# 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

- To study deleterious variants in a maize diallel population and their contribution 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 material and phenotypic data

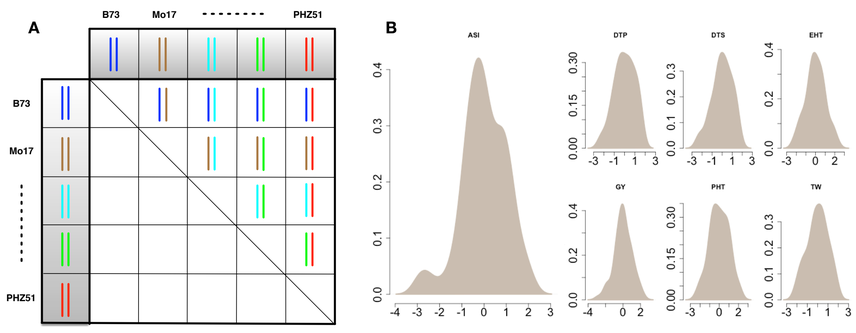


**Figure 1: Diallel experimental design and distribution of phenotypic data.** (A) Twelve maize inbred lines were selected and crossed in a partial diallel fashion without considering reciprocal effects. Ten of these inbreds (LH1, LH123HT, LH82, PH207, 4676A, PHG39, PHG47, PHG84, PHJ40, PHZ51) are proprietary 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, and two are predominant public inbreds B73 and Mo17. (B) Phenotypic data was 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).

### Sequencing and SNP conservation annotation

DNA from the twelve inbred lines was isolated and sequenced to an average coverage of 10X. Read pairs, kept after filtering, were mapped to the maize B73 reference genome (AGPv2) with bwa-mem. Reads, with mapping quality (MAPQ) higher than 10 and with a best alignment score higher than the second best one, were kept for further analyses. After filtering 13,782,809 are kept for further comparisons, including 1,909,416 genic SNPs and 361,280 in protein coding regions). Genome-wide deleterious variants were characterized by using genomic evolutionary rate profiling (GERP) [1, 2]. For each position of the multiple alignment, GERP scores were obtained by computing the regected substitutions subtracted by the neutral rate in a set of sequenced plant genomes.

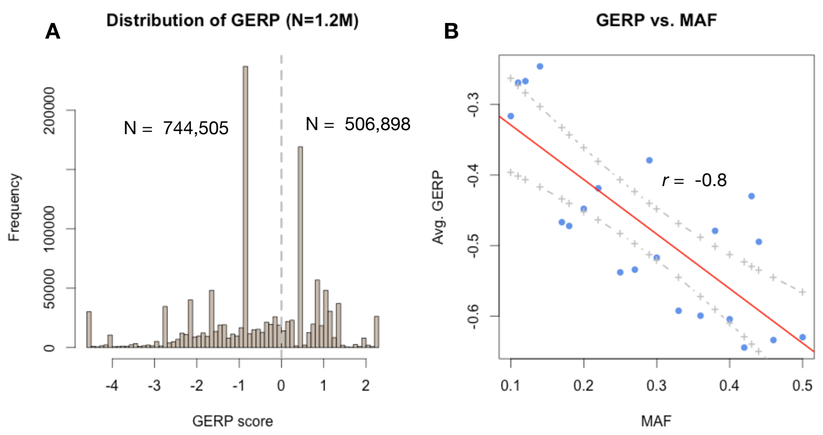


**Figure 2: Conservation scores distribution of diallel SNPs and relationship between GERP and minor allele frequency.** (A) Among the ∼14 million SNPs identified in the diallel population, ∼1.2 million (∼10%) sites could obtain their evolutionary conservation information using GERP. Of these sites, 506,898 (42%) are evolutionary constraint; variants at these sites were considered as deleterious. (B) Mean GERP scores were calculated for each bin (bin size = 0.01) of minor allele frequency (MAF); a regression analysis indicated that variants at conserved site tend to be maintained in a low frequency. Red line denotes the regression line and grey lines define its 95% confidence interval.

### Computing conservation score for IBD blocks

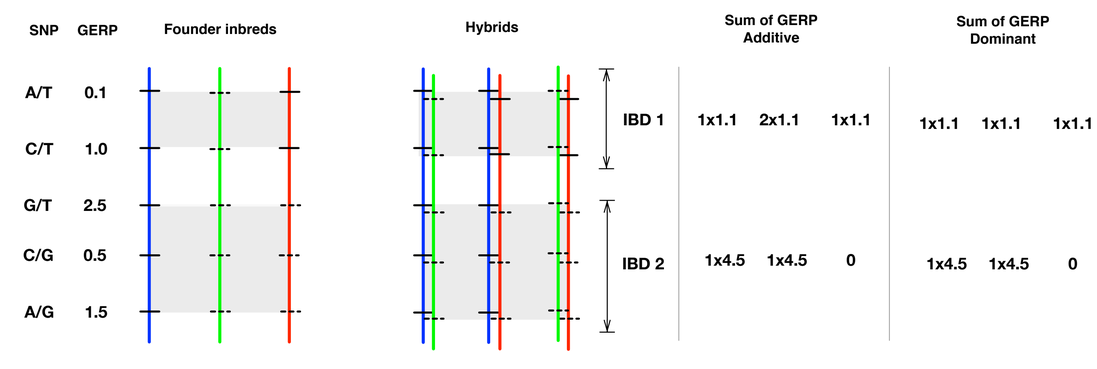


**Figure 3: Incoporation of conservation information into IBD blocks.** SNPs in a IBD block were added up with their GERP scores as the conservation estimates of the IBD block. This estimation was calculated with both additive and dominant models. Under the additive model, 2 x GERP score was assigned to the homozygous loci with non-reference SNP calls; 1 x GERP score was assigned to the heterozygous loci; and 0 was assigned to the homozygous loci with reference SNP calls. Under the dominant model, 1 x GERP score was assigned to both the homozygous loci with non-reference SNP calls and heterozygous loci; 0 was assigned to homozygous loci with reference SNP calls.

## Results

A haplotype based genomic selection strategy was conceived by using the IBD blocks as the explanatory variables. SNP loci with GERP score 0 were considered as evolutionary conserved sites and genomic variations detected at these sites were thought to be deleterious. After coding the summary statistics of the IBD with SNPs' GERP scores, a Bayesian-based approach (BayesC) [] was employed for the genomic selection experiments. To rule out the possibility that SNPs with high GERP scores tend to be enriched in the genic regions, To rule out the arguement that the prediction were taken advantage of the more informative genic SNPs. We conducted the real and shuffled experiment with a same set of genic SNPs. Therefore, the controlled variables in the circular shuffled data are the GERP score only.

As a result, we found that for the trait per se, the prediction accuracies were singificantly improved for ASI, DTS and PHT with the additive model. Prediction accuracies were significantly improved for ASI, DTP, GY and TW with the dominant model. In general, the average prediction rates are higher using additive model (r = 0.8) than dominant model (r = 0.7). For the heterosis transformation traits, for example percent high parental heterosis (pHPH), only one prediction rate was significantly increased for the GY heterosis.

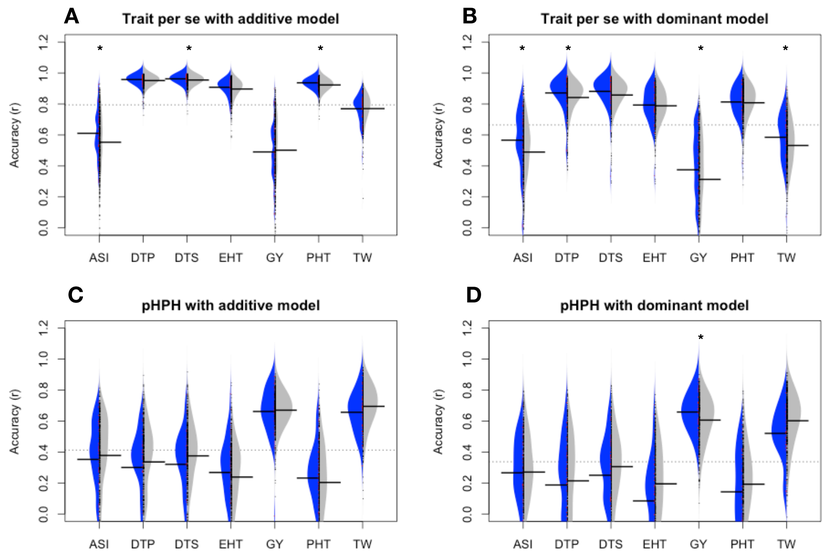


**Figure 4: Beanplots of the prediction accuracies.** The cross-validation was conducted using genic SNPs and circular shuffled data from the same set of the genic SNPs for trait per se with additive model (A), trait per se with dominant model (B), pHPH with the additive model (C) and pHPH with the dominant model (D). Accuaries drived from the real data were plotted on the left side of the bean (blue) and accuracies derived from the circular shuffled data were plotted on the right side of the bean (grey). Horizotal bars on beans indicate the mean accuracies. Grey dashed line indicate overall average. Stars on top of the beans indicate significantly improved cross-validation accuracies.

## Conclusions

- Large number (N=506,898) of deleterious alleles were identified in elite maize lines. The GERP enabled us to identify deleterious alleles beyond the protein coding region.
- A genomic selection pipeline was developed, which utilized evolutionary conservation information in the model.
- Cross-validation results suggested the prediction accuracies could be significantly improved for some of the traits with additive and dominant models.

## References

[1] Gregory M Cooper, Eric a Stone, George Asimenos, Eric D Green, Serafim Batzoglou, and Arend Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7):901–13, July 2005.

[2] Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025, January 2010.

[3] Sofiane Mezmouk and Jeffrey Ross-Ibarra. The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda, Md.)*, 4(January):163–71, 2014.

## Acknowledgements

Thanks for the funding!