Biochemical Networks and Epistasis Shape the *Arabidopsis thaliana* Metabolome

Heather C. Rowe,a,1 Bjarne Gram Hansen,b,1 Barbara Ann Halkier,b and Daniel J. Kliebenstein,a,2

a Genetics Graduate Group and Department of Plant Sciences, University of California Davis, Davis, California 95616
b Plant Biochemistry Laboratory, Department of Plant Biology, Faculty of Life Sciences, University of Copenhagen, 1871 Frederiksberg C, Copenhagen, Denmark

Genomic approaches have accelerated the study of the quantitative genetics that underlie phenotypic variation. These approaches associate genome-scale analyses such as transcript profiling with targeted phenotypes such as measurements of specific metabolites. Additionally, these approaches can help identify uncharacterized networks or pathways. However, little is known about the genomic architecture underlying data sets such as metabolomics or the potential of such data sets to reveal networks. To describe the genetic regulation of variation in the *Arabidopsis thaliana* metabolome and test our ability to integrate unknown metabolites into biochemical networks, we conducted a replicated metabolomic analysis on 210 lines of an *Arabidopsis* population that was previously used for targeted metabolite quantitative trait locus (QTL) and global expression QTL analysis. Metabolic traits were less heritable than the average transcript trait, suggesting that there are differences in the power to detect QTLs between transcript and metabolite traits. We used statistical analysis to identify a large number of metabolite QTLs with moderate phenotypic effects and found frequent epistatic interactions controlling a majority of the variation. The distribution of metabolite QTLs across the genome included 11 QTL clusters; 8 of these clusters were associated in an epistatic network that regulated plant central metabolism. We also generated two de novo biochemical network models from the available data, one of unknown function and the other associated with central plant metabolism.

INTRODUCTION

Understanding the molecular and genetic bases of complex traits like disease resistance, growth, and development is a unifying goal in diverse scientific fields. Genetic variation regulating complex traits in natural populations is largely quantitative and polygenic and can interact with environmental, epigenetic, and other genetic factors (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Quantitative trait mapping, the most common approach to the analysis of complex traits, measures the association of genetic markers with phenotypic variation, delineating quantitative trait loci (QTL) (Liu, 1998; Lynch and Walsh, 1998). Computational and genomic advances have generated increasingly precise QTL maps for a wide array of traits, ranging from development and morphology to metabolism and disease resistance (Kliebenstein et al., 2002b; Lexer et al., 2005; Symonds et al., 2005; Anderson et al., 2006; Keurentjes et al., 2006; Hoffmann and Weeks, 2007; Yagil et al., 2007). However, the molecular bases of many quantitative traits remain unknown despite the long history of QTL identification (Sax, 1923).

Quantitative trait analysis is enhanced by the use of microarray technology to measure global transcript levels in mapping populations and to map expression QTLS (eQTLS) (Jansen and Nap, 2001; Doerge, 2002; Schadt et al., 2003). Whole genome eQTL analysis in multiple organisms has revealed that gene expression traits are highly heritable, with a complex genetic architecture (Brem et al., 2002; Schadt et al., 2003; Morley et al., 2004; Keurentjes et al., 2007; West et al., 2007). These studies found large numbers of both cis- and trans-acting eQTLS, with evidence of nonadditive genetic variation such as epistasis and transgressive segregation as well as genetic variation altering entire transcriptional networks (Kliebenstein et al., 2006; Keurentjes et al., 2007; Potokina et al., 2008). Recent work directly links eQTLS to phenotypic alterations in specific metabolic pathways, highlighting the complexity of interactions between transcript and metabolite variation (Sønderby et al., 2007; Wentzell et al., 2007; Hansen et al., 2008). These analyses suggest significant differences between the organization of genetic regulation of transcripts and metabolites for a specific subset of *Arabidopsis thaliana* secondary metabolites (Wentzell et al., 2007), but this hypothesis has not been broadly tested within the plant metabolome.

Quantitative genetic analysis in plants has enabled the detailed molecular dissection of several secondary metabolite biosynthetic pathways (Magrath et al., 1993; McMullen et al., 1998; Kliebenstein et al., 2001b, 2001c; Szalma et al., 2005). In addition, broad-spectrum metabolite analyses now allow QTL mapping of an expanded portion of the plant metabolome (Keurentjes et al., 2006; Schauer et al., 2006; Meyer et al., 2007). This
approach has confirmed known secondary metabolite QTLs and aided the identification of a new enzymatic step in a known secondary metabolite pathway (Kliebenstein et al., 2001a; Kroymann et al., 2003; Keurentjes et al., 2006). It remains to be tested whether metabolomic analyses can generate hypotheses regarding new pathways or recreate linkages between known metabolic pathways.

Comparative analysis of metabolite and developmental variation suggests an integral link between plant central metabolism and development/physiology, but QTLs for metabolite and developmental traits were not colocalized more than expected by chance (Keurentjes et al., 2006; Meyer et al., 2007). This lack of overlap between known development and metabolite QTLs may indicate that genetic regulation of plant metabolism is more complex than presumed, such that current studies lack sufficient power to detect the majority of metabolite QTLs present in a population. The level of genetic complexity regulating plant metabolism will determine the size of the structured mapping populations necessary for effective QTL analyses and could potentially affect the methodology and interpretation of association mapping studies (Beavis, 1994, 1998; Nordborg et al., 2005; Clark et al., 2007).

To begin describing the genetic regulation of variation in the Arabidopsis metabolome, we conducted metabolomic analyses on the Arabidopsis Bayreuth-0 (Bay) × Shahdara (Sha) recombinant inbred line (RIL) population (Loudet et al., 2002). This population has previously been utilized for targeted metabolite QTL and global eQTL analysis (Loudet et al., 2003; Calenge et al., 2006; Kliebenstein et al., 2006; Wentzell et al., 2007; West et al., 2007). Metabolic traits were less heritable than were global transcript levels, suggesting that metabolite accumulation may be more susceptible to environmental influence. Statistical analysis identified a large number of metabolite QTLs with moderate phenotypic effects, and informed pairwise marker tests showed that epistasis strongly influences the genetic architecture underlying the Arabidopsis metabolome. Eleven QTL clusters influenced more metabolites than expected. Eight of these clusters are associated in an epistatic network that appears to regulate plant central metabolism. Two clusters were associated with previously identified secondary metabolite loci (Kliebenstein et al., 2001b, 2001c). These results show that the genetic architecture underlying the Arabidopsis metabolome is highly complex and governed by numerous epistatic interactions. Interpreting relationships among these QTLs will require the analysis of significantly larger mapping populations. In spite of the limited power available in 210 RILs, we were able to identify two de novo biochemical networks, one of unknown function and the other associated with central plant metabolism.

RESULTS

Metabolite Distribution and Detection

We utilized the University of California Davis Metabolomics Core gas chromatography–time of flight–mass spectrometry (GC-TOF-MS) metabolomics platform to measure metabolite accumulation in the Arabidopsis accessions Bay and Sha, the parents of the Bay × Sha RIL population (Loudet et al., 2002; Nikiforova et al., 2005). This GC-TOF-MS platform is believed to detect predominantly primary metabolites within plant samples, and metabolites are identified based on comparison with reference spectra (Roessner et al., 2001; Meyer et al., 2007; http://fiehnlab.ucdavis.edu/Metabolite-Library-2007). The Bay and Sha parents significantly differed in metabolite accumulation for 61 of 396 metabolites detected in this preliminary experiment (see Supplemental Data Sets 1 and 2 online). Metabolites differing significantly between Bay and Sha included phosphoric acid and a diverse set of amino acids, sugars, and fatty acids. Each parent had the highest levels of roughly equal numbers of metabolites, with Sha showing higher levels of tricarboxylic acid cycle (TCA) and pentose phosphate–associated metabolites, while Bay was typically higher in amino acids and storage sugars (see Supplemental Data Set 2 online). Consistent with a global analysis of variable transcript accumulation in these same parents, the average metabolite heritability as measured using just the Bay and Sha parents was 7% (Figure 1A). The distribution of metabolite and transcript heritabilities was similar using these

![Figure 1. Metabolite and Transcript Level Heritabilities.](image-url)

(A) Histogram of estimated broad-sense heritability values in RIL parental accessions Bay and Sha for metabolites (black) and transcripts (gray). Values are presented as percentages of the 397 total metabolites or 22,746 total transcripts with a given heritability.

(B) Histogram of estimated broad-sense heritability values in 210 Bay × Sha RILs for metabolites (black) and transcripts (gray). Values are presented as percentages of the 371 total metabolites or 22,746 total transcripts with a given heritability.
two parents, with transcripts having a slightly heavier tail (Figure 1A).

To measure metabolomic variation in the progeny from Bay and Sha, four biological replicate samples from each of 210 Bay × Sha RILs were analyzed via a GC-TOF-MS platform. The majority of the metabolites (330) were detected in both parents and >90% of the RILs, while a subset of metabolites were detected in one parent but not the other (40 in Bay only and 28 in Sha only). As found in previous Arabidopsis metabolite QTL studies, a significant number of metabolites (226) was detected in the RILs only (Kliebenstein et al., 2001c, 2002a; Lambrix et al., 2001; Keurentjes et al., 2006) (see Supplemental Figure 1 online). The presence or absence of metabolites was likely caused by a mixture of polymorphisms with qualitative and quantitative effects, as over half of the known metabolites were detected in fewer than half of the RILs. This included pyruvate, salicylic acid, citrulline, and Met, all metabolites with central roles in plant metabolism. These metabolites are likely present in all of the RILs but below the detection threshold in some lines.

In a prior study of two specific Arabidopsis secondary metabolite pathways, average heritability estimates for metabolite traits were less than those for the transcripts encoding the biosynthetic enzymes (Wentzell et al., 2007). Our analysis of heritability for all metabolites detected in both replicate RIL experiments showed that the average heritability was 25% and the highest heritability was 55% (Figure 1B; see Supplemental Data Sets 3 and 4 online; these data sets also include measures of all other estimatable sources of variance for this experiment). This range of metabolite heritability was much lower than the distribution of global transcript heritability, suggesting that these metabolite traits in Arabidopsis were subject to more influence and/or noise from both the internal (physiological) and external plant environment (Figure 1B).

The difference in metabolite heritability, in terms of both qualitative metabolite presence and quantitative metabolite content, between the RILs and parents resulted from transgressive segregation. It increased the level of genetically regulated phenotypic variation within the RILs in comparison with the parents, thus leading to an increase in genetic variance or heritability as measured in the RILs. Transgressive segregation for metabolite presence was manifested in the significant fraction of metabolites found in both parents that were not detected in all RILs as well as in metabolites found in neither parent but detected in a number of the RILs (see Supplemental Figure 1 online). Analysis of the RILs for 330 metabolites found in both parents identified positive and negative transgressive segregation for metabolite accumulation (see Supplemental Data Figure 2 and Supplemental Data Set 5 online). Positive transgressive segregation was found for 143 metabolites, in which at least half of the RILs had metabolite accumulation more than twice the value of the highest parent and <5% of the RILs had values lower than the lowest parent. One hundred thirty-eight metabolites showed negative transgressive segregation, in which half of the RILs had metabolite accumulation lower than half of the lowest parent and <5% of the RILs accumulated metabolite levels higher than the highest parent. Of the remaining metabolites, the majority showed transgressive segregation in both directions (see Supplemental Data Set 5 online). Thus, Bay and Sha possess significant genetic variation for metabolite accumulation that is not evident in the parental phenotypes.

Metabolomic QTL Location and Effect

Data from 557 metabolites present in the 210 Bay × Sha RILs were used to map QTLs. The composite interval mapping (CIM) algorithm Within QTL Cartographer identified 438 QTLs affecting 243 metabolites (Figure 2; see Supplemental Data Set 6 online). This included 77 putatively identified metabolites and 166 unknown compounds (http://fiehnlab.ucdavis.edu/Metabolite-Library-2007). For 315 metabolites, no QTLs were detected using CIM: these were equally split between metabolites detected and not detected in the parents. Metabolic QTLs were not equally distributed across the genome, and we identified 11 regions that contained more metabolite QTLs than expected by random chance (permutation estimated P < 0.05; Figure 2). Five of these metabolite QTL clusters (Met.chromosome_centimorgan [cM], Met.II.15, Met.II.47, Met.III.04, Met.IV.65, and Met.V.67) colocalated with previously identified eQTL hot spots, suggesting a link with transcript variation (West et al., 2007). The remaining loci did not associate with eQTL hot spots. However, colocalization of the AOP and Elong loci with known QTLs that determine aliphatic glucosinolate accumulation, glucosinolate structure, and transcript accumulation for aliphatic glucosinolate biosynthetic genes suggests that eQTL hot spots can be pathway- or network-specific and may not appear within genome-scale analyses (Wentzell et al., 2007).

To further investigate the allelic effects of these metabolite QTL clusters, we utilized analysis of variance (ANOVA) to test the marker closest to the peak of each metabolite QTL cluster for association with metabolite accumulation. We individually analyzed all 557 metabolites detected in the RILs to test whether variation in their accumulation was associated with any of the metabolite QTL clusters. After correction for multiple comparisons, this analysis showed that 372 metabolites associated with at least one QTL hot spot (see Supplemental Data Set 6 online). Increased QTL detection via this approach could be attributed to the inclusion of significant QTLs in the model, thus decreasing error variance and increasing the power to detect QTLs that may have been marginal in the CIM analysis. This approach had crude similarities to multiple-interval mapping approaches. By contrast, the CIM analysis focused on individual loci and did not allow for the error term to be adjusted downward for significant QTLs.

The ANOVA provided two estimates of allelic effect for each significant QTL. First, the genetic r² estimated the proportion of phenotypic variance attributable to genotype at that specific locus. Second was the allelic substitution effect, which expressed the change in phenotype associated with substitution of one parental allele for the other at the queried locus (Figure 3). In a population possessing a broad range of phenotypic values, allelic substitution effects may be more reflective of the biological impact of a given locus than genetic r². For example, if a population has a 10-fold range in phenotype, a specific QTL of 10% r² will have a 1-fold phenotypic difference or 100% allelic substitution effect. Since the phenotypic range for many metabolites in this RIL population was quite large, the majority of
detected QTLs had an allelic substitution effect size of 30% or greater despite small \( r^2 \) values (Figure 3; see Supplemental Data Set 5 online). These low estimates of genetic \( r^2 \) also suggested that numerous unidentified genetic loci influenced metabolite concentrations in this population.

**HIF195 and Genetic Limitation on Detection Power**

The power to detect a QTL is predominantly determined by its genetic \( r^2 \), the amount of phenotypic variance apportioned to a specific QTL within the given population. Within our analysis of...
the Bay × Sha population, the wide range of metabolite concentrations caused QTLs with strong allelic effects to have low genetic \( r^2 \) per QTL. This suggested that our analysis may have been underpowered and that numerous QTLs were not detected in this replicated analysis of 210 Bay × Sha RILs. To test this possibility, we utilized a higher power analysis focusing upon a single genomic region to test whether we could identify QTLs that were not observed in the whole genome analysis. To do this, we obtained HIF195, a heterogenous inbred family (HIF) previously used to validate a fructose QTL on chromosome IV in the Bay × Sha RILs (Calenge et al., 2006). The two HIF195 genotypes are isogenic except for a small region on chromosome IV where one line is Bay and the other is Sha (see black box in Figure 2A). These lines allow a specific test of the allelic effect of a substitution at this genomic position.

The analysis of the 210 RILs revealed 10 metabolite QTLs in this region, including the expected fructose QTL, but this region did not exceed the significance threshold delineating metabolite QTL clusters (Figure 2A). We measured metabolite accumulation in the two HIF195 genotypes using a replicated design with the same growth conditions and analytical protocols as for the 210 RIL analysis and measured 297 metabolites. Scrutiny of the 10 metabolites with QTLs identified in the HIF region by genome-wide analysis showed that 8 metabolites differed between the two HIF genotypes at \( P = 0.05 \), with 3 of those, including fructose, being significant at \( P = 0.01 \) (see Supplemental Data Set 7 online). Expanding the analysis to the 287 other metabolites detected in this experiment identified significant differences in the accumulation of 42 additional metabolites between the two HIF195 genotypes, with an average allelic substitution effect of 28%. These differences were significant at the \( P = 0.01 \) level, where we expect three false-positives in 297 tests (64 total metabolites were significant at \( P = 0.05 \); see Supplemental Data Set 7 online). Polymorphism(s) in this region of the genome, therefore, altered diverse aspects of primary metabolism (Figure 4). Given the broad impact of this single region, which was not initially identified as a metabolic QTL hot spot, we conclude that recombination in the 210 RILs was likely insufficient to allow the detection of all metabolite QTLs present in this population (Beavis, 1994, 1998). HIF analysis may also increase QTL detection power if HIF195 possesses a genetic background that optimizes alleles present at loci epistatic to those directly tested by HIF195 to increase measured QTL effects. However, because genotypes at nontarget loci are fixed randomly in the HIF, any specific HIF has an equal chance of increasing or decreasing the power to detect significant differences in an epistatic combination. Where resources are available, testing multiple HIFs for a given region would allow the discrimination of genetic background effects.

**Metabolomic QTLs and Metabolic Pathways**

To better understand the relationships among metabolite QTL clusters, we clustered metabolites based on QTL position and allelic effect, thus connecting metabolite QTL clusters based on shared regulation of specific metabolites. For instance, the \( AOP \) locus altered a predictable set of metabolites, given its role in glucosinolate metabolism, with metabolites such as Met and glucose-1-phosphate (G-1-P) forming a discernible cluster defined by a strong effect of \( AOP \) (Figure 2B; see Supplemental Data Set 8 online). Clustering analysis suggested that polymorphisms at the Met.I.42, Met.I.80, Met.II.15, and Met.II.47 loci affected a core set of metabolic pathways (see white box in Figure 2B). The other identified metabolite QTL clusters did not show specific metabolic pathway associations.

To visualize QTL effects, we developed a rough map of central plant metabolism (Mueller et al., 2003; Zhang et al., 2005) and plotted the known metabolites influenced by each QTL cluster (Figure 4). This showed that the previously associated Met.I.42, Met.I.80, Met.II.15, and Met.II.47 loci had global effects on central metabolism (Figures 2A and 4). While these four loci altered the accumulation of numerous unknown metabolites, they showed a significant enrichment for known metabolites in central metabolism \( (\chi^2 \text{ test } P < 0.001 \text{ for all four QTLs}; \text{ Figure 4}; \text{ see Supplemental Data Set 8 online}) \). Most central metabolites
Figure 4. Metabolite QTL Hot Spots Alter Primary Metabolic Networks.
showed a similar direction of allelic effect for a given QTL, such that for Met.I.42 and Met.II.15, the Bay allele led to higher accumulation for all of the plotted metabolites, while in Met.II.47, the Sha allele led to elevated metabolite levels (Figure 4). A similar bias was detected for the unknown metabolites (see Supplemental Data Set 4 online). In Met.I.80, the Sha allele led to higher accumulation of all metabolites except glucose and fructose, which had decreased accumulation (Figure 4). This suggested that a polymorphism in Met.I.80 might have altered the interconversion of glucose-6-phosphate and fructose-6-phosphate into glucose and fructose, respectively. However, none of the Arabidopsis hexose kinases are located in this region, suggesting that Met.I.80 was not caused by a hexose kinase polymorphism.

In contrast with the above four metabolite QTL clusters with primarily central metabolism effects, the other seven clusters showed no significant bias for known or unknown metabolites (χ² test P > 0.10 for all seven QTLs) nor any significant metabolite groupings suggestive of particular functions (Figure 4; see Supplemental Data Set 8 online). For instance, the AOP locus altered the accumulation of Met and G-1-P, as would be expected given the accumulation for all of the plotted metabolites, while in Met.I.80, the Sha allele led to elevated metabolite levels (Figure 4). A similar pattern of diffuse network effects on central and unknown metabolites was found for the other metabolite QTL clusters (see Supplemental Data Set 8 online).

Logic Approach to Biochemical Network Generation with QTls

Combining natural genetic variation with quantitative metabolite analysis enabled the identification and validation of parts of the aliphatic glucosinolate biosynthetic pathway (Magrath et al., 1993; Kliebenstein et al., 2001a, 2001c; Kroymann et al., 2003; Keurentjes et al., 2006; Wentzell et al., 2007). However, the AOP QTL was additionally associated with altered aromatic amino acid and fatty acid accumulation. The AOP QTL also altered the accumulation of numerous metabolites of unknown association with the aliphatic glucosinolate pathway. A similar pattern of diffuse network effects on central and unknown metabolites was found for the other metabolite QTL clusters (see Supplemental Data Set 8 online).
was not a component of the reference metabolomics database, so it is possible that it was one of the unknown compounds. Additionally, the presence of succinate suggested that this biochemical network also involved the TCA cycle. Identification of these two networks suggested that a logic-based approach could derive biochemical networks directly from metabolomics QTL data. However, these associations did not provide direct evidence for the molecular nature of the QTL or the biochemical association.

The AOP and Elong QTLs and Biosynthetic Linkages

In addition to the logic approach to biosynthetic pathway analysis, previous work has shown that epistatic QTLs can provide insight into biochemical relationships between metabolites within a pathway (Mithen and Campos, 1996; Kliebenstein et al., 2001b; Lambrix et al., 2001; Keurentjes et al., 2006; Zhang et al., 2006; Wentzell et al., 2007). These studies focused on epistasis between the AOP and Elong QTLs, in which all detected aliphatic glucosinolates identified an interaction between AOP and Elong (Figure 6). We tested whether it was possible to link metabolites to the glucosinolate biosynthetic pathway based on their regulation by an epistatic interaction between AOP and Elong. We hypothesized that regulation of two metabolites by the same QTL-QTL interaction was less likely to result from chance or false detection than from shared regulation of two metabolites by a single QTL.

This epistatic association test identified 31 metabolites whose accumulation was determined by an epistatic interaction between these two loci (Figure 6). The majority of these 31 metabolites were unidentified compounds that share common QTLs with known glucosinolates and showed either positive or negative epistasis between AOP and Elong (Figure 6). This suggested that these metabolites, if not aliphatic glucosinolates themselves, may be intermediates in the biosynthetic pathway. Since some intermediates in this pathway are elongated Met

Figure 5. QTL Utilization for Pathway Generation.

(A) The hypothetical impact on the accumulation of two metabolites in a biochemical pathway if the pathway linking the two metabolites is altered via QTLs at two different positions. One metabolite's accumulation is positively affected, while the other's is negatively affected by QTLX. Both metabolites are positively affected by QTLY.

(B) to (D) Arrows link metabolites showing a positive or negative association with a QTL. Arrows are double-headed to show that biochemical directionality cannot be ascertained from these data. Labels above or next to the arrows show the chromosome (Roman numerals) and centimorgan (Arabic numerals) positions of QTLs. Vertical and horizontal lines link metabolites that show a similar direction of effect for a given QTL. Metabolite labels are described in Supplemental Data Set 1 online. Abbreviations for glucosinolates are as follows: 8MSO, 8-methylsulfinyloctyl; 8MT, 8-methylthiooctyl; 4MT, 4-methylthiobutyl; 4MSO, 4-methylsulfinylbutyl; 3MT, 3-methylthiopropyl; 3OHP, 3-hydroxypropyl. (B) shows glucosinolate metabolite linkages. (C) represents metabolites showing pathway linkages with 4-picolinate/isonicotinate. (D) represents metabolites showing pathway linkages with shikimate. Known primary metabolites were used to separate the metabolites into previously defined central metabolic pathways.
derivatives and related organic acids, they are likely detectable with the GC-TOF-MS platform (Textor et al., 2007). Failure to identify QTLs at the ESP locus (controlling glucosinolate hydrolysis) for these unknown metabolites suggests that they are not hydrolysis products. Interestingly, this genetic approach identified metabolites peripherally associated with glucosinolate synthesis, such as glucose, glucose-1-phosphate, and Gly, as well as metabolites such as uracil and homoglutamine with no obvious linkage to aliphatic glucosinolate metabolism. Some of these metabolite associations could be explained by these additional metabolites having QTLs linked to, but not caused by, AOP and Elong. However, the requirement for epistatic interaction between these hypothetical additional QTLs diminishes the likelihood that this was a random association. As such, known epistatic loci facilitated building hypotheses regarding the association of additional metabolites with glucosinolate biosynthesis. While further work is required to validate the association of these metabolites with aliphatic glucosinolates, the association of unknown metabolites with known biosynthetic pathways enhances the utility of metabolite databases and suggests approaches to determining a biological role for these unknown metabolites.

**Pairwise Epistasis**

QTL studies focused on the accumulation of specific metabolites in Arabidopsis frequently identify epistatic interactions between QTLs with significant main effects (Kliebenstein et al., 2001b, 2002a; Loudet et al., 2003; Calenge et al., 2006; Pfalz et al., 2007). To test whether the global metabolome showed a similar enrichment in epistatic interactions, we formally tested all 55 pairwise epistatic interactions between the 11 metabolite QTL clusters against the average accumulation of 557 metabolites within the RILs. This informed model used less than one-third of the available degrees of freedom. Using a false discovery rate (FDR)–adjusted P value (<0.05), 240 metabolites were identified with between one and five significant epistatic interactions, for a total of 328 significant interactions (FDR < 0.1 identified 497 metabolites with 1538 significant interactions; see Supplemental Data Sets 9 and 10 online). For the majority of metabolites, the significant epistatic terms explained as much or more of the genetic variance, in terms of genetic $r^2$, as the combined main effect QTLs (Figure 7A). However, individual main effect QTLs showed larger effects than did the individual epistatic terms (Figure 7B). This shows that epistasis was widely prevalent in the determination of metabolite levels within this Arabidopsis RIL population. We did not conduct an unsupervised search of all possible pairwise epistatic interactions, as the use of only 210 RILs did not provide sufficient degrees of freedom.

**Pairwise Epistasis and QTL Association**

Research on mutants in yeast metabolic genes identified epistasis more frequently between loci that function within the same or closely related metabolic pathways (Segre et al., 2005). As such, QTLs whose pairwise epistatic interactions regulate an enhanced number of metabolites may regulate a metabolic network. Because the pairwise marker comparisons have a

---

**Figure 6.** Epistatic Interaction of AOP and Elong.

The average metabolite accumulations in terms of percentage difference from the RIL population average are shown for all metabolites detecting a significant epistatic interaction between the AOP and Elong loci. Glucosinolate abbreviations are as described for Figure 5. The RILs were divided into four genotypic classes at AOP and Elong, and the average metabolite accumulation per class was determined. Red shows that the value was greater than the RIL population average, while green shows that the value was less than the RIL population average. Black shows that the value was not different from the average. Glucosinolate metabolites measured in a different experiment (Wentzell et al., 2007) are shown in boldface underline. Unknown metabolites are labeled with their database numbers. The average trait divergence from the RIL population across the four genotypic groups was used to estimate distances between the metabolites using the absolute value of the Pearson moment correlation. Clustering was conducted using the weighted group average linkage. Maximal distance is 1.48 on this diagram. The dominant class of paired alleles for each major cluster is distinguished by a separate line color, and the pairwise AOP and Elong genotypes are as labeled. The bar at left of the heat plot indicates whether a given metabolite exhibits positive (red) or negative (blue) epistasis (e) between the AOP and Elong QTLs.
identified metabolite QTL clusters. The presented pairwise marker interactions connected 8 of the 11 metabolites than would be expected by chance (Figure 8). The overrepresentation of pairwise epistatic interactions as affecting more metabolites than expected. This identified 9 of the 55 possible pairwise interactions was utilized to estimate the type III sums-of-squares for ANOVA including all 11 main-effect markers and their 55 pairwise interactions. For all significant epistatic terms, the genetic variance was divided by the number of metabolites ($r^2$ per each term). This gave an estimate of the genetic value of the individual epistatic and main-effect term, the genetic $r^2$ every fifth percentile ($y$ axis) is presented. Values $>1$ indicate that significant epistatic terms explained more of the genetic variance than did significant main-effect terms.

For each interaction, we estimated a measure of epistasis ($\epsilon$). A positive value of $\epsilon$ indicated that recombinant genotypes had less metabolite than the parental genotypes; negative $\epsilon$ indicated elevated metabolite content in recombinant genotypes (Figure 8). The $AOP \times Elong$ interaction displayed a mixture of positive and negative epistasis. Such a mixture of epistatic effects on a set of metabolites suggested that the polymorphisms creating these epistatic QTLs were located within the biosynthetic pathway. We applied this interpretation to other, less characterized epistatic interactions. In the $Met.II.47 \times Met.V.67$ interaction, positive epistasis only affected fatty acid accumulation, while negative epistasis affected all other measures of central metabolism (see Supplemental Figure 3F online). This suggested that $Met.II.47 \times Met.V.67$ interacted to regulate the partitioning of carbon between fatty acids and other components of central carbon metabolism. Similarly, in the $Met.I.80 \times Met.II.47$ interaction, the sign of epistasis changed between short- and long-chain fatty acids, suggesting that this interaction may regulate fatty acid elongation (see Supplemental Figure 3E online). These analyses suggested that the $Met.II.47$ QTL affects the partitioning of carbon within the plant and may be a major regulator of Arabidopsis metabolism. The remaining pairwise marker interactions showed exclusively positive (see Supplemental Figures 3A and 3B online) or negative (Figure 8B; see Supplemental Figures 3B and 3G online) epistasis. All together, this analysis showed that considering the sign of epistatic interactions can enhance the interpretation of metabolite QTL effects.

Three-Way Epistasis

Most studies of plant quantitative traits are limited to the analysis of pairwise epistasis, but this lack of information does not preclude the potential for higher order interactions. Analysis of the main effects of three central metabolism QTLs, $Met.I.42$, $Met.II.15$, and $Met.II.47$ (Figure 2), and their pairwise interactions showed that these QTLs may coordinate central metabolism through metabolites associated with the TCA cycle (Figure 8; see Supplemental Figure 3 online). We used an ANOVA model to test whether TCA metabolites identify a three-way epistatic interaction between $Met.I.42$, $Met.II.15$, and $Met.II.47$. Five of the eight TCA metabolites that biosynthetically link citrate to Gln showed significant three-way epistasis between $Met.I.42$, $Met.II.15$, and $Met.II.47$ (Figure 9). Interestingly, a single recombinant genotype had elevated content of all five metabolites (Figure 9). All individuals in this genotypic class, $Met.I.42^{BAY}$, $Met.II.15^{BAY}$, and

---

**Figure 7.** Comparison of Main QTL versus Epistatic Effects.

ANOVA including all 11 main-effect markers and their 55 pairwise interactions was utilized to estimate the type III sums-of-squares for each term. This gave an estimate of the genetic $r^2$ per each term. The nonsignificant terms were omitted from further analyses.

(A) Histogram of per metabolite QTL $r^2$ values. For each metabolite, $r^2$ estimates for all significant main-effect terms were summed. $r^2$ estimates for all significant epistatic terms were also summed. The $r^2$ due to epistatic terms was divided by the $r^2$ due to the main-effect terms. The number of metabolites ($y$ axis) with a given ratio ($x$ axis) is presented. Values $>1$ indicate that significant epistatic terms explained more of the genetic variance than did significant main-effect terms.

(B) Quantile–quantile plot of per individual QTL epistatic $r^2$ ($y$ axis) versus per individual main-effect $r^2$ ($x$ axis). For each significant epistatic or main-effect term, the genetic $r^2$ was determined. The plot shows the value of the individual epistatic and main-effect $r^2$ every fifth percentile across both entire distributions. The diagonal line shows the expected plot if the epistatic and main-effect $r^2$ distributions were identical.

A broad range of two-way allele frequencies, we used a permutation approach to identify pairwise interactions regulating more metabolites than expected. This identified 9 of the 55 possible pairwise epistatic interactions as affecting more metabolites than would be expected by chance (Figure 8). The overrepresented pairwise marker interactions connected 8 of the 11 identified metabolite QTL clusters. The $Met.II.47$ QTL was the main hub of this network, connecting directly to five other metabolite QTL clusters and apparently altering the accumulation of metabolites associated with the TCA cycle and sulfur metabolism. Interestingly, the interaction between $Met.II.47$ and $Met.I.16$ altered the accumulation of nearly all fatty acids (Figure 8B). This suggests that the often overlooked effect of epistasis can be utilized to aid in network generation with large genomic data sets. The epistatic interaction between $AOP$ and $Elong$, while previously experimentally validated, did not exceed the permutation-derived significance threshold, suggesting that our permutation approach to determine significance level left a number of false-negatives and that the observed epistasis network was likely conservative.

For each interaction, we estimated a measure of epistasis ($\epsilon$). A positive value of $\epsilon$ indicated that recombinant genotypes had less metabolite than the parental genotypes; negative $\epsilon$ indicated elevated metabolite content in recombinant genotypes (Figure 8). The $AOP \times Elong$ interaction displayed a mixture of positive and negative epistasis. Such a mixture of epistatic effects on a set of metabolites suggested that the polymorphisms creating these epistatic QTLs were located within the biosynthetic pathway. We applied this interpretation to other, less characterized epistatic interactions. In the $Met.II.47 \times Met.V.67$ interaction, positive epistasis only affected fatty acid accumulation, while negative epistasis affected all other measures of central metabolism (see Supplemental Figure 3F online). This suggested that $Met.II.47 \times Met.V.67$ interacted to regulate the partitioning of carbon between fatty acids and other components of central carbon metabolism. Similarly, in the $Met.I.80 \times Met.II.47$ interaction, the sign of epistasis changed between short- and long-chain fatty acids, suggesting that this interaction may regulate fatty acid elongation (see Supplemental Figure 3E online). These analyses suggested that the $Met.II.47$ QTL affects the partitioning of carbon within the plant and may be a major regulator of Arabidopsis metabolism. The remaining pairwise marker interactions showed exclusively positive (see Supplemental Figures 3A and 3B online) or negative (Figure 8B; see Supplemental Figures 3B and 3G online) epistasis. All together, this analysis showed that considering the sign of epistatic interactions can enhance the interpretation of metabolite QTL effects.
**DISCUSSION**

This study identified numerous QTLs regulating plant metabolism (Figure 2). The average metabolite’s accumulation was affected by multiple QTLs of moderate effect with epistatic interactions, such that the contribution of pairwise epistatic effects frequently equaled those of the main-effect QTLs. The detection of at least one higher order (three-way) epistatic interaction implies that fully querying this complexity will require larger mapping populations than are currently available (Figure 9). Identification of the molecular bases of these metabolic QTLs, especially QTLs controlling epistatic interactions, will provide useful potential to modify plant central metabolism. This will also improve our understanding of the molecular basis of epistasis as well as the integration of plant central metabolism with development and physiology. Given this estimated complexity, the medium-sized population (210 RILs) utilized for this study likely revealed only a fraction of the natural variation modulating metabolic diversity between the Bay and Sha accessions, let alone within Arabidopsis as a species. Thus, the genetic architecture regulating the Arabidopsis metabolome is highly complex.

The identified metabolite QTLs formed 11 clusters (Figure 2). Four of these clusters directly associated with central metabolism (Figure 4), and four different clusters epistatically associated with central metabolism (Figure 8). Two additional clusters, AOP and Elong, were associated with secondary metabolism, specifically aliphatic glucosinolate metabolism (Figure 6) (Kliebenstein et al., 2006). This contrasts with a previous report that three of four metabolite clusters detected in Arabidopsis, including the AOP and Elong loci, were secondary metabolite-associated (Keurentjes et al., 2006). This difference is likely caused by platform-specific detection bias, as the GC-TOF-MS platform detects predominantly primary metabolites in addition to numerous unknown metabolites (our study), while the liquid chromatography–mass spectrometry platform detects numerous secondary metabolites (Fiehn, 2002; Keurentjes et al., 2006).

**Metabolic Pathway versus Physiology**

Central metabolism QTL clusters frequently displayed directional bias in their allelic effects, such that all metabolites were either increased or decreased for a given genotype (Figure 4). Simple enzyme polymorphisms affecting the conversion of one metabolite to another would lead to differential allelic effects upon metabolites, depending on the relative positions of the enzyme and metabolites within the pathway (Figures 5 and 6). The unidirectionality of allelic effects for the central metabolism QTL clusters suggests a more complex basis for these QTLs.
Several of the identified metabolite QTL clusters also associate with major gene expression polymorphisms (West et al., 2007). These pleiotropic QTLs may regulate developmental or physiological differences in the plant that subsequently alter central metabolism. Alternatively, the effect of the QTLs upon central metabolism might indirectly alter development or physiology. An association between central metabolism and physiology/development has been found in other *Arabidopsis* RIL studies. In two studies, a major metabolite QTL cluster associated with an experimentally induced polymorphism in the developmental regulator protein *ERECTA* (Keurentjes et al., 2006; Meyer et al., 2007). Explaining the bases of central metabolism QTL clusters will require detailed analyses of physiological and developmental traits in the Bay × Sha population. At this stage, the association between growth and central metabolism is purely correlational and does not facilitate speculation regarding the specific direction of the causal relationships between metabolic polymorphisms and development/physiology polymorphisms. Determining the direction of these relationships will require cloning the underlying loci and validating their molecular functions. These relationships, however, do suggest that RIL analysis directed toward specific analysis of metabolomic variation may be most successful in

**Figure 9.** Three-Way Epistasis and the TCA Cycle.

ANOVA was utilized to test the eight TCA cycle metabolites for a three-way epistatic interaction between the Met.I.42, Met.II.15, and Met.II.47 metabolite QTL hot spots (abbreviations are as described for Figure 2). Each metabolite graph shows the average and SE of the metabolite accumulation in terms of percentage of the RIL population average (x axis) for the eight genotypic classes (y axis; genotypes labeled in the order Met.I.42, Met.II.15, Met.II.47, where B and S represent Bay and Sha, respectively). Statistical differences between the genotypic classes were estimated by pairwise t tests within the model. Letters indicate three-way genotypes that are statistically different at P < 0.05 using Tukey’s honestly significant difference test. The inset graph shows the distribution of Gln accumulation within the eight genotypic classes in terms of the RIL population average and SE.
populations that do not segregate for variation in genes known to regulate developmental or physiological phenotypes, such as flowering time. Eliminating known developmental variation would reduce the effects of developmental or physiological variation on the metabolome, but it would not eliminate the potential for metabolic variation to alter development or physiology.

Generating Metabolic Pathways via Logic Algorithms

Previous research suggested that natural metabolic variation may be useful in predicting and cloning genes in metabolic pathways (Kliebenstein et al., 2001b, 2001c, 2002b; Lambrix et al., 2001; Kroymann et al., 2003; Benderoth et al., 2006; Keurentjes et al., 2006; Wentzell et al., 2007). Within our QTL data set, unknown metabolites were sometimes paired such that one unknown was positively affected by a QTL while the other was negatively affected by the same QTL (see Supplemental Data Set 6 online). Application of a logic-based approach to the QTL data generated two putative biochemical networks (Figure 5). One network involved the pentose phosphate and TCA cycles but contained a number of unidentified metabolites. These unknown metabolites may be intermediates in these pathways or altered forms of known metabolites. The second network centered on 4-picoline, a compound that has not been detected previously within Arabidopsis. As such, further work is required to determine whether this is a breakdown product of Lys biosynthetic intermediates or it represents a new pathway for a previously undetected metabolite.

The arrows linking metabolites in models generated by this approach do not imply a single enzymatic step between these metabolites. Nor do they imply that the QTL is caused by polymorphism within a single enzyme. As found for glucosinolates, some major QTLs affect individual enzymes and also alter the expression of the entire biosynthetic pathway (Wentzell et al., 2007). Connection of metabolites by this approach indicates only their potential relationship via enzymatic interconversion. This relationship may be mediated by a single biosynthetic enzyme, an entire pathway, or, as found with the AOP locus, opposing reactions using the same unmeasured precursor (Figure 5B). While specific relationships between metabolites are not revealed, this approach allows the generation of new metabolic hypotheses and provides an ability to link unknown metabolites with known biosynthetic pathways through metabolomic QTLs. Increasing the power of QTL detection and mapping precision could greatly expand our ability to generate networks and explore their relationships.

Glucosinolate Bias?

Research suggesting the utility of genomics and QTLs for pathway generation has largely focused on the aliphatic glucosinolate biosynthetic pathway within Arabidopsis (Kliebenstein et al., 2001a, 2001c; Hirai et al., 2005, 2007; Keurentjes et al., 2006; Hansen et al., 2007; Sonderby et al., 2007). Recent results show altered sequence and gene expression polymorphism patterns in this pathway, indicative of selective pressures driving the formation and maintenance of structural diversity (Kliebenstein et al., 2001b, 2001c, 2002b; Lambrix et al., 2001; Kroymann et al., 2003; Benderoth et al., 2006; Keurentjes et al., 2006; Wentzell et al., 2007; Kliebenstein, 2008). Closer analysis of glucosinolate-associated unknown metabolites showed that they had significantly elevated heritability in comparison with the rest of the metabolites (glucosinolate metabolite heritability = 35%, metabolomic average = 25%; P < 0.001), providing greater detection power. Within Arabidopsis, this combination of high QTL detection power and diversifying selection may be unique to glucosinolate metabolites. Nevertheless, it appears that our logic approach to network generation can be extended to other metabolites, including unknowns (Figure 5). Other plant species may possess at least one secondary metabolite pathway with similar genetic characteristics to the glucosinolate pathway, potentially allowing these specialized pathways to be rapidly identified and characterized via similar genomics approaches.

Genetic Power

QTL studies frequently identify one or two large-effect QTLs per trait, along with a suite of small-effect QTLs. It has been proposed that creating a large number of small structured populations (<200 lines per parental pair) from a diverse set of germplasm will allow the characterization of most large-effect QTLs in a species. Our data, however, suggest that the average allelic effect is less limiting for QTL detection than is the number of recombination events (Beavis, 1994, 1998). Several metabolite QTL clusters were located <30 cm apart, diminishing our ability to discern their independent effects on the metabolome (Figure 2). The total genetic r² explained by main-effect and epistatic QTLs was on average <30%, suggesting the contribution of multiple unidentified genetic loci with moderate phenotypic effects, as found in a recent yeast eQTL study (Brem and Kruglyak, 2005). Strikingly, focused analysis of the HIF195 genotypes identified a large number of new metabolite QTLs not found in the 210 RILs, with most showing moderate (20 to 50%) allelic substitution effects (see Supplemental Data Set 8 online). The knowledge that this locus, previously identified as a fructose QTL (Calenge et al., 2006), alters additional metabolites may aid in the identification of candidate genes with the potential to influence multiple metabolites.

The prevalence of pairwise and potentially higher order epistasis also diminishes QTL detection power in small to medium populations. For instance, testing for a pairwise epistatic interaction splits the population into four genotypic classes versus only two groups for a main-effect QTL. Fewer measurements per genotypic class decrease statistical power. The potential impact of epistasis is magnified by the existence of three-way epistasis, such as found in this study (Figure 9). Reanalysis of total aliphatic glucosinolate data from 403 Bay × Sha RILs identified three different three-way interactions influencing >20% of trait variation (P < 0.001 for AOP × Elong × MYB28, AOP × Elong × GSL_ALIPH.III.15, and AOP × GSL_ALIPH.III.15 × MYB28), while no three-way interactions were detected in a previous study using a 210 RIL subset of the same population (Wentzell et al., 2007). This supports our hypothesis that higher order epistasis is common in the Arabidopsis metabolome and highlights the need for larger mapping populations to detect these interactions. This high level of epistasis may be a consequence of the interconvertability of metabolites within the metabolome, suggesting that
epistasis might be of greater consequence in metabolomics than transcriptomics.

Metabolite versus Transcript Heritabilities

Previous work had shown that for glucosinolates, the average metabolite had a lower heritability than did the transcripts encoding the biosynthetic pathway (Wentzell et al., 2007). This GC-TOF-MS metabolomics data set shows that the Arabidopsis metabolome as measured in the Bay × Sha RILs shows lower heritability than the transcriptome (Figure 1). Published estimates of technical error for the GC-TOF-MS and Affymetrix microarray platforms suggest that differences in the level of phenotypic variance explained are not caused predominantly by differences in measurement error between these two technologies (Weckwerth et al., 2004b). Additionally, heritability for the putative glucosinolate-related metabolites detected by the GC-TOF-MS platform (30%; Figure 6) was similar to the heritability for the glucosinolates in these same lines as measured by HPLC in a larger experiment using this same population (40%; Wentzell et al., 2007), which further supports the notion that this difference is not purely technical.

Differences in the genetic regulation of these two classes of trait may be explained by the simple fact that transcripts are functionally linked to potential DNA polymorphisms in their genes or regulators and, as such, have less potential for the introduction of biological noise between genetic variation and the measured transcript variation. In comparison, the measured metabolite variation requires a DNA polymorphism to be processed via transcription and translation. This extra step may introduce more biological noise. Another explanation for the difference in heritability between transcripts and metabolites is that metabolites are often interrelated, such that the atoms in one metabolite can potentially be transferred to a different metabolite in relatively few steps (Fell and Wagner, 2000; Jeong et al., 2000; Arita, 2004). This metabolic interconnectedness may increase the effect of small biological perturbations, introducing more noise in metabolic networks than in corresponding transcripts. These possibilities could be tested in a replicated experiment measuring global transcripts and metabolites in a structured population in the presence and absence of a controlled environmental perturbation (Van Leeuwen et al., 2007). If the metabolic network amplifies biological noise, then the genotype × environment interaction term should be greater for metabolites than for global transcripts.

METHODS

Plant Material and Experimental Conditions

Seeds for Arabidopsis thaliana accessions Bay, Sha, and the F8 generation of a Bay × Sha RIL population (Loudet et al., 2002) were obtained from The Arabidopsis Information Resource (stock number CSS57920; www.arabidopsis.org). The advanced generation RILs were created by single-seed descent from the F2 generation (Loudet et al., 2002). The RIL (F8) plants and parental accessions were grown in a single growth chamber and allowed to self-pollinate; sufficient seed was harvested from individual plants to conduct our replicated experiments as well as a previous microarray experiment utilized to map eQTL (Kliebenstein et al., 2006; West et al., 2007).

We conducted two independent measurements of the Bay × Sha RIL metabolites, separated by 2 months. Within each experiment, two plants for each of 210 RILs, plus parental controls, Bay and Sha, were grown using the growth chambers and conditions previously utilized for seed production and eQTL studies (West et al., 2006). Briefly, seeds were sown directly onto Premier ProMix B potting soil (Premier Brands) in 36-cell (125 cm³) soil volume per cell flats and grown in controlled-environment chambers at 20°C with 8 h of light at 100 to 120 µE. Within an experiment, two randomized blocks each contained one plant per RIL. Prior to planting, seeds were imbibed and cold-stratified at 4°C for 3 d to break dormancy. Seedlings were thinned to one plant per cell at 7 d after planting. At 6 weeks after germination, the plants were harvested and extracted for metabolomics analysis.

Metabolomics Analysis

For each metabolomics sample, one leaf disc from each of two leaves per plant was harvested, providing two leaf discs of ~20 mg total weight. All harvesting started at subjective midday, finishing within 2 h. The harvest order of the four blocks was randomized to minimize circadian effects. Each plant was independently harvested and extracted as per published protocols, providing a total of 840 samples, 24 for parental genotypes and 816 for the RILs (Weckwerth et al., 2004a, 2004b; Meyer et al., 2007). The samples were stored dry at 80°C until automated derivatization and GC-TOF-MS analysis at the University of California Davis Genome Center Metabolomics Facility (http://metabolomics-core.ucdavis.edu; Fiehn et al., 2005). Metabolite identity was determined by comparing retention time and mass with the University of California Davis Genome Center Metabolomics Facility metabolites database (http://fiehnlab.ucdavis.edu/Metabolite-Library-2007/; Fiehn et al., 2005). This library contains reference spectra for 713 known metabolites, generated by the analysis of purified reference compounds. Metabolites not contained within this library are listed as unknown or unidentified metabolites. This analysis identified 636 putative metabolites within the entire experiment (see Supplemental Data Set 1 online). After analysis, samples were subjected to quality control by further analysis of samples with internal standard values of >3 s from the experimental mean within a given experiment, suggesting faulty derivatization or GC-TOF-MS analysis of that specific sample. Given the number of samples, we chose 3 s rather than 2 s to exclude only extreme (biologically improbable) outlier samples. Because we had 210 lines with replication, the inclusion of 2 s outliers should have minimally affected the overall outcome. Excluding real 2 s outliers (caused by biology, not error) could significantly diminish the power of the experiment. Outlier samples were then tested for poor extraction quality by screening for samples with extreme (>3 s) outlier values for the average accumulation of amino acids and TCA metabolites. The flagged samples were distributed randomly among the RILs, suggesting that they were outliers due to technical error rather than genetic segregation. Removal of these samples from our analysis left a total of 769 metabolite measurements across 210 lines.

Estimation of Heritability

We calculated estimates of broad-sense heritability (H) for each metabolite as $H = \frac{\sigma_M^2}{\sigma_P^2}$, where $\sigma_M^2$ was the estimated metabolite level genetic variance among different genotypes in this sample of 210 RILs or two parental genotypes and $\sigma_P^2$ was the phenotypic variance for a trait (Liu, 1998). To estimate heritability between the Bay and Sha parental genotypes, we utilized the 371 metabolites detected in the parents in both independent RIL experiments plus an additional 26 metabolites that were detected in common between at least one of the two RIL experiments and a third parental experiment (see Supplemental Data Set 2 online). This
data set contained 18 biological replicates per parent (Bay and Sha): 6 replicates of each parent grown concurrently with each RIL experiment and 6 replicates of each parent in a third experiment containing just the parents. Heritability within the RILs was estimated for the 371 metabolites detected in both independent RIL experiments (see Supplemental Data Set 3 online). This data set included four biological replicates for each of the 210 RILs, permitting the estimation of heritability for each metabolite within the RIL population. Similar information is provided for those metabolites identified in only one of the two replicates, but these were not used for heritability comparisons (see Supplemental Data Set 4 online). Metabolite H was compared with previously estimated heritability values (West et al., 2007) for all transcripts using the same 210 lines.

QTL Analysis

The Bay × Sha RIL population has been genotyped previously (Loudet et al., 2002), and additional markers were obtained from the expression QTL analysis (West et al., 2006) as well as for the GSL-AOP and GSL-Elong loci (Kliebenstein et al., 2001b, 2002b; Textor et al., 2004; Benderoth et al., 2006; Wentzell et al., 2007). To detect metabolomic QTLs, we utilized the average metabolite accumulation per RIL across all experiments (see Supplemental Data Set 5 online). We tested only the 557 metabolites present in at least 5% of the RILs for QTLs. Seventy-five percent of those metabolites were detected in at least 25% of the RILs, and 232 of the metabolites detected in >80% of the RILs. There was a slight positive but insignificant correlation between the number of RILs in which a metabolite occurred and the number of QTLs identified for that metabolite (P = 0.08, r² = 0.04), suggesting that including metabolites present in a minority of RILs did not introduce significant bias.

The average per RIL accumulation for each metabolite was used to map QTLs within Windows QTL Cartographer version 2.5 (Basten et al., 1999; Zeng et al., 1999; Wang et al., 2006). CIM was implemented using Zmap (model 6) with a 10-cM window and an interval-mapping increment of 1 cM. Forward regression was used to identify five cofactors per metabolite trait. The declaration of statistically significant QTLs was based on permutation-derived empirical thresholds using 1000 permutations for each trait map (Churchill and Doerge, 1994; Doerge and Churchill, 1996). CIM using permutations to assign significance based on the underlying trait distribution is robust at handling normal or nearly normal trait distributions (Rebai, 1997), as found for most of our metabolites. The Eqtl module of QTL Cartographer identified the location, additive effect, and 1-logarithm of odds interval of each significant QTL for each trait (see Supplemental Data Set 6 online) (Wang et al., 2006).

QTL clusters were identified using a QTL summation approach in which the position of each QTL for each trait × experiment combination was indicated by a 1 and the number of traits regulated by a QTL at a given position was summed (Kliebenstein et al., 2006). This summation was conducted within a 10-cM sliding window. Permutation analysis with 1000 permutations determined a cluster size of 12 metabolite QTLS as the threshold for cluster significance (α = 0.05) (West et al., 2007). The QTL clusters identified defined genetic positions that were named with respect to their positions using the Arabidopsis Genome Initiative code of the marker closest to the peak. The last three digits of the Arabidopsis Genome Initiative code were removed to avoid confusion with the specific gene and to highlight that it was an approximate position. The QTLS detected at the previously characterized and cloned AOP and Elong loci were named as such (Magrath et al., 1994; Kliebenstein et al., 2001a, 2001b; Kroymann et al., 2003).

Metabolite QTL Hot Spot ANOVA

To test directly the effect of each identified QTL cluster on all metabolites, we used an ANOVA model containing the markers most closely associated with each of the 11 significant metabolite QTL clusters as individual main-effect terms. For each metabolite, the average accumulation in lines of genotype g at marker m was shown as ygm. The model for each metabolite was as follows:

\[ y_{gm} = \mu + \sum_{m=1}^{11} \sum_{n=1}^{11} M_{mg} + \sum_{m=1}^{11} \sum_{n=1}^{11} M_{mg} M_{np} + e_{gm}, \]

where \( g = \text{Bay or Sha} \) and \( m = 1, \ldots, 11 \). The main effect of the 11 markers was denoted as \( \mu \). The error, \( e_{gm} \), was assumed to be normally distributed, with mean 0 and variance \( \sigma^2 \). An automated SAS script tested all metabolites within the model and returned all P values, type III sums-of-squares for the complete model and each main effect, and QTL main-effect estimates (in terms of allelic substitution values) (see Supplemental Data Set 8 online). Significance values were corrected for multiple testing within each model using FDR < 0.05 in the automated script.

Logic Algorithm for Pathway Generation

To identify putative biochemical associations between metabolites, we utilized a logic algorithm. This logic algorithm first compartmentalized all metabolites into bins based on the metabolites having QTL peaks no farther apart than 2 cM, based on the QTL Cartographer output (see Supplemental Data Set 6 online). Assuming that a single locus is affecting the metabolites at a given QTL, that the metabolites are measured in the same samples, and that all statistics are performed with the same marker genotypes, the QTL peaks caused by a single locus should be very tightly linked. The bins were then tested for the presence of metabolites with allelic effects in the opposite direction, and only those containing metabolites with opposite allelic effects were kept. We then required the average absolute value of the allelic substitution effect within a bin to be >25%, using the glucosinolate QTLs as guides to set this threshold. Maintaining the threshold at this level allowed us to focus on the larger effect QTLs. Omitting this threshold led to the presence of more bins and a more complicated biochemical network but did not destroy the presented networks. This generated a list of potential biochemical associations between metabolites (see Supplemental Data Set 6 online). To generate pathways, we identified metabolites present in multiple bins and used them to link metabolites into potential biochemical pathways. All QTL information used to generate these putative networks was statistically validated as described in the QTL section. This approach provided a putative network that allowed the inclusion of known and unknown metabolites.

Metabolite QTL Epistasis Analysis

To test directly for epistatic interactions between the metabolite QTL clusters, we conducted an ANOVA using a model containing all 11 metabolite QTL clusters as individual main-effect terms using the markers most closely associated with each significant QTL cluster. Additionally, we tested all 55 possible pairwise interactions between these markers. For each metabolite, the average accumulation in lines of genotype g at marker m was shown as ygm. The model for each metabolite was as follows:

\[ y_{gm} = \mu + \sum_{m=1}^{11} \sum_{n=1}^{11} M_{mg} + \sum_{m=1}^{11} \sum_{n=1}^{11} M_{mg} M_{np} + e_{gm}, \]

where \( g = \text{Bay or Sha} \) and \( m = 1, \ldots, 11 \). The main effect of the 11 markers was denoted as \( \mu \). The error, \( e_{gm} \), was assumed to be normally distributed with mean 0 and variance \( \sigma^2 \). An automated SAS script was developed to test all metabolites within this ANOVA and return all P values, type III sums-of-squares for the complete model and each individual term, and QTL pairwise effect estimates in terms of allelic substitution values (see Supplemental Data Sets 9 and 10 online). Pairwise interactions were
defined as having a significant epistatic cluster if they affected >5% of the tested metabolites. Significance values were corrected for multiple testing within a model using FDR < 0.05 in the automated script.

To quantify the epistatic interactions, we adapted an approach utilized to quantify epistatic interactions altering fitness (Elena and Lenski, 1997; Segre et al., 2005). For fitness, nonscaled epistasis is measured as 

\[ \varepsilon = W_{XY} - W_X W_Y \]

where \( W_X \) and \( W_Y \) are the fitness values of the single mutants and \( W_{XY} \) is the fitness of the double mutant. Fitness was measured as the ratio of growth between the mutant genotype and the wild type. We adapted this measure of epistasis to quantify the epistasis for each metabolite’s accumulation as regulated by two QTLs, X and Y, so that 

\[ \varepsilon = M_{XY} - M_X M_Y \]

represented the ratio of a metabolite’s average accumulation in the RILs containing a Sha allele at QTL X and a Bay allele at QTL Y to the average metabolite accumulation in the RILs containing a Bay allele at both QTLs. Conversely, \( M_Y \) was the ratio of a metabolite’s average accumulation in the RILs containing a Bay allele at QTL Y and a Bay allele at QTL X to that metabolite’s average accumulation in the RILs containing a Bay allele at both QTLs. The sign of \( \varepsilon \) was independent of which parental genotype was used in the denominator. Negative values of \( \varepsilon \) indicated negative epistasis, whereby the recombinant progeny had lower metabolite accumulation than the parental genotypes. Positive \( \varepsilon \) showed positive epistasis, whereby the recombinant progeny had greater metabolite accumulation than the parental genotypes (Elena and Lenski, 1997; Segre et al., 2005).

To test the three main TCA cycle QTLs for a three-way epistatic interaction regulating the TCA cycle, we conducted an ANOVA using a model containing all 11 main-effect terms, all 55 pairwise interactions, and a single three-way epistatic term. An automated SAS script then individually tested the eight detected metabolites central to the TCA cycle against the combined model. The script returned all P values, type III sums-of-squares for the complete model and each individual term, and means for the eight genotypic classes in the three-way epistatic interaction. Significance values were corrected for multiple testing within a model using FDR < 0.05 in the automated script. Analyzing all other metabolites with the same model detected no additional metabolites showing this significant three-way interaction.

### HIF195 Analysis


Meyer, R.C., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H.,


