GigaScience

Association Mapping Across a Multitude of Traits Collected in Diverse Environments in Maize

--Manuscript Draft--

Manuscript Number:	GIGA-D-22-00083R1
Full Title:	Association Mapping Across a Multitude of Traits Collected in Diverse Environments in Maize
Article Type:	Research
Funding Information:	Advanced Research Projects Agency - Energy (DE-AR0001064) — Dr James C Schnable Advanced Research Projects Agency - Energy (DE-AR0001367) — Dr James C Schnable National Science Foundation (OIA-1557417) — Dr James C Schnable National Science Foundation (OIA-1826781) — Dr James C Schnable USDA-NIFA under the AI Institute: for Resilient Agriculture (2021-67021-35329) — Dr James C Schnable Foundation for Food and Agriculture Research (602757) — Dr James C Schnable US. Department of Agriculture (5030-21000-068-00D) — Dr James C Schnable
Abstract:	Classical genetic studies have identified many cases of pleiotropy where mutations in individual genes alter many different phenotypes. Quantitative genetic studies of natural genetic variants frequently examine one or a few traits, limiting their potential to identify pleiotropic effects of natural genetic variants. Widely adopted community association panels have been employed by plant genetics communities to study the genetic basis of naturally occurring phenotypic variation in a wide range of traits. High-density genetic marker data -- 18M markers -- from two partially overlapping maize association panels comprising 1,014 unique genotypes grown in field trials across at least seven US states and scored for 162 distinct trait datasets enabled the identification of of 2,154 suggestive marker-trait associations and 697 confident associations in the maize genome using a resampling-based genome-wide association strategy. The precision of individual marker-trait associations was estimated to be three genes based a reference set of genes with known phenotypes. Examples were observed of both genetic loci associated with variation in diverse traits (e.g. above-ground and below-ground traits), as well as individual loci associated with the same or similar traits across diverse environments. Many significant signals are located near genes whose functions were previously entirely unknown or estimated purely via functional data on homologs. This study demonstrates the potential of mining community association panel data using new higher density genetic marker sets combined with resampling-based genome-wide association tests to develop testable hypotheses about gene functions, identify potential pleiotropic effects of natural genetic variants, and study genotype by environment interaction.
Corresponding Author:	James Schnable University of Nebraska-Lincoln Lincoln, NE UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Nebraska-Lincoln
Corresponding Author's Secondary Institution:	
First Author:	Ravi V Mural, Ph.D

First Author Secondary Information:	
Order of Authors:	Ravi V Mural, Ph.D
	Guangchao Sun, Ph.D
	Marcin Grzybowski, Ph.D
	Michael C Tross
	Hongyu Jin
	Christine Smith
	Linsey Newton
	Carson M Andorf
	Margaret R Woodhouse
	Addie M Thompson, Ph.D
	Brandi Sigmon, Ph.D
	James C Schnable, Ph.D
Order of Authors Secondary Information:	
Response to Reviewers:	Reviewer reports: Reviewer #1: Mural et al. reported a large-scale association analysis based on publicly published genotype and phenotype datasets and a meta-GWAS. This study provides a good example for mining community association panel data and further identifying candidate genes, pleiotropic loci and G x E. Actually, meta-analysis of GWAS has been used in humans and animals. However, I have some major concerns as follows. 1. This study only used three association panels (MAP, SAM, and WiDiv), as I know, some publicly available genotype and phenotype could be obtained for other association panels, for example the association panel including 368 inbred lines (Li et al., 2013, Nat Genetics, 45(1):43-50. doi: 10.1038/ng.2484), which was used widely in GWAS studies in maize. Can other association panels be integrated into this research, which would provide a rich genetic resource for maize research groups. We agree with the reviewer that the integration of additional panels has the potential to increase the power of massive community GWAS in maize, a point we now touch on in the discussion section of this revised manuscript (Page 10, Right column, Second paragraph). In the specific case of the AM508 panel from Li et al 2013 (now cited in the introduction (Page 2, Left column, Third paragraph) and discussion (Page 10, Right column, Second paragraph) of our revised manuscript) we faced two challenges. The first is that only a minority of published GWAS papers for this study include phenotype data for the individual lines. The greater challenge is that the current genetic marker set used for GWAS with the AM508 panel is from an earlier version of the maize genome -- the website from Li et al 2013 was not accessible to us, but based on gene names and publication data we can narrow it down to B73 RefGen_v2 or v_3. Unfortunately this means the genetic markers used for the AM508 GWAS will not share markers/coordinates with the marker set employed in this study which was based on the B73_RefGen_v4 reference genome. In the future this could be addressed either through reanalysis published resequencing of the AM508 panel, enabling integrated analysis of North American and Chinese maize diversity panels. We now discuss these options in our revised discussions section (Page 10, Right column, Second paragraph). 2. For association analysis, a total of 1014 unique inbred lines and 162 distinct traits from different association panels were used, but these traits were not measured for each of 1041 inbreds. For example, cellular-related traits were mainly measured in the

SAM association panel. Hence, association analysis for cellular-related traits were conducted in SAM or 1014 inbreds. If 1014 inbreds were used to perform association analysis for cellular-related traits, how did you analyze the phenotype data? Please describe the method of phenotype data analysis in the Method section.

We apologize for this confusion. For each trait genetic data was subset to only that subset of individuals for which measured trait values were present. We have revised our methods section to make this less ambiguous (Page 12, Left column, Second paragraph of the “Quantitative Genetic Analysis of Trait Data” subsection).

3. Authors used RMIP values to identify significant association signals, please add more details about the RMIP method. What advantages of the resampling-based genome-wide association strategy over other methods?

You are right, this was a significant omission. We have added a new paragraph to the introduction discussing the RMIP method (Page 2, Right Column, Second Paragraph). Briefly, RMIP allows us to employ the FarmCPU algorithm which provides greater power to detect additional trait associated loci which might be masked by the effects of one or more large effect genes without the instability of results found in single individual runs of the FarmCPU GWAS algorithm. As one of our big goals in conducting this analysis was to generate a dataset that would enable other researchers to compare their results to. For this purpose the stability of marker trait associations is of particular value.

4. Although some important functional genes could be identified, were some new candidate genes obtained in this study functionally verified by the mutants or overexpression experiments.

Unfortunately no, the scope of our study did not permit us to conduct transgenic or gene knock out validation of new candidate genes. We have revised our discussion section to emphasize some of our highest priority candidates for future validation.

5. The authors identified pleiotropic loci based on categories of phenotypes associated with the same peak. For example, the phenotypes associated with the pleiotropic peak on chromosome 8 from 134,706,389 to 134,759,977 bp belongs to Flowering Time, Root and Vegetative categories, thus the locus was associated with different traits. Do you have any ideas on pleiotropic genes based on the results?

We have revised our manuscript to include discussion of both cases where a known causal gene is associated with a pleiotropic locus identified in our study (e.g. MADS69, Liang et al 2018), as well as to discuss in more detail potential candidate genes for two loci with pleiotropic effects without previously known and validated candidate genes associated (Page 10, Right Column, Third Paragraph)

Reviewer #2: The authors described a study of integrating multiple published datasets for reanalysis. They combined previously community panel data and newly collected data in the present study, finally assembling 1014 accessions with 18M SNP markers and 162 traits at different environments. They used a resample-based GWAS method to reanalyze this assemble dataset, and identified 2154 suggestive associations and 697 confident associations. They found genetic loci were pleiotropic to multiple traits. As the authors mentioned, I acknowledge their efforts for collecting and assembling different sources of previously datasets, which should be useful for the maize community. However, to the manuscript per se, I feel the paper seems not to be sufficiently quantified regarding the novelty and significance of reported findings.
If the authors could present several novel results because the previous studies had the limitations on population size, diversity, trait dimensions and environments. In this study, the authors seemed trying to present like this, but it may be improved further and more. It's hard to let me understand there are some novel things which was found due to the merged large dataset. On the other hand, using this assembled dataset, I'm not very clear what's the scientific questions that the authors want to address.

This is a very insightful comment that speaks to the difference approaches research groups take to science. Our goal in assembling this dataset was not to address a

	specific scientific question, but to generate a dataset/resource which would be valuable and reusable for multiple scientific research avenues. Essentially this boils down the distinction between hypothesis testing research and hypothesis generating research. In terms of what novel things that were found as a result of our initial proof of concept analyses of this large dataset our intent was to emphasize both the quantitative genetic evidence for pleiotropy between above ground and below ground traits. In this revised manuscript we have revised and added text to emphasize the novelty of this finding. See also our response to reviewer #1 point #4. In technical sense, I'm wondering how did authors deal with the batch effects when merging datasets phenotype from different environments? It's not comparable for the phenotypes from different accessions collected in different environments. It's hard to figure out the phenotypic difference is caused by genotype, environment, or their interaction. We apologize that this was not explained sufficiently clear in our previous version. We did not merge data across multiple environments, unless the merging was conducted by the authors prior to making them public by the authors. We have revised our methods section to clarify this that we analyzed these datasets separately rather than merging and producing batch effects (Page 12, Left column, Second paragraph of “Quantitative Genetic Analysis of Trait Data” subsection). This means when a peak is identified in one dataset and not another for same phenotype from different studies is because of genotype by environment interacts that we now discuss in greater length in the discussion section (Page 10, Left Column, Third Paragraph). The introduction section lacked the proper review for the project background, related progress and publications and findings. We were sorry to read that the reviewer feels we did not provide sufficient background and citations in the previous version of this manuscript, and assure him or her that it was not our intent to minimize or omit references to the work of other researchers within this field. While we are unsure what specific publications and findings the reviewer feels were improperly omitted by us, this revised manuscript includes an expanded introduction and discussion and 14 additional new citations not present in the original manuscript. Reviewer #3: The manuscript "Association Mapping Across a Multitude of Traits Collected in Diverse Environments in Maize" by Ravi V. Mural et al. reported the application of high-density genetic marker data from two partially overlapping maize association panels, comprising 1,014 unique genotypes grown in seven US states, allowing the identification 2,154 suggestive marker-trait associations and 697 confident associations and suggesting the possible application to study gene functions, pleiotropic effects of natural genetic variants and genotype by environment interaction. The background data are well documented, experimental data are convincing, clearly presented and well discussed, the paper is suitable for publication in Giga Science in its present form. Thank your for your kind words regarding our manuscript and work.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes

Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

GigaScience, 2017, 1–29

doi: xx.xxxx/xxxx
Manuscript in Preparation
Paper

PAPER

Association Mapping Across a Multitude of Traits Collected in Diverse Environments in Maize

Ravi V. Mural [0000-0002-5489-9918][1,2], Guangchao Sun [0000-0003-3942-6175][1,2], Marcin Grzybowski [0000-0002-6303-5433][1,2], Michael C. Tross [0000-0002-4410-9679][1,2], Hongyu Jin [0000-0002-4562-0639][1,2], Christine Smith [0000-0001-8128-2924][1], Linsey Newton [0000-0002-7149-4752][3], Carson M. Andorf [0000-0003-2380-6704][5,6], Margaret R. Woodhouse [0000-0003-2164-8300][5], Addie M. Thompson [0000-0002-4442-6578][3], Brandi Sigmon [0000-0003-3914-6894][4] and James C. Schnable [0000-0001-6739-5527][1,2,*]

[1]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE, 68588 USA and [2]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, 68588 USA and [3]Department of Plant Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824 USA and [4]Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln, NE, 68588 USA and [5]USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA, 50010 USA and [6]Department of Computer Science, Iowa State University, Ames, IA 50011, USA

[*]schnable@unl.edu; +1 (402) 472-4540

Abstract

Classical genetic studies have identified many cases of pleiotropy where mutations in individual genes alter many different phenotypes. Quantitative genetic studies of natural genetic variants frequently examine one or a few traits, limiting their potential to identify pleiotropic effects of natural genetic variants. Widely adopted community association panels have been employed by plant genetics communities to study the genetic basis of naturally occurring phenotypic variation in a wide range of traits. High-density genetic marker data – 18M markers – from two partially overlapping maize association panels comprising 1,014 unique genotypes grown in field trials across at least seven US states and scored for 162 distinct trait datasets enabled the identification of of 2,154 suggestive marker-trait associations and 697 confident associations in the maize genome using a resampling-based genome-wide association strategy. The precision of individual marker-trait associations was estimated to be three genes based a reference set of genes with known phenotypes. Examples were observed of both genetic loci associated with variation in diverse traits (e.g. above-ground and below-ground traits), as well as individual loci associated with the same or similar traits across diverse environments. Many significant signals are located near genes whose functions were previously entirely unknown or estimated purely via functional data on homologs. This study demonstrates the potential of mining community association panel data using new higher density genetic marker sets combined with resampling-based genome-wide association tests to develop testable hypotheses about gene functions, identify potential pleiotropic effects of natural genetic variants, and study genotype by environment interaction.

Key words: Quantitative Genetics; Community Association Populations; Pleiotropy; Maize

Introduction

Association mapping, initially on a gene-by-gene level and later at a genome-wide scale, has been widely adopted as a tool to identify natural genetic variants controlling variation in both quantitative and qualitative traits. In the plant genetics community, logistical and scientific constraints have driven the development and widespread adoption of community association panels comprising sets of distinct plant genotypes which can be propagated and shared, whether through the use of homozygous inbred lines or clonal propagation. In the earlier era where association mapping was conducted on a gene-by-gene level, the use of community association panels allowed the work of estimating population structure within the population to be conducted once rather than for each independent study. In the later era of genome-wide association studies, the use of community association panels again provided substantial practical benefits: the time-consuming and expensive process of genotyping hundreds of thousands or millions of genetic markers across hundreds of individuals had to be undertaken only once to enable an effectively infinite number of studies on the genes controlling different traits by different research groups.

The use of community association panels by many independent research groups to investigate diverse research questions results in data on a wide range of individual traits for genetically identical individuals across one or more environments. This provides significant opportunities to investigate both pleiotropy, the effect of a single genetic locus on multiple phenotypes, and genotype by environment interactions, where the same allele influences the same phenotype in different ways in different environments. In addition, associations between a given genetic locus and a particular trait which are only marginally significant in several individual studies can often be assigned a higher degree of confidence when the same association is identified across multiple studies. Finally, marginal statistical signals from genome-wide association studies can assist in the interpretation of later mutant mapping, gene expression, or selection scans, but only if the initial studies results are made available in a method which is easy to capture and cross reference.

In maize, an early widely adopted community association panel was the Maize Association Panel (MAP), also referred to variously as the maize 282 panel and the Buckler-Goodman Association panel (Table S1). MAP initially consisted of 302 diverse inbreds estimated to represent 80% of the genetic diversity within maize, although several of these were dropped in later years based on poor seed increasability or other factors [1, 2]. Slow decreases in population size over time are a common feature of many community association panels. Two challenges were observed with the initial maize association panel. Firstly, while the MAP panel captured a large proportion of total maize genetic diversity, it consisted primarily of older public sector lines with limited representation of current temperate elite germplasm. Secondly, many of the MAP lines were difficult to grow and increase in the northern US Corn belt. As a result, two additional panels were generated #1: 1) The Shoot Apical Meristem association panel (SAM panel) which included many of the MAP genotypes augmented with expired Plant Variety Patent lines generated by the major seed companies in the USA [3], and 2) The Wisconsin Diversity Panel (WiDiv) developed by selecting non-redundant and diverse genotypes which were able to complete their life cycle and produce significant amounts of seed when grown in Madison, Wisconsin [4]. #1

As sequencing technologies have improved, new sets of genetic markers have been deployed for existing community association panels with increasing degrees of genetic resolution. The MAP was initially genotyped with a modest number of (<100) SSR markers to estimate population structure as a potential confounder for single gene association tests [2]. The WiDiv panel was initially genotyped with 1,536 microarray-based markers [4]. The SAM panel was initially genotyped using sequencing of mRNA samples from each line, enabling the identification and scoring of 1.2M segregating SNP markers [3]. A new high-density marker set for the WiDiv panel was also generated by sequencing mRNA samples, providing a set of 900k segregating genetic markers in this population [5, 6]. The original MAP population which shares many genotypes with both the SAM and WiDiv panels was resequenced as part of the Maize HapMap3 project increasing the number of segregating genetic markers to 83M [7]. A subset of lines from the WiDiv panel were resequenced, resulting in a set of 3.1M SNPs scored across 511 genotypes [8].

#1

Results

Properties of widely studied maize association panels

Three maize association panels were identified in the literature: the Maize Association Panel (MAP) [2], the Shoot Apical Meristem (SAM) panel (369 lines) [3] and the Wisconsin Diversity (WiDiv) panel consisting of either 627 or 942 lines [4, 6]. The latter two populations are largely supersets of MAP, with SAM excluding 10 MAP lines present in the WiDiv panel and WiDiv excluding 67 lines retained in both the SAM and MAP populations (Figure 1A). The SAM and WiDiv populations also share 95 lines not present in the MAP population that served as a partial progenitor for both. These are predominantly more recently released lines developed in the private plant breeding sector and released when the associated plant variety parents expired (expired Plant Variety Patents or exPVPs). The total overlap between the SAM and WiDiv populations was 297, sufficient to enable joint analyses of trait datasets collected in these two populations. The union of the SAM and WiDiv populations included 1,014 unique maize genotypes with both genetic marker information at least one phenotypic record. An additional 35 unique maize genotypes where included in one or both populations, had at least one source of genetic marker data but no phenotypic records and so were excluded from downstream analyses.

The set of approximately 200 papers citing either the SAM panel [3] or either iteration of the WiDiv panel [4, 6] were screened to identify studies which conducted GWAS and published trait datasets collected from one or both of these populations. A total of 21 papers were identified which included GWAS results generated from these populations. After excluding studies where we were unable to locate trait data for individual maize lines, excluding traits which failed initial QC, and condensing studies which utilized previously published data, 132 unique trait datasets drawn from 15 separate published studies remained. This included 55 trait datasets collected from the SAM panel, 66 trait datasets collected from the WiDiv panel, and 11 trait datasets collected from an even larger population of 2,815 maize lines with substantial overlap with these two populations [9]. An additional 30 phenotypes scored in Lincoln, Nebraska in 2020 were included for a final set of 162 trait datasets employed for downstream analyses (Table S3). Individual trait datasets included data values for between 222 and 817 maize lines (Figure 1B) and were collected from field or controlled environment studies conducted in seven states (Figure 1C). Measurements related to inflorescence architecture were the most abundant category among these 162 trait datasets (Figure 1D). SNP-based estimates of narrow-sense heritability for individual trait datasets were variable, with a median value of 0.527 and mean value of 0.523 across all traits. Traits related to flowering time (e.g. timing of

Compiled on: July 13, 2022.
Draft manuscript prepared by the author.

Table 1. Studies from which maize trait datasets were drawn

Reference	Study Type	Phenotypes Scored[a]	Accessions Evaluated[b]	Panel
Peiffer *et al.* 2014[16]	Reproductive & Vegetative	11	737	Ames Panel
Hirsch *et al.* 2014[5]	Reproductive & Vegetative	3	427	WiDiv-503
Leiboff *et al.* 2015[3]	Agronomic, Cellular/Biochemical, & Vegetative	9	378	SAM
Lin *et al.* 2017[17]	Cellular/Biochemical, Root, & Vegetative	16	363	SAM
Gustafson *et al.* 2018 [18]	Disease	7	447	WiDiv-503
Gage *et al.* 2018[19]	Reproductive	16	817	WiDiv-942
Mazaheri *et al.* 2019[6]	Cellular/Biochemical & Vegetative	5	788	WiDiv-942
Qiao *et al.* 2019[20]	Cellular/Biochemical	4	429	WiDiv-503
Sekhon *et al.* 2019[21]	Agronomic	3	364	WiDiv-503
Zheng *et al.* 2019[22]	Agronomic, Root	13	359	SAM
Azodi *et al.* 2020[23]	Reproductive & Vegetative	3	388	WiDiv-503
Lin *et al.* 2020[24]	Cellular/Biochemical & Reproductive	8	439	WiDiv-503
Renk *et al.* 2021[25]	Seed Composition	16	499	WiDiv-503
Schneider *et al.* 2021[26]	Root	1	599	WiDiv-503
Zhou *et al.* 2021[27]	Reproductive	17	339	SAM
Sun *et al.* 2021[11]	Disease	1	687	WiDiv-942
Previously Unpublished	Agronomic, Disease, Reproductive, & Vegetative	29	752	WiDiv-942

*Note: "WiDiv": Wisconsin Diversity Panel, "SAM": Shoot Apical Meristem Panel, "UNL": University of Nebrasks-Lincoln, "a": Phenotypes used in this study from the phenotypes scored in respective studies (after removing exact same phenotype values if used in another study).
*Note: The highest number of accessions with phenotype data used in this study from the respective publication.

anthesis, timing of silking, or anthesis-silking-interval) was the category which exhibited the highest median heritability of 0.762 (Figure 1E, and Table S3). Flowering time traits collected in different environments were correlated with each other and also exhibited notable correlations with a subset of both above-ground vegetative and below-ground root related traits (Figure 1F).

The total number of unique maize line names observed across these 162 trait datasets was 1,118 which was modestly more than the set of 1,014 unique genotypes present across the WiDiv and SAM mapping populations. We speculate that this difference may result from the inclusion of local checks or lines-of-interest or changes in naming convention or transcription errors which we were unable to resolve. For the 1,014 unique genotypes named as part of the SAM population [3] or WiDiv population [6], raw whole-genome sequencing or RNA-seq sequence data was aggregated from a number of sources (Table S2) [7, 10, 6] as described in Sun *et al.* 2021[11]. Alignment of published sequence data from these sources to the maize reference genome (B73_RefGen_V4) [12, 13], scoring of *a priori* segregating SNPs from HapMap3 [7], imputation, and filtering resulted in a set of 17,717,568 with minor allele frequency >0.01 and heterozygosity rate of <0.1, leading to an average of one SNP per 120 bp (see Methods).

The 17,717,568 polymorphic markers chosen for downstream analysis were distributed roughly evenly across the ten chromosomes of maize, with local reductions in SNP density around centromeres/pericentromeric regions of each chromosome (Figure S3A). Rare SNPs with minor allele frequencies <0.1 were modestly more abundant than common SNPs (Figure 2A). Linkage disequilibrium decayed rapidly, with the average r^2 between two SNPs separated by 10 kilobases being approximately 0.18 (Figure 2B) similar to previous reports [14, 15]. The first three principal components of variation explained approximately 10% of total variance among genotypes (Figure S3B). Principal coordinate (PCo) analysis using this SNP set separated lines with known assignments to major heterotic groups (Figure 2C and D). The same set of PCo analyses did not identify obvious biases in the distribution of lines present in different association panels (Figure S3C).

Unified marker-trait analyses

Genome-wide association studies conducted using FarmCPU with the 162 traits and about 18M markers described in the previous section and employing a RMIP cutoff of 5 for a suggestive association identified 2,154 signals across 151 traits (Figure 3A). Among traits with one or more suggestively significant hits, the median number of hits was 12 (mean 12.57), the maximum was 33 and the minimum was 1.

Consolidation of 2,154 SNPs with at least suggestive statistical significant associations with phenotypes ($\geq$5 RMIP) into distinct peaks based on physical distance and LD (See Methods) reduced the number of associations to 1,466 peaks distributed across phenotypes assigned to eight categories (Table 2). Of these 1,466 peaks, 161 peaks were associated with 11 agronomic traits, 92 peaks were associated with 17 of the 21 total cellular/biochemical traits, 72 peaks are associated with eight disease traits, 176 peaks are associated with 15 flowering time traits, 459 peaks were associated with 41 of 47 total inflorescence traits, 113 peaks were associated with 15 root traits, 128 with 16 seed composition traits and 295 with 28 of 29 total vegetative traits (Figure 3B, and Table 2).

A wide range of approaches are employed in the literature to define the set of annotated gene models adjacent to a significant GWAS peak which should be labeled as "candidate genes". These can include both fixed windows around the peak, examining an arbitrary number of the closest annotated gene models to the peak, or adaptive windows defined based on local levels of linkage disequilibrium or haplotype blocks. To assess the precision provided by the peaks identified in this study, we utilized a set of 604 gene models recorded in the MaizeGDB database[28] as associated with one or more phenotypes. These gene models constituted 1.5% of the total set of 39,498 annotated gene models present on the B73_v4 reference genome [13]. The first three genes closest to GWAS peaks identified above were more likely to be associated with reports of phenotypes in MaizeGDB than the expected background rate, and this pattern became stronger at more stringent RMIP cutoffs (Figure 4A). When employing physical distance rather than rank order, the greatest enrichment of genes with reported phenotypes in the MaizeGDB database was observed in the categories "within gene" "closer than 10 kilobases" and "10-40 kilobases", although noticeable enrichment remained observable at greater distances from the GWAS peak (Figure 4B).

Flowering time trait datasets tended to identify a disproportionately high number of independent GWAS peaks (Figure 3C), potentially as a result of the greater proportion of variance among these traits explained by genetic factors (Figure 1E). Overall, in 1,252 cases (85.4%) a peak was identified only in the analysis of a single trait dataset (Table 2). The remaining 214 peaks were identified

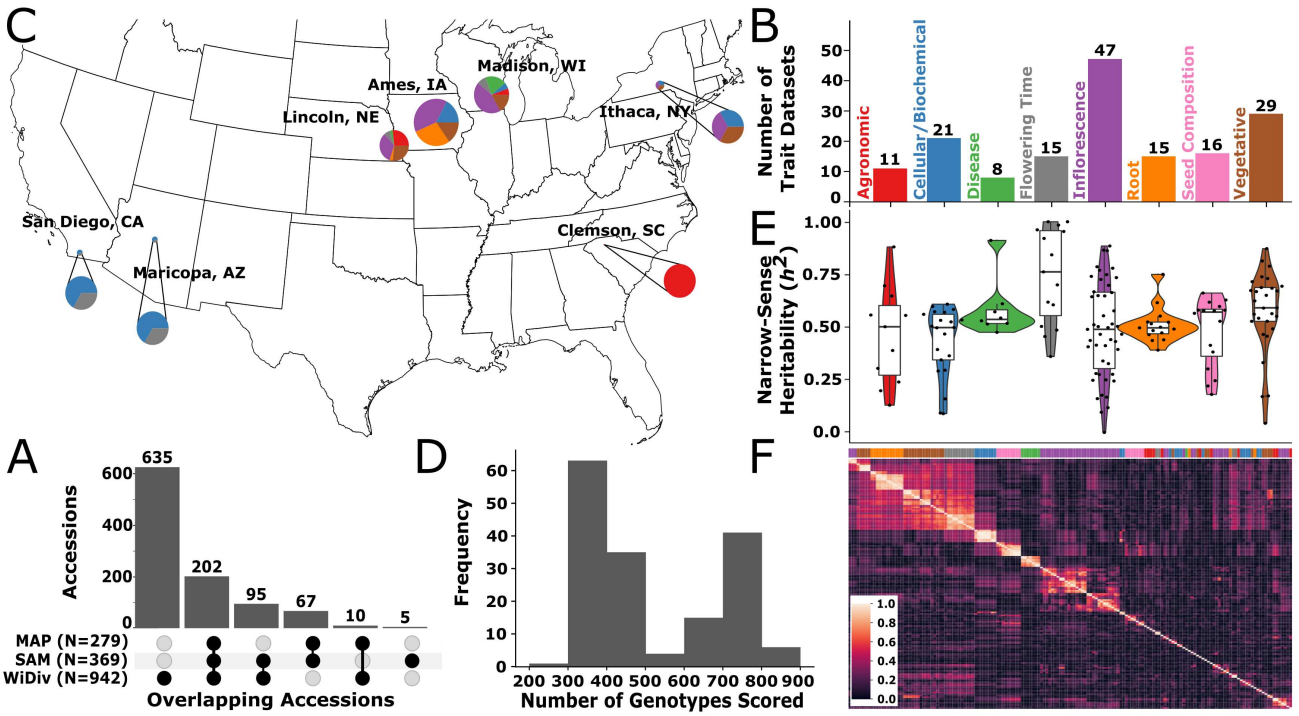

Figure 1. Characteristics of Maize Association Panel trait datasets. A) Number of accessions which are represented in any of the three diversity panels. B) Representation of eight broad phenotypic categories among the 162 traits collected here. Category assignments for individual traits are provided in Table S3. C) Geographic distribution of trials where trait datasets were collected. Size of circles indicates number of traits collected at a specific geographic location. Colors of circles indicate types of trait datasets collected at that location. Labels for which colors correspond to which types of traits are given in Panel B. D) Distribution of the number of genotypes scored for a given trait. E) Distributions of narrow-sense heritability values, across the same eight broad phenotypic categories shown in panel B. Colors corresponding to the color key for phenotype classes are provided in Panel B. F) Correlations among the 162 trait datasets analyzed in this study. Trait datasets are clustered based upon absolute spearman correlation value. Phenotype classes are indicated with color bar on top x-axis with colors corresponding to the color key for phenotype classes provided in Panel B.

Table 2. Summary of unique associations with RMIP $\geq$ 5 within each of the eight phenotypic groups analyzed.

Phenotype Group	# of Phenotypes analyzed	# of Phenotypes with hits	# of peaks	# of single trait peaks	# of Multi Trait Peaks	# of Multi Trait Peaks within each Category [a]	# of Peaks Associated Across Category
Agronomic	11	11	161	155	6	4	2
Cellular/Biochemical	21	17	92	69	23	20	3
Disease	8	8	72	44	28	28	0
Flowering Time	15	15	176	128	48	32	16
Inflorescence	47	41	459	420	39	32	7
Root	15	15	113	81	32	25	7
Seed Composition	16	16	128	108	20	19	1
Vegetative	29	28	295	247	48	28	20
Total Unique	162	151	1466*	1252	214*	188	26*

[a] Excluding 26 peaks that overlap between two or more phenotype groups/categories. Of these 26 peaks 22 peaks are associated with traits belonging to two phenotype categories and 4 peaks are associated with phenotype traits belonging to three phenotype categories.

* The value is less than the total of all the values in respective columns because of peaks associated with phenotypes in multiple categories are depicted in each category they show significance.

in analyses of two or more separate trait datasets. In 188 cases the same peak was identified in the analysis of two or more phenotypes belonging to the same general category. For example, a peak consisting of four SNPs in high LD with each other on chromosome 6 spanning from 108,211,603 to 108,213,234 bp with the single highest RMIP SNP located at 108,212,338 bp was identified in analysis of both kernel starch abundance (Starch_K) (RMIP=52) and kernel fat abundance (Fat_K) (RMIP=23) within the overall category of "seed composition" traits [25]. The peak spans the 5' end of the gene model Zm00001d036982 (108,212,462 to 108,219,350 on chromosome 6) which encodes *DGAT1-2* (Diacylglycerol O-acyltransferase 1-2)/ *ln1* (linoleic acid1). *DGAT1-2* substantially increases the seed oil and oleic-acid contents [29]. Largely, oils are stored in the form of triacylglycerol (TAG) and DGAT catalyzes the final step of TAG biosynthesis by transferring an acyl group from acyl-CoA to the sn-3 position of 1,2-diacylglycerol (DAG) thus, acting as the rate-limiting enzymes for TAG biosynthesis [30, 31] (Figure 5A&B). The rarer allele ("T") is associated with an increase in seed fat and decreases in seed starch (Figure 5C). The starch-promoting allele was more abundant in iodent subpopulations and less abundant in sweet corn subpopulations (Figure S4). The original analysis of these two datasets employed the FarmCPU algorithm, but without resampling [25] and did not identify these two associations, consistent with observed RMIP values which suggest only a one in two chance of detecting the DGAT/starch association and only a one in four chance of detecting the DGAT/fat association in a single round of GWAS.

In the remaining 26 cases where the same genomic interval was identified in the analysis of multiple trait datasets (Figure 6A-D, and S5-S26), the trait datasets involved spanned two or more categories, with 22 peaks associated with trait datasets spanning two categories and four peaks associated with trait datasets spanning

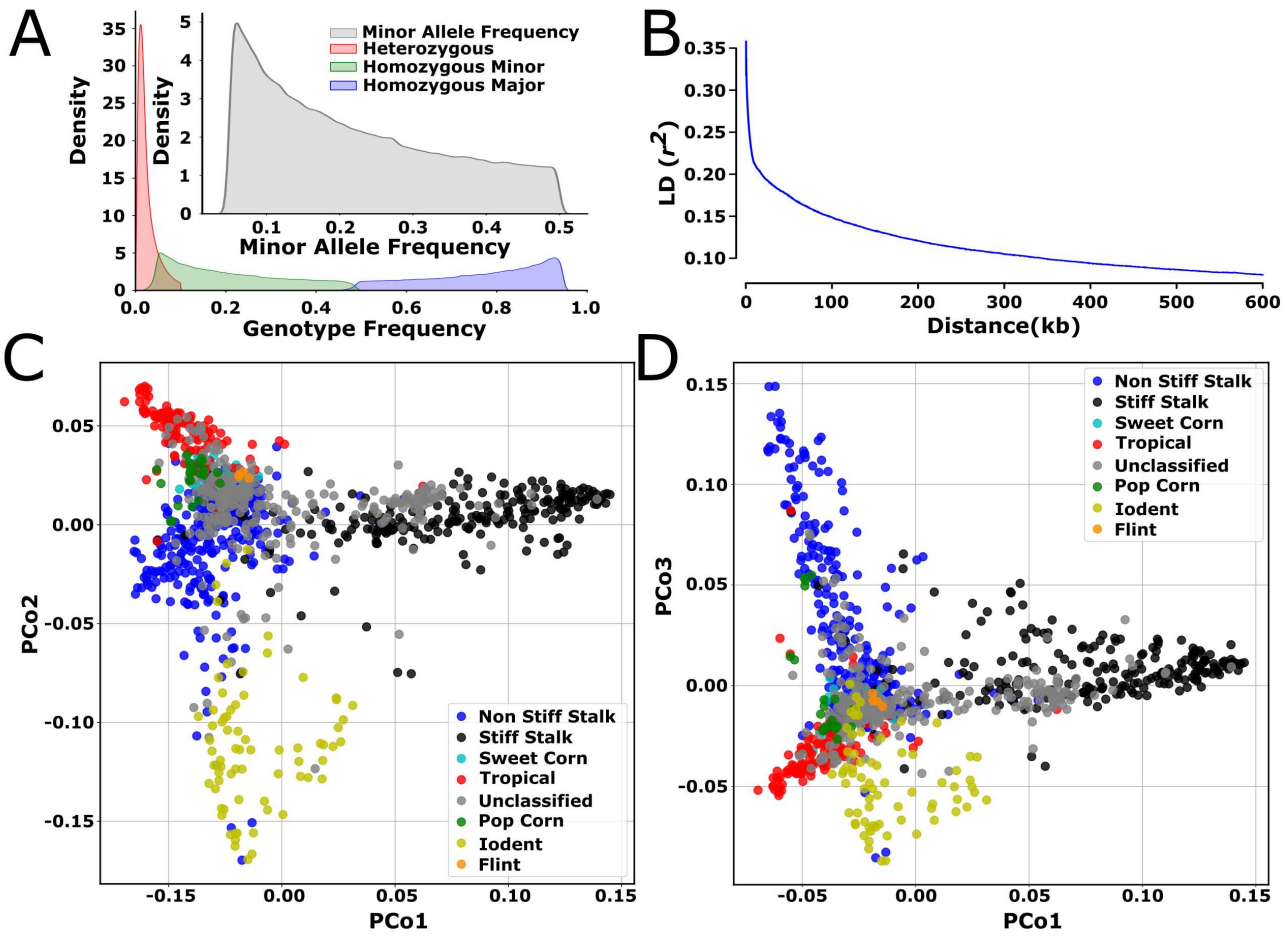

Figure 2. Characteristics of Maize Association Panel Marker datasets. A) Genotype frequency and minor allele frequency of the marker dataset. B) The genome-wide LD decay with maximum distance of 600 kilobases between two SNPs. C) Genetic relationship among the accessions used in this study and visualized using multidimensional scaling/principal coordinate analysis of the distance matrix. The X- and Y-axis represent first and the second principal component coordinates. Each point is color coded by the heterotic group each accession belongs to. D) Genetic relationship among the accessions used in this study and visualized using multidimensional scaling/principal coordinate analysis of the distance matrix. The X- and Y-axis represent first and the third principal component coordinates. Each point is color coded by the heterotic group each accession belongs to.

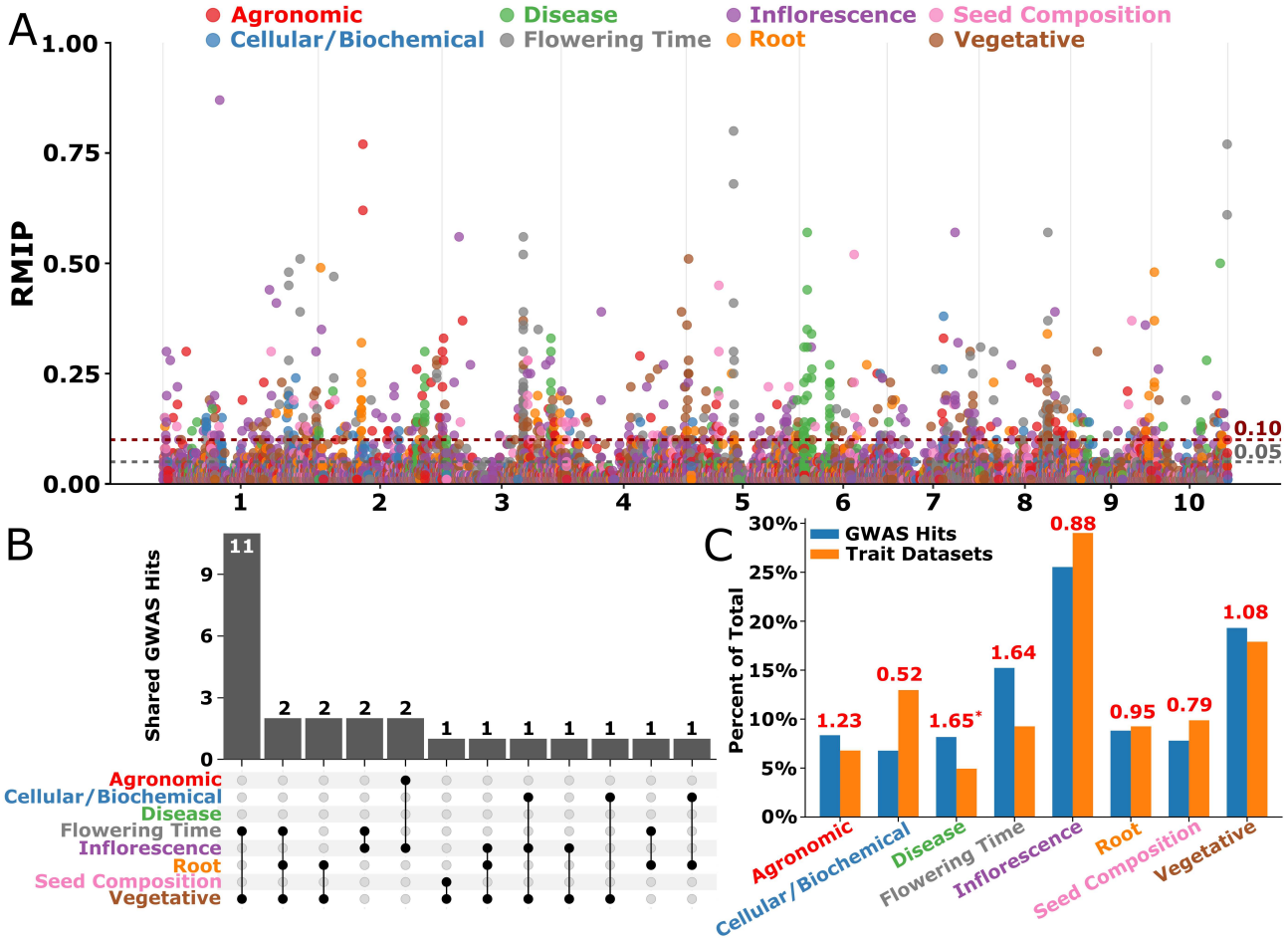

Figure 3. GWAS Summary: Multi-trait peaks detected across phenotypic categories. A) Combined Manhattan plot for GWAS using all 1014 individuals screened using 18M markers. Dashed grey and red lines indicates the cutoff of 5% and 10% for statistical significance calculated based on RMIP value. Each chromosome is shown in the X-axis. The Y-axis is the resampling model inclusion probability (RMIP) values ranging from 0 to 1. B) An upset plot showing number of shared GWAS hits between various phenotypic categories. C) Percent representation of GWAS hits for the number of trait datasets analyzed. Number on top of each pair of bars in each phenotypic category corresponds to the ratio of GWAS hits:number of trait datasets analyzed in each category. *Note: The ratio was higher for the disease traits but the traits in this category are essentially the same trait analyzed at different timepoints in a time series manner, thus most of the hits overlap among the traits leading to an inflated ratio.

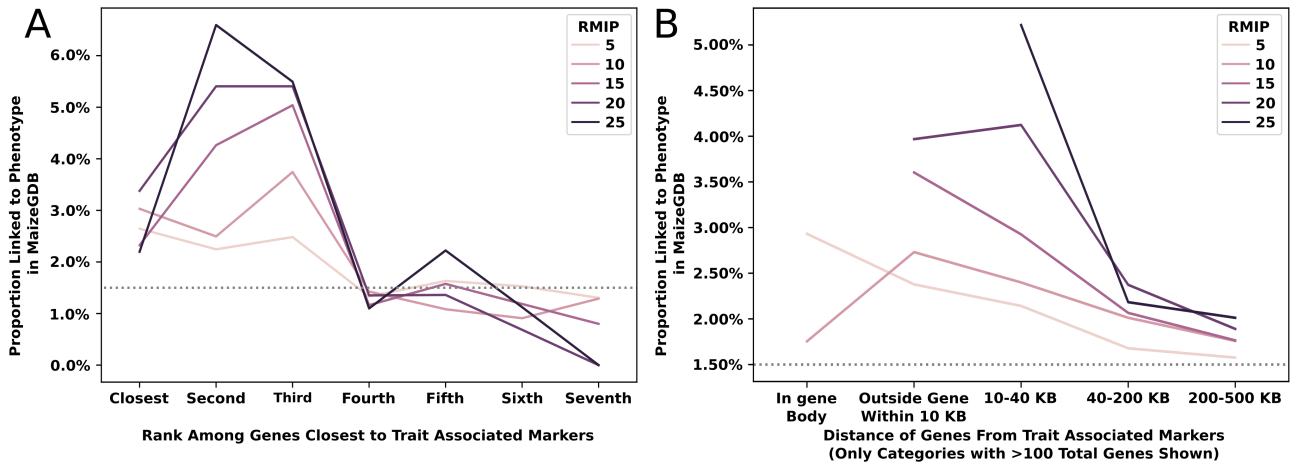

Figure 4. Probability of genes at different distances from peak SNP from GWAS are linked to phenotypes. A) Gene positions of unique trait associations. First seven genes closest to the GWAS peaks were selected and shown on X-axis. B) Gene order of unique trait associations. The distance of the genes from the trait associated markers are shown on X-axis.

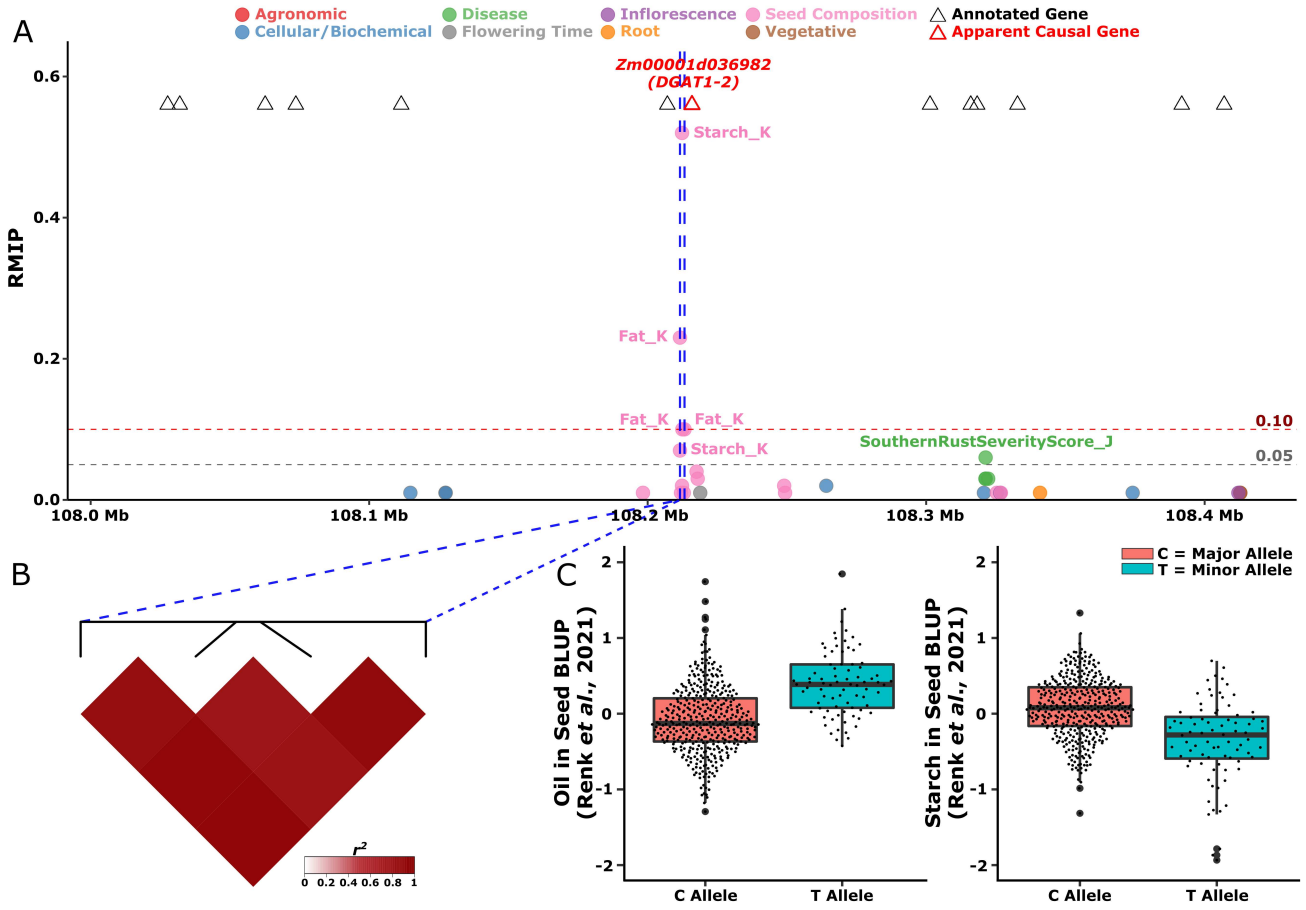

Figure 5. Combined GWAS identifies peak associated with seed starch and fat. A) View of resampling marker inclusion probability values for markers in a window from 108,211,603-108,213,234 on chromosome 6 spanning 200 kilobases upstream and downstream of the pleiotropic peak identified for seed starch and oil content. Only markers with resampling marker inclusion probability values $\geq$ 0.01 are shown. B) The LD relationships between the significant SNPs within the peak. C) Distributions of observed oil and starch content values reported in [25] for lines carrying either allele of the peak SNP located at position 108,212,338 bp.

three categories (Figure 3B). Genomic intervals associated with flowering time were disproportionately more likely to be associated with phenotypes from at least one other category. Sixteen of 176 unique peaks identified for flowering time were also associated with one or more phenotypes from other categories (9%), while only 10 of 1,290 unique peaks (0.8%) identified for non-flowering time traits were associated with traits from two or more of the remaining seven categories (Figure 3B).

An illustrative example of the potential for genes influencing flowering time to be identified in genome-wide association studies for other traits is the case of *ZmMADS69*. *ZmMADS69* (syn *Zmm22*) (Zm00001d042315) is a MADS-box transcription factor located between 160,564,021 bp and 160,591,933 bp on maize chromosome 3, which has been shown to function as flowering activator, with a derived allele conferring earlier flowering in many maize lines relative to its wild progenitor teosinte [32, 6]. A peak consisting of 26 SNPs in high LD with each other was consistently identified for multiple flowering time related traits including seven measurements of anthesis (male flowering) in different environments (Anthesis_A, Anthesis_G Anthesis1_L, Anthesis4_H, Anthesis6_H, Anthesis7_H, Anthesis_J) and three measures of silking in different environments (Silking_A, Silking_L, Silking_J). The same peak was also identified in the analysis of multiple vegetative traits including measurements of plant height in two environments (PlantHeight_D, PlantHeight_G), extant leaf number, stalk diameter, and biomass yield (Figure 6A). *ZmMADS69* has been shown to downregulate the expression of *ZmRap2.7*, which relieves repression of the florigen gene *ZCN8*, causing/resulting in early flowering [32]. Both *ZmRap2.7* and *ZCN8* are located on chromosome 8 [33, 34, 35] and both of these genes are also associated with GWAS peaks. A peak on chromosome 8 consisting of four SNPs between 126,884,534 bp to 126,891,234 bp was separated by only two kilobases from the gene model encoding *ZCN8* (Zm00001d010752, located between 126,880,531 and 126,882,389 bp) and was associated with anthesis in three environments (Anthesis_G, Anthesis7_H, Anthesis_J). *ZmRap2.7* (Zm00001d010987) is located approximately 10 megabases away from *ZCN8* on chromosome 8 (between 136,009,216 and 136,012,084 bp) and is associated with a peak consisting of 13 SNPs that was detected for seven measurements of anthesis (Anthesis_A, Anthesis_G, Anthesis1_L, Anthesis5_H, Anthesis6_H, Anthesis7_H, Anthesis_J), two approaches to measuring silking in the same environment (Silking_L, and SilkingGDD_L), and a number of vegetative traits including extant leaf number (ExtantLeafNumber1_J), leaf width (LeafWidth_J), and plant height (PlantHeight_D).

In addition to the three peaks discussed above, thirteen other peaks were also associated with both flowering time traits and traits from other categories (Figure 3.A, Figure 3.B, and Table S4). In two cases a significant signal for flowering time was co-located with significant signals for inflorescence architecture traits. The first, located on chromosome 1 between 102,077,749 and 102,120,437 bp, consists of two SNPs in high LD with each other and is significantly associated with both date of silking (Silking_J) and ear length (EarLength_O) in separate environments (Table S4, and Figure S5). The second, located on chromosome 4 between 78,020,118 and 78,451,569 bp, consists of two SNPs and showed significant associations with both male flowering in one environment (Anthesis5_H) and the length of the central spike of the tassel in another environment (SpikeLength1_C) (Table S4, and Figure S14). In eleven cases a significant association for flowering time was co-located in the genome with a signal for an above-ground vegetative trait dataset. These were typically vegetative traits with known links to flowering time including leaf/node number, and plant or ear height (Table S4, and Figure 6D, S6, S12, S16 - S20, S22, S24, and S26).

The potential pleiotropy of genes linked to flowering time was not confined to above-ground traits. In three cases a significant signal for flowering time was also associated with one or more datasets describing variation in root phenotypes. A signal on chromosome 5 between 94,710,702 and 94,712,951 bp was associated with flowering time across a wide range of environments (Anthesis_A, Anthesis1_L, Anthesis7_H, Anthesis_J, Silking_J, Silking_L, SilkingGDD_L), with other above-ground vegetative traits (LeafAreaIndex_J, LeafLength_J) and with many root architecture traits (RootArea1_O, RootArea2_O, RootArea4_O, RootWidth4_O) (Table S4 and Figure S19). The specific SNPs which define the peak are all located within Zm00001d015513 which encodes a cinnamoyl-CoA reductase expressed primarily in leaves and leaf meristems [36]. A signal on chromosome 8 between 28,727,658 and 28,769,198 bp, was associated with both male and female flowering time in Nebraska (Anthesis_J, Silking_J) and variation in root depth in Iowa (RootDepth1_O, RootDepth2_O) (Table S4 and Figure S22). The last of the three signals, associated with flowering time (Anthesis_A, Anthesis_G, Anthesis4_H, Anthesis7_H, Anthesis_J, Silking_A, Silking_J), leaf number, and root (RootArea1_O, RootArea2_O, RootArea4_O, RootWidth3_O) traits is also located on chromosome 8, between 134,706,389 to 134,759,977 bp. This 54 kilobase interval is entirely free of annotated genes, but ends 600 bp upstream of classical mutant *liguleless4* (Zm00001d010948) (Figure 6D). *Liguleless4* (synonym *knox11*) encodes a knox transcription factor which is highly expressed in the SAM, seed radicle, internode tissues, crown roots, pericarp of seed and the endosperm of maize [36]. A dominant allele of *liguleless4* abolishes the ligule and alters the sheath-blade boundary in maize leaves [37] likely via ectopic expression [38], however phenotype of loss of function alleles, if any, remains uncharacterized.

Discussion

The widespread adoption of diverse association panels in plant biology has enabled a wide range of research and discovery by researchers working on diverse phenotypes, species, and research questions [39]. Beyond lists of specific candidate genes identified in the main text or supplemental figures, the reuse of GWAS results can sometimes be challenging. Changes in genome versions, gene model annotations, or genetic marker datasets, as well as changes in best practices and algorithms for conducting genome-wide association tests can all hinder comparisons with and/or reuse of previously published GWAS results. In the field of human genetics where the release of individual-level trait and genetic data could raise privacy and ethics concerns, the field has converged on a standard of releasing summary statistics but not individual-level trait and genetic data [40]. Working with plant data, privacy concerns do not typically preclude the release of individual-level data, and the reuse of the same genotypes across independent studies increases the potential value of individual-level data.

We identified 21 published papers describing phenotypes or GWAS conducted using one or more of two widely adopted association panels in maize [4, 3]. In 18 cases (85%) it was possible to locate raw trait values for individual lines used in the published analyses, including the 16 studies summarized in Table 1 and the two additional papers not included in our analyses. The high frequency with which trait data is being released for published GWAS studies is encouraging as it indicates the maize quantitative genetics community is adopting similar norms and practices to the genomics community which has long been a leader in promoting strong best practices for raw data deposition and dissemination [41]. Unlike the genomics community, the plant quantitative genetics community does not have access to widely used and standardized data repositories. Challenges in integration of these data included inconsistent naming, extra lines that were not part of panel, repeated traits across papers, and data distributed across supplemental files or Figshare. Metadata for how and when individual traits were collected typically was present but often needed to be manually extracted from reading the manuscript text. Information that would further increase the value of released trait data such as the GPS co-

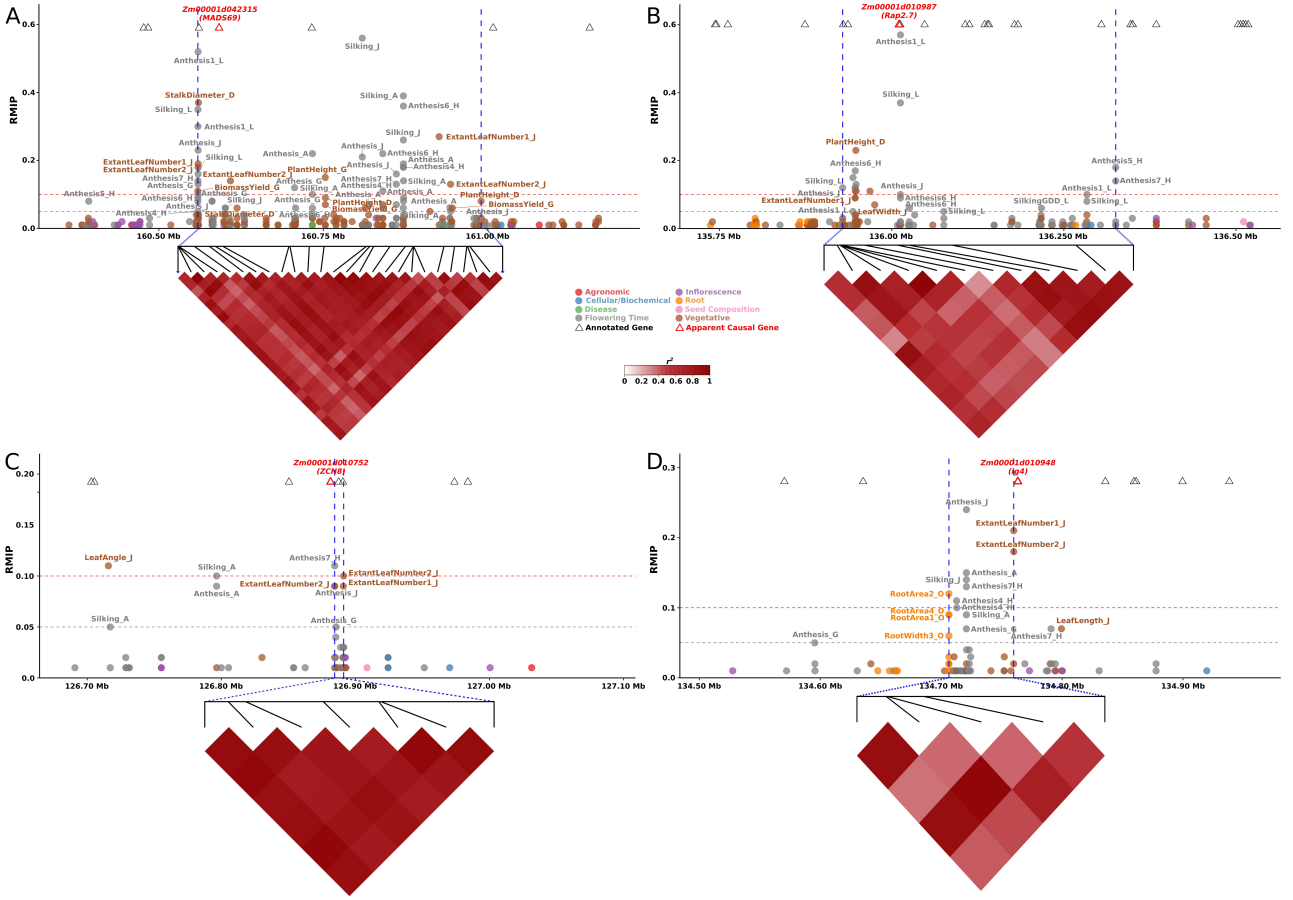

Figure 6. GWAS peaks associated with multiple traits. A) Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 3 from 160,559,294 to 160,989,691 bp. This peak is associated with *MADS69* (Zm00001d042315). The phenotypes associated with this peak belongs to Flowering Time and Vegetative categories. The phenotypes associated with this peak are Anthesis1_L, Anthesis4_H, Anthesis6_H, Anthesis7_H, Anthesis_A, Anthesis_G, Anthesis_J, BiomassYield_G, ExtantLeafNumber1_J, ExtantLeafNumber2_J, PlantHeight_D, PlantHeight_G, Silking_A, Silking_J, Silking_L, and StalkDiameter_D. The vertical dashed lines shows the peak boundary. B) Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 135,928,821 to 136,325,345 bp. This peak is associated with *Rap2.7* (Zm00001d010987). The phenotypes associated with this peak belongs to Flowering Time and Vegetative categories. The phenotypes associated with this peak are Anthesis1_L, Anthesis5_H, Anthesis6_H, Anthesis7_H, Anthesis_A, Anthesis_G, Anthesis_J, ExtantLeafNumber1_J, LeafWidth_J, PlantHeight_D, SilkingGDD_L, Silking_L. The vertical dashed lines shows the peak boundary. C) Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 126,884,534 to 126,891,234 bp. This peak is associated with *ZCN8* (Zm00001d010752). The phenotypes associated with this peak belongs to Flowering Time and Vegetative categories. The phenotypes associated with this peak are Anthesis7_H, Anthesis_G, Anthesis_J, ExtantLeafNumber1_J, ExtantLeafNumber2_J. The vertical dashed lines shows the peak boundary. D) Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 134,706,389 to 134,759,977 bp. This peak is associated with *lg4* (Zm00001d010948). The phenotypes associated with this peak belongs to Flowering Time, Root and Vegetative categories. The phenotypes associated with this peak are Anthesis4_H, Anthesis7_H, Anthesis_A, Anthesis_G, Anthesis_J, ExtantLeafNumber1_J, ExtantLeafNumber2_J, RootArea1_O, RootArea2_O, RootArea4_O, RootWidth3_O, Silking_A, Silking_J. The vertical dashed lines shows the peak boundary.

ordinates and planting dates of individual field trials was provided in some cases but not others. The identification of a single common repository, standards for metadata on individual field trials, and for the preservation of a single unique identifier for each genotype included in a community association population would all lower barriers to the reuse of trait datasets. However, despite these current challenges both the overall consistency and quality of data release and documentation was exceptional, enabling the investigation of multi-environment and multi-trait genetic associations.

One desirable outcome of having access to raw trait data is that it enables reanalysis of existing trait datasets, each of which represents a substantial investment of both finance resources and human effort/suffering, as new higher resolution genetic marker datasets and new analysis algorithms become available. Here we employed a RMIP-based filter to the FarmCPU GWAS algorithm [42, 43, 44] and were able to identify 2,154 suggestive associations (RMIP$\geq$5) and 697 confident associations (RMIP$\geq$10) across 162 traits collected in 33 environments spanning at least seven states. These signals included new associations identified as a result of either new genetic marker data and/or the FarmCPU/RMIP approach (Figure 5). Overall 1,466 and 468 unique sites in the genome tagged with a suggestive or confident association, respectively. These associations were enriched near genes with previously reported phenotypic effects (Figure 4). However, many of these signals are located near genes whose functions were previously entirely unknown or estimated purely via functional data on homologs which is typically useful for inferring molecular function of a protein encoded by a given gene but can produce misleading information on the specific biological processes a given gene is involved in [45].#1 These new layers of functional data will be most useful if they can be integrated into community genomics repositories. In this case the functional data generated as part of this project have been integrated as browser tracks and downloads at the maize community repository MaizeGDB [28] to enable maize researchers to quickly access these data, cross reference them with other data types, and compare with mutant or QTL mapping results.

#1

The significant signals identified for flowering time and both above-ground and below-ground plant architectural traits adjacent to *liguleless4*/*knox11* are an example of an intermediate case between confirmation of known functions and assigning potential functions to previously uncharacterized genes or regions of the genome. *Liguleless4* belongs to the knox class-I gene family [38] a family of genes involved in regulating plant development via expression in apical meristems [46]. The *liguleless4* gene itself was identified via a dominant allele *Lg4-O* which alters development at the leaf blade/sheath boundary [37] and associated with ectopic expression [38]. Loss of function alleles of the *liguleless3* gene, a paralog of *liguleless4*, do not exhibit any obvious phenotype [38]. A role for *liguleless4* in determining flowering time and above/below-ground plant architecture is consistent with the reported expression pattern of the wild type allele in tissues including the shoot apex, root tips, and developing inflorescence [38]. #1

One striking observation from the colocalization of association signals across multiple trait datasets was how common the re-identification of shared signals was. 14.5% (214/1466) of all suggestive associations and 16.2% (76/468) of all confident associations were identified in at least two trait datasets. One utility of reanalyzing published trait datasets is that variants with consistent but moderate effects across many studies can be distinguished from, and assigned higher confidence, than signals of equivalent statistical significance which are identified in only a single study in a single environment. Another lesson to take away from the same colocalization data is how common it was for the same locus to be identified for traits belonging to separate categories of phenotypes. The interpretation of a genetic locus with a significant association with root area will be quite different depending on whether than same locus is also associated with flowering time [47] (Figure 6D, S19 and S22). In both cases the key take away is that researchers do not have to analyze or interpret GWAS in a vacuum but instead are able to interpret their results in the context of the rich datasets of previously scored phenotypes and previous GWAS analyses. Our understandings of genetics, genotype by environment interactions, and pleiotropy will all benefit from the broad use of these rich datasets.

Potential implications

Logistical and financial limitations often constrain quantitative genetic analyses of plant populations to collecting data on a single phenotype or a small suite of related phenotypes, limiting our capacity to identify and study the ways individual genetic variants can control multiple phenotypic outcomes. The dataset described in this manuscript, including 162 traits scored across different subsets of 1,014 immortalized maize inbred genotypes and associated with a high density marker set of 18M segregating markers dramatically lowers the barriers for further quantitative genetic studies of both pleiotropy and genotype by environment interactions in maize. In addition, by identifying more than 2,000 confident or suggestion genetic associations in the maize genome, this dataset means researchers do not have to analyze or interpret GWAS in a vacuum but instead can interpret their results in the context of the rich datasets of previously scored phenotypes and previous GWAS analyses.

Materials & Methods

Collection of Published Trait Data:

Papers publishing maize analyses were identified by searching papers citing the initial description of the first iteration of the WiDiv panel[4], the initial description of the SAM diversity panel[3], or the expanded WiDiv panel[6]. Screening of studies citing one or more of these papers concluded on 25th June 2021. Published studies were excluded if we were unable to locate de-anonymized trait values for individual accessions or if less than 200 total accessions were phenotyped. If two studies indicated that the same trait was collected from the same lines in the same location in the same year, only one version of the data was retained. If a study published both data from individual environments and aggregated estimates across environments (e.g. averages, Best Linear Unbiased Predictions (BLUPs), or Best Linear Unbiased Estimates (BLUEs)) only individual environment trait data was retained. If only aggregate estimates across environments were published, aggregated traits were employed. After preliminary analysis with MLM-based GWAS several other trait datasets were discarded when it proved impossible to effectively control false discoveries across the genome. The final data file of all accession-level trait values employed in this study is provided as Table S2.

Trait Data Not Previously Published

A set of 752 maize genotypes, which were a strict subset of the WiDiv panel, and included 254 of 369 genotypes from the SAM diversity panel were evaluated in a field experiment conducted in Lincoln, Nebraska in the summer of 2020. The experimental design of this field experiment has been previously described [11]. Briefly, the field was laid out in a randomized complete block design with two blocks of 840 plots including a repeated check genotype (B97). Each plot was two rows, 7.5 feet (approximately 2.3 meters) long with 30 inch row spacing (approximately 0.76 meters), 4.5 inch spacing between sequential plants (approximately 11.5 centimeters) and 30 inch alleyways between sequential plots (approximately 0.76 meters). The field was planted on May 6th, 2020 and was located at the University of Nebraska-Lincoln's Havelock Farm (40.852 N,

96.616 W).

Tassel architecture phenotypes were collected once tassels had fully emerged for three randomly selected plants per plot, avoiding edge plants. Tassel lengths were measured from the basal primary tassel branch to the tip of the tassel spike. Branch zone length was defined as the length from the basal primary tassel branch to the top primary tassel branch. Tassel spike length was defined as the length from the top primary tassel branch to the tip of the tassel spike. The total number of primary tassel branches were also counted as well as the number of these primary tassel branches that were initiated, but later aborted (Figure S1).

Male and female flowering times for each plot were scored on the first day that 50% of plants has visible pollen shed or visible silks, respectively. Root and stalk lodging were scored at the end of the growing season as a percent of extant plants in the plot, following the published Genomes to Fields phenotyping protocol for both traits [48]. Leaf phenotypes – leaf length, leaf width, and leaf angle – were measured for two plants per plot and collected from each plot after anthesis and silking. One plant was randomly selected from each of the two rows for measurement, avoiding edge plants when possible. Leaf length was measured from the leaf ligule to the leaf tip on the adaxial surface of the first leaf above the top ear of the plant. Leaf width was measured on the same leaf at the midpoint between the ligule and the leaf tip. Extant leaf number was determined by counting the number of visible leaf collars on the same two plants. Plant height was measured between the soil surface and the flag leaf collar using a marked pole. Leaf Area Index values were estimated using a LAI-2200C Plant Canopy analyzer (LI-COR, Inc). For each plot one above canopy and three below canopy measurements were collected using the LAI-2200C's 270-degree view cap. The three below canopy measurements collected diagonally in the space between the two rows of the plot. The first measure adjacent to one row, the second equidistant between the rows and the third adjacent to the second row. Leaf Area Index measurements were collected between July 28th - August 12th, 2020.

All ears were harvested from eight semi-randomly selected plants per plot with edge plants being excluded when possible. Ear length, ear width, length of fill, kernel row number and number of kernels per row were hand measured or hand counted for six ears per plot or all ears when less than six ears were present (Figure S2). The average of individual ears were used to calculate plot level values. All harvested ears were weighed, shelled, and the resulting pooled grain was also weighed. Cob weight was calculated as the difference between ear weight and grain weight. Initial hundred kernel weight was calculated by counting and weighing 100 kernels per plot after shelling and pooling of grain. Grain moisture was measured using a Dickey-John GAC® 2500-AGRI Grain Analysis Computer (Dickey-John® Corporation, Auburn, IL). Total grain weight and hundred kernel weight was recalculated to a standardized 15.5% moisture content. When insufficient grain was harvested to collect accurate grain moisture data using the GAC® 2500-AGRI, a default value of 8.25% moisture, corresponding to the approximate median of all grain moisture values was employed to calculate moisture standardized grain weight and hundred kernel weight. Further, the BLUPs for each phenotype was calculated by fitting a linear mixed model using R package lme4[49] with genotypes fit as random variable for the traits with data from 2020.

Unified Genetic Marker Data

A single set of markers scored across 1,049 accessions were employed for downstream analyses. These genotypes were determined based on marker data aggregated from three published sources: resequencing data for the WiDiv-503 panel (454 individuals) [10], resequencing data generated as part of the HapMap3 project (141 individuals) [7], RNA-seq data for the WiDiv-942 panel (399 individuals) [5, 6]. The specific NCBI SRA ID numbers of the files used to call SNPs for each of the accessions are provided in Table S2.

Both genome resequencing data and RNA-seq data were quality trimmed using Trimmomatic (v0.33) [50]. BWA-MEM (v0.7) with default parameter settings [51] was employed to align the resulting trimmed resequencing data to v4 of the B73 maize reference genome [12, 13]. STAR (v2.7) [52] was used to align the trimmed RNA-seq reads to v4 of the B73 maize reference genome in two rounds as described in Sun et al [53]. Apparent PCR duplicates were marked within the resulting BAM alignments using picard (v2.22) [54]. *A priori* segregating genetic markers identified in maize HapMap3 [7, 11] scored for each individual using GATK toolkit (v5.1) [55]. Missing values were imputed using beagle/5.01 with the HapMap3 population treated as a reference panel and parameters settings: 'window=1 overlap=0.1 ne=1200' [56]. The imputed genetic marker dataset was filtered to remove markers with a minor allele frequency less than 0.01 or proportion of site heterozygous calls greater than 0.1 to produce the final set of 17,717,568 SNP markers.

Quantitative Genetic Analysis of Trait Data

A kinship matrix for the complete set of 1,049 genotyped accessions, including all maize lines included in the SAM or WiDiv panels and 35 additional maize lines for which sequence data was generated and released as part of Mazaheri *et al.* 2019 [6] were calculated using the first method described by VanRaden (2008) [57] as implemented in rMVP (v1.0.5) [58]. Narrow-sense heritability for each trait was calculated using this kinship matrix and the R package sommer (v4.1.1) [59]. Multidimensional scaling or the Principal Coordinate (PCo) Analysis was performed with –mds-plot 2 and –cluster options within plink v1.90 [60]. Genome-wide patterns of linkage disequilibrium decay were estimated by calculating (LD/r2) for all pairs of genetic markers where both genetic markers exhibited minor allele frequency greater than 0.05 and were separated up to a physical distance of less than 600 kilobases using PopLDdecay (v3.41) [61].

Marker-trait associations were identified using 100 iterations of the FarmCPU algorithm as implemented in the R package rMVP v1.0.5 with parameter settings maxLoop = 10; method.bin = "FaST-LMM" [44, 58]. For each iteration, the first three principal components calculated from the genetic marker dataset were used as fixed effect and the kinship matrix calculated internally by the FarmCPU algorithm was fitted as random effects. #1 The overall marker file was filtered on a per-GWAS basis to retain only those markers with a minor allele frequency >0.05 among the lines phenotyped a given trait prior to association testing for that trait. For each trait, 100 analyses were run, each incorporating data from a different randomly selected subset of phenotyped lines [43]. In each resampling analysis, the overall threshold for statistical significance was the bonferroni corrected p-value at 5%. Resampling model inclusion probability (RMIP) values for each marker were calculated as the proportion of the 100 analyses in which that marker was significantly associated with the target trait [43, 42, 62, 63, 64].

Linkage disequilibrium was calculated among all genetic markers with RMIP values $\geq$ 5 for at least one trait. Genetic markers with linkage disequilibrium >0.5 were merged into single peaks for downstream analyses, unless the markers were separated by >1 Mb. When two or more markers were merged into a single peak, the marker with the greatest RMIP value was selected as representative of the entire peak.

Availability of supporting data and materials

Phenotypic values for all trait datasets employed in this study for all maize accessions evaluated are provided in Table S2 and S3. The sources of sequence data used to call genetic marker genotypes

for each maize accession were NCBI BioProjects PRJNA661271, PRJNA189400 & PRJNA437324 [8, 6, 5]. The specific SRA IDs for individual maize accessions are indicated in Table S2. The locations of GWAS peaks, the traits associated with each peak and the genes adjacent to each peak are provided in Table S4 and S5. The VCF file used for genetic analyses has been deposited at FigShare [65]. Scripts and code used to implement various analyses described in the methods section above have been deposited in github [66]. All supporting data and materials are available in the GigaScience GigaDB database [67].

Declarations

List of abbreviations

BLUPs: Best Linear Unbiased Predictions, BLUEs: Best Linear Unbiased Estimates, DAG: Diacylglycerol, exPVPs: expired Plant Variety Patents, LD: Linkage Disequilibrium, GWAS: Genome-wide Association Studies, MAP: Maize Association Panel, MLM: Mixed Linear Model, NCBI: National Center for Biotechnology Information, PCo: Principal coordinate, PCR: Polymerase Chain Reaction, QTL: Quantitative Trait Locus, RMIP: Resampling Model Inclusion Probability, SAM: Shoot Apical Meristem, SNP: Single Nucleotide Polymorphism, SRA: Sequence Read Archive TAG: Triacylglycerol, WiDiv: Wisconsin Diversity Panel.

Ethical Approval (optional)

Not applicable

Consent for publication

Not applicable

Competing Interests

James C. Schnable has equity interests in Data2Bio, LLC; Dryland Genetics LLC; and EnGeniousAg LLC. He is a member of the scientific advisory board of GeneSeek and currently serves as a guest editor for The Plant Cell. The authors declare no other conflicts of interest.

Funding

This material is based upon work supported by the US Department of Energy Advanced Research Projects Agency-Energy (ARPA-E) under Award Nos. DE-AR0001064 and DE-AR0001367, the National Science Foundation under grants OIA-1557417 and OIA-1826781 and USDA-NIFA under the AI Institute: for Resilient Agriculture, Award No. 2021-67021-35329 and the Foundation for Food and Agriculture Research Award No. 602757. This research was supported in part by the US. Department of Agriculture, Agricultural Research Service Project #5030-21000-068-00D. 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.

Author's Contributions

RVM, and JCS conceived of the study. RVM, GS, MG, MCT, HJ, CS, LN, AMT, and BS designed and carried out experiments to generate data. RVM, MG, and GS analyzed the data. RVM and JCS wrote the initial draft of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Thomas Hoban, Isabel Sigmon, Alice Guo, Nathaniel Pester, Leighton Wheeler, Sierra Conway, Isaac Stevens, and Olivier N. Mizero for assistance with field maintenance, harvest, and trait data collection.

References

1. Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 2003;165(4):2117–2128.
2. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. The plant journal 2005;44(6):1054–1064.
3. Leiboff S, Li X, Hu HC, Todt N, Yang J, Li X, et al. Genetic control of morphometric diversity in the maize shoot apical meristem. Nature communications 2015;6(1):1–10.
4. Hansey CN, Johnson JM, Sekhon RS, Kaeppler SM, De Leon N. Genetic diversity of a maize association population with restricted phenology. Crop Science 2011;51(2):704–715.
5. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights into the maize pan-genome and pan-transcriptome. The Plant Cell 2014;26(1):121–135.
6. Mazaheri M, Heckwolf M, Vaillancourt B, Gage JL, Burdo B, Heckwolf S, et al. Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC plant biology 2019;19(1):1–17.
7. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third-generation Zea mays haplotype map. Gigascience 2018;7(4):gix134.
8. Qiu Y, Oâ€™Connor CH, Della Coletta R, Renk JS, Monnahan PJ, Noshay JM, et al. Whole-genome variation of transposable element insertions in a maize diversity panel. G3 Genes|Genomes|Genetics 2021 06;11(10). https://doi.org/10.1093/g3journal/jkab238, jkab238.
9. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome biology 2013;14(6):1–18.
10. O'Connor C, Qiu Y, Della Coletta R, Renk JS, Monnahan P, Noshay JM, et al. Population level variation of transposable elements in a Maize diversity panel. Biorxiv 2020;.
11. Sun G, Mural RV, Turkus JD, Schnable JC. Quantitative resistance loci to southern rust mapped in a temperate maize diversity panel. bioRxiv 2021;.
12. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. science 2009;326(5956):1112–1115.
13. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature 2017;546(7659):524–527.
14. Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS one 2009;4(12):e8451.
15. Chia JM, Song C, Bradbury PJ, Costich D, De Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nature genetics 2012;44(7):803–807.
16. Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z, Millard MJ, et al. The genetic architecture of maize height. Genetics 2014;196(4):1337–1356.

17. Lin Hy, Liu Q, Li X, Yang J, Liu S, Huang Y, et al. Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. Genome biology 2017;18(1):1–14.
18. Gustafson TJ, de Leon N, Kaeppler SM, Tracy WF. Genetic analysis of sugarcane mosaic virus resistance in the Wisconsin diversity panel of maize. Crop Science 2018;58(5):1853–1865.
19. Gage JL, White MR, Edwards JW, Kaeppler S, de Leon N. Selection signatures underlying dramatic male inflorescence transformation during modern hybrid maize breeding. Genetics 2018;210(3):1125–1138.
20. Qiao P, Lin M, Vasquez M, Matschi S, Chamness J, Baseggio M, et al. Machine learning enables high-throughput phenotyping for analyses of the genetic architecture of bulliform cell patterning in maize. G3: Genes, Genomes, Genetics 2019;9(12):4235–4243.
21. Sekhon RS, Saski C, Kumar R, Flinn BS, Luo F, Beissinger TM, et al. Integrated genome-scale analysis identifies novel genes and networks underlying senescence in maize. The Plant Cell 2019;31(9):1968–1989.
22. Zheng Z, Hey S, Jubery T, Liu H, Yang Y, Coffey L, et al. Shared genetic control of root system architecture between Zea mays and Sorghum bicolor. Plant physiology 2020;182(2):977–991.
23. Azodi CB, Pardo J, VanBuren R, de Los Campos G, Shiu SH. Transcriptome-based prediction of complex traits in maize. The Plant Cell 2020;32(1):139–151.
24. Lin M, Matschi S, Vasquez M, Chamness J, Kaczmar N, Baseggio M, et al. Genome-wide association study for maize leaf cuticular conductance identifies candidate genes involved in the regulation of cuticle development. G3: Genes, Genomes, Genetics 2020;10(5):1671–1683.
25. Renk JS, Gilbert AM, Hattery TJ, O'Connor CH, Monnahan PJ, Anderson N, et al. Genetic control of kernel compositional variation in a maize diversity panel. The Plant Genome 2021;p. e20115.
26. Schneider HM, Lor VSN, Hanlon MT, Perkins A, Kaeppler SM, Borkar AN, et al. Root angle in maize influences nitrogen capture and is regulated by calcineurin B-like protein (CBL)-interacting serine/threonine-protein kinase 15 (ZmCIPK15). Plant, Cell & Environment 2021;.
27. Zhou Y, Kusmec A, Mirnezami SV, Attigala L, Srinivasan S, Jubery TZ, et al. Identification and utilization of genetic determinants of trait measurement errors in image-based, high-throughput phenotyping. The Plant Cell 2021;33(8):2562–2582.
28. Woodhouse MR, Cannon EK, Portwood JL, Harper LC, Gardiner JM, Schaeffer ML, et al. A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biology 2021;21(1):1–10.
29. Zheng P, Allen WB, Roesler K, Williams ME, Zhang S, Li J, et al. A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nature genetics 2008;40(3):367–372.
30. Bewley JD, Bradford K, Hilhorst H, et al. Seeds: physiology of development, germination and dormancy. Springer Science & Business Media; 2012.
31. dos Santos Maraschin F, Kulcheski FR, Segatto ALA, Trenz TS, Barrientos-Diaz O, Margis-Pinheiro M, et al. Enzymes of glycerol-3-phosphate pathway in triacylglycerol synthesis in plants: Function, biotechnological application and evolution. Progress in lipid research 2019;73:46–64.
32. Liang Y, Liu Q, Wang X, Huang C, Xu G, Hey S, et al. Zm MADS 69 functions as a flowering activator through the ZmRap2. 7-ZCN 8 regulatory module and contributes to maize flowering time adaptation. New Phytologist 2019;221(4):2335–2347.
33. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proceedings of the National Academy of Sciences 2007;104(27):11376–11381.
34. Lazakis CM, Coneva V, Colasanti J. ZCN8 encodes a potential orthologue of Arabidopsis FT florigen that integrates both endogenous and photoperiod flowering signals in maize. Journal of experimental botany 2011;62(14):4833–4842.
35. Meng X, Muszynski MG, Danilevskaya ON. The FT-like ZCN8 gene functions as a floral activator and is involved in photoperiod sensitivity in maize. The Plant Cell 2011;23(3):942–960.
36. Woodhouse MR, Sen S, Schott D, Portwood JL, Freeling M, Walley JW, et al. qTeller: a tool for comparative multi-genomic gene expression analysis. Bioinformatics 2022;38(1):236–242.
37. Fowler JE, Freeling M. Genetic analysis of mutations that alter cell fates in maize leaves: dominant Liguleless mutations. Developmental genetics 1996;18(3):198–222.
38. Bauer P, Lubkowitz M, Tyers R, Nemoto K, Meeley RB, Goff SA, et al. Regulation and a conserved intron sequence of liguleless3/4 knox class-I homeobox genes in grasses. Planta 2004;219(2):359–368.
39. Tibbs Cortes L, Zhang Z, Yu J. Status and prospects of genome-wide association studies in plants. The Plant Genome 2021;14(1):e20077.
40. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research 2019;47(D1):D1005–D1012.
41. Kaye J, Heeney C, Hawkins N, De Vries J, Boddington P. Data sharing in genomics—re-shaping scientific practice. Nature Reviews Genetics 2009;10(5):331–335.
42. Huang GJ, Shifman S, Valdar W, Johannesson M, Yalcin B, Taylor MS, et al. High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. Genome research 2009;19(6):1133–1140.
43. Valdar W, Holmes CC, Mott R, Flint J. Mapping in structured populations by resample model averaging. Genetics 2009;182(4):1263–1277.
44. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS genetics 2016;12(2):e1005767.
45. Du H, Feng BR, Yang SS, Huang YB, Tang YX. The R2R3-MYB transcription factor gene family in maize. PloS one 2012;7(6):e37463.
46. Hay A, Tsiantis M. KNOX genes: versatile regulators of plant development and diversity. Development 2010;137(19):3153–3165.
47. Mural RV, Grzybowski M, Miao C, Damke A, Sapkota S, Boyles RE, et al. Meta-Analysis Identifies Pleiotropic Loci Controlling Phenotypic Trade-offs in Sorghum. bioRxiv 2020;.
48. McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, et al. Maize genomes to fields (G2F): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. BMC research notes 2020;13(1):1–6.
49. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:14065823 2014;.
50. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–2120.
51. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997 2013;.
52. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.
53. Sun G, Yu H, Wang P, Guerrero MGL, Mural RV, Mizero ON, et al. A role for heritable transcriptomic variation in maize adaptation to temperate environments. bioRxiv 2022;.
54. The Picard toolkit;. Accessed: 2021-03-22. http://broadinstitute.github.io/picard/.

55. 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 research 2010;20(9):1297–1303.
56. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics 2018;103(3):338–348.
57. VanRaden PM. Efficient methods to compute genomic predictions. Journal of dairy science 2008;91(11):4414–4423.
58. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. rmvp: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics, proteomics & bioinformatics 2021;.
59. Covarrubias-Pazaran G. Genome-assisted prediction of quantitative traits using the R package sommer. PloS one 2016;11(6):e0156744.
60. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 2007;81(3):559–575.
61. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 2019;35(10):1786–1788.
62. Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, et al. Distinct genetic architectures for male and female inflorescence traits of maize. PLoS genetics 2011;7(11):e1002383.
63. Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, et al. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant physiology 2012;158(2):824–834.
64. Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES. Association mapping across numerous traits reveals patterns of functional variation in maize. PLoS genetics 2014;10(12):e1004845.
65. Mural RV, Sun G, Grzybowski M, Tross MC, Jin H, Smith C, et al. Maize WiDiv SAM 1051 Genotypes in VCF format. FigShare 2022 2;https://figshare.com/articles/dataset/Maize_WiDiv_SAM_1051Genotype_vcf_gz_genotype_file/19175888.
66. Mural RV, Maize resampling-based GWAS. GitHub; 2022. https://github.com/ravimural/Maize_resampling-based_GWAS.
67. Mural RV, Sun G, Grzybowski M, Tross MC, Jin H, Smith C, et al. Supporting data for "Association Mapping Across a Multitude of Traits Collected in Diverse Environments in Maize". GigaScience Database 2022;http://dx.doi.org/10.5524/102234.
68. Yu J, Holland JB, McMullen MD, Buckler ES. Genetic design and statistical power of nested association mapping in maize. Genetics 2008;178(1):539–551.
69. Li C, Li Y, Sun B, Peng B, Liu C, Liu Z, et al. Quantitative trait loci mapping for yield components and kernel-related traits in multiple connected RIL populations in maize. Euphytica 2013;193(3):303–316.
70. Yang X, Gao S, Xu S, Zhang Z, Prasanna BM, Li L, et al. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. Molecular Breeding 2011;28(4):511–526.
71. Revilla P, Rodríguez VM, Ordás A, Rincent R, Charcosset A, Giauffret C, et al. Cold tolerance in two large maize inbred panels adapted to European climates. Crop Science 2014;54(5):1981–1991.
72. Gouesnard B, Negro S, Laffray A, Glaubitz J, Melchinger A, Revilla P, et al. Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. Theoretical and Applied Genetics 2017;130(10):2165–2189.
73. Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, et al. Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 2014;198(1):3–16.

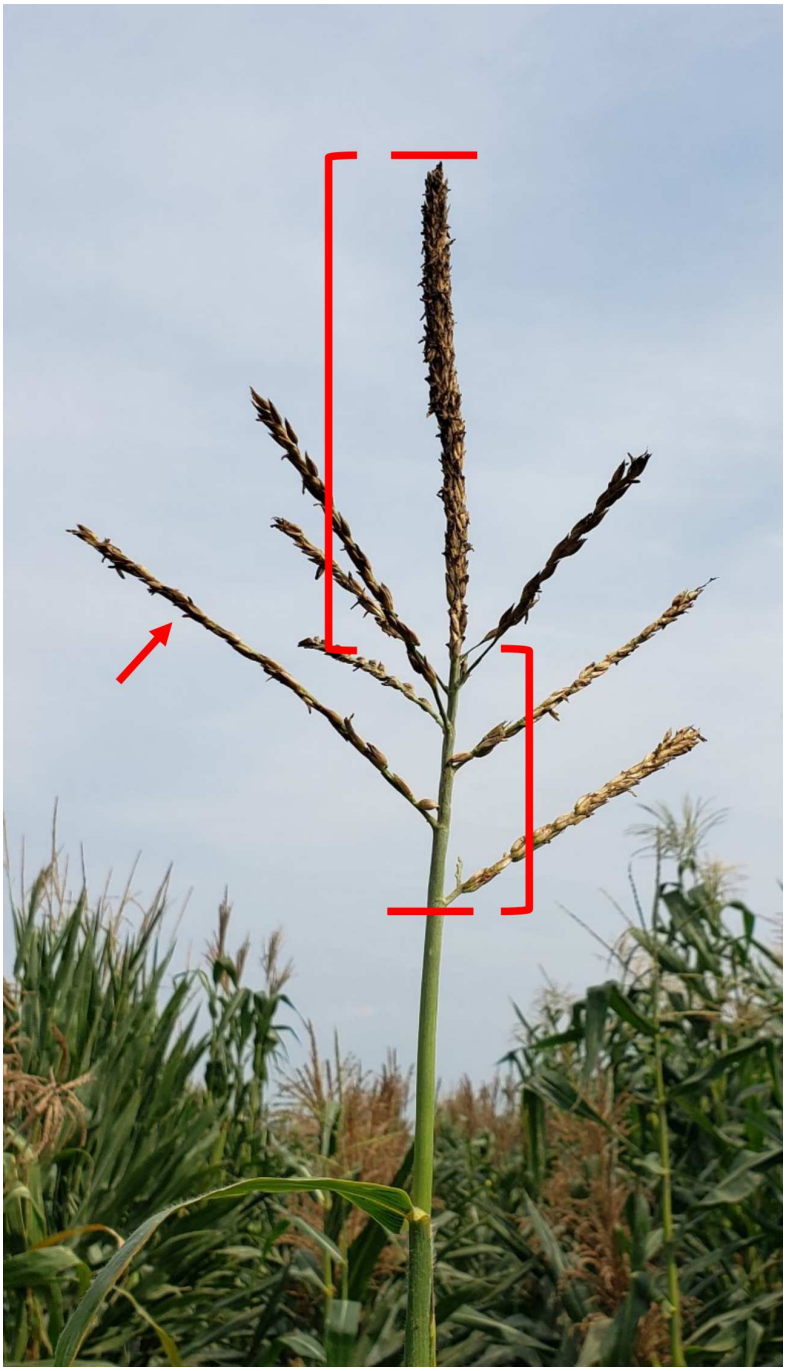

Figure S1. Phenotyping of Tassel Architecture. A) Tassel lengths were measured from the bottom-most primary tassel branch to the tip of the tassel spike (red bars). B) The branch zone length was defined as the length from the bottom-most primary tassel branch to the upper-most primary tassel branch (red left-facing bracket). C) The tassel spike length was defined as the length from the upper-most primary tassel branch to the tip of the tassel spike (red right-facing bracket). D) The total number of primary tassel branches were also counted (example arrowed) as well as the number of these primary tassel branches that were initiated, but later aborted (not pictured).

Figure S2. Phenotyping of Cob Traits. A) Ear lengths were measured as distance from base to the tip of each cob in centimeters, highlighted in green. B) ear fill was measured as distance on the cob from base to the tip where the seeds were set, highlighted in red. C) Ear width was measured as distance of diagonal of the cob, highlighted by blue color D) Kernels per row corresponds to the number of kernels on each line when cobs were set vertically, highlighted in pink color. E) Kernel row number correspond to the number of rows, highlighted in Brown color.

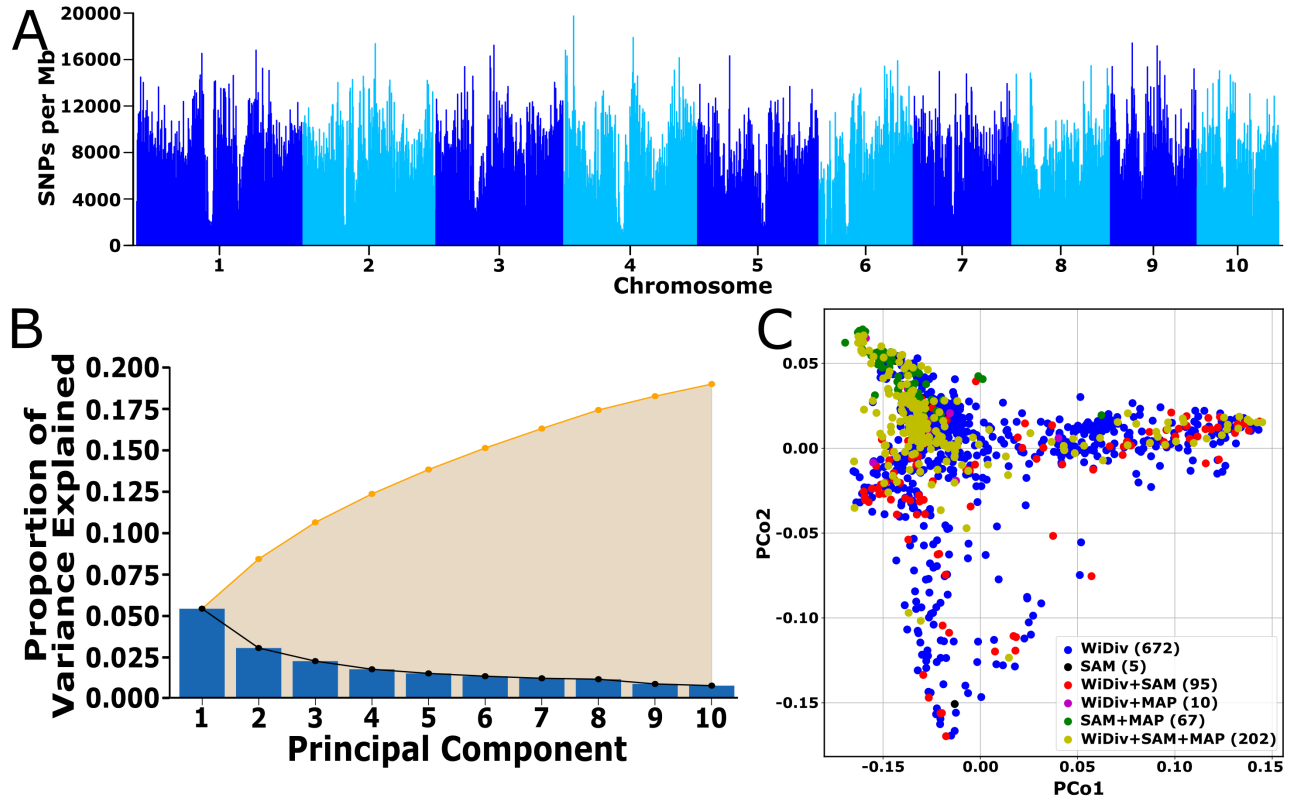

Figure S3. Multi Dimension Scaling or Principal Coordinate Analysis. A) Distribution of SNP density across the sorghum genome in 1 megabase sliding windows. B) Scree plot of eigen values for the principal components estimated from the marker data used in this study. C) Genetic relationship among the accessions used in this study and visualized using multidimensional scaling/principal coordinate analysis of the distance matrix. The X- and Y-axis represent first and the second principal component coordinates. Each point is color coded by the community association panels each accession belongs to.

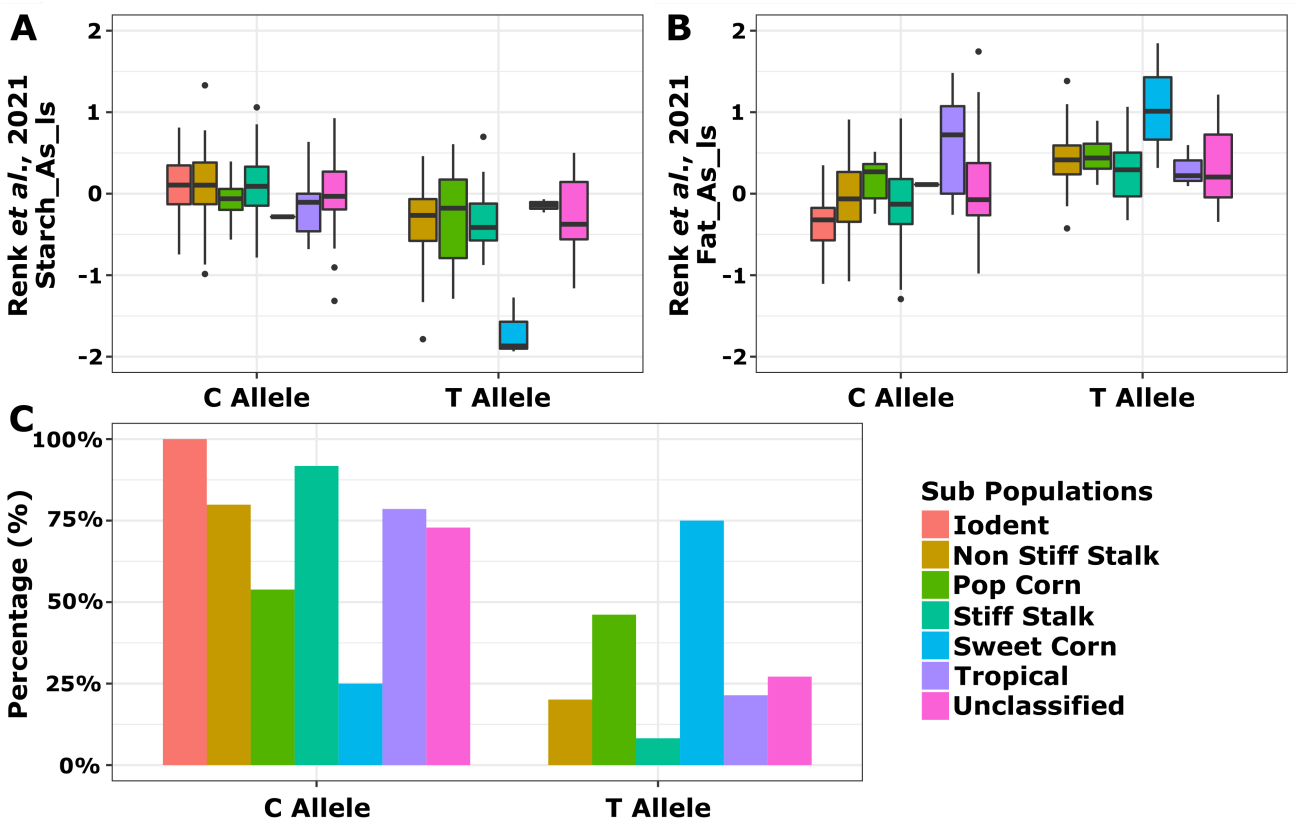

Figure S4. Allele frequencies of the top SNP associated with the DGAT-2 gene. A) The allele frequency for the starch content in each subpopulation of top SNP associated with DGAT gene. B) The allele frequency for the fat content in each subpopulation of top SNP associated with DGAT gene. C) Percentage of alleles in each subpopulation: The starch promoting allele was more abundant in iodent subpopulations and less abundant in sweet corn subpopulations.

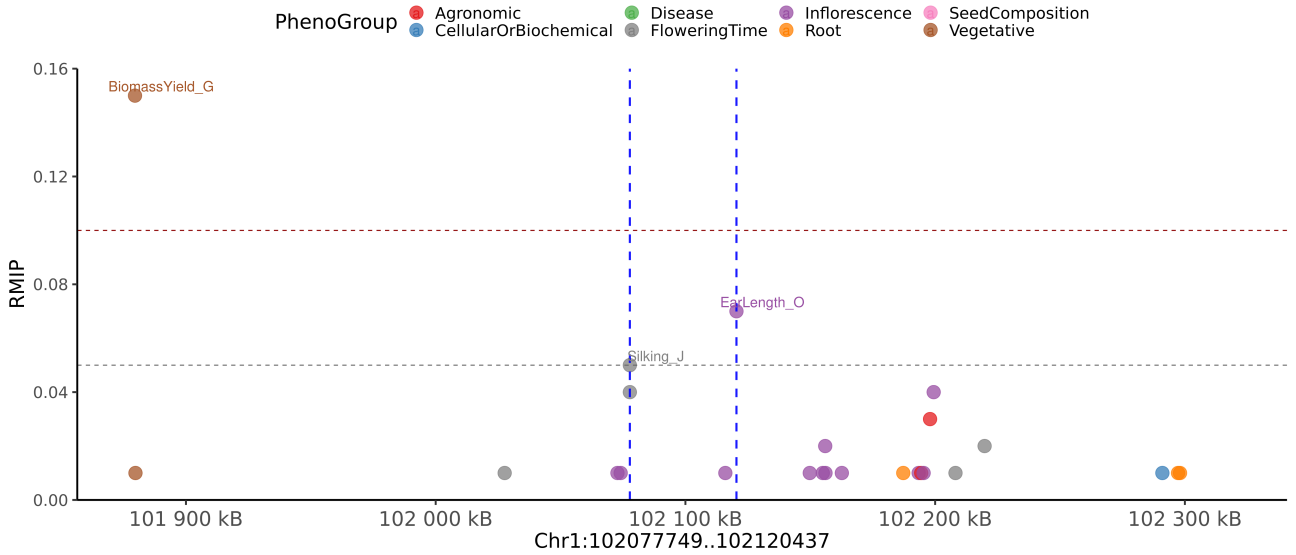

Figure S5. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 1 from 102,077,749 bp and 102,120,437. This peak is associated with the phenotypes belonging to Inflorescence and Flowering Time categories. The phenotypes associated with this group are EarLength_O and Silking_J.

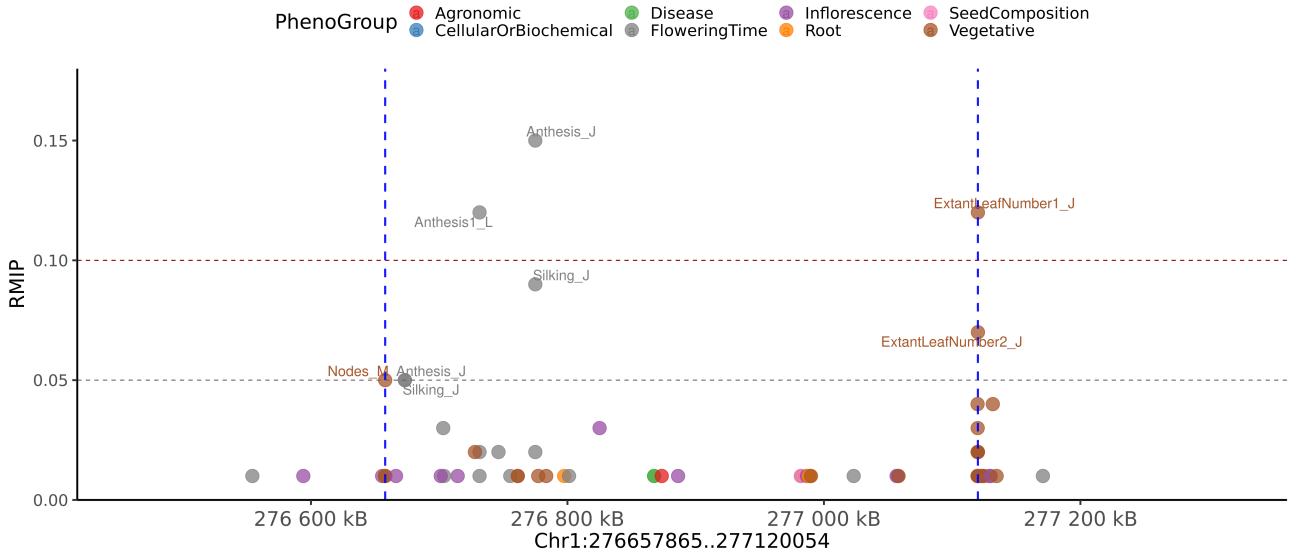

Figure S6. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 1 from 276,657,865 to 277,120,054. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are Anthesis1_L, Anthesis_J, ExtantLeafNumber1_J, ExtantLeafNumber2_J, Nodes_M, and Silking_J.

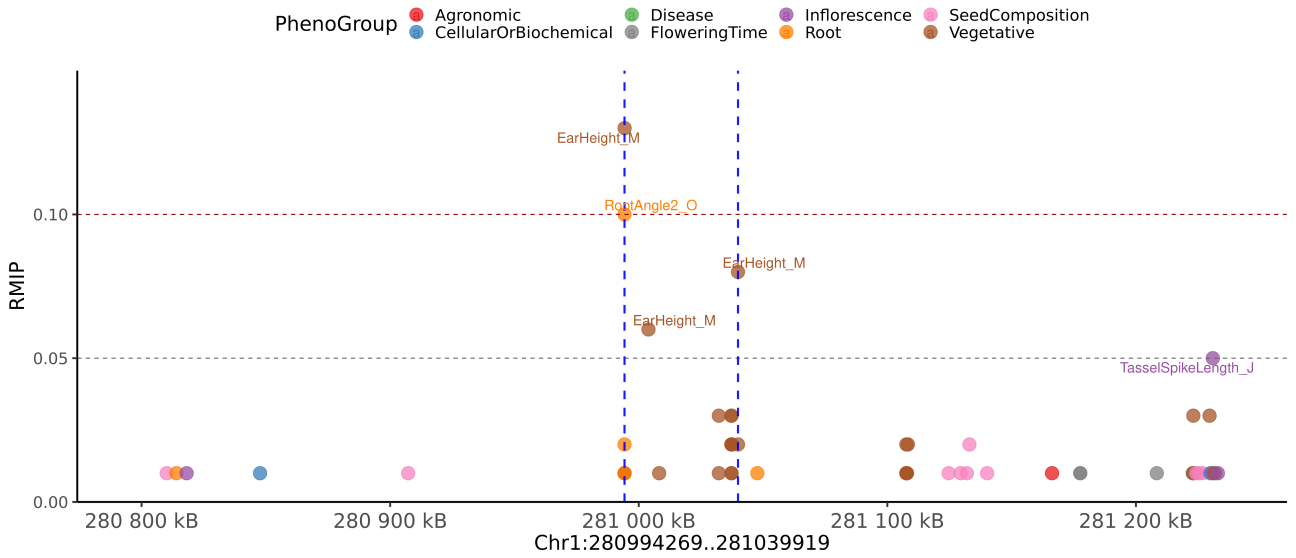

Figure S7. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 1 from 280,994,269 to 281,039,919. This peak is associated with the phenotypes belonging to Root and Vegetative categories. The phenotypes associated with this group are EarHeight_M and RootAngle2_O.

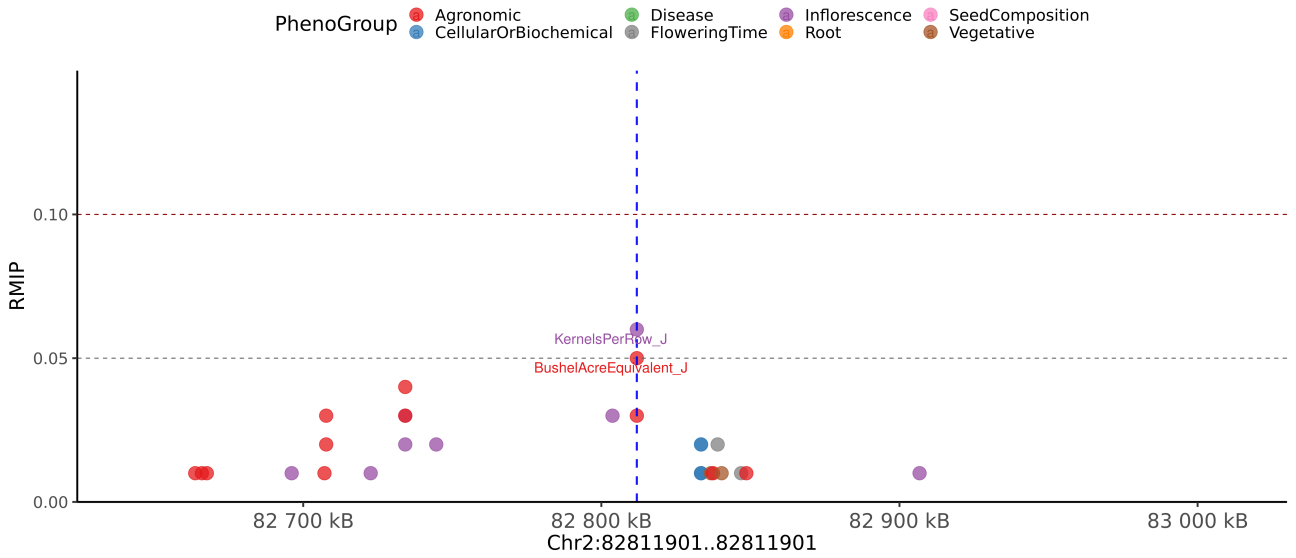

Figure S8. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 2 from 82,811,901 to 82,811,901. This peak is associated with the phenotypes belonging to Agronomic and Inflorescence categories. The phenotypes associated with this group are BushelAcreEquivalent_J and KernelsPerRow_J.

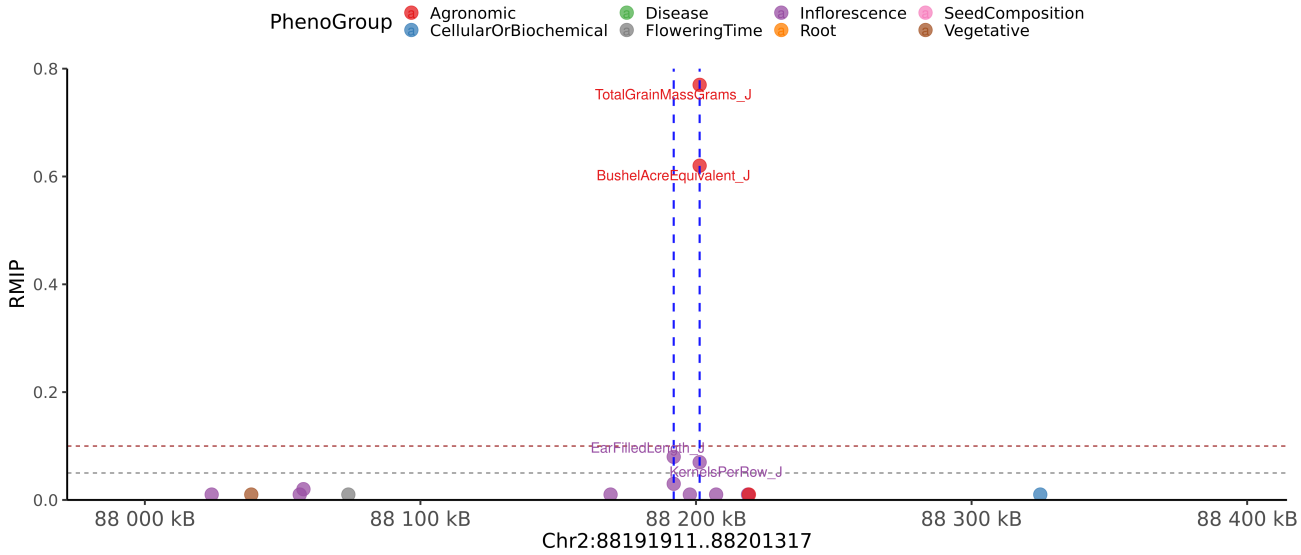

Figure S9. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 2 from 88,191,911 to 88,201,317. This peak is associated with the phenotypes belonging to Agronomic and Inflorescence categories. The phenotypes associated with this group are BushelAcreEquivalent_J, EarFilledLength_J, KernelsPerRow_J and TotalGrainMassGrams_J.

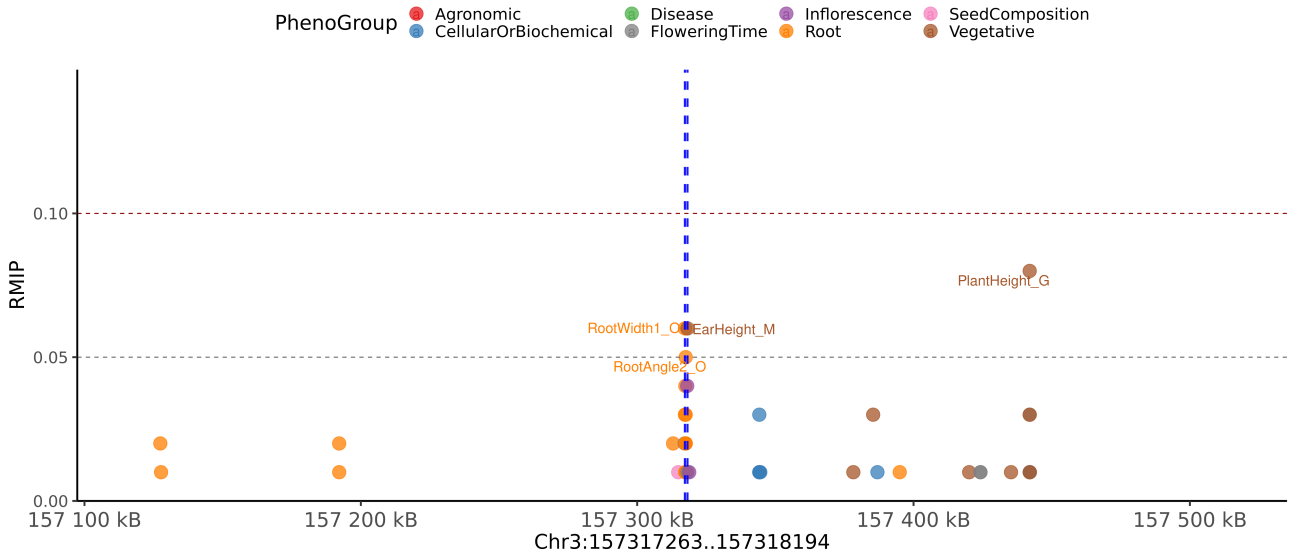

Figure S10. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 3 from 157,317,263 to 157,318,194. This peak is associated with the phenotypes belonging to Root and Vegetative categories. The phenotypes associated with this group are EarHeight_M, RootAngle2_O and RootWidth1_O.

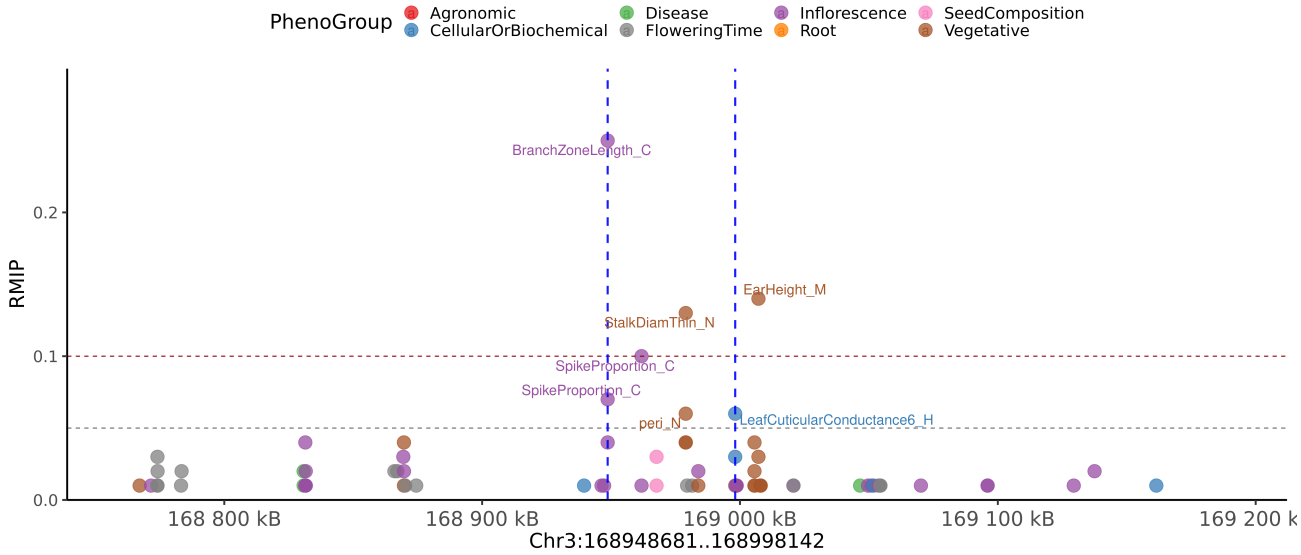

Figure S11. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 3 from 168,948,681 to 168,998,142. This peak is associated with the phenotypes belonging to Cellular/Biochemical, Inflorescence and Vegetative categories. The phenotypes associated with this group are BranchZoneLength_C, LeafCuticularConductance6_H, SpikeProportion_C, StalkDiamThin_N and peri_N.

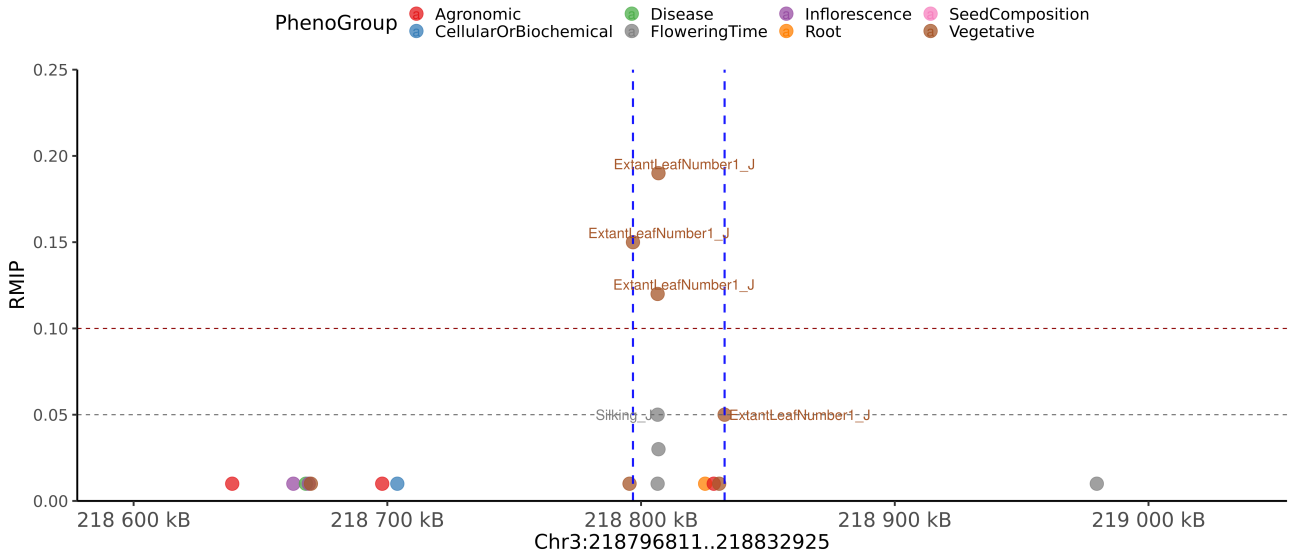

Figure S12. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 3 from 218,796,811 to 218,832,925. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are ExtantLeafNumber1_J and Silking_J.

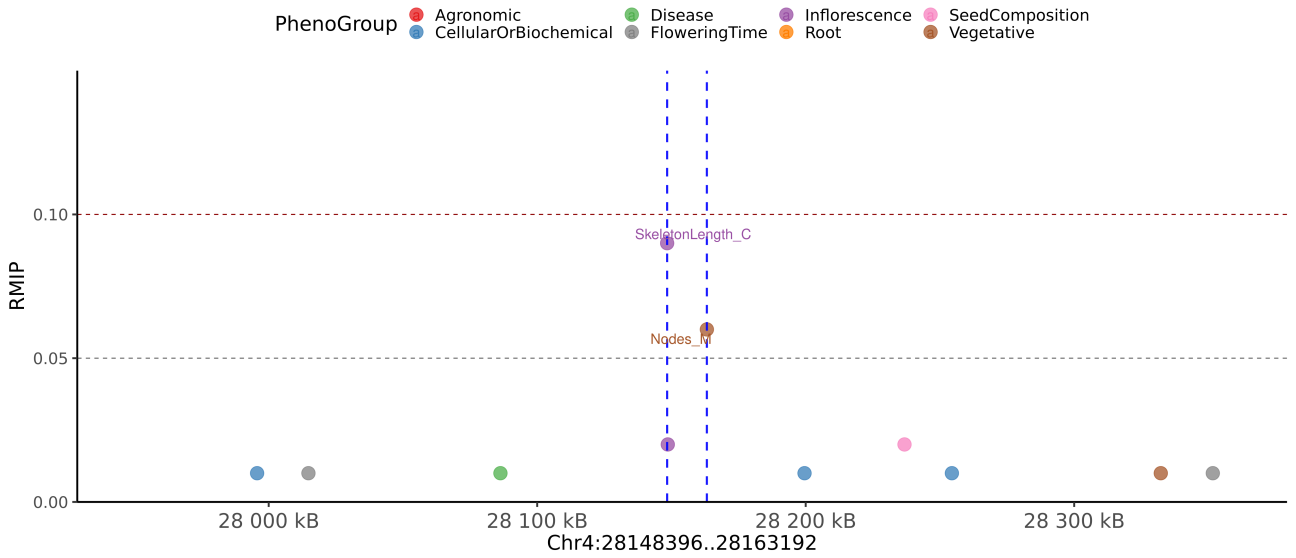

Figure S13. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 4 from 28,148,396 to 28,163,192. This peak is associated with the phenotypes belonging to Inflorescence and Vegetative categories. The phenotypes associated with this group are Nodes_M and SkeletonLength_C.

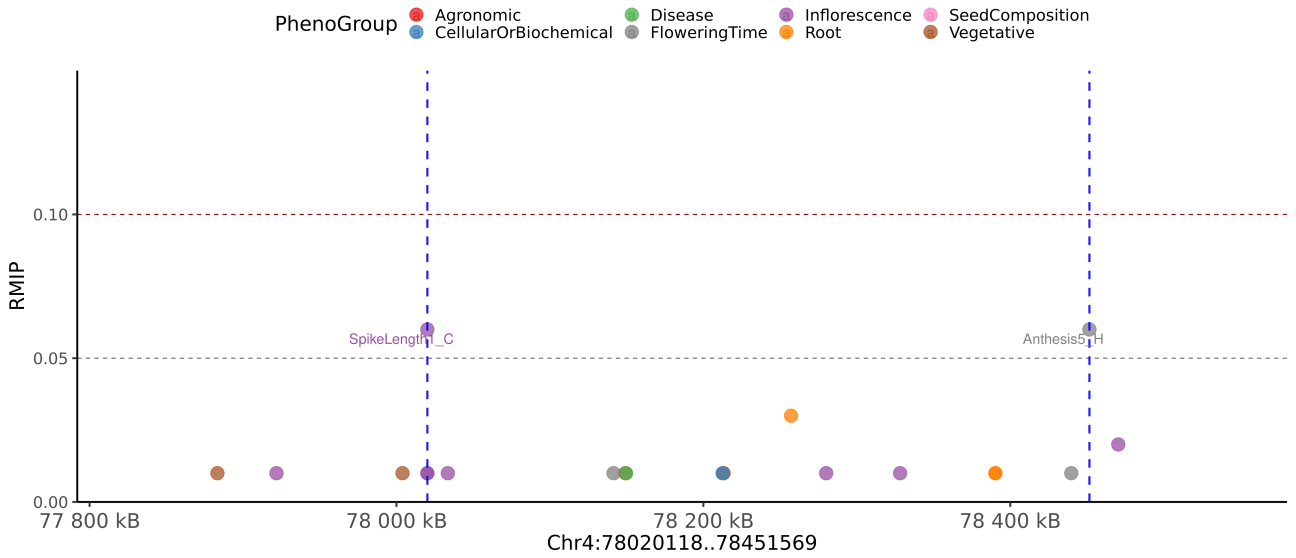

Figure S14. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 4 from 78,020,118 to 78,451,569. This peak is associated with the phenotypes belonging to Flowering Time and Inflorescence categories. The phenotypes associated with this group are Anthesis5_H and SpikeLength1_C.

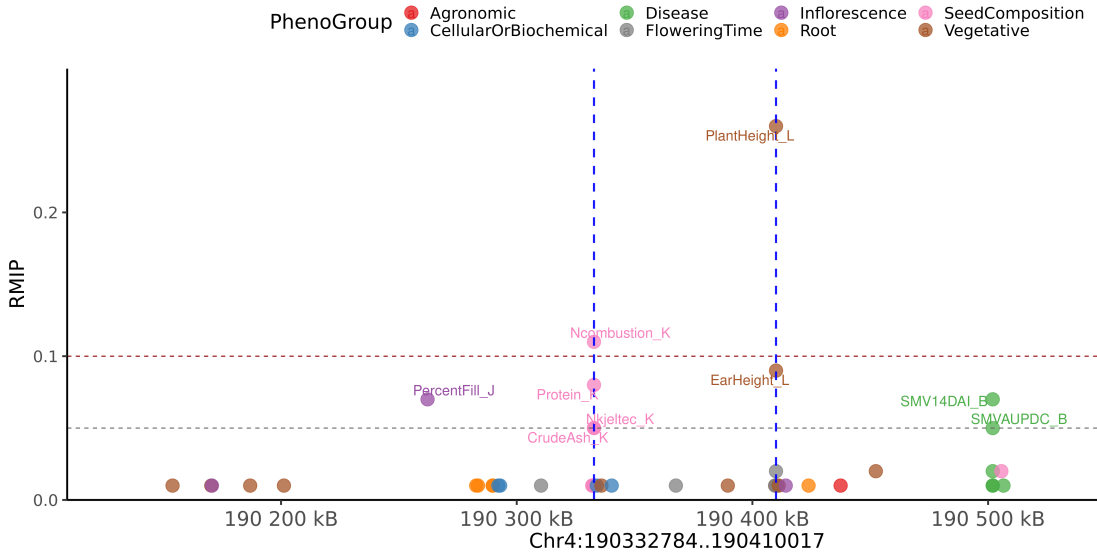

Figure S15. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 4 from 190,332,784 to 190,410,017. This peak is associated with the phenotypes belonging to Seed Composition and Vegetative categories. The phenotypes associated with this group are CrudeAsh_K, EarHeight_L, Ncombustion_K, Nkjeltec_K, PlantHeight_L, and Protein_K.

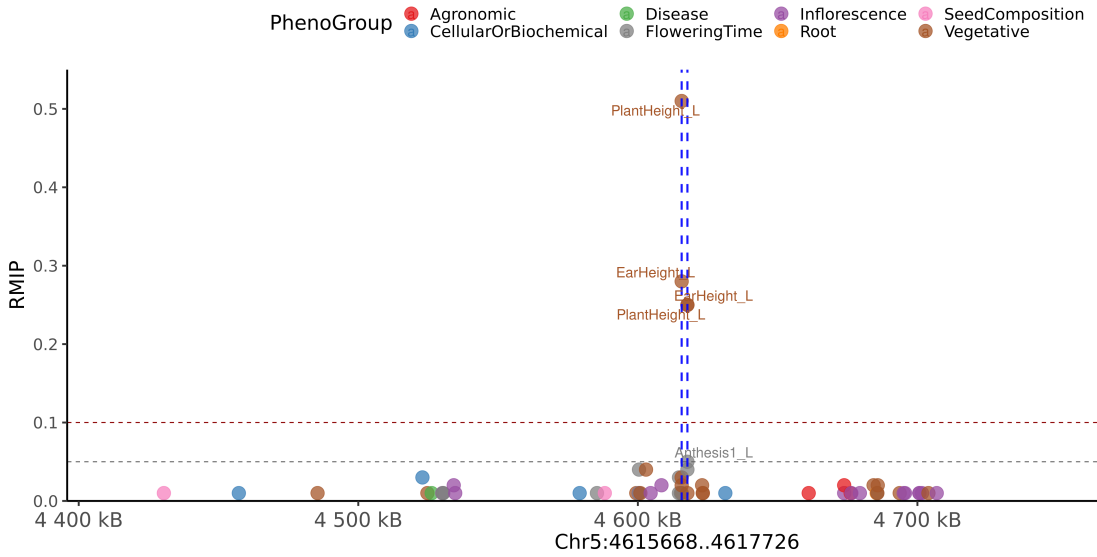

Figure S16. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 5 from 4,615,668 to 4,617,726. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are Anthesis1_L, EarHeight_L and PlantHeight_L.

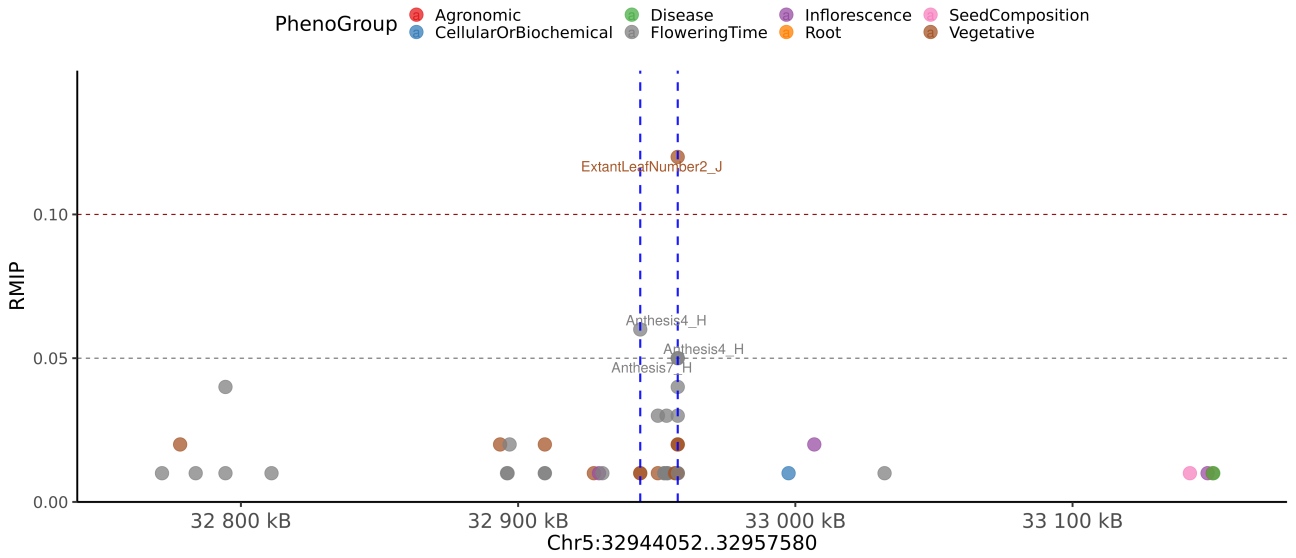

Figure S17. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 5 from 32,944,052 to 32,957,580. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are Anthesis4_H, Anthesis7_H and ExtantLeafNumber2_J.

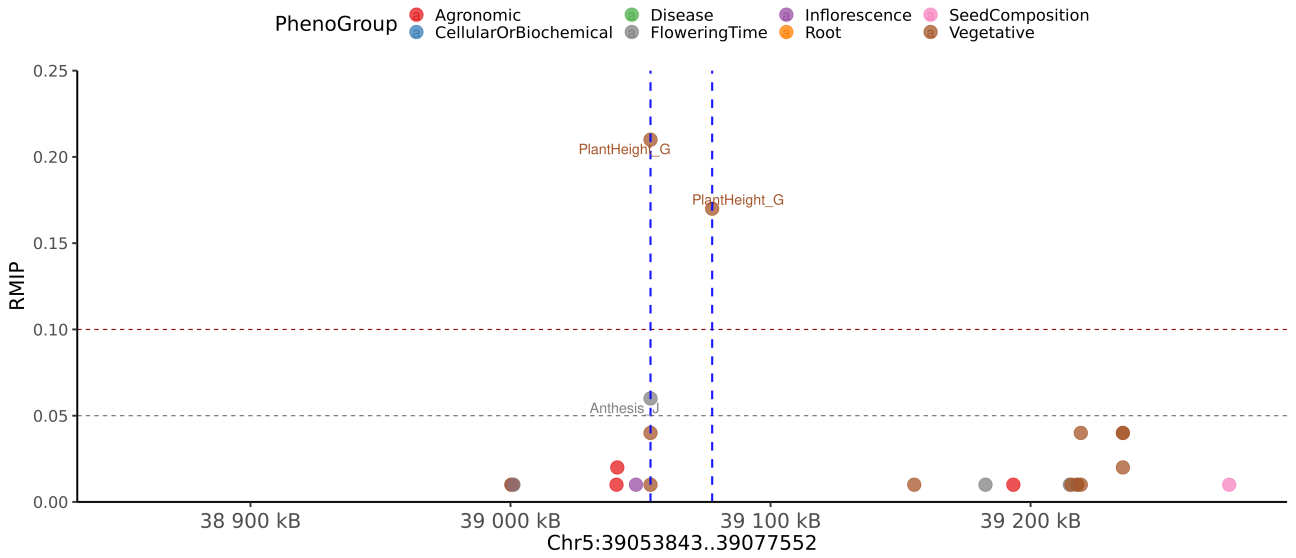

Figure S18. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 5 from 39,053,843 to 39,077,552. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are Anthesis_J and PlantHeight_G.

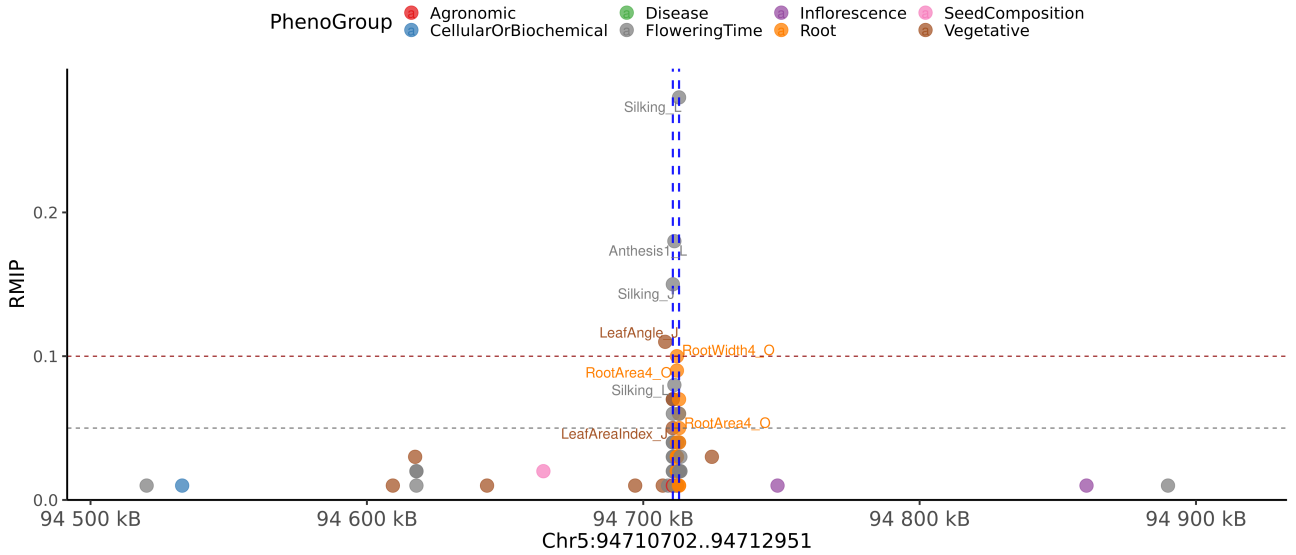

Figure S19. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 5 from 94,710,702 to 94,712,951. This peak is associated with the phenotypes belonging to Flowering Time, Root and Vegetative categories. The phenotypes associated with this group are Anthesis1_L, Anthesis7_H, Anthesis_A, Anthesis_J, LeafAreaIndex_J, LeafLength_J, RootArea1_O, RootArea2_O, RootArea4_O, RootWidth4_O, SilkingGDD_L, Silking_J and Silking_L.

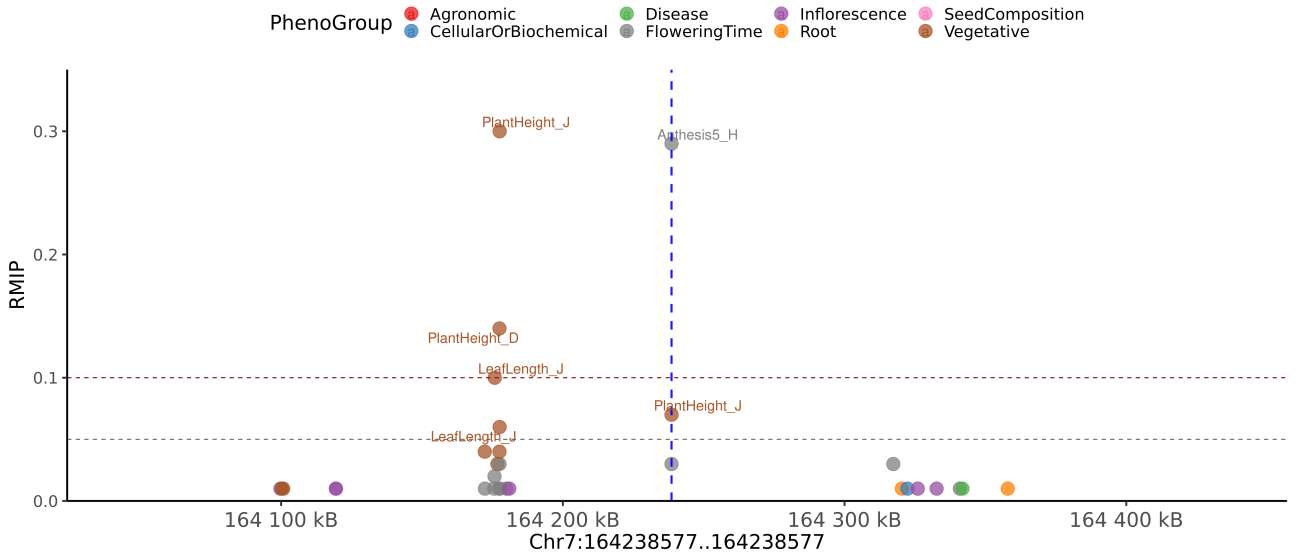

Figure S20. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 7 from 164,238,577 to 164,238,577. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are Anthesis5_H and PlantHeight_J.

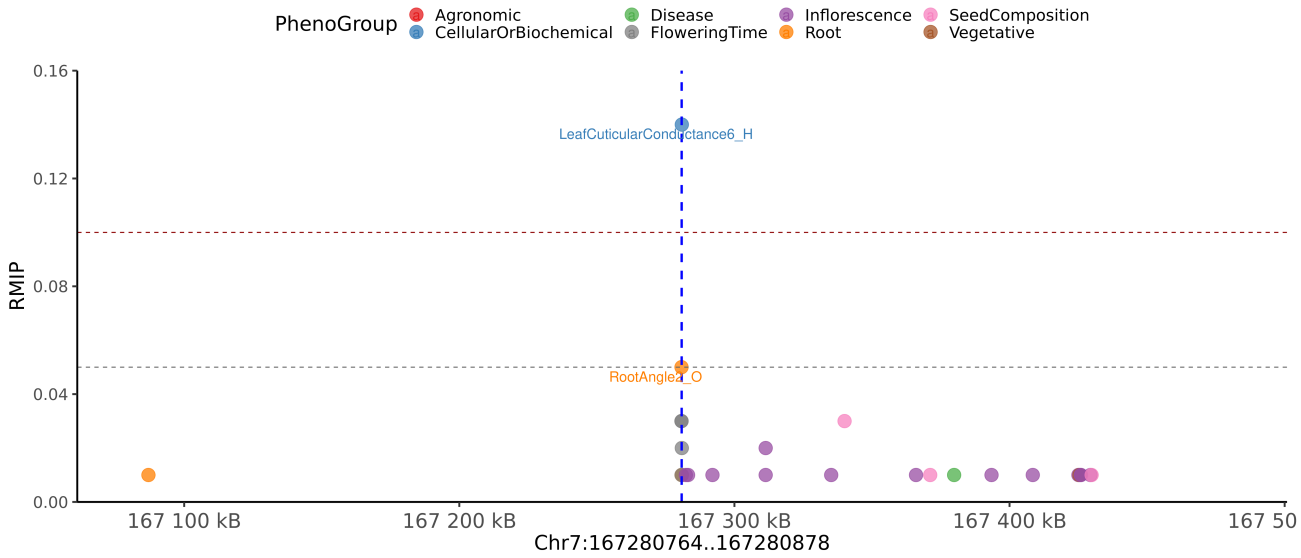

Figure S21. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 7 from 167,280,764 to 167,280,878. This peak is associated with the phenotypes belonging to Cellular/Biochemical and Root categories. The phenotypes associated with this group are LeafCuticularConductance6_H and RootAngle2_O.

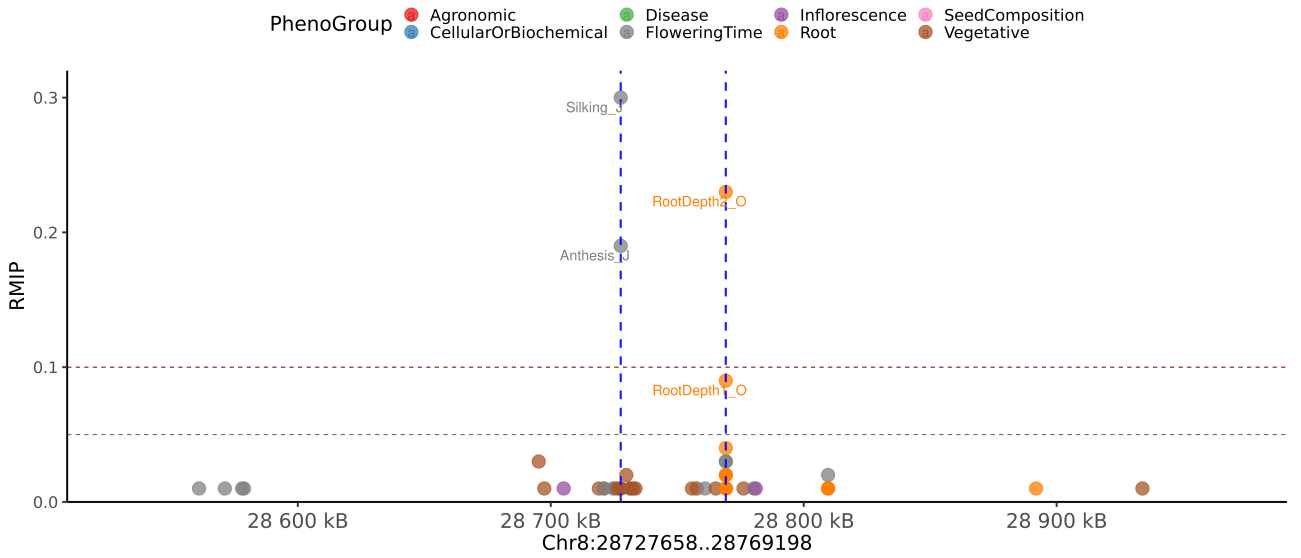

Figure S22. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 28,727,658 to 28,769,198. This peak is associated with the phenotypes belonging to Flowering Time and Root categories. The phenotypes associated with this group are Anthesis_J, RootDepth1_O, RootDepth2_O and Silking_J.

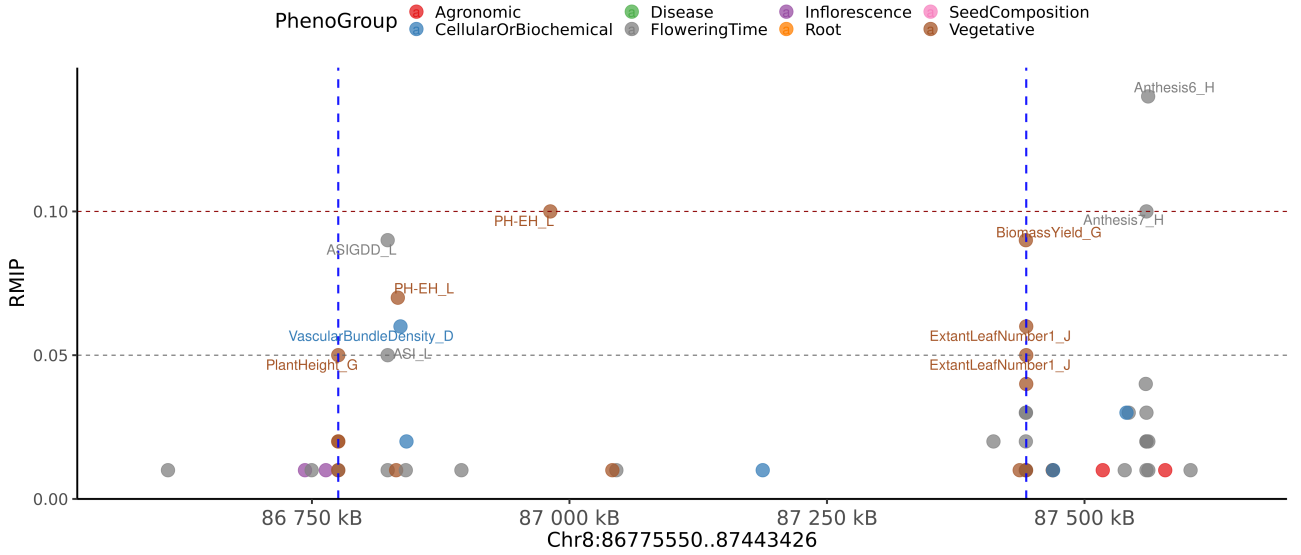

Figure S23. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 86,775,550 to 87,443,426. This peak is associated with the phenotypes belonging to Cellular/Biochemical and Vegetative categories. The phenotypes associated with this group are BiomassYield_G, ExtantLeafNumber1_J, PH-EH_L, PlantHeight_G and VascularBundleDensity_D.

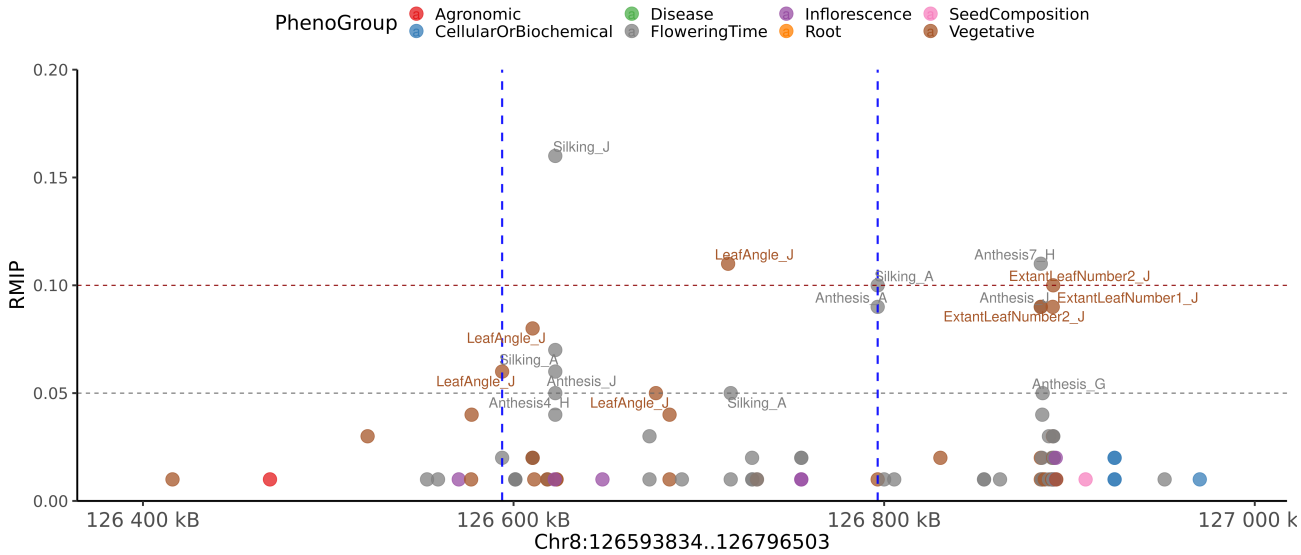

Figure S24. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 126,593,834 to 126,796,503. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are Anthesis4_H, Anthesis_A, Anthesis_J, LeafAngle_J, Silking_A and Silking_J.

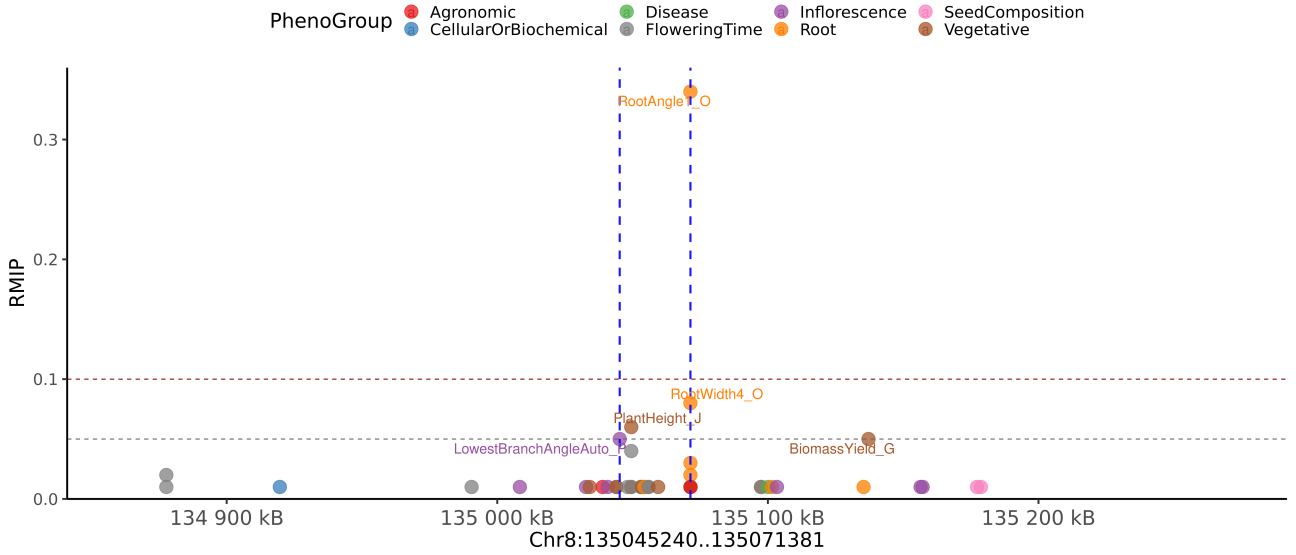

Figure S25. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 135,045,240 to 135,071,381. This peak is associated with the phenotypes belonging to Inflorescence, Root and Vegetative categories. The phenotypes associated with this group are LowestBranchAngleAuto_P, PlantHeight_J, RootAngle1_O and RootWidth4_O.

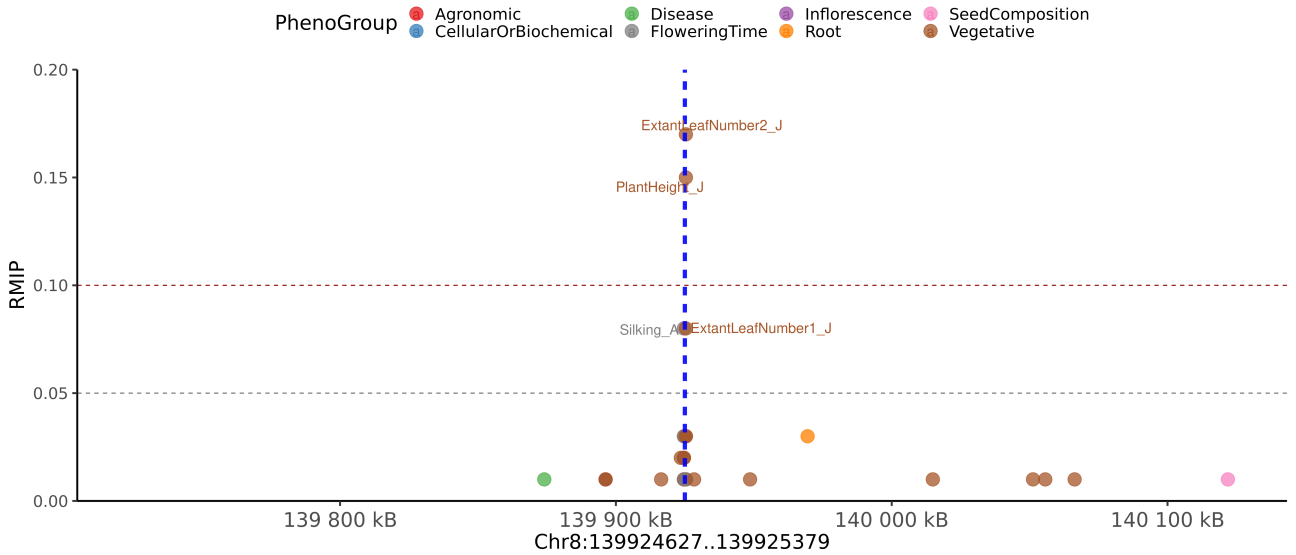

Figure S26. Local Manhattan plot with +/-200 kilobases of pleiotropic peak on chromosome 8 from 139,924,627 to 139,925,379. This peak is associated with the phenotypes belonging to Flowering Time and Vegetative categories. The phenotypes associated with this group are ExtantLeafNumber1_J, ExtantLeafNumber2_J, PlantHeight_J and Silking_A.

Table S1. Widely Used Maize Community Association Panels.

Panel Name	Panel Abbreviation	Country	# of Accessions [a]	Publication
Maize Association Panel	MAP	USA	282-312	Flint-Garcia *et al.* 2005[2]
Maize Nested Association Mapping Population	US-NAM	USA	5000	Jianming *et al.* 2008[68]
Ames Inbred Lines	NCRPIS	USA	2815	Romay *et al.* 2013[9]
Maize Shoot Apical Meristem	SAM	USA	369	Leiboff *et al.* 2015[3]
Wisconsin Diversity Panel	WiDiv	USA	503-942	Hanesy *et al.* 2011[4], and Mazaheri *et al.* 2019[6]
China Nested Association Mapping Population	CN-NAM	China	1971	Li *et al.* 2013[69]
Global Diverse Lines			527	Yang *et al.* 2011[70]
Dent and Flint Panel		Europe	306 + 292	Revilla *et al.* 2014[71]
European Flint Collection		Europe	1191	Gouesnard *et al.* 2017[72]
European Flint Nested Association Mapping	EUNAM-Flint	Europe	811	Lehermeier *et al.* 2014[73]
European Dent Nested Association Mapping	EUNAM-Dent	Europe	841	Lehermeier *et al.* 2014[73]

[a] The number of lines have changed over time due to the loss of some genotypes and/or the addition of others.

Table S2. Trait values for all phenotypes per accession and associated information per accession. Provided as included excel file.

Table S3. Phenotypes and associated information. Provided as included excel file.

Table S4. The locations of GWAS peaks, the traits associated with each peak and the genes adjacent to each peak. Provided as included excel file.

Table S5. Individual GWAS hits. Provided as included excel file.

Click here to access/download

Supplementary Material

SupplementalDataFileS2_GenotypesAndTraitValues.xlsx

Click here to access/download

Supplementary Material

SupplementalDataFileS3_PhenotypeMetaData.xlsx

Dear Hongfang Zhang,

Thank you very much for securing these three reviews for our manuscript and your encouragement to submit a revised manuscript for further consideration at Gigascience. I and my co-authors have worked to address the comments raised by all reviewers and indicated in the attached manuscript with new and revised text marked in blue and as explained in our point by point summary of the revisions made in response to each reviewer comment below.

Sincerely,
James

Original reviews in Black
Our responses in Blue

Reviewer reports:

Reviewer #1: Mural et al. reported a large-scale association analysis based on publicly published genotype and phenotype datasets and a meta-GWAS. This study provides a good example for mining community association panel data and further identifying candidate genes, pleiotropic loci and G x E. Actually, meta-analysis of GWAS has been used in humans and animals. However, I have some major concerns as follows.

1. This study only used three association panels (MAP, SAM, and WiDiv), as I know, some publicly available genotype and phenotype could be obtained for other association panels, for example the association panel including 368 inbred lines (Li et al., 2013, Nat Genetics, 45(1):43-50. doi: 10.1038/ng.2484), which was used widely in GWAS studies in maize. Can other association panels be integrated into this research, which would provide a rich genetic resource for maize research groups.

We agree with the reviewer that the integration of additional panels has the potential to increase the power of massive community GWAS in maize, a point we now touch on in the discussion section of this revised manuscript **(Page 10, Right column, Second paragraph).**

In the specific case of the AM508 panel from Li et al 2013 (now cited in the introduction **(Page 2, Left column, Third paragraph)** and discussion **(Page 10, Right column, Second paragraph)** of our revised manuscript) we faced two challenges. The first is that only a minority of published GWAS papers for this study include phenotype data for the individual lines.

The greater challenge is that the current genetic marker set used for GWAS with the AM508 panel is from an earlier version of the maize genome -- the website from Li et al 2013 was not accessible to us, but based on gene names and publication data we can narrow it down to B73 RefGen_v2 or v_3. Unfortunately this means the genetic markers used for the AM508 GWAS will not share markers/coordinates with the marker set employed in this study which was based on the B73_RefGen_v4 reference genome.

In the future this could be addressed either through reanalysis published resequencing of the AM508 panel, enabling integrated analysis of North American and Chinese maize diversity panels. We now discuss these options in our revised discussions section **(Page 10, Right column, Second paragraph).**

2. For association analysis, a total of 1014 unique inbred lines and 162 distinct traits from different association panels were used, but these traits were not measured for each of 1041 inbreds. For example, cellular-related traits were mainly measured in the SAM association panel. Hence, association analysis for cellular-related traits were conducted in SAM or 1014 inbreds. If 1014 inbreds were used to perform association analysis for cellular-related traits, how did you analyze the phenotype data? Please describe the method of phenotype data analysis in the Method section.

We apologize for this confusion. For each trait genetic data was subset to only that subset of individuals for which measured trait values were present. We have revised our methods section to make this less ambiguous (**Page 12, Left column, Second paragraph** of the "Quantitative Genetic Analysis of Trait Data" subsection).

3. Authors used RMIP values to identify significant association signals, please add more details about the RMIP method. What advantages of the resampling-based genome-wide association strategy over other methods?

You are right, this was a significant omission. We have added a new paragraph to the introduction discussing the RMIP method (**Page 2, Right Column, Second Paragraph**). Briefly, RMIP allows us to employ the FarmCPU algorithm which provides greater power to detect additional trait associated loci which might be masked by the effects of one or more large effect genes without the instability of results found in single individual runs of the FarmCPU GWAS algorithm. As one of our big goals in conducting this analysis was to generate a dataset that would enable other researchers to compare their results to. For this purpose the stability of marker trait associations is of particular value.

4. Although some important functional genes could be identified, were some new candidate genes obtained in this study functionally verified by the mutants or overexpression experiments.

Unfortunately no, the scope of our study did not permit us to conduct transgenic or gene knock out validation of new candidate genes. We have revised our discussion section to emphasize some of our highest priority candidates for future validation.

5. The authors identified pleiotropic loci based on categories of phenotypes associated with the same peak. For example, the phenotypes associated with the pleiotropic peak on chromosome 8 from 134,706,389 to 134,759,977 bp belongs to Flowering Time, Root and Vegetative categories, thus the locus was associated with different traits. Do you have any ideas on pleiotropic genes based on the results?

We have revised our manuscript to include discussion of both cases where a known causal gene is associated with a pleiotropic locus identified in our study (e.g. MADS69, Liang et al 2018), as well as to discuss in more detail potential candidate genes for two loci with pleiotropic effects without previously known and validated candidate genes associated (**Page 10, Right Column, Third Paragraph**)

Reviewer #2: The authors described a study of integrating multiple published datasets for reanalysis. They combined previously community panel data and newly collected data in the present study, finally assembling 1014 accessions with 18M SNP markers and 162 traits at different environments. They used a resample-based GWAS method to reanalyze this assemble dataset, and identified 2154 suggestive associations and 697 confident associations. They found genetic loci were pleiotropic to multiple traits. As the authors mentioned, I acknowledge their efforts for collecting and assembling different sources of previously datasets, which should be useful for the maize community. However, to the manuscript per se, I feel the paper seems not to be sufficiently quantified regarding the novelty and significance of reported findings.
If the authors could present several novel results because the previous studies had the limitations on population size, diversity, trait dimensions and environments. In this study, the authors seemed trying to present like this, but it may be improved further and more. It's hard to let me understand there are some novel things which was found due to the merged large dataset. On the other hand, using this assembled dataset, I'm not very clear what's the scientific questions that the authors want to address.

This is a very insightful comment that speaks to the difference approaches research groups take to science. Our goal in assembling this dataset was not to address a specific scientific question, but to generate a dataset/resource which would be valuable and reusable for multiple scientific research avenues. Essentially this boils down the distinction between hypothesis *testing* research and hypothesis *generating* research.

In terms of what novel things that were found as a result of our initial proof of concept analyses of this large dataset our intent was to emphasize both the quantitative genetic evidence for pleiotropy between above ground and below ground traits. In this revised manuscript we have revised and added text to emphasize the novelty of this finding. See also our response to reviewer #1 point #4.

In technical sense, I'm wondering how did authors deal with the batch effects when merging datasets phenotype from different environments? It's not comparable for the phenotypes from different accessions collected in different environments. It's hard to figure out the phenotypic difference is caused by genotype, environment, or their interaction.

We apologize that this was not explained sufficiently clear in our previous version. We did not merge data across multiple environments, unless the merging was conducted by the authors prior to making them public by the authors. We have revised our methods section to clarify this that we analyzed these datasets separately rather than merging and producing batch effects (**Page 12, Left column, Second paragraph** of "Quantitative Genetic Analysis of Trait Data" subsection). This means when a peak is identified in one dataset and not another for same phenotype from different studies is because of genotype by environment interacts that we now discuss in greater length in the discussion section (**Page 10, Left Column, Third Paragraph**).

The introduction section lacked the proper review for the project background, related progress and publications and findings.

We were sorry to read that the reviewer feels we did not provide sufficient background and citations in the previous version of this manuscript, and assure him or her that it was not our intent to minimize or omit references to the work of other researchers within this field.

While we are unsure what specific publications and findings the reviewer feels were improperly omitted by us, this revised manuscript includes an expanded introduction and discussion and 14 additional new citations not present in the original manuscript.

Reviewer #3: The manuscript "Association Mapping Across a Multitude of Traits Collected in Diverse Environments in Maize" by Ravi V. Mural et al. reported the application of high-density genetic marker data from two partially overlapping maize association panels, comprising 1,014 unique genotypes grown in seven US states, allowing the identification 2,154 suggestive marker-trait associations and 697 confident associations and suggesting the possible application to study gene functions, pleiotropic effects of natural genetic variants and genotype by environment interaction.

The background data are well documented, experimental data are convincing, clearly presented and well discussed, the paper is suitable for publication in Giga Science in its present form.

Thank your for your kind words regarding our manuscript and work.