Core ideas

- Maize grain contains tocochromanols, essential micronutrients in the human diet as vitamin E
- Genomic prediction of tocochromanols can enhance breeding for biofortification
- Exotic germplasm can enhance genetic diversity but is challenging to predict
- Prediction accuracy is modest within populations, but can be low across populations
- Optimal training population design facilitates prediction of tocochromanols

# Genomic prediction of tocochromanols in exotic-derived maize

Laura E. Tibbs-Cortes[1], Tingting Guo[2,3], Xianran Li[4], Ryokei Tanaka[5], Adam E. Vanous[6], David Peters[6], Candice Gardner[6], Maria Magallanes-Lundback[7], Nicholas T. Deason[7], Dean DellaPenna[7], Michael A. Gore[5], Jianming Yu[1]

[1]Department of Agronomy, Iowa State University, Ames, IA; [2]Hubei Hongshan Laboratory, Wuhan, China; [3]Huazhong Agricultural University, Wuhan, China; [4]USDA-ARS, Pullman, WA; [5]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY; [6]USDA-ARS, Ames, IA; [7]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI

## ABBREVIATIONS

αT, α-tocopherol; αT3, α-tocotrienol; AP, Ames Diversity Panel; BL, Bayesian Lasso; BRR, Bayesian ridge regression; BLUE, best linear unbiased estimate; CV, cross-validation; δT, δ-tocopherol; δT3, δ-tocotrienol; DH, doubled haploid; ExPVP, expired Plant Variety Protection; FURS, fast and unique representative subset selection; γT, γ-tocopherol; γT3, γ-tocotrienol; GBS, genotyping by sequencing; GEM, germplasm enhancement of maize; GP, genomic prediction; HPLC, high-performance liquid chromatography; IBS, identity-by-state; MaxCD,

maximization of connectedness and diversity; MSE, mean square error; NSS, non-Stiff-Stalk; OTP, optimal training population; PAM, partitioning around medoids; PC, principal component; ΣT, total tocopherols; ΣT3, total tocotrienols; ΣTT3, total tocochromanols; SS, Stiff-Stalk.


## ABSTRACT

Tocochromanols (vitamin E) are an essential part of the human diet. Plant products including maize grain are the major dietary source of tocochromanols; therefore, breeding maize with higher vitamin content (biofortification) could improve human nutrition. Incorporating exotic germplasm in maize breeding for trait improvement including biofortification is a promising approach and an important research topic. However, information about genomic prediction of exotic-derived lines using available training data from adapted germplasm is limited. In this study, genomic prediction was systematically investigated for nine tocochromanol traits within both an adapted (Ames Diversity Panel) and an exotic-derived (BGEM) maize population. While prediction accuracies up to 0.79 were achieved using gBLUP when predicting within each population, genomic prediction of BGEM based on an Ames Diversity Panel training set resulted in low prediction accuracies. Optimal training population (OTP) design methods FURS, MaxCD, and PAM were adapted for inbreds and, along with the methods CDmean and PEVmean, often improved prediction accuracies compared to random training sets of the same size. When applied to the combined population, OPT designs enabled successful prediction of the rest of the exotic-derived population. Our findings highlight the importance of leveraging genotype data in training set design to efficiently incorporate new exotic germplasm into a plant breeding program.

# 1. INTRODUCTION

Vitamin E is an essential nutrient in the human diet. The term vitamin E collectively refers to a total of eight different fat-soluble molecules, called tocochromanols. The more common and dietarily active group of tocochromanols are the tocopherols, which have a saturated tail, followed by the tocotrienols, which have a tail containing three unconjugated double bonds. Based on the degree and position of methylation, both tocopherols and tocotrienols are divided into α, β, γ, and δ species, with α-tocopherol (αT) having the most and δ-tocotrienol (δT3) having the least vitamin E activity (DellaPenna & Mène-Saffrané, 2011; Lipka et al., 2013).

This vitamin is critical for maintaining the integrity of cell membranes and enabling healthy erythrocyte and nervous function, as well as providing important antioxidant activity. While clinical vitamin E deficiency is rare today, 79% of individuals in a global survey had below the recommended desirable blood serum levels for this vitamin, potentially resulting in chronic health consequences including increased risk of Alzheimer's and cardiovascular disease (Péter et al., 2015). Because tocochromanols can only be synthesized by plants and other photosynthetic organisms, plant products are the major dietary source of vitamin E (DellaPenna & Mène-Saffrané, 2011). These include the staple crop maize (Chander, Guo, Yang, Yan, et al., 2008), raising the possibility of improving human nutrition by breeding maize with higher vitamin content. This approach is a type of biofortification, a process that aims to increase the bioavailable micronutrient content of food crops (Rosell, 2016). Previous work has shown substantial genetic variation for tocochromanol content exists in maize (Chander, Guo, Yang, Yan, et al., 2008; Chander, Guo, Yang, Zhang, et al., 2008; Diepenbrock et al., 2017; Li et al., 2012; Lipka et al., 2013; Shutu et al., 2012; Weber, 1987; Wong, Lambert, Tadmor, & Rocheford, 2003), providing the necessary foundation for biofortification.

The previously studied maize germplasm represents a small subset of global maize diversity, raising the possibility of identifying additional favorable alleles within exotic maize. The Germplasm Enhancement of Maize (GEM) project is an important source of exotic germplasm adapted to long day environments and has released a collection of doubled haploid (DH) lines known as the BGEM panel. The two recurrent parents of this population represent the two major heterotic pools commonly used in maize breeding, with PHZ51 representing the non-Stiff-Stalk (NSS) group and PHB47 the Stiff-Stalk (SS) group, while the exotic parents include accessions from Argentina, Bolivia, Brazil, Chile, Colombia, Cuba, Ecuador, Guatemala, Martinique, Mexico, Paraguay, Peru, and Venezuela (Vanous et al., 2018).

Genomic prediction, in which statistical models trained on genotyped and phenotyped individuals (the training set) are used to predict phenotypes of other individuals that have only been genotyped (the validation set) by leveraging shared genetic information, is critical to modern plant breeding (Crossa et al., 2017). Genomic prediction enables breeders to save both the time and money that would otherwise be required to phenotype all lines of interest and has been applied to molecular (Yu et al., 2020), agronomic (Dzievit, Guo, Li, & Yu, 2021), and nutritional (Owens et al., 2014) traits in maize. Notably, genomic prediction has been successfully applied to maize grain tocochromanol traits within a diverse panel of sweet corn (Baseggio et al., 2019; Hershberger et al., 2022).

Many methods for genomic prediction have been created, each of which has its own assumptions about trait architecture that may make it more or less suitable for a particular situation (Habier, Fernando, Kizilkaya, & Garrick, 2011; Wang et al., 2018). However, genomic prediction's requirement of shared genetic information means that a training set should ideally cover the potential genetic space of the validation set. When the training and validation sets are

not closely related, as in the case of predicting exotic-derived from adapted germplasm, differences in alleles present in the populations and their linkage disequilibrium with markers will lead to extrapolation. This extrapolation of the genomic prediction model beyond the genetic space in which it was trained may result in reduced prediction accuracy (Crossa et al., 2017; Dzievit et al., 2021; Yu et al., 2016).

Optimal training population (OTP) design aims to improve genomic prediction accuracy by identifying an OTP of a given size that best covers the available genetic space. Some methods can take into account the genetic information of the proposed validation set to identify a training set best suited to its prediction, potentially improving prediction accuracies across populations (Akdemir, 2018; Akdemir, Sanchez, & Jannink, 2015). These methods use a genetic algorithm to identify a training set that minimizes mean prediction error variance (PEVmean) or maximizes mean coefficient of determination (CDmean) in the validation set (Akdemir, 2018; Akdemir et al., 2015). PEVmean has previously been used to improve accuracy when predicting tropical maize from publicly available training data (Pinho Morais et al., 2020). Based on data mining techniques and genetic design, three new OTP methods were developed and examined for hybrid performance prediction: fast and unique representative subset selection (FURS), maximization of connectedness and diversity (MaxCD), and partitioning around medoids (PAM) (Guo et al., 2019). FURS is based on graphic network analysis. In this method, representative nodes from a graph derived from the genetic correlation matrix are selected as the training set. MaxCD is a method based on the population's mating scheme, selecting a set of hybrids with non-overlapping parents followed by additional hybrids from pairs of inbreds most distantly related to one another. Finally, PAM is based on a clustering algorithm in which the individuals are grouped into the desired number of clusters using their genetic covariance matrix; then, the

medoid of each cluster forms the training set (Guo et al., 2019). The methods FURS, MaxCD, and PAM have not previously been examined in inbreds.

While it is desirable to incorporate exotic germplasm in breeding and specifically biofortification efforts, doing so is a challenge. In a breeding program, phenotyping the entire selection population would be expensive in both time and money, particularly for vitamin traits measured by high-performance liquid chromatography (HPLC), while genomic prediction using models established with existing data from the adapted germplasm would require extrapolation. Therefore, we investigated this dilemma through a systematic examination of genomic prediction of grain tocochromanol traits in both the Ames Diversity Panel (AP, a diverse panel of adapted inbreds) and BGEM (a panel of exotic-derived DH lines) (Fig. 1) (Dzievit et al., 2021; Gianola, 2021; Yu et al., 2016). In this study, we first demonstrated the accuracy of genomic prediction for nine grain tocochromanol traits within AP and BGEM individually and the challenge of using adapted inbreds in AP to predict exotic-derived lines in BGEM. Next, we validated the OTP design methods FURS, MaxCD, and PAM in inbreds and compared with PEVmean and CDmean. Finally, we applied the OTP design methods to the combined AP-BGEM data to identify training sets that could generate accurate predictions of tocochromanols for the exotic-derived germplasm.

## 2. MATERIALS AND METHODS

### 2.1 Experimental Material

Experimental materials were from the Ames Diversity Panel (AP) and BGEM populations. AP consists of 2,815 diverse maize inbreds sampled from breeding programs around the world (Romay et al., 2013). BGEM consists of 252 DH lines created by crossing 54 exotic maize accessions representing 52 exotic maize races with the temperate-adapted expired Plant Variety

Protection (ExPVP) lines PHZ51 and/or PHB47 to create $F_1$ plants. Each $F_1$ was backcrossed to its ExPVP parent, producing 71 unique $BC_1F_1$ populations, which were crossed to a haploid inducer to create haploid plants. These haploid plants were self-pollinated to create the 252 BGEM lines (Vanous et al., 2018).

In 2015 and 2017, 1,815 maize inbreds from AP that are adapted to the U.S. Corn Belt (able to flower and set seed in Iowa) were grown in Boone, Iowa in a randomized augmented block design. In 2018, a subset of 1,023 of these inbreds were again grown in Boone, Iowa in a randomized augmented block design. Based on this data, the AP lines grown in 2015 and 2017 but not 2018 were designated the AP training set, while remaining AP lines (grown in 2015, 2017, and 2018) were designated the AP validation set. This enabled prediction of the validation set in the same environment as the training set (2015 and 2017) as well as in a different environment (2018). A single replication of 225 and 224 BGEM lines were grown in 2016 and 2018, respectively, in Ames, Iowa, for a total of 236 distinct BGEM lines. The two recurrent parents of the BGEM population (PHZ51 and PHB47) were also grown in both of those years.

**2.2 Phenotype data**

Each year, mature grain was harvested and tocochromanol traits were measured in the ground kernels by HPLC and fluorometry as previously described (Lipka et al., 2013). Six tocochromanol traits were measured as μg/g of dry seed: α-tocopherol (αT), δ-tocopherol (δT), γ-tocopherol (γT), α-tocotrienol (αT3), δ-tocotrienol (δT3), and γ-tocotrienol (γT3). From these, three additional traits were calculated: total tocopherols (αT + δT + γT, denoted ΣT), total tocotrienols (αT3 + δT3 + γT3, denoted ΣT3), and total tocochromanols (ΣT + ΣT3, denoted ΣTT3).

Mature grain was successfully harvested and measured for tocochromanol traits from 215 and 202 of the BGEM lines grown in 2016 and 2018, respectively, as well as the two recurrent parents. After excluding sweet corn and popcorn lines, which have unique kernel structures that distort metabolite measurements, 1,444 AP lines were measured in 2015, 1,436 in 2017, and 888 in 2018. Phenotypic data was processed separately for each of four data sets (2015 and 2017 AP training set, 2015 and 2017 AP validation set, 2018 AP validation set, and 2016 and 2018 BGEM). For each tocochromanol trait, best linear unbiased estimates (BLUEs) were calculated by fitting the following mixed linear model was fit using the lme4 package in R:

$$Y_{ijklmn} = genotype_i + check_i + year_j + group \times year_{ij} + genotype \times year_{ij} + tier(year)_{jk} + pass(tier \times year)_{jkl} + range(tier \times year)_{jkm} + plate(year)_{jn} + \varepsilon_{ijklmn}$$

In this model, $Y_{ijklmn}$ is a single phenotypic observation; $genotype_i$ is the fixed effect of the $i^{th}$ genotype, set to zero for check genotypes; $check_i$ is the fixed effect of the check, set to zero when the $i^{th}$ genotype is not a check; $year_j$ is the random effect of the $j^{th}$ year; group is an indicator variable indicating whether the $i^{th}$ genotype is check or non-check; group × $year_{ij}$ is the random interaction term between the group of the $i^{th}$ genotype and the $j^{th}$ year; $tier(year)_{jk}$ is the random effect of the $k^{th}$ tier within the $j^{th}$ year; $pass(tier \times year)_{jkl}$ and $range(tier \times year)_{jkm}$ are the random effects of the $l^{th}$ pass (field column) and the $m^{th}$ range (field row), respectively, within the $k^{th}$ field tier within the $j^{th}$ year; $plate(year)_{jn}$ is the random effect of the $n^{th}$ HPLC autosampler plate used for tocochromanol measurements in the $j^{th}$ year; and $\varepsilon_{ijklmn}$ is the residual term, assumed to be normally distributed with mean of 0 and variance of $\sigma_\varepsilon^2$.

Studentized residuals were calculated using these models, and one round of outlier removal was conducted with a Bonferroni-corrected 0.05 threshold. After removing these

outliers, the models were fitted again to calculate BLUEs, which were used in subsequent analyses (Table S1).

### 2.3 Genotype data

For AP, imputed genotyping by sequencing (GBS) data were obtained from Panzea (file name ZeaGBSv27_publicSamples_imputedV5_AGPv4-181023.vcf). For accessions with more than one entry in this data set, pairwise identity-by-state (IBS) was calculated. If mean IBS within a given accession was less than 95%, that accession was dropped. For the remaining repeated accessions, a consensus sequence was generated for each accession using a custom R script. This GBS data was then filtered as described in (Wu et al., 2022), leaving 257,995 total SNPs for further analyses.

For BGEM, GBS data containing 370,630 SNPs was used (Vanous et al., 2018). These unimputed data contained only biallelic SNPs, and were in AGPv2 coordinates, so they were first uplifted to v4 coordinates. The GBS data were filtered to exclude SNPs with missing data rates above 80% or minor allele frequency below 0.5%, then imputed in Beagle 5.0 using default settings and no map. Finally, SNPs with minor allele frequency below 1% were excluded, leaving 164,530 total SNPs for further analyses.

After filtering, 257,995 and 164,530 SNPs were present in the AP and BGEM genotype data, respectively, and were used for within-population predictions. Of these, 68,444 SNPs were present in both data sets, and were used for cross-population predictions. Lines with phenotype data but no genotype data were removed from the analysis, creating a final data set of 607 lines in the AP testing set, 855 in the AP validation set, and 201 in BGEM.

## 2.4 Genomic prediction

Genomic prediction was conducted using eight methods: gBLUP, sBLUP, and cBLUP (Wang et al., 2018) implemented in GAPIT (Wang & Zhang, 2020); and BayesA, BayesB (Meuwissen, Hayes, & Goddard, 2001), BayesC (Kizilkaya, Fernando, & Garrick, 2010), Bayesian Ridge Regression (BRR) (Meuwissen et al., 2001), and Bayesian Lasso (BL) (de los Campos et al., 2009) implemented in BGLR (Pérez & de los Campos, 2014).

The available data enabled the following prediction scenarios (Fig. 1A):

I. AP CV: Ten-fold cross-validation (CV) within the AP training set.

II. BGEM CV: Ten-fold CV within BGEM.

III. AP, common environment: The AP training set used to predict the AP validation set grown in the same environment (2015 and 2017).

IV. AP, new environment: The AP training set used to predict the AP validation set genotypes grown in a different environment (2018).

V. AP predicting BGEM in new environment: The AP training set used to predict BGEM grown in a different environment.

These prediction scenarios include within-population predictions both in a common environment (Scenarios I-III) and across different environments (Scenario IV) as well as the more challenging across-population prediction in a different environment (Scenario V). Prediction accuracy was calculated as the correlation between the observed and predicted values.

## 2.5 Principal components analysis

Principal components (PCs) were calculated using the *prcomp* function in R (version 4.0.3) (R Core Team 2020) using the overlapping 68,444 SNPs found in both BGEM and AP. First, genotype data for the 201 BGEM lines and the AP training set were used to calculate PCs

(Fig. 2A). Additionally, PCs were calculated using only the AP training set lines, then the BGEM lines were projected onto these coordinates using the R function *predict* (Fig. 2B).

**2.6 Optimal training population design**

Optimal training populations were examined in five scenarios (Fig. 1B):

A. AP OTP: OTP design used to identify a subset of the AP training set; this subset used to predict the remainder of the AP training set.
B. BGEM OTP: OTP design used to identify a subset of BGEM; this subset used to predict the remainder of BGEM.
C. AP and BGEM predicting AP validation set (2015 and 2017): OTP design used to identify a subset of the combined AP training set and BGEM data; this subset used to predict the AP validation set grown in 2015 and 2017.
D. AP and BGEM predicting AP validation set (2018): OTP design used to identify a subset of the combined AP training set and BGEM data; this subset used to predict the AP validation set grown in 2018.
E. AP and BGEM predicting remaining BGEM using OTP: OTP design used to identify a subset of the combined AP training set and BGEM data; this subset used to predict remaining BGEM.

Five methods were used: PEVmean and CDmean (Akdemir et al., 2015) implemented in STPGA (Akdemir, 2018); and MaxCD, PAM, and FURs (Guo et al., 2019). For PEVmean and CDmean, the function *GenAlgForSubsetSelctionNoTest* was used. Except for the PAM method which produces a single unique training population in the case of inbreds, OTP methods were replicated 50 times to create 50 distinct training populations for each prediction scenario. For

each method, OTPs consisting of 2.5%, 5%, 10%, and 15% of the full training set were identified. The range of 2.5%-15% of the training set was chosen based on Guo et al. (2019).

While PEVmean and CDmean were developed for use in inbreds, the other three methods were developed for hybrids and so had to be adapted for this purpose (Fig. S1). For the MaxCD method, a Euclidean distance matrix was calculated from the training set kinship, which was then used to draw a hierarchical tree. The desired number of inbreds were then chosen, evenly spaced, from the lowest level of this tree. For each replicate, the tree was shuffled. To enable multiple replicates for the FURS method, this method was updated to randomly select among equally-good candidates, defined as genotypes with an equal number of connections within the graphic network, when identifying additional genotypes to add to the training set. Finally, the PAM method was able to be applied to inbreds without modification, but is the only OTP method discussed here that produces a unique solution in the case of inbreds. The prediction accuracies when using these OTPs for gBLUP genomic prediction were compared with accuracies from random training sets of the same size.

# 3. RESULTS

## 3.1 PREDICTION WITHIN POPULATIONS

### 3.1.1 Prediction within the Ames Diversity Panel

In all three within-population prediction scenarios in AP (Scenarios I, III, and IV, Fig. 1A), prediction accuracy was respectable across all nine traits for seven of the eight genomic prediction methods (Fig. 3, S2-S4). Excluding sBLUP, prediction accuracies ranged from a minimum of 0.33 for the Scenario IV prediction of δT3 by BayesA to a maximum of 0.65 for the Scenario IV prediction of αT by BayesB (Fig. S4). The method sBLUP was a notable exception, with consistently lower prediction accuracies than other methods in all cases. Otherwise,

prediction accuracies were remarkably consistent for each trait across genomic prediction methods. Prediction accuracies were also largely consistent across the cross-validation (Scenario I), within-environment (Scenario III), and across-environment (Scenario IV) prediction scenarios for each trait (Fig. 3, S2-S4).

### 3.1.2 Prediction within BGEM

Ten-fold cross-validation within BGEM (Scenario II) achieved mean prediction accuracy ranging from 0.41 for ΣTT3 predicted by BL to 0.79 for δT predicted by BL or BRR (Fig. S5). Again, sBLUP was a notable outlier with significantly lower prediction accuracies than other methods for all traits. For most traits and prediction methods, prediction accuracies were significantly higher in BGEM CV (Scenario II) than in AP CV (Scenario I) (Fig. 3); for example, when excluding sBLUP, the trait with the highest prediction accuracies in BGEM CV was δT, with mean prediction accuracies of 0.78-0.79 (Fig. S5), as compared to 0.44-0.47 in AP CV (Fig. S2).

## 3.2 PREDICTION ACROSS POPULATIONS

To examine the challenge of predicting novel, exotic-derived germplasm from adapted germplasm, the AP training set was used to predict BGEM (Scenario V) (Fig. 3). For a few traits, similar prediction accuracies were observed to those from within-population prediction scenarios; for example, prediction accuracy of 0.67 was achieved for δT using gBLUP in Scenario V, which is comparable to the accuracy values of 0.46 and 0.79 observed for this trait in Scenarios I and II, respectively (Fig. 3). For most traits, though, prediction accuracies were very erratic across methods and poor overall, including many negative accuracies. The sBLUP method was no longer the consistently worst method and in fact was the best method for one trait

(αT3) (Fig. S6). In general, there was no consistently best or worst genomic prediction method, in part because no single method was able to provide positive prediction accuracies for all traits.

This poor prediction accuracy likely reflects the genetic distance between the two populations and the resulting extrapolation of the genomic prediction model when predicting across populations. In PCs based on the combined data of the AP training set and BGEM, clear separation was visible between AP and BGEM, as well as between BGEM lines with the PHB47 parent and those with the PHZ51 parent (Fig. 2A). When the BGEM lines were projected onto PCs based only on AP training set genotypes, all BGEM lines were clustered around (0,0), indicating that the observed genetic diversity in AP used to construct these PCs did not well reflect that found in BGEM (Fig. 2B).

Because results from different prediction methods are generally consistent (with the exception of sBLUP), the rest of this paper focuses on gBLUP genomic prediction because of its superior computational speed (Wang et al., 2018) rather than running all eight prediction methods.

Including PCs in genomic prediction models to account for population structure may improve prediction across populations (Dadousis, Veerkamp, Heringstad, Pszczola, & Calus, 2014); therefore, PCs were added to the gBLUP model used in Scenario V. We found that including PCs in the genomic prediction models, does have the potential to improve prediction accuracies (Fig. S7). However, this potential would only be usable if an appropriate number of PCs could be identified without using validation set phenotypes. Three methods were used to select the number of PCs to include: identifying the elbow in the scree plot (Cattell, 1966), a BIC-based model selection implemented in GAPIT (Wang & Zhang, 2020), and identifying the number of PCs that minimizes the mean square error (MSE) of predictions within the training set

using ten-fold cross validation (Dadousis et al., 2014). However, these methods provided inconsistent results (Table S2). Overall, no single method consistently identified the best-performing model, and for three traits (αT, γT3, ΣT3), all models returned negative prediction accuracies (Table S2, Fig. S7).

## 3.3 OPTIMAL TRAINING POPULATION DESIGN

### 3.3.1 Optimal training population validation

A small, optimally-selected training population was sufficient to achieve prediction accuracy comparable to ten-fold cross-validation within a given population in both AP (Scenario A) and BGEM (Scenario B) (Fig. S8, S9) using gBLUP. In fact, some traits and training set design methods reached the same or better mean accuracy compared to the corresponding CV while requiring only a fraction of the phenotyping effort. For example, prediction accuracy for γT was even higher when predicted by 10% of BGEM selected by PAM (0.68, Scenario B) than in BGEM CV (0.65, Scenario II) (Fig. S5, S9). Notably, the PAM-selected training set provided the best (or tied for best) prediction accuracy in 52 out of 72 cases examined and was second-best (or tied for second) in an additional 13 cases. It was significantly better than the random training set in all cases in Scenario B and in all but one case (δT predicted by 15%) in Scenario A, although the relative advantage of PAM over the random training set tended to decrease as the size of the chosen training population increased.

### 3.3.2 Optimal training population for prediction across populations

Because some traits (e.g., αT3) had extremely low and even negative prediction accuracies when predicting across populations, even after adding PCs to control for population structure or using OTP within the AP training set to predict BGEM (Fig. S10), it seems likely that BGEM has genetic diversity for these traits that is not present in AP and therefore cannot be

predicted using AP alone. This was also supported by the PCA, as all BGEM lines clustered around zero when projected onto PCs calculated based on AP data (Fig 2B), but formed distinct clusters when BGEM genetic data were included in PC construction (Fig. 2A). Therefore, it is logical to create a training population including both AP and BGEM lines, but to use OTP design to decrease total phenotyping investment.

Doing so enabled successful prediction of the remaining BGEM lines for all traits (Scenario E) (Fig. 4, Fig. S11). Even difficult traits like αT and αT3 had good prediction accuracies using only 2.5% of the combined AP training set and BGEM data as a training population. Again, PAM gave consistently good results, providing the best mean prediction accuracy in 21 out of the 36 cases and second best in seven cases; however, it is worth noting that as the training population size increased, the relative advantage of PAM over random selection decreased. In fact, in nine cases, PAM did not perform significantly better than random. Not all training set selection methods performed comparably; for example, FURS performed poorly in several traits, notably ΣTT3.

Additionally, to verify that the combined OTPs could predict both populations, they were also used to predict the AP validation set in 2015 and 2017 (Scenario C, Fig. S12) and in 2018 (Scenario D, Fig. S13). While the OTPs did not predict the AP validation set as well as did the full AP training set, prediction accuracies up to 0.41 were achieved. Again, PAM was the best OTP method in most cases, and OTP design generally provided a greater advantage over random design in smaller training population sizes.

## 4. DISCUSSION

As expected, genomic prediction for tocochromanol traits achieved generally high accuracies within both the BGEM and the AP population (Scenarios I-IV). In addition,

prediction accuracies with AP for a given trait across CV, common environment, and different environment prediction scenarios (Scenarios I, III, and IV) were consistent (Fig. S2-S4). The consistent accuracies across the common and different environment prediction scenarios could indicate that tocochromanol content is a relatively stable trait, or alternatively that the environments studied were too similar in important factors for noticeable genotype by environment interaction to occur. The consistency across all three prediction scenarios indicates that the AP training set contains a good representation of the diversity present in AP for these traits. Together, this suggests that tocochromanol content is well-suited to improvement by genomic selection within a population, facilitating biofortification.

The widespread application of genomic prediction in crop breeding has prompted the development of many different prediction models (e.g., (Kizilkaya et al., 2010; Meuwissen et al., 2001; Wang et al., 2018). Because these different methods make different assumptions about the true genetic architecture of a trait, they are expected to have different prediction accuracies depending on how closely those assumptions correspond to the reality for a given trait. For example, the sBLUP method is best suited for prediction of simple traits controlled by few genes (Wang et al., 2018). The poor relative performance of sBLUP in the tocochromanol traits analyzed in this study may suggest that the true genetic architecture of these traits in AP and BGEM is more complex. This is supported by existing literature, as 52 QTLs have previously been reported for tocochromanol content in maize grain (Diepenbrock et al., 2017). Barring mismatches between the assumptions of the genomic prediction model and the true genetic architecture, any modern genomic prediction model will typically produce similarly good results (e.g., (Calus et al., 2014; Daetwyler, Calus, Pong-Wong, de los Campos, & Hickey, 2013). Therefore, when selecting a genomic prediction method from among several with assumptions

that fit a given situation, computational efficiency and ease of implementation may become the decisive factor.

However, prediction accuracies drop substantially with all methods when predicting across populations (Fig. 3, Fig. S6). Despite previous principal coordinate analyses grouping GEM lines between ExPVP and tropical lines, suggesting some shared genetics between Corn Belt and exotic lines (Romay et al., 2013), as well as generally overlapping phenotypic ranges for tocochromanol traits (Table S1), PCA of BGEM and AP in this study indicated that these two populations had different patterns of genetic diversity (Fig. 2). This leads to extrapolation when using genomic prediction across these populations. While some traits (e.g., ΣT) achieved comparable prediction accuracy across populations as within populations, most traits had substantially poorer and even negative (e.g., αT, γT3) prediction accuracies when predicting across populations (Fig. 3, S6). Incorporating PCs in the model to improve genomic prediction accuracy in the presence of population structure has been suggested and has been successful in some cases (Azevedo et al., 2017; Dadousis et al., 2014) but not all (Lyra et al., 2018). This did improve prediction accuracies in some traits, notably αT3 when using PCs based on the combined AP and BGEM data (Fig. S7). However, the available methods of selecting PCs *a priori* for inclusion in the model often provide very different recommendations and resulting prediction accuracies, and in fact rarely select the model with the highest prediction accuracy (Table S2).

Optimal training population design improved or maintained prediction accuracy while reducing required investment in phenotyping. The training population design methods used in this analysis were developed for use in diverse panels of inbreds (PEVmean and CDmean) (Akdemir et al., 2015) or in hybrids (MaxCD, FURS, and PAM) (Guo et al., 2019). Here, all

methods were validated in BC DHs for the first time in Scenario B, and MaxCD, FURS, and PAM were validated in diverse inbreds for the first time in Scenario A. While PEVmean and CDmean could be directly applied to the BGEM and AP data using existing R functions (Akdemir, 2018), MaxCD, FURS, and PAM methods were adapted for use in non-hybrids. Compared to CV, OTP reached comparable or occasionally even better prediction accuracy when predicting within both AP and BGEM while requiring only a fraction of the phenotyping effort. PAM performed well in both Scenario A and B validation and has the advantage of recommending a single optimal training population, providing a clear recommendation of which individuals to phenotype to form the training population. Notably, the relative advantage of optimal training population design over random selection was greatest in small training population sizes, as seen in previous literature (Pinho Morais et al., 2020).

Optimal training population design also performed well in across population prediction. When using only AP lines in the training set, prediction accuracies were similar overall whether using an optimal training population (Fig. S10) or the full AP training set (Fig S6), despite OTP using a much smaller training population and therefore requiring much less resources for phenotyping. However, these still included very poor and even negative accuracies. Instead, creating an optimally selected training set from the combined adapted and exotic-derived populations provided a solution to improve prediction accuracy while minimizing additional phenotyping effort required compared to a random approach. Using PAM to identify an OTP consisting of only 2.5% of the combined AP training set and BGEM data (Scenario E) enabled notably improved prediction accuracies of 0.25-0.78 when predicting the remaining BGEM lines (Fig. 4, S11). This OTP was not limited to predicting BGEM. It was able to predict both component populations, as shown by prediction accuracies up to 0.41 when used to predict the

AP validation set (Scenarios C-D, Fig. S12, S13). Of course, this approach does require the growing and phenotyping of some members of the new population. However, in the case of exploiting novel, exotic germplasm, this small additional investment is worthwhile as it results in a significant increase in prediction accuracy by avoiding extrapolation.

The diverse germplasm available from gene banks and exotic-derived panels such as the BGEM panel studied here are a critical resource for breeders, especially as they continue to improve both yield and nutritional content in the face of rapid changes in climate coupled with increasing global population (Dyer, López-Feldman, Yúnez-Naude, & Taylor, 2014; McLean-Rodríguez, Costich, Camacho-Villa, Pè, & Dell'Acqua, 2021; Vanous et al., 2018; Yu et al., 2016). Notably, exotic donor lines have already been used to increase provitamin A in maize grain (Menkir, Rocheford, Maziya-Dixon, & Tanumihardjo, 2015). Genomic prediction has been recommended as an approach to turbocharge gene banks, enabling assessment and utilization of these genetic resources (Yu et al., 2016). OTP design methods will facilitate this process, enabling the initial design of the training set and efficient updates of existing training sets as additional diverse germplasm is genotyped.

## 5. CONCLUSION AND PERSPECTIVES

Genomic prediction is a critical tool in crop breeding, enabling rapid prediction and selection of new germplasm. It relies on shared genetic information between the training and testing set, but in the case of a new exotic-derived validation set, this assumption may not be justified for all traits, limiting the utility of genomic prediction in exotic germplasm. However, incorporating exotic germplasm can bring new diversity and potentially beneficial alleles into the breeding program, creating a dilemma. In this study, we investigated this dilemma and found that

OTP design using only 2.5% of the combined adapted and exotic germplasm sets enabled acceptable prediction accuracy in the rest of the exotic-derived population.

Therefore, when incorporating a new exotic population into a breeding program or genomic prediction model, we recommend using PAM or a similar optimal training population design method to identify an optimal subset of the combined adapted and exotic lines to phenotype. This combined training population will facilitate the combination of both populations into a single breeding program, enabling prediction of members of both populations. The training population can be made larger or smaller depending on available phenotyping resources and will enable genomic prediction without extrapolation. This approach will facilitate utilization of exotic germplasm in maize breeding projects including vitamin E biofortification.

## DATA AND CODE ACCESS

Ames Diversity Panel genotype data was obtained from Panzea.org/genotypes (file ZeaGBSv27_publicSamples_imputedV5_AGPv4-181023.vcf). BGEM genotype data as well as Ames Diversity Panel and BGEM tocochromanol data are available at https://github.com/LTibbs/vit_predict. This GitHub also contains R scripts for optimal training population design methods MaxCD, PAM, and FURS applied to inbreds as well as the custom R script used to create consensus sequences.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

LTC: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft, Writing – Review & Editing; TG:

Methodology, Writing – Review & Editing; XL: Methodology, Writing – Review & Editing; RT: Writing – Review & Editing; AEV: Data Curation, Writing – Review & Editing; DP: Resources, Writing – Review & Editing; CG: Resources, Writing – Review & Editing; MML: Investigation, Data Curation, Writing – Review & Editing; NTD: Investigation, Writing – Review & Editing; DDP: Funding Acquisition, Writing – Review & Editing; MAG: Funding Acquisition, Writing – Review & Editing; JY: Conceptualization, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

## ACKNOLWEDGEMENTS

This research was supported by the National Science Foundation (IOS-1546657 to DDP, MAG, and JY); the National Institute of Food and Agriculture of the USDA Hatch under accession numbers 1013641 (MAG), 1023660 (MAG), and 1021013 (JY); HarvestPlus (MAG); and the Iowa State University Plant Sciences Institute. LTC was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. 1744592). This study was also made possible by the support of the American People provided to the Feed the Future Innovation Lab for Crop Improvement through the United States Agency for International Development (USAID) (M.A.G.). The contents are the sole responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government. Program activities are funded by the United States Agency for International Development (USAID) under Cooperative Agreement No. 7200AA-19LE-00005.

## ORCID

Laura E. Tibbs-Cortes https://orcid.org/0000-0003-3188-6820

Tingting Guo https://orcid.org/0000-0002-6647-6998

Xianran Li https://orcid.org/0000-0002-4252-6911

Ryokei Tanaka https://orcid.org/0000-0002-3479-377X

Adam E. Vanous https://orcid.org/0000-0003-0079-7286

David Peters https://orcid.org/0000-0001-9148-1660

Candice Gardner https://orcid.org/0000-0002-5334-6123

Maria Magallanes-Lundback https://orcid.org/0000-0001-5826-6897

Nicholas T. Deason https://orcid.org/0000-0002-2218-2699

Dean DellaPenna https://orcid.org/0000-0001-9505-7883

Michael A. Gore https://orcid.org/0000-0001-6896-8024

Jianming Yu https://orcid.org/0000-0001-5326-3099

## REFERENCES

Akdemir, D. (2018). STPGA: Selection of Training Populations by Genetic Algorithm. R package version 5.2.1. Retrieved from https://cran.r-project.org/package=STPGA

Akdemir, D., Sanchez, J. I., & Jannink, J. L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, *47*, 38. https://doi.org/10.1186/s12711-015-0116-6

Azevedo, C. F., de Resende, M. D. V., Fonseca e Silva, F., Nascimento, M., Viana, J. M. S., & Valente, M. S. F. (2017). Population structure correction for genomic selection through eigenvector covariates. *Crop Breeding and Applied Biotechnology*, *17*(4), 350–358. https://doi.org/10.1590/1984-70332017V17N4A53

Baseggio, M., Murray, M., Magallanes-Lundback, M., Kaczmar, N., Chamness, J., Buckler, E. S., … Gore, M. A. (2019). Genome-Wide Association and Genomic Prediction Models of Tocochromanols in Fresh Sweet Corn Kernels. *The Plant Genome*, *12*(1), 180038. https://doi.org/10.3835/PLANTGENOME2018.06.0038

Calus, M. P., Huang, H., Vereijken, A., Visscher, J., ten Napel, J., & Windig, J. J. (2014). Genomic prediction based on data from three layer lines: a comparison between linear methods. *Genetics Selection Evolution 2014 46:1*, *46*, 57. https://doi.org/10.1186/S12711-014-0057-5

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chander, S., Guo, Y. Q., Yang, X. H., Yan, J. B., Zhang, Y. R., Song, T. M., & Li, J. S. (2008). Genetic dissection of tocopherol content and composition in maize grain using quantitative trait loci analysis and the candidate gene approach. *Molecular Breeding*, *22*(3), 353–365. https://doi.org/10.1007/S11032-008-9180-8

Chander, S., Guo, Y. Q., Yang, X. H., Zhang, J., Lu, X. Q., Yan, J. B., … Li, J. S. (2008). Using molecular markers to identify two major loci controlling carotenoid contents in maize grain. *Theoretical and Applied Genetics*, *116*(2), 223–233. https://doi.org/10.1007/s00122-007-0661-7

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., … Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Dadousis, C., Veerkamp, R. F., Heringstad, B., Pszczola, M., & Calus, M. P. (2014). A comparison of principal component regression and genomic REML for genomic prediction across populations. *Genetics Selection Evolution*, *46*, 60. https://doi.org/10.1186/S12711-014-0060-X

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., & Hickey, J. M. (2013). Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics*, *193*(2), 347–365. https://doi.org/10.1534/GENETICS.112.147983

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., … Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375–385. https://doi.org/10.1534/genetics.109.101501

DellaPenna, D., & Mène-Saffrané, L. (2011). Vitamin E. In Fabrice Rébeillé & Roland Douce (Eds.), *Advances in Botanical Research* (Vol. 59, pp. 179–227). Academic Press. https://doi.org/10.1016/B978-0-12-385853-5.00002-7

Diepenbrock, C. H., Kandianis, C. B., Lipka, A. E., Magallanes-Lundback, M., Vaillancourt, B., Góngora-Castillo, E., … Dellapenna, D. (2017). Novel Loci Underlie Natural Variation in Vitamin E Levels in Maize Grain. *The Plant Cell*, *29*, 2374–2392. https://doi.org/10.1105/tpc.17.00475

Dyer, G. A., López-Feldman, A., Yúnez-Naude, A., & Taylor, J. E. (2014). Genetic erosion in maize's center of origin. *Proceedings of the National Academy of Sciences*, *111*(39), 14094–14099. https://doi.org/10.1073/pnas.1407033111

Dzievit, M. J., Guo, T., Li, X., & Yu, J. (2021). Comprehensive analytical and empirical evaluation of genomic prediction across diverse accessions in maize. *The Plant Genome*, *14*, e20160. https://doi.org/10.1002/tpg2.20160

Gianola, D. (2021). Opinionated Views on Genome-Assisted Inference and Prediction During a Pandemic. *Frontiers in Plant Science*, *12*, 717284. https://doi.org/10.3389/FPLS.2021.717284/BIBTEX

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., … Yu, J. (2019). Optimal Designs for Genomic Selection in Hybrid Crops. *Molecular Plant*, *12*(3), 390–401. https://doi.org/10.1016/j.molp.2018.12.022

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*, 186. https://doi.org/10.1186/1471-2105-12-186

Hershberger, J., Tanaka, R., Wood, J. C., Kaczmar, N., Wu, D., Hamilton, J. P., … Gore, M. A. (2022). Transcriptome-wide association and prediction for carotenoids and tocochromanols in fresh sweet corn kernels. *The Plant Genome*, e20197. https://doi.org/10.1002/TPG2.20197

Kizilkaya, K., Fernando, R. L., & Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of Animal Science*, *88*(2), 544–551. https://doi.org/10.2527/jas.2009-2064

Li, Q., Yang, X., Xu, S., Cai, Y., Zhang, D., Han, Y., … Yan, J. (2012). Genome-Wide Association Studies Identified Three Independent Polymorphisms Associated with α-Tocopherol Content in Maize Kernels. *PLOS ONE*, *7*(5), e36807. https://doi.org/10.1371/JOURNAL.PONE.0036807

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., … DellaPenna, D. (2013). Genome-Wide Association Study and Pathway-Level Analysis of Tocochromanol Levels in Maize Grain. *Genes Genomes Genetics*, *3*(8), 1287–1299. https://doi.org/10.1534/g3.113.006148

Lyra, D. H., Granato, Í. S. C., Morais, P. P. P., Alves, F. C., dos Santos, A. R. M., Yu, X., … Fritsche-Neto, R. (2018). Controlling population structure in the genomic prediction of tropical maize hybrids. *Molecular Breeding*, *38*, 126. https://doi.org/10.1007/s11032-018-0882-2

McLean-Rodríguez, F. D., Costich, D. E., Camacho-Villa, T. C., Pè, M. E., & Dell'Acqua, M. (2021). Genetic diversity and selection signatures in maize landraces compared across 50 years of in situ and ex situ conservation. *Heredity*, *126*, 913–928. https://doi.org/10.1038/s41437-021-00423-y

Menkir, A., Rocheford, T., Maziya-Dixon, B., & Tanumihardjo, S. (2015). Exploiting natural variation in exotic germplasm for increasing provitamin-A carotenoids in tropical maize. *Euphytica*, *205*(1), 203–217. https://doi.org/10.1007/S10681-015-1426-Z/TABLES/6

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Owens, B. F., Lipka, A. E., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C. H., Kandianis, C. B., … Rocheford, T. (2014). A foundation for provitamin A biofortification of maize: Genome-wide association and genomic prediction models of carotenoid levels. *Genetics*, *198*(4), 1699–1716. https://doi.org/10.1534/genetics.114.169979

Pérez, P., & de los Campos, G. (2014). BGLR : A Statistical Package for Whole Genome Regression and Prediction. *Genetics*, *198*(2), 483–495.

Péter, S., Friedel, A., Roos, F. F., Wyss, A., Eggersdorfer, M., Hoffmann, K., & Weber, P. (2015). A Systematic Review of Global Alpha-Tocopherol Status as Assessed by Nutritional Intake Levels and Blood Serum Concentrations. *Int. J. Vitam. Nutr. Res*, *85*, 5–

608 6. https://doi.org/10.1024/0300-9831/a000102

609 Pinho Morais, P. P., Akdemir, D., Braatz de Andrade, L. R., Jannink, J. L., Fritsche-Neto, R.,
610 Borém, A., … Granato, Í. S. C. (2020, August 1). Using public databases for genomic
611 prediction of tropical maize lines. *Plant Breeding*. Blackwell Publishing Ltd.
612 https://doi.org/10.1111/pbr.12827

613 RCoreTeam. (2020). R: A language and environment for statistical computing. Vienna, Austria:
614 Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

615 Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., …
616 Gardner, C. A. (2013). Comprehensive genotyping of the USA national maize inbred seed
617 bank. *Genome Biology*, *14*, R55. https://doi.org/10.1186/gb-2013-14-6-r55

618 Rosell, C. M. (2016). Fortification of Grain-Based Foods. In *Reference Module in Food Science*.
619 Elsevier. https://doi.org/10.1016/b978-0-08-100596-5.00074-3

620 Shutu, X., Dalong, Z., Ye, C., Yi, Z., Shah, T., Ali, F., … Jianbing, Y. (2012). Dissecting
621 tocopherols content in maize (Zea mays L.), using two segregating populations and high-
622 density single nucleotide polymorphism markers. *BMC Plant Biology 2012 12:1*, *12*(1), 1–
623 14. https://doi.org/10.1186/1471-2229-12-201

624 Vanous, A., Gardner, C., Blanco, M., Martin-Schwarze, A., Lipka, A. E., Flint-Garcia, S., …
625 Lübberstedt, T. (2018). Association Mapping of Flowering and Height Traits in Germplasm
626 Enhancement of Maize Doubled Haploid (GEM-DH) Lines. *The Plant Genome*, *11*(2),
627 170083. https://doi.org/10.3835/PLANTGENOME2017.09.0083

628 Wang, J., & Zhang, Z. (2020). GAPIT Version 3: Boosting Power and Accuracy for Genomic

Association and Prediction. *BioRxiv*. https://doi.org/10.1101/2020.11.29.403170

Wang, J., Zhou, Z., Zhang, Z., Li, H., Liu, D., Zhang, Q., … Zhang, Z. (2018). Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity*, *121*, 648–662. https://doi.org/10.1038/s41437-018-0075-0

Weber, E. J. (1987). Carotenoids and tocols of corn grain determined by HPLC. *Journal of the American Oil Chemists' Society 1987 64:8*, *64*, 1129–1134. https://doi.org/10.1007/BF02612988

Wong, J. C., Lambert, R. J., Tadmor, Y., & Rocheford, T. R. (2003). QTL Associated with Accumulation of Tocopherols in Maize. *Crop Science*, *43*(6), 2257–2266. https://doi.org/10.2135/CROPSCI2003.2257

Wu, D., Li, X., Tanaka, R., Wood, J. C., Tibbs-Cortes, L. E., Magallanes-Lundback, M., … Gore, M. A. (2022). Combining GWAS and TWAS to identify candidate causal genes for tocochromanol levels in maize grain. *BioRxiv*. https://doi.org/10.1101/2022.04.01.486706

Yu, X., Leiboff, S., Li, X., Guo, T., Ronning, N., Zhang, X., … Yu, J. (2020). Genomic prediction of maize microphenotypes provides insights for optimizing selection and mining diversity. *Plant Biotechnology Journal*, *18*, 2456–2465. https://doi.org/10.1111/PBI.13420

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S. E., … Yu, J. (2016). Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nature Plants*, *2*, 16150. https://doi.org/10.1038/nplants.2016.150

## FIGURES AND TABLES

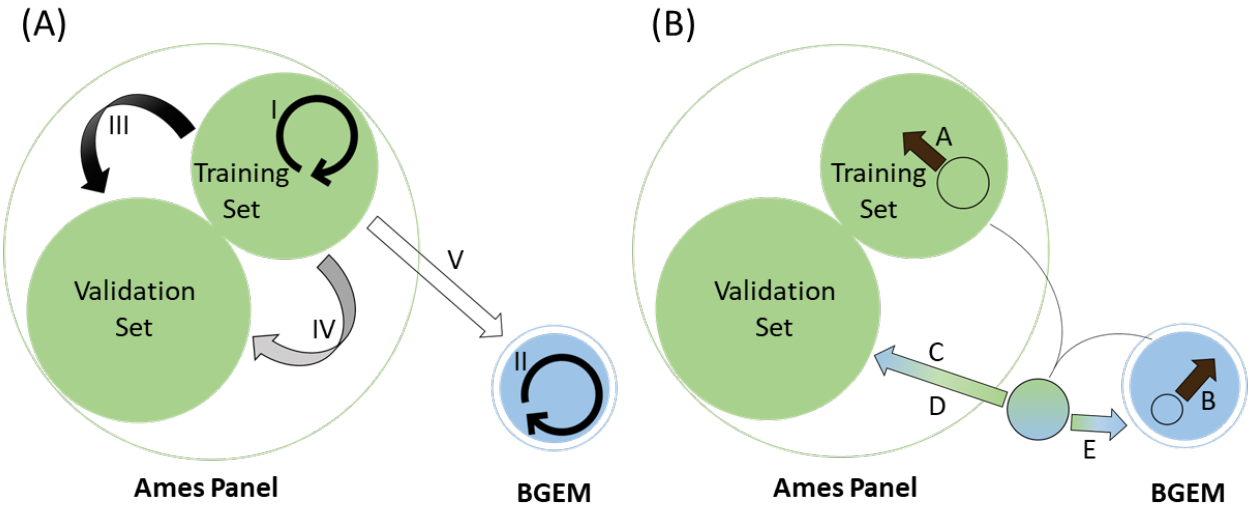


Figure 1. Overview of genomic prediction scenarios. (A) Five prediction scenarios: I. AP CV; II. BGEM CV; III. AP, common environment; IV. AP, new environment; V. AP predicting BGEM in new environment. (B) Five Optimal Training Population (OTP) design scenarios: A. AP OTP; B. BGEM OTP; C. AP and BGEM predicting AP validation set (2015 and 2017); D. AP and BGEM predicting AP validation set (2018); E. AP and BGEM predicting remaining BGEM. In both panels, the area of each circle is approximately proportional to the number of lines included; filled circles represent members of the population phenotyped in this study. Arrows show predictions; circular arrows denote ten-fold cross-validation. Arrow fill denotes the amount of information shared between training and validation set; black fill denotes the case with the most shared data (common environment, within population), grey fill the moderate case (new environment, within population), white fill the least shared data (new environment, across populations), and blue/green gradient fill the case in which members of the training set vary in the amount of information they share with the validation set (Scenarios C-E ).

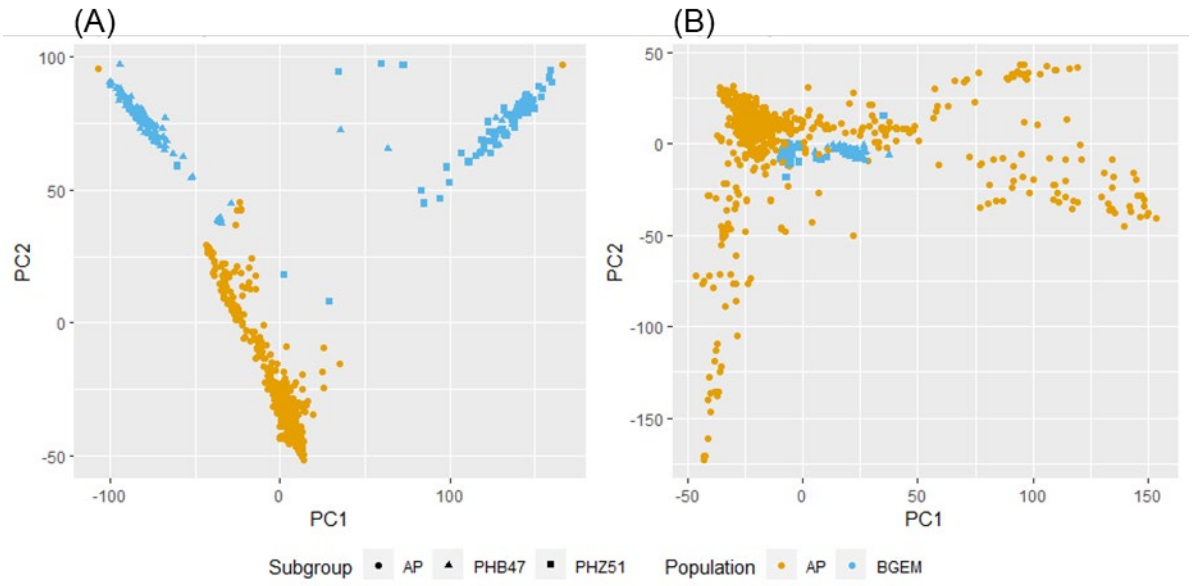


Figure 2. Genetic relationship shown with the first two principal components (PCs). (A) Clear separation is visible between the Ames Diversity Panel (AP, orange) and BGEM (blue) lines when using PCs calculated from the combined genotype data of the AP training set and BGEM. Additionally, PC1 clearly separates the BGEM lines by recurrent parent (subgroup, denoted by shape). The two recurrent parents of BGEM, present in AP, cluster with their respective subgroups of BGEM. (B) The BGEM lines cluster at (0,0) on PCs calculated from the genotype data of the AP training set alone.

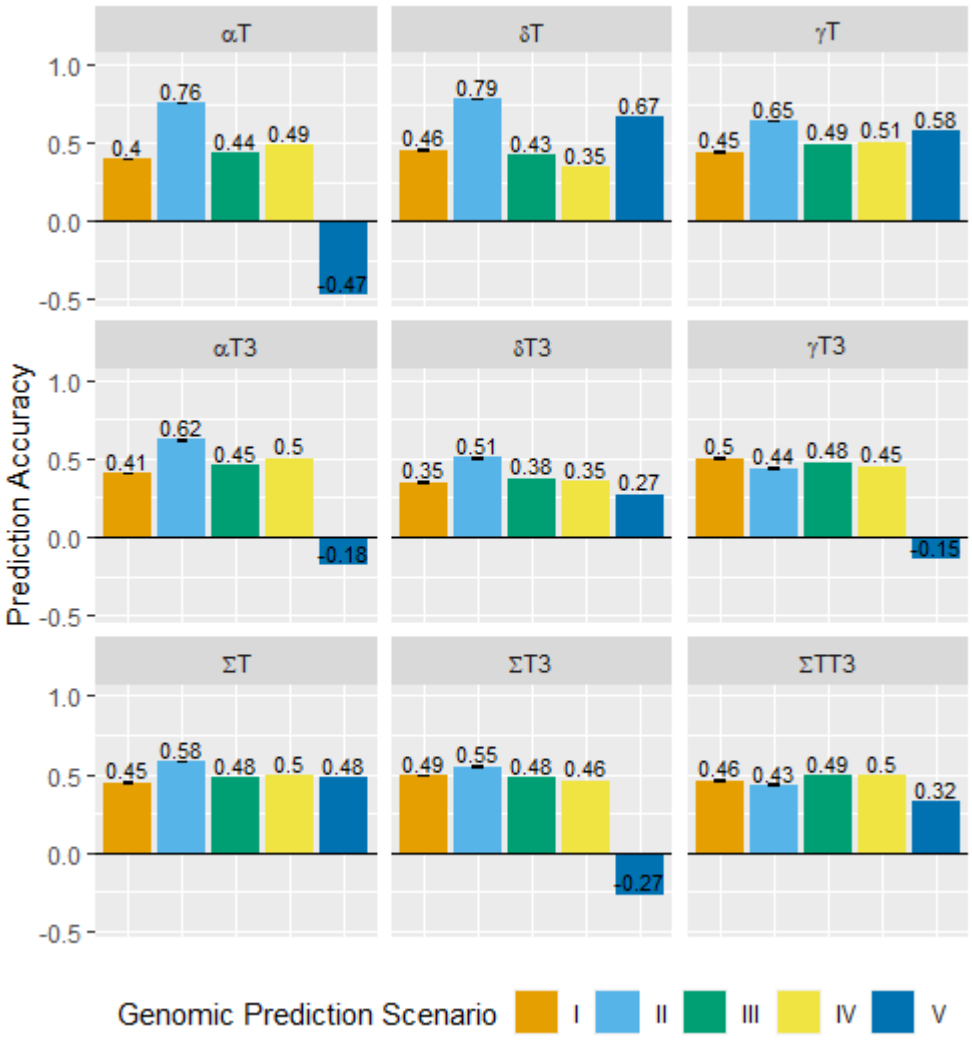


Figure 3. Genomic prediction accuracies. Prediction accuracies for all five genomic prediction scenarios (I-V) using gBLUP. For cross-validation scenarios (I and II), ten replicates were conducted; error bars show standard error.

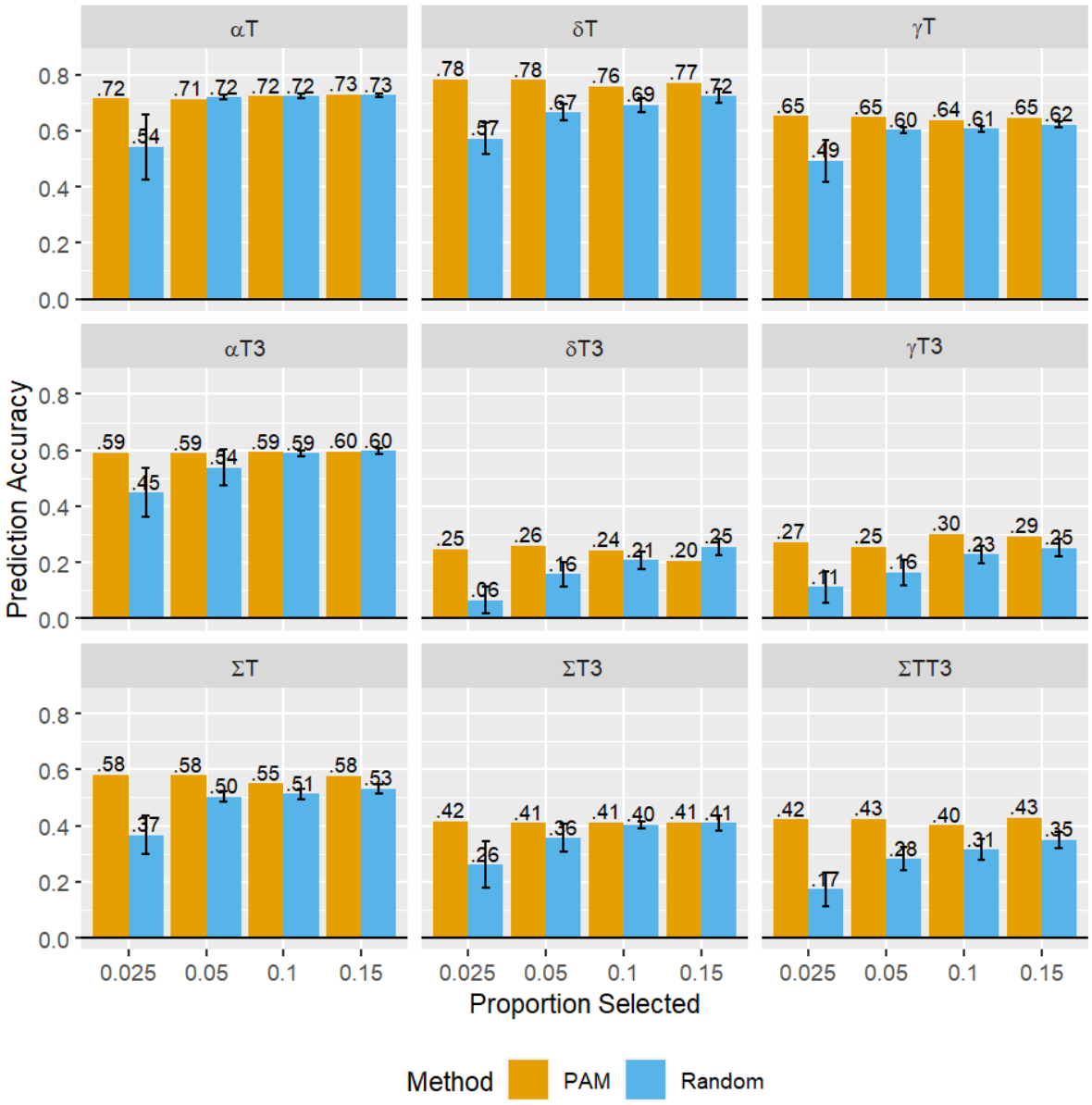


Figure 4. Prediction accuracy of PAM vs. random selection in Scenario E. The Optimal Training Population (OTP) design method PAM as well as random selection were used to select training sets of a given proportion (x axis) of the combined data of the AP training set and BGEM, which were then used to predict the remaining BGEM lines using gBLUP. Error bars show standard error for prediction accuracy based on 50 replicates.

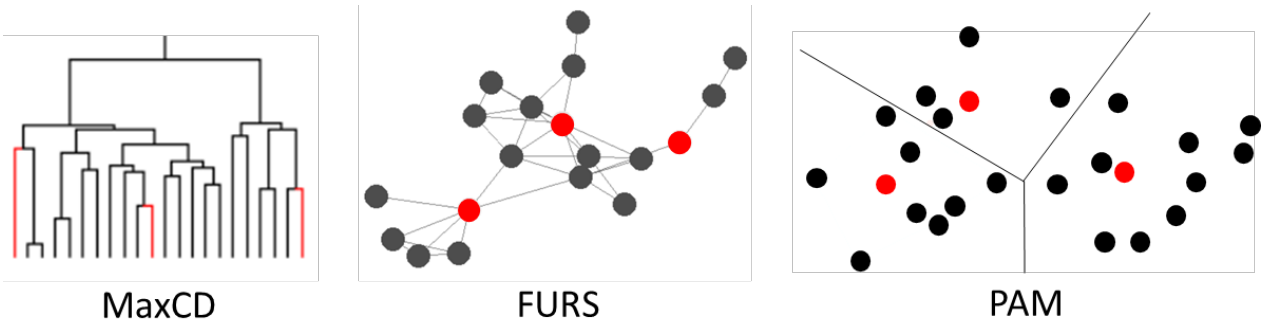


Figure S1. Schematic representations of Optimal Training Population (OTP) design algorithms. Red indicates inbreds chosen for the training population using MaxCD (maximization of connectedness and diversity), PAM (partitioning around medoids), and FURS (fast and unique representative subset selection). Two other OTP design methods, CDmean and PEVmean, are difficult to visualize due to their iterative nature of search procedure.

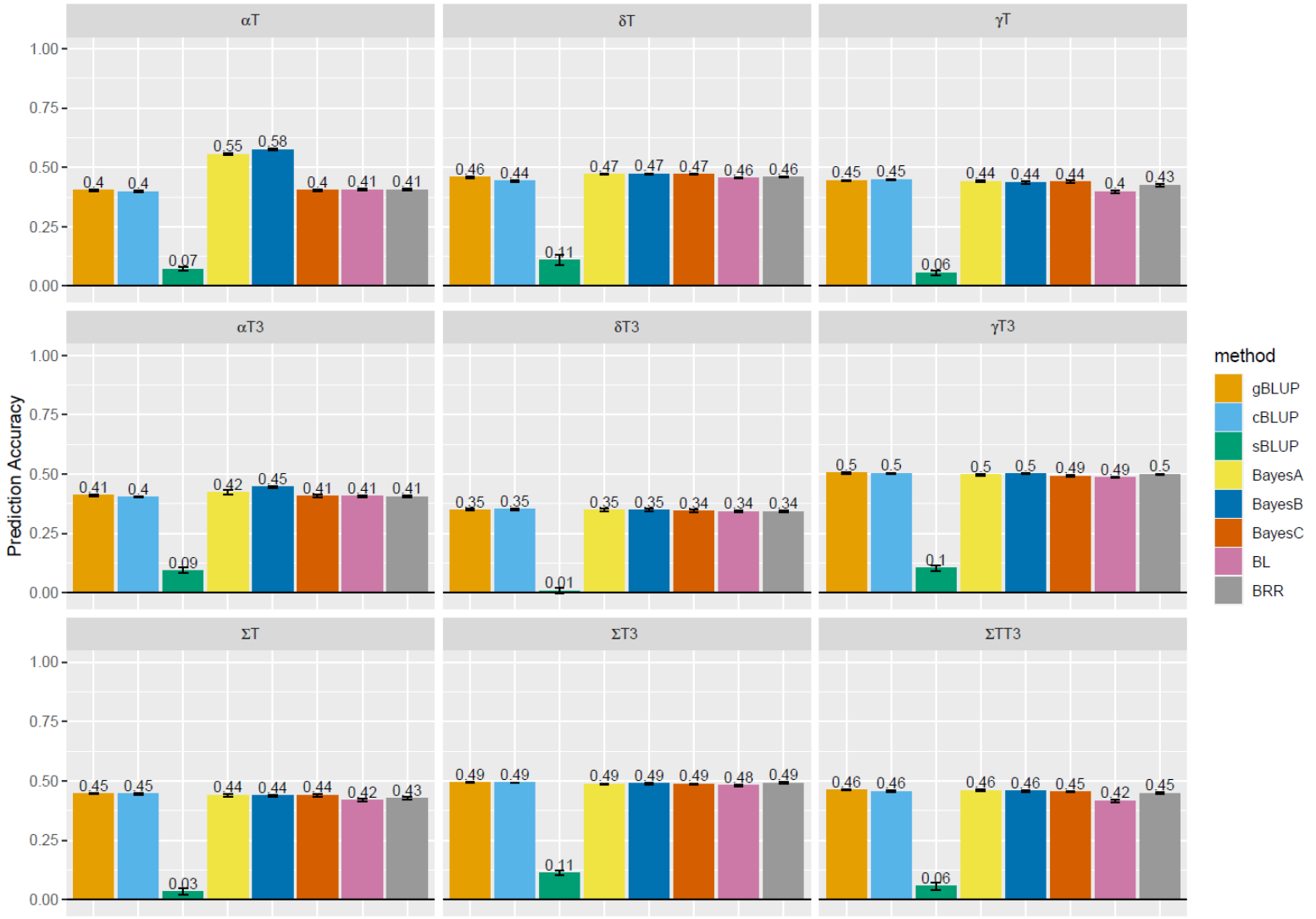


Figure S2. Genomic prediction Scenario I results. Genomic prediction accuracies using eight methods to predict nine traits. Ten replicates of ten-fold cross-validation were conducted; error bars show standard error.

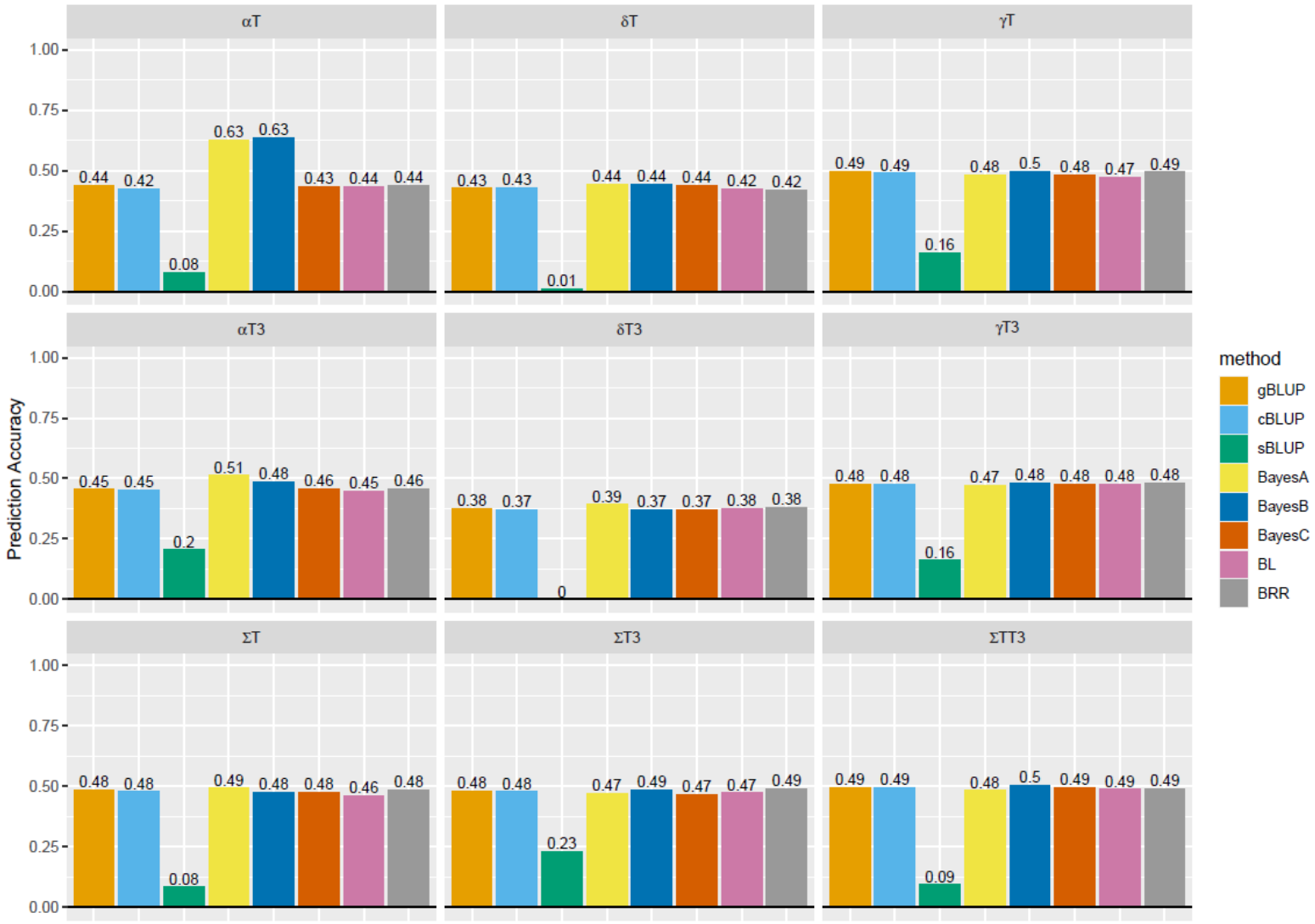


Figure S3. Genomic prediction Scenario III results. Genomic prediction accuracies using eight methods to predict nine traits.

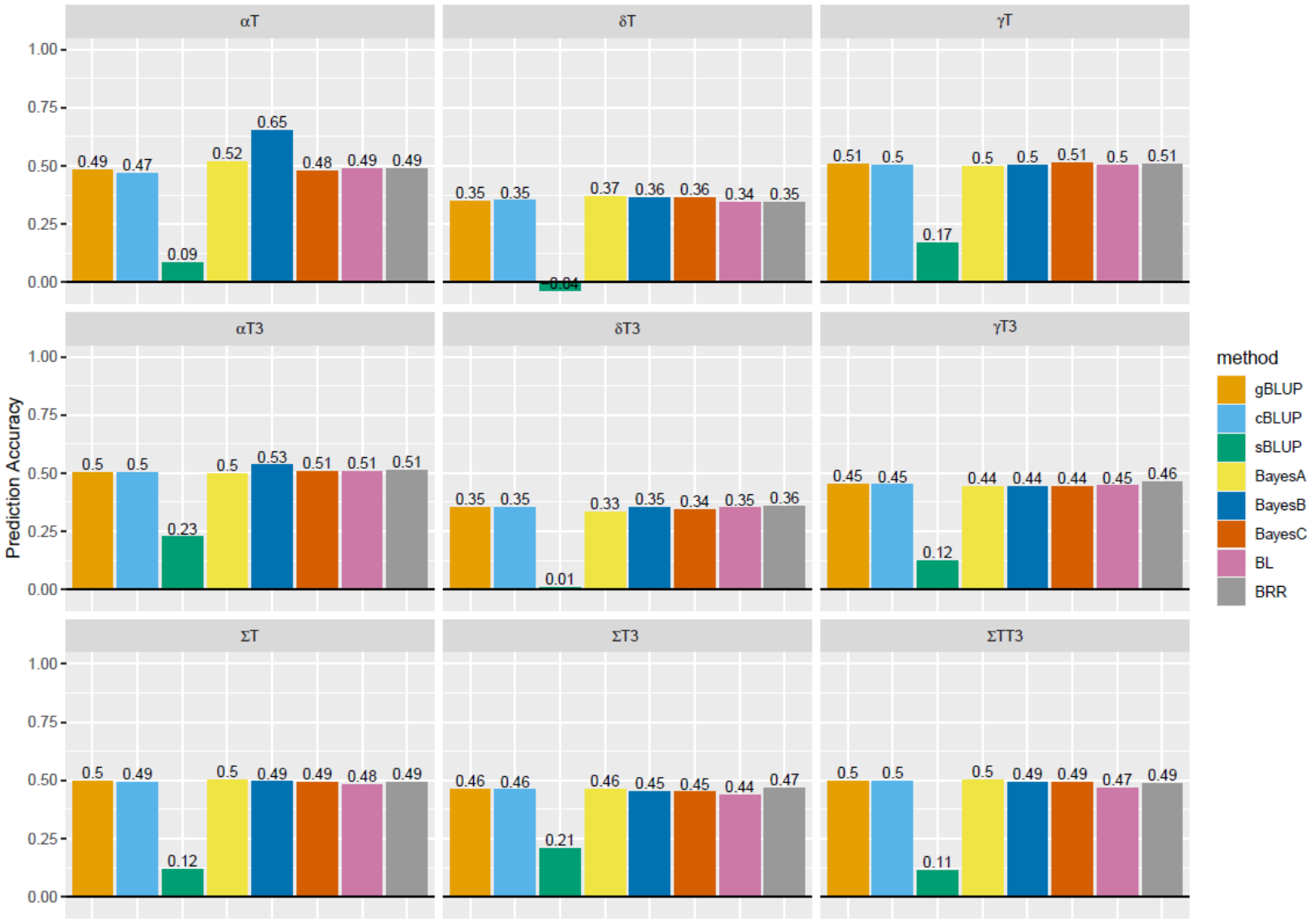


Figure S4. Genomic prediction Scenario IV results. Genomic prediction accuracies using eight methods to predict nine traits.

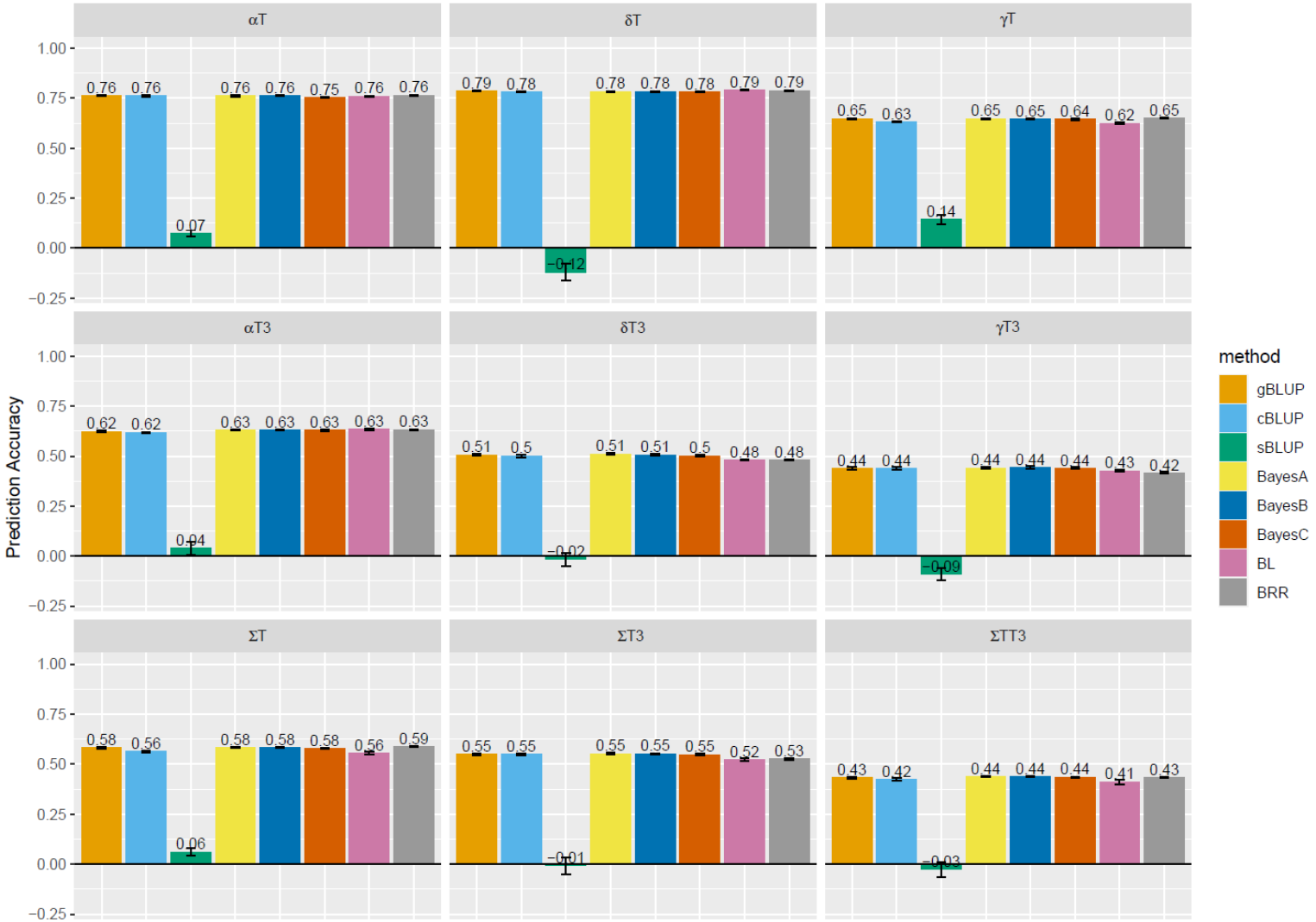


Figure S5. Genomic prediction Scenario II results. Genomic prediction accuracies using eight methods to predict nine traits. Ten replicates of ten-fold cross-validation were conducted; error bars show standard error.

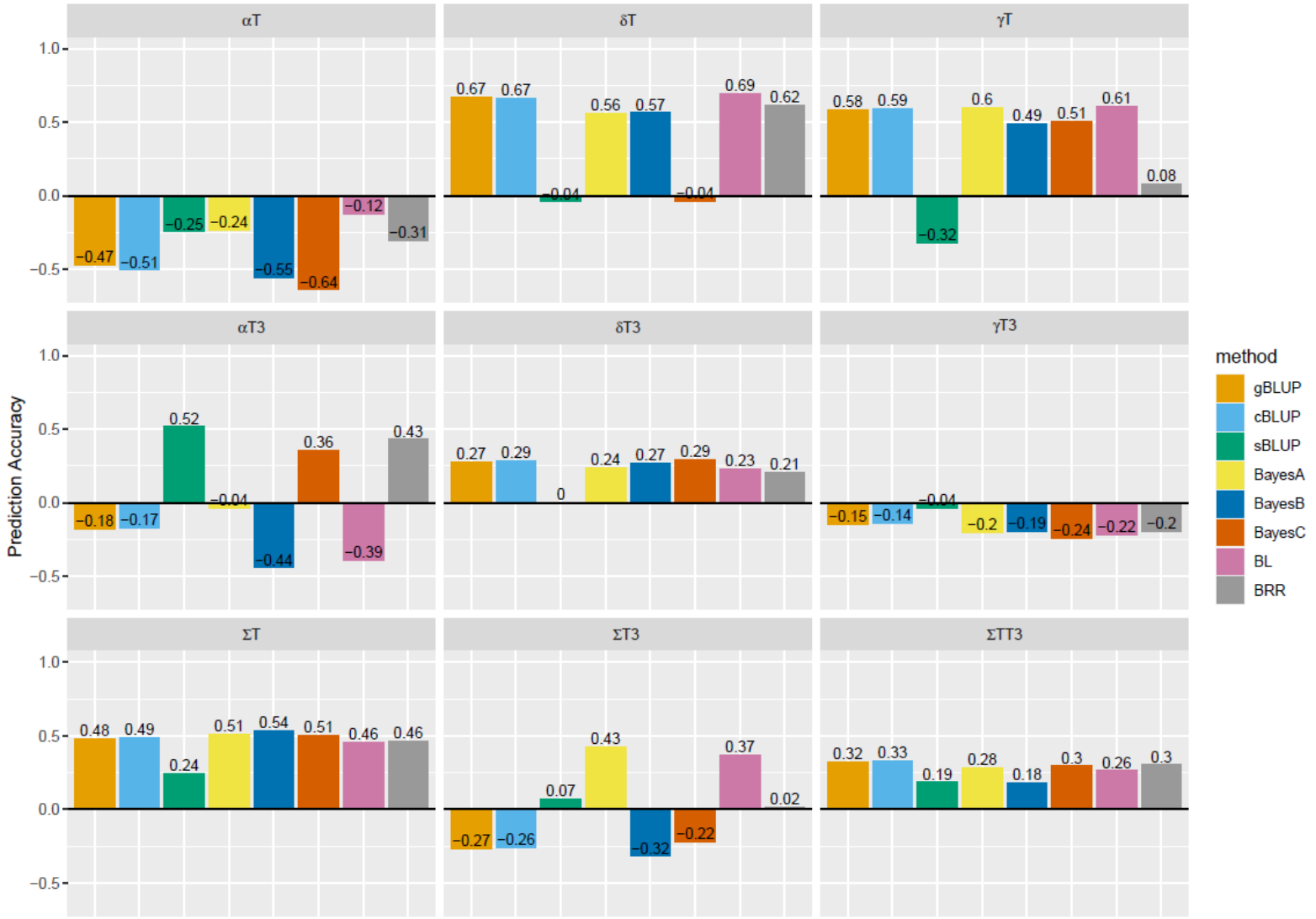


Figure S6. Genomic prediction Scenario V results. Genomic prediction accuracies using eight methods to predict nine traits.

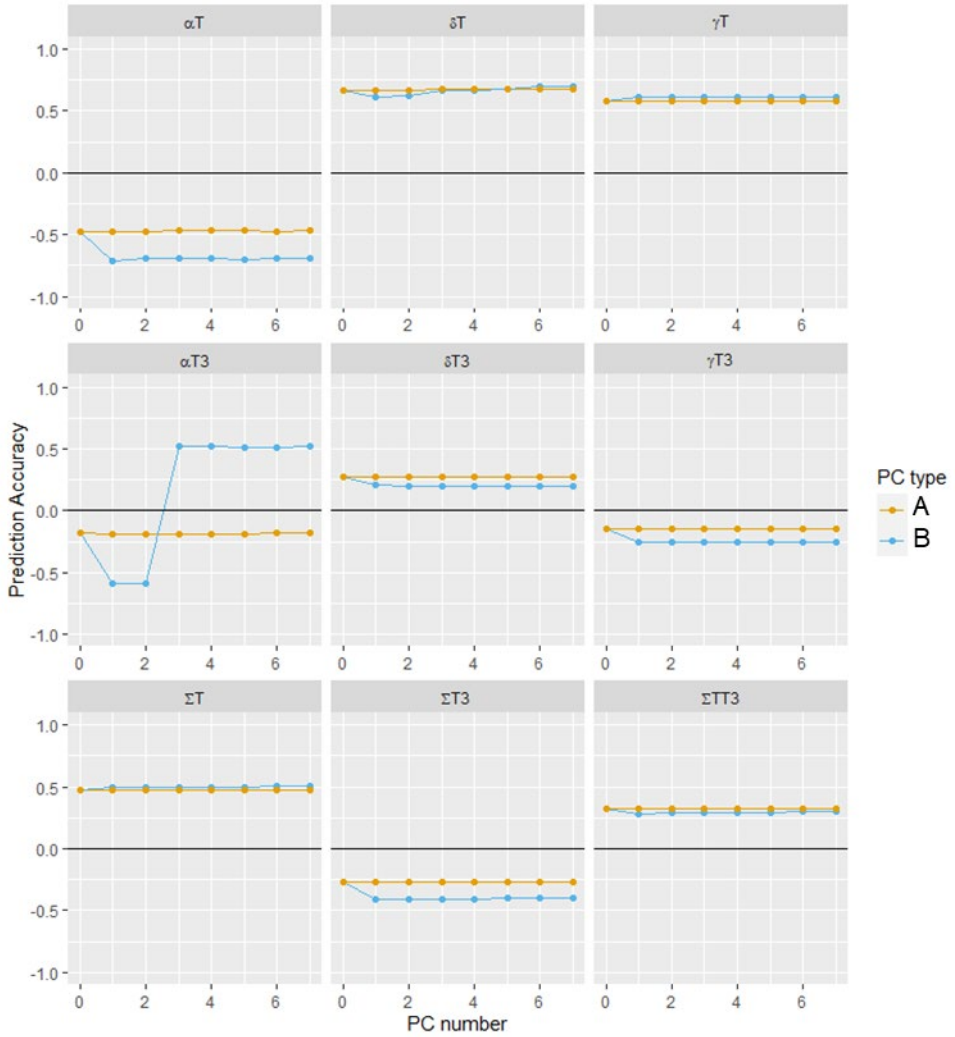


Figure S7. Prediction accuracy incorporating principal components in Scenario V. The prediction accuracy (y axis) achieved in Scenario V by a gBLUP model including a given number of principal components (PCs, x axis) is shown for each trait (panel labels). Results are shown for PCs calculated from the combined genotype data from AP and BGEM (denoted A, see Fig. 2A) and from the AP genotype data only (denoted B, see Fig. 2B), shown in blue and orange, respectively.

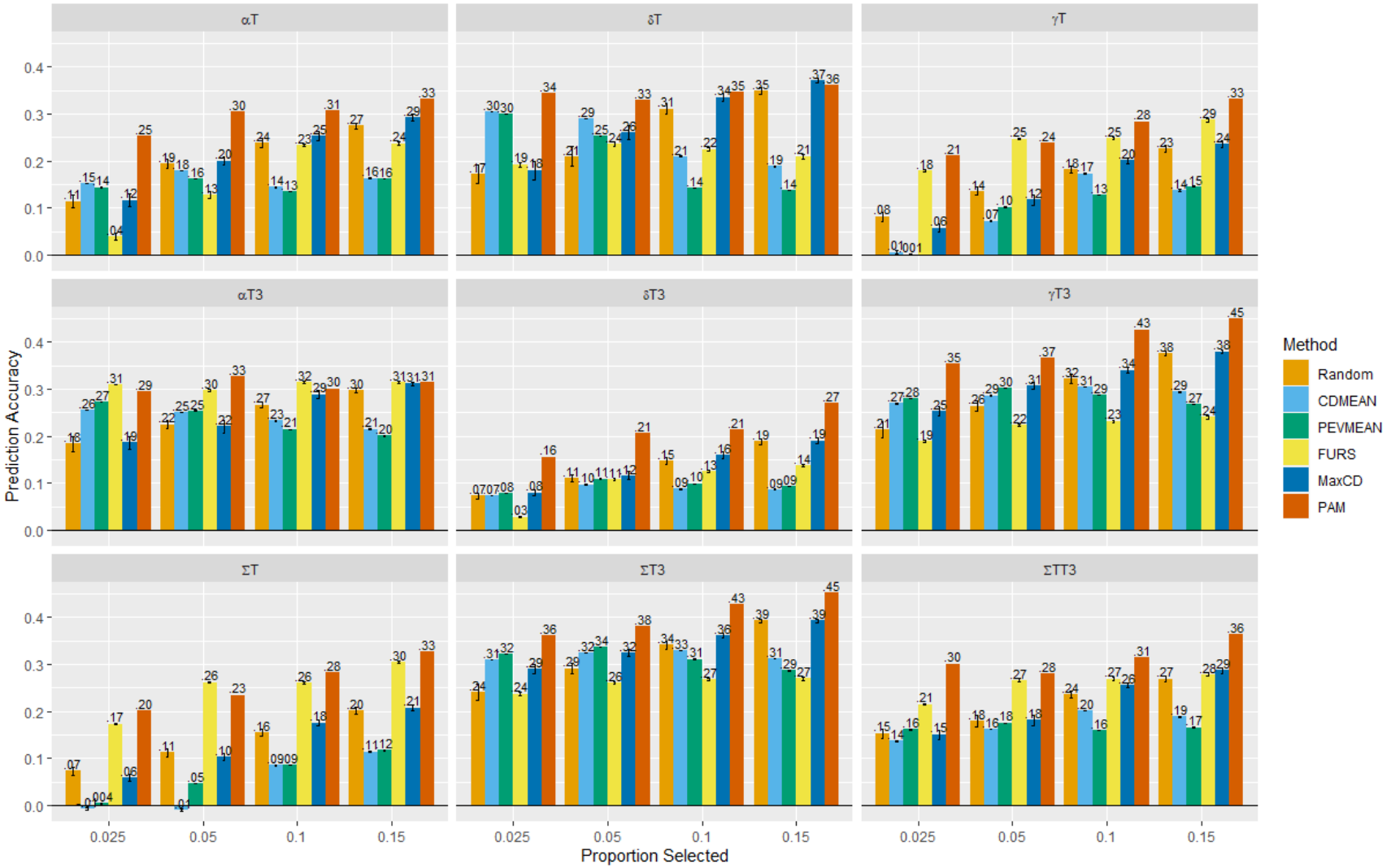


Figure S8. OTP Scenario A results. Five different OTP design methods as well as random selection were used to select training sets of a given size (x axis) from the original Ames Diversity Panel (AP) training set, which was then used to predict the remaining AP training set lines. Error bars show standard error for prediction accuracy based on 50 replicates.

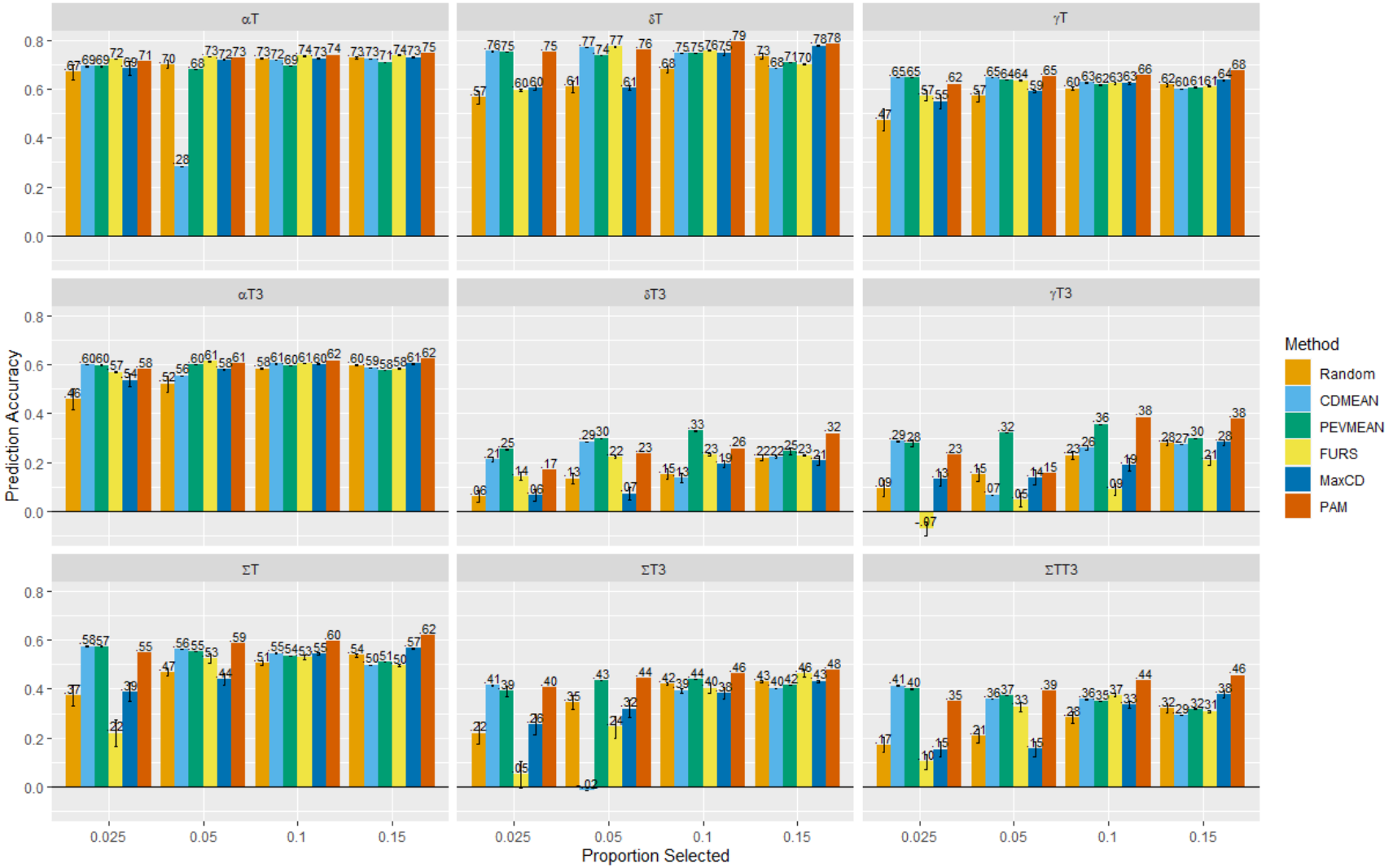


Figure S9. OTP Scenario B results. Five different OTP design methods as well as random selection were used to select training sets of a given size (x axis) from the available BGEM data, which was then used to predict the remaining BGEM lines. Error bars show standard error for prediction accuracy based on 50 replicates.

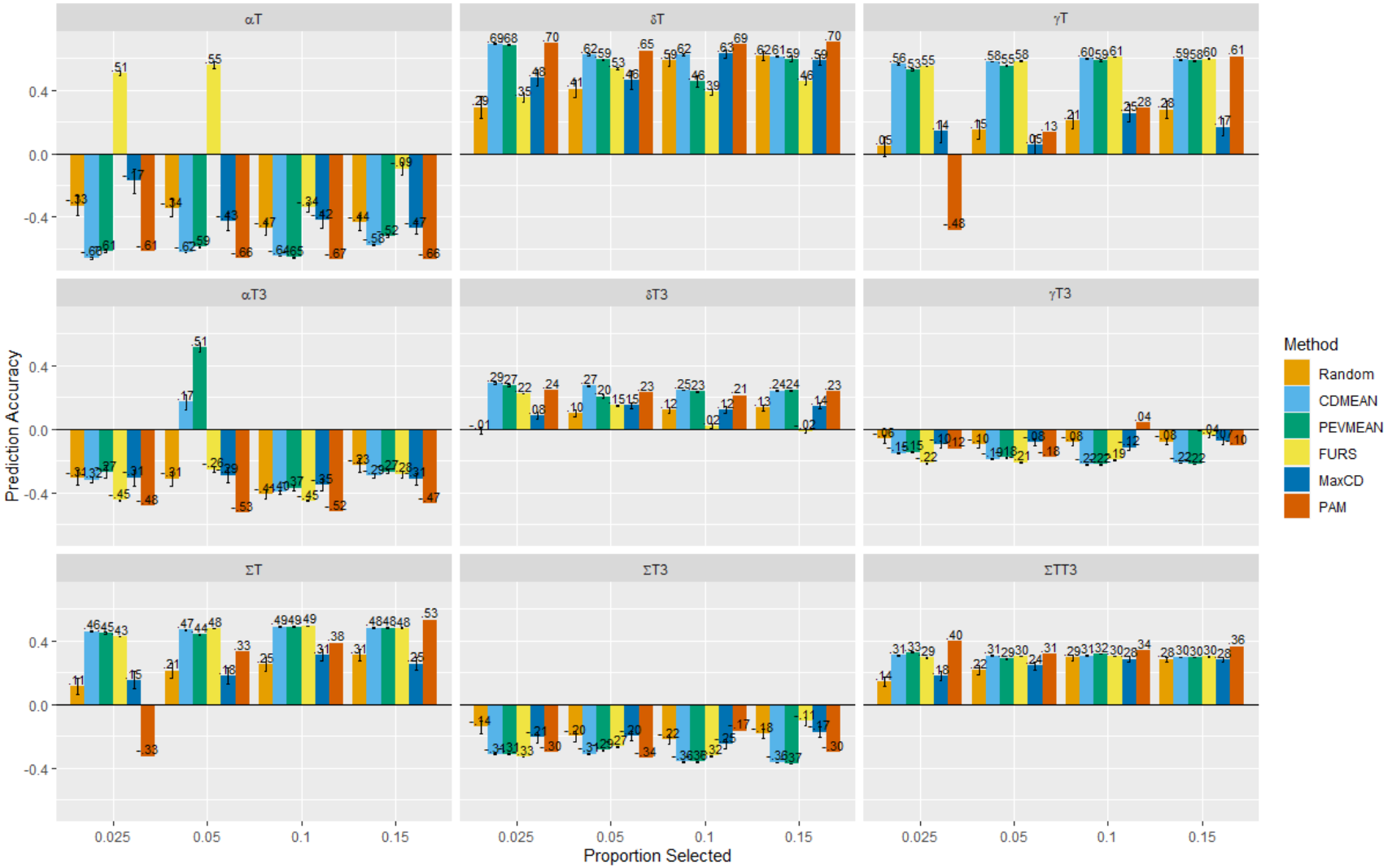


Figure S10. Five different OTP design methods as well as random selection were used to select training sets of a given size (x axis) from the AP full training set data, which was then used to predict the BGEM lines. Error bars show standard error for prediction accuracy based on 50 replicates.

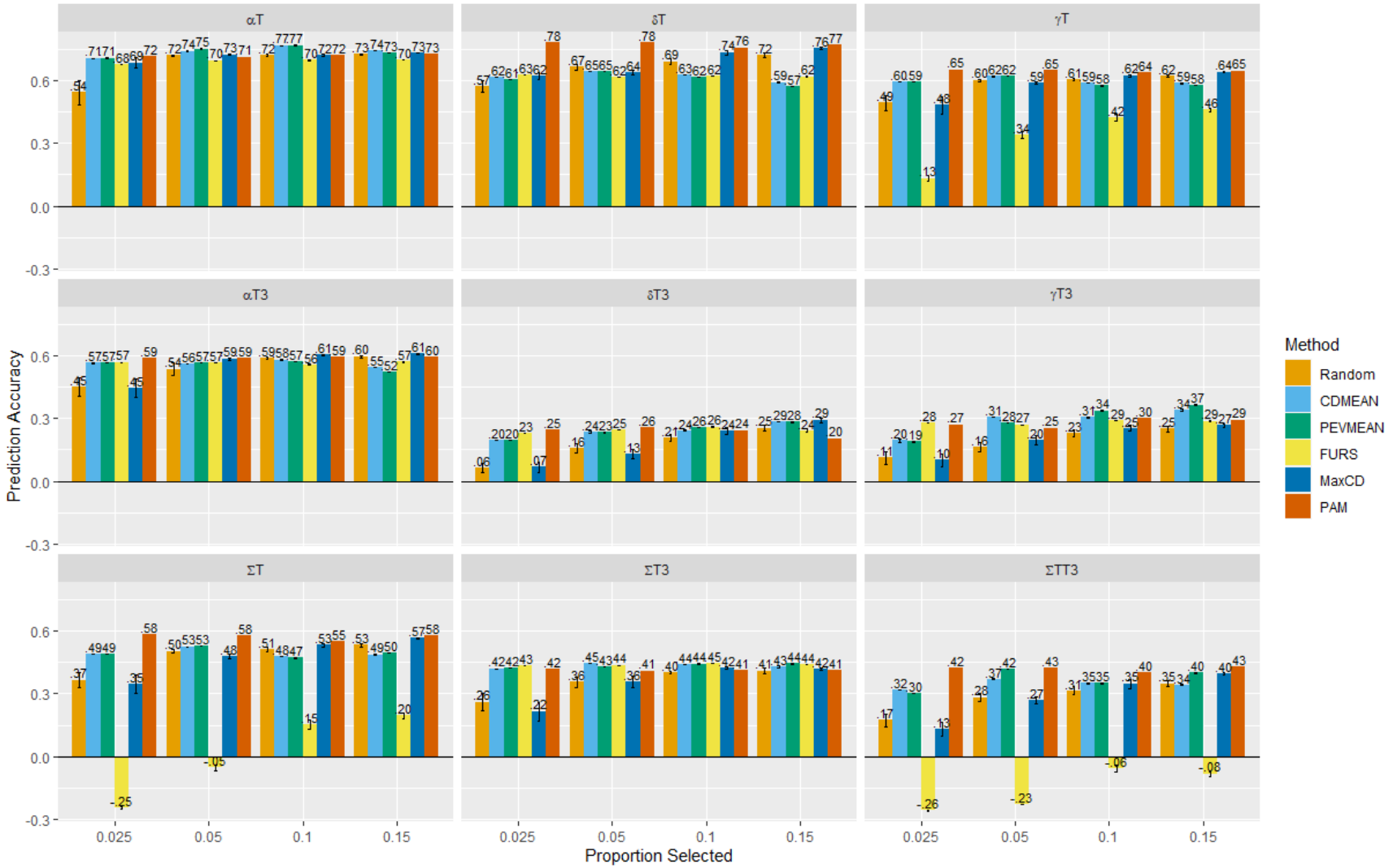


Figure S11. OTP Scenario E results. Five different OTP design methods as well as random selection were used to select training sets of a given proportion (x axis) of the combined AP full training set and BGEM data, which was then used to predict the remaining BGEM lines. Error bars show standard error of prediction accuracy based on 50 replicates.

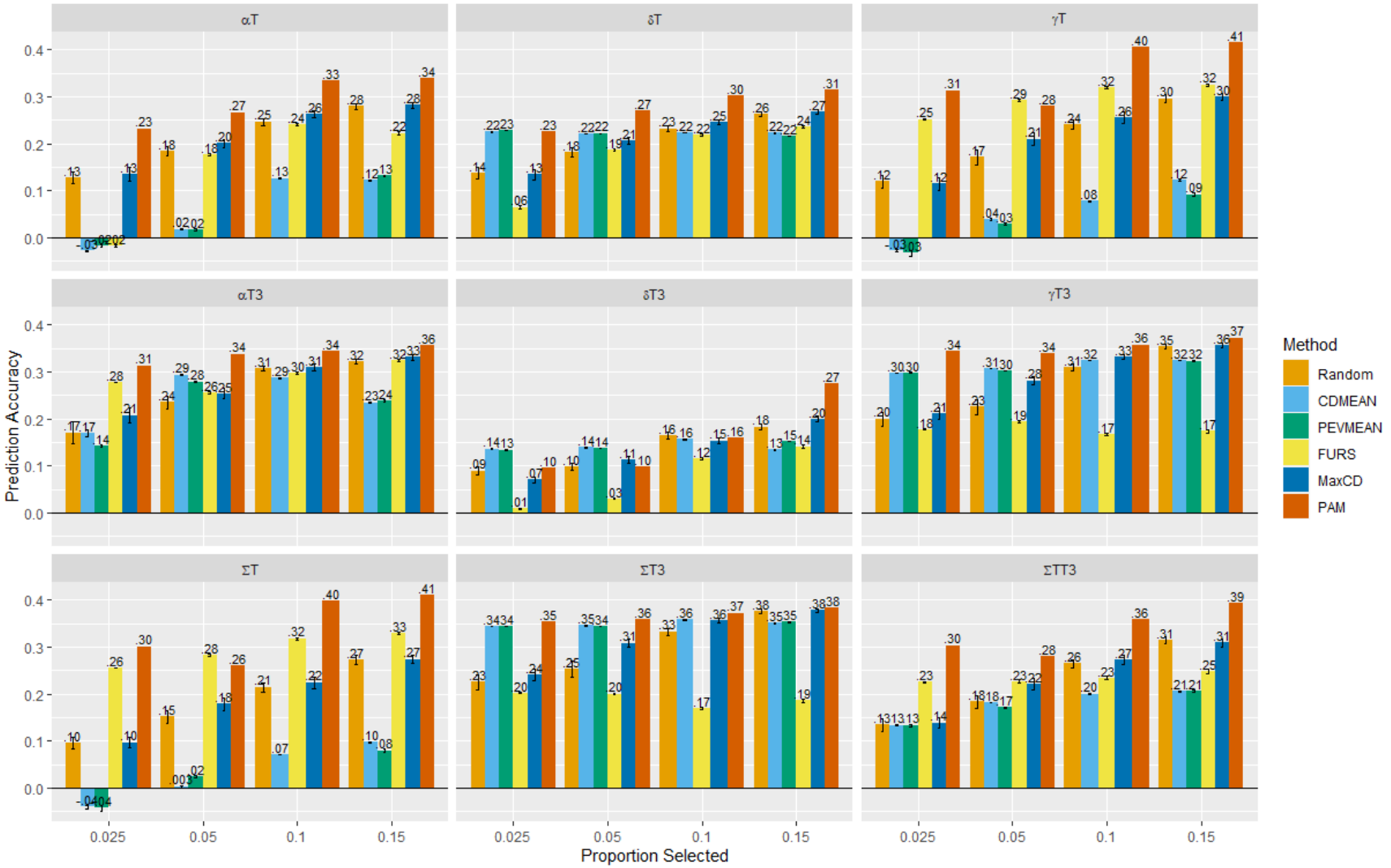


Figure S12. OTP Scenario C results. Five different OTP design methods as well as random selection were used to select training sets of a given proportion (x axis) of the combined AP full training set and BGEM data, which was then used to predict the AP validation set grown in 2015 and 2017. Error bars show standard error for prediction accuracy based on 50 replicates.

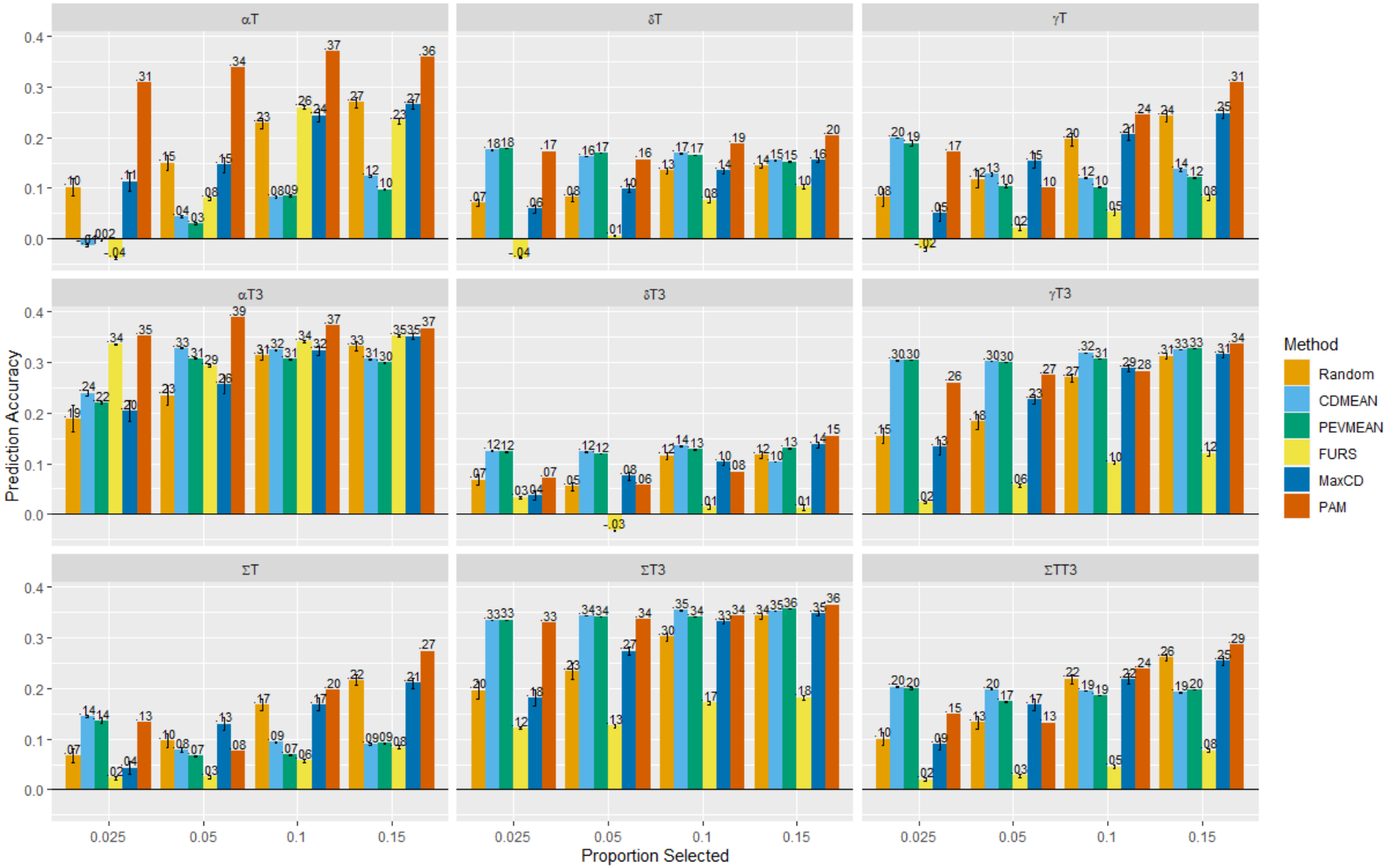


Figure S13. OTP Scenario D results. Five different OTP design methods as well as random selection were used to select training sets of a given proportion (x axis) of the combined AP full training set and BGEM data, which was then used to predict the AP validation set grown in 2018. Error bars show standard error for prediction accuracy based on 50 replicates.

Table S1: Phenotypic summary. Mean, standard deviation, minimum, and maximum BLUE values in μg per gram of dry seed for tocochromanol traits in the Ames Diversity Panel Training Set (AP) and BGEM.

| Population | Trait | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| AP | αT | 5.90 | 4.69 | -1.77 | 41.36 |
| BGEM | αT | 7.12 | 4.39 | 0.24 | 21.57 |
| AP | αT3 | 7.88 | 3.11 | 0.88 | 21.60 |
| BGEM | αT3 | 9.19 | 2.93 | 1.33 | 17.20 |
| AP | δT | 1.73 | 1.52 | -0.21 | 9.67 |
| BGEM | δT | 2.02 | 2.01 | -0.65 | 10.81 |
| AP | δT3 | 0.88 | 0.82 | 0.01 | 6.64 |
| BGEM | δT3 | 1.17 | 0.67 | 0.09 | 3.64 |
| AP | γT | 41.26 | 20.19 | 1.14 | 115.83 |
| BGEM | γT | 41.90 | 16.83 | 6.41 | 84.52 |
| AP | γT3 | 17.04 | 11.28 | 0.48 | 62.16 |
| BGEM | γT3 | 17.11 | 6.13 | 3.21 | 36.83 |
| AP | ΣT | 49.12 | 21.92 | 5.39 | 135.01 |
| BGEM | ΣT | 51.05 | 16.47 | 14.55 | 100.96 |
| AP | ΣT3 | 25.94 | 12.55 | 3.88 | 72.45 |
| BGEM | ΣT3 | 27.36 | 7.57 | 11.62 | 49.90 |
| AP | ΣTT3 | 75.44 | 26.77 | 19.00 | 170.52 |
| BGEM | ΣTT3 | 78.57 | 17.43 | 26.73 | 130.98 |

Table S2: Optimal number of principal components (PCs) selected for inclusion in the gBLUP model for prediction of BGEM using the AP subset as a training set. Based on the scree plot, 0 to 7 PCs were considered for inclusion in the model. Results are shown for PCs calculated from the combined genotype data from AP and BGEM (denoted A, see Fig. 2A) and from the AP genotype data only (denoted B, see Fig. 2B). Three methods were used: identifying the elbow in the scree plot (denoted Scree plot) (Cattell, 1966), a BIC-based model selection implemented in GAPIT (denoted BIC) (Wang & Zhang, 2020), and identifying the number of PCs that minimizes the mean square error (MSE) of predictions within the training set using ten-fold cross validation (denoted MSE) (Dadousis et al., 2014). For GAPIT, the model-selection procedure was run separately for each trait but the same result was found for all traits.

| Method | Trait | Optimal # PCs (A) | Optimal # PCs (B) |
|---|---|---|---|
| **Scree plot** | All | 5 | 7 |
| **BIC** | All | 0 | 0 |
| **MSE** | αT | 0 | 0 |
| **MSE** | αT3 | 0 | 0 |
| **MSE** | δT | 0 | 0 |
| **MSE** | δT3 | 0 | 0 |
| **MSE** | γT | 3 | 1 |
| **MSE** | γT3 | 0 | 1 |
| **MSE** | ΣT | 3 | 2 |
| **MSE** | ΣT3 | 1 | 1 |
| **MSE** | ΣTT3 | 3 | 6 |