Evaluation of association methods

The GWAS analysis in this study was performed using TASSEL v5.2.84 statistical software (Bradbury et al., 2007). TASSEL provides the opportunity to include indels (insertion & deletions) in the analysis, which most other GWAS software ignore. Given the presence of indels in our dataset, we used TASSEL for our study.

Before we decided on the most appropriate method to conduct GWA studies, we evaluated four models using a publicly available flowering time dataset, from which we extracted data corresponding to the landrace collection that we evaluated throughout our study: (Mansueto et al., 2017): i) a general linear model (GLM), ii) a general linear model containing a principal component analysis (PCA) correction for population structure (GLM + PCA), iii) a mixed linear model (MLM) including a kinship matrix (MLM + K) as a random effect, and iv) an MLM including both population structure and kinship relations (MLM + PCA + K). For iii and iv, the MLM used the Centered IBS matrix calculated in Tassel. GWAS results are not sensitive to the number of PCs (Price et al., 2006). For this reason, in the GLM + PCA analysis and MLM + PCA + K, we chose the first five components as the population structure matrix to conduct our GWAS (Zhao et al., 2018). The kinship matrix in the MLM models was constructed using the 'Centered IBS' kinship method in TASSEL. The MLM analysis was conducted with the options of compression and re-evaluation of variances at each marker.

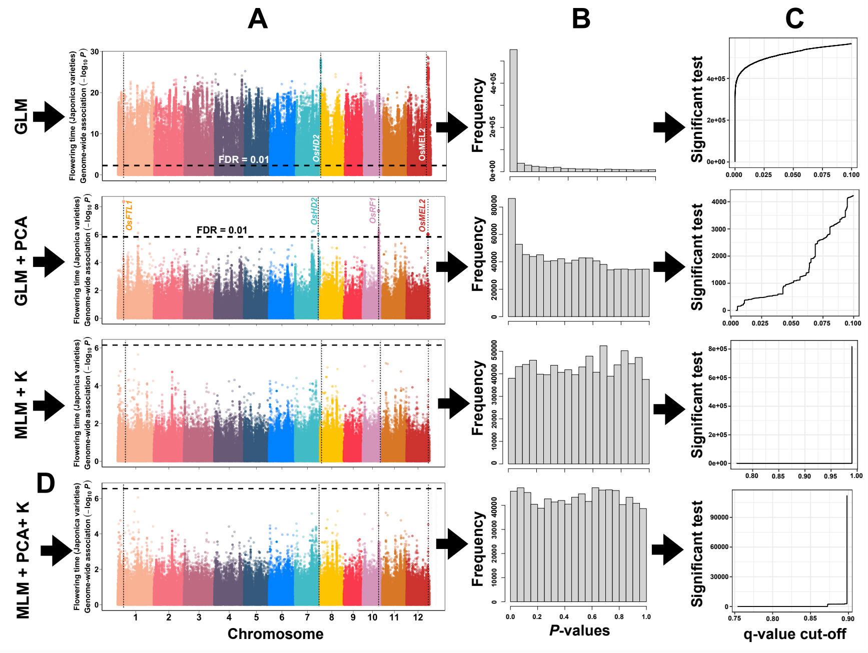

Figure1. GLM + PCA is the most appropriate model to conduct GWAS in our dataset. A. Manhattan plots displaying the results from four GWAS models using publicly available flowering time data. See text for a description of Manhattan plots. **B.** Distribution of *P*-values with the results of these four GWAS models. **C.** Number of significant associations obtained from these four models as a function of the q-value cut-off.

The Manhattan plots represent the *P*-values of the four GWAS models we tested (Fig. 1A). Datapoints depicting individual genetic variants are represented in genomic order by color-demarcated chromosome and position (x-axis). The value on the y-axis represents the score of the association of each variant with natural variation in flowering time for Japonica landraces. The score is the $-\log_{10}$ of the *P*-value, so the higher the score (and therefore, the lower the *P*-value), the stronger the association of any given variant with flowering time. As a result of the local correlation of causative genetic variants with neighboring variants arising from the relationship between genetic distance and recombination, groups of significant *P*-values tend to rise up on the plot, making the graph look like a Manhattan skyline, hence the name.

Flowering time, also known in rice as the heading date, is an appropriate trait to test the effectiveness of these different GWAS models in our dataset because the switch from vegetative to reproductive development is an important adaptive trait that is well-studied, and the causality of several QTLs previously identified through GWAS analysis has been demonstrated. From the study of the resulting Manhattan plots resulting from these four GWAs methods, we can identify in the GLM + PCA method four clear peaks within the FDR < 0.01 threshold in the regions of: 1) *OsFTL1* (*Os01g0218500*; *FLOWERING LOCUS T* (*FT*)-Like homolog), 2) *OsHD2* (*Os07g0695100*; *HEADING DATE 2*), 3) *OsRF1* (*Os10g0497432*; *RESTORATION OF FERTILITY 1*), and 4) *OsMEL2* (*Os12g0572800*; *MEIOSIS ARRESTED AT LEPTOTENE 2*). These genes have previously been functionally associated with the transition front vegetative to reproductive phase in rice (Komori et al., 2004; Nonomura et al., 2011; Ogiso-Tanaka et al., 2013; Liu et al., 2021), increasing confidence in the robustness of the GLM + PCA model for our application.

These peaks in the Manhattan plot are also present in the GLM model, but we cannot identify these 'towers' due to the confounding effects introduced by false positives arising from the structure of the population. Indeed, a major problem in association mapping is controlling the spurious associations that can arise from population structure. This becomes evident when we observe the distinct inflation of *P*-values that approach statistical significance in the GLM method (Fig. 1A and B). This inflation in *P*-values makes it impossible to reliably identify causal variants. The GLM with PCA model is expected to reduce false positives that arise due only to population structure (Price et al., 2006), as we demonstrate in Figure 1 for flowering time. As a result, the PCA correction for population structure allows us to identify the appropriate candidates.

The MLM with PCA and K model includes the kinship matrix in the model and is expected to further reduce false positives that arise from family relatedness (Yu et al., 2006). The MLM model alone has been reported to perform better than the GLM model alone by controlling false positives (Yu et al., 2006). The advantage of the MLM model to prevent false positives disappears for complex traits when these are associated with population structure resulting from extensive genetic divergence. The MLM model controls the *P*-value inflation well but also increases false negatives, weakening the identification of true associations (Zhang et al., 2010). Indeed, our analysis shows that the MLM model introduces an over-correction, i.e. a type II error, that renders this analysis problematic by completely removing identification of true positives. This is evident from the inspection of the Manhattan plots, where there are no significant associations using a significance threshold of FDR of 1% in the MLM models -- the four 'towers' that we identify in the GLM + PCA model, which correspond to four genes known to be flowering-related, are not present.

To further illustrate this, we applied the 'qvalue' package in R (R Core Team) to the *P*-values resulting from each of these four GWAS models to determine the number of significant associations versus each q-value cut-off (Fig. 1C). As a result, we can observe that in the GLM model, there is an inflation in the number of significant tests. This inflation is corrected once we correct for population structure (GLM + PCA). MLM models result in overcorrection and the absence of significant associations regardless of the significance threshold we impose (Figure 1C for MLM + PCA + K and MLM + K).

After careful consideration, we determined that GLM + PCA is the most appropriate model for our study. In support of this decision, previous work utilizing this GWA model in rice has proven its efficacy (Sales et al., 2017; Mogga et al., 2019; Vargas et al., 2020; Bai et al., 2022). We acknowledge that despite the imposition of correction for population structure (compared to the inflation of significant *P*-values present in the GLM model that we correct in the GLM + PCA model), our results, as is inherently true for all GWA studies, will unavoidably have a number of spurious associations. When interpreting our results, it is important to remember that correlation does not equal causation, and any associations found here cannot be assumed to be definitely adaptive or causative. We have reduced type I error, but the presence of spurious associations among the identified causative variants, while minimized, is to be expected.

MLM models are computationally demanding and require substantial computational resources and time. In our study, we conducted GWA analysis in SNPs and INDEL datasets from 658 Indica and 283 Japonica rice landraces to obtain genotype–environment associations using our dataset of 413 environmental variables. We note in passing that the computational power needed to compute the 1,656 GWAS analyses that were necessary to complete this study was alleviated by using the GLM + PCA model.

Literature cited

Bai, S., Hong, J., Su, S., Li, Z., Wang, W., Shi, J., Liang, W., and Zhang, D. (2022). Genetic basis underlying tiller angle in rice (Oryza sativa L.) by genome-wide association study. *Plant Cell Rep.* **41**:1707–1720.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**:2633–2635.

Komori, T., Ohta, S., Murai, N., Takakura, Y., Kuraya, Y., Suzuki, S., Hiei, Y., Imaseki, H., and Nitta, N. (2004). Map-based cloning of a fertility restorer gene, Rf-1, in rice (Oryza sativa L.). *Plant J.* **37**:315–325.

Liu, X., Liu, H., Zhang, Y., He, M., Li, R., Meng, W., Wang, Z., Li, X., and Bu, Q. (2021). Fine-tuning Flowering Time via Genome Editing of Upstream Open Reading Frames of Heading Date 2 in Rice. *Rice* **14**:59.

Mansueto, L., Fuentes, R. R., Borja, F. N., Detras, J., Abriol-Santos, J. M., Chebotarov, D., Sanciangco, M., Palis, K., Copetti, D., Poliakov, A., et al. (2017). Rice SNP-seek database update: new SNPs, indels, and queries. *Nucleic Acids Res.* **45**:D1075–D1081.

Mogga, M., Sibiya, J., Shimelis, H., Mbogo, D., Muzhingi, T., Lamo, J., and Yao, N. (2019). Correction: Diversity analysis and genome-wide association studies of grain shape and eating quality traits in rice (Oryza sativa L.) using DArT markers. *PLoS One* **14**:e0212078.

Nonomura, K.-I., Eiguchi, M., Nakano, M., Takashima, K., Komeda, N., Fukuchi, S., Miyazaki, S., Miyao, A., Hirochika, H., and Kurata, N. (2011). A novel RNA-recognition-motif protein is required for premeiotic G1/S-phase transition in rice (Oryza sativa L.). *PLoS Genet.* **7**:e1001265.

Ogiso-Tanaka, E., Matsubara, K., Yamamoto, S.-I., Nonoue, Y., Wu, J., Fujisawa, H., Ishikubo, H., Tanaka, T., Ando, T., Matsumoto, T., et al. (2013). Natural variation of the RICE FLOWERING LOCUS T 1 contributes to flowering time divergence in rice. *PLoS One* **8**:e75959.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**:904–909.

R Core Team *R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria URL (2020).*

Sales, E., Viruel, J., Domingo, C., and Marqués, L. (2017). Genome wide association analysis of cold tolerance at germination in temperate japonica rice (Oryza sativa L.) varieties. *PLoS One* **12**:e0183416.

Vargas, Y., Rubio, A., Petro, E. E., and Camila Rebolledo, M. (2020). GWAS for low radiation tolerance during grain filling in rice (Oryza sativa L.). In *2020 Virtual Symposium in Plant Omics Sciences (OMICAS)*, pp. 1–5. ieeexplore.ieee.org.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**:203–208.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**:355–360.

Zhao, H., Mitra, N., Kanetsky, P. A., Nathanson, K. L., and Rebbeck, T. R. (2018). A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS). *Stat. Appl. Genet. Mol. Biol.* **17**.