Predicting $G \times E$ through phenomic prediction: A benchmark study across three species

Clément Bienvenu[1,2], Hugues de Verdal[1,2,†], Vincent Segura[2,3,†], and David Pot[1,2,†,*]

[1] *CIRAD, UMR AGAP Institut, Montpellier, F-34398, France*
[2] *UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, F-34398, France*
[3] *Geno-Vigne, IFV-INRAE-Institut Agro, F-34398, Montpellier, France*
[†] *Hugues de Verdal, David Pot and Vincent Segura should be considered joint senior authors*
[*] *Corresponding author: David Pot, email: david.pot@cirad.fr, tel: +33 6 51 75 13 76*

Theoretical and applied genetics

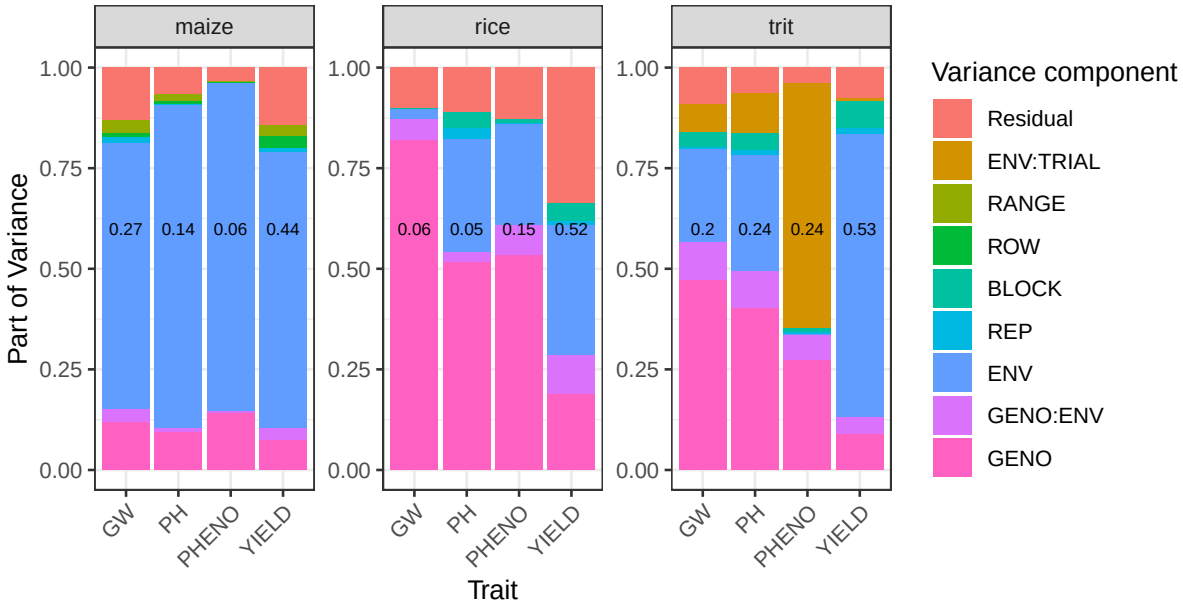


Figure S1: **Variance components for each trait in each dataset.** Black values indicate the ratio between GxE and G variances. For the triticale dataset (trit), all three populations were considered together in the model estimating the variance components. Different models were used for each dataset to take each specific experimental design into account (see material and methods section for details).

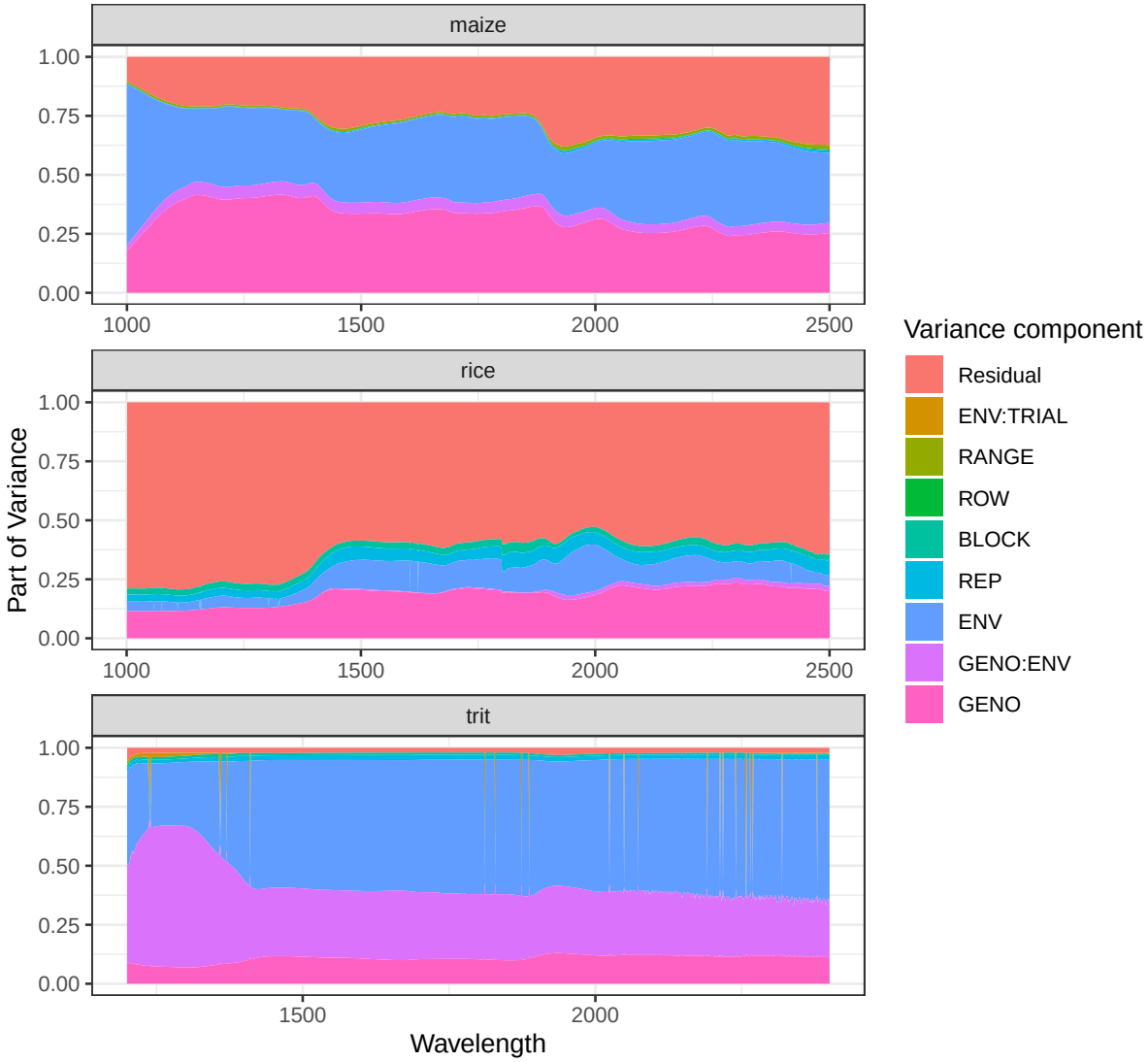


Figure S2: **Variance components for each wavelength in each dataset.** See material and methods section for details.

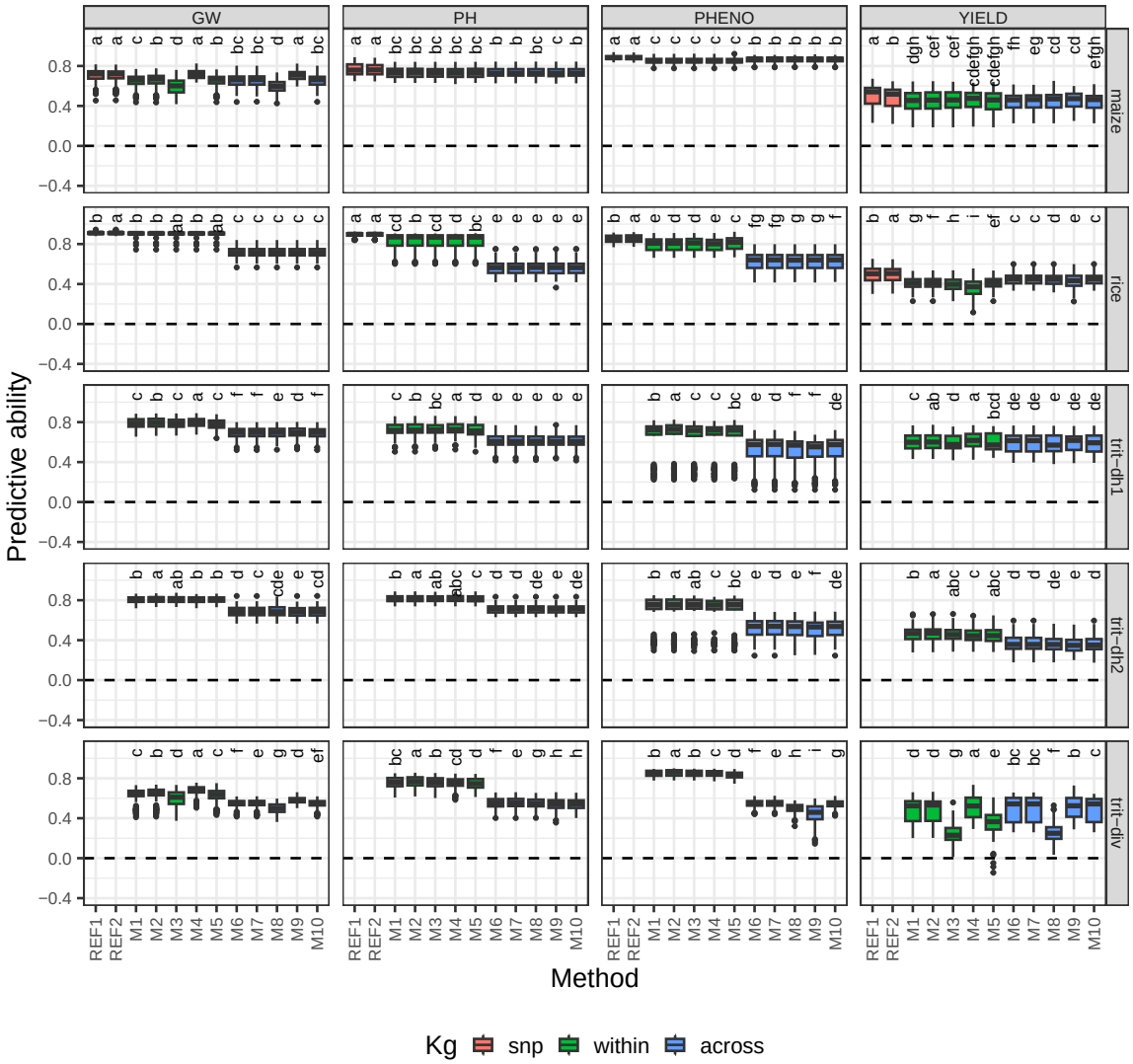


Figure S3: **Comparison between reference methods in the oGnE scenario.** Letters indicate groups of non-significantly different methods. GW = grain weight, PH = plant height, PHENO = phenology trait. snp = genomic prediction method, within = phenomic prediction with the main genetic relationship matrix being calculated as the cell by cell average of environment specific genetic relationship matrices which are calculated with within environment genetic BLUEs of spectra, across = phenomic prediction with the main genetic relationship matrix being calculated using across environment BLUEs of spectra.

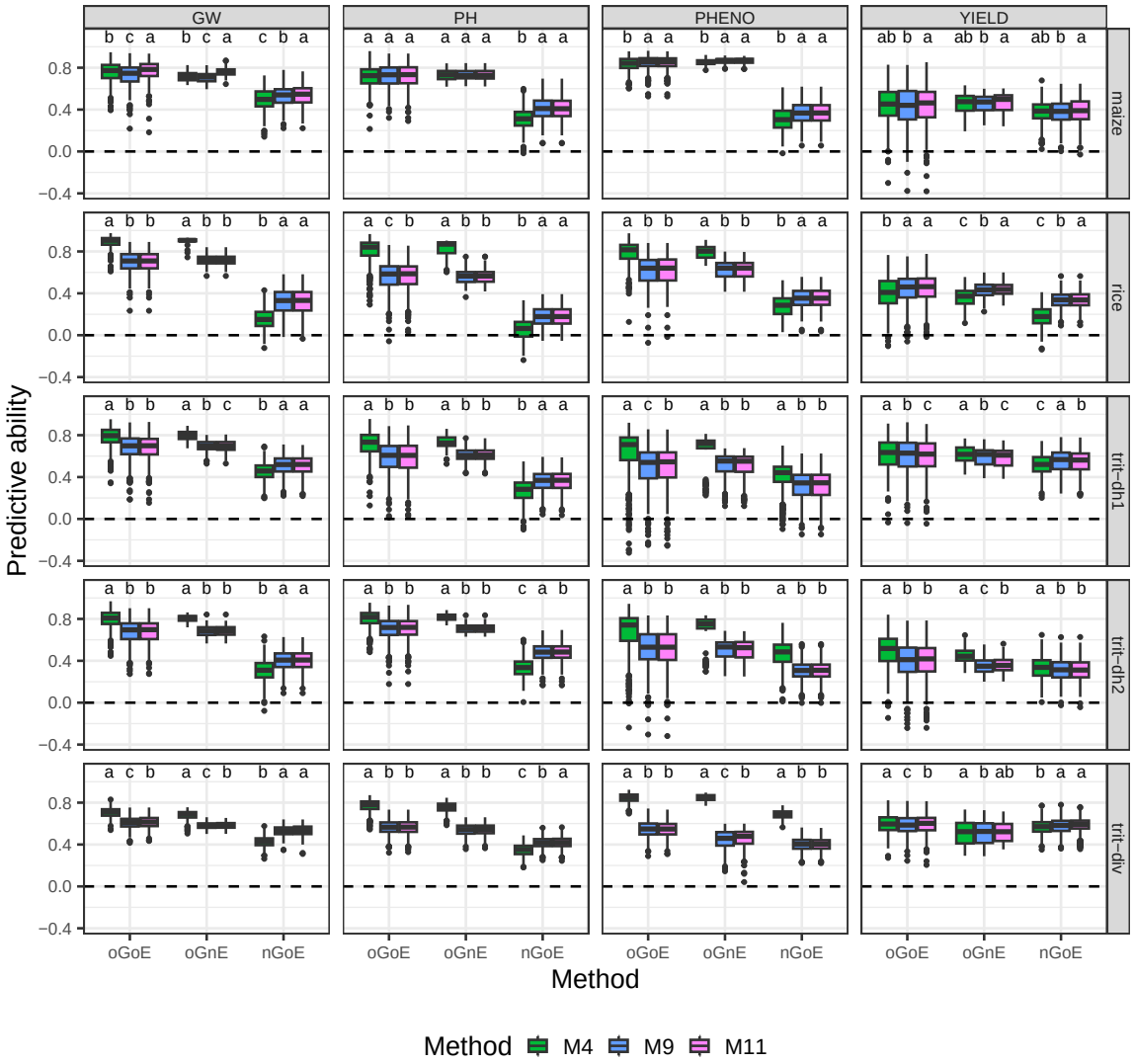


Figure S4: **Comparison between reference methods and the decomposition method.** Letters indicate groups of non-significantly different methods. GW = grain weight, PH = plant height, PHENO = phenology trait. Models M4 and M9 were the most successful models including a G and a $G \times E$ matrix found in PP literature, models M11 used a new way of computing the $G \times E$ matrix. See material and methods section for details.

Table S1: **Performances of methods M11 to M14 compared to M4 and M9 across datasets, cross validation scenario and traits.** Results arise from pairwise paired Wilcoxon tests between methods. Tests for M11 were done separately from test for M12 to M14. The "Method" column indicates the methods that is considered. The "Higher" and "Lower" columns indicate the number of times the method had a predictive ability significantly higher or significantly lower than both M4 and M9 methods. Column "No difference" indicate the number of times the method was not significantly different from M4 or M9, or had a predictive ability in between the ones of M4 and M9. M12 has less comparisons as this method produced models that did not converge in many occurrences.

Methods	Higher	No difference	Lower
M11	7 (11.6%)	51 (85%)	2 (3.3%)
M12	0	8 (20%)	32 (80%)
M13	1 (1.7%)	11 (18.3%)	48 (80%)
M14	0	3 (5%)	57 (95%)

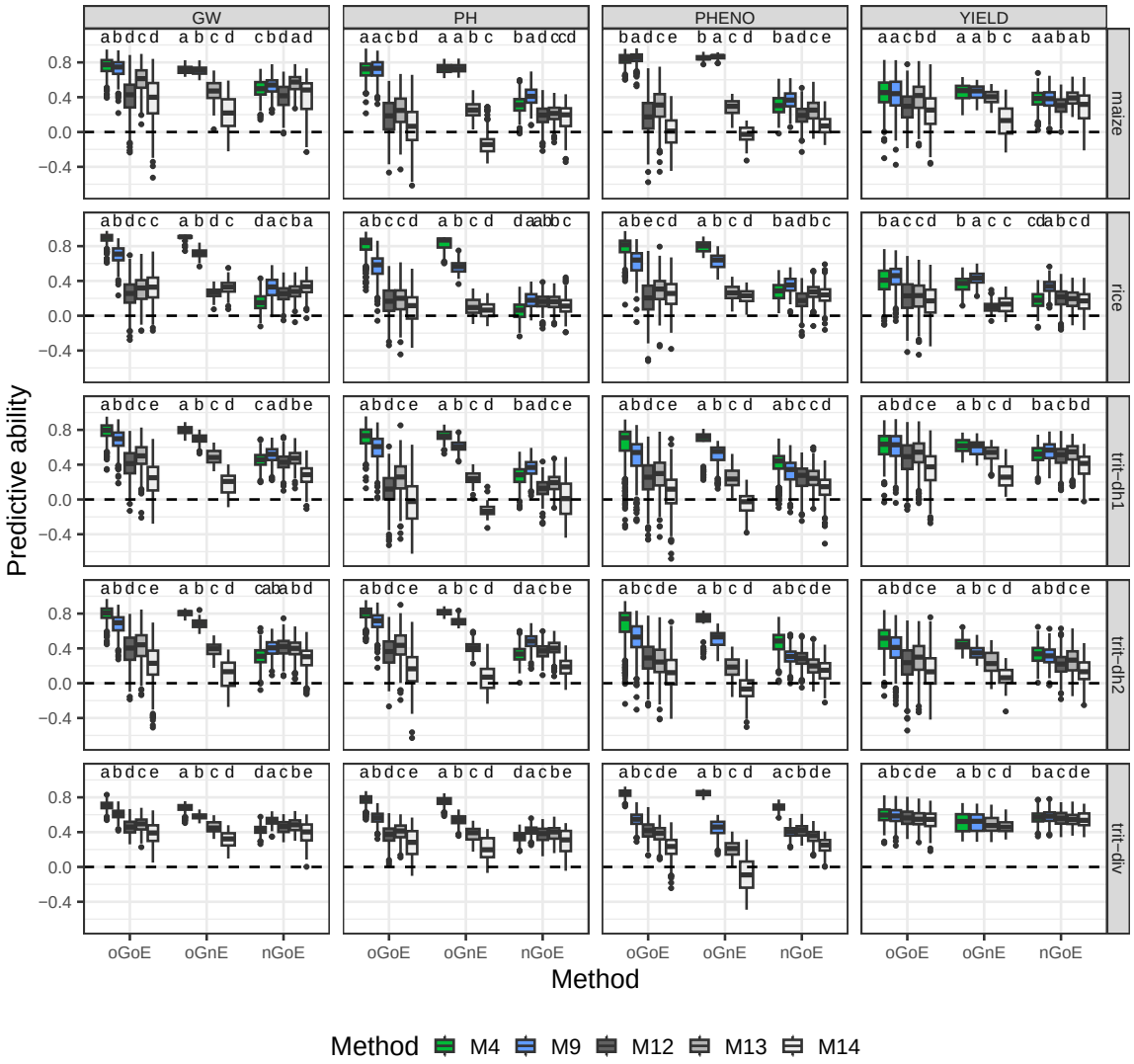


Figure S5: **Comparison between reference methods and methods with no genetic main effect.** Letters indicate groups of non-significantly different methods. GW = grain weight, PH = plant height, PHENO = phenology trait. Models M4 and M9 were the most successful models including a G and a $G \times E$ matrix found in PP literature, models M12 to M14 only have a $G \times E$ matrix.

Table S2: **Number of times methods M15 to M18 outperformed REF2 and have the best predictive ability across traits, datasets and cross validation scenarios.** Results arise from pairwise paired Wilcoxon tests between methods. The "Method" column indicates the methods that is considered. The "Best" column indicates the number of times the method had the best predictive ability in the cases when there were significant differences between methods (if multiple methods had the best PA and were not significantly different they were all granted a point). The "Higher", "No difference" and "Lower" columns indicate the number of times the method had a predictive ability significantly higher, not significantly different, or significantly lower compared to the method stated in the "Reference" column. Percentages are calculated within each method relatively to the total number of times each method was tested (which is 24).

Reference	Method	Best	Higher	No Difference	Lower
REF2	REF2	6 (25%)	\	\	\
	M15	3 (12.5%)	4 (16.7%)	16 (66.7%)	4 (16.7%)
	M16	8 (33.3%)	8 (33.3%)	10 (41.7%)	6 (25%)
	M17	14 (58.3%)	15 (62.5%)	7 (29.2%)	2 (8.3%)
	M18	9 (37.5%)	7 (29.2%)	13 (54.2%)	4 (16.7%)

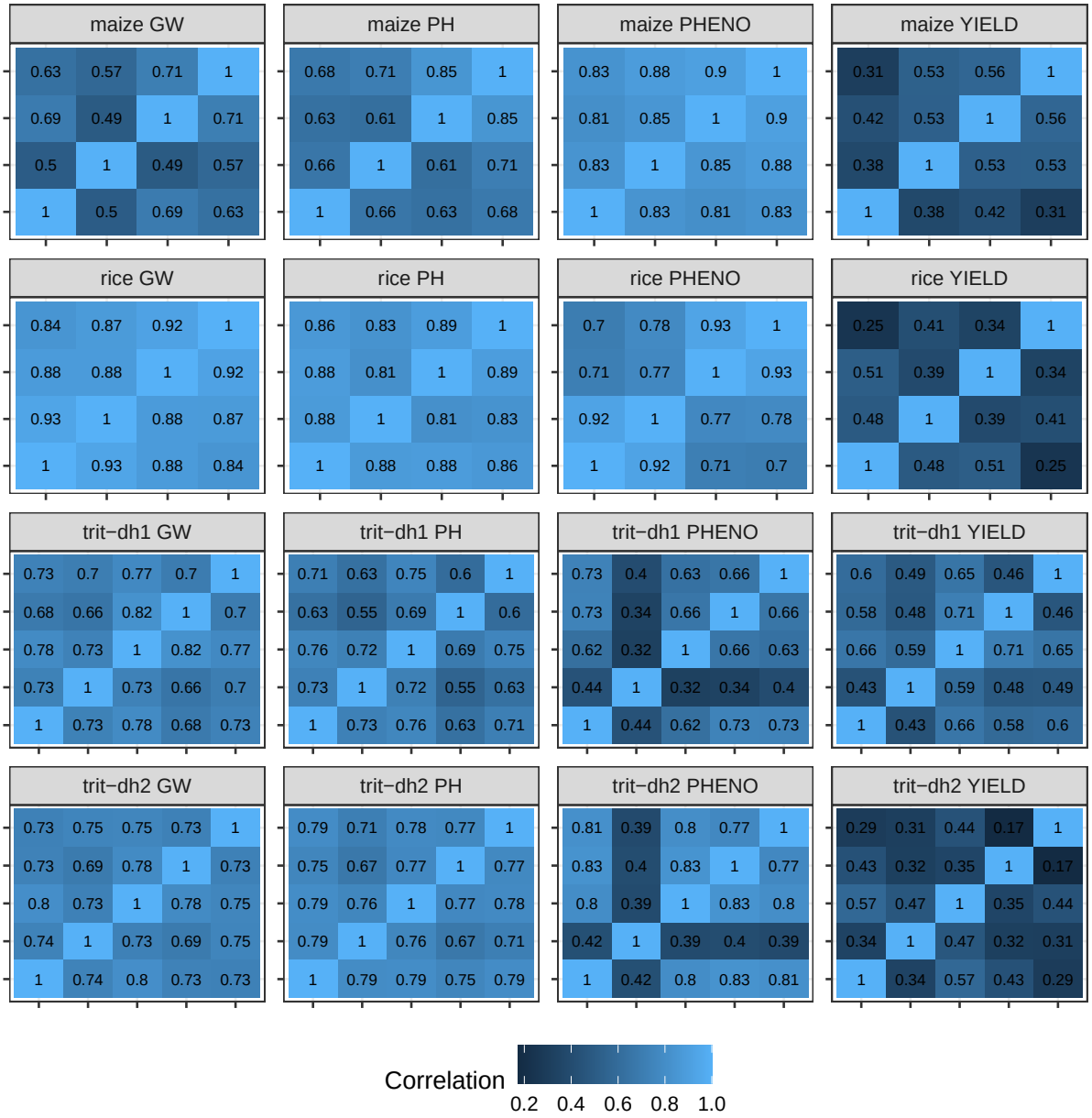


Figure S6: **Pearson correlations between BLUEs of the different environments for the different datasets and traits.** The trit-div population was omitted because the number of environments made it illegible.