DR. TAO SANG (Orcid ID : 0000-0003-0111-9065)

Article type : Original Article

Article subject: Population and Conservation Genetics

Population transcriptomic characterization of the genetic and expression variation of a candidate progenitor of *Miscanthus* energy crops

JUAN YAN,*[1] ZHIHONG SONG,†‡[1] QIN XU,† LIFANG KANG,† CAIYUN ZHU,‡§ SHILAI XING,‡§ WEI LIU,§ JOSEF GREIMLER,¶ TOBIAS ZÜST□, JIANQIANG LI*[2] and TAO SANG,†§[2]

*Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China;

†Key Laboratory of Plant Resources and Beijing Botanical Garden, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China;

‡University of Chinese Academy of Sciences, Beijing 100049, China

§State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China;

¶University of Vienna, Department of Botany and Biodiversity Research, Rennweg 14, Vienna 1030, Austria;

□ Institute of Plant Sciences, University of Bern, Altenbergrain 21, Bern 3013, Switzerland

[1]These two authors contributed equally to this work.

[2] Correspondence: Jianqiang Li. fax: +86-27-87510251, E-mail: lijq@wbgcas.cn

Tao Sang. fax: +86-10-62590843, E-mail: sang@ibcas.ac.cn

Keywords: *Miscanthus*, transcriptome, gene expression, nucleotide diversity, population genetics

Running title: POPULATION TRANSCRIPTOMICS OF MISCANTHUS

Abstract

The use of transcriptome data in the study of the population genetics of a species can capture faint signals of both genetic variation and expression variation, and can provide a broad picture of a species' genomic response to environmental conditions. In this study, we characterized the genetic and expression diversity of *Miscanthus lutarioriparius* by comparing more than 16,225 transcripts obtained from 78 individuals, belonging to 10 populations distributed across the species' entire geographic range. We only observed a low level of nucleotide diversity ($\pi = 0.000434$) among the transcriptome data of these populations, which is consistent with highly conserved sequences of functional elements and protein-coding genes captured with this method. Tests of population divergence using the transcriptome data were consistent with previous microsatellite data but proved to be more sensitive, particularly if gene expression variation was considered as well. For example, the analysis of expression data showed that genes involved in photosynthetic processes and responses to temperature or reactive oxygen species stimuli were significantly enriched in certain populations. This differential gene expression was primarily observed among populations and not within populations. Interestingly, nucleotide diversity was significantly negatively correlated with expression diversity within populations, while this correlation was positive among populations. This suggests that genetic and expression variation play separate roles in adaptation and population persistence. Combining analyses of genetic and gene expression variation represents a promising approach for studying the population genetics of wild species, and may uncover both adaptive and non-adaptive processes.

Introduction

Rapid developments in sequencing technologies are beginning to offer powerful tools for studying genomic patterns in non-model organisms at the population level, and promise to provide new insights into the fundamental processes underlying genetic differentiation and adaptation of species. Starting from the late 1960s, non-coding or anonymous gene markers have been used to analyze genetic data for populations of species with the goal of understanding demographic processes such as mutation-selection balance, bottlenecks and migrations *etc*. (Slatkin 1987; Orr 1998; Hedrick 2009). Beyond these classical applications, the need to study process of adaptation and adaptive evolution has resulted in the development of novel method for population genetics. For example, new genomic and transcriptomic methods provide population geneticists with tools to more accurately characterize patterns of variation in natural populations and study the ecological and evolutionary processes that underlie species adaptations (Hartl & Clark 2006; Barrett & Schluter 2008; Stapley *et al*. 2010; Ekblom & Galindo 2011; Lamichhaney *et al.* 2012; Tine *et al*. 2014; Pan *et al*. 2017; Bowsher *et al*. 2017).

RNA-Seq is one of the most commonly used next-generation sequencing (NGS) approaches in non-model organisms (Shendure *et al*. 2008; Robertson *et al.* 2010; Grabherr *et al.* 2011). It accurately quantifies the relative levels of each transcript present in a sample by mapping reads to a reference sequence assembly (Morin *et al*. 2008; Harrison *et al.* 2012), yielding high-throughput gene expression data at low effort (Gahlan *et al.* 2012; Annadurai 2012). In addition, single nucleotide polymorphisms (SNP) identified from RNA-Seq data can potentially capture key sequence variation involved in the adaptation of a species to its environment. Because variation in gene expression could similarly be involved in local adaptation, and gene expression levels are at least in part affected by regulatory elements or epigenetic mechanisms, several studies have indicated that gene expression levels may be heritable and be acted upon via natural selection (Ellison *et al*. 2011; Guggisberg *et al*. 2013; Nabholz *et al*. 2014). Variation in genetic expression may thus be altered even before genetic variants arise in the population (West-Eberhard 2003), hence such changes could reflect the early processes that underlie adaptive divergence (e.g., in Oleksiak *et al*. 2002; Derome *et al*. 2006; Jeukens *et al*. 2010).

Transcriptomes represent a cost-effective method to explore the genetic mechanisms that govern ecological interactions and that underlie adaptive divergence at levels of sequences, genes or metabolic pathways in non-model organisms (Alvarez *et al*. 2015). For example, Derome *et al*. (2006) sequenced the transcriptomes of lake whitefish (*Coregonus clupeaformis*) from two separate lakes and found that sixteen genes related to energetic metabolism and regulation of muscle contraction showed true parallelism of transcription, i.e., parallel directional changes in expression in independently evolving populations. Harris *et al*. (2015) compared the transcriptomes of the white-footed mouse (*Peromyscus leucopus*) from urban and rural populations, and found the two populations to differ in 104,655 high-quality SNPs and 65 single sequence repeats (SSRs), as well as having 19 differentially expressed contigs related to the modification of proteins and immune function. These studies provide insights into the divergence of genetic variation and the level of gene expression that may drive the responses to different ecological environments. Therefore, variation in gene sequences and gene expression may be complementary mechanisms for a species to respond to its environment, and each mechanism may have unique roles in governing species adaptation (Whitehead *et al*. 2012). However, few studies to date have investigated the patterns and relationships between genetic variation and gene expression within and among populations across the geographic distribution of a species.

Endemic species provide an ideal opportunity to investigate the patterns and relationships of genetic variation and gene expression diversity in response to dynamic environments as they are restricted to narrowly defining geographic regions. *Miscanthus lutarioriparius* is endemic to central China and found on seasonally flooded river banks, farm lands and ruderal land in cities. The species has a distinct distribution range from *Miscanthus sacchariflorus* and produces thicker, taller tillers (Chen & Renvoize 2005). *M. lutarioriparius* plays a key role in the Yangtze River ecosystem and constitutes the major food source and habitat for many types of birds, mammals, and other animals. *Miscanthus* species are considered promising second-generation energy crops (Jorgensen & Schwarz 2000; Clifton-Brown *et al.* 2004; Somerville *et al.* 2010; Clark *et al*. 2014; 2015), and *M. lutarioriparius* could be developed as a bioenergy resource because of its considerable biomass, presumably adapted to marginal lands, and abundant genetic variation found in natural populations (Yan *et al*. 2012; 2016).

Similar to other *Miscanthus* species, *M. lutarioriparius* can spread clonally from rhizomes, and propagate sexually from seeds (Quinn *et al.* 2010; Nishiwaki *et al.* 2011; Yan *et al.* 2012). Once established, the plant develops strong rhizomes and an intense root system that enhances drought tolerance and carbon sequestration of the ecosystem (Sang & Zhu 2011; Mi *et al*. 2014). Despite this clonality, morphological surveys and physiological measurements in different common gardens at different latitudes have indicated that there is considerable variation in *M. lutarioriparius*, and that there is evidence for local adaptation of plant genotypes (Yan *et al.* 2012; 2015). When *M. lutarioriparius* was transplanted from its native habitat to the Loess Plateau (Fig. 1), the low survival rate in the new environment drastically decreased genetic diversity of the transplanted population, while in the surviving plants the expression of stress-tolerance genes showed significantly increased variation (Fan *et al.* 2015; Xu *et al.* 2015; Xing *et al.* 2016; Song *et al.* 2017). However, while these studies characterized the genetic responses of *M. lutarioriparius* under extremely novel conditions, we still lack information on the performance of the species in its natural populations.

In this study, we used population RNA-Seq data from ten geographically isolated populations of *M. lutarioriparius* to examine the patterns of nucleotide and transcriptional variation across the natural range of this species, and to identify the potential mechanisms underlying the plant's strong adaptive capacity. Additionally, we proposed to test 1) whether transcriptome data may be used as a new tool to explore population genetics via comparisons of diversity patterns obtained from SNP and expression data and 2) whether pairing nucleotide diversity with expression data from the same set of genes is a useful strategy for studying evolutionary and ecological genomics. The results of this study are expected to potentially reveal novel insights on local adaptations and evolutionary processes.

Materials and Methods

Sample collection, cDNA preparation and RNA-Seq

M. lutarioriparius is a riparian plant that can be established in semiarid regions, and which has a fast-growing vegetative stage and high photosynthetic activity during the flood season (from May to August) (Yan *et al*. 2015). Additionally, the species has a simple genetic structure as well as long-distance bidirectional gene flow with certain genetic barriers between populations, as observed using microsatellite data to estimate genetic information for populations across the native distribution range (Yan *et al*. 2016). To identify underlying characteristics of wild populations at the transcriptomic level, we sampled 80 individuals (8 individuals per population) from ten populations representing different habitats in the same region at noon from 18 to 28 June 2013 (Fig. 1). Plants were sampled at regular intervals of 25 m along transects in each population. This minimized the chance of sampling clonal individuals, as the largest clonal root network of *M. lutarioriparius* in these habitats was found to be 21.213 m in diameter (Yan *et al*. 2016). The fourth mature leaf from the apex of each plant was cut and immediately placed into liquid nitrogen for storage until RNA extraction.

Total RNA was isolated from each individual using Trizol reagent (Invitrogen, Carlsbad, California, United States), purified using the RNAeasy Mini RNA Kit (Qiagen, Schnackenburgallee, Hamburg, Germany), quantified with a NanoDrop 1000 instrument (Thermo Scientific) and stored at -80 °C. The mRNA of each individual was isolated from 5 μg of purified total RNA using one round of purification with oligo d (T) beads (Dynabeads®

mRNA Purification Kit, Invitrogen). The cDNA libraries were prepared with the NEBNext Ultra RNA Library Prep Kit for Illumina (New England BioLabs). First-strand cDNA was synthesized using random hexamer-primed reverse transcription. After second-strand cDNA synthesis and adaptor ligation, approximately 450-bp cDNA fragments were isolated using Ampure XP beads (Beckman). The isolated cDNA fragments were amplified in 10 PCR cycles. The concentration and size of the library were assayed using the Agilent 2100 Bioanalyzer (Agilent Technologies) and Qubit® 2.0 fluorometer (Life Technologies), respectively, and 100-bp paired-end sequencing was performed on an Illumina HiSeq 2500.

Pre-processing, quantifying gene expression, and calling SNPs

The RNA-Seq data were filtered and trimmed to control the quality of raw reads using FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit). Because of the unstable sequence content in the first 9 bases and certain sharp declines in quality in the last 11 bases, the data were trimmed for all samples, and a total of 231 Gb of sequence was generated from 2,457,041,020 raw reads (Table S1). Although no reference genome sequence was available for *M. lutarioriparius*, Xu *et al*. (2015) assembled a high-resolution reference transcriptome consisting of 18,503 unigenes based on reference genomes of related species (*Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, and *Zea mays*). Therefore, the trimmed reads of each individual could be mapped to the Bowtie-build index of the reference transcriptome of *M. lutarioriparius* with an N50 of 1425 bp, and the total mapped reads were calculated using TopHat with the default settings (Langmead *et al*. 2009; Trapnell *et al*. 2009). The proportion of the reference transcriptome that was detected in each individual ranged from 55.0% to

87.4%, with an average of 74.5% for the 80 individuals of *M. lutarioriparius*. After eliminating two individuals with low coverage (< 60%), the transcriptomic coverage of remaining individuals (78 samples) subjected to further analysis ranged from 65.1% to 87.4% (Fig. 2).

The expression abundance of each gene was described in terms of the expected fragments per kilobase of transcript per million fragments mapped (FPKM) using Cufflinks (v2.2.1) with the minimum number of fragments needed for new transfrags set at 2 (Trapnell *et al.* 2010). To filter out extreme values, the FPKM values per gene were added and then $\log_2$ transformed (Xu *et al.* 2013). The transformed values were then subjected to a quartile analysis to filter out values greater than 1.5 times the inter-quartile range. After a quartile analysis of the filtered values, 16,225 transcripts were obtained with at least 40 $\log_2$-transformed FPKMs larger than 0.

Hidden paralogous genes might lead to spurious SNPs and an excess of heterozygosity; therefore, the READS2SNP software was used to exclude spurious SNPs (Gayral *et al.* 2013). First, raw data were mapped onto the reference transcriptome using the BWA software (v0.7.12) with mismatch penalty set at 2, gap open penalty set at 6, and gap extension penalty set at 4 (Li & Durbin 2009), and then the aligned reads were sorted and indexed using SAMtools (v1.3.1) with default settings (Li *et al.* 2009). The BAM files for each individual were used to screen putative paralogous loci with a likelihood ratio test (Gayral *et al.* 2013). Candidate SNPs were identified with READS2SNP (Gayral *et al.* 2013) and SAMtools (Li *et al.* 2009). To ensure the accuracy of SNPs identification, only SNPs identified by both methods were kept, and SNPs with a quality score ≤ 10, missing data, or a minor allele frequency (MAF) ≤ 0.05 were deleted. All subsequent analyses were carried out on the remaining high-quality SNPs.

Genetic variation and population structure based on SNP data

Based on the SNP data, the nucleotide diversity per site (π) (Nei 1983), observed heterozygosity (H_O) (Brookfield 1996), Nei's genetic diversity (H_E) (Nei 1983) and genetic differentiation (F_{ST}) (Nei 1973) of the transcripts in each population were calculated using the formulas below: $\pi = \frac{\sum_{i=1}^{n}(\sum_{j=1}^{i} x_i x_j \pi_{ij})}{L}$, where π_{ij} is the proportion of different nucleotides between the *i*th and *j*th sequences, x_i and x_j are the frequencies of these sequences, respectively, and L is the length of the gene sequences; $H_O = \frac{\sum_{i=1}^{m}(1\ if a_{i1} \neq a_{i2})}{n}$, where *m* is the number of individuals in the population and a_{i1} and a_{i2} are the alleles of individual *i* at the target locus; $H_E = 1 - \sum_{i=1}^{n} p_i^2$, where *n* is the number of alleles at the target locus and p_i is the allele frequency of the *i* allele at the target locus; $F_{ST} = \frac{H_T - H_S}{H_T}$, where H_S is based on the expected heterozygosity in subpopulations and H_T is based on the expected heterozygosity for the overall total population (Nei 1973); $H_S = \frac{H_{E1}\ X\ N_1\ +\ H_{E2}\ X\ N_2\ +\cdots+ H_{En}\ X\ N_n}{Nn}$, where H_{En} is the expected heterozygosity of the population *n* and N_n is the number of population *n*; and $H_T = 1 - \sum(p^2 + q^2)$, where *p* and *q* represent the allele frequency of the total population.

The genetic structure of the species based on SNP data was investigated by Principal component analysis (PCA) and Bayesian clustering. The PCA was performed for 78 individuals using the program SMARTPCA of the EIGENSOFT software package (v4.2) (Patterson *et al*. 2006). Bayesian clustering was performed with a series of predefined populations (*K* = 2-10) using STRUCTURE v2.3 (Pritchard *et al*. 2000). Overall, 500,000

Markov chain Monte Carlo (MCMC) iterations were employed after a burn-in period of 100,000 iterations. Each *K* value was repeated five times to assign individuals to clusters with the maximum likelihood. Structure Harvester was used to detect the optimal value of *K* based on the recommended Δ*K* (Earl & Vonholdt 2012).

Mantel tests were performed in IBD 2.0 (Bohonak 2002) with 10,000 permutations between the pair-wise genetic distance [$F_{\mathrm{ST}}/(1-F_{\mathrm{ST}})$] and geographic distance (km) in all populations. We used the nonparametric Analysis of Molecular Variance (AMOVA) to describe the variability within and among populations and loci. A hierarchical analysis of variance and locus-by-locus analysis was performed with ARLEQUIN 3.5.1.2 (Excoffier & Lischer 2010), and significance tests were performed for 1,000 permutations.

Variation in gene expression and population differentiation based on FPKM

Xu *et al*.'s (2015) method was used to evaluate variation in gene expression in the populations. The population gene expression (E_{p}) was calculated as the average FPKM of the individuals sampled from the population using $E_P = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} E_i}{n}$, where *n* represents the number of individuals sampled from the population and E_i represents the FPKM of a given gene of the *i*th individual in the population. The variation of gene expression in the population, which is denoted as the expression diversity (E_d), was calculated using $E_d = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} |E_i - E_p|}{(\mathrm{n}-1)E_p}$.

PCAs and Pearson correlation analyses were performed between individuals to characterize the relationships among the 78 samples based on the normalized FPKMs of each gene. The PCAs were performed in the R 3.2.1 missMDA package (Josse & Husson 2016) using the normalized FPKMs of each gene. The Pearson correlation coefficients among individuals were calculated in the R 3.2.1 Hmisc package using the normalized FPKMs of each gene. Then, to estimate the gene expression relationship among populations (E_p similarity), we calculated Pearson's correlation coefficients based on the average correlation coefficients of individuals and used the method proposed by Wolf *et al*. (2010) for standardization. Mantel tests were performed in IBD 2.0 (Bohonak 2002) with 10,000 permutations between E_p similarity and geographic distance (km) in all populations.

The Kolmogorov-Smirnov test (*K*-*S* test) as implemented in R 3.2.1 was applied to measure the differentially expressed genes (DEGs) among the populations. To further identify genes that were differentially expressed among the 10 populations, one-way ANOVAs were performed for 16,225 genes (transcripts) of the *M. lutarioriparius* transcriptome obtained from the 78 individuals of 10 populations, and the Bonferroni correction was used to adjust each *P* value with a multiplier of 16, 225. During the one-way ANOVA, the requirements of normality and homogeneity of variance were fulfilled. Adjusted *P-values* of 0.05 were used to identify genes that were significantly differentially expressed among populations.

Additionally, we also identified differentially expressed genes among the ten populations using the DESeq package (v1.26.0) because DESeq is for pairwise comparisons, whereas ANOVA is for any number of populations. First, filtered and trimmed reads were aligned to the reference transcriptome using Bowtie2 (Langmead *et al*. 2009) and TopHat (Trapnell *et al*. 2009). Then, Cufflinks, cuffmerge (Trapnell *et al*. 2012), and HTSeq (Anders *et al*. 2015) were used to calculate the read counts. DEGs among populations were then identified according to a false discovery rate (FDR) of 5% (Anders & Huber 2010). The significant DEGs for each pair-wise comparison were analyzed using the functional annotation tools of clusterProfiler in R (Yu *et al*. 2012) at FDR < 0.05.

Detecting the relationship between expression diversity and nucleotide diversity among populations

To detect if the gene expression diversity was consistent with the nucleotide diversity for the 16,225 genes within and among populations, we calculated the correlation coefficients (*r*) of the E_d and π for these genes in each population. Finally, a Pearson's correlation analysis was performed between the median E_d and π of each population to explain the relationship between genetic variation and gene expression diversity. R 3.2.1 was used for the analysis and *P*-values were estimated from 1000 permutations.

Results

Genetic variation and population structure based on SNP data

Overall, 12,285 SNPs from 5,920 transcripts were identified across the 78 individuals. The average nucleotide diversity (π) of the populations was 0.000434 (0.000418 (LU24)-0.000444 (LU5)) (Table 1). The majority of transcripts harbored a low level of genetic variation (Fig. 3A). The π quantiles in each population were similar (Fig. S1A), and the π distributions were similar among the following seven populations according to the *K-S* test: LU10, LU17, LU5, LU7, LU14, LU19, and LU24 (Table S2). The π distributions of LU4 and LU9 were both significantly different from these seven populations, and in addition the distribution of π in LU23 was significantly different from the populations LU7, LU14, and LU19 (Table S2). The genetic diversity (H_O and H_E) of the 10 populations varied from 0.4636 (LU10) to 0.5144 (LU4) and from 0.3596 (LU10) to 0.3704 (LU9), respectively (Table 1). The pair-wise F_{ST} values among the 10 populations ranged from 0.0274 to 0.0578 (Table S3).

The STRUCTURE analysis revealed a consistent increasing trend in the mean lnP(K) value from $K = 2$ to $K = 10$, and low variance was observed across replicate runs. $K = 3$ was found to be the most appropriate K value (ΔK), which suggested that a model with three gene pools captured a major split (Fig. S1B). The PCA showed that the populations could cluster in three groups that corresponded to populations in the western, central, and eastern portions of the Yangtze River (Fig. 3B). Moreover, a significant association was observed between the level of genetic differentiation and geographic distance among the 10 populations by IBD analysis ($r = 0.318$, $P = 0.033$) (Fig. 3C).

A locus-by-locus AMOVA of the SNP data (12,285 loci) showed that only 0.83% of the total genetic variation partitioned in the 10 *M. lutarioriparius* populations was attributed to differences among populations (*d. f.* = 9; $P < 0.0001$), while 99.17% was attributed to the differences within populations (*d. f.* = 150; $P < 0.0001$) (Fig. 3D).

Variation in gene expression within and among populations based on FPKM

For the distribution of expression density, the E_p shapes showed a right-skewed distribution and peak at the expression level intervals of 10-30 FPKM (Fig. 4A), and the *K-S* test showed that 43 pairs (excluding the pairs LU19-LU5 and LU19-LU4) were significantly different

(Table S4). Based on the quantiles of the seven populations, the $\log_2 E_p$ quantile shifted down in LU23 and LU24 and up in LU9 (Fig. S2A). The E_d of these populations also showed a right-skewed distribution and peak at 0.2 - 0.3 (Fig. 4B). The average E_d values of 10 populations ranged from 0.304 (LU24) to 0.426 (LU19) (Fig. S2B). Each pair-wise comparison of the E_d among the 10 populations was significantly different (Table S5).

The PCA analysis of the expression level of the 16,225 genes from all individuals showed that these individuals were distributed to some degree based on their geographic locations (Fig. 4C). Moreover, a significantly negative correlation was observed between the population similarity and geographic distance among the 10 populations (r = -0. 437, P = 0.003) (Fig. 4D). According to the one-way ANOVA, most of variation in the gene expression occurred within populations (69.64%), although the variation range was wider among populations ($0.80e^{-4}$-32.59) than within populations (0.02-26.47). 1,570 genes showed significantly differentiated expression, with the majority of differences occurring among populations (FDR < 0.05) (Fig. 4E). All significant enriched terms for each pair-wise comparison are listed in Table S6.

Relationship between gene expression variations and genetic variations within and among populations

A comparison of variation in nucleotide diversity and gene expression revealed a contrasting pattern within and among populations. At the genes level, the relationship between the E_d and π of 16,225 genes in each population was significantly negatively correlated (r = -0.03 ~ -0.10; P < 0.0001 ~ 0.019) (Fig. 5). At the population level, the relationship between average

E_d and π of each population was significantly positively correlated (*r* = 0.054; *P* = 0.046) (Fig. 6A). Among populations, the expression similarity matrix (Table S7) was significantly negatively correlated with the genetic distance matrix (*r* = -0.576; *P* = 0.001) (Fig. 6B).

Discussion

Transcriptome analysis performed across the natural range of a species can reveal a broad and detailed picture of the species' genomic response to environmental differences (Gibson 2008; Whitehead *et al*. 2012). Our results for *M. lutarioriparius* provide several notable insights. First, variation among transcriptomes revealed population structures consistent with those previously reported from nuclear microsatellites data. Specifically, low levels of genetic differentiation among populations were found by both transcriptome and microsatellite data, suggesting high levels of gene flow or recent historical connections among certain populations (McGlashan *et al.* 2001). Second, differentiation among populations was greater for gene expression data than for nucleotide diversity. Variation in gene expression is affected by the environment but has at least some heritable component (Ellison et al. 2011; Guggisberg et al. 2013; Nabholz et al. 2014). Thus, gene expression data is a compelling complement to SNP markers and has the potential to use as an additional tool to explore population genetics. Moreover, gene expression diversity and nucleotide diversity are significantly negatively correlated within all populations, while they are positively correlated among populations. These differences could provide insights into driving mechanisms underlying plant adaptations to natural habitats.

Based on the population transcriptomes, *M. lutarioriparius* had a lower nucleotide diversity ($\pi = 0.000434$) than other crop grasses, such as *Z. mays* (Wright *et al*. 2005; Hufford *et al*. 2012) or *Pennisetum glaucum* (Clotault *et al*. 2012), though similarly low nucleotide diversity has been reported in *Oryza barthii* (Nabholz *et al*. 2014). The alternate measure of genetic diversity H_{E} was also lower than the diversity estimated by microsatellite analyses across the entire distribution range ($H_{\mathrm{E\ Transcriptome}}$: 0.3596-0.3704 vs. $H_{\mathrm{E\ Microsatellites}}$: 0.682-0.786) (Yan *et al*., 2016). The low nucleotide variation and genetic diversity observed in the study could have several explanations. First, we used highly conservative criteria of SNP calling, and SNPs with a quality score ≤ 10, having missing data, or a minor allele frequency (MAF) ≤ 0.05 were all eliminated, which decreases the number of SNPs from transcripts in the study. Second, RNA-Seq analysis attempts to reconstruct the transcribed sequence of a given sample from short transcript fragments, but a bias towards highly expressed transcripts and current technological limitations often mean that the transcriptome cannot be retrieved in full (Novembre & Stephens 2008). Third, transcribed regions of genes targeted by RNA-Seq are likely subject to stronger negative selection than introns or intergenic regions (Gossmann *et al.* 2010). Although SNPs from RNA-seq data might therefore underestimate genetic variation, they have several advantages for population genetics, including higher genomic density, data quality, reproducibility, and genotyping efficiency (Schunter *et al*. 2014).

The genetic differentiation of *M. lutarioriparius*, as determined by SNPs was much lower than that of the close relative *Miscanthus sinensis* (Clark *et al*. 2015). This can be explained by distinct differences in the species' ecologies, such as the narrow distribution along riparian corridors and the frequent human activity in the native habitats of *M. lutarioriparius*, which has been shown to increase gene flow among populations (Yan *et al*. 2016). From the evolutionary and ecological perspectives, the lower genetic differentiation may have been caused by geographic variations in the Yangtze River region, which indicates that the species was exposed to stressors or environmental factors that produced local genetic differentiation and genetic homogeneity. Critically, natural selection that favored adaptations to local environmental conditions also affected the genetic differentiation of the populations (Slatkin 1987).

We could demonstrate that SNPs from RNA-Seq data revealed local adaptations by plants to different environments. Specifically, population LU23 had a different distribution of π compared with the other populations (Table S2), and the results further demonstrated the genetic disconnection of LU23 from adjacent populations (Yan *et al*. 2016). Moreover, the π distributions showed that several populations, such as LU4 and LU9, were considerably different from the other populations, which suggests that different environments could cause the divergence of populations under non-controlled conditions (Cheviron *et al*. 2008). LU4 is located close to Dongting Lake and endures flood stress throughout the growing season; LU9 is located at the junction of Dongting Lake and the Yangtze River; and LU23 is located along a dam, which is the source of intensive anthropogenic activities. Thus, the study of transcriptomic variation could provide new insights into the adaptive strategies of *M. lutarioriparius* in distinct geographical regions, which will help in the designation of conservation units based on the degree of variation among populations.

Variation in gene expression is produced by a combination of genetic and environmental factors, and can be very sensitive to environmental differences; therefore, it has the potential to reveal early phases of species divergence (Gibson 2008; Wolf *et al*. 2010). Indeed, all E_p and E_d pairs of the 10 populations were significantly different, whereas the levels of nucleotide diversity in the populations were almost indistinguishable. Studying expression patterns using transcriptome techniques therefore represents an important complementary approach for studying population divergence, even though the relative contribution of such gene expression divergence to overall species differentiation remains unclear, and should be addressed in future studies.

The PCA of gene expression revealed strong differentiation of the population structure, for example, LU5, LU9, LU14, and LU19 were clearly distinguished with the longitude and latitude of the populations. In addition, we found a significant pattern of isolation by distance (IBD) at the kilometers scale, and increasingly strong correlations between genetic distance/ expression similarity and geographic distance were observed for microsatellite data (r = 0.295) (Yan *et al*. 2016), SNP data (r = 0.318), and gene expression data (r = -0.437). The ANOVA of gene expression showed the same pattern as the AMOVA of genetic variation, with the majority of variation occurring within populations. According to the functional annotations for biological processes (differential expression analysis), we found that LU19 exhibits greater differentiation than the other populations, while LU4 and LU5 did not show significantly enriched terms compared with those of the other populations. All of the above analyses suggest that variation in gene expression is more sensitive to detect weak differentiation than genetic diversity (Microsatellite and SNP).

Intriguingly, gene expression related to photosynthetic processes was significantly enriched between LU4 and LU14, LU7 and LU23, LU9 and LU23, and LU14 and LU23. In addition, the genes responding to temperature stimuli were enriched between LU9 and LU14 as well as LU14 and LU19. Between LU10 and LU19, the responses to temperature stimulus, hydrogen peroxide and reactive oxygen species were enriched. Specifically, the differential gene expression was primarily observed among populations and not within populations, and this pattern appears to have been the result of natural selection (Whitehead & Crawford 2006a; 2006b). As such, analyses of gene expression in population genetics research could be used to more precisely characterize population differentiation among populations and identify genes influenced by natural selection.

Variation in gene expression among populations might result from genetic, environmental, developmental or random biological effects, which are important for adaptive evolution (Oleksiak *et al*. 2005). Understanding the patterns of gene expression and genetic variation in populations from different habitats could shed light on the plant responses to novel environments via variations in allelic properties and/or gene expression diversity. We found a significantly negative correlation between nucleotide diversity and gene expression diversity at the gene level in all populations (Fig. 5), which suggests that the expression and nucleotide diversity of genes have a reciprocal relationship when a gene adapts to an organism's surroundings. The pattern might be attributed to *Miscanthus* suffering from the chromosomal duplication in its evolution history (Hodkinson *et al*. 2015). Indeed, duplicated genes usually increase gene expression diversity, but with slightly weaker genetic diversity than single-copy genes (Gu *et al*. 2004; Jordan *et al*. 2004). Additionally, divergence in gene expression may well be an important contributor to adaptation in different environments and could thereby maintain the stability of phenotype in a population via genetic canalization (Wagner *et al*. 1997; Wilkins 1997).

Plants respond to stressors either via plasticity (i.e., phenotypic changes that do not depend on immediate heritable genetic change) within the lifetimes of individuals or via evolutionary adaptations over multiple generations (Harrisson *et al*. 2013; Lempe *et al*. 2013). Regardless of the pathways involved, the response depends on altering intrinsic materials (e.g., gene expression or allelic frequencies) responsible for adaptations that enhance the fitness and survival of individuals. For example, variation in DNA sequence (genetic variation) likely causes changes in gene expression, and in turn variation in gene expression reflects underlying genetic variations. In the study, we found a significantly positive correlation between nucleotide diversity and gene expression diversity (Fig. 6A), and a significantly negative correlation between similarity of expression diversity and genetic distance among populations (Fig. 6B), which indicate gene expression variation represents an important component in maintaining a stable population (Roger & Bernatchez 2005; 2007; Scott *et al*. 2009; Fu *et al*. 2012; Ackermann *et al*. 2013; Leder *et al*. 2015). Thus, inter-individual variations in gene expression are likely heritable and provide genomic evidence of the maintenance of genetic differentiation via natural selection as the natural counterpart of genetic variation.

In conclusion, gene expression data can be used as a powerful tool to explore population genetics by comparing the diversity patterns obtained from SNPs and expression data. The results obtained using SNPs from the transcriptome were consistent with results from microsatellites in the nuclear genome, but the gene expression data revealed additional levels of population structure not captured with conventional methods. The reciprocal relationship between genetic variation and gene expression is an important pattern that reflects the responses of a species to its environment, and both are likely involved in local adaptation. In short, RNA-Seq is a useful tool for assessing the variations in gene expression that underlies phenotypic differences among wild populations, and the SNP library and extensive transcriptome sequences developed in this study will facilitate future functional analyses of molecular signatures capable of indicating adaptations in populations of *M. lutarioriparius*.

Acknowledgements

We thank Mingdong Zhu for sampling the population materials; the Beijing Center for Physical and Chemical Analysis for generating the sequencing data, the Beijing Computing Center for assisting with computational infrastructure for data analysis. The work was supported by grants from the Key Program of the National Natural Science Foundation of China (No. 91131902) and the Project for Autonomous Deployment of the Wuhan Botanical Garden (55Y755271G02).

References

Ackermann, M., Sikora-Wohlfeld, W., & Beyer, A. (2013). Impact of natural genetic variation on gene expression dynamics. Plos Genetics, 9(6), e1003514. doi:10.1371/journal.pgen.1003514

Alvarez, M., Schrey, A. W., & Richards, C. L. (2015). Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Molecular Ecology, 24(4), 710-725. doi:10.1111/mec.13055

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology, 11(10). doi:ARTN R10610.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166-169. doi:10.1093/bioinformatics/btu638

Annadurai, R., Jayakumar, V., Mugasimangalam, R., Katta, M., Anand, S., Gopinathan, S., . . . Rao, S. (2012). Next generation sequencing and *de novo* transcriptome analysis of *Costus pictus* D. Don, a non-model plant with potent anti-diabetic properties. Bmc Genomics, 13(1), 663.

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. Trends in Ecology & Evolution, 23(1), 38-44. doi:http://dx.doi.org/10.1016/j.tree.2007.09.008

Bohonak, A. J. (2002). IBD (isolation by distance): A program for analyses of isolation by distance. Journal of Heredity, 93(2), 153-154. doi:DOI 10.1093/jhered/93.2.153

Bowsher, A. W., Shetty, P., Anacker, B. L., Siefert, A., Strauss, S. Y., & Friesen, M. L. (2017). Transcriptomic responses to conspecific and congeneric competition in co-occurring Trifolium. Journal of Ecology, 105. doi:10.1111/1365-2745.12761

Brookfield, J. F. Y. (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. Molecular Ecology, 5(3), 453-455. doi:DOI 10.1046/j.1365-294X.1996.00098.x

Chen, S. L., & Renvoize, S. A. (2005). A new species and a new combination of *Miscanthus* (Poaceae) from China. Kew Bulletin, 60(4), 605-607. doi:10.2307/25070249

Cheviron, Z. A., Whitehead, A., & Brumfield, R. T. (2008). Transcriptomic variation and plasticity in rufous-collared sparrows (*Zonotrichia capensis*) along an altitudinal gradient. Molecular Ecology, 17(20), 4556-4569. doi:10.1111/j.1365-294X.2008.03942.x

Clark, L. V., Brummer, J. E., Glowacka, K., Hall, M. C., Heo, K., Peng, J. H., . . . Sacks, E. J. (2014). A footprint of past climate change on the diversity and population structure of Miscanthus sinensis. Annals of Botany, 114(1), 97-107. doi:10.1093/aob/mcu084

Clark, L. V., Stewart, J. R., Nishiwaki, A., Toma, Y., Kjeldsen, J. B., Jørgensen, U., . . . Sacks, E. J. (2015). Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression. Journal of Experimental Botany. doi:10.1093/jxb/eru511

Clifton-Brown, J. C., Stampfl, P. F., & Jones, M. B. (2004). *Miscanthus* biomass production for energy in Europe and its potential contribution to decreasing fossil fuel carbon emissions. Global Change Biology, 10(4), 509-518. doi:10.1111/j.1529-8817.2003.00749.x

Clotault, J., Thuillet, A. C., Buiron, M., De Mita, S., Couderc, M., Haussmann, B. I. G., . . . Vigouroux, Y. (2012). Evolutionary history of pearl millet (*Pennisetum glaucum* [L.] R. Br.) and selection on flowering genes since its domestication. Molecular Biology and Evolution, 29(4), 1199-1212. doi:10.1093/molbev/msr287

Derome, N., Duchesne, P., & Bernatchez, L. (2006). Parallelism in gene transcription among sympatric lake whitefish (*Coregonus clupeaformis* Mitchill) ecotypes. Molecular Ecology, 15(5), 1239-1249. doi:10.1111/j.1365-294X.2005.02968.x

Dyer, R. J., Nason, J. D., & Garrick, R. C. (2010). Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. Molecular Ecology, 19(17), 3746-3759. doi:10.1111/j.1365-294X.2010.04748.x

Earl, D., & vonHoldt, B. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, 4(2), 359-361. doi:10.1007/s12686-011-9548-7

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107(1), 1-15. doi:10.1038/hdy.2010.152

Ellison, C. E., Hall, C., Kowbel, D., Welch, J., Brem, R. B., Glass, N. L., & Taylor, J. W. (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proceedings of the National Academy of Sciences, 108(7), 2831-2836. doi:10.1073/pnas.1014971108

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources, 10(3), 564-567. doi:10.1111/j.1755-0998.2010.02847.x

Fan, Y., Wang, Q., Kang, L., Liu, W., Xu, Q., Xing, S., . . . Sang, T. (2015). Transcriptome-wide characterization of candidate genes for improving the water use efficiency of energy crops grown on semiarid land. Journal of Experimental Botany, 66(20), 6415-6429. doi:10.1093/jxb/erv353

Fu, J., Wolfs, M. G. M., Deelen, P., Westra, H.-J., Fehrmann, R. S. N., te Meerman, G. J., . . . Franke, L. (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. Plos Genetics, 8(1), e1002431. doi:10.1371/journal.pgen.1002431

Gahlan, P., Singh, H., Shankar, R., Sharma, N., Kumari, A., Chawla, V., . . . Kumar, S. (2012). De novo sequencing and characterization of *Picrorhiza kurrooa* transcriptome at two temperatures showed major transcriptome adjustments. Bmc Genomics, 13(1), 126.

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., . . . Galtier, N. (2013). Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. Plos Genetics, 9(4), e1003457. doi:10.1371/journal.pgen.1003457

Gibson, G. (2008). The environmental contribution to gene expression profiles. Nature Reviews Genetetics, 9(8), 575-581.

Gossmann, T. I., Song, B.-H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., . . . Eyre-Walker, A. (2010). Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Molecular Biology and Evolution, 27(8), 1822-1832. doi:10.1093/molbev/msq079

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech, 29(7), 644-652.

Gu, Z. L., Rifkin, S. A., White, K. P., & Li, W. H. (2004). Duplicate genes increase gene expression diversity within and between species. Nature Genetics, 36(6), 577-579. doi:10.1038/ng1355

Guggisberg, A., Lai, Z., Huang, J., & Rieseberg, L. H. (2013). Transcriptome divergence between introduced and native populations of Canada thistle, Cirsium arvense. New Phytologist, 199(2), 595-608. doi:10.1111/nph.12258

Harris, S. E., O'Neill, R. J., & Munshi-South, J. (2015). Transcriptome resources for the white-footed mouse (*Peromyscus leucopus*): new genomic tools for investigating ecologically divergent urban and rural populations. Molecular Ecology Resources, 15(2), 382-394. doi:10.1111/1755-0998.12301

Harrison, P. W., Wright, A. E., & Mank, J. E. (2012). The evolution of gene expression and the transcriptome–phenotype relationship. Seminars in Cell & Developmental Biology, 23(2), 222-229.

Hartl DL, Clark AG (2006) Principles of population genetics, 4th ed. Sinauer Ass., Sunderland MA.

Harrisson, K. A., Pavlova, A., Amos, J. N., Takeuchi, N., Lill, A., Radford, J. Q., & Sunnucks, P. (2013). Disrupted fine-scale population processes in fragmented landscapes despite large-scale genetic connectivity for a widespread and common cooperative breeder: the superb fairy-wren (Malurus cyaneus). Journal of Animal Ecology, 82(2), 322-333. doi:10.1111/1365-2656.12007

Hedrick PW (2009) Population genetics and ecology. In *The Princeton Guide to Ecology*; Lewin, S.A., Carpenter, S.R., Eds.; Princeton University Press: Princeton, NJ, USA; p. 737.

Hodkinson, T. R., Klaas, M., Jones, M. B., Prickett, R., & Barth, S. (2015). Miscanthus: a case study for the utilization of natural genetic variation. Plant genetic resources-characterization and utilization, 13(3), 219-237. doi:10.1017/S147926211400094x

Hufford, M. B., Bilinski, P., Pyhajarvi, T., & Ross-Ibarra, J. (2012). Teosinte as a model system for population and ecological genomics. Trends in Genetics, 28(12), 606-615. doi:10.1016/j.tig.2012.08.004

Jeukens, J., Renaut, S., St-Cyr, J., Nolte, A. W., & Bernatchez, L. (2010). The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. Molecular Ecology, 19(24), 5389-5403. doi:10.1111/j.1365-294X.2010.04934.x

Jordan, I. K., Wolf, Y. I., & Koonin, E. V. (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evolutionary Biology, 4. doi:10.1186/1471-2148-4-22

Jorgensen, U., & Schwarz, K. U. (2000). Why do basic research? A lesson from commercial exploitation of miscanthus. New Phytologist, 148(2), 190-193.

Josse, J., & Husson, F. (2016) missMDA: a package for handling missing values in multivariate data analysis. Journal of Statistical Software, 70(1): 1-31.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3). doi: 10.1186/gb-2009-10-3-r25

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundström, G., Rubin, C.-J., Gilbert, E. R., . . . Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences, 109*(47), 19345-19350. doi:10.1073/pnas.1216128109

Lempe, J., Lachowiec, J., Sullivan, A. M., & Queitsch, C. (2013). Molecular mechanisms of robustness in plants. Current Opinion in Plant Biology, 16(1), 62-69. doi:10.1016/j.pbi.2012.12.002

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Proc, G. P. D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Liu, W., Mi, J., Song, Z., Yan, J., Li, J., & Sang, T. (2014). Long-term water balance and sustainable production of Miscanthus energy crops in the Loess Plateau of China. Biomass and Bioenergy, 62, 47-57.

Liu, W., & Sang, T. (2013). Potential productivity of the Miscanthus energy crop in the Loess Plateau of China under climate change. Environmental Research Letters, 8(4), 044003. doi:10.1088/1748-9326/8/4/044003

Liu, W., Yan, J., Li, J., & Sang, T. (2012). Yield potential of Miscanthus energy crops in the Loess Plateau of China. GCB Bioenergy, 4(5), 545-554. doi:10.1111/j.1757-1707.2011.01157.x

Ma, X.-F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Thomas-Jones, S., . . . Flavell, R. (2012). High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Miscanthus sinensis*. PLoS ONE, 7(3), e33821. doi:10.1371/journal.pone.0033821

Mi, J., Liu, W., Yang, W. H., Yan, J., Li, J. Q., & Sang, T. (2014). Carbon sequestration by *Miscanthus* energy crops plantations in a broad range semi-arid marginal land in China. Science of the Total Environment, 496, 373-380. doi: 10.1016/j.scitotenv.2014.07.047

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., . . . Marra, M. A. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research, 18(4), 610-621. doi:10.1101/gr.7179508

Nabholz, B., Sarah, G., Sabot, F., Ruiz, M., Adam, H., Nidelet, S., . . . Glemin, S. (2014). Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*). Molecular Ecology, 23(9), 2210-2227. doi:10.1111/mec.12738

Nei, M. (1973). Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences of the United States of America, 70(12), 3321-3323. doi:DOI 10.1073/pnas.70.12.3321

Nei, M., Tajima, F., & Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular-data .2. Gene-frequency data. Journal of Molecular Evolution, 19(2), 153-170. doi:Doi 10.1007/Bf02300753

Nishiwaki, A., Mizuguti, A., Kuwabara, S., Toma, Y., Ishigaki, G., Miyashita, T., . . . Stewart, J. R. (2011). Discovery of natural *Miscanthus* (Poaceae) triploid plants in sympatric populations of Miscanthus sacchariflorus and Miscanthus sinensis in southern Japan. American Journal of Botany, 98(1), 154-159. doi:10.3732/ajb.1000258

Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. Nature Genetics, 40(5), 646-649. doi:10.1038/ng.139

Oleksiak, M. F., Churchill, G. A., & Crawford, D. L. (2002). Variation in gene expression within and among natural populations. Nature Genetics, 32(2), 261-266. doi:Doi 10.1038/Ng983

Oleksiak, M. F., Roach, J. L., & Crawford, D. L. (2005). Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. Nature Genetics, 37(1), 67-72.

Pan, S., Zhang, T., Rong, Z., Hu, L., Gu, Z., Wu, Q., . . . Zhan, X. (2017). Population transcriptomes reveal synergistic responses of DNA polymorphism and RNA expression to extreme environments on the Qinghai–Tibetan Plateau in a predatory bird. *Molecular Ecology*, n/a-n/a. doi:10.1111/mec.14090

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. Plos Genetics, 2(12), 2074-2093. doi: 10.1371/journal.pgen.0020190

Pratlong, M., Haguenauer, A., Chabrol, O., Klopp, C., Pontarotti, P., & Aurelle, D. (2015). The red coral (*Corallium rubrum*) transcriptome: a new resource for population genetics and local adaptation studies. Molecular Ecology Resources, 15, 1205-1215. doi:10.1111/1755-0998.12383

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.

Quinn, L. D., Allen, D. J., & Stewart, J. R. (2010). Invasiveness potential of Miscanthus sinensis: implications for bioenergy production in the United States. GCB Bioenergy, 2(6), 310-320. doi:10.1111/j.1757-1707.2010.01062.x

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3), R25. doi:10.1186/gb-2010-11-3-r25

Rogers, S. M., & Bernatchez, L. (2005). Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). Molecular Ecology, 14(2), 351-361. doi:DOI 10.1111/j.1365-294X.2004.02396.x

Rogers, S. M., & Bernatchez, L. (2007). The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonas sp Salmonidae*) species pairs. Molecular Biology and Evolution, 24(6), 1423-1438. doi:10.1093/molbev/msm066

Sang, T. A. O., & Zhu, W. (2011). China's bioenergy potential. GCB Bioenergy, 3(2), 79-90. doi:10.1111/j.1757-1707.2010.01064.x

Schunter, C., Garza, J. C., Macpherson, E., & Pascual, M. (2014). SNP development from RNA-seq data in a nonmodel fish: how many individuals are needed for accurate allele frequency prediction? Molecular Ecology Resources, 14(1), 157-165. doi:10.1111/1755-0998.12155

Scott, C. P., Williams, D. A., & Crawford, D. L. (2009). The effect of genetic and environmental variation on metabolic gene expression. Molecular Ecology, 18(13), 2832-2843. doi:10.1111/j.1365-294X.2009.04235.x

Shendure, J., & Ji, H. L. (2008). Next-generation DNA sequencing. Nature Biotechnology, 26(10), 1135-1145. doi:10.1038/nbt1486

Slatkin, M. (1987). Gene Flow and the Geographic Structure of Natural-Populations. Science, 236(4803), 787-792. doi:DOI 10.1126/science.3576198

Somerville, C., Youngs, H., Taylor, C., Davis, S. C., & Long, S. P. (2010). Feedstocks for lignocellulosic biofuels. science, 329(5993), 790-792. doi:10.1126/science.1189268

Song, Z., Xu, Q., Lin, C., Tao, C., Zhu, C., Xing, S., . . . Sang, T. (2017). Transcriptomic characterization of candidate genes responsive to salt tolerance of *Miscanthus* energy crops. GCB Bioenergy, n/a-n/a. doi:10.1111/gcbb.12413

Stapley, J., Reger, J., Feulner, P. G. D., Smadja, C., Galindo, J., Ekblom, R., . . . Slate, J. (2010). Adaptation genomics: the next generation. *Trends in Ecology & Evolution, 25*(12), 705-712. doi:http://dx.doi.org/10.1016/j.tree.2010.09.002

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R. S. T., . . . Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat Commun*, *5*. doi:10.1038/ncomms6770

Todesco, M., Balasubramanian, S., Hu, T. T., Traw, M. B., Horton, M., Epple, P., . . . Weigel, D. (2010). Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*. Nature, 465(7298), 632-636.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), 562-578. doi:10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28(5), 511-515.

Wagner, G. P., Booth, G., Bagheri-Chaichian, H. (1997) A population genetic theory of canalizaiton. Evolution, 51(2), 329-347

West-Eberhard, M. J. (2003). Phenotypic accommodation: Adaptive innovation due to developmental plasticity, with or without genetic change. Integrative and Comparative Biology, 43(6), 970-970.

Whitehead, A., & Crawford, D. L. (2006a). Neutral and adaptive variation in gene expression. Proceedings of the National Academy of Sciences, 103(14), 5425-5430. doi:10.1073/pnas.0507648103

Whitehead, A., & Crawford, D. L. (2006b). Variation within and among species in gene expression: raw material for evolution. Molecular Ecology, 15(5), 1197-1211. doi:10.1111/j.1365-294X.2006.02868.x

Whitehead, A., Dubansky, B., Bodinier, C., Garcia, T. I., Miles, S., Pilley, C., . . . Galvez, F. (2012). Genomic and physiological footprint of the Deepwater Horizon oil spill on resident marsh fishes. Proceedings of the National Academy of Sciences, 109(50), 20298-20302. doi:10.1073/pnas.1109545108

Wilkins, A. S. (1997) Canalization: a molecular genetic perspective. BioEssays, 19(3), 257-262. doi: 10.1002/bies.950190312

Wolf, J. B. W., Bayer, T., Haubold, B., Schilhabel, M., Rosenstiel, P., & Tautz, D. (2010). Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. Molecular Ecology, 19, 162-175. doi:10.1111/j.1365-294X.2009.04471.x

Wright, S. I., & Gaut, B. S. (2005). Molecular Population genetics and the search for adaptive evolution in plants. Molecular Biology and Evolution, 22(3), 506-519. doi:10.1093/molbev/msi035

Xu, Q., Xing, S. L., Zhu, C. Y., Liu, W., Fan, Y. Y., Wang, Q., . . . Sang, T. (2015). Population transcriptomics reveals a potentially positive role of expression diversity in adaptation. Journal of Integrative Plant Biology, 57(3), 284-299. doi:Doi 10.1111/Jipb.12287

Yan, J., Chen, W., Luo, F. A. N., Ma, H., Meng, A., Li, X., . . . Sang, T. A. O. (2012). Variability and adaptability of Miscanthus species evaluated for energy crop domestication. GCB Bioenergy, 4(1), 49-60. doi:10.1111/j.1757-1707.2011.01108.x

Yan, J., Zhu, C. Y., Liu, W., Luo, F., Mi, J., Ren, Y. J., . . . Sang, T. (2015). High photosynthetic rate and water use efficiency of Miscanthus lutarioriparius characterize an energy crop in the semiarid temperate region. Global Change Biology Bioenergy, 7(2), 207-218. doi:Doi 10.1111/Gcbb.12118

Yan, J., Zhu, M. D., Liu, W., Xu, Q., Zhu, C. Y., Li, J. Q., & Sang, T. (2016). Genetic variation and bidirectional gene flow in the riparian plant Miscanthus lutarioriparius, across its endemic range: implications for adaptive potential. Global Change Biology Bioenergy, 8(4), 764-776. doi:10.1111/gcbb.12278

Yu, G. C., Wang, L. G., Han, Y. Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics-a Journal of Integrative Biology, 16(5), 284-287. doi:10.1089/omi.2011.0118

Data accessibility

The Illumina short reads are available from the NCBI sequence read archive (Project ID: SRP066219, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA301483).

Author contributions:

Sang T and Li JQ funded project and designed the study. Yan J & Song ZH conducted analyses of data and wrote the manuscript. Kang LF & Xing SL carrie out the experiments and data analysis. Züst T, Greimler J,Liu W, Xu Q and Zhu CY contributed to the manuscript revisions.

Figure legends

Figure 1 Ten locations of the studied *Miscanthus lutarioriparius* populations across the species distribution range. These populations were sampled because of their special geographic position, genetic variation or adaptability (Yan *et al.* 2012; 2015; 2016). The population identities correspond to the codes used in Yan *et al*. 2016. In the upper left corner, the gray colors in the map correspond to locations where *M. lutarioriparius* has been transplanted from warm and wet sites to semiarid regions and was able to adapt to the environment (Yan *et al.* 2012). The degree of suitability for growing M. lutarioriparius in those environments is indicated by darker shades of gray.

Figure 2 Coverage of the reference transcriptome of the 78 *M. lutarioriparius* samples. The dots represent an individual with a certain number of reads after rigorous quality testing. The regression line is based on the best least squares fit to the function at $r^2 = 0.269$.

Figure 3 Genetic analysis based on SNPs in *M. lutarioriparius*: (A) The distribution of nucleotide diversity; (B) PCA of SNP data in each individual; (C) Relationship between genetic distance and geographic distance between each two populations; and (D) Distribution of genetic variation within populations and among populations based on SNP data.

Figure 4 Expression analysis based on FPKM in *M. lutarioriparius*: (A) The distribution of population expression (E_p); (B) Expression diversity (E_d) relationship between genetic distance and geographic distance; (C) PCA of gene expression (FPKM); (D) Relationship between population expression similarity and geographic distance; and (E) Distribution of expression variation within populations and among populations based on FPKM. Genes with differential expression are indicated by red dots.

Figure 5 Correlation between the nuclear diversity and expression diversity of all genes in each population.

Figure 6 Transcriptomic patterns of *M. lutarioriparius* across the distribution range: (A) relationship between nucleotide diversity and expression diversity; and (B) relationship between genetic distance and population expression similarity.

Supporting information

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

Table S1 The RNAseq data after being filtered and trimmed.

Table S2 *K-S* test of nucleotide diversity (π) between populations.

Table S3 The pair-wise F_{ST} values among the 10 populations based on SNP data.

Table S4 *K-S* test of population expression (E_{p}) between populations.

Table S5 *K-S* test of expression diversity (E_{d}) between populations.

Table S6 Significantly enriched GO terms in biological processes of DEGs between populations.

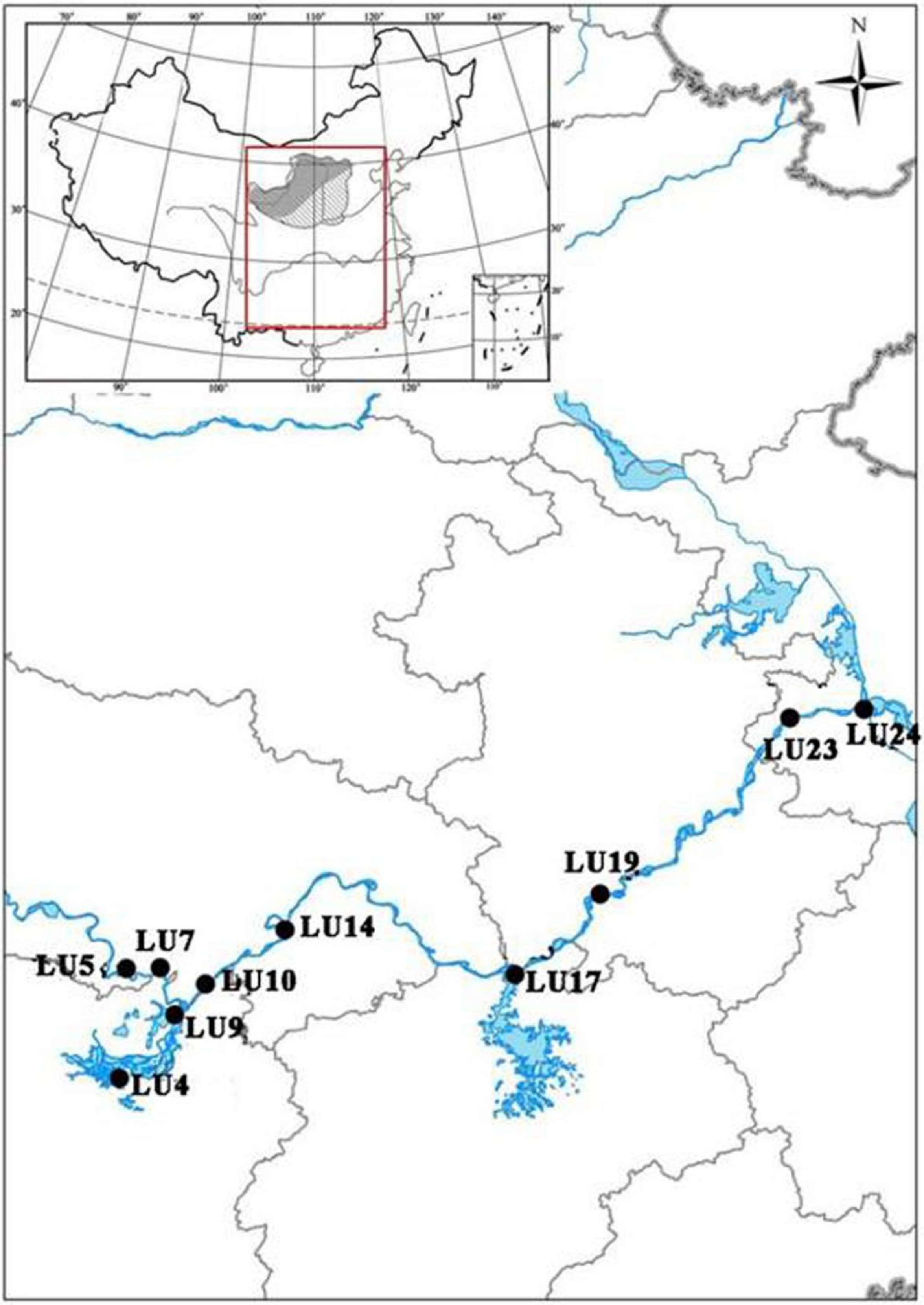
Table S7 Similarity of population expression (E_{p}) among populations. Pearson correlation coefficients based on scaled E_{p} of 10 populations.

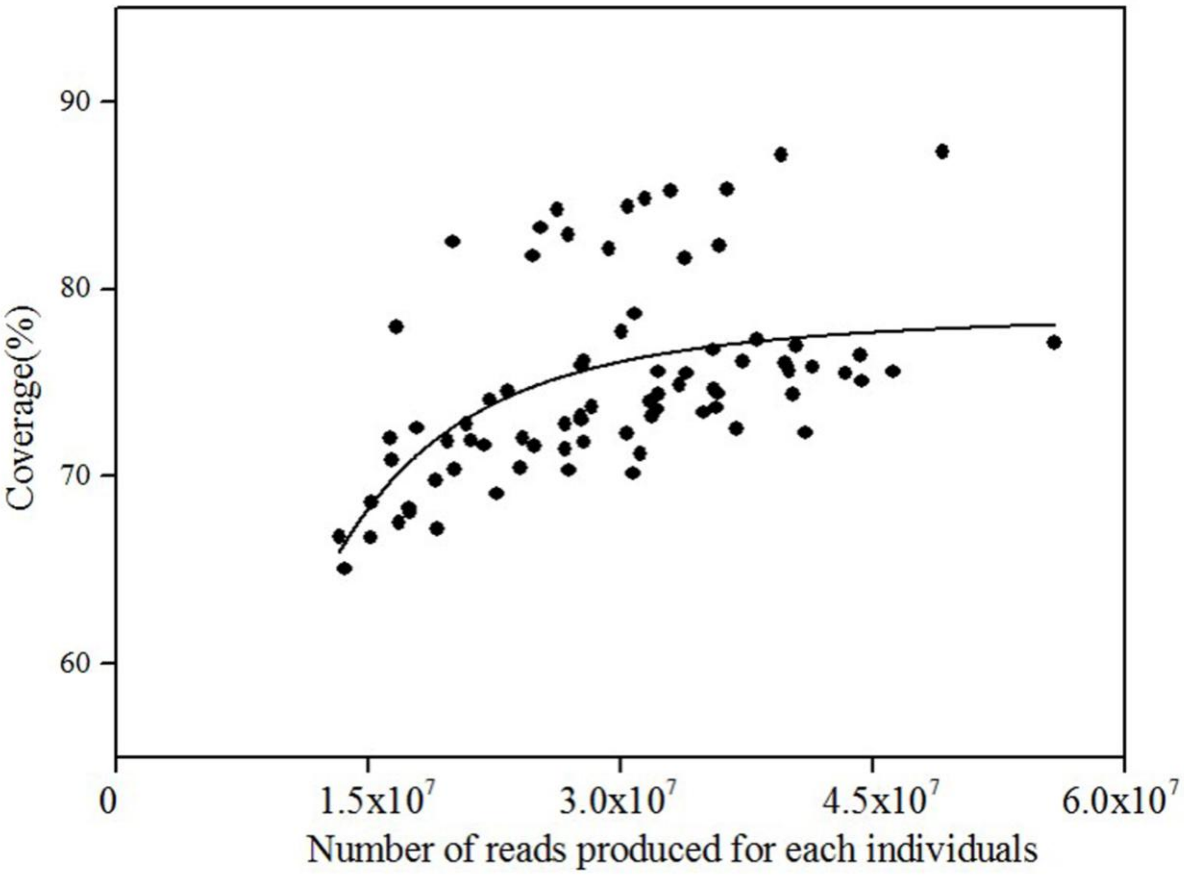
Fig. S1.Summary of genetic variation and gene structure in *M. lutarioriparius* (A) nucleotide diversity, (B) population genetic structure based on STRUCTURE.

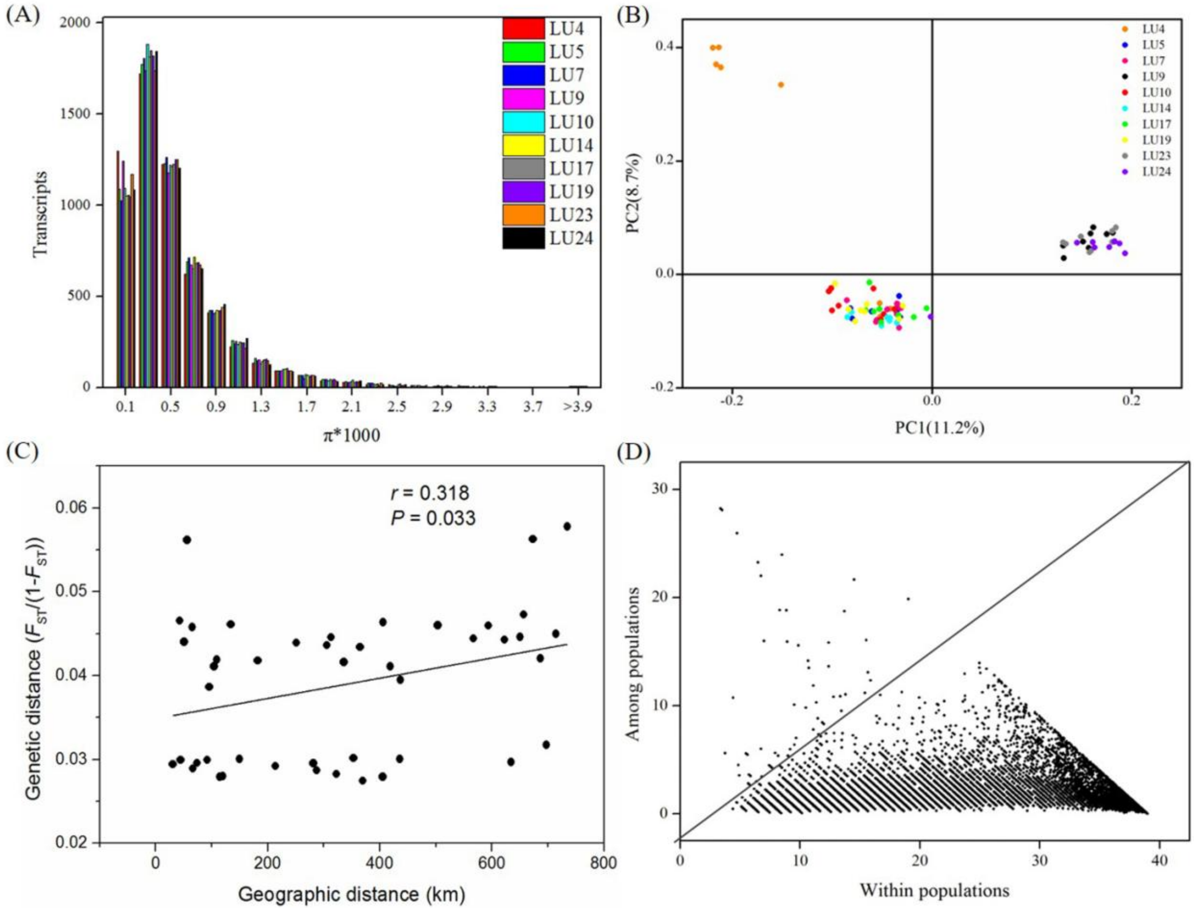
Fig. S2.Summary of gene expression in *M. lutarioriparius*: (A) population expression (E_{p}), (B) expression diversity (E_{d}).

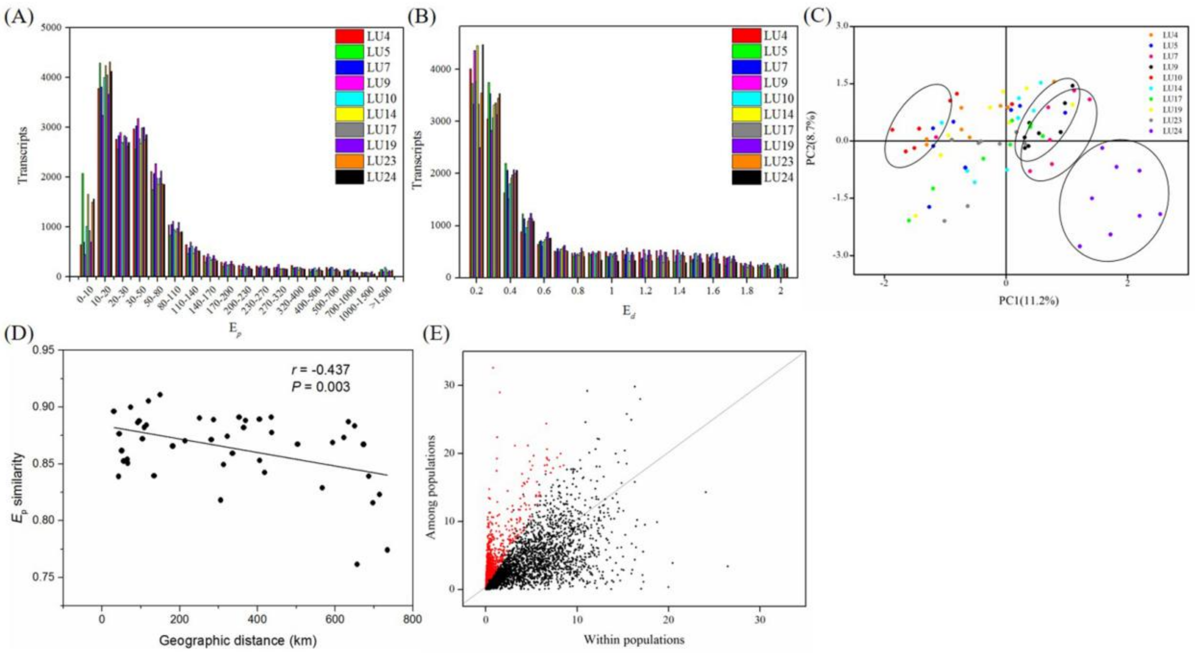
Table 1 Measures of genetic diversity for each population of *M. lutarioriparius* based on population transcriptome data.

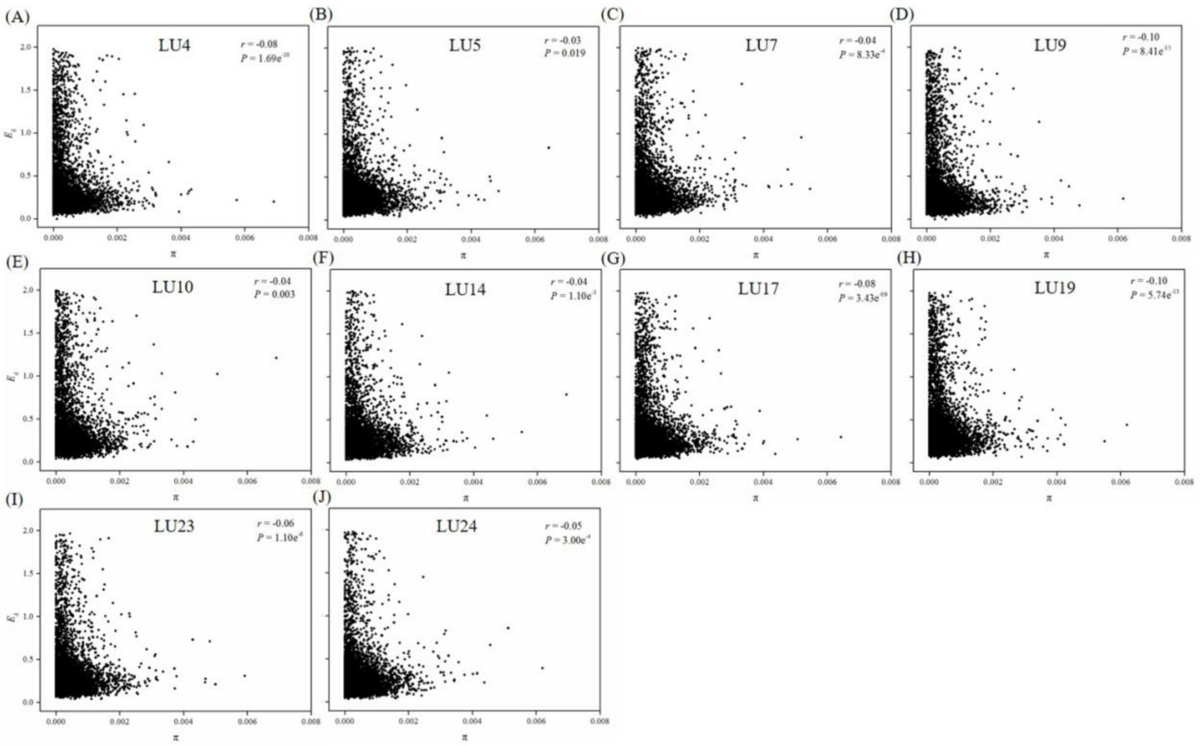
	ID	N	π	H_E	H_O
1	LU4	8	0.000428	0.3675	0.5144
2	LU5	7	0.000444	0.3676	0.4957
3	LU7	8	0.000443	0.3615	0.4775
4	LU9	8	0.000425	0.3704	0.5141
5	LU10	8	0.000429	0.3596	0.4636
6	LU14	7	0.000443	0.3679	0.5035
7	LU17	8	0.000438	0.3601	0.4759
8	LU19	8	0.000440	0.3602	0.4715
9	LU23	8	0.000434	0.3659	0.5083
10	LU24	8	0.000418	0.3702	0.5080

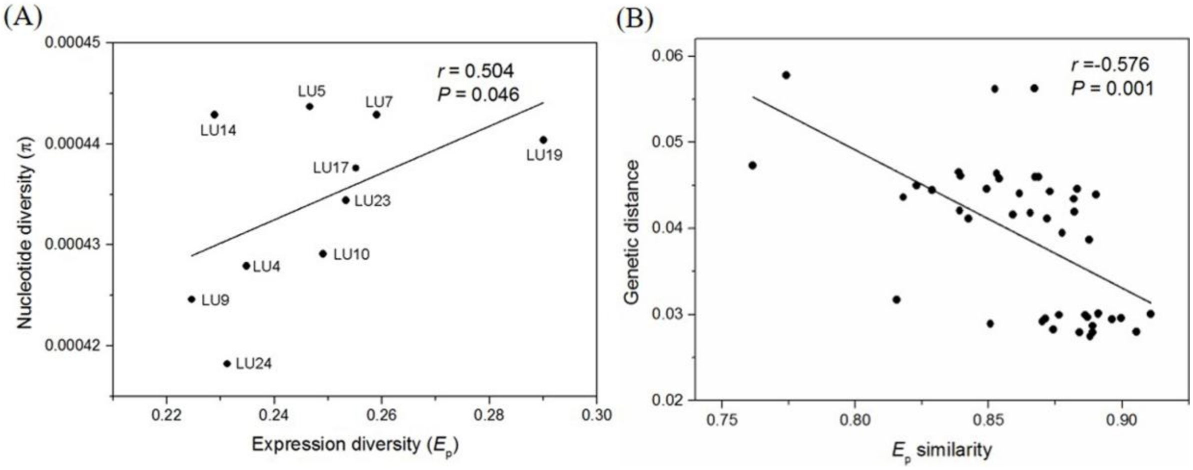

N
70°
80°
90°
100°
110°
120°
130°
140°
50°
40°
30°
20°
10°
LU5
LU7
LU10
LU9
LU4
LU14
LU17
LU19
LU23
LU24


90
80
70
60
Coverage(%)
0
1.5x10^7
3.0x10^7
4.5x10^7
6.0x10^7
Number of reads produced for each individuals


(A)
Transcripts
2000
1500
1000
500
0
π*1000
0.1
0.5
0.9
1.3
1.7
2.1
2.5
2.9
3.3
3.7
>3.9
LU4
LU5
LU7
LU9
LU10
LU14
LU17
LU19
LU23
LU24
(B)
PC2(8.7%)
PC1(11.2%)
0.4
0.2
0.0
-0.2
-0.2
0.0
0.2
LU4
LU5
LU7
LU9
LU10
LU14
LU17
LU19
LU23
LU24
(C)
r = 0.318
P = 0.033
Genetic distance (F_{ST}/(1-F_{ST}))
0.06
0.05
0.04
0.03
0.02
Geographic distance (km)
0
200
400
600
800
(D)
Among populations
Within populations
30
20
10
0
0
10
20
30
40


(A)
Transcripts
0
1000
2000
3000
4000
5000
E_p
0-10
10-20
20-30
30-50
50-80
80-110
110-140
140-170
170-200
200-230
230-270
270-320
320-400
400-500
500-700
700-1000
1000-1500
>1500
LU4
LU5
LU7
LU9
LU10
LU14
LU17
LU19
LU23
LU24
(B)
Transcripts
0
1000
2000
3000
4000
E_d
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
LU4
LU5
LU7
LU9
LU10
LU14
LU17
LU19
LU23
LU24
(C)
PC2(8.7%)
3.0
1.5
0.0
-1.5
-3.0
PC1(11.2%)
-2
0
2
LU4
LU5
LU7
LU9
LU10
LU14
LU17
LU19
LU23
LU24
(D)
E_p similarity
0.95
0.90
0.85
0.80
0.75
r = -0.437
P = 0.003
Geographic distance (km)
0
200
400
600
800
(E)
Among populations
0
10
20
30
Within populations
0
10
20
30


(A)
LU4
r = -0.08
P = 1.69e^{-10}
(B)
LU5
r = -0.03
P = 0.019
(C)
LU7
r = -0.04
P = 8.33e^{-4}
(D)
LU9
r = -0.10
P = 8.41e^{-13}
(E)
LU10
r = -0.04
P = 0.003
(F)
LU14
r = -0.04
P = 1.10e^{-3}
(G)
LU17
r = -0.08
P = 3.43e^{-09}
(H)
LU19
r = -0.10
P = 5.74e^{-13}
(I)
LU23
r = -0.06
P = 1.10e^{-6}
(J)
LU24
r = -0.05
P = 3.00e^{-4}
E_i
0.0
0.5
1.0
1.5
2.0
π
0.000
0.002
0.004
0.006
0.008


(A)
r = 0.504
P = 0.046
LU5
LU7
LU14
LU19
LU17
LU23
LU10
LU4
LU9
LU24
Nucleotide diversity (π)
0.00042
0.00043
0.00044
0.00045
Expression diversity (E_p)
0.22
0.24
0.26
0.28
0.30
(B)
r =-0.576
P = 0.001
Genetic distance
0.02
0.03
0.04
0.05
0.06
E_p similarity
0.75
0.80
0.85
0.90