Historical genomics of North American maize

Joost van Heerwaarden[a,b,1], Matthew B. Hufford[a], and Jeffrey Ross-Ibarra[a,c,1]

[a]Department of Plant Sciences and [c]Center for Population Biology and The Genome Center, University of California, Davis, CA 95616; and [b]Biometris, Wageningen University, Postbus 100, 6700 AC, The Netherlands

Edited* by M. T. Clegg, College of Natural and Agricultural Sciences, Irvine, CA, and approved June 25, 2012 (received for review June 1, 2012)

Since the advent of modern plant breeding in the 1930s, North American maize has undergone a dramatic adaptation to high-input agriculture. Despite the importance of genetic contributions to historical yield increases, little is known about the underlying genomic changes. Here we use high-density SNP genotyping to characterize a set of North American maize lines spanning the history of modern breeding. We provide a unique analysis of genome-wide developments in genetic diversity, ancestry, and selection. The genomic history of maize is marked by a steady increase in genetic differentiation and linkage disequilibrium, whereas allele frequencies in the total population have remained relatively constant. These changes are associated with increasing genetic separation of breeding pools and decreased diversity in the ancestry of individual lines. We confirm that modern heterotic groups are the product of ongoing divergence from a relatively homogeneous landrace population, but show that differential landrace ancestry remains evident. Using a recent association approach, we characterize signals of directional selection throughout the genome, identifying a number of candidate genes of potential agronomic relevance. However, overall we find that selection has had limited impact on genome-wide patterns of diversity and ancestry, with little evidence for individual lines contributing disproportionately to the accumulation of favorable alleles in today's elite germplasm. Our data suggest breeding progress has mainly involved selection and recombination of relatively common alleles, contributed by a representative but limited set of ancestral lines.

Society depends critically on agricultural production. In most crops, yield under high-input conditions has increased dramatically over the past hundred years, providing inexpensive food and feed that form the basis of today's industrialized economies. Sustained increases in productivity have been possible in part thanks to the continued release of new crop varieties by public and private plant breeders. Nowhere is this development more apparent than in North American maize, where constant genetic gains in yield have been documented since the early 20th century (1).

Institutional maize breeding gained traction in the 1930s, when inbred lines derived from open-pollinated Corn Belt Dent varieties became the source of the first successful double-cross hybrids (2). In the late 1950s, transition to more productive single-cross hybrids marked the inception of three so-called heterotic groups, Iowa Stiff Stalk Synthetic (SS), Non-Stiff Stalk (NSS), and Iodent (IDT) (3), which today constitute genetically distinct breeding pools providing superior hybrid performance (4). A last major change occurred in the 1980s, when breeding became increasingly privatized and reliant on high-yielding, elite commercial lines (3).

Although developments in breeding practice, pedigree, and phenotype associated with historical breeding progress are well documented (5–8), little is known about the underlying genomic changes. Knowledge of genome-wide responses to artificial selection is becoming ever more important now that genomic data increasingly form the basis for selection decisions in breeding programs (9, 10). Previous marker studies have addressed relationships between modern maize lines (11–14), and some have described historical genetic changes using a limited number of markers (15–17), but so far there has not been a genome-wide account of breeding history. The ancestral origin of modern genetic differences, such as observed among heterotic groups, and the role of artificial selection in determining the composition of the genome therefore remain largely unknown.

Here we present an in-depth analysis of genomic history in North American maize, using a dataset of ∼46,000 single nucleotide polymorphism (SNP) markers genotyped in a large number of accessions, spanning four eras of maize breeding: open-pollinated landraces (pre-1930s, hereafter era 0), early inbred lines (pre-1950, era 1), advanced public inbred lines (pre-1980, era 2) and elite commercial inbred lines (post-1985, era 3). We provide a comprehensive analysis of changes in genome-wide patterns of diversity and ancestry precipitated by almost a century of breeding. By identifying individual SNPs with evidence of directional selection, we define regions and genes of potential importance to genetic improvement. We characterize patterns of ancestry at selected sites and determine the ancestral sources of favorable alleles to shed light on the effects of artificial selection on the genomic evolution of modern maize.

Results

Historical Developments in Population Structure and Genomic Ancestry. Principal component analysis (PCA) (Fig. 1) reveals pronounced patterns of population structure caused by genetic differentiation. We find 39 significant principal components (PCs) (18), of which the first three clearly relate to historical patterns of differentiation (Fig. 1*A*). Clustering of era 0 and 1 genotypes around the origin of the first three axes (Fig. 1*A*) suggests limited population structure is present within the first two eras of maize breeding, a fact that is confirmed by the occurrence of mixed clusters in the dendrogram based on all significant PCs (Fig. S1). By contrast, era 3 lines form three distinct, perpendicular clusters (Fig. 1*A*) that separate known members of the three heterotic groups (i.e., SS, NSS, and IDT). Era 2 lines mostly group closer to the origin, with exception of three genotypes at the apices of the heterotic clusters (14) that represent three historically important breeding lines for elite germplasm (2) (Fig. 1*A*). A similar pattern is observed for PCs 4–6 (Fig. S2) where four important lines mark the separation of clusters within the SS and NSS heterotic groups. Together, these results point to an increase in genetic structure over time and to a dominant role of a limited number of lines to the ancestry of the commercial inbreds in era 3.

Differentiation within the four eras, as measured by divergence of genetic groups from inferred ancestral allele frequencies (19), indeed increases steadily with time from 0.06 in era 0 to 0.38 in era 3 (Fig. 2*A*). Small divergent groups are present within eras 0 and 1, but only the last two eras show high divergence of major groups (Fig. S3). Within era 3, high levels of differentiation of the SS and IDT groups (0.27 and 0.23, respectively, compared with 0.07 for the NSS group) indicate that differentiation among heterotic groups is an important component of current-day population structure. Differentiation among the four eras as such is modest, however, with overall allele frequency divergence ranging from 0.01 for era 0 to 0.07 for era 3. This means that allele frequencies

Author contributions: J.v.H. and J.R.-I. designed research; J.v.H., M.B.H., and J.R.-I. performed research; J.v.H. and J.R.-I. analyzed data; and J.v.H. and J.R.-I. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The genotype data reported in this paper has been made available at http://www.rilab.org/resources/historicalgenomics2012.zip.

[1]To whom correspondence may be addressed. E-mail: joost.vanheerwaarden@wur.nl or rossibarra@ucdavis.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1209275109/-/DCSupplemental.

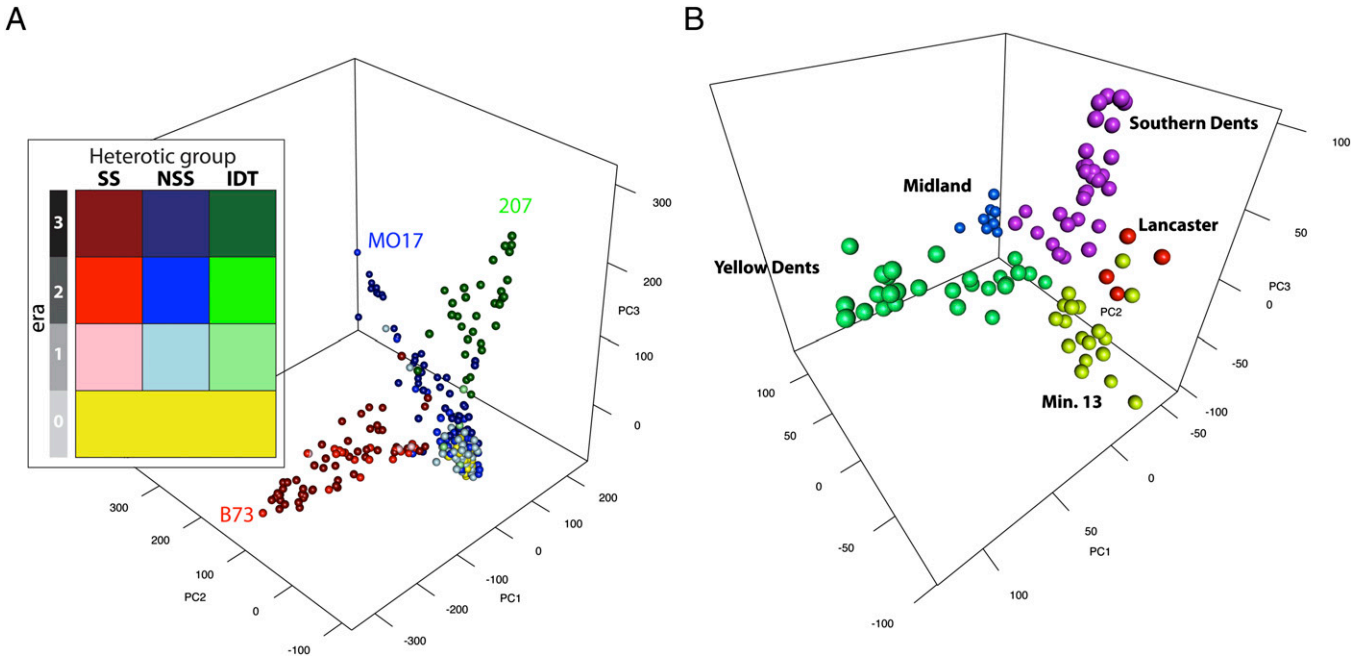

Fig. 1. (*A*) Genetic structure described by the first three genetic PCs obtained from PCA on the whole sample. Era 0 landraces are shown in yellow; remaining colors indicate heterotic groups (red, SS; green, IDT; blue, NSS) with darker tones representing later eras (e.g., light red, era 1; red, era 2; dark red, era 3). Colored labels indicate the position of the three historically important era 2 lines within each heterotic group. (*B*) First three genetic PCs obtained from PCA on Corn Belt Dent landraces; different genetic groups are shown in different colors (green, Yellow Dents; red, Lancaster; blue, Midland; yellow, Minnesota 13; purple, Southern Dents).

have remained relatively constant in North American breeding material as a whole, but have diverged strongly between breeding pools within the most recent eras.

The historical increase in within-era differentiation is associated with clear shifts in direct ancestry, which we define on the basis of haplotype sharing with lines from the same era and before (*Methods*). Direct ancestry from earlier (public) lines, although still common in eras 1 and 2, is rare in era 3, in which most ancestry is traced to other (proprietary) era 3 lines (Fig. 2*B*). Similarly, direct era 0 ancestry decreases sharply with time, being almost absent in era 3. Parallel to the increasing differentiation among heterotic groups, there is sharp decline in shared direct ancestral contributions, culminating in a virtual absence of interheterotic ancestry in era 3 (Fig. S4). In fact, ancestry of the three groups in era 3 is marked by large contributions from the same three era 2 lines identified by PCA (Fig. 1 and Fig. S2) whose cumulative proportion, estimated on the basis of ancestry from eras 0–2 (i.e., public ancestry), is as high as 0.32, 0.18, and 0.47 for SS, NSS, and IDT, respectively, confirming their historical importance.

The ancestral composition of the different heterotic groups thus seems to have narrowed over time. This is confirmed by a decrease in genetic diversity in the background of individual lines, measured by pairwise genetic differences between ancestral contributors (Fig. 2*C*) and by the effective number of contributing lines (Fig. 2*D*). In addition, linkage disequilibrium and shared haplotype length increase substantially from era 0 to era 3 (Fig. S5), suggesting that the genomes of era 3 lines consist of much larger haplotype blocks than those of earlier lines.

The recent separation of the three heterotic groups, combined with the low differentiation in era 0 confirm recent claims (4) that heterotic groups are the product of modern breeding rather than of historical divergence among era 0 landrace founders, as previously thought. However, we nonetheless find that differentiation among era 0 landraces makes a small but detectable contribution to the distinction between heterotic groups. Despite weak population structure, five well-defined era 0 variety types are distinguished by PCA-based clustering (Fig. 1*B* and Dataset S1), two of which, Yellow Dent and Lancaster, are traditionally thought to form the basis of the SS/NSS heterotic distinction (4, 20). Although a genomic assignment test of era 3 lines identifies Yellow Dent as the main ancestral contributor to all three heterotic groups, notable differences in contribution of Yellow Dent and Lancaster to SS and NSS (Fig. 2*E*) lend some justification to the traditional distinction between these two heterotic groups on the basis of their landrace ancestry.

Evidence for Directional Selection. The systematic allele frequency differences implied by the pronounced population structure in our sample presents a challenge to the detection of selected loci. We therefore implement a recent Bayesian method (21) to detect allele frequency correlations with time, while taking explicit account of genetic structure. We define the four breeding eras as different levels of an ordinal variable and test for consistent frequency changes at each individual SNP, using a genome-wide genetic covariance matrix to correct for genetic structure.

We identify 236 candidate regions with maximum Bayes factors (Bf) ranging from 21 to 5,586 and a median width of five SNPs or 93 kb (Fig. 3). Candidate SNPs generally exhibit strong directional shifts in frequency without reaching fixation (Fig. S6). Overall, only 5% of SNPs consistently have a Bf > 1 and thus show some evidence of directional selection. A total of 1,021 genes overlap with candidate regions, 715 of which have known orthologs. Notable candidate genes with clearly defined putative functions (Dataset S2) include those involved in shade and stress response, lignin biosynthesis, and auxin response and synthesis. One of the top candidates (GRMZM2G113583, Bf 548), is similar to organ-size controlling genes (ARGOS) that have been patented for increasing biomass and yield in maize (US patent 7834240); a second candidate (GRMZM2G463904, Bf 28) is orthologous to ERECTA genes patented for similar purposes (US patent 7847158). Several other genes for auxin responsive growth and stress response are also among our candidates (Dataset S2).

Limited Genomic Effects of Selection. Selection on rare favorable alleles can leave marked genomic signatures, as neighboring SNPs are swept to high frequencies (e.g., ref. 22). If such sweeps

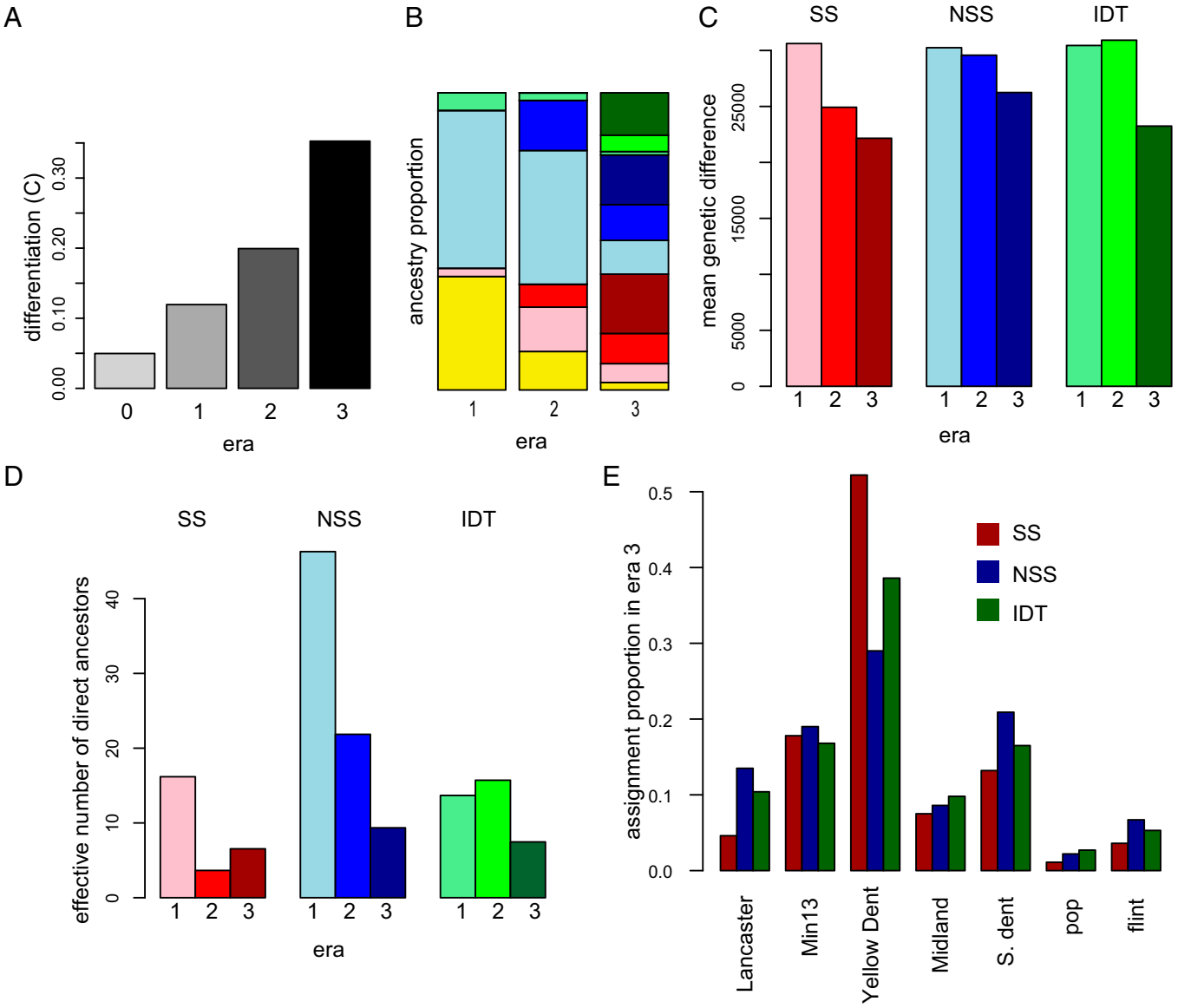

Fig. 2. Historical developments in genetic differentiation and ancestry. (*A*) Mean differentiation among genetic groups in eras 0–3. (*B*) Changes in ancestral composition from era 1–3 (colors as in Fig. 1*A*, with era 0 in yellow and tones from light to dark for eras 1–3). (*C*) Average number of differences between ancestral haplotypes within individual inbred lines. (*D*) Weighted average of the effective number of direct ancestors contributing to individual inbred lines. (*E*) Differential assignment of era 3 heterotic groups to different era 0 landraces.

are common, we expect candidate sites in era 3 to display extended shared haplotypes, reduced haplotype diversity, and local distortions of ancestry caused by the frequency increase of the selected haplotype. We define ancestral haplotypes by stretches of shared identity with specific era 0 or 1 haplotypes, which we call basal ancestry. We evaluate basal ancestry patterns by quantifying the distortion of local ancestry relative to genome-wide ancestry, as summarized by PCA, and by quantifying the diversity of ancestral haplotypes across the genome (*Methods*).

Selection seems to have had little effect on genome-wide ancestry patterns. Contrary to what is expected for selective sweeps, we do not observe extended haplotypes (Fig. S5), distorted basal ancestry, or a substantial reduction in ancestral haplotype diversity (Fig. 3) associated with selection candidates. Distortion of basal ancestry fluctuates across the genome (Fig. 3) but is not increased at candidate SNPs (Kruskal–Wallis test, $P < 0.53$). A positive correlation ($r = 0.51$) between ancestry distortion and ancestral haplotype diversity furthermore shows that distortion associates with higher rather than lower ancestral diversity and is not caused by strong frequency shifts of individual haplotypes. Reduction in haplotype diversity at selected sites is only 5% (5.9 vs. 6.2, $P = 0.006$, Kruskal–Wallis test) and in instances where low diversity does coincide with candidate SNPs (Fig. 3, black circles), putative donors of the favored alleles are lines that are common in each heterotic group, rather than rare ancestral contributors as expected under a selected sweep (Dataset S3).

Although we find no evidence for selective fixation of ancestral haplotypes at individual candidate loci, selection by breeders may still have affected the genome by favoring era 1 lines with superior multilocus genotypes. In this case, we would expect to observe era 1 lines with disproportionate ancestral contributions to favorable alleles in era 3 or, if multilocus genotype were a main determinant of a line's success, find enrichment for favorable alleles in lines that contributed most to era 3 ancestry. Neither effect is observed in our data however (Fig. 4), although some lines that are known to have been popular with breeders in the past (2) show significant enrichment for favorable alleles.

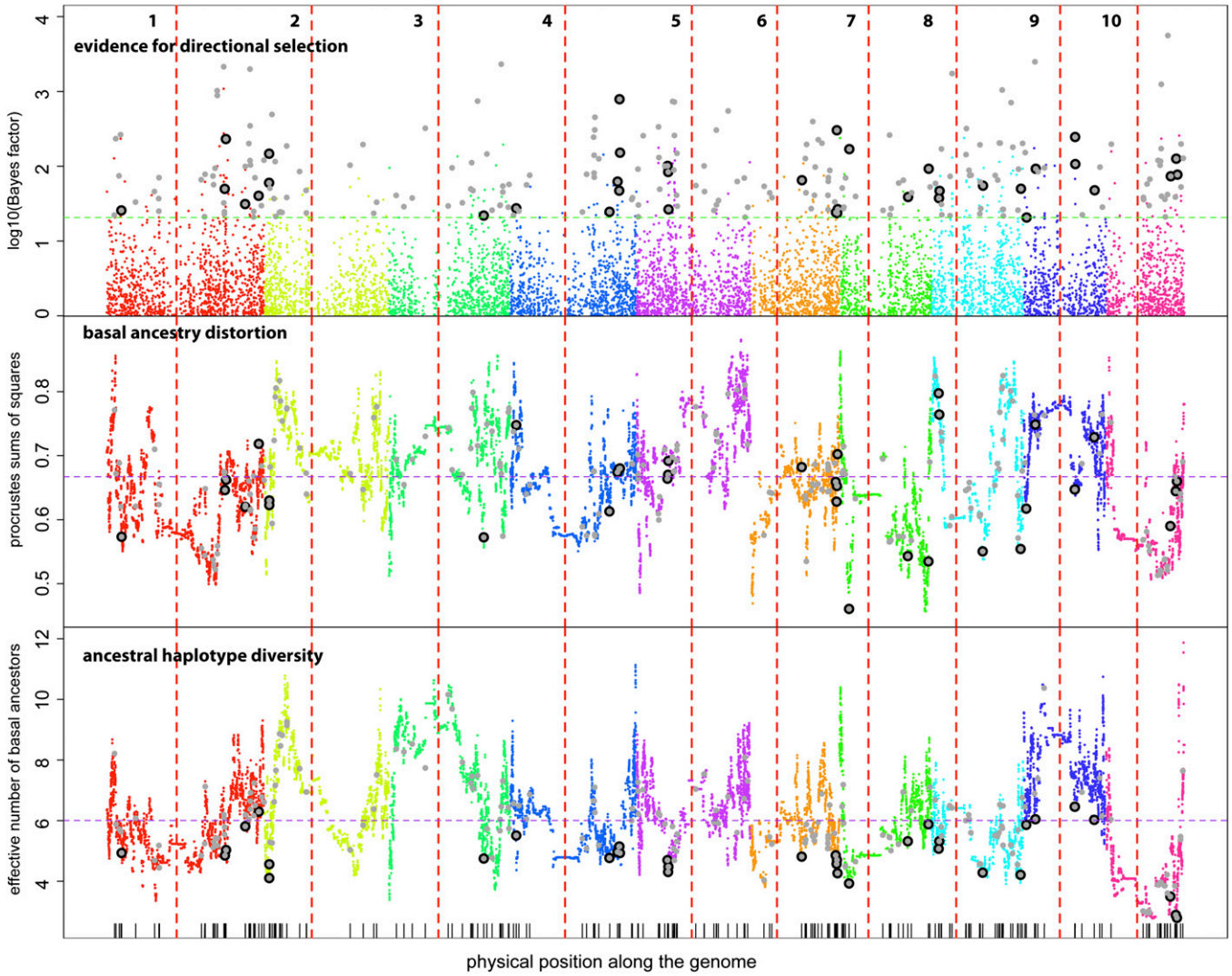

Fig. 3. Evidence for directional selection (*Top*), basal ancestry distortion (*Middle*), and ancestral haplotype diversity (*Bottom*) across the genome. Colors indicate the separate chromosomes with red vertical lines marking the centromeres. Green dashed horizontal line marks the 99th percentile of Bayes factors; purple dashed horizontal lines indicate median values of ancestry distortion and effective number of basal ancestors. Black vertical ticks mark selected features. Gray dots mark candidate SNPs. Black circles mark candidates that coincide with sites of low ancestral diversity.

Discussion

The genomics of breeding history is of great importance to understanding the genetic basis of crop improvement and is instrumental to the identification of molecular targets of artificial selection. The current state of marker technology has granted us an unprecedented look across eight decades of breeding and selection, providing insight into historical developments in diversity, ancestry, and the effects of selection across the genome.

The transition from open-pollinated varieties to inbred lines and the emergence of heterotic groups have caused profound changes in population structure, linkage disequilibrium, and ancestry patterns. Differentiation in the first two eras, although significant, is weak and our results support pedigree analyses (4) that suggest current population structure is mainly due to recent divergence of breeding pools rather than to different landrace origins. The strong differentiation observed in the modern era 3 lines is likely the result of the use of smaller numbers of more closely related breeding lines and limited genetic exchange among heterotic groups in the last two eras. Nonetheless, differential landrace ancestry remains detectable in elite material, providing some justification for the use of the traditional designations Reid (Yellow Dent) and Lancaster for the SS and NSS heterotic groups.

Compared with the dramatic shifts in ancestry, directional selection has had limited effect on the genome, with only 5% of SNPs showing some evidence of consistent selection. Candidate sites, apart from a slight reduction in ancestral diversity, do not deviate meaningfully from genome-wide patterns of haplotype length and ancestry. A potential caveat regarding this observation is that our selection scan is most sensitive to cumulative changes in allele frequency, possibly missing alleles fixed in the early stages of maize breeding. To account for this potential bias, we measured ancestry distortion and haplotype diversity at the 236 SNPs with highest frequency differentiation between eras 0 and 3, finding similar results as for our candidate SNPs (i.e., no increase in distortion and only 12% diversity reduction). Our results are also consistent with a recent resequencing study showing modest genome-wide effects of recent selection in a limited but geographically diverse sample of maize accessions (23). Nonetheless, a considerable number of candidate regions are identified across the genome, containing many genes affecting processes of agronomic relevance such as lignin synthesis (24) and response to auxin (25) and stress (1). It must also be noted that we have mapped selection associated with breeding progress per se, and that further analyses may detect selective changes specific to individual heterotic groups.

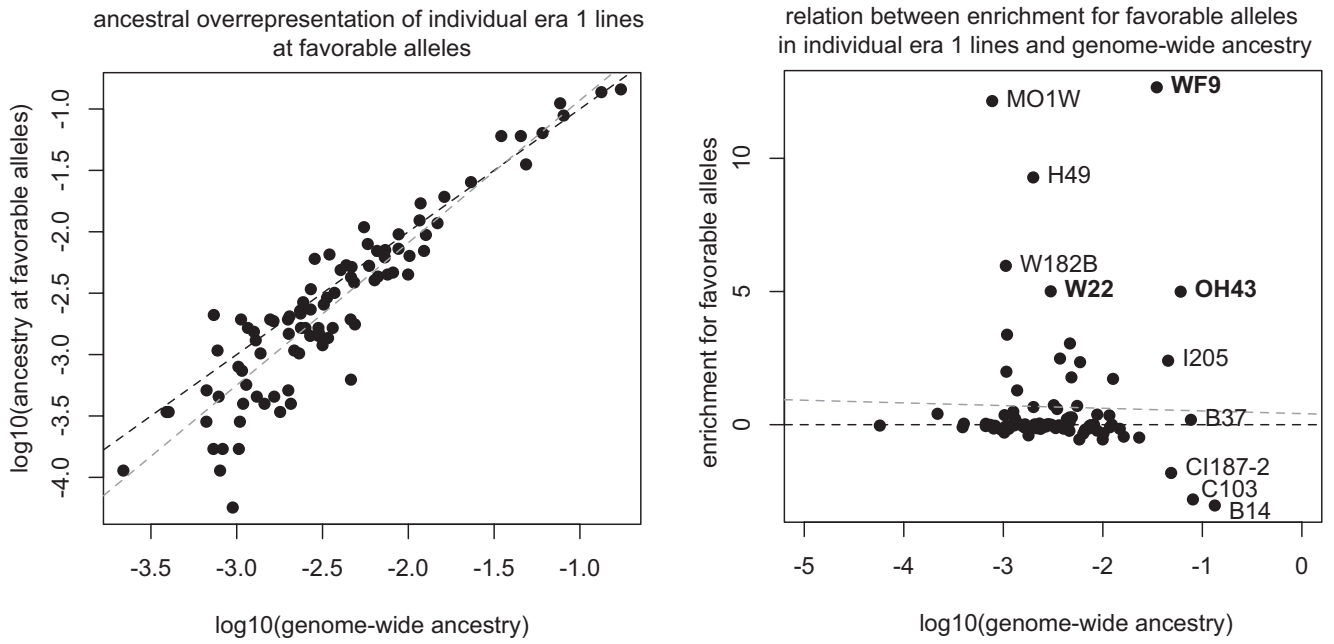

Fig. 4. Analysis of disproportionate ancestral contributions of individual era 1 lines to favorable alleles in era 3. *Left*: Overrepresentation of individual era 1 lines in the ancestry of favorable alleles, estimated by plotting the average ancestry proportion at favorable alleles against the genome-wide proportion. *Right*: Enrichment (as defined by the log probability ratio (LPR) with respect to noncandidate SNP) of favorable alleles in era 1 lines as a function of their average ancestral contribution to era 3. Black dotted lines represent the 1:1 diagonal and 0 horizontal, respectively. Gray dotted lines are regression lines (slope/r^2: 1.15/0.85 and −0.1/0.00). Line names on the *Right* are shown for lines with LPR values higher than 4 or ancestry proportion above 0.03. Labels in boldface mark breeding lines of known historic popularity.

The genomic signature of selection is informative of the genetic architecture of breeding progress. Two issues of obvious interest are the selective importance of rare alleles of large effect and the contribution of dominant ancestors with superior multilocus genotypes. The infrequent occurrence of rare ancestral contributors and absence of extended haplotypes at candidate loci favor a model of selection on common variants rather than one of strong selective sweeps (26, 27), and we find no evidence of the long-term success of specific lines being determined by their multilocus genotype. This being said, the exceptionally favorable genotypes observed for some era 1 inbreds suggests that selection of outstanding lines may have occurred, albeit with limited effect on future genomic composition.

In all, our results suggest that genetic gain achieved by plant breeding has been a complex process, involving a steady accumulation of changes at multiple loci (28), combined with heterosis due to differentiation of breeding pools (29). We thereby support the notion that selected traits of agronomic importance are predominantly quantitative in nature (30), with relatively few dominant contributions from individual alleles or lines. It will therefore be interesting to see whether our candidates prove useful in defining improved multilocus targets for genomic selection. Although challenging, the application of historical genomics to crop improvement is a tantalizing prospect that we hope breeders will soon put to the test.

Methods

Samples and Genotyping. We obtained a total of 400 accessions from US Department of Agriculture (USDA)'s National Plant Germplasm System and collaborators. Lines were chosen by a combination of literature research, consultation with plant breeders, and by querying the stock database hosted at maizegdb.org for accessions with a large number of references. Approximate ages of the selected lines were similarly obtained from the literature and germplasm databases. Accessions were divided into 99 classic North American landraces (era 0), 94 early inbreds from before the 1950s (era 1), 70 advanced public lines from the 1960s and 70s (era 2), and 137 elite commercial lines from the 1980s and 90s (era 3) that are no longer under plant variety protection (ex-PVP).

For each accession, DNA was extracted by a standard cetyltrimethyl ammonium bromide (CTAB) protocol (31) for genotyping on the Illumina MaizeSNP50 Genotyping BeadChip platform using the clustering algorithm of the GenomeStudio Genotyping Module v1.0 (Illumina). Of the total of 56,110 markers contained on the chip, 45,997 polymorphic SNPs were genotyped successfully with less than 10% missing data for use in subsequent analysis. SNPs were of diverse origins and discovery schemes. We evaluated the effects of ascertainment by comparing results for 33,575 SNPs derived from more diverse discovery panels to 12,422 SNPs that were discovered between the advanced public lines B73 and Mo17. Effects on differentiation and selection inference were found to be statistically significant but modest (*SI Text*).

Diversity, Linkage, and Ancestry Analysis. Diversity analyses followed (32, 33). Briefly, PCA was performed on normalized genotype matrices and the number of significant eigenvalues determined by comparison with a Tracy–Widom (TW) distribution (18). Genotypes were assigned to k groups by Ward clustering on the Euclidean distance calculated from the $k-1$ significant PCs. PCA-based clustering into groups was done separately for each era. To improve clustering within era 0, Corn Belt Dents were analyzed separately from Northern Flints and a divergent group containing a popcorn and a Cherokee Flower Corn (referred to here as popcorn). Genetic differentiation within each era was measured as the weighted mean of Nicholson's population-specific differentiation parameter C (19), a measure of allele frequency divergence from an estimated base population frequency, calculated for each genetic group using the popdiv function of the R (34) package popgen.

For linkage and ancestry analysis, era 0 genotypes were converted to phased haplotypes using the program fastPHASE (35). To correct for background linkage caused by genetic differentiation, linkage disequilibrium (LD) between SNPs was calculated as the squared correlation (r^2) between inverse logit-transformed residuals of a multiple logistic regression on each SNP, using the first six genetic PCs as covariates to correct for population structure. LD decay was described by nonlinear regression as in ref. 36. Mean haplotype length was calculated at 1,000 random positions across the genome and compared with the expected length obtained by randomizing SNPs within each genetic group around the same positions. Linkage disequilibrium between closely spaced SNPs was accounted for by randomizing blocks of SNPs separated by more than 4 kb.

We estimated direct genomic ancestry by shared haplotype analysis. For each line, the longest shared haplotype with lines from the same era or older

was iteratively determined until the whole genome was assigned a closest relative. Identity was assumed for sites that were heterozygous or had missing data, both of which occurred infrequently in the data. Basal and public ancestries were calculated using the same procedure, but restricting possible ancestry to groups 0 and 1 or 0, 1, and 2, respectively.

Diversity of ancestral contributions to each line was summarized by the effective number of ancestors (i.e., the inverse Simpson index) (37) as well as the number of SNP differences (i.e., Manhattan distance) among contributing lines, weighted by the number of SNPs contributed by each ancestor.

We characterized genome-wide patterns of basal ancestry distortion by the sum of squared differences between matrices of basal ancestry and genome-wide ancestry compared by Procrustes analysis. We compared the matrix of PC scores (PC 1–6) belonging to the inferred basal (era 0 or 1) ancestors at each SNP to a reference matrix of equal dimensions obtained by taking the genome-wide, per-line average of ancestral PC scores. Ancestral haplotype diversity across the genome was calculated as the weighted per-heterotic group average of the effective number of basal ancestors. To elucidate the most probable landrace population of origin of the different genomic segments, we performed a simple likelihood-based assignment test that assigned each basal ancestry segment to the most likely landrace population of origin (38) based on the observed frequencies in each population.

Selection. We performed a genome-wide scan for evidence of positive selection across the four sampled eras using a Bayesian method developed for environmental association analysis (21). By defining time as an ordinal environmental variable, this method gives a posterior probability that the frequency of a SNP correlates with time. The calculation of a Bf for selection takes explicit account of population structure by comparing allele frequency differences to those expected on the basis of a genome-wide genetic covariance matrix. To this end we defined populations on the basis of PCA, where genetic groups were split according to era. The genetic covariance matrix was estimated using 5,000 randomly selected SNPs. For each SNP, we performed five replicates of 30,000 iterations and considered SNPs within the 99th percentile of average Bf and consistently in the 95th percentile of each replicate to be potentially under selection. SNPs with Bf $>$ 1 in each replicate were considered to show evidence of selection. We tested for spatial dependence between putatively selected SNPs using spatial autocorrelation analysis of Bf as a function of genomic distance using Moran's *I* statistic. Because of autocorrelation of Bf across the genome (Fig. S5), we grouped adjacent SNPs with Bf above the median that contained at least one potentially selected site into independent candidate regions. Within each region, the SNP with the maximum Bf was considered a candidate SNP in further analyses. All genes from the high-quality filtered gene set (based on the maize reference genome release 5b.60) contained within these regions were considered selection candidates. Functional description of candidate genes was performed by searching for orthologous sequences in other species as defined at maizesequence.org.

At every SNP, we defined the modern allele as the most common allele in era 3, which for candidate SNPs was assumed to be the allele favored by selection (i.e., favorable allele). For each era 1 line, we then calculated enrichment for favorable alleles as the log probability ratio (LPR) $\log10(p_0/p_1)$, where p_1 is the probability of containing the observed number x of favorable alleles among n candidate SNPs, and p_0 is the probability of finding x modern alleles among n randomly selected control SNPs, averaged over 1,000 replicates. Probabilities were calculated using the normal approximation $x \approx N(n\overline{p}, \sum p(1-p))$, where p is the frequency of a modern allele in era 1. Control SNPs were sampled from a subset of SNPs with similar genetic differentiation among eras 1 and 3 to minimize biases in the calling of modern alleles between candidate and noncandidate SNPs. The relation between overrepresentation of favorable alleles and proportion of ancestry in era 3 lines was tested by linear regression.

ACKNOWLEDGMENTS. We thank Lauren Sagara for technical assistance with SNP genotyping and Justin Gerke and two anonymous reviewers for helpful comments on an earlier version of this manuscript. We thank Major Goodman, Jode Edwards, Mark Millard, and the US Department of Agriculture (USDA)'s North Central Regional Plant Introduction Station for providing maize accessions. This project was supported by Agriculture and Food Research Initiative Competitive Grant 2009-01864 from the USDA's National Institute of Food and Agriculture.

1. Duvick DN (2005) The contribution of breeding to yield advances in maize (Zea mays L.). *Adv Agron* 86:83–145.
2. Troyer AF (1999) Background of US hybrid corn. *Crop Sci* 39:601–626.
3. Mikel MA, Dudley JW (2006) Evolution of North American dent corn from public to proprietary germplasm. *Crop Sci* 46:1193–1205.
4. Tracy WF, Chandler MA (2006) The historical and biological basis of the concept of heterotic patterns in corn belt dent maize. *Plant Breeding: The Arnel R Hallauer International Symposium*, eds Lamkey KR, Lee M (Blackwell Publishing, Ames, IA), pp 219–233.
5. Troyer AF (2009) in *Handbook of Maize Genetics and Genomics: Volume II: Genetics and Genomics*, eds Bennetzen JL, Hake S (Springer Science + Business Media LLC, New York).
6. Smith JSC, Duvick DN, Smith OS, Cooper M, Feng LZ (2004) Changes in pedigree backgrounds of pioneer brand maize hybrids widely grown from 1930 to 1999. *Crop Sci* 44:1935–1946.
7. Smith S (2007) Pedigree background changes in US hybrid maize between 1980 and 2004. *Crop Sci* 47:1914.
8. Duvick DN, Smith JSC, Cooper M (2004) Long-Term Selection in a Commercial Hybrid Maize Breeding Program. *Plant Breed Rev* 24:109–151.
9. Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: From theory to practice. *Brief Funct Genomics* 9:166–177.
10. Morrell PL, Buckler ES, Ross-Ibarra J (2011) Crop genomics: Advances and applications. *Nat Rev Genet* 13:85–96.
11. Mumm RH, Hubert LJ, Dudley JW (1994) A classification of 148 US maize inbreds: II. Validation of cluster analysis based on RFLPs. *Crop Sci* 34:852.
12. Romero-Severson J, et al. (2001) Pedigree analysis and haplotype sharing within diverse groups of Zea mays L. inbreds. *Theor Appl Genet* 103:567–574.
13. Ho JC, Kresovich S, Lamkey KR (2005) Extent and distribution of genetic variation in US maize: Historically important lines and their open-pollinated dent and flint progenitors. *Crop Sci* 45:1891–1900.
14. Nelson PT, et al. (2008) Molecular characterization of maize inbreds with expired US plant variety protection. *Crop Sci* 48:1673–1685.
15. Messmer MM, et al. (1991) Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population: Comparison of allozyme and RFLP data. *Theor Appl Genet* 83:97–107.
16. Hagdorn S, Lamkey KR, Frisch M, Guimaraes PEO, Melchinger AE (2003) Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations. *Crop Sci* 43:474–482.
17. Feng L, Sebastian S, Smith S, Cooper M (2006) Temporal trends in SSR allele frequencies associated with long-term selection for yield of maize. *Maydica* 51:293–300.
18. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.
19. Nicholson G, et al. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J R Stat Soc Series B Stat Methodol* 64:695–715.
20. Gerdes JT, Tracy WF (1993) Pedigree diversity within the Lancaster surecrop heterotic group of maize. *Crop Sci* 33:334–337.
21. Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185:1411–1423.
22. Palaisa K, Morgante M, Tingey S, Rafalski A (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc Natl Acad Sci USA* 101:9885–9890.
23. Hufford MB, et al. (2012) Comparative population genomics of maize domestication and improvement. *Nat Genet* 44:808–811.
24. Flint-Garcia SA, Jampatong C, Darrah LL, McMullen MD (2003) Quantitative trait locus analysis of stalk strength in four maize populations. *Crop Sci* 43:13–22.
25. Yamasaki M, et al. (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* 17:2859–2872.
26. Brown PJ, et al. (2011) Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet* 7:e1002383.
27. Tian F, et al. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162.
28. Moose SP, Dudley JW, Rocheford TR (2004) Maize selection passes the century mark: A unique resource for 21st century genomics. *Trends Plant Sci* 9:358–364.
29. Troyer AF, Wellin EJ (2009) Heterosis decreasing in hybrids: Yield test inbreds. *Crop Sci* 49:1969.
30. Goodman MM (2004) Plant breeding requirements for applied molecular biology. *Crop Sci* 44:1913–1914.
31. Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc Natl Acad Sci USA* 81:8014–8018.
32. Van Heerwaarden J, et al. (2010) Fine scale genetic structure in the wild ancestor of maize (Zea mays ssp. parviglumis). *Mol Ecol* 19:1162–1173.
33. van Heerwaarden J, et al. (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc Natl Acad Sci USA* 108:1088–1092.
34. R Development Core Team (2009) R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna).
35. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78:629–644.
36. Remington DL, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 98:11479–11484.
37. Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738.
38. Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* 4:347–354.

Supporting Information

van Heerwaarden et al. 10.1073/pnas.1209275109

SI Text

SNP data are subject to bias in the allele frequency spectrum due to marker discovery in small and/or unrepresentative sets of individuals. When severe, such bias may affect inference of genetic differentiation and selection. SNPs on the Illumina genotyping array were provided by a number of contributors using a variety of ascertainment schemes. We obtained a measure of the severity of ascertainment bias by comparing results for 33,575 reference SNPs of varying origin to those for 12,422 SNPs that were known to have been exclusively ascertained as polymorphic between the legacy inbred lines B73 and Mo17.

We evaluated the effects on genetic differentiation by comparing correlations of the Euclidean distance along the first six genetic principal components (PCs) between B73/Mo17 SNPs and the reference set of SNPs. Correlation between principal component analysis (PCA) distances calculated on B73/Mo17 SNPs and random draws of 12,422 reference SNPs were 0.96 compared with 0.99 for the average correlation between two random draws from the reference SNP set. Although significant ($P < 0.01$, based on 100 random samples), the effect of ascertainment on inferred patterns of differentiation thus appears to be relatively weak.

Of our 236 candidate SNPs, 34.7% were B73/Mo17 markers. This represents a small but significant (binomial test $P = 0.01$) enrichment over the expected 27%. This overrepresentation may be due to the slightly higher (0.38 vs. 0.35, Wilcoxon two-sample test: $P < 0.0001$) expected heterozygosity for B73/Mo17 SNPs, because it is easier to detect frequency shifts in markers at intermediate frequencies than in markers close to fixation.

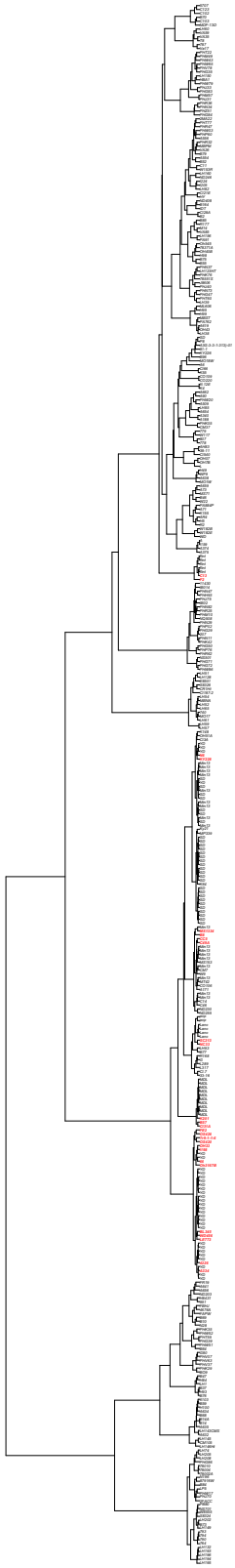

Fig. S1. Ward dendrogram based on the Euclidean distance on 39 PCs. Era 1 lines clustering with landraces are marked in red. (YD, Yellow Dents; Lanc, Lancaster; Min13, Minnesota 13; MDL, Midland; SD, Southern Dents).

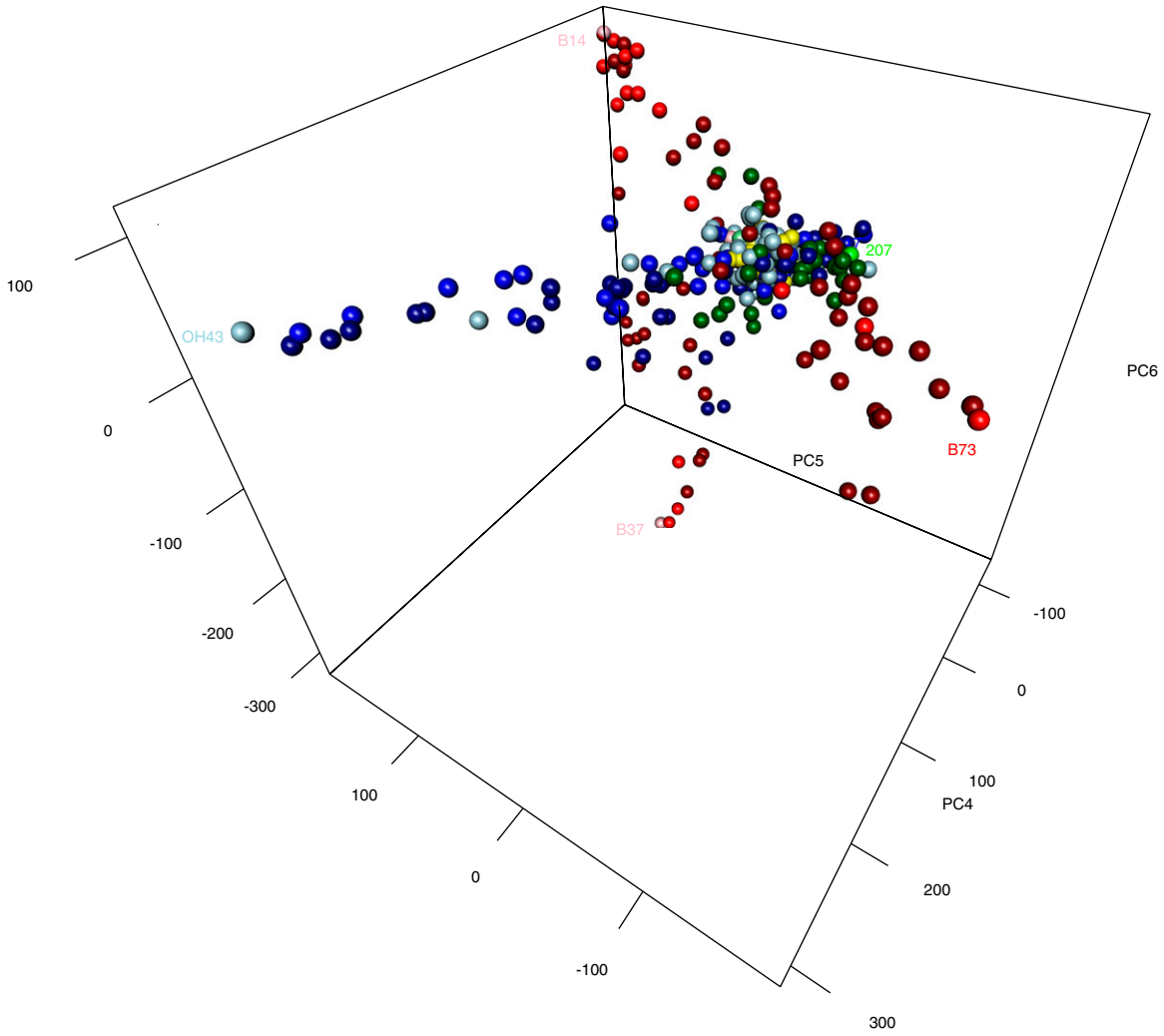

Fig. S2. PCs 3–6, obtained from PCA on all lines. Colors represent heterotic groups [red, Iowa Stiff Stalk Synthetic (SS); green, Iodent (IDT); blue, Non-Stiff Stalk (NSS)], and darker colors represent later eras (e.g., light red, era 1; red, era 2; dark red, era 3).

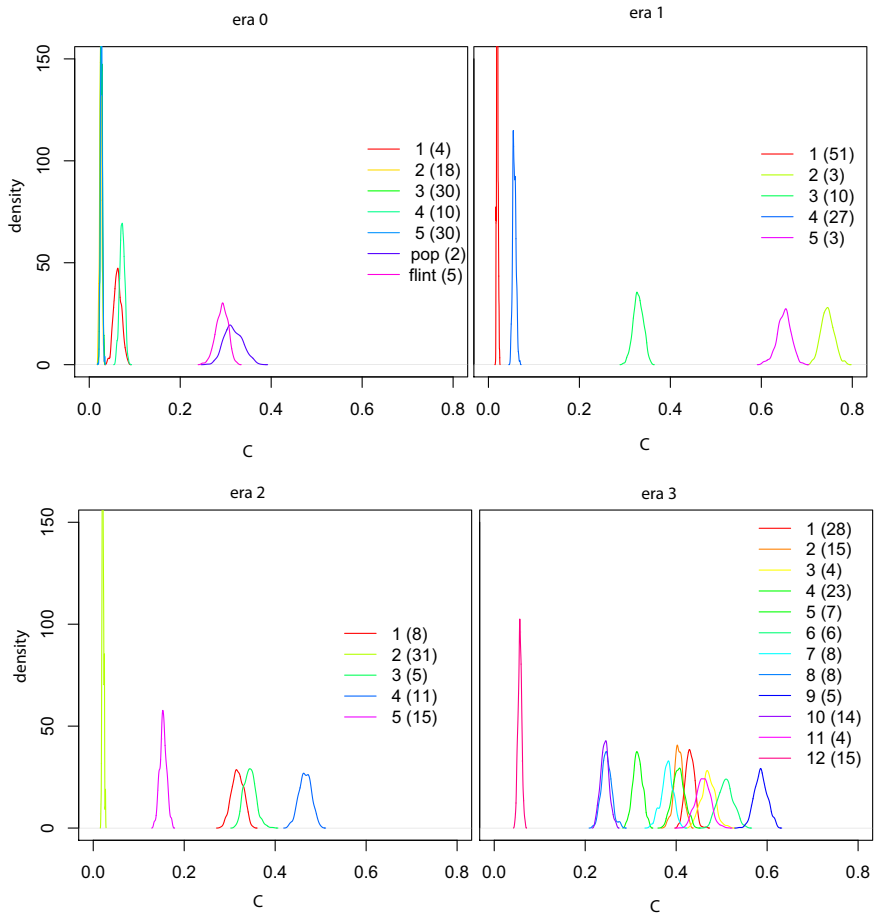

Fig. S3. Divergence from a common ancestor, C, of genetic groups within eras 0–3. Numbers between parentheses denote the number of individuals in each group.

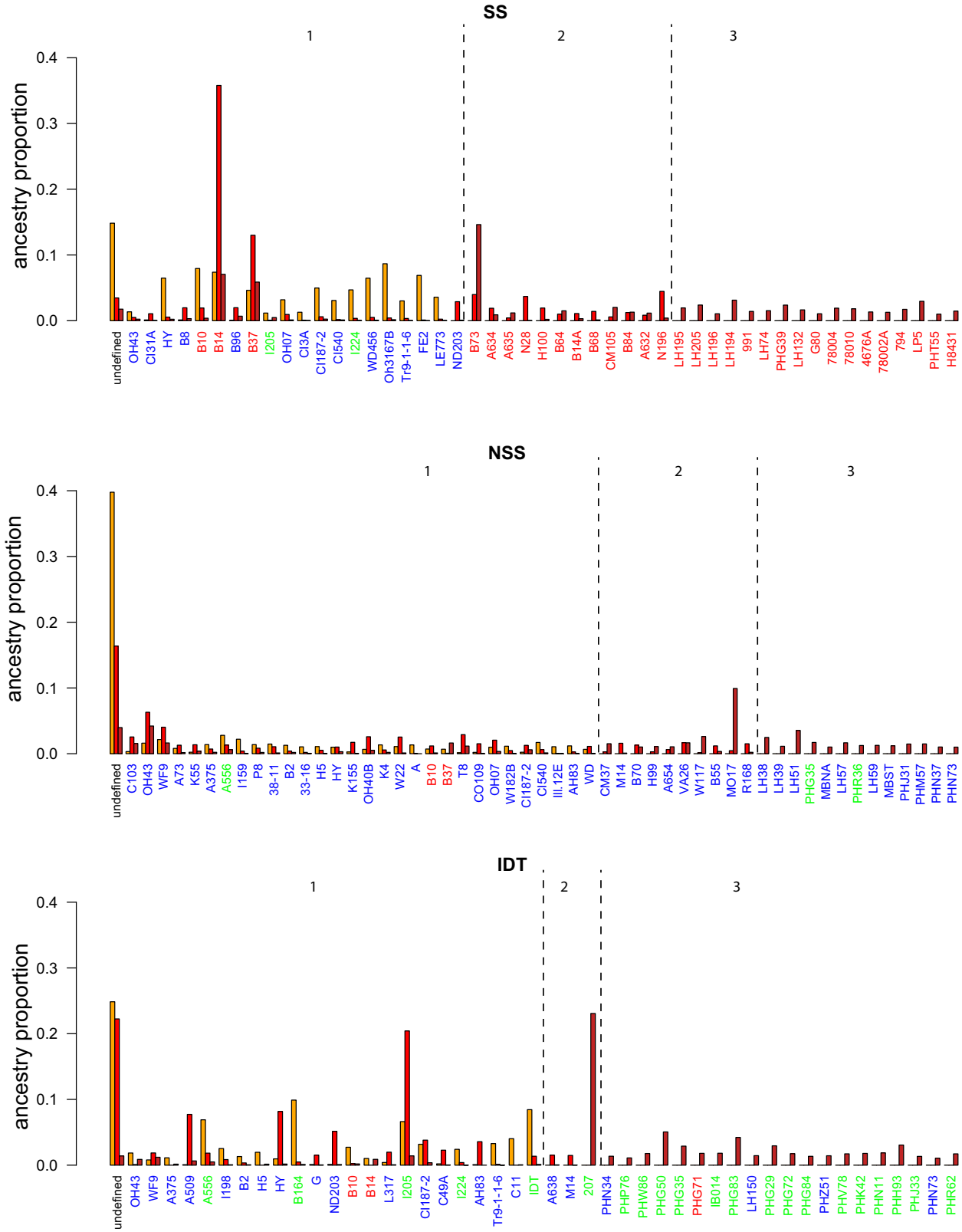

Fig. S4. Barplot of the fraction of ancestry of different lines in the different eras (orange, era 1; red, era 2; brown, era 3). Label colors indicate heterotic groups (red, SS; blue, NSS; green, IDT).

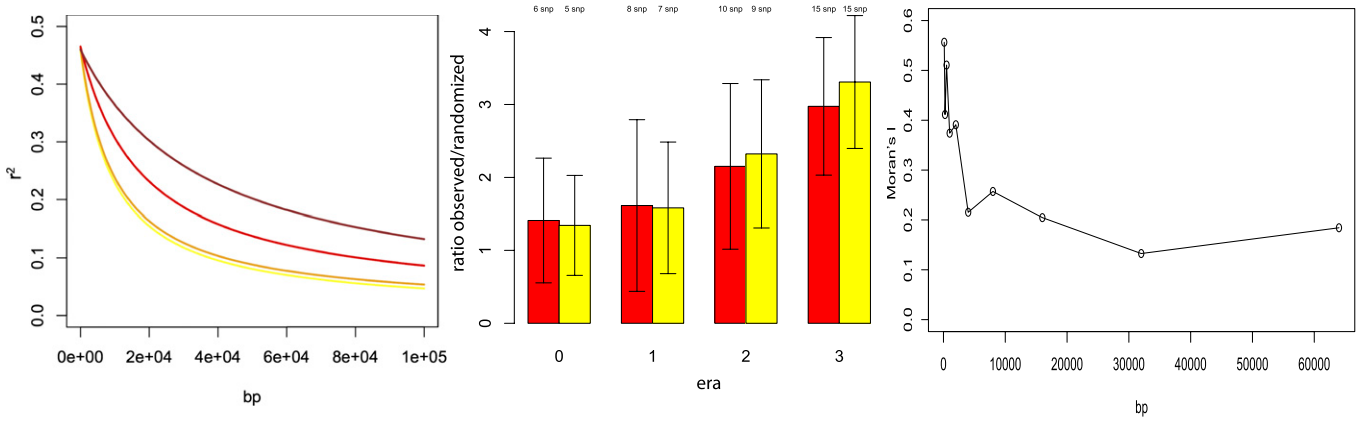

Fig. S5. Patterns of linked variation. *Left*: Linkage disequilibrium (r^2) as a function of physical distance in eras 0–3 (era 0, yellow; 1, orange; 2, red; 3, brown). *Center*: Ratio of mean observed haplotype length to that measured using randomized SNPs. The genome as a whole is indicated in red and selected regions in yellow. The mean haplotype length in SNPs of each category is shown above the bar; error bars represent one SD. *Right*: Spatial autocorrelation analysis (Moran's *I*) of evidence of selection (Bayes factor) across the genome.

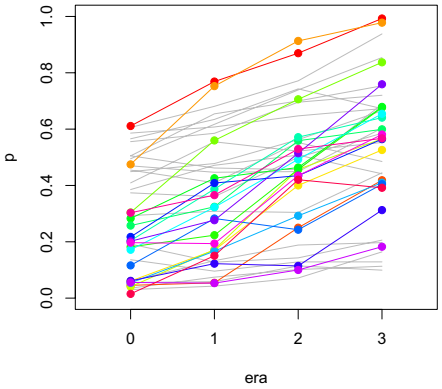

Fig. S6. Frequency change of the top 20 candidate SNPs (colored lines) across the four eras. Gray lines represent 20 random SNPs with similar frequencies in era 0.

Dataset S1. List of accessions and assigned genetic group

Dataset S1

Dataset S2. List of candidate genes with available functional information

Dataset S2

Dataset S3. Table showing the most common basal ancestor at candidate SNPs that display a reduced number of effective ancestors

Dataset S3