Unlocking Diversity in Germplasm Collections via Genomic Selection: A Case Study Based on Quantitative Adult Plant Resistance to Stripe Rust in Spring Wheat

Kebede T. Muleta, Peter Bulli, Zhiwu Zhang, Xianming Chen, and Michael Pumphrey*

Abstract
Harnessing diversity from germplasm collections is more feasible today because of the development of lower-cost and higher-throughput genotyping methods. However, the cost of phenotyping is still generally high, so efficient methods of sampling and exploiting useful diversity are needed. Genomic selection (GS) has the potential to enhance the use of desirable genetic variation in germplasm collections through predicting the genomic estimated breeding values (GEBVs) for all traits that have been measured. Here, we evaluated the effects of various scenarios of population genetic properties and marker density on the accuracy of GEBVs in the context of applying GS for wheat (Triticum aestivum L.) germplasm use. Empirical data for adult plant resistance to stripe rust (Puccinia striiformis f. sp. tritici) collected on 1,163 spring wheat accessions and genotypic data based on the wheat 9K single nucleotide polymorphism (SNP) iSelect assay were used for various genomic prediction tests. Unsurprisingly, the results of the cross-validation tests demonstrated that prediction accuracy increased with an increase in training population size and marker density. It was evident that using all the available markers (5,619) was unnecessary for capturing the trait variation in the germplasm collection, with no further gain in prediction accuracy beyond 1 SNP per 3.2 cM (~1,850 markers), which is close to the linkage disequilibrium decay rate in this population. Collectively, our results suggest that larger germplasm collections may be efficiently sampled via lower-density genotyping methods, whereas genetic relationships between the training and validation populations remain critical when exploiting GS to select from germplasm collections.

Core Ideas
- Larger germplasm collections may be efficiently sampled via lower-density genotyping methods.
- The selection population needs to be well represented in the training set.
- Good prospects for applying genomic selection to efficiently unlock the potential of plant genetic resources exist.

Several recent advances in genotyping technology and statistical modeling have enabled the development of low-cost efficient selection methods for complex traits (Tanksley, 1993; Dwivedi et al., 2007; Jannink et al., 2010; Le Gouis et al., 2012; Poland et al., 2012; Bentley et al., 2014). Genomic selection, which performs whole-genome scans and selects for large numbers of both minor and large effect quantitative trait loci (QTL) without prior knowledge of their genomic location (Meuwissen et al., 2001).
The potential for using GS depends on the accuracy of predicting GEBVs. With higher accuracy of predicting GEBVs, GS can increase the frequency of target genotypes in a population and accelerate gain from selection (Pérez-Cabal et al., 2012; Desta and Ortiz, 2014). Prediction accuracy, measured as a correlation between the GEBVs and the observed phenotypes of the validation population, is affected by several factors. In particular, the presence of a strong population structure can significantly affect the accuracy of predicting GEBVs, as marker effects estimated within a training population can have low predictive accuracy when they are used to estimate GEBVs in a genetically distant validation population (Schulz-Streeck et al., 2012; Zhao et al., 2012). In breeding populations, the genetic distance between the training and testing populations tends to diverge and, consequently, the accuracy of predicting GEBVs decreases over time (Jannink et al., 2010), indicating the need for proper assessment of the genetic relationships between the training and validation populations (Asoro et al., 2011; Sallam et al., 2015). Previous simulation and empirical studies have indicated that prediction accuracy generally improves with a larger training population size (Heffner et al., 2011a; Nakaya and Isobe, 2012). Similarly, higher marker densities, to a point, ensure the presence of strong and robust marker–trait associations and hence improve prediction accuracy. The training population size and marker density required for optimal prediction accuracy depend on the level of genetic diversity and the extent of linkage disequilibrium (LD) in a population and the genetic architecture of the target trait.

Although the potential for increasing genetic gain in wheat through GS has been recognized (Heffner et al., 2010, 2011b; Rutkoski et al., 2010; Ornella et al., 2012), there is still limited information on the application of GS for exploiting the genetic diversity of complex traits in large diverse populations. Moderate to high prediction accuracies, ranging from 0.3 to 0.8, have been achieved when using relatively low marker densities and small population sizes (Daetwyler et al., 2014). To date, most of the GS studies in wheat have been based on simulation and cross-validation studies using biparental and multi-family breeding populations (Crossa et al., 2010; Heffner et al., 2010, 2011b). Thus the relatively higher prediction accuracies with lower marker densities is probably caused by the lower genetic diversity resulting from the narrow genetic differences between parental lines and the associated long-range LD in the biparental populations. The level of prediction accuracy that can be achieved in pedigree breeding populations with low marker densities and smaller training population sizes may not be attainable with large collections of diverse accessions because of extensive LD decay and the wide range of genetic variation. Assessing the accuracy of GEBVs in a more representative context of population genetic features such as a larger sample size and broader genetic diversity is therefore essential for exploring the potential of using GS in wheat improvement.

The USDA National Small Grains Collection wheat germplasm consists of 42,138 T. aestivum accessions from all representative wheat-producing areas of the world (Bonman et al., 2015). A core subset of the germplasm collection has been genotyped and is a good reference population for performing genomic predictions and association studies for important traits. Selection of accessions based on GEBVs for numerous traits of importance to specific breeding programs should enhance the efficient use of the available genetic resources of wheat. Although genotyping of core collections has been applied in many crop species, genotyping entire collections is the next likely step because of the rapidly reducing cost of genotyping. Given that national and international seed banks contain hundreds of thousands of accessions, we believe efficient genotyping and sampling strategies are needed. In this study, we determined the optimum marker density required for GS for resistance to stripe rust across a range of wheat training population sizes with varying degrees of genetic relationships, and investigated the effect of genetic distance between the training and validation populations on the prediction accuracy of GS.

The wheat rusts (Puccinia spp.) are among the most damaging diseases of wheat and have caused massive losses to wheat production globally (Chen, 2005; Kolmer, 2005; Ellis et al., 2014; Hulbert and Pumphrey, 2014). Commensurate with their economic importance, the wheat–rust pathosystems are one of the best characterized systems of plant–pathogen interactions, where monogenic, oligogenic, and polygenic trait architectures are common. According to the 2013 Catalogue of Gene Symbols for Wheat (McIntosh et al., 2013) and the 2013–2014 Supplement (http://wheat.pw.usda.gov/GG2/Triticum/wgc/2013-2014_Supplement.pdf, accessed 4 July 2017), there are 67, 58, and 74 officially named stripe rust (Puccinia striiformis), stem rust (Puccinia graminis), and leaf rust (Puccinia triticina) resistance genes. The vast majority of the officially named wheat rust resistance genes are based on control by a single (monogenic) or a few (oligogenic) resistance loci, although a few of them are based on truly polygenic quantitative trait loci.
Recent association genetics studies have indicated that polygenic resistance to wheat rust diseases is paramount (Bajgain et al., 2015; Maccaferri et al., 2015; Bulli et al., 2016; Aoun et al., 2016). Depending on the race composition in a field, environmental conditions for disease development, and the specific trait being measured (assessment method), the wheat rust pathosystem may provide the opportunity to study the effects of trait genetic architecture and assessment methods on GS prediction accuracy. Accordingly, we assessed the effect of trait architecture as affected by the difference in race composition between the environments and trait assessment methods, comparing a semiquantitative measure of disease reaction using a 0–9 scoring scale with a strictly quantitative method based on visual estimates of the percentage of infected leaf area (disease severity).

Materials and Methods
Wheat Germplasm Panel and Phenotypic Trait Evaluation
A set of 1163 hexaploid spring wheat accessions from the wheat core collection of the USDA-ARS National Small Grains and Potato Germplasm Research Unit, Aberdeen, ID, was used in this study. The accessions were landraces, cultivated lines, breeding lines, and cultivars originating from 91 countries representing diverse geographic regions of the world. For quality control, accessions with more than 10% missing genotypic data were excluded from the final analyses. Genetically identical accessions were also identified on the basis of kinship analysis and were represented by only one individual in the subsequent analyses. Accordingly, a total of 959 nonduplicate accessions were used for the genome-wide association study (GWAS) and GS analyses.

The panel was evaluated for response to stripe rust infection in five field experiments at two locations in the Pacific Northwest region of the United States. The field experiments were conducted at Mount Vernon (MTV, 48°25’12”N; 122°19’34”W) during the 2012–2014 crop seasons and at Pullman (46°43’59”N; 117°10’00”W) during the 2012 and 2014 crop seasons. The nursery locations are subject to high disease pressure on an annual basis but vary in stripe rust populations and weather patterns. Mount Vernon is a cooler and high rainfall area located west of the Cascade Mountain range, which is conducive for a year-round occurrence of stripe rust inoculation and infection. Pullman is a semiarid wheat belt area located east of the Cascade Mountain range with the greatest number of races and races with the widest range of virulence (Chen, 2005). The accessions were planted as single rows in nonreplicated trials. The highly susceptible cultivar Avocet S was planted every 20 entries and on each side of the plot to ensure uniform disease pressure across the experimental plots. Infection type (IT) was estimated by using a 0 to 9 scale (Line and Qayoum, 1992), and disease severity was recorded as the percentage of leaf area showing disease symptoms.

Single Nucleotide Polymorphism Genotyping
Genotyping of the 1163 spring wheat accessions was performed with the Illumina iSelect 9K wheat assay (Illumina Inc., San Diego, CA) at the USDA-ARS Genotyping Laboratory, Fargo, ND. Genotype calling of the original Illumina data was performed using the computer software GenomeStudio version 2011.1 (Illumina Inc.) to optimize the SNP call rates for misclassification and ambiguous clustering. After removing SNPs with low-quality clustering and minor allele frequencies less than 5%, a total of 5619 high-quality SNPs with positions on the wheat 9K SNP consensus map (Cavanagh et al., 2013) were retained for GS and GWAS analyses.

Population Structure, Kinship, and LD Analyses
Population structure and kinship relationships among the spring wheat accessions were determined to perform a cluster-based grouping of the accessions and ultimately examine the effect of population structure and genetic diversity on the prediction accuracy of GEBVs. The Bayesian model-based clustering algorithm implemented in STRUCTURE software (version 2.2.3; Pritchard et al., 2000) was used to investigate the population structure of the 1163 accessions with a set of 425 SNP markers distributed across the entire wheat genome with an intermarker distance of >10 cM apart. A burn-in of 50,000 iterations and 100,000 Monte Carlo Markov chain replicates were used to determine the K values (number of subpopulations) in the range of 1 to 10, with five independent runs for each K. The Evanno method (Evanno et al., 2005) was used to determine the likely number of subpopulations via the STRUCTURE HARVESTER program (Earl and VonHoldt, 2012). Principal component analysis (PCA) was also performed by using JMP GENOMICS version 6.1 software (SAS Institute Inc., Cary, NC) to further analyze the population structure. A marker-based kinship matrix for the 959 accessions was also generated based on the Fast Ward identity-by-descent clustering algorithm implemented in JMP Genomics version 6.0.

Linkage disequilibrium between pairs of SNP markers were estimated by using JMP GENOMICS version 6.1. Linkage disequilibrium was estimated as the squared allele frequency correlations ($r^2$) between pairs of intrachromosomal SNPs with known chromosomal positions and minor allele frequencies of ≥0.05. To determine the average pattern of genome-wide LD decay over genetic distance, a scatterplot of $r^2$ values against the corresponding genetic distance between markers was constructed. The second-degree locally weighted polynomial regression-based curve was fitted to estimate the extent of LD decay (Cleveland, 1979). In addition, the critical $r^2$ value that indicated the demarcation beyond which LD was caused by true physical linkage was determined by taking the 95th percentile of the square root of transformed $r^2$ data of unlinked markers (Breseghello and Sorrells, 2006).
Phenotypic Data Analysis

Stripe rust IT and severity from the field experiments were subjected to ANOVAs via the SAS Mixed Procedure (SAS Institute Inc., Cary, NC). The variance components were calculated via the restricted maximum likelihood method (Corbeil & Searle 1976), in which all effects except the mean were considered to be random. Estimates of the variance components were calculated for each location and across all locations with datasets from the multiple seasons as replicates. Variance components were estimated according to the following model:

\[ y_{ij} = \mu + g_i + e_j + g_e_{ij} + r_{ij}, \]  

where \( y_{ij} \) is the observation for accession \( i \) at environment \( j \), \( \mu \) is the overall mean, \( g_i \) is the effect of the accession \( i \), \( e_j \) is the effect of environment \( j \), \( g_e_{ij} \) is the interaction between accession \( i \) and environment \( j \), and \( r_{ij} \) is the residual. Broad-sense heritability (\( H^2 \)) was calculated according to the following formula:

\[ H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2 + \sigma_{G\times E}^2}, \]

where \( \sigma_G^2 \) is the genotypic variance, \( \sigma_E^2 \) is the environment variance, \( \sigma_{G\times E}^2 \) is the genotype \( \times \) environment interaction variance, \( \sigma_{err}^2 \) is the residual error variance, and \( y \) is the number environments. Genotype-adjusted means of IT and severity at each location were computed from the best linear unbiased predictors (BLUPs). The normality of the BLUP values were examined by constructing the normal quantile plots. The BLUP values of IT and severity plotted against the theoretical quantiles approximated the diagonal straight line, indicating that the residuals were reasonably normally distributed. Hence, the BLUPs of the original data were used as the phenotype-based estimates of breeding values for the subsequent model training and validation tests in GS.

Prediction of GEBVs and Assessment of Accuracy

Assessing the Effect of Training Population Size and Composition

The effect of training population sizes on the accuracy of GEBVs for quantitative adult plant resistance to stripe rust were assessed. In addition to all 959 accessions in the spring wheat panel, three additional training population sizes (NT) of 210, 478, and 640 randomly selected accessions were also sampled to assess the effect of training population size on prediction accuracy. The names of the 959 accessions, their improvement status, and countries of origin are listed in Supplemental Table S1. Cross-validation tests were performed for each of the four populations to estimate the genome-wide marker effects and predict the breeding values. For each of four population sizes, cross-validation tests were repeated 500 times and the mean of each prediction accuracy across the 500 repeats was reported.

We assessed the effect of training population composition in relation to the validation population on prediction accuracies in separate cross-validation tests. Population structure analyses (both the model-based Bayesian clustering algorithm and PCA) as well as the hierarchical Fast Ward clustering algorithm revealed the presence of two subpopulation (SP) groups (SP1 and SP2) in the panel of 959 accessions. Thus, subset (within clusters) and interset (across clusters) cross-validation tests were performed to assess the effect of genetic distance or relatedness between individuals in the training set (TS) and validation set (VS). Five scenarios of the cross-validation design were performed. In the first and second scenarios, a subset validation was performed within SP1 and SP2. In the third scenario, SP1 and SP2 were mixed and completely randomized to ensure random assignment of genetically related individuals across the TS and VS. The accuracies of GEBVs were then estimated under each scenario. Finally, interset validations were performed in which SP1 was used as TS and SP2 as VS, and vice versa in the fourth and fifth scenarios, respectively.

Genomic Selection with Marker Panels of Different Densities

To determine the marker density required for obtaining the maximum attainable prediction accuracy (i.e., the marker density beyond which no or little gain in prediction accuracy is reached), we performed cross-validation tests using four marker panels selected from the 5619 SNPs to represent different marker densities. The four marker densities were constructed by systematically removing markers from the original genotypic data on the basis of their genetic position in the 21 wheat chromosomes (Cavanagh et al., 2013). The four levels of marker densities include the total marker set (M\(_t\)) in which all 5619 SNPs were included, and three marker densities constructed by removing SNPs that are located at ≤1.0, ≤5.0, and ≤10.0 cM from the preceding adjacent marker, designated as M\(_{1cM}\), M\(_{5cM}\), and M\(_{10cM}\), respectively. A total of 1849, 543, and 322 SNP markers were retained for M\(_{1cM}\), M\(_{5cM}\), and M\(_{10cM}\), with an average of 1 SNP per 3.2, 9.4, and 14.8 cM, respectively. The marker density (all 5619 SNPs) covered a total of 4645 cM of the wheat genome, with an average linkage map size of 221.2 cM and average distribution of 1 SNP per 1.3 cM based on the consensus map developed by Cavanagh et al. (2013). Table 1 summarizes the marker distribution across the 21 wheat chromosomes for the four marker densities in terms of the average number of markers and the intermarker distance between adjacent markers.

Association Analyses and the Joint GWAS–GS Strategy

Marker–trait association analysis was performed by using the compressed mixed linear model method (Yu et al., 2006; Zhang et al., 2010) implemented in the
The objective of the marker–trait association analysis in this study was to select a low-density panel of markers in strong association with stripe rust resistance and use them in a cross-validation test. In addition, the prediction accuracy of the selected low-density panel of markers was compared with that of MT and the other three marker panels representing different densities. The GWAS was performed without the validation population to test for associations between markers and the trait. We then performed cross-validation test using only the significant markers to predict trait values. Single nucleotide polymorphism markers significantly associated with stripe rust IT or severity at a nominal probability ($P < 0.05$) in at least two of the five environments were used to represent low-density marker panel.

### Genomic Prediction Model and Cross-Validation Analysis

Ridge-regression–best linear unbiased prediction (RR-BLUP) was used to estimate genome-wide marker effects and predict GEBVs. We used rr-BLUP version 3.8 (Endelman, 2011) implemented in the R package. In RR-BLUP, SNP effects are assumed to be random and can be modeled as:

$$y = \mu + \sum_{k=1}^{p} X_{ik}\beta_k + e_i$$  \[3\]

where $y_i$ is the vector of phenotypes for individual $i$, $\mu$ is the overall mean, $X_{ik}$ is the genotype of individual $i$ for SNP marker $k$, $\beta_k$ is the additive effect of SNP marker $k$, and $e_i$ is the residual effect. The breeding values (GEBVs) were predicted via the formula $y_v = MB_v$, where $y_v$ is an $N_v \times 1$ vector of predicted trait values (GEBVs) for the accessions in the validation set, $M$ is an $N_v \times N_M$ matrix.
of genotype indicators for the validation set, $\beta$ is an $N_M \times 1$ vector of RR-BLUP marker effects.

Cross-validation analysis was performed to evaluate the accuracy of the predicted breeding values compared with the observed phenotypes of the individuals in the VS. Average correlation between marker-predicted genotypic value (i.e., GEBV) and the observed performance of accessions in the VS were used to calculate the accuracy of genome-wide prediction. Each of the four populations was split into the TS and VS, in which three-fourths of the individuals were assigned to TS and one-fourth to VS to determine accuracy of the GEBVs. For each of the four $N_r$ scenarios, phenotypic data on stripe rust were used to calculate the accuracy of the predicted breeding values compared to the VS to determine accuracy of the GEBVs. For each of the four $N_r$ scenarios, phenotypic data on stripe rust (as BLUP values) and genotypic data were available for the TS to estimate SNP effects, whereas phenotypic data were masked in the VS and marker genotypes were used to predict the GEBVs of the accessions in the VS on the basis of the model trained with the TS.

Results

Estimates of Variance Components and Trait Heritability

The accessions were evaluated for response to stripe rust in multi-environment field trials between 2012 and 2014. Mean stripe rust responses, estimates of variance components, and broad sense heritabilities are presented in Table 2. Levels of stripe rust infection were high in all of the five testing environments, allowing for an unambiguous assessment of the field reaction of the accessions to stripe rust. Based on the BLUP values of the stripe rust response data, 18% of the accessions displayed a high level of resistance (IT = 0–3) and 24% of the accessions displayed susceptible reactions (IT = 7–9) across all environments (Supplemental Fig. S1). The mixed-model ANOVA components revealed highly significant ($P < 0.0001$) differences among the genotypes and genotype × environment interactions for severity and IT, whereas the variance components for environment were not significant across all analyses. High heritability ($H^2$) values, ranging from 0.8 to 0.9, were observed for both IT and severity in all locations.

Population Structure and Genetic Diversity

Both STRUCTURE and PCA analyses indicated structuring of the germplasm panel into two major clusters (Fig. 1A, Supplemental Fig. S4). The first and second principal components explained 47.5 and 12.2% genetic variation in the germplasm panel, respectively. Further characterization of the accessions using the Ward hierarchical clustering analysis revealed a grouping pattern similar to those of PCA and structure analysis (Fig. 1B). The analyses of population structure and Ward clustering also indicated the grouping pattern of the accessions based on geographic origin and their improvement status. A major division between accessions from Asia (SP1) and accessions from Europe, North America, and South America (SP2) was evident. Kinship analysis also revealed that accessions in SP1 were more genetically related to each other (an average kinship coefficient of 0.18) than accessions in SP2, which had distant genetic relationships (an average kinship coefficient of 0.13). The two distinct clusters of the spring wheat germplasm (SP1 and SP2) were used to construct training and validation populations to assess the effect of population structure and genetic relatedness on GS prediction accuracy.

Clustering related to stripe rust resistance was assessed by performing correlation analyses between the population structure and BLUP values of IT and severity. The analyses revealed a significant correlation between population subclustering and response to stripe rust resistance ($P < 0.001$; $r$ ranging from 0.11 to 0.30, data not shown). Accordingly, accessions in SP1 displayed a higher proportion of susceptible to highly susceptible reactions to stripe rust. Conversely, accessions in SP2 showed uniformly distributed stripe rust IT and severity ratings.

Linkage Disequilibrium

The $r^2$ values among the 21 chromosomes averaged 0.55 for the completely linked marker pairs. The value of $r^2$ decreased to 0.3 among the incompletely linked marker pairs with a genetic distance of <1.0 cM, indicating a decrease in $r^2$ value to approximately half of its initial value within a 1.0-cM distance. Based on the fitted non-linear model, genome-wide LD was predicted to decay below the critical $r^2$ of 0.36 at an intermarker distance of 2.5 cM. Among the marker pairs that showed true physical linkage ($r^2 > 0.36$), chromosome 2B contained the highest percentage (17.2%) of these marker pairs, whereas chromosome 4D contained the lowest percentage of marker pairs (less than 1%). The proportion of marker pairs in the A, B, and D genomes in LD greater than the critical $r^2$ of 0.36 were 55.3, 41.7, and 7.0%, respectively (Supplemental Fig. S3).

Table 2. Estimates of variance components and heritability for resistance to Puccinia striiformis f. sp. tritici in spring wheat accessions.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Mount Vernon</th>
<th>Pullman</th>
<th>Across environments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IT†</td>
<td>SEV</td>
<td>IT</td>
</tr>
<tr>
<td>Mean</td>
<td>5.0</td>
<td>53.5</td>
<td>4.9</td>
</tr>
<tr>
<td>Max.</td>
<td>9.0</td>
<td>100.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Min.</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$\sigma^2_g$</td>
<td>5.0***</td>
<td>750.2***</td>
<td>5.09***</td>
</tr>
<tr>
<td>$\sigma^2_i$</td>
<td>0.4ns</td>
<td>7.3ns</td>
<td>0.01ns</td>
</tr>
<tr>
<td>$\sigma^2_{g \times e}$</td>
<td>0.9***</td>
<td>240.4***</td>
<td>0.95***</td>
</tr>
<tr>
<td>$\sigma^2_{e}$</td>
<td>1.1ns</td>
<td>1.0ns</td>
<td>1.02ns</td>
</tr>
<tr>
<td>$h^2$</td>
<td>0.86</td>
<td>0.90</td>
<td>0.84</td>
</tr>
</tbody>
</table>

***, Significant at the <0.0001 probability level.
† IT, infection type; SEV, disease severity; $\sigma^2_g$, genotype variance; $\sigma^2_i$, environment variance; $\sigma^2_{g \times e}$, genotype × environment variance; $\sigma^2_e$, error variance; $h^2$, heritability; ns, not significant.
Prediction of GEBVs for the Quantitative Adult Plant Resistance to Stripe Rust

Estimates of GEBVs for stripe rust resistance and their accuracies were assessed under scenarios of different population sizes, degrees of genetic relatedness within a population, and marker densities using empirical data from multi-environment field trials. The results of the cross-validation tests generally demonstrated moderate to high average prediction accuracies across the different training population parameters and marker profiles. The lowest prediction accuracy (0.45) was recorded when the smallest training population size ($N_T = 210$) and the lowest marker density ($M_{10cM}$) were combined. Optimum levels of prediction accuracy were attained with an optimized training population of accessions with higher level of genetic relatedness than with a random subset or the entire population.

Effects of Training Population Size on GEBV Accuracy

Prediction accuracy increased significantly with an increase in the training population size (Table 3). Prediction accuracies were 0.50, 0.57, 0.60, and 0.63 for the respective population sizes of 210, 478, 640, and 959 accessions when cross-validation tests were performed by splitting each population into three-fourths as the TS and one-fourth as the VS. Accuracy increased by an average of 1% for every increase of 50 individuals in the training population size. Standard deviations and confidence intervals of the means associated with prediction accuracies using smaller training population sizes were higher than the larger training population sizes (Fig. 2, Table 3).

Genetic Relationships and Effect on GEBV Accuracy

Genetic relatedness and population structure were shown to be the most critical factors affecting the reliability of GEBVs. Optimizing the composition of a training population on the basis of genetic relatedness improved prediction accuracy compared with the use of populations comprised of randomly selected accessions. Prediction accuracy ranged from 0.75 to 0.79 and from 0.51 to 0.58 for predictions within the kinship and structure-based optimized subpopulations SP1 and SP2, respectively. The levels of prediction accuracies within SP1 and SP2 were consistent with the level of genetic relatedness among accessions within each subpopulation (Fig. 3). The kinship analysis indicated that accessions within SP1 were more highly related to each other than accessions within SP2. The average prediction accuracy was reduced to 0.08

Fig. 1. Population structure of the spring wheat accessions based on principal component analysis (PCA) and hierarchical clustering of the spring wheat panel. (A) Scatterplot of the first two principal components (PCs) of the accessions overlaid with geographic origin and improvement status of the accessions. PCA classified the accessions into two major subpopulations, labeled as Subpopulation 1 (SP1) and Subpopulation 2 (SP2). (B) Heat map and dendrogram of the pairwise kinship coefficients estimated on the basis of identity-by-descent using the Ward clustering algorithm. Vertical dotted lines indicate the genetic similarity thresholds used to classify accessions into SP1 and SP2.
and 0.04 for the inter set cross-validation test in which SP1 was used as TS and SP2 as VS, and vice versa, respectively. When both subpopulations were mixed and completely randomized to ensure uniform assignment of the genetically related accessions in SP1 and SP2 across the TS and VS, accuracy ranged from 0.70 to 0.73 (Table 4).

Predicting GEBVs by using Marker Panels of Various Densities

Cross-validation tests using the RR-BLUP method of GS revealed that all of the 5619 SNPs were more than necessary to capture the variation for stripe rust resistance in the germplasm panel. For almost all training population sizes (N_t), there were no significant differences in prediction accuracy between the marker densities of M_M (5619 SNPs at a density of 1 SNP per 1.3 cM) and M_lm (1849 SNPs at a density of 1 SNP per 3.2 cM) in the prediction model (Fig. 4). With the smallest training population size (N_t = 210), the accuracy of GEBVs at M_sm (543 genome-wise markers with 1 SNP per 9.4 cM) was slightly lower than that of both M_M and M_lm. However, the decrease in the accuracy became significantly higher at M_sm compared with M_M and M_lm as the population size increased to 478 and greater. Average prediction accuracy values at N_t = 210 corresponding to marker densities of M_M, M_lm, M_sm, and M_lm were 0.50, 0.51, 0.48, and 0.46, respectively, whereas the respective accuracy values at N_t = 959 were 0.63, 0.63, 0.59, and 0.56. Thus the effect of marker density on prediction accuracy was minimal at the smallest N_t but increased with increases in N_t (Fig. 4), highlighting the need for higher marker density for larger training population sizes. Furthermore, the accuracies of the GEBVs were similar for marker densities at M_M and M_lm across the different training population sizes, suggesting that a marker density of ~1849 SNPs was adequate to capture the QTL for stripe rust resistance in this population.

Joint GWAS–GS Strategy

We assessed the effect of selecting the markers significantly associated with resistance to stripe rust on the basis of GWAS in relation to GS prediction accuracy. Genome-wide association study using only accessions from the training populations, identified an average of 421 SNPs associated with stripe rust resistance at a marker-wise P < 0.05 in at least two of the five environments. Of the 421 loci, only 11 were significant according to the false discovery rate multiple correction method with a P-value of <0.1. The prediction accuracies of the 421 SNPs from the cross-validation tests were compared with those of the full marker panel (5619 SNPs) and other marker panels.

The cross-validation tests with the markers most strongly associated with resistance to stripe rust yielded an average prediction accuracy of 0.57 to 0.60, which was comparable to that of M_s (Table 3). Compared with the optimum marker density (1 SNP per 3.2 cM), at which prediction accuracy reached a plateau, prediction with the lower marker density of 421 SNPs selected via the GWAS test showed a decrease in accuracy of approximately 4%.

Effects of Trait Assessment Methods and Environments on GS Accuracy

Disease assessment methods (IT versus disease severity) and the environments (Pullman versus MTV) were used to assess the effects of trait assessment methods and environments on trait architecture and prediction accuracies. Consistent with the difference in stripe rust race composition and disease pressure between the two environments and assessment methods, severity at Pullman seems to be the most polygenic trait, as indicated by the nearly normal frequency distribution and normal quantile plots (Supplemental Fig. S1 and Supplemental Fig. S2). Similarly, the prediction accuracy response curve for severity at Pullman showed a linear increase with an increase in population density.
Fig. 2. Relationships between the size of the training population ($N_T$) and the accuracy of predicting genomic estimated breeding values (GEBVs) for quantitative adult wheat plant resistance to stripe rust *Puccinia striiformis* f. sp. *tritici*. Accuracies for the stripe rust infection type (IT) and disease severity (SEV) were assessed using four different training population sizes ($N_T = 210, 478, 640, \text{ and } 959$). Cross-validation tests were performed by splitting each of the four populations into the training set (TS) and the validation set (VS), with three-fourths of the individuals assigned to TS and one-fourth to VS. Prediction accuracies for stripe rust IT at Mount Vernon (MTV_IT), SEV at Mount Vernon (MTV_SEV), IT at Pullman (PLM_IT), and SEV at Pullman (PLM_SEV) were plotted against population size.

Fig. 3. Relationships between prediction accuracy and the extent of coefficient of identity-by-descent (IBD) kinship among accessions in wheat Subpopulation 1 (SP1), Subpopulation 2 (SP2), and their mixture (SP1+SP2). Accessions in SP1 exhibited a higher coefficient of kinship than those in SP2. The average prediction accuracies obtained from the cross-validation tests using SP1, SP2, and their mixtures (SP1+SP2) were plotted against the average IBD values for the respective populations.
size, whereas those of severity at MTV, IT at MTV, and IT at Pullman plateaued at the largest population size (Fig. 1). On the other hand, IT at Pullman showed significantly higher prediction accuracy than IT at MTV across all population sizes, except at the smallest population size \( N_T = 210 \). Although nonsignificant, severity at MTV gave a higher prediction accuracy than IT at MTV.

### Table 4. Effect of population structure and relatedness between individuals in the training and validation population on genomic estimated breeding value accuracy in wheat. Five cross-validation tests were performed with Subpopulation 1 (SP1), Subpopulation 2 (SP2), and the mixture of SP1 and SP2 to assess the impact of population structure and genetic relationships between individuals in the training set (TS) and validation set (VS) on prediction accuracy.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Parameter</th>
<th>Within SP1</th>
<th>Within SP2</th>
<th>Mixed (SP1+SP2)†</th>
<th>SP1_TS, SP2_VS‡</th>
<th>SP2_TS, SP1_VS§</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTV_IT</td>
<td>Mean</td>
<td>0.79</td>
<td>0.51</td>
<td>0.70</td>
<td>0.05</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.047</td>
<td>0.083</td>
<td>0.04</td>
<td>0.126</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td>SEM</td>
<td>0.003</td>
<td>0.006</td>
<td>0.003</td>
<td>0.009</td>
<td>0.007</td>
</tr>
<tr>
<td>MTV_SEV</td>
<td>Mean</td>
<td>0.79</td>
<td>0.53</td>
<td>0.72</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.051</td>
<td>0.081</td>
<td>0.04</td>
<td>0.118</td>
<td>0.145</td>
</tr>
<tr>
<td></td>
<td>SEM</td>
<td>0.004</td>
<td>0.006</td>
<td>0.003</td>
<td>0.008</td>
<td>0.01</td>
</tr>
<tr>
<td>PLM_IT</td>
<td>Mean</td>
<td>0.78</td>
<td>0.55</td>
<td>0.72</td>
<td>0.12</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.053</td>
<td>0.072</td>
<td>0.042</td>
<td>0.12</td>
<td>0.141</td>
</tr>
<tr>
<td></td>
<td>SEM</td>
<td>0.004</td>
<td>0.005</td>
<td>0.003</td>
<td>0.008</td>
<td>0.01</td>
</tr>
<tr>
<td>PLM_SEV</td>
<td>Mean</td>
<td>0.75</td>
<td>0.58</td>
<td>0.73</td>
<td>0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.058</td>
<td>0.066</td>
<td>0.039</td>
<td>0.134</td>
<td>0.163</td>
</tr>
<tr>
<td></td>
<td>SEM</td>
<td>0.004</td>
<td>0.005</td>
<td>0.003</td>
<td>0.009</td>
<td>0.012</td>
</tr>
</tbody>
</table>

† Prediction based on a mixture of SP1 and SP2, mixed and completely randomized to allow uniform assignment of each accession across TS and VS.
‡ Prediction accuracies generated by using SP1 as the TS and SP2 as the VS.
§ Prediction accuracies generated by using SP2 as the TS and SP1 as the VS.
¶ SEM, SE of the mean. MTV, Mount Vernon; PLM, Pullman; IT, infection type; SEV, severity.
Discussion

The collection of wheat genetic resources held in germplasm banks represent an enormous range of genetic diversity that has been underexploited in current wheat breeding programs. One of the major bottlenecks hindering the efficient use of these genetic resources is the cost of generating evaluation data for the substantially large number of available plant genetic materials (Endresen et al., 2011). A better sampling strategy is required to unlock the potential of plant genetic resources efficiently. Accordingly, the present study was performed with the aim of assessing the potential of GS to enhance the efficient use of the wheat genetic resources available in germplasm collections. Using empirical data on the quantitative adult plant resistance to stripe rust in a global collection of the USDA NSGC spring wheat accessions genotyped with a 9000 SNP Illumina assay, we explored the effects of different training population sizes, population structures and different marker densities on GS prediction accuracy. We also integrated the output of GWAS into a GS model to determine the possible use of a smaller number of markers that are strongly associated with stripe rust resistance to obtain adequate prediction accuracies.

Effects of Population Size on GEBV Prediction Accuracy

The accuracy of predicting GEBVs obtained via cross-validation analysis of empirical data is known to be affected by population size, population structure, marker density, and the genetic architecture of the traits (Daetwyler et al., 2010; Lorenz et al., 2011; Lorenz, 2013; Bassi et al., 2016). Most of the earlier GS studies conducted to examine the effect of these factors on prediction accuracy were based on a relatively restricted genetic diversity (Yu et al., 2016). To explain the relationship between GEBV prediction accuracy and population size in the context of broader genetic diversity in germplasm collection, we used differently sized training and validation populations. Prediction accuracies increased with an increase in training population sizes, corroborating previously reported findings both in elite germplasm and diversity panels (Heffner et al., 2011a; Tayeh et al., 2015; Combs and Bernardo, 2013, Crossa et al., 2016). The average prediction accuracies in this study for population sizes of 210, 478, 640, and 959 were 0.50, 0.57, 0.60, and 0.63, respectively. In empirical studies of other plant species such as in biparental populations of maize (Zea mays L.), Arabidopsis thaliana (L.) Heyn., and barley (Hordeum vulgare L.), the highest population size also resulted in the highest accuracy of GEBV prediction (Lorenzana and Bernardo 2009; Guo et al., 2012; Combs and Bernardo, 2013). In this study, it was notable that the increase in prediction accuracy with an increase in the training population size showed signs of reaching a plateau at either the largest training population size only or no sign of reaching a plateau at all, depending on the environment and the specific trait measured. The reason for achieving the plateau of prediction at such large population sizes may arise from the high level of genetic diversity in the worldwide collection of wheat accessions we used. Using high-density genomic data and historical phenotypic data collected from a large USDA Soybean germplasm collection, Jarquin et al. (2016) reported median prediction accuracies in the range of 0.56–0.69 for yield, protein content, and oil content, which supports the conclusion that large population sizes are required for achieving adequate levels of prediction accuracy when sampling from diverse germplasm collections. We recently reported a GWAS for resistance to stripe rust in the first core subset of the NSGC spring wheat germplasm collection (Maccaferri et al., 2015). With this study, the entire NSGC spring wheat core collection (~1800 accessions) has now been screened for stripe rust resistance and genotyped with the 9K SNP assay. This combined dataset should accelerate and enhance the efficiencies of wheat breeding programs by providing a robust training population for predicting GEBVs for stripe rust resistance.

Effect of Population Structure on GEBV Prediction Accuracy

Accessions in germplasm collections may consist of unique subgroupings caused by nonrandom mating between individuals that leads to the presence of an unequal distribution of alleles and LD level between these subpopulations (Flint-Garcia et al., 2003). To understand the effect of population structure and genetic relatedness on GS accuracy, we performed subset and interset validation experiments using two genetically distinct subpopulations (SP1 and SP2) from the core collection. The results of the cross-validation tests for predictions within SP1 and SP2 revealed differences in the accuracy of GEBVs that can be explained mainly by the degree of relationship between the training and validation populations. In a USDA soybean [Glycine max (L.) Merr.] germplasm collection, Jarquin et al. (2016) investigated the effect of predicting within and across subpopulations. They reported that predictive abilities were moderate to high (>0.6) for all the traits they assessed, whereas predictive abilities were highly variable when predicting across subpopulations. Hayes et al. (2009), Habier et al. (2010), and Asoro et al. (2011) also reported that groups of genotypes that are more genetically related predicted each other better than less related ones. Combining genetically distant clusters (SP1 and SP2) into one cross-validation population and ensuring a uniform assignment of genetically related individuals across TS and VS could provide an alternative design for optimized prediction accuracy. Accuracy was higher for the randomized combination of SP1 and SP2 than the accuracy of predicting within SP2 and the average of the two single clusters. This may indicate that accessions from SP1 included in the TS of the later cross-validation test (the randomized combination of SP1 and SP2) improved the prediction accuracy of the VS individuals from SP2. These results are in line with the report that a
higher prediction accuracy for cross-validation tests can be obtained by using genetically more related clusters than distantly related ones (Asoro et al., 2011). Asoro et al. (2011) also found that combining less related clusters gave a prediction accuracy that was better than the average of the two single clusters. Akdemir et al. (2015) used a genetic algorithm scheme to establish an optimized TS from a larger set of candidate individuals and observed improved the accuracy when the subsets were selected by the algorithm over random subsets of the same size. It was apparent that understanding the patterns of population structure in the germplasm panel and ensuring uniform assignment of genetically related individuals across the TS and VS can resolve the problem of genetic relatedness of accessions or a structured population and maximize prediction accuracy.

Effects of Marker Density on GEBV Prediction Accuracy
The ability of GS to enhance the use of crop germplasm collections rests mainly on the concept that estimation of breeding values can be done more cost-effectively by using the markers than field phenotyping (Meuwissen et al., 2001; Beaulieu et al., 2014; Jarquin et al., 2016; Yu et al., 2016). Hence, determining the appropriate density of markers required to capture the variation in the study population may enable lower-density genotyping and optimize costs to GS programs. In this study, we assessed the effect of different densities of markers on the accuracy of predicting breeding values across different training population sizes. The results showed that all 5619 available SNPs were more than necessary to obtain the highest achievable prediction accuracy, even with all 959 accessions. No further gain in prediction accuracy was observed beyond 1 SNP per 3.2 cM (~1850 markers), which is close to the LD decay rate in this population. However, the number of markers at which prediction accuracy reached a plateau was higher than the size reported by previous GS studies in wheat. The marker density required for attaining the maximum attainable prediction accuracy depends on the extent of LD and genetic diversity in the population, the genetic architecture of the trait, and the method of prediction. In a mixed population of wheat, Combs and Bernardo (2013) observed no difference in prediction accuracy between marker densities of 4.5 and 3.5 cM apart. Working on elite North American oats (Avena sativa L.), Asoro et al. (2011) reported that traits responded differently to an increase in marker density. According to this report, groat percentage reached a plateau in accuracy at 600 markers, whereas β-glucan, days to heading, plant height, and yield accuracy continued to increase to the maximum of 900 markers. The most likely explanations for the higher marker density required to reach the maximum possible prediction accuracy in the present study compared with most of the previous studies are the presence of extensive LD caused by historical recombination events as well as the high genetic diversity in the germplasm panel we used. Meuwissen (2009) also suggested that for predicting unrelated individuals, a substantially higher marker density is required to obtain high accuracies. Similarly, Asoro et al. (2011) reported that the marker density requirements for biparental populations are much lower than those for a set of lines with broad genetic diversity. In general, our results suggest that although the genotyping costs per marker data point are declining rapidly and facilitating the availability of abundant molecular marker data, marker density should be carefully considered to make screening large or complete germplasm collections more cost-effective.

Accuracy of Predicting GEBVs: Whole-Genome SNPs Versus SNPs in LD with QTL
Because high-density marker genotyping of large collections of germplasm may not be cost-effective, and because specific traits are often the focus of research projects, we evaluated the use of low-density markers that showed strong associations with resistance to stripe rust based on GWAS. Prediction accuracies obtained when using the markers in LD with QTL for stripe rust resistance were slightly lower than the optimum marker density but comparable to the accuracy achieved with a panel of genome-wide markers with density of 1 SNP per 9.4 cM. Habier et al. (2009) implemented a GS approach that used a panel of evenly spaced low-density SNPs across the genome to estimate the GEBVs of selection candidates in pedigreed populations and reported a slight loss of accuracy when a panel with evenly spaced SNPs at 10 cM was used compared with all high-density SNPs.

Effects of Environment and Trait Assessment Method on Prediction Accuracy
Prediction accuracy varied significantly depending on the environments where the experiments were conducted. These variations are consistent with the difference in trait genetic architecture caused by the difference in stripe rust race composition, environmental conditions for disease development, and assessment methods. With reduced race diversity and highly conducive conditions for stripe rust development, MTV gave the lowest prediction accuracy for IT compared with severity at MTV, IT at Pullman, and severity at Pullman. The effect of the difference between the environments and trait assessment methods on prediction accuracy was also shown by the linearly increasing response curve when prediction accuracy was plotted against population size for severity at Pullman. Conversely, GS accuracy tended to plateau at the largest population size for IT at Pullman, IT at MTV, and severity at MTV. The relatively low race diversity and the highly conducive conditions at MTV (Chen, 2005) might have led to the effectiveness of having a limited number of major-effect resistance loci, unlike Pullman, which had a more diverse race population and moderately conducive conditions for stripe rust development, leading to the effectiveness of having a higher number of QTL. These results are in
agreement with previous studies that reported that trait genetic architecture affects the performance of different GS methods differently: RR-BLUP works well for genetic architectures consisting of many loci of small effects (Jannink et al., 2010; Lorenz et al., 2011; Zhang et al., 2011). Our study provides empirical insights into the effect of the underlying genetic architecture on the performance of GS methods.

Conclusion

This study was conducted to investigate the potential of genomic prediction models to exploit the genetic diversity of complex traits in a large global collection of USDA wheat accessions. We assessed the effect of different population genetic parameters (population size and population structure) and marker properties (marker density) for their effect on the accuracy of predicting GEBVs in the context of exploiting wheat genetic diversity in the USDA germplasm collection. The overall levels of prediction accuracy for stripe rust resistance within the germplasm collection are similar to values reported for traits with moderate to high heritability in wheat and other crops (Heffner et al., 2011a, 2011b; Combs and Bernardo, 2013; Crossa et al., 2016). Average prediction accuracies ranged from 0.45 for the smallest population size and lowest marker density to 0.65 for the largest population size and marker density, whereas accuracies of up to 0.80 were obtained with an optimized training population when predicting within a specific subpopulation. With the rapidly declining costs of genotyping, the level of prediction accuracy observed in the present study implies the prospect of evaluating large numbers of accessions and selection of lines that combine desirable alleles based only on predicted breeding values (without being tested in the field) more efficiently. It also implies that GS would accelerate genetic gain by increasing selection intensity (through genotypic evaluation of a larger number of candidates) and reducing costs of selection. The possibility of improving the costs of GS experiments was further shown by analyzing the effect of marker density on prediction accuracy. Our results suggested that larger germplasm collections may be efficiently sampled based on lower-density genotyping methods or with smaller panels of trait-associated markers. The accuracy of GEBVs can be further improved by integrating the analyses of appropriate population genetic parameters (genetic structure and relationships) into GS experiments. Collectively, this study revealed that GS could provide an efficient and cost-effective sampling strategy of unlocking the potential of wheat genetic resources and accelerating the rate of genetic gain in wheat breeding programs.

Studies on the genomic prediction of traits measured in crop germplasm collection have been limited thus far. The results of a few published reports, such as studies on wheat (Crossa et al., 2016), sorghum [Sorghum bicolor (L.) Moench.] (Yu et al., 2016), and soybean (Jarquin et al., 2016) gene bank accessions, support the findings of the current GS study, collectively suggesting that genomic prediction can provide a promising global strategy for mining useful alleles from crop germplasm collections.

Once the genomic predictions of the GEBVs of accessions are possible on a large scale, a crucial step forward is to design efficient prebreeding strategies for mobilizing the useful alleles in applied breeding programs.

Availability of Data and Materials


Supplemental Information

Supplemental Fig. S1. Phenotypic distributions of stripe rust severity (SEV) and infection types (IT) based on BLUP values within each environment. MTV_IT, Mount Vernon IT; PLM_IT, Pullman IT; MTV_SEV, Mount Vernon SEV; PLM_SEV, Pullman SEV.

Supplemental Fig. S2. Normal quantile plots of the BLUP values of stripe rust severity (SEV) at Pullman (PLM) (A), SEV at Mount Vernon (MTV) (B), IT at PLM (C), and IT at MTV.

Supplemental Fig. S3. Scatter plot of average genome-wide linkage disequilibrium (LD) ($r^2$) as a function of the genetic distance between markers (cM) for the spring wheat association mapping. (A) The solid gray curve represents the model fitted to LD decay and the broken gray color represents the mean LD fitted to the model. (B) Chromosome-wise distribution of the number of marker pairs that showed LD arising from physical linkage.

Supplemental Fig. S4. Estimation of the most likely number of clusters in the germplasm panel based on the Bayesian clustering model in STRUCTURE software. Means of the estimated log-likelihood and the magnitude of $\Delta K$ were plotted against $K$ values based on five independent runs and with $K$ ranging from 1 to 10.

Supplemental Table S1. Origin, plant ID, and accession type of the 959 USDA NSGC spring wheat accessions.

Supplemental Table S2. Average prediction accuracy as affected by trait assessment method and environment.

Conflict of Interest Disclosure

The authors declare that they have no conflict of interest.

Acknowledgments

This project was funded by the National Research Initiative Competitive Grant 2011-68002-30029 (Triticeae-CAP) from the USDA National Institute of Food and Agriculture, and stipend support to KM from the Monsanto Beachell Borlaug International Scholars Program. MP and KTM conceived and designed the experiments. KTM performed the experiments. KTM analyzed the data. MP contributed reagents, materials, and analysis tools. KTM wrote the paper. MP, PB, ZZ, and XC revised the paper.