7-30-2024

Identifying Genes Linked to Variation in Metabolic and Whole Plant Phenotypes using Data from Genome Resequencing, Transcriptomics, and Metabolic Profiling of a Field-Grown Maize Diversity Panel

Ramesh Kanna Mathivanan
University of Nebraska-Lincoln

IDENTIFYING GENES LINKED TO VARIATION IN METABOLIC AND WHOLE PLANT PHENOTYPES USING DATA FROM GENOME RESEQUENCING, TRANSCRIPTOMICS, AND METABOLIC PROFILING OF A FIELD-GROWN MAIZE DIVERSITY PANEL

by

Ramesh Kanna Mathivanan

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Agronomy

Under the Supervision of James C. Schnable

Lincoln, Nebraska

August,2024

IDENTIFYING GENES LINKED TO VARIATION IN METABOLIC AND WHOLE PLANT PHENOTYPES USING DATA FROM GENOME RESEQUENCING, TRANSCRIPTOMICS, AND METABOLIC PROFILING OF A FIELD-GROWN MAIZE DIVERSITY PANEL

Ramesh Kanna Mathivanan, M.S.

University of Nebraska, 2024

Adviser: James C. Schnable

Maize metabolism is highly complex and influenced by genetic variation, yet the specific genes contributing to this variation and their links to non-metabolic traits remain less understood. To address this knowledge gap, we identified genes involved in maize metabolic variation and linked them to non-metabolic traits. We utilized a quadruplicate dataset of whole genome resequencing, transcriptomic, metabolic, and whole plant phenotype data from a single common field experiment of 660 diverse maize inbred lines. Leaf samples were collected shortly before flowering and analyzed using GC-MS for 26 metabolites. A Resampling Model Inclusion Probability Genome-Wide Association Study (RMIPGWAS) of approximately 2.6 million SNPs was conducted for these metabolites, identifying 155 candidate genes, with 17 showing particularly strong signals. A parallel Transcriptome Wide Association Study (TWAS) identified 6 candidate genes. A random forest feature importance-based approach identified one overlapping gene, Cu(2+)-exporting ATPase, and other genes not found by TWAS. Three loci associated with metabolite traits were also linked to non-metabolic traits in RMIPGWAS of 41 non-metabolic traits, including whole plant phenotypes, hyperspectral, and photosynthetic traits. Key genes identified include Zm00001eb270570 (Theobromine synthase), Zm00001eb354560 (Ubiquitin carboxyl-terminal hydrolase), and Zm00001eb051410 (N-acetyl-gamma-glutamyl-phosphate reductase). Our analysis showed that each method identified unique sets of genes associated with metabolite variation, demonstrating the complementary nature of different genomic approaches. The use of machine learning techniques like RF is crucial for identifying genes from gene expression data. These findings facilitate further studies on the roles of these metabolites and genes in plant growth and development.

Key Words: Maize, Plant Metabolism, GWAS, TWAS and RF.

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my beloved Appa, Amma, Anna, Anni and all my family members and well-wishers for their unwavering support and encouragement. A special thanks to Nikee Shrestha, Waqar Ali, Sai Subhash, Honyu Jin, Michael Tross, Jensina Davis, Vladimir Torres, Jon Turkus, Chidanand Ullagaddi, Harshita Mangal, and Kyle Linders for their assistance and companionship.

I am deeply grateful to Dr. James C. Schnable for his exceptional guidance and the positive impact he has made on my life and I would also like to extend my heartfelt thanks to my committee members Dr. Toshiro Obata, Dr. Jinliang Yang, and Dr. Souparno Ghosh for their invaluable insights throughout my journey.

GRANT INFORMATION

This research is supported by ARPA-E grant (ID: AR0001367), NSF and USDA NIFA under AI Institute: for Resilient Agriculture, Award No.2021-67021-35329 and U.S. Department of Energy, Grant no: DE-SC0023138.

Contents

List of Figures

List of Tables:

Introduction

Maize and Diversity panel

Maize (Zea mays L.) was domesticated from its wild relative, the lowland grass teosinte, around 9,000 years ago in southwestern Mexico [109, 90]. Today, maize is the most widely cultivated grain globally, with the United States Department of Agriculture (USDA) estimating production of about 1.23 billion metric tons for the 2023/2024 period [152]. Maize's ability to thrive under diverse environmental conditions suggests that its metabolism, particularly the varied metabolites it produces, may contribute to its growth, development, and adaptability across different climates [38, 130]. Despite the reduction in genetic diversity through domestication, maize retains substantial metabolic variability among its genotypes. This metabolic variation is crucial for studying plant metabolism and its influence on various plant phenotypes, as it provides insights into how different genotypes adapt to environmental stresses and affect plant traits [154, 38, 130, 132].

There are various methods to link genes to metabolic variation, including mutant studies, expressing genes in heterologous systems, and characterizing their effects. However, a powerful approach is to use association panels to find an association between genes and traits of interest [137]. Diversity panels play a critical role in crop research by offering a unique perspective on genetic variability and its implications for plant breeding and conservation. These panels, consisting of a wide array of genotypes, help identify genetic variants linked to important plant traits, facilitating the discovery of novel alleles and gene variants that contribute to traits such as disease resistance, yield potential, and environmental adaptability [53]. In crops like rice, maize, wheat, and soybean, diversity panels have been instrumental in unraveling the genetic basis of phenotypic variation, contributing to the development of improved cultivars for different climatic conditions and agronomic needs [92]. Additionally, diversity panels help understand evolutionary processes and guide conservation strategies for wild relatives and landraces, preserving genetic resources crucial for future breeding efforts [3].

There are multiple maize diversity panels, including the MAP/282 and SAM panels. However, the Wisconsin Diversity (WiDiv) panel, with 752 lines, is one of the largest and most widely used panels in North America. This panel includes a diverse collection of public, expired plant variety protection (exPVP), and germplasm enhancement of maize (GEM)-derived inbreds representing major North American field corn heterotic groups, such as stiff stalk, non-stiff stalk, and Iodent. It also contains unselected inbreds from synthetic populations and landraces, including 54 inbreds from the Iowa stiff stalk cycle 0 (BSSSC0) synthetic population [91]. The panel's whole-genome resequencing data is available and provides a comprehensive genetic resource for researchers [48]. Furthermore, various traits have been scored across the entire panel or subsets of it, providing valuable phenotypic data [100]. In addition, RNA-seq data has been profiled for the entire panel using tissue collected at the same time and from the same experiment as this study [146, 147]. The extensive genetic diversity of the WiDiv panel makes it an essential tool for researchers and breeders working on developing improved maize varieties. It offers an extensive range of genetic diversity and is an invaluable resource for researchers and breeders working towards the development of improved maize varieties [91].

The Significance of Plant Metabolism

Plants can produce a wide array of metabolites with diverse structures that perform essential roles in growth, cellular regeneration, resource allocation, development, and responses to biotic and abiotic stresses. Additionally, these metabolites are valuable resources for human nutrition, bioenergy, medicine, flavorings, and more [27]. While the total number of metabolites across all plant species can range from a few thousand to one million, each plant species typically synthesizes a specific subset, ranging from about 5,000 to several tens of thousands [115]. Understanding plant metabolism is crucial for sustainable agriculture and resource conservation, especially under changing climate conditions [97].

Metabolites can be classified into two main types: primary metabolites and specialized metabolites. Primary metabolites include proteins, amino acids, nucleotides, lipids, and carbohydrates, which are closely related to crop performance due to their essential roles in growth, development, and life cycle. When interacting with the environment, plants rely heavily on specialized metabolites. These compounds help plant growth and adaptation to changing environments and provide valuable resources for human health [38]. Many drugs are derived from plants, with plant metabolites playing a pivotal role. In the last two decades, many new medicines, such as alkaloids, terpenoids, flavonoids, and glycosides have been based on these metabolites. These compounds, essential for plant defense and other functions, have proven crucial for developing therapeutic agents for various diseases [106].

Genetic Bases of Plant Metabolic Diversity

Plant metabolic diversity is driven by the complex interactions of numerous genes that regulate the biosynthesis of a wide range of metabolites. The diversity in metabolic pathways enables plants to respond to biotic and abiotic stresses. While the metabolic diversity between different plant species has been widely studied, the diversity within species is now being explored more extensively. Understanding the genetic basis of this intra-species diversity is crucial for improving plant resilience and productivity. Recent advances in genomics and metabolomics have enabled the identification and characterization of key genes involved in plant metabolism, providing new insights into this complex process [39]. These developments have uncovered the genetic foundations contributing to variations in metabolite production among different plant species and varieties. Understanding these genetic mechanisms not only enhances our knowledge of plant metabolism but also opens up opportunities for developing new plant varieties with improved traits through selective breeding and biotechnological approaches [44,19].

Studies have helped uncover the crucial genes associated with metabolite variation in different crop species, especially in maize [36]. Through comprehensive genetic and molecular analyses, including GWAS and expression QTL analysis, researchers have uncovered the roles and mechanisms of two key genes, Bx12 and ZmGLK44, in regulating metabolite biosynthesis and improving drought tolerance in maize. Bx12 increases the production of protective metabolites, particularly within the benzoxazinoid pathway, during drought conditions. In contrast, ZmGLK44 coordinates the expression of multiple genes involved in the flavonoid and phenylpropanoid pathways, which respond to drought by influencing metabolic and physiological reactions [172]. Research on maize revealed 1,459 significant genetic loci associated with metabolic traits across 983 metabolite features in 702 varieties, underscoring the metabolic genetic diversity influencing maize kernel metabolism [162]. A study emphasized the integration of genetic and metabolomic approaches in maize, where more than half of the identified loci were validated by expression QTLs and linkage mapping, leading to the discovery and confirmation of five crucial genes associated with metabolic traits [162]. Another piece of investigation in maize related research, which involved 289 inbred lines, identified significant genetic variants associations with 26 metabolites. Notably, it established links between lignin precursors and a key gene on chromosome 9, connecting these metabolites to essential agronomic

traits [117]. A GWAS study combining metabolic, and phenotypes analyzed 2,980 metabolic features in 299 cassava accessions, identifying 18,218 significant marker-metabolite associations and 12 candidate genes. Functional validation confirmed the roles of the genes such as Me3GT, MeMYB4, and UGT85K4/UGT85K5 in flavone, anthocyanin, and cyanogenic glucoside metabolism [31]. A study identified a tomato gene cluster on chromosome 7 that is responsible for medium chain acylsugar biosynthesis. This cluster includes acyl-CoA synthetase and enoyl-CoA hydratase genes and co-localizes with a steroidal alkaloid gene cluster. The formation of this gene cluster correlates with medium chain acylsugar accumulation, illustrating gene cluster evolution and metabolite diversity in Solanaceae [35]. A study employing GWAS linked 123 SNPs to drought-related metabolic traits in 318 maize lines, uncovering 23 loci associated with metabolite variation under drought conditions [174]. Comparative GWAS between rice and maize have identified 420 and 292 loci associated with 123 metabolites in rice and maize, respectively. This study revealed 42 homologous loci that influence the abundance of approximately 19% of the detected metabolites across both species [23]. GWAS across a diverse range of plant species demonstrated that plant metabolism is moderately heritable, consistent with the polygenic nature of these traits [22]. These studies explore the genetic basis of metabolic variations in plants, identifying key genes responsible for the regulation of metabolite biosynthesis and stress responses. They highlight the complex interplay between genetic factors and metabolic diversity, offering insights into the mechanisms underlying metabolite variation and providing potential targets for crop improvement.

Linking Metabolic Variation to Plant Phenotypes

Investigating the genetic and biochemical underpinnings of plant metabolism provides essential insights into the development and variation of phenotypic traits. These studies help us understand how specific metabolic pathways and genetic variations contribute to observable characteristics in plants [36, 93]. For instance, research on Arabidopsis and maize has shown considerable overlap between the levels of primary metabolites and lignin precursors with biomass production [81, 18, 117]. Research on potatoes has identified that loci associated with metabolite production share locations with those involved in starch content and the process of cold sweetening [18]. A study on tomato introgression lines (ILs) explains the genetic and environmental determinants of seed metabolic traits, revealing 30 metabolite quantitative trait loci (mQTLs) and highlighting the central role of amino acids in the seed metabolic network's topology [149]. These studies explored how genetic variants influence both metabolites and other plant phenotypes.

Genome Wide Association Study (GWAS)

Advances in genomic technologies, novel methodological developments, and a keen interest in exploring trait variations across diverse genetic backgrounds have driven the early surge in association mapping studies in various crop species [179]. The creation of reference genomes and the rise of high-density genotyping have greatly enhanced the accessibility and richness of genomic data [95]. Genome-wide association Studies (GWAS) are a powerful method used to identify genetic variants associated with specific traits. In a GWAS, researchers test hundreds of thousands to millions of genetic variants across the genomes of many individuals to find associations between specific genotypes and phenotypes [140]. This method helps pinpoint genetic loci that contribute to important traits. GWAS have been extensively used to investigate agriculturally essential traits in many major crop species, including maize (Zea mays L.) [144].

The theoretical underpinnings of GWAS, initially discussed in the mid-1990s, were focused on understanding the genetics of heritable human diseases [69, 118]. The practical application of these theories became feasible following the release of the preliminary version of the human genome and the establishment of SNP datasets [70]. The first GWAS publication in 2002 identified genetic markers associated with the risk of myocardial infarction, examining around 65,000 genome-wide SNPs in 94

individuals [105]. Subsequently, GWAS has facilitated the identification of thousands of genetic variants linked to various diseases, leading the way for the development of diverse treatments [155]. In 2005, GWAS was first applied outside human medical genetics, with a study on Arabidopsis thaliana confirming genes associated with flowering time and disease resistance [5]. The initial GWAS in a crop species was published in 2008, identifying a variant of the fad2 gene related to increased oleic acid in maize [8].

GWAS involves conducting an analysis of variance (ANOVA) on each SNP in the genome to determine if there is a difference in phenotypic mean between AA, Aa, and aa genotypes [15]. However, this approach faces several challenges. First, the sheer number of SNPs tested simultaneously leads to multiple testing problems, increasing the likelihood of false positives. Second, population structure and kinship can confound results, as individuals with similar genetic backgrounds might show associations that are not due to the SNPs being tested but rather their relatedness. To address the multiple testing problem, researchers use statistical corrections such as the false discovery rate (FDR) and the Bonferroni correction to adjust the significance threshold, thereby reducing the number of false positives [9, 14]. To correct population structure and kinship, methods such as principal component analysis (PCA) and mixed linear models (MLMs) are employed. These methods account for the genetic relatedness among individuals, ensuring that the associations detected are due to the SNPs themselves and not confounding factors.

The structured association method, also known as STRUCTURE, uses null markers to identify subpopulations within genetic data. These subpopulations are then used as covariates within the association model to control for population structure [113]. The general linear model (GLM) further controls population structure by adding subpopulation covariates [113]. Additionally, the mixed linear model (MLM) incorporates both population structure (Q) and kinship (K) measures into the association model [171]. Population structure can be defined using STRUCTURE or by principal component analysis [113, 112]. Kinship, an estimate of relatedness among individuals in the population, can be calculated using various algorithms that use allele frequencies and identity-by-state to estimate identity-by-descent and define kinship coefficients [134]. By incorporating a fixed effect of population structure and a random effect of kinship under the MLM framework, false positives are controlled more effectively [171]. This enhanced control has led MLM to replace previous methods.

Various methods have been developed to address relatedness issues in the population for solving mixed linear model (MLM) equations, including efficient mixedmodel association (EMMA) [58], factored spectrally transformed linear mixed models (FaST-LMM) [80], and genome-wide efficient mixed model analysis (GEMMA) [175]. To improve GWAS power, methods such as compressed MLM (CMLM) [76], enriched CMLM (ECMLM) [76], FaST-LMM-Select [80], and settlement of MLM under progressively exclusive relationship (SUPER) [175] have been developed. Multi-locus GWAS methods like multi-locus mixed model (MLMM) [126] and fixed and random model circulating probability unification (FarmCPU) [82] simultaneously use multiple markers as covariates, particularly benefiting complex traits with closely linked large-effect loci. FarmCPU is popular among researchers due to its combination of power and efficiency.

However, GWAS is often time-intensive and may not always meet expectations. It has several notable limitations. Firstly, GWAS may fail to identify genes critical to a phenotype if the functional variation of these genes is absent or present at low frequencies in the population being studied, thereby reducing the statistical power needed to detect these variants [147]. Secondly, while GWAS identifies relatively small genomic regions compared to QTL mapping, these regions often still include multiple candidate genes. This makes it challenging to pinpoint the specific gene linked to the observed GWAS signal without conducting further extensive and resource-demanding experiments. We aim to identify the specific gene because understanding the exact genetic basis of a trait can lead to more precise breeding and biotechnological

interventions. Despite these challenges, the reusable nature of genetic marker data across multiple traits provides a cost-effective strategy, leveraging initial investments over several projects and enhancing the logistical feasibility of GWAS [147].

Transcriptome Wide Association Study (TWAS)

Transcriptome-Wide Association Studies (TWAS) is an influential method used to link gene expression levels with variations in plant phenotypes. By focusing on the impact of genetic variants on gene expression, TWAS offers a deeper exploration into the genetic basis of complex traits [50]. It provides a detailed understanding of how changes in gene expression are connected to phenotypic variations [147]. TWAS uncovers sets of genes that complement rather than duplicate those found by GWAS focusing on the same characteristics in identical populations. The expression of a single gene can aggregate signals from various upstream regulatory variants, each too minor or infrequent to be directly associated with changes in the relevant phenotype [75]. When relying on direct gene expression measurements, TWAS usually pinpoints precise candidate genes instead of regions containing several genes. This remains accurate even for species or populations characterized by a slight decrease in linkage disequilibrium across their genomes [75]. Numerous barriers have restricted the extensive use of TWAS studies in plants, such as the diurnal variation that impacts a significant number of transcripts across various species, including Arabidopsis [94], rice, poplar [40], maize, sorghum, and foxtail millet [68]. Quick freezing of samples is necessary to prevent changes in gene expression, making it challenging to collect large enough samples quickly. Additionally, the high cost of RNA sequencing versus cheaper DNA genotyping limits TWAS's use [147]. While 3' tail RNA-sequencing is a cost-effective, reliable alternative, it falls short in detecting differentially expressed genes and lacks information for detailed gene analysis [83]. Normally, GWAS offers a logistical advantage by allowing the reuse of genetic marker data for mapping various traits, making the initial high cost more manageable across numerous studies. This contrasts with gene expression data, which varies across tissues, stages, and environmental conditions, limiting its reuse for identifying genes related to different traits under varying conditions. Despite this, studies have shown that even gene expression data from non-target tissues and environments can identify relevant causal genes [74, 75]. However, reusing transcriptome-wide expression data is less common than reusing genetic marker data in research [147].

TWAS has proven instrumental in identifying candidate genes associated with trait variation in various plants, as evidenced by various studies on different crops, though it faces some disadvantages. For instance, in research focusing on rice, TWAS revealed the genetic variants that influence panicle traits, identifying essential genes such as OsSh1 and OsMADS1, which are crucial in understanding rice panicle architecture [98]. Another rice related combined GWAS and TWAS to explore diverse root phenotypes and uncover genes related to root architecture [161]. A comprehensive multi-omics approach study, including metabolomic and transcriptomic analyses on maize, identified 13 key genes affecting tocochromanol (vitamin E) levels. This includes five novel genes linked to its biosynthesis and transport [164]. Research on the heterophylly of paper mulberry through GWAS and transcriptome analysis pinpointed the candidate genes Bp07g0981 (WOX), and Bp07g0920 (HHO), which are associated with leaf shape [52]. In Brassica napus, critical insights from GWAS and TWAS revealed four gene modules significantly associated with seed oil content, highlighting the BnPMT6s genes as critical negative regulators, and the integration of multi-omics data has identified key metabolic pathways and regulatory networks involved in oil accumulation [141]. The TWAS approach in self-pollinating soybean populations notably identified pod color L2 as a novel gene influencing trait variation, alongside the detection of diverse causal variations and alternative splicing for flowering time [75]. A study on maize using RNA-Seq data showed that selective gene expression could predict flowering time more accurately than whole-genome SNPs. Notably, this study highlighted MADS-transcription factors 69 and 67 (Mads69 and Mads67), both previously linked to

flowering time, as key among the most informative genes for this prediction [146]. In an RNA-Sequencing study of 693 maize genotypes, TWAS outperformed GWAS by identifying tenfold more genes associated with flowering time, including both known and previously unidentified genes, through mature leaf tissue analysis and eQTL interactions [147]. A sorghum study integrating GWAS and TWAS analyses across 869 accessions identified 394 unique genes related to water use efficiency traits [37]. In a maize study, combining gene expression and metabolic data with genetic markers enhanced genomic prediction, with the effectiveness varying by trait and data volume [49]. These studies highlight the effectiveness of TWAS in identifying novel genes associated with various plant traits. By integrating gene expression data, TWAS provides a more precise understanding of the genetic basis of complex traits, complementing GWAS findings. The ability of TWAS to pinpoint candidate genes offers significant potential for advancing plant breeding and genetic research.

Random Forest (RF)

The random forest (RF) algorithm is a machine-learning tool that creates multiple decision trees to produce accurate and robust predictions. First introduced by Leo Breiman in 2001 [13], RF has transformed predictive modeling by employing an ensemble learning technique. During the training phase, RF generates a 'forest' of decision trees, with each tree created based on a random subset of the data. The final output of the RF model is determined by combining the predictions from all the trees, which inherently reduces the risk of overfitting and improves the model's transferability to new, unseen data [129].

RF is popular for its ease of use, minimal input requirements, and ability to handle various dependent variables, including binary, categorical, count, and continuous data. It is computationally efficient as it uses only a fraction of independent variables and performs automatic variable selection, offering a ranking of feature importance. This feature importance scoring is particularly advantageous as it helps identify the most influential variables in a dataset, providing insights into the underlying data structure and facilitating model interpretation. One of the key strengths of RF is its ability to provide feature importance scores. These scores are derived from the aggregated information gain or decrease in impurity across all trees in the forest, giving a clear indication of which variables have the most impact on the model's predictions. This process is not easily achievable with most other machine learning and deep learning approaches, which often act as black boxes and do not readily offer interpretable importance metrics [13, 129, 99].

By ranking features based on their importance, RF allows researchers to focus on the most significant predictors, potentially reducing the dimensionality of the dataset and improving the performance of subsequent analyses. It also helps in understanding the contribution of each variable to the model's predictions, which can be crucial for domains where interpretability is key, such as genomics and medicine. Additionally, feature importance scores can be used to detect and remove irrelevant or redundant features, leading to simpler and more interpretable models. This capability enhances the robustness and generalizability of the models, especially in scenarios with high dimensional data and a limited number of samples. The ability to extract feature importance from RF models sets it apart from other ML/DL approaches, making it a valuable tool for both predictive accuracy and interpretability in various research fields. Although some algorithms may outperform RF, the differences are typically marginal, and RF's relatively simple and fast implementation makes it a preferred choice. Furthermore, RF models are robust, with few hyperparameters that need tuning, making them less prone to overfitting than more complex models like deep neural networks. This is particularly true for biological datasets, where sample sizes are often much smaller, and we don't have N=millions [99]. The ease of implementation, with many available free and open-source options, and the potential for parallelization due to the independent nature of decision tree growth further contribute to RF's widespread application in various fields [99]. This is especially true for its efficacy in managing complex genomic data [24]. Unlike other statistical approaches such as GWAS

and TWAS, which typically focus on linear associations, RF is superior in capturing both linear and nonlinear interactions frequently observed in biological datasets [78, 129].

RF's adaptability has been demonstrated in various plant science applications, including trait and genomic prediction, where its ability to manage high-dimensional data effectively [24, 122, 135, 156, 73]. For example, Gonza´lez-Recio and Forni (2011) found that RF performed better than Bayesian regression methods in detecting resistant and susceptible animals based on genetic markers. They reported that RF produced the most consistent results with excellent predictive ability, highlighting its robustness and accuracy in the context of genomic selection (GS) [45]. The study by Shah et al. (2019) showcases the use of RF for predicting wheat leaf chlorophyll content through hyperspectral sensors, underscoring RF's effectiveness in handling spectral data [129]. Gill et al. (2022) found that certain machine-learning models, including Random Forest and XGBoost, outperformed specific deep-learning architectures in predicting soybean traits. These machine learning models demonstrated strong abilities in feature selection and model interpretability, pinpointing essential SNPs that matched previously identified loci in GWAS studies [43]. Toubiana et al. (2019) demonstrated the utility of RF in identifying metabolic pathways in tomatoes, significantly predicting pathways like β alanine degradation and tryptophan degradation via indole three pyruvate. Their approach combined correlation-based network analysis with machine learning, validating the presence of these pathways through in vivo assays [148]. Integrating RF with gene expression analysis effectively pinpointed the genes associated with flower color in Platycodon grandiflorus [170]. This approach demonstrates the utility of RF in analyzing gene expression data to identify relevant genes. Additionally, a study found that tree-based machine learning methods, such as RF and Gradient Boosting, significantly outperformed deep learning algorithms in predicting wheat grain yield [131]. Gradient Boosting, similar to RF, builds an ensemble of trees but does so sequentially, where each tree attempts to correct the errors of the previous one, enhancing model accuracy and robustness [41]. In another study, RF was compared with artificial neural networks (ANN) to predict yield and quality in winter rapeseed. The findings revealed that RF outperformed ANN in predicting key output variables such as seed yield and oil content [116]. A study demonstrated that integrating RF models with crop growth models (CGMs), using climate variables and the normalized difference vegetation index (NDVI), significantly enhances yield prediction accuracy for winter wheat and oilseed rape in Bavaria compared to traditional CGMs alone [30]. RF exhibited superior performance to multiple linear regressions in accurately forecasting global and regional yields of wheat, maize, and potato, highlighting RF's robustness in managing complex ecological datasets for agricultural yield prediction at both macro and micro levels [57]. A study found that RF performs better than other modeling techniques for binary traits when the sample size is large, and the percentage of missing data is low [42]. Another study found that, for binary traits, RF outperformed the genomic BLUP (GBLUP) method only in a scenario combining the highest heritability ($h2 = 0.30$), the most extensive dense marker panel (50K SNP chip), and the most significant number of QTL (725) [101]. These studies demonstrated the utility of incorporating machine learning approaches in genomic studies, particularly in identifying genes associated with trait variation. RF's ability to capture both linear and non-linear associations between genes and traits enhances our understanding of complex biological interactions relative to purely linear models.

Numerous efforts have been made to identify genes that control variation in metabolism. However, capturing the complex genetic architecture underlying metabolic traits often succeeds when using multiple or different statistical approaches as these approaches lead to more robust and reliable findings. This suggests that incorporating mixed or alternative methods, such as machine learning approaches, may help identify potential genes that traditional quantitative genetics methods alone may miss [129, 79]. On the other hand, the application of gene expression data to investigate the association between genes and metabolite variation in field-grown diversity panels remains underexplored. Studies focusing on the genetic

control of metabolites highlight the need for further exploration through gene expression data to advance our understanding of the diversity of metabolites and their genetic underpinnings [53, 92, 178].

Our study aimed to fill this research gap by utilizing genomic data and field-based gene expression data to investigate the genes associated with metabolite variation through quantitative genetics approaches such as GWAS and TWAS, as well as machine learning approaches such as RF in a maize diversity panel. These approaches allowed us to capture both the linear and non-linear associations between the genes and the traits studied, making our findings more comprehensive [78, 129]. By integrating these methods, we were able to identify a total of 240 genes associated with metabolite variations and one gene that was identified between TWAS and RF. This study not only highlighted genes directly linked to metabolite production, such as N-acetyl-gamma-glutamyl-phosphate reductase with L-glutamic acid but also revealed other genes crucial for various aspects of plant metabolism, especially resistance against biotic and abiotic stresses. Additionally, we identified three loci associated with variation in both metabolic and non-metabolic traits. These results enable future studies to delve deeper into the roles of these metabolites and genes in plant growth and development processes.

Materials and Methods

Field experiments and trait scoring

The maize field experiment in the summer of 2020 at the Havelock farm of the University of Nebraska-Lincoln (40.852◦N, 96.616◦W). The experimental design and trait evaluation methodology conducted in the Lincoln, Nebraska field trial was previously described in past research [100, 138, 147, 120]. Our field experiment consisted of 750 genotypes, representing 699 unique genotypes, with 51 genotypes represented twice, a subset of the Wisconsin Diversity Panel [91]. Initially, data was collected from 750 plants. After applying quality control (QC) procedures, data from 660 genotypes were successfully retained. Among these, 47 genotypes were used twice for metabolite quantification. The field was laid out in a randomized complete block design on May 6, 2020, consisting of two replications of each genotype. A total of 1680 plots, with each block consisting of 840 plots and with repeated checks of the B97 genotype. The layout for each plot consisted of two rows, each 7.5 (about 2.3 meters) feet long, with rows spaced 30 (roughly 0.76 meters) inches apart. Plants within the rows were placed 4.5 (approximately 11.5 centimeters) inches apart from each other, and the plots were separated by 30-inch (around 0.76 meters) alleyways in between.

Metabolites/Biochemical Quantification

We collected leaf samples for metabolite profiling during the flowering stage. The methodology for collecting these samples was previously described in Torres et al. (2023) [147]. On July 8th, 2020, samples were collected from one out of two duplicate blocks, specifically Block 1, the block furthest west of the Lincoln, Nebraska field experiment. A representative plant was chosen for sampling from each plot, ideally avoiding those at the plot edges. We collected five leaf disks from the pre-antepenultimate leaf and the fourth leaf down from the top visible leaf of the chosen plant. The leaf tissue was immediately subjected to flash freezing in liquid nitrogen and subsequently stored on dry ice until it could be transferred to a freezer at -80°C. This collection process was executed efficiently by seven researchers in parallel, ensuring that all samples were gathered within two hours and completed before midday on the collection date.

We quantified metabolite abundance following a method described previously by Wase et al. (2022) [160]. Collected leaf samples were ground to a fine powder while being kept under liquid nitrogen using TissueLyser II (Qiagen). We added 730 µL of methanol premixed with 20 mg/mL ribitol in water at a 700:30 ratio in a 2 ml Eppendorf microfuge tube containing approximately 25 mg of ground tissue of each sample,

vortexed immediately, and kept it on ice immediately afterward. Sample tubes were then transferred to a thermomixer at 70°C and shaken for 15 minutes at 950 rpm. Following this step, we centrifuged the tube at 17,000 X g and transferred the supernatant to the new Eppendorf tube at room temperature. Afterward, 325 µL chloroform and 750 µL water were added and vortexed for 30 seconds. Samples were then centrifuged at 1500 X g for 15 minutes. Finally, we transferred an aliquot of 50 µL from the upper polar phase into a fresh 2 mL Eppendorf tube and dried the samples with a centrifugal vacuum concentrator. After vacuum drying, each tube was filled with argon gas and tightly closed to prevent the oxidation of metabolites.

We used Gas Chromatography-Mass Spectrometry (GC-MS) for metabolite profiling as described in Wase et al. (2022) [160]. The dried metabolite extracts were derivatized by methoxyamination in 20 mg/mL methoxyamine hydrochloride in pyridine for two hours at 37°C. The samples were further trimethylsylilated for 30 minutes at 37◦C with 70 µL N-Methyl-N-(trimethylsilyl) trifluoroacetamide (Millipore Sigma). A fatty acid methyl ester mixture was added to the trimethylsylilation solution for retention time calibration. One microliter of each sample was injected into GC-MS (7200 GC-QTOF system, Agilent) equipped with a HP5msUI (30 m length, 0.25 mm diameter, 0.25 µm thickness) column. GC and MS parameters are exactly as described in Wase et al. (2022) [160]. Chromatographic peaks were annotated to metabolites by MassHunter Unknows (Agilent) based on the retention time and mass spectrum matching with data in the Fiehn Metabolomics Library (Agilent). MassHunter Quantitation (Agilent) calculated the representative ion's peak heights. Based on careful manual curation, we quantified only the metabolite peaks, which allowed us to confidently identify unique peaks across all samples. Ultimately, we measured 26 distinct metabolites: Phosphoric acid, Glyceric acid, Serine, L-alanine, L-threonine, Betaalanine, Malic acid, Aspartic acid, Glutamic acid, Trans-Aconitic acid, Glycerol 1phosphate, Shikimic acid, Citric acid, Quinic acid, Fructose, D-glucose, Tyrosine, Galactonic acid, Mucic acid, Myo-inositol, Caffeic acid, D-glucose 6-phosphate, Sucrose, Loganin, Chlorogenic acid, and Raffinose. After subtracting the background noise, the peak heights of the representative ions for each metabolite were normalized against the internal standard ribitol and adjusted for the exact fresh weight of the samples used for extraction. The initial estimates of relative metabolite content were log-transformed.

Whole plant phenotypes

We used 28 whole plant phenotypes that were previously published [100]. The traits are days to pollen, days to silk, root lodging percentage, stalk lodging percentage, leaf angle, ears per plant, cob weight (grams), hundred kernel mass (grams), total grain mass (grams), bushel per acre equivalent, grain percent moisture, ear length (cm), ear width (cm), ear filled length (cm), kernel row number, kernels per row, percent fill, southern rust severity score, leaf area index, leaf length (cm), leaf width (cm), plant height (cm), extant leaf number, nodes with brace roots, tillers per plant, branches per tassel, branch zone length (cm), and tassel spike length (cm). Male and female flowering times were recorded when 50% of plants showed pollen shed or silks, respectively. Root and stalk lodging percentages were assessed at the season's end, alongside leaf traits (length, width, angle) for two plants per plot post-anthesis. The leaf area index was estimated using an LAI-2200C analyzer. For yield attributes, ear length, ear width, ear filled length, kernel row number, and kernels per row were measured for six ears per plot from eight semi-randomly selected plants, excluding edge plants. Grain and cob weights were determined post-harvest, with grain moisture standardized to 15.5%.

Hyperspectral Traits

We utilized the hyperspectral data that had been previously published [150]. The hyperspectral reflectance data was collected over nine days within 13 days, from July 8th to July 20th, 2020. This data was acquired from a fully expanded leaf of a representative plant in each plot, taking care to avoid edge plants wherever feasible, and using a benchtop spectroradiometer [87] equipped with a contact probe. Reflectance measurements recorded 2,151 values at one-nanometer intervals across a 350 to 2500-nanometer spectrum, taken from the adaxial side of leaves at the tip, middle, and base. These measurements were averaged over nine days, resulting in a composite spectrum for each plot. To reduce the dimensionality of the 1,658 plot-level hyperspectral reflectance values, both principal component analysis (PCA) and a trained autoencoder neural network were employed. Using the sci-kit-learn package [107], PCA reduced the dataset to 10 principal components (LV1 to LV10), summarizing 99% of the variance.

Photosynthetic Traits

We used four photosynthetic traits that were previously published [120]. The traits FvP/FmP, relative chlorophyll and leaf temperature were measured in plants using a spectrophotometer [1] using the protocol previously published [66]. The measurements were conducted during the pre-flowering growth stage, specifically from the youngest fully expanded leaves of the plants, over a six-day period from July 23rd to 28th, 2020.

Data Integrity and Quality Assessment

Our field experiment consisted of 750 genotypes, representing 699 unique genotypes, with 51 genotypes represented twice, a subset of the Wisconsin Diversity Panel [91]. Initially, data were collected from 750 plants. After applying quality control (QC) procedures, data from 660 genotypes were successfully retained. Among these, 47 genotypes were used twice for metabolite quantification. Extreme outliers were then removed through manual examination of histograms and scatter plots. Best Linear Unbiased Predictors (BLUPs) for each metabolite were estimated using a mixed linear model, generated using the lme4 package [7] implemented in R v4.2.1 [114] with the equation: $y_i = \mu + 1|\text{Genotype}_i + 1|\text{Batch}_j + 1|\text{RunOrder}_k + \text{error}_{ijk}$ where yi is the mean value for the metabolite of interest in the ith genotype ran in jth batch and kth run order during GC-MS pipeline. The variance explained by each factor included in the model was extracted. We calculated the repeatability using 47 duplicated genotypes out of a total of 795 genotypes using the following equation: where $\sigma G2$ was the total variance explained by the genotype in each mixed linear model and $\sigma e2$ is the total residual variance in each mixed linear model. Importantly, we calculated the repeatability without removing outliers before calculating GWAS, as outliers were removed for GWAS analysis. We used principal component analysis (PCA) to identify critical patterns and relationships among metabolites by using the Factominer [71] function implemented in R. We also used the Pearson correlation method to examine how metabolites are related to each other using the corrplot [125] function implemented in R.

Resampling Model Inclusion Probability Genome-Wide Association Study (RMIPGWAS)

Genome-wide association studies reported here were conducted using published resequencing based on 46 million high-confidence genetic marker data for 752 genotypes drawn from the Wisconsin Diversity panel [48]. The dataset was filtered to retain only markers with minor allele frequency >0.05 among the 660 unique genotypes included in this study using plink2 (v2.0a1) [20], resulting in a final dataset of 2,688,200 genetic markers. We performed the GWAS using the Fixed and Random model Circulating Probability Unification (FarmCPU) algorithm [82] implemented in the rMVP package [169] as FarmCPU offers enhanced detection of true-positive genetic associations [82]. To ensure the reliability of these signals, we

used resampling techniques like the resample model inclusion probability (RMIP), which quantifies signal stability across multiple bootstraps [153, 138]. Marker trait associations were considered significant if identified in at least 10 out of 100 conducted resampling-based GWAS analyses. In each iteration or GWAS analysis, a random 10% of total genotypes corresponding to 66 genotypes were masked, and a separate FarmCPU analysis was performed, keeping significant markers with a p-value of less than a threshold of 2 ×10-8. The significance threshold was set using a Bonferroni correction, calculated as 0.05 divided by the total number of SNPs used in the analysis [33].

Transcriptome-Wide Association Study (TWAS)

Our study utilizes a previously published gene expression dataset quantifying the expression of 24,585 gene models [147] to conduct Transcriptome-Wide Association Studies (TWAS) with the compressed mixed linear model as implemented in Genomic Association and Prediction Integrated Tool (GAPIT) [176].To obtain the TPM (Transcript per million) file, we took leaf tissue samples from the pre-ante-penultimate leaf (used for scoring the metabolites) of maize plants and rapidly froze them to maintain RNA integrity. RNA extraction was performed using a Kingfisher Flex automated extraction robot (ThermoFisher Scientific; 5400630) with the MagMax Plant RNA Isolation Kit (ThermoFisher Scientific; A47157). We assessed the quality and quantified the RNA using the Quant-IT Broad Range RNA Assay Kit (ThermoFisher Scientific; Q10213). We created libraries with the TruSeq kit from Illumina and sequenced them on the NovaSeq 6000 Illumina platform. We processed the raw sequence data by filtering it for quality using trimmomatic (v0.33) [11] and estimating gene expression with Kalliso (v0.46)[12]. We used the B73 RefGen V5 sequence file [124, 54] from Phytozome [46] as a reference. Finally, we filtered the data based on principal component analysis and transcript abundance by removing genes with TPM < 0.1 in at least 50% of the final samples. The first three principal components of variation calculated by GAPIT from the expression data were included as covariates. A gene was considered significantly associated with the trait of interest when the associated p-value was less than $2.03\times10-5$, with the significance threshold set using a Bonferroni [33].

Random forest (RF)

We employed the Random Forest (RF) model [13] as implemented in the R libraries randomForest [78] and caret [67] to predict the genes associated with the metabolites using gene expression of genotypes used in our study [147]. RF models with five different tree counts (100, 200, 300, 400, and 500) were employed and validated using 5-fold cross-validation. Within each tree count parameter, the model's performance was evaluated using the root mean square error (RMSE), and the importance of the particular gene for prediction was assessed based on the increase in mean squared error (IncMSE) when the gene was not included in the model. To identify the significant genes associated with metabolite concentrations, we compared the original dataset with control sets created by shuffling taxa order while keeping other variables constant across shuffled datasets. We used shuffled data as control sets to validate the association findings and provide a baseline to assess the significance of our observed gene-metabolite associations against what could be expected by chance. We extracted feature importance scores from the original and shuffled datasets and generated density plots to visualize the distributions. We considered different feature importance scores and selected a threshold that corresponded to achieving an FDR level of approximately 0.05 for all metabolites (B.3) (A.11), as described by Benjamini and Hochberg [9]. The threshold is calculated by dividing the number of genes identified in the shuffled dataset by the total number of shuffled datasets and comparing this ratio to the count from the original dataset. Finally, we retained the top genes with importance scores significantly higher in the original dataset than in the shuffled dataset for their crucial role in influencing metabolite variations.

Results

This study extensively explored the genetic factors influencing maize metabolism and their impact on non-metabolic traits. Using a dataset of 660 diverse maize inbred lines from the WiDiv panel [91], we employed GC-MS to quantify metabolite peaks, which allowed us to confidently identify 26 unique metabolites across all samples. Through the application of quantitative genetics and machine learning methods, we successfully identified both well-known and previously uncharacterized genes associated with variation in metabolite and non-metabolite traits in maize.

Repeatability (*r*)

We observed a wide range of repeatability values for 26 metabolites, with an average repeatability of 0.36 across all metabolites, indicating a modest genetic influence on metabolism as a whole (F.1A). Notably, galactonic acid showed the highest repeatability (r = 0.67), suggesting strong genetic control. Other metabolites such as glycerol 1-phosphate (r=0.48) and chlorogenic acid (r=0.48) also displayed moderate repeatability. Conversely, some metabolites exhibited low repeatability, emphasizing the significant influence of environmental and other non-genetic factors. For non-metabolite traits, repeatability (r) was quantified for 41 traits, including whole plant phenotypes, hyperspectral traits, and photosynthetic traits. Repeatability values for whole plant phenotypes ranged from 0.36 to 1.00 (A.2), for hyperspectral traits from 0.60 to 0.75 (A.4), and for photosynthetic traits from 0.58 to 0.83 (A.3). These findings underscore the complex interplay between genetics and the environment in determining maize phenotypes.

Variance partitioning

In our study, we employed variance partitioning analysis to distinguish the contributions of genetic and non-genetic factors to the total observed phenotypic variance. This method quantified the influence of four primary factors: genotype (genetic diversity among maize lines), batch (variations in sample processing), run order (analysis sequence effects due to instrument sensitivity shifts), and residual factors (all other unaccounted variances like minor environmental or experimental discrepancies) on the total observed variance for each metabolite (F.1b). This approach allowed us to determine how these factors influenced each analyzed metabolite. Our results revealed that metabolites like Shikimic acid, Tyrosine, Trans-Aconitic acid, Glycerol 1-phosphate, Chlorogenic acid, D-glucose, and Loganin displayed considerable genetic control. In contrast, others were more influenced by non-genetic factors such as environmental conditions and experimental procedures, such as batch and run order. It was observed that non-genetic factors contributed notably to the variance, highlighting the significance of metabolite expression in a field-grown maize diversity panel [172, 77].

Principle Component Analysis (PCA)

Principal Component Analysis (PCA) indicated a diverse distribution of metabolite profiles, with the first two principal components capturing a combined 25% of the variance (A.5).

Correlation analysis (*r*)

We found a strong positive correlation between D-glucose and fructose (r=0.62), consistent with the known enzymatic conversion of glucose-6-phosphate to fructose-6phosphate by glucose isomerase in glycolysis [102] and moderate positive correlation between Shikimic acid and Quinic acid (r = 0.51), consistent with their known common origin through the shikimate pathway [165]. Moreover, we also discovered other significant positive correlations, such as between D-malic acid and Lthreonine (r=0.55), Beta-alanine and L-glutamic acid (r=0.46), and Aspartic acid and L-glutamic acid (r=0.44). On the other hand, a negative

correlation was found between L-glutamic acid and L-alanine (r = -0.14) and between Trans-Aconitic acid and Galactonic acid (r = -0.11), indicating inverse relationships in their metabolic concentrations (A.6). We also extended our correlation analysis to see the positive and negative correlation between metabolites and non-metabolite traits. Notably, we found a negative correlation between Shikimic acid and Percent fill (r = -0.25) and positive correlations between Shikimic acid and plant height (r = 0.23), Trans-Aconitic acid and days to pollen (r = 0.23), Quinic acid and days to silk (r = 0.28), Quinic acid and plant height (r = 0.22), Beta-alanine and 100 kernel mass (r = 0.22), and citric acid and days to silk (r = 0.23). These correlations reveal complex interactions within maize metabolism and their effects on various plant characteristics (A.7).

Resampling Model Inclusion Probability Genome Wide Association Study (RMIPGWAS) for Metabolites

Our GWAS study aimed to identify genetic variants associated with 26 metabolites using approximately 2.6 million single nucleotide polymorphisms (SNPs)[48] using the FarmCPU algorithm, which offers enhanced detection of true-positive genetic associations [82]. To ensure the reliability of these signals, we used resample model inclusion probability (RMIP), which quantifies signal stability across multiple bootstraps [153, 138]. We identified a total of 155 genetic variants associated with 26 metabolites at RMIP significance thresholds of $x \geq 0.1$. Among these variants, 17 strong signals were identified at RMIP significance thresholds of $x \geq 0.3$. Among the 17 variants which exceed RMIP significance thresholds of $x \geq 0.3$ and we found 1 marker each linked to variation in Glyceric acid, Shikimic acid, L-serine, Quinic acid, Raffinose, Sucrose, Tyrosine, D-glucose, and Fructose and 2 markers each linked to variation in Phosphoric acid, Chlorogenic acid, Galactonic acid, and Trans-Aconitic acid (F.2).

We specifically highlighted seven candidate genes within 100 kb intervals centered around seven significant SNPs associated with five metabolites (A.8). These genes show promising linkage disequilibrium (table 2) and are known to play significant roles in plant metabolism, though in none of these cases were the genes identified previously known to be directly involved in the production or catabolism of the metabolite the GWAS they were adjacent to was linked to. The genes are as follows:Theobromine synthase (TS) and Benzoate O-methyltransferase (BMT) are associated with Galactonic acid; Phosphoglycolate phosphatase (PGP) and Peroxidase (POX) are associated with Chlorogenic acid; Ubiquitin Carboxyl-terminal Hydrolase (UCH) is associated with Phosphoric acid; a Protein kinase domain (PK) is associated with D-glucose; and Dimethylaniline Monooxygenase (DMAO) is associated with L-serine. Our LD analysis revealed moderate to high linkage disequilibrium (LD) for these key metabolic genes (TS, BMT, PGP, POX, UCH, PK, and DMAO), indicating that they are co-localized and inherited together. This co-localization confirms the association between these genes and the metabolites they influence.

Briefly, TS is a key enzyme in the biosynthesis of theobromine, catalyzing the methylation of 7-methylxanthine to produce theobromine in cacao plants. This enzyme plays a critical role in the pathway that converts xanthosine to theobromine, an important alkaloid for plant defense mechanisms. [34, 151, 65, 143]. On the other hand, BMT catalyzes the methylation of benzoic acid to form methyl benzoate, a key volatile compound involved in plant defense and attraction of pollinators. This enzyme plays a crucial role in the formation of floral scent and the defense mechanism against herbivores and pathogens [25, 128, 103, 163]. PGPase is a crucial enzyme in the photorespiratory pathway of photosynthetic organisms, including algae, plants, and cyanobacteria. It catalyzes the hydrolysis of phosphoglycolate (PG), a byproduct of the oxygenase activity of ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco). This enzyme is essential for the growth of photosynthetic organisms under light conditions by facilitating the recycling of carbon and preventing the accumulation of toxic intermediates. [28, 32, 88]. POX is an enzyme that catalyzes the reduction of hydrogen peroxide using various electron donors. It plays a critical role in plant

metabolism by detoxifying reactive oxygen species, contributing to lignin biosynthesis, and helping in defense responses against pathogens. It helps in maintaining cellular redox balance and facilitating the cross-linking of cell wall components, which is essential for plant growth and stress tolerance [51, 60, 4]. UCH enzymes are critical for protein regulation in plants by removing ubiquitin from target proteins, thus preventing their degradation. UCHs are involved in various cellular processes, including protein quality control, signaling, and stress responses. By maintaining protein homeostasis, UCH enzymes support plant growth and development under varying environmental conditions [110, 139, 157, 166]. PK enzymes are crucial regulators in plant metabolism. They catalyze the transfer of phosphate groups to target proteins, altering their activity, stability, and localization. PKs play essential roles in signaling pathways that control plant growth, development, and responses to environmental stresses. By modulating various metabolic and physiological processes, PKs help plants adapt to changing conditions and optimize their survival and productivity [133, 142, 168]. DMAO also known as flavin-containing monooxygenase (FMO), is an enzyme that catalyzes the oxygenation of various substrates containing nitrogen, sulfur, or phosphorus. In plant metabolism, DMAO plays a vital role in detoxifying xenobiotics and endogenous compounds, converting them into more water-soluble forms that can be easily excreted. This enzymatic activity helps plants manage oxidative stress and maintain cellular homeostasis, contributing to overall plant health and resilience [64]. These findings highlight the essential roles of highlighted genes in plant metabolism. These genes may influence metabolite levels indirectly through their broader metabolic functions.

Transcriptome Wide Association Study (TWAS)

Our TWAS study focused on identifying genes that are associated with variations in 26 metabolites using a previously published gene expression dataset from leaf tissue [147], which featured 24,585 gene models. As a result of our analysis, we identified six genes linked to three specific metabolites (F.3). One gene was linked to L-glutamic acid, one to Quinic acid, and four to Glycerol 1-phosphate, with no significant associations found for other metabolites. Notably, we highlighted two genes with well-established roles in plant metabolism, despite not being directly involved in the production of the metabolites they are associated with. The first gene is a Multi-copper oxidase (MCO), which is linked to Quinic acid. The second gene is a Cu(2+)-exporting ATPase (CUEA), associated with both Glycerol 1-phosphate and L-glutamic acid. Interestingly, the CUEA gene was identified in both our TWAS and random forest analyses, highlighting its potential significance in the regulation of Glycerol 1-phosphate. Briefly, CUEA is an enzyme that plays a crucial role in the homeostasis of copper ions within plant cells. This ATPase actively transports excess Cu(2+) ions out of the cytoplasm, either into the vacuole for storage or out of the cell, thereby preventing toxic levels of copper from accumulating [29, 72, 96, 62]. MCOs are enzymes that facilitate the oxidation of a wide range of substrates, including polyphenols, aromatic polyamines, L-ascorbate, and metal ions. These enzymes function by transferring electrons from the substrates to a copper cluster within the enzyme. This copper cluster subsequently uses the electrons to reduce oxygen molecules to water, completing the catalytic cycle [121]. This enzymatic activity is essential for processes such as lignin degradation, iron homeostasis, and defense mechanisms against pathogens [47, 59, 10]. These findings underscore the integral roles of copper-related genes, such as CUEA and MCOs, in plant metabolism. Their involvement in critical processes like ion homeostasis, lignin degradation, and defense mechanisms suggests that they may influence metabolite levels indirectly through their broader metabolic functions.

Random forest (RF)

In our study, we applied the RF feature importance method to identify genes associated with metabolite production using gene expression data. Unlike GWAS and TWAS, which typically focus on linear relationships, RF might capture both linear and nonlinear relationships in biological data [129]. In our RF study, we focused on three metabolites: L-glutamic acid, Glycerol 1-phosphate, and Quinic acid. These metabolites were selected based on their significant associations found in both RMIPGWAS and TWAS. Through this approach, we identified 79 genes significantly associated with these metabolites at an approximate false discovery rate (FDR) of 5% [9] (F.4). Notably, this includes one overlapping gene, CUEA, identified between TWAS and RF. Each metabolite association had an FDR ranging from 0.03 to 0.06, suggesting confidence in our results (S.3). Specifically, the number of genes (n) identified in each metabolite was as follows: L-glutamic acid (n=26 genes), Quinic acid (n=29), and Glycerol 1-phosphate (n=24). Among the genes identified, a notable one is an N-acetyl-gamma-glutamyl-phosphate reductase (argC), which is associated with L-glutamic acid. ArgC is a key enzyme in the arginine biosynthesis pathway, catalyzing the reduction of N-acetylglutamate 5-phosphate to N-acetylglutamate 5semialdehyde. L-glutamic acid serves as a precursor in this pathway, providing the initial substrate. This highlights the essential role of argC in glutamic acid metabolism [6, 86, 21]. This finding underscores the effectiveness of RF in identifying genes directly linked to metabolite production. It also demonstrates the importance of incorporating methods like RF, which can capture both linear and nonlinear associations, alongside TWAS, which is limited to linear relationships. Using RF analysis, we may gain a more comprehensive understanding of the complex interactions between gene expression and metabolite profiles, revealing insights that may not be uncovered through TWAS.

Resampling Model Inclusion Probability Genome-Wide Association Study (RMIPGWAS) for Non-Metabolites

We carried out a GWAS employing the RMIP method and the FARMCPU algorithm to analyze around 2.6 million SNPs [48] to identify genetic variants linked to 42 non-metabolite traits. These traits include 28 different plant phenotypes, 10 hyperspectral traits, and 4 photosynthetic traits. Our analysis identified 223 genetic markers associated with these non-metabolite traits. Out of these markers, 135 are linked to whole plant phenotypes, 55 to hyperspectral traits, and 33 to photosynthetic traits (A.12). We identified these markers at RMIP significance thresholds of $x \geq 0.1$, and we also found the closest candidate genes within a 100-kb interval centered around each significant SNP associated with these markers. We compared genetic variants linked to both metabolite and non-metabolite traits identified through RMIPGWAS and discovered three overlapping genetic markers between these traits within a 100-kb interval. These overlapping genes were found for the following metabolite and nonmetabolite trait pairs: L-serine vs. percent fill (distance between the markers: 31,392 bp), Mucic acid vs. LV9 (distance between the markers: 64,110 bp), and Chlorogenic acid vs. branches per tassel (distance between the markers: 92,950 bp) (Figure 6,7,8). Additionally, we identified three candidate genes associated with these genetic markers at 100 kb intervals that exhibit promising LD and have established roles in plant metabolism though there was no established relationship between the overlapping genes and the traits they associated with. These genes are Aldehyde dehydrogenase (ALDH) found associated with both L-serine and Percent fill, Myb/SANT-like DNA-binding domain (MYB) found associated with both Mucic acid and LV9, cyclin-dependent kinase (CDKs) found associated with both Chlorogenic acid and Branches per tassel.

Briefly, ALDH is an enzyme family essential in plant metabolism, primarily involved in detoxifying reactive aldehydes produced during various metabolic processes. ALDH converts these toxic aldehydes into less harmful carboxylic acids, thus maintaining cellular homeostasis and protecting plant cells from

oxidative stress. This detoxification is crucial for plant development, growth, and stress response mechanisms. By detoxifying harmful aldehydes, ALDH might support efficient metabolic processes such as L-serine synthesis and overall plant health [136, 177, 145, 56, 173, 158]. The MYB gene family plays a significant role in plant metabolism by regulating gene expression involved in various biological processes. These include plant growth, development, stress response, and secondary metabolism, which involves the biosynthesis of compounds such as flavonoids, phenolic acids, and monolignols. These compounds contribute to cell wall structure, UV protection, and defense mechanisms. Based on their overall role in plant metabolism, MYB genes might indirectly be associated with the traits they found associated with [16, 167, 2, 111]. CDKs are crucial enzymes that regulate the cell cycle in plants by phosphorylating target proteins, thereby controlling cell division and growth. They play a vital role in various metabolic processes, including DNA replication, repair, and transcription. CDKs also play crucial roles in stress responses by controlling cell cycle checkpoints and modulating stress-responsive genes, enabling plants to adapt to environmental stresses such as drought, salinity, and pathogen attack. Although CDKs have a well-established role in plant metabolism, there is no direct relationship between the traits (Chlorogenic acid and branches per tassel) they are associated with [108, 63, 104, 84, 26].

Discussion

Maize produces various metabolites during its growth cycle, which not only enable growth and development but also enhance its adaptability and resistance against biotic and abiotic stresses [38, 130]. Even after domestication, maize still possesses significant genetic diversity, making it an ideal model for studying plant metabolism and its impact on different phenotypes [154, 38, 130, 132]. While GWAS has been extensively used to investigate plant metabolism, the potential of leveraging gene expression data from field-grown diversity panels to explore associations between genes and metabolites remains largely unexplored [53, 92, 178]. These diversity panels capture a wide range of variations within the population, making them ideal for studying the genetic control of specialized metabolites. Research on the genetic control of these metabolites underscores the need for further exploration through gene expression data to enhance our understanding of metabolite diversity and its genetic associations [53, 92, 178, 127]. Additionally, by employing both quantitative genetics and machine learning methods, researchers may uncover complex interactions, including linear and non-linear relationships, between genes and metabolites [129, 79]. Furthermore, the link between genes associated with metabolite variation and other plant traits remains largely unexplored, presenting a promising area for future research [19, 55, 31]. Our study aimed to address these gaps by using 2.6 million SNP data and 24 thousand field-based gene expression data to explore the genetic determinants of metabolite and non-metabolite variation through quantitative genetic approaches such as GWAS, TWAS, and machine learning approaches such as RF. Out of 155 genetic variants identified by RMIPGWAS at RMIP significance thresholds of $x \geq 0.1$, we focused on 17 strong signals identified at RMIP significance thresholds of $x \geq 0.3$ in our study. A parallel TWAS identified 6 candidate genes. A random forest feature importance-based approach was conducted for the specific metabolites where at least one hit was observed in TWAS identified one overlapping gene, Cu(2+)-exporting ATPase, and a number of other genes not identified by TWAS. Three of the loci found associated with variation in metabolite traits were also linked to non-metabolic traits in an RMIPGWAS involving a total of 42 non-metabolic traits.

This study initially aimed to characterize the properties of metabolites using Principal Component Analysis (PCA) and correlation analysis (A.5, A.6). Next, we used repeatability and variance partitioning to further explore the connections and the genetic and environmental factors influencing metabolite variation in maize (F.1A, F.1B). We observed significant variability in metabolite profiles across different genotypes through

PCA. This variability may be influenced by genetic and environmental factors, indicating a potential interplay that requires further investigation.

We observed a wide range of repeatability values for 26 metabolites, with an average repeatability of 0.36 across all metabolites, indicating a modest genetic influence on metabolism as a whole. Notably, galactonic acid showed the highest repeatability (r = 0.67), suggesting strong genetic control. Other metabolites such as glycerol 1-phosphate (r = 0.48) and chlorogenic acid (r = 0.48) also displayed moderate repeatability. Conversely, some metabolites exhibited low repeatability, emphasizing the significant influence of environmental and other non-genetic factors. For non-metabolite traits, repeatability (r) was quantified for 41 traits, including whole plant phenotypes, hyperspectral traits, and photosynthetic traits (A.2,3,4). Repeatability values for whole plant phenotypes ranged from 0.36 to 1.00, for hyperspectral traits from 0.60 to 0.75, and for photosynthetic traits from 0.58 to 0.83. These findings underscore the complex interplay between genetics and the environment in determining maize phenotypes, with some traits being more strongly influenced by genetic factors while others are more susceptible to environmental variations and complex metabolic pathways involving multiple genes.

Through variance partitioning analysis, it was discovered that non-genetic factors (e.g., batch effects, run order, and residuals) significantly contribute to the total phenotypic variance. These findings highlight the importance of considering genetic and non-genetic factors when studying plant metabolism, as both factors play critical roles in shaping phenotypic outcomes [22, 172, 77]. Our correlation analysis identified several significant associations among metabolites, both positive and negative. We found a significant positive correlation between D-glucose and fructose (r=0.62), which is consistent with the known enzymatic conversion of glucose-6-phosphate to fructose-6-phosphate by glucose isomerase in glycolysis. This process is crucial for energy production and highlights the tight metabolic connection between these sugars [102]. A moderate positive correlation was also observed between shikimic acid and quinic acid (r = 0.51), suggesting a biosynthetic relationship through the shikimate pathway [159]. Our analysis also revealed other notable positive and negative correlations between different metabolites that have not yet been explored. This indicates potential areas for further investigation, which could uncover new insights into plant metabolism and biochemical interactions. Our correlation analysis also revealed several associations between the metabolites and non-metabolite traits. Notably, correlations between Shikimic acid and plant height (r=0.23) and between Quinic acid and days to silk (r=0.28) suggest underlying metabolic influences on phenotypic traits. However, there is a lack of direct evidence in the current literature that establishes the connection between these metabolites and non-metabolite traits. This gap in knowledge highlights the need for more detailed research to investigate the mechanisms underlying these associations in order to gain a better understanding of the metabolic contributions to plant biochemistry and development.

Among the 17 identified genes from RMIPGWAS at a significant threshold (RMIP x $\geq$ 0.3), we highlighted seven genes (TZ, BMT, PGPase, POX, UCH, PK, and DMAO) that play a significant role in plant metabolism as supported by literature and previous research (F.2). TS is a key enzyme in theobromine biosynthesis, catalyzing the methylation of 7-methylxanthine to produce theobromine in cacao plants, an essential alkaloid for plant defense mechanisms [34, 65, 151, 143]. BMT catalyzes the methylation of benzoic acid to form methyl benzoate, a critical volatile compound for plant defense and pollinator attraction [25, 128, 103, 163]. PGPase is a crucial enzyme in the photorespiratory pathway of photosynthetic organisms which catalyzes the hydrolysis of phosphoglycolate (PG), a byproduct of the oxygenase activity of Rubisco. This enzyme is essential for the growth of photosynthetic organisms under light conditions by facilitating the recycling of carbon and preventing the accumulation of toxic intermediates [28, 32, 88]. POX reduces hydrogen peroxide using various electron donors, detoxifying reactive oxygen species and contributing to lignin biosynthesis and pathogen defense responses [51, 60, 4]. UCH enzymes regulate

protein degradation by removing ubiquitin from target proteins, maintaining protein homeostasis, and supporting plant growth, development, and stress responses [110, 139, 157, 166].PK enzymes regulate plant metabolism by transferring phosphate groups to target proteins, influencing their activity, stability, and localization, and playing essential roles in signaling pathways that control plant growth and responses to environmental stresses [133, 142, 168].DMAO is a flavin-containing monooxygenase which involve in oxygenates nitrogen, sulfur, and phosphorus-containing compounds. It detoxifies xenobiotics and endogenous substances and helps in oxidative stress management [64]. Overall, our GWAS identified strong signals associated with specific metabolites, highlighting the critical roles these genes play in responding to biotic and abiotic stresses. While there is no direct evidence linking these genes to the associated metabolites, it is known that they play crucial roles in plant metabolism and adaptability. These findings emphasize the essential roles of various genes in plant metabolism, suggesting that they may indirectly influence metabolite levels through their broader metabolic functions.

Among the six identified genes from TWAS, we notably identified two genes which are MCOs associated with Quinic acid and CUEA associated with both Glycerol 1phosphate and L-glutamic acid (F.3). Although the direct relationship between these genes and their associated metabolites is not well-established, it is worth noting that the CUEA gene was also identified in the random forest results where it was found to be associated with Glycerol 1-phosphate. CUEA plays a crucial role in maintaining copper ion homeostasis within plant cells by actively transporting excess Cu(2+) ions out of the cytoplasm, either into the vacuole for storage or out of the cell, preventing toxic levels of copper from accumulating [29, 96, 72, 62]. MCOs catalyze the oxidation of various substrates, including polyphenols, aromatic polyamines, Lascorbate, and metal ions, and facilitate the reduction of dioxygen to water [121]. This activity is essential for lignin degradation, iron homeostasis, and defense against pathogens [47, 59, 10]. These TWAS findings especially underscore the integral roles of copper-related genes, such as CUEA and MCOs, in plant metabolism. Their involvement in crucial processes like ion homeostasis, lignin degradation, and defense mechanisms indicates that these genes may influence metabolite levels indirectly through their broader metabolic functions. This highlights the complex interactions within plant metabolic networks and the importance of copper-related genes in maintaining metabolic balance and adaptability.

Among the 79 genes identified through RF analysis, a notable one is N-acetylgamma-glutamyl-phosphate reductase (argC), which is associated with L-glutamic acid (F.4). ArgC is a key enzyme in the arginine biosynthesis pathway, catalyzing the reduction of N-acetylglutamate 5-phosphate to N-acetylglutamate 5-semialdehyde. This step is critical as it represents a committed stage in the production of arginine, an essential amino acid. L-glutamic acid serves as a precursor in this pathway, providing the initial substrate, and thus its metabolism is tightly linked to argC activity. The proper functioning of argC is crucial not only for arginine synthesis but also for overall nitrogen metabolism in plants, impacting growth and stress responses. ArgC has been extensively studied in E. coli, but its role in plants remains less explored, highlighting an area for further research. Understanding the function of argC in plants could provide insights into optimizing nitrogen use efficiency and improving stress tolerance. These findings underscore the essential role of argC in glutamic acid metabolism and its broader implications for plant health and development [6, 86, 21]. Additionally, We identified one overlapping gene, CUEA which is found in both TWAS and RF, and a number of other genes not identified by TWAS. These findings underscore the effectiveness of RF in identifying genes directly linked to metabolite production. This also demonstrates the importance of incorporating methods like RF, which can capture both linear and nonlinear associations, alongside TWAS, which is limited to linear relationships. By using RF, we may gain a more comprehensive understanding of the complex interactions between gene expression and metabolite profiles. This approach provides deeper insights into the metabolic networks and enhances our ability to uncover significant gene-metabolite associations.

From our RMIPGWAS analysis for non-metabolite traits, we found that three genetic markers associated with variations in metabolite traits were also linked to non-metabolite traits, suggesting these loci are might be potential key players in both plant metabolism and other plant characteristics (F.6,7,8). Additionally, we identified three candidate genes associated with these genetic markers that exhibit strong LD and have well-established roles in plant metabolism. ALDH is an enzyme family essential in plant metabolism, involved in detoxifying reactive aldehydes produced during various metabolic processes. This detoxification helps maintain cellular homeostasis, protects against oxidative stress, and supports plant development, growth, and stress response mechanisms [136, 177, 145, 56, 173, 158]. The MYB gene family regulates gene expression related to plant growth, development, stress response, and secondary metabolism, including the biosynthesis of flavonoids, phenolic acids, and monolignols. These compounds contribute to the cell wall structure, UV protection, and defense mechanisms [16, 167, 2, 111]. CDKs regulate the cell cycle by phosphorylating target proteins and controlling cell division and growth. They are involved in DNA replication, repair, transcription, and stress responses, enabling plants to adapt to environmental stresses like drought, salinity, and pathogen attack [108, 63, 104, 84, 26]. These overlapping genes identified between these traits need to be studied extensively. Along with RMIPGWAS, incorporating other statistical analyses such as machine learning approaches may be essential to identify key genetic variants associated with both metabolite and non-metabolite traits. Understanding the genes controlling variation in both types of traits may also reveal how these traits are associated, providing deeper insights into their interconnected biological pathways [61].

Our study utilized extensive genomic and gene expression data, employing a combination of quantitative genetics and machine learning methods, to identify key genes influencing essential metabolic processes and provide protection against various biotic and abiotic stresses, as well as their impact on both metabolite and non-metabolite traits. Through this approach, we have uncovered significant insights into the genetic architecture underlying plant metabolism. Our analysis showed that each method identified unique sets of genes associated with metabolite variation, demonstrating the complementary nature of different genomic approaches and their distinct input data. Interestingly, our study found that RF analysis, using the same expression data applied in TWAS, identified genes with direct associations to metabolites that were not detected by TWAS. These findings suggest that employing machine learning techniques like RF, which can captures both linear and nonlinear relationships, may be important for identifying genes associated with metabolites using gene expression data. Integrating machine learning analysis with quantitative genetics analyses such as GWAS and TWAS could enhance our ability to uncover complex genetic influences on metabolic traits. The identification of the key gene which is N-acetylgamma- glutamyl-phosphate reductase (argC) associated with L-glutamic acid and overlapping gene Cu(2+)-exporting ATPase (CUEA) associated with both Glycerol 1-phosphate identified across different methodologies point to their potential as key players in plant metabolism and development and highlights the importance of integrating multiple analytical approaches to reveal the gene metabolite association.

Our RMIPGWAS analysis uncovered that specific genetic markers associated with variations in metabolite traits are also linked to non-metabolite traits. This dual association indicates that these loci may play crucial roles in both plant metabolism and other plant characteristics. This dual functionality underscores the complex network of interactions that govern plant traits and metabolite traits, highlighting the importance of thoroughly studying these overlapping genes. By doing so, we can gain deeper insights into their interconnected biological pathways, ultimately enhancing our understanding of plant biology as a whole.

Conclusion

In conclusion, our study underscores the complexity of plant metabolism and its genetic basis, highlighting the importance of using both quantitative genetics and machine learning analyses to uncover the full range of gene-metabolite associations. This comprehensive approach is crucial for enhancing our understanding of plant metabolic networks and their genetic regulation, which is essential for addressing challenges related to plant resilience and productivity. Continued research in this area, incorporating diverse statistical and machine learning methods, will enhance our ability to identify potential gene-metabolite associations with improved stress tolerance and metabolic efficiency. The promising results from our study pave the way for future research to explore the roles of these identified metabolites and genes in plant growth and development processes, ultimately contributing to the development of more resilient and productive crops.

References

[1] MultispeQ V2.0. PhotosynQ, East Lansing, MI, USA, 2023. Available from: https://photosynq.org/.

[2] Farhat Abbas, Yanguo Ke, Yiwei Zhou, Yunyi Yu, Muhammad Waseem, Umair Ashraf, Chutian Wang, Xiaoyu Wang, Xinyue Li, Yuechong Yue, et al. Genomewide analysis reveals the potential role of myb transcription factors in floral scent formation in hedychium coronarium. Frontiers in Plant Science, 12:623742, 2021.

[3] Fred W Allendorf, Paul A Hohenlohe, and Gordon Luikart. Genomics and the future of conservation genetics. Nature reviews genetics, 11(10):697–709, 2010.

[4] L Almagro, LV G´omez Ros, S Belchi-Navarro, R Bru, A Ros Barcelo´, and MA Pedren˜o. Class iii peroxidases in plant defence reactions. Journal of experimental botany, 60(2):377–390, 2009.

[5] Mar´ıa Jos´e Aranzana, Sung Kim, Keyan Zhao, Erica Bakker, Matthew Horton, Katrin Jakob, Clare Lister, John Molitor, Chikako Shindo, Chunlao Tang, et al. Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS genetics, 1(5):e60, 2005.

[6] Annette Baich and Henry J Vogel. N-acetyl-γ-glutamokinase and n-acetylglutamic γ-semialdehyde dehydrogenase: Repressible enzymes of arginine synthesis in escherichiacoli. Biochemical and biophysical research communications, 7(6):491–496, 1962.

[7] Douglas M Bates. lme4: Mixed-effects modeling with r, 2010.

[8] Andr´e Bel´o, Peizhong Zheng, Stanley Luck, Bo Shen, David J Meyer, Bailin Li, Scott Tingey, and Antoni Rafalski. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Molecular Genetics and Genomics, 279:1–10, 2008.

[9] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300, 1995.

[10] Mar´ıa Bernal and Ute Kr¨amer. Involvement of arabidopsis multi-copper oxidase-encoding laccase12 in root-to-shoot iron partitioning: a novel example of copper-iron crosstalk. Frontiers in Plant Science, 12:688318, 2021.

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

[12] Nicolas L Bray, Harold Pimentel, Pa´ll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. Nature biotechnology, 34(5):525–527, 2016. [13] Leo Breiman. Random forests. Machine learning, 45:5–32, 2001.

[14] Damian Brzyski, Christine B Peterson, Piotr Sobczyk, Emmanuel J Cand`es, Malgorzata Bogdan, and Chiara Sabatti. Controlling the rate of gwas false discoveries. Genetics, 205(1):61–75, 2017.

[15] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. PLoS computational biology, 8(12):e1002822, 2012.

[16] Yunpeng Cao, Kui Li, Yanli Li, Xiaopei Zhao, and Lihu Wang. Myb transcription factors as regulators of secondary metabolism in plants. Biology, 9(3):61, 2020.

[17] Fabrizio Carbone, Anja Preuss, Ric CH De Vos, ELEONORA D'AMICO, Gaetano Perrotta, Arnaud G Bovy, Stefan Martens, and Carlo Rosati. Developmental, genetic and environmental factors affect the expression of flavonoid genes, enzymes and metabolites in strawberry fruits. Plant, Cell & Environment, 32(8):1117–1131, 2009.

[18] Natalia Carreno-Quintero, Animesh Acharjee, Chris Maliepaard, Christian WB

Bachem, Roland Mumm, Harro Bouwmeester, Richard GF Visser, and Joost JB Keurentjes. Untargeted metabolic quantitative trait loci analyses reveal a relationship between primary metabolism and potato tuber quality. Plant physiology, 158(3):1306–1318, 2012.

[19] Eva KF Chan, Heather C Rowe, Bjarne G Hansen, and Daniel J Kliebenstein. The complex genetic architecture of the metabolome. PLoS genetics, 6(11):e1001198, 2010.

[20] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. Gigascience, 4(1):s13742–015, 2015.

[21] Daniel Charlier and Indra Bervoets. Regulation of arginine biosynthesis, catabolism and transport in escherichia coli. Amino Acids, 51:1103–1127, 2019.

[22] Wei Chen, Yanqiang Gao, Weibo Xie, Liang Gong, Kai Lu, Wensheng Wang,

Yang Li, Xianqing Liu, Hongyan Zhang, Huaxia Dong, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nature genetics, 46(7):714–721, 2014.

[23] Wei Chen, Wensheng Wang, Meng Peng, Liang Gong, Yanqiang Gao, Jian Wan, Shouchuang Wang, Lei Shi, Bin Zhou, Zongmei Li, et al. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. Nature communications, 7(1):12767, 2016.

[24] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. Genomics, 99(6):323–329, 2012.

[25] Julie Chong, Marie-Agnees Pierrel, Rossitza Atanassova, Daniele WerckReichhart, Bernard Fritig, and Patrick Saindrenan. Free and conjugated benzoic acid in tobacco plants and cell cultures. induced accumulation upon elicitation of defense responses and role as salicylic acid precursors. Plant Physiology, 125(1):318–328, 2001.

[26] Leelyn Chong, Xiaoning Shi, and Yingfang Zhu. Signal integration by cyclindependent kinase 8 (cdk8) module and other mediator subunits in biotic and abiotic stress responses. International Journal of Molecular Sciences, 22(1):354, 2020.

[27] Vincenzo De Luca, Vonny Salim, Sayaka Masada Atsumi, and Fang Yu. Mining the biodiversity of plants: a revolution in the making. Science, 336(6089):1658– 1661, 2012.

[28] Youn`es Dellero, Mathieu Jossier, Jessica Schmitz, Veronica G Maurino, and Michael Hodges. Photorespiratory glycolate–glyoxylate metabolism. Journal of Experimental Botany, 67(10):3041–3052, 2016.

[29] Fenglin Deng, Naoki Yamaji, Jixing Xia, and Jian Feng Ma. A member of the heavy metal p-type atpase oshma5 is involved in xylem loading of copper in rice. Plant physiology, 163(3):1353–1362, 2013.

[30] Maninder Singh Dhillon, Thorsten Dahms, Carina Kuebert-Flock, Thomas Rummler, Joel Arnault, Ingolf Steffan-Dewenter, and Tobias Ullmann. Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape. Frontiers in Remote Sensing, 3:1010978, 2023.

[31] Zehong Ding, Lili Fu, Bin Wang, Jianqiu Ye, Wenjun Ou, Yan Yan, Meiying Li, Liwang Zeng, Xuekui Dong, Weiwei Tie, et al. Metabolic gwas-based dissection of genetic basis underlying nutrient quality variation and domestication of cassava storage root. Genome Biology, 24(1):289, 2023.

[32] Laure Dumont, Mark B Richardson, Phillip van der Peet, Danushka S Marapana, Tony Triglia, Matthew WA Dixon, Alan F Cowman, Spencer J Williams, Leann Tilley, Malcolm J McConville, et al. The metabolite repair enzyme phosphoglycolate phosphatase regulates central carbon metabolism and fosmidomycin sensitivity in plasmodium falciparum. MBio, 10(6):10–1128, 2019.

[33] Olive Jean Dunn. Multiple comparisons among means. Journal of the American statistical association, 56(293):52–64, 1961.

[34] MU Eteng, EU Eyong, EO Akpanyung, MA Agiang, and CY Aremu. Recent advances in caffeine and theobromine toxicities: a review. Plant foods for human nutrition, 51:231–243, 1997.

[35] Pengxiang Fan, Peipei Wang, Yann-Ru Lou, Bryan J Leong, Bethany M Moore, Craig A Schenck, Rachel Combs, Pengfei Cao, Federica Brandizzi, Shin-Han Shiu, et al. Evolution of a plant gene cluster in solanaceae and emergence of metabolic diversity. Elife, 9:e56717, 2020.

[36] Chuanying Fang, Alisdair R Fernie, and Jie Luo. Exploring the diversity of plant metabolism. Trends in Plant Science, 24(1):83–98, 2019.

[37] John N Ferguson, Samuel B Fernandes, Brandon Monier, Nathan D Miller, Dylan Allen, Anna Dmitrieva, Peter Schmuker, Roberto Lozano, Ravi Valluru, Edward S Buckler, et al. Machine learning-enabled phenotyping for gwas and twas of wue traits in 869 field-grown sorghum accessions. Plant Physiology, 187(3):1481–1500, 2021.

[38] Alisdair R Fernie and Nicolas Schauer. Metabolomics-assisted breeding: a viable option for crop improvement? Trends in genetics, 25(1):39–48, 2009.

[39] Alisdair R Fernie and Takayuki Tohge. The genetics of plant metabolism. Annual Review of Genetics, 51:287–310, 2017.

[40] Sergei A Filichkin, Ghislain Breton, Henry D Priest, Palitha Dharmawardhana, Pankaj Jaiswal, Samuel E Fox, Todd P Michael, Joanne Chory, Steve A Kay, and Todd C Mockler. Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. PloS one, 6(6):e16907, 2011.

[41] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.

[42] Manuel Garc´ıa-Magarin˜os, Inaki Lo´pez-de Ullibarri, Ricardo Cao, and Antonio Salas. Evaluating the ability of tree-based methods and logistic regression for the detection of snp-snp interaction. Annals of human genetics, 73(3):360–369, 2009.

[43] Mitchell Gill, Robyn Anderson, Haifei Hu, Mohammed Bennamoun, Jakob Petereit, Babu Valliyodan, Henry T Nguyen, Jacqueline Batley, Philipp E Bayer, and David Edwards. Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. BMC plant biology, 22(1):1–8, 2022.

[44] Liang Gong, Wei Chen, Yanqiang Gao, Xianqing Liu, Hongyan Zhang, Caiguo Xu, Sibin Yu, Qifa Zhang, and Jie Luo. Genetic analysis of the metabolome exemplified using a rice population. Proceedings of the National Academy of Sciences, 110(50):20320–20325, 2013.

[45] Oscar Gonza´lez-Recio and Selma Forni. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genetics Selection Evolution, 43:1–12, 2011.

[46] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane,

Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. Nucleic acids research, 40(D1):D1178–D1186, 2012.

[47] Gregor Grass and Christopher Rensing. Cueo is a multi-copper oxidase that confers copper tolerance in escherichia coli. Biochemical and biophysical research communications, 286(5):902–908, 2001.

[48] Marcin W Grzybowski, Ravi V Mural, Gen Xu, Jonathan Turkus, Jinliang Yang, and James C Schnable. A common resequencing-based genetic marker data set for global maize diversity. The Plant Journal, 113(6):1109–1121, 2023.

[49] Zhigang Guo, Michael M Magwire, Christopher J Basten, Zhanyou Xu, and Daolong Wang. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theoretical and applied genetics, 129:2413–2427, 2016.

[50] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics, 48(3):245–252, 2016.

[51] Susumu Hiraga, Katsutomo Sasaki, Hiroyuki Ito, Yuko Ohashi, and Hirokazu Matsui. A large family of class iii plant peroxidases. Plant and Cell Physiology, 42(5):462–468, 2001.

[52] Yanmin Hu, Feng Tang, Dan Zhang, Shihua Shen, and Xianjun Peng. Integrating genome-wide association and transcriptome analysis to provide molecular insights into heterophylly and eco-adaptability in woody plants. Horticulture Research, 10(11):uhad212, 2023.

[53] Xuehui Huang, Xinghua Wei, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, Chuanrang Zhu, Tingting Lu, Zhiwu Zhang, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nature genetics, 42(11):961–967, 2010.

[54] Matthew B Hufford, Arun S Seetharam, Margaret R Woodhouse, Kapeel M Chougule, Shujun Ou, Jianing Liu, William A Ricci, Tingting Guo, Andrew Olson, Yinjie Qiu, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science, 373(6555):655–662, 2021.

[55] Tasiu Isah. Stress and defense responses in plant secondary metabolites production. Biological research, 52, 2019.

[56] Md Sifatul Islam, Munira Mohtasim, Tahmina Islam, and Ajit Ghosh. Aldehyde dehydrogenase superfamily in sorghum: genome-wide identification, evolution, and transcript profiling during development stages and stress conditions. BMC Plant Biology, 22(1):316, 2022.

[57] Jig Han Jeong, Jonathan P Resop, Nathaniel D Mueller, David H Fleisher, Kyungdahm Yun, Ethan E Butler, Dennis J Timlin, Kyo-Moon Shim, James S Gerber, Vangimalla R Reddy, et al. Random forests for global and regional crop yield predictions. PloS one, 11(6):e0156571, 2016.

[58] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. Genetics, 178(3):1709–1723, 2008.

[59] Kavleen Kaur, Aarjoo Sharma, Neena Capalash, and Prince Sharma. Multicopper oxidases: Biocatalysts in microbial pathogenesis and stress management. Microbiological research, 222:1–13, 2019.

[60] Tomonori Kawano. Roles of the reactive oxygen species-generating peroxidase reactions in plant defense and growth induction. Plant cell reports, 21:829–837, 2003.

[61] Mirezhatijiang Kayoumu, Asif Iqbal, Noor Muhammad, Xiaotong Li, Leilei Li, Xiangru Wang, Huiping Gui, Qian Qi, Sijia Ruan, Ruishi Guo, et al. Phosphorus availability affects the photosynthesis and antioxidant system of contrasting low-p-tolerant cotton genotypes. Antioxidants, 12(2):466, 2023.

[62] Habib Khoudi. Significance of vacuolar proton pumps and metal/h+ antiporters in plant heavy metal tolerance. Physiologia Plantarum, 173(1):384–393, 2021.

[63] Georgios Kitsios and John H Doonan. Cyclin dependent protein kinases and stress responses in plants. Plant signaling & behavior, 6(2):204–209, 2011.

[64] Lingling Kong, Pingping Liu, Moli Li, Huizhen Wang, Jiaoxia Shi, Jingjie Hu, Yueru Li, and Xiaoli Hu. Transcriptional responses of flavin-containing monooxygenase genes in scallops exposed to pst-producing dinoflagellates implying their involvements in detoxification. Frontiers in Marine Science, 8:732000, 2021.

[65] Yoko Koyama, Yoshihisa Tomoda, Misako Kato, and Hiroshi Ashihara. Metabolism of purine bases, nucleosides and alkaloids in theobromine-forming theobroma cacao leaves. Plant Physiology and Biochemistry, 41(11-12):977–984, 2003.

[66] Sebastian Kuhlgert, Greg Austic, Robert Zegarac, Isaac Osei-Bonsu, Donghee Hoh, Martin I Chilvers, Mitchell G Roth, Kevin Bi, Dan TerAvest, Prabode Weebadde, et al. Multispeq beta: a tool for large-scale plant phenotyping connected to the open photosynq network. Royal Society open science, 3(10):160592, 2016.

[67] Max Kuhn. Building predictive models in r using the caret package. Journal of statistical software, 28:1–26, 2008.

[68] Xianjun Lai, Claire Bendix, Lang Yan, Yang Zhang, James C Schnable, and Frank G Harmon. Interspecific analysis of diurnal gene regulation in panicoid grasses identifies known and novel regulatory motifs. BMC genomics, 21:1–17,2020.

[69] Eric Lander and Leonid Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature genetics, 11(3):241– 247, 1995.

[70] Eric S Lander and Nicholas J Schork. Genetic dissection of complex traits. Focus, 265(3):2037–458, 2006.

[71] S´ebastien Lˆe, Julie Josse, and Fran¸cois Husson. Factominer: an r package for multivariate analysis. Journal of statistical software, 25:1–18, 2008.

[72] Xiangpeng Leng, Qian Mu, Xiaomin Wang, Xiaopeng Li, Xudong Zhu, Lingfei Shangguan, and Jinggui Fang. Transporters, chaperones, and p-type atpases controlling grapevine copper homeostasis. Functional & integrative genomics, 15:673–684, 2015.

[73] Bo Li, Nanxi Zhang, You-Gan Wang, Andrew W George, Antonio Reverter, and Yutao Li. Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. Frontiers in genetics, 9:237, 2018.

[74] Delin Li, Qiang Liu, and Patrick S Schnable. Twas results are complementary to and less affected by linkage disequilibrium than gwas. Plant physiology, 186(4):1800–1811, 2021.

[75] Delin Li, Qi Wang, Yu Tian, Xiangguang Lyu, Hao Zhang, Yinglu Sun, Huilong Hong, Huawei Gao, Yan-Fei Li, Chaosen Zhao, et al. Transcriptome brings

variations of gene expression, alternative splicing, and structural variations into gene-scale trait dissection in soybean. bioRxiv, pages 2023–07, 2023.

[76] Meng Li, Xiaolei Liu, Peter Bradbury, Jianming Yu, Yuan-Ming Zhang, Rory J Todhunter, Edward S Buckler, and Zhiwu Zhang. Enrichment of statistical power for genome-wide association studies. BMC biology, 12:1–10, 2014.

[77] Zhiyong Li, Chunhui Li, Yaxing Shi, Hui Dong, Senlin Xiao, Ruyang Zhang, Hui Liu, Yanyan Jiao, Aiguo Su, Xiaqing Wang, et al. Full-scale landscape metabolome map provides insights to convergent metabolite divergence and promotes edible maize breeding. 2023.

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

[79] Fan Lin, Jue Fan, and Seung Y Rhee. Qtg-finder: A machine-learning based algorithm to prioritize causal genes of quantitative trait loci in arabidopsis and rice. G3: Genes, Genomes, Genetics, 9(10):3129–3138, 2019.

[80] 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):833–835, 2011.

[81] Jan Lisec, Rhonda C Meyer, Matthias Steinfath, Henning Redestig, Martina Becher, Hanna Witucka-Wall, Oliver Fiehn, Otto´ T¨orj´ek, Joachim Selbig, Thomas Altmann, et al. Identification of metabolic and biomass qtl in arabidopsis thaliana in a parallel analysis of ril and il populations. The Plant Journal, 53(6):960–972, 2008.

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

[83] Feiyang Ma, Brie K Fuqua, Yehudit Hasin, Clara Yukhtman, Chris D Vulpe, Aldons J Lusis, and Matteo Pellegrini. A comparison between whole transcript and 3'rna sequencing methods using kapa and lexogen library preparation methods. BMC genomics, 20:1–12, 2019.

[84] Xiaoyan Ma, Zhu Qiao, Donghua Chen, Weiguo Yang, Ruijia Zhou, Wei Zhang, and Mei Wang. Cyclin-dependent kinase g2 regulates salinity stress response and salt mediated flowering in arabidopsis thaliana. Plant Molecular Biology, 88:287–299, 2015.

[85] Hiroshi Maeda and Natalia Dudareva. The shikimate pathway and aromatic amino acid biosynthesis in plants. Annual review of plant biology, 63:73–105, 2012.

[86] Rajtilak Majumdar, Boubker Barchi, Swathi A Turlapati, Maegan Gagne, Rakesh Minocha, Stephanie Long, and Subhash C Minocha. Glutamate, ornithine, arginine, proline, and polyamine metabolic interactions: the pathway is regulated at the post-transcriptional level. Frontiers in plant Science, 7:78, 2016.

[87] Malvern Panalytical Ltd. Fieldspec4 spectroradiometer. Previously Analytical Spectral Devices, 2023. Available from: https://www.malvernpanalytical. com.

[88] T Mamedov, G Zakiyeva, F Demirel, G Mammadova, And G Hasanova. Isolation, cloning, and gene expression analysis of phosphoglycolate phosphatase from green alga chlamydomonas reinhardtii. Photosynthetica, 62(1):90–101, 2024.

[89] Andrea Matros, Guozheng Liu, Anja Hartmann, Yong Jiang, Yusheng Zhao, Huange Wang, Erhard Ebmeyer, Viktor Korzun, Ralf Schachschneider, Ebrahim Kazman, et al. Genome–metabolite associations revealed low heritability, high genetic complexity, and causal relations for leaf metabolites in winter wheat (triticum aestivum). Journal of experimental botany, 68(3):415– 428, 2017.

[90] Fumio Matsuda, Yozo Okazaki, Akira Oikawa, Miyako Kusano, Ryo Nakabayashi, Jun Kikuchi, Jun-Ichi Yonemaru, Kaworu Ebana, Masahiro Yano, and Kazuki Saito. Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. The Plant Journal, 70(4):624–636, 2012.

[91] Mona Mazaheri, Marlies Heckwolf, Brieanne Vaillancourt, Joseph L Gage, Brett Burdo, Sven Heckwolf, Kerrie Barry, Anna Lipzen, Camila Bastos Ribeiro, Thomas JY Kono, et al. Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC plant biology, 19:1–17, 2019.

[92] Susan R McCouch, Mark H Wright, Chih-Wei Tung, Lyza G Maron, Kenneth L McNally, Melissa Fitzgerald, Namrata Singh, Genevieve DeClerck, Francisco Agosto-Perez, Pavel Korniliev, et al. Open access resources for genome-wide association mapping in rice. Nature communications, 7(1):10532, 2016.

[93] David B Medeiros, Yariv Brotman, and Alisdair R Fernie. The utility of metabolomics as a tool to inform maize biology. Plant Communications, 2(4), 2021.

[94] Todd P Michael, Ghislain Breton, Samuel P Hazen, Henry Priest, Todd C Mockler, Steve A Kay, and Joanne Chory. A morning-specific phytohormone gene expression program underlying rhythmic plant growth. PLoS biology, 6(9):e225, 2008.

[95] Todd P Michael and Scott Jackson. The first 50 plant genomes. Plant Genome, 6(2):1–7, 2013.

[96] Magdalena Migocka, Ewelina Posyniak, Ewa Maciaszczyk-Dziubinska, Anna Papierniak, and Anna Kosieradzaka. Functional and biochemical characterization of cucumber genes encoding two copper atpases cshma5. 1 and cshma5. 2.Journal of Biological Chemistry, 290(25):15717–15729, 2015.

[97] Ron Milo and Robert L Last. Achieving diversity in the face of constraints: lessons from metabolism. Science, 336(6089):1663–1667, 2012.

[98] Luchang Ming, Debao Fu, Zhaona Wu, Hu Zhao, Xingbing Xu, Tingting Xu, Xiaohu Xiong, Mu Li, Yi Zheng, Ge Li, et al. Transcriptome-wide association analyses reveal the impact of regulatory variants on rice panicle architecture and causal gene regulatory networks. Nature Communications, 14(1):7501, 2023.

[99] Osval Antonio Montesinos L´opez, Abelardo Montesinos Lo´pez, and Jose Crossa. Random forest for genomic prediction. In Multivariate statistical machine learning methods for genomic prediction, pages 633–681. Springer, 2022.

[100] Ravi V Mural, Guangchao Sun, Marcin Grzybowski, Michael C Tross, Hongyu Jin, Christine Smith, Linsey Newton, Carson M Andorf, Margaret R Woodhouse, Addie M Thompson, et al. Association mapping across a multitude of traits collected in diverse environments in maize. GigaScience, 11:giac080, 2022.

[101] S Naderi, T Yin, and S K¨onig. Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. Journal of Dairy Science, 99(9):7261–7273, 2016.

[102] Ki Hyun Nam. Glucose isomerase: functions, structures, and applications. Applied Sciences, 12(1):428, 2022.

[103] Yasser Nehela, Naglaa A Taha, Abdelnaser A Elzaawely, Tran Dang Xuan, Mohammed A. Amin, Mohamed E Ahmed, and Asmaa El-Nagar. Benzoic acid and

its hydroxylated derivatives suppress early blight of tomato (alternaria solani) via the induction of salicylic acid biosynthesis and enzymatic and nonenzymatic antioxidant defense machinery. Journal of Fungi, 7(8):663, 2021.

[104] Sophia Ng, Estelle Giraud, Owen Duncan, Simon R Law, Yan Wang, Lin Xu, Reena Narsai, Chris Carrie, Hayden Walker, David A Day, et al. Cyclindependent kinase e1 (cdke1) provides a cellular switch in plants between growth and stress responses. Journal of Biological Chemistry, 288(5):3449–3459, 2013.

[105] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nature genetics, 32(4):650–654, 2002.

[106] Rafael Melo Palhares, Marcela Gon¸calves Drummond, Bruno dos Santos Alves Figueiredo Brasil, Gustavo Pereira Cosenza, Maria das Grac¸as Lins Brand˜ao, and Guilherme Oliveira. Medicinal plants recommended by the world health organization: Dna barcode identification associated with chemical analyses guarantees their quality. PloS one, 10(5):e0127866, 2015.

[107] Fabian Pedregosa, Ga¨el Varoquaux, Alexandre Gramfort, Vincent Michel,Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.

[108] Adrian Peres, Michelle L Churchman, Srivaidehirani Hariharan, Kristiina Himanen, Aurine Verkest, Klaas Vandepoele, Zoltan Magyar, Yves Hatzfeld, Els Van Der Schueren, Gerrit TS Beemster, et

al. Novel plant-specific cyclin-dependent kinase inhibitors induced by biotic and abiotic stresses. Journal of Biological Chemistry, 282(35):25588–25596, 2007.

[109] Dolores R Piperno, Anthony J Ranere, Irene Holst, Jose Iriarte, and Ruth Dickau. Starch grain and phytolith evidence for early ninth millennium bp maize from the central balsas river valley, mexico. Proceedings of the National Academy of Sciences, 106(13):5019–5024, 2009.

[110] Lutz Pollmann and Michael Wettern. The ubiquitin system in higher and lower plants—pathways in protein metabolism. Botanica acta, 102(1):21–30, 1989.

[111] Durvasula Sumana Pratyusha and Dronamraju VL Sarada. Myb transcription factors—master regulators of phenylpropanoid biosynthesis and diverse developmental and stress responses. Plant Cell Reports, 41(12):2245–2260, 2022.

[112] 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– 909, 2006.

[113] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. Genetics, 155(2):945–959, 2000.

[114] R R Core Team et al. R: A language and environment for statistical computing. 2013.

[115] Amit Rai, Kazuki Saito, and Mami Yamazaki. Integrated omics analysis of specialized metabolism in medicinal plants, 2017.

[116] Dragana Rajkovi´c, Ana Marjanovi´c Jeromela, Lato Pezo, Biljana Lonˇcar, Federica Zanetti, Andrea Monti, and Ankica Kondi´c Spika. Yield and quality pre-ˇ diction of winter rapeseed—artificial neural network and random forest models. Agronomy, 12(1):58, 2021.

[117] Christian Riedelsheimer, Jan Lisec, Angelika Czedik-Eysenberg, Ronan Sulpice, Anna Flis, Christoph Grieder, Thomas Altmann, Mark Stitt, Lothar Willmitzer, and Albrecht E Melchinger. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proceedings of the National Academy of Sciences, 109(23):8872–8877, 2012.

[118] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. Science, 273(5281):1516–1517, 1996.

[119] Mariana Rosa, Carolina Prado, Griselda Podazza, Roque Interdonato, Juan A Gonza´lez, Mirna Hilal, and Fernando E Prado. Soluble sugars: Metabolism, sensing and abiotic stress: A complex network in the life of plants. Plant signaling & behavior, 4(5):388–393, 2009.

[120] Seema Sahay, Marcin Grzybowski, James C Schnable, and Katarzyna G lowacka. Genetic control of photoprotection and photosystem ii operating efficiency in plants. New Phytologist, 2023.

[121] Takeshi Sakurai and Kunishige Kataoka. Structure and function of type i copper in multicopper oxidases. Cellular and Molecular Life Sciences, 64:2642–2656, 2007.

[122] Rupam Kumar Sarkar, AR Rao, Prabina Kumar Meher, T Nepolean, and T Mohapatra. Evaluation of random forest regression for prediction of breeding value from genomewide snps. Journal of genetics, 94:187–192, 2015.

[123] Paul Schmidt, Jens Hartung, Jo¨rn Bennewitz, and Hans-Peter Piepho. Heritability in plant breeding on a genotype-difference basis. Genetics, 212(4):991– 1008, 2019.

[124] 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.

[125] Philip Sedgwick. Pearson's correlation coefficient. Bmj, 345, 2012.

[126] Vincent Segura, Bjarni J Vilhja´lmsson, Alexander Platt, Arthur Korte, Umit¨ Seren, Quan Long, and Magnus Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics, 44(7):825–830, 2012.

[127] Rajandeep S Sekhon, Christopher Saski, Rohit Kumar, Barry S Flinn, Feng Luo,

Timothy M Beissinger, Arlyn J Ackerman, Matthew W Breitzman, William C Bridges, Natalia de Leon, et al. Integrated genome-scale analysis identifies novel genes and networks underlying senescence in maize. The Plant Cell, 31(9):1968–1989, 2019.

[128] Tissa Senaratna, David Merritt, Kingsley Dixon, Eric Bunn, Darren Touchell, and KJPGR Sivasithamparam. Benzoic acid may act as the functional group in salicylic acid and derivatives in the induction of multiple stress tolerance in plants. Plant Growth Regulation, 39:77–81, 2003.

[129] Syed Haleem Shah, Yoseline Angel, Rasmus Houborg, Shawkat Ali, and Matthew F McCabe. A random forest machine learning approach for the retrieval of leaf chlorophyll content in wheat. Remote Sensing, 11(8):920, 2019.

[130] Bekele Shiferaw, Boddupalli M Prasanna, Jonathan Hellin, and Marianne Ba¨nziger. Crops that feed the world 6. past successes and future challenges to the role played by maize in global food security. Food security, 3:307–327, 2011.

[131] Manisha Sanjay Sirsat, Paula Rodrigues Oblessuc, and Ricardo S Ramiro. Genomic prediction of wheat grain yield using machine learning. Agriculture, 12(9):1406, 2022.

[132] Elpiniki Skoufogianni, Alexandra Solomou, Georgios Charvalas, and Nicholaos Danalatos. Maize as energy crop. In Maize-Production and use. IntechOpen London, UK, 2019.

[133] Sudhir K Sopory and Meenakshi Munshi. Protein kinases and phosphatases and their role in cellular signaling in plants. Critical Reviews in Plant Sciences, 17(3):245–318, 1998.

[134] Doug Speed and David J Balding. Relatedness in the post-genomic era: is it still useful? Nature Reviews Genetics, 16(1):33–44, 2015.

[135] Johannes Stephan, Oliver Stegle, and Andreas Beyer. A random forest approach to capture genetic effects in the presence of population structure. Nature communications, 6(1):7432, 2015.

[136] Naim Stiti, Tagnon D Missihoun, Simeon O Kotchoni, Hans-Hubert Kirch, and Dorothea Bartels. Aldehyde dehydrogenases in arabidopsis thaliana: biochemical requirements, metabolic pathways, and functional analysis. Frontiers in plant science, 2:65, 2011.

[137] Karsten Suhre and Christian Gieger. Genetic variation in metabolic phenotypes: study designs and applications. Nature reviews genetics, 13(11):759–769, 2012.

[138] Guangchao Sun, Ravi V Mural, Jonathan D Turkus, and James C Schnable. Quantitative resistance loci to southern rust mapped in a temperate maize diversity panel. Phytopathology®, 112(3):579–587, 2022.

[139] Mari Suzuki, Rieko Setsuie, and Keiji Wada. Ubiquitin carboxyl-terminal hydrolase l3 promotes insulin signaling and adipogenesis. Endocrinology, 150(12):5230–5239, 2009.

[140] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Boss´e, Guillaume Par´e, and David Meyre. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics, 20(8):467–484, 2019.

[141] Shan Tang, Hu Zhao, Shaoping Lu, Liangqian Yu, Guofang Zhang, Yuting Zhang, Qing-Yong Yang, Yongming Zhou, Xuemin Wang, Wei Ma, et al. Genome-and transcriptome-wide association studies provide insights into the genetic basis of natural variation of seed oil content in brassica napus. Molecular Plant, 14(3):470–487, 2021.

[142] Guillaume Tena, Marie Boudsocq, and Jen Sheen. Protein kinase signaling networks in plant innate immunity. Current opinion in plant biology, 14(5):519– 529, 2011.

[143] Jie Teng, Changyu Yan, Wen Zeng, Yuqian Zhang, Zhen Zeng, and Yahui Huang. Purification and characterization of theobromine synthase in a theobromine-enriched wild tea plant (camellia gymnogyna chang) from dayao mountain, china. Food chemistry, 311:125875, 2020.

[144] Laura Tibbs Cortes, Zhiwu Zhang, and Jianming Yu. Status and prospects of genome-wide association studies in plants. The plant genome, 14(1):e20077, 2021.

[145] Adesola J Tola, Amal Jaballi, Hugo Germain, and Tagnon D Missihoun. Recent development on plant aldehyde dehydrogenase enzymes and their functions in plant development and stress signaling. Genes, 12(1):51, 2020.

[146] J Vladimir Torres-Rodriguez, Guangchao Sun, Ravi V Mural, and James C Schnable. Measurement of expression from a limited number of genes is sufficient to predict flowering time in maize. bioRxiv, pages 2022–12, 2022.

[147] Jorge Vladimir Torres-Rodriguez, Delin Li, Jonathan Turkus, Linsey Newton,

Jensina Davis, Lina Lopez-Corona, Waqar Ali, Guangchao Sun, Ravi V Mural, Marcin W Grzybowski, et al. Population level gene expression can repeatedly link genes to functions in maize. bioRxiv, pages 2023–10, 2023.

[148] David Toubiana, Rami Puzis, Lingling Wen, Noga Sikron, Assylay Kurmanbayeva, Aigerim Soltabayeva, Maria del Mar Rubio Wilhelmi, Nir Sade, Aaron Fait, Moshe Sagi, et al. Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. Communications biology, 2(1):214, 2019.

[149] David Toubiana, Yaniv Semel, Takayuki Tohge, Romina Beleggia, Luigi Cattivelli, Leah Rosental, Zoran Nikoloski, Dani Zamir, Alisdair R Fernie, and Aaron Fait. Metabolic profiling of a mapping population exposes new insights in the regulation of seed metabolism and seed, fruit, and plant relations. PLoS genetics, 8(3):e1002612, 2012.

[150] Michael C Tross, Marcin W Grzybowski, Talukder Z Jubery, Ryleigh J Grove, Aime V Nishimwe, J Vladimir Torres-Rodriguez, Guangchao Sun, Baskar Ganapathysubramanian, Yufeng Ge, and James C

Schnable. Data driven discovery and quantification of hyperspectral leaf reflectance phenotypes across a maize diversity panel. The Plant Phenome Journal, 7(1):e20106, 2024.

[151] Hirotaka Uefuji, Shinjiro Ogita, Yube Yamaguchi, Nozomu Koizumi, and Hiroshi Sano. Molecular cloning and functional characterization of three distinct nmethyltransferases involved in the caffeine biosynthetic pathway in coffee plants.Plant physiology, 132(1):372–380, 2003.

[152] United States Department of Agriculture, Foreign Agricultural Service. Corn:World markets and trade, 2023. Accessed: 2024-04-01.

[153] William Valdar, Christopher C Holmes, Richard Mott, and Jonathan Flint. Mapping in structured populations by resample model averaging. Genetics, 182(4):1263–1277, 2009.

[154] Y Vigouroux, Michael McMullen, CT Hittinger, K Houchins, L Schulz, S Kresovich, Y Matsuoka, and J Doebley. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proceedings of the National Academy of Sciences, 99(15):9650–9655, 2002.

[155] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics, 101(1):5–22, 2017.

[156] Patrik Waldmann. Genome-wide prediction using bayesian additive regression trees. Genetics Selection Evolution, 48:1–12, 2016.

[157] Dong-Hui Wang, Wei Song, Shao-Wei Wei, Ya-Feng Zheng, Zhi-Shan Chen, Jing-Dan Han, Hong-Tao Zhang, Jing-Chu Luo, Yong-Mei Qin, Zhi-Hong Xu, et al. Characterization of the ubiquitin c-terminal hydrolase and ubiquitinspecific protease families in rice (oryza sativa). Frontiers in Plant Science, 9:1636, 2018.

[158] Haibo Wang, Yong Gao, and Junyun Guo. Comprehensive analysis of dicer-like, argonaute, and rna-dependent rna polymerase gene families and their expression analysis in response to abiotic stresses in jatropha curcas. Journal of Plant Interactions, 19(1):2282432, 2024.

[159] Liang Wang, Xiaoqi Pan, Lishi Jiang, Yu Chu, Song Gao, Xingyue Jiang, Yuhui Zhang, Yan Chen, Shajie Luo, and Cheng Peng. The biological activity mechanism of chlorogenic acid and its applications in food industry: A review. Frontiers in Nutrition, 9:943911, 2022.

[160] Nishikant Wase, Nathan Abshire, and Toshihiro Obata. High-throughput profiling of metabolic phenotypes using high-resolution gc-ms. In High-Throughput Plant Phenotyping: Methods and Protocols, pages 235–260. Springer, 2022.

[161] Shujun Wei, Ryokei Tanaka, Taiji Kawakatsu, Shota Teramoto, Nobuhiro Tanaka, Matthew Shenton, Yusaku Uga, and Shiori Yabe. Genome-and transcriptome-wide association studies to discover candidate genes for diverse root phenotypes in cultivated rice. Rice, 16(1):55, 2023.

[162] 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(1):3438, 2014.

[163] Saskia Windisch, Anja Walter, Narges Moradtalab, Frank Walker, BirgitHo¨glinger, Abbas El-Hasan, Uwe Ludewig, Gu¨nter Neumann, and Rita Grosch. Role of benzoic acid and lettucenin a in the defense response of lettuce against soil-borne pathogens. Plants, 10(11):2336, 2021.

[164] Di Wu, Xiaowei Li, Ryokei Tanaka, Joshua C Wood, Laura E Tibbs-Cortes,

Maria Magallanes-Lundback, Nolan Bornowski, John P Hamilton, Brieanne Vaillancourt, Christine H Diepenbrock, et al. Combining gwas and twas to identify candidate causal genes for tocochromanol levels in maize grain. Genetics, 221(4):iyac091, 2022.

[165] Sijia Wu, Wenjuan Chen, Sujuan Lu, Hailing Zhang, and Lianghong Yin. Metabolic engineering of shikimic acid biosynthesis pathway for the production of shikimic acid and its branched products in microorganisms: advances and prospects. Molecules, 27(15):4779, 2022.

[166] Hui Xie, Yu Wang, Yiqian Ding, Chen Qiu, Litao Sun, Zhongshuai Gai,

Honglian Gu, and Zhaotang Ding. Global ubiquitome profiling revealed the

roles of ubiquitinated proteins in metabolic pathways of tea leaves in responding to drought stress. Scientific reports, 9(1):4286, 2019.

[167] Huiling Yan, Xiaona Pei, Heng Zhang, Xiang Li, Xinxin Zhang, Minghui Zhao, Vincent L Chiang, Ronald Ross Sederoff, and Xiyang Zhao. Myb-mediated regulation of anthocyanin biosynthesis. International Journal of Molecular Sciences, 22(6):3103, 2021.

[168] Fengbo Yang, Yuchen Miao, Yuyue Liu, Jose R Botella, Weiqiang Li, Kun Li, and Chun-Peng Song. Function of protein kinases in leaf senescence of plants. Frontiers in Plant Science, 13:864215, 2022.

[169] Lilin Yin, Haohao Zhang, Zhenshuang Tang, Jingya Xu, Dong Yin, Zhiwu Zhang, Xiaohui Yuan, Mengjin Zhu, Shuhong Zhao, Xinyun Li, et al. rmvp: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics, proteomics & bioinformatics, 19(4):619–628, 2021.

[170] Go-Eun Yu, Younhee Shin, Sathiyamoorthy Subramaniyam, Sang-Ho Kang, Si-Myung Lee, Chuloh Cho, Seung-Sik Lee, and Chang-Kug Kim. Machine learning, transcriptome, and genotyping chip analyses provide insights into snp markers identifying flower color in platycodon grandiflorus. Scientific Reports, 11(1):8019, 2021.

[171] 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):203–208, 2006.

[172] Fei Zhang, Jinfeng Wu, Nir Sade, Si Wu, Aiman Egbaria, Alisdair R Fernie, Jianbing Yan, Feng Qin, Wei Chen, Yariv Brotman, et al. Genomic basis underlying the metabolome-mediated drought adaptation of maize. Genome Biology, 22:1–26, 2021.

[173] Xiaoming Zhang, Jingwen Zhong, Liang Cao, Chunyuan Ren, Gaobo Yu, Yanhua Gu, Jingwen Ruan, Siqi Zhao, Lei Wang, Haishun Ru, et al. Genome-wide characterization of aldehyde dehydrogenase gene family members in groundnut (arachis hypogaea) and the analysis under saline-alkali stress. Frontiers in Plant Science, 14:1097001, 2023.

[174] Xuehai Zhang, Marilyn L Warburton, Tim Setter, Haijun Liu, Yadong Xue, Ning Yang, Jianbing Yan, and Yingjie Xiao. Genome-wide association studies of drought-related metabolic changes in maize using an enlarged snp panel. Theoretical and Applied Genetics, 129:1449–1463, 2016.

[175] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K

Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. Nature genetics, 42(4):355–360, 2010.

[176] Zhiwu Zhang et al. GAPIT Version 3.1: Genomic Association and Prediction

Integrated Tool, 2021. Accessed: 2024-06-16.

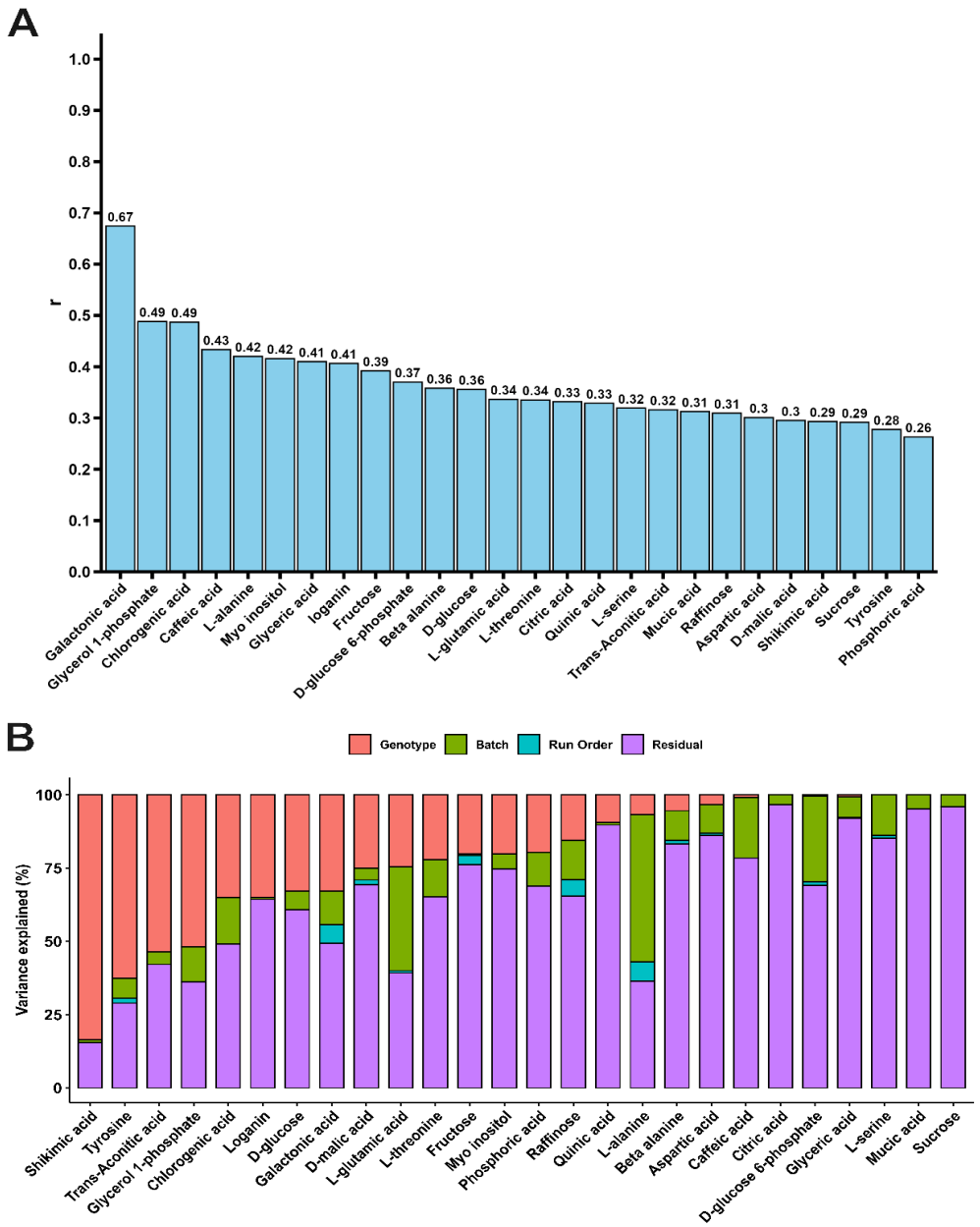
[177] Junyi Zhao, Tagnon D Missihoun, and Dorothea Bartels. The role of arabidopsis aldehyde dehydrogenase genes in response to high temperature and stress combinations. Journal of Experimental Botany, 68(15):4295–4308, 2017.

[178] Shaoqun Zhou, Karl A Kremling, Nonoy Bandillo, Annett Richter, Ying K

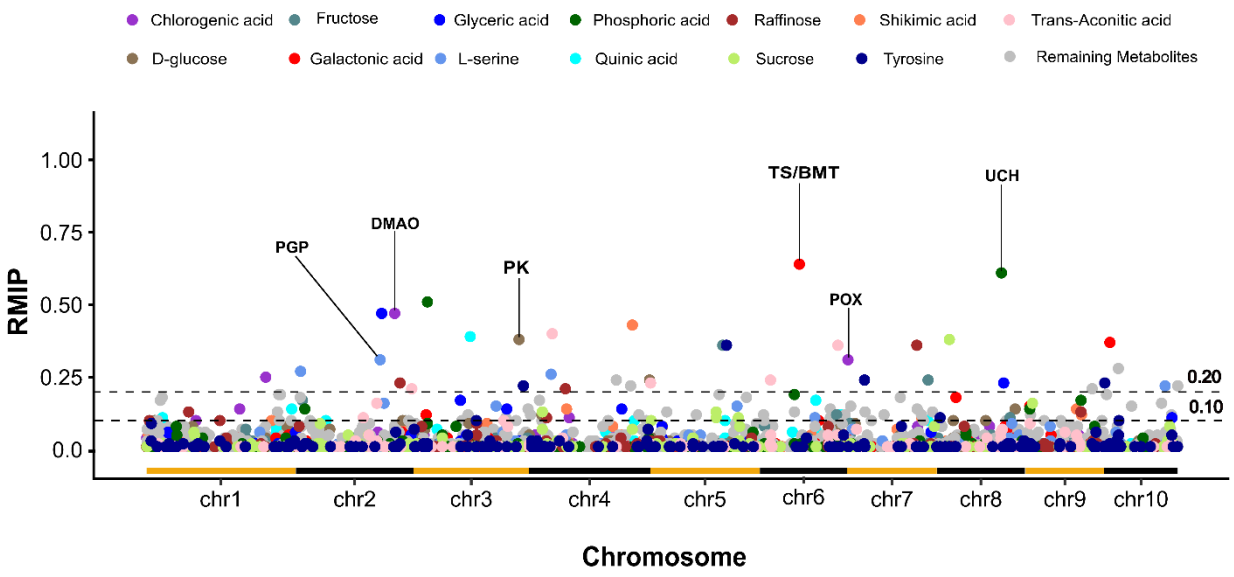
Zhang, Kevin R Ahern, Alexander B Artyukhin, Joshua X Hui, Gordon C Younkin, Frank C Schroeder, et al. Metabolome-scale genome-wide association studies reveal chemical diversity and genetic control of maize specialized metabolites. The Plant Cell, 31(5):937–955, 2019.

[179] Chengsong Zhu, Michael Gore, Edward S Buckler, and Jianming Yu. Status and prospects of association mapping in plants. The plant genome, 1(1), 2008.

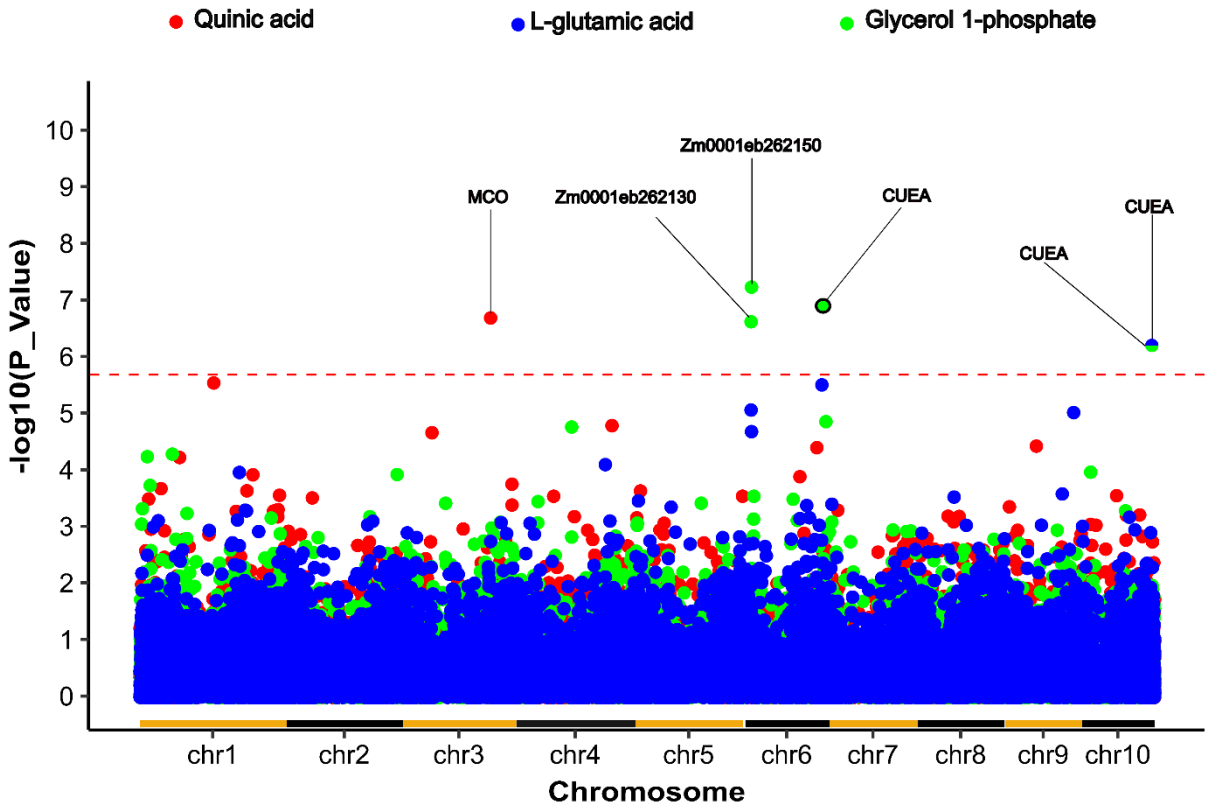
Main Figures



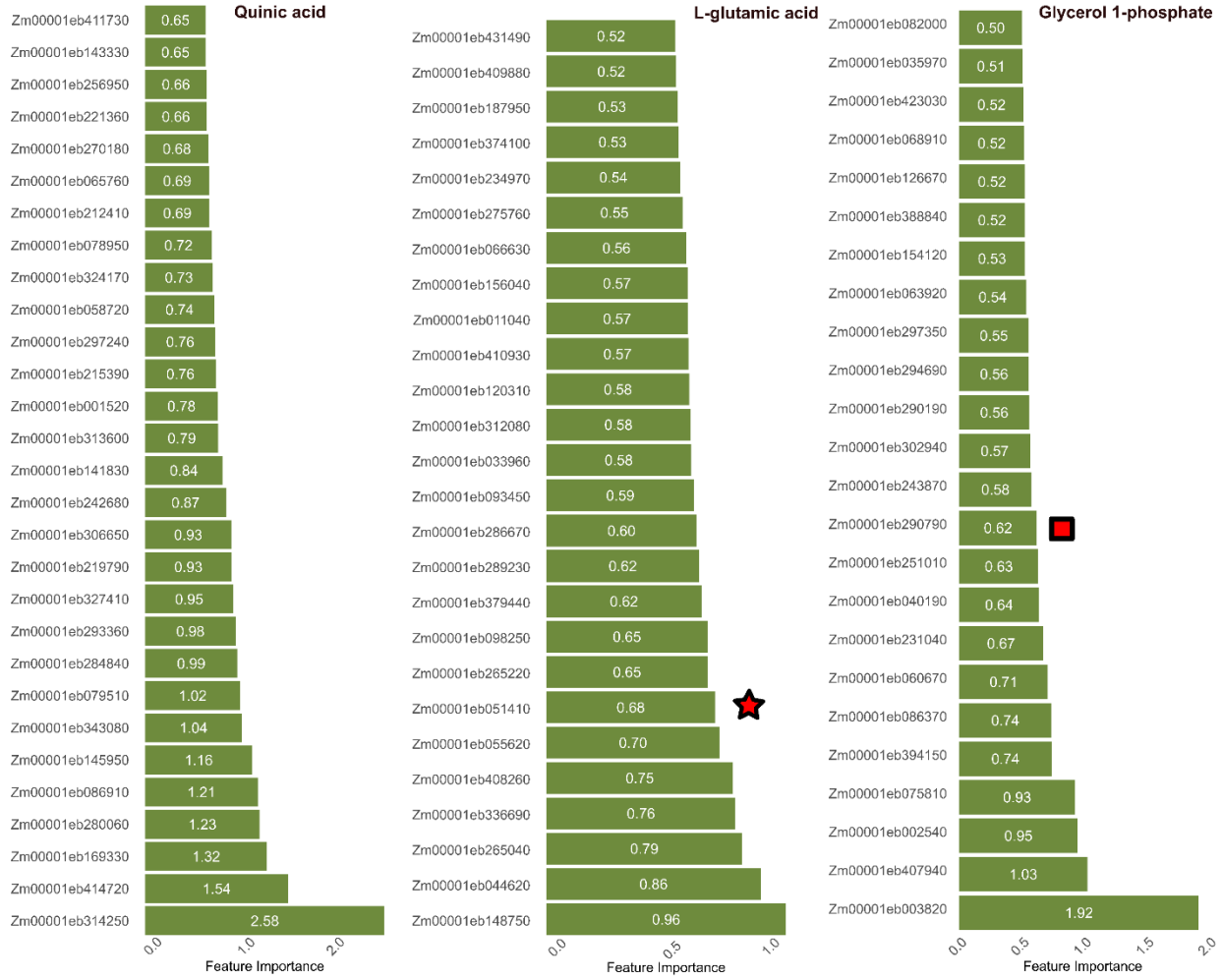
F.1: Property of the Metabolite Variation in Maize. A) Repeatability (r): Bar graph quantifying the Repeatability for various metabolites, where a higher bar suggests a stronger genetic influence on the metabolite's expression, and a lower bar suggests a lesser genetic influence. B) Variance Partitioning: A stacked bar chart shows the contribution of different factors to the total variance for each metabolite. The colors in the bars correspond to the proportion of variance attributed to Genotype, Batch, Run Order, and Residual factors.



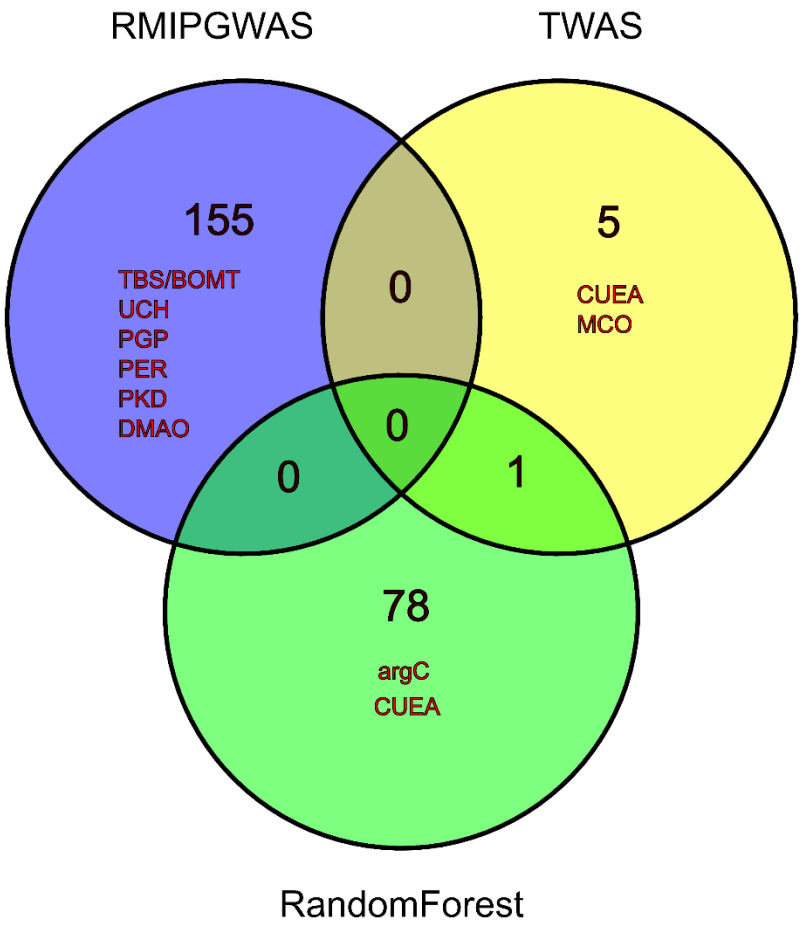
F.2: Genes Associated with Metabolite Variation via Resampling Model In- clusion Probability Genome-Wide Association Study (RMIPGWAS). A result of a RMIPGWAS conducted using the FarmCPU algorithm. The x-axis represents maize chromosomes, while the y-axis indicates RMIP values, reflecting the probability of SNP associations with the metabolite traits under study. Different colored dots rep- resent SNPs associated with specific metabolites, as indicated in the legend. These specific metabolites, chosen for their associations with an RMIP value of 0.3 and above, are highlighted in various colors, while remaining metabolites are highlighted in grey. The plot includes two horizontal dashed lines marking RMIP significance thresholds: the upper line at 0.2 (indicating SNPs significant in at least 20% of re- sampled datasets) and the lower line at 0.1 (indicating SNPs significant in at least 10% of resampled datasets). Key genes are highlighted above the plot. The physical positions between the chromosomes are marked with horizontal lines in two different colors.



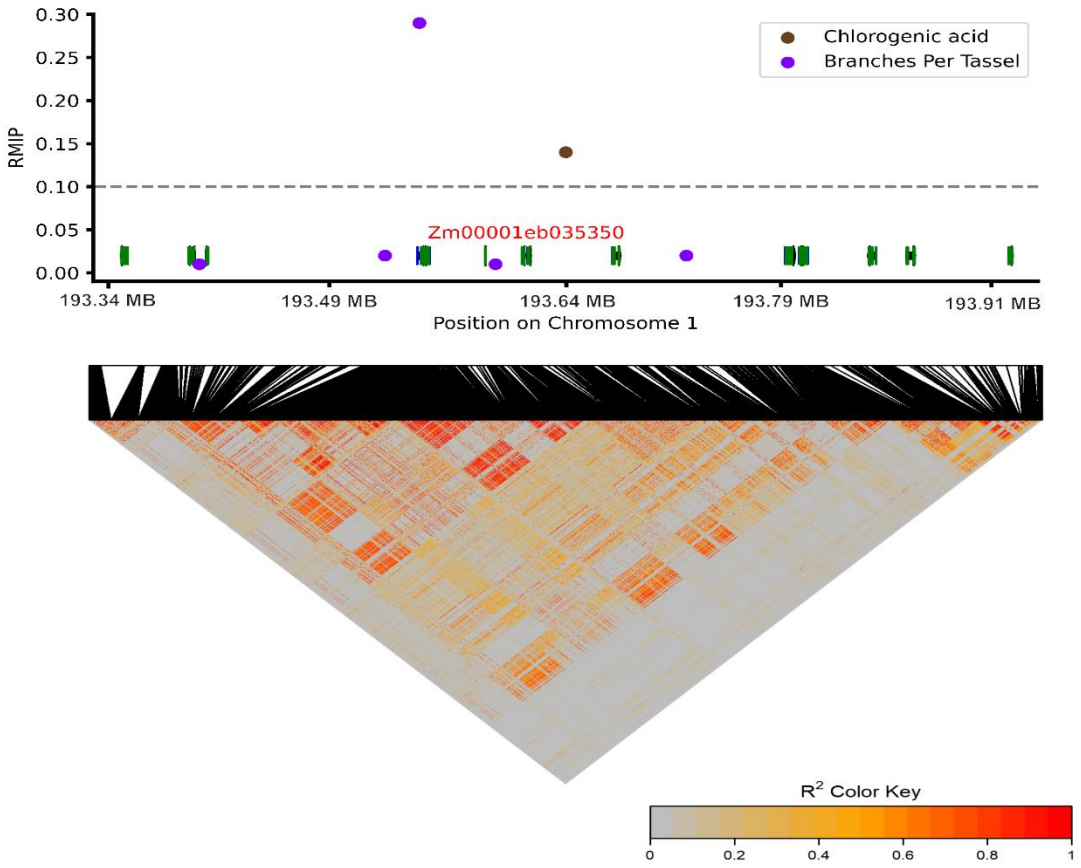
F.3: Genes Associated with Metabolite Variation via Transcriptome-wide Association Studies (TWAS). The TWAS results use transcript abundance data to identify genes associated with metabolite variation. The x-axis represents maize chromosomes, while the y-axis shows the -log10(p-values) for associations between gene expression levels and metabolite concentrations. Individual genes are represented by dots, with those above the dashed red line meeting the Bonferroni-corrected significance threshold, indicating a significant association with metabolite variations. Different colored dots correspond to genes associated with specific metabolites, as detailed in the legend. Key genes are highlighted above the plot, with overlapping genes identified between TWAS and RF analysis circled in two round circles.



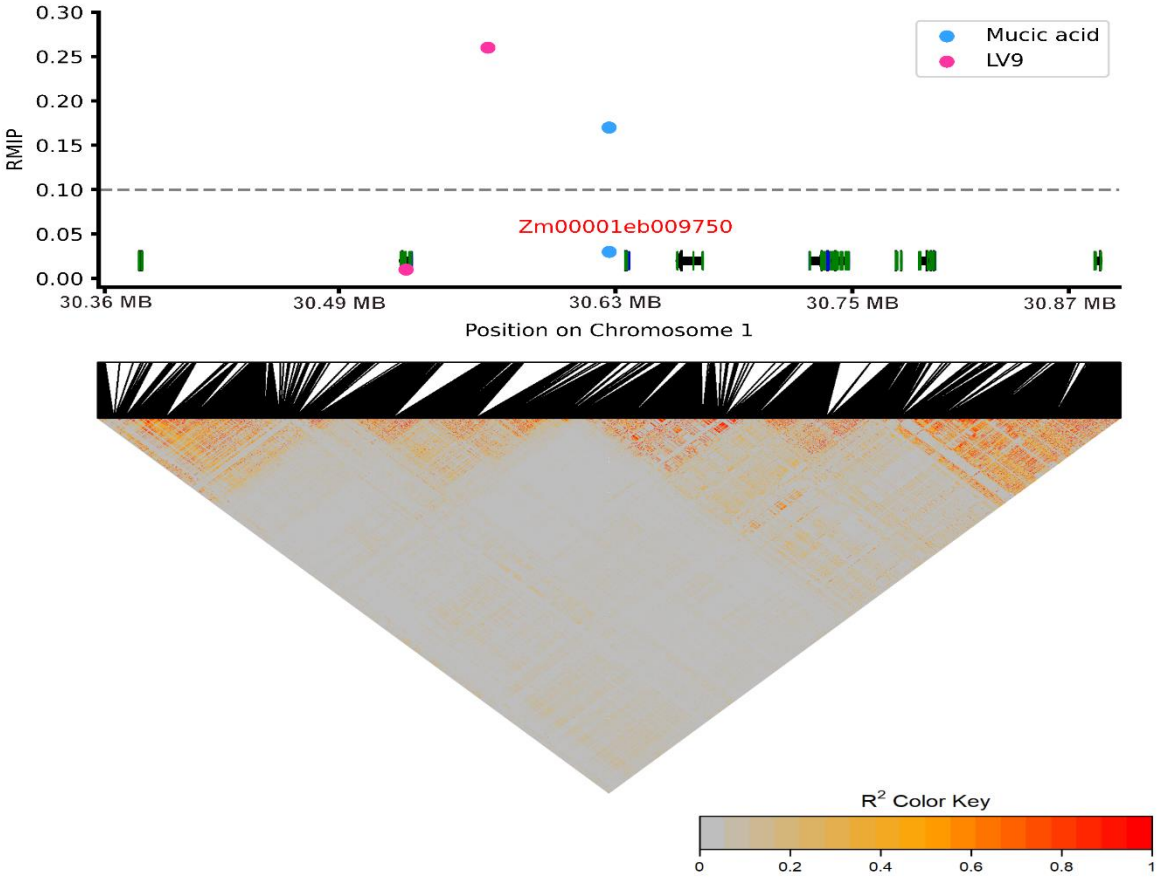
F.4: A Random Forest (RF) feature importance-based approach was conducted for three specific metabolites, each with at least one significant association found in both RMIPGWAS and TWAS. This panel displays a series of bar charts, where the x-axis represents the feature importance scores (as numerical values) and the y-axis lists the genes identified by the Random Forest analysis. Two key genes are highlighted with distinct shapes for emphasis: a star shape indicates a gene directly associated with metabolites, while a square shape indicates genes that overlap with both TWAS and RF results. The feature importance scores provide insight into the significance of each gene in relation to metabolite variation, underscoring their critical roles in plant metabolism and development.



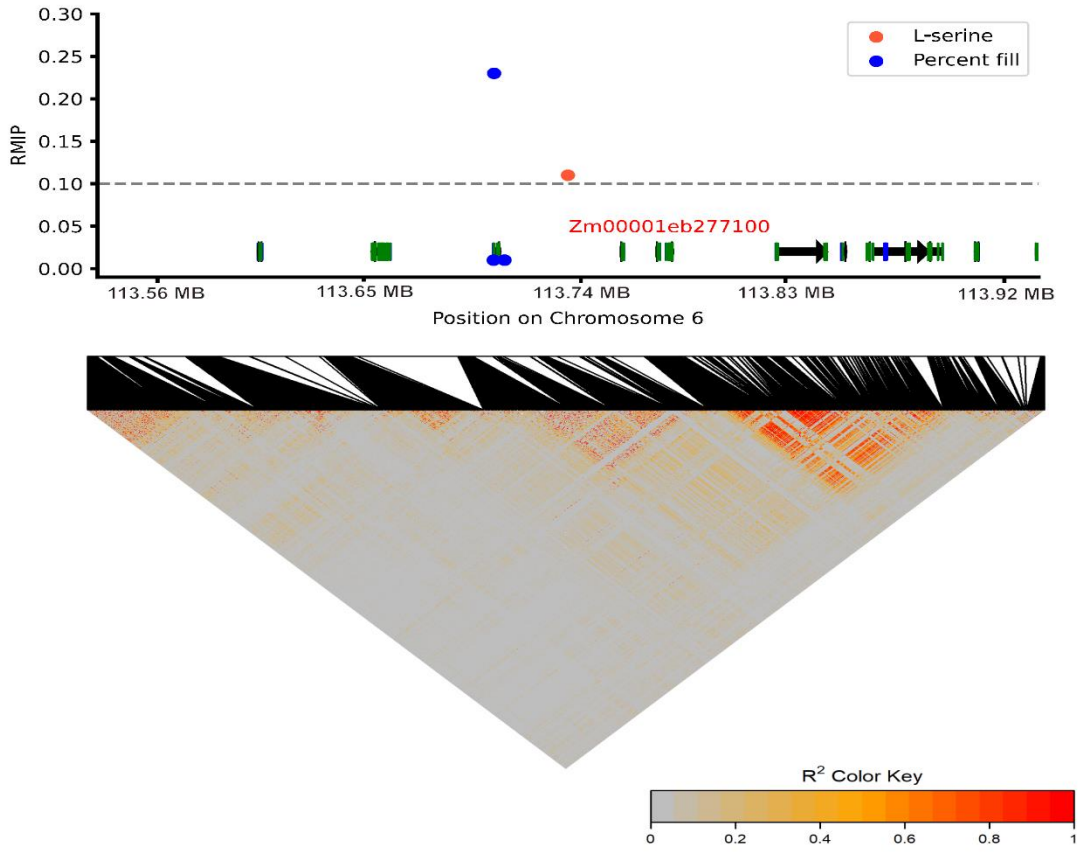
F.5: Venn Diagram shows the numbers of shared and uniquely identified genes associated with metabolite variation using quantitative genetics and machine learning methods with key genes highlighted mentioned in each circle.



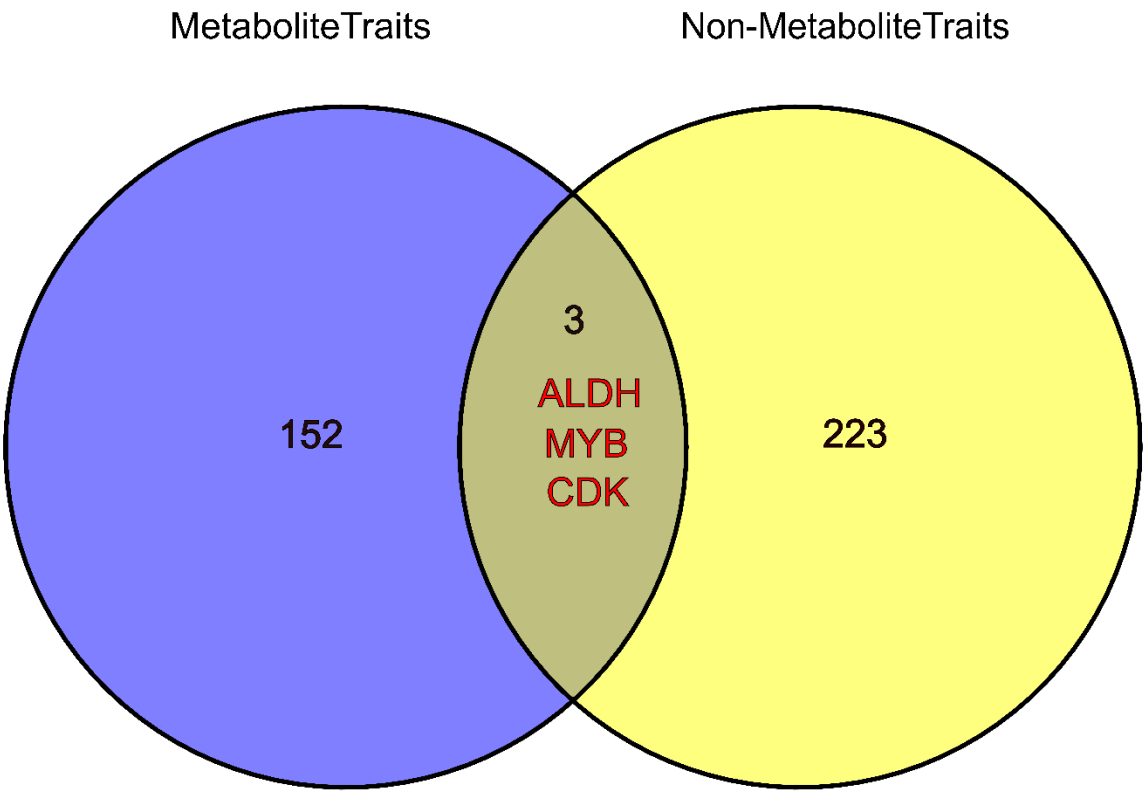
F.6: The figure displays results from an RMIPGWAS conducted using the FarmCPU algorithm, highlighting genes associated with metabolite and non-metabolite variation. The x-axis represents the physical position of a particular chromosome in maize, measured in megabases (MB), while the y-axis indicates RMIP values, reflecting the probability of SNP associations with the traits under study. Different colored dots represent SNPs associated with specific traits. These traits were chosen for their associations with both metabolite and non-metabolite traits. A horizontal dashed line at RMIP = 0.1 marks the significance threshold, indicating SNPs are significant in at least 10% of resampled datasets. The gene of interest is highlighted above the x-axis in red. Below the x-axis, an LD (Linkage Disequilibrium) plot is included for the same chromosome region, showing the LD levels of the gene and associated markers, with a color key indicating high LD in red and lower LD in yellow. This comprehensive figure illustrates the genetic associations and linkage disequilibrium relevant to the traits under study.



F.7: The figure displays results from an RMIPGWAS conducted using the FarmCPU algorithm, highlighting genes associated with metabolite and non-metabolite variation. The x-axis represents the physical position of a particular chromosome in maize, measured in megabases (MB), while the y-axis indicates RMIP values, reflecting the probability of SNP associations with the traits under study. Different colored dots represent SNPs associated with specific traits. These traits were chosen for their associations with both metabolite and non-metabolite traits. A horizontal dashed line at RMIP = 0.1 marks the significance threshold, indicating SNPs are significant in at least 10% of resampled datasets. The gene of interest is highlighted above the x-axis in red. Below the x-axis, an LD (Linkage Disequilibrium) plot is included for the same chromosome region, showing the LD levels of the gene and associated markers, with a color key indicating high LD in red and lower LD in yellow. This comprehensive figure illustrates the genetic associations and linkage disequilibrium relevant to the traits under study.



F.8: The figure displays results from an RMIPGWAS conducted using the FarmCPU algorithm, highlighting genes associated with metabolite and non-metabolite variation. The x-axis represents the physical position of a particular chromosome in maize, measured in megabases (MB), while the y-axis indicates RMIP values, reflecting the probability of SNP associations with the traits under study. Different colored dots represent SNPs associated with specific traits. These traits were chosen for their associations with both metabolite and non-metabolite traits. A horizontal dashed line at RMIP = 0.1 marks the significance threshold, indicating SNPs are significant in at least 10% of resampled datasets. The gene of interest is highlighted above the x-axis in red. Below the x-axis, an LD (Linkage Disequilibrium) plot is included for the same chromosome region, showing the LD levels of the gene and associated markers, with a color key indicating high LD in red and lower LD in yellow. This comprehensive figure illustrates the genetic associations and linkage disequilibrium relevant to the traits under study.



F.9: Venn Diagram shows the numbers of shared and uniquely identified genes associated with metabolite and non-metabolites variation using RMIPGWAS with shared genes highlighted mentioned in the middle circle.

Main Tables:

T.1: Genes Associated with Metabolite variation via Resampling Model Inclusion Probability Genome-Wide Association Study (RMIPGWAS) at RMIP significance thresholds of x ≥ 0.3

Metabolites	SNP	Distance from the marker	Gene_ID	Gene_Name	Gene description	RMIP
Galactonic_acid	chr6_81874806	38129	Zm00001eb270570	TBS	Theobromine synthase	0.64
Phosphoric_acid	chr8_133024421	32734	Zm00001eb354560	UCH	Ubiquitin carboxyl terminal hydrolase	0.61
Phosphoric_acid	chr3_28733643	761	Zm00001eb126310	NA	NA	0.51
Glyceric_acid	chr2_177080341	29285	Zm00001eb097540	NA	Endonuclease	0.47
Chlorogenic_acid	chr2_203757097	9983	Zm00001eb104230	PGP	Phosphoglycolate phosphatase	0.47
Shkimic_acid	chr4_212836690	32290	Zm00001eb201130	NA	NA	0.43
Trans_aconitic_acid	chr4_48138869	8808	Zm00001eb175400	NA	uridylyltransferase	0.4
Quinic_acid	chr3_116669684	53271	Zm00001eb135350	NA	NA	0.39
Sucrose	chr8_26033461	14868	Zm00001eb338670	NA	PWWP domain (PWWP)	0.38
D_Glucoe	chr3_216820299	19238	Zm00001eb157570	PKD	Protein kinase domain	0.38
Galactonic_acid	chr10_11900210	81008	Zm00001eb408510	NA	PPR repeat (PPR)	0.37
Raffinose	chr7_142676377	10013	Zm00001eb317720	NA	Proteasome subunit beta type-1	0.36
Fructose	chr5_149005008	NA	NA	NA	NA	0.36
Tyrosine	chr5_156613475	44418	Zm00001eb239880	NA	NA	0.36
Trans_aconitic_acid	chr6_161018577	34988	Zm00001eb289030	NA	A/G-specific adenine glycosylase	0.36
Chlorogenic_acid	chr7_1204204	22715	Zm00001eb298230	PER	Peroxidase	0.31
L_serine	chr2_173821786	53115	Zm00001eb096820	DMAO	Dimethylaniline monooxygenase	0.31

(*Candidate genes identified within 100 kb intervals centered around seven significant SNPs**)**

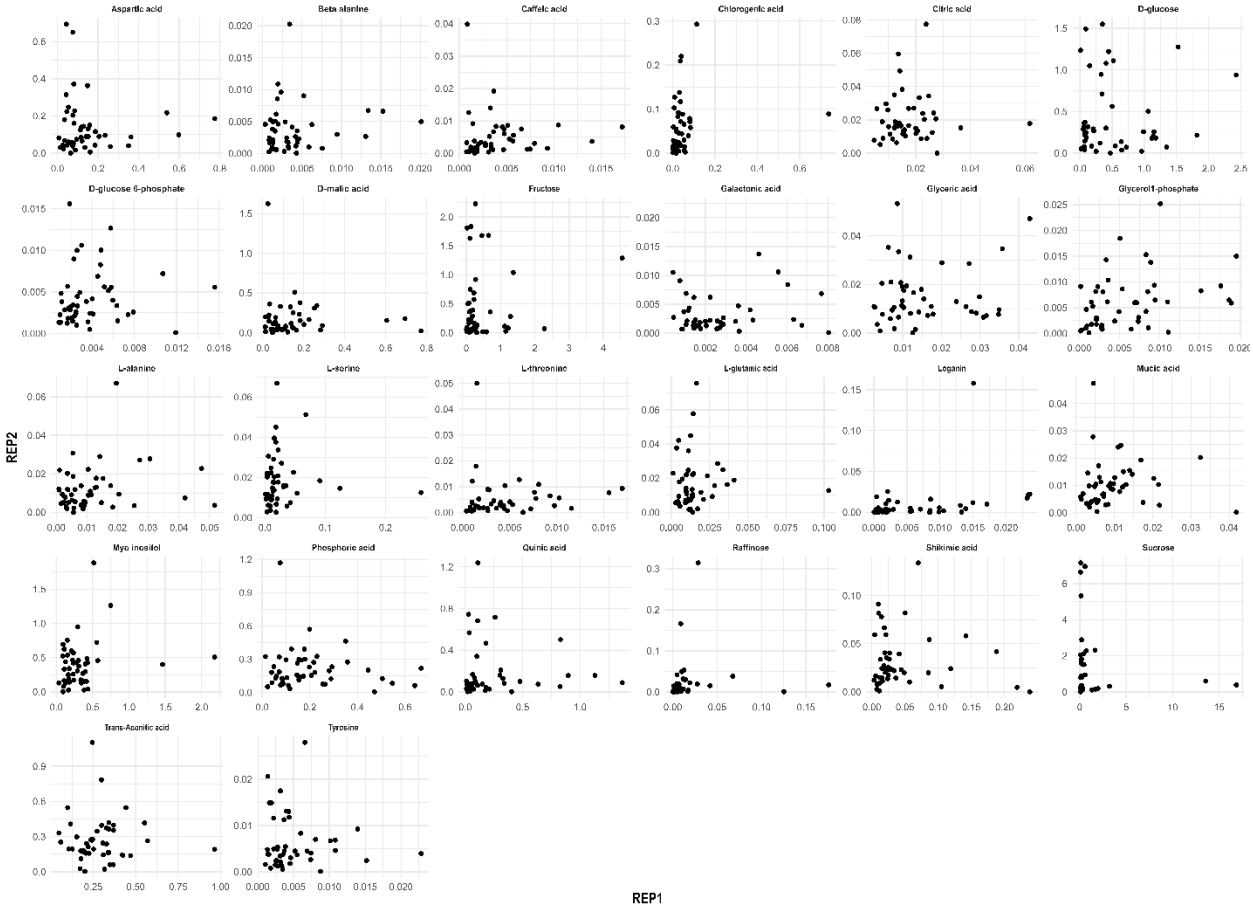
T.2: Genes Associated with Metabolite Variation via Transcriptome-wide Association Studies (TWAS)

Traits	SNP	Gene_ID	P_Value	R_square (pearson)	Gene_Name	Gene_Description
L_glutamic_acid	10_145065893	Zm00001eb431150	5.70E-07	0.050616345	CUEA	Cu(2+)-exporting ATPase
Quinic_acid	3_184155685	Zm00001eb147850	1.95E-07	0.070810384	MCO	MULTI-COPPER OXIDASE
Glycerol_1_Phosphate	6_16935828	Zm00001eb262130	2.28E-07	0.06172314	NA	NA
Glycerol_1_Phosphate	6_17825994	Zm00001eb262520	5.56E-08	0.063609407	NA	NA
Glycerol_1_Phosphate	6_165558187	Zm00001eb290790	1.21E-07	0.063337317	CUEA	Cu(2+)-exporting ATPase
Glycerol_1_Phosphate	10_145065893	Zm00001eb431150	6.35E-07	0.056087551	CUEA	Cu(2+)-exporting ATPase

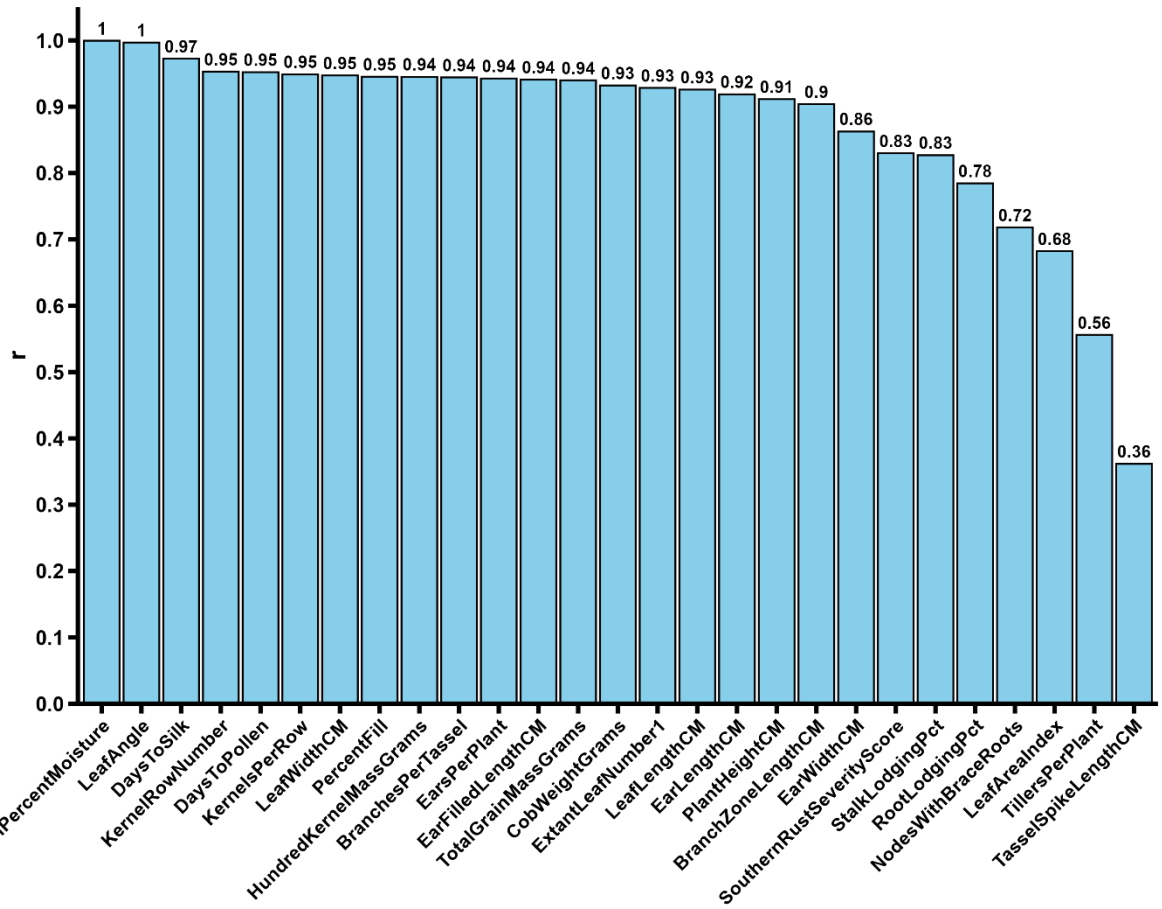
T.3: Genes Associated with both Metabolite and Non-Metabolite Variation via Resampling Model Inclusion Probability Genome-Wide Association Study (RMIPGWAS)

Metabolites	Non-Metabolites	Metabolite SNP	Non-Metabolite SNP	Metabolite RMIP	Non-Metabolite RMIP	Distance between the marker	Gene of interest	Gene description
L-serine	Percent_fill	chr6_113734505	ch6_113703113	0.11	0.23	31392	Zm00001eb277100	Aldehyde Dehydrogenase
Mucic acid	LV9	chr1_30626701	chr1_30562591	0.17	0.26	64110	Zm00001eb009750	Myb/SANT-Like DNA-Binding Domain
Chlorogenic acid	Branches per tassel	chr1_193639633	chr1_193546683	0.14	0.29	92950	Zm00001eb035350	Cyclin-Dependent Kinase

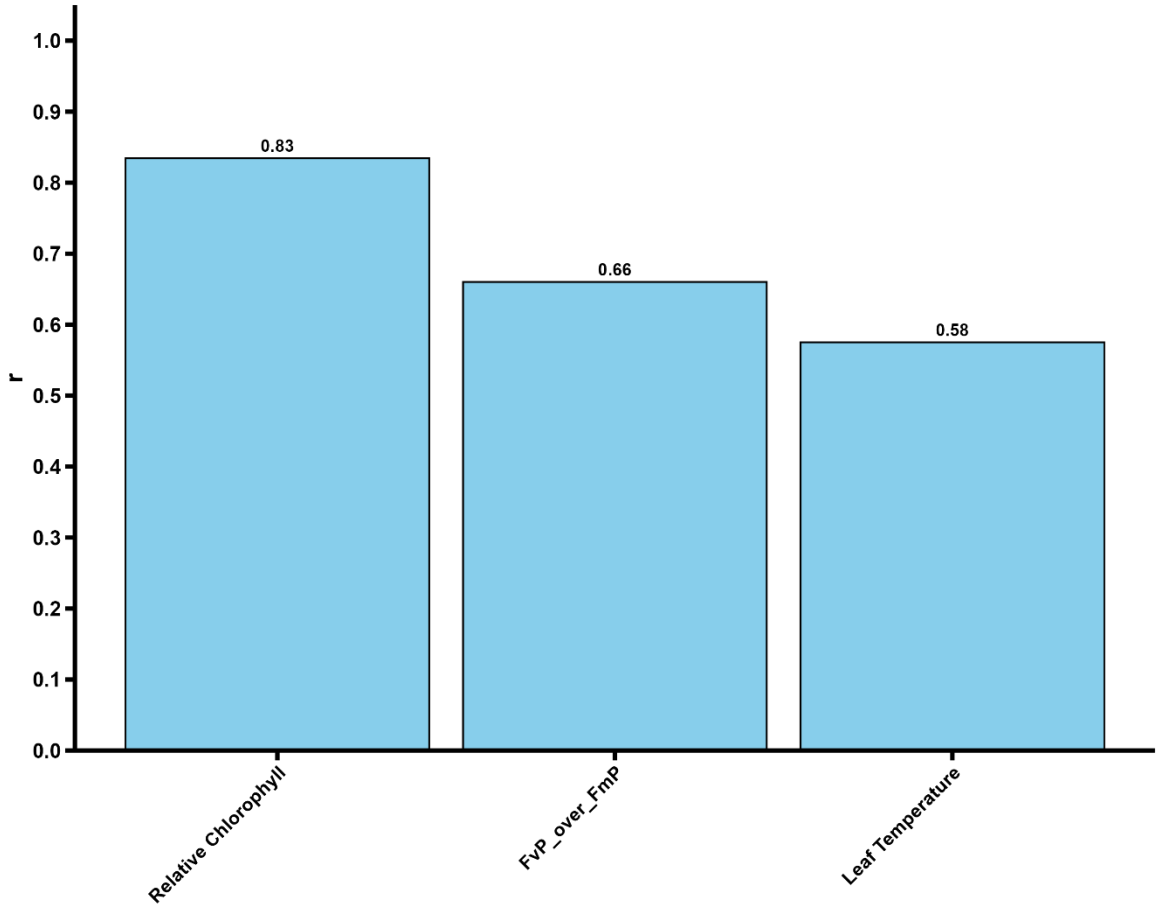
Appendix A: Supplementary Figures



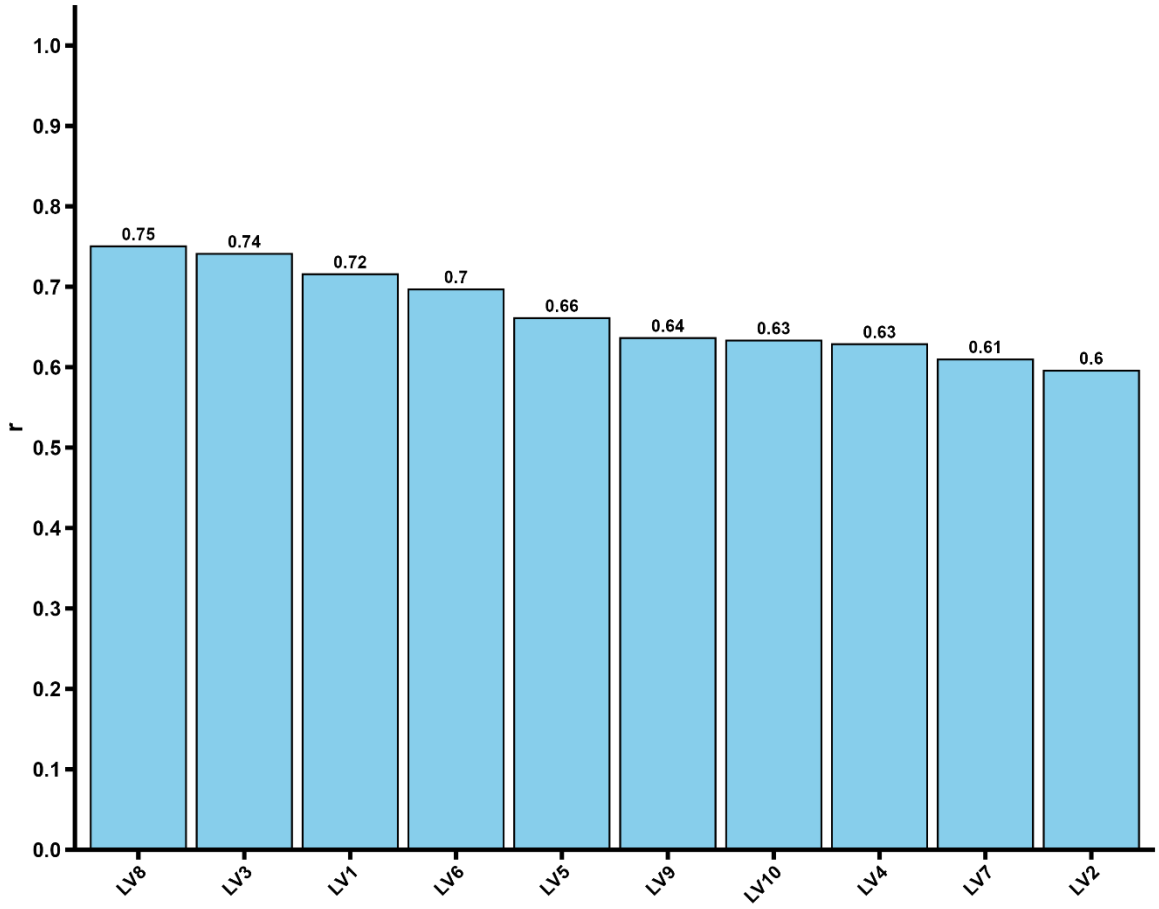
A.1: The figure displays a series of scatter plots, each representing the relationship between replicate measurements (REP1 and REP2) for different metabolites without removing the outliers. The data points on each plot correspond to the replicated genotypes from the metabolite study. The x-axis represents the measurements from the first replicate (REP1), and the y-axis represents the measurements from the second replicate (REP2).



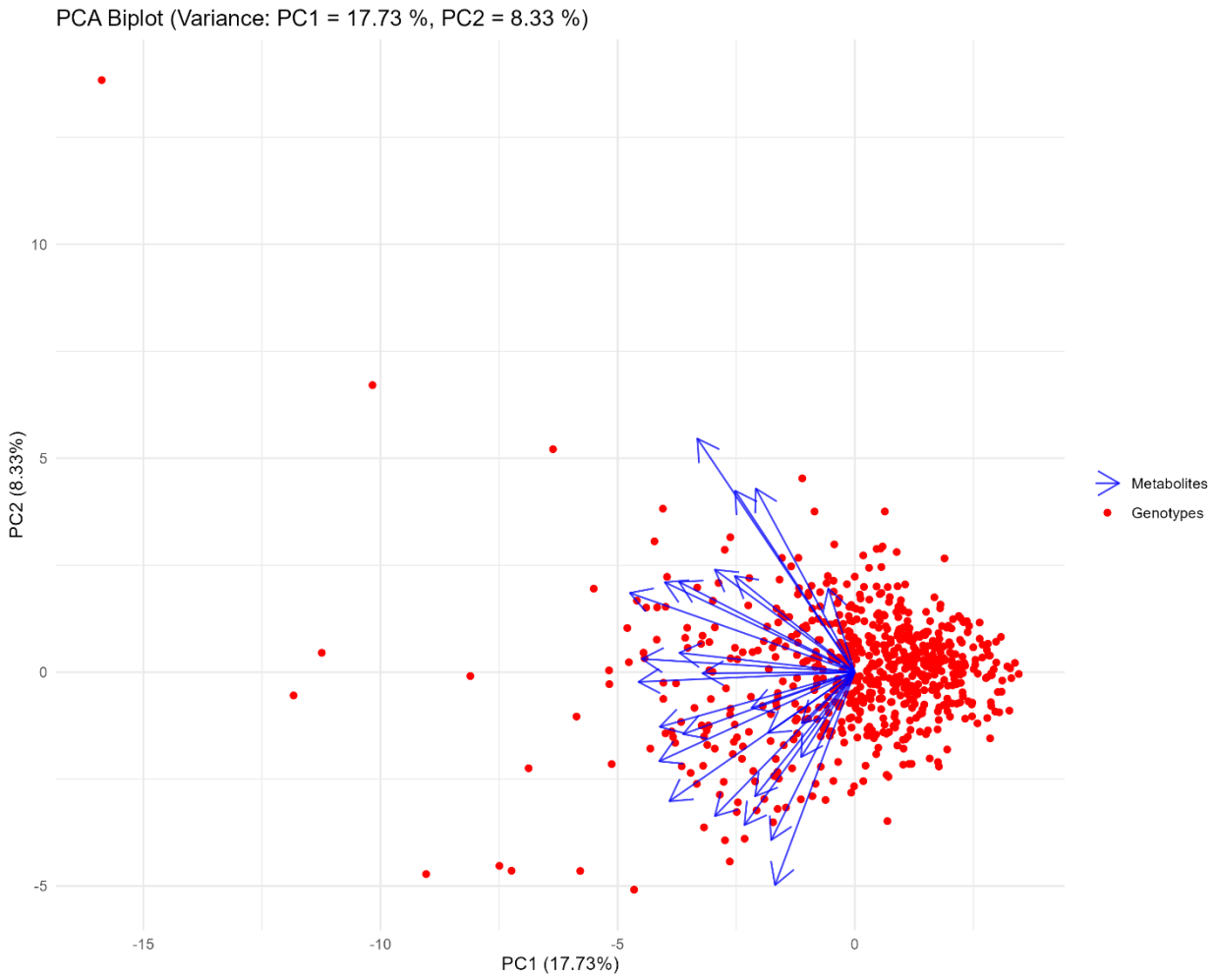
A.2: Repeatability (r): Bar graph quantifying the repeatability for various plant phenotypes, where a higher bar suggests a stronger genetic influence on the metabolite's expression, and a lower bar suggests a weaker genetic influence.



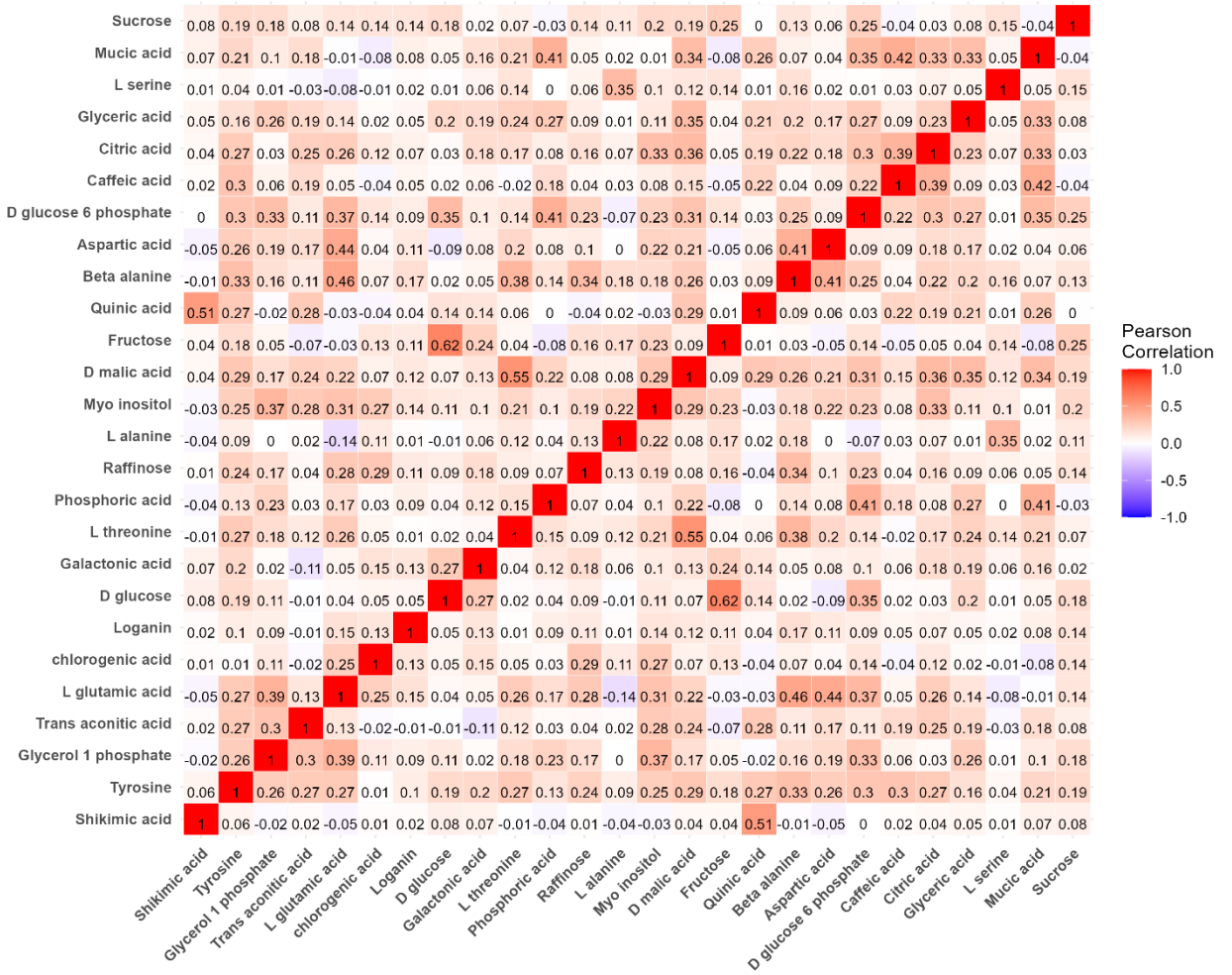
A.3: Repeatability (r): Bar graph quantifying the repeatability for various photosynthetic traits, where a higher bar suggests a stronger genetic influence on the metabolite's expression, and a lower bar suggests a weaker genetic influence.



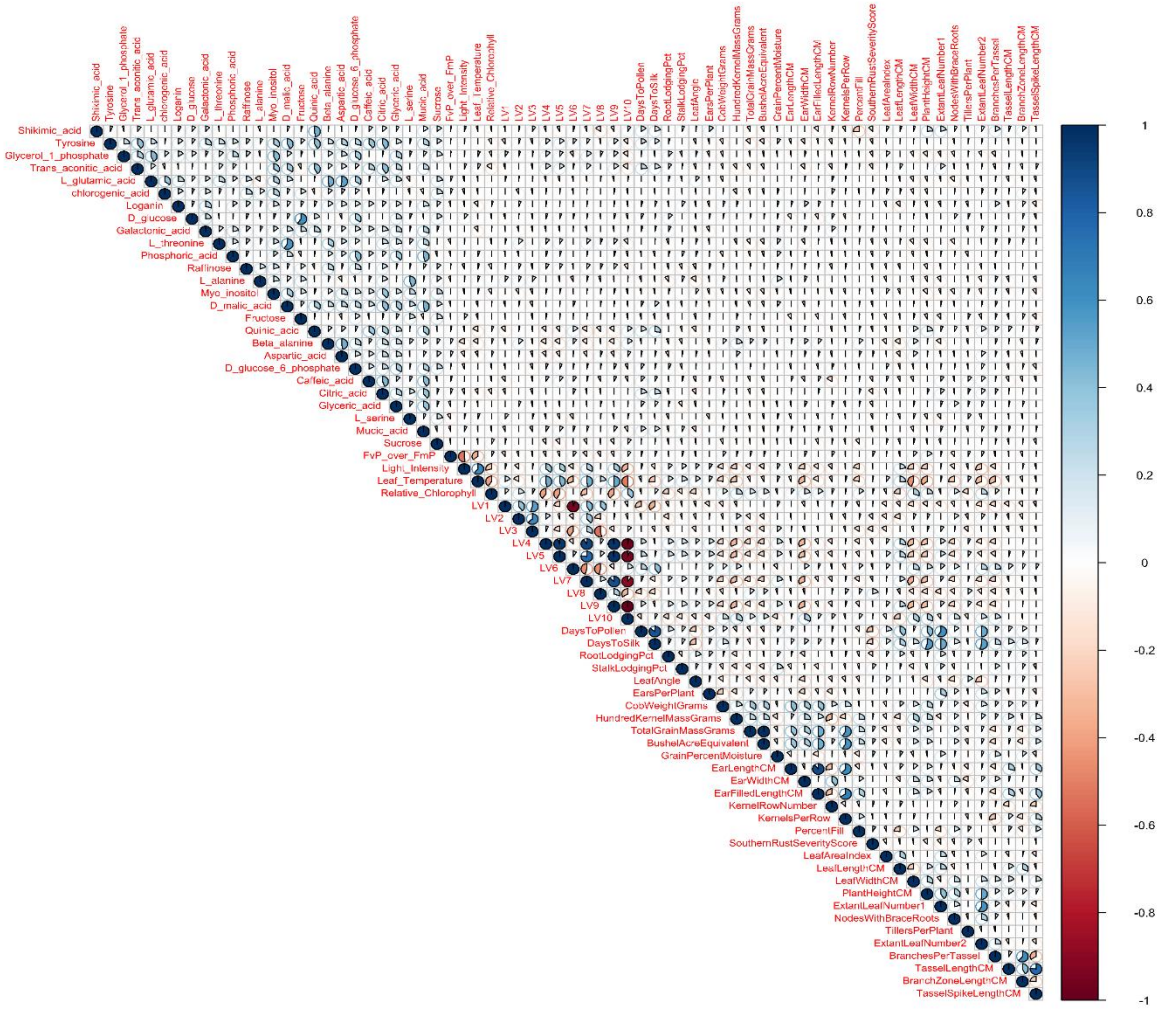
A.4: Repeatability (r): Bar graph quantifying the repeatability for various hyperspectral traits, where a higher bar suggests a stronger genetic influence on the metabolite's expression, and a lower bar suggests a weaker genetic influence.



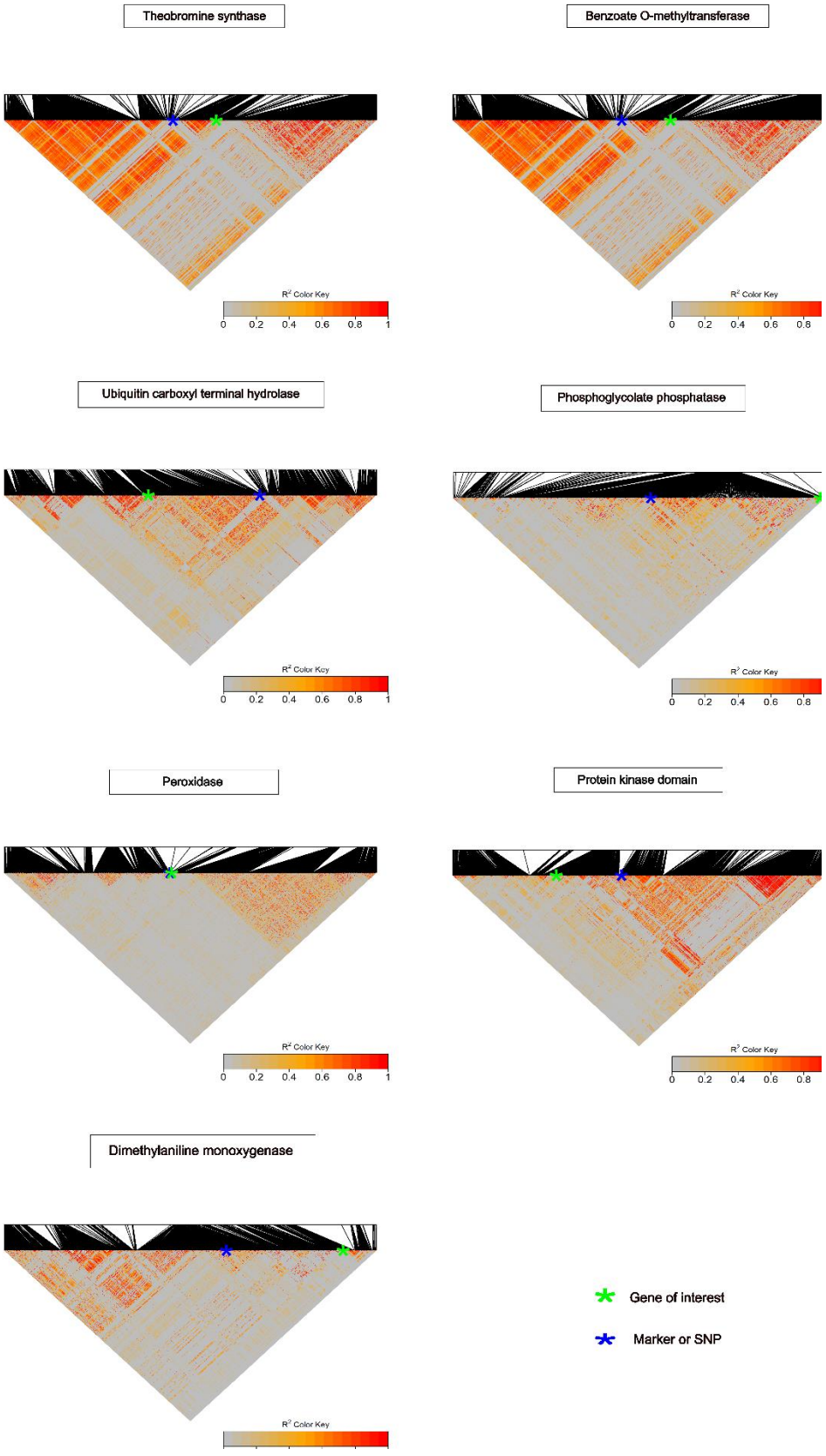
A.5: Principal Component Analysis (PCA): This scatter plot visualizes the principal component analysis for maize genotypes based on metabolite data. Each dot on the plot represents an individual maize genotype, positioned according to its scores on the first two principal components which explain a certain percentage of the variation in the dataset. The lines, or vectors, extend from the plot origin to labels indicating specific metabolites. These lines show the loadings of each metabolite, illustrating their influence on the principal components. The direction and length of the lines suggest how each metabolite correlates with the principal components and with each other.



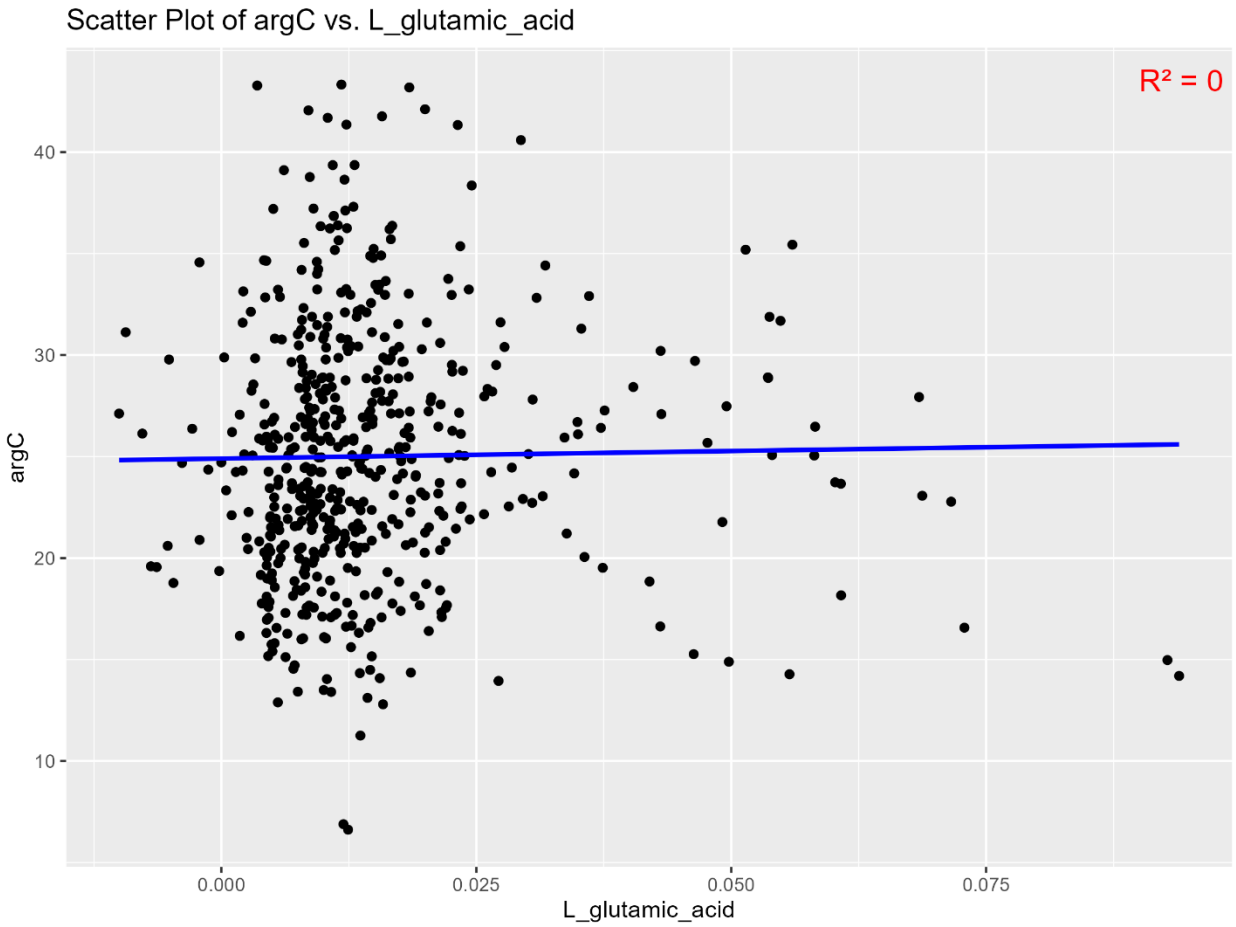
A.6: Correlation (r): Heatmap presenting the correlation coefficients between metabolites. Dark red squares indicate a strong positive correlation, while dark blue squares indicate a strong negative correlation, and lighter colors suggest weaker correlations. Coefficient values are displayed inside the squares for clarity.



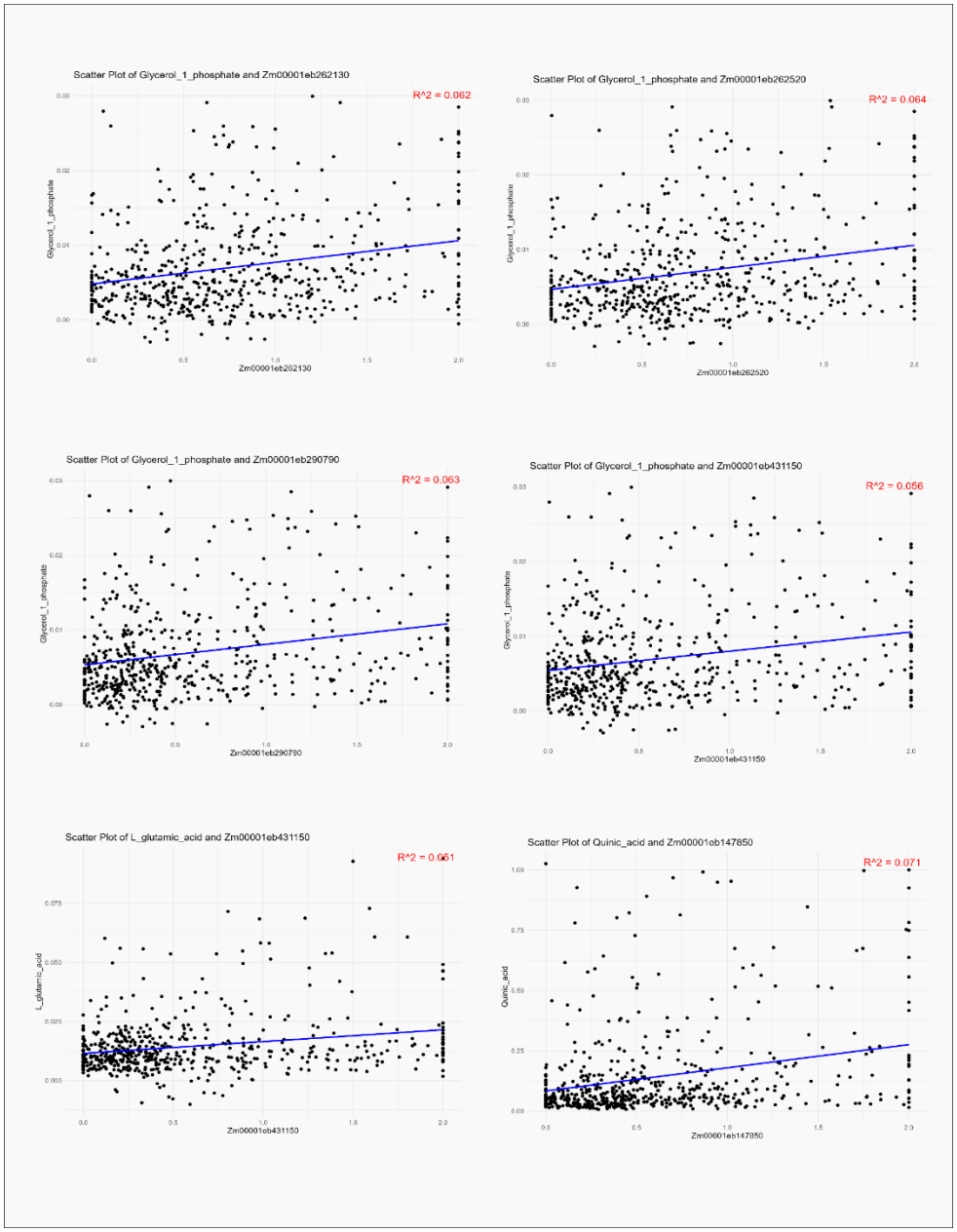
A.7: Correlation (r): Heat map presenting the correlation coefficients between the metabolites and non-metabolite traits. Dark red squares indicate a strong positive correlation, while dark blue squares indicate a strong negative correlation and lighter colors suggest weaker correlations.


Theobromine synthase
Benzoate O-methyltransferase
Ubiquitin carboxyl terminal hydrolase
Phosphoglycolate phosphatase
Peroxidase
Protein kinase domain
Dimethylaniline monoxygenase
R² Color Key
0 0.2 0.4 0.6 0.8 1
Gene of interest
Marker or SNP

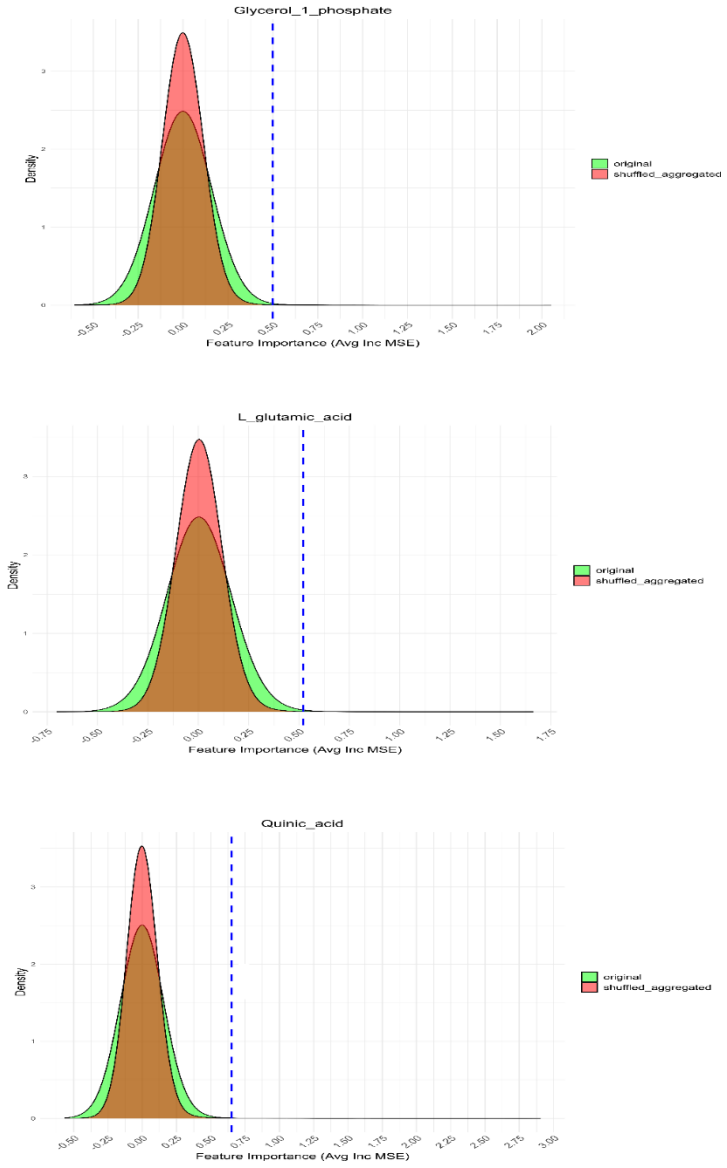
A.8: The figure displays Linkage Disequilibrium (LD) plots for seven key genes associated with metabolites identified from RMIPGWAS. Each subplot represents the LD structure surrounding one of the genes: Theobromine synthase, Benzoate O-methyltransferase, Ubiquitin carboxyl-terminal hydrolase, Phosphoglycolate phosphatase, Peroxidase, Protein kinase domain, and Dimethylaniline monooxygenase. The LD levels are depicted with a color gradient, where red indicates high LD (r^2 value close to 1) and yellow indicates lower LD. The green cross marks the gene of interest, while the blue cross marks a specific marker or SNP (Single Nucleotide Polymorphism). This comprehensive figure provides a visual representation of the genetic associations and LD patterns relevant to the traits under study.



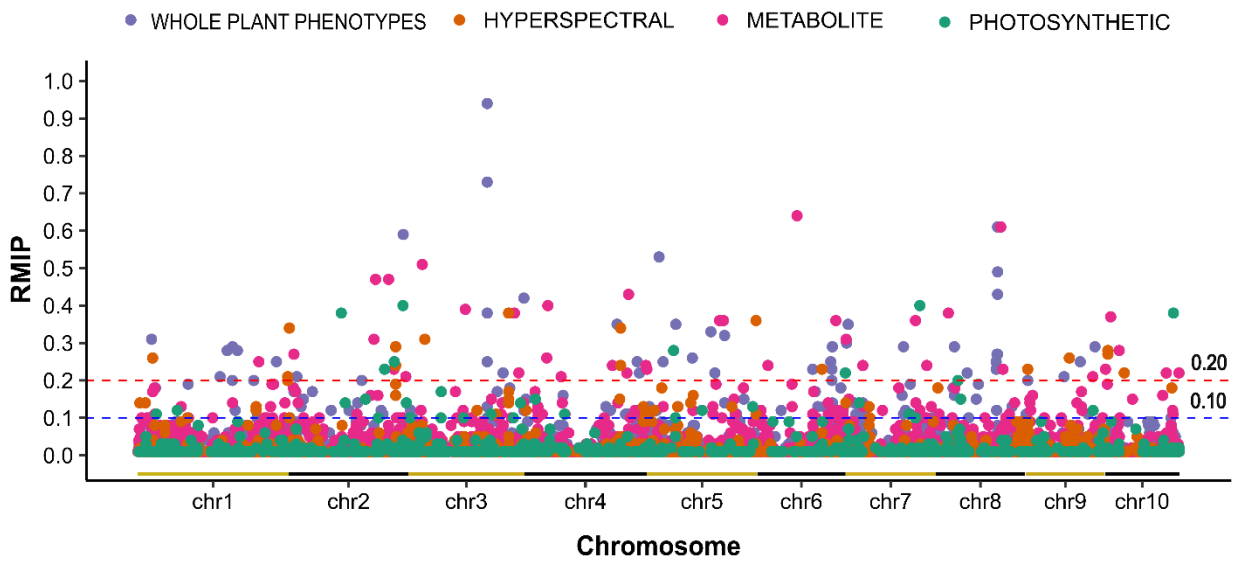
A.9: The figure presents a series of scatter plots illustrating the relationship between metabolite levels and gene expression for argC gene identified through Random Forest (RF).



A.10: The figure presents a series of scatter plots illustrating the relationship between metabolite levels and gene expression for significant genes identified through Transcriptome-Wide Association Studies (TWAS). Each plot compares the expression of a specific gene with the level of a corresponding metabolite, with pairings as follows: Glycerol_1_phosphate with Zm00001eb262130 (R^2 = 0.062), Zm00001eb262520 (R^2 = 0.064), Zm00001eb260790 (R^2 = 0.063), and Zm00001eb431150 (R^2 = 0.056); L_glutamic_acid with Zm00001eb431150 (R^2 = 0.061); and Quinic_acid with Zm00001eb147850 (R^2 = 0.071). In each scatter plot, the x-axis represents gene expression levels and the y-axis represents metabolite levels, with a blue line indicating the linear regression fit. The R^2 value, shown in red, quantifies the strength of the relationship, with higher values indicating stronger correlations. This figure effectively illustrates the associations between significant genes and their corresponding metabolites, highlighting the genetic influences on metabolite levels.



A.11: Combined Profile of Feature Importance Scores from Original and Shuffled Data: This panel showcases a feature importance plot derived from a Random Forest analysis of gene expression data. The x-axis represents feature importance scores where a score of 0 represents the average feature importance while the y-axis indicates the density of these scores. The green area indicates the distribution of feature importance scores based on original gene expression data, while the red area represents the importance scores from a shuffled dataset used as a control. The blue dotted lines indicate the established threshold for significant feature importance. Scores that surpass this blue threshold in the original data are considered biologically significant, as they exceed what would be expected by chance. The threshold is set at a level where features exceeding a score of certain feature importance score correspond to an FDR of approximately 0.05, highlighting genes that are considered to have a significant impact on metabolites.



A.12: Genes Associated with Metabolite and Non-metabolite Variation via Resampling Model Inclusion Probability Genome-Wide Association Study (RMIPGWAS). A result of a RMIPGWAS conducted using the FarmCPU algorithm. The x-axis represents maize chromosomes, while the y-axis indicates RMIP values, reflecting the probability of SNP associations with the traits under study. Colored dots distinguish SNPs associated with specific traits: purple for Whole Plant Phenotypes, orange for Hyperspectral traits, pink for Metabolite traits, and green for Photosynthetic traits. The plot includes two horizontal dashed lines marking RMIP significance thresholds: the upper red dashed line at 0.20 (indicating SNPs significant in at least 20% of resampled datasets) and the lower blue dashed line at 0.10 (indicating SNPs significant in at least 10% of resampled datasets). The physical positions between the chromosomes are marked with horizontal lines in two different colors. The physical positions between the chromosomes are marked with horizontal lines in two different colors.

APPENDIX B: Supplementary Tables

All supplementary tables are available on Figshare and can be accessed through the provided links (DOI).

Table B.1: Genes Associated with Both Metabolite and Non-Metabolite Variation via Resampling Model Inclusion Probability Genome-Wide Association Study (RMIPGWAS) at RMIP significance thresholds of x≥0.1x \geq 0.1x≥0.1
DOI: 10.6084/m9.figshare.26384044

Table B.2: Genes Associated with Three Specific Metabolites via Random Forest (RF)
DOI: 10.6084/m9.figshare.26384068

Table B.3: Analysis of False Discovery Rate (FDR) for Selecting Promising Genes from the Random Forest Findings
DOI: 10.6084/m9.figshare.26384071