Published December 12, 2012

ORIGINAL RESEARCH

Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing

Jesse Poland,* Jeffrey Endelman, Julie Dawson, Jessica Rutkoski, Shuangye Wu, Yann Manes, Susanne Dreisigacker, José Crossa, Héctor Sánchez-Villeda, Mark Sorrells, and Jean-Luc Jannink

Abstract

Genomic selection (GS) uses genomewide molecular markers to predict breeding values and make selections of individuals or breeding lines prior to phenotyping. Here we show that genotyping-by-sequencing (GBS) can be used for de novo genotyping of breeding panels and to develop accurate GS models, even for the large, complex, and polyploid wheat (*Triticum aestivum* L.) genome. With GBS we discovered 41,371 single nucleotide polymorphisms (SNPs) in a set of 254 advanced breeding lines from CIMMYT's semiarid wheat breeding program. Four different methods were evaluated for imputing missing marker scores in this set of unmapped markers, including random forest regression and a newly developed multivariate-normal expectation-maximization algorithm, which gave more accurate imputation than heterozygous or mean imputation at the marker level, although no significant differences were observed in the accuracy of genomic-estimated breeding values (GEBVs) among imputation methods. Genomic-estimated breeding value prediction accuracies with GBS were 0.28 to 0.45 for grain yield, an improvement of 0.1 to 0.2 over an established marker platform for wheat. Genotyping-by-sequencing combines marker discovery and genotyping of large populations, making it an excellent marker platform for breeding applications even in the absence of a reference genome sequence or previous polymorphism discovery. In addition, the flexibility and low cost of GBS make this an ideal approach for genomics-assisted breeding.

Published in The Plant Genome 5:103–113.
doi: 10.3835/plantgenome2012.06.0006

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

GENOMIC SELECTION (GS) uses genomewide molecular markers to predict complex, quantitative traits in animal and plant breeding (Meuwissen et al., 2001). The underlying concept of GS is to model the entire complement of quantitative trait loci effects across the genome to produce a genomic estimated breeding value (GEBV), from which candidates can be selected by genotyping before phenotypic evaluation. The GS modeling approach was revolutionary in the sense that individual genetic effects were not identified but all markers were incorporated into the model to generate a prediction that was the sum total of all genetic effects, regardless of how minor. Genomic selection models have proven to be advantageous for complex traits such as grain yield where many loci of small effects control the trait (Burgueño et al., 2012; Crossa et al., 2010; de los Campos et al., 2009; González-Camacho et al., 2012; Jannink et al., 2010). Selection on single or limited numbers of markers for quantitative traits often misses a substantial portion of

J. Poland, USDA-ARS and Dep. of Agronomy, Kansas State Univ. (KSU), 4011 Throckmorton Hall, Manhattan KS, 66506; S. Wu, Dep. of Agronomy, Kansas State Univ., 4008 Throckmorton Hall, Manhattan KS, 66506; J. Endelman and J.-L. Jannink, USDA-ARS R.W. Holley Center, Cornell Univ., Ithaca, NY 14853; J. Dawson, J. Rutkoski, and M. Sorrells, Dep. of Plant Breeding and Genetics, Cornell Univ., 240 Emerson Hall, Ithaca NY 14853; Y. Manes, S Dreisigacker, J. Crossa, and H Sanchez-Villeda, International Maize and Wheat Improvement Center (CIMMYT), Int. Apdo. Postal 6-641, 06600 Mexico, DF, Mexico. Jesse Poland and Jeffrey Endelman contributed equally to this study. Received 4 June 2012. *Corresponding author (jesse.poland@ars.usda.gov and jpoland@ksu.edu).

Abbreviations: σ_e^2, error variance; BLUE, best linear unbiased estimate; BLUP, best linear unbiased prediction; DArT, Diversity Array Technology; DTH, days to heading; EM, expectation maximization; GBS, genotyping-by-sequencing; GEBV, genomic-estimated breeding value; GS, genomic selection; H^2, broad-sense heritability; het, heterozygote; MVN, multivariate normal; NGS, next-generation sequencing; RF, random forest; SAWSN, Semi-Arid Wheat Screening Nursery; SNP, single nucleotide polymorphism; TKW, thousand-kernel weight.

the genetic variance contributed by loci of small effects. Genomic selection modeling therefore takes advantage of the increasing abundance of molecular markers through modeling of many genetic loci with small effects.

One premise of using GS in applied breeding programs is the availability of high-density genomewide molecular markers at a cost that is comparable to (or lower than) the cost of phenotyping (Goddard and Hayes, 2007; Heffner et al., 2009; Jannink et al., 2010; Meuwissen et al., 2001). High costs of phenotyping and the availability of inexpensive genotyping make GS a more feasible and attractive alternative selection method. This balance can be seen in the first large-scale implementation of GS: dairy cattle (*Bos primigenius taurus*) breeding using single nucleotide polymorphism (SNP) arrays. This was clearly a situation where breeders were working with very expensive phenotypes and increasingly inexpensive molecular markers. Dairy bulls are selected based on daughter progeny testing, a phenotype that takes years to evaluate and typically cost tens of thousands of dollars per bull (Schaeffer, 2006). In contrast, GS using SNP arrays is considerably less expensive with a typical array priced at a few hundred dollars or less. In plant breeding, inbred line testing has typically been relatively inexpensive in the range of tens to hundreds of dollars depending on the plot size, level of replications, and number of locations evaluated (Heffner et al., 2010). As marker technologies also lagged behind in many crop species due to large and complex genomes and the lack of a reference sequence, the tipping point for large-scale application of GS is just now being met.

Wheat, a staple crop of global economic importance, has a very large polyploid genome. The wheat genome is hexaploid and roughly 16 Gbp in size (Arumuganathan and Earle, 1991), both of which have hindered molecular marker development for this crop. Recently, new approaches to genotyping using next-generation sequencing (NGS) have been demonstrated in a range of species as an effective tool to generate high-density genomewide markers at a low per-sample cost (Elshire et al., 2011). Genotyping-by-sequencing (GBS) takes advantage of restriction enzymes to capture a reduced representation of the target genome and DNA barcoded adapters to sequence multiple samples (96 to 384) in parallel on a single run of NGS platforms (multiplexing). Genotyping-by-sequencing has recently been applied in the large barley (*Hordeum vulgare* L.) and wheat genomes and shown to be an effective tool to rapidly generate molecular markers for these species (Poland et al., 2012).

Here we show that GBS can be applied directly to breeding programs and produce de novo molecular markers suitable for whole-genome predictions and GS. We applied GBS to a set of 254 elite breeding lines from the International Wheat and Maize Improvement Center (CIMMYT) and developed GS models for yield, days to heading (DTH), and thousand-kernel weight (TKW). Cross-validation was used to determine the accuracy of GEBVs. We investigated the marker imputation accuracy of missing data in the GBS datasets using random forest (RF) regression and a newly developed kinship-based imputation algorithm. Using the same cross validation approach, prediction accuracy using GBS markers was compared with a currently used array-based platform.

Materials And Methods

Germplasm and Phenotypes

The germplasm consisted of 254 advanced breeding lines from the CIMMYT Cycle 29 Semi-Arid Wheat Screening Nursery (SAWSN). The lines were F_6 derived from a set of 122 unique crosses, of which 68 were represented by a single line in the set of 254. The remaining families contributed 2 to 12 lines to the nursery.

The breeding lines were evaluated in 2010 in small yield plots of two beds measuring 0.8 by 3 m each (total plot of 1.6×3 m = 4.8 m^2) at the CIMMYT research station Campo Experimental Norman Ernest Borlaug (CENEB) in Ciudad Obregon, Mexico. Due to space limitations, the breeding lines were split into seven trials, each with three replicates, for both the irrigated and managed drought environments. Sister lines from the same cross were grouped in the same trial and evaluated together. Under full irrigation conditions, yield was measured for all three replicates and TKW for two replicates. Days to heading was also measured for two replicates but only the average was recorded. In the drought nursery yield was measured in three replications.

Genotypes

Genomic DNA was extracted from bulked leaves of ten 2-wk-old seedlings using a cetyltrimethylammonium bromide procedure (Saghai-Maroof et al., 1984) modified based on CIMMYT protocols (CIMMYT, 2005). As part of an earlier study, the lines had been genotyped using the Diversity Array Technology (DArT) platform (Diversity Arrays Technology Pty Ltd.) (Wenzl et al., 2004), which produced 1726 markers for this population.

The GBS libraries were constructed in 95-plex using the P384A adaptor set (Poland et al., 2012). For each plate a single random blank well was included for quality control to ensure that libraries were not switched during construction and sequencing. Genomic DNA was codigested with the restriction enzymes *Pst*I (CTGCAG) and *Msp*I (CCGG) and barcoded adapters were ligated to individual samples. Samples were pooled by plate into a single library and polymerase chain reaction amplified. Detailed protocols can be found in Poland et al. (2012) and the latest updates on the GBS approach for wheat can be found on the USDA Wheat Genetics and Germplasm Improvement website (http://www.wheatgenetics.org/research/). Each library was sequenced on a single lane of Illumina HiSeq 2000 (Cornell Life Science Core Laboratory Center).

We used a population-based SNP calling approach to identify informative SNP markers in the data set. Sequence tags (unique sequences within the full set of tags) were internally aligned using allowable mismatch of 3 bp in a 64 bp sequence. For each position in the tag, putative SNPs were identified. Single nucleotide polymorphisms were

then filtered using a Fisher exact test to determine if the SNP alleles were independent in a population of inbred lines (Fig. 1). For each putative SNP allele the number of individuals in the population with that allele was counted and a 2 × 2 table constructed of the number of individuals with one or the other allele, both, or neither. A Fisher exact test was then used to determine if the two alleles were independent. If the null hypothesis of independence for the putative SNP was rejected ($p < 0.001$) we assumed that the tags were allelic in the population (and therefore that the putative SNP was a true SNP). For putative SNPs due to sequencing errors, duplications, and homologous sequence on different genomes, the two alleles are often found in the same individual. For inbred lines as examined here, this would be evidenced by an excess number of individuals called heterozygous and failure to reject the null hypothesis of independence. For this study the SNPs were unordered and curated to include only one SNP per tag.

Imputation of Genotypic Data

The DArT markers had 2% missing data, which were imputed with the population mean for each marker. Four different imputation methods were evaluated for the GBS data, which had up to 80% missing data per marker: (i) using the marker mean value (mean), (ii) calling missing genotypes as heterozygotes (hets), (iii) using RF regression (Breiman, 2001), and (iv) using a multivariate normal (MVN)-expectation maximization (EM) algorithm.

Random Forest Imputation

Random forest is a machine-learning algorithm that uses an ensemble of decision trees, taking a quorum vote or average of the multiple decision trees to determine a classification or a prediction value for new instances. It is a very robust algorithm for classification and regression when there are thousands of input variables. In this study an ensemble average for 100 regression trees was used to impute the missing values for each marker with the RandomForest package (Liaw and Wiener, 2002) in R 2.14.1 (R Development Core Team, 2011) using R package multicore for parallelization. For each marker, the set of lines available for training consisted of those for which marker scores were available. For each regression tree, the algorithm generated a bootstrap sample as the training population, and the number of markers randomly sampled at each split was (by default) two-thirds of the total available. The missing genotypes for that marker were then predicted as the ensemble average of the 100 trees applied to the other markers (imputing missing values with the mean). In principle this process could be repeated using the first-round predictions instead of the mean in the training population. Based on cross-validation exercises (data not shown), we determined that additional iterations did not improve the imputation accuracy.

Multivariate Normal Expectation Maximization Algorithm

We developed a novel kinship-based imputation for GS using the EM algorithm. The EM algorithm represents a general approach to calculating maximum likelihood estimates of unknown parameters when data are missing (Dempster et al., 1977). To use the algorithm, a model must be specified for how the data were generated. Our results are based on the assumption that marker genotypes follow a MVN distribution. While this is obviously not exact—only three outcomes are possible for biallelic markers in diploid individuals—it is a useful approximation in the context of breeding value predictions with a realized (additive) relationship matrix.

To illustrate, let $\mathbf{X} \in \{0,1,2\}^{n\times m}$ be the genotype matrix for n lines and m biallelic markers with alleles designated α and β and marker scores coded αα = 0, αβ = 1, and ββ = 2. The symbol $\mathbf{W}$ denotes the "centered" genotype matrix constructed by subtracting the marker mean from each data point: $W_{ik} = X_{ik} - 2p_k$, in which p_k is the frequency of the β allele. Using the current population as the "base" (Powell et al., 2010), the realized relationship matrix is (VanRaden, 2008)

$$\mathbf{A} = \mathbf{WW}' / \left[2\sum\nolimits_k p_k \left(1 - p_k\right)\right]. \qquad [1]$$

From matrix algebra we can rewrite the outer product $\mathbf{WW}'$ in terms of the sample covariance matrix $\mathbf{S}$:

$$\begin{aligned} m^{-1}\mathbf{WW}' &= m^{-1}\sum_{k=1}^{m}\left(\mathbf{W}_{\bullet k} - \bar{\mathbf{w}}\right)\left(\mathbf{W}'_{\bullet k} - \bar{\mathbf{w}}'\right) + \bar{\mathbf{w}}\bar{\mathbf{w}}' \\ &= \mathbf{S} + \bar{\mathbf{w}}\bar{\mathbf{w}}' \end{aligned} \qquad [2]$$

in which $\bar{\mathbf{w}}$ denotes the row mean of $\mathbf{W}$. For an infinitesimal genetic model, the true relationship matrix is achieved in the large m limit. In this limit we can replace the sample statistics $\mathbf{S}$ and $\bar{\mathbf{w}}$ by the corresponding parameters for the distribution of marker genotypes, denoted by $\boldsymbol{\Sigma}$ (covariance) and $\boldsymbol{\mu}$ (mean):

$$\mathbf{A} = [\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']/E[2p(1-p)], \qquad [3]$$

in which $E[\cdot]$ denotes the expectation operator.

Thus far we have not specified the distribution of marker genotypes. If we now assume that the centered genotypes are MVN: $\mathbf{w} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, the EM algorithm can be used to estimate $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and in turn calculate the realized relationship matrix. From Eq. [3] the MVN assumption means markers are imputed based on realized relationships. While the MVN-EM algorithm is a common technique in statistics (Little and Rubin, 1987), concise matrix formulations of the problem are harder to find. The following description is taken from Schneider (2001).

In the maximization (M) step the missing data for each marker are imputed by multiple regression. This involves partitioning the current estimates of the mean and variance based on the pattern of observed (subscript 1) and missing (subscript 2) data for each marker:

$$\mathbf{W}_{2k} = \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\left(\mathbf{W}_{1k} - \hat{\boldsymbol{\mu}}_1\right). \qquad [4]$$

In the expectation (E) step, the estimates of the mean and variance are updated from the imputed marker matrix:

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \bar{\mathbf{w}} \\ \hat{\boldsymbol{\Sigma}} &= \left(m-1\right)^{-1}\sum_k\left(\hat{\mathbf{T}}_k - \bar{\mathbf{w}}\bar{\mathbf{w}}'\right) \end{aligned} \qquad [5]$$

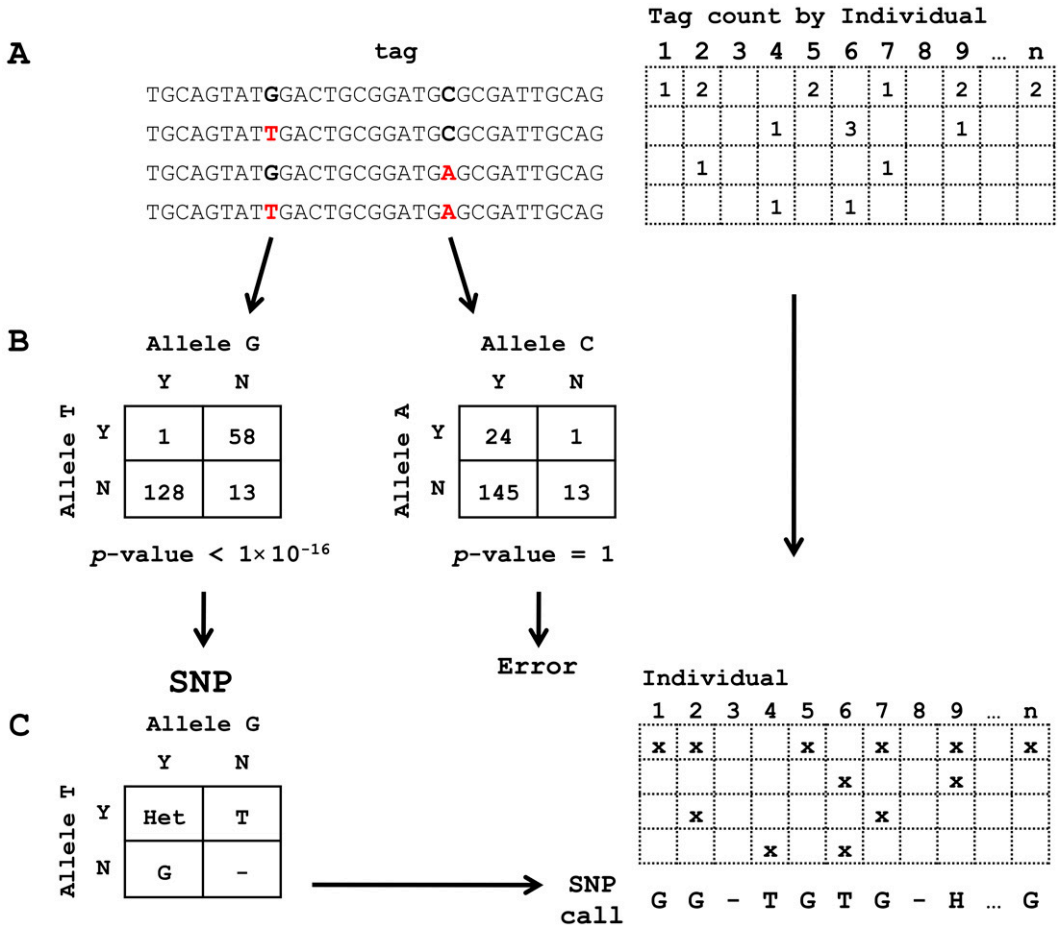

Figure 1. Single nucleotide polymorphism (SNP) calling from genotyping-by-sequencing tags. For reference-independent SNP calling, a population-based filtering approach was used. (A) Putative SNPs were first identified by internal alignment of sequence tags allowing 1 to 3 bp mismatch in a 64 bp tag. (B) The number of individuals (samples) in the population with each SNP allele were tallied and a Fisher exact test was conducted to test if the two alleles were independent. Within an inbred line, alleles at a biallelic SNP locus should be mutually exclusive (i.e., the inbred line should not have both alleles). Putative SNPs that failed the Fisher test (p-value < 0.001) were considered biallelic SNPs in the population and converted to SNP calls. (C) Based on presence–absence of the different tags in the individuals across the population, genotype scores were assigned. By incrementally increasing the stringency of the alignments, paralogous sequence on the alternate genomes could be filtered through genome-specific SNPs.

in which $\hat{\mathbf{T}}_k$ is the conditional expectation of the cross-products:

$$\hat{\mathbf{T}}_k = \begin{bmatrix} \mathbf{W}_{1k}\mathbf{W}'_{1k} & \mathbf{W}_{1k}\mathbf{W}'_{2k} \\ \mathbf{W}_{2k}\mathbf{W}'_{1k} & \mathbf{W}_{2k}\mathbf{W}'_{2k} + \hat{\mathbf{U}}_k \end{bmatrix}$$

$$\hat{\mathbf{U}}_k = \hat{\boldsymbol{\Sigma}}_{22} - \hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}. \quad [6]$$

We have written Eq. [6] as a partitioned matrix to show the presence of the imputation uncertainty term $\hat{\mathbf{U}}_k$ in the cross-product between alleles with missing data; in reality the row and column order is the same for all k.

The EM algorithm was initialized by imputing with the mean for each marker and estimating ($\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$) by the sample statistics. The convergence error at iteration t was calculated from the root mean-squared error:

$$n^{-1}\left\{\sum_{ij}\left[A_{ij}^{(t)} - A_{ij}^{(t-1)}\right]^2\right\}^{1/2}, \quad [7]$$

which represents the estimation precision for each element of the relationship matrix. A convergence threshold of 0.02 was used for all results. The MVN-EM algorithm is available through R package rrBLUP version 3.8 or higher (Endelman, 2011).

Marker Imputation Accuracy

Marker imputation accuracy was estimated by randomly masking 25 (nonmissing) genotypes per marker for 50 random SNPs with minor allele frequency ≥0.05. This process was repeated five times, using different SNPs each time, so that statistics are based on 250 unique markers. The accuracy for each marker was calculated as the mean absolute error for the masked genotypes: $(1/25)\sum_k |x_k - \hat{x}_k|$, which ranges from 0 to 2. Because the distribution of accuracies was nonnormal, we report the median and first and third quartile.

Statistical Analysis of Wheat Phenotypes

Both genotypic and breeding values (GEBV) were calculated by mixed model analysis in ASReml 3.0 (VSN International, 2009). Genotypic values (i.e., line means) were estimated as fixed effects (best linear unbiased estimate [BLUE]) with a random effect for replicates nested within trials. Breeding values were predicted as random effects (best linear unbiased prediction [BLUP]), with covariance proportional to the realized relationship matrix, and including random effects for trial and replication within trial. Broad-sense heritability (H^2) was calculated on a plot basis as $H^2 = V_g/(V_g + V_{error})$, in which V_g and V_{error} are the estimated genetic and residual variance components, respectively, when modeling the line effects as independent.

Genomic prediction accuracy was calculated by cross-validation using each of the seven trials as validation sets (folds). Because sister lines were only evaluated in the same trial, this partitioning ensured that no line in the training population had a full-sib line in the validation set. This gives a more stringent test of GS prediction by removing full-sib lines (most closely related) from the training population. Within each training population (consisting of six trials), GEBVs were calculated using the same mixed model method described above. Accuracy was calculated as the correlation between the GEBV and genotypic value (BLUE) in the validation set, and the mean reported is the average across the seven folds. This analysis was repeated for each of the four imputation methods as well as for the relationship matrix based on a reduced set of GBS markers and the relationship matrix from the DArT markers. The reduced GBS marker set was chosen to have approximately the same number of markers as the DArT set. By limiting the maximum percentage of missing data to 20%, we were left with 1827 GBS markers (vs. 1726 DArT).

To test for significance differences in prediction accuracy between the imputation methods, the cross-validation results were analyzed with SAS PROC GLM (SAS Institute, 2010), using fold as a blocking factor. The REGWQ procedure was used to control the strong familywise error rate at the 0.05 level. The same means comparison procedure was used to test for significant differences between the full GBS marker set, the reduced GBS marker set, and the DArT markers. We recognize that the ANOVA assumption of independence of errors is violated due to correlations between the testing and training sets in each fold and therefore *p*-values are not exact under the null hypothesis. The purpose of this ANOVA is not to test specific null hypotheses but simply to help quantify the relative magnitudes of the factors affecting accuracy.

Bias and Accuracy of Genomic-Estimated Breeding Values

The accuracy and bias of breeding value predictions were evaluated by simulation for the mean, heterozygote, and MVN-EM imputation methods. Random forest imputation was not included due to its large computational demand. Our analysis focused on the 100 lines with the least missing data, for which there were 5360 SNPs with less than 5% missing data. At this low level of missing data the imputation method is not important, so we called them as heterozygotes. The relationship matrix was calculated using all 5360 markers and then used as the covariance matrix ($\mathbf{G} = \mathbf{A}\sigma^2$) to simulate breeding values from the MVN distribution. Phenotypes were simulated by adding independent normal deviates (with variance σ_e^2) to each breeding value. Results are shown for $\sigma_e^2 = 0.8$ and $\sigma^2 = 1$, for which the expected heritability (regressing breeding values on phenotypes) is $\sigma^2(1 + f)/[\sigma^2(1 + f) + \sigma_e^2] \approx 0.7$. This heritability was confirmed in the simulation.

In each simulation, four of the seven trials were randomly assigned to the training population, and the remaining three trials were used as a validation set. This 4:3 split led to training populations with 47 to 67 lines and validation sets with 33 to 53 lines, with no sister lines shared between them. Genomic-estimated breeding values were predicted by BLUP using four different relationship matrices. The first was the same as that used to simulate the breeding values. For the other three relationship matrices we first randomly masked 50% of the genotypes per marker and then applied each of the imputation methods. The true (simulated) breeding values were then regressed on the GEBVs, for which we report the mean regression coefficient and the mean accuracy based on 1000 simulations.

Results

Genotypes and Phenotypes

We used GBS to genotype 254 lines from the Cycle 29 SAWSN. In this breeding panel we identified a set of 41,371 SNPs that were at an allele frequency greater than 1% and had more than 20% data present (Supplemental File S2). Removing multiple SNPs in the same tag reduced the marker number to 34,749 SNPs that were used for subsequent analysis. As is typical of sequence-based genotyping at low coverage, many markers had a large proportion of missing data. There was limited power to confirm low-frequency alleles in the presence of sequencing errors. This was evidenced by a decrease in the number of identified SNPs with minor allele frequency below 5% (data not shown).

Four quantitative traits from the 2010 season in Obregon, Mexico, were analyzed including yield under irrigated and drought conditions, TKW, and DTH (Fig. 2). The mean irrigated yield was 7.2 Mg ha^{-1} (SD = 0.6). Due to higher than normal precipitation, stress in the managed drought trial was mild, and the mean yield was 4.2 Mg ha^{-1} (SD = 0.3). Thousand kernel weight ranged from 37 g to 60 g, with a mean of 49 g (SD = 4). Heading was observed over a 3-wk period (73–93 d after planting), with an average of 85 days (SD = 4) after planting.

Table 1 shows the H^2 of the yield and TKW traits (only a single average measurement was available for heading date). For both irrigated and drought yield, $H^2 = 0.62$. The heritability of TKW was higher at $H^2 = 0.95$.

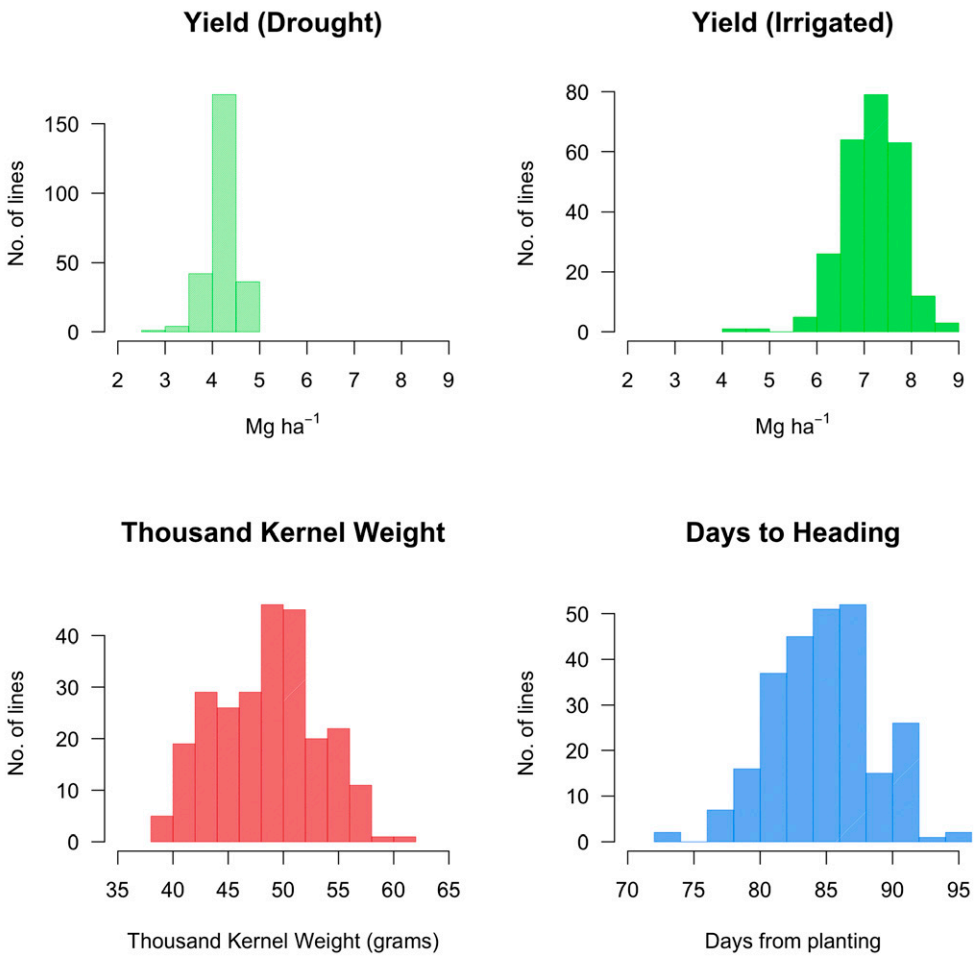

Figure 2. Phenotypic distribution of four agronomic traits on Cycle 29 Semi-Arid Wheat Screening Nursery (SAWSN). Each panel shows the distribution of best linear unbiased estimates of 254 lines from the Cycle 29 SAWSN used for this study.

Table 1. Broad-sense heritability (H^2) of agronomic traits evaluated in Cycle 29 Semi-Arid Wheat Screening Nursery.

Trait	H^2
Yield (irrigated)	0.62
Yield (drought)	0.62
Thousand kernel weight	0.95

Marker Imputation Accuracy

For most genomic prediction models, imputation of the missing genotypes is a necessary first step. We compared the accuracy of four different imputation methods by randomly masking 25 (nonmissing) genotypes per marker, imputing, and then calculating the mean absolute error for each marker (Fig. 3). The highest error was observed when the missing genotypes were called as heterozygotes.

Imputing with the population mean had the second-to-highest error. The strong dependence on minor allele frequency reflects the fact that as the minor allele frequency decreases, the population mean approaches the genotypic value of the majority of the lines and hence the error decreases. There was no relationship between error and the amount of missing data when imputing with the mean.

As expected, imputations based on RF or the realized relationship matrix (MVN-EM) were more accurate than the other two methods. We found that RF imputation was more accurate than MVN-EM imputation across the range of missing data but computational time was considerably increased. For both of these methods, the median error increased slightly with minor allele frequency and with the amount of missing data. The trend for missing data reflects the fact that markers with more missing data have smaller training populations and therefore larger error (Fig. 3).

Genomic Prediction Accuracy and Bias

Although the imputation methods were clearly differentiated in terms of marker imputation error, this made little to no difference with respect to genomic prediction accuracy. Genomic prediction accuracies, defined here as the correlation between GEBV and phenotype, were in the range of 0.3 to 0.5 for all traits (Table 2). The two simple imputation methods (heterozygote and mean) tended to have slightly higher accuracy, on the order of a few points, but for no trait was this difference in accuracy statistically significant (p-value > 0.05).

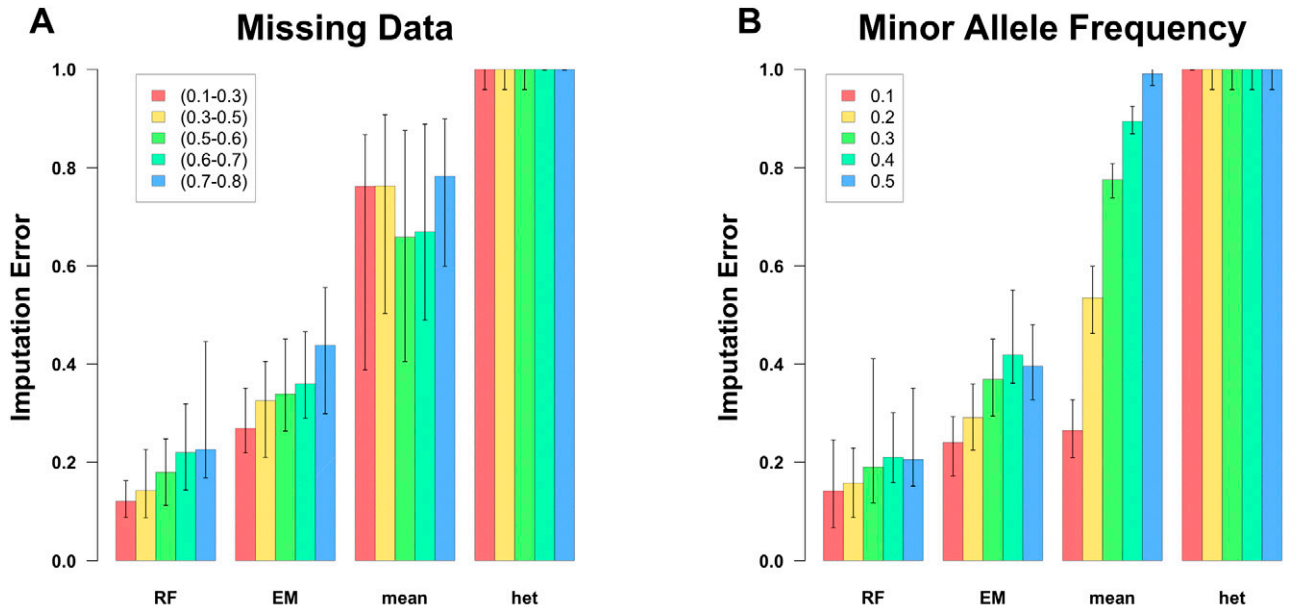

Figure 3. Marker imputation error on 254 breeding lines in the Cycle 29 Semi-Arid Wheat Screening Nursery. For each of 250 randomly chosen markers from the full set of 34,749 genotyping-by-sequencing (GBS) markers, 25 genotypes were masked and the imputed genotypes were compared to observed. Panel A shows imputation error at different levels missing data. The colors indicate what fraction of the 254 genotypes was missing before masking the 25 additional genotypes. The upper and lower limits for the range of different of missing data for different tests are shown in the legend. In panel B, results are shown as a function of the minor allele frequency. The median (column height) and first and third quartile (error bars) statistics are shown for four imputation methods: (i) heterozygote (het), (ii) population mean, (iii) multivariate normal expectation maximization (EM), and (iv) random forest (RF) regression.

Table 2. Prediction accuracy for agronomic traits in the Cycle 29 Semi-Arid Wheat Screening Nursery using different marker imputation methods. Cross-validation with seven folds was used and sister lines from the same cross were grouped in the same fold. No significant differences in prediction accuracy were observed among the imputation methods.

Trait	MVN-EM†	RF	Mean	Het
Yield (irrigated)	0.32	0.32	0.28	0.33
Yield (drought)	0.42	0.40	0.45	0.44
Thousand kernel weight	0.33	0.34	0.38	0.36
Days to heading	0.35	0.36	0.37	0.37

†Imputation methods are multivariate normal expectation maximization (MVN-EM), random forest (RF), mean, and heterozygote (Het).

Table 3. Prediction accuracy and bias for simulated phenotypes.

	Complete†	MVN-EM	Mean	Het
Accuracy	0.477 a	0.466 b	0.457 c	0.444 d
Regression coefficient‡	1.03 (0.02)	0.99 (0.02)	1.19 (0.02)	1.31 (0.03)

†Imputation methods are complete genotypic data, multivariate normal expectation maximization (MVN-EM), mean, and heterozygote (Het). Significant differences between imputation methods are shown by letters.

‡Regression of true breeding value on genomic estimated breeding value with standard error shown in parenthesis.

Our finding that the imputation methods were comparable with respect to breeding value prediction accuracy was confirmed in simulation. As shown in Table 3, there was one percentage point separating the heterozygote method ($r = 0.444$) from the mean method ($r = 0.457$), which was in turn one point lower than the MVN-EM algorithm ($r = 0.466$), and the MVN-EM accuracy was one point lower than with the complete marker data ($r = 0.477$). Although these differences are very small, they were significant based on 1000 simulations ($p < 10^{-10}$ in pairwise t tests). The ranking of the imputation methods (MVN-EM > mean > het) mirrors that observed for the marker imputation error.

Although comparable in accuracy, we observed significant bias in the GEBVs when using the mean or heterozygous imputation methods (Table 3). For these two methods, the mean regression coefficient was greater than one, indicating the GEBV tended to underestimate the true breeding values (in magnitude). As would be expected, this bias was more severe with the heterozygote method than with the mean method. The regression coefficient was not significantly different than one with either the complete marker data or when markers were imputed with the MVN-EM algorithm.

Comparison with Diversity Array Technology Markers

This population had previously been genotyped with 1729 DArT markers, an established platform for wheat (Akbari et al., 2006). Principal component analysis with the GBS and DArT markers produced similar results (Fig. 4). The two largest principal components accounted for 17 and 11% of the total variation in the realized relationship matrix with the GBS markers (calculated by

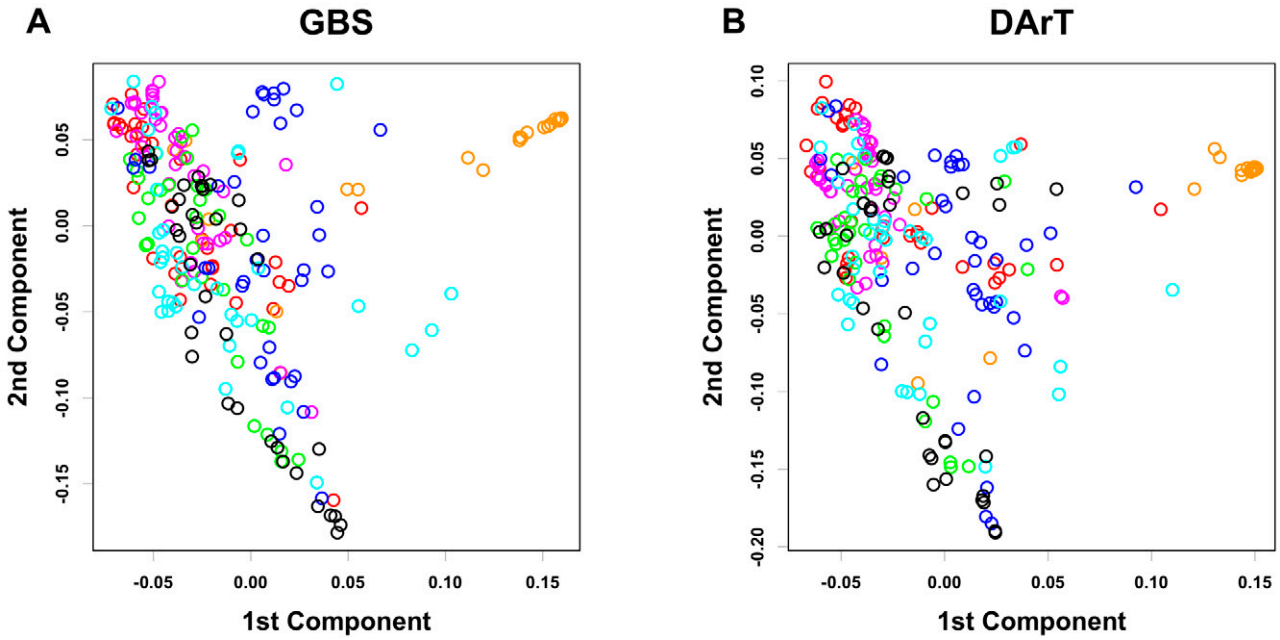

Figure 4. Principal component analysis of breeding lines from the Cycle 29 Semi-Arid Wheat Screening Nursery. Position of 254 wheat lines in the coordinate system defined by the first two principal components using multivariate normal expectation maximization imputed genotypes. The points are color coded according to the seven folds used in the genomic prediction cross-validation scheme. Panel A is based on genotyping-by-sequencing (GBS) markers and panel B is with Diversity Array Technology (DArT) markers.

MVN-EM) compared to 18 and 10% with the DArT markers. Although we did not quantify it, the general pattern of population structure was similar (Fig. 4).

Despite these similarities, the GBS markers led to higher genomic prediction accuracies (Fig. 5; Table 4). For both yield traits and heading date, the accuracy gain was in the range of 0.13 of 0.24. For TKW the increase was smaller (0.05) and not significant (p-value > 0.05). To investigate whether the higher accuracy with the GBS markers was due simply to higher density, we repeated the analysis using only the 1827 GBS markers with less than 20% missing data, which is comparable in number to the 1726 DArT markers. Compared to the full set of GBS markers, the mean accuracy with the reduced marker set was not significantly different for yield and TKW (0.07). Even with a comparable number of markers, the GBS platform led to significantly higher accuracy (gains of approximately 0.15) for drought yield and heading date when compared to the DArT markers.

Discussion

Our finding that imputation with RF led to lower error compared to imputation with MVN-EM is expected from the nature of the two algorithms. Whereas the MVN-EM algorithm imputes based on the realized relationship (averaged over all markers), the regression trees in the RF method can capture specific patterns of linkage disequilibrium. This increased power comes at a computational cost. Imputing all 34,749 markers with RF required 22 h when fully parallelized on a workstation with two 2.95 GHz 6-core Intel Xeon processors and 32 GB of RAM. By contrast, parallel execution of the MVN-EM algorithm converged after five iterations and took only 3 min. Although the simple imputation methods performed as well as RF and MVN-EM imputation for prediction accuracy of GEBV, the lower imputation error on a marker basis makes RF and MVN-EM imputation methods preferable.

We found that a GBS marker set of 34,749 produced significantly more accurate GEBVs than a DArT data set of 1729 markers. However, we also observed that a reduced set of only 1827 GBS markers still gave prediction accuracies better than the DArT markers and as good as the full GBS data set. The accuracy from 1829 GBS markers was lower, although not significantly different, than the accuracy from 34,749 GBS markers. The comparable performance of a limited number of GBS markers relative to the complete GBS data set of 34,749 markers indicates that (i) the population under study has relatively close relationships resulting in only a limited number of markers being need for full characterization, (ii) since the true breeding values remain unknown, uncertainty in the phenotypic observations limits the prediction accuracy, which was measured as the correlation between GEBVs and the observed phenotypes rather than the true breeding values, and/or (iii) the addition of GBS markers with higher levels of missing data does little to improve the characterization of kinship among the breeding lines. Simulation studies on this dataset indicated that increased accuracy due to higher heritability was large relative to the effect of increasing marker numbers (data not shown). Although not exclusive of the other conclusions, this supports the first conclusion that a limited number of markers are needed for this population to produce accurate predictions.

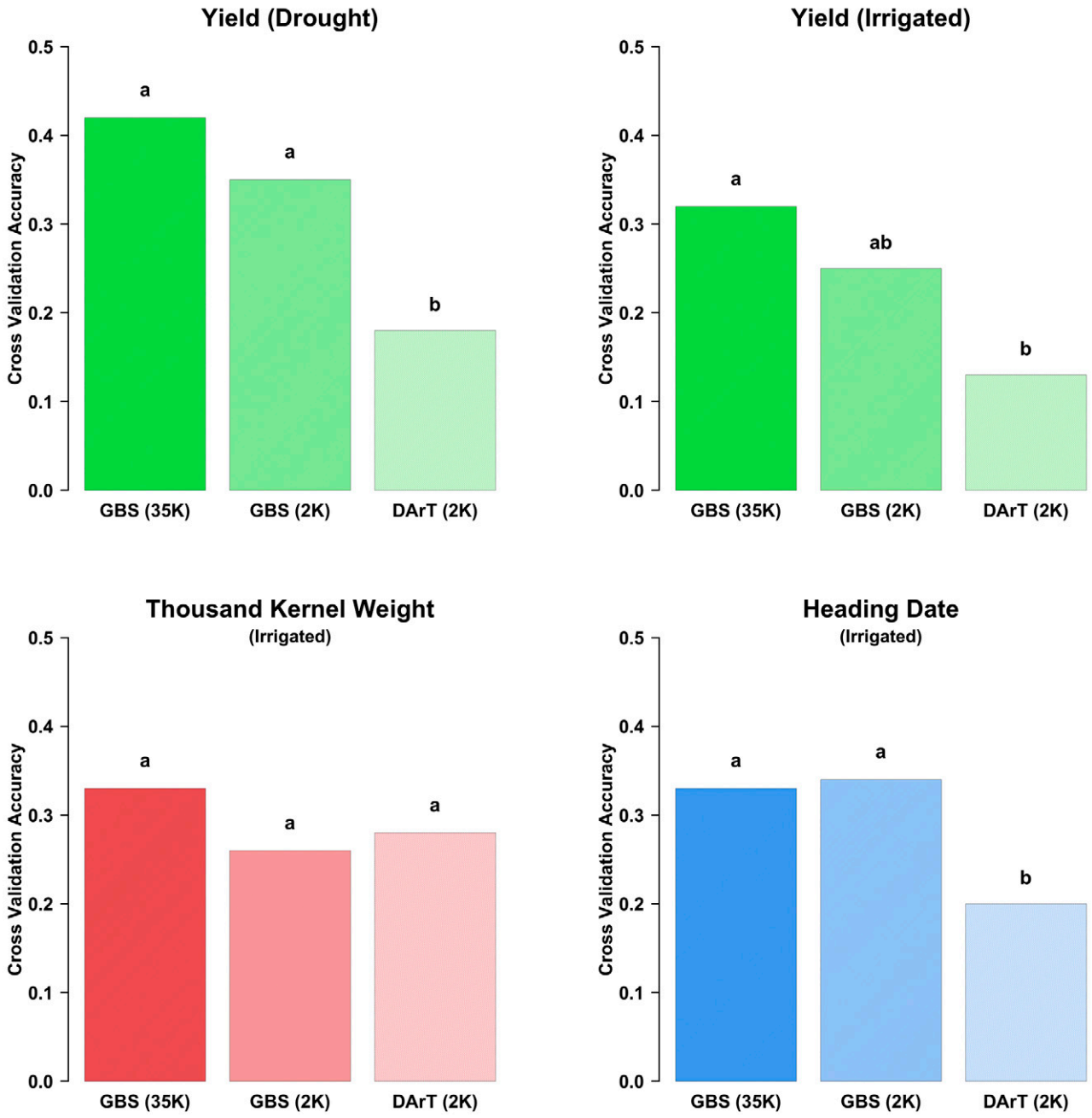

Figure 5. Cross-validation accuracy of genomic selection models for predicting line performance in the Cycle 29 Semi-Arid Wheat Screening Nursery, CIMMYT, using genotyping-by-sequencing (GBS) and Diversity Array Technology (DArT) markers on 254 elite breeding lines. Each trait was evaluated using sevenfold cross validation with sister lines from a single cross being grouped in the same fold. Significant differences among marker types within traits are denoted by letters above the bars. The approximate number of markers for each set are in parentheses. The actual numbers of markers are 1729 for DArT, 1827 for GBS (2K), and 34,749 for GBS (35K). Genotyping-by-sequencing (2K) markers have up to 20% missing data per marker and GBS (35K) have up to 80% missing data per marker.

It is unclear why more accurate predictions were observed with GBS than with DArT, even when controlling for marker number. One possibility is that the GBS markers are free of the genotypic ascertainment bias that is found with fixed array genotyping. For this study, the GBS markers were discovered concurrently with genotyping the set of germplasm of interest. As array-based genotyping often relies on a reference set of germplasm to discover, validate, and design markers, application of such platforms to different sets of germplasm could result in bias due to the nonrepresentativeness of the reference set in assaying polymorphisms in new sets of genotypes. Another possibility is that the GBS markers are more uniformly distributed across the genome than the DArT markers, which tend to cluster and show low density in the centromeric regions and the D genome (Akbari et al., 2006; Poland et al., 2012). This could lead to improved genome coverage with the same number of markers.

Table 4. Cross validation prediction accuracy for genomic selection models using genotyping-by-sequencing (GBS) and Diversity Array Technology (DArT) markers on 254 elite breeding lines from the Cycle 29 Semi-Arid Wheat Screening Nursery, CIMMYT. Letters denotes significant differences within traits.

Trait	Marker type and number		
	DArT 2000†	GBS 2000‡	GBS 35000§
Yield (irrigated)	0.13 b	0.25 ab	0.32 a
Yield (drought)	0.18 b	0.35 a	0.42 a
Thousand kernel weight	0.28 a	0.26 a	0.33 a
Days to heading	0.20 b	0.34 a	0.33 a

†Number of markers: 1729 for DArT, 1827 for GBS (2K), and 34,749 for GBS (35K).

‡Markers with up to 20% missing data.

§Markers with up to 80% missing data.

The prediction accuracies found in this study are sufficiently high to merit implementation of GS in applied breeding programs. Heffner et al. (2010) found that an accuracy of only 0.3 would be needed for GS to increase rates of gain on a per-year basis in a winter wheat breeding program relative to a marker-assisted breeding program. In the CIMMYT spring wheat breeding programs, GS will need to have higher accuracy to outperform phenotypic selection due to faster selection cycles. In applied breeding programs, iterative testing and model updating is expected to produce more accurate models as the training population builds each year with new entries (Jannink et al., 2010; VanRaden et al., 2009). It is also important to note that we are only comparing the GEBVs in this study to observed phenotypes and the true breeding values remain unknown. Given true breeding values, we would expect a higher correlation to the GEBVs. We have, however, focused only on within-environment predictions. Across different environments it is expected that heritability along with prediction accuracy would decrease as genotype × environment interactions are introduced. Phenotypic selection, however, faces this same limitation, which has always been a challenge for plant breeders.

Conclusions

Rapid advances in output from NGS platforms with corresponding decreases in cost have made sequence-based genotyping a very attractive and practical approach to rapidly characterize genomes and populations. Previously it has been shown that GBS can be used to generate high-density markers efficiently and inexpensively. Here we have shown that GBS can be used to generate markers to characterize wheat breeding lines and develop accurate GS models. We have concurrently developed a novel EM algorithm to impute unordered markers and shown that this method produces unbiased GEBVs. Prediction accuracies from GBS were consistently higher than those using an established marker platform for important agronomic traits, including grain yield, and in the range needed to apply GS in breeding programs.

By applying GBS directly to elite breeding lines we have demonstrated that it is both a cost-effective and robust marker platform for genomics-assisted breeding even for a species with a genome as challenging as wheat. In this study the cost of producing GBS data was less than US$20 per sample. Advances in NGS platforms, even in the short time since these data were generated, would now permit collection of the same amount of data per sample at a cost of $10 (192-plexing). In fact, the current cost for GBS is well below that of replicated yield testing (Heffner et al., 2010), showing that it is very practical for applied breeding programs to start selecting for complex traits before advanced testing.

Genotyping-by-sequencing is a suitable marker platform for generating robust, genomewide molecular markers for a low per-sample cost for wheat breeding programs. De novo marker discovery in the GBS datasets also makes this an excellent tool for new species or understudied crops with limited genomics research. Genotyping-by-sequencing can be applied to different populations or even different species without any prior genomic knowledge as marker discovery is simultaneous with the genotyping of the population. The use of GBS for GS, therefore, should be applicable to a range of model and nonmodel crop species to implement genomics-assisted breeding. Continual improvements in NGS capacity will make this genotyping approach even more attractive by further decreasing sample costs while improving genotyping output. Coupled with available reference genomic sequence, GBS markers become even more powerful when physically mapped, allowing the use of more accurate imputation algorithms. The combination of low per-sample costs and flexibility make GBS an ideal tool for genomics-assisted breeding in crops.

Supplemental Information Available

Supplemental material is available at http://www.crops.org/publications/tpg.

Supplemental File S1. Best linear unbiased estimates (BLUEs) for yield, heading date, and thousand kernel weight and DArT markers for CIMMYT Cycle 29 Semi-Arid Wheat Screening Nursery.

Supplemental File S2. Genotyping-by-sequencing data for Cycle 29 Semi-Arid Wheat Screening Nursery.

Acknowledgments

The Cornell University Life Sciences Core Laboratory Center conducted Illumina sequencing for all of the materials. The Biochemistry core facility at KSU provided florescence plate reader for quantifying DNA plates. Funding for this research was provided by the Bill & Melinda Gates Foundation through a grant to Cornell University for "Genomic Selection: The next frontier for rapid gains in maize and wheat improvement" and the United States Department of Agriculture-Agricultural Research Service (Appropriation #5430-21000-006-00D). J. Rutkoski is supported through USDA National Needs Fellowship Grant 2008-38420-04755. This research was supported in part by Triticeae-CAP USDA-NIFA-AFRI grant 2011-68002-30029, Hatch project 149-449, Australian Grains Research & Development Corporation, Kansas Wheat Alliance, Kansas State University, and Cornell University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Mention of trade names

or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer. Contribution no. 13-079-J from the Kansas Agricultural Experiment Station.

References

Akbari, M., P. Wenzl, V. Caig, J. Carling, L. Xia, S. Yang, G. Uszynski, V. Mohler, A. Lehmensiek, H. Kuchel, M. Hayden, N. Howes, P. Sharp, P. Vaughan, B. Rathmell, E. Huttner, and A. Kilian. 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor. Appl. Genet. 113:1409–1420. doi:10.1007/s00122-006-0365-4

Arumuganathan, K., and E. Earle. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9:208–218. doi:10.1007/BF02672069

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32. doi:10.1023/A:1010933404324

Burgueño, J., J. Crossa, J.M. Cotes, F.S. Vicente, and B. Das. 2012. Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. Crop Sci. 52:707–719.

CIMMYT. 2005. Laboratory protocols: CIMMYT applied molecular genetics laboratory. 3rd ed. CIMMYT, Mexico, D.F., Mexico.

Crossa, J., G. de los Campos, P. Perez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385. doi:10.1534/genetics.109.101501

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Society. Series B (Methodological) 39:1–38.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. doi:10.1371/journal.pone.0019379

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen. 4:250–255. doi:10.3835/plantgenome2011.08.0024

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124:323–330. doi:10.1111/j.1439-0388.2007.00702.x

González-Camacho, J., G. de los Campos, P. Pérez, D. Gianola, J. Cairns, G. Mahuku, R. Babu, and J. Crossa. 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125:759–771. doi:10.1007/s00122-012-1868-9

Heffner, E.L., A.J. Lorenz, J.-L. Jannink, and M.E. Sorrells. 2010. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. 50:1681–1690. doi:10.2135/cropsci2009.11.0662

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Briefings in Functional Genomics 9:166–177. doi:10.1093/bfgp/elq001

Liaw, A., and M. Wiener. 2002. Classification and regression by RandomForest. R News 2:18–22.

Little, R.J.A., and D.B. Rubin. 1987. Statistical analysis with missing data. 1st ed. John Wiley & Sons, New York, NY.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253. doi:10.1371/journal.pone.0032253

Powell, J.E., P.M. Visscher, and M.E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800–805. doi:10.1038/nrg2865

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Saghai-Maroof, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA 81:8014–8018. doi:10.1073/pnas.81.24.8014

SAS Institute. 2010. The SAS system for Windows. Release 9.3. SAS Inst., Cary, NC.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223 doi:10.1111/j.1439-0388.2006.00595.x

Schneider, T. 2001. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. J Clim. 14:853–871. doi:10.1175/1520–0442(2001)014

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24. doi:10.3168/jds.2008-1514

VSN International. 2009. ASReml 3. VSN Intl., Hemel Hempstead, UK.

Wenzl, P., J. Carling, D. Kudrna, D. Jaccoud, E. Huttner, A. Kleinhofs, and A. Kilian. 2004. Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc. Natl. Acad. Sci. USA 101:9915–9920. doi:10.1073/pnas.0401076101