Jarquín et al. BMC Genomics 2014, 15:740
http://www.biomedcentral.com/1471-2164/15/740

BMC Genomics

RESEARCH ARTICLE Open Access

Genotyping by sequencing for genomic prediction in a soybean breeding population

Diego Jarquín[1], Kyle Kocak[1], Luis Posadas[1], Katie Hyma[2], Joseph Jedlicka[1], George Graef[1] and Aaron Lorenz[1*]

Abstract

Background: Advances in genotyping technology, such as genotyping by sequencing (GBS), are making genomic prediction more attractive to reduce breeding cycle times and costs associated with phenotyping. Genomic prediction and selection has been studied in several crop species, but no reports exist in soybean. The objectives of this study were (i) evaluate prospects for genomic selection using GBS in a typical soybean breeding program and (ii) evaluate the effect of GBS marker selection and imputation on genomic prediction accuracy. To achieve these objectives, a set of soybean lines sampled from the University of Nebraska Soybean Breeding Program were genotyped using GBS and evaluated for yield and other agronomic traits at multiple Nebraska locations.

Results: Genotyping by sequencing scored 16,502 single nucleotide polymorphisms (SNPs) with minor-allele frequency (MAF) > 0.05 and percentage of missing values ≤ 5% on 301 elite soybean breeding lines. When SNPs with up to 80% missing values were included, 52,349 SNPs were scored. Prediction accuracy for grain yield, assessed using cross validation, was estimated to be 0.64, indicating good potential for using genomic selection for grain yield in soybean. Filtering SNPs based on missing data percentage had little to no effect on prediction accuracy, especially when random forest imputation was used to impute missing values. The highest accuracies were observed when random forest imputation was used on all SNPs, but differences were not significant. A standard additive G-BLUP model was robust; modeling additive-by-additive epistasis did not provide any improvement in prediction accuracy. The effect of training population size on accuracy began to plateau around 100, but accuracy steadily climbed until the largest possible size was used in this analysis. Including only SNPs with MAF > 0.30 provided higher accuracies when training populations were smaller.

Conclusions: Using GBS for genomic prediction in soybean holds good potential to expedite genetic gain. Our results suggest that standard additive G-BLUP models can be used on unfiltered, imputed GBS data without loss in accuracy.

Background

Marker-assisted selection (MAS) has played an important role in soybean breeding, particularly for traits that are difficult to evaluate phenotypically such as soybean cyst nematode (SCN) resistance. An early demonstration of successful MAS for SCN resistance allowed accurate identification of resistant lines using microsatellite markers [1]. Use of MAS for improving grain yield, however, has been met with limited success in soybean. A host of QTL mapping studies have reported QTL for grain yield in exotic soybean populations [2-5], but introgression of yield QTL has not been consistent across different genetic backgrounds [6]. Moreover, the QTL mapping – introgression approach is difficult to justify unless large effect QTL alleles are identified, which is rarely the case for grain yield.

Sebastian et al. [7] reported on a MAS approach that involved the sub-lining of existing soybean elite cultivars derived from single F3 or F4 plants. The authors called this approach context-specific MAS. Essentially, MAS is performed within narrow populations (i.e., elite cultivars with residual heterogeneity) with the goal of obtaining more precise estimates of genetic value in early field trials consisting of only a single replication. Lines were selected and advanced for further testing on the basis of marker scores calculated using significant marker effects estimated within populations. Significant superiority in grain yield for some of the selected sublines,

* Correspondence: alorenz2@unl.edu
[1]Department of Agronomy and Horticulture, University of Nebraska, 363 Keim Hall, Lincoln, NE 68583, USA
Full list of author information is available at the end of the article

relative to their "mother lines" (i.e., the elite cultivars with residual heterogeneity), was demonstrated using this approach [7].

This common-sense approach is ideal for increasing accuracy of preliminary yield tests: marker effects are estimated more accurately because marker alleles are highly replicated across individuals comprising a large population, whereas phenotypic values are estimated using only a single observation. Marker effects, however, can only be applied for calculating marker scores within a single biparental population and therefore, the current generation. It would be desirable to predict breeding values after one or more generations of recombination and selection in order to facilitate rapid cycling of parents. Furthermore, pooling genotypic and phenotypic information across populations could allow for more populations to be evaluated for grain yield within the same field space. Fewer individuals could be phenotypically evaluated in each biparental population and the breeding value of remaining (non-evaluated) individuals could be predicted using markers. Current trends strongly indicate plant breeding programs will be limited by phenotyping capacity, not genotyping capacity, thus increasing the attractiveness of this strategy through time.

Genomic selection (GS) has become the predominant method of applying molecular markers for selection of complex traits in plant breeding programs [8]. Briefly, genomic selection entails building a prediction model through associating marker information with phenotypic information in a "model training" step. Individuals that have been genotyped and phenotyped comprise the "training population" or "calibration set". The prediction model is applied to a set of selection candidates that have been genotyped but not evaluated phenotypically. The primary difference between GS and traditional forms of MAS is that QTL mapping is not performed and markers are not chosen for inclusion in the model based on a statistical analysis, but rather all marker information is used simultaneously. The types of models used to deal with the "large p, small n" problem created by the genomic approach to prediction have been reviewed and compared elsewhere [8-10].

Dramatic advances in sequencing technologies are providing highly dimensional molecular marker information at low cost. Genotyping by sequencing [11] is a method well described by its name: polymorphisms are scored using next-generation sequencing technologies followed by a bioinformatics pipeline. The advantage of GBS is that it reduces cost through an enzyme-based genomic complexity reduction step and the use of barcoded adapters for multiplexing [12]. Genotyping by sequencing has been applied to investigations of genetic diversity in maize [13] as well as to studies on GS [14-16]. Working in soybean, Sonah et al. [17] developed a novel GBS protocol and reliably called 10,120 high-quality SNPs among eight diverse lines. These authors called high-quality SNPs displaying only a small percentage of missing data, whereas many applications of GBS [16] have tolerated SNPs with very high frequencies of missing data, sometimes up to 80%.

Given this high rate of missing values in GBS data, imputation of marker scores is typically performed. The best imputation method, and whether imputated GBS data provides better predictions than simply selecting SNPs with low rates of missing data, however, is not known. Rutkoski et al. [18] showed a slight advantage to using imputation when markers were used with high missing data rates. In maize, however, Crossa et al. [16] failed to find any improvement in prediction accuracy using a haplotype-based imputation method on GBS data. Poland et al. [14] showed that a random forest imputation method provided the most accurate imputations, but the effect on genomic prediction accuracy was not significant.

A large number of studies on GS in multiple crops has been reported [2,19-22]. A study on GS in soybean, however, has not. Moreover, there are only a few reports on the use of GBS for GS [14-16]. In light of the current research on GS and the dearth of reported research on GS and GBS in soybean breeding, the objectives of this study were (i) evaluate prospects for GS using GBS in a typical soybean breeding program and (ii) evaluate the effect of GBS marker selection and imputation on genomic prediction accuracy. To achieve these objectives, a set of soybean experimental lines sampled from the University of Nebraska-Lincoln Soybean Breeding Program were genotyped using GBS and evaluated for grain yield and other agronomic traits at multiple Nebraska locations. Reported findings are important to the application of GBS to selection for grain yield in future soybean breeding efforts.

Methods

Germplasm and phenotyping

Three hundred and one soybean experimental lines currently in advanced stages of the University of Nebraska-Lincoln Soybean Breeding Program were sampled. Two hundred and seventy-five lines were in the $F_{5:8}$ generation and twenty-six lines were in the $F_{4:7}$ generation. Soybean lines belonged to maturity groups I (N = 64), II (N = 213), and III (N = 24) (Table 1) and represent 34 biparental families ranging in size from one to 28 lines per family. Median family size was 8.

During the summer of 2011, soybean lines were grown in two-row plots (0.76 m apart, 2.9 m long) seeded to a density of 26 seeds per square meter. Plots were arranged in an augmented incomplete block design with two replications. Blocks consisted of 27 – 39 experimental entries

Table 1 Number of lines belonging to each maturity group (MG) and grown at each Nebraska location

	Beemer	Phillips	Cotesfield	Mead	Lincoln	Clay center
MG 1	64	64	64	64	0	0
MG 2	213	213	213	213	0	0
MG 3	0	24	0	24	24	24
Total	277	301	277	301	24	24

and three check cultivars. Lines belonging to maturity groups I and II were evaluated at the Nebraska locations Beemer, Phillips, Cotesfield, and Mead. Lines belong to maturity group III were evaluated at the Nebraska locations Phillips, Mead, Lincoln, and Clay Center (Table 1). Grain yield (GY; Mg ha^{-1}) was measured at all locations, plant height (PH; cm) was measured only at Mead, and days to maturity (MD) was measured at Beemer, Phillips and Mead. Grain yield was recorded as machine harvestable grain yield adjusted to 13% moisture. Plant height was measured as the distance (in centimeters) between the surface of the soil and the main-stem apical meristem. Days to maturity was defined as the number of days from planting until the R8 stage when 95% of the pods were mature and brown in color.

Phenotypes were adjusted to remove location and block effects according to the model:

$$y_{ijkl} = \mu + g_i + e_j + r_{k(j)} + b_{l(k)} + ge_{ij} + \varepsilon_{ijkl}$$

where g_i represents the effect of the i^{th} genotype (i.e., soybean line), e_j represents the effect of the j^{th} location, $r_{k(j)}$ represents the effect of the k^{th} replicate nested in location j, $b_{l(k)}$ represents the effect of the l^{th} incomplete block nested within replicate k, and. Best linear unbiased estimates were calculated for soybean lines and input into the genomic prediction models described below. The genotype effect was also treated as random in order to estimate variance components for the purpose of estimating heritability. Broad-sense heritability (H^2) on an entry-mean basis was calculated as $H^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_{GE}/_{er} + \sigma^2_\varepsilon/_r}$ where σ^2_G is the variance among soybean lines, σ^2_{GE} is the genotype-by-environment interaction variance, σ^2_ε is the residual variation, and e and r in this context are the number of environements and replications within environments, respectively.

Genotyping

Leaf discs were collected from 12 random plants of each soybean line at approximately the V6 growth stage. DNA isolation was performed using the Qiagen DNeasy Plant 96 kit. Samples were sent to the Institute of Genomic Diversity at Cornell University for genotyping by sequencing as described in [11] and at www.maizegenetics.net/gbs-overview. Briefly, DNA samples were digested with the ApeKI restriction enzyme followed by ligation of adapters to fragment ends. Adapters consisted of Illumina sequencing primers and a barcode adapter. After adapter ligation, samples are combined into pools consisting of 96 samples. A PCR amplification is carried out to create the GBS libraries, which are submitted to a single Illumina HiSeq2000 flow cell for sequencing. Four sequenced libraries produced on average 247,255,883 reads, of which on average 219,580,690 were good, barcoded reads.

The GBS analysis pipeline as implemented in Tassel Version 3.0.156 was used to call SNPs. Briefly, 1) tag counts were generated from fastq files with the FastqToTagCountPlugin (options: –s 300000000, –c 1), 2) tag counts were merged with the MergeMultipleTagCountPlugin (options: –c 5), 3) tags were aligned to the reference genome Gmax_109_softmasked.fa.gz, which was downloaded from ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Gmax/assembly/ on 16 July 2012 and indexed for use with bwa version 0.6.1-r104. BWA version 0.7.3a-r367 was used for alignment (–n 0.04). Chromosomes were renamed for compatibility with the GBS pipeline by removing the leading 'Gm' and 'scaffold_', and then converted to the tags-on-physical-map format using the SAMConverterPlugin 4). Counts of tags per individual (taxa) were generated with the FastqToTBTPlugin (options: –c 1 –s 300000000, –y), 5) Counts of tags per individual were merged with the MergeTagsByTaxaFilesPlugin (options: –s 300000000, –x), 6) SNPs were called using the TagsToSNPByAlignmentPlugin (options: –mnMAF 0.01, –mnLCov 0.1, –mnMAC 10, –mxSites 1,000,000). Duplicate sites were merged with the MergeDuplicateSNPsPlugin (options: –callHets, –misMat 0.05), and duplicated taxa were merged with the MergeIdenticalTaxaPlugin (options: –hetFreq 0.8). Scaffolds were ignored for SNP calling.

Genomic prediction models

Base pair calls contained in the hapmap file obtained from the GBS analysis pipeline were converted to numerical genotype scores, $x \in \{0, 1, 2\}$, where x is the number of copies of the major allele.

Two genomic prediction models were studied: a standard G-BLUP model including only additive effects, and an extended version of the G-BLUP model also including additive-by-additive effects. Two different formulations of additive-by-additive effects have been presented in the literature [23,24] and both of them were considered.

The standard G-BLUP model including additive effects only is

$$y_i = \mu + g_i + \varepsilon_i \quad (1)$$

where y_i represents the phenotype of the i^{th} line, μ represents the intercept, g_i represents the additive genetic value, and e_i represents the residual. The additive genetic value can be estimated as $g_i = \sum_{j=1}^{p} x_{ij} b_j$, where x_{ij} is the genotype score at the j^{th} ($j = 1,\ldots,p$) locus in the i^{th} ($i = 1,\ldots,n$) line, and b_j is the allelic substitution effect (marker effect) at the j^{th} marker locus. Marker effects were considered as *IID* random variables from a normal distributions such that $b_j \overset{IID}{\sim} N(0, \sigma_b^2)$. From properties of the multivariate normal distribution the vector $\mathbf{g} = \mathbf{Xb}$, ($\mathbf{g} = [g_1, \ldots, g_n]'$), follows a multivariate normal distribution with null mean and covariance matrix $Cov(\mathbf{g}) = \mathbf{XX}'\sigma_b^2 = \mathbf{G}\sigma_g^2$ where $\mathbf{G} = \frac{1}{p}\mathbf{XX}'$ and $\sigma_g^2 = p\sigma_b^2$. Hereafter, we centered and standardized genotype scores by dividing by $\sqrt{2\theta_j(1-\theta_j)}$, where θ_j is the estimated allele frequency at the j^{th} marker locus. The **G** matrix is a genomic realized relationships matrix whose entries are given by $G_{ii'} = \frac{\sum_{j=1}^{p}(x_{ij}-2\theta_j)(x_{i'j}-2\theta_j)}{\sum_{j=1}^{p}2\theta_j(1-\theta_j)}$. Summarizing above stated assumptions the model (1) becomes a mixed effects model with $\underline{\mathbf{g}} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\varepsilon_i \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$. Using this model, the lines are related through the off-diagonal values of **G** matrix, allowing the borrowing of information between lines to predict performance of lines not phenotyped.

Additive-by-additive effects were modeled using two different covariance structures among lines. Several authors [25,26] modeled additive-by-additive epistasis through a $\mathbf{G} \circ \mathbf{G}$ matrix following Cockerham (1954) and Kempthorne (1954), where $\circ$ represents the Hadamard, or element-wise, multiplication operation. The first model including additive-by-additive epistasis was

$$y_i = \mu + gg_i + \varepsilon_i \quad (2)$$

with $\underline{\mathbf{gg}} \sim N(\mathbf{0}, \mathbf{G}\circ\mathbf{G}\sigma_{gg}^2)$ and $\varepsilon_i \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$.

More recently, Xu [24] proposed an alternative way to include these interaction effects using the covariance structure given by $\mathbf{K}_{aa} = \frac{1}{C_{aa}}\mathbf{K}^*_{aa}$, with $\mathbf{K}^*_{aa} = \sum_{j=1}^{p}\sum_{j'=k+1}^{p-1}(Z_k \circ Z_{k'})(Z_k \circ Z_{k'})'$, $C_{aa} = mean[diag(\mathbf{K}^*_{aa})]$ and Z_k is the j^{th} marker locus such that $Z_{ij} = \begin{cases} +1 & \text{for A} \\ 0 & \text{for H} \\ -1 & \text{for B} \end{cases}$. Using this assumption a different version of the common epistasis model is given by:

$$y_i = \mu + k_i + \varepsilon_i \quad (3)$$

with $\underline{\mathbf{k}} \sim N(\mathbf{0}, \mathbf{K}\sigma_{aa}^2)$ and $\varepsilon_i \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$.

Modeling additive and additive-by-additive components was conducted to assess the proportion of the phenotypic variance accounted for by these effects and improvements in accuracy of genomic prediction. By combining models (1) and (2), a model including additive and epistatic effects was formulated:

$$y_i = \mu + g_i + gg_i + \varepsilon_i \quad (4)$$

with $\underline{\mathbf{g}} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$, $\underline{\mathbf{gg}} \sim N(\mathbf{0}, \mathbf{G}\circ\mathbf{G}\sigma_{gg}^2)$ and $\varepsilon_i \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$.

The alternative additive and additive-by-additive model following Xu [24] was built combining models (1) and (3):

$$y_i = \mu + g_i + k_i + \varepsilon_i \quad (5)$$

with $\underline{\mathbf{g}} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$, $\underline{\mathbf{k}} \sim N(\mathbf{0}, \mathbf{K}\sigma_{aa}^2)$ and $\varepsilon_i \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$.

Models (1)-(5) were fitted to the full data set using computational methods described in [27]. All the statistical analyses were implemented in the R statistical software [28].

Marker imputation

Genotyping-by-sequencing data sets typically have high rates of missing data [14,16]. Three imputation methods were considered to impute missing values of the soybean GBS data set. (*i*) Naïve imputation substitutes missing values at each locus with $2\theta_j$, where θ_j is the estimated frequency of the major allele at the j^{th} locus. This method is not expected to add information, but rather serves the purpose of ensuring unchanged allele frequencies after imputation and provides a marker matrix containing no missing data so that analytical operations can be performed. (*ii*) Random forest imputation is based on random forest regression introduced by Breiman [29]. Marker imputation for this study was performed using the MissForest R package according to [18]. The algorithm was performed chromosome-wise and for each PMV and MAF combination in parallel. (*iii*) FILLIN (HI) is a novel imputation method based on haplotypes, which is implemented in TASSEL 5.0. Default settings were used with the exception of maximum heterozygosity, which was set to 0.30 using the option -mxHet. Detailed information can be found in the TASSEL 5.0 User Guide at www.maizegenetics.net.

Varying factors affecting genomic prediction accuracy

To evaluate the effects of GBS marker selection and imputation methods on genomic prediction accuracy, two criteria for filtering SNPs were considered. Filtering of GBS SNPs was done sequentially, first filtering based on percent missing values (PMV), and then, for minor-allele frequency (MAF). Levels for PMV (27) and MAF (12) were (*l*) 1–20, 25, 30, 40, 50, 60, 70, and 80%, and (*m*) 0.05-0.1, 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40, respectively. Markers were filtered based on all possible

combinations of PMV and MAF. After filtering, remaining missing values were imputed using each of the three imputation methods described above. This produced 972 marker datasets (e.g., 27 PMV levels × 12 MAF levels × 3 imputation methods).

All comparisons were made on the basis of the correlation between the observed phenotype and the predicted breeding value, which is referred to as *predictive ability*, following [30]. We reserve the term *prediction accuracy* (r_{gp}) for the correlation between the prediction and the true breeding value. Prediction accuracy can be approximated by dividing predictive ability by $\sqrt{H^2}$ [8,31]. Predictive ability ($r_{\hat{g}p}$) of each marker filtering criteria was evaluated using 10-fold cross validation replicated 200 times. The mean predictive ability across the 200 replicates was calculated and bootstrap confidence intervals.

The impact of training population size on prediction accuracy was evaluated using a validation set comprised of 50 randomly selected lines and training sets of variable sizes. The validation set was formed by randomly sampling 50 lines without replacement. From the remaining 251 lines, the training population of size n was formed sequentially such that its size ranged between 2 and 251. First, two lines were sampled (i.e., $n = 2$) without replacement, then, from the remaining 251-n lines, additional lines were incorporated to the training set, by increments of one. Once a line was sampled, it remained in the training set. The validation set was held constant with the initial 50 lines. Two GBS marker subsets were used to evaluate training population size effect: 1) PMV ≤ 5% and MAF > 0.05; and 2) PMV ≤ 80% and MAF > 0.3. This procedure was repeated 1000 times and accuracies at each training population size were averaged across replicates.

Results

A total of 5,770,366 unique 64-bp sequence tags were identified across all four soybean libraries, of which 68.75% aligned uniquely to the reference genome, 11.32% aligned to multiple positions and 19.92% could not be aligned. The mean (median) sequencing depth per SNP locus was 11 (6), with mean (median) proportion "missingness" per SNP locus of 0.18 (0.08).

Unique tag counts and SNP density were higher towards the chromosome ends compared to pericentromeric regions (Figure 1). The GBS protocol targets gene-rich regions, such as the chromosome ends, through the use of methylation sensitive restriction enzymes. Related to the distribution of uniqe tag counts, percent missing values were lower towards chromosome ends and higher towards the pericentromeric regions. There was no apparent pattern regarding MAF with the exception of some chromosomes harboring more diversity than others (e.g., chromosomes 11 and 20 versus chromosomes 15 and 18) (Figure 1). The number of SNPs remaining after filtering by MAF and PMV is shown in Figure 2. The number of SNPs available with cuttoff values of PMV ≤ 80% and MAF > 0.05 was 52,349. There were 16,502 SNPs with PMV ≤ 5% and MAF > 0.05.

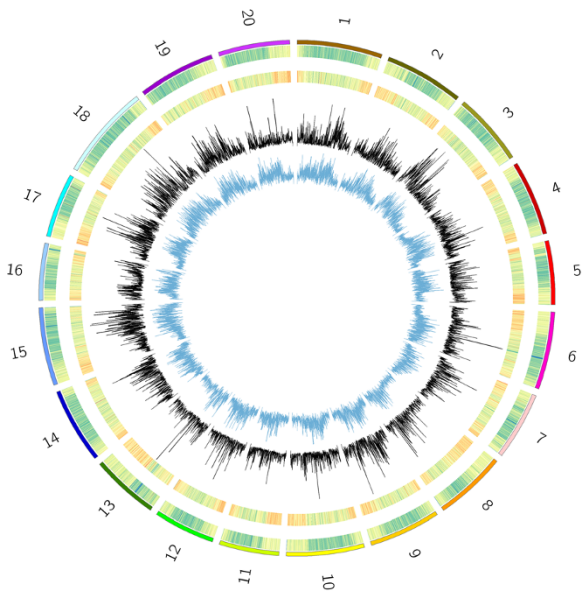

Figure 1 Genoptyping by sequencing parameters on 301 elite soybean breeding lines. Parameters were calculated using a 100 kbp window with a 50 kbp slide. From outside to inside: 1) Unique 64-bp sequence tags per window; 2) SNP density; 3) Minor-allele frequency; 4) Percent missing values. For the 64-bp sequence tag and SNP density heatmaps, hot colors represent larger values on a log base three scale.

The high quality of the phenotypic data was reflected with relatively high heritabilities of 0.69 for GY and 0.94 for MD (Table 2). As expected, the genotype-by-environment interaction source of variation was greater for GY compared to MD. The overall $r_{\hat{g}p}$ of G-BLUP using a SNP dataset with cuttoff values of PMV ≤ 80% and MAF ≥ 0.05, and the Naïve imputation method, was 0.565 for GY, 0.374 for PH, and 0.644 for MD. Prediction accuracy estimates for GY, PH, and MD were, 0.64, 0.42 and 0.65, respectively.

The effect of SNP filtering on $r_{\hat{g}p}$ was assessed by building a series of G-BLUP models using SNP datasets created by applying combinations of MAF and PMV filtering criteria. Number of SNPs is quickly reduced as SNPs are filtered based on MAF and PMV (Figure 2). Overall, marker filtering criteria did not have a large effect on $r_{\hat{g}p}$ for GY, but some important effects were observed for PH and MD (Figure 3). For PH, $r_{\hat{g}p}$ was

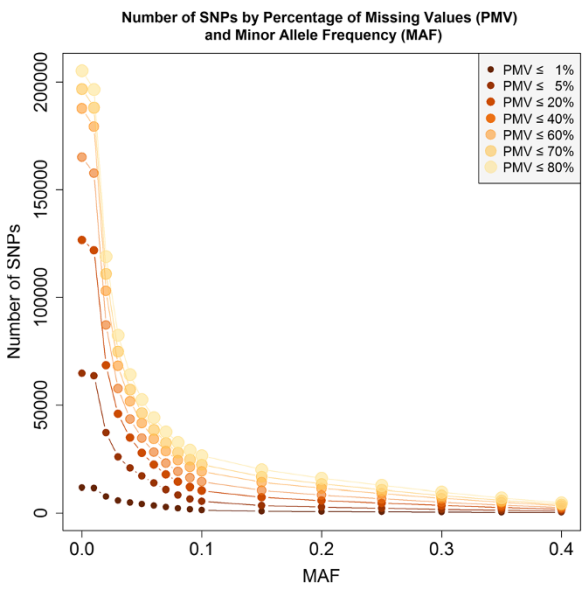

Figure 2 Number of SNPs remaining after applying filtering by combinations of minor-allele frequency and percent missing values.

greater when markers with MAF between 0.08 and 0.10 were used compared to all other MAF cutoffs. When considering jointly both filtering criteria, the $r_{\hat{g}p}$ of a trait were maximized at different combinations between PMV and MAF. For GY the maximum $r_{\hat{g}p}$ (0.59) was obtained with a marker dataset including SNPs with up to 80 PMV and MAF greater than 0.30. For PH and MD, $r_{\hat{g}p}$ was maximized when only SNPs with lower PMV were included (Figure 3).

Imputation

No significant differences in $r_{\hat{g}p}$ were found among the imputation methods (Figure 4). When Naïve imputation was used, $r_{\hat{g}p}$ was slightly reduced by including SNPs with high levels of PMV. When random forest imputation was used, however, $r_{\hat{g}p}$ was maximized when all SNPs were included in the model. A random forest imputation with 80 PMV provided the highest $r_{\hat{g}p}$ overall (Figure 4). The random forest method provided numerically higher $r_{\hat{g}p}$ than the HI method at high PMV levels, but differences were not statistically significant.

Model comparison

Contribution of polygenic additive-by-additive epistatic interactions to total phenotypic variation was assessed by constructing epistatic relationship matrices using the Hadamard product of the additive relationship matrix [23], as well as the marker-generated additive-by-additive relationship matrix as described by Xu [24]. For these comparisons, a marker set including SNPs with PMV ≤ 80% and MAF > 0.05 was used. Missing values were imputed using the Naïve method.

The percentages of phenotypic variation accounted for by each model term varied across traits. For GY, the realized additive relationship matrix captured 91.2% of total phenotypic variance. Since the $\mathbf{G}^{\circ}\mathbf{G}$ and $\mathbf{K}_{aa}$ matrices are highly collinear with $\mathbf{G}$ (data not shown), the epistatic relationship matrices accounted for similar amounts of variation (Table 3). When combining both additive and epistatic effects into the same model, the epistatic effects account for variable amounts of phenotypic variation. Nevertheless, the percentage of residual variation is fairly constant (Table 3), indicating that including an additive-by-additive epistasis relationship matrix provides no improvement over standard additive G-BLUP models. This was also observed using the cross-validation approach to evaluate $r_{\hat{g}p}$. No difference was observed among the models regarding $r_{\hat{g}p}$ (Table 4).

Training population size

For a set of SNPs with PMV ≤5% and MAF > 0.05, $r_{\hat{g}p}$ plateaued around a training population size just greater than 100 (Figure 5). Predictive ability, however, did steadily increase up until the maximum training population size possible in the cross validation strategy. Predictive ability was improved at lower TP sizes only when a MAF > 0.30 was used to construct $\mathbf{G}$. For example, using a MAF > 0.05 and TP size of 50, $r_{\hat{g}p}$ was only 0.28,

Table 2 Summary of phenotypic data analysis for grain yield (GY), plant height (PH) and days to maturity (MD)

Trait	Units	Mean	SD[†]	Range	Variance component[‡]			H^2
					G	G × E	Residual	
GY	Mg ha^{-1}	4505	377.3	2836–5624	12.9	7.28	31.4	0.69
PH	cm	100.4	11.28	61.00–121.9	67.0	NA[§]	33.0	0.80
MD	days	134	4.07	121–141	76.3	5.49	8.94	0.94

[†]Standard deviation.
[‡]G, soybean genotype; GxE, genotype-by-environment interaction.
H^2 – Broad-sense heritability on an entry-mean basis.
[§]Plant height was measured at only one location.

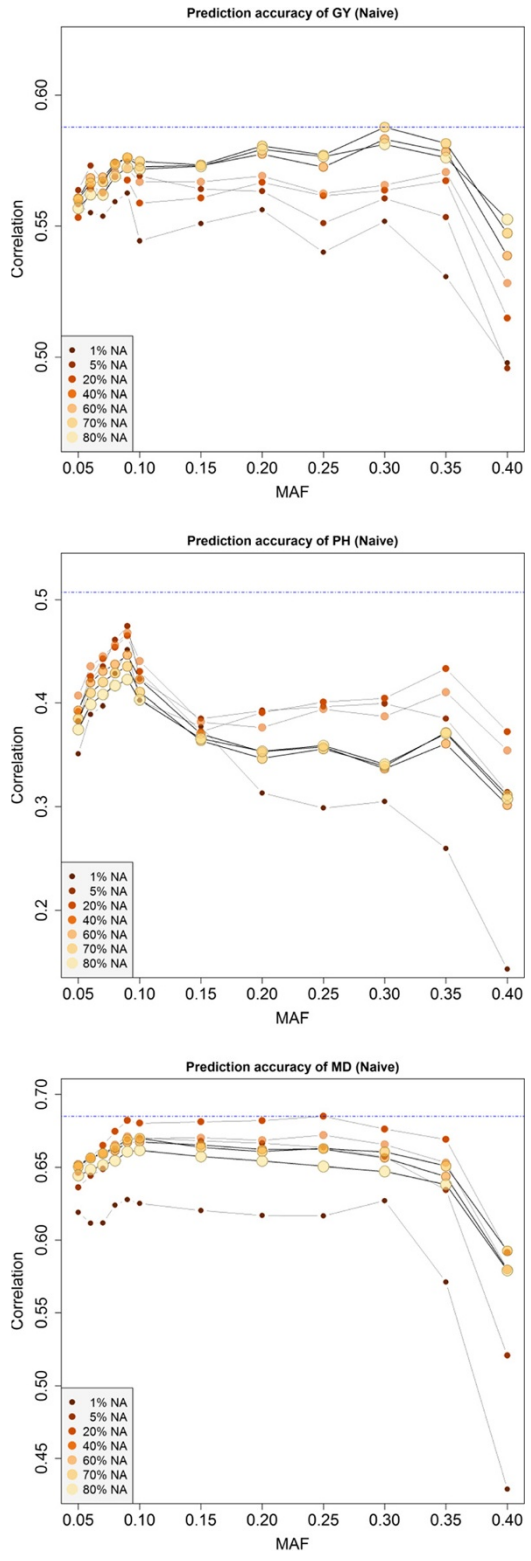

Figure 3 Average predictive abilities (y-axis) at each combination of minor-allele frequency (MAF) and percent missing value (PMV) for grain yield (GY), plant height (PH) and days to maturity (MD). Naïve imputation was used to fill in missing values.

but when a MAF > 0.30 was used, $r_{\hat{g}p}$ was increased to 0.47. A TP consisting of at least 100 individuals was required to reach this $r_{\hat{g}p}$ using SNPs with MAF > 0.05 (Figure 5).

Discussion

The results of this study suggest that the use of GBS for genomic selection holds good potential for improving soybean grain yield. Using a cross-validation approach, genomic predictions explained ~32% of the variation in yield phenotypes. After using the phenotype heritability to correct for random environmental deviations included in the validation phenotypes, approximately 41% of the variation in genotypic values was explained by genomic predictions. Because validation phenotypes (i.e., soybean line means) include both additive and non-additive effects, and genomic predictions using the G-BLUP model include only additive effects, this estimated prediction accuracy is likely biased downward.

In order to quantify the relative benefit of genomic selection over phenotypic selection, Technow et al. [32] rearranged the formula for relative response to indirect selection to obtain the inequality $L_Y < \frac{r_A}{H_X} L_X$, where L_Y is the cycle length of genomic selection, r_A is the genomic prediction accuracy, H_X is the phenotypic selection accuracy, and L_X is the cycle length of phenotypic selection [32]. Substituting the values estimated herein for grain yield into this formula indicates that genomic selection is expected to be superior to phenotypic selection in terms of genetic gain per unit time if the cycle length of genomic selection is less than 77% the cycle length of phenotypic selection. It is entirely possible for a genomic selection cycle to be 66% of a phenotypic selection cycle based on the structure of the University of Nebraska Soybean Breeding Program. On top of this, the above formula assumes equal selection intensities for genomic and phenotypic selection. As genotyping costs continue to decline, selection intensity for genomic prediction could be increased compared to phenotypic selection at equal cost. Before soybean breeding programs incorporate genomic selection on a large scale, these results need to be validated through comparisons of phenotypes and genomic predictions across years, as well as by comparison of progenies from phenotypic and genomic selection programs.

The high genomic prediction accuracy observed was fairly consistent across SNP datasets with differing levels of PMV. More than 16,000 SNPs were scored with less than 5 PMV using GBS, which is surely more than is needed to ensure good SNP-QTL linkage disequilibrium among elite soybean breeding progenies [33]. Little to no sacrifice in accuracy was observed when SNPs with up to 80 PMV were included. It might be desirable to

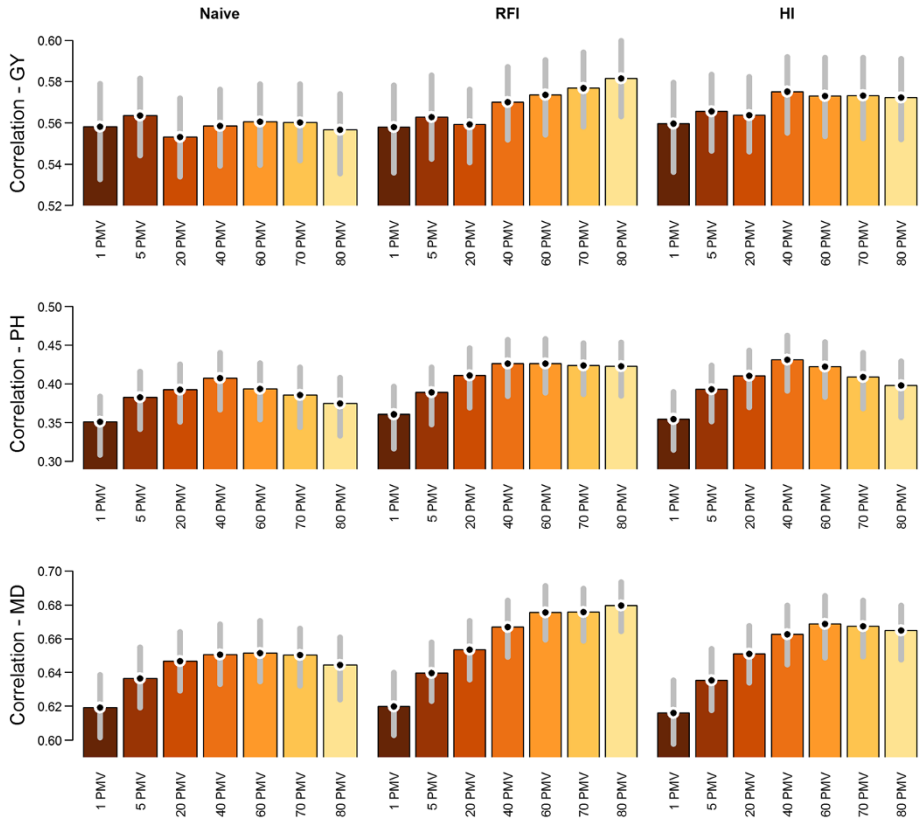

Figure 4 Average predictive ability and corresponding 95% bootstrap confidence intervals for multiple levels of percent missing values (PMV) and three imputation methods: Naïve, random forest imputation (RFI), and haplotype-based imputation (HI).

reduce the SNP numbers to ease computational requirements when predicting individual SNP effects and summing effects to calculate genomic predictions. However, more saturated SNP datasets may be more desirable for computing genomic predictions of multi-family selection schemes of more diverse germplasm. The G-BLUP approach is more computationally efficient with computational demands scaling with individual number rather than marker number. Knowing that data filtering steps are not likely needed for using GBS data for genomic prediction reduces the number of optimization steps and simplifies the process.

We failed to find significant differences among imputation methods, including differences between Naïve imputation and the other two which use covariance information between nearby SNPs. While not significantly better, the machine learning, non-parametric method called random forest performed best when SNPs with up to 80 PMV were included. This was also observed by Rutkoski et al. [18], but these authors did not compare random forest imputation with a method using marker order information. We observed that a haplotype-based method, which used marker order information from the soybean physical map, was not superior to random forest imputation. Random

Table 3 Percentage of phenotypic variation in grain yield (GY), plant height (PH), and days to maturity (MD) explained by additive and non-additive effects included in models 1 – 5

Model	Percentage of phenotypic variance accounted for by each component											
	GY				PH				MD			
	G	G°G	K_{aa}	Res	G	G°G	K_{aa}	Res	G	G°G	K_{aa}	Res
[1] G	91.2			8.8	91.8			8.2	94.2			5.8
[2] G°G		86.7		13.3		86.4		13.6		90.0		10.0
[3] K_{aa}			86.9	13.1			86.3	13.7			90.1	9.9
[4] G_G°G	49.0	39.7		11.3	73.7	16.3		9.9	28.7	62.2		9.1
[5] G_K_{aa}	69.7		19.9	10.4	74.7		15.9	9.5	49.1		42.9	8.0

Table 4 Predictive abilities for grain yield (GY), plant height (PH) and days to maturity (MD) under models [1] – [5]

Model	GY	PH	MD
[1] G	0.60	0.45	0.67
[2] G°G	0.58	0.43	0.68
[3] K_{aa}	0.58	0.43	0.68
[4] G_G°G	0.59	0.45	0.68
[5] G_K_{aa}	0.59	0.44	0.68

forest has often been used for imputing markers for genomic selection in plant breeding, especially when a reference genome is not available [10,15,18]. A haplotype-based method was also used for GBS data by Crossa et al. [16], but these authors also observed very little to no advantage over Naïve imputation. This general lack of benefit to imputation is likely due to the fact the genomic prediction is robust to missing marker data [18] and the number of markers with relatively low PMV provided by GBS is more than enough to cover the linkage disequilibrium space in crop breeding germplasm.

Rather than compare shrinkage models and Bayesian variable selection models for prediction accuracy as has been frequently performed previously (e.g., [10,21]), we compared G-BLUP models including additive effects only against those also including additive-by-additive effects. Additive-by-additive interaction effects were incorporated into the model in the Cockerham-Kempthorn fashion by including a random polygenic interaction effect with a covariance structure specified as the Hadamard product of the additive genomic relationship matrix [34]. This model makes many assumptions, and the soybean population clearly violates the assumptions of linkage equilibrium between loci and randomly mating individuals. Because of this violation, another formulation of the additive-by-additive relationship matrix according to Xu [24] was used. It turned out the $\mathbf{K}_{aa}$ matrix calculated according to Xu [24] was highly collinear with the simple Cockerham-Kempthorn Hadamard product and explained similar amounts of phenotypic variation. Neither $\mathbf{G} \circ \mathbf{G}$ nor $\mathbf{K}_{aa}$ was orthogonal to the $\mathbf{G}$ matrix as can be seen by the variance component estimation. Similar amounts of variation were explained when any of these effects were included in the model alone or together. The amount of residual variation was actually slightly smaller when only $\mathbf{G}$ was modeled and genomic prediction accuracies were not enhanced by including additive-by-additive effects using either the Cockerham-Kempthorn or Xu [24] formulation.

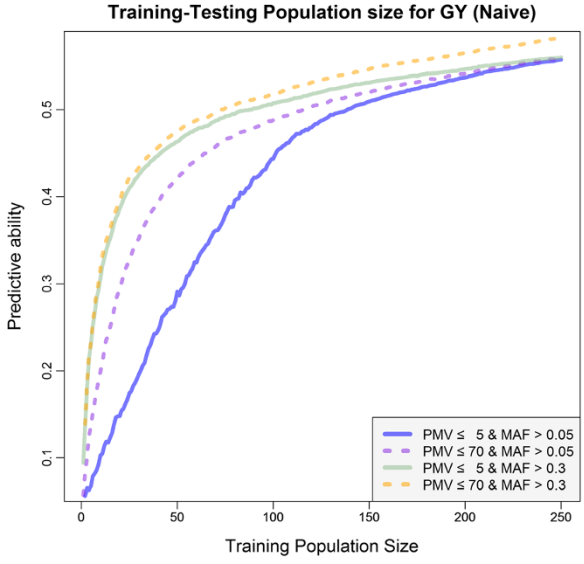

Figure 5 Relationship between predictive ability and training population size for multiple levels of percent missing values (PMV) and minor-allele frequency (MAF). The trait displayed here is grain yield (GY).

Low to moderately sized training populations could be used in a soybean breeding program to achieve adequate prediction accuracies (Figure 5). Although it's probably not necessary to reduce TP sizes down to such a low level, training population sizes could be reduced further if only SNPs with higher MAF are included. The underlying reason for this is not clear. It is possible that SNPs with low MAF are not sampled by chance when small training populations are sampled and phenotyped. If they are not polymorphic in the TP, they cannot contribute information to the relationships between individuals, which is the basis of predictions in G-BLUP. When TP size is increased, SNP alleles with low frequency are more likely to be adequately represented in the TP. When MAF threshold is higher, this problem is reduced by the fact that all SNP alleles have a reasonable chance of contributing to relationship even when TPS sizes are small.

Conclusions

This first look at GBS for genomic prediction in soybean suggests GBS holds good potential to enhance genetic gain in soybean. Over 16,000 SNPs were scored with less than 5% missing data. Filtering markers based on amount of missing data had little to no effect. No differences were observed among imputation methods. The highest accuracies were observed when random forest imputation was used on all SNPs, but differences were not significant. A standard additive G-BLUP model was robust; modeling additive-by-additive epistasis did not provide any improvement in prediction accuracy.

Abbreviations

MAF: Minor-allele frequency; PMV: Percent missing values; RFI: Random forest imputation; MAS: Marker-assisted selection; GS: Genomic selection; G-BLUP: Genomic best linear unbiased prediction; r_{PP}: Predictive ability; GY: Grain yield; PH: Plant height; MD: Days to maturity; SCN: Soybean cyst

nematode; GBS: Genotyping by sequencing; SNP: Single nucleotide polymorphism; HI: Haplotype-based imputation.

Competing interests
The authors declare they have no competing interests.

Authors' contributions
DJ performed the genomic predictions, statistical analysis, and drafted the manuscript. KK participated in the collection of phenotype and genotype data, analyzed the phenotypic data, and participated in drafting the manuscript. LP participated in the collection of genotypic data and edited the manuscript. KH performed the bioinformatics, summarized the SNP data, and edited the manuscript. JJ participated in the collection of phenotype and genotype data. GG managed the data collection, participated in the design of the study, and edited the manuscript. AL participated in the design of the study, managed the statistical analysis, and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The authors are grateful to the Nebraska Soybean Board for providing funding.

Author details
[1]Department of Agronomy and Horticulture, University of Nebraska, 363 Keim Hall, Lincoln, NE 68583, USA. [2]Institute of Genomic Diversity, Cornell University, Ithaca, NY, USA.

Received: 5 June 2014 Accepted: 22 August 2014
Published: 29 August 2014

References
1. Mudge J, Cregan PB, Kenworthy JP, Kenworthy WJ, Orf JH, Young ND: **Two Microsatellite Markers That Flank the Major Soybean Cyst Nematode Resistance Locus.** *Crop Sci* 1997, **37:**1611.
2. Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, Longin FH, Ranc N, Reif JC: **Accuracy of genomic selection in European maize elite breeding populations.** *Theor Appl Genet* 2012, **124:**769–776.
3. Guzman PS, Diers BW, Neece DJ, St. Martin SK, LeRoy AR, Grau CR, Hughes TJ, Nelson RL: **QTL Associated with Yield in Three Backcross-Derived Populations of Soybean.** *Crop Sci* 2007, **47:**111.
4. Orf JH, Chase K, Jarvik T, Mansur LM, Cregan PB, Adler FR, Lark KG: **Genetics of soybean agronomic traits. I. Comparison of three related recombinant inbred populations.** *Crop Sci* 1999, **39**(6):1642–1651.
5. Yuan J, Njiti VN, Meksem K, Iqbal MJ, Triwitayakorn K, Kassem MA, Davis GT, Schmidt ME, Lightfoot DA: **Quantitative trait loci in Two Soybean Recombinant Inbred Line Populations Segregating for Yield and Disease Resistance.** *Crop Sci* 2002, **42:**271–277.
6. Concibido VC, La Vallee B, McLaird P, Pineda N, Meyer J, Hummel L, Yang J, Wu K, Delannay X: **Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars.** *Theor Appl Genet* 2003, **106:**575–582.
7. Sebastian SA, Streit LG, Stephens PA, Thompson JA, Hedges BR, Fabrizius MA, Soper JF, Schmidt DH, Kallem RL, Hinds MA, Feng L, Hoeck JA: **Context-Specific Marker-Assisted Selection for Improved Grain Yield in Elite Soybean Populations.** *Crop Sci* 2010, **50:**1196.
8. Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrels ME, Jannink J-L: **Chapter Two – Genomic Selection in Plant Breeding: Knowledge and Prospects.** *Adv Agron* 2011, **110:**77–123.
9. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL: **Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding.** *Genetics* 2013, **193:**327–345.
10. Heslot N, Rutkoski J, Poland J, Jannink J-L, Sorrells ME: **Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity.** *PLoS One* 2013, **8:**e74612.
11. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species.** *PLoS One* 2011, **6:**e19379.
12. Poland JA, Rife TW: **Genotyping-by-Sequencing for Plant Breeding and Genetics.** *Plant Gen* 2012, **5:**92.
13. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA: **Comprehensive genotyping of the USA national maize inbred seed bank.** *Genome Biol* 2013, **14:**R55.
14. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink J-L: **Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing.** *Plant Gen* 2012, **5:**103.
15. Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch HG, Okechukwu R, Dixon AGO, Kulakow P, Jannink J-L: **Relatedness and Genotype × Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study in Cassava.** *Crop Sci* 2013, **53:**1312.
16. Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, Jannink J-L, Cruz MAL, Babu R: **Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing.** *G3* 2013, **3:**1903–1926.
17. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M, Belzile F: **An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping.** *PLoS One* 2013, **8:**e54603.
18. Rutkoski JE, Poland J, Jannink J-L, Sorrells ME: **Imputation of unordered markers and the impact on genomic selection accuracy.** *G3* 2013, **3:**427–439.
19. Burgueño J, de los Campos G, Weigel K, Crossa J: **Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers.** *Crop Sci* 2012, **52:**707.
20. Heslot N, Yang H-P, Sorrells ME, Jannink J-L: **Genomic Selection in Plant Breeding: A Comparison of Models.** *Crop Sci* 2012, **52:**146.
21. Lorenz AJ, Smith KP, Jannink J-L: **Potential and Optimization of Genomic Selection for Fusarium Head Blight Resistance in Six-Row Barley.** *Crop Sci* 2012, **52:**1609.
22. Massman JM, Jung H-JG, Bernardo R: **Genomewide Selection versus Marker-assisted Recurrent Selection to Improve Grain Yield and Stover-quality Traits for Cellulosic Ethanol in Maize.** *Crop Sci* 2012, **53:**58–66.
23. Cockerham CC: **An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present.** *Genetics* 1954, **39:**859–882.
24. Xu S: **Mapping Quantitative Trait Loci by Controlling Polygenic Background Effects.** *Genetics* 2013, **195:**1209–1222.
25. Henderson CR: *Applications of Linear Models in Animal Breeding.* Ontario, Canada: University of Guelph; 1984.
26. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, Stricker C, Gianola D, Schlather M, Mackay TFC, Simianer H: **Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster.** *PLoS Genet* 2012, **8:**e1002685.
27. de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J: **Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods.** *Genet Res* 2010, **92:**295–308.
28. R Core Team: *A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing; 2013.
29. Breiman L: **Random Forests.** *Mach Learn* 2001, **45:**5–32.
30. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE: **Genomic and metabolic prediction of complex heterotic traits in hybrid maize.** *Nat Genet* 2012, **44:**217–220.
31. Legarra A, Robert-Granié C, Manfredi E, Elsen J-M: **Performance of Genomic Selection in Mice.** *Genetics* 2008, **180:**611–618.
32. Technow F, Bürger A, Melchinger AE: **Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups.** *G3* 2013, **3:**197–203.
33. Hyten DL, Choi I-Y, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB: **Highly Variable Patterns of Linkage Disequilibrium in Multiple Soybean Populations.** *Genetics* 2007, **175:**1937–1944.
34. Gianola D, de los Campos G: **Inferring genetic values for quantitative traits non-parametrically.** *Genet Res* 2008, **90:**525–540.

doi:10.1186/1471-2164-15-740
Cite this article as: Jarquín *et al.*: Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 2014 **15**:740.