Considerations and Limitations

Genome-wide association studies (GWAS) have become an essential tool for mapping genetic associations. Oryza GenoCLIM and CLIMGeno rely on Genome-Wide Association analysis to retrieve associations between genetic variants and climate variables that are representative of the areas from which the sequenced landraces originated. We recommend that users of these tools become familiar with the particularities of GWA approaches to better interpret the data that we present here (Burghardt, Young, and Tiffin 2017). In this document, we discuss and acknowledge the general limitations of GWA analyses and describe our approaches to address these to the extent possible. The benefits and caveats associated with GWAS have been extensively described in the literature (Tam et al. 2019; Korte and Farlow 2013; Atwell et al. 2010; Platt, Vilhjálmsson, and Nordborg 2010). Here, we explain these limitations in the context of our study. For the non-expert, the most essential point is that **correlation should not be mistaken for causation: GWA studies demonstrate associations whose potential causative relationships require validation**.

1. Multiple testing

For every hypothesis tested, there is an inherent risk of erroneously accepting a false hypothesis. To reduce the probability of accepting a false positive, researchers usually define a level of significance for which they determine that the probability of obtaining a type I error is acceptably low. An issue with all GWA studies is that they do not test a single hypothesis, but rather a vast number of hypotheses, typically testing the association of hundreds of thousands or even millions of genetic variants (SNPs and INDELs) with the trait of interest. Accordingly, the presence of false positives in any GWAS is unavoidable.

Full-genome sequencing data, such as utilized here, provided by the 3K Genomes Project (Wang et al. 2018), facilitates the discovery of causal variants, but there is also a downside: the use of full-genome sequences vastly increases the number of variants tested, increasing the chances of uncovering causative variants, but also increasing false positives (type I error).

Given this, it becomes important to estimate and control for the rate of false positives in such studies (Johnson et al. 2010). Accordingly, we followed a number of steps to reduce the number of tests (genetic variants) considered. To prioritize the most likely causal variants in the association lists that we provide, we filtered out variants in intergenic regions. To prioritize the genetic variants that are more likely to be adaptive, we filtered out those with a frequency < 5%. Furthermore, we include information on the predicted effects of each SNP (e.g., synonymous vs. non-synonymous), locus annotations, its frequency in the population, as well as its likelihood to alter RNA structure, to facilitate an informed prioritization of candidates. Most importantly, to control type I error, we applied the "qvalue" (Johnson et al. 2010) package in R using the Benjamini-Hochberg approach to calculate and provide q-values of the associated SNPs. Oryza CLIMtools provides information on significant genome by environment associations using < 0.01 as the FDR threshold. However, the user can impose a more stringent FDR threshold if desired. Despite our measures to reduce, describe, and control type I error derived from multiple comparisons, users of our data should be aware that it is never possible to completely eliminate false positives resulting from multiple testing in GWAS.

2. Population structure

Perhaps the most studied limitation of GWA studies is the significant presence of false positives resulting from the confounding effects of population structure (Campbell et al. 2005; Lander and Schork 2006; Hey and Machado 2003). Population structure refers to the existing difference in allele frequencies that is observed across populations of the same species over their distribution range. This difference in allele frequencies can be the result of different demographic processes, such as the existing differences among populations in their life histories, population size, isolation by distance, recombination, gene flow, and population bottlenecks. Another factor that shapes the genetic structure of different populations is the different selection pressures that those populations encounter in their native range, including in the case of crop landraces such as in this study, cultural and historical processes that shaped the distribution of germplasm (Gutaker et al. 2020).

It is not trivial to address this issue. For instance, rice temperate and tropical Japonica groups diverged more than 4,200 years ago as a consequence of a global cooling event (Gutaker et al. 2020). When we try to identify the genetic variations that facilitated the adaptation of the temperate Japonica population to colder environments, the challenge comes from the divergence of the temperate and tropical groups and subsequent isolation by distance, as well as other demographic events. This then results in a challenge to identify adaptive genetic variants vs. covarying non-adaptive or neutral variants that emerged through time as a result of the divergent demographic histories of these two populations.

The 3K rice population displays a well-studied population structure (Wang et al. 2018). Mixed models and principal component analysis can be used to correct for population structure. Here, we used Principal Component Analysis (PCA), a widely used method to detect and account for population structure in association analysis. The inferred principal components capturing the genetic ancestry of each genotype are included as fixed effects in a regression-based test of association in order to account for population structure (Patterson, Price, and Reich 2006; Price et al. 2006). In the S1 document of the article describing these tools, we describe in detail the considerations we followed to correct for population structure in our analysis (Ferrero-Serrano et al. 2023).

Even after implementing a correction for population structure, results are expected to include significant statistical inflation that introduces false positives (type I error). At the same time, one can also expect that there will be some number of false negative results (type II error) arising from over-correction when addressing the confounding effects of population structure, especially for any environmental pattern that follows a discrete geographical pattern. For example, as reported by Lasky, Josephs, and Morris (2022) (Lasky, Josephs, and Morris 2022) in African sorghum landraces, the loss of photoperiod sensitivity caused by natural variants in *MATURITY1* increases with decreasing latitude. However, this association was not significant in a GWA analysis after accounting for population structure (Lasky et al. 2015).

In summary, the absence of correction for population structure increases the chances of type I error, and correction for population structure increases the chances of missing true associations (type II error). Because we were more concerned in this study with reducing the chance of type I error (false positives) than type II error (false negatives), we retained the mixed model approach of correction for population structure.

3. Sampling

The outcome of a GWA analysis is conditioned by the sampled individuals that are sequenced and tested for association. Because populations from the 3K landrace population are not evenly collected throughout their distribution range, we can expect that the existing genetic variation is best described for those populations that may be over-represented.

While the sampling is therefore not perfect, from the observation of the latitudinal and longitudinal distributions of the included landraces (Figure 1) we can conclude that they are fairly evenly distributed. This is especially true considering the enormous distribution range covered. Given the large sample size and distribution, including very geographically distant accessions, our analysis has an advantage and a caveat. The massive amount of genetic and environmental variance included in the analysis, derived from a very large and widespread sample, increases the power and confidence in the resulting candidates. The caveat has to do, as discussed earlier regarding population structure, with an increased proportion of false negatives due to genetic heterogeneity. This is because, in some cases, more than one co-correlated genetic variant, which differs in frequencies among different populations, underlies the same trait. Such would be the case for two different SNPs affecting the same codon resulting in a weakened correlation for each SNP with the tested climate.

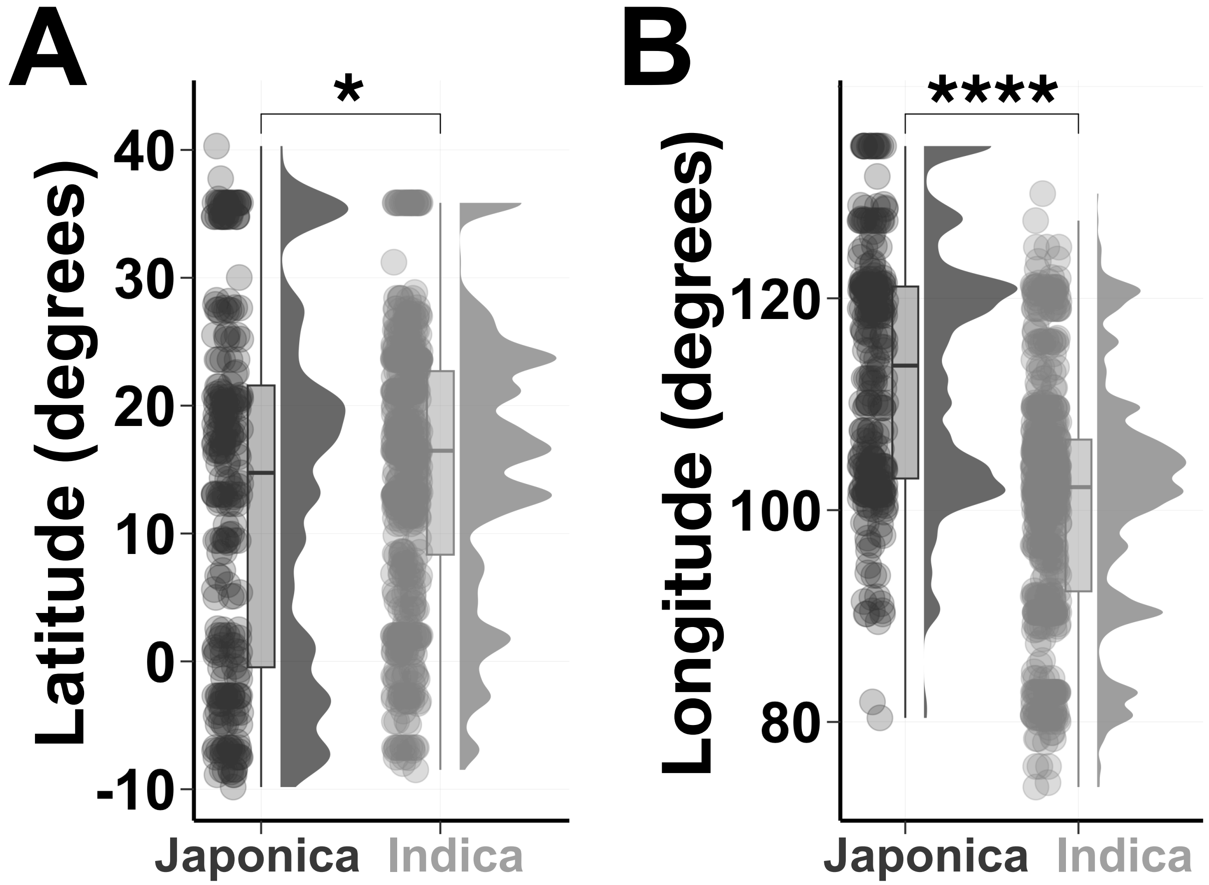

Figure 1. Latitudinal (A) and longitudinal (B) distribution of the Japonica and Indica landrace samples included in this study. Pairwise nonparametric Wilcoxon tests were conducted to assess differences between alleles between subspecies. **** (p <0.0001); * (0.01 < p < 0.05).

4. Synthetic associations

Synthetic associations are the association of non-causative markers that are in linkage disequilibrium with causative markers. In CLIMGeno, we provide the tools to explore the scores, q-values, allele frequencies, and the predicted effect of any identified variant. Importantly, to facilitate the identification of synthetic associations, we also provide the ability to select a genomic window of a customizable length around any variant of interest to visually explore the possible presence of synthetic associations.

For example, in Arabidopsis within the *FRIGIDA* (FRI) gene, there is a variety of naturally occurring loss-of-function alleles, occurring at different frequencies in natural populations, that affect the expression of *FLOWERING LOCUS C*, a key regulator of flowering time that was first identified through natural variation before the implementation of GWA studies in Arabidopsis (Johanson et al. 2000). Despite the fact that a considerable fraction of the variation in flowering time is explained by the existing natural variation in FRI, it has proven difficult to identify *FRI* using GWAS (Atwell et al. 2010), mostly because it has such high allelic heterogeneity (Sasaki et al. 2021), resulting in a false negative.

Similarly, synthetic associations can also result in false positives. A previously published GWA study highlights the *AOP2/AOP3* cluster of glucosinolate biosynthesis genes as candidates to regulate flowering time in Swedish Arabidopsis populations. However, this was later determined to be a spurious, “synthetic” association, derived from the existence of two statistically associated causative loci within linkage disequilibrium (Sasaki et al. 2021).

5. Limitations addressed by removing rare variants

The power of GWAS depends on the phenotypic (and/or environmental) variance within the studied population, as explained by the SNPs found in association. For this reason, rare variants, present in a limited number of individuals in the population, present problems for GWAS (Asimit and Zeggini 2010; Gibson 2012). Mixed models, which are effective in finding associations with common variants, have been found to be susceptible to spurious associations with variants with rare allele frequencies (Price et al. 2010). Additionally, rare variants are prone to be found in association with many other rare variants more often than with common variants (Dickson et al. 2010). For example, all variants present in just one rice landrace will necessarily be found in association with each other. Because here we are interested in likely adaptive variants, found in higher frequencies, we addressed this limitation by filtering out uncommon variants (MAF < 5%). Despite great improvements in sequencing technologies, the error rate of Next Generation Sequencing is approximately 1% (Fox et al. 2014), to which is added the inherent human error rate derived from obtaining and manipulating sequencing data from diverse research facilities (Buell 2018; Asimit and Zeggini 2010). Our removal of variants with MAF < 5% also reduces the chances of inclusion of these errors in our datasets.

Literature Cited

Asimit, Jennifer, and Eleftheria Zeggini. 2010. “Rare Variant Association Analysis Methods for Complex Traits.” *Annual Review of Genetics* 44: 293–308.

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. 2010. “Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines.” *Nature* 465 (7298): 627–31.

Buell, C. Robin. 2018. “Trust but Verify: A Lesson in Technology Limitations and Error Propagation.” *The Plant Cell* 30 (3): 515–16.

Burghardt, Liana T., Nevin D. Young, and Peter Tiffin. 2017. “A Guide to Genome-Wide Association Mapping in Plants.” *Current Protocols in Plant Biology* 2 (1): 22–38.

Campbell, Catarina D., Elizabeth L. Ogburn, Kathryn L. Lunetta, Helen N. Lyon, Matthew L. Freedman, Leif C. Groop, David Altshuler, Kristin G. Ardlie, and Joel N. Hirschhorn. 2005. “Demonstrating Stratification in a European American Population.” *Nature Genetics* 37 (8): 868–72.

Dickson, Samuel P., Kai Wang, Ian Krantz, Hakon Hakonarson, and David B. Goldstein. 2010. “Rare Variants Create Synthetic Genome-Wide Associations.” *PLoS*

Biology 8 (1): e1000294.

Ferrero-Serrano, Ángel, David Chakravorty, Kobie J. Kirven, and Sarah M. Assmann. 2023. “Oryza CLIMtools: An Online Portal for Investigating Genome-Environment Associations in Rice.” *BioRxiv : The Preprint Server for Biology*, June. https://doi.org/10.1101/2023.05.10.540241.

Fox, Edward J., Kate S. Reid-Bayliss, Mary J. Emond, and Lawrence A. Loeb. 2014. “Accuracy of Next Generation Sequencing Platforms.” *Next Generation, Sequencing & Applications* 1. https://doi.org/10.4172/jngsa.1000106.

Gibson, Greg. 2012. “Rare and Common Variants: Twenty Arguments.” *Nature Reviews. Genetics* 13 (2): 135–45.

Gutaker, Rafal M., Simon C. Groen, Emily S. Bellis, Jae Y. Choi, Inês S. Pires, R. Kyle Bocinsky, Emma R. Slayton, et al. 2020. “Genomic History and Ecology of the Geographic Spread of Rice.” *Nature Plants* 6 (5): 492–502.

Hey, Jody, and Carlos A. Machado. 2003. “The Study of Structured Populations — New Hope for a Difficult and Divided Science.” *Nature Reviews. Genetics* 4 (7): 535–43.

Johanson, Urban, Joanne West, Clare Lister, Scott Michaels, Richard Amasino, and Caroline Dean. 2000. “Molecular Analysis of *FRIGIDA*, a Major Determinant of Natural Variation in *Arabidopsis* Flowering Time.” *Science* 290 (5490): 344–47.

Johnson, Randall C., George W. Nelson, Jennifer L. Troyer, James A. Lautenberger, Bailey D. Kessing, Cheryl A. Winkler, and Stephen J. O’Brien. 2010. “Accounting for Multiple Comparisons in a Genome-Wide Association Study (GWAS).” *BMC Genomics* 11 (December): 724.

Korte, Arthur, and Ashley Farlow. 2013. “The Advantages and Limitations of Trait Analysis with GWAS: A Review.” *Plant Methods* 9 (1): 1–9.

Lander, Eric S., and Nicholas J. Schork. 2006. “Genetic Dissection of Complex Traits.” *FOCUS* 4 (3): 442–58.

Lasky, Jesse R., Emily B. Josephs, and Geoffrey P. Morris. 2022. “Genotype–Environment Associations to Reveal the Molecular Basis of Environmental Adaptation.” *The Plant Cell* 35 (1): 125–38.

Lasky, Jesse R., Hari D. Upadhyaya, Punna Ramu, Santosh Deshpande, C. Tom Hash, Jason Bonnette, Thomas E. Juenger, et al. 2015. “Genome-Environment Associations in Sorghum Landraces Predict Adaptive Traits.” *Science Advances* 1 (6): e1400218.

Patterson, Nick, Alkes L. Price, and David Reich. 2006. “Population Structure and Eigenanalysis.” *PLoS Genetics* 2 (12): e190.

Platt, Alexander, Bjarni J. Vilhjálmsson, and Magnus Nordborg. 2010. “Conditions under Which Genome-Wide Association Studies Will Be Positively Misleading.” *Genetics* 186 (3): 1045–52.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. “Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies.” *Nature Genetics* 38 (8): 904–9.

Price, Alkes L., Noah A. Zaitlen, David Reich, and Nick Patterson. 2010. “New Approaches to Population Stratification in Genome-Wide Association Studies.” *Nature Reviews. Genetics* 11 (7): 459–63.

Sasaki, Eriko, Thomas Köcher, Danièle L. Filiault, and Magnus Nordborg. 2021. "Revisiting a GWAS Peak in Arabidopsis Thaliana Reveals Possible Confounding by Genetic Heterogeneity." *Heredity* 127 (3): 245–52.

Tam, Vivian, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. 2019. "Benefits and Limitations of Genome-Wide Association Studies." *Nature Reviews. Genetics* 20 (8): 467–84.

Wang, Wensheng, Ramil Mauleon, Zhiqiang Hu, Dmytro Chebotarov, Shuaishuai Tai, Zhichao Wu, Min Li, et al. 2018. "Genomic Variation in 3,010 Diverse Accessions of Asian Cultivated Rice." *Nature* 557 (7703): 43–49.