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

Background

Cereal grains are a cornerstone of global food and feed security, as well as a key raw material for bioenergy production. Dissecting the molecular networks and metabolic pathways that shape grain composition is critical for advancing agricultural and nutritional sciences. Sorghum (*Sorghum bicolor*), a climate-resilient C4 cereal with exceptional tolerance to heat and drought, is an important model for metabolic genomics. However, the genetic basis of natural variation in sorghum grain metabolites, key determinants of nutritional value and end-use quality, remains largely unexplored.

Results

Our integrative metabolomic, genomic analyses and machine learning reveal extensive metabolic diversity in sorghum grain and its underlying genetic architecture. We performed large-scale untargeted metabolomic profiling of mature grains from the Sorghum Association Panel, detecting 4,877 metabolites with extensive variability across accessions. More than 36% of metabolites, particularly those linked to amino acid and polyphenol biosynthesis, exhibited high natural diversity, underscoring sorghum's potential for nutritional enhancement. Metabolite-based genome-wide association studies (mGWAS) identified approximately 4.15 million significant SNP–metabolite associations enriched in regulatory regions, thereby revealing the genetic architecture of metabolic diversity. We further identified 38 metabolite–gene clusters, highlighting coordinated regulation of key pathways. By integrating machine learning, we uncovered major metabolites determining grain color and pinpointed associated genes, demonstrating predictive frameworks for phenotype-genotype-metabolite relationships.

Conclusions

This study provides the first population-scale map of sorghum grain metabolomic and genetic diversity. To enable broad access and translational use, we established the Sorghum Grain Metabolite Diversity Atlas (SorGMDA), an open database integrating metabolomic and genomic variation. Together, these resources offer a foundation for comparative metabolic genomics across cereals, enable systems-level dissection of metabolic regulation, and support breeding of nutrient-dense, climate-resilient grains to address global food and agriculture challenges.

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[1, 2]. Sorghum is cultivated for human food, animal feed, and bioenergy production, but its full nutritional potential remains unrealized amid escalating global food demands[3]. Sorghum is a gluten-free carbohydrate source rich in bioactive metabolites such as phenolic acids, flavonoids, and tannins. Each of these molecules contributes to its high antioxidant capacity and potential health benefits, including anti-inflammatory[4, 5] and anti-carcinogenic properties[6, 7]. These nutritional attributes uniquely position sorghum as a low-input crop that can be leveraged to address both global food security and dietary health challenges, particularly in gluten-sensitive populations[8].

Sorghum accessions exhibit remarkable genetic and phenotypic diversity[9] in the concentration of bioactive metabolites in their grain [10, 11]. For instance, variation in grain color and fiber content is linked to polyphenol levels that can modulate gut microbiota composition[12-14]. As a result, enhancing sorghum's nutrient content and bioavailability has become a key target in crop improvement programs aiming to meet consumer demand for functional, health-promoting foods[15-17].

Recent advances in sorghum breeding underscore the importance of grain bioactive metabolites[18-20], which are increasingly recognized for their dual roles in promoting human health[21] and enhancing crop resilience and nutritional quality[22]. For instance, phenolic metabolites not only support human health by reducing the risk of chronic diseases-including diabetes, obesity, cancer, and cardiovascular disease[18, 23-26], but also play key roles in plant defense against biotic and abiotic stresses by detoxification of Reactive Oxygen Species (ROS)[27] [28]. Elucidating the genetic basis of metabolite variation is critical for

improving sorghum grain quality through molecular breeding and genetic engineering[29]. Each of these approaches will allow targeted manipulation of metabolic pathways to breed sorghum varieties with enhanced nutritional profiles, advancing the sustainability of agriculture and public health.

To facilitate these efforts, 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[30, 31], protein content[32], and micronutrient diversity[33]. By leveraging high-throughput genotyping and phenotyping technologies, GWAS has identified numerous genetic loci associated with traits like amylose-amylopectin ratios[34], amino acid profiles[35], and concentrations of essential micronutrients such as iron and zinc[36, 37]. 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[38], 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[39], 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[40]. Similarly, protein content and micronutrient diversity are influenced by genetic background, polygenic interactions, and environmental factors, complicating the identification of causal variants[41, 42]. 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[43, 44].

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[45-47]. 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. Their proportionate expression across developmental stages captures a dynamic snapshot of the plant's physiological state and serves as a crucial bridge between genotype and phenotype[48]. This integrative approach to data analysis enables the identification of genetic variants that regulate the diversity and accumulation of metabolites in plant tissues. In addition, the metabolome offers valuable insights into the interplay between genetic and environmental factors, helping to elucidate the genetic basis of complex traits that influence crop performance and nutritional quality[49].

Recent advances in untargeted metabolome profiling utilizing high-performance liquid chromatography (HPLC) coupled with a Q Exactive Orbitrap mass spectrometer[50] have revolutionized our ability to characterize complex metabolite mixtures with ultra-high resolution and mass accuracy[51], enabling the effective separation and identification of structurally similar isomers[52, 53]. Unlike targeted metabolomics, which focuses on quantifying a predefined set of known metabolites, untargeted approaches aim to capture a broad and unbiased spectrum of metabolites, including intermediary, unknown or unexpected compounds, making them particularly valuable for exploratory studies and hypothesis generation[54].

When integrated with high-throughput genotyping through metabolome genome-wide association studies (mGWAS), genetic variation is linked to metabolite profiles, clarifying the genetic basis of key traits. The rapid expansion of high-throughput omics (genomics, metabolomics, transcriptomics) and phenotyping technologies has generated vast, complex datasets for cereal crops like sorghum[46, 55]. However, extracting meaningful insights from these data requires advanced analytical approaches. Machine learning (ML) tools, such as random forests and neural networks, excel at identifying non-linear patterns and predicting complex traits (e.g., interactions between grain nutritional content, stress tolerance, and yield), accelerating trait discovery[56]. The ability of ML to support 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[57]. In wheat, the 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 metabolites as substrates[58]. In rice, mGWAS identified *Os02g57760*, a previously unknown nicotinic acid N-methyltransferase gene associated with grain size[49]. Similarly, in barley, mGWAS associated variation in fructan composition with specific metabolic pathways and known fructan biosynthesis genes, further connecting metabolite profiles to grain quality traits [59]. These studies underscore the utility of mGWAS in clarifying the genetic control of metabolism during seed development[60] and to dissect complex traits[61].

In addition, mGWAS have successfully identified significant loci associated with metabolites involved in a wide array of physiological processes[62, 63]. These processes include stress response mechanisms, secondary metabolite biosynthesis, and nutrient assimilation, which collectively influence crop resilience and quality. For instance, in rice, mGWAS identified 233 drought-responsive metabolites (DRMs) in 510 diverse accessions at the reproductive stage. A transferase gene was identified, which is responsible for the metabolic variation of N-feruloylputrescine, a metabolite linked to improved drought resistance[64]. 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. While mGWAS have provided significant insights into the genetic basis of nutritionally important metabolites in rice and maize[65, 66], the genetic basis of key grain metabolites, such as flavonoids and tannins, remains poorly characterized, limiting targeted improvement of nutritional quality.

The Sorghum Association Panel (SAP) includes genetically diverse accessions[9] with wide variation in grain quality traits[31], 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. 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 and provide a foundation for metabolite-informed breeding strategies aimed at enhancing grain quality and nutritional value in sorghum.

RESULTS

Metabolite Diversity in Sorghum Grain

To capture the metabolic diversity of sorghum grain represented in the SAP, we performed untargeted metabolomic profiling on mature grains from 266 accessions from SAP. Our annotation workflow combined high-confidence spectral matching, pathway-based classification, and multiple reference databases to enhance metabolite identification (Supplementary Fig.1). A total of 4,877 metabolites were identified, demonstrating high dataset coverage, with 3,527 detected in positive-mode electrospray ionization (ESI+) and 1,350 in negative-mode electrospray ionization (ESI−). Among these, 89.6% were annotated, 52.8% had known biological functions, and 36.7% were mapped to established metabolic pathways (Supplementary Fig.2A and Supplementary file 1). To evaluate the extent of variability among metabolites across the population, coefficient of variation (CV) analysis was conducted (Supplementary Fig.2B). The majority of metabolites exhibited high variability among genotypes (CV > 80%), accounting for 51.11% of the total metabolites, while intermediate variability among genotypes was observed in 23.17% (CV > 60% and ≤ 80%)

and 16.37% (CV > 40% and ≤ 60%) of metabolites. Only 9.34% displayed low variability among genotypes (CV<40%). Principal Component Analysis (PCA) further highlighted extensive genotypic diversity, with the first two components explaining 42% and 38% of the total variance, respectively (Supplementary Fig.3A, Supplementary file 2). 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%) being the most abundant (top five) categories (Supplementary Fig.3B). Metabolite profiling across accessions revealed extensive variation in abundance patterns (Supplementary Fig.3C), with distinct groups of accessions sharing similar metabolic signatures. Metabolites grouped by class, such as amino acids, lipids, and polyphenols, indicated coordinated accumulation patterns (Supplementary Fig.3D).

To quantify metabolic diversity across all SAP accessions, we applied the combined Shannon and Simpson Diversity Indices, widely used metrics that capture both the richness and evenness within the metabolite landscape [67, 68]. The analysis revealed that 36% of the detected metabolites showed high diversity (coefficient of variation, CV > 80%) across genotypes, indicating widespread variation. These metabolites also displayed high heritability, with estimates reaching up to 1 (Fig.1A, Supplementary File 2). A 99% confidence interval (CI) confirmed the statistical reliability of the observed diversity trends, i.e., the patterns of variation in metabolite levels across genotypes-highlighting the robustness of the findings. Diversity scores across genotypes varied by metabolite class, with amino acids, carbohydrates, nucleotides, polyphenols, polyketides, terpenoids, and organic acids each exhibiting distinct diversity patterns and median heritability values that reflect the typical genetic influence within each class (Fig.1B). Metabolites were stratified into three diversity categories-high, medium, and low-based on their diversity index across genotypes (Supplementary File 2). The most diverse (top 5%) metabolites across genotypes were enriched in biosynthetic pathways such as alanine, aspartate, and glutamate metabolism, which showed strong statistical significance (FDR = 1×10^{-1}; $-\log_{10}$[FDR] = 10) (Fig.1C). Additional highly variable metabolites were associated with lysine metabolism (FDR = 1×10^{-10}) and the pentose

phosphate pathway (FDR = 3×10^{-10}). 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).

Given the observed diversity in metabolite accumulation across genotypes, we next explored how metabolites with similar accumulation patterns are organized within the broader metabolic network. Metabolites exhibiting correlated accumulation patterns are often co-regulated within shared biosynthetic pathways, suggesting coordinated metabolic control[69]. To identify such co-regulated molecular clusters, we constructed correlation-based networks linking signals (nodes) with similar accumulation profiles at the $P \leq 0.01$ level of significance (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, Supplementary file 1). This approach enabled tentative structural identification based on exact mass, retention time (RT), and database matches (Supplementary file 1). The resulting network revealed distinct clusters associated with both primary (e.g., carbohydrates, organic acids, nucleic acids) and secondary metabolism (e.g., alkaloids, benzenoids, polyketides) (Fig.2). For instance, Cluster I (alkaloids) contained metabolites such as indoline (peak ID 38562), atropine (peak ID 167180), and quinolin-8-ol (peak ID 67541); and MS/MS reference spectra confirmed their nitrogen-containing amine structures (Supplementary File 1). 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 involvement in stress-related biochemical pathways, underscoring the potential of MS/MS-based clustering to reveal

unexpected biochemical relationships. These findings indicate a high degree of coordinated metabolite accumulation within discrete biochemical pathways, highlighting the role of tightly regulated metabolic networks in shaping sorghum's biochemical landscape and ultimately contributing to phenotypic diversity.

Genetic variation associated with metabolome diversity

To explore the genetic architecture underlying metabolic variation, used 38.8 million high-quality SNPs (minor allele frequency > 0.05) obtained from whole-genome resequencing of the SAP[9]. To assess the robustness of the mGWAS approach, we first evaluated two representative metabolites: (–)-epigallocatechin ($C_{15}H_{14}O_7$) and D-(–)-fructose ($C_6H_{12}O_6$), both of which were associated with significant SNPs (Supplementary Fig.4, Supplementary File 3). (–)-Epigallocatechin, a plant-derived polyphenol and condensed tannin, is known for its antioxidant properties and potential role in grain pathogen defense mechanisms[70, 71]. D-fructose, a simple sugar central to primary metabolism and energy storage, is a major component of the sorghum grain carbohydrate profile and yield [72, 73]. For (–)-epigallocatechin, mGWAS identified 47 significant SNPs ($P \leq 10\text{-}8$), including genetic variants located ~ 14 kb downstream of the well-characterized *Tannin1* (*Tan1*, *Sobic.004G280800*) gene on Chromosome 4[74] (Supplementary Fig.4A and Supplementary file 3). It encodes a WD40-repeat transcription factor that regulates the expression of genes in the proanthocyanidin (condensed tannin) biosynthetic pathway[74] and plays a critical role in seed coat pigmentation and pathogen defense[74]. While *Tan1* showed the strongest association with tannin content ($P = 4.85 \times 10^{-15}$), an additional signal was detected ~15 kb upstream of *TRANSPARENT TESTA 6* (*TT6*, *Sobic.003G418000*) on Chromosome 3 ($P = 1.60 \times 10^{-11}$), which encodes flavanone 3-hydroxylase (F3H), a 2-oxoglutarate-dependent Fe(II) oxygenase (Supplementary Fig.4A). F3H catalyzes the stereospecific hydroxylation of flavanones to dihydroflavonols (an essential step in flavonoid biosynthesis), as observed in *Arabidopsis thaliana*[75], rice[76] and maize[77]. For D-(–)-fructose, 29 significant SNPs were identified ($P \leq 10^{-8}$), including a major peak on Chromosome 4 located ~28kb

upstream of *Starch Synthase IIb* (*SSIIb, Sobic.006G221000*)43 (P = 1.53 × 10□18), a known regulator of starch biosynthesis (Supplementary Fig.4B and Supplementary file 3). Hence, the strong associations at biologically relevant loci for both metabolites confirmed the sensitivity and accuracy of our mGWAS framework.

A total of 2,206 metabolites that were detected in over 95% of SAP accessions and exhibited a normal distribution (Shapiro-Wilk test P > 0.05; within ±3 standard deviations) were retained for the mGWAS analysis (Supplementary Fig.5A). Using a significance threshold of P ≤ 1 × 10-8, we identified 4,151,451 unique SNPs, of which 2,221,701 showed robust associations (RMIP ≥ 0.05) (Supplementary Fig.5B, Supplementary file 4). 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.

Intriguingly, our genotypic diversity assessment identified lysine as one of the top 5% most diverse metabolites within the SAP (Fig.1C). The lysine-specific mGWAS (C□H□□N□O□) revealed 152 significant loci (P < 10□□) (Fig.3A, Supplementary File 3), 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 × 10□^{12}, *Sobic.003G412100*, P = 1.93 × 10□□ and *Sobic.002G155100*, P = 4.49 × 10□□), dihydrodipicolinate synthase (DHDPS, *Sobic.006G186600*, P = 3.00 × 10□□), and lysine decarboxylase (LDC, *Sobic.001G127700*, P = 3.93 × 10□□), 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 is essential in the lysine biosynthesis pathway, converting L-aspartate into L-aspartyl-4-phosphate, the first step in lysine production[78]. DHDPS catalyzes the conversion of L-aspartate-semialdehyde into (2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate, a key intermediate in the pathway known to influence lysine level in

maize[79], rice[80, 81] and *Camelina sativa* (camelina)[82]. LDC, on the other hand, catalyzes the conversion of L-lysine into cadaverine, a key intermediate in alkaloid synthesis that contributes to the complexity of plant metabolism by enabling the production of a diverse range of specialized metabolites with various biological functions[83]. AK converts L-aspartate into L-aspartyl-4-phosphate, a first step in lysine synthesis. DHDPS catalyzes the transformation of L-aspartate-semialdehyde into (2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate, an intermediate that makes the pathway energetically favorable. Finally, LDC catalyzes lysine degradation by transforming L-lysine into cadaverine, a versatile molecule that serves as a precursor for a variety of alkaloids, adding a layer of complexity to the plant's metabolic landscape. In addition to the known genes, the 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 ($P = 3.00 \times 10^{\square\square}$) (Fig.3A, Supplementary File 3). *PL-1* is known for its role in cell wall modification and fruit softening[84], suggesting a potential indirect link to lysine accumulation through modulation of cell wall-associated metabolic processes.

Genomic Hotspots for Metabolite Variation in Sorghum

We compared SNPs from two datasets: the input set of all genome-wide SNPs in the SAP lines and significant mGWAS SNPs to determine the enrichment of the mGWAS signals in specific genomic features. Comparison between the input SNPs (n = 38,873,896) and mGWAS SNPs (n = 4,151,451) 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; Supplementary File 5). While downstream variants constituted 18.0% of the input SNPs, their proportion increased to 20.6% among mGWAS SNPs. Similarly, upstream variants rose from 18.3% to 23.8% among mGWAS SNPs. In contrast, intergenic variants were markedly underrepresented, decreasing from 49.0% of input

SNPs to 24.7% of mGWAS SNPs. Fold-enrichment analysis using chi-square (χ^2) tests indicated significant overrepresentation of missense (7.1-fold; $P = 6.32 \times 10^{-2}$), synonymous (6.6-fold; $P = 8.61 \times 10^{-1}$), and UTR variants (~2-fold; 5′-UTR: $P = 7.27 \times 10$, 3′-UTR: $P = 1.07 \times 10^{1}$) (Fig.4C; Supplementary File 5). Modest enrichment was also observed for upstream, downstream, and intronic variants (<2-fold), whereas intergenic variants were significantly reduced (0.50-fold; log FC = -0.99). Positional analysis of non-genic SNPs revealed a pronounced enrichment near promoter regions of the nearest gene, suggesting potential cis-regulatory roles where variants may directly influence gene transcription. A secondary enrichment peak was observed approximately 100 kb from genes, likely representing distal enhancers that can regulate gene expression from a distance via. chromatin interactions (Fig.4D). Collectively, these patterns indicate that metabolite-associated variants are preferentially located in genic and regulatory regions. The reduced frequency of significant intergenic SNPs in the mGWAS suggests that metabolic diversity is largely shaped by loci with functionally relevant loci, rather than by distal non-coding regions without known regulatory roles.

Metabolic Gene Clusters (MGCs) Related to Sorghum Grain Nutrients

Metabolic Gene Clusters (MGCs) are groups of physically clustered genes that encode enzymes involved in the same, co-regulated metabolic pathway[85, 86]. A total of 38 gene clusters distributed across all 10 chromosomes were identified in *Sorghum bicolor* genome harboring 434 biosynthetic genes (Fig.5A and Supplementary Fig.6, Supplementary file 6). 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. The highest number of gene clusters (8 in total) were identified on chromosome 5, primarily associated with polyphenol metabolism (5 clusters), as well as amino acid (1) and carbohydrate metabolism (1), suggesting that this region may play a central role in secondary metabolism (Fig.5A; Supplementary File 6). In addition, smaller but distinct clusters were observed on other chromosomes, indicating that MGCs are distributed

throughout the genome (Fig.5A). Notably, these clusters were located within intergenic distances ranging from 26.95 kb to 806.04 kb (Fig.5A; Supplementary Fig.6; Supplementary File 6). Across 35 clusters, a total of 85 sequence variants were identified within ±50 kb of biosynthetic genes involved in these pathways, distributed across chromosomes 1, 2, 3, 4, 5, 6, 7, 8, and 10 (Supplementary Fig.6; Supplementary File 7).

Lignin is crucial for seed development[87]; it protects the seed [88] and enhances water regulation[89], longevity[90], and resistance to environmental stress[91, 92], insects, and pathogens. Further analysis of the lignin MGC cluster (No. 24) on Chromosome 5 identified variants associated with two polyphenols lignin precursors: 3-coumaric acid and quercetin-3-(6"-malonyl)-glucoside (Fig. 5B). These variations were located near two genes: *Sobic.005G218100* (P = 9.51 × 10□□), encoding a flavin-containing amine oxidoreductase with a methyltransferase domain, and *Sobic.005G218500* (P = 8.71 × 10□□), a NUCLEOPORIN-RELATED gene encoding a dirigent enzyme (Supplementary Fig. 6, Supplementary File 7). Both enzymes are known to be involved in lignin biosynthesis during seed development[93-96]. A total of 20 genes within this lignin MGC, located within ±50 kb of these variants-including methyltransferase loci and the 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 (Supplementary File 8). 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; Supplementary File 8). 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.

The Metabolic Basis of Complex Phenotypes: Kernel Color

Kernel color in sorghum and other cereals is a complex phenotype conferred by the morphologies of pericarp and endosperm tissues, as well as the accumulation of various classes of pigment-related metabolites, including flavonoids (e.g., anthocyanins), carotenoids, and phlobaphenes**[97]**. Machine learning approaches have proven especially useful in identifying complex, non-linear relationships between metabolite profiles and seed traits[98, 99]. To dissect the metabolic basis of kernel color, we applied machine learning to compare metabolomic profiles among 211 SAP accessions, categorized into four grain color groups based on GRIN kernel color codes: brown (77 lines), red (22 lines), yellow (10 lines), and white (102 lines) (Supplementary File 9).

Hierarchical clustering of metabolite abundances across kernel color groups revealed distinct metabolic signatures, including differences in the composition of 49 secondary metabolites (Fig.6A, Supplementary Fig.7A, Supplementary File 10). 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. A distinct cluster accounted for 43% of the variation among PC1 and 37% among PC2 (Supplementary Fig. 7B). Owing to their antioxidant properties, polyphenols also influence seed development, germination, growth, and overall quality[13, 100, 101].

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; Supplementary Table 1; Supplementary File 10; Supplementary Fig. 7C). ROC curve analysis evaluated the

discriminatory power of these metabolites for grain color classification. Among them, isorhamnetin-3-glucoside (AUC = 0.881, 95% CI: 0.837–0.925) and genistein (AUC = 0.866, 95% CI: 0.819–0.914) showed the strongest statistical separation between four grain color groups based on metabolite profiles. Moderate classification performance was observed for 3-coumaric acid (AUC = 0.803, 95% CI: 0.728–0.879), 6-hydroxy-7-methoxycoumarin (AUC = 0.794, 95% CI: 0.718–0.870), and quercetin-3-(6''-malonyl)-glucoside (AUC = 0.762, 95% CI: 0.697–0.828). In contrast, kaempferol showed limited discriminatory potential (AUC = 0.640, 95% CI: 0.555–0.725) (Supplementary Fig.8; Supplementary File 12).

The mGWAS analysis of these seven polyphenols identified 214 significant genetic variants ($P < 10^{-11}$) associated with grain color variation (Fig. 6C, Supplementary File 11). Several loci were mapped within a 50-kb flanking region of significant mGWAS signals near genes previously implicated in pigment regulation, including *Y1* (*Sobic.001G398100*, $P = 7.76 \times 10^{-21}$), *Tan2* (*Sobic.002G076600*, $P = 3.24 \times 10^{-11}$), *TT2* (*Sobic.003G183800*, $P = 2.55 \times 10^{-[illegible]}$), and *TT10* (*Sobic.004G236000*, $P = 3.57 \times 10^{-[illegible]}$), as well as key enzymes in the phenylpropanoid and lignin biosynthetic pathways-cinnamate-4-hydroxylase (C4H, *Sobic.002G126600*, $P = 2.36 \times 10^{-22}$), caffeoyl-CoA O-methyltransferase (CCoAOMT, *Sobic.002G242300*, $P = 4.74 \times 10^{-[illegible]}$), and cinnamyl alcohol dehydrogenase (CAD, *Sobic.010G071800*, $P = 3.54 \times 10^{-13}$) (Supplementary Fig.7D, Supplementary File 11). (Supplementary Fig.7D, Supplementary File 11). The QQ plot showed a clear deviation from the expected null distribution, reflecting an enrichment of significant associations across all seven polyphenols analyzed. This pattern highlights the robustness of the mGWAS approach and helps clarify the polygenic control of variation in grain color (Supplementary Fig.7E). These enzymes play critical roles in polyphenol and lignin biosynthesis. C4H[102] catalyzes an early step in the phenylpropanoid pathway and CCoAOMT[103] and CAD[104] contribute to lignin formation by modifying phenolic intermediates. 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 (Supplementary Fig. 7E). These enzymes play critical roles in

polyphenol and lignin biosynthesis. C4H[102] catalyzes an early step in the phenylpropanoid pathway and CCoAOMT[103] and CAD[104] contribute to lignin formation by modifying phenolic intermediates. 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 support molecular marker development for metabolite-driven grain color variation, we analyzed allele frequency distributions of the top mGWAS signals across the four grain color classes in Table 1. This analysis revealed a strong association between haplotypes defined by the eleven most-impactful SNPs for the four grain color groups. Each SNP exhibited two haplotypes (Hap1 and Hap2), which aligned with dark- (red and brown) and light-colored (yellow and white) grains, respectively. Metabolite abundance, measured as log□-fold change, showed significant differences between Hap1 and Hap2, with Hap1 associated with elevated polyphenol accumulation (Table 1). Hap1 was predominant in red (71–78%) and brown (9.8–22%) grains but occurred at significantly lower frequencies in yellow (4.6–10%) and white (4–8.5%) grains (Supplementary Fig.9; $\chi^2 = 13.82$, df = 3, $P = 0.00316$). This pattern suggests that Hap1, enriched in dark-colored grains, is strongly linked to enhanced polyphenolic metabolite production, particularly within the lignin and flavonoid metabolic pathways. To validate the role of these seven metabolites in discriminating among four grain color groups, we performed a joint haplotype analysis using the top eleven mGWAS signals alongside kernel color phenotypes 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). Notably, lines carrying Hap1 showed higher levels of the identified polyphenols and exhibited darker grain colors, whereas Hap2 was linked to lower metabolite levels and lighter grain colors. These results support the hypothesis that genetic variation within polyphenol biosynthesis pathways contributes to the diversity of grain color in sorghum.

Discussion

SorGMDA: Sorghum Grain Metabolite Database Enabling Post-GWAS Breeding and Genetic Analysis

Sorghum's inherent nutritional richness and resilience to abiotic stresses make it a cornerstone for global food security, particularly in marginal environments. Understanding its genetic and metabolic diversity provides unique opportunities to enhance crop vigor and seed quality through targeted breeding. In this study, untargeted UPLC–MS profiling of 266 lines from the Sorghum Association Panel (SAP) identified 2,206 high-confidence metabolites in mature grains, with >95% coverage across accessions. This comprehensive catalog highlights the breadth of sorghum's metabolomic landscape and provides a foundation for translational research.

Our mGWAS identified candidate genes underlying chemical signatures that shape both qualitative traits (e.g., flavor, color) and quantitative traits (e.g., yield, nutrient content). To maximize the utility of these findings, we developed the Sorghum Grain Metabolite Diversity Atlas (SorGMDA), a publicly accessible database that integrates metabolomic profiles with genetic variation. SorGMDA allows users to query metabolites, filter accessions by concentration, and explore genotype–metabolite relationships through interactive visualization tools, including Manhattan plots. This integration of large-scale metabolomic and genomic datasets provides an unprecedented resource for dissecting the genetic architecture of sorghum's metabolic diversity and applying these insights in crop improvement.

The post-GWAS framework, exemplified by SorGMDA, translates genetic and metabolomic insights into practical breeding strategies. The database's query functionality allows breeders to select SAP lines with elevated levels of health-beneficial metabolites (e.g., antioxidants, essential amino acids), facilitating the development of nutrient-rich sorghum varieties. In tomato, for instance, mGWAS studies identified loci controlling malic acid[105] lycopene

and carotenoid content[106], enabling breeders to prioritize lines with flavor and enhanced nutritional value respectively. Similarly, in maize, loci associated with kernel tocopherol (vitamin E) content were used to select for antioxidant-rich varieties[107-109]. These examples illustrate how metabolite data empower breeders to target traits that are challenging to assess visually or through traditional yield-based metrics[110]. Beyond phenotyping, mGWAS facilitates marker-assisted selection (MAS) by identifying single nucleotide polymorphisms (SNPs) linked to metabolite traits. In rice, SNPs associated with amylose content were converted into PCR-based markers[111], allowing breeders to efficiently select for low-amylose, sticky varieties favored in certain markets. These markers, often developed using high-throughput platforms like KASP or TaqMan[112], streamline breeding by enabling early selection, reducing reliance on costly and environmentally sensitive phenotyping assays. The precision and efficiency of MAS make it a cornerstone of modern breeding programs aiming to enhance metabolite-related traits. SorGMDA sets the stage for integrating metabolomics with genomics, transcriptomics, and proteomics to achieve a systems-level understanding of seed development and grain quality. Expanding mGWAS to include additional tissues (e.g., leaves, roots) and environmental conditions could further elucidate metabolite dynamics under stress, enhancing the efficiency of plant breeding by clarifying genotype-by-environment-by-management interactions (G×E×M)[63]. In particular, leveraging machine learning to predict genotype-metabolite-environment interactions could refine marker development and accelerate both cultivar and hybrid improvement.

In summary, this study delivers the first population-scale map of sorghum grain metabolomic and genetic diversity and establishes SorGMDA as a resource for the plant science and breeding communities. By linking metabolic signatures with genetic determinants, SorGMDA enables discovery, prediction, and application of metabolite traits, supporting the development of nutrient-rich, climate-resilient sorghum varieties. These advances highlight the potential of integrative omics frameworks to address global challenges in food security and agricultural sustainability.

The importance of mGWAS approaches in crop improvement

GWAS has revolutionized our ability to dissect the genetic architecture of complex traits in crops. mGWAS has emerged as an updated multi-omic approach to GWAS linking natural genetic variation with biochemical diversity[113, 114]. By linking metabolite variation to QTLs, mGWAS narrows down the list of candidate genes impacting trait expression and enables researchers to more quickly identify genes regulating or participating in biosynthetic pathways, such as those for sugars[115], amino acids[116], and flavonoids[117]. These compounds underpin stress tolerance, nutrition, functional food quality, and flavor. For example, identifying loci associated with flavonoid accumulation can elucidate pathways enhancing antioxidant properties, a key factor in sorghum's health benefits[118]. Functional validation of these loci using tools like CRISPR/Cas9 can confirm gene-metabolite relationships, advancing our understanding of metabolic regulation[119]. Comparative analyses with other crops, such as maize[49, 57, 120-123], rice[49, 124, 125], barley[59, 126] and tomato[106, 127] highlight the universal applicability of mGWAS as a tool in uncovering conserved and species-specific metabolic pathways to support sorghum genetic enhancement and breeding.

From a genomics perspective, mGWAS enables high-resolution mapping of metabolite traits onto candidate genes, paving the way for fine-mapping and functional studies. In maize, for example, mGWAS led to the discovery of a starch synthase gene regulating kernel starch content, which was subsequently validated with CRISPR/Cas9-mediated gene editing[128]. Similarly, in tomato, loci linked to anthocyanin content were mapped to transcription factors regulating flavonoid biosynthesis, providing actionable targets for metabolic engineering[129]. These efforts are complemented by pathway analyses that integrate mGWAS data with transcriptomics, proteomics, and metabolomics to construct comprehensive metabolic models that can be time-staged to provide even more information about biosynthetic processes.

In *Arabidopsis*, mGWAS was used to identify interactions between biosynthetic enzymes and regulatory elements in the glucosinolate pathway, informing engineering strategies in

Brassicaceae crops[130]. In rice, mGWAS was used to link amino acid content to nitrogen metabolism, guiding efforts to boost grain protein content. Such pathway-level insights enable geneticists to identify key regulatory nodes, facilitating broader metabolic improvements to crops, such as upregulating entire classes of metabolites like flavonoids[49, 125]. By combining gene mapping with multi-omic systems-level analyses, breeders have the tools and knowledge needed to implement modern genomics-assisted breeding strategies.

The mGWAS approach can be used in any species. In wheat, mGWAS identified loci for grain phenolic content, leading to markers for selecting antioxidant-rich varieties and the mapping of a phenylalanine ammonia-lyase gene[131]. In soybean, mGWAS for isoflavone content enabled marker development for health-beneficial lines and elucidated the role of a chalcone synthase gene in the isoflavonoid pathway[132]. Barley mGWAS for β-glucan content supported the breeding of heart-healthy varieties and identified β-glucan synthase genes for genetic engineering[133]. These cross-crop examples underscore the universal applicability of mGWAS in linking genotypes to metabolite phenotypes, with downstream benefits for both breeding and genetic research.

Nevertheless, several challenges remain. First, the confounding effects of population structure can obscure true genotype–metabolite associations, leading to spurious correlations if not properly controlled in statistical models. Second, metabolomic analyses remain relatively expensive and resource-intensive, particularly when large and diverse panels are involved, limiting their routine integration into breeding programs. Finally, there is a critical need to validate metabolite–trait associations across diverse environmental conditions, as metabolite accumulation is highly sensitive to environmental variation[110, 134]. Finally, there is a critical need to validate metabolite–trait associations across diverse environmental conditions, as metabolite accumulation is highly sensitive to environmental variation[110, 134].

Emerging statistical approaches commonly used by breeders-such as mixed linear models (MLMs), multi-omics integration, and machine learning-based predictive modeling-can be leveraged to address these obstacles. MLMs, which incorporate both fixed and random effects, are widely used in genome-wide association studies (GWAS) to control population structure

and relatedness, thereby reducing false-positive associations[135, 136]. For instance, the FarmCPU and BLINK algorithms extend the MLM framework to improve power and computational efficiency in large panels[137]. Machine learning models, such as random forests or deep neural networks, can improve the predictive accuracy of metabolite-based traits under varying environmental conditions by capturing nonlinear interactions and high-dimensional patterns that traditional models might miss[138]. These tools are increasingly used in genomic selection to forecast performance across environments and genotypes with reduced phenotyping costs.[139]Emerging statistical approaches commonly used by breeders-such as mixed linear models (MLMs), multi-omics integration, and machine learning-based predictive modeling-can be leveraged to address these obstacles. MLMs, which incorporate both fixed and random effects, are widely used in genome-wide association studies (GWAS) to control for population structure and relatedness, thereby reducing false-positive associations[135, 136]. For instance, the FarmCPU and BLINK algorithms extend the MLM framework to improve power and computational efficiency in large panels[137]. Machine learning models, such as random forests or deep neural networks, can improve the predictive accuracy of metabolite-based traits under varying environmental conditions by capturing nonlinear interactions and high-dimensional patterns that traditional models might miss[138]. These tools are increasingly used in genomic selection to forecast performance across environments and genotypes with reduced phenotyping costs.[139]. Advances in the efficiency and throughput of genetic engineering and CRISPR-based gene editing can use mGWAS data to develop metabolite-enhanced crops[117].

Methods

Sample collection

The original SAP collection (n = 400) was obtained from the U.S. National Plant Germplasm System (NPGS) and is documented in 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 using a benchtop freeze dryer (Labconco FreeZone 2.5 Liter, –84 °C) connected to a vacuum pump (Labconco Model 117) to prevent degradation and minimize heat generation during grinding. The dried grain was then finely ground in 2 mL microcentrifuge tubes using a bead mill homogenizer (e.g., Precellys 24) with appropriate ceramic or metal beads.

UPLC-MS and metabolite 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-second 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 UPLC-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 files (.raw) were converted into .mzXML format using MSConvert, with an MS-level filter applied[140]. Metabolite quantification was achieved by analyzing signal intensities through the XCMS and CAMERA data processing pipelines[141]. 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 Mm,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 *Mm,l*,1, *Mm,l*,2, *Mm,l*,3, are the normalized metabolite intensities from the three biological replicates. Metabolite annotation was performed using multiple databases, including HMDB[142], MassBank[143], ChemSpider and KEGG[144], allowing metabolite identification and classification.

Phenotypic data normalization and heritability estimation

Each phenotypic dataset, consisting of 2,206 distinct metabolite traits, was normalized prior to GWA analysis. Normalization was performed 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 high dimensionality of the dataset, 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:

$$Y_{ij} = \mu + \delta_i + \epsilon_{ij}$$

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:

$$H2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{1}{n}\sigma_R^2}$$

Metabolite diversity analysis

Following normalization, metabolite intensities were adjusted to ensure comparability across samples. Metabolites detected in fewer than 5% of the samples were excluded to minimize 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[145, 146]. 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. Following normalization, metabolite intensities were adjusted to ensure comparability across samples. The Shannon method[147, 148] was utilized to emphasize richness and evenness across metabolites, while the Simpson method highlighted dominance[149, 150]. 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 SNPs from the 266 SAP accessions were used[9]. Marker sets were filtered by removing those with minor allele frequencies (< 5%) among the set of genotypes employed for a given analysis. This filtering criteria resulted in a final set of 38,873,876 SNP markers for use in the mGWAS. Metabolites were selected for GWAS if they were detected in at least 90% of the 266 phenotyped verities and passed normality criteria to ensure robust statistical analysis. GWAS was independently performed on each of the 2,206 qualifying metabolites using a mixed linear model (MLM)[136] with principal components (PCs) and the kinship matrix-and-one multi-locus model, Fixed and random model Circulating Probability Unification (FarmCPU)[135]. These models were implemented with the R packages GAPIT[151] and rMVP (v1.0.1)[152]. 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"[153]. The variance component analysis method, vc.methods, was set to "GEMMA"[154]. 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)[155]. A stringent genome-wide significance threshold of $1\times10^{-}$ 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)[152]. 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. Candidate genes within these defined

regions were annotated using the B.Tx623 reference genome (v3.1) from Phytozome v10[156]. Genomic regions containing SNPs were further annotated using the Ensembl Variant Effect Predictor (VEP)[157].

Identification of the metabolite gene clusters

Metabolic (biosynthetic) gene clusters (MGCs) in *Sorghum bicolor* were identified using the plantiSMASH web server (version 2.0-beta4), following default analysis protocols[158]. Advanced parameters included a CD-HIT cutoff of 0.5 to cluster protein sequences sharing ≥50% identity, and a minimum of two unique biosynthetic domains required 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[60].

Time-staged tissue txpression 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, we combined our mGWAS results with previously published RNA-Seq expression profiles (E-MTAB-13406)[60] 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[159], and correlation networks among genes were visualized using Cytoscape v3.10.1[160] .

Analysis of kernel color related metabolite

Kernel color information for the SAP accessions was obtained from the U.S. National Plant Germplasm System (NPGS)[161, 162]. 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. To identify key polyphenolic metabolites associated with variation in seed color, we applied supervised machine learning models-Random Forest (RF)[163, 164] and Extreme Gradient Boosting (XGBoost)[165]. These models used metabolite abundance data as input features and learned to classify samples according to their grain color group. 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 greater influence on the model's classification performance. The two models were implemented using the scikit-learn and xgboost Python libraries[166], respectively, and trained using 10-fold stratified cross-validation to ensure robust performance and account for potential class imbalance[167], where one class (e.g., positive or negative cases) is underrepresented relative to the other, which can bias model training and evaluation. Feature importance was calculated using the mean decrease in Gini impurity (RF) and gain-based importance (XGBoost)[168, 169].

Polyphenols that consistently ranked among the top features across cross-validation folds in both models were considered robust discriminators, highlighting their contribution to the metabolic differentiation among grain color groups. Receiver Operating Characteristic (ROC) curve analysis[170, 171] was performed to evaluate the discriminatory ability of individual polyphenolic metabolites between the sorghum groups. The area under the curve (AUC) was calculated for each metabolite using the pROC package in R, with values closer to 1.0 indicating stronger classification performance. Ninety-five percent confidence intervals (95 % CI) were computed to assess the statistical robustness of the AUC estimates. 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[172]. Functionally linked SNPs were

extracted and analyzed for their correlation with phenotypic traits using the SKAT package[173, 174]. Differential metabolomic abundance for each haplotype were visualized using the ggplot2 package[175]. 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[176]. 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[160] 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 signal data 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 pipelines, as detailed in the Methods section.

Contribution

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 experiment. DN performed the data analysis. PK and DN designed the database. DN, YJ and MY drafted the manuscript. All authors 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: Metabolite Diversity Across SAP accessions. A) Diversity index of metabolites, 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, where 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: Metabolomic Networks within SAP Grain. A) Metabolite networks where each node represents a metabolite signal. Grey edges illustrate the molecular MS/MS network based on spectral similarity, and red edges highlight co-accumulating grain metabolites that

share similar accumulation patterns across 266 sorghum accessions. The networks were constructed using a similarity score threshold of 0.7. Clusters of metabolites are outlined with structures of representative metabolites shown for each cluster. Metabolite names within each cluster are provided in Supplementary Table 1. 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: Functional Insights into the Lysine Biosynthetic Pathway 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 mapped 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 bicolor*, 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 Analysis in mGWAS

Distribution of SNPs across different genomic regions in the SAP original dataset (input). B) SNPs significantly associated with 2,206 metabolites identified through mGWAS. The panel compares the distribution of input SNPs with mGWAS-associated SNPs across various

genomic regions, including upstream, downstream, and intergenic regions, highlighting differences in their proportions. 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. A prominent peak is observed near promoter regions, with a secondary peak around 100 kb from the nearest gene, suggesting potential enrichment in regulatory elements.

Fig. 5: Chromosomal Distribution and Functional Analysis of Metabolic Gene Clusters in Sorghum bicolor

A) Distribution of metabolic gene clusters (MGCs) along the 10 chromosomes of Sorghum bicolor. Each cluster type is color-coded by biosynthetic pathway, including polyketides, lignin, polyphenols, amino acids, sugars, terpenes, alkaloids, and their combinations. The bar plot inset displays the total number of MGCs identified per chromosome. The x-axis indicates chromosomal position (base pairs), and the y-axis shows the chromosome number.

B) Detailed analysis of the lignin metabolic gene cluster (Cluster No. 24) on chromosome 5. 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 gene expression (FPKM > 1), showing distinct seed-specific expression patterns supported by RNA-seq data[60]. Genes involved in flavonoid biosynthesis are marked in blue while those 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). Clustering was performed using the average linkage method.

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.

Supplementary figures:

Supplementary Figure 1: 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 in 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.

Supplementary Figure 2: Metabolite Annotation and Variability Analysis. A) Bar graph showing the total number of detected metabolites with subsets representing annotated metabolites, those with known functions, and those with defined pathways. B) Bar plot depicting the distribution of the coefficient of variation (CV%) for the detected metabolites, illustrating the variability range across the SAP lines.

Supplementary Figure 3: 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.

Supplementary Figure 4: 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 ($C_{6}H_{12}O_{6}$), both used as positive controls. Each point represents a SNP with chromosomal position shown on the *x*-axis and 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.

Supplementary Figure 5. 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 exhibiting 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.

Supplementary Figure 6: Genomic Distribution of Metabolic Gene Clusters (MGCs) and SNP Variation in Sorghum. 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.

Supplementary Figure 7: Grain Polyphenol Diversity and Enzymatic Pathway Variation in 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.

Supplementary Figure 8: 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 respective true positive and false positive rates.

Supplementary Figure 9: 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 (χ^2 = 13.82, df = 3, p = 0.00316).

Supplementary Figure 10: Density Distribution of Metabolite Coverage across Sorghum Accessions in the SAP. 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.

Supplementary Figure 11: 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□□(*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 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.

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**).

Supplementary Table

Supplementary Table 1. 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).

Supplementary files:

Supplementary File 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 (PIXXXXXX), representing relative abundance values derived from the metabolomics dataset.

Supplementary File 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%).

Supplementary File S3.

List of candidate genes associated with the accumulation of (-)-epigallocatechin ($C_{15}H_{14}O_{7}$), D-(-)-fructose ($C_{6}H_{12}O_{6}$), and lysine ($C_{6}H_{14}N_{2}O_{2}$) 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.

Supplementary File 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.

Supplementary File 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.

Supplementary File 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 proximity of genes.

Supplementary File 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.

Supplementary File 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.

Supplementary File S9.

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

Supplementary File 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□□(p-value), which facilitates interpretation of statistical significance. Quantitative values reflect normalized metabolite abundances across samples within each color group.

Supplementary File 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).

Supplementary File 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

1. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A: **The Sorghum bicolor genome and the diversification of grasses.** *Nature* 2009, **457:**551-556.
2. Ali AEE, Sharp RE, Greeley L, Peck SC, Tabb DL, Ludidi N: **Proteomic dataset of sorghum leaf and root responses to single and combined drought and heat stress.** *Scientific Data* 2025, **12:**97.
3. Liaqat W, Altaf MT, Barutçular C, Mohamed HI, Ahmad H, Jan MF, Khan EH: **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* 2024, **24:**74-101.
4. Burdette A, Garner PL, Mayer EP, Hargrove JL, Hartle DK, Greenspan P: **Anti-inflammatory activity of select sorghum (Sorghum bicolor) brans.** *Journal of medicinal food* 2010, **13:**879-887.
5. Zhang Y, Li M, Gao H, Wang B, Tongcheng X, Gao B, Yu L: **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* 2019, **7:**949-958.
6. Mazumder S, Bhattacharya D, Lahiri D, Moovendhan M, Sarkar T, Nag M: **Harnessing the nutritional profile and health benefits of millets: a solution to global food security problems.** *Critical Reviews in Food Science and Nutrition* 2024**:**1-22.
7. Collins A, Santhakumar AB, Francis N, Blanchard C, Chinkwo K: **Impact of sorghum (Sorghum bicolor L. Moench) phenolic compounds on cancer development pathways.** *Food Bioscience* 2024**:**104177.
8. Kaur S, Kumar K, Singh L, Sharanagat VS, Nema PK, Mishra V, Bhushan B: **Gluten-free grains: Importance, processing and its effect on quality of gluten-free products.** *Critical reviews in food science and nutrition* 2024, **64:**1988-2015.
9. Boatwright JL, Sapkota S, Jin H, Schnable JC, Brenton Z, Boyles R, Kresovich S: **Sorghum Association Panel whole□genome sequencing establishes cornerstone resource for dissecting genomic diversity.** *The Plant Journal* 2022, **111:**888-904.
10. Habyarimana E, Dall'Agata M, De Franceschi P, Baloch FS: **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* 2019, **14:**e0225979.

11. Kaufman RC, Herald TJ, Bean SR, Wilson JD, Tuinstra MR: **Variability in tannin content, chemistry and activity in a diverse group of tannin containing sorghum cultivars.** *Journal of the Science of Food and Agriculture* 2013, **93:**1233-1241.
12. Yang Q, Van Haute M, Korth N, Sattler SE, Toy J, Rose DJ, Schnable JC, Benson AK: **Genetic analysis of seed traits in Sorghum bicolor that affect the human gut microbiome.** *nature communications* 2022, **13:**5641.
13. Rhodes DH, Hoffmann Jr L, Rooney WL, Ramu P, Morris GP, Kresovich S: **Genome-wide association study of grain polyphenol concentrations in global sorghum [Sorghum bicolor (L.) Moench] germplasm.** *Journal of agricultural and food chemistry* 2014, **62:**10916-10927.
14. Punia H, Tokas J, Malik A, Satpal, Sangwan S: **Characterization of phenolic compounds and antioxidant activity in sorghum [Sorghum bicolor (L.) Moench] grains.** *Cereal Research Communications* 2021, **49:**343-353.
15. Lividini K, Masters WA: **Tracing global flows of bioactive compounds from farm to fork in nutrient balance sheets can help guide intervention towards healthier food supplies.** *Nature food* 2022, **3:**703-715.
16. Sigalas PP, Shewry PR, Riche A, Wingen L, Feng C, Siluveru A, Chayut N, Burridge A, Uauy C, Castle M: **Improving wheat grain composition for human health by constructing a QTL atlas for essential minerals.** *Communications Biology* 2024, **7:**1001.
17. Sánchez-Rodríguez AR, Rey M-D, Nechate-Drif H, Castillejo MÁ, Jorrín-Novo JV, Torrent J, Del Campillo MC, Sacristán D: **Combining P and Zn fertilization to enhance yield and grain quality in maize grown on Mediterranean soils.** *Scientific reports* 2021, **11:**7427.
18. Girard AL, Awika JM: **Sorghum polyphenols and other bioactive components as functional and health promoting food ingredients.** *Journal of Cereal Science* 2018, **84:**112-124.
19. Przybylska-Balcerek A, Frankowski J, Stuper-Szablewska K: **Bioactive compounds in sorghum.** *European Food Research and Technology* 2019, **245:**1075-1080.
20. Vanamala JK, Massey AR, Pinnamaneni SR, Reddivari L, Reardon KF: **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* 2018, **58:**2867-2881.
21. Farrar JL, Hartle DK, Hargrove JL, Greenspan P: **A novel nutraceutical property of select sorghum (Sorghum bicolor) brans: inhibition of protein glycation.** *Phytotherapy Research* 2008, **22:**1052-1056.
22. Mukherjee A, Maheshwari U, Sharma V, Sharma A, Kumar S: **Functional insight into multi-omics-based interventions for climatic resilience in sorghum (Sorghum bicolor): a nutritionally rich cereal crop.** *Planta* 2024, **259:**91.

23. Awika JM, Yang L, Browning JD, Faraj A: **Comparative antioxidant, antiproliferative and phase II enzyme inducing potential of sorghum (Sorghum bicolor) varieties.** *LWT-Food Science and Technology* 2009, **42:**1041-1046.
24. Dykes L, Rooney WL, Rooney LW: **Evaluation of phenolics and antioxidant activity of black sorghum hybrids.** *Journal of Cereal Science* 2013, **58:**278-283.
25. 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: **Extruded sorghum flour (Sorghum bicolor L.) modulate adiposity and inflammation in high fat diet-induced obese rats.** *Journal of functional foods* 2018, **42:**346-355.
26. Chen X, Shen J, Xu J, Herald T, Smolensky D, Perumal R, Wang W: **Sorghum phenolic compounds are associated with cell growth inhibition through cell cycle arrest and apoptosis in human hepatocarcinoma and colorectal adenocarcinoma cells.** *Foods* 2021, **10:**993.
27. Cingoz GS, Gurel E: **Effects of salicylic acid on thermotolerance and cardenolide accumulation under high temperature stress in Digitalis trojana Ivanina.** *Plant Physiology and Biochemistry* 2016, **105:**145-149.
28. Tak Y, Kumar M: **Phenolics: a key defence secondary metabolite to counter biotic stress.** *Plant Phenolics in Sustainable Agriculture: Volume 1* 2020**:**309-329.
29. 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: **Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor).** *Genome biology* 2011, **12:**1-15.
30. 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: **Genome-wide association study for starch content and constitution in sorghum (Sorghum bicolor (L.) Moench).** *Journal of Integrative Agriculture* 2019, **18:**2446-2456.
31. Boyles RE, Pfeiffer BK, Cooper EA, Rauh BL, Zielinski KJ, Myers MT, Brenton Z, Rooney WL, Kresovich S: **Genetic dissection of sorghum grain quality traits using diverse and segregating populations.** *Theoretical and applied genetics* 2017, **130:**697-716.
32. Li J, Tang W, Zhang Y-W, Chen K-N, Wang C, Liu Y, Zhan Q, Wang C, Wang S-B, Xie S-Q: **Genome-wide association studies for five forage quality-related traits in sorghum (Sorghum bicolor L.).** *Frontiers in plant science* 2018, **9:**1146.
33. Thakur NR, Gorthy S, Vemula A, Odeny DA, Ruperao P, Sargar PR, Mehtre SP, Kalpande HV, Habyarimana E: **Genome-wide association study and expression of candidate genes for Fe and Zn concentration in sorghum grains.** *Scientific Reports* 2024, **14:**12729.
34. Li C, Huang Y, Huang R, Wu Y, Wang W: **The genetic architecture of amylose biosynthesis in maize kernel.** *Plant biotechnology journal* 2018, **16:**688-695.
35. Angelovici R, Batushansky A, Deason N, Gonzalez-Jorge S, Gore MA, Fait A, DellaPenna D: **Network-guided GWAS improves identification of genes affecting free amino acids.** *Plant physiology* 2017, **173:**872-886.

36. Izquierdo P, Sadohara R, Wiesinger J, Glahn R, Urrea C, Cichy K: **Genome-wide association and genomic prediction for iron and zinc concentration and iron bioavailability in a collection of yellow dry beans.** *Frontiers in Genetics* 2024, **15:**1330361.

37. Hao Y, Kong F, Wang L, Zhao Y, Li M, Che N, Li S, Wang M, Hao M, Zhang X: **Genome-wide association study of grain micronutrient concentrations in bread wheat.** *Journal of Integrative Agriculture* 2024, **23:**1468-1480.

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

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

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

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

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

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

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

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

46. Chen J, Xue M, Liu H, Fernie AR, Chen W: **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* 2021, **2**.

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

48. Edwards D, Batley J: **Plant bioinformatics: from genome to phenome.** *TRENDS in Biotechnology* 2004, **22:**232-237.

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

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

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

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

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

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

55. Raffler J, Friedrich N, Arnold M, Kacprowski T, Rueedi R, Altmaier E, Bergmann S, Budde K, Gieger C, Homuth G: **Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality.** *PLoS genetics* 2015, **11:**e1005487.

56. Singhal R, Izquierdo P, Ranaweera T, Segura Abá K, Brown BN, Lehti-Shiu MD, Shiu S-H: **Using supervised machine-learning approaches to understand abiotic stress tolerance and design resilient crops.** *Philosophical Transactions B* 2025, **380:**20240252.

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

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

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

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

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

62. Fang C, Luo J: **Metabolic GWAS based dissection of genetic bases underlying the diversity of plant metabolism.** *The Plant Journal* 2019, **97:**91-100.

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

64. Guo Z, Wang S, Zhang F, Xiang D, Yang J, Li D, Bai B, Dai M, Luo J, Xiong L: **Common and specific genetic basis of metabolite-mediated drought responses in rice.** *Stress Biology* 2024, **4:**6.

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

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

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

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

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

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

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

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

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

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

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

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

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

78. Galili G: **The aspartate-family pathway of plants: linking production of essential amino acids with energy and stress regulation.** *Plant signaling & behavior* 2011, **6:**192-195.

79. Varisi VA, Medici LO, van der Meer I, Lea PJ, Azevedo RA: **Dihydrodipicolinate synthase in opaque and floury maize mutants.** *Plant Science* 2007, **173:**458-467.

80. Lee SI, Kim HU, Lee Y-H, Suh S-C, Lim YP, Lee H-Y, Kim H-I: **Constitutive and seed-specific expression of a maize lysine-feedback-insensitive dihydrodipicolinate synthase gene leads to increased free lysine levels in rice seeds.** *Molecular Breeding* 2001, **8:**75-84.

81. Yang QQ, Yu WH, Wu HY, Zhang CQ, Sun SSM, Liu QQ: **Lysine biofortification in rice by modulating feedback inhibition of aspartate kinase and dihydrodipicolinate synthase.** *Plant Biotechnology Journal* 2021, **19:**490-501.

82. Huang A, Coutu C, Harrington M, Rozwadowski K, Hegedus DD: **Engineering a feedback inhibition-insensitive plant dihydrodipicolinate synthase to increase lysine content in Camelina sativa seeds.** *Transgenic Research* 2021**:**1-18.

83. Zhu X, Galili G: **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* 2003, **15:**845-853.

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

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

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

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

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

89. Nguyen T-N, Son S, Jordan MC, Levin DB, Ayele BT: **Lignin biosynthesis in wheat (Triticum aestivum L.): its response to waterlogging and association with hormonal levels.** *BMC plant biology* 2016, **16:**1-16.

90. Choudhary P, Pramitha L, Aggarwal PR, Rana S, Vetriventhan M, Muthamilarasan M: **Biotechnological interventions for improving the seed longevity in cereal crops: progress and prospects.** *Critical Reviews in Biotechnology* 2023, **43:**309-325.

91. Riaz MW, Yousaf MI, Hussain Q, Yasir M, Sajjad M, Shah L: **Role of lignin in wheat plant for the enhancement of resistance against lodging and biotic and abiotic stresses.** *Stresses* 2023, **3:**434-453.

92. Khan P, Tong L, Khan S, Zhang C, Wang W: **Lignin: a defensive shield halting the environmental stresses-a review.** *Applied Ecology & Environmental Research* 2022, **20**.

93. Ushimaru R: **Functions and mechanisms of enzymes assembling lignans and norlignans.** *Current Opinion in Chemical Biology* 2024, **80:**102462.

94. Fornalé S, Rencoret J, García-Calvo L, Encina A, Rigau J, Gutiérrez A, Del Río JC, Caparros-Ruiz D: **Changes in cell wall polymers and degradability in maize mutants lacking 3′-and 5′-O-methyltransferases involved in lignin biosynthesis.** *Plant and Cell Physiology* 2017, **58:**240-255.

95. Cona A, Rea G, Angelini R, Federico R, Tavladoraki P: **Functions of amine oxidases in plant development and defence.** *Trends in plant science* 2006, **11:**80-88.

96. Yonekura-Sakakibara K, Yamamura M, Matsuda F, Ono E, Nakabayashi R, Sugawara S, Mori T, Tobimatsu Y, Umezawa T, Saito K: **Seed-coat protective neolignans are produced by the dirigent protein AtDP1 and the laccase AtLAC5 in Arabidopsis.** *The Plant Cell* 2021, **33:**129-152.

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

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

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

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

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

102. Zhang B, Lewis KM, Abril A, Davydov DR, Vermerris W, Sattler SE, Kang C: **Structure and function of the cytochrome P450 monooxygenase cinnamate 4-hydroxylase from Sorghum bicolor.** *Plant physiology* 2020, **183:**957-973.

103. Rakoczy M, Femiak I, Alejska M, Figlerowicz M, Podkowinski J: **Sorghum CCoAOMT and CCoAOMT-like gene evolution, structure, expression and the role of conserved amino acids in protein activity.** *Molecular Genetics and Genomics* 2018, **293:**1077-1089.

104. Saballos A, Ejeta G, Sanchez E, Kang C, Vermerris W: **A genomewide analysis of the cinnamyl alcohol dehydrogenase family in sorghum [Sorghum bicolor (L.) Moench] identifies SbCAD2 as the brown midrib6 gene.** *Genetics* 2009, **181:**783-795.

105. Gai W, Yang F, Yuan L, Ul Haq S, Wang Y, Wang Y, Shang L, Li F, Ge P, Dong H: **Multiple-model GWAS identifies optimal allelic combinations of quantitative trait loci for malic acid in tomato.** *Horticulture Research* 2023, **10:**uhad021.

106. Burgos E, Belen De Luca M, Diouf I, de Haro LA, Albert E, Sauvage C, Tao ZJ, Bermudez L, Asís R, Nesi AN: **Validated MAGIC and GWAS population mapping reveals the link between vitamin E content and natural variation in chorismate metabolism in tomato.** *The Plant Journal* 2021, **105:**907-923.

107. Li Q, Yang X, Xu S, Cai Y, Zhang D, Han Y, Li L, Zhang Z, Gao S, Li J: **Genome-wide association studies identified three independent polymorphisms associated with α-tocopherol content in maize kernels.** *PLoS One* 2012, **7:**e36807.

108. Xiao Y, Yu Y, Li G, Xie L, Guo X, Li J, Li Y, Hu J: **Genome-wide association study of vitamin E in sweet corn kernels.** *The Crop Journal* 2020, **8:**341-350.

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

110. Fernie AR, Schauer N: **Metabolomics-assisted breeding: a viable option for crop improvement?** *Trends in genetics* 2009, **25:**39-48.

111. Reddappa SB, Chhabra R, Muthusamy V, Zunjare RU, Talukder ZA, Mishra SJ, Katral A, Singh AK, Hossain F: **Characterization of Sub□Tropically Adapted Maize Breeding Lines for Loci Governing Kernel Amylose and Resistant Starch.** *Starch□Stärke* 2024, **76:**2300289.

112. Semagn K, Babu R, Hearne S, Olsen M: **Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the**

technology and its application in crop improvement. *Molecular breeding* 2014, **33:**1-14.

113. Soto-Cerda BJ, Vasudevan A, Laroche A, Ragupathy R: **Genome-wide association studies (GWAS) in cereals.** *Accelerated Breeding of Cereal Crops* 2022**:**119-132.

114. Huang X, Han B: **Natural variations and genome-wide association studies in crop plants.** *Annual review of plant biology* 2014, **65:**531-551.

115. Slaten ML, Yobi A, Bagaza C, Chan YO, Shrestha V, Holden S, Katz E, Kanstrup C, Lipka AE, Kliebenstein DJ: **mGWAS uncovers Gln-Glucosinolate seed-specific interaction and its role in metabolic homeostasis.** *Plant physiology* 2020, **183:**483-500.

116. Peng Y, Liu H, Chen J, Shi T, Zhang C, Sun D, He Z, Hao Y, Chen W: **Genome-wide association studies of free amino acid levels by six multi-locus models in bread wheat.** *Frontiers in plant science* 2018, **9:**1196.

117. Chen J, Zhang Y, Wei J, Hu X, Yin H, Liu W, Li D, Tian W, Hao Y, He Z: **Beyond pathways: Accelerated flavonoids candidate identification and novel exploration of enzymatic properties using combined mapping populations of wheat.** *Plant Biotechnology Journal* 2024, **22:**2033-2050.

118. Xiong Y, Zhang P, Warner RD, Fang Z: **Sorghum grain: From genotype, nutrition, and phenolic profile to its health benefits and food applications.** *Comprehensive Reviews in Food Science and Food Safety* 2019, **18:**2025-2046.

119. Weldemichael MY, Gebremedhn HM, Teklu TH: **Advances in genome editing and future prospects for Sorghum improvement: A review.** *Plant Gene* 2024**:**100464.

120. Kaňovská I, Biová J, Škrabišová M: **New perspectives of post-GWAS analyses: From markers to causal genes for more precise crop breeding.** *Current Opinion in Plant Biology* 2024, **82:**102658.

121. Shu G, Wang A, Wang X, Ding J, Chen R, Gao F, Wang A, Li T, Wang Y: **Identification of southern corn rust resistance QTNs in Chinese summer maize germplasm via multi-locus GWAS and post-GWAS analysis.** *Frontiers in Plant Science* 2023, **14:**1221395.

122. Liang X, Liu S, Wang T, Li F, Cheng J, Lai J, Qin F, Li Z, Wang X, Jiang C: **Metabolomics□driven gene mining and genetic improvement of tolerance to salt□induced osmotic stress in maize.** *New Phytologist* 2021, **230:**2355-2370.

123. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N: **Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels.** *Nature genetics* 2013, **45:**43-50.

124. Guo T, Yang J, Li D, Sun K, Luo L, Xiao W, Wang J, Liu Y, Wang S, Wang H: **Integrating GWAS, QTL, mapping and RNA-seq to identify candidate genes for seed vigor in rice (Oryza sativa L.).** *Molecular Breeding* 2019, **39:**1-16.

125. Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, Li Y, Liu X, Zhang H, Dong H: **Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism.** *Nature genetics* 2014, **46:**714-721.

126. Thabet SG, Alqahtani MD, Jabbour AA, Alqudah AM: **Genetic associations underpinning the metabolite-mediated salt stress tolerance in barley.** *Plant molecular biology reporter* 2024, **42:**410-423.
127. Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M: **Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits.** *Plant physiology* 2014, **165:**1120-1132.
128. Luo J, He C, Yan S, Jiang C, Chen A, Li K, Zhu Y, Gui S, Yang N, Xiao Y: **A metabolic roadmap of waxy corn flavor.** *Molecular Plant* 2024, **17:**1883-1898.
129. Butelli E, Titta L, Giorgio M, Mock H-P, Matros A, Peterek S, Schijlen EG, Hall RD, Bovy AG, Luo J: **Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors.** *Nature biotechnology* 2008, **26:**1301-1308.
130. Chan EK, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ: **Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana.** *PLoS biology* 2011, **9:**e1001125.
131. Nigro D, Laddomada B, Mita G, Blanco E, Colasuonno P, Simeone R, Gadaleta A, Pasqualone A, Blanco A: **Genome-wide association mapping of phenolic acids in tetraploid wheats.** *Journal of Cereal Science* 2017, **75:**25-34.
132. Yasmin F, Zhang H, Leamy L, Wang B, Winnike J, Reid RW, Brouwer CR, Song B-H: **Genetic basis and selection of glyceollin elicitation in wild soybean.** *Frontiers in Plant Science* 2024, **15:**1240981.
133. Geng L, Li M, Xie S, Wu D, Ye L, Zhang G: **Identification of genetic loci and candidate genes related to β-glucan content in barley grain by genome-wide association study in International Barley Core Selected Collection.** *Molecular Breeding* 2021, **41:**1-12.
134. Schauer N, Fernie AR: **Plant metabolomics: towards biological function and mechanism.** *Trends in plant science* 2006, **11:**508-516.
135. Liu X, Huang M, Fan B, Buckler ES, Zhang Z: **Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies.** *PLoS genetics* 2016, **12:**e1005767.
136. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM: **Mixed linear model approach adapted for genome-wide association studies.** *Nature genetics* 2010, **42:**355-360.
137. Poudel S: **Optimizing Gwas in Barley: Trade-Offs Between Statistical Power and Computational Efficiency Across Different Models.** 2025.
138. Sen P, Lamichhane S, Mathema VB, McGlinchey A, Dickens AM, Khoomrung S, Orešič M: **Deep learning meets metabolomics: a methodological perspective.** *Briefings in Bioinformatics* 2021, **22:**1531-1542.
139. Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu S-H: **Benchmarking parametric and machine learning models for genomic prediction of complex traits.** *G3: Genes, Genomes, Genetics* 2019, **9:**3691-3702.

140. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J: **A cross-platform toolkit for mass spectrometry and proteomics.** *Nature biotechnology* 2012, **30:**918-920.
141. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G: **XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.** *Analytical chemistry* 2006, **78:**779-787.
142. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, Dizon R, Sayeeda Z, Tian S, Lee BL: **HMDB 5.0: the human metabolome database for 2022.** *Nucleic acids research* 2022, **50:**D622-D631.
143. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K: **MassBank: a public repository for sharing mass spectral data for life sciences.** *Journal of mass spectrometry* 2010, **45:**703-714.
144. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28:**27-30.
145. Gorelick R: **Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices.** *Ecography* 2006, **29:**525-530.
146. DeJong TM: **A comparison of three diversity indices based on their components of richness and evenness.** *Oikos* 1975**:**222-227.
147. Pallister T, Jackson MA, Martin TC, Zierer J, Jennings A, Mohney RP, MacGregor A, Steves CJ, Cassidy A, Spector TD: **Hippurate as a metabolomic marker of gut microbiome diversity: Modulation by diet and relationship to metabolic syndrome.** *Scientific reports* 2017, **7:**13670.
148. Menni C, Zhu J, Le Roy CI, Mompeo O, Young K, Rebholz CM, Selvin E, North KE, Mohney RP, Bell JT: **Serum metabolites reflecting gut microbiome alpha diversity predict type 2 diabetes.** *Gut microbes* 2020, **11:**1632-1642.
149. Salgado AL, Glassmire AE, Sedio BE, Diaz R, Stout MJ, Čuda J, Pyšek P, Meyerson LA, Cronin JT: **Metabolomic evenness underlies intraspecific differences among lineages of a wetland grass.** *Journal of Chemical Ecology* 2023, **49:**437-450.
150. Bi B, Wang K, Zhang H, Wang Y, Fei H, Pan R, Han F: **Plants use rhizosphere metabolites to regulate soil microbial diversity.** *Land Degradation & Development* 2021, **32:**5267-5280.
151. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z: **GAPIT: genome association and prediction integrated tool.** *Bioinformatics* 2012, **28:**2397-2399.
152. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X: **rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study.** *Genomics, Proteomics and Bioinformatics* 2021, **19:**619-628.
153. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D: **FaST linear mixed models for genome-wide association studies.** *Nature methods* 2011, **8:**833-835.

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

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

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

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

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

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

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

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

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

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

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

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

166. Brownlee J: *XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn.* Machine Learning Mastery; 2016.

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

168. Pardeshi[1] NG, Patil DV: **Applying Gini Importance and RFE Methods.** In *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*. Springer Nature; 2023: 214.

169. Ben Jabeur S, Stef N, Carmona P: **Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering.** *Computational Economics* 2023, **61:**715-741.

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

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

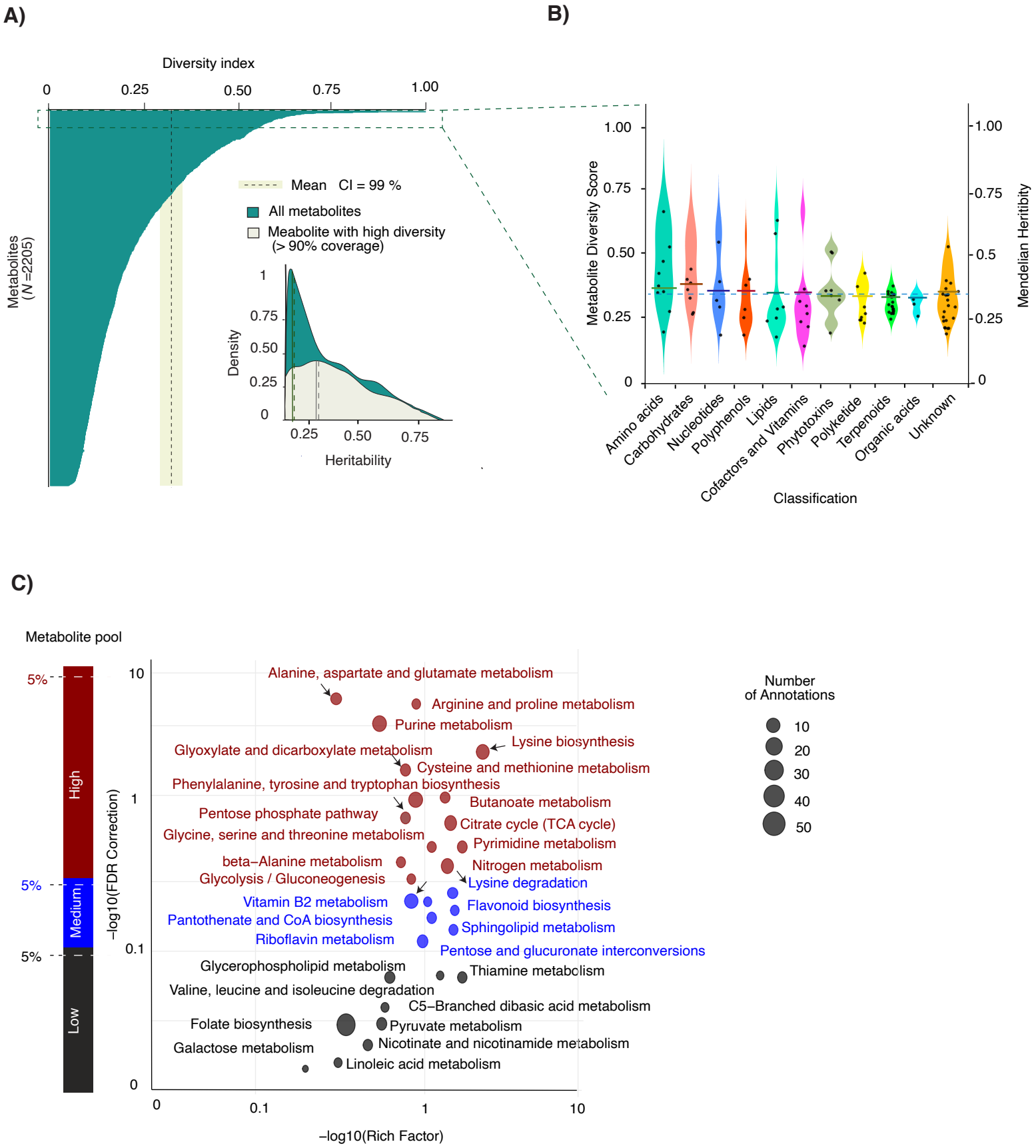
172. Zhang R, Jia G, Diao X: **geneHapR: an R package for gene haplotypic statistics and visualization.** *BMC bioinformatics* 2023, **24:**199.

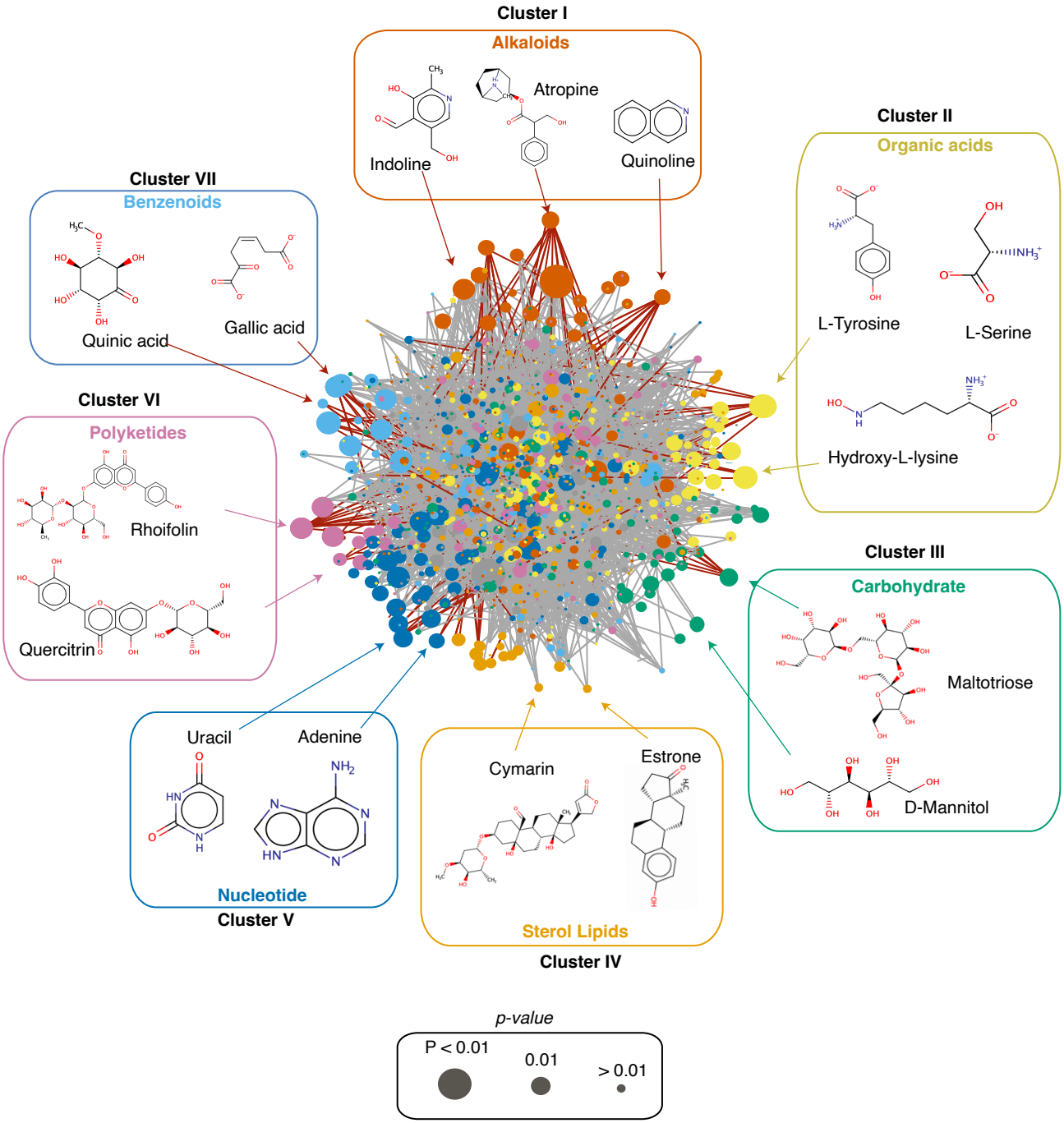
173. Lee S, Miropolsky L, Wu M, Lee M, CompQuadForm D: **Package "SKAT.".** In *CRAN*. 2011

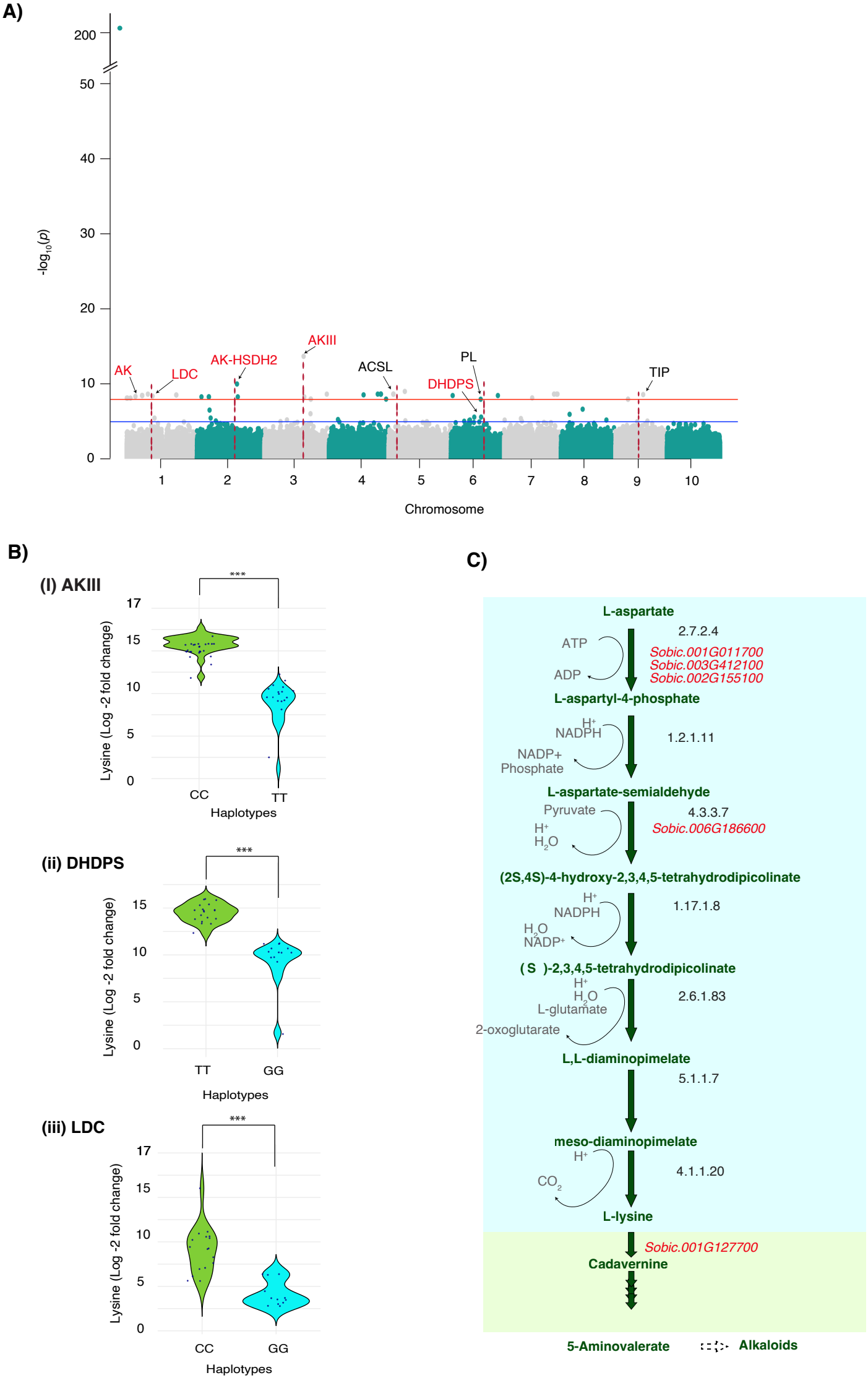
174. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare-variant association testing for sequencing data with the sequence kernel association test.** *The American Journal of Human Genetics* 2011, **89:**82-93.

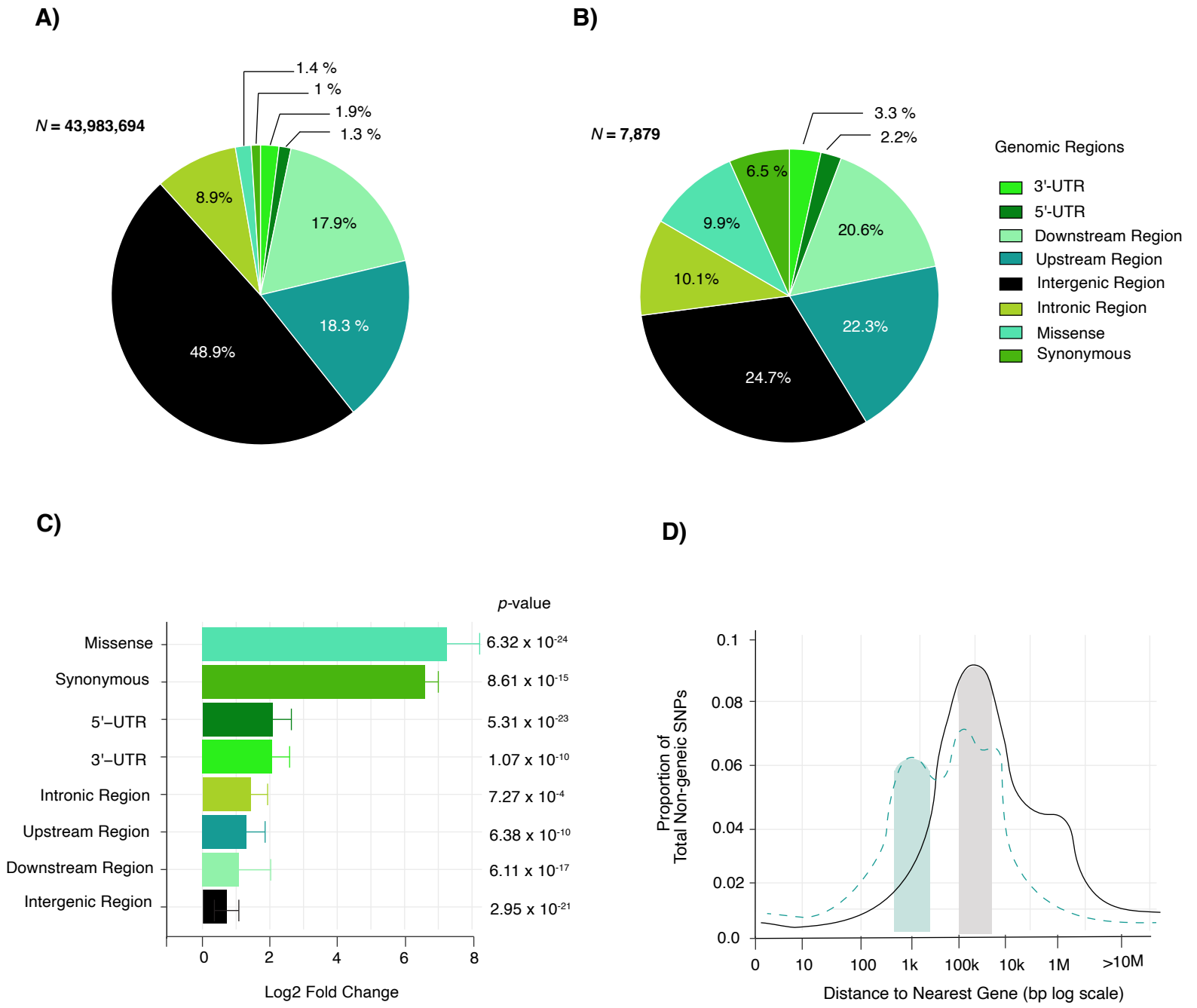
175. Wickham H: **ggplot2.** *Wiley interdisciplinary reviews: computational statistics* 2011, **3:**180-185.

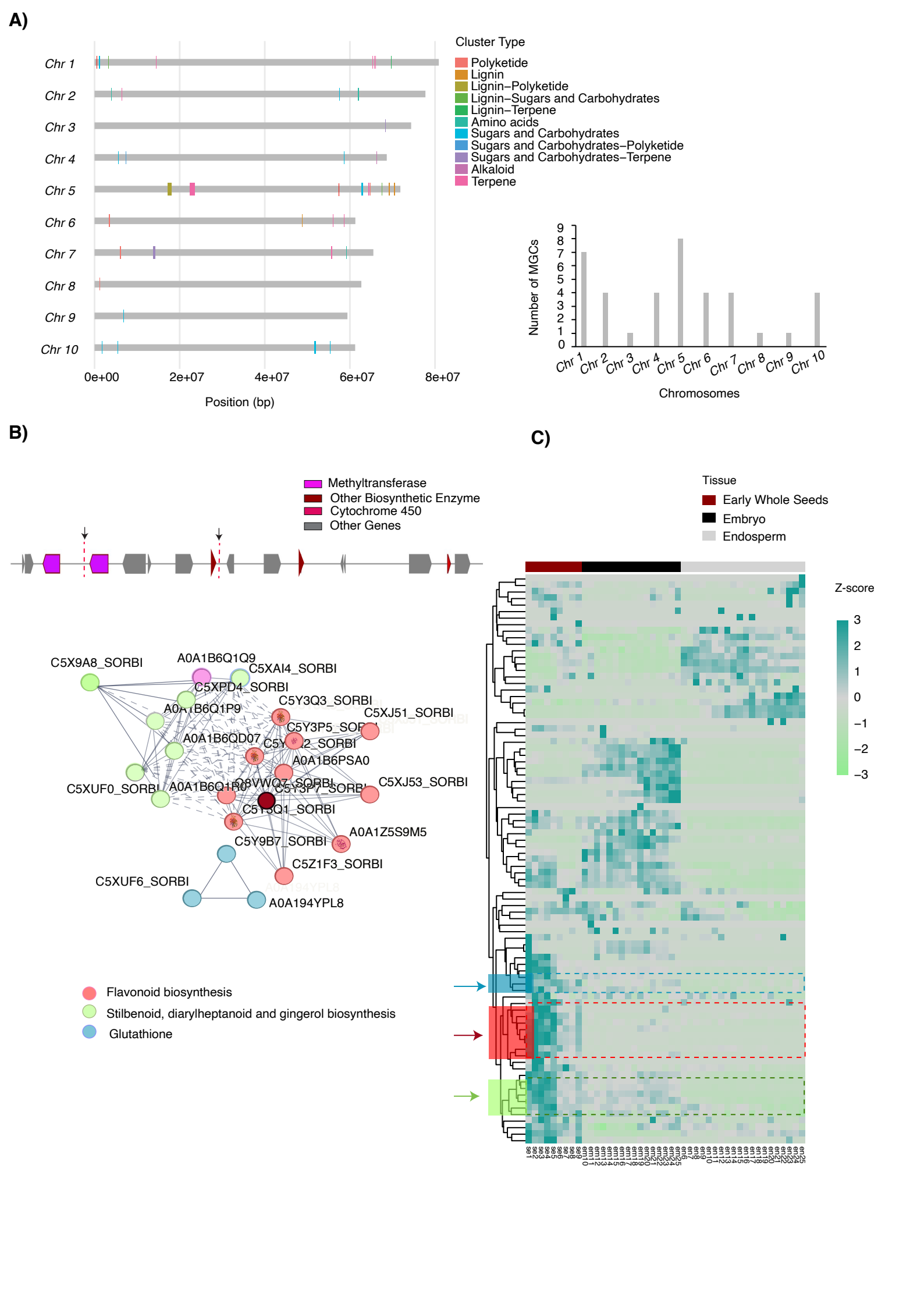
176. Kim S: **ppcor: an R package for a fast calculation to semi-partial correlation coefficients.** *Communications for statistical applications and methods* 2015, **22:**665.


A)
Diversity index
0
0.25
0.50
0.75
1.00
Metabolites
(N =2205)
Mean CI = 99 %
All metabolites
Meabolite with high diversity
(> 90% coverage)
Density
Heritability
B)
Metabolite Diversity Score
Mendelian Heritibity
Amino acids
Carbohydrates
Nucleotides
Polyphenols
Lipids
Cofactors and Vitamins
Phytotoxins
Polyketide
Terpenoids
Organic acids
Unknown
Classification
C)
Metabolite pool
5%
High
Medium
Low
–log10(FDR Correction)
–log10(Rich Factor)
Alanine, aspartate and glutamate metabolism
Arginine and proline metabolism
Purine metabolism
Lysine biosynthesis
Glyoxylate and dicarboxylate metabolism
Cysteine and methionine metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
Butanoate metabolism
Pentose phosphate pathway
Citrate cycle (TCA cycle)
Glycine, serine and threonine metabolism
Pyrimidine metabolism
beta–Alanine metabolism
Nitrogen metabolism
Glycolysis / Gluconeogenesis
Lysine degradation
Vitamin B2 metabolism
Flavonoid biosynthesis
Pantothenate and CoA biosynthesis
Sphingolipid metabolism
Riboflavin metabolism
Pentose and glucuronate interconversions
Glycerophospholipid metabolism
Thiamine metabolism
Valine, leucine and isoleucine degradation
C5–Branched dibasic acid metabolism
Folate biosynthesis
Pyruvate metabolism
Galactose metabolism
Nicotinate and nicotinamide metabolism
Linoleic acid metabolism
Number
of Annotations
10
20
30
40
50


Cluster I
Alkaloids
Indoline
Atropine
Quinoline
Cluster II
Organic acids
L-Tyrosine
L-Serine
Hydroxy-L-lysine
Cluster III
Carbohydrate
Maltotriose
D-Mannitol
Cluster IV
Sterol Lipids
Cymarin
Estrone
Cluster V
Nucleotide
Uracil
Adenine
Cluster VI
Polyketides
Rhoifolin
Quercitrin
Cluster VII
Benzenoids
Quinic acid
Gallic acid
p-value
P < 0.01
0.01
> 0.01


A)
-log10(p)
200
50
40
30
20
10
0
AK
LDC
AK-HSDH2
AKIII
ACSL
DHDPS
PL
TIP
1
2
3
4
5
6
7
8
9
10
Chromosome
B)
(I) AKIII
***
Lysine (Log -2 fold change)
17
15
10
5
0
CC
TT
Haplotypes
(ii) DHDPS
***
Lysine (Log -2 fold change)
15
10
5
0
TT
GG
Haplotypes
(iii) LDC
***
Lysine (Log -2 fold change)
17
15
10
5
0
CC
GG
Haplotypes
C)
L-aspartate
2.7.2.4
ATP
ADP
Sobic.001G011700
Sobic.003G412100
Sobic.002G155100
L-aspartyl-4-phosphate
H+
NADPH
NADP+
Phosphate
1.2.1.11
L-aspartate-semialdehyde
Pyruvate
H+
H2O
4.3.3.7
Sobic.006G186600
(2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate
H+
NADPH
H2O
NADP+
1.17.1.8
(S)-2,3,4,5-tetrahydrodipicolinate
H+
H2O
L-glutamate
2-oxoglutarate
2.6.1.83
L,L-diaminopimelate
5.1.1.7
meso-diaminopimelate
H+
CO2
4.1.1.20
L-lysine
Sobic.001G127700
Cadavernine
5-Aminovalerate
Alkaloids


A)
N = 43,983,694
1.4 %
1 %
1.9%
1.3 %
8.9%
17.9%
18.3 %
48.9%
B)
N = 7,879
3.3 %
2.2%
6.5 %
9.9%
20.6%
10.1%
22.3%
24.7%
Genomic Regions
3'-UTR
5'-UTR
Downstream Region
Upstream Region
Intergenic Region
Intronic Region
Missense
Synonymous
C)
p-value
Missense
Synonymous
5'–UTR
3'–UTR
Intronic Region
Upstream Region
Downstream Region
Intergenic Region
6.32 x 10-24
8.61 x 10-15
5.31 x 10-23
1.07 x 10-10
7.27 x 10-4
6.38 x 10-10
6.11 x 10-17
2.95 x 10-21
0
2
4
6
8
Log2 Fold Change
D)
Proportion of Total Non-geneic SNPs
0.1
0.08
0.06
0.04
0.02
0.0
0
10
100
1k
100k
10k
1M
>10M
Distance to Nearest Gene (bp log scale)


A)
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
0e+00
2e+07
4e+07
6e+07
8e+07
Position (bp)
Cluster Type
Polyketide
Lignin
Lignin−Polyketide
Lignin−Sugars and Carbohydrates
Lignin−Terpene
Amino acids
Sugars and Carbohydrates
Sugars and Carbohydrates−Polyketide
Sugars and Carbohydrates−Terpene
Alkaloid
Terpene
Number of MGCs
Chromosomes
B)
Methyltransferase
Other Biosynthetic Enzyme
Cytochrome 450
Other Genes
C5X9A8_SORBI
A0A1B6Q1Q9
C5XAI4_SORBI
C5XPD4_SORBI
C5Y3Q3_SORBI
C5XJ51_SORBI
A0A1B6QD07
A0A1B6PSA0
C5XUF0_SORBI
C5XJ53_SORBI
C5Y9B7_SORBI
A0A1Z5S9M5
C5Z1F3_SORBI
C5XUF6_SORBI
A0A194YPL8
Flavonoid biosynthesis
Stilbenoid, diarylheptanoid and gingerol biosynthesis
Glutathione
C)
Tissue
Early Whole Seeds
Embryo
Endosperm
Z-score

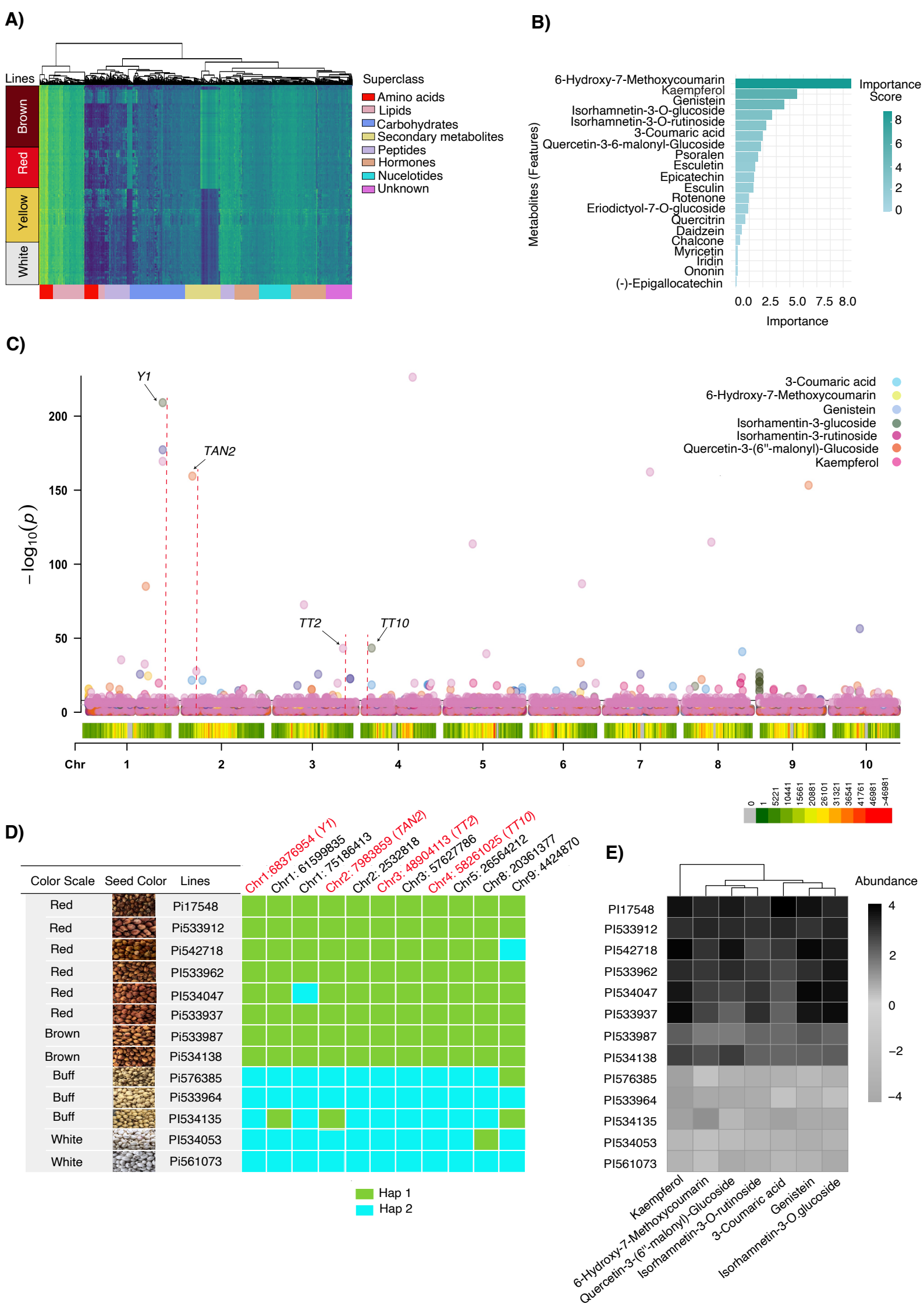
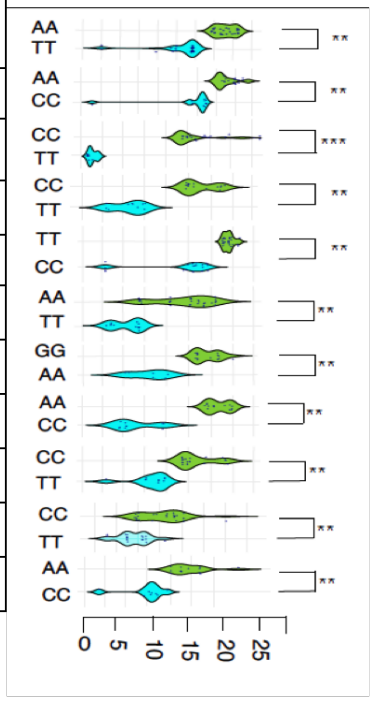

A)
Lines
Brown
Red
Yellow
White
Superclass
Amino acids
Lipids
Carbohydrates
Secondary metabolites
Peptides
Hormones
Nucelotides
Unknown
B)
Metabolites (Features)
6-Hydroxy-7-Methoxycoumarin
Kaempferol
Genistein
Isorhamnetin-3-O-glucoside
Isorhamnetin-3-O-rutinoside
3-Coumaric acid
Quercetin-3-6-malonyl-Glucoside
Psoralen
Esculetin
Epicatechin
Esculin
Rotenone
Eriodictyol-7-O-glucoside
Quercitrin
Daidzein
Chalcone
Myricetin
Iridin
Ononin
(-)-Epigallocatechin
0.0 2.5 5.0 7.5 8.0
Importance
Importance Score
8
6
4
2
0
C)
Y1
TAN2
TT2
TT10
$-\log_{10}(p)$
200
150
100
50
0
Chr
1
2
3
4
5
6
7
8
9
10
3-Coumaric acid
6-Hydroxy-7-Methoxycoumarin
Genistein
Isorhamentin-3-glucoside
Isorhamentin-3-rutinoside
Quercetin-3-(6''-malonyl)-Glucoside
Kaempferol
0
1
5221
10441
15661
20881
26101
31321
36541
41761
46981
>46981
D)
Color Scale
Seed Color
Lines
Chr1:68376954 (Y1)
Chr1: 61599835
Chr1: 75186413
Chr2: 7983859 (TAN2)
Chr2: 2532818
Chr3: 48904113 (TT2)
Chr3: 57627786
Chr4: 58261025 (TT10)
Chr5: 26564212
Chr8: 20361377
Chr9: 4424870
Red Pi17548
Red Pi533912
Red Pi542718
Red PI533962
Red PI534047
Red Pi533937
Brown Pi533987
Brown Pi534138
Buff Pi576385
Buff Pi533964
Buff PI534135
White PI534053
White Pi561073
Hap 1
Hap 2
E)
Abundance
4
2
0
−2
−4
PI17548
PI533912
PI542718
PI533962
PI534047
PI533937
PI533987
PI534138
PI576385
PI533964
PI534135
PI534053
PI561073
Kaempferol
6-Hydroxy-7-Methoxycoumarin
Quercetin-3-(6''-malonyl)-Glucoside
Isorhamnetin-3-O-rutinoside
3-Coumaric acid
Genistein
Isorhamnetin-3-O-glucoside

Table 1. Allelic variation at candidate SNPs influences accumulation of seven polyphenolic compounds in sorghum grains. The table summarizes significant SNP–metabolite associations, indicating the chromosomal position, allele types, 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 ($p < 0.01$ **).

S.No.	Polyphenolic Compounds	SNP (Chr:Pos)	Alleles	Allele frequence of Hap1 (in %)				Metabolite Fold change
				Red	Brown	Yellow	White	
1	Kaempferol	Chr1:68376954 *(Y1)*	A/T	71.00%	12.50%	10.42%	6.25%	
2	Genistein	Chr1:61599835	A/C	76.00%	12.12%	6.06%	6.06%	
3	6-Hydroxy-7-Methoxycoumarin	Chr1:75186413	C/T	72.00%	12.07%	10.34%	5.17%	
4	Quercetin-3-(6"-malonyl)-Glucoside	Chr2:7983859 *(TAN2)*	C/T	78.00%	12.50%	4.69%	4.69%	
5	Kaempferol	Chr2:2532818	T/C	76.00%	12.00%	8.00%	4.00%	
6	Genistein	Chr3:48904113 *(TT2)*	A/T	72.00%	12.50%	9.38%	6.25%	
7	Isorhamnetin-3-O-glucoside	Chr3:57627786	G/A	75.00%	9.80%	9.80%	5.88%	
8	6-Hydroxy-7-Methoxycoumarin	Chr4:58261025 *(TT10)*	A/T	71.00%	16.44%	5.48%	6.85%	
9	Quercetin-3-(6"-malonyl)-Glucoside	Chr6:14231414	G/A	72.00%	12.77%	6.38%	8.51%	
10	Quercetin-3-(6"-malonyl)-Glucoside	Chr9:4424870	A/C	73.00%	12.20%	7.32%	7.32%	
11	3-Coumaric acid	Chr8:20361377	C/T	71.00%	14.71%	8.82%	5.88%