The concepts of general combining ability (GCA) and specific combining ability (SCA), developed by Sprague and Tatum (1942), greatly enhanced the design of hybrid breeding programs (Hallauer and Miranda, 1988; Fu et al., 2014). General combining ability describes the average performance of an inbred line when crossed to many other inbred lines, whereas SCA describes the deviation in performance of specific parental combinations from what is expected based on parental GCA. Typically, GCA variation results from additive gene action, and SCA variation results from dominance gene action, whereas additive-by-additive epistasis contributes to both (Lynch and Walsh, 1998). Although research has shown that variation in GCA is the predominant source of variation among maize (Zea mays L.) hybrids, variance in SCA is still important, especially for grain yield (GY) (Hallauer and Miranda, 1988; Bernardo, 1996; Technow et al., 2014). Variation in SCA necessitates the evaluation of specific parental combinations (single crosses) to identify superior hybrids.

Evaluation of single crosses would ideally be performed using all possible crosses among candidate parental lines. However, the sheer number of possible crosses—$N_1 \times N_2$ for a factorial or $N(N - 1)/2$ for a partial diallel—is far too large to evaluate in the early stages of hybrid breeding programs. Recent simulation and field studies have shown great promise of genomic prediction of single-cross performance. These previous studies, however, have primarily focused on parametric genomic prediction models. This study tested three nonparametric models—reproducing kernel Hilbert spaces, support vector regression, and neural networks—for prediction of early-stage single crosses. Two separate datasets, consisting of 481 and 312 single crosses, were used to evaluate models. Single crosses were made by randomly crossing inbred progenies between heterotic groups. Genomic prediction models were trained to directly predict single-cross performance, or to predict general combining ability (GCA) of inbred parents and specific combining abilities (SCA) of single crosses between them. Using cross-validation, genomic predictions were compared with predictions using phenotypes of single crosses with a common parent (common-parent single crosses), as well as phenotypic estimates of GCA. Of these three options for predicting single-cross performance, genomic prediction resulted in the highest correlation between observed and predicted values. Predictive abilities of parametric and nonparametric genomic prediction models were nearly identical. All genomic prediction models displayed good ability to predict GCA effects, but none could predict SCA effects. Our results suggest that nonparametric models do not provide an advantage over parametric models for prediction of early-stage, single-cross performance using modestly sized training populations like those used here.

**ABSTRACT**

Prediction of single-cross performance in a hybrid breeding program is extremely important because it is not feasible to evaluate all parental combinations. Recent simulation and field studies have shown great promise of genomic prediction of single-cross performance. These previous studies, however, have primarily focused on parametric genomic prediction models. This study tested three nonparametric models—reproducing kernel Hilbert spaces, support vector regression, and neural networks—for prediction of early-stage single crosses. Two separate datasets, consisting of 481 and 312 single crosses, were used to evaluate models. Single crosses were made by randomly crossing inbred progenies between heterotic groups. Genomic prediction models were trained to directly predict single-cross performance, or to predict general combining ability (GCA) of inbred parents and specific combining abilities (SCA) of single crosses between them. Using cross-validation, genomic predictions were compared with predictions using phenotypes of single crosses with a common parent (common-parent single crosses), as well as phenotypic estimates of GCA. Of these three options for predicting single-cross performance, genomic prediction resulted in the highest correlation between observed and predicted values. Predictive abilities of parametric and nonparametric genomic prediction models were nearly identical. All genomic prediction models displayed good ability to predict GCA effects, but none could predict SCA effects. Our results suggest that nonparametric models do not provide an advantage over parametric models for prediction of early-stage, single-cross performance using modestly sized training populations like those used here.
stages of a breeding program, as the number of candidate parental lines is typically very large (in the 100s to 10,000s). Various approaches have been proposed to evaluate the performance of candidate parents to be used in single crosses, including inbred per se performance, crossing to a tester line (topcross test) (Jenkins and Brunson, 1932), genetic distance estimated with molecular markers (Melchinger, 1999), best linear unbiased prediction using pedigree or marker data (Bernardo, 1994, 1996), and markers associated with hybrid performance (Vuylsteke et al., 2000). These approaches, however, have been found to have limited utility, and newer methods of hybrid prediction need to be developed (Schrag et al., 2009).

Models based on whole-genome marker data, known as genomic prediction models, have been developed in recent years to predict genetic values for complex traits (Meuwissen et al., 2001). Prospects of genomic prediction for single-cross performance have been investigated recently in maize (Massman et al., 2013; Technow et al., 2014). Genomic prediction for single-cross performance involves the prediction of genetic values of all possible single-cross combinations based on the genotypic data of all inbreds and phenotypic data on a subset of single crosses between them. Massman et al. (2013) used ridge regression best linear unbiased prediction (RRBLUP) and genomic best linear unbiased prediction (GBLUP) to predict the performance of untested single crosses. The mean prediction accuracies were 0.87 for GY, 0.90 for grain moisture, 0.69 for stalk lodging, and 0.84 for root lodging. In another study, Technow et al. (2014) used GBLUP and BayesB with prediction accuracies ranging from 0.75 to 0.92 for GY, and 0.59 to 0.95 for grain moisture.

The parental inbred lines used in the abovementioned studies on genomic hybrid prediction can be described as being from the advanced stages of a maize hybrid breeding pipeline. That is, they were either elite parental lines already used in commercial hybrid production, or they resulted from multiple stages of selection based on topcross performance. Very few, if any, of the inbred lines used in these studies were from the same breeding cross, as would be the case during the early stages of a maize hybrid breeding pipeline. The full potential of genomic prediction in this context would be better assessed by predicting and testing all possible single crosses among many random inbred progenies derived from breeding crosses before selection to capture as much of the genetic space of the breeding program as possible. Kadam et al. (2016) used GBLUP to predict the performance of early-stage single crosses made among random recombinant inbred lines (RILs) and doubled haploid lines (DHLs) from three biparental families from each heterotic group. Observed prediction accuracies for GY ranged from 0.39 to 0.77 depending on the model used and number of parents represented in the training set. The variance captured by modeling the SCA component was 12% of the sum of the GCA variance. Therefore, further investigation into the potential of genomic prediction for single-cross performance with specific emphasis on prediction of SCA is desirable.

Parametric genomic prediction models are not well suited to capture nonadditive genetic effects (Gianola et al., 2006; Howard et al., 2014). These models regress the phenotype on marker covariates using some type of regularization or variable selection procedure (de los Campos et al., 2013). Although both additive and nonadditive effects can be included in these models by adding appropriate interactions between marker covariates, the partitioning of genetic value into additive, dominance, and epistasis used in parametric models holds only under idealized conditions (Gianola et al., 2006). These conditions often do not hold in a typical breeding program, limiting the effectiveness of these models to capture nonadditive effects. Moreover, the sheer dimensionality of the marker data used in genomic prediction could easily result in hundreds of millions or billions of interaction effects to estimate, which is a challenge not easily met by parametric methods (Gianola et al., 2010).

Nonparametric models have been proposed to exploit nonadditive genetic effects in genomic prediction (Gianola et al., 2006, 2010). These models focus on prediction and relax assumptions about the form of the relationship between markers and phenotype. These models rather seek a form that best fits the training data while maintaining some generality for new data. This is in contrast with parametric models, where the focus is on parameter estimation rather than prediction. This distinctive feature of nonparametric models is expected to better enable accounting for nonadditive genetic effects without explicitly modeling them, thus possibly improving the ability to predict phenotypes for complex traits (Gianola et al., 2010). Common nonparametric models used in the context of genomic prediction include reproducing kernel Hilbert spaces (RKHS) (Gianola et al., 2006, 2010), support vector regression (SVR) (Long et al., 2011), and neural networks (NN) (Gianola et al., 2011). Empirical studies using these models have reported accuracy similar to or greater than the benchmark models RRBLUP and GBLUP (Heslot et al., 2012; Pérez-Rodríguez et al., 2012; Crossa et al., 2013). Crossa et al. (2013) obtained a 5 to 18% improvement in prediction accuracies using RKHS as compared with GBLUP for GY in a testcross population of maize lines. Heslot et al. (2012) compared different parametric and nonparametric models for predicting various quantitative traits over eight different datasets. In their study, the predictive ability of RKHS was greater than RRBLUP in 16 of 18 the comparisons made. In a study by Pérez-Rodríguez et al. (2012), mean prediction accuracies of RKHS and NN were better than Bayesian ridge regression for predicting
days to heading and GY in a wheat (*Triticum aestivum* L.) dataset consisting of elite lines.

The suggested potential of nonparametric models to capture nonadditive effects makes them interesting candidates for genomic prediction of single-cross performance, which is considerably influenced by SCA (Massman et al., 2013; Technow et al., 2014; Kadam et al., 2016). However, to our knowledge, no studies have yet compared these models for single-cross prediction. Maenhout et al. (2007) compared GBLUP with SVR and found little advantage of SVR in terms of predictive ability of single crosses, but the number of markers used in this study was minimal (101 microsatellite markers). With these considerations, the main objective of the present study was to evaluate three nonparametric genomic prediction models—RKHS, SVR, and NN—for prediction of single crosses among random inbred lines from a limited number of families. Genomic prediction of single crosses was first compared with phenotypic GCA-based predictions and performance of single crosses with a common parent (common-parent single crosses) to establish a baseline. Four genomic prediction models were subsequently investigated in detail for single-cross prediction. Two different datasets consisting of 481 and 312 single crosses made by randomly crossing RILs or DHLs belonging to Iowa Stiff Stalk Synthetic (BSSS) and Non-Stiff Stalk Synthetic (NSSS) heterotic groups were used for this study. The parents of tested single crosses were genotyped using genotyping-by-sequencing, and single cross performance was predicted based on observed performances of tested single crosses or combining abilities of their parents.

## MATERIALS AND METHODS

### Plant Materials and Field Experiments

#### Dataset I

The germplasm in Dataset I consisted of 481 single crosses created by crossing 89 RILs derived from six BSSS biparental families to 103 RILs derived from six NSSS biparental families. The parents of the biparental families were Plant Variety Protection expired (ex-PVP) lines (Table 1). The six biparental families from the BSSS heterotic group were created by crosses among eight ex-PVP parents, and the six biparental families from the NSSS heterotic group were created by crosses among six ex-PVP parents. The RILs were $F_1$ derived with no intentional artificial selection applied. Parent information, number of RILs per family, and single crosses per family are displayed in Table 1. The number of single crosses made per single BSSS and NSSS RIL ranged from 1 to 14 and 1 to 8, respectively. The mean number of single crosses per BSSS RIL was 5.4, and for NSSS it was 4.7. The BSSS RILs were designated as females, and the NSSS RILs were designated as males. Field trials to evaluate agronomic traits of single crosses were conducted in Mead and York, NE, during 2014, and Havelock and York, NE, during 2015, for a total of four distinct environments. Trials included 450 single crosses in 2014 and 467 single crosses in 2015, with 436 single crosses evaluated in both the years. The total number of single crosses evaluated in at least 1 yr was 481. The experimental design was a randomized complete block with two replications at each environment. Plots consisted of two rows, 4.46 m in length and 0.76 m apart, planted to a density of 88,506 seeds ha$^{-1}$. Plant height (PH) was measured from the base of the plant to the collar of the flag leaf postanthesis. Three plants were randomly chosen per row, and the mean of the six PH measurements was taken to represent the plot PH. Plots were machine harvested to determine grain weight per plot, which was converted to megagrams per hectare on a 155-g kg$^{-1}$ moisture basis. Grain yield data points from plots having $>10$ lodged plants were discarded.

#### Dataset II

This dataset was used previously in Kadam et al. (2016), from which detailed information on experimental design and measurement can be found. Briefly, the germplasm in Dataset II consisted of 312 single crosses made using an incomplete factorial design between 46 RILs or DHLs belonging to the BSSS heterotic group and 171 RILs or DHLs belonging to the NSSS heterotic group. The RILs or DHLs were derived from three interconnected biparental families per heterotic group. Approximately 10% of these lines were RILs and 90% were DHLs. Henceforth, the RILs and DHLs will be collectively referred to as “RILs” to simplify the discussion of Datasets I and II together. Parent information and number of RILs and single crosses per family are displayed in Table 2. The mean number of single crosses per

### Table 1. The parents of each biparental family, number of recombinant inbred lines (RILs) in each biparental family (given in parentheses), and number of single crosses for each of 36 family-wise cross combinations in Dataset I. The total number of single crosses per biparental family are listed in the margins.

<table>
<thead>
<tr>
<th>NSSS‡</th>
<th>BSSS†</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHG50×LH123 (20)</td>
<td>PHZ51×LH123 (14)</td>
</tr>
<tr>
<td>LHI32×B73 (16)</td>
<td>26</td>
</tr>
<tr>
<td>PHB47×PHK29 (19)</td>
<td>19</td>
</tr>
<tr>
<td>PHB47×PHW52 (17)</td>
<td>29</td>
</tr>
<tr>
<td>PHJ40×PHW52 (3)</td>
<td>6</td>
</tr>
<tr>
<td>LH74×LH132 (14)</td>
<td>8</td>
</tr>
<tr>
<td>PHG86×PHW52 (20)</td>
<td>25</td>
</tr>
<tr>
<td>Total (89)</td>
<td>113</td>
</tr>
</tbody>
</table>

† BSSS, Iowa Stiff Stalk Synthetic.
‡ NSSS, Non-Stiff Stalk Synthetic.
individual BSSS RIL was 6.9, whereas it was 1.8 for and individual NSSS RIL. The BSSS RILs were designated as females, and the NSSS RILs were designated as males. Single crosses comprising Dataset II were evaluated in Illinois in 2012 at two locations, and in 2013 at three locations. Two locations were common between the years. The five location–year combinations were defined as separate environments. The experimental design was an α(0, 1) incomplete block design (Patterson and Williams, 1976) with three replications at each environment. The phenotypic data were recorded for several agronomic traits. Grain yield was converted to megagrams per hectare (Mg/ha) on a moisture basis. Plant height was measured post-anthesis on a single representative plant determined by visually surveys the entire plot before measurement.

Genotyping by Sequencing
Genotyping of RILs included as part of Dataset I was performed using genotyping-by-sequencing (Elshire et al., 2011). Briefly, five seeds of each RIL were planted in the greenhouse for leaf sample collection and pooled leaf tissue was immediately frozen in liquid N. DNA was extracted from lyophilized leaf samples using the Qagen DNaseasy Plant 96 kit. Library preparation and sequencing were performed at the Institute for Genomic Diversity at Cornell University as described by Elshire et al. (2011). Single nucleotide polymorphisms (SNPs) were called from the raw sequence data using the TASSLE GBS Pipeline version 3.0 (Glaubitz et al., 2014) on the combined set of BSSS and NSSS RILs. Marker profiles of single crosses were inferred from their parental SNP information. The SNPs with >20% missing values and <5% minor allele frequency were removed wherein heterozygous SNPs calls were set to missing values. Missing values were subsequently imputed using naive imputation (i.e., missing values were replaced by mean coded value for the marker). The final marker dataset consisted of 23,923 SNPs common across the BSSS RILs, NSSS RILs, and single crosses. The procedure used for DNA extraction, genotyping, and SNP calling in Dataset II was described in Kadam et al. (2016). The marker profiles of single crosses were inferred from their parental SNP information. The SNPs were filtered for 5% minor allele frequency. Only SNPs that were common to all three sets (BSSS RILs–doubled haploids, NSSS RIL–doubled haploids, and single crosses) were retained for further analysis. A total of 2273 SNPs remained after filtering.

Phenotypic Data Analysis
Analysis of phenotypic data across the environments was performed for each dataset using the following statistical model:

\[
y_{ijkl} = \mu + g_i + e_j + g_e_{ik} + r_{ij} + b_{q(k)} + \varepsilon_{ijkl} \quad [1]
\]

where \(y_{ijkl}\) is the phenotypic observation of the \(i\)th single cross evaluated in the \(k\)th environment in the \(j\)th complete block (i.e., replication) and \(q\)th incomplete block, \(\mu\) is the grand mean, \(g_i\) is the effect of the \(i\)th single cross, \(e_j\) represents the effect of the \(j\)th environment, \(g_e_{ik}\) represents the interaction effect between the \(i\)th single cross and \(k\)th environment, \(r_{ij}\) represents the effect of the \(i\)th complete block nested within the \(j\)th environment, \(b_{q(k)}\) represents the effect of the \(q\)th incomplete block nested within the \(k\)th complete block in the \(j\)th environment, and \(\varepsilon_{ijkl}\) represents the residual. Environment and replication nested within environment effects were modeled as fixed effects. All other effects were treated as random. Single cross (\(g\)) and single cross-by-environment interaction effects (\(g_e\)) effects were assumed to be sampled from independent normal distributions with common variances among environments. Incomplete block \(b_{q(k)}\) and residual effects, however, were assumed to be sampled from normal distributions with heterogeneous variances among environments. Stand count was included as a covariate in Eq. [1] for analysis of GY in Dataset I. The stand count covariate was significant \((P < 0.05)\) and inclusion of stand count as a covariate reduced the error variance by 4%.

The restricted maximum likelihood (REML) estimates of all variance components were obtained using ASReml-R software (Butler et al., 2009). Significance of the variance components was determined using likelihood ratio tests at 0.001 level of significance. The entry–mean heritability of each trait was estimated as

\[
H^2 = \frac{\sigma^2_s}{\sigma^2_e + \sigma^2_{e_x}}
\]

where \(\sigma^2_s\) represents the variance among single crosses, \(\sigma^2_e\) represents the residual variance, \(\sigma^2_{e_x}\) represents the variance of single cross \(\times\) environment interaction effects, \(e\) is the number of environments, and \(r\) is the number of replications per environment. The phenotypic data were unbalanced due to missing observations in Dataset II. Therefore, \(e\) and \(r\) were substituted by the harmonic mean of number of observations per single cross within an environment, and harmonic mean of total number of observations per single cross as suggested by Holland et al. (2003).

The GCA and SCA effects were estimated using the following model:

\[
y_{ijkl} = \mu + f_i + m_j + e_{ij} + e_k + f_e_{ik} + m_e_{jk} + s_{e_{ijk}} + r_{q(k)} + b_{q(k)} + \varepsilon_{ijkl} \quad [2]
\]
where $y_{ijg}$ is the phenotypic observation for single cross between $i$th female RIL and $j$th male RIL evaluated in the $k$th environment in the $l$th complete block (i.e., replicate) and $q$th incomplete block, $\mu$ is the grand mean, $f_i$ is the GCA effect of $i$th female, $m_j$ is the GCA effect of $j$th male, $s_{ijklq}$ represents the effect of the $ijklq$th environment, and $f_{ik}$, $m_{jk}$, and $s_{ijklq}$ represent the interaction effects of female GCA, male GCA, and cross SCA with the $k$th environment. The distributions of $f_i$, $m_j$, and $s_{ijklq}$ were assumed as follows: $f_i \sim N(0, \sigma^2_{GCA,f})$, $m_j \sim N(0, \sigma^2_{GCA,m})$, and $s_{ijklq} \sim N(0, \sigma^2_{SCA})$ where $\sigma^2_{GCA,f}$ is the variance of GCA effects of females, $\sigma^2_{GCA,m}$ is the variance of GCA effects of males, and $\sigma^2_{SCA}$ is the variance of SCA effects. All remaining terms are as described in Eq. [1].

**Genomic Best Linear Unbiased Prediction**

The GBLUP model with additive and dominance effects of single crosses has the form

$$y = 1_n \mu + Z_A \hat{u}_A + Z_D \hat{u}_D + e$$  \[3\]

where $y$ is the phenotype vector, $1_n$ is an $n$-length vector of ones, $\mu$ stands for the grand mean, and $Z_A$ and $Z_D$ are the design matrices for the vectors of random additive and dominance genetic effects of $n$ single crosses. The genetic effects have the covariance structures $\hat{u}_A \sim N(0, G_A \sigma^2_A)$ and $\hat{u}_D \sim N(0, G_D \sigma^2_D)$, where $\sigma^2_A$ and $\sigma^2_D$ are the additive and dominance genetic variances, and $G_A$ and $G_D$ represent the $n \times n$ additive and dominance genomic relationship matrices, respectively. $G_A$ was calculated following VanRaden (2008) as

$$G_A = \frac{WW'}{2\sum_{k=1}^p P_k (1-p_k)}$$

where $W$ is an $n \times m$ matrix with $w_{ij} = x_j - 2p$, and $x_j$ is marker genotype of the $j$th single cross for the $i$th marker, which is coded as 0, 1, or 2 for homozygous major allele, heterozygous, and homozygous minor allele states, respectively. $G_D$ was calculated according to Vitezica et al. (2013) as

$$G_D = \frac{HH'}{2\sum_{k=1}^p P_k (1-p_k)^2}$$

where $H$ is an $n \times m$ matrix with $h_{ij} = -2(1-p)^2$ if the individual is homozygous for major allele, $h_{ij} = 2p(1-p)$ if the individual is heterozygous, and $h_{ij} = -2p^2$ if the individual is homozygote for the minor allele. The mixed model equations to obtain the solutions of additive and dominance effects were as given in Zhao et al. (2013):

$$\begin{bmatrix}
\hat{\mu} \\
\hat{\mu}_A \\
\hat{\mu}_D
\end{bmatrix} =
\begin{bmatrix}
I_n \\
Z_A^{\top}
\end{bmatrix}^{\top}
\begin{bmatrix}
1_n \\
Z_A^n
\end{bmatrix}
\begin{bmatrix}
Z_A^{\top} Z_A + G_A \sigma^2_A \\
G_A \sigma^2_A + G_D \sigma^2_D
\end{bmatrix}^{-1}
\begin{bmatrix}
y \\
Z_A^n y
\end{bmatrix}$$

The summation of estimates of additive genetic effect ($\hat{u}_A$) and dominance genetic effects ($\hat{u}_D$) was considered as the predicted single-cross performance.

The GBLUP model for combining ability-based prediction was as follows:

$$y_{CA} = 1_n \mu + Z_{g_f} \hat{g}_f + Z_{m_m} \hat{g}_m + Z_s + e$$  \[4\]

where $Z_{g_f}$, $Z_{m_m}$, and $Z_s$ are the design matrices for the vectors of random GCA effects of females (i.e., $g_f$), males (i.e., $g_m$) and the SCA effect (i.e., $s$), respectively. The random effects were assumed to be distributed as $g_f \sim N(0, G_{g_f} \sigma^2_{g_f})$, $g_m \sim N(0, G_{g_m} \sigma^2_{g_m})$, and $s \sim N(0, S \sigma^2_s)$ where $\sigma^2_{g_f}$, $\sigma^2_{g_m}$, and $\sigma^2_s$ are the variances of GCA effects of females, GCA effects of males, and SCA effects, respectively; $G_{g_f}$ and $G_{g_m}$ are the additive genomic relationship matrices among female and male parents; and $S$ is a matrix of coefficients.
of covariances among SCA effects. \( \mathbf{G}_f \) and \( \mathbf{G}_m \) were calculated for the female and male parents separately following VanRaden (2008) as described above. The elements of \( \mathbf{S} \) were calculated as the products of the elements of \( \mathbf{G}_f \) and \( \mathbf{G}_m \) corresponding to the parents of the single crosses as described in Bernardo (2010) and Technow et al. (2014). The mixed model equations to obtain \( \mathbf{g}_r \) and \( \mathbf{g}_m \), and \( \mathbf{s} \) were (Bernardo, 2010)

\[
\begin{align*}
&\mathbf{Z}'_1 \mathbf{y} = \mathbf{Z}'_2 \mathbf{Z} \mathbf{g}_r + \mathbf{Z}'_3 \mathbf{g}_m + \mathbf{Z}'_4 \mathbf{s} + \mathbf{Z}'_5 \text{vec} \left( \mathbf{R} \right) \mathbf{b} + \mathbf{e} \\
&\mathbf{b} = \left( \mathbf{K} + \mathbf{I} \right)^{-1} \mathbf{K} \mathbf{g}_r \end{align*}
\]

Both GBLUP models were implemented using ASReml-R software (Butler et al., 2009) to obtain REML of all variance components and solve the mixed linear model equations.

**Reproducing Kernel Hilbert Space**

Gianola et al. (2006) introduced the RKHS model for genomic prediction. In this model, the genomic relationship matrix used in GBLUP is replaced by a kernel matrix, which enables nonlinear regression in a higher-dimensional feature space. The RKHS model can be represented as follows:

\[
y = \mathbf{1}_n \mu + \mathbf{K} \alpha + \mathbf{e}
\]

where \( \mathbf{K} \) is an \( n \times n \) reproducing kernel matrix whose entries are functions of marker genotypes of pairs of individuals, \( \alpha \) is an \( n \times 1 \) vector of regression coefficients, and \( \mathbf{e} \) is a vector of residuals of length \( n \). The distributions of \( \alpha \) and \( \mathbf{e} \) are \( \alpha \sim N(0, \mathbf{K}^{-1} \sigma^2_\alpha) \) and \( \mathbf{e} \sim N(0, \mathbf{I} - \mathbf{K}^{-1} \sigma^2_\mathbf{e}) \). The values of \( \alpha \) are estimated by minimizing the objective function \((y - \mathbf{K}\alpha)'(y - \mathbf{K}\alpha) + \lambda \alpha' \mathbf{K} \alpha\), where \( \lambda \) is a regularization parameter. The solution to minimizing the above function is (Morota and Gianola, 2014)

\[
\hat{\alpha} = \left( \mathbf{K} + \mathbf{I} \right)^{-1} \mathbf{y}
\]

where \( \mathbf{I} \) is an identity matrix. We used a Gaussian kernel with \( \mathbf{K}_y = \exp(-h ||x_i - x_j||^2) \), where \( ||x_i - x_j|| \) is the Euclidean distance between individuals \( i \) and \( j \) normalized to range from 0 to 1, and \( h \) is a bandwidth parameter that controls the rate of decay of \( \mathbf{K}_y \) with increasing Euclidean distance.

The RKHS model was implemented using the R package GBLUP (de los Campos and Perez Rodriguez, 2017). The optimum value for \( h \) was chosen by performing cross validation using the whole dataset over a range of values from \( 10^{-6} \) to 10. The parameters \( \sigma \) and \( \mathbf{b} \) in the \( \mathbf{K} \) are specified using \( \mathbf{b} \) and \( \mathbf{b} \) specifying total number of iterations for Gibbs-sampler and initial number of iterations to be discarded were set at 40,000 and 1000, respectively. All other parameters were specified according to the default settings.

**Support Vector Regression**

Support vector regression is a machine learning algorithm for classification and regression problems. The SVR has been evaluated for genomic prediction by Maenhout et al. (2007), and Long et al. (2011) in plant breeding and Moser et al. (2009) in animal breeding. The SVR is a particular case of RKHS regression where the quadratic error loss function in RKHS is replaced by \( \varepsilon \)-insensitive loss function (González-Recio et al., 2014). The \( \varepsilon \)-insensitive loss function has the following form:

\[
L[|y_i - f(x_i)|] = \begin{cases} 0, & \text{if } |y_i - f(x_i)| < \varepsilon \\ |y_i - f(x_i)| - \varepsilon, & \text{otherwise} \end{cases}
\]

If the errors are \( \leq \varepsilon \), the loss function assigns zero loss. If the errors are \( > \varepsilon \), the loss is equal to the difference between absolute error and \( \varepsilon \). Thus, \( \varepsilon \) determines the number of support vectors used in the regression function. The objective function to be minimized with \( \varepsilon \)-insensitive loss is

\[
C \sum_{i=1}^{n} (\xi_i - \xi_i^*) + \frac{1}{2} \mathbf{w}' \mathbf{w}
\]

where \( \varepsilon \) is a penalty parameter, \( \xi_i \geq y_i - f(x_i); \xi_i^* \geq f(x_i) - y_i - \varepsilon \), and \( \mathbf{w} \) is a vector of unknown weights (i.e., regression coefficients). The minimizing solution to this objective function is given by

\[
\hat{f}(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x)
\]

where \( k() \) is a kernel of choice, and \( \alpha_i \) and \( \alpha_i^* \) are solutions to a nonlinear system of equations (Moser et al., 2009). In our implementation of SVR, we used Gaussian radial basis kernel which has the form \( k(x_i, x_j) = \exp(-\sigma ||x_i - x_j||^2) \), where \( \sigma \) is the bandwidth parameter.

The SVR was implemented using R package kernlab (Zeileis et al., 2004). The specified parameters were ‘eps-svr’ (epsilon regression) as the type and ‘rbfdot’ (radial basis kernel “Gaussian”) as the kernel. The optimal value of three tuning parameters required to solve the SVR regression (i.e., the insensitivity zone \( \varepsilon \), penalty parameter \( C \), and the bandwidth parameter \( \sigma \)) were determined by grid search over a three dimensional parameter space). The grid values for \( \varepsilon \) ranged from \( 10^{-6} \) to 1; grid values of \( C \) and \( \sigma \) ranged from \( 10^{-6} \) to 10. All other parameters were set to the default values.

**Neural Network**

In the field of statistical prediction, a NN is a form of two-step nonlinear regression (Hastie et al., 2002). Gianola et al. (2011) first used NN for genomic prediction. One of the basic and most commonly used forms of NN for genomic prediction is the single hidden layer feed-forward NN. This form consists of an “input layer,” a “hidden layer,” and an “output layer.” Predictions from this form of NN are obtained in two steps. In the first step, inputs are nonlinearly transformed in the hidden layer. This is accomplished by combining the inputs (i.e., coded marker genotypes \( x_i \) with weights \( \beta \)) and an intercept \( b \) in each of \( t \) \( [1, 2, 3, \ldots, s] \) neurons. This is followed by transformation of a linear score \( y_i = b + \sum_{j=1}^{p} x_j \beta_j \) through a nonlinear activation function, \( y_i = g(y_i) \). In the second step, the response variable (i.e., phenotype) is linearly regressed on the data derived predictors, \( y_i \). The output function of NN can be represented as

\[
y_i = \mu + \sum_{i=1}^{p} w_i y_i + \varepsilon_i = \mu + \sum_{i=1}^{p} w_i g_i \left( b_i + \sum_{j=1}^{p} x_j \beta_j \right) + \varepsilon_i
\]
where \( w_t \) is the weight of \( t \)th neuron to the output, \( g(.) \) is the activation function, and \( \varepsilon_t \sim N(0, \sigma^2_u) \).

The NN model was implemented in this study using the R package \textit{brnn} (Pérez-Rodriguez and Gianola, 2013). A tangent hyperbolic activation function, \( g(s) = [\exp(2s) - 1]/[\exp(2s) + 1] \), was used in this implementation. The model was fitted using a genomic relationship matrix rather than a SNP incidence matrix as the predictor (Gianola et al., 2011). The genomic relationship matrices were calculated as described in the GBLUP section. The number of epochs to train the model was set to be 30 except for SCA prediction, where <10 epochs were used. More epochs generated singularities due to less variability in SCA covariance matrix. The optimal number of neurons was determined by cross-validation using the whole dataset. Neuron numbers ranged from one to six in the cross validation. All other parameters were set to the default values.

Cross-Validation to Evaluate Predictive Ability

The predictive ability for genetic value of single crosses was assessed using phenotypic GCA, common-parent single cross, and genome-based prediction. The phenotypic GCA predictions were calculated using parents GCA estimated from the phenotype data only. Genome-based single-cross predictions were calculated by summing the effects estimated from Eq. [4].

“Common-parent single crosses” were considered to be those that shared one parent. The performance of an untested single cross was predicted based on one randomly sampled tested single cross involving either male or female parent in common with the untested single cross. For each untested single cross, one tested single cross representing the female parent and one tested single cross representing the male parent were sampled to form two vectors of tested single crosses. These were used separately to estimate the performance of untested single crosses by correlating the trait values, and the mean correlation was taken as the common-parent-single-cross predictive ability. Random sampling of a tested single cross for each parent was repeated 10 times. It is important to note that correlations between common-parent tested and untested single crosses do not approximate correlations between topcrosses and untested single crosses because of differences in variances between the two types of populations. The denominator of the correlation coefficient between tested and untested single crosses is \( \sqrt{\sigma^2_{T1} \sigma^2_{U}} \), where \( \sigma^2_{T1} \) and \( \sigma^2_{U} \) are the variances between tested and untested single crosses, respectively. These variances can be considered equal if the single crosses are sampled from the same population. The denominator of the correlation coefficient between untested single crosses and topcross progeny is \( \sqrt{\sigma^2_{U} \sigma^2_{T1}} \), with \( \sigma^2_{U} \) expected to be greater than \( \sigma^2_{T1} \) because the variance in male GCA effects would not contribute to \( \sigma^2_{T1} \). Therefore, it can be seen that “common-parent single cross predictive ability” is expected to be lower than that of a topcross test. An expectation for \( \sigma^2_{U} \) was not derived because this quantity is highly dependent on the chosen tester, and testers are not randomly sampled from a population of male lines but rather carefully selected based on their performance as male parents and testers in practical hybrid breeding program.

To further compare genome and phenotypic GCA prediction, four types of single crosses depending on the tested or untested parents were distinguished for cross-validations. These single crosses were named as T2 = both parents tested in other single-cross combinations, TIF = only female parent tested in other single-cross combinations, TIM = only male parent tested in other single-cross combinations, and T0 = neither parent tested in other single-cross combinations. Also, the notation T1 was used to denote singles crosses having either male or female parent tested.

Leave-one-out cross-validations were performed to evaluate phenotypic GCA, common-parent single cross, and genome-based predictive abilities. For all cross-validations, predictive abilities were estimated as the Pearson’s correlation coefficient between the observed and predicted single-cross performance. The SEs of the predictive abilities were estimated using the bootstrap method implemented in the R package \textit{boot} (Canty and Ripley, 2012). The number of bootstrap samples used was 200.

RESULTS

Variance Components

The variance among single crosses (\( \sigma^2_{S} \)), GCA variances of females (\( \sigma^2_{GCA,F} \)) and males (\( \sigma^2_{GCA,M} \)), and SCA variances (\( \sigma^2_{SCA} \)) for GY and PH were all significantly different from zero in both datasets (Tables 3 and 4). As expected, variance in GCA was far more important than variance in SCA, but SCA variance was still 12 to 14% of the total GCA variance for GY, and 4 to 9% for PH (Table 4). These proportions were similar to Technow et al. (2014) and Massman et al. (2013).

Comparisons of Phenotypic GCA, Common-Parent Single Cross-, and Genome-Based Predictive Abilities

We initially explored single-cross predictive abilities based on phenotypic GCA, common-parent single cross performance, and genome-based prediction using a GBLUP model. A total of 192 (89 females and 103 males) and 217 (46 females and 171 males) single crosses are required for common-parent single cross prediction in Datasets I and II, respectively. To compare the common-parent single cross-based prediction with phenotypic GCA- and genome-based prediction, we calculated the single-cross predictive abilities of the latter two approaches with training set sizes of 192 and 217 in Datasets I and II, respectively, to eliminate confounding effects of population size. The highest predictive abilities were obtained with genome-based prediction followed by phenotypic GCA prediction, whereas the lowest predictive abilities were found using the common-parent single crosses (Table 5). The genome-based approach outperformed the phenotypic GCA approach, amounting to a 9% improvement for GY on average across the two datasets (Table 5).
Phenotypic GCA and Genome-Based Predictive Abilities for T2, T1, and T0 Single Crosses

We further investigated the phenotypic GCA and genome-based predictive abilities specifically for T2, T1, and T0 single crosses. Genome-based predictive abilities considerably differed depending on the number of tested parents of single crosses (Fig. 1). The predictive abilities for T2 single crosses were highest, followed by T1F, T1M, and T0 for both traits in each dataset. The mean genome-based predictive abilities for T2 single crosses were 85 and 69% of the mean predictive abilities for T2 single crosses for GY. Similarly, for PH, the mean genome-based predictive abilities for T1 and T0 single crosses were 87 and 73% of mean predictive abilities for T2 single crosses.

The genome and phenotypic GCA predictive abilities were nearly identical for T2 single crosses. The genome-based prediction abilities, however, were higher than phenotypic GCA prediction when one parent was untested. For T1 single crosses, the mean genome-based predictive abilities were 27 and 32% better than phenotypic GCA predictive abilities for GY and PH, respectively. Moreover, the mean genome-based predictive abilities for T0 single crosses were 0.45 for GY and 0.62 for PH. Single crosses in the T0 set cannot be predicted based on phenotype data because phenotype data are not available for either parent.

Influence of Tuning Parameters on Predictive Abilities of Nonparametric Models

Ten-fold cross-validations in replicates of five were performed separately for GY and PH in each dataset to find optimum values of tuning parameters for each nonparametric model. The sensitivity of these models to the corresponding tuning parameter values was explored. Based on a preliminary analysis, appropriately spaced grid values were selected and a cross-validation-based grid search was performed on the whole dataset. Predictive abilities of the three models varied considerably over the range of tuning parameter values investigated (Fig. 2). However, maximum predictive ability (±0.01) was observed over a broad range of tuning parameters (RKHS and NN) or with many different combinations of tuning parameters (SVR). In the case of RKHS, the

Table 3. Mean, range, variance components and heritabilities for grain yield (GY) and plant height (PH) in Dataset I and Dataset II.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Grain yield</th>
<th>Plant height</th>
<th>Grain yield</th>
<th>Plant height</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dataset I</td>
<td></td>
<td>Dataset II</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>Mg ha⁻¹</td>
<td>cm</td>
<td>Mg ha⁻¹</td>
<td>cm</td>
</tr>
<tr>
<td>Range</td>
<td>10.97</td>
<td>243.48</td>
<td>8.67</td>
<td>210.1</td>
</tr>
<tr>
<td>σ²g†</td>
<td>8.95–13.24</td>
<td>203.25–277.89</td>
<td>7.14–10.2</td>
<td>191–231</td>
</tr>
<tr>
<td>H²§</td>
<td>0.71 ± 0.06</td>
<td>158.48 ± 10.81</td>
<td>0.50 ± 0.07</td>
<td>118.08 ± 10.61</td>
</tr>
<tr>
<td>s²g</td>
<td>0.76 ± 0.02</td>
<td>0.95 ± 0.004</td>
<td>0.58 ± 0.04</td>
<td>0.89 ± 0.01</td>
</tr>
</tbody>
</table>

† σ²g, variance among single crosses.
‡ Significant at α = 0.001.
§ H², entry-mean heritability.

Table 4. General combining ability variance of females (σ²f), males (σ²m), and specific combining ability variance (σ²s) of single crosses between them in Dataset I and Dataset II.

<table>
<thead>
<tr>
<th>Variance components</th>
<th>Grain yield</th>
<th>Plant height</th>
<th>Grain yield</th>
<th>Plant height</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dataset I</td>
<td></td>
<td>Dataset II</td>
<td></td>
</tr>
<tr>
<td>σ²f</td>
<td>Mg ha⁻¹</td>
<td>cm</td>
<td>Mg ha⁻¹</td>
<td>cm</td>
</tr>
<tr>
<td></td>
<td>0.20 ± 0.05</td>
<td>50.16 ± 9.70</td>
<td>0.22 ± 0.07</td>
<td>28.66 ± 7.74</td>
</tr>
<tr>
<td>σ²m</td>
<td>0.09 ± 0.03</td>
<td>51.86 ± 9.15</td>
<td>0.20 ± 0.06</td>
<td>34.48 ± 5.92</td>
</tr>
<tr>
<td>σ²s</td>
<td>0.04 ± 0.01</td>
<td>9.53 ± 1.00</td>
<td>0.05 ± 0.01</td>
<td>2.6 ± 0.77</td>
</tr>
<tr>
<td>σ²f/(σ²g + σ²s)</td>
<td>0.14</td>
<td>0.09</td>
<td>0.12</td>
<td>0.04</td>
</tr>
</tbody>
</table>

† Significant at α = 0.001.

Table 5. Phenotypic general combining ability (GCA), common-parent single-cross, and genome-based predictive abilities (r) and their SEs for grain yield (GY) and plant height (PH) in Datasets I and II.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Trait</th>
<th>Phenotypic GCA prediction</th>
<th>Common-parent single-cross prediction</th>
<th>Genome-based prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>SE</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Dataset I</td>
<td>GY</td>
<td>0.66</td>
<td>0.02</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>PH</td>
<td>0.71</td>
<td>0.02</td>
<td>0.38</td>
</tr>
<tr>
<td>Dataset II</td>
<td>GY</td>
<td>0.48</td>
<td>0.04</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>PH</td>
<td>0.75</td>
<td>0.03</td>
<td>0.39</td>
</tr>
</tbody>
</table>
predictive ability for both GY and PH in Datasets I and II was maximized between $h$ values 0.005 and 2 (Fig. 2). Similarly, in the case of SVR, many different combinations of $\varepsilon$, $C$, and $\sigma$ values provided maximum predictive ability. However, it appears that the combinations of three parameter values providing maximum predictive abilities were specific to dataset and trait. The predictive ability of NN varied only slightly over different $s$ values from 1 to 6, with a maximum difference of 0.04 across traits and datasets. The optimum values of different tuning parameters for GY and PH in each dataset used for further analysis are provided in Table 6. We also compared the predictive abilities of the three models with the optimum values of tuning parameters against the default values provided by the software packages used for implementing these models (Table 6). Optimum tuning parameters improved the predictive abilities of SVR and NN, but many times the improvement was only very slight. The optimum and default $h$ values were equal (0.5) for RKHS.

**Comparison of GBLUP and Nonparametric Models**

The nonparametric models evaluated in this study provided no advantage in predictive ability of single-cross performance. This was true for models built on single-cross performance, as well as those built on combining ability (Fig. 1). The predictive ability of GBLUP, RKHS, and SVR were comparable, whereas NN provided lower predictive abilities. The difference in predictive ability between GBLUP and the three nonparametric methods was minimal regardless of whether predictions were made in the T2, T1, or T0 scenarios. The predictive abilities of RKHS, SVR, and NN were generally the same between GCA-only and GCA + SCA models, except in some cases (e.g., GY, Dataset II, T0) where the GCA + SCA model actually resulted in slightly lower predictive ability.

**DISCUSSION**

In the early stages of hybrid breeding, RILs or DHLs are generated from several biparental families for testing in hybrid combinations. As the number of possible single crosses is too large to evaluate at this stage, initial selection of lines for hybrid performance is traditionally performed based on a topcross test using a single inbred tester from the opposite heterotic group, with evaluation of specific hybrid combinations occurring in more advanced stages (Hallauer and Miranda, 1988). Although the topcross test is easy compared with making and evaluating many pairwise crosses, the additional generations of topcross testing can increase the time required for commercial hybrid development. There is also a possibility of losing some unique potential single crosses due to discarding of lines in the early
stages based only on topcross test data. Therefore, it would be desirable to evaluate the lines based on predicted single-cross performance in the early stages. In this study, we initially evaluated three approaches for early-stage single-cross prediction: (i) phenotypic GCA-based prediction, (ii) common-parent single-cross prediction, and (iii) genome-based prediction. Phenotypic GCA and genome-based predictive abilities were substantially higher than those from common-parent single crosses. Genome-based prediction provided an advantage over phenotypic GCA prediction for T2, T1F, and T1M single crosses. Moreover, the genome-based approach predicted T0 single crosses with moderate to high accuracy for which the phenotypic GCA approach has no ability to predict.

Table 6. Predictive abilities of three nonparametric models obtained with optimum values of tuning parameters selected through cross validation and default values provided by respective software packages used to implement these models. Optimum and default tuning parameter values are shown in parentheses for each model. The tuning parameters optimized were the bandwidth parameter \((h)\) for reproducing kernel Hilbert spaces (RKHS); the complexity parameter \((C)\), \(\varepsilon\)-insensitivity, and bandwidth parameter \((s)\) for support vector regression (SVR); and number of neurons \((s)\) for neural networks (NN).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Trait</th>
<th>RKHS ((h))</th>
<th>SVR ((C, \varepsilon, \sigma))</th>
<th>NN ((s))</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>GY</td>
<td>0.78 (0.5)</td>
<td>0.78 (5, 0.005, 0.00005)</td>
<td>0.76 (4)</td>
</tr>
<tr>
<td></td>
<td>PH</td>
<td>0.86 (0.5)</td>
<td>0.86 (10, 0.05, 0.000005)</td>
<td>0.84 (1)</td>
</tr>
<tr>
<td>II</td>
<td>GY</td>
<td>0.56 (0.5)</td>
<td>0.58 (5, 0.05, 0.0001)</td>
<td>0.51 (1)</td>
</tr>
<tr>
<td></td>
<td>PH</td>
<td>0.82 (0.5)</td>
<td>0.83 (10, 0.0005, 0.0001)</td>
<td>0.79 (1, 0.0002)</td>
</tr>
</tbody>
</table>
Table 7. Correlation between observed and predicted general (GCA) and specific combining ability (SCA) effects with genomic best linear unbiased prediction (GBLUP) and three nonparametric methods reproducing kernel Hilbert spaces (RKHS), support vector regression (SVR), and neural networks (NN) for grain yield.

<table>
<thead>
<tr>
<th>Hybrid type</th>
<th>GBLUP GCA</th>
<th>GBLUP SCA</th>
<th>RKHS GCA</th>
<th>RKHS SCA</th>
<th>SVR GCA</th>
<th>SVR SCA</th>
<th>NN GCA</th>
<th>NN SCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>0.931</td>
<td>0.125</td>
<td>0.896</td>
<td>-0.131</td>
<td>0.961</td>
<td>-0.070</td>
<td>0.916</td>
<td>-0.164</td>
</tr>
<tr>
<td>T1F</td>
<td>0.855</td>
<td>0.021</td>
<td>0.824</td>
<td>-0.022</td>
<td>0.865</td>
<td>-0.033</td>
<td>0.786</td>
<td>-0.137</td>
</tr>
<tr>
<td>T1M</td>
<td>0.822</td>
<td>0.154</td>
<td>0.776</td>
<td>-0.127</td>
<td>0.829</td>
<td>-0.075</td>
<td>0.781</td>
<td>-0.103</td>
</tr>
<tr>
<td>T0</td>
<td>0.737</td>
<td>0.048</td>
<td>0.704</td>
<td>-0.093</td>
<td>0.720</td>
<td>-0.033</td>
<td>0.656</td>
<td>-0.107</td>
</tr>
</tbody>
</table>

Supplemental Material Available
Supplemental material is available online for this article.
Conflict of Interest
The authors declare that there is no conflict of interest.

Acknowledgments
This research was partially supported by a grants from the Nebraska Corn Board and National Research Initiative or Agriculture and Food Research Initiative Competitive Grants Program (Grant no. 2012–67013–19460) from the USDA National Institute of Food and Agriculture. D.C. Kadam was supported by an Indian Council of Agricultural Research International Fellowship. We are indebted to Ronal Haarman of Dupont Pioneer for contributing phenotypic data to Dataset I, as well as to Rex Bernardo of the University of Minnesota for helpful suggestions that enhanced the accuracy and clarity of the final manuscript.

References
Bernardo, R. 2010. Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN.


