Metabolome-Scale Genome-Wide Association Studies Reveal Chemical Diversity and Genetic Control of Maize Specialized Metabolites

Shaoqun Zhou,a,b,1 Karl A. Kremling,c,2 Nonoy Bandillo,c,3 Annett Richter,a Ying K. Zhang,a,d Kevin R. Ahern,a,c Alexander B. Artyukhin,a,d Joshua X. Hui,a Gordon C. Younkin,a,b Frank C. Schroeder,a,d Edward S. Buckler,c,e and Georg Jandera,b,5

1 Boyce Thompson Institute, Ithaca, New York 14853
2 Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, New York 14853
3 Present address: Department of Agronomy and Horticulture, University of Nebraska, Lincoln, Nebraska 68583
4 Present address: EAG Laboratories, Columbia, Missouri 65202
5 Address correspondence to: gj32@cornell.edu.

ORCID IDs: 0000-0002-3643-7875 (S.Z.); 0000-0001-5068-0372 (K.A.K.); 0000-0002-5941-9447 (N.B.); 0000-0003-0930-4265 (A.R.); 0000-0002-6062-7652 (Y.K.Z.); 0000-0002-8680-0687 (K.R.A.); 0000-0002-1734-266X (A.B.A.); 0000-0002-2932-8988 (J.X.H.); 0000-0002-3735-3534 (G.C.Y.); 0000-0002-4420-0237 (F.C.S.); 0000-0002-3100-371X (E.S.B.); 0000-0002-9675-934X (G.J.)

Cultivated maize (Zea mays) has retained much of the genetic diversity of its wild ancestors. Here, we performed nontargeted liquid chromatography-mass spectrometry metabolomics to analyze the metabolomes of the 282 maize inbred lines in the Goodman Diversity Panel. This analysis identified a bimodal distribution of foliar metabolites. Although 15% of the detected mass features were present in >90% of the inbred lines, the majority were found in <50% of the samples. Whereas leaf bases and tips were differentiated by flavonoid abundance, maize varieties (stiff-stalk, nonstiff-stalk, tropical, sweet maize, and popcorn) showed differential accumulation of benzoxazinoid metabolites. Genome-wide association studies (GWAS), performed for 3,991 mass features from the leaf tips and leaf bases, showed that 90% have multiple significantly associated loci scattered across the genome. Several quantitative trait locus hotspots in the maize genome regulate the abundance of multiple, often structurally related mass features. The utility of maize metabolite GWAS was demonstrated by confirming known benzoxazinoid biosynthesis genes, as well as by mapping isomeric variation in the accumulation of phenylpropanoid hydroxycitric acid esters to a single linkage block in a citrate synthase-like gene. Similar to gene expression databases, this metabolomic GWAS data set constitutes an important public resource for linking maize metabolites with biosynthetic and regulatory genes.

INTRODUCTION

Plants produce a wide variety of metabolites that are not directly related to their central energy metabolism or structural integrity. The distribution and diversity of these specialized metabolites are reflective of their essential functions in plant stress responses, particularly in their interactions with microbial pathogens and insect herbivores. For human societies, plant-derived specialized metabolites have long been valuable sources of flavor, nutrition, and pharmaceutical products. More recently, advances in genetics and molecular biology have led to the clarification of the complete biosynthetic pathways of plant specialized metabolites such as glucosinolates (Halkier and Gershenzon, 2006) and benzoxazinoids (Zhou et al., 2018). This knowledge has made it possible to manufacture some plant specialized metabolites at industrial scales, as well as to genetically improve crop species for increased pest and disease resistances.

The productivity of maize (Zea mays), the world’s most economically important crop species, with >700 million metric tons harvested each year (Ranum et al., 2014), is often limited by pathogens and insect pests (Mueller, 2017). For instance, in parts of Africa, ongoing epidemics of fall armyworm (Spodoptera frugiperda) have devastated local maize production, with far-reaching socioeconomic ramifications (Stokstad, 2017). These problems highlight the need for continuous genetic improvement of pest and disease resistance in the commercial maize germplasm to cope with the spatiotemporal fluctuations of biotic stresses. Even after millennia of artificial selection, maize is known for its genetic diversity at the population level (Buckler et al., 2006; Jiao et al.,...
IN A NUTSHELI

**Background:** Plants defend themselves against herbivores and pathogens by producing a myriad of specialized metabolites. This physiological process is encoded by biosynthetic and regulatory genes that are positioned throughout the plant genome. In staple crop species like maize (*Zea mays*), understanding how these metabolites are synthesized, regulated, and distributed across diverse cultivars will facilitate the development of novel pest-resistant cultivars and hence promote yield in stressful environments.

**Question:** The production of individual specialized metabolites has previously been associated with specific loci in the maize genome. In this study, we aimed to provide a metabolome-scale association resource that includes thousands of both known and as yet uncharacterized maize metabolites. We achieved this goal by measuring metabolite abundance in the seedling leaf tips and bases from more than 220 genetically diverse maize cultivars.

**Findings:** We established a high-resolution resource for associating maize metabolite abundance with genetic loci, validated the quality of these associations with known metabolites and their biosynthetic genes, and demonstrated the utility of this resource by discovering a likely biosynthetic gene for a relatively uninvestigated class of maize metabolites. Furthermore, by compiling associations between metabolites and genetic elements, we identified hotspots in the maize genome that have disproportional impacts on numerous metabolites and demonstrated that metabolites associated with shared genetic elements tend to be similar in their chemical structures. Finally, we demonstrated that both tissue type (leaf tips vs. bases) and maize population structure have a significant impact on the abundance of specialized metabolites.

**Next steps:** Our high-resolution metabolite mapping resource leads to testable hypotheses about the association of genetic elements with metabolite abundance. For each specific metabolite, further functional validation studies, such as experiments with genetic knockouts, are required to demonstrate the causal relationships between metabolite abundance and specific loci in the maize genome.

**RESULTS**

**Comparisons of Maize-Seedling-Leaf-Specialized Metabolomes between Tissue Types and Genetic Subpopulations**

We planted the 282-line Goodman Diversity Panel, along with inbred line B73 controls, and harvested seedling leaf tips and leaf bases for reversed-phase ultrahigh-performance liquid chromatography (UPLC)/high-resolution-MS analysis of 50% (v/v) methanol extracts, which measures a wide range of midpolarity metabolites. Due to the lack of seed germination for some maize inbred lines, losses during sample processing, and occasional low-quality UPLC-MS runs, full spectra were obtained for the following: leaf tips, negative ionization (221 inbred lines and 17 B73 control); leaf tips, positive ionization (258 inbred lines and 25 B73 control); leaf bases, negative ionization (220 inbred lines and 22 B73 control); and leaf bases positive ionization (223 inbred lines and 22 B73 control). Raw MS data are available at the Cyverse Discovery Environment (Zhou, 2019).

After filtering, >7,000 mass features were detected in at least three of the samples (see “Methods”; Supplemental Data Sets 1.
and 2). Principal component analysis (PCA) demonstrated that tissue type explained >30% of the observed variance (Figure 1A). Two-way analyses of variance (ANOVA) on the same data set showed that >97% of all the mass features analyzed were significantly influenced by tissue type (false discovery rate [FDR] < 0.05; Supplemental Data Set 3). By contrast, genetically defined maize population structure did not make a significant contribution to the variance (Figure 1B; Supplemental Data Set 3) and failed to separate in PCA, even when metabolomics data were analyzed independently within each tissue type (Figures 1C and 1D). Similarly, PCA within either tissue type showed no systematic bias introduced by the different blocks in which each maize inbred line was planted (Supplemental Figure 1).

Metabolomic Differentiation Based on Tissue Type and Genetic Subpopulation Are Driven by Different Classes of Specialized Metabolites

In the 200- to 400-nm UV light absorption chromatogram, neighboring peaks tended to have similar UV absorbance profiles. Specifically, peaks eluting between 240 and 360 s had UV light absorbance profiles resembling phenylpropanoids, peaks eluting between 360 and 460 s had typical benzoxazinoid-like UV light absorbance profiles, and those eluting after 460 s were flavonoid-like (Figures 2A and 2B). Measurement of a narrower window of UV light absorption also showed a distinct pattern in the three different elution periods (Supplemental Figure 2).

We plotted the extent of differentiation for each mass feature based on tissue type, genetic subpopulation, or their interactive effect, as measured by the negative logarithm of P values from two-way ANOVA, against their retention time (Figure 2C). These plots demonstrated that mass features from distinct ranges of the chromatogram, and hence different classes of specialized metabolites, were responsible for metabolomic differentiation by tissue and subpopulation, respectively. Specifically, mass features that were significantly different between leaf tips and bases were present in all three examined time intervals of the chromatograms, but were predominant in the flavonoid range (Figure 2C). By contrast, metabolites under significant influence from the maize subpopulation or its interaction with tissue type were almost

Figure 1. The Maize Specialized Metabolome Significantly Differentiates Leaf Tips and Bases, But Not Genetic Subpopulations. (A) PCA differentiates the metabolomes of maize seedling leaf tips and bases. (B) Consistently, more mass features are significantly different by tissue type than between subpopulations (two-way ANOVA, FDR < 0.05). Number of mass features that differ by tissue type (red), subpopulation (yellow), or their interactive effect (blue) are shown in the colored circles, with overlaps. (C) and (D) Within either tissue type, genetic subpopulations cannot be differentiated by PCA based on their overall metabolomic fingerprint.
exclusively found among the benzoxazinoids (Figure 2C). These visual patterns were confirmed with statistical comparisons of the extent of differentiation between the retention time groups (Figure 2D). Consistent with our visual assessment of the chromatograms, the 460- to 570-s time interval containing flavonoids showed the strongest differentiation between tissue types (Figure 2D, top). Together, these observations indicate that (1) flavonoid abundance is significantly different between the maize leaf tip and leaf base, and (2) benzoxazinoid content is different between lines but not enough to cluster subpopulations together when all the mass features are included in the analysis.

In support of the first observation, all major flavonoid-like UV absorption peaks were completely absent in leaf base samples and were only found in the more developmentally advanced leaf tips (Figure 3A). There are five maize genes that encode chalcone synthases, the enzyme catalyzing the first committing step in flavonoid biosynthesis (based on the B73 reference genome v4; Jiao et al., 2017). Analysis of previously published transcriptomic data (Kremling et al., 2018) showed that the two most strongly expressed chalcone synthase genes (GRMZM2G422750 and GRMZM2G380650) are expressed at a significantly higher level in the leaf tips than in the leaf bases in the GWAS panel (Figure 3B). Consistent with the UV light absorbance pattern, tandem MS (MS/MS) analyses of B73 whole seedling leaf extracts demonstrated that repeated fragment patterns are found within certain ranges of retention time (Supplemental Data Set 4). Queries of an online phytochemical MS/MS spectra library, along with prior experience with the benzoxazinoid compounds, led to the identification of 26 of the 94 metabolites detected under negative mode of electron spray ionization (Supplemental Data Set 4). Queries of an online phytochemical MS/MS spectra library, along with prior experience with the benzoxazinoid compounds, led to the identification of 26 of the 94 metabolites detected under negative mode of electron spray ionization (Supplemental Data Set 4). Among these identified metabolites, known phenylpropanoids eluted between 240 and 330 s, all of the flavonoids eluted after 450 s, and the most abundant benzoxazinoids eluted between 300 and 360 s, with two low-concentration compounds eluting at an earlier retention time.
The distribution of benzoxazinones, hydroxycitric acid, and naringenin as commonly occurring fragments in different retention time windows is illustrated in Figure 3C. As confirmation for the second observation, we identified mass features representing the most abundant benzoxazinoid compounds in maize seedling leaves, 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one-β-D-glucopyranose (DIMBOA-Glc) and its methylated glucoside derivative, 2-(2-hydroxy-4,7-dimethoxy-1,4-benzoxazin-3-one)-β-D-glucopyranose (HDMBOA-Glc). Consistent with prior observations (Meihls et al., 2013), DIMBOA-Glc was significantly depleted in tropical inbred lines, which instead contained significantly more HDMBOA-Glc ($P < 0.05$, ANOVA; Figure 4).

Structurally Related Metabolites Tend To Be Coregulated

The abundance of structurally related metabolites, which often arise from shared metabolic pathways, tends to be coregulated in plants. To investigate this phenomenon on a global scale in the maize metabolome, we constructed mutual rank-based correlation networks with the metabolomic data set using an exponential decay function ($\lambda = 50$) and detected overlapping coregulative clusters using the ClusterONE algorithm (Nepusz et al., 2012; Wisecaver et al., 2017). This analysis identified a similar number of significant clusters in leaf tips and bases ($P < 0.05$, Mann Whitney U-test; 15 in leaf tips and 16 in leaf bases). Consistent with the larger number of mass features detected in the leaf tip samples, clusters found in leaf tips were significantly larger than those found in leaf bases (mean = 100 versus mean = 58; $P < 0.005$, Student’s t test). We plotted the distribution of the retention times of mass features belonging to each coregulative network in 10-s bins and assessed the extent of retention time clustering of each network by calculating the cumulative frequency of the top three bins (Figure 5; Supplemental Data Sets 5 and 6). In 24 of the 31 detected coregulative networks, at least half of the mass features were located in the top three bins, suggesting that these coregulated mass features are structurally related. Interestingly, we found that the cumulative frequency of the top-three 10-s bins in coregulative networks derived from the leaf tip metabolome (57%) was significantly lower than that of leaf base metabolome-derived networks (75%; $P < 0.05$, Student’s t-test).

The Maize Metabolome Is Skewed toward Rare Metabolites

Our data set provides an opportunity to examine the diversity of specialized metabolites in maize. In both tissue types, there was a bimodal frequency distribution of the mass feature occurrence rate, as measured by the percent of maize genotypes where a mass feature was detected. Whereas 15% of mass features in either tissue type were present in >90% of all the genotypes,
>63% of mass features were found in less than half of the examined genotypes (Figure 6A). Mass features representing actual maize metabolites, rather than background noise in the MS assay, should have larger variance across genotypes than within the same genotype. The experimental design allowed us to calculate the between- and within-genotype variance of those mass features that were detected in the replicated B73 control samples planted in each flat. Together, these two variances can be used to estimate the broad sense heritability \( H^2 = \frac{\text{Var}_{\text{total}} - \text{Var}_{\text{B73}}}{\text{Var}_{\text{total}}} \) of each mass feature, assuming that within-genotype variance in B73 is a proxy for environmental variance (Figure 6B). Based on this assessment, we found that 25% of the mass features in leaf tips and 40% of those leaf bases had \( H^2 < 0.2 \), suggesting a relatively small component of genetic variation. The overall bimodal distribution pattern of mass feature occurrence remained intact after removing the low heritability mass features (\( H^2 < 0.2 \); Figure 6A).

If the less common mass features were the result of background variation in the MS data set, we would expect them to have a lower signal intensity than mass features resulting from actual maize metabolites. The mean nonzero intensity of each mass feature showed significant positive correlation \( (R^2 > 0.96) \) with its occurrence rate in both tissue types (Figure 6C), suggesting that less common mass features were indeed lower in abundance. However, given the slope of the regression line, a mass feature detected in only 10% of the genotypes would be on average <10-fold lower in intensity than a ubiquitous mass feature. By contrast, mass features of any given occurrence rate showed a 100-fold range in peak intensity (Figure 6C). Therefore, many or most of the less common mass features are likely to represent actual maize metabolites that are present in only a subset of tested inbred lines, rather than being noise in the MS chromatograms.

**The Genetic Architecture of Specialized Metabolites Is Complex and Is Strongly Influenced by Tissue Type, But Not Occurrence Rate**

The existing genotype data for the Goodman Diversity Panel (Bukowski et al., 2018; Kremling et al., 2018) make it possible to perform GWAS with each mass feature as an independent trait to...
understand its genetic architecture. Given the large number of traits to be analyzed, we employed a rapid recursive GWAS pipeline that was recently developed using an optimized general linear model (Kremling et al., 2018). Before this computation-intensive analysis, the LC-MS data set was further filtered by the rate of occurrence (detected in $10\%$ of all genotypes) and the broad sense heritability ($H^2$), using B73 to estimate environmental variation. Given the size of the inbred line population, it would not be possible to obtain accurate genetic mapping data for metabolites that are present in $<10\%$ of the tested genotypes, i.e. present in $<25$ inbred lines. Filtering for $H^2 \geq 0.2$ was done only for those mass features that were present in B73. Mass features that were not present in B73 did not have an estimate of heritability and were all included in the analysis. Altogether, 1,320 mass features from the leaf bases and 2,554 mass features from the leaf tips remained after filtering (Supplemental Data Sets 7 to 10), and GWAS was performed for each metabolite using 29 million SNPs (Bukowski et al., 2018). The raw GWAS results including all mass feature-SNP associations with $–\log(P) \geq 5$ are available at the Cyverse Discovery Environment (Zhou, 2019).

To investigate the complexity of metabolite regulation in maize seedlings, we collected the top-10 most strongly associated SNP markers for each mass feature and counted the SNPs in 10-kb segments spanning the maize genome. This showed that, in both leaf tips and leaf bases, the 10 most significant SNP associations were in an average of 7.4 distinct 10-kb blocks (Figure 7A). If the size of the scanned chromosomal segments was increased to 60 kb or 360 kb (Supplemental Figure 3), the average number of distinct blocks with significant SNP associations decreased to 6.8 and 6.2, respectively, but the overall shape of distribution was not

Figure 6. Mass Feature Occurrence Rates Are Bimodally Distributed and Are Positively Correlated With Their Average Nonzero Intensity.

(A) Distribution of mass features before (white) and after (gray) filtering by broad sense heritability in inbred line B73 ($H^2 > 0.2$) in either tissue type plotted in 10% incremental bins.
(B) Average heritability of mass features within each 10% occurrence incremental bins with exclusive lower boundaries and inclusive upper boundaries. Error bars = SE
(C) Each mass feature in either tissue type was plotted based on its occurrence rate (x axis) and the log of average nonzero intensity scale (y axis). Significant positive linear correlations between the two variables are found in both tissue types and indicated by blue dashed lines. Mass features that are above the 99% confidence interval of the overall linear correlation patterns are marked in red.
affected. Less than 9% of all mass features analyzed in either tissue type had their top-10 most strongly associated SNP markers located in less than four 10-kb blocks. We aligned 455 mass features detected in both leaf tips and leaf bases and compared their top-50 most strongly associated SNP markers in leaf tips and bases. The majority of these traits (405 out of 455) showed no overlap in their top-50 most strongly associated SNP markers, indicating that metabolic traits can be under distinct genetic regulatory mechanisms in different maize tissues, as has also been observed for glucosinolates in Arabidopsis (Chan et al., 2011). The most prevalent mass features (occurrence in >90% of inbred lines) mapped to significantly more loci than the less common ones in the population (Figure 7B), suggesting that components of central metabolism that are found in all maize plants are subject to more complex regulation than specialized metabolites that are not essential for maize survival under all environmental conditions and are not present in all maize inbred lines. No additional pattern could be clearly identified between the occurrence rate and genetic complexity of mass features. Together, these results indicate that maize metabolic traits have a complex genetic architecture that is under the control of numerous interacting genetic loci and varies by tissue type, as has been shown previously with Arabidopsis (Chan et al., 2010, 2011).

**Structurally Related Metabolites Tend To Be Coregulated**

In addition to identifying candidate genes significantly associated with individual metabolites of interest, the GWAS results can be used to look at the overall distribution of metabolite QTL. As has been reported previously for Arabidopsis (Chan et al., 2010), there were genomic hotspots that control the abundance of multiple maize metabolites. When the distributions of the most significantly associated SNP markers for 4,859 mass features were plotted in 10-kbp intervals across the maize genome, there were several loci to which a disproportionate number of metabolites were mapped (Figures 8A and 8C). In both leaf bases and leaf tips, three loci on chromosomes 1, 4, and 10, respectively, showed a large number of metabolite GWAS hits (Figures 8A and 8C). Additionally, there were genomic hotspots specific to either tissue type. The locations of these hotspots were consistent when the analysis included either the 10- or 50-most significantly associated SNP markers for each mass feature, as well as when varying the size of chromosomal blocks used to plot the QTL distribution (increasing from 10 to 60 or 360 kbp; Supplemental Figure 4).

We hypothesized that the genomic hotspots would contain one or more loci that regulate multiple structurally related metabolites derived from the same biosynthetic pathway. To test this hypothesis, we ordered the mass features based on the locations of their most strongly associated SNP markers in the maize genome and calculated the variance in retention time with a sliding window of 100 mass features with adjacent QTL in the genome. Because most mass features have their most strongly associated SNP markers at multiple positions in the genome, their retention times were included in the calculations more than once. Across the entire maize genome, there was a stable background level of retention time variance (Figures 8B and 8D). However, there were clear dips, i.e. lower variance in the retention time, below the background level at some loci. When results from this analysis were aligned to the previously generated plots of mass features per locus (Figures 8A and 8C), there was colocalization of dips in

![Figure 7](image-url)

**Figure 7.** Metabolic Traits Tend to Have Complex Genetic Architecture Irrespective of Their Heritability or Occurrence Rate. (A) and (B) Distribution of mass features in leaf tips and bases plotted based on the number of 10-kb LD blocks that contain one of their top-10 strongest associated SNP markers. Statistical mean of each distribution is given and marked by an arrow. This measurement was then compared across different occurrence rate bins (B) by one-way ANOVA and Tukey’s Honestly Significant Difference (HSD) test. Groups significantly different from each other (*P* < 0.05) are denoted with different letters above their respective columns. Error bars = SE.
features that have similar retention times.

Figure 8. Metabolite GWAS Hotspots Tend to Be Associated With Mass Features That Have Similar Retention Times.

(A) and (C) The number of mass features with at least one of their top-10 or top-50 most strongly associated SNP marker located in each 10-kbps block plotted for either tissue type. Results of neighboring chromosomes are shown in different colors, and results based on different top SNP threshold (10 or 50) are indicated by different color shades.

(B) and (D) Variance in the retention time of 100 mass features with adjacent GWAS hits in a sliding window across the genome were calculated and mapped based on the physical locations of the top SNP hits.

retention time variance with the genomic QTL hotspots (Figures 8B and 8D), indicating that the abundance of structurally related metabolites tends to be coregulated by the same genetic loci. This pattern was true for all three genomic hotspots shared by both tissue types, but was not necessarily valid all the time. For example, the second dip in retention time variance on chromosome 1 for the leaf tip data did not correspond to any increase in the number of mass features mapped to that locus. Conversely, mass features mapped to the leaf base-specific hotspot on chromosome 3 did not have similar retention times.

GWAS Reveal Both Known and Novel Genetic Loci Affecting Benzoxazinoid Accumulation

To determine the efficacy of gene identification by maize metabolite GWAS, we genetically mapped the abundance of two benzoxazinoid compounds, 2-(2,4-dihydroxy-7,8-dimethoxy-1,4-benzoxazin-3-one)-β-D-glucopyranose (DIM,BOA-Glc) and HDMBOA-Glc. GWAS with both metabolites confirmed known QTL-containing biosynthetic genes: Bx10 on chromosome 2 for DIM,BOA-Glc (Figure 9A; Handrick et al., 2016) and Bx10-12 on chromosome 1 for HDMBOA-Glc (Figure 9B; Meihls et al., 2013), with the most significantly associated SNPs being in linkage disequilibrium (LD) with the respective biosynthetic genes. The Bx10-12 genomic region also corresponds to the metabolite QTL hotspot found on chromosome 1 in both leaf tips and leaf bases (Figure 8B). Interestingly, in addition to the SNP markers in LD with the known biosynthetic genes, GWAS also identified SNP markers associated with the metabolites of interest in adjacent LD blocks, suggesting the presence of cis-regulatory loci at some distance from the genes of interest (Figures 9A and 9B).

A previously unknown locus affecting natural variation in HDMBOA-Glc was found on chromosome 9, with the most significantly associated SNPs located in a single 25-kb LD block (Figure 9B). We inferred biallelic haplotypes at the mapped loci on chromosome 1 and chromosome 9 based on SNPs within each locus and assigned inbred lines to one of the two haplotypes using a nearest-neighbor cladogram. Bx10-12 and the newly identified locus on chromosome 9 had an additive effect on HDMBOA-Glc content (Figure 9C). The 25-kb LD block on chromosome 9 contained the region immediately 3’ of GRMZM2G108309, a gene model encoding a predicted protein phosphatase 2C family protein. Transcript profiling data (Kremling et al., 2018) showed that GRMZM2G108309 expression levels were significantly different between the inbred lines carrying one or the other allele of the 25-kb linkage block on chromosome 9 (Figure 9D). We detected a significant difference in HDMBOA-Glc content when comparing the 20 inbred lines with the highest and lowest GRMZM2G108309 expression levels, respectively (Figure 9E; Supplemental Figure 5A). However, correlation analysis with the entire inbred line population showed no significant relationship between gene expression level and benzoxazinoid content (Supplemental Figure 5B)

Phenylpropanoid Hydroxycitric Acid Ester Isomers Found in Distinct Maize Subpopulations Are Associated with a Predicted Citrate Synthase

One of the patterns in our analyses of specialized metabolite diversity was that there were clear outliers to the overall positive correlation between the occurrence rate and mean nonzero intensity of mass features (Figure 6B). The majority of these outliers were concentrated in the high occurrence rate range, where the linear correlative relationship was capped by maximal occurrence rate. However, in both leaf tips and bases, a group of high intensity mass features were detected in 20% or fewer of the examined genotypes (red dots in the left of the graphs in Figure 6B). Among these outliers, there were three mass features with characteristic phenylpropanoid-like UV light absorbance profiles and two common daughter ions with m/z = 189.004 and m/z = 127.003 under negative electron spray ionization (Figure 10A). Furthermore, the MS data indicated that the phenylpropanoid moieties in these three metabolites differed by masses consistent with the addition of hydroxyl (m/z 15.99) and methyl groups (m/z 14.01), respectively (Figure 10A).

In maize inbred lines where these predicted phenylpropanoid metabolites were not detected, at least one additional peak was present in each of the three m/z channels, all of which had earlier
retention times than those that were detected in <20% of maize lines (Figure 10B). These earlier-eluting peaks also had phenylpropanoid-like UV absorption peaks and had the same daughter ions in MS/MS. The earlier elution times of the peaks with higher occurrence rate suggest that they are structural isomers with higher polarity relative to the three high-abundance peaks that are present in <20% of maize inbred lines (Figure 10A). Two-dimensional nuclear magnetic resonance (NMR) spectra of the purified higher-polarity peaks, which are present in B73 and the majority of other inbred lines, showed that they are ester conjugates of coumaric acid, caffeic acid, and ferulic acid with 2-hydroxycitric acid, respectively (Figure 10B; Supplemental Figure 6; Supplemental Data Set 11). However, when attempting to isolate the corresponding lower-polarity isomers, which were found in <20% of inbred lines (CML247 is shown as an example in Figure 10B), the isolated samples rapidly degraded in the NMR solvent, and hence their exact chemical structures could not be elucidated.

To examine how the pairs of phenylpropanoid hydroxycitric acid ester isomers were distributed among maize genotypes, we constructed a dendrogram of the Goodman Diversity Panel using a 66,000 SNP data set (derived from Samayoa et al., 2015; Figure 11A) and plotted the abundance of the three pairs of structural isomers (Figure 11B) relative to this tree. In both tissue types, the rare isomers tended to co-occur and were overrepresented in the tropical inbred lines. Furthermore, the presence of the two groups of isomers was generally mutually exclusive. However, these trends were not perfect, particularly in the case of the isomers with m/z = 369.046, both of which were sporadically distributed across the population in the leaf bases without necessarily co-occurring with the other metabolites. The metabolism of these pairs of phenylpropanoid-containing isomers is likely also under developmental regulation, as demonstrated by the different distribution patterns in leaf tips and leaf bases.

GWAS showed that, for all three pairs of phenylpropanoid-hydroxycitric acid ester isomers, the most strongly associated...
Figure 10. Phenylpropanoid-Containing Mass Features Coelute With Common Daughter Ions.

(A) MS scans of three mass features coeluting with phenylpropanoid-like UV absorbance peaks are shown. The parental ions and the two shared daughter ions are labeled with their exact m/z measurement.

(B) The predominant peaks at each specific m/z range eluting at different retention times likely represent different structural isomers of the same compound in inbred lines B73 and CML247.

(C) Structures of ester conjugates of coumaric acid, caffeic acid, and ferulic acid, with 2-hydroxycitric acid.
Three Pairs of Hydroxycitric Acid Conjugates Have Complementary Distribution Patterns and Common Regulation of Abundance in the Maize Diversity Panel.

(A) Dendrogram of the 282 maize inbred lines included in the GWAS panel constructed with the distance matrix calculated from 66,000 SNP markers. The estimated concentration of three pairs of phenylpropanoid-containing structural isomers are shown in a color scale (blue, low abundance; yellow, high abundance; white, no sample measured) for each maize inbred line. Each monophyletic group was assigned to a genetic subpopulation as defined in Flint-Garcia et al. (2005) based on the predominant group assignment for the individuals within that clade.

(B) The pairs of mass features shown in (A) have different retention times in minutes and were detected in negative ionization mode (m/z).

Figure 11.
SNP markers were located within a 10-kb LD block on Chromosome 4 (Figure 11C), in the same position as a metabolite QTL hotspot for both leaf tips and leaf bases that is shown in Figure 8. In the B73 reference genome, this LD block was contained within a single gene model, GRMZM2G063909, which was annotated as an ortholog of Arabidopsis and rice citrate synthase genes (Figures 11D and 11E). GRMZM2G063909 expression was not significantly different between maize inbred lines accumulating different structural isomers of the phenylpropanoid hydroxycitric acid esters (Figure 11F; data from Kremling et al., 2018), suggesting that, consistent with all linked SNPs being in the coding region, structural variation in the encoded enzyme is more likely to be responsible for the observed metabolic differences.

To independently verify the genetic association between GRMZM2G063909 and the phenylpropanoid hydroxycitric acid ester isomers, we examined two sets of sixth-generation recombinant inbred lines (RILs) derived from Ki11 × B73 and CML247 × B73 (McMullen et al., 2009) to identify RILs with residual heterozygosity at GRMZM2G063909. Whereas B73 encodes the temperate maize isomers of the phenylpropanoid hydroxycitric acid esters, Ki11 and CML247 encode the tropical maize isomers. Progeny of these RIL families segregated near isogenic lines that were either heterozygous or homozygous for one or the other parental allele (Figures 12A and 12B). MS assays showed perfect cosegregation between the genotypic markers and the two classes of phenylpropanoid hydroxycitric acid esters (Supplemental Figure 7). Whereas the two tropical inbred lines also accumulated small amounts of the more polar isomers that are characteristic of temperate inbred lines, B73 tissues did not contain any of the less polar phenylpropanoid hydroxycitric acid esters (Figures 12C to 12E). Furthermore, heterozygous lines showed intermediate phenotypes, producing both isomers, but in lesser abundance than either homozygote.

**DISCUSSION**

Technological advances in MS and accumulating high-density genotype data are enabling metabolome-scale quantitative genetics studies. Prior studies of this type have focused on topics ranging from primary metabolites of nutritional interest to known specialized metabolites in both model organisms and economically relevant crop species (Chan et al., 2010, 2011; Riedelsheimer et al., 2012; Chen et al., 2014; Wen et al., 2014; Matsuda et al., 2015). However, unlike transcriptomic data, where each transcript can be functionally annotated to at least some extent based on sequence homology and structural features, most mass features from nontargeted metabolomics datasets represent unknown metabolites, and MS data provide incomplete information about their structures. Our metabolome-scale correlation network analyses (Figure 5) and GWAS (Figure 8; Supplemental Data Sets 12 and 13) provide a basis for structural and functional assignments of the many unknown metabolites in maize seedlings. These metabolomic genetic mapping data complement other currently available approaches to metabolite identification, including large-scale coelution tests with known compounds and the construction of molecular networks based on shared MS/MS fragments, which are indicative of structural similarity (Nguyen et al., 2013; Matsuda et al., 2015).

We conducted this GWAS of maize metabolomic data with a single individual plant of each maize inbred line. The replicates did not comprise individual plant lines, but rather the different alleles at each locus in the genome. Because all loci are biallelic (nonbiallelic ones were filtered out), each allele was sampled an average of >100 times in our genetic mapping panel. There are thousands of available maize inbred lines, and many of these have been fully genotyped. Therefore, our experiments were not limited by maize genetics, but rather by the resources that were available for growing plants and running MS assays. If more resources had been available, we would have analyzed additional independent lines rather than replicates of this set, thereby gaining a greater amount of genetic mapping resolution in the GWAS.

Transcriptome data for the Goodman Diversity Panel were also collected with single replicates to conduct a GWAS of maize gene expression levels (Kremling et al., 2018). In addition to the genetic mapping data, this publication also provides a resource for other maize researchers who want to compare maize gene expression levels to other data that they have generated. Similar gene expression resources are available and commonly for Arabidopsis and other plant species (e.g. www.arabidopsis.org). There will always be potential problems in the application of such gene expression resources to plants that were grown in different environments. However, using the same maize growth conditions as Kremling et al. (2018), we have been able to minimize such variation in our comparisons of maize metabolite content and gene expression.

Our data sets allowed us to assess variation in the maize specialized metabolome in two tissue types across a diverse population of inbred lines. The metabolomes of both leaf tips and leaf bases demonstrated bimodal distributions, with a relatively small core component and a large number of rare mass features (Figure 6A). Compared with the presence/absence distribution of gene expression in the maize pan-transcriptome (Hirsch et al., 2018), GWAS identified a common locus on chromosome 4 that regulates the abundance of all of the identified mass features from (A). Only the results from the more polar isomers with structural confirmations are shown.

**Figure 11.** (continued).

(C) GWAS identified a common locus on chromosome 4 that regulates the abundance of all of the identified mass features from (A). Only the results from the more polar isomers with structural confirmations are shown.

(D) SNP markers most strongly associated with the phenylpropanoid hydroxycitric acid esters plotted based on their physical location in the maize genome (x axis) and level of association with the metabolites (y axis), and overlaid on the predicted transcripts of GRMZM2G063909 located at the same locus.

(E) Pairwise correlation coefficients between SNP markers around the candidate gene were calculated to demonstrate that the significantly associated SNP markers are not in LD with any adjacent gene model.

(F) GRMZM2G063909 expression in leaf tips and bases was obtained from Kremling et al. (2018) and compared between maize inbred lines accumulating different phenylpropanoid hydroxycitric acid ester isomers. No significant difference (N.S.) in expression was found in either tissue type (P > 0.05; Student’s t test).
2014), the profiles of our metabolomic data are much more left-skewed, i.e. the majority of maize metabolites are present in <50% of the inbred lines. This is perhaps reflective of the more commonly nonessential nature of specialized metabolites relative to transcripts, which contain large numbers of housekeeping genes that are involved not only in primary metabolism but also other essential cellular functions. However, the observed distribution differences could also result from the greater sensitivity of RNA-sequencing–based transcriptomics compared with metabolomics, which would allow detection of rare transcripts in a larger number of maize inbred lines.

Broad sense heritability of metabolite content, estimated as

\[
\Phi^2 = \frac{\text{Variance(total)} - \text{Variance(B73)}}{\text{Variance(total)}}
\]

differs according to metabolite prevalence in the population. In particular, the on-average lower heritability of less common mass features suggests that some of them may be artifacts of the MS assay. Nevertheless, there are a significant number of uncommon metabolites with high heritability. The somewhat paradoxical observation of negative heritability is the result of different sample sizes. Whereas only ~20 B73 samples were used to calculate environmental variance, total population variance was calculated based on >200 inbred lines. Estimating environmental variance based on only one genotype is a somewhat imperfect approach. However, the resources for measuring multiple replicates of all maize inbred lines by MS were not available.

PCA of the metabolome clearly differentiates leaf tips and leaf bases, but not different maize subpopulations (Figure 1). In the case of the two tissue types, one factor that contributes to their separation by PCA is the presence and absence of flavonoids. Prior studies also have documented such developmental regulation of flavonoids and other maize metabolites (Jahne et al., 1993; Pick et al., 2011). It is likely that the more exposed position of leaf tips requires more flavonoids for defense against biotic and abiotic stress. Despite the different benzoxazinoid (Figure 4) and phenylpropanoid hydroxycitric acid ester (Figure 11) profiles in tropical and temperate maize, PCA did not differentiate maize subpopulations (Figure 1). A possible cause of this effect could be a random presence/absence distribution of the large number of uncommon maize metabolites in the different maize subpopulations.

Our study provides a metabolome-scale evaluation of the complex genetic architecture of metabolic traits in maize seedling leaves. As observed previously in Arabidopsis (Chan et al., 2010), most maize metabolites have multiple biosynthetic or regulatory loci significantly associated with them. Moreover, with the exception of the most common metabolites, we observed no consistent correlative relationship between genetic architecture complexity and occurrence rate of metabolic traits (Figure 7). We speculate that individual metabolic traits are regulated by different sets of genetic loci in different subsets of the maize population. This observation also could explain the significantly higher number of mapped loci associated with the most ubiquitous mass features (Figure 7B), which are more likely to be involved in primary metabolism.

Another omic-scale pattern identified from our study is tissue-specific and shared metabolite QTL hotspots (Figure 8). This nonuniform distribution of significant GWAS hits is comparable to results from published Arabidopsis and rice metabolite GWAS.
(Chan et al., 2010; Chen et al., 2014). Similar MS fragmentation and UV light absorbance profiles of metabolites in the QTL hotspots indicate that structurally related metabolites tend to be coregulated by shared genomic loci. The presence of these metabolite QTL hotspots generates hypotheses for the regulation of specialized metabolism both for specific metabolites and at a system scale. Further studies of these loci could lead to the elucidation of the underlying physiological mechanisms of these genetic associations.

The QTL hotspot on Chromosome 1 (Figure 8) represents a 110-kb region containing the paralogous BX10, BX11, and BX12 genes, encoding O-methyltransferases that catalyze the biosynthesis of HDMBOA-Glc (Meihls et al., 2013; Handrick et al., 2016). Mass features mapped to this locus include HDMBOA-Glc, DIMBOA, and other benzoazinoid compounds. However, several mass features that were not associated with known benzoazinoid compounds also mapped to this locus, suggesting the regulation of other classes of maize specialized metabolites. Such regulation could be indirect, as benzoazinoid degradation has been shown to induce other maize defense responses (Ahmad et al., 2011; Meihls et al., 2013).

The identification of a HDMBOA-Glc regulatory locus on chromosome 9, which was not identified in several other bi-parental mapping studies (Meihls et al., 2013; Zheng et al., 2015; Handrick et al., 2016), highlights the power of investigating a population with broader genetic diversity and denser SNP markers. However, it also illustrates one of the potential shortcomings of this GWAS approach. The absence of a large part of the maize pan-genome in any given inbred line, in combination with the use of the B73 genome sequence as the basis of our GWAS, likely has led to incomplete identification of metabolite QTL. The expression level of GRMZM2G108309 on chromosome 9 is associated with HDMBOA-Glc content (Figure 9), but all of the associated SNPs are downstream of this candidate gene. Although this could represent 3′ regulation of GRMZM2G108309 expression, another possibility is that we have mapped a locus that is not present in B73 but regulates the accumulation of HDMBOA-Glc in as-yet unsequenced maize inbred lines. Similarly, the metabolite QTL hotspot on chromosome 10, which influences a dozen mass features in both leaf tips and bases, spans a 30-kb region containing seven retroelements and a low confidence gene model in the B73 genome. The abundance of transposon genes in this region of the B73 genome suggests that there may be absence/absence variation among the diverse maize inbred lines, and that the causative gene may not be present in the B73 reference genome. As more high-quality maize genome sequences become available in the next few years, it will be possible to look for such genes that may be missing from B73 and other currently available maize genomes.

The presence of other metabolite QTL hotspots may also lead to the identification of previously unknown regulatory loci of maize metabolism. For instance, nine mass features found in leaf tips had at least one of their 10 most-associated SNP markers located within a 20-kb region on chromosome 3 (Figure 8). This genomic region contains a single gene model, GRMZM2G143723, which is analogous to a rice Cys2His2-like fold group zinc finger protein and thus may represent a regulatory gene for this group of metabolites.

Although phenylpropanoid hydroxycitric acid esters were previously identified as maize metabolites (Ozawa et al., 1977; Plenchamp, 2013), their biosynthesis and structural diversity have not been investigated. Coding-sequence variation in the identified citrate synthase-like gene (GRMZM2G063909; Figure 11) likely leads to the formation of multiple isomers of coumaroyl-, caffeoyl-, and feruloyl hydroxycitric acid. Further experiments will be needed to confirm the effects of specific SNPs, both in vivo using transgenic maize plants and in vitro with enzyme activity assays.

We identified the more polar phenylpropanoid hydroxycitric acid esters as the 2-o-acylated derivatives of hydroxycitric acid (Figure 10C). Due to their instability, it was not possible to determine the structures of the less polar isomers that are typical of tropical maize. However, given that 3-o-acylated hydroxycitric acid esters are prone to acid- or base-catalyzed elimination of the 3-o-acyl moiety, we hypothesize that these later-eluting isomers represent the corresponding 3-o-acylated hydroxycitric acid esters of coumaric acid, caffeic acid, and ferulic acid, respectively. The accumulation of phenylpropanoid hydroxycitric acid esters is induced by the soil bacterium Pseudomonas putida (Plenchamp, 2013). Thus, it is tempting to speculate that these metabolites have a defensive function and that the two groups of isomers represent different defensive properties of this pathway that have been selected during the breeding of temperate and tropical maize varieties, respectively.

By demonstrating the use of maize inbred lines from the Goodman Diversity Panel to map metabolite quantitative traits to the single-gene or near-single-gene level (Figures 9 and 11), we have generated a rich resource of high-resolution associations between maize metabolic phenotypes and genetic loci. Large gene expression data sets generated with DNA microarrays or Illumina-based sequencing (RNA sequencing) are frequently used for experimental validation and to generate ideas for further research. In a similar manner, our metabolomic association mapping data constitute a community resource that will allow for the formulation of testable hypotheses and functional analysis of diverse maize metabolites. Future researchers who are investigating maize metabolites LC-MS will be able to link their identified mass features with our genetic mapping data to identify potential biosynthetic and regulatory loci. For instance, if our mapping data (Supplemental Data Sets 12 and 13) had been available, the authors who previously reported the discovery of phenylpropanoid hydroxycitric acid esters in maize (Ozawa et al., 1977; Plenchamp, 2013) could have immediately associated their metabolites with GRMZM2G063909, the citrate synthase-like gene that regulates their relative abundance (Figures 8 and 9). Conversely, someone investigating the function of GRMZM2G063909 could look up this gene in our data tables to identify mass features whose abundance is influenced by this locus. Furthermore, the availability of our raw MS data sets and genetic mapping data in Zhou (2019) will enable future analyses beyond what we have done for this project.

Our metabolomic assays, which were focused on midpolarity metabolites isolated in a single extraction of maize seedling leaves, provide only a snapshot of the total maize metabolome. Future research will need to be directed at identifying metabolites that are extracted by other methods, from other maize tissues and growth stages, as well as under biotic stress conditions that are likely to induce the production of metabolites that are otherwise not abundant enough to be reliably detected. Nevertheless, the genetic loci and alleles that we have identified will be useful for
marker assisted breeding to increase the production of targeted maize metabolites, thereby promoting pathogen resistance or other important agronomic traits.

METHODS

Plant Growth and Tissue Collection

All maize (*Zea mays*) seeds were originally obtained from the Maize Genetics Cooperation Stock Center. To ensure comparability of our metabolomics data with previous published transcriptomics data collected in the same tissue types, the same seed stocks were used and the growth conditions were replicated in the same greenhouse space at the same time of the year, early June (Krenling et al., 2018). Eight seeds of each maize genotype were planted in ~50 cm³ vermiculite, and the entire diversity panel was fitted into twenty-six 30-cm × 60-cm 96-cell flats. Plants were grown in a greenhouse under natural sunlight. To control for microenvironmental variation, eight B73 seeds were included in each flat, and all flats were randomized daily. When the third leaf had visibly emerged from the whorl, 2 cm of tissue from the leaf tips and bases of these leaves were collected. Tissue was collected only from maize inbred lines where at least two of the planted seeds had germinated. For leaf base tissues, seedlings were cut at the soil line and unrolled to expose the leaf base. For each maize inbred line, tissues from two seedlings (19 to 120 mg) were pooled, weighed, snap-frozen in liquid nitrogen, and stored at ~80°C for later metabolite extraction. The time between cutting of the maize seedlings and placing the weighed samples into liquid N was <3 min. To minimize the effects of diurnal variation in the maize metabolome, all samples were harvested in a 2-h time window between 10 AM and noon. With the exception of inbred line B73, single replicates were collected for each maize line that was analyzed.

Metabolomics Analyses and Data Preprocessing

Frozen seedling leaf tissues were extracted with 200 μL of 50% (v/v) methanol acidified with 0.1% (v/v) formic acid, and analyzed on a Supelco reverse-phase C18 column (Sigma-Aldrich) on a Dionex 3000 Ultimate UPLC-diode array detector system coupled to a Q Exactive mass spectrometer (Thermo Fisher Scientific). The two mobile phase solvents were water (Solvent A) and acetonitrile (Solvent B), both acidified with 0.1% (v/v) formic acid. The mobile phase gradient ran from 95% (v/v) Solvent A at 0 min to 100% Solvent B at 10.5 min with a curvature of 2 to optimize compound separation while reducing the runtime of each individual analysis to accommodate our large sample size. Each extract was separately analyzed with both positive and negative modes of electron spray ionization. A blank sample (0.1-μL 100% methanol) was added at the beginning of each batch and between every 60 runs to wash potential residuals off the LC column and to allow compensation for background signals. Raw MS output files were converted to “mzxml” formats with the MSConvert tool using an inclusive MS level filter (Chambers et al., 2012). Metabolite quantification was estimated with signal intensity acquired through the XCMS-CAMERA mass scan data processing pipeline (Tautenhahn et al., 2008; Benton et al., 2010; Kuhl et al., 2012). To account for potential rare metabolites occurring in this diverse population, the minimal sample threshold for keeping a mass feature was set at 3” at the grouping step of the XMCS processing. For initial chemical diversity analyses, LC-MS results from different tissue types were processed together to allow comparison across tissue types. For tissue-type-specific statistical analyses and GWAS, only LC-MS results from the same tissue type were aligned to one another and processed as a group to avoid widespread zero values introduced by tissue-specific mass features.

Mass features detected by the XCMS-CAMERA pipeline were filtered based on their retention times (60 to 630 s) and exact masses (m/z < 0.5 at first decimal point), and peaks annotated as naturally occurring isotopes were removed. Specific parameters that were used are described in the Supplemental Methods. Peaks annotated as MS adducts were retained because we had observed a high rate of false annotation of real metabolites into this category. Mass feature quantification was then corrected by tissue fresh-weight and normalized by the total ion concentration of each sample to account for technical variation.

Chemical Diversity Analyses

Measurement of each mass feature across the diversity panel was log-transformed for multivariate analyses. Zero values were changed to “1” before log-transformation. This data set was uploaded to the MetaboAnalyst 3.0 online tool platform for PCA and two-way ANOVA (Xia et al., 2015). The mass feature list was further filtered by interquartile range and Pareto-scaled before these analyses. In both tissue types, a small number of genotypes had only data available from either positive or negative ionization mode analysis due to a failed run under the other mode. These missing data were imputed to population medians to minimize their influence on the overall data structure without losing the usable data. Each maize inbred line was assigned to a genetic subpopulation as defined in Flint-Garcia et al. (2005). All other statistical analyses and data visualization were performed in R (R Core Team, 2018) and Excel (Microsoft).

MS/MS Analysis

B73 seedlings were grown in ~50 cm³ vermiculite under natural sunlight, and the third leaves from three seedlings were independently extracted and subjected to the same LC inlet method as described above in the Metabolics Analysis and Data Processing section. In addition to the full MS scan in the original method under either positive or negative mode of electron spray ionization (collision energy = 20 V), another MS scan focused on the top-5 features from each previous scan was included to provide extra fragmentation. MS/MS spectra from the results were extracted and the major fragment peaks were manually identified. Each MS/MS spectrum was queried to the ReSpect for Phytochemicals online (Sawada et al., 2012), and “three or more exact matches between input and a reference chemical” was set as the criterion for a peak identification.

Benzoazainozid Identification

Benzoazainoids were identified based on known masses, purified standards, known benzoazainoid profiles of 25 maize inbred lines that are included in the GWAS panel (Meihls et al., 2013), and several years of experience in the identification and analysis of maize benzoazainoids (Meihls et al., 2013; Mijares et al., 2013; Betsiashvili et al., 2015; Tzin et al., 2015a, 2015b, 2017; Zhou et al., 2018).

Structural Confirmation of Phenylpropanoid Hydroxycitric Acid Esters

The three phenylpropanoid hydroxycitric acid esters examined in this study were extracted overnight at 4°C from bulk snap-frozen B73 seedling leaves with 50% (v/v) methanol acidified with 0.1% (v/v) formic acid. Solid debris was removed through centrifugation and the crude extract was concentrated with a Buchi Rotovapor. Target compounds were separated with a water:acetonitrile gradient on an Eclipse XDB C18 column (ZORBAX) using a model no. 1100 HPLC system (Agilent). Purified compounds were dried, weighed, and redissolved in pure methanol. NMR spectroscopy analyses were performed on a Unity INOVA 600 instrument (Varian Medical Systems) with the following conditions: 256 scans for 1H NMR; number of scans for each point (nt) = 16 and number of points acquired in the second
Local LD Estimation, Haplotype Inference, and Inbred Line Relationships

SNP marker data across the same GWAS diversity panel around the most strongly associated SNP markers for each trait were downloaded from the Cyverse Discovery Environment (plant/home/shared/panzea/hapmap3/hmp321) and used to estimate local LD with the pairwise correlation with sliding window algorithm implemented in TASSEL 5.2.40 (Bradbury et al., 2007). Biallelic haplotypes at genetic loci associated with HDMBOA-Glc on Chromosome 1 and Chromosome 9 were inferred based on SNP data at either locus with a nearest-neighbor cladogram also implemented in TASSEL 5.2.40. A smaller SNP data set (Samayoa et al., 2015) with filtering for maximal missing data (<80%), maximal heterozygosity level (<50%), and minimal minor allele frequency (>30%) was used to estimate the phylogenetic relationship among the maize inbred lines included in this study. Approximately 66,000 SNP markers were retained after the filtering process and used to calculate a pairwise distance matrix with TASSEL 5.2.40. This distance matrix was then used to construct a dendrogram using a hierarchical clustering algorithm with the Ward method implemented by the “hclust” function in R.

Accession Numbers

Sequence data discussed in this article can be found in the GenBank library/MaizeGDB gene records under the following accession numbers: NC_024465.2 (35957502..35959127)/GRMZM2G108894; NC_024462.2 (204719770..204722586)/GRMZM2G114471; NC_024462.2 (196893216..196896503)/GRMZM2G422750; NC_024462.2 (230953451..230957085)/GRMZM2G151227; NC_024459.2 (233440789..233442623)/GRMZM2G380650; NC_024458.2 (67312119..67314042)/GRMZM2G311036; NC_024459.2 (67228930..67231301)/GRMZM2G336824; NC_024459.2 (67109115..67116836)/GRMZM2G023325; NC_024460.2 (238143556..238145183)/AC148152.3 _FG005; NC_024467.2 (128358117..128362499)/GRMZM2G108309; NC_024462.2 (238977913..23898897)/GRMZM2G063909; NC_024461.2 (5057135..5060050)/GRMZM2G114372.3.

Raw LC-MS result files, LC-MS/MS results files, full FastGlm GWAS results files for all filtered mass features, and raw 1D proton, 2D COSY (Homonuclear Correlation Spectroscopy), HSQC (Heteronuclear Single Quantum Coherence Spectroscopy), and HMBC (Heteronuclear Multiple Bond Correlation) NMR spectra are freely accessible through the Cyverse Discovery Environment (doi.org/10.25739/9dsj-kw33). The method used to extract annotated peak intensity from mass spectrometry raw files with the XCMS-CAMERA pipeline is also available under the same directory in the Cyverse Discovery Environment.

Supplemental Data

Supplemental Figure 1. Maize seedling leaf specialized metabolomes are not significantly different among experimental blocks.

Supplemental Figure 2. Major peaks from distinct ranges of the chromatogram share characteristic UV absorbance profiles.

Supplemental Figure 3. Frequency distributions of genetic architecture complexity of metabolic traits are consistent across different LD window sizes.

Supplemental Figure 4. Presence and locations of metabolite QTL hotspots are consistent across different LD window sizes.

Supplemental Figure 5. Correlation of GRMZM2G108309 expression and HDMBOA-Glc content.

Supplemental Figure 6. NMR spectra.

Supplemental Figure 7. Chromatograms of isomers of phenylpropanoid hydroxycitric acid ester.
Supplemental Data Set 1. Mass features detected from positive electron spray ionization MS in seedling leaves of diverse maize inbred lines.

Supplemental Data Set 2. Mass features detected from negative electron spray ionization MS in seedling leaves of diverse maize inbred lines.

Supplemental Data Set 3. Two-way ANOVA results of mass features based on tissue type and genetic subpopulation.

Supplemental Data Set 4. Major fragments from MS/MS analysis of B73 seedling leaf extract under negative mode of electron spray ionization.

Supplemental Data Set 5. Overlapping significant correlative networks of mass features detected in tips and bases of maize seedling leaves.

Supplemental Data Set 6. Retention time distribution of mass features in each correlative network.

Supplemental Data Set 7. Mass features detected from negative electron spray ionization MS in seedling leaves tips of diverse maize inbred lines.

Supplemental Data Set 8. Mass features detected from positive electron spray ionization MS in seedling leaves tips of diverse maize inbred lines.

Supplemental Data Set 9. Mass features detected from negative electron spray ionization MS in seedling leaves bases of diverse maize inbred lines.

Supplemental Data Set 10. Mass features detected from positive electron spray ionization MS in seedling leaves bases of diverse maize inbred lines.

Supplemental Data Set 11. 2D-NMR data of maize phenylpropanoid hydroxycytic acid esters.

Supplemental Data Set 12. Top-50 most significantly associated SNP markers of mass features detected in maize seedling leaf tips.

Supplemental Data Set 13. Top-50 most significantly associated SNP markers of mass features detected in maize seedling leaf bases.

Supplemental Methods. Extracting annotated peak intensity from MS raw files with the XCMS-CAMERA pipeline.

ACKNOWLEDGMENTS

This work was supported by the U.S. National Science Foundation (awards 1139329 and 1339237 to G.J.), the Graduate Research Fellowship Program (award DGE-1650441 to K.A.K.), the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy (award DEAR0000598 to N.B.), the Howard Hughes Medical Institute (Faculty Scholar to F.C.S.), and the U.S. Department of Agriculture – Agricultural Research Service (to E.S.B.). This work was also partially supported, indirectly, by the U.S. National Science Foundation (Major Research Instrumentation award CHE-1531632 to Cornell University NMR Facility where some work was performed). We thank Zack Miller and Lynn Johnson for programming and members of the Jander lab for assistance in plant growth and sample collection.

AUTHOR CONTRIBUTIONS

All experiments and analyses were conceived and performed by S.Z.; K.R.A., J.X.H., and G.C.Y. assisted in plant material and LC-MS sample preparations; K.A.K., N.B., and E.S.B. provided advice in experimental design, plant preparation protocols, and performed bioinformatic analysis; A.R. performed experiments with near-isogenic lines; Y.K.Z., A.B.A., and F.C.S. performed 2D-NMR spectroscopic analyses; G.J. provided advice in all experiments and analyses; S.Z. and G.J. prepared the article.

Received October 15, 2018; revised May 3, 2019; accepted March 27, 2019; published March 28, 2019.

REFERENCES


Metabolome-Scale Genome-Wide Association Studies Reveal Chemical Diversity and Genetic Control of Maize Specialized Metabolites

*Plant Cell* 2019;31;937-955; originally published online March 28, 2019; DOI 10.1105/tpc.18.00772

This information is current as of March 12, 2021

| Supplemental Data | /content/suppl/2019/03/28/tpc.18.00772.DC1.html  
/content/suppl/2019/04/03/tpc.18.00772.DC2.html |
|-------------------|-------------------------------------------------|
| References        | This article cites 43 articles, 12 of which can be accessed free at:  
/content/31/5/937.full.html#ref-list-1 |
| eTOCs             | Sign up for eTOCs at:  
http://www.plantcell.org/cgi/alerts/ctmain |
| CiteTrack Alerts  | Sign up for CiteTrack Alerts at:  
http://www.plantcell.org/cgi/alerts/ctmain |
| Subscription Information | Subscription Information for *The Plant Cell* and *Plant Physiology* is available at:  
http://www.aspb.org/publications/subscriptions.cfm |