving rise to a single band against its original BES in e-PCR software. For further limitation of the number of candidates, we chose primer pairs that amplified PCR products 50 to 1,500 bp long. In addition, the INDEL size was set at 10% or more of each PCR product for easy detection of the polymorphism using agarose gel electrophoresis.

Amplicon polymorphism assay

PCR was performed under conditions of 94°C for 2 min and a subsequent 35 rounds of 94°C for 1 min, 56°C for 1 min and 72°C for 1 min, followed by 72°C for 2 min, using GoTaq Green Master Mix kit (Promega) and a T1 Thermocycler (Biometra). Amplicon polymorphisms of the INDEL markers were identified by electrophoresis in 3% gels of Certified low range ultra agarose (Bio-Rad) followed by ethidium bromide staining.

To detect minute differences in electrophoretic distances more precisely, we also performed fragment analysis of the PCR products using the fluorescent-labeling fragment analyzing system of a capillary sequencer (Applied Biosystems (ABI) PRISM 3130xl) in accordance with the instructions for microsatellite analysis by ABI. ABI PRISM fluorescent primers, fluorescently 5′-labeled with PET, 6-FAM, NED, or VIC dyes, were used for forward INDEL primers. To avoid the "additional A" problem, in which unexpected addition of an adenine residue at the tail of PCR products often makes results unstable, seven nucleotides were artificially added to the 5′ end of reverse primers as recommended by ABI. The PCR product lengths were analyzed using a 3130xl Genetic Analyzer and GeneMapper software (ABI).

DNA sequencing

PCR products amplified with naked INDEL primer pairs were directly sequenced using a BigDye Terminator v3.1 Cycle Sequencing Kit (ABI) according to the manufacturer's instructions. Some products were cloned into vector pCRII (Invitrogen), then amplified and sequenced with universal M13 forward and reverse primers.

Phylogenetic analysis

A rooted phylogenetic tree was drawn by the UPGMA method (Sokal and Michener 1958) on the basis of the PCR product lengths of INDEL markers determined manually following gel electrophoresis. Data from the fragment analysis were used only to judge the size of DNA fragments showing similar, but slightly different mobilities in the gel. The tree construction was based on Nei's chord distance (Nei *et al.* 1983) and performed using Populations version 1.2.31 software (http://bioinformatics.org/~tryphon/populations/) with 1,000 bootstraps.

Table 1. The accessions of wild *Oryza* species used in this study

Acc No.	Species	Genome type	Origin	Acc No.	Species	Genome type	Distribution
W0106	*rufipogon*	AA	India	W1361	*officinalis*	CC	Malaysia
W0120	*rufipogon*	AA	India	W1830	*officinalis*	CC	unknown
W1294	*rufipogon*	AA	Philippines	W1131	*officinalis*	CC	India
W1866	*rufipogon*	AA	Thailand	W1302	*officinalis*	CC	Philippines
W1921	*rufipogon*	AA	Thailand	W1814	*officinalis*	CC	Sri Lanka
W2003	*rufipogon*	AA	India	W1805	*rhizomatis*	CC	Sri Lanka
W0630	*rufipogon*	AA	Myanmar	W1527	*eichingeri*	CC	Uganda
W1236	*rufipogon*	AA	unknown	W1519	*eichingeri*	CC	Uganda
W1807	*rufipogon*	AA	Sri Lanka	W1522	*eichingeri*	CC	Uganda
W1945	*rufipogon*	AA	unknown	W1525	*eichingeri*	CC	Uganda
W2051	*rufipogon*	AA	unknown	W1166	*latifolia*	CCDD	Mexico
W2078	*rufipogon*	AA	Australia	W1197	*latifolia*	CCDD	Colombia
W2263	*rufipogon*	AA	Cambodia	W2200	*latifolia*	CCDD	Brazil
W0652	*barthii*	AA	Sierra Leone	W0019	*latifolia*	CCDD	Cuba
W1588	*barthii*	AA	Cameroun	W1181	*latifolia*	CCDD	Panama
W0698	*barthii*	AA	Guinea	W0017	*alta*	CCDD	Surinam
W0720	*barthii*	AA	Mali	W1182	*alta/latifolia*	CCDD	Guyana
W0747	*barthii*	AA	Mali	W0018	*alta*	CCDD	Paraguay
W1646	*barthii*	AA	Tanzania	W1147	*alta*	CCDD	Surinam
W1169	*glumaepatula*	AA	Cuba	W0613	*grandiglumis*	CCDD	Brazil
W2145	*glumaepatula*	AA	Brazil	W1194	*grandiglumis*	CCDD	Brazil
W2199	*glumaepatula*	AA	Brazil	W2220	*grandiglumis*	CCDD	Brazil
W1185	*glumaepatula*	AA	Suriname	W1476	*grandiglumis*	CCDD	Brazil
W1187	*glumaepatula*	AA	Brazil	W1480(B)	*grandiglumis*	CCDD	Brazil
W1196	*glumaepatula*	AA	Colombia	W0008	*australiensis*	EE	Australia
W1625	*meridionalis*	AA	Australia	W1628	*australiensis*	EE	Australia
W1635	*meridionalis*	AA	Australia	W1632	*australiensis*	EE	Australia
W1297	*meridionalis*	AA	Australia	W2086	*australiensis*	EE	Australia
W1627	*meridionalis*	AA	Australia	W2104	*australiensis*	EE	Australia
W2069	*meridionalis*	AA	Australia	W1401	*brachyantha*	FF	Sierra Leone
W2079	*meridionalis*	AA	Australia	W1711	*brachyantha*	FF	Cameroun
W2103	*meridionalis*	AA	Australia	W1407(B)	*brachyantha*	FF	Mali
W1413	*longistaminata*	AA	Sierra Leone	W1706	*brachyantha*	FF	Chad
W1508	*longistaminata*	AA	Madagascar	W0003	*granulata*	GG	India
W0643	*longistaminata*	AA	Gambia	W0067(B)	*granulata*	GG	Thailand
W0708	*longistaminata*	AA	Guinea	W0005	*granulata*	GG	Ceylon
W1540	*longistaminata*	AA	Congo	W0615	*granulata*	GG	Myanmar
W1624	*longistaminata*	AA	Cameroun	W1356	*meyeriana*	GG	Malaysia
W1514	*punctata(2x)*	BB	Kenya	W1348	*meyeriana*	GG	Malaysia
W1590	*punctata(2x)*	BB	Cameroun	W1352	*meyeriana*	GG	Malaysia
W1024	*punctata(4x)*	BBCC	Ghana	W1354	*meyeriana*	GG	Malaysia
W1408	*punctata(4x)*	BBCC	Nigeria	W1360	*meyeriana*	GG	Malaysia
W1474(B)	*punctata(4x)*	BBCC	Chad	W2068	*meyeriana*	GG	unknown
W1213	*minuta*	BBCC	Philippines	W0001	*ridleyi*	HHJJ	Thailand
W1331	*minuta*	BBCC	Philippines	W0604	*ridleyi*	HHJJ	Malaya
W0016	*minuta*	BBCC	Surinam	W2033	*ridleyi*	HHJJ	Thailand
W1319	*minuta*	BBCC	Philippines	W2035	*ridleyi*	HHJJ	Thailand
W1323	*minuta*	BBCC	Philippines	W1220	*longiglumis*	HHJJ	Indonesia
W1328	*minuta*	BBCC	Philippines	W1215	*longiglumis*	HHJJ	Indonesia
W1329	*minuta*	BBCC	Philippines	W1224	*longiglumis*	HHJJ	Indonesia
W0002	*officinalis*	CC	Thailand	W1228	*longiglumis*	HHJJ	Indonesia

Flow cytometry

In addition to wild species, we used a haploid plant, produced from *O. sativa* cv. Nipponbare by the slightly modified method of the anther culture described by Niizeki and Oono (1968), to normalize the nuclear DNA contents of wild species. The preparation of samples and flow cytometric analysis was performed in accordance with Miyabayashi *et al.* (2007) with slight modifications. The sample nuclei were extracted from a piece of adult leaves (5 × 5 mm). The leaves were chopped with a razor blade in 400 μL of extraction

Marker ID	wild *Oryza* species																			
	AA					BB	BBCC		CC			CCDD			EE	FF	GG		HHJJ	
	ruf	bar	glu	mer	lon	pun	pun	min	off	eic	rhi	lat	alt	gra	aus	bra	gta	mey	rid	log
Ch01-301W	352			375		410													296/410	
Ch02-269W	259			285		124											115		120	
Ch02-308W	112			160		176														
Ch02-342W	209															405				
Ch02-343W	380			210, 370	135* 210*	135											140		120/136	
Ch03-128W	298					280											420		222/298	
Ch03-173W	369			130											137		153		135	
Ch03-363W	461								550							363			430/461	
Ch04-276G	132															150	360		150	
Ch04-312W	178				200							160*	160*		160	396				
Ch05-067W	195					145													135	
Ch05-070W	263											371*	263* 371*	371	365	310	320		336	
Ch05-109G	398					180									500		122			
Ch05-202W	148															130	65			
Ch05-277W	326																		362	
Ch06-269W	299			600		357														
Ch06-300W	180															210	360		180*	
Ch06-306W	124						185*		182							140	120		145	150
Ch07-233W	281			440		200			300	300*		281*	281*						181/350	
Ch08-006W	234				270											372	245		410	
Ch09-037G	161															140	363		330	
Ch10-044G	121															150	553			

Fig. 1. A diagram of the polymorphic PCR band pattern of 20 wild *Oryza* species obtained using 22 INDEL markers. The color pattern indicates how many polymorphic bands were detected in each of 22 INDEL markers for 20 wild *Oryza* species. The species indicated with a same color exhibited an identical band size for each marker on the gel electrophoresis. The number in each color indicates the estimated band size (bp). For example, the Ch7-233W marker gave rise to 5 different bands: 281 bp (magenta), 440 bp (yellow), 200 bp (green), 300 bp (light blue) and 181/350 bp (dark blue). In BBCC and CCDD tetraploid species, this marker gave two bands from both BB and CC types, colored green and light blue. Two numbers with a slash (ex., 181/350) indicate that in tetraploid species (for example, HHJJ), the genome types from which both bands were raised are uncertain. Two numbers with a comma (ex., 210, 370) indicate that the marker amplifies two PCR bands in a diploid species. The numbers with asterisks indicate that those bands are polymorphic in several accessions within the same species. ruf, *rufipogon*; bar, *barthii*; glu, *glumaepatula*; mer, *meridionalis*; lon, *longistaminata*; pun, *punctata*; min, *minuta*; off, *officinalis*; eic, *eichingeri*; rhi, *rhizomatis*; lat, *latifolia*; alt, *alta*; gra, *grandiglumis*; aus, *australiensis*; bra, *brachyantha*; gta, *granulata*; mey, *meyeriana*; rid, *ridleyi*; log, *longiglumis.* The seven marker IDs colored red are a minimum marker set useful for classifying species and genome species in the genus *Oryza*.

solution A (Partec) and incubated for 30 min. The supernatant was filtered with 50 µm- and subsequently with 20 µm-CellTrics filters. 160 µL of staining solution B (Partec) was added to the supernatant and incubated for 30 min. The extract from the wild species sample was mixed with an equal volume of that from the haploid plant and supplied for the measurement of nuclear DNA contents by ploidy analyzer PA system (Partec) according to the manufacturer's instruction.

Online disclosure of marker information

Information on the 22 INDEL markers obtained in this study is open access on the integrated rice science database, Oryzabase (Yamazaki *et al.* 2010, http://www.shigen.nig.ac.jp/rice/oryzabaseV4/).

Results

Selection of INDEL markers to discriminate wild Oryza species

To design PCR primers for discriminating wild relatives of rice, BESs derived from twelve wild species, including *O. nivara* and *O. coarctata* in accordance with the taxonomy in Ammiraju *et al.* (2006), were selected from public databases and compared to the reference *O. sativa* cv. Nipponbare genome. This permitted the identification of

Table 2. The number of putative INDELs detected *in silico* in this study

Oryza species*	Genome type	No. INDELs candidate sites in this study	No. BESs used in this study
rufipogon	AA	439	70,982
nivara	AA	462	106,124
glaberrima	AA	556	66,821
punctata	BB	1,270	68,384
minuta	BBCC	1,911	169,460
officinalis	CC	1,504	101,091
alta	CCDD	1,370	128,732
australiensis	EE	705	135,769
brachyantha	FF	550	67,364
granulata	GG	538	138,171
ridleyi	HHJJ	831	204,729
coarctata	HHKK	1,971	195,285
		12,107	1,452,912

* The taxonomy is in accordance with Ammiraju *et al.* (2006)

12,107 putative INDELs in the genomes of wild species compared with cultivated rice (Table 2). 3,244 INDEL loci were shared by 3 to 10 species each (referred to hereafter as multi-INDELs) (data not shown). The multi-INDELs were further reduced to 120 loci based on the degree of conservation of INDEL-flanking sequences and examination of the lengths of PCR products (see Methods). In addition, we selected 104 INDELs, each of which was detected with reference to a single BES from *O. rufipogon*, *O. punctata*, *O. brachyantha*, or *O. granulata* (single INDEL). In total, 224 markers were investigated for their applicability in classification of wild rice relatives. A total of 40 rice microsatellite (RM) markers (McCouch *et al.* 2002) were also investigated. However, all of the RM markers gave intraspecifically polymorphic PCR fragments only in AA-genome species (data not shown); therefore, they were excluded from subsequent analyses.

A total of 22 wild accessions representing the 20 *Oryza* species in addition to two rice varieties, *O. sativa* ssp. *japonica* cv. Nipponbare and ssp. *indica* cv. Kasalath, were used for the first assessment of INDEL markers. We screened the 224 primer sets, focusing on their ability to amplify PCR products in all genome types, especially in distantly related EE, FF, GG or HHJJ species, since this study aimed to establish PCR-based markers applicable for all *Oryza* species. In this screening, 60 of 120 multi- and 13 of 104 single INDELs were selected and further analyzed (Supplemental Fig. 1). More efficient amplification in distantly related species by multi-INDELs than by single INDELs suggested that INDEL-flanking sequences are diverse among species, and that careful mining of flanking sequences conserved across multiple species is important for efficient selection of INDEL markers. In the second assessment, 22 out of 73 INDELs that gave clear PCR bands in 102 wild accessions were selected (Table 3 and Supplemental Fig. 1) and further analyzed. The 22 INDEL loci were widely distributed on all Nipponbare chromosomes except for chromosomes 11 and

Table 3. The 22 INDEL markers developed to discriminate wild *Oryza* species in this study

Marker ID*	Primer sequences		INDEL size (bp)	BES ID
	Forward	Reverse		
Ch01-301W	TTTGTTCATCTGCATCAACTCA	TGATCATACGATGGAAAGGTAGA	56	OR_ABa0229J12.r
Ch02-269W	TTCGGTAAAGAACCTCTTGAGTG	GATTCTAGTGCCATTTCGCC	135	OP__Ba0011A10.r
Ch02-308W	CCTTAAGAAATTGTTAGTTCAGGCA	CTTCTTCTTGCTAGCAGTTGTCT	64	OP__Ba0008P06.f
Ch02-342W	TGGCAGACCATCTGAGAGAG	CCAGAAATCAGCAATCTGCAA	196	OB__Ba0075C24.f
Ch02-343W	TTCTCCACCCTCGTCTTCTC	TTGCTCTCCAGCTTCTCCTC	245	OB__Ba0060I13.r
Ch03-128W	TGATTCCTTGGTAGTCTTCCC	TGCTCACCATAGACTCTTCCA	76	OR_ABa0265L24.r
Ch03-173W	AGGCAAAGTTCAGAATGCAA	AGCGGCAATAGCCATCTAAG	217	OG_ABa0011K22.r
Ch03-363W	TTTCCGTCAGATTGCACATT	TGATTTACCACCAAACAGTAAGTCA	98	OB__Ba0073N03.f
Ch04-276G	GGTACCTCCAGGAATCCCAT	CCAATGTGCATGGCATTTAG	229	OG_ABa0018K10.r
Ch04-312W	TTCTTTGTCGTGATCGCAAG	TTTCATTCAACGTGGTGGTT	218	OR_ABa0207K04.f
Ch05-067W	CCCATTCCCTATACCTGTGTAAA	AGAATCACAGAGGATCCGAA	60	OR_ABa0151A19.r
Ch05-070W	GGAAGAAAGCAAGGATGCAA	TCTGCTGTCATATGCTTGGG	108	OA_BBa0041D18.f
Ch05-109G	TGATGATGAAATACCTTGCCC	TGTATGGCTGCATTTGCACT	276	OG_ABa0128B15.f
Ch05-202W	TCTTCAAGAAACCAGAAGATCTGA	TGGATGTGCTTCTGACGCTA	83	OG_ABa0074B07.r
Ch05-277W	CCAGAACCGTTGTTCCTGTT	GGATGTTGAGAAGGGTGGAA	64	OR_ABa0019H23.r
Ch06-269W	CCAATGAAATGCAGTCGAGA	GGACACATTCAACCCTCACA	79	OC__Ba0241E15.f
Ch06-300W	CACAGACAGTGTCCAGAGTTCAG	GCCTTGAAAGTTGAAACCCA	180	OG_ABa0049K22.f
Ch06-306W	GAGCCCTCGGTTAGATGTGA	CGTGCCGTATATGTCCTGAA	61	OM__Ba0207M10.f
Ch07-233W	TCACCAAGCCAATTCTTCTTC	CCTCCTAAACCAGACTGCACA	100	OR_ABa0269N20.f
Ch08-006W	TTGCATTGAATCAGTAGGTCAA	AGGACCTTGATTTGCCATGT	138	OB__Ba0054E21.f
Ch09-037G	CCGGAGTCCTATCCACAGGT	TTGGGCATCACCTGATAAGA	202	OG_ABa0001G03.f
Ch10-044G	CTTGTTTCCAGCAAGGTTGG	TACCAGGTTTGCTGCATGTT	432	OG_ABa0068J15.r

* The suffixes W and G indicate that the markers are multi- and single (*granulata*)-INDELs, respectively.

12 and eighteen of them were multi-INDELs (suffixed "W" in Table 3). The size of the 22 INDELs ranged from 56 to 432 bp and averaged 150.8 bp (Table 3).

Validation of 22 INDEL markers in discrimination of wild Oryza species

Each of the 22 selected INDEL markers exhibited various interspecific amplicon polymorphisms in the genus *Oryza* (Fig. 1). To validate the reliability of these markers for discriminating *Oryza* species, they were applied to phylogenetic analysis of 102 wild accessions and two cultivars (Fig. 2). The band sizes of PCR products were manually determined following agarose gel electrophoresis. In addition, for detecting minute differences in electrophoretic distances more precisely, we attempted fragment analysis. However, it was difficult to completely replace the manual method by fragment analysis, because the size of amplicons varied even among wild accessions from the same species. This may be due to numerous inter- and intraspecific SNPs and small INDELs in the sequences flanking the targeted INDEL being responsible for minute differences in electrophoretic distances. Thus, we decided to evaluate PCR band sizes mainly manually, using fragment analysis to support the manual method. All raw data obtained by either method are compiled in Supplemental Table 1. Furthermore, DNA sequencing confirmed that each of the 22 INDEL primer sets enabled amplifying PCR products from a syntenic locus shared by multiple genome types of *Oryza* species (an example of sequence alignments is shown in Supplemental Fig. 2), indicating the reliability of these markers. Only one of the 22 INDELs was derived from an exonic region (Ch04-276G), four were from intergenic regions (Ch04-312W, Ch05-109G, Ch05-202W and Ch06-300W) and all remaining INDELs were from intronic regions (data not shown).

Phylogenetic analysis

Phylogenetic analysis using the 22 INDEL markers selected above successfully divided 102 wild accessions into 9 genome groups (Fig. 2). Here, the taxonomy of the genus *Oryza* depended on Vaughan and Morishima (2003). Thirty-eight accessions of AA-genome species were divided into 3 phylogenetic groups: the group including *rufipogon*, *barthii* and *glumaepatula*, the *meridionalis* group and the *longistaminata* group. Thirty-seven accessions of BB, CC, BBCC and CCDD genome species, the "*officinalis* complex" (see below), were divided into five independent but closely related groups (Fig. 2). Fourteen accessions of *O. latifolia*, *O. alta* and *O. grandiglumis*, all classified into the CCDD tetraploids, were divided into 4 closely-related groups (Fig. 2). The accessions of *latifolia* and/or *alta* segregated into all 4 groups. All *O. grandiglumis* accessions were in a single group with two accessions of *alta* or *latifolia* (Fig. 2). The *officinalis* complex groups was first connected with EE-genome species and then with the AA-genome in the phylogenetic tree (Fig. 2). In contrast, GG, FF and HHJJ taxa were distant to AA taxa (Fig. 2).

A minimum set of INDELs to discriminate 9 genome types

This study revealed that seven INDEL markers (colored red in Fig. 1) were sufficient to discriminate 9 genome types and several species in the genus *Oryza*.

Four INDELs enabled classification of wild species into each genome group by PCR and agarose gel electrophoresis (Fig. 1 and Supplemental Fig. 1). First, Ch06-306W gave PCR bands characteristic of the CC, FF, GG and HHJJ genome groups. HHJJ species were separable into *O. ridley* and *O. longiglumis*, allowing consequent determination of *O. brachyantha*, since the FF genome is composed of a single species. Second, Ch07-233W characterized diploid *O. punctata* (BB), diploid CC genome species and tetraploid BBCC species. This marker also discriminates *O. meridionalis* from other AA species. Third, Ch04-312W clearly discriminated the EE species, *O. australiensis* and CCDD species from other species. *O. longistaminata* was discriminated from other AA species by this marker. Additional use of Ch05-070W enabled *O. grandiglumis* to be distinguished from other CCDD species.

Validation by fragment analysis further enabled several genome groups and species to be distinguished (Supplemental Table 1). Ch02-343W made *O. longistaminata* clearly separable from other AA species, and Ch06-300W and Ch10-044G distinguished tetraploid *O. punctata* from *O. minuta* (both BBCC).

Discussion

In this study, we developed the 22 INDEL markers from the public information of over a million of BESs by means of *in silico* selection and subsequent PCR-based assessments. They are co-dominant markers reliably and rapidly to discriminate all genome types and several species in the genus *Oryza*.

Phylogenetic analysis of 102 wild accessions also confirmed the reliability of the markers. In this study, *O. rufipogon*, *O. barthii* and *O. glumaepatula* were unable to be separated (Fig. 2). This result is comparable to those of previous studies, in which the AA-genome species lack clear morphological characteristics, except for *O. longistaminata* (reviewed in Vaughan and Morishima (2003)). A sole grouping of *O. meridionalis* accessions is consistent with the results obtained using restriction fragment length polymorphisms (RFLPs) and short interspersed elements (SINEs) (Doi *et al.* 1995, Wang *et al.* 1992, Xu *et al.* 2005). The *longistaminata* accessions displayed intraspecific amplicon polymorphisms for an INDEL marker (asterisked in Fig. 1), probably representing genetic diversity of this species previously reported (Kiambi *et al.* 2005, Oka 1988).

BB, CC, BBCC and CCDD genome species are distributed widely in Asia, Australia, Central and South America and Africa and have been called the "*officinalis* complex" because of diverse but relatively similar morphology (Tateoka 1962). RFLP analysis in a previous study revealed that these species are closely related (Wang *et al.* 1992), corresponding

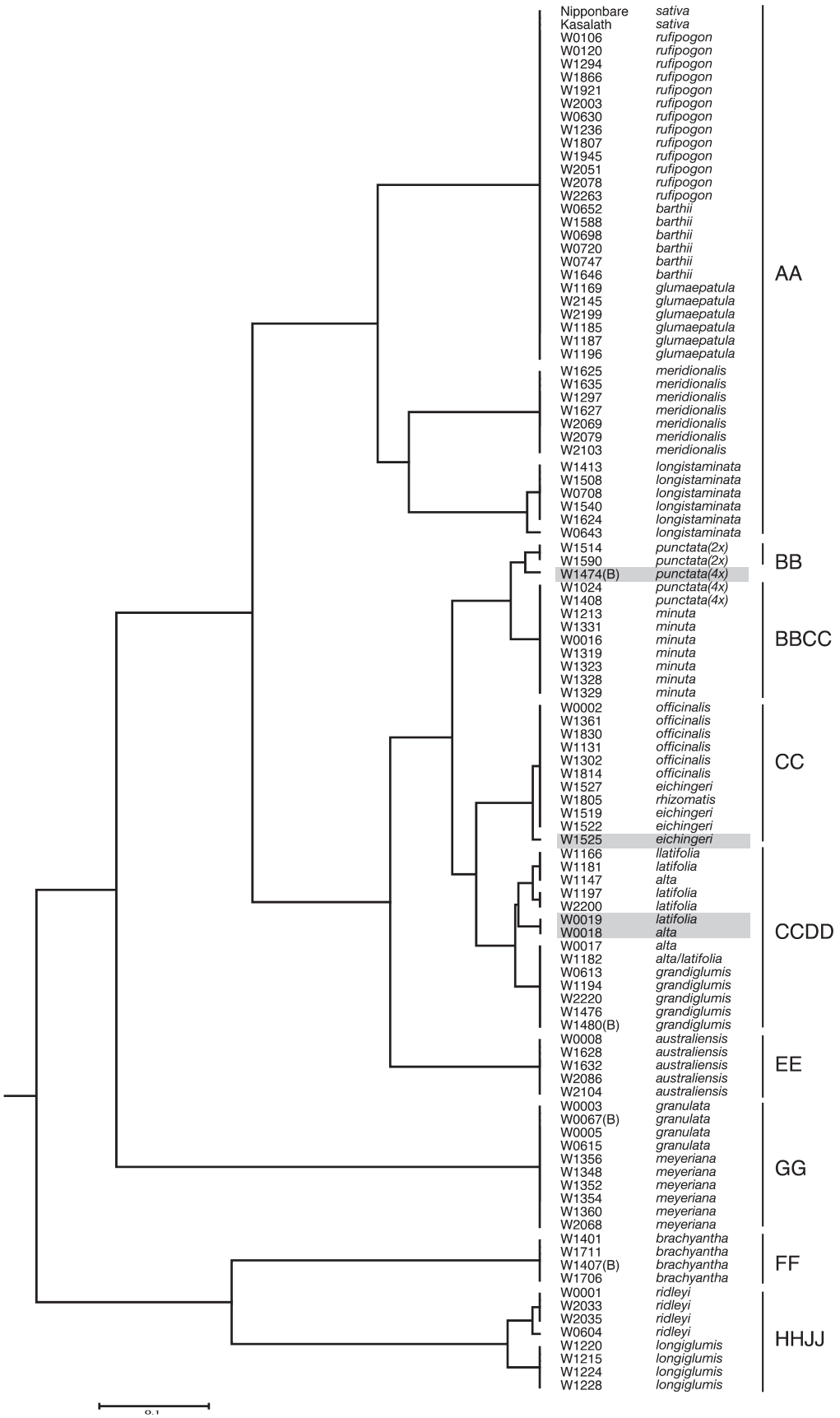

Fig. 2. Phylogenetic tree of 102 wild *Oryza* accessions and two cultivars obtained using 22 INDEL markers. This phylogenetic tree was depicted by the UPGMA method (Sokal and Michener 1958), based on data shown in Supplemental Table 1.

to the results of this study. Three CCDD tetraploids, the *O. latifolia*, *O. alta* and *O. grandiglumis*, shared almost INDEL patterns with each other, resulting in all these accessions being in 4 closely-related phylogenetic groups (Fig. 2). This result seems to be consistent to a previous report that these three species are closely related in terms of taxonomy; *latifolia* and *alta* were distinguished from each other only by the size of spikelet, and that *grandiglumis* was distinguished from *latifolia* and *alta* only by large glumes (Morishima and Martin 1994). EE genome species were closer to the *officinalis* complex species than GG, FF and HHJJ species, consistent with a previous proposal in which the EE genome is closely related to the DD genome progenitor (Ge *et al.* 1999, Wang *et al.* 1992). The phylogenetic results in this study largely corresponded to those in the previous studies, indicating the reliability of the INDEL markers in identification of wild species of rice.

Phylogenetic analysis in this study also revealed possibility of several mistakes in registration of wild accessions conserved in the NIG collection. W1525 was classified as diploid *O. eichingeri* (CC) and W0018 and W0019 as tetraploid *O. alta* and *O. latifolia* (CCDD), respectively. However, the pattern of PCR amplification by three INDEL markers, Ch07-233W, Ch03-128W and Ch04-312W (Supplemental Fig. 1), suggests the possibility that these three accessions should be classified as tetraploid BBCC species. In addition, W1474, deposited as tetraploid *O. punctata* (BBCC), exhibited the INDEL pattern of diploid *O. punctata* (BB) (Supplemental Fig. 1). These problem accessions were involved in the clades distinct from the expected clades in the phylogenetic tree (shaded in Fig. 2). In case of W1525 and W1474, we extended the number of plants examined and confirmed by INDEL marker-assisted and flow-cytometric analyses that each of these accessions actually segregated into two different ploidies (Supplemental Fig. 3). These results clearly demonstrate the reliability of the INDEL markers developed in this study. In the current database, W1805 from Sri Lanka is deposited as *O. rhizomatis*, but was previously deposited as *O. eichingeri* (http://www.shigen.nig.ac.jp/rice/oryzabaseV4/strain/wildCore/detail/36). However, the haplotype of the 22 INDELs in this accession is identical to that in African *O. eichingeri* accessions (Fig. 2). These problematic accessions need to be reclassified more carefully.

Genetic resources are the common heritage of humankind. However, some natural populations of *Oryza* species have been lost from their original habitats, mainly due to drastic environmental changes and human activities (Akimoto *et al.* 1996, Nonomura *et al.* 2010). *In situ* and *ex situ* conservation strategies have become important globaly. Wild *Oryza* species are quite diverse genetically and physiologically and sometimes form mixed populations of different species in their original habitats. Especially for *ex situ* conservation, it is critical to pay attention to maintaining the genetic reliability of strains by excluding any contamination from different species. The 22 PCR-based and co-dominant INDEL markers developed in this study will be powerful tools to help determinate species identity and genome types easily and to establish germplasm stocks with corroborative genetic information useful for experimental studies and rice breeding.

Acknowledgments

We thank Yurie Tamaki and Akemi Ishii for assistance in INDEL evaluation and Sonoko Saeki (NIG) for growing rice plants. We also thank Yukiko Yamazaki, Rie Tsuchiya and Shingo Sakaniwa (NIG) for updating INDEL marker information on Oryzabase. This work was funded by a grant from the National Bioresource Project (NBRP) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

Literature Cited

Akimoto, M., Y. Shimamoto and H. Morishima (1996) A case study of genetic erosion of wild rice in Thailand. Rice Genet. Newsl. 13: 60–62.

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.

Ammiraju, J.S., M. Luo, J.L. Goicoechea, W. Wang, D. Kudrna, C. Mueller, J. Talag, H. Kim, N.B. Sisneros, B. Blackmon *et al.* (2006) The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. Genome Res. 16: 140–147.

Doi, K., A. Yoshimura, M. Nakano, N. Iwata and D.A. Vaughan (1995) Polygenetic study of AA genome species of genus *Oryza* using nuclear RFLP. Rice Genet. Newsl. 12: 160–162.

Ge, S., T. Sang, B.R. Lu and D.Y. Hong (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc. Natl. Acad. Sci. USA 96: 14400–14405.

Jena, K.K. (2010) The species of the genus *Oryza* and transfer of useful genes from wild species into cultivated rice, *O. sativa*. Breed. Sci. 60: 518–523.

Kiambi, D.K., H.J. Newbury, B.V. Ford-Lloyd and I. Dawson (2005) Contrasting genetic diversity among *Oryza longistaminata* (A. Chev et Roehr) populations from different geographic origins using AFLP. African J. Biotech. 4: 308–317.

Kottapalli, K.R., M. Lakshmi Narasu and K.K. Jena (2010) Effective strategy for pyramiding three bacterial blight resistance genes into fine grain rice cultivar, Samba Mahsuri, using sequence tagged site markers. Biotechnol. Lett. 32: 989–996.

McCouch, S.R., L. Teytelman, Y. Xu, K.B. Lobos, K. Clare, M. Walton, B. Fu, R. Maghirang, Z. Li, Y. Xing *et al.* (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res. 9: 257–279.

McCouch, S.R., K. Zhao, M. Wright, C.W. Tung, K. Ebana, M. Thomson, A. Reynolds, D. Wang, G. DeClerck, M.L. Ali *et al.* (2010) Development of genome-wide SNP assays for rice. Breed. Sci. 60: 524–535.

Michaels, S.D. and R.M. Amasino (1998) A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. Plant J. 14: 381–385.

Miyabayashi, T., K.I. Nonomura, H. Morishima and N. Kurata (2007) Genome size of twenty wild species of *Oryza* determined by flow cytometric and chromosome analyses. Breed. Sci. 57: 73–78.

Morishima, H. and P.S. Martin (1994) Investigations of plant genetic resources in the Amazon basin with the emphasis on the genus *Oryza*—Report of study-tour 1992/93. *In*: Morishima, H., Y. Shimamoto, Y. Sano and Y.I. Sato (eds.) Reports of the Study-tours for Investigation of Wild and Cultivated Rice Species. Part II., National Institute of Genetics, Japan, pp. 251–357. (The pdf is available on the website; http://www.shigen.nig.ac.jp/rice/oryzabaseV4/asset/ricereport/Part_II.pdf)

Neff, M.M., J.D. Neff, J. Chory and A.E. pepper (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana*. Plant J. 14: 387–392.

Nei, M., F. Tajima and Y. Tateno (1983) Accuracy of estimated phylogenetic trees from molecular data. J. Mol. Evol. 19: 153–170.

Niizeki, H. and K. Oono (1968) Induction of haploid rice plant from anther culture. Proc. Jpn. Acad. 44: 554–557.

Nonomura, K.I., H. Morishima, T. Miyabayashi, S. Yamaki, M. Eiguchi, T. Kubo and N. Kurata (2010) The wild *Oryza* collection in National BioResource Project (NBRP) of Japan: History, biodiversity and utility. Breed. Sci. 60: 502–508.

Oka, H.I. (1988) Origin of cultivated rice. J. Sci. Soc. Press/Elsevier, Tokyo.

Pacurar, D.I., M.L. Pacurar, N. Street, J.D. Bussell, T.I. Pop, L. Gutierrez and C. Bellini (2012) A collection of INDEL markers for map-based cloning in seven *Arabidopsis* accessions. J. Exp. Bot. 63: 2491–2501.

Rogers, S.O. and A.J. Bendich (1988) Plant molecular biology manual. Kluwer Academic Publishers, Dordrecht. pp. 1–10.

Sokal, R.R. and C.D. Michener (1958) A statistical method for evaluating systematic relationships. The Univ. Kansas Sci. Bullet. 38: 1409–1438.

Tateoka, T. (1962) Taxonomic studies of *Oryza* I. *O. latifolia* complex. Bot. Mag. Tokyo 75: 418–427.

Vaughan, D.A. and H. Morishima (2003) Biosystematics of the genus *Oryza*. *In*: Smith, C.W. (ed.) Rice: Origin, History, Technology, and Production, John Wiley & Sons, Inc., New York. pp. 27–65.

Wang, Z.Y., G. Second and D. Tanksley (1992) Polymorphism and phylogenetic relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. Theor. Appl. Genet. 83: 565–581.

Wing, R.A., J.S. Ammiraju, M. Luo, H. Kim, Y. Yu, D. Kudrna, J.L. Goicoechea, W. Wang, W. Nelson, K. Rao *et al.* (2005) The *Oryza* Map Alignment Project: the golden path to unlocking the genetic potential of wild rice species. Plant Mol. Biol. 59: 53–62.

Xu, J.H., N. Kurata, M. Akimoto, H. Ohtsubo and E. Ohtsubo (2005) Identification and characterization of Australian wild rice strains of *Oryza meridionalis* and *Oryza rufipogon* by SINE insertion polymorphism. Genes Genet. Syst. 80: 129–134.

Yamazaki, Y., S. Sakaniwa, R. Tsuchiya, K.I. Nonomura and N. Kurata (2010) Oryzabase: an integrated information resource for rice science. Breed. Sci. 60: 544–548.