Theor Appl Genet (2008) 116:739–744
DOI 10.1007/s00122-007-0707-x

ORIGINAL PAPER

Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice

Vanessa Prigge · Hans Peter Maurer · David J. Mackill · Albrecht E. Melchinger · Matthias Frisch

Received: 16 May 2007 / Accepted: 18 December 2007 / Published online: 31 January 2008

Abstract Computer simulations are useful tools to optimize marker-assisted breeding programs. The objective of our study was to investigate the closeness of computer simulations of the recurrent parent genome recovery with experimental data obtained in two marker-assisted backcrossing programs in rice (*Orzya sativa* L.). We simulated the breeding programs as they were practically carried out. In the simulations we estimated the frequency distributions of the recurrent parent genome proportion in the backcross populations. The simulated distributions were in good agreement with those obtained practically. The simulation results were also observed to be robust with respect to the choice of the mapping function and the accuracy of the linkage map. We conclude that computer simulations are a useful tool for pre-experiment estimation of selection response in marker-assisted backcrossing.

Introduction

Marker-assisted backcrossing (MAB) is used to transfer genes conferring favorable agronomic traits into the genetic background of a recipient parent (Hospital 2005; Frisch 2005). Common applications are the introgression of resistance genes or transgenes in maize, rice, and other crops (Ragot et al. 1995; Neeraja et al. 2007; Lecomte et al. 2004). Important factors that determine the efficiency of a MAB program are, for example, the number of target genes to be transferred, the marker map, the crossing scheme, and the applied selection strategy (Frisch 2005).

Computer simulations are a useful tool to investigate and optimize MAB programs with respect to the above factors. Employing molecular markers with known map position speeds up recovery of the recurrent parent genome (RPG) by about two to three generations (Hospital et al. 1992; Ribaut and Hoisington 1998). Hospital and Charcosset (1997) investigated marker-assisted introgression of quantitative trait loci (QTL) combining selection for the target gene to be introgressed with that for the genetic background of the recipient parent. Selection strategies and breeding plans for the simultaneous introgression of two genes (Frisch and Melchinger 2001a) or a recessive gene (Frisch and Melchinger 2001b) were compared regarding the required marker data points and the RPG recovered. Ribaut et al. (2002) investigated the efficiency of marker-assisted selection for the recurrent parent alleles (background selection) in different generations of a MAB program. However, to our knowledge no results are published investigating the fit of simulated to experimentally obtained frequency distributions of the RPG recovery in MAB programs.

The objective of our study was to compare the simulated probability distributions of the RPG contribution in MAB with the experimental data in two MAB programs for the introgression of submergence-tolerance QTL *Sub1* into two rice cultivars. Further, the robustness of the simulations with respect to mapping function and linkage map was considered.

Communicated by D. A. Hoisington.

Vanessa Prigge and Hans Peter Maurer contributed equally to this work.

V. Prigge · H. P. Maurer · A. E. Melchinger (✉) · M. Frisch
Institute of Plant Breeding, Seed Science,
and Population Genetics, University of Hohenheim,
70593 Stuttgart, Germany
e-mail: melchinger@uni-hohenheim.de

D. J. Mackill
International Rice Research Institute,
DAPO Box 7777, Metro Manila, The Philippines

Materials and methods

Experimental data

We investigated two MAB programs carried out in rice by the International Rice Research Institute, the Phillipines, to introduce the submergence-tolerance QTL designated *Sub1* (Xu and Mackill 1996) from donor parent IR49830-7-1-2-3 into the recipients, cultivars Swarna and Samba Mahsuri. The *Sub1* QTL is located within a chromosome interval of 0.06 cM length, which is $d = 3.3$ cM distant from the telomere on the short arm of chromosome nine (Xu et al. 2006).

The MAB program in Swarna was discussed in detail by Neeraja et al. (2007). Here, we only briefly summarize the information on the linkage maps and crossing schemes required to conduct the simulations; as well as the relevant details of the MAB in cultivar Samba Mahsuri. Both programs were initiated with a cross of the recipient cultivars with the donor. Subsequently, the cultivars Swarna and Samba Mahsuri were used as recurrent parent (RP).

In the Swarna MAB program (Neeraja et al. 2007; Fig. 1), a BC_1 population of size $n = 697$ was generated, and selection for *Sub1* was carried out with the two target markers RM464 ($d = 3.3$ cM) and SSR1 ($d = 4$ cM). Further, the marker RM219 ($d = 11.7$ cM) was used to identify $n' = 20$ individuals with a small donor genome segment carrying the target gene. These 20 plants, heterozygous at markers RM464 and SSR1 and homozygous for the RP allele at RM219, were evaluated for the genetic background of the RP at $b = 56$ loci. Plant no. 242 was selected and backcrossed to the RP to generate the population BC_2 ($n = 320$). These 320 plants were analyzed for markers RM464 and SSR1 and $n' = 158$ plants heterozygous at these target markers were selected and evaluated at $b = 64$ background markers (8 additional to the 56 background markers employed in BC_1). Plant no. 246, which was heterozygous at the target markers and showed the highest RPG percentage at the markers used for background selection, was selfed to generate the $BC_2{-}S_1$ population ($n = 420$). In generation $BC_2{-}S_1$, selection for *Sub1* was carried out with the target markers RM464 and SSR1. In addition, the target marker RM464A was employed to verify that no double crossover occurred in the target region. For background selection, five polymorphic markers were used. On the basis of this analysis, $n' = 4$ $BC_2{-}S_1$ plants were preselected and a total of $b = 95$ background markers were used to confirm their genotypes for RP. Plant no. 237 was homozygous for the donor allele at the target markers and homozygous for the RP alleles at 94 of the 95 background markers. It was selected in $BC_2{-}S_1$ as final product of the Swarna MAB program.

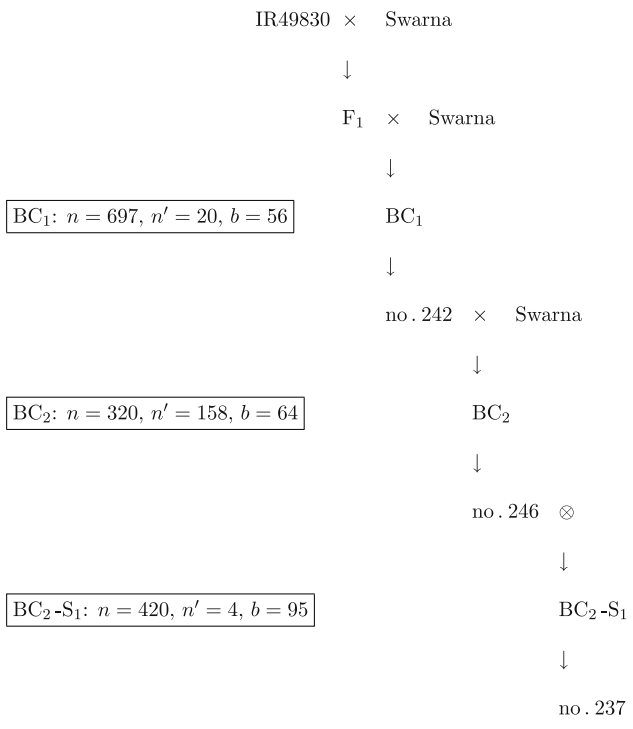

Fig. 1 Breeding scheme for introgression of the submergence-tolerance *Sub1* from the donor line IR49830 into the recipient line Swarna. n is the population size before and n' that after preselection for presence of donor alleles at the target markers, and b is the number of background selection markers analyzed in the n' plants

In the Samba Mahsuri MAB program, a BC_1 population of size $n = 388$ was generated (D. Mackill, personal communication; Fig. 2). In the BC_1 population selection for *Sub1* was carried out with the marker GnS2 ($d = 6.8$ cM), which was not available when the marker analyses were started. Therefore, the entire population ($n = n' = 388$) was analyzed with $b = 47$ background selection markers and BC_1 plant no. 166 was selected. The selected plant was backcrossed to the RP to generate the BC_2 population ($n = 80$). In the BC_2 population, selection for *Sub1* was carried out with the marker RM8300 ($d = 6.9$ cM), and $n' = 37$ plants thus selected were subjected to analysis of $b = 57$ markers for background selection. On these bases, BC_2 plants no. 46 and 62 were selected. Plant no. 62 was selfed to generate the $BC_2{-}S_1$ population ($n = 130$), which was analysed for the target markers GnS2 and RM8300. One plant thus identified (plant no. 36) was analysed for $b = 59$ background markers and was finally selected. In addition, the second selected plant in BC_2 generation, plant no. 46 was backcrossed to generate the BC_3 population with $n = 14$ plants. The marker genotype of these 14 plants was assessed and $n' = 4$ plants carrying the donor alleles at the target markers GnS2 and RM8300 were selected and further analyzed for $b = 67$ background selection markers. The BC_3 plant no. 12 was thus selected and selfed to generate population $BC_3{-}S_1$. In the $BC_3{-}S_1$ generation, stepwise

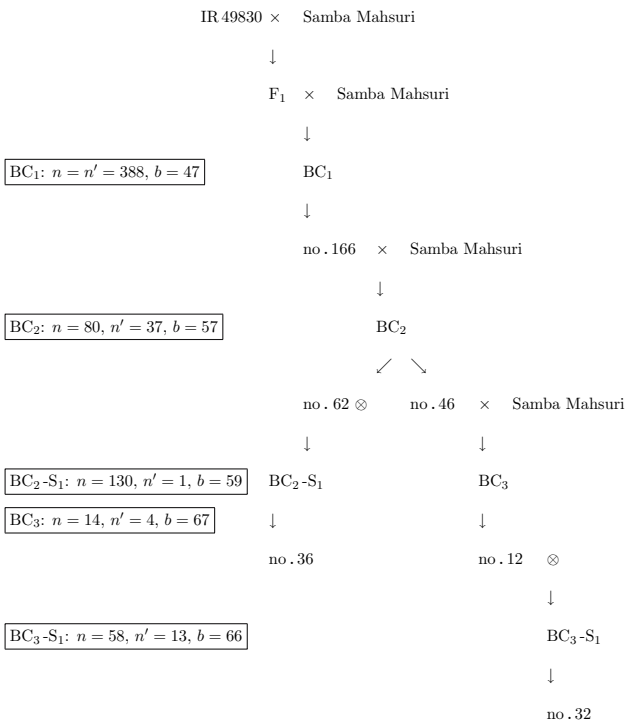

Fig. 2 Breeding scheme for introgression of the submergence-tolerance QTL *Sub1* from the donor line IR49830 into the recipient line Samba Mahsuri. The symbols n, n', and b are explained in Fig. 1

selection, as outlined above, was carried out for donor target markers and $b = 66$ background selection markers, n and n' being 58 and 13, respectively, and plant no. 32 was selected as another final product of the Samba Mahsuri MAB program.

Computer simulations

The simulations were based on two linkage maps constructed from the pairwise recombination frequencies of the Swarna and Samba Mahsuri BC_1 datasets, respectively. We employed the least-squares based approach of Stam (1993) using Haldane's (1919) mapping function. The calculations were done with the mathematical software Maple (http://www.maplesoft.com). Markers additionally analyzed in generations BC_2, $BC_2{-}S_1$, BC_3, and $BC_3{-}S_1$ were integrated into the map by using map distances from the literature (McCouch et al. 1997, 2002; Temnykh et al. 2000).

We simulated the two backcross programs described in Figs. 1 and 2 using software Plabsoft (Maurer et al. 2008), assuming no interference in crossover formation (Haldane 1919, Stam 1979). In the simulations, we used the marker-genotypes of the individuals which were actually selected in the MAB programs to simulate the advanced generations. Each introgression program was simulated 10,000 times. For each simulation run, the percentage of RPG at the marker loci in (a) the entire set of n' plants analyzed with background selection markers and (b) the selected individuals of each generation was assessed by dividing the number of loci homozygous for the recurrent parent allele by the total number of loci monitored.

To assess the fit of the simulated probability distributions to the experimental data, we carried out a χ^2-goodness-of-fit test (cf. Armitage and Berry 1994) for the n' analyzed plants of the BC_1 and BC_2 populations. For the test, the data were divided in classes of equal class width and at least five observations per class. Due to the small number n' of plants analyzed with background markers, the test was not carried out in four populations in the $BC_2{-}S_1$, BC_3, and the $BC_3{-}S_1$ generations. To compare the RPG values of the selected plants in the experimental data with the computer simulations, we determined the 2.5 and 97.5% quantiles of the simulated RPG distributions for the selected plants.

Results

Comparison of the simulated and observed probability distributions of the RPG showed a high degree of closeness (Fig. 1). The χ^2 test was applied in four populations in the BC_1 and BC_2 generations. It was not significant in three cases but was significant ($P < 0.01$) in the Swarna BC_2 population. In the remaining four populations, in which the χ^2 was not carried out, the simulated and the observed values agreed well, in spite of the very small population sizes (1, 4, 4, and 13). In the Swarna BC_2 population, wherein the χ^2 test was significant, the observed mean of the RPG was slightly smaller than the simulated value. This trend, however, could not be confirmed across populations. In comparison to the simulated value, the observed mean was smaller in four populations, higher in two populations, and in the remaining two populations, the means were extremely close. No systematic differences between the experimental and simulated values were observed regarding the standard deviations of the RPG.

The selected individuals had RPG values, which were near the 2.5% quantiles of their simulated distributions in all populations with the exception of Swarna BC_2-S_1 and Samba Mahsuri BC_3, where the values were closer to the means than to the 2.5% quantiles (Table 1).

Discussion

Mapping function and linkage map

Computer simulations are conducted on the basis of linkage maps, modelling the sequence of loci on a chromosome.

Table 1 Observed RPG values of the selected individuals as well as mean, 2.5% ($Q_{2.5}$), and 97.5% ($Q_{97.5}$) quantiles of the simulated distribution of the RPG in the selected individuals

Generation	Simulated			Observed
	$Q_{2.5}$	Mean	$Q_{97.5}$	
Swarna				
BC_1	0.82	0.86	0.91	0.83
BC_2	0.96	0.97	0.99	0.95
BC_2-S_1	0.96	0.98	0.99	0.97
Samba Mahsuri				
BC_1	0.86	0.89	0.93	0.85
BC_2	0.96	0.97	0.98	0.95
BC_3	0.95	0.96	0.97	0.96
BC_2-S_1	0.89	0.91	0.93	0.88
BC_3-S_1	0.96	0.97	0.99	0.96

Therefore, we assessed the robustness of the simulations with respect to the mapping function employed and estimation errors of map distances between adjacent loci.

We used Haldane's (1919) mapping function, which assumes no interference in crossover formation. Under positive interference, as assumed in the model underlying Kosambi's (1944) mapping function, crossovers are distributed more evenly on the chromosome than under the assumption of no interference. This is a lower probability of very short and very long chromosome segments, and consequently in a smaller variance of the parental genome contribution. We carried out additional simulations of the BC_1 and BC_2 populations using Kosambi's mapping function, and found that the mapping function did not improve the closeness of observed and simulated value (results not shown). We conclude that our simulation results are robust with respect to the choice of the mapping function.

The linkage maps employed in the simulations were generated from the observed pairwise recombination frequencies in the MAB programs using specifically written routines for least-squares estimation (Stam 1993). This high input effort was undertaken to use as accurate maps as possible. The map distances between adjacent markers in our linkage map vs. those in published maps (McCouch et al. 1997, 2002; Temnykh et al. 2000) differed up to 20 cM in case of the Swarna dataset. We compared the results of simulations based on our custom made map and the published maps, and observed the simulation results to be only marginally different (results not shown). We conclude that simulations are robust with respect to the accuracy of the linkage map and that differences of the observed magnitude have no severe effects on simulation results. Therefore, it may not be necessary to develop custom made linkage maps for each simulation study.

RPG content of the BC populations

There was a good agreement between the observed and simulated RPG values except for the Swarna BC_2 population, wherein the observed mean RPG content was significantly lower than the simulated mean (Fig. 3). This could not be attributed to sampling effects, as n' was as large as 158 in this population. Further, there was good agreement between observed and expected means in other populations having sample size as small as 1–13. A systematic estimator error of the simulation algorithm could also not be the reason for the deviations, because the simulations fit quite well the experimental data for the other populations and the deviations have been observed in both, positive and negative directions. The deviations may have arisen from discrepancies in scoring the marker bands.

Prediction of the selection response

The number of marker-assayed individuals was largest for the Samba Mahsuri BC_1 population ($n = n' = 388$) among all marker-assayed populations. Here, the simulated and observed probability distributions were in good agreement (Fig. 3). On the other hand, due to the small sample size of $n' = 20$ in the Swarna BC_1 population, the sampling effect has a geater impact on the actually observed RPG distribution (Fig. 3). The effect of sampling on the variance of the predicted response to selection is well known from classical selection theory, and is similarly expected for prediction of response to selection in MAB.

An a posteriori model was used for the simulations in this study, that is, we used the marker genotypes of the plants actually selected in the experimental MAB program to simulate the corresponding backcross. Simulations are often used as an a priori approach, that is, the selection response for a breeding program is assessed with simulations before starting the program and, hence, before actually knowing the genotype of the marker genotypes of the selected individuals. Therefore, the RPG of the selected individuals in a MAB program may show larger deviation from the simulated mean than those obtained in the present study.

In a practical breeding program, the sample size decreases with the advancement of the filial generation and selection program. The quantiles of the simulated distributions take into account the large sampling variance of the selection response due to small sample size. Therefore, for obtaining realistic predictions it seems of particular importance to use the quantiles of the simulated probability distribution of the selection response as suggested by Frisch et al. (1999), but not its expectation as is commonly used in analytical approaches for standard selection theory.

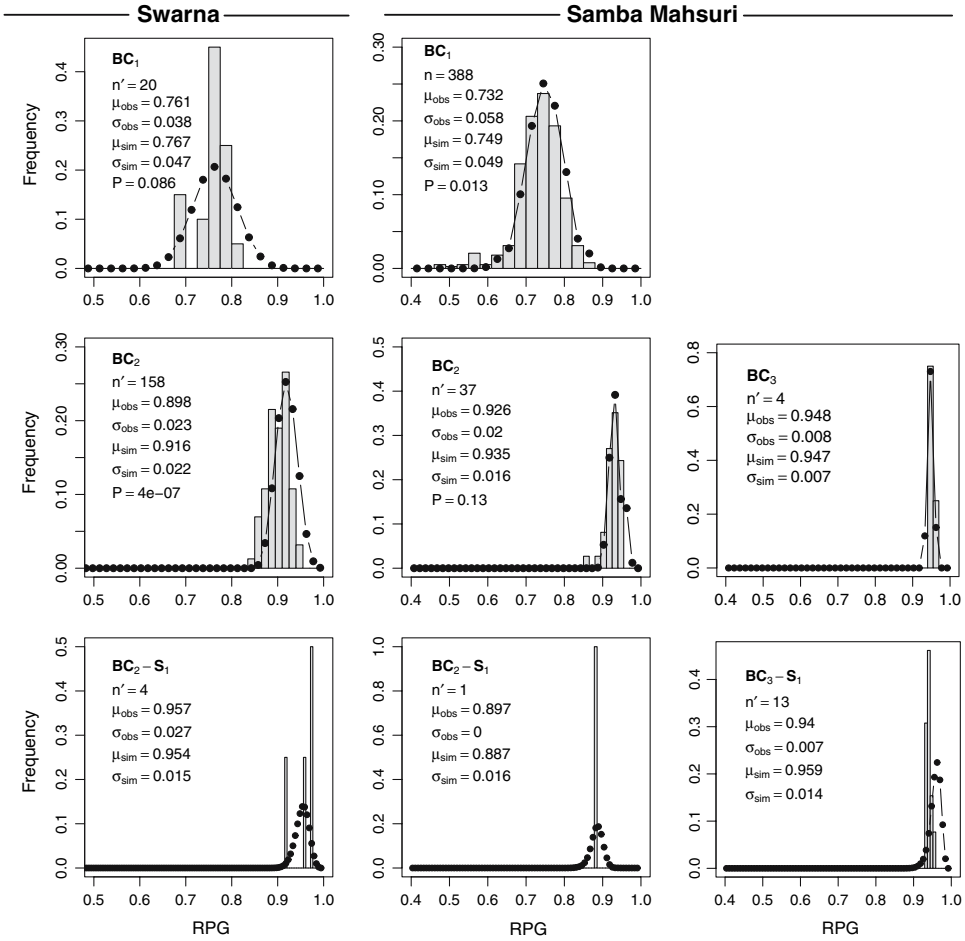

Fig. 3 Frequency distribution of experimental (*bars*) and simulated (*lines*) values for the RPG percentage recovered in the two backcross programs. For rice cultivar Swarna (*left*), the results of BC_1, BC_2,and BC_2-S_1 are presented. For rice cultivar Samba Mahsuri (*center and right*), the results of BC_1, BC_2, BC_3, BC_2-S_1, and BC_3-S_1 are presented. The acronyms denote: μ_{sim} and μ_{obs}: mean of the simulated and experimentally observed values; σ_{sim} and σ_{obs}: standard deviation for simulated and experimentally observed values; n is the population size before and n' that after preselection for presence of donor alleles at the target markers; P: P-value of the χ^2 test

In the selection experiments under evaluation, the RPG values of the selected individuals were near the 2.5% quantile in six out of eight populations (Table 1). An RPG value near the 2.5% quantile may be expected to occur due to random sampling, but the probability of such sampling effects in 75% of populations, as in the present study, is low. More experimental data needs to be generated, to evaluate whether there is any systematic underestimation in the present dataset and/or whether the simulation algorithm needs improvement.

Summarizing, within the limitations of the sample size used and the assumptions made, the simulated probability distributions matched well with the experimental data, and the RPG values of the selected plants were within or close to the simulated quantiles of their distribution. We conclude that computer simulations are a useful tool for the estimation of expected response to selection in MAB and, therefore, can greatly facilitate the development of an optimal selection program.

Acknowledgments We thank Dr. E. M. Septiningsih, Dr. B. Collard, and Prof. B. S. Dhillon for helpfull comments on the manuscript and appreciate the editorial work of Dr. J. Muminović. We thank the anonymous reviewers for their comments and suggestions, which helped to improve the manuscript. The financial support from the Bundesministerium für Wirtschaftliche Zusammenarbeit (BMZ, Germany) is gratefully acknowledged.

References

Armitage P, Berry G (1994) Statistical methods in medical research, 3rd edn. Blackwell, Oxford

Frisch M (2005) Optimum design of marker-assisted backcross programs. In: Loerz H, Wenzel G (eds) Biotechnology in agriculture and forestry vol 55. Springer, Berlin, pp 319–334

Frisch M, Melchinger AE (2001a) Marker-assisted backcrossing for simultaneous introgression of two genes. Crop Sci 41:1716–1725

Frisch M, Melchinger AE (2001b) Marker-assisted backcrossing for introgression of a recessive gene. Crop Sci 41:1485–1494

Frisch M, Bohn M, Melchinger AE (1999) Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. Crop Sci 39:967–975

Haldane JBS (1919) The combination of linkage values and the calculation of distance between the loci of linkage factors. J Genet 8:299–309

Hospital F (2005) Selection in backcross programmes. Phil Trans R Soc B 360:1503–1512

Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469–1485

Hospital F, Chevalet C, Mulsant P (1992) Using markers in gene introgression breeding programs. Genetics 132:1199–1210

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172–175

Lecomte L, Duffe P, Buret M, Servin B, Hospital F, Causse M (2004) Marker-assisted introgression of five QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. Theor Appl Genet 109:658–668

Maurer HP, Melchinger AE, Frisch M (2008) Population genetic simulation and data analysis with Plabsoft. Euphytica (in press). doi:10.1007/s10681-007-9493-4

McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y, Cho YG, Huang N, Ishii T, Blair M (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol 35:89–99

McCouch SR, Teytelman L, Xu Y, Lobos KB, Clare K, Walton M, Fu B, Maghirang R, Li Z, Xing Y, Zhang Q, Kono I, Yano M, Fjellstrom R, DeClerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res 9:199–207

Neeraja CN, Maghirang-Rodriguez R, Pamplona A, Heuer S, Collard BCY, Septiningsih EM, Vergara G, Sanchez D, Ismail AM, Mackill DJ (2007) A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. Theor Appl Genet 115:767–776

Ragot M, Biasiolli M, Delbut MF, Dell'Orco A, Malgarini L, Thevenin P, Vernoy J, Vivant J, Zimmermann R, Gay G (1995) Marker-assisted backcrossing: a practical example. In: Techniques et utilisations des marqueurs moleculaires. Montepellier, France. INRA, Paris, 29–31 March 1994

Ribaut J-M, Hoisington D (1998) Marker-assisted selection: new tools and strategies. Trends Plant Sci 3:236–239

Ribaut J-M, Jiang C, Hoisington D (2002) Simulation experiments on efficiencies of gene introgression by backcrossing. Crop Sci 42:557–565

Stam P (1979) Interference in genetic crossing over and chromosome mapping. Genetics 92:573–594

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Temnykh S, Park WD, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100:697–712

Xu K, Mackill DJ (1996) A major locus for submergence tolerance mapped on rice chromosome 9. Mol Breeding 2:219–224

Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R, Heuer S, Ismail AM, Mailey-Serres J, Ronald PC, Mackill DJ (2006) *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442:705–708