Pattern and distribution of deleterious mutations in maize

Jeffrey Ross-Ibarra

Abstract

Most nonsynonymous mutations are thought to be deleterious because of their effect on protein sequence. Such polymorphisms are expected to be removed or kept at low frequency by the action of natural selection, but in small or inbred populations the effects of genetic drift may also be important. Rare deleterious variants have been implicated as a possible explanation for the 'missing heritability' seen in many studies of complex traits. Here, we make use of genome-wide genotyping data to assess the evolution of 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 genes showing associations 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 contributes to hybrid vigor.

Introduction

Mutation is the driving force behind much of the genetic variation which forms the basis of evolutionary change. While a small minority of new mutations will be beneficial, many may have little consequence for an organisms's fitness, and a large proportion are likely to be deleterious. The mutation-selection balance maintain these deleterious alleles at low frequencies. They can, however, reach moderate to high frequencies and drift to fixation if the selection pressure (s) is small compared to the effective population size ($s < 1/2N_e$; **?**). Different deleterious mutations could thus be fixed in different populations or genetic groups (**??**).

In addition to the mutation effect and effective population size, a number of other factors affect the destiny of deleterious alleles, such as the mating system and the recombination rate.

Selfing species and inbreeding within populations will expose the lethal mutations to selection faster than in an outcrossing population (**?**). The slightly deleterious mutations will however be maintained at moderate frequencies, even with the presence of gene flow between populations (**?**). In low recombination regions, deleterious mutations could hitchhike with advantageous alleles at linked loci (**??**). Fixation of deleterious mutations is slower with an increased recombination rate in a finite population (**?**)

The presence of deleterious allele within loci associated to quantitative traits may explain part of there determinism and heritability. It was also suggested that complementation at deleterious SNPs could explain a non negligible part of heterosis (**?**). Evaluating the pattern of deleterious mutations is thus of interest and has been investigated in the Human genome (**??????**), yeast (**?**), bacteria (**?**), RNA viruses (**?**) and different plant species including rice, *Arabidopsis thaliana* (**???**), and tomatoes (**?**).

? estimated, in 182 Human genes, the fraction of amino acid mutations that

are deleterious to around 80% with only 20% of them being slightly deleterious. However, **?** estimated to only 20% the variants that are damaging to the protein, in 25 genes associated with sex steroid biosynthesis. At a genome level, **?** observed higher proportion of SNPs predicted to be deleterious in african americans in comparison to european americans; they mainly explained this observation by the out of africa bottelneck. **?** examined the temporal pattern of deleterious SNPs at internal and terminal of the human tree in four human populations. The percentage of deleterious mutations seen reached 48% of the amino acid variant specific to a given genome. Also they observed that deleterious fraction of genome specific non synonymous SNPS is up to 7 times higher compared to that shared between human.

In a plant genome wide predictions, **?** observed different proportion of deleterious SNPs within *Arabidopsis thaliana* accessions and fewer deleterious SNPs in wild rice in comparison to the cultivated accessions. Similarly, **?** estimated that around 25% of the differences between rice cultivars are deleterious. Within coding regions of eight house keeping genes, **?** estimated to 90% the proportion of non-synonymous SNPs under selection, suggesting that they may be deleterious.

Maize is a worldwide economically important cereal with the highest yield and largest cultivated area within cereal (FAO statistics, `http://faostat.fao.org`), distributed in a broad range of latitudes and environments (**?**) to which it adapted given its tremendous genetic diversity (**?**).

The presence of a very strong structure into heterotic groups of the maize breeding material and the high observed levels of heterosis makes it interesting to analyze the distribution of the deleterious mutations

The aim of the current study was to make use of the availability of the maize genome sequence, high density single nucleotide polymorphisms (SNPs) and phenotypic data for an enough large sample of inbred lines and hybrids to (1) carry

out a genome wide scan for deleterious mutations using the whole maize B73 reference genes, (2) analyze their distribution within the genome and within different genetic groups and (3) test for enrichment of these loci in the results of genome wide association mapping with both the genetic values of inbred lines and the hybrid vigor for different quantitative traits of interest.

Our results showed that a majority of deleterious alleles are segregating in maize; these alleles are in general at very low frequencies as expected from theory and very few are differentially fixed within different genetic groups. Genome wide association mapping results showed a small enrichment enrichment of these deleterious loci.

Materials and methods

Plant material and phenotypic data

Phenotypic data from 247 maize inbred lines from the diversity panel described by **?** were analyzed in the current study (see supplemental data for a list of inbred lines). Each inbred lines was crossed to the stiff-stalk inbred B73 (population A) and both the inbred lines and their B73-hybrids were evaluated in three environments in 2003 (**?**). A subset of 102 inbreds were additionally crossed to both B73 (population B1) and Mo17 (population B2) and evaluated in a single environment in 2006 (**?**).

Traits measured in both populations include cob diameter (cm), cob weight (g), ear length (cm), plant height (cm), individual kernel weight (g) and total kernel weight (g/ear). Additional traits including days to anthesis, plant yield (g/plant), tassel length (cm), tassel branch count, tassel angle, upper leaf angle, leaf width (cm), leaf length (cm), stem puncture resistance (kg/section), stem width (cm), 10 kernel weight (g) and kernel height (cm) were collected for population A, and seed

number per ear was collected for populations B1 and B2. Details of the phenotypes and measurements can be found in (**?**).

Genotypic data

We made use of genotypic data from (**?**) for the full set of 247 lines, available to download from `http://www.panzea.org/lit/data_sets.html`. Lines were genotyped using the genotype-by-sequencing approach (GBS; **?**) approach, resulting in a total of 437,650 SNPs that were partially imputed. 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 (Supp Fig 1-A), and 0.18% had more than 10% heterozygous genotypes (Supp Fig 1-B).

We estimated error rates by first comparing our genotyped inbred B73 to the B73 reference genome, then by comparison of all our genotypes to those from 7,225 overlapping SNPs on the maize SNP50 bead chip (**?**). Compared to the reference genome, our B73 genotype differed 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 **?**.

Statistical analyses

SNP annotation

SNPs were annotated as synonymous and nonsynonymous using the software polydNdS from the analysis package of libsequence (**?**). The deleterious effects of amino acid changes were predicted for proteins derived from the first transcript of each gene in the B73 5.b filtered gene set using both the SIFT (**??**) and MAPP

(**?**) 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 (**?**), 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 criteria 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 a minimum 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* (**?**). For each set of proteins, ClustalW2 (**?**) was used to align the sequences and build a neighbour-joining tree. A custom R script (available **X**) was used to link amino acid positions to SNP positions and to link the amino acid polymorphisms to MAPP and SIFT predictions.

Phenotypic data analyses

Genetic values of inbreds and hybrids in population B were taken from **?**. Genetic values for population A were estimated from the raw phenotypic data

using the model:

$$Y = \mathbf{1}\mu + ZG + \varepsilon$$

where Y is the vector of phenotypic values, μ is the mean of Y, Z is an incidence matrix, G is the vector of fixed individual effects and ε are the $N(0, \sigma_\varepsilon^2 I)$ residuals.

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, upper leaf angle and rind penetrometer resistance.

Association mapping

SNP association with the genetic values of the inbred lines were tested using the mixed linear model:

$$\hat{G} = \mathbf{1}\mu + M\vartheta + S\beta + Zu + \varepsilon$$

where $\hat{G}$ is the vector of estimated genetic values for inbred lines, μ is the mean of $\hat{G}$, M is the tested SNP, ϑ is the SNP effect, S is the structure covariates estimated by **?**, β is the fixed structure effects, Z is an incidence matrix, u is a random effect vector assumed $N(0, \sigma_\varepsilon^2 K)$ and ε are the model residuals assumed $N(0, \sigma_{\varepsilon'}^2 I)$. 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.

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 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, ϑ the effect of the locus, and ε' is the vector of residuals assumed $N(0, \sigma^2_{\varepsilon'} I)$. SNPs were deemed to be statistically significant at $p \leq 0.001$; analyses were also conducted controlling the false discovery rate (**?**) at 10%.

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 (**??**) and MAPP (**?**) to the 39,656 genes in version 5b.60 of the maize filtered gene set (`http://www.maizesequence.org`; **?**). 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). Among the codons covered by GBS SNPs, predictions were made by both SIFT and MAPP for 20,195 genes (95%).

More than 80% of predictions were congruent between the two approaches; an overlap similar to what has been seen in *Arabidopsis thaliana* and rice (**?**). 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 (**?**) data to survey potentially deleterious mutations across a panel of 247 diverse maize inbred lines (**?**). The genotyping data covered 112,326 and 107,472 codons representing 19,145 and 18,255 genes in the SIFT and MAPP data, 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 1).

Characterization of deleterious SNPs in a diversity panel

Across all lines, the site frequency spectrum (SFS) of coding SNPs showed an excess of rare variants compared to neutral expectations, with 45% of SNPs at a frequency lower than 5% across all lines. Even so, nonsynonymous SNPs showed an excess of rare variants when compared to synonymous SNPs (Mann-Whitney U test p-value $< 2.2\ 10^{-16}$; Figure 1-A), and putatively deleterious SNPs showed a marked excess of rare variants (Mann-Whitney U test p-value $< 2.2\ 10^{-16}$; Figure 1-B) compared to other nonsynonymous variants. These observations are consistent with the action of weak purifying selection (**?**) and provide a measure of independent corroboration of the utility of MAPP and SIFT in predicting deleterious variants.

Although most predicted deleterious alleles were rare, 923 were found segre-

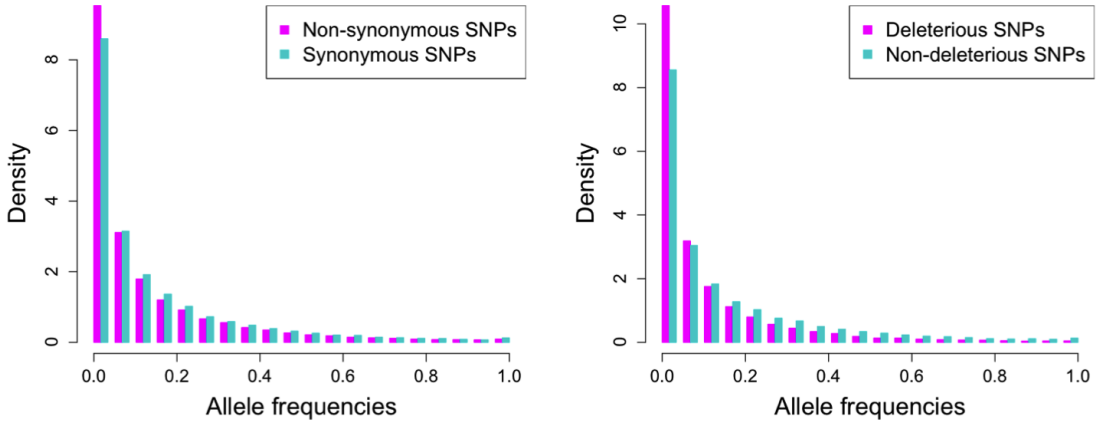


Figure 1: Site frequency spectrum of (A) synonymous *vs* non synonymous SNPs and (b) non-synonymous non-deleterious *vs* non-synonymous deleterious SNPs

gating at high frequency (≥ 0.80) across all lines. To test whether these alleles may have been driven to high frequency by selection during domestication (**?**) we analyzed the pattern of haplotype sharing across the genome (**?**) within each of the tropical, stiff-stalk, non-stiff stalk and mixed genetic groups as defined by **?**. Only 87 SNPs (9.4% of all tests) showed signs of positive selection in at least one of the genetic groups with only 16 SNPs in candidate regions for selection during maize domestication or improvement (**?**), providing little evidence, in the analyzed material, to support linked hitchhiking during domestication as major influence on deleterious alleles in the genome.

Across the genome, the proportion of genic SNPs predicted to be deleterious appeared relatively uniform (Figure 2 and Supplemental Figure 3), a very low but slightly significant correlation was observed between the proportion of predicted deleterious SNPs and recombination rate (Pearson r of 0.06; p-value =0.005), and explicit comparison of 1,778 nonsynonymous 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).

The negative correlation between recombination and residual heterozygosity 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 (**??**). Our results do not provide strong 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 sever reduction in the efficacy of selection, little difference was observed between regions with high and low rates of recombination (**?**).

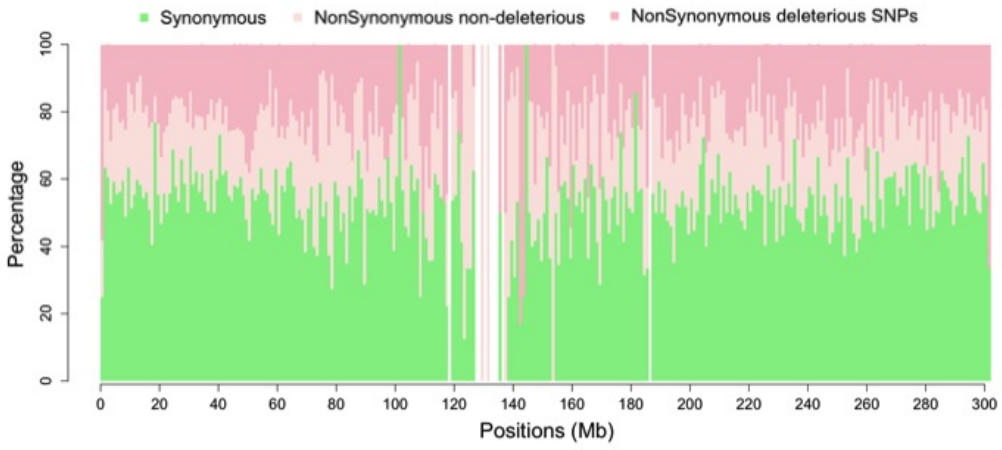


Figure 2: Proportion of genic SNPs predicted to be synonymous, non-deleterious nonsynonymous and deleterious nonsynonymous in 1Mb windows along chromosome 1

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 heterotic group carried on average fewer deleterious mutations (9%) than did lines from other groups (14-15%). Although a historically low N_e (**?**) 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 alleles found in the reference B73 genome which belongs to the stiff stalk heterotic group; similar bias has recently been described in analyses of the human genome (**?**).

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 heterotic groups, correlations were generally high (Pearson 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 25 - 52 %), however, reflecting correlations seen between IBS and heterosis observed at SSR loci (**?**). The "mixed" (within group $r = 0.22$, $r = -0.05$ *to* 0.36 with other groups) and stiff stalk (within-group $r = 0.15$, $r = -0.65$ *to* 0.16 with other groups) groups appeared exceptions to this rule, perhaps due to previously unrecognized population substructure (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 $< 2.2\ 10^{-16}$; Figure 3). After correcting for allele frequencies in both classes, however, these differences disappeared, and the proportion of deleterious SNPs in the top 1% 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). Nonetheless, after controlling for allele frequency a number of predicted deleterious SNPs do show signs of significant differentiation among groups (Figure 4). These SNPs are located in 287 genes (30 genes with 2 or more deleterious SNPs in the

top 1%) including 11 and 9 genes previously identified as candidates or in selected regions during respectively maize improvement and domestication (**?**).

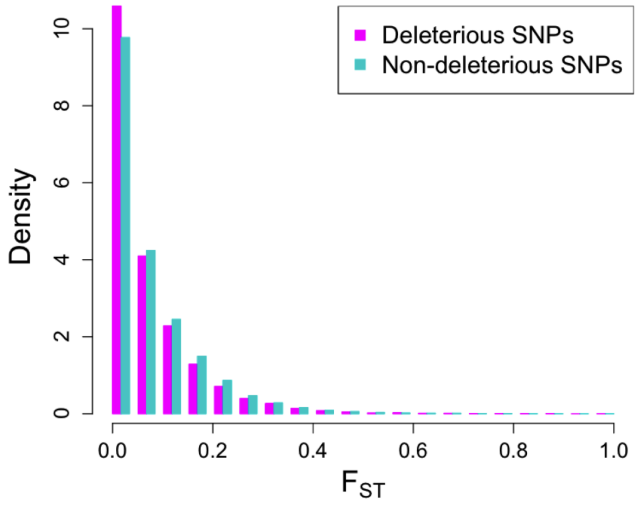


Figure 3: F_{ST} distribution for deleterious and non-deleterious SNPs

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 deleterious SNPs at generally low frequencies and few fixed differences as well as higher differentiation in comparisons including the stiff stalk group (Figure 5). Non fixation of deleterious alleles within groups is in agreement with observed heterosis in crosses of inbred lines from the same group. Although, frequency differences of deleterious alleles leads to higher heterosis of between-group crosses as a result of less deleterious alleles shared and in agreement with previous studies (**?**).

Effect of deleterious mutations on traits of interest

We performed a genome wide association analysis to investigate the role of predicted deleterious alleles in observed levels of heterosis and inbreeding depression. We looked for association of all SNPs with a $MAF > 0.05$ with 17 traits evaluated

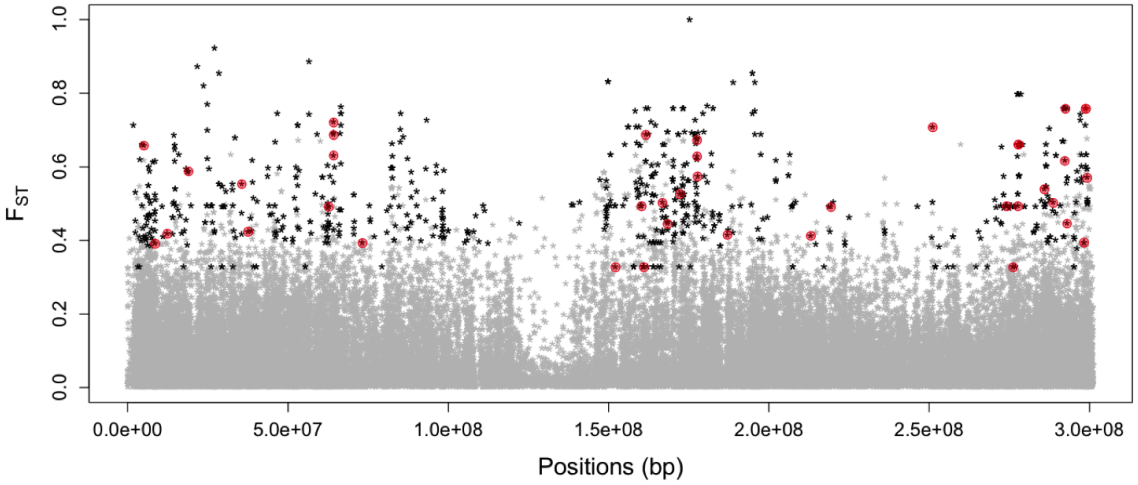


Figure 4: Distribution of F_{ST} along chromosome 1; black dots represent top 1% SNPs, the predicted deleterious are surrounded in red.

in two populations while controlling for population structure (see Methods). Analyses were carried out using the genetic values of inbred lines and both mid-parent and best-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 (up to 70%) of genic loci. Predicted deleterious SNPs showed, however, little evidence for significant enrichments among genome wide association results (Table 1 and Supplemental Table 3).

Association results between heterozygosity and heterosis showed highly variable numbers of significant loci (Tables 1 and Supplemental Table 3) with, as observed for inbred values, a high proportion of genic SNPs. The highest number of associated loci were observed for plant height and yield-related traits. Virtually almost all traits exhibited some enrichment ($5-45\%$) of predicted deleterious SNPs among significant heterozygous loci associated with hybrid vigor, but only for whole plant yield and days to tasseling was the observed enrichment statistically significant.

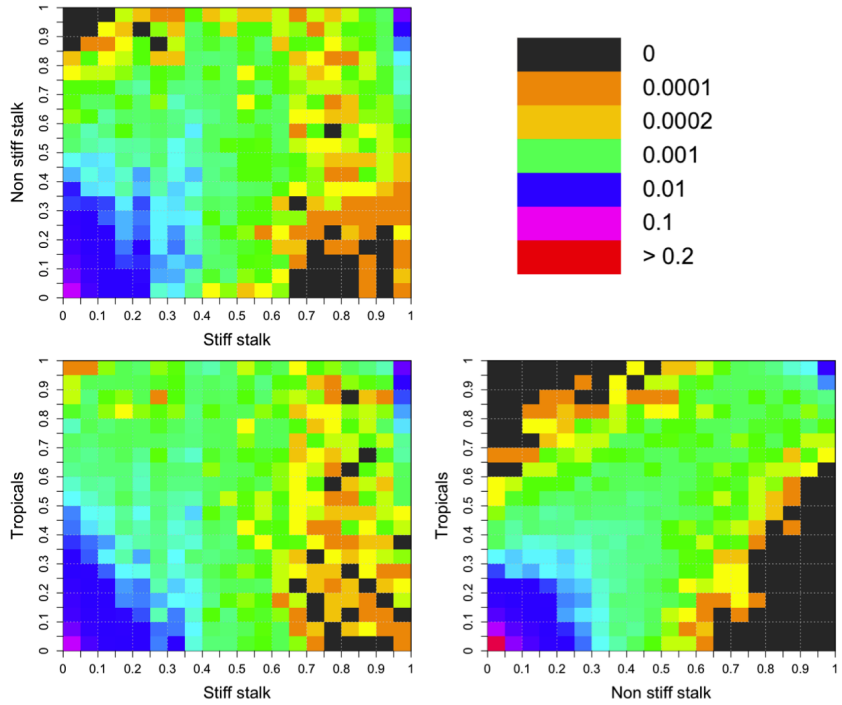


Figure 5: Joint site frequency spectrum of stiff-stalk, non stiff-stalk and tropical inbred lines

Individual significant SNPs explained 4 to 40% of heterosis observed with a given traits which is, however, over-estimated due to Beavis effect (**?**).

Because many deleterious SNPs are at frequencies too low for inclusion in association analysis, we expanded our test of enrichment to the gene level. We asked 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. A number of traits show statistically significant enrichment in population A, but nearly all traits in both populations show a positive enrichment for genes with predicted deleterious SNPs (Tables 2 and Supplemental Table 4), a result that is highly unlikely by chance (sign test

p-value=3 10^{-5} for population A and 0.01 for population B).
This observation may be due to so-called synthetic associations between rare deleterious loci and a common locus at high enough frequency to be included in the association mapping analyses (**??**). Recent work suggests that this sort of association is only likely to hold for deleterious SNPs with relatively small effect on phenotype (**?**). An observation in agreement, under an additive model, with the expected weak to intermediate effects of the deleterious loci involved in heterosis, while deleterious mutations of large effect are likely to be purged from the population due to their effects on inbreeding depression (**???**).

Although only 124,129 protein coding SNPs were covered by our analyses, which is by no means a full accounting of the genic SNPs in the maize genome (see **?** for comparison), this first scan for deleterious mutations in maize brings evidence for the contribution of complementation at deleterious mutations to heterosis. The dominance hypothesis is one of the several hypotheses (reviewed by **?**) proposed for explaining heterosis. Using conservation based approaches to predicted deleterious mutations, we showed here the involvement of complementation of predicted deleterious mutations in heterosis. Including, in future analyses, a higher marker density with estimates of deleterious SNPs within both genie and non genie regions, may increase the power to detect deleterious SNPs and enrichments in association mapping results bringing an important information for maize breeding and heterosis understanding.

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, and G. Coop, J. Gerke, P. Morrell, and P. Ralph for helpful comments. This project was supported by Agriculture and Food Research Initiative Competitive Grant 2009-01864 from

the USDAs National Institute of Food and Agriculture as well as a grant from DuPont Pioneer.

Tables

Table 1: Total number of significant SNPs (n) and fold enrichment (f) in genie regions, for loci with deleterious mutations in population A. Numbers marked with * are statistically significant.

	Inbreds		BPH		MPH	
Traits	n	f	n	f	n	f
Traits	S*	f	S*	f	S*	f
DTT	475	1.05	3372	1.15*	1123	1.12
TSLLEN	458	0.81	297	1.21	365	1.16
TSLBCHCNT	300	0.98	4077	0.98	1257	1.12
TSLANG	244	1.11	490	0.93	646	1.18
PLTHT	282	0.92	18068	0.98	9712	0.93
UPLFANG	415	1.20	8927	0.99	2266	1.12
LFWDT	289	1.21	1064	1.16	1051	1.01
LFLEN	389	1.14	4256	0.93	2257	1.07
KNLHGT	292	1.10	8752	1.08	4512	1.01
RPR	258	0.79	359	1.30	375	0.93
PLTYLD	257	1.50	7440	1.12*	7007	1.14*
EARLGH	231	0.89	605	1.11*	907	1.00
10KWT	298	1.29	709	1.15	761	1.30
COBDIA	219	1.04	4363	1.16*	405	0.88
COBWT	228	1.09	1746	0.93	519	0.69
TOTKNLWT	256	0.88	3781	0.98	2045	0.95

Table 2: Total number of genes with significant SNPs (n) and fold enrichment for genes with predicted deleterious SNPs(f) in population A

	Inbreds		BPH		MPH	
Traits	n	f	n	f	n	f
DTT	176	1.11	1137	1.12*	429	1.15*
TSLLEN	173	1.08	128	1.14	154	1.20
TSLBCHCNT	114	1.02	1257	1.13*	472	1.14*
TSLANG	103	1.03	177	1.10	254	1.15
PLTHT	128	1.22	4529	1.10*	2741	1.10*
UPLFANG	166	1.13	2553	1.11*	810	1.15*
LFWDT	112	1.27	379	1.05	375	1.14
LFLEN	141	1.18	1290	1.13*	821	1.20*
KNLHGT	123	1.09	2633	1.13*	1506	1.14
RPR	99	1.24	150	1.15	145	1.07
PLTYLD	117	1.22	2440	1.14*	2302	1.14*
EARLGH	84	1.02	230	1.20	333	1.15
10KWT	137	1.18	288	1.17	308	1.13
COBDIA	90	1.10	1419	1.13*	162	1.12
COBWT	99	1.19	548	1.07	176	1.13
TOTKNLWT	101	1.18	1228	1.11*	714	1.07

Supplementals

List of the inbred lines used

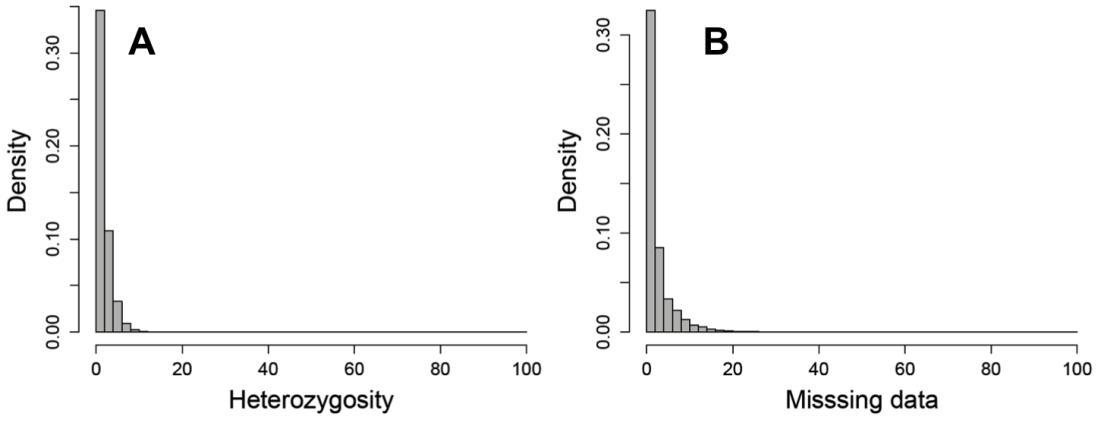
PopulationA

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,

X33.16, X38.11, X4226, X4722

PopulationB

B73, MO17, X33.16, 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, Thellunigiella halophila, Vitis vinifera, Volvox carteri.

Sup. Table 1: 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 2: Comparion of the results of MAPP prediction 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)	-

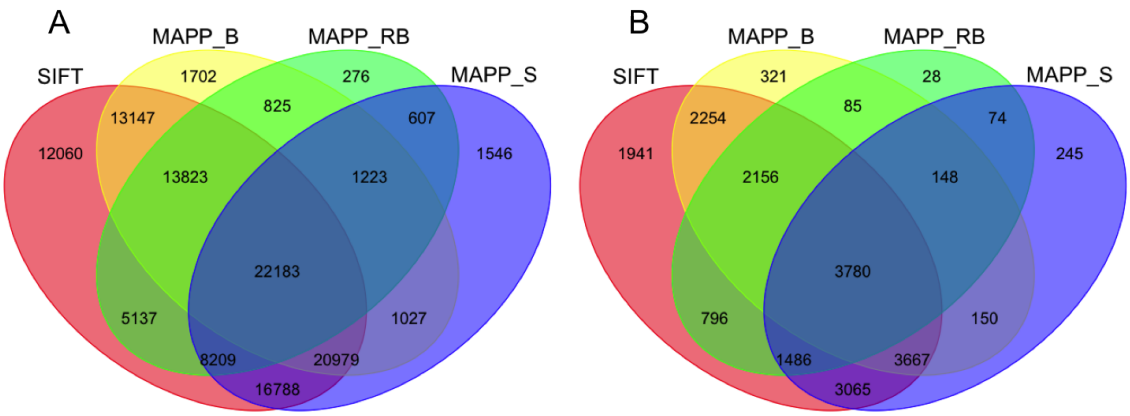
The lower triangle indicates the number of amino acid positions predicted with two given gene sets and covered by GBS SNPs (number of genes between brackets); the upper triangle indicates the percentage of amino acids with the same predictions.

Sup. Table 3: Total number of significant SNPs (n) and fold enrichment (f), in genic regions, for loci with deleterious mutations in population B. Numbers marked with * are statistically significant.

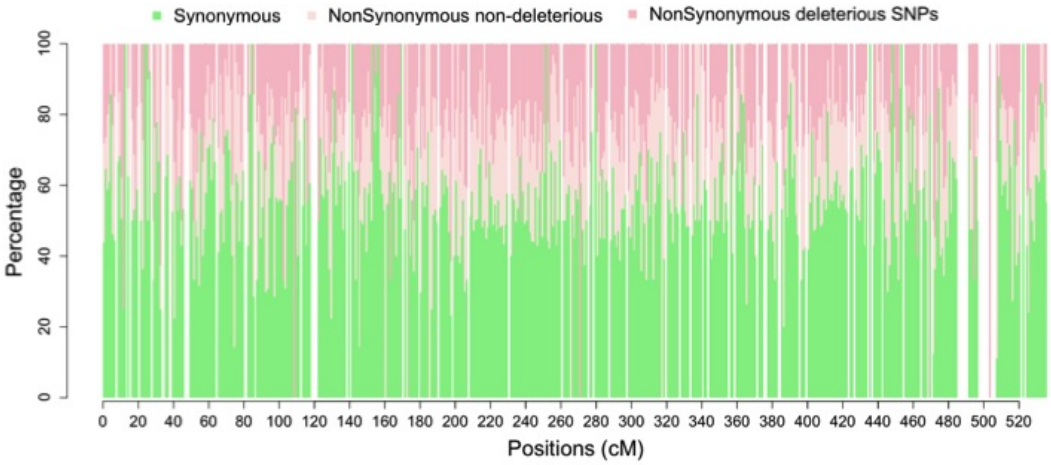
	Inbreds		BPH_B73		MPH_B73		BPH_Mo17		MPH_Mo17	
Traits	*n*	*f*	*n*	*f*	*n*	*f*	*n*	*f*	*n*	*f*
10KWT	310	0.77	404	1.17*	257	0.86	698	0.83	723	0.98
COBWT	313	0.62	941	1.15*	387	0.69	257	1.33	532	0.95
COBDIA	226	1.49	159	1.25*	236	1.06*	349	0.78	615	0.72
COBLEN	598	1.08	239	1.20*	97	0.24	280	1.08	140	0.92
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 4: Total number of genes with significant SNPs (n) and fold enrichment for genes with predicted deleterious SNPs (f) in population B

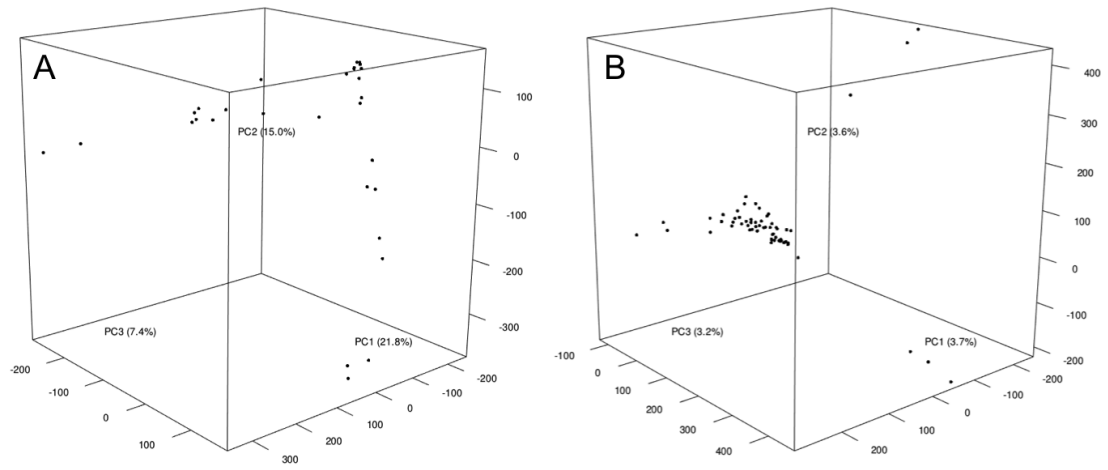
	Inbreds		BPH_B73		MPH_B73		BPH_Mo17		MPH_Mo17	
Traits	n	f	n	f	n	f	n	Enri.	n	f
10KWT	73	1.17	169	1.14	95	1.11	246	1.11	274	1.11
COBWT	71	1.13	316	1.08	128	1.04	94	1.10	204	1.10
COBDIA	81	1.07	57	1.08	86	1.11	134	1.03	234	1.14
COBLEN	203	1.09	89	1.24	30	1.17	110	1.17	51	1.21
SEEDWT	138	1.10	146	1.14	50	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-deleterious nonsynonymous and deleterious nonsynonymous 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