Supplementary Material for 'Genomic mating in outbred species: predicting cross usefulness with additive and total genetic covariance matrices'

Marnin Wolfe, Ariel Chan, Peter Kulakow, Ismail Rabbi and Jean-Luc Jannink.

Appendix

Does the validation-data type (i.i.d. BLUPs vs. GBLUPs) make a difference? Most often, cross-validation done to test genomic prediction accuracy uses validation data (the stand-in for "truth") consisting of adjusted values, (e.g. BLUPs or BLUEs) for total individual performance, not including genomic relatedness information. In our study we set-up cross-validation folds that enable us to predict the GEBV and GETGV (GBLUPs) of validation family-members, and to subsequently compute their sample means, variances and usefulness. This approach has the added advantage of expanding the available sample size of validation progeny with complete data across traits. Nevertheless, we made some comparison to results using BLUPs that do not incorporate genomic relatedness information; in other words, independent and identically distributed (i.i.d.) BLUPs.

Prediction accuracy for family means were nearly uniformly higher using GBLUPs compared to iidBLUPs (median 0.18 higher). The Spearman rank correlation between prediction accuracies based on iidBLUPs and GBLUPs was high (median 0.75, range 0.55-0.84). Similar to the means, accuracy using GBLUP-validation-data appeared mostly higher compared to iidBLUPs (median difference GBLUPs-iidBLUPs = 0.07, interquartile range -0.002-0.14). The Spearman rank correlations of iidBLUP and GBLUP-validation-based accuracies was positive for family (co)variances, but smaller compared to family means (mean correlation 0.5, range 0.04-0.89). Supplementary plots comparing validation-data accuracies for means and (co)variances were inspected (Figure S6-S7). Based on this, we conclude that we would reach similar though more muted conclusions about which trait variances and trait-trait covariances are best or worst predicted, if restricted to iidBLUPs for validation data.

What if we consider only families with greater than a threshold size? In our primary analysis, we computed (co)variance prediction accuracies with weighted correlations, considering any family with more than one member. We also considered a more conservative alternative approach of including only families with $\geq$ 10 (n=112); we thought beyond that was too stringent as at $\geq$ 20 only 22 families remain. The Spearman rank correlation between accuracy estimates when all vs. only families with more than 10 members was 0.89. There should therefore be good concordance with our primary conclusions, depending on the family size threshold we impose. The median difference in accuracy ("threshold size families" minus "all families' ') was 0.01. Considering only size 10 or greater families noticeably improved prediction accuracy for several trait variances and especially for two covariances (DM-TCHART and logFYLD-MCMDS) (Figure S8).

Comparing posterior mean variance (PMV) to variance of posterior mean (VPM) predictions: Variances and covariances were predicted with the computationally intensive PMV method. Population variance estimates based on PMV were consistently larger than VPM, but the correlation of those estimates is 0.98 (Figure S9). Using the predictions from the cross-validation results, we further observed that the PMV predictions were consistently larger and most notably that the correlation between PMV and VPM was very high (0.995). Some VPM prediction accuracies actually appear better than PMV predictions (Figure S10).

The critical point is that VPM and PMV predictions should have very similar rankings. In our primary analysis, we focus on the PMV results with the only exception being the exploratory predictions where we saved time/computation and used the VPM. If implementing mate selections via the usefulness criteria, choosing the VPM method would mostly have the consequence of shrinking the influence on selection decisions towards the mean.

Comparing the directional dominance to the "classic" model: Our focus in this article was not in finding the optimal or most accurate prediction model for obtaining marker effects. However, genome-wide estimates of directional dominance have not previously been made in cassava. For this reason, we make some brief comparison to the standard or "classic" additive-dominance prediction model, where dominance effects are centered on zero. Overall, the ranking of models and predictions between the two models were similar, as indicated by a rank correlation between model accuracy estimates of 0.98 for family means and 0.94 for variances and covariances. Three-quarters of family-mean and almost half of (co)variance accuracy estimates were higher using the directional dominance model. The most notably improved predictions were for the family-mean logFYLD TGV (Figure S11-S12). There was also an overall rank correlation of 0.98 between models in the prediction of untested crosses.

Supplementary Tables

Most Supplementary Tables are included as worksheets in the file **SupplementaryTables.xlsx**. Very large ones are included as separate CSV files.

Table S1: Selection indices. For each trait, the standard deviation of BLUPs (blupSD), which were divided by "unscaled" index weights for the StdSI and BiofortSI indices to get StdSI and BiofortSI weights used throughout the study.

Table S2: Summary of cross-validation scheme. For each fold of each Rep, the number of parents in the test-set (Ntestparents) is given along with the number of clones in the corresponding training (Ntraintset) and testing (Ntestset) datasets and the number of crosses to predict (NcrossesToPredict).

Table S3: Test-parents. For each fold of each cross-validation repeat, the set of parents whose crosses are to be predicted is listed.

Table S4: Training-Testing partitions of germplasm. For each fold of each repeat, the genotype ID (germplasmName) of all clones in the "trainset" and "testset" are given.

Table S5: Crosses to predict each fold. For each fold of each repeat, the sireID and damID are given for each cross-to-be-predicted.

Table S6: Predicted and observed cross means. For each fold of each repeat, each cross distinguished by a unique pair of sireID and damID is given. The genetic model used (Models A, AD, DirDomAD, DirDomBV), whether the prediction is of mean breeding value (predOf=MeanBV) or mean total genetic value (predOf=MeanTGV), the trait (BiofortSI or StdSI), type of observation (ValidationData: GBLUPs or iidBLUPs) and corresponding prediction (predMean) and observations (obsMean) are shown.

Table S7: Predicted cross variances. All predictions of cross-variance from the cross-validation scheme are detailed. For each fold of each repeat and each unique cross (sireID x damID). Both variances (Trait1==Trait2) and co-variances (Trait1!=Trait2) are given. The genetic model used (Model: A, AD, DirDomAD, DirDomBV), the variance component being predicted (VarComp=VarA or VarD), along with the number of segregating SNPs in the family (Nsegsnps) and the time taken in seconds for computation, per family (totcomputetime) are given. The predictions based on the variance of posterior means (VPM) and the posterior mean variances (PMV) are both shown.

Table S8: Predicted versus observed cross variances. From the cross-validation analysis. For each fold of each repeat, each cross distinguished by a unique pair of sireID and damID is given. The genetic model used (Model: A, AD, DirDomAD, DirDomBV), whether the prediction is of family variance in breeding value (predOf=VarBV) or variance in total genetic value (predOf=VarTGV), the trait (BiofortSI or StdSI), type of observation (ValidationData: GBLUPs or iidBLUPs) and corresponding prediction (predVar)

and observations (obsVar) are shown. The predictions are based on either only the variance of posterior means (VarMethod=VPM) or the posterior mean variances (VarMethod=PMV). The family size (number of genotyped offspring, FamSize) or number of offspring with direct phenotypes (Nobs) are used to weight the correlation (CorrWeight) between observed and predicted family variances.

Table S9: Predicted versus observed UC. For each fold of each repeat, each cross distinguished by a unique pair of sireID and damID is given. The predicted usefulness criterion (predUC) was computed as the predMean + realIntensity*predSD, where predMean is the predicted family mean and predSD is the predicted genetic standard deviation. The genetic model used (Model: A, AD, DirDomAD, DirDomBV), whether the prediction is of family variance in breeding value (predOf=VarBV) or variance in total genetic value (predOf=VarTGV), the trait (BiofortSI or StdSI) and corresponding prediction (predUC) and observations (obsUC) are shown. The family size (number of genotyped offspring, FamSize) is shown along with the realized selection intensity (realIntensity) for each selection stage in the breeding pipeline (Parent, CET, PYT, AYT, UYT) and also a constant intensity value (Stage=ConstIntensity).

Table S10: Accuracies predicting the mean. For each fold of each repeat, the accuracy predicting family means (Accuracy) is given. The genetic model used (Model: A, AD, DirDomAD, DirDomBV), whether the prediction is of mean breeding value (predOf=MeanBV) or mean total genetic value (predOf=MeanTGV), the trait (BiofortSI or StdSI), type of observation (ValidationData: GBLUPs or iidBLUPs) are shown.

Table S11: Accuracy of predicting the variances. For each fold of each repeat the estimated accuracy of predicting family variances is given. Accuracy was computed the correlation between predicted and observed variance, either weighted by family size (AccuracyWtCor) or not (AccuracyCor). The genetic model used (Model: A, AD, DirDomAD, DirDomBV), whether the prediction is of family variance in breeding value (predOf=VarBV) or variance in total genetic value (predOf=VarTGV), the trait (BiofortSI or StdSI), type of observation (ValidationData: GBLUPs or iidBLUPs) are shown. The predictions are based on either only the variance of posterior means (VarMethod=VPM) or the posterior mean variances (VarMethod=PMV).

Table S12: Accuracy predicting the usefulness criteria. For each fold of each repeat the estimated accuracy of predicting family usefulness criteria is given. Accuracy was computed as the correlation between predicted UC and observed UC (mean of selected offspring), either weighted by family size (AccuracyWtCor) or not (AccuracyCor). The genetic model used (Model: A, AD, DirDomAD, DirDomBV), whether the prediction is of UC in breeding value (predOf=VarBV) or UC in total genetic value (predOf=VarTGV), the trait (BiofortSI or StdSI), type of observation (ValidationData: GBLUPs or iidBLUPs) are shown. The predictions of cross variance used to compute the UC are based on either only the variance of posterior means (VarMethod=VPM) or the posterior mean variances (VarMethod=PMV).

Table S13: Realized within-cross selection metrics. Table summarizing measurements made of selection within each cross (unique sireID-damID). Summaries included: family size (FamSize), number (NmembersUsedAsParent) and proportion of members used as parents (propUsedAsParent), mean GEBV and GETGV of top 1% of each family (meanTop1pctGEBV, meanTop1pctGETGV), for each selection index Trait (Trait: BiofortSI, StdSI), proportion of each family that has been phenotyped (propPhenotyped, NmembersPhenotyped) and past each stage of the breeding pipeline (propPast and NmembersPast CET, PYT, AYT) and finally the corresponding realized intensity of selection for each stage (e.g. realIntensityAYT).

Table S14: Genome-wide proportion of SNPs that are homozygous, for each clone (GID=germplasmName).

Table S15: Variance-covariance estimates for each genetic group. Summary of the population-level genetic variance estimates in each genetic group (GG=C0, TMS13=C1, TMS14=C2, TMS15=C3), for each genetic model (Model: A, AD, DirDomA, DirDomAD), each variance (Trait1==Trait2) and covariance (Trait1!=Trait2). The estimates are computed both based on the variance of posterior means (VarMethod=VPM) and the posterior mean variances (VarMethod=PMV).

Table S16: Directional dominance effects estimates. Based on the directional dominance model, the genome-wide posterior mean (InbreedingEffect) and posterior standard deviation (InbreedingEffectSD) inbreeding effect is given. Estimates are provided for each trait, genetic group (Group), and each repeat-fold of the cross-validation study.

Table S17: Predictions of untested crosses. Compiled predictions of 47,083 possible crosses (sireID x damID) of 306 parents. Predictions were made with two additive-dominance genetic models: either with (Model=DirDomAD) or without (Model=ClassicAD) a directional dominance term. The predictions included are the cross mean (predMeanBV,predMeanGV), standard deviation (predSdBV,predSdGV) and usefulness (predUCparent,predUCvariety) in terms of breeding (BV) and total genetic (GV) value. Additional information provided for each cross include: whether the cross is a self (IsSelf=T/F), has previously been made (CrossPrevMade=Yes/No), the number of segregating SNPs expected in the family (Nsegsnps) and the parental GEBV (sireGEBV, damGEBV).

Table S18: Long-form table of predictions about untested crosses. Compiled predictions of 47,083 possible crosses (sireID x damID) of 306 parents. Predictions were made with two additive-dominance genetic models: either with (Model=DirDomAD) or without (Model=ClassicAD) a directional dominance term. The predictions (Pred) included are of the cross mean (PredOf=Mean), standard deviation (PredOf=Sd) and usefulness (PredOf=UC) in terms of breeding (Component=BV) and total genetic (Component=GV) value. Additional information provided for each cross include: whether the cross is a self (IsSelf=T/F) and has previously been made (CrossPrevMade=Yes/No).

Table S19: Top 50 crosses (sireID x damID) selected by each of 16 predictions of 47,083 crosses. Predictions for each trait were made with two additive-dominance genetic models: either with (Model=DirDomAD) or without (Model=ClassicAD) a directional dominance term. The predictions (Pred) selected on are of the cross mean (PredOf=Mean), standard deviation (PredOf=Sd) and usefulness (PredOf=UC) in terms of breeding (Component=BV) and total genetic (Component=GV) value. Additional information provided for each cross include: whether the cross is a self (IsSelf=T/F) and has previously been made (CrossPrevMade=Yes/No).

Supplementary Figures (Main)

Figure S01: Genome-wide proportion homozygous

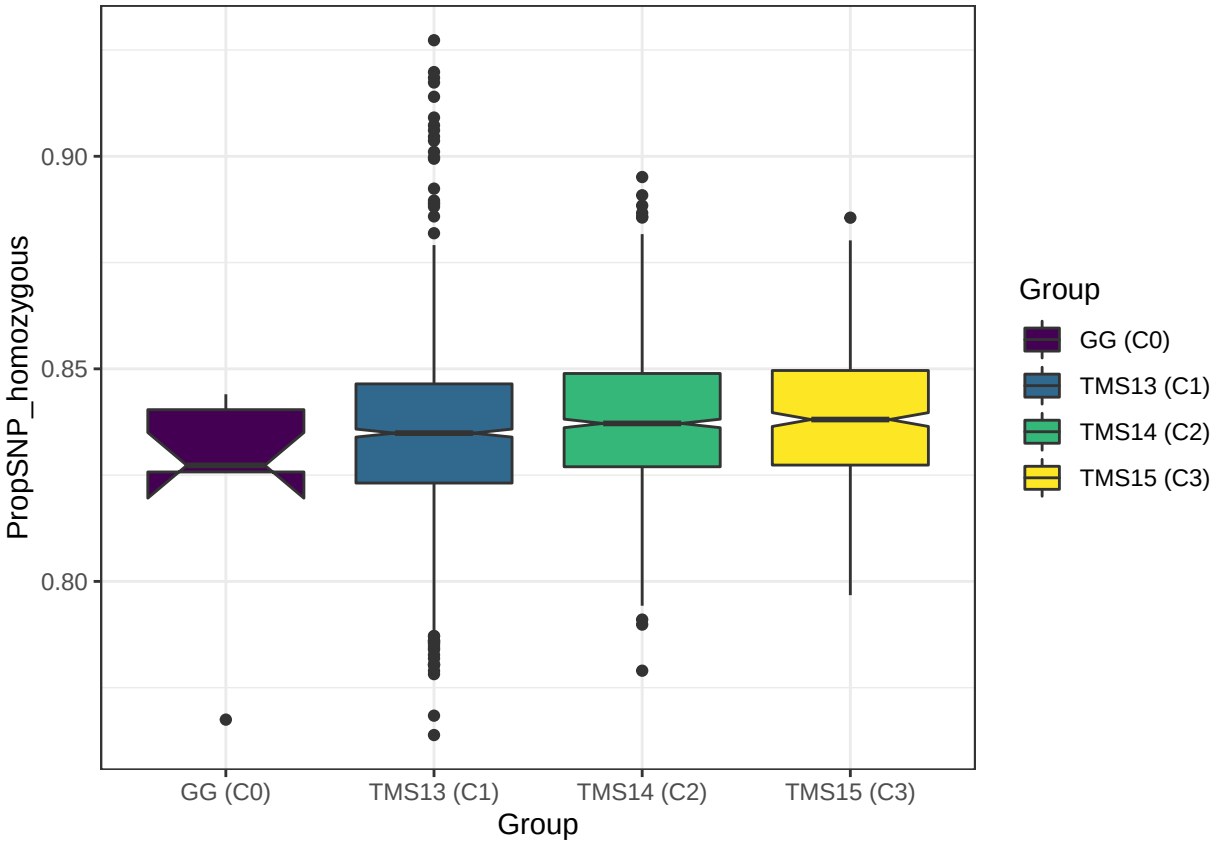

Figure S01: Boxplot of the genome-wide proportion of homozygous SNPs in each of four genetic groups comprising the study pedigree.

Figure S02: Correlations among phenotypic BLUPs (including Selection Indices)

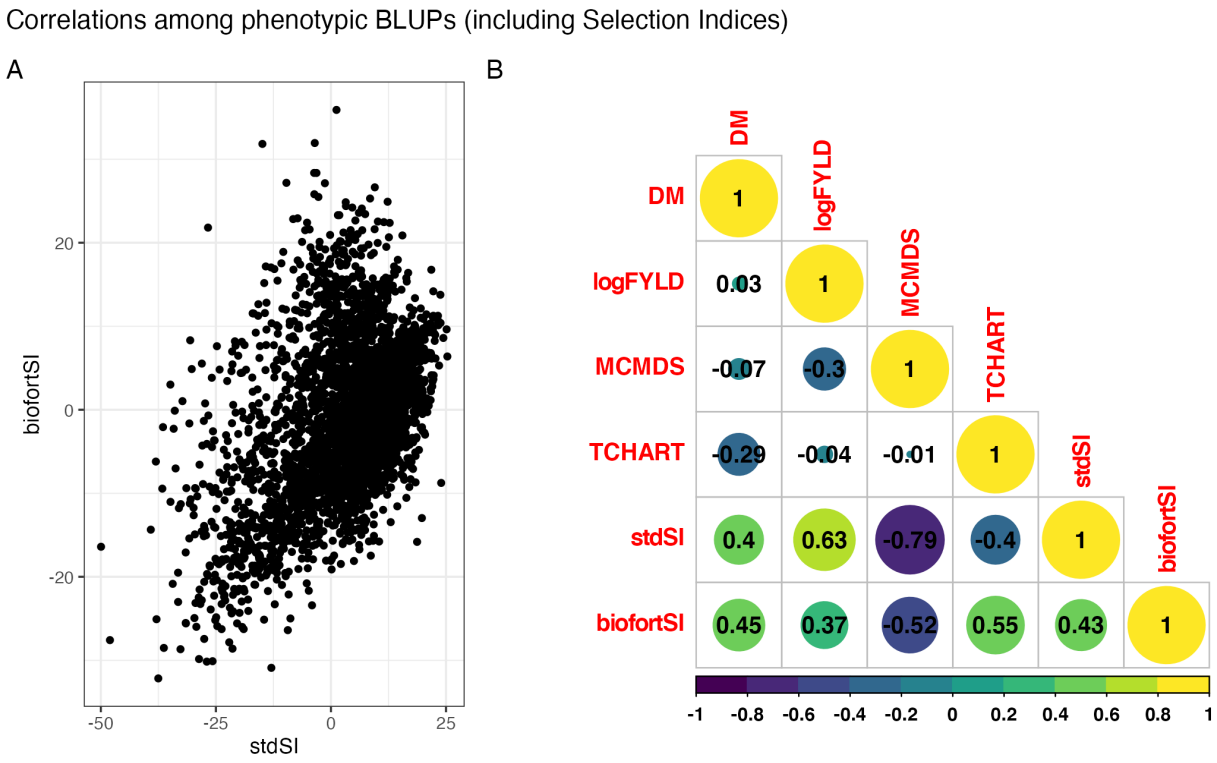

Figure S02: Correlations among BLUPs (including Selection Indices). (A) StdSI vs. BiofortSI computed from i.i.d. BLUPs. (B) Heatmap of the correlation among BLUPs for each of four component traits and two derived selection indices.

Figure S03: Realized selection intensities: measuring post-cross selection

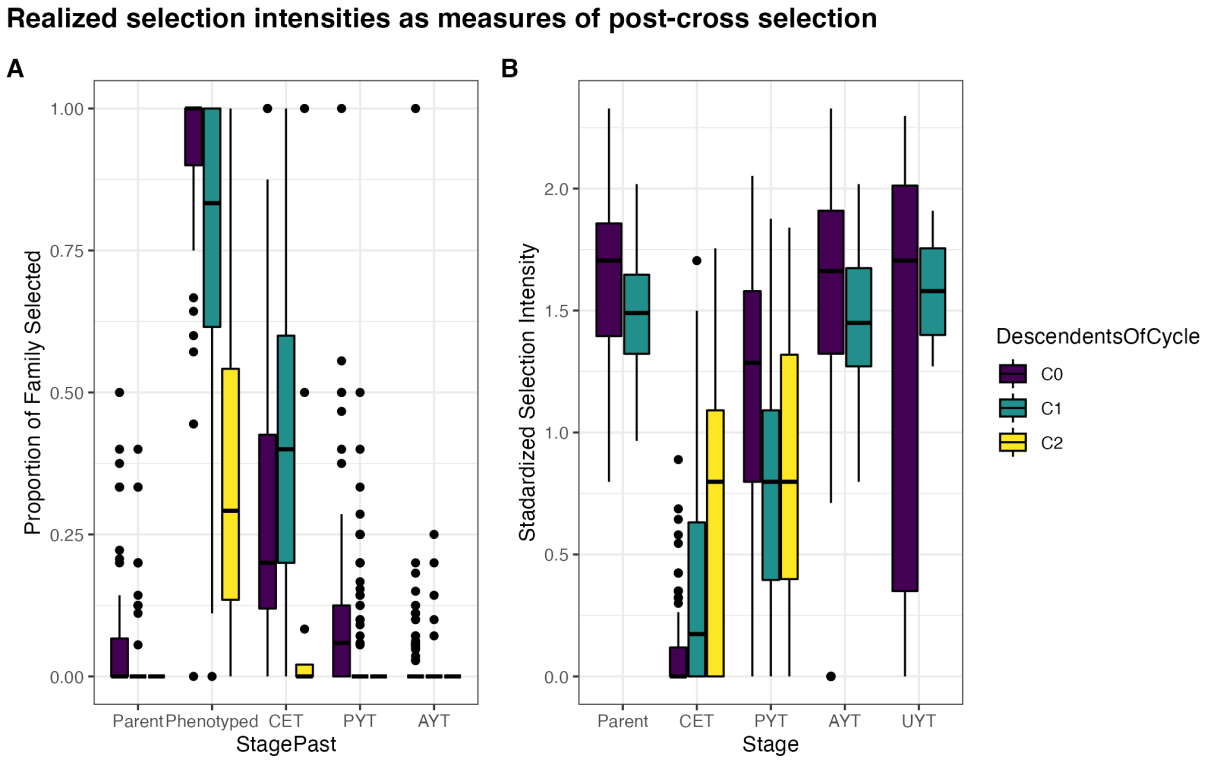

Figure S03: Realized selection intensities: measuring post-cross selection. Boxplots showing (A) the proportion of each family selected and (B) the standardized selection intensity for each stage of the breeding pipeline, in each genetic group.

Figure S04: Correlation matrix for predictions on the StdSI

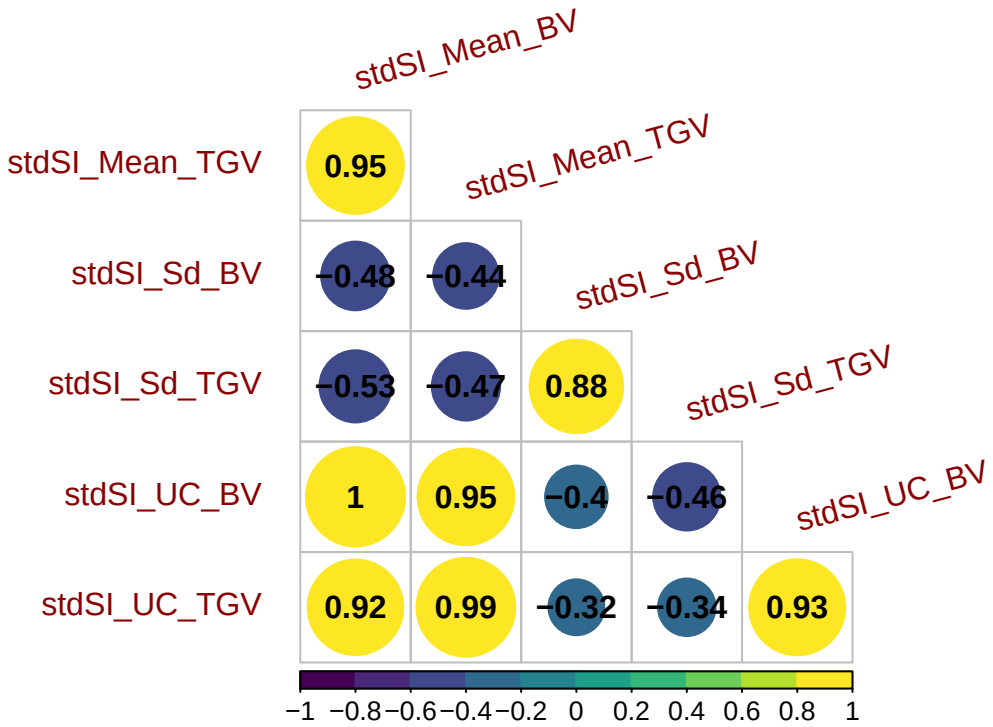

Figure S04: Correlation matrix for predictions on the StdSI. Heatmap of the correlations between predictions of mean, standard deviation, and usefulness in terms of BV and TGV. Predictions were made for 47,083 possible pairwise crosses of 306 parents with a directional dominance model.

Figure S05: Correlation matrix for predictions on the BiofortSI

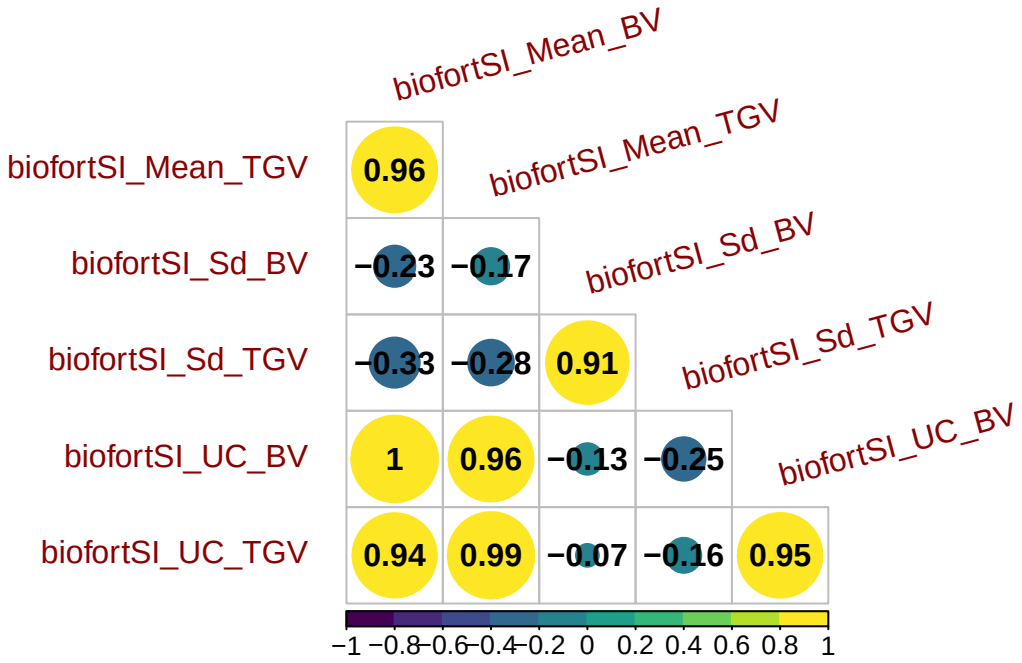

Figure S05: Correlation matrix for predictions on the BiofortSI. Heatmap of the correlations between predictions of mean, standard deviation, and usefulness in terms of BV and TGV. Predictions were made for

47,083 possible pairwise crosses of 306 parents with a directional dominance model.

Supplementary Figures (Appendix)

Figure S06: Contrasting GBLUPs and iidBLUPs as validation data for measuring family mean prediction accuracy

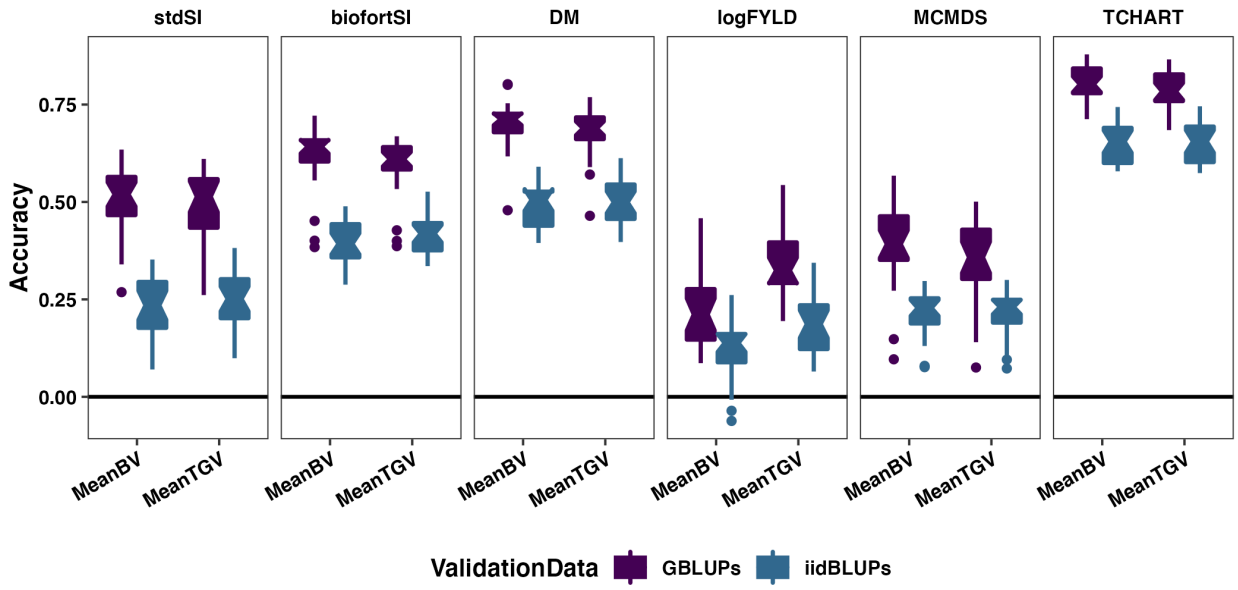

Figure S06: Contrasting GBLUPs and iidBLUPs as validation data for measuring family mean prediction accuracy. The cross mean prediction accuracy based on fivefold parent-wise cross-validation is shown using boxplots. Each panel contains results for one of the selection indices (stdSI and biofortSI) and for the component traits (DM, logFYLD, MCMDS, TCHART). Prediction accuracies are on the y-axis and cross mean GEBV (MeanBV) and GETGV (MeanTGV) are on the x-axis. Colors distinguish the validation data used to estimate the accuracy.

Figure S07: Contrasting GBLUPs and iidBLUPs as validation data for measuring family *co)variance prediction accuracy

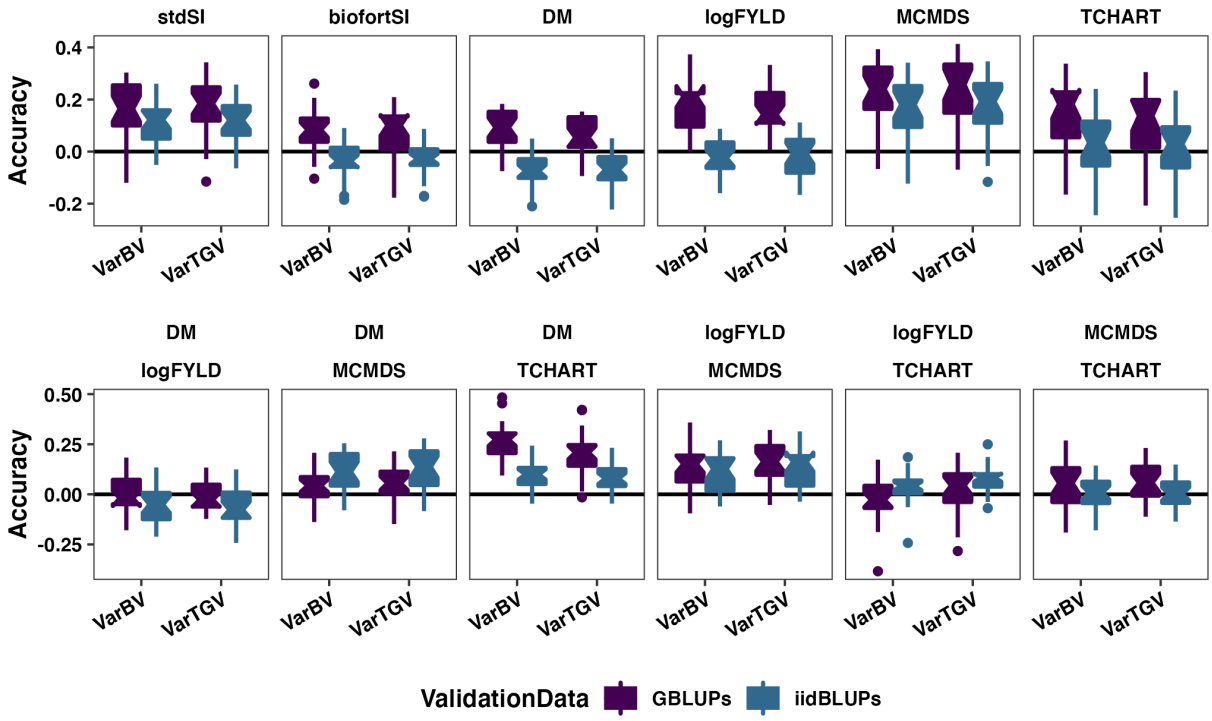

Figure S07: Contrasting GBLUPs and iidBLUPs as validation data for measuring family (co)variance prediction accuracy. The cross variance and covariance prediction accuracy based on fivefold parent-wise cross-validation is shown using boxplots. Each panel in the top row contains results for either a selection index (stdSI and biofortSI) or a component trait (DM, logFYLD, MCMDS, TCHART) variance. Each panel on the bottom row contains one of the six pairwise covariances between the four component traits. Prediction accuracies are on the y-axis and cross variance/covariance for GEBV (VarBV) and GETGV (VarTGV) are on the x-axis. Colors distinguish the validation data used to estimate the accuracy.

Figure S08: Variance-covariance Accuracy considering only families with 10+ members?

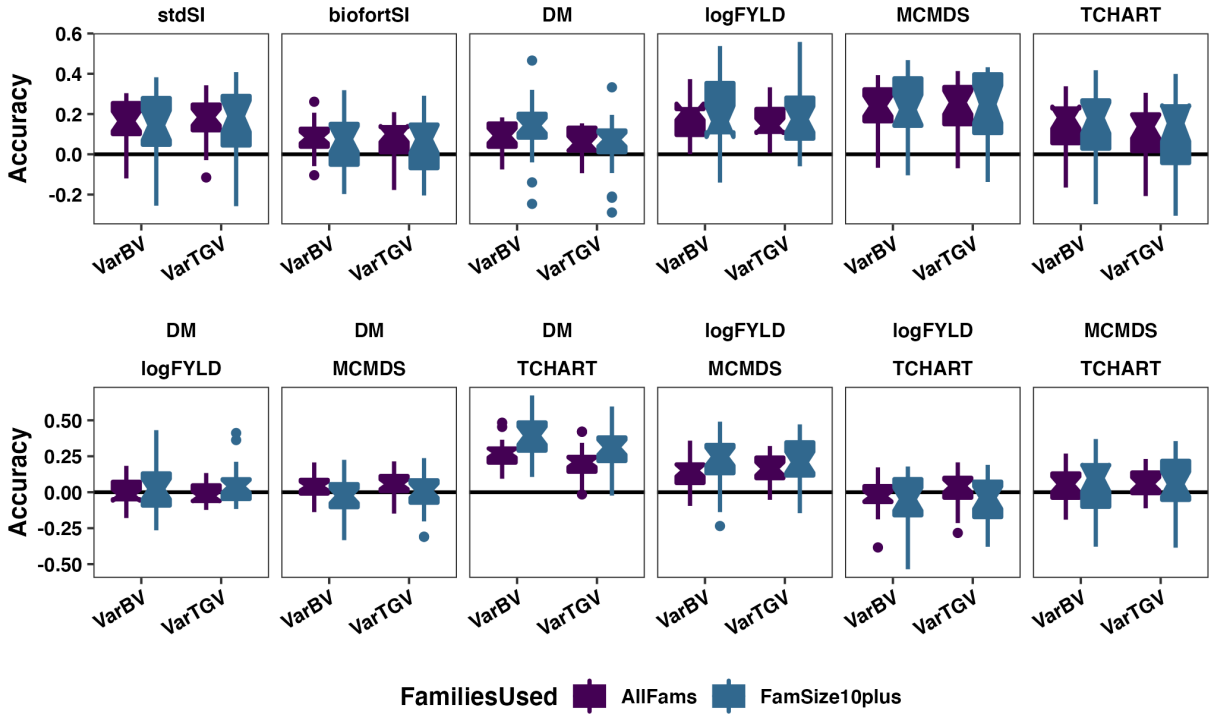

Figure S08: Variance-covariance Accuracy considering only families with 10+ members? The cross variance and covariance prediction accuracy based on fivefold parent-wise cross-validation is shown using boxplots. Results are shown based on the directional dominance model. Each panel in the top row contains results for either a selection index (stdSI and biofortSI) or a component trait (DM, logFYLD, MCMDS, TCHART) variance. Each panel on the bottom row contains one of the six pairwise covariances between the four component traits. Prediction accuracies are on the y-axis and cross variance/covariance for GEBV (VarBV) and GETGV (VarTGV) are on the x-axis. Colors distinguish whether all families (AllFams) or just the ones with at least 10 members were included in the accuracy estimates.

Figure S09: Population estimates of genetic variance parameters - PMV vs. VPM

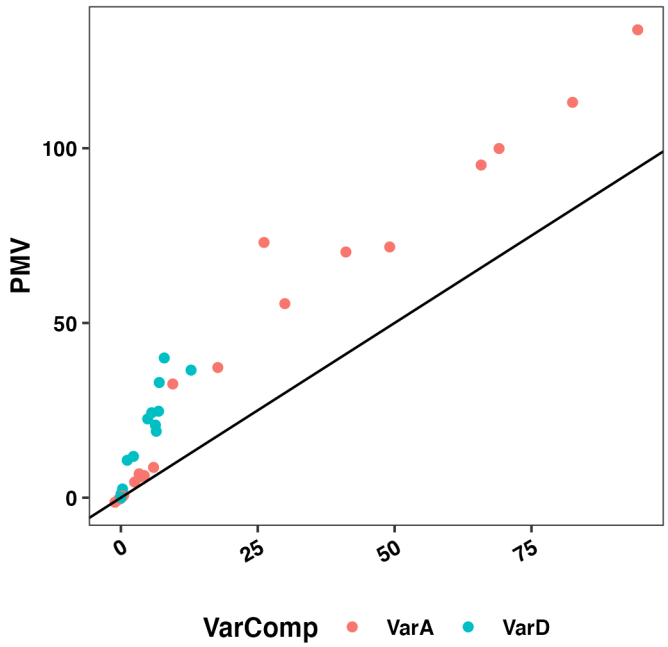

Figure S09: Population estimates of genetic variance parameters - PMV vs. VPM. We contrasted the variance of posterior means (VPM; x-axis) to the less biased, more intensive-to-compute posterior mean variance (PMV; y-axis). Each point is a trait variance or trait-trait covariance estimate from the directional dominance model. Colors distinguish additive variance (VarA) and dominance variance(VarD).

Figure S10: PMV vs. VPM - Compare prediction accuracy

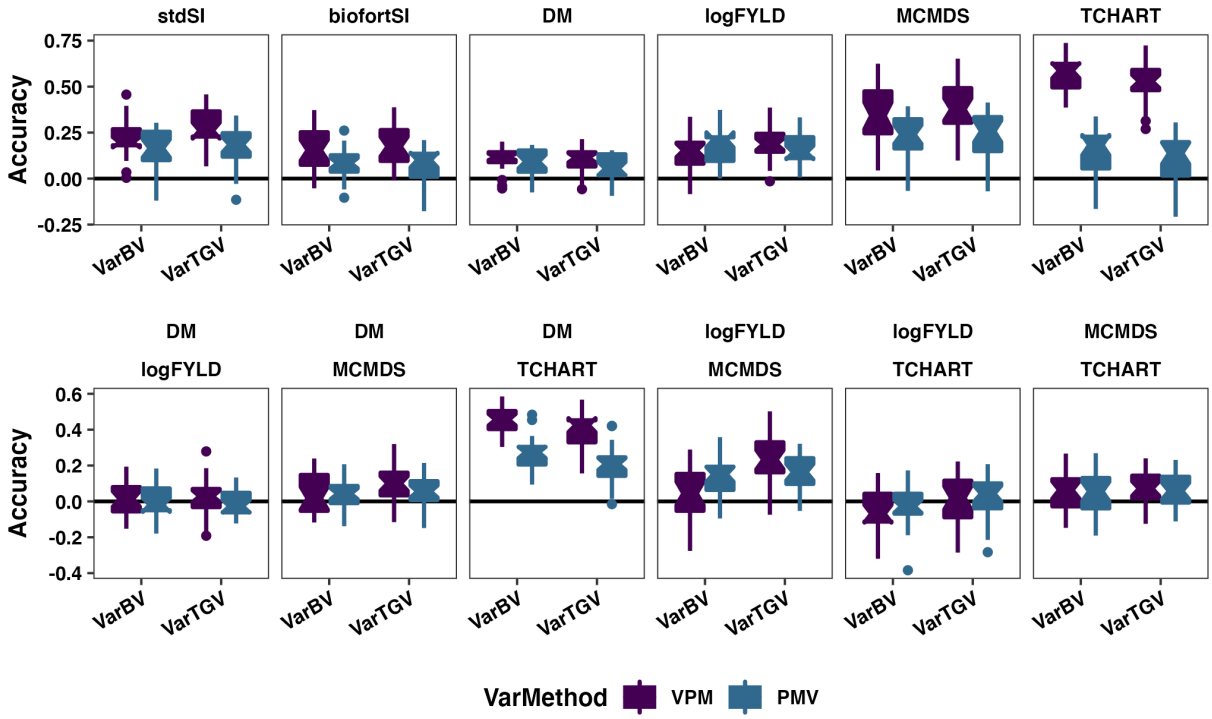

Figure S10: Variance-covariance Accuracy comparing VPM vs. PMV. The cross variance and covariance prediction accuracy based on fivefold parent-wise cross-validation is shown using boxplots. Results are shown based on the directional dominance model. Each panel in the top row contains results for either a selection index (stdSI and biofortSI) or a component trait (DM, logFYLD, MCMDS, TCHART) variance. Each panel on the bottom row contains one of the six pairwise covariances between the four component traits. Prediction accuracies are on the y-axis and cross variance/covariance for GEBV (VarBV) and GETGV (VarTGV) are on the x-axis. Colors distinguish whether predictions were based on the VPM or the PMV.

Figure S11: Directional Dominance vs. Classic Model - Family Mean Prediction Accuracy

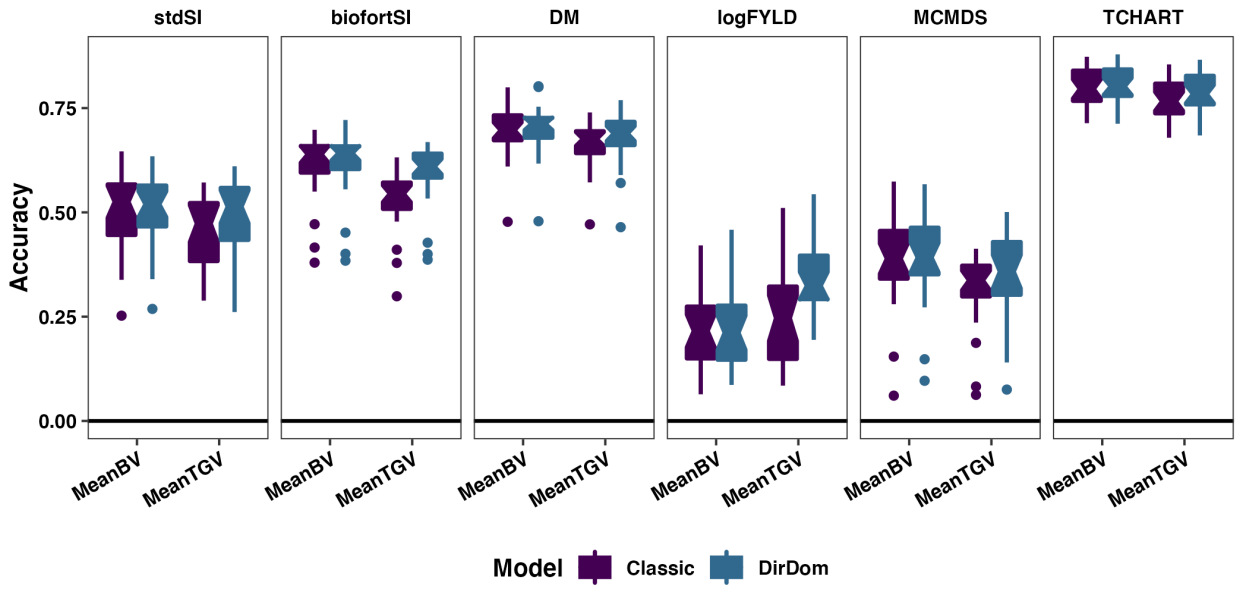

Figure S11: Contrasting directional and non-directional dominance models accuracy predicting family means. The cross mean prediction accuracy based on fivefold parent-wise cross-validation is shown using boxplots. Each panel contains results for one of the selection indices (stdSI and biofortSI) and for the component traits (DM, logFYLD, MCMDS, TCHART). Prediction accuracies are on the y-axis and cross mean GEBV (MeanBV) and GETGV (MeanTGV) are on the x-axis. Colors distinguish whether results are based on the directional dominance (DirDom) model or not (Classic).

Figure S12: Directional Dominance vs. Classic Model - Family (Co)variance Prediction Accuracy

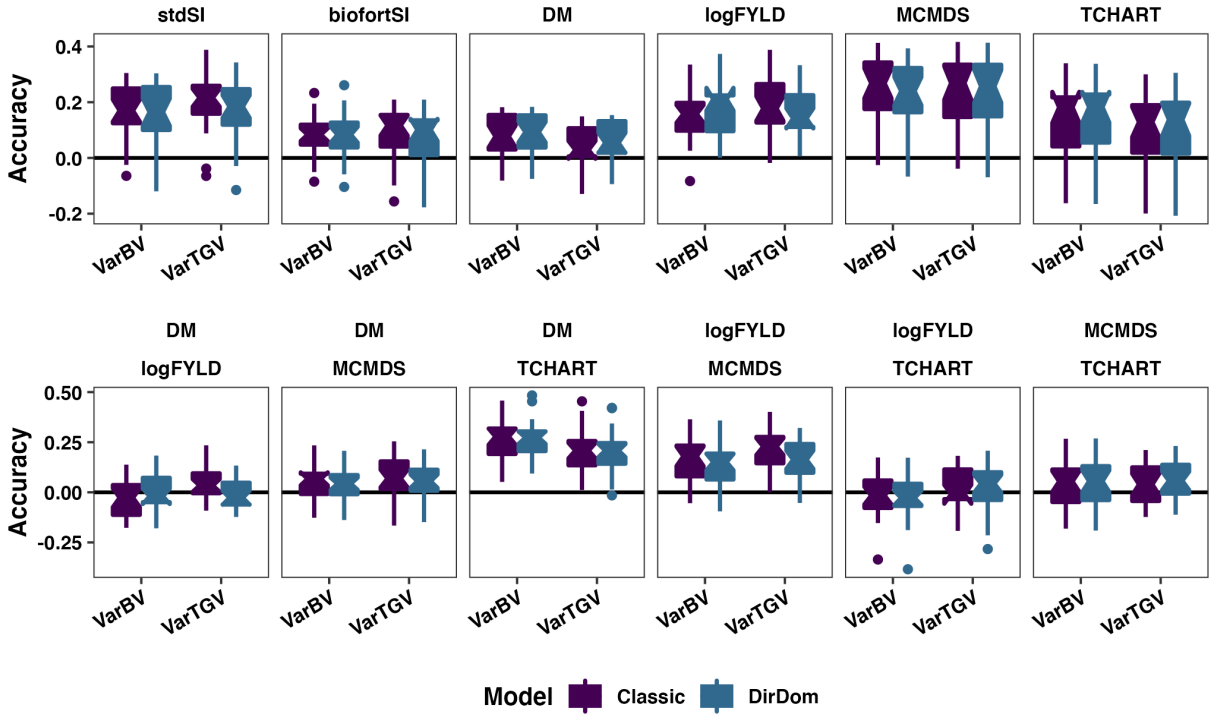

Figure S12: Contrasting directional and non-directional dominance models accuracy predicting family (co)variances. The cross variance and covariance prediction accuracy based on fivefold parent-wise cross-validation is shown using boxplots. Results are shown based on the directional dominance model. Each panel in the top row contains results for either a selection index (stdSI and biofortSI) or a component trait (DM, logFYLD, MCMDS, TCHART) variance. Each panel on the bottom row contains one of the six pairwise covariances between the four component traits. Prediction accuracies are on the y-axis and cross variance/covariance for GEBV (VarBV) and GETGV (VarTGV) are on the x-axis. Colors distinguish whether results are based on the directional dominance (DirDom) model or not (Classic).