# **Detection of locally adapted genomic regions in wild rice (*Oryza rufipogon*) using environmental association analysis**

James A. Bedford1,2, Mark Carine2, and Mark A. Chapman1.

1. Biological Sciences, Life Sciences Building 85, University of Southampton, Hampshire, SO17 1BJ, United Kingdom

2. Life Sciences, The Natural History Museum, Cromwell Road, London, SW7 5BD, United Kingdom

**Contacts**

J.A.Bedford@soton.ac.uk

M.Chapman@soton.ac.uk

**Keywords**

Local adaptation; *Oryza rufipogon*; Environmental Association Analysis; Climate; Abiotic stress; Genomics; Domestication; Wild rice

**Abstract**

*Oryza rufipogon* is the wild progenitor of cultivated rice *O. sativa* and exhibits high levels of genetic diversity across its distribution, making it a useful resource for the identification of abiotic stress tolerant varieties and genes that could limit future climate-changed induced yield losses. To investigate local adaptation in *O. rufipogon*, we analyzed single nucleotide polymorphism (SNP) data from a panel of 286 samples located across a diverse range of climates. Environmental association analysis (EAA), a genome wide association study (GWAS)-based method, was used and revealed 15 regions of the genome significantly associated with various climate factors. Genes within these environmentally associated regions have putative functions in abiotic stress response, phytohormone signaling and the control of flowering time. This provides an insight into potential local adaptation in *O. rufipogon* and reveals possible locally adaptive genes that may provide opportunities for breeding novel rice varieties with climate change-resilient phenotypes.

# Introduction

Domestication and selective breeding are necessary to produce high crop yields, however, crops tend to have a restricted genetic diversity when compared to their wild relatives (Jovovic *et al.*, 2020) due to large genetic bottlenecks and selection (Flint-Garcia, 2013). The reduction in diversity in domesticates is concerning for future food security as population genetic diversity is generally regarded as beneficial for adaptation to changing environments (Jump *et al.*, 2009).

Rice (*Oryza sativa* L.) provides food for approximately 3.5 billion people, accounting for an estimated 50% of the caloric intake of half a billion people living in poverty in Asia (Muthayya *et al.*, 2014). Globally, approximately 750 million tonnes of rice are produced per year, comparable with other cereal crops such as wheat (FAOSTAT, 2020). The high yields of cultivated rice are a result of selective breeding and *O. rufipogon* Griff., the wild progenitor of *O. sativa,* has greater levels of sequence diversity than both subspecies of *O. sativa* (subspecies *japonica* and *indica*; Caicedo *et al.*, 2007; Huang *et al.*, 2012). Analysis of rice yields in recent history (1979–2008) revealed that fluctuations in temperature and precipitation have had a considerable impact on yield (Ray *et al.*, 2015) and climate modelling indicates that a 3°C increase in global mean temperature could reduce rice yield by up to 19% (Tao *et al.*, 2011).

*Oryza rufipogon* has a perennial life cycle and occurs in permanent wetland environments, including ponds, marshes, rivers and streams (Gao *et al.*, 2012), preferring relatively deep water between 0.2 and 4 m (Vaughan, 1994). An annual ecotype, referred to by some authors as *O. nivara* S.D. Sharma & Shastry, is found in shallow water and seasonally wet environments (Chang, 1976; Li *et al.*, 2006) and is recognized in this study as an *O. rufipogon* ecotype. Outcrossing rates of annual *O. rufipogon* ranges from 4% to 25.5% (Phan *et al*., 2012), whereas in perennial forms this can reach 56% (Barbier, 1989). The native range of *O. rufipogon* in its broad sense extends across East Asia, Indonesia and northern Australia, located between latitudes of approximately 20° S and 30° N (Vaughan, 1994), although the range is hypothesized to have shifted throughout its evolutionary history, coinciding with changes in climate patterns during the Holocene period (Dodson *et al.*, 2021).

*Oryza rufipogon* is distributed widely across a range of ecological environments, potentially generating selection pressures resulting in locally adaptive variants in the genome, that is, genetic variants linked to increased fitness of populations in their native environment when compared to other introduced populations. This could stem from resistance to pests and pathogens, or the abiotic environment, for example, seasons, temperature and precipitation. In cultivated rice, crown root density is associated with high fitness under drought conditions (Groen *et al.*, 2021) and flowering time varies across latitude, with locally adapted loss of function variants detected in key flowering time genes (reviewed by Shrestha *et al.*, 2014). Investigations into the genetics of local adaptation in wild rice are limited but there may be some overlap of genes involved in local adaptation with those in cultivated rice, although this has not been investigated.

Climate modelling and archaeobotanical information have revealed that both domesticated and wild rice have been influenced by shifts in global climate patterns. Spatial and temporal niche modelling of the *japonica* subspecies and integration with archaeological data suggest that declines in temperature played a role in shaping its distribution with minor changes in temperature predicted to have caused a large decline of tropical *japonica* in Northeast China (d’Alpoim Guedes *et al.*, 2015). It is also suggested that these changes in temperature generated adaptive pressure resulting in the formation of temperate *japonica* from the tropical subspecies (d’Alpoim Guedes *et al.*, 2015) which is supported by modelling and archaeological rice records (Gutaker *et al.*, 2020). Overall, previous studies have thus highlighted the significant impact of climate on rice evolution and distribution and have identified traits linked to local adaptation in the domesticated *O. sativa* and its wild relative *O. rufipogon*.

High genetic diversity is a main characteristic of crop wild relatives (Zhang *et al.*, 2017) and makes *O. rufipogon* a suitable model for the identification of abiotic stress tolerant varieties and genes, with implications for the improvement of rice and other crops. There are multiple approaches to investigate the genetics of local adaptation in plant species, for example quantitative trait loci (QTL) mapping (Savolainen *et al.*, 2013) or analysis of genetic markers though environmental association analysis (EAA) (Hoban *et al*., 2016). EAA uses statistical models to associate environmental variation with genetic variation, typically single nucleotide polymorphisms (SNPs), across hundreds of accessions or populations of a species (Rellstab *et al.*, 2015). Genes for environmental adaptation will be located in, or near to, associated regions. These locally adaptive genes may be offset from significantly associated SNPs, as a result of linkage disequilibrium, and require further analysis to be confident of the adaptive value.

Applications of EAA to investigate local adaptation and identify genetic variation associated with environmental conditions has been demonstrated in a few crop wild relatives, for example the wild progenitors of barley (*Hordeum vulgare* subsp. *spontaneum* (K. Koch) Thell; Lei *et al.*, 2019) and soybean (*Glycine soja* Siebold & Zucc; Anderson *et al.*, 2016). Therefore, EAA can be used for identification of genetic markers for crop breeding or identification of genes with potential functions in environmental response.

In this study we investigate local adaptation in *O. rufipogon* with a view to investigating the genetic basis of local adaptations associated with aspects of the abiotic environment, with potential for rice breeding for future or more varied climates. We analyzed genome-wide SNP data from a subset of the wild *O. rufipogon* accessions sequenced by Huang *et al.* (2012), using EAA to identify environmentally associated genomic regions. Genes within these regions were collated and their functions analyzed in the context of previous literature, with the aim of identifying consistencies or novel differences between domesticated rice and its wild relative. The study provides an insight into potential local adaptation in wild rice and highlights several genes that may be involved in adaptation to environmental conditions.

# Methods

**Diversity Panel**

The diversity panel of wild rice used in this investigation is derived from a set of 446 *O. rufipogon* *sensu lato* accessions (i.e., including samples referrable to *O. nivara*) located in the OryzaGenome database release 1.0 (Ohyanagi *et al.*, 2015) and originally sequenced by Huang *et al.* (2012). This sequencing data is a collection of low-coverage (average 1.9x) whole-genome sequence data, for which SNPs were previously called by alignment to the *Oryza sativa* reference genome (IRGSP 4.0) with Smalt (v 0.4), keeping uniquely mapping reads with >96% identity, and identified by the Ssaha Pileup package (v 0.5), removing low quality bases. Missing genotypes were imputed using a k-nearest neighbor imputation method with a reported accuracy of approximately 98% (Huang *et al.*, 2010; Huang *et al.*, 2012).

## Climate data

Initial checks of the 446 accessions resulted in the removal of 18 accessions without location data and correcting the country of origin for an additional 11 samples. Climate data was obtained from the WorldClim database version 1.4 (Hijmans *et al.*, 2005) using the R package Raster (Hijmans *et al.*, 2020) in RStudio (R Core Team, 2022). The WorldClim dataset contains global climate information interpolated from observed data collected between approximately 1960 and 1990 and organized into ca. 1 km2 grids (30 arc-second resolution grid, 0.86 km2 at the equator). Nineteen bioclimatic variables (nine temperature-associated, six precipitation-associated variables and four associated with both temperature and precipitation; Table S1) were retrieved from the dataset using the accession location coordinates from Huang *et al.* (2012) to download specific grid values. Elevation values for each sample were obtained using the Raster “getData” function and the SRTM 90 m resolution dataset. Three accessions were removed as elevation data could not be extracted for these locations.

Rellstab *et al.*, (2015) recommended that populations showing extreme environmental values should not be included in analyses to reduce false positives arising through chance association with outliers. Consequently, rice accessions that were sampled from extreme climates (i.e., at least one variable was outside the 1.5X IQR) were identified and removed. This reduced the number of accessions to 286.

Composite climate variables, principal component 1 (PC1) and PC2, were derived from the climatic data via principal component analysis (PCA) on all climate variables using the RStudio core stats function “princomp” with parameters cor and scores enabled. The first and second principal component values were then extracted. Correlations between variables at each sample location were analyzed with Spearman’s rank correlation and visualized using the R package “ggcorrplot” (Kassambara, 2019). Based on the correlation data, groups of highly correlated variables (ρ ≤ -0.8; ρ ≥ 0.8) were reduced to a single variable to limiting the effects of multiple testing. Therefore, thirteen environmental variables were analyzed here.

## *Oryza rufipogon* accessions and processing of genetic data

SNP data from the 286 accessions was processed in RStudio version 3.6.0. Missing data was re-coded from “-” to “N” and the dataset was converted to “HapMap” style. SNPs were allocated IDs, e.g., ORRUF01\_00001445T:A, detailing the species (ORRUF), chromosome (01), the location within the chromosome (00001445) and the reference/SNP base present (T:A).

On average, accessions were missing 6.78% of SNPs after imputation, with a maximum of 13% missing data, therefore no further accessions were removed from the analysis because of missing data, resulting in a SNP count of 2,463,549. The proportion of missing data for each SNP ranged from 0% to 77.3% and SNPs with greater than 20% missing data were removed, in-line with previous studies (e.g., Lei *et al.*, 2019), resulting in 2,267,618 SNPs. Finally, SNP minor allele frequency (MAF) was calculated and SNPs with a MAF of ≤0.05 were excluded, resulting in 1,898,737 SNPs for the EAA.

## Environmental association analysis

The genome-wide association study (GWAS) program GAPIT3 (Lipka *et al.*, 2012) was selected for the environmental association analysis using the single locus mixed linear model (MLM). This approach has been used in other similar EAA studies (e.g., De La Torre *et al.*, 2019; Lei *et al.*, 2019). Typically, GAPIT identifies associations between phenotype and genotype data collected from a genetically and phenotypically diverse accession panel. This is achieved using mixed-linear models which account for underlying population structure using PCA of SNP data, in our case the first three principal components, and incorporating estimates of relatedness between accessions by producing a VanRaden kinship matrix. However, in EAA using GWAS programs, phenotype data is substituted with environmental data, adding the assumption that local adaptation is occurring in the studied population. Significant SNPs were corrected for multiple testing with False Discovery Rate (FDR) adjusted p values and use of the Bonferroni multiple test correction threshold. Manhattan plots (Figure S1 in File S1) and quantile-quantile (QQ) plots (Figure S2 in File S1) were produced to visualize the model outputs.

## EAA SNP regions, gene identification and analysis

SNPs significantly associated with the climate variables are often clustered into regions due to linkage disequilibrium (LD), including putatively neutral SNPs that are in close proximity to a selectively advantageous SNP variant. Regions were identified based on the 0.05 FDR corrected p-value and Bonferroni significance thresholds. Additional regions were identified from single SNPs if they were highly significant, passing the Bonferroni threshold. For gene identification, regions were extended by 10 kb to capture those that may be linked to the region by LD. Genes within regions were determined by comparing the chromosomal locations to the IRGSP Ver1.0 genome browser with MSU Osa1 gene model tracks (<http://viewer.shigen.info/oryzagenome2/mapview/MapView.do>).

Gene ontology (GO) term enrichment analysis was carried out on the genes present in the genomic regions identified from the EAA using AgriGO V2.0 (Tian *et al.*, 2017), selecting singular enrichment analysis (SEA), an FDR significance threshold of 0.05 and using the MSU 7.0 gene ID reference consisting of 24075 genes with annotated “Complete GO” GO terms. Two sets of genes were analyzed, those associated with at least one temperature variable (BIO1, 4, 5, 8 and PC1) and those associated with at least one precipitation variable (BIO8, 13, 14 and PC1). Some overlap occurs between these categories due to a subset of the environmental variables being linked to both temperature and precipitation metrics. The temperature category input 59 genes, with 28 being annotated with GO terms, and the precipitation group inputting 52 genes with 24 annotated. Genic SNPs were also identified through alignment to the MSU Osa1 gene model and non-synonymous SNPs assigned a Grantham score to determine the impact of amino acid sequence changes (Grantham, 1974).

# Results

## *Oryza rufipogon* is distributed across a diverse range of environments

The climate variables comprise 19 bioclimatic variables and elevation. Across the sample locations, average annual temperature (BIO1) ranges from 21.4°C to 28.1°C, generally decreasing with latitude (Figure 1a). Annual precipitation (BIO12) displays a more heterogenous pattern than the temperature variables and varies 3-fold across the sample range, between ca. 820 mm and 2430 mm (Figure 1b). There is a high degree of correlation between the climatic variables, latitude and longitude, with strong correlations between most temperature-associated variables and between precipitation variables (Figure S3 in File S1). Based on this correlation data, groups of highly correlated variables (ρ ≤ -0.8; ρ ≥ 0.8) were reduced to a single variable for the EAA, resulting in analysis of 13 variables.

Composite variables were derived from the climate data using PCA, and the first and second PCs used in the EAA. The greatest loadings for PC1 includes BIO1, 4, 9, 11, 14 and 17 (Figure 2; Table S2), which are measurements of temperature or precipitation in dry periods. Loadings for PC2 include BIO3, 6, 15 and 19 which are a variety of measurements related to isothermality and seasonality. In the environmental PCA (Figure 2), samples with low PC1/intermediate PC2 values are predominantly from China, with a cooler climate, large annual temperature variation and consistently higher precipitation. Samples with high PC1/intermediate PC2 values are from hot climates with average annual temperature range and maximum temperature, and low precipitation. The samples with the greatest PC2 have consistently warm temperatures throughout the year, and samples with low PC2 values are from environments with high annual temperature variation and the highest maximum temperatures.

Previous population structure analyses on the SNP dataset using neighbor-joining method revealed three genetic groups, Or-I, II and III (Huang *et al*., 2012). These show patterns of geographic distribution, with Or-III predominantly found in China and west India, and Or-I and II displaying a more mixed distribution. This is reproduced in a PCA of the SNP data, revealing that the samples originating from China are more genetically distant and that the two genetic groups display mixed origins (Figure S4 in File S1).

## Identification of genomic regions linked to environmental variables

The EAA was run using GAPIT, based on 1,898,737 SNPs across 286 individuals. Overall, 8 of 13 variables produced significant associations (Figure S1 in File S1). Variables that failed to produce significant associations in the EAA included PC2, elevation, annual range in temperature (BIO7), annual precipitation (BIO12) and precipitation seasonality (BIO15). PC1 produced associations similar to other highly correlated temperature variables and no genomic regions were associated with PC2.

Regions (R) were named by chromosome and then sequentially if more than one region was present on a chromosome. In total, 15 regions across 9 of the 12 chromosomes were identified, with each passing at least the FDR threshold and most passing the more stringent Bonferroni threshold (Table 1). Two of these associations represent significant associations for two or more of the environmental variables. For example, environmentally associated region 2.1 (R2.1) was identified on chromosome 2 for variables PC1, BIO1 and BIO4 (Table 1), all of which are various measurements of temperature. Similarly, four variables were associated with region R5.1.

The size of each significant region varied considerably, from a single significantly associated SNP to large regions of over 100 kb containing several significant SNPs after FDR correction. This could be a result of differences in sequence coverage or imputation in these regions, degree of linkage disequilibrium, or the presence of multiple genes that are associated with the environmental variables.

**Identification of genes linked to environmental variables**

Overall, 86 genes were present within or close to (within 10 kb) the 15 genomic regions (Table S3). Several of these genes demonstrate putative functions that could be involved in local adaptation. Of the genes located in the environmentally associated regions, seven contain an FDR-corrected SNP within their coding sequence, and three of these contain one or more non-synonymous SNP (LOC\_Os01g05640, LOC\_Os01g67500 and LOC\_Os05g27670; Table 2). Grantham scores indicate that two SNPs present in a receptor-like protein kinase (LOC\_Os01g05640) will have low impact on the structure, whereas the serine to cysteine substitution for the armadillo/beta-catenin repeat family protein (LOC\_Os01g67500) and premature stop codon in the uncharacterized protein (LOC\_Os05g27670) will likely have a larger effect on the protein function. A GO enrichment analysis was also conducted, comparing the GO terms of genes located within regions associated with temperature or precipitation variables against a reference dataset of GO terms. However, no significant enrichment was detected after correcting for multiple tests (Table S4).

**Chromosome 1**

Three environmentally associated regions were detected in Chromosome 1. R1.1 was associated with the precipitation of the driest month, another (R1.2) shows association with precipitation of the wettest month, and the third (R1.3) is in association with the warmest month.

R1.1 contains a receptor-like protein kinase of unknown function, a disease resistance (R) gene *PYRICULARIA ORYZAE RESISTANCE T* (*PIT*; LOC\_Os01g05630) and a metallothionein gene (*MT2a*; LOC\_Os01g05650) ca. 2kb downstream of this region. R1.2 is ca. 6 kb from a sulphate transporter (*SULTR3;6*; LOC\_Os01g52130), which has previously been shown to be upregulated in drought and salinity stresses (Kumar *et al.*, 2011). Additional genes within the region include those encoding hypothetically expressed proteins and a heavy metal associated protein (HMP5; LOC\_Os01g52160). Finally, R1.3 contains three genes with the annotations DNA binding protein (LOC\_Os01g67480), OTU-like cysteine protease (LOC\_Os01g67490), armadillo/beta-catenin repeat family protein (LOC\_Os01g67500) and a GDP-L-galactose phosphorylase (*GGP*; LOC\_Os01g67520) is located downstream of this region and is predicted to function in the biosynthesis of ascorbic acid (Akram *et al.*, 2017).

**Chromosome 2**

Three environmentally associated regions were detected in chromosome 2. R2.1 is associated with several highly correlated temperature-associated variables. This region contains several retrotransposons and a gene encoding an F-box/Kelch repeat protein (*OsFBOX84*; LOC\_Os02g21110), which is differentially expressed under light and dark conditions (Jain *et al.*, 2007). The non-reference allele is predominantly in regions with low PC1 values, which corresponds to regions with lower temperatures.

The final two regions in chromosome 2 are associated with temperature of the wettest quarter and mean diurnal range, respectively. R2.2 contains an uncharacterized gene encoding an expressed protein (LOC\_Os02g35860) and R2.3 has a terpene synthase gene, ent-isokaur-15-ene synthase (*OsKS6*; LOC\_Os02g36264), which is located 1.9 kb upstream of the single significant SNP in this region.

**Chromosome 3**

In chromosome 3, a single significant SNP peak was detected and associated with the mean temperature of the wettest quarter (BIO8). This SNP is near a gene encoding an ethylene receptor, *ETHYLENE RESPONSE SENSOR 1* (*OsERS1*; LOC\_Os03g49500).

**Chromosome 5**

R5.1 is associated with the precipitation of the driest month (BIO14), PC1, annual temperature and temperature seasonality. This region contains a single gene encoding an expressed protein (LOC\_Os05g27680) and a transposon protein.

**Chromosome 6**

In chromosome 6, two significant regions were detected. R6.1, associated with the precipitation of the wettest month (BIO13), contains a predicted growth regulator, O-fucosyltransferase (LOC\_Os06g13215) and R6.2, associated with the mean temperature of the wettest quarter (BIO8), is close to a predicted MADS-box transcription factor (LOC\_Os06g40609) and an ESCRT-III complex SNF7 component (LOC\_Os06g40620).

**Chromosome 7**

R7.1 was found to be associated with the maximum temperature of the warmest month (BIO5). This large 147 kb region contains 26 genes, 22 of which are genes of unknown function or transposons. Genes with functional annotations include genes putatively related to flowering time (*OsGHD7-like*; LOC\_Os07g15770), reactive oxygen species (ROS) antioxidant activity (LOC\_Os07g15670) and leaf and spikelet development (*NAL8*; LOC\_Os07g15880). *OsGhd7* is linked to adaptation of rice to higher latitudes (Koo *et al.*, 2013) and a SNP within *Ghd7* associated with BIO5 is found in accessions located in environments with high maximum temperatures (Figure 3).

**Chromosome 8**

R8.1 is associated with precipitation of the wettest month (BIO13) and contains a glutamate decarboxylase (*OsGAD1;* LOC\_Os08g36320), with the SNP allele present in accessions found in climates with relatively low precipitation in their wettest month. *OsGAD1* has high sequence similarity to *AtGAD1* (AT5G17330: 94% coverage; 86.97% identity), which is implicated in the response to heat, drought and hypoxia stresses, through production of GABA (Kinnersley and Turano, 2000).

**Chromosome 10**

Analysis of the precipitation of the driest month produced an additional SNP peak in chromosome 10 (R10.1) near genes encoding ubiquitin-conjugating enzyme 10 (*OsUBC10*; LOC\_Os10g31000) and mRNA adenosine methylase 4 (*OsMTA4*; LOC\_Os10g31030). Although the function of these proteins in rice is largely unknown, expression of *OsUBC10* is upregulated by the phytohormone abscisic acid (E *et al.*, 2015).

**Chromosome 12**

In chromosome 12 there are two significant regions both associated with BIO14. R12.1 consists of three genes with unknown functions and R12.2 does not contain any genes within 10 kb. The closest gene is 11.2 kb downstream and is predicted to encode a 9-cis-epoxycarotenoid dioxygenase (NCED2; LOC\_Os12g42280).

Overall, a large proportion of the environmentally associated SNPs were near genes with unknown functions, highlighting the importance of gene annotation in model systems. Despite this, 11 of the 19 regions identified contain genes with putative functions linked to response to abiotic/biotic conditions or hormone signaling. This included genes known to be regulated by or implicated in ethylene and abscisic acid (ABA) signaling and the GABA shunt pathway. Genes functioning in response to temperature, drought, UV and hypoxia were present in several of the significant regions. Whilst there was no significant enrichment of GO terms, this may indicate that the various adaptations are controlled by a range of processes.

# Discussion

The aim of the analysis was to identify genetic markers associated with variation in environmental conditions across the native range of *O. rufipogon*. Genes predicted to be in close proximity to these markers were identified and their biological functions analyzed bioinformatically. This has allowed us to detect genomic regions potentially involved in the local adaptation of *O. rufipogon* and to highlight genes and mechanisms predicted to contribute to response to environmental conditions.

The diversity panel used in the analysis extends across most of the species’ range, although with limited sampling in Indonesia and west India. The environmental dataset was summarized with a PCA, revealing that variables contributing the most to the PC1 loadings include those relating to temperature. This is similar to a previous analysis of Chinese *O. rufipogon*, which found that the variables annual temperature, temperature seasonality and mean annual radiation were the main factors contributing to variation in climate across the population (Zhou *et al.*, 2013). Whilst some of the genome regions identified were associated with a single variable, others were associated with multiple (e.g., R2.1 and R5.1), meaning the specific environmental factor underlying the association cannot be confirmed, and in many cases multiple overlapping EAA regions are likely due to the correlation between variables. It should also be noted that the climate variables may also correlate with variables not analyzed in this investigation, such as soil factors, exposure or salinity. Here, we focus on genes potentially underlying local adaptation to temperature and precipitation.

## Detection of the heading date gene *Ghd7*, known to be involved in local adaptation in domesticated rice

One of the candidate genes in a region associated with temperature variables encodes the heading date gene *Ghd7*. In *O. sativa, Ghd7* is a key component of the flowering time pathway (Shrestha *et al.*, 2014). *S*equence differences in *Ghd7* are associated with adaptation to long-day conditions (Shrestha *et al.*, 2014) and naturally occurring haplotypes of *OsGhd7* have been associated with variation in heading dates and latitudes (Xue *et al.*, 2008; Koo *et al.*, 2013; Cui *et al.*, 2020). Cultivars from the most northern latitudes (beyond the northern limits of *O. rufipogon*) are often associated with non-functional *Ghd7* variants (Xue *et al.*, 2008).

In our analysis, populations with the non-reference allele were from locations with high maximum temperatures. Because high temperature stress during reproductive stages causes sterility (at least in domesticated rice; Jagadish *et al.*, 2007), natural variation in *Ghd7* could alter flowering time, functioning to escape high temperature stress. Lifecycle shifts for stress avoidance are known to occur in *A. thaliana* in response to a broad range of abiotic and biotic factors (Kazan and Lyons, 2015), and in the wild oat species *Avena barbata* Pott ex Link (Sherrard and Maherali, 2006). *OsGhd7* is also responsive to other stress conditions (Weng *et al.*, 2014; Du *et al.*, 2018) and hence could be involved in adaptation to other conditions.

## Detection of phytohormone-associated genes

Some of the EAA candidate genes are associated with phytohormones, which are often involved in stress response and developmental processes, for example ABA (Shi and Yang, 2014; Muhammad Aslam *et al*., 2022) and ethylene (Müller and Munné-Bosch, 2015). *OsGhd7*, mentioned before, is repressed by ABA and jasmonic acid (Weng *et al.*, 2014). The gene *ERS1* is a predicted negative regulator of ethylene signaling, influences root length in *O. sativa* (Ma *et al.*, 2014) and is upregulated in salinity (Hossain *et al.*, 2016).

Finally, the ubiquitin conjugating enzyme UBC10 is associated with precipitation of the driest month and quarter. UBCs are involved in the ubiquitination of target proteins (Ciechanover, 1994) and can function in a variety of processes including developmental processes, hormone signaling (Dreher and Callis, 2007), flowering time (Xu *et al.*, 2009) and abiotic stress response (Lyzenga and Stone, 2011). In *O. sativa*, *UBC10* expression is increased in response to ABA (E *et al.*, 2015), which is usually associated with the regulation of developmental processes (Rodríguez-Gacio *et al.*, 2009) and response to abiotic stresses (Seki *et al.*, 2002). In rice, ABA has also been linked to drought escape pathways through regulation of flowering time genes (Du *et al.*, 2018). Therefore, UBC10 may function in ABA signaling to regulate developmental or stress response processes.

## Detection of abiotic stress related genes

Some of the EAA candidate genes have possible functions in response to various abiotic stresses. A previous drought study in *O. sativa* identified large transcriptional changes in root and shoot tissue between drought and control conditions (Groen *et al.*, 2021). Four of these differentially expressed genes were identified in this EAA, *OsMT2a*, a cytochrome p450, an expressed protein and *GAD1*. In addition to differential expression under drought, two genes are induced under other stress conditions too; *OsMT2a* expression is induced by heat shock (Hsieh *et al.*, 1996) and *SULTR3;6* is responsive to salinity stresses (Kumar *et al.*, 2011). It is possible that these genes confer local adaptation in *O. rufipogon*.

*GAD1* encodes an enzyme in the GABA-shunt pathway and was detected in association with precipitation. The GABA-shunt pathway is branched from the TCA cycle and is thought to function to support the respiratory metabolic system, as well as during stress responses by promoting seedling growth (Lee *et al.*, 2021) and though ROS scavenging (Bouché and Fromm, 2004), a mechanism used to maintain ROS homeostasis (Gill and Tuteja, 2010). *GAD1* in *O. sativa* is upregulated during heat (Cao *et al.*, 2013; Liao *et al.*, 2015), drought (Groen *et al.*, 2021) and cold stresses (Zeng *et al.*, 2022), supporting the association identified in the EAA.

Finally, *FBOX84*, predominantly associated with temperature variables in our EAA, is downregulated in white light compared to dark conditions, but not for other stresses (Jain *et al.*, 2007). It is possible that these environments correlate with UV, exposure or shading, although the specific function of this gene remains unknown.

## Evaluation of the EAA method

It should be noted that GWAS-based methods have reduced detection power when analyzing complex traits as these are likely controlled by multiple additive loci with weak effects. Although the number of accessions is reasonably high in this study, the SNP sequence data is limited by low coverage, which would have provided less detection power in certain regions. This would especially occur if these regions are characterized by higher recombination rates breaking up the association between SNP alleles and causative genes. Due to the low coverage, the SNP data does not include heterozygous sites, which will be present in a wild species such as this, and so missing possible associations. The proportion of heterozygous sites has previously been calculated to range between 0.1% and 5.4% for five *O. rufipogon* accessions with greater than 9x sequencing depth (Huang *et al*., 2012). Further, the resequencing data was mapped to an older congener reference genome than is available now (e.g., Xie *et al*., 2021;Que *et al*., 2022) and so there may be genes in *O. rufipogon* that are missing from our analysis, although this can occur even within species (Zhao *et al*., 2018). Therefore, in future investigations, greater sequencing depth and either a new reference genome or conspecific reference could be used. Despite these limitations, the *O. rufipogon* SNP dataset has previously been successfully analyzed using GWAS for traits such as tiller angle (Huang *et al*., 2012). The removal of multiple accessions during filtering steps may also reduce detection of associations, however, it increases the likelihood that the significant SNPs detected here are true positive results, providing a more conservative approach of detecting SNP-environment associations. Optimization of EAA analyses more broadly in terms of outlier removal, both in terms of SNPs and individuals, should be prioritized to ensure these analyses achieve the optimal balance between removing false positives and reducing false negatives.

The EAA only detects associated genomic loci and is unable to detect transcriptional responses to the environment. Related to this, linkage disequilibrium (LD) can result in the causative gene being distant from a significantly associated SNP, adding to the challenge of confidently selecting candidate genes. LD decays relatively quickly in rice, dropping to half at around 20 kb (Huang *et al*., 2012), therefore we assume that the causative genetic changes are physically relatively close to the outlier SNPs identified. The association between *Ghd7*, a gene known to be involved in local adaptation, and maximum temperature in the EAA (see Figure 3) provides confidence that at least a subset of our associated genes and regions are true positive associations.

The absence of significant GO terms may be explained by LD; only a single gene may be causative in each environmentally associated region, but our GO analysis necessarily included all genes in EAA regions. In addition, for a significant portion of the associated genes the annotations were weak, hence some adaptively important genes may have been uncovered but remain with unknown function. Regardless of this limitation, associated regions could be used in marker assisted selective breeding.

Taking these caveats into account, the *O. rufipogon* EAA revealed 15 regions of the genome associated with environmental variables and are predicted to contain several genes with functions potentially associated with abiotic stress, flowering, developmental processes and hormone signaling. The investigation looks at the association between environmental conditions and genetic variation, and consequently focuses on the abiotic component of local adaptation. Therefore, additional genomic regions may be detected in association with biotic factors. Although requiring confirmatory analysis, these loci provide significant potential for adaptive variation in the wild progenitor of one of the world’s most widely grown crops. Going forward, alleles at these genes have the potential for being incorporated into breeding programs to ensure rice adaptation for a future climate.

# References

Akram, N. A., Shafiq, F. & Ashraf, M. (2017). Ascorbic acid-a potential oxidant scavenger and its role in plant development and abiotic stress tolerance. *Frontiers in Plant Science,* 8.

Anderson, J. E., Kono, T. J. Y., Stupar, R. M., Kantar, M. B. & Morrell, P. L. (2016). Environmental association analyses identify candidates for abiotic stress tolerance in *Glycine soja*, the wild progenitor of cultivated soybeans. *G3 Genes|Genomes|Genetics,* 6**,** 835-843.

Barbier, P. (1989). Genetic variation and ecotypic differentiation in the wild rice species *Oryza rufipogon*. II. Influence of the mating system and life-history traits on the genetic structure of populations. *Japanese Journal of Genetics*, 64, 273-285.

Bouché, N. & Fromm, H. (2004). GABA in plants: just a metabolite? *Trends in Plant Science,* 9**,** 110-115.

Caicedo, A. L., Williamson, S. H., Hernandez, R. D., Boyko, A., Fledel-Alon, A., York, T. L.*, et al.* (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS genetics,* 3**,** 1745-1756.

Cao, Y., Zhang, Q., Chen, Y., Zhao, H., Lang, Y., Yu, C.*, et al.* (2013). Identification of differential expression genes in leaves of rice (*Oryza sativa* L.) in response to heat stress by cDNA-AFLP analysis. *BioMed Research International,* 2013**,** 576189.

Chang, T. (1976). The origin, evolution, cultivation, dissemination, and diversification of Asian and African rices. *Euphytica,* 25**,** 425-441.

Ciechanover, A. (1994). The ubiquitin-proteasome proteolytic pathway. *Cell,* 79**,** 13-21.

Cui, Y., Wang, J., Feng, L., Liu, S., Li, J., Qiao, W.*, et al.* (2020). A combination of long-day suppressor genes contributes to the northward expansion of rice. *Frontiers in Plant Science,* 11.

D’alpoim Guedes, J., Jin, G. & Bocinsky, R. K. (2015). The Impact of climate on the spread of rice to North-Eastern China: A new look at the data from Shandong province. *PLOS ONE,* 10**,** e0130430.

De La Torre, A. R., Wilhite, B. & Neale, D. B. (2019). Environmental genome-wide association reveals climate adaptation is shaped by subtle to moderate allele frequency shifts in Loblolly pine. *Genome Biology and Evolution,* 11**,** 2976-2989.

Dodson, J., Hung, H., Li, C., Li, J., Lu, F. & Yan, H. (2021). The Probable Critical Role of Early Holocene Monsoon Activity in Siting the Origins of Rice Agriculture in China. *Frontiers in Earth Science,* 9.

Dreher, K. & Callis, J. (2007). Ubiquitin, hormones and biotic stress in plants. *Annals of Botany,* 99**,** 787-822.

Du, H., Huang, F., Wu, N., Li, X., Hu, H. & Xiong, L. (2018). Integrative regulation of drought escape through ABA-dependent and -independent pathways in rice. *Molecular Plant,* 11**,** 584-597.

E, Z., Zhang, Y., Li, T., Wang, L. & Zhao, H. (2015). Characterization of the ubiquitin-conjugating enzyme gene family in rice and evaluation of expression profiles under abiotic stresses and hormone treatments. *PLOS ONE,* 10**,** e0122621.

FAOSTAT. (2020). www.fao.org

Flint-Garcia, S. A. (2013). Genetics and consequences of crop domestication. *Journal of Agricultural and Food Chemistry,* 61**,** 8267-8276.

Gao, L., Li, D., Wu, X., Wen, C., Huang, Z. & Wei, X. (2012). *In situ* conservation of wild rice populations: A targeted study of common wild rice *Oryza rufipogon* from China. *American Journal of Plant Sciences,* 3**,** 854-868.

Gill, S. S. & Tuteja, N. (2010). Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. *Plant Physiology and Biochemistry,* 48**,** 909-930.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science,* 185**,** 862-4.

Groen, S. C., Joly-Lopez, Z., Platts, A. E., Natividad, M., Fresquez, Z., Mauck, W. M., Iii*, et al.* (2021). Evolutionary systems biology reveals patterns of rice adaptation to drought-prone agro-ecosystems. *The Plant Cell,* 34**,** 759-783.

Gutaker, R. M., Groen, S. C., Bellis, E. S., Choi, J. Y., Pires, I. S., Bocinsky, R. K.*, et al.* (2020). Genomic history and ecology of the geographic spread of rice. *Nature Plants,* 6**,** 492-502.

Hijmans, R., Etten, J. V., Sumner, M., Cheng, J., Baston, D., Bevan, A.*, et al.* (2020). Geographic data analysis and modeling. Version 3.4-5. https://rspatial.org/raster

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology,* 25**,** 1965-1978.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., *et al*. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*. 188, 379-397.

Hossain, M. R., Bassel, G. W., Pritchard, J., Sharma, G. P. & Ford-Lloyd, B. V. (2016). Trait specific expression profiling of salt stress responsive genes in diverse rice genotypes as determined by modified significance analysis of microarrays. *Frontiers in Plant Science,* 7.

Hsieh, H., Liu, W., Chang, A. & Huang, P. C. (1996). RNA expression patterns of a type 2 metallothionein-like gene from rice. *Plant Molecular Biology,* 32**,** 525-529.

Huang, X., Kurata, N., Wei, X., Wang, Z. X., Wang, A., Zhao, Q.*, et al.* (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature,* 490**,** 497-501.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y.*, et al.* (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics,* 42**,** 961-967.

Jagadish, S., Craufurd, P. & Wheeler, T. (2007). High temperature stress and spikelet fertility in rice (Oryza sativa L.). *Journal of Experimental Botany,* 58**,** 1627-1635.

Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P.*, et al.* (2007). F-Box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant Physiology,* 143**,** 1467-1483.

Jovovic, Z., Andjelkovic, V., Przulj, N. & Mandic, D. (2020). Untapped genetic diversity of wild relatives for crop improvement. *In:* SALGOTRA, R. K. & ZARGAR, S. M. (eds.) *Rediscovery of Genetic and Genomic Resources for Future Food Security.* Singapore: Springer Singapore.

Jump, A. S., Marchant, R. & Peñuelas, J. (2009). Environmental change and the option value of genetic diversity. *Trends in Plant Science,* 14**,** 51-58.

Kassambara, A. (2019). Visualization of a correlation matrix using ‘ggplot2’. Version 0.1.3. http://www.sthda.com/english/wiki/ggcorrplot

Kazan, K. & Lyons, R. (2015). The link between flowering time and stress tolerance. *Journal of Experimental Botany,* 67**,** 47-60.

Kinnersley, A. M. & Turano, F. J. (2000). Gamma Aminobutyric Acid (GABA) and plant responses to stress. *Critical Reviews in Plant Sciences,* 19**,** 479-509.

Koo, B., Yoo, S., Park, J., Kwon, C., Lee, B., An, G.*, et al.* (2013). Natural variation in OsPRR37 regulates heading date and contributes to rice cultivation at a wide range of latitudes. *Molecular Plant,* 6**,** 1877-1888.

Kumar, S., Asif, M. H., Chakrabarty, D., Tripathi, R. D. & Trivedi, P. K. (2011). Differential expression and alternative splicing of rice sulphate transporter family members regulate sulphur status during plant growth, development and stress conditions. *Functional & Integrative Genomics,* 11**,** 259-273.

Lee, J., Kim, S., Kim, S. & Shim, I. (2021). Production of γ-Aminobutyric acid and its supplementary role in the TCA cycle in rice (*Oryza sativa* L.) seedlings. *Journal of Plant Growth Regulation,* 40**,** 78-90.

Lei, L., Poets, A. M., Liu, C., Wyant, S. R., Hoffman, P. J., Carter, C. K.*, et al.* (2019). Environmental association identifies candidates for tolerance to low temperature and drought. *G3 Genes|Genomes|Genetics,* 9**,** 3423-3438.

Li, C., Zhou, A. & Sang, T. (2006). Genetic analysis of rice domestication syndrome with the wild annual species, *Oryza nivara*. *New Phytologist,* 170**,** 185-194.

Liao, J., Zhou, H., Peng, Q., Zhong, P., Zhang, H., He, C.*, et al.* (2015). Transcriptome changes in rice (*Oryza sativa* L.) in response to high night temperature stress at the early milky stage. *BMC Genomics,* 16**,** 18.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J.*, et al.* (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics,* 28**,** 2397-2399.

Lyzenga, W. J. & Stone, S. L. (2011). Abiotic stress tolerance mediated by protein ubiquitination. *Journal of Experimental Botany,* 63**,** 599-616.

Ma, B., Yin, C., He, S., Lu, X., Zhang, W., Lu, T.*, et al.* (2014). Ethylene-induced inhibition of root growth requires abscisic acid function in rice (*Oryza sativa* L.) seedlings. *PLOS Genetics,* 10**,** e1004701.

Muhammad Aslam, M., Waseem, M., Jakada, B. H., Okal, E. J., Lei, Z., Saqib, H. S. A.*, et al.* (2022). Mechanisms of abscisic acid-mediated drought stress responses in plants. *International Journal of Molecular Sciences,* 23**,** 1084.

Müller, M. & Munné-Bosch, S. (2015). Ethylene response factors: A key regulatory hub in hormone and stress signaling. *Plant physiology,* 169**,** 32-41.

Muthayya, S., Sugimoto, J. D., Montgomery, S. & Maberly, G. F. (2014). An overview of global rice production, supply, trade, and consumption. *Annals of the New York Academy of Sciences,* 1324**,** 7-14.

Ohyanagi, H., Ebata, T., Huang, X., Gong, H., Fujita, M., Mochizuki, T.*, et al.* (2015). OryzaGenome: Genome diversity database of wild *Oryza* species. *Plant and Cell Physiology,* 57**,** e1-e1.

Phan, P. D. T., Kageyama, H., Ishikawa, R. and Ishii, T. (2012). Estimation of the outcrossing rate for annual Asian wild rice under field conditions. *Breed Sci*. 62, 256-262

Que, Z., Lu, Q. and Shen, C. (2022). Chromosome-level genome assembly of Dongxiang wild rice (*Oryza rufipogon*) provides insights into resistance to disease and freezing. *Frontiers in Genetics*, 13:1029879.

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available: www.R-project.org/

Ray, D. K., Gerber, J. S., Macdonald, G. K. & West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nature Communications,* 6**,** 5989.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology,* 24**,** 4348-4370.

Rodríguez-Gacio, M. D. C., Matilla-Vázquez, M. A. & Matilla, A. J. (2009). Seed dormancy and ABA signaling: the breakthrough goes on. *Plant signaling & behavior,* 4**,** 1035-1049.

Savolainen, O., Lascoux, M. & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics,* 14**,** 807-820.

Seki, M., Ishida, J., Narusaka, M., Fujita, M., Nanjo, T., Umezawa, T.*, et al.* (2002). Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray. *Functional & Integrative Genomics,* 2**,** 282-291.

Sherrard, M. E. & Maherali, H. (2006). The adaptive significance of drought escape in *Avena barbata*, an annual grass. *Evolution,* 60**,** 2478-2489.

Shi, Y. & Yang, S. (2014). ABA regulation of the cold stress response in plants. *In:* ZHANG, D.P. (ed.) *Abscisic Acid: Metabolism, Transport and Signaling.* Dordrecht: Springer Netherlands.

Shrestha, R., Gómez-Ariza, J., Brambilla, V. & Fornara, F. (2014). Molecular control of seasonal flowering in rice, arabidopsis and temperate cereals. *Annals of Botany,* 114**,** 1445-1458.

Tao, F., Zhang, Z. & Yokozawa, M. (2011). Dangerous levels of climate change for agricultural production in China. *Regional Environmental Change,* 11**,** 41-48.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z.*, et al.* (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research,* 45**,** W122-W129.

Vaughan, D. A. (1994). *The wild relatives of rice: a genetic resources handbook*, International Rice Research Institute.

Weng, X., Wang, L., Wang, J., Hu, Y., Du, H., Xu, C.*, et al.* (2014). Grain Number, Plant Height, and Heading Date7 Is a Central Regulator of Growth, Development, and Stress Response. *Plant Physiology,* 164**,** 735-747.

Xie, X., Du, H., Tang, H., Tang, J., Tan, X, Liu, W. *et al*. (2021). A chromosome-level genome assembly of the wild rice *Oryza rufipogon* facilitates tracing the origins of Asian cultivated rice. *Science China Life Sciences*, 64, 282-293.

Xu, L., Ménard, R., Berr, A., Fuchs, J., Cognat, V., Meyer, D.*, et al.* (2009). The E2 ubiquitin-conjugating enzymes, AtUBC1 and AtUBC2, play redundant roles and are involved in activation of FLC expression and repression of flowering in Arabidopsis thaliana. *The Plant Journal,* 57**,** 279-288.

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L.*, et al.* (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nature Genetics,* 40**,** 761-767.

Zeng, Z., Zhang, S., Li, W., Chen, B. & Li, W. (2022). Gene-coexpression network analysis identifies specific modules and hub genes related to cold stress in rice. *BMC Genomics,* 23**,** 251.

Zhang, H., Mittal, N., Leamy, L. J., Barazani, O. & Song, B. H. (2017). Back into the wild—Apply untapped genetic diversity of wild relatives for crop improvement. *Evolutionary Applications,* 10**,** 5-24.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q. *et al*. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, 50, 278-284.

Zhou, W., Wang, Z., Davy, A. J. & Liu, G. (2013). Geographic variation and local adaptation in *Oryza rufipogon* across its climatic range in China. *Journal of Ecology,* 101**,** 1498-1508.

# Tables

|  |
| --- |
| **Table 1 –** Fifteen environmentally associated regions, each assigned a chromosome/region ID. Listed are the associated trait(s), the genomic position of region boundaries, whether the region is formed from a single SNP or forms a peak region of multiple SNPs, the number of genes found within each region or in/within 10 kbp of the region, excluding transposable elements, details on the most significant SNP in each region and the most stringent threshold passed. |
| **Chr/Region ID** | **Associated trait** | **Region start position (kbp)** | **Region end position (kbp)** | **Region size (kbp)** | **Region type** | **Gene count (in region)** | **Gene count (within 10 kb)** | **Lowest FDR adjusted p value** | **Threshold passed** |
| R1.1 | BIO14 | 2688 | 2691 | 3 | region | 1 | 5 | 2.8x10-4 | Bonferroni |
| R1.2 | BIO13 | 30002 | 30003 | 1 | single | 0 | 5 | 1.2x10-2 | Bonferroni |
| R1.3 | BIO5 | 39215 | 39235 | 20 | region | 1 | 5 | 8.4x10-3 | FDR |
| R2.1 | PC1; BIO1; BIO4 | 12523 | 12550 | 27 | region | 1 | 2 | 1.1x10-2 | Bonferroni1 |
| R2.2 | BIO8 | 21547 | 21548 | 1 | region | 0 | 2 | 6.2x10-3 | Bonferroni |
| R2.3 | BIO2 | 21901 | 21902 | 1 | region | 0 | 3 | 2.2x10-3 | Bonferroni |
| R3.1 | BIO8 | 28175 | 28176 | 1 | single | 0 | 3 | 1.7x10-2 | FDR |
| R5.1 | PC1; BIO1; BIO4; BIO14 | 16122 | 16140 | 18 | region | 2 | 3 | 2.8x10-4 | Bonferroni2 |
| R6.1 | BIO13 | 7257 | 7258 | 1 | single | 0 | 4 | 1.2x10-2 | Bonferroni |
| R6.2 | BIO8 | 24219 | 24220 | 1 | single | 0 | 6 | 7.4x10-3 | Bonferroni |
| R7.1 | BIO5 | 9098 | 9247 | 147 | region | 7 | 9 | 1.0x10-3 | Bonferroni |
| R8.1 | BIO13 | 22882 | 22883 | 1 | single | 1 | 4 | 1.2x10-2 | Bonferroni |
| R10.1 | BIO14 | 16192 | 16195 | 3 | region | 0 | 2 | 2.8x10-4 | Bonferroni |
| R12.1 | BIO14 | 4400 | 4425 | 25 | region | 3 | 4 | 2.6x10-3 | Bonferroni |
| R12.2 | BIO14 | 26256 | 26257 | 1 | region | 0 | 0 | 4.3x10-3 | Bonferroni |
| 1. PC1 (Bonferroni); BIO1 (Bonferroni); BIO4 (FDR)2. PC1 (Bonferroni); BIO1 (FDR); BIO4 (Bonferroni); BIO14 (Bonferroni) |

|  |
| --- |
| **Table 2 -** details about the FDR-significant SNPs located within predicted coding sequences of seven genes and predicted sequence changes for the non-synonymous SNPs. |
| **Gene model** | **Strand** | **Annotation** | **SNP** | **Amino acid change** | **Grantham score** |
| LOC\_Os01g05640 | - | receptor-like protein kinase 5 precursor | ORRUF01\_02688980G:C |  |  |
| ORRUF01\_02689466G:A |  |  |
| ORRUF01\_02689630A:T | phe > ile | 21 |
| ORRUF01\_02690264T:C | lys > arg | 26 |
| ORRUF01\_02690615C:T |  |  |
| ORRUF01\_02690861A:G |  |  |
| ORRUF01\_02690885G:A |  |  |
| LOC\_Os01g67500 | - | armadillo/beta-catenin repeat family protein | ORRUF01\_39223051G:C | ser > cys | 112 |
| ORRUF01\_39224857G:A |  |  |
| LOC\_Os02g35860 | + | expressed protein | ORRUF02\_21547147T:C |  |  |
| LOC\_Os05g27670 | + | expressed protein | ORRUF05\_16124784G:A | trp > stop | \* |
| LOC\_Os06g13215 | + | growth regulator related protein | ORRUF06\_07257355C:T |  |  |
| LOC\_Os07g15680 | - | phospholipase D | ORRUF07\_09109350G:A |  |  |
| LOC\_Os08g36320 | - | Glutamate decarboxylase | ORRUF08\_22882132G:A |  |  |
| \*Premature stop codon results in a predicted 11 amino acid truncation. |

# Figures

**Figure 1 - (a)** Mean annual temperature (°C; BIO1) and **(b)** Mean annual precipitation (mm; BIO12) across the 286 *O. rufipogon* sample locations.

**Figure 2 -** Accession distribution based on primary and secondary principal components derived from the environmental data. Points correspond to individual rice accessions and arrows are the variable loadings, annotated with the variable name. Percentages represent the amount of variation explained by each axis.

**Figure 3 - (a)** Distribution of SNPs across the diversity panel. Points represent the approximate locations of accessions (overlapping points offset for clarity), separated by the SNP ORRUF07\_09153554T:C in the region R7.1 which is associated with BIO5 and contains 26 genes including the flowering time gene Ghd7. **(b)** Association between the SNP and BIO5 maximum temperature of the warmest month (°C). Each point represents an individual accession, with mean +/- 1 SD overlayed.

# Data availability

The full genome sequencing data from Huang *et al.* (2012) are available from OryzaGenome (http://viewer.shigen.info/oryzagenome/mapview/Top.do). The processed SNP and climate data from the subset of individuals we used in our analysis are available online (doi: 10.6084/m9.figshare.21916038.v1).

# Acknowledgements

The authors acknowledge the use of the IRIDIS High Performance Computing Facility at the University of Southampton in the completion of this work.

# Funding

This work was supported by the Natural Environmental Research Council for the INSPIRE DTP at the University of Southampton (grant number NE/S007210/1).

# Supplementary material

**File S1 -** Contains Figures S1 to S6

**Figure S1 -** SNPs associated with the 10 bioclimatic variables, principal components one and two, and elevation. Each point represents a SNP. The dashed line is the Bonferroni significance threshold and colored SNPs pass FDR.

**Figure S2 -** Observed versus expected SNP p-values for each environmental variable input into the GAPIT mixed linear model. The red line illustrates a 1:1 relationship.

**Figure S3 -** Correlations between environmental and location variables. The scale corresponds to the Spearman’s rank correlation coefficient rho values. T = temperature-associated variable; P = precipitation; B = both temperature and precipitation; O = other variables.

**Figure S4 -** Population structure of the 286 *O. rufipogon* accessions, resulting from a PCA of the SNP data. Each accession is colored by its country of origin.

**Table S1 -** Bioclimatic and elevation variable keys and descriptions.

**Table S2 -** Loading scores for each bioclimatic variable across the first 20 principal components

**Table S3 -** Details on the 86 genes identified in or within 10kb of a significant environmentally associated region. Region ID corresponds to the chromosome (R1.1 and R1.2 are both in chromosome 1). Region boundaries are to the nearest kilobase.

**Table S4 -** Enrichment scores for GO terms associated with genes found in or within 10 kb of temperature-associated genomic regions and precipitation-associated regions.