The pattern and distribution of deleterious mutations in maize

Sofiane Mezmouk*[1] and Jeffrey Ross-Ibarra†[1,2]

[1]Department of Plant Sciences, University of California Davis

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

Abstract

Most non-synonymous mutations are thought to be deleterious because of their effect on protein sequence, and are expected to be removed or kept at low frequency by the action of natural selection. Nonetheless, the effect of positive selection on linked sites or drift in small or inbred populations may also impact the evolution of deleterious alleles. In spite of their potential to affect complex trait phenotypes, deleterious alleles are difficult to study precisely because they are often at low frequency. Here, we made use of genome-wide genotyping data to characterize deleterious variants in a large panel of maize inbred lines. We show that, in spite of small effective population sizes and inbreeding, most putatively deleterious SNPs are indeed at low frequencies within individual genetic groups. We find that

*smezmouk@ucdavis.edu

†rossibarra@ucdavis.edu

genes associated with a number of complex traits are enriched for deleterious variants. Together these data are consistent with the dominance model of heterosis, in which complementation of numerous low frequency, weak deleterious variants contribute to hybrid vigor.

SNP data is available at http://www.panzea.org/dynamic/derivative_data/genotypes/Maize282_GBS_genos_imputed_20120110.zip.

Phenotypic data is available at http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0007433.s001 **and** http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0007433.s002

Introduction

The effect of new mutations on organismal fitness is not well understood, but both theoretical considerations (Fisher, 1930) and empirical estimates (Joseph and Hall, 2004) suggest that most new mutations are deleterious and only a small minority are beneficial. Strongly deleterious mutations are expected to be kept at low frequencies by natural selection, whereas weakly deleterious alleles may be effectively neutral (Ohta, 1973; Kimura, 1983) and subject to the effects of genetic drift (Lynch and Gabriel, 1990; Lande, 1994; Whitlock et al., 2003). In addition to selection and drift, a number of other factors such as mating system and recombination rate also impact the evolution of deleterious alleles. Selfing species and inbreeding within populations will expose lethal mutations to selection faster than in an outcrossing population (Wang et al., 1999; Glémin et al., 2003). Moreover, in genomic regions with low levels of recombination, selection against deleterious mutations will be less effective (Charlesworth et al., 1993) and the potential exists for deleterious mutations to rise to high frequency due to the effects of linked selection on beneficial mutations (Felsenstein, 1974; Hill and Robertson, 1966; Chun and Fay, 2011).

Deleterious alleles may play an important functional role in affecting the phenotype of traits of interest, and complementation between haplotypes carrying different deleterious alleles may explain much of the observation of hybrid vigor or heterosis (Charlesworth and Willis, 2009). In human disease studies, significant correlation was observed between the deleterious predictions of SNPs and their association with cancer (Zhu et al., 2004); predicted rare deleterious SNPs were also shown to be involved in common diseases (Cohen et al., 2004; Smigrodzki et al., 2004). Furthermore, rare deleterious SNPs have gained interest due to their potential role in explaining quantitative trait variation (Gibson, 2012), especially in populations that have experienced recent growth (Lohmueller, 2013) .

Evaluating the abundance and frequency of deleterious mutations is thus of considerable interest and has been investigated in a wide range of species. These analyses have varied in terms of the percentage of non-synonymous sites estimated to be deleterious, from 3% in bacterial populations (Hughes, 2005) to 80% in the human genome (Fay et al., 2001). They have also shown that recently bottlenecked populations may have a higher abundance of deleterious sites and that heterozygosity at deleterious SNPs is lower than at synonymous SNPs (Lohmueller et al., 2008). In plants, Gossmann et al. (2010) found that most new mutations in plants are strongly deleterious, with only 25% acting as effectively neutral. Cao et al. (2011) show that the abundance of deleterious variants correlates with effective population size in *Arabidopsis thaliana*, and a demographic bottleneck appears to have relaxed purifying selection in *Capsella rubella* (Brandvain et al., 2013). Other analyses of purifying selection in plants have implicated a role for environmental differences (Tellier et al., 2011) and identified differences among genes based on their level of expression (Paape et al., 2013). In natural populations of *Arabidopsis thaliana*, selection appears to act to maintain variants that are locally adaptive but deleterious elsewhere (Fournier-Level et al., 2011), whereas positive selection on domestication genes may have increased the abundance of deleterious variants in domesticated genomes such as rice (Günther and Schmid, 2010; Lu et al., 2006). While these studies have provided insight into the evolutionary fate of deleterious mutations, we still understand relatively little about the role of deleterious variants in effecting phenotypic traits.

Maize (*Zea mays*) is a economically important cereal worldwide, with the highest yield and one of largest cultivated areas (FAO statistics, `http://faostat.fao.org`); it is also an important model for basic and applied research (Strable and Scanlon, 2009). Maize was traditionally cultivated in open pollinated populations (landraces) but, after the first documented observations of hybrid vigor in

this species (East, 1908; Shull, 1908), inbred lines were developed and structured into heterotic groups that maximize inter-group combining ability. The transition from heterozygous populations to strongly structured heterotic groups of inbred lines makes maize of interest for analyzing the distribution and frequency of deleterious mutations. Furthermore, high observed values of hybrid vigor or heterosis in maize hybrids makes it an excellent system for studying the effects of deleterious mutations and their contribution to heterosis. The dominance model of heterosis posits that inbred lines are homozygous for a number of recessive deleterious alleles and that crosses between inbreds carrying different complements of deleterious alleles will result in heterozygous progeny with higher fitness than either parent.

The aim of the current study was to (1) carry out a genome-wide scan for deleterious mutations in a maize diversity panel, (2) analyze their distribution across the genome and within different genetic groups, and (3) test for enrichment of deleterious loci in the results of genome wide association mapping. High density single nucleotide polymorphisms (SNPs) and phenotypic data available for a large sample of inbred lines and hybrids were used to address these questions. Our results showed that maize inbred lines are segregating for a large number of predicted deleterious variants (20 to 40 % of protein coding SNPs were predicted to have a deleterious allele), and that these alleles are generally at very low frequencies with few fixed differences observed among different genetic groups. Genome-wide association analysis of hybrid vigor finds little evidence for enrichment of individual deleterious SNPs, but significant enrichment for genes containing deleterious SNPs, suggesting a meaningful role for dominance and complementation in explaining observations of hybrid vigor.

Materials and methods

Plant material and phenotypic data

We utilized phenotypic data published in Flint-Garcia et al. (2005) for 247 maize inbred lines (see supplemental data for a list of inbred lines). Each inbred line was crossed to the stiff-stalk inbred B73 (population A) and both the inbred lines and their B73-hybrids were evaluated in 2003, in adjacent blocks within three environments with a single replicate in each (Flint-Garcia et al., 2009). A subset of 102 inbreds were additionally crossed to both B73 (population B1) and Mo17 (population B2); both inbred lines and hybrids were evaluated in a single environment in 2006 (Flint-Garcia et al., 2009). Supplemental Table 1 lists the analyzed traits which are detailed in Flint-Garcia et al. (2009).

The panel structure was previously analyzed (Flint-Garcia et al., 2005) and inbred lines were attributed to the following subpopulations: stiff-stalk (27 inbred lines), non stiff-stalk (90 inbred lines), tropicals (60 inbred lines), popcorns (8 inbred lines), sweet (6 inbred lines) and mixed (56 inbred lines). For the main temperate inbred lines, these subpopulations corresponds to the different heterotic groups.

Genotypic data

We made use of genotypic data from Larsson et al. (2013) for the full set of 247 lines. The latter were genotyped using the genotyping-by-sequencing approach (GBS; Elshire et al., 2011), resulting in a total of 437,650 partially imputed SNPs . Of these SNPs, 127,994 mapped to protein coding sequences representing 123,289 codons in 21,064 genes. The median (mean) percentage of missing data per SNP, including triallelic sites, was 1.06% (2.52%), while the percentage of heterozygous sites was 1.08% (2.52%). Only 4.5% of SNPs had more than 10% missing data

(Supplemental Figure 1-A), and 0.18% had more than 10% heterozygous genotypes (Supplemental Figure 1-B).

We estimated error rates by first comparing our genotyped inbred B73 to the B73 reference genome, then by comparing all our genotypes to those from 7,225 overlapping SNPs on the maize SNP50 bead chip (Cook et al., 2012). Compared to the reference genome, our B73 genotype differed (alternative homozygote allele) at 1.75% of SNPs, and across all lines our genotypes differed at a median (mean) rate of 1.83% (4.62%) from the maize SNP50 data (Cook et al., 2012).

Statistical analyses

SNP annotation and analyses

The first transcript of each gene in the B73 5b filtered gene set was used to annotate SNPs as synonymous and non-synonymous with the software polydNdS from the analysis package of libsequence (Thornton, 2003). The deleterious effects of amino acid changes were then predicted for proteins derived from the first transcript of each gene with both the SIFT (Ng and Henikoff, 2003, 2006) and MAPP (Stone and Sidow, 2005) software packages.

SIFT uses homologous sequences identified by PSI-BLAST against protein databases to identify conserved amino acids. The software provides a scaled score of the putative deleterious effect of a particular amino acid at a position along a protein.

MAPP predicts deleterious amino acid polymorphisms from a user-defined alignment of protein homologs. It uses the phylogenetic relatedness among sequences and the physicochemical properties of amino acids to quantify the potential deleterious effect of a given amino acid change. We created alignments for MAPP using three different methods. First, we made BLASTX comparisons of

protein sequences from maize against the TrEMBL database (Boeckmann et al., 2003) retaining all proteins with an e-value $\leq 10^{-40}$ and at least 60% identity with the query. Second, we used a reciprocal best BLAST criterion to compare protein sequences of maize against protein sequences from 31 plant genomes (supplemental data) from Phytozome version 8.0 (`http://www.phytozome.net`), retaining the best hit protein from each of the other genomes with an e-value $\leq 10^{-100}$ and $\geq 70\%$ coverage of the query length. Finally, we made use of a set of syntenic genes from the grasses *Zea mays*, *Sorghum bicolor*, *Oryza sativa* and *Brachypodium distachyon* (Schnable et al., 2012). For each set of proteins, ClustalW2 (Larkin et al., 2007) was used to align the sequences and build a neighbor-joining tree. Custom R code (`https://github.com/RILAB/siftmappR`) was used to link amino acid positions to SNP positions and to link the amino acid polymorphisms to MAPP and SIFT predictions.

The derived site frequency spectrum was calculated for all protein coding SNPs using *Tripsacum* (Chia et al., 2012) to determine ancestral state. The pattern of haplotype sharing across the genome (PHS statistics; Toomajian et al., 2006) was analyzed within each of the tropical, stiff-stalk, non-stiff stalk and mixed subpopulations as defined by Flint-Garcia et al. (2005). We will refer to these subpopulations as "genetic groups".

Phenotypic data analyses

Genetic values (the average phenotypic value of all individuals with the same genotype) of inbreds and hybrids in population B were taken from Flint-Garcia et al. (2009). Genetic values for population A were estimated from the raw phenotypic data using the model:

$$\mathbf{y} = \mathbf{1}\mu + X\mathbf{g} + \varepsilon$$

where $\mathbf{y}$ is the vector of phenotypic values, μ is the mean of $\mathbf{y}$, X is an incidence matrix, $\mathbf{g}$ is the vector of fixed individual effects and ε are the residuals assumed to be $\mathcal{N}(0, \sigma_\varepsilon^2 I)$.

Hybrid vigor for each individual was 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, tassel branch count, tassel angle and upper leaf angle.

Association mapping

SNP association with the genetic values of the inbred lines were tested with the R package EMMA (Kang et al., 2008), following a mixed linear model similar to Yu et al. (2006):

$$\hat{\mathbf{g}} = \mathbf{1}\mu + M\vartheta + S\beta + Z\mathbf{u} + \varepsilon$$

where $\hat{\mathbf{g}}$ is the vector of estimated genetic values for inbred lines, μ is the mean of $\hat{\mathbf{g}}$, M is the tested SNP, ϑ is the SNP effect, S are the structure covariates estimated by Flint-Garcia et al. (2005) using the STRUCTURE software (Pritchard et al., 2000), β are the fixed structure effects, Z is an incidence matrix, $\mathbf{u}$ is a random effect vector assumed to be $\mathcal{N}(0, \sigma_u^2 K)$ and ε are the model residuals assumed to be $\mathcal{N}(0, \sigma_\varepsilon^2 I)$. The coancestry matrix K among inbred lines was approximated by an identity by state matrix calculated with the SNPs. Only SNPs with a minor allele frequency ≥ 0.05 were used for association mapping tests.

In hybrids, we tested the effect of heterozygosity at a given locus on observed heterosis. Each SNP was assigned numerical values corresponding to 0 if the hybrid is homozygous or 1 if the hybrid is heterozygous. The association mapping tests were thus carried out between heterozygosity at a given locus and hybrid vigor:

$$PH = \mathbf{1}\mu' + D\beta + H\vartheta + \varepsilon'$$

where PH is **the vector of heterosis values** (either MPH, BPH_{max} or BPH_{min}), μ' is the mean of PH, D is the genetic distance between the tester (B73 or Mo17) and each inbred line, β is the fixed effect of that distance, H is the tested locus **(1 if heterozygote and 0 if homozygote)**, ϑ the effect of the locus, and ε' is the vector of residuals assumed to be $\mathcal{N}(0, \sigma^2_{\varepsilon'} I)$. SNPs were deemed to be statistically significant at $p \leq 0.001$. Analyses were also conducted controlling for a false discovery rate (Benjamini and Hochberg, 1995) at 20%.

Results and Discussion

Prediction of deleterious mutations

In order to investigate deleterious mutations in a diverse set of maize inbred lines, we first applied two complementary approaches to predict deleterious mutations across the maize genome. We applied the software packages SIFT (Ng and Henikoff, 2003, 2006) and MAPP (Stone and Sidow, 2005) to the 39,656 genes in version 5b of the maize filtered gene set (`http://www.maizesequence.org`; Schnable et al., 2009). SIFT predicted amino acid change consequences for nearly 12 million codons in 32,000 genes, while MAPP obtained predictions for a total of 11 million codons in 29,000 genes combined across the three ortholog datasets used (see Methods). More than 80% of predictions were congruent between the two

approaches, similar to what has been seen in *Arabidopsis* and rice (Günther and Schmid, 2010). SIFT and MAPP respectively identified ∼80 % and 60 % of amino acid polymorphisms as "tolerated", with the remainder predicted to be premature stop codons or "non-tolerated" amino acid changes; we will refer to these latter categories as predicted deleterious SNPs.

We then took advantage of recently published genotyping-by-sequencing (GBS; Elshire et al., 2011) data to survey potentially deleterious mutations across a panel of 247 diverse maize inbred lines (Larsson et al., 2013; Romay et al., 2013). The genotyping data include a total of 437,650 SNPs covering 123,289 codons. SIFT and MAPP predictions were obtained for 112,326 and 107,472 codons representing 19,145 and 18,255 genes, respectively. Nearly 50% of these codons showed no amino acid polymorphism in each dataset; while the vast majority of these monomorphic amino acids were due to synonymous polymorphisms in the GBS data, several hundred predicted deleterious amino acids were fixed across all maize lines analyzed (Supplemental Table 2). Combining results from both SIFT and MAPP, our data consist of 25,352 predicted deleterious SNPs in 11,034 genes.

Characterization of deleterious SNPs in a diversity panel

Across all lines, the derived site frequency spectrum (SFS) of coding SNPs showed an excess of rare variants compared to neutral expectations, with 45% of predicted deleterious SNPs occurring at derived frequencies less than 5% in the SFS across all lines. Even so, non-synonymous SNPs showed an excess of rare variants when compared to synonymous SNPs (Mann-Whitney U test p-value $< 10^{-15}$), and predicted deleterious SNPs showed a marked excess of rare variants compared to both synonymous and non-deleterious non-synonymous variants (Mann-Whitney U test p-value $< 10^{-15}$ for both comparisons; Figure 1). The SFS of non-deleterious non-synonymous was not distinguishable from that of syn-

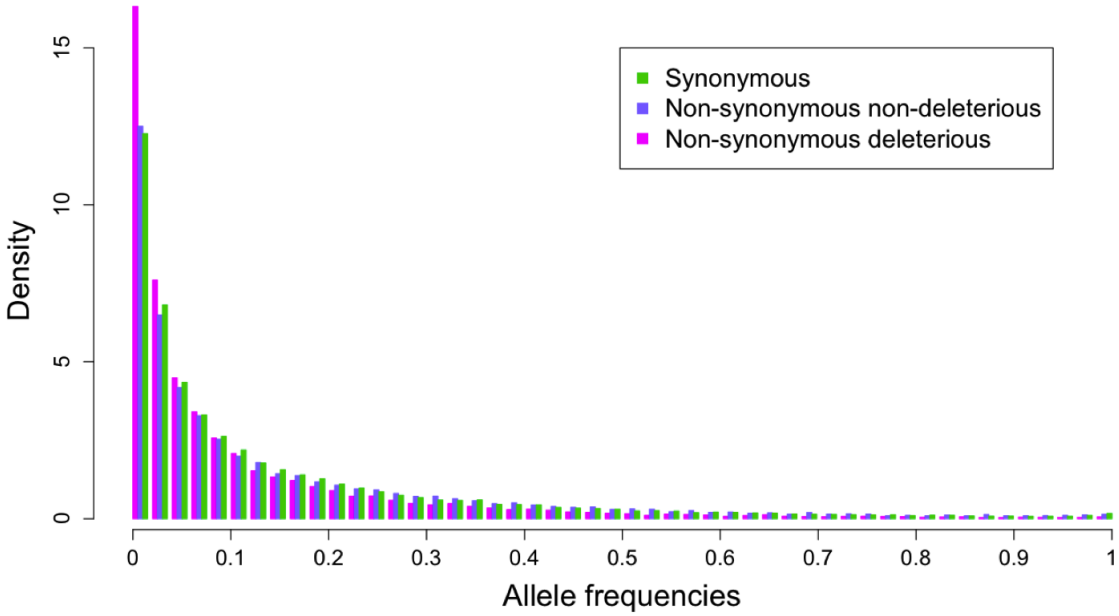


Figure 1: Derived site frequency spectrum of synonymous, non-synonymous non-deleterious and non-synonymous deleterious SNPs. *Tripsacum* was used as outgroup for identifying the derived allele.

onymous variants (Mann-Whitney U test p-value= 0.07). These observations are consistent with the action of weak purifying selection (Cummings and Clegg, 1998; Fay et al., 2001) and independent corroboration of the utility of MAPP and SIFT in predicting deleterious variants.

Although most predicted deleterious alleles were rare, 923 were found segregating at high frequency (≥ 0.80) across all lines. To test whether these alleles may have been driven to high frequency by selection at linked loci during domestication (Lu et al., 2006), we analyzed the pattern of haplotype sharing across the genome (PHS statistics; Toomajian et al., 2006). Only 87 of these SNPs (9.4 % of all tests) showed signs of positive selection in at least one of the genetic groups, and only 25 (2.7 %) were found in candidate regions for selection during maize domestica-

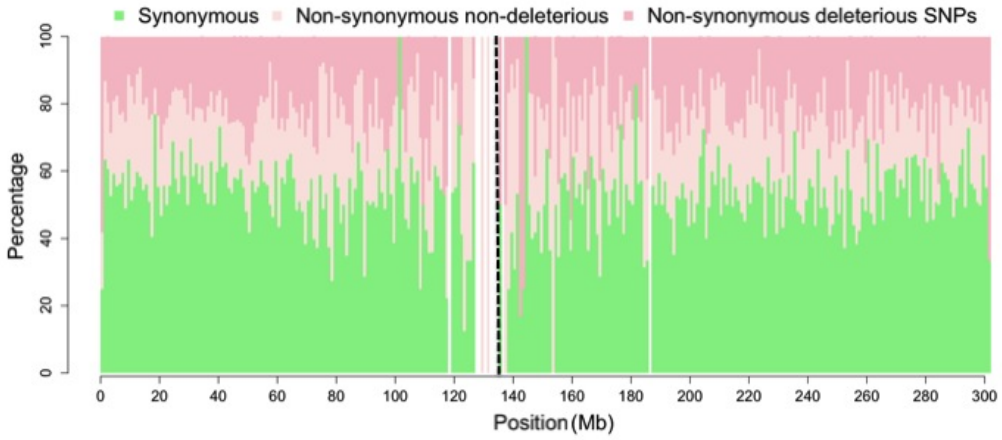


Figure 2: Proportion of genic SNPs predicted to be synonymous, non-synonymous non-deleterious and non-synonymous deleterious in 1 Mb windows along chromosome 1. The vertical dashed black line indicates the centromere position and blank lines indicate windows with missing data.

tion (Hufford et al., 2012), providing little evidence to support hitchhiking during domestication as a major influence on the distribution of deleterious alleles in the genome.

The proportion of genic SNPs predicted to be deleterious appeared relatively uniform (Figure 2 and Supplemental Figure 3) across the genome, showing a very low correlation with recombination rate (Pearson's r of 0.06; p-value =0.005) from the IBM (Intermated B73xMo17) genetic map (Gerke et al., 2013). Explicit comparison of 1,778 non-synonymous pericentromeric ($\pm$ 5 cM around the functional centromere) SNPs did not show an elevated proportion of predicted deleterious SNPs in comparison to the whole genome (Fisher's Exact Test p-value = 0.68) and no correlation was observed between gene density and the proportion of predicted deleterious mutation in 1 Megabase windows (Pearson's r of -0.06; p-value =0.01). The negative correlation between recombination and residual heterozygos-

ity observed in recombinant inbred lines of the maize nested association mapping population has been attributed to the inefficiency of selection against deleterious alleles in low recombination regions of the genome (McMullen et al., 2009; Gore et al., 2009). Our results do not provide support for this explanation, perhaps suggesting that recombination in these regions over longer periods of time is sufficient to avoid the accumulation of deleterious alleles. Consistent with this idea, while regions of the *Drosophila* genome completely lacking in recombination showed a severe reduction in the efficacy of selection, little difference was observed between regions with high and low rates of recombination (Haddrill et al., 2007).

Individual lines varied considerably in their content of predicted deleterious alleles, carrying between 4 and 16% of all predicted deleterious alleles. Lines from the stiff stalk group carried on average fewer deleterious mutations (9%) than did lines from other groups (14-15%), even after weighting by the total SNPs in each group (data not shown). Although drift due to a historically low N_e (Messmer et al., 1991) could explain this observation, other groups with low N_e such as the popcorns do not show such a trend. Instead, we posit that both the SIFT and MAPP algorithms may be biased against identifying deleterious alleles found in the reference B73 genome. Because B73 is a stiff stalk line and both programs use the reference allele in identifying deleterious alleles, non-synonmyous SNPs at appreciable frequency in the stiff stalk group may be more likely falsely identified as tolerated. Similar bias has recently been described in analyses of the human genome (Simons et al., 2013).

Allele sharing at predicted deleterious SNPs generally followed genome-wide patterns of identity by state (IBS). Within the non-stiff stalk, tropical, popcorn and sweet groups, correlations were generally high (Pearson's r of 0.75-0.99) between numbers of shared predicted deleterious alleles (mean of 5 -10%) and IBS. Correlations between inbreds from different genetic groups were much lower (r of

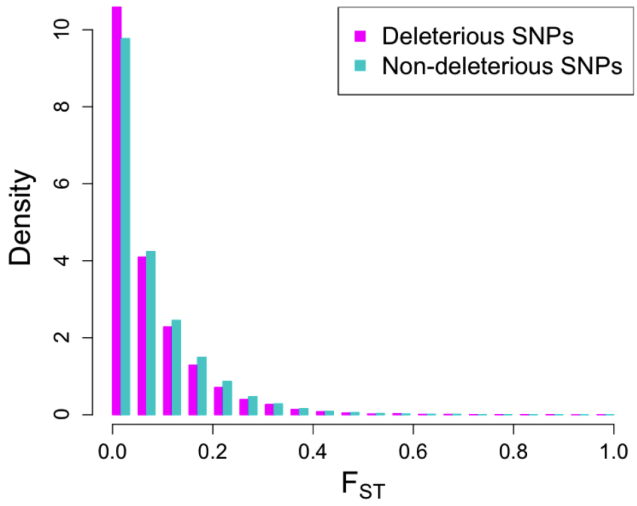


Figure 3: F_{ST} distribution for deleterious and non-deleterious SNPs

0.25-0.52), however, as has been previously seen in correlations between IBS and heterosis observed at SSR loci (Flint-Garcia et al., 2009). The "mixed" (within group $r = 0.22$ and $r = -0.05$ *to* 0.36 with other groups) and stiff stalk (within-group $r = 0.15$ and $r = -0.65$ *to* 0.16 with other groups) groups appeared exceptions to this pattern, perhaps due to the aforementioned ascertainment bias or previously unrecognized population substructure within these groups (Supplemental Figure 4).

Across all genetic groups, levels of population differentiation were slightly lower for predicted deleterious (mean $F_{ST} = 0.07$) than non-deleterious (mean $F_{ST} = 0.08$) SNPs (Mann-Whitney U test p-value $< 10^{-15}$; Figure 3). After correcting for allele frequencies in both classes, however, these differences disappeared and the proportion of deleterious SNPs in the top 1% of F_{ST} was not significantly different from the proportion observed for synonymous SNPs (Fisher's Exact Test p-value = 0.94) or all SNPs in genic regions (Fisher's Exact Test p-value = 0.51). After allele frequency correction, 287 genes had a predicted deleterious SNP in

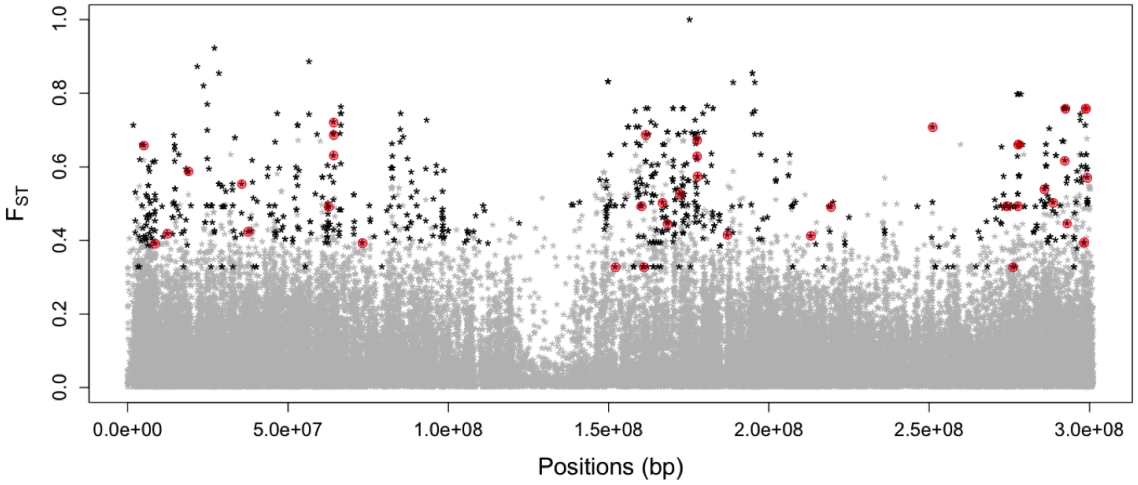


Figure 4: Distribution of F_{ST} along chromosome 1. Black dots represent SNPs in the top 1% of F_{ST} and those predicted to be deleterious are surrounded in red.

the top 1% of Fst among genetic groups, and 30 genes had 2 or more high-Fst predicted deleterious SNPs (see Figure 4 for chromosome 1). Only eleven genes (4%) with high-Fst deleterious SNPs are found in regions thought to be selected during maize improvement (Hufford et al., 2012) and only 44 of the 287 genes (15%) show significant signs of positive selection with the PHS statistic. Neither result provides much evidence that selection on linked beneficial mutations strongly impacts frequencies of deleterious alleles.

Comparisons of the predicted deleterious SFS between stiff stalk, non stiff stalk, and tropical groups (Figure 5) mirrored patterns of between-group F_{ST}, revealing few fixed differences between groups and generally low frequencies within groups, as well as higher differentiation in comparisons involving the stiff stalk group.

Observed frequencies of deleterious SNPs in different populations (Figure 5) may help explain patterns of hybrid vigor. Though F_{ST} is generally low, inbreds from different genetic groups are nonetheless likely to share fewer deleterious variants than inbreds from the same group, and heterosis is higher among crosses be-

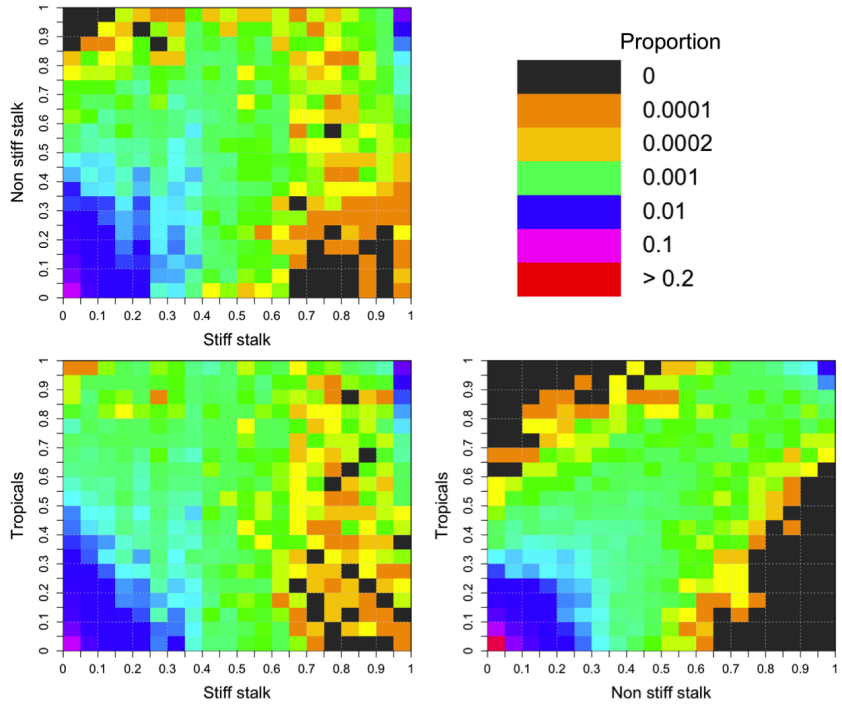


Figure 5: Joint site frequency spectrum of stiff-stalk, non stiff-stalk and tropical genetic groups. Axes represent the frequency of the predicted deleterious alleles in a group and colors show the proportion of SNPs at a given frequency.

tween groups (Supplemental Figure 5). Nonetheless, even crosses among inbreds from the same genetic group show evidence of heterosis (Supplemental Figure 1-A), likely due to the large number of deleterious SNPs segregating at low frequencies within individual populations.

Effect of deleterious mutations on traits of interest

To investigate the contribution of predicted deleterious alleles to observed levels of heterosis and inbreeding depression, we performed a genome wide association analysis of 17 traits evaluated in two populations (see Methods). Analyses were carried out using the genetic values of inbred lines and both mid-parent and best-

Table 1: Total number of significant SNPs in genic regions (n) and fold enrichment (f) for deleterious SNPs in population A. Numbers marked with "*" are statistically significant (Fisher's exact test p-value< 0.05).

	Inbreds		BPH		MPH	
Traits	n	f	n	f	n	f
Days to tasseling	475	1.05	3372	1.15*	1123	1.12
Tassel length	458	0.81	297	1.21	365	1.16
Tassel branch count	300	0.98	4077	0.98	1257	1.12
Tassel angle	244	1.11	490	0.93	646	1.18
Plant height	282	0.92	18068	0.98	9712	0.93
Upper leaf angle	415	1.20	8927	0.99	2266	1.12
Leaf width	289	1.21	1064	1.16	1051	1.01
Leaf length	389	1.14	4256	0.93	2257	1.07
Kernel height	292	1.10	8752	1.08	4512	1.01
Stem puncture resistance	258	0.79	443	1.04	375	0.93
Plant yield	257	1.50	7440	1.12*	7007	1.14*
Ear length	231	0.89	605	1.11*	907	1.00
10 kernel weight	298	1.29	709	1.15	761	1.30
Cob diameter	219	1.04	4363	1.16*	405	0.88
Cob weight	228	1.09	1746	0.93	519	0.69
Kernel weight	256	0.88	3781	0.98	2045	0.95

parent heterosis. Genome wide association results using the genetic values of inbred lines identified between 219 (cob diameter) and 598 (cob length) significant SNPs with a high proportion of genic loci (up to 70%) but little evidence for significant enrichment of predicted deleterious SNPs (Table 1 and Supplemental Table 4).

Results for association between SNP heterozygosity and heterosis showed highly variable numbers of significant loci (Table 1 and Supplemental Table 4), also with a high proportion of genic SNPs (up to 74%). Significant loci explained between 4 and 40% of the observed phenotypic variation in heterosis; though these values

are likely inflated due to small sample size (Beavis, 1994). The highest numbers of associated SNPs were observed for plant height and yield-related traits which also showed the highest levels of observed heterosis. Furthermore, most traits exhibited some enrichment ($5-45\%$) of predicted deleterious SNPs and the enrichment was statistically significant for whole plant yield and days to tasseling. These enrichment results hold even after an FDR control at 20% and similar enrichments were observed when comparing non-synonymous deleterious to non-synonymous non-deleterious SNPs (data not shown).

Because most deleterious SNPs are at frequencies too low for inclusion in association analyses (Figure 1), we expanded our test of enrichment to the gene level, asking whether genes with predicted deleterious SNPs were more likely than random to have SNPs significantly associated with traits of interest. At this level we see much stronger evidence of enrichment, even with an FDR control at 20%: a number of traits show statistically significant enrichment in population A, but virtually all traits in both populations show a positive enrichment for genes with predicted deleterious SNPs (Tables 2 and Supplemental Table 5), a result that is highly unlikely by chance (sign test p-value=3 10^{-5} for population A and 0.01 for population B). Similar tests of low-frequency synonymous SNPs show no evidence of enrichment (p-value ≈ 1), and the low correlation between total SNPs in a gene and the number of significant associations ($r \leq 0.2$) suggests that our observation is not an artifact of the number of SNPs analyzed per gene. Furthermore, the enrichment result holds for groups of genes with similar numbers of SNPs.

We posit that the observed excess of significant associations in genes with predicted deleterious variants may be due to so-called synthetic associations between rare deleterious alleles and a common allele at a linked locus at high enough frequency to be included in association mapping tests (Dickson et al., 2010; Goldstein, 2009). Recent work suggests that this sort of association is only likely to hold for

Table 2: Total number of genes with significant SNPs (n) and fold enrichment for genes with predicted deleterious SNPs(f) in population A. Numbers marked with "*" are statistically significant (Fisher's exact test p-value< 0.05).

	Inbreds		BPH		MPH	
Traits	n	f	n	f	n	f
Days to tasseling	176	1.11	1137	1.12*	429	1.15*
Tassel length	173	1.08	128	1.14	154	1.20
Tassel branch count	114	1.02	1257	1.13*	472	1.14*
Tassel angle	103	1.03	177	1.10	254	1.15
Plant height	128	1.22	4529	1.10*	2741	1.10*
Upper leaf angle	166	1.13	2553	1.11*	810	1.15*
Leaf width	112	1.27	379	1.05	375	1.14
Leaf length	141	1.18	1290	1.13*	821	1.20*
Kernel height	123	1.09	2633	1.13*	1506	1.14
Stem puncture resistance	99	1.24	164	1.10	145	1.07
Plant yield	117	1.22	2440	1.14*	2302	1.14*
Ear length	84	1.02	230	1.20	333	1.15
10 kernel weight	137	1.18	288	1.17	308	1.13
Cob diameter	90	1.10	1419	1.13*	162	1.12
Cob weight	99	1.19	548	1.07	176	1.13
Kernel weight	101	1.18	1228	1.11*	714	1.07

deleterious alleles with a relatively small effect on phenotype (Thornton et al., 2013), which is consistent with the expected weak to intermediate effects of deleterious alleles likely to be involved in heterosis (Charlesworth and Charlesworth, 1987; Whitlock et al., 2000; Glémin et al., 2003; Charlesworth and Willis, 2009). Strongly deleterious alleles, though potentially playing a role in inbreeding depression (Whitlock et al., 2000), are less likely to be observed in our study as selection should effectively remove them from our panel of inbred lines.

Although we have analyzed only a relatively small subset of the genome-wide diversity of maize (Chia et al., 2012), our data nonetheless present the first genome-

wide scan of deleterious coding variants in maize. Our results provide evidence for the contribution of deleterious mutations to heterosis via complementation, consistent with the dominance hypothesis. The weak expected effects of these deleterious SNPs, combined with their low frequencies, make their detection difficult using conventional approaches. *A priori* prediction of the potential effect of rare polymorphisms, however, may improve predictions of inbred line breeding values and combining ability. Future analysis of full genome sequence data, allowing for the inclusion of all coding SNPs and noncoding variants, will provide an even richer catalog of variants that will expand our understanding of the role of rare deleterious variants in maize breeding.

Acknowledgments

We would like to thank S. Flint-Garcia and S. Takuno for help with data analysis; E.S. Buckler for early access to the genotyping data; C. Romay and J. Glaubitz for bioinformatics support; G. Coop, J. Gerke, P. Morrell, P. Ralph, and O. Smith for helpful comments on an earlier version of the manuscript; and two anonymous reviewers for their constructive comments. This project was supported by Agriculture and Food Research Initiative Competitive Grant 2009-01864 from the USDA National Institute of Food and Agriculture as well as a grant from DuPont Pioneer.

References

Beavis, W. (1994). *The power and deceit of QTL experiments: lessons from comparitive QTL studies.*

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*, 57:289–300.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The swiss-prot protein knowledgesbase and its supplement trembl in 2003. *Nucleic Acids Res*, 31(1):365–370.

Brandvain, Y., Slotte, T., Hazzouri, K. M., Wright, S. I., and Coop, G. (2013). Genomic identification of founding haplotypes reveals the history of the selfing species capsella rubella. *PLoS Genet*, 9(9):e1003754.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple arabidopsis thaliana populations. *Nat Genet*, 43(10):956–63.

Charlesworth, D. and Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Ann. Rev. Ecol. Syst.*, 18:237–68.

Charlesworth, D., Morgan, M. T., and B, C. (1993). Mutation accumulation in finite populations. *Journal of Heredity*, 84:321–325.

Charlesworth, D. and Willis, J. H. (2009). The genetics of inbreeding depression. *Nat Rev Genet*, 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.

Chun, S. and Fay, J. C. (2011). Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genet*, 7(8):e1002240.

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., and Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of hdl cholesterol. *Science*, 305(5685):869–72.

Cook, J. P., McMullen, M. D., Holland, J. B., Tian, F., Bradbury, P., Ross-Ibarra, J., Buckler, E. S., and Flint-Garcia, S. A. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol*, 158(2):824–34.

Cummings, M. P. and Clegg, M. T. (1998). Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (hordeum vulgare ssp. spontaneum): an evaluation of the background selection hypothesis. *PNAS*, 95:5637–5642.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol*, 8(1):e1000294.

East, E. (1907-1908). *Reports of the Connecticut Agricultural Experiment Station for Years*, volume Inbreeding in corn,. Connecticut Agricultural Experi-

ment Station, New Haven, Connecticut Agricultural Experiment Station, New Haven,CT.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLoS One*, 6(5):e19379.

Fay, J., Wyckoff, G., and CI, W. (2001). Positive and negative selection on the human genome. *Genetics*, 158:1227–1234.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, 78(2):737–756.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Flint-Garcia, S. A., Buckler, E. S., Tiffin, P., Ersoz, E., and Springer, N. M. (2009). Heterosis is prevalent for multiple traits in diverse maize germplasm. *PLoS One*, 4(10):e7433.

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J*, 44(6):1054–64.

Fournier-Level, A., Korte, A., Cooper, M., Nordborg, M., Schmitt, J., and Wilczek, A. (2011). A map of local adaptation in arabidopsis thaliana. *Science*, 334(6052):86–89.

Gerke, J. P., Edwards, J. W., Guill, K. E., Ross-Ibarra, J., and McMullen, M. D. (2013). The genomic impacts of drift and selection for hybrid performance in maize. *http://arxiv.org/abs/1307.7313*.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat Rev Genet*, 13(2):135–45.

Glémin, S., Ronfort, J., and Bataillon, T. (2003). Patterns of inbreeding depression and architecture of the load in subdivided populations. *Genetics*, 165(4):2193–212.

Goldstein, D. B. (2009). Common genetic variation and human traits. *N Engl J Med*, 360(17):1696–8.

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*, 326(5956):1115–7.

Gossmann, T., Song, B., Windsor, A., Mitchell-Olds, T., Dixon, C., Kapralov, M., Filatov, D., and Eyre-Walker, A. (2010). Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol*, 27(8):1822–1832.

Günther, T. and Schmid, K. J. (2010). Deleterious amino acid polymorphisms in arabidopsis thaliana and rice. *Theor Appl Genet*, 121(1):157–68.

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 Biol*, 8(2):R18.

Hill, W. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet. Res.*, 8(3):269–294.

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., Elshire, R. J., Glaubitz, J. C., Guill, K. E., Kaeppler, S. M.,

Lai, J., Morrell, P. L., Shannon, L. M., Song, C., Springer, N. M., Swanson-Wagner, R. A., Tiffin, P., Wang, J., Zhang, G., Doebley, J., McMullen, M. D., Ware, D., Buckler, E. S., Yang, S., and Ross-Ibarra, J. (2012). Comparative population genomics of maize domestication and improvement. *Nat Genet*, 44(7):808–11.

Hughes, A. L. (2005). Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics*, 169(2):533–8.

Joseph, S. B. and Hall, D. W. (2004). Spontaneous mutations in diploid saccharomyces cerevisiae: more beneficial than expected. *Genetics*, 168(4):1817–25.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press, UK: Cambridge.

Lande, R. (1994). Risk of population extinction from fixation of new deleterious mutations. *Evolution*, 48:1460–1469.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–8.

Larsson, S. J., Lipka, A. E., and Buckler, E. S. (2013). Lessons from dwarf8 on the strengths and weaknesses of structured association mapping. *PLoS Genet*, 9(2):e1003246.

Lohmueller, K. E. (2013). The impact of population demography and selectionon genetic architecture of complex traits. *arXiv.org*.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., and Bustamante, C. D. (2008). Proportionally more deleterious genetic variation in european than in african populations. *Nature*, 451(7181):994–7.

Lu, J., Tang, T., Tang, H., Huang, J., Shi, S., and Wu, C.-I. (2006). The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet*, 22(3):126–31.

Lynch, M. S. and Gabriel, W. (1990). Mutationload and the survical of small populations. *Evolution*, 44:1725–1737.

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*, 325(5941):737–40.

Messmer, M., Melchinger, A., Lee, M., Woodman, W., and Lamkey, K. (1991). Genetic diversity among progenitors and elite lines from the iowa stiff stalk synthetic (bsss) maize population: comparison of allozyme and rflp data. *Theor Appl Genet*, 38(1):97–107.

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.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246:96–98.

Paape, T., Bataillon, T., Zhou, P., J, Y. K. T., Briskine, R., Young, N., and Tiffin, P. (2013). Selection, genome-wide fitness effects and evolutionary rates in the model legume medicago truncatula. *Mol Ecol*, 22(13):3525–3538.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

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

Schnable, J. C., Freeling, M., and Lyons, E. (2012). Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol*, 4(3):265–77.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla,

A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009). The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–5.

Shull, G. H. (1908). The composition of a field of maize. *The Journal of Heridity*, 4(1):296–301.

Simons, Y. B., Turchin, M. C., and Pritchard, J. K. (2013). The deleterious mutation load is sensitive to recent population history. *http://arxiv.org/abs/1305.2061*.

Smigrodzki, R., Parks, J., and Parker, W. D. (2004). High frequency of mitochon-

drial complex i mutations in parkinson's disease and aging. *Neurobiol Aging*, 25(10):1273–81.

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.

Strable, J. and Scanlon, M. J. (2009). Maize (zea mays): a model organism for basic and applied research in plant biology. *Cold Spring Harb Protoc*, 2009(10):pdb.emo132.

Tellier, A., Fischer, I., Merino, C., Xia, H., Camus-Kulandaivelu, L., Städler, T., and Stephan, W. (2011). Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity (Edinb)*, 107(3):189–99.

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

Thornton, K. R., Foran, A. J., and Long, A. D. (2013). Properties and modeling of gwas when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. *PLoS Genet*, 9(2):e1003258.

Toomajian, C., Hu, T. T., Aranzana, M. J., Lister, C., Tang, C., Zheng, H., Zhao, K., Calabrese, P., Dean, C., and Nordborg, M. (2006). A nonparametric test reveals selection for rapid flowering in the arabidopsis genome. *PLoS Biol*, 4(5):e137.

Wang, J., Hill, W., Charlesworth, D., and Charlesworth, B. (1999). Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. *Genet. Res.*, 74:165–178.

Whitlock, M., Ingvarsson, P., and Hatfield, T. (2000). Local drift load and the heterosis of interconnected populations. *Heridity*, 84:452–457.

Whitlock, M. C., Grisworld, C. K., and Peters, A. D. (2003). Compensating for meltdown: The critical effective size of a population with deleterious and compendatory mutations. *Ann. Zool. Fennici*, 40:169–183.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 38(2):203–8.

Zhu, Y., Spitz, M., Amos, C., Lin, J., Schabath, M., and Wu, X. (2004). An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. *Cancer Res*, 64(6):2251–2257.

Supplementals

List of the inbred lines used

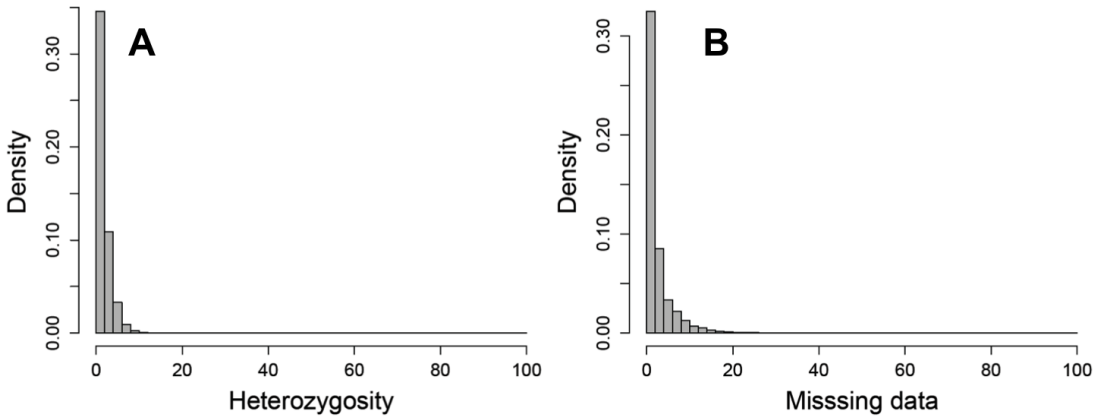
PopulationA

B73, A214N, A441.5, A554, A556, A6, A619, A632, A634, A635, A641, A654, A659, A661, A679, A680, A682, AB28A, B10, B104, B105, B109, B115, B14A, B164, B2, B37, B46, B57, B64, B68, B73HTRHM, B75, B76, B77, B79, B84, B97, CH701.30, CH9, CI187.2, CI21E, CI28A, CI31A, CI3A, CI64, CI66, CI7, CI90C, CI91B, CM174, CM37, CM7, CML10, CML103, CML108, CML11, CML14, CML154Q, CML157Q, CML158Q, CML218, CML220, CML228, CML238, CML247, CML258, CML261, CML264, CML277, CML281, CML287, CML311, CML314, CML321, CML322, CML323, CML328, CML331, CML332, CML333, CML341, CML38, CML5, CML52, CML69, CML77, CML91, CML92, CMV3, CO255, D940Y, DE1, DE2, DE811, E2558W, EP1, F2834T, F44, F6, GA209, GT112, H105W, H84, H91, H95, H99, HI27, HP301, HY, I137TN, I205, I29, IA2132, IA5125, IDS28, IDS69, IDS91, IL101T, IL14H, IL677A, K148, K4, K55, K64, KI11, KI14, KI2021, KI21, KI3, KI43, KI44, KY21, KY226, KY228, L317, L578, M14, M162W, M37W, MEF156.55.2, MO17, MO18W, MO1W, MO24W, MO44, MO45, MO46, MOG, MP339, MS1334, MS153, MS71, MT42, N192, N28HT, N6, N7A, NC222, NC230, NC232, NC236, NC238, NC250, NC258, NC260, NC262, NC264, NC294, NC296, NC296A, NC298, NC300, NC302, NC304, NC306, NC310, NC314, NC318, NC320, NC324, NC326, NC328, NC33, NC336, NC338, NC342, NC344, NC346, NC348, NC350, NC352, NC354, NC356, NC358, NC360, NC362, NC364, NC366, NC368, ND246, OH40B, OH43E, OH603, OH7B, OS420, P39, PA762, PA875, PA880, PA91, R168, R177, R229, R4, SA24, SC357, SC55, SD44, SG1533, SG18, T232, T8, TX303, TZI10, TZI11, TZI16, TZI18, TZI25, TZI8, TZI9, U267Y, VA102, VA14, VA22, VA35, VA59, VA99, VAW6, W117HT, W153R, W182B, W64A, WD,

33.16, 38.11, X4226, X4722

PopulationB

B73, MO17, 33.16, A188, A239, A619, A632, A634, A635, A641, A654, A661, A679, A680, A682, B103, B104, B109, B115, B14A, B37, B46, B52, B57, B64, B68, B73, B73HTRHM, B75, B76, B77, B79, B84, C103, C49A, CH701.30, CM105, CM174, CO125, DE.2, DE1, DE811, EP1, H105W, H49, H84, H91, H95, H99, HP301, IL101, IL14H, K148, KY226, M14, MEF156.55.2, MO44, MO45, MO46, MO47, MS1334, MS153, MS71, N192, N28HT, N6, NC262, NC264, NC294, NC306, NC310, NC314, NC324, NC326, NC328, NC342, NC364, ND246, OH43, OH43E, OS420, P39, PA762, PA875, PA880, PA91, R168, R177, R4, SD40, SD44, SG18, VA102, VA14, VA17, VA22, VA35, VA85, VA99, W182B, W22, W64A, WF9, YU796.NS.



Sup. Fig. 1: Histograms of the percentage of (A) heterozygosity and (B) missing data per SNP

List of genomes used for reciprocal BLAST

Aquilegia coerulea, Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Brassica rapa, Capsella rubella, Carica papaya, Chlamydomonas reinhardtii, Citrus clementina, Citrus sinensis, Cucumis sativus, Eucalyptus grandis, Glycine max, Linum usitatissimum, Malus domestica, Manihot esculenta, Medicago truncatula, Mimulus guttatus, Oryza sativa, Panicum virgatum, Phaseolus vulgaris, Physcomitrella patens, Populus trichocarpa, Prunus persica, Ricinus communis, Selaginella moellendorffii, Setaria italica, Sorghum bicolor, Thellungiella halophila, Vitis vinifera, Volvox carteri.

Sup. Table 1: List of Analyzed traits

Traits	Abbreviation	Populations
Days to tasseling	DTT	A
Tassel length (cm)	TSLLEN	A
Tassel branch count	TSLBCHCNT	A
Tassel angle	TSANG	A
Plant height (cm)	PLTHT	A & B
Upper leaf angle	UPLFANG	A
Leaf width (cm)	LFWDT	A
Leaf length (cm)	LFLEN	A
Kernel height	KNLHGT	A
Kernel weight	TOTKNLWT	A
Stem puncture resistance (kg/section)	RPR	A
Plant yield (g/plant)	PLTYLD	A
Ear length (cm)	EARLGH	A & B
10 kernel weight (g)	10KWT	A
Cob diameter (cm)	COBDIA	A & B
Cob weight (g)	COBWT	A & B
Seed number per ear	SEEDNB	B

Sup. Table 2: Detailed results of the prediction of deleterious amino acids with MAPP, using the different gene sets, and with SIFT

	MAPP			SIFT
Gene sets	BLASTX	Reciprocal BLAST	Syntenic genes	PSI-BLAST
Total a.a. positions with predictions	7,746,638	5,570,035	6,869,010	11,906,167
Total number of genes	20,348	11,918	17,957	31,843
Number of positions covered by SNPs	74,909	52,283	72,562	112,326
Number of genes covered by SNPs	12,561	8,553	12,615	19,145
Monomorphic tolerated	39,009	25,270	39,300	58,685
Monomorphic not tolerated*	144	3470	14	387
Polymorphic tolerated	18,379	10,753	17,792	42,606
Polymorphic not tolerated*	17,377	12,790	15456	10,648

*Includes premature stop codons

Sup. Table 3: Comparion of the results of MAPP predictions with the different gene sets.

Gene sets	BLASTX	Reciprocal BLAST	Syntenic genes
BLASTX	-	80.1%	78.2%
Reciprocal BLAST	38,054 (6,169)	-	79.8%
Syntenic genes	45,412 (7,745)	32,222 (5,488)	-

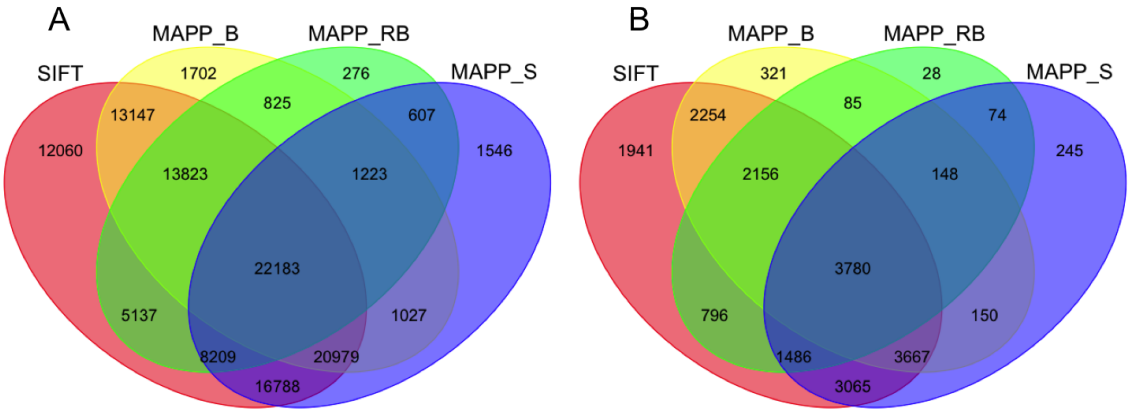
The lower triangle indicates the number of amino acid positions predicted with two given gene sets and covered by GBS SNPs (number of genes between brackets); the upper triangle indicates the percentage of amino acids with the same predictions.

Sup. Table 4: Total number of significant SNPs in genic regions (n) and fold enrichment (f) for loci with deleterious mutations in population B. Numbers marked with "*" are statistically significant (Fisher's exact test p-value< 0.05).

	Inbreds		BPH_B73		MPH_B73		BPH_Mo17		MPH_Mo17	
Traits	*n*	*f*	*n*	*f*	*n*	*f*	*n*	*f*	*n*	*f*
10KWT	310	0.77	404	1.17*	257	0.86	698	0.83	723	0.98
COBWT	313	0.62	941	1.15*	387	0.69	257	1.33	532	0.95
COBDIA	226	1.49	159	1.25*	236	1.06*	349	0.78	615	0.72
COBLEN	598	1.08	239	1.20*	97	0.24	280	1.08	140	0.92
SEEDWT	362	1.09	378	1.32*	118	1.23*	1043	0.92	1080	0.78
SEEDNB	373	0.99	320	0.86	251	0.92	348	1.06	454	0.82
PLTHT	505	1.02	261	0.89	143	1.45*	1022	1.08	156	1.16

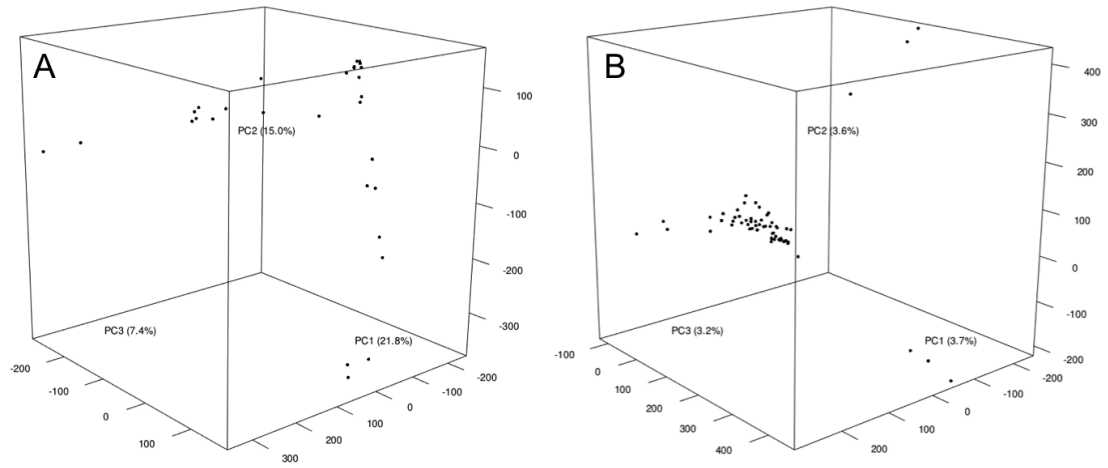
Sup. Table 5: Total number of genes with significant SNPs (n) and fold enrichment for genes with predicted deleterious SNPs (f) in population B. Numbers marked with "*" are statistically significant (Fisher's exact test p-value< 0.05).

	Inbreds		BPH_B73		MPH_B73		BPH_Mo17		MPH_Mo17	
Traits	*n*	*f*	*n*	*f*	*n*	*f*	*n*	*f*	*n*	*f*
10KWT	73	1.17	169	1.14	95	1.11	246	1.11	274	1.11
COBWT	71	1.13	316	1.08	128	1.04	94	1.10	204	1.10
COBDIA	81	1.07	57	1.08	86	1.11	134	1.03	234	1.14
COBLEN	203	1.09	89	1.24	30	1.17	110	1.17	51	1.21
SEEDWT	138	1.10	146	1.14	50	0.97	371	1.09	389	1.09
SEEDNB	106	1.15	128	1.13	116	0.98	130	1.12	166	1.09
PLTHT	169	1.15	112	1.09	65	1.13	348	1.15	65	1.15

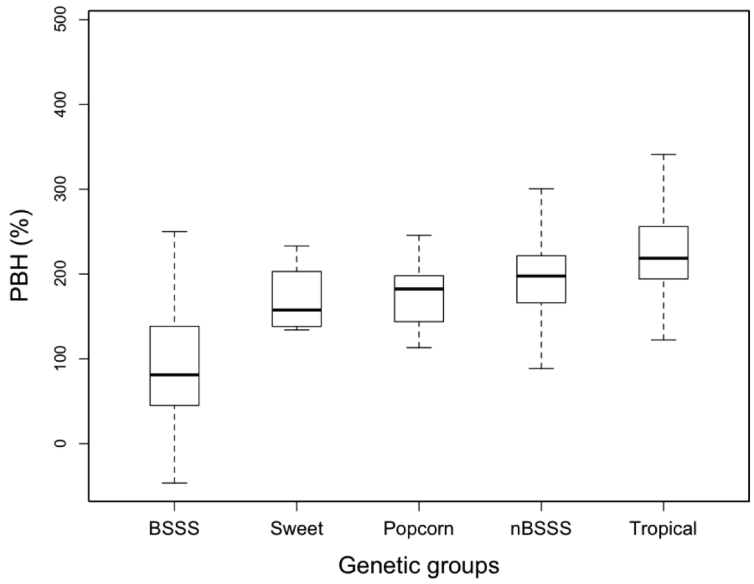


Sup. Fig. 2: Comparison of the number of predicted (A) amino acids and (B) genes, covered by SNP data. For MAPP, 3 gene sets were used: BLASTX (MAPP_B), reciprocal BLAST (MAPP_RB) and syntenic genes (MAPP_S)

Sup. Fig. 3: Proportion of genic SNPs predicted to be synonymous, non-synonymous non-deleterious and non-synonymous deleterious in 1 cM windows along chromosome 1



Sup. Fig. 4: Projection of the (A) stiff stalk and (B) mixed inbred lines on the three first axes of a principal component analysis



Sup. Fig. 5: Distribution of best parent heterosis (BPH) for plant yield in population A. BSSS and nBSSS indicate the stiff stalk and non-stiff stalk groups.