A Genome-Wide Scan for Evidence of Selection in a Maize Population Under Long-Term Artificial Selection for Ear Number

Timothy M. Beissinger,^{*,†} Candice N. Hirsch,[†] Brieanne Vaillancourt,^{§,**} Shweta Deshpande,^{††}

Kerrie Barry,^{††} C. Robin Buell,^{§,**} Shawn M. Kaeppler,^{*,**} Daniel Gianola,^{†,§§} and Natalia de Leon^{*,**}

^{*}Department of Agronomy, [†]Animal Sciences Department, ^{††}Department of Energy, Great Lakes Bioenergy Research Center, and ^{§§}Department of Dairy Science, and University of Wisconsin, Madison, Wisconsin 53706, [‡]Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul, Minnesota 55108, [§]Department of Plant Biology, and ^{**}Department of Energy, Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan 48824, ^{††}Department of Energy, Joint Genome Institute, Walnut Creek, California 94598

ABSTRACT A genome-wide scan to detect evidence of selection was conducted in the Golden Glow maize long-term selection population. The population had been subjected to selection for increased number of ears per plant for 30 generations, with an empirically estimated effective population size ranging from 384 to 667 individuals and an increase of more than threefold in the number of ears per plant. Allele frequencies at >1.2 million single-nucleotide polymorphism loci were estimated from pooled whole-genome resequencing data, and F_{ST} values across sliding windows were employed to assess divergence between the population preselection and the population postselection. Twenty-eight highly divergent regions were identified, with half of these regions providing gene-level resolution on potentially selected variants. Approximately 93% of the divergent regions do not demonstrate a significant decrease in heterozygosity, which suggests that they are not approaching fixation. Also, most regions display a pattern consistent with a soft-sweep model as opposed to a hard-sweep model, suggesting that selection mostly operated on standing genetic variation. For at least 25% of the regions, results suggest that selection operated on variants located outside of currently annotated coding regions. These results provide insights into the underlying genetic effects of long-term artificial selection and identification of putative genetic elements underlying number of ears per plant in maize.

CHANGES in allele frequency occur in populations undergoing selection (e.g., Wright 1931; Crow and Kimura 1970). Understanding the patterns of such changes can provide a wealth of information regarding the genetic factors that control traits under selection. By comparing the allelic composition of a population pre- and postselection, the genetic control of a trait may be revealed through the discovery of altered allele frequencies, assuming that selection

effects can be statistically separated from random genetic drift (e.g., Krimbas and Tsakas 1971; Parts *et al.* 2011). Additionally, an improved understanding of the processes that take place during selection will contribute to answering long-standing genetic questions. For instance, the relative levels of diversity around selected sites may demonstrate whether selection has operated primarily on long-standing variation or on relatively new mutations (Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski *et al.* 2005). Work in this area has demonstrated that soft sweeps, or selection on standing variation, may often be found in cases of polygenic traits (Strasburg *et al.* 2012) and are expected to be common in human populations (reviewed by Pritchard *et al.* 2010). Another persistent question involves how rapidly selected sites approach fixation (Kimura 1962), which can be addressed by an analysis of selected sites once they are identified. Kelly *et al.* (2013) recently demonstrated a selection experiment in *Mimulus* for which

Copyright © 2014 by the Genetics Society of America
doi: 10.1534/genetics.113.160655

Manuscript received August 19, 2013; accepted for publication December 29, 2013; published Early Online December 30, 2013.

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160655/-DC1>.

Sequence data from this article have been deposited in the National Center for Biotechnology Information Sequence Read Archive under BioProject accession no. PRJNA194561.

[†]Corresponding author: Department of Agronomy, University of Wisconsin, 1575 Linden Dr., Moore Hall Room 459, Madison, WI 53706.

E-mail: ndeleongatti@wisc.edu

numerous partial sweeps, or selective sweeps that have not reached fixation, were found. The relative importance of nongenic DNA is another long-standing question that may be addressed by the identification of selected regions (e.g., King and Wilson 1975; Wray 2007).

Previously, assessing allele frequencies in selected populations has been feasible only for an *a priori* set of candidates or for a limited set of random loci, due to the limited number of markers available and the cost of conducting the assays. However, the recent reduction in the cost of DNA sequencing and single-nucleotide polymorphism (SNP) detection (reviewed by Metzker 2010) now allows genome-wide characterization of allelic variants. Accordingly, multiple experiments utilizing high-density SNP or sequence data to identify selected sites in both naturally and artificially selected populations and in both sexual and asexual species have been conducted (e.g., Akey *et al.* 2002; Parts *et al.* 2011; Bigham *et al.* 2010; Turner *et al.* 2011). The goals of these experiments ranged from localizing selected sites for unknown traits in natural populations (Voight *et al.* 2006) to identifying quantitative trait loci for specific traits in experimentally derived populations or crosses (Parts *et al.* 2011).

The methods employed to identify selected sites in natural populations include assessment of variation between vs. within populations (Lewontin and Krakauer 1973; Akey *et al.* 2002), detection of abnormalities in the site frequency spectrum (SFS) (Payseur *et al.* 2002), and assessment of local patterns of linkage disequilibrium (LD) (Sabeti *et al.* 2002, 2007; Voight *et al.* 2006) to find recent selective sweeps (Maynard Smith and Haigh 1974). In natural populations, it is often difficult or impossible to evaluate the phenotypic effects of selected polymorphisms because many traits are simultaneously selected in such populations and the relative intensity of selection is unknown. In experimental populations, however, selection is often deliberately conducted under controlled conditions, allowing for better inference of the strength of selection and biological role of genes localized within potentially selected sites. Methods for identifying selection in these artificially selected populations may include any of the methods utilized for natural populations, but a benefit of these types of studies is that samples of the progenitor population are frequently available, which allows for direct measurement of allele frequency changes. Separation of selection vs. genetic drift effects has been performed by comparing allele frequencies to simulations of drift (Wisser *et al.* 2008) and by developing significance tests based on replicated or control populations (Parts *et al.* 2011; Turner *et al.* 2011).

Long-term breeding projects in agricultural species, both plants and animals, have generated excellent resources that can be leveraged for identifying loci that were affected by artificial selection. In animals, for instance, Johansson *et al.* (2010) worked with a population of chickens divergently selected for body size and found that the majority of changes can be attributed to selection on standing genetic variation

vs. new mutations. Another study using chickens identified 82 putatively selected regions with reduced levels of heterozygosity (Qanbari *et al.* 2012). Similarly in cattle, Flori *et al.* (2009) found 13 regions that were under selection in recent history, a subset of which included genes previously known to affect milk production. Also, Pan *et al.* (2013) identified selected regions in cattle based on LD and then verified the functional roles of several genes based on a review of genome annotation, gene ontology enrichment analysis, and pathway enrichment analysis. Another interesting study in cattle was conducted by Qanbari *et al.* (2011), which employed a multifaceted approach including both allele frequency- and LD-based methods to identify signatures of selection.

Several studies scanning for selection in agricultural crop species have also been conducted. For instance, Wright *et al.* (2005) looked for evidence of selection across a set of 774 maize genes and found that 2–4% had undergone selection. Recently, whole-genome studies have been conducted as well; both Jiao *et al.* (2012) and Hufford *et al.* (2012) looked for signatures of selection by investigating diverse sets of maize lines and highly dense marker sets. These studies have also been conducted with other important crops, including soybeans (Lam *et al.* 2010) and rice (He *et al.* 2011). Often, plant species have the advantage that remnant seeds representing a population before selection began often remain available for years or decades following the selection process itself (e.g., Odhiambo and Compton 1987). This characteristic was utilized by Wisser *et al.* (2008), who compared marker data gathered from samples before and after several generations of selection to identify loci affecting northern leaf blight resistance in closed populations of maize that had undergone selection.

Maize is an important crop species that has been subjected to artificial selection for ~9000 years (Matsuoka *et al.* 2002). Modern research and breeding investments have provided numerous examples of existing maize populations that have been selected for a particular trait over time spans ranging from only a few cycles to >100 generations (e.g., Odhiambo and Compton 1987; Coors and Mardones 1989; Ross *et al.* 2006; Dudley 2007; Wisser *et al.* 2008). One such example involves the Golden Glow maize selection project (Coors and Mardones 1989), which has undergone selection for a specific yield component, prolificacy, defined as the number of ears per plant. Selection for an increase in number of ears per plant was accomplished using recurrent mass selection for 30 generations, maintaining a large effective population size (N_e) and strong selection intensity in the process. Selection succeeded in increasing the mean number of ears per plant from 1.6 at cycle 0 to 4.9 by cycle 24 (de Leon and Coors 2002). Number of ears per plant is a trait of particular interest to maize breeders because it is highly correlated with grain yield and density tolerance (Russell 1984; Carlone and Russell 1987; Subandi 1990; Duvick 1997; Ahmad *et al.* 2011). In fact, Coors and Mardones (1989) reported a correlation between

ears per plant and grain yield per plant of 0.90 through cycle 12 of the Golden Glow population. Maita and Coors (1996) still found the correlation to be positive after 20 cycles of selection ($r = 0.71$) and reported that increased number of ears per plant may improve the population's ability to yield in stress conditions. Additionally, number of ears per plant is of interest as a model trait because it is correlated with other important agronomic traits, including lodging and moisture at harvest (Cross *et al.* 1987), and has been shown to be a secondary effect of maize domestication (Doebley *et al.* 1990). Overall, the combination of large N_e , strong selection intensity, substantial phenotypic response to selection, and practical and biological relevance of the trait makes Golden Glow an ideal crop model population to evaluate allele frequency changes resulting from selection.

The objectives of this study were to (1) estimate SNP allele frequencies in the cycle-30 selected population relative to the initial population by pooled whole-genome resequencing to scan for signatures of selection and (2) evaluate the putatively selected regions to assess whether selected sites are approaching fixation, estimate the extent of selective sweeps and genetic hitchhiking, and explore the proportion of sites for which selection may have operated on intergenic as opposed to genic regions.

Materials and Methods

Germplasm

Selection for increased number of ears per plant in the Golden Glow maize population was initiated by J. H. Lonnquist at the University of Wisconsin in 1971. For the first 12 cycles of selection, selection intensity was maintained at ~2.5–5%. From the 13th cycle onward, the selection intensity was made stronger, to between ~0.5 and 1%. A complete description of the selection process was provided by de Leon and Coors (2002).

For the present experiment, 48 randomly chosen plants from each of cycles 0 and 30 were utilized for analysis. To preserve population seed samples over the decades, remnant seed from the original cycles was occasionally increased through random mating of individual plants, utilizing large population sizes to minimize unwanted changes in allele frequency due to drift or unintentional selection. While genetic drift was minimized during this process, it could not be completely eliminated. The sample taken from cycle 0 had incurred five generations of seed increase, utilizing on average 110 individuals each generation, while that from cycle 30 had incurred two generations of increase, utilizing on average 130 individuals each generation.

DNA extraction, SNP genotyping, and sequencing

DNA extraction for array-based SNP genotyping was performed for each individual sampled. Leaf tissue was harvested from 96 plants (48 from each population), followed by DNA extractions using the cetyl(trimethyl)ammonium bro-

mid (CTAB) method (Saghai-Maroo *et al.* 1984). Genotyping was performed on the individual samples by Pioneer Hi-Bred International (Johnston, IA), using a 768-marker multiplex assay on the Illumina (San Diego) BeadArray platform (Jones *et al.* 2009). These array-based SNPs were used only for the determination of effective population size.

For the whole-genome resequencing, an equal amount of tissue from 48 seedlings from each population cycle was harvested and pooled. From each pool, DNA was extracted using the CTAB method (Saghai-Maroo *et al.* 1984). Libraries with a target insert size of 270 bp were prepared according to the Illumina protocol. Libraries were sequenced using the Illumina HiSeq at the Department of Energy Joint Genome Institute to generate 2×100 nucleotide pair-end (PE) sequence reads. Sequences are available in the Sequence Read Archive at the National Center for Biotechnology Information (BioProject accession no. PRJNA194561). Sequence read quality was evaluated using the FastQC program (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and sequencing lanes with insufficient quality were not used in the analysis. In total, 555,078,520 read pairs from eight sequencing lanes of cycle 0 and 652,901,808 read pairs from nine sequencing lanes of cycle 30 were generated. Prior to mapping, reads from high-quality lanes were cleaned using the fastx clipper program from the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), which removed the Illumina adapter sequences, and required a minimum length of 20 bp after trimming.

Two mapping pipelines were used to establish a high-confidence SNP set. In the “SE pipeline,” all reads that passed through the cleaning step above were mapped as single-end (SE) reads, using Bowtie version 0.12.7 (Langmead *et al.* 2009), to the B73 version 2 reference sequence [AGPv2; <http://ftp.maizesequence.org> (Schnable *et al.* 2009)]. An alignment was considered valid if there were two or fewer mismatches relative to the reference sequence ($-v\ 2$) and a read was required to have only one valid alignment ($-m\ 1$). All other parameters were set to the default values. In the “mixed pipeline” read pairs for which both reads passed through the cleaning step were mapped as PE and read pairs for which only one read passed the cleaning step were mapped as SE. For the PE mapping, in addition to the $-m\ 1$ and $-v\ 2$ options used in the SE pipeline mapping, the minimum insertion size was set to 0 bp and the maximum insertion size was set to 1000 bp. All other parameters were set to the default values.

The same SNP detection pipeline was used for alignments from both the SE pipeline and the mixed pipeline. Within each population (cycle 0 and cycle 30), all valid alignments were processed using SAMtools version 0.1.12a (Li *et al.* 2009) sort, merge, index, and pileup programs to generate unfiltered pileup files. For the pileup program, the $-B$ option was used to disable BAQ computation. Nucleotide frequencies (A, T, C, and G) were determined at each position, requiring a quality score of at least 20 for a base within a read

to be included. For a particular position, if at least two nucleotides were supported by at least two reads each across the two populations, that position was considered polymorphic. Only positions that were identified as polymorphic in both the SE pipeline and the mixed pipeline were included for further analysis, and allele frequency estimates were based on the SE pipeline.

SNP filtering and estimating allele frequencies

In total, 8,128,042 SNPs were identified from sequencing, but the set of SNPs selected for analysis was filtered to include only high-confidence sites. Only SNPs with two alleles were included, due to the increased likelihood that multiallelic SNPs included sequencing errors and because of complications related to assessing allele frequency changes at multiallelic loci. It was required that every SNP location included in the analysis was observed at least 20 times in each population and no more than 89 times based on the SE pipeline mapping. An observation of 89 corresponded to the mean SNP coverage plus two standard deviations. SNPs read at more than this level are more likely to be from organellar or repetitive DNA that is inaccurately represented by a single position in the reference genome. After filtering, 1,211,745 high-confidence SNPs were retained for analysis and the allele frequencies at each SNP were calculated in each population. Allele frequencies were computed according to their maximum-likelihood estimate. Thus, the number of times a particular allele at a position was observed in the population was divided by the total number of times any allele was observed at that position.

Estimating effective population size

The effective population size over the course of the selection program was evaluated in two ways. First, because the number of breeding males and females used and selected in the experiment was known and held relatively constant across cycles, the simple relationship $N_e = 4N_mN_f/(N_m + N_f)$, where N_m and N_f are the numbers of mating males and females, respectively, was used. Next, an estimate was made using the 768 array-based SNPs, according to the relationship $N_e = 1/2(1 - \sqrt{H_t/H_0})$, where H_t and H_0 are the mean levels of heterozygosity in the t th ($t = 30$ in this case) and 0th generations, respectively. Both equations for effective population size are provided by Crow and Kimura (1970).

Scan for selection

A genome-wide scan for selection similar to the Lewontin and Krakauer (1973) test was conducted. Note that because of the simple population structure involved in this experiment, with there being only two subpopulations and no migration between them, it is not expected that the shortcomings of the Lewontin–Krakauer test, with respect to migration (Nei and Maruyama 1975) or correlations between subpopulations (Robertson 1975), will be detrimental. Unfortunately, however, the implementation of pooled sequencing precludes the ability to accurately estimate LD in

the population and therefore makes a simulation-based approach for establishing precise significance levels intractable. Instead, the scan described is used to classify genomic regions as empirically divergent or not divergent over the course of the experiment. The most divergent sites, based on a sliding-window estimate of F_{ST} , are highlighted as the most promising candidates for selection. This approach is justified based on the documented process of strong selection, coupled with the dramatic phenotypic changes that have accumulated during the experiment, which leads to the conclusion that selection is expected to have changed the allele frequency in regions of the Golden Glow population's genome associated with the selection target. As such, those sites demonstrating the greatest levels of divergence are those most likely to have been affected by selection.

Because of the substantial sampling error that is inherent to pooled sequencing, a sliding-window approach was implemented to evaluate divergence. First, SNP-specific estimates of F_{ST} were computed within R version 3.0.2 (R Core Team 2013) according to $F_{ST} = s^2/(\bar{p}(1 - \bar{p}) + s^2/r)$, where s^2 is the sample variance of allele frequency between populations, $\bar{p}$ is the mean allele frequency across populations, and r is the number of populations (Weir and Cockerham 1984). This formula assumes a large sample size, which is met by the previously described filtering step where loci observed <20 times were removed. The formula also corrects for bias based on the small number of populations sampled (in this case two). F_{ST} values were averaged over sliding windows of 25 SNPs. Thus, each SNP locus was assigned a new value based on the average of itself along with the 12 upstream and 12 downstream SNPs. A window size of 25 SNPs was chosen because such a size appeared to maximize signal from sites while minimizing noise from sampling and sequencing error. When smaller window sizes were employed, F_{ST} values behaved erratically, and employing larger window sizes led to F_{ST} values that were unrealistically homogenous. Such an approach for determining sliding-window boundaries has been previously demonstrated (e.g., Myles *et al.* 2008; Akey 2009).

Outlying window-based F_{ST} values that exceeded the 99.9% or 99.99% level of the empirical window-based F_{ST} distribution were identified. It should be noted that these outlier levels were not chosen due to a connection with a specific level of significance, but instead because they provide a reasonable number of candidates for strong or extremely strong selection, respectively, which are used in downstream analyses. Once SNPs exceeding the outlier thresholds were identified, overall boundaries for divergent regions were defined by taking the set of all SNPs that exceeded the specified threshold and identifying groups of SNPs that likely correspond to the same divergent region. This was achieved by deeming outlying SNPs that were within 5 Mb of one another as belonging to a single region. Five megabases was used because at distances greater than this, the likelihood of LD between SNPs is minimal, yet when a distance of <5 Mb was employed, it was clear that some of the resulting regions were likely correlated with the

same selection event (Supporting Information, Figure S1). Also, utilizing this relatively large window provides a conservative estimate of divergent region boundaries. It should be emphasized that with this approach the identified regions could be (and usually were) <5 Mb (Figure S2).

Each group of outlying SNPs was considered a divergent group, and the position of the 12th SNP upstream of the group start and that of the 12th SNP downstream of the group end were added to the group to define the boundaries of each divergent region. The up- and downstream additions were incorporated because the sliding-window method included information from up to the 12th SNP distal from the region in either direction.

Testing for allele fixation

For each identified region that displayed evidence of divergence (exceeded the 99.9% F_{ST} level), a test was performed to determine whether the pattern demonstrated was consistent with a drive toward fixation. To test for a loss of variability, the expected level of heterozygosity (based on the expectation from Hardy–Weinberg equilibrium) was computed for each SNP within a region at cycle 0 and at cycle 30. Next, a t -test was performed to determine whether the expected heterozygosity at cycle 30 was significantly different than the heterozygosity at cycle 0. A significant reduction in expected heterozygosity was interpreted as evidence for a tendency toward fixation at the divergent region, while no change or an increase in heterozygosity was interpreted as evidence that the region has not yet approached fixation.

Extent of hitchhiking

Genetic hitchhiking is the process by which the frequency of a neutral locus is altered due to being in LD with a locus under selection (Maynard Smith and Haigh 1974). Hitchhiking can occur in the instance of a “hard sweep,” where selection operates on a newly arisen allele that is immediately beneficial, or from a “soft sweep,” in which an allele previously segregating in the population becomes advantageous due to new selective pressures (Hermisson and Pennings 2005). It has been shown that the genomic footprint of a selective sweep is expected to extend substantially farther from the selected locus under a hard sweep model than in a soft sweep (Innan and Kim 2004; Hermisson and Pennings 2005). To investigate whether the divergent, and potentially selected, regions identified by F_{ST} were consistent with soft or hard sweeps, the size of selected regions was used as an indicator of hitchhiking. A K -means clustering algorithm (Hartigan and Wong 1979) was performed to group divergent regions based on size. Two centers were employed, so that identified clusters were classified into one group of small size and one group of large size regions, corresponding to those likely to depict hard sweeps and soft sweeps, respectively.

Also, data from the intermated B73 $\times$ Mo17 (IBM) population (Lee *et al.* 2002) were used to test whether the recombination rate is the main contributor to region size

rather than the type of sweep experienced. Liu *et al.* (2009) estimated a map of centimorgans per megabase in the IBM population. Although the IBM population will not necessarily display identical recombination patterns to those of the Golden Glow maize population at all positions in the genome, the overall patterns are expected to be similar. The physical position of each divergent region in the Golden Glow was anchored to the nearest physical position with a given centimorgan per megabase in the Liu *et al.* (2009) map. In cases for which multiple IBM positions with reported centimorgans per megabase were within a single Golden Glow selected region, the level of centimorgans per megabase across all of these positions was averaged. Thereby, every highly divergent region identified in the Golden Glow maize population was assigned a single value for centimorgans per megabase. A significant product–moment correlation was tested for using raw data as well as log-transformed data.

Results

Effective population size

A total of ~ 4250 plants for cycles 1–12 and 14,250 plants for cycles 13–30 were evaluated in the selection plots, but ~ 1000 males and 200 females were selected in each cycle, leading to an N_e that is smaller than the total census size. Assuming plants were randomly mating over the course of the experiment, it was estimated based on population demographics that the effective population size was expected to be ~ 667 . However, preferential pollen flow among neighboring plants and assortative mating among plants flowering on the same day may have prevented truly random mating; thus the N_e was also estimated from markers. Based on 768 array-based SNP markers, the effective population size estimate was 378. It is worth noting that due to the effects of selection, the marker-based estimate is expected to be biased downward. Therefore, true N_e for the Golden Glow population over the course of the selection experiment is likely somewhere between 378 and 667 individuals.

Twenty-eight genomic regions were identified as substantially divergent

To identify the specific genomic regions most likely to have been affected by selection, an outlier-based approach that scanned for regions exceeding the 99.9% or 99.99% levels of the empirical distribution, based on 25-SNP sliding windows, was employed. Using the 25-SNP sliding-window statistic, specific genomic regions that were most likely to have been affected by selection were apparent (Figure S3). Twenty-eight regions were identified as divergent at the 99.9% outlier level. Three of these regions also exceeded the 99.99% level (Figure 1; Table 1). Regions identified at the 99.9% level were found on all 10 of the maize chromosomes. The regions ranged in size from 4251 bp to 9.2 Mb and encompassed from 0 to 73 predicted B73 5b annotated genes (<http://ftp.maizesequence.org>) (Schnable *et al.*

Figure 1 Physical location of 28 regions identified as divergent and potentially under selection for number of ears per plant based on changes in allele frequency between cycle 0 and cycle 30 of the Golden Glow maize population, using 25-bp sliding-windows estimates of F_{ST} .

2009). Assuming that there was limited unintentional selection for other traits during the course of the selection experiment, genes in these regions can be considered candidates for control of number of ears per plant in maize. Of the regions identified, 22 (79%) included ≤ 5 annotated genes, and 14 regions (50%) included 1 or 0 annotated genes. As an example, a region on maize chromosome 6 that encompasses ~ 10 kb (AGPv2 position 119,682,711–119,692,810) falls entirely within a single predicted gene, GRMZM2G368678 (Figure 2). This gene is annotated as an “androgen-induced inhibitor of proliferation” based on sequence similarity to the *Sorghum bicolor* gene Sb10g010710.1 and is expressed in shoot apical meristem and multiple other tissues (Sekhon *et al.* 2011).

Few regions show evidence of fixation

Each of the genomic regions that were identified as highly divergent was tested for a change in expected heterozygosity between cycle 0 and cycle 30. A decrease in expected heterozygosity suggests that strong selection has taken place and that allele frequencies are being driven toward fixation. Conversely, an increase in expected heterozygosity at the selected site may be observed in the case where the initial favorable allele frequency was < 0.5 and selection has taken place but was not strong enough or has not been occurring for a long enough time to move the allele frequency close to fixation. Other explanations for an increase in heterozygosity can include overdominance, complex linkage relationships between multiple selected sites, and variable selection environments. It is also possible that selection has occurred but there is no change in expected heterozygosity; this would be the case, for example, if allele frequency changed from 0.4 at cycle 0 to 0.6 at cycle 30.

It was observed that only 2 of the 28 divergent regions demonstrated a statistically significant reduction in expected heterozygosity from cycle 0 to cycle 30 (two-tailed Bonferroni-corrected P -value = $0.025/28 = 0.0009$). However, 10 regions (35.7%) displayed a significant increase in expected heterozygosity. The change in the level of expected heterozygosity across the remaining 16 regions was not significant (Table 1). Examples of divergent regions that displayed an increase, a decrease, and no change in expected heterozygosity are provided (Figure 3). These observations suggest that although selection was strong, it was not strong enough over the course of 30 generations to drive favorable alleles to fixation at the majority of sites that display evidence of strong divergence. Instead, most sites are still segregating in the population and, for a substantial subset of identified sites, the genetic variability in the population has increased as a result of selection. Kelly *et al.* (2013) demonstrated that increased heterozygosity across a region as a result of selection is expected in the situation where not only has selection led to more intermediate frequencies at the selected variant, but also the selected variant is positively associated with rare variants in the region. Additionally, it may be predicted that the two regions that did display significant reductions in heterozygosity are those that experienced the strongest selection.

Selection mostly operated on standing variation

Studies have shown that selection on rare or new alleles is most likely to cause a hard sweep and lead to long-range hitchhiking and that the hitchhiking pattern may be mostly or completely absent in the case of a soft sweep, when selection operates on an allele that was segregating in the population at the onset of selection (Innan and Kim 2004;

Table 1 Position, size, and expected heterozygosity information for each of the 28 highly divergent regions identified

Region no.	Chromosome	Start position	End position	99.99% significance	Genes contained	Heterozygosity test significance ^a	Mean change in heterozygosity
1	1	11,588,371	11,892,655	N	8	<i>0.000</i>	0.132
2	1	122,802,601	122,831,005	N	0	0.473	0.024
3	1	164,947,151	165,229,053	N	12	0.191	0.052
4	2	35,519,192	35,682,346	Y	3	<i>0.000</i>	0.173
5	2	41,731,365	41,755,299	N	2	0.595	−0.019
6	2	71,306,928	71,378,431	N	3	<i>0.000</i>	−0.253
7	2	101,062,088	101,069,759	N	0	0.071	0.076
8	2	160,786,800	160,802,631	N	2	0.608	0.026
9	3	177,548,249	177,681,538	N	2	0.026	−0.047
10	3	215,594,013	215,778,968	N	4	0.014	−0.111
11	4	66,924,240	66,935,990	N	0	<i>0.000</i>	0.196
12	4	82,825,221	82,858,997	N	0	0.006	−0.131
13	4	113,455,144	122,680,452	Y	73	<i>0.000</i>	0.080
14	4	191,396,139	191,400,390	N	1	0.298	0.051
15	5	30,083,952	30,139,317	N	1	0.868	0.005
16	6	41,490,195	45,914,266	Y	42	<i>0.000</i>	0.122
17	6	75,749,792	76,382,768	N	5	0.003	0.099
18	6	119,682,711	119,692,810	N	1	<i>0.000</i>	0.229
19	7	146,671,419	146,771,150	N	1	<i>0.000</i>	0.211
20	7	167,742,364	167,809,449	N	1	0.484	−0.034
21	8	92,876,772	94,647,137	N	26	<i>0.000</i>	0.025
22	8	118,681,864	118,767,444	N	3	0.106	0.069
23	9	26,149,935	26,181,104	N	0	0.809	−0.010
24	9	101,071,793	101,097,690	N	1	<i>0.000</i>	−0.243
25	10	7,635,223	8,719,903	N	13	<i>0.000</i>	0.056
26	10	18,846,988	19,024,881	N	1	0.931	−0.004
27	10	25,251,913	25,264,660	N	0	0.032	0.089
28	10	97,503,134	97,542,318	N	0	<i>0.000</i>	0.171

^a Values for regions that displayed significant changes in heterozygosity (P -value = 0.0009) are in italics.

Hermisson and Pennings 2005; Przeworski *et al.* 2005). More precisely, Hermisson and Pennings (2005) showed that a soft sweep from standing variation is expected to display a narrower footprint than that of a hard sweep. This is because for a new allele, LD between the favorable polymorphism and the genetic background in which it resides is likely to extend a longer range than if the allele were at an intermediate frequency and therefore was present in a variety of different haplotypes. To investigate the prevalence of hard vs. soft sweeps during this selection program, the size of divergent regions was used to indicate the extent of hitchhiking and thus the type of sweep that may have occurred; larger regions suggest longer-range hitchhiking and therefore indicate a hard sweep (or the possibility of favorable alleles at multiple loci in proximity to each other), while smaller regions suggest less hitchhiking and likely a soft sweep.

Substantial variability in the size of divergent regions was observed (Table 1), with a range of 4251 bp to 9.2 Mb. The median region size was 69.3 kb. A K -means clustering algorithm with two centers was employed to separate regions into two groups based on size (Hartigan and Wong 1979). The results were that 26 of the 28 regions were placed into a small-size cluster and only 2 regions were placed into a large-size cluster (Figure S2). The median region size for the small-size cluster was 61.2 kb, while that of the large-size

cluster was 6.8 Mb. Because of the method by which regions were identified, the possibility that the large regions are the result of multiple independently selected sites in close proximity cannot be ruled out, although this appears to be less likely due to the size of the gap assumed (Figure S1).

There is also the possibility that the size of regions is heavily influenced by the variability of recombination rates across the genome. This was tested by utilizing a recombination map developed from the IBM population (Lee *et al.* 2002). No evidence for a correlation between recombination rate and region size was found in the raw data ($\rho = -0.126$, P -value = 0.5215) or by utilizing log-transformed region sizes ($\rho = 0.172$, P -value = 0.3807; Figure S4). Therefore, it is likely that the 2 large regions correspond to a hard sweep model with a large amount of hitchhiking due to selection on rare, relatively new variants and the remaining 26 regions demonstrate selection on standing variation. Consequently, the vast majority of selection is not consistent with selection on new variants but instead on existing genomic variants that were already segregating well before cycle 0. This observation is also consistent with the large phenotypic response seen in a relatively small number of generations of selection; rare variants (e.g., $<P = 0.01$), even of substantial magnitude, would take multiple generations of selection before they began to contribute meaningful variation to the selection response.

Figure 2 F_{ST} across a single significantly selected region of chromosome 6. The selected region is included entirely within a predicted gene, GRMZM2G368678, which has been annotated as an androgen-induced inhibitor of proliferation in *Sorghum bicolor*.

Selection on genes or intergenic regions

Functional alleles can involve changes in the coding sequence, transcriptional control of genes by nearby promoter and controlling elements, and nontranslated controlling sequences. To determine the potential importance of genic vs. nongenic variants underlying phenotypic variation for number of ears per plant, each position of the 28 divergent regions was classified as nongenic or genic (containing one or more annotated gene models). Of these, 7 (25%) regions neither contain currently annotated genes nor are located within 5 kb of a 5b reference gene (Schnable *et al.* 2009). This suggests either that the population harbors selected genes that are not present or annotated in the reference sequence or that a sizeable subset of the selection in the Golden Glow population has operated on nongenic regions or a combination thereof.

Discussion

This analysis provides insight into the genetic processes that take place during long-term experimental selection as well as the genetic control of number of ears per plant in maize, based on a long-term selection program. A total of 28 highly divergent regions were identified and therefore likely to have been under selection, with representation on all 10 of the maize chromosomes. Among these, 22 contain five or fewer annotated gene models and 14 contain one or zero annotated genes. Moreover, evidence from past studies helps to corroborate the potential role of some of the identified regions. For instance, GRMZM2G368678 is annotated as an androgen-induced inhibitor of proliferation based on sequence similarity to a *S. bicolor* gene. Separately, a quantitative trait locus (QTL) study was performed by de Leon *et al.* (2005) based on a mapping population derived from the Golden Glow population at cycle 23. The study identified a QTL on chromosome 6 for ear number that closely corresponds to one of the divergent regions that was identified. Additionally, the maize gene *zcn15* was found within a divergent region on a different area of chro-

mosome 6 that contains a total of five annotated genes. Danilevskaya *et al.* (2008) report that this is among the most favorable candidates for function as a promoter of the floral transition. Additional research is needed to determine whether the genes and regions putatively subjected to selection are directly involved in meristem function resulting in increased number of ears per plant in this population.

A detailed analysis of F_{ST} values and expected heterozygosity across the selected regions demonstrated that in the majority of cases, selection did not drive variants toward fixation. Such a result appears to coincide with findings from other studies involving numerous different species. In *Drosophila*, for example, studies based on reverse evolution (Teotónio *et al.*, 2009) as well as long-term evolution (Burke *et al.* 2010) have shown little or no evidence of fixation or substantial changes in diversity. Also, Parts *et al.* (2011) found minimal evidence of fixation in yeast populations that had been selected for heat tolerance for 288 generations. Interestingly, for 10 of the identified Golden Glow regions, heterozygosity significantly increased as a result of selection, compared to only 2 where it decreased. Although this observation is consistent with certain models of selection (Crow and Kimura 1970), it is often forgotten as a potential consequence of selection. Specifically, increased regional heterozygosity after selection is expected if the selected variant has been driven to an intermediate frequency and is positively associated with rare alleles at neighboring loci, which have also, therefore, been driven to more intermediate frequencies (Kelly *et al.* 2013). One possibility for a positive association between rare alleles is that occasional historical outcrossing has introduced haplotypes consisting of an abundance of rare alleles into the population.

The finding of increased heterozygosity resulting from selection may be important when choosing appropriate methods to use to scan for selection. For instance, methods that scan for selective sweeps by looking for a loss of variability (e.g., Kim and Stephan 2002) would have no power to detect selection from such a signature. Therefore, it is important to take into consideration that selection is an ongoing process and that, even in a simple fully additive

Figure 3 Expected heterozygosity and F_{ST} for three example selected regions of the Golden Glow maize population. Expected heterozygosity was calculated using individual-SNP values of $2p(1-p)$, where p is the allele frequency of the minor allele. For the purpose of plotting, values were averaged over 25-SNP sliding windows. (A) Expected heterozygosity and F_{ST} over a region that demonstrates a loss in variability between cycle 0 and cycle 30. (B) Expected heterozygosity and F_{ST} over a region that demonstrates a gain in variability between cycle 0 and cycle 30. (C) Expected heterozygosity and F_{ST} over a region that demonstrates an insignificant change in variability between cycle 0 and cycle 30, even while F_{ST} increased, demonstrating that allele frequencies were changing.

model, selected loci for which the favorable allele has initial frequency of <0.5 will show an increase in heterozygosity before it begins to decrease as the allele moves closer to fixation and that the same may be observed at neighboring sites depending on the initial haplotype structure of the population. Another possibility is that overdominant gene action is present for several selected sites, driving alleles to equilibrium at an intermediate frequency instead of to fixation. If such is the case, it would provide evidence in favor of the overdominant theory to explain heterosis (reviewed by Schnable and Springer 2013). More likely, however, is that this result is simply a function of 30 generations being too short a time for a substantial loss in heterozygosity for all but the most strongly selected sites.

It is notable that regions with a substantial amount of long-range hitchhiking, demonstrating hard sweeps, were

rare in this experiment. Selection on relatively new mutations or rare alleles, which are in high LD with the genetic background in which they reside, is the situation that leads to long-range hitchhiking. Conversely, selection on relatively common alleles, the model of a soft sweep, is not expected to display a substantial pattern of hitchhiking (Hermisson and Pennings 2005; Przeworski *et al.* 2005). Two of the 28 divergent regions identified in this experiment were clustered separately from the remaining 26 regions, suggesting that they may be cases of long-range hitchhiking and thus hard sweeps. This implies that the majority of selection operated on standing variation for which beneficial alleles were segregating in the Golden Glow population before selection began, suggesting that although most of the polymorphisms capable of generating a high number of ears per plant were present at cycle 0, it was not until selection

incrementally increased the frequency of these variants within individuals that highly prolific phenotypes emerged.

This observation is consistent with one made by Coop *et al.* (2009), regarding human populations. The authors found that while positive selection in the human genome may be common, such selection driving new mutations to fixation is exceedingly rare. Similarly, Innan and Kim (2004) investigated selective sweeps on standing variation, as may have occurred during a domestication event. The authors focused particularly on maize. Their finding was that selection on standing variation may not be identifiable, because genetic variation at linked loci surrounding the selected site will not necessarily be reduced. This finding agrees with one reported by Teshima *et al.* (2006), who found that for an initially neutral mutation that had drifted to frequency 0.05 when it became beneficial, allele frequencies in the selected population for loci surrounding the selected site are likely to be intermediate. The onset of selection at cycle 0 of the Golden Glow experiment parallels what occurs at the onset of domestication, where the fitness of individuals suddenly and dramatically changes due to new selective pressures; thus the patterns of variation may be similar. However, because in this study the main approach for identifying selection was allele frequency divergence between the selected and nonselected populations rather than finding regions with reduced variation, this approach is not limited by the potential lack of reduced variation. Yet upon further exploring each of the selected regions to identify those consistent with long-range hitchhiking, the findings here match expectations; the overwhelming tendency was that selection modified allele frequencies at isolated sites rather than across wide spans, suggesting that most of the observed selection operated on standing genetic variation.

Finally, several of the divergent sites (25%) contain no currently annotated genes nor are they in close proximity to any annotated gene models. While this could be due to Golden Glow genes that are not present in the reference genome, it is also possible that these are instances of selection on nongenic DNA. The possibility of expression-controlling regions leading to major phenotypic differences between organisms was discussed decades ago by authors such as King and Wilson (1975). Since then, a multitude of studies have identified such regions across a wide array of species (reviewed by Wray 2007). In maize, the expression of the *tb1* gene has been shown to be affected by intergenic sequence tens of kilobases away from the gene itself (Clark *et al.*, 2006). Similarly, the finding appears to be pervasive in domesticated animal species, where studies involving horses (Gu *et al.*, 2009), cattle (Qanbari *et al.* 2011), and chickens (Qanbari *et al.* 2012) have all identified selection in gene-poor regions. Likewise, in human populations it has been observed that at a minimum, 14% of selected regions identified across multiple studies result from selection on noncoding material (Akey 2009). In *Drosophila*, various regulatory changes that modify phenotype have been found (Sucena and Stern 2000; Prud'homme *et al.* 2006). The incomplete nature of the maize reference genome (Schnable

et al. 2009), coupled with this study's inability to precisely isolate the causative sites that were selected down at the nucleotide level, precludes firm conclusions regarding the proportion of selection that operated on nongenic material. For instance, even within selected regions that do include genes it is possible that the causative variant was not one of those genes but instead a regulatory variant. The findings here imply that at least for a subset of sites, noncoding polymorphisms are selectively relevant.

In summary, important insight into the putative control of number of ears per plant was gained by scanning for signatures of selection based on differences in allele frequency between selected and unselected cycles in a maize population subjected to artificial selection for a number of generations. Furthermore, the findings show that, at least for the Golden Glow population, soft sweeps appear to be more common than hard sweeps, the rate of allele fixation is relatively slow for regions under selection, and changes in allele frequencies in noncoding polymorphisms that have effects on the phenotype can be generated by selection.

Acknowledgments

This work was funded by the Department of Energy (DOE) Great Lakes Bioenergy Research Center (DOE BER Office of Science grant DE-FC02-07ER64494). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. DOE under contract no. DE-AC02-05CH11231. Simulations were performed using resources and the computing assistance of the University of Wisconsin, Madison (UW-Madison) Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison and the Wisconsin Alumni Research Foundation and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. DOE's Office of Science. DuPont-Pioneer provided SNP genotyping with the Illumina Golden Gate assay. T.M.B. was supported by the University of Wisconsin Graduate School and by a gift to the UW-Madison Plant Breeding and Plant Genetics program from Monsanto.

Literature Cited

- Ahmad, M., R. Ahmad, M. Ishaque, and A. U. Malik, 2011 Why do maize hybrids respond differently to variations in plant density? *Crop Environ.* 2: 52–60.
- Akey, J. M., 2009 Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res.* 19: 711–722.
- Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12: 1805–1814.
- Bigham, A., M. Bauchet, D. Pinto, Z. Mao, J. M. Akey *et al.*, 2010 Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* 6: e1001116.

- Burke, M. K., J. P. Dunham, P. Shahrenstani, K. R. Thornton, M. R. Rose *et al.*, 2010 Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* 467: 587–590.
- Carlone, M. R., and W. A. Russell, 1987 Response to plant densities and nitrogen levels for four maize cultivars from different eras of breeding. *Crop Sci.* 28: 465–470.
- Clark, R. M., T. N. Wagler, P. Quijada, and J. Doebley, 2006 A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet.* 38(5): 594–597.
- Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li *et al.*, 2009 The role of geography in human adaptation. *PLoS Genet.* 5: e1000500.
- Coors, J. G., and M. C. Mardones, 1989 Twelve cycles of mass selection for prolificacy in maize I. Direct and correlated responses. *Crop Sci.* 29: 262–266.
- Cross, H. Z., J. T. Kamen, and L. Brun, 1987 Plant density, maturity and prolificacy effects on early maize. *Can. J. Plant Sci.* 67: 35–42.
- Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetic Theory*. Harper & Row, New York.
- Danilevskaya, O. N., X. Meng, Z. Hou, E. V. Ananiev, and C. R. Simmons, 2008 A genomic and expression compendium of the expanded PEBP gene family from maize. *Plant Physiol.* 146: 250–264.
- de Leon, N., and J. G. Coors, 2002 Twenty-four cycles of mass selection for prolificacy in the Golden Glow maize population. *Crop Sci.* 42: 325–333.
- de Leon, N. J., J. G. Coors, and S. M. Kaeppler, 2005 Genetic control of prolificacy and related traits in the Golden Glow maize population II: genotypic analysis. *Crop Sci.* 45: 1370–1378.
- Doebley, J., A. Stec, J. Wendel, and M. Edwards, 1990 Genetic and morphological analysis of a maize-teosinte F2 population: implications for the origin of maize. *Proc. Natl. Acad. Sci. USA* 87: 9888–9892.
- Dudley, J. W., 2007 From means to QTL: the Illinois long-term selection experiment as a case study in quantitative genetics. *Crop Sci.* 47: S20–S31.
- Duvick, D. N., 1997 What is yield? pp. 332–335 in *Developing Drought and Low N-Tolerant Maize. Proceeding of a Symposium*, edited by G. O. Edmeades, M. Bänziger, H. R. Mickelson, and C. B. Peña-Valdivia CIMMYT, El Batán, Mexico.
- Flori, L., S. Fritz, F. Jaffrezic, M. Boussaha, I. Gut *et al.*, 2009 The genome response to artificial selection: a case study in dairy cattle. *PLoS ONE* 4: e6595.
- Gu, J., N. Orr, S. D. Park, L. M. Katz, G. Sulimova *et al.*, 2009 A genome scan for positive selection in thoroughbred horses. *PLoS ONE* 4: e5767.
- Hartigan, J. A., and M. A. Wong, 1979 A K-means clustering algorithm. *Appl. Stat.* 28: 100–108.
- Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169: 2335–2352.
- Hufford, M. B., X. Xu, J. Heerwaarden, T. Pyhajarvi, J. Chia *et al.*, 2012 Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44(7): 808–811.
- Innan, H., and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. USA* 101: 10667–10672.
- Jiao, Y., H. Zhao, L. Ren, W. Song, B. Zeng *et al.*, 2012 Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* 44(7): 812–815.
- Johansson, A. M., M. E. Pettersson, P. B. Siegel, and O. Carlborg, 2010 Genome-wide effects of long-term divergent selection. *PLoS Genet.* 6: e1001188.
- Jones, E., W. Chu, M. Ayele, J. Ho, E. Bruggeman *et al.*, 2009 Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm. *Mol. Breed.* 24: 165–176.
- Kelly, J. K., B. Koseva, and J. P. Mojica, 2013 The genomic signal of partial sweeps in *Mimulus guttatus*. *Genome Biol. Evol.* 5(8): 1457–1469.
- Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765–777.
- Kimura, M., 1962 On the probability of fixation of mutant genes in a population. *Genetics* 47: 713–719.
- King, M., and A. C. Wilson, 1975 Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116.
- Krimbas, C. B., and S. Tsakas, 1971 The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control—selection or drift? *Evolution* 25: 454–460.
- Lam, H., X. Xu, X. Liu, W. Chen, G. Yang *et al.*, 2010 Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42(12): 1053–1059.
- Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10: R25.
- Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74: 175–195.
- Lee, M., N. Sharapova, W. D. Beavis, D. Grant, M. Katt *et al.*, 2002 Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol. Biol.* 48: 453–461.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Liu, S., C. Yeh, T. Ji, K. Ying, H. Wu *et al.*, 2009 *Mu* transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.* 5(11): e1000733.
- Maita, R., and J. G. Coors, 1996 Twenty cycles of biparental mass selection for prolificacy in the open-pollinated maize population Golden Glow. *Crop Sci.* 36: 1527–1532.
- Matsuoka, Y., Y. Vigouroux, and M. M. Goodman, J. Sanchez G., E. Buckler *et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. USA* 99: 6080–6084.
- Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* 23: 23–35.
- Metzker, M. L., 2010 Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11: 31–46.
- Myles, S., K. Tang, M. Somel, R. E. Green, J. Kelso *et al.*, 2008 Identification and analysis of genomic regions with large between-population differentiation in humans. *Ann. Hum. Genet.* 72: 99–110.
- Nei, M., and T. Maruyama, 1975 Letters to the editors: Lewontin-Krakauer test for neutral genes. *Genetics* 80: 395.
- Odhambo, M. O., and W. A. Compton, 1987 Twenty cycles of divergent mass selection for seed size in corn. *Crop Sci.* 27: 1113–1116.
- Pan, D., S. Zhang, J. Jiang, Q. Zhang, and J. Liu, 2013 Genome-wide detection of selective signature in Chinese Holstein cattle. *PLoS ONE* 8: e60440.
- Parts, L., F. A. Cubillos, J. Warringer, K. Jain, F. Salinas *et al.*, 2011 Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Res.* 21: 1131–1138.
- Payseur, B. A., A. D. Cutter, and M. W. Nachman, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* 19: 1143–1153.

- Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* 20: R206–R215.
- Prud'homme, B., N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams *et al.*, 2006 Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* 440: 1050–1053.
- Przeworski, M., G. Coop, and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. *Evolution* 59: 2312–2323.
- Qanbari, S., D. Gianola, B. Hayes, F. Schenkel, S. Miller *et al.*, 2011 Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics* 12: 318.
- Qanbari, S., T. M. Strom, G. Haberer, S. Weigend, A. A. Gheyas *et al.*, 2012 A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. *PLoS ONE* 7: e49525.
- R Core Team, 2013 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at: <http://www.R-project.org/>.
- Robertson, A., 1975 Gene frequency distributions as a test of selective neutrality. *Genetics* 81: 775–785.
- Ross, A. J., A. R. Hallauer, and M. Lee, 2006 Genetic analysis of traits correlated with maize ear length. *Maydica* 51: 301–313.
- Russell, W. A., 1984 Agronomic performance of maize cultivars representing different eras of breeding. *Maydica* 24: 375–390.
- Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832–837.
- Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913–918.
- Saghai-Marouf, M. A., K. M. Soliman, R. W. Jorgensen, and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. USA* 81: 8014–8018.
- Schnable, P. S., and N. M. Springer, 2013 Progress toward understanding heterosis in crop plants. *Annu. Rev. Plant Biol.* 64: 71–88.
- Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stean, F. Wei *et al.*, 2009 The B73 maize genome: complexity, diversity and dynamics. *Science* 326: 1112–1115.
- Sekhon, R. S., H. Lin, K. L. Childs, C. N. Hansey, C. R. Buell *et al.*, 2011 Genome-wide atlas of transcription during maize development. *Plant J.* 66: 553–563.
- Strasburg, J. L., N. A. Sherman, K. M. Wright, L. C. Moyle, J. H. Willis *et al.*, 2012 What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philos. Trans. R. Soc. B* 367: 364–373.
- Subandi, 1990 Ten cycles of selection for prolificacy in a composite variety of maize. *Indones. J. Crop Sci.* 5: 1–11.
- Sucena, E., and D. L. Stern, 2000 Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc. Natl. Acad. Sci. USA* 97: 4530–4534.
- Teotónio, H., I. M. Chelo, M. Bradic, M. R. Rose, and A. D. Long, 2009 Experimental evolution reveals natural selection on standing genetic variation. *Nat. Genet.* 41(2): 251–257.
- Teshima, K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genome scans for selective sweeps? *Genome Res.* 16: 702–712.
- Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice, and A. M. Tarone, 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet.* 7: e1001336.
- Voight, B. F., S. Kudavalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. *PLoS Biol.* 4: e72.
- Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. *Evolution* 38(6): 1358–1370.
- Wisser, R. J., S. C. Murray, J. M. Kolkman, H. Ceballos, and R. J. Nelson, 2008 Selection mapping of loci for quantitative disease resistance in a diverse maize population. *Genetics* 180: 583–599.
- Wray, G. A., 2007 The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8: 206–216.
- Wright, S., 1931 Evolution in Mendelian populations. *Genetics* 16: 97–159.
- Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection on the maize genome. *Science* 308: 1310–1314.
- Ziwen, H., W. Zhai, H. Wen, T. Tang, Y. Wang *et al.*, 2011 Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genetics* 7(6): e1002100.

Communicating editor: S. I. Wright

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160655/-/DC1>

A Genome-Wide Scan for Evidence of Selection in a Maize Population Under Long-Term Artificial Selection for Ear Number

Timothy M. Beissinger, Candice N. Hirsch, Brieanne Vaillancourt, Shweta Deshpande,
Kerrie Barry, C. Robin Buell, Shawn M. Kaeppler, Daniel Gianola, and Natalia de Leon

Figure S1 An illustration of the rationale behind using a gap of up to 5 MB between divergent SNPs within a single declared region. Depicted is a region on chromosome six that was identified as potentially under selection. The Grey bar represents the genome, while the black bars show the declared selected region. The top example shows that when a 5 MB gap was permitted, a single region was identified. As the permitted gap between SNPs belonging to the region was decreased, multiple regions were identified. It is more likely that there was a single selected variant and that LD across the region resulted in high F_{ST} values for multiple SNPs than that multiple independent selected variants were in such close proximity to one another. Limitations in our ability to accurately estimate LD from pooled sequencing data prevent us from estimating precise gap width per region.

Figure S2 The physical size of each region identified as potentially under selection based on F_{ST} is plotted, along with the size-group that the region was placed in according to k-means clustering. Regions are numbered in the same order as Table 1.

Figure S3 Sliding window F_{ST} values, computed by comparing cycle 0 and cycle 30 of the Golden Glow maize population, for each SNP on all chromosomes.

Figure S4 Recombination rate in the IBM population, a set of recombinant inbred lines derived from the maize inbreds B73 and Mo17, is plotted against the physical size of potentially selected regions that were found in the Golden Glow maize population. Each point depicts a putatively selected region. A: The physical size of each region was compared to the approximate recombination rate for that area of the genome as approximated by the IBM population (correlation = -0.126, p-value = 0.5215). B: The log of the physical size of each region was compared to the approximate recombination rate for that area of the genome as approximated by the IBM population (correlation = 0.172, p-value = 0.3807).