Summer 8-1-2019

New Approaches to Use Genomics, Field Traits, and High-throughput Phenotyping for Gene Discovery in Maize (*Zea mays*)

Zhikai Liang
University of Nebraska-Lincoln

New Approaches to Use Genomics, Field Traits, and High-throughput Phenotyping for Gene Discovery in Maize (*Zea mays*)

by

Zhikai Liang

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy

Under the Supervision of Professor James C. Schnable

Lincoln, Nebraska

August, 2019

New Approaches to Use Genomics, Field Traits, and High-throughput Phenotyping for Gene Discovery in Maize (*Zea mays*)

Zhikai Liang, Ph.D.

University of Nebraska, 2019

Adviser: James C. Schnable

Maize is one of major crop species over the world. With lots of genetic resources and genomic tools, maize also serves as a model species to understand genetic diversity, facilitate the development of trait extraction algorithms and map candidate functional genes. Since the first version of widely used B73 reference genome was released, independent research groups in the maize community propagated seeds themselves for further research purposes. However, unexpected or occasional contamination may happen during this process. The first study in this thesis used public RNA-seq data of B73 from 27 research groups across three countries for calling single nucleotide polymorphisms (SNP). Those SNPs were applied for investigating the distance of 27 maize B73 samples from the reference genome and three major clades were defined for determining their original sources. On the other side, maize is a plant with clear plant architecture. The second study was to employ the high-throughput plant phenotyping to dissect plant phenotypes using computer vision methods. A total of 32 maize inbreds distributed from the Genomes to Fields project were captured images in daily by 4 types of cameras (RGB, Hyperspectral, Fluorescence and Thermal-IR) for approximate 1 month. Differences between computer vision measurements and manual measurements about the plant fresh biomass were evaluated. Broad-sense heritability was estimated for extracted measurements from images. The expanded types of plant phenotype from the perspective of imaging provided a broader range of opportunities for connecting phenotypic variants with genetic variants. The third study utilized the phenome-wide variants in maize Goodman-Buckler 282 association panel to scan and associate with

genetic variants of annotated genes along the maize genome. Genes detected by the proposed model, Genome-Phenome Wide Association Study (GPWAS), are significantly different from conventional GWAS detected genes. GPWAS genes tend to be more functionally conserved and more similar as classical maize mutants with known functions. Results from these researches assist to answer question about the genetic purity of same maize genotype. Methods developed in this thesis can also provide the valuable reference for trait discoveries from images and candidate functional gene identification using a broad set of phenotypes.

ACKNOWLEDGMENTS

I would like to express my honest gratitude to my academic advisor Dr. James C. Schnable. These works could not be completed without his patient guidance and insightful directions. He not only enlightened me on my interested research directions, but also widened my researches from multiple perspectives. My doctoral committee members: Dr. Stephen Baenziger, Dr. David Holding and Dr. Jennifer Clarke gave me precious comments during my PhD program and committee meetings. My collaborator, Dr. Yumou Qiu, suggested me statistical approaches in analyzes for my several research projects. Dr. Yufeng Ge and Piyush Pandey from Department of Biological and Agricultural Engineering in UNL demonstrated me ideas for image analysis. Dr. Jinliang Yang from Department of Agronomy and Horticulture suggested me methods in quantitative genetics. I would also acknowledge all past and current lab members in Dr. James Schnable's lab for valuable discussions and ideas for inspiration. My mentored undergraduate, Thomas Hoban, assisted me on processing experiments. Greenhouse manager and staff took care of my plants and enabled my projects to go smoothly. Various funding sources from University of Nebraksa-Lincoln supported my study during my PhD program.

I appreciated the honest friendship in two states I have lived in, Mississippi and Nebraska, since I came to US. The cat that I raised temporarily gave me endlessly wonderful memory. Beautiful places I have traveled during my PhD in US relaxed me and made my life enjoyable.

My wife, Dr. Xiaoxi Meng, who was always there for me and accompanied with me during the whole journey of my PhD study, encouraged and motivated me to achieve the degree eventually. My parents supported my every decision in last several years. My big family in China gave me the kindest care all the time. Those people firmly supported me to successfully accomplish my dissertation.

PREFACE

The research discussed in Chapter 2 using RNA-seq data to investigate the genetic identity of maize inbred B73 has been published in PloS one. (Liang Z and Schnable J.C., 2016. PloS one, 11(6), p.e0157942)

The high-throughput plant phenotyping experiment and methods for processing maize images discussed in Chapter 3 has been published in Gigascience. (Liang, Z., Pandey, P., Stoerger, V., Xu, Y., Qiu, Y., Ge, Y. and Schnable, J.C., 2017. GigaScience, 7(2), p.gix117)

The theoretical model of Genome-Phenome Wide Association Study (GPWAS) and its comparison with conventional Genome-Wide Association Study (GWAS) model written in Chapter 4 has been published in preprint sever bioRxiv and is under preparation for publication in official scientific journal. Liang, Z., Qiu, Y. and Schnable, J.C., 2019. bioRxiv, p.534503)

Table of Contents

List of Figures

List of Tables

1:Literature Review

1.1 Introduction

The origins of maize genetics research can be traced back to Rollins A. Emerson in the 1900's. One of the reasons maize emerged as an early genetic model is that maize produces separate male and female flowers on separate reproductive structures, which makes manual controlled crosses much more practical on a large scale than in many species. As a result, a large number of progeny can be produced from an pair of parents and complex studies of complementation, epistasis, and quantitative genetics are particularly feasible in maize. Barbara McClintock, the winner of the 1983 Nobel Prize in Physiology or Medicine, was recognized by her discovery of transposable elements in maize.[1] However, maize has also served as a model for addressing many other biological questions, particularly in the fields of genetics, genome biology, selection, and evolution. The closely evolutionary distances among maize and other grass species result from shared conserved genomic regions, which enable syntenic analyses.[2] Because of an ancient whole genome duplication in maize, there are generally two co-orthologous syntenic regions in maize which correspond to single regions in related grass species like rice, foxtail millet, or sorghum.[3–5] With added comparable species, this boosts the statistical power to detect shared conserved information across species such as syntenic genes,[6] differentially regulated orthologs,[7] and conserved non-coding sequences.[8] At the population level, abundant phenotypic and genotypic datasets from maize diversity and association populations have been collected and stored in diverse public depositories (PanZea, Cyverse, NCBI, Gramene). For example, in the PanZea

(https://www.panzea.org/) database, agronomic traits such as grain yield, plant height and flowering time have been measured and recorded across different environments for individual genotypes in the maize 282 association panel.[9] When combined with published high density genotypic datasets, these resources enable researchers to connect phenotype with genotype and determine potentially causal loci for traits of interest using algorithms such as GWAS (Genome-Wide Association Study). Identifying candidate loci/genes may narrow down the total annotated genes from ~40,000 to a much smaller range. However, utilizing these public resources requires grappling with the twin problems of missing data, and the potential for mislabeling or inconsistent genetics of the same genetic line in different environments.

The thesis consists of three studies: 1) Using RNA-seq data to investigate the genetic purity and consistency of the maize inbred B73, which is widely used in maize genetic studies in many countries; 2) Generation and processing of high-throughput phenotyping data from maize as a prelude to expanding the number of potential phenotypes which can be efficiently measured; 3) The description and evaluation of a new approach, the Genome-Phenome Wide Association Study (GPWAS) which I show can identify functionally conserved genes.

1.2 Maize is a model species in genetics studies

1.2.1 Reference genome

B73 is the most commonly used variety in the maize community and was first registered in 1972.[10] With more erect leaf architecture and superior performance in grain yield, B73 was broadly used as a parental line for generating new varieties.[10] The first version of the B73 genome assembly was completed in 2009.[11] This published genome accelerated genomics research in maize. Raw sequenced reads could be aligned against this reference genome to detect polymorphisms or infer gene expression levels in various maize varieties. Different genetic backgrounds could also produce alignment

mismatches or gaps. This could due to remaining heterozygous loci, large introgressions of exotic genome, repetitive regions and distant genetic relationship with reference genome. The first two issues could be avoided through careful investigation of the genetic purity of input materials. In the maize genome, around 85% of the genome consists of repetitive DNA.[11] Of these 85% repetitive DNA, more than 75% maize genome consists of LTR (Long Terminal Repeat) retrotransposons,[11] ranging from several hundreds of base pairs to tens of thousands of base pairs. The first version of maize genome assembly was completed by Sanger sequencing. To reduce the cost of sequencing, Illumina sequencing technology was developed to sequence millions of short fragments in parallel and achieves high-throughput of sequence generation. However, many genomic regions consist of repetitive sequences. Short reads will lead to high computational cost to reveal sequences in these regions. The development of single molecule sequencing extends the length of raw reads to more than 20kb.[12] Using single molecule technologies has produced better and more complete assemblies of heterochromatic regions as seen in the fourth release of the B73 reference genome based on PacBio sequencing.[13] To increase the alignment rate between sequenced data generated from diverse maize varieties and the reference genome, *de novo* genome assemblies of multiple maize varieties representing different heterotic groups have been produced[14–16] (Table 1.1; Data source: MaizeGDB). Clearly understanding the genetic distance between known samples and reference genome will boost the accuracy of downstream analyzes, such as SNP calling, transposonal elements detection and expression abundance determination per gene. Improved sequencing technologies have made feasible the genome assembly of the 26 NAM (Nested Association Mapping) founders feasible (a project currently being conducted by Matthew Hufford, an Associate Professor working on evolutionary genomics and population genetics at Iowa State University), seeks to capture as much of the diverse genetic background of maize varieties as possible. More sequenced genomes will provide opportunities for researchers

Table 1.1: Common maize inbred lines have been sequenced to date.

Sequenced Inbred	Genome Released Time	Sequencing Platform	Genome Coverage
B73 (AGPv1)[11]	2009	Sanger	4x-6x
CML247[17]	2016	Illumina	130x
PH207[14]	2016	Illumina	230x
B104 (Unpublished)	2017	Illumina	50x
Mo17[15]	2017	PacBio Illumina	>120x
B73 (AGPv4)[13]	2017	PacBio Illumina	60x
W22[16]	2018	Illumina	210x
SK[18]	2019	PacBio Illumina	166x

to identify high-confidence molecular markers with precise associations to phenotypes.

1.2.2 Molecular markers

Prior to the application of molecular markers to identify crop varieties, visible phenotypes were used as markers to distinguish different plants. The "father of modern genetics", Gregor Mendel, initially used visible traits such as pod color in garden peas to discover the genetic basis of inheritance.[19] Even in modern agricultural production, using visible traits to evaluate breeding lines is still the most direct and efficient method when selecting for straightforward traits. For example, PHW30 (a patent-off inbred) can be easily distinguished from Mo17 (Non-Stiff Stalk), because of the distinct leaf architecture.[20] However, the genome sequence itself of a specific variety is the most unique feature distinguishing from other varieties. Markers to dissect maize genotypes such as RFLP (Restriction Fragment Length Polymorphism), SSR (Simple Sequence Repeats) or AFLP (Amplified Fragment Length Polymorphism)[21] require running a great number of electrophoresis genes.

A SNP (Single Nucleotide Polymorphism) represents a specific genomic site within a population where two or more different nucleotides are present in different individuals

or haplotypes. Genotyping-By-Sequencing (GBS) involves digestion of DNA into small fragments using restriction enzymes in order to obtain reads covering identical genomic positions in each sequenced individual in a given population.[22] Using GBS technology, thousands of markers can be detected and imputed (i.e. LinkImpute,[23] Beagle[24]) in each inbred line in a given population. Genotypes of F1 hybrids can also be inferred using genotypes from parental lines.[25] Combined with recorded agronomic data, methods like genomic selection (GS) can be employed to speed up the breeding and selection process. However, the low coverage of GBS sequencing data limits its use in detecting variations along the genome (i.e. gene regions, regulation regions) as well as structural variation. Whole Genome Sequencing (WGS) provides much deeper coverage for studied samples. For example, in maize, the recently completed Hapmap3 project gathered more than 1,200 individuals and performed resequencing which lead to the identification and scoring of over 83 million polymorphisms along the chromosomes of maize.[26] This abundant resource of polymorphic markers in maize can be used in studies of phenomena such as genotype-phenotype associations and the identification of evolutionarily conserved sites in the genome.

1.2.3 Genetic sources

Maize is widely grown across various geographical locations and more than 13% of the world's total cropland is planted with maize.[27] Maize lines are often grouped into different categories such as NSS (Non-Stiff Stalk), SS (Stiff Stalk), TS (Tropical or Semitropical), sweet corn and popcorn.[28,29] To utilize this diversity a large number of populations have been generated by different research groups. To understand associations between genetic loci and investigated traits, one of several widely used approaches is to create bi-parental populations. Because of the segregation of large blocks of parental haplotypes into progeny and the high frequently of the parental haplotypes in the resulting population RIL populations are powerful to detect

co-segregation signals with investigated phenotypes. In general, a single cross is performed between two selected parental lines to produce F1 seeds. The segregation will then occur after self-pollinating the F1. The Single Seed Descent (SSD) method is used to propagate each single seed from F1 plants for the generation of RIL (Recombinant Inbred Line) populations. However, this biparental population has a limited number of generations for informative recombinations to occur and it can be hard to use these populations for mapping a source of phenotypic variation to a precise genomic region. This is where we stopped editing In order to break large haplotypes in a single line of RIL population, the IBM (Intermated B73 x Mo17) population was generated by randomly crossing F2 individuals with no prior phenotype targets.[30] However, the biparental population contains limited genetic variations and therefore only can map genes to a certain number of traits. The effort spent on natural genetic resources collection brings the opportunity of linking genomic variations with phenotypic variations at the single nucleotide level. Since the first version of the B73 reference genome was published, a broad set of applications of GWAS in maize has accelerated the process of revealing the genetic architectures in a wide range of traits.[17, 31–33] However, individuals in a GWAS population share genetic relatedness. Many subpopulations are both genetically distinct (i.e. have different allele frequencies for many markers) and have different average values for a wide range of traits. Failing to control for the population structure will lead to many false discoveries which actually associates with the population structure rather than the studied trait (Figure 1.1A, B). The Nested Association Mapping (NAM) population was designed to select 26 representative founders and produced 25 RILs (Recombination Inbred Lines) with B73 after generations of propagation for dissecting the genetic architecture of complex traits.[34] However, to generate and maintain this population, it requires a lot of effort. The progeny of each RIL shares pedigree from parental lines but still produces sub-population structure in the NAM population. To address the sub-population issue, the Multiple-parent Advanced-Generation Inter-Cross

(MAGIC) was developed to create higher chances of recombination through genomes using a multi-parental intermating strategy.[35] With different research purposes, there are a broad set of populations being generated. They are well maintained and stored by organizations such as USDA, CIMMYT and Scuola Superiore Sant Anna (IT). These materials can facilitate researches in the maize community.

1.3 Plant phenotyping

Plant genomes can be generated more efficiently than ever before. The number of different maize lines with complete resequencing data is expanding expontentially and the number of different maizde lines with independently assembled and annotated reference genomes is beginning to follow the same trajectory. As a result, the number of phenotypic measurements which can be realistically obtained is emerging as the new limitation for plant biologists when they seek to explain the function, if any, of specific genomic variants. Massive amounts of effort are invested in breeding new varieties mainly targeted at specific traits like grain yield, plant height, flowering time or stress resistance. However measuring these traits are time consuming and significant variance can be present in measurements of individual plants or plots, necessitating large replicated experiments with thousands or tens of thousands of individual measurements. Using an unified criterion for measuring a specific phenotype of plants is needed to standardize this process. On the other side, ways of defining and measuring phenotypes have been generally accepted by the community for a long time. Given that only around 1% of annotated genes have been functionally characterized in maize,[36] it is likely the way that we define a phenotype still has room for improvement. Other than traditional phenotypes, the integration of interdisciplinary technologies and collaborations bring opportunities to investigate phenotypic measurements, such as intermediate phenotypes across plant development of stages, plant images captured by broader wavelengths (i.e. hyperspectral cameras, infrared cameras), the same phenotype

measured in different environments (GxE interactions or plasticity[37,38] contributes the variation of the same trait for one genotype) and phenotypes at the molecular level (i.e. gene expression, metabolites, nutrient content). Accurate measurements of these traits provide a way to inspect genes having not only large effects but also subtle effects on the investigated trait. Thus, there is potential to discover previously unknown gene functions.

1.3.1 Automated and high-throughput plant phenotyping

High-throughput phenotyping (HTP) platforms are being developed to accurately measure dynamic phenotypes not easily measured manually before. These include traditional phenotypes such as seedling vigor, days to flowering time, and terminal plant height. In addition to being relatively straightforward to define and measure, these traits also have clear links to overall yield and plant performance. Given a population, QTL-mapping or GWAS is widely adapted to explain these phenotypic variations.[39–41] The development of HTP is represented in both controlled environments and field conditions. The utility of HTP can expand to phenotypes in higher dimensions. In controlled environments, these expanded dimensions are mainly represented in three aspects, 1) The single plant can be imaged from multiple side views, as well as the top view. The combination of images for these angles can evaluate plant phenotypes from 3-D dimensions, which is helpful to get a more accurate estimation of traits such as biomass; 2) HTP provides the way to trace plant development and capture images in time-lapse;[42,43] 3) Except for images taken by visible wavelengths, hyperspectral, multispectral or infrared cameras can capture plant images from invisible bands.[44] Overall, high-resolution cameras can generate a single image with summed pixels from several millions of pixels to tens of thousands of pixels.[45] The basic pipeline to process these images is to extract plant pixels from background, then to perform binarization and segmentation. To provide this pipeline, several kinds of software were developed

(Table 1.2). These software process images to provide measured traits such as plant height,[46] root architecture,[47] and ear length.[48] From original measurements, potential "traits" could also be derived through mathematical transformations (i.e. ratio between traits,[49] principal components[50]). The increase in the number of potential traits open possibilities to investigate the limited number of variations of genetic markers, and therefore understanding underlying functions of annotated genes. Also, different from 2D images, reconstructed 3D images can better reflect the volume of plant architecture and therefore are more like real plants. Complex traits extracted from 3D images demonstrated trait measurements, like leaf growth tracking,[51] surface areas,[52] whole plant skeletonization[53] and whole system architecture of the root.[54] In field environments, sensors installed on unmanned systems (i.e. unmanned aerial vehicle, field based phenotyping robot) can score traits in a block of dozens of plants simultaneously[55] and give an average value.[56] This technology can save labor during field season and provide more accurate numeric values for phenotypic diversity investigations. Many systems use RGB cameras which capture three sets of light intensities per pixel to approximate the the way humans visually perceive the world. However, other types of cameras or sensors are also used which can either more precisely capture differences in light which would appear identical to the human eye or an RGB camera and/or capture and measure light from wavelengths outside of the range perceived by humans and RGB cameras (generally (380nm-740nm)). Values extracted from hyperspectral images, a type of camera which measurement the intensity of many more wavelengths of light than the human eye or an RGB camera can be used to accurately predict the nutrient and water content of plants.[57] In general, the physiological changes of plants are not easily quantified. Based upon specific wavelength signatures, level of responses to environmental stress could be more accurately quantified[44,58] and associated with genetic variants. These nondestructive methods can monitor dynamic changes of a single plant over time in a more efficient way.

Table 1.2: Examples of software developed for processing plant images

Software	Implemented Language	Measured Organs/Task
PlantCV[59]	Python	Whole above-ground plant traits
DIRT[47]	Python	Root architecture
phytomorph[48]	MatLab	Maize ear, cob and kernel
DeepPlantPhenomics[60]	Python	Mutant classification, Leaf counting
3D modelling code[51]	R	Tracking leaf growth
IAP[50, 61]	Java, R	Plant morphological traits

1.3.2 Molecular phenotyping

Plant phenotypes are not limited to traits that can be measured visibly. The abundance of expressed transcripts and metabolite compounds often act as intermediates between genetic sequence variants and visible plant phenotypes. These invisible pathways can change and potentially produce visible phenotypic changes. Understanding molecular phenotypes prior to observations of conventional phenotypes could be used for monitoring early response of plants to external stimuli or plant organ determinations. Large-scale gene expression data could be collected from three aspects. The first is to measure as many tissues as possible for a given genotype, like maize B73.[62, 63] This type of data could serve as a standard for evaluating gene expression levels comprehensively for a given species. Second is to measure gene expressions of selected tissues in a large number of diverse varieties. A recently published result in maize generated RNA-seq data of 255 varieties in seven tissues.[64] Produced expression levels of each individual gene could be considered as a molecular trait and associated with genomic variants, where introduced variants in transcription to better explain phenotypic consequences. Thirdly, with the precise dissection of changes over time with a resolution of days or even hours profiling transcript abundance in the same tissues at different time points aid in understanding dynamic processes that occur during development or response to stress. Yi et al.[65] split maize seed development during first the six days into every four hours and sequenced RNA samples, which provides higher resolution for gene functional

classification based on time-course data. With applications of flow cytometry and laser-capture microdissection, it becomes possible to isolate single cells from a tissue and then perform single cell sequencing. Other than the traditional sequencing with mixed cells in the analyzed tissue, single cell sequencing can distinguish information of individual cells from a mixture. These more precise results generated from single cell sequencing technology could monitor the dynamic and developmental changes of gene expression, methylation modification and open chromatin.[66,67] Similar to sequencing data, metabolic compounds are measured in comparable ways in maize.[68] The rich set of invisible phenotypic data at molecular levels expands possibilities to explain underlying biological pathways in maize.

1.4 Genotype-phenotype association models

1.4.1 Early explorations in genotype-phenotype associations

Bridging gaps between genotypes and phenotypes is a consistent topic for plant biologists to reveal underlying complex genetic mechanisms affecting observed phenotypes. Francis Crick described the nbow well-known central dogma of biology in 1957.[69] Proteins are encoded by genes. These proteins can form complex structures (i.e. transcription factors, enzymes, hormones) and are involved in cellular processes in various tissues that ultimately determine phenotype. This suggests that observed phenotypes and genotypes are associated with each other. The initial approach to test associations between individual molecular markers and individual qualitative phenotype (i.e. root hair, seed color) is based on the chi-squared test of independence to access how genotypes can co-segregate with these binary phenotypes.[70] However, phenotypes in the real world can also include traits which are best represented by continuous values, which causes the problem of testing for association to be more complex. To dissect the genetic architecture of quantitative traits, a set of genetic markers could be used to construct a linkage map based on the genetic linkage between markers. As genes physically located

close to each other on chromosomes will tend to be inherited together from parents, linkage analysis uses this to detect co-segregation signals with investigated phenotypes.[71] Although linkage analysis can detect markers with large effects associated with the phenotype, the low resolution of this analysis is an obvious limitation.

1.4.2 Genome Wide Association Study

The diverse genetic background of maize mapping population contains enough variation to associate with phenotypic variants statistically using Genome Wide Association Study (GWAS). Based on the linkage disequilibrium (LD) decay in a certain population, candidate genes could be sought in the LD decay surrounding ranges of detected SNP positions. However, in many cases divisity panels can exhibit population structure where some individuals tend to share both common alleles and common phenotypic values as a result of either reproductive isolation, selection, or recent common ancestors (Figure 1.1). If there is no control for this, a large number of detected SNPs are likely to be co-segregated with common ancestors which are considered as false positives. To ameliorate this issue, the General Linear Model (GLM) (using principal components as additional covariates in fixed effects, also called Q model[72]) or Mixed Linear Model (MLM) (using kinship matrix as additional covariates in random effects, also known as K model, or plus principal components as additional covariates, which is called as Q+K model[73]) was developed to control false positives. However, if genetic markers are truly co-segregated with the studied phenotype in two sub-populations, over-correcting population structure will lead to false negatives. The ideal trait in a GWAS population should be segregated without the strong influence of population structure (Figure 1.1C). Nevertheless, in the past several years in the maize community, GWAS assists researchers on narrowing down candidate genes. Genes/QTLs may control traits including morphology (i.e. leaf architecture,[74] shoot development[75]), metabolites biosysnthesis (i.e. seed oil synthesis[33]), stress resistance (i.e. drought resistance,[76] head

smut resistance[77]) or flowering time.[40] Although GWAS can generate lots of genotype-phenotype associations, failed validations from some of these associations when they are tested in individual experiments not surprising. On one side, theoretical algorithms of many developed GWAS models still can yield false positives in situations with a relatively high detection power. In these cases, false discoveries can still happen in detected signals from GWAS.[78,79] On the other side, association does not mean causation (Ziegler and Van Steen, Brazil 2010). Many confounding variables, such as environmental factors and many regulators, could be involved in the underlying genetic architecture of the investigated trait, which could potentially play a role in determining the phenotype. One of disadvantages in GWAS that is debated a lot is the missing heritability issue. Signals in GWAS will almost never fully explain the phenotypic variation observed in a population, whether because of rare alleles, less representative populations, epigenetic effects, the limitations of additive genetic models or other factors.[80] To improve the detection power of association signals and deal with missing heritability issues, the multivariate GWAS model[81] and multi-locus GWAS model[82] were proposed, which show stronger powers than a simple univariate trait model. To validate detected association signals for some traits from other aspects, transcriptomic level association[83] or selective sweeps were applied to investigate co-detected signals.[84]

Generating precise genotype data and expanding the dimensionality of phenotypic data through measuring more traits in more environments at more timepoints enables the investigation of associations between genotypes and phenotypes in ways never before possible. This new methods in turn provide new opportunities to identify a subset of potentially functional genes among the total set of all annotated gene models. This thesis will demonstrate the importance of using new methods to confirm the identity of maize genotypes used in association studies using a case study of samples of the inbred B73 grown at different locations around the world, high-throughput phenotyping, and illustrate a new statistical method for using phenome-wide data to uncover functionally

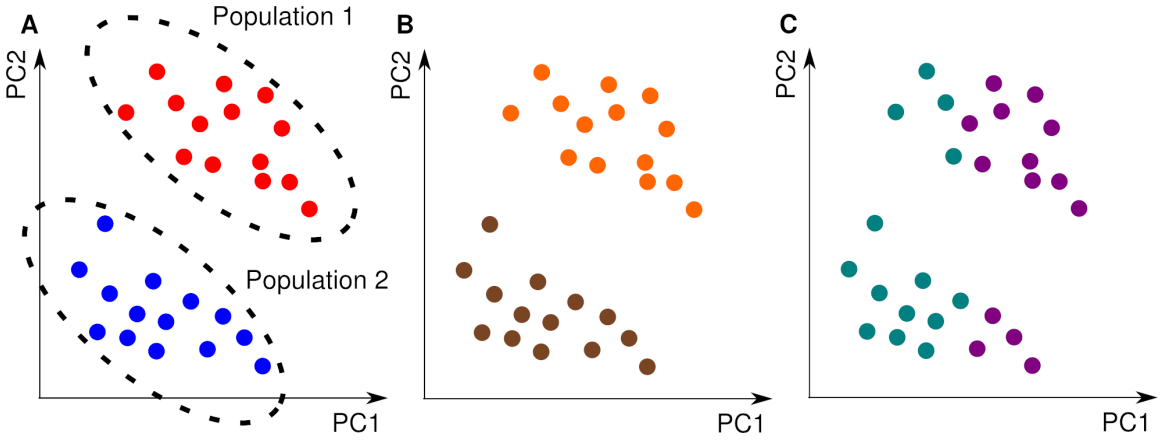

Figure 1.1: Population structure in GWAS study. (A) Two populations are separated by genetic markers via PCA analysis. Each dot represents an individual which owns unique values on first two PCs; (B) Individuals in the same two populations with observed binary traits are separated in the same way as owned population structure. Each color represents for one of binary traits; (C) Individuals in the same two populations with observed binary traits are separated without influence of population structure. Each color represents one value of a binary trait.

conserved genes in maize.

2:RNA-Seq Based Analysis of Population Structure within the Maize Inbred B73

2.1 Introduction

A great deal of biological research depends on reference genotypes that allow researchers around the world on work with material that is genetically identical or nearly identical. For many decades, assessing whether samples labeled as coming from genetically identical sources truly were identical was a costly, time consuming, and often inconclusive process.[85,86] However, recent advances in genotyping and sequencing technology have revealed a number of cases where sample names and sequence information significantly different stories. One study of human cell cultures found that 18% of cell lines were either contaminated or something entirely different from what they were labeled as[87] with the widely used HeLa cell line being one of the most frequent offenders.[88] Among plants, a recent resequencing study of *Arabidopsis* demonstrated that a line believed to carry a mutation for the ABP1 gene in an otherwise *Col-0* background actually contained a wide range of other nonsense and missense mutations as well as a large region on chromosome 3 which came from a different arabidopsis accession.[89] In soybean (*Glycine max*), segregating variation covering ~3.1% of the soybean genome assembly was observed between two sources of the reference genotype used in the construction of the soybean reference genome.[90] Resequencing of multiple plants from a single batch of Columbia-0 seed in *Arabidopsis* identified multiple haplotypes present in areas that summed up to ~20% of the total reference genome.[91] The problem of contaminated or mislabeled samples is a very real one in plant biology, and can invalidate the results of experiments in which substantial time and resources have been invested.[92]

Here we set out to quantify how severely these issues of divergence among samples labeled as belonging to the same genetic background impact maize (*Zea mays*), a preeminent model for plant genetics over the past century. Unlike soybean and *Arabidopsis*, maize is a naturally outcrossing species, so reference genotypes must be maintained by manually controlled self-pollination in each generation. Previous studies using small sets of individually scored markers have identified genetic variation between different sources of the same maize inbred.[85] This study focuses specifically on the maize reference genotype B73, which was developed in Iowa and first registered in 1972,[10] widely used in commercial hybrid seed production across the United States for much of the 1970s and 1980s[93] and is represented in the parentage of many elite lines even today.[94] B73 has also been widely used by plant biologists conducting basic genetic research in maize, and was employed in the sequencing and assembly of the first maize reference genome.[11]

2.2 Materials and Methods

2.2.1 Data sources

A search of NCBI's sequence read archive identified 25 Illumina RNA-seq data sets deposited by 19 independent research group in three countries (Table 1). Two additional RNA-seq data sets were constructed from B73 seed requested from Iowa State and the USDA's Germplasm Resources Information Network (Control 1 and Control 2 respectively). For these two samples RNA was extracted from 10-day old B73 seedlings grown at the University of Nebraska-Lincoln (Table 1). In four cases where the total amount of data per run was limited (USA 6, USA 8, USA 9 and USA 17), data from multiple sequencing runs labeled as coming from the same sample were grouped together for analysis. In one case, SRR514100, the total quantity of data was excessive, so only 1/10th of the total data set was employed.

2.2.2 Alignment and initial SNP calling

Low quality sequences were removed using Trimmomatic-0.33 with settings LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:36.[95] Trimmed reads were aligned to the repeat masked version of the maize reference genome (version B73 RefGen v3)[11] downloaded from Ensemble ($ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/$) using GSNAP in version 2014-12-29 (with parameters -N 1,-n 2,-Q).[96] Output files were converted from SAM to BAM format, sorted, and indexed using SAMtools.[97] SNPs were called in parallel along ten chromosomes of the maize version 3 using SAMtools mpileup (-I -F 0.01) and bcftools call (-mv -Vindels -Ob).

2.2.3 SNP list generation

The view function of Bcftools was combined with in-house Python scripts to extract the content of bcf files and classify SNPs based on the number of reference and non-references alleles on every screened SNP locus. In detail, if the total number of reads covering a particular SNP in a particular sample was below 5, then the site was treated as missing data. When 99% reads on the locus of a sample were from the non-reference allele the sample was coded as homozygous non-reference allele. The same criteria were used for calling a site as homozygous reference allele. When the reads containing reference and non-reference alleles totaled more than 90% of all reads and each allele was represented by more than 20% of aligned reads the site was coded as heterozygous. If two or more alleles were present at $>$1% of aligned reads but the above criteria were not satisfied, the site was also coded as missing data. To reduce the prevalence of false SNPs resulting from the alignment of reads from multiple paralogous loci to a single position in the reference genome, sites which were scored as heterozygous in more than 20% of all genotyped individuals were discarded. In total, 13,360 SNPs were used in downstream analysis. For each of these SNPs, the impact of the SNP on gene function was estimated

using SnpEff v4.1 and SnpEff databases ($AGPv3.26$).[98]

2.2.4 Population structure analysis

The distribution of the three possible genotypes (homozygous reference allele, homozygous non-referenece allele and heterozygous allele) over each of the ten chromosomes of maize was visualized using matplotlib. PhyML 3.0[99] was used to construct a phylogenetic tree with 100 bootstrap replicates, and 13,360 SNPs in total of 27 data sets. The maximum parsimony tree was constructed using Phylip-3.696[100] and the full set of 13,360 SNPs with missing data imputed by LinkImpute.[23]

2.2.5 Expression bias test

Individual FPKM (Frequency per kilobase of exon per million reads) value for each gene in each data set was calculated using Cufflinks v2.2.1.[101] Expression values were averaged across all China and USA South samples (excluded USA 12 sample that contained a unique introgressed region) separately. Only genes with average FPKM values $>= 10$ in both groups were retained for testing expression bias. The remaining genes were sorted into two groups: genes located in the 7 chromosome intervals where USA South and China showed different haplotypes and genes outside these intervals. The median gene expression value on behalf of each group was used to be compared.

2.2.6 Origins of haplotype blocks

The origins of haplotype blocks observed in some B73 accessions but not in the published reference genome were investigated using data from diverse maize lines in the HapMap2 project.[102] In order to make comparisons to these data, alignments and SNP calling were performed a second time as above using B73 RefGen v2. All of samples in China or USA North clade were combined to generate a consensus sets of SNP calls with reduced missing data. In examining region c2r2, sample USA 12 was used individually in

addition to the combined China and USA North sequences. In the analysis of region c5r2, USA 10, USA 14 and USA 15 were combined to generate a consensus set of SNP calls for the UC-Berkeley clade. The resulting SNP sets were employed for phylogenetic analysis as described above, with the alteration that the an approximate likelihood ratio test (aLRT) method with SH-like was employed. The resulting trees were visualized using FigTree v1.4.2 ($http://tree.bio.ed.ac.uk/software/figtree/$).

2.3 Results

2.3.1 Relationship among accessions labeled as B73

After alignment, SNP calling, and filtering (see Methods), a total of 13,360 high confidence segregating SNPs were identified among the 27 RNA-seq samples labeled as B73 employed in this study, substantially lower than the ~64,000 high quality SNPs identified by RNA-seq in a population segregating for a single non-B73 haplotype.[103] Phylogenetic analysis identified three distinct clades of samples separated by long branches with 100% bootstrap support (Fig 2.1). One clade consisted entirely of Chinese samples, one clade of samples from US research groups from Minnesota and Wisconsin, and the final clade encompassed the majority samples from US research groups as well as all German samples and the published reference genome for B73. We designated these clades "China", "USA North", and "USA South" respectively. Notably, the USA North clade is paraphyletic with respect to the China clade, suggesting B73 samples in China are likely derived from this group while both German samples are clearly part of the USA South Clade.

The USA South clade was somewhat arbitrarily divided into three subclades with at least 60% bootstrap support, as well as a number of singleton lineages (USA 1, USA 13, USA 19). Two of these clades contained control samples generated for this study, one from B73 seed requested through the USDA Germplasm Resource Network, and one from B73 seed requested from Iowa State. The subclade containing the known USDA B73

sample also contained the B73 reference genome sequence, consistent with the reported seed source for the B73 used in the construction of the reference genome. The final subclade did not contain any control samples. However, it was notable that four of the six samples placed in this clade originated in research groups whose PIs had conducted either PhD or Postdoctoral training with Michael Freeling at UC-Berkeley, and none of the samples outside of this clade originated in research groups linked to UC-Berkeley. Based on these, we designated the final USA South subclade "UC-Berkeley". This accessions has also been described as "Freeling B73".[104] In addition, the three major clades were also recovered in a parallel analysis using a tree generated using maximum parsimony, however the three subclades within USA South subclades were not fully recovered with identical membership. The consistency index (CI) and retention index (RI) for this tree was 0.825 and 0.861 respectively. Gene flow can product significant amounts of apparent homoplasy when constructing trees from multiple accessions of the same species. Therefore, these values were relatively higher than expected.

2.3.2 Genomic distribution of within-B73 polymorphisms

The polymorphic SNPs identified in this study could originate from one of several sources including *de novo* mutations or the introgression of non-B73 haplotypes in one or more lineages. SNPs originating from de novo mutations would be expected to show a distribution approximating that of gene density across the maize chromosomes. SNPs resulting from introgression of other haplotypes into B73 should be tightly clustered.

When the positions of the SNPs identified in this study were plotted it became clear that 55.3% SNPs identified in this study fall within a small number of dense genomic blocks on chromosomes 2, 4, 5, and 6 (Fig 2.2). The distribution of non-reference-genome-like haplotype blocks is consistent with the clade relationships identified above. The USA North clade can be defined by a large block of SNPs on chromosome 2, and smaller blocks on chromosomes 2 and 5, all of which are shared with

the China B73 clade. In addition to the blocks shared with the USA North B73 clade, samples from the China B73 clade all share a number of additional non-reference-genome-like blocks on chromosomes 2, 4, and 6. There are no non-reference-genome-like blocks shared by all members of the USA South clade, however a single non-reference-genome-like block on chromosome 5 is shared by the UC-Berkeley subclade of USA South. This block appears to share one breakpoint but not both with a block present in the USA North and China samples. Based on the location of this block, it is likely the same divergent haplotype region identified between the B73 reference genome and the B73 sample used to construct HapMap1.[105] The large block non-reference-genome-like block like SNPs observed only on chromosome 2 on USA 12 can likely be explained by the unique origin of this sample from wild type siblings of knotted1 mutants backcrossed into B73.[106] The remaining USA South samples, including the USDA GRIN, Iowa State, and German samples do not contain any obvious SNP blocks.

2.3.3 Functional impact of within-B73 polymorphism

Because the data used here came entirely from RNA-seq studies, our ability to detect SNPs was limited to genes which were consistently expressed at high enough levels to provide coverage of target regions. A total of 25,644 genes were expressed at levels >10 FPKM when at least one of data sets analyzed in this study. Of these genes, 633 (2.5%) fell within regions with non-reference-genome-like SNP blocks in one or more B73 clades. Using SnpEff, we identified 10 cases where SNPs produced "high impact" change such as the gain or loss of a stop code or the alteration of a splice donor or splice acceptor site and 396 cases which produced missense mutations which altered protein sequence. Only three genes with reported mutant phenotypes (whp1, mop1, and gol1) were in these regions, which only constituted at 2.7% of 112 classical identified maize genes with reported mutant phenotypes.[36] However, it must be noted that this is likely an

underestimate of the true number of changes, nonsense mediated decay may reduce or eliminate the expression of alleles of genes containing high impact SNPs, reducing the chances these SNPs will be detected from RNA-seq data.

2.3.4 Impact of within-B73 polymorpism on estimated gene expression

Overall, limited correlation was observed between gene expression level and detected SNP density. The correlation coefficient r between SNP density (number of snps per 1000 bases of exon sequence) and median gene expression across all analyzed datasets was 0.018 and 0.211 for genes outside and inside of block regions respectively. A previous study found that alignment rate for RNA-seq data from non-B73 genotypes to the B73 reference genome is approximately 13% lower than the alignment rate of RNA-seq data generated from B73 plants.[107] To test whether the introgression of non-reference genome like blocks created a bias towards lower estimated expression of genes in those blocks, for each gene within a block, the the median gene expression value observed across all datasets containing the block was compared to the median gene expression value across datasets where the same genomic region matched the reference genome. The comparison of global patterns across large populations of genes controls for experiment specific changes in the regulation of individual genes. Genes within introgressed regions showed a 5.6% reduction on expression relative to a control set of genes outside introgressed regions in this comparison between B73 USA South and B73 China (see Methods). This reduction was approximately half as large as would be predicted if the reduced alignment rate of data from non-B73 samples resulted solely from the increased difficulty of aligning reads containing SNPs to the reference genome. Potentially, the other half of the reduced alignment rate for non-B73 samples is the result of reads originating from transcripts of lineage specific genes, as previously suggested.[107]

2.3.5 Origins of polymorphic regions in B73 accessions

A total of 7 chromosome intervals (referred to here as c2r1, c2r2, c4r1, c5r1, c5r2, c6r1 and c6r2) containing non-reference genome haplotypes were identified in two or more samples (Table 2). SNP calls were extracted from individual non-reference-genome-like blocks using the previous version of the maize reference genome (B73 RefGen v2) and compared to genotype calls generated from 103 diverse inbreds resequenced by the Maize HapMap2 project.[102] One example, c2r1 is shown in Fig 2.3A. The non-reference genome haplotype present in this block for the Chinese samples clusters very closely with W22 (Fig 2.4), an older inbred developed in Wisconsin which has also been widely used in the maize genetics research community. Analysis of the other six large haplotype blocks produced longer branch lengths relative to the accessions represented in the Maize HapMap2 dataset (Table 2). However, in each case the haplotypes generated from each clade containing a non-reference-genome-like block clustered together, confirming that these regions did not result from parallel introgressions covering the same regions of the genome. Consensus SNP calls from the UC-Berkeley, USA North, and China B73 samples all clustered together with the HapMap2 B73 accession, but not with the B73 reference genome sequence which suggests that the source of B73 seed used for HapMap2 – like HapMap 1[105] – likely belonged to the UC-Berkeley subclade. Constraining the c2r2 region to only cover that portion of the genome which contained a block of SNPs in the USA North clade, the China clade and sample USA 12 revealed that USA North and China clustered together while USA 12 was placed at a different location on the tree. Interestingly, the only separation case of B73 RefGen and B73 HapMap2 in the origin tree of c5r2 indicated B73 seed in HapMap2 came from the UC-Berkeley sub-clade. In addition, for two cases, c2r1 and c5r2, we validated our haplotype assignments using an orthogonal analytical method, kmeans analysis. SNP data was first imputed using Linkimpute,[23] and then grouped into two clusters using kmeans function in R with k=2. For c2r1, the analysis was entirely consistent with the results presented above with

samples classified as China placed in one cluster with W22 and samples classified as USA North and South placed in the other. For c5r2, as expected all samples classified as China, USA North, and UC-Berkeley subclade were placed in a cluster with the B73 sample resequenced by the Hapmap2 project. In addition, one sample classified as USA South (USA 19) was placed in this cluster. Manual reexamination determined that USA 19 was heterozygous from the c5r2 SNP block (Fig 2).

2.4 Discussion

The maize community has long speculated that significant differences exist among B73 from different sources. Several previous studies have confirmed that genetic variation exists between different sources of the same maize inbreds,[85,105,108] yet due to constraints of cost and seed avaliability these comparisons were genrally able to compare only a small subset of potential seed sources for a given inbred. The avaliability of previously published RNA-seq data sets from a large number of independent research groups has made it possible to conduct a broad survey of the diversity among B73 accessions. No cases of samples which were labeled as originated from B73 but were clearly not B73 based on SNP data were identified in this study. Despite a 40+ generation reproductive history distributed across at least three continents, this analysis shows that 97.7% of the gene space of the maize genome is represented by a single consistent haplotype across all B73 accessions represented here. This compares favorably to approximately 20% of the genome showing multiple haplotypes in a single seed batch of the reference genotype of arabidopsis *Columbia-0*.[91] One potential explanation is that maize geneticists, always aware of the significant risk of pollen contamination, have had to be alert for signs of hybrid vigor or unexpected phenotypes when propogating inbred lines.[92]

In soybean, the published reference genome was found to consist of a mosaic of sequences observed in two separate sources of the reference variety and likely is not representative of the haplotype present in any individual plant.[90] In maize, a number of

samples classified into the USDA GRIN subclade (Fig 2.1) are largely consistent with the reference genome suggesting that the maize reference genome sequence likely is representative of a specific plant.

The interspersed SNPs distributed over ten chromosomes of maize may result from *de novo* mutations, segregation of heterozygous loci in the original B73 founder accession,[90] or false positive SNP calling errors. However, the majority of polymorphisms identified among B73 samples in this study primarily fell into a small number of dense non-reference-genome-like blocks, consistent with introgression of non-B73 germplasm into a B73 background. It is important to note that the B73 reference genome was sequenced relatively recently compared to the total age of the B73 accession. Therefore, it is not possible to infer whether a given non-reference-genome-like block originated from introgression into the line in which the non-reference-genome SNPs are observed or introgression into the B73 lineage which was ultimately employed in the creation of the B73 reference genome. However, in either case the relatively small size of these non-reference genome like blocks suggests multiple generations of backcrossing to the original B73 line, which would not be consistent with an origin as unrecognized pollen contamination.

Instead we propose a model based on the results from Sample USA 12. USA 12 consists of homozygous wild-type plants selected from family segregating for the Knotted1.[109] Therefore the block on chromosome 2 (~1% of the total maize genome) likely represents residual sequence from the knotted1 mutant donor parent line and is consistent with at least 5 generations of backcrossing (expected contribution of the donor parent = ~1.56%). Similar accidental fixations of unlinked regions may have occurred during the intentional introgression of other traits into a B73 background, such as disease resistance genes.[110]

The monophyletic placement of Chinese B73 datasets suggests that the B73 seed available in China likely originated from a single transfer from the USA, apparently of

seed belonging to the USA North clade and is an indicator of current tight controls on seed import/export which limit the ease with which seed change be exchanged between collaborators in China and the United States. Samples from Germany did not consistently form a monophyletic group. The concordance of academic lineages and genomic relationships in the UC Berkeley subclade acts as a notable positive control. More extensive sampling of B73 samples from many labs which employ this genotype in maize genetics research but have not, to date, published RNA-seq datasets may identify further B73 clades and subclades and additional cases where specific genomic variations have dispersed across the country as graduate students and postdocs leave a given lab for faculty positions of their own.

2.5 Conclusions

The existence of genomic variation among samples labeled as belonging to the same accession creates barriers to reproducbility, one of the core requirements of the scientific method.[92] In this study no examples of sample mislabeling were identified, however the possibility of ascertainment bias, with samples mislabeled as B73 being identified prior to publication must be acknowledged. A number of non-reference-genome-like blocks were identified in B73 samples originating from some sources. These blocks were shown to contain missense and nonsense mutations and measurably lower estimated expression values for genes in these regions. The identification of the relationships among different variants of B73 and the genomic locations of non-reference-genome-like regions will allow these differences to be controlled for future studies. With the rapid rise of sequencing-based assays such as RNA-seq, the strategy employed here may be a good one to apply in any case where one or more reference genotypes are widely employed in research across institutions, countries, and continents.

Table 2.1: B73 RNA-seq data sets sources.

Sample Name	Run Accession	Library Layout (bp)	Institute
Control 1	SRR3372478	Paired (101)	University of Nebraska - Lincoln
Control 2	SRR3371876	Single (51)	University of Nebraska - Lincoln
USA 1[111]	SRR651051	Paired (51)	University of Minnesota
USA 2[112]	SRR1819621	Paired (52)	University of Minnesota
USA 3[113]	SRR404150	Single (76)	University of Wisconsin - Madison
USA 4[114]	SRR514100	Paired (151)	University of Wisconsin - Madison
USA 5[63]	SRR940300	Single (101)	University of Wisconsin - Madison
USA 6[115]	SRR395191 SRR395192 SRR395194 SRR395208	Single (40)	Iowa State University
USA 7	SRR445245	Paired (102)	Iowa State University
USA 8[116]	SRR039505 SRR039506	Single (35)	Danold Danforth Center
USA 9[117]	SRR755252 SRR762349 SRR762350 SRR762351 SRR764626 SRR764627	Single (35)	Danold Danforth Center
USA 10[118]	SRR1656746	Single (101)	University of Nebraska - Lincoln
USA 11[119]	SRR1567899	Paired (50)	Iowa State University
USA 12[109]	SRR504480	Single (100)	University of California - Berkeley
USA 13[120]	SRR1587038	Single (101)	University of Wisconsin - Madison
USA 14[121]	SRR1231518	Single (100)	Cornell University
USA 15[122]	SRR1272115	Paired (50)	DuPont Pioneer
USA 16[123]	SRR640263	Single (35)	Yale University
USA 17[124]	SRR520998 SRR520999	Paired (51)	Cold Spring Harbor Laboratory
USA 18[125]	SRR536834	Single (76)	Virginia Tech
USA 19[126]	SRR999052	Paired (50)	Cold Spring Harbor Laboratory
USA 20[127]	SRR248565	Paired (81)	Stanford University
CHN 1[128]	SRR491307	Paired (76)	China Agricultural University
CHN 2[129]	SRR1522119	Paired (102)	China Agricultural University
CHN 3[130]	SRR910231	Paired (91)	China Academy of Agricultural Sciences
DEU 1[131]	SRR924107	Single (96)	MPIPZ
DEU 2[132]	SRR1030995	Single (85)	University of Bonn

*USA 12 harbors a long introgression on chromosome 2.

Table 2.2: Relationship of Non-Reference-Genome Like SNP Blocks to Haplotypes Surveyed by HapMap2.

Genomic blocks	Chr	Start (kb)	Stop (kb)	Closest haplotypes	Branch length	Present in
c2r1	2	40000	44300	W22	0.00000018	China
c2r2	2	212450	224250	BKN010	0.41156403	China
				BKN010	0.41156407	USA North
				M162W	0.32027864	USA 12
c4r1	4	169650	191550	CAU178	0.64099035	China
c5r1	5	201200	203000	no single best match	-	China
				no single best match	-	USA North
c5r2	5	209732	211540	B73 HapMap2	0.00000001	China
				B73 HapMap2	0.00000021	USA North
				B73 HapMap2	0.00000001	UC Berkeley
c6r1	6	120	8800	CML511	0.59542615	China
c6r2	6	20900	24670	OH7B	0.08905230	China

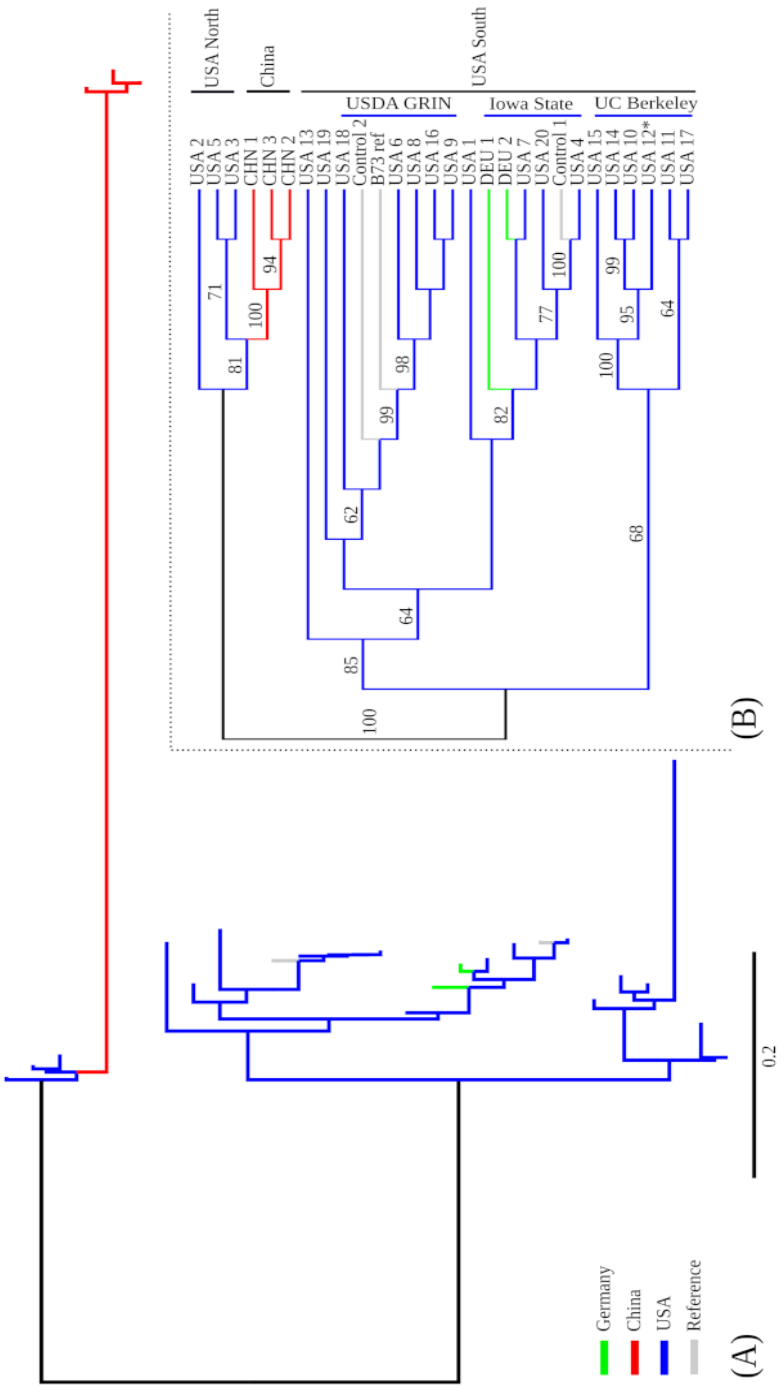

Figure 2.1: (A) Distance-scaled branch lengths; (B) Unscaled tree. Only bootstrap values greater than or equal to 60 are displayed.

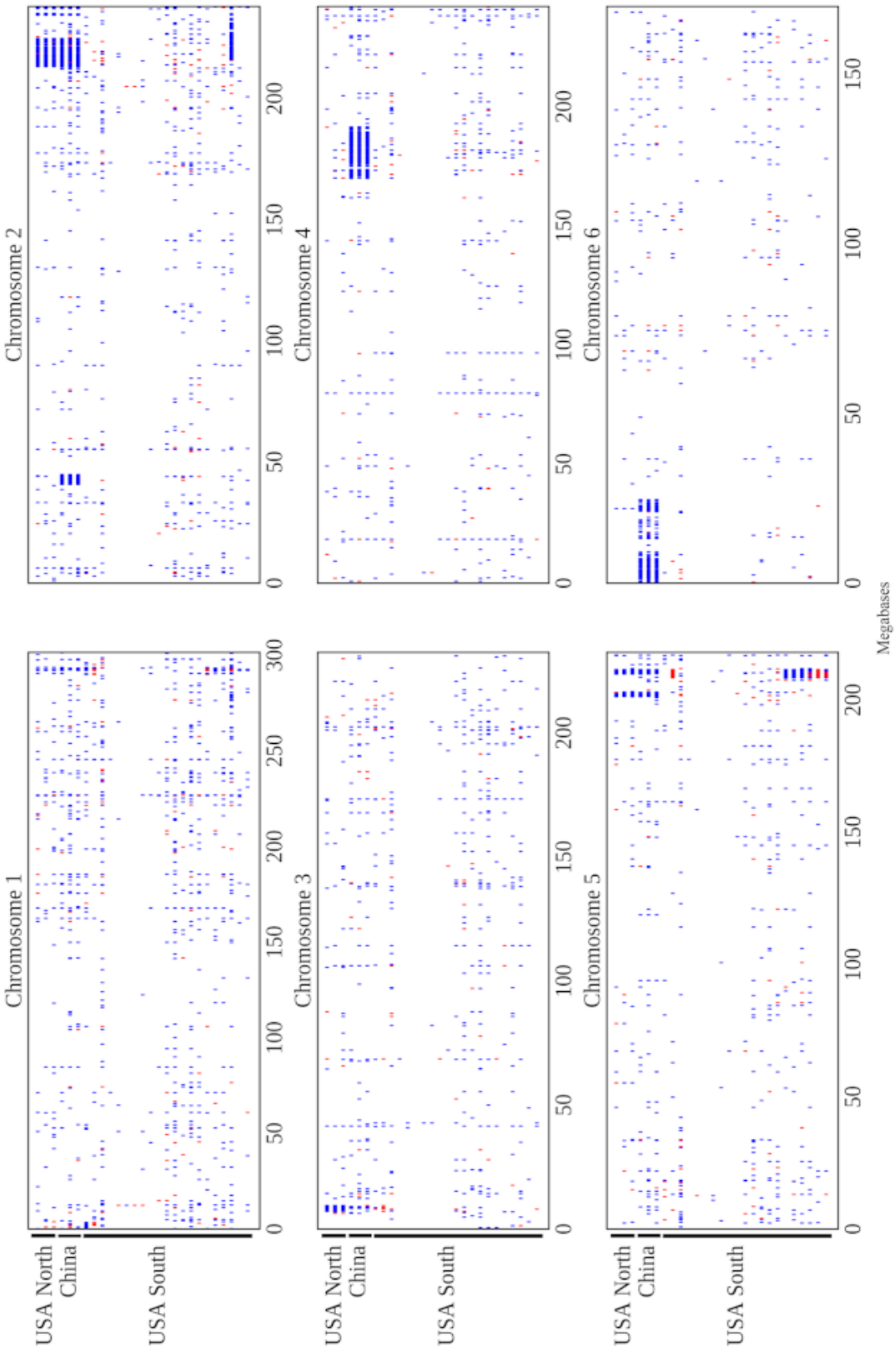

Figure 2.2: Non-reference-like homozygous genotypes are indicated in blue and heterozygous genotypes in red. The sample order from top to bottom on Y-axis in each sub-figure is the same order displayed as in Fig 1B.

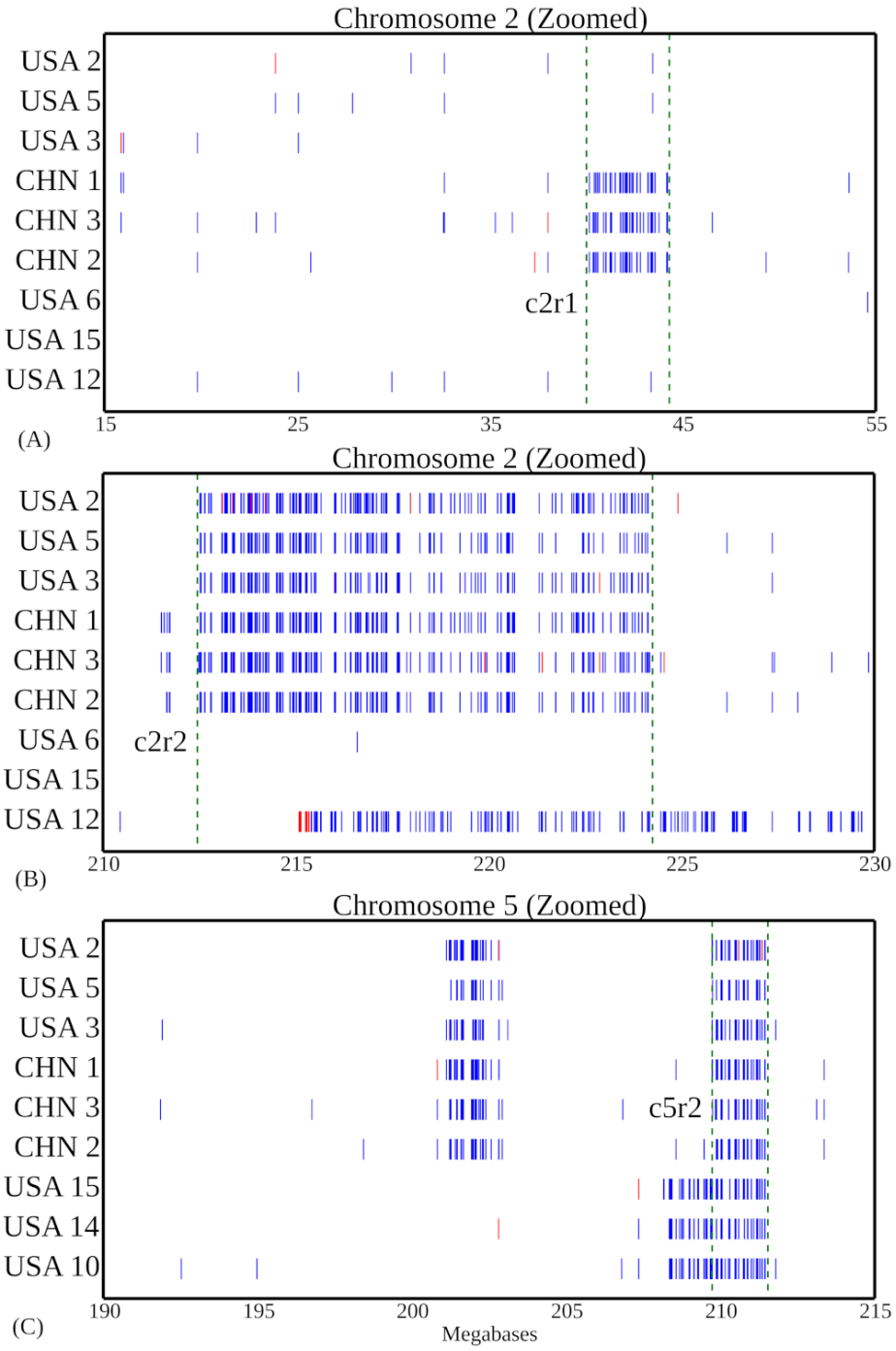

Figure 2.3: (A) Haplotype region c2r1 on Chromosome 2; (B) Haplotype region c2r2 on Chromosome 2; (C) Haplotype region c5r2 on Chromosome 5. Non-reference-like homozygous genotypes are indicated in blue and heterozygous genotypes in red. Named haplotype regions are those between the green bars.

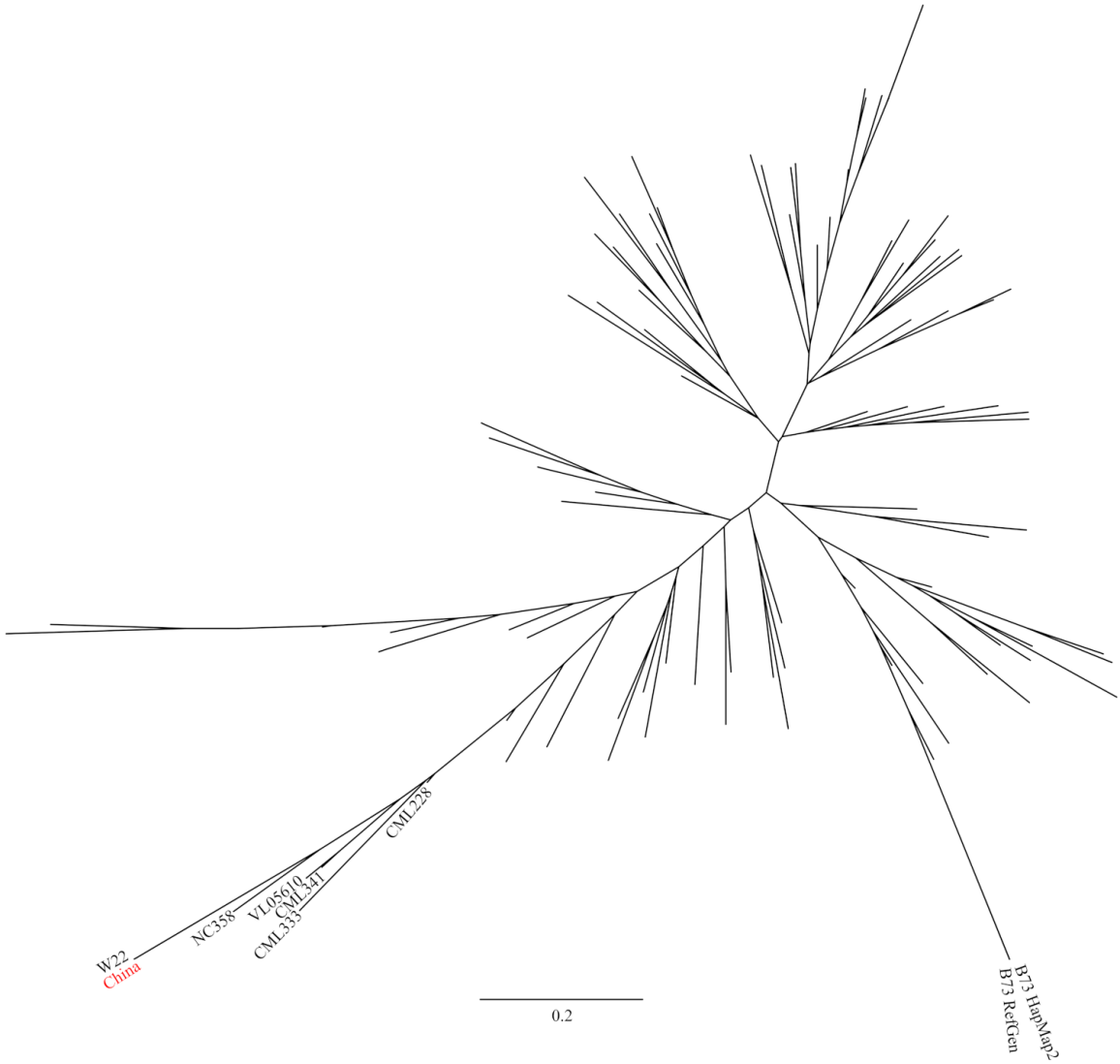

Figure 2.4: Relationship of the China B73 version of haplotype region c2r1 to the maize HapMap2 varieties.

3:Conventional and hyperspectral time-series imaging of maize lines widely used in field trials

3.1 Data Description

3.1.1 Background

The green revolution created a significant increase in the yields of several major crops in the 1960s and 1970s, dramatically reducing the prevalence of hunger and famine around the world, even as population growth continued. One of the major components of the green revolution was new varieties of major grain crops produced through conventional phenotypic selection with higher yield potentially. Since the green revolution, the need for food has continued to increase, and a great deal of effort in the public and private sectors is devoted to developing crop varieties with higher yield potential. However, as the low hanging fruit for increased yield vanish, each new increase in yield requires more time and resources. Recent studies have demonstrated that yield increases may have slowed or stopped for some major grain crops in large regions of the world.[133] New approaches to plant breeding must be developed if crop production continues to grow to meet the needs of an increasing population around the world.

The major bottleneck in modern plant breeding is phenotyping. Phenotyping can be used in two ways. Firstly, by phenotyping a large set of lines, a plant breeder can identify those lines with the highest yield potential and/or greatest stress tolerance in a given environment. Secondly, sufficiently detailed phenotyping measurements from enough different plants can be combined with genotypic data to identify regions of the genome of a particular plant species which carry beneficial or deleterious alleles. The breeder can

then develop new crop varieties which incorporate as many beneficial alleles and exclude as many deleterious alleles as possible. Phenotyping tends to be expensive and low throughput, yet as breeders seek to identify larger numbers of alleles each with individually smaller effects, the amount of phenotyping required to achieve a given increase in yield potential is growing. High throughput computer vision based approaches to plant phenotyping have the potential to ameliorate this bottleneck. These tools can be used to precisely quantify even subtle traits in plants and will tend to decrease in unit cost with scale, while conventional phenotyping, which remains a human labor intensive processes, does not.

Several recent pilot studies have applied a range of image-processing techniques to extract phenotypic measurements from crop plants. RGB (R: Red channel; G: Green channel; B: Blue channel) camera technology, widely used in the consumer sector, has also been the most widely used tool in these initial efforts at computer vision based plant phenotyping.[47, 54, 134, 135] Other types of cameras including fluoresence[44, 50] and NIR (near-infrared)[46, 50, 136] have also been employed in high throughput plant phenotyping efforts, primarily in studies of the response of plant to different abiotic stresses.

However, the utility of current studies is limited in two ways. Firstly, current analysis tools can extract only a small number of different phenotypic measurements from images of crop plants. Approximately 150 tools for analyzing plant image data are listed in a field specific database, however the majority of these are either developed specifically for *Arabidopsis thaliana* which is a model plant, or are designed specifically to analyze images of roots.[137] Secondly, a great deal of image data is generated in controlled environments, however, there are comparatively few attempts to link phenotypic measurements in the greenhouse to performance in the field. However, one recent report in maize suggested that more than 50% of the total variation in yield under field conditions could be predicted using traits measured under controlled environments.[135]

Advances in computational tools for extracting phenotypic measurements of plants

from image data and statistical models for predicting yield under different field conditions from such measurements requires suitable training datasets. Here, we generate and validate such a dataset consisting of high throughput phenotyping data from 32 distinct maize (*Zea mays*) accessions drawn primarily from recently off-patent lines developed by major plant breeding companies. These accessions were selected specifically because paired data from the same lines exists for a wide range of plant phenotypes collected in 54 distinct field trials at locations spanning 13 North American states or provinces over two years.[138] This extremely broad set of field sites captures much of the environmental variation among areas in which maize are cultivated with total rainfall during the growing season ranging from 133.604 mm to 960.628 mm (excluding sites with supplemental irrigation) and peak temperatures during the growing season ranging from $23.5°$C to $34.9°$C. In addition, the same lines have been genotyped for approximately 200,000 SNP markers using GBS.[138] Towards these existing data, we added RGB, thermal infra-red, fluorescent and hyperspectral images collected once per day per plant, as well as detailed water-use information (single day, single plant resolution). At the end of the experiment, 12 different types of ground-truth phenotypes were measured for individual plants including destructive measurements. A second experiment focused on interactions between genotype and environmental stress, collecting the same types of data described above from two maize genotypes under well watered and water stressed conditions.[139] We are releasing this curated dataset of high throughput plant phenotyping images from accessions where data on both genotypic variation and agronomic performance under field conditions is already available. All data was generated using a Lemnatec designed high throughput greenhouse-based phenotyping system constructed at the University of Nebraska-Lincoln. This system is distinguished from existing public sector phenotyping systems in North America by both the ability to grow plants to a height of 2.5 meters and the incorporation of a hyperspectral camera.[46] Given the unique properties described above, this

comprehensive data set should lower the barriers to the development of new computer vision approaches or statistical methodologies by independent researchers who do not have the funding or infrastructure to generate the wide range of different types of data needed.

3.1.2 Methods

3.1.2.1 Greenhouse Management

All imaged plants were grown in the greenhouse facility of the University of Nebraska-Lincoln's Greenhouse Innovation Center (Latitude: 40.83, Longitude: -96.69) between October 2nd, 2015 to November 10th, 2015. Kernels were sown in 1.5 gallon pots with Fafard germination mix supplemented with 1 cup (236 mL) of Osmocote plus 15-9-12 and one tablespoon (15 mL) of Micromax Micronutrients per 2.8 cubic feet (80 L) of soil. The target photoperiod was 14:10 with supplementary light provided by LED growth lamps from 07:00 to 21:00 each day. The target temperature of the growth facility was between $24 - 26^{\circ}$C. Pots were weighed once per day and watered back to a target weight of 5,400 grams from 10-09-2015 to 11-07-2015 and a target weight of 5,500 grams from 11-08-2015 to the termination of the experiment.

3.1.2.2 Experimental Design

A total of 156 plants, representing the 32 genotypes listed in Table 3.1 were grown and imaged, as well as 4 pots with soil but no plant which serve as controls for the amount of water lost from soil as a result of non-transpiration mechanisms (e.g. evaporation). The 156 plants plus control pots were arranged in a ten row by sixteen column grid, with 0.235 meter spacing between plants in the same row and 1.5 meters spacing between rows (Table **??**). Sequential pairs of two rows were consisted of a complete replicate with either 31 genotypes and one empty control pot, or 32 genotypes. Within each pair of rows, genotypes were blocked in groups of eight (one half row), with order randomized within

blocks between replicates in order to maximize statistical power to analyze within-greenhouse variation.

Plant Imaging

The plants were imaged daily using four different cameras in separate imaging chambers. The four types of cameras were thermal infrared, fluorescence, conventional RGB, and hyperspectral.[139] Images were collected in the order that the camera types are listed in the previous sentence. On each day, plants were imaged sequentially by row, starting with row 1 column 1 and concluding with row 10, column 16 (Table **??**).

Plants were imaged from the side at two angles offset 90 degrees from each other as well as a top down view. On the first day of imaging or when plants reached the two leaf stage of development, the pot was rotated so that the major axis of leaf phylotaxy was parallel to the camera in the PA0 orientation and perpendicular to the camera in the PA90 orientation. This orientation is consistent for all cameras and was not adjusted again for the remainder of the experiment. The fluorescence camera captured images with a resolution of 1038 $\times$ 1390 pixels and measures emission intensity at wavelengths between 500-750 nm based on excitation with light at 400-500 nm. Plants were imaged using the same three perspectives employed for the thermal infrared camera. The RGB camera captured images with a resolution of 2454 $\times$ 2056 pixels. Initially the zoom of the RGB camera in side views was set such that each pixel corresponds to 0.746 mm at the distance of the pot from the camera. Between 2015-11-05 and 2015-11-10, the zoom level of the RGB camera was reduced to keep the entire plant in the frame of the image. As a result of a system error, this same decreased zoom level was also applied to all RGB images taken on 2015-10-20. At this reduced zoom level, each pixel corresponds to 1.507 mm at the distance of the pot from the camera, an approximate 2x change. Plants were also imaged using the same three perspectives employed for the thermal infrared camera. The hyperspectral camera captured images with a resolution of 320 horizontal

pixels. As a result of the scanning technology employed, vertical resolution ranged from 494 to 499 pixels. Hyperspectral imaging was conducted using illumination from halogen bulbs (Manufacturer Sylvania, model # ES50 HM UK 240V 35W 25Âř GU10). A total of 243 separate intensity values were captured for each pixel spanning a range of light wavelengths between 546nm-1700nm. Data from each wavelength was stored as a separate grayscale image.

Ground Truth Measurement

Ground truth measurements were collected at the termination of data collection on November 11-12, 2015. Manually collected phenotypes included plant height, total number of visible leaves, number of total fully extended leaves, stem diameter at the base of the plant, stem diameter at the collar of the top fully extended leaf, length and width of top fully extended leaf, and presence/absence visible anthocyanin production in the stem. After these measurements, total above-ground fresh weight biomass was measured for four out of five replicates, resulting in the destruction of the plants. Ground truth data for the drought stressed subset of this dataset was collected following the procedure previously described in.[139]

RGB Image Processing

Pixels covering portions of the plant were segmented out of RGB images using a green index ((2×G)/(R+B)). Pixels with an index value greater than 1.15[139] were considered to be plant pixels. This method produced some false positive plant pixels within the reflective metal columns at the edge of the image. To reduce the impact of false positives, these areas were excluded from the analysis. Therefore, when plant leaves cross the reflective metal frame, some true plant pixels were excluded. If no plant pixels were identified in the image – often the case in the first several days when the plant had either not germinated or had not risen above the edge of the pot – the value was recorded as "NA" in

the output file.

Heritability Analysis

A linear regression model was used to analyze the genotype effect (excluding genotype ZL22 which lacked replication) and greenhouse position effect on plant traits. The responses were modeled independently for each day as

$$y_{h,ij,t} = \mu_{h,t} + \alpha_{h,i,t} + \gamma_{h,\nu(i,j),t} + \epsilon_{h,ij,t}, \tag{3.1}$$

where the subscript $h = 1, \ldots, 6$ denotes the three responses extracted from the images: plant height, width and size for the two views 0 and 90 degree. The subscripts i, j and t denote the jth plant in the ith row and day t, respectively, and $\nu(i, j)$ stands for the genotype at this pot. The parameters α and γ denote row effect and genotype effect, respectively. The error term is $\epsilon_{h,ij,t}$. Let $\text{SS}_{\alpha,t}$, $\text{SS}_{\gamma,t}$ and $\text{SS}_{\epsilon,t}$ be the sum of squares of the regression model (3.1) for the row effect, genotype effect and the error at time t, respectively. Let $\text{SS}_t = \text{SS}_{\alpha,t} + \text{SS}_{\gamma,t} + \text{SS}_{\epsilon,t}$ be the total sum of squares at time t. The heritability HR_t (3.2) of a given trait within this population was defined as the ratio of the genotype sum of squares over the sum of genotype and error sum of squares. For the estimate of the heritability of measurement error, the row effect term was replaced by a replicate effect (each replicate consisted of two sequential rows) with exclusion of ZL22 as only one plant of this genotype was grown.

$$\text{HR}_t = \frac{\text{SS}_{\gamma,t}}{\text{SS}_{\epsilon,t} + \text{SS}_{\gamma,t}}. \tag{3.2}$$

As the heritability index may change over the growth of the plant, an nonparametric smoothing method was provided for analyzing the time varying heritability of plants. The definition in (3.3) excludes the variation brought by the greenhouse row effect, which can be considered as the percentage of the variation in plant response that can be

explained by the genotype effect after adjusting the environmental effect. To compare with this definition of heritability (3.2), the response in the model without considering the row effect was constructed as

$$y_{h,ij,t} = \mu_{h,t} + \gamma_{h,\nu(i,j),t} + \epsilon_{h,ij,t}, \tag{3.3}$$

where similarly as (3.1), $\nu(i,j)$ is the genotype of the jth plant in the ith row. Let $\widetilde{\text{SS}}_{\gamma,t}$ and $\widetilde{\text{SS}}_t$ be the genotype sum of squares and total sum of squares under (3.4). The classical heritability is defined as

$$\widetilde{\text{HR}}_t = \frac{\widetilde{\text{SS}}_{\gamma,t}}{\widetilde{\text{SS}}_t}. \tag{3.4}$$

Hyperspectral Image Processing

Two methods and thresholds were used to extract plant regions of interest from hyperspectral images. First, the commonly used NDVI (normalized difference vegetation index) formula was applied to all pixels using the formula (R_{750nm}-R_{705nm})/(R_{750nm}+R_{705nm}), and pixels with a value greater than 0.25 were classified as originating from the plant.[140] Second, based on the difference in reflectance between stem and leaves at wavelengths of 1056nm and 1151nm, the stem was segmented from other part of plants by selecting pixels where (R_{1056nm}/R_{1151nm}) produced a value greater than 1.2. Leaf pixels were defined as pixels identified as plant pixels based on NDVI but not classified as stem pixels. In addition to the biological variation between individual plants, overall intensity variation existed both between different plants imaged on the same day and the same plant on different days as a result of changes in the performance of the lighting used in the hyperspectral imaging chamber. To calibrate each individual image and make the results comparable, a python script (hosted on Github; see code availability section) was used to normalize the intensity values of each

plant pixel using data from the non-plant pixels in the same image.

In order to visualize variation across 243 separate wavelength measurements across multiple plant images, we used a PCA (Principal Component Analysis) based approach. After the normalization described above, PCA analysis of intensity values for individual pixels was conducted. PCA values of each individual plant pixel per analyzed plant were translated to intensity values using the formula [x-min(x)]/[max(x)-min(x)]. False color RGB images were constructed with the values for the first principal component stored in the red channel, the second principal component in the green channel and the third principal component stored in the blue channel.

Fluorescence Image Processing

A consistent area of interest was defined for each zoom level to exclude the pot and non-uniform areas of the imaging chamber backdrop. Within that area, pixels with an intensity value greater than 70 in the red channel were considered to be plant pixels. The aggregate fluorescence intensity was defined as the sum of the red channel intensity values for all pixels classified as plant pixels within the region of interest, and the mean fluorescence intensity as the aggregate fluorescence intensity value divided by the number of plant pixels within the region of interest.

Plant Biomass Prediction

Two methods were used to predict plant biomass. The first was a single variable model based on the number of zoom level adjusted plant pixels identified in the two RGB side view images on a given day. The second was a multivariate model based upon the sum of plant pixels identified in the two RGB side views, sum of plant pixels identified in the two RGB side views plus the RGB top view, aggregate fluorescence intensity in the two side views, aggregate fluorescence intensity in the two side views plus the top view, number of plant stem pixels identified in the hyperspectral image and number of plant

leaf pixels identified in the hyperspectral image. Traits were selected to overlap with those employed by[141] where possible. This multivariate dataset was used to predict plant biomass using linear modeling as well as MARS, Random Forest and SVM.[141] MARS analysis was performed using the R package earth,[142] Random Forest with the R package randomForest[143] and SVM with the R package e1071.[144]

Data Validation and Quality Control

Validation against ground truth measurements

A total of approximately 500 GB of image data was initially generated by the system during the course of this experiment consisting of RGB images (51.1%), fluorescence images (4.3%), and hyperspectral images (44.6%). A subset of the RGB images within this dataset were previously analyzed in,[145] and were made available for download from http://plantvision.unl.edu/dataset under the terms of the Toronto Agreement. To validate the dataset and ensure plants had been properly tracked through both the automated imaging system and ground truth measurements, a simple script was written to segment images into plant and not-plant pixels (Figure 3.1). Source codes for all validation analysis are posted online (https://github.com/shanwai1234/Maize_Phenotype_Map).

Based on the segmentation of the image into plant and non-plant pixels, plant height was scored as the y axis dimension of the minimum bounding box. Plant area was scored as the total number of plant pixels observed in both side view images after correcting for the area of each pixel at each zoom employed (See Methods). Similar approaches to estimate plant biomass have been widely employed across a range of grain crop species including rice,[146] wheat,[147] barley,[147,148] maize,[139] sorghum[149] and seteria.[46] Calculated values were compared to manual measurements of plant height and plant fresh biomass which were quantified using destructive methods on the last day of the experiment. In both cases manual measurements and image derived estimates were

highly correlated, although the correlation between manual and estimated height was greater than the correlation between manually measured and estimated biomass (Figure 4.1A,B). Using the PlantCV software package,[59] equivalent correlations between estimated and ground truth biomass were obtained (r=0.91). Estimates of biomass using both software packages were more correlated with each other (r=0.96) than either was with ground truth measurements. This suggests that a significant fraction of the remaining error is the result of the expected imperfect correlation between plant size and plant mass, rather than inaccuracies in easimating plant size using individual software packages. Recent reports have suggested that estimates of biomass incorporating multiple traits extracted from image data can increase accuracy.[141] We tested the accuracy of biomass prediction of four multivariate estimation techniques on this dataset (see Methods). The correlation coefficient (r value) of the estimated biomass measures with ground truth data was 0.949, 0.958, 0.925 and 0.951 for multivariate linear model, MARS, Random Forest and SVM respectively.

The residual value – difference between the destructively measured biomass value and the predicted biomass value based on image data and the linear regression line equation – was calculated for each individual plant (Figure 4.1C). Using data from the multiple replicates of each individual accession, the proportion of error which is controlled by genetic factors rather than random error can be ascertained. We determined that 58% of the total error in biomass estimate was controlled by genetic variation between different maize lines. As such, this error is systematic rather than random and thus more likely to produce misleading downstream results when used in quantitative genetic analysis. As mentioned above, biomass and plant size are imperfectly correlated, as different plants can exhibit different densities, for example as a result of different leaf to stem ratios. Recent reports have suggested that estimates of biomass incorporating multiple traits extracted from image data can increase accuracy.[141] We tested the accuracy of biomass prediction of four multivariate estimation

techniques on this dataset (see Methods). The correlation of the estimated biomass measures with ground truth data was 0.949, 0.958, 0.925 and 0.951 for multivariate linear model, MARS, Random Forest and SVM respectively. However, even when employing the most accurate of these four methods (MARS), 63% of the error in biomass estimation could be explained by genetic factors. This source of error, with the biomass of some lines systematically underestimated and the biomass of other lines systematically overestimated presents a significant challenge to downstream quantitative genetic analysis. Given the prevalence of plant pixel counts as a proxy for biomass.[46, 139, 146–149]

Patterns of change over time

One of the desirable aspects of image based plant phenotyping is that, unlike destructively measured phenotypes, the same plant can be imaged repeatedly. Instead of providing a snapshot in time this allows researchers to quantify rates of change in phenotypic values over time, providing an additional set of derived trait values. Given the issues with biomass quantification presented above, measurements of plant height were selected to validate patterns of change in phenotypic values over time. As expected, height increases over time, and the patterns of increase tended to cluster together by genotype (Figure 3.3A). Increases in height followed by declines, as observed for ZL26, were determined to be caused by a change in the angle of the main stalk. While the accuracy of height estimates was assessed by comparison to physical ground truth measurements only on the last day, the height of three randomly selected plants (Plant 007-26, Plant 002-7 and Plant 041-29) were manually measured from image data and compared to software based height estimates, and no significant differences were observed between the manual and automated measurements (Figure 3.3B; Supplementary Table 1). To perform a similar test of the accuracy of biomass estimation at different stages in the maize life cycle, a set of existing ground truth measurements for two genotypes under two stress treatments[139] were combined with additional later

grow stage data (Supplemental Table 2). The correlation between total plant pixels observed in the two side views and plant biomass was actually substantially higher in this dataset (r=0.97) than the primary dataset, likely as a result of the smaller amount of genetic variability among these plants (Supplementary Figure 1).

Heritability of phenotypes

The proportion of total phenotypic variation for a trait controlled by genetic variation is referred to as the heritability of that trait and is a good indicator of how easy or difficult it will be to either identify the genes which control variation in a given trait, or to breed new crop varieties in which a given trait is significantly altered. Broad-sense heritability can be estimated without the need to first link specific genes to variation in specific traits.[150] Variation in a trait which is not controlled by genotype can result from environmental effects, interactions between genotype and environment, random variance, and measurement error. Controlling for estimated row effects on different phenotypic measurements significantly increased overall broad sense heritability (Figure 3.4A,B). This result suggests that even within controlled environments such as greenhouses, significant micro-environmental variation exists and that proper statistically based experimental design remains critical importance in even controlled environment phenotyping efforts.

If the absolute size of measurement error was constant in this experiment, as the measured values for a given trait became larger, the total proportion of variation explained by the error term should decrease and, as a result, heritability should increase as observed (Figure 3.4A). This trend was indeed observed across six different phenotypic measurements (three traits calculated from each of two viewing angles (Figure 3.4B). Plant height also exhibited significantly greater heritability than plant area or plant width and greater heritability when calculated solely from the 90 degree side angle photo than when calculated solely from to 0 degree angle photo.

In previous studies, fluorescence intensity has been treated as an indicator for plant abiotic stress status[44,151–153] or chlorophyll content level.[154,155] Using the fluorescence images collected as part of this experiment, the mean fluorescence intensity value for each plant image was calculated (see Methods). We found that this trait exhibited moderate heritability, with the proportion of variation controlled by genetic factors increasing over time and reaching approximately 60% by the last day of the experiment (Figure 3.4B).

Hyperspectral image validation

Hyperspectral imaging of crop plants has been employed previously in field settings using airborne cameras.[156–158] As a result of the architecture of grain crops such as maize, aerial imagery will largely capture leaf tissue during vegetative growth, and either tassels (maize) or seed heads (sorghum, millet, rice, oats, etc) during reproductive growth. The dataset described here includes hyperspectral imagery taken from the side of individual plants, enabling quantification of the reflectance properties of plant stems in addition to leaf tissue.

Many uses of hyperspectral data reduce the data from a whole plant or whole plot of genetically identical plants to a single aggregate measurement. While these approaches can increase the precision of intensity measurements for individual wavelengths, these approaches also sacrifice spatial resolution and can in some cases produce apparent changes in reflectivity between plants that result from variation in the ratios of the sizes of different organs with different reflective properties. To assess the extent of variation in the reflectance properties of individual plants, a principal component analysis of variation in intensity values for individual pixels was conducted. After non-plant pixels were removed from the hyperspectral data cube (Figure 3.5A) (See Methods), false color images were generated encoding the intensity values of the first three principal components of variation as the intensity of the red, green, and blue channels respectively

(Figure 3.5B, C and D). The second principal component (green channel) marked boundary pixels where intensity values likely represent a mixture of reflectance data from the plant and from the background. The first principal component (red channel) appeared to indicate distinctions between pixels within the stem of the plant and pixels within the leaves.

Based on this observation, an index was defined which accurately separated plant pixels into leaf and stem (see Methods). Stem pixels were segmented from the rest of the plant using an index value derived from the difference in intensity values observed in the 1056nm and 1151nm hyperspectral bands. This methodology was previously described.[139] The reflectance pattern of individual plant stems is quite dissimilar from the data observed from leaves and exhibits significantly different reflective properties in some areas of the near infrared (Figure 3.6). Characteristics of the stem are important breeding targets for both agronomic traits (lodging resistance, yield for biomass crops) and value added traits (biofuel conversion potential for bioenergy crops, yield for sugarcane and sweet sorghum). Hyperspectral imaging of the stem has the potential to provide nondestructive measurements of these traits. The calculated pattern of leaf reflectance for the data presented here are comparable with those observed in field-based hyperspectral studies,[159–161] providing both external validation and suggesting that the data presented here may be of use in developing new indices for use under field conditions.

In conclusion, while the results presented above highlight some of the simplest traits which can be extracted from plant image data, these represent a small fraction of the total set of phenotypes for which image analysis algorithms currently exist, and those in turn represent a small fraction of the total set of phenotypes which can potentially be scored from image data. Software packages already exist to measure a range of plant architectural traits such as leaf length, angle, and curvature from RGB images.[50,61] Tools are also being developed to extract phenotypic information on abiotic stress response

patterns from fluorescence imaging.[44,50] The analysis of plant traits from hyperspectral image data, while common place in the remote sensing realm where an entire field may represent a single data point, is just beginning for single plant imaging. Recent work as highlighted the potential of hyperspectral imaging to quantify changes in plant composition and nutrient content throughout development.[57,139] While these techniques have great potential to accelerate efforts to link genotype to phenotype through ameliorating the current bottleneck of plant phenotypic data collection, it will be important to balance the development of new image analysis tools with the awareness of the potential for systematic error resulting from genetic variation between different lines of the same crop species.

3.2 Availability of source code and requirements

- Project name: Maize Phenotype Map
- Project home page: `https://github.com/shanwai1234/Maize_Phenotype_Map`
- Operating system(s): Linux
- Programming language: Python 2.7
- Other requirements: OpenCV module 2.4.8, Numpy >1.5, CMake > 2.6, GCC > 4.4.x, Scipy 0.13
- License: BSD 3-Clause License

3.3 Availability of supporting data and materials

The image data sets from four types of cameras, pot weight records per day and ground truth measurements with corresponding documentation for 32 maize inbreds and same types of image data for two maize inbreds under two stress treatments were deposited in

the CyVerse data commons under a CC0 license with.[162] All image data were stored in the following data structure: Genotype $->$ Plant $->$ Camera type $->$ Day. For the hyperspectral camera each photo is stored as 243 sub images, each image representing intensity values for a given wavelength, so these require one additional level of nesting in the data structure Day $->$ wavelength. The grayscale images from the IR camera and the hyperspectral imaging system are stored as three-channel images with all three channels in a given pixel set to identical values. The fluorescence images contain almost all information in the red channel with the blue and green channel having intensities equal to or very close to zero, but data all three channels exist. Genotype data of 32 inbreds were generated as part of a separate project and SNP calls for individual inbred lines were made available either through[163] or the ZeaGBSv2.7 GBS SNP dataset stored in Panzea. Measurements for thirteen core phenotypes at each field trial as well as local weather data can be retrieved from publicly released Genomes 2 Fields datasets released on CyVerse.[163,164] Data from the 2014 G2F field trials is posted[163] and data from the 2015 G2F field trials is posted.[164] Genetically identical seeds from the majority of the accessions used in creating both this dataset and the Genomes 2 Fields field trial data can be ordered from public domain sources (e.g. USDA GRIN) and are listed in Table 3.1. Further supporting metadata and snapshots of the Maize Phenotype Map code are available in the GigaScience database, GigaDB.[165]

3.4 Declarations

3.4.1 List of abbreviations

DAP: Days after planting

GBS: Genotyping by Sequencing

LED: Light-emitting diode

MARS: Multivariate Adaptive Regression Splines

NDVI: Normalized difference vegetation index

NIR: Near-infrared

RGB: An image with separate intensity values for the red, blue and green channels

SNP: Single Nucleotide Polymorphism

SVM: Support Vector Machines

UNL: University of Nebraska-Lincoln

PA0: Plant Area calculated from a 0 degree image. Plants were initially orientated then leaves would be arranged parallel to the camera at 0 degrees.

PA90: Plant Area calculated from a 90 degree image. Plants were initially orientated then leaves would be arranged perpendicular to the camera at 90 degrees.

PCA: Principal Component Analysis

PH0: Plant Height calculated from a 0 degree image

PH90: Plant Height calculated from a 90 degree image

PW0: Plant Width calculated from a 0 degree image

PW90: Plant Width calculated from a 90 degree image

PF0: Average of plant fluorescence intensity in 0 degree

PF90: Average of plant fluorescence intensity in 90 degree.

Table 3.1: 32 genotypes in maize phenotype map

Genotype ID	Genotype	Source	Released Year
ZL1	740	Novartis Seeds	1998
ZL2	2369	Cargill	1989
ZL3	A619	Public Sector	1992
ZL4	A632	Public Sector	1992
ZL5	A634	Public Sector	1992
ZL6	B14	Public Sector	1968
ZL7	B37	Public Sector	1971
ZL8	B73	Public Sector	1972
ZL9	C103	Public Sector	1991
ZL10	CM105	Public Sector	1992
ZL11	LH123HT	Holden's Foundation	1984
ZL12	LH145	Holden's Foundation	1983
ZL13	LH162	Holden's Foundation	1990
ZL14	LH195	Holden's Foundation	1989
ZL15	LH198	Holden's Foundation	1991
ZL16	LH74	Holden's Foundation	1983
ZL17	LH82	Holden's Foundation	1985
ZL18	Mo17	Public Sector	1964
ZL19*	DKPB80	DEKALB Genetics	?
ZL20	PH207	Pioneer Hi-Bred	1983
ZL21	PHB47	Pioneer Hi-Bred	1983
ZL22**	PHG35	Pioneer Hi-Bred	1983
ZL23	PHG39	Pioneer Hi-Bred	1983
ZL24	PHG47	Pioneer Hi-Bred	1986
ZL25	PHG83	Pioneer Hi-Bred	1985
ZL26	PHJ40	Pioneer Hi-Bred	1986
ZL27	PHN82	Pioneer Hi-Bred	1989
ZL28	PHV63	Pioneer Hi-Bred	1988
ZL29	PHW52	Pioneer Hi-Bred	1988
ZL30	PHZ51	Pioneer Hi-Bred	1986
ZL31	W117HT	Public Sector	1982
ZL32	Wf9	Public Sector	1991

* Not currently available for order.

** Genotype represented by only a single plant in the dataset.

Table 3.2: Experimental layout (ID: ZL1-ZL32). At the time this experiment was conducted, the total size of the UNL greenhouse system was ten rows by twenty columns. Positions marked with UP indicate pots filled with plants from an unrelated experiment, while positions marked with NA indicate pots which had no plants. The first complete replicate is shown in color, and the four incomplete blocks within the first replicate are marked in different colors. * marks empty pots within the experimental design.

29	15	25	8	19	25	12	29	11	9
13	10	30	1	32	9	23	31	16	7
23	5	4	17	29	21	32	15	1	3
14	32	9	23	24	27	16	13	32	10
27	31	16	21	16	28	7	1	17	23
7	21	32	5	13	12	28	17	27	25
11	16	14	7	3	5	2	25	6	26
30	26	20	24	8	11	18	9	22	19
12	2	*	27	17	15	10	21	24	13
1	18	10	18	14	6	11	30	31	5
28	9	6	3	18	*	8	3	14	29
4	25	29	11	30	7	26	5	30	21
3	6	28	31	10	4	27	*	15	2
20	8	12	15	26	23	4	19	28	4
17	24	26	19	1	31	20	14	8	18
19	*	13	2	2	20	24	6	12	20
NA	NA	NA	NA	UP	UP	UP	UP	UP	UP
NA	NA	NA	NA	UP	UP	UP	UP	UP	UP
NA	NA	NA	NA	UP	UP	UP	UP	UP	UP
NA	NA	NA	NA	UP	UP	UP	UP	UP	UP

Figure 3.1: An example of plant segmentation. Segmentation of images into plant and not plant pixels for one representative plant (Path to this image in the released dataset: Genotype_ZL019 − > Plant_008-19 − > Image_Type − > Day_32). The area enclosed by green border is composed of pixels scored as "plant", the area outside the green border s composed of pixels scored as "not-plant". Minimum bounding rectangle of plant pixels is shown in red. (A) Side view, angle 1; (B) Side view, 90 degree rotation relative to A; (C) Top View.

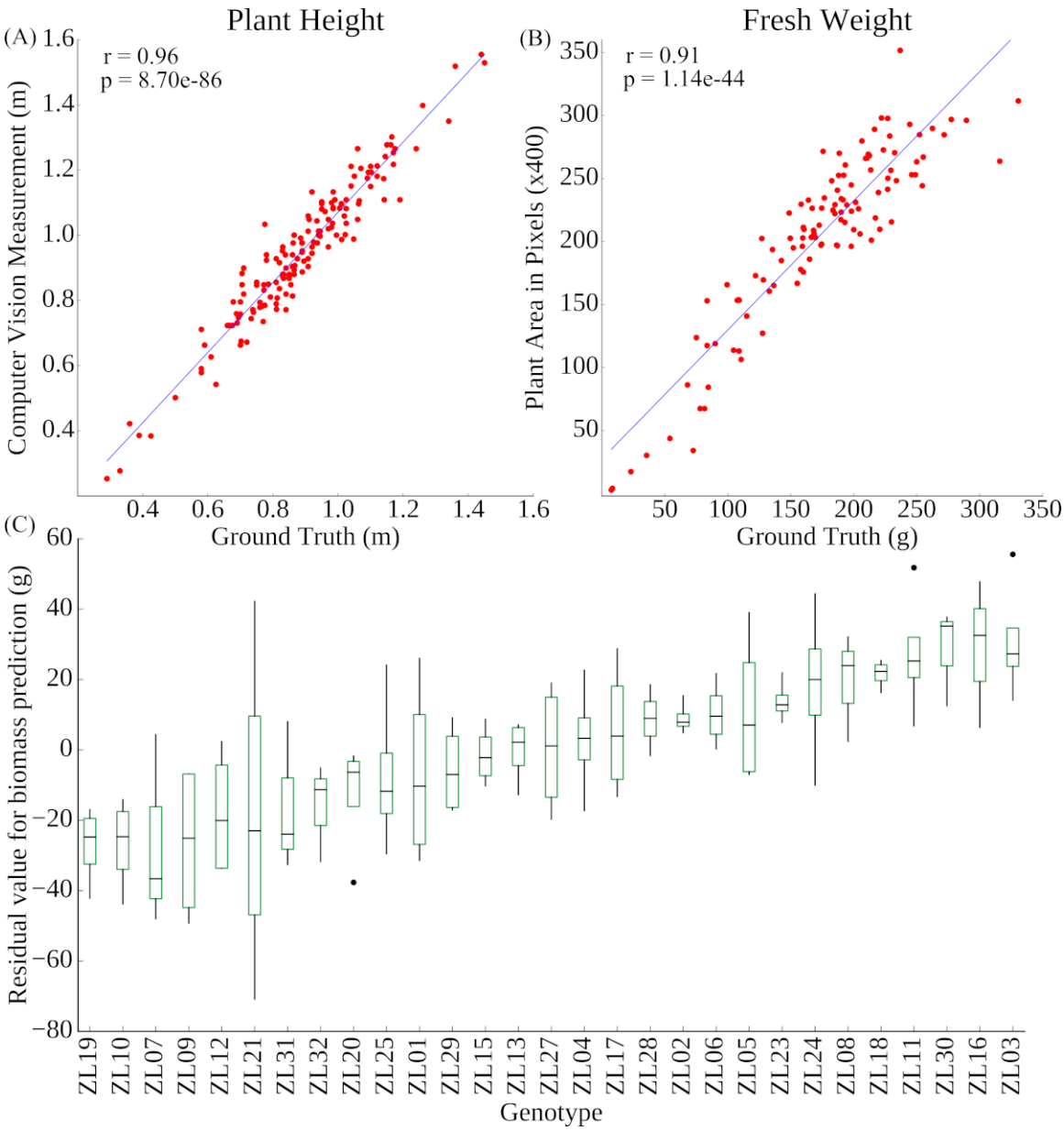

Figure 3.2: Correlation between image-based and manual measurements of individual plants. (A) Plant height; (B) Plant fresh biomass; (C) Variation in the residual between estimated biomass and ground truth measurement of biomass across inbreds.

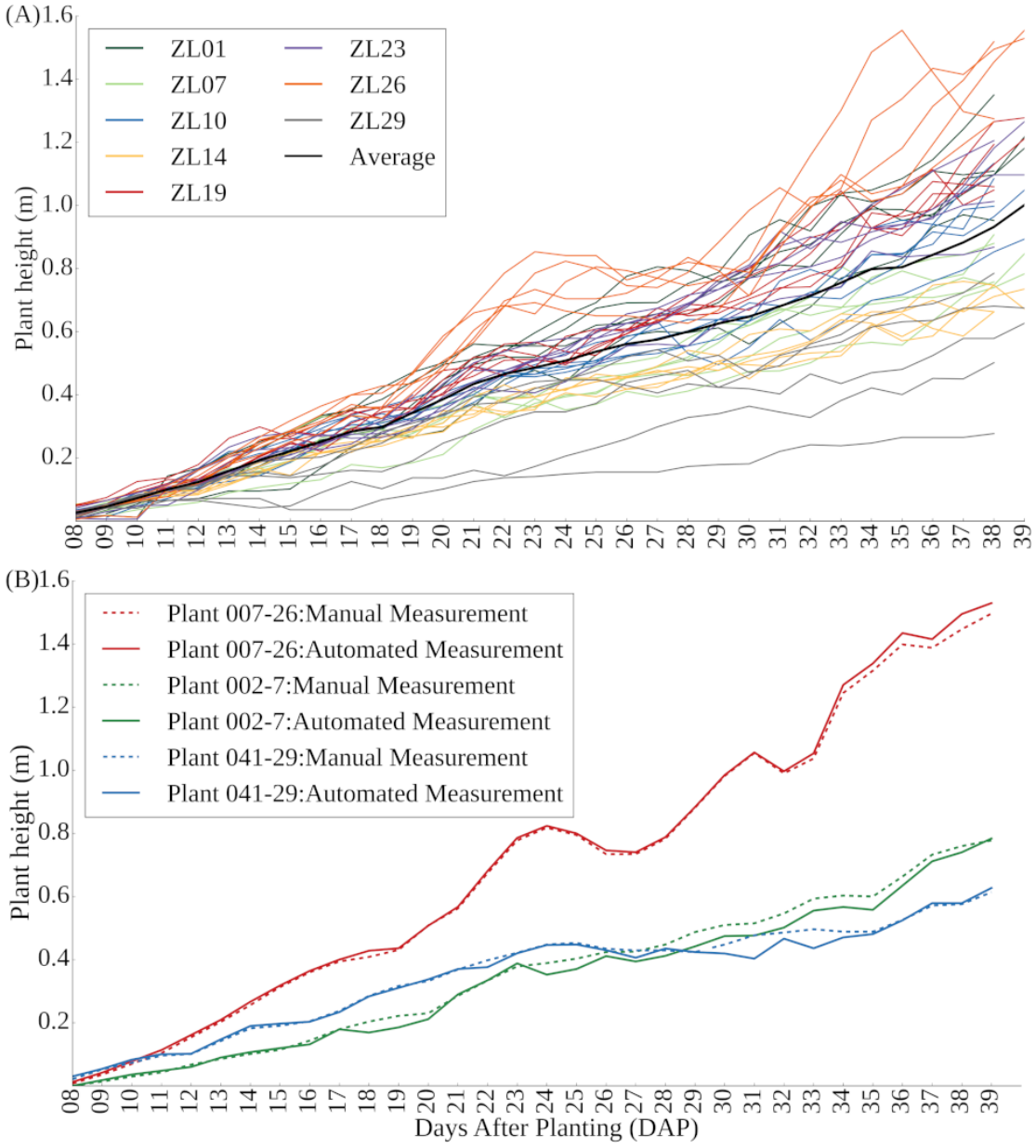

Figure 3.3: Time-series plant heights extracted from images. (A) Plant growth curves of each of five replicates of eight selected genotypes; (B) Comparison of manual measurements of plant height from image data with automated measurements for three randomly selected plants on each day of the experiment.

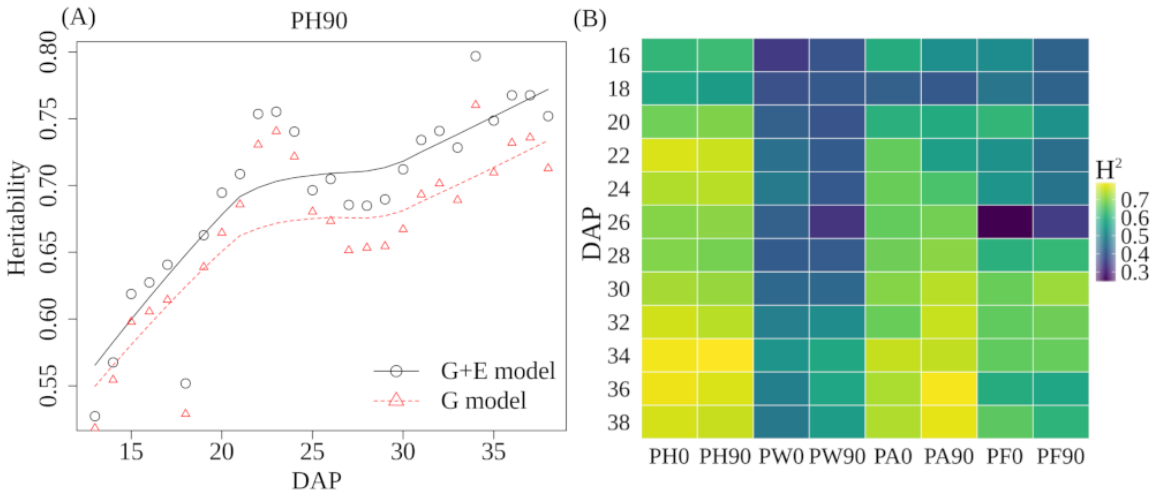

Figure 3.4: Time course heritability of extracted traits. (A) The time course broad sense heritability of PA90 before and after controlling for the row effect. The heritability in the G model was calculated using a linear model that only considers the effect of genotype with residual values in the error term while heritability in the G + E model was calculated using a linear model that considers the effect of both genotype and environment (row effect) with residual values in the error term; (B) Variation in broad-sense heritability (H^2) after controlling row effects for 6 trait measurements every second day across the phenotyping cycle. PA0: Plant Area in 0 degree (The major axis of leaf phylotaxy was parallel to the camera at 0 degree); PA90: Plant Area in 90 degree (The major axis of leaf phylotaxy was perpendicular to the camera at 90 degree); PH0: Plant Height in 0 degree; PH90: Plant Height in 90 degree; PW0: Plant Width in 0 degree; PW90: Plant Width in 90 degree; PF0: Average of plant fluorescence intensity in 0 degree; PF90: Average of plant fluorescence intensity in 90 degree.

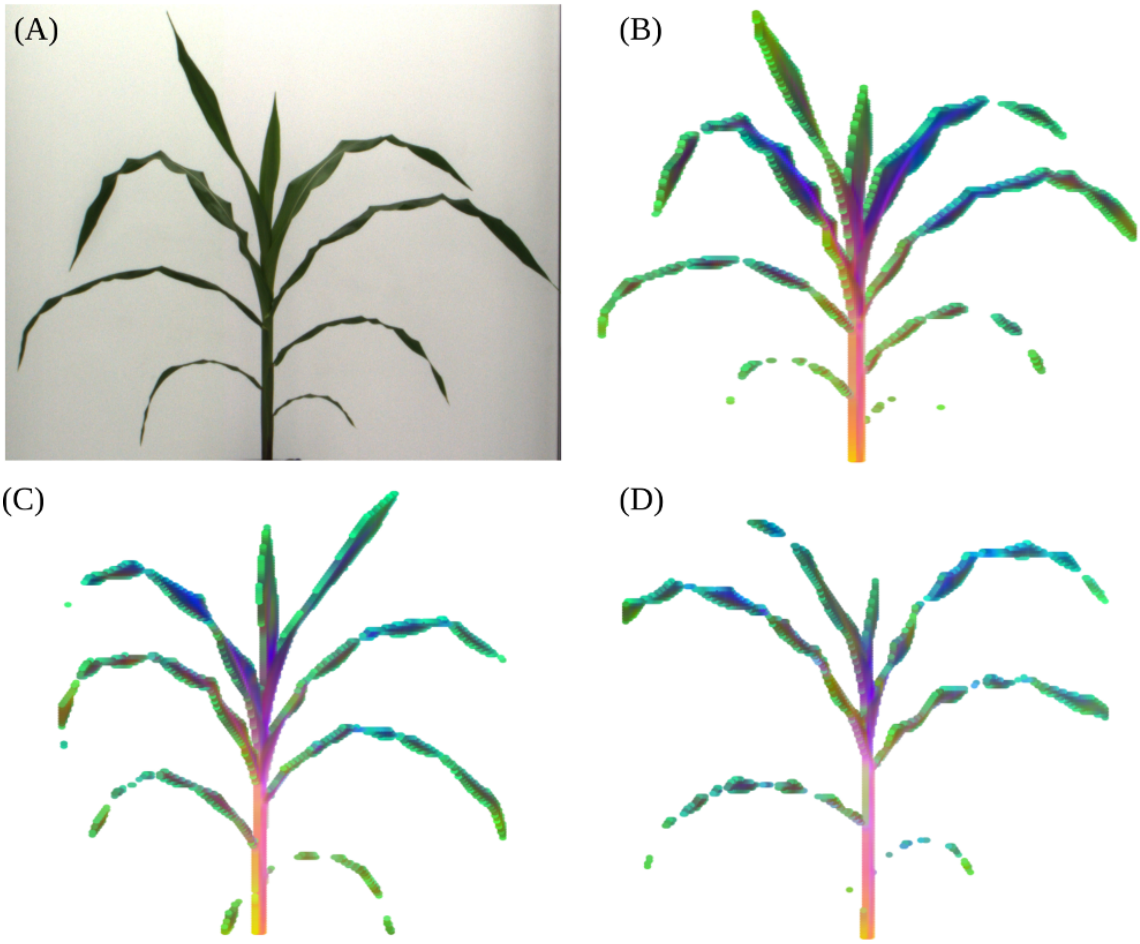

Figure 3.5: Segmentation and visualization of variation in hyperspectral signatures of representative maize plant images. (A) RGB photo of Plant 013-2 (ZL02) collected on DAP 37; (B) False color image constructed of the same corn plant from a hyperspectral photo taken on the same day. For each plant pixel the values for each of the first three principal components of variation across 243 specific wavelength intensity values are encoded as one of the three color channels in the false image; (C) Equivalent visualization for Plant 048-9 (ZL09); (D) Equivalent visualization for Plant 008-19 (ZL19).

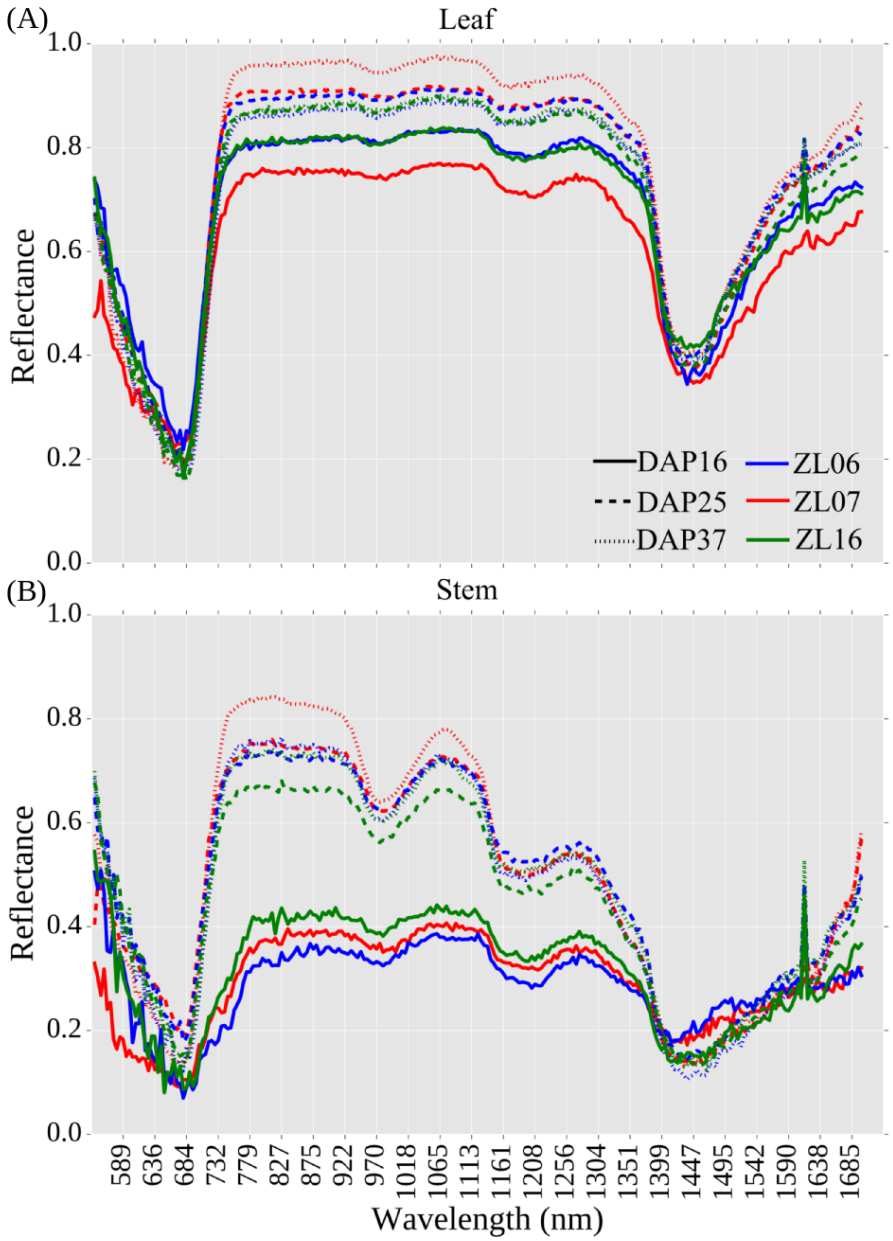

Figure 3.6: Reflectance values for three plants. Plant 090-6 (ZL06), Plant 002-7 (ZL07), and Plant 145-16 (ZL16) on three days across development. (A) Reflectance values for non-stem plant pixels (i.e. leaves) (B) Reflectance values for pixels within the plant stem.

4:Genome-phenome wide association in maize identifies a molecularly, structurally, and evolutionarily distinct set of genes

4.1 Introduction

Many approaches can be taken to achieve the goal of linking individual genes to their roles in determining the characteristics of an organism. One widely used approach is to employ natural functional variation between alleles in populations. Individual genetic markers are tested for association with differences in phenotype. Arguably, the first such association was the identification of a seed size QTL in dry bean (*Phaseolus vulgaris*) in 1923. This study used a single genetic marker, which was a qualitative trait controlled by a single gene.[166] Soon after, quantitative trait variation could be linked directly to chromosome structural markers.[167] Technology for scoring genetic markers continued to advance, making it possible to genotype markers covering the entire genome across a population. This enabled Genome Wide Association Study (GWAS) employing the linkage disequilibrium (LD) present in natural populations to identify functionally variable alleles of a gene influencing variation in a target trait.[168–170]

It is now feasible to collect data for thousands of intermediate molecular phenotypes, such as transcript, protein, or metabolite abundance, from entire association populations. These data can be incorporated into GWAS and Phenome Wide Association Study (PheWAS) analysis as either explanatory[83,171] or response variables.[64,172–174] Advances in high-throughput plant phenotyping have expanded the capacity of these techniques to score dozens or hundreds of whole-organism phenotypes across multiple time points and environments.[175,176] Multivariate GWAS methodologies

increase the power to detect true positives relative to single-trait methods,[177–184] however, current multivariate GWAS approaches face computational chanllenges related to scaling to hundreds of traits simultaneously. Statistical methods for PheWAS or "reverse GWAS"[185–187] seek to identify traits showing a statistical association with either a given marker, or all polymorphisms present in a given target gene.[188,189] Attempts have been made to unify GWAS and PheWAS in animals[190] and plants,[191] however, the rapid scaling of multiple testing makes it challenging to retain appropriate statistical power.

Here we employ a published dataset of 260 distinctly scored traits for 277 resequenced maize inbred lines[9,26] to develop and evaluate a novel approach to identify the links between genes and quantitative phenotypic variation using a multi-trait multi-SNP framework. We demonstrate that the genes identified using this method, which we call Genome-Phenome Wide Association Study (GPWAS), show substantially greater cross-validation in an independent study using data from approximately 20 times as many individuals[192] than do genes identified using conventional GWAS analysis of the same dataset. For a wide range of features, including expression level and breadth, syntenic conservation, purifying selection in related species, and the prevalence of presence-absence variation (PAV) across diverse maize lines, the genes identified using this multi-trait multi-SNP approach appear more similar to genes identified using forward mutagenesis and less similar to the overall population of annotated maize gene models.

4.2 Results

Genetic marker data were obtained from resequencing data of 277 inbred lines from the Buckler-Goodman maize association panel.[9] These lines are part of Maize HapMap3, which contains data for a total of 81,687,392 SNPs.[26] After removing the SNPs with high levels of missing data, those that were not polymorphic among the 277 individuals employed here, and several other quality filtering parameters, 12,411,408 SNPs remained.

Of these, 1,904,057 SNPs were assigned to 32,084 annotated gene models from the B73 RefGenV4 genome release. Filtering to eliminate redundancy between SNPs assigned to the same gene in high LD with each other reduced this number to 557,968 highly informative SNPs. A phenotypic dataset consisting of 57 specific traits scored for the Buckler-Goodman maize association panel across 1 to 16 distinct environments for a total of 285 unique phenotypic datasets was obtained from Panzea.[193] Removing datasets with extremely high levels of missing data resulted in 260 trait datasets with a median missing data rate of 18%. Of the total 72,020 potential trait datapoints (277 inbred lines $\times$ 260 traits), 23.6% or 16,963 trait datapoints were unobserved. Unobserved trait datapoints were imputed using a kinship-based method,[194] and the estimated imputation accuracies for the individual traits are reported in Supplementary Table 1.

A conventional GWAS analysis generally employs either empirically determined statistical significance cutoffs,[192] or a Bonferroni correction based on the total number of hypothesis tests conducted.[195] For the above dataset, employing a Bonferroni correction would mean each individual analysis would be conducted using a multiple-testing corrected p-value cutoff of 8.96e-08, while a sequential analysis of all 260 traits should employ a multiple-testing corrected p-value of 3.45e-10. As shown in Figure 4.1a, a given gene might be identified in multiple independent GWAS analyses for individual traits but not be considered significantly associated with any traits when correcting for the total number of traits analyzed. In the example given, Zm00001d002175 shows a statistically significant association with flowering time in multiple environments, yet none of these associations are individually significant enough to meet the threshold for the full multiple testing correction.

Bonferroni multiple testing correction assumes that each test is independent of all other tests, however, the different trait datasets collected from the Buckler-Goodman association panel exhibited significant correlation (Figure 4.1c), including three large blocks of traits related to flowering time, plant architectural traits, and tassel structure

traits respectively. To address the challenges of partially correlated traits and partially correlated genotype matrices, we developed an approach based upon a stepwise regression model fitting. In this model, the SNPs inside a gene body region are treated as response variables, and both population structure and individual trait datasets are employed to explain the patterns of SNP variance across the population. The significance of the association between each gene and the population of plant phenotypes is determined through a comparison of the final model with an initial model incorporating only the population structure variables (see Methods and Supplementary Figure 4.3).

Multiple testing was corrected using a permutation-based method (see Methods), which controls for the complexities introduced by iterative model selection. Although computationally expensive, permutation has been shown to be robust for controlling false positives in both GWAS and PheWAS studies.[74,196] Based on the permutation analysis, a p-value cutoff of 1.00e-23 resulted in the classification of 1,776 genes as being significantly associated with phenomic variation in the Buckler-Goodman association panel, resulting in an estimated false discovery rate (FDR) $<$ 1.00e-3. For comparison purposes, the same set of traits and genotypes was also tested for associations using three conventional GWAS algorithms: a general linear model (GLM GWAS),[72] a mixed linear model (MLM GWAS),[72] and FarmCPU GWAS[197] (See Methods). Applying an equivalent permutation based FDR threshold to each conventional GWAS algorithm removed the vast majority of positive signals (Supplementary Figure 4.4). Therefore, for GWAS models, a conventionally multiple testing corrected p-value cutoff was employed (Supplementary Table 2).

4.2.1 Validation of Gene-Phenome Associations

A second published dataset of genes identified as being associated with variation in trait values in the maize nested association mapping (NAM) population, which includes approximately 5,000 lines,[198] was employed to assess the relative power and accuracy of

three conventional GWAS algorithms as well as the GPWAS algorithm.[192,198] As the published data for the NAM population used B73 RefGenV2, all comparisons employed only the subset of 29,372 gene models with a clear 1:1 correspondence between gene models included in the B73 RefGenV2 and B73 RefGenV4 annotation versions. Of these, 4,227 of these genes were identified as being associated with at least one trait in the NAM dataset.[192] Genes identified using GPWAS showed significantly higher cross-validation in the NAM dataset than the sets of genes identified using GLM GWAS ($p = 2.05e\text{-}5$; Chi-squared test), MLM GWAS ($p = 0.010$; Chi-squared test), or FarmCPU GWAS ($p = 0.013$; Chi-squared test) (Figure 4.2a; Supplementary Figure 4.5; Supplementary Table 2). Filtering to remove signals from rare SNPs where the minor allele was present in only one or two of the NAM population founder lines reduced the total number of genes identified in that study to 3,621. However, the overall trend observed remained consistent and statistically significant, with the genes identified using the GPWAS algorithm continuing to show statistically significantly higher rates of identification in the reduced NAM dataset (GLM GWAS, $p=1.63e\text{-}4$; MLM GWAS, $p=0.002$; FarmCPU GWAS, $p=0.025$; Chi-squared test) (Supplementary Table 2). Analyses with two smaller real-world datasets for biochemical traits related to vitamin A (24 traits) and vitamin E (20 traits) metabolism[174,199] did not reveal any significant increase in the number of *a priori* gene candidates identified as showing a link to phenotypic variation relative to conventional GWAS approaches (Supplementary Figure 4.6). This was consistent with the results of the simulation analyses, for which GPWAS showed a significant increase in power/false discovery trade-offs for datasets with 100 simulated phenotypes, even including many with low heritability, but did not show substantial advantages relative to conventional GWAS for datasets with smaller numbers of traits (Supplementary Figure 4.7).

Our GPWAS algorithm also produces a list of the specific traits included in the model for a given gene. For example, in Figure 4.1b, the overall association between Zm00001d002175 and the trait dataset was statistically significant. The 11 individual

traits included in the Zm00001d002175 model included both flowering time measured in multiple locations, as well as additional traits with indirect links to flowering time (e.g. number of leaves, Summer 2008, Cayuga, NY), and others with no obvious links to flowering time. These included the total kernel volume in one year in one location and kernel proteins as estimated using near infrared imaging in another year in a different location.

4.2.2 GPWAS Accurately Predicts Pleiotropic Consequences of Gene Knockouts

It is important to keep in mind that the associations of individual phenotypes identified within the model are not rigorously controlled for false discovery. We therefore sought to qualitatively evaluate whether traits included in the model for an individual gene make sense in the context of existing detailed biological knowledge about the function of a given gene. One such gene was *anther ear1* (*an1*), a classical maize gene encoding an ent-copalyl diphosphate synthase involved in gibberellic acid biosynthesis, for which knockout alleles have been shown to reduce or abolish tassel branching, reduce plant height, delay growth, and delay flowering.[200] In a separate analysis of the 5,000 individual maize NAM lines, *an1* was identified as being associated with one trait, tassel spike length,[32] however, it was not found to be associated with any individual traits through a conventional GWAS analysis of the Buckler-Goodman 282 dataset. GPWAS identified a statistically significant link between *an1* and a model incorporating multiple phenotypes including flowering time, plant height, and tassel branch number, all consistent with the known mutant phenotypes (Supplementary Figure 4.8). At least one additional phenotype included in the GPWAS model – germination count (Summer 2006, Johnston, NC) – was not supported by direct reports of characterization of the *an1* knockout allele, but is consistent with the role of *an1* in gibberellic acid metabolism.[201, 202] Overall, the set of phenotypes identified using GPWAS for the *an1* gene appeared to be consistent with previously reports based on either the characterization of the knockout

allele or quantitative genetic analyses of natural populations.

The GPWAS model also identified *liguleless2* (*lg2*), another classical maize mutant with a well characterized knockout mutant phenotype.[203] The *lg2* encodes a bZIP transcription factor.[204] The loss of *lg2* function disrupts the establishment of the ligule and auricle of the maize leaf and results in plants with extremely erect leaves.[203, 205] Lines carrying *lg2* knockout alleles have been reported to exhibit substantially (10-50%) higher grain yield than otherwise isogenic hybrids,[206, 207] reduced tassel branch numbers,[207, 208] and moderately increased central spike length.[208] Quantitative genetic analyses have identified signals for leaf angle, tassel branch number, and kernel row number associated with the *lg2* locus,[32, 74, 209] although the effect on kernel row number was not significant in at least one study utilizing null alleles of *lg2*.[208] In this study, GPWAS identified a statistically significant link between *lg2* and a model incorporating multiple phenotypes including upper leaf angle, leaf length, central spike length, kernel weight (a yield component trait), and cob diameter. Cob diameter exhibits substantial correlation and overlapping genetic architecture with kernel row number[210] (Supplementary Figure 4.9). The GPWAS model for *lg2* also incorporated a number of flowering-time related traits, which do not have consistent support in either the characterization of *lg2* knockout mutants, or previous quantitative genetic analyses of flowering time in maize. Despite this, knockout alleles of *lg2* have been reported to alter the vegetative-to-reproductive phase transition in maize and produce increased numbers of leaves on the main stalk, which would be consistent with its altered flowering time.[208] As in the case of *an1*, the traits identified as being associated with *lg2* using GPWAS appear to be largely consistent with previous characterization of the functional roles of *lg2* in maize.

4.2.3 Greater Functional Specificity of Genes Identified Using GPWAS

Genes identified using GPWAS appear to be a significantly less random sample of total gene models than the set of genes identified using GLM GWAS. A set of 1,406 genes were

uniquely identified using GPWAS but not GLM GWAS. An equivalent set of 1,630 genes were identified using GLM GWAS but not GPWAS. In the larger unique-to-GLM GWAS gene set, a single Gene Ontology (GO) term showed a statistically significant bias towards being associated with phenotypic variation (GO:0046034: ATP metabolic process), and two GO terms with nearly identical gene assignments showed a statistically significant bias towards not being associated with phenotypic variation (GO:0000723: Telomere maintenance and GO:003220 Telomere organization). However, the moderately smaller set of genes uniquely identified using GPWAS was enriched or purified for the presence of many more GO terms. A total of 71 GO terms were overrepresented in the unique-to-GPWAS (relative to GLM GWAS) gene set to a statistically significant degree, including numerous terms linked to development, hormone signalling, response to different stimuli, and cell growth (Supplementary Table 4). The 13 GO terms that were underrepresented among genes uniquely identified using the GPWAS algorithm were generally associated with DNA conformation and replication (Supplementary Table 4). A similar comparison was made between genes uniquely identified using GPWAS and FarmCPU GWAS. In this case only 706 genes were uniquely identified using FarmCPU. As it is more likely for an enrichment or purification to be statistically significant in larger populations, only the 706 most significant unique-to-GPWAS (relative to FarmCPU GWAS) genes were evaluated in this comparison to eliminate any potential bias. Among the unique-to-FarmCPU GWAS gene set, only a single GO term was overrepresented to a statistically significant degree (GO:0051707: Response to other organism). However, among the the unique-to-GPWAS (relative to FarmCPU GWAS) gene set of equal size, 39 GO terms showed a statistically significant overrepresentation, while another 4 were statistically underrepresented (Supplementary Table 4).

Several potential factors could explain the large difference in GO enrichment purification we observed between genes identified solely using GWAS and genes identified solely using GPWAS. A number of factors, including the number of GO terms

per gene and the proportion of genes with no assigned GO term, differed modestly between the different populations of genes (Supplementary Table 5). The specificity of GO terms varied somewhat between the two populations. The median GO term assigned to a gene identified only using GLM GWAS was assigned to 514 total distinct gene models in B73 RefGenV4. For genes identified only using GPWAS, this decreased to 430 gene models. This difference in the number of genes that a given GO term is assigned does not appear to explain the differences observed in the enrichment or purification (Supplementary Figure 4.10). Rather, the large differences observed here are consistent with GWAS identifying a more random subset of annotated genes as being associated with phenotypic variation than did GPWAS.

4.2.4 Molecular, Structural, and Evolutionary Features of Genes Identified Using GPWAS

Genes identified using the GPWAS algorithm differed from the overall population of annotated maize gene models in a number of characteristics, as well as from the populations of genes identified using conventional GWAS. In many cases, the properties of genes identified using GPWAS appeared more similar to the population of genes with validated loss-of-function phenotypes.[36] Slightly less than half of all annotated maize genes were expressed to a level above 1 fragment per kilobase of transcript per million mapped reads (FPKM) in at least one of the 92 tissues/time points assayed.[63] This figure was greater than 2/3 for the genes identified using the three conventional GWAS algorithms, and approximately 3/4 for genes identified using the GPWAS algorithm and maize genes with validated loss-of-function phenotypes (Supplementary Table 2). Genes identified using GLM GWAS, MLM GWAS, FarmCPU GWAS, GPWAS, and the classical mutants all exhibited greater breadths of expression across tissues, larger numbers of genes with observed evidence of translation, and greater gene lengths than the population of annotated genes as a whole (Supplementary Table 2; Supplementary Table

6). The number of associated SNPs was positively correlated with the log-transformed inverse p-value assigned to genes using both GWAS (r = 0.566) and GPWAS (r = 0.625) (Supplementary Figure 4.11; Supplementary Table 6). However, this association declined dramatically in the permuted data for GPWAS (median permuted r = 0.155), but remained high for GWAS (median permuted r = 0.626) (Supplementary Table 7). This suggests that the high number of SNPs per gene for GPWAS (median: 43 SNPs, mean: 47.3 SNPs) relative to the overall gene set (median: 12 SNPs, mean: 17.4 SNPs) is a biological property of the genes controlling phenotypic variation in this population, rather than reflecting a bias in the GPWAS algorithm.

On a population and comparative genomics level, genes identified using the GPWAS algorithm also differed from the overall population of annotated maize gene models, and looked more like genes with validated loss-of-function phenotypes. Genes identified using both the conventional GWAS and GPWAS algorithms were significantly less likely to exhibit PAV in the maize populations (Figure 4.2b) than the overall population of maize gene models. The reduction in PAV frequency for genes identified using GPWAS (7.0%) was statistically significantly greater than for genes identified only using GWAS (10.4%) (p=0.0015; Chi-squared test), and not statistically significantly different from low level of presence absence variation observed for maize genes with validated loss of function phenotypes genes (4.1%) (p=0.36; Chi-squared test) (Supplementary Table 3). Genes identified using either conventional GWAS and GPWAS algorithms were significantly more likely to be conserved at syntenic orthologous locations in sorghum than the overall set of maize gene models (Figure 4.2c). Genes uniquely identified using GPWAS were more likely to be conserved at syntenic locations in the genome of sorghum *Sorghum bicolor* (91.8%) than those uniquely identified using GWAS (74-85%; see Supplementary Table 3). This difference was statistically significant in comparisons to all three GWAS algorithms tested and was comparable to the likelihood of syntenic conservation for maize genes with known loss of function mutant

phenotypes (93.9%) (Supplementary Table 3).

The genes identified as being associated with phenotypic variation using GPWAS also appeared to be under stronger purifying selection than either the overall population of maize gene models or those identified using any of the three conventional GWAS algorithms (Figure 4.2d; Supplementary Table 3). This analysis was constrained to the subset of gene models with conserved orthologs in sorghum (*Sorghum bicolor*), and foxtail millet (*Setaria italica*). Among these genes, those uniquely identified using GPWAS showed a reduced ratio of nonsynonymous substitution rate to synonymous substitution rate (Ka/Ks) (median: 0.168-0.169; mean 0.208-0.210), relative to the overall population of syntenically conserved maize gene models (median: 0.200; mean: 0.246), while those uniquely identified using GWAS showed elevated rates (median: 0.202-0.233; mean: 0.251-0.261) relative to the same overall population (Supplementary Table 3). Among the maize genes with characterized loss-of-function phenotypes, this ratio declined even further (median: 0.144; mean: 0.177). In short, the typical annotated gene appears to experience notably less purifying selection than those associated with organismal-level phenotypic variation based on either characterized loss-of-function mutant phenotypes or those identified using the GPWAS, but not a GWAS, algorithm.

4.3 Discussion

Complex datasets can contain scores for dozens or hundreds of traits across the same populations. The prevalence of these datasets and the challenges and opportunities they present is expected to grow in the coming years. Here, we developed an approach for identifying genotype-phenotype associations that can scale to the analysis of datasets containing hundreds, or potentially even thousands, of traits. The statistical tests upon which the GPWAS approach is built become unstable once the number of traits exceeds the number of individuals scored, therefore, scaling to high numbers of traits would require the use of larger association populations than many of the most widely used

plant populations today.[9,168,211,212] Multicollinearity in either the predictor or response variables can make the statistical estimation and inference procedures we employed unstable.[213] One common approach for reducing the total number of traits in a multi-year and/or multi-field site trial is to calculate the best linear unbiased predictors (BLUPs), which provide a single value for a given trait in a given line across multiple environments.[214] However, this approach strips out information on trait plasticity across environments, controlled by distinct sets of genes from those controlling multiple environment mean values[38] and is thus likely to bias the downstream analysis away from a large class of genes involved in determining organismal phenotypes across changing environments. In cases where the number of measured traits exceeds the number of environments, it would be advisable to employ alternative approaches to reduce the dimensionality of the trait dataset, whether that be an *ad hoc* approach such as selecting a subset of representative traits from highly correlated blocks, or dimensional reduction analyses such as a principal component analysis or multidimensional scaling. The automatic application of variable selection and/or dimensional reduction in such scenarios could be incorporated into future GPWAS implementations.

Another challenge for the present implementation of GPWAS is that it requires regions of interest to be defined across the genome. In this study, annotated gene models were used to define these regions, however, approximately 40% of of the phenotypic variation in maize has been estimated to be explained by noncoding regulatory regions.[215] These regions can be separated from the genes whose expression they control by many kilobases,[216,217] while LD in maize generally decays within one to several kilobases.[108,218] Both sequence conservation and chromatin mark data could be used to define additional regions of interest likely to represent regulatory sequences.[8,215,219–222] Similar approaches could also be employed to identify currently unannotated regions of the genome with a high potential of containing cryptic genes, including functional long noncoding RNAs.[223]

We found that genes with statistically significant links to phenotypic variation exhibit substantial differences from the overall population of annotated genes in the maize genome for a number of characteristics. They are more likely to be transcribed to significant levels, more likely to be conserved at syntenic orthologous positions in the genomes of related species, dramatically less likely to exhibit PAV across diverse maize inbred lines, and appear to be subjected to notably stronger purifying selection than the overall population of annotated genes. In all these cases, the genes identified using GPWAS are less like the overall population of annotated gene models and more like the small subset of genes in the maize genome whose functions have been characterized using loss of function alleles.[36] The distinct features shared by both genes identified using classical forward genetics and now using GPWAS suggest that it is unlikely that all annotated genes in the maize genome contribute to organismal phenotypes. 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.[224–226] However, many analyses still implicitly assume that genes annotated in the genome based on homology and/or expression evidence must play a role in determining organismal phenotypes. The absence of evidence for a role in determining a phenotype is interpreted as a failure to find either the correct trait to measure or the correct environment in which to measure it. Improved approaches to distinguish which annotated gene models are more likely to contribute to controlling organismal phenotypes will be critical to future efforts to guide gene-by-gene functional characterization efforts.

4.4 Methods

4.4.1 Genotype and Phenotype Sources, Filtering, and Imputation

Raw genotype calls from the resequencing of the maize 282 association panel[26] were retrieved from Panzea in AGPv4 coordinates. Missing genotypes were imputed using

Beagle (version: 2018-06-10).[227] Only biallelic SNPs with fewer than 20% missing data points were subjected to imputation. After imputation, SNPs with a minor allele frequency (MAF) of less than 0.05 or which were scored as 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, which was downloaded from Ensembl. SNPs were clustered based on $R^2 > 0.8$ and only one randomly selected SNP per cluster was retained. If, after collapsing the highly correlated clusters, the number of SNPs exceeded 138 (50% of the number of inbred lines scored), a random subsample of 138 SNPs was employed for the downstream analyses. Identical final SNP sets were employed for the GPWAS and GWAS analyses.

Of the 285 initial trait datasets, 25 were removed because the data file contained a recorded trait value for only one individual, leaving a total of 260 trait datasets. Using a Bayesian multiple-phenotype mixed model,[194] missing phenotypes were imputed based on a kinship matrix calculated from 1.24 million SNPs generated using GEMMA.[181] For those traits with a sufficient numbers of real observations to enable evaluation, the accuracy of the phenotypic imputation was assessed independently by masking 1% of available records for each trait and comparing the imputed and masked values. This process was repeated 10x for each trait.

4.4.2 GPWAS Analysis

All the operations for the GPWAS analyses are detailed in the R source code used to conduct the analysis – and associated documentation – which has been made available online (https://github.com/shanwai1234/GPWAS). Briefly, we employed a model

selection approach to adaptively select the most significant phenotypes associated with each gene. A F-test was used to compare a model to explain variation in SNPs based solely on population and a model which incorporated both population structure and trait data. The significance in the difference of the goodness of fit between these two models was used to determine the significance of the association of individual genes with phenotypic variation in the dataset.

The first stage is a stepwise selection procedure. The procedure iterates over all phenotypes in order to select individual phenotypes to incorporate into the model. This approach models all the SNP markers assigned to a given gene jointly with multiple responses. During each iteration, the association between each single trait and all of the evaluated SNPs are determined using a F-test which incorporates the dependence among the SNPs (see provided R code for details). If at least one trait passes a set threshold (in the analyses presented in this paper a threshold of $p < 0.01$ was employed), the single most significant trait is added to the model. If at least one trait was not significant based on the same threshold employed above, the single least significantly associated trait was removed from consideration. This process is repeated for a configurable number of iterations. For the analyses presented in this paper, the number of iterations was set to 35 as, given this number of iterations, none of the models for any gene included the maximum of 35 distinct traits.

After the number and identity of the phenotypes included in the model for a particular gene is finalized, the next stage is to evaluate how much the inclusion of phenotypic data improves model fit, relative to a purely population structure based model. To do this, two separate models are fit. The first model (initial model or IM) uses only population structure principal components to predict the values for all SNP markers associated with the target gene. The second model (GPWAS model or GM) uses both population structure and the phenotypes selected in stage one to predict the values for the same set of SNP markers. The goodness of fit of these two models is compared using

a F-test. The final result of the F-test takes into account all of the SNPs included from the target interval, as well as the degree of correlation between these SNPs. One of the criteria of those F-tests is that multiple response variables should not exhibit strong correlations with each other. This is the reason that the set of SNPs within each gene/interval were first filtered to select only one representative SNP from groups of SNPs in high linkage disequilibrium with each other.

In order to calculate the principal components used above, a separate PCA analysis was conducted for genes on each of the 10 chromosomes of maize. For analysis of the given gene on each chromosome, markers solely from the other 9 chromosomes were used to reduce the endogenous correlations between genes and principal components.[228] 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.

The final model can be represented as:

$$g_{k,i} = PC_{k,1}\beta_{i1} + PC_{k,2}\beta_{i2} + PC_{k,3}\beta_{i3} + \sum_{j=1}^{v_i} Phe_{k,(j)}\tau_{i(j)} + \epsilon_{k,ij}. \tag{4.1}$$

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 each scanned gene. The p-value of each gene was determined using 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.

FDR cutoffs for the partial F-test were based on the results from 20 permutation analyses, for which the values for each trait were independently shuffled among the 277 genotyped individuals and the entire GPWAS pipeline was rerun for all genes. Selected significant GPWAS genes with incorporated phenotypes are listed in Supplementary Table 8.

4.4.3 GWAS Analysis

GLM GWAS and MLM GWAS analyses were conducted using the algorithm first defined by Price and coworkers.[72] The FarmCPU GWAS with the algorithm was defined by Liu and colleagues.[197] All of 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). FarmCPU analysis method was run using maxLoop = 10 and the variance component method method.bin = "Fast-LMM".[229] The first three principal components were considered to be additional covariates for the population structure control in all of analyses. The same kinship matrix used in the phenotype imputation was also used for controlling the genotype relationship in the MLM GWAS model, while the method for analyzing variance components (vc.method) was set to GEMMA.[230] To enable a comparison with the GPWAS results, each gene was assigned the p-value of the single most significant SNP among all the SNPs assigned to that gene across the 260 analyzed phenotypes in the GWAS model.

4.4.4 Nested Association Mapping Comparison

Published associations identified for 41 phenotypes scored across ~5,000 maize recombinant inbred lines were retrieved from Panzea ($http:$ $//cbsusrv04.tc.cornell.edu/users/panzea/download.aspx?filegroupid = 14$).[192] Following the thresholding proposed in that paper, a SNP and CNV (copy number

variant) hits with a resample model inclusion probability $\geq$ 0.05, which were either within the longest annotated transcript for each gene (AGPv2.16) or within 15kb upstream or downstream of the annotated transcription start or stop sites were assigned to that gene respectively. Gene models were converted from the B73 RefGenV2 to B73 RefGenV4 using a conversion list published on MaizeGDB (https://www.maizegdb.org/search/gene/ download_gene_xrefs.php?relative=v4).

4.4.5 Gene Expression Analysis

Raw reads from the a published maize expression atlas generated for the inbred line B73 were downloaded from the NCBI Sequence Read Archive PRJNA171684.[63] Reads were trimmed using Trimmomatic-0.38 with default setting parameters.[95] Trimmed reads were aligned to the maize B73 RefGenV4 reference genome using GSNAP version 2018-03-25.[96] Alignment results were converted to a sorted BAM file format using Samtools 1.6,[97] and the FPKM values where calculated for each gene in the AGPv4.39 maize gene models in each sample using Cufflinks v2.2.[101] Only annotated genes located on 10 maize pseudomolecules were used for downstream analyses and the visualization of the FPKM distribution.

4.4.6 Ka/Ks Calculations

For each gene listed in a public syntenic gene list,[231] the coding sequence for the single longest transcript per locus was downloaded from Ensembl Plants. There sequences were each aligned to the single longest transcript of genes annotated as syntenic orthologs in *Sorghum bicolor* v3.1[232] and *Setaria italica* v2.2,[233] retrieved from Phytozome v12.0 using a codon-based alignment as described previously.[7] 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 automatically calculated using an in-house software pipeline posted to github

(https://github.com/shanwai1234/Grass-KaKs). Genes with a synonymous substitution rate less than 0.05 were excluded from the analyses, because their extremely small number of synonymous substitutions tended to produce quite extreme Ka/Ks ratios. Genes with multiple tandem duplicates were also excluded from the Ka/Ks calculations. The calculated Ka/Ks ratios of maize genes are provided in Supplementary Data 1.

4.4.7 Presence/Absence Variation (PAV) Analysis

PAV data were downloaded from a published data file.[234] Following the thresholding proposed in that paper, a gene was considered to exhibit presence absence variance if at least one inbred line had a coverage of less than 0.2.

4.4.8 Gene Ontology Enrichment Analysis

All GO analyses used the maize-GAMER GO annotations for B73 RefGenV4 gene models.[235] Statistical tests for GO term enrichment and purification were performed using the goatools software package (v0.8.12),[236] with support for a two-sided Fisher's exact test provided by the fisher_exact function in SciPy. To determine the median information content of the GO term, each was assigned a score based on the total number of gene models to which this GO term was assigned to in the maize-GAMER dataset. This analysis considered only gene models to which 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. Genes in B73 RefGenV4 Zm00001d.2 that employed in maize-GAMER GO annotations (∼40,000 genes) were used as the background population.

4.4.9 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. A sample of 2,000 randomly selected genes associated with 30,547 SNP markers were employed for the downstream simulations. In each simulation, 100 genes (5%) were randomly selected as causal genes. For each causal gene in each simulation, a causal SNP was selected to simulate the phenotypic effects. A total of 100 phenotypic traits were simulated with heritability equaling to 0.5 in each permutation of the analysis. 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 (v1.91.6).[237]

The resulting simulated trait data and genuine genotype calls were analyzed using GLM GWAS, FarmCPU GWAS, and GPWAS as described above, with the exception that the population structure PCs were calculated using a sample (1% or 191,856 SNPs) of the total SNPs remaining after filtering, rather than only using 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 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). **Total 100 simulated phenotypes were split into 1, 5, 10, 20, 50 and 100 subgroups for running GPWAS.** new added The power evaluation for GPWAS was defined as the number of true positive genes relative to the total number of causal genes, and FDR was defined as the number of false positive genes relative to the total number of positive genes. Power and FDR were calculated in a stepwise manner (step size: five genes) from five total positive genes to 500 (i.e. {5,10,...,495,500}).

4.5 Acknowledgements

This work is supported by National Science Foundation Awards MCB-1838307 and OIA-1826781 to JCS. In addition, we received support from the Quantitative Life Sciences Initiative at the University of Nebraska-Lincoln, which in turn received support from the

University of Nebraska Program of Excellence. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative. The authors thank Andy Dahl for the advice and instruction in the use of phenotype imputation software, Zheng Xu and Wenlong Ren for their 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.

All of supplementary tables are deposited in FigShare under the link of https://figshare.com/s/c2f6a2d2003227740a83.

Supplementary Table 1: The 260 phenotypes employed in this study with corresponding missing data rates, imputation accuracies and classified phenotype classes.

Supplementary Table 2: Expression characteristics, protein abundance and NAM gene validation among gene populations.

Supplementary Table 3: Conversation features for unique gene sets between each of GWAS models (GLM GWAS, MLM GWAS and FarmCPU GWAS) and GPWAS.

Supplementary Table 4: GO terms enriched and purified in gene populations uniquely identified in GPWAS.

Supplementary Table 5: Statistics of GO terms assigned to each gene population.

Supplementary Table 6: Gene length and SNP density in each gene population.

Supplementary Table 7: Correlation between significance levels and SNP numbers per gene for the genes generated from permuted and real data in GPWAS and GLM GWAS.

Supplementary Table 8: Significant genes detected using GPWAS and the phenotypes selected for each gene model.

Supplementary Data 1: Categories of annotated maize genes (AGPv4.39).

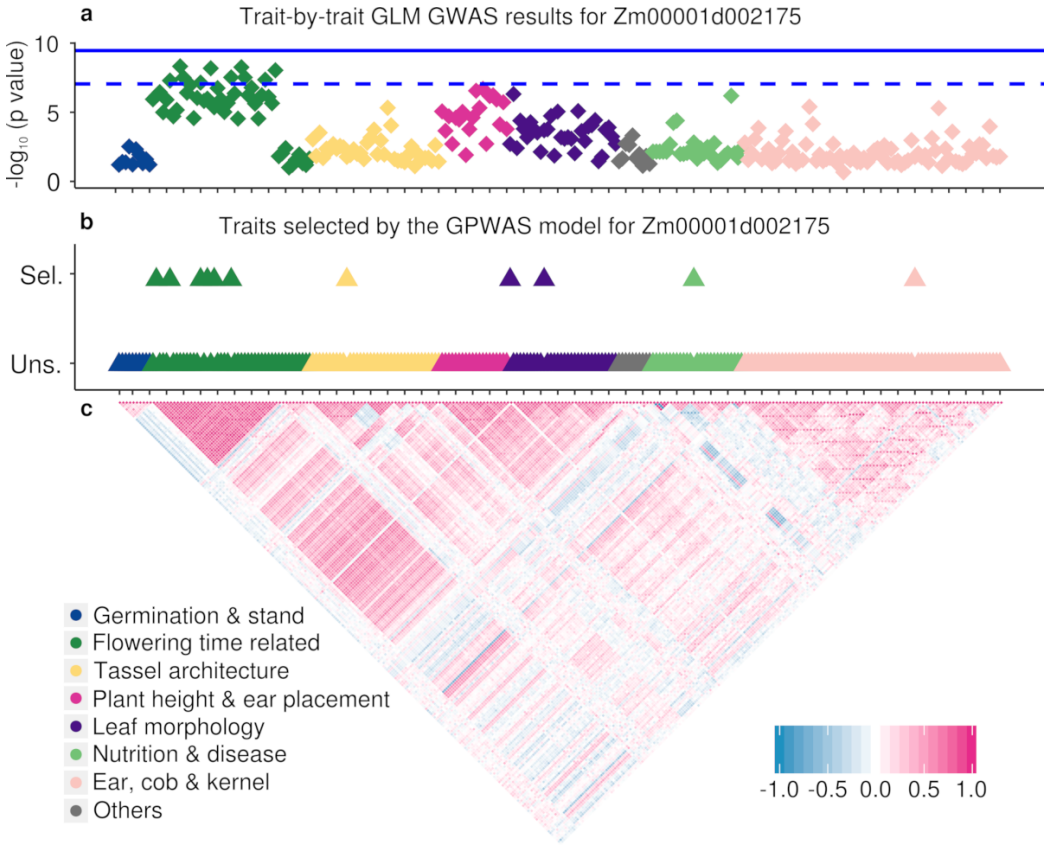

Figure 4.1: Each diamond or triangle represents one specific phenotypic dataset. Symbol colors indicate the broad categories into which each specific phenotype falls. The specific identities of each phenotype ordered from left to right are given in Supplementary Table 1. (a) The position of each diamond on the y-axis indicates the negative log10 p-value of the most statistically significant SNP assigned to that gene in a GLM GWAS analysis for that single trait. The dashed blue line indicates a p = 0.05 cutoff after Bonferroni correction for multiple testing based on the number of statistical tests in a single GWAS analysis (8.96e-8). The solid line indicates a p = 0.05 cutoff after Bonferroni correction for multiple testing based on the number of statistical tests in GWAS for all 260 traits (3.45e-10). (b) The placement of each triangle on the y-axis indicates whether a given phenotype was included in (Sel.) or excluded from (Uns.) the final GPWAS model constructed for this gene. The complete list of phenotypes incorporated into the GPWAS model for Zm00001d002175 is as follows: days to silk (Summer 2006, Cayuga, NY; Summer 2007, Johnston, NC), days to tassel (Summer 2007, Johnston, NC; Summer 2008, Cayuga, NY), GDD (Growing Degree Days) day 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 (Near InfraRed)-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 the top and middle panels indicates a distance of five phenotype datasets.

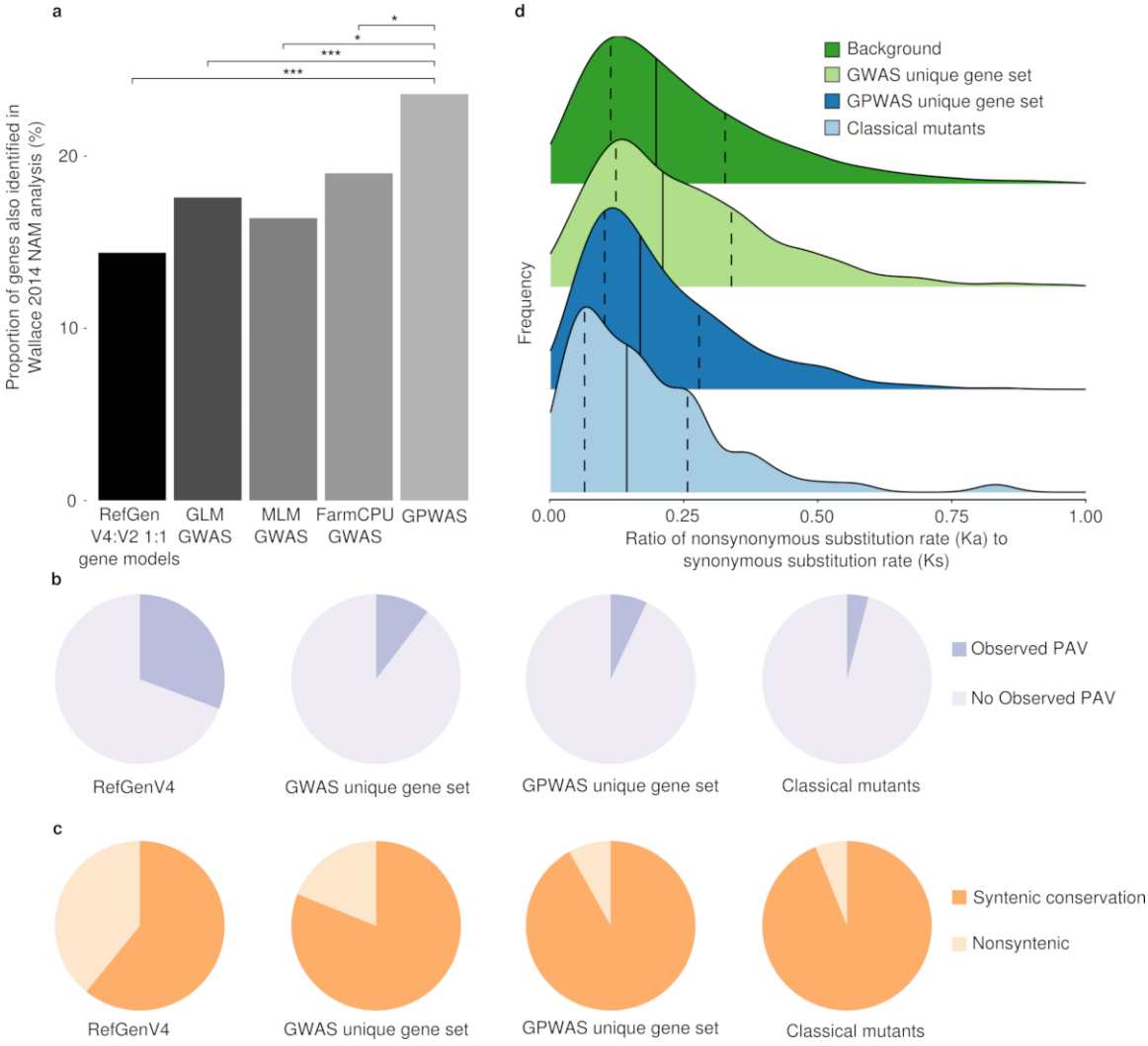

Figure 4.2: (a) Proportions of genes detected using various GWAS models (GLM, MLM, and FarmCPU), which overlap with genes detected by Jason Wallace et al.[192] *: p value $\leq$ 0.05; ***: p value $\leq$ 1e-3 (Chi-squared test). (b) Ratio of detected genes with PAVs to genes without PAVs. (c) Ratio between detected genes with syntenic conservation features relative to sorghum and genes without syntenic conservation features. The proportions of genes identified using MLM GWAS and FarmCPU GWAS with features of PAV, synteny conservation and Ka/Ks can be found in Supplementary Table 3. (d) Distribution of Ka/Ks values for different populations of genes within the maize genome. The background set comprises all maize genes with syntenic orthologs in sorghum (*Sorghum bicolor*) and foxtail millet (*Setaria italica*) after the exclusion of genes with tandem duplicates and genes with extremely few synonymous substitutions identified in the original alignment. The kernel density plots for genes uniquely identified using either GWAS or GPWAS, as well as by the use of 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 the 25th and 75th percentiles of the distribution. GLM GWAS was used to represent the GWAS model in panels b, c, and d.

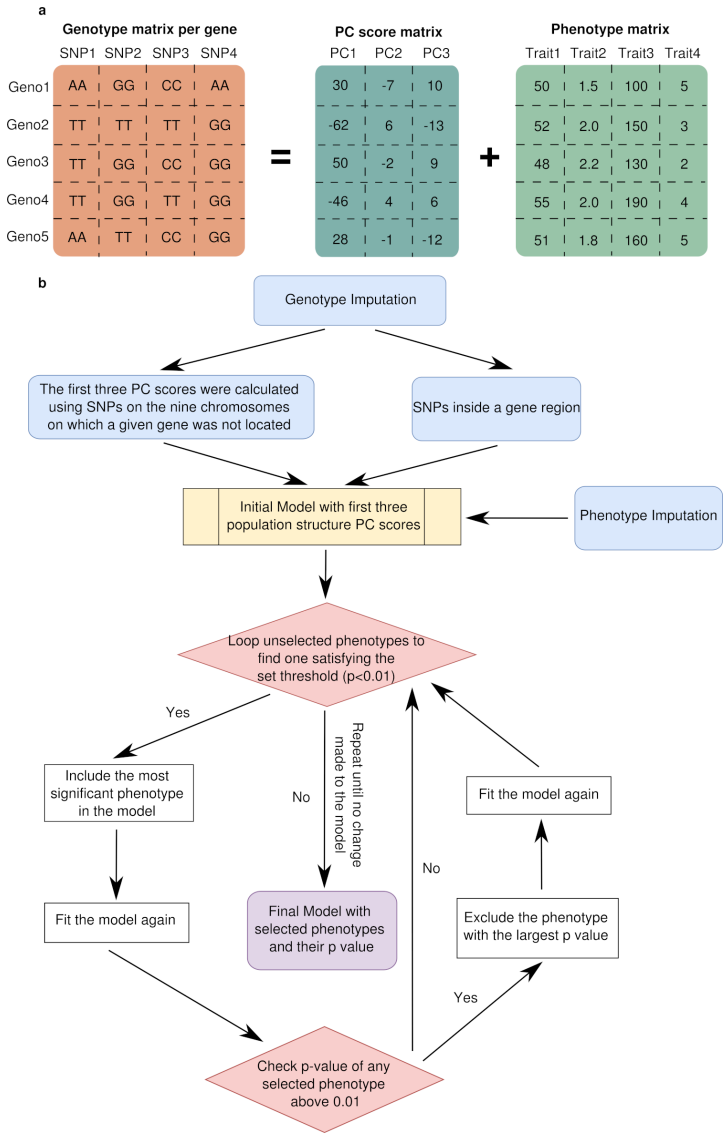

Figure 4.3: GPWAS algorithm implementation. (a) Example of trait and genotype matrices employed for GPWAS. (b) Flow chart of the initial data processing and the forward selection process within the GPWAS algorithm.

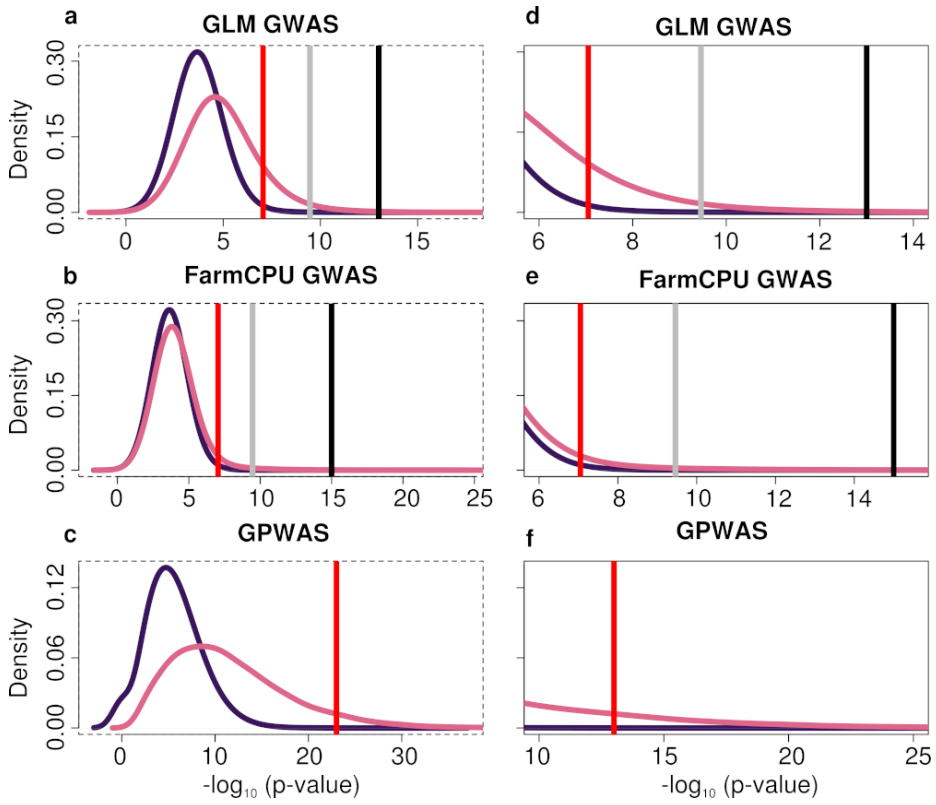

Figure 4.4: Permutation testing based estimation of false discovery rates for GLM GWAS, FarmCPU, and GPWAS. For each panel, the dark curve shows the distribution of per gene p-values obtained from 20 permutations of genotype and trait data (see Methods), while the light curve indicates the distribution 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 top=8.96e-8 for GLM and FarmCPU and an estimated FDR < 0.001 for GPWAS. Genes assigned p-values on the right side of each red line were employed for all downstream analyses in the main text. Panels a-c show the entirety of the distributions, while panels d-f display a magnified view of the regions of the curve where the p-value threshold is employed. When these data were used to estimate the p-value cut off corresponding to an estimated FDR < 0.001 for GLM GWAS, this was found to correspond to an uncorrected p-value of approximately 1e-14, resulting in 31 genes would remain statistically significantly associated with traits. For FarmCPU GWAS, the minimum FDR achieved was FDR < 0.029 at a p-value threshold of 1e-15, resulting in 38 genes remaining statistically significantly associated with traits.

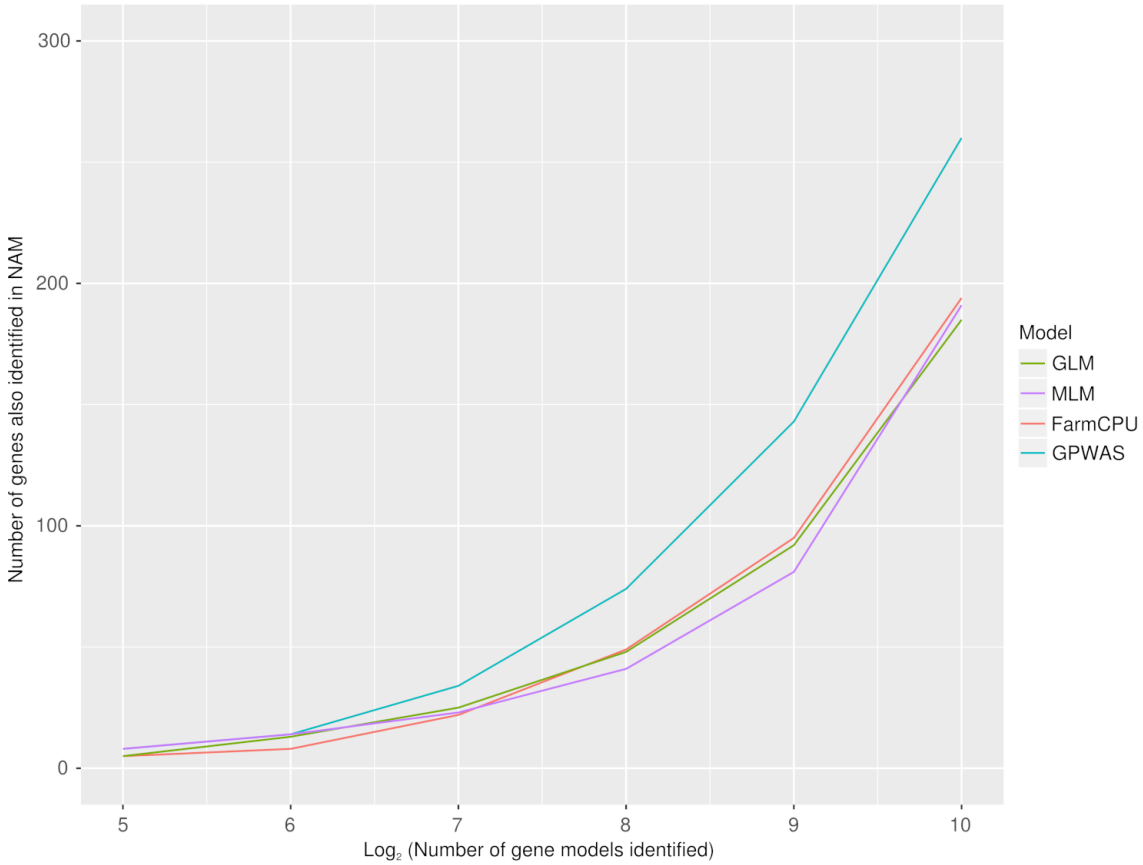

Figure 4.5: Comparison of the performance of GPWAS and conventional GWAS methods in the identification of candidate genes identified by Wallace et al[192] using genotypic and phenotypic data from the maize NAM population. Genes were sorted by p-value, and the genes with the most significant p-values were selected at each threshold number of significant genes listed on the x-axis.

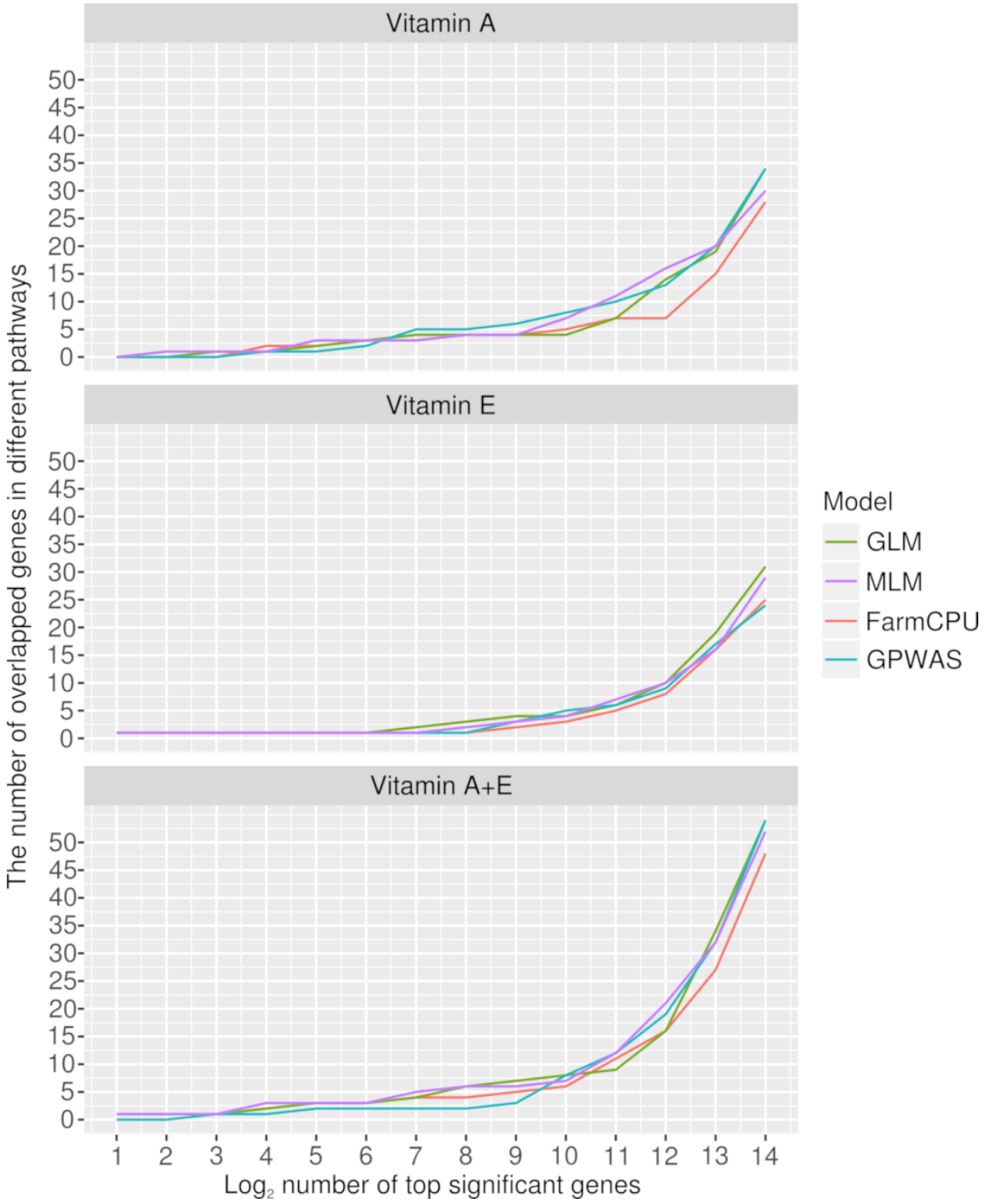

Figure 4.6: Comparison of the performance of GPWAS and conventional GWAS methods in the identification of *a prior* candidate genes involved in vitamin A and E biosynthesis. Phenotypic data and published *a priori* candidate gene lists for vitamin A and vitamin E were taken from previous studies.[174,199] The methodology used here was otherwise identical to that employed for Supplementary Figure 4.5.

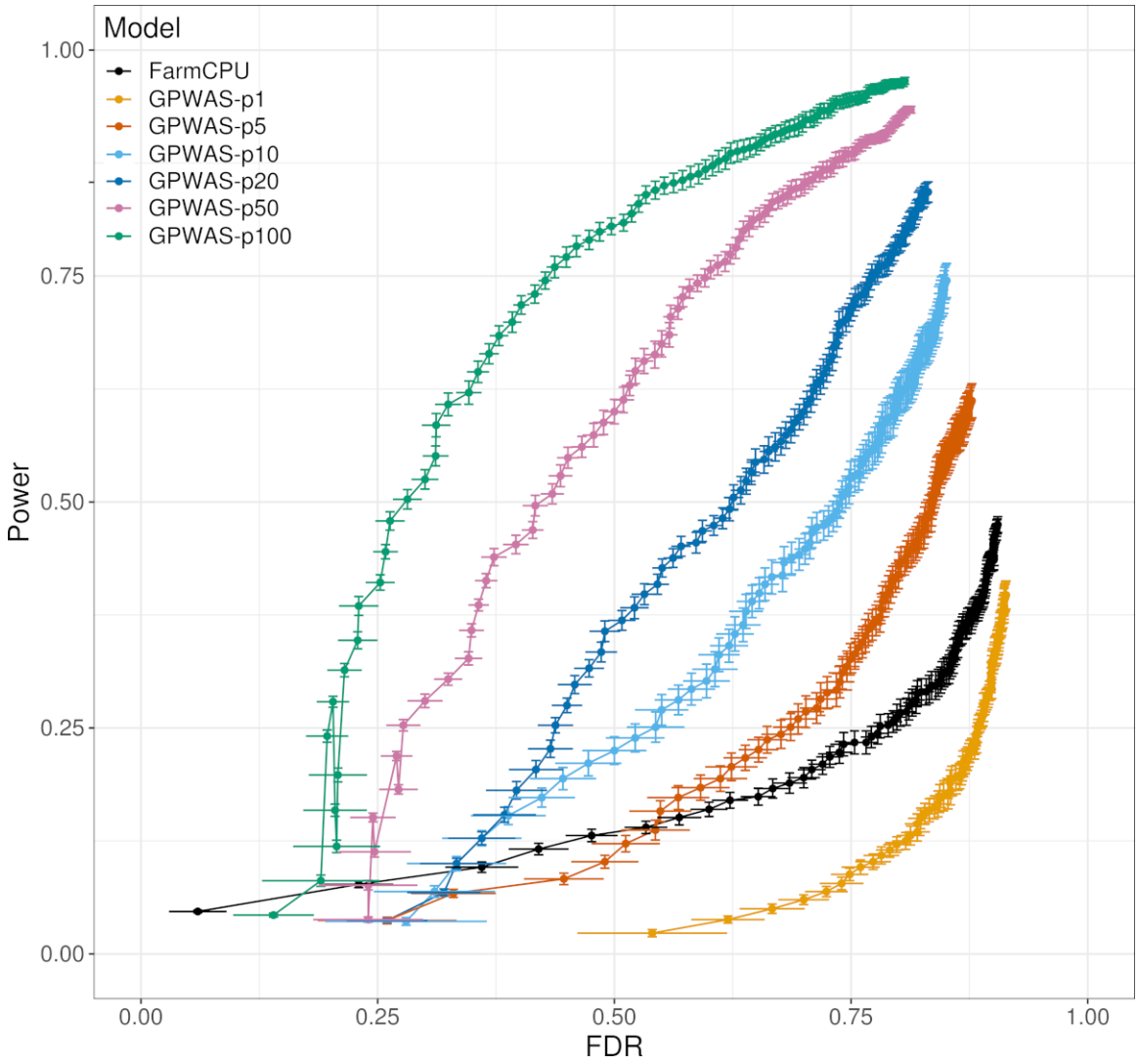

Figure 4.7: Power and FDR evaluation of the GPWAS model compared to the GWAS model based on simulated phenotypes. Ten random sets of 100 quantitative trait nucleotides (QTNs) were used to simulate 100 replicated phenotypes with h^2 of 0.5. For one simulated phenotype set, the positive genes were defined as the top m most significant of 2,000 genes. Each dot was represented as mean values of power and FDR of 10 replicates in each rank. Error bars in both vertical and horizontal ways were represented by standard errors of 10 replicates for power and FDR in each dot. The curve of power to FDR of GLM model is under FarmCPU (data not shown). GPWAS-p1 stands for using 1 simulated phenotype for running GPWAS, GPWAS-p2 stands for using 2 simulated phenotypes for running GPWAS. The same naming standard can be applied on GPWAS-p5, GPWAS-p10, GPWAS-p20, GPWAS-p50 and GPWAS-p100. Although more phenotypic information was incorporated into GPWAS model, it demonstrated a better power/false discovery trade-off relative to FarmCPU with only 1 trait.

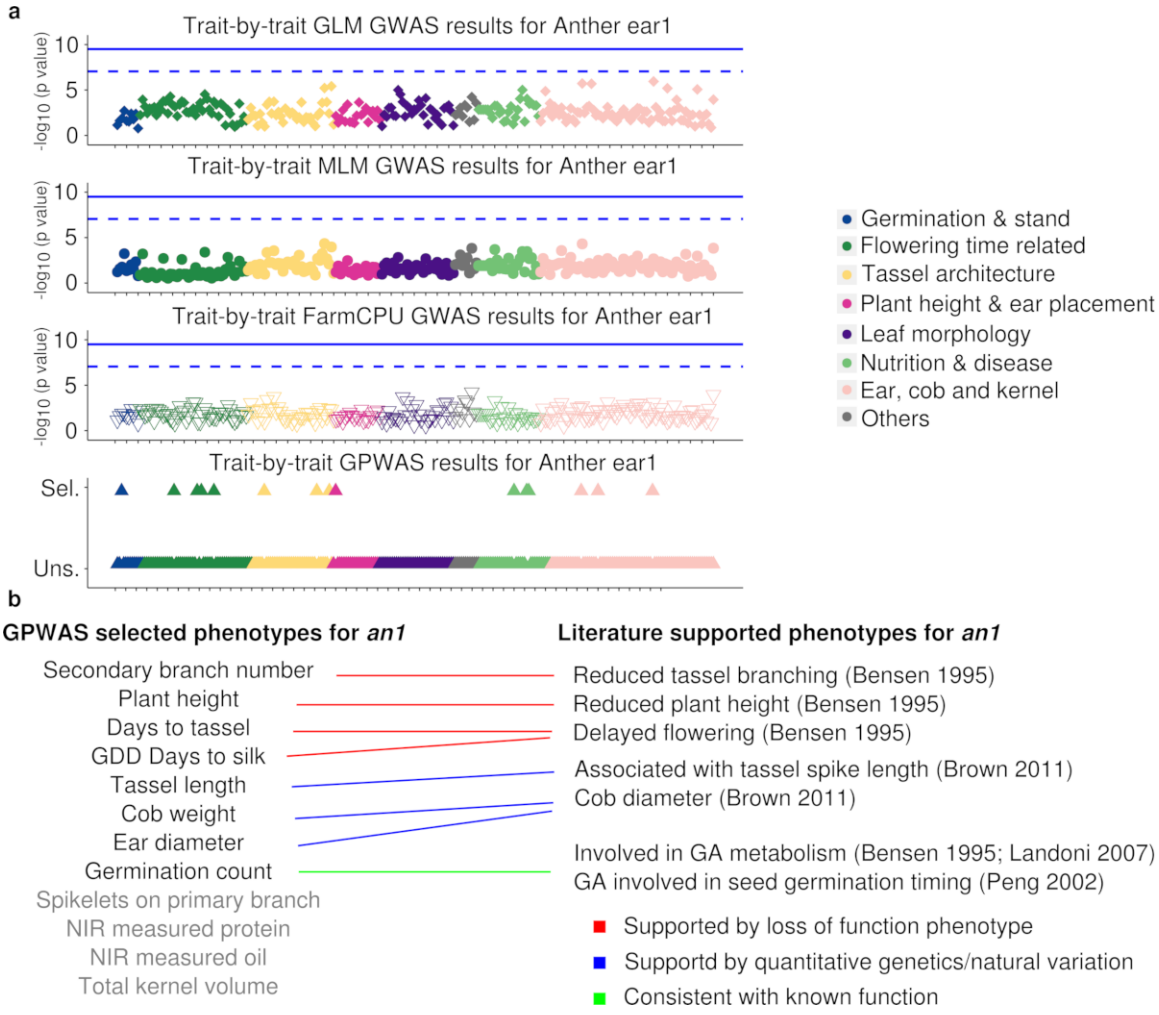

Figure 4.8: Evaluation of GLM GWAS, FarmCPU GWAS, and GPWAS using the known maize gene *Anther ear1* (*an1*) (Zm00001d032961). (a) The dashed lines indicates a p-value corresponding to 0.05 after a Bonferroni correction for independent tests on 557,968 (SNPs). Solid lines indicate the stricter multiple testing corrected threshold, which considers both the number of SNPs and the number of phenotypes tested. In the GPWAS panel, Sel. and Uns. indicate traits that were selected and unselected respectively, in the model GPWAS fit for this particular gene. Phenotypes are ordered along the x-axis in the same order used for Figure 1, with each tick mark indicating a distance of five phenotypes. Phenotypes incorporated in the GPWAS model for *an1* were as follows: germination count (Summer 2006, Johnston, NC), days to tassel (Summer 2007, Cayuga, NY), GDD days to silk (Summer 2007, Johnston, NC; Summer 2007, Champaign, IL; 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). (b) The potential correspondence between phenotypes selected using the GPWAS model for *an1* using the GPWAS model and phenotypes either reported for loss of function *an1* mutants or previous quantitative genetic analyses.[32, 200–202]

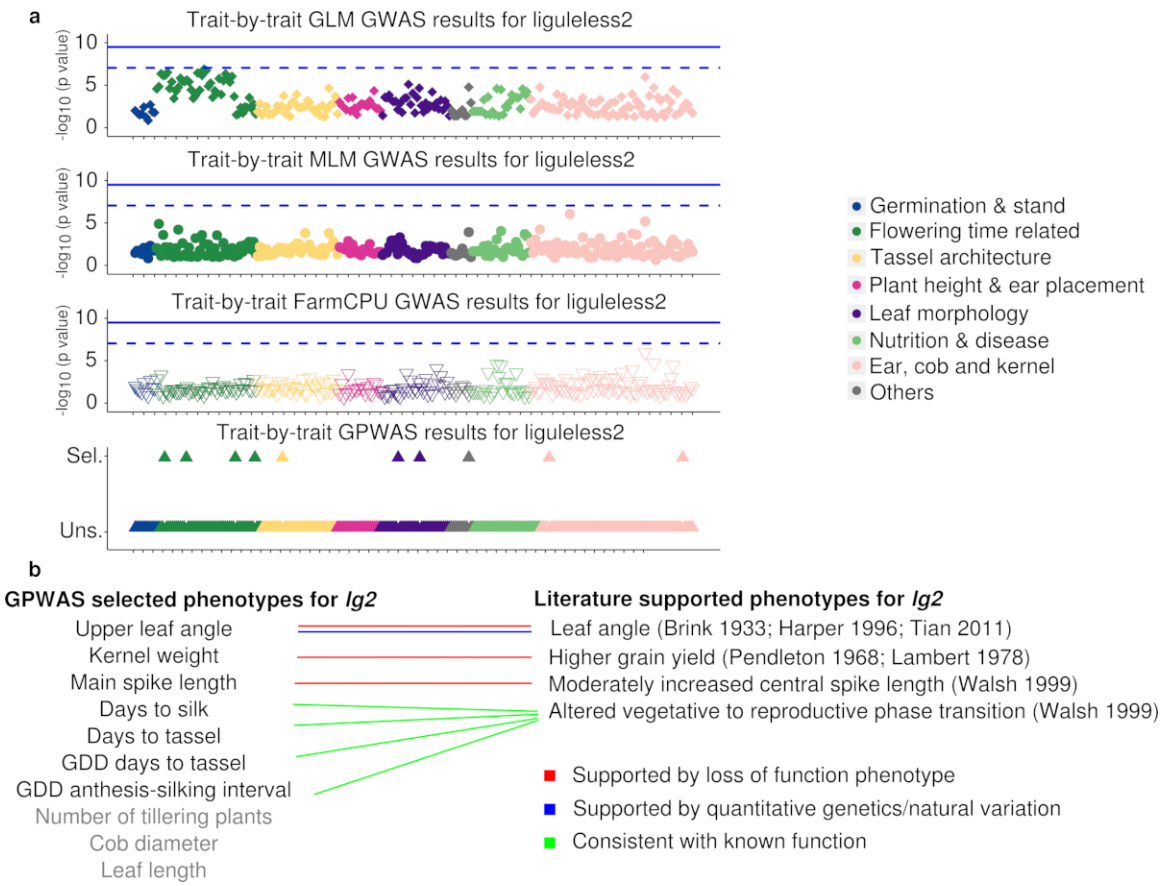

Figure 4.9: Evaluation of GLM GWAS, FarmCPU GWAS, and GPWAS using the known maize gene *liguleless2* (*lg2*) (Zm00001d042777). (a) The dashed lines indicates a p value corresponding to 0.05 after a Bonferroni correction for independent tests on 557,968 (SNPs). Solid lines indicate the stricter multiple testing corrected threshold which considers both the number of SNPs and the number of phenotypes tested. In the GPWAS panel, Sel. and Uns. indicate traits that were selected and unselected respectively, in the model GPWAS fit for this particular gene. Phenotypes are ordered along the x-axis in the same order used for Figure 1, with each tick mark indicating a distance of five phenotypes. Phenotypes incorporated in the GPWAS model for *lg2* were as follows: days to silk (Summer 2006, Johnston, NC), days to tassel (Winter 2006, Ponce, PR), GDD days to tassel (Summer 2007, Champaign, IL), GDD anthesis-silking interval (Winter 2007, Miami-Dade, FL), main spike length (Summer 2006, Johnston, NC), leaf length (Summer 2006, Boone, MO), upper leaf angle (Summer 2006, Cayuga, NY), number of tillering plants (Summer 2007, Cayuga, NY), cob diameter (Winter 2006, Ponce, PR) and kernel weight (Summer 2007, Cayuga, NY). (b) The potential correspondence between phenotypes selected by the GPWAS model for *lg2*, and phenotypes either reported for loss of function *lg2* mutants or previous quantitative genetic analyses.[74, 203, 205–208]

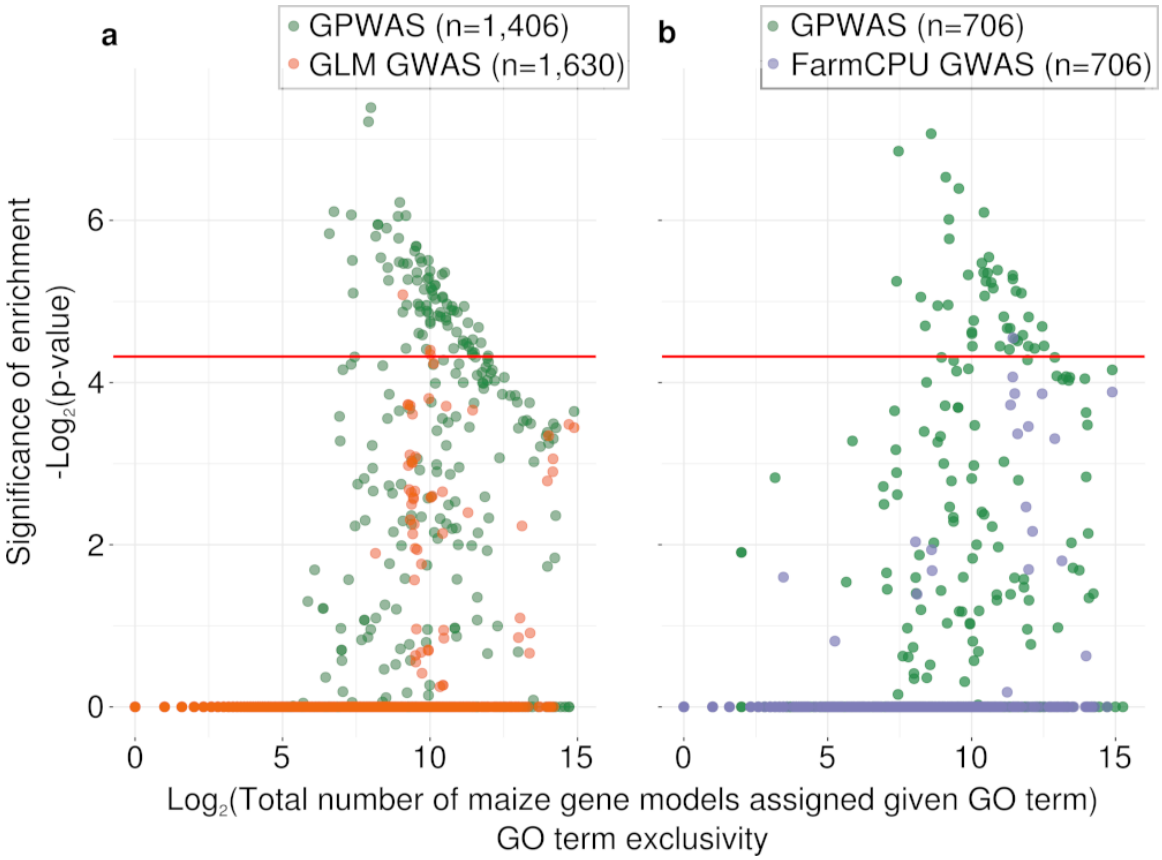

Figure 4.10: Comparison of GO enrichment/purification among genes uniquely identified as being associated with phenotypic variation using different statistical approaches. Each circle represents a single GO term in a single analysis. The position of each circle on the x axis indicates the total number of maize gene models which were assigned to this GO term in the maize GAMER dataset.[235] 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. Red lines indicate the threshold for determining a significant GO term after a Bonferroni correction. (a) Comparison of the patterns of GO term enrichment/purification among genes either uniquely identified as being associated with phenotypic variation using a GLM GWAS analysis or uniquely identified as being associated with phenotypic variation in a GPWAS analysis. (b) As in panel a, but the comparison is between genes uniquely identified as being associated with phenotypic variation using a FarmCPU analysis or uniquely identified as being associated with phenotypic variation in a GPWAS analysis. Only the 706 genes uniquely identified using GPWAS with the strongest statistical signal were employed in panel b, to prevent any bias towards more significant p-values resulting from an analysis using a larger population of genes identified using GPWAS than those identified using FarmCPU.

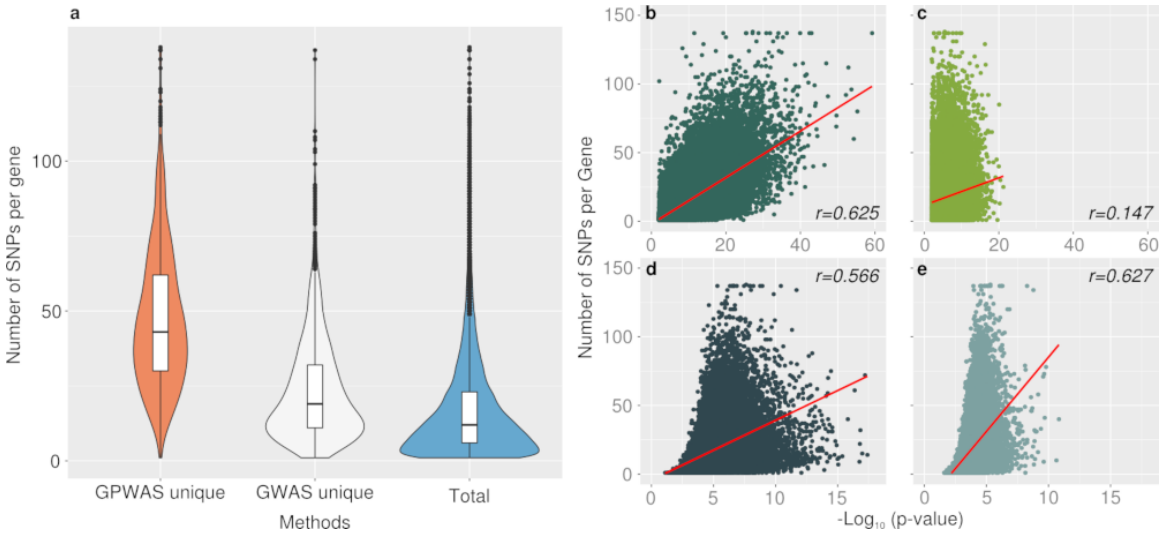

Figure 4.11: Number of SNPs identified per gene and the p-value of genes identified using different models. (a) The number of SNPs assigned to genes uniquely identified using either GPWAS or GLM GWAS, as well as the total number of genes with identified SNPs. SNPs assigned to gene regions were filtered and employed in all analyses. The maximum remaining number of SNPs per gene was 138. The distributions of the genes uniquely identified using GLM GWAS or GPWAS were statistically significantly different, $p < 2.2e-16$ (Mann-Whitney U test). (b) Correlations between the SNP number per gene and the -log10 p-value of the total number of genes identified using GPWAS on real phenotype data. (c) Correlations between the SNP number per gene and the -log10 p-value of the total genes identified using GPWAS on randomly selected phenotype data from 20 permutations. (d) Correlations between the SNP number per gene and the -log10 p-value of the total genes identified using GLM GWAS on real phenotype data. (e) Correlations between the SNP number per gene and -log10 p-value of total genes identified using GLM GWAS on randomly selected phenotype data from 20 permutations. Spearman correlation methods were employed for the correlation test between SNP number and -log10 transformed p-value for each gene. Full statistical reports are presented in Supplementary Table 7.

5:Summary

Genomic researches accelerate the process for understanding genetic basis of observed phenotypes in maize and other crop species. However, only a limited number of inbreds have been sequenced to serve as reference genomes. We need to align raw reads against reference genome to detect molecular signals in specific genomic regions, such gene expression level, single nucleotide polymorphisms and non-coding regulators. To accomplish these tasks, both assembled genome in high-quality and seed materials in pure genetic background are needed. The method we demonstrated in maize B73 population only requires RNA-seq data for investigating differences against the reference genome. This could potentially provides an approach to answer unexpected observations during genomic/transcriptomic studies, such as low correlations among biological replicates, introgressions into inbreds during propagation and extremely low expressions for certain genes.

With seed resources in high-confident genetic identity and advanced genotyping technologies, we can acquire high-confident genotype markers for research materials. To understand gene functions, precisely measured phenotypes are also needed, either at molecular level or visible level. High-throughput phenotyping technologies (HTP) enables us to measure dozens or hundreds of phenotypes per plant in an unified standard and high efficient way. Imaging is one of broadly applied methods for this phenotype collection process. Under well-controlled environment, we can extract diverse sets of numeric values from images in different types across the plant developmental stage. Potentially novel "trait", as the measurement bias between manually and

computationally estimated plant biomass, could be an example to show the application of HTP on trait discoveries. The future application of HTP measured trait associated with genetic variants could be verified from experiments for more functional gene mining.

In population level, a nature diversity panel consists of a large number of genotypes from different geographical sources. This big genetic pool gives the accessibility for evaluating associations between candidate genetic loci with investigated phenotypes. The application of GWAS can dissect the genetic architecture of a certain trait. However, except for the connection between genotype and phenotype, underlying relationships between phenotypes and phenotypes, or within molecular markers are highly complex. The emerging research direction in plant phenotyping can generate much more phenotypic measurements than before. A broad set of phenomic data enable us to test the null hypothesis to see if any of annotated genes can be significantly associated with these phenome-wide variants. The developed GPWAS model detected a distinct set of maize gene population from both background and conventional GWAS detected genes. The GPWAS genes are more conserved in functions and have closer genetic distance with classical maize knock-out gene mutants, which have been studied by many research groups. Overall, this dissertation highlights methods for utilizing precise and high-throughput genotype and phenotype data for functional gene discoveries in maize.

References

[1] Barbara McClintock. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355, 1950.

[2] Graham Moore, KM Devos, Z Wang, and MD Gale. Cereal genome evolution: grasses, line up and form a circle. *Current biology*, 5(7):737–739, 1995.

[3] Zuzana Swigoňová, Jinsheng Lai, Jianxin Ma, Wusirika Ramakrishna, Victor Llaca, Jeffrey L Bennetzen, and Joachim Messing. Close split of sorghum and maize genome progenitors. *Genome research*, 14(10a):1916–1923, 2004.

[4] James C Schnable, Nathan M Springer, and Michael Freeling. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences*, 108(10):4069–4074, 2011.

[5] Zhikai Liang and James C Schnable. Functional divergence between subgenomes and gene pairs after whole genome duplications. *Molecular plant*, 11(3):388–397, 2018.

[6] Eduard D Akhunov, Sunish Sehgal, Hanquan Liang, Shichen Wang, Alina R Akhunova, Gaganpreet Kaur, Wanlong Li, Kerrie L Forrest, Deven See, Hana Šimková, et al. Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant physiology*, 161(1):252–265, 2013.

[7] Yang Zhang, Daniel W Ngu, Daniel Carvalho, Zhikai Liang, Yumou Qiu, Rebecca Roston, and James Schnable. Differentially regulated orthologs in sorghum and the subgenomes of maize. *The Plant Cell*, pages tpc–00354, 2017.

[8] Xianjun Lai, Sairam Behera, Zhikai Liang, Yanli Lu, Jitender S Deogun, and James C Schnable. Stag-cns: An order-aware conserved noncoding sequences discovery tool for arbitrary numbers of species. *Molecular plant*, 10(7):990–999, 2017.

[9] Sherry A Flint-Garcia, Anne-Céline Thuillet, Jianming Yu, Gael Pressoir, Susan M Romero, Sharon E Mitchell, John Doebley, Stephen Kresovich, Major M Goodman, and Edward S Buckler. Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal*, 44(6):1054–1064, 2005.

[10] WA Russell. Registration of b70 and b73 parental lines of maize1 (reg. nos. pl16 and pl17). *Crop Science*, 12(5):721–721, 1972.

[11] Patrick S Schnable, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A Graves, et al. The b73 maize genome: complexity, diversity, and dynamics. *science*, 326(5956):1112–1115, 2009.

[12] Simon Ardui, Adam Ameur, Joris R Vermeesch, and Matthew S Hestand. Single molecule real-time (smrt) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research*, 46(5):2159–2168, 2018.

[13] Yinping Jiao, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C Stitzer, Bo Wang, Michael S Campbell, Joshua C Stein, Xuehong Wei, Chen-Shan Chin, et al. Improved maize reference genome with single-molecule technologies. *Nature*, 546(7659):524, 2017.

[14] Candice N Hirsch, Cory D Hirsch, Alex B Brohammer, Megan J Bowman, Ilya Soifer, Omer Barad, Doron Shem-Tov, Kobi Baruch, Fei Lu, Alvaro G Hernandez,

et al. Draft assembly of elite inbred line ph207 provides insights into genomic and transcriptome diversity in maize. *The Plant Cell*, 28(11):2700–2714, 2016.

[15] Ning Yang, Xi-Wen Xu, Rui-Ru Wang, Wen-Lei Peng, Lichun Cai, Jia-Ming Song, Wenqiang Li, Xin Luo, Luyao Niu, Yuebin Wang, et al. Contributions of zea mays subspecies mexicana haplotypes to modern maize. *Nature communications*, 8(1):1874, 2017.

[16] Nathan M Springer, Sarah N Anderson, Carson M Andorf, Kevin R Ahern, Fang Bai, Omer Barad, W Brad Barbazuk, Hank W Bass, Kobi Baruch, Gil Ben-Zvi, et al. The maize w22 genome provides a foundation for functional genomics and transposon biology. *Nature genetics*, 50(9):1282, 2018.

[17] Fei Lu, Maria C Romay, Jeffrey C Glaubitz, Peter J Bradbury, Robert J Elshire, Tianyu Wang, Yu Li, Yongxiang Li, Kassa Semagn, Xuecai Zhang, et al. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature communications*, 6:6914, 2015.

[18] Ning Yang, Jie Liu, Qiang Gao, Songtao Gui, Lu Chen, Linfeng Yang, Juan Huang, Tianquan Deng, Jingyun Luo, Lijuan He, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nature genetics*, 51(6):1052, 2019.

[19] Gregor Mendel. *Experiments in plant hybridisation*. 1866.

[20] Matthew J Dzievit, Xianran Li, and Jianming Yu. Dissection of leaf angle variation in maize through genetic mapping and meta-analysis. *The Plant Genome*, 12(1), 2019.

[21] I Pejic, P Ajmone-Marsan, M1 Morgante, V Kozumplick, P Castiglioni, G Taramino, and M1 Motto. Comparative analysis of genetic similarity among maize inbred lines detected by rflps, rapds, ssrs, and aflps. *Theoretical and Applied genetics*, 97(8):1248–1255, 1998.

[22] Robert J Elshire, Jeffrey C Glaubitz, Qi Sun, Jesse A Poland, Ken Kawamoto, Edward S Buckler, and Sharon E Mitchell. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PloS one*, 6(5):e19379, 2011.

[23] Daniel Money, Kyle Gardner, Zoë Migicovsky, Heidi Schwaninger, Gan-Yuan Zhong, and Sean Myles. Linkimpute: fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genetics*, 5(11):2383–2390, 2015.

[24] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[25] Zhikai Liang, Shashi K Gupta, Cheng-Ting Yeh, Yang Zhang, Daniel W Ngu, Ramesh Kumar, Hemant T Patil, Kanulal D Mungra, Dev Vart Yadav, Abhishek Rathore, et al. Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. *G3: Genes, Genomes, Genetics*, 8(7):2513–2522, 2018.

[26] Robert Bukowski, Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, Dawen Xu, Bicheng Yang, Chuanxiao Xie, et al. Construction of the third-generation zea mays haplotype map. *Gigascience*, 7(4):gix134, 2017.

[27] Billie Leff, Navin Ramankutty, and Jonathan A Foley. Geographic distribution of major crops across the world. *Global Biogeochemical Cycles*, 18(1), 2004.

[28] Kejun Liu, Major Goodman, Spencer Muse, J Stephen Smith, ED Buckler, and John Doebley. Genetic structure and diversity among maize inbred lines as inferred from dna microsatellites. *Genetics*, 165(4):2117–2128, 2003.

[29] Mark A Mikel and John W Dudley. Evolution of north american dent corn from public to proprietary germplasm. *Crop science*, 46(3):1193–1205, 2006.

[30] Michael Lee, Natalya Sharopova, William D Beavis, David Grant, Maria Katt, Deborah Blair, and Arnel Hallauer. Expanding the genetic map of maize with the intermated b73× mo17 (ibm) population. *Plant molecular biology*, 48(5-6):453–461, 2002.

[31] Jason P Cook, Michael D McMullen, James B Holland, Feng Tian, Peter Bradbury, Jeffrey Ross-Ibarra, Edward S Buckler, and Sherry A Flint-Garcia. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant physiology*, 158(2):824–834, 2012.

[32] Patrick J Brown, Narasimham Upadyayula, Gregory S Mahone, Feng Tian, Peter J Bradbury, Sean Myles, James B Holland, Sherry Flint-Garcia, Michael D McMullen, Edward S Buckler, et al. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS genetics*, 7(11):e1002383, 2011.

[33] Hui Li, Zhiyu Peng, Xiaohong Yang, Weidong Wang, Junjie Fu, Jianhua Wang, Yingjia Han, Yuchao Chai, Tingting Guo, Ning Yang, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nature genetics*, 45(1):43, 2013.

[34] Jianming Yu, James B Holland, Michael D McMullen, and Edward S Buckler. Genetic design and statistical power of nested association mapping in maize. *Genetics*, 178(1):539–551, 2008.

[35] Matteo DellâĂŹAcqua, Daniel M Gatti, Giorgio Pea, Federica Cattonaro, Frederik Coppens, Gabriele Magris, Aye L Hlaing, Htay H Aung, Hilde Nelissen, Joke Baute, et al. Genetic properties of the magic maize population: a new platform for high definition qtl mapping in zea mays. *Genome biology*, 16(1):167, 2015.

[36] James C Schnable and Michael Freeling. Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PloS one*, 6(3):e17855, 2011.

[37] Joseph L Gage, Diego Jarquin, Cinta Romay, Aaron Lorenz, Edward S Buckler, Shawn Kaeppler, Naser Alkhalifah, Martin Bohn, Darwin A Campbell, Jode Edwards, et al. The effect of artificial selection on phenotypic plasticity in maize. *Nature communications*, 8(1):1348, 2017.

[38] Aaron Kusmec, Srikant Srinivasan, Dan Nettleton, and Patrick S Schnable. Distinct genetic architectures for phenotype means and plasticities in zea mays. *Nature plants*, 3(9):715, 2017.

[39] A Hund, Y Fracheboud, A Soldati, E Frascaroli, S Salvi, and P Stamp. Qtl controlling root and shoot traits of maize seedlings under cold stress. *Theoretical and applied genetics*, 109(3):618–629, 2004.

[40] Edward S Buckler, James B Holland, Peter J Bradbury, Charlotte B Acharya, Patrick J Brown, Chris Browne, Elhan Ersoz, Sherry Flint-Garcia, Arturo Garcia, Jeffrey C Glaubitz, et al. The genetic architecture of maize flowering time. *Science*, 325(5941):714–718, 2009.

[41] Jason A Peiffer, Maria C Romay, Michael A Gore, Sherry A Flint-Garcia, Zhiwu Zhang, Mark J Millard, Candice AC Gardner, Michael D McMullen, James B Holland, Peter J Bradbury, et al. The genetic architecture of maize height. *Genetics*, 196(4):1337–1356, 2014.

[42] Moses M Muraya, Jianting Chu, Yusheng Zhao, Astrid Junker, Christian Klukas, Jochen C Reif, and Thomas Altmann. Genetic variation of growth dynamics in maize (zea mays l.) revealed through automated non-invasive phenotyping. *The Plant Journal*, 89(2):366–380, 2017.

[43] Sruti Das Choudhury, Vincent Stoerger, Ashok Samal, James C Schnable, Zhikai Liang, and Jin-Gang Yu. Automated vegetative stage phenotyping analysis of maize plants using visible light images. In *KDD workshop on data science for food, energy and water, San Francisco, California, USA*, 2016.

[44] Malachy T Campbell, Avi C Knecht, Bettina Berger, Chris J Brien, Dong Wang, and Harkamal Walia. Integrating image-based phenomics and association analysis to dissect the genetic architecture of temporal salinity responses in rice. *Plant physiology*, 168(4):1476–1489, 2015.

[45] Zhikai Liang, Piyush Pandey, Vincent Stoerger, Yuhang Xu, Yumou Qiu, Yufeng Ge, and James C Schnable. Conventional and hyperspectral time-series imaging of maize lines widely used in field trials. *GigaScience*, 7(2):gix117, 2017.

[46] Noah Fahlgren, Maximilian Feldman, Malia A Gehan, Melinda S Wilson, Christine Shyu, Douglas W Bryant, Steven T Hill, Colton J McEntee, Sankalpi N Warnasooriya, Indrajit Kumar, et al. A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in setaria. *Molecular plant*, 8(10):1520–1535, 2015.

[47] Abhiram Das, Hannah Schneider, James Burridge, Ana Karine Martinez Ascanio, Tobias Wojciechowski, Christopher N Topp, Jonathan P Lynch, Joshua S Weitz, and Alexander Bucksch. Digital imaging of root traits (dirt): a high-throughput computing and collaboration platform for field-based root phenomics. *Plant methods*, 11(1):51, 2015.

[48] Nathan D Miller, Nicholas J Haase, Jonghyun Lee, Shawn M Kaeppler, Natalia de Leon, and Edgar P Spalding. A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images. *The Plant Journal*, 89(1):169–178, 2017.

[49] Matthias Eberius and José Lima-Guerra. High-throughput plant phenotyping–data acquisition, transformation, and analysis. In *Bioinformatics*, pages 259–278. Springer, 2009.

[50] Dijun Chen, Kerstin Neumann, Swetlana Friedel, Benjamin Kilian, Ming Chen, Thomas Altmann, and Christian Klukas. Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. *The Plant Cell*, 26(12):4636–4655, 2014.

[51] Ben Ward, Chris Brien, Helena Oakey, Allison Pearson, Sónia Negrão, Rhiannon K Schilling, Julian Taylor, David Jarvis, Andy Timmins, Stuart J Roy, et al. High-throughput 3d modelling to dissect the genetic control of leaf elongation in barley (hordeum vulgare). *The Plant Journal*, 2019.

[52] Nathan Hughes, Hugo R Oliveira, Nick Fradgley, Fiona Corke, James Cockram, John H Doonan, and Candida Nibau. μct trait analysis reveals morphometric differences between domesticated temperate small grain cereals and their wild relatives. *The Plant Journal*, 2019.

[53] Sheng Wu, Weiliang Wen, Boxiang Xiao, Xinyu Guo, Jianjun Du, Chuanyu Wang, and Yongjian Wang. An accurate skeleton extraction approach from 3d point clouds of maize plants. *Frontiers in plant science*, 10, 2019.

[54] Christopher N Topp, Anjali S Iyer-Pascuzzi, Jill T Anderson, Cheng-Ruei Lee, Paul R Zurek, Olga Symonova, Ying Zheng, Alexander Bucksch, Yuriy Mileyko, Taras Galkovskyi, et al. 3d phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. *Proceedings of the National Academy of Sciences*, 110(18):E1695–E1704, 2013.

[55] Simon Madec, Fred Baret, Benoît De Solan, Samuel Thomas, Dan Dutartre, Stéphane Jezequel, Matthieu Hemmerlé, Gallian Colombeau, and Alexis Comar.

High-throughput phenotyping of plant height: comparing unmanned aerial vehicles and ground lidar estimates. *Frontiers in plant science*, 8:2002, 2017.

[56] Xiaqing Wang, Ruyang Zhang, Wei Song, Liang Han, Xiaolei Liu, Xuan Sun, Meijie Luo, Kuan Chen, Yunxia Zhang, Hao Yang, et al. Dynamic plant height qtl revealed in maize through remote sensing phenotyping using a high-throughput unmanned aerial vehicle (uav). *Scientific reports*, 9(1):3458, 2019.

[57] Piyush Pandey, Yufeng Ge, Vincent Stoerger, and James C Schnable. High throughput in vivo analysis of plant leaf chemical properties using hyperspectral imaging. *Frontiers in Plant Science*, 8, 2017.

[58] Andrew J Cal, Millicent Sanciangco, Maria Camila Rebolledo, Delphine Luquet, Rolando O Torres, Kenneth L McNally, and Amelia Henry. Leaf morphology, rather than plant water status, underlies genetic variation of rice leaf rolling under drought. *Plant, cell & environment*, 2019.

[59] Malia A Gehan, Noah Fahlgren, Arash Abbasi, Jeffrey C Berry, Steven T Callen, Leonardo Chavez, Andrew N Doust, Max J Feldman, Kerrigan B Gilbert, John G Hodge, et al. Plantcv v2. 0: Image analysis software for high-throughput plant phenotyping. Technical report, PeerJ Preprints, 2017.

[60] Jordan R Ubbens and Ian Stavness. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Frontiers in plant science*, 8:1190, 2017.

[61] Christian Klukas, Dijun Chen, and Jean-Michel Pape. Integrated analysis platform: an open-source information system for high-throughput plant phenotyping. *Plant physiology*, 165(2):506–518, 2014.

[62] Rajandeep S Sekhon, Haining Lin, Kevin L Childs, Candice N Hansey, C Robin Buell, Natalia de Leon, and Shawn M Kaeppler. Genome-wide atlas of transcription during maize development. *The Plant Journal*, 66(4):553–563, 2011.

[63] Scott C Stelpflug, Rajandeep S Sekhon, Brieanne Vaillancourt, Candice N Hirsch, C Robin Buell, Natalia de Leon, and Shawn M Kaeppler. An expanded maize gene expression atlas based on rna sequencing and its use to explore root development. *The plant genome*, 9(1), 2016.

[64] Karl AG Kremling, Shu-Yun Chen, Mei-Hsiu Su, Nicholas K Lepak, M Cinta Romay, Kelly L Swarts, Fei Lu, Anne Lorant, Peter J Bradbury, and Edward S Buckler. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, 555(7697):520, 2018.

[65] Fei Yi, Wei Gu, Jian Chen, Ning Song, Xiang Gao, Yingsi Zhou, Xuxu Ma, Weibin Song, Haiming Zhao, Eddi Esteban, et al. High-temporal-resolution transcriptome landscape of early maize seed development. *The Plant Cell*, pages tpc–00961, 2019.

[66] Stephen J Clark, Heather J Lee, Sébastien A Smallwood, Gavin Kelsey, and Wolf Reik. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology*, 17(1):72, 2016.

[67] Kook Hui Ryu, Ling Huang, Hyun Min Kang, and John Schiefelbein. Single-cell rna sequencing resolves molecular relationships among individual plant cells. *Plant physiology*, 179(4):1444–1456, 2019.

[68] Shaoqun Zhou, Karl Kremling, Nonoy Bandillo, Annett Richter, Ying K Zhang, Kevin R Ahern, Alexander B Artyukhin, Joshua X Hui, Gordon C Younkin, Frank C Schroeder, et al. Metabolome-scale genome-wide association studies reveal chemical diversity and genetic control of maize specialized metabolites. *The Plant Cell*, pages tpc–00772, 2019.

[69] Matthew Cobb. 60 years ago, francis crick changed the logic of biology. *PLoS biology*, 15(9):e2003243, 2017.

[70] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[71] Joan E Bailey-Wilson and Alexander F Wilson. Linkage analysis in the next-generation sequencing era. *Human heredity*, 72(4):228–236, 2011.

[72] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.

[73] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203, 2006.

[74] Feng Tian, Peter J Bradbury, Patrick J Brown, Hsiaoyi Hung, Qi Sun, Sherry Flint-Garcia, Torbert R Rocheford, Michael D McMullen, James B Holland, and Edward S Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics*, 43(2):159, 2011.

[75] Samuel Leiboff, Xianran Li, Heng-Cheng Hu, Natalie Todt, Jinliang Yang, Xiao Li, Xiaoqing Yu, Gary J Muehlbauer, Marja CP Timmermans, Jianming Yu, et al. Genetic control of morphometric diversity in the maize shoot apical meristem. *Nature communications*, 6:8974, 2015.

[76] Xianglan Wang, Hongwei Wang, Shengxue Liu, Ali Ferjani, Jiansheng Li, Jianbing Yan, Xiaohong Yang, and Feng Qin. Genetic variation in zmvpp1 contributes to drought tolerance in maize seedlings. *Nature genetics*, 48(10):1233, 2016.

[77] Ming Wang, Jianbing Yan, Jiuran Zhao, Wei Song, Xiaobo Zhang, Yannong Xiao, and Yonglian Zheng. Genome-wide association study (gwas) of resistance to head smut in maize. *Plant science*, 196:125–131, 2012.

[78] Inke R König. Validation in genetic association studies. *Briefings in bioinformatics*, 12(3):253–258, 2011.

[79] Zhikai Liang, Yumou Qiu, and James C Schnable. Distinct characteristics of genes associated with phenome-wide variation in maize (zea mays). *bioRxiv*, page 534503, 2019.

[80] Benjamin Brachi, Geoffrey P Morris, and Justin O Borevitz. Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology*, 12(10):232, 2011.

[81] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407, 2014.

[82] Yuan-Ming Zhang, Zhenyu Jia, and Jim M Dunwell. The applications of new multi-locus gwas methodologies in the genetic dissection of complex traits. *Frontiers in plant science*, 10, 2019.

[83] Hung-ying Lin, Qiang Liu, Xiao Li, Jinliang Yang, Sanzhen Liu, Yinlian Huang, Michael J Scanlon, Dan Nettleton, and Patrick S Schnable. Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by erd-gwas. *Genome biology*, 18(1):192, 2017.

[84] Yinping Jiao, Hainan Zhao, Longhui Ren, Weibin Song, Biao Zeng, Jinjie Guo, Baobao Wang, Zhipeng Liu, Jing Chen, Wei Li, et al. Genome-wide genetic changes during modern breeding of maize. *Nature genetics*, 44(7):812, 2012.

[85] James G Gethi, Joanne A Labate, Kendall R Lamkey, Margaret E Smith, and Stephen Kresovich. Ssr variation in important us maize inbred lines. *Crop Science*, 42(3):951–957, 2002.

[86] Sofía E Olmos. Genetic variability within accessions of the b73 maize inbred line. *Maydica*, 59(3):1–8, 2016.

[87] Roderick AF MacLeod, Wilhelm G Dirks, Yoshinobu Matsuo, Maren Kaufmann, Herbert Milch, and Hans G Drexler. Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *International Journal of Cancer*, 83(4):555–563, 1999.

[88] Brendan P Lucey, Walter A Nelson-Rees, and Grover M Hutchins. Henrietta lacks, hela cells, and cell culture contamination. *Archives of pathology & laboratory medicine*, 133(9):1463–1467, 2009.

[89] Tara A Enders, Sookyung Oh, Zhenbiao Yang, Beronda L Montgomery, and Lucia C Strader. Genome sequencing of arabidopsis abp1-5 reveals second-site mutations that may affect phenotypes. *The Plant Cell*, 27(7):1820–1826, 2015.

[90] William J Haun, David L Hyten, Wayne W Xu, Daniel J Gerhardt, Thomas J Albert, Todd Richmond, Jeffrey A Jeddeloh, Gaofeng Jia, Nathan M Springer, Carroll P Vance, et al. The composition and origins of genomic variation among individuals of the soybean reference cultivar williams 82. *Plant physiology*, 155(2):645–655, 2011.

[91] Mon-Ray Shao, Vikas Shedge, Hardik Kundariya, Fredric R Lehle, and Sally A Mackenzie. Ws-2 introgression in a proportion of arabidopsis thaliana col-0 stock seed produces specific phenotypes and highlights the importance of routine genetic verification. *The Plant Cell*, 28(3):603–605, 2016.

[92] Joy Bergelson, Edward S Buckler, Joseph R Ecker, Magnus Nordborg, and Detlef Weigel. A proposal regarding best practices for validating the identity of genetic stocks and the effects of genetic variants. *The Plant Cell*, 28(3):606–609, 2016.

[93] LL Darrah and MS Zuber. 1985 united states farm maize germplasm base and commercial breeding strategies 1. *Crop Science*, 26(6):1109–1113, 1986.

[94] Mark A Mikel. Genetic diversity and improvement of contemporary proprietary north american dent corn. *Crop science*, 48(5):1686–1695, 2008.

[95] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[96] Thomas D Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

[97] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[98] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[99] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Systematic biology*, 59(3):307–321, 2010.

[100] Jacques D Retief. Phylogenetic analysis using phylip. In *Bioinformatics methods and protocols*, pages 243–258. Springer, 2000.

[101] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562, 2012.

[102] Jer-Ming Chia, Chi Song, Peter J Bradbury, Denise Costich, Natalia De Leon, John Doebley, Robert J Elshire, Brandon Gaut, Laura Geller, Jeffrey C Glaubitz, et al. Maize hapmap2 identifies extant variation from a genome in flux. *Nature genetics*, 44(7):803, 2012.

[103] Sanzhen Liu, Cheng-Ting Yeh, Ho Man Tang, Dan Nettleton, and Patrick S Schnable. Gene mapping via bulked segregant rna-seq (bsr-seq). *PloS one*, 7(5):e36406, 2012.

[104] Robyn Johnston, Minghui Wang, Qi Sun, Anne W Sylvester, Sarah Hake, and Michael J Scanlon. Transcriptomic analyses indicate that maize ligule development recapitulates gene expression patterns that occur during lateral organ initiation. *The Plant Cell*, 26(12):4718–4732, 2014.

[105] Michael A Gore, Jer-Ming Chia, Robert J Elshire, Qi Sun, Elhan S Ersoz, Bonnie L Hurwitz, Jason A Peiffer, Michael D McMullen, George S Grills, Jeffrey Ross-Ibarra, et al. A first-generation haplotype map of maize. *Science*, 326(5956):1115–1117, 2009.

[106] Erik Vollbrecht, Leonore Reiser, and Sarah Hake. Shoot meristem size is dependent on inbred background and presence of the maize homeobox gene, knotted1. *Development*, 127(14):3161–3172, 2000.

[107] Candice N Hirsch, Jillian M Foerster, James M Johnson, Rajandeep S Sekhon, German Muttoni, Brieanne Vaillancourt, Francisco Peñagaricano, Erika Lindquist, Mary Ann Pedraza, Kerrie Barry, et al. Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell*, 26(1):121–135, 2014.

[108] Maria C Romay, Mark J Millard, Jeffrey C Glaubitz, Jason A Peiffer, Kelly L Swarts, Terry M Casstevens, Robert J Elshire, Charlotte B Acharya, Sharon E Mitchell, Sherry A Flint-Garcia, et al. Comprehensive genotyping of the usa national maize inbred seed bank. *Genome biology*, 14(6):R55, 2013.

[109] Nathalie Bolduc, Alper Yilmaz, Maria Katherine Mejia-Guerra, Kengo Morohashi, Devin O'Connor, Erich Grotewold, and Sarah Hake. Unraveling the knotted1 regulatory network in maize meristems. *Genes & development*, 26(15):1685–1690, 2012.

[110] PE Lipps, RC Pratt, and JJ Hakiza. Interaction of ht and partial resistance to exserohilum turcicum in maize. *Plant disease*, 81(3):277–282, 1997.

[111] Steven R Eichten, Roman Briskine, Jawon Song, Qing Li, Ruth Swanson-Wagner, Peter J Hermanson, Amanda J Waters, Evan Starr, Patrick T West, Peter Tiffin, et al. Epigenetic and genetic influences on dna methylation variation in maize populations. *The Plant Cell*, 25(8):2783–2797, 2013.

[112] Irina Makarevitch, Amanda J Waters, Patrick T West, Michelle Stitzer, Candice N Hirsch, Jeffrey Ross-Ibarra, and Nathan M Springer. Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS genetics*, 11(1):e1004915, 2015.

[113] Rajandeep S Sekhon, Roman Briskine, Candice N Hirsch, Chad L Myers, Nathan M Springer, C Robin Buell, Natalia de Leon, and Shawn M Kaeppler. Maize gene atlas developed by rna sequencing and comparative evaluation of transcriptomes based on rna sequencing and microarrays. *PLoS One*, 8(4):e61005, 2013.

[114] Jeffrey A Martin, Nicole V Johnson, Stephen M Gross, James Schnable, Xiandong Meng, Mei Wang, Devin Coleman-Derr, Erika Lindquist, Chia-Lin Wei, Shawn Kaeppler, et al. A near complete snapshot of the zea mays seedling transcriptome revealed from ultra-deep sequencing. *Scientific reports*, 4:4519, 2014.

[115] Anja Paschold, Yi Jia, Caroline Marcon, Steve Lund, Nick B Larson, Cheng-Ting Yeh, Stephan Ossowski, Christa Lanz, Dan Nettleton, Patrick S Schnable, et al. Complementation contributes to transcriptome complexity in maize (zea mays l.) hybrids relative to their inbred parents. *Genome research*, 22(12):2445–2454, 2012.

[116] Pinghua Li, Lalit Ponnala, Neeru Gandotra, Lin Wang, Yaqing Si, S Lori Tausta, Tesfamichael H Kebrom, Nicholas Provart, Rohan Patel, Christopher R Myers,

et al. The developmental dynamics of the maize leaf transcriptome. *Nature genetics*, 42(12):1060, 2010.

[117] Lin Wang, Angelika Czedik-Eysenberg, Rachel A Mertz, Yaqing Si, Takayuki Tohge, Adriano Nunes-Nesi, Stephanie Arrivault, Lauren K Dedow, Douglas W Bryant, Wen Zhou, et al. Comparative analyses of c 4 and c 3 photosynthesis in developing leaves of maize and rice. *Nature biotechnology*, 32(11):1158, 2014.

[118] Malachy T Campbell, Christopher A Proctor, Yongchao Dou, Aaron J Schmitz, Piyaporn Phansak, Greg R Kruger, Chi Zhang, and Harkamal Walia. Genetic and molecular characterization of submergence response identifies subtol6 as a major submergence tolerance locus in maize. *PLoS One*, 10(3):e0120385, 2015.

[119] Gibum Yi, Anjanasree K Neelakandan, Bryan C Gontarek, Erik Vollbrecht, and Philip W Becraft. The naked endosperm genes encode duplicate indeterminate domain transcription factors required for maize endosperm cell patterning and differentiation. *Plant physiology*, 167(2):443–456, 2015.

[120] Zachary H Lemmon, Robert Bukowski, Qi Sun, and John F Doebley. The role of cis regulatory evolution in maize domestication. *PLoS genetics*, 10(11):e1004745, 2014.

[121] Margaret H Frank, Molly B Edwards, Eric R Schultz, Michael R McKain, Zhangjun Fei, Iben Sørensen, Jocelyn KC Rose, and Michael J Scanlon. Dissecting the molecular signatures of apical cell-type shoot meristems from two ancient land plant lineages. *New Phytologist*, 207(3):893–904, 2015.

[122] Shawn R Thatcher, Wengang Zhou, April Leonard, Bing-Bing Wang, Mary Beatty, Gina Zastrow-Hayes, Xiangyu Zhao, Andy Baumgarten, and Bailin Li. Genome-wide analysis of alternative splicing in zea mays: landscape and genetic regulation. *The Plant Cell*, 26(9):3472–3487, 2014.

[123] Guangming He, Beibei Chen, Xuncheng Wang, Xueyong Li, Jigang Li, Hang He, Mei Yang, Lu Lu, Yijun Qi, Xiping Wang, et al. Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome biology*, 14(6):R57, 2013.

[124] Michael Regulski, Zhenyuan Lu, Jude Kendall, Mark TA Donoghue, Jon Reinders, Victor Llaca, Stephane Deschamps, Andrew Smith, Dan Levy, W Richard McCombie, et al. The maize methylome influences mrna splice sites and reveals widespread paramutation-like switches guided by small rna. *Genome research*, 23(10):1651–1662, 2013.

[125] Akshay Kakumanu, Madana MR Ambavaram, Curtis Klumas, Arjun Krishnan, Utlwang Batlang, Elijah Myers, Ruth Grene, and Andy Pereira. Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by rna-seq. *Plant Physiology*, 160(2):846–867, 2012.

[126] Andrea L Eveland, Alexander Goldshmidt, Michael Pautler, Kengo Morohashi, Christophe Liseron-Monfils, Michael W Lewis, Sunita Kumari, Susumu Hiraga, Fang Yang, Erica Unger-Wallace, et al. Regulatory modules controlling maize inflorescence architecture. *Genome research*, 24(3):431–443, 2014.

[127] Antony M Chettoor, Scott A Givan, Rex A Cole, Clayton T Coker, Erica Unger-Wallace, Zuzana Vejlupkova, Erik Vollbrecht, John E Fowler, and Matthew MS Evans. Discovery of novel transcripts and gametophytic functions via rna-seq analysis of maize gametophytic transcriptomes. *Genome biology*, 15(7):414, 2014.

[128] Mei Zhang, Shaojun Xie, Xiaomei Dong, Xin Zhao, Biao Zeng, Jian Chen, Hui Li, Weilong Yang, Hainan Zhao, Gaokui Wang, et al. Genome-wide high resolution parental-specific dna and histone methylation maps uncover patterns of imprinting regulation in maize. *Genome research*, 24(1):167–176, 2014.

[129] Jian Chen, Biao Zeng, Mei Zhang, Shaojun Xie, Gaokui Wang, Andrew Hauck, and Jinsheng Lai. Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant physiology*, 166(1):252–264, 2014.

[130] Junjie Fu, Yanbing Cheng, Jingjing Linghu, Xiaohong Yang, Lin Kang, Zuxin Zhang, Jie Zhang, Cheng He, Xuemei Du, Zhiyu Peng, et al. Rna sequencing reveals the complex regulatory network in the maize kernel. *Nature communications*, 4:2832, 2013.

[131] Claude Urbany, Andreas Benke, Johanna Marsian, Bruno Huettel, Richard Reinhardt, and Benjamin Stich. Ups and downs of a transcriptional landscape shape iron deficiency associated chlorosis of the maize inbreds b73 and mo17. *BMC plant biology*, 13(1):213, 2013.

[132] Nina Opitz, Anja Paschold, Caroline Marcon, Waqas Ahmed Malik, Christa Lanz, Hans-Peter Piepho, and Frank Hochholdinger. Transcriptomic complexity in young maize primary roots in response to low water potentials. *BMC genomics*, 15(1):741, 2014.

[133] Patricio Grassini, Kent M Eskridge, and Kenneth G Cassman. Distinguishing between yield advances and yield plateaus in historical crop production trends. *Nature communications*, 4:2918, 2013.

[134] Anja Hartmann, Tobias Czauderna, Roberto Hoffmann, Nils Stein, and Falk Schreiber. Htpheno: an image analysis pipeline for high-throughput plant phenotyping. *BMC bioinformatics*, 12(1):148, 2011.

[135] Xuehai Zhang, Chenglong Huang, Di Wu, Feng Qiao, Wenqiang Li, Lingfeng Duan, Ke Wang, Yingjie Xiao, Guoxing Chen, Qian Liu, et al. High-throughput phenotyping and qtl mapping reveals the genetic architecture of maize plant growth. *Plant physiology*, pages pp–01516, 2017.

[136] Rana Munns, Richard A James, Xavier RR Sirault, Robert T Furbank, and Hamlyn G Jones. New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. *Journal of Experimental Botany*, 61(13):3499–3507, 2010.

[137] Guillaume Lobet, Xavier Draye, and Claire Périlleux. An online database for plant image analysis software tools. *Plant methods*, 9(1):38, 2013.

[138] Darwin Campbell, Natalia de Leon, Jode Edwards, Jack Gardiner, Naser Al Khalifah, Carolyn Lawrence-Dill, Jane Petzoldt, Cinta Romay, Renee Walton, and the Genomes to Fields Cooperators (http://www.genomes2fields.org/). Genomes to fields 2016 data release. *CyVerse Data Commons*, 2016.

[139] Yufeng Ge, Geng Bai, Vincent Stoerger, and James C Schnable. Temporal dynamics of maize plant growth, water use, and leaf water content using automated high throughput rgb and hyperspectral imaging. *Computers and Electronics in Agriculture*, 127:625–632, 2016.

[140] JA Gamon and JS Surfus. Assessing leaf pigment content and activity with a reflectometer. *New Phytologist*, 143(1):105–117, 1999.

[141] Dijun Chen, Rongli Shi, Jean-Michel Pape, and Christian Klukas. Predicting plant biomass accumulation from image-derived parameters. *bioRxiv*, page 046656, 2016.

[142] Stephen Milborrow. Earth: multivariate adaptive regression spline models. *R package version*, 3:2–7, 2014.

[143] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[144] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel. Misc functions of the department of statistics (e1071), tu wien. *R package version*, pages 1–5, 2005.

[145] Sruti Das Choudhury, Vincent Stoerger, Ashok Samal, James C Schnable, Zhikai Liang, and Jin-Gang Yu. Automated vegetative stage phenotyping analysis of maize plants using visible light images. *Data Science for Food, Energy and Water workshop, San Francisco, California, USA, August 2016*.

[146] Nadia Al-Tamimi, Chris Brien, Helena Oakey, Bettina Berger, Stephanie Saade, Yung Shwen Ho, Sandra M Schmöckel, Mark Tester, and Sónia Negrão. Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. *Nature communications*, 7, 2016.

[147] Mahmood R Golzarian, Ross A Frick, Karthika Rajendran, Bettina Berger, Stuart Roy, Mark Tester, and Desmond S Lun. Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant methods*, 7(1):2, 2011.

[148] Nora Honsdorf, Timothy John March, Bettina Berger, Mark Tester, and Klaus Pillen. High-throughput phenotyping to detect drought tolerance qtl in wild barley introgression lines. *PLoS one*, 9(5):e97047, 2014.

[149] Elizabeth Heather Neilson, AM Edwards, CK Blomstedt, B Berger, B Lindberg Møller, and RM Gleadow. Utilization of a high-throughput shoot imaging system to examine the dynamic phenotypic responses of a c4 cereal crop plant to nitrogen and water deficiency over time. *Journal of experimental botany*, 66(7):1817–1832, 2015.

[150] James B Holland, Wyman E Nyquist, and Cuauhtemoc T Cervantes-Martínez. Estimating and interpreting heritability for plant breeding: an update. *Plant breeding reviews*, 22:9–112, 2003.

[151] Olaf Van Kooten and Jan FH Snel. The use of chlorophyll fluorescence nomenclature in plant stress physiology. *Photosynthesis research*, 25(3):147–150, 1990.

[152] Y Fracheboud, P Haldimann, J Leipner, and P Stamp. Chlorophyll fluorescence as a selection tool for cold tolerance of photosynthesis in maize (zea mays l.). *Journal of experimental botany*, 50(338):1533–1540, 1999.

[153] Hazem M Kalaji, Anjana Jajoo, Abdallah Oukarroum, Marian Brestic, Marek Zivcak, Izabela A Samborska, Magdalena D Cetner, Izabela Łukasik, Vasilij Goltsev, and Richard J Ladle. Chlorophyll a fluorescence as a tool to monitor physiological status of plants under abiotic stress conditions. *Acta Physiologiae Plantarum*, 38(4):102, 2016.

[154] Erik H Murchie and Tracy Lawson. Chlorophyll fluorescence analysis: a guide to good practice and understanding some new applications. *Journal of experimental botany*, 64(13):3983–3998, 2013.

[155] Luis Guanter, Yongguang Zhang, Martin Jung, Joanna Joiner, Maximilian Voigt, Joseph A Berry, Christian Frankenberg, Alfredo R Huete, Pablo Zarco-Tejada, Jung-Eun Lee, et al. Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence. *Proceedings of the National Academy of Sciences*, 111(14):E1327–E1333, 2014.

[156] PJ Zarco-Tejada, A Catalina, MR González, and P Martín. Relationships between net photosynthesis and steady-state chlorophyll fluorescence retrieved from airborne hyperspectral imagery. *Remote Sensing of Environment*, 136:247–258, 2013.

[157] Mainassara Zaman-Allah, O Vergara, JL Araus, A Tarekegne, C Magorokosho, PJ Zarco-Tejada, A Hornero, A Hernández Albà, B Das, P Craufurd, et al. Unmanned aerial platform-based multi-spectral imaging for field phenotyping of maize. *Plant methods*, 11(1):35, 2015.

[158] Craig Yendrek, Tiago Tomaz, Christopher M Montes, Youyuan Cao, Alison M Morse, Patrick J Brown, Lauren McIntyre, Andrew Leakey, and Elizabeth

Ainsworth. High-throughput phenotyping of maize leaf physiology and biochemistry using hyperspectral reflectance. *Plant physiology*, pages pp–01447, 2016.

[159] KL Smith, MD Steven, and JJ Colls. Use of hyperspectral derivative ratios in the red-edge region to identify plant stress responses to gas leaks. *Remote sensing of environment*, 92(2):207–217, 2004.

[160] Duli Zhao, K Raja Reddy, Vijaya Gopal Kakani, and VR Reddy. Nitrogen deficiency effects on plant growth, leaf photosynthesis, and hyperspectral reflectance properties of sorghum. *European Journal of Agronomy*, 22(4):391–403, 2005.

[161] Piotr Baranowski, Malgorzata Jedryczka, Wojciech Mazurek, Danuta Babula-Skowronska, Anna Siedliska, and Joanna Kaczmarek. Hyperspectral and thermal imaging of oilseed rape (brassica napus) response to fungal species of the genus alternaria. *PloS one*, 10(3):e0122913, 2015.

[162] Maize image phenotype dataset released in association with this paper.

[163] Publicly released genomes 2 fields 2014 field trial dataset.

[164] Publicly released genomes 2 fields 2015 field trial dataset. `https://doi.org/10.7946/P24S31`.

[165] Zhikai Liang, Piyush Pandey, Vincent Stoerger, Yuhang Xu, Yumou Qiu, Yufeng Ge, and James C Schnable. Supporting data for "conventional and hyperspectral time-series imaging of maize lines widely used in field trials". *GigaScience Database*, 2017.

[166] Karl Sax. The association of size differences with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genetics*, 8(6):552, 1923.

[167] GF Sprague. The location of dominant favorable genes in maize by means of an inversion. *Genetics*, 26(170):143–149, 1941.

[168] Susanna Atwell, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, Tina T Hu, et al. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627, 2010.

[169] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[170] Andrew DeWan, Mugen Liu, Stephen Hartman, Samuel Shao-Min Zhang, David TL Liu, Connie Zhao, Pancy OS Tam, Wai Man Chan, Dennis SC Lam, Michael Snyder, et al. Htra1 promoter polymorphism in wet age-related macular degeneration. *Science*, 314(5801):989–992, 2006.

[171] Karl Kremling, Christine Diepenbrock, Michael Gore, Edward Buckler, and Nonoy Bandillo. Transcriptome-wide association supplements genome-wide association in zea mays. *bioRxiv*, page 363242, 2018.

[172] Weiwei Wen, Dong Li, Xiang Li, Yanqiang Gao, Wenqiang Li, Huihui Li, Jie Liu, Haijun Liu, Wei Chen, Jie Luo, et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature communications*, 5:3438, 2014.

[173] Fumio Matsuda, Ryo Nakabayashi, Zhigang Yang, Yozo Okazaki, Jun-ichi Yonemaru, Kaworu Ebana, Masahiro Yano, and Kazuki Saito. Metabolome-genome-wide association study dissects genetic architecture for

generating natural variation in rice secondary metabolism. *The Plant Journal*, 81(1):13–23, 2015.

[174] Christine H Diepenbrock, Catherine B Kandianis, Alexander E Lipka, Maria Magallanes-Lundback, Brieanne Vaillancourt, Elsa Góngora-Castillo, Jason G Wallace, Jason Cepela, Alex Mesberg, Peter Bradbury, et al. Novel loci underlie natural variation in vitamin e levels in maize grain. *The Plant Cell*, pages tpc–00475, 2017.

[175] Achim Walter, Frank Liebisch, and Andreas Hund. Plant phenotyping: from bean weighing to image analysis. *Plant methods*, 11(1):14, 2015.

[176] José Luis Araus, Shawn C Kefauver, Mainassara Zaman-Allah, Mike S Olsen, and Jill E Cairns. Translating high-throughput phenotyping into genetic gain. *Trends in plant science*, 23(5):451–466, 2018.

[177] Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, 44(9):1066, 2012.

[178] Paul F OâĂŹReilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

[179] Sophie Van der Sluis, Danielle Posthuma, and Conor V Dolan. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS genetics*, 9(1):e1003235, 2013.

[180] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8(7):e65245, 2013.

[181] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821, 2012.

[182] Yifan Wang, Aiyi Liu, James L Mills, Michael Boehnke, Alexander F Wilson, Joan E Bailey-Wilson, Momiao Xiong, Colin O Wu, and Ruzong Fan. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic epidemiology*, 39(4):259–275, 2015.

[183] Patrick Turley, Raymond K Walters, Omeed Maghzian, Aysu Okbay, James J Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A Furlotte, et al. Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature genetics*, 50(2):229, 2018.

[184] William Pitchers, Jessica Nye, Eladio J Márquez, Alycia Kowalski, Ian Dworkin, and David Houle. A multivariate genome-wide association study of wing shape in drosophila melanogaster. *Genetics*, pages genetics–301342, 2019.

[185] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[186] SA Pendergrass, K Brown-Gentry, SM Dudek, ES Torstenson, JL Ambite, CL Avery, S Buyske, C Cai, MD Fesinmeyer, C Haiman, et al. The use of phenome-wide association studies (phewas) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic epidemiology*, 35(5):410–422, 2011.

[187] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, et al. Systematic comparison of phenome-wide association study of electronic

medical record data and genome-wide association study data. *Nature biotechnology*, 31(12):1102, 2013.

[188] Sarah A Pendergrass, Kristin Brown-Gentry, Scott Dudek, Alex Frase, Eric S Torstenson, Robert Goodloe, Jose Luis Ambite, Christy L Avery, Steve Buyske, Petra Bžková, et al. Phenome-wide association study (phewas) for detection of pleiotropy within the population architecture using genomics and epidemiology (page) network. *PLoS genetics*, 9(1):e1003087, 2013.

[189] Joshua C Denny, Lisa Bastarache, and Dan M Roden. Phenome-wide association studies as a tool to advance precision medicine. *Annual review of genomics and human genetics*, 17:353–373, 2016.

[190] Khader Shameer, Joshua C Denny, Keyue Ding, Hayan Jouni, David R Crosslin, Mariza De Andrade, Christopher G Chute, Peggy Peissig, Jennifer A Pacheco, Rongling Li, et al. A genome-and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Human genetics*, 133(1):95–109, 2014.

[191] Yaping Lu, Yemao Liu, Xiaohui Niu, Qingyong Yang, Xuehai Hu, Hong-Yu Zhang, and Jingbo Xia. Systems genetic validation of the snp-metabolite association in rice via metabolite-pathway-based phenome-wide association scans. *Frontiers in plant science*, 6:1027, 2015.

[192] Jason G Wallace, Peter J Bradbury, Nengyi Zhang, Yves Gibon, Mark Stitt, and Edward S Buckler. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS genetics*, 10(12):e1004845, 2014.

[193] Wei Zhao, Payan Canaran, Rebecca Jurkuta, Theresa Fulton, Jeffrey Glaubitz, Edward Buckler, John Doebley, Brandon Gaut, Major Goodman, Jim Holland, et al.

Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Research*, 34(suppl_1):D752–D757, 2006.

[194] Andrew Dahl, Valentina Iotchkova, Amelie Baud, Åsa Johansson, Ulf Gyllensten, Nicole Soranzo, Richard Mott, Andreas Kranis, and Jonathan Marchini. A multiple-phenotype imputation method for genetic studies. *Nature genetics*, 47(3):466, 2015.

[195] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

[196] Bahram Namjou, Keith Marsolo, Robert Carroll, Joshua Denny, Marylyn D Ritchie, Todd Lingren, Aleksey Porollo, Cassandra Perry, Leah Claire Kottyan, Ingrid Adele Holm, et al. Phenome-wide association study (phewas) in emr-linked pediatric cohorts. *Frontiers in genetics*, 5:401, 2014.

[197] Xiaolei Liu, Meng Huang, Bin Fan, Edward S Buckler, and Zhiwu Zhang. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS genetics*, 12(2):e1005767, 2016.

[198] Michael D McMullen, Stephen Kresovich, Hector Sanchez Villeda, Peter Bradbury, Huihui Li, Qi Sun, Sherry Flint-Garcia, Jeffry Thornsberry, Charlotte Acharya, Christopher Bottoms, et al. Genetic properties of the maize nested association mapping population. *Science*, 325(5941):737–740, 2009.

[199] Brenda F Owens, Alexander E Lipka, Maria Magallanes-Lundback, Tyler Tiede, Christine H Diepenbrock, Catherine B Kandianis, Eunha Kim, Jason Cepela, Maria Mateos-Hernandez, C Robin Buell, et al. A foundation for provitamin a biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. *Genetics*, 198(4):1699–1716, 2014.

[200] Robert J Bensen, Gurmukh S Johal, Virginia C Crane, John T Tossberg, Patrick S Schnable, Robert B Meeley, and Steven P Briggs. Cloning and characterization of the maize an1 gene. *The Plant Cell*, 7(1):75–84, 1995.

[201] Jinrong Peng and Nicholas P Harberd. The role of ga-mediated signalling in the control of seed germination. *Current opinion in plant biology*, 5(5):376–381, 2002.

[202] Michela Landoni, Francesca Dalla Vecchia, Giuseppe Gavazzi, Anna Giulini, Nicoletta La Rocca, Nicoletta Rascio, Monica Colombo, Monica Bononi, and Gabriella Consonni. The an1-4736 mutation of anther ear1 in maize alters scotomorphogenesis and the light response. *Plant science*, 172(1):172–180, 2007.

[203] RA Brink. Heritable characters in maize: XlviâĂŤligu1eless-2. *Journal of Heredity*, 24(8):325–326, 1933.

[204] Justine Walsh, Cynthia A Waters, and Michael Freeling. The maize geneliguleless2 encodes a basic leucine zipper protein involved in the establishment of the leaf blade–sheath boundary. *Genes & Development*, 12(2):208–218, 1998.

[205] Lisa Harper and Michael Freeling. Interactions of liguleless1 and liguleless2 function during ligule induction in maize. *Genetics*, 144(4):1871–1882, 1996.

[206] JW Pendleton, GE Smith, SR Winter, and TJ Johnston. Field investigations of the relationships of leaf angle in corn (zea mays l.) to grain yield and apparent photosynthesis 1. *Agronomy Journal*, 60(4):422–424, 1968.

[207] RJ Lambert and RR Johnson. Leaf angle, tassel morphology, and the performance of maize hybrids 1. *Crop Science*, 18(3):499–502, 1978.

[208] Justine Walsh and Michael Freeling. The liguleless2 gene of maize functions during the transition from the vegetative to the reproductive shoot apex. *The Plant Journal*, 19(4):489–495, 1999.

[209] Manfei Li, Wanshun Zhong, Fang Yang, and Zuxin Zhang. Genetic and molecular mechanisms of quantitative trait loci controlling maize inflorescence architecture. *Plant and Cell Physiology*, 59(3):448–457, 2018.

[210] Lei Liu, Yanfang Du, Xiaomeng Shen, Manfei Li, Wei Sun, Juan Huang, Zhijie Liu, Yongsheng Tao, Yonglian Zheng, Jianbing Yan, et al. Krn4 controls quantitative variation in maize kernel row number. *PLoS genetics*, 11(11):e1005670, 2015.

[211] Geoffrey P Morris, Punna Ramu, Santosh P Deshpande, C Thomas Hash, Trushar Shah, Hari D Upadhyaya, Oscar Riera-Lizarazu, Patrick J Brown, Charlotte B Acharya, Sharon E Mitchell, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences*, 110(2):453–458, 2013.

[212] Xuehui Huang, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, Chuanrang Zhu, Tingting Lu, Zhiwu Zhang, Meng Li, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics*, 42(11):961, 2010.

[213] Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.

[214] HP Piepho, J Möhring, AE Melchinger, and A Büchse. Blup for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209–228, 2008.

[215] Eli Rodgers-Melnick, Daniel L Vera, Hank W Bass, and Edward S Buckler. Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences*, 113(22):E3177–E3184, 2016.

[216] Anthony Studer, Qiong Zhao, Jeffrey Ross-Ibarra, and John Doebley. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature genetics*, 43(11):1160, 2011.

[217] Sara Castelletti, Roberto Tuberosa, Massimo Pindo, and Silvio Salvi. A mite transposon insertion is associated with differential methylation at the maize flowering time qtl vgt1. *G3: Genes, Genomes, Genetics*, 4(5):805–812, 2014.

[218] David L Remington, Jeffry M Thornsberry, Yoshihiro Matsuoka, Larissa M Wilson, Sherry R Whitt, John Doebley, Stephen Kresovich, Major M Goodman, and Edward S Buckler. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences*, 98(20):11479–11484, 2001.

[219] Wenli Zhang, Yufeng Wu, James C Schnable, Zixian Zeng, Michael Freeling, Gregory E Crawford, and Jiming Jiang. High-resolution mapping of open chromatin in the rice genome. *Genome research*, 22(1):151–162, 2012.

[220] Gina Turco, James C Schnable, Brent Pedersen, and Michael Freeling. Automated conserved non-coding sequence (cns) discovery reveals differences in gene content and promoter evolution among grasses. *Frontiers in plant science*, 4:170, 2013.

[221] Rurika Oka, Johan Zicola, Blaise Weber, Sarah N Anderson, Charlie Hodgman, Jonathan I Gent, Jan-Jaap Wesselink, Nathan M Springer, Huub CJ Hoefsloot, Franziska Turck, et al. Genome-wide mapping of transcriptional enhancer candidates using dna and chromatin features in maize. *Genome biology*, 18(1):137, 2017.

[222] Zefu Lu, William A Ricci, Robert J Schmitz, and Xiaoyu Zhang. Identification of cis-regulatory elements by chromatin structure. *Current opinion in plant biology*, 42:90–94, 2018.

[223] John P Lloyd, Zing Tsung-Yeh Tsai, Rosalie P Sowers, Nicholas L Panchy, and Shin-Han Shiu. A model-based approach for identifying functional intergenic

transcribed regions and noncoding rnas. *Molecular biology and evolution*, 35(6):1422–1436, 2018.

[224] Jeffrey L Bennetzen, Craig Coleman, Renyi Liu, Jianxin Ma, and Wusirika Ramakrishna. Consistent over-estimation of gene number in complex plant genomes. *Current opinion in plant biology*, 7(6):732–736, 2004.

[225] Mark B Gerstein, Can Bruce, Joel S Rozowsky, Deyou Zheng, Jiang Du, Jan O Korbel, Olof Emanuelsson, Zhengdong D Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-encode? history and updated definition. *Genome research*, 17(6):669–681, 2007.

[226] James C Schnable. Genome evolution in maize: from genomes back to genes. *Annual Review of Plant Biology*, 66:329–343, 2015.

[227] Brian L Browning and Sharon R Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[228] Jin Zhang, JY Feng, YL Ni, YJ Wen, Y Niu, CL Tamba, C Yue, Q Song, and YM Zhang. plarmeb: integration of least angle regression with empirical bayes for multilocus genome-wide association studies. *Heredity*, 118(6):517, 2017.

[229] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833, 2011.

[230] Xiang Zhou. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics*, 11(4):2027, 2017.

[231] James C Schnable. Sorghum version 3, maize versions 3 and 4 syntenic gene list. *FigShare*.

[232] Ryan F McCormick, Sandra K Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, Mojgan Amirebrahimi, Brock D Weers, Brian McKinley, et al. The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, 93(2):338–354, 2018.

[233] Jeffrey L Bennetzen, Jeremy Schmutz, Hao Wang, Ryan Percifield, Jennifer Hawkins, Ana C Pontaroli, Matt Estep, Liang Feng, Justin N Vaughn, Jane Grimwood, et al. Reference genome sequence of the model plant setaria. *Nature biotechnology*, 30(6):555, 2012.

[234] Alex B Brohammer, Thomas JY Kono, Nathan M Springer, Suzanne E McGaugh, and Candice N Hirsch. The limited role of differential fractionation in genome content variation and function in maize (zea mays l.) inbred lines. *The Plant Journal*, 93(1):131–141, 2018.

[235] Kokulapalan Wimalanathan, Iddo Friedberg, Carson M Andorf, and Carolyn J Lawrence-Dill. Maize go annotation-methods, evaluation, and review (maize-gamer). *Plant Direct*, 2(4):e00052, 2018.

[236] DV Klopfenstein, Liangsheng Zhang, Brent S Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J Mungall, Jeffrey M Yunes, Olga Botvinnik, Mark Weigel, et al. Goatools: A python library for gene ontology analyses. *Scientific reports*, 8(1):10872, 2018.

[237] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.