Utilizing evolutionary conservation information to improve prediction accuracy in genomic selection

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

[*]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. 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 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 for many species across species, from yeast (Shapira *et al.* 2014) to vertebrates (Gama *et al.* 2013). Recent studies indicated that the complementation of the deleterious alleles, which fit the classical dominance genetic model, may play an important role in determining heterosis (Charlesworth and Willis 2009). 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 maintained in a low frequency (Eyre-Walker and Keightley 2007).

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. In practice, the relatively homogeneous maize germplasm pool was artificially 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

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Wednesday 29th April, 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

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. Here, we employed a genomic selection (GS) approach to simultaneously estimate genome-wide deleterious variants in a half diallel population. The diallel population composed of a set of hybrids 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 previous study, we found the enrichment of deleterious SNPs 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, we demonstrated the prediction accuracies were significantly improved not only for some traits *per se*, but for some heterosis transformations. Further studies indicated that the genetic architectures varied among traits with different levels of heterosis; and the prediction accuracies with different modes of inheritance would perform differently.

Materials and Methods

need more work, ignore this section when editing...

Plant Material and Phenotypic Data

Twelve maize inbred lines were selected and crossed in a half diallel fashion without considering reciprocal effects **(CITE)**. The experimental design includes the 66 F1 hybrids, the 12 inbred parents, and 2 current commercial check hybrids grown 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. Plots consisted of 4 rows, with all observations taken from the inside 2 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 (TWT, in pounds) **(CITE)**.

Best Linear Unbiased Estimation (BLUE) of the genetic effects were calculated with ASReml-R **(CITE)** 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 both best- and mid-parent heterosis (BPH and MPH, respectively):

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

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

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

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.

Sequencing of Founder Lines and SNP Callings

DNA from the twelve inbred lines was CTAB extracted (Doyle and Doyle 1987) and Covaris sheared for Illumina library preparation. The DNA libraries were then sequenced to an average coverage of 10X **(say where the sequencing was done?)**.

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 (Li and Durbin 2009). 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 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 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 (Browning and Browning 2009) and 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.

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 2003, 2006) and MAPP (Stone and Sidow 2005) software packages as described by (Mezmouk and Ross-Ibarra 2014).

Genomic evolutionary rate profiling (GERP) **CITE**, which estimates the evolutionary constraint by quantifying substitution deficits after multiple genome alignments, was obtained from **cite eli2015** 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-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, SNPs falling into a given IBD block were added up using their GERP scores as the conservation estimates of the IBD block. This estimation was calculated using a python script gerpIBD (https://github.com/RILAB/pvpDiallel) with 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 homzygous loci with 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 **CITE**, was employed for the GS experiments. To conduct predict, the diallel population was randomly divided into training and validation sets for 10 times according to a 5-fold cross-validation method. First, the BayesC model was trained independently on each of the training set. Second, the prediction accuracy was obtained by comparing the predicted and observed phenotypes on the conresponding validation set.

In addition, the GERP scores were circularly shuffled. The 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 Genetics, 184: 1). 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 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) and percent better-parental heterosis (pBPH). Because the selected inbred lines are commercial relevant and fairly elite in performance, hybrids in this population exhibit relative low hybrid vigor (overall mean pBPH = 0.3% $\pm$ 0.4%) compared to their parents for most of the traits except GY (mean pBPH = 95% $\pm$ 16%, Figure S2). Finally, general and specific combining ability (GCA ans SCA) were estimated following (Falconer and Mackay 1996). 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.

Genomic sites under evolutionary constraint

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 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 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 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.

Evolutionary constraint information for genomic sites was obtained by computing the rejected substitution rate relative to the neutral rate after multiple genome alignment, or genomic evolutionary rate profiling (GERP) (Davydov *et al.* 2010). This approach could estimate the GERP score at a base-pair level, extending the definition of deleterious variants to intergenic regions compared to previous approaches, such as SIFT (**citation**) or MAPP (**citation**). In B73 reference genome AGPv2, a total of 86,006,888 sites (4.2% of the genome) **how this compared to human**, were detected having GERP scores />0; and these sites were determined as evoulationary constraint sites. Genomic variants occured on them were potentially deleterious. From genome-wide of view, generally, genomic sequences are evolutionarily conserved near the telomeric regions and the conserveness dropped toward conteromic regions (Figure S2); there are some exceptions, such as the long arm of chr4, chr5 and chr1.

Although genomic sites with GERP >0 were conservationary constraint, they are not immune to mutation. Indeed, in our diallel population, N SNPs were detected at sites with GERP score />0. Consistent with previous study (())**eli**, the minor allele frequcy (MAF) was negatively correlated with the mean GERP

scores (correlation P value =). This observation indicated the putative deleterious alleles tend to be purged and maintained in a low frequency in the population.

how many in genic and how many in non-genic

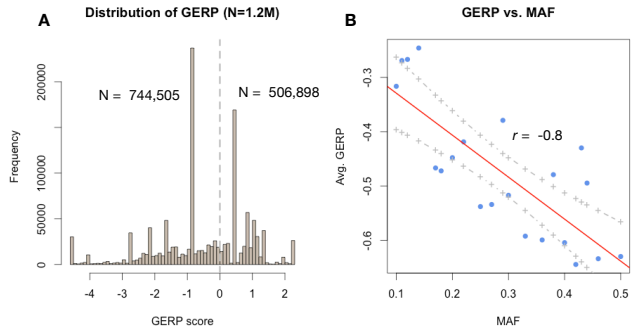

Figure 1 GERP distribution of SNPs and relationship between GERP and minor allele frequency. GERP scores were obtained for ∼1.2 million (∼10%) SNPs. The spikes of the MAF distribution. Of these, 506,898 (42%) were under evolutionary constraint and considered as deleterious variants. **(B)** Mean GERP 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.

Supplementary Figure and a little bit story of the conserved regions.

IBD region size and general statistics

The phenotypic effects of the genetic load of these potentially deleterious alleles have not been explored. To estimate their phenotypic effects and their contributions to heterosis, we conceived a genomic selection approach to predict their joint effects in the population. However, because SNPs with high GERP score are negatively correlated with the MAF (Figure and citation), they become less useful due to statistical limitation (citation). To capture the information carried by these potentially deleterious sites, we conducted the genomic-enabled prediction with the SNP's GERP score.

Identity by descent was esimated with fasibd. Average size 44,980 bp (36 to 10,320,000 bp) Figure 1.

SNPs in IBD regions IBD: ranged from 1 to 370,152, with the mean of 10,167 and median 4162. Number of SNPs in the IBD blocks: 1-1600, mean 22.83 and median 9.

To estimate the phenotypic effects of the genetic loads, we conceived the Genomic selection approach to evaluate whether incorporaing the GERP score would potentially improve the prediction accucy. However, the population in this study is relative small with only 66 individuals. The 20k conserved SNPs are highly colinear with others. Simpe SNP based analysis would suffer a lot from so called big p small n problem. To alliviate this problem, we employed a haplotype-based approach for genomic selection, where IBD blocks were considered as a haplotype (**citation**).

With pairwise comparisons, IBD regions were chopped into 55,000 IBD blocks. These IBD blocks have the average size of 44,980 bp (ranged from 36 to 10,320,000 bp, Figure S3).

For IBD blocks > 1-kb and contains at least one deleterious alleles (SNP variation with GERP >0), they were used for the analysis. Two modes, additive and dominant, were used to code the IBD blocks as the measurement of the conserveness of haplotypes.

Evolutionary conservation information improved prediction accuracies

With 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 score information in the IBD blocks under the additive model; prediction accuracy was significantly improved for ASI under the dominant model. For heterosis transformation traits (measured by 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, Supplementary table N). In general, the average prediction accuracis are higher using the additive model (r = 0.81, 0.49 and 0.29 for traits *per se*, BPH and pBPH) than the dominant model (r = 0.70, 0.42 and 0.24). And the prediction accuracies decreased for predicting heterosis transformations (BPH and pBPH) as compared to predictions for traits *per se*.

Incorporation of GERP scores showed the potential to improve prediction accuarices. However, because SNPs in genic regions tend to evolutionarily consertive and therefore had a relative higher GERP scores. Our circular shuffling approach had a chance to shift the high GERP score to intergenic regions. To rule out the prossibility that the method is not weight higher on genic SNPs. We elected the SNPs in genic regions and did the circular shuffling to random assign GERP scores the the same set of the selected SNPs. By doing this, but becasue they all from the genic regions, circular shuffling would not take positional advantages any more. Note that less conserved SNPs were selected (N=), for traits *per se*, model prediction accuracies were significantly improved GY under the additive model (FDR < 0.001). Prediction accuracies were significantly improved for GY and PHT under the additive model (Figure S4).

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 permuta- tions were used both considering all SNPs or considering only genic SNPs to control for differences between genic and nongenic regions.

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 aver- age 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.

The top predictors are enriched in centeromeic regions

To learn why the prediction performace varied among traits and heterosis transformations. First of all, we derived the posterior variance explained by the IBD coded with GERP score and IBD coded with genic SNPs. As expected, higher posterior variance explained by genome-wide IBD blocks were observed using all

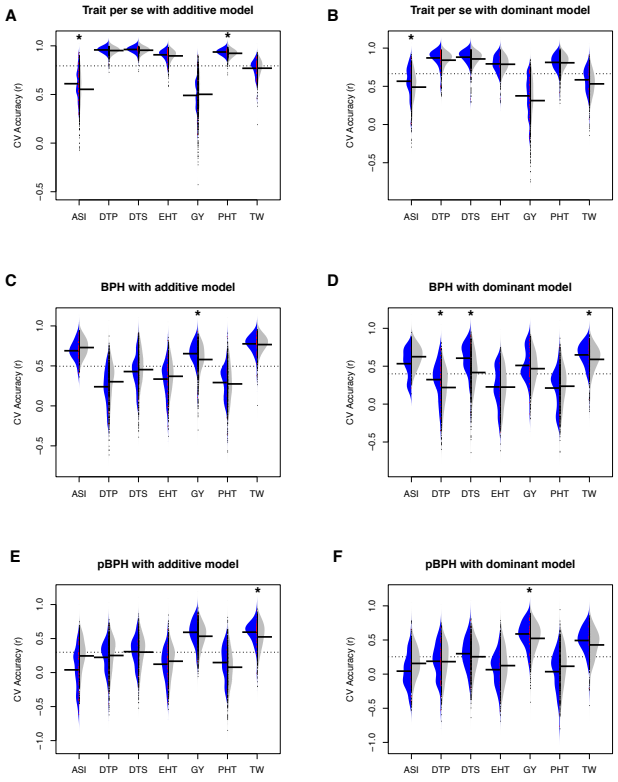

Figure 2 Beanplots of cross-validation accuracies using SNPs with positive GERP score. Cross-validation experiments were conducted using selected SNPs and circular shuffled data from the same set of SNPs for traits *per se* (**A, B**), HPH (**C, D**) and pHPH (**E, F**) under additive (**A, C, E**) and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

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.

GCA and SCA

and the breeding values.

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.

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 regions of the genome. Majority of them were maintained in a low frequency, which consistent with the previous observation **Eli PNAS, 2015** and indicate the deleteriousness of the variants in the conserved sites. 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.

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.

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.

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.

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

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

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

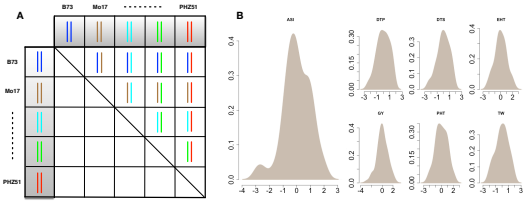

Figure S1 Diallel experimental design and distribution of phenotypic data. (A) Twelve maize inbred lines were selected and crossed in a half diallel. Ten of these (LH1, LH123HT, LH82, PH207, 4676A, PHG39, PHG47, PHG84, PHJ40, PHZ51) are proprietary inbreds that have expired from Plant Variety Protection (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. **(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).

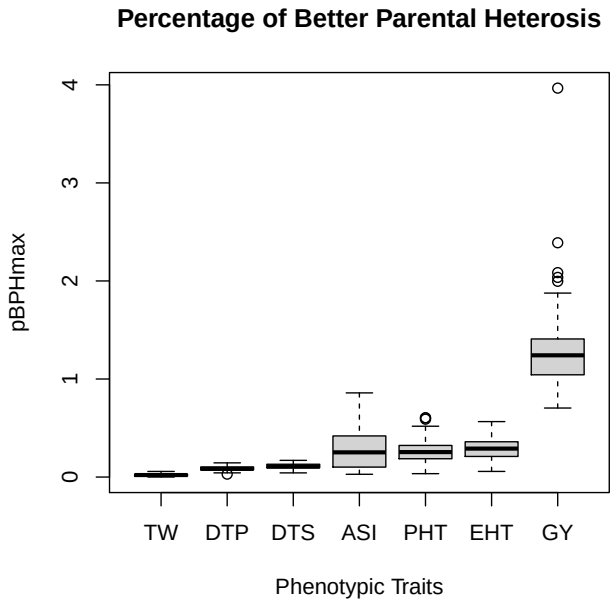

Figure S2 Boxplot of the percent better parental heterosis (pBPH). In the plot, ASI was calculated using pBPHmin and the other six traits were calculated using pBPHmax.

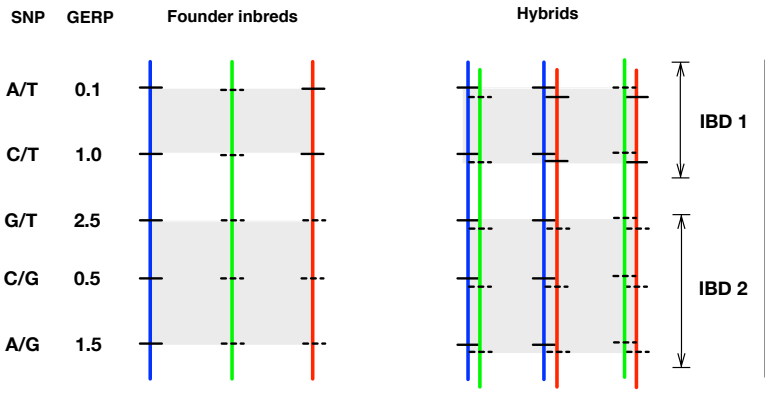

Figure S3 Incoporation of conservation information into IBD blocks. Regions of the genome that are identical by descent (IBD) among the 12 inbreds were identified using Beagle (Browning and Browning 2009). The GERP scores of SNPs in an IBD block were summed under both additive and dominant models. Under 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.

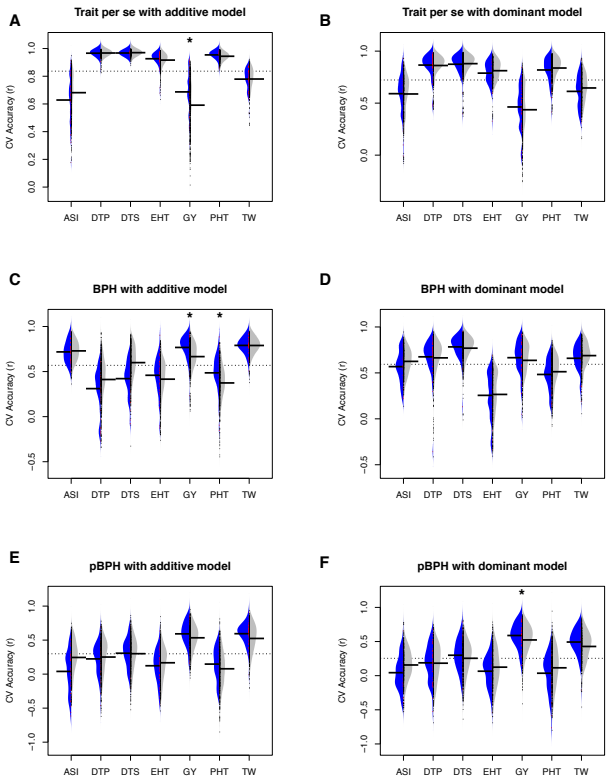

Figure S4 Beanplots of cross-validation accuracies using genic SNPs. Cross-validation experiments were conducted using genic SNPs and circular shuffled data from the same set of the genic SNPs for traits *per se* (**A, B**) and pHPH (**C, D**) under additive (**A, C**) and dominant (**B, D**) models. 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.