Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing

Jesse Poland\textsuperscript{a}, Jeffrey Endelman\textsuperscript{a}, Julie Dawson, Jessica Rutkoski, Shuangye Wu, Yann Manes, Susanne Dreisigacker, Jose Crossa, Hector Sanchez-Villeda, Mark Sorrells, Jean-Luc Jannink

Author Affiliations:
J. Poland, United States Department of Agriculture – Agricultural Research Service (USDA-ARS) and Department of Agronomy, Kansas State University (KSU), 4011 Throckmorton Hall, Manhattan KS, 66506; S. Wu, Department of Agronomy, KSU; J. Endelman and J-L. Jannink, USDA-ARS and Department of Plant Breeding and Genetics, Cornell University, 240 Emerson Hall, Ithaca, NY 14853, J. Dawson, J. Rutkoski, and M. Sorrells, Dept. of Plant Breeding and Genetics, Cornell University, 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

\textsuperscript{a} These authors contributed equally to this study

* Corresponding author: jesse.poland@ars.usda.gov

Received:

Journal: The Plant Genome
Abstract

Genomic selection (GS) uses genome-wide 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 genome. With GBS we discovered 41K SNPs in a set of 254 advanced breeding lines from CIMMYT’s semi-arid 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, though no significant differences were observed in the accuracy of genomic-estimated breeding values (GEBVs). GEBV prediction accuracies with GBS were 0.28 – 0.45 for grain yield, an improvement of 0.1-0.2 over an established marker platform for wheat. GBS combines marker discovery and genotyping of large populations making it an excellent marker platform for breeding applications even in the absence of 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.
Genomic selection (GS) uses genome-wide 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 QTL effects across the genome to produce a genomic estimated breeding value (GEBV) from which candidates can be selected by genotyping, prior to 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. GS models have proven to be advantageous for complex traits such as grain yield where many loci of small effects control the trait (Burgueno et al., 2011; 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 the genetic variance contributed by loci of small effects. GS modeling therefore takes advantage of the increasing abundance of molecular markers through modeling of many genetic loci with small effects.

One premise of using genomic selection in applied breeding programs is the availability of high-density genome-wide molecular markers at a cost that is comparable (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 makes GS a more feasible and attractive alternative selection method. This balance can be seen in the first large-scale implementation of GS: dairy cattle breeding using 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, genomic selection 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 genomic selection 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 giga-basepairs 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 genome-wide 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 in parallel multiple samples (96 to 384) on a single run of next-generation sequencing platforms (multi-plexing). GBS has recently been applied in the large barley 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 genomic selection. We applied GBS to a set of 254 elite breeding lines from the International Wheat and Maize Improvement Center (CIMMYT) and developed genomic selection models for yield, days to heading, and thousand-kernel weight. 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 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

Germlasm and phenotypes

The germplasm consisted of 254 advanced breeding lines from the CIMMYT Cycle 29 Semi-Arid Wheat Screening Nursery (C29SAWSN). 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 m x 3 m each (total plot of 1.6 m x 3 m = 4.8 m²) at the CIMMYT research station Campo Experimental Norman Ernest Borlaug (CENEB) in Cd. 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 thousand-kernel weight (TKW) for two replicates. Days to heading (DTH) 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 two-week-old seedlings using a CTAB procedure (Saghai-Maroof et al., 1984) modified based on CIMMYT protocols (http://www.cimmyt.cgiar.org/ABC/Protocols/manualABC.html). As part of an earlier study, the lines had been genotyped using the Diversity Array Technology (DArT) platform (Diversity Arrays Technology Pty Ltd, Yarralumla, ACT 2600, Australia) (Wenzl et al., 2004), which produced 1726 markers for this population.

The GBS libraries were constructed in 95-plex using the P384A adapter set (Supplemental Dataset S1) (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 co-digested with the restriction enzymes PstI (CTGCAG) and MspI (CCGG) and barcoded adapters were ligated to individual samples. Samples were pooled by plate into a single library and PCR amplified. Detailed protocols can be found in Poland et al. (2012) and at www.wheatgenetics.org. 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. SNPs were then filtered using a Fisher
Exact test to determine if the SNP alleles were independent in a population of inbred lines (Figure 1). For each putative SNP allele the number of individuals in the population with that allele was counted and a 2 x 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 thus 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: (1) using the marker mean value (mean); (2) calling missing genotypes as heterozygotes (het); (3) using random forest regression (RF; Breiman, 2001), and (4) using a multivariate normal expectation-maximization algorithm (MVN-EM).

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, 2009) using R package multicore for parallelization (Urbanek, 2011). 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) 2/3 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.

**MVN EM 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 multivariate normal distribution. While this is obviously not exact—only three outcomes are possible for bi-allelic 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 $X \in \{0,1,2\}^{n \times m}$ be the genotype matrix for $n$ lines and $m$ bi-allelic markers with alleles designated $\alpha$ and $\beta$ and marker scores coded $\alpha\alpha = 0$, $\alpha\beta = 1$, $\beta\beta = 2$. The symbol $W$ denotes the “centered” genotype matrix, constructed by subtracting the marker mean from each data point: $W_{ik} = X_{ik} - 2p_k$, where $p_k$ is the frequency of the $\beta$ allele. Using the current population as the “base” (Powell et al., 2010), the realized relationship matrix is (VanRaden, 2008):

$$\Lambda = \frac{WW'}{2 \sum_{k=1}^{m} p_k (1 - p_k)}$$

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

$$WW' = \Lambda S$$
$m^{-1}WW' = m^{-1} \sum_{k=1}^{m} (W_{sk} - \langle W_{sk} \rangle)(W_{sk} - \langle W_{sk} \rangle)' + \langle W_{sk} \rangle \langle W_{sk} \rangle'$

where $\langle W_{sk} \rangle$ denotes the row mean of $W$. 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 $S$ and $\langle W_{sk} \rangle$ by the corresponding parameters for the distribution of marker genotypes, denoted by $\Sigma$ (covariance) and $\mu$ (mean):

$$A = \frac{\Sigma + \mu \mu'}{2E[p(1-p)]}$$

where $E[\cdot]$ denotes the expectation operator.

Thus far we have not specified the distribution of marker genotypes. If we now assume that the centered genotypes are multivariate normal (MVN): $w \sim N(\mu, \Sigma)$, the EM algorithm can be used to estimate $(\mu, \Sigma)$ and in turn calculate the realized relationship matrix. From Equation 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 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 (1) and missing (2) data for each marker:

$$W_{2k} = \hat{\mu}_2 + \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} (W_{2k} - \hat{\mu}_1)$$

$$W_{*k} = \begin{bmatrix} W_{1k} \\ W_{2k} \end{bmatrix} \hat{\mu} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} \hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix}$$
In the E-step, the estimates of the mean and variance are updated from the imputed marker matrix:

\[ \hat{\mu} = \langle W_{ik} \rangle \]

\[ \hat{\Sigma} = (m-1)^{-1} \sum_k \hat{T}_k - \langle W_{ik} \rangle \langle W_{ik} \rangle \]  

where \( \hat{T}_k \) is the conditional expectation of the cross-products:

\[ \hat{T}_k = \begin{bmatrix} W_{1k} W'_{1k} & W_{1k} W'_{2k} \\ W_{2k} W'_{1k} & W_{2k} W'_{2k} + \hat{U}_k \end{bmatrix} \]

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

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

\[ n^{-1} \sqrt{\sum_{\theta} [A^{(t)}_{\theta} - A^{(t-1)}_{\theta}]^2} \]

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 (non-missing) genotypes per marker for 50 random SNPs with minor allele frequency \( \geq 0.05 \). This process was repeated 5 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: \[
\left( \frac{1}{25} \right) \sum_k |x_k - \hat{x}_k|,
\]
which ranges from 0 to 2.

Because the distribution of accuracies was non-normal, we report the median and first/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, UK). Genotypic values (i.e., line means) were estimated as fixed effects (BLUE) with a random effect for replicates nested within trials. Breeding values were predicted as random effects (BLUP), with covariance proportional to the realized relationship matrix, and including random effects for trial and replication within trial. For broad-sense heritability calculations the genotypic effects were modeled as independently and identically distributed, and \( \text{H}^2 = \frac{V_g}{V_g + V_{error}} \).

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, Cary, NC), using fold as a blocking factor. The REGWQ procedure was used to control the strong family-wise 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 thus 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 GEBVs**

The accuracy and bias of breeding value predictions were evaluated by simulation for the mean, heterozygote, and MVN-EM imputation methods. RF 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 5,360 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 5,360 markers and then used as the covariance matrix \( \mathbf{G} = \mathbf{A} \sigma^2_e \) to simulate breeding values from the multivariate normal distribution. Phenotypes were simulated by adding independent normal deviates (with variance \( \sigma^2_e \)) to each breeding value. Results are shown for \( \sigma^2_e = 0.8 \) and \( \sigma^2 = 1 \), for which the expected heritability (regressing breeding values on phenotypes) is \( \frac{\sigma^2}{\sigma^2(1+f) + \sigma^2_e} \approx 0.7 \). This heritability was confirmed in the simulation.

In each simulation, 4 of the 7 trials were randomly assigned to the training population, and the remaining 3 trials were used as a validation set. This 4:3 split led to training populations with 47–67 lines and validation sets with 33–53 lines, with no sister lines shared between them. GEBVs 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 applying each of 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 Dataset 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 frequency 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, thousand kernel weight, and days to heading (Figure 2). The mean irrigated yield was 7.2 (SD 0.6) Mg ha\(^{-1}\). Due to higher than normal precipitation, stress in the managed drought trial was mild, and the mean yield was 4.2 (SD 0.3) Mg ha\(^{-1}\). Thousand kernel weight (TKW) ranged from 37 g to 60 g, with a mean of 49 (SD 4) g. Heading was observed over a three-week period (73–93 days after planting), with an average of 85 (SD 4) days after planting.

Table 1 shows the broad-sense heritability 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 \(H^2 = 0.95\).

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 (non-missing) genotypes per marker, imputing, and then calculating the mean absolute error for each marker (Figure 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 thus larger error (Figure 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 selection prediction accuracies, defined here as the correlation between GEBV and phenotype, was in the range 0.3–0.5 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).

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 DArT markers**

This population had previously been genotyped with 1,729 DArT markers, an established platform for wheat (Akbari et al., 2006). Principal component analysis with the GBS and DArT markers produced similar results (Figure 4). The two largest PCs accounted for 17% and 11% of the total variation in the realized relationship matrix with the GBS markers (calculated by MVN-EM) compared to 18% and 10% with the DArT markers. Though we were not able to quantify it, the general pattern of population structure was similar (Figure 4).

Despite these similarities, the GBS markers led to higher genomic prediction accuracies (Figure 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 1,827 GBS markers with less than 20% missing data, which is comparable in number to the 1,726 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 ≈ 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 35K markers with RF
required 22 hours 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 three minutes. Although, the simple imputations 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 35K produced significantly more accurate GEBVs than a DArT data set of 1,700 markers. However, we also observed that a reduced set of only 1,800 GBS markers still gave prediction accuracies better than the DArT markers and as good as the full GBS data set. The accuracy from 1,800 GBS markers was lower, though not significantly different, than the accuracy from 35K GBS markers. The comparable performance of a limited number of GBS markers relative to the complete GBS data set of 35K markers indicates that 1) the population under study has relatively close relationships resulting in only a limited number of markers being need for full characterization, 2) 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 3) 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 $h^2$ was large relative to the effect of increasing marker numbers (data not shown). Though 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.

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 non-representativeness 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), which could lead to improved genome coverage for the same number of markers.

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, only focused on within environment predictions. Across different environments it is expected that heritability along with prediction accuracy would decrease as genotype-by-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 genomic selection 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 sequence-based genotyping 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 $20 per sample. Advances in NGS platforms, even in the short time since this data was 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 prior to advanced testing.

Genotyping-by-sequencing is a suitable marker platform for generating robust, genome-wide 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. GBS 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 non-model 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 data becomes even more powerful with physical based ordering of markers and additional algorithms for imputation. The combination of low per sample costs and flexibility make GBS an ideal tool for genomics-assisted breeding in crops.

**ACKNOWLEDGEMENTS**

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 and Melinda Gates Foundation and the United States Department of
Agriculture – Agricultural Research Service (Appropriation No. 5430-283 21000-005-00). J Rutkoski is supported through USDA National Needs Fellowship Grant 2008-38420-04755. This research was supported in part by Triticaceae-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.
REFERENCES


Urbanek, S. 2011. Multicore: Parallel processing of R code on machines with multiple cores or CPUs, R package version 0.1-7.


FIGURES

Figure 1. 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 one to three bp mismatch in a 64 bp tag. 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 bi-allelic 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 bi-allelic 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.

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

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 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. In panel B, results are shown as a function of the minor allele frequency. The median (column height) and first/third quartile (error bars) statistics are shown for four imputation methods: (1) heterozygote (het), (2) population mean, (3) multivariate normal expectation maximization (EM), and (4) random forest (RF) regression.
Figure 4. Principal component analysis (PCA) of breeding lines from the Cycle 29 SAWSN. Position of 254 wheat lines in the coordinate system defined by the first two principal components using MVN-EM 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 GBS markers and panel B is with DArT markers.

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 DArT markers on 254 elite breeding lines. Each trait was evaluated using 7-fold 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 1,729 for DArT, 1,827 for GBS (2K), and 34,749 for GBS (35K). GBS (2K) markers have up to 20% missing data per marker and GBS (35K) have up to 80% missing data per marker.

SUPPLEMENTAL DATA

Dataset S1 - Information on elite breeding lines in the Cycle 29 Semi-Arid Wheat Screening Nursery use for this study including phenotypic data and genotyping-by-sequencing library information.

Dataset S2 – Genotyping-by-sequencing data for Cycle 29 Semi-Arid Wheat Screening Nursery
Table 1. Broad-sense heritability of agronomic traits evaluated in Cycle 29 SAWSN.

<table>
<thead>
<tr>
<th>Trait</th>
<th>( H^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (Irrigated)</td>
<td>0.62</td>
</tr>
<tr>
<td>Yield (Drought)</td>
<td>0.62</td>
</tr>
<tr>
<td>Thousand kernel weight</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Table 2. Prediction accuracy for agronomic traits in the Cycle 29 SAWSN using different marker imputation methods. Cross-validation with 7-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.

<table>
<thead>
<tr>
<th>Trait</th>
<th>MVN-EM†</th>
<th>RF</th>
<th>mean</th>
<th>het</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (Irrigated)</td>
<td>0.32</td>
<td>0.32</td>
<td>0.28</td>
<td>0.33</td>
</tr>
<tr>
<td>Yield (Drought)</td>
<td>0.42</td>
<td>0.40</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>Thousand kernel weight</td>
<td>0.33</td>
<td>0.34</td>
<td>0.38</td>
<td>0.36</td>
</tr>
<tr>
<td>Days to heading</td>
<td>0.35</td>
<td>0.36</td>
<td>0.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

† Imputation methods are: Multivariate Normal Expectation Maximization (MVN-EM), Random Forest (RF); mean; and heterozygote.
Table 3. Prediction accuracy and bias for simulated phenotypes.

<table>
<thead>
<tr>
<th></th>
<th>Complete</th>
<th>MVN-EM</th>
<th>Mean</th>
<th>Het</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.477^a</td>
<td>0.466^b</td>
<td>0.457^c</td>
<td>0.444^d</td>
</tr>
<tr>
<td>Regression Coefficient ‡</td>
<td>1.03 (0.02)</td>
<td>0.99 (0.02)</td>
<td>1.19 (0.02)</td>
<td>1.31 (0.03)</td>
</tr>
</tbody>
</table>

^ Imputation methods are: Complete genotypic data; Multivariate Normal Expectation Maximization (MVN-EM); mean; and heterozygote. Significant differences between imputation methods are shown by superscript letters.

‡ Regression of true breeding value on genomic estimated breeding value.
Table 4. Cross validation prediction accuracy for genomic selection models using genotyping-by-sequencing (GBS) and DArT markers on 254 elite breeding lines from the Cycle 29 Semi-Arid Wheat Screening Nursery, CIMMYT. Superscript letters denotes significant differences within traits.

<table>
<thead>
<tr>
<th>Trait</th>
<th>DArT 2K†</th>
<th>GBS 2K‡</th>
<th>GBS 35K§</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield (Irrigated)</td>
<td>0.13ⁱᵇ</td>
<td>0.25ᵃᵇ</td>
<td>0.32ᵃ</td>
</tr>
<tr>
<td>Yield (Drought)</td>
<td>0.18ⁱᵇ</td>
<td>0.35ᵃ</td>
<td>0.42ᵃ</td>
</tr>
<tr>
<td>Thousand kernel weight</td>
<td>0.28ᵃ</td>
<td>0.26ᵃ</td>
<td>0.33ᵃ</td>
</tr>
<tr>
<td>Days to heading</td>
<td>0.20ⁱᵇ</td>
<td>0.34ᵃ</td>
<td>0.33ᵃ</td>
</tr>
</tbody>
</table>

†Number of markers: 1,729 for DArT; 1,827 for GBS (2K); 34,749 for GBS (35K)
‡ markers with up to 20% missing data
§ markers with up to 80% missing data
A

SNP

error

<table>
<thead>
<tr>
<th>Tag count</th>
<th>SNP sequence</th>
<th>Allele G</th>
<th>Allele C</th>
</tr>
</thead>
<tbody>
<tr>
<td>120</td>
<td>TGCAGTATGACTGCGATGCGGATTGCA</td>
<td>x x</td>
<td>x x</td>
</tr>
<tr>
<td>55</td>
<td>........G...........C...........</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>10</td>
<td>........T...........C...........</td>
<td></td>
<td>x x</td>
</tr>
<tr>
<td>5</td>
<td>........T...........A...........</td>
<td></td>
<td>x x</td>
</tr>
</tbody>
</table>

B

Allele G

<table>
<thead>
<tr>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>128</td>
<td>12</td>
</tr>
</tbody>
</table>

Allele C

<table>
<thead>
<tr>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>163</td>
<td>12</td>
</tr>
</tbody>
</table>

p-val<1E-16

C

Allele G

<table>
<thead>
<tr>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Het</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>-</td>
</tr>
</tbody>
</table>

Individual

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>...</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>...</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>