INSUSFAR

INnovative approaches to optimize genetic diversity for SUStainable FARming systems of the future

UNIKASSEL VERSITÄT

Association analysis in lines derived from winter wheat CCPs—comparing four different populations stratification methods

Dominic Dennenmoser, Jelena Baćanović-Šišić, Gunter Backes

University of Kassel, Faculty of Organic Agricultural Sciences, ddennenmoser@uni-kassel.de

Background

- Genome-wide association studies (GWAS) attempt to identify links between gene loci and trait expressions. In order to avoid false positives, GWAS methods use information about population structure, which might have disadvantageous effects in association studies. Several methods are used to describe and integrate this additional information in GWAS.
- However, structures might feature discrete as well as continuous patterns of variation which cannot be identified sufficiently by current (linear) analysis approaches (Diaz-Papkovich et al. 2019). Therefore, GWAS models using non-linear methods (msMDS, NMDS and UMAP) were compared with those using linear methods (PCA, PCoA, iMDS) by calculating the pairwise correlation coefficient of the p-values yielded from GWAS models and the resulting relationships were visualised by UMAP.

Material and Methods

- DNA was isolated from 184 CCP lines derived from two winter wheat CCPs and genotyped using a 20k wheat SNP array (TraitGenetics).
- The genotyping data, together with the phenotypic data are being used for the GWAS to link allelic changes to trait expressions.
- GAPIT-related GWAS: general linear model (GLM), mixed linear model (MLM), multi-locus mixed model (MLMM), fixed and random models circulating probability unification (FarmCPU) included K and Q matrix (Wang and Zhang 2019).
- Covariates were calculated using GAPIT-based PCA on 5822 selected SNPs. Additionally, principal coordinate analysis (PCoA), interval, M-spline, and ordinal (non-metric) multi-dimensional scaling (MDS) using MM algorithm initialised by Torgerson configuration (de Leeuw and Mair 2009), as well as uniform manifold approximation and projection (UMAP) initialised by spectral embedding (McInnes et al. 2018) were calculated using 583 SNPs selected by clumping.
- Altogether, 76 combinations (Table 1) were compared by calculating Pearson correlation coefficient of the p-values yielded from the GWAS models, converted to Euclidian distances ($\delta_r = \sqrt{1-\rho}$).

Results

- The results of GLM-, MLM, and MLMM-based models tend to cluster together, whereas FarmCPU shows different outcomes.
- UMAP yielded the best results for correcting PS used in GLM for the plant height.
- PCA outperformed MDS-based PS methods, and little differences were observed between PS configurations for MLM- and MLMM-based models. In contrast, FarmCPU-based models tend to be conservative: the correction for PS with PCA tends to be too strong.

Conclusions

- The preliminary results are promising and show a potential to use additional covariate methods for GWAS when analysing data derived from diverse CCP lines of wheat.
- Therefore, further tests and comparison with different environments, GWAS methods, and settings are needed, especially for the fine-tuning of UMAP-based methods.

Table 1: Combination of GWAS methods (GLM, MLM, MLMM, and FarmCPU) and populations stratification methods (PCA, PCoA, interval/M-spline/ordinal MDS with TC, and UMAP-Sp).

PS method[a]	GWAS method[b]				No. of components
	GLM	*MLM*	*MLMM*	*FarmCPU*	
None	○	○	○	○	–
PCA	○	○	○	○	2–9
PCoA	○	○	○	○	2–3
(i/ms/N)MDS-TC	○/○/○	○/○/○	○/○/○	○/○/○	2–3
UMAP-Sp	○	○	○	○	2–3
Total number	19	19	19	19	76[c]

[a] PS: populations stratification, PCA: principal component analysis, PCoA: principal coordination analysis, MDS: multi-dimensional scaling, iMDS: interval MDS, msMDS: M-spline MDS, NMDS: ordinal (aka. non-metric) MDS, TC: Torgerson (initial) configuration, UMAP: uniform manifold approximation and projection, Sp: spectral embedding (initiation). [b] GLM: general linear model, MLM: mixed linear model, MLMM: multi-locus mixed model, FarmCPU: fixed and random models circulating probability unification. [c] Total number of calculated models for each trait.

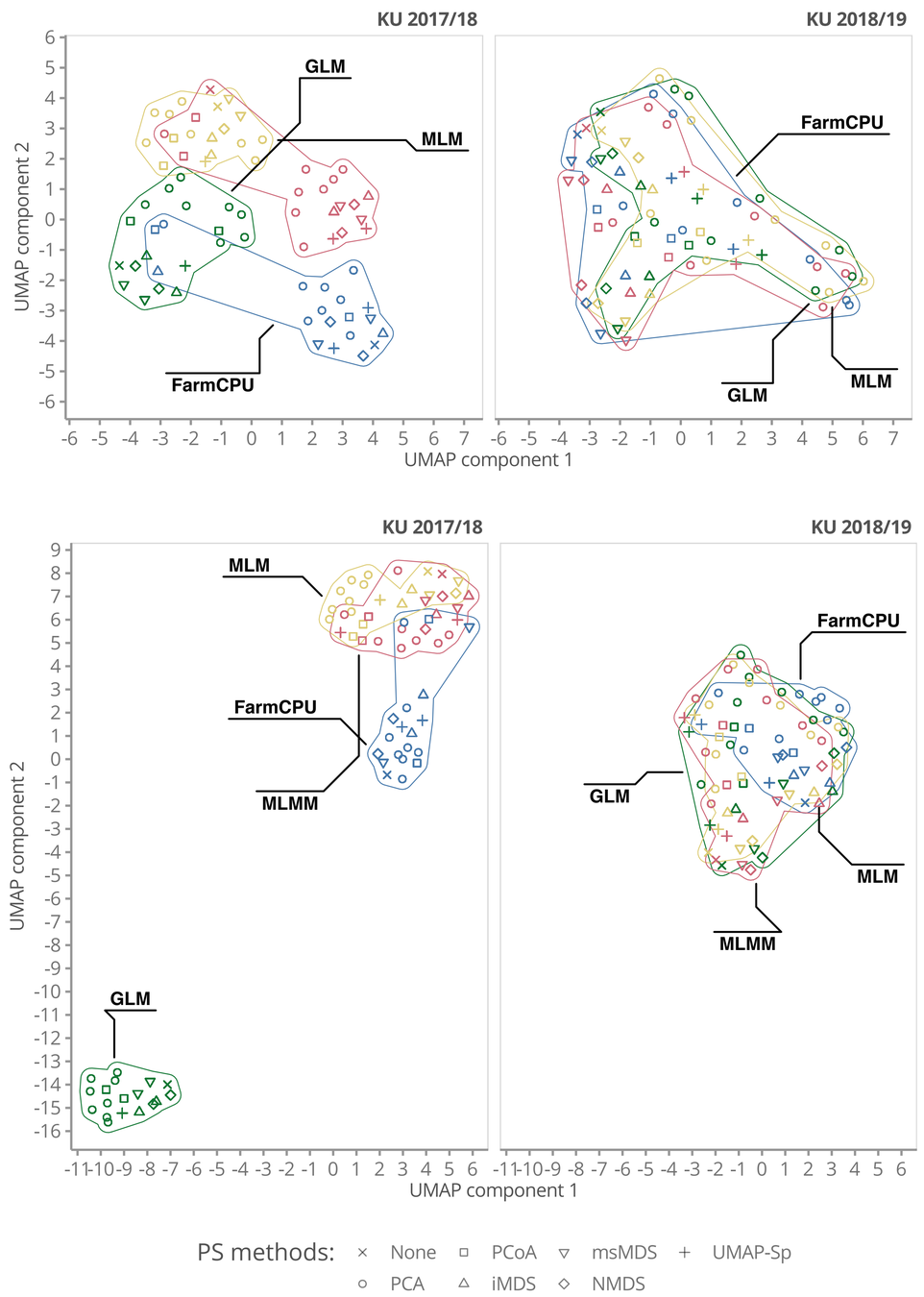

Figure 1: Comparison distances δ_r obtained from p-value (bottom) and negative logarithmic p-value (top) of GWAS models (GLM, MLM, MLMM, FarmCPU) using different population stratification methods (PCA, PCoA, [i/ms/o]MDS, UMAP) exemplarily for the trait plant height (KU 2017/18 and 2018/19).

References

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *J. Stat. Softw.*, *31*(3), 1–30.

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics*, *15*, 1–24.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv 1802.03426.

Wang, J., & Zhang, Z. (2019). GAPIT version 3: An interactive analytical tool for genomic association and prediction. Retrieved February 19, 2019, from https://github.com/jiabowang/GAPIT3

Scan the QR code to get more information about the project.

UNIKASSEL VERSITÄT

Julius Kühn-Institut

Projektträger Jülich Forschungszentrum Jülich

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung