Fine mapping of the awn gene on chromosome 4 in rice by association and linkage analyses

HU GuangLong[1*], ZHANG DongLing[1*], PAN HuiQiao[1], LI Ben[1], WU JianTao[1], ZHOU XueYa[2], ZHANG QunYuan[1,3], ZHOU Lei[1], YAO GuoXin[1], LI JunZhou[1], LI JinJie[1], ZHANG HongLiang[1†] & LI ZiChao[1†]

[1] *Key Laboratory of Crop Genomics and Genetic Improvement of Ministry of Agriculture, Beijing Key Laboratory of Crop Genetic Improvement, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China;*

[2] *Key Laboratory of Bioinformatics and Bioinformatics Division of Ministry of Education, Tsinghua National Laboratory for Information Sciences and Technology, Department of Automation, Tsinghua University, Beijing 100084, China;*

[3] *Division of Statistical Genomics, Washington University School of Medicine, St Louis, MO, USA*

Awnness is a key trait in rice domestication, yet no studies have been conducted on fine mapping or association mapping of the rice awn gene. In this study, we investigated the awnness and genotype of a core collection of 303 cultivated rice varieties and a BC$_5$F$_2$ segregating population of 200 individuals. Combining association and linkage analyses, we mapped the awnness related genes to chromosome 4. Primary association analysis using 24 SSR markers revealed five loci significantly associated with awnness on chromosome 4. The associated markers cover previously identified regions. Fine association mapping was conducted using another 29 markers within a 4-Mb region, covering the associated marker in34, which is close to the awn gene *Awn4.1*. Seven associated markers were revealed, distributed over an 870-kb region. Combining the fine association mapping and linkage analysis of awnness in the 200 BC$_5$F$_2$ segregating population, we finally identified a 330-kb region as the candidate region for *Awn4.1*. The results indicate that combining association mapping and linkage mapping provides an efficient and precise approach to both genome-wide mapping and fine mapping of rice genes.

rice, awn, association mapping, linkage analysis, fine mapping

Awn is an important trait in rice evolution and production. For example, wild rice has long awns, which are beneficial to seed dispersal and protect rice grains from animal attack. By contrast, most cultivated rice varieties do not have long awns for convenience of harvesting. Efforts to uncover the genetic mechanisms underlying the development of rice awnness began in the 1960s–1970s [1–3]. The results suggest that rice awnness is a complicated trait regulated by multiple genes, and that their expressions are affected by the environment. The first quantitative trait loci (QTLs) for the awn gene were reported in 1963 [2]. Three genes, *An-1*, *An-2* and *An-3* mapped to Chr 3, Chr 4 and Chr 5, respectively, were reported to control various degrees of awning via additive effects [2,4]. It was not until 1999 that the first molecularly mapped awn gene, *an-5(t)*, was mapped to a 33.70–75.70 cM region of Chr 4 by Xiong et al. [5], using an F$_2$ population derived from a cross between wild rice P16 and *indica* Var. Aijiaonante. Thomson et al. [6] mapped an awn gene, *Awn4.1*, in the 0–14.6 cM region of Chr 4. Further major genes/QTLs have been mapped using cross populations. A total of 31 loci were found to be associated with awn presence or awn length in rice (http://www.gra-

mene. org/), and seven of these loci were mapped to chromosome 4 (Table 1). However, none of these QTLs were fine mapped nor identified via association analysis.

Linkage analysis (LA) can only detect recombination events of a few relevant loci with two alleles in a segregating population produced by bi-parental crossing. By contrast, association mapping can detect recombination events accumulated in natural populations during long-term evolution and domestication, and has the advantage of higher mapping resolution, and even using a smaller sample population can be equally as effective [10–13]. Moreover, association mapping can detect the majority of loci controlling the same complex traits and multiple alleles of the same locus. Association mapping was first used to detect alleles in human disease and has now been widely applied to gene mapping in plants [14–21].

In this study, the population used for association mapping consisted of 303 cultivated rice varieties in a rice mini core collection (including 204 Chinese varieties [22] and 99 varieties from other countries, Table S1). The awnness of each variety was evaluated in Sanya (18°09′N) and Hangzhou (29°44′N) in 2006 and in Beijing (39°56′N) in 2009. If over 30% of the florets for any variety had awn in any of the experimental locations, the variety was considered as an awned variety. Eighty awned varieties were observed (Table S1) and 40% of these showed awn in all 3 locations.

Population structure can produce stronger linkage disequilibrium (LD) between non-linked loci, because of geographical origins, local adaptation and breeding history; therefore, association mapping has a higher probability of type I errors than linkage analysis. Thus, we incorporated the individual's membership (the Q value of the subpopulations) as the covariate in the association analysis. The model-based (Bayesian) cluster software STRUCTURE 2.2 [23] was used to estimate the population structure of the 303 rice varieties with 60 unlinked polymorphic markers distributed across all the rice chromosomes. Using a burn-in of 10000 and a run length of 100000, with an admixture model and correlated allele frequencies, each K (the inferred number of subpopulations), from $K=1$ to $K=10$, was run independently ten times. We selected the value of K by investigating the value of $\mathrm{Ln}P(D)$ and the value of ΔK [24] ($\Delta K = m(|L(k+1) - 2L(k) + L(k-1)|)/s[L(k)]$). Our results [25] indicated

that there were two distinctly divergent subpopulations, for which the membership coefficient of each variety, i.e. the Q value, was estimated.

Awnness is a binomial trait; therefore, in the association analysis of awnness, we employed a logistic model: $\mathrm{Ln}(P/1-P) = \beta_0 + \beta_1 Q + \beta_2 M$, where P is the probability of awned varieties, Q is the individual's membership in the first subpopulation, M is the marker's indicative function, and β_0, β_1, and β_2 are unknown parameters, estimated by the maximum likelihood method provided by Proc Logistic in SAS. Most of our SSR markers are multiple alleles; therefore, we transformed the multi-allelic data into biallelic data using Plink (http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml). Meanwhile, we excluded the rare alleles (frequency less than 5%). To control the false positives, we calculated the false discovery rate (FDR) [26] and the adjusted P-value (P_{adj}) using Proc Multtest in SAS and selected the loci with $P_{adj} < 0.05$ as significant associated loci. In addition, we investigated the collinearity of the population structure and the markers using Proc REG in SAS. The results show that the tolerances between population structure and the markers were all above 0.42, much higher than the maximum tolerance (0.2) for multicollinearity to exist [27]. This indicates that the association excludes the influence of multicollinearity.

To reduce the workload, we implemented the association analysis step by step. First we conducted a preliminary association using 24 markers distributed evenly on Chr 4 (Figure 1), and then we increased the marker density around the significantly associated loci obtained in the preliminary scan and carried out a further fine association mapping. In the preliminary association, six loci were significantly associated with rice awnness ($P<0.01$) and five of these exhibited a false discovery rate of less than 0.05 after multiple testing (Figure 1 and Table 2). Our association analysis detected the majority of the QTLs previously reported on Chr 4 (Figure 1). For example, marker in34 was located in the region of *Awn4.1* and close to *qAl4-1*; Rm5320 was mapped in the region of *An5* and close to *An1* and *qAL4-2*. In addition, we found three new loci associated with awnness on Chr 4. This suggests that we can achieve higher efficiency in gene mapping using association mapping than linkage analysis, which requires a segregating population.

Table 1　Loci on rice chromosome 4 associated with awnness (http://www.gramene.org)

Gene code/QTL	Gene/QTL	Alias	Chromosome	Genetic location (cM)	Physical location (cM)
GR:0060048	*AN1*	*An1, Awn-1*	4	65.6–78.4 (–) [2]	22.8–24.7
GR:0060052	*AN5*	*an5, awn-5*	4	33.7–75.7 [5]	20.0–31.6
GR:0061161	*An12*	*An8(t)*, Awn-8*	4	23.2–47.6 [7]	20.1–22.2
AQGD021	*qAL4-2*		4	71.1–95.3 [8]	26.8
AQFE112	*Awn4.1*		4	0–14.6 [6]	6.5–18.5
AQGF014	*qAL4-2*		4	107.7–107.7 [9]	26.8
AQGD020	*qAL4-1*		4	16.8–71.1 [8]	18.8
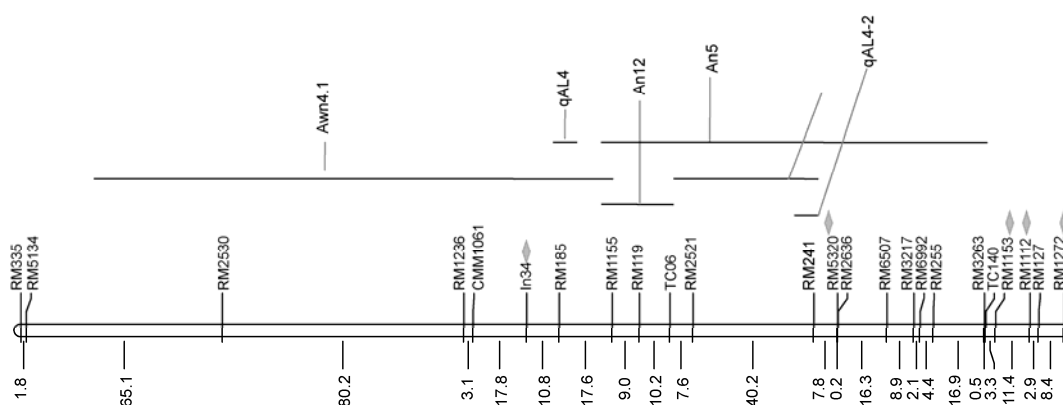
Figure 1   Preliminary association mapping of awnness on chromosome 4. Solid black bars indicate previously identified QTLs (http://www.gramene.org). Physical distances (in 100-kb) between adjacent loci are shown below the chromosome. The diamond marks indicate the associated markers.

**Table 2**   Associated loci detected by logistic regression[a)]

Marker genotype	P_{raw}	FDR	P_{adj}	Tolerance	Odds ratio
in24_137	0.0001	0.00313	0.005	0.92032	3.307
in24_215	0.0001	0.00313	0.005	0.92177	0.262
RM1272_102	0.0001	0.00313	0.005	0.9773	0.359
RM1272_166	0.0001	0.00313	0.005	0.89652	0.252
in34_180	0.0002	0.00417	0.00455	0.96993	0.354
in34_200	0.0002	0.00417	0.00455	0.96993	2.823
in60_200	0.0006	0.01071	0.01071	0.90044	0.184
cmm1157_312	0.001	0.015	0.01563	0.87628	2.586
in60_183	0.0012	0.015	0.01579	0.89772	4.391
RM1153_184	0.0012	0.015	0.01579	0.65651	0.305
in33_172	0.0016	0.01667	0.01739	0.96391	0.412
in33_202	0.0016	0.01667	0.01739	0.96391	2.43
RM1112_108	0.0019	0.01827	0.01827	0.82792	0.321
indel114_136	0.0029	0.02417	0.025	0.42861	3.049
indel114_210	0.0029	0.02417	0.025	0.42861	0.328
in26_158	0.0042	0.02986	0.03182	0.97848	0.442
in26_212	0.0042	0.02986	0.03182	0.97848	2.264
RM5320_101	0.0043	0.02986	0.02986	0.80776	2.343
cmm1157_358	0.0078	0.05132	0.05132	0.96078	0.474
Rm127_500	0.0097	0.06063	0.06063	0.99112	0.239

a) P_{raw}, *P*-value of logistic regression; FDR, false discovery rate; P_{adj}, adjusted *P*-value; Tolerance, tolerance between variables; Odds ratio, point value of odds ratio.

Of the associated loci depicted above, in34 falls in the region of a QTL that was identified in the segregating population BC_5F_2, whose donor and recurrent parents were, respectively, awned Gaoli upland rice and awnless Nipponbare. We conducted fine association mapping in this region by designing 29 additional markers within the 4-Mb region around in34 using the same approach as in the preliminary association mapping. In addition to in34, 6 markers were detected to be significantly associated with awnness ($P<0.01$, $P_{adj}<0.05$) (Table 2, Figure 2). These awnness-associated markers (in60, indel114, cmm1157, in34, in33, in26, and in24; Table 2 and Figure 2) covered an 870-kb fragment, which made it impossible to confirm further which of these seven markers were closest to the awn gene

or to narrow the mapping region via association analysis alone. Consequently, we turned to linkage analysis on a specific segregation population.

In the BC_5F_2 population, the awnness was segregated in the ratio 3 (148 awned plants): 1 (52 awnless plants) ($\chi^2=0.054$). This suggests that the awnness in this BC_5F_2 population is controlled by a dominant gene (Figure 3). Using bulked segregant analysis (BSA) [28], 5 (Rm1236, cmm1061, cmm1304, in60, and in34) of the 53 markers on Chr 4 displayed polymorphism between the awned group and the awnless group. Utilizing MapMaker/EXP3.0 [29], we mapped the awn gene to a 4.5-cM region between in60 and in34, with map distances of 1.2 cM from the former marker and 3.3 cM from the latter (Figure 4). In this region,
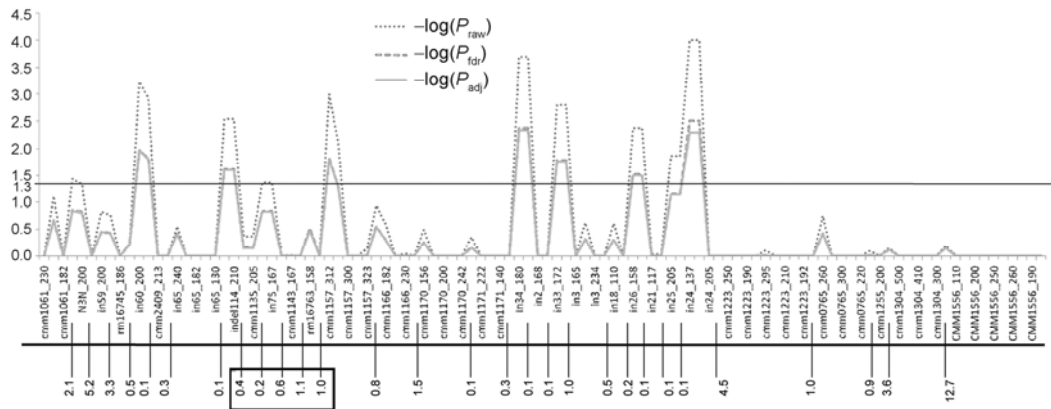
Figure 2 Fine association mapping in the 4-Mb region centered around in34 on Chr 4. Physical distances (in 100-kb) between adjacent loci are shown at the bottom. The number following the "–" in the marker name denotes its genotype. $-\log(P_{raw})$, $-\log(P_{fdr})$, and $-\log(P_{adj})$ are, respectively, the negative logarithms to base 10 of P_{raw}, P_{fdr} and P_{adj}. The black rectangle shows the final candidate region in which A*wn4.1* is located.


Figure 3 The panicle phenotype of Nipponbare (middle), near isogenic awned (left) and awnless (right) lines.

4 markers, including in60, indel114, cmm1157, and in34 are significantly associated with the awnness of rice grains (Figure 2). Linkage and association mapping (illustrated in Figures 2 and 4, respectively) show that indel114 and cmm1157 are closer to the mapped gene than in60 and in34. Thus the candidate gene region could be narrowed down to a 330-kb region between indel114 and cmm1157.

How do we explain the occurrence of significantly associated loci outside the 330-kb candidate region? Is this a false positive association caused by linkage disequilibrium

between markers rather than association with the awn gene? Or does it indicate the existence of other awn-related genes? To address these issues, we analyzed the linkage disequilibrium of seven significantly associated loci (in60, indel114, cmm1157, in34, in33, in26 and in24) using Tassel (http://www.maizegenetics.net/tassel/). Significant pairwise LDs among markers (R^2>0.1, P<0.01) were observed in both *indica* and *japonica* subpopulations (Table 3). This suggests that the association significance of the loci (in60, in34, in33, in24 and in26) outside the candidate region is caused by linkage disequilibrium of neighboring loci and does not imply the existence of further awn genes.

In summary, the logistic regression model, based on population structure, has been shown to be reliable and highly efficient in association analysis of genes regulating complex traits. However, a combination of association and linkage analyses can increase mapping efficiency and accuracy. Association mapping using a natural population can be applied to fine mapping of the results obtained by linkage analysis when a QTL region is large, e.g. 5 cM, and no segregating population or polymorphic markers are available. Additionally, fine association mapping combined with linkage mapping can exclude false positive associated loci caused by high linkage disequilibrium. Our preliminary association analysis detected five loci (in34, RM5320, RM1153, RM1112 and Rm1272) associated with awnness on Chr 4. Further fine mapping combining association and linkage analysis located Awn4.1 within a 330-kb fragment. In addition, multicollinearity analysis of the population
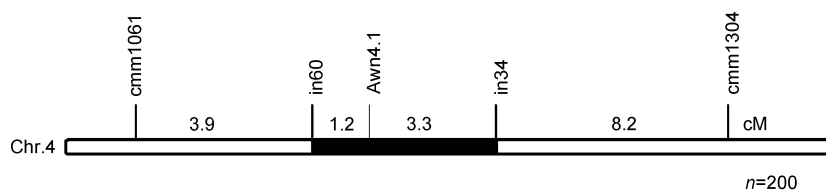
Figure 4 Linkage mapping of *Awn4.1*. Genetic distances (in cM) are shown by the numbers between the markers.

Table 3 Linkage disequilibrium between significantly associated loci

Locus 1	Locus 2	R^2	P	Subpopulation
cmm1157	in34	0.38150726	0	*indica*
cmm1157	in26	0.39275635	0	*indica*
cmm1157	in33	0.40080162	0	*indica*
in26	in24	0.31988456	0.000001	*indica*
in34	in24	0.47181408	0	*indica*
in33	in24	0.53060596	0	*indica*
in33	in26	0.74835585	0	*indica*
in34	in26	0.76392140	0	*indica*
in34	in33	0.96517857	0	*indica*
indel114	in26	0.13799918	0.01580254	*japonica*
indel114	in34	0.18148148	0.00258702	*japonica*
in34	in26	0.92410714	0	*japonica*

structure and markers suggested that the parameter estimation of the logistic regression model utilizing population structure as co-variant is reliable. To reduce the risk of false positives, we also calculated the false discovery rate.

1   Hsieh S C. Bot Bull Acad Sin, 1960, 1: 117–132
2   Nagao S, Takahashi M E. J Fac Agr Hokkaido Univ, 1963, 53: 72–130
3   Iwata N, Omura T. Sci Bull Fac Agr Kyushu Univ, 1971, 25: 137–153
4   Takamure I, Kinoshita T. Rice Genet Newsl, 1991, 8: 98–99
5   Xiong L Z, Liu K D, Dai X K, et al. Theor Appl Genet, 1999, 98: 243–251
6   Thomson M, Tai T, McClung A, et al. Theor Appl Genet, 2003, 107: 479–493
7   Kurakazu T, Yoshimura A. Rice Genet Newsl, 2001, 18: 28–30
8   Gu X Y, Kianian S F, Hareland G A, et al. Theor Appl Genet, 2005, 110: 1108–1118
9   Gu X Y, Kianian S F, Foley M E. Genetics, 2005, 171: 695–704
10 Flint-Garcia S A, Thornsberry J M, Buckler E S. Annu Rev Plant Biol, 2003, 54: 357–374
11 Salvi S, Tuberosa R. Trends Plant Sci, 2005, 10: 297–304
12 Wu R, Zeng Z B. Genetics, 2001, 157: 899–909
13 Yu J, Pressoir G, Briggs W H, et al. Nat Genet, 2006, 38: 203–208
14 Thornsberry J M, Goodman M M, Doebley J, et al. Nat Genet, 2001, 28: 286–289
15 Kraakman A T, Niks R E, van den Berg P M, et al. Genetics, 2004, 168: 435–446
16 Aranzana M J, Kim S, Zhao K, et al. PLoS Genet, 2005, 1: 531–539
17 Mazzucato A, Papa R, Bitocchi E, et al. Theor Appl Genet, 2008, 116: 657–669
18 Zhu C S, Gore M, Buckler E S, et al. Plant Genome, 2008, 1: 5–20
19 Silke S, Grit H, Jean-Baptiste V, et al. Theor Appl Genet, 2009, 118: 259–273
20 Yao J, Wang L X, Liu L H, et al. Genetica, 2009, 137: 67–75
21 Wen W, Mei H, Feng F, et al. Theor Appl Genet, 2009, 119: 459–470
22 Zhang H L, Zhang D L, Wang M X, et al. Theor Appl Genet, 2010, doi: 10.1007/s00122-010-1421-7
23 Pritchard J K, Stephens M, Donnelly P. Genetics, 2000, 155: 945–959
24 Evanno G, Regnaut S, Goudet J. Mol Ecol, 2005, 14: 2611–2620
25 Zhang D L. Dissertation for the Doctoral Degree. Beijing: China Agricultural University, 2007
26 Benjamini Y, Hochberg Y. J Roy Stat Soc Ser B, 1995, 57: 289–300
27 Menard S W. Thousand Oaks, CA: SAGE Publication, Inc., 1995
28 Michelmore R W, Paran I, Kesseli R V. Proc Natl Acad Sci USA, 1991, 88: 9828–9832
29 Lincoln S, Daly M, Lander E. Genomics, 1992, 1: 174–181

Supporting Information

Table S1   PDB IDs for antigen-antibody complexes included in the testing dataset

The supporting information is available online at csb.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.