# Utilizing Evolutionary Conservation Information to Improve Prediction Accuracy in Genomic Selection

**Jinliang Yang**[1], **Sofiane Mezmouk**[1,2], **Rita Mumm**[3], **and Jeffrey Ross-Ibarra**[1]

[1]Department of Plant Sciences, University of California, Davis, CA 95616, USA
[2]Current address: KWS SAAT AG, Grimsehlstr. 31, 37555 Einbeck, Germany
[3]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

## ABSTRACT

Genomic selection (GS) has gained popularity recently as the availability of genome-wide markers has increased. Current methods for GS weigh all the available SNPs equally in model training, without considering their functional differences. Genetic variations detected at evolutionary conserved sites tend to be deleterious and, thus, may be more informative for GS. To utilize this kind of information as a prior into the GS model, we proposed a method to put more weight on evolutionarily constrained sites. As a proof-of-concept, a half diallel population based on 12 diverse inbred lines was used, from which seven phenotypic traits were collected. Some of these traits show high levels of heterosis. After sequencing the 12 founder lines, about 14 million SNPs were discovered and the SNPs were used to identify 502,913 haplotype blocks shared through identity by descent (IBD). A five fold cross-validation experiment was conducted using the summary statistics of the SNP conservation scores, which were computed by evaluating sequences similarity of multiple divergent species, in the IBD blocks as explanatory variables. The results show that the prediction accuracies are significantly better than shuffled data with randomly assigned conservation scores. This study demonstrates the importance of incorporating evolutionary information in GS and its potential use in plant breeding.

## Objectives

- Study deleterious variants in a maize diallel population and their contributions to heterosis.
- Quantify IBD blocks by using evolutionary conservation information.
- Test whether it is useful to incorporate evolutionary conservation information in genomic selection.

## Materials and Methods

### Plant materials and phenotypic data

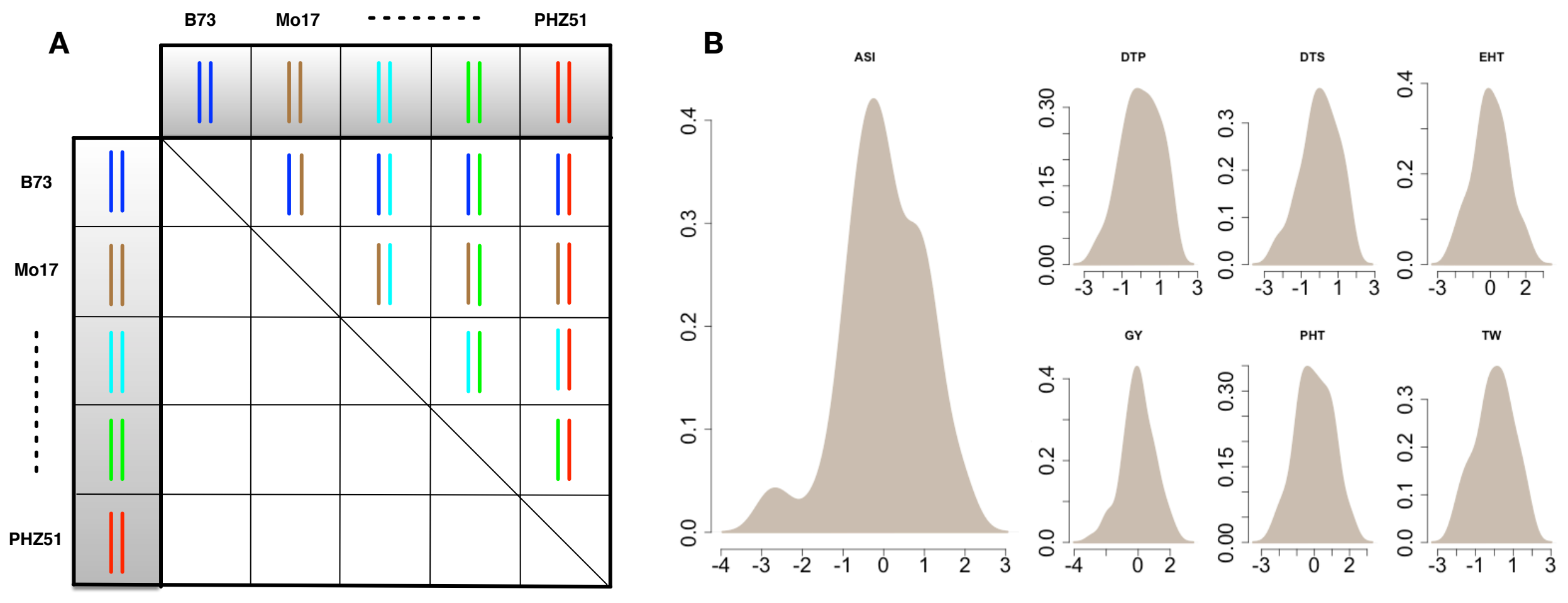

**Figure 1: Diallel experimental design and distribution of phenotypic data. (A)** Twelve maize inbred lines were selected and crossed in a half diallel. Ten of these (LH1, LH123HT, LH82, PH207, 4676A, PHG39, PHG47, PHG84, PHJ40, PHZ51) are proprietary inbreds that have expired from Plant Variety Protection (PVP) and represent the lineage of key heterotic germplasm pools used in present-day commercial corn hybrids. Two of them are important public inbreds, B73 and Mo17. **(B)** Phenotypic data were collected for anthesis-silking interval (ASI, in days), days to 50% pollen shed (DTP), days to 50% silking (DTS), ear height (EHT, in cm), grain yield adjusted to 15.5% moisture (GY, in bu/A), plant height (PHT, in cm), and test weight (TW, in pounds). Analyses were carried out on the traits per se as well as percent high parent heterosis (pHPH).

### Sequencing and SNP conservation annotation

All twelve inbreds were sequenced to an average depth of ∼10X. Reads were mapped to the maize B73 reference genome (AGPv2) with bwa-mem. After filtering, 13.8M SNPs were kept for further comparisons, including 1.9M genic SNPs and 361,280 in protein coding regions. Genome-wide deleterious variants were characterized by using genomic evolutionary rate profiling (GERP) (Davydov et al., 2010).

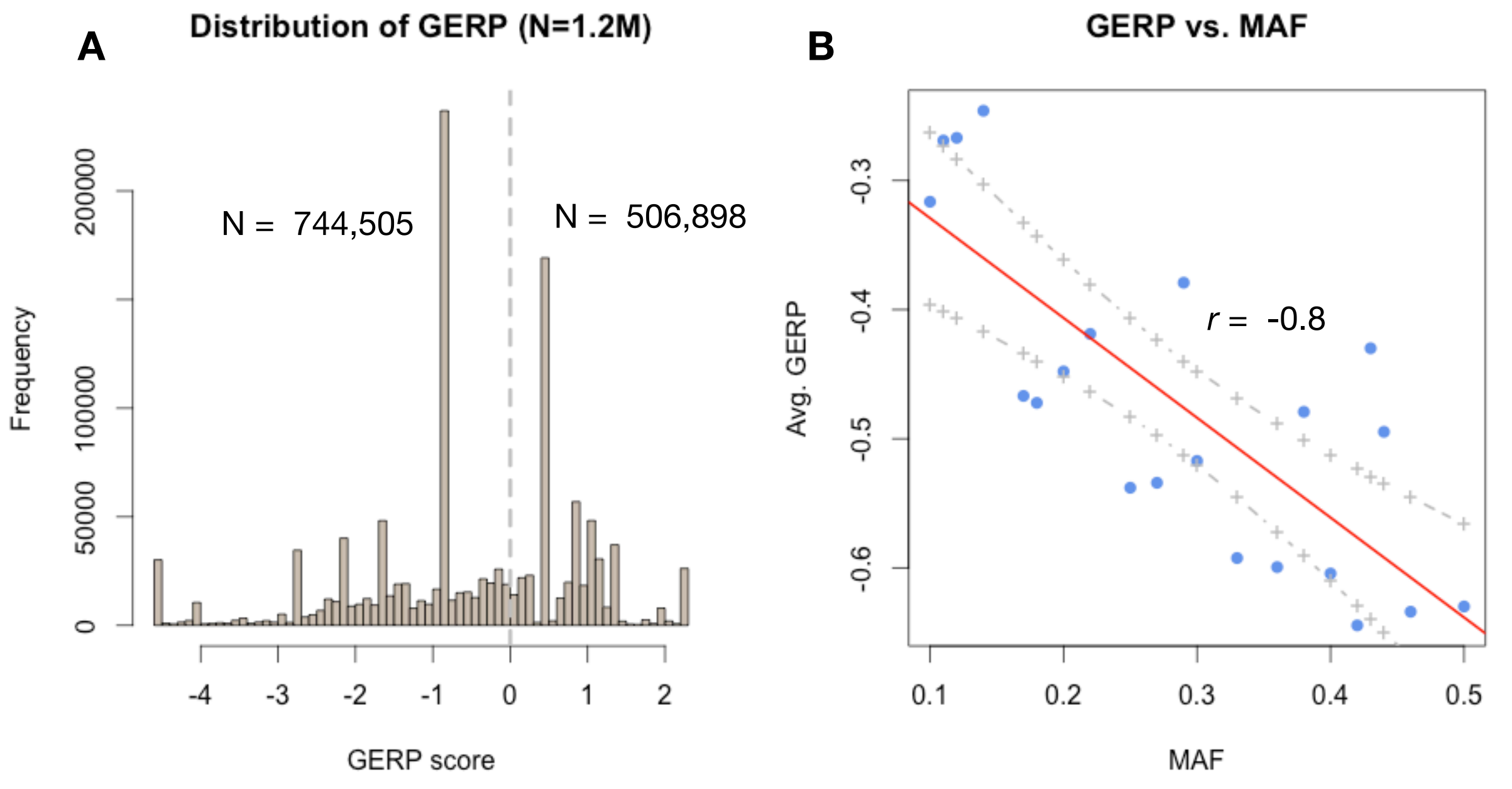

**Figure 2: GERP distribution of SNPs and relationship between GERP and minor allele frequency. (A)** GERP scores were obtained for ∼1.2 million (∼10%) SNPs. Of these, 506,898 (42%) were under evolutionary constraint and considered as deleterious variants. **(B)** Mean GERP scores were calculated for each bin (bin size = 0.01) of minor allele frequency (MAF). It shows that variants at conserved sites are maintained at low frequency. The red line and grey lines define the regression and its 95% confidence interval.

## Acknowledgements

Funding for this project comes from the USDA-AFRI, USDA Hatch, DuPont Pioneer, and N2 Genetics. We would like to thank Eli Rodgers-Melnick for allowing us to use his GERP rate estimations.

### Computing conservation score for IBD blocks

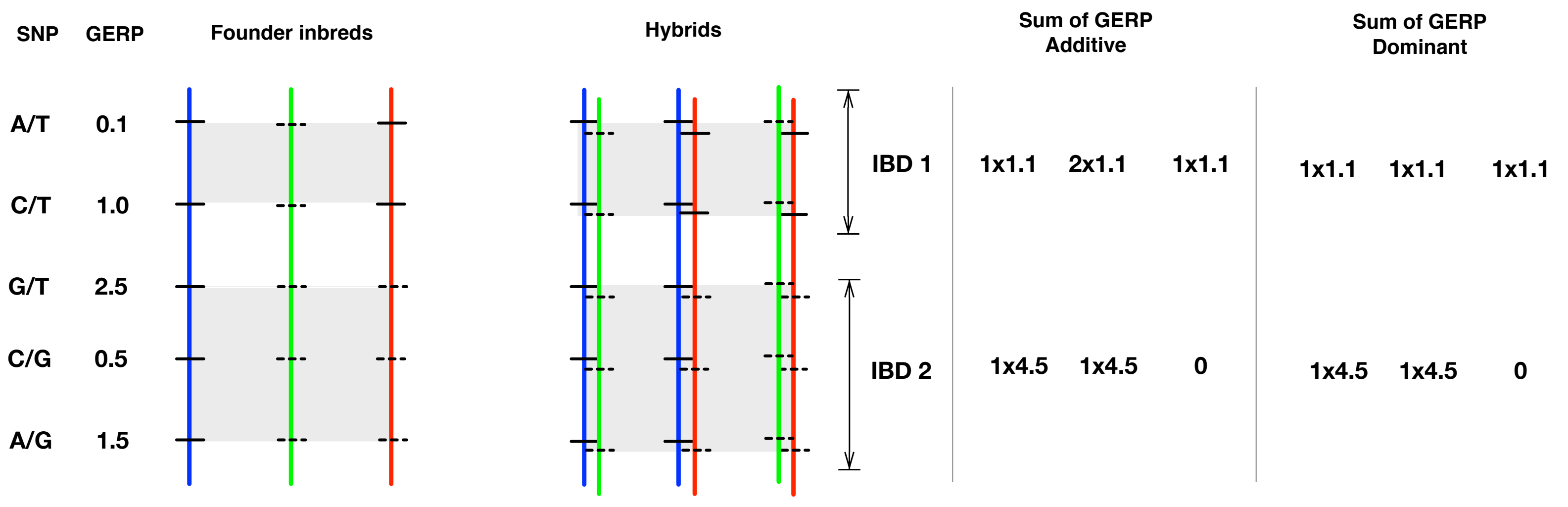

**Figure 3: Incoporation of conservation information into IBD blocks.** Regions of the genome that are identical by descent (IBD) among the 12 inbreds were identified using Beagle (Browning and Browning, 2009). The GERP scores of SNPs in an IBD block were summed under both additive and dominant models. Under the additive model, 2 x GERP score was assigned to genotypes homozygous for the non-reference allele, 1 x GERP score was assigned to heterozygotes, and 0 was assigned to the homozygous reference genotype. Under the dominant model, 1 x GERP score was assigned to both genotypes with a nonreference allle and 0 to the homozygous reference genotype.

### Genomic selection models

A Bayesian-based approach (BayesC) (Habier et al., 2011) was employed for the genomic selection (GS) experiments. To estimate predict accuracies, the diallel population was randomly divided into training and validation sets for 10 times using a 5-fold cross-validation method. Circular permutations were used both considering all SNPs or considering only genic SNPs to control for differences between genic and nongenic regions.

## Results

For traits *per se*, model prediction accuracies were significantly improved for ASI, DTS and PHT when incorporating GERP score information under the additive model. Prediction accuracies were significantly improved for ASI, DTP, GY and TW under the dominant model. In general, the average prediction rates are higher using the additive model ($r$ = 0.8) than the dominant model ($r$ = 0.7). For heterosis, incorporation of GERP scores only improved grain yield prediction and only under a dominant model.

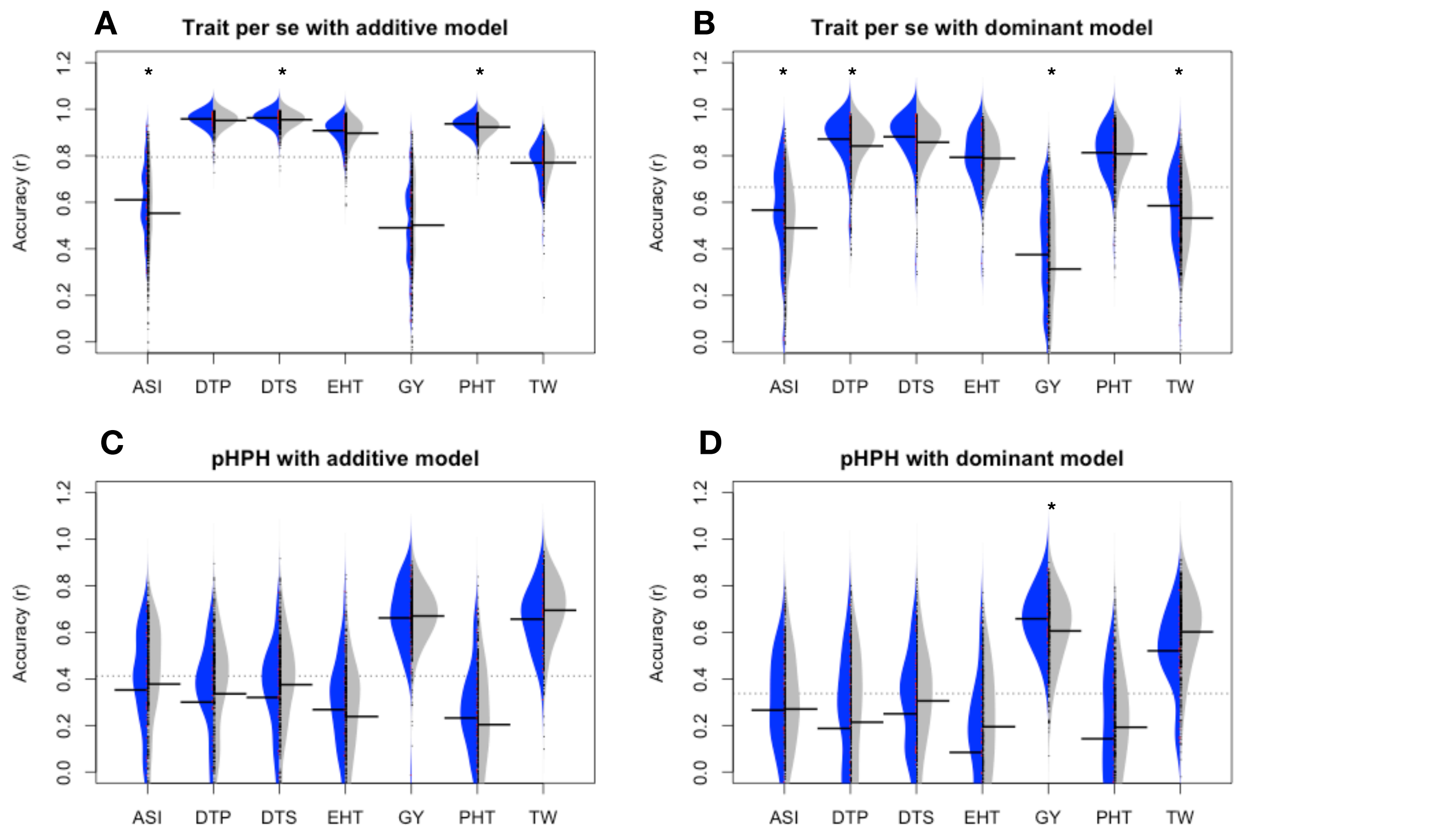

**Figure 4: Beanplots of cross-validation accuracies.** Cross-validation experiments were conducted using genic SNPs and circular shuffled data from the same set of the genic SNPs for traits *per se* (**A, B**) and pHPH (**C, D**) under additive (**A, C**) and dominant (**B, D**) models. Accuaries from the real data were plotted on the left side of the bean (blue) and permutation results plotted on the right (grey). Horizotal bars on beans indicate mean accuracies. The grey dashed line indicates the overall average accuracy. Stars indicate significantly improved cross-validation accuracies.

## Conclusions

- More than 500,000 deleterious SNPs were identified in elite maize lines including in noncoding regions of the genome.
- A genomic selection pipeline was developed, which utilized evolutionary conservation information in the model.
- Cross-validation results suggested prediction accuracies for some traits could be significantly improved by incorporating GERP scores.

## References

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84(2):210–23.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1):186.