Heterosis and genetic load in an ExPVP partial diallel cross

Jinliang Yang[*1], Sofiane Mezmouk[†1,2,3], Rita H. Mumm[‡4], and Jeffrey Ross-Ibarra[§1,2]

[1]Department of Plant Sciences, University of California, Davis, CA 95616, USA

[2]Center for Population Biology and Genome Center, University of California Davis

[3]Current address: KWS SAAT AG, Grimsehlstr. 31, 37555 Einbeck, Germany

[4]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[*]jolyang@ucdavis.edu

[†]smezmouk@ucdavis.edu

[‡]ritamumm@illinois.edu

[§]rossibarra@ucdavis.edu

1

Abstract

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

Introduction

Why we care about deleterious variants. Discuss results of Mezmouk 2014. We are extending this in three ways

- all deleterious SNPs not just coding

- genome-wide not just reduced representation

- using GS to test whether they improve prediction

The phenomenon of heterosis has been observed for many species across taxa, from yeast (Shapira et al., 2014) to vertebrates (Gama et al., 2013). Although heterosis has been broadly applied for food production and has been extensively studied by many researchers, the underlying mechanism controlling for it remains elusive. 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 deemed to be 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 selected against. Therefore, deleterious alleles were mainly 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. By doing this, the maize yield has been steadily improved since the early 20th century **citation**. However, the removing of deleterious mutations in low recombination regions or in tightly linked regions is 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 pericentromeric region and the vigorous performance could only be realized by combining two sets of non-deleterious or beneficial alleles in repulsion state, thus lead to pesudo-overdomination.

Recent QTL study identified loci controlling for heterosis are enriched in centromeric regions (Larièpe et al., 2012), which may support this pesudo-overdominance hypothesis.

To study the relationship between deleterious variants, especially those with minor effect or rare in the population, and their contribution to heterosis, a diallel population was employed (). This diallel population enabled us to invistigate the different modes of inheritance of the deleterious variant in a hybrid population with super dense SNP markers; while only relative little sequencing efforts need to be conducted on founder lines. Our previous study observed that an excess of deleterious SNPs were enriched in GWAS set (Mezmouk and Ross-Ibarra, 2014). In that study, deleterious variants were defined as nonsynonymous mutations. Clearly, deleterious varinats are not limitted 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). GERP scores were obtained by computing the regected substitutions subtracted by the neutral rate after multiple sequence alignment of a set of related species (Davydov et al., 2010). **Sofina's GWAS results..** Because most of the deleterious alleles are rare, their effects are difficult to be detected by traditional GWAS approach. To overcome it, a genomic selection approach was conceived, which enable us to estimate the combined effects of all the possible deleterious variants simuteously. As the results, genomic selection with GERP score incorporated in the prediction model show better performance than circular shuffled data for some phentoypic traits. This study demonstrates the importance of incorporating evolutionary information in genomic selection and its potential use in plant breeding.

Results

- Genetic values, heritability, heterosis and combining ability

- SNP calls and annotation; distribution of deleterious mutations along the genome

- correlation between complementation at deleterious SNPs with heterosis and SCA for the different

- IBD region size and general statistics

- IBD correlation with heterosis and SCA

- GWAS results at an SNP level and then comparison with haplotypic results

- Analyses of the significant haplotypic bloc to see if there is any pattern

Genetic values, heritability, heterosis and combining ability of a partial diallel population

A half diallel population was created using 12 maize inbred lines. Two of them are important public inbreds, B73 and Mo17. And the other ten of them are (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. This diallel facilitated the creation of F1 hybrids that are adapted to the U.S., fairly elite in performance, and commercially relevant. The set is diverse enough to facilitate a broad sweep of the heterotic sub-groups that comprise U.S. commercial germplasm.

IBD region size and general statistics

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

Figure S1: **??**

GERP, SIFTMAPP and general statistics

Introduce GERP SNP and GERP elements. For each position of the multiple alignment, the conservation score is computed in rejected substitution by subtracting the estimated evolutionary rate from the neutral rate (citation). 224,087 GERP elements were identified in B73v2. 29,869,451 bps (1.4% of the maize genome).

Figure S2. distribution of GERP, GERP vs. MAF.

Figure S3. GERP in gene, exon and intron.

Evolutionary conservation information improved prediction accuracies

Figure S4. GERP in IBD design with additive and dominance model.

Some prior knowledge have not been used for model training. SNPs with high GERP score were considered as evolutionary conserved. The mutation at these sites tend to be deleterious. 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.

The population in this study is relative small with only 66 individuals. The 20k conserved SNPs are highly colinear with others. To solve this problem, we use the haplotype-based approach for model training, where the haplotype was coded with the SNP conservation score as the explainatory variables.

With a 5-fold cross-validation approach, the prediction accuracies of the real data and cicularly shuffled data were compared. To rule out the prossibility that genic SNPs are more conserved than intergenic ones. 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.

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

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 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.

Materials and methods

Plant material and phenotypic data

Twelve maize inbred lines were selected and crossed in a partial diallel fashion without considering reciprocal effects **(CITE)**. Ten of these inbreds (LH1, LH123HT, LH82, PH207, 4676A, PHG39, PHG47, PHG84, PHJ40, PHZ51) are proprietary inbreds that have expired from Plant

Variety Protection (PVP) and represent the lineage of key heterotic germplasm pools used in present-day commercial corn hybrids, and two are predominant public inbreds B73 and Mo17.

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)**.

Analyses

Inbred line sequencing 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).

After filtering 13,782,809 are kept for further comparisons, including 1,909,416 genic SNPs 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 (Romay et al., 2013). The comparisons showed 99.12% allele similarity with shared SNPs previously identified.

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.

Phenotypic data analyses

Best Linear Unbiased Estimation (BLUE) of the genetic effects were calculated with ASReml-R **(CITE)** following the linear model **check the following mixed 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 lth genotype evaluated in the kth block of the jth replicate within the ith environment; μ, the overall mean; ς_i, the fixed effect of the ith environment; δ_{ij}, the fixed effect of the jth replicate nested in the ith environment; β_{jk}, the random effect of the kth block nested in the jth block; α_l, the the fixed genetic effect of the lth individual; $\varsigma_i \cdot \alpha_l$, the interaction effect of the lth individual with the ith environment; ε, the model residuals.

All fixed effects, including the genotype onces, 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.

Heterosi 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.

Finally, general and specific combining ability (GCA ans SCA) respectively were estimated following Falconer and Mackay (1996).

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

9

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 conrresponding 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.

References

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

Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *Am J Hum Genet*, 88(2):173–82.

Charlesworth, D. and Willis, J. H. (2009). The genetics of inbreeding depression. *Nature reviews. Genetics*, 10(11):783–96.

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

Clark, R. M., Tavaré, S., and Doebley, J. (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., Stone, E. a., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7):901–13.

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

Doyle, J. J. and Doyle, J. (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(1):69–74.

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

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

Gama, L. T., Bressan, M. C., Rodrigues, E. C., Rossato, L. V., Moreira, O. C., Alves, S. P., and Bessa, R. J. B. (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(1):98–104.

Gore, M. a., Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. a., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H., and Buckler, E. S. (2009). A first-generation haplotype map of maize. *Science (New York, N.Y.)*, 326(5956):1115–7.

Haddrill, P. R., Halligan, D. L., Tomaras, D., and Charlesworth, B. (2007). Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome biology*, 8(2):R18.

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

Larièpe, a., Mangin, B., Jasson, S., Combes, V., Dumas, F., Jamin, P., Lariagon, C., Jolivot, D., Madur, D., Fiévet, J., Gallais, A., Dubreuil, P., Charcosset, A., and Moreau, L. (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(2):795–811.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., Pressoir, G., Romero, S., Oropeza Rosas, M., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J. C., Goodman, M., Ware, D., Holland, J. B., and Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science (New York, N.Y.)*, 325(5941):737–40.

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

Ng, P. C. and Henikoff, S. (2003). Sift: predicting amino acid changes that affect protein function. *Nucl Acids Res*, 31(13):3812–3814.

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

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., Elshire, R. J., Acharya, C. B., Mitchell, S. E., Flint-Garcia, S. A., McMullen, M. D., Holland, J. B., Buckler, E. S., and Gardner, C. A. (2013). Comprehensive genotyping of the usa national maize inbred seed bank. *Genome Biol*, 14(6):R55.

Shapira, R., Levy, T., Shaked, S., Fridman, E., and David, L. (2014). Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models. *Heredity*, 113(4):1–11.

Stone, E. A. and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res*, 15(7):978–86.

Thornton, K. (2003). Libsequence: a c++ class library for evolutionary genetic analysis. *Bioinformatics*, 19(17):2325–2327.

van Heerwaarden, J., Hufford, M. B., and Ross-Ibarra, J. (2012). Historical genomics of north american maize. *Proc Natl Acad Sci U S A*, 109(31):12420–5.
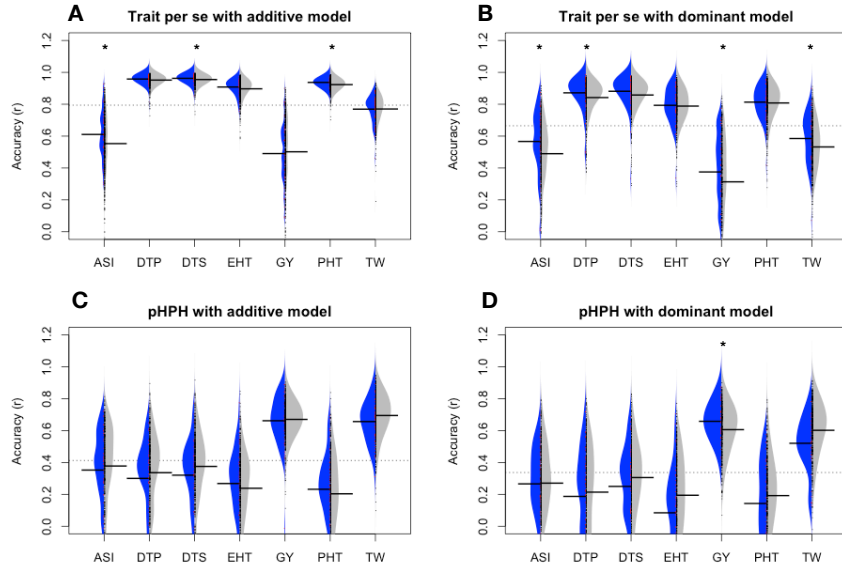
Figures and Tables

Figure 1: **Beanplots of cross-validation accuracies.** Cross-validation experiments were conducted using genic SNPs and circular shuffled data from the same set of the genic SNPs for traits *per se* (**A, B**) and pHPH (**C, D**) under additive (**A, C**) and dominant (**B, D**) models. Accuaries from the real data were plotted on the left side of the bean (blue) and permutation results plotted on the right (grey). Horizotal bars on beans indicate mean accuracies. The grey dashed line indicates the overall average accuracy. Stars indicate significantly improved cross-validation accuracies.