Deleterious genetic loads and their contributions to heterosis 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 Complementation of the deleterious alleles carried by the inbred parents may contribute to the vigorous performance, or heterosis, of the hybrid progenies. The detection of deleterious alleles was previous limited to the protein-coding regions. With the genomic evolutionary rate profiling (GERP), we extended the definition of deleterious alleles to the genome-wide. In total, about 86 million base-pairs, which make up 4.2% of the maize genome, were detected as evolutionary conserved sequences in both genic and non-genic regions. Here we take advantage of evolutionary measures 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 and phenotyped both parents and hybrids for seven phenotypic traits across one environment in three 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. We find that 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. Because most variation at conserved sites is deleterious, we interpret these results as consistent with the simple complementation model for heterosis. Overall, this work demonstrates the importance of incorporating evolutionary information in GS and its potential usage in plant breeding.

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

The phenomenon of heterosis or hybrid vigor has been observed ~~for many speciesacross species~~across many species, from yeast (Shapira *et al.* 2014) to plants **(CITE)** and vertebrates (Gama *et al.* 2013). ~~Recent studies indicated that the complementation of the~~ A number of hypotheses have been put forth to explain the phenomenon, including gene dosage **(CITE)**, **X (CITE)**, and **X (CITE)**. Complementation of recessive deleterious alleles, ~~which fit the classical dominance genetic model, may play an important role in determining heterosis (Charlesworth and Willis 2009).~~ however, remains the simplest genetic explanation (Charlesworth and Willis 2009), and is supported by considerable empirical evidence **(CITE)**.

Deleterious alleles were arisen from new mutations during meiosis. In maize, about 90 new mutations were generated per meiosis (Clark *et al.* 2005), majority of which were deleterious according to empirical estimates (Joseph and Hall 2004). In a natural outcross population, the negative effects on fitness of these deleterious alleles make them subject to be selection against~~. Therefore, deleterious alleles were~~, which lead the deleterious alleles to be maintained in a low frequency (Eyre-Walker and Keightley 2007). But the deleterious alleles could not be completerly purged.

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 ~~important traits~~purposes. In practice, the relatively homogeneous maize germplasm pool was artifi-

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Monday 22nd June, 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

cially divided into different heterotic groups (van Heerwaarden *et al.* 2012). 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 (Duvick 2001). 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 (Gore *et al.* 2009; McMullen *et al.* 2009) and the low recombination is effective over long period of time (Haddrill *et al.* 2007). 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 (Larièpe *et al.* 2012), 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 ~~minor effects individually~~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 ~~a~~ our previous study, ~~we found the enrichment of deleterious SNPs~~ deleterious SNPs were found to be enriched in a SNP set identified by GWAS (Mezmouk and Ross-Ibarra 2014). 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 ~~by~~ using genomic evolutionary rate profiling (GERP) (Cooper *et al.* 2005). By incorporating ~~the~~ GERP information in ~~the GS model~~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 ~~the genetic architectures varied among traits with different levels of heterosis; and the prediction accuracies with different~~ joint effects of deleterious alleles with additive and dominant modes of inheritance ~~would perform differently~~may contribute to heterosis.

Materials and Methods

Plant ~~Material~~ material and ~~Phenotypic Data~~phenotypic data

~~Twelve~~ We selected 12 maize inbred lines~~were selected and crossed in a half diallel fashion without considering reciprocal effects . The experimental design includes the~~ , broadly representative of corn belt maize germplasm (Mikel and Dudley 2006) , as parents of a partial diallel population. Each parent in a cross was used as both male and female and the resulting seed was bulked. We evaluated the 66 F1 hybrids, ~~the~~ 12 inbred parents ~~, and 2~~ and two current commercial check hybrids ~~grown in~~ in the field in Urbana, IL over three years (2009-2011) in an incomplete block design with ~~3 replications; hybrids and inbreds were grouped separately. The test was grown at Urbana, IL in 2009, 2010, and 2011.~~ three replicates each year. Plots consisted of ~~4~~ four rows, with all observations taken from the inside ~~2~~ 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~~ We measured 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 (~~TWT~~TW, in pounds)~~.~~. Overall mean phenotypic values for each cross can be found at **(CITE)** .

We estimated Best Linear Unbiased ~~Estimation (BLUE~~Estimates (BLUEs) of the genetic effects ~~were calculated with~~ in ASReml-R ~~following the~~ (Gilmour *et al.* 2009) with the following linear model:

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

where ~~y_{ijkl}~~ 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~~year; μ, the overall mean; ς_i, the fixed effect of the i^{th} ~~environment~~year; δ_{ij}, the fixed effect of the j^{th} replicate nested in the i^{th} ~~environment~~year; β_{jk}, the random effect of the k^{th} block nested in the j^{th} ~~block~~replicate; 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~~year; ε, the model residuals.

~~Heterosis for each hybrid was then estimated by both best- and mid-parent heterosis (*BPH* and *MPH*, respectively)~~ We estimated best-parent heterosis (BPH) as:

$$MPH_{ij} = \hat{G_{ij}} - \frac{1}{2}(\hat{G_i} + \hat{G_j})$$

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

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

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. *what about ear height and DTS?*

Sequencing ~~of Founder Lines~~ and ~~SNP Callings~~Genotyping

We extracted DNA from the ~~twelve inbred lines was CTAB extracted (Doyle and Doyle 1987) and Covaris sheared for Illumina~~ 12 inbred lines following Doyle and Doyle (1987) and sheared the DNA on a Covaris (Woburn, Massachusetts) for library preparation. ~~The DNA libraries~~ *do we need details on library prep? at least a citation?* Libraries were then sequenced ~~to an average coverage of 10X~~ *where? what length reads? insert size?* .

~~Raw paired reads (reverse and forward for each sequence), were trimmed for~~ We trimmed raw sequence reads 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~~ and for quality and sequence length (≥ 20 nucleotides) with Sickle ~~package~~ (https://github.com/najoshi/sickle). ~~Read pairs, kept after filtering, were mapped~~ We mapped filtered reads to the maize B73 reference genome (AGPv2) with bwa-mem (Li and Durbin 2009)~~. Reads,~~ , keeping 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~~ We called single nucleotide polymorphisms (SNPs) ~~were called~~

~~with mpileup~~ using the *mpileup* function from samtools utilities (Li *et al.* 2009). To deal with ~~paralogy, which is a major problem in maize sequence mapping (Chia *et al.* 2012) , all~~ known issues with paralogy in maize (Chia *et al.* 2012) , 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~~ fewer 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 (Browning and Browning 2009) and~~ We used the fastIBD method implemented in BEAGLE (Browning and Browning 2009, 2011) to impute missing data and identify regions of identity by descent ~~regions~~ (IBD) between the 12 inbred lines~~were identified with BEAGLE's fastIBD method (Browning and Browning 2011) . The pairwise IBD region starts and ends were used to delimit haplotipic blocs where several inbred lines shared a homogenous haplotypes.~~. We then defined haplotype blocks as contiguous regions within which there were no IBD break points across all pairwise comparisons of the parental lines (Figure S2).

~~The SNPs were annotated as synonymous and non-synonymous with the software polydNdS from the analysis package of libsequence (Thornton 2003) using the first transcript of each gene in B73 5b filtered gene set. Deleterious effects of amino acid changes were then predicted with both SIFT (Ng and Henikoff 2003a, 2006) and MAPP (Stone and Sidow 2005a) software packages as described by (Mezmouk and Ross-Ibarra 2014) . Genomic evolutionary rate profiling (GERP) , which estimates the evolutionary constraint by quantifying substitution deficits after multiple genome alignments, was obtained from for AGPv2.~~

~~Association mapping~~

~~SNP association with heterosis (BPH and MPH) was tested assuming dominance/recessivity of the reference allele or assuming overdominance where only the heterozygote alleles are expected to be significant. For each SNP, root mean square error were used to select the best fitting model. Haplotype association with heterosis were tested comparing the heterozygote alleles to all homozygote ones all confounded.~~

Genomic selection using IBD blocks incorporated with GERP scores

~~Genomic-enabled prediction with GERP score~~

~~A haplotype based genomic selection (GS) strategy was conceived by using the IBD blocks as the explanatory variables containing conservation information. The reference genome sites with GERP score />0 were considered as conserved sites and genomic variations at these sites were deemed as deleterious. To incoporate the conservation information into IBD blocks, SNPsfalling into a given IBD block were added up using their GERP scores as the conservation estimates of the IBD block~~We used genome-wide estimates of evolutionary constraint (GERP Davydov *et al.* 2010) estimated by Rodgers-Melnick *et al.* (2015) . Haplotype blocks were weighted by the summed GERP scores of all deleterious (GERP score > 0) SNPs. This estimation was calculated ~~using a python script gerpIBD ()with additive and dominant models. Under the additive model, 2 x GERP score was assigned to the homozygous loci with~~ under both additive and dominant modes of inheritance using a custom python script available at (https://github.com/yangjl/zmSNPtools). For a particular SNP with a GERP score g, the 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.~~ homozygote was assigned a value of $2g$, the heterozygote a value of g, and the reference homozygote a value of 0. Under the dominant model, ~~1 x GERP score was assigned to both the homzygous loci with~~ both the heterozygote and the non-reference ~~SNP calls and heterozygous loci; 0 was assigned to homozygous loci with reference SNP calls. seven phenotypic traits were trained with the conservation statistics in IBD block as genotype. A Bayesian-based approach, BayesC , was employed for the GS experiments. To conduct predict,~~ homozygote were assigned a value of g, with the reference homozygote again assigned a value of 0. To conduct prediction, a 5-fold cross-validation method was used, dividing the diallel population ~~was randomly divided into training~~ randomly into training (80%) and validation sets ~~for~~ (20%) 10 times~~according to a 5-fold cross-validation method. First, the BayesC model was trained independently on each of the trainingset. Second, the prediction accuracy was~~ . The BayesC option from GenSel4 (Habier *et al.* 2011) was used for model training, using 41,000 iterations and removing the first 1,000 as burn-in. *this said chains but I think you mean iterations?* After model training, prediction accuracies were obtained by comparing the predicted ~~and observed phenotypes on the conresponding validation set. In addition, the~~ breeding values with the observed phenotypes in the corresponding validation sets. For comparison, GERP scores were circularly shuffled ~~. The cross-validation~~ by 50k SNPs windows ($>$ 100Mb) 10 times to estimate a null conservation score for each IBD blocks. Cross-validation experiments using the circularly shuffled data were conducted on the same training and validation sets.

~~Data Access~~

~~*GENETICS* is committed to the open access to all primary data (see). Please indicate where data can be found (supplemental files, public repository, or published with another paper).~~

Results

Genetic values, heritability and heterosis~~transformations of a half diallel population~~

A ~~half~~ partial 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 much of 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~~ 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~~< 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~~ > 0.05). Broad sense heritability for these traits ranged from 0.65 for ASI to 0.95 for PHT. ~~With~~ Using the parental phenotypic data, we ~~conducted heterosis transformations using better-parental heterosis~~ then estimated best-parent heterosis (BPH) ~~and percent better-parental heterosis (pBPH)~~for each trait. Because the selected inbred lines are ~~commercial~~ commercially relevant and fairly elite in performance, hybrids in this population exhibit ~~relative~~ relatively low hybrid vigor (overall mean ~~pBPH~~ percent BPH = 0.3% ± 0.4%) ~~compared to their parents for most of the~~ for most traits except GY (mean ~~pBPH~~ percent BPH = 95% ± 16%, Figure S3). Finally, general and specific combining ability (GCA ~~ans~~ and SCA) were estimated following (Falconer and Mackay 1996). ~~The~~ GCA and SCA varied ~~according to the~~ among traits (Table S2) ~~. For the GY,~~ *please add ref to actual table* , but B73, PHG47 and PHG39 ~~are the top three inbred lines that combining better with others~~showed the greatest GCA *or should this say SCA?* for grain yield.

Evolutionary constraint information for genomic variants

~~In this study, all~~ 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,~~ , resulting in a filtered set of 13.8 million SNPs~~were kept for further analysis, 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~~ using three independent data sets: for all individuals using 41,292 overlapping SNPs on the maize SNP50 bead chip (van Heerwaarden *et al.* 2012); ~~we then compared our SNP alleles for B73 and Mo17 with the 10,426,715 SNP previously identified in HapMap2 (Chia *et al.* 2012) ; finally, we compared our SNPs to~~ for all individuals using 180,313 overlapping SNPs identified through genotyping by sequencing (GBS) (Romay *et al.* 2013)~~. The comparisons showed 99.12% allele similarity with shared SNPs previously identified.~~ ; and for B73 and Mo17 using the 10,426,715 SNP from the HapMap2 project (Chia *et al.* 2012) . *add a sentence or two describing the error rates!*

~~Evolutionary constraint information of GERP was obtained from (Rodgers-Melnick *et al.* 2015) , which was computed by using rejected substitution rate relative to the neutral rate after multiple genome squences alignment (Davydov *et al.* 2010) . 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 (Ng and Henikoff 2003b) or MAPP (Stone and Sidow 2005b) . In AGPv2, more~~ More than 86 million ~~bases (~4.2%~~ bp of the 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 S4) 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,~~ were annotated as conserved, with GERP scores > 0 (Figure 1A and S4). Nonetheless, 506,898 ~~SNP variants were detected at sites where the GERP scores are bigger than zero (Figure 1A). Consistent with previous study (Rodgers-Melnick *et al.* 2015) , the minor allele frequencies (MAFs) of these SNPs were~~ of these sites were found to segregate among the 12 inbred parents of our diallel. The minor allele frequency of SNPs at conserved sites was negatively correlated with ~~their GERP scores (Pearson's correlation test~~ GERP score (Figure 1B; P value < 0.05, r = ~~−0.08), indicating that the putative deleterious alleles tend to be purged and maintained in a low frequency in the population~~−0.8), consistent with the idea that variants at sites with more positive GERP scores are more deleterious and more strongly impacted by purifying selection.

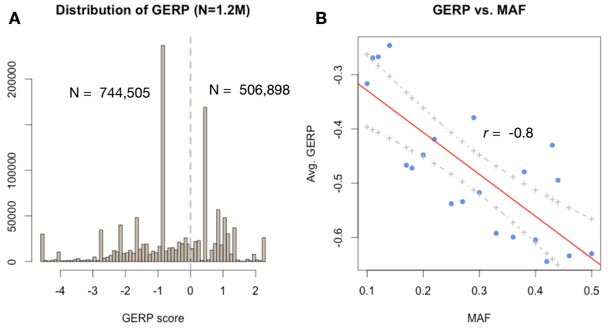


Figure 1 Distribution of GERP scores and relationship between GERP scores and MAFs. **(A)** GERP scores ~~were~~ extracted ~~for~~ from ~1.2 million SNPs ~~(~10%~~ of the ~~total SNPs) of the~~ diallel population. ~~The spikes in the histogram were because only limited genomes (N < 10?) were used in GERP calculation.~~ *1.2M is just the sites where GERP scores are available, correct?* **(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 ~~occured~~ occurred at the evolutionary constraint sites were potentially deleterious. The phenotypic effects of these genetic loads and their contributions to heterosis become ~~interesting~~an interesting area to explore. However, the population size in this study is relative small and SNPs detected at ~~high GERP sites~~ sites containing high GERP scores are generally in low ~~frequcies. To alliviate~~ 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 ~~genomic selection (GS)~~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 ~~at least one~~ > 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 ~~conserveness~~ conservation of the haplotypes. More details of this procedure were illustrated in Figure S5.

A Bayesian-based statistical method (BayesC) (Habier *et al.* 2011) 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~~and pBPH~~), incorporation of GERP scores improved BPH of GY under the additive model and improved BPH of DTP, DTS and TW under the dominant model ~~; and the method improved predictions for pBPH of TW under the additive model and pBPH of GY under the dominant model~~ (Figure 2 ~~C-F~~C and D, Supplementary table 3). In general, the average prediction accuracis were higher using the additive model (mean $r = 0.81$ ~~, 0.49 and 0.29~~ and 0.49 for traits *per se* ~~, BPH and pBPH~~and BPH) than the dominant model (mean $r = 0.70$ ~~, 0.42 and 0.24~~and 0.42). And the prediction accuracies decreased for predicting heterosis transformations (BPH~~and pBPH~~) 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 S6 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 ~~expected, for most of the phenotypes, higher posterior variance could be explained for~~ shown in Figure 3, additive models explained more phenotypic variance for traits ~~*per se* than BPHand pBPH using~~ *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.

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 GERP scores.~~ 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.

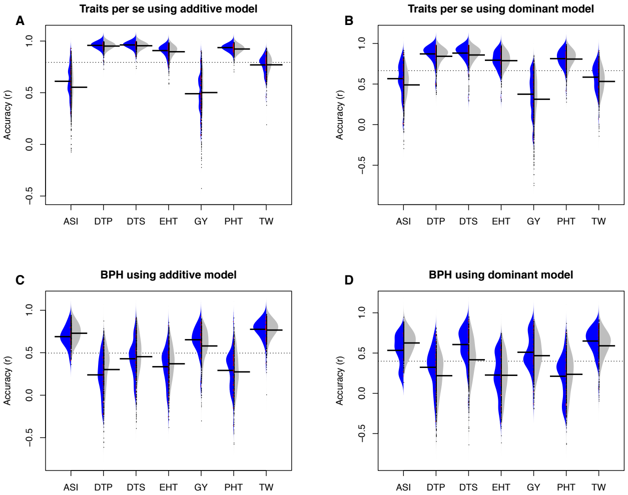


Figure 2 Beanplots of cross-validation accuracies using SNPs with positive GERP score. Cross-validation experiments were conducted using selected SNPs and circular shuffled data from the same set of SNPs for traits *per se* (**A, B**), ~~HPH~~ BPH (**C, D**) and ~~pHPH~~ pBPH (**E, F**) under additive (**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.

~~were observed using all the conserved SNPs as compared to the genic SNPs (Table SN) . For most of the traits *per se*, phenotypic variance could be largely explained (posterior variance explained by markers range from 0.9 to 0.8 with additive model; with dominant model). But, it became difficult to explain heterosis transformations. Figure 3 GCA and SCA and the breeding values. Breeding values from the BeyesC models were extracted. The below four models were compared.~~

$$Y_{ij} = \mu + GCA_i + GCA_j + \varepsilon \tag{1}$$

$$Y_{ij} = \mu + GCA_i + GCA_j + G_{ij} + \varepsilon \tag{2}$$

$$Y_{ij} = \mu + GCA_i + GCA_j + SCA_{ij} + \varepsilon \tag{3}$$

$$Y_{ij} = \mu + GCA_i + GCA_j + SCA_{ij} + G_{ij} + \varepsilon \tag{4}$$

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.

~~1. It capture high order of interactions. 2. Explainary variables came from the centeromic regions. TO DO: 0. preparing all the traits (per se, HPH, MPH, pHPH and pMPH) 1. comparing same no. of random SNPs vs. GERP SNPs without considering their SCORE. 2. comparing MAF in different caterigories 3. training all the data and get the idea of significant ones.~~

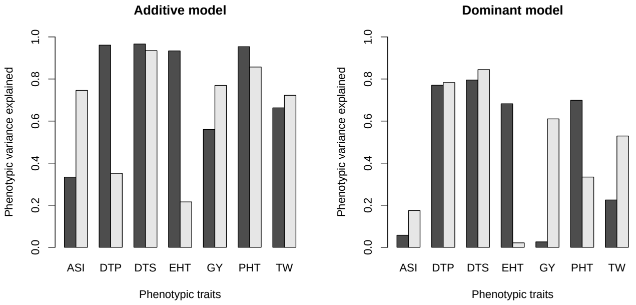


Figure 3 Barplots of the phenotypic variance explained by IBD blocks incorporated with GERP scores.

Discussion

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

~~Schmitt: How to explain the prediction difference? - First, 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.~~ In this study, more than 500,000 deleterious SNPs were identified in elite maize lines including in ~~noncoding~~ non-coding regions of the genome. Majority of them were maintained in a low frequency, which consistent with the previous observation ~~and indicate~~ (Rodgers-Melnick *et al.* 2015) and indicated the deleteriousness of the variants in the conserved sites. ~~A~~ To estimate the joint effects of the potential deleterious alleles for the phenotype, a genomic selection pipeline was developed, which utilized evolutionary conservation information in the ~~model. Cross-validation~~ 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.

Literature Cited

Browning, B. L. and S. R. Browning, 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**: 210–23.

Browning, B. L. and S. R. Browning, 2011 A fast, powerful method for detecting identity by descent. Am J Hum Genet **88**: 173–82.

Charlesworth, D. and J. H. Willis, 2009 The genetics of inbreeding depression. Nature reviews. Genetics **10**: 783–96.

Chia, J.-M., C. Song, P. J. Bradbury, D. Costich, N. de Leon, J. Doebley, R. J. Elshire, B. Gaut, L. Geller, J. C. Glaubitz, M. Gore, K. E. Guill, J. Holland, M. B. Hufford, J. Lai, M. Li, X. Liu, Y. Lu, R. McCombie, R. Nelson, J. Poland, B. M. Prasanna, T. Pyhäjärvi, T. Rong, R. S. Sekhon, Q. Sun, M. I. Tenaillon, F. Tian, J. Wang, X. Xu, Z. Zhang, S. M. Kaeppler, J. Ross-Ibarra, M. D. McMullen, E. S. Buckler, G. Zhang, Y. Xu, and D. Ware, 2012 Maize hapmap2 identifies extant variation from a genome in flux. Nat Genet **44**: 803–7.

Clark, R. M., S. Tavaré, and J. Doebley, 2005 Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Molecular Biology and Evolution **22**: 2304–2312.

Cooper, G. M., E. a. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow, 2005 Distribution and intensity of constraint in mammalian genomic sequence. Genome research **15**: 901–13.

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

Doyle, J. J. and J. Doyle, 1987 Genomic plant dna preparation from fresh tissue-ctab method. Phytochem Bull **19**: 11–15.

Duvick, D. N., 2001 Biotechnology in the 1930s: the development of hybrid maize. Nature Reviews Genetics **2**: 69–74.

Eyre-Walker, A. and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nature reviews. Genetics **8**: 610–618.

Falconer, D. and T. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman.

Gama, L. T., M. C. Bressan, E. C. Rodrigues, L. V. Rossato, O. C. Moreira, S. P. Alves, and R. J. B. Bessa, 2013 Heterosis for meat quality and fatty acid profiles in crosses among Bos indicus and Bos taurus finished on pasture or grain. Meat Science **93**: 98–104.

Gilmour, A. R., B. Gogel, B. Cullis, and R. Thompson, 2009 Asreml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK .

Gore, M. a., J.-M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz, B. L. Hurwitz, J. a. Peiffer, M. D. McMullen, G. S. Grills, J. Ross-Ibarra, D. H. Ware, and E. S. Buckler, 2009 A first-generation haplotype map of maize. Science (New York, N.Y.) **326**: 1115–7.

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

Haddrill, P. R., D. L. Halligan, D. Tomaras, and B. Charlesworth, 2007 Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. Genome biology **8**: R18.

Joseph, S. B. and D. W. Hall, 2004 Spontaneous mutations in diploid Saccharomyces cerevisiae: More beneficial than expected. Genetics **168**: 1817–1825.

Larièpe, a., B. Mangin, S. Jasson, V. Combes, F. Dumas, P. Jamin, C. Lariagon, D. Jolivot, D. Madur, J. Fiévet, A. Gallais, P. Dubreuil, A. Charcosset, and L. Moreau, 2012 The genetic basis of heterosis: multiparental quantitative trait loci mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (Zea mays L.). Genetics **190**: 795–811.

Li, H. and R. Durbin, 2009 Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics **25**: 1754–60.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, 2009 The sequence alignment/map format and samtools. Bioinformatics **25**: 2078–9.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li, Q. Sun, S. Flint-Garcia, J. Thornsberry, C. Acharya, C. Bottoms, P. Brown, C. Browne, M. Eller, K. Guill, C. Harjes, D. Kroon, N. Lepak, S. E. Mitchell, B. Peterson, G. Pressoir, S. Romero, M. Oropeza Rosas, S. Salvo, H. Yates, M. Hanson, E. Jones, S. Smith, J. C. Glaubitz, M. Goodman, D. Ware, J. B. Holland, and E. S. Buckler, 2009 Genetic properties of the maize nested association mapping population. Science (New York, N.Y.) **325**: 737–40.

Mezmouk, S. and J. Ross-Ibarra, 2014 The pattern and distribution of deleterious mutations in maize. G3 (Bethesda, Md.) **4**: 163–71.

Mikel, M. A. and J. W. Dudley, 2006 Evolution of north american dent corn from public to proprietary germplasm. Crop science **46**: 1193–1205.

Ng, P. C. and S. Henikoff, 2003a Sift: predicting amino acid changes that affect protein function. Nucl Acids Res **31**: 3812–3814.

Ng, P. C. and S. Henikoff, 2003b Sift: Predicting amino acid changes that affect protein function. Nucleic acids research **31**: 3812–3814.

Ng, P. C. and S. Henikoff, 2006 Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet **7**: 61–80.

Rodgers-Melnick, E., P. J. Bradbury, R. J. Elshire, J. C. Glaubitz, C. B. Acharya, S. E. Mitchell, C. Li, Y. Li, and E. S. Buckler, 2015 Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceedings of the National Academy of Sciences **112**: 3823–3828.

Romay, M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts, T. M. Casstevens, R. J. Elshire, C. B. Acharya, S. E. Mitchell, S. A. Flint-Garcia, M. D. McMullen, J. B. Holland, E. S. Buckler, and C. A. Gardner, 2013 Comprehensive genotyping of the usa national maize inbred seed bank. Genome Biol **14**: R55.

Shapira, R., T. Levy, S. Shaked, E. Fridman, and L. David, 2014 Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models. Heredity **113**: 1–11.

Stone, E. A. and A. Sidow, 2005a Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res **15**: 978–86.

Stone, E. A. and A. Sidow, 2005b Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome research **15**: 978–986.

Thornton, K., 2003 Libsequence: a c++ class library for evolutionary genetic analysis. Bioinformatics **19**: 2325–2327.

van Heerwaarden, J., M. B. Hufford, and J. Ross-Ibarra, 2012

Historical genomics of north american maize. Proc Natl Acad Sci U S A **109**: 12420–5.

Supporting Information

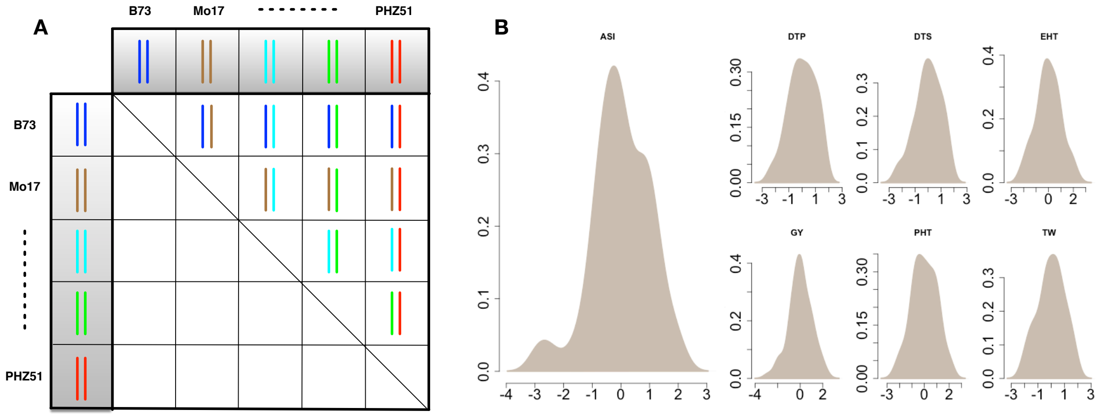
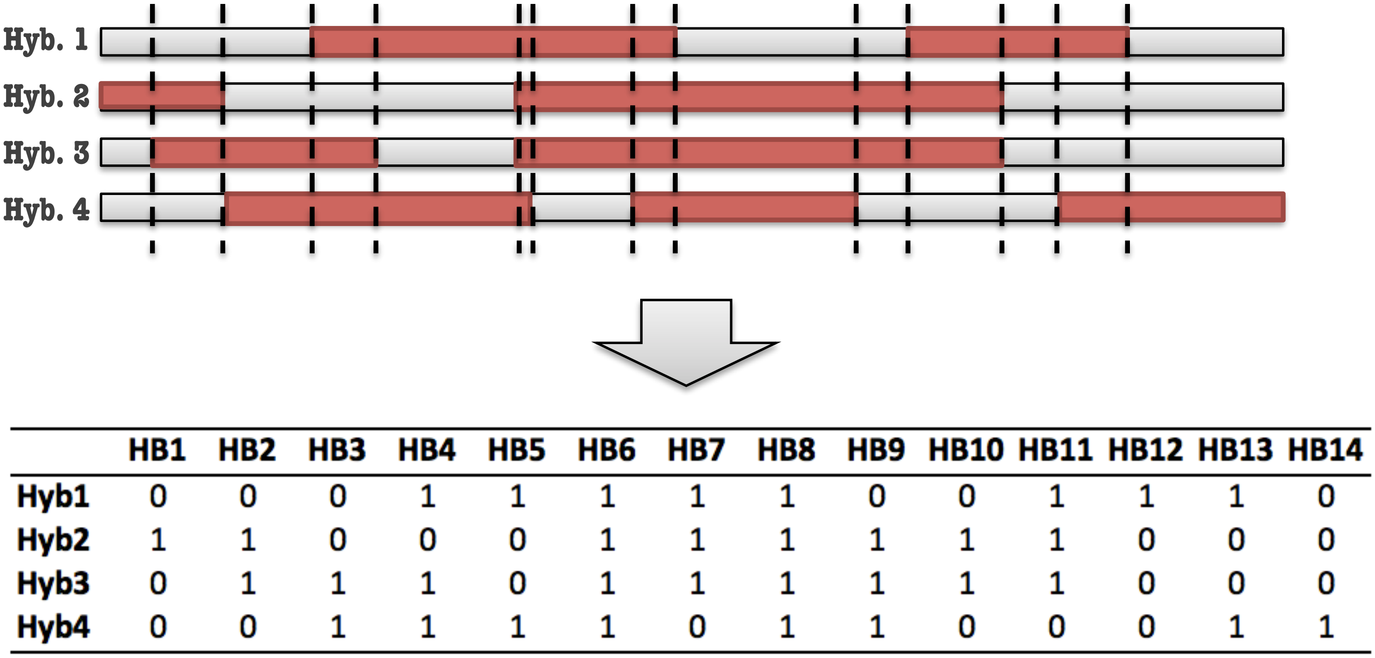


Figure S1 **Diallel experimental design and distribution of phenotypic data.** **(A)** Twelve maize inbred lines were selected and crossed in a ~~half diallel. Ten of these (LH1, LH123HT, LH82, PH207, 4676A, PHG39, PHG47, PHG84, PHJ40, PHZ51) are proprietary inbreds that have expired from Plant Variety 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.~~ partial diallel. *do we need to modify this diagram now that we now reciprocal crosses were pooled?* **(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).~~ Density plots of the phenotypic distributions.



	HB1	HB2	HB3	HB4	HB5	HB6	HB7	HB8	HB9	HB10	HB11	HB12	HB13	HB14
Hyb1	0	0	0	1	1	1	1	1	0	0	1	1	1	0
Hyb2	1	1	0	0	0	1	1	1	1	1	1	0	0	0
Hyb3	0	1	1	1	0	1	1	1	1	1	1	0	0	0
Hyb4	0	0	1	1	1	1	0	1	1	0	0	0	1	1

Figure S2 Haplotype block identification using an IBD approach. In the upper panel, regions in red are IBD blocks identified by pairwise comparison of the two parental lines of a hybrid. The vertical dashed lines define haplotype blocks. In the lower panel, hybrid genotypes at each block are coded as heterozygotes (0) or homozygotes (1).

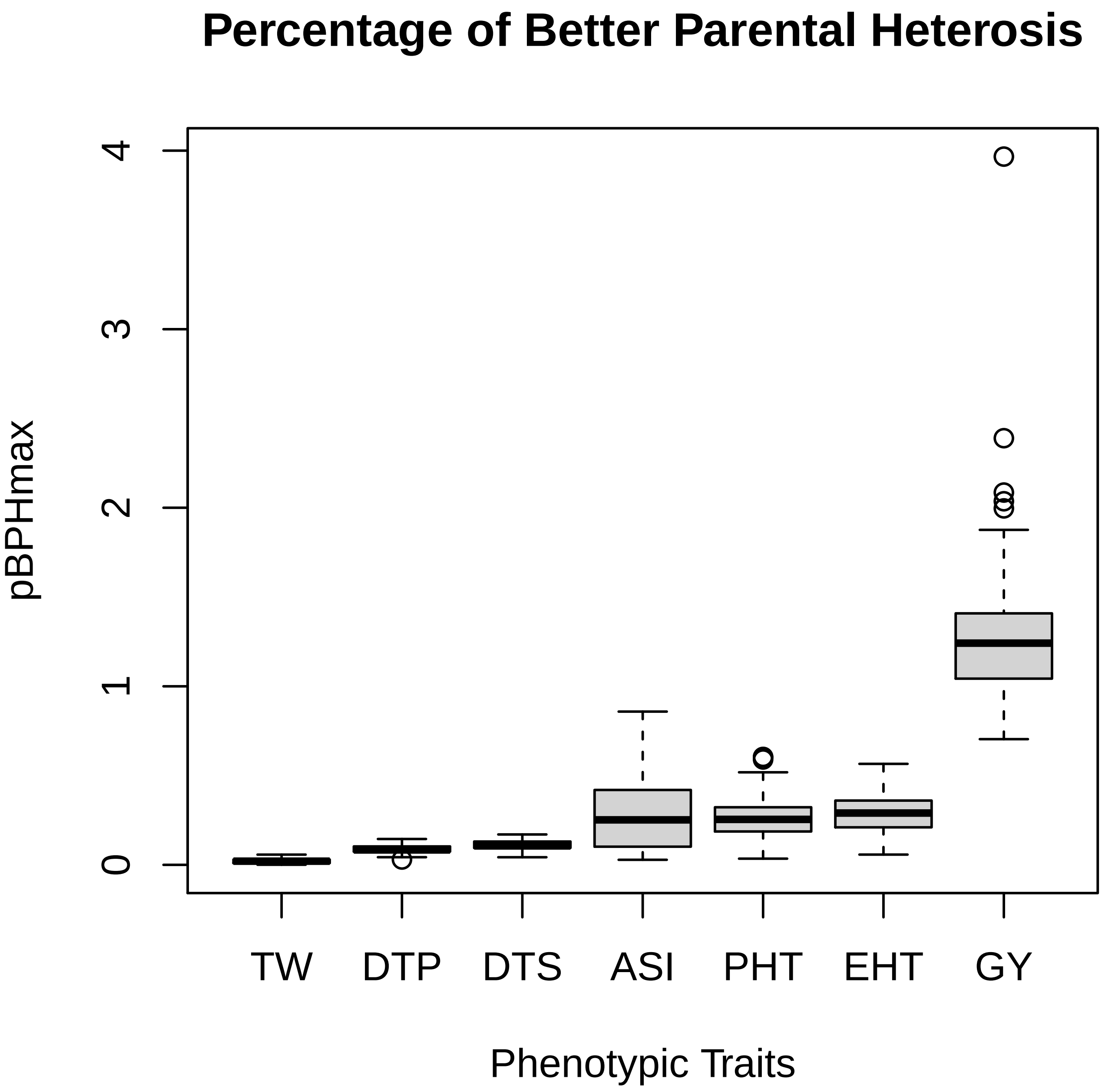


Figure S3 Boxplot of ~~the~~ percent ~~better parental~~ best parent heterosis (pBPH). In the plot, ASI was calculated using pBPHmin and the other six traits were calculated using pBPHmax.

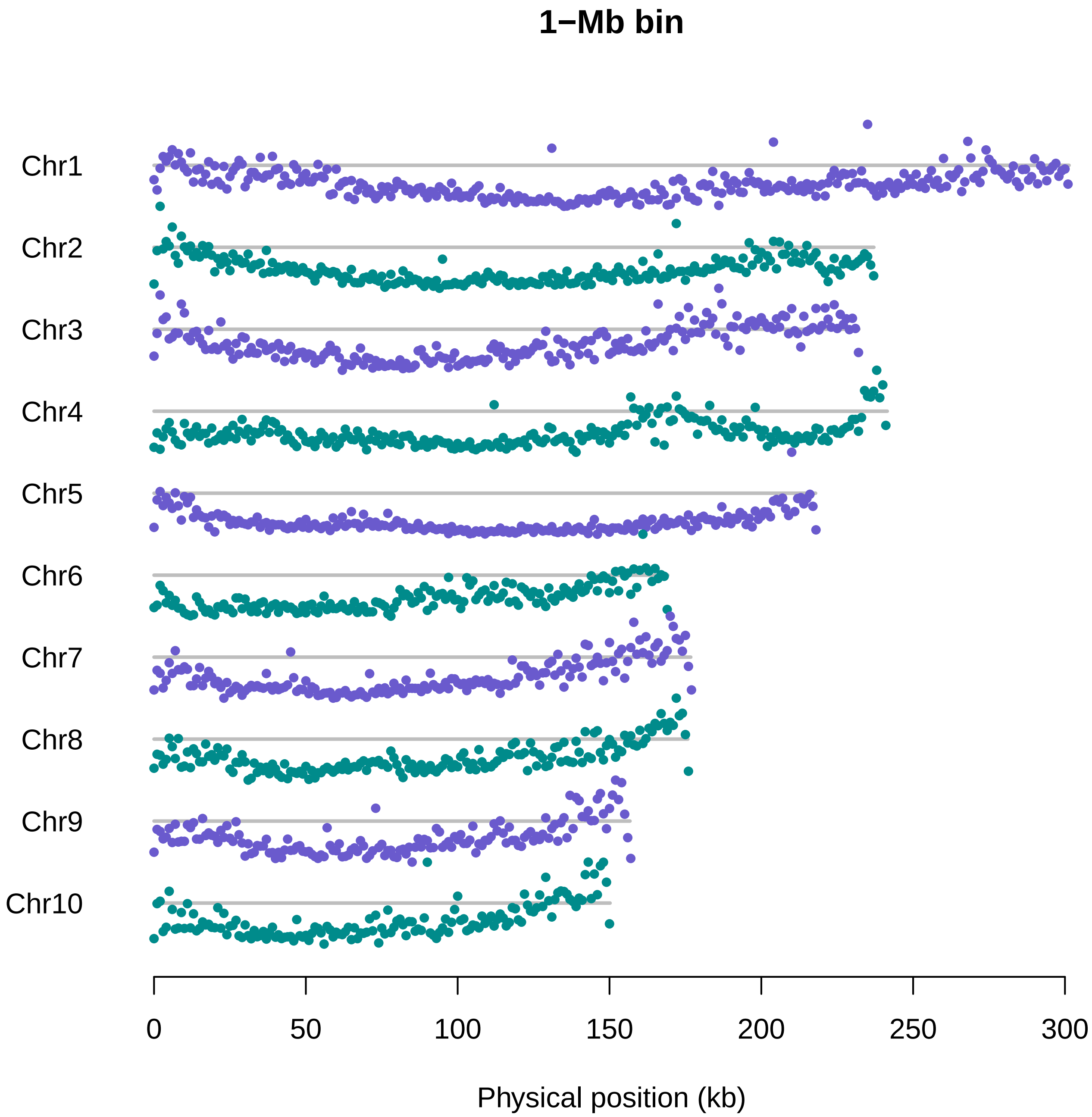


Figure S4 GERP score distribution across the genome. ~~On the y-axis~~ Shown are ~~the~~ mean GERP scores in a 1-Mb bin region.

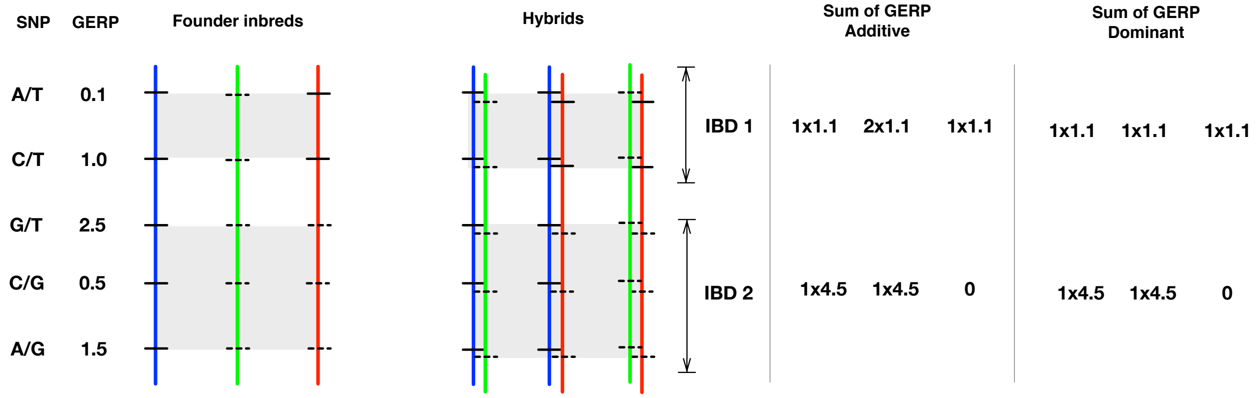


Figure S5 Incoporation of conservation information into IBD blocks. Regions of the genome that are identical by descent (IBD) among the 12 inbreds were identified using Beagle (Browning and Browning 2009). The GERP scores of SNPs in an IBD block were summed under both additive and dominant models. ~~Under the additive model, 2 x~~ For a particular SNP with GERP score ~~was assigned to genotypes homozygous for~~ g, the homozygous non-reference ~~allele, 1 x GERP score~~ genotype was assigned ~~to heterozygotes~~a value of 2g, ~~and 0 was~~ the heterozygote assigned ~~to~~ a value of g, and the ~~homozygous~~ reference ~~genotype.~~ homozygote a value of 0. Under the dominant model, ~~1 x GERP score was assigned to~~ both ~~genotypes with a nonreference allle~~ the heterozygote and ~~0 to~~ the ~~homozygous~~ non-reference homozygote were assigned a value of g, with the reference ~~genotype.~~homozygote again assigned a value of 0.

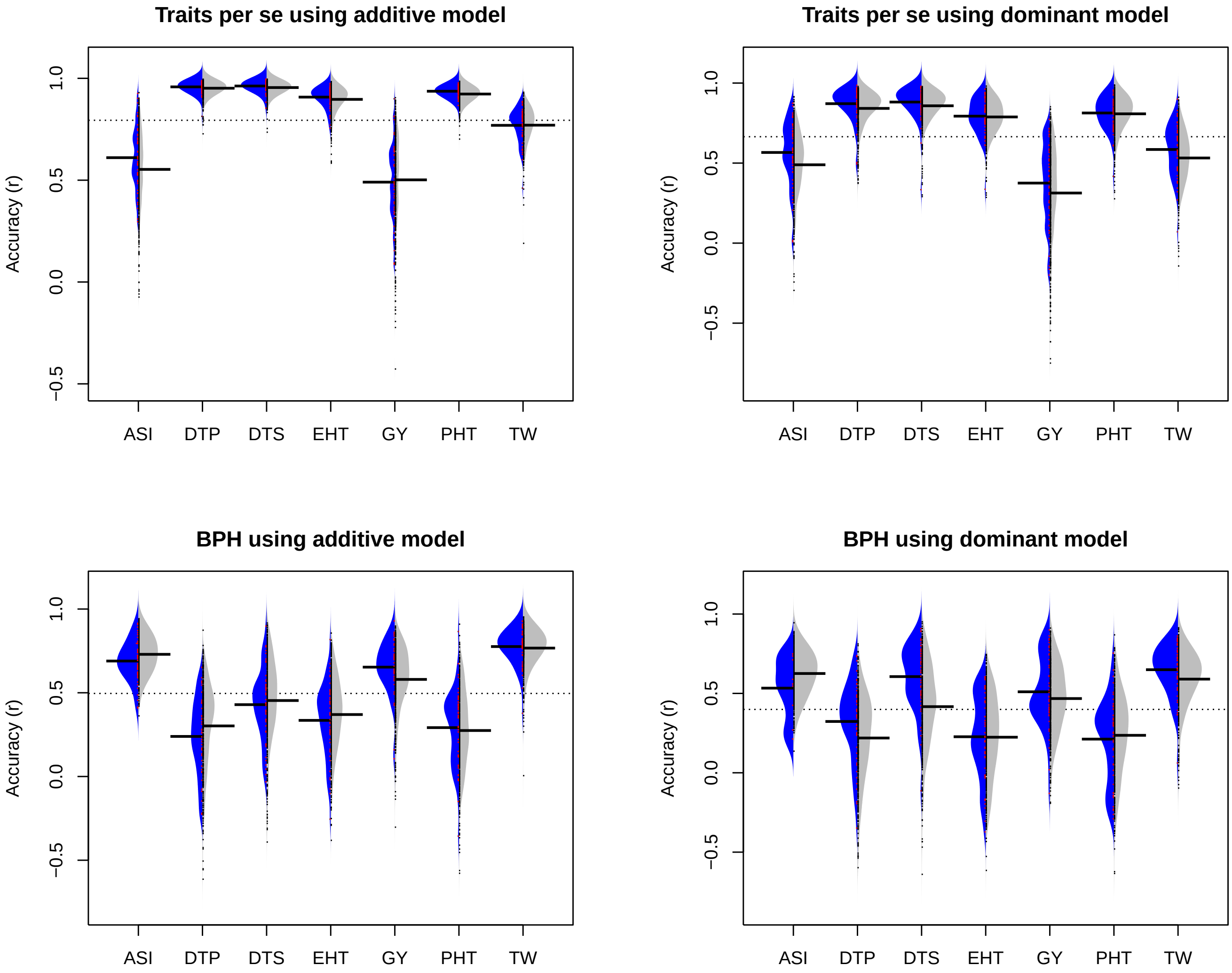


Figure S6 ~~Beanplots of cross-validation~~ Cross-validation accuracies using genic SNPs. Cross-validation experiments were conducted using genic SNPs and ~~circular shuffled~~ compared to circular-shuffled data ~~from the same set of the genic SNPs~~ for traits *per se* (**A, B**) and ~~pHPH~~ pBPH (**C, D**) under additive (**A, C**) and dominant (**B, D**) models. ~~Accuaries~~ Distirbutions show accuracty of prediction from ~~the~~ real data ~~were plotted on the left side of the bean~~ (blue) and ~~permutation results plotted on the right~~ permutations (grey)~~. Horizotal~~ , with horizontal bars ~~on beans~~ to indicate mean ~~accuracies. The grey dashed line indicates the overall average~~ accuracy. Stars indicate significantly ~~improved~~ higher cross-validation ~~accuracies~~accuracy for the real data. The average accuracy across all traits is shown with the grey dotted line.