Journal of Advanced Research 42 (2022) 117–133

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Article

A novel Synthetic phenotype association study approach reveals the landscape of association for genomic variants and phenotypes

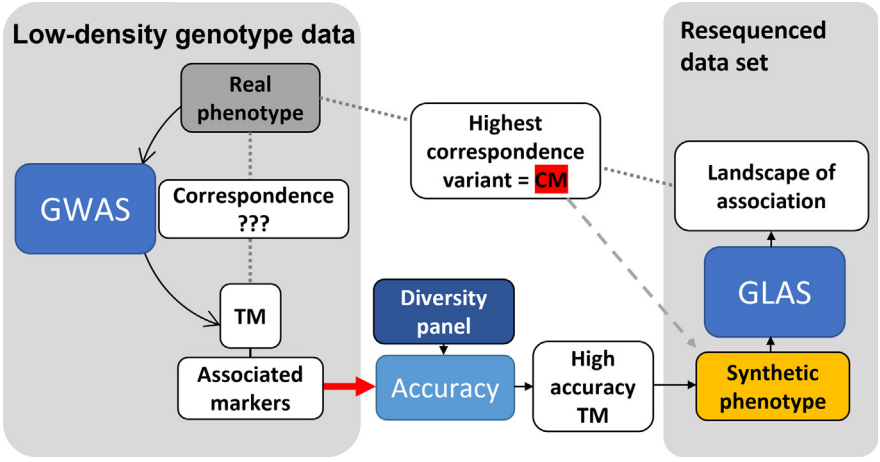
Mária Škrabišová [a], Nicholas Dietz [b], Shuai Zeng [c,d], Yen On Chan [d,e], Juexin Wang [c,d], Yang Liu [d,e], Jana Biová [a], Trupti Joshi [c,d,e,f,*], Kristin D. Bilyeu [g,*]

[a] Department of Biochemistry, Faculty of Science, Palacky University Olomouc, Olomouc 78371, Czech Republic
[b] Division of Plant Sciences, University of Missouri, Columbia, MO 65201, USA
[c] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65212, USA
[d] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65212, USA
[e] MU Data Science and Informatics Institute, University of Missouri, Columbia, MO 65212, USA
[f] Department of Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO 65212, USA
[g] Plant Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, University of Missouri, Columbia, MO 65211, USA

HIGHLIGHTS

- This study proposes additional post-GWAS evaluation criteria.
- Accuracy serves as a measure of direct correspondence between variant positions and phenotypes.
- Every genomic variant position can be used as a Synthetic phenotype in GWAS.
- SPAS reveals the landscape of association for genomic variants.
- Synthetic phenotype leverages resequenced data set information.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 4 November 2021
Revised 14 February 2022
Accepted 8 April 2022
Available online 12 April 2022

Keywords:
GWAS

ABSTRACT

Introduction: Genome-Wide Association Studies (GWAS) identify tagging variants in the genome that are statistically associated with the phenotype because of their linkage disequilibrium (LD) relationship with the causative mutation (CM). When both low-density genotyped accession panels with phenotypes and resequenced data accession panels are available, tagging variants can assist with post-GWAS challenges in CM discovery.

Objectives: Our objective was to identify additional GWAS evaluation criteria to assess correspondence between genomic variants and phenotypes, as well as enable deeper analysis of the localized landscape of association.

Peer review under responsibility of Cairo University.

* Corresponding authors at: Department of Health Management and Informatics, School of Medicine, 1201 E Rollins St, 271B Life Science Center, Columbia, MO 65201, USA (T. Joshi). Plant Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, 110 Waters Hall, University of Missouri, Columbia, MO 65211, USA (K.D. Bilyeu).

E-mail addresses: joshitr@missouri.edu (T. Joshi), kristin.bilyeu@usda.gov, bilyeuk@missouri.edu (K.D. Bilyeu).

https://doi.org/10.1016/j.jare.2022.04.004
2090-1232/© 2022 The Authors. Published by Elsevier B.V. on behalf of Cairo University.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Soybean
Genomics
Resequencing
Genotyping
Phenotyping

Methods: We used genomic variant positions as Synthetic phenotypes in GWAS that we named "Synthetic phenotype association study" (SPAS). The extreme case of SPAS is what we call an "Inverse GWAS" where we used CM positions of cloned soybean genes. We developed and validated the Accuracy concept as a measure of the correspondence between variant positions and phenotypes.
Results: The SPAS approach demonstrated that the genotype status of an associated variant used as a Synthetic phenotype enabled us to explore the relationships between tagging variants and CMs, and further, that utilizing CMs as Synthetic phenotypes in Inverse GWAS illuminated the landscape of association. We implemented the Accuracy calculation for a curated accession panel to an online Accuracy calculation tool (AccuTool) as a resource for gene identification in soybean. We demonstrated our concepts on three examples of soybean cloned genes. As a result of our findings, we devised an enhanced "GWAS to Genes" analysis (Synthetic phenotype to CM strategy, SP2CM). Using SP2CM, we identified a CM for a novel gene.
Conclusion: The SP2CM strategy utilizing Synthetic phenotypes and the Accuracy calculation of correspondence provides crucial information to assist researchers in CM discovery. The impact of this work is a more effective evaluation of landscapes of GWAS associations.
© 2022 The Authors. Published by Elsevier B.V. on behalf of Cairo University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Even though GWAS have successfully identified many thousands of genetic associations, the ratio of conducted GWAS to successful GWAS-derived findings is far from being balanced [1,2]. This is indeed caused by the nature of genetic interconnections such as pleiotropy and/or epistasis that lead to statistical restrictions due to high dimensionality and multicollinearity [3]. For GWAS, identification of a phenotype-associated haplotype-tagging variant position with high correspondence among other variant positions is crucial for correct identification of the associated genomic region. The fact that the highest associated tagging variant is not the physically closest variant to a CM, but instead is in strong LD with it, is a key feature of GWAS that is strikingly neglected [4,5]. This is especially problematic when low-density genotype data (less than 1% of total variants) is used that is unlikely to have the CM present in the genotype data set. Low-density genotype data sets are relatively inexpensive and therefore widely used in GWAS. Genotyping efforts must balance the cost and effort in capturing the genomic variation with the size and power needed for association panels. It is therefore of great interest to improve strategies to identify candidate genes containing the CM starting with low-density genotype data sets when whole genome resequenced data sets are available, as they are for many species including soybean (*Glycine max* [L.] Merr.) [6–11].

Our first attempts to explore the landscape of genomic variation resulted in the online tool, named SNPViz [12], that enables haplotype visualizations and was recently enhanced with new features and data sets (SNPViz v2.0 [13,14]). In the process of expanding the capabilities of SNPViz, we developed a novel analysis method enhanced by a GWAS to Genes concept. In this work, we present demonstrative experiments that show the usefulness of the new concepts. Fundamentally, GWAS requires a defined accession panel that has both determined phenotypes and genotype data for each accession; we have redefined the concept of what can be represented as phenotype data and refer to that as a Synthetic phenotype. A Synthetic phenotype can be any single genomic variant position present in the data set. For low-density genotypes, each marker position can be used as a Synthetic phenotype. Analogously for resequenced data sets, the Synthetic phenotype can be derived from any variant position that can be a SNP or an insertion/deletion (InDel), or other sequence structural variation that is of binary nature.

In a perfect GWAS of a simplified example of a qualitative phenotype (real phenotype), a bi-allelic genomic variant at a certain position present in resequenced genotype data would exactly match the binary variation in the phenotype [15]. In such a simplified case, this genomic variant would be the CM underlying the phenotype. Therefore, the allele status (reference and alternate) of such a variant position could be used as a Synthetic phenotype in GWAS (SPAS) and illuminate all the other associated variants that are in LD with the phenotype with the same statistical significance as the original real/observed phenotype. Applying the concept of the Synthetic phenotype to the CM variant position and running GWAS backward will produce a landscape of variation for all other variant positions in LD with the CM, and thus the phenotype (Fig S1a). Using the CM as a Synthetic phenotype is the extreme case of SPAS that we refer to as Inverse GWAS. Thus, in this perfect GWAS example, the landscape of association for the real phenotype would be identical with the landscape of association for the CM produced by Inverse GWAS (Fig. S1a). On the contrary, in the low-density genotype-based GWAS discoveries, because of the nature of low-density genotype data, the correspondence of an identified low-density tagging variant (tagging marker, TM) to the phenotype is rarely perfect but always unknown to the CM *a priori* (Fig. S1b). We successfully used the CM as a Synthetic phenotype to identify markers corresponding to the CM (Proxy markers) in our former work [16,17].

Improving GWAS is an active area of research taking on many different approaches. Besides fitting statistical models, there are a number of emerging within-GWAS methods that aim to rapidly identify a casual gene by putting weight on various trait-genome characteristics (such as LD [18]). Although LD-based weighting is a common approach to evaluate GWAS results, accuracy and efficiency [19,20] were also used in previous studies. Nevertheless, these approaches did not adequately characterize variants in terms of sensitivity and specificity to express how well the allele status of a variant position relates to a phenotype. We extended the standard GWAS statistical significance of a variant position by a parameter that is a simple mathematical measure of direct correspondence, a concept that we refer to herein as Accuracy.

Soybean offers several high-quality resequenced data sets [6,7,9,21] as well as reference genomes [22]. The USDA Soybean Germplasm Collection (GRIN, Urbana, IL) repository contains ~20,000 accessions genotyped with the low-density Illumina SoySNP50K bead chip (including accessions from the resequenced data sets) [23]. Therefore, the GRIN collection serves as an immense pool of available soybean GWAS results for many phenotypes in which TMs were identified [24,25]. Thus, soybean data enabled us to test our SPAS on a palette of various traits, phenotypes and data sets, and here we present our findings on a selection of three demonstrative soybean phenotypes for pod shattering (quantitative phenotype) as well as flower color and stem termination traits (qualitative phenotypes). Furthermore, we used our

approach to identify a CM for a novel pod shatter gene *NST1A* [26]. Here we show how our Synthetic phenotype and Accuracy calculation concepts can be utilized in CM identification. We created a novel SP2CM strategy where knowledge gained from GWAS with low-density genotypes for a large set of phenotyped accessions can be used to leverage an association panel of a limited number of resequenced accessions for which the desired phenotype information may not be available. Thus, the value of high-density information of an association panel of resequenced accessions can be exploited without phenotyping or resequencing another panel of accessions. The SP2CM strategy improves the effectiveness of TMs to identify a CM - the fundamental goal in GWAS. To measure direct correspondence using Accuracy calculations with the highest possible confidence for each type of association study, we created a curated panel of accessions for soybean (Soy775 accession panel) that consists of all publicly available otherwise independent resequenced data sets. We demonstrate utilization of the SP2CM strategy on CM identification of a novel pod shattering gene *NST1A*. We envision that our strategy could directly improve low-density genotype-based GWAS discoveries that could be potentially utilized in every other species as well.

Material and methods

Defining SPAS, Inverse GWAS, SP2CM and workflow

SPAS follows the same principles and rules as any other GWAS; it requires two essential components: genotype input data and phenotype input data. SPAS, as well as Inverse GWAS, can be performed on either low-density genotyping data or on resequenced data depending on what landscape is going to be explored. Inverse GWAS on low-density data can be performed on a data set of resequenced accessions where a CM is present in the genotype and thus, can be transformed into the Synthetic phenotype and used in GWAS on all chromosomes with genotypes from either the SoySNP50K bead chip or a subset of the SoySNP50K positions directly extracted from the resequenced data. We also performed Inverse association on resequenced genotype data but only on a localized part of a single chromosome; therefore, we refer to it as to the Inverse GLAS (genome-localized association study). Proxy markers are generated from resequenced data in Inverse association. SP2CM consists of two parts. Both parts include an association and Accuracy calculation step. Part 1 associates a real phenotype with low-density genotyping data of a panel of accessions in GWAS with all chromosomes, and Part 2 associates a Synthetic phenotype in GLAS.

Data sets

Basically, in this work, we used two types of genotype data: publicly available soybean resequenced data sets and genotyped accessions form the USDA germplasm collection (GRIN, Urbana, IL). The majority of accessions from the resequenced data sets were also genotyped at low-density - the USDA Agricultural Research Service Soybean Genomics Group has genotyped the entire USDA Soybean Germplasm Collection counting over 22,000 accessions with the Illumina Infinium SoySNP50K Illumina Infinium BeadChip [23] (https://www.soybase.org/snps). We downloaded SoySNP50K haplotypes (https://soybase.org/snps/download.php) for all accessions with known phenotype information for our three demonstrative traits: pod shattering score, flower color and stem termination; the downloaded data was used for GWAS of SP2CM-Part 1.

The resequenced soybean data sets used in this work were previously published [6,7,11]: 302 wild and cultivated accessions (Zhou, sequencing coverage 11x) [6], 106 soybean genomes genomic diversity and trait discovery project (MSMC, sequencing coverage 17x) [7], and 481 diverse soybean accessions from genetic variation project [11] two data sets with sequencing coverage 15x: USB15x, Soja15x; and one data set with sequencing coverage 40x - USB40x. For simplicity, the data sets were named in the following format: Name + Average sequencing depth+(number of unique accessions in the data set). We aggregated all the accessions from these data sets into a single panel of curated accessions (Soy775 accession panel; described below). We used this Soy775 accession panel for Accuracy calculation and for Inverse GWAS on low-density genotype data. Otherwise, we used the USB15x (302) resequenced data set that served as a model data set representing an ideal compromise between sequencing depth and number of samples with a sufficient portion of alternate phenotypes (302 unique Plant Introduction accessions at 15x average depth). We used this USB15x(302) data set to demonstrate the SP2CM strategy for stem termination.

Data curation for aggregation of resequenced data sets

To increase power of our Inverse GWAS and Accuracy calculations, we aggregated all publicly available soybean resequenced data sets into one curated panel of resequenced accessions. Since one of the data sets, Zhou11x(293) [6], was assembled in a different genome assembly version than the others, we remapped it from *Glycine max* Williams 82 a1.v1.1 (Wm82.a1.v1.1) to Wm82.a2.v1. The Zhou11x(293) data set raw read files were acquired from SRA as published in the paper [6]. Our PGen [27] workflow, which incorporates all genotyping required tools into one automated pipeline for SNP and InDel calling, was utilized to perform analysis on an XSEDE computation resource. The analysis was conducted using batch procedure (~50 accessions), and it has quality and filtration steps to filter out low quality reads. It utilizes the Burrows-Wheeler Aligner (BWA) [28] to align with the *G. max* Williams 82 (Wm82.a2.v1) reference genome and Genome Analysis Toolkit (GATK) [29] to do SNP and InDel calling. Separate Genomic Variant Call Format (GVCF) files from each batch were combined using the CombineGVCF argument in GATK and filtered to create the VCF file, using quality by depth (QD), Fisher strand values (FS), and mapping quality of variants (MQ) for SNPs and InDels. The FastQC reports, filtered SNP and Indel VCF files are available in SoyKB [28,30] via the NGS resequencing data browser. The Zhou dataset [6] which was remapped using the Wm82.a2.v1 genome assembly version (named Zhou302v2 in short) has 3.78% more positions compared to the Zhou dataset mapped using the Wm82.a1.v1.1 genome assembly version (named Zhou302v1 in short). Among the positions, the Zhou302v2 dataset also has>32 million SNPs while the Zhou302v1 dataset has only around 31 million SNPs.

Aggregation of resequenced data sets into the Soy775 accession panel

The Soy775 accession panel is an aggregation of all nonredundant resequenced soybean accessions. It is comprised of data sets from the USB-481 resequencing project [11] and Zhou302 data set [6] remapped to Wm82.a2.v1. VCF files for the USB15x(302), USB40x(42), Zhou11x(293), Soja15x(43) and MSMC17x(95) resequenced data sets were aggregated into a diverse set of 775 soybean accessions that we named the Soy775 accession panel (See Figure S2 for detailed composition of the Soy775 accession panel). This aggregation step was performed via a “fast merge” method, whereby any variant that was present in one dataset, but absent from another, was assigned as missing data in the data set in which it was absent. All 35,718,025 variants (SNP and InDel positions) from the combined Soy775 accession panel were run through the

SNPEff v4.3T software [31], using the Ensembl *Glycine max* v2.1.47 gtf annotation file, to obtain the predicted impact of each variant. We performed testing of the Soy775 data set and concluded that the error in respect to SoySNP50k marker positions was less than 4.7%; those marker positions were omitted from the curated accession panel due to high missing data values (above 40%). This aggregated file is publicly available for download on the SoyKB [28,30] server (https://soykb.org/public_data.php). The Soy775 data set represents the data set used for Inverse GWAS for the pod shattering trait, and it is used for the Accuracy calculation in the AccuTool (described further in the Methods).The final list of accessions is available at https://github.com/nad7wf/AccuTool/tree/master/publication_files.

Phenotypes for TM identification

Real (observed) phenotypes for soybean pod shattering, flower pigmentation and stem termination traits were used for GWAS that aimed to identify SoySNP50K TMs. For all accessions in this analysis, we used phenotypes that are publicly available in the GRIN database (https://www.ars-grin.gov) that are downloadable at Soybase (https://soybase.org/grindata/). For the pod shattering trait, we used the same categorization as in the recent work of Zhang and Singh [26] where late shattering score [32] was grouped into two phenotypes: shattering resistant (scale 1; n = 3,446) and shattering susceptible (scale 2-–5; n = 8,749). For GWAS on the flower pigmentation trait, we used white and purple flower color as phenotypes for the subset of USB15x(302) accessions with available SoySNP50K genotyping data (purple: n = 114; white: n = 166; unknown or other flower color: n = 18). The reason for using the subset of the USB15x(302) accessions for flower pigmentation trait instead of using all accessions with that available phenotype was to enable comparison between GWAS and Inverse GWAS on the same data set. For GWAS on the stem termination trait, we used 16,475 GRIN accessions that were either determinate (n = 8,771)) or indeterminate (n = 7,705). The phenotype files used in this work are available at https://github.com/nad7wf/AccuTool/tree/master/publication_files.

Synthetic phenotypes for SPAS

All Synthetic phenotypes used in this work were prepared as follows: A particular position of a variant or a marker on a chromosome in the soybean genome was extracted from SNP matrices of the original data sets (USB15x(302) [11]) or from SoySNP50K genotype data of our Soy775 accession panel that were downloaded at Soybase (https://soybase.org/snps/download.php). In the Synthetic phenotype files, the reference and alternate alleles were coded numerically as wild-type (WT, 1) or mutant (MUT, 2) phenotypes for each accession. The reference genotype Williams 82 was set as WT or MUT genotype (depending on the ancestral phenotype). Accessions with missing data for the position were coded as an unknown phenotype (NA).

For the Inverse GWAS on pod shattering trait, we derived the Synthetic phenotype from resequenced data of our Soy775 accession panel where we numerically coded the *pdh1* CM on chromosome 16 at position 29,944,393 *Glycine max* Williams 82.a2.v1 for all SoySNP50K genotyped accessions of the Soy775 accession panel as described in our earlier work [16].

For Inverse GWAS on flower pigmentation trait, we selected a variant position of the SNP associated with the deletion/substitution of the CM in the flavonoid 3′5′-hydroxylase (F3′5′H) *W1* gene [33] at chr13: 17,316,756 as the Synthetic phenotype. We performed the Inverse GWAS for USB15x(302) and Soy775 accession panels, respectively, using SoySNP50K genotyping data. We aimed to compare the landscape of the CM association between the USB15x(302) data set and the Soy775 accession panel. Besides the Inverse GWAS for flower pigmentation on SoySNP50K genotyping data, we also performed the Inverse GLAS on the resequenced genotype data of the Soy775 accession panel. This analysis aimed to reveal whether and how different the landscape of the *w1* CM associated SoySNP50K markers is in presence/absence of the other genomic variants (in the resequenced genotype of the Soy775 accession panel). For simplicity of the Inverse GWAS/GLAS associated SoySNP50K markers, these were extracted from the GLAS result file (using AccuTool described below) and displayed in a separate visual output.

To demonstrate how our SP2CM strategy can identify a CM from a TM, we performed GLAS of SP2CM – Part 2 for stem termination *Dt1* where we used the Synthetic phenotype of ss715635425 (chr19: 45,204,441) *Dt1/dt1* haplotype TM [25].

All Synthetic phenotypes used in this work are available at https://github.com/nad7wf/AccuTool/tree/master/publication_files.

GWAS

In this work, we performed genome association studies that were either genome-wide or genome-localized, but irrespective of the type of association, we conducted all the analyses employing a mixed linear model (MLM) [34] using the Genome Association and Prediction Integrated Tool (GAPIT) software implemented in R as previously described [35] and Trait Analysis by aSSociation, Evolution and Linkage (TASSEL) [36]. We used the default GAPIT setup for MLM except GWAS for pod shattering, where we performed the association with population structure correction for the first three principal components, in contrast to previously published Bayesian Information Criterion test [26]. For GLAS, we applied a limit-to-known-associated-one-chromosome-region approach where we performed the association on a 3 M bp wide region covering the *W1*-associated region for flower color and a 2 M bp wide region covering the *Dt1*-associated region for stem termination. All GWAS files generated within this work are available at https://github.com/nad7wf/AccuTool/tree/master/publication_files.

Accuracy calculation

Accuracy is an essential component of this work and therefore is calculated for every association study here. We implemented several comprehensive descriptors as additional post-GWAS selection criteria. The two key descriptors are Average accuracy and Combined accuracy pessimistic that were described previously [37].

For Accuracy calculations, correct association indicates an exact match between the genotype and phenotype case. Average accuracy (the balanced accuracy) combines WT accuracy (substitutes original Sensitivity or True Positive Rate) and MUT accuracy (substitutes original Specificity or True Negative Rate) into the following equation:

$$Average\ accuracy\ (\%) = \left(\frac{WT\ Accuracy + MUT\ Accuracy}{2}\right) \times 100$$

where WT accuracy is given by:

$$WT\ accuracy\ (\%) = \left(\frac{Number\ of\ accessions\ with\ correct\ WT\ association}{Number\ of\ accessions\ with\ correct\ and\ incorrect\ WT\ associations}\right) \times 100$$

and MUT accuracy is given by:

$$MUT\ accuracy\ (\%) = \left(\frac{Number\ of\ accessions\ with\ correct\ MUT\ association}{Number\ of\ accessions\ with\ correct\ and\ incorrect\ MUT\ associations}\right) \times 100$$

We derived an equation for calculation of combined accuracy (binary classification accuracy) from the concept of combined sensitivity and specificity [37] according to the standard equation:

$$Combined\ accuracy\ realistic\ (\%) = \left(\frac{Number\ of\ true\ positives + number\ of\ true\ negatives}{Sum\ of\ true\ positives + true\ negatives + false\ negatives + false\ positives}\right) \times 100$$

where true positives are accessions with MUT phenotype corresponding with MUT genotype, true negatives are accessions with WT phenotype corresponding with WT genotype, false positives are accessions with WT phenotype corresponding with MUT genotype and false negatives are accessions with MUT phenotype corresponding with WT genotype. Unlike Average accuracy, the Combined accuracy calculation produces imbalanced values based on the frequency of WT and MUT cases.

To better understand missing data values in Combined accuracy realistic, we calculate Combined accuracy pessimistic that represents a "worst-case scenario" that also incorporates associations where phenotype and/or genotype information is unknown. The worst-case scenario accuracy is therefore calculated based on a 3x3 contingency table (WT, MUT, unknown phenotype and/or genotype) where all unknown genotypes or phenotypes are considered as mismatches (false positives and false negatives) according to the following equation [37]:

$$Combined\ accuracy\ pessimistic\ (\%) = \left(\frac{Sum\ of\ accessions\ with\ correct\ WT\ association + correct\ MUT\ association}{total\ number\ of\ accessions}\right) \times 100$$

In this work, we use Average accuracy and Combined accuracy pessimistic (displayed in parentheses in all plots).

AccuTool construction

The AccuTool (https://soykb.org/AccuTool/index.php) is a web application written in R v3.5.1 using the R Shiny v1.3.2 package [38] and Perl v5.16.3. Specifically, R Shiny was used to create the interactive, front-facing website graphical user interface (GUI) and server communication, while Perl scripting was used to manipulate the underlying data and calculate Accuracy values. The AccuTool uses the soybean Williams 82 a2.v1 reference genome and the Soy775 accession panel to calculate accuracies for variants present in a user-defined genomic region. The tool enables upload of a phenotype file and a GWAS statistics file or to input a variant position as a Synthetic phenotype. The full description of the AccuTool with demo files is available at https://github.com/-nad7wf/AccuTool. The AccuTool functionalities that we used in this work include filtering by SoySNP50K marker, *p*-value, and sorting Accuracy by descent. To calculate Accuracy for a genomic position (either a CM or a TM, specified earlier) the position was selected in the tool as a phenotype as described in the AccuTool. The AccuTool input and output files are available at https://github.com/nad7wf/AccuTool/tree/master/publication_files.

AccuTool streamlined SP2CM for a novel gene identification

To identify a CM for the novel pod shatter gene *NST1A* with a streamlined SP2CM strategy, we performed an Accuracy analysis with the AccuTool using the genomic position of the TM ss715598106 [26] at chromosome 07 (chr07: 4,277,666) as Synthetic phenotype (tagging variant). We analyzed a 4 M bp landscape around the TM associated genomic region where we focused on modifying variants by selecting "Return only amino acid-modifying variants". The full AccuTool output for the landscape of ss715598106 is available at https://github.com/nad7wf/AccuTool/tree/master/publication_files.

Proxy markers

The AccuTool was used to identify the five highest Accuracy SoySNP50K markers for each of the three study cases as well as for the novel gene CM identification and other important cloned soybean genes. The CM position was determined from the report of gene cloning, and it was used as the Synthetic phenotype in Inverse GWAS, where the reference genotype Williams 82 was set as WT or MUT genotype (depending on the ancestral phenotype) with an arbitrary range of plus and minus 2 M bp with filtering for 70–100% Average accuracy and return of SoySNP50K positions only. The results were sorted by descending Average accuracy, and the first five markers were extracted along with the marker name and position, Average accuracy, and Combined accuracy pessimistic.

Data visualization

For GWAS data visualization, the R packages Sushi [39] and ggplot2 [40] were used.

Data availability

All data generated or analyzed during this study are included in this published article, its supplementary information files and at https://github.com/nad7wf/AccuTool/tree/master/publication_files. The Zhou11x(293) data set remapped to *Glycine max* Williams 82.a2.v1 as well as the aggregated Soy775 accession panel are publicly available at https://soykb.org/public_data.php.

The code for the AccuTool analysis as well as all the analyses outputs and supporting data used in this work are available at https://github.com/nad7wf/AccuTool.

Results

The Synthetic phenotype, Inverse GWAS, and Accuracy analysis concepts

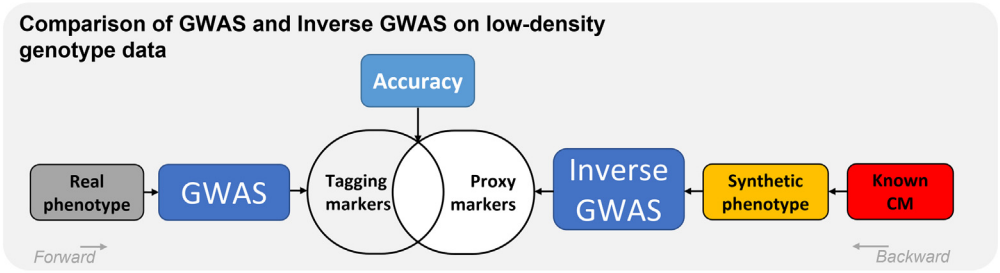
We developed an alternative association strategy to link phenotypes with genotypes and explore LD relationships independent of the frequencies of alleles. The Synthetic phenotype is a genomic position that can be converted into a phenotype by simple transposition of the base at the defined position into the reference or alternate case (WT or MUT phenotypes) (Fig. 1a). Every genomic variant position can be transformed into a Synthetic phenotype. For example, the soybean stem termination CM was readily transformed into a Synthetic phenotype (Fig. 1a).

The concept of Inverse GWAS uses a genomic position of a known CM (from a published cloned gene) as a Synthetic phenotype and enables deeper exploration of directly associated variant positions in the genome (the landscape of association) when applied to resequenced genotype data without the burden of mis-phenotyped artifacts or pleiotropic/epistatic effects of other genes. Fig. 1b illustrates the difference between standard (forward) GWAS where a real phenotype is used for the association, and our Inverse (backward) GWAS where a CM is used as the Synthetic phenotype. In contrast to standard GWAS, where one or more TMs in LD with the CM controlling the phenotype are identified, Inverse GWAS identifies Proxy markers – the most highly associated markers to a CM. Inverse GWAS outcomes are an LD landscape for corresponding positions and Proxy markers for the CM.

a

	Real phenotype (Observed)		Synthetic phenotype (CM position genotype)		
	Stem architecture	**Numerical**	**Position**	**Genotype**	**Numerical**
REF (Wm82)	Indeterminate	**1**	chr19:45,183,701	**T**	**1**
ALT	Determinate	**2**	chr19:45,183,701	**A**	**2**

b



c

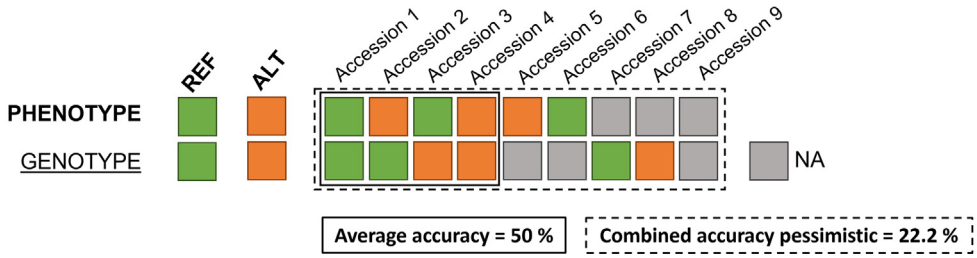


Fig. 1. Scheme that highlights key points of the Inverse GWAS approach: GWAS and Inverse GWAS, shows how a Synthetic phenotype can be utilized and created, and defines Accuracy. a, Representative table that shows the difference between real and Synthetic phenotype. Synthetic phenotype can be created by transforming the genotype of a variant position into a numerical binary phenotype (reference variant as 1 and alternate variant as 2). b, GWAS on low-density genotype data associates a real/observed phenotype with markers where those with the highest *p*-value are recognized as TMs. Inverse GWAS uses a known CM as a Synthetic phenotype and therefore highlights CM-associated markers – Proxy markers. Accuracy calculation enables identification of the most accurate TM based on direct correspondence to a phenotype. c, Accuracy calculation scheme. NA: not analyzed; unknown variable - phenotype or genotype.

We adopted an "Accuracy analysis" as post-GWAS or stand-alone mathematical evaluation criteria that assesses how well a genomic variant corresponds to a real or Synthetic phenotype. The two key Accuracy calculations are Average accuracy and Combined accuracy pessimistic (Fig. 1c). Average accuracy is a strict measure of correspondence that ignores missing data and frequencies (phenotype or genotype) and is defined as the mean percentage of the accessions with a match between the user-selected genomic position or phenotype and each of the resequenced data-derived variant positions in a selected range where the phenotype and genotype of the reference genome and alternate case are used for comparison (Fig. 1c). Combined accuracy pessimistic takes missing data into account and thus, enables comparison of various data set sizes and missing data percentages. Any accession with either missing genotype or missing phenotype data is considered a mismatch and therefore penalized in the Combined accuracy pessimistic calculations (Fig. 1c). When Accuracy is calculated for GWAS, the highest Accuracy markers among the associated markers can be selected (the Proxy markers of Inverse GWAS, the intersection in Fig. 1b). For soybean, we added power to our Accuracy calculations by aggregating a curated panel of 775 resequenced soybean accessions (the Soy775 accession panel, Fig. S1a, b). We implemented the Accuracy analysis to every GWAS in this work by coupling GWAS outputs with our automated Accuracy calculation tool, the AccuTool (described below).

Coalescing these new concepts, tools, and resources, we devised the SP2CM strategy that aims to identify CM (Fig. 2). SP2CM consists of two parts that each employ an association step and an Accuracy analysis. Part 1 associates a real phenotype with low-density genotype data of a panel of accessions in GWAS with all chromosomes, whereas Part 2 associates a Synthetic phenotype with localized genotype data of a resequenced data set (GLAS), further delineating the associated region as part of a broader strategy. Both parts include the Accuracy calculation step. In Part 1, Accuracy enables selection of the highest correspondence marker among the associated markers (TM). In Part 2, Accuracy identifies candidate genes and putative CMs.

To enable automated Accuracy analysis, we created the AccuTool (https://soykb.org/AccuTool/index.php) to calculate and explore the mathematical correspondence (Accuracy) between resequenced data-derived variant positions in the genome and user-defined positions or phenotypes. AccuTool results enable exploration of the LD landscape in a focused region, identification of Proxy markers for CMs, and accentuation of GWAS visuals of the three demonstrative example cases and one example of the novel gene. For soybean, the AccuTool calculates Accuracy against a single genomic position or phenotype in a selected genomic range for each position in the resequenced data curated from our analysis of 775 soybean accessions. The Soy775 accession panel currently consists of 110 wild (*G. soja*) soybean accessions, 475

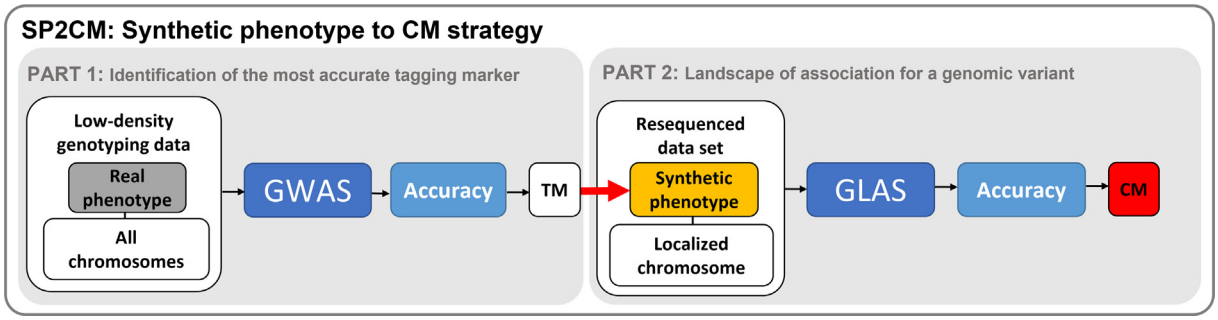


Fig. 2. The Synthetic phenotype to CM (SP2CM) strategy. The pipeline illustrates the SP2CM process that consists of two parts: Part 1 and Part 2 (grey boxes). In Part 1, GWAS is performed on low-density genotype data of accessions with a known real phenotype. GWAS-identified associated markers are then tested for Accuracy where the highest Accuracy/-$\log_{10}p$-value marker is the TM. Part 2 starts with transformation of the genotype of the TM variant position present in a resequenced data set into a Synthetic phenotype (yellow box). In Part 2, GLAS identifies the TM-associated genomic variants, where those with the highest Accuracy are candidates for CM (red box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

soybean (*G. max*) landraces, and 190 improved (*G. max*) soybean accessions, with ~ 35.7 M variant positions derived from independent resequencing projects that encompasses 5 individual data sets (Fig. S2a, b) [6,7].

Inverse GWAS illuminates multiple near-perfect TM for a pod shattering CM

In soybean, seed dispersal susceptibility is an ancestral trait, whereas pod shattering resistance is caused (among others) by disfunction of a dirigent-like protein *Pdh1* that otherwise controls pod wall torsion after dehiscence [41]. Quantitative traits present challenges that make it difficult to identify the associated genes in part because each associated region controls an unknown portion of the phenotype. The non-functional *pdh1* allele contributing to pod shatter resistance is caused by a nonsense mutation on chromosome 16 at position 29,944,393 *G. max* Williams 82.a2.v1 that leads to a premature stop codon in the *Pdh1* protein [41]. Besides the SoySNP50K low-density marker ss715624199 Proxy that we described previously from our Inverse GWAS analysis [16], the ss715624201 marker was recently also associated with *Pdh1* in GWAS on an independent panel of soybean accessions [26]. Since both studies were performed on a limited number of accessions in the data sets (~500 and ~ 800), we maximized the power of the GWAS by including all accessions with available pod shattering phenotypes (n = 12,195). Genome-wide evidence for association to the pod shattering phenotype is documented in Fig. 3a where the ss715624199 marker was identified as an isolated TM with the highest -$\log_{10}(p)$. To compare the landscape of association for the real pod shattering score phenotype and for the *pdh1* CM genomic variant as a Synthetic phenotype, we conducted an Inverse GWAS analysis using the *pdh1* CM genotype from the Soy775 accession panel as a Synthetic phenotype on SoySNP50k genotype data. We identified three Proxy markers to the *pdh1* CM with nearly identical maximized -$\log_{10}(p)$ values in the landscape of other associated markers (Fig. 3b). To understand why the Zhang & Singh [26] TM for *Pdh1* was different from our *Pdh1* Proxy marker identified in GWAS here and in our previous study [16], we zoomed into the Inverse GWAS *pdh1* CM-associated region (Fig. 3c) and accentuated the plot with Accuracy values to the *pdh1* CM for every SoySNP50K marker in the region calculated by the AccuTool (Fig. 3d). Based on the accentuated Accuracy information, we determined that in the SoySNP50K marker case of *Pdh1*, there were several markers that exposed the CM with very high Accuracy (five markers with > 95% Accuracy to the CM), along with a few SNPs in close physical proximity to the CM that had lower -$\log_{10}p$-values as well as low Accuracy. To assess direct correspondence between a real phenotype and a CM (or any other genomic position), the AccuTool enables upload of a user-defined phenotype for the Soy775 accession panel. Here, to ascertain the direct correspondence between the *pdh1* CM and the pod shattering score, we used this AccuTool option for the quantitative pod shatter trait and revealed that even though the most highly associated markers with the *pdh1* CM have Accuracy values close to 100%, the correspondence with the observed pod shattering score phenotype is only 87.5%, reflecting the power of determining the landscape of association in an isolated genomic region, which can be very meaningful in dissecting multi-genic traits. Although this pod shatter result revealed multiple near-perfect Proxy markers for the *pdh1* CM, that result is not typical for GWAS on low-density genotype data (data not shown).

Accuracy analysis can serve as a component for prioritizing tagging variants

When Inverse GWAS is performed on resequenced data, it reveals the whole landscape of association for the CM where the low-density markers are interspersed with the other genomic variants present in the resequenced data set. Thus, Inverse GWAS with a resequenced data set provides a more accurate view of the low-density markers within the context of associated genomic variants.

Soybean flower pigmentation is a qualitative trait that is controlled by the *W1* locus [42]. The *W1* allele is a gene for flavonoid 3′5′-hydroxylase (F3′5′H) that is essential for completing the biosynthetic pathway of anthocyanins resulting in purple flower color in wild soybeans (*Glycine soja* [Siebold & Zucc.]). The non-functional *w1* allele is caused by a small insertion and substitution leading to a premature stop codon in the F3′5′H gene (*Glyma.13G072100*) and white flowers [33]. In our curated Soy775 accession panel resequenced data, the complicated genomic rearrangement resulted in five *w1* CM InDel variant positions on chromosome 13 between 17,316,723 and 17,316,758 as well as a SNP that mapped to position 17,316,756 in the rearrangement and was in nearly perfect association with flower color phenotype (99.4% Accuracy, See Data availability) and the other functional InDels (Table S1). We utilized the SNP position as the *w1* CM in this work due to the complexity of the InDel *w1* positions in the resequence data. There were five other modifying variants in the transcript region of *Glyma.13G072100* with low Accuracy to flower pigmentation (Table S1).

For this analysis, we utilized the USB15x(302) resequenced data set containing 302 soybean accessions at 15x average depth to illustrate the relative effectiveness of smaller data sets with high-density information [11]. The GWAS for flower color phenotype with SoySNP50K marker genotype on the USB15x(302) accessions

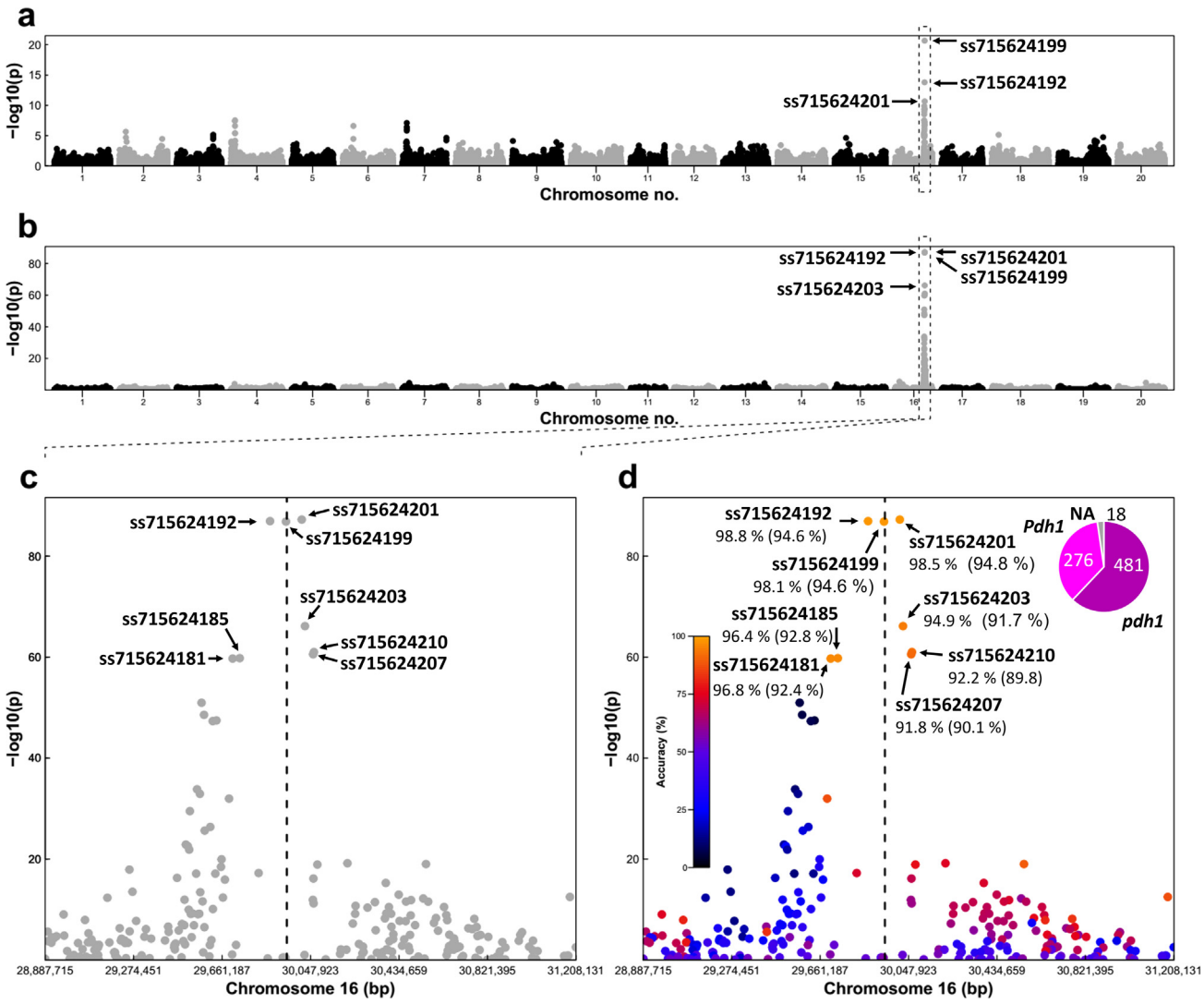


Fig. 3. GWAS and Inverse GWAS for the pod shattering phenotype demonstrating example of multiple near-perfect TMs for the pod shattering CM.Manhattan plot depicting the evidence of association (-$\log_{10}p$-value) across soybean GRIN accessions genotyped by low-density SoySNP50K bead chip for pod shattering score phenotype in GWAS (a) and an Inverse GWAS using the *pdh1* CM as Synthetic phenotype with the SoySNP50K marker subset from the AccuTool Soy775 accession panel (b). c, Zoomed Inverse GWAS *Pdh1*-CM associated 2.32 M bp region of (b). d, Accentuation of Inverse GWAS results in (c) using the AccuTool to generate Accuracy values for the markers against the *Pdh1*/*pdh1* CM allele status (color scale bar). The most highly associated markers are labeled with their corresponding ss code and Accuracy values, with the value in parentheses indicating the Combined accuracy pessimistic output that accounts for allele frequency and penalizes missing data. Dashed line intercept on x axis indicates the *pdh1* CM position at chr16: 29,944,393 Wm82.a2.v1. The insert pie chart indicates the count of *Pdh1*/*pdh1* CM alleles in the Soy775 accession panel.

resulted in two highly associated TMs, ss715616657 (reported by Bandillo et al. [25]) and ss715616658 located 6,787 or 9,493 bp upstream of the *w1* SNP CM, respectively (Fig. 4a). Fig. 4b-c show the distribution of real flower color phenotype and *w1*/*W1* allele counts from the Soy775 accession panel that were used for Accuracy accentuation. When zoomed in on the GWAS associated region on chromosome 13 and accentuating with Accuracy values to the *w1* CM from the Soy775 accession panel AccuTool data, there are several associated variant positions in close proximity to the *w1* CM, but the two SoySNP50K markers that are within 10 kb of the *w1* CM emerge as the most highly associated TMs and those with the highest Accuracy to the *w1* CM (Fig. 4d). It was notable that ss715616654 was the closest SoySNP50K marker in proximity to the *w1* CM (325 bp). Despite statistically significant association to the phenotype using the USB15x(302) data set (-log10(p) was 11.96), ss715616654 had low Accuracy to the *w1* CM (Fig. 4d). ss715616654 is the only SoySNP50K marker in *Glyma.13G072100* coding sequence, a silent mutation in close vicinity to the functional InDels (Table S1). We performed the Inverse GWAS also on the Soy775 accession panel genotyped with SoySNP50K chip to maximize the GWAS power by doubling the number of accessions in the data set and observed a very similar pattern (Fig. 4e). We further investigated the surrounding landscape of association by performing Inverse GLAS on the USB15x(302) data set. Surprisingly, when Inverse GLAS was conducted using the *w1* CM as a Synthetic phenotype with USB15x(302)-derived resequencing variants, ss715616654 was the highest associated SoySNP50K marker (Fig. 4f). Upon dissection of the AccuTool Soy775 Accuracy calculations (filtering for SoySNP50K marker genotypes only, Fig. 4g), the ss715616654 marker at position chr13: 17,316,431 had near perfect correspondence with the *w1* mutant allele, but low Accuracy for the *W1* functional allele (Table S1). For the USB15x(302) data set, the ss715616654 marker would have been identified as a Proxy marker based on Inverse GWAS. However, accentuated Accuracy information for the variant positions from the Soy775 data set clearly showed the best Proxy markers to the *w1* CM identified with the AccuTool, regardless of their physical distance to the *w1* functional InDels (Fig. 4f-g).

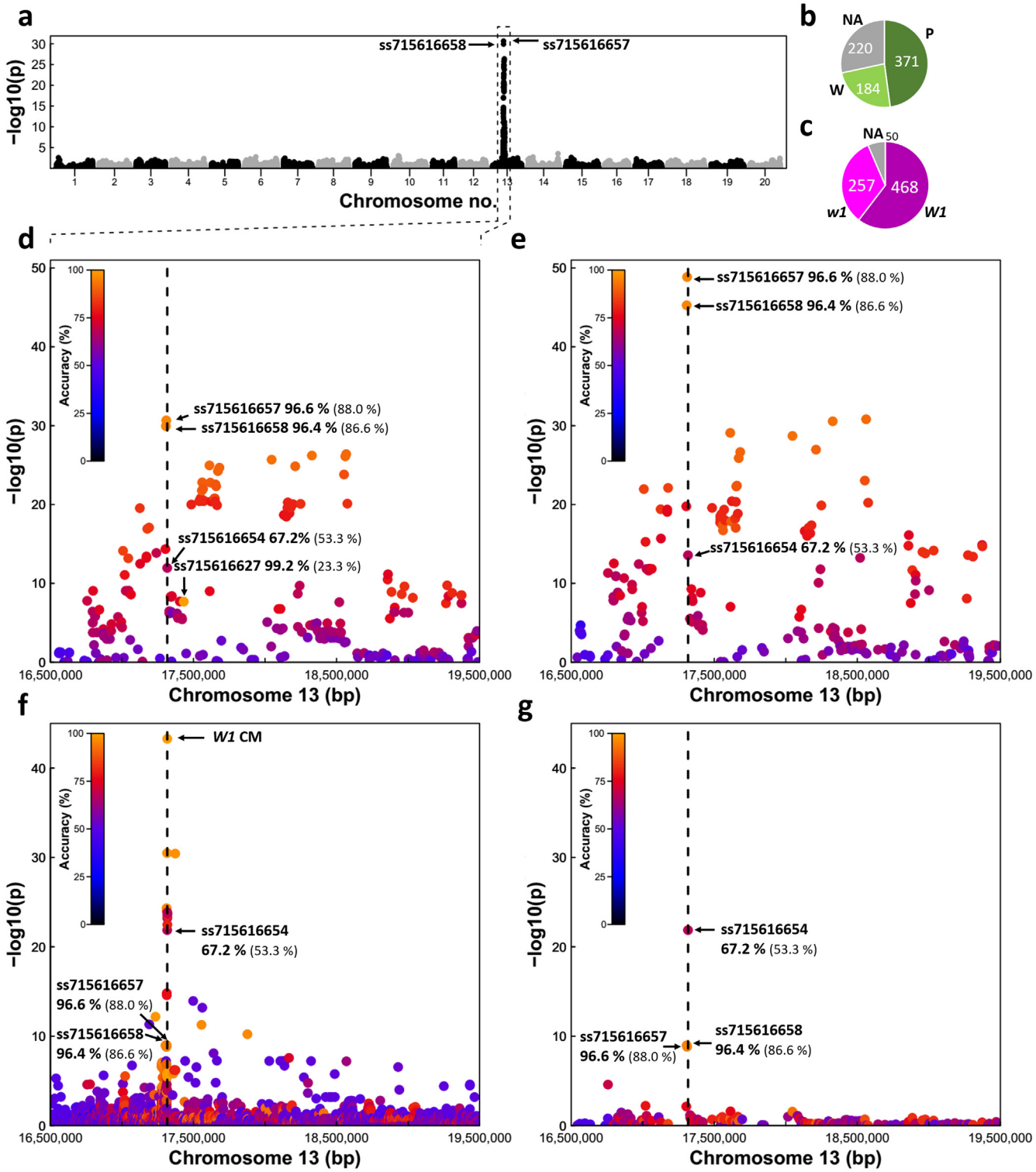


Fig. 4. GWAS, Inverse GWAS and GLAS for *w1* flower color CM shows that Accuracy can serve as a component for prioritizing TMs.a, Manhattan plot from GWAS depicting the evidence for association (-$\log_{10}p$-value) across the soybean USB15x(302) data set accessions for white or purple flower color phenotype genotyped by low-density SoySNP50K bead chip. b-c, pie charts representing distribution of white (W), purple (P), and unknown (NA) flower color phenotypes (b) and *w1*/*W1* allele counts (c) of Soy775 accession panel that were used for Accuracy accentuation in the following plots. d, Zoomed *W1*-associated region of (a) on chromosome 13 (3 M bp range) where the color coding represents the AccuTool Soy775 accession panel outputs for Accuracy using the *w1*/*W1* allele status on the associated SoySNP50K markers. Selected markers are labeled with their identifier and Accuracy values, with the value in parentheses indicating the Combined accuracy pessimistic output that accounts for allele frequency and penalizes missing data. Dashed line intercept on x axis indicates the *w1* CM position at chr13: 17,316,756 Wm82.a2.v1. e, Soy775 accession panel data set Inverse GWAS results with the *w1* CM Synthetic phenotype and SoySNP50K genotypes zoomed to *W1*-associated region on chromosome 13 (3 M bp range) where the color coding represents the AccuTool Soy775 accession panel outputs for Accuracy using the *w1*/*W1* allele status on the associated SoySNP50K markers. f, USB15x(302) data set Inverse GLAS results with the *w1* CM Synthetic phenotype across USB15x(302) all resequenced data positions as genotype zoomed on chromosome 13 (3 M bp range) where the color coding represents the AccuTool Soy775 accession panel outputs for Accuracy using the *w1* CM on the associated SoySNP50K markers. g, same as f showing just the SoySNP50K markers. Dashed line intercept on × axis of d-g highlights *w1* CM position at chr13: 17,316,756 Wm82.a2.v1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Marker ss715616627 was an outlier for high Accuracy to the *w1* CM with lower statistical significance for association than many other markers in the region (Fig. 4d). Further investigation revealed a quality issue with the ss715616627 marker genotype such that the accessions had very high missing genotype information in the data set for that marker position, and missing data was reflected in the AccuTool Combined accuracy pessimistic value of 23.3%.

The SP2CM strategy discriminates multiple alleles of a candidate gene for stem termination

In this example, we present our application of the SP2CM strategy on the indeterminate or determinate stem termination trait conditioned by the *Dt1* locus. The *Dt1* gene (*Glyma.19g194300*) encodes *GmTFL1b*, a positive regulator of the shoot apical meristem [43], and the missense R166W allele was responsible for the determinate plant (*dt1*) type [43,44]; three additional missense alleles (R62S, P113L, and R130K) were identified in the gene [44]. The functional and missense alleles of *Dt1*/*dt1 Glyma.19g194300* are present in the Soy775 accession panel, with the reference and R166W alleles being the most frequent.

The ultimate prerequisite for amplification of GWAS power on genotype data is a large number of accessions with a known real phenotype. For GWAS for the SP2CM Part 1 (Fig. 2), we used stem termination phenotypes and SoySNP50K marker genotypes on the GRIN collection of accessions (n = 16,475). This GWAS produced a relatively isolated highly associated TM, ss715635425 (Fig. 5a). To zoom into the associated region and to accentuate the variants with Accuracy, we used stem termination phenotype and *dt1*/*Dt1* allele status: Fig. 5b-c show their distribution in the Soy775 accession panel. When zoomed in on the associated region and accentuating with the AccuTool Accuracy values using the real/observed phenotypes (433 available from 775 accessions), the TM with high statistical significance had only modest Accuracy values (79.0% for Accuracy and 45.3% for Combined accuracy pessimistic; Fig. 5d). Replacing the observed phenotype Accuracy values with AccuTool calculations utilizing the reported *dt1* R166W CM (chr19:45,183,701) as a Synthetic phenotype revealed the relationship between the TM and the most frequent *dt1* CM for the stem termination phenotype (Fig. 5e). The stem termination TM is a Proxy marker for the *dt1* R166W CM, but it is not a Proxy marker for the other three *dt1* missense alleles (data not shown). For GLAS of the SP2CM Part 2 we used USB15x(302) resequenced accessions with the ss715635425 TM as the Synthetic phenotype and all USB15x(302) resequencing variants in the 2 M bp surrounding region as genotype. The GLAS produced a cluster of associated variant positions that included the *dt1* R166W CM (Fig. 5f). Further zooming into the ss715635425-associated region and accentuating with AccuTool Accuracy values using the *dt1* R166W CM as phenotype revealed the variant positions in LD with the *dt1* TM that have high accuracies to the CM (Fig. 5g, AccuTool outputs are available at https://github.com/nad7wf/AccuTool/tree/master/publication_files). High Accuracy values extended throughout most the 250 kb-associated region for a subset of the variant positions, despite fluctuation in -log10(p) values (Fig. 5g).

The SP2CM strategy assists in identification of a CM for an uncloned gene contributing to the pod shattering phenotype

The SP2CM strategy utilizes accuracy to identify LD between variant positions associated with a phenotype and serves as a bridge between low-density and resequence data sets. In our GWAS with the quantitative pod shatter phenotype (Fig. 3a) we identified the previously published low-density ss715598106 as the *NST1A* locus TM [26]. The previous study used the TM as a Proxy marker for pod shatter in their analysis, because they identified a modest association for four of 32 additional variant positions focused on the candidate *NST1A* gene (*Glyma.07g050600*) approximately 55,000 bp downstream from the TM [26]. Therefore, we followed a streamlined SP2CM strategy, where we directly analyzed the landscape of the ss715598106 associated genomic region for the Soy775 accession panel. We selected all modifying variants in the region and analyzed the positions with the highest Average Accuracy to the ss715598106 synthetic phenotype position. A position with the third highest Accuracy was a SNP (chr07: 4,332,840) that creates a stop-lost mutation for the alternate allele in *Glyma.07g050600,* a NAC Secondary wall Thickening Promoting Factor 1 ortholog (*NST1A*) (Table 1). The high-accuracy modifying variants in other surrounding genes were in LD with the candidate CM position and the TM, while no other modifying variants in *NST1A,* including the variants identified in previous work [26] had high Accuracy (Table 2; data not shown). The direct correspondence between the TM and the stop-lost mutation in *NST1A* implicates that position as the pod shattering CM.

Using the AccuTool to generate a Proxy marker resource for cloned genes

The SP2CM strategy exploits the overlap in TMs and Proxy markers (Fig. 1b). Our concepts, tools, and resources assist in connecting phenotypes to genotypes, and an additional application of the AccuTool is to generate viable Proxy markers for cloned genes. A Proxy marker resource can therefore be used as a first check for new GWAS results to more confidently assign identified loci to known genes. For soybean, we used the AccuTool to extract the five highest Accuracy low density Proxy markers for the CMs for both pod shatter loci, the flower color pigmentation gene, the stem termination gene, and the other cloned soybean genes for which a CM was available (Table 3 [26,33,50–58,41,43–49]). The top five Proxy markers ranged from 73.8 to 98.8% Accuracy. One notable feature of Proxy markers is that the absolute physical distance to the CM was variable.

Discussion

Here we demonstrated that Inverse GWAS with a CM as a Synthetic phenotype associates variants that are in LD with the CM, and, that Accuracy analysis can illuminate variants with the highest direct correspondence to the phenotype among them. Thus, correlation of SPAS-associated genomic positions with direct correspondence to real or Synthetic phenotypes and *vice versa* led us to the following conclusion: The Synthetic phenotype empowers the researcher to employ this connection of phenotype to genotype association; a highly accurate low-density marker position associated to a real phenotype can be used as a Synthetic phenotype in GWAS with a resequenced data set where it subsidizes otherwise missing phenotype information. We further showed that the concept of Synthetic phenotype and Accuracy calculation is a key to further success in genotype-based GWAS discoveries if a panel of resequenced accessions is available. We demonstrated the utility of both a 775-member accession panel and a 302-member accession panel of resequenced data sets. The benefit of this approach is in creation of critical information that can further assist in identifying CM.

Each of our three demonstrative cases served a particular purpose and represent the most common GWAS complicating factors that can be solved by our proposed concepts. These factors are: GWAS analyses on different sets of limited-number accessions identify different TMs; CMs can be positionally distant to TMs and thus, evaluation based on LD can be insufficient or even mis-

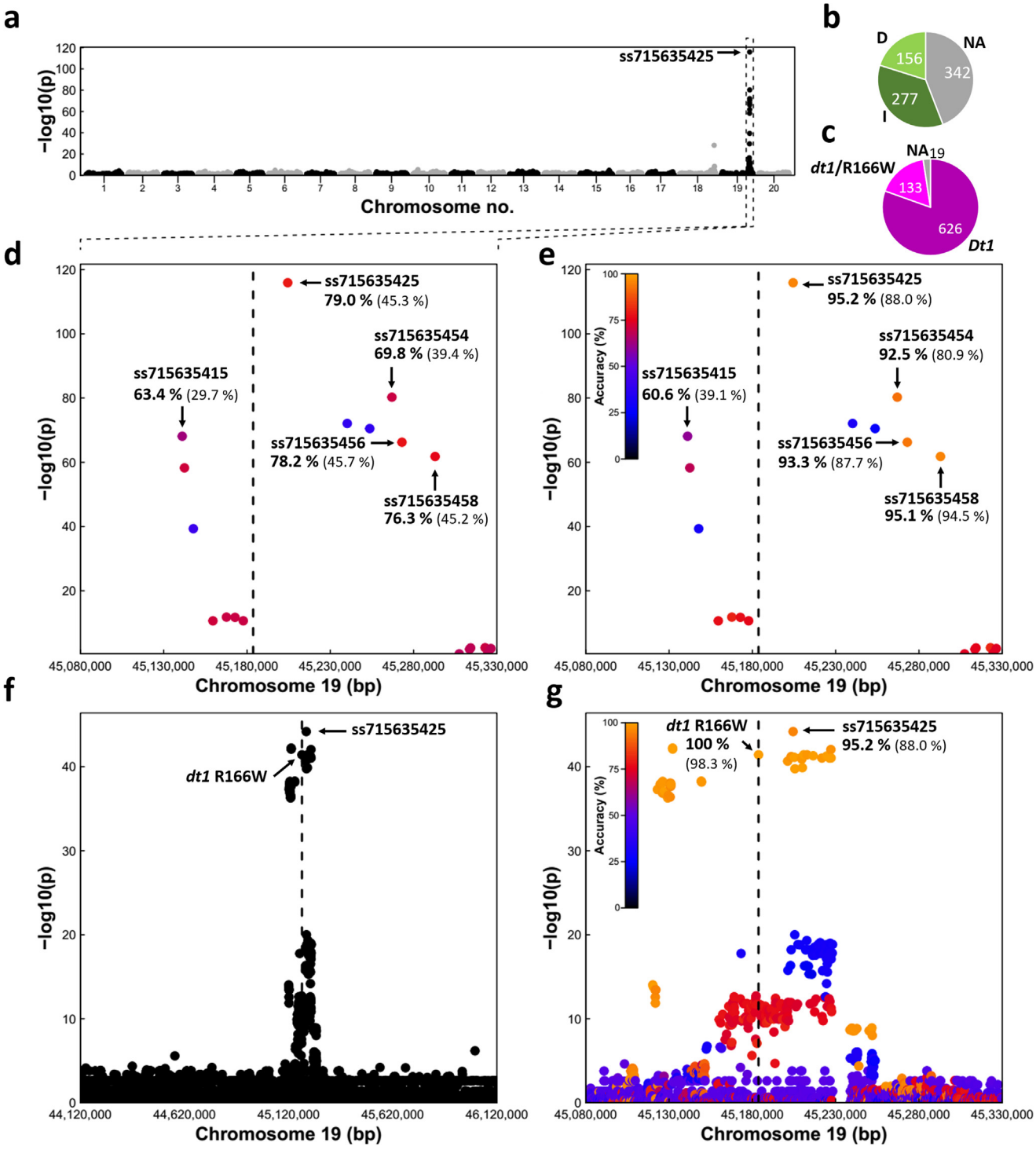


Fig. 5. Identification of the *dt1* R166W determinate stem termination CM through the SP2CM strategy, an example of multiple alleles of a candidate gene. a, SP2CM Part 1: Manhattan plot depicting the evidence for association (-$\log_{10}$*p*-value) across the soybean GRIN collection of accessions genotyped by low-density SoySNP50K markers for determinate (D) or indeterminate (I) stem termination phenotype. b-c, pie charts representing distribution of determinate (D), indeterminate (I), and unknown (NA) stem termination phenotypes (b) and *dt1*/*Dt1* allele counts (c) of Soy775 accession panel that were used for Accuracy accentuation in the following plots. d, Zoomed GWAS of SP2CM Part 1 *Dt1*-associated 250 kb region on chromosome 19 with accentuated Accuracy to the stem termination phenotype (based on the AccuTool Soy775 panel with distribution of the phenotypes described in the pie chart inset). Dashed line intercept on x axis highlights *dt1*/R166W CM position at chr19: 45,183,701 Wm82.a2.v1. e, The same as d except with the Accuracy calculated to the highest frequency determinate allele *dt1* R166W/*Dt1* Synthetic phenotype (pie chart in the inset documents the distribution of the Synthetic phenotype in the AccuTool Soy775 panel). Section f depicts GLAS results of SP2CM Part 2 on USB15x(302) resequenced data set with the *Dt1* haplotype TM ss715635425 as a Synthetic phenotype (2 M bp). g zooms into the 250 kb *Dt1* associated region from f and shows the landscape of the associated variants with their AccuTool-generated Accuracy to the *Dt1/dt1(R166W)* allele status (distribution of the Synthetic phenotype in the AccuTool Soy775 panel is documented in the inset pie chart). The values in parentheses document uncertainty expressed as Combined accuracy pessimistic that accounts for allele frequency and penalizes missing data.

Table 1
AccuTool output for selected modifying variant positions of *NST1A* associated locus. Accuracy was calculated to the ss715598106 marker position at chr07: 4,277,666 as Synthetic phenotype with reference allele as WT (n = 476) and alternate as MUT (n = 285). Only 1.8% of Soy775 accessions were with missing genotype. Only positions with Average accuracy ≥ 90.0% are shown and sorted descendent by Combined accuracy pessimistic. Candidate gene *NST1A* CM is in bold. (Distance to TM is calulated in base pairs with upstream locations in paretheses; Avg_Accu, Average accuracy expressed in percentage; Comb_Acc_Pess, Combined accuracy pessimistic expressed in percentage; Effect, Effect on amino acid change; WT_Accu, Accuracy of accessions with WT allele expressed in percentage; MUT_Accu, Accuracy of accessions with MUT alle expressed in percentage; Miss_genot_WT, missing genotype of WT accessions expressed in percentage; Miss_genot_MUT, missing genotype of MUT accessions expressed in percentage).

Position at chr07	Distance to ss715598106	Avg_Accu	Comb_ Accu_Pess	SoySNP50K	Gene	Effect	WT_Accu	Miss_ Genot_WT	MUT_ Accu	Miss_ Genot_MUT
4,239,995	37,671	97.0	93.2	.	*Glyma.07g049800*	G>A|G563E	95.1	1.3	98.9	2.5
4,232,393	45,273	97.1	92.9	.	*Glyma.07g049700*	A>C|I191S	94.6	1.9	99.6	2.1
4,332,840	**(55,174)**	**96.1**	**92.6**	**.**	***Glyma.07g050600***	**T>A|stoplost 401R**	**95.1**	**1.5**	**97.1**	**1.8**
4,231,740	45,926	97.2	92.4	.	*Glyma.07g049700*	T>C|T409A	94.7	1.5	99.6	4.2
4,297,714	(20,048)	94.5	92.4	.	*Glyma.07g050500*	G>A|A51V	98.7	0.6	90.2	3.2
4,240,492	37,174	94.7	91.7	.	*Glyma.07g049800*	G>C|V702L	95.1	0.6	94.2	2.8
4,291,502	(13,836)	94.9	91.6	.	*Glyma.07g050400*	C>A|L238I	99.4	1.5	90.4	5.3
4,238,298	39,368	97.2	91.0	.	*Glyma.07g049800*	T>C|L41P	95.4	3.2	98.9	6.0
4,240,329	37,337	93.4	90.2	.	*Glyma.07g049800*	C>A|N647K	96.8	1.9	90.1	3.9
4,238,391	39,275	97.0	89.9	.	*Glyma.07g049800*	T>A|L72Q	95.6	3.8	98.5	7.7
4,238,327	39,339	94.7	89.4	.	*Glyma.07g049800*	T>C|Y51H	95.7	2.5	93.6	6.7
4,232,510	45,156	91.4	89.4	ss715598067	*Glyma.07g049700*	C>A|G152V	96.4	1.5	86.4	2.1
4,224,511	53,155	91.2	89.4	.	*Glyma.07g049700*	T>A|T1080S	96.2	0.6	86.2	3.2
4,220,519	57,147	92.3	89.3	.	*Glyma.07g049700*	A>T|S1883N	94.5	1.5	90.2	3.2
4,224,562	53,104	91.3	89.0	.	*Glyma.07g049700*	A>T|S1063T	96.4	1.3	86.2	3.5
4,297,080	(19,414)	92.5	88.4	.	*Glyma.07g050500*	C>A|K262N	99.2	0.8	85.8	11.2
4,238,249	39,417	90.1	87.0	.	*Glyma.07g049800*	C>A|L25I	97.4	2.9	82.7	4.9
4,238,751	38,915	95.8	82.7	.	*Glyma.07g049800*	T>C|V192A	95.9	2.3	95.6	28.4
4,296,169	(18,503)	94.2	78.2	.	*Glyma.07g050500*	T>G|N375H	98.7	21	89.7	8.1

Table 2
AccuTool output for all modifying variant positions in *NST1A* gene. Accuracy was calculated to the ss715598106 marker position at chr07: 4,277,666 as a Synthetic phenotype with the reference allele as WT (n = 476) and alternate as MUT (n = 285). Only 1.8% of Soy775 accessions were with missing genotype. The candidate CM of *NST1A* is in bold. (Distance to TM is calulated in base pairs with upstream locations in paretheses; Avg_Accu, Average accuracy expressed in percentage; Comb_Acc_Pess, Combined accuracy pessimistic expressed in percentage; Effect, Effect on amino acid change; WT_Accu, Accuracy of accessions with WT allele expressed in percentage; MUT_Accu, Accuracy of accessions with MUT allele expressed in percentage; Miss_genot_WT, missing genotype of WT accessions expressed in percentage; Miss_genot_MUT, missing genotype of MUT accessions expressed in percentage).

Position at chr07	Distance to ss715598106	Avg_Accu	Comb_ Accu_Pess	Gene	Effect	WT_Accu	Miss_ Genot_WT	MUT_Accu	Miss_ Genot_MUT
4,332,204	(54,538)	49.9	36.9	*Glyma.07g050600*	G>A|G189R	99.7	39.7	0	33
4,332,246	(54,580)	49.9	36.9	*Glyma.07g050600*	G>A|G203R	99.7	39.7	0	33
4,332,444	(54,778)	50.4	61.5	*Glyma.07g050600*	C>G|H269D	100	0.2	0.7	0
4,332,451	(54,785)	50.2	37.2	*Glyma.07g050600*	C>CCAA|inframe insertion 275dup	100	39.7	0.5	33
4,332,604	(54,938)	56.6	65.3	*Glyma.07g050600*	CCAACAA>CCAACAA| inframe insertion 328dup	99.4	1.1	13.8	3.2
4,332,604	(54,938)	56.6	65.3	*Glyma.07g050600*	CCAA>C|inframe deletion 328del	99.4	1.1	13.8	3.2
4,332,612	(54,946)	50.2	37.2	*Glyma.07g050600*	A>G|N325D	100	39.7	0.5	33
4,332,763	(55,097)	50.2	37.2	*Glyma.07g050600*	T>C|M375T	100	39.7	0.5	33
4,332,784	(55,118)	50.2	37.2	*Glyma.07g050600*	A>G|H382R	100	39.7	0.5	33
4,332,797	(55,131)	49.8	61.2	*Glyma.07g050600*	A>T|Q386H	99.6	0.0	0	0.4
4,332,840	**(55,174)**	**96.1**	**92.6**	***Glyma.07g050600***	**T>A|stop lost 401R**	**95.1**	**1.5**	**97.1**	**1.8**

leading [59]; evolutionary unrelated multiple CMs in one gene/locus are tagged by a marker corresponding to the most frequent CM in the context of its specific haplotype. The case study on the pod shatter *Pdh1* gene has shown that Accuracy can be used for TM evaluation even in cases of multiple near-perfect TMs and demonstrated that Accuracy can be used to compare TMs gained from different studies. Accuracy can reveal an underpowered GWAS, or Accuracy analysis can serve as a measure of GWAS analyses performance. In the current available literature, there was no alternative that could be used to describe GWAS performance other than *p*-value. Our example on flower pigmentation provides insight as to why physical distance or genomic vicinity to a TM might be misleading in GWAS-driven discoveries and, we extend this knowledge by listing Proxy markers for cloned genes of major soybean traits (Table 3.) The case study on stem termination has shown that the SP2CM strategy can be utilized in more complicated cases such as multiple alleles in the causal gene. Also, results of this example show that even though the TM is of relatively lower Accuracy to the real phenotype, that TM can be still utilized in our SP2CM strategy and successfully identify the most prevalent of the multiple allelic CMs. We performed a streamlined SP2CM strategy to identify the CM of the novel pod shatter gene *NST1A*. By this example we aimed to encourage other soybean researchers that SP2CM can be performed in a streamlined way where the AccuTool can calculate Accuracy for the aggregated Soy775 accession panel for every genomic variant position as the Synthetic phenotype. In this example, we showed that we were able to evaluate the previously proposed list of CMs and select the most highly accurate CM within

Table 3
Proxy marker analysis for a selection of important soybean genes. The top five Proxy SoySNP50K markers were identified by Inverse GWAS Accuracy from the AccuTool Soy775 accession panel. Proxy markers of the five highest Accuracy values to the CM using the AccuTool are shown in the table. (Distance to CM, calulated in base pairs with upstream locations in paretheses; Avg_Accu, Average accuracy expressed in percentage; Comb_Acc_Pess, Combined accuracy pessimistic expressed in percentage; NA, Pdh1 gene that is annotated as *Glyma16g25580* in Wm82.a1.v1 of the Williams 82 soybean genome reference sequence, but it is not annotated in Wm82.a2.v1; Proxy markers in bold indicate concordance with the TM identified by Bandillo and co-authors [25]).

Trait	REF/ALT allele	Williams 82.a2.v1	Chromo-some	CM position	Reference	SoySNP50K Position	Proxy marker	Distance to CM	Avg_ Accu	Comb_ Acc_Pess
Pod shatter	*pdh1/Pdh1*	NA	16	29,944,393	[41]	29,870,849	ss715624192	(73,544)	98.8	94.6
						30,009,486	ss715624201	65,093	98.5	94.8
						29,940,504	ss715624199	(3,889)	98.1	94.6
						29,707,327	ss715624181	(237,066)	96.8	92.4
						29,738,349	ss715624185	(206,044)	96.4	92.8
Flower color	*w1 SNP/W1*	*Glyma.13g072100*	13	17,316,756	[33]	17,309,969	**ss715616657**	(6,787)	96.6	88.0
						17,307,263	ss715616658	(9,493)	96.4	86.6
						18,046,553	ss715616090	729,797	92.0	82.3
						18,327,972	ss715615785	1,011,216	92.0	84.1
						18,567,932	ss715615487	1,251,176	91.8	83.1
Stem termination	*Dt1/dt1* R166W	*Glyma.19g194300*	19	45,183,701	[43,44]	45,204,441	**ss715635425**	20,740	95.2	88.0
						45,292,930	ss715635458	109,229	95.1	94.5
						44,937,972	ss715635400	(245,729)	94.4	88.5
						45,273,019	ss715635456	89,318	93.3	87.7
						45,266,984	ss715635454	83,283	92.5	80.9
Pod shatter	*nst1a/NST1A*	*Glyma.07g050600*	7	4,332,840	[26], this work	4,231,147	ss715598065	(101,693)	98.4	93.7
						42,37,522	ss715598070	(95,318)	98.2	93.7
						4,247,302	ss715598081	(85,538)	96.2	90.7
						4,277,666	ss715598106	(55,174)	95.2	92.6
						4,265,150	ss715598093	(67,690)	93.9	90.8
Pod shatter	*nst1b/SHAT1-5*	*Glyma.16g019400*	16	1,727,642	[45]	1,725,360	ss715623567	(2,282)	94.4	90.8
						1,533,772	ss715623488	(193,870)	79.5	80.0
						1,950,911	ss715623640	223,269	78.5	76.4
						1,953,870	ss715623642	226,228	78.2	67.9
						1,498,023	ss715623475	(229,619)	77.6	65.8
Green seed coat	*g/G*	*Glyma.01G198500*	1	53,229,579	[46]	53,151,056	ss715580344	(78,523)	90.2	85.8
						53,141,084	**ss715580343**	(88,495)	88.5	84.3
						53,378,518	ss715580361	148,939	86.8	85.3
						53,335,309	ss715580358	105,730	77.4	79.7
						54,250,600	ss715580443	1,021,021	75.3	80.0
Hard seed	*hs1-1/Hs1-1*	*Glyma.02g269500*	2	45,379,743	[47]	44,975,684	ss715583161	(404,059)	87.5	83.6
						46,823,818	ss715583338	1,444,075	84.6	88.9
						46,873,183	ss715583343	1,493,440	83.8	31.9
						45,207,651	ss715583177	(172,092)	83.2	82.8
						45,708,763	ss715583227	329,020	80.2	68.3
Pubsecence color light tawny	*Td/td*	*Glyma.03G258700*	3	45,301,350	[48]	45,243,426	ss715586624	(57,924)	88.7	84.1
						45,385,087	ss715586641	83,737	84.3	67.9
						45,586,075	ss715586661	284,725	77.9	85.0
						45,462,321	ss715586654	160,971	77.8	84.8
						45,159,972	ss715586611	(141,378)	73.8	52.1
flowering time/maturity	*E1la/e1la* K82E	*Glyma.04g156400*	4	36,758,368	[49]	38,617,883	ss715587795	1,859,515	98.5	92.5
						38,690,809	ss715587797	1,932,441	98.2	92.6
						37,366,196	ss715587592	607,828	97.8	90.7
						37,750,626	ss715587601	992,258	97.8	92.0
						37,650,714	ss715587599	892,346	97.7	90.8
Maturity	*e1-as/E1*	*Glyma.06g207800*	6	20,207,322	[50]	20,966,229	ss715593867	758,907	98.0	72.5
						19,913,355	**ss715593833**	(293,967)	97.9	92
						20,633,420	ss715593856	426,098	97.0	93.9
						19,858,251	ss715593832	(349,071)	96.9	70.8
						20,596,465	ss715593854	389,143	96.7	93.4
Pubsecence color tawny	*T/t*	*Glyma.06g202300*	6	18,737,366	[51]	18,315,510	ss715593787	(421,856)	96.2	94.1
						18,446,052	**ss715593791**	(291,314)	96.1	94.3
						18,970,072	ss715593807	232,706	93.1	92.5
						18,896,222	ss715593805	158,856	88.4	81.4
						18,788,512	ss715593801	51,146	86.5	79.2

(*continued on next page*)

Table 3 (*continued*)

Trait	REF/ALT allele	Williams 82.a2.v1	Chromo-some	CM position	Reference	SoySNP50K Position	Proxy marker	Distance to CM	Avg_ Accu	Comb_ Acc_Pess
Black/brown seed coat	*R/r*	*Glyma.09g235100*	9	45,759,137	[52]	45,815,773	ss715604620	56,636	85.8	72.9
						45,739,541	ss715604613	(19,596)	84.0	70.1
						45,694,733	ss715604610	(64,404)	82.7	67.9
						46,035,289	ss715604640	276,152	81.8	82.3
						45,781,506	ss715604617	22,369	79.9	65.0
Maturity	*E2/e2*	*Glyma.10g221500*	10	45,310,798	[53]	44,852,490	ss715607431	(458,308)	90.3	85.7
						44,920,131	ss715607435	(390,667)	87.8	84.8
						45,269,968	ss715607475	(40,830)	87.3	82.5
						45,250,482	ss715607471	(60,316)	86.3	78.8
						44,622,989	ss715607402	(687,809)	86.1	78.3
Flowering time/maturity	*tof12-1/Tof12*	*Glyma.12g073900*	12	5,520,945	[54]	5,677,390	ss715613198	156,445	95.1	93.9
						5,409,612	ss715613172	(111,333)	93.4	33.9
						5,400,963	ss715613171	(119,982)	92.2	91.7
						5,502,184	**ss715613180**	(18,761)	91.7	89.9
						5,724,257	ss715613204	203,312	89.2	90.3
Seed coat luster	*b1/B1*	*Glyma.13G241700*	13	35,163,354	[55]	35,301,446	ss715615610	138,092	98.2	97.0
						35,307,166	**ss715615611**	143,812	98.2	97.4
						35,077,503	ss715615595	(85,851)	97.3	97.8
						34,620,193	ss715615548	(543,161)	89.5	93.8
						35,584,259	ss715615638	420,905	88.5	80.5
Semi-determinate	*dt2/Dt2*	*Glyma.18g273600*	18	55,642,486	[56]	55,622,046	**ss715632223**	(20,440)	95.7	97.8
						55,675,146	ss715632229	32,660	88.0	76.4
						55,279,467	ss715632165	(363,019)	77.2	74.5
						55,700,093	ss715632231	57,607	75.0	49.5
						55,643,993	ss715632225	1,507	74.4	49.4
Narrow leaves/3 seeded pods	*Ln*/ln	*Glyma.20g116200*	20	35,828,042	[57]	36,074,213	ss715637655	246,171	93.5	86.7
						35,645,532	ss715637614	(182,510)	91.8	31.2
						35,616,323	ss715637607	(211,719)	79.8	69.4
						37,091,712	ss715637807	1,263,670	79.8	73.4
						36,052,996	ss715637652	224,954	79.1	59.7

the landscape of *NST1A* ss715598106 TM associated genomic region.

The most widely used approach for GWAS as a step in CM identification of various phenotypes in crop species is the practice of using low-density genotyping data for phenotypes obtained on relatively small sets of accessions and association panels [60,61]. Even though GWAS methodology is being continuously improved [1], GWAS is rarely perfect and therefore, identification of a suboptimal haplotype tagging variant in under-powered GWAS may lead to the wrong conclusions without any additional GWAS selection criteria. Here we demonstrated that the Accuracy calculation for every variant position in a phenotype-associated genomic region increased efficiency when using low-density genotypes for GWAS. A key to this approach is access to a quality resequenced data set. In soybean, there are tremendous resources for phenotype and genotype data. The publicly available GRIN collection of accessions with SoySNP50K marker data along with resequenced data sets for an increasing number of accessions, many of which overlap with the GRIN collection, were utilized here. The approach of leveraging resequenced data for low-density marker GWAS is broadly applicable to other species that have been limited by current GWAS power. For many species, the decision to invest additional resources in phenotyping accessions that have already been resequenced or to resequence accessions that have already been phenotyped must be seriously considered. In the decision making, among other factors, the frequency of the targeted phenotypes will play an important role. And, undoubtedly, both approaches are laborious resource-intensive undertakings. In plants, other species that could benefit immediately from our strategy have large accession panels with low-density genotypes available, a reference genome sequence along with a set of at least 200 resequenced accessions that overlap with the phenotyped accession panels, and broad LD, such as is the case for rice (*Oryza sativa* L.) [62], sorghum (*Sorghum bicolor* [L.] Moench) [63,64], and *Arabidopsis thaliana* [65]. Most major crop species now have at least one reference genome sequence available. Crop species with broad LD (typically self-pollinating species) with a record of GWAS would require additional investments in generating and analyzing resequenced data sets as would be the case for most of the crop legumes [66].

Calculating Accuracy enables a direct assessment of correspondence and a novel view of the landscape of LD in focused genomic regions utilizing our AccuTool. A more visual representation of haplotype blocks in user-defined regions is possible using our SNPViz v2.0 tool [12–14]. When phenotypes are available for at least a subset of resequenced accessions, Accuracy accentuation of associated variant positions increases the efficiency of selecting tagging variants with stronger LD to the CM. When setting up a new GWAS panel of accessions for phenotyping, including accessions that have resequence data will increase the power to leverage low-density markers for ultimate identification of the CM. Our proposed Accuracy calculation is limited by the binomial distribution of studied phenotypes that goes hand in hand with the bi-allelic nature of genetic variants in the vast majority of cases [15]; however, in the final step of deciding between a functional and a non-functional gene variant, the binomial concept is valid. Recently, a binomial categorization of phenotypes was applied in FPCA-based GWAS on quantitative traits with successful outcome on sorghum [67].

Because CMs have been identified in a multitude of cloned genes responsible for a variety of traits in soybean, we were able to investigate the LD landscape in the associated region around the CM in a modified GWAS approach that we call Inverse GWAS,

the extreme variant of the SPAS. Using the allele status of the CM as a phenotype and calculating Accuracy with the AccuTool for SoySNP50K markers in the CM region exposed the low-density Proxy markers with the best ability to predict the CM status, typically with Accuracies ranging from 99% to 74% for the top five Proxy markers for each CM (Table 3 and Table S1). However, the distance to the CM was quite variable for the Proxy markers; for 17 cloned soybean genes, Proxy markers with > 90% Accuracy averaged over 340,000 bp from the CM, and this could be due to the density of the SoySNP50K markers in any particular region or evolutionary aspects of the CM. For soybean, the low-density SoySNP50K markers represent only about 0.1% of the currently defined variant positions in the soybean genome. Inverse GWAS therefore provides evidence that caution should be used when generating candidate gene lists with speculative CMs from GWAS based on proximity to highly associated tagging variants. In contrast, GWAS tagging variant results that overlap with our Proxy markers are likely pointing to the CMs of the cloned genes.

Recent advances in post-GWAS methodology, especially coupling with transcriptomics, eQTL or even gene expression in association analyses (TWAS) have shown that this strategy can complicate the hunt for CMs with false positive associations [5,68]. Therefore, here we emphasize that our concept of Accuracy calculation can be used in any GWAS as an additional evaluation criterion that adds no false positives or negatives and thus, directly improves every GWAS. This study aimed to reveal the landscapes of association, and it resulted in the creation of the novel GWAS to Genes strategy, the SP2CM. However, utilization of the SP2CM strategy in identification of uncloned genes will need further investigation that we plan to focus on in our future work.

Conclusion

Our results on SPAS pointed us to direct application of the Synthetic phenotype, Accuracy calculation and the use of aggregated panels of resequenced accessions in what we call the SP2CM strategy. This strategy can benefit future GWAS to Genes discoveries across species that are currently limited by insufficient GWAS power.

Additional Information

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Compliance with Ethics Requirement

This article does not contain any studies with human or animal subjects.

CRediT authorship contribution statement

Mária Škrabišová: Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Nicholas Dietz:** Data curation, Formal analysis, Methodology, Software, Validation. **Shuai Zeng:** Data curation, Formal analysis, Methodology, Software, Validation, Visualization. **Yen On Chan:** Data curation, Formal analysis, Methodology, Software, Validation. **Juexin Wang:** Data curation, Formal analysis, Methodology, Software, Validation. **Yang Liu:** Data curation, Formal analysis, Methodology, Software, Validation. **Jana Biová:** Formal analysis, Methodology, Software. **Trupti Joshi:** Project administration, Funding acquisition, Resources, Supervision, Writing – review & editing. **Kristin D. Bilyeu:** Investigation, Project administration, Funding acquisition, Resources, Supervision, Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research has been conducted using the USDA Soybean Germplasm Collection (Urbana, IL, application note). The research was supported using Missouri soybean farmers' checkoff dollars provided by the Missouri Soybean Merchandising Council (MSMC) and the United Soybean Board (USB): MSMC (KB and TJ: GWAS to Genes, Project #385), USB (TJ and KB: Applied Genomics to Improve Soybean Seed Protein, #1920-152-0131-C, #2220-152-0202) and IGA (MS: Palacký University Internal Grant Agency #IGA_2020_013) projects. Authors wish to thank the following contributors to this work from Palacký University in Olomouc, Faculty of Science, Czech Republic: Tomáš Fürst (Department of Mathematical Analysis and Applied Mathematics) for his invaluable insight into Bayesian approach to high dimensional genetic multicollinearity and Jana Slivková (Department of Biochemistry) for critical reading of the manuscript. USDA is an equal opportunity provider and employer.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2022.04.004.

References

[1] Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet 2019;20(8):467–84. doi: https://doi.org/10.1038/s41576-019-0127-1.
[2] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 2017;101(1):5–22. doi: https://doi.org/10.1016/j.ajhg.2017.06.005.
[3] Filho DF, Filho JS de SB, Regitano LC de A, Alencar MM de, Alves RR, Meirelles SLC. Tournaments between markers as a strategy to enhance genomic predictions. PLoS One 2019;14:e0217283. https://doi.org/10.1371/journal.pone.0217283.
[4] Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet 2015;24(R1):R111–9. doi: https://doi.org/10.1093/hmg/ddv260.
[5] Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. Nat Genet 2019;51(5):768–9. doi: https://doi.org/10.1038/s41588-019-0404-0.
[6] Zhou Z, Jiang Yu, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 2015;33(4):408–14. doi: https://doi.org/10.1038/nbt.3096.
[7] Valliyodan B, Dan Qiu, Patil G, Zeng P, Huang J, Dai Lu, et al. Landscape of genomic diversity and trait discovery in soybean. Sci Rep 2016;6(1). doi: https://doi.org/10.1038/srep23598.
[8] Kim JY, Jeong S, Kim KH, Lim WJ, Lee HY, Jeong N, et al. Dissection of soybean populations according to selection signatures based on whole-genome sequences. GigaScience 2019;8:1–19. doi: https://doi.org/10.1093/gigascience/giz151.
[9] Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. Cell 2020;182(1):162–176.e13. doi: https://doi.org/10.1016/j.cell.2020.05.023.
[10] Zhang H, Jiang H, Hu Z, Song Q, An YC. A versatile resource of 1500 diverse wild and cultivated soybean genomes for post-genomics research. BioRxiv 2020:2020.11.16.383950.
[11] Valliyodan B, Brown AV, Wang J, Patil G, Liu Y, Otyama PI, et al. Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. Sci Data 2021;8(1). doi: https://doi.org/10.1038/s41597-021-00834-w.

[12] Langewisch T, Zhang H, Vincent R, Joshi T, Xu D, Bilyeu K, et al. Major soybean maturity gene haplotypes revealed by SNPViz analysis of 72 sequenced soybean genomes. PLoS ONE 2014;9(4):e94150. doi: https://doi.org/10.1371/journal.pone.0094150.
[13] Zeng S, Skrabisova M, Lyu Z, Chan YO, Bilyeu K, Joshi T. SNPViz v2.0: A web-based tool for enhanced haplotype analysis using large scale resequencing datasets and discovery of phenotypes causative gene using allelic variations. In: Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020, Institute of Electrical and Electronics Engineers Inc.; 2020, p. 1408–15. https://doi.org/10.1109/BIBM49941.2020.9313539.
[14] Zeng S, Škrabišová M, Lyu Z, Chan YO, Dietz N, Bilyeu K, et al. Application of SNPViz v2.0 using next-generation sequencing data sets in the discovery of potential causative mutations in candidate genes associated with phenotypes. IJDMB 2021;25(1/2):65. doi: https://doi.org/10.1504/IJDMB.2021.116886.
[15] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409(6822):928–33. doi: https://doi.org/10.1038/35057149.
[16] Miranda C, Culp C, Škrabišová M, Joshi T, Belzile F, Grant DM, et al. Molecular tools for detecting Pdh1 can improve soybean breeding efficiency by reducing yield losses due to pod shatter. Mol Breed 2019;39:1–9. doi: https://doi.org/10.1007/s11032-019-0935-1.
[17] Langewisch T, Lenis J, Jiang G-L, Wang D, Pantalone V, Bilyeu K. The development and use of a molecular model for soybean maturity groups. BMC Plant Biol 2017;17:91. doi: https://doi.org/10.1186/s12870-017-1040-4.
[18] Li X, Shi Z, Qie Q, Gao J, Wang X, Han Y. CandiHap: a toolkit for haplotype analysis for sequence of samples and fast identification of candidate causal gene(s) in genome-wide association study. Cold Spring Harbor Laboratory; 2020. http://doi.org/10.1101/2020.02.27.967539.
[19] Shi A, Buckley B, Mou B, Motes D, Morris JB, Ma J, et al. Association analysis of cowpea bacterial blight resistance in USDA cowpea germplasm. Euphytica 2016;208(1):143–55. doi: https://doi.org/10.1007/s10681-015-1610-1.
[20] Ravelombola WS, Qin J, Shi A, Nice L, Bao Y, Lorenz A, et al. Genome-wide association study and genomic selection for tolerance of soybean biomass to soybean cyst nematode infestation. PLoS ONE 2020;15(7):e0235089. doi: https://doi.org/10.1371/journal.pone.0235089.
[21] Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol 2017;18(1). doi: https://doi.org/10.1186/s13059-017-1289-9.
[22] Valliyodan B, Cannon SB, Bayer PE, Shu S, Brown AV, Ren L, et al. Construction and comparison of three reference-quality genome assemblies for soybean. Plant J 2019;100(5):1066–82. doi: https://doi.org/10.1111/tpj.14500.
[23] Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, et al. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS ONE 2013;8(1):e54985. doi: https://doi.org/10.1371/journal.pone.0054985.
[24] Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, et al. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 2015;8(3). doi: https://doi.org/10.3835/plantgenome2015.04.0024.
[25] Bandillo NB, Lorenz AJ, Graef GL, Jarquin D, Hyten DL, Nelson RL, et al. Genome-wide association mapping of qualitatively inherited traits in a germplasm collection. Plant Genome 2017;10(2). doi: https://doi.org/10.3835/plantgenome2016.06.0054.
[26] Zhang J, Singh AK. Genetic control and geo-climate adaptation of pod dehiscence provide novel insights into soybean domestication. G3 2020;10:545–54. doi: https://doi.org/10.1534/g3.119.400876.
[27] Liu Y, Khan SM, Wang J, Rynge M, Zhang Y, Zeng S, et al. PGen: Large-scale genomic variations analysis workflow and browser in SoyKB. BMC Bioinf 2016;17(S13). doi: https://doi.org/10.1186/s12859-016-1227-y.
[28] Joshi T, Patil K, Fitzpatrick MR, Franklin LD, Yao Q, Cook JR, et al. Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. BMC Genomics 2012;13(S1). doi: https://doi.org/10.1186/1471-2164-13-S1-S15.
[29] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–303. doi: https://doi.org/10.1101/gr.107524.110.
[30] Joshi T, Wang J, Zhang H, Chen S, Zeng S, Xu B, et al. The evolution of soybean knowledge base (SoyKB). Methods Mol. Biol., vol. 1533, Humana Press Inc.; 2017, p. 149–59. https://doi.org/10.1007/978-1-4939-6658-5_7.
[31] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6(2):80–92. doi: https://doi.org/10.4161/fly.19695.
[32] Hill JL, Peregrine EK, Sprau GL, Cremeens CR, Nelson RL, Kenty MM, et al. Evaluation of the USDA soybean germplasm collection: maturity groups 000-IV (PI 578371-PI 612761). US Dep Agric Tech Bull 2001:1894.
[33] Zabala G, Vodkin LO. A rearrangement resulting in small tandem repeats in the F3′5′H gene of white flower genotypes is associated with the soybean W1 locus. Crop Sci 2007;47:113–24. doi: https://doi.org/10.2135/cropsci2006.12.0838tpg.
[34] Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet 2010;42(4):355–60. doi: https://doi.org/10.1038/ng.546.
[35] Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F, et al. GAPIT version 2: An enhanced integrated tool for genomic association and prediction. Plant Genome 2016;9(2). doi: https://doi.org/10.3835/plantgenome2015.11.0120.
[36] Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 2007;23(19):2633–5. doi: https://doi.org/10.1093/bioinformatics/btm308.
[37] Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8(4):283–98. doi: https://doi.org/10.1016/S0001-2998(78)80014-2.
[38] Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: web application framework for R. R package version 1.3.2. 2019.
[39] Phanstiel DH, Boyle AP, Araya CL, Snyder M. Sushi: An R/Bioconductor package for visualizing genomic data. R Packag Version 1260 2020.
[40] Wickham H. Programming with ggplot2. Springer Science & Business Media; 2016.
[41] Funatsuki H, Suzuki M, Hirose A, Inaba H, Yamada T, Hajika M, et al. Molecular basis of a shattering resistance boosting global dissemination of soybean. Proc Natl Acad Sci U S A 2014;111(50):17797–802. doi: https://doi.org/10.1073/pnas.1417282111.
[42] Palmer RG, Pfeiffer TW, Buss GR, Kilen TC. Qualitative genetics. In: Shibles RM, Harper JE, Wilson RF, Shoemaker RC, editors. Soybeans Improv. Prod. Uses. 3rd ed., John Wiley & Sons, Ltd; 2016, p. 137–233. https://doi.org/10.2134/agronmonogr16.3ed.c5.
[43] Liu B, Watanabe S, Uchiyama T, Kong F, Kanazawa A, Xia Z, et al. The soybean stem growth habit gene Dt1 is an ortholog of arabidopsis TERMINAL FLOWER1. Plant Physiol 2010;153(1):198–210. doi: https://doi.org/10.1104/pp.109.150607.
[44] Tian Z, Wang X, Lee R, Li Y, Specht JE, Nelson RL, et al. Artificial selection for determinate growth habit in soybean. Proc Natl Acad Sci U S A 2010;107(19):8563–8. doi: https://doi.org/10.1073/pnas.1000088107.
[45] Dong Y, Yang X, Liu J, Wang B-H, Liu B-L, Wang Y-Z. Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nat Commun 2014;5:3352. doi: https://doi.org/10.1038/ncomms4352.
[46] Wang M, Li W, Fang C, Xu F, Liu Y, Wang Z, et al. Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat Genet 2018;50(10):1435–41. doi: https://doi.org/10.1038/s41588-018-0229-2.
[47] Sun L, Miao Z, Cai C, Zhang D, Zhao M, Wu Y, et al. GmHs1-1, encoding a calcineurin-like protein, controls hard-seededness in soybean. Nat Genet 2015;47(8):939–43. doi: https://doi.org/10.1038/ng.3339.
[48] Yan F, Githiri SM, Liu Y, Sang Y, Wang Q, Takahashi R. Loss-of-Function Mutation of Soybean R2R3 MYB Transcription Factor Dilutes Tawny Pubescence Color. Front Plant Sci 2020;10:1–12. doi: https://doi.org/10.3389/fpls.2019.01809.
[49] Xia Z, Zhai H, Wu H, Xu K, Watanabe S, Harada K. The Synchronized Efforts to Decipher the Molecular Basis for Soybean Maturity Loci E1, E2, and E3 That Regulate Flowering and Maturity. Front Plant Sci 2021;12. doi: https://doi.org/10.3389/FPLS.2021.632754.
[50] Xia Z, Watanabe S, Yamada T, Tsubokura Y, Nakashima H, Zhai H, et al. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc Natl Acad Sci U S A 2012;109(32). doi: https://doi.org/10.1073/pnas.1117982109.
[51] Zabala G, Vodkin L. Cloning of the Pleiotropic T Locus in Soybean and Two Recessive Alleles That Differentially Affect Structure and Expression of the Encoded Flavonoid 3′ Hydroxylase. Genetics 2003;163:295–309. doi: https://doi.org/10.1093/genetics/163.1.295.
[52] Gillman JD, Tetlow A, Lee J-D, Shannon J, Bilyeu K. Loss-of-function mutations affecting a specific Glycine max R2R3 MYB transcription factor result in brown hilum and brown seed coats. BMC Plant Biol 2011;11(1):155. doi: https://doi.org/10.1186/1471-2229-11-155.
[53] Watanabe S, Xia Z, Hideshima R, Tsubokura Y, Sato S, Yamanaka N, et al. A Map-Based Cloning Strategy Employing a Residual Heterozygous Line Reveals that the GIGANTEA Gene Is Involved in Soybean Maturity and Flowering. Genetics 2011;188(2):395–407. doi: https://doi.org/10.1534/genetics.110.125062.
[54] Lu S, Dong L, Fang C, Liu S, Kong L, Cheng Q, et al. Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. Nat Genet 2020;52(4):428–36. doi: https://doi.org/10.1038/s41588-020-0604-7.
[55] Zhang D, Sun L, Li S, Wang W, Ding Y, Swarm SA, et al. Elevation of soybean seed oil content through selection for seed coat shininess. Nat Plants 2018;4(1):30–5. doi: https://doi.org/10.1038/s41477-017-0084-7.
[56] Ping J, Liu Y, Sun L, Zhao M, Li Y, She M, et al. Dt2 is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. Plant Cell 2014;26(7):2831–42. doi: https://doi.org/10.1105/tpc.114.126938.
[57] Jeong N, Suh SJ, Kim M-H, Lee S, Moon J-K, Kim HS, et al. Ln is a key regulator of leaflet shape and number of seeds per pod in soybean. Plant Cell 2012;24(12):4807–18. doi: https://doi.org/10.1105/tpc.112.104968.
[58] Fang C, Li W, Li G, Wang Z, Zhou Z, Ma Y, et al. Cloning of Ln Gene Through Combined Approach of Map-based Cloning and Association Study in Soybean. J Genet Genomics 2013;40(2):93–6. doi: https://doi.org/10.1016/j.jgg.2013.01.002.
[59] Sesia M, Bates S, Candès E, Marchini J, Sabatti C. False discovery rate control in genome-wide association studies with population structure. Proc Natl Acad Sci U S A 2021;118(40). doi: https://doi.org/10.1073/pnas.2105841118.

[60] Deng Y, Pan W. Improved use of small reference panels for conditional and joint analysis with gwas summary statistics. Genetics 2018;209(40):401–8. doi: https://doi.org/10.1534/genetics.118.300813.
[61] Benner C, Havulinna AS, Järvelin M-R, Salomaa V, Ripatti S, Pirinen M. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. Am J Hum Genet 2017;101 (4):539–51. doi: https://doi.org/10.1016/j.ajhg.2017.08.012.
[62] Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, et al. RiceVarMap: A comprehensive database of rice genomic variations. Nucleic Acids Res 2015;43 (D1):D1018–22. doi: https://doi.org/10.1093/nar/gku894.
[63] Valluru R, Gazave EE, Fernandes SB, Ferguson JN, Lozano R, Hirannaiah P, et al. Deleterious mutation burden and its association with complex traits in sorghum (Sorghum bicolor). Genetics 2019;211(3):1075–87. doi: https://doi.org/10.1534/genetics.118.301742.
[64] Mockler T. A Complete-Sequence Population for Pan-Genome Analysis of Sorghum 2016. https://doi.org/10.25585/1488180.
[65] Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 2016;166(2):481–91. doi: https://doi.org/10.1016/j.cell.2016.05.063.
[66] Bauchet GJ, Bett KE, Cameron CT, Campbell JD, Cannon EKS, Cannon SB, et al. The future of legume genetic data resources: Challenges, opportunities, and priorities. Legum Sci 2019;1(1). doi: https://doi.org/10.1002/leg3.16.
[67] Miao C, Xu Y, Liu S, Schnable PS, Schnable JC. Increased power and accuracy of causal locus identification in time-series genome-wide association in sorghum. Plant Physiol 2020;183(4):1898–909. doi: https://doi.org/10.1104/pp.20.00277.
[68] Li D, Liu Q, Schnable PS. TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. Plant Physiol 2021;186(4):1800–11. doi: https://doi.org/10.1093/plphys/kiab161.