New Phytologist

Research

Complex evolution of *S5*, a major reproductive barrier regulator, in the cultivated rice *Oryza sativa* and its wild relatives

Hongyi Du*, Yidan Ouyang*, Chengjun Zhang and Qifa Zhang

National Key Laboratory of Crop Genetic Improvement and National Centre of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

Author for correspondence:
Qifa Zhang
Tel: +86 27 87282429
Email: qifazh@mail.hzau.edu.cn

Received: *24 November 2010*
Accepted: *31 January 2011*

New Phytologist (2011) **191**: 275–287
doi: 10.1111/j.1469-8137.2011.03691.x

Key words: evolution, hybrid sterility, *Oryza sativa* (rice), reproductive isolation, speciation gene, wide compatibility.

Summary

- The hybrid sterility gene *S5* comprises three types of alleles in cultivated rice. Such tri-allelic system provided a unique opportunity to study the molecular bases of evolutionary changes underlying reproductive isolation in plants.
- We analysed the sequence diversity and evolutionary history of *S5* in 138 *Oryza* accessions. We also examined the effect of the two functional variations (C819A and C1412T) in determining hybrid sterility by transformation.
- Nineteen haplotypes were identified, which were classified into the *indica*-like, the *japonica*-like and the wide-compatibility gene (WCG)-like group, according to the sequence features of the tri-allelic system. The origin and evolutionary course of the three allelic groups were investigated, thus confirming the independent origins of *indica* and *japonica* subspecies. There were perfect associations between C819A and C1412T in the rice germplasm assayed, and the combination of C819 and C1412 was required for hybrid sterility. Evidence of positive selection in the WCG-like alleles suggested that they might have been favored by selection for higher compatibility in hybrids.
- The complex evolution of *S5* revealed the counteractive function of the three allelic groups at the species level. *S5* might perform an important primary function in an evolutionary scale, and hybrid sterility acts as a 'byproduct' of this speciation gene.

Introduction

The irreversible process of speciation has attracted the attention of biologists since the work of Darwin (Darwin, 1859). The population undergoing divergent evolution can ultimately result in reproductive isolation, which establishes and maintains the species to be genetically distinct (Coyne & Orr, 2004). Speciation genes are involved in the formation of new species or subspecies and are responsible for genetic incompatibilities in hybrids thus causing reproductive isolation. The evolutionary history of such special group of genes is generally accomplished by the origin and differentiation events between incipient species. In plants, hybrid sterility is the most common type of postzygotic reproductive isolation. The best known example is perhaps the hybrid sterility between *indica* and *japonica* subspecies of Asian cultivated rice *Oryza sativa* L. (Ouyang *et al.*, 2010).

Oryza sativa was domesticated in Asia from the wild progenitor *Oryza rufipogon* and/or *Oryza nivara* (Oka, 1988; Dally & Second, 1990; Ge *et al.*, 1999). Classical studies in the subpopulation structure of *O. sativa* have identified two primary subspecies, namely *indica* and *japonica* (Kato *et al.*, 1928; Oka, 1988; Zhang *et al.*, 1992, 1997). These two subspecies have been recorded as distinct rice groups in the literature in China since the Han Dynasty (> 2000 yr ago), and are referred to as *hsien* and *keng*, respectively (Ting, 1949a,b). The two subspecies differ markedly both in phenotypic adaptations and in molecular characteristics. Differentiation between *indica* and *japonica* has resulted in various forms of hybrid sterility including embryo sac abortion and pollen sterility (Ouyang *et al.*, 2009). A number of loci conferring hybrid male or female sterility (in a few cases, both) have been identified in rice (Ouyang *et al.*, 2009). Two genes, *S5* and *Sa*, conditioning female and male sterility, respectively, in *indica–japonica* hybrids were recently cloned and molecularly characterized (Chen *et al.*,

---

*These authors contributed equally to this work.

2008; Long *et al.*, 2008). Another study identified that epistatic interaction between two duplicated genes, *S27* and *S28*, can cause F_1 hybrid male sterility between the cultivated rice *O. sativa* and the wild rice *Oryza glumaepatula* (Yamagata *et al.*, 2010).

S5 functions in megaspore survival, encoding an aspartic protease with relatively high expression in ovule tissues. There are three alleles at the *S5* locus: an *indica* allele (*S5*-i), a *japonica* allele (*S5*-j), and a neutral allele (*S5*-n) also referred to as the wide-compatibility gene (WCG) (Ikehashi & Araki, 1986). The *S5*-i and *S5*-j differ by two nucleotides, both of which cause amino acid substitutions, while, the *S5*-n has a 115-aa deletion at the *N*-terminus (Chen *et al.*, 2008). Sterility occurs only when the plants have *S5*-i and *S5*-j alleles simultaneously, whereas plants carrying *S5*-n (referred to as wide-compatibility varieties (WCVs)) with either *S5*-i or *S5*-j would be fully fertile (Ikehashi & Araki, 1986; Yanagihara *et al.*, 1995; Chen *et al.*, 2008). It was inferred that the *S5*-i and *S5*-j alleles have acted as important promoting factors for the genetic differentiation between *indica* and *japonica* during the evolution, whereas *S5*-n, which enables hybridization, provides an opposing force for holding the differentiated groups together (Ouyang *et al.*, 2010). Thus, the coexistence of *S5*-i, *S5*-j and *S5*-n in rice provides an excellent model system for studying the evolutionary processes of reproductive isolation and speciation.

While both *indica* and the *japonica* subspecies of the cultivated rice have made great contributions to food production globally, there are still controversies over the origin and evolutionary history of the two rice groups. There are two competing hypotheses regarding the origin of the two subspecies. One hypothesis proposes that *japonica* was derived from *indica* (Chang, 1976; Oka, 1988), while the alternative hypothesis suggests independent origins of *indica* and *japonica* from their wild ancestors (Second, 1982; Bautista *et al.*, 2001; Cheng *et al.*, 2003). Although accumulating data seem to favor one hypothesis over the other (Kovach *et al.*, 2007; Sang & Ge, 2007; Sweeney & McCouch, 2007), all the studies are based on evidence from either archaeological analysis or genetic markers; none of the studies involved speciation genes. Speciation genes in rice could have been the primary causes of differentiation between *indica* and *japonica* subspecies. Therefore, the evolutionary history of the speciation gene *S5* provides a rare opportunity for understanding the genetic and molecular evidence regarding the origin and evolution of *indica* and *japonica* subspecies.

Complete understanding of the evolutionary mechanism of reproductive isolation requires answers to several questions. What are the evolutionary processes for establishing the subspecies and species accomplished with the origin of the speciation genes? Have the speciation genes changed their protein sequences at the causative mutation sites in ancient time? How do speciation genes function in determining the hybrid sterility? During the past decade, an increasing number of studies in speciation genes have focused on function, mechanism and molecular evolution, as well as their selective pressures in animals. However, few evolutionary studies concerning speciation genes have been reported in plants. In this study, we address these questions using the hybrid sterility gene *S5*, an interesting system for studying the molecular evolution of reproductive isolation and speciation in *O. sativa*. The results will provide unequivocal evidence of the evolutionary history of this hybrid sterility gene with direct implications for the mechanisms in reproductive isolation. The results will also help understanding the dynamic process of rice speciation.

Materials and Methods

Plant materials

All seeds or DNA used in this study were obtained from our own laboratory or provided by the International Rice Research Institute (IRRI, Los Banos, the Philippines), including 44 accessions of *O. sativa*, 40 accessions of *O. rufipogon* and 38 accessions of *O. nivara* (see the Supporting Information Table S1). Rice seeds for each accession were treated at 50–55°C for 5 d and 4°C for 3 d to break dormancy. The seeds were then germinated in the half-strength Murashige–Skoog medium (Murashige & Skoog, 1962) to obtain seedlings.

PCR amplification and DNA sequencing

DNA was extracted from fresh leaves according to Doyle & Doyle (1987). Primers (Table 1) that amplified the 4.7-kb fragment of *S5* region were designed according to *S5* alleles in cultivated rice (GenBank accession nos. EU889293 (*S5*-n), EU889294 (*S5*-j) and EU889295 (*S5*-i)). Two PCR systems were used with total DNA as the template. The 15 μl volume system contained 30 ng DNA template, together with: 0.2 μl of the forward and reverse primers (both 10 μM), 1.5 μl of 2 mM dNTP, 1.5 μl of 25 mM $MgCl_2$, 0.2 μl of 5 U μl^{-1} rTaq polymerase (TaKaRa Biotechnology, Dalian, China), and 1.5 μl 10× rTaq buffer. The 20 μl volume system contained 60 ng DNA template, together with: 0.2 μl of the forward and reverse primers (both 10 μM), 1.2 μl of 2 mM dNTP, 1 μl of 50% glycerol, 0.2 μl of 5 U μl^{-1} ExTaq polymerase (Takara Biotechnology), and 2 μl of 10× ExTaq buffer. All PCR amplifications were repeated three times independently on Gene AMP PCR system 9700 (Applied Biosystems, Carlsbad, CA), with the following profile: 4 min at 94°C for pre-denaturation, followed by 30 cycles of 1 min at 94°C, 1 min at 59°C, and 2 min at 72°C, and a final 7 min extension at 72°C.

Table 1 Primers and the expected amplicon size

Primer name	Sequence	Product size (bp)
F18-02	Forward 5′-AACGATGCTCATGCATGCTGAGGT	1157
R10-02	Reverse 5′-CCTCTGCTGCCTCTGTGTCTACGT	
F18-03	Forward 5′-GTGACGTCGAGATAAACCTTGGCA	1207
R10-03	Reverse 5′-TTCGGTCGCACAATGGACGCAACA	
F18	Forward 5′-TGTCAACGCCGCATGGTTCTGAGA	1241
R10	Reverse 5′-CAGGCAGTCAAACGTAGGAAAGGA	
F19	Forward 5′-TAATCGATCGGCCATTCCTTCCGA	1262
R11	Reverse 5′-ATGTGTAGGATCTGCCGGGATCGA	
F20	Forward 5′-GATCGAAGACAGCAGCATCAACGA	1226
R12	Reverse 5′-GAAACGAGGACATGCATGGACAGA	
F21	Forward 5′-TTGCTCAGAATCCTGCTCTCAGGT	1226
R13	Reverse 5′-ATTAATCTGGCGCCTAAGCTCGCA	
PS5HF	Forward 5′-TAAAAAGCTTCATCATGGGCTGCTGCTAGATGGA	2822
R11	Reverse 5′-ATGTGTAGGATCTGCCGGGATCGA	
F20	Forward 5′-GATCGAAGACAGCAGCATCAACGA	1012
S5-R5	Reverse 5′-AGCAGGATTCTGAGCAAAGGTC	
S5-RACE1	Forward 5′-TTATCGGAGCACGACTATTCTGGTT	1389
R13BR	Reverse 5′-TACGGATCCATTAATCTGGCGCCTAAGCTCGCA	
dCAPS-F	Forward 5′-GCATGGATGTCAAGTACAGCG	138
dCAPS-R	Reverse 5′-CGTCAGTGGGCAAGCAGTAG	
GUS1.6F	Forward 5′-CCAGGCAGTTTTAACGATCAGTTCGC	1576
GUS1.6R	Reverse 5′-GAGTGAAGATCCCTTTCTTGTTACCG	

For DNA sequencing, 5 µl of amplified PCR products was first digested in 0.18 µl of 25 mM $MgCl_2$, 0.3 µl of 10× rTaq buffer, 0.13 µl of SAP, 0.25 µl Exo I, filled with double-distilled water to a total 8 µl volume. The reaction was conducted for 1 h at 37°C and 20 min at 80°C in water bath. The 3 µl digested PCR product was then used as the template, together with 0.16 µl each of forward and reverse primers (both 10 µM), 0.5 µl ABI BigDye (Life Technologies Corporation, Carlsbad, CA, USA) 3.6 premixture, and 1.75 µl of 5 × BigDye buffer, and filled with double-sterilized water to a final volume of 10 µl. The reactions were conducted on Gene AMP PCR system 9700 (Applied Biosystems, Carlsbad, CA), with the following profile: 2 min at 96°C for pre-denaturation, followed by 28 cycles of 10 s at 96°C, 10 s at 50°C, 4 min at 60°C and a final 5 min extension at 72°C. The sequencing was conducted on ABI 3730 DNA analyser, and the sequence data were assembled and aligned using the DATACOLLECTION V3.6 software (Life Technologies Corporation, Carlsbad, CA, USA).

The 122 available DNA sequences have been deposited in the GenBank under accession codes HQ846206 to HQ846327. The *S5* sequence from *O. barthii* has been deposited in the GenBank under accession code of JF298922.

Sequence analysis

Sequences were inspected using SEQUENCHER V4.5 program (Gene Codes Corporation, Ann Arbor, MI, USA) and manually edited using the CONTIGEXPRESS program in Invitrogen Vector NTI advance V.10 package (Life Technologies Corporation, Carlsbad, CA, USA). Sequences from different *Oryza* accessions were aligned using CLUSTAL X version 1.83 (Thompson *et al.*, 1997) and adjusted manually with GENEDOC 2.7 (Nicholas & Nicholas, 1997).

Two measures of nucleotide variability, average number of nucleotide differences per site (π) (Nei, 1987) and nucleotide diversity based on the proportion of segregating sites (θw) (Watterson, 1975), were calculated using DNASP v5.0 (Librado & Rozas, 2009). Genetic distances between different populations were calculated with MEGA 4.1. Haplotype diversity analysis was conducted using MEGA 4.1 and DNASP v5.0 (Librado & Rozas, 2009) separately. DNASP v5.0 (Librado & Rozas, 2009) was used to perform tests of selection, including Tajima's D test (Tajima, 1989), and Fu and Li's D*, F* tests (Fu, 1997). The sliding-window method was employed via SWAAP v1.0.3 (Stanford University, Stanford, CA, USA) to analyse the polymorphisms across the 4.7-kb coding sequence of *S5*, using window size 100 and step size 20 in which pairwise insertions/deletions (InDels) were removed. A haplotype flowchart, representing unique protein sequences separated by mutational steps, was constructed with the computer program TCS 1.21 (Clement *et al.*, 2000) using genetic distance and statistical parsimony methods.

Transformation and fertility examination

Transformation was conducted following the method of Lin & Zhang (2005). Six constructs (Fig. S1) were transformed into *japonica* recipient Balilla. For preparing the constructs, recombinant fragments were amplified from the genomic DNA of Nanjing 11 and Balilla, using the primers in Table 1. The amplified fragments were ligated to the vector pCAMBIA1301 (Hajdukiewicz *et al.*, 1994) and transformed into the *Agrobacterium* strain EHA105.

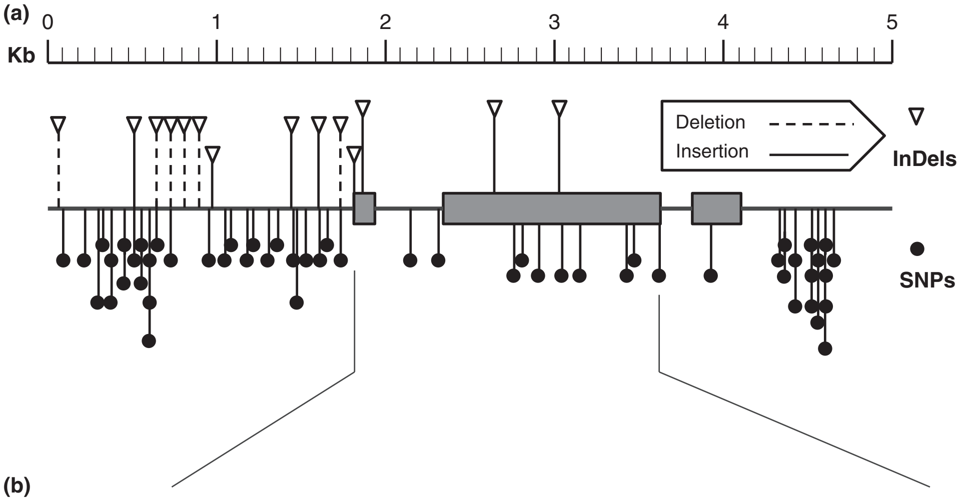
Copy number of the transgene plants was determined using Southern blot hybridization by the DIG-High Prime DNA Labeling and Detection Starter Kit I (No. 11745832910; Roche). For segregation analysis, the progeny plants were stained for glucuronidase (GUS) activity and PCR amplified for the GUS gene fragment, using primers GUS1.6F and GUS1.6R (Table 1).

Transgenic plants were grown in the summer in Wuhan, China. Panicles were harvested to examine spikelet fertility and were scored as the ratio of the number of filled grains to the total spikelets.

Results

Nucleotide variations and protein divergences in *S5* region

A sample of 122 *Oryza* accessions of three species of the AA genome *O. sativa*, *O. rufipogon* and *O. nivara*, from 15 countries representing a diverse range of the occurrence and distribution of these species, were taken from the International Genetic Resources Center maintained by IRRI. Sequences of the *S5* locus spanning approx. a 4.7-kb region from the promoter to the 3′-flanking regions were obtained for each accession. In addition, data for 16 accessions were obtained from a previously reported study (Chen



(b)

Haplotype (frequency %)	InDel1 136 bp	InDel2 351 bp	InDel3 6 bp	InDel4 103 bp	Site 1	Site 2	Site 3	InDel5 275 bp	Site 4	Site 5	Site 6	Site 7	Site 8
H1 (0.72)	no	no	yes	no	T	G	A	no	A	T	T	C	T
H2 (1.45)	no	no	yes	no	T	G	G	no	A	T	T	C	T
H3 (1.45)	yes	no	no	no	T	G	G	no	A	T	T	C	T
H4 (0.72)	yes	no	no	no	C	G	G	no	A	T	T	C	T
H5 (16.67)	no	no	no	no	T	G	G	no	A	T	T	C	T
H6 (1.45)	no	no	no	no	T	G	A	no	C	T	T	C	C
H7 (1.45)	no	no	no	yes	T	G	G	no	C	T	T	C	C
H8 (0.72)	no	no	no	no	T	G	G	no	C	T	T	T	C
H9 (5.07)	no	no	no	no	T	A	G	no	C	T	T	C	C
H10 (0.72)	no	no	no	no	T	G	G	yes	C	T	T	C	C
H11 (1.45)	yes	no	no	no	C	G	G	no	C	T	T	C	C
H12 (7.25)	yes	no	no	no	C	G	G	no	C	T	T	T	C
H13 (3.62)	yes	no	no	no	C	G	G	no	C	C	T	T	C
H14 (0.72)	yes	no	no	no	T	G	G	no	C	C	T	T	C
H15 (5.80)	no	no	yes	no	T	G	A	no	C	T	C	T	C
H16 (0.72)	yes	no	no	no	T	G	G	no	A	T	T	T	T
H17 (9.42)	yes	no	no	no	T	G	G	no	C	T	T	C	C
H18 (0.72)	no	yes	no	no	T	G	G	no	C	T	T	C	C
H19 (39.88)	no	no	no	no	T	G	G	no	C	T	T	C	C
Codon difference					CTG/CCG	AAC/GAC	TCG/TCA		TTA/TTC	ATG/ACG	CTG/CCG	CCA/TCA	GTT/GCT
AA difference					L/P	N/D	S/S		F/L	M/T	L/P	P/S	V/A
Mutation type					Replace	Replace	Silent		Replace	Replace	Replace	Replace	Replace
Position (from ATG)	–69	1	66	435	542	589	678	797	819	950	1211	1252	1412

Fig. 1 Schematic drawing of *S5* structure with the summary of DNA and protein polymorphisms in *Oryza rufipogon*, *Oryza nivara* and *Oryza sativa*. (a) The gene model of *S5*. Shaded boxes indicate three exons, and lines represent two introns and other non-coding regions. Numbering begins from the left border of promoter region. Single nucleotide polymorphisms (SNPs) are indicated by lines below the gene model with solid circles. InDels are indicated by lines above the gene model with triangles. (b) Protein variations are summarized. The site 4 and site 8 marked with shaded rectangles correspond to the two functional mutations that distinguish the *indica*-like and *japonica*-like alleles.

et al., 2008). Thus, in total, data for 138 accessions were used in this study, including 60 accessions of *O. sativa*, 40 of *O. rufipogon* and 38 of *O. nivara* (Table S1).

In total, 55 single nucleotide polymorphisms (SNPs) and 14 insertions/deletions (InDels) were found within this 4.7-kb region (Fig. 1). The average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population (π) was used to estimate the polymorphisms within *O. sativa*, *O. nivara* and *O. rufipogon* (Fig. 2a) (Nei, 1987). The *S5* sequence showed higher nucleotide polymorphism in *O. rufipogon* (π = 0.00458) than in *O. sativa* and *O. nivara* (π = 0.00207 and 0.00218, respectively; Table 2). The majority of the variable sites were found outside the open reading frame (ORF), that is, sequences in promoters or the 3′-flanking regions, while only eight SNPs and five InDels were located in coding regions.

A total of 19 haplotypes were observed for the predicted S5 proteins based on the polymorphisms described earlier (Figs 1, S2). Eight haplotypes had a 115-aa deletion in the *N*-terminus of the protein, which is a feature of the *S5*-n sequence (Chen *et al.*, 2008). Eight of the remaining 11 haplotypes shared an identical sequence featuring the amino acid combination of Phe-273 (F) and Ala-471 (A), while the other three haplotypes had the Leu-273 (L) and Val-471 (V) combination at the corresponding sites. These two functional variations encoding Phe273Leu and Ala471Val, or C819A and C1412T in terms of DNA sequence, were the characteristics of *S5*-i and *S5*-j, respectively (Chen *et al.*, 2008). Thus, the 19 haplotypes were classified into three groups, namely the WCG-like group, the *indica*-like group, and the *japonica*-like group, according to the features described. We found relatively high levels of *S5* variation within the WCG-like group (π = 0.00421,

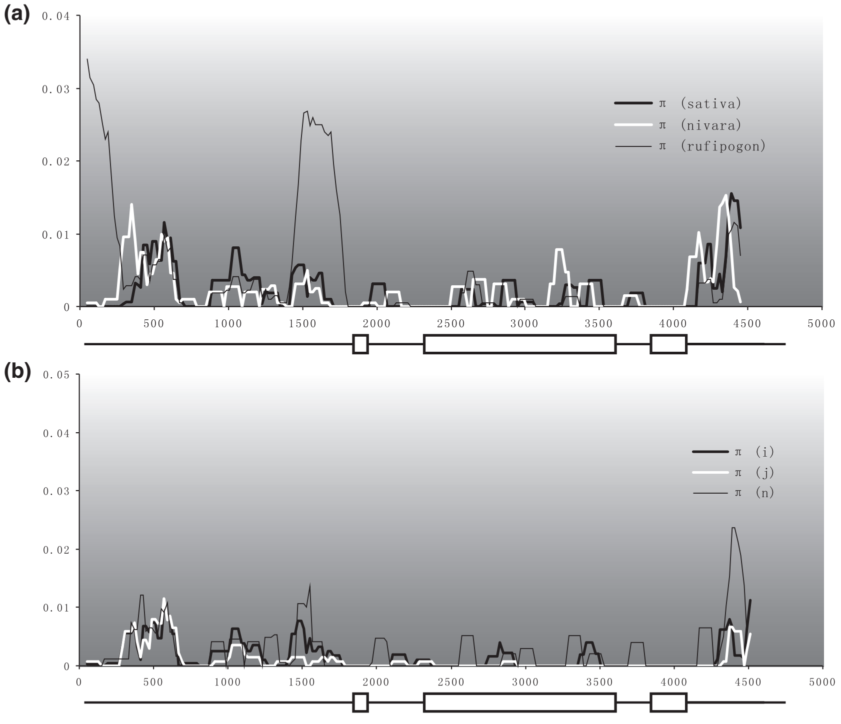


Fig. 2 Sliding-window analysis for the *S5* region spanning from promoter area to 3′-UTR in (a) three *Oryza* species (*O. sativa*, *O. nivara* and *O. rufipogon*) and (b) in three populations with different genotypes (i,j,n). The genomic structure is shown at the bottom, where the boxes indicate exons and the lines indicate introns and other noncoding regions.

Table 2 Polymorphisms and neutrality tests of different species and genotypes in *S5*

Sample population	π	θw	Tajima's D	Fu and Li's D*	Fu and Li's F*
Oryza sativa (*n* = 44)	0.00207	0.00216	−0.1498	0.39814	0.24317
O. nivara (*n* = 38)	0.00218	0.00255	−0.5863	−0.47736	−0.61127
O. rufipogon (*n* = 40)	0.00458	0.01594	**−2.66962***	**−5.82403***	**−5.58691***
indica-like accessions (*n* = 72)	0.00172	0.00181	−0.15388	0.84118	0.55065
japonica-like accessions (*n* = 22)	0.00121	0.00170	−1.10606	−0.73292	−0.98799
WCG-like accessions (*n* = 28)	0.00421	0.00899	**−2.08536***	**−3.8864***	**−3.88908***

Significance levels are determined by 10 000 random coalescent simulations based on the number of alleles and the observed number of segregating sites. Bold type indicates the significant statistics at the 95% level. *, $P < 0.05$.

Fig. 2b, Table 2), compared with those of the *indica*-like and *japonica*-like groups (π = 0.00172 and 0.00121, respectively).

A haplotype flowchart was constructed to describe the evolutionary relationships and mutational steps of these 19 haplotypes (Fig. 3). The flowchart analysis also illustrated that the haplotypes of S5 were grouped into well-defined clades, which was consistent with the tri-allelic system classification. The *indica*-like group could be further divided into three subgroups. The predominant subgroup was made up of a single haplotype H19, which was found in 39.88% of the 138 accessions. The second subgroup had five haplotypes (H6, H8, H9, H15 and H18) differentiated either by InDels or SNPs (Figs 1, 3). A 351-bp insertion (InDel 2) occurred in the 5′-UTR region of H18, and another 6-bp insertion (InDel 3) occurred in the exon 1 of H15. Within the third subgroup, premature terminations caused by frameshift insertions (InDel 4 and InDel 5) were found in two haplotypes (H7 and H10), both in exon 2. The predicted proteins encoded by H7 and H10 may not have functions because of the premature terminations. However, these two haplotypes were still classified into the *indica*-like group based on their sequence features.

The *japonica*-like group could be divided into two subgroups, whose products varied by one or two amino acids (Figs 1, 3). The first subgroup (H5) represented the more frequent haplotype and the latter (H1 and H2) comprised the rare ones. The H5 subgroup was found in 16.67% of the whole sample. Interestingly, except for the two functional variations, the predicted protein sequence of H5 was exactly the same as that of the *indica*-like H19.

A 136-bp deletion (InDel 1) involving the translation start site ATG generated the WCG-like *S5* allele, causing a frameshift and a delayed origination of translation. The WCG-like allele could also be classified into two subgroups. The first subgroup (H11-14 and H17) shared the amino acid sequence combination of Phe-273 (F) and Ala-471 (A), which was the same as the *indica*-like allele, while the latter (H3, H4 and H16) comprised the Leu-273 (L) and Val-471 (V) combination at the corresponding sites, which was the same as the *japonica*-like allele.

All these predicted proteins encoded by the *indica*-like or *japonica*-like alleles contained either 472 or 474 amino acids with almost the same overall domain structure. The only exceptions were two proteins (H7 and H10) encoded by truncated *indica*-like alleles and H18 with the 351-bp

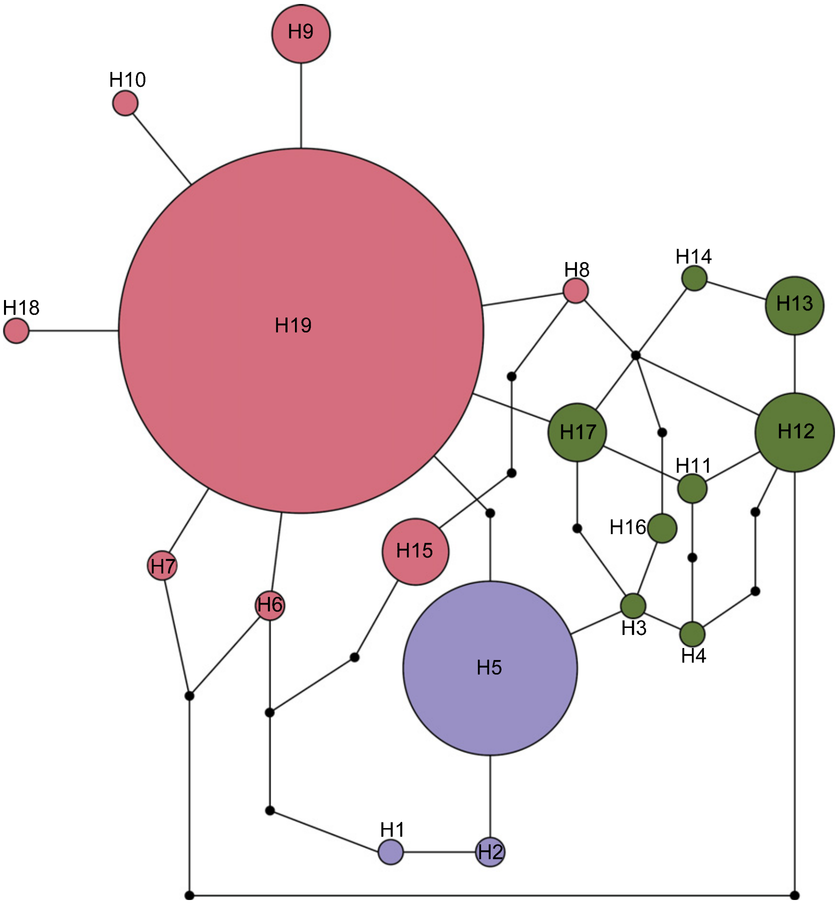


Fig. 3 The flowchart of 19 haplotypes of S5 (H1–H19). Each circle represents a unique haplotype. The size of circle corresponds to the frequency of each haplotype. The name of the 19 haplotypes are used for indicating the *indica*-like (H6–H10, H15, H18, H19), *japonica*-like (H1, H2, H5) and wide-compatibility gene (WCG)-like groups (H3, H4, H11–H14, H16, H17) of each sampled alleles. Each solid line represents one mutational step that interconnects two haplotypes, while closed dots represent possible missing haplotypes.

insertion. The WCG-like proteins contained 357 amino acids, all of which had a 115-aa deletion in the *N*-terminus.

Haplotype structure and distribution: the evolutionary history of *S5*

The closest wild relatives of *O. sativa* are *O. nivara* and *O. rufipogon*, although which of them is the immediate progenitor of the cultivated rice remains controversial. Some researchers suggested that *O. sativa* was domesticated from *O. nivara* based on the phenotypic similarity between them (Khush, 1997; Sharma *et al.*, 2000; Chang, 2003; Grillo *et al.*, 2009). Others favor *O. rufipogon* as the ancestor of Asian cultivated rice, as a direct wild progenitor (Oka, 1988).

To investigate the ancestry and evolutionary history of S5, the 19 haplotypes were aligned to identify the events that differentiate *O. sativa*, *O. nivara* and *O. rufipogon* (Fig. 4). Based on this analysis, it could be inferred that the S5 haplotypes H19 and H5, respectively, were the ancestral types in the *indica*-like group and *japonica*-like group, as they existed in all three species. To determine which of these two is more ancestral, we further sequenced *S5* from an accession of *O. barthii*, another member of the AA genome group. The predicted S5 protein from *O. barthii* was identical to haplotype H19. This result suggested that H19 was the ancestral type. H5, found in 88.46% (23/26) of the *japonica*-like accessions, differed from H19 (*indica*-like) at C819A and C1412T (Fig. 1). This clearly indicates that the divergence of the two functional variations in *S5* occurred before the speciation of *O. sativa*, *O. nivara* and *O. rufipogon*. The WCG-like H17 existed in all three species, and was identical to *indica*-like H19 across the entire sequence of the S5 protein, except for the 115-aa deletion. We can therefore deduce that H17 might be the ancestral type of the WCG-like group, and this derived 115-aa deletion in S5 conferring wide-compatibility arose in the *indica*-like clade.

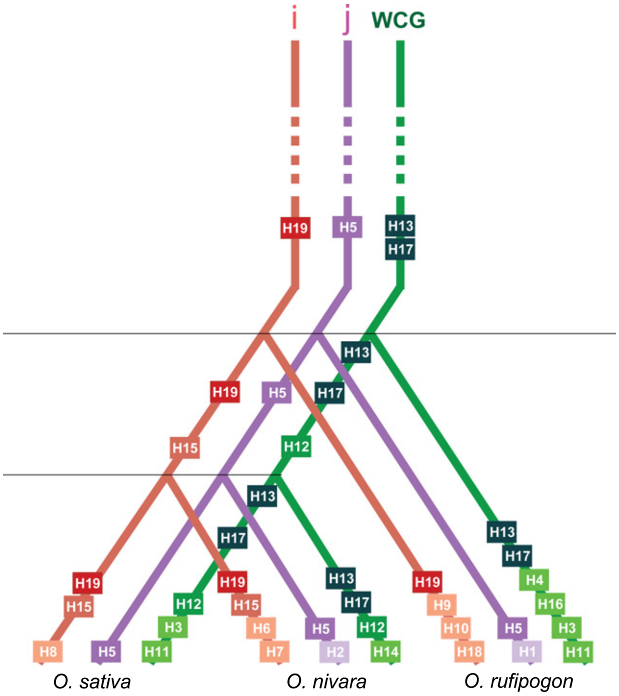


Fig. 4 Schematic drawing of the evolution history of S5 aligned with the lineage phylogeny of *Oryza sativa*, *Oryza nivara* and *Oryza rufipogon*. The thick lines indicate the lineage phylogeny of *O. sativa*, *O. nivara* and *O. rufipogon*. The *indica*-like (left), *japonica*-like (middle), and wide-compatibility gene (WCG)-like (right) S5 haplotypes are aligned in the branches according to the putative differentiate events.

In order to obtain a more comprehensive view in *S5* evolution, we further analysed the mutation events in *O. sativa*, *O. nivara* and *O. rufipogon* (Fig. 4). The *japonica*-like clade was simple and comprised only a predominant H5 and two minor variant types (H1, with one amino acid difference with H2). H1 was unique in *O. rufipogon* while H2 was only found in *O. nivara*. It may be inferred that H2 originated directly from H5, and H1 might originate from H2 after one-nucleotide substitution in the predicted coding sequence (Figs 1, 4).

The *indica*-like group comprised numerous rare haplotypes that differed by divergent mutational steps from the most common and presumed ancestral haplotype H19. H9 found in *O. rufipogon* might have directly originated from H19 after a one-nucleotide substitution in the predicted coding sequence, after the divergence of *O. rufipogon* (Figs 1, 4). Two other haplotypes H10 and H18 were also unique in *O. rufipogon*. The result suggested that they arose in *O. rufipogon* after a 275-bp insertion and a 351-bp insertion in the ancestral haplotype H19, respectively (Figs 1, 4). Therefore, the occurrences of H9, H10 and H18 suggested distinct events in the *rufipogon* lineage. H15 might have originated from H19 after a series of mutational steps before the divergence between *O. sativa* and *O. nivara*, as it was found in both species. H6 and H7 in *O. nivara* and H8 in *O. sativa* each exhibited only one mutation relative to the ancestral haplotype H19 (Figs 1, 4), suggesting that these three haplotypes were derived from H19, which might occur specifically in one species recently. Alternatively, the low detection frequency of these types may also be the result of sample size.

The WCG-like group seems to be more complex. In addition to the ancestral haplotype H17, H13 also existed in all three species and displayed the same SNPs as the *indica*-like allele like in the two functional variations (Figs 1, 4). Therefore, H13 might originate from H17 before the divergence of *O. sativa*, *O. nivara* and *O. rufipogon*. H12 and H14 were closely related to H13 (Fig. 3), and H12 was found both in *O. sativa* and in *O. nivara*, whereas H14 was unique in *O. nivara*. Therefore, both H12 and H14 might have directly originated from H13 by a one-nucleotide substitution before and after the divergence between *O. sativa* and

O. nivara (Figs 1, 4). H11 in *O. sativa* and *O. rufipogon* was closely related to H17 (Fig. 3). Therefore, H11 might directly originate from H17 by a one-nucleotide substitution. Interestingly, H3, H4 and H16 displayed the same SNPs as the *japonica*-like allele in the two functional variations. It was noteworthy that all these three haplotypes existed in *O. rufipogon* and were more related to the *japonica*-like H5 (Figs 1, 3).

Concurrence of the two functional variations is essential in determining hybrid sterility

There was perfect concurrence between the two functional variations (C819A and C1412T) in *S5*, and no recombinants between these two or other substitution at either site were found in the rice germplasm assayed. However, no such association was observed between other polymorphic sites. This suggested that the concurrence of these two sites might be relevant to the function of *S5*.

To test this hypothesis, we transformed six recombinant fragments containing various combination of genome sequences derived from different regions of *S5*-i and *S5*-j into a *japonica* variety (Balilla) (Fig. S1, Table 3). Based on present understanding (Chen *et al.*, 2008), if the transformed fragment was functional as the *indica* allele for reproductive isolation, significantly reduced spikelet fertility would be expected in the transformants.

Examination of the spikelet fertility of the T_0 plants detected no significant difference between the transgene-positive and transgene-negative plants of constructs 1–5 (Table 3), indicating that none of them was functional in reproductive isolation. By contrast, transgene-positive and transgene-negative plants of the construct 6 (the *japonica*-type promoter + *indica*-type C819 and C1412) showed a highly significant reduction in spikelet fertility; the average spikelet fertility (79.30%) of the negative plants was much higher than that of the positive plants (12.25%) (Table 3).

Single-copy transgenic T_0 plants have been obtained from constructs 1–3 and 6, and the fertility of their T_1 progenies was further examined. The spikelet fertility of the T_1 plants of constructs 1–3 also showed no statistically significant difference between the transgene-positive and transgene-negative plants. By contrast, analysis of a T_1 family from constructs 6 showed that spikelet fertility of the positive plants was greatly reduced compared with the negative segregants (Table 3). Such perfect cosegregation between the transgene and spikelet fertility confirmed that this construct has function for hybrid fertility indicating that these two amino acids have to be together for such function.

Table 3 Fertility of plants transformed with different constructs and their T_1 progenies

Transgene	Generation	Genotype	Number of plants	Spikelet fertility (%)[a]
Construct 1 *indica* promoter + C + T	T_0	Positive	7	67.50 ± 8.55
		Negative	4	81.85 ± 0.85
		t		2.262157
		P		0.246985
	T_1	Positive	23	79.50 ± 1.16
		Negative	9	77.87 ± 1.30
		t		0.800538
		P		0.214847
Construct 2 *indica* promoter + A + C	T_0	Positive	10	78.25 ± 5.84
		Negative	3	65.80 ± 16.60
		t		2.200985
		P		0.381127
	T_1	Positive	20	82.21 ± 1.36
		Negative	12	81.87 ± 2.25
		t		0.138786
		P		0.445273
Construct 3 *indica* promoter + A + T	T_0	Positive	4	62.83 ± 19.62
		Negative	3	73.63 ± 4.60
		t		2.570582
		P		0.665378
	T_1	Positive	23	77.62 ± 1.08
		Negative	9	76.97 ± 1.70
		t		0.323635
		P		0.37423
Construct 4 *japonica* promoter + A + C	T_0	Positive	5	69.82 ± 9.91
		Negative	14	78.11 ± 0.89
		t		1.739607
		P		0.171367
Construct 5 *japonica* promoter + C + T	T_0	Positive	6	60.05 ± 12.19
		Negative	3	66.87 ± 16.85
		t		2.364624
		P		0.754769
Construct 6 *japonica* promoter + C + C	T_0	Positive	17	12.25 ± 2.01
		Negative	2	79.30 ± 0.80
		t		2.109816
		P		**2.99E-09***
	T_1	Positive	26	12.86 ± 0.38
		Negative	8	79.13 ± 1.10
		t		−72.4942
		P		**2.27E-37***

Bold type indicates the highly significant statistics at the 95% level.

[a] Mean ± SEM; *, $P < 0.05$.

Geographic distribution of *S5* alleles

The geographic distributions of the 19 haplotypes in the 15 sampled areas are provided in Fig. 5. H19 and H5 were the most common haplotypes with the widest geographic distributions in the *indica*-like group and the *japonica*-like group, respectively, both of which were identified in 11 different countries. Twelve S5 haplotypes were found in *O. rufipogon*, suggesting that this species displayed more allele diversity than *O. sativa* and *O. nivara*. India was generally the most diversified area and displayed more allele variations for *S5*. Some haplotypes were found in a single area, for example H1 and H2 were found in India or in Nepal, respectively. This may be because of the small sample size. In addition, there was haplotype-sharing between areas across large geographical distances (Fig. 5). For example, H9 was found in South

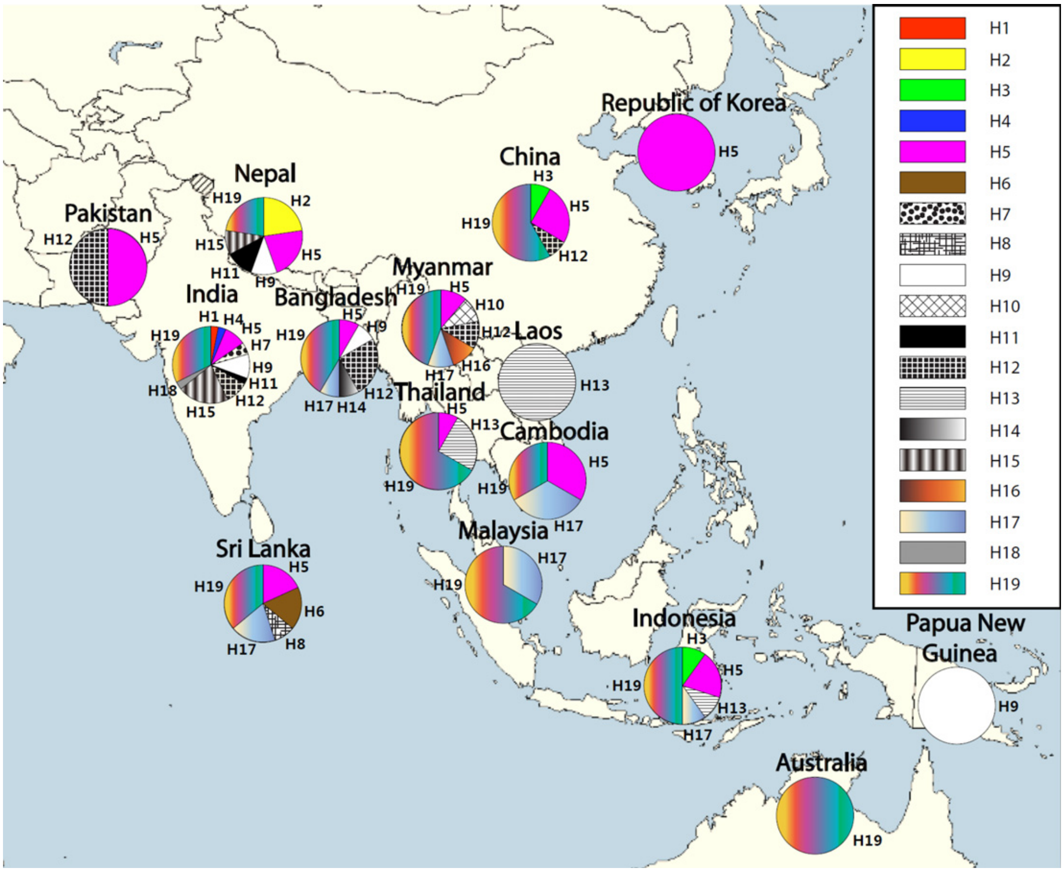


Fig. 5 Geographic distributions of different haplotypes of S5 among the 15 areas sampled. Each circle is constructed with the respective frequencies of these haplotypes at each locality. The haplotypes are indicated outside the cycle.

Asia such as India, Nepal and Bangladesh, as well as Papua New Guinea in the southern Pacific Ocean. The most ancestral haplotypes based on our simple genealogical hypotheses, that is H19 for the *indica*-like clade and H5 for the *japonica*-like clade, were the most frequent ones in all the areas investigated. However, the most frequent haplotype in the WCG-like clade was not the ancestral one (Fig. 1). This may be because of the different evolutionary process in the origination of WCG-like haplotypes.

Tests for selection in the rice lineages

We examined the selection pattern in the three rice species and populations carrying different types of *S5* alleles. Three tests – Tajima's D (Tajima, 1989), Fu and Li's D* and F* (Fu, 1997) – were employed to examine deviation from neutrality. Significant negative values for these parameters would indicate an excess of low-frequency polymorphisms that might have resulted from either population expansion or deviation from neutral evolution, while positive values signify evidence of balancing selection or a decrease in population size.

In the *rufipogon* population, the *S5* gene showed significant negative values in all the tests ($P < 0.05$, Table 2). Tests for WCG-like alleles also produced significant negative values, thus suggesting that this class of haplotypes deviated from the expectation under neutrality. The sequence data in *S5* suggested that seven of the eight SNPs found in the coding sequence were non-synonymous substitutions, indicating a high enrichment of diversity in S5 protein sequences. Therefore, the *S5* gene in the *rufipogon* population might have been under positive selection as there was an excess of non-synonymous mutations. Similarly, the WCG-like alleles also exhibited apparently enrichment of non-synonymous mutations; it seems that populations carrying such *S5* alleles underwent accelerated evolution when the 115-aa deletion emerged.

Discussion

The analyses revealed several noticeable features of the complex evolution of the *S5* locus, which concern major issues surrounding the origin and evolution of the cultivated rice.

The complex evolutionary course of the *S5*

The haplotypes of the S5 sequences uncovered a complex evolutionary course of the S5 gene. The ancestral H19 and

H5 were found in high frequencies in the *indica*-like and the *japonica*-like haplotypes in all three species, suggesting that their origins were more ancient than the formation of these three species. The two additional *japonica*-like haplotypes (H1 and H2), both of which were derived from H5, were found only in the two wild rice species, with very low frequencies (Figs 3, 4). Six of the seven remaining *indica*-like haplotypes, except H15, arose directly from H19 each by a single mutation (Fig. 3). Interestingly, the *indica*-like H15 seems to originate from H19 after a series of mutational steps, while the intermediate haplotypes have not been found, indicating the likely existence of more *indica*-like alleles.

The evolutionary course of the WCG-like haplotypes, featured by the 136-bp deletion, seems more complex. The first subgroup of the WCG-like alleles (H11-14 and H17), accounting for the majority of the WCG-like haplotypes, displayed C819 and C1412 in the two functional variations, the same as the *indica*-like ones. Therefore, it could be inferred that these WCG-like alleles arose directly from the *indica*-like ones by a 136-bp deletion followed by additional nucleotide substitutions. The second subgroup of the WCG-like alleles (H3, H4 and H16) displayed the same SNPs as the *japonica*-like ones in the two functional variations, likely resulted from a recombination between a WCG-like allele in the first subgroup and a *japonica*-like allele. Such inference is supported by the fact that H13 and H17 were found in all three species, thus having more ancient origin than the formation of these species.

Ancient history of the *S5* tri-allelic system confirmed independent origins of *indica* and *japonica* subspecies

The results clearly showed that the tri-allelic system of the *S5* locus, an *indica* allele, a *japonica* allele and a neutral allele (WCG), has an ancient history, as shown by its existence in all three species. The inference of antiquity of the system is further enhanced by the finding of allelic diversity within each of the allelic groups – *indica*-like, *japonica*-like and WCG-like – in each of the species, especially in the two wild species. A similar tri-allelic system is also found in *Sa*, a locus for reproductive isolation by regulating pollen fertility in *indica–japonica* hybrids (Long *et al.*, 2008). This suggests that reproductive isolation was already well established in the wild rice species.

Studies of genetic diversity in cultivated rice have established that *indica–japonica* differentiation at the whole-genome level comprises the major source of genetic diversity in cultivated rice (Zhang *et al.*, 1992; Han & Xue, 2003). A question along this line is that whether the *indica* and *japonica* differentiation had already occurred in the wild rice species. Multiple studies demonstrated that *indica* and *japonica* accessions showed closer affinity with different accessions of *O. rufipogon* than to each other (Second, 1982; Wang *et al.*, 1992; Caicedo *et al.*, 2007). Phylogeographic analysis at three genetic loci revealed that differentiated gene pools already existed in *O. rufipogon* (Londo *et al.*, 2006). These results suggest that the genetic backgrounds were already differentiated in wild species before utilization by humans. This is consistent with the inference based on DNA sequences that *indica* and *japonica* subgroups diverged 0.2–0.4 million yr ago (Ma & Bennetzen, 2004; Vitte *et al.*, 2004; Zhu & Ge, 2005), while rice domestication was thought to start *c.* 9000 yr ago (Khush, 1997).

Such evolutionary history of the speciation gene *S5* together with the evidence from other studies provided unambiguous evidence that *indica*-like and *japonica*-like rice groups may already exist in wild ancestors. Therefore, *indica* and *japonica* subspecies of the cultivated rice *O. sativa* arose independently in wild species rather than through derivation of one from another.

The C819 and C1412 combination is required for hybrid sterility

Previously it was identified that the *indica* and *japonica* alleles of *S5* differed by two nucleotides: C in *S5i* to A in *S5j* at site 819 (referred to as C819A) and C to T at site 1412 (C1412T), both of which caused amino acid substitutions, Phe-273 (F) to Leu-273 (L) and Ala-471 (A) to Val-471 (V). These changes were regarded as part of the cause of hybrid sterility. The transformation experiment of the present study showed that only a CC combination at the two variant sites could cause hybrid sterility when transformed to the *japonica* variety Balilla, while any other combinations could not. This result confirmed that these two variant sites as the cause of hybrid sterility in *indica–japonica* crosses.

According to crystal structure analysis, an AP protein has three domains, the central domain, the *N*-terminal lobe, and the *C*-terminal lobe (Fujinaga *et al.*, 1995; Kervinen *et al.*, 2004); both of the mutant sites in S5, amino acids 273 and 471, are located in the central domain (Chen *et al.*, 2008). By sequence alignment analysis, Chen *et al.* (2008) found that Phe-273 (F) was conserved in APs across a large range of organisms from plants to animals and humans, whereas amino acid 471 was highly variable. However, the conserved Phe-273 (hydrophobic and aromatic) is replaced by Leu (hydrophobic but nonaromatic) in S5-j, which they speculated may reduce the stability and activity of the enzyme. Loss of function by substituting A with V at amino acid 471 of S5 as a hybrid sterility regulator further suggested that the Ala residue is necessary for the function of the protein. However, several questions have to be answered before we can understand how such likely structural difference is related to the embryo sac fertility.

The primary function of *S5* remains puzzling

The alleles of speciation gene *S5* function counteractively at the species level. Reproductive isolation conferred by S5i-S5j would promote genetic differentiation between subspecies that might have also enhanced genetic diversity in the evolutionary scale. By contrast, the WCG or S5n may provide a coherent force to hold the differentiated populations together at the species level by enabling gene flow.

However, as was frequently reiterated, reproductive isolation is only a 'byproduct' of the speciation gene (Coyne & Orr, 2004), implying that hybrid sterility is only a secondary function of the speciation gene. A question for this study then arises: What is the primary function of *S5*? The fact that all the wide-compatible lines homozygous for the S5n allele are phenotypically normal suggests that S5 may not be essential for rice growth, development or reproduction (Chen *et al.*, 2008).

However, such an inference is challenged by the results of the present study. First, there was a perfect association between the two functional variations, C819A and C1412T in *S5*, that is, concurrence of CC or AT, while other combinations were not found in the rice germplasm assayed. Second, only the CC haplotype could cause hybrid sterility when transformed to a *japonica* line, while other combinations could not cause hybrid sterility. If the function of CC or AT is only to regulate hybrid sterility, which seems to be disadvantageous for the survivorship of the gene itself, such a close linkage is unnecessary and thus should not exist, which is not the case. Thus, it is highly likely that the *S5* gene actually performs an important primary function at least in an evolutionary scale, and the two haplotypes CC and AT are advantageous for the fitness of rice plants, while other combinations of these two variant sites have been disfavored.

The dynamic process of *S5* evolution under selection

It seems that natural selection acting on hybrid incompatibility genes could be a factor causing barriers to reproduction, and eventually speciation. Previous studies have indicated that many positively selected genes are responsible for reproductive isolation, and these proteins have been reported as rapidly diverged genes among species in many cases. Four hybrid incompatibility genes in *Drosophila* (*OdsH*, *Hmr*, *Nup96* and *Lhr*) showed high levels of amino acid variation and have been attributed to positive selection (Ting *et al.*, 1998; Presgraves *et al.*, 2003; Barbash *et al.*, 2004; Brideau *et al.*, 2006). Concerted evolution and positive selection have also rapidly altered the sequence of a hybrid sterility gene *Prdm9* in mice (Oliver *et al.*, 2009). Therefore, multiple substitutions driven by positive selection may be a general phenomenon required to generate speciation genes.

However, beyond our expectation, we did not detect selection acting at the incompatible *indica*-like and *japonica*-like alleles. Conversely, there was seemingly strong selection acting at the WCG-like alleles, as indicated by the neutrality tests. Therefore, we supposed that the tri-allelic system of *S5* might have a more ancient history than the speciation of *O. sativa*, *O. nivara*, and *O. rufipogon*, and the selective pressure of the system in nature might be a dynamic process. One might speculate that the incipient incompatible alleles emerging in the wild progenitors underwent natural selection in very ancient times, which somewhat caused genetic differentiation of the rice population into primitive *indica*-like and *japonica*-like types, and such differentiation would be gradually enhanced as the evolution proceeded. Subsequently, the compatible alleles emerged and provided an opposing force to hold the differentiated populations together. These alleles might have been favored by selection gradually, likely because of higher reproduction rates in the hybrids than the other two allelic groups.

Therefore, the reason why we did not detect selection in the incompatible *indica*-like and *japonica*-like alleles might be that the samples we used diverged more recently, while the incipient incompatible alleles emerged longer ago. We inferred that selection in incompatible alleles might be detected using accessions across broader species in *Oryza* genus, as selection driving reproductive isolation between *indica* and *japonica* subpopulations occurred at a more ancient time. This work will be done in future studies for further elucidating the counteractive process of *S5* evolution.

Thus, the *S5* locus is under a complex interplay of evolutionary forces involving both primary and secondary functions of the three allelic groups subjected to both natural selection and artificial breeding. Complete understanding of the functions and their impacts require detailed characterization of the alleles at various levels and at an evolutionary scale.

Acknowledgements

We thank Professor Hanhui Kuang for helpful discussions, and Dr D.S. Brar of the International Rice Research Institute for providing the rice seeds. This research was supported by grants from the National Program on Key Basic Research Project, National Special Key Project on Functional Genomics of Major Plants and Animals, and National Natural Science Foundation of China.

References

Barbash DA, Awadalla P, Tarone AM. 2004. Functional divergence caused by ancient positive selection of a *Drosophila* hybrid incompatibility locus. *PLoS Biology* **2**: e142.

Bautista NS, Solis R, Kamijima O, Ishii T. 2001. RAPD, RFLP and SSLP analyses of phylogenetic relationships between cultivated and wild species of rice. *Genes & Genetics Systems* **76**: 71–79.

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA. 2006. Two Dobzhansky–Muller genes interact to cause hybrid lethality in *Drosophila*. *Science* **314**: 1292–1295.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR *et al.* 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genetics* **3**: 1745–1756.

Chang TT. 1976. Origin, evolution, cultivation, dissemination, and Asian and African rice. *Euphytica* **25**: 425–441.

Chang TT. 2003. Origin, domestication, and diversification. In: Smith CW, Dilday RH, eds. *Rice: origin, history, technology, and production.* Hoboken, NJ, USA: John Wiley & Sons, Inc., 3–25.

Chen J, Ding J, Ouyang Y, Du H, Yang J, Cheng K, Zhao J, Qiu S, Zhang X, Yao J *et al.* 2008. A triallelic system of *S5* is a major regulator of the reproductive barrier and compatibility of *indica–japonica* hybrids in rice. *Proceedings of the National Academy of Sciences, USA* **105**: 11436–11441.

Cheng C, Motohashi R, Tsuchimoto S, Fukuta Y, Ohtsubo H, Ohtsubo E. 2003. Polyphyletic origin of cultivated rice: based on the interspersion pattern of SINEs. *Molecular Biology and Evolution* **20**: 67–75.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**: 1657–1659.

Coyne JA, Orr HA. 2004. *Speciation.* Sunderland, MA, USA: Sinauer Associates.

Dally AM, Second G. 1990. Chloroplast DNA diversity in wild and cultivated species of rice (genus *Oryza*, section *Oryza*). Cladistic-mutation and genetic-distance analysis. *Theoretical and Applied Genetics* **80**: 209–222.

Darwin C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* London, UK: J. Murray.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**: 11–15.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925.

Fujinaga M, Chernaia MM, Tarasova NI, Mosimann SC, James MN. 1995. Crystal structure of human pepsin and its complex with pepstatin. *Protein Science* **4**: 960–972.

Ge S, Sang T, Lu BR, Hong DY. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proceedings of the National Academy of Sciences, USA* **96**: 14400–14405.

Grillo MA, Li C, Fowlkes AM, Briggeman TM, Zhou A, Schemske DW, Sang T. 2009. Genetic architecture for the adaptive origin of annual wild rice, *Oryza nivara*. *Evolution* **63**: 870–883.

Hajdukiewicz P, Svab Z, Maliga P. 1994. The small, versatile pPZP family of *Agrobacterium* binary vectors for plant transformation. *Plant Molecular Biology* **25**: 989–994.

Han B, Xue Y. 2003. Genome-wide intraspecific DNA-sequence variations in rice. *Current Opinion in Plant Biology* **6**: 134–138.

Ikehashi H, Araki H. 1986. Genetics of F_1 sterility in remote crosses of rice. In: **IRRI, eds. *Rice genetics*. Manila, Philippines: Philippines International Rice Research Institute, 119–130.**

Kato S, Kosaka H, Hara S. 1928. On the affinity of rice varieties as shown by fertility of hybrid plants (Japanese and English). *Bulletin of Sciences of Faculty of Agriculture, Kyushu University* **3**: 132–147.

Kervinen J, Wlodawer A, Zdanov A. 2004. 17. Phytepsin. In: Barrett A, Rawlings N, Woessner J, eds. *Handbook of proteolytic enzymes, 2nd edn.* Amsterdam, the Netherlands: Academic Press, 77–84.

Khush GS. 1997. Origin, dispersal, cultivation and variation of rice. *Plant Molecular Biology* **35**: 25–34.

Kovach MJ, Sweeney MT, McCouch SR. 2007. New insights into the history of rice domestication. *Trends in Genetics* **23**: 578–587.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Lin YJ, Zhang Q. 2005. Optimising the tissue culture conditions for high efficiency transformation of indica rice. *Plant Cell Reports* **23**: 540–547.

Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA. 2006. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences, USA* **103**: 9578–9583.

Long Y, Zhao L, Niu B, Su J, Wu H, Chen Y, Zhang Q, Guo J, Zhuang C, Mei M *et al.* 2008. Hybrid male sterility in rice controlled by interaction between divergent alleles of two adjacent genes. *Proceedings of the National Academy of Sciences, USA* **105**: 18871–18876.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences, USA* **101**: 12404–12410.

Murashige T, Skoog F. 1962. A revised medium for rapid growth and bioassays with tobacco tissue culture. *Physiologia Plantarum* **15**: 473–497.

Nei M. 1987. *Molecular evolutionary genetics.* New York, NY, USA: Columbia University Press.

Nicholas KB, Nicholas HBJ. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author [WWW document]. URL http://www.psc.edu/biomed/genedoc.

Oka HI. 1988. *Origin of cultivated rice.* Tokyo, Japan: Japan Scientific Societies Press.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics* **5**: e1000753.

Ouyang Y, Chen J, Ding J, Zhang Q. 2009. Advances in the understanding of inter-subspecific hybrid sterility and wide-compatibility in rice. *Chinese Science Bulletin* **54**: 2332–2341.

Ouyang Y, Liu YG, Zhang Q. 2010. Hybrid sterility in plant: stories from rice. *Current Opinion in Plant Biology* **13**: 186–192.

Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* **423**: 715–719.

Sang T, Ge S. 2007. Genetics and phylogenetics of rice domestication. *Current Opinion in Genetics and Development* **17**: 533–538.

Second G. 1982. Origin of the genetic diversity of cultivated rice: study of the polymorphism scored at 40 isozyme loci. *Japanese Journal of Genetics* **57**: 25–57.

Sharma SD, Tripathy S, Biswal J. 2000. Origin of *O. sativa* and its ecotypes. In: Nanda JS, ed. *Rice breeding and genetics: research priorities and challenges.* Enfield, NH, USA: Science Publications, 349–369.

Sweeney M, McCouch S. 2007. The complex history of the domestication of rice. *Annals of Botany* **100**: 951–957.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**: 4876–4882.

Ting CT, Tsaur SC, Wu ML, Wu CI. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **282**: 1501–1504.

Ting Y. 1949a. Chronological studies of the cultivation and the distribution of rice varieties, Keng and Sen (in Chinese). *Agricultural Bulletin of the College of Agriculture, Sun Yatsen University* **6**: 1–32.

Ting Y. 1949b. A preliminary report on the cultivation and distribution of *hsien* and *keng* rices in ancient China and the classification of current cultivars (in Chinese). *Memoirs of the College of Agriculture, Sun Yatsen University* **6**: 1–32.

Vitte C, Ishii T, Lamy F, Brar D, Panaud O. 2004. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). *Molecular Genetics and Genomics* **272**: 504–511.

Wang Z, Second G, Tanksley S. 1992. Polymorphism and phylogenetic relationship among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. *Theoretical and Applied Genetics* **113**: 885–894.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**: 256–276.

Yamagata Y, Yamamoto E, Aya K, Win KT, Doi K, Sobrizal, Ito T, Kanamori H, Wu J, Matsumoto T *et al*. 2010. Mitochondrial gene in the nuclear genome induces reproductive barrier in rice. *Proceedings of the National Academy of Sciences, USA* **107**: 1494–1499.

Yanagihara S, McCouch SR, Ishikawa K, Ogi Y, Maruyama K, Ikehashi H. 1995. Molecular analysis of the inheritance of the *S-5* locus, conferring wide compatibility in *indica/japonica* hybrids of rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* **90**: 182–188.

Zhang Q, Liu KD, Yang GP, Saghai Maroof MA, Xu CG, Zhou ZQ. 1997. Molecular marker diversity and hybrid sterility in *indica–japonica* rice crosses. *Theoretical and Applied Genetics* **95**: 112–118.

Zhang Q, Saghai Maroof MA, Lu TY, Shen BZ. 1992. Genetic diversity and differentiation of *indica* and *japonica* rice detected by RFLP analysis. *Theoretical and Applied Genetics* **83**: 495–499.

Zhu Q, Ge S. 2005. Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. *New Phytologist* **167**: 249–265.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 A diagram of the recombinant genomic fragments used for preparing the transformation constructs of the *S5* gene.

Fig. S2 Predicted sequences of the S5 protein for each of the 19 haplotypes.

Table S1 *S5* information from 122 *Oryza* accessions and *Oryza barthii*, together with data of 16 published sequences

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.