G3 Genes | Genomes | Genetics

INVESTIGATION

Generating High Density, Low Cost Genotype Data in Soybean [*Glycine max* (L.) Merr.]

Mary M. Happ, Haichuan Wang, George L. Graef, and David L. Hyten[1]
University of Nebraska-Lincoln, Lincoln, NE 68503
ORCID IDs: 0000-0002-5897-2617 (M.M.H.); 0000-0001-6324-9389 (D.L.H.)

ABSTRACT Obtaining genome-wide genotype information for millions of SNPs in soybean [*Glycine max* (L.) Merr.] often involves completely resequencing a line at 5X or greater coverage. Currently, hundreds of soybean lines have been resequenced at high depth levels with their data deposited in the NCBI Short Read Archive. This publicly available dataset may be leveraged as an imputation reference panel in combination with skim (low coverage) sequencing of new soybean genotypes to economically obtain high-density SNP information. Ninety-nine soybean lines resequenced at an average of 17.1X were used to generate a reference panel, with over 10 million SNPs called using GATK's Haplotype Caller tool. Whole genome resequencing at approximately 1X depth was performed on 114 previously ungenotyped experimental soybean lines. Coverages down to 0.1X were analyzed by randomly subsetting raw reads from the original 1X sequence data. SNPs discovered in the reference panel were genotyped in the experimental lines after aligning to the soybean reference genome, and missing markers imputed using Beagle 4.1. Sequencing depth of the experimental lines could be reduced to 0.3X while still retaining an accuracy of 97.8%. Accuracy was inversely related to minor allele frequency, and highly correlated with marker linkage disequilibrium. The high accuracy of skim sequencing combined with imputation provides a low cost method for obtaining dense genotypic information that can be used for various genomics applications in soybean.

KEYWORDS
imputation
high density SNP data
skim sequencing
low cost genotyping
soybean

Genomics research has yielded a variety of tools which allow for more efficient and precise translation of genetic variation into crop improvements. Panels of single nucleotide polymorphisms (SNPs) obtained through SNP arrays or genotyping-by-sequencing (GBS) are the most common tool used to explore and make associations between genetic and phenotypic variation. Genomics-assisted crop breeding continues to demand increasing densities of genotype information to successfully dissect and predict genetically complex traits (Hamblin *et al.* 2011; Lorenz *et al.* 2011). Current approaches of directly ascertaining a high density of SNP genotype data on large populations are cost prohibitive or fall short of being able capture the maximum amount of genetic space.

doi: https://doi.org/10.1534/g3.119.400093
Manuscript received February 13, 2019; accepted for publication May 1, 2019; published Early Online May 9, 2019.

Supplemental material available at FigShare: https://doi.org/10.25387/g3.7975751.
[1]Correspondence: Department of Agronomy & Horticulture, 322 Keim Hall, 1875 North 38th Street, Lincoln, NE 68503. E-mail: david.hyten@unl.edu

Fixed SNP arrays and GBS are popular options for SNP genotyping in crops. Panels ranging in densities of up to ~600,000 variants are now common in several crop species (Rasheed *et al.* 2017). However, recent genomics studies are utilizing datasets consisting of one million or more markers to answer complex, quantitative genetic questions. The need for this high density of markers is rendering current arrays and GBS approaches inadequate to generate the magnitude of data modern genomic studies require (Tian *et al.* 2011; Patil *et al.* 2016; Li *et al.* 2018). High-depth whole genome sequencing can achieve these marker densities. One study utilizing high-depth whole genome sequencing in soybean found 9,107,000 high quality SNPs (Valliyodan *et al.* 2016). Despite advances and the plummeting cost of next generation sequencing (NGS) data, this approach still presents a heavy financial burden, as several reads are required at each variant site to ensure data quality and completeness.

Decreasing genome coverage in the interest of cost savings introduces missing data, which decreases power and can produce biased results. Imputation of missing data has the potential to allow the researcher to recover nearly all of the missing data points resulting from skim sequencing, drastically reducing genotyping expenses associated generating complete, high quality, high resolution SNP datasets. By predicting the unobserved genotypes based on the surrounding

variants and their correlation to a complete reference panel, missing data can be amended to the correct allele genotype. This technique has been developed and extensively used in human genomic research, and is now commonly extended to other organisms (Pei *et al.* 2008; Howie *et al.* 2009; Howie *et al.* 2011). Seen frequently in plants is the use of imputation to fill missing data points in GBS data (Chan *et al.* 2016; Chung *et al.* 2017). Specially designed populations such as bi-parental, nested, and multi-parent where the founders are genotyped to a high depth and used for the reference haplotypes has been shown to boost accuracy (Tian *et al.* 2011; Swarts *et al.* 2014; Huang *et al.* 2014; Bayer *et al.* 2015; Cericola *et al.* 2018).

Crop breeding programs working with inbred species and/or inbred lines are uniquely positioned to leverage imputation algorithms in an extremely accurate manner. Near complete homozygosity through inbreeding or double haploids allows calling of genotypes despite having sampled one allele at the site. Large haplotype blocks in historically inbred crops theoretically permit imputation accuracy to extend across large physical regions, where genotyped markers are sparse but in high correlation with each other. Success with such a combinatorial approach has been reported in rice, using ~1X coverage sequence data of 517 individuals. Imputation of the missing genotypes in these individuals without a reference panel to produce a SNP panel of ~3.6 million markers with >98% accuracy (Huang *et al.* 2010). This was confirmed in a later study that also included simulations performed down to 0.1X depth. Falling below a depth of 0.5X resulted in steep accuracy consequences, with concordance falling to 76% at the 0.1X level. (Wang *et al.* 2016).

Incorporation of a reference panel has been shown to result in large accuracy improvements at sequencing coverage less than <1X in humans, where imputation at the 0.1X level was improved from less than 5% accuracy to ~70% (Pasaniuc *et al.* 2012). With the growing amount of sequence data present in public databases for many common crops, it is possible to generate an extensive reference panel that might improve accuracy at ultra-low sequence coverage and further cut per sample genotyping cost. In this study, we report on a low coverage whole genome sequencing with imputation approach in a naturally inbred crop, soybean, for producing a low cost, high quality, high density SNP dataset. A reference panel was generated using publicly available high-depth sequencing data for 106 lines, and employed for imputing the missing genotypes of 114 lines sequenced at ultra-low depth. Coverages from 0.1X – 1X depth at intervals of 0.1X were evaluated. The factors influencing error rates and extensibility within/outside soybean were investigated, and the consequences of error rates and types of error on a typical genome-wide association study (GWAS) were explored.

MATERIALS & METHODS

Reference Panel

The reference panel for genotype imputation was generated using publicly available sequence data deposited in the NCBI Short Read Archive from study number SRP062245 (Valliyodan *et al.* 2016). This unfiltered, raw dataset consisted of 106 *Glycine max* lines sequenced at an average of 17.1X coverage (Supplementary Table 1). The raw reads were filtered for adapter sequence contamination, base quality, and truncated reads using Trimmomatic (Bolger *et al.* 2014). Bowtie2 was used to map reads to the *Glycine max* Wm82.a2.v1 reference genome with the "very sensitive" option (Langmead and Salzberg 2012). Reads with a mapping quality score of less than 20 were discarded. SNPs were called using the GATK3.7 HaplotypeCaller tool for an initial panel of 13,052,759 SNPs across all lines (Poplin *et al.* 2017). SNP calls with five or less reads supporting the call were filtered out, as well as calls with a confidence score of less than 20. To control for potential sample contamination/mixing, the inbreeding coefficient, also called the F statistic (Jain and Workman 1967), was calculated using the software Plink1.9 (Purcell *et al.* 2007). As soybean is historically an inbred crop, one can expect F statistics close to one in *Glycine max*. Seven samples fell below a cutoff of 0.9 and were discarded from the final reference panel. All heterozygous calls in the remaining 99 lines were filtered, leaving only biallelic SNPs for consideration. The final reference panel spanned 10,803,148 biallelic homozygous SNPs in 99 lines compared to 10,417,285 SNPs found by Valliyodan *et al.* using the same data set.

Imputation Panel

To generate a low sequence coverage panel for imputation, 114 experimental lines selected from the University of Nebraska soybean breeding program (Supplementary Table 1) were sequenced to a depth of 1X or greater on an Illumina NextSeq 500 (Illumina Hayward, Hayward, CA) using the manufacturer's protocol and 150 base pair paired end reads. DNA was isolated from lyophilized leaf tissue collected from twenty plants per genotype using a CTAB based extraction method (Keim 1988) scaled down for a 96 well plate by dividing all reagent volumes by 40. Extracted genomic DNA was fragmented using a Covaris S220 with the manufacturer's recommended settings for generating ~350 base pair length fragments (Covaris, Inc., Woburn, MA 01801). Double sided size selection was performed using KAPA Pure Beads to retain only fragments within the 250-450 base pair range using the manufacturer's protocol and eluted in 40 μl of TE buffer (Roche Sequencing Solutions, Santa Clara, CA 95050). After testing DNA concentration, samples were standardized to 62.5 ng /μl. Libraries were prepared using a custom protocol adapted from literature to perform A-tailing and end-repair in one reaction, and avoid PCR after adapter ligation by extending the incubation time (Kozarewa and Turner 2011; Knapp *et al.* 2012) . To perform end repair and A-tailing, 16 μl of fragmented genomic DNA for each sample was combined with 1 μl of T4 polynucleotide kinase (PNK) (10U/μl), 1 μl of T4 DNA polymerase (5U/μl), 1 μl of DreamTaq DNA Polymerase (5U/ μl), 2.7 μl of Cut Smart Buffer (10x), 2.2 μl of dATP (10mM), 0.8 μl of dNTP (10 mM), and 0.3 μl of ATP (10mM). Samples were incubated in a thermocycler for 30 min at 20°, and then immediately ramped to 65° and held at this temperature for 30 min. Samples then proceeded immediately to adapter ligation. To the 25 μl of end repaired and A-tailed product the following was added: 10 μl of T4 DNA Ligase Buffer, 3 μl of T4 DNA Ligase (2000U/μl), 3 μl of PEG 6000, 2 μl of PCR grade water, and 2 μl of uniquely barcoded adapters (30mM). Samples were incubated on a thermocycler for 45 min at 20°. After this time, samples were immediately cleaned using KAPA Pure Beads to retain fragments within the 350-550 base pair range and eluted in 20 μl of TE buffer. Multiplexing was performed by combining 5 μl of each individual library. Libraries were quantified using the KAPA Library Quantification Kit for Illumina platforms.

To create subsets simulating depths from 0.1X to 1X at intervals of 0.1X, reads were randomly selected from the raw datasets based upon the total number of reads obtained for each genotype. Each dataset was trimmed for adapter contamination, base quality and truncated reads using Trimmomatic, and then mapped to the *Glycine max* Wm82.a2.v1 reference genome with Bowtie2 using the "very sensitive" option. Mapped reads below a quality score of 20 were filtered. The genotypes at all 10,803,148 SNP positions in the reference panel were called in the low coverage imputation panel using GATK3.7 Haplotype Caller. Genotyping SNPs from a single read has been found accurate in rice whole genome sequencing and maize GBS applications

(Swarts *et al.* 2014; Wang *et al.* 2016) . Any heterozygous calls were discarded, as well as calls not matching the two allele options at that position. For each subset, a random 5% of calls were masked and considered "true" genotypes for evaluating imputation accuracy.

To characterize how genetically distinct the experimental lines were from one another, a genomic relatedness matrix was constructed according to the van Raden metric using the R package "synbreed". Prior to calculation, the imputed dataset was filtered to retain variants with a Beagle posterior genotype probability (GP) score above 0.9, pairwise r^2 LD metric below 0.4, and variant site missing rate below 5% using Plink1.9 (Purcell *et al.* 2007).

Imputation Concordance Evaluation

For the sake of computational efficiency, imputation was performed on a per chromosome basis using Beagle 4.1 (Browning and Browning 2016) with the low memory option. To assess accuracy, the imputed genotype calls were compared to the masked calls, and the percent of those in agreement constituted overall concordance using GATK 3.7's Genotype Concordance tool (McKenna *et al.* 2010). This accuracy assessment was performed across sequencing depths and minor allele frequencies. Three post imputation datasets were considered to quantify any accuracy improvement obtained by filtering poorly imputed sites. This included the raw imputed dataset, and two datasets filtered on GP. Values with GP scores under 0.45 and 0.9 were filtered for the latter two evaluation panels, respectively. VCFtools0.01.12a was used to bin by minor allele frequency, and Plink1.9 (Purcell *et al.* 2007; Danecek *et al.* 2011) was used to filter on GP score. GP score filtering thresholds were determined after examining their relationship to error rate (Supplementary Table 2)

Error and Linkage Disequilibrium

Error in relationship to linkage disequilibrium (LD) was examined as a potential metric of extensibility to other soybean population and crop species. D' and r^2 statistics were calculated for all pairwise reference panel SNPs using Plink1.9 (Gaunt *et al.* 2007; Purcell *et al.* 2007). Proportion of errors made at each SNP site across was calculated by comparing the imputed values to the masked values across all subsets of depths. To reduce noise, data were smoothed through the application of a rolling average window with a width of 1500 SNPs after ordering by the respective pairwise LD metric. A second order polynomial was fit to describe the D' and error relationship, and a simple linear regression was fit to describe the relationship between r^2 and error.

Relationship Between Samples & Reference Panel

Close relatedness between the sample and reference genotypes has been previously reported to increase imputation precision. Relatedness matrices were generated based on five different coefficients and averaged the top five scores from each sample genotype as a metric for gauging degree of relatedness to the reference panel. These measures were plotted against concordance scores from the imputed data filtered for GP scores above 0.9 and averaged across all depth levels. A simple linear regression model was fit to assess potential correlation. Relatedness matrices were calculated using the R package "synbreed", using options corresponding to measures described by vanRaden, Astle and Balding, Reif, Hayes and Goddard, and Euclidean distances (Wimmer *et al.* 2012).

Genome Representation

Genomic studies improve as the linkage between the genotyped polymorphism and underlying causative gene increases. The extent of LD between two markers therefore constitutes proxy for the correlation of the marker and underlying gene(s) of interest. To assess how well the panel represented variation across the genome, the distribution of LD in the imputed experimental dataset was compared to the SoySNP50k Array positions extracted from the imputed experimental dataset (Song *et al.* 2015). SNPs with MAF below 0.05 were filtered out, a quality control step implemented in most genomic studies. Both D' and r^2 were calculated using Plink1.9, and distributions plotted in R3.4 (Team 2017).

Error and Beagle Posterior Genotype Probability

To explore the possibility of using GP values as a post imputation filtering metric, proportion of error across depth subsets was plotted against GP. A rolling average window with a width of 500 SNPs was applied to the proportion error after ordering by GP, and a second degree polynomial was fit to describe the relationship in R3.4.

Error Type: Allele frequencies exhibit some degree of influence on the results of many genomics studies. Therefore, how imputation error skews this metric is of significant interest. Masked and imputed datasets were coded according to the major allele in the reference dataset. Errors were binned into four categories, homozygous major to minor, homozygous minor to homozygous major, homozygous major to heterozygous, and homozygous minor to heterozygous, based on which allele was incorrectly imputed and which allele was true. Because all heterozygous calls were filtered in the initial data generation, no heterozygous to major, or heterozygous to minor category exists.

Power Analysis

In the interest of determining the potential cost of imputation error, a basic power calculation for minor to major and major to minor errors in a GWAS was performed. Using an R implementation of Purcell's "Genetic Power Calculator" (Purcell *et al.* 2003), power was calculated to detect a moderate effect QTL across minor allele frequency bins from <0.025 to 0.5. Simulations assumed an additive genetic model, 300 genotypes, LD between the QTL and marker of 0.8 D', a significance threshold that mirrored the Bonferroni correction for 1,716,234 SNPs (the final size of the SNP dataset after quality control filtering), and a QTL effect size of 1 standard deviation. Error rates from 1–10% were tested at intervals of 1%, with 100 iterations of the simulation performed at each error level. To investigate the possibility of including more genotypes to overcome power losses associated with imputation error, simulations were also performed for 150, 500, and 1000 genotypes for a 5% error rate at the same conditions as specified above.

Cost Analysis

Decreasing cost per sample allows a researcher to expand a study to overcome power loss introduced through the imputation error. To illustrate the impact of this, per sample sequencing costs were calculated using current Illumina NextSeq500 high throughput 300 cycle sequencing kit prices, cost analysis of a custom library prep protocol, and CTAB DNA extraction method (Supplementary Table 3). The retained cost per sample and average raw concordance were plotted as depth decreased.

Data Availability

Raw sequencing data directly generated by this project for use in creating the study panel has been deposited in the NCBI Short Read Archive under accession number PRJNA512147. The reference panel used for genotype imputation was generated using previously publicly available sequence data deposited in the NCBI Short Read Archive from study number SRP062245 (Valliyodan *et al.* 2016). Supplementary figures

Table 1 The number of markers and genotyping rate in each low coverage subset from 0.1X to 1X sequencing depth. As coverage decreases, the total number of markers captured and completeness of the SNP panel decreases.

Mean Depth	Genotyping Rate	Number of SNPs	Reads	Base Pairs
1	32.44%	1,288,463	6,327,889	949,183,385
0.9	30.41%	1,240,823	5,695,100	854,265,047
0.8	27.77%	1,174,619	5,062,311	759,346,708
0.7	24.91%	1,097,843	4,429,522	664,428,370
0.6	21.80%	1,005,880	3,796,734	569,510,031
0.5	18.47%	895,596	3,163,945	474,591,693
0.4	14.98%	760,167	2,531,156	379,673,354
0.3	11.40%	590,786	1,898,367	284,755,016
0.2	7.85%	375,343	1,265,578	189,836,677
0.1	4.74%	133,747	632,789	94,918,339

and tables can be found in "Supplementary Figures and Tables." The 114 × 114 relatedness matrix is available in Supplementary File 1. Supplemental material available at FigShare: https://doi.org/10.25387/g3.7975751.

RESULTS

SNP Genotyping & Imputation

The reference panel for imputation was constructed using 106 *Glycine max* lines sequenced at an average of 17.1X coverage using publicly available sequencing data deposited in the NCBI Short Read Archive (Valliyodan *et al.* 2016) (Supplementary Table 1). After quality control measures were applied to the raw and mapped sequence data (see Materials & Methods), a final reference panel of 10,803,148 biallelic homozygous SNPs across 99 lines was generated. SNPs discovered in the reference panel were used to genotype experimental lines in the study panel. This consisted of 114 lines that were sequenced to a depth of at least 1X. Coverages from 0.1X to 1X were analyzed by randomly subsetting reads from the raw sequence data. Of the 10,803,148 million markers discovered in the reference panel, the number of SNPs genotyped by this low coverage study panel subsets ranged from 133,747 to 1,288,463 markers. These subsets also ranged in missing data rates from 95.26 to 67.56% for those markers (Table 1). Using the reference panel, genotype values for all missing positions were imputed.

Close relatedness between the lines in the experimental panel may bias the overall accuracy of the imputation results. To evaluate this, vanRaden relatedness scores were calculated using real and imputed genotypes. The resulting values ranged from -0.44858 to 0.91112, with a median value of -0.02851, mean of -0.00286, and standard deviation of 0.17650. Strong relationships are generally indicated by values over 0.4. Our experimental panel exhibits few strongly related lines, with only 2.7% of all possible pairwise combinations showing a relationship above this threshold. Therefore, we would conclude the majority of our experimental genotypes to be distally/non-related (Supplementary File 1).

An alternative to this whole genome sequencing approach are fixed SNP arrays. However, this method provides less total SNPs for genomic studies and may not capture as much of the genome. High LD between SNPs can be extended to assume a strong correlation to other genomic variation between them. Plotting the density distributions of r^2 and D' LD measures for the Soy50KSNP Array and imputed dataset demonstrated that whole genome sequencing with imputation had a greater concentration of values toward higher linkage values. Generally, a D' or r^2 of over 0.8 between is considered "strong linkage". The imputed dataset provided 1,716,234 SNPs after common quality control filters, with 36.00% and 85.66% of r^2 and D' values above 0.8, respectively. This is in comparison to the 42,133 SNPs in the fixed array, where 24.20% and 80.00% of r^2 and D' values are above 0.8 (Figure 1). If high LD indicates a better tagging of underlying variation, the imputed dataset captures the genome's SNP variation better than the Soy50KSNP Array.

Imputation Accuracy

Prior to imputation, 5% of genotype calls from the skim sequencing data were withheld to assess accuracy. Overall imputation accuracy was consistent for raw and filtered datasets as sequencing depth decreased from 1X until 0.3X, where accuracy drops off by an average of 3.5% from 0.3X to 0.1X (Figure 2A, Supplementary Table 2). Assessing the error type of this study showed that 53.13% of the errors made were incorrect imputation of the minor allele when the major allele was true. Of the remaining errors, 35.10% were incorrect imputation of the major allele when the minor allele was true, and 11.77% were incorrect imputation of heterozygous calls. No heterozygous to major/minor errors exist as at heterozygous calls were filtered in the initial panels (Figure 3).

Filtering on Beagle's posterior genotype probability (GP) to improve dataset quality was successful. When imputed positions with a GP score of less than 0.45 were discarded, accuracy improved by an average of 2.50% across sequencing depths. A more stringent filter that only kept positions with a GP score over 0.9 resulted in a 4.26% increase in

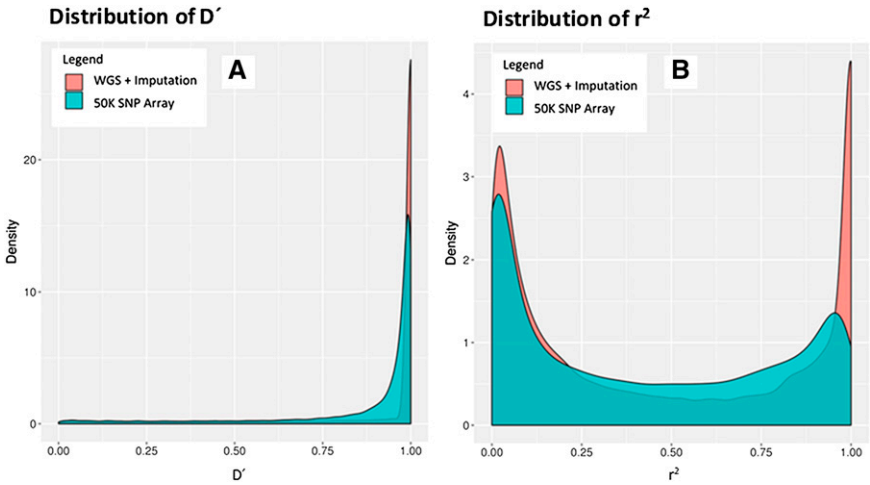

Figure 1 Comparing density plots for LD measures D' (A) and r^2 (B) demonstrates that using whole genome sequencing with imputation results in a dataset that has a higher proportion of SNPs is strong pairwise linkage with each other, represented in the heavier tails in red near D' and r^2 values of 1.

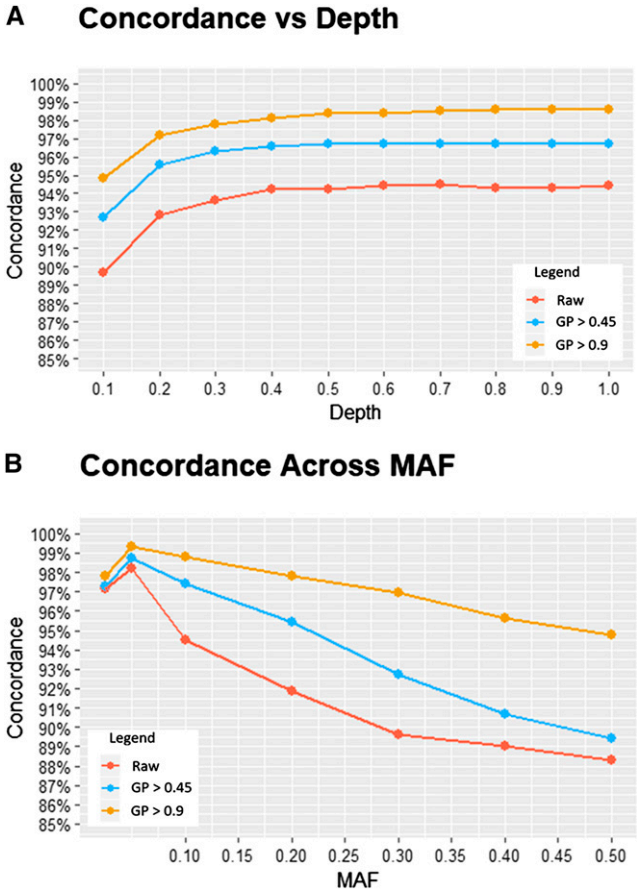

Figure 2 A) Overall accuracy of filtered and raw imputed datasets were plotted across the evaluated depths. For all study panels, concordance rapidly erodes below a sequencing depth of ~0.3X. B) Examining accuracy in the context of minor allele frequency (MAF) reveals that error occurs at higher rates as MAF approaches a maximum of 0.5.

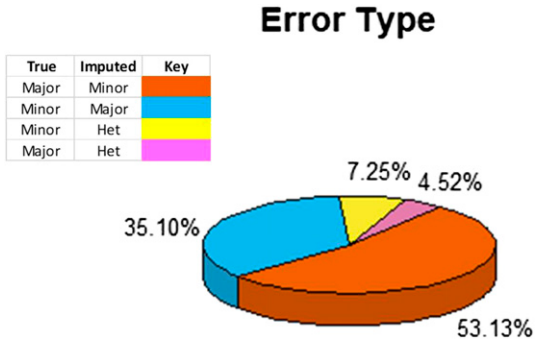

Figure 3 Proportion of errors made as categorized by whether the minor/major/heterozygous alleles was misimputed. In over half of all the errors made, Beagle overimputes the minor allele when the major allele is the true genotype. Incorrect heterozygous imputations make up a minor proportion of the total error and would likely be filtered out in inbred panels.

accuracy (Supplementary Table 2). This practice did reintroduce some missing data, which varied across depth and filtering level. Data loss as a result of post imputation filtering was below 5% for all depths at a filtering level of GP $>$ 0.45, but quickly inflates when filtering for imputation quality of GP $>$ 0.9 to a missing data rate of 20.82% at 0.1X (Supplementary Figure 1). While filtering on Beagle's posterior genotype probability may reduce falsely imputed genotypes, it must be balance with the reintroduction of missing data it causes.

The error rate at individual marker loci may not be well captured by the overall concordance across all SNPs. Examining concordance in the context of minor allele frequency (MAF) reveals as MAF values approach a maximum of 0.5, concordance decreases. Application of post imputation filters of GP values increases overall accuracy through improved concordance at these increased MAFs (Figure 2B). This trend is uniform across all sequencing depths (Supplementary Figure 2). Through examining imputation accuracy in this manner, it is apparent that higher error rates are occurring at SNP positions at MAFs nearest 0.5 than is described by the average concordance measure.

Error rates in imputation may be influenced by characteristics specific to the population and crop species to which it is applied. The correlation between variants is a cornerstone to the success of imputation. If the correlation between alleles is high then imputation accuracy should also be high and as the correlation between alleles decrease then the accuracy of imputation should also decrease. This correlation between alleles can be measured with LD. Soybean is a historically inbred crop with long ranging LD (Zhou *et al.* 2015). As D' and r^2 approach 1, where neighboring SNPs are in perfect linkage with each other, error rates are at their lowest. Both relationships demonstrate a very strong correlation with R^2 values of 0.98 and 0.89 for r^2 and D' respectively (Figure 4), indicating LD is an important factor to consider when applying this technique to other soybean populations or other crop species.

Relationship of the study genotypes to the reference panel genotypes has been suggested as a strong influencer of imputation accuracy. Plotting calculated values for five unique kinship metrics against concordance for each genotype did not demonstrate any strong linear relationships. The maximum correlation for any of the measures was for Reif's method, at an R^2 of 0.26. Examination of the standard error, shows that the study population varies narrowly in terms of relatedness to the reference panel. Additionally, assessing the raw values suggests that the study population is weakly related to reference genotypes. This is best illustrated with the vanRaden and Astle & Balding measurements, where a "strong" relationship is usually indicated by values approximately >= 0.4. In both these cases, the largest measure does not exceed 0.18 and 0.16 (VanRaden 2008; Astle and Balding 2009). The combination of diminished values and narrow standard error indicates a weak relationship of the study panel to the reference panel (Supplementary Figure 3). The evidence of a weak relationship suggests that relationship was not a strong influencer of the high imputation accuracies obtained.

GWAS Power

Understanding the effect error rate has on genomic studies is important when selecting an appropriate genotyping technique. To determine the effect of the error rate of skim sequencing and imputation has on GWAS we performed power simulations of detecting a moderate effect QTL in a panel of 300 individuals. This power study showed significantly decreased power to detect QTL with increasing errors at MAF from 0.1-0.3. This was most pronounced when the minor allele was incorrectly imputed as the major allele. Above 0.3 MAF, power for QTL detection was minimally affected by error (Figure 5A). Studying the effect of three additional sample sizes, while assuming a 5% error rate demonstrates the potential for experimenters to recover power losses through inclusion of more genotypes. Including 500 individuals at this fixed error rate recovers and even slightly improves power at mid-range MAFs over studying 300 genotypes with no genotyping error (Figure 5B).

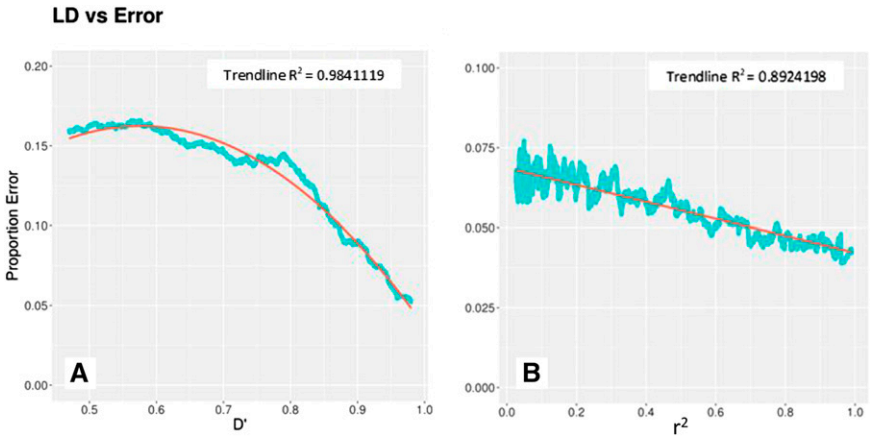

Figure 4 Comparing the smoothed frequency of errors made at individual SNP sites with LD measures D' (A) and r^2 (B) demonstrates the strong influence of linkage disequilibrium on imputation accuracy.

DISCUSSION

This study illustrates the potential of low coverage sequencing with imputation as an economical approach to obtaining high density SNP genotype information in soybean. Accelerating improvement of complex phenotypes through genomics necessitates high quality, high resolution marker data. However, studies are often limited by the cost required to obtain this information through high coverage sequencing. The combination of low coverage sequencing with imputation presents an option that drastically cuts costs while retaining a high level of accuracy. Implementing a similar method in rice allowed researchers to generate a high quality, dense SNP dataset using 1X depth whole genome sequence (Huang *et al.* 2010; Wang *et al.* 2016). This analysis in soybean, which differs in the inclusion of a reference panel for imputation, determined sequencing depth could be reduced to 0.3X with no significant accuracy losses. Analogous results have been demonstrated in humans, where it was concluded that a reasonably accurate and dense dataset could be obtained from 0.2X coverage supplemented with imputation using a reference panel (Pasaniuc *et al.* 2012). To our knowledge, this is the first work to examine using imputation with real sequence data at less than 1X coverage in the construction of a high quality, highly affordable SNP dataset in plants. The effect of imputation method and structure of the reference panel have not been specifically examined in the context of application to skim sequencing, providing future avenues for research and improvement.

While SNP arrays and GBS are popular options for obtaining genotype information, high precision genomics demands markers to be in close linkage to the contributing genes. Regions of the genome with sparsely correlated markers may therefore contain overlooked causal variation (Hirschhorn and Daly 2005; Witte 2010). Skim sequencing with imputation, as investigated here, tags a significantly larger portion of the genome in tighter LD than the current soybean 50k array. This effect may be presumed to extend to GBS datasets of a similar size. Such a boost in resolution may therefore reveal QTL in regions of the genome that would not have been captured through smaller datasets.

The accuracy and extensibility of this approach in other soybean populations, as well as other crops is based on several factors. To explore potential limitations in this method, population LD and the relatedness of reference panel to study lines were examined. Both of these factors have been implicated as strong influencers of imputation accuracy due to the innate reliance of the technique on the presence of sample haplotypes within the reference panel, as well as the extent of correlation between observed markers (Hickey *et al.* 2012; He *et al.* 2015). The strong

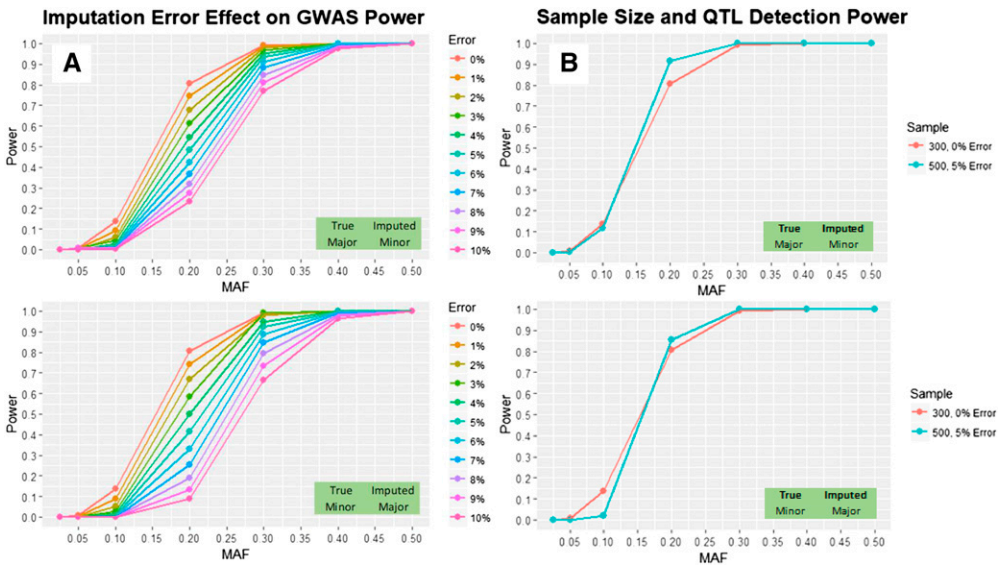

Figure 5 A) The power to detect a moderate effect QTL becomes increasingly sensitive to error for both major to minor and *vice versa* errors at intermediary MAFs. B) Comparing the power to detect the same QTL with 300 samples at a 0% genotyping error *vs.* 500 samples with a 5% error rate demonstrates that cost savings can be used to increase study sizes in order to recover power losses introduced by the imputation error of both major and minor alleles.

inverse relationship observed between the proportion of SNPs incorrectly imputed at a given position and LD measurements suggests that for soybean populations and other crops with shorter range LD, imputation accuracy will likely decrease. There was no significant relationship detected between kinship measures and accuracy. However, the study genotypes exhibited little variation for any of the calculated metrics, which can be seen in the low standard deviations. Without a wide range of values to examine, identifying a clear trend is unlikely. The positive effect of relatedness on imputation accuracy is documented in other literature (Hickey *et al.* 2012; Ma *et al.* 2013; Boison *et al.* 2015), and should therefore be a consideration in expanding this method to other soybean populations and crop species. The overall weak kinship between study and reference panels in this data may also be viewed as a positive, since high levels of imputation accuracy were achieved despite this populations being interpreted as distally related.

The power to detect a QTL is partially dependent on the allele frequency at that loci (Ardlie *et al.* 2002; Tabangin *et al.* 2009). Therefore, the relationship between imputation accuracy, minor allele frequency (MAF), and statistical power may be considered particularly important. In agreement with an analysis performed with maize, the data showed steadily decreasing imputation accuracy as MAF increased with the exception of very rare alleles (MAF $<$ 0.05) (Hickey *et al.* 2012). An opposite tendency was observed with respect to statistical power losses across MAF, so it can be interpreted that at the loci a SNP dataset would display the highest imputation error rates, the GWAS is least affected by them. This trend has also been supported in human imputation analyses looking at sample size inflation factors under different imputation error types (Huang *et al.* 2009). In both cases, power consequences were greater for incorrect imputation of the minor allele. It is unclear how a combination of error types at a SNP locus would influence genomic studies. Decisions on the level of decreased coverage that can be tolerated should consequently be made not on the overall average concordance, but by examining the concordance across minor allele frequencies in relation to the maximum allowable error to retain power.

The cost savings associated with this method can be used to include more sample genotypes, not only recovering power losses at low minor allele frequencies, but potentially increasing total power. Similar results in humans have indicated sampling more genotypes with small error is more beneficial over fewer genotypes with perfect accuracy (Pasaniuc *et al.* 2012). Comparing the raw accuracy along sequencing depths along with per sample costs, shows that at the previously identified critical threshold of 0.3X coverage, there is only a 0.85% loss of accuracy relative to using a 1X sequence, while costs decreased 57% (Supplementary Figure 4). Moreover, the use of public sequence data to construct a broad reference panel eliminates the cost and limitations of assembling special populations and sequencing the founders to a high coverage to serve as the reference haplotypes.

CONCLUSION

Here it is demonstrated that low coverage sequencing accompanied with imputation from a reference panel can be extended below 1X depth in soybean to capture high density, reasonably accurate SNP genotype information economically. The tremendous drop in per sample sequencing cost over high depth methods may allow researchers to expand the number of study genotypes in their investigations, while representing a larger portion of the genome than fixed SNP arrays and GBS data. The potential for success of this genotyping method within and outside of soybean is highly reliant on population LD. Furthermore, researchers should examine accuracy and power within the context of minor allele frequency to make informed decisions about sequencing depth tolerances. As genomics demands increasing SNP panel densities across a wide range of genotypes, skim sequencing with imputation constitutes a financially feasible and highly accurate way to meet these requirements.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the Nebraska Soybean Board project #1726. The authors also acknowledge Dr. Reka Howard and Dr. Keenan Amundsen for providing their technical perspective during the compilation of project results and manuscript drafting. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

LITERATURE CITED

Ardlie, K. G., K. L. Lunetta, and M. Seielstad, 2002 Testing for Population Subdivision and Association in Four Case-Control Studies. Am. J. Hum. Genet. 71: 304–311. https://doi.org/10.1086/341719

Astle, W., and D. J. Balding, 2009 Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat. Sci. 24: 451–471. https://doi.org/10.1214/09-STS307

Bayer, P. E., P. Ruperao, A. S. Mason, J. Stiller, C.-K. K. Chan *et al.*, 2015 High-Resolution Skim Genotyping by Sequencing Reveals the Distribution of Crossovers and Gene Conversions in Cicer Arietinum and Brassica Napus. Theor. Appl. Genet. 128: 1039–1047. https://doi.org/10.1007/s00122-015-2488-y

Boison, S. A., D. J. A. Santos, A. H. T. Utsunomiya, R. Carvalheiro, H. H. R. Neves *et al.*, 2015 Strategies for Single Nucleotide Polymorphism (SNP) Genotyping to Enhance Genotype Imputation in Gyr (Bos Indicus) Dairy Cattle: Comparison of Commercially Available SNP Chips. J. Dairy Sci. 98: 4969–4989. https://doi.org/10.3168/jds.2014-9213

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Browning, B. L., and S. R. Browning, 2016 Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. 98: 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Cericola, F., I. Lenk, D. Fè, S. Byrne, C. S. Jensen *et al.*, 2018 Optimized Use of Low-Depth Genotyping-by-Sequencing for Genomic Prediction Among Multi-Parental Family Pools and Single Plants in Perennial Ryegrass (Lolium Perenne L.). Front. Plant Sci. 9: 369. https://doi.org/10.3389/fpls.2018.00369

Chan, A. W., M. T. Hamblin, and J. L. Jannink, 2016 Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data. PLoS One 11: e0160733. https://doi.org/10.1371/journal.pone.0160733

Chung, Y. S., S. C. Choi, T.-H. Jun, and C. Kim, 2017 Genotyping-by-Sequencing: A Promising Tool for Plant Genetics Research and Breeding. Hortic. Environ. Biotechnol. 58: 425–431. https://doi.org/10.1007/s13580-017-0297-8

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The Variant Call Format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Huang, B. E., C. Raghavan, R. Mauleon, K. W. Broman, and H. Leung, 2014 Efficient Imputation of Missing Markers in Low-Coverage Genotyping-by-Sequencing Data from Multi-Parental Crosses. Genetics 197: 401–404. https://doi.org/10.1534/genetics.113.158014

Gaunt, T. R., S. Rodríguez, and I. N. Day, 2007 Cubic Exact Solutions for the Estimation of Pairwise Haplotype Frequencies: Implications for Linkage Disequilibrium Analyses and a Web Tool 'CubeX.'. BMC Bioinformatics 8: 428. https://doi.org/10.1186/1471-2105-8-428

Hamblin, M. T., E. S. Buckler, and J.-L. Jannink, 2011 Population Genetics of Genomics-Based Crop Improvement Methods. Trends Genet. 27: 98–106. https://doi.org/10.1016/j.tig.2010.12.003

He, S., Y. Zhao, M. F. Mette, R. Bothe, E. Ebmeyer *et al.*, 2015 Prospects and Limits of Marker Imputation in Quantitative Genetic Studies in

European Elite Wheat (Triticum Aestivum L.). BMC Genomics 16: 168. https://doi.org/10.1186/s12864-015-1366-y

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos, 2012 Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52: 654. https://doi.org/10.2135/cropsci2011.07.0358

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-Wide Association Studies for Common Diseases and Complex Traits. Nat. Rev. Genet. 6: 95–108. https://doi.org/10.1038/nrg1521

Howie, B., J. Marchini, and M. Stephens, 2011 Genotype Imputation with Thousands of Genomes. G3 (Bethesda) 1: 457–470. https://doi.org/10.1534/g3.111.001198

Howie, B. N., P. Donnelly, and J. Marchini, 2009 "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies." PLoS Genet. 5: e1000529. https://doi.org/10.1371/journal.pgen.1000529

Huang, B. E., C. Raghavan, R. Mauleon, K. W. Broman, and H. Leung, 2014 Efficient Imputation of Missing Markers in Low-Coverage Genotyping-by-Sequencing Data from Multi-Parental Crosses. Genetics 197: 401–404, .https://doi.org/10.1534/genetics.113.158014

Huang, L., C. Wang, and N. A. Rosenberg, 2009 The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations. Am. J. Hum. Genet. 85: 692–698. https://doi.org/10.1016/j.ajhg.2009.09.017

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng *et al.*, 2010 Genome-Wide Association Studies of 14 Agronomic Traits in Rice Landraces. Nat. Genet. 42: 961–967. https://doi.org/10.1038/ng.695

Jain, S. K., and P. L. Workman, 1967 Generalized F-Statistics and the Theory of Inbreeding and Selection. Nature 214: 674–678. https://doi.org/10.1038/214674a0

Keim, P., 1988 "A Rapid Protocol for Isolating Soybean DNA." Soybean Genet. Newsl. 15: 150–152. https://ci.nii.ac.jp/naid/10015372412/.

Knapp, M., M. Stiller, and M. Meyer, 2012 Generating Barcoded Libraries for Multiplex High-Throughput Sequencing. Methods Mol. Biol. 840: 155–170. https://doi.org/10.1007/978-1-61779-516-9_19

Kozarewa, I., and D. J. Turner, 2011 Amplification-Free Library Preparation for Paired-End Illumina Sequencing. Methods Mol. Biol. 733: 257–266. https://doi.org/10.1007/978-1-61779-089-8_18

Langmead, B., and S. L. Salzberg, 2012 "Fast Gapped-Read Alignment with Bowtie 2." Nature Methods 9: 357–359. https://doi.org/10.1038/nmeth.1923

Li, F., J. Xie, X. Zhu, X. Wang, Y. Zhao *et al.*, 2018 Genetic Basis Underlying Correlations Among Growth Duration and Yield Traits Revealed by GWAS in Rice (Oryza Sativa L.). Front. Plant Sci. 9: 650. https://doi.org/10.3389/fpls.2018.00650

Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 Genomic Selection in Plant Breeding: Knowledge and Prospects. Adv. Agron. 110: 77–123. https://doi.org/10.1016/B978-0-12-385531-2.00002-5

Ma, P., R. F. Brøndum, Q. Zhang, M. S. Lund, and G. Su, 2013 Comparison of Different Methods for Imputing Genome-Wide Marker Genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96: 4666–4677. https://doi.org/10.3168/jds.2012-6316

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. Genome Res. 20: 1297–1303. https://doi.org/10.1101/gr.107524.110

Pasaniuc, B., N. Rohland, P. J. McLaren, K. Garimella, N. Zaitlen *et al.*, 2012 Extremely Low-Coverage Sequencing and Imputation Increases Power for Genome-Wide Association Studies. Nat. Genet. 44: 631–635. https://doi.org/10.1038/ng.2283

Patil, G., T. Do, T. D. Vuong, B. Valliyodan, J.-D. Lee *et al.*, 2016 Genomic-Assisted Haplotype Analysis and the Development of High-Throughput SNP Markers for Salinity Tolerance in Soybean. Sci. Rep. 6: 19199. https://doi.org/10.1038/srep19199

Pei, Y.-F., J. Li, L. Zhang, C. J. Papasian, and H.-W. Deng, 2008. "Analyses and Comparison of Accuracy of Different Genotype Imputation Methods." PLoS ONE 3: e3551. https://doi.org/10.1371/journal.pone.0003551

Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro *et al.*, 2017 Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples. bioRxiv. https://doi.org/10.1101/201178

Purcell, S., S. S. Cherny, and P. C. Sham, 2003 "Genetic Power Calculator: Design of Linkage and Association Genetic Mapping Studies of Complex Traits." *BIOINFORMATICS APPLICATIONS NOTE.* Vol. 19. http://svn.donarmstrong.com/don/trunk/projects/research/linkage/papers/genetic_power_calculator_purcell_sham_bioinfor_19_1_149_2003_pmid_12499305.pdf. https://doi.org/10.1093/bioinformatics/19.1.149

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81: 559–575. https://doi.org/10.1086/519795

Rasheed, A., Y. Hao, X. Xia, A. Khan, Y. Xu *et al.*, 2017 Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. Mol. Plant 10: 1047–1064. https://doi.org/10.1016/j.molp.2017.06.008

Song, Q., D. L. Hyten, G. Jia, C. V. Quigley, E. W. Fickus *et al.*, 2015 Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3 (Bethesda) 5: 1999–2006. https://doi.org/10.1534/g3.115.019000

Swarts, Kelly, Huihui Li, J. Alberto Romero Navarro, Dong An, Maria Cinta Romay, Sarah Hearne, Charlotte Acharya *et al.* 2014 "Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants." *The Plant Genome* 7 (3). https://doi.org/10.3835/plantgenome2014.05.0023

Tabangin, M. E., J. G. Woo, and L. J. Martin, 2009 "The Effect of Minor Allele Frequency on the Likelihood of Obtaining False Positives." *BMC Proceedings* 3 Suppl 7 (Suppl 7): S41. https://doi.org/10.1186/1753–6561–3-S7–S41

Team, R. C., 2017 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria., https://scholar.google.com/scholar?hl=en&as_sdt=0,28&cluster=8918609904990403039.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun *et al.*, 2011 Genome-Wide Association Study of Leaf Architecture in the Maize Nested Association Mapping Population. Nat. Genet. 43: 159–162. https://doi.org/10.1038/ng.746

Valliyodan, B., D. Qiu, G. Patil, P. Zeng, J. Huang *et al.*, 2016 Landscape of Genomic Diversity and Trait Discovery in Soybean. Sci. Rep. 6: 23598. https://doi.org/10.1038/srep23598

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wang, H., X. Xu, F. G. Vieira, Y. Xiao, Z. Li *et al.*, 2016 The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication. Mol. Plant. 9: 975–985. https://doi.org/10.1016/j.molp.2016.04.018

Wimmer, V., T. Albrecht, H.-J. Auinger, and C.-C. Schön, 2012 Synbreed: A Framework for the Analysis of Genomic Prediction Data Using R. Bioinformatics 28: 2086–2087. https://doi.org/10.1093/bioinformatics/bts335

Witte, John S. 2010 "Genome-Wide Association Studies and Beyond." *Annual Review of Public Health* 31: 9–20 4 p following 20. https://doi.org/10.1146/annurev.publhealth.012809.103723

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu *et al.*, 2015 Resequencing 302 Wild and Cultivated Accessions Identifies Genes Related to Domestication and Improvement in Soybean. Nat. Biotechnol. 33: 408–414 (erratum: Nat. Biotechnol. 34: 441). https://doi.org/10.1038/nbt.3096

Communicating editor: D. J. de Koning