Distinct characteristics of genes associated with phenome-wide variation in maize (*Zea mays*)

Zhikai Liang[1,2], **Yumou Qiu**[3], **and James C. Schnable**[1,2*]

[1]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, USA
[2]Plant Science Innovation Center, University of Nebraska-Lincoln, Lincoln, NE, USA
[3]Department of Statistics, Iowa State University, Ames, IA, USA
[*]Corresponding author: schnable@unl.edu

ABSTRACT

Naturally occurring functional genetic variation is often employed to identify genetic loci that regulate specific traits. Existing approaches to link functional genetic variation to quantitative phenotypic outcomes typically evaluate one or several traits at a time. Advances in high throughput phenotyping now enable datasets which include information on dozens or hundreds of traits scored across multiple environments. Here, we develop an approach to use data from many phenotypic traits simultaneously to identify causal genetic loci. Using data for 260 traits scored across a maize diversity panel, we demonstrate that a distinct set of genes are identified relative to conventional genome wide association. The genes identified using this many-trait approach are more likely to be independently validated than the genes identified by conventional analysis of the same dataset. Genes identified by the new many-trait approach share a number of molecular, population genetic, and evolutionary features with a gold standard set of genes characterized through forward genetics. These features, as well as substantially stronger functional enrichment and purification, separate them from both genes identified by conventional genome wide association and from the overall population of annotated gene models. These results are consistent with a large subset of annotated gene models in maize playing little or no role in determining organismal phenotypes.

Main

Genetics seeks to link individual genes to their roles in determining the characteristics of an organism. Early QTL studies utilized individual phenotypic[1] or chromosomal markers[2]. Now Genome Wide Association Study (GWAS) can scan 100,000's or millions of markers for association with a target trait[3–5]. Statistical methods for Phenome Wide Association Study (PheWAS) have also developed[6–8]. Unifying GWAS and PheWAS produces multiple testing problems which make retaining statistical power challenging[9,10]. Multivariate GWAS methodologies have been shown to increase power[11–17]. However, these approaches face challenges scaling to hundreds or thousands of traits. Medical record data mining[18,19], high throughput phenotyping[20], and scoring of molecular traits such as transcript and metabolite abundance[21–24] are making high dimensional trait datasets increasingly common. Here we employ a published dataset of 260 distinct traits of maize (*Zea mays*)[25,26] to evaluate a multi-trait multi-SNP framework to identify genotype-phenotype associations. Genes identified by our model show greater independent validation[27] and increased similarity to a gold standard set of genes characterized by knockout phenotypes in maize[28].

Maize HapMap3 SNPs[25,26] were imputed and filtered based on minor allele frequency, linkage disequilibrium, and distance to annotated gene models to produce a set of 557,968 unique SNPs associated with 32,084 maize gene models (See Methods). Filtering of a set of 57 unique traits scored across up to 16 environments resulted in a set of 260 trait datasets with a median missing data rate of 18%[29]. Unobserved trait datapoints were imputed using PHENIX[30] (Supplementary Table S1).

Two widely used GWAS methodologies – GLM and FarmCPU[31,32] (Table 1) – were employed to identify gene-trait associations (Table 1). A given gene may be identified as statistically significantly associated with a phenotype in a single analysis, but fail to survive multiple testing correction when many trait datasets are analyzed sequentially (Figure 1a). We developed an approach based upon stepwise regression model – Genome-Phenome Wide Association Study (GPWAS) – fitting where the set of SNPs fallen in a gene treated as the response variable, and both population structure and individual trait datasets are employed to explain the patterns of genetic variance across the population (See Methods; Figure S1). In principle, this approach should address the challenges of partially correlated traits and genotype matrices (Figure 1c). Given the complexities introduced by the iterative model selection step, we chose to correct for multiple testing using a permutation-based method (See Methods) which has been shown to be robust in controlling false positives in both GWAS and PheWAS studies[33,34]. With an estimated false discovery rate (FDR) $<$ 1.00e-3, 1,776 genes were classified as significantly associated with phenome-wide

variation (Figure S2). Comparison gene sets were identified by GLM[31] and FarmCPU[32] using a same dataset (See Methods; Table 1).

The accuracy of each of these three approaches was validated using data from a second much larger dataset, the maize nested association mapping (NAM) population[27,35]. The comparison employed the subset of 29,430 gene models with clear 1:1 correspondence between RefGenV4 and RefGenV2 maize gene models. GPWAS identified genes showed substantially higher overlapped with the validation dataset than GLM GWAS (p=2.15e-5; Chi-squared test) and FarmCPU GWAS (p=9.17e-3; Chi-squared test).

GPWAS produces a list of the specific traits which have been included in the model for a particular gene (Figure 1b). However, the associations of individual phenotypes identified within the GPWAS model for a given gene are not rigorously controlled for false discovery. Anther ear1 (*an1*) is a classical maize gene shown to encode a Ent-Copalyl diphosphate synthase involved in gibberellic acid biosynthesis. Knockout alleles of *an1* have been show to produce reduced or abolished tassel branching, reduced plant height, delayed growth, and delayed flowering[36]. Anther ear1 is also associated with quantitative variation in tassel spike length[37]. The *an1* gene was not significantly associated with any individual traits in conventional GWAS in this dataset. GPWAS found the association between this gene and a group of 11 traits was statistically significant (FDR < 0.001). Many traits incorporated into this model were consistent with the characterized function of this gene (Figure S3).

Maize genes with known mutant phenotypes – classical mutants[28] – are more likely to show significant mRNA expression in at least one tissue, and tend to be expressed across a broader set of tissues than other gene models (Table 1; Supplementary Table S2). Genes identified by GPWAS were more likely to be expressed > 1 FPKM in at least one of the 92 tissues/timepoints than those identified by GLM and FarmCPU, although the latter difference was not statistically significant (p=0.007, p=0.085; chi-squared test)[39,40] (Table 1). Genes identified by GLM GWAS, FarmCPU GWAS and GPWAS all showed much higher breadth of expression than other gene models (Supplementary Table S2). Genes identified by GPWAS were longer, contained both more total SNPs and a higher density of SNPs per KB of gene space than genes identified by conventional GWAS (Supplementary Table S3). But both of populations showed the same bias towards greater polymorphism rates relative to all annotated genes (Figure S4; Supplementary Table S3). Permutation testing revealed a modest bias towards the identification of genes with greater numbers of SNPs by GPWAS, however, this bias was insufficient to explain the patterns observed in real data (Supplementary Table S4). Classical maize mutants exhibited the opposite pattern, showing less polymorphism than the overall gene set (Supplementary Table S3). Both genes identified through forward genetics screens and through analysis of natural variation must play a role in specifying the characteristics of an organism. However, unlike genes identified through forward genetics, genes identified through analysis of natural variation must also be represented by functionally variable alleles in the studied population. It may be this second criteria which explains the additional bias towards the identification of genes with high polymorphism rates in GPWAS and GWAS.

Table 1. Expression Characteristics of Different Gene Populations

	Total Gene Models	Expressed Gene Models (Average FPKM >1)	RefGenV4:RefGenV2 1:1 Gene Models	Overlap with Gene Models Identified in NAM
RefGenV4	45,045	22,463 (49.9%)	29,430	4,227 (14.4%)
GLM GWAS	2,000	1,393 (69.7%)	1,712	301 (17.6%)
FarmCPU GWAS	880	620 (70.5%)	783	149 (19.0%)
GPWAS	1,776	1,309 (73.7%)	1,615	381 (23.6%)
Classical Mutants	99	75 (75.8%)	99	18 (18.2%)

Maize classical mutants are substantially less likely to exhibit presence absence variation (PAV) (Table 2). Gene models identified by both GWAS and GPWAS were much less likely to exhibit PAV than other gene models (Table 2; Supplementary Table S5). GPWAS was less likely to identify gene models with PAV than conventional GWAS (p=1.55e-3; Chi-squared test). Genes exhibiting PAV in maize are less likely to be conserved at syntenic locations in other species[41]. The frequency of syntenic conservation was the inverse of the pattern observed for PAV (Table 2; Supplementary Table S5), and the increased syntenic conservation of GPWAS identified genes relative to conventional GWAS was statistically significant (p<2.2e-16; p=1.46e-6; Chi-squared test; GLM and FarmCPU respectively). It was also possible to calculate Ka/Ks ratios for maize genes with syntenic orthologs in sorghum. Classical mutants show much lower Ka/Ks ratios – a sign of stronger purifying selection – than the overall population of conserved genes (Figure 2; Table 2). The Ka/Ks ratios of gene models identified by GLM GWAS were not significantly different from the overall population of conserved gene models (Figure 2; Table 2). Gene models identified by GPWAS showed significantly lower Ka/Ks ratios than all conserved gene models (p=1.24e-9), gene models identified by GLM GWAS (p=1.09e-9) and FarmCPU GWAS (p=4.24e-5; Mann–Whitney U test) (Figure 2; Table 2).

A set of 137 GO terms showed statistically significant (Bonferroni corrected p-value < 0.05) enrichment (119 terms) or

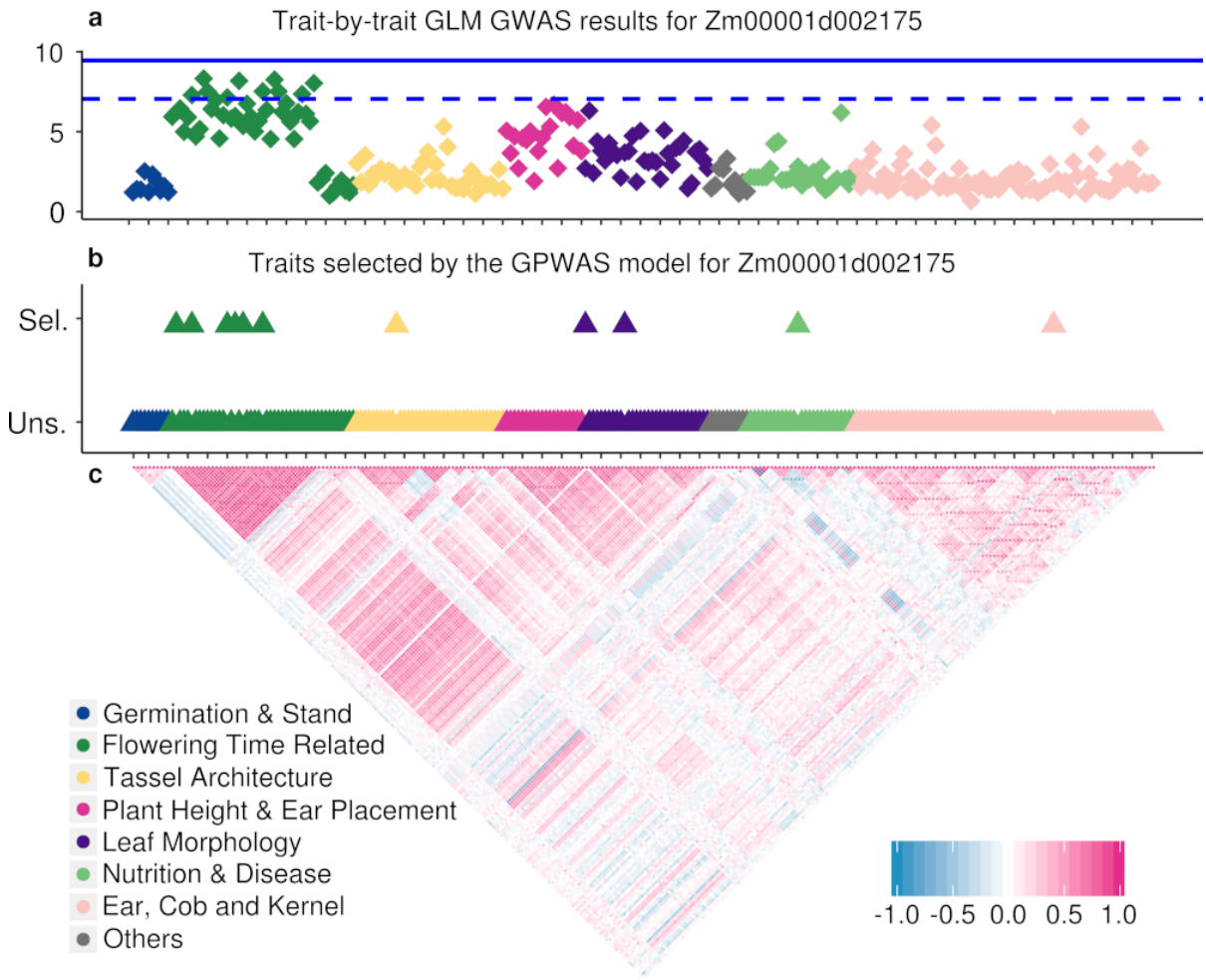

Figure 1. Statistically association between the maize gene Zm00001d002175 and 260 distinct phenotypes. Each diamond or triangle represents one specific phenotypic dataset. Colors of diamonds and triangles indicate the broad categories each specific phenotype falls into (see legend within figure). The specific identities of each phenotype ordered from left to right are given in Supplementary Table S1. (a) The height of each diamond indicates the negative $\log_{10}$ p-value of the most statistically significant SNP among the SNPs assigned to that gene in a GLM GWAS analysis for that single trait. Conventional GWAS analysis generally employs either empirically determined statistical significance cutoffs[27], or employs Bonferroni correction based on the total number of tests conducted[38]. Employing Bonferroni correction, in the dataset above, each individual analysis would be conducted using a multiple testing corrected p-value cut off of 8.96e-08, while sequential analysis of all 260 traits should employ a multiple testing corrected p-value of 3.45e-10. The dashed blue line in the top panel indicates a p=0.05 cut off after Bonferonni correction for multiple testing based on the number of statistical tests in a single GWAS analysis (8.96e-8). The solid line in the top panel indicates a p=0.05 cut off after Bonferonni correction for multiple testing based on the number of statistical tests in GWAS for all 260 traits (3.45e-10). (b) The vertical placement of each triangle indicates whether a given phenotype was included (Sel.) or excluded (Uns.) from the final GPWAS model constructed for this gene. The complete list of genotypes incorporated into the GPWAS model for Zm00001d002175 are Days to Silk (Summer 2006, Cayuga, NY; Summer 2007, Johnston, NC), Days to Tassel (Summer 2007, Johnston, NC; Summer 2008, Cayuga, NY), GDDDay to Silk (Summer 2006, Cayuga, NY; Summer 2007, Johnston, NC), Main Spike Length (Summer 2006, Johnston, NC), Number of Leaves (Summer 2008, Cayuga, NY), Leaf Width (Summer 2006, Champaign, IL), NIR Measured Protein (Summer 2006, Johnston, NC) and Ear Weight (Summer 2006, Champaign, IL). (c) The panel indicates the pairwise pearson correlation coefficient between each pair of measured phenotypes. Clustering based on phenotypic correlation was used to determine the ordering of phenotypes along the x-axis. Each tick mark on the x-axes of a and b panels indicates a distance of five phenotypes. Detailed phenotypes from left to right were in the same order as in Supplementary Table S1 from top to bottom.

purification (18 terms) among genes uniquely identified by GPWAS relative to GLM GWAS. In contrast only 15 GO terms

Table 2. Differences Among Gene Populations[a]

	Presence Absence Variation (Percent)[b]	Syntenic Conservation (Percent)[c]	Ka/Ks Median; Mean
All Genes	11,971/39,005 (30.7%)	27,735/45,578 (60.9%)	0.200; 0.246
Uniquely Identified by GLM GWAS	165/1,591 (10.4%)	1,322/1,630 (81.1%)	0.210; 0.251
Uniquely Identified by GPWAS	98/1,397 (7.0%)	1,292/1,406 (91.9%)	0.169; 0.208
Classical Mutants	4/98 (4.1%)	93/99 (93.9%)	0.144; 0.177

[a] Comparison for the same set of features in genes Uniquely Identified by FarmCPU relative to GPWAS and Uniquely Identified by GPWAS relative to FarmCPU are provided in Supplementary Table S5.
[b] Genes not included in[42] were excluded from this analysis.
[c] The syntenic conservation is defined as orthologs in maize relative to sorghum based on a previous publication[43].

(11 enriched and 4 purified) were identified in the corresponding set of genes uniquely identified by GLM GWAS (Figure 3a, Supplementary Table S6). Genes annotated as involved in development, response to stimuli, cell wall and cell membrane metabolism, hormone signalling, disease response, and transport were all over represented among genes associated with phenome-wide variation, while those associated with nucleotide metabolism, DNA replication, translation, and telomere organization were disproportionately unlikely to show such associations. Relative to the total number of gene models a given GO term is assigned to (e.g. information content), p values of enriched GO terms tends to be more significant (Figure 3a). Similar results were obtained for FarmCPU even after controlling for total number of genes identified (see Supplementary Information; Figure 3b; Supplementary Table S6). Number of GO terms per gene and proportion of genes with no assigned GO term did not differ dramatically between gene populations, however the median GO term assigned to a gene uniquely identified by GPWAS had higher information content than the median GO term assigned to a gene uniquely identified by GLM GWAS (See Methods; Supplementary Table S7). These results are consistent with unified GPWAS identifying a less random subset of annotated genes than sequential GWAS for each trait.

The statistical method we employ for GPWAS requires a complete absence of missing data. It is only because advances in kinship based phenotypic imputation approach is now available[30]. It also requires binning individual genetic markers into groups associated with individual genes. This binning is likely to be imperfect, as regulatory regions of genes can be separated from coding sequence by tens of Kb in maize[45,46]. Noncoding regulatory sequences, many distant from annotated genes, have been shown to explain approximately 40% of the total phenotypic variation in maize[47]. Finally, the present GPWAS algorithm and implementation is quite computationally expensive. We estimate the GPWAS analyses presented here required a total of approximately 6.9 years of CPU time.

Today, there are only a few datasets which contain as many different traits scored for the same population across multiple environments as the Panzea dataset. However, the rapid emergence of high throughput plant phenotyping technologies make it likely that high dimensional trait datasets – where the number of measured phenotypes exceeds even the number of individuals in the population – will become much more common in the future[20]. Increases in the total number of phenotypes should increase the power and accuracy of GPWAS. However, if many highly correlated traits are included, the result can be issues with multicollinearity that makes the statistical estimation and inference procedures employed unstable. The current statistical procedure also encounters challenges once the number of input traits exceed the number of individuals in the population. In these cases, it would be best to avoid the common practice of employing BLUP scores[48] as this approach strips out information on trait plasticity across environments, and trait plasticity is often controls by distinct sets of genes from genes controlling variation in multi-environment trait means[49]. Automatic variable selection and/or dimensional reduction could be incorporated into future GPWAS implementations. Here we have developed a new approach to identify genes with statistical links to a variation in a large set of diverse plant traits scored for a maize diversity panel across multiple environments, and showed that it exhibits greater consistency with genes identified as controlling organismal phenotypes in an independent population than do genes identified using two conventional GWAS approaches. We also showed that gene models identified by GPWAS exhibit greater structural, molecular, and evolutionary similarity to gold standard maize genes identified through forward genetics than genes identified by conventional trait-by-trait GWAS. Over the past three decades, without substantial discussion or debate, many in the scientific community have moved from a definition of genes that was based on organismal function, to one which is based on molecular features[50–52]. It may be possible to combine all of these data types to quantitatively determine the subset gene models likely involved in specifying the characteristics of organisms. These patterns could also guide prioritization in

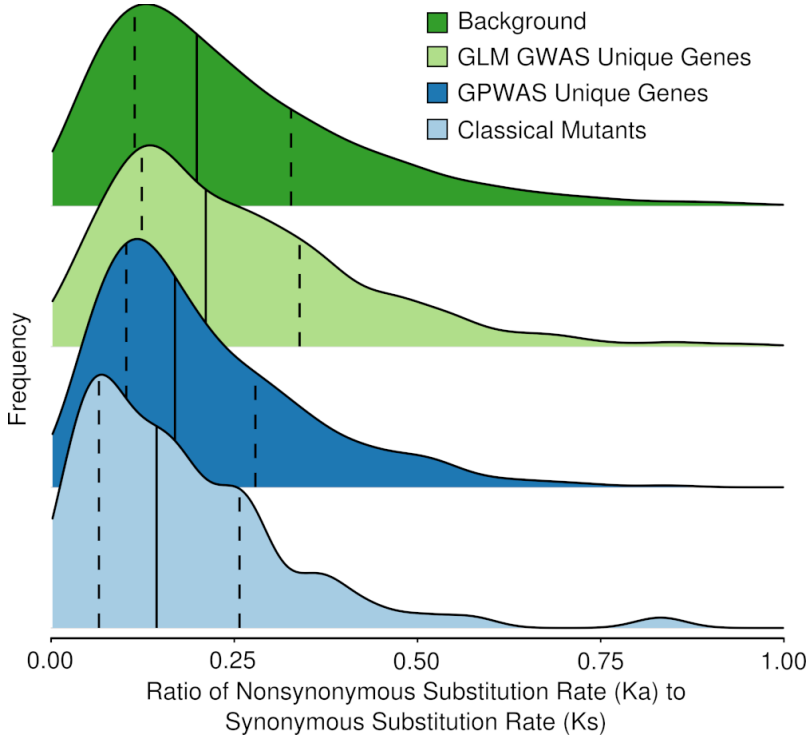

Figure 2. Distribution of Ka/Ks values for different populations of genes within the maize genome. The background set is composed of all maize genes with syntenic orthologs in sorghum and setaria after genes with tandem duplicates and genes with extremely few synonymous substitutions identified in the original alignment were excluded (See methods). The kernel density plots for genes uniquely identified by either GLM GWAS or GPWAS, as well as classical mutants are the subsets of each of these categories which also met the criteria for inclusion in the background gene set. For each population of genes the median value is indicated with a solid black line, and dashed black lines indicate 25th and 75th percentiles of the distribution.

future reverse genetics efforts.

Methods

Genotype and Phenotype Sources, Filtering, and Imputation

Raw genotype calls in RefGenV4 coordinates from resequencing data of the maize 282 association panel[26] were retrieved from PanZea. Missing genotypes were imputed using Beagle (version: 2018-06-10)[53]. Only biallelic SNPs with less than 80% missing points were input for imputation. After imputation, SNPs with MAF (Minor Allele Frequency) less than 0.05 or which were scored at heterozygous in more than 10% of samples were discarded. A phenotype file (traitMatrix_maize282NAM_v15-130212.txt) containing total of 285 traits, corresponding to 57 unique types of phenotypes scored in 1 to 16 environments was downloaded from PanZea. A set of 277 accessions with identical names in the HapMap3 data release and the PanZea trait data were employed for all downstream analyses.

Maize gene regions were extracted from AGPv4.39 downloaded from Ensembl. SNPs were clustered based on $R^2 > 0.8$ and only one randomly selected SNP per cluster was retained. If the number of SNPs after collapsing highly correlated clusters exceeded 138 (50% of the number of inbreds scored), a random subsample of 138 SNPs was employed for downstream analyses. Identical final SNP sets were employed for GPWAS and GWAS analyses.

Of the 285 initial trait datasets, 25 were removed because the data file contained a recorded trait value for only a single individual among the 277 maize inbreds genotyped, leaving a total of 260 trait datasets. Missing phenotypes were imputed based on a kinship matrix calculated from 1.24 million SNPs calculated in GEMMA[15] and using a Bayesian multiple-phenotype mixed model[30]. Accuracy of phenotypic imputation was assessed independently for each trait with sufficient number of real observations to evaluation using ten iterations of masking 1% of available records for each trait and comparing imputed and masked values for each trait (Supplementary Table S1).

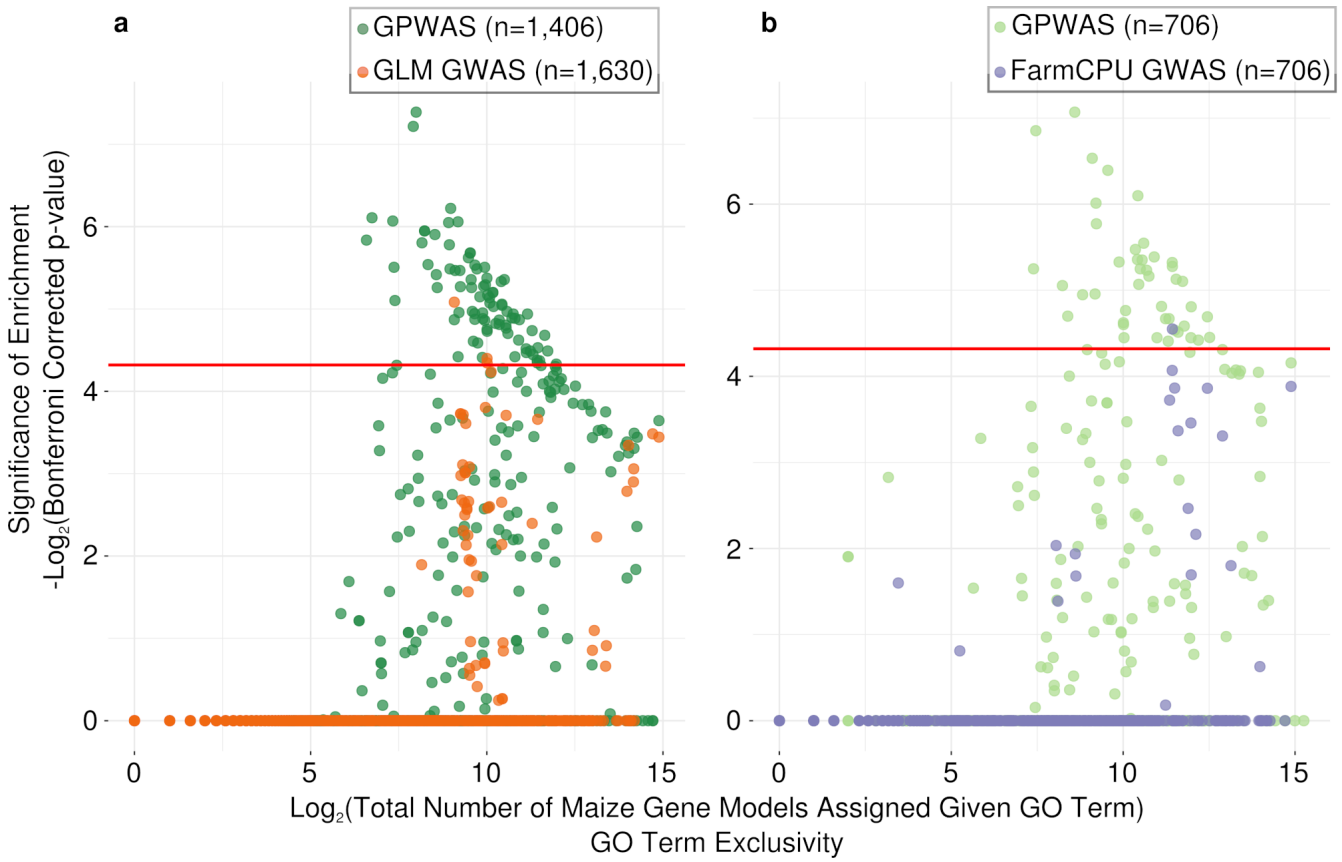

Figure 3. Comparison of GO enrichment/purification among genes uniquely identified as associated with phenotypic variation using different statistical approaches. Each circle is a single GO term in a single analysis. The position of each circle on the x axis indicates the total number of maize gene models which were assigned this GO term in the maize GAMER dataset[44]. The position of each circle on the y axis indicates the statistical significance of the enrichment or purification of this GO term in the given gene population relative to the background set of all annotated maize gene models. (a) Comparison of the patterns of GO term enrichment/purification among genes either either uniquely identified as associated with phenotypic variation by GLM GWAS analysis or uniquely identified as associated with phenotypic variation by GPWAS analysis. (b) As in panel A, but comparing genes uniquely identified as associated with phenotypic variation by FarmCPU analysis or uniquely identified as associated with phenotypic variation by GPWAS analysis. Only to 706 genes uniquely identified by GPWAS with the strongest statistical signal were employed in panel B, to prevent any bias towards more significant p-values which would result from conducting the analysis with a larger population of genes for GPWAS than for FarmCPU.

GPWAS Analysis

We propose a model selection approach to adaptively choose the most significant phenotypes associated with each gene. Given a gene, we consider all the SNPs as the multi-responses. For analysis of the given gene on each chromosome, a separate principal component analysis (PCA) was conducted using markers solely from the other 9 chromosomes to reduce the endogenous correlations between genes and principal components[54]. A subset of 1.24 million SNPs distributed across both intragenic and intergenic regions on all 10 chromosomes was used to perform PCA for both GPWAS and GWAS. The first three PCs were calculated using R prcomp function and included in GPWAS analysis. Let α_{in} and α_{out} be the criterion thresholds for the p-values of the phenotypes. If a phenotype with p-value smaller than α_{in}, we consider it as potentially significant and should be added into the regression model. Whereas, if the p-value of an existing phenotype in the model is larger than α_{out}, we consider it as insignificant and exclude it from the model. As a default, we choose $\alpha_{in} = \alpha_{out} = 0.01$ for each gene.

The stepwise selection procedure is as follows:

1. Start with the multi-response model with all the SNPs as responses and the first three PC scores as covariate. Search for the the most significant phenotype across all the phenotype measurements. Include this phenotype into the model if its p-value is below α_{in}. Otherwise, declare no phenotype is significant for this gene.

2. For the ℓth step, add each one of the remaining phenotypes into the existing model with the covariates that have already been selected, and calculate its p-value. This p-value measures the effect of this phenotype on the responses given all the selected phenotypes from the previous steps.

3. Find the remaining phenotype with the minimum p-value. Include this phenotype into the model if its p-value is below α_{in}. Otherwise, declare no new add in.

4. The newly added covariate may be correlated with the existing covariates in the model. This may change their corresponding p-values. Fit all the selected phenotypes jointly in the model and drop the phenotype with the largest p-value that is greater than the cutoff value α_{out}.

5. Repeat steps (2), (3) and (4) until no phenotypes can be added or removed from the model. This is considered as the final model for the targeted gene.

The final model can be represented as:

$$g_{k,i} = PC_{k,1}\beta_{i1} + PC_{k,2}\beta_{i2} + PC_{k,3}\beta_{i3} + \sum_{j=1}^{v_i} Phe_{k,(j)}\tau_{i(j)} + \varepsilon_{k,ij}. \quad (1)$$

Here, the subscript k and i represent the kth observation and the ith gene, respectively. There are v_i selected phenotypes for the ith gene, where $v_i \leq 260$. The selected phenotypes $\{Phe_{k,(j)}\}$ are a subset of the collection of all the phenotypes $\{Phe_{k,1}, Phe_{k,2}, \ldots, Phe_{k,260}\}$, where $\tau_{i(j)}$ is the corresponding coefficients for the selected phenotype $Phe_{k,(j)}$ of the ith gene. The first three PC scores PC_1, PC_2 and PC_3 were always included in the model with their effects β_{i1}, β_{i2} and β_{i3}. Note that $g_{k,i}$, $\beta_{i1}, \beta_{i2}, \beta_{i3}$ and $\tau_{i(j)}$ could be vectors corresponding to the multiple SNPs within the ith gene. Total phenotypes was iteratively selected for 35 times for each scanned gene. All the unselected phenotypes were considered as insignificant for a particular gene. The p-value of each gene was determined by the partial F test through comparing the final model containing both the first three PCs and the selected phenotypes with the initial model containing only the PCs. Of 32,084 gene models, genomic data of every 200 genes was extracted and submitted to cluster (Intel Xeon E5-2670 2.60GHz 2 CPU/16 cores per node) in Holland Computing Center at University of Nebraska-Lincoln for processing GPWAS with input phenotypes.

FDR cut offs of partial F-test were based on the results from 20 permutation analysis where the values for each trait were independently shuffled among the 277 genotyped individuals and the entire GPWAS pipeline rerun for all genes. The code implementing the above analyses in R and associated documentation has been published as the "GPWAS" which is avaliable from the following link: https://github.com/shanwai1234/GPWAS. Selected significant GPWAS genes with incorporated phenotypes were listed in Supplementary Table S8.

GWAS Analysis

GLM GWAS analyses were conducted using the algorithm first defined by Price and coworkers[31] and FarmCPU GWAS with the algrothm defined by Liu and colleagues[32]. Both algorithms were run using the R-based software rMVP (A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool For Genome-Wide Association Study) (https://github.com/XiaoleiLiuBio/rMVP). Both analysis methods were run using maxLoop = 10 and the variance component method method.bin = "Fast-LMM"[55]. The first three principal components were considered as additional covariates for population structure control. For comparison to GPWAS results, each gene was assigned the p-value of the single most significant SNP among all the SNPs assigned to that gene across 260 analyzed phenotypes in the GWAS model.

Nested Association Mapping Comparison

Published associations identified for 41 phenotypes scored across 5,000 maize recombinant inbred lines were retrieved from Panzea ($http://cbsusrv04.tc.cornell.edu/users/panzea/download.aspx?filegroupid=14$)[27]. Following the thresholding proposed in that paper a SNP and CNV (copy number variant) hits with a resample model inclusion probability ≥ 0.05 which were either within the longest annotated transcript for each gene AGPv2.16 or within 15kb upstream or downstream from the annotated transcription start and stop sites were assigned to that gene. Gene models were converted from B73 RefGenV2 to B73 RefGenV4 using a conversion list published on MaizeGDB (https://www.maizegdb.org/search/gene/download_gene_xrefs.php?relative=v4).

Gene Expression Analysis

Raw reads from the a published maize expression atlas generated for the inbred B73 were downloaded from the NCBI Sequence Read Archive PRJNA171684[39]. Reads were trimmed using Trimmomatic-0.38 with default setting parameters[56]. Trimmed reads were aligned to the maize B73 RefGenv4 reference genome using GSNAP version 2018-03-25[57]. Alignment results were converted to sorted BAM file format using Samtools 1.6[58] and Fragments Per Kilobase of transcript per Million mapped reads

(FPKM) where calculated for each gene in the AGPv4.39 maize gene models in each sample using Cufflinks v2.2[59]. Only annotated genes located on 10 maize pseudomolecules were used for downstream analyses and the visualization of FPKM distribution.

Ka/Ks Calculations

For each gene listed in a public syntenic gene list,[60], the coding sequence for the single longest transcript per locus was downloaded from Ensembl Plants and aligned to the single longest transcript of genes annotated as syntenic orthologs in Sorghum bicolor in v3.1[61] and Setaria italica v2.2[62] were retrieved from Phytozome v12.0 using a codon based alignment as described previously[43]. The calculation of the ratio of the number of nonsynonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) was automated using in-house constructed software pipeline posted to github (https://github.com/shanwai1234/Grass-KaKs). Genes with synonymous substitution rate less than 0.05 were excluded from the analyses as the extremely small number of synonymous substitutions tended to produce quite extreme Ka/Ks ratios and genes with multiple tandem duplicates were also excluded from Ka/Ks calculations. Calculated Ka/Ks ratios of maize genes were provided in Supplementary Table S9.

Presence/Absence Variation (PAV) Analysis

PAV data was downloaded from a published data file[42]. Following the thresholding proposed in that paper, a gene was considered to exhibit presence absence variance if at least one inbred line with coverage less than 0.2.

Gene Ontology Enrichment Analysis

All GO analyses used the maize-GAMER GO annotations for B73 RefGenV4 gene models[44]. Statistical tests for GO term enrichment and purificiation were performed using the goatools software package[63] with support for the Fisher Exact test provided by the fisher_exact function in SciPy. For determining median information content individual GO terms, each GO term was assigned a score based on the total number of gene models this GO term was assigned to in the maize-GAMER dataset. This analysis considered only gene models a GO term was specifically applied to in the dataset, but not gene models where the assignment of the GO term may have been implied by the assignment of a child GO term.

Power and FDR evaluation of GPWAS and GWAS using simulated data

SNP calls for the entire set of 1,210 individuals included in Maize HapMap3 were retrieved from Panzea[26], filtered, imputed, and assigned to genes as described above resulting in 1,648,398 SNPs assigned to annotated gene body regions in B73 RefGenV4. 2,000 randomly selected genes associated with 30,547 SNP markers were employed for downstream simulations. In each simulation, 100 genes (5%) were selected as causal genes. For each causal gene in each simulation, a causal SNP was selected for simulating phenotypic effects. A total of 100 phenotypes were simulated in each permutation of the analysis, with 10 traits simulated with heritability of 0.7, 30 traits simulated with heritability of 0.5 and 60 traits simulated with heritability of 0.3. Effect sizes for each SNP for each phenotype in each permutation were drawn from a normal distribution centered on zero using the additive model in GCTA (version 1.91.6)[64].

The resulting simulated trait data and genuine genotype calls were analyzed using GLM GWAS, FarmCPU GWAS, and GPWAS as described above with the exception of calculating population structure principal components using a sample (1% or 157,748 SNPs) of the total SNPs remaining after filtering, rather than only the subset of SNPs assigned to the 2,000 randomly selected genes included in this analysis. For each analysis, the set of 2,000 genes was ranked from most to least statistically significant based on the significance of the single most significantly associated SNP (for GLM and FarmCPU GWAS) or the significance of the overall model fit relative to a population structure only model (for GPWAS). Power evaluation for GPWAS was defined as the number of true positive genes to the total number of causal genes and FDR was defined as the number of false positive genes to the total number of positive genes. Power and FDR were calculated in a step size of five genes from 5 total positive genes to 500 (i.e. $\{5,10,...,450,500\}$).

References

1. Sax, K. The association of size differences with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genetics* **8**, 552 (1923).
2. Sprague, G. The location of dominant favorable genes in maize by means of an inversion. *Genetics* **26**, 143–149 (1941).
3. Klein, R. J. *et al.* Complement factor h polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
4. DeWan, A. *et al.* Htra1 promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989–992 (2006).
5. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature* **465**, 627 (2010).

6. Denny, J. C. *et al.* Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

7. Pendergrass, S. *et al.* The use of phenome-wide association studies (phewas) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. epidemiology* **35**, 410–422 (2011).

8. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. biotechnology* **31**, 1102 (2013).

9. Shameer, K. *et al.* A genome-and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. genetics* **133**, 95–109 (2014).

10. Lu, Y. *et al.* Systems genetic validation of the snp-metabolite association in rice via metabolite-pathway-based phenome-wide association scans. *Front. plant science* **6**, 1027 (2015).

11. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. genetics* **44**, 1066 (2012).

12. O'Reilly, P. F. *et al.* Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one* **7**, e34861 (2012).

13. Van der Sluis, S., Posthuma, D. & Dolan, C. V. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS genetics* **9**, e1003235 (2013).

14. Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PloS one* **8**, e65245 (2013).

15. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. genetics* **44**, 821 (2012).

16. Wang, Y. *et al.* Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genet. epidemiology* **39**, 259–275 (2015).

17. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using mtag. *Nat. genetics* **50**, 229 (2018).

18. Pendergrass, S. A. *et al.* Phenome-wide association study (phewas) for detection of pleiotropy within the population architecture using genomics and epidemiology (page) network. *PLoS genetics* **9**, e1003087 (2013).

19. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-wide association studies as a tool to advance precision medicine. *Annu. review genomics human genetics* **17**, 353–373 (2016).

20. Walter, A., Liebisch, F. & Hund, A. Plant phenotyping: from bean weighing to image analysis. *Plant methods* **11**, 14 (2015).

21. Wen, W. *et al.* Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. communications* **5**, 3438 (2014).

22. Matsuda, F. *et al.* Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *The Plant J.* **81**, 13–23 (2015).

23. Diepenbrock, C. H. *et al.* Novel loci underlie natural variation in vitamin e levels in maize grain. *The Plant Cell* tpc–00475 (2017).

24. Kremling, K. A. *et al.* Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520 (2018).

25. Flint-Garcia, S. A. *et al.* Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant J.* **44**, 1054–1064 (2005).

26. Bukowski, R. *et al.* Construction of the third-generation zea mays haplotype map. *GigaScience* **7**, gix134 (2017).

27. Wallace, J. G. *et al.* Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS genetics* **10**, e1004845 (2014).

28. Schnable, J. C. & Freeling, M. Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PloS one* **6**, e17855 (2011).

29. Zhao, W. *et al.* Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.* **34**, D752–D757 (2006).

30. Dahl, A. *et al.* A multiple-phenotype imputation method for genetic studies. *Nat. genetics* **47**, 466 (2015).

31. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. genetics* **38**, 904 (2006).

32. Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS genetics* **12**, e1005767 (2016).

33. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. genetics* **43**, 159 (2011).

34. Namjou, B. *et al.* Phenome-wide association study (phewas) in emr-linked pediatric cohorts. *Front. genetics* **5**, 401 (2014).

35. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).

36. Bensen, R. J. *et al.* Cloning and characterization of the maize an1 gene. *The Plant Cell* **7**, 75–84 (1995).

37. Brown, P. J. *et al.* Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS genetics* **7**, e1002383 (2011).

38. Consortium, I. H. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299 (2005).

39. Stelpflug, S. C. *et al.* An expanded maize gene expression atlas based on rna sequencing and its use to explore root development. *The plant genome* **9** (2016).

40. Walley, J. W. *et al.* Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).

41. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci.* **108**, 4069–4074 (2011).

42. Brohammer, A. B., Kono, T. J., Springer, N. M., McGaugh, S. E. & Hirsch, C. N. The limited role of differential fractionation in genome content variation and function in maize (zea mays l.) inbred lines. *The Plant J.* **93**, 131–141 (2018).

43. Zhang, Y. *et al.* Differentially regulated orthologs in sorghum and the subgenomes of maize. *The Plant Cell* tpc–00354 (2017).

44. Wimalanathan, K., Friedberg, I., Andorf, C. M. & Lawrence-Dill, C. J. Maize go annotation—methods, evaluation, and review (maize-gamer). *Plant Direct* **2**, e00052 (2018).

45. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. genetics* **43**, 1160 (2011).

46. Castelletti, S., Tuberosa, R., Pindo, M. & Salvi, S. A mite transposon insertion is associated with differential methylation at the maize flowering time qtl vgt1. *G3: Genes, Genomes, Genet.* **4**, 805–812 (2014).

47. Rodgers-Melnick, E., Vera, D. L., Bass, H. W. & Buckler, E. S. Open chromatin reveals the functional maize genome. *Proc. Natl. Acad. Sci.* **113**, E3177–E3184 (2016).

48. Piepho, H., Möhring, J., Melchinger, A. & Büchse, A. Blup for phenotypic selection in plant breeding and variety testing. *Euphytica* **161**, 209–228 (2008).

49. Kusmec, A., Srinivasan, S., Nettleton, D. & Schnable, P. S. Distinct genetic architectures for phenotype means and plasticities in zea mays. *Nat. plants* **3**, 715 (2017).

50. Bennetzen, J. L., Coleman, C., Liu, R., Ma, J. & Ramakrishna, W. Consistent over-estimation of gene number in complex plant genomes. *Curr. opinion plant biology* **7**, 732–736 (2004).

51. Gerstein, M. B. *et al.* What is a gene, post-encode? history and updated definition. *Genome research* **17**, 669–681 (2007).

52. Schnable, J. C. Genome evolution in maize: from genomes back to genes. *Annu. Rev. Plant Biol.* **66**, 329–343 (2015).

53. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *The Am. J. Hum. Genet.* **98**, 116–126 (2016).

54. Zhang, J. *et al.* plarmeb: integration of least angle regression with empirical bayes for multilocus genome-wide association studies. *Heredity* **118**, 517 (2017).

55. Lippert, C. *et al.* Fast linear mixed models for genome-wide association studies. *Nat. methods* **8**, 833 (2011).

56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

57. Wu, T. D. & Nacu, S. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).

58. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

59. Trapnell, C. *et al.* Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nat. protocols* **7**, 562 (2012).

60. Schnable, J. C. Sorghum version 3, maize versions 3 and 4 syntenic gene list. *FigShare* .

61. McCormick, R. F. *et al.* The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant J.* **93**, 338–354 (2018).

62. Bennetzen, J. L. *et al.* Reference genome sequence of the model plant setaria. *Nat. biotechnology* **30**, 555 (2012).

63. Klopfenstein, D. *et al.* Goatools: A python library for gene ontology analyses. *Sci. reports* **8**, 10872 (2018).

64. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *The Am. J. Hum. Genet.* **88**, 76–82 (2011).

65. Peng, J. & Harberd, N. P. The role of ga-mediated signalling in the control of seed germination. *Curr. opinion plant biology* **5**, 376–381 (2002).

Acknowledgements

This work is supported by the Quantitative Life Sciences Initiative at the University of Nebraska-Lincoln, which receives support from a University of Nebraska Program of Excellence and by the National Science Foundation Awards MCB-1838307 and OIA-1826781 to JCS. The authors thank Andy Dahl advice and instruction in the use of phenotype imputation, Zheng Xu and Wenlong Ren for consultation on the design of the association study, and the PanZea project (http://www.panzea.org) for gathering the phenotypic and genotypic data employed in this study. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

Supplementary Information

Genetic marker data was obtained from resequencing data of 277 inbreds from the Buckler-Goodman maize association panel which was published as part of the maize HapMap3 project[25,26]. Maize HapMap3 contains data for a total of 81,687,392 SNPs, however, after removing SNPs with high missing data, those which were not polymorphic among the 277 specific individuals employed here, and a number of other quality filtering parameters, a total of 12,411,408 SNPs remained, of which 1,904,057 SNPs were assigned to 32,084 annotated gene models in the B73 RefGenV4 genome release (See Methods). Filtering to eliminate redundancy between SNPs in high LD with each other assigned to the same gene, it further reduced this number to 557,968 unique SNPs.

A phenotypic dataset consisting of 57 specific traits scored for the Buckler-Goodman maize association panel across one to sixteen distinct environments for a total of 285 unique phenotypic datasets was obtained from Panzea[29]. Filtering to remove phenotypic datasets with extremely high missing data rates left a total of 260 trait datasets with a median missing data rate of 18%. Of the total 72,020 potential trait datapoints (277 inbreds $\times$ 260 traits) 23.6% or 16,963 trait datapoints were unobserved.

Similar GO analysis results were obtained for FarmCPU even after controlling for total number of genes identified (see Supplementary Information). Analysis using the 706 genes uniquely detected genes by FarmCPU GWAS found only a single significantly enriched GO term GO:0009987 ”cellular process”, while, even when the number of uniquely GPWAS identified genes was constrained to be identical to the number of uniquely identified FarmCPU GWAS genes, 67 GO terms still showed significant enrichment (58 terms) or purification (9 terms) (Figure 3b, Supplementary Table S6).

Supplementary Table 1: 260 phenotypes employed in this study with corresponding missing data rate, imputation accuracy and classified phenotype group.

Supplementary Table 2: Expressed genes and expression breath of different gene populations.

Supplementary Table 3: Gene length and SNP density in each gene population.

Supplementary Table 4: Correlation between significant level and SNP number per gene of genes generated from permuted and real data in GPWAS and GLM GWAS.

Supplementary Table 5: Conservation features for unique gene sets between FarmCPU GWAS and GPWAS.

Supplementary Table 6: GO terms enriched and purified in each gene population.

Supplementary Table 7: Statistics of GO terms assigned to each gene population.

Supplementary Table 8: Selected significant genes with incorporated phenotypes in GPWAS model.

Supplementary Table 9: Ka/Ks value per gene in maize version 4.

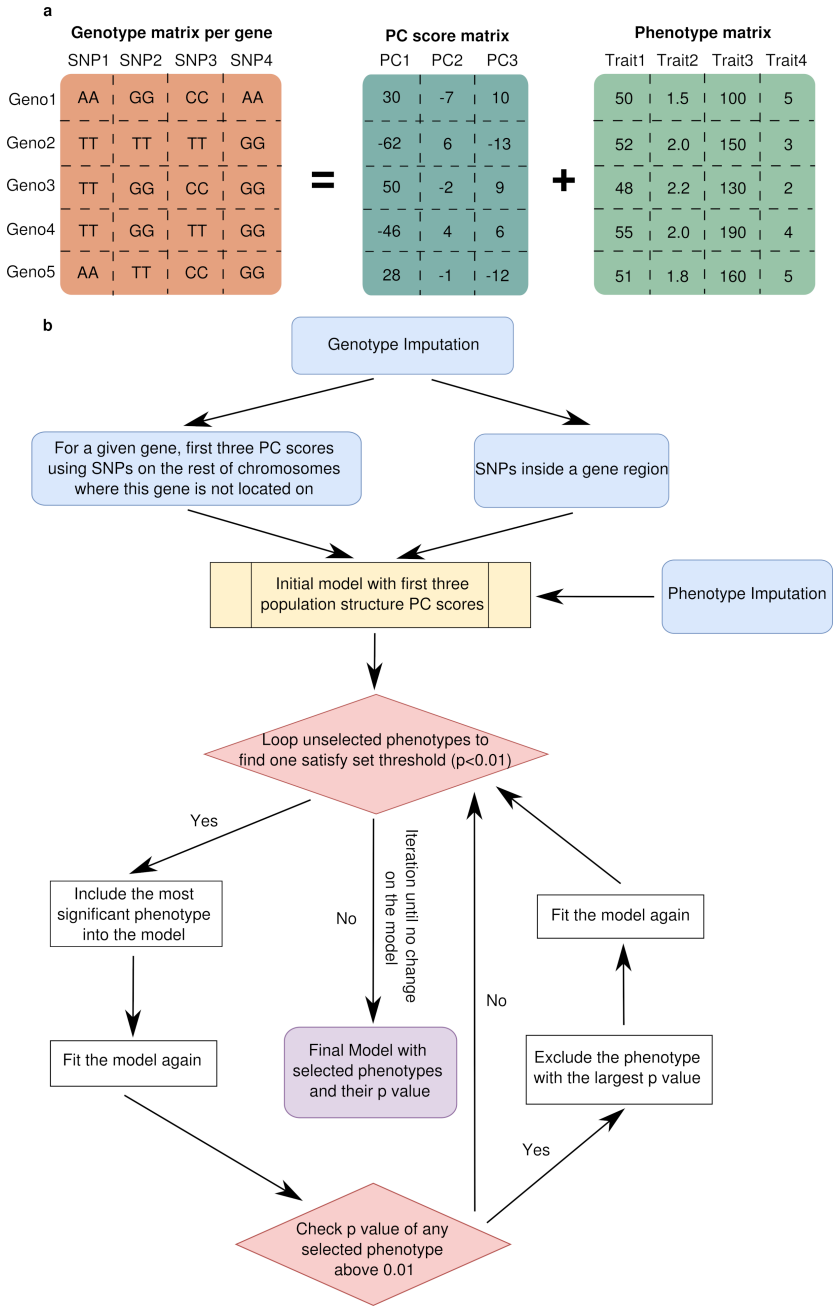

Figure S1. Description of GPWAS algorithm implementation (a) Example of trait and genotype matrices employed for GPWAS. (b) Flow chart schematization of the GPWAS algorithm.

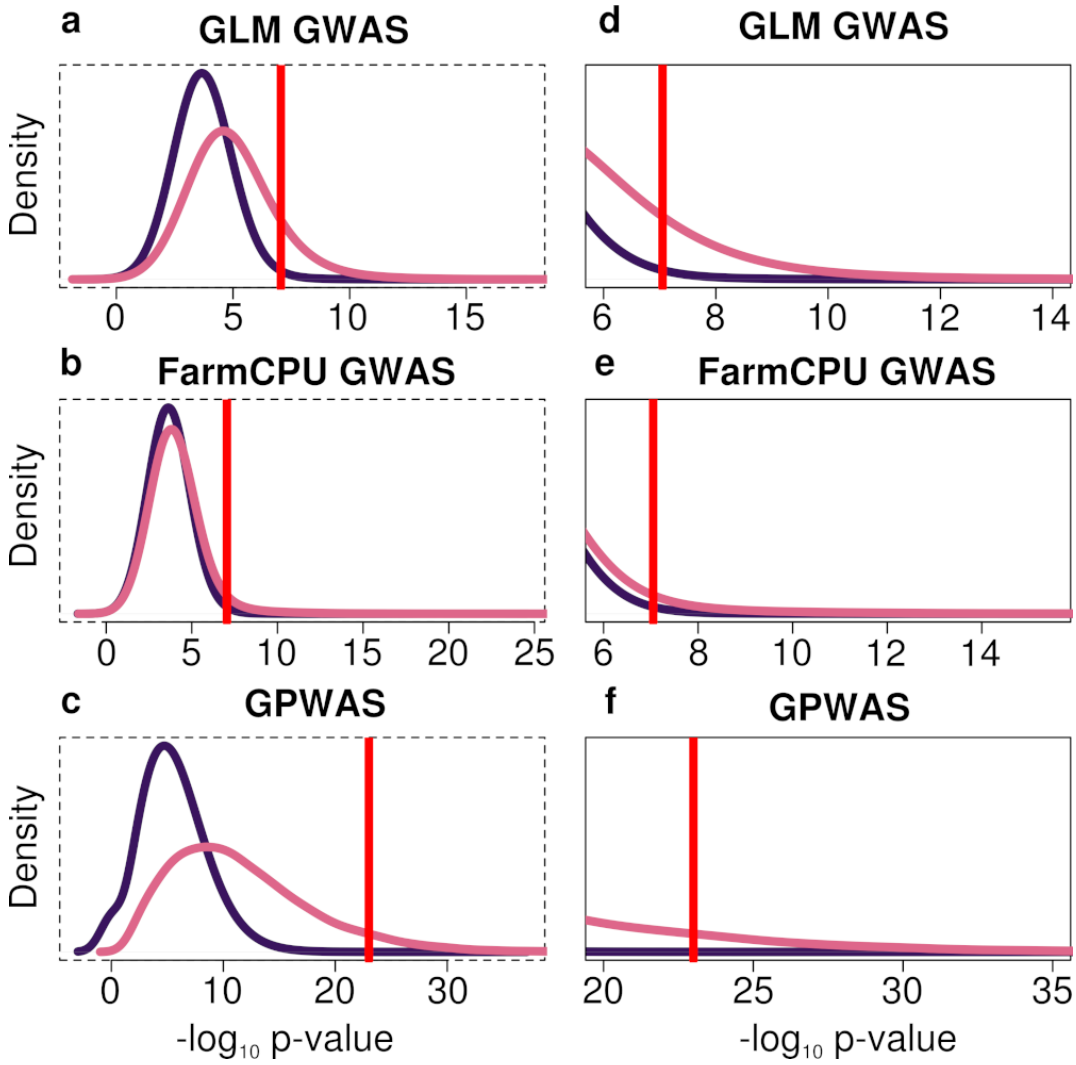

Figure S2. Permutation testing based estimation of false discovery rates for GLM GWAS, FarmCPU, and GPWAS. For each panel, the dark curve shows the distribution of per gene p-values obtained from twenty permutations of genotype and trait data (See Methods), while the light curve indicates the distribution of per gene p-values obtained from the analysis of the non-permuted dataset. Red lines indicate the p-value analyses employed in these analysis, corresponding to p=8.96e-8 for GLM and FarmCPU, calculated based on 557,968 SNPs and p=1.00e-23 calculated based on on a target FDR < 0.001 for GPWAS. Genes on the right side of red line were pulled for analyzes. An equivalent FDR cut off for GLM GWAS would require an uncorrected p-value cut off of approximate 1e-14, and 31 genes would remain statistically significantly associated with traits at this threshold. For FarmCPU GWAS, the minimum FDR achieved was FDR < 0.029 at a p-value threshold of 1e-15, with 38 genes remaining statistically significantly associated with traits at this threshold. Left Panels (a-c) show the entirely of the distributions, while right panels (d-f) display a zoomed in view of the regions of the curve where the p-value threshold employed.

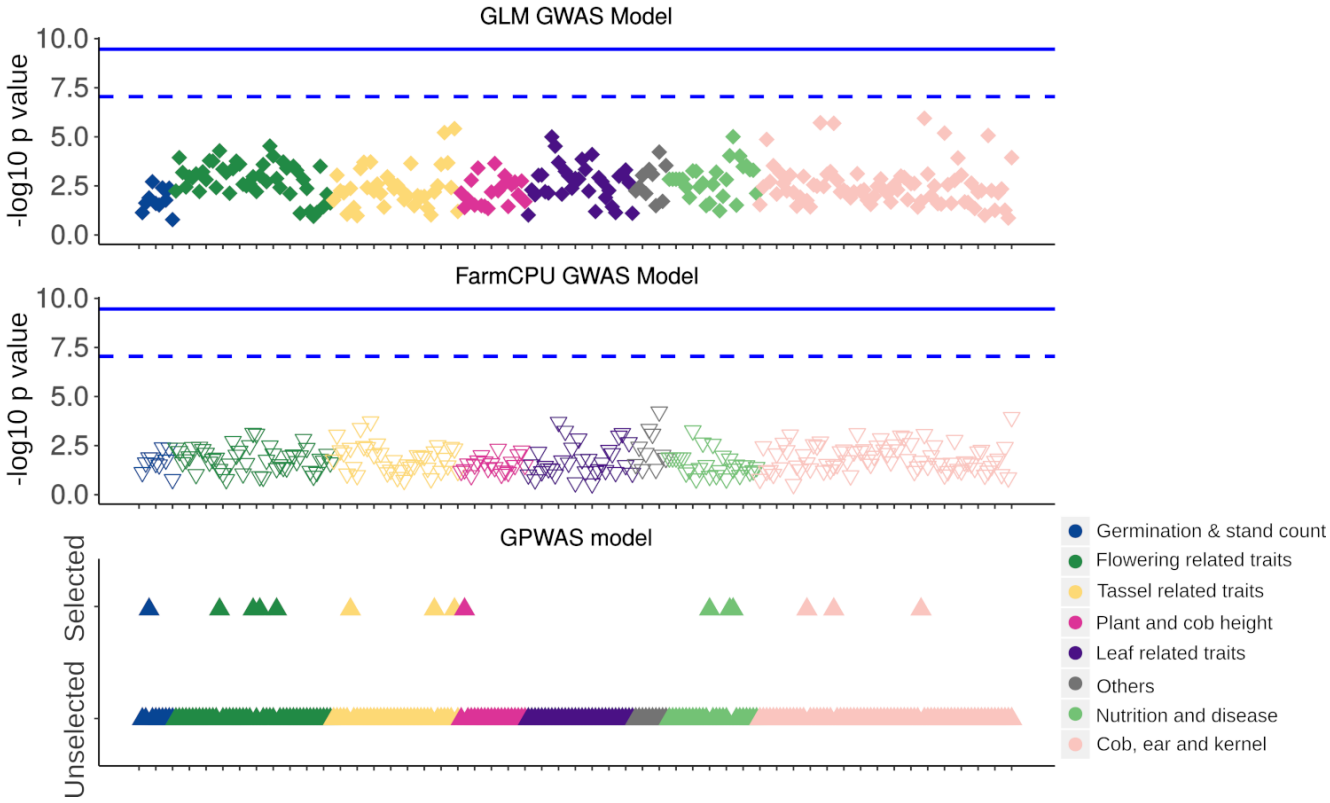

Figure S3. The power of GLM GWAS model, FarmCPU GWAS model and GPWAS model on detecting maize Anther ear1 ((*an1*)) gene (Zm00001d032961). Dashed line is p value after Bonferroni correction with 557,968 (SNPs) hypothesized testings. Solid lines in GLM GWAS and FarmCPU are stricter Bonferroni corrected p-value with their original number of hypothesis multiplied by the number of phenotypes (260) tested. Scales on ytick labels are -log10 p values. Sel. and Uns. standed for phenotypes selected and unselected by the GPWAS model. Phenotypes incorporated in the GPWAS model are Germination count (Summer 2006, Johnston, NC), Days to Tassel (Summer 2007, Cayuga, NY; Summer 2007, Johnston, NC), GDDDays to silk (Winter 2006, Miami-Dade, FL), Tassel Length (Summer 2007, Cayuga, NY), Spikelets Primary Branch (Summer 2006, Champaign, IL), Secondary Branch Number (Summer 2006, Boone, MO), Plant Height (Summer 2006, Cayuga, NY), NIR measured protein (Summer 2006, Johnston, NC), NIR measured oil (Summer 2006, Johnston, NC; Winter 2006, Miami-Dade, FL), Cob weight (Summer 2007, Johnston, NC), Ear diameter (Summer 2007, Johnston, NC) and Total kernel volume (Summer 2006, Cayuga, NY). Among them, flowering time, plant height, and tassel branch number are all consistent with the known mutant phenotype[36]. Germination count (Summer 2006, Johnston, NC) was also identified as part of the model for the an1 gene in the GPWAS analysis. While there are no published reports of altered germination in the an1 mutant, such a phenotype would be consistent with the role of an1 in gibberellic acid metabolism[65]. Every five phenotypes were added a xtick label.

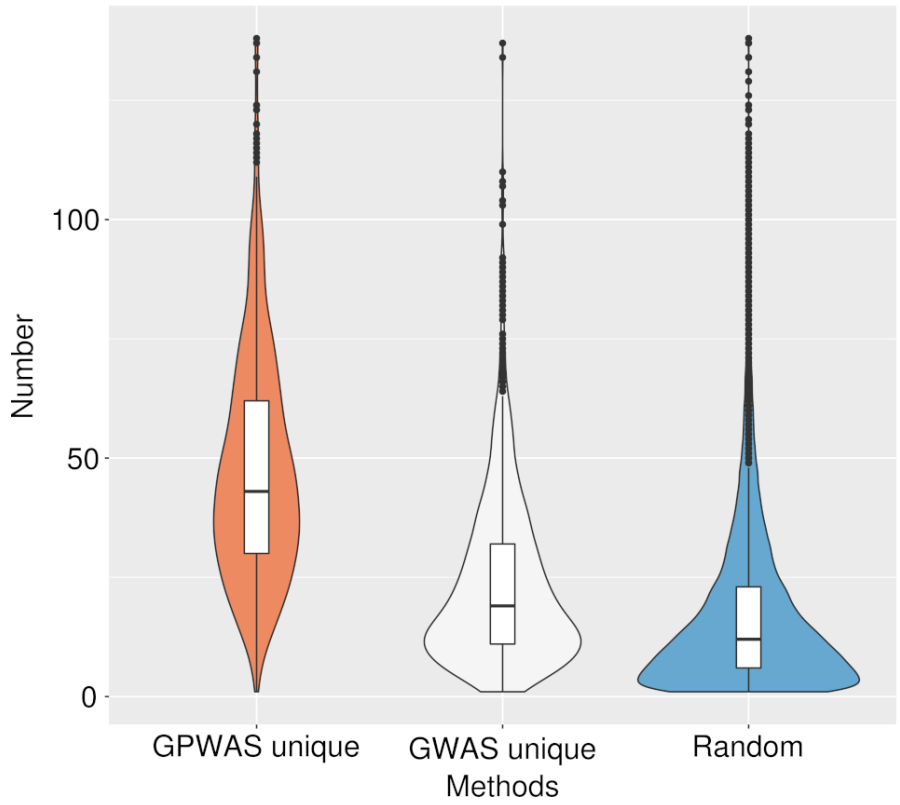

Figure S4. Numbers of SNPs per gene in uniquely identified genes by GPWAS, uniquely identified genes by GLM GWAS, and total (random) genes with identified SNPs. ($p<2.2e-16$ between uniquely identified genes by GLM GWAS and between uniquely identified genes by GPWAS, Mann–Whitney U test).

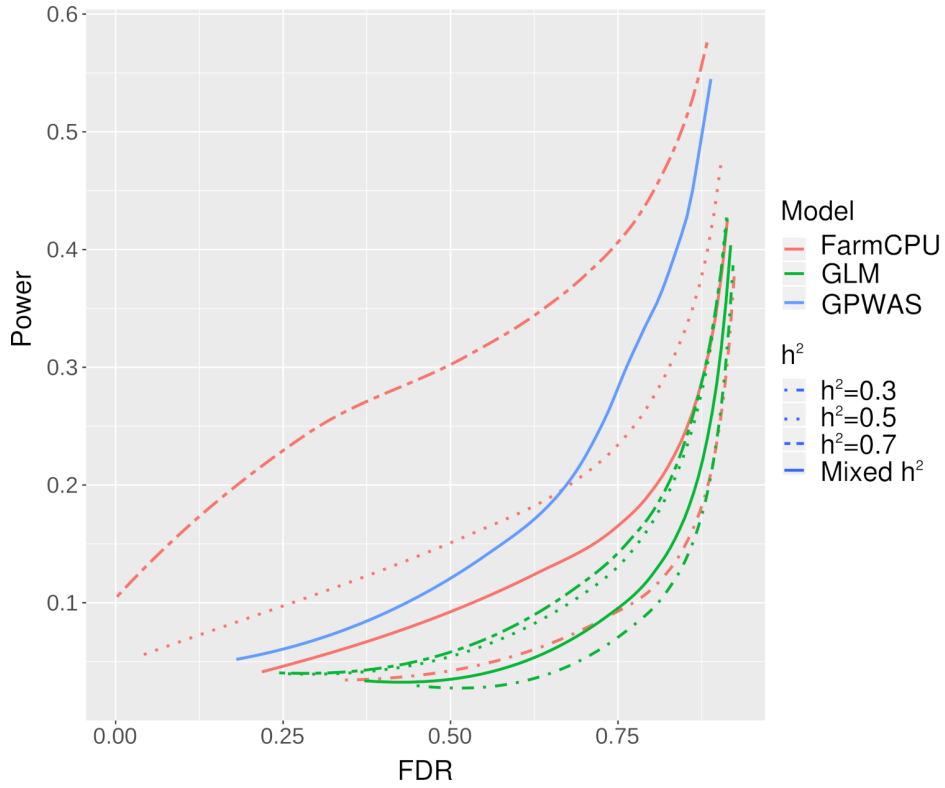

Figure S5. Power and FDR evaluation of GPWAS model with GLM and FarmCPU GWAS models on simulated phenotypes from variable heritabilities. Ten random sets of 100 QTNs were used to simulate 100 replicated phenotypes with 10% h^2 as 0.7, 30% h^2 as 0.5 and 60% h^2 as 0.3. For one simulated phenotype set, positive genes were defined as top m ranked significant genes of 2,000 genes. Ratios of power to FDR in GWAS models were calculated as the mean value of total simulated phenotypes under different heritabilities (h^2) in each rank, while these ratios were calculated using all 100 simulated phenotypes in GPWAS model in each rank.