Supplementary Methods: The genomic basis for short-term evolution of environmental adaptation in maize

Randall J. Wisser, Zhou Fang, James B. Holland, Juliana E. C. Teixeira, John Dougherty, Teclemariam Weldekidan, Natalia de Leon, Sherry Flint-Garcia, Nick Lauter, Seth C. Murray, Wenwei Xu, and Arnel Hallauer

Contents

1 Genotyping of variants upstream of ZmCCT10

1.1 KASP assays for SNPs

Two SNPs were genotyped using Kompetitive Allele-Specific PCR (KASPTM, V3.0 Master-Mix; LGC Genomics LLC, Beverly, MA) according to the manufacturer's protocol, except an additional 12 cycles of amplification were used. These SNPs included: (i) SNP495, a T/C polymorphism at AGPv4 position 94, 434, 777; (ii) SNP1520, a C/T polymorphism at AGPv4 position 94, 434, 924. KASP products were assayed using an ABI 7500 fast real-time qPCR instrument.

1.2 PCR assay for the presence-absence variant

ZmCCT10_CACTA, a marker for the *ZmCCT10*-associated presence-absence causal variant for photoperiodism was genotyped by scoring PCR amplicons separated in 2.0% agarose gels stained with SYBR Gold (Life Technologies, Carlsbad, CA). Genotypes were determined from two separate reactions (insertion assay primer pair: 5'-AAACGCTGACACTTCCGACT and 5'-AGCTTCGAATTTTGCTGCTG; deletion assay primer pair: 5'-GCGTACCCGAATCAAATCAA and 5'-CGTATGTGCATCCATCAGGA) amplified using standard PCR reagents (M0273L, New England BioLabs) with 0.2 μM primer, 0.1 μM dNTPs, 5 M betaine and 20 ng DNA. Thermal cycling was performed for 1 cycle for 15 min at 94 °C, 10 cycles for (i) 20 s at 94 °C and (ii) 1 min at 65 °C with each additional cycle reducing by -0.8 °C down to 57 °C, 26 cycles for (i) 20 s at 94 °C and (ii) 1 min at 57 °C, followed by 4 °C.

2 Quality Control

Custom scripts were used to perform quality control on the genotype data as follows. Based on inspection of pairwise genotypic correlations we identified one sample pair, g_0 (C0_165_2) and g_4 (C4_840_1), with an unusually high genotypic correlation ($r = 0.85$). The putative g_4 sample was most highly related to g_0 samples rather than other g_4 samples, while the g_0 sample was most related to the other g_0 samples; therefore, genotype data from C4_840_1 was removed. Five samples (C0_062314_075, C0_062314_079, C0_062314_085, C6_853_1, C6_853_4) with call rates $< 85\%$ were also removed. With the remaining samples, marker filters were sequentially applied as indicated in Table S1.

Physical map data from the B73 AGPv2 reference assembly was originally used for quality control filtering and some analyses of the MaizeSNP50 SNPs (Tables S1 and S2). We later received AGPv4 map information (Illumina Inc., San Diego, CA) and used this for subsequent analyses (Table S2). Supplemental File S1 contains all map information and indicator variables for the analysis-specific subset to which markers belonged.

3 Genetic Map Projection

To generate a genetic map including markers having only physical coordinates, centimorgan (cM) positions were projected onto the maize nested association mapping (NAM) population consensus linkage map [McMullen et al., 2009]. Using a core set of 1106 markers with physical and genetic coordinates, for each sequential interval delimited by the i^{th} pair of core markers, the cM position for each j^{th} marker nested within the interval (having only a physical coordinate) was estimated as:

$$\text{cM} = y_i + ([x_{j(i)} - x_i][y_{i+1} - y_i / x_{i+1} - x_i]); \quad (1)$$

where y_i and y_{i+1} correspond to the cM positions for core markers delimiting the i^{th} interval, x_i and x_{i+1} correspond to physical positions for the same markers, and $x_{j(i)}$ corresponds to the physical position of the j^{th} marker nested within the i^{th} interval.

4 STRUCTURE analysis

For values of $K = 1 - 8$, ten replicate runs were performed with $20,000$ burn-in iterations (exception for $K = 1$, which required $50,000$ iterations for alpha to converge) and $50,000$ MCMC iterations. At $K = 3$ and $K = 8$, some replicate runs converged on a separate optimum that was very different from the modal trend in $\mathrm{Ln}(K)$ among replicated runs; for these values of K, additional replicates were executed to obtained 10 that converged on the more frequently similar $\mathrm{Ln}(K)$ values. *CLUMPP* [Jakobsson and Rosenberg, 2007] was used to consolidate replicate runs of *STRUCTURE*, and samples were assigned to the subpopulation for which their admixture proportion was greatest (Table S3). *DISTRUCT* [Rosenberg, 2004] was used to plot the results from *CLUMPP*.

5 SIM test

A whole genome simulator was used to model the expected variation in SNP frequencies across generations under the breeding scheme for Hallauer's Tusón without selection (i.e., neutral allele frequency change). The simulation was implemented using SAEGUS (`https://github.com/maizeatlas/saegus`), an extension of simuPOP [Peng and Kimmel, 2005]. Fixed recombination rates between markers were assumed as the difference in cM values for the NAM-projected genetic map of the MaizeSNP50 markers.

The simulation was initiated using the observed genotype matrix of g_0, which captures the starting LD and structure in the population. However, because the g_0 sample was less than the original census size of the population ($n = 10,000$), a sample expansion step (next paragraph) was added to the simulation to recreate the g_0 base population. Using g_0 samples to reproduce the g_0 base population will be somewhat imprecise (sampling effect) and one additional meiosis was simulated that did occur originally; however, the population structure $\mathbf{Q}$ matrix estimated for g_0 (Table S3) was used for this step to provide a close approximation of the actual scenario.

According to the analysis of structure, g_0 was formed from six ancestral subpopulations. Using the genotype matrix of g_0 as input, an in silico population of $10,000$ individuals was formed by repeating the following steps: (i) the real individuals listed in Table S3 were assigned to one subpopulation, according to their maximum subpopulation assignment; (ii) a real individual was then drawn at random and artificially designated a female plant; (iii) the subpopulation admixture profile $\mathbf{Q}$ of the selected female was used as a probability mass function to determine the subpopulation from which a mate (male) was randomly drawn; and (iv) ignoring $\mathbf{Q}$ for the selected mate, the chosen female and male were mated in silico to produce a single offspring using the corresponding genotype data on those particular individuals. Note that this process will include within subpopulation matings according to the $\mathbf{Q}$-conditioned probability mass function. This procedure was repeated until 10,000 offspring were produced.

From these $10,000$ simulated genotypes, 400 individuals were chosen at random (without replacement) and designated as females. Simulated crosses to these females were made with a pool of 800 males that included the same 400 females and an additional set of 400 randomly chosen individuals. Therefore, mating occurred among some individuals (as males) that were not advanced as females and also included some selfing, as would be expected. Random mating among this group of females and males proceeded until $10,000$ individuals were formed. This

process was repeated for 10 generations. From among the 10,000 simulated individuals per generation $\{0, 2, ..., 10\}$, samples were taken with sizes corresponding to the real data set. For each locus and each replication of simulation, generation-specific allele and genotype frequencies were then recorded.

A function was written to determine the marker-specific two-tailed probability of the sum of sequential allele frequency differences across g_0 to g_{10} for the observed data relative to the distribution from 10,000 replications of simulation.

The SIM test was also used to examine whether SIM$^+$ markers were the same across generations. In this case, the test was applied to sequential pairs of generations. This identified a total of 2,416 SIM$^+$ markers (1% FDR), of which 84% were specific to one pair of generations (Figure S3).

6 Estimation of genetic parameters

6.1 Q_{ST}

Genetic variances for $\hat{Q}_{\mathrm{ST}} = \hat{\sigma}^2_{\mathrm{GB}}/(\hat{\sigma}^2_{\mathrm{GB}} + 2\overline{\hat{\sigma}^2_{\mathrm{GB}}})$ [Spitze, 1993, Leinonen et al., 2013] were estimated using phenotype data from [Teixeira et al., 2015], where $\hat{\sigma}^2_{\mathrm{GB}}$ and $\overline{\hat{\sigma}^2_{\mathrm{GB}}}$ are the among-generation and average within-generation additive genetic variance. The data was subset to include observations from Iowa and Delaware and check entries were removed. Iowa was the original site of selection, and these data were highly correlated with Delaware [Teixeira et al., 2015], providing additional observations for more precise estimation. When estimating $\hat{\sigma}^2_{\mathrm{GB}}$, six additional columns were added to the input file defining a grouping structure for generation-sets, where all families in a given pair of generations g_0,$g_{i=\{2,4,6,8,10\}}$ were assigned the factor level 1 and all other families were assigned the factor level 0. The following mixed linear models were fit to the data:

$$\mathbf{y} = \mathbf{X}_m\boldsymbol{\beta} + \mathbf{X}_p\boldsymbol{\psi} + \mathbf{Z}_E\mathbf{e} + \mathbf{Z}_{I(R*E)}\mathbf{i} + \mathbf{Z}_{G(P)}\mathbf{g} + \mathbf{Z}_{F(G(P))}\mathbf{f} + \mathbf{Z}_{F(G(P))*E}\mathbf{f}*\mathbf{e} + \boldsymbol{\varepsilon}; \quad (2)$$

$$\mathbf{y} = \mathbf{X}_m\boldsymbol{\beta} + \mathbf{X}_G\boldsymbol{\gamma} + \mathbf{Z}_E\mathbf{e} + \mathbf{Z}_{I(R*E)}\mathbf{i} + \mathbf{Z}_{F(G)}\mathbf{f} + \mathbf{Z}_{F(G)*E}\mathbf{f}*\mathbf{e} + \boldsymbol{\varepsilon}; \quad (3)$$

Equation 2 was used to estimate $\hat{\sigma}^2_{\mathrm{GB}}$ and equation 3 was used to estimate $\hat{\sigma}^2_{\mathrm{GW}}$. The design matrices X_M, X_P, Z_E, $Z_{I(R*E)}$, $Z_{G(P)}$, $Z_{F(G(P))}$, $Z_{G(P)*E}$ (equation 2) and X_M, X_G, Z_E, $Z_{I(R*E)}$, $Z_{F(G)}$, $Z_{F(G)*E}$ (equation 3) relate the vector of observations, $\mathbf{y}$, to the corresponding vector of effects. The fixed effects are: $\boldsymbol{\beta}$ is the overall mean; $\boldsymbol{\psi}$ is the vector of "generation-set" effects (equation 2); $\boldsymbol{\gamma}$ is the vector of generation effects (equation 3). The random effects are: $\mathbf{e}$ is the vector of environment effects; $\mathbf{i}$ is the vector of incomplete block nested in replication-by-environment interaction effects (fit for Delaware only, see below); $\mathbf{g}$ is the vector of generation nested in generation-set effects (equation 2); $\mathbf{f}$ is the vector of family nested in generation within generation-set effects (3) or the vector of family nested in generation effects (equation 3); $\mathbf{f}*\mathbf{e}$ is the vector of family nested in generation-by-environment interaction within generation-set effects (equation 2) or family nested in generation-by-environment effects (equation 3); $\boldsymbol{\varepsilon}$ is the vector of residual effects. Replications for Iowa and Delaware and incomplete blocks for Iowa were not included in the model since these were previously found to be non-significant according to the likelihood ratio test [Teixeira et al., 2015]

For equation 2, random effects were assumed to be distributed $\mathbf{MVN} \sim (0, \hat{\sigma}^2_{\{\mathbf{e},\mathbf{i}_e,\mathbf{g}_s,\mathbf{f}_s,\mathbf{f}_{s*e},\boldsymbol{\varepsilon}_r\}})$, where $\hat{\sigma}^2_{\mathbf{e}}$ is the variance among environments, $\hat{\sigma}^2_{\mathbf{i}_e}$ is the Delaware-specific variance among incomplete blocks, $\hat{\sigma}^2_{\mathbf{g}_s}$ is the genetic variance among generations in the s^{th} generation-set—$\hat{\sigma}^2_{\mathrm{GB}}$ forQ_{ST}

(for each set, the variance estimated at factor level 1 is the set-specific variance) $\hat{\sigma}^2_{\mathbf{f}_s} = \mathbf{G}\hat{\sigma}^2_{\mathbf{g}_s}$ is the s^{th} generation-set-specific additive genetic variance among families nested in generations (this is the "pooled" additive genetic variance among all families for each s_{th} set; this is not $\hat{\sigma}^2_{\mathrm{GW}}$ for each generation [see below]), $\hat{\sigma}^2_{\mathbf{f}_{s*e}}$ is the genotype-by-environment variance among families nested in generations within the s^{th} generation-set, and $\hat{\sigma}^2_{\boldsymbol{\varepsilon}_r}$ is the variance in environment-specific residuals.

For equation 3, random effects were assumed to be distributed $\mathbf{MVN} \sim (0, \hat{\sigma}^2_{\{\mathbf{e},\mathbf{i}_e,\mathbf{f}_g,\mathbf{f}_{g*e},\boldsymbol{\varepsilon}_r\}})$, where $\hat{\sigma}^2_{\mathbf{e}}$ is the variance among environments, $\hat{\sigma}^2_{\mathbf{i}_e}$ is the Delaware-specific variance among incomplete blocks, $\hat{\sigma}^2_{\mathbf{f}_g} = \mathbf{G}\hat{\sigma}^2_{\mathbf{g}_s}$ is the g^{th} generation-specific additive genetic variance among families nested within generations—$\hat{\sigma}^2_{\mathrm{GW}}$ forQ_{ST}, $\hat{\sigma}^2_{\mathbf{f}_{g*e}}$ is the generation-specific genotype-by-environment variance among families nested in generations, and $\hat{\sigma}^2_{\boldsymbol{\varepsilon}_r}$ is the variance in environment-specific residuals.

The **G** matrix was estimated using markers in set C (Table S2). For equations 2 and 3, because no covariance between individuals in different generations is assumed for $\hat{F}_{\mathrm{ST}}$, covariance in **G** between individuals from different generations was set to 0.

6.2 Estimating variance per chromosome

Equation 2 in the main text was extended to partition the additive, dominance and residual genetic variances for (i) each chromosome and (ii) SIM^+ markers. In each case, the genomic relationship matrices used to partition the variance were constructed from among set C markers (Table S2). For the former estimation, 10 chromosome-specific **G** and **D** matrices were computed from markers on each chromosome, in addition to 10 complementary **G** and **D** matrices computed from all markers on the remaining chromosomes. Ten separate models were fit to the data that each included the chromosome-specific and remaining chromosome **a** and **d** terms. For the latter estimation, separate **G** and **D** matrices were computed using set C markers classified as either SIM^- or SIM^+. These were fit in a single model that included separate **a** and **d** terms for SIM^- and SIM^+ markers. For all models, each of the random genetic effects was assumed to be distributed mutually independent.

Estimates of genetic variance components per chromosome are reported in Table S6. The sum of genetic variances in each row of the table is approximately equal to one another and to the total, genome-wide estimate for genetic variance of 32.7 (based on the model from the main text). While comparisons may be made in terms of relative amounts of variance per chromosome, we note that the chromosome-specific additive and dominance variances (column sums) are upward biased. That is, the column sum for chromosome-specific additive variance is 30.6 while the genome-wide estimate is 27.6. Similarly, the column sum for chromosome-specific dominance variance is 10.4 while the genome-wide estimate is 5.1. This indicates the residual chromosome variances (i.e., "sans chromosome" cell values in Table S6) are downward biased. We speculate this is due to covariance that is not accounted for between the specific chromosome being modeled and the residual set of chromosomes; i.e., the chromosome-specific genomic relationship matrices capture the covariance within that chromosome, while the residual genomic relationship matrices capture the covariance only within and between the remaining chromosomes.

References

[Jakobsson and Rosenberg, 2007] Jakobsson, M. and Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14):1801–1806.

[Leinonen et al., 2013] Leinonen, T., McCairns, R. J. S., O'Hara, R. B., and Merilä, J. (2013). Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nature reviews. Genetics*, 14:179–90.

[McMullen et al., 2009] McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H. H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., Pressoir, G., Romero, S., Rosas, M. O., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J. C., Goodman, M., Ware, D., Holland, J. B., and Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science*, 325(5941):737–740.

[Peng and Kimmel, 2005] Peng, B. and Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687.

[Rosenberg, 2004] Rosenberg, N. A. (2004). DISTRUCT: A program for the graphical display of population structure. *Molecular Ecology Notes*, 4:137–138.

[Spitze, 1993] Spitze, K. (1993). Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. *Genetics*, 135(2):367–374.

[Teixeira et al., 2015] Teixeira, J. E. C., Weldekidan, T., de Leon, N., Flint-Garcia, S., Holland, J. B., Lauter, N., Murray, S. C., Xu, W., Hessel, D. a., Kleintop, a. E., Hawk, J. a., Hallauer, A., and Wisser, R. J. (2015). Hallauer's Tusón: a decade of selection for tropical-to-temperate phenological adaptation in maize. *Heredity*, 114(2):229–240.