Genetic Analysis of Wild and Cultivated Populations of Northern Wild Rice (*Zizania palustris* L.) Reveal New Insights into Gene Flow and Domestication

Matthew W. Haas[1], Lillian McGilp[1], Mingqin Shao[1†], Reneth Millas[1], Claudia Castel-Miller[2], Laura M. Shannon[3], Anthony Kern[4], Jennifer A. Kimball[1*]

[1] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA;

[2] Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, USA;

[3] Department of Horticultural Science, University of Minnesota, St. Paul, MN 55108, USA;

[4] Department of Math and Science, University of Minnesota, Crookston, MN 56716, USA;

[†] Current address: Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

* Corresponding Author: jkimball@umn.edu

Keywords: *Zizania palustris*, Northern Wild Rice, genetic diversity, genotyping-by-sequencing, spatio-temporal diversity, linkage disequilibrium, selection

Abstract

Northern Wild Rice (NWR; *Zizania palustris* L.) is an aquatic annual grass with significant ecological, cultural, and economic importance to the Great Lakes region of the United States and Canada. Understanding the genetic diversity of NWR is essential to conservation efforts. In this study, we assembled and genotyped a diverse collection of 789 NWR individuals using genotyping-by-sequencing and obtained 5,955 single-nucleotide polymorphisms (SNPs). Our collection consisted of samples from 12 wild NWR populations, some of which were collected temporally, a representative sampling of cultivated NWR germplasm, and a *Zizania aquatica* population (outgroup). Using these data, we characterized the genetic diversity, relatedness, and population structure of this broad collection of NWR genotypes. We found that wild populations of NWR clustered primarily by their lake or river of origin as well as larger clustering patterns based on watershed assignment. Contextualizing some clustering patterns with historical knowledge of ecosystem management helped to unravel some of the complexity of the population structure of wild populations. Cultivated materials were genetically distinct from wild populations, suggesting limited gene flow between the semi-domesticated crop and its wild counterparts. Estimates of linkage disequilibrium (LD) in NWR demonstrated that LD decays quickly across the genome and provided insights into the quality of the reference genome assembly of NWR. The first genome-wide scans of putative selection events in cultivated NWR were also evaluated. Overall, this study presents a large set of new SNP markers for use in NWR genetic studies as well as new insights into the gene flow, history, and complexity of wild and cultivated populations of NWR.

Introduction

Northern wild rice (NWR; *Zizania palustris* L.) is an annual, diploid ($2n=2x=30$), aquatic grass native to the Eastern Temperate and Northern Forest ecoregions of North America (Elliot 1980; Grombacher et al., 1993). Predominantly found in shallow, slow-moving waterways surrounding the Great Lakes of the United States and Canada, NWR's regional significance in these areas is vast and complex (De Wet and Oelke, 1978; Drewes and Silbernagel, 2005; Lu et al., 2005; Myrbo et al., 2017; Biesboer, 2019; Desmarais, 2019). To begin, NWR is a natural resource that provides food and substantial habitat for a wide range of wildlife (Moyle, 1944; Fannucchi, 1983) as well as important ecosystem services, such as anchoring soils and inhibiting algal blooms (Rogosin, 1954). This abundant and nutritional grain did not go unnoticed by humans and has been hand-harvested from regional lakes and rivers for centuries (Grombacher, 1997). The First Peoples to consume the grain were Dakota and Anishinaabe. The grain is of particular importance to the Anishinaabe as NWR plays a central role in their origin story (Matson et al., 2021) and remains an integral component of their lives and culture today. The species has also been partially domesticated and commercialized, starting in the 1950s when Minnesota farmers began growing NWR in dike-irrigated paddies (Oelke, 1982). Today, cultivated NWR, sometimes called paddy rice, is considered a high-valued small commodity crop, grown primarily in Minnesota and California. Additionally, as a part of the *Oryzae* tribe in the *Poaceae* family, *Zizania* species are also considered crop wild relatives of *Oryza sativa* L. (white rice). Due to the many services described above, the conservation of NWR serves as an important intersection between our ecosystems, our cultures and food, and our economies.

Worldwide, plant species are experiencing declines and extinction events as a result of human activities that alter natural ecosystems. It has been widely reported that for at least the last

century NWR has been experiencing such declines in its native habitats (Norrgard, 2006; Hansen, 2008) and the species appears to be slowly migrating northward (Terrell et al., 1997). Recent reports have stated that NWR is at high risk of loss, and the International Union for Conservation of Nature has added NWR to their red list of threatened species (Maiz-Tome, 2016). Threats to the species include local hydrology changes due to damming and channelization, recreational water activity, shoreline development, and water pollution from industrial activities (Hansen, 2008; Myrbo et al., 2017). The species is particularly sensitive to high levels of sulfates in its water supply and acts as an important indicator species of water quality (Fort et al., 2014; Myrbo et al., 2017). In addition to these challenges, NWR seed is recalcitrant or desiccation intolerant, which reduces seed longevity in storage to 1-2 years (Probert et al., 1989) and eliminates the feasibility of storing NWR seed in *ex-situ* seed banks for its conservation and preservation. Therefore, the genetic diversity of NWR can currently only be preserved *in-situ* within its natural range (Porter, 2019).

Genetic diversity represents the extent of heritable variation within and among populations of a species and is necessary for populations to adapt to changes in their environment and maintain long-term viability (Frankel, 1970; Ellegren and Galtier, 2016). Biologists and conservationists have widely recognized the value of characterizing genome-wide diversity within a species for use in conservation efforts (Sgro et al., 2011; Santamaria, 2012). However, such studies in NWR have been limited and have heavily relied on outdated molecular marker systems (Lu et al., 2005; Kahler et al., 2014; Diller et al., 2018), which are laborious to produce and often limited in number (Collard et al., 2005). The advent of low-cost high-throughput sequencing has revolutionized the ability to characterize the genetic diversity in large germplasm collections through genotyping-by-sequencing (GBS) technologies, which produce

prolific numbers of co-dominant, single-nucleotide polymorphism (SNP) markers (Elshire et al., 2011). In 2019, a study with limited sample size demonstrated the potential for GBS to be applied to NWR (Shao et al., 2019).

In this study, we used GBS data to identify SNPs and conduct population genetic analyses on a diverse collection of NWR genotypes. The collection consisted of wild NWR populations gathered across northern Minnesota and Wisconsin (referred to here as Natural Stands), cultivars and germplasm from the University of Minnesota (UMN) NWR breeding, genetics, and conservation program (referred to here as Cultivated Material), and a population of *Zizania aquatica* L. A Temporal panel was also generated to look at the change in diversity over time between two lakes. Cultivated materials were included in this study as the crop is grown primarily within the species centers of origin and diversity and gene flow from cultivated materials is a primary concern to Dakota and Anishinaabe First Peoples.

Materials and Methods

Plant materials. The data we present here consisted of 789 individuals representing three different collections including a Natural Stand collection, a Temporal collection, and a Cultivated collection. Our Natural Stand collection consisted of 530 individual samples collected from 10 wild populations across northern Minnesota (50 samples per lake/river) and 2 wild populations from western Wisconsin (10 from Mud Hen Lake; 20 from Phantom Lake) (Figure 1; Table S1). An additional 50 *Z. aquatica* samples, also collected in northern Minnesota, were used as an outgroup in this study. Three major HUC-8 watersheds defined by the Minnesota Department of Natural Resources were represented in this collection including seven populations from the Upper Mississippi River (UMR) watershed, three populations from the Red River of the

North (RRN) watershed, and two populations from the St. Croix River (SCR) watershed (Figure 1; Table S1). A distance (km) matrix for all Natural Stand populations is available in Table S2. Our Temporal collection consisted of 200 samples collected from Garfield and Shell Lakes in 2010 as well as 2018 (50 samples/lake/collection date). GPS coordinates were used to ensure the same population at the same precise location were collected at both time points (Table S1). To avoid confusion, we would like to note that samples from Garfield and Shell Lakes collected in 2018 were also a part of our Natural Stand collection. Our Cultivated collection consisted of 209 individual open-pollinated samples, representing 3 cultivars and 15 breeding populations from the UMN NWR breeding program, which were collected at the UMN North Central Research and Outreach Center (NCROC) in Grand Rapids, MN in 2018 (Table S1; Figure S1).

DNA extraction and sequencing. During collection, 8 cm of leaf tissue was harvested on ice and then lyophilized using a TissueLyser II (Qiagen, Valencia, CA, USA). Genomic DNA was extracted using a Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. After checking DNA concentration and mass with a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA), samples were submitted to the UMN Genomics Center for library preparation and sequencing. The digestion step was performed using two restriction enzymes, *Btg1* (5'-C/CRYGG-3') and *TaqI* (5'-T/CGA-3'), initially evaluated in NWR in Shao et al. (2019). After digestion, unique barcodes were ligated to DNA fragments for sample identification and pooling. Single-end 150-bp sequencing to a depth of 2.5 million reads per sample was performed on an Illumina NovaSeq machine (Illumina, San Diego, CA, USA). Raw data were deposited in the National Center for Biotechnology

Information Short Read Archive (NCBI SRA) under accession number PRJNA774842. BioSample accession numbers for individual samples are provided in Table S3.

Read mapping. Quality control was initially performed on the FASTQ files using fastQC version 0.11.7 (Andrews, 2010) using default parameters to check for quality and adapter contamination. After adapter trimming with Cutadapt version 1.18 (Martin, 2011), reads were mapped to the reference genome v1.0 of NWR cultivar 'Itasca-C12' (Haas et al., 2021) using the default parameters of BWA-MEM version 0.7.13 (Li, 2013). Using default parameters, SNPs were called using the mpileup function from BCFtools version 2.3 (Li et al., 2011) and sorted with the sort function from samtools version 1.9 (Li et al., 2009), resulting in 2,183 variant call format (VCF) files, one for each scaffold of the NWR genome. The 17 largest VCF files representing the 15 NWR chromosomes and 2 additional large scaffolds of the NWR reference genome were merged into a single VCF file with the concat function from BCFtools version 1.2 (Li et al., 2011). Filtering of the merged VCF file was carried out using VCFtools with default parameters (Danecek et al., 2011). A maximum missing rate of 20% across all samples and a minimum depth of 4 reads per variant site was used to obtain our final SNP set, which was deposited at the Data Repository for the University of Minnesota (DRUM) under Digital Object Identifier XXX.

Marker statistics. Summary statistics were calculated for each SNP including their distribution across chromosomes/scaffolds as well as genic regions, their polymorphism information content (PIC) value, and their transition/transversion (TsTv) ratio. Polymorphism information content (PIC) was calculated using the *snpReady* R package (Granato et al., 2018). The number and type of transitions and transversions were calculated using VCFtools.

Genetic Diversity Assessments. To assess the structure and distribution of genetic variation within our collections, we conducted principal component (PC) analysis (PCA). To generate the PCA, the VCF file resulting from our SNP calling pipeline was imported into PLINK version 1.90b6.10 (Purcell et al., 2007) and the PCA function was used to calculate eigenvectors and eigenvalues. These data were imported into the R statistical environment (R Core Team, 2013) for further analysis and visualization. Neighbor-joining trees were created using the R packages *adegenet* (Jombart, 2008; Jombart and Ahmed, 2011), *ape* (Paradis and Schliep, 2019), *poppr* (Kamvar et al., 2014; Kamvar et al., 2015), and *vcfR* (Knaus et al., 2017). We used the `aboot()` function (bootstrapped dendrograms) from *poppr* with the unweighted pair group method with arithmetic mean (UPGMA) algorithm with Prevosti's genetic distance ("bitwise.dst"), 100 bootstrap replicates, and a cutoff value of 50.

The population structure of our Natural Stand and Cultivated collections was determined using Bayesian clustering implemented in STRUCTURE version 2.3.4 (Pritchard et al., 2000). Genotypic data for these individuals and our final bi-allelic SNP set were loaded into STRUCTURE and analyzed with the admixture model. The markov chain monte carlo was run for K=2 to K=6 and subsequently from K=10 to K=15 with a burn-in length of 1,000 followed by 10,000 iterations. Lower K values were used to look for larger structural patterns while the higher K values were chosen to test if we were able to separate each population (K=14=12 Natural Stands + 1 Cultivated + 1 *Z. aquatica*) into their own STRUCTURE-assigned cluster. We were particularly interested in looking for admixture between specific populations using this level of resolution. Each K value was run 10 times. The ideal number of clusters was determined using the DeltaK statistic from the Structure Harvester web tool version 0.6.94 (Earl and

vonHoldt, 2012), which uses the Evanno et al. (2005) method for determining the number of clusters.

Analysis of Molecular Variance (AMOVA) was performed using the `poppr.amova()` function from the *poppr* R package (Kamvar et al., 2014; Kamvar et al., 2015) based on the work of Excoffier et al. (1992). We defined groups in a hierarchy based on their collection membership (e.g., lake/river of origin or cultivar/breeding line identity), species (*Z. palustris* or *Z. aquatica*), and, more broadly, their germplasm type (e.g., Natural Stand or Cultivated). We also conducted an AMOVA based on STRUCTURE-assigned population membership. For each AMOVA, we used the "farthest neighbor" algorithm and the "quasieuclid" (default) method to correct for non-Euclidean distance implemented in the *ade4* R package using default parameters (Dray and Dufour, 2007). The significance was determined using the `randtest()` function with 999 repetitions. In order to test for correlation between geographic and genetic distances in our Natural Stand collection, we conducted a Mantel test using the `mantel.rtest()` function from the *ade4* R package (Paradis and Schliep, 2019). We used the same individual genetic distances that were used to create the UPGMA trees, averaged by population membership. Geographic distances were calculated using the `distm()` function from the *geosphere* R package (Hijmans et al., 2011) using default parameters. The regression equation was found by fitting the genetic and geographic distances to a linear model using the following code in R: `lm(genet ~ geo)`.

Gene flow. Pairwise estimations of genetic differentiation (F_{ST}) between different subgroups based on geographic origin, germplasm type (cultivated population, natural stand, species), and STRUCTURE-defined population subgroups, according to the method of Weir and Cockerham

(1984), were calculated in VCFtools (Danecek et al., 2011). We also used Dsuite version 0.4 (Malinsky et al., 2021) to calculate Patterson's *D*-statistics, also known as the ABBA-BABA statistic, in order to test for introgressions between groups. Since the analysis requires four groups, we defined the groups based on their membership in STRUCTURE-assigned groups when *K*=4. This approach split the Natural Stands into two groups (described in the Results section as Natural Stand I and Natural Stand II) while the Cultivated collection formed a third group. The fourth and final group (the outgroup) was *Z. aquatica*.

Linkage Disequilibrium. To calculate LD, we used the LD.decay() function from the *sommer* R package (Covarrubias-Pazaran, 2016) using default parameters (gamma = 0.95). We analyzed Natural Stand and Cultivated collections separately. Since the LD.decay() function cannot handle missing data, BEAGLE version 5.1 (Browning and Browning, 2007, Browning et al., 2018) was used to impute missing genotype data for this analysis. The results (r^2 and pairwise-distance) were plotted in R. We calculated LD for both individual chromosomes as well as the genome as a whole. In order to convert the *r* measure of LD into the genomic distance at which LD has effectively decayed, we used a known value of y (r^2) = 0.20 (Rafalski et al., 2002; Vos et al., 2017) to solve for the value for x (genomic position). Using the genome-wide LD results (*r2*, D, and the fitted loess curve results), we selected all rows where *r* was between 0.20 and 0.21 and calculated basic statistics (e.g., mean, median) for the data points that fit our criteria (~100 points for each).

Genome-wide Scans for Signatures of Selection. To begin to identify putative selection events in cultivated NWR, we tested for deviations from the model of neutrality in a variety of ways. First,

we calculated site frequency spectra (SFS) for cultivated NWR and the Natural Stands (excluding *Z. aquatica*) across the NWR genome by calculating the frequency of each allele (reference or alternate) for every SNP and plotting the distribution of the derived (alternate) allele frequencies with bin sizes of 0.02 (2%). This was done separately for Cultivated and Natural Stands. We calculated nucleotide diversity (π; Nei, 1987) and Tajima's D (Tajima, 1989) for each SNP using VCFtools (Danecek et al., 2011) by carrying out calculations separately for Natural Stand and Cultivated collections. The results were plotted in the R statistical environment (R Core Team, 2013) by aggregating π values into 1-Mb bins using the mean π value for each bin. In order to explore putative selection events in greater detail, we used the Cross-Population Composite Likelihood Ratio (XP-CLR) test (Chen et al., 2010) to support or refute the potential for the regions of low nucleotide diversity to be caused by selection events.

Results

Genotyping-by-sequencing. Using GBS technology, 739 samples representing our diversity collection along with 50 samples of *Z. aquatic*a, which served as an outgroup, were sequenced to a depth of 2.5 million reads per sample. Samples were categorized into three collections including a Natural Stand collection, which consisted of 530 *Z. palustris* samples collected from Minnesota and Wisconsin; a Cultivated collection, which consisted of 209 samples from the UMN NWR breeding program; and a Temporal collection, which consisted of 200 samples collected from two Minnesota lakes in 2010 and 2018. A total of 1,833,504,458 reads were generated with an average of 2,100,455 reads per sample. From these data, a total of 5,955 SNP markers were identified that met our filtering criteria, 3,005 (50.1%), of which, reside in genes. Basic marker statistics are shown in Table 1. Using 1-Mb bins, the SNP density ranged from

1.15 - 6.37 SNPs/Mb per chromosome with a genome-wide average of 4.85 SNPs/Mb. In genic regions, the SNP density ranged from 10-20 SNPs/Mb per chromosome. The genome-wide TsTv ratio was 3.15 with a minimum of 1.50 (ZPchr0458) and a maximum of 4.18 (ZPchr0016) (Table S4). The average PIC value was 0.217 (range: 0.030-0.313) (Table S5).

Genetic Diversity of NWR Collections. To visualize the variation within our diversity collection, we first performed principal component analysis on the Natural Stand and Cultivated collections together (Figure 2a). The first PC explained 23% of the variance, with the cultivated collection at one extreme and the Shell Lake population at the other. The Natural Stand collection exhibited a larger range of continuous variation, with samples mostly clustering according to their lake or river of origin. *Zizania aquatica* fell directly in the center of both PCs. When analyzed separately from the Cultivated collection, the Natural Stand collection formed four distinct clusters with the first two PCs explaining 19% and 13% of the variance, respectively (Figure 2b). The largest cluster, Natural Stand I (NS I), consisted of six populations including Bass Lake, Dahler Lake, and Decker Lake from the UMR watershed; Mud Hen Lake and Phantom Lake from the SCR watershed; and Upper Rice Lake from the RRN. The second largest cluster, Natural Stand II (NS II), consisted of Clearwater River and Ottertail River from the RRN, and Lake Plantagenet and Shell Lake from the UMR. The third cluster, Natural Stand III (NS III), consisted of Garfield Lake and Necktie River from the UMR. The fourth cluster, Natural Stand IV (NS IV), consisted of the Z. *aquatica* outgroup, which previously clustered with Upper Rice Lake samples when Natural Stand and Cultivated collections were analyzed together (Figure 2a). It is distinguished by the second PC. In comparison to the Natural Stands collection, there was far less variation and structure evident among the 3 cultivars and 15 breeding lines of the Cultivated collection (Figure

2a and c). When the Cultivated collection was analyzed alone, the first and second PCs explained 12% and 7% of the variation, respectively (Figure 2c). The most distinct cluster within the Cultivated collection consisted of 'K2' samples. While the beginnings of other clusters (i.e., 'FY-C20', Itasca-C12, 'PM3E') can be seen, there was an overall lack of population structure among Cultivated materials.

The UPGMA clustering analysis supported some of the clustering observed in the PCA (Figure S2). In particular, monophyletic clades, which could be defined for many of the sampling locations, included the majority of the samples from those populations. Notably, all samples from Clearwater River clustered, as did 92% of the samples from Lake Plantagenet and 86% of the samples from Dahler Lake. Over 90% of NSIII clustered in a single clade. Similarly, 88% of the samples from Shell Lake and 82% of the samples from Ottertail River identified as NSII clustered together. However, NSIII clustered most closely with Lake Plantagenet and Clearwater formed a separate clade. NSI is spread throughout the tree with Decker Lake and Bass Lake each forming two distinct unrelated clades. Upper Rice Lake, Phantom Lake, and Mud Hen Lake were all also distributed through the tree. Consistent with the PCA, there was minimal structure within the breeding program. However, unlike the PCA samples from Shell Lake, Ottertail River, Phantom Lake, Decker Lake, Necktie River, Dahler Lake, Bass Lake, Plantagenet Lake and Upper Rice Like as well as a single sample of *Z. aquatica* fell within the clade containing 82% of the breeding germplasm. A second breeding germplasm cluster at the base of the tree consists entirely of a subset of K2EF-C16 samples (10/20). Perhaps most surprising, *Z. aquatica* did not form a basal clade. Although the vast majority of the *Z. aquatica* samples formed a single clade, six breeding lines, three samples each from Shell Lake and Phantom Lake, two samples each

from Garfield Lake and Dahler Lake, and a sample from Mud Hen Lake and Upper Rice Lake, were basal to the *Z. aquatica* cluster.

STRUCTURE analysis (Figure 3) recapitulated the results observed in the PCA (Figure 2a) and UPGMA plots (Figures S2). Assuming a *K* value of 2, we observed one cluster consisting of NS II and another consisting of Cultivated Material and NS I (Figure 3a). As expected, the results from the DeltaK plot (Figure S3) of the Evanno analysis (Evanno, 2005) confirm that a *K*=2 is not appropriate for our collections. At *K*=3, we observed two clusters representing NS I and NS II, and the third representing the Cultivated material (Figure 3b). When higher values of *K* were considered, additional clusters with biological meaning were observed. At *K*=4, we observed NS I, NS II, Cultivated material, and the addition of a *Z. aquatica* cluster (NS IV) (Figure 3c) along with a few minor exceptions. For example, while Phantom Lake consistently clustered with NS I at lower *K* values, some samples at *K*=4 showed admixture with the Cultivated group and the *Z. aquatica* group (NS IV). Additionally, Mud Hen Lake clustered with NS IV at *K*=4 but previously clustered with NS I populations at lower *K* values. At *K*=5, we observed the four groups represented at *K*=4 (Cultivated material, NS I, NS II, and NS IV) with the exception that we observed a new NS III cluster (Garfield Lake and Necktie River), which previously clustered with the NS II group at lower *K* values (Figure 3d). Beyond *K*=5, the value of DeltaK began to increase (Figure S3), so there was no benefit to considering higher *K* values. These results suggest that a *K* value between 3 and 5 is appropriate. We selected *K*=5 as the most likely value as it minimized DeltaK (Figure S4).

We also evaluated *K*=14, which represented the 12 individual Natural Stand populations, the Cultivated materials, and *Z. aquatica* to further evaluate admixture among samples (*data not shown*). The vast majority of the collections showed limited admixture (<1%) between different

populations with the exceptions of Upper Rice Lake and Phantom Lake. Upper Rice Lake showed heavy admixture with Decker Lake, Dahler Lake, Bass Lake, Ottertail River, and Shell Lake populations. Phantom Lake showed an average of 21.43% admixture with the Cultivated materials.

Analysis of Molecular Variance. Analysis of Molecular Variance (AMOVA) was conducted between several different groupings including: 1) Natural Stands vs. Natural Stands; 2) Cultivated vs. Cultivated; 3) Natural Stands vs. Cultivated; 4) STRUCTURE-defined populations (K=5); and 5) *Z. palustris* vs. *Z. aquatica*. The AMOVA results revealed more variation within rather than between groups (Table 2). All of the 5 comparative groupings listed above identified 3.37%-8.10% of the variation could be attributed to differences among individuals in the groupings rather than differences within individuals (91.90%-96.63% attributed variation). The highest and lowest variation among individuals was identified in the Natural Stand vs Natural Stand analysis (8.10%) and the Cultivated vs Cultivated analysis (3.37%), respectively (Table 2).

Gene Flow. Genetic differentiation (F_{ST}) results are shown in Table 3 and include comparisons of our Natural Stand collection (Table 3a), our Cultivated collection (Table 3b), and a comparison of the two (Tables 3c). Overall, genetic differentiation between our collection of Natural Stand populations, including *Z. aquatica*, was 0.0913 (0.0886 without *Z. aquatica*). The lowest and highest F_{ST} values were between Necktie River and Garfield Lake (F_{ST} = 0.0387; 8.9 km between the two populations), and Shell Lake and Mud Hen Lake (F_{ST} = 0.1578; 265.9 km between the two populations), respectively (Table 3a). Mud Hen Lake was highly diverged from all other Natural Stand pairwise comparisons (range: F_{ST}=0.0861-0.1578). *Z. aquatica* pairwise

comparisons with the Natural Stand populations ranged from 0.0734 - 0.1291. In contrast, F_{ST} values between Cultivated populations were much lower (average: 0.0162; range: 0.0058 – 0.0337; Table 3b). When comparing Natural Stands to Cultivated material, the average F_{ST} value was 0.0488. The lowest and highest F_{ST} values were between PM3E and Upper Rice Lake (F_{ST} = 0.0240), and Itasca-C20 and Mud Hen Lake (F_{ST} = 0.0803), respectively (Table 3c).

To further examine the basis for the observed genetic differentiation, we used a Mantel test to look for correlation between genetic and geographic distances of the Natural Stand collection. A positive correlation between the two was observed (r^2 = 0.3981) and the best fit line yielded a regression equation of y = 78.22 + 0.0086x (Figure S5). Permutation tests showed that the observed results were unlikely to have occurred by chance (Figure S6). We were also interested in whether we could detect any admixture between our Natural Stand and Cultivated collections. Using *D*-statistics, we analyzed three groups including Cultivated Material, NS I, NS II, along with *Z. aquatica*, which served as an outgroup, and found that there was no significant difference between the groups (Z=1.66, *p*=0.098; Table S6).

The Temporal Collection. We were also interested in how genetic diversity changes over time in wild populations of NWR. Using two populations, Garfield Lake and Shell Lake, collected in 2010 and 2018, we identified distinct clustering patterns between the two lakes (Figure 2d). We also observed changes in Shell Lake samples collected over time with samples collected in 2010 showing a wider range of variation than those collected in 2018. In contrast, Garfield Lake samples did not show a considerable change in genetic variation over time.

Linkage Disequilibrium. We calculated the rate of LD decay genome-wide (Figure 4) as well as on a per chromosome/scaffold basis (Figure S7) for both Natural Stand and Cultivated collections. For both collections, the LD decay curves on ZPchr0002, ZPchr0009, ZPchr00010, ZPchr00011, and ZPchr00013 chromosomes did not fit the standard LD curves for this type of analysis (Figure S7). On these chromosomes, LD initially decreased but then rose again with increasing distance. For the Natural Stands, the genome-wide rate of LD decay was 0.018 and ranged from 0.013 (ZPchr0007) to 0.0226 (ZPchr0008) for all seventeen chromosomes and scaffolds evaluated. The median distance at which LD decayed to r^2=0.2 was 28,389 bp (range: 1-468,619 bp). For the Cultivated collection, the genome-wide rate of LD decay was 0.034 and ranged from 0.031 (ZPchr0017) to 0.036 (ZPchr0008) for all seventeen chromosomes and scaffolds evaluated. The median distance at which LD decayed to r^2=0.2 was 63,452 bp (range: 1-625,024 bp).

Scanning for Signatures of Selection. To identify genomic regions that have putatively undergone selection in cultivated NWR, we plotted the site frequency spectrum (SFS) for both Natural Stand and Cultivated collections (Figure S7). The SFS plots for both collections feature a peak of alternate (derived) allele frequencies at values beneath 0.10. The peak gradually declines in the Cultivated material whereas the drop is more pronounced in the Natural Stands. In both collections, there is a slight increase in derived allele frequency between 0.4 and 0.5 before declining to a lower, but evenly distributed, distribution of derived allele frequencies at values above 0.5. Both SFS Natural Stand and Cultivated histograms were similar and neither showed a fixation peak of derived alleles associated with evidence of selection or domestication. However, while the Natural Stand SFS fits the expected pattern of the neutrality model, the Cultivated SFS

appeared flatter with some higher peaks in the middle, which may be indicative of balancing selection.

We also calculated genome-wide statistics for nucleotide diversity (π), F_{ST}, and XP-CLR tests (Figure 5). Nucleotide diversity (π), F_{ST}, and XP-CLR scores were evaluated genome-wide in 1-Mb increments for a total of 1,223 bins. A total of 1,012 (83%) bins contained at least one SNP. While we observed considerable similarity between π values of the Natural Stand and Cultivated collections across the genome (Figure 5a), we also observed 25 bins with negative Tajima D values in the Cultivated collection (Table S7). The most negative Tajima's D was -0.192 at ~50 Mb on ZPchr0009. Genome-wide scans of F_{ST} values calculated between Natural Stand and Cultivated collection identified 166 individual pairwise comparisons above 0.10; 63 above 0.20; 24 above 0.30; 9 above 0.40; and 3 above 0.50 (Figure 5b). Genomic positions for these hits are listed in Table S7. The 3 highest F_{ST} values were identified on ZPchr0007, ZPchr0009, and ZPchr0013. Finally, we identified 9 regions of the genome with XP-CLR scores above 40 and 5 regions with scores above 60 (Figure 5c). Genomic positions for these hits are listed in Table S7. The 5 regions with XP-CLR scores above 60 were found on ZPchr0005, ZPchr0006, ZPchr0011, and ZPchr0013. Allele frequencies of SNPs in these regions were all fixed in the Cultivated materials and ~0.50/0.50 in the Natural Stand collection. Interestingly, we did not observe the same high XP-CLR scores for regions on ZPchr0009 that we observed in our nucleotide diversity and F_{ST} analyses.

To identify candidate genes that are under selection pressure in cultivated NWR, we looked into genomic regions of interest with high F_{ST} (>0.40) and/or XP-CLR scores (>60), which were located on ZPchr0004, ZPchr0005, ZPchr0006, ZPchr0007, ZPchr0009, ZPchr0011, ZPchr0013, and ZPchr0015 (Table S7). For many of these regions, we found genes without

annotation, which hampered our ability to conduct enrichment analyses. However, numerous genes known to confer abiotic stress tolerance, disease resistance, nitrogen use efficiency, and male sterility in other crops fell within these regions. Eleven genes specifically related to drought stress and abscisic acid (ABA) signaling in other species included the following: a calmodulin-binding receptor-like cytoplasmic kinase 3 (Yang et al., 2010; Sun et al., 2021) on ZPchr0004; a calcium-dependent protein kinase family protein (Campo et al., 2014; Wei et al., 2014) on ZPchr0005; two copies of coatomer subunit beta'-1 isoform X1 (Mirzaei et al., 2014), a zinc finger CCCH domain-containing protein 43 (Peng et al., 2012), and a receptor-like protein kinase At5g2005 (Marshall et al., 2012) on ZPchr0006; a CBL-interacting protein kinase 32 (Tai et al., 2016) on ZPchr0011; two copies of ras-related protein RABC2a (Mérida-Garcia et al., 2020) and E3 ubiquitin-protein ligase RZFP34 isoform X2 (Ding et al., 2015; Shu and Yang, 2017) on ZPchr0013; methyltransferase-like protein 13 isoform X2 (Gupta et al., 2020) on ZPchr0015. We also identified five genes in these regions known to confer disease resistance in other species including a LRR receptor-like serine/threonine-protein kinase At1g06840 isoform X1 (Kumar et al., 2018; Guo et al., 2022), a respiratory burst oxidase homolog protein H (Sagi et al., 2004; Wang et al., 2018), a hydrolase, NUDIX family protein (Bartsch et al., 2006; Ge and Xia, 2008), a 60S ribosomal protein L3 (Di et al., 2010; Kant et al., 2012), and the barley B recombinant-like protein A (Ameen et al., 2020).

For nitrogen use efficiency-related genes, we identified a urease accessory protein G on ZPchr0005 (Witte, 2011) and five copies of an early nodulin-93 gene on ZPchr0006 (Bi et al., 2009). Two genes, pentatricopeptide repeat-containing protein At1g11900 isoform X1 (Bentolila et al., 2002) and cytochrome c oxidase subunit 5b (Isaac et al, 1985), were identified that are associated with male sterility.

Discussion

NWR is a species with ecological, cultural, and agricultural significance. For these reasons, NWR merits heightened attention for conservation of its natural habitat. In this study, we used GBS to genotype a diverse collection of NWR individuals representing wild populations from across its native range and time as well as cultivated materials from the UMN NWR breeding program. Compared to previous studies based on isozymes (Lu et al., 2005), AFLPs (Diller et al., 2018), and SSRs (Kahler, 2010, Kahler et al., 2014), we generated orders of magnitude more markers. Our expectation was that we would be able to discern fine-scale population structure and reaffirm previous findings due to the scale of our genotyping data. Additionally, knowledge of gene flow from cultivated NWR to natural stands and vice versa is particularly important, especially as the cultivated industry and breeding programs strive to be better stewards of their domesticated plants.

Assessment of Genetic Diversity of Northern Wild Rice via Genotyping-By-Sequencing. Cost-effective sequencing technologies that are capable of generating robust sets of genome-wide molecular markers, such as GBS, are providing researchers, especially those working with complex plant genomes and limited public resources, an avenue for rapid variant detection. The large, informative datasets that can be generated via GBS have been used for a wide array of studies in plant species including genetic diversity assessments (Schreiber et al., 2018, Serba et al., 2019), quantitative trait loci (QTL) mapping (Sonah et al., 2015; Pereira et al., 2018), and genome-wide association mapping (Sonah et al., 2014). These studies can inform a variety of plant breeding and conservation processes including parental selection and targeting of

populations of interest, respectively. In this study, we provide a large genome-wide SNP dataset, aligned to the NWR reference genome, for future use and as evidence that GBS is a useful platform for genetic diversity assessments in NWR. However, we want to emphasize that this GBS approach, which is based on *Btg1* and *TaqI* restriction enzymes, may have introduced a bias in allele frequencies due to polymorphisms in restriction sites (Arnold et al., 2013) and therefore, may have slightly skewed the inferences we made regarding the population genetics of NWR.

Gene Flow in Wild Populations of Northern Wild Rice is Tied to Geography and a Complex History of Ecosystem Management. In natural populations, gene flow can be regarded as either a constraining or creative force in the evolution of a species, which determines the extent to which genetic changes in individual populations are independent from outside sources (Slatkin, 1985; Slatkin, 1987). Gene flow can prevent populations from adapting to local environmental conditions as new alleles adapted to other conditions are introduced but conversely, can also introduce superior alleles that increase a population's fitness. While the role that gene flow plays is largely species-dependent, the geographic distribution of a species (fragmented vs. continuous) is particularly important (Lenormand, 2002). For example, habitat fragmentation can lead to smaller populations and declines in genetic diversity that can incur fitness costs (Ewers and Didham, 2006; Keyghobadi, 2007). As natural ecosystem landscapes continue to be fragmented and threatened by human activity, evaluating the extent of gene flow across natural populations of a species is critical to understand how a species may be impacted by extensive environmental change.

Previous genetic analysis of NWR wild populations have indicated limited gene flow between populations and a lack of data to support a correlation between genetic variation and

geographic distance (Lu et al., 2005; Kahler et al., 2014; Xu et al., 2015). In this study, we observed wild NWR samples generally clustering according to their lake or river of origin in PCA, UPGMA, and STRUCTURE analyses, indicating that they are largely genetically distinct from one another. However, we also observed a continuous range of variation between some populations and larger clustering patterns that can be explained by watershed membership (Figure 1, Figure 2b). These results along with those from the Mantel test (Figure S4), which identified a positive correlation between genetic and geographic distance (r^2 = 0.3918), imply that while gene flow may be limited between distant drainages, it does in fact occur across populations in close proximity to one another. This is especially evident in the genetic similarity between Garfield Lake and Necktie River populations (Figure 2b; Table 3a), which are close to one another geographically (8.9 km). For future studies looking into the relationships between watershed membership and gene flow, collection of samples in the Rainy River and Great Lakes/Laurentian watersheds will also be important to include as those watersheds were not represented in this study.

Historical knowledge of the management and development of lakes across NWR's natural range can lend insight into the population structure we identified in this study. Efforts to establish new locales of NWR and address declining population sizes has resulted in large reseeding efforts across the species natural range (Porter, 2019), which can complicate population genetic studies in NWR. For example, Upper Rice Lake in the RRN basin, known for large reseeding efforts (A. Kern, *personal knowledge*), clustered primarily with UMR populations in NS I, while also showing a continuous variation with other RRN basin populations (NS II) (Figure 2b; Figure 3). Upper Rice Lake also showed heavy admixture with a number of lakes at *K*=14 in STRUCTURE analyses. Taken together, these results appear to

confirm historic reseeding efforts of Upper Rice Lake. Additionally, Phantom Lake of the SCR watershed also showed heavy admixture, not from other Natural Stand populations but with Cultivated materials. These results were surprising as Phantom Lake is one of the most geographically distant Natural Stand populations from cultivated NWR production and multiple Natural Stand populations much closer to cultivated production showed little to no admixture with Cultivated materials. However, Phantom Lake is part of the Crex Meadows State Wildlife Area and was artificially created in the 1950s when a series of dykes were installed. NWR restoration began in 1991 when 500 lbs (227 kg) of wild rice were seeded in this area over the course of three years (Thompson and Sorenson, 2000). We hypothesize that at least a portion of the seed utilized in these efforts came from cultivated NWR production, which further highlights the complexity of population genetic studies in NWR as well as the importance of documenting seed sources for reseeding efforts. We also suggest that future reseeding efforts should not use Phantom Lake populations as a seed source. While additional research with more balanced distribution of watershed membership is necessary to improve our understanding of the population structure of wild populations of NWR, contextualizing the data with historic management of NWR ecosystems, when possible, appears to be important in unraveling the history of NWR diversity.

The data we present here for 12 wild populations of NWR demonstrates that each population harbors unique genetic variation that is not found in other populations. This indicates that for conservation efforts, each individual population of NWR is incredibly important to consider as each harbor unique alleles and importantly, some may be less adaptive to environmental change. While we currently don’t understand whether gene flow between NWR populations can be regarded as a constraining or creative evolutionary force for change, it stands

to reason that improving our knowledge of the population structure and genetic relationships between wild populations could aid in decision-making for reseeding efforts.

Spatio-Temporal Genetic Diversity Analyses Can Aid in Conservation Efforts. Genetic variation determines the extent to which a species is able to survive and thrive across the variability of both the spatial and temporal conditions of a landscape. Species-specific characteristics, such as mating systems; population dynamics, such as genetic drift and natural selection; availability of key nutrients, such as light; and anthropogenic-induced factors, such as pollution, can all contribute to the extent and structure of natural populations across space and over time (Ruokolainen, 2013). Therefore, comprehensive monitoring of the spatio-temporal genetic diversity of a species could provide a better understanding of the evolutionary changes a species can undergo over time, and importantly, help to identify targets for conservations efforts. While many studies offer a snapshot of the present-day genetic diversity for a wide array of plant species, the majority of spatio-temporal diversity assessments have focused on large agricultural commodity crops (Le Clerc et al., 2006; Deu et al., 2010; Arya et al., 2019). Few studies have focused on natural populations (Ozbek et al., 2007; Fačkovcová et al., 2020) and many, only theoretically (Gray, 1996; González-Megías et al., 2011).

As a species of great conservation interest, we suggest that monitoring the spatio-temporal diversity of wild NWR populations would provide impactful data for resource managers and environmental agencies interested in the health and preservation of NWR populations across the species natural range. As a preview of what a more extensive study on the spatio-temporal genetic diversity of NWR could provide, we evaluated two wild populations from Shell Lake and Garfield Lake over time (8 years). Across time, we identified a reduction of

diversity in samples collected from Shell Lake in 2018 compared to those collected in 2010, while observing limited change in the Garfield Lake population across time (Figure 2d). This suggests that Shell Lake has experienced a loss of genetic diversity during the eight years between collection times. At the same time, the 2010 Garfield Lake samples were already less diverse than the 2010 Shell Lake samples. Of course, a wide array of factors could have contributed to this reduction in diversity including the 3-5 year boom and bust cycles of NWR (Waheed, 2021) or favorable environmental conditions for specific genotypes in the population's seed bank. Anecdotally, however, Shell Lake harbors multiple campgrounds and resorts that could represent significant shoreline development and recreational activity, which are both known to negatively impact NWR populations (Hansen, 2008). Conversely, Garfield Lake has seen no to little development in the 8 years between collection times. Clearly, a deeper understanding of all factors affecting genetic diversity is needed to aid in the conservation of NWR and research featuring additional lakes and more frequent sampling time points is needed to answer these questions. New research in remote sensing and landscape modeling techniques could also help to identify specific drivers of change in wild populations of NWR (O'Shea et al., 2021; Cavender-Bares et al., 2022).

Cultivated Northern Wild Rice is Genetically Distinct from Wild Populations. Concerns regarding gene flow between domesticated crops and their wild counterparts often include an increase in weediness and the potential effects on the genetic diversity and identity of wild populations and compatible species relatives (Gepts and Papa, 2003). For these reasons, the extent of gene flow between crops and their wild cohorts has been evaluated in numerous species and found to be dependent on a variety of factors including, but not limited to, a species mating

system (e.g. out-crossing vs selfing), the type and frequency of pollination (e.g. insect vs wind), the selective (dis)advantage of particular domestication traits (e.g. seed shattering resistance reducing seed dispersal), genetic drift, and genotype × environment interactions (Sonnante et al., 1994; Evans and Kermicle, 2001; Gepts, 2002; Gepts and Papa, 2003). Some studies, such as those in soybean (*Glycine max*), have identified limited gene flow and monophyletic domesticated clades separate from their wild counterparts (Li et al., 2010; Jeong et al., 2019). Other studies have identified significant historical gene flow during domestication, such as Emmer wheat (*Triticum dicoccon*; Luo et al., 2007) as well as on-going gene flow between crop-weed complexes, such as those in cowpea (*Vigna unguiculata* (L.) Walp; Coulibaly et al., 2002), pearl millet (*Pennisetum glaucum*; Mariac et al., 2006), and species in the *Sorghum* genus (Arriola et al., 1996; Sagnard et al., 2011). Cultivated NWR is commercially produced within both the centers of origin and diversity of *Z. palustris*. Given the out-crossing nature of NWR, a solid understanding of gene flow between cultivated and wild populations is important for both breeders and conservationists alike.

In this study, Cultivated and Natural Stand collections are genetically distinct from one another based on PCA (Figure 2a), UPGMA tree (Figure S2), and STRUCTURE analysis (Figure 3). These findings indicate that there is little to no gene flow between these two groups (with the exception of Phantom Lake), agreeing with previous diversity studies in NWR using different marker systems (Lu et al., 2005; Kahler et al., 2014; Diller et al., 2018). Limited gene flow in NWR is further supported by recent pollen-mediated gene flow studies that estimated the majority of NWR pollen only travels short distances (7 m) (Gietzel et al., 2021). Limited pollen travel has been reported for a close relative of NWR, *Zizania texana*, as well (Oxley et al., 2008). ABBA-BABA (*D*-statistics; Table S6) also support the lack of significant recent introgression

between the two largest Natural Stand groups (NS I and NS II) and the Cultivated collection, although there may not be enough time since the cultivation of NWR began (~65 years) to accurately evaluate introgressions between wild and cultivated NWR. However, many domestication traits, such as seed shattering resistance, heavily selected for in cultivated NWR, are recessively inherited (Kennard et al., 2002), further reducing the likelihood that such agronomic traits would become fixed in wild populations. These findings should alleviate some concerns that cultivated NWR is affecting the genetic identity of wild NWR populations.

Cultivated Northern Wild Rice Lacks Distinct Population Structure. Assessing the genetic diversity and population structure within and among available germplasm is an important piece of the plant breeding process to develop new and improved varieties with desirable traits. Only one such study has been conducted in cultivated NWR using eleven SSRs transferred from *Z. texana*, which revealed a lack of genetic differentiation between cultivated populations (Kahler, 2014). Our analyses in this study reaffirm these findings. While the lack of population structure among Cultivated material is in stark contrast to the structure observed among Natural Stand populations of NWR, this is likely due to the common ancestry of populations in the breeding program (Kahler, 2014) as well as high levels of pollen-mediated gene flow due to the close proximity in which cultivated populations are grown (Grombacher, 1993). A major exception is the clear separation of K2 from the rest of the Cultivated materials (Figure 2c). Special efforts have been made to keep K2 from cross-pollinating with other breeding populations through spatial separation in small, isolated paddies. These findings along with those from Gietzel et al. (2021), suggesting the majority of pollen-mediated gene flow occurs within 7 m from the pollen source, indicate that increasing spatial separation between Cultivated material would increase

genetic differentiation and maintain the purity of populations in NWR breeding programs. Buffer zones between populations grown in close proximity (within the same paddy) and growing small populations in stock tanks that are spatially separated could help this effort.

Linkage Disequilibrium Decays Quickly in Northern Wild Rice. Understanding LD, or the nonrandom association of alleles at different loci, and its rate of decay is important for numerous genetic analyses as the rate of LD decay determines the number and density of molecular markers as well as the experimental design required to detect marker-trait associations (Flint-Garcia et al., 2003). Several evolutionary forces contribute to and maintain LD in a species (Hartel and Clark, 1997; Rakshit et al., 2007), including a species mating system, wherein LD generally decays more rapidly in outcrossing species, like NWR, in comparison to selfing species (Nordborg, 2000). In outcrossing maize (*Zea mays*) and selfing white rice (*O. sativa*), for example, LD is estimated to decay at a rate of 1-10 kb (Yan et al., 2009) and 75-500 kb (Mather et al., 2007), respectively. Additionally, larger haplotype blocks in domesticated crops in comparison to their wild counterparts has been observed in several crop species, such as maize (Remington et al., 2001), soybean (Hyten et al., 2007), and white rice (Mather et al., 2007). We found similar patterns of a rise in LD in cultivated NWR (63.4 kb rate of LD decay) in comparison to wild NWR populations (28.4 kb rate of LD decay), which suggest the ongoing domestication of cultivated NWR is affecting LD patterns in the species.

We also identified atypical LD patterns on several chromosomes where LD decayed and then began to rise gradually as genomic distance increased (Figure 4; Figure S6). We suggest a likely explanation for this observation is the draft status of the NWR genome (Haas et al., 2021), where several chromosomes (ZPchr0002, ZPchr0006, ZPchr0009, ZPchr0010, ZPchr0011, and

ZPchr0013) exhibit high gene density in the centromeric regions, rather than their typical location in the telomeric regions of eukaryotic species (Sidhu and Gill, 2004). High BUSCO scores indicated that the gene space of NWR was well assembled in v1.0 of the NWR reference assembly (Haas et al., 2021). However, a low Long-Terminal Repeat (LTR) Assembly Index (LAI) score (Ou et al., 2018) indicated that the repetitive regions of the genome were only indicative of a draft assembly, which could have resulted in misassemblies of various scaffolds, coverage gaps, or difficulty collapsing alternative contigs in heterozygous regions (Haas et al., 2021). The unexpected gene density distribution and low LAI scores of the NWR reference assembly along with the unexpected LD patterns we identified in this study suggest that improvements to the assembly are needed. Optical maps have helped to assemble genomic regions difficult to capture by DNA sequencing in other species and could help to resolve the completeness of the NWR genome assembly (Aston et al., 1999; Yuan et al., 2020). Despite the concerns regarding LD decay on several chromosomes, our data still suggest LD decays quickly across the NWR genome and that dense marker sets are needed to conduct future genetic analyses in NWR.

Genome-wide Scans for Selection Signatures in Cultivated Northern Wild Rice Reveal the Crop is only Semi-Domesticated. Domestication is typically viewed as a process rather than a specific event (Zeder et al., 2006) and thus, different species exhibit varying levels of domestication. In cereals and other major agricultural crops, a common set of traits artificially selected for during their domestication processes include those associated with seed retention and size, seed dormancy and germination, plant growth habit, and plant size (Harlan, 1992). The comparative analysis of these common traits identified across multiple taxa is known as the domestication

syndrome, which differentiates domesticated species from their wild counterparts. While many of today's largest agricultural commodity crops have undergone mass selection for thousands of years, the advent of new technologies, such as genomic sequencing, provide today's plant breeders with new opportunities for the rapid, targeted domestication of new crops. Additionally, these technologies afford researchers the opportunity to study the domestication process in real-time.

To begin exploring the domestication process of cultivated NWR, we evaluated changes in

nucleotide diversity levels, allele frequency distributions, and LD using a wide array of genetic tools and tests. We found limited evidence for large selective sweeps in cultivated NWR. The pattern of allele frequencies in the SFS histogram of our Cultivated collection (Figure S7) suggest that while selection is affecting allele frequencies in cultivated NWR, it is not driving alleles to fixation. The process of domesticating NWR is a recent one (~65 years of selection) and it is possible there simply hasn't been enough time for the fixation of alleles to occur. The pattern in the SFS histogram also suggests the effectiveness of selection in NWR may be limited and that a heterozygous state is being maintained, creating what is in effect balancing selection. Given the spatial limitations of the paddies in which cultivated breeding germplasm is grown, it is likely that selection has been hindered by pollen travel between open-pollinated populations grown in very close proximity to one another.

Aside from large selective sweeps, we did identify some evidence for putative smaller selection events based on reduced nucleotide diversity values as well as the highest F_{ST} values and XP-CLR scores. A preliminary scan of genes in these regions show a preponderance of drought tolerance-related genes. Unlike wild populations of NWR, cultivated NWR is grown in

man-made irrigated paddies, which are drained during or shortly after flowering (principal phenological stage 6; Duquette and Kimball, 2020) to allow for mechanical harvesting of the grain. Therefore, during production, cultivated NWR experiences conditions similar to upland crops where standing water is not available during the development of fruit, ripening, and senescence. It is possible that drought tolerance genes were heavily selected for in cultivated NWR to allow for this drastic change in environmental conditions. Other genes associated with disease resistance and nitrogen use efficiency were also identified and have been selected for in breeding programs. The detection of male sterility genes in these regions of interest may also reflect the significant selection of a male panicle type called “bottlebrush”, which is linked to male sterility, for several varieties released in the 1990s (Cregan, 2004). Today, this panicle type is still present in numerous breeding populations of cultivated NWR. Interestingly, these regions were not associated with shattering resistance despite the notable differences in seed shattering between cultivated NWR and wild NWR populations and the heavy selection for seed retention in cultivated NWR. Despite these limits to our current study, we think these regions merit further investigation in future studies.

Domestication indices to account for varying levels of domestication have been purposed for several species and typically include the following: the extent of phenotypic differentiation between the domesticated species and its wild counterparts; the length of a species domestication history; if major genetic changes to the domestication species has been identified; if the species has been adapted to agricultural settings through targeted breeding efforts; and the extent of the species cultivation (Clement, 1999; Dempewolf et al., 2008; Hammer and Khoshbakht, 2015). Cultivated NWR is somewhat phenotypically distinct from wild NWR, mainly in its growth habit and seed retention characteristics, which have been made possible through a short history of

concentrated breeding efforts. The species has also been cultivated for a short time, but production has expanded to California, which is outside the species natural range. For these reasons, we suggest cultivated NWR be classified as semi-domesticated.

Conclusions.

Northern Wild Rice is an enigmatic species with ecological, cultural, and economic importance to the Great Lakes region of the United States and Canada. Our results suggest that wild and cultivated NWR are genetically distinct and that gene flow between the two groups of NWR is limited. Wild NWR populations are genetically distinct from each other, but on a larger scale, appear to be structured based on watershed membership. With the exception of K2, cultivated NWR has no such structure. Compared to other commercial crops, cultivated NWR has experienced a shorter period of domestication pressures. Therefore, it is unsurprising that there are no strong selection signals that are often found in major crops. Nevertheless, we found putative selection signals that may be associated with traits that are unique to cultivated NWR including drought tolerance and the bottlebrush panicle type.

Code and Data Availability. All data generated from this project have been deposited at the NCBI Sequence Read Archive under BioProject PRJNA774842. Other supporting data are available at the Data Repository for the University of Minnesota under the DOI (xxx). All code for the analysis described can be found at https://github.com/UMNKimballLab/genetic_diversity.

Acknowledgements. Our team would like to acknowledge the extensive and tireless efforts of Dakota and Anishinaabe First Peoples to protect and conserve this precious natural resource. The

authors would also like to thank Calvin Lawrence from the Wisconsin Department of Natural Resources for his assistance in obtaining and formatting the Wisconsin watershed data that we used to create Figure 1.

Conflict of Interest. The authors declare no conflict of interest.

References

Aagaard, K., **Eash, J., Ford, W., Heglund, P.J., McDowell, M. & Thogmartin, W.E.** (2018) Modeling the relationship between water level, wild rice abundance, and waterfowl abundance at a central North American wetland. *Wetlands* **39**, 149-160.

Achaz, G. (2009) Frequency spectrum neutrality tests: one for all and all for one. *Genetics* **183**, 249-258.

Ameen, G., Solanki, S., Drader, T., Sager-Bittara, L., Steffenson, B., Kleinhofs, A., Vogiatzis, C. & Brueggeman, R.S. (2020) rcs5-mediated spot blotch resistance in barley is conferred by wall-associated kinases that resist pathogen manipulation. *BioRxiv*, 2020.04.13.040238

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 25 May 2021.

Arnold, B., Corbett□Detig, R.B., Hartl, D. & Bomblies, K. (2013) RAD seq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular ecology* **22**, 3179-3190.

Arriola, P.E & Ellstrand, N.C. (1996). Crop□to□weed gene flow in the genus Sorghum (Poaceae): Spontaneous interspecific hybridization between johnsongrass, *Sorghum halepense*, and crop sorghum, *S. bicolor*. *American Journal of Botany* **83**, 1153-1159.

Arya, L., Verma, M., Singh, S.K. & Verma, R.P.S. (2019) Spatio-temporal genetic diversity in Indian barley (*Hordeum vulgare* L.) varieties based on SSR markers. *Indian Journal of Experimental Biology* **57**, 545-552.

Aston, C., Mishra, B. & Schwartz, D.C. (1999) Optical mapping and its potential for large-scale sequencing projects. *Trends in Biotechnology* **17**, 297-302.

Bartsch, M., Gobbato, E., Bednarek, P., Debey, S., Schultze, J.L., Bautor, J. & Parker, J.E. (2006) Salicylic acid–independent ENHANCED DISEASE SUSCEPTIBILITY1 signaling in Arabidopsis immunity and cell death is regulated by the monooxygenase FMO1 and the nudix hydrolase NUDT7. *The Plant Cell* **18**, 1038-1051.

Bentolila, S., Alfonso, A.A. & Hanson, M.R. (2002) A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10887-10892.

Bi, Y.M., Kant, S., Clark, J., Gidda, S., Ming, F., Xu, J., Rochon, A., Shelp, B.J., Hao, L., Zhao, R., Mullen, R.T., Zhu, T. & Rothstein, S.J. (2009) Increased nitrogen□use efficiency in transgenic rice plants over□expressing a nitrogen□responsive early nodulin gene identified from rice expression profiling. *Plant, cell & environment* **32**, 1749-1760.

Biesboer, D.D. (2019) The ecology and conservation of wild rice, *Zizania palustris* L., in North America. *Acta Limnologica Brasiliensia* **31**, e102.

Bernardo, R. (2016) Bandwagons I, too, have known. *Theoretical and Applied Genetics* **129**, 2323-2332.

Browning, S.R. & Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *American Journal of Human* Genetics **81**, 1084-1097.

Browning, B.L., Zhou, Y. & Browning, S.R. (2018) A one-penny imputed genome from next-generation reference panels. *American Journal of Human Genetics* **103**, 338-384.

Campo, S., Baldrich, P., Messeguer, J., Lalanne, E., Coca, M. & San Segundo, B. (2014) Overexpression of a calcium-dependent protein kinase confers salt and drought tolerance in rice by preventing membrane lipid peroxidation. *Plant Physiology* **165**, 688-704.

Cavender-Bares, J., Schneider, F D., Santos, M.J., Armstrong, A., Carnaval, A., Dahlin, K. M., Fatoyinbo, L., Hurtt, G.C., Schimel, D., Townsend, P.A., Ustin, S.L., Wang, Z. & Wilson, A.M. (2022). Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nature Ecology and Evolution*, 1-14.

Chambliss, C.E. (1940) The botany and history of *Zizania aquatica* L. ("wild rice"). *Journal of the Washington Academy of Sciences* **30**, 185-205.

Chen, H., Patterson, N. & Reich, D. (2010) Population differentiation as a test for selective sweeps. *Genome Research* **20**, 393-402.

Clement, C.R. (1999).1492 and the loss of Amazonian crop genetic resources. I. The relation between domestication and human population decline. *Economic Botany* **53**, 188-202.

Collard, B.C.Y., Jahufer, M.Z.Z., Brouwer, J.B. & Pang, E.C.K. (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142**, 169-196.

Coulibaly, S., Pasquet, R.S., Papa, R. & Gepts, P. (2002). AFLP analysis of the phenetic organization and genetic diversity of *Vigna unguiculata* L. Walp. reveals extensive gene flow between wild and domesticated types. *Theoretical and Applied Genetics* **104**, 358-366.

Covarrubias-Pazaran, G. (2016) Genome-assisted prediction of quantitative traits using the R package *sommer*. *PLoS ONE* **11**, e0156744.

Cregan, J.L. (2004) Aspects of seed storage, pollen travel, and population dynamics of wildrice (*Zizania palustris).* M.Sc. thesis, University of Minnesota

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. & Durbin, R. (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

Davies, M.S. (1993) Rapid evolution in plant populations. Evolutionary Patterns and Processes, Blackwell, Oxford. 171-188.

Dempewolf, H., Rieseberg, L.H. & Cronk, Q.C. (2008) Crop domestication in the Compositae: a family-wide trait assessment. *Genetic Resources and Crop Evolution* **55,** 1141–1157.

Desmarais, S. (2019) Returning the rice to the wild: revitalizing wild rice in the Great Lakes region through indigenous knowledge governance and establishing a geographical indication. *Lakehead Law Journal* **3**, 36.

Deu, M., Sagnard, F., Chantereau, J., Calatayud, C., Vigouroux, Y., Pham, J. L., Mariac, C., Kapran, I., Mamadou, A., Gérard, B., Ndjeunga, J. & Bezançon, G. (2010) Spatio-

temporal dynamics of genetic diversity in Sorghum bicolor in Niger. *Theoretical and Applied Genetics* **120**, 1301-1313.

De Wet, J.M. & Oelke, E.A. (1978) Domestication of American wild rice (*Zizania aquatica* L., Gramineae). *Journal d'agriculture traditionnelle et de botanique appliquée* **25**, 67-84.

Di, R., Blechl, A., Dill-Macky, R., Tortora, A. & Tumer, N.E. (2010) Expression of a truncated form of yeast ribosomal protein L3 in transgenic wheat improves resistance to Fusarium head blight. *Plant Science* **178**, 374-380.

Ding, S., Zhang, B. & Qin, F. (2015) Arabidopsis RZFP34/CHYR1, a ubiquitin E3 ligase, regulates stomatal movement and drought tolerance via SnRK2. 6-mediated phosphorylation. *The Plant Cell* **27**, 3228-3244.

Diller, S.N., McNaught, A.S., Swanson, B.J., Dannenhoffer, J.M. & Ogren, S. (2018) Genetic structure and morphometric variation among fragmented Michigan wild rice populations. *Wetlands* **38**, 793-805.

Dray, S. & Dufour, A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* **22**, 1-20.

Duquette, J. & Kimball, J.A. (2020) Phenological stages of cultivated northern wild rice according to the BBCH scale. *Annals of Applied Biology* **176**, 350-356.

Duvall, M. & Biesboer, D. (1988) Anatomical distinctions between the pistillate spikelets of the species of wild-rice (*Zizania*, Poaceae). *American Journal of Botany* **75**, 157-159.

Duvall, M., Peterson, P., Terrell, E. & Christensen, A. (1993) Phylogeny of North American oryzoid grasses as construed from maps of plastid DNA restriction sites. *American Journal of Botany* **80**, 83-88.

Earl, D.A. & vonHoldt, B.M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetic* Resources **4**, 359-361.

Eberhardt, R.T. & Riggs, M. (1995) Effects of sex and reproductive status on the diets of breeding ring-necked ducks (*Aythya collaris*) in north-central Minnesota. *Canadian Journal of Zoology* **73**, 392-399.

Ellegren, H. & Galtier, N. (2016) Determinants of genetic diversity. *Nature Reviews Genetics* **17**, 422-433.

Elliot, W.A. (1980) Chapter 52: Wild-rice. In: Fehr WR, Hadley HH (eds) Hybridization of crop plants. ASA-CSSA-SSSA Books. Madison, WI, pp 721-731.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. & Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, e19379.

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

Evans, M.M.S. & Kermicle, J.L. (2001) Teosinte crossing barrier1, a locus governing hybridization of teosinte with maize. *Theoretical and Applied Genetics* **103**, 259–265.

Ewers, R.M. & Didham, R.K. (2006). Confounding factors in the detection of species responses to habitat fragmentation. *Biological Reviews* **81**, 117–142.

Excoffier, L., Smouse, P.E. & Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491.

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**, 564-567.

Fačkovcová, Z., Slovák, M., Vďačný, P., Melichárková, A., Zozomová-Lihová, J. & Guttová, A. (2020) Spatio-temporal formation of the genetic diversity in the Mediterranean dwelling lichen during the Neogene and Quaternary epochs. *Molecular phylogenetics and evolution* **144**, 106704.

Fannucchi, W.A. (1983) Wildlife use of wild rice beds and the impact of rice harvesting on wildlife in east central Minnesota. M.Sc. thesis. University of Wisconsin-Stevens Point, Stevens Point, WI.

Fischer, M., Hock, M. & Paschke, M. (2003) Low genetic variation reduces cross-compatibility and offspring fitness in populations of a narrow endemic plant with a self-incompatibility system. *Conservation Genetics* **4**, 325-336.

Flint-Garcia, S.A., Thornsberry, J.M. & Buckler IV, E.S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357-374.

Fort, D.J., Mathis, M.B., Walker, R., Tuominen, L.K., Hansel, M., Hall, S., Richards, R., Grattan, S.R. & Anderson, K. (2014) Toxicity of sulfate and chloride to early life stages of wild rice (*Zizania palustris*). *Environmental Toxicology* **33**, 2802-2809.

Frankel, O.H. (1970) Sir William Macleay memorial lecture 1970: variation—the essence of life. Proc Linnean Soc NSW **95**, 158–169.

Frankham, R., Briscoe, D.A. & Ballou, J.D. (2002) Introduction to conservation genetics. Cambridge, UK: Cambridge University Press.

Garrison, E. & Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*

Ge, X. & Xia, Y. (2008) The role of AtNUDT7, a Nudix hydrolase, in the plant defense response. *Plant Signaling & Behavior* **3**, 119-120.

Gepts, P (2002) A comparison between crop domestication, classical plant breeding, and genetic engineering. *Crop Science* **42**, 1780–1790.

Gepts, P. & Papa, R. (2003) Possible effects of (trans)gene flow from crops on the genetic diversity from landraces and wild relatives. *Environmental Biosafety Research* **2**, 89-103.

Gietzel, C., Duquette, J., McGilp, L. & Kimball, J. (2021) Recessive male floret color for tracking gene flow in cultivated northern wild rice (*Zizania palustris* L.) *Crop Science* **62**, 157-166.

González-Megías, A., Gómez, J.M. & Sánchez-Piñero, F. (2011). Spatio-temporal change in the relationship between habitat heterogeneity and species diversity. *Acta Oecologica* **37**, 179-186.

Goudet, J. (2005) HIERFSTAT, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* **5**, 184-186.

Granato, I.S.C., Galli, G., de Oliveira Couto, E.G., e Souza, M.B., Mendonça, L.F. & Fritsche-Neto, R. (2018) *snpReady*: a tool to assist breeders in genomic analysis. *Molecular Breeding* **38**, 102.

Gray, A.J. (1996) Genetic diversity and its conservation in natural populations of plants. *Biodiversity* Letters **3**, 71-80.

Grombacher, A.W., Porter, R.A. & Everett, L.A. (1997) Breeding wild rice. *Plant Breeding* Reviews **14**, 237–265.

Gruber, B., Unmack, P.J., Berry, O.F. & Georges, A. (2017) DARTR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology* Resources **18**, 691-699.

Guo, D., Xia, Z., Jiang, X., Huang, H., Yang, Y., Jiang, S., Wang, D. & Chen, Z. (2022) Sequencing and Functional Annotation of Competing Endogenous RNAs and MicroRNAs in Tea Leaves During Infection by *Lasiodiplodia theobromae*. *PhytoFrontiers* **X**, xxx-xxx.

Gupta, S., Mishra, S.K., Misra, S., Pandey, V., Agrawal, L., Nautiyal, C.S. & Chauhan, P.S. (2020) Revealing the complexity of protein abundance in chickpea root under drought-stress using a comparative proteomics approach. *Plant Physiology and Biochemistry* **151**, 88-102.

Haas, M., Kono, T., Macchietto, M., Millas, R., McGilp, L., Shao, M., Duquette, J., Qiu, Y., Hirsch, C. & Kimball, J. (2021) Whole genome assembly and annotation of Northern Wild Rice, *Zizania palustris* L., supports a whole genome duplication in the *Zizania* genus. *The Plant Journal* **107**, 1802-1818.

Hammer, K., & Khoshbakht, K. (2015) A domestication assessment of the big five plant families. *Genetic Resources and Crop Evolution* **62**, 665-689.

Hansen, D. (2008) Natural wild rice in Minnesota. *A Wild Rice Study document submitted to the Minnesota Legislature by the Minnesota Department of Resources*.

Harlan, J.R. (1992) Crops and Man, 2nd ed. American Society of Agronomy and Crop Science Society of America. Madison, WI, USA.

Hartl, D.L. & Clark A.G. *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA

Hijmans, R.J., Williams, E. & Vennes, C. (2011) Package "geosphere". Available: http://cran.r-project.org/web/packages/geosphere. Accessed 2021 October 25

Hyten, D.L., Choi, I.Y., Song, Q., Shoemaker, R.C., Nelson, R.L., Costa, J.M., Specht, J.E. & Cregan, P.B. (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**, 1937-1944.

Isaac, P.G., Jones, V.P. & Leaver, C.J. (1985) The maize cytochrome c oxidase subunit I gene: sequence, expression and rearrangement in cytoplasmic male sterile plants. *The EMBO Journal* **4**, 1617-1623.

Jeong, S.C., Moon, J.K., Park, S.K., Kim, M.S., Lee, K., Lee, S.R., .Jeong, N., Choi, M.S., Kim, N., Kang, S.-T. & Park, E. (2019). Genetic diversity patterns and domestication origin of soybean. *Theoretical and Applied* Genetics **132**, 1179-1193.

Jombart, T. (2008) *adegenet*: an R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Jombart, T. & Ahmed, I. (2011) *adegenet 1.3-1*: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070-3071.

Kahler, A.L. (2010) Genome organization and genetic diversity of wildrice (*Zizania palustris* L.) Ph.D. thesis. University of Minnesota

Kahler, A.L., Kern, A.J., Porter, R.A. & Phillips, R.L. (2014) Chapter 9: Maintaining food value of wild rice (*Zizania palustris* L.) using comparative genomics. In: Tuberosa R et al. (eds) Genomics of Plant Genetic Resources. Springer Science+Business Media, Dordrecht, The Netherlands, pp 233-248

Kamvar, Z.N., Tabima, J.F. & Grünwald, N.J. (2014) Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281.

Kamvar, Z.N., Brooks, J.C. & Grünwald, N.J. (2015) Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics* **6**, 208.

Kant, P., Gulati, A., Harris, L., Gleddie, S., Singh, J. & Pauls, K.P. (2012) Transgenic corn plants with modified ribosomal protein L3 show decreased ear rot disease after inoculation with *Fusarium graminearum*. *Australian Journal of Crop Science* **6**, 1598-1605.

Keyghobadi, N. (2007). The genetic implications of habitat fragmentation for animals. *Canadian Journal of Zoology* **85**, 1049–1064.

Knaus, B.J. & Grünwald, N.J. (2017) VCFR: A package to manipulate and visualize variant call format in R. *Molecular Ecology Resources* **17**, 44-54.

Kumar, I.S., Amran, N.A. & Nadarajah, K. (2018) In silico analysis of qBFR4 and qLBL5 in conferring quantitative resistance against rice blast. *Journal of Pure Applied Microbiology* **12**, 1703-1718.

Le Clerc, V., Cadot, V., Canadas, M., Lallemand, J., Guerin, D. & Boulineau, F. (2006) Indicators to assess temporal genetic diversity in the French Catalogue: no losses for maize and peas. *Theoretical and Applied Genetics* **113**, 1197-1209.

Lenormand, T. (2002) Gene flow and the limits to natural selection. *Trends in Ecology and Evolution* **17**, 183-189.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Li, Y.-H., Li, W., Zhang, C., Yang, L., Chang, R.-Z., Gaut, B.S. & Qiu, L.-J. (2010), Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytologist* **188**, 242-253.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping, and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 1303-3997.

Lu, Y., Waller, D.M. & David, P. (2005) Genetic variability is correlated with population size and reproduction in American wild-rice (*Zizania palustris* var. *palustris*, Poaceae) populations. *American Journal of Botany* **92**, 990-997.

Luo, M.C., Yang, Z.L., You, F.M., Kawahara, T., Waines, J.G. & Dvorak, J. (2007). The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theoretical and Applied Genetics* **114**, 947-959.

Malinsky, M., Matschiner, M. & Svardal, H. (2021) Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* **21**, 584-595.

Manching, H., Sengupta, S., Hopper, K.R., Polson, S.W., Ji, Y. & Wisser, R.J. (2017) Phased genotyping-by-sequencing enhances analysis of genetic diversity and reveals divergent copy number variants in maize. *G3-Genes Genomes Genetics* **7**, 2161-2170,

Mariac, C., Robert, T., Allinne, C., Remigereau, M.S., Luxereau, A., Tidjani, M., Seyni, O., Bezancon, G., Pham, L. & Sarr, A. (2006). Genetic diversity and gene flow among pearl millet crop/weed complex: a case study. *Theoretical and Applied Genetics* **113**, 1003-1014.

Marshall, A., Aalen, R.B., Audenaert, D., Beeckman, T., Broadley, M.R., Butenko, M.A., Caño-Delgado, A.I., de Vires, S., Dresselhaus, T., Felix, G., Graham, N.S., Foulkes, J., Granier, C., Greb, T., Grossniklaus, U., Hammond, J.P., Heidstra, R., Hodgman, C., Hothorn, M., Inzé, D., Østergaard, L., Russinova, E., Simon, R., Skirycz, A., Stahl, Y., Zipfel, C. & De Smet, I. (2012) Tackling drought stress: receptor-like kinases present new approaches. *The Plant Cell* **24**, 2262-2278.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12.

Mather, K.A., Caicedo, A.L., Polato, N.R., Olsen, K.M., McCouch, S. & Purugganan, M.D. (2007). The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* **177**, 2223-2232.

Matson, L., Ng, G.-H.C., Dockry, M., Nyblade, M., King, H.J., Bellcourt, M., Bloomquist, J., Bunting, P., Chapman, E., Dalbotten, D., Davenport, M.A., Diver, K., Duquain, M., Graveen, W., Hagsten, K., Hedin, K., Howard, S., Howes, T., Johnson, J., Kesner, S., Kojola, E., LaBine, R., Larkin, D.J., Montano, M., Moore, S., Myrbo, A., Northbird, M., Porter, M., Robinson, R., Santelli, C.M., Schmitter, R., Shimek, R., Schuldt, N., Smart, A., Strong, D., Torgeson, J., Vogt, D. & Waheed, A. (2021) Transforming research and relationships through collaborative tribal-university partnerships on Manoomin (wild rice). *Environmental Science and Policy* **115**, 108-115.

McCouch, S.R., Baute, G.J., Bradeen, J., Bramel, P., Bretting, P.K., Buckler, E., Burke, J.M., Charest, D., Cloutier, S., Cole, G., Dempewolf, H., Dingkuhn, M., Feuillet, C., Gepts, P., Grattapaglia, D., Guarino, L., Jackson, S., Knapp, S., Langridge, P., Lawton-Rauh, A., Lijua, Q., Lusty, C., Michael, T., Myles, S., Naito, K., Nelson, R.L., Pontarollo, R., Richards, C.M., Rieseberg, L., Ross-Ibarra, J., Rounsley, S., Hamilton, R.S., Schurr, U., Stein, N., Tomooka, N., van der Knapp, E., van Tassel, D., Toll, J., Valls, J., Varshney, R.K., Ward, J., Waugh, R., Wenzl, P. & Zamir, D. (2013) Feeding the future. *Nature* **499**, 23-24.

Mérida-García, R., Gálvez, S., Paux, E., Dorado, G., Pascual, L., Giraldo, P. & Hernandez, P. (2020) High resolution melting and insertion site-based polymorphism markers for wheat variability analysis and candidate genes selection at drought and heat MQTL loci. *Agronomy* **10**, 1294.

Minnesota Department of Natural Resources. (2021). *Watersheds, MN DNR*. https://resources.gisdata.mn.gov/pub/gdrs/data/pub/us_mn_state_dnr/geos_dnr_watersheds/metadata/metadata.html5. Accessed 24 September 2021

Mirzaei, M., Soltani, N., Sarhadi, E., George, I.S., Neilson, K.A., Pascovici, D., Shahbazian, S., Haynes, P.A., Atwell, B.J. & Salekdeh, G.H. (2014) Manipulating root water supply elicits major shifts in the shoot proteome. *Journal of Proteome Research*, **13**, 517-526.

Moyle, J.B. (1944) Wild rice in Minnesota. *Journal of Wildlife Management* **8**, 177-184.

Myrbo, A., Swain, E.B., Engstrom, D.R., Wasik, J.C., Brenner, J., Shore, M.D., Peters, E.B. & Blaha, G. (2017) Sulfide generated by sulfate reduction is a primary controller of the occurrence of wild rice (*Zizania palustris*) in shallow aquatic ecosystems. *Journal of Geophysical Research: Biogeosciences* **122**, 2736-2753.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3321-3323.

Nordborg, M. (2000) Linkage disequilibrium, gene trees, and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923-929.

O'Shea, K., LaRoe, J., Vorster, A., Young, N., Evangelista, P., Mayer, T., Carver, D., Simonson, E., Martin, V., Radomski, P., Knopik, J., Kern, A. & Khoury, C. K. (2020) Improved remote sensing methods to detect northern wild rice (*Zizania palustris* L.). *Remote Sensing* **12**, 3023.

Oxley, F.M., Echlin, A., Power, P., Tolley-Jordan, L. & Alexander, M.L. (2008) Travel of pollen in experimental raceways in the endangered Texas wild rice (*Zizania texana*). *The Southwestern Naturalist* **53**, 169-174.

Ou, S., Chen, J. & Jiang, N. (2018) Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research* **46**, e126.

Ozbek, O., Millet, E., Anikster, Y., Arslan, O. & Feldman, M. (2007) Spatio-temporal genetic variation in populations of wild emmer wheat, *Triticum turgidum* ssp. *dicoccoides*, as revealed by AFLP analysis. *Theoretical and Applied Genetics* **115**, 19-26.

Paradis, E. & Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528.

Pearson, J., Ipsen, J., Sutherland, S., Wegerson, K. & Onello, E. (2019) Risks and costs to human health of sulfide-ore mining near the Boundary Waters Canoe Area Wilderness. *Human Ecology Risk Assessment: An International Journal* ***26***, 1329-1340.

Peng, X., Zhao, Y., Cao, J., Zhang, W., Jiang, H., Li, X., Ma, Q., Zhu, S. & Cheng, B. (2012) CCCH-type zinc finger family in maize: genome-wide identification, classification and expression profiling under abscisic acid and drought treatments. *PloS ONE* **7**, e40120.

Pereira, L., Ruggieri, V., Pérez, S., Alexiou, K. G., Fernández, M., Jahrmann, T., Pujol, M. & Garcia-Mas, J. (2018) QTL mapping of melon fruit quality traits using a high-density GBS-based genetic map. *BMC Plant Biology* **18**, 1-17.

Pillsbury, R.W. & McGuire, M.A. (2009) Factors affecting the distribution of wild rice (*Zizania palustris*) and the associated macrophyte community. *Wetlands* **29**, 724-734.

Pollman, C.D., Swain, E.B., Bael, D., Myrbo, A., Monson, P. & Shore, M.D. (2017) The evolution of sulfide in shallow aquatic ecosystem sediments: An analysis of the roles of sulfate, organic carbon, and iron and feedback constraints using structural equation modeling. *JGR Biogeosciences* **122**, 2719-2735.

Porter, R. (2019) Wildrice (*Zizania* L.) in Norther America: Genetic resources, conservation, and use. In: *North American Crop Wild Relatives: Important species*. Springer International Publishing, 83-97.

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

Probert, R.J. & Longley, P.L. (1989) Recalcitrant seed storage physiology in three aquatic grasses (*Zizania palustris*, *Spartina anglica*, and *Porteresia coarctata*) *Annals of Botany* **63**, 53064.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. & Sham, P.C. (2007) PLINK: A toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559-575.

R Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5**, 94-100.

Rakshit, S., Rakshit, A., Matsumura, H., Takahashi, Y., Hasegawa Y., Ito, A., Takashige, I., Miyashita, N.T. & Terauchi, R. (2007) Large-scale DNA polymorphism study of *Oryza sativa* and *O. rufipogon* reveals the origin and divergence of Asian rice. *Theoretical and Applied Genetics* **114**, 731-743.

Ramanatha Rao, V. & Hodgkin, T. (2002) Genetic diversity and conservation and utilization of plant genetic resources. *Plant Cell, Tissue and Organ Culture* **68**, 1-19.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M. & Buckler, E.S. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of The United States of America* **98**, 11479-11484.

Richards, C.M., Reilley, A., Touchell, D., Antolin, M.F. & Walters, C. (2004) Microsatellite primers for Texas wild rice (*Zizania texana*), and a preliminary test of the impact of cryogenic storage on allele frequency at these loci. *Conservation Genetics* **5**, 853-859.

Rogosin, A. (1954) An ecological history of wild rice. Minnesota Department of Conservation, Division of Game and Fish. Saint Paul, MN

Ruokolainen, L., Lindén, A., Kaitala, V. & Fowler, M.S. (2009) Ecological and evolutionary dynamics under coloured environmental variation. *Trends in Ecology and Evolution* **24**, 555-563.

Ruokolainen, L. (2013) Spatio-temporal environmental correlation and population variability in simple metacommunities. *PLoS One* **8**, e72325.

Sagi, M., Davydov, O., Orazova, S., Yesbergenova, Z., Ophir, R., Stratmann, J.W. & Fluhr, R. (2004) Plant respiratory burst oxidase homologs impinge on wound responsiveness and development in Lycopersicon esculentum. *The Plant Cell* **16**, 616-628.

Sagnard, F., Deu, M., Dembélé, D., Leblois, R., Touré, L., Diakité, M., Calatayud, C., Vaksmann, M., Bouchet, S., Mallé, Y., Togola, S. & Traoré, P.C.S. (2011) Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in a western African region. *Theoretical and Applied Genetics* **123,** 1231-1246.

Santamaria, L. & Mèndez, P.F. (2012) Evolution in biodiversity policy–current gaps and future needs. *Evolutionary Applications* **5**, 202–218.

Schreiber, M., Himmelbach, A., Böner, A. & Mascher, M. (2019) Genetic diversity and relationship between domesticated rye and its wild relatives as revealed through genotyping-by-sequencing. *Evolutionary Applications* **12**, 66-77.

Seabloom, E.W., Bjørnstad, O.N., Bolker, B.M. & Reichman, O.J. (2005) Spatial signature of environmental heterogeneity, dispersal, and competition in successional grasslands. *Ecological monographs* **75**, 199-214.

Serba, D.D., Muleta, K.T., St. Amand, P., Bernardo, A., Bai, G., Perumal, R. & Bashir, E. (2019) Genetic diversity, population structure, and linkage disequilibrium of pearl millet. *Plant Genome* **12**, 180091.

Sgro, C.M., Lowe, A.J. & Hoffmann, A.A. (2011) Building evolutionary resilience for conserving biodiversity under climate change. *Evolutionary Applications* **4**, 326–337.

Shu, K. & Yang, W. (2017) E3 ubiquitin ligases: ubiquitous actors in plant development and abiotic stress responses. *Plant and Cell Physiology* 58, 1461-1476.

Sidhu, D. & Gill, K.S. (2004) Distribution of genes and recombination in wheat and other eukaryotes. *Plant Cell, Tissue, and Organ Culture* **79**, 257-270.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212.

Slatkin, M. (1985) Gene flow in natural populations. *Annual Review of Ecology and Systematics* **16**, 393-430.

Slatkin, M. (1987) Gene flow and the geographic structure of natural populations. *Science* **236**, 787-792.

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I. & Belzile, F. (2015) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnology Journal* **13**, 211-221.

Sonnante, G., Stockton, T., Nodari, R.O., Becerra Velásquez, V.L. & Gepts, P. (1994) Evolution of genetic diversity during the domestication of common-bean (*Phaseolus vulgaris* L.). *Theoretical and Applied Genetics* **89**, 629–635.

Sun, X., Cai, X., Yin, K., Gu, L., Shen, Y., Hu, B., Wang, Y., Chen, Y., Zhu, Y., Jia, B. & Sun, M. (2021) Wild soybean SNARE proteins BET1s mediate the subcellular localization of the cytoplasmic receptor□like kinases CRCK1s to modulate salt stress responses. *The Plant Journal* **105**, 771-785.

Tai, F., Yuan, Z., Li, S., Wang, Q., Liu, F. & Wang, W. (2016) ZmCIPK8, a CBL-interacting protein kinase, regulates maize response to drought stress. *Plant Cell, Tissue and Organ Culture* **124**, 459-469.

Tajima, F. (1989) Statistical method for resting the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Tanksley, S.D. & McCouch, S.R. (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science **277**, 1063-1066.

Thompson, A.L., Luthin C.S., Griffin, M. & Watermolen, L.D. (2010) Wetland restoration handbook for Wisconsin landowners.

Thomson, M.J. (2014) High-throughput SNP genotyping to accelerate crop improvement. *Plant Breeding Biotechnology* **2**, 195-212.

Vos, P.G., Paulo, M.J., Voorrips, R.E., Visser, R.G.F, van Eck, H.J. & van Eeuwijk, F.A. (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SN-array data of tetraploid potato. *Theoretical and Applied Genetics* **130**, 123-135.

Waheed, A.K. (2021) *Manoomin (wild rice) and environmental change at a significant river system of the Lac du Flambeau Band of Lake Superior Chippewa*. M.Sc. dissertation. Twin Cities, MN: University of Minnesota.

Wang, W., Chen, D., Zhang, X., Liu, D., Cheng, Y. & Shen, F. (2018) Role of plant respiratory burst oxidase homologs in stress responses. *Free Radical Research* **52**, 826-839.

Warwick, S. & Aiken, S. (1986) Electrophoretic evidence for the recognition of two species in annual wild rice (*Zizania*, Poaceae). *Systematic Botany* **11**, 464-473.

Webster, C.G. (1964) Fall foods of soras from two habitats in Connecticut. *Journal of Wildlife Management* **28**, 163-165.

Wei, S., Hu, W., Deng, X., Zhang, Y., Liu, X., Zhao, X., Luo, Q., Jin, Z., Li, Y., Zhou, S., Sun, T., Wang, L., Yang, G. & He, G. (2014) A rice calcium-dependent protein kinase OsCPK9 positively regulates drought stress tolerance and spikelet fertility. *BMC Plant Biology* **14**, 1-13.

Weir, B.S. & Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358.

Witte, C.P. (2011) Urea metabolism in plants. *Plant Science* **180**, 431-438.

Xu, X.W., Wu, J.W., Qi, M.X., Lu, Q.X., Lee, P.F., Lutz, S., Ge, S. & Wen, J. (2015) Comparative phylogeography of the wild-rice genus *Zizania* (Poaceae) in eastern Asia and North America. *American Journal of Botany* **102**, 239-247.

Yan, J., Shah, T., Warburton, M.L., Buckler, E.S., McMullen, M.D. & Crouch, J. (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* **4**, e8451.

Yang, L., Ji, W., Zhu, Y., Gao, P., Li, Y., Cai, H., Bai, X. & Guo, D. (2010) GsCBRLK, a calcium/calmodulin-binding receptor-like kinase, is a positive regulator of plant tolerance to salt and ABA stress. *Journal of Experimental Botany* **61**, 2519-2533.

Yuan, Y., Chung, C.Y.L. & Chan, T.F. (2020) Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal* **18**, 2051-2062.

Zeder, M.A., Emshwiller, E. & Smith, B.D. (2006) Documenting domestication: the intersection of genetics and archaeology. *Trends in Genetics* **22**, 139–155.

Tables

Table 1. Summary marker statistics for 5,955 Northern Wild Rice (NWR; *Zizania palustris* L.) bi-allelic single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

Chr/Scaffold	Size (Mbp)	# of SNPs	% of total SNPs	# of SNPs in genic regions	% of SNPs in genic regions per chromosome
ZPchr0001	95.4	483	8.11%	74	15.32%
ZPchr0002	103.4	435	7.30%	56	12.87%
ZPchr0003	58.8	302	5.07%	37	12.25%
ZPchr0004	98.7	491	8.25%	24	4.89%
ZPchr0005	66.6	350	5.88%	41	11.71%
ZPchr0006	118.0	598	10.04%	126	21.07%
ZPchr0007	42.6	181	3.04%	66	36.46%
ZPchr0008	75.7	367	6.16%	38	10.35%
ZPchr0009	95.1	481	8.08%	25	5.20%
ZPchr0010	111.4	568	9.54%	75	13.20%
ZPchr0011	63.2	285	4.76%	41	14.39%
ZPchr0012	105.9	410	6.88%	86	20.98%
ZPchr0013	111.3	583	9.79%	66	11.32%
ZPchr0014	24.0	91	1.53%	28	30.77%
ZPchr0015	39.1	237	3.98%	12	5.06%
ZPchr0016	13.8	88	1.48%	6	6.82%
ZPchr0458	4.3	5	0.08%	3	60.00%

Table 2. Analysis of molecular variance (AMOVA) of a diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.) grouped by germplasm type (Natural Stand and Cultivated collections), STRUCTURE groupings at *K*=5, and species type (*Z. palustris* vs *Zizania aquatica*) based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

Grouping	Source of variation	df	MS	Sigma	% of total	Probability
Natural stand vs. Natural stand	Variations between individuals	12	7440.64	133.71	8.10	0.001
	Variations within individuals	567	1516.08	1516.08	91.90	0.001
	Total of variation	579	1638.87	1649.78	100.00	0.001
Cultivated vs. Cultivated	Variations between individuals	20	1841.49	49.61	3.37	0.001
	Variations within individuals	166	1421.51	1421.51	96.63	0.001
	Total of variation	186	1466.67	1471.11	100.00	0.001
Natural stand vs. Cultivated	Variations between individuals	1	20,955.05	68.42	4.09	0.001
	Variations within individuals	765	1604.18	1604.17	95.91	0.001
	Total of variation	766	1629.44	1672.60	100.00	0.001
STRUCTURE grouping *K*=5	Variations between individuals	5	8508.68	57.96	3.53	0.001
	Variations within individuals	761	1584.24	1584.24	96.47	0.001
	Total of variation	766	1629.44	1642.20	100.00	0.001
Z. palustris vs. *Z. aquatica*	Variations between individuals	1	9517.98	86.37	5.05	0.001
	Variations within individuals	578	1625.24	1625.23	94.95	0.001
	Total of variation	579	1638.87	1711.61	100.00	0.001

df, degrees of freedom; MS, mean of squares

Table 3. Fixation index (F_{ST}) values derived using the weighted Weir and Cockerham method (Weir and Cockerham, 1984) based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS) for a diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.) consisting of a.) A Natural Stand collection; b.) A Cultivated collection; and c.) A comparison of Natural Stand and Cultivated collections. Sample sizes can be found in Table S1.

a.

	Clearwater River	Dahler Lake	Decker Lake	Garfield Lake	Mud Hen Lake	Necktie River	Ottertail River	Phantom Lake	Plantagenet	Shell Lake	Upper Rice Lake	*Z. aquatica*
Bass Lake	0.1040	0.0722	0.0487	0.0802	0.1529	0.0979	0.1201	0.0678	0.1080	0.1254	0.0434	0.1201
Clearwater River		0.1060	0.0934	0.0679	0.1406	0.0734	0.0670	0.0902	0.0410	0.0553	0.0493	0.1004
Dahler Lake			0.0782	0.0996	0.1519	0.1156	0.1233	0.0563	0.1132	0.1255	0.0483	0.1135
Decker Lake				0.0696	0.1490	0.0824	0.1076	0.0680	0.0931	0.1034	0.0388	0.1037
Garfield Lake					0.1515	0.0387	0.0791	0.0885	0.0589	0.0731	0.0459	0.1100
Mud Hen Lake						0.1580	0.1353	0.0861	0.1468	0.1578	0.1107	0.1291
Necktie River							0.0923	0.0996	0.0545	0.0838	0.0646	0.1221
Ottertail River								0.0952	0.0668	0.0494	0.0638	0.1072
Phantom Lake									0.0944	0.1077	0.0417	0.0734
Plantagenet										0.0538	0.0570	0.1024
Shell Lake											0.0628	0.1175

Upper Rice Lake												0.0790

b.

	FY-C20	Itasca-C12	Itasca-C20	K2EF-C16	PM3E
Barron	0.0099	0.0058	0.0155	0.0273	0.0111
FY-C20		0.0063	0.0065	0.0233	0.0086
Itasca-C12			0.0078	0.0279	0.0121
Itasca-C20				0.0337	0.0190
K2EF-C16					0.0276

c.

	Barron	FY-C20	Itasca-C12	Itasca-C20	K2EF-C16	PM3E
Bass Lake	0.0362	0.0342	0.0342	0.0436	0.0440	0.0276
Clearwater River	0.0456	0.0464	0.0472	0.0595	0.0584	0.0431
Dahler Lake	0.0412	0.0418	0.0398	0.0504	0.0515	0.0369
Decker Lake	0.0412	0.0424	0.0422	0.0559	0.0466	0.0338
Garfield Lake	0.0542	0.0500	0.0540	0.0663	0.0571	0.0435
Mud Hen Lake	0.0709	0.0648	0.0694	0.0804	0.0684	0.0693
Necktie River	0.0600	0.0574	0.0589	0.0734	0.0667	0.0510
Ottertail River	0.0530	0.0505	0.0517	0.0586	0.0562	0.0485
Phantom Lake	0.0313	0.0267	0.0316	0.0360	0.0342	0.0262
Plantagenet	0.0495	0.0469	0.0492	0.0587	0.0532	0.0441
Shell Lake	0.0566	0.0561	0.0574	0.0719	0.0621	0.0504
Upper Rice Lake	0.0298	0.0305	0.0296	0.0423	0.0352	0.0240
Z. aquatica	0.0487	0.0479	0.0479	0.0572	0.0518	0.0445

Figures

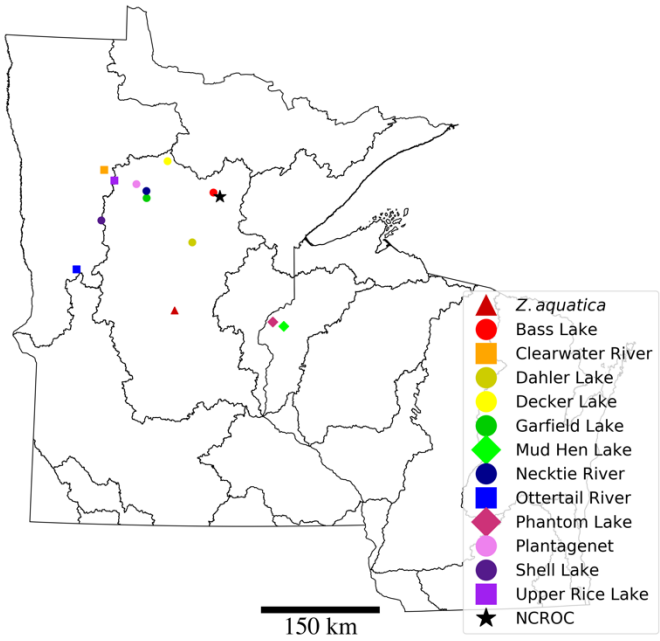

Figure 1. A HUC-8 watershed map of the US states of Minnesota and western Wisconsin showing where leaf tissue samples of a Northern Wild Rice (NWR; *Zizania palustris* L.) diversity collection were collected. Colors and shapes correspond to those featured in the principal component analysis (PCA) plots (Figure 2a-b). Each Natural Stand population has its own color and shapes correspond to watershed membership (circles=Upper Mississippi River; squares=Red River of the North; diamonds=St. Croix River). Exact GPS coordinates can be found in Table S1. The Unweighted Pair Group Method with Arithmetic averaging (UPGMA) cluster analysis with bootstrapping of our Natural Stands and Cultivated collections of Northern Wild Rice (NWR; *Zizania palustris* L.) is also shown to the right of the map.

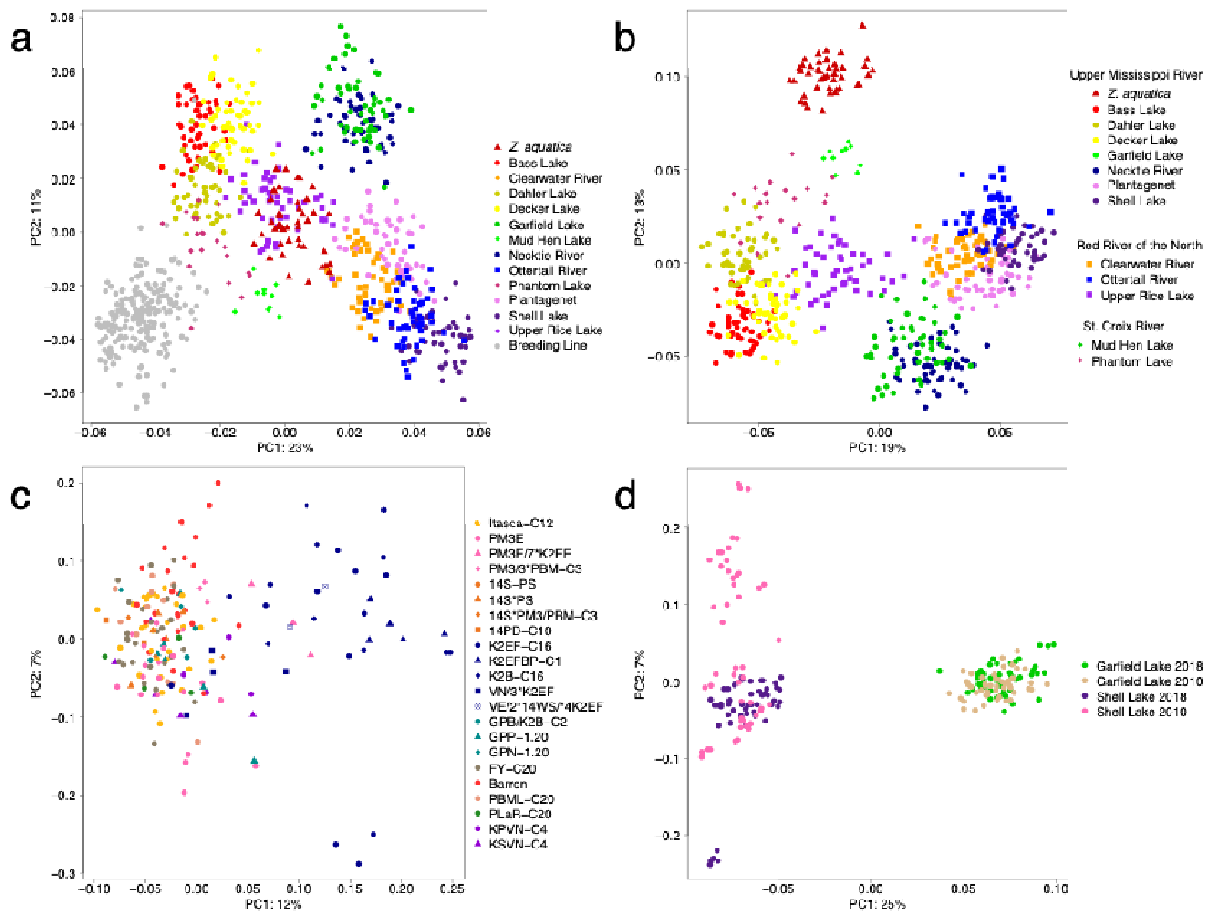

Figure 2. Principal component (PC) analysis (PCA) plots showing the differentiation of the 1st and 2nd PCs of a.) the Natural Stand and Cultivated collections of Northern Wild Rice (NWR; *Zizania palustris* L.); b.) the Natural Stand collection; c.) the Cultivated collection; and d.) the Temporal collection based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

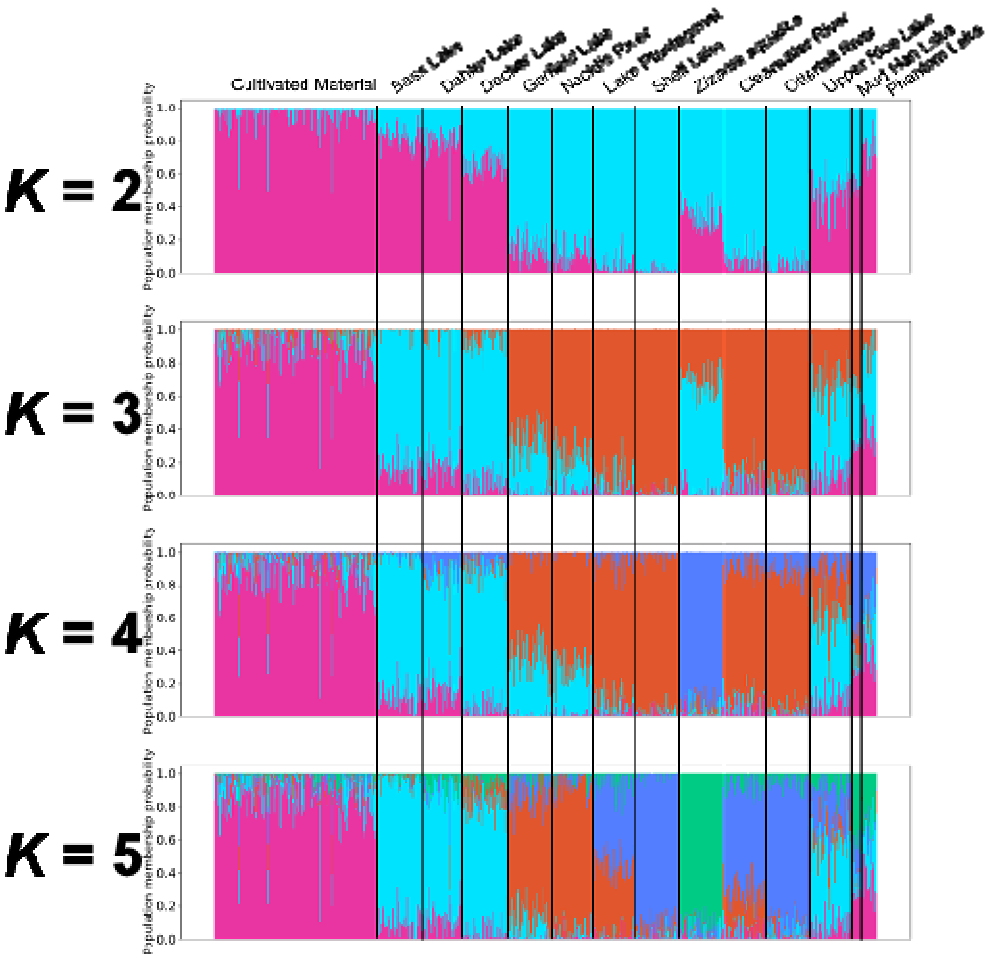

Figure 3. Population structure analysis of Northern Wild Rice (NWR; *Zizania palustris* L.) Natural Stand and Cultivated collections using the program STRUCTURE with 10,000 reps and a burn-in length of 1,000 including: a.) *K*=2; b.) *K*=3; c.) *K*=4; d.) *K*=4, and e.) *K*=5 based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

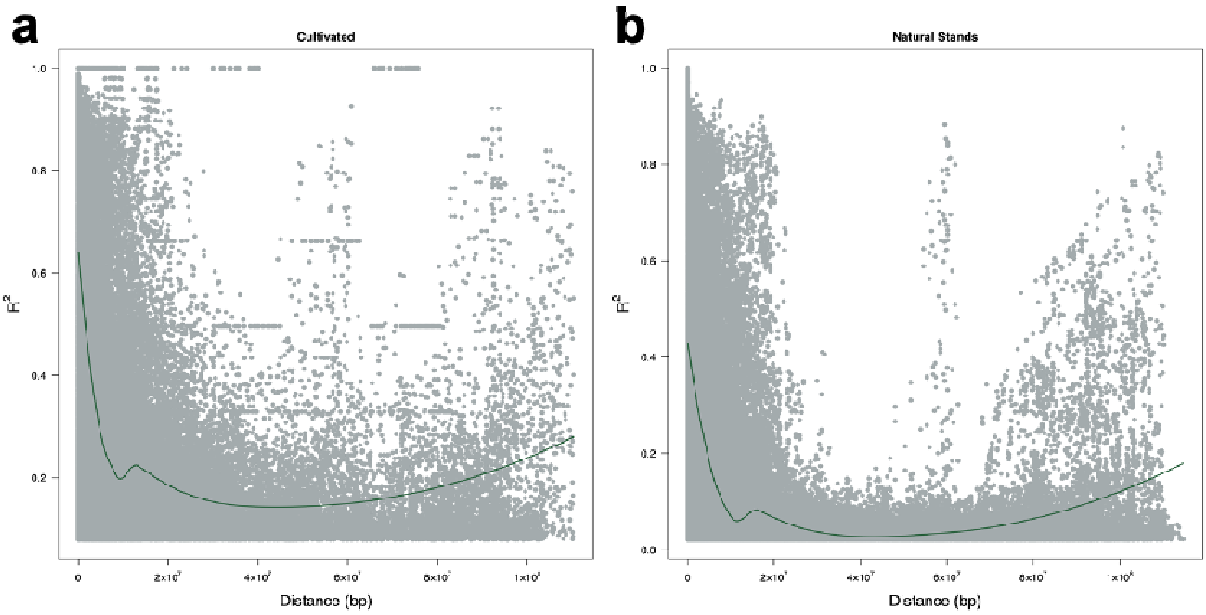

Figure 4. Genome-wide linkage disequilibrium (LD) decay of a.) A Natural Stand collection and b.) A Cultivated collection of Northern Wild Rice (NWR; *Zizania palustris* L.) based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

Figure 5. Genome-wide scans of a.) nucleotide diversity on a per-site basis (π; Nei, 1987); b.) Fixation index (F_{ST}) values derived using the weighted Weir and Cockerham method (Weir and Cockerham, 1984); and c.) XP-CLR scores for a diverse collection of Northern Wild Rice (NWR; *Zizania palustris* L.) based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

Supporting Tables

Table S1. List of samples in our diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.) genotyped with 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS). HUC 8 watershed designations include Upper Mississippi River (UMR), Red River of the North (RRN), and St. Croix River (SCR) basins.

Sample ID	Sample #	Collection Type	GPS Coordinates	Watershed
Bass Lake	50	Natural Stand	47.28716 -93.63132	UMR
Clearwater River	50	Natural Stand	47.51838 -95.46291	RRN
Dahler Lake	50	Natural Stand	46.71888 -93.97281	UMR
Decker Lake	50	Natural Stand	47.63515 -94.40492	UMR
Garfield Lake	50	Natural Stand/Temporal	47.21361 -94.74380	UMR
Mud Hen Lake	10	Natural Stand	45.77005 -92.47173	SCR
Necktie River	50	Natural Stand	47.29363 -94.74724	UMR
Ottertail River	50	Natural Stand	46.38005 -95.86430	RRN
Phantom Lake	20	Natural Stand	45.81839 -92.65074	SCR
Lake Plantagenet	50	Natural Stand	47.36758 -94.91755	UMR
Shell Lake	50	Natural Stand/Temporal	46.94510 -95.48486	UMR
Upper Rice Lake	50	Natural Stand	47.40030 -95.28659	RRN
Zizania aquatica	50	Natural Stand	45.94473 -94.24821	UMR
Garfield Lake 2010	50	Temporal	47.21361 -94.74380	UMR
Shell Lake 2010	50	Temporal	46.94510 -95.48486	UMR
14PD-C10	2	Cultivated	n/a	n/a
14S*PM3/PBM-C3	3	Cultivated	n/a	n/a
14S*PS-C3	5	Cultivated	n/a	n/a
Barron	25	Cultivated	n/a	n/a
Itasca-C12 (Dovetail)	15	Cultivated	n/a	n/a
Itasca-C12	31	Cultivated	n/a	n/a
Itasca-C20	11	Cultivated	n/a	n/a
FY-C20	29	Cultivated	n/a	n/a
GPB/K2B-C2	3	Cultivated	n/a	n/a
GPN-1.20	2	Cultivated	n/a	n/a
GPP-1.20	6	Cultivated	n/a	n/a
K2B-C16	4	Cultivated	n/a	n/a
K2EF-C16	20	Cultivated	n/a	n/a
K2EFBP-C1	5	Cultivated	n/a	n/a
KPVN-C4	4	Cultivated	n/a	n/a
KSVN-C4	5	Cultivated	n/a	n/a
PLaR-C20	4	Cultivated	n/a	n/a
PM3/3*PBM-C3	6	Cultivated	n/a	n/a
PM3/7*K2EF	3	Cultivated	n/a	n/a
PM3E	17	Cultivated	n/a	n/a
VE/2*14WS/*4K2EF	2	Cultivated	n/a	n/a
VN/3*K2EF	4	Cultivated	n/a	n/a

Table S2. Geographic distance (km) matrix of US states of Minnesota and Wisconsin lakes and rivers where Northern Wild Rice (NWR; *Z. palustris* L.) leaf tissue samples were collected along with watershed designations for the Upper Mississippi River (UMR), Red River of the North (RRN), and St. Croix River (SCR).

Natural Population (Watershed)	*Z. aquatica*	Bass Lake	Clearwater River	Dahler Lake	Decker Lake	Garfield Lake	Mud Hen Lake	Necktie River	Ottertail Lake	Phantom Lake	Lake Plantagenet	Shell Lake	Upper Rice Lake
Z. aquatica (UMR)		156.7	198.2	88.7	188.6	146.3	139.1	154.9	133.7	124.6	166.4	146.3	180.4
Bass Lake (UMR)	156.7		140.4	68.4	69.9	84.5	190.8	84.3	197.8	179.9	97.5	145.5	125.5
Clearwater River (RRN)	198.2	140.4		143.7	80.5	64.0	300.2	59.4	130.3	286.3	44.4	63.8	18.7
Dahler Lake (UMR)	88.7	68.4	143.7		107.1	80.4	156.6	86.9	149.6	142.8	101.7	117.9	125.2
Decker Lake (UMR)	188.6	69.9	80.5	107.1		53.4	254.7	45.9	178.3	242.5	48.7	112.0	71.3
Garfield Lake (UMR)	146.3	84.5	64.0	80.4	53.4		236.9	8.9	126.1	223.2	21.6	63.6	45.9
Mud Hen Lake (SCR)	139.1	190.8	300.2	156.6	254.7	236.9		243.2	270.6	14.9	258.2	265.9	281.6
Necktie River (UMR)	154.9	84.3	59.4	86.9	45.9	8.9	243.2		132.6	229.6	15.3	68.0	42.4
Ottertail Lake (RRN)	133.7	197.8	130.3	149.6	178.3	126.1	270.6	132.6		255.8	131.4	69.3	121.8
Phantom Lake (SCR)	124.6	179.9	286.3	142.8	242.5	223.2	14.9	229.6	255.8		244.5	251.2	267.6
Lake Plantagenet (UMR)	166.4	97.5	44.4	101.7	48.7	21.6	258.2	15.3	131.4	244.5		63.7	28.1
Shell Lake (UMR)	146.3	145.5	63.8	117.9	112.0	63.6	265.9	68.0	69.3	251.2	63.7		52.8
Upper Rice Lake (RRN)	180.4	125.5	18.7	125.2	71.3	45.9	281.6	42.4	121.8	267.6	28.1	52.8	

Table S3. List of samples in our diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.) sorted according to the National Center for Biotechnology Information Short Read Archive (NCBI SRA) BioSample accession numbers. The BioProject ID for the collection is PRJNA774842.

See External Excel File 'Table S3'

Table S4. Marker statistics including the-transition/transversion (TsTv) ratios for 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS) using our diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.)

	ZPchr0001	**ZPchr0002**	**ZPchr0003**	**ZPchr0004**	**ZPchr0005**	**ZPchr0006**	**ZPchr0007**	**ZPchr0008**	**ZPchr0009**
Scaffold size (bp)	95,470,783	103,377,072	58,865,324	98,769,966	66,616,710	118,006,097	42,614,780	75,690,682	95,187,837
Number of SNPs	483	435	302	491	350	598	181	367	481
SNP density (SNPs/Mb)	5.06	4.21	5.13	4.97	5.25	5.07	4.25	4.85	5.05
A/C	40	33	30	38	17	60	23	29	44
A/G	169	171	108	205	147	197	57	128	181
A/T	22	16	12	18	17	26	10	13	16
C/G	19	19	17	19	13	34	17	16	17
C/T	197	174	109	178	131	233	58	152	196
G/T	36	22	26	33	25	48	16	29	27
Ts	366	345	217	383	278	430	115	280	377
Tv	117	90	85	108	72	168	66	87	104
TsTv ratio	3.13	3.83	2.55	3.55	3.86	2.56	1.74	3.22	3.63
	ZPchr0010	**ZPchr0011**	**ZPchr0012**	**ZPchr0013**	**ZPchr0014**	**ZPchr0015**	**ZPchr0016**	**ZPchr0458**	**Genome-wide**
Scaffold size (bp)	111,429,323	63,217,741	105,879,435	111,260,184	24,058,870	39,125,943	13,817,767	4,333,358	1,227,721,872
Number of SNPs	568	285	410	583	91	237	88	5	5955
SNP density (SNPs/Mb)	5.10	4.51	3.87	5.24	3.78	6.06	6.37	1.15	4.85
A/C	49	21	25	43	7	15	6	0	480

A/G	213	109	149	234	30	92	40	2	2232
A/T	25	18	21	23	3	9	1	1	251
C/G	31	8	28	28	4	10	4	0	284
C/T	210	114	154	212	39	98	31	1	2287
G/T	40	15	33	43	8	13	6	1	421
Ts	423	223	303	446	69	190	71	3	4519
Tv	145	62	107	137	22	47	17	2	1436
TsTv ratio	2.92	3.60	2.83	3.26	3.14	4.04	4.18	1.50	3.15

Table S5. Polymorphic Information Content (PIC) values for 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS) using our diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.)

	Mean	Median	Minimum	Maximum	Standard Deviation
Barron	0.1649	0.1557	0.0216	0.3125	0.1043
FY-C20	0.1639	0.1557	0.0199	0.3125	0.1052
Itasca-C12	0.1531	0.1379	0.0166	0.3125	0.1064
Itasca-C20	0.1840	0.1829	0.0449	0.3125	0.0981
K2EF-C16	0.1803	0.1829	0.0261	0.3125	0.1053
PM3E	0.1774	0.1829	0.0292	0.3125	0.1041
Bass Lake	0.1484	0.1287	0.0010	0.3125	0.1093
Clearwater River	0.1523	0.1369	0.0113	0.3125	0.1084
Dahler Lake	0.1541	0.1373	0.0142	0.3125	0.1079
Decker Lake	0.1540	0.1395	0.0091	0.3125	0.1111
Garfield Lake	0.1471	0.1263	0.0091	0.3125	0.1090
Mud Hen Lake	0.1924	0.1999	0.0493	0.3125	0.0982
Necktie River	0.1500	0.1260	0.0113	0.3125	0.1108
Ottertail River	0.1528	0.1353	0.0010	0.3125	0.1104
Phantom Lake	0.1637	0.1702	0.0248	0.3125	0.1049
(Lake) Plantagenet	0.1534	0.1353	0.0010	0.3125	0.1105
Shell Lake	0.1477	0.1221	0.0010	0.3125	0.1116
Upper Rice Lake	0.1420	0.1141	0.0010	0.3125	0.1088
Zizania aquatica	0.1496	0.1313	0.0010	0.3125	0.1107

Table S6. *D*-statistics (ABBA-BABA) results for a diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.).

P1	P2	P3	ABBA[a]	BABA[a]	BBAA[a]	D	p-value	F4-ratio	Z[b]
Cultivated	Natural Stands I	Natural Stands II	348.181	336.828	367.406	0.017	0.098	0.193	1.655

[a] Number of ABBA and BABA sites. *Z. aquatica* was used as an outgroup for comparisons.
[b] Z: Z-score

Table S7. Significant values for Tajima's D, F_{ST}, and XP-CLR scores for a diversity collection of Northern Wild Rice (NWR; *Zizania palustris* L.) based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

See External Excel File 'Table S7'

Supporting Figures

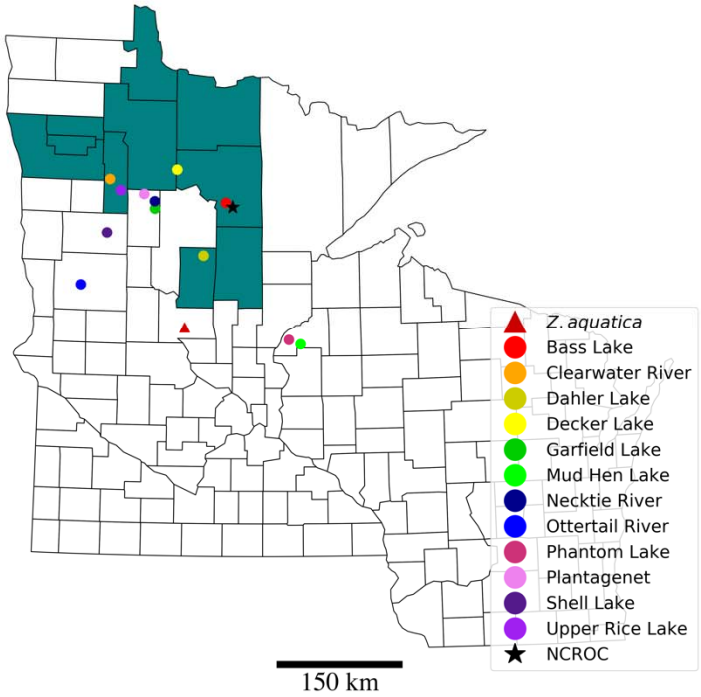

Figure S1. A county-level map of the United States of Minnesota and western Wisconsin showing where leaf tissue samples of our Northern Wild Rice (NWR; *Zizania palustris* L.) diversity collection were collected and highlighting counties with significant production of cultivated NWR. Colors correspond to those featured in the principal component analysis (PCA) plots (Figure 2a-b).

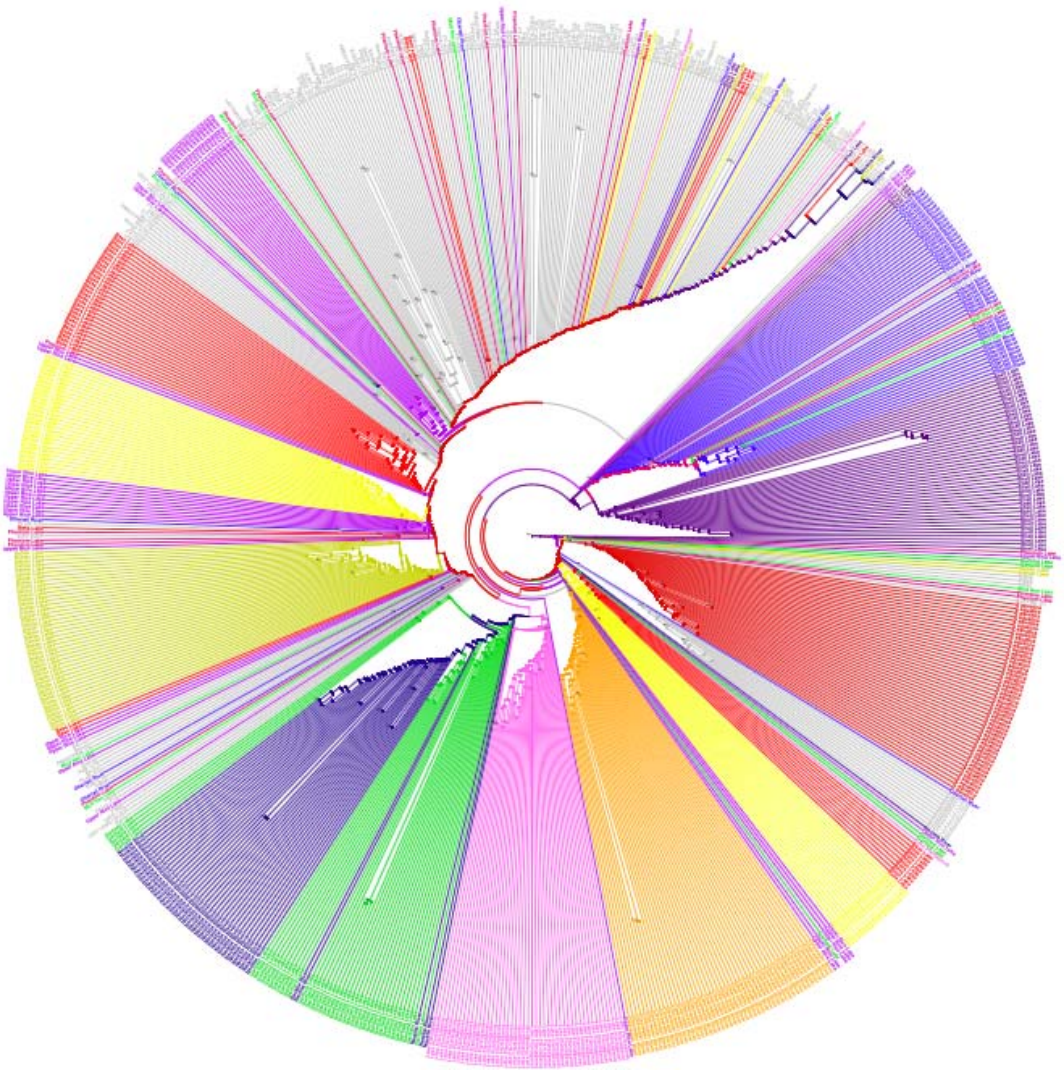

Figure S2. Unweighted pair group method with arithmetic averaging (UPGMA) cluster analysis with bootstrapping of our Natural Stands and Cultivated collections of Northern Wild Rice (NWR; *Zizania palustris* L.).

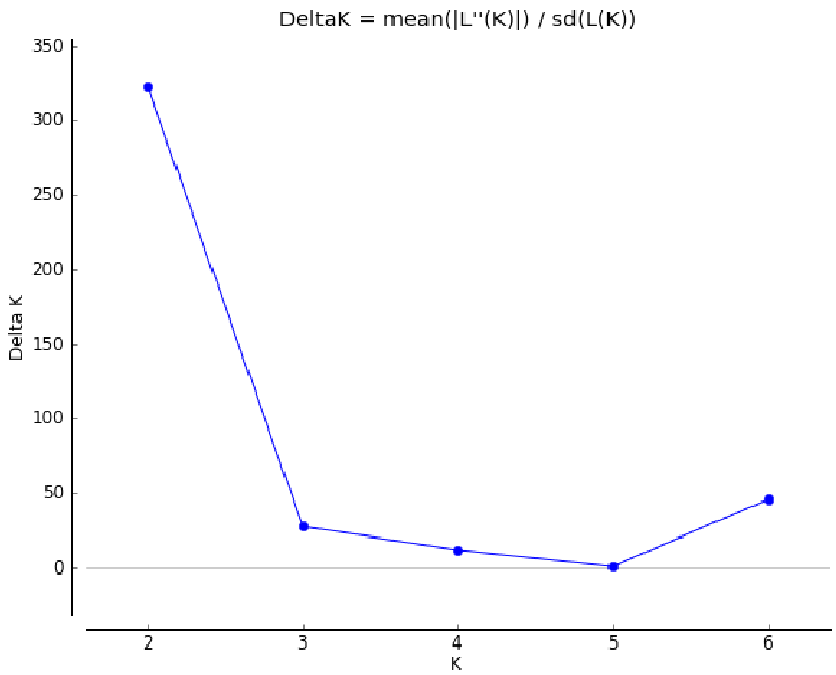

Figure S3. The Evanno method (Evanno 2005) was carried out by uploading our results from STRUCTURE (Pritchard et al. 2000) into the program STRUCTURE harvester (Earl and vonHoldt, 2012). DeltaK is minimized at *K*=5, suggesting that there are 5 subpopulations present in our diversity panel.

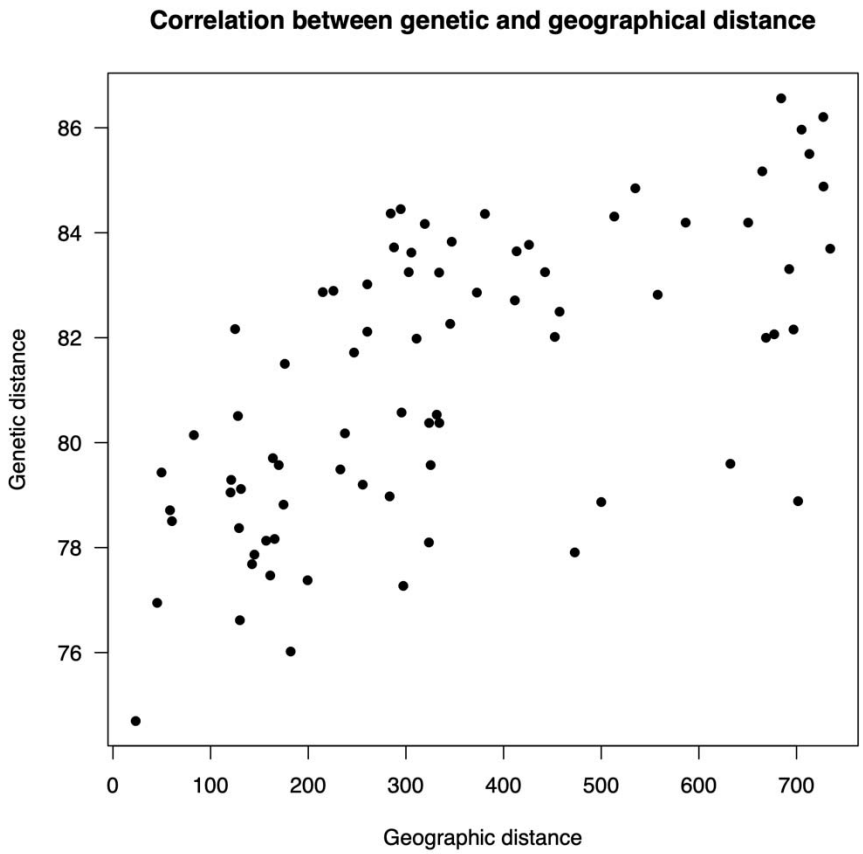

Figure S4. Mantel test results. The plot shows the correlation between geographic distance (x-axis) and genetic distance (y-axis). Geographic distance (in kilometers) and genetic distance (as Prevosti's distance, also used in our UPGMA trees) were used as input. Units shown are not the same as the input units.

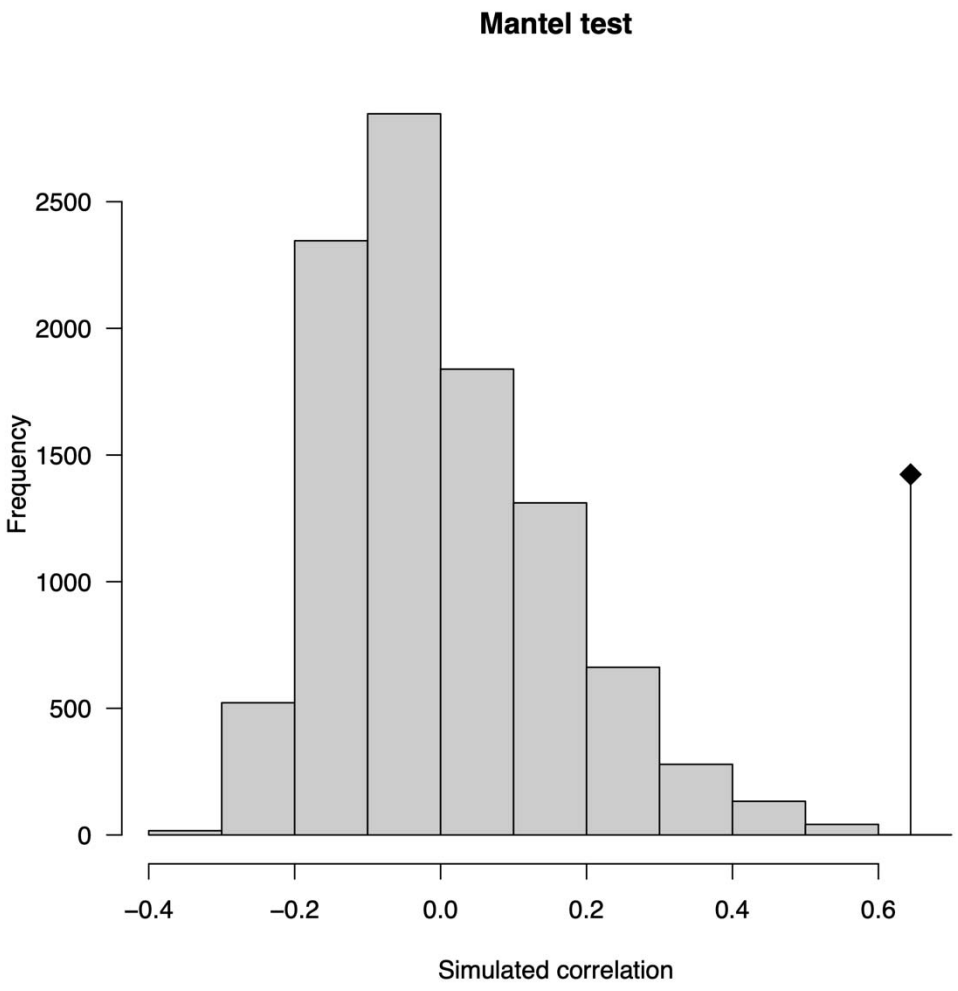

Figure S5. Results of permutation testing. The histogram shows the frequency of simulated correlation tests resulting from permutation tests. The black diamond with a vertical line beneath it shows the actual correlation value from our Mantel (Figure S4) test using real data. This signifies that our results are unlikely to have been reached by chance.

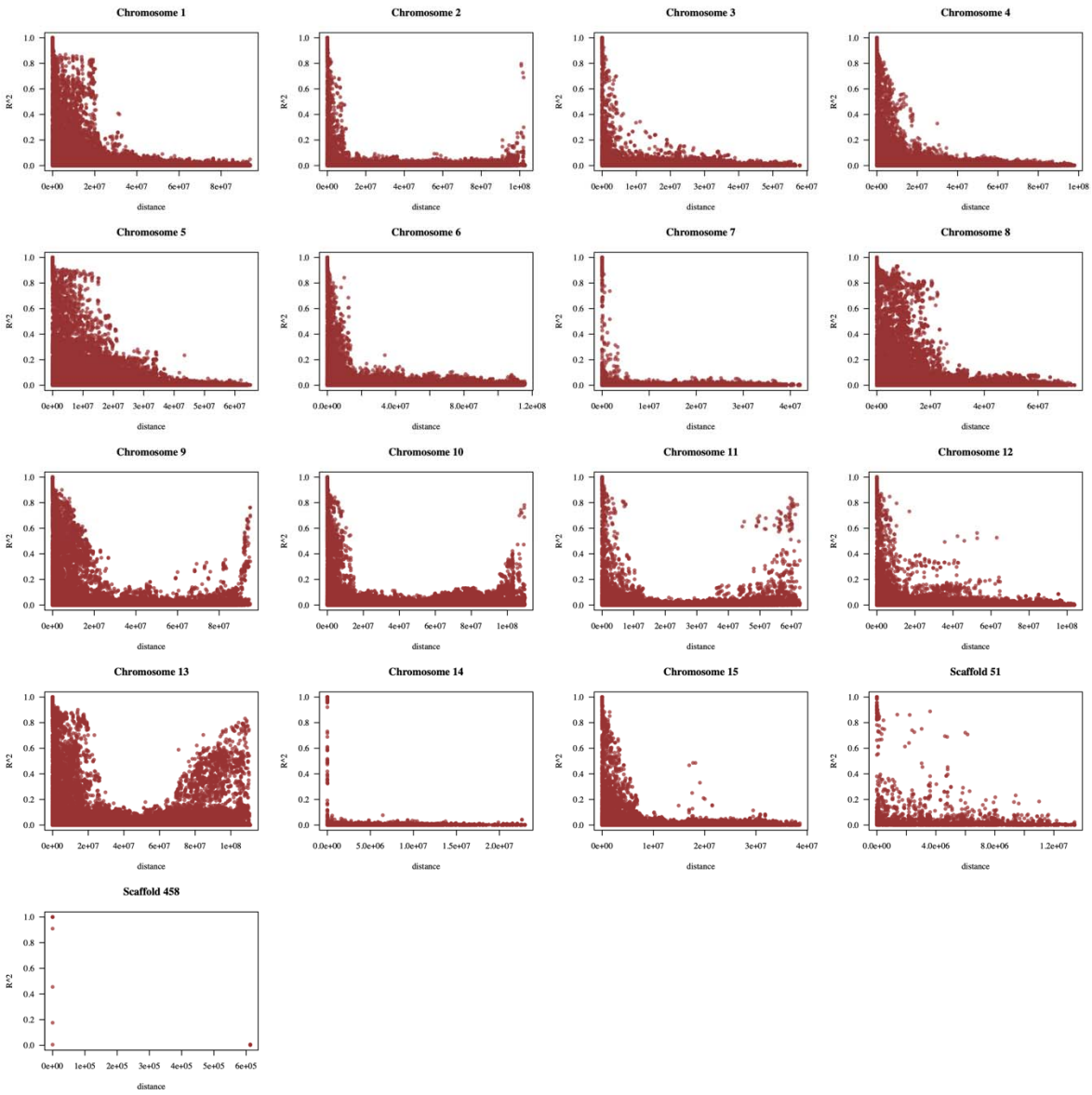

Figure S6. Linkage disequilibrium decay for all chromosomes and scaffolds of Northern Wild Rice (NWR; *Zizania palustris* L.) based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).

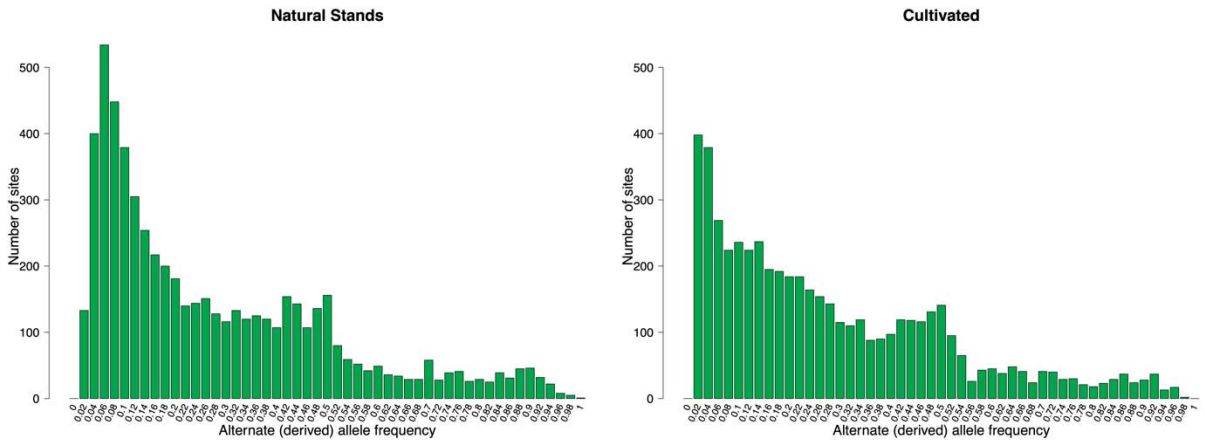

Figure S7. Site spectrum frequency histograms for our a.) Natural Stand collection and b.) Cultivated collection of Northern Wild Rice (NWR; *Zizania palustris* L.) based on 5,955 single nucleotide polymorphism (SNP) markers generated via genotyping-by-sequencing (GBS).