CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

Mapping the *Arabidopsis* Metabolic Landscape by Untargeted Metabolomics at Different Environmental Conditions

Si Wu[1], Takayuki Tohge[1], Álvaro Cuadros-Inostroza[1,2], Hao Tong[1], Hezi Tenenboim[1,2], Rik Kooke[3], Michaël Méret[2], Joost B. Keurentjes[3], Zoran Nikoloski[1], Alisdair R. Fernie[1], Lothar Willmitzer[1] and Yariv Brotman[1,4,*]

[1]Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

[2]MetaSysX GmbH, Am Mühlenberg 11, 14476 Potsdam-Golm, Germany

[3]Laboratory of Genetics, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

[4]Department of Life Sciences, Ben Gurion University of the Negev, Beersheva, Israel

***Correspondence: Yariv Brotman (brotmany@post.bgu.ac.il)**

https://doi.org/10.1016/j.molp.2017.08.012

ABSTRACT

Metabolic genome-wide association studies (mGWAS), whereupon metabolite levels are regarded as traits, can help unravel the genetic basis of metabolic networks. A total of 309 *Arabidopsis* accessions were grown under two independent environmental conditions (control and stress) and subjected to untargeted LC–MS-based metabolomic profiling; levels of the obtained hydrophilic metabolites were used in GWAS. Our two-condition-based GWAS for more than 3000 semi-polar metabolites resulted in the detection of 123 highly resolved metabolite quantitative trait loci ($p \leq$ 1.0E-08), 24.39% of which were environment-specific. Interestingly, differently from natural variation in *Arabidopsis* primary metabolites, which tends to be controlled by a large number of small-effect loci, we found several major large-effect loci alongside a vast number of small-effect loci controlling variation of secondary metabolites. The two-condition-based GWAS was followed by integration with network-derived metabolite–transcript correlations using a time-course stress experiment. Through this integrative approach, we selected 70 key candidate associations between structural genes and metabolites, and experimentally validated eight novel associations, two of them showing differential genetic regulation in the two environments studied. We demonstrate the power of combining large-scale untargeted metabolomics-based GWAS with time-course-derived networks both performed under different abiotic environments for identifying metabolite–gene associations, providing novel global insights into the metabolic landscape of *Arabidopsis*.

Key words: untargeted metabolomics, GWAS, network analysis, different environments, secondary metabolism

Wu S., Tohge T., Cuadros-Inostroza Á., Tong H., Tenenboim H., Kooke R., Méret M., Keurentjes J.B., Nikoloski Z., Fernie A.R., Willmitzer L., and Brotman Y. (2018). Mapping the *Arabidopsis* Metabolic Landscape by Untargeted Metabolomics at Different Environmental Conditions. Mol. Plant. **11**, 118–134.

INTRODUCTION

Plants produce large arsenals of structurally and biologically diverse metabolites. The plant metabolome is often regarded as the terminal downstream product of the genome and, thus, as the bridge between the genotypes and the phenotypes of the plant (Agrawal et al., 2012; Navarova et al., 2012; Prasad et al., 2012; Kerwin et al., 2015). Therefore, unraveling the genetic control underlying metabolite abundance can remarkably enhance our understanding of plant integral regulatory systems for key metabolic traits. Leveraging advances in high-throughput sequencing, genome-wide association study (GWAS) exploits natural genotypic variation and enables the analysis of associations between hundreds of thousands of single-nucleotide polymorphisms (SNPs) and specific phenotypes (Nielsen et al., 2011; Weigel, 2012; Soltis and Kliebenstein, 2015). This powerful genetic resource, along with the development of advanced mass spectrometry (MS) platforms, has made it possible to regard detected metabolites as phenotypic traits for conducting metabolite-based GWAS

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

(mGWAS), resulting in high-resolution maps of genomic regions associated with metabolite variation, namely metabolic quantitative trait loci (mQTL) (Kliebenstein et al., 2002; Chan et al., 2010b, 2011; Riedelsheimer et al., 2012; Li et al., 2013, 2014; Chen et al., 2014; Wen et al., 2014; Matsuda et al., 2015; Wu et al., 2016).

mGWAS has been successfully applied to detect links between targeted secondary metabolites and structural genes, involved in glucosinolate biosynthesis (Kliebenstein et al., 2002; Hansen et al., 2008; Chan et al., 2010b, 2011), flavonoid metabolism (Routaboul et al., 2012; Bac-Molenaar et al., 2015; Ishihara et al., 2016), and the phenylpropanoid pathway (Li et al., 2014) in *Arabidopsis thaliana*, as well as in several crop species (Riedelsheimer et al., 2012; Li et al., 2013; Chen et al., 2014; Sauvage et al., 2014; Wen et al., 2014; Matsuda et al., 2015). Despite the wide application of conventional methods (e.g., transcriptome co-expression and genome-wide similarity analysis) in the identification of novel genes, mGWAS detects key genes involved in naturally occurring limiting steps. For example, mGWAS led to identification of genes encoding two 2-oxoglutarate-dependent dioxygenases (*AOP2* and *AOP3*) that are responsible for natural variation in the presence of methylsulfinylalkyl, alkenyl, and hydroxyalkyl glucosinolates (Kliebenstein et al., 2001). The mapping-based discovery of *BGLU6*, a glycoside hydrolase family 1-type gene involved in flavonoid metabolism, elucidated naturally occurring loss-of-function alleles in some *Arabidopsis* accessions, explaining natural variation in flavonol glycoside accumulation in *Arabidopsis* strains (Ishihara et al., 2016). Moreover, mGWAS has been increasingly applied to wider scope by using untargeted metabolomics in order to uncover new and uncharacterized genes/pathways. Gas chromatography time-of-flight MS (GC–TOF-MS)-based untargeted metabolomics was successfully applied to conduct mGWAS in 96 *A. thaliana* accessions, mainly focusing on primary metabolites, to query the genotypic components controlling the diversity of the *Arabidopsis* metabolome (Chan et al., 2010a). While some mapping studies performed cross-validation for QTL identification in different environments/locations (Korte et al., 2012; Wen et al., 2014), only very few pioneering investigations induced artificial stresses to increase the number of identified new loci (Chan et al., 2011; Davila Olivas et al., 2016). That said, the application of untargeted liquid chromatography–MS (LC–MS)-based metabolomics to *Arabidopsis* GWAS panels, especially concentrating on causal-locus identification in different environments for secondary metabolite levels, remains lacking.

In our previous report (Wu et al., 2016), we were able to provide improved detection of causal genes for 94 primary metabolites in *Arabidopsis* by integrating quantitative genetics with metabolite–transcript correlation-network analysis. Here, applying a novel strategy, we subjected the same collection of 309 *A. thaliana* accessions grown in two distinct environmental conditions (control and stress) to untargeted metabolomics-based GWAS for more than 3000 LC–MS-measured semi-polar metabolites (mainly secondary metabolites). The two-condition-based GWAS resulted in the detection of 123 highly resolved mQTLs, 24.39% of which were environment-specific. Using a statistical framework for five distinct metabolite classes, we demonstrated that the stress GWAS displayed increased accuracy and sensitivity in true-causal-gene discovery, revealing the power of conducting GWAS in different environments. In parallel, metabolite–transcript correlation networks were constructed based on a time-course stress experiment featuring eight different light/temperature conditions. The combination of GWAS and network analysis allowed the identification of 42 key trait–locus associations, leading to 70 candidate genes. Besides well-characterized secondary metabolite–gene associations, we discovered a substantial number of novel associations, part of which were validated by genetic analyses. Our study demonstrates the merit of conducting untargeted metabolomics-based GWAS in multiple different environments as an unbiased approach for uncovering new regulatory genes.

RESULTS

Comprehensive Metabolic Profiling of *Arabidopsis* Accessions under Control and Stress Conditions

The induced biosynthesis of many metabolites solely under stress conditions hinders our efforts to obtain a complete picture of plant metabolic pathways. Considering this, we conducted untargeted metabolomics-based mGWAS in two different environmental conditions (control and stress). Abiotic stress (darkness and 32°C), along with normal condition, were applied in our two-condition-based GWAS. This particular stress was selected from our previous time-course stress experiment (Caldana et al., 2011). In that experiment we observed the highest number of significantly changed secondary metabolites across 23 time points in the darkness and 32°C conditions compared with seven other stress conditions.

Using high-throughput LC–MS analysis, we detected and relatively quantified 4182 and 3968 distinct metabolite features in 309 *Arabidopsis* accessions (Supplemental Table 1) under control and stress conditions, respectively, using positive and negative ionization modes. Most of metabolite features were detected in both control and stress conditions, but 1443 and 1178 were control specific and stress specific, respectively (Supplemental Figure 1). Next, we grouped the metabolite features according to accurate mass difference, retention time, and correlation between metabolite features across all accessions, resulting in 2916 (control) and 2463 (stress) unique metabolites. Chemical structures of 128 metabolites were identified or putatively annotated with information of metabolite identification confidence level (Supplemental Table 2), following a previous publication (Sumner et al., 2007).

Normalized metabolite data identified under the control and stress conditions are provided in Supplemental Table 3. The levels of each metabolite feature varied widely across the natural accessions in control and stress conditions, with a higher proportion of metabolite features showing >10-fold difference under stress than in the control condition (Supplemental Figure 2). Among the 2790 common (detected in both condition) metabolite features, 2148 (77%) of the features displayed broad-sense heritability (H^2) greater than 0.5, while 1618 (58%) had heritability greater than 0.7 (Supplemental Figure 3). On the basis of metabolite-feature levels (peaks in LC–MS datasets), principal component analysis (PCA) clearly separates these natural ecotypes into two different groups in accordance with the two environmental conditions (Supplemental Figure 4).

Genetic Basis Underlying the *Arabidopsis* Metabolome

Using the GWAS, more than 56% of the metabolite features in two different conditions had at least one associated locus at a genome-wide significance level of $p \leq 5.01E\text{-}06$ (LOD ≥ 5.3). In total, 5210 and 5182 distinct trait–locus associations were identified for the control and stress conditions, respectivey, resulting in, on average, around two associated loci for each metabolite feature (Supplemental Table 4). Manhattan plots of the significant loci detected repeatedly are illustrated, including 174 loci corresponding to glucosinolates, flavonoids, phenylpropanoids, amino acids and their derivatives, nucleic acids and their derivatives, and other known metabolites, as well 437 loci corresponding to currently unknown metabolites for control condition; in the stress condition, there are 167 and 443 identified loci corresponding to known and currently unknown metabolites, respectively (Figure 1A).

To obtain robust mQTL, we applied a higher LOD threshold of 8.0 for further investigation, resulting in 123 highly resolved mQTLs. We then compared the mQTLs in the control condition (LOD $\geq$ 8.0) with the ones in the stress condition (LOD $\geq$ 5.3), as well as vice versa (stress at LOD $\geq$ 8.0; control at LOD $\geq$ 5.3). We defined environment-specific mQTLs as those either detected in only one specific condition or those detected in both conditions but with ΔLOD > 3. Using this definition, we detected 13 and 17 control- and stress-specific mQTLs, respectively, totaling 24.39% of all the detected mQTLs in both conditions at a LOD threshold of 8.0 (Figure 1B). The full lists of significant associations between metabolite traits and genes ($p \leq 1.0E\text{-}08$, LOD $\geq$ 8.0) are presented in Supplemental Table 5.

Evaluation of the Performance of GWAS Conducted in Different Environments

We used five distinct metabolite classes (glucosinolates, flavonoids, phenylpropanoids, amino acids, and amines) to test the performance of our GWAS and to demonstrate the power of conducting GWAS in different environments in identifying causal loci. For each metabolite class, we generated two gene lists (''actual gene list'' and ''reference gene list''). The actual gene list refers to a list of genes from our actual GWAS results that each metabolite class mapped to in only control, only stress, and control + stress datasets, using LOD thresholds ranging from 5.3 to 10.0 with an interval of 0.1. The reference gene list stands for the published inventory gene lists (Supplemental Table 6) containing all the experimentally characterized or putatively annotated genes related to the five metabolite classes (Kanehisa and Goto, 2000; Thimm et al., 2004; Fraser et al., 2007; Chan et al., 2011; Saito et al., 2013). By comparing these two gene lists for each tested metabolite class, we assessed the performance of our two-condition GWAS by calculating *precision*, *recall*, and *F-measure*, three widely applied statistics for scoring metrics in pattern recognition and information retrieval (Powers, 2011), and by calculating the significance of overlap using LOD thresholds ranging from 5.3 to 10.0. As shown in Figure 2, the stress GWAS displayed higher *precision*, *recall*, *F-measure*, and more significant enrichment *p* values in comparison with the control GWAS across all the tested LOD thresholds, demonstrating the merit of conducting GWAS in different environments in causal-locus identification with increased accuracy and sensitivity, in comparison with single-condition GWAS. In addition, all the enrichment values for the overlap between the actual and reference gene lists were statistically significant for all five metabolite classes when using the selected LOD threshold of 8.0, with the exception of amino acids in the control condition.

Metabolite–Transcript Correlation-Network Analysis

Transcript data were obtained from our previous time-course stress experiment (Caldana et al., 2011). The detailed information about the experimental setup is provided under ''Time-Course Stress Experiment'' in Methods. In brief, wild-type *A. thaliana* Columbia-0 (Col-0) was exposed to eight environmental conditions differing in light and temperature. Temperature- and light-stress treatments were conducted as follows: aside from the control condition (21°C and 150 μE m^{-2} s^{-1}, abbreviated as 21-L), the plants were exposed to seven different environmental conditions: (i) 4°C and darkness (4-D), (ii) 21°C and darkness (21-D), (iii) 32°C and darkness (32-D), (iv) 4°C and 85 μE m^{-2} s^{-1} (normal light; 4-L), (v) 21°C and 75 μE m^{-2} s^{-1} (low light; 21-LL), (vi) 21°C and 300 μE m^{-2} s^{-1} (high light; 21-HL), and (vii) 32°C and 150 μE m^{-2} s^{-1} (normal light; 32-L). We used the material harvested in this previous experiment to perform an untargeted LC–MS metabolite profiling. Both metabolite and transcript data were used for correlation-network analysis. The detailed process of metabolite–transcript correlation-network analysis, including selection of correlation thresholds, condition-specific network construction, and network quality evaluation, is provided in Supplemental Note 1.

Well-Characterized Genes Were Detected by GWAS and Network Analysis

The integration of the two-condition-based GWAS and the network analysis allowed the identification and refinement of 42 unique key trait–locus associations, in turn giving rise to 70 candidate genes known or putatively annotated as enzymes taking part in metabolic processes (Supplemental Table 12). For these 70 candidate genes, we used 200K SNP data across accessions in comparison with the Col-0 reference genome to detect polymorphic variants causing amino acid sequence change or truncation caused by the introduction of a premature stop codon (Supplemental Table 13).

We briefly list examples of the major loci detected for well-characterized secondary metabolites in Supplemental Note 2, with one typical example, the *OMT1* locus, combining GWAS, network analysis, and metabolite annotation by isotope labeling and MS/MS fragmentation analysis, described in the following (Figure 3).

A metabolite feature (*m/z* 371.0985, retention time = 4.69, negative mode) exhibited high positive correlations with candidate gene *OMT1* (caffeic acid/5-hydroxyferulic acid *O*-methyltransferase, *AT5G54160*) in three darkness-related conditions (Figure 3A). GWAS results indicated that this metabolite trait was strongly associated with the genetic locus on chromosome five harboring *OMT1* (p = 1.51E-15) (Figure 3B). The metabolite was putatively annotated as 5-hydroxyferulic acid glucoside (5HFAG) based on isotope labeling, limiting the possible chemical formula to $C_{16}H_{20}O_{10}$ (Figure 3C and 3D), and based on public database search (KNApSAck, Metabolome.JP, PubChem, and KEGG) and MS/MS fragmentation profiles (Figure 3E). We

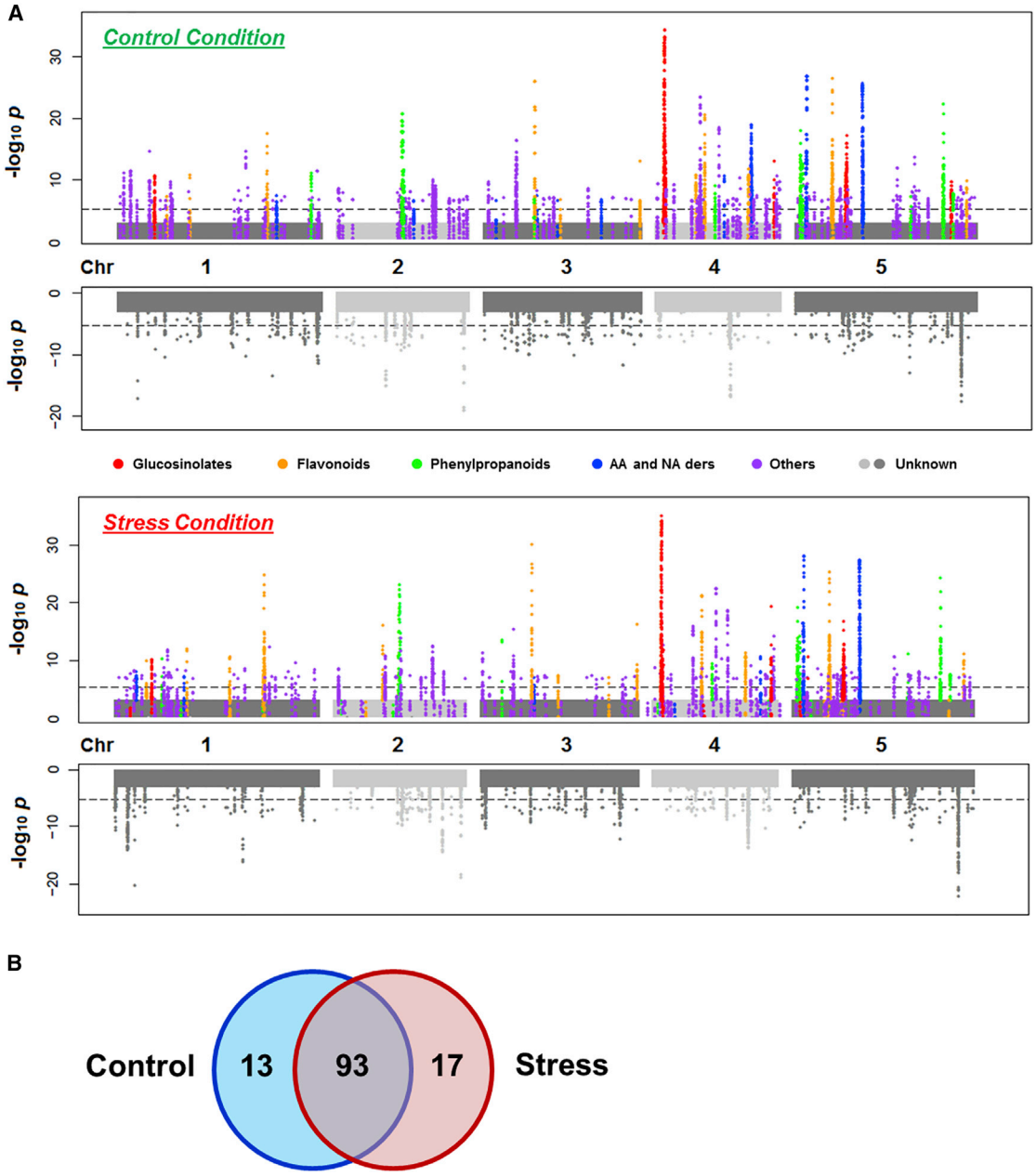

Figure 1. Summary of mGWAS Results Conducted in Two Different Environmental Conditions.

(A) Manhattan plots of mGWAS results with genetic associations at a genome-wide significance level of $p \leq$ 5.01E-06 (LOD $\geq$ 5.3). The strength of association for known (top) and unknown (bottom) metabolites is indicated as the negative logarithm of the *p* values for the compressed mixed linear model in the control (upper panel) and stress (bottom panel) conditions. All metabolite–SNP associations with *p* values below 5.01E-06 (horizontal dashed line in all Manhattan plots) are plotted against genome location in intervals of 1 Mb. AA and NA ders, amino acid and nucleic acid derivatives.

(B) Venn diagram for the comparison of detected loci in GWAS for control and stress conditions using the selected LOD threshold of 8.0 ($p \leq$ 1.0E-08).

further validated the association between 5HFAG and *OMT1* using *omt1* mutants (Figure 3F). 5HFAG was significantly decreased in Col-0 plants under stress condition, mimicking our results in the time-course stress experiment (Figure 3A). In parallel, *OMT1* mutant plants displayed remarkably higher levels of 5HFAG compared with Col-0 plants in both conditions (Figure 3F). Additionally, a different metabolite feature, putatively annotated as coumaric acid glucoside, also mapped to the *OMT1* locus, exhibiting metabolic behaviors similar to that of 5HFAG in *OMT1* mutant plants (Figure 3F). Although *OMT1* had already been characterized by recombinant protein (Muzac et al., 2000) and knockout (KO) (Tohge et al., 2007) analyses, our integrative approach further provides genetic and co-regulation evidence for *OMT1*.

Identification and Experimental Validation of Novel Associations

Aside from the well-characterized secondary metabolites described above, we applied our integrative approach to the

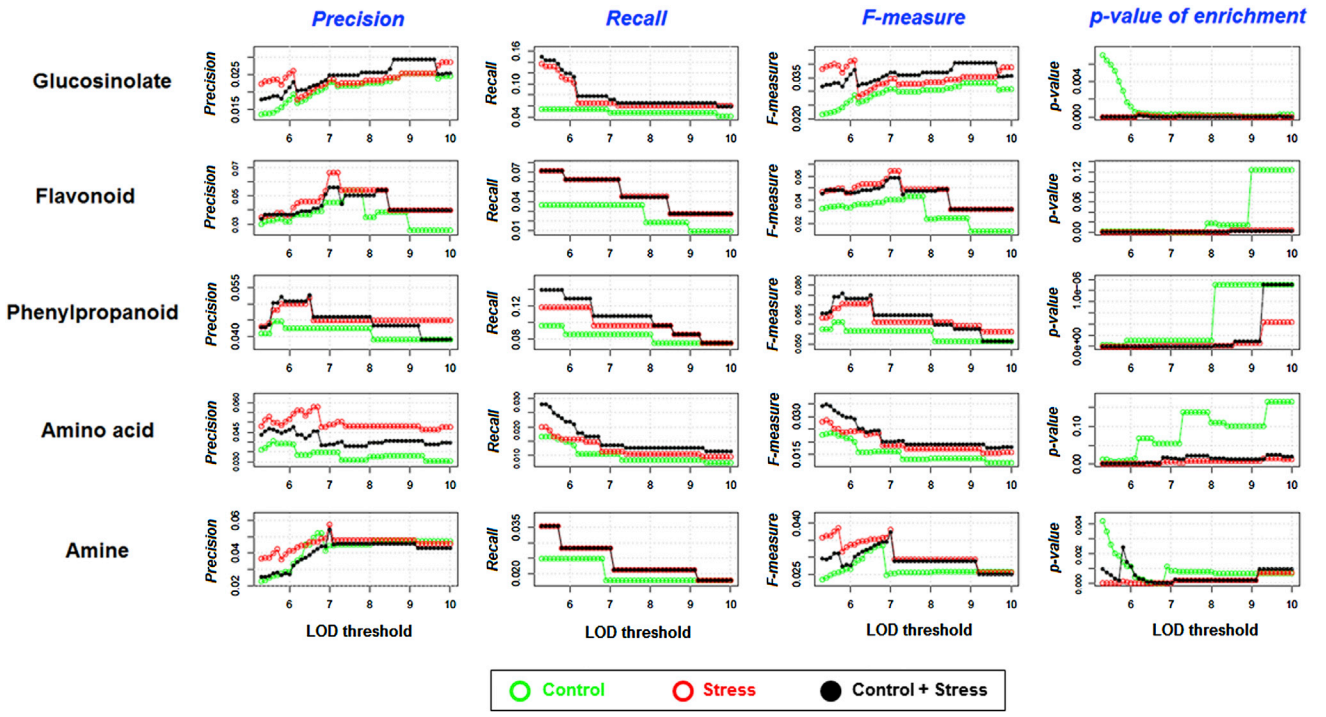

Figure 2. Statistical Analysis to Evaluate GWAS Performance in Different Environments Using Five Distinct Metabolite Classes.
Panels from left to right: *precision*, *recall*, *F-measure*, and *p* value of enrichment analysis.

realm of unidentified metabolites. In doing so we discovered novel associations involved in glycosylation, flavonoid metabolism, and nicotinate metabolism.

AT3G55700, annotated as a UDP-glycosyltransferase superfamily protein (*UGT*), was significantly associated with four unidentified metabolites (metabolites A–D) in the GWAS of both control and stress conditions (Figure 4A). There are four SNP markers in *UGT* (Figure 4B), giving rise to 12 haplotypes, which can be further classified into three main clusters (Figure 4C). Based on haplotype analysis, the levels of metabolite B show significant differences between the three main clusters, exhibiting a gradually increasing trend from cluster I to cluster III (Figure 4C) (metabolite B serves as a representative for the other three metabolites, whereby the other three metabolites also demonstrate significant differences between the three clusters). All four SNPs lead to amino acid substitutions, but only the second SNP (m119016, A/T) clearly distinguishes cluster I from clusters II and III, leading to a substitution from a polar amino acid (serine) to a non-polar one (cysteine). This amino acid substitution results in significant difference in the levels of metabolite B (p = 1.33E-12) (Figure 4D). To further investigate the importance of this residue, we compared the amino acid sequence in *A. thaliana* with that from 16 other species, including three Brassicaceae. We found serine to be conserved in this position in all 16 species, indicating the importance of this amino acid for the enzymatic activity that the candidate gene (*AT3G55700*) encodes. This result also suggests that the mutational event that led to cysteine in some *Arabidopsis* accessions occurred after the formation of the *A. thaliana* species. Next, we used two independent T-DNA insertion lines to validate the association with *UGT*. The levels of metabolite A were significantly increased in the two KO lines, whereas the levels of metabolite B and metabolite C were undetectable, and metabolite D had significantly decreased levels compared with Col-0 plants (Figure 4E). Therefore, we propose a possible reaction scheme in which metabolites A–D may be involved (Figure 4E). Of note, residual levels of metabolite D in the KO lines suggest possible gene redundancy or alternative biosynthetic routes for this metabolite's synthesis. Isotope-labeling results indicate 13 and 12 for metabolite D and metabolites B and C (the latter two sharing the same *m/z*), respectively. The likely chemical formulae for metabolites B and C are $C_{13}H_{24}O_9$ and $C_{13}H_{24}O_9$, and for metabolite D $C_{12}H_{22}O_9$. Additionally, similar MS/MS fragmentation patterns were detected for metabolites B, C, and D (Supplemental Figure 13), suggesting that they share similar structures. Intriguingly, we also observed that the *m/z* difference between metabolite A and metabolites B and C is 162.0534, the characteristic fragment loss of one hexose, usually from one glycoside to an aglycone, further supporting the possible role of *UGT* as a UDP-glycosyltransferase involved in glycosylation.

We additionally observed a metabolite feature (*m/z* 741.2220, retention time = 6.29, positive mode) that mapped to a locus on chromosome one harboring candidate gene *BGLU1* (β-glucosidase 1, *AT1G45191*) with high LOD scores of 8.78 and 10.22 in control and stress conditions, respectively (Figure 5A). Lead SNP m27661 and seven other significantly associated SNPs are located in *BGLU1*. Among the eight SNP markers, three lead to changes in the amino acid sequence. The first and last polymorphism variants (C/G, m27656; C/A, m27663) result in substitutions from non-polar to polar amino acids (Figure 5B). The mapped metabolite is significantly different among the

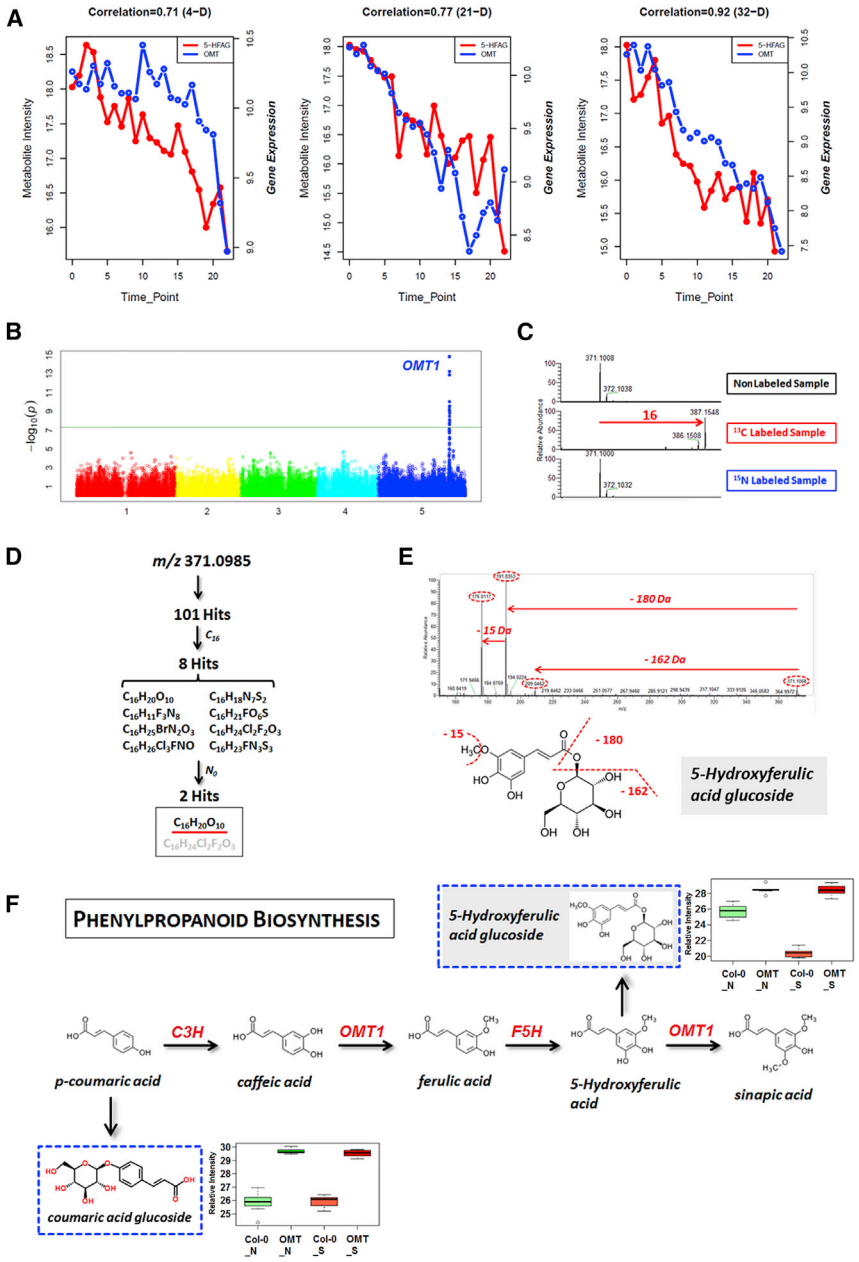

Figure 3. Validation of the Integrative GWAS and Network Analysis for *OMT1* (Caffeic Acid/5-Hydroxyferulic Acid *O*-Methyltransferase, *AT5G54160*).

(A) Detected positive correlations between the unidentified metabolite and the candidate gene *OMT1* in three darkness-related conditions (4-D, 0.71; 21-D, 0.77; 32-D, 0.92).

(B) Manhattan plot displaying GWAS results for the levels of the unidentified metabolite trait in the control condition.

(C) Isotope-labeling results using ^{13}C- and ^{15}N-labeled samples compared with a non-labeled sample. A 16.0540 mass shift and no mass shift are observed comparing ^{13}C- and ^{15}N-labeled samples, respectively, with the non-labeled sample.

(D) The process of narrowing down the scope of possible chemical formulae for the unidentified metabolite based on the isotope-labeling results. Two possible formulae ($C_{16}H_{20}O_{10}$ and $C_{16}H_{24}Cl_2F_2O_3$) fit the constraints derived from the labeling (16 carbons, no nitrogen). Considering the relative rarity of chlorine and fluorine in natural products, the most plausible formula for this unidentified feature was tentatively annotated as $C_{16}H_{20}O_{10}$.

(E) MS/MS fragmentation profiles and putative annotation for the unidentified metabolite. The observed neutral losses of 180 Da and 162 Da indicate a hexose and an aromatic attachment in the metabolite. The fragment ion *m/z* 176.0117 was speculated to be produced by loss of $–CH_3$ from the target metabolite. Based on the above fragmentation patterns, the metabolite was putatively annotated as 5-hydroxyfuric acid glucoside.

(F) The phenylpropanoid pathway, in which the candidate gene *OMT1* and the mapped metabolite traits (5-hydroxyferulic acid glucoside and coumaric acid glucoside) are involved, and the *OMT1* knockout experiment conducted in the control (N) and stress (S; 32°C + darkness) conditions. The enzymes and their abbreviations are *p*-coumarate 3-hydroxylase (*C3H*), caffeate *O*-methyltransferase 1 (*OMT1*), and ferulate 5-hydroxylase (*F5H*).

holotypic forms (Figure 5C). Furthermore, its levels decrease in *bglu1* (p = 6.55E-03; Figure 5D). This unidentified metabolite feature shared the same *m/z* with the well-characterized flavonol glycoside kaempferol 3-*O*-[2″-*O*-(rhamnosyl) glucoside]-7-*O*-rhamnoside (Tohge et al., 2005), but was considerably lower in abundance. The two metabolites elute at different retention times and exhibit different MS/MS fragmentation patterns, although they share the same characteristic fragment ion 287, representing the existence of a kaempferol backbone as aglycone. Moreover, the consecutive observed losses of 146 Da, 146 Da, and 162 Da indicate the attachment of two deoxy-hexoses and one hexose to the kaempferol skeleton of the mapped metabolite (Figure 5E). This unknown metabolite feature also co-mapped to the well-characterized locus *UGT78D1*, catalyzing the transfer of UDP-rhamnose to the 3-OH position of kaempferol and quercetin (Jones et al., 2003). This co-mapping observation further supported the attachment of deoxy-hexoses and hexose to the kaempferol backbone. Although the final identification of this unknown metabolite for its sugar types and exact positions of sugars is yet to be achieved, this example demonstrates the likely presence of the unidentified minor flavonol glycoside, and that *BGLU1* is likely involved in the flavonol glycosylation pathway.

Three unknown metabolite features (named metabolites 1–3) were significantly associated with two related candidate genes, namely *BGLU7* (β-glucosidase 7, *AT3G62740*) and *BGLU8* (β-glucosidase 8, *AT3G62750*) (Supplemental Figure 14A). Besides *BGLU7* and *BGLU8*, metabolites 1 and 2 additionally co-mapped to the well-studied *UGT78D1* locus, a UDP-rhamnose transferase active at the 3-OH position of kaempferol and quercetin (Jones et al., 2003). With differentiating genetic

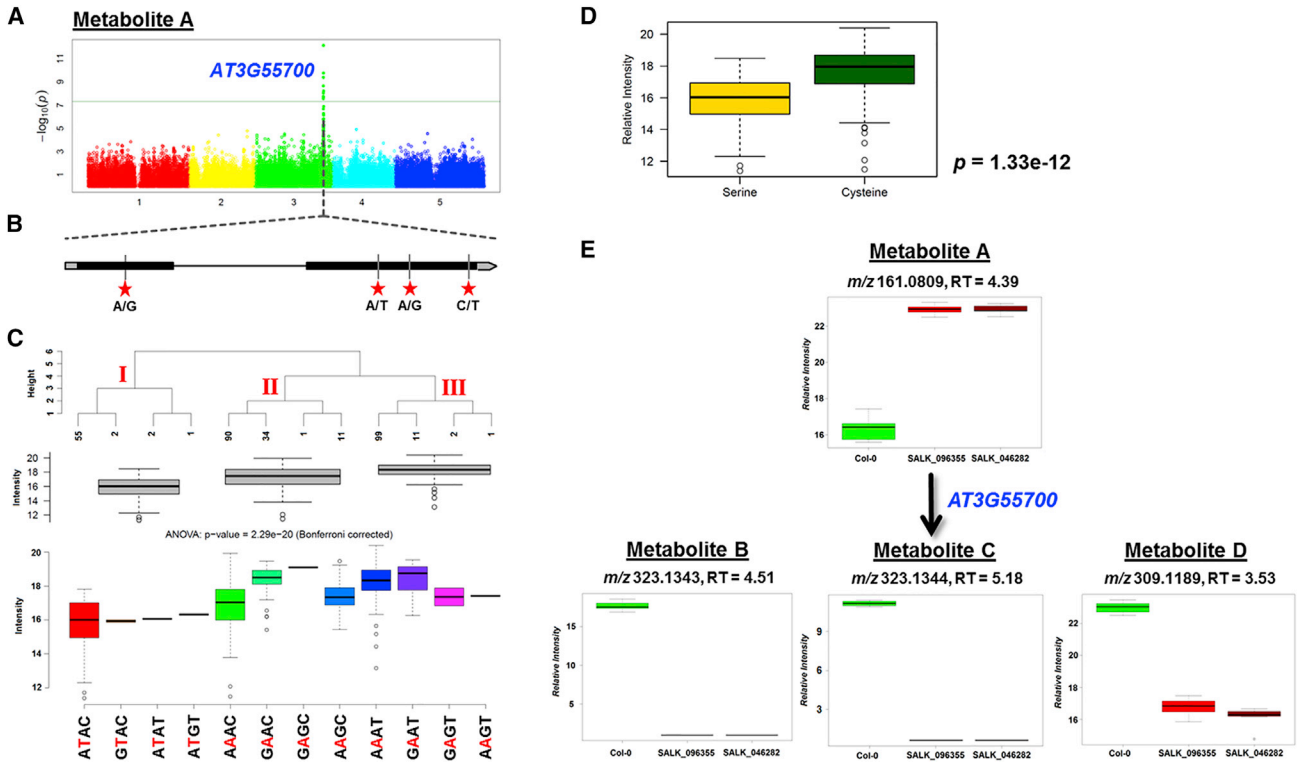

Figure 4. Functional Identification of the Candidate Associations between *UGT* (UDP-Glycosyltransferase, *AT3G55700*) and Four Unknown Metabolites (Metabolites A–D).

(A) Manhattan plot of metabolite A in the stress condition as the representative for the other three co-mapped metabolite traits (metabolites B–D). *p* Values are shown on a $\log_{10}$ scale; the *x* axis shows the physical positions on five chromosomes in *A. thaliana*. *p* = 3.36E-10, 6.39E-08, 4.59E-11, and 4.24E-07 in the control condition, and *p* = 7.23E-13, 7.20E-09, 5.16E-08, and 4.56E-07 in the stress condition for the four metabolites, respectively.

(B) Gene model of *AT3G55700*. Filled black boxes represent coding sequence. The light-gray vertical lines mark polymorphic sites identified by high-throughput genotyping; stars represent the SNP markers resulting in changed amino acid sequence.

(C) Haplotype analysis for four SNPs genotyped in *UGT*. Haplotypes were clustered to three main groups according to their sequence similarities based on Ward's minimum variance method (upper panel). Box plots show the intensity of metabolite B for these three different clusters (middle panel; box width represents number of accessions in each cluster) and for the various haplotypes (bottom panel). One-way ANOVA was applied to detect differences between cluster means, followed by Bonferroni's correction for multiple comparisons (*p* = 2.29E-20).

(D) The levels of metabolite B are significantly different between the accessions divided by the different amino acid residues (serine and cysteine) resulting from the proposed functional site (the second SNP marker in *UGT*, m119016, A/T) (*p* = 1.33E-12).

(E) Proposed chemical scheme that the four mapped metabolite traits (metabolites A–D) are involved in and the changed levels of metabolites A–D in the loss-of-function mutant experiment using two independent KO lines (SALK_096355 and SALK_046282).

regulations, Metabolite 3 co-mapped to the *UGT79B2* and *UGT79B3* locus, recently reported as UDP-rhamnose transferases using cyanidin and cyanidin 3-*O*-glucoside as acceptors (Li et al., 2016). We observed concomitant metabolic changes for these three traits in three independent KO lines for *BGLU7* and *BGLU8* (Supplemental Figure 14B). In addition, we noted that metabolites 1 and 2 shared the same *m/z* with the well-characterized kaempferol 3-*O*-[2″-*O*-(rhamnosyl) glucoside]-7-*O*-rhamnoside (K3RG7R) (Tohge et al., 2005). Taken together, the above evidence led us to predict two different reaction schemes in which these three associations may be involved (Supplemental Figure 14B).

A strong link between an unknown metabolite trait (*m/z* 256.0810, retention time = 1.05) and the candidate gene *GC1* (guanylyl cyclase 1 [Wong and Gehring, 2013], *AT5G05930*) was supported by both GWAS (Supplemental Figure 15A) and network analysis in 21-D and 32-D conditions (Supplemental Figure 15B). In *GC1*, two SNPs (m164251, G/T, lead SNP; m164253, C/A) result in an altered protein amino acid sequence (Supplemental Figure 15C). Together with the results from linkage disequilibrium (LD) (Supplemental Figure 15D) and haplotype (Supplemental Figure 15E) analyses, this finding suggests that these polymorphic variants are likely to constitute the functional variation underlying this association. Isotope-labeling results suggested the possible chemical formula $C_{11}H_{13}NO_6$ (Supplemental Figure 15F). Assaying the standard compound nicotinate D-ribonucleoside revealed that it shares the same MS/MS fragmentation pattern as our metabolite (Supplemental Figure 15G), albeit with a slight retention-time shift, suggesting that our metabolite is structurally highly similar to nicotinate D-ribonucleoside.

Besides the novel associations described above, we also detected two additional associations involved in tyrosine degradation and glucosinolate biosynthesis derived from branched-chain amino acids. Using mutant analysis, we provide new insights into these two well-studied pathways. A detailed

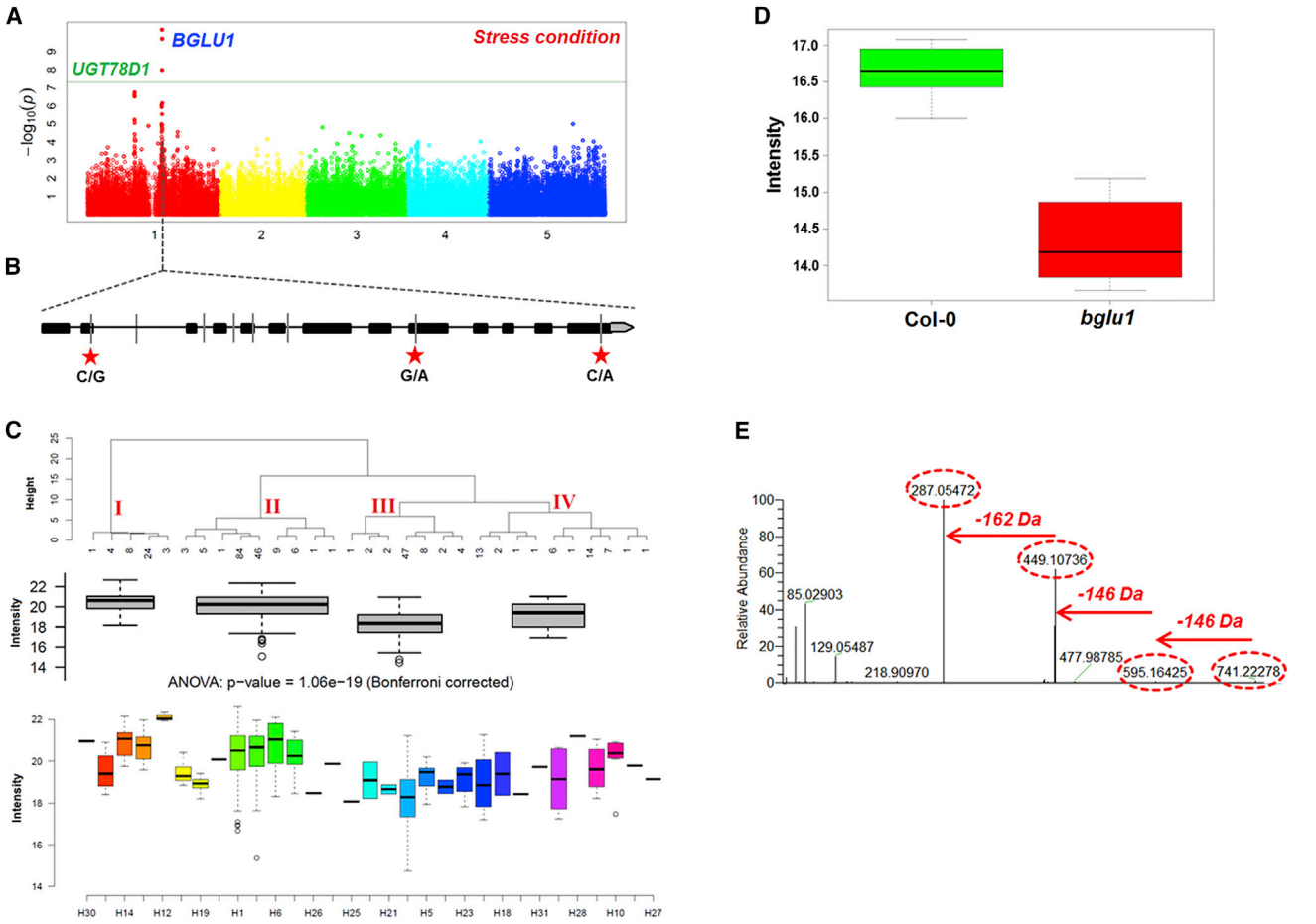

Figure 5. An Exemplary Association Found by GWAS between *BGLU1* and the Metabolite Trait Proposed as a Flavonol Glycoside.
(A) Manhattan plot for the unidentified metabolite trait and the significant association signals in the stress condition (p = 5.98E-11).
(B) Gene model of *BGLU1*, stars represent the SNP markers resulting in changed amino acid sequence.
(C) Haplotype analysis for eight SNPs genotyped in candidate gene *BGLU1*. Metabolite-trait levels are significantly different among the three main clusters based on haplotype sequence similarity (p = 1.06E-19).
(D) Metabolite changes for the unidentified metabolite trait in the KO experiment (p = 6.55E-03).
(E) MS/MS fragmentation analysis for the metabolite trait. The characteristic fragment ion 287 represents the existence of a kaempferol backbone as aglycone. The consecutive losses of 146 Da, 146 Da, and 162 Da indicate the attachment of two deoxy-hexoses and one hexose on the kaempferol skeleton.

analysis of these two associations is described in Supplemental Note 3.

Gene–Environment Interplay Leads to Discovery of Novel Associations

We detected some metabolite–locus associations specific to one of the GWAS conditions tested (Figure 1B), and highlighted several of such cases in the following subsections.

Saccharopine Specifically Accumulates due to the Activation of LKR/SDH in Darkness

We detected a strong association between saccharopine (confirmed by authentic standard) and the *LKR/SDH* locus (saccharopine dehydrogenase; *AT4G33150*), involved in lysine degradation (Zhu et al., 2001; Serrano et al., 2012), only in the stress condition (Figure 6A). In our time-course stress experiment, saccharopine exclusively accumulated in the low-light- or darkness-related conditions (21-D, 21-LL, and 32-D), but was absent in other conditions (Supplemental Figure 18), which accounts for the lack of mapping results for saccharopine in the control condition. Lead SNP m156605 is in a high LD with SNP m156606 (r^2 > 0.90, p < 0.0001) (Figure 6B), leading to an amino acid substitution from phenylalanine to leucine. We used *lkr/sdh* mutants to investigate whether the function of *LKR/SDH* is specifically triggered under stress. Saccharopine was not detected in the control condition (Figure 6C). In the stress condition, saccharopine was notably increased in Col-0 plants, consistent with our time-course stress results; in parallel, significantly lower levels were detected in the *lkr/sdh* mutant (p = 7.89E-05) (Figure 6C). The expression profiles of *LKR/SDH* in the time-course stress experiment show increasing trends in 21-D and 32-D conditions (Supplemental Figure 19). This finding suggests that the route of lysine degradation in which *LKR/SDH* is involved is a non-essential pathway under standard conditions, in line with the previous publication (Zhu et al., 2001), but may be activated under stress conditions (Figure 6D). However, saccharopine was not completely depleted in the *lkr/sdh* KO (Figure 6C), which may indicate gene redundancy or alternative pathways that can also produce this metabolite.

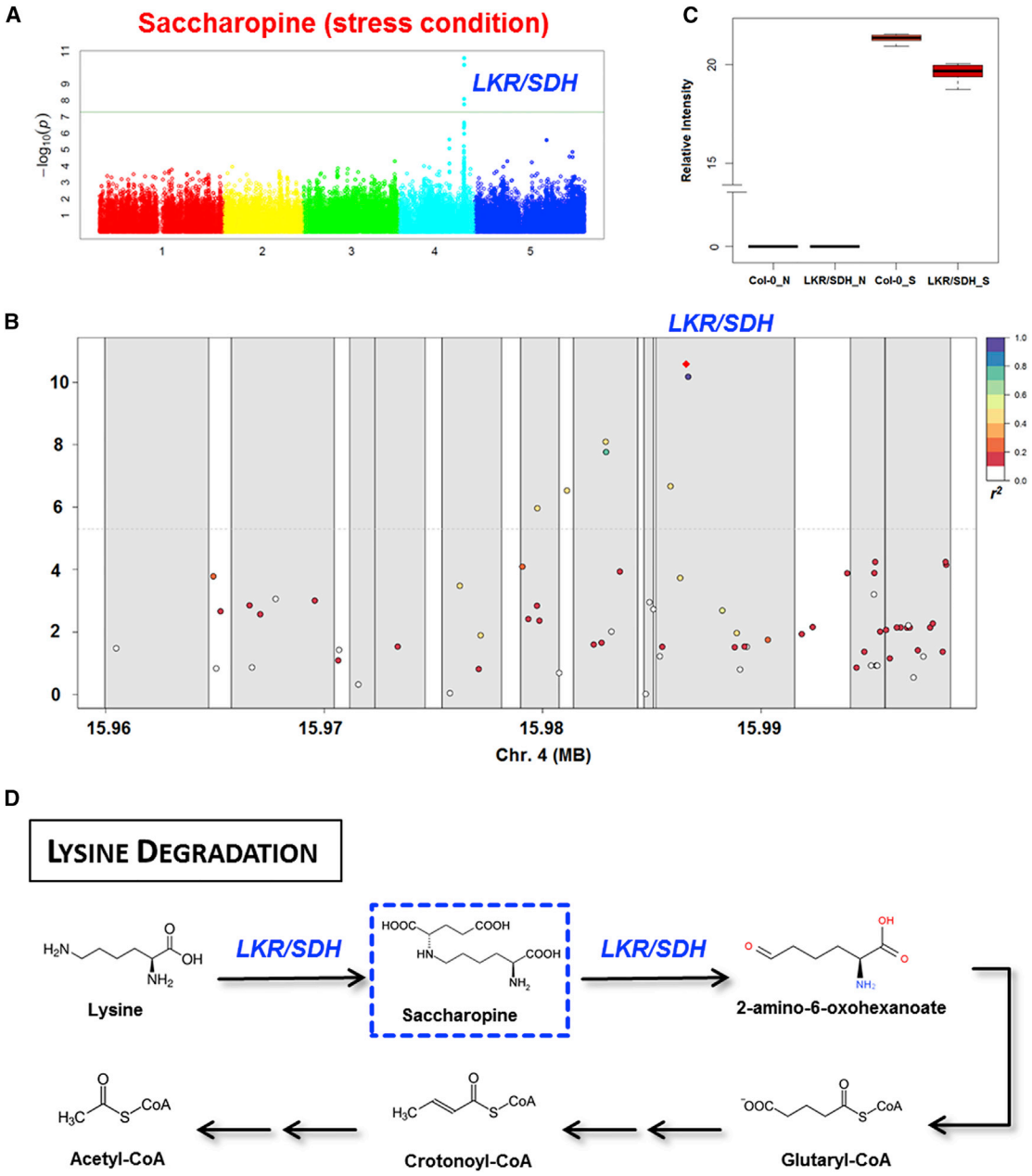

Figure 6. A Representative Association Involved in Lysine Catabolism Discovered by GWAS Conducted in Two Different Environments.

(A) Manhattan plot for the metabolite trait saccharopine in the stress condition (p = 2.60E-11).

(B) Linkage disequilibrium (LD) plot for the locus associated with saccharopine levels on a zoomed-in Manhattan plot. The x axis shows the physical positions in this LD block on chromosome 4 and the y axis shows the significance levels with p values on a $\log_{10}$ scale. Each gray block denotes a gene in the locus to which the unknown metabolite trait mapped. Each dot serves as one SNP marker, with the lead SNP (with highest LOD) shown as a red diamond. Imputation revealed several closely located SNPs in strong LD (r^2) with the lead SNP.

(C) KO experiment with *lkr/sdh* mutants conducted in the control (21°C light) and stress (32°C + darkness) conditions. Saccharopine lacked completely in Col-0 and *lkr/sdh* mutants in the control condition. In the stress condition, saccharopine accumulated in Col-0 plants, but was significantly decreased in *lkr/sdh* mutant plants compared with Col-0 plants (p = 7.89E-05).

(D) The lysine degradation pathway in which the candidate association between saccharopine and *LKR/SDH* is involved.

An Improved Mapping Signal in Stress Condition Leads to the Discovery of an Adenylosuccinate Lyase

Differential levels of significance for an association between succinoadenosine and *AT4G18440* (putatively annotated as an adenylosuccinate lyase) were detected in the control and stress conditions (control, p = 2.29E-06; stress, p = 1.70E-10) (Figure 7A). Haplotype analysis showed that accessions sharing haplotypes "GCAG" and "GGAG" featured significantly lower levels of succinoadenosine than accessions of the other four haplotypes (p = 5.17E-07) (Figure 7B). The mapped metabolite

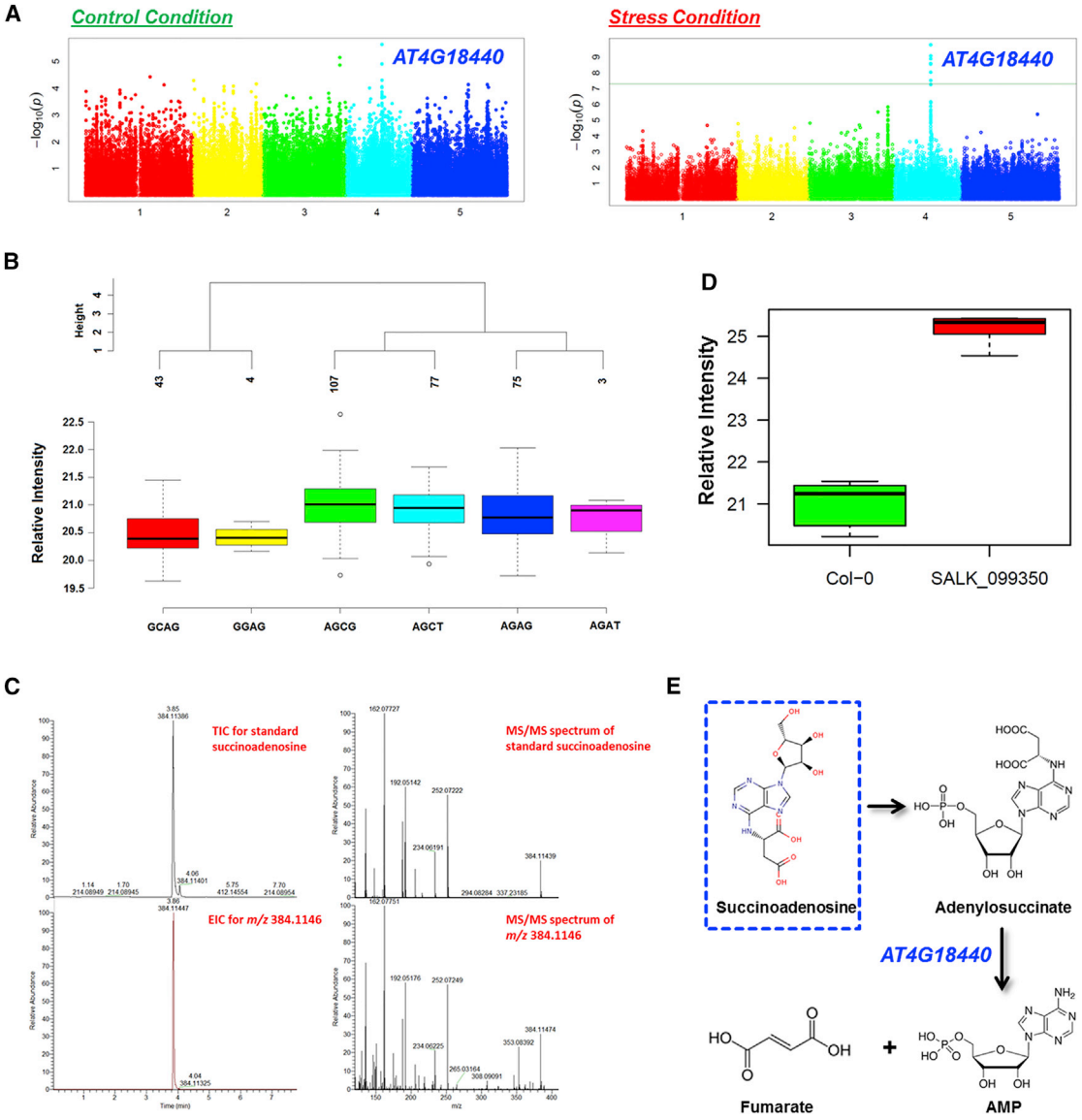

Figure 7. An Improved Mapping Signal in Stress Condition Leads to the Discovery of an Adenylosuccinate Lyase.

(A) The metabolite trait succinoadenosine has different genetic regulations in the control and stress conditions shown by the Manhattan plots (control, p = 2.29E-06; stress, p = 1.68E-10).

(B) Haplotype analysis for four SNPs genotyped in candidate gene AT4G18440. The levels of succinoadenosine differ significantly among the six groups of accessions distinguished by the different haplotypes.

(C) The annotation of the metabolite trait as succinoadenosine was confirmed with authentic standard by comparing the retention time and MS/MS fragmentation pattern.

(D) KO plants (SALK_099350) for candidate gene *AT4G18440* show significant increased levels of succinoadenosine in comparison with Col-0 plants (p = 5.34E-10).

(E) The purine nucleoside pathway in which the candidate association between succinoadenosine and *AT4G18440* is involved.

was confirmed as succinoadenosine based on accurate mass, labeling results, and MS/MS fragmentation analysis, as well as by comparison with a standard compound (Figure 7C). Furthermore, we confirmed the association by metabolic profiling of a KO line, exhibiting significantly higher succinoadenosine levels (p = 5.34E-10) (Figure 7D). Based on our results and the gene's putative annotation, we suggest that *AT4G18440* is involved in purine nucleotide metabolism, catalyzing the conversion of adenylosuccinate to AMP rather than acting directly on succinoadenosine (Figure 7E).

DISCUSSION

Recent years have witnessed the successful application of mGWAS to *Arabidopsis* (Kliebenstein et al., 2002; Keurentjes et al., 2006; Hansen et al., 2008; Chan et al., 2010a, 2010b, 2011; Routaboul et al., 2012; Angelovici et al., 2013; Li et al., 2014; Bac-Molenaar et al., 2015; Ishihara et al., 2016; Wu et al., 2016) and crop species (Riedelsheimer et al., 2012; Li et al., 2013; Chen et al., 2014; Sauvage et al., 2014; Wen et al., 2014; Matsuda et al., 2015). However, most of these studies,

especially those in *Arabidopsis*, largely focused on one or several classes of target metabolites. In this study, we applied a comprehensive untargeted metabolomics-based GWAS approach with high coverage, sensitivity, and accuracy to simultaneously detect thousands of semi-polar metabolite features, mainly focusing on secondary metabolites, in a collection of 309 *A. thaliana* accessions, grown under two different environmental conditions, thus facilitating the discovery of 123 robust mQTLs with the selected LOD threshold of 8.0 ($p \leq$ 1.0E-08) (Figure 1B). Interestingly, different from our previous published observation that natural variation in *Arabidopsis* primary metabolites tends to be controlled by a large number of small-effect loci (Wu et al., 2016), here we found several major large-effect loci alongside a vast number of small-effect loci for variation of secondary metabolites. This observation is highly consistent with previous results (Chan et al., 2011) and is likely caused in part by the more linear/hierarchical genetic architecture of some secondary metabolites in comparison with the intricate and multi-layered regulation inherent in primary metabolism (Chen et al., 2014). However, it may also reflect the highly enriched diversity of secondary metabolites, both qualitatively and quantitatively (Rowe et al., 2008; Keurentjes, 2009; Chan et al., 2010a; Joseph et al., 2013).

Plants facing environmental challenges undergo remarkable reprogramming of their transcriptomes and metabolomes. Several pioneering GWAS have taken different environments into account (Chan et al., 2011; Korte et al., 2012; Davila Olivas et al., 2016). Chan et al. (2010a), focusing mainly on primary metabolites, successfully applied untargeted GC–MS-based metabolomics on a GWAS panel to investigate the influence of environmental variation on genetic architecture. Therefore, the application of a systems-based approach in this vein to secondary metabolites for the improvement of causal-locus identification in a non-biased manner remains lacking. Given that a vast number of secondary metabolites are only produced under conditions of (a)biotic stress, it follows that comparing the genetic regulation of plant metabolism in different environments can accelerate the discovery of the novel metabolic pathways and signaling mechanisms that are involved in the complicated gene-by-environment interactions tailoring the plants' adaptive responses to stresses.

The two-condition-based GWAS allowed us to obtain a total of 123 robust mQTLs (LOD $\geq$ 8.0), 24.39% of which were specific to one of the environmental conditions (mQTL only detected in one specific condition; or detected in both conditions but with ΔLOD > 3). Furthermore, our statistical analysis (Figure 2) for five different metabolite classes demonstrated that conducting GWAS in different environments strongly enhanced the discovery of causal genes with increased accuracy and sensitivity, in comparison with single-condition GWAS. Of note, not every secondary metabolite can be induced under specific growth conditions; thus our study illustrates the importance of exploring the influence of different conditions (e.g., stresses, tissues, development stages) on plant natural variation, providing multiple dimensions of biological insight.

In parallel, the integration of additional forms of genome-scale data derived from our time-resolved stress experiment was applied to directly detect metabolite–gene correlations. The subsequent integration of the two-condition-based GWAS and the independent network analysis using different environments allowed us to provide a global landscape of *Arabidopsis* metabolism in three dimensions: genome, metabotype, and environments (Figure 8).

Our strategy facilitates the discovery of novel and underexplored candidate associations. We found in this study several novel associations related to glycosylation. Glycosylation enhances the solubility of aglycones, and thus is likely to be essential for their synthesis, transport, and storage in their final destination in the vacuole or cell wall (Ishihara et al., 2016). The widest use of glycosylation identified in *Arabidopsis* so far is involved in flavonoid modification by flavonol glycosyltransferase (GT) (Yonekura-Sakakibara and Hanada, 2011; Hectors et al., 2014). So far several *Arabidopsis* GTs have been identified, largely by transcriptome co-expression networks, then functionally characterized by loss-of-function mutants or recombinant-protein assays (Tohge et al., 2005; Yonekura-Sakakibara et al., 2007, 2008, 2012). Some of them were also detected in our GWAS: a flavonol 3-*O*-rhamnosyltransferase (*UGT78D1*), a flavonoid 3-*O*-glucosyltransferase (*UGT78D2*), a flavonol 3-*O*-glucoside, 6″-*O*-glucosyltransferase (*BGLU6*), and UDP-glycosyltransferases acting on cyanidin and cyanidin 3-*O*-glucoside (*UGT79B2* and *UGT79B3*). In addition to the aforementioned genes, we discovered several hitherto non-reported associations involved in glycosylation: the candidate gene *AT3G55700* with four unknown metabolites, and *BGLU1*, *BGLU7*, and *BGLU8* with minor flavonol glycosides. Metabolite identification is one of the bottlenecks in untargeted metabolomic studies. Although the state-of-the-art isotope-labeling experiment and MS/MS fragmentation analysis we applied in this study have provided valuable insights into possible formulae or structures for these unknown metabolites, further structural confirmation regarding the sugar donors and the exact sugar positions in the glycosides needs to be undertaken. Of note, our genetic evidence and the loss-of-function mutant results indicate the likely presence of unidentified minor (flavonol) glycosides. It seems highly likely that these related candidate genes encode the decoration enzymes in their formation (Figures 4 and 5). Our findings thus lead to a more complete understanding of the glycosylation process in *Arabidopsis*, facilitating the reconstruction of biosynthetic pathways, which may in turn benefit metabolic engineering of nutritionally important compounds in plants.

The association between saccharopine, a product of lysine degradation, and *LKR/SDH* was only identified by applying GWAS in the stress environment. In agreement with our time-course stress experiment, we observed significantly higher lysine levels in the 21-D and 32-D stress conditions (Supplemental Figure 20), whereas saccharopine was exclusively accumulated in 21-LL, 21-D, and 32-D stress conditions but was absent in all other conditions, including 21-L (Supplemental Figure 18). Moreover, *LKR/SDH* expression showed enhanced levels in the 21-D and 32-D conditions (Supplemental Figure 19). Our KO results (Figure 6C) proved the association between saccharopine and *LKR/SDH* as reported previously (Zhu et al., 2001), and further dissected the environmental regulation of the gene activity. It is well documented that lysine degradation in plants serves as a stress-associated metabolic pathway (Galili et al., 2001). This example not only to a large extent increases our understanding of the vital role of lysine degradation related

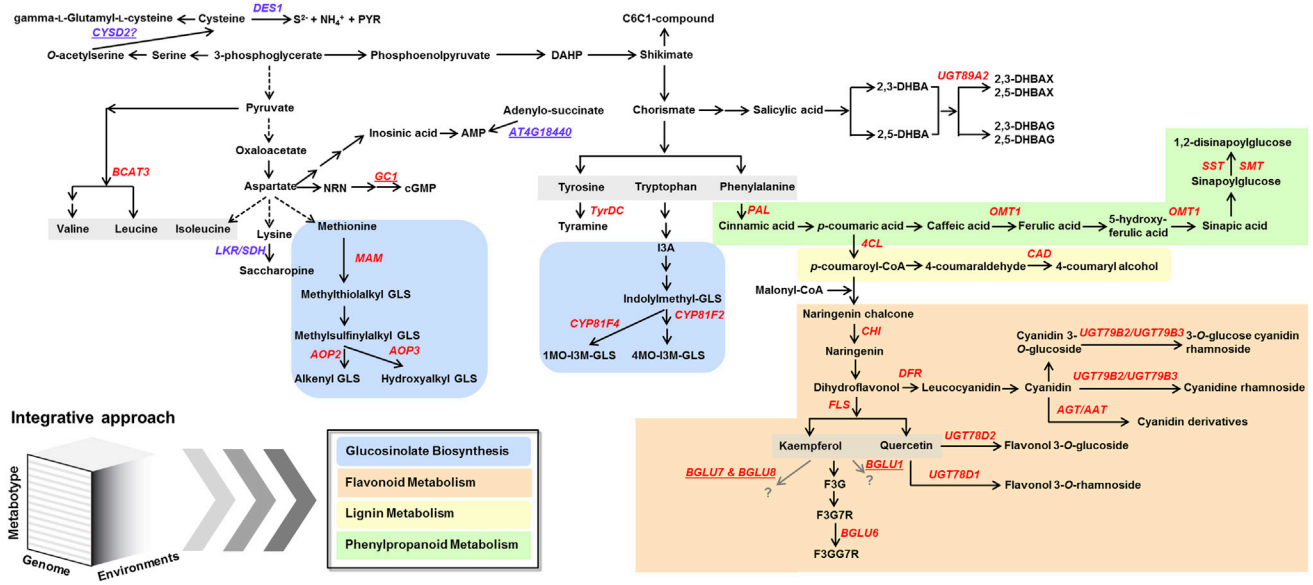

Figure 8. A Global *Arabidopsis* Metabolic Landscape at Three-Dimensional View (Genome, Metabotype, and Environments) Based on an Integrative Approach Combining Untargeted Metabolomics-Based GWAS and Network Analysis.

In red are key candidate genes that were detected in both control and stress conditions; in purple are key candidate genes that were detected only in one condition. Without underline: key candidate genes that were previously characterized and also detected in this study; underlined: previously uncharacterized genes first described in this study. The putatively proposed reactions are in gray. Metabolites are abbreviated as follows: DAHP, 3-deoxy-D-arabino-heptulosonic acid 7-phosphate; 2,3-DHBA, 2,3-dihydroxybenzoic acid; 2,5-DHBA, 2,5-dihydroxybenzoic acid; 2,3-DHBAX, 2,3-dihydroxybenzoic acid 5-*O*-β-D-xyloside; 2,5-DHBAX, 2,5-dihydroxybenzoic acid 5-*O*-β-D-xyloside; 2,3-DHBAG, 2,3-dihydroxybenzoic acid 5-*O*-β-D-glucoside; 2,5-DHBAG, 2,5-dihydroxybenzoic acid 5-*O*-β-D-glucoside; NRN, nicotinate D-ribonucleoside; GLS, glucosinolates; 1MO-I3M-GLS, 1-methoxy-3-indolylmethyl-glucosinolate; 4MO-I3M, 4-methoxy-3-indolylmethyl-glucosinolate; F3G, flavonol 3-*O*-glucoside; F3G7R, flavonol 3-*O*-glucosyl-7-*O*-rhamnoside; F3GG7R, flavonol 3-*O*-glucosyl-glucoside 7-*O*-rhamnoside; K3RRG, kaempferin 3-*O*-rhamnosyl-rhamnosyl-glucoside. Genes are abbreviated as follows: *DES1*, L-cysteine desulfhydrase 1; *CYSD2*, cysteine synthase D2; *BCAT*, branch-chain aminotransferase; *LKR/SDH*, lysine-ketoglutarate reductase/saccharopine dehydrogenase; *GC1*, guanylyl cyclase 1; *MAM*, methylthioalkylmalate synthase; *AOP2*, alkenyl hydro alkyl producing 2; *AOP3*, alkenyl hydro alkyl producing 3; *TyrDC*, tyrosine decarboxylase; *CYP81F4*, cytochrome P450, family 81, subfamily F, polypeptide 4; *CYP81F2*, cytochrome P450, family 81, subfamily F, polypeptide 2; *PAL*, phenylalanine ammonia-lyase; *UGT89A2*, UDP-glycosyltransferase 89A2; *OMT1*, caffeate *O*-methyltransferase 1; *SST*, sinapoyl-Glc:sinapoyl-Glc sinapoyltransferase; *SMT*, sinapoyl-Glc:malate sinapoyltransferase; *CAD*, cinnamyl alcohol dehydrogenase; *4CL*, 4-coumarate-CoA ligase; *CHI*, chalcone isomerase; *DFR*, dihydroflavonol reductase; *FLS*, flavonol synthase; *UGT79B2*, UDP-glycosyltransferase 79B2; *UGT79B3*, UDP-glycosyltransferase 79B3; *UGT78D2*, UDP-glycosyltransferase 78D2; *UGT78D1*, UDP-glycosyltransferase 78D1; *BGLU6*, β-glucosidase 6; *BGLU1*, β-glucosidase 1; *BGLU7*, β-glucosidase 7; *BGLU8*, β-glucosidase 8.

with plants' adaptive responses to the stresses, but also manifests the merit of conducting GWAS in different environments.

In summary, we present here an untargeted metabolomics-based GWAS conducted under two different environmental conditions. Our two-condition-based GWAS for more than 3000 semi-polar metabolites (mainly secondary metabolites) resulted in the detection of 123 highly resolved mQTLs, 24.39% of which were environment specific, with a LOD threshold of 8.0. The statistical analysis demonstrated that conducting GWAS in different environments largely boosted the discovery of causal genes with enhanced accuracy and sensitivity. In parallel, metabolite–transcript correlation-network analysis was applied as a further filter to cross-validate with the two-condition-based GWAS. The combined approach promotes the selection of candidate associations and provides functional and biological insights into the holistic landscape of *Arabidopsis* metabolism, mainly focusing on secondary metabolism. Based on the integrative approach, we screened 42 key trait–locus associations, resulting in 70 candidate genes, not only including the ones involved in well-characterized secondary metabolite biosynthesis but also shedding light on novel, previously uncharacterized connections. Using mutant plants we validated eight of the novel associations, including two with differential genetic regulations, involved in lysine degradation and purine nucleotide metabolism, detected in two GWAS environments. To the best of our knowledge, this is the first report to implement an untargeted metabolomics-based GWAS, as well as network analysis of a time-course experiment, both conducted in different abiotic environments, to comprehensively select and prioritize candidate associations in the realm of global *Arabidopsis* secondary metabolism. In addition, we were able to provide a considerably broader GWAS of *Arabidopsis* secondary metabolism than has been presented to date and, hence, to compare and contrast the genetic architecture of constitutive "housekeeping" and inducible "defense" metabolites.

METHODS

Plant Materials

Time-Course Stress Experiment

Time-resolved stress experiments using different light and temperature conditions were conducted in a previous study (Caldana et al., 2011). In

brief, wild-type *A. thaliana* Col-0 was grown in soil (potting compost) in short days (8 h light) for 4 weeks, then transferred to long days (16 h light) at day/night temperature of 21°C/18°C for 2 weeks. Temperature- and light-stress treatments were conducted as follows: aside from the control condition (21°C and 150 μE m^{-2} s^{-1}, abbreviated as 21-L), the plants were exposed to seven different environmental conditions: (i) 4°C and darkness (4-D), (ii) 21°C and darkness (21-D), (iii) 32°C and darkness (32-D), (iv) 4°C and 85 μE m^{-2} s^{-1} (normal light; 4-L), (v) 21°C and 75 μE m^{-2} s^{-1} (low light; 21-LL), (vi) 21°C and 300 μE m^{-2} s^{-1} (high light; 21-HL), and (vii) 32°C and 150 μE m^{-2} s^{-1} (normal light; 32-L). It should be noted that a reduced light intensity of 85 μE m^{-2} s^{-1} was used in conjunction with the 4°C treatment to prevent a secondary stress caused by excess light (Bieniawska et al., 2008). The 4°C condition can therefore not be regarded as merely different in temperature compared with the 21-L or 32-L conditions.

Plant material was sampled at 20-min intervals for a total of 360 min to yield a 19 data-point linear series (including 0 min). Additional samples were taken after 5, 10, 640, and 1280 min to obtain 10 data points (including 0 min) in a logarithmic time series. For each condition and each time point, three independent plants were sampled and analyzed for metabolites and transcripts.

Natural Population and Stress Experiment

A previously described diverse collection of 314 natural *A. thaliana* accessions was used to measure secondary metabolites for GWAS with existing SNP data (Li et al., 2010; Horton et al., 2012). Seeds were sown directly to soil in 6-cm pots for each ecotype, and stratified at long-day with cold-night condition (16 h LD, 250 μE m^{-2} s^{-1} at day/night temperature of 20°C/6°C and humidity 60%/75%). After 2 weeks, the seedlings were pricked to separate pots with six replicates for each accession. The plants then grew in short days (8 h light) for another 2 weeks. Climate in the culture room was converted to long-day condition (16 h light) for the next 2 weeks. Plants were placed randomly to avoid block effects during growth. All plants were watered daily for 5 min with 1/1000 Hyponex solution (Hyponex, Osaka, Japan), and the trays with plants were rotated horizontally every 2 days to prevent positional light effects.

To investigate the influence of the abiotic stress that the plants were exposed to in the aforementioned dynamic time-course stress experiment (Caldana et al., 2011), we chose the most severe stress (32°C + darkness) among all abiotic stresses to conduct the stress experiment on the natural panel of 309 *A. thaliana* accessions. We randomly divided six plant replicates for each accession to two equal groups, one remaining in the control untreated condition (16 h light, 150 μE m^{-2} s^{-1} at day/night temperature of 20°C/16°C and humidity 60%/75%, abbreviated as control condition), and the other exposed to stress (darkness, temperature of 32°C and humidity 75%, abbreviated as stress condition) for 1280 continuous minutes, completely mimicking the stress condition in our previous time-course stress experiment (Caldana et al., 2011). At 42 days post germination, plants from the control group were harvested within 1 h, starting 4 h after the beginning of the light period in random order to minimize any variation due to harvest order. Next, after 1280 min of stress treatment, the stress-group plants were harvested within 1 h. For both conditions, three independent plants were pooled together to make one biological replicate of each sample, and frozen in liquid nitrogen. All the samples were stored at −80°C until subsequent LC–MS metabolite profiling.

Knockout Mutant Lines: Selection, Genotyping, and Growth Conditions

A. thaliana Col-0 (wild-type) plants were used as control throughout the experiment. We obtained 11 SALK lines, one SAIL line, and one GABI line from the Arabidopsis Stock Center, with T-DNA insertions in the following genes: *UGT89A2* (*AT5G03490*; SALK_081110 and SALK_085860), *TyrDC* (*AT4G28680*; SALK_106437, SALK_135982, SALK_120028, and SALK_090725), *BGLU1* (*AT1G45191*; SALK_060948), *AT3G55700* (SALK_096355 and SALK_046282), *BGLU7* (*AT3G62740*; GABI_612_C01), *BGLU8* (*AT3G62750*; SALK_036368 and SAIL_731_D02), and *AT4G18440* (SALK_099350). Knockout lines were selected on plates supplemented with kanamycin for SALK lines, BASTA for SAIL lines, and sulfadiazine for GABI lines. Non-segregating homozygous lines were then genotyped. Left primer (LP), right primer (RP), and border primer (BP) were designed using the Primer Design Tool provided by the Salk Institute Genomic Analysis Laboratory (http://signal.salk.edu/tdnaprimers.2.html) and used for PCR analysis to assay for the presence of the T-DNA and for zygosity in the offspring of the delivered seeds. qPCR analysis of the mutant lines was performed with gene-specific primers. All the primers used in this study are listed in Supplemental Table 14. All of the T-DNA insertion mutants were demonstrated to have completely knocked out the expression of the gene.

Knockout lines and control plants (Col-0) were grown, 12 biological replicates from each line, in short-day condition for 4 weeks, then transferred to long-day condition for another 2 weeks. Next, we randomly divided the plants into two equal groups, one remaining in control untreated condition and the other exposed to stress (32-D) for 1280 continuous minutes, mimicking the stress condition in the time-course stress experiment. The rosettes of all plants in normal and stress conditions were harvested and frozen in liquid nitrogen, then stored at −80°C until subsequent LC–MS measurement.

Untargeted Metabolite Profiling by Liquid Chromatography–Mass Spectrometry

Semi-polar metabolite extraction and measurement from *A. thaliana* leaves using LC–MS were performed as described by Giavalisco et al. (2011). In brief, approximately 100 mg of frozen *Arabidopsis* rosettes were homogenized twice for 1 min at maximum speed using a Retschmill (MM 301, Retsch, http://www.retsch.com). The metabolites were extracted in 1 ml of a homogeneous mixture of pre-cooled methanol/methyl-tert-butyl-ether/water (1:3:1), with shaking for 10 min at 4°C. This was followed by another 10 min of incubation in an ice-cooled ultrasonication bath. The homogenate was then supplemented with 650 μl of UPLC-grade methanol/water (1:3), and was vortexed and spun for 5 min at 4°C. The addition of methanol/water resulted in a phase separation, with the polar and semi-polar metabolites in the lower aqueous phase. The separate phase was isolated and dried down in a SpeedVac and stored at −80°C until subsequent LC–MS analysis. LC–MS data were obtained using a Waters Acquity UPLC system (Waters, http://www.waters.com), coupled to an Exactive mass spectrometer (Thermo Fisher, http://www.thermofisher.com). Instrumental settings were previously described (Giavalisco et al., 2011). Chromatograms from the UPLC–FT/MS runs were analyzed and processed with REFINER MS 10.0 (GeneData, http://www.genedata.com). Molecular masses, retention times, and associated peak intensities for each sample were extracted from the .raw files. The chemical noise was subtracted automatically. The chromatogram alignments were performed using a pairwise alignment-based tree using *m/z* windows of five points and retention-time windows of five scans within a sliding frame of 200 scans. The further processing of the MS data included isotope clustering, adduct detection, and library search. Resulting data matrices with peak ID, retention time, and peak intensities in each sample were generated. Day-normalization and sample-median-normalization were conducted, and the resulting data matrices were used for further analysis.

Metabolite Annotation and Identification

To facilitate the annotation of detected metabolite features by our comprehensive metabolomics approach, we obtained accurate *m/z* and retention time of each metabolite feature. Considering that the annotation of all the detected metabolite features is beyond the scope of this study, we only focused on metabolite features that were biologically interesting. We first checked each metabolite feature in the chromatograms of ^{13}C-, ^{15}N-, and ^{34}S-labeled *Arabidopsis* Col-0 leaves, described previously (Giavalisco et al., 2011), to narrow down the possible elemental formula.

The result matrix of isotope-shifted secondary metabolite data is provided in this previous publication. Next, for each target metabolite feature, a fragmentation pattern was obtained by running the analysis under Data Dependent tandem mass spectrometry Top-3 MS/MS mode using the normalized collision energy 25. The metabolite information, including accurate *m/z*, possible chemical formula, and the fragmentation pattern, was searched against different databases (PubChem, ChemSpider, KEGG, METLIN, MassBank, HMDB, and KNApSAcK). Based on the putative annotation, commercially available standards were purchased and analyzed using the same profiling procedure as described above. By comparing the accurate *m/z*, retention time, and the fragmentation patterns with the standards, the annotation for the key metabolite traits that were in the candidate associations was confirmed. For the metabolites that could not be identified by available standards, their putative annotation was obtained with the best matches for the possible structures by searching the aforementioned databases and the literature.

Metabolite datasets in positive and negative ionization modes were merged to single datasets in the control and stress conditions, following the method proposed by Calderón-Santiago et al. (2016). Next, we grouped together the metabolite features that are actually derived from the same metabolite to generate unique metabolites based on the following criteria: (i) tolerance of mass difference as 5 ppm; (ii) retention-time tolerance as 0.05 min; (iii) correlation between metabolite features across all accessions higher than 0.90.

Genome-wide Associations

Population-Structure Analysis by Metabolomics

PCA was used to infer the structure of the diverse worldwide *A. thaliana* population. After exclusion of five accessions that had poor germination from a total of 314 *A. thaliana* accessions, the data matrix was generated from 309 *A. thaliana* varieties and 4182 and 3968 detected metabolite features in control and stress conditions, respectively. PCA for metabolite profiles in control and stress samples was conducted to also show the influence of stress conditions on the natural panel of *A. thaliana* accessions.

Statistical Analysis of Metabolic Traits

Fold-change values were calculated independently for each metabolite feature in the control and stress datasets, using the maximal intensity to divide the minimal intensity of a given metabolite feature across all the used accessions. Broad-sense heritability (H^2) was calculated using the following equation by treating genotype and environment as random effects, applying a mixed linear model: $H^2 = var_{(G)}/(var_{(G)} + var_{(E)})$, where $var_{(G)}$ and $var_{(E)}$ represent the variance derived from genotypic and environmental effects, respectively, adapted from Chen et al. (2014).

Data Acquisition for GWAS and Mapping

200K SNP data for 309 *A. thaliana* accessions, obtained using Affymetrix GeneChip Array 6.0, were taken from previous publications (Li et al., 2010; Horton et al., 2012). Metabolic profiling was performed using LC–MS as described above. To avoid spurious false-positive associations due to small sample sizes, we included in the data preprocessing only metabolic traits with non-missing values across at least 50% of the accession samples. Following this initial quality control, 4182 and 3968 metabolite features were detected for control and stress conditions. Metabolite intensities were log-transformed. Genome-wide association analysis for metabolite traits was performed using 199 455 SNPs with minor allele frequency >1% across 309 accessions to investigate the associations between metabolite traits and SNPs. At each of these SNPs, a compressed mixed linear model (Zhang et al., 2010) was fitted for each trait in the Genome Association and Prediction Integrated Tool (GAPIT) R package (Lipka et al., 2012). This model includes principal components as fixed effects to account for population structure (commonly called the "Q" matrix) (Price et al., 2006), and a kinship matrix (commonly called the "K" matrix) (Eu-Ahsunthornwattana et al., 2014) to account for family relatedness across the accessions. The SNP fraction parameter was set to 0.1 to avoid excessive computation, as recommended by the GAPIT user manual. Other parameters were set to default values.

Locus Identification

The following procedure was applied to identify genomic regions associated with metabolite traits. First, all the SNPs with LOD value > $-\log_{10}(1/N)$ (N is the number of SNPs used in the study) were extracted as described previously (Wen et al., 2014). LOD threshold was set at 5.3 by using this method. The resulting SNPs with LOD $\geq$ 5.3 were then assigned to the same group if the genomic distance between them was less than 10 kb. Finally, all the genes within the resulting groups were taken into account as putative candidates.

mQTL Comparison

To find environment-specific mQTL, for a certain metabolite trait we compared mQTL from the control and stress conditions. Two resulting loci were regarded as identical if the genes in these two loci shared more than 70% overlap.

Linkage Disequilibrium Analysis

To assist in identification of causal genes in the mapped genomic regions for a given metabolite trait, we calculated the correlations between each SNP marker in the mapped locus and the lead SNP (with highest LOD) to determine LD block.

Performance Evaluation of GWAS Conducted in Different Environments

We used five distinct metabolite classes (glucosinolates, flavonoids, phenylpropanoids, amino acids, and amines) to evaluate and compare the performance of GWAS in the two different environments. We first generated the reference gene lists (Supplemental Table 5) containing all the experimentally characterized or putatively annotated genes related to the aforementioned five metabolite classes (Kanehisa and Goto, 2000; Thimm et al., 2004; Chan et al., 2011; Fraser and Chapple, 2011; Saito et al., 2013). Next, we obtained the gene lists (actual gene lists) from the resulting loci that glucosinolates, flavonoids, phenylpropanoids, amino acids, and amines mapped to in only control, only stress, and control + stress datasets. By comparing the actual gene lists and the reference gene list for the five metabolite classes using LOD thresholds ranging from 5.3 to 10.0 with an interval of 0.1, we obtained the number of common genes between the actual and reference gene lists.

The performance of our GWAS conducted in different environments was evaluated by calculating *precision*, *recall*, and *F-measure* (Powers, 2011). The parameter *precision* represents the positive predictive value of the method; *recall* is equivalent to sensitivity. The two metrics are often combined as their harmonic mean, known as the *F-measure*:

$$precision = \frac{N_c}{N_g}$$

$$recall = \frac{N_c}{N_r}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall},$$

where N_r is the number of genes provided by the reference gene lists for each metabolite class, N_g is the number of genes provided by GWAS, and N_c is the number of common genes between the actual and reference gene lists.

The significance of overlap between the actual gene list obtained from GWAS and the reference gene list for the five metabolite classes was calculated by Fisher's exact test (Rivals et al., 2007) by using the "fisher.test" function in R.

Network Analysis

Transcript and Metabolite Data Acquisition from Time-Course Stress Experiments

Transcript data from time-course stress experiments were derived from our previous work (Caldana et al., 2011), resulting in 15 089 transcripts

for further network analysis. We used the same plant material from our previous work (Caldana et al., 2011) to perform untargeted metabolomic profiling. Metabolite extraction, measurement, and data processing followed the same process mentioned above. Significantly changed metabolites across 23 time points in each condition were selected by ANOVA using the "aov" function in R (http://www.r-project.org/) at a significance level of 0.05 with a multiple correction test, using false discovery rate (FDR) estimation by comparing three replicates at all time points. The metabolites that changed significantly in each condition were used for the construction of condition-specific networks. PCA for metabolite and transcript profiles was performed using the "pcaMethods" Bioconductor package (Stacklies et al., 2007).

Condition-Specific Network Construction and Network Quality Assessment

Pearson correlation coefficient (PCC) between metabolite and transcript features was calculated in R. PCC thresholds for building edges between features (metabolites and transcripts) in networks in each condition were obtained based on a permutation test (FDR < 0.05). The performance of each condition-specific network was assessed by *precision*, *recall*, and *F-measure*, which followed the procedure described previously (Wu et al., 2016). The optimal PCC thresholds were selected when *F-measure* reached the highest values. Based on the optimal PCC thresholds, undirected networks for each condition were constructed with nodes representing metabolite and transcript features and edges connecting the nodes between features, with a PCC passing the threshold using the igraph package (Csardi and Nepusz, 2006) in R.

Permutation Test for the Network Analysis for Integration with GWAS

All metabolite features in the GWAS and network datasets were matched to obtain common metabolite features for further data integration. We regarded metabolite features from these two datasets as identical if (i) their mass difference was less than 5 ppm, (ii) their retention-time difference was less than 0.05 min, and (iii) they shared similar spectra. For a given metabolite shared between the GWAS and network datasets, the number of genes from GWAS that were also supported by network analysis was defined as the true number of confirmed genes (x). To check whether the network analysis can support the GWAS in selection and prioritization of candidate genes better than random networks, we applied a permutation test. We used the same number of randomly selected genes as the number of genes provided by network analysis. The randomly selected genes were used to compare and support the genes provided by GWAS, and the permuted number of confirmed genes (yk) was obtained. To estimate a p value empirically, we then compared the true number of confirmed genes (x) and permuted number of confirmed genes (yk) in k permutations (k = 10 000):

$$p = \frac{1}{n}\sum_{k=1}^{n} F(x, y_k),$$

$$F(x,y) = \begin{cases} 0 & \text{for } x > y_k \\ 1 & \text{else} \end{cases}.$$

Hence, if the true number of confirmed genes is higher than the permuted number of confirmed genes for 9500 of the 10 000 permutations, we obtain a p value estimate of 0.05.

Comparison of AT3G55700 Protein Sequence with Its Orthologs

We compared the identity of the No. 294 amino acid in the sequence of AT3G55700 with that from 16 other species (*Arabidopsis lyrata*, *Capsella rubella*, *Brassica rapa*, *Gossypium raimondii*, *Theobroma cacao*, *Citrus sinensis*, *Eucalyptus grandis*, *Prunus persica*, *Malus domestica*, *Fragaria vesca*, *Medicago truncatula*, *Glycine max*, *Ricinus communis*, *Manihot esculenta*, *Populus trichocarpa*, and *Vitis vinifera*).

Statistics for Knockout Validation Experiment

Metabolite intensity data after transformation and normalization were used for ANOVA to test the significance levels of metabolite changes in knockout and Col-0 plants under normal and stress conditions, following by correction for multiple comparisons using the "p.adjust" function in R (http://www.r-project.org/). Subsequently, pairwise comparison was conducted by the Tukey's HSD tests using the "TukeyHSD" function in R.

Data Availability

All data generated or analyzed during this study are available as a public resource. The raw metabolomics datasets from the two-condition-based GWAS and the time-course stress experiment are provided in Supplemental Tables 3 and 7. The gene expression dataset from the time-course stress experiment was obtained from our previous publication (Caldana et al., 2011), which was deposited in the Array-Express repository (http://www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-375.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

AUTHOR CONTRIBUTIONS

S.W. conducted experiments. S.W. analyzed and interpreted the data and results with input from T.T., A.C.-I., and H.Tong. M.M. performed MS/MS experiments. Y.B. and L.W. conceived and supervised the study. R.K. and J.B.K. provided some of the plant materials for the experiments. S.W. wrote the manuscript with input from H.Tenenboim, Z.N., A.R.F., L.W., and Y.B. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Änne Michaelis and Gudrun Wolter for excellent technical assistance. We acknowledge Dr. Patrick Giavalisco, Dr. Corina M. Fusari, Dr. Sebastian Klie, Dr. Marek Mutwil, Dr. Sabrina Kleeßen, and Dr. Nooshin Omranian for critical and helpful discussions. We thank Urszula Luzarowska, Niklas Endres, Francisco de Abreu e Lima, Marcin Luzarowski, Weronika Jasinska, Sarah Khedhayir, Anastasiya Kuhalskaya, and Philip Hofmann for help in harvesting plant material. We acknowledge Dr. Masami Yokota Hirai from RIKEN Center and Dr. Kyoungwhan Back from Chonnam National University for kindly providing *mam1*, *mam3*, and *omt1* KO seeds. No conflict of interest declared.

Received: June 24, 2017
Revised: August 16, 2017
Accepted: August 23, 2017
Published: August 31, 2017

REFERENCES

Agrawal, A.A., Hastings, A.P., Johnson, M.T., Maron, J.L., and Salminen, J.P. (2012). Insect herbivores drive real-time ecological and evolutionary change in plant populations. Science **338**:113–116.

Angelovici, R., Lipka, A.E., Deason, N., Gonzalez-Jorge, S., Lin, H., Cepela, J., Buell, R., Gore, M.A., and Dellapenna, D. (2013). Genome-wide analysis of branched-chain amino acid levels in *Arabidopsis* seeds. Plant Cell **25**:4827–4843.

Bac-Molenaar, J.A., Fradin, E.F., Rienstra, J.A., Vreugdenhil, D., and Keurentjes, J.J. (2015). GWA mapping of anthocyanin accumulation reveals balancing selection of MYB90 in *Arabidopsis thaliana*. PLoS One **10**:e0143212.

Bieniawska, Z., Espinoza, C., Schlereth, A., Sulpice, R., Hincha, D.K., and Hannah, M.A. (2008). Disruption of the *Arabidopsis* circadian clock is responsible for extensive variation in the cold-responsive transcriptome. Plant Physiol. **147**:263–279.

Caldana, C., Degenkolbe, T., Cuadros-Inostroza, A., Klie, S., Sulpice, R., Leisse, A., Steinhauser, D., Fernie, A.R., Willmitzer, L., and Hannah, M.A. (2011). High-density kinetic analysis of the metabolomic

and transcriptomic response of *Arabidopsis* to eight environmental conditions. Plant J. **67**:869–884.

Calderón-Santiago, M., Fernández-Peralbo, M.A., Priego-Capote, F., and Luque de Castro, M.D. (2016). MSCombine: a tool for merging untargeted metabolomic data from high-resolution mass spectrometry in the positive and negative ionization modes. Metabolomics **12**:43.

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

Chan, E.K., Rowe, H.C., and Kliebenstein, D.J. (2010b). Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. Genetics **185**:991–1007.

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., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat. Genet. **46**:714–721.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Systems, 1695.

Davila Olivas, N.H., Kruijer, W., Gort, G., Wijnen, C.L., van Loon, J.J., and Dicke, M. (2016). Genome-wide association analysis reveals distinct genetic architectures for single and combined stress responses in *Arabidopsis thaliana*. New Phytol. **213**:838–851.

Eu-Ahsunthornwattana, J., Miller, E.N., Fakiola, M., Jeronimo, S.M., Blackwell, J.M., and Cordell, H.J. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data. PLoS Genet. **10**:e1004445.

Fraser, C.M., and Chapple, C. (2011). The phenylpropanoid pathway in *Arabidopsis*. Arabidopsis Book **9**:e0152.

Fraser, C.M., Thompson, M.G., Shirley, A.M., Ralph, J., Schoenherr, J.A., Sinlapadech, T., Hall, M.C., and Chapple, C. (2007). Related *Arabidopsis* serine carboxypeptidase-like sinapoylglucose acyltransferases display distinct but overlapping substrate specificities. Plant Physiol. **144**:1986–1999.

Galili, G., Tang, G., Zhu, X., and Gakiere, B. (2001). Lysine catabolism: a stress and development super-regulated metabolic pathway. Curr. Opin. Plant Biol. **4**:261–266.

Giavalisco, P., Li, Y., Matthes, A., Eckhardt, A., Hubberten, H.M., Hesse, H., Segu, S., Hummel, J., Kohl, K., and Willmitzer, L. (2011). Elemental formula annotation of polar and lipophilic metabolites using (13)C, (15)N and (34)S isotope labelling, in combination with high-resolution mass spectrometry. Plant J. **68**:364–376.

Hansen, B.G., Kerwin, R.E., Ober, J.A., Lambrix, V.M., Mitchell-Olds, T., Gershenzon, J., Halkier, B.A., and Kliebenstein, D.J. (2008). A novel 2-oxoacid-dependent dioxygenase involved in the formation of the goiterogenic 2-hydroxybut-3-enyl glucosinolate and generalist insect resistance in *Arabidopsis*. Plant Physiol. **148**:2096–2108.

Hectors, K., Van Oevelen, S., Geuns, J., Guisez, Y., Jansen, M.A., and Prinsen, E. (2014). Dynamic changes in plant secondary metabolites during UV acclimation in *Arabidopsis thaliana*. Physiol. Plant. **152**:219–230.

Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N.W., Platt, A., Sperone, F.G., Vilhjalmsson, B.J., et al. (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat. Genet. **44**:212–216.

Ishihara, H., Tohge, T., Viehover, P., Fernie, A.R., Weisshaar, B., and Stracke, R. (2016). Natural variation in flavonol accumulation in *Arabidopsis* is determined by the flavonol glucosyltransferase BGLU6. J. Exp. Bot. **67**:1505–1517.

Jones, P., Messner, B., Nakajima, J., Schaffner, A.R., and Saito, K. (2003). UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in *Arabidopsis thaliana*. J. Biol. Chem. **278**:43910–43918.

Joseph, B., Corwin, J.A., Li, B., Atwell, S., and Kliebenstein, D.J. (2013). Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. eLife **2**:e00776.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. **28**:27–30.

Kerwin, R., Feusier, J., Corwin, J., Rubin, M., Lin, C., Muok, A., Larson, B., Li, B., Joseph, B., Francisco, M., et al. (2015). Natural genetic variation in *Arabidopsis thaliana* defense metabolism genes modulates field fitness. eLife **4**:e05604.

Keurentjes, J.J. (2009). Genetical metabolomics: closing in on phenotypes. Curr. Opin. Plant Biol. **12**:223–230.

Keurentjes, J.J., Fu, J., de Vos, C.H., Lommen, A., Hall, R.D., Bino, R.J., van der Plas, L.H., Jansen, R.C., Vreugdenhil, D., and Koornneef, M. (2006). The genetics of plant metabolism. Nat. Genet. **38**:842–849.

Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J., and Mitchell-Olds, T. (2001). Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. Plant Cell **13**:681–693.

Kliebenstein, D., Pedersen, D., Barker, B., and Mitchell-Olds, T. (2002). Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in *Arabidopsis thaliana*. Genetics **161**:325–332.

Korte, A., Vilhjalmsson, B.J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. **44**:1066–1071.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., and Borevitz, J.O. (2010). Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **107**:21199–21204.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. **45**:43–50.

Li, X., Svedin, E., Mo, H., Atwell, S., Dilkes, B.P., and Chapple, C. (2014). Exploiting natural variation of secondary metabolism identifies a gene controlling the glycosylation diversity of dihydroxybenzoic acids in *Arabidopsis thaliana*. Genetics **198**:1267–1276.

Li, P., Li, Y.J., Zhang, F.J., Zhang, G.Z., Jiang, X.Y., Yu, H.M., and Hou, B.K. (2016). The *Arabidopsis* UDP-glycosyltransferases UGT79B2 and UGT79B3, contribute to cold, salt and drought stress tolerance via modulating anthocyanin accumulation. Plant J. **89**:85.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics **28**:2397–2399.

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, J., 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.

Muzac, I., Wang, J., Anzellotti, D., Zhang, H., and Ibrahim, R.K. (2000). Functional expression of an *Arabidopsis* cDNA clone encoding a

flavonol 3′-O-methyltransferase and characterization of the gene product. Arch. Biochem. Biophys. **375**:385–388.

Navarova, H., Bernsdorff, F., Doring, A.C., and Zeier, J. (2012). Pipecolic acid, an endogenous mediator of defense amplification and priming, is a critical regulator of inducible plant immunity. Plant cell **24**:5123–5141.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. **12**:443–451.

Powers, D.M.W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Tech. **2**:37–63.

Prasad, K.V., Song, B.H., Olson-Manning, C., Anderson, J.T., Lee, C.R., Schranz, M.E., Windsor, A.J., Clauss, M.J., Manzaneda, A.J., Naqvi, I., et al. (2012). A gain-of-function polymorphism controlling complex traits and fitness in nature. Science **337**:1081–1084.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**:904–909.

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A.E. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc. Natl. Acad. Sci. USA **109**:8872–8877.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics **23**:401–407.

Routaboul, J.M., Dubos, C., Beck, G., Marquis, C., Bidzinski, P., Loudet, O., and Lepiniec, L. (2012). Metabolite profiling and quantitative genetics of natural variation for flavonoids in *Arabidopsis*. J. Exp. Bot. **63**:3749–3764.

Rowe, H.C., Hansen, B.G., Halkier, B.A., and Kliebenstein, D.J. (2008). Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. Plant Cell **20**:1199–1216.

Saito, K., Yonekura-Sakakibara, K., Nakabayashi, R., Higashi, Y., Yamazaki, M., Tohge, T., and Fernie, A.R. (2013). The flavonoid biosynthetic pathway in *Arabidopsis*: structural and genetic diversity. Plant Physiol. Biochem. **72**:21–34.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P.T., Nikoloski, Z., Fernie, A.R., and Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol. **165**:1120–1132.

Serrano, G.C., Rezende e Silva Figueira, T., Kiyota, E., Zanata, N., and Arruda, P. (2012). Lysine degradation through the saccharopine pathway in bacteria: LKR and SDH in bacteria and its relationship to the plant and animal enzymes. FEBS Lett. **586**:905–911.

Soltis, N.E., and Kliebenstein, D.J. (2015). Natural variation of plant metabolism: genetic mechanisms, interpretive caveats, and evolutionary and mechanistic insights. Plant Physiol. **169**:1456–1468.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics **23**:1164–1167.

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W., Fiehn, O., Goodacre, R., Griffin, J.L., et al. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). Metabolomics **3**:211–221.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. **37**:914–939.

Tohge, T., Nishiyama, Y., Hirai, M.Y., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D.B., Kitayama, M., et al. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. Plant J. **42**:218–235.

Tohge, T., Yonekura-Sakakibara, K., Niida, R., Watanabe-Takahashi, A., and Saito, K. (2007). Phytochemical genomics in *Arabidopsis thaliana*: a case study for functional identification of flavonoid biosynthesis genes. Pure Appl. Chem. **79**:811–823.

Weigel, D. (2012). Natural variation in *Arabidopsis*: from molecular genetics to ecological genomics. Plant Physiol. **158**:2–22.

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

Wong, A., and Gehring, C. (2013). Computational identification of candidate nucleotide cyclases in higher plants. Methods Mol. Biol. **1016**:195–205.

Wu, S., Alseekh, S., Cuadros-Inostroza, A., Fusari, C.M., Mutwil, M., Kooke, R., Keurentjes, J.B., Fernie, A.R., Willmitzer, L., and Brotman, Y. (2016). Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in *Arabidopsis thaliana*. PLoS Genet. **12**:e1006363.

Yonekura-Sakakibara, K., and Hanada, K. (2011). An evolutionary view of functional diversity in family 1 glycosyltransferases. Plant J. **66**:182–193.

Yonekura-Sakakibara, K., Tohge, T., Niida, R., and Saito, K. (2007). Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in *Arabidopsis* by transcriptome coexpression analysis and reverse genetics. J. Biol. Chem. **282**:14932–14941.

Yonekura-Sakakibara, K., Tohge, T., Matsuda, F., Nakabayashi, R., Takayama, H., Niida, R., Watanabe-Takahashi, A., Inoue, E., and Saito, K. (2008). Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in *Arabidopsis*. Plant Cell **20**:2160–2176.

Yonekura-Sakakibara, K., Fukushima, A., Nakabayashi, R., Hanada, K., Matsuda, F., Sugawara, S., Inoue, E., Kuromori, T., Ito, T., Shinozaki, K., et al. (2012). Two glycosyltransferases involved in anthocyanin modification delineated by transcriptome independent component analysis in *Arabidopsis thaliana*. Plant J. **69**:154–167.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. **42**:355–360.

Zhu, X., Tang, G., Granier, F., Bouchez, D., and Galili, G. (2001). A T-DNA insertion knockout of the bifunctional lysine-ketoglutarate reductase/saccharopine dehydrogenase gene elevates lysine levels in *Arabidopsis* seeds. Plant Physiol. **126**:1539–1545.