Utilizing evolutionary conservation information to improve prediction accuracy in genomic selection

Jinliang Yang[*,1], **Sofiane Mezmouk**[*,1,2], **Andy Baumgarten**[†], **Rita H. Mumm**[‡] **and Jeffrey Ross-Ibarra**[*,§,3]

[*]Department of Plant Sciences, University of California, Davis, CA 95616, USA, [§]Center for Population Biology and Genome Center, University of California, Davis, CA 95616, USA, [†]DuPont Pioneer, Johnston, IA 50131, USA, [‡]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ABTRACT 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. Here we take advantage of evolutionary measure of sequence conservation to ask whether sites with prior evidence of functionality can inform GS models. We tested this idea using a partial diallel cross of 12 maize inbred lines. We sequenced the genomes of the parents of the diallel and phenotyped them for seven phenotypic traits across **X**environments in **X**years. We made use of an identity-by-decent analysis of the parents to identify haplotype blocks, and scored blocks in hybrids using a weighted sum of the GERP conservation score. Incorporating sequence conservation improves prediction accuracies in a five-fold cross-validation experiment for several traits *per se* as well as heterosis for those traits. 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 for some traits *per se* and heterosis transformations 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 usage in plant breeding.

KEYWORDS genomic selection; diallel; GERP; deleterious; heterosis; maize

The phenomenon of heterosis has been observed across many species, from yeast (**?**) to vertebrates (**?**). Recent studies indicated that the complementation of the deleterious alleles, which fit classic dominant model, may play an important role in determining heterosis (**?**). Deleterious alleles were arisen from new mutations during meiosis. In maize, about 90 new mutations were generated per meiosis (**?**), majority of which were deleterious according to empirical estimates (**?**). In a natural outcross population, the negative effects on fitness of these deleterious alleles make them subject to be selection against, which lead the deleterious alleles to be maintained in a low frequency (**?**). But the deleterious alleles could not be completerly purged.

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Monday 4th May, 2015%
[1]Department of Plant Sciences, University of California, Davis, CA 95616, USA. Email: rossibarra@ucdavis.edu
[2]These authors contributed equally to this work
[3]Current address: KWS SAAT AG, Grimsehlstr. 31, 37555 Einbeck, Germany

In maize, the total number of mildly deleterious mutations is substantial because of the exponential growth of population size after domestication. The modern breeding probably aims to remove these deleterious mutations and pyramiding beneficial alleles for agronomical purposes. In practice, the relatively homogeneous maize germplasm pool was artificially divided into different heterotic groups (**?**). It enabled the improvement of germplasm pools to be conducted in a parallel fashion, and therefore, facilitated the breeding efficiency. Using this hybrid breeding approach, the maize yield has been steadily improved since the early 20th century (**?**). However, removing deleterious mutations in low recombination regions or in tightly linked regions become less effective. Studies indicated that residual heterozygosity correlates negatively with recombination (**??**) and the low recombination is effective over long period of time (**?**). As a consequence, the deleterious alleles would be accumulated in the low recombination regions, such as the pericentromeric regions in maize, and the vigorous performance could be realized by combining two sets of non-deleterious or beneficial

alleles in repulsion state, thus lead to pesudo-overdominance. A recent QTL study identified loci controlling for heterosis are enriched in centromeric regions (**?**), which partly support this pesudo-overdominance hypothesis.

Despite the importance of deleterious alleles in contributing to heterosis, they have not been systematically investigated probably because of their low frequencies in the population and mostly exhibiting minor effects. Here, we employed a genomic selection (GS) approach to simultaneously estimate genome-wide deleterious variants in a half diallel population. The diallel population was composed of a set of hybrids, which enabled us to explore different modes of inheritance of the deleterious variants. And the study can be conducted with millions of variants but using relative little sequencing efforts. In our previous study, deleterious SNPs were found to be enriched in a SNP set identified by GWAS (**?**). The deleterious variants in the study were defined as non-synonymous mutations in the coding regions. Clearly, deleterious variants are not limited to coding regions. Here, we expanded the characterization of deleterious variants to genome-wide using genomic evolutionary rate profiling (GERP) (**?**). By incorporating GERP information in GS models, we demonstrated the prediction accuracies were significantly improved not only for some traits *per se*, but aslo for some heterosis transformations (especially for traits exhibiting high levels of hereosis). Further studies indicated that joint effects of deleterious alleles with additive and dominant modes of inheritance may contribute to heterosis.

Materials and Methods

Plant material and phenotypic data

Twelve maize inbred lines were selected and crossed in a half diallel fashion without considering reciprocal effects. Using this half diallel population, a field experiment was conducted in an incomplete block design with three replications. In the experiment, in addition to the 66 F1 hybrids, the 12 inbred parents and two current commercial check hybrids were planted; and hybrids and inbreds were grouped separately. The plants were grown at Urbana, IL in 2009, 2010 and 2011. Plots consisted of four rows, with all observations taken from the inside two rows to minimize effects of shading and maturity differences from adjacent plots. Both inbred lines and the 66 resulting hybrids were field evaluated. Phenotypic data was collected for plant height (PHT, in cm), ear height (EHT, in cm), days to 50% silking (DTS), days to 50% pollen shed (DTP), anthesis-silking interval (ASI, in days), grain yield adjusted to 15.5% moisture (adj GY, in bu/A), and test weight (TW, in pounds).

Best Linear Unbiased Estimation (BLUE) of the genetic effects were calculated with ASReml-R (**?**) following the linear model:

$$Y_{ijkl} = \mu + \varsigma_i + \delta_{ij} + \beta_{jk} + \alpha_l + \varsigma_i \cdot \alpha_l + \varepsilon$$

where Y_{ijkl} is the phenotypic value of the l^{th} genotype evaluated in the k^{th} block of the j^{th} replicate within the i^{th} environment; μ, the overall mean; ς_i, the fixed effect of the i^{th} environment; δ_{ij}, the fixed effect of the j^{th} replicate nested in the i^{th} environment; β_{jk}, the random effect of the k^{th} block nested in the j^{th} block; α_l, the the fixed genetic effect of the l^{th} individual; $\varsigma_i \cdot \alpha_l$, the interaction effect of the l^{th} individual with the i^{th} environment; ε, the model residuals.

Heterosis for each hybrid was then estimated by better-parental heterosis (BPH):

$$BPH_{min,ij} = \hat{G_{ij}} - min(\hat{G_i}, \hat{G_j})$$

$$BPH_{max,ij} = \hat{G_{ij}} - max(\hat{G_i}, \hat{G_j})$$

where $\hat{G_{ij}}$, $\hat{G_i}$ and $\hat{G_j}$ are the genetic values of the hybrid and its two parents i and j. BPH_{min} was used instead of BPH_{max} for days to anthesis.

Sequencing and SNP calling

DNA from the twelve inbred lines was CTAB extracted (**?**) and Covaris sheared for Illumina library preparation. The DNA libraries were then sequenced to an average coverage of 10X.

Raw paired reads (reverse and forward for each sequence), were trimmed for adapter contamination with Scythe package (https://github.com/vsbuffalo/scythe) which calculate the probability of having a contamination given the adapter sequence, the number of mismatches and sequence quality. The reads were then trimmed for quality and sequence length ($\geq$ 20 nucleotides) with Sickle package (https://github.com/najoshi/sickle) .

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.

Single nucleotide polymorphisms (SNPs) were called with *mpileup* function from samtools utilities (**?**). To deal with paralogy, which is a major problem in maize sequence mapping (**?**), all SNPs were filtered to a) be heterozygote in less than 3 inbred lines, b) have a mean minor allele depth over all genomes of at least 4, c) have a mean depth over all individuals lower than 30 and d) have missing/heterozygote alleles in less than 6 inbred lines (allelic information for at least 15 hybrids in the partial diallel design).

Haplotype identification and SNP annotation

All missing alleles were then imputed with BEAGLE package (**?**) and identity by descent (IBD) regions between the 12 inbred lines were identified with BEAGLE's fastIBD method (**?**). The pairwise IBD region starts and ends were used to delimit haplotipic blocks where several inbred lines shared a homogenous haplotypes.

Genomic evolutionary rate profiling (GERP) (**?**), which estimates the evolutionary constraint by quantifying substitution deficits after multiple genome alignments, was obtained from (**?**) for AGPv2.

Genomic selection using IBD blocks incorporated with GERP scores

A haplotype based genomic selection (GS) strategy was conceived by using the IBD blocks as the explanatory variables, where the IBD blocks were coded with evolutionary conservation information. The reference genome sites with GERP score >0 were considered as conserved sites and genomic variants detected at these sites were determined as candidate deleterious alleles. To incorporate the conservation information into IBD blocks, SNPs falling into a given IBD block were added up using their estimated GERP scores. The sum of the score for a given IBD block were used as the conservation estimates of the IBD block. This estimation was calculated using a python script gerpIBD (https://github.com/RILAB/pvpDiallel) with both additive and dominant modes of inheritance. 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.

To conduct prediction, a 5-fold cross-validation method was used, where the diallel population was randomly divided into training (80%) and validation sets (20%) for 10 times. The BayesC option in the GS software package of GenSel4 (**?**) was used for the model training, where 41,000 chains of iteration were used and the first 1000 chains were removed as a burn-in. After model training of the seven phenotypic traits *per se* and their heterosis transformations, the prediction accuracies were obtained by comparing the predicted breeding values with the observed phenotypes in the corresponding validation sets. In addition, the GERP scores were circularly shuffled using 10-Mb window for 100 times as the null conservation estimates of the IBD blocks. Note that the cross-validation experiments using the circularly shuffled data were conducted on the same training and validation sets.

Results

Genetic values, heritability and heterosis transformations

A half diallel population was created using 12 maize inbred lines (Figure S1a). Two of them are important public inbreds, B73 and Mo17. And the other ten of them are proprietary inbreds (LH1, LH123HT, LH82, PH207, 4676A, PHG39, PHG47, PHG84, PHJ40, PHZ51) that have expired from Plant Variety Protection (PVP) and represent the lineage of key heterotic germplasm pools used in present-day commercial corn hybrids. The set is diverse enough to facilitate a broad sweep of the heterotic sub-groups that comprise U.S. commercial germplasm. From this population, phenotypic data were collected for seven traits of interest during 2009-2011. The phenotypic traits are 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).

The best linear unbiased estimators (BLUEs) for genotypes of the seven traits were derived from mixed linear models (Table S1). In the models, all fixed effects were significant (Wald test P value < 0.05) for all traits except ASI for which the effect of the replicates within environments were not significant. As shown in the Figure S1b, the BLUE values were normally distributed (normality test P values > 0.05). The broad sense heritability of the traits ranged from 0.65 for ASI to 0.95 for PHT. With the parental phenotypic data, we conducted heterosis transformations using better-parental heterosis (BPH). Because the selected inbred lines are commercial relevant and fairly elite in performance, hybrids in this population exhibit relative low hybrid vigor (overall mean percent BPH = 0.3% $\pm$ 0.4%) normalized by their better parental data for most of the traits except GY (mean percent BPH = 95% $\pm$ 16%, Figure S2). Finally, general and specific combining ability (GCA ans SCA) were estimated following (**?**). The GCA and SCA varied according to the traits (Table S2). For the GY, B73, PHG47 and PHG39 are the top three inbred lines that combining better with others.

Evolutionary constraint information for genomic variants

In this study, 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 of depth, heterozygosity and missingness, 13.8 million SNPs were kept, including 1.9 million SNPs in genic regions and 361,280 in protein coding regions. We estimated the allelic error rate by first comparing our genotype calls to those of 41,292 overlapping SNPs on the maize SNP50 bead chip (**?**); and then compared our SNP alleles for B73 and Mo17 with the 10,426,715 SNP previously identified in HapMap2 (**?**); finally, we compared our SNPs to 180,313 overlapping SNPs identified through genotyping by sequencing (GBS) (**?**). The comparisons showed 99.12% allele similarity with shared SNPs previously identified.

Evolutionary constraint information of GERP was obtained from (**?**), which was computed by using rejected substitution rate relative to the neutral rate after multiple genome squences alignment (**?**). This approach could estimate the conservation information at the base-pair level, extending the definition of deleterious variants to non-genic regions compared to previous approaches of SIFT (**?**) or MAPP (**?**). In AGPv2, more than 86 million bases (~4.2% of the maize genome) were detected as evolutionary constraint bases with GERP score >0. From genome-wide of view, genomic sequences are evolutionarily constraint near the telomeric regions and the mean GERP scores dropped, in general, toward centromeric regions (Figure S3) with some exceptions at long arm of chromosome 1 and 4.

Although genomic sites with GERP score >0 showed evidence of under evolutionary constraint, they are not immune to mutate. Indeed, in our diallel population, 506,898 SNP variants were detected at sites where the GERP scores are bigger than zero (Figure 1A). Consistent with previous study (**?**), the minor allele frequencies (MAFs) of these SNPs were negatively correlated with their GERP scores (Pearson's correlation test P value < 0.05, $r = -0.8$), indicating that the putative deleterious alleles tend to be purged and maintained in a low frequency in the population.

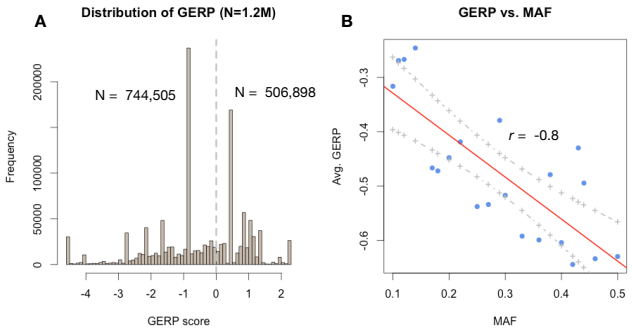

Figure 1 Distribution of GERP scores and relationship between GERP scores and MAFs. **(A)** GERP scores were extracted for ~1.2 million SNPs (~10% of the total SNPs) of the diallel population. The spikes in the histogram were because only limited genomes (N < 10? *ask for this number*) were used in GERP calculation. **(B)** Mean GERP scores were calculated for each bin (bin size = 0.01) of minor allele frequency (MAF). The red line and grey lines define the regression and its 95% confidence interval.

Incorporatipn of GERP information improved prediction accuracies

Genomic variants occurred at the evolutionary constraint sites were potentially deleterious. The phenotypic effects of these genetic loads and their contributions to heterosis become an interesting area to explore. However, the population size in

this study is relative small and SNPs detected at sites containing high GERP scores are generally in low frequencies. The statistical power to detect the separate effects of these putative deleterious alleles becomes very low. To alleviate the statistical limitations, we conceived a haplotype-based approach for GS, which could add up individual effects of deleterious alleles in IBD blocks and estimate these IBD blocks simultaneously. To conduct the analysis, first of all, 55,000 IBD blocks were identified, which had an average size of 44,980 bp (ranged from 36 to 10,320,000 bp). IBD blocks having > 1-kb in size and containing > 1 deleterious alleles (SNPs at sites with GERP scores >0) were kept for further analysis. Secondly, the GERP scores of SNPs in an IBD block were summed under both additive and dominant models. Those summed GERP scores on IBD blocks were considered as the measurements of the conservation of the haplotypes. More details of this procedure were illustrated in Figure S4.

A Bayesian-based statistical method (BayesC) (**?**) was employed for model training. Using a 5-fold cross-validation approach, the prediction accuracies of the real data and cicularly shuffled data were compared. As shown in Figure 2, for traits *per se*, prediction accuracies were significantly (FDR < 0.05) improved for ASI and PHT when incorporating GERP information in the IBD blocks under the additive model; prediction accuracy was significantly improved for ASI under the dominant model. For heterosis transformation traits (BPH), incorporation of GERP scores improved BPH of GY under the additive model and improved BPH of DTP, DTS and TW under the dominant model (Figure 2 C and D, Supplementary table 3). In general, the average prediction accuracis were higher using the additive model (mean r = 0.81 and 0.49 for traits *per se* and BPH) than the dominant model (mean r = 0.70 and 0.42). And the prediction accuracies decreased for predicting heterosis transformations (BPH) as compared to the predictions for traits *per se*.

It was argued that SNPs in genic regions might have higher GERP scores than those in non-genic regions. The circular shuffling permutations may shift the high GERP scores to non-genic regions. If that is the case, the approach tended to weigh more on genic SNPs. To rule out this possibility, we elected SNPs with GERP scores >0 in genic regions only and did the circular shuffling to assign GERP scores to the same set of the selected SNPs. By doing this, the method will not take advantage of genomic positional information any more. Noted that in this study less number of SNPs was selected (N = 316, 983). Nevertheless, model prediction accuracies were significantly improved for traits *per se* of GY under the additive model. For heterosis transformations, prediction accuracies were significantly improved for BPH of GY and PHT under the additive model and the prediction accuracy was significantly improved for pBPH of GY (Figure S5 and Table S4).

Posterior phenotypic variance explained and model comparisons

To learn why the prediction performace varied among traits *per se* and heterosis transformations, we obtained the posterior variance explained by our models using the complete set of data. As shown in Figure 3, additive models explained more phenotypic variance for traits *per se* of DTP, DTS, EHT and PHT; but explained less phenotypic variance for heterosis transformations (BPH) of ASI, GY and TW. On the contrary, lager proportions of the phenotypic variance could be explained by the dominant models for heterosis transformations (BPH) of ASI, GY and TW. For the GY in particular, 50% of the heterosis could be explained by dominant model.

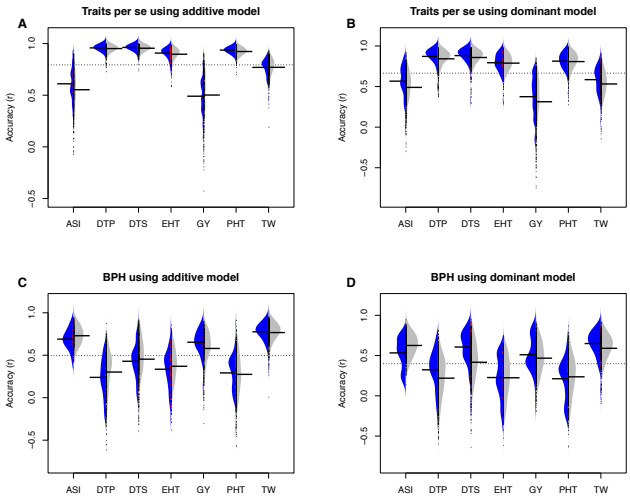

Figure 2 Beanplots of cross-validation accuracies using SNPs with positive GERP score. Cross-validation experiments were conducted using selected SNPs and circular shuffled data from the same set of SNPs for traits *per se* (**A, B**), HPH (**C, D**) and pHPH (**E, F**) under additive (**A, C, E**) and dominant (**B, D, F**) models. Accuraries 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 lines indicate the overall average accuracies. Stars indicate significantly improved cross-validation accuracies with FDR < 0.05.

Heterosis transformations were largely determined by the accuracies of the parental phenotypes. To take the uncertainty of the parental phenotypes into consideration, we estimated the combining abilities from the hybrid population itself to investigate which modes of inheritance perform better than the null models. We extracted the breeding values estimated with both additive and dominant models using the genome-wide IBD blocks incorporated with the GERP scores. Consistent with above analysis, IBD blocks coded with dominant mode of inheritance significantly (equation 1 vs. equation 2, ANOVA P value < 0.05) improved model fitting for ASI and GY. We also compared model 3 and model 4, ANOVA results indicated that 4 performed almost as good as 3, indicating that specific combining ability captured most of the parental interactions and the current method could not detect higher order of interactions.

$$Y_{ij} = \mu + GCA_i + GCA_j + \varepsilon \quad (1)$$

$$Y_{ij} = \mu + GCA_i + GCA_j + G_{ij} + \varepsilon \quad (2)$$

$$Y_{ij} = \mu + GCA_i + GCA_j + SCA_{ij} + \varepsilon \quad (3)$$

$$Y_{ij} = \mu + GCA_i + GCA_j + SCA_{ij} + G_{ij} + \varepsilon \quad (4)$$

where Y_{ij} is the BLUE value of the hybrid crossed between the i^{th} inbred and j^{th} inbred; μ, the overall mean; GCA_i, the general combining ability of the i^{th} inbred; GCA_j, the general combining ability of the j^{th} inbred; SCA_{ij}, the specific combining ability of between the i^{th} and j^{th} inbreds; ε, the model residuals.

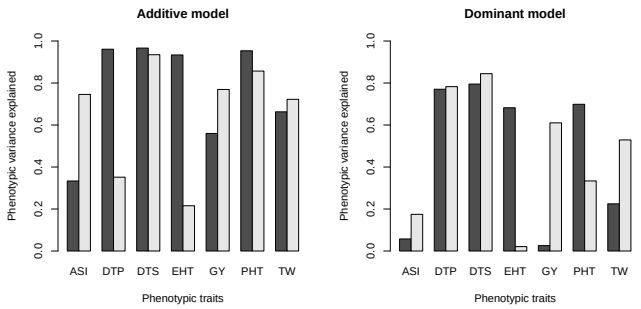

Figure 3 Barplots of the phenotypic variance explained by IBD blocks incorporated with GERP scores.

Discussion

- how do results match with heritability and heterosis?
- do we support deleterious model of Mezmouk et al.?

In this study, more than 500,000 deleterious SNPs were identified in elite maize lines including in non-coding regions of the genome. Majority of them were maintained in a low frequency, which consistent with the previous observation *Eli PNAS, 2015* and indicated the deleteriousness of the variants in the conserved sites. To estimate the joint effects of the potentical deleterious alleles for the phenotype, a genomic selection pipeline was developed, which utilized evolutionary conservation information in the IBD blocks as explainatroy variables. After model training, cross-validation results suggested prediction accuracies for some traits *per se* and heterosis transformations could be significantly improved by incorporating GERP scores.

Paritcally, with dominant model, up to 20% of the phenotypic variance could be explained for the heterosis traits. Theoritically, BPH transformation subtracts the joint effects of the additive and dominant alleles in the best parents as residule, the substantiall variance of these redidules explained by the additive or dominant models in our studies indicated that genetic components controlling for heterosis might in linked state. Note that the current model simply assume the traits were simply determined by complete additive or complete dominant. In the real case, the phenotype most likely be controlled by some degree of mixture of additive and dominant complenents.

Schmitt: How to explain the prediction difference? The variation of the prediction accuacies were relative large in this study. First of all, broad sense heritability of the traits are different. Second, from the simulation we learned that different traits may controlled by different proportion of additive, dominant and even recessive gene actions. Our naive model only built the pure additive and pure dominant effects in. For the more complicated cases, the models may not work very well.

Supporting Information

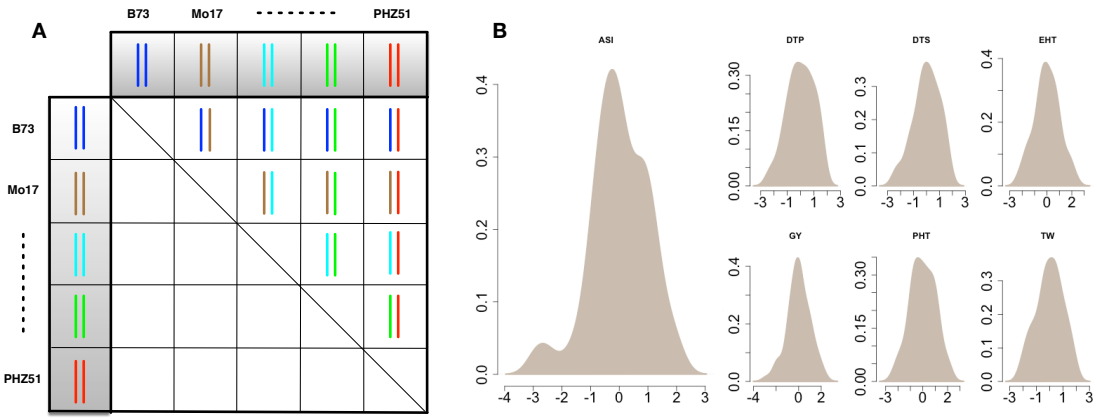

Figure S1 Diallel experimental design and distribution of phenotypic data. (A) Twelve maize inbred lines were selected and crossed in a half diallel. **(B)** Density plots of the phenotypic distributions.

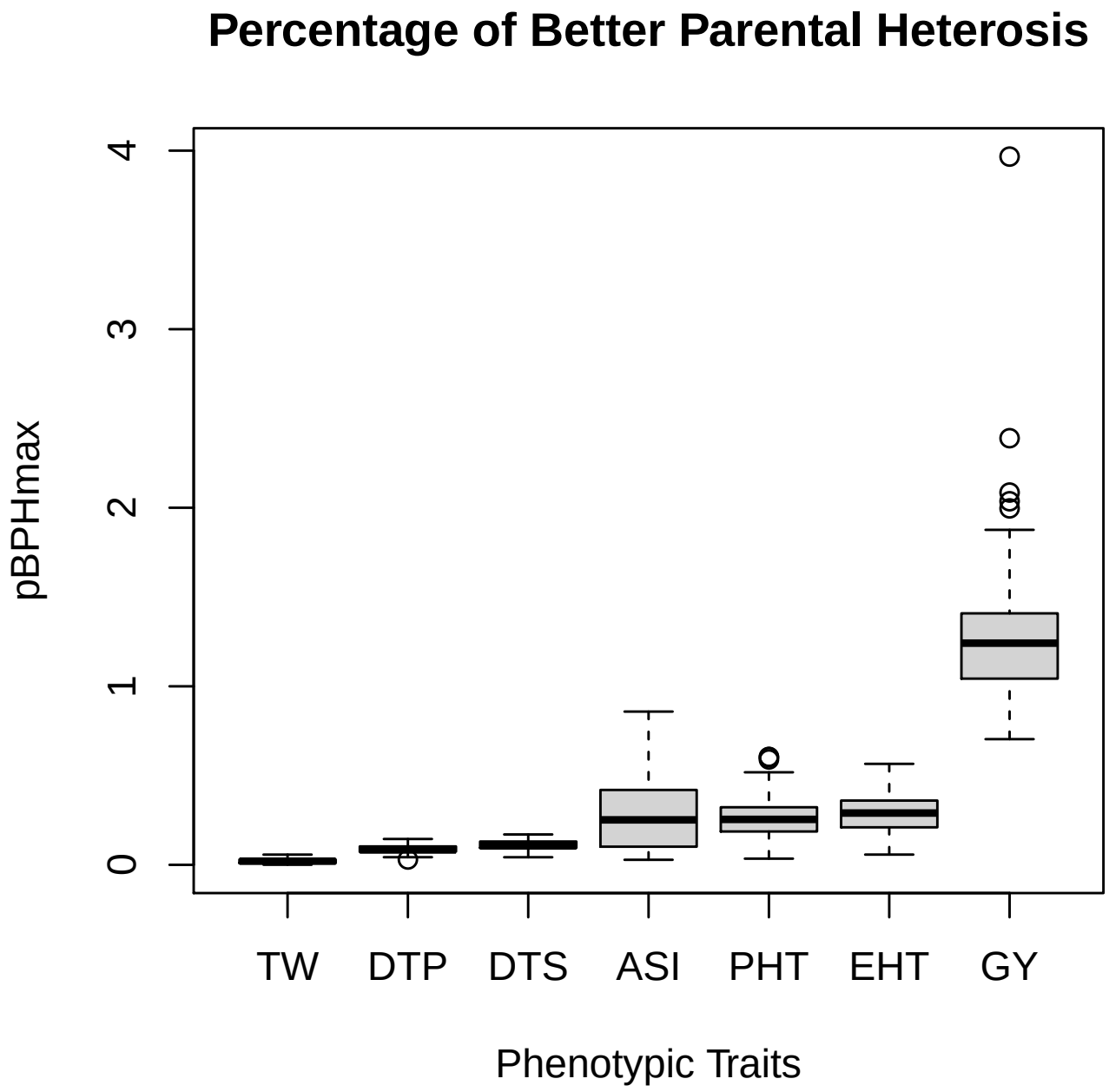

Figure S2 Boxplot of the percent better parental heterosis (pBPH). In the plot, ASI was calculated using pBPHmin and the other six traits were calculated using pBPHmax.

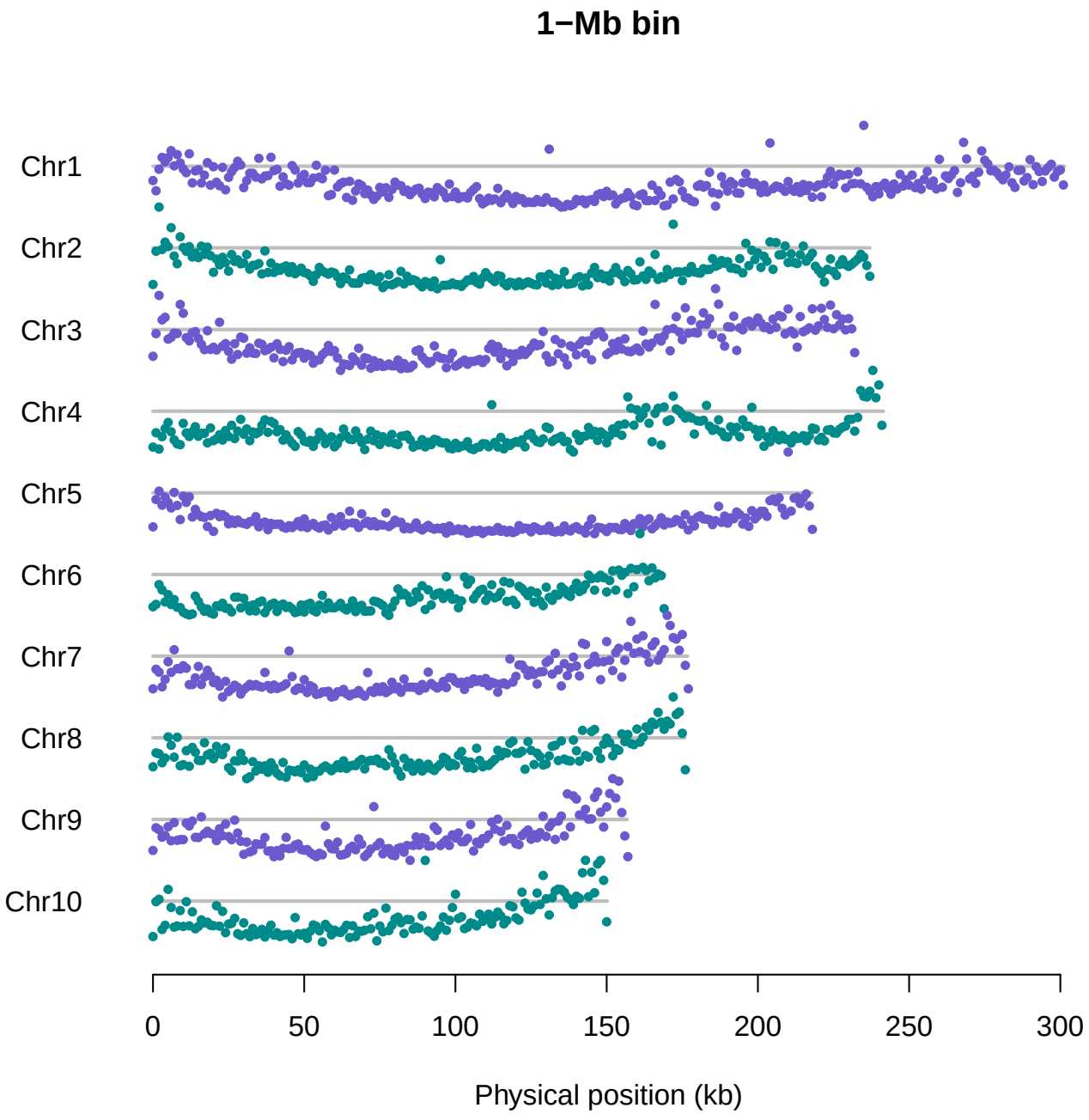

Figure S3 GERP score distribution across the genome. On the y-axis are the mean GERP scores in a 1-Mb bin region.

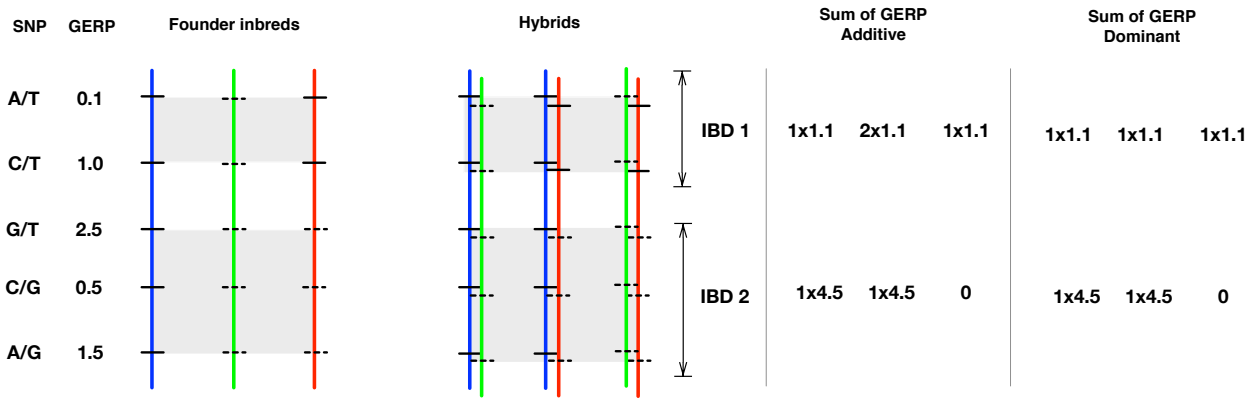

Figure S4 Incoporation of conservation information into IBD blocks. Regions of the genome that are identical by descent (IBD) among the 12 inbreds were identified using Beagle (**?**). The GERP scores of SNPs in an IBD block were summed under both additive and dominant models. Under the additive model, 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.

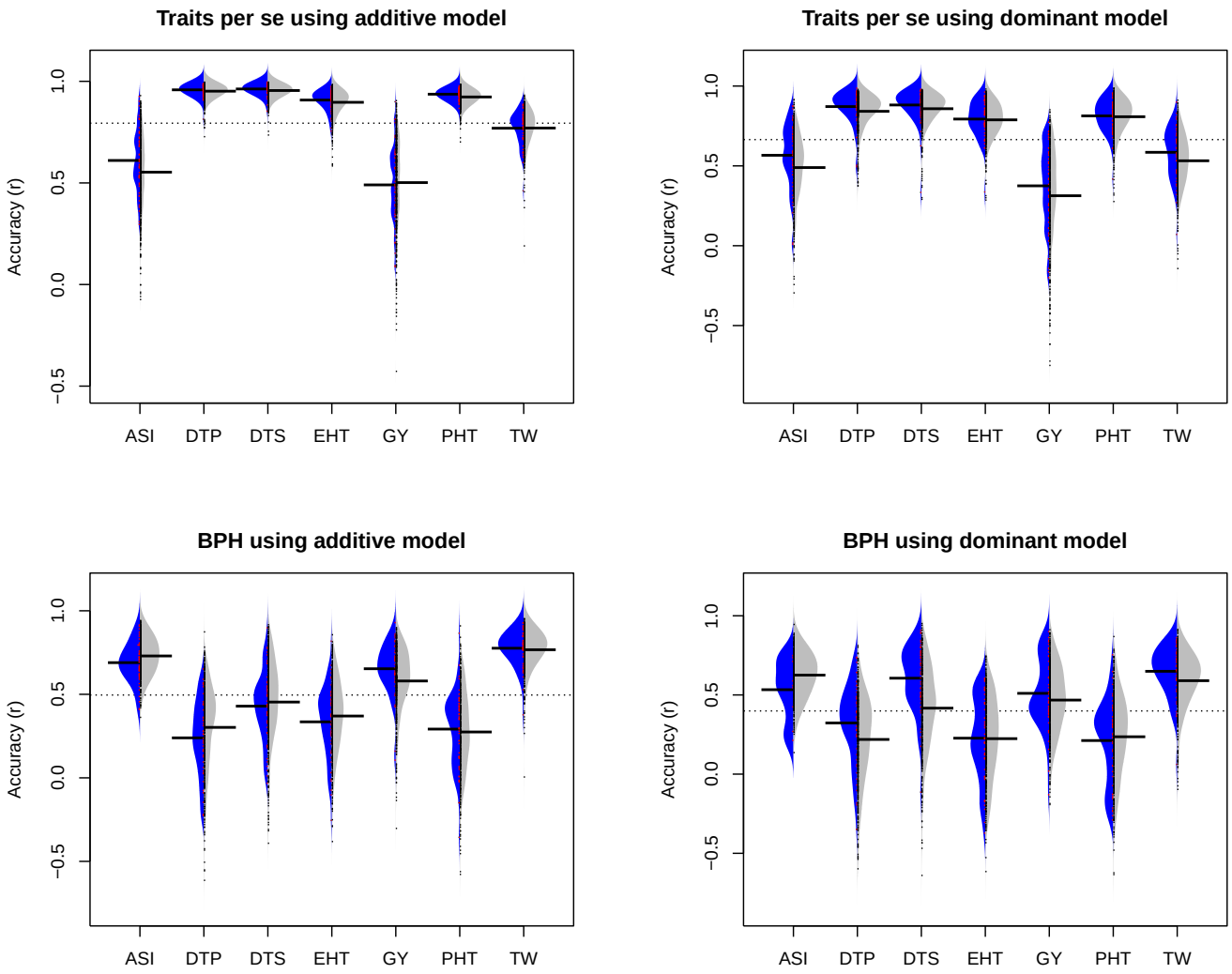

Figure S5 Beanplots of cross-validation accuracies using genic SNPs. Cross-validation experiments were conducted using genic SNPs and circular shuffled data from the same set of the genic SNPs for traits *per se* (**A, B**) and pHPH (**C, D**) under additive (**A, C**) and dominant (**B, D**) models. 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.