Heredity (2017) 119, 88–94

www.nature.com/hdy

ORIGINAL ARTICLE

Myosin XI is associated with fitness and adaptation to aridity in wild pearl millet

IS Ousseini[1,2,3,4], Y Bakasso[2], NA Kane[4,5], M Couderc[1], L Zekraoui[1,4,5], C Mariac[1], D Manicacci[6], B Rhoné[1,7], A Barnaud[1,4,5], C Berthouly-Salazar[1,4,5], A Assoumane[2,8], D Moussa[8], T Moussa[8] and Y Vigouroux[1,3,5]

Phenotypic changes in plants can be observed along many environmental gradients and are determined by both environmental and genetic factors. The identification of alleles associated with phenotypic variations is a rapidly developing area of research. We studied the genetic basis of phenotypic variations in 11 populations of wild pearl millet (*Pennisetum glaucum*) on two North-South aridity gradients, one in Niger and one in Mali. Most of the 11 phenotypic traits assessed in a common garden experiment varied between the populations studied. Moreover, the size of the inflorescence, the number of flowers and aboveground dry mass co-varied positively with a decrease in rainfall. To decipher the genetic basis of these phenotypes, we used an association mapping strategy with a mixed model. We found two SNPs on the same myosin XI contig significantly associated with variations in the average number of flowers. Both the allele frequency of the two SNPs and the average number of flowers co-varied with the rainfall gradient on the two gradients. Interestingly, this gene was also a target of selection during domestication. The Myosin XI gene is thus a good candidate for fitness-related adaptation in wild populations.
Heredity (2017) **119,** 88–94; doi:10.1038/hdy.2017.13; published online 15 March 2017

INTRODUCTION

A major goal in evolutionary biology is to understand how phenotypes and genotypes change in response to selection. Adaptation is characterized by a change in the phenotype of individuals in a population towards a phenotype that best fits the present environment (Orr, 2005; Tiffin and Ross-Ibarra, 2014). One important link that remains to be made is how selection on phenotypes relates to genetic and genomic changes (Ehrenreich and Purugganan, 2006). Considerable attention is currently being paid to understanding the genetic basis of adaptation of living organisms to their environment (Pritchard and Di Rienzo, 2010). The study of the phenotypic and genetic consequences of adaptation calls for different sets of approaches (Franks *et al.*, 2014). The study of selection has generally been based on either the genetic and genomic signature of selection or the phenotypic signature, and more rarely both.

The study of phenotype variation along environmental gradients is certainly one of the oldest approaches used to decipher the action of natural selection and its consequences in terms of phenotypic adaptation (Endler, 1986). However, because *in situ* observation cannot distinguish phenotypic variation potentially associated with selection from phenotypic plasticity, common garden or reciprocal transplants are generally used to confirm adaptation (West-Eberhard, 2003; Merilä and Hendry, 2014). With the increasing availability of genomic sequencing, it became possible to directly trace the signature of selection at the molecular level (Nielsen, 2001). These approaches use statistical methods that allow identification of an outlier locus (Lewontin and Krakauer, 1973; Watterson, 1979; Hudson *et al.*, 1987; Tajima, 1989; McDonald and Kreitman, 1991). More recent approaches have been developed based on the allele frequency spectrum (Nielsen *et al.*, 2005), or haplotype homozygosity (Sabeti *et al.*, 2002) but also methods that use environmental data to correlate genetic variation and environmental variables (Sgrò and Hoffmann, 2004; Coop *et al.*, 2010; De Mita *et al.*, 2013; Günther and Coop, 2013; McGaughran *et al.*, 2014). These methods led to the identification of several candidate markers and genes, but one always has to keep in mind that selection signatures could be false positives (Sabeti *et al.*, 2006; Hancock and Di Rienzo, 2008; Pérez O'Brien *et al.*, 2014).

Once selection signatures at the molecular level identify candidate genes, their link to phenotypic variation remains to be demonstrated. Different methods have been developed to identify this link. Two of the available methods to validate genotype/phenotype links are linkage mapping approaches using crosses, that is, QTL mapping (Ehrenreich and Purugganan, 2006) and population association mapping. In association mapping, the genomic region controlling variation is identified using existing populations (Bergelson and Roux, 2010). This approach exploits both historical recombination and the natural diversity built up within populations during the evolutionary history of each species (Yu and Buckler, 2006; Beló *et al.*, 2007). Because it is used on actual populations, it is easier to link to the result of detection of selection signature approaches also based on population diversity.

[1]Institut de Recherche pour le Développement, Montpellier, France; [2]Université Abdou Moumouni de Niamey, Niamey, Niger; [3]Université de Montpellier, Montpellier, France; [4]Institut Sénégalais de la Recherche Agronomique, Campus de Bel Air, Dakar, Sénégal; [5]Laboratoire Mixte International Adaptation des Plantes et Microorganismes Associés aux Stress Environnementaux (LMI LAPSE), Centre de Recherche de Bel Air, Dakar, Sénégal; [6]Université Paris-Sud, UMR 0320 / UMR 8120 Génétique Quantitative et Évolution – Le Moulon, Gif-sur-Yvette, France; [7]Centre National de la Recherche Scientifique, Lyon, France and [8]Institut de Recherche pour le Développement, Niamey, Niger
Correspondence: Dr Y Vigouroux, Institut de Recherche pour le Développement, 911, avenue Agropolis, BP 64501, 34394 Montpellier, cedex 5, France.
E-mail: yves.vigouroux@ird.fr
Received 3 May 2016; revised 9 January 2017; accepted 12 January 2017; published online 15 March 2017

This population association method has been successfully applied to search for genes underlying variations in traits in several plant species, including *Arabidopsis thaliana* (Atwell *et al.*, 2010; Brachi *et al.*, 2010; Li *et al.*, 2010), pearl millet (Saïdou *et al.*, 2009; Mariac *et al.*, 2011), and maize (Remington *et al.*, 2001; Yan *et al.*, 2011; Wallace et al., 2014).

Identifying markers under potential selection, linking their genotype to a phenotype and studying the evolution of this phenotypic trait along an environmental gradient is providing stronger support for ongoing environmental selection (Hoffmann and Willi, 2008), and is a first step toward *in situ* validation.

In this study, we assessed phenotypic variability in a common garden experiment using wild pearl millet (*Pennisetum glaucum*) populations sampled along environmental gradients. Wild pearl millet grows up to the limit of the Sahara, in extreme rainfall and temperature environments. This species is the closest wild relative of a cereal that plays an important role in food security in sub-Saharan Africa. Using an association mapping framework, we assessed the link between phenotypic variability and SNP variation in 181 previously identified selected candidate genes. Among the most interesting genes, a Myosin XI was associated with the number of flowers, and consequently could be related to the adaptation of pearl millet to aridity.

MATERIALS AND METHODS

Samples, field experiments and genetic data

We studied 11 populations of wild pearl millet sampled along a North-South aridity gradient, six populations from Niger and five from Mali (Figure 1, Supplementary Table S1). Phenotypic variations in the 11 populations were evaluated in three different trials in Niger and Senegal in 2013 and 2014. The first two field trials were performed in Senegal and Niger during the rainy season in 2013. The first trial was conducted at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) field station in Sadore (13°14′N, 2°17′E), Niger. The second was conducted at the *Institut Sénégalais de Recherche Agricole* (ISRA) field station in Bambey (14°70′N, -16°47′W) in Senegal. A total of 550 individuals, with 50 per population, were sown in each field trial. Plants were randomized; spacing was 1 × 0.8 m in the trial in Sadore and 1 × 1 m in the trial in Bambey. To avoid side effects, cultivated pearl millet lines were used to border the plots used in both trials. Sowing dates were 15 July 2013 in Sadore and 20 July 2013 in Bambey. The trials were conducted under rainfed conditions with supplementary sprinkler irrigation if necessary. The fungicide Eperon (3.88% Metalaxyl-M+64% Mancozeb) was used at the seedling stage. In 2013, for each plant, five inflorescences from the Sadore experiment were selfed and bulked. These progenies were used for the 2014 field trial also in Sadore. Ten progenies per single selfed plant were sown on 3 July 2014 using the exact same protocol of the experimental conditions of 2013. In all three experiments, we phenotyped 11 traits associated with plant morphology and fitness (Supplementary Table S2, Supplementary Figure S1): time from sowing to heading (HT), average spike length (SLA) estimated on five spikes, average spike diameter (SDA) estimated on five spikes, the number of spikes per individual (SNI) at maturity, average number of involucres per spike (ANIS) estimated on five spikes, main stem length (MSL), main stem diameter (MSD), the number of basal branches, the number of aerial branches (NAB) and dry matter weight (DMW) in grams. The average number of involucres per individual (ANII), was obtained by multiplying the ANIS by the NSI. In the 2014 experiment, the phenotypic values were averaged across progeny.

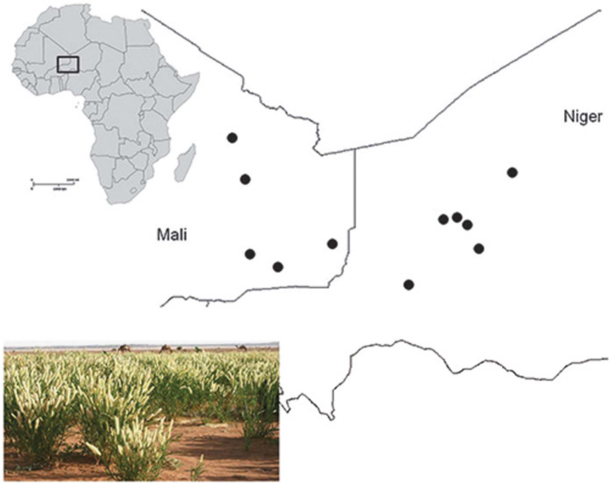

Figure 1 Population sampling site. Each circle corresponds to the sample site of one population. The sampling covered a gradient of aridity from latitude 15 to 19 in Niger and Mali. Population latitude and longitude are in decimal degree. The map of Africa at the top left shows the sampling area in Niger and Mali.

Genotyping. A previous study identified 540 contigs having evidences of signature of selection out of 11 155 contigs (Berthouly-Salazar *et al.*, 2016). Theses contigs were consequently good candidates for genes associated with adaptation. This initial study was done on two extreme populations along the aridity gradient in Niger, and in Mali. The 11 populations we studied here were sampled along the two same gradients but do not include these extreme populations. Genotyping was performed on all the individuals from the 2013 field experiment from the 11 populations. The exact same individuals used in the phenotyping in 2013 were genotyped. We used 181 SNPs data (Supplementary Table S3) derived from 113 contigs showing the strongest evidence of signature of selection (Berthouly-Salazar *et al.*, 2016). We also used 35 SNPs (Supplementary Table S3) randomly drawn from the 35 non-selected contigs (Berthouly-Salazar *et al.*, 2016).

Inference of population structure and genetic relatedness

To infer population structure, we used the genotypic data from 35 random 'neutral' SNP markers from all individuals phenotyped in the 2013 experiments in Niger and Senegal. We used the Bayesian method STRUCTURE with the admixture model, which considers that the genes / alleles of one individual can have different origins (Pritchard *et al.*, 2000; Falush *et al.*, 2003). A total of 100 000 burn-ins and 100 000 iterations were performed per run. We evaluated from 2 to 20 possible clusters (K) with five runs for each K value. We used STRUCTURE HARVESTER (http://taylor0.biology.ucla.edu/structureHarvester/) to handle files and multiple runs (Earl and vonHoldt, 2012), and to calculate the *ad hoc* method based on the second-order rate of change of likelihood (Evanno *et al.*, 2005). We used CLUMMP (Jakobsson and Rosenberg, 2007) and DISTRUCT (Rosenberg, 2004) to represent and to obtain coefficients of ancestry for each individual. Here, coefficient of ancestry is a statistical measure of the proportions of alleles for each individual that could be traced to each of a set of K populations. Genetic relatedness between individuals was calculated using the efficient mixed-model association (EMMA) package (Kang *et al.*, 2008) in R (R Core Team, 2015). We used the IBS Kinship matrix method to calculate this matrix using the 35 random 'neutral' SNP markers (Kang *et al.*, 2008). The matrix was calculated between all individuals without considering the population level (named KINSHIP). But, we also calculated a second matrix considering relatedness inside population only (named KINSHIPpop).

Analysis of the phenotypic data and association mapping

We first investigated variations in the 11 traits across the 11 populations studied.

We used an analysis of variance (AOV) in R (R Core Team, 2015) to test the effects of population and experimental environment on phenotypic variability. We used the model $y_{ijk}=\mu+\alpha_i+\beta_j+\lambda_{ij}+\epsilon_{ijk}$, where y_{ijk} is the phenotype of individual, α is the effect of population, β is the effect of experimental environment and λ is the effect of interaction between α and β. The value ϵ is the residual error effect and μ the grand mean. The indices *i*, *j*, *k* are respectively the population, the experimental environment and the individual studied. We visualized the error distribution of all traits studied using R. We also used a Box-Cox transformation in R to normalize our phenotypic data for new AOV. We compared the AOV results obtained for the original phenotypic data and the transformed data for this experiment.

Next, we investigated how the phenotypic variation was distributed along the two gradients. We extracted 19 climatic variables (Supplementary Table S4) from WORDCLIM (http://www.worldclim.org; Hijmans *et al.*, 2005). We performed a principal component analysis (PCA) using climatic and geographic coordinates to synthesize climatic information (Supplementary Table S4). We then estimated correlations between phenotypic variation and latitude, longitude and the climate data, including axis 1 and axis 2 of the PCA using the Kendall non-parametric method of correlation available in R.

The STRUCTURE software analysis identified 11 clusters; they did not perfectly fit the 11 populations. AOV analysis showed that original population controls (Model 'POPULATION') improved the number of false positive compared to Bayesian clustering (Model 'STRUCTURE'). For that reason, for the identification of SNPs, which are significantly associated with variations in morphological traits, we first extracted residuals of the AOV, which only takes the original population into account. We used these corrected phenotypes (residues) and the matrix of kinship coefficient (K) calculated between all individuals to apply a linear mixed model association (Model 'POPULATION +KINSHIP') via *t*-test with REML estimates (emma.REML.t) from package EMMA (Kang *et al.*, 2008) in R (R Core Team, 2015). To assess robustness of our results to this model choice, we compared the result of our analysis first with a linear mixed model with the matrix of kinship coefficient only (Model 'KINSHIP') and then with a model with a correction for population only (Model 'POPULATION'). We used the Quantile-Quantile (Q-Q) plot method to compare the *P*-values obtained with the four models previously cited and the naïve or 'NULL' model without any corrections. We also assessed significance of our SNPs with another linear mixed model association (Model 'POPULATION+KINSHIPpop'), which used the corrected phenotypes (residues) and the matrix of kinship coefficient considering relatedness inside population only.

We used a false discovery rate (FDR) of 5% for the control of the error rate under multiple testing (Benjamini and Hochberg, 1995). We then also only considered SNPs that had a significant effect (5% FPR threshold) on phenotypes in both the 2013 trials (Sadore+Bambey) and 2014 (Sadore).

For these significant SNPs, we calculated and statistically assessed the correlation of the SNP frequency with latitude separately for the populations from Niger and Mali, and compared the two correlations. We also calculated the correlation for all the SNP frequencies ('neutral' and 'selected candidate') with the latitude and the first axis of the PCA on all the 11 populations using Pearson's method in R (R Core Team, 2015). To assess the significance of correlation coefficients, we compared them to the histogram of distribution of R^2, showing all SNP frequencies with the latitude and the first axis of PCA, respectively. We constructed a histogram of the frequency distribution of the correlation coefficient (R^2) of 35 'neutral' and 216 'selected candidate' SNPs with latitude and the first axis of PCA. We also assessed the significance of each correlation with a *t*-test in R (R Core Team, 2015).

Interesting contigs, that is, that contain SNP(s) with significant association, were BLASTed (BLASTN, http://blast.ncbi.nlm.nih.gov) against the National Center for Biotechnology Information (NCBI) database. We used the nucleotide collection (nr/nt) database and an algorithm for highly similar sequences.

Contig nucleotide sequences (Supplementary Table S5) were translated using the ExPASy server (http://web.expasy.org/translate/; Gasteiger *et al.*, 2003). Polymorphism changes were assessed for synonymous or non-synonymous substitution.

RESULTS

Phenotypic analysis

With lower rainfall and higher temperature, we observed key phenotypic changes in wild pearl millet populations. In the Niger and Senegal trials in 2013, there were 430 and 204 survivors, respectively. In the 2013 experiments, significant effects of population ($P<0.001$) and strong environment effects ($P<0.001$) were detected for all traits except the number of aerial branches (Supplementary Table S6-2013). Less variance was explained by the interaction between environment and population than by direct population and environment effects ($F_{1,\ 718}=2.27$ to 9.58, $P<0.001$). No significant P × E interaction was found for heading date, main stem and spike diameter (Supplementary Table S6). With this AOV, we also found that the distribution of the errors for all traits studied were rather similar to a normal distribution (Supplementary Figure S2). Nevertheless, using a Box-Cox transformation gave the same results (Supplementary Table S7). A total of 260 progenies had enough seed to be included in the 2014 experiment. In the 2014 experiment, we still detected a significant population effect on heading date, average number of involucres per spike, spike length and diameter, and main stem length ($P<0.01$). The lack of significance for the other traits may be due to the smaller number of individuals available in the 2014 experiment (260 compared to 634 in 2013), which reduced statistical power.

We found that the plants tended to flower early, reduce seed production and overall dry mass at higher latitudes, with lower rainfall and higher temperatures. Most of the traits showed a significant negative correlation with latitude in both 2013 and 2014 (Supplementary Table S8). Similar results were obtained in the Bambey experiment. In all three experiments, the strongest and negative correlations were found for average spike length.

Correlations between phenotypes and longitude were less significant (Supplementary Table S9). Almost no significant correlations were found in the 2014 experiment. Only heading date, number of spikes per individual and dry matter weight were found to be significantly correlated to longitude in both Sadore and Bambey in 2013; but heading date was positively correlated in Sadore and negatively correlated in Bambey.

These correlation studies showed that the observed phenotypic variation is mainly explained by the effect of latitude (Supplementary Table S8). The effect of longitude was much less pronounced (Supplementary Table S8).

The first axis of the PCA (Supplementary Table S10) mainly explained rainfall variables and to a lesser extent temperature seasonality. The rainfall variables that contributed to the formation of this axis were annual precipitation, precipitation in the wettest month and the wettest quarter (Supplementary Table S10). The first axis explained 59% of the variance and the second axis 25%. In the 2013 experiment, the first axis of the PCA revealed significant positive correlations among all the traits studied with the exception of heading date, which was correlated with the second axis of the PCA (Supplementary Table S11 and S12). However in the 2014 experiment, a positive correlation was detected for heading date with first axis of PCA (Supplementary Table S10). In all three experiments, the strongest correlation was found for average spike length. Overall, the second axis of PCA explained smaller effects (Supplementary Table S11 and S12).

Analysis of the genetic structure of wild pearl millet populations

A total of 35 random 'neutral' SNP markers on 634 individuals were used for population structure studies. The second order rate of change of the likelihood showed a maximum for K = 11 (Supplementary Figure S3), supporting 11 clusters as one of the possible scenarios. At K = 2, the two major clusters mainly separated the two gradients (one from Mali and the other from Niger; Supplementary Figure S4). Individuals from Niger tended to be in the first cluster (Proportion correctly assigned) and individuals from Mali in the second cluster (Proportion correctly assigned). From K = 3 to K = 11, different populations appeared mainly corresponding to populations as sampling units. At K = 11, the clusters tended to correspond to the sampled populations, but individual coefficient of ancestry was noisy. Indeed, inside a population, each individual was not perfectly assigned to its respective population (Supplementary Figure S4).

Association study

The control of population structure was mostly captured by the population of origin rather than by inferring the coefficient of ancestry. Both the individual coefficient of ancestry (Model STRUCTURE) and the individual population of origin (Model 'POPULATION') corrected genotype/phenotype associations better than a naive or 'NULL' model (Figure 2, Supplementary Figure S5), but we found that for most traits, control was better using the population of origin (Supplementary Figure S5). Very few traits, including the number of basal branches, did not covary with population structure and consequently all models gave very similar results (Supplementary Figure S5). In our final association study model, we consequently used the population of origin rather than inferred coefficient of ancestry. We also added a kinship matrix between individuals to the statistical analysis (Model 'POPULATION+KINSHIP'). The Q-Q plot distribution for all five models also showed that the model 'POPULATION +KINSHIP' reduced the number of false positives better than the 'NULL','STRUCTURE', 'POPULATION' and 'KINSHIP' model (Supplementary Figure S6). This last model corrected only for kinship between individuals.

We tested 216 SNP associations using model 'POPULATION +KINSHIP' with the 11 phenotypic traits in 2013 and 2014 (Supplementary Table S13a). Using a 5% FPR threshold, we detected a total of 86 and 81 SNPs associated with at least one trait in 2013 and 2014, respectively (Supplementary Table S13a). The traits with the highest number of associations were the number of spikes per individual (15 SNP) in 2013 and the average spike diameter (17 SNPs) in 2014 (Supplementary Table S13a). Considering the two experiments, the number of spikes per individual had the largest number of associations (23 SNPs). Results were similar if we used the KINSHIPpop matrix (Supplementary Table S13d)

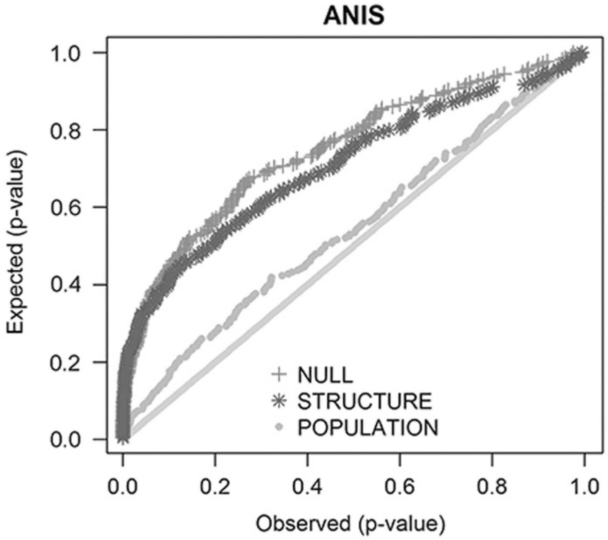

Figure 2 The Q-Q plot method distribution. The quantile-quantile (Q-Q plot) method distribution was shown for a null model, considering structure and considering the original source population of the individual. Phenotype is average number of involucres per spike (ANIS) estimated on five spikes in 2013 experiments. Axes represent the observed *P*-values versus the expected *P*-values. NULL corresponds to the model where there is no correction for structure and the original source population of the individual, STRUCTURE to the model with correction for structure and POPULATION to the model with correction for original source population. The gray line corresponds to distribution of observed *P*-value equal to expected *P*-value. The three different models were performed based on analysis of variance (AOV). The *P*-values were calculated using R.

When we used a FDR of 5%, SNP22 was associated with the length of the main stem, the diameter of the main stem and dry matter weight in 2013, and SNP210 was associated with dry matter weight in the 2014 experiments ($P<0.05$). However, the allele frequency of these SNPs was low, so these associations might be spurious (Supplementary Table S13a). We also found that SNP20 and SNP21 associated significantly with the average number of involucres per spike in both 2013 and 2014 ($P< 0.05$). These two SNPs are located on the same contig and show high LD ($r^2=0.82$). The same two SNPs were also associated with significant *P*-values in analysis using 'POPULATION' or 'KINSHIP' (Supplementary Table S13b, c).

The allele C frequency of SNP20 and SNP21 (Supplementary Table S14) increased ($R^2>0.79$, $P<0.007$) with latitude (Figure 3, Supplementary Table S15) and decreased ($R^2>0.7$, $P<0.001$) with the first axis of PCA (Figure 4, Supplementary Table S15). The R^2 values of SNP20 and SNP21 obtained with the correlation of the latitude and the first axes of PCA were extreme compared to the overall distributions, and fell out of the neutral SNP distribution (Supplementary Figure S7, Supplementary Table S15).

The same two alleles (C and T) were identified for SNP20 and SNP21. For the two SNPs, genotype T/T had a greater average number of involucres per spike than genotype C/C and T/C ($P<0.05$). The average number of involucres per spike of SNP20 was 164.9 (SE ± 4.33) for C/C, 200.8 (SE ± 9.73) for T/C and 273.6 (SE ± 34.40) for T/T in 2013. Similar trends were observed in the two experiments (2013 and 2014) and for the two SNPs (Figure 5, Supplementary Figure S8).

The average number of involucres per spike ($P<0.001$) and frequency of allele T ($P<0.05$) decreased with latitude. So southern individuals had a high probability of having a T/T genotype and consequently a high average number of involucres per spike, whereas northern individuals had a high probability of having a C/C or T/C genotype with the small average number of involucres per spike. When performing a BLASTN against GenBank database, we found that the sequence of the contig carrying both SNPs (Supplementary Table S3) shared 83% identity with the sequence of the gene myosin XI identified in *Oryza brachyantha*. The two mutations are synonymous or located in the 3′untranslated region (3′UTR).

DISCUSSION

More compact phenotype with drier climate

In pearl millet, more compact phenotypes are associated with lower rainfall and higher temperatures during the growing season. Wild pearl millet originating from the drier area flowers earlier has shorter spikes with a smaller diameter, a smaller total number of spikes and spikelets, and a shorter stem with a smaller diameter. In the present study, we found that dry mass decreased with increasing latitude and a drier environment. This change was observed in the two field studies in 2013, but not all the traits showed a statistically significant pattern in 2014. This difference can certainly be explained by the fact that fewer plants were studied in 2014, and we consequently had less statistical power.

Plants typically express phenotypic variation along environmental gradients (Teklehaimanot et al., 1998; Ivancich et al., 2012).

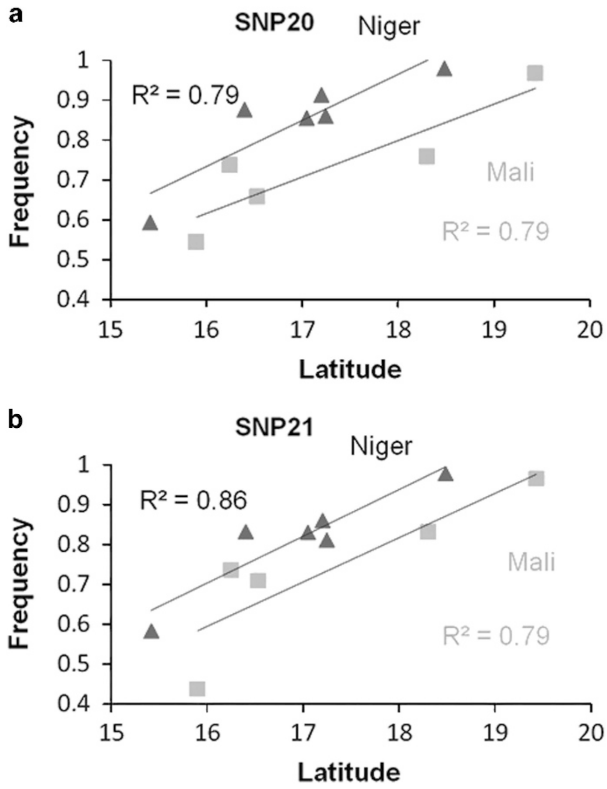

Figure 3 Correlation of SNP frequency with latitude. The correlation of SNP20 (**a**) and SNP21 (**b**) frequency in the populations from Niger and Mali with the latitude. $R^2=0.79$ for SNP20 (Niger, $P<0.05$), 0.79 for SNP20 (Mali, $P<0.05$), 0.86 for SNP21 (Niger, $P<0.05$) and 0.79 for SNP21 (Mali, $P<0.01$).

With the lower rainfall and higher temperatures, as expressed by the first PCA axis, the phenotypic traits of these pearl millet populations allow seeds to be produced with less overall investment in aboveground biomass. This is a well-known trade-off in evolutionary biology for an adaptation to a more stressfull environment (Chapin *et al.*, 1993): rapid initial growth but relatively less production of aboveground biomass. In the present case, if the total number of flowers (and ultimately seed) is changed, it is because the number of flower per inflorescence is changed not the number of inflorescence overall. Producing an inflorescence more rapidly certainly implies producing less flower per inflorescence. So rapid initial growth (of an inflorescence) is certainly key feature for this adaptation. Similar observations have been made in temperate climates, with in this case, a shorter growing season associated with winter and frost (Jonas *et al.*, 2008; McKown *et al.*, 2014). In *Populus trichocarpa*, biomass accumulation and growth rates and ecophysiological traits correlated strongly with latitude, maximum day length and temperature in the area of origin of the tree (McKown *et al.*, 2014). In dryland areas, the height of Eucalyptus trees for a given trunk diameter declines with decreasing rainfall from 2000 to 300 mm and increasing dry season length (Cook *et al.*, 2015). In *Chaetanthera moenchioides* populations derived from the drier gradient showed significantly shorter flowering and fruiting phenology and smaller capitula than the other populations (Bull-Hereñu and Arroyo, 2009). Taken together, these studies suggest that in harsher environments, reduced development and less investment in aboveground biomass may be a widespread strategy (Chapin *et al.*, 1993).

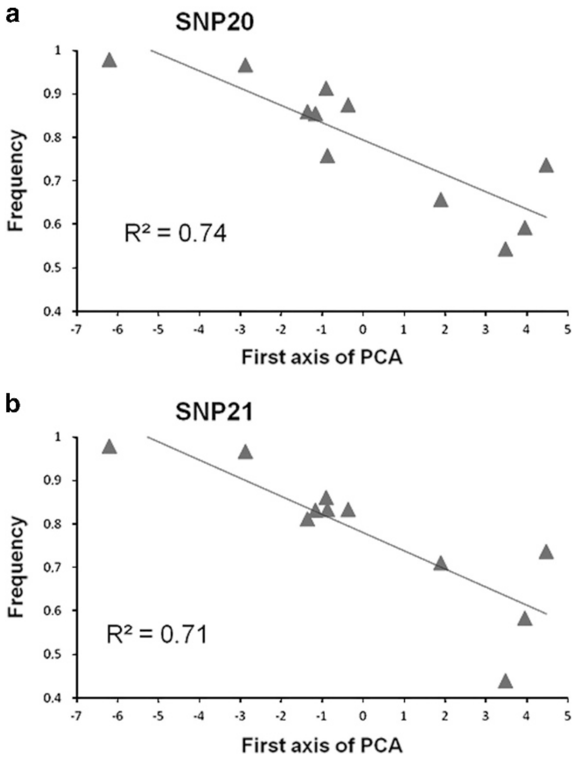

Figure 4 Correlation of SNP associated frequency in the populations with the Axis 1 of PCA. The correlation of SNP20 (**a**) and SNP21 (**b**) allele frequency in the populations with the first axis of PCA notably explained by rainfall variables. $R^2=0.74$ for SNP20 ($P<0.01$) and 0.71 for SNP21 ($P<0.01$).

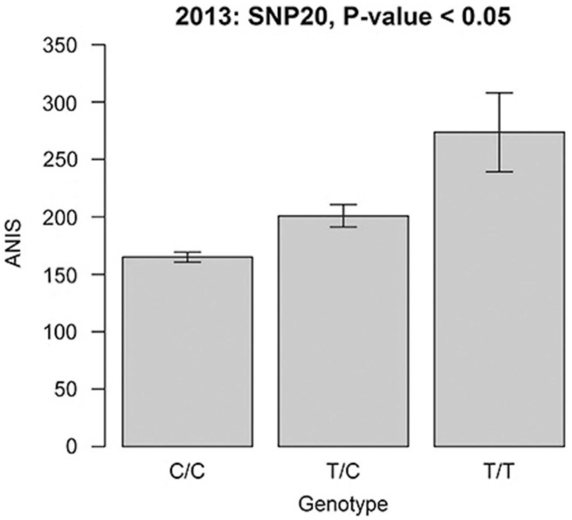

Figure 5 The average number of involucres per spike (ANIS) by genotypes. We presented here the average number of involucres per spike (ANIS) by genotypes (C/C, T/C and T/T) for SNP20. The SNP20 showed a significant effect on ANIS. On the *y* axis, The means value and standard errors of ANIS for each genotype of the SNP20 are given.

In conclusion, we highlighted a decrease in the total number of flowers due to the decrease in the size of the inflorescence. We showed non-significant changes in the number of inflorescences. Among the different phenotypic trade-offs, only a decrease in the size of the inflorescence appears to have been selected.

Gradient, pseudo-replication and association genotype /phenotype
The genetic differentiation between the two gradients was identified by the STRUCTURE analysis. We can therefore consider that the experiment with individual plants from Niger was a pseudo replication of the experiment with individual plants from Mali. Consequently, correlations of the SNP in both gradients are a somewhat natural repetition of evolution against similar climatic conditions.

We also showed that information concerning the original sampling of individuals was better than coefficient of ancestry at controlling genetic structure (Astle and Balding, 2009; Hubisz *et al.*, 2009). This better correction is likely due to the small number of 'neutral' SNPs used here and the consequently poorer estimation of plant individual coefficient of ancestry. Accordingly, the direct use of the population of origin to control for false positive associations is recommended in this particular case (Zhao *et al.*, 2007; Kang *et al.*, 2008; Simko and Hu, 2008). We refined the model to include the population of origin and a kinship matrix as this approach is widely recommended (Kang *et al.*, 2008; Yu *et al.*, 2006; Saïdou *et al.*, 2009; MacKenzie and Hackett, 2011).

Two SNPs (SNP20 and SNP21) on the same contig were shown to be significantly associated with the average number of involucres per spike (ANIS) in both the 2013 and 2014 experiments. Several results suggest that these SNPs are of particular interest. Their frequencies were significantly associated with latitude in both Niger and Mali gradients. Their association was stronger than any other SNP in both the 'neutral' and 'selected candidate' histogram of the distribution of correlation coefficients. We highlighted the fact that the distribution of 'neutral' and 'selected' candidates was similar with the exception of a particular bump in the distribution in the 'selected' candidates with a higher correlation with latitude and with the first axis of the PCA. Those SNPs might be in genes and alleles of interest for the study of adaptation to these gradients. These adaptations may also correspond to phenotypic variations that were not studied here, since they did not show up in our association analyses.

To sum up, we found an association with phenotypic traits, a correlation with the latitude/environment data and we also observed that the average number of involucres per spike (ANIS) in populations decreased significantly with latitude. The sequence surrounding these two SNPs shared 83% identity with a myosin XI gene. The myosin XI is responsible for cytoplasmic streaming and transport of intracellular organelles (Shimmen and Yokota, 2004). When RNA interference was used to specifically silence the myosin XI gene, an effect on tip growth was demonstrated in *Physcomitrella patens* (Vidali *et al.*, 2010). Moreover, the use of transgenic *Arabidopsis thaliana* plants expressing different amounts of myosin XI showed that this gene plays an important role in variations in plant size (Tominaga *et al.*, 2013). The transgenic plants that expressed high- and low-speed moving myosin XI along the actin bundle produced respectively large and small plants compared to the wild control (Tominaga *et al.*, 2013). Although the study of this gene requires further validation, the literature suggests that changes in this gene may actually affect growth.

One SNP is a synonymous mutation and the other a non-coding mutation. We could not rule out the two SNPs might be the real 'causal' polymorphism affecting, for example, expression regulation, or simply being in linkage disequilibrium with real causal SNPs. A recent study of pearl millet showed that the association signal rarely goes further than the genes studied (Saïdou *et al.,* 2014). Consequently, there is a high probability that the causal SNP lies within the myosin XI gene itself. Final validation of the gene might be achieved through a finer study of the region, and/or by functional validation.

A very recent study of pearl millet domestication highlighted a signature of selection associated with domestication around the myosin XI gene (Varshney *et al.,* comm. Pers.). In the cultivated millet, diversity was depleted and strong differentiation was observed between a representative sample of wild and cultivated plants. This result strongly suggests that the polymorphism found in the wild relative was targeted during the domestication of pearl millet. We still do not know for what specific characteristic myosin XI was selected during pearl millet domestication, but the present study suggests that myosin XI is associated with the increase in the number of flowers. The number of flowers is one of the strong and important traits selected during crop domestication, including that of pearl millet.

CONCLUSION

In this study, we have demonstrated that wild pearl millet shows significant phenotypic variation along environmental gradients and that the control of population structure is mostly captured by the population of origin rather than by inferring coefficient of ancestry. We identified two SNPs on the same contig associated with the average number of involucres per spike (ANIS). The sequence of this contig shares 83% identity with the myosin XI gene. The involvement of myosin XI in variations in the average number of involucres per spike could now be validated by functional studies such as the study of variation in the expression of this gene along the environmental gradient.

DATA ARCHIVING

The data is available at the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.mn3g7

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported from a grant from the *Agence Nationale de la Recherche scientifique* to YV and DM (ANR-12-ADAP-0002). ISO was funded by a grant from the French Embassy in Niger and the *Institut de Recherche pour le Développement*. CBS was funded by the Agence Nationale de la Recherche (grant ANR-12-PDOC-0009-01). ISO and AA are funded by the 'jeune équipe associée' à l'IRD (JEAI AVACLI).

AUTHOR CONTRIBUTIONS

YV and DM designed the study; ISO, YB, NAK, MC, LZ, CM, AB, CBS, AA, DM, TM, YV coordinated the different experiments and generated the data; ISO, BR and YV performed the analyses; ISA and YV wrote the manuscript, all the authors commented on the manuscript.

Astle W, Balding DJ (2009). Population structure and cryptic relatedness in genetic association studies. *Stat Sci* **24**: 451–471.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y *et al.* (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**: 627–631.

Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B *et al.* (2007). Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. *Mol Genet Genomics* **279**: 1–10.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* **57**: 289–300.

Bergelson J, Roux F (2010). Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat Rev Genet* **11**: 867–879.

Berthouly-Salazar C, Thuillet A-C, Rhoné B, Mariac C, Ousseini IS, Couderc M *et al.* (2016). Genome scan reveals selection acting on genes linked to stress response in wild pearl millet. *Mol Ecol.* **25**: 5500–5512.

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M *et al.* (2010). Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLoS Genet* **6**: e1000940.

Bull-Hereñu K, Arroyo MTK (2009). Phenological and morphological differentiation in annual Chaetanthera moenchioides (Asteraceae) over an aridity gradient. *Plant Syst Evol* **278**: 159–167.

Chapin FSI, Autumn K, Pugnaire F (1993). Evolution of suites of traits in response to environmental stress. *Am Naturalist* **142**: S78–S92.

Cook GD, Liedloff AC, Cuff NJ, Brocklehurst PS, Williams RJ (2015). Stocks and dynamics of carbon in trees across a rainfall gradient in a tropical savanna. *Austral Ecol* **40**: 845–856.

Coop G, Witonsky D, Rienzo AD, Pritchard JK (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**: 1411–1423.

De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J *et al.* (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol Ecol* **22**: 1383–1399.

Earl DA, von Holdt BM (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* **4**: 359–361.

Ehrenreich IM, Purugganan MD (2006). The molecular genetic basis of plant adaptation. *Am J Bot* **93**: 953–962.

Endler JA (1986). *Natural Selection in the Wild*. Princeton University Press: Princeton, NJ, USA.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol* **14**: 2611–2620.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Franks SJ, Weber JJ, Aitken SN (2014). Evolutionary and plastic responses to climate change in terrestrial plant populations. *Evol Appl* **7**: 123–139.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**: 3784–3788.

Günther T, Coop G (2013). Robust identification of local adaptation from allele frequencies. *Genetics* **195**: 205–220.

Hancock AM, Di Rienzo A (2008). *Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data*. Social Science Research Network: Rochester, NY, USA.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005). Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* **25**: 1965–1978.

Hoffmann AA, Willi Y (2008). Detecting genetic responses to environmental change. *Nat Rev Genet* **9**: 421–432.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009). Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* **9**: 1322–1332.

Hudson RR, Kreitman M, Aguadé M (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.

Ivancich HS, Lencinas MV, Pastur GJM, Esteban RMS, Hernández L, Lindstrom I (2012). Foliar anatomical and morphological variation in Nothofagus pumilio seedlings under controlled irradiance and soil moisture levels. *Tree Physiol* **32**: 554–564.

Jakobsson M, Rosenberg NA (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinforma Oxf Engl* **23**: 1801–1806.

Jonas T, Rixen C, Sturm M, Stoeckli V (2008). How alpine plant growth is linked to snow cover and climate variability. *J Geophys Res Biogeosciences* **113**: G03013.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al.* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**: 1709–1723.

Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**: 175–195.

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO (2010). Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proc Natl Acad Sci* **107**: 21199–21204.

MacKenzie K, Hackett CA (2011). Association mapping in a simulated barley population. *Euphytica* **183**: 337–347.

Mariac C, Jehin L, Saïdou A-A, Thuillet A-C, Couderc M, Sire P *et al.* (2011). Genetic basis of pearl millet adaptation along an environmental gradient investigated by a combination of genome scan and association mapping. *Mol Ecol* **20**: 80–91.

McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652–654.

McGaughran A, Morgan K, Sommer RJ (2014). Environmental variables explain genetic structure in a beetle-associated nematode. *PLoS One* **9**: e87317.

McKown AD, Guy RD, Klápště J, Geraldes A, Friedmann M, Cronk QCB *et al.* (2014). Geographical and environmental gradients shape phenotypic trait variation and genetic structure in Populus trichocarpa. *New Phytol* **201**: 1263–1276.

Merilä J, Hendry AP (2014). Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. *Evol Appl* **7**: 1–14.

Nielsen R (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**: 641–647.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005). Genomic scans for selective sweeps using SNP data. *Genome Res* **15**: 1566–1575.

Orr HA (2005). The genetic theory of adaptation: a brief history. *Nat Rev Genet* **6**: 119–127.

Pérez O'Brien AM, Utsunomiya YT, Mészáros G, Bickhart DM, Liu GE, Van Tassell CP *et al.* (2014). Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. *Genet Sel Evol GSE* **46**: 19.

Pritchard JK, Di Rienzo AD (2010). Adaptation – not by sweeps alone. *Nat Rev Genet* **11**: 665–667.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

R Core Team (2015). R: A language and environment for statistical computing. Vienna, Austria. Available at https://www.r-project.org.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J *et al.* (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci* **98**: 11479–11484.

Rosenberg NA (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**: 137–138.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O *et al.* (2006). Positive natural selection in the human lineage. *Science* **312**: 1614–1620.

Saïdou A-A, Mariac C, Luong V, Pham J-L, Bezançon G, Vigouroux Y (2009). Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. *Genetics* **182**: 899–910.

Saïdou A-A, Clotault J, Couderc M, Mariac C, Devos KM, Thuillet A-C *et al.* (2014). Association mapping, patterns of linkage disequilibrium and selection in the vicinity of the PHYTOCHROME C gene in pearl millet. *Theor Appl Genet* **127**: 19–32.

Sgrò CM, Hoffmann AA (2004). Genetic correlations, tradeoffs and environmental variation. *Heredity* **93**: 241–248.

Shimmen T, Yokota E (2004). Cytoplasmic streaming in plants. *Curr Opin Cell Biol* **16**: 68–72.

Simko I, Hu J (2008). Population structure in cultivated lettuce and its impact on association mapping. *J Am Soc Hortic Sci* **133**: 61–68.

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

Teklehaimanot Z, Lanek J, Tomlinson HF (1998). Provenance variation in morphology and leaflet anatomy of Parkia biglobosa and its relation to drought tolerance. *Trees* **13**: 96–102.

Tiffin P, Ross-Ibarra J (2014). Advances and limits of using population genetics to understand local adaptation. *Trends Ecol Evol* **29**: 673–680.

Tominaga M, Kimura A, Yokota E, Haraguchi T, Shimmen T, Yamamoto K *et al.* (2013). Cytoplasmic streaming velocity as a plant size determinant. *Dev Cell* **27**: 345–352.

Vidali L, Burkart GM, Augustine RC, Kerdavid E, Tüzel E, Bezanilla M (2010). Myosin XI is essential for tip growth in *Physcomitrella patens*. *Plant Cell* **22**: 1868–1882.

Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES (2014). Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet* **10**: e1004845.

Watterson GA (1979). Estimating and testing selection: tshe two-alleles, genic selection diffusion model. *Adv Appl Probab* **11**: 14–30.

West-Eberhard MJ (2003). *Developmental plasticity and evolution*. Oxford University Press: New York, NY, USA.

Yan J, Warburton M, Crouch J (2011). Association mapping for enhancing maize (L.) genetic improvement. *Crop Sci* **51**: 433.

Yu J, Buckler ES (2006). Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol* **17**: 155–160.

Yu J, Pressoir G, Briggs WH, Vroh BiI, Yamasaki M, Doebley JF *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C *et al.* (2007). An arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**: e4.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)