**Population Genetic Structure of Wild Rice (*Oryza rufipogon* Griff.) in Indochina Peninsula**

**Abstract**

Adaptation…. Reproductive strategy….. level of genetic variation between Perennial and annual

**Introduction**

Population of a species inhabit in natural habitat is a good system to understanding the consequences of what evolutionary processes in which the populations have been gone through. The populations experienced heterogeneity microenvironments which created different selection pressures lead to evolution of local adaptation.

Kawecki and Ebert (2004) explained the term “local adaptation” for patterns and processes observed across local populations of the same species connected, at least potentially, by dispersal and gene flow”.

Plants that experience selection pressure by local microenvironment conditions and have limited genetic exchange among populations may lead to genetic divergence in different populations due to different genetically controlled character states (Galloway and Fenster, 2000).

Schoville et al., (2012) documented that adaptive genetic variation is defined as the variation found between the genomes of individuals and resulting from natural selection.

In addition to local adaptation, gene flow also plays a significant rule in plants population genetic structure especially genetic exchange between wild ancestral and its domesticated species

Wild rice (*Oryza rufipogon* Griff.) is commonly known as the ancestor of the Asian cultivated rice (ref)

Adaptation to local conditions

The levels and patterns of genetic variation in natural populations of wild progenitors of domesticated plants is in a dynamics system especially

Genetic variation among and within populations of a species

In addition to local environments, the role of gene flow between wild ancestor and its domesticate species have been addressed in many plants species i.e. maize (ref), beets (ref), soy bean (ref), and

Habitat of wild rice (Oryza rufipogon Griff.)

We examined the 47 newly collection of wild rice (*O. rufipogon* Griff.) collected in natural habitats in Cambodia, Laos, Myanmar and Thailand. We sampled nearly a thousand individuals. Our specific questions in the present study are: 1)

**Study system**

*Oryza rufipogon* Griff. or Asian wild rice

**Materials and Methods**

**Population collection**

Forty-eight newly natural wild rice (*O. rufipogon* Griff.) populations were collected in 4 countries: Cambodia (CB), Laos (LS), Myanmar (MM), and Thailand (TH) (Table 1 and Fig 1). Ten of these populations were from Cambodia (CB1 – CB10), eight from Laos (LS1 – LS8), sixteen from Myanmar (MM1 – MM16), and fourteen from Thailand (TH1 – TH14). Geographical reference (UTM), habitat type, distance from cultivated rice fields, and life-history trait types were recorded (Table 1). Leaves of at least 20 individuals per population were collected within the fields: the distance between individuals varied between 5 to 10 meters to prevent clonally collection since perennial wild rice mostly propagate by vegetative and stolen. Leaf samples were silica-dried and kept at room temperature until DNA extraction.

**DNA extraction and Microsatellite analysis**

Total DNA of silica-dried leaf tissue was individually extracted using modified CTAB method (Doyle and Doyle 1987) and/or Viogene plant genomic DNA extraction system (Viogene, Taipei, Taiwan). A total of twenty-four microsatellite markers that are distributed across the 12 rice chromosomes were chosen at random (Table 2). Polymerase chain reactions (PCR) were performed in a total volume of 10 µl containing 10-20 ng of template DNA, 2 pM fluorescence-labeled (6FAM, HEX, NED; Applied Biosystems, USA) forward primer and 2 pM unlabelled reverse primer, 2 µl *Taq* MgCl2-free 5x buffer, 250 µM deoxyribonucleotides (dNTP), 1.25 mM MgCl2, and 0.1 unit of *Taq* polymerase (Promega, USA). Reactions were performed by denaturing at 94ºC (4 min) followed by 40 cycles of 94ºC, 50-68 ºC or 67ºC, 72 ºC, each for 30 sec intervals followed by a final 72 ºC extension for 5 min, and a 4 ºC hold. Amplified products of different sizes and contrasting fluorescent labels were multiplexed and run on a capillary electrophoresis sequencer/genotyper ABI 3130xl Genetic Analyser (Applied Biosystems, USA) using GENESCAN 400HD ROX as an internal size standard. Microsatellite alleles were scored using GENEMAPPER v3.7 (Applied Biosystems, USA) and proof checked manually.

**Population structure analysis**

To investigate population structure and infer the most likely number of clusters (K), we used a model-based clustering algorithm implemented in STRUCTURE v2.3 (Pritchard and Wen, 2004). Since *O. rufipogon* Griff. is a cross-pollination species, we used the admixture model with correlated allele frequencies setting. In this approach, multilocus genotypic data were used to define a set of clusters with distinct allele frequencies and to assign individuals probabilistically to them. The analysis were run for up to twelve putative number of clusters (*K*= 1 to *K*=12). The six independent runs for each *K* value with 200000 burn-ins and 200000 iterations after burn-ins. The probability of best fit into each number of assumed clusters (K) was estimated by an ad hoc statistics ∆K based on the rate of change in the log probability of data between consecutive K-values (Evanno et al., 2005). The parameters of the methods of Evanno et al., (2005) were calculated using the program STRUCTURE HARVESTER v0.6.1 (Earl and vonHoldt, 2012) and the clusters was visualized using program DICSTRUCT v1.1 (Rosenberg, 2004).

To take into account the spatial coordinates, we used the spatial Bayesian clustering algorithm implemented in TESS v2.3 (Chen et al., 2007), which assigns individuals to geographical cluters and estimates their probabilities of membership in clusters. The program allows individuals to be distributed over Kmax cluster, estimating the most likely value for the number of clusters as a value K less than or equal to Kmax. The TESS runs with the smallest values of the Deviance Information Criterion (DIC), a penalized measure of how well the model underlying TESS fits the data. We used the admixture model with an interaction parameter Ψ =0.6 and 100 runs with 120 000 sweeps, a 20 000 sweep burn-in, and the default values for the allele frequency model and admixture model parameters. The top 10-20% runs of the highest likelihood were then analyzed in CLUMPP v1.0 (Jakobsson and Rosenberg, 2007) to correct for between-run discrepancies common to cluster analyses. The results of CLUMPP were then visualized in DISTRUCT v1.1 (Rosenberg, 2004). The spatial Bayesian clustering algorithm implemented by TESS is supposed to perform better than STRUCTURE when genetic differentiation is low and when the genetic units are closely distributed in a continuous or quasi-continuous habitat (Chen et al., 2007).

To explore multilevel analysis of molecular variance (AMOVA) we used GENALEX v6.5 (Peakall and Smouse 2012) to perform on all samples and individual country or on the perennial and annual. This allowed differentiation between ecotypes to be assessed while taking into account the subpopulation structure within each habitat type. Shannon’s index offers a means to partition diversity within and among groups, which complements AMOVA and F-statistics. This analysis, also performed in GENALEX, permits a goodness-of-fit test of differentiation, as well as population-specific indices of information.

To illustrate the two-dimensional spatial representation of genetic differentiation we used principal coordinate analysis (PCoA). The PCoA implemented in GENALEX was used among wild rice populations on all samples, individual country or the perennial and annual.

To determine genetic relationship among forty-eight wild rice populations we constructed neighbor-joining tree based on microsatellite genotype data, inferred from genetic distance matrix using Nei’s distance measure *DA* (Nei et al., 1983), as calculated by the software POWERMARKER V3.25 (Liu and Muse 2005).

**Genetic diversity analysis**

Genotypic data from each population were checked for Hardy-Weinberg Equilibrium (HWE) across loci and populations according to Guo and Thompson (1992), with 100 000 Markov chain step and 10000 dememorization steps. Llinkage disequilibrium (LD) between pairs of loci was tested using the likelihood ratio test. All analyses were conducted using ARLEQUIN v3.5 ([Excoffier and](http://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2012.05587.x/full" \l "b30" \o "Link to bibliographic citation) Lischer, 2010).

Genetic diversity within populations was measured as the number of alleles per locus (Ac) and per population (A), the percentage of polymorphic loci (%P), fixation index (FIS), observed (Ho) and expected (He) heterozygosity, using ARLEQUIN and GENALEX softwares. The outcrossing rate were calculated using the following equation t = (1-FIS)/(1+FIS) while FIS is fixation index (Weir, 1996).

Analysis of isolation by distance (IBD) was carried out using the Mantel test by regressing pairwise FST/(1-FST) with log of pairwise geographical distance between populations, using the program GENALEX. Statistic significance of the regression coefficient was tested with 9999 bootstraps.

**Bottleneck test**

To detect recent population bottlenecks, we used two approaches. First, the approach based on heterozygosity excess, we used the coalescent approach implemented in the BOTTLENECK v1.2.02 (Piry et al. 1999). The detection is based on the expectation that a population bottleneck will lead to a rapid loss of rare alleles producing an excess of heterozygotes and a shift in allele frequency proportions (Cornuet and Luikart, 1996); the population is thus not in mutation-drift equilibrium. The analyses were performed under all three microsatellite mutational models available: IAM, SMM and two phase mutational model (TPM) with 20% SMM and 80% IAM and a variance of 10 as suggested by the program authors (Piry et al. 1999). To determine if the number of loci exhibiting heterozygosity excess was signiﬁcant, we applied the one-tailed Wilcoxon signed rank test for heterozygote excess. BOTTLENECK was run for 10000 iterations. The mode-shift test as implemented in BOTTLENECK to determine if the allele frequency distribution has been shifted towards more common alleles with fewer low frequency alleles as would be expected in the case of a bottleneck.

Second, the approach based on the reduction in effective population size by calculating the *M-statistic* value for each population is according to Garza and Williamson (2001) using the software ARLEQUIN. The significance was assessed by comparison between the mean value *M* across all loci and the value *M*=0.680, the threshold value below which a population can reasonably be assumed to have undergone a reduction in population size. Garza and Williamson (2001) showed that the mean ratio of the number of alleles to the range of allele size (*M*) is sensitive to population bottlenecks, dropping in size when a population is reduced. Therefore the Garza and Williamson statistic is supposed to be very small in population having been through a bottleneck and close to one in stationary populations (manual Arlequin 3.5).

Maps showing the geographical distribution of samples, genetic allocation to clusters in the K=… model in InStruct, and alleles at \_\_\_\_\_ and \_\_\_\_\_ were plotted in ArcMap 10.1. Precise locations of origin were unknown for many samples. For the purposes of plotting data, these were roughly estimated from the geographic information available using GoogleEarth.

**Results**

***Population structure***

were obtained for Kmax greater than three. In Fig. 3, we report results for Kmax= 2 clusters. The cluster membership coefficients estimated for the southeast and northwest accessions suggest that clinal variation occurs

The TESS

and TESS the number of cluster (K) was varied from one to sixteen, with 5 replicate runs performed for all K values. The highest likelihood was obtained when K was set to \_\_\_. However, using the method of Evano et al., (2005), maximal deltaK occurred at K=\_\_ (Fig 2), with the nest largest peak at K=\_\_. At K=2, wild rice were divided into Perennial and Annual

*AMOVA*

Hierarchical analysis of molecular variance suggested low, but significant, differentiation between life-history traits; overall, 4% of variation occurred among habitat types (P<0.001), 16% among subpopulations (P<0.001), and 80% within subpopulation (P<0.001).

*Principal coordinates analysis*

The proportion of variance explained by each principal coordinates axis dropped off after axis 1, which represented 24.78% of the variation (axes 2 and 3 explained 17.86% and \_\_\_%, respectively). Principal coordinate 1was strongly associated with life-history traits (Fig.\_\_). The perennial was clustered in the left of the graph while the annual was clustered in the right of the graph (Fig. \_\_).

*NJ tree*

8yyfty

***Genetic diversity***

The genetic diversity of the perennial type was clearly higher than that of the annual type (Fig 2a and 2b).

***Genetic bottleneck***

**Discussion**