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.

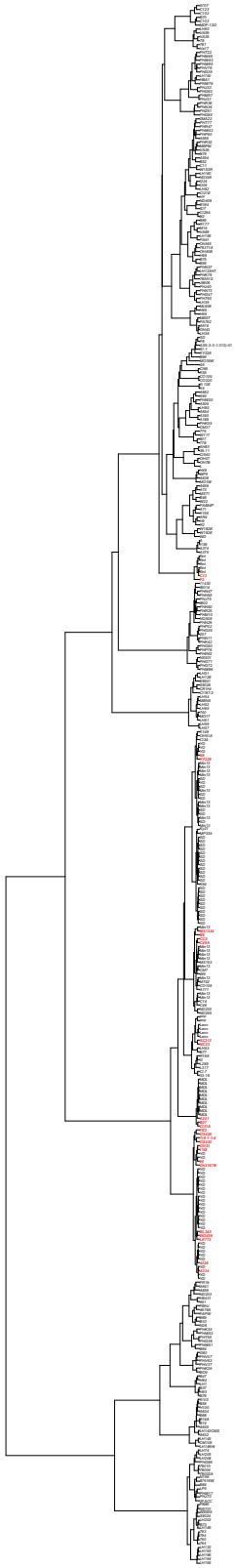

Fig. S1. Ward dendrogram based on the Euclidean distance on 39 PCs. Era 1 lines clustering with landraces are marked in red. (YD, Yellow Dents; Lanc, Lancaster; Min13, Minnesota 13; MDL, Midland; SD, Southern Dents).

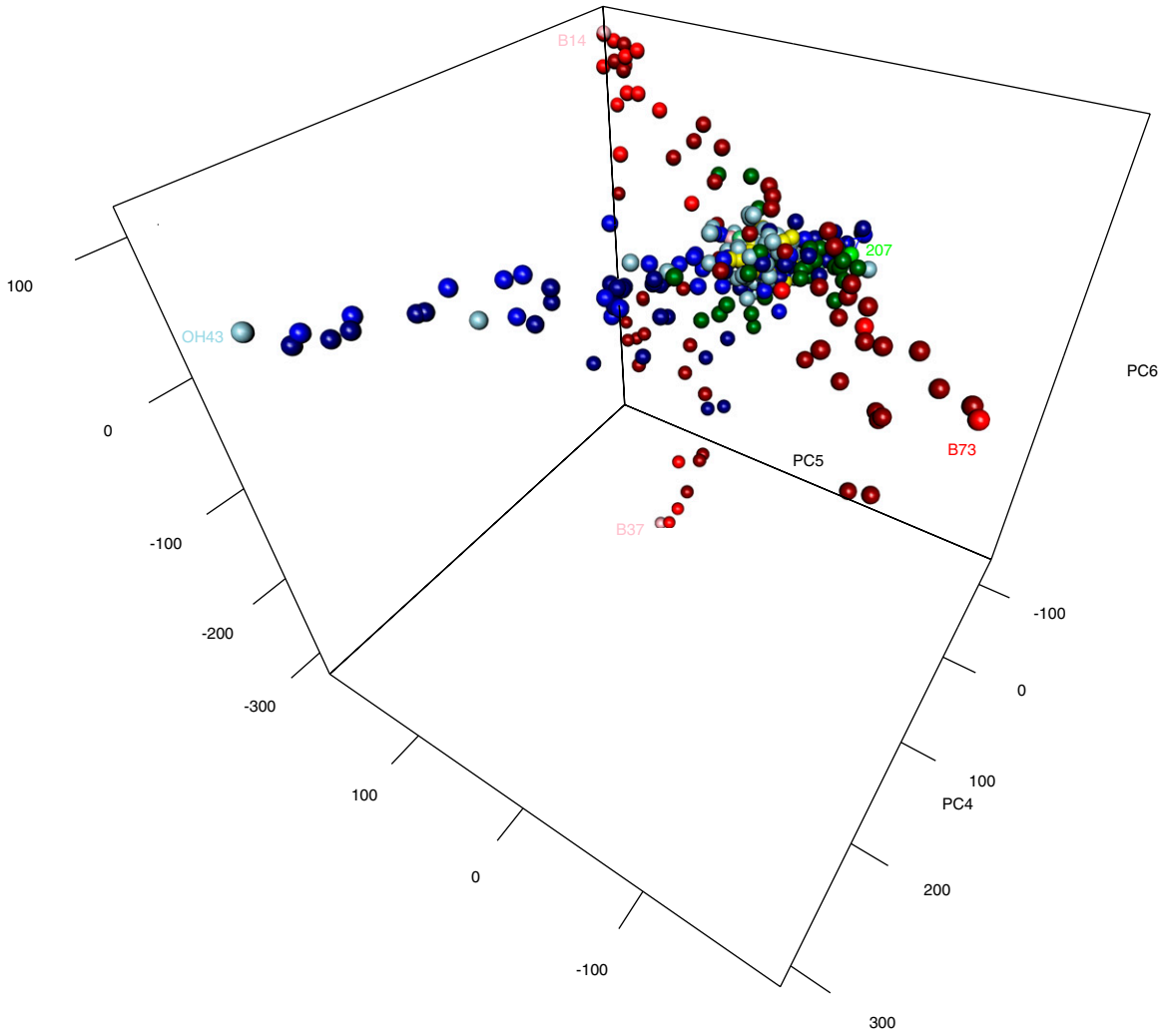

Fig. S2. PCs 3–6, obtained from PCA on all lines. Colors represent heterotic groups [red, Iowa Stiff Stalk Synthetic (SS); green, Iodent (IDT); blue, Non-Stiff Stalk (NSS)], and darker colors represent later eras (e.g., light red, era 1; red, era 2; dark red, era 3).

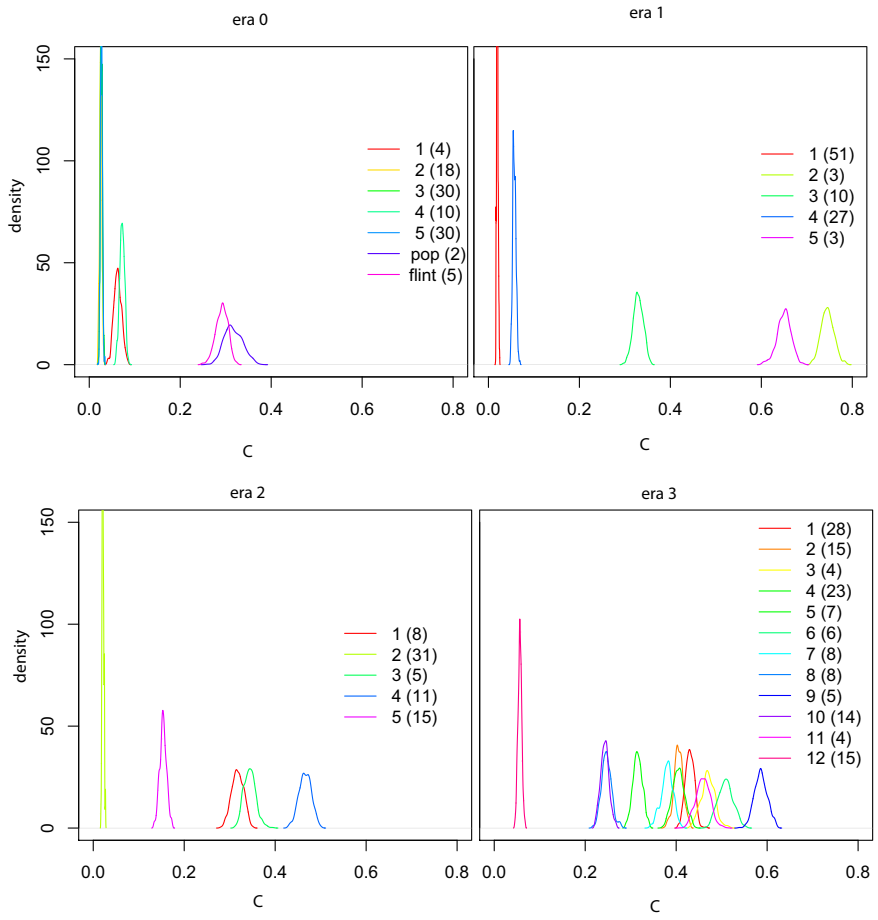

Fig. S3. Divergence from a common ancestor, C, of genetic groups within eras 0–3. Numbers between parentheses denote the number of individuals in each group.

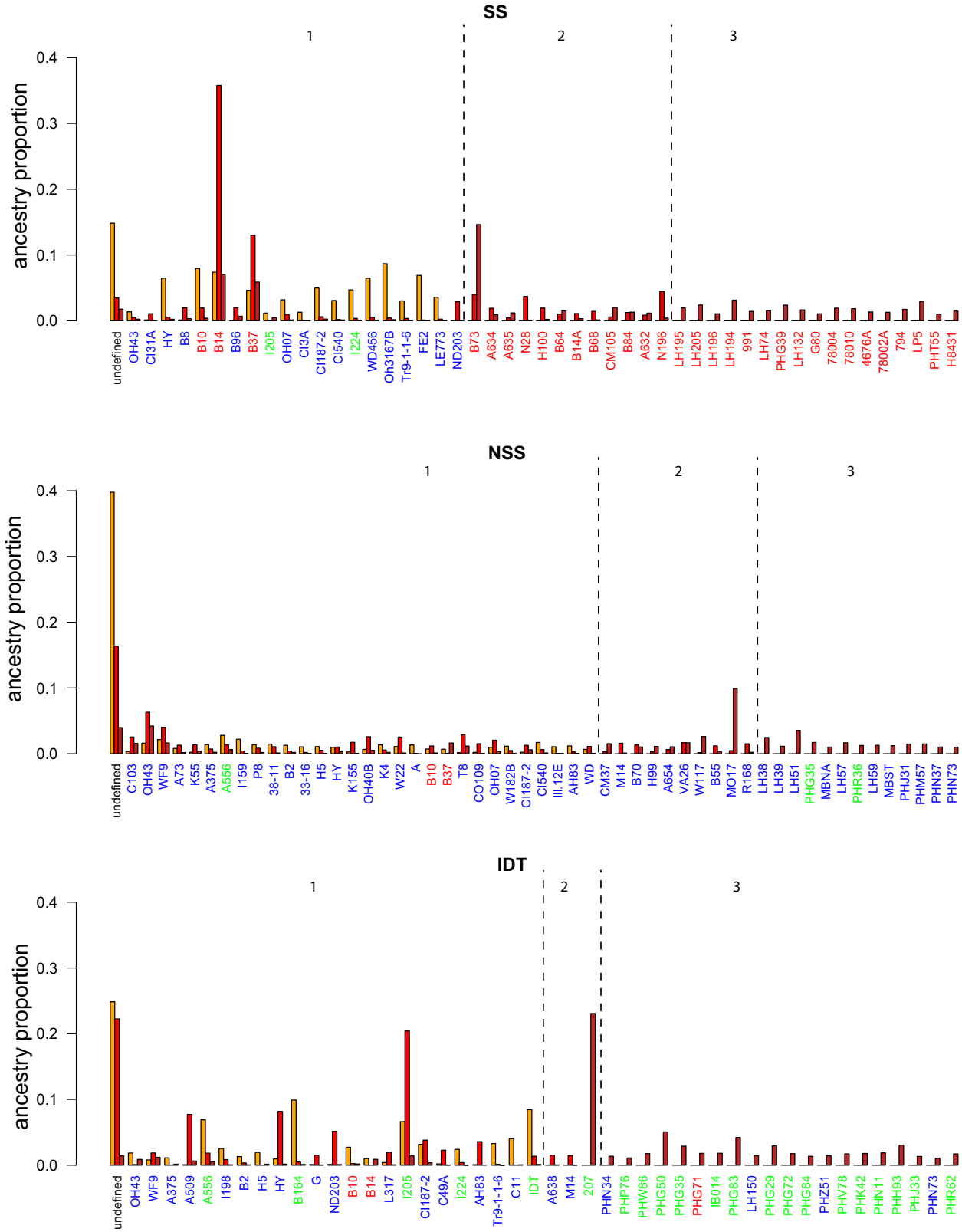

Fig. S4. Barplot of the fraction of ancestry of different lines in the different eras (orange, era 1; red, era 2; brown, era 3). Label colors indicate heterotic groups (red, SS; blue, NSS; green, IDT).

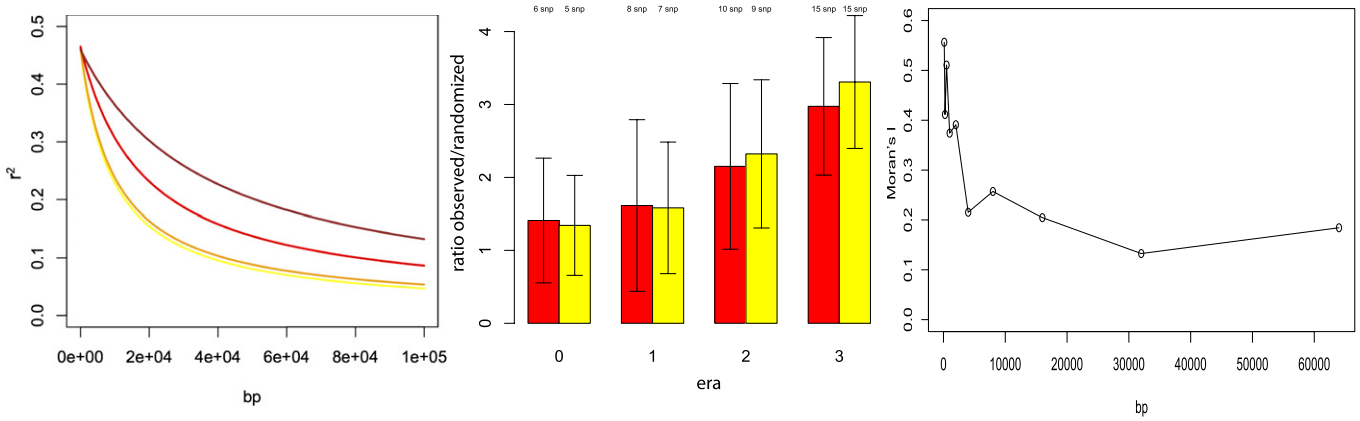

Fig. S5. Patterns of linked variation. *Left*: Linkage disequilibrium (r^2) as a function of physical distance in eras 0–3 (era 0, yellow; 1, orange; 2, red; 3, brown). *Center*: Ratio of mean observed haplotype length to that measured using randomized SNPs. The genome as a whole is indicated in red 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.

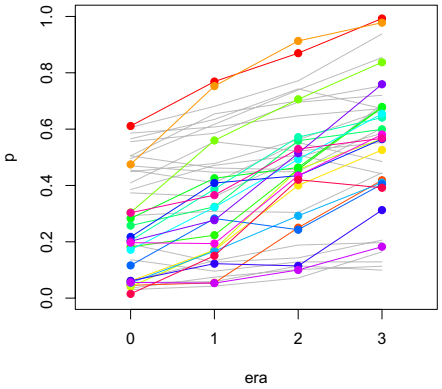

Fig. S6. Frequency change of the top 20 candidate SNPs (colored lines) across the four eras. Gray lines represent 20 random SNPs with similar frequencies in era 0.

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