TITLE: Integrative metabolome-genome analysis reveals the genetic architecture of metabolic diversity in sorghum grain

Deepti Nigam[1], Sarah Metwally[2], Songyue Shi[2], Priscilla Kolagani[1], Nasir Ali Khan[1], Ran Tian[1], Melinda Yerka[3], Fang Chen[2], Yinping Jiao[1]#

[1]Institute of Genomics for Crop Abiotic Stress Tolerance (IGCAST), Department of Plant and Soil Science, Texas Tech University, Lubbock, Texas, 79409 USA

[2]Center for Biotechnology and Genomics, Texas Tech University, Lubbock, Texas, 79409 USA

[3]Department of Agriculture, Veterinary and Rangeland Sciences, University of Nevada, Reno, Reno, Nevada 89557 USA

#For correspondence: Yinping Jiao: yijiao@ttu.edu

Abstract

Cereal grains are fundamental to global food security and bioenergy production, yet the genetic and molecular bases of grain metabolic diversity remain unresolved. Sorghum, a climate-resilient C4 grass with exceptional tolerance to heat and drought stress, provides an effective system for metabolic genomics. Here, we integrated large-scale untargeted metabolomic profiling, population genomics, and machine learning to systematically dissect grain metabolic diversity and its genetic determinants in sorghum. Untargeted metabolomic analysis of mature grains from the Sorghum Association Panel detected 4,877 metabolites exhibiting extensive quantitative variation across accessions, with pronounced diversity in amino acid and polyphenol biosynthetic pathways, indicating substantial standing variation relevant to nutritional improvement. Metabolite-based genome-wide association studies (mGWAS) identified ~4.15 million significant SNP-metabolite associations, significantly enriched in noncoding regulatory regions, consistent with a highly polygenic architecture of metabolic regulation. We further delineated 38 metabolite-gene clusters that reveal

coordinated genetic control of core metabolic pathways, and used machine learning to pinpoint key metabolites underlying grain color variation and to prioritize associated candidate genes, demonstrating predictive models that integrate genotype, metabolome, and phenotype. This work establishes the first population-scale atlas of sorghum grain metabolomic and genetic diversity and introduces the Sorghum Grain Metabolite Diversity Atlas (SorGMDA), an open resource integrating metabolomic and genomic variation. These resources enable comparative metabolic genomics across cereals and support systems-level breeding strategies for improving grain nutritional quality and climate resilience.

Introduction

Sorghum (*Sorghum* bicolor (L.) Moench) is a staple C4 monocotyledon crop in the *Poaceae* family. It exhibits high heat and drought tolerance, making it a cornerstone of agricultural systems in arid and semi-arid regions worldwide (Paterson et al. 2009; Ali et al. 2025). Sorghum is cultivated for human food, animal feed, and bioenergy production, but its full nutritional potential remains unrealized amid escalating global food demands (Liaqat et al. 2024). Sorghum is a gluten-free carbohydrate source rich in bioactive metabolites, including phenolic acids, flavonoids, and tannins (Tanwar et al. 2023). These molecules contribute to its high antioxidant capacity and potential health benefits, including anti-inflammatory (Burdette et al. 2010; Zhang et al. 2019) and anti-carcinogenic properties (Collins et al. 2024; Mazumder et al. 2024). These nutritional attributes uniquely position sorghum as a low-input crop that can be leveraged to address global food security and dietary health challenges, particularly in gluten-sensitive populations (Kaur et al. 2024).

Sorghum populations exhibit high genetic and phenotypic diversity (Boatwright et al. 2022) in grain bioactive metabolite concentrations (Kaufman et al. 2013; Habyarimana et al. 2019). Recent advances in sorghum breeding underscore the importance of grain bioactive metabolites (Girard and Awika 2018; Vanamala et al. 2018; Przybylska-Balcerek et al. 2019), which are increasingly recognized for their dual roles in promoting human health (Farrar et al. 2008) and enhancing crop resilience and nutritional quality (Mukherjee et al. 2024). For instance, phenolic metabolites not only support human health by reducing the risk of chronic diseases, including diabetes, obesity, cancer, and cardiovascular disease (Awika et al. 2009; Dykes et al. 2013; Arbex et al. 2018; Girard and Awika 2018; Chen et al. 2021b), but also play key roles in plant defense against biotic and abiotic stresses by detoxifying reactive oxygen species (ROS) (Cingoz and Gurel 2016).

Elucidating the genetic basis of metabolite variation is critical for improving sorghum grain quality through molecular breeding and genetic engineering (Zheng et al. 2011). Genome-wide association studies (GWAS) have emerged as a powerful tool for dissecting the genetic architecture of grain quality traits in sorghum, including starch composition (Boyles

et al. 2017; Chen et al. 2019), protein content (Li et al. 2018), and micronutrient diversity (Thakur et al. 2024). However, the resolution of these studies is often constrained by several factors, including the inability to capture intermediary metabolites, the polygenic nature of many grain quality phenotypes (Hamblin et al. 2007), and extensive linkage disequilibrium (LD) in elite temperate sorghum germplasm around loci for height, maturity, cytoplasmic male sterility systems (the most common basis of heterotic groups), and photoperiod insensitivity. For example, starch content, a key determinant of grain quality (Thitisaksakul et al. 2012), is typically measured as a single trait, yet its biosynthesis involves intricate biochemical pathways regulated by multiple gene networks, including those controlling amylose synthesis and starch branching enzymes (López-González et al. 2019). Similarly, protein content and micronutrient diversity are influenced by genetic background, polygenic interactions, and environmental factors, complicating the identification of causal variants (Ng'Uni et al. 2016; Duressa et al. 2018). The low resolution of GWAS limits the pinpointing of precise genetic markers, necessitating complementary approaches such as fine-mapping, haplotype-based analyses, and functional validation studies (McCarthy et al. 2008; Santure and Garant 2018).

To overcome these limitations, metabolomics-based genome-wide association studies (mGWAS) offer a powerful framework for linking metabolic traits to their genetic determinants by integrating high-resolution metabolite profiling with genome-wide genotyping (Luo 2015; Zhou et al. 2019; Chen et al. 2021a). The plant metabolome includes a diverse array of complex metabolites, including starches, proteins, micronutrients, and bioactive secondary metabolites such as phenolic acids and flavonoids, which collectively determine its nutritional value, end-use functionality, and health benefits. Recent advances in untargeted metabolome profiling utilizing high-performance liquid chromatography (HPLC) coupled with a Q Exactive Orbitrap mass spectrometer (Michalski et al. 2011) have revolutionized our ability to characterize complex metabolite mixtures with ultra-high resolution and mass accuracy (Eiler et al. 2017), enabling the effective separation and identification of structurally similar isomers (Siddique 2021; Qin et al. 2022). Unlike targeted

metabolomics, which focuses on quantifying a predefined set of known metabolites, untargeted approaches aim to capture a broad, unbiased spectrum of metabolites, including intermediates, unknown, and unexpected compounds, making them particularly valuable for exploratory studies and hypothesis generation (Ribbenstedt et al. 2018). The utility of machine learning (ML) tools in supporting mGWAS studies has been demonstrated in other major cereal crops, including *Zea mays* (maize), *Triticum aestivum* (wheat), *Oryza sativa* (rice), and *Hordeum vulgare* (barley). For instance, a multi-environment mGWAS of 983 metabolite features across 702 genotypes identified putative causal variants in five candidate genes linked to metabolic traits (Wen et al. 2014). In wheat, profiling of 805 metabolites across 182 accessions using >14,000 SNP markers revealed >1,000 significant associations and identified multiple candidate genes associated with variation in flavonoid metabolite substrates (Chen et al. 2020). In rice, mGWAS identified *Os02g57760,* a previously unknown nicotinic acid N-methyltransferase gene associated with grain size (Chen et al. 2016). Similarly, in barley, mGWAS identified associations between fructan composition and specific metabolic pathways and known fructan biosynthesis genes, further connecting metabolite profiles to grain quality traits (Matros et al. 2021). These studies underscore the utility of mGWAS in elucidating the genetic control of metabolism during seed development (Khan et al. 2024) and dissecting complex traits (Ramakrishna et al. 2018). Notably, in foxtail millet, the integration of GWAS, mGWAS, and microbiome-wide association studies (MWAS) revealed genotype-dependent effects on root-associated microbial communities and metabolite profiles, which could help plant breeders select for desirable crop-microbiome interactions based on host genetics (Wang et al. 2022). While mGWAS has provided significant insights into the genetic basis of nutritionally important metabolites in rice and maize (Zhao et al. 2024; Zhu et al. 2024), the genetic basis of key grain metabolites, such as flavonoids and tannins, remains poorly characterized, thereby limiting targeted improvement of nutritional quality. Addressing this gap requires high-resolution mapping of grain metabolites in genetically diverse sorghum germplasm.

The Sorghum Association Panel (SAP) includes genetically diverse accessions (Boatwright et al. 2022) with wide variation in grain quality traits (Boyles et al. 2017), making it an ideal resource for genetic studies. Recent advancements in deep whole-genome sequencing of the SAP have generated a high-density map of genetic variants, providing a robust foundation for high-resolution association mapping (Boatwright et al. 2022). In this study, we explored the metabolic diversity of sorghum grains in 266 SAP accessions using untargeted ultra-performance liquid chromatography-mass spectrometry (UPLC-MS). Our mGWAS identified 4.15 million genetic variants significantly associated with 2,206 metabolites, highlighting the complex genetic architecture underlying sorghum grain metabolism. These findings demonstrate the power of mGWAS to unravel genotype–metabolite relationships at scale in sorghum grain and provide a foundation for metabolite-informed breeding strategies aimed at enhancing grain quality and nutritional value in this staple crop.

RESULTS

Metabolite Annotation and Quantitative Assessment in Sorghum Grain

To capture the metabolic diversity of sorghum grain, we performed untargeted metabolomic profiling on mature grains from 266 diverse accessions of the Sorghum Association Panel (SAP). Metabolite annotation confidence was supported by high-quality MS/MS spectral matches (mass accuracy ≤ 5 ppm and spectral similarity score ≥ 0.7) against multiple databases, ensuring reliable metabolite identification for downstream genetic analyses (Supplemental Fig. S1). In total, 4,877 metabolites were identified, including 3,527 detected in positive-mode electrospray ionization (ESI+) and 1,350 in negative-mode electrospray ionization (ESI-). Among these metabolites, 89.6% were successfully annotated, 52.8% were assigned known biological functions, and 36.7% were mapped to established metabolic pathways (Supplemental Fig. S2A; Supplemental Table S1).

To assess metabolite abundance variation across the SAP population, coefficients of variation (CVs) were calculated from normalized intensity values after batch correction and

median normalization (Supplemental Fig. S2B, Supplemental Table S2). A total of 51.11% of metabolites exhibited high variability among genotypes (CV > 80%), reflecting substantial genetic diversity in metabolite accumulation, while an additional 23.17% and 16.38% of metabolites displayed intermediate (CV = 60-80%) and moderate (CV = 40-60%) variability, respectively, and only 9.34% showed low variability (CV < 40%). The prevalence of highly variable metabolites supports their suitability as quantitative traits for mGWAS.

Principal Component Analysis (PCA) further highlighted extensive metabolomic diversity within the SAP population, with the first two principal components explaining 42% and 38% of the total variance, respectively (Supplemental Fig. S3A). This indicates that much of the variation in sorghum grain metabolite profiles can be summarized by a small number of principal components. Collectively, these results demonstrate that the untargeted metabolomics pipeline provides robust, quantitative metabolite traits suitable for downstream mGWAS and integrative genetic analyses.

Landscape of metabolite diversity in sorghum grain

Identified metabolites were classified into distinct biochemical categories, with carbohydrates (21%), benzenoids (17.5%), organo-heterocyclic metabolites (10.2%), fatty acyls (9.5%), and nucleic acids (9.3%) representing the five most abundant classes (Supplemental Fig. S3B). Comparative profiling across the SAP accessions revealed substantial diversity in metabolite abundance patterns (Supplemental Fig. S3C), with subsets of accessions exhibiting similar overall metabolite profiles. When metabolites were grouped by biochemical class, compounds such as amino acids, lipids, and polyphenols showed correlated abundance patterns across accessions, reflecting class-specific variation within each class (Supplemental Fig. S3D).

The quantitative reliability of the untargeted metabolomics was validated by targeted LC-MS/MS quantification of 10 flavonoids and phenylpropanoid-derived compounds across 12 sorghum lines (36 samples, three biological replicates per line; Supplemental Fig. S2C, Supplemental Table S3). The analyzed metabolites included luteolinidin, apigeninidin,

7-methoxyapigenidin, luteolinidin-5-O-glucoside and apiforol, which contribute to sorghum pericarp coloration (Nip and Burns 1969; Wu et al. 2016), and additional flavonoids and phenylpropanoid derivatives such as naringenin, eriodictyol, luteolin, apigenin and 5,7-dihydroxycoumarin, which participate in broader defense-related and antioxidant pathways (Pontieri et al. 2021). Comparison of metabolite abundances obtained from targeted and untargeted approaches revealed strong concordance (Supplemental Fig. S2C, Supplemental Table S3). Pearson correlation coefficients ranged from 0.762 to 0.996, confirming the robustness of relative and quantitative metabolite measurements.

To quantify metabolic diversity in a sorghum population, we used the combined Shannon and Simpson diversity indices, widely used metrics that capture both richness and evenness within the metabolite landscape (Buzas and Hayek 2005; Halliday and Rohr 2019). A total of 36% of the detected metabolites showed high diversity (coefficient of variation, CV > 80%) across genotypes, and these metabolites also displayed high heritability, with estimates reaching up to 1 (Fig.1A, Supplemental Table S2). A 99% confidence interval (CI) around the diversity estimates supported the reliability of the observed patterns of variation in metabolite levels across genotypes. Amino acids, carbohydrates, nucleotides, polyphenols, polyketides, terpenoids, and organic acids each exhibited distinct distributions of diversity and median heritability values, indicating class-specific patterns of variation and genetic control (Fig. 1B).

Metabolites were further classified into high-, medium-, and low-diversity index groups across genotypes (Supplemental Table S2). The most diverse metabolites (top 5%) across genotypes were enriched in biosynthetic pathways such as alanine, aspartate, and glutamate metabolism, which showed strong statistical significance (FDR = 1×10^{-10}; $-\log_{10}$[FDR] = 10) (Fig.1C). Additional highly variable metabolites mapped to lysine metabolism (FDR = 1×10^{-7}) and the pentose phosphate pathway (FDR = 3×10^{-6}). Medium-diversity metabolites were linked to vitamin B2 metabolism (FDR = 0.01), lysine degradation (FDR = 0.003), and flavonoid biosynthesis (FDR = 0.005). In contrast, the least variable metabolites across genotypes were associated with glycerophospholipid metabolism, thiamine metabolism, and folate biosynthesis (Fig.1C).

Metabolites exhibiting correlated accumulation patterns are often co-regulated within shared biosynthetic pathways, suggesting coordinated metabolic control (Sawada et al. 2009; Fukushima et al. 2011). To identify such co-regulated molecular clusters, we constructed correlation-based networks linking signals (nodes) with similar accumulation profiles at the $P \leq 0.01$ (Fig.2). In total, 3,845 metabolite signals were annotated using MS1-based matching against database standards, while an additional 660 metabolites were characterized by comparing MS/MS spectra with previously recorded entries in external databases (Fig.2, Supplemental Table S1). This approach enabled tentative structural identification based on exact mass, retention time (RT), and database matches (Supplemental Table S1). The resulting network revealed distinct clusters associated with both primary (e.g., carbohydrates, organic acids, nucleic acids) and secondary metabolism (e.g., flavonoids, benzenoids, polyketides) (Fig.2). For instance, Cluster I (flavonoids) contained metabolites such as luteolin (peak ID F121951, m/z = 286.0477) and apigenin (peak ID F6485, m/z = 270.05); MS/MS reference spectra confirmed their oxygen-containing phenolic flavone structures (Supplemental Table S1). Additional clusters were defined by their shared MS/MS spectral features in a similar manner. Notably, metabolites such as L-serine, L-tyrosine, and hydroxy-L-lysine co-clustered within a single organic acid-related group (Fig.2). This observation is noteworthy because these amino acids differ structurally and functionally, yet their co-clustering suggests that they may share key spectral features or structural subcomponents-such as hydroxyl or amine groups-that contribute to similar MS/MS fragmentation patterns. Moreover, their grouping may reflect coordinated regulation within amino acid metabolism or shared biochemical pathways, underscoring the potential of MS/MS-based clustering to reveal unexpected biochemical relationships.

Genetic variation associated with metabolome diversity in sorghum

To explore the genetic architecture underlying metabolic diversity, mGWAS was performed using 38.8 million high-quality SNPs (minor allele frequency > 0.05) obtained from whole-genome resequencing of the SAP panel (Boatwright et al. 2022). As positive controls, successful GWAS signals were recovered for known genes associated with two representative

metabolites, (-)-epigallocatechin ($C_{15}H_{14}O_7$) and D-(-)-fructose ($C_6H_{12}O_6$) demonstrating the robustness of our metabolome profiling and mGWAS framework (Supplemental Fig. S4, Supplemental Table S4). (-)-Epigallocatechin (EGC) is a flavan-3-ol that functions as an antioxidant in plants and has been implicated in defense-related responses, reactive oxygen species (ROS) scavenging, cell-wall strengthening, and inhibition of fungal growth in various plant species (Grace 2005). Although direct functional studies in sorghum remain limited, its role as a precursor of condensed tannins supports a potential contribution to grain protective mechanisms (Bröhan et al. 2011; Rao et al. 2018).

The mGWAS of EGC identified 47 significant SNPs ($P \leq 10^{-8}$), including genetic variants located ~14 kb downstream ($P = 4.85 \times 10^{-45}$), of the well-characterized *Tannin1* (*Tan1*, *Sobic.004G280800*) gene on Chromosome 4 (Supplemental Fig.4A and Supplemental Table S4). *Tan1* encodes a WD40-repeat transcription factor that regulates the expression of genes in the proanthocyanidin (condensed tannin) biosynthetic pathway, seed coat pigmentation, and pathogen defense (Wu et al. 2012). An additional signal was detected ~15 kb upstream of *TRANSPARENT TESTA 6* (*TT6*, *Sobic.003G418000*) on Chromosome 3 ($P = 1.60 \times 10^{-10}$), which encodes flavanone 3-hydroxylase (F3H), a 2-oxoglutarate-dependent Fe(II) oxygenase (Supplemental Fig.4A and Supplemental Table S4). F3H catalyzes the stereospecific hydroxylation of flavanones to dihydroflavonols, an essential step in flavonoid biosynthesis, as observed in *Arabidopsis thaliana* (Pelletier and Shirley 1996), rice (Lam et al. 2014) and maize (Falcone Ferreyra et al. 2015). As a second positive control, D-fructose, a simple sugar central to primary metabolism and energy storage, is a major component of the sorghum grain carbohydrate profile and yield (Almodares and Hadi 2009; Simeone et al. 2017). Our mGWAS identified 29 significant SNPs ($P \leq 10^{-8}$), including a major association peak on chromosome 4 located approximately 28 kb upstream of Starch Synthase IIb (*SSIIb*; Sobic.006G221000) ($P = 1.53 \times 10^{-18}$) (Supplemental Fig. S4B), a well-characterized enzyme involved in starch biosynthesis (Campbell et al. 2016).

A total of 2,206 metabolites detected in more than 95% of the 266 SAP accessions and exhibiting a normal distribution (Shapiro-Wilk test, $P > 0.05$; within ±3 standard

deviations) were retained for mGWAS analysis (Supplemental Fig. S5A and B). Using a significance threshold of $P \leq 1 \times 10^{-8}$, we identified 4,151,451 unique SNPs, of which 2,221,701 showed robust associations (resample model inclusion probability [RMIP] ≥ 0.05 (Supplemental Table S5). This measure reflects how consistently a SNP is selected across multiple resampling iterations. As a result, SNPs with higher values are considered more reliable in their association.

Among the SNPs identified through mGWAS, many were located near known biosynthetic enzymes, while others represented novel associations, highlighting the utility of mGWAS for uncovering previously unrecognized genetic mechanisms. For example, lysine ranks among the top 5% most diverse metabolites within the SAP (Fig. 1C). The mGWAS of lysine ($C_6H_{14}N_2O_2$) revealed 77 significant SNPs ($P < 10^{-8}$) (Fig.3A, Supplemental Table S4), including four loci located within ±50 kb of genes encoding key enzymes in the lysine biosynthetic pathway. These included three aspartokinases (AK) (*Sobic.001G011700*, $P = 6.65 \times 10^{-12}$, *Sobic.003G412100*, $P = 1.93 \times 10^{-9}$ and *Sobic.002G155100*, $P = 4.49 \times 10^{-9}$), dihydrodipicolinate synthase (DHDPS, *Sobic.006G186600*, $P = 3.00 \times 10^{-9}$), and lysine decarboxylase (LDC, *Sobic.001G127700,* $P = 3.93 \times 10^{-9}$), each corresponding to distinct steps in the pathway (Fig.3B). Three of these enzymes revealed specific haplotypes at their respective loci: a CC-to-TT transition on Chromosome 3 for AKIII, TT-to-GG on Chromosome 6 for DHDPS, and CC-to-GG on Chromosome 1 for LDC (Fig.3C). These haplotypes were associated with significant differences in lysine accumulation, indicating that allelic variation at these loci contributes to the natural variation in lysine content observed across the SAP. AK catalyzes the first committed step from L-aspartate, DHDPS catalyzes a key branch-point reaction in the diaminopimelate pathway, and LDC converts L-lysine to cadaverine, linking primary lysine metabolism to alkaloid biosynthesis (Zhu and Galili 2003; Jander and Joshi 2009; Galili et al. 2016; Bunsupa et al. 2017). In addition to the known genes, mGWAS also identified genetic variants located in regions distant from any previously characterized gene involved in lysine biosynthesis. For example, we identified a SNP ~40 kb upstream of PECTATE LYASE 1-RELATED (*PL-1*; *Sobic.006G014400*) on Chromosome 6

($P = 3.00 \times 10^{-9}$) (Fig.3A, Supplemental Table S4). *PL-1* is known for its role in cell wall modification and fruit softening (Xu et al. 2022), suggesting a potential indirect link to lysine accumulation through modulation of cell wall-associated metabolic processes.

Genomic hotspots for metabolite variation in sorghum

To identify genomic regions enriched for metabolite-associated variation, we compared SNPs from two datasets: the input set of genome-wide SNPs in the SAP panel and the significant mGWAS SNPs. Comparison between the input SNPs (n = 38,873,896) and mGWAS SNPs (n = 4,151,451) revealed significant enrichment of variants in non-coding regulatory regions, particularly within the 5′ untranslated region (5′-UTR) and promoter-proximal upstream regions (Fig.4A and B; Supplemental Table S6). Fold-enrichment analysis using chi-square (χ^2) tests indicated significant overrepresentation of missense (7.1-fold; $P = 6.32 \times 10^{-24}$), synonymous (6.6-fold; $P = 8.61 \times 10^{-15}$), and UTR variants (~2-fold; 5′-UTR: $P = 7.27 \times 10^{-4}$, 3′-UTR: $P = 1.07 \times 10^{-10}$) (Fig.4C; Supplemental Table S6). Modest enrichment was also observed for upstream, downstream, and intronic variants (< 2-fold), whereas intergenic variants were significantly reduced (0.50-fold; $\log_2$FC = -0.99). To investigate the spatial distribution of non-genic metabolite-associated variants, the distance from each significant non-genic mGWAS SNP to its nearest annotated gene was quantified (Fig. 4D; Supplemental Table S6). Significant non-genic SNPs showed a strong enrichment near genes relative to the genomic background, with a major peak within ~1 kb of the nearest gene, consistent with promoter-proximal regulatory regions. A secondary enrichment peak was observed at ~100 kb from genes, a distance consistent with many distal regulatory elements, such as enhancers that act through long-range chromatin looping. By contrast, SNPs located far from genes (>1 Mb) were depleted compared with background non-genic variants (Fig. 4D; Supplemental Table S6). Together, these patterns indicate that metabolite-associated variants preferentially localize to genic and nearby regulatory regions, supporting a model in which metabolic diversity is primarily shaped by functionally relevant loci rather than by distant non-coding regions lacking known regulatory activity

Metabolic gene cluster variants underlying specialized metabolism and nutrients in sorghum grains

Given the strong enrichment of metabolite-associated variants within gene-dense and regulatory regions, we next examined whether these variants preferentially localize to higher-order genomic features involved in metabolic regulation, particularly metabolic gene clusters (MGCs). MGCs comprise physically co-located genes that encode enzymes functioning in the same, co-regulated metabolic pathway, and they play key roles in determining the nutritional and biochemical composition of sorghum grains (Nützmann et al. 2016; Liang et al. 2021). Across the sorghum genome, we identified 38 MGCs distributed across 10 chromosomes, collectively harboring 434 biosynthetic genes (Fig.5A and Supplemental Fig. S6, Supplemental Table S7). These clusters belong to diverse metabolic classes, including polyketide, lignin, lignan-polyketide, lignin-saccharide, lignin-terpene, amino acids, saccharide-polyketide, saccharide-terpene, alkaloid, and terpene. Notably, these clusters were located within intergenic distances ranging from 26.95 kb to 806.04 kb, and Chromosome 5 harbors the highest number of gene clusters (eight in total), including five polyphenol, one amino acid, and one carbohydrate metabolic cluster, suggesting that this chromosome region serves as a hub for secondary metabolic pathways (Fig.5A; Supplemental Table S6).

The mGWAS identified 85 associated SNPs within ±50 kb of biosynthetic genes in 35 of the 38 MGCs identified above (Supplemental Fig. S6; Supplemental Table S7). We further focused on SNPs associated with lignin content variation within MGCs.Lignin plays a critical role in seed development (Frei 2013) and contributes to tolerance against both abiotic and biotic stresses (Li et al. 2022), including insect herbivory and pathogen infection (Dowd et al. 2016). A total of 20 genes within this lignin MGC, located within ±50 kb of these variants including methyltransferase loci and a nucleoporin-related gene, exhibit seed-specific expression and strong co-expression with other cluster genes involved in specialized metabolism, collectively forming a distinct functional network (Fig.5B). Within this network, two major metabolic modules were identified: one related to flavonoid biosynthesis (7 genes) and another encompassing stilbenoid, diarylheptanoid, and gingerol biosynthesis (10 genes),

representing a broader phenylpropanoid-derived pathway (Supplemental Table S8). A third, smaller module (3 genes) associated with glutathione metabolism was also detected; however, this module showed no correlation with the other two (Fig.5B). Expression profiling of these genes during seed development revealed three distinct clusters, indicating that they are specifically co-expressed in seed tissues (Fig.5C; Supplemental Table S8). The enrichment of biosynthetic pathway genes within this co-expression network, along with their similar expression patterns, suggests that genetic variation near methyltransferase loci may contribute to broader regulatory control of specialized metabolism during seed development.

Metabolic regulation of kernel color

Metabolomic diversity underlies phenotypic variation in complex end-use traits and enables detailed dissection of their biochemical bases (Carreno-Quintero et al. 2013; Karakas et al. 2025; Rai et al. 2025). Here, we use kernel color as a representative trait to investigate the metabolic regulation of a key seed phenotype in sorghum. Kernel color in sorghum and other cereals is a complex trait determined by the morphology of the pericarp and endosperm tissues, as well as by the accumulation of diverse classes of pigment-related metabolites, including flavonoids, e.g., anthocyanins, carotenoids, and phlobaphenes (Dykes et al. 2009; Su XiaoYu et al. 2017; Kathuria et al. 2024). Given the multifactorial and non-linear nature of this trait, machine learning approaches are particularly well-suited to identify complex relationships between metabolite profiles and phenotypic outcomes (Tiozon et al. 2023; Zhang et al. 2023b).

We applied machine learning to compare metabolomic profiles across 211 SAP accessions, grouped into four grain color categories: brown (77 lines), red (22 lines), yellow (10 lines), and white (102 lines) (Supplemental Table S9). Hierarchical clustering of metabolite abundances across kernel color groups revealed distinct metabolic signatures, including differences in the composition of 49 secondary metabolites (Fig. 6A, Supplemental Fig. S7A, Supplemental Table S10). These were predominantly polyphenols exhibiting clear patterns of accumulation that correspond to variation in grain color. PCA supported a key role

for polyphenols in driving kernel pigment diversity, with PC1 and PC2 explaining 43% and 37% of the variance among accessions, respectively (Supplemental Fig. S7B). Owing to their antioxidant properties, polyphenols also influence seed development, germination, growth, and overall quality (Salunkhe et al. 1983; Rhodes et al. 2014; Lu et al. 2025).

Feature selection analyses identified seven top-ranking polyphenols with high feature importance scores strongly associated with grain pigmentation: 3-coumaric acid, 6-hydroxy-7-methoxycoumarin, genistein, isorhamnetin-3-glucoside, isorhamnetin-3-rutinoside, kaempferol, and quercetin-3-(6"-malonyl)-glucoside (Fig.6B; Supplemental Table 1; Supplemental Table S11; Supplemental Fig. S7C).

ROC curve analysis evaluated the discriminatory power of these metabolites for grain color classification (Supplemental Fig. S8, Supplemental Table S11). Among them, isorhamnetin-3-glucoside (Bordiga et al. 2014; Sytar et al. 2019; Coklar and Akbulut 2021) and genistein (Bursać et al. 2017; Lee et al. 2017) showed the strongest statistical separation between four grain color groups based on metabolite profiles, consistent with their roles as flavonoid-derived pigments within the phenylpropanoid pathway that directly influence grain pigmentation. Moderate classification performance was observed for 3-coumaric acid (Noel et al. 2005), 6-hydroxy-7-methoxycoumarin (Kai et al. 2006) , and quercetin-3-(6"-malonyl)-glucoside (Ojwang et al. 2012), which are phenylpropanoid and flavonol intermediates that may contribute more broadly to secondary metabolism rather than specifically to grain color contrast. In contrast, kaempferol (Park et al. 2006) showed limited discriminatory potential, indicating that its abundance is either relatively similar across color classes or that it plays a more general antioxidant or structural role not tightly coupled to visible grain pigmentation. The mGWAS of these seven polyphenols identified 217 significant genetic variants ($P < 10^{-14}$) associated with grain color variation (Fig. 6C, Supplemental Table S12). A total of 214 genes were mapped within a 50-kb flanking region of significant mGWAS signals near genes previously implicated in pigment regulation, such as *Y1* (*Sobic.001G398100*, $P = 7.76 \times 10^{-21}$), *Tan2* (*Sobic.002G076600*, $P = 3.24\times 10^{-16}$), *TT2* (*Sobic.003G183800*, $P = 2.55\times 10^{-49}$), and *TT10* (*Sobic.004G236000*, $P = 3.57 \times$

10^{-50}) along with enzymes know to regulate the phenylpropanoid, flavonoid, and lignin biosynthetic pathways (Supplemental Fig. S6D, Supplemental Table S12).

The QQ plot showed a clear deviation from the expected null distribution, reflecting an enrichment of significant associations across all seven polyphenols analyzed (Supplemental Fig.7E). This pattern underscores the robustness of the mGWAS approach in elucidating the polygenic control underlying variation in grain color. The proximity of these genes to mGWAS signals suggests that genetic variation in regulatory or coding regions may modulate polyphenol accumulation, thereby influencing differential pigmentation among sorghum accessions.

To identify the key haplotypes associated with metabolite-driven grain color variation, we performed a joint haplotype analysis of the top 11 mGWAS signals of these seven compounds across the four grain color classes (Table 1). For each associated SNP, two haplotypes were defined: Hap1, derived from dark-colored (red and brown) grains, and Hap2, derived from light-colored (yellow and white) grains. Metabolite abundances differed significantly between Hap1 and Hap2, with Hap1 consistently associated with elevated polyphenol accumulation (Table 1). Hap1 at these 11 SNPs predominated in red (71-78%) and brown (9.8-22%) grains but occurred at much lower frequencies in yellow (4.6-10%) and white (4-8.5%) grains (Supplemental Fig. S9; $\chi^2 = 13.82$, df = 3, P = 0.00316). This pattern suggests that sorghum lines carrying a higher number of Hap1 alleles at these 11 SNPs, which are linked to lignin and flavonoid metabolism, tend to exhibit deeper grain coloration. To validate the role of these seven metabolites in discriminating among four grain color groups, we analyzed the haplotypes from 13 randomly selected sorghum lines, ranging from white to dark red grains. This analysis revealed a significant correlation between specific haplotypes, metabolite abundance, and grain color phenotypes (Fig. 6D and E), highlighting the power of metabolomics as a tool for dissecting complex traits.

Discussion

SorGMDA: an interactive sorghum grain metabolite and mGWAS database

The sorghum grain metabolome exhibits remarkable chemical diversity, with more than 4,800 detected features and 2,206 high-confidence metabolites quantified, covering 95% of the SAP population across 266 genetically diverse accessions (Supplemental Fig. S10). Over 50% of these metabolites displayed high genotypic variation, underscoring the strong genetic control of grain biochemical composition in sorghum. By integrating untargeted metabolomics with whole-genome resequencing and mGWAS, we identified millions of robust metabolite-QTLs and pinpointed causal candidate genes for several nutritionally and agronomically important pathways.

The power of mGWAS lies in its ability to move beyond traditional phenotypic selection toward direct selection on biochemical composition. In sorghum, this is particularly relevant for traits that are difficult or expensive to phenotype (e.g., antioxidant polyphenols, free lysine, condensed tannins) yet critically influence nutritional quality, food functionality, and pre-harvest sprouting resistance. Comparable successes in other crops- including maize vitamin E content (Lipka et al. 2013), rice 2-acetyl-1-pyrroline (2-AP) aroma (Li et al. 2024), tomato lycopene (Sauvage et al. 2014), and wheat phenolic acids (Chen et al. 2020)- illustrate that metabolite-guided breeding can rapidly deliver consumer- and climate-oriented varieties.

A major outcome of this work is SorGMDA, the first public, interactive atlas that links sorghum grain metabolite profiles, mGWAS results, and candidate genes. SorGMDA enables breeders and geneticists to (i) query metabolites of interest, (ii) identify accessions with extreme concentrations of beneficial compounds, (iii) retrieve diagnostic SNPs, and (iv) download ready-to-use marker sets for marker-assisted selection. For example, (–)-epigallocatechin can be explored in SorGMDA to identify high-accumulating lines and their associated candidate genes for breeding purposes (Supplemental Fig. S11). Increasing grain (–)-epigallocatechin content is valuable due to its antioxidant activity, contributions to seed

defense, and potential benefits for nutritional quality and human health, making it an important target for sorghum improvement (Awika and Rooney 2004; Mokra et al. 2022). This resource directly translates complex metabolomic and genomic data into practical breeding tools.

Despite these advances, challenges remain. Metabolite levels are highly sensitive to the environment, and our study was conducted on field-grown material from a single season and location, so future multi-environment trials will be essential to dissect G×E interactions and develop robust, environment-resilient markers. Second, although untargeted metabolomics provides extraordinary coverage, absolute quantification and full structural elucidation of all features remain resource-intensive.

By establishing SorGMDA and providing the first genome-to-metabolome map of sorghum grain, this study delivers an immediately actionable platform for nutritional enhancement and climate-resilient breeding, and the framework is readily transferable to other orphan and major crops where biochemical quality is a breeding objective. In an era of accelerating climatic stress and rising demand for nutrient-dense foods, integrating metabolomics with genomics offers a direct route to redesign crop composition for human and planetary health.

Post-GWAS gene function studies

Our metabolite-based GWAS (mGWAS) identified more than four million genetic variants associated with key metabolites, encompassing both well-characterized genes and a substantial number of previously unannotated loci. The recovery of known loci such as *Tannin1, TT6, and SSIIb* demonstrates the robustness and high quality of both the metabolomic and genetic datasets. Notably, significant variants were enriched in promoter-proximal and 5′ untranslated regions, highlighting the prominent role of cis-regulatory variation in shaping metabolite abundance. In addition, the identification of metabolic gene clusters containing co-expressed biosynthetic enzymes underscores the importance of physical clustering in the coordinated regulation of specialized metabolic pathways. The large

number of uncharacterized candidate genes uncovered by this study provides a valuable opportunity to expand our understanding of the biosynthetic networks governing sorghum grain composition. However, the functional characterization of these genes remains a long-term challenge that will require systematic, multifaceted investigation. To support community-wide efforts, we have made our mGWAS results openly available through the SorGMDA database, providing a foundation for downstream functional analyses.

Dissecting gene function will require integrating complementary approaches. Fine mapping of associated loci can refine candidate gene lists and identify causal variants. Subsequent multi-layered analyses, including transcriptomics, metabolomics, proteomics, and targeted biochemical assays, can elucidate the mechanistic roles of candidate genes within metabolic pathways. Ultimately, functional validation through genetic manipulation, such as gene knockouts or overexpression, will be essential to confirm causal relationships and assess their potential utility in crop improvement. Collectively, these post-GWAS studies advance our understanding of sorghum specialized metabolism and establish a framework for translating genetic discoveries into precision breeding and metabolic engineering strategies to improve grain quality and nutritional value.

Methods

Sample collection

The original SAP was obtained from the U.S. National Plant Germplasm System (NPGS) via the Germplasm Resources Information Network (GRIN). Plants were grown at the Texas Tech University Research Farm in Lubbock, Texas (33°35′52.9″N, 101°54′21.4″W; elevation 992 m). The site is characterized by a semi-arid climate, with an average annual rainfall of approximately 469 mm, and fine-loamy, mixed, superactive, thermic Aridic Paleustalf soils typical of the Amarillo series. Supplemental irrigation (~2.56 cm per week) was provided throughout the growing season. Panicles were covered with pollination bags prior to flowering to prevent outcrossing. For metabolite profiling, a subset of 266 SAP accessions was randomly selected based on their ability to reach physiological maturity within the local

growing season. Grain samples (~20 seeds per plant) were collected from three independent plants per accession to serve as biological replicates. Grains were freeze-dried for 12 h using a benchtop freeze dryer (Labconco FreeZone 2.5 Liter, -84 °C) connected to a vacuum pump (Labconco Model 117) to prevent degradation. To minimize heat generation during milling, tubes were frozen in liquid nitrogen before grinding. The dried grains were then finely ground in 2-mL microcentrifuge tubes using a bead-mill homogenizer.

Untargeted UHPLC-MS metabolite profiling and data processing

For metabolite extraction, 35 mg of each freeze-dried sample was transferred into 2-mL microcentrifuge tubes containing two ceramic zirconium oxide beads. Samples were immersed in a pre-cooled extraction solvent consisting of methyl tert-butyl ether (MTBE) and methanol (3:1, v/v), supplemented with internal standards: 100 µL of 1,2-diheptadecanoyl-sn-glycero-3-phosphocholine (for lipid analysis) and 50 µL each of corticosterone and ampicillin (for polar metabolite analysis). Samples were homogenized using a Precellys 24 bead mill at 5000 Hz for two 45-s intervals. Homogenates were incubated on an orbital rotator at 100 rpm for 45 minutes at room temperature to facilitate extraction. Phase separation was induced by adding a water/methanol mixture (3:1, v/v), followed by additional homogenization and centrifugation at 20,000 × *g* for 5 minutes at 4 °C. The resulting upper (lipid-containing) and lower (polar metabolite) phases were transferred into separate tubes, dried using a SpeedVac, or stored at -80 °C for future analysis. Prior to UHPLC-MS analysis, the lipid extracts were reconstituted in acetonitrile:2-propanol:water (65:30:5, v/v), and the polar extracts in UPLC-grade methanol:water (1:1, v/v). Both fractions were centrifuged to remove particulates and transferred into autosampler vials. To monitor instrument stability and analytical repeatability, a mixed quality control (QC) sample was prepared by pooling 20 µL aliquots from all samples across accessions and biological replicates.

Metabolite separation and detection were performed in both ESI positive and negative ion modes using a Waters 2777c UPLC system coupled with a Q Exactive HF high-resolution mass spectrometer (Thermo Fisher Scientific, USA) for metabolic profiling. The resulting

raw mass spectrometry (MS) data (.raw) files were converted to .mzXML format using MSConvert, with an MS-level filter applied (Chambers et al. 2012). Metabolite quantification was achieved by analyzing signal intensities using the XCMS and CAMERA data processing pipelines(Smith et al. 2006). A minimum sample threshold of three was set during the XCMS grouping phase to capture rare metabolites across this diverse sample set. Missing values were handled using median imputation and interpolation methods to preserve data integrity and ensure completeness. Mass features identified by XCMS-CAMERA were filtered based on retention times (60-630 seconds) and precise mass ($m/z < 0.5$ at the first decimal point), excluding naturally occurring isotopes. Although false annotations were frequent, MS adducts were retained to minimize the risk of discarding true metabolites. Feature quantification was adjusted for tissue fresh weight and normalized against the total ion count of each sample to account for technical variability. The resulting data were log-transformed to enable multivariate analyses. To further refine the dataset, interquartile range filtering and Pareto scaling were applied before conducting statistical analysis. All additional statistical assessments and visualizations were performed using the ggplot2 package in R. The metabolic trait (m-trait) data for the SAP were calculated as the average of three biological replicates measured using UPLC-MS. Specifically, for each metabolite m (where m=1,2,...,4,877) in each accession l (where l=1,2,...,266), the m-trait value $M_{m,l}$ was computed as:

$$M_{m,l} = \frac{1}{3}\left(M_{m,l,1} + M_{m,l,2} + M_{m,l,3}\right)$$

where $M_{m,l,1}$, $M_{m,l,2}$, $M_{m,l,3}$, are the normalized metabolite intensities from the three biological replicates.

Metabolite identification confidence and validation

Metabolite annotation was conducted using multiple complementary databases, including HMDB (Wishart et al. 2022), MassBank (Horai et al. 2010), ChemSpider (Pence and Williams 2010), and KEGG (Kanehisa and Goto 2000), enabling broad metabolite assignment across diverse chemical classes. Due to the large-scale, population-level nature

of this untargeted metabolomics study and the lack of a comprehensive set of authentic standards for sorghum grain metabolites, no MSI Level 1 identifications were performed. Instead, all metabolites were assigned putative identities corresponding to Metabolomics Standards Initiative (MSI) Levels 2-3 (Sansone et al. 2007; Fiehn et al. 2008; Spicer et al. 2017). MSI Level 2 assignments were supported by high-resolution MS/MS spectral similarity to reference databases, while MSI Level 3 annotations were based on precursor mass accuracy and diagnostic fragmentation patterns consistent with specific metabolite classes. Confidence in biologically important metabolites, such as proanthocyanidin monomers, flavonoids, and phenolic acids-was strengthened through multiple lines of evidence, including high MS/MS spectral similarity scores (> 0.85), high-resolution precursor mass accuracy (≤ 5 ppm), co-localization of mGWAS signals with previously characterized biosynthetic loci (e.g., *Tannin1*), and agreement with published metabolome datasets. Together, these criteria provide robust and reproducible metabolite annotation suitable for downstream genetic and biochemical analyses while maintaining broad, unbiased coverage of the sorghum grain metabolome.

Targeted quantification of sorghum flavonoids using parallel reaction monitoring

To quantify a group of flavonoids in sorghum seeds, polar extracts from 12 selected lines, each analyzed in three biological replicates, were subjected to targeted Parallel Reaction Monitoring (PRM) analysis using a Thermo Scientific Vanquish UHPLC system coupled to a Q Exactive HF high-resolution mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). Chromatographic separation was performed on a Waters Acquity UPLC HSS T3 column (1.7 µm, 2.1 × 100 mm). The mobile phases consisted of 0.1% formic acid in water (A) and 0.1% formic acid in methanol (B), delivered at a flow rate of 0.40 mL/min. The gradient program was as follows: 5% B initially; 5-50% B from 1-12 min; 50-98% B from 12-13 min; held at 98% B from 13-16 min; returned to 0.5% B from 16-17 min; and re-equilibrated at 0.5% B from 17-20 min. The injection volume was 5 µL. PRM data were acquired in both positive and negative ionization modes. Instrument parameters included a

resolution of 30,000, AGC target of 3×10^6, maximum injection time of 100 ms, and an isolation window of 1.0 m/z. Compound identification and quantification were based on accurate mass, retention time, and compound-specific precursor-product ion transitions. Method performance was assessed for calibration linearity, and relative quantification was performed using external calibration curves. Quality control samples were analyzed throughout the analytical sequence to monitor signal stability and instrument performance, ensuring data reliability across sample batches.

Phenotypic data normalization and heritability estimation

The 2,206 distinct metabolite traits were normalized using the bestNormalize R package (v1.4.3), which systematically evaluates a range of transformation techniques, including Lambert W × F, Box-Cox, Yeo-Johnson, and Ordered Quantile, to identify the method that best improves normality, as measured by the Pearson *P* test statistic. Given the dataset's high dimensionality, manual outlier curation was infeasible. Instead, a rules-based automated approach was employed to remove extreme values. After transformation, values exceeding ±1.5 times the interquartile range (IQR) from the first or third quartile of each trait distribution were considered outliers and treated as missing. Broad-sense heritability (H^2) was then estimated for each metabolite using replicate-level averages per genotype. Each genotype was represented by three biological replicates, and values beyond predefined biological thresholds were excluded prior to heritability estimation. A linear mixed model was fitted, treating genotype as a random effect, as follows:

$$Y_{\mathrm{ij}} = \mu + \delta_{\mathrm{i}} + \epsilon_{\mathrm{ij}}$$

where Y_{ij} represents the normalized metabolite value for the *j*-th replicate of the *i*-th genotype, μ is the overall mean, δ_i is the random effect associated with the *i*-th genotype, and ϵ_{ij} is the residual error term. Variance components (σ_G^2 for genotype and σ_R^2 for residual error) were extracted from the model, and broad-sense heritability was calculated as:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{n}\sigma_R^2}$$

where n is the number of biological replicates per genotype (here, $n = 3$).

Metabolite diversity analysis

Following normalization, metabolite intensities were scaled to ensure comparability across samples. Metabolites detected in fewer than 5% of the samples were excluded to reduce noise. For each sample, the relative abundance of each metabolite (p_i) was calculated by dividing its intensity by the total intensity of all metabolites in that sample. To assess metabolic diversity, we employed two complementary indices: the Shannon Diversity Index and the Simpson Diversity Index (DeJong 1975; Gorelick 2006). Metabolite diversity was quantified in two ways: (a) metabolite richness, defined as the total number of metabolites detected in each accession; and (b) abundance-weighted diversity, which accounts for both the number and relative abundances of metabolites across accessions. These indices were selected to capture different aspects of diversity. The Shannon method (Pallister et al. 2017; Menni et al. 2020) was used to emphasize richness and evenness across metabolites, while the Simpson method highlighted dominance (Bi et al. 2021; Salgado et al. 2023). The Shannon Diversity Index (*H'*) was computed as:

$$H' = -\sum p_i \cdot \ln p_i,$$

where pi represents the relative abundance of metabolite i, with higher *H'* values indicating greater diversity.

The Simpson Diversity Index (D) was calculated using the formula:

$$D = 1 - \sum p_i^2,$$

where *pi* represents the proportional abundance of the ith metabolite in an accession. This index gives more weight to abundant metabolites. Higher D values indicate that a few metabolites dominate the sample, while lower values suggest a more even distribution across metabolites. The index was computed for each accession, and the distribution of D values across all accessions was visualized using density plots.

Metabolite-based genome-wide association study (mGWAS)

Only the 38,873,876 SNPs from the 266 SAP accessions after filtering minor allele frequencies (≥ 0.05) were used for mGWAS (Boatwright et al. 2022). Metabolites were selected for GWAS if they were detected in at least 90% of the 266 phenotyped accessions and passed normality criteria. GWAS was independently performed on each of the 2,206 qualifying metabolites using a mixed linear model (MLM) (Zhang et al. 2010) with principal components (PCs) and the kinship matrix and one multi-locus model, Fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al. 2016). These models were implemented with the R packages GAPIT (Lipka et al. 2012) and rMVP (v1.0.1) (Yin et al. 2021). In the MLM, a kinship matrix was integrated and implemented in rMVP. FarmCPU was run with the first three PCs as covariates, and the kinship matrix was calculated internally by the algorithm as a random effect. FarmCPU was executed with maxLoop = 10 and the MVP option for bin selection using method.bin = "FaST-LMM" (Lippert et al. 2011). The variance component analysis method, vc.method, was set to "GEMMA" (Team et al. 2024).

Bonferroni corrections were applied based on the effective number of independent markers in each dataset, calculated using the genetic type I error calculator (GEC v0.2) (Li et al. 2012). A stringent genome-wide significance threshold of 1×10^{-5} was used, corresponding to the Bonferroni-corrected cutoff to control the family-wise error rate. Due to instability in individual FarmCPU runs, the model was rerun 100 times with 90% subsampling, and Resample Model Inclusion Probability (RMIP) was calculated for each variant. Only variants with RMIP > 0.05 were considered robust and retained for downstream interpretation. Manhattan plots were created using the CMplot R package (v3.6.2) (Yin et al. 2021). Statistically significant markers within 1 MB on the same chromosome were merged into single peaks. Population structure and kinship were accounted for using PCAs and kinship matrices, which helped reduce false associations, as confirmed by the Q-Q plots. Based on the observed genome-wide linkage disequilibrium (LD) decay of ~50 kb ($r^2 = 0.2$) in this population, we grouped significant SNPs within ±50 kb into single loci. Lead SNPs were identified by explicitly accounting for LD among significant variants within each locus,

ensuring that only independent associations were retained. Candidate genes within these defined regions were annotated using the BTx623 reference genome (v3.1) from Phytozome v10 (Goodstein et al. 2012). Genomic regions containing SNPs were further annotated using the Ensembl Variant Effect Predictor (VEP) (McLaren et al. 2016).

Identification of metabolite gene clusters

Metabolic (biosynthetic) gene clusters (MGCs) in sorghum were identified using the plantiSMASH web server (version 2.0-beta4) according to default analysis protocols (Kautsar et al. 2017). Advanced parameters included a CD-HIT cutoff of 0.5 to cluster protein sequences with ≥50% identity and a minimum of 2 unique biosynthetic domains per candidate cluster. Built-in profile Hidden Markov Models (pHMMs) enhanced the detection of plant-specific biosynthetic domains. Candidate clusters were further manually validated based on their domain composition and gene co-expression patterns, using seed development RNAseq data (Khan et al. 2024). Significant mGWAS variants were then mapped onto the validated clusters to identify metabolite-associated loci and the genes harboring these variants.

Time-staged tissue expression profiling defining lignin biosynthesis during seed development

To demonstrate the potential for mGWAS to elucidate trait architecture when integrated with time-resolved multi-omics data, mGWAS results were combined with previously published RNA-Seq expression profiles (E-MTAB-13406) (Khan et al. 2024) covering sorghum seed development stages. RNA-Seq data from five developmental time points (5, 10, 15, 20, and 25 days post-anthesis [dpa]) representing early (≤6 dpa), middle (6-24 dpa), and late (25-35 dpa) seed developmental phases were analyzed to investigate lignin biosynthesis in embryo, endosperm, and whole-grain tissues. Protein-protein interaction networks were inferred using STRING v12.0 (Von Mering et al. 2005), and correlation networks among genes were visualized using Cytoscape v3.10.1 (Shannon et al. 2003).

Machine learning-based analysis of kernel color-associated metabolites

Kernel color information for the SAP accessions was obtained from the U.S. National Plant Germplasm System (NPGS) (Council et al. 1990; Byrne et al. 2018). According to the Germplasm Resources Information Network (GRIN) database naming convention, grain color was categorized as follows: 1 = white, 2 = red, 3 = yellow, and 4 = brown, encompassing a total of 211 accessions. Prior to modeling, the metabolomic dataset (1,428 polyphenolic features) was pre-processed as follows: probabilistic quotient normalization, $\log_2$ transformation, removal of features with >20% missing values across samples, and k-nearest neighbor imputation (k = 5) for the remaining missing values using the impute R package.

To identify key polyphenolic metabolites associated with variation in seed color, we applied supervised machine learning models-Random Forest (RF) (Liaw and Wiener 2002; Erban et al. 2019) and Extreme Gradient Boosting (XGBoost) (Chen et al. 2017), in a multi-class classification framework (four color classes). These models used metabolite abundance data as input features and were trained to predict the grain color category. For full reproducibility, random seeds were fixed (np.random.seed(42), random.seed (42), and XGBoost seed = 42). This ensures that the ML methods are fully repeatable. Each metabolite was assigned an importance score by the models, indicating how strongly it contributes to distinguishing grain color categories. Metabolites with higher scores have a greater influence on the model's classification performance. Models were trained and evaluated using repeated stratified 10-fold cross-validation (10 repeats, yielding 100 total folds) via RepeatedStratifiedKFold (scikit-learn), preserving the original class distribution in each fold (~50% white, ~25% red, ~15% yellow, ~10% brown) and mitigating class imbalance effects (Fontanari et al. 2022), where one class (e.g., positive or negative cases) is underrepresented relative to the other, which can bias model training and evaluation. Random Forest was run with the following hyperparameters: n_estimators = 2000, max_features = 'sqrt', min_samples_split = 2, min_samples_leaf = 1, bootstrap = True, class_weight = 'balanced_subsample'. Feature importance was calculated using the mean decrease in Gini

impurity averaged across all trees and all 100 folds. XGBoost was trained with objective = 'multi:softprob', num_class = 4, eta = 0.05, max_depth = 6, min_child_weight = 3, subsample = 0.8, colsample_bytree = 0.8, gamma = 0.1, alpha = 0.01, lambda = 1.0, tree_method = 'hist', and early_stopping_rounds = 50. Feature importance was based on the “gain” metric, again averaged over the 100 folds. Metabolites that consistently ranked in the top 30 most important features in ≥70% of the cross-validation folds in both RF and XGBoost models were considered robust discriminators of grain color (Supplemental Table S7). Receiver Operating Characteristic (ROC) curve analysis (Hajian-Tilaki 2013; Burke 2023) was subsequently performed in R v4.3.1 using the pROC package to evaluate the discriminatory ability of each of the 27 individual polyphenolic metabolites across the four color groups (one-vs-rest and pairwise comparisons). The area under the curve (AUC) and 95% confidence intervals were calculated using 2000 bootstrap replicates.

Lead SNPs associated with seven polyphenolic compounds were then analyzed to assess allelic variation. Alleles were numerically encoded as follows: the reference allele was assigned a value of 1, the alternative allele a value of 2, and the heterozygous allele a value of 1.5. Haplotype analysis was performed using the R package geneHapR (Zhang et al. 2023a). Functionally linked SNPs were extracted and analyzed for their correlation with phenotypic traits using the SKAT package (Lee et al. 2011; Wu et al. 2011). Differential metabolomic abundance for each haplotype was visualized using the ggplot2 package (Wickham 2011). To explore broader haplotype architectures, joint haplotypes were constructed by clustering haplotype blocks spanning multiple genomic regions. Subsequent association analyses examined the relationships between these joint haplotypes and candidate polyphenolic metabolites implicated in grain color variation, assessing their potential functional relevance. Additionally, a genome-wide correlation network analysis was conducted to identify SNPs with the highest connectivity to polyphenolic metabolites, providing insights into key genetic components influencing pigment biosynthesis. Partial correlation coefficients (r) and P-values were computed pairwise across all SNPs using the ppcor package (Kim 2015). SNP pairs with significant positive correlations ($r = 0.5$, $P \leq 0.05$) were filtered and visualized as a

correlation network in Cytoscape v3.10.1 (Shannon et al. 2003) to assess their potential roles in polyphenol-associated pathways.

Database development

SorGMDA was developed using MySQL and PHP to enable users to search metabolome profiling data acquired from the SAP accessions studied. The database also integrates mGWAS signals with P-values $\leq 10^{-3}$ mapped across all chromosomes.

Data availability

All metabolic traits and corresponding GWAS results can be accessed through SorGMDA (https://www.depts.ttu.edu/igcast/SorbGMDA/index.php).

Code availability

All analyses were performed using publicly available software and standard computational workflows as described in the Methods section. The full machine-learning pipeline used in this study, including data preprocessing, model training, hyperparameter settings, cross-validation procedures, feature selection, ROC analysis, and downstream genetic association scripts, is openly accessible on GitHub at: https://github.com/Deepti098c/sorghum_kernel_color_ML.

Author contributions

YJ conceptualized and designed the project. MY contributed to project discussions. NK and RT prepared samples for metabolomic analysis. SM, QS, and FC conducted the HPLC experiments. DN performed all data analyses. PK and DN designed the database. DN, YJ, and MY drafted the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

Acknowledgments

This research was supported by the U.S. Department of Agriculture (USDA)'s Intramural Research Program (3096-21000-024-00D), the National Institute of Food and Agriculture's

Agriculture and Food Research Initiative (AFRI) under award number 2023-67013-39631, and the Non-Assistance Cooperative Agreement 58-3020-2-024 between Texas Tech University and USDA-ARS (3020-43440-002-00D). The authors are grateful to the publicly available USDA Germplasm Resources Information Network (GRIN) database for providing accessions of the SAP used in this study.

Competing Interests

The authors declare no conflict of interest.

Figure Legends

Main Figures:

Fig. 1: Sorghum grain metabolite diversity. A) Metabolite diversity index, reflecting the variability across samples. B) Violin plots showing metabolite diversity scores across different superclasses, indicating the distribution and density of diversity within each class. C) Scatter plot of the top 5% most diverse metabolites with significant annotations; the *x*-axis represents the -log10(*p*-value) and the *y*-axis shows fold change. Red and blue dots denote metabolites with higher and lower diversity, respectively.

Fig. 2: Sorghum grain metabolomic networks. Metabolite networks where each node represents a metabolite signal. Grey edges illustrate the molecular MS/MS network based on spectral similarity, while red edges highlight co-accumulating grain metabolites that share similar accumulation patterns across 266 sorghum accessions. Metabolite clusters are outlined, with structures of representative metabolites shown for each cluster. Metabolite names within each cluster are provided in Supplemental Table S1. Node size reflects *p*-values from the Kolmogorov-Smirnov test: larger nodes ($p < 0.01$) represent metabolites with higher statistical significance, while smaller nodes ($p > 0.01$) represent those with lower statistical significance.

Fig. 3: The mGWAS results for lysine biosynthesis **in sorghum grain. A)** Manhattan plot showing significant associations for lysine accumulation. The *x*-axis represents the 10 sorghum chromosomes, while the *y*-axis shows -log10(*p*-value). Associations are assigned to genes located within ±50 kb of lead SNPs. Red dotted lines indicate genomic loci near candidate genes, including AK (aspartate kinase*)* and LDC (lysine decarboxylase) on chromosome 1; AK-HSD2 (aspartate kinase-homoserine dehydrogenase 2) on chromosome 2; AKIII (aspartate kinase III) on chromosome 3; ACSL (acyl-CoA synthetase long-chain family member) on chromosome 5; DHDPS (dihydrodipicolinate synthase) and PL (putative peptidase-like gene) on chromosome 6; and TIP (tonoplast intrinsic protein) on chromosome 9. **B)** Haplotypes identified for three key genes, AKIII, DHDPS, and LDC, showing allelic variants. **C)** Lysine biosynthetic pathway in sorghum, depicting major intermediates, enzymatic steps, and corresponding genes. Enzyme Commission (EC) numbers are provided where applicable. Genes encoding key enzymes identified are highlighted in red, including *Sobic.001G011700* and *Sobic.003G412100* (aspartate kinases), and *Sobic.006G186600* (*DHDPS*).

Fig. 4: Genomic enrichment of mGWAS signals. **A)** Distribution of SNPs across different genomic regions in the SAP original dataset (input). **B)** SNPs significantly associated with 2,206 metabolites identified through mGWAS. **C)** Fold enrichment analysis of SNPs across genomic regions, showing notable enrichment of missense, synonymous, and UTR variants among mGWAS-associated SNPs compared to the input set. D) Distance of non-genic SNPs from the nearest gene.

Fig. 5: Chromosomal distribution and functional analysis of MGCs. A) Distribution of metabolic gene clusters (MGCs) across 10 sorghum chromosomes. Cluster types are color-coded by biosynthetic pathway. The inset bar plot shows the total number of MGCs identified per chromosome. The x-axis represents chromosomal position (base pairs), and the y-axis indicates chromosome number. **B)** The lignin metabolic gene cluster. The top panel depicts

the gene arrangement and location of sequence variants near methyltransferase-encoding genes and a dirigent enzyme gene. The bottom panel shows the co-expression network of genes within this cluster, highlighting functionally related genes involved in flavonoid biosynthesis (red nodes), stilbenoid, diarylheptanoid, and gingerol biosynthesis (green nodes), and glutathione biosynthesis (blue nodes). **C)** Heatmap of co-expressed lignin biosynthesis genes with expression levels (FPKM > 1). Genes involved in flavonoid biosynthesis are marked in blue, while genes associated with stilbenoid, diarylheptanoid, and gingerol biosynthesis are highlighted in red and green, respectively.

Fig. 6: Metabolomic clustering and loci mapping of key grain polyphenols among sorghum accessions. **A)** Hierarchical clustering of 2,206 metabolites profiled across 211 sorghum accessions grouped by GRIN classification: blue (77 lines), red (22 lines), yellow (10 lines), and white (102 lines). Metabolite superclasses are color-coded, revealing distinct metabolic signatures among genetic groups. **B)** Feature importance plot highlighting the top-ranking polyphenols consistently identified by both Random Forest (RF) and XGBoost models, based on their predictive importance scores. **C)** Composite Manhattan plot of the top seven polyphenols, showing significant genomic loci associated with their variation. Previously reported genes were annotated at their corresponding chromosomal positions, supporting the validity of the associations. **D)** Distribution of 13 super-haplotypes carrying combinations of alleles associated with seed color, arranged from dark to light pigmentation. **E)** Normalized metabolite abundance of seven key polyphenols across the 13 super-haplotypes, illustrating combined genetic effects.

Supplemental figures:

Supplemental Figure S1: Workflow for mass spectrometry-based metabolomics analysis in the Sorghum Association Panel (SAP). The workflow for mass spectrometry-based metabolomics analysis covered data filtration, peak annotation, and downstream analyses. In

Step A, raw data from Thermo Fisher Scientific mass spectrometers (.raw format) are converted to standardized formats (.mzML/.mzXML) for further processing. Step B details several stages of data processing: peak detection identifies metabolites by their mass-to-charge ratio (*m/z*) and intensity; retention time alignment ensures consistency across samples; signal filtration improves data quality by evaluating peak shape and count; intensity noise filtration removes low-intensity noise; gap filtration addresses missing data points; and feature grouping organizes the processed data into an intensity matrix. In Step C, processed peaks are annotated using established databases, such as HMDB, MassBank, ChemSpider, and KEGG. The workflow concludes with downstream analyses, including the calculation of metabolite diversity, mGWAS, and validation of results through further experimental or computational methods.

Supplemental Figure S2: Multivariate and functional categorization of identified metabolites. A) Principal Component Analysis (PCA) of 2,206 metabolites, with PC1 and PC2 explaining 42% and 38% of the total variance, respectively. Each dot represents a metabolite: dark green indicates non-variable metabolites, while light green denotes those with variability. B) Treemap visualization of the identified metabolites, categorized into distinct chemical classes. Each class is represented by a different color, with the size of each box proportional to the number of metabolites it contains. C) Hierarchical clustering heatmap showing relationships among metabolites based on their abundance profiles across the 266 SAP accessions. D) Correlation plot showing the metabolites grouped into major functional categories, including amino acids, lipids, carbohydrates, nucleotides, and polyphenols, as highlighted by color-coded boxes.

Supplemental Figure S3: Genome-wide association signals for (-)-Epigallocatechin and D-(-)-Fructose. Manhattan plots depicting GWAS results for (-)-epigallocatechin ($C_{15}H_{14}O_7$) and D-(-)-fructose ($C_6H_{12}O_6$), both used as positive controls. Each point represents an SNP, with the chromosomal position on the x-axis and the statistical significance ($-\log_{10}$ p-value)

on the y-axis. Peaks indicate SNPs with strong associations with the respective traits, marking genomic regions potentially involved in their biosynthesis or regulation.

Supplemental Figure S4. Distribution of metabolite values and SNP association prioritization. A) Distribution of 2206 metabolite values demonstrating normality across samples (Shapiro-Wilk test $P > 0.05$; within ±3 standard deviations). The histogram displays the frequency of metabolite measurements, with an overlaid normal distribution curve (red line), illustrating the retention of metabolites that exhibit normal distribution characteristics. B) Comparison of the total number of single nucleotide polymorphisms (SNPs) identified per metabolite and the subset of SNPs with resample model inclusion probability (RMIP) ≥ 0.05. The RMIP threshold was used to prioritize SNPs consistently detected across multiple resampling iterations, enhancing the reliability and reproducibility of association signals.

Supplemental Figure S5: Genomic distribution of Metabolic Gene Clusters (MGCs) and SNP variation. Metabolic gene clusters (total 38) in the sorghum genome, labeled from cluster 1 to cluster 38, along with SNPs associated with metabolites via mGWAS. The red dotted vertical line indicates the location of identified SNPs. On the right-hand side, the figure shows the location and size of each gene cluster within the genome.

Supplemental Figure S6: Grain polyphenol diversity and enzymatic pathway variation across SAP accessions. A) 49 Polyphenols identified across four grain color groups. B) PCA of metabolite profiles across 211 accessions. PC1 and PC2 explain 43% and 37% of the total variance, respectively. Accessions are color-coded by grain color groups: blue (77 lines), red (22 lines), yellow (10 lines), and white (102 lines). C) Metabolites are shown as dots with grey representing all detected metabolites, green indicating secondary metabolites, and yellow highlighting 49 polyphenols exhibiting diversity across the SAP. D) Schematic representation of the phenylpropanoid, flavonoid, and lignin biosynthetic pathways. Enzymes

detected near regions of genetic variation (±50 kb) are indicated, linking metabolite diversity with potential regulatory loci. E) Quantile-quantile (QQ) plot of GWAS results for seven selected flavonoids. The plot displays the observed versus expected $-\log_{10}(p)$ values for SNP associations with metabolite abundances. Each color represents a different metabolite, including 3-Coumaric acid (light blue), 6-Hydroxy-7-Methoxycoumarin (green), Genistein (grey), Isorhamnetin-3-glucoside (blue), Isorhamnetin-3-rutinoside (purple), Quercetin-3-(6''-malonyl)-Glucoside (pink), and Kaempferol (red). The deviation from the diagonal red line indicates SNPs with stronger-than-expected associations, highlighting potential genetic loci controlling metabolite variation.

Supplemental Figure S7: Receiver Operating Characteristic (ROC) curves for seven polyphenols showing their performance in distinguishing between 4 groups. Each panel displays the ROC curve for an individual polyphenol, with the shaded area representing the 95% confidence interval around the curve. The diagonal dashed line indicates the performance of a random classifier (AUC = 0.5). These curves illustrate the discriminatory power of individual compounds based on their true-positive and false-positive rates.

Supplemental Figure S8: Allele frequency distribution of Hap1 across four grain color groups.

Violin plots display the distribution of Hap1 allele frequencies across red, brown, yellow, and white grain color groups. Hap1 was predominantly associated with darker grains (red and brown) and occurred at much lower frequencies in lighter grains (yellow and white). A Chi-square test confirmed a significant association between Hap1 frequency and grain color ($\chi^2 = 13.82$, $df = 3$, $p = 0.00316$).

Supplemental Figure S9: Density distribution of metabolite coverage across SAP accessions. The *x*-axis represents the percentage of accessions in which each metabolite is detected, while the *y*-axis shows the density of metabolites at each coverage level.

Supplemental Figure 10: Phenotypic variation and genetic architecture of grain (–)-epigallocatechin accumulation in sorghum. A) Distribution of (-)-epigallocatechin across GRIN-based SAP line numbers, with corresponding abundance values measured by UPLC-MS, obtained using SorGMDA. B) Manhattan plot of GWAS results for grain (-)-epigallocatechin accumulation in SAP accessions. The plot displays -$\log_{10}$(*p*-value) for each SNP across all sorghum chromosomes (Chr 1-10) with alternating colors used to distinguish between chromosomes. Only associations with *p*-values $\leq 10^{-3}$ are shown. Each dot represents an SNP, with the vertical axis indicating the statistical significance of association and the horizontal axis representing genomic position. A strong association signal is observed on chromosome 3, with a peak at position 165.4528 Mb showing a -$\log_{10}$(p-value) of 45. The red and blue dashed lines represent suggestive and genome-wide significance thresholds, respectively.

Main Table

Table 1. Allelic variation at candidate SNPs influences the accumulation of seven polyphenolic compounds in sorghum grain. The table summarizes significant SNP-metabolite associations, indicating the chromosomal position, allele type, and allele frequency of the top haplotype (Hap1) across different grain color groups (red, brown, yellow, white). Violin plots (right) show the distribution of metabolite fold changes across genotypes for each SNP-metabolite pair. Notable associations include SNPs near previously known genes such as *Y1*, *TAN2*, *TT2*, and *TT10*, which influence the levels of metabolites like kaempferol, genistein, and quercetin derivatives. Asterisks indicate significance levels based on statistical comparisons of allele groups ($p < 0.01$**).

Supplemental Table

Table S1. Feature Importance and Statistical Significance of the Top Seven Polyphenols. A total of seven metabolites were identified using XGBoost and Random Forest (RF) as key discriminators of grain pigment groups among accessions selected based on grain color (red, brown, yellow, and white).

Supplemental files:

Supplemental Table S1.

Table listing the metabolite features detected via LC-MS, including Feature ID, m/z (mass-to-charge ratio), retention time (rtime), molecular formula, and compound name. Peak intensities are shown for each sorghum accession, representing relative abundance values derived from the metabolomics dataset.

Supplemental Table S2.

List of metabolite features annotated with mass-to-charge ratio (m/z), retention time (rtime), molecular formula, and compound name. Each metabolite is characterized by its coefficient of variation (CV), diversity classification based on CV thresholds, and broad-sense heritability. Diversity categories are defined as follows: Low (CV < 40%), Medium (CV 40-80%), and High (CV > 80%).

Supplemental Table S3.

List of candidate genes associated with the accumulation of (-)-epigallocatechin ($C_{15}H_{14}O_7$), D-(-)-fructose ($C_6H_{12}O_6$), and lysine ($C_6H_{14}N_2O_3$) metabolites, identified through GWAS and supported by functional annotation. The table includes metabolite names, gene identifiers, chromosomal locations, SNP positions, gene boundaries (start and end coordinates), *p*-values, and corresponding gene annotations. Genes annotated as "putative" or "uncharacterized"

reflect limited or predicted functional information as currently available in the Phytozome database.

Supplemental Table S4.

Summary of SNP discovery and binning across ~18 Mb genomic intervals for ten selected metabolites. The table includes compound names, total SNPs identified, SNPs with a final Resample Model Inclusion Probability (RMIP) > 0.05, genomic bins (each approximately 18 Mb in size), total SNP counts per bin, and the count of SNPs within each bin that meet the RMIP threshold. RMIP values reflect the robustness of SNP-trait associations across resampling iterations in GWAS models.

Supplemental Table S5.

Comparison of the genomic distribution of SNPs identified through mGWAS versus the total input SNPs used in the analysis. Genomic regions include 3' and 5' UTRs, upstream and downstream regulatory regions, intergenic areas, introns, missense, and synonymous coding sites. Counts and percentages reflect the proportion of SNPs within each genomic category out of the total SNPs detected in mGWAS (left) and the full input SNP set (right). This analysis highlights the genomic context of SNP-trait associations and their relative enrichment in functional or regulatory regions.

Supplemental Table S6.

Summary of biosynthetic gene clusters identified in the sorghum genome detailing their genomic coordinates, cluster type, and key functional domains. Each cluster is defined by its chromosome position (NCBI accession), start and end coordinates, approximate size in kb, core biosynthetic or regulatory domains, and the number of gene clusters grouped by CD-HIT. Cluster classification is based on domain composition and genes proximity.

Supplemental Table S7.

List of 83 genes located within or near Metabolic Gene Clusters (MGCs) in the sorghum genome that harbor significant SNPs. The table includes gene names, chromosomal coordinates, variant site locations, distances between variants and gene boundaries (in kb), and corresponding GWAS p-values. Each gene is annotated with available functional information, including Gene Ontology (GO), Pfam, InterPro, KEGG, KOG, Panther, and Enzyme classifications.

Supplemental Table S8.

Normalized expression profiles (FPKM) of flavonoid biosynthesis genes in sorghum across 45 seed development RNA-seq samples. Samples are categorized into early whole seeds (se1-se9), embryo (em10-em25), and endosperm (en6-en25). The expression values represent the relative transcriptional activity of each gene under different developmental stages and tissue types. Pathway annotations are based on KEGG and other curated databases. Zero values indicate undetectable gene expression.

Supplemental Table S9.

Classification of SAP accessions based on four grain color groups, as recorded in the Germplasm Resources Information Network (GRIN) database.

Supplemental Table S10.

Selected secondary metabolites identified in sorghum seeds, highlighting their relative importance and statistical significance in relation to seed color. Sorghum accessions were classified into four color categories: white, red, brown, and yellow. Each row represents a specific metabolite, accompanied by its importance score, associated p-value, and the corresponding $-\log_{10}$(p-value), which facilitates interpretation of statistical significance. Quantitative values reflect normalized metabolite abundances across samples within each color group.

Supplemental Table S11.

Metabolites and genetic mapping results identifying candidate genes associated with variation among four grain color groups. For each metabolite, the table lists the statistical model(s) used for association analysis (FarmCPU, MLM), chromosome number, standardized effect size, lead SNP position (in base pairs), p-value of association, mapped gene ID, gene start and end positions, gene annotation, and reference genome version (BTx623 v3.1).

Supplemental Table S12

SAP accessions, grain color category (white), binary true labels, and normalized abundances for seven key metabolites associated with grain color variation. Metabolite values represent relative abundances normalized across all samples. The true labels column indicates binary classification (e.g., presence or absence of a trait) used in predictive modeling or phenotype association analyses.

References

Ali AEE, Sharp RE, Greeley L, Peck SC, Tabb DL, Ludidi N. 2025. Proteomic dataset of sorghum leaf and root responses to single and combined drought and heat stress. *Scientific Data* **12**: 97.

Almodares A, Hadi M. 2009. Production of bioethanol from sweet sorghum: A review. *African journal of agricultural research* **4**: 772-780.

Arbex PM, de Castro Moreira ME, Toledo RCL, de Morais Cardoso L, Pinheiro-Sant'ana HM, dos Anjos Benjamin L, Licursi L, Carvalho CWP, Queiroz VAV, Martino HSD. 2018. Extruded sorghum flour (Sorghum bicolor L.) modulate adiposity and inflammation in high fat diet-induced obese rats. *Journal of functional foods* **42**: 346-355.

Awika JM, Rooney LW. 2004. Sorghum phytochemicals and their potential impact on human health. *Phytochemistry* **65**: 1199-1221.

Awika JM, Yang L, Browning JD, Faraj A. 2009. Comparative antioxidant, antiproliferative and phase II enzyme inducing potential of sorghum (Sorghum bicolor) varieties. *LWT-Food Science and Technology* **42**: 1041-1046.

Bi B, Wang K, Zhang H, Wang Y, Fei H, Pan R, Han F. 2021. Plants use rhizosphere metabolites to regulate soil microbial diversity. *Land Degradation & Development* **32**: 5267-5280.

Boatwright JL, Sapkota S, Jin H, Schnable JC, Brenton Z, Boyles R, Kresovich S. 2022. Sorghum Association Panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. *The Plant Journal* **111**: 888-904.

Bordiga M, Gomez-Alonso S, Locatelli M, Travaglia F, Coïsson JD, Hermosin-Gutierrez I, Arlorio M. 2014. Phenolics characterization and antioxidant activity of six different pigmented Oryza sativa L. cultivars grown in Piedmont (Italy). *Food Research International* **65**: 282-290.

Boyles RE, Pfeiffer BK, Cooper EA, Rauh BL, Zielinski KJ, Myers MT, Brenton Z, Rooney WL, Kresovich S. 2017. Genetic dissection of sorghum grain quality traits using diverse and segregating populations. *Theoretical and applied genetics* **130**: 697-716.

Bröhan M, Jerkovic V, Wilmotte R, Collin S. 2011. Catechins and derived procyanidins in red and white sorghum: Their contributions to antioxidant activity. *Journal of the Institute of Brewing* **117**: 600-607.

Bunsupa S, Yamazaki M, Saito K. 2017. Lysine-derived alkaloids: overview and update on biosynthesis and medicinal applications with emphasis on quinolizidine alkaloids. *Mini Reviews in Medicinal Chemistry* **17**: 1002-1012.

Burdette A, Garner PL, Mayer EP, Hargrove JL, Hartle DK, Greenspan P. 2010. Anti-inflammatory activity of select sorghum (Sorghum bicolor) brans. *Journal of medicinal food* **13**: 879-887.

Burke HB. 2023. Use the receiver operating characteristic to assess model accuracy. *JAMA cardiology* **8**: 998-998.

Bursać M, Krstonošić MA, Miladinović J, Malenčić Đ, Gvozdenović L, Cvejić JH. 2017. Isoflavone composition, total phenolic content and antioxidant capacity of soybeans with colored seed coat. *Natural Product Communications* **12**: 1934578X1701200417.

Buzas MA, Hayek L-AC. 2005. On richness and evenness within and between communities. *Paleobiology* **31**: 199-220.

Byrne PF, Volk GM, Gardner C, Gore MA, Simon PW, Smith S. 2018. Sustaining the future of plant breeding: The critical role of the USDA-ARS National Plant Germplasm System. *Crop Science* **58**: 451-468.

Campbell BC, Gilding EK, Mace ES, Tai S, Tao Y, Prentis PJ, Thomelin P, Jordan DR, Godwin ID. 2016. Domestication and the storage starch biosynthesis pathway: signatures of selection from a whole sorghum genome sequencing strategy. *Plant biotechnology journal* **14**: 2240-2253.

Carreno-Quintero N, Bouwmeester HJ, Keurentjes JJ. 2013. Genetic analysis of metabolome–phenotype interactions: from model to crop species. *Trends in Genetics* **29**: 41-50.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* **30**: 918-920.

Chen B-R, Wang C-Y, Ping W, Zhu Z-X, Ning X, Shi G-S, Miao Y, Nai W, Li J-H, Hou J-M. 2019. Genome-wide association study for starch content and constitution in sorghum (Sorghum bicolor (L.) Moench). *Journal of Integrative Agriculture* **18**: 2446-2456.

Chen J, Hu X, Shi T, Yin H, Sun D, Hao Y, Xia X, Luo J, Fernie AR, He Z. 2020. Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. *Plant Biotechnology Journal* **18**: 1722-1735.

Chen J, Xue M, Liu H, Fernie AR, Chen W. 2021a. Exploring the genic resources underlying metabolites through mGWAS and mQTL in wheat: From large-scale gene identification and pathway elucidation to crop improvement. *Plant Communications* **2**.

Chen L, Zhang T, Li T. 2017. Gradient boosting model for unbalanced quantitative mass spectra quality assessment. In *2017 International conference on security, pattern analysis, and cybernetics (SPAC)*, pp. 394-399. IEEE.

Chen W, Wang W, Peng M, Gong L, Gao Y, Wan J, Wang S, Shi L, Zhou B, Li Z. 2016. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nature communications* **7**: 12767.

Chen X, Shen J, Xu J, Herald T, Smolensky D, Perumal R, Wang W. 2021b. Sorghum phenolic compounds are associated with cell growth inhibition through cell cycle arrest and apoptosis in human hepatocarcinoma and colorectal adenocarcinoma cells. *Foods* **10**: 993.

Cingoz GS, Gurel E. 2016. Effects of salicylic acid on thermotolerance and cardenolide accumulation under high temperature stress in Digitalis trojana Ivanina. *Plant Physiology and Biochemistry* **105**: 145-149.

Coklar H, Akbulut M. 2021. Changes in phenolic acids, flavonoids, anthocyanins, and antioxidant activities of Mahonia aquifolium berries during fruit development and elucidation of the phenolic biosynthetic pathway. *Horticulture, Environment, and Biotechnology* **62**: 785-794.

Collins A, Santhakumar AB, Francis N, Blanchard C, Chinkwo K. 2024. Impact of sorghum (Sorghum bicolor L. Moench) phenolic compounds on cancer development pathways. *Food Bioscience*: 104177.

Council NR, Resources CoMGG, Imperatives A. 1990. The US national plant germplasm system.

DeJong TM. 1975. A comparison of three diversity indices based on their components of richness and evenness. *Oikos*: 222-227.

Dowd PF, Funnell-Harris DL, Sattler SE. 2016. Field damage of sorghum (Sorghum bicolor) with reduced lignin levels by naturally occurring insect pests and pathogens. *Journal of Pest Science* **89**: 885-895.

Duressa D, Weerasoriya D, Bean SR, Tilley M, Tesso T. 2018. Genetic basis of protein digestibility in grain sorghum. *Crop Science* **58**: 2183-2199.

Dykes L, Rooney WL, Rooney LW. 2013. Evaluation of phenolics and antioxidant activity of black sorghum hybrids. *Journal of Cereal Science* **58**: 278-283.

Dykes L, Seitz LM, Rooney WL, Rooney LW. 2009. Flavonoid composition of red sorghum genotypes. *Food Chemistry* **116**: 313-317.

Eiler J, Cesar J, Chimiak L, Dallas B, Grice K, Griep-Raming J, Juchelka D, Kitchen N, Lloyd M, Makarov A. 2017. Analysis of molecular isotopic structures at high precision and accuracy by Orbitrap mass spectrometry. *International Journal of Mass Spectrometry* **422**: 126-142.

Erban A, Fehrle I, Martinez-Seidel F, Brigante F, Más AL, Baroni V, Wunderlin D, Kopka J. 2019. Discovery of food identity markers by metabolomics and machine learning technology. *Scientific Reports* **9**: 9697.

Falcone Ferreyra ML, Emiliani J, Rodriguez EJ, Campos-Bermudez VA, Grotewold E, Casati P. 2015. The identification of maize and Arabidopsis type I flavone synthases links flavones with hormones and biotic interactions. *Plant Physiology* **169**: 1090-1107.

Farrar JL, Hartle DK, Hargrove JL, Greenspan P. 2008. A novel nutraceutical property of select sorghum (Sorghum bicolor) brans: inhibition of protein glycation. *Phytotherapy Research* **22**: 1052-1056.

Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee DY, Lu Y, Moon S, Nikolau B. 2008. Quality control for plant metabolomics: reporting MSI-compliant studies. *The Plant Journal* **53**: 691-704.

Fontanari T, Fróes TC, Recamonde-Mendoza M. 2022. Cross-validation strategies for balanced and imbalanced datasets. In *Brazilian Conference on Intelligent Systems*, pp. 626-640. Springer.

Frei M. 2013. Lignin: characterization of a multifaceted crop component. *The Scientific World Journal* **2013**: 436517.

Fukushima A, Kusano M, Redestig H, Arita M, Saito K. 2011. Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC systems biology* **5**: 1.

Galili G, Amir R, Fernie AR. 2016. The regulation of essential amino acid synthesis and accumulation in plants. *Annual review of plant biology* **67**: 153-178.

Girard AL, Awika JM. 2018. Sorghum polyphenols and other bioactive components as functional and health promoting food ingredients. *Journal of Cereal Science* **84**: 112-124.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**: D1178-D1186.

Gorelick R. 2006. Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices. *Ecography* **29**: 525-530.

Grace SC. 2005. Phenolics as antioxidants. *Antioxidants and reactive oxygen species in plants*: 141-168.

Habyarimana E, Dall'Agata M, De Franceschi P, Baloch FS. 2019. Genome-wide association mapping of total antioxidant capacity, phenols, tannins, and flavonoids in a panel of Sorghum bicolor and S. bicolor× S. halepense populations using multi-locus models. *PLoS One* **14**: e0225979.

Hajian-Tilaki K. 2013. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine* **4**: 627.

Halliday FW, Rohr JR. 2019. Measuring the shape of the biodiversity-disease relationship across systems reveals new findings and key gaps. *Nature communications* **10**: 5032.

Hamblin MT, Salas Fernandez MG, Tuinstra MR, Rooney WL, Kresovich S. 2007. Sequence variation at candidate loci in the starch metabolism pathway in sorghum: prospects for linkage disequilibrium mapping. *Crop Science* **47**: S-125-S-134.

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry* **45**: 703-714.

Jander G, Joshi V. 2009. Aspartate-derived amino acid biosynthesis in Arabidopsis thaliana. *The arabidopsis book/American society of plant biologists* **7**: e0121.

Kai K, Shimizu B-i, Mizutani M, Watanabe K, Sakata K. 2006. Accumulation of coumarins in Arabidopsis thaliana. *Phytochemistry* **67**: 379-386.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**: 27-30.

Karakas E, Bulut M, Fernie A. 2025. Metabolome guided treasure hunt-learning from metabolic diversity. *Journal of Plant Physiology*: 154494.

Kathuria D, Thakur S, Singh N. 2024. Advances of metabolomic in exploring phenolic compounds diversity in cereal and their health implications. *International Journal of Food Science and Technology* **59**: 4213-4233.

Kaufman RC, Herald TJ, Bean SR, Wilson JD, Tuinstra MR. 2013. Variability in tannin content, chemistry and activity in a diverse group of tannin containing sorghum cultivars. *Journal of the Science of Food and Agriculture* **93**: 1233-1241.

Kaur S, Kumar K, Singh L, Sharanagat VS, Nema PK, Mishra V, Bhushan B. 2024. Gluten-free grains: Importance, processing and its effect on quality of gluten-free products. *Critical reviews in food science and nutrition* **64**: 1988-2015.

Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. 2017. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic acids research* **45**: W55-W63.

Khan A, Tian R, Bean SR, Yerka M, Jiao Y. 2024. Transcriptome and metabolome analyses reveal regulatory networks associated with nutrition synthesis in sorghum seeds. *Communications Biology* **7**: 841.

Kim S. 2015. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods* **22**: 665.

Lam PY, Zhu F-Y, Chan WL, Liu H, Lo C. 2014. Cytochrome P450 93G1 is a flavone synthase II that channels flavanones to the biosynthesis of tricin O-linked conjugates in rice. *Plant Physiology* **165**: 1315-1327.

Lee J, Hwang Y-S, Kim ST, Yoon W-B, Han WY, Kang I-K, Choung M-G. 2017. Seed coat color and seed weight contribute differential responses of targeted metabolites in soybean seeds. *Food Chemistry* **214**: 248-258.

Lee S, Miropolsky L, Wu M, Lee M, CompQuadForm D. 2011. Package “SKAT.”. In *CRAN*.

Li J, Tang W, Zhang Y-W, Chen K-N, Wang C, Liu Y, Zhan Q, Wang C, Wang S-B, Xie S-Q. 2018. Genome-wide association studies for five forage quality-related traits in sorghum (Sorghum bicolor L.). *Frontiers in plant science* **9**: 1146.

Li M-X, Yeung JM, Cherny SS, Sham PC. 2012. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human genetics* **131**: 747-756.

Li Q, Fu C, Liang C, Ni X, Zhao X, Chen M, Ou L. 2022. Crop lodging and the roles of lignin, cellulose, and hemicellulose in lodging resistance. *Agronomy* **12**: 1795.

Li Y, Miao Y, Yuan H, Huang F, Sun M, He L, Liu X, Luo J. 2024. Volatilome-based GWAS identifies OsWRKY19 and OsNAC021 as key regulators of rice aroma. *Molecular Plant* **17**: 1866-1882.

Liang J, Shen Q, Wang L, Liu J, Fu J, Zhao L, Xu M, Peters RJ, Wang Q. 2021. Rice contains a biosynthetic gene cluster associated with production of the casbane-type diterpenoid phytoalexin ent-10-oxodepressin. *New Phytologist* **231**: 85-93.

Liaqat W, Altaf MT, Barutçular C, Mohamed HI, Ahmad H, Jan MF, Khan EH. 2024. Sorghum: a star crop to combat abiotic stresses, food insecurity, and hunger under a changing climate: a review. *Journal of Soil Science and Plant Nutrition* **24**: 74-101.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R news* **2**: 18-22.

Lipka AE, Gore MA, Magallanes-Lundback M, Mesberg A, Lin H, Tiede T, Chen C, Buell CR, Buckler ES, Rocheford T. 2013. Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3: Genes, Genomes, Genetics* **3**: 1287-1299.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**: 2397-2399.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nature methods* **8**: 833-835.

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

López-González C, Juárez-Colunga S, Morales-Elías NC, Tiessen A. 2019. Exploring regulatory networks in plants: transcription factors of starch metabolism. *PeerJ* **7**: e6841.

Lu Y, Wang K, Hu L. 2025. Advancements in delivery systems for dietary polyphenols in enhancing radioprotection effects: challenges and opportunities. *npj Science of Food* **9**: 51.

Luo J. 2015. Metabolite-based genome-wide association studies in plants. *Current opinion in plant biology* **24**: 31-38.

Matros A, Houston K, Tucker MR, Schreiber M, Berger B, Aubert MK, Wilkinson LG, Witzel K, Waugh R, Seiffert U. 2021. Genome-wide association study reveals the genetic complexity of fructan accumulation patterns in barley grain. *Journal of Experimental Botany* **72**: 2383-2402.

Mazumder S, Bhattacharya D, Lahiri D, Moovendhan M, Sarkar T, Nag M. 2024. Harnessing the nutritional profile and health benefits of millets: a solution to global food security problems. *Critical Reviews in Food Science and Nutrition*: 1-22.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics* **9**: 356-369.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. *Genome biology* **17**: 1-14.

Menni C, Zhu J, Le Roy CI, Mompeo O, Young K, Rebholz CM, Selvin E, North KE, Mohney RP, Bell JT. 2020. Serum metabolites reflecting gut microbiome alpha diversity predict type 2 diabetes. *Gut microbes* **11**: 1632-1642.

Michalski A, Damoc E, Hauschild J-P, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M, Horning S. 2011. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & cellular proteomics* **10**.

Mokra D, Joskova M, Mokry J. 2022. Therapeutic effects of green tea polyphenol (–)-Epigallocatechin-3-Gallate (EGCG) in relation to molecular pathways controlling inflammation, oxidative stress, and apoptosis. *International journal of molecular sciences* **24**: 340.

Mukherjee A, Maheshwari U, Sharma V, Sharma A, Kumar S. 2024. Functional insight into multi-omics-based interventions for climatic resilience in sorghum (Sorghum bicolor): a nutritionally rich cereal crop. *Planta* **259**: 91.

Ng'Uni D, Shargie N, Andersson S, Van Biljon A, Labuschagne M. 2016. Genetic variation and trait associations of yield, protein and grain micronutrients for identification of promising sorghum varieties. *Cereal Research Communications* **44**: 681-693.

Nip W, Burns E. 1969. Pigment characterization in grain sorghum. I. Red varieties. *Cereal Chem* **46**: 490-495.

Noel JP, Austin MB, Bomati EK. 2005. Structure–function relationships in plant phenylpropanoid biosynthesis. *Current opinion in plant biology* **8**: 249-253.

Nützmann HW, Huang A, Osbourn A. 2016. Plant metabolic clusters–from genetics to genomics. *New phytologist* **211**: 771-789.

Ojwang LO, Dykes L, Awika JM. 2012. Ultra performance liquid chromatography–tandem quadrupole mass spectrometry profiling of anthocyanins and flavonols in cowpea (Vigna unguiculata) of varying genotypes. *Journal of agricultural and food chemistry* **60**: 3735-3744.

Pallister T, Jackson MA, Martin TC, Zierer J, Jennings A, Mohney RP, MacGregor A, Steves CJ, Cassidy A, Spector TD. 2017. Hippurate as a metabolomic marker of gut microbiome diversity: Modulation by diet and relationship to metabolic syndrome. *Scientific reports* **7**: 13670.

Park JS, Rho HS, Kim DH, Chang IS. 2006. Enzymatic preparation of kaempferol from green tea seed and its antioxidant activity. *Journal of agricultural and food chemistry* **54**: 2951-2956.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**: 551-556.

Pelletier MK, Shirley BW. 1996. Analysis of flavanone 3-hydroxylase in Arabidopsis seedlings (Coordinate regulation with chalcone synthase and chalcone isomerase). *Plant physiology* **111**: 339-345.

Pence HE, Williams A. 2010. ChemSpider: an online chemical information resource. ACS Publications.

Pontieri P, Pepe G, Campiglia P, Merciai F, Basilicata MG, Smolensky D, Calcagnile M, Troisi J, Romano R, Del Giudice F. 2021. Comparison of content in phenolic compounds and antioxidant capacity in grains of white, red, and black sorghum varieties grown in the mediterranean area. *ACS Food Science & Technology* **1**: 1109-1119.

Przybylska-Balcerek A, Frankowski J, Stuper-Szablewska K. 2019. Bioactive compounds in sorghum. *European Food Research and Technology* **245**: 1075-1080.

Qin Sh, Yan F, E S, Xiong P, Tang Sn, Yu Kq, Zhang M, Cheng Yc, Cai W. 2022. Comprehensive characterization of multiple components of Ziziphus jujuba Mill using UHPLC-Q-Exactive Orbitrap Mass Spectrometers. *Food Science & Nutrition* **10**: 4270-4295.

Rai M, Dutta M, Saito K, Rai A. 2025. A deep dive into plant metabolomics: Milestones, technologies, and translational impact. *Plant Physiology* **199**: kiaf408.

Ramakrishna G, Kaur P, Nigam D, Chaduvula PK, Yadav S, Talukdar A, Singh NK, Gaikwad K. 2018. Genome-wide identification and characterization of InDels and SNPs in Glycine max and Glycine soja for contrasting seed permeability traits. *BMC plant biology* **18**: 141.

Rao S, Santhakumar AB, Chinkwo KA, Wu G, Johnson SK, Blanchard CL. 2018. Characterization of phenolic compounds and antioxidant activity in sorghum grains. *Journal of Cereal Science* **84**: 103-111.

Rhodes DH, Hoffmann Jr L, Rooney WL, Ramu P, Morris GP, Kresovich S. 2014. Genome-wide association study of grain polyphenol concentrations in global sorghum [Sorghum bicolor (L.) Moench] germplasm. *Journal of agricultural and food chemistry* **62**: 10916-10927.

Ribbenstedt A, Ziarrusta H, Benskin JP. 2018. Development, characterization and comparisons of targeted and non-targeted metabolomics methods. *PloS one* **13**: e0207082.

Salgado AL, Glassmire AE, Sedio BE, Diaz R, Stout MJ, Čuda J, Pyšek P, Meyerson LA, Cronin JT. 2023. Metabolomic evenness underlies intraspecific differences among lineages of a wetland grass. *Journal of Chemical Ecology* **49**: 437-450.

Salunkhe D, Jadhav S, Kadam S, Chavan J, Luh B. 1983. Chemical, biochemical, and biological significance of polyphenols in cereals and legumes. *CRC Critical Reviews in Food Science and Nutrition* **17**: 277-305.

Sansone S-A, Fan T, Goodacre R, Griffin JL, Hardy NW, Kaddurah-Daouk R, Kristal BS, Lindon J, Mendes P, Morrison N et al. 2007. The Metabolomics Standards Initiative. *Nature Biotechnology* **25**: 846-848.

Santure AW, Garant D. 2018. Wild GWAS—association mapping in natural populations. *Molecular ecology resources* **18**: 729-738.

Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M. 2014. Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant physiology* **165**: 1120-1132.

Sawada Y, Akiyama K, Sakata A, Kuwahara A, Otsuki H, Sakurai T, Saito K, Hirai MY. 2009. Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant and Cell Physiology* **50**: 37-47.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**: 2498-2504.

Siddique I. 2021. Unveiling the power of high-performance liquid chromatography: Techniques, applications, and innovations. *European Journal of Advances in Engineering and Technology* **8**: 79-84.

Simeone MLF, Parrella RA, Schaffert RE, Damasceno CM, Leal MC, Pasquini C. 2017. Near infrared spectroscopy determination of sucrose, glucose and fructose in sweet sorghum juice. *Microchemical Journal* **134**: 125-130.

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry* **78**: 779-787.

Spicer RA, Salek R, Steinbeck C. 2017. A decade after the metabolomics standards initiative it's time for a revision. *Scientific Data* **4**: 1-3.

Su XiaoYu SX, Rhodes D, Xu JingWen XJ, Chen Xi CX, Davis H, Wang DongHai WD, Herald TJ, Wang WeiQun WW. 2017. Phenotypic diversity of anthocyanins in sorghum accessions with various pericarp pigments.

Sytar O, Zivcak M, Neugart S, Toutounchi PM, Brestic M. 2019. Precultivation of young seedlings under different color shades modifies the accumulation of phenolic compounds in Cichorium leaves in later growth phases. *Environmental and Experimental Botany* **165**: 30-38.

Tanwar R, Panghal A, Chaudhary G, Kumari A, Chhikara N. 2023. Nutritional, phytochemical and functional potential of sorghum: A review. *Food Chemistry Advances* **3**: 100501.

Team G, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, Sifre L, Rivière M, Kale MS, Love J. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:240308295*.

Thakur NR, Gorthy S, Vemula A, Odeny DA, Ruperao P, Sargar PR, Mehtre SP, Kalpande HV, Habyarimana E. 2024. Genome-wide association study and expression of candidate genes for Fe and Zn concentration in sorghum grains. *Scientific Reports* **14**: 12729.

Thitisaksakul M, Jiménez RC, Arias MC, Beckles DM. 2012. Effects of environmental factors on cereal starch biosynthesis and composition. *Journal of cereal science* **56**: 67-80.

Tiozon RJN, Sreenivasulu N, Alseekh S, Sartagoda KJD, Usadel B, Fernie AR. 2023. Metabolomics and machine learning technique revealed that germination enhances the multi-nutritional properties of pigmented rice. *Communications biology* **6**: 1000.

Vanamala JK, Massey AR, Pinnamaneni SR, Reddivari L, Reardon KF. 2018. Grain and sweet sorghum (Sorghum bicolor L. Moench) serves as a novel source of bioactive compounds for human health. *Critical Reviews in Food Science and Nutrition* **58**: 2867-2881.

Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. 2005. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research* **33**: D433-D437.

Wang Y, Wang X, Sun S, Jin C, Su J, Wei J, Luo X, Wen J, Wei T, Sahu SK. 2022. GWAS, MWAS and mGWAS provide insights into precision agriculture based on genotype-dependent microbial effects in foxtail millet. *Nature Communications* **13**: 5913.

Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J. 2014. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature communications* **5**: 3438.

Wickham H. 2011. ggplot2. *Wiley interdisciplinary reviews: computational statistics* **3**: 180-185.

Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, Dizon R, Sayeeda Z, Tian S, Lee BL. 2022. HMDB 5.0: the human metabolome database for 2022. *Nucleic acids research* **50**: D622-D631.

Wu G, Johnson SK, Bornman JF, Bennett SJ, Clarke MW, Singh V, Fang Z. 2016. Growth temperature and genotype both play important roles in sorghum grain phenolic composition. *Scientific reports* **6**: 21835.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**: 82-93.

Wu Y, Li X, Xiang W, Zhu C, Lin Z, Wu Y, Li J, Pandravada S, Ridder DD, Bai G. 2012. Presence of tannins in sorghum grains is conditioned by different natural alleles of Tannin1. *Proceedings of the National Academy of Sciences* **109**: 10281-10286.

Xu Z, Dai J, Kang T, Shah K, Li Q, Liu K, Xing L, Ma J, Zhang D, Zhao C. 2022. PpePL1 and PpePL15 are the core members of the pectate lyase gene family involved in peach fruit ripening and softening. *Frontiers in Plant Science* **13**: 844055.

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X. 2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics and Bioinformatics* **19**: 619-628.

Zhang R, Jia G, Diao X. 2023a. geneHapR: an R package for gene haplotypic statistics and visualization. *BMC bioinformatics* **24**: 199.

Zhang T, Lu L, Yang N, Fisk ID, Wei W, Wang L, Li J, Sun Q, Zeng R. 2023b. Integration of hyperspectral imaging, non-targeted metabolomics and machine learning for vigour prediction of naturally and accelerated aged sweetcorn seeds. *Food Control* **153**: 109930.

Zhang Y, Li M, Gao H, Wang B, Tongcheng X, Gao B, Yu L. 2019. Triacylglycerol, fatty acid, and phytochemical profiles in a new red sorghum variety (Ji Liang No. 1) and its antioxidant and anti-inflammatory properties. *Food Science & Nutrition* **7**: 949-958.

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics* **42**: 355-360.

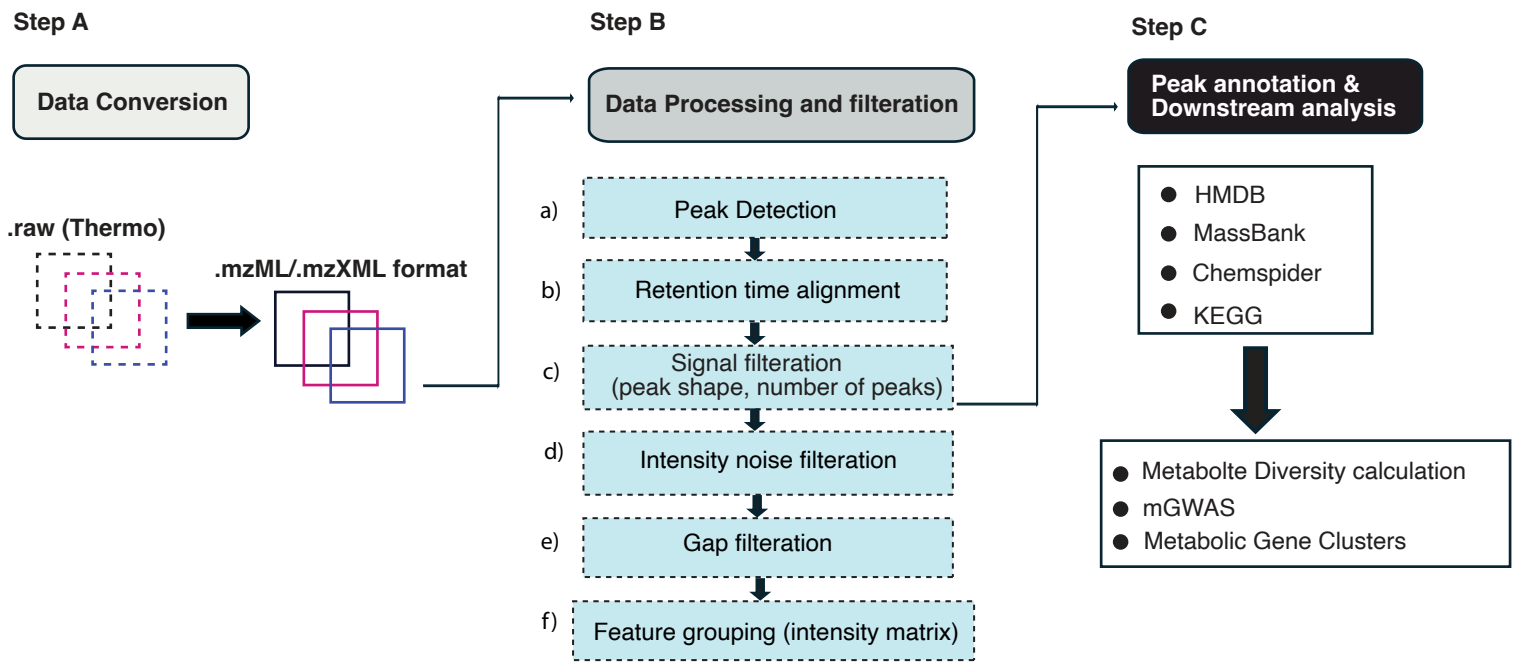
Zhao J, Xu Y, Li H, An W, Yin Y, Wang B, Wang L, Wang B, Duan L, Ren X. 2024. Metabolite-based genome-wide association studies enable the dissection of the genetic bases of flavonoids, betaine and spermidine in wolfberry (Lycium). *Plant Biotechnology Journal* **22**: 1435-1452.

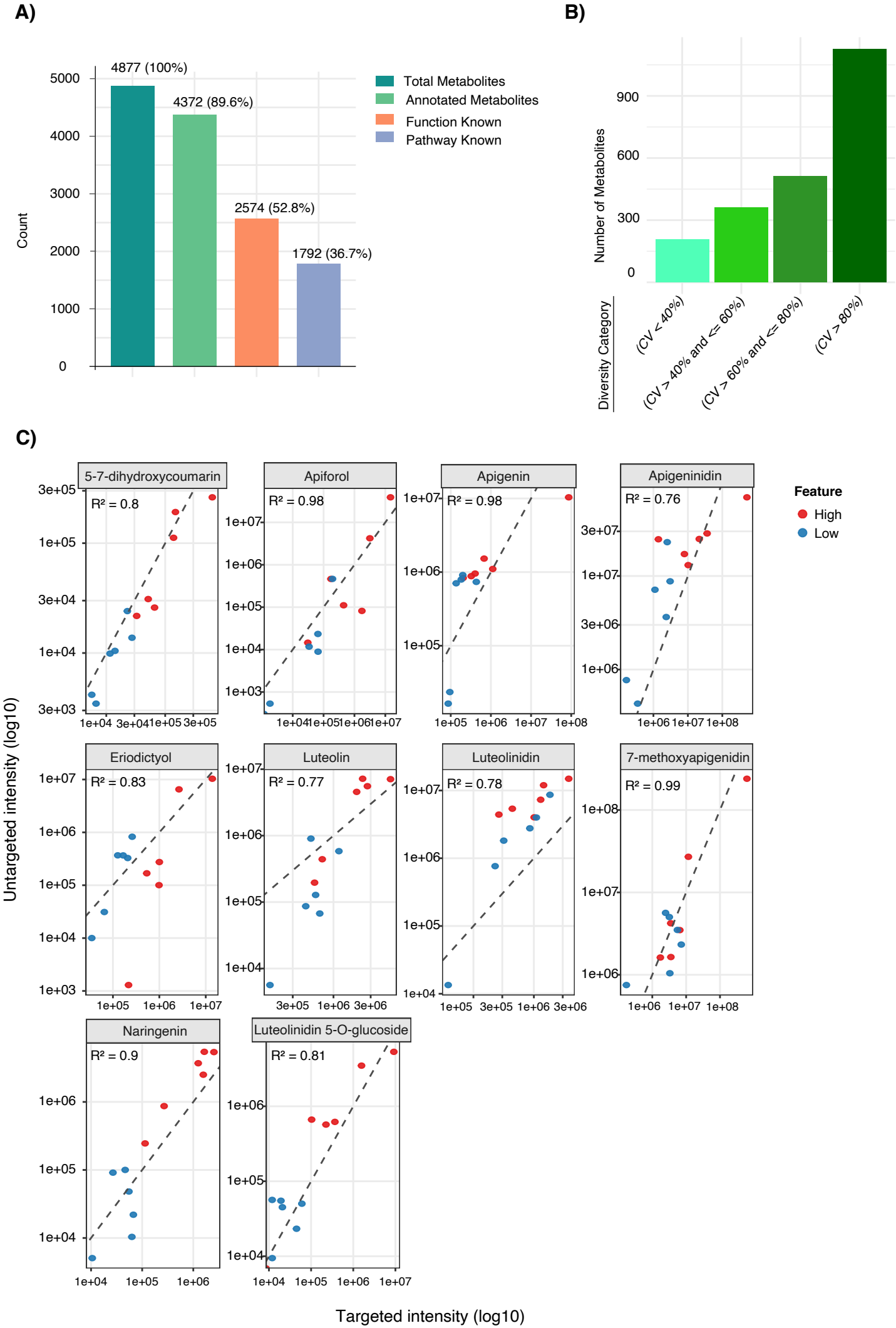
Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, Liu T-F, Jiang S, Ramachandran S, Liu C-M. 2011. Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome biology* **12**: 1-15.

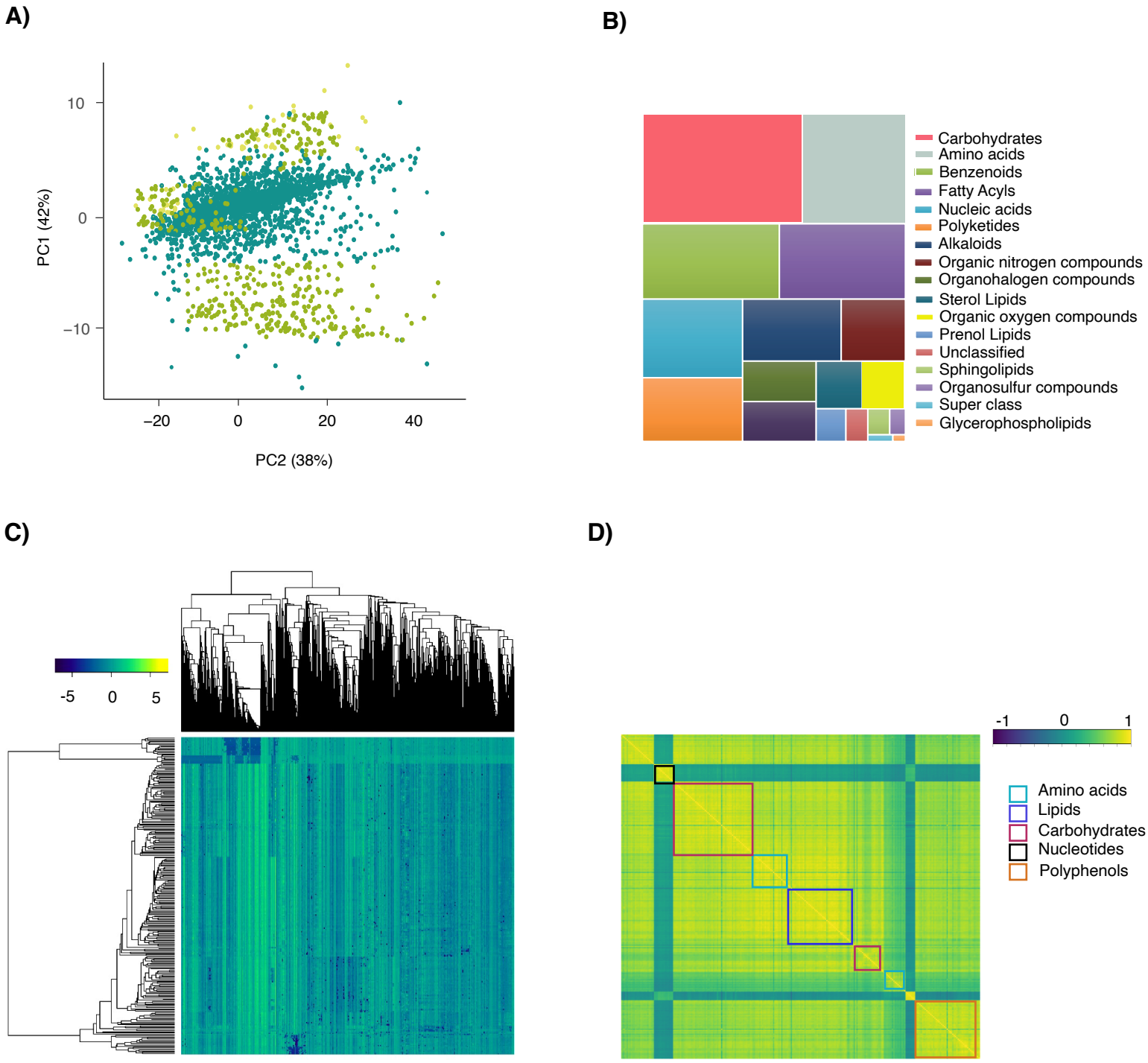
Zhou S, Kremling KA, Bandillo N, Richter A, Zhang YK, Ahern KR, Artyukhin AB, Hui JX, Younkin GC, Schroeder FC. 2019. Metabolome-scale genome-wide association studies reveal chemical diversity and genetic control of maize specialized metabolites. *The Plant Cell* **31**: 937-955.

Zhu A, Liu M, Tian Z, Liu W, Hu X, Ao M, Jia J, Shi T, Liu H, Li D. 2024. Chemical-tag-based semi-annotated metabolomics facilitates gene identification and specialized metabolic pathway elucidation in wheat. *The Plant Cell* **36**: 540-558.

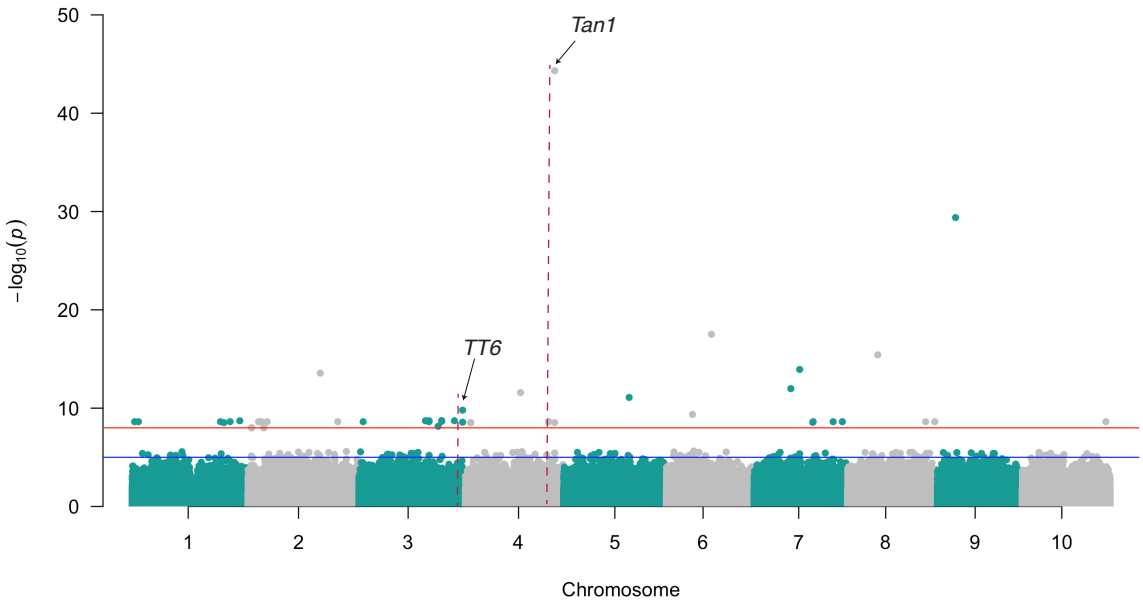
Zhu X, Galili G. 2003. Increased lysine synthesis coupled with a knockout of its catabolism synergistically boosts lysine content and also transregulates the metabolism of other amino acids in Arabidopsis seeds. *The Plant Cell* **15**: 845-853.


Step A
Data Conversion
.raw (Thermo)
.mzML/.mzXML format
Step B
Data Processing and filteration
a)
Peak Detection
b)
Retention time alignment
c)
Signal filteration
(peak shape, number of peaks)
d)
Intensity noise filteration
e)
Gap filteration
f)
Feature grouping (intensity matrix)
Step C
Peak annotation &
Downstream analysis
HMDB
MassBank
Chemspider
KEGG
Metabolte Diversity calculation
mGWAS
Metabolic Gene Clusters

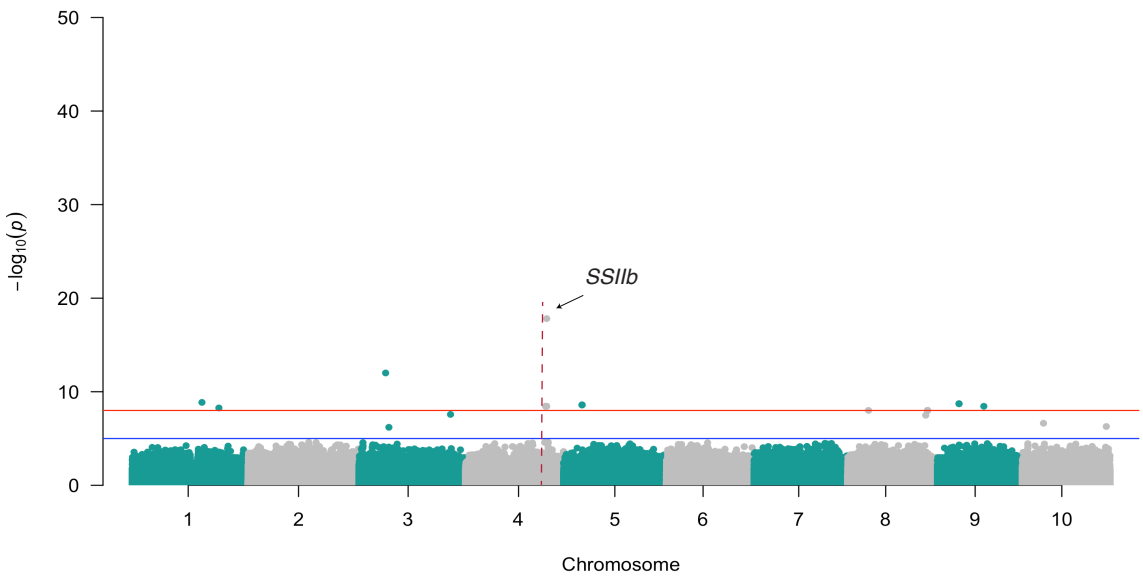

A)
Count
4877 (100%)
4372 (89.6%)
2574 (52.8%)
1792 (36.7%)
Total Metabolites
Annotated Metabolites
Function Known
Pathway Known
B)
Number of Metabolites
Diversity Category
(CV < 40%)
(CV > 40% and <= 60%)
(CV > 60% and <= 80%)
(CV > 80%)
C)
5-7-dihydroxycoumarin
R² = 0.8
Apiforol
R² = 0.98
Apigenin
R² = 0.98
Apigeninidin
R² = 0.76
Eriodictyol
R² = 0.83
Luteolin
R² = 0.77
Luteolinidin
R² = 0.78
7-methoxyapigenidin
R² = 0.99
Naringenin
R² = 0.9
Luteolinidin 5-O-glucoside
R² = 0.81
Feature
High
Low
Untargeted intensity (log10)
Targeted intensity (log10)


A)
10
0
−10
PC1 (42%)
−20
0
20
40
PC2 (38%)
B)
Carbohydrates
Amino acids
Benzenoids
Fatty Acyls
Nucleic acids
Polyketides
Alkaloids
Organic nitrogen compounds
Organohalogen compounds
Sterol Lipids
Organic oxygen compounds
Prenol Lipids
Unclassified
Sphingolipids
Organosulfur compounds
Super class
Glycerophospholipids
C)
-5
0
5
D)
-1
0
1
Amino acids
Lipids
Carbohydrates
Nucleotides
Polyphenols

A)


Tan1
TT6
50
40
30
20
10
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
Chromosome

B)


SSIIb
50
40
30
20
10
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
Chromosome

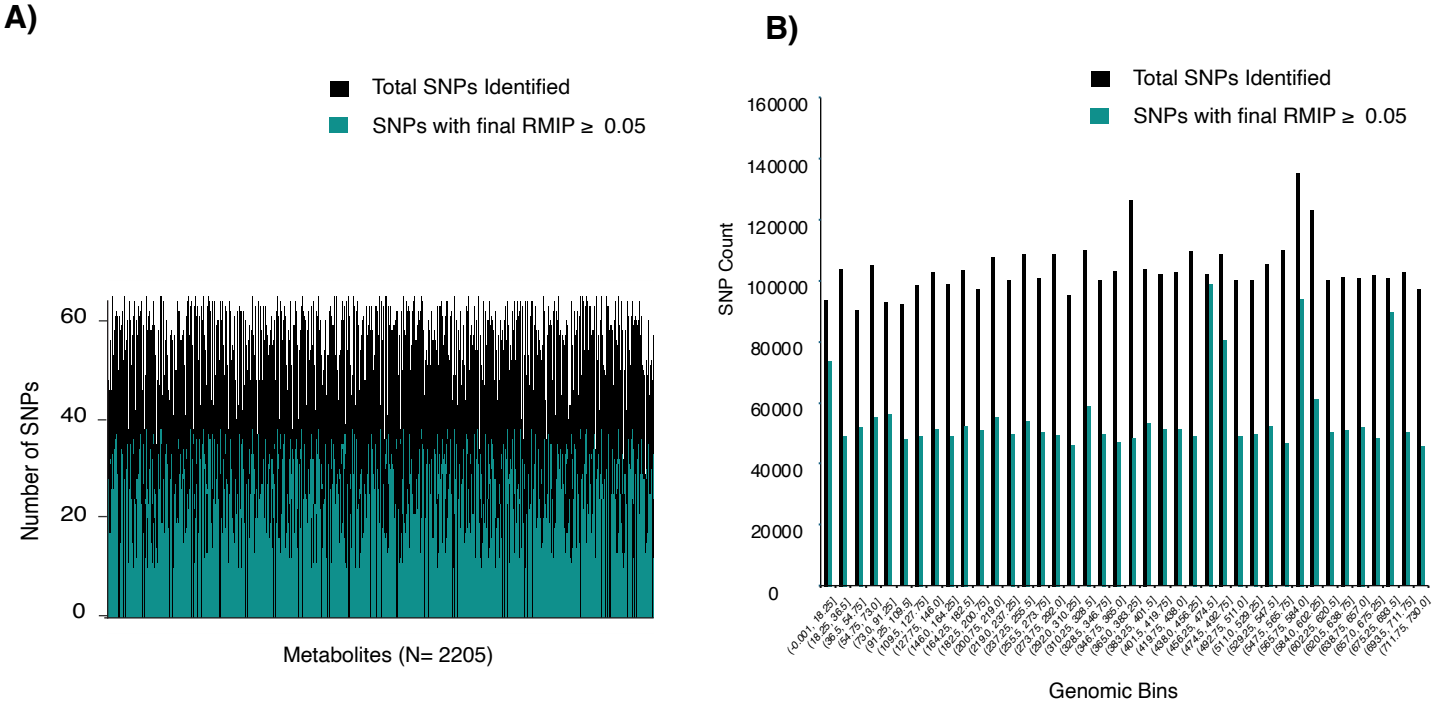
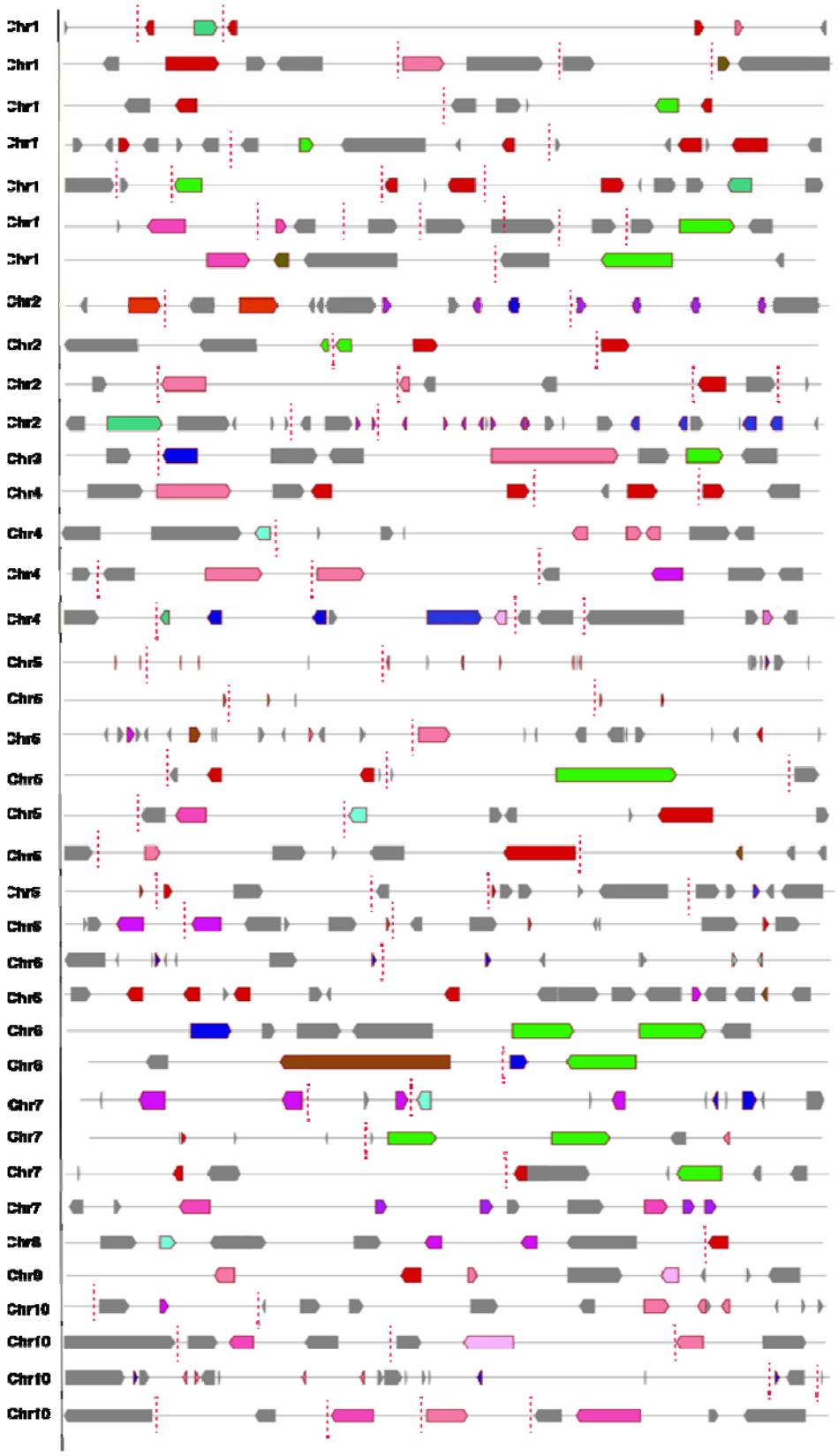
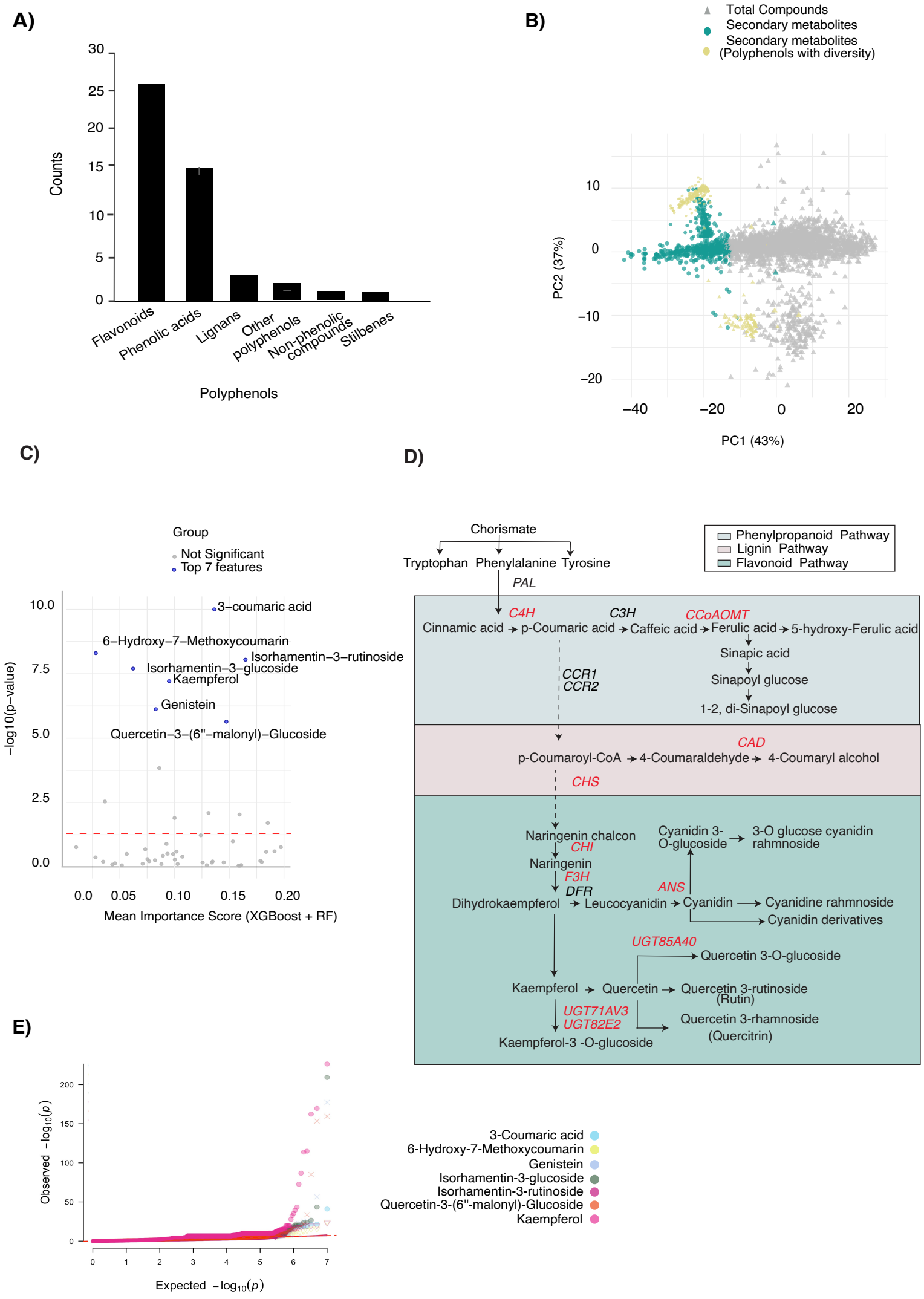

A)
Total SNPs Identified
SNPs with final RMIP ≥ 0.05
Number of SNPs
0
20
40
60
Metabolites (N= 2205)
B)
Total SNPs Identified
SNPs with final RMIP ≥ 0.05
SNP Count
0
20000
40000
60000
80000
100000
120000
140000
160000
Genomic Bins

Fig. S5. Genomic Distribution of Metabolic Gene Clusters and SNP Variation in *Sorghum bicolor.* Metabolic gene clusters (total 38) in the Sorghum bicolor genome, labeled from cluster 1 to cluster 38, along with genomic variation associated with metabolites via mGWAS. The red dotted vertical line indicates the location of identified SNPs. On the right-hand side, the figure shows the location and size of each gene cluster within the genome.

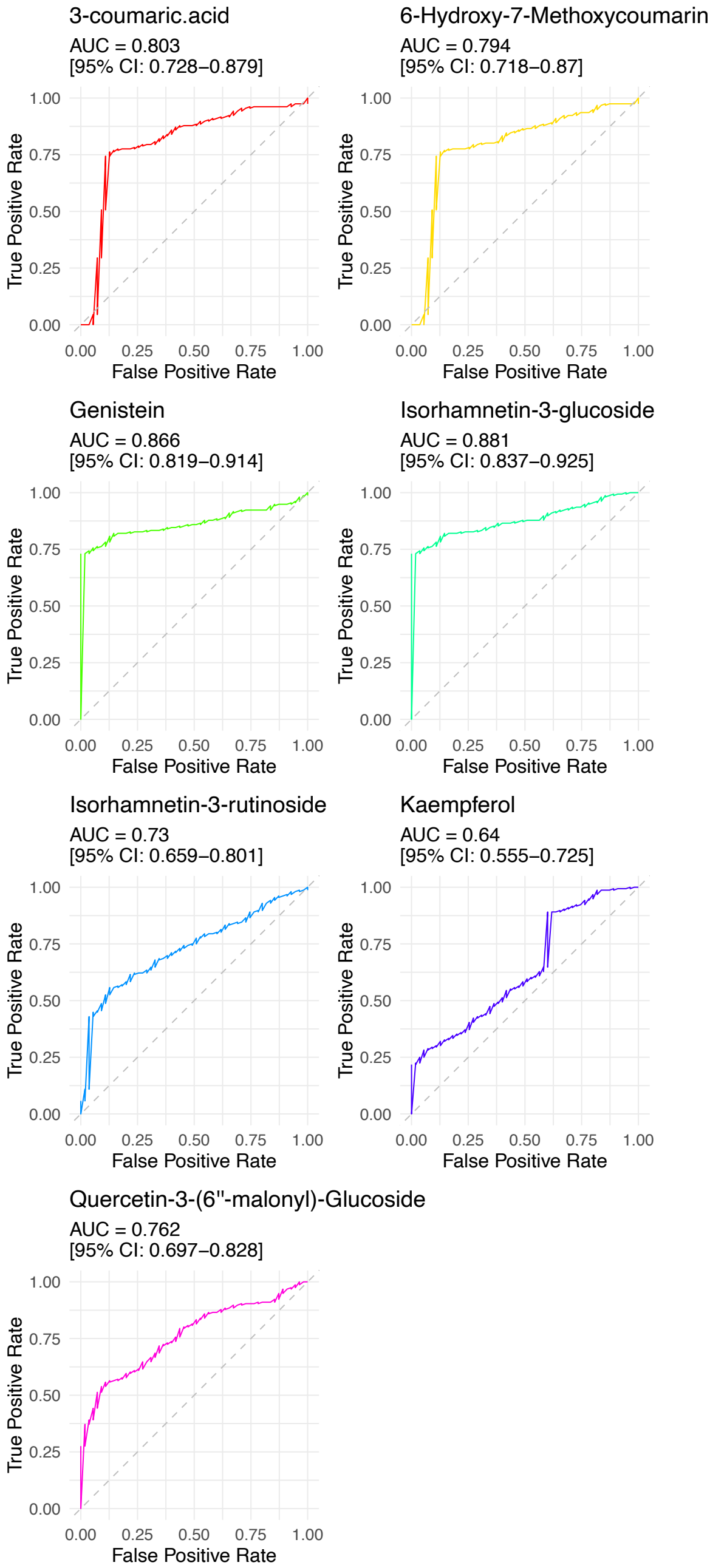


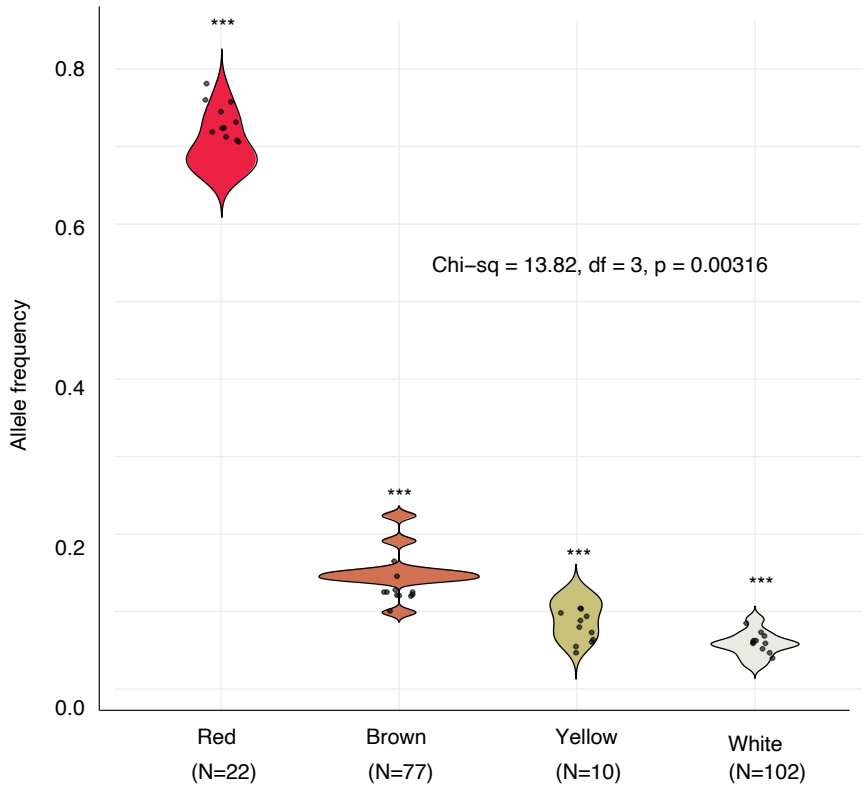
	Location From	Location To	Cluster Size (kb)
Chr1	1016097	1169880	153.78
Chr1	3292513	3330439	37.93
Chr1	17685101	17819900	134.9
Chr1	65226896	65347607	120.71
Chr1	65755901	65855510	99.61
Chr1	66043026	66133935	90.91
Chr1	69633872	69682476	48.8
Chr2	3947713	4047444	99.73
Chr2	6462489	6510841	47.85
Chr2	57443829	57560934	117.31
Chr2	61868057	62070583	202.53
Chr3	68380815	68441563	60.75
Chr4	5564025	5628714	64.69
Chr4	7283793	7358302	74.51
Chr4	58633369	58683790	50.42
Chr4	66262526	66361329	78.6
Chr5	17258080	18118109	860.04
Chr5	5559616	5594836	35.2
Chr5	62696059	62995686	299.63
Chr5	64393393	64529531	136.14
Chr5	64724026	64810990	86.96
Chr5	67564361	67642514	78.13
Chr5	69188562	69370262	183.7
Chr5	70387619	70627798	240.18
Chr6	3318320	3556279	237.96
Chr6	48791658	48872417	80.73
Chr6	56050196	56098097	47.9
Chr6	58634377	58691552	57.17
Chr7	6002761	6236433	233.67
Chr7	13853873	14289676	435.8
Chr7	55648377	55821589	173.19
Chr7	59141972	59209470	67.5
Chr8	1131970	1195454	63.48
Chr9	6842073	6900598	58.62
Chr10	1791465	1884189	92.71
Chr10	5451172	5490963	39.79
Chr10	51662049	51943022	280.97
Chr10	55408700	55435652	26.95

- Cytochrome 450
- Glycosyltransferase
- Oxidoredustase
- Other biosynthetic genes
- Other genes
- Co-esterase
- BAHD Acyltranferase
- Amino-transferase
- Dioxygenase
- Strictosidine synthase like
- Ketosynthase
- Methyltransferase
- Scl acyltransferase


A)
Counts
30
25
20
15
10
5
0
Flavonoids
Phenolic acids
Lignans
Other polyphenols
Non-phenolic compounds
Stilbenes
Polyphenols
B)
Total Compounds
Secondary metabolites
Secondary metabolites (Polyphenols with diversity)
PC2 (37%)
10
0
−10
−20
−40
−20
0
20
PC1 (43%)
C)
Group
Not Significant
Top 7 features
−log10(p−value)
10.0
7.5
5.0
2.5
0.0
3−coumaric acid
6−Hydroxy−7−Methoxycoumarin
Isorhamentin−3−rutinoside
Isorhamentin−3−glucoside
Kaempferol
Genistein
Quercetin−3−(6''−malonyl)−Glucoside
0.0
0.05
0.10
0.15
0.20
Mean Importance Score (XGBoost + RF)
D)
Chorismate
Tryptophan
Phenylalanine
Tyrosine
PAL
Phenylpropanoid Pathway
Lignin Pathway
Flavonoid Pathway
C4H
C3H
CCoAOMT
Cinnamic acid
p-Coumaric acid
Caffeic acid
Ferulic acid
5-hydroxy-Ferulic acid
Sinapic acid
Sinapoyl glucose
1-2, di-Sinapoyl glucose
CCR1
CCR2
CAD
p-Coumaroyl-CoA
4-Coumaraldehyde
4-Coumaryl alcohol
CHS
Naringenin chalcon
CHI
Naringenin
F3H
DFR
Cyanidin 3-O-glucoside
3-O glucose cyanidin rahmnoside
ANS
Dihydrokaempferol
Leucocyanidin
Cyanidin
Cyanidine rahmnoside
Cyanidin derivatives
UGT85A40
Quercetin 3-O-glucoside
Kaempferol
Quercetin
Quercetin 3-rutinoside (Rutin)
UGT71AV3
UGT82E2
Quercetin 3-rhamnoside (Quercitrin)
Kaempferol-3 -O-glucoside
E)
Observed −log10(p)
200
150
100
50
0
0
1
2
3
4
5
6
7
Expected −log10(p)
3-Coumaric acid
6-Hydroxy-7-Methoxycoumarin
Genistein
Isorhamentin-3-glucoside
Isorhamentin-3-rutinoside
Quercetin-3-(6''-malonyl)-Glucoside
Kaempferol

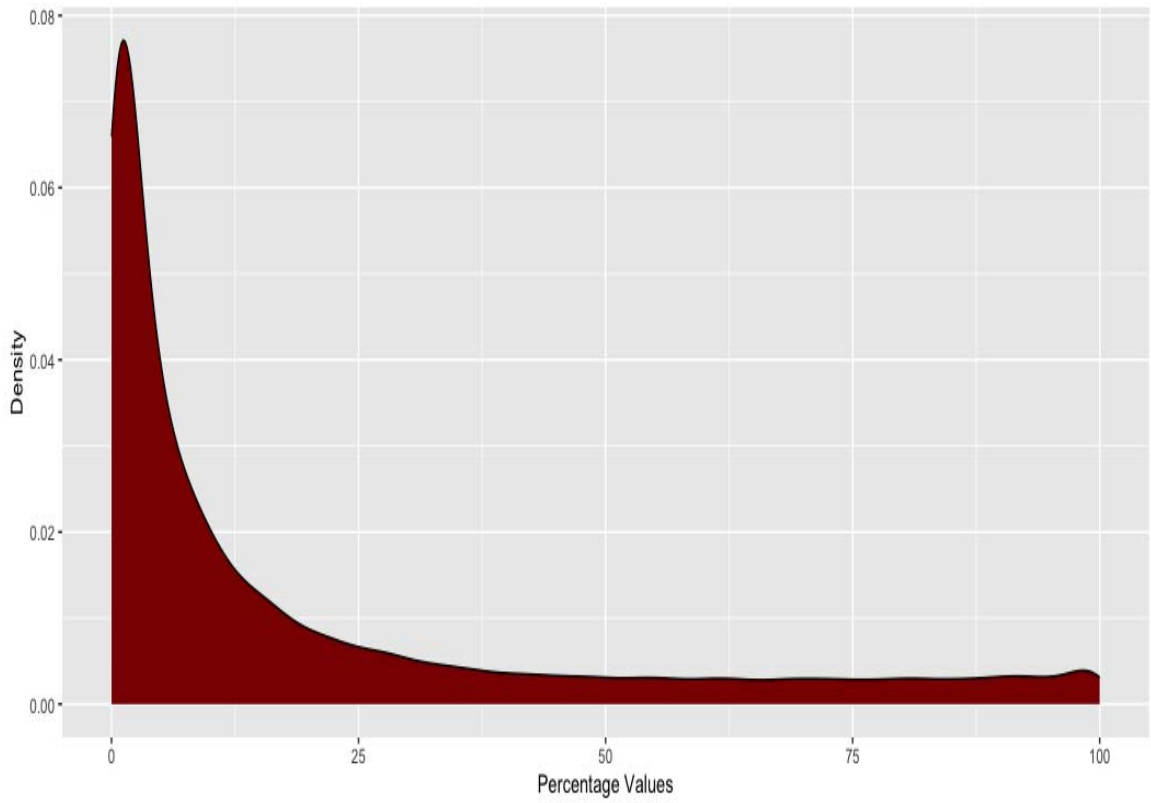
ROC Curves for 7 Polyphenols with AUC and 95% CI




Allele frequency
0.0
0.2
0.4
0.6
0.8
***
***
***
***
Chi−sq = 13.82, df = 3, p = 0.00316
Red
(N=22)
Brown
(N=77)
Yellow
(N=10)
White
(N=102)

Supplementary fig S11.

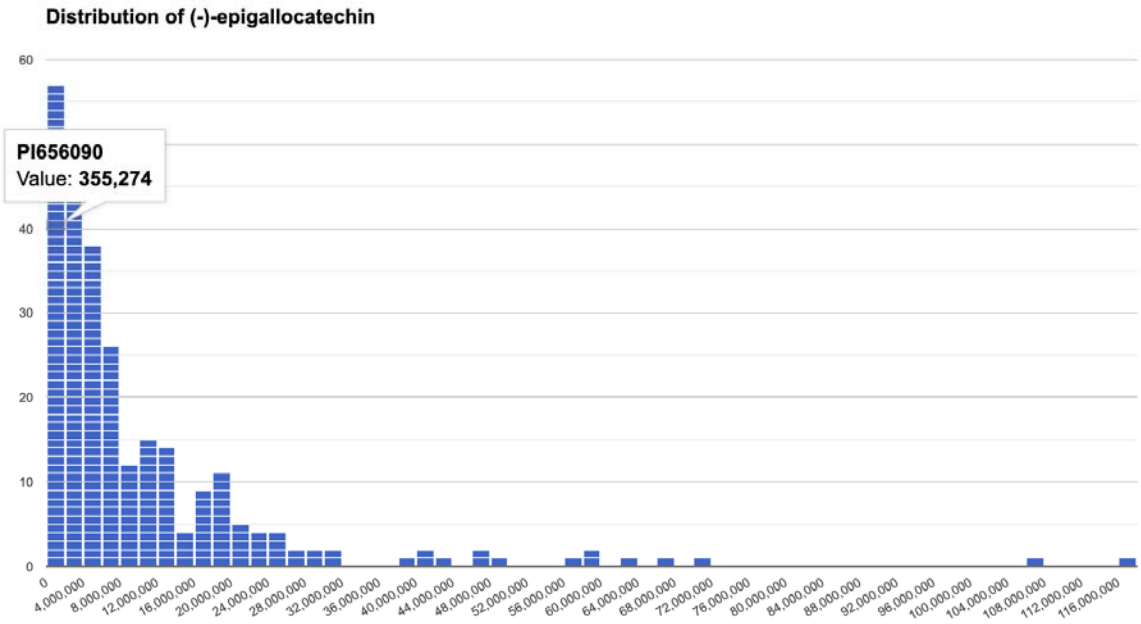
Distribution and Coverage of Total Metabolites Identified in Sorghum Accessions and Phenotypic Variants (SAP)



Coverage (%) is SAP population	Total count of metabolites
>50%	33198
>60%	26590
>70%	20041
>80%	12853
>90%	4877
>95%	2471
100%	268

Supplementary Fig. S10:

A) Distribution of (-)-epigallocatechin across GRIN-based SAP line numbers, with corresponding abundance values measured by UPLC-MS, obtained using SorGMDA.



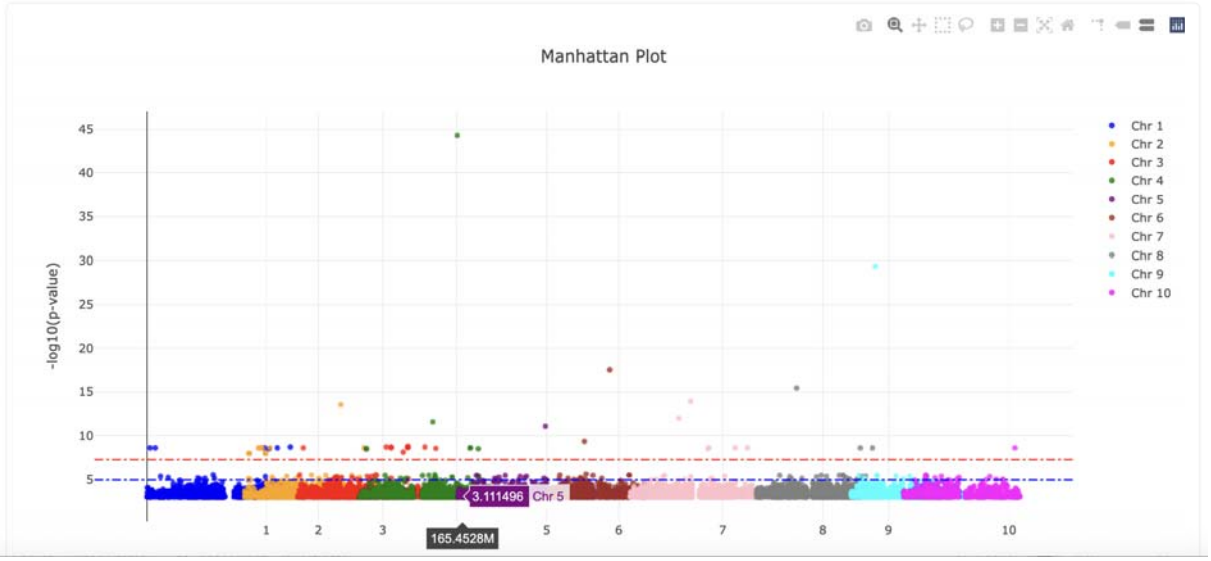
B. Manhattan plot of genome-wide association study (GWAS) results for (–)-epigallocatechin accumulation in the SAP population. The plot displays -log□□(p-value) for each SNP across all sorghum chromosomes (Chr 1–10), with alternating colors used to distinguish between chromosomes. Only associations with p-values ≤ 10□³ are shown. Each dot represents a single nucleotide polymorphism (SNP), with the vertical axis indicating the statistical significance of association and the horizontal axis representing genomic position. A strong association signal is observed on chromosome 3, with a peak at position 165.4528 Mb showing a –log□□(p-value) of 45. The red and blue dashed lines represent suggestive and genome-wide significance thresholds, respectively.

» Page: Home 2 3

GWAS Manhattan Plot

Search compound... Search **Select Table:** (-)-epigallocatechin Generate Plot Download Plot

Download CSV Note: Our database is only showing $p \times 10^{-3}$



Supplementary Table 1. Consensus polyphenol features identified by both XGBoost and Random Forest that distinguish grain color groups (red, brown, yellow, and white) in GRIN-selected lines, along with corresponding importance scores.

Compound	Score		Importance	p-value	$-\log_{10}$(p-value)	Consensus Feature
3-coumaric acid	0.084	0.072	0.078	0.002	2.7	Yes
6-Hydroxy-7-Methoxycoumarin	0.066	0.054	0.06	0.015	1.82	Yes
Genistein	0.091	0.077	0.084	0.001	3	Yes
Isorhamnetin-3-glucoside	0.074	0.062	0.068	0.008	2.1	Yes
Isorhamnetin-3-rutinoside	0.079	0.069	0.074	0.009	2.05	Yes
Kaempferol	0.096	0.083	0.09	0.0005	3.3	Yes
Quercetin-3-(6"-malonyl)-Glucoside	0.088	0.071	0.08	0.003	2.52	Yes

Note: Importance scores are normalized and represent the relative contribution of each compound in distinguishing between pigment groups. Adjusted p-values were calculated using [statistical test method, ANOVA], and multiple testing correction was applied using the Benjamini-Hochberg method. The **$-\log_{10}$(p-value)** is used for visualization, where higher values indicate stronger statistical significance.