Plant Cell Advance Publication. Published on March 28, 2019, doi:10.1105/tpc.18.00772

LARGE-SCALE BIOLOGY ARTICLE

Metabolome-scale Genome-wide Association Studies Reveal Chemical Diversity and Genetic Control of Maize Specialized Metabolites

Shaoqun Zhou[1,2,6], Karl A. Kremling[3,7], Nonoy Bandillo[3,8], Annett Richter[1], Ying K. Zhang[1,4], Kevin R. Ahern[1,3], Alexander B. Artyukhin[1,9], Joshua X. Hui[1], Gordon C. Younkin[1,2], Frank C. Schroeder[1,4], Edward S. Buckler[3,5], Georg Jander[1,*]

[1]Boyce Thompson Institute, 533 Tower Road, Ithaca, NY 14853
[2]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY14853
[3]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY14853
[4]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY14853
[5]United States Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, New York 14853
[6]Present address: Elo Life Systems, 5 Laboratory Drive, Research Triangle Park, NC 27709
[7]Present address: Inari Agriculture, Cambridge, MA 02139
[8]Present address: Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68583
[9]Present address: EAG Laboratories, Columbia, MO 65202
*Corresponding Author: gj32@cornell.edu

Short title: Maize metabolomic GWAS

One-sentence summary: Metabolite profiling combined with genome-wide association studies provides a resource for structural and functional assignments of the many unknown metabolites in maize seedlings.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Georg Jander (gj32@cornell.edu).

ABSTRACT

Cultivated maize (*Zea mays*) has retained much of the genetic diversity of its wild ancestors. Here, we performed non-targeted high-performance 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, non-stiff-stalk, tropical, sweet corn, 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 metabolically 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 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 more than 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 frugipeda*) 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., 2017). Similarly, different maize inbred lines possess distinct, tissue-specific profiles of specialized metabolites (Meihls et al., 2013; Wen et al., 2014; Handrick et al., 2016; Wen et al., 2016). Therefore, combining high-throughput metabolite profiling, existing genetic resources, and genotypic data for a metabolome-scale genome-wide association studies (GWAS) will allow large-scale identification of candidate genes and loci involved in maize specialized metabolism, opening up the possibility of harnessing the natural biochemical defenses found in the broader maize germplasm for improved pest and disease resistance.

Pioneering GWAS studies conducted with *Arabidopsis thaliana* showed that the plant metabolome has a complex genetic architecture. Genetic mapping of 327 metabolites that were detected by non-targeted metabolomics identified "hotspots" in the Arabidopsis genome that regulate the abundance of multiple metabolites (Chan et al., 2010). The regulation of glucosinolates, a predominant class of Arabidopsis defensive metabolites, showed strong growth stage-specific effects in an analysis of 96 Arabidopsis ecotypes (Chan et al., 2011). Similarly, studies with both rice (*Oryza sativa*) seedling shoots and maize kernels have led to the genome-wide identification of metabolic quantitative trait loci (QTL; Wen et al., 2014; Matsuda et al., 2015; Wen et al., 2016).

In this study, we performed liquid chromatography-mass spectrometry (LC-MS) analysis of the tips and bases of the emerging third leaves of maize inbred lines from the Goodman Diversity Panel (Flint-Garcia et al., 2005). These two tissue types were chosen because 1) they represent distinct stages of differentiation, and 2) constitutive concentrations of specialized metabolites tend to decrease as plants age (Cambier et al., 2000; Zheng et al., 2015). The Goodman Diversity Panel contains 282 maize inbred lines belonging to five genetic subpopulations and has been genotyped with over 29 million single nucleotide polymorphism (SNP) markers (Bukowski et al., 2018). More recently, this genetic mapping panel was analyzed by whole transcriptome profiling of eight distinct tissue-environment combinations (Kremling et al., 2018), including the two tissue types used in the current study. Through metabolomic GWAS of maize seedling leaves, we not only provide important insights into the nature of the maize metabolome, but we also developed a public resource that can be used to associate both known metabolites and previously unidentified LC-MS mass features with specific regulatory and biosynthetic loci in the maize genome.

RESULTS

Comparisons of maize seedling leaf specialized metabolomes between tissue types and genetic sub-populations

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 UPLC/high-resolution-MS analysis of 50% methanol extracts, which measures a wide range of mid-polarity 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 (doi.org/10.25739/9dsj-kw33).

After filtering, more than 7000 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 over 30% of the observed variance (Figure 1A). Two-way analyses of variance (ANOVA) on the same data set showed that more than 97% of all the mass features analyzed were significantly influenced by tissue type (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 (Figure 1C,D). 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 ultraviolet (UV) absorption chromatogram, neighboring peaks tended to have similar UV absorbance profiles. Specifically, peaks eluting between 240 and 360 s had UV absorbance profiles resembling phenylpropanoids, peaks eluting between 360 and 460 s had typical benzoxazinoid-like UV absorbance profiles, and those eluting after 460 s were flavonoid-like (Figure 2A,B). Measurement of a narrower window of UV 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 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 460-570 s time interval containing flavonoids showed the strongest differentiation between tissue types (top panel, Figure 2D). 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 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 absorbance pattern, tandem 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 tandem 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 seconds, all of the flavonoids eluted after 450 seconds, and the most abundant benzoxazinoids eluted between 300 and 360 seconds, with two low-concentration compounds eluting at an earlier retention time (210 and 246 seconds). 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, DIMBOA-Glc (2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one-β-D-glucopyranose) and its methylated glucoside derivative, HDMBOA-Glc (2-(2-hydroxy-4,7-dimethoxy-1,4-benzoxazin-3-one)-β-D-glucopyranose). 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 co-regulated

The abundance of structurally related metabolites, which often arise from shared metabolic pathways, tends to be co-regulated 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 correlative 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 vs. mean = 58; $p < 0.005$, Student's *t*-test). We plotted the distribution of the retention times of mass features belonging to each correlative network in 10-second 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 correlative networks, at least half of the mass features were located in the top three bins, suggesting that these co-regulated mass features are structurally related. Interestingly, we found that the cumulative frequency of the top three 10-second-bins in correlative 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 towards 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 more than 90% of all the genotypes, more than 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 = (Var$_{\text{total}}$ - Var$_{\text{B73}}$)/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 less than 0.2, suggesting a relatively small component of genetic variation. The overall bi-modal 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 non-zero 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 less than ten-fold lower in intensity than a ubiquitous mass feature. By contrast, mass features of any given occurrence rate showed a hundred-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). Prior to 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 ($H^2 \geq 0.2$), 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 less than 10% of the tested genotypes, *i.e*. present in less than ~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 7–10), and GWAS was performed for each metabolite using 29 million SNPs (Bukowski et al., 2018; data are available at Cyverse Discovery Environment, doi.org/10.25739/9dsj-kw33).

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 ten 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 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 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; Chan et al., 2011).

Structurally related metabolites tend to be co-regulated

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 (Figure 8A,C). In both leaf bases and leaf tips, three loci on chromosomes 1, 4, and 10, respectively, showed a large number of metabolite GWAS hits (Figure 8A,C). 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. Since 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 (Figure 8B,D). 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 (Figure 8A,C), there was co-localization of dips in retention time variance with the genomic QTL hotspots (Figure 8B,D), indicating that the abundance of structurally related metabolites tends to be co-regulated 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.

Genome-wide association studies 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$_2$BOA-Glc) and HDMBOA-Glc. GWAS with both metabolites confirmed known QTL containing biosynthetic genes: *Bx13* on chromosome 2 for DIM$_2$BOA-Glc (Handrick et al., 2016; Figure 9A) and *Bx10-12* on chromosome 1 for HDMBOA-Glc (Meihls et al., 2013; Figure 9B), 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 8). 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 (Figure 9A,B).

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 bi-allelic 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 non-zero 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 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 less than 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 less than 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 less than 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 over-represented 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 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 (Figure 11D,E). 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 x B73 and CML247 x 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 RILs families segregated near-isogenic lines that were either heterozygous or homozygous for one or the other parental allele (Figure 12A,B). MS assays showed perfect co-segregation 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 (Figure 12C-E). Furthermore, heterozygous lines showed intermediate phenotypes, producing both isomers, but in lesser abundance than either homozygote.

DISCUSSION

Technological advances in mass spectrometry 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 plants and economically relevant crop species (Chan et al., 2010; Chan et al., 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 non-targeted metabolomics data sets represent unknown metabolites, and mass spectrometry data provide incomplete information about their structures. Our metabolome-scale correlation network analyses (Figure 5) and genome-wide association studies (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 co-elution tests with known compounds and the construction of molecular networks based on shared tandem mass spectrometry (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. Since all loci are bi-allelic (non-biallelic ones were filtered out), each allele was sampled an average of more than 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 the current 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, by using the same maize growth conditions as Kremling et al., 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 bi-modal distributions, with a relatively small core component and a large number of rare mass features (Figure 6A). Compared to the presence/absence distribution of gene expression in the maize pan-transcriptome (Hirsch et al., 2014), the profiles of our metabolomic data are much more left-skewed, $i.e.$ the majority of maize metabolites are present in less than 50% of the inbred lines. This is perhaps reflective of the more commonly non-essential 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-seq-based transcriptomics compared to metabolomics, which would allow detection of rare transcripts in a larger number of maize inbred lines.

Broad sense heritability of metabolite content, estimated as

$$\left[H^2 = \frac{\mathrm{Variance(total)} - \mathrm{Variance(B73)}}{\mathrm{Variance(total)}}\right]$$

, 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 approximately 20 B73 samples were used to calculate environmental variance, total population variance was calculated based on more than 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 sub-populations (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 sub-populations (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 sub-populations.

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 non-uniform 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 absorbance profiles of metabolites in the QTL hotspots indicate that structurally related metabolites tend to be co-regulated 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 benzoxazinoid compounds. However, several mass features that were not associated with known benzoxazinoid compounds also mapped to this locus, suggesting the regulation of other classes of maize specialized metabolites. Such regulation could be indirect, as benzoxazinoid 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 the current 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 our GWAS, may have 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 B73. The abundance of transposon genes in this region in the B73 genome suggests that there may be presence/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 C2H2 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 lined from the Goodman Diversity Panel to map metabolite quantitative traits to the single-gene or near single gene level (Figures 9, 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-seq) 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 or our raw MS data files and genetic mapping data in the Cyverse Discovery Environment (doi.org/10.25739/9dsj-kw33) will enable future analyses beyond what we have done for the current project.

Our metabolomic assays, which were focused on mid-polarity 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 (Champaign-Urbana, IL, USA). 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 (Kremling et al., 2018). Eight seeds of each maize genotype were planted in approximately 50 cm^3 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 micro-environmental 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, two centimeters 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 nitrogen was less than three minutes. To minimize the effects of diurnal variation in the maize metabolome, all samples were harvested in a two-hour 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% methanol acidified with 0.1% formic acid, and analyzed on a Sigma Supelco reverse phase C18 column on a Dionex 3000 Ultimate UPLC-diode array detector system coupled to a Thermo Q Exactive mass spectrometer. The two mobile phase solvents were water (Solvent A) and acetonitrile (Solvent B), both acidified with 0.1% formic acid. The mobile phase gradient ran from 95% Solvent A at 0 minutes to 100% Solvent B at 10.5 minutes with 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 mass spectrometry 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 three 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 (60-630 seconds) 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 prior to log-transformation. This data set was uploaded to the MetaboAnalyst 3.0 online tool platform for principal component analysis 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 failed run under the other mode. These missing data were replaced by zeros 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 carried out in R and Microsoft Excel.

MS2 analysis

B73 seedlings were grown under comparable growth conditions described above, and the third leaves from three seedlings were independently extracted and subjected to the same LC inlet method as described above. In addition to the full MS scan in the original method under either positive or negative mode of electron spray ionization (collision energy = 20V), another MS scan focused on the top 5 features from each previous scan was included to provide extra fragmentation. MS2 spectra from the results were extracted and the major fragment peaks were manually identified. Each MS2 spectrum was queried to the ReSpect for Phytochemical 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.

Benzoxazinoid identification

Benzoxazinoids were identified based on known masses, purified standards, known benzoxazinoid 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 benzoxazinoids (Meihls et al., 2013; Mijares et al., 2013; Betsiashvili et al., 2015; Tzin et al., 2015b; Tzin et al., 2015a; Tzin et al., 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% methanol acidified with 0.1% 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 a ZORBAX Eclipse XDB C18 column on an Agilent 1100 HPLC system (Agilent, Santa Clara, CA). Purified compounds were dried, weighed, and re-dissolved in pure methanol. NMR spectroscopy analyses were carried out on a Unity INOVA 600 instrument (Varian Medical Systems, Palo Alto, CA) with the following conditions: 256 scans for ^{1}H NMR; nt = 16 and ni = 800 for COSY and nt = 32 and ni > 800 for HSQC and HMBC.

Genotyping of maize recombinant inbred lines

B73 x Ki11 and B73 x CML247 RILs from a maize nested association mapping population (McMullen et al., 2009) were identified based on their residual heterozygosity in the area of the GRMZM2G063909 gene (based on genotyping data at www.panzea.org). DNA was extracted from three-week-old seedlings using KAZU DNA extraction buffer according to the manufacturer's instructions (www.kerafast.com). A polymorphic marker near the GRMZM2G063909 gene was amplified using the primers F: GACAGGGAAGGTATATGC and R: GATAGAGCATCAACTTGATC in a reaction containing 7.5 µl GoTaq MasterMix, 2.5 µl of each primer, and 2.5 µl maize genomic DNA. The PCR conditions were 95 °C 3 min, 34 cycles

(95 °C 30 sec, 60 °C 30 sec, 72 °C 30 sec), and 5 min 72 °C. Amplified products were separated on 1% agarose gels and detected by ethidium bromide staining.

Correlative network analyses

The metabolomic data sets were used to calculate pairwise Pearson correlation matrices and then mutual rank matrices for the two tissue types separately. Pairwise mutual rank indices were converted to edge weights by an exponential decay functions, with $\lambda = 50$ as previously described (Wisecaver et al., 2017). For each conversion, edges with weight lower than 0.01 were filtered out. These edge lists were imported into Cytoscape v 3.4.0 (Shannon et al., 2003), and overlapping clusters were detected with the ClusterONE app (Nepusz et al., 2012).

Genome-wide association study with metabolic traits

The signal intensity of each mass feature across the population was log-transformed. Box-cox transformation was skipped, as it distorted the distribution of the rare mass features with many zero values. Mass features were filtered based on estimated broad sense heritability and rate of occurrence as described in the Results section, and the remaining 3,991 mass features were analyzed with the fast GWAS pipeline (Kremling et al., 2018). To reduce data storage to a realistic level, only SNPs with $-\log(p) \geq 5$ for each mass feature were recorded. The top 50 most significantly associated SNP markers of each mass feature from leaf tips (Supplemental Data Set 12) and leaf bases (Supplemental Data Set 13) were extracted for easier reference.

To survey the genetic architectures of metabolic traits and investigate their relationship with trait heritability and occurrence rate at a metabolomic scale, the top 10 most strongly associated SNP markers for each metabolic trait to three different fixed size LD blocks, namely 10 kb, 60 kb, and 360 kb were mapped. As expected, more metabolic traits have their top GWAS SNP hits located within fewer number of LD blocks as the estimated LD size increases, but the overall shape of distribution was not affected (Supplemental Figure 3). The same LD-block assigning process was used to generate the overview of GWAS hits distribution across the maize genome, by counting the numbers of mass features mapped to each LD block, and plotting them according to the physical location of the LD blocks in the maize genome. Similarly, the locations of metabolite QTL genomic hotspots are consistent across different window sizes of LD (Supplemental Figure 4). Finally, GWAS hits were ordered based on their physical location in the maize genome, and the log variance of mass feature retention time of 100 adjacent hits was calculated using a sliding window algorithm.

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 under the following directory (iplant/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). Bi-allelic 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 use 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_024460.2 (230953451..230957085)/GRMZM2G151227; NC_024459.2 (233440789..233442623)/GRMZM2G380650; NC_024459.2 (67312119..67314042)/GRMZM2G311036; NC_024459.2 (67228930..67231301)/GRMZM2G336824; NC_024459.2 (67109158..67116836)/GRMZM2G023325; NC_024460.2 (238143556..238145183)/AC148152.3_FG005; NC_024467.2 (128358117..128362499)/GRMZM2G108309; NC_024462.2

(238977913..238988971)/GRMZM2G063909; NC_024461.2 (5057135..5060050)/GRMZM2G143723.

Raw LC-MS result files, LC-MS2 results files, full FastGlm GWAS results files for all filtered mass features, and raw 1-D proton, 2-D COSY, HSQC, HMBC 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 mass spectrometry in seedling leaves of diverse maize inbred lines.

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

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

Supplemental Data Set 4. Major fragments from MS2 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 mass spectrometry in seedling leaves tips of diverse maize inbred lines.

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

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

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

Supplemental Data Set 11. 2D-NMR data of maize phenylpropanoid hydroxycitric 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 mass spectrometry raw files with the XCMS-CAMERA pipeline.

ACKNOWLEDGMENTS

This research was funded by US National Science Foundation awards 1139329 and 1339237 to G.J., and US National Science Foundation Graduate Research Fellowship Program award DGE-1650441 to K.A.K. N.B. was funded by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DEAR0000598. F.C.S. is a Faculty Scholar of the Howard Hughes Medical Institute. E.S.B. is funded by the United States Department of Agriculture – Agricultural Research Service. This work made use of the Cornell University NMR Facility, which is partially supported by Major Research Instrumentation award CHE-1531632 from the US National Science Foundation. 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.A., J.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 manuscript.

REFERENCES

Ahmad, S., Veyrat, N., Gordon-Weeks, R., Zhang, Y., Martin, J., Smart, L., Glauser, G., Erb, M., Flors, V., Frey, M., and Ton, J. (2011). Benzoxazinoid metabolites regulate innate immunity against aphids and fungi in maize. Plant Physiol **157,** 317-327.

Benton, H.P., Want, E.J., and Ebbels, T.M.D. (2010). Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data. Bioinformatics **26,** 2488-2489.

Betsiashvili, M., Ahern, K.R., and Jander, G. (2015). Additive effects of two quantitative trait loci that confer *Rhopalosiphum maidis* (corn leaf aphid) resistance in maize inbred line Mo17. J Exp Bot **66,** 571-578.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23,** 2633-2635.

Buckler, E.S., Gaut, B.S., and McMullen, M.D. (2006). Molecular and functional diversity of maize. Curr Opin Plant Biol **9,** 172-176.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C., Fan, L., Gao, S., Xu, X., Zhang, G., Li, Y., Jiao, Y., Doebley, J.F., Ross-Ibarra, J., Lorant, A., Buffalo, V., Romay, M.C., Buckler, E.S., Ware, D., Lai, J., Sun, Q., and Xu, Y. (2018). Construction of the third-generation Zea mays haplotype map. Gigascience **7,** 1-12.

Cambier, V., Hance, T., and de Hoffmann, E. (2000). Variation of DIMBOA and related compounds content in relation to the age and plant organ in maize. Phytochemistry **53,** 223-229.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., and Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol **30,** 918-920.

Chan, E.K., Rowe, H.C., Hansen, B.G., and Kliebenstein, D.J. (2010). The complex genetic architecture of the metabolome. PLoS Genet **6,** e1001198.

Chan, E.K., Rowe, H.C., Corwin, J.A., Joseph, B., and Kliebenstein, D.J. (2011). Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. PLoS Biol **9,** e1001125.

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., Zhang, W., Zhang, L., Yu, S., Wang, G., Lian, X., and Luo, J. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet **46,** 714-721.

Flint-Garcia, S.A., Thuillet, A.C., Yu, J., Pressoir, G., Romero, S.M., Mitchell, S.E., Doebley, J., Kresovich, S., Goodman, M.M., and Buckler, E.S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J **44,** 1054-1064.

Halkier, B.A., and Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. Annu Rev Plant Biol **57,** 303-333.

Handrick, V., Robert, C.A., Ahern, K., Zhou, S., Machado, R.A., Maag, D., Glauser, G., Fernandez-Penny, F.E., Chandran, J.N., Rodgers-Melnick, E., Schneider, B., Buckler, E.S., Boland, W., Gershenzon, J., Jander, G., Erb, M., and Köllner, T.G. (2016). Biosynthesis of 8-*O*-methylated benzoxazinoid defense compounds in maize. Plant Cell **28,** 1682-16700.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Penagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., de Leon, N., Kaeppler, S.M., and Buell, C.R. (2014). Insights into the maize pan-genome and pan-transcriptome. Plant Cell **26,** 121-135.

Jahne, A., Fritzen, C., and Weissenbock, G. (1993). Chalcone synthase and flavonoid products in primary-leaf tissues of rye and maize. Planta **189,** 39-46.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K.L., Wolfgruber, T.K., May, M.R., Springer, N.M., Antoniou, E., McCombie, W.R., Presting, G.G., McMullen, M., Ross-Ibarra, J., Dawe, R.K., Hastie, A., Rank, D.R., and Ware, D. (2017). Improved maize reference genome with single-molecule technologies. Nature **546,** 524-527.

Kremling, K.A.G., Chen, S.Y., Su, M.H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J., and Buckler, E.S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature **555,** 520-523.

Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T.R., and Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Analytical Chemistry **84,** 283-289.

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, J.I., Ebana, K., Yano, M., and Saito, K. (2015). Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. Plant J **81,** 13-23.

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S.E., Peterson, B., Pressoir, G., Romero, S., Oropeza Rosas, M., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith,

S., Glaubitz, J.C., Goodman, M., Ware, D., Holland, J.B., and Buckler, E.S. (2009). Genetic properties of the maize nested association mapping population. Science **325,** 737-740.

Meihls, L.N., Handrick, V., Glauser, G., Barbier, H., Kaur, H., Haribal, M.M., Lipka, A.E., Gershenzon, J., Buckler, E.S., Erb, M., Köllner, T.G., and Jander, G. (2013). Natural variation in maize aphid resistance is associated with 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one glucoside methyltransferase activity. Plant Cell **25,** 2341-2355.

Mijares, V., Meihls, L., Jander, G., and Tzin, V. (2013). Near-isogenic lines for measuring phenotypic effects of DIMBOA-Glc methyltransferase activity in maize. Plant Signal Behav.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods **9,** 471-472.

Nguyen, D.D., Wu, C.H., Moree, W.J., Lamsa, A., Medema, M.H., Zhao, X., Gavilan, R.G., Aparicio, M., Atencio, L., Jackson, C., Ballesteros, J., Sanchez, J., Watrous, J.D., Phelan, V.V., van de Wiel, C., Kersten, R.D., Mehnaz, S., De Mot, R., Shank, E.A., Charusanti, P., Nagarajan, H., Duggan, B.M., Moore, B.S., Bandeira, N., Palsson, B.O., Pogliano, K., Gutierrez, M., and Dorrestein, P.C. (2013). MS/MS networking guided analysis of molecule and gene cluster families. Proc Natl Acad Sci U S A **110,** E2611-2620.

Ozawa, T., Nishikiori, T., and Takino, Y. (1977). Three new substituted cinnamoyl hydroxycitricacids from corn plant. Ag Biol Chem **42,** 359-367.

Pick, T.R., Brautigam, A., Schluter, U., Denton, A.K., Colmsee, C., Scholz, U., Fahnenstich, H., Pieruschka, R., Rascher, U., Sonnewald, U., and Weber, A.P. (2011). Systems analysis of a maize leaf developmental gradient redefines the current C4 model and provides candidates for regulation. Plant Cell **23,** 4208-4220.

Plenchamp, C. (2013). Direct and indirect effects of the rhizobacteria *Pseudomonas putida* KT2440 on maize plants In Institute of Biology (Neuchâtel University of Neuchâtel), pp. 157.

Ranum, P., Pena-Rosas, J.P., and Garcia-Casal, M.N. (2014). Global maize production, utilization, and consumption. Ann N Y Acad Sci **1312,** 105-112.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A.E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet **44,** 217-220.

Samayoa, L.F., Malvar, R.A., Olukolu, B.A., Holland, J.B., and Butron, A. (2015). Genome-wide association study reveals a set of genes associated with resistance to the Mediterranean corn borer (*Sesamia nonagrioides* L.) in a maize diversity panel. BMC Plant Biol **15,** 35.

Sawada, Y., Nakabayashi, R., Yamada, Y., Suzuki, M., Sato, M., Sakata, A., Akiyama, K., Sakurai, T., Matsuda, F., Aoki, T., Hirai, M.Y., and Saito, K. (2012). RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. Phytochemistry **82,** 38-45.

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

Stokstad, E. (2017). New crop pest takes Africa at lightning speed. Science **356,** 473-474.

Tautenhahn, R., Bottcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. Bmc Bioinformatics **9**.

Tzin, V., Lindsay, P., Christensen, S.A., Meihls, L.N., Blue, L.B., and Jander, G. (2015a). Genetic mapping shows intraspecific variation and transgressive segregation for caterpillar-induced aphid resistance in maize. Mol Ecol **24,** 5739-5750.

Tzin, V., Hojo, Y., Strickler, S.R., Bartsch, L.J., Archer, C.M., Ahern, K.R., Zhou, S., Christensen, S.A., Galis, I., Mueller, L.A., and Jander, G. (2017). Rapid defense responses in maize leaves induced by Spodoptera exigua caterpillar feeding. Journal of Experimental Botany **68,** 4709-4723.

Tzin, V., Fernandez-Pozo, N., Richter, A., Schmelz, E.A., Schoettner, M., Schaefer, M., Ahern, K.R., Meihls, L.N., Kaur, H., Huffaker, A., Mori, N., Degenhardt, J., Mueller, L.A., and Jander, G. (2015b). Dynamic maize responses to aphid feeding are revealed by a time series of transcriptomic and metabolomic assays Plant Physiol **169,** 1727-1743.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., and Yan, J. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun **5,** 3438.

Wen, W., Liu, H., Zhou, Y., Jin, M., Yang, N., Li, D., Luo, J., Xiao, Y., Pan, Q., Tohge, T., Fernie, A.R., and Yan, J. (2016). Combining quantitative genetics approaches with regulatory network analysis to dissect the complex metabolism of the maize kernel. Plant Physiol **170,** 136-146.

Wisecaver, J.H., Borowsky, A.T., Tzin, V., Jander, G., Kliebenstein, D.J., and Rokas, A. (2017). A Global Co-Expression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell **29,** 944-959.

Zheng, L., McMullen, M.D., Bauer, E., Schon, C.C., Gierl, A., and Frey, M. (2015). Prolonged expression of the BX1 signature enzyme is associated with a recombination hotspot in the benzoxazinoid gene cluster in *Zea mays*. J Exp Bot **66,** 3917-3930.

Zhou, S., Richter, A., and Jander, G. (2018). Beyond defense: Multiple functions of benzoxazinoids in maize metabolism. Plant Cell Physiol **59,** 1528-1537.

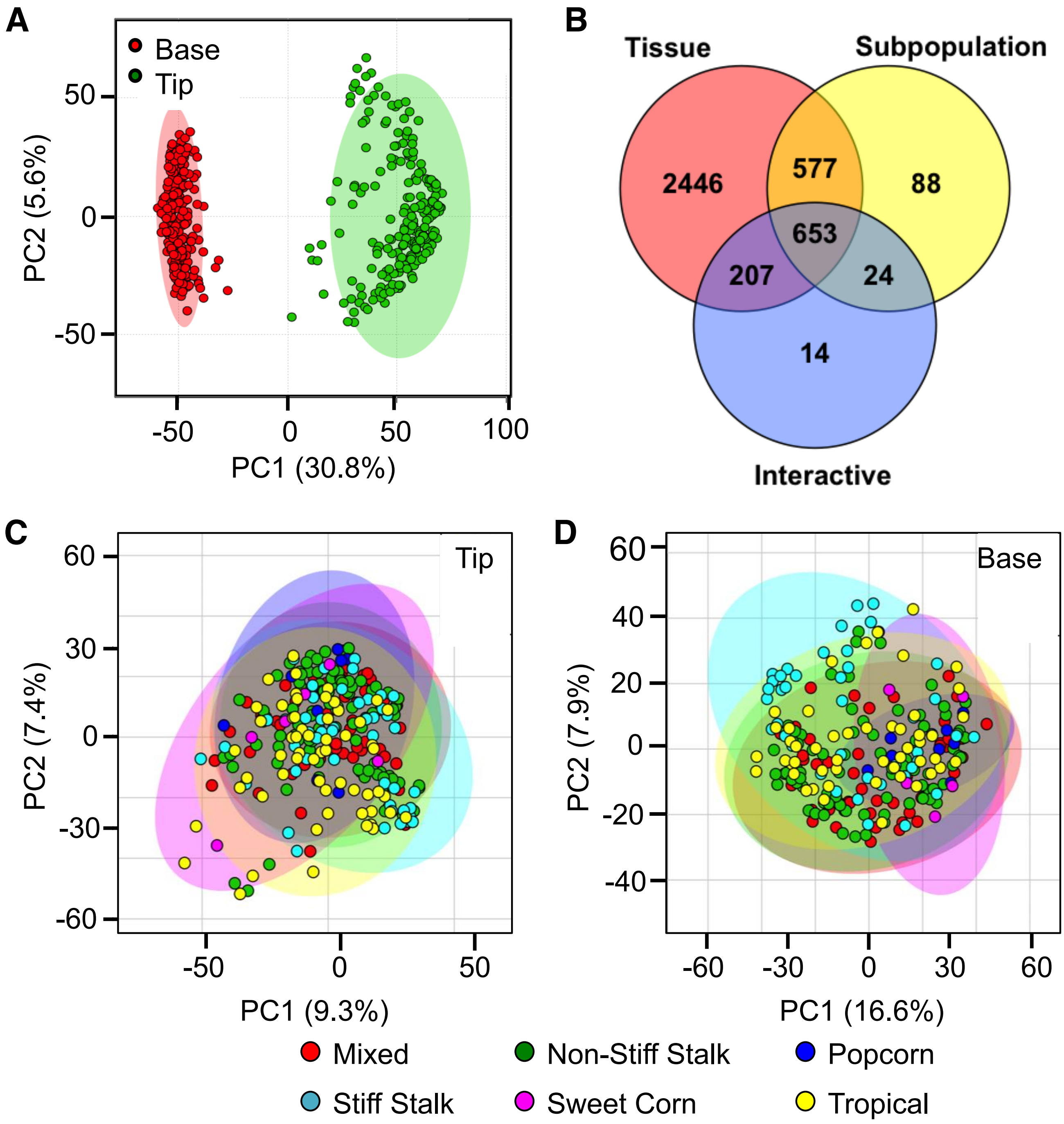


Figure 1. Maize specialized metabolome significantly differentiates leaf tips and bases, but not genetic subpopulations. (A) Principal component analysis differentiates the metabolomes of maize seedling leaf tips and bases. (B) Consistently, more mass features are significantly different by tissue type 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,D) Within either tissue type, genetic subpopulations cannot be differentiated by principal component analysis based on their overall metabolomic fingerprint.

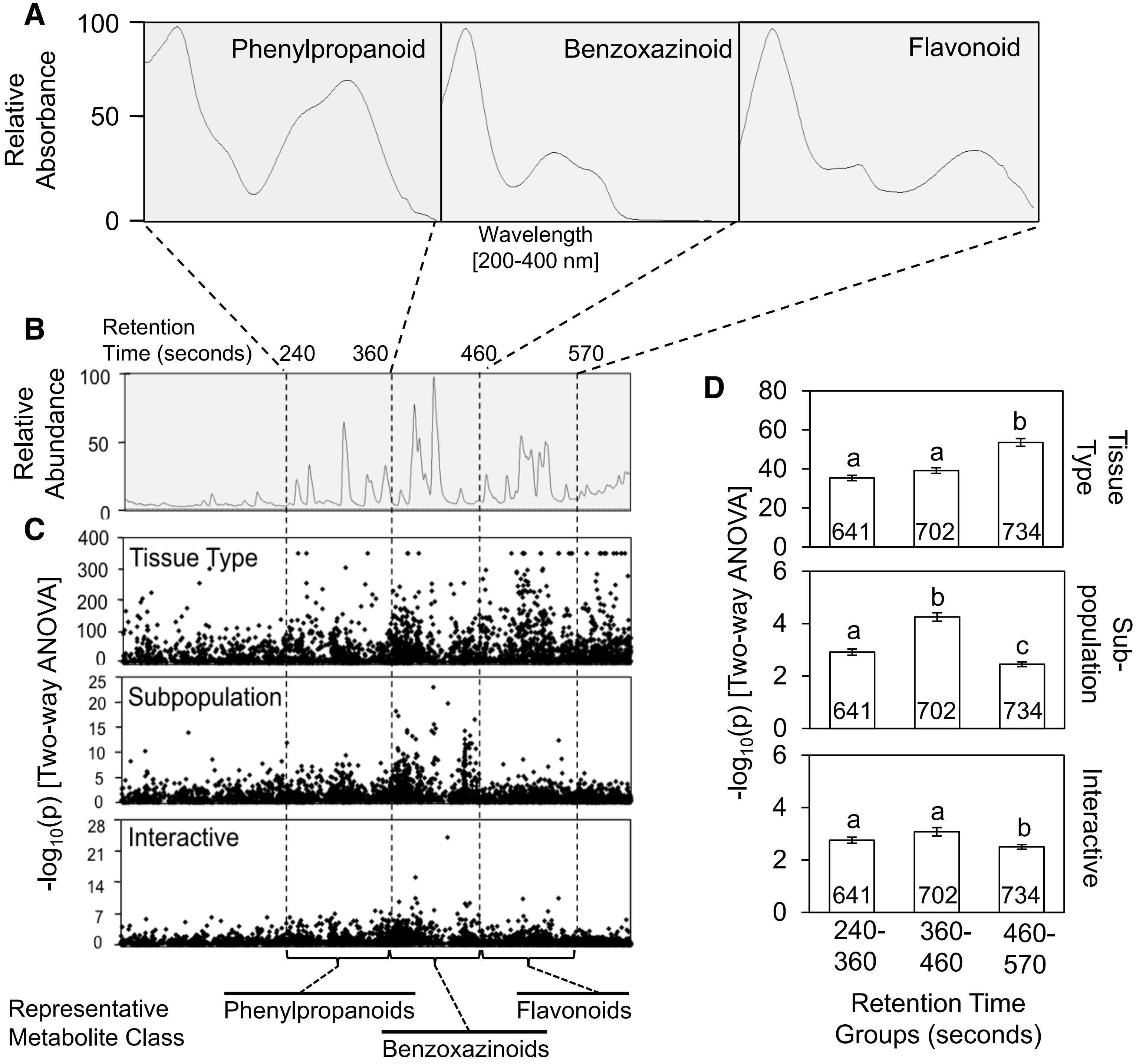


Figure 2. Metabolomic differentiation between tissue types and subpopulations is driven by different classes of specialized metabolites. Mass features with characteristic UV absorbance profiles of phenylpropanoids, benzoxazinoids, and flavonoids (A) are found in distinct segments of the MS chromatogram (B). (C) Each mass feature was plotted based on its retention time (x-axis) and –log(p) by either tissue type, subpopulation, or their interactive effect (y-axes; Two-way ANOVA), and aligned to a sample total ion chromatogram. (D) Average –log(p) from two-way ANOVA by each variable is compared among three retention time ranges, corresponding to three classes of specialized metabolites. Different letters indicate $P < 0.05$, ANOVA followed by Tukey's HSD test. Numbers in the bars indicate sample sizes. Error bars = standard error.

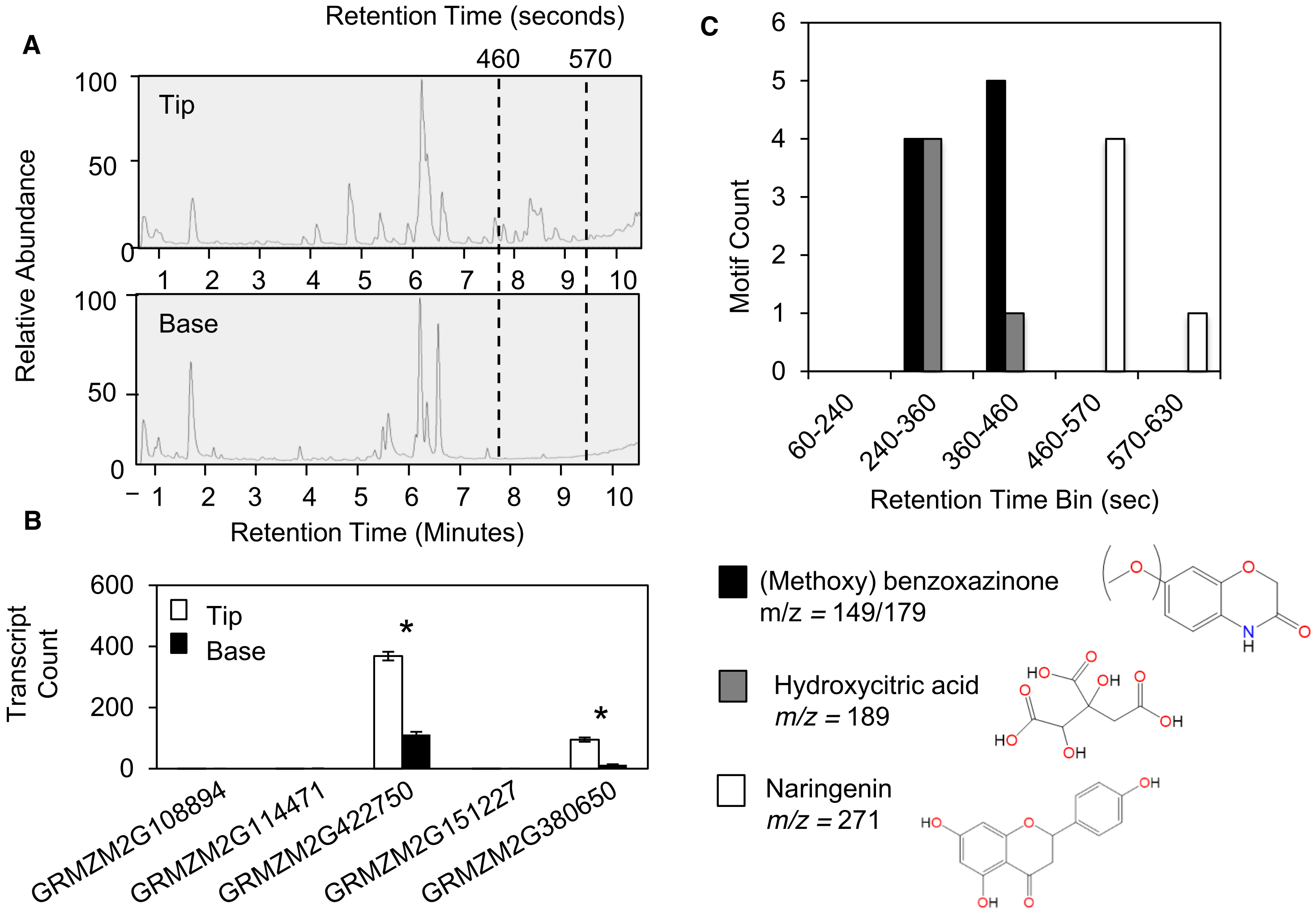


Figure 3. Flavonoids are absent and chalcone synthases expression is low in seedling leaf base tissues. (A) Sample UV absorbance chromatograms of the seedling leaf tip and base of the same genotype show a lack of peaks in the flavonoid time range (460-570 seconds) in the leaf tip and base. (B) Average expression levels of five chalcone synthase-encoding gene models in the B73 reference genome v4 (Annotation 5b+) across the Goodman diversity panel are compared between these two tissue types with Student's t-tests (*FDR < 0.05; N = 300 for each tissue type/gene model combination). Error bars = standard errors. Expression data were obtained from Kremling et al., 2018. (C) Distribution of MSMS fragmentation motifs representative of three structural backbones is clustered in each retention time bin defined by characteristic peaks in the UV absorbance chromatograms.

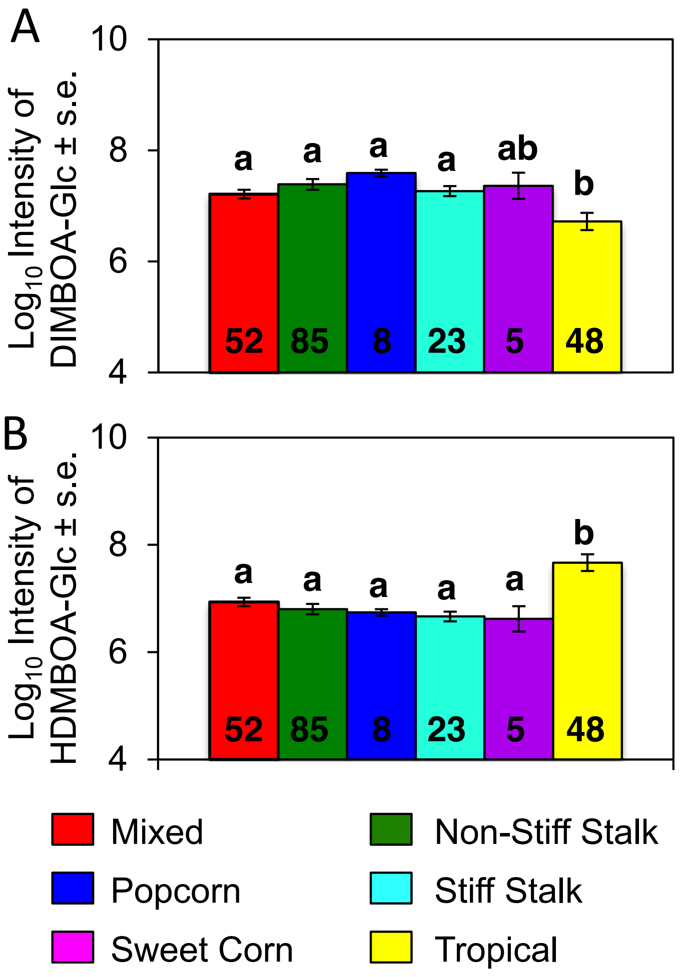


Figure 4. Tropical and temperate maize lines accumulate different benzoxazinoid compounds. Maize inbred lines were assigned to genetic subpopulations defined in Flint-Garcia et al., 2005. (A) DIMBOA-Glc and (B) HDMBOA-Glc in seedling leaf tips were estimated based on their respective representative mass features detected under negative electron spray ionization mode (DIMBOA-Glc [M]-: *mz* = 372.09; HDMBOA-Glc [M+FA]-: *mz* = 432.11). Error bars = standard error. Different letters indicate significant difference ($p < 0.05$), ANOVA followed by Tukey's HSD test. The number of genotypes in each sub-population is indicated at the base of the columns.

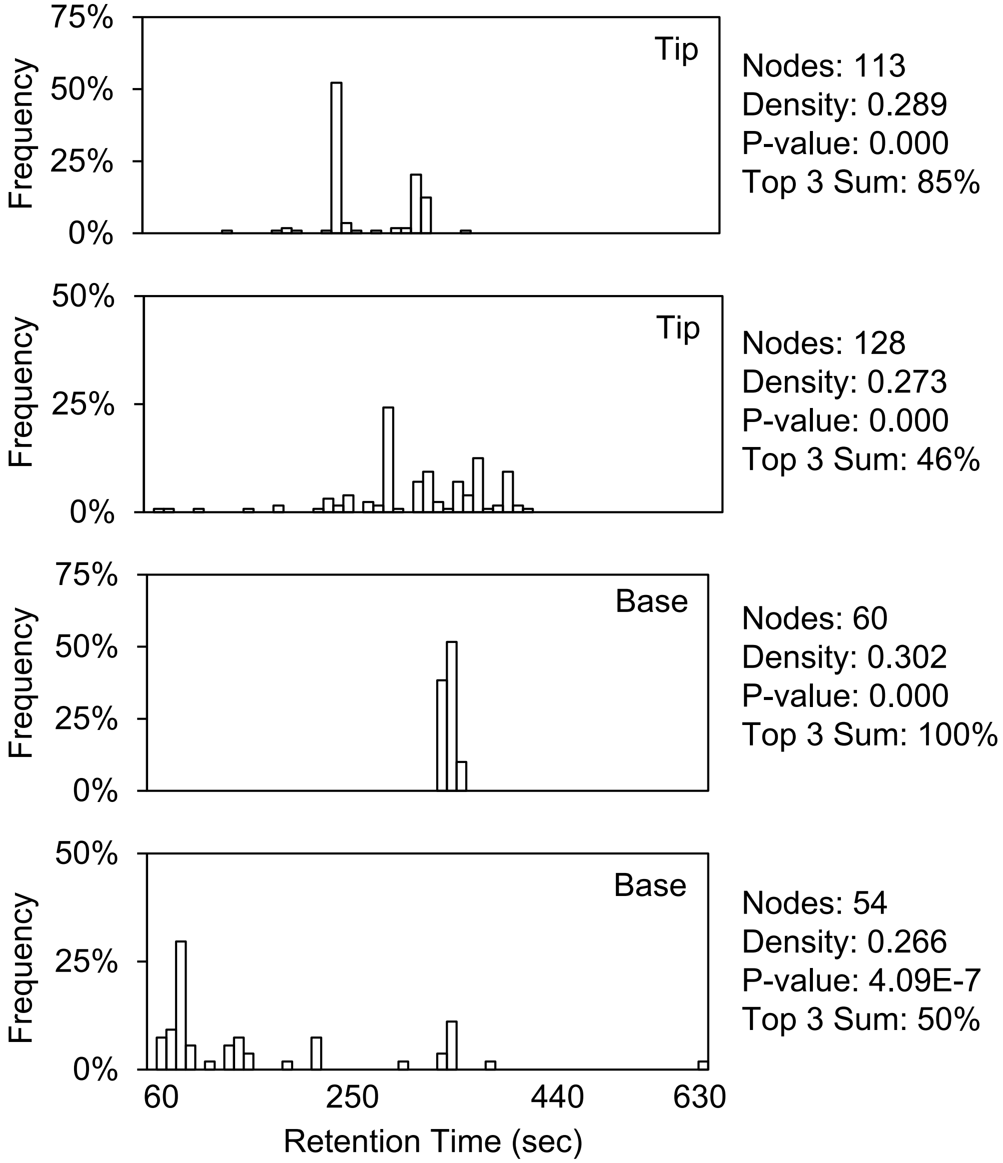


Figure 5. Mass features in the same correlation network tend to have similar retention times. Distributions of retention times of mass features of each correlation network plotted in ten-second increment bins. Density and p-value (one-sided Mann-Whitney *U*-test) of each network were calculated using the graph-clustering algorithm ClusterOne. The top 3 sum is the accumulative percentage frequency of the top 3 ten-second bins, which is used to assess the level of clustering in retention time within each network. Only two significant networks with contrasting level of retention time clustering from either tissue type are shown. All other significant networks ($p < 0.05$) are listed in Supplemental Data Sets 4 and 5.

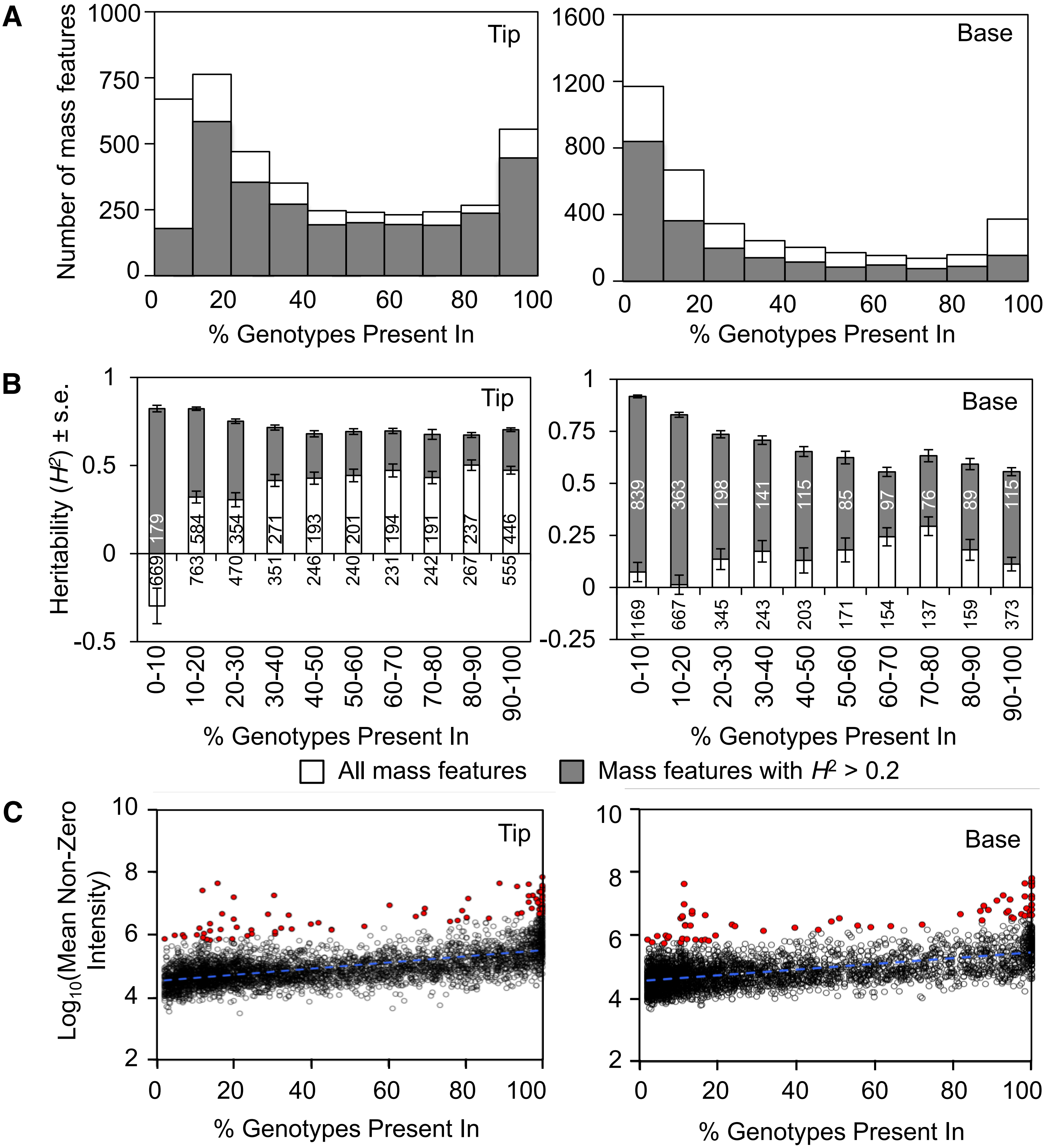


Figure 6. Mass feature occurrence rates are bimodally distributed and are positively correlated with their average non-zero intensity. (A) Distribution of mass features before (white) and after (grey) filtering by broad sense heritability in inbred line B73 ($H^2 > 0.2$) in either tissue type plotted in 10% incremental bins. (B) Average heritability of mass features within each 10% occurrence incremental bins with exclusive lower boundaries and inclusive upper boundaries. The numbers of mass features in each bin before and after filtering by heritability are shown below and above the x-axis in each column, respectively. (C) Each mass feature in either tissue type was plotted based on its occurrence rate (x-axis) and the log of average non-zero 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.

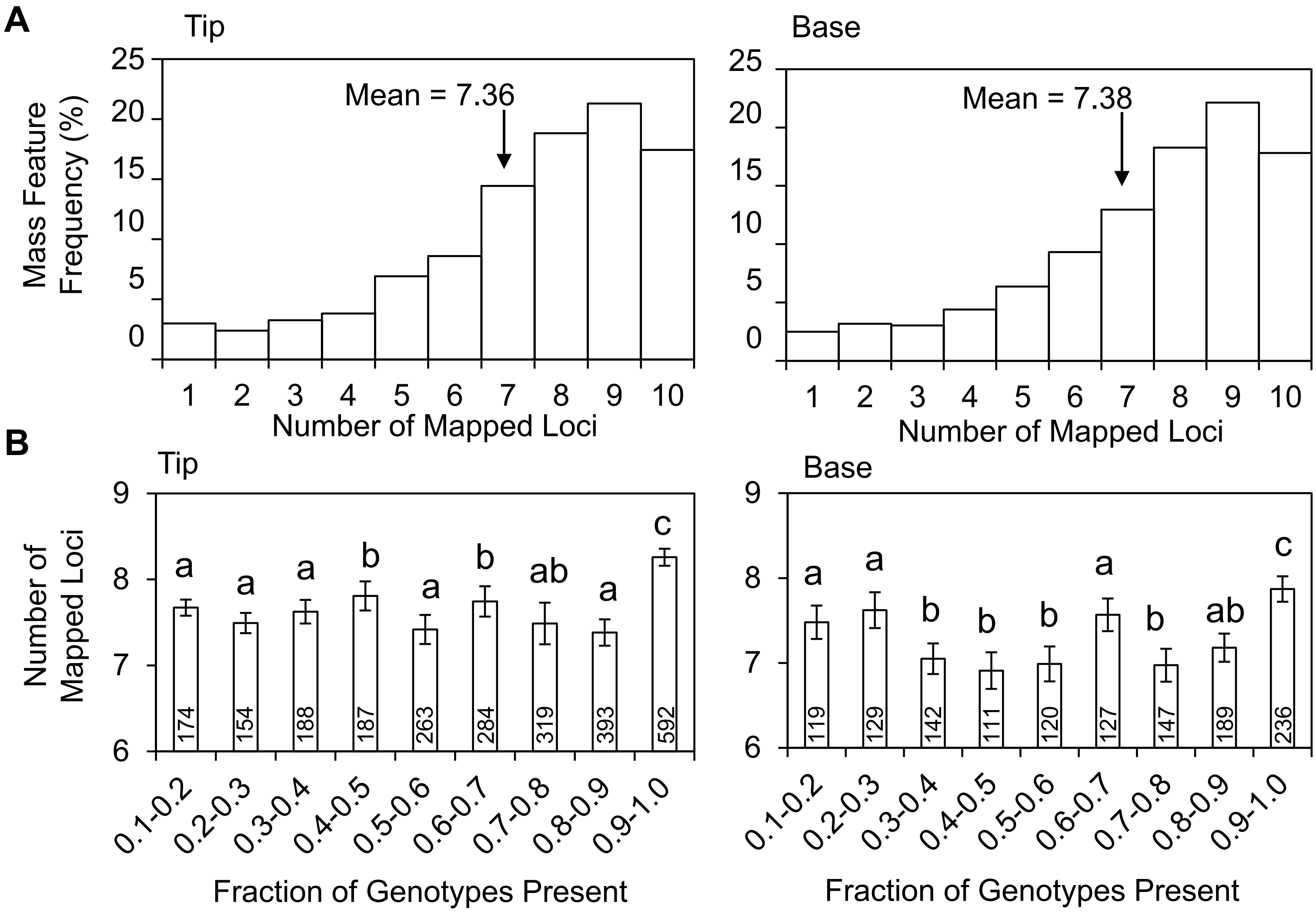


Figure 7. Metabolic traits tend to have complex genetic architecture irrespective of their heritability or occurrence rate. (A) Distribution of mass features in leaf tips and bases plotted based on the number of 10 kb LD blocks that contain one of their top 10 strongest associated SNP markers. Statistical 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 HSD. Groups significantly different from each other ($p < 0.05$) are denoted with different letters above their respective columns. Numbers in the bars indicate number of mass features in each bin. Error bars = standard errors.

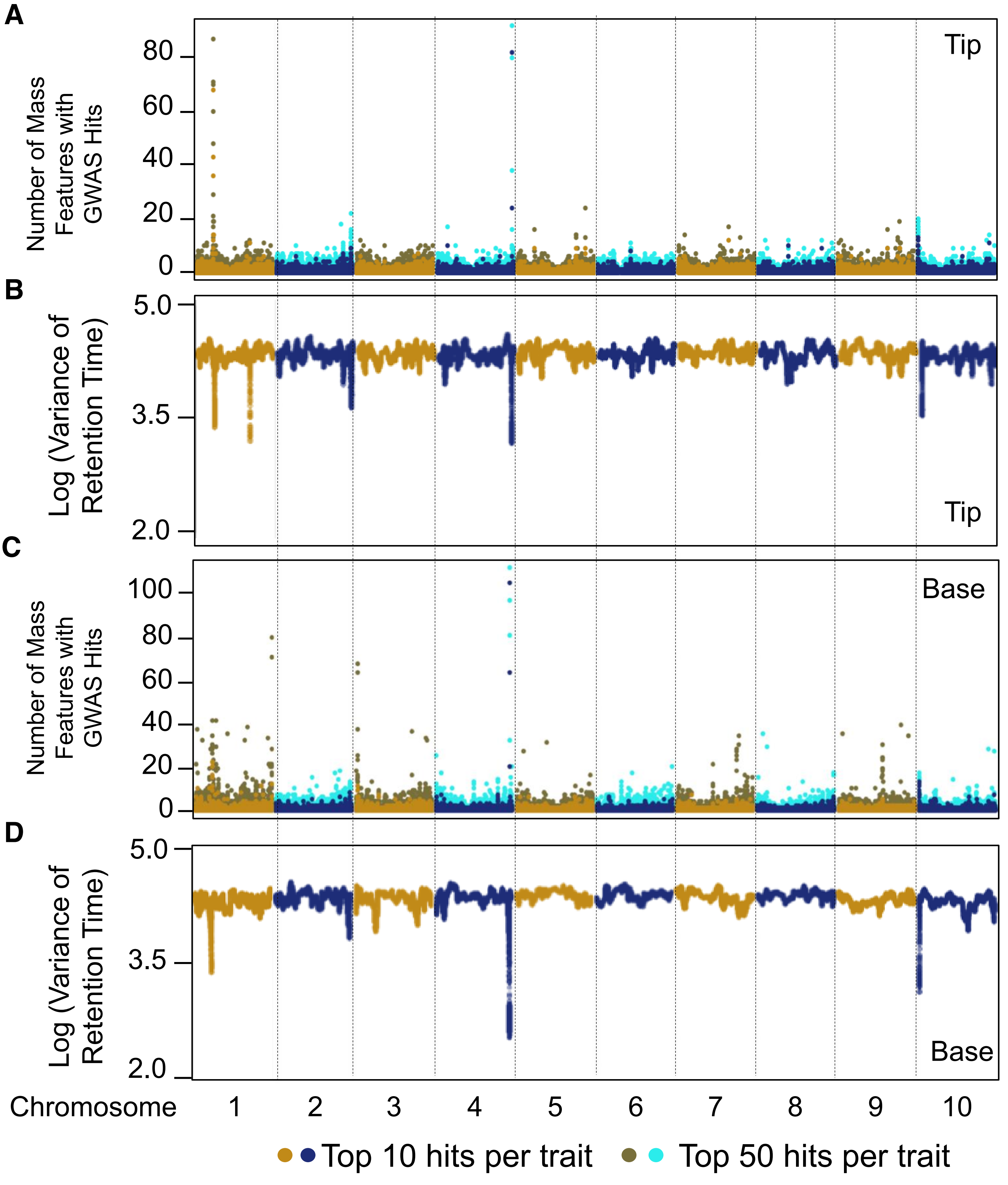


Figure 8. Metabolite GWAS hotspots tend to be associated with mass features that have similar retention times. (A,C) The number of mass features with at least one of their top 10 or top 50 most strongly associated SNP marker located in each 10 kbps block plotted for either tissue type. Results of neighboring chromosomes are shown in different colors, and results based on different top SNP threshold (10 or 50) are indicated by different color shades. (B,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.

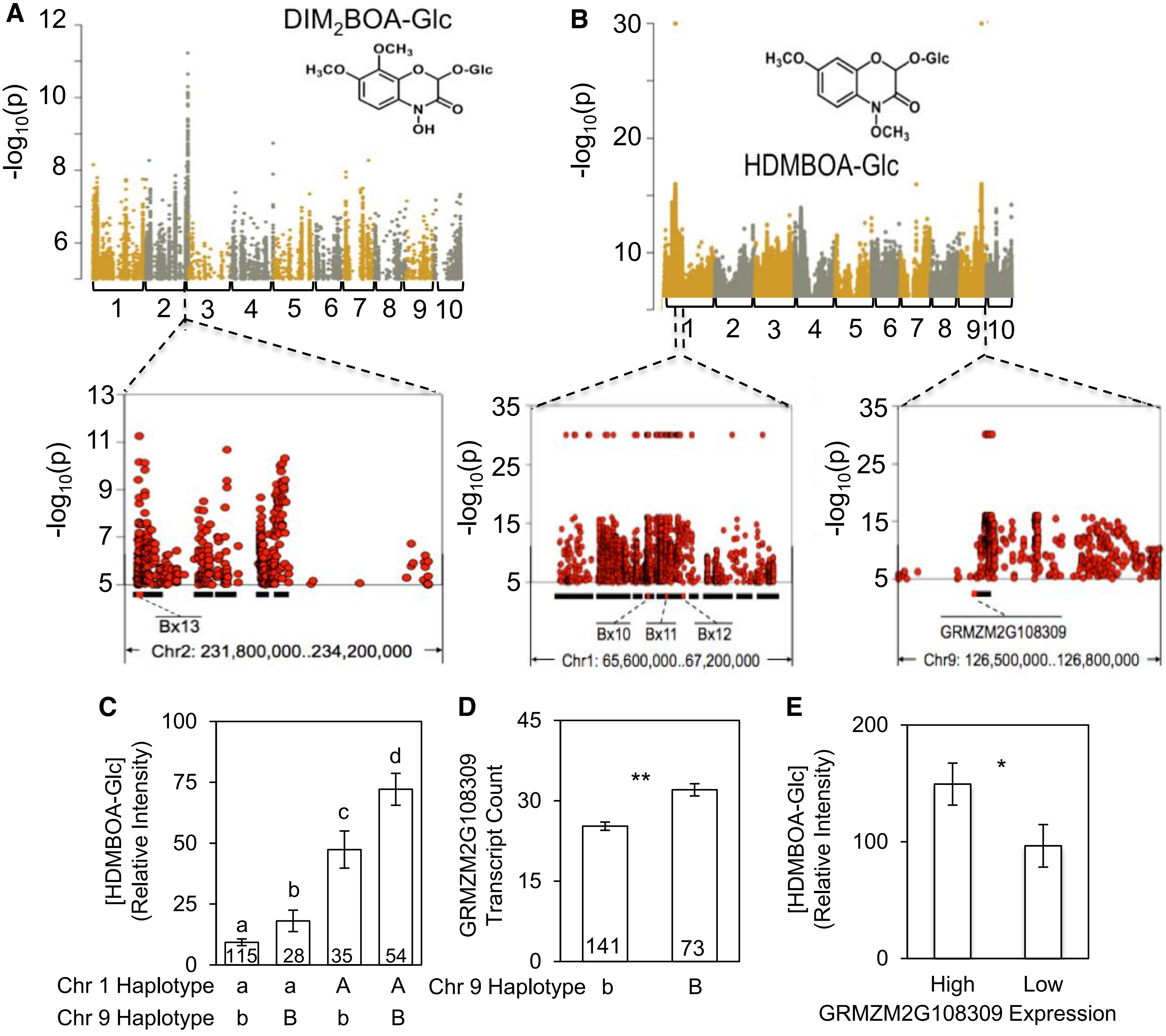


Figure 9. Genome-wide association analysis with HDMBOA-Glc identifies known biosynthetic genes and a previously unknown locus. (A) Natural variation in the abundance of (A) DIM2BOA-Glc and (B) HDMBOA-Glc was mapped by GWAS. Each SNP marker was plotted based on its physical location in the maize genome (x-axis) and level of association with HDMBOA-Glc abundance (y-axes). SNP markers perfectly associated with the phenotype (*i.e.* p = 0) were rounded down to 30 on the y-axes for graphical representation. SNP markers on adjacent chromosomes (labeled at the bottom) are shown in different colors. Only SNP markers with –log10(p) > 5 were plotted. Local linkage disequilibrium blocks around the most highly associated markers, calculated from the same SNP data set, are indicated by black bars at the bottoms of the plots, and known benzoxazinoid biosynthetic genes are highlighted in red. (C) Additive effect on HDMBOA-Glc abundance of the two loci on chromosome 1 and chromosome 9. Mean +/- s.e., different letters indicate significant differences, P < 0.05, ANOVA followed by Tukey's HSD test. Numbers in bars indicate sample sizes (D) Effect of haplotypic segregation on the expression of the candidate gene. Mean +/- s.e., **P < 0.005, two-tailed t-test. Numbers in bars indicate sample sizes. (E) Comparison of HDMBOA-Glc abundance in the 20 inbred lines with the highest and lowest GRMZM2G108309 gene expression levels. Mean +/- s.e., *P < 0.05, two-tailed *t*-test.

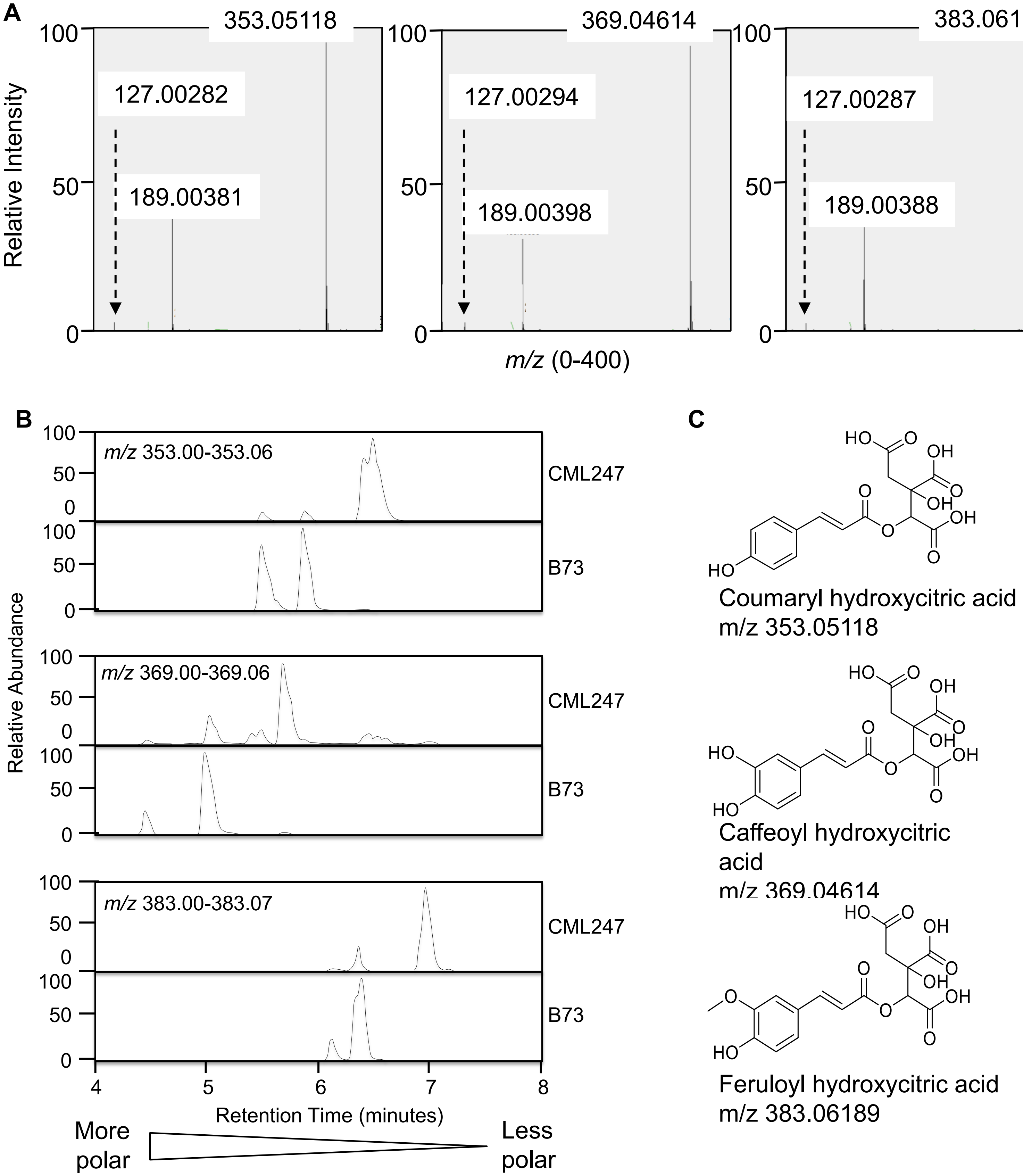
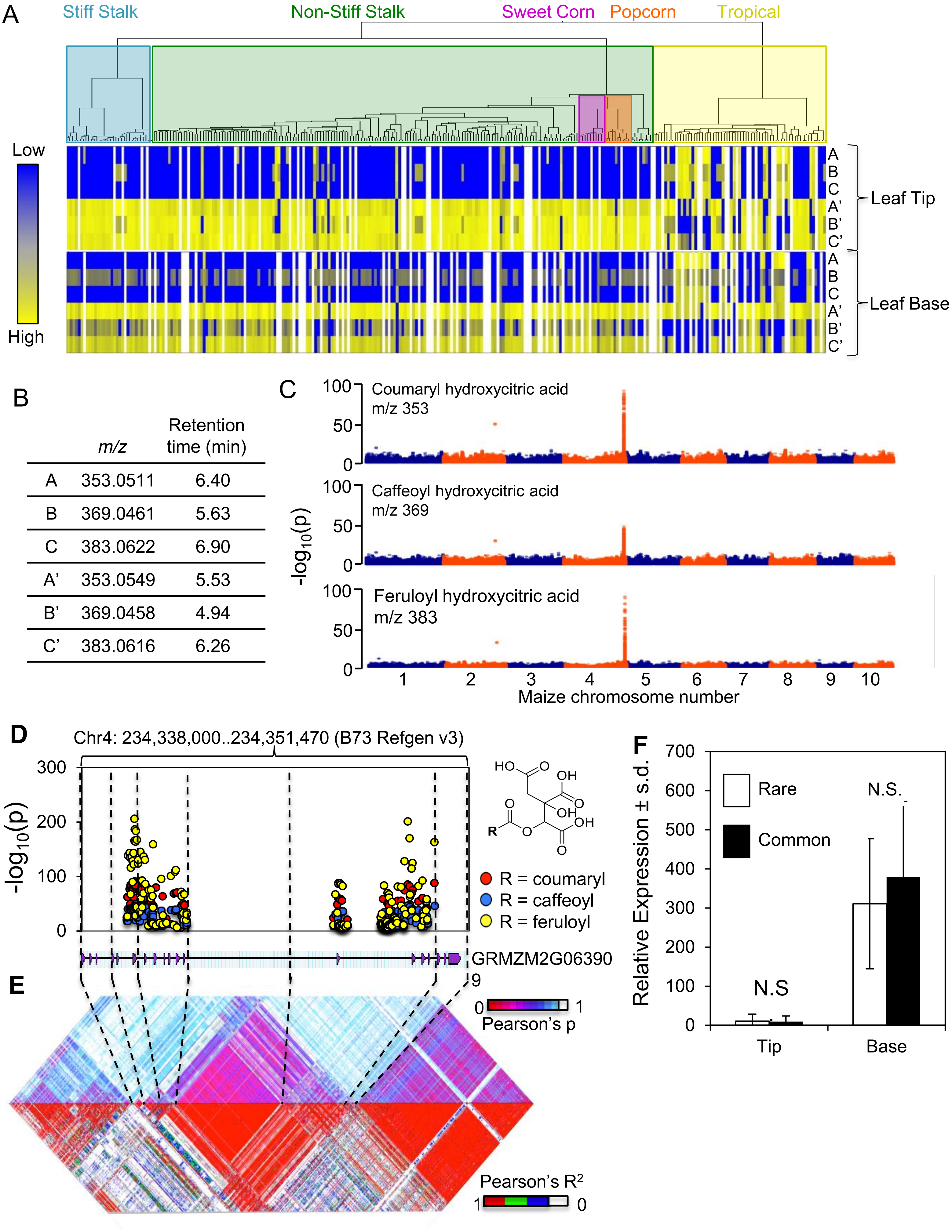


Figure 10. Phenylpropanoid-containing mass features co-elute with common daughter ions. (A) Mass spectrum scans of three mass features co-eluting with phenylpropanoid-like UV absorbance peaks are shown. The parental ions and the two shared daughter ions are labeled with their exact *m/z* measurement. (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.



	m/z	Retention time (min)
A	353.0511	6.40
B	369.0461	5.63
C	383.0622	6.90
A'	353.0549	5.53
B'	369.0458	4.94
C'	383.0616	6.26

Figure 11. Three pairs of hydroxycitric acid conjugates have complementary distribution 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 panel A have different retention times in minutes and were detected in negative ionization mode (*m*/*z*). (C) GWAS identified a common locus on chromosome 4 that regulates the abundance of all of the identified mass features from panel 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 linkage disequilibrium 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 the rare (N = 72) and common (N = 207) 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).

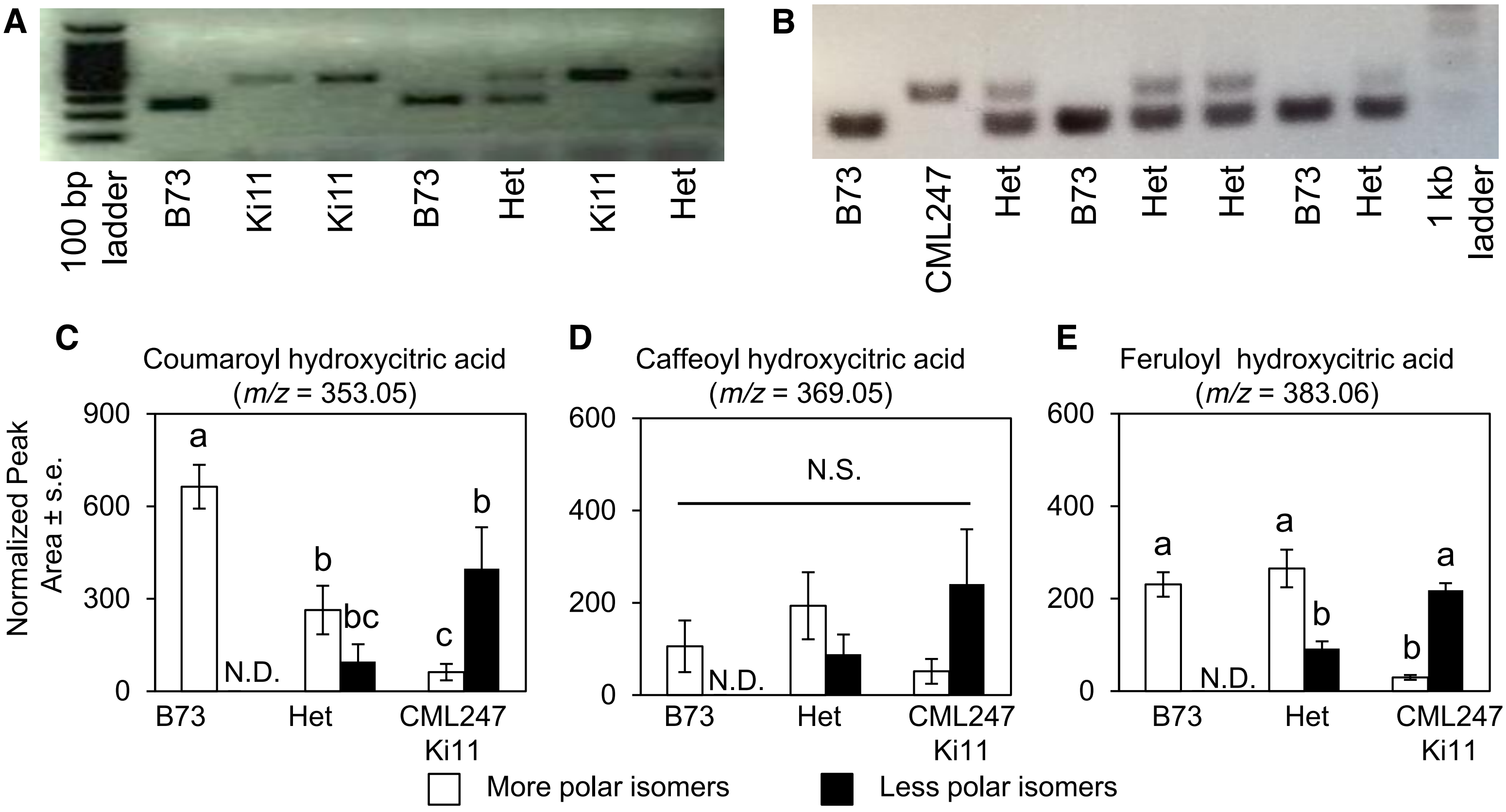


Figure 12. Different isomers of phenylpropanoid hydroxycitric acid esters co-segregate with genetic markers at QTL on Chromosome 4 across near isogenic lines. PCR-based genotyping of B73 x Ki11 (A) and B73 x CML247 (B) near-isogenic lines. Graphical representations of all (C) coumaroyl hydroxycitric acid, (D) caffeoyl hydroxycitric acid, and (E) feruloyl hydroxycitric acid isomers found in near isogenic lines of different genotypes normalized by the total ion concentration of each sample and the total normalized peak area compared across genotypes and between each other by two-way ANOVA followed by Tukey HSD. Groups of different significance levels are indicated by different letters ($p < 0.05$). N = 14 (B73), 3 (Het), and 4 (CML247/Ki11). N.S. = not significant, N.D. = not detected, Het = heterozygote.