# Supplemental Methods

The following methods were used for data cleaning and processing before statistical analyses.

## Marker data

Low coverage marker data were generated through Genotyping-By-Sequencing (GBS)(Elshire *et al.* 2011) for inbred parental lines at the Cornell University Institute for Genomic Diversity and at the University of Wisconsin Biotechnology Center using the TASSEL-GBS Pipeline (Bradbury *et al.* 2007; Glaubitz *et al.* 2014) to create a TagsOnAPhysicalMap (TOPM) file from version 2 of the maize reference genome (TOPM file AGPv2). SNPs aligned to the TOPM were later uplifted to version 4 of the maize reference genome. The raw GBS dataset available on Cyverse (Cyverse) contained 955,690 markers and 1,517 inbred parental lines, including duplicates of some lines. These data underwent a number of cleaning and filtering techniques to arrive at a dataset of 1,918 hybrid crosses with 20,373 markers distributed across the genome (Figure S7).

First, markers with greater than 40% missing data were removed in TASSEL (Bradbury *et al.* 2007) using the Filter Genotype Table Sites option. Next, an analysis of depth per marker was performed to check for markers with very high sequencing depth, suggesting mapping of paralogs to a single marker location. A custom Python script was used to extract read depths associated with each genotype call from a VCF file. Depth analysis was performed in R using a custom script to examine the distribution of read depth calls for data in two ways (R Core Team). First, total depth per marker was examined to see if any markers had particularly high depth as a first pass searching for possible alignment of paralogs. Second, a matrix of depth values for each inbred line and marker combination was examined to determine a reasonable cut off for removal. Markers with any read depth over 40 corresponding to an individual genotype call were considered possible cases of alignment of paralogs and were removed from analysis, leaving 188,799 markers in the dataset.

Next, markers with greater than 20% missing data were removed using TASSEL to minimize the need for imputation of missing values (retaining 59,122 markers after this step). Heterozygosity was analyzed in TASSEL using the Analyze Heterozygosity function, and markers with greater than 5% heterozygosity among the inbred parents were removed, leaving 57,536 markers. Heterozygosity was analyzed a second time by inbred line. Inbred lines with greater than 5% heterozygosity were removed, leaving 1,371 parental lines. After this, imputation was performed using LD KNNi (Money *et al.* 2015) in TASSEL with default settings for High LD Sites = 30; Number of nearest neighbors = 10; Max distance between site to find LD = 10,000,000. This method was chosen because some samples were in the GBS dataset multiple times and samples in the dataset were known to have population structure from past studies (Romay *et al.* 2013; Peiffer *et al.* 2014). After imputation, 1.54% marker scores remained missing and were filled in as the major allele homozygote.

Using this complete data matrix, we performed principal component analysis (PCA) to check if duplicate samples properly clustered together post-imputation, and if similarity of PCA scores matched known pedigree relationships. Data passed this check and subsequently markers were thinned by position, resulting in 25,709 markers with a minimum distance of 100 bp between markers. Next, duplicate lines were removed, keeping only the observation with the highest coverage from each set of duplicates, leaving 1,186 unique parental lines. From here, monomorphic markers were removed, leaving 22,214 markers. Discrepancies in line names between genotype and phenotype data were harmonized to create a consensus list of names to be used for both datasets. PCA was performed again on the filtered parental dataset as a check to ensure data integrity and matched expectations for population structure.

The cleaned inbred genotype data were used to generate hybrid genotypes using TASSEL (Bradbury *et al.* 2007). To generate hybrid genotypes, inbred parent heterozygous calls (< 0.08% of the data) were first set to homozygous calls of one of the two alleles at the marker at random, with equal probability using a custom R script. Then TASSEL’s Create Hybrid Genotypes function was used to generate genotype scores for hybrids present in the 2014-2016 data. Of the 2,196 hybrids included in the phenotypic data, genotypes were produced for 1,918 (96%) hybrids, the remaining hybrids had at least one parent that was not genotyped or was removed from the genotype data set during filtering. Reciprocal crosses were merged into a single common hybrid entry. Genotypes were numericalized to counts of the minor allele with respect to the full panel of hybrids for all subsequent analyses. Finally, any missing genotype scores (2.13%) were imputed to 0, homozygous for the major allele. Realized additive (**A**) and dominance genomic relationship (**D**) matrices were estimated using the hybrid marker data in TASSEL (Bradbury *et al.* 2007) using centered IBS (Endelman and Jannink 2012) and dominance centered IBS methods (Muñoz *et al.* 2014), respectively (Files S13, S14, and S15). To identify genetic groups within the hybrids, cluster analysis was performed on the hybrid genotype data using Ward’s Method for agglomerative clustering (Ward 1963) with complete linkage metric. To compare the genetic relationships of hybrids and their inbred parents together, we combined the genotypic data for two sets and performed principal components analysis.

## Phenotype data

Hybrid phenotype data from each location were submitted to a common data curation team, who assembled complete data sets for all locations in each year. Trait data included days from planting to anthesis (DTA) and silk emergence (DTS), height to topmost node (plant height) and to primary ear node (ear height), stand count, number of lodged plants, plot grain yield, grain moisture at harvest, and test weight as described on in the G2F website (“The Genomes to Fields (G2F) Initiative”). Environments varied for which traits were recorded; however, all environments have measurements for grain yield, stand counts, plant heights, and ear heights.

Prior to release of the compiled trait data sets, the curators of the G2F hybrid data performed the following filtering to remove individual trait values with low biological probability. Days to anthesis or silking fewer than 20 or more than 100 days after planting were set to missing. Ear height less than 20 cm were set to missing. Yield, test weight, and grain moisture content were all set to missing if plot weight was less than 454 g. The original dataset also contains the local check hybrids, which are selected by the primary investigator at that location, known to perform well in the region, but which were not genotyped, so we removed them from our data set. Notes on a variety of data reliability issues were included with the trait data, including comments from investigators regarding damage to plots from machinery and animals, plots that suffered flooding, areas with low stand count, misplantings, and hybrid substitutions. Comments were examined to determine reliability of data, and unreliable data were discarded. Information on blocks of the experimental designs was not included in the original compiled data files, we added blocking information based on known experimental design features. In 2014-2015 fields were arranged with incomplete block designs and replications. In 2016 incomplete blocks were not part of the field experimental design, so a constant block value of 1 was assigned to all 2016 plots.

We removed all data from plots with stand counts equivalent to planting densities of less than 37,066 plants per ha-1. Plot grain yields were converted to Mg ha-1 and yields greater than 18.83 Mg ha-1 were set to missing. Days to silking less than 40 days after planting were set to missing. The ratio between ear and plant height was calculated and the plant and ear height values were set to missing wherever the ratio was below 0.25 or above 0.75. All phenotypic data from plots in an intersection of 2.5 inter-quartile range (IQR) and more than 30% above or below the mean stand count within a location were removed from further analysis. Outlier filtering was also performed on yield data by censoring yields that were outliers by both 2.5 IQR and > 30% different from the mean (Figure S8). After filtering, 31,787 plots with phenotypic observations remained. All environments data for grain yield, but some missing values were present for other traits.

During the initial Stage 1 analysis of individual environments, we discovered that the within-site heritability for all traits from Delaware in 2014 had zero heritability. Inspection of the raw data revealed that during the compilation of data from different locations, the data from this location had undergone a permutation from the original plot numbers. Information provided directly by the collaborators at the University of Delaware was used to correct the assignments of hybrids to plots. We also discovered that all traits from KSH1\_2015 and KSH1\_2016 had heritabilities less than 10%, but we were unable to identify a data compilation error. Therefore, these two sites were removed from all further analyses.

Echidna software achieved convergence relatively quickly for complex models with multiple relationship matrices or FA structures for Stage 2 analyses. Even complex G×E models were able to run in just over a day on eight 3.5 GHz processors and 64 GB RAM using Echidna Version 0.92 without specifying initial parameter values. All models reported here would be considered feasible for a breeding program to implement for use in making selection decisions in a timely fashion.

## Weather Data

Spectrum (Aurora, IL) watchdog weather stations were placed adjacent to fields to collect readings on temperatures, humidity, dew point, precipitation, wind direction and speed, and solar radiation every thirty minutes. However, some stations were not deployed until up to 6 weeks after planting, and some stations malfunctioned, resulting in large amounts of missing data in the raw weather datasets. To remedy this, values for missing weather variables except solar radiation, were imputed using daily measurements from the nearest local airport weather station. First, 30-minute resolution Spectrum station weather data was summarized to daily values (File S16). Maximum, minimum, and mean daily temperatures were calculated, along with daily mean dewpoints, humidity, wind speed, and total precipitation. Missing values for temperatures, dew point, windspeed, and total precipitation were imputed using information scraped from Weather Underground’s (https://www.wunderground.com/) historical database from the closest airport (“Weather Underground”). Wind speed and direction measurements from all Spectrum stations were considered to be unreliable because the winds were partly blocked once adjacent maize plants grew to the height of the anemometer. Therefore, all wind speed values were imputed using web scraping from Weather Underground (“Weather Underground”). As a check on weather data reliability, the weather data for each environment was compared to data from same dates at nearest weather stations included in the Iowa Environmental Mesonet (IEM) ASOS (“Iowa Environmental Mesonet” 2019). Daily Spectrum station and IEM station data were correlated highly in 2014 and 2015 but had approximately zero correlation in the 2016 data set (File S17). Further investigation revealed a permutation in the initially released 2016 environmental data, rendering it unusable. For this reason, all 2016 Spectrum station environmental data were discarded and replaced with the nearest 2016 IEM data for respective locations as a best proxy for field data. Data from 2014 and 2015 was checked for outliers where G2F and IEM data for temperature related variables differed by more than 7 °C. In these cases, three data sources closest to the environment (Spectrum data, IEM, and Weather Underground were compared to determine a consensus value. Where Weather Underground agreed with the Spectrum data, the Spectrum data point was kept, and where Weather Underground agreed with IEM, the IEM values were kept. Temperature, dew point, and relative humidity are all interrelated, providing a means to impute missing data for one of the values when the other two measurements were recorded. Dew point () is a function of relative humidity (; the ratio of actual water vapor pressure to saturation vapor pressure) and mean daily temperature () to dew point as follows (Lawrence 2005):

This equation was tested on observations where all three values were present to check for accuracy. Finding high accuracy (), this calculation was used to impute dew point for observations where readings for and were present, but was missing.

A formula to calculate relative humidity was derived from equation 1 and is given in equation 2 by solving algebraically for relative humidity. This equation was tested on observations where all three values were present from the in-field weather station to check for accuracy and was found to be highly accurate (). Equation 2 was used to impute relative humidity for observations where readings for and were present, but was missing.

After web scraping and imputation, growing degree days () were calculated for all daily data points using equations from the NDAWN(North Dakota Agricultural Weather Network). is a measure used to understand plant growth and development, calculated using temperature data. An indicator variable for rain was created with a value of zero on days when no precipitation was captured by the Spectrum Watchdog, and one otherwise. Irrigation was captured by the Spectrum watchdog weather stations in most locations. For locations noted to not have irrigation collected by the watchdog, the amount of irrigation applied was added as precipitation for consistency amongst environments. In 2016, irrigation values were added to the imputed weather data from IEM.

Solar radiation is known to follow a cyclical pattern with little to no detectable radiation activity when no light is present (i.e. night), and activity peaking around midday when the sun reaches its zenith (Myers 2013). Field station data showed high radiation activity in many sites during nighttime daytime hours, indicating that the sensors likely were not functioning properly. Data from the National Solar Radiation Database (<https://maps.nrel.gov/nsrdb-viewer/>) were used to impute solar radiation for each site by accessing the data from the closest collecting station to the field. These data were then used to estimate the mean Global Horizontal Irradiance (GHI) for each day along with the maximum GHI observed for each day. The maximum GHI will be referred to as peak GHI hereafter.

Effective photoperiod was calculated using tables of sunrise and sunset times from the United States Naval Observatory under the recommended guidelines of computing the length in minutes between sunrise and sunset times and adding one hour to give effective day length (United States Naval Observatory).

Field soil samples were taken during the 2016 field season, but no soil sample measurements existed for the 2014 or 2015 seasons. For 2016, we took the soil composition measurements (sand, silt, clay percentages) for locations where information was available. For locations where no such information was available, data from the USDA Natural Resources Conservation Services Web Soil Survey (United States Department of Agriculture) was used to impute soil type. For imputation, web soil survey soil types and percentages were recorded and used to compute approximate soil composition values for sand silt and clay in R (File S16).

Daily weather data for each year was processed in R using a custom sliding window function (File S16) to reduce dimensionality into multiple datasets with resolution ranging between five-day and 30-day periods. The mean values for each variable were averaged over days within each window. All three years of periodic mean weather data from all years were then combined into a common data set that was centered and scaled by variable prior to analysis (File S18).

Citations

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al.*, 2007 TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

Cyverse www.cyverse.org.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A Robust , Simple Genotyping-by-Sequencing ( GBS ) Approach for High Diversity Species. PLoS One 6: 1–10.

Endelman, J. B., and J. Jannink, 2012 Shrinkage Estimation of the Realized Relationship Matrix. G3 Genes, Genomes, Genet. 2: 1405–1413.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS : A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS One 9:.

Iowa Environmental Mesonet, 2019.

Lawrence, M. G., 2005 The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. Bull. Am. Meteorol. Soc. 86: 225–233.

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.-Y. Zhong *et al.*, 2015 LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3 Genes, Genomes, Genet. 5: 2383–2390.

Muñoz, P. R., M. F. R. Resende, S. A. Gezan, M. Deon Vielela Resende, G. de los Campos *et al.*, 2014 Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. Genetics 198: 1759–1768.

Myers, D. R., 2013 Fundamentals of Solar Radiation, pp. 1–13 in *Solar Radiation: Practical Modeling for Renewable Energy Applications*, edited by A. Ghassemi. CRC Press, Boca Raton, FL.

North Dakota Agricultural Weather Network Corn Growing Degree Days (GDD).

Peiffer, J. A., M. C. Romay, M. A. Gore, S. A. Flint-Garcia, Z. Zhang *et al.*, 2014 The genetic architecture of maize height. Genetics 196: 1337–1356.

R Core Team R: A language and environment for statistical computing.

Romay, M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts *et al.*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14: R55.

The Genomes to Fields (G2F) Initiative.

United States Department of Agriculture Natural Resources Conservation Service: Web Soil Survey.

United States Naval Observatory Sun or Moon Rise/Set Table for One Year.

Ward, J. H. J. ., 1963 Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58: 236–244.

Weather Underground.