# PATRIOT: A pipeline for tracing identical-by-descent chromosome segments to improve genomic prediction in self-pollinating crop species

Johnathon M. Shook[1], Daniela Lourenco[2], Asheesh K. Singh[1,*]

[1]Department of Agronomy, Iowa State University, IA, USA; [2]Department of Animal and Dairy Science, University of Georgia, GA, USA.

[*] Corresponding author (singhak@iastate.edu)

**Abbreviations:**

GP= Genomic prediction

GS= Genomic selection

rrBLUP= Ridge regression best linear unbiased predictors

PATRIOT= "Parental Allele Tracing, Recombination Identification, and Optimal predicTion"

IBD= Identical-by-descent

IBS= Identical-by-state

LD= Linkage disequilibrium

QTL= Quantitative trait locus

MAS= Marker-assisted selection

USDA= United States Department of Agriculture

NAM= Nested Association Mapping (population)

SNP= Single nucleotide polymorphism

## ABSTRACT

The lowering genotyping cost is ushering in a wider interest and adoption of genomic prediction and selection in plant breeding programs worldwide. However, improper conflation of historical and recent linkage disequilibrium between markers and genes restricts high accuracy of genomic prediction (GP). Multiple ancestors may share a common haplotype surrounding a gene, without sharing the same allele of that gene. This prevents parsing out genetic effects associated with the underlying allele of that gene among the set of ancestral haplotypes. We present 'Parental Allele Tracing, Recombination Identification, and Optimal predicTion' (i.e., PATRIOT) approach that utilizes marker data to allow for a rapid identification of lines carrying specific alleles, increases the accuracy of genomic relatedness and diversity estimates, and improves genomic prediction. Leveraging identity by descent, PATRIOT showed an improvement in GP accuracy by 16.6% compared to the traditional rrBLUP method. This approach will help to increase the rate of genetic gain and allow available information to be more effectively utilized within breeding programs.

## INTRODUCTION

Crop domestication has caused extreme genetic bottleneck, with a reduction in genetic diversity in domesticated crops compared to wild ancestors including in soybean (*Glycine max* L. Merr.) (Hyten et al. 2006). Consequently, the number of ancestral individuals that are represented in modern cultivars is quite low (Gizlice et al. 1994). For example, 17 founding lines contributed 75% of the genes in modern U.S. soybean cultivars, and 95% of genes could be traced to 35 ancestral lines, demonstrating an extremely narrow genetic variation challenging breeding progress. This is not confined to soybean alone, as other crops have similar challenges (Bennett et al. 2012, Smith 2007).

The narrow genetic variability within modern breeding programs is a concern for breeders, as low diversity implies an incomplete sampling of favorable alleles as breeders attempt to improve crop performance and plasticity (Kisha et al. 1998). Furthermore, the likelihood of untapped resistance to biotic and abiotic stresses and unavailability of favorable genes is high (Burdon 2001). Low genetic diversity also negatively influences the response to selection, i.e., genetic gain (Tanksley and McCouch 1997). In soybean, reliance on a single source for soybean cyst nematode (SCN) resistance has increased the ability of SCN to reproduce at a higher rate, necessitating the need to bring in additional sources of resistance (Tylka 2007). Tracking identity-by-descent (IBD) presents unique advantages that can benefit ongoing plant breeding efforts in utilizing the narrow genetic germplasm pool within modern varieties effectively.

Within breeding populations, genes or marker alleles can be expressed as either IBD (individuals share nucleotide sequence; marker allele is the same by inheritance from a shared ancestor) or identical-by-state (IBS) (individuals share nucleotide sequence; marker allele is the

same independent of the origin) (Lynch and Walsh 1998). IBD data provides greater information than IBS, as the nucleotide sequence between two adjacent IBD marker alleles from one parent in an individual is also inherited from that same parent at a high probability, barring mutation or double recombination. This physical linkage can extend across multiple loci on a segment of chromosome, depending on local recombination rates; and can be captured using genetic markers. Further, linkage disequilibrium (LD) between a set of markers or gene loci due to physical linkage, mutation, population stratification, or recombination rates may be present (Slatkin 2008). When recombination is low within a region of multiple marker loci, it becomes possible to identify haplotypes, or runs of multiple markers which are consistently inherited together (Daly et al. 2001).

Current genomic selection models effectively model IBS relationships between lines and utilize historic linkage between markers and the trait of interest. This approach has worked reasonably well (Sorrells 2015), which can be attributed to the similarity between IBS and IBD relatedness metrics. In cases where LD is high locally, IBS relationships are more similar to those calculated based on IBD. The use of IBD can improve relationship estimation (Li et al. 2014), kinship and population structure (Morrison 2013), and also is useful for genetic mapping (Dawn Teare and Barrett 2005). The haplotype information from IBD due to inheritance from a recent common ancestor can therefore enable more accurate relationship estimates and improve the effectiveness of genomic selection with IBD-based genomic selection approaches. However, in order to take full advantage of the benefits of IBD data, it is first necessary to track true IBD segments within the population.

The advancements in genetic marker technology have revolutionized the understanding of existing breeding germplasm, allowing the identification of QTL responsible for variation in a

trait of interest using genome-wide association or QTL mapping studies, as well as selection of genotypes which contain a QTL of interest in marker-assisted selection (MAS) or marker assisted backcrossing (MABC). Nearly the entire USDA collection of soybean varieties has been genotyped using the SoySNP50K single nucleotide polymorphism (SNP) array containing genome wide markers (Song et al. 2015). Large-scale genotyping efforts have been shown to be useful for scanning germplasm collections in multiple crops (Moellers et al. 2017, de Azevedo Peixoto *et al.* 2017, Yu et al. 2016), and for conducting genome wide association studies using the germplasm collection accessions (Coser et al. 2017, Moellers et al. 2017, Zhang et al. 2017). Alternatively, genome-wide markers can be used to predict performance of untested lines (i.e., genomic prediction (GP)) and subsequently select new varieties with the greatest expected genotypic values (genomic selection (GS)). GS is becoming mainstream in mid- to large breeding programs (Hickey et al. 2017), as it unlocks new opportunities to perform selections in early generations and predict parental suitability (Battenfield et al. 2016, Yao et al. 2018). This leads to the ability to select improved lines reliably with less field testing and speed their re-use as parents in a breeding program. However, GS models require continual model training and validation as a reduction in prediction accuracy has been reported after two to three generations (Jannick 2010). Widespread adoption of GS may require higher prediction accuracy to make up for the additional cost of generating marker data and phenotyping a training population. Therefore, numerous efforts have been made to improve the prediction accuracy of GS models (Jia et al. 2012, Solberg et al. 2008, Habier et al. 2011). One such approach is to better utilize LD information to track chromosome regions inherited from each parent in regions of high LD (Thompson 2013). In this approach, fewer markers are needed for filial generations as so-called "tag" markers can be used to elucidate haplotypes in lines and progenies (Johnson et al. 2001).

While previous efforts have relied on using haplotypes based on observed LD between markers, we explore an alternative approach of tracking the parental source of each allele. Two main distinctions between the approaches should be noted: 1) our approach does not assume any previous evidence of haplotypes or LD, instead utilizing markers which could only have been inherited from exactly one of the direct parents to define IBD segments, and 2) individuals which would otherwise have the same estimated effect from a shared haplotype can now be assigned different estimated effects due to tracking exactly which ancestral line a haplotype was inherited from.

We test an approach hereafter named "Parental Allele Tracing, Recombination Identification, and Optimal predicTion" (PATRIOT) that utilizes raw marker data for tracking IBD inheritance of chromosome segments, enabling the rapid identification of lines carrying specific alleles, increasing the accuracy of genomic relatedness and diversity estimates, and improving genomic prediction and selection performance. Using the SoyNAM population (Song et al. 2017), which includes 39 parents crossed to a common parent and 5176 recombinant inbred lines, we explored the effectiveness of GS with additional information conferred with IBD. We traced chromosome segments from parent to progeny, followed by the calculation of an allelic score for each parental source of each SNP. These allelic scores were used in place of the raw marker data in order to allow the incorporation of IBD data into a GS pipeline.

## MATERIALS AND METHODS

### Pedigree records

Pedigrees for public breeding lines tested in the Uniform Soybean Tests were recorded based on reporting in their last year of testing in the Northern tests (https://www.ars.usda.gov/midwest-area/west-lafayette-in/crop-production-and-pest-control-research/docs/uniform-soybean-tests-northern-region/) or Southern tests (https://www.ars.usda.gov/southeast-area/stoneville-ms/crop-genetics-research/docs/uniform-soybean-tests/). Additional breeding records were recorded from cultivar release papers, primarily from Crop Science (https://acsess.onlinelibrary.wiley.com/journal/14350653), the Journal of Plant Registrations (https://acsess.onlinelibrary.wiley.com/journal/19403496), and Canadian Journal of Plant Science (https://www.nrcresearchpress.com/journal/cjps). Pedigree information for other lines in the NPGS soybean germplasm collection were downloaded from https://npgsweb.ars-grin.gov/gringlobal/search.aspx?. The pedigree information used in this study is provided **Supplemental File 1** and is also available from GitHub (https://github.com/SoylabSingh/PATRIOT).

### Marker data

**Soybean Nested Association Mapping (SoyNAM) panel:** SNP marker data for 5149 SoyNAM RILs, as well as their parents, were downloaded from SoyBase (https://soybase.org/SoyNAM/index.php), using the Wm82.a2 reference genome for downloading. For the SoyNAM panel, 4289 SNP markers were used in the analysis. Markers were re-ordered prior to tracing and imputation based on the composite linkage map created in previous work (Song et al. 2017). The ancestral source of each chromosome segment was identified using the pipeline illustrated in Figure 1.

**Released cultivars and isolines:** We used 868 lines that were the progeny of parental lines, and wherein both parent and progeny were genotyped with the SoySNP50k SNP set. Near-isogenic lines derived from backcrossing schema were also included in this pipeline, which may make some statistical metrics not applicable for this panel. SNP marker data for all accessions in the GRIN database were downloaded from Soybase.org (https://soybase.org/dlpages/#snp50k) as a VCF file, with positions annotated based on the Wm82.a2 reference genome. Pre-processing to remove SNPs aligned to scaffolds or the mitochondria left 42,080 SNP markers aligned to the Wm82.a2 reference genome and used in further analysis. Missing SNP data were imputed using Beagle 4.0 with default settings (Browning and Browning 2007). This panel will be referred to as the "868/50K panel" for brevity.

**Performance data**

Phenotypic records for the SoyNAM recombinant inbred line mapping population were downloaded from SoyBase (https://soybase.org/SoyNAM/index.php). Replicated entries' phenotypic records from within a single environment were used to calculate BLUPs for those lines, while unreplicated entries were incorporated using the raw phenotypic values. The "Corrected Strain" column was used to connect phenotypes with genotypic records. Phenotypic records were available from 2011 (IL and NE), 2012 (IA, IL, IN, KS, MI, MO, NE, OH[1], and OH[2]), and 2013 (IA, IL, IN, KS, and MO).

*Phytophthora* root rot resistance ratings were queried from the National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search.aspx) for each of the ancestors of the modern cultivar "Rend" (Nickell et al. 1999). "Rend" was selected for demonstration of the multi-generation chromosome segment tracing code due to both parents and all four grandparents

being genotyped with the same platform, as well as the major resistance gene segregating within the pedigree.

**PATRIOT workflow and code development**

PATRIOT workflow utilizes LD and haplotype in a novel way to improve genomic prediction. Specifically, this system allows for the tracing of chromosomal segments from the immediate parents to the offspring, and also to trace chromosomal segments through multiple generations. The allele tracing code outputs can be used as inputs into a modified genomic selection code, wherein the nominal class data are converted to numeric through the use of differences from the mean. Custom R scripts were developed to identify SNPs which could only come from one of the listed parents (hereafter "anchor markers", **Figure 1a, 1b**), followed by imputation of SNPs of fixed markers based on surrounding anchor markers (**Figure 1c**). Code for identifying anchor markers, imputation, multi-generation tracing, and recombination zone identification are available as **Code 1**, **Code 2**, **Code 3**, **Code 4**, respectively (**Supplementary File 2**). Genomic prediction was evaluated using rrBLUP in R with raw marker data and allele tracing alternatives (**Code 5**) (**Supplementary file 2**).

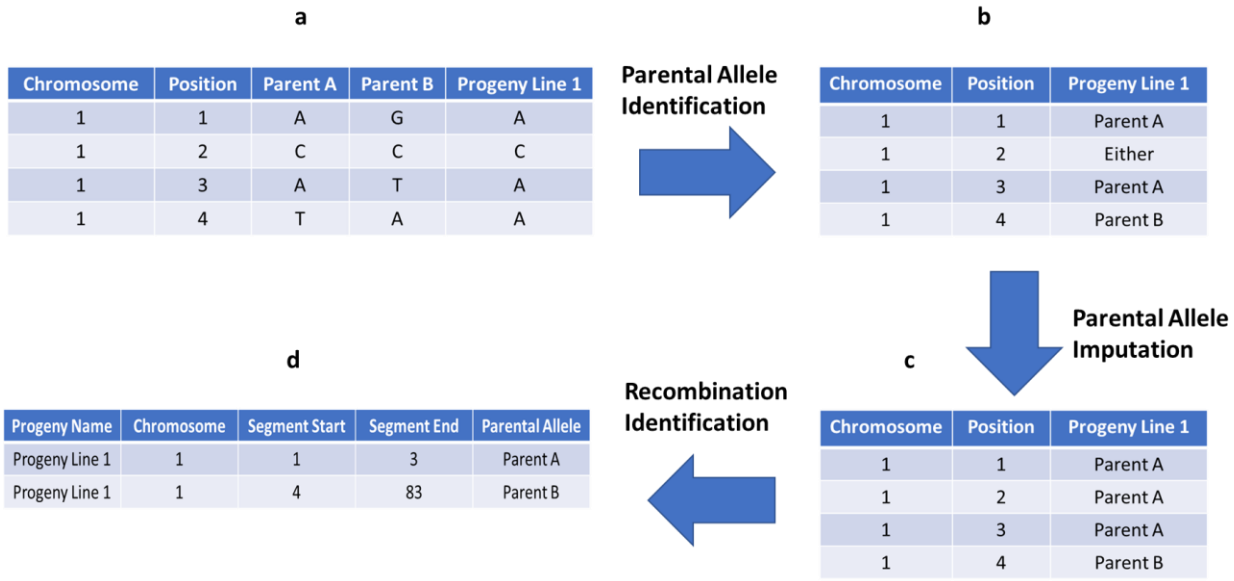



**Figure 1**. General workflow of Parental Allele Tracing, Recombination Identification, and Optimal predicTion (PATRIOT) input feature preparation for implementation in genomic selection: (a) Raw marker data is provided for both parent and progeny genotypes, b) Parental alleles encoded for those markers which can be conclusively traced to a specific parent, c) Alleles previously not assigned to a specific parent are imputed based on flanking markers, d) Those chromosome segments identical-by-descent from each parent are compiled. The "Position" column refers to the marker order and is provided only for demonstration purposes.

## Chromosomal tracing and Identity By Descent (IBD)

As a proof of concept, tracing of chromosome segment inheritance within the pedigree of soybean cultivar "Rend" was performed. After ensuring consistency between expected results and the outputs, chromosome tracing was performed on the remainder of the 868/50K panel. Following completion of the single-generation tracing pipeline, the multi-generation tracing script was run on traced lines to allow visualization of multiple generations of inheritance and recombination.

In addition to the 868/50K panel, SoyNAM project parents and RILs were investigated with the chromosome tracing pipeline. The A/B genotype representation data available from SoyBase was utilized to impute chromosomal segments. Even with a sparse marker coverage, recombination events were still identifiable (**Supplemental File 2**).

**Genomic prediction models**

To expand on the usefulness of the chromosome tracing pipeline outlined in **Figure 1**, we used the SoyNAM panel to evaluate accuracy of genomic prediction using ancestral alleles. Genomic prediction was evaluated for multiple traits using the 39 SoyNAM RIL populations based on the phenotypic records available from the SoyNAM project and all 4289 available markers. All comparisons were made using 80% of individuals for training and predicting on the remaining 20% of individuals. For each marker, an allelic differential estimator (ADE) was calculated as:

$$ADE_{ijk} = Average_{ik|SNP_j} - Average_{ik}\ , \quad [1]$$

for environment i, marker j, and trait k. The ADE value therefore is not regressed towards the mean to account for multiple regression. Instead, these values replace the marker representation as an input to GS models that evaluate the performance of this new approach (**Table 1**). They allow for the use of many distinct ancestral haplotypes in linear regression-based models based on the sign and relative scale of the estimated haplotype effect.

Traditional rrBLUP performance was evaluated using mixed.solve, a function in "rrBLUP" (Endelmann 2011). The rrBLUP-PATRIOT analysis was performed using mixed.solve, but replacing the marker input data (0,1,2) with a matrix of ADEs calculated in PATRIOT. The mean observed phenotype of lines with top 10% of predicted performance using rrBLUP and PATRIOT were compared, as well as the difference in phenotype between selected

lines and the base population. For yield, five-fold cross-validation was used to reduce sampling bias in the estimation of GP accuracy for each method.

**Table 1**. Simplified matrix showcasing 5 potential progeny, parents, and their ADEs. ADEs were calculated using the full panel (more than one family) and with unequal population sizes, so marker effects are not necessarily equal and opposite. Progeny 2 and 3 differ based on the site of recombination around SNP 3. Progeny 4 represents the optimal combination of ancestral alleles available from within that population, but not necessarily within the full panel. Progeny 5 represents the global optimum progeny from within the panel; however, this progeny would require multiple crosses to introgress segments from all three parents. ADE scale is shown based on potential values for yield in terms of kg $ha^{-1}$ in soybean.

| | **SNP 1** | **SNP 2** | **SNP 3** | **SNP 4** | **SNP 5** | **SNP 6** | **SNP 7** |
|---|---|---|---|---|---|---|---|
| Parent 1 | 45 | -23 | 70 | 14 | -56 | 73 | 15 |
| Parent 2 | -40 | 17 | -65 | -50 | -15 | -51 | 70 |
| Parent 3 | -53 | 20 | -71 | 106 | 69 | -36 | -43 |
| Progeny 1 | 45 | -23 | 70 | -50 | -15 | -51 | 70 |
| Progeny 2 | -40 | 17 | -65 | 14 | -56 | 73 | 15 |
| Progeny 3 | -40 | 17 | 70 | 14 | -56 | 73 | 15 |
| Progeny 4 | 45 | 17 | 70 | 14 | -15 | 73 | 70 |
| Progeny 5 | 45 | 20 | 70 | 106 | 69 | 73 | 70 |

# RESULTS

## Recombination Identification

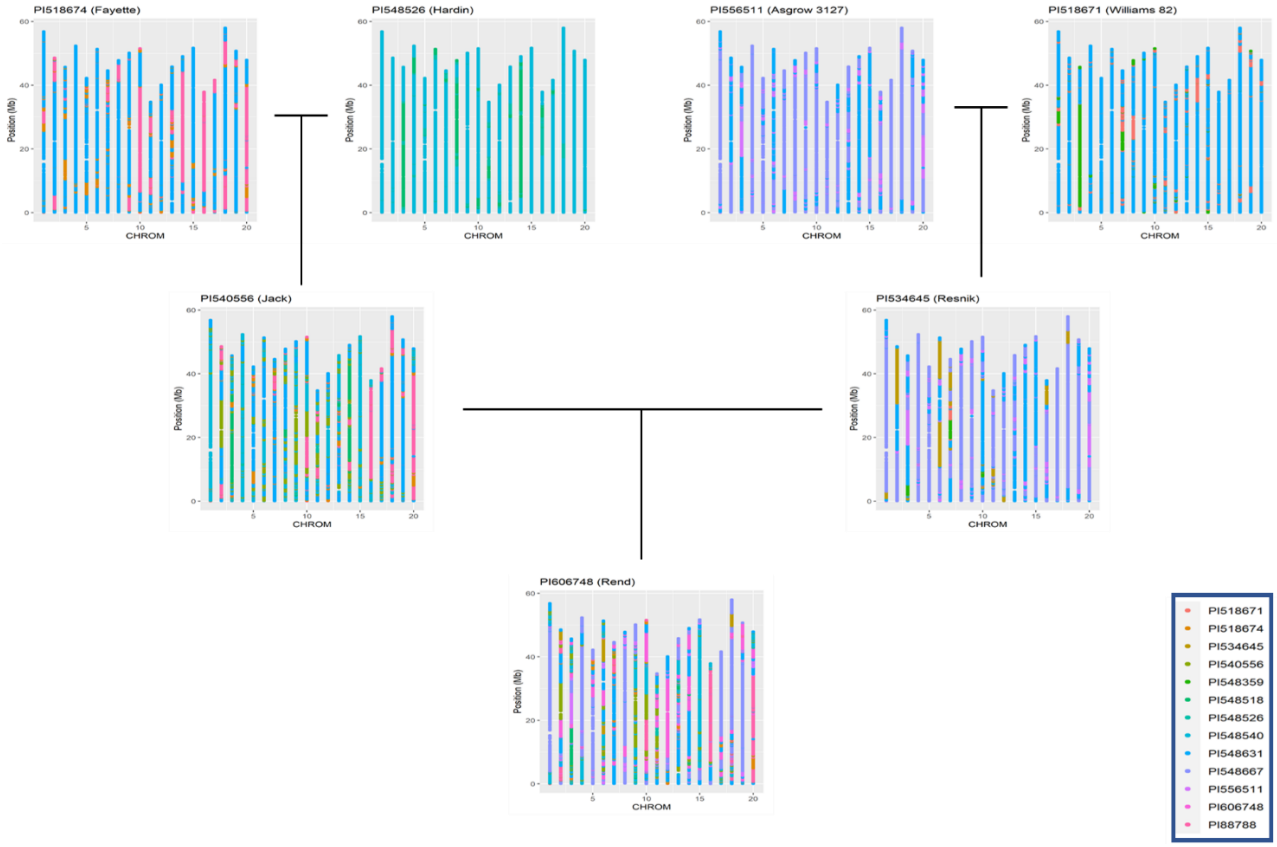
For the 868/50Kpanel, 13.14% of all SNPs were unassigned to a specific parent. For the SoyNAM panel, 6.78% of all SNPs were unable to be assigned to a specific parent. Using the SoyNAM panel marker data after PATRIOT IBD tracing and imputation, we examined the rates of recombination throughout the genome. Of the 5149 RILs examined, we found total recombinations per line ranged from 10 to 557, with an average of 50.9 recombinations per line. 18808/102960 (18.3%) chromosomes were inherited intact from one parent or another. 5011 RILs inherited at least one intact chromosome from a parent.

## Chromosomal Segment Tracing and Recombination Events

Chromosomal segments were traced in the 868/50K panel using the PATRIOT framework. To demonstrate the PATRIOT workflow, we trace the inheritance of the major *Phytophthora* root rot (PRR) resistance locus *Rps1* (**Figure 2**). Williams 82 (i.e., PI518671) inherited the *Rps1k* allele (that confers PRR resistance), as a long introgression (shown in green) on chromosome 3 from Kingwa (i.e., PI548359). This allele is then transmitted from Williams 82 to Resnik (i.e., PI534645) in a smaller chromosomal segment around *Rps1k*. However, the resistance allele was not passed on to Resnik's progeny, Rend (i.e., PI606748). Resnik is therefore more suitable than Rend to breed for Phytophthora resistance. Chromosomal tracing over multiple generations allows presence/absence characterization for the *Rps1k* allele without the need for allele-specific markers and can reduce the need for phenotyping in disease nurseries, as allele state is known by virtue of IBD. **Figure 2** gives a visual chromosomal segment tracing that is applicable to all varieties with available pedigree records that have been genotyped.

Recombination events can be visually identified when examining multiple generations within **Figure 2** (or similar plots) in two ways using the chromosome 3 example: (i) between Williams 82 and Resnik, the length of the green segment surrounding *Rps1k* is greatly reduced in Resnik, indicating recombination during the cross of Asgrow 3127[4] x Williams 82, and (ii) a segment of the soft red "AmbiguousParentage" class appears in the progeny, which indicates that recombination occurs somewhere within this region, but could not be delimited between two adjacent markers due to multiple markers being alike by state in the parents. This occurs in Asgrow 3127 (i.e., PI556511) on chromosome 3, separating large segments inherited from Williams and Essex.

While the *Rps1k* example is provided, the PATRIOT framework is applicable to trace chromosomal regions and for IBD characterization of important genes through generations, as well as to visualize nearby recombination events.



276 **Figure 2**. Scatterplot maps of chromosome segments inherited from ancestral sources, traced
277 through progenitors of soybean cultivar Rend (i.e., PI606748). Chromosome number (based on
278 Wms82.a2 reference genome) is plotted left to right on the x-axis, while position is plotted on the
279 y-axis.
280
281 **Comparison of genomic prediction accuracy using SoyNAM**
282 To examine the relative effectiveness of rrBLUP with PATRIOT (PATRIOT GS)
283 compared to traditional rrBLUP (rrBLUP GS), yield predictions for 16 environments from each
284 model were generated using the same randomized testing set for each model (Table 2). Results
285 from the two GS approaches are presented in **Table 2**. A 16.6% increase was attained in genomic
286 prediction accuracy by using PATRIOT GS compared with traditional rrBLUP (0.557 vs. 0.478).
287 Using a scenario of selecting 10% (and discarding 90%) from the SoyNAM RIL population and
288 comparing to the overall SoyNAM RIL population mean, PATRIOT GS had a 8.6% greater
289 selection differential among the selected RILs over basic rrBLUP GS (an increase of +538.7 in
290 PATRIOT GS vs +496.1 kg $ha^{-1}$ in rrBLUP GS).
291
292

293 **Table 2**. Comparison of the effectiveness of genomic selection methods rrBLUP GS and
294 PATRIOT GS for yield. In each environment, the best model for each metric is highlighted in
295 bold. Standard deviation given in parentheses. Each row contains results from a single site/year
296 combination; "2011 IL" is therefore from an experiment in Illinois during 2011.

| **Environment** | **Testing set mean** | **rrBLUP GS** | | **PATRIOT GS** | | **Marker-based heritability ($h^2$)** |
|---|---|---|---|---|---|---|
| | Testing set mean (kg $ha^{-1}$) | Average Yield, Top 10% (kg $ha^{-1}$) | Correlation between observed phenotype and BLUP | Average Yield, Top 10% (kg $ha^{-1}$) | Correlation between observed phenotype and BLUP | |
| 2011 IL | 2786.51 (12.26) | **3162.15 (59.99)** | 0.51 (0.02) | 3147.67 (59.58) | **0.53 (0.01)** | 0.471 |
| 2011 NE | 5048.98 (11.09) | **5538.78 (54.92)** | **0.62 (0.03)** | 5420.83 (47.24) | 0.57 (0.03) | 0.614 |
| 2012 IA | 2777.59 (10.38) | 3263.39 (47.66) | 0.43 (0.02) | **3313.28 (43.48)** | **0.51 (0.02)** | 0.438 |
| 2012 IL | 3390.61 (12.17) | 3766.04 (75.62) | 0.37 (0) | **3887.74 (45.01)** | **0.48 (0.02)** | 0.415 |
| 2012 IN | 4238.95 (12.81) | 4761.52 (29.91) | 0.5 (0.01) | **4773.01 (58.21)** | **0.56 (0.01)** | 0.508 |
| 2012 KS | 3875.14 (23.95) | 4187.91 (39.67) | 0.47 (0.03) | **4204.66 (33.05)** | **0.6 (0.02)** | 0.548 |
| 2012 MI | 2361.89 (32.5) | 2785.12 (139.67) | 0.57 (0.05) | **2849.09 (132.92)** | **0.73 (0.04)** | 0.613 |
| 2012 MO | 3458.43 (32.13) | 4254.08 (65.28) | 0.62 (0.04) | 4244.01 (79.29) | **0.64 (0.03)** | 0.603 |
| 2012 NE | 4723.31 (18.08) | 5251.94 (45.91) | 0.44 (0.03) | **5288.3 (52.85)** | **0.51 (0.02)** | 0.588 |
| 2012 OH[1] | 3402 (27.94) | 3826.54 (133.36) | 0.29 (0.04) | **3979.25 (167.09)** | **0.4 (0.05)** | 0.285 |
| 2012 OH[2] | 2811.38 | 3577.66 | 0.58 (0.04) | **3709.85** | **0.71 (0.04)** | 0.664 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | (16.06) | (99.49) | | **(103.35)** | | |
| 2013 IA | 2865.03 (15) | 3216.05 (13.77) | 0.39 (0.04) | **3266.87 (39.46)** | **0.47 (0.02)** | 0.435 |
| 2013 IL | 3113.94 (8.64) | 3434.06 (33.06) | 0.44 (0.03) | **3467.78 (28.75)** | **0.51 (0.02)** | 0.459 |
| 2013 IN | 5062.98 (17.78) | 5492.83 (49.1) | 0.38 (0.03) | **5521.03 (42.71)** | **0.45 (0.01)** | 0.427 |
| 2013 KS | 2759.03 (37.33) | 3323.76 (67.07) | 0.44 (0.02) | **3441.16 (63.83)** | **0.53 (0.01)** | 0.548 |
| 2013 MO | 4075.8 (11.47) | 4848.04 (39.33) | 0.61 (0.04) | **4855.79 (35.65)** | **0.71 (0.04)** | 0.604 |

[1] and [2] represent two separate tests grown in Ohio in 2012.

To help explain the cause of the difference in performance improvement between genomic prediction accuracy (+16.6%) and genomic selection effectiveness (+8.6%) (both compared to rrBLUP), we further examined the yield predictions from the 2012 OH[1] environment, which showed a large increase in GP accuracy (+39.5%) but only slight increase in genomic selection effectiveness (+3.8%). When examining the bottom 10% of predicted lines (rather than top 10% as before), the genomic selection effectiveness was 52.7% greater using PATRIOT than rrBLUP. This finding, coupled with smaller average absolute error terms using PATRIOT, suggests that the GP accuracy increase came from decreased error terms throughout the full range of phenotypes, allowing for better rankings. Indeed, using a 5% selection level for high GEBVs using PATRIOT resulted in a 29.8% increase in average observed phenotype compared to rrBLUP in the 2012 OH[1] set.

## DISCUSSION

Some of the earlier efforts in soybean chromosomal tracing involved RFLP markers, as researchers traced chromosome segments in 67 genotypes through generations (Lorenzen 1994). The transition to SNP markers as more mainstream marker technology enables better genome coverage to trace chromosomal segments from progenitors (Letcher and King 2001), with increased resolution for recombination identification (Yu et al. 2011). However, the biallelic nature of SNP markers is a limitation for more refined haplotype generation. In the 868/50K panel, 13.14% of all markers could not be definitively traced back to their ancestral source. While some portion of this unassigned group can be attributed to heterozygous allele state in either one of the parents or the progeny, a substantial portion is due to recombination in the affected area in which both parents are IBS at several consecutive markers. Less singletons were identified in the SoyNAM panel, possibly due to the non-identification of small double-recombination events, as well as due to the use of a linkage map for marker ordering, which by its nature reduces the likelihood of mis-ordering markers along a chromosome or linkage group. The genome tracing of large segments through multiple generations enables breeders to follow genes of interest throughout the pedigrees of modern lines (Bruce et al. 2020). This allows for a rapid identification of lines containing the desired allele even if allele specific markers are not available. Visualization of relatedness of lines based on IBD metrics similar to **Figure 2** allows breeders to rapidly identify pairings of lines with high genetic diversity as parents to create breeding families (Liu & Anderson 2003).

While IBD can be traced in many released public cultivars on the basis of markers from the SoySNP50K chip in soybean, applicability to breeding programs during the development of new pure lines requires a cost-effective genotyping system to allow genotyping of these lines at

an earlier stages of development. This can be achieved by utilizing a smaller, less expensive genotyping array such as the SoyNAM6K BeadChip (Song et al. 2017) to genotype experimental lines.

The PATRIOT framework facilitates the identification of lines for breeding purposes that have favorable genes linked in coupling, as well as in situations where breaking the linkage drag is imperative. For example, SCN resistance from PI494182 was determined to carry a risk of linkage drag (St-Amour et al. 2020). Likewise, SCN resistance from the commonly used donor PI88788 was initially associated with considerable linkage drag (Cregan et al 1999). With the use of PATRIOT, parents can be readily identified which contain the gene(s) of interest with the least amount of additional introgressed region(s), thereby reducing the likelihood of linkage drag, and concurrently deploy it in a GS pipeline. With an additional generation of traced progeny, those regions negatively associated with another trait can be identified to inform marker based decisions.

Much like genome-wide association studies (GWAS), genomic prediction and genomic selection models rely on the association between markers and the phenotype of interest. However, the association between marker and phenotype decays in subsequent generations, leading to reduced accuracy without re-training of the model (Habier et al. 2007, Hayes et al. 2009, Jannick 2010). With the chromosome tracing approach, the linkage between marker and phenotype withstands this shift, since parental allele representation is directly incorporated into the marker data. The prediction accuracy decays much more slowly with chromosome tracing because the linkage between marker and phenotype decays only as recombination between marker and underlying gene(s) occurs. Furthermore, multi-generation tracing allows the

357 preservation of information on lineage-specific marker association which can better model the
358 differences in genes linked to a particular marker or set of markers.

359 The widespread use of PATRIOT GS would be encouraged by the establishment of a
360 fully connected pedigree (fully known relationships between all germplasm utilized) and
361 development of base population resources with equal and ample representation of each parental
362 source within the breeding pool. For example, while the SoyNAM panel can be readily used as a
363 training set for materials derived from any combination of the 40 parents, it's efficacy is limited
364 to that context, with the exception of including a small number of the parents' ancestors within
365 the pool. Instead, in some situations, breeding applications would benefit from the development
366 of fully inter-related populations derived from the founder lines, such as through MAGIC design
367 (Xiao et al 2013, Matteo et al 2015) or a NAM population created with founder parents (Yu et al.
368 2008) that can happen in different crossing cycles. Moreover, most breeding programs have an
369 inherent nested design especially when a few superior parents are used extensively in the
370 development of breeding populations, therefore this effort is not incremental.

371 The multi-generation chromosome segment tracing aspect of PATRIOT can also be used
372 as a tool to connect QTL mapping studies among related populations. In addition to tracing
373 chromosomal regions within a pedigree, this framework can be used to connect linkage mapping
374 studies using related lines as parents by tracing QTL regions identified in related parents in
375 separate studies to their ancestral sources. This allows for a meta-analysis to utilize the increased
376 power which comes from having multiple mapping populations with common ancestry to map
377 marker-trait associations.

378 However, there are challenges to the PATRIOT framework. In crosses where parents
379 share large runs of IBS or IBD based on marker data, it is difficult to determine which parent is

contributing each allele to the progeny. However, if these runs are IBD, the effect on allele estimation is equivalent, regardless of which parent is assigned to the allele. Additionally, a surprising number of singleton marker calls suggests that either double recombination is occurring at a much higher rate than previously believed, or that the reference genome assembly order does not agree with the true marker order. Increased marker density can overcome some of these challenges. Likewise, uncertain regions can be assigned new allele effect classes. For example, Williams 82 (PI518671) has 3,399 out of 42,080 markers which could not be assigned with certainty to a specific parent (Williams or Kingwa). To circumvent this challenge, each of these markers was assigned a new parent class of "PI518671" when tracking segments passed on to progeny, but continue to use ADEs based on the average ADE of parents Williams and Kingwa when predicting its own performance.

PATRIOT genomic prediction accuracy for yield using all populations was greater than the calculated marker-based heritability of the trait in 13 of 16 environments (**Table 2**), suggesting that genomic prediction using ancestral allele tracing can perform better than traditional genomic prediction. Generating separate prediction models in this way for each environment can also be used to reduce the number of environments needed for phenotypic evaluation, as the prediction accuracy very nearly reaches the heritability of the trait itself. The fact that this high level of prediction accuracy was possible with a 6K SNP chip in the SoyNAM populations suggest significant potential cost savings, as the cost of genotyping at this density is less expensive than growing and phenotyping in replicated field plots (Xu et al. 2020). More generally speaking, if small arrays are to continue to be used in community research projects, the array needs to be carefully designed to provide adequate coverage throughout the genome. Consideration of both linkage distance and optimal SNP selection in genic regions should be

made a priority. Alternatively, other genotyping platforms such as genotyping-by-sequencing (GBS) can be used to implement this approach, which is able to decrease the negative impact of missing data that is common from GBS (Gardner et al. 2014).

While our genomic prediction models utilized only the immediate parents for calculating allele effect estimates, it is possible to expand the method by combining with the multi-generation IBD tracing script. This approach has an added benefit of bridging the gap between populations that do not share a direct parent but share ancestors in previous generations. By doing so, an increased number of lines can be used for allele effect estimation, further improving the accuracy of these values.

IBD-based genomic selection has the clear potential to improve selection accuracy over existing genomic selection approaches. However, there is a tradeoff due to the significant increase in computational time. While the chromosome segment tracing portion of the workflow need only be run once for any particular genotype, the ADE matrix must be calculated separately for each trait and environment. Fortunately, this calculation can be parallelized, and only needs to be performed for the training population. Typical computation time on an AMD Ryzen Threadripper 1950X for ADE matrix calculation was on the order of one minute without parallelization of the code, while the genomic prediction itself took on the order of three minutes for a dataset with 2500 individuals and 4289 markers. Computation time for the tracing and imputation of alleles within the SoyNAM study totaled 7 hours 41 minutes. However, minor modifications to run each chromosome in parallel on a different computational thread can reduce the wall time to around 35 minutes.

## CONCLUSION

PATRIOT provides a framework for identifying, tracking, and applying IBD information in order to increase effectiveness of genomic selection. Tracking IBD with PATRIOT enables pedigree-based gene tracking through generations, which can be useful for parental selection, as well as for predicting phenotypes for monogenic and oligogenic traits. Relatedness metrics within breeding populations can also be improved due to the specification of IBD allele sharing rather than IBS. The IBD information also works to improve genomic prediction and selection results. This improvement was shown in first-cycle genomic prediction, but should provide additional benefits in later cycles due to the donor-specific allele effect estimation, which does not suffer from the problem of population shift between training and testing sets. The large and consistent benefit shown suggests that chromosome tracing is a quick and efficient way to increase the accuracy of genomic selection models, with no additional cost beyond modestly increased computational time.

**Contributions:** JMS conceptualized the project with AKS; JMS conducted the statistical analysis with suggestions from AKS and DL. JMS and AKS prepared the first draft. All authors contributed in writing and editing the manuscript.

**Acknowledgements:** Authors sincerely appreciate inputs from Dr. David Grant (USDA-ARS, retired), and Dr. Rex Nelson (USDA-ARS) for assistance with pedigree compilation and suggestions on potential applications for the method. We thank Dr. Kulbir Sandhu, Sarah Jones, and Liz van der Laan for reviewing the manuscript draft.

**Funding:** Authors sincerely appreciate the funding support from Iowa Soybean Association, R F Baker Center for Plant Breeding, Bayer Chair in Soybean Breeding and USDA CRIS project

(IOW04314). Part of JMS graduate assistance was provided by the NSF NRT (graduate fellowship).

**Pedigree and Code availability:** https://github.com/SoylabSingh/PATRIOT

## REFERENCES

Avolio, M.L., J.M. Beaulieu, E.Y.Y. Lo, and M.D. Smith, 2012 Measuring genetic diversity in ecological studies. 213 (7):1105-1115.

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., … Poland, J. A. (2016). Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. *Plant Genome*, 9, 1– 12.

S R Browning and B L Browning (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097. doi:10.1086/521987

Bruce, R.W., Torkamaneh, D., Grainger, C.M. *et al.* Haplotype diversity underlying quantitative traits in Canadian soybean breeding germplasm. *Theor Appl Genet* **133,** 1967–1976 (2020). https://doi.org/10.1007/s00122-020-03569-1

Burdon, R.D. 2001. Genetic diversity and disease resistance: some considerations for research, breeding, and deployment. *Can J For Res.* 31:596-605.

Coser SM, Chowda Reddy RV, Zhang J, Mueller DS, Mengistu A, Wise KA, Allen TW, Singh A and Singh AK (2017) Genetic Architecture of Charcoal Rot (*Macrophomina phaseolina*) Resistance in Soybean Revealed Using a Diverse Panel. *Front. Plant Sci.* 8:1626. doi: 10.3389/fpls.2017.01626

Cregan, PB, Mudge, J, Ficus, EW, Danesh, D, Denny, R, and Young ND. 1999 Two simple sequence repeat markers to select for soybean cyst nematode resistance conditioned by the rhg1 locus. Their Appl Genet 99:811-818.

Cui, Z, T.E. Carter, Jr., J. Gai, J. Qiu, and R.L. Nelson. 1999. Origin, Description, and Pedigree of Chinese Soybean Cultivars from 1923 to 1995. U.S. Department of Agriculture, Agricultural Research Service, Technical Bulletin No. 1871. 263 pp.

Daly, M., Rioux, J., Schaffner, S. *et al.* High-resolution haplotype structure in the human genome. *Nat Genet* **29,** 229–232 (2001). https://doi.org/10.1038/ng1001-229

Dawn Teare, M., & Barrett, J. H. (2005). *Genetic linkage studies. The Lancet, 366(9490), 1036–1044.*

de Azevedo Peixoto, L., T.C. Moellers, J. Zhang, A.J. Lorenz, L.L. Bhering et al., 2017 Leveraging genomic prediction to scan germplasm collection for crop improvement. PLOS ONE 12 (6):e0179191.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255.

Gardner, K.M., P. Brown, T.F. Cooke, S. Cann, F. Costa *et al.*, 2014 Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing. *G3: Genes|Genomes|Genetics* 4 (9):1681.

Garrido-Cardenas, J.A., C. Mesa-Valle, and F. Manzano-Agugliaro, 2018 Trends in plant research using molecular markers. 247 (3):543-557.

Gizlice, Z., T.E. Carter Jr, and J.W. Burton, 1994 Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988. Crop Science 34 (5):cropsci1994.0011183X003400050001x.

494 Graham, J., 2011 Molecular Plant Breeding By Y. Xu. Wallingford, UK: CABI (2010), pp.
495 734. ISBN 978-184593-392-0. Experimental Agriculture 47 (1):173-173.
496 Habier, D., Fernando, R.L., and Dekkers, J.C.M. 2007. Genomic selection across multiple
497 breeding cycles in applied bread wheat breeding. *Genetics* 177(4): 2389-2397.
498 Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the Bayesian
499 alphabet for genomic selection. *BMC bioinformatics*, *12*(1), 186.
500 Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the
501 realized relationship matrix. *Genetics research*. 2009;91:47–60.
502 Hickey, J., Chiurugwi, T., Mackay, I. *et al.* Genomic prediction unifies animal and plant
503 breeding programs to form platforms for biological discovery. *Nat Genet* **49,** 1297–1303
504 (2017). https://doi.org/10.1038/ng.3920
505 Hyten DL, Song Q, Zhu Y, et al. Impacts of genetic bottlenecks on soybean genome diversity.
506 *Proc Natl Acad Sci U S A*. 2006;103(45):16666-16671. doi:10.1073/pnas.0604379103
507 Jannick, J.L. 2010. Dynamics of long-term genetic selection. *Genet Sel Evol.* 42(1):35
508 Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic
509 value prediction accuracy. *Genetics*, *192*(4), 1513-1522.
510 Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H.,
511 Cordell, H.J., Eaves, I.A., Dudbridge, F. et al.(2001) Haplotype tagging for the
512 identification of common disease genes. Nature Genet., 29, 233–237.
513 Kisha, T.J., Diers, B.W., Hoyt, J.M. and Sneller, C.H. (1998), Genetic Diversity among
514 Soybean Plant Introductions and North American Germplasm. Crop Science, 38: 1669-
515 1680

516 Langewisch T, Zhang H, Vincent R, Joshi T, Xu D, Bilyeu K (2014) Major Soybean Maturity
517 Gene Haplotypes Revealed by SNPViz Analysis of 72 Sequenced Soybean Genomes.
518 PLoS ONE 9(4): e94150. https://doi.org/10.1371/journal.pone.0094150
519 Letcher B.H., King T.L., Parentage and grandparentage assignment with known and unknown
520 matings: application to Connecticut River Atlantic salmon restoration, Can. J. Fish.
521 Aquat. Sci. 58(2001)1812–1821
522 Levings C.S., Siedow J.N. (1992) Molecular basis of disease susceptibility in the Texas
523 cytoplasm of maize. In: Schilperoort R.A., Dure L. (eds) 10 Years Plant Molecular
524 Biology. Springer, Dordrecht
525 Li, H., Glusman, G., Hu, H., Caballero, J., Hubley, R., Witherspoon, D., Guthery, S.L.,
526 Mauldin, D.E., Jorde, L.B., Hood, L. and Roach, J.C., 2014. Relationship estimation from
527 whole-genome sequence data. *PLoS Genet*, *10*(1), p.e1004144.
528 Liu S.X, Anderson J.A. Marker assisted evaluation of *Fusarium* head blight resistant wheat
529 germplasm. Crop Sci. 2003;43:760–766.
530 Lorenzen L.L. 1994, Soybean cultivar development: a genome perspective. Ph.D. Dissertation,
531 Iowa State University, Ames, IA. Available at:
532 http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=11625&context=rtd (12 July 2020,
533 date last accessed).
534 Lynch, M. and Walsh, B. 1998. *Genetics and Analysis of Quantitative Traits.* Sinauer
535 Associates, Sunderland, MA.
536 Matteo, D. A. et al., Genetic properties of the MAGIC maize popu-lation: a new platform for
537 high definition QTL mapping in Zea mays. Genome Biol., 2015, 16, 167;
538 doi:10.1186/s13059-015-0716-z

539 Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001. Prediction of total genetic value
540 using genome-wide dense marker maps. Genetics 157:1819-1829.
541 Moellers, T.C., Singh, A., Zhang, J. *et al.* Main and epistatic loci studies in soybean for
542 *Sclerotinia sclerotiorum* resistance reveal multiple modes of resistance in multi-
543 environments. *Sci Rep* **7,** 3554 (2017). https://doi.org/10.1038/s41598-017-03695-9
544 Morrison, J. (2013), Characterization and Correction of Error in Genome-Wide IBD
545 Estimation for Samples with Population Structure. Genet. Epidemiol., 37: 635-641.
546 Nickell, C.D., G.R. Noel, T.R. Cary, D.J. Thomas, D.D. Hoffman. 1999. Rend Soybean. Crop
547 Sci. (Madison) 39(5):1533
548 Patil, G., Do, T., Vuong, T. *et al.* Genomic-assisted haplotype analysis and the development of
549 high-throughput SNP markers for salinity tolerance in soybean. *Sci Rep* **6,** 19199 (2016).
550 https://doi.org/10.1038/srep19199
551 Riquet, J., W. Coppieters, N. Cambisano, J.-J. Arranz, P. Berzi *et al.*, 1999 Fine-mapping of
552 quantitative trait loci by identity by descent in outbred populations: Application to milk
553 production in dairy cattle. *Proceedings of the National Academy of Sciences* 96
554 (16):9252.
555 Sebastian, S, Streit, LG, Stephens, PA, Thompson, JA, Hedges, BR, Fabrizius, MA, Soper, JF,
556 Schmidt, DH, Kallem, RL, Hings, MA, Feng, L, and Hoeck, JA. 2010. Context-specific
557 marker-assisted selection for improved grain yield in elite soybean populations. Crop Sci
558 50(4):1196-1206.
559 Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the
560 medical future. *Nat Rev Genet*. 2008;9(6):477-485. doi:10.1038/nrg2361

561 Smith, S. Pedigree background changes in U.S. hybrid maize between 1980 and 2004. *Crop*
562 *Sci.* 2007 47 1914– 1926
563 Solberg, T. R., Sonesson, A. K., Woolliams, J. A., & Meuwissen, T. H. E. (2008). Genomic
564 selection using different marker types and densities. *Journal of animal science*, *86*(10),
565 2447-2454.
566 Song, Q, Hyten, D.L., Jia, G., Quigley, C.V., Fickus, E.W., Nelson, R.L., and Cregan, P.B. G3:
567 Genes, Genomes, Genetics October 1, 2015 vol. 5 no. 10;
568 https://doi.org/10.1534/g3.115.019000
569 Song, Q., L. Yan, C. Quigley, B.D. Jordan, E. Fickus et al., 2017 Genetic Characterization of
570 the Soybean Nested Association Mapping Population. The Plant Genome 10
571 (2):plantgenome2016.2010.0109.
572 Sorrells M.E. (2015) Genomic Selection in Plants: Empirical Results and Implications for
573 Wheat Breeding. In: Ogihara Y., Takumi S., Handa H. (eds) Advances in Wheat
574 Genetics: From Genome to Field. Springer, Tokyo. https://doi.org/10.1007/978-4-431-
575 55675-6_45
576 Tanksley, S.D., and McCouch, S.L. 1997. Seed banks and molecular maps: unlocking genetic
577 potential from the wild. Science 277, 1063-1066.
578 Thompson E.A. 2013 Identity by Descent: Variation in Meiosis, Across Genomes, and in
579 Populations. Genetics, 194, 301:326.
580 Tylka G, 2007. Current Status of the Soybean Cyst Nematode as aThreat to Soybean
581 Production in the Midwest. Proceedings of the Integrated Crop Management Conference.
582 https://doi.org/10.31274/icm-180809-892

583 Xiao, F. L., Zhi, X. L., Dong, B. L., Yan, Z. L., Xing, X. M., Zhi, X. L. and Hua, J. L.,
584 Development and evaluation of multi-genotype varieties of rice derived from MAGIC
585 lines. Euphytica, 2013, 192, 77–86; doi:10.1007/s10681-013-0879-1.
586 Xu, Y., X. Liu, J. Fu, H. Wang, J. Wang *et al.*, 2020 Enhancing Genetic Gain through Genomic
587 Selection: From Livestock to Plants. *Plant Communications* 1(1):100005
588 Yao, J., Zhao, D., Chen, X., Zhang, Y., and Wang, J. 2018. Use of genomic selection and
589 breeding simulation in cross prediction for improvement of yield and quality in wheat
590 (*Triticum aestivum* L.). *The Crop Journal* 6(4):353-365.
591 Yu, J., Holland, J.B., McMullen, M.D., and Buckler, E.S.. 2008. Genetic design and statistical
592 power of nested association mapping in maize. Genetics 178:539–551
593 Yu, H. *et al.* Gains in QTL detection using an ultra-high density SNP map based on population
594 sequencing relative to traditional RFLP/SSR markers. *PLoS ONE* **6**, e17595 (2011).
595 Yu, X., Li, X., Guo, T. *et al.* Genomic prediction contributing to a promising global strategy to
596 turbocharge gene banks. *Nature Plants* **2,** 16150 (2016).
597 https://doi.org/10.1038/nplants.2016.150
598 Zhang, J., Naik, H., Assefa, T. *et al.* Computer vision and machine learning for robust
599 phenotyping in genome-wide studies. *Sci Rep* **7,** 44048 (2017).
600 https://doi.org/10.1038/srep44048