Haplotypes Phased from Population Transcriptomes Detecting Selection in the Initial Adaptation of Miscanthus lutarioriparius to Stressful Environments

Cai-yun Zhu, Wei Liu, Li-Fang Kang, Qin Xu, Shi-Lai Xing, Yang-Yang Fan, Zhi-Hong Song, Juan Yan,* Jian-Qiang Li, and Tao Sang*

Abstract
Adaptation is a characteristic that enhances the survival or reproduction of organisms; selection is the critical process leading to adaptive evolution. Therefore, detecting selection is important in studying evolutionary biology. Changes in allele frequency are fundamental to adaptive evolution. The allele frequency of entire genes at the genomic scale is more intensive and precise for analyzing selection effects, compared with simple sequence repeat and single nucleotide polymorphism (SNP) alleles from nuclear gene fragments. Here, we analyzed 29,094 SNPs derived from 80 individuals of 14 Miscanthus lutarioriparius L. Liou ex S.L. Chen & Renvoize populations planted near their native habitat (Jiangxia, Hubei Province, JH) and a stressful environment (Qingyang, Gansu Province, QG) to detect selection during initial adaptation. The nucleotide diversity of over 60% of genes was decreased in QG compared with JH, suggesting that most genes were undergoing selection in the stressful environment. We explored a new approach based on haplotype data inferred from RNA-seq data to analyze the change in frequency between two sites and to detect selection signals. In total, 402 and 51 genes were found to be targets of positive and negative selection, respectively. Among these candidate genes, the enrichment of abiotic stress-response genes and photosynthesis-related genes might have been responsible for establishment in the stressful environment. This is the first study assessing the change in allele frequency at the genomic level during adaptation. The method in which allele frequency detects selection during initial adaptation using population RNA-seq data would be useful for developing evolutionary biology.

Core Ideas
- A new method based on allele frequency to detect selection is proposed.
- Haplotypes were inferred from transcriptomes of 80 individuals.
- In total, 401 and 52 genes were targets of positive and negative selection respectively.
- Abiotic-related and photosynthesis-related genes were enriched in targeted genes.

A D A P T A T I O N is a characteristic that enhances the survival or reproduction of organisms and thus evolutionary biologists always focus on studies on the critical resources of adaptation that have evolved by selection. For example, quantitative trait loci, genome scans of population diversity and divergence, and the ratio of

Published in Plant Genome
Volume 10. doi: 10.3835/plantgenome2016.11.0119

© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
This is an open access article distributed under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
nonsynonymous to synonymous substitutions have been implemented in the discovery of loci involved in local adaptation (Hamblin et al., 2004; Vonholdt et al., 2010; Shimada et al., 2011; Jones et al., 2012; Ágren et al., 2013; Axelsson et al., 2013; Koenig et al., 2013). Actually, selection is the critical mechanism leading to the evolution of adaptation (Simpson, 1994; Futuyma, 2009). Therefore, it is critically important to choose a comparative method to detect selection and then to shed light on the adaptive process in new ecological environments.

Single nucleotide polymorphisms are among the most widely used genetic markers used to analyze selection and underlying process of adaptation. Certainly, the phased haplotype route has more advantages than SNPs because genetic variation is likely to be transmitted as a linkage unit by structuring into a haplotype, which is formed from the combined effect of mutation, genetic drift, selection, migration, and recombination (Clark, 2004; Crawford and Nickerson, 2005). However, the haplotype-based statistical methods based on linkage disequilibrium or population differentiation between populations or species are limited to species with available reference genomes, and genetic differentiation that is supposed to be large enough to be detected between two populations of interest.

Populations adapt to novel environments through selection on standing variations or new mutations. Adaptation from standing variations is likely to be faster than that from new mutations because beneficial alleles are already available and they usually start at a higher frequency (Innan and Kim, 2004). Many previous studies of quantitative genetics have established the pervasiveness of standing variation in natural populations, which predicts short-term responses to selection (Hill, 1982; Roff, 2007). With recent advances in next-generation sequencing and the accompanying bioinformatic tollkits, it is possible to sensitively detect a slight change in preexisting genetic variation and the signal of selection for nonmodel organisms, even at early stages of adaptation to a new environment (Metzker, 2010). The population transcriptome has been applied for elucidating the slight genetic differences underlying the adaptation of plant species to new habitats (Guggisberg et al., 2013; Bedada et al., 2014), but few studies have sought to understand what and how selective effects change the standing genetic variation for initial adaptation to the new environment. Miscanthus lutarioriparius, an endemic species in warm and wet central China, has been established for 3 yr (one generation) in the semiarid Loess Plateau, which is colder and drier than their native habitat (Yan et al., 2012), which provides good material for studying the selective effect of plants during initial adaptation.

As a C$_4$ perennial grass capable of producing high biomass in cool climates and poor soil conditions, Miscanthus serves as a promising candidate as an energy crop for domestication (Beale and Long, 1995; Naidu et al., 2003; Clifton-Brown et al., 2004; Clifton-Brown et al., 2007; Wang et al., 2008). In our previous field study, M. lutarioriparius was determined to be the wild progenitor for domestication, producing a higher biomass and photosynthetic rate than Miscanthus sacchariflorus (Maxim.) Hack. and Miscanthus sinensis Andersson in the domestic site (QG) and the site near the native habitat (JH) (Yan et al., 2012; Liu et al., 2012; Yan et al., 2015). For detecting the selection signals and candidate genes of M. lutarioriparius accurately during early-stage adaptation, we used SNPs and haplotypes phased from transcriptomic data to characterize the transcriptome-wide genetic variation and genetic differentiation of 80 individuals from 14 populations planted in JH and QG (Xu et al., 2015).

In the garden design of this study, JH is near the native habitat and the whole environment is similar to the collection sites of M. lutarioriparius. It is reasonable to assume that genetic drift mainly contributed to the change in allele frequency between the native habitat and JH. In contrast, the domestic site at QG, located in the Loess Plateau of China, is a stressful environment that is much colder and drier than JH. The altitudes of JH and QG are 45 and 1 258 m asl, respectively. The average annual temperature was 16.7 and 9.3°C in JH and QG, respectively. The average annual precipitation was 1319 and 556.5 mm in JH and QG, respectively. This environmental stress can shape selective pressure and cause decreased genetic diversity within species (Bijlsma and Loeschcke, 2013). Thus we hypothesized that individuals in QG experienced stronger selective pressure than those in JH, and that both selection and genetic drift have shaped allele frequency in the QG population. Assuming that the effect of genetic drift is irrelevant to environmental differences, we could treat the allele frequency in JH as a “control” site and expect to detect the signature of selection through a comparison between QG and JH to eliminate the effect of random factors as much as possible. We proposed a new haplotype frequency-based approach for the first time, using haplotype obtained from RNA-seq data to detect the signature of selection on standing genetic variation during early-stage adaptation. It would be feasible to apply this approach to assess selection during initial adaptation in widespread nonmodel species.

**Materials and Methods**

**Sampling and Qualifying SNPs**

Mature seeds were collected from 14 populations of M. lutarioriparius, representing the main distributional range of the endemic species along the Yangtze River (Xu et al., 2015). In April 2009, the seeds of these populations were grown into seedlings and around 15 seedlings from each population were planted in two field sites, one near the native habitat in JH and the other in QG, located in the Loess Plateau of China where energy crop domestication has been planned (Yan et al., 2012). Plants that grew to the end of the next season after planting were all robust and considered to be established in terms of energy crop production. After the growing season of 2010, 156 out of total 180 individuals survived and became established in JH, whereas 138 of 198 individuals survival and became
established in QG. The survival rate in QG decreased about 12% compared with that in JH (Supplemental Fig. S1A). Furthermore, in 2010, 93.4% of plants flowered and produced seeds in JH, but the flowering rate was only 7.2% and none of the plants produced seeds in QG (Supplemental Fig. S1B). Therefore, the population in QG was considered to be under selection during the adaptation process, and we focused on the selection imposed by the harsh environment on the initial adaption of *M. lutarioriparius* in this study. Furthermore, a previous study (Yan et al., 2012) found that the height and biomass in QG were both significantly higher than that in JH (Supplemental Fig. S1C, D), which indicated that the growth of plants can be shaped by the environment.

The growing season in QG was about 1 mo later than that in JH. Thus, in June and July of 2012, leaves from the same developmental stages were collected from individuals planted in these two experimental fields and stored in liquid N for RNA-seq analysis. To represent the genetic characteristics of the populations as much as possible, we used a random sampling strategy to select samples for RNA-seq analysis. All individuals in each field were labeled. All individuals of *M. lutarioriparius* in each field site were labeled. Three individuals of each *M. lutarioriparius* population were randomly sampled for RNA-seq analysis. Thus 84 individuals from 14 populations were sampled from two field sites for RNA-seq analysis. Because of the poor quality of sequencing, the reads of four samples (JS6, WQ4, YP3, and ZJ1) were excluded for further analysis. Detailed information about the sequenced samples can be found in Xu et al. (2015). Thus RNA-seq data were obtained from 80 individuals sampled from both fields. With the high quality reference transcriptome assembled with the custom pipeline Population RNA-seq to Assemble Reference Transcriptome (Xu et al., 2015), SNPs were identified using SAMtools (Li et al., 2009) after mapping short reads of each sample to the reference transcriptome with the Burrows–Wheeler Aligner. After removing SNPs with quality score of <10, ≥10% missing data, and a minor allele frequency of <0.01, a total of 29,094 SNPs were found among the 80 individuals and they were located on 7429 assembled transcripts. Transcripts with sites that were heterozygous in all individuals were supposed to be from paralogs and were excluded during the assembly processes (Xu et al., 2015).

Detecting Selection on the Basis of SNPs

Forty individuals sampled from each one field site were treated as a single population on the basis of the previous finding that a population structure with the species was lacking (Xu et al., 2015). Nucleotide diversity and expected heterozygosity (*H*<sub>e</sub>) within each of the two populations (sites) were evaluated in a custom Perl script (Nei, 1978). The index of genetic differentiation (*F*<sub>ST</sub>) between the two populations was estimated. Single nucleotide polymorphisms were plotted with the corresponding *H*<sub>e</sub> and *F*<sub>ST</sub> values to evaluate whether a SNP might have experienced selection in QG.

We also used BayeScan version 2.1 to detect the signature of selection (Foll and Gaggiotti, 2008). BayeScan is a Bayesian method using the *F*<sub>ST</sub> coefficient to identify candidate loci under selection and is robust to complex demographic scenarios (Foll and Gaggiotti, 2008). The *F*<sub>ST</sub> coefficient is decomposed into a population-specific component (*β*) shared by all loci and a locus-specific component (*α*) shared by all populations with a logistic regression. When *α* is significantly different from 0, the corresponding locus is supposed to be a candidate locus. The Q-value is used for multiple-testing correction to control the false positive rate in BayeScan. BayeScan was implemented with SNPs from two field sites with a thinning interval of 10 and 100,000 iterations plus 50,000 iterations for burn-in.

Inferring Haplotypes

We attempted to detect the signature of selection based on the change in allele frequency from JH to QG by taking advantage of population transcriptome data. At first, it was necessary to construct alleles by determining the SNP combination within an allele or a haplotype. We used PHASE version 2.1.1 (http://stephenslab.uchicago.edu/phase/download.html, accessed 16 June 2017) to analyze the transcriptome data of the 80 individuals and infer haplotypes (Stephens and Donnelly, 2003). PHASE is based on the approximate coalescent prior, which considers that the gene sequence of a mutant offspring differs only slightly from the progenitor sequence, as changes in haplotypes only occur through recombination and mutation. Thus unresolved haplotypes tend to be similar to the known haplotypes. Both homozygous and heterozygous positions are taken into account when considering whether close-matching haplotypes are sought. PHASE was originally applied to genome sequences. Here, we adopted PHASE for analyzing the population transcriptome data with previously assembled reference-quality transcripts (Xu et al., 2015), because *Miscanthus* is a non-model species without a genome reference.

Considering the high heterozygosity of *M. lutarioriparius*, we inferred the haplotype for 7122 transcripts with no more than 10 SNPs (95.7% of 7429 transcripts with SNPs) to ensure the accuracy of inference. For each transcript, SNPs of all 80 individuals were pooled as input data to be ordered in PHASE according to their relative positions on the transcript. The new model, which makes explicit allowance for recombination, was chosen for haplotype reconstruction with the default value for the initial estimate of the background recombination parameter being 0.0004. We carried out a preperation for 100 random selected genes with 200, 500, and 1000 iterations plus another 200 iterations for burn-in, and found that the best reconstruction group of this combination of parameters was the same as that with the default set. Therefore, we implemented PHASE with the default number of iterations (100) plus another 100 iterations for burn-in and a thinning interval of 1 to obtain the haplotype group with the highest probability.
For each gene, the inferred haplotypes can be considered as alleles of the gene. Thus the number of alleles could be estimated for each gene whose transcripts were identified in the transcriptomes. The change in frequency of these alleles from the population grown in JH to that in QG could be used as an indication of the presence of selection at the loci.

**Detecting Selection on the Basis of Alleles (Haplotypes)**

For an allele of any given gene, its presence or absence in each individual fits a Bernoulli distribution. Our sampling strategy involved individuals being randomly sampled from 14 populations, suggesting that samples from one site can be considered to be independent. Because no obvious population structure was observed across all individuals, 40 individuals in one field site were considered to come from one large population (Xu et al., 2015). Thus we assumed that the probability of the presence of a given allele was the same among individuals in one field site. Therefore, the number times a given allele was present in 40 individuals fitted a binomial distribution. The null hypothesis was that the number of times a give allele was likely to be present was equal in both field sites. Thus we could test the hypothesis using the binomial test to compare the allele frequency of each allele in JH and QG.

A one-tailed binomial test was performed with R version 3.0.1 (R Core Development Team, 2013) and a 99.9% confidence interval was used after false discovery rate correction for multiple testing. To obtain a powerful binomial test, alleles with a frequency less than 0.1 or higher than 0.9 in either field were not included in the test or in further evaluations. Alleles with a significantly increased frequency in QG compared with JH were denoted AI, whereas alleles with significantly decreased frequency in QG compared with JH were denoted AD. Alleles with increased and decreased frequency might have experienced positive and negative selection when plants were transplanted from their native habit to the Loess Plateau. Genes containing alleles with a significantly changed frequency between JH and QG were divided into three categories: genes with only AI (GI), genes with only AD (GD), and genes with both AI and AD (GB).

Because the change in gene frequency could also result from random factors including drift, we made the follow corrections to avoid falsely detecting positive selection. The population in QG had a lower survival rate than that in JH (JH: 85%; QG: 73%); consequently, it had a lower π overall (Yan et al., 2012; Xu et al., 2015). Genes with a higher π in QG than in JH were not included in the selection test because they were likely to be affected by random factors.

The filtered candidate genes were used for following analyses. We compared the distributions of genetic differentiation ($F_{ST}$) of SNPs distributed in GI, GD, and GB plus the total pool (genes included in the previous phasing process) using a one-tailed nonparametric Wilcoxon rank-sum test. We randomly sampled a certain amount of SNPs from all loci 10,000 times, and the result was the same regarding the number of sites distributed across GI, GD, and GB separately. Then the same test and resampling times were used to compare the $F_{ST}$ distribution of loci between GB and GI, although the sample size was the same as the number of loci in GB because there were many more SNPs in GI. The comparison between GD and GI was implemented in the same way. The test between GB and GD was performed with the all original data for the number of loci distributed across GB and GD. The difference in $F_{ST}$ among the three gene categories and the total pool may be helpful for illustrating the selection acting on these candidate genes. We also compared the distributions of $F_{ST}$ among GI, GD, GB, and the total pool, in which nonsynonymous loci and synonymous loci were treated separately.

Furthermore, we annotated the corresponding protein sequences (derived from a previous study; Xu et al., 2015) of GI via functional enrichment analysis and the Pfam database (http://pfam.xfam.org/, accessed 16 June 2017) using Fisher’s exact test. Functional annotation was not performed for the GB genes because of the undetermined direction of selection, or for the GD genes because of the small dataset. We also compared the three gene classes with genes detected on the basis of the performance of SNPs ($F_{ST}$ and $H_π$).

**Results**

**Genetic Variation of *M. lutarioriparius* in JH and QG**

A total of 29,094 SNPs located in 7429 transcripts were found among 80 individuals of *M. lutarioriparius*, with the number of SNPs in a gene ranging from 1 to 31 (Xu et al., 2015). Overall, 7112 (95.7%) genes with no more than 10 SNPs were subjected to further analysis for detecting selection. For each gene, the π of the populations in JH and QG was calculated and compared between the two field sites (Fig. 1). In comparison to JH, π in QG was decreased and increased for 61.4 and 37.0% of genes, respectively. The $H_π$ of SNPs were calculated and compared between the JH and QG values. The mean $H_π$ in QG (0.29) was significantly lower than that in JH (0.31) (paired t-test, $P < 2.2 \times 10^{-16}$), with 56% having higher $H_π$ in JH than in QG. This result is consistent with the trend of a decrease in π from JH to QG (Fig. 1). We then calculated the $F_{ST}$ between the two field sites for each SNP. A total of 1454 and 291 SNPs had $F_{ST}$ values ranked in the top 5 and 1%, respectively.

**The SNP-Based Signature of Selection**

If we consider $H_π$ and $F_{ST}$ together, the distribution of the differences in $H_π$ and $F_{ST}$ values are shown in Fig. 2. As a result, 948 SNPs on 761 genes and 187 SNPs on 155 genes were found to have the top 5% and 1% $F_{ST}$ values, respectively, and decreased $H_π$ in QG.

In total, 376 genes were detected as candidate genes under selection with a Q-value threshold of 0.001. Over
30% candidate genes derived from the method based on $F_{ST}$ (top 5%) and $H_e$ were also detected with BayeScan.

We further examined the function of genes derived on the basis of SNPs with the top 5% and 1% $F_{ST}$ values and decreased $H_e$ in QG. Here, 9.09% of the genes that were found to have the top 1% $F_{ST}$ values and decreased $H_e$ in QG were concentrated to six Pfam families with more than three members. Three Pfam families were involved in the response to stress, including the RNA recognition motif, the bZIP transcription factor, and glutathione S-transferase. In addition, 29.28% of the genes found to have the top 5% $F_{ST}$ scores and decreased $H_e$ in QG were concentrated to 45 Pfam families with more than three members; 22.22% of the 45 Pfam families were responsible for stress resistance, such as the RNA recognition motif, the WD domain, the DnaJ domain, the EF-hand domain, the bZIP transcription factor, and glutathione S-transferase.

Fig. 1. Distribution of nucleotide diversity ($\pi$) for Miscanthus genes in two sites. (A) Distribution of $\pi$ in Jiangxia, Hubei Province (JH) and Qingyang, Gansu Province (QG), showed that most genes were surrounded by the red square line ($y = x$), indicating that major genes had similar diversity in two sites. (B) Distribution of the $\pi$ ratio of genes. The $\pi$ ratio was estimated by dividing $\pi$ in QG by that in JH, indicating that most genes had a reduced diversity after being transplanted in QG.

Fig. 2. Three-dimensional bar plot of proportion of Miscanthus genes in a specified index of genetic differentiation ($F_{ST}$) range and a specified expected heterozygosity ($H_e$) difference range. The red line shows that Jiangxia, Hubei Province (JH) and Qingyang, Gansu Province (QG) had the same expected heterozygosity.
The Allele-Based Signature of Selection

We found that the number of alleles reached 80 after PHASE analysis for genes with no more than 10 SNPs (Fig. 3). The counts of SNPs and alleles for each gene had a significant correlation ($r^2 = 0.91, P < 2.2 \times 10^{-16}$). The number of SNPs and alleles within a gene ranged from 1 to 10, and 2 to 80, respectively. The average number of SNPs and alleles was 3.5 and 10 per gene, respectively. About 33% of the genes had one SNP and two alleles, and about 20% of the genes had more than 15 alleles. The alleles of 7112 genes (95.7% of total genes) of 80 individuals in the two populations were inferred with PHASE version 2.1.1.

To evaluate the change in allele frequency from JH to QG, the ratio of the allele frequency between QG and JH was calculated for all 72,014 alleles. The ratios ordered from the lowest to the highest are shown in Fig. 4. Of these alleles, 17.9% alleles had an unchanged frequency. About 41.5 and 40.6% of alleles had an increased and decreased frequency from JH to QG, respectively. A total of 11,072 alleles of 5717 genes with a frequency between 0.1 and 0.9 were subjected to the one-tailed binomial test. Of these, 624 alleles (5.6%) and 398 alleles (3.6%) had a significantly increased or decreased frequency from JH to QG ($P \leq 0.001$).

The average frequency of 11,072 alleles in JH was 0.37. The average frequency of AI and AD alleles was 0.36, and 0.52, respectively (Table 1). The distribution of AI was similar to the overall allele distribution, with a large portion of alleles with a frequency between 0.1 and 0.3. For AD, the peak allele frequency distribution was between 0.3 and 0.7, within which more than 70% of alleles fell.

Genes containing AI or AD alleles could have experienced positive or negative selection in QG compared with JH. Genes containing AI, AD, or both were classified as GI, GD, and GB, with the number of each type being 428, 208, and 186, respectively. After removing genes with increased $\pi$ in QG, which was possibly caused by random factors, 402 GI, 51 GD, and 139 GB genes were obtained. Of these, 66.42% GI, 65.93% GD, and 95.89% GB had no more than 10 alleles (Fig. 5). Each gene in GI and GD had at least three alleles, and 56.12% genes in GB had two alleles. The allele frequency in GI, GD, GB, and the total pool were plotted in Fig. 6, which shows that AI and AD had a greater difference in allele frequency between the two field sites (insignificant alleles were distributed around the diagonal line).
Zhu et al.: Detecting Selection for Early-Stage Adaptation of Miscanthus

G. I. Table 1. The frequency distribution of classified alleles in Miscanthus plants from Jiangxia, Hubei Province (JH).

<table>
<thead>
<tr>
<th>Allele frequency</th>
<th>Total (11,072)†</th>
<th>AI (624)‡</th>
<th>AD (398)§§</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.1, 0.2)</td>
<td>27.93%</td>
<td>28.21%</td>
<td>0%</td>
</tr>
<tr>
<td>(0.2, 0.3)</td>
<td>20.24%</td>
<td>18.27%</td>
<td>8.79%</td>
</tr>
<tr>
<td>(0.3, 0.4)</td>
<td>15.26%</td>
<td>12.50%</td>
<td>27.89%</td>
</tr>
<tr>
<td>(0.4, 0.5)</td>
<td>12.27%</td>
<td>14.26%</td>
<td>21.11%</td>
</tr>
<tr>
<td>(0.5, 0.6)</td>
<td>7.89%</td>
<td>13.46%</td>
<td>10.55%</td>
</tr>
<tr>
<td>(0.6, 0.7)</td>
<td>6.66%</td>
<td>11.54%</td>
<td>11.31%</td>
</tr>
<tr>
<td>(0.7, 0.8)</td>
<td>5.79%</td>
<td>1.78%</td>
<td>8.79%</td>
</tr>
<tr>
<td>(0.8, 0.9)</td>
<td>3.96%</td>
<td>0%</td>
<td>11.56%</td>
</tr>
</tbody>
</table>

† All alleles from genes with 10 SNPs at most were filtered according to their frequencies. Only alleles with frequencies ranging from 0.1 to 0.9 in both sites were used for further analysis and considered as the total.
‡ AI, alleles with a significantly increased frequency in Qingyang, Gansu Province (QG) compared with JH (binomial test, \( P < 0.001 \)).
§ AD, alleles with a significantly decreased frequency in QG compared with JH (binomial test, \( P < 0.001 \)).
¶ Counts of each kind of alleles are shown in parentheses in the column heads.

To compare the genetic differences among loci located in GI, GD, and GB, we conducted a one-tailed nonparametric Wilcoxon rank-sum test for the \( F_{ST} \) of these three categories of genes. The \( F_{ST} \) distributions of 1576, 212, and 246 SNPs were obtained for GI, GD, and GB, which were compared with those generated from 10,000 samples of the same number of SNPs from the total of 22,430 SNPs. The tests indicated that all three categories of genes had significantly different distributions leaning toward higher \( F_{ST} \) values, with all comparisons having \( P < 0.01 \) (Fig. 7). Among the SNPs located in 5717 genes retained in the SNP-based PHASE analysis, synonymous loci and nonsynonymous loci accounted for 51.69% (9,580) and 48.31% (8,953), respectively. The relative proportion of synonymous SNPs in 402 GI, 51 GD, and 139 GB were 51.38 (6,727), 47.15 (91), and 56.60% (120), respectively. Both synonymous SNPs and nonsynonymous SNPs in GI, GD, and GB had a significantly higher level of genetic difference than the total pool, suggesting that our allele-based method was appropriate for detecting potential candidate genes that are undergoing selection between two distinct environments (Table 2). No significant difference was observed between synonymous loci and nonsynonymous loci located in GI, GD, GB, and the total pool, which was reasonable, considering that Miscanthus is still at the initial adaptation stage.

The functional categories of GI genes were compared with those of the total pool. It was found that GI was enriched in nine Pfam families \((P < 0.05)\), four of which were involved in abiotic stress response, including glutathione S-transferase, the bZIP transcription factor, the zinc knuckle, and topoisomerase II-associated protein PAT1. Leucine-rich repeat (LRR) and reversibly glycosylated polypeptides were associated with biotic resistance. Chlorophyll ab binding protein and the pentatricopeptide repeat family (PPR) are involved in photosynthesis, and PPR is also related to restoring fertility. The remaining family was ribosomal protein S2.

Discussion

Genetic Variation on the Basis of SNPs

According to a previous study (Xu et al., 2015), a high-quality transcriptome reference was assembled based on the population RNA-seq data. The average sequence coverage of 80 individuals of Miscanthus was 60.4%. Only about 8.4% regions of the leaf reference transcriptome for Miscanthus were sequenced in all individuals, suggesting the completeness of the assembled reference (Xu et al., 2015). The high-quality transcriptome reference provided an accurate template for SNP calling for each sample of Miscanthus.

A radical change in environmental conditions can reduce the survival and reproduction of species, which may affect genetic variation in response to selection (Fitter and Hay, 2012; Shaw and Etterson, 2012). For example, after Miscanthus was transplanted to QG (the domestication site) from its native habitat, the survival rate of the species decreased about 12%, and the flowering rate was only 7.2% in QG compared with 93.4% in JH (Yan et al., 2012). Given the much colder and drier conditions in QG than in JH (near the native habitat), individuals carrying some alleles failed to survive, resulting in a decrease in nucleotide diversity for most genes in QG, which resulted in decreased \( H_e \) in QG compared with that in JH for 56% of SNPs.

Wright’s \( F_{ST} \) indicates genetic differentiation among populations and is often used as an index to screen potential SNPs undergoing selection (Wright, 1951). In
In our study, genes consisting of SNPs with relatively high $F_{st}$ and decreased $H_e$ may be targets of selection, which might be related to drought- and cold- tolerance. Indeed, for 155 genes with SNPs with the top 1% $F_{st}$ scores and decreased $H_e$ in QG, only 9.09% were concentrated to six Pfam families with more than three members. These three Pfam families were detected to be involved in stress resistance, including glutathione S-transferase, the bZIP transcription factor, and the RNA recognition motif. Genes belonging to glutathione S-transferase and the bZIP transcription factor could be involved in stress resistance associated with the cold and dry climate and poor soil conditions in the Loess Plateau (Dixon et al., 2002; Jakoby et al., 2002). Proteins with the RNA recognition motif are supposed to be associated with plants' ability to respond to environmental stimuli, including abiotic and biotic stress (Lorkovic, 2009; Kang and Kwak, 2012; Kang et al., 2013).

For genes with SNPs with the top 5% $F_{st}$ and decreased $H_e$ in QG compared with JH, 29.28% of the genes were concentrated to 45 Pfam families with more than three members. As well as the three Pfam families described above, another seven Pfam families were detected to be involved in stress resistance, including 2OG-Fe(II) oxygenase, the EF-hand domain, the DnaJ domain, the LRR, the RING finger, the Zinc knuckle, and the WD domain. 2OG-Fe(II) oxygenase might respond to intensified UV radiation at the higher altitude (Sakamoto et al., 2004). Genes containing the EF-hand domain could have contributed to effective control of stomatal opening for higher water use efficiency and more adequate C source transportation during the increased photosynthetic rate (Williams et al., 2000; Day et al.,...
The DnaJ domain is associated with the hsp70 heat-shock system, which functions to protect proteins from irreversible aggregation during synthesis and cells from thermal or oxidative stress, which is beneficial for *M. lutarioriparius* photosynthesizing under intensified UV radiation at the higher elevation (Qiu et al., 2006). The LRR was associated with plant disease resistance. Many proteins containing a RING finger can mediate the activity of ubiquitin ligase, which is important in regulating abscisic acid signals transduction (Joazeiro and Weissman, 2000). Additionally, despite the functional diversity of the Zinc knuckle and the WD domain, some proteins embracing these domains have been detected to respond to plant abiotic stress such as drought and nutrient stress, which is important for *M. lutarioriparius* to adapt to the harsh environment in the Loess Plateau (Laity et al., 2001; Xiong et al., 2005; Kim et al., 2010; Xu and Min, 2011; Mishra et al., 2012).

The Allele-Based Signature of Selection

Genotypes only provide information at each SNP locus separately, whereas the combined information at multiple sites (alleles) is unknown, and patterns of alleles will obviously change even when selection is acting on multiple sites (Ferrer-Admetlla et al., 2014). According to a previous study, for genes with a low level of genetic diversity, there were no significant correlations between the expression level and genetic diversity in two field sites (Xu et al., 2015). Furthermore, the expression data were extracted to calculate the correlation between the change in genetic diversity and the change in expression level between the two field sites. We found that the change in expression level was not significantly correlated with the change in genetic diversity ($r^2 = -0.23, P = 0.053$). This result suggested that minor allele frequency would not be affected by expression levels in this study. Thus we compared the allele frequency between the two field sites to detect targets of selection. Through comparisons of allele frequency between the two field sites, 624 AI and 398 AD alleles were identified to be targets of positive selection and negative selection, respectively. The number of AI alleles was nearly twice the number of AD alleles, which may be attributed to AD alleles being part of genes undergoing negative selection, because the stronger negative genes were eliminated when some individuals failed to survive. Considering that genes with decreased nucleotide diversity in the stressful environment might be driven by both selection and genetic drift, more information was needed to assess the relative effects of selection and genetic drift during adaptation. Thus, 402 GI and 51 GD genes were identified as candidates of positive selection and negative selection, respectively, by filtering genes with increased nucleotide diversity in QG; 139 GB genes were also supposed to be under selection but the direction is unknown for now.

With the strict reference transcriptome obtained from a previous study (Xu et al., 2015), we also compared the functional categories of the genes between GI and total genes. The GI genes were enriched in nine Pfam
families, eight of which were related to abiotic and biotic stress response, and photosynthesis, which may contribute to the higher photosynthesis and biomass of *M. lutarioriparius* in QG than in JH (Yan et al., 2012, 2015; Xu et al., 2015). Four Pfam families have been described above, including glutathione S-transferase, the bZIP transcription factor, the Zinc knuckle and topoisomerase II-associated protein PAT1. Additionally, topoisomerase II-associated protein PAT1 was found to be a potential interactor of abscisic acid deficient 2, an enzyme in the abscisic acid biosynthesis pathway, thus participating in responding to abiotic stress (Qin et al., 2011; Zhang, 2013). These gene families could be involved in stress resistance associated with the cold and dry climate and poor soil conditions in the Loess Plateau. Reversibly glycosylated polypeptides were shown to be plasmodesmal-associated proteins, and the overexpression of plasmodesmal proteins has proven to result in a cell-to-cell transport block, evoking the plant defense response (Zavaliev et al., 2010). Chlorophyll ab-binding protein is the light-harvesting antenna, and has been proven to be involved in pigment biosynthesis or in assembly of the thylakoid membrane, which are vital for photosynthesis (Jansson, 1994; Horton et al., 1996). Recent evidence shows that PPR proteins are involved in the processing of chloroplast RNA, and can function in stabilizing fertility (Brown et al., 2003; Nakamura et al., 2003). These gene families are beneficial for improving photosynthesis efficiency and biomass in the Loess Plateau (Yan et al., 2012).

Selection and drift are involved in shaping the genetic variation of a population during the process of adaptation to a new environment. After eliminating the effect of drift as much as possible, the signal of selection can be detected in genes. We found that the *F*<sub>ST</sub> of SNPs in the three genes categories were significantly higher than that in the total pool, which indicates that selection pressure applied on these genes has given rise to the significant skewing of allele frequency. No significant difference was observed between synonymous and non-synonymous loci, indicating selection at the early stage of adaptation has not left a detectable signature in a single locus. Difference among the classified genes and the total pool are predicted to be larger until populations achieve the optimum adaptive peak.

**Comparison of Methods based on SNPs and Alleles**

The initial strong negative selection has led to the decreased survival rate in QG, but still, the new approach based on alleles can detect genes undergoing weak negative selection. The direction of selection also can be inferred, which is essential for directing the future domestication of *M. lutarioriparius* but the direction of selection is insoluble if based on SNPs only. Furthermore, among the genes derived from SNPs with the 5% top *F*<sub>ST</sub> scores and decreased *H*<sub>E</sub>, 196, 10, and 55 genes (accounting for 34.3% of 761 genes) were observed to overlap with 402 GI, 51 GD, and 139 GB genes, respectively. The corresponding counts were 56, 0, and 16 genes (accounting for 46.45% of 155 genes) were observed in 402 GI, 51 GD, and 139 GB genes, considering the 1% top *F*<sub>ST</sub> scores and decreased *H*<sub>E</sub>. These data indicate the limited adaptability inferred by using the method based on *F*<sub>ST</sub> scores for capturing the signature of selection, especially for populations with relatively low *F*<sub>ST</sub> values. Therefore, the new approach based on alleles from a population is supposed to be much better than the method based on SNPs in our study, but having 34.3 and 46.45% overlapping genes for different thresholds and the complementary functional annotation also provided a way to improve the performance of this new method.

**Conclusions**

In this study, we described an effective approach based on allele data inferred from RNA-seq data for detecting the signals of selection in a stressful environment, and detected a total of 402 and 51 genes as targets of positive and negative selection, respectively. Among these candidate genes, the enrichment in abiotic stress-response genes and photosynthesis-related genes might have been responsible for the successful establishment of *M. lutarioriparius* in QG. This is the first study trying to detect multiple genes possibly responding to selection during initial adaptation by assessing the change in allele frequency. An effective approach based on population RNA-seq data would be useful for developing evolutionary biology. Adaptation can be shaped by many factors, such as selection and epigenetic mechanisms (Richards et al., 2010), and other mechanisms underlying the adaptability of *M. lutarioriparius* await further investigation when the genome reference becomes available.

**Supplemental Information**

The datasets supporting the results of this article are available in the Dryad Digital Repository at http://datadryad.org/doi/short/10.5061/dryad.qc54k (accessed 13 June 2017) before the formal DOI (10.5061/dryad.qc54k) is fully registered with the DOI system.

**Conflict of Interest Disclosure**

The authors declare that there is no conflict of interest.

**Acknowledgments**

The work was financed by the National Natural Science Foundation of China (Nos 31400284 and 31500186), the National Key Research and Development Program of China (No. 2016YFC0500905), and the Science and Technology Service Network Initiative of the Chinese Academy of Sciences (KFJ-EW-STS-119). The authors thank the Beijing Center for Physical and Chemical Analysis for generating the sequencing data and the Beijing Computing Center for assisting with computational infrastructure for data analysis.

**References**


Axelsson, E., A. Ratnakumar, M.L. Arendt, K. Maqbool, M.T. Webster, M. Perloski, et al. 2013. The genomic signature of dog domestication...


Zhang, Y. 2013. Characterization of abscisic acid deficient 2 (ABA2) and its protein interactors in Arabidopsis thaliana. MSc, thesis, Univ. of Toronto, Toronto, ON.