mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis

Marianne L. Slaten, a Abou Yobi, a Clement Bagaza, a, 2 Yen On Chan, a Vivek Shrestha, a Samuel Holden, a Ella Katz, b Christa Kanstrup, c Alexander E. Lipka, d Daniel J. Kliebenstein, b Hussam Hassan Nour-Eldin, c and Ruthie Angelovici a, 3, 4

a Division of Biological Sciences, Interdisciplinary Plant Group, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211
b Department of Plant Sciences, University of California Davis, Davis, California 95616
c DynaMo Center, Copenhagen Plant Science Centre, Department of Plant and Environmental Sciences, University of Copenhagen, 1871 Frederiksberg C, Denmark
d Department of Crop Sciences, University of Illinois, Urbana, Illinois 61801

Gln is a key player in plant metabolism. It is one of the major free amino acids that is transported into the developing seed and is central for nitrogen metabolism. However, Gln natural variation and its regulation and interaction with other metabolic processes in seeds remain poorly understood. To investigate the latter, we performed a metabolic genome-wide association study (mGWAS) of Gln-related traits measured from the dry seeds of the Arabidopsis (Arabidopsis thaliana) diversity panel using all potential ratios between Gln and the other members of the Glu family as traits. This semicombinatorial approach yielded multiple candidate genes that, upon further analysis, revealed an unexpected association between the aliphatic glucosinolates (GLS) and the Gln-related traits. This finding was confirmed by an independent quantitative trait loci mapping and statistical analysis of the relationships between the Gln-related traits and the presence of specific GLS in seeds. Moreover, an analysis of Arabidopsis mutants lacking GLS showed an extensive seed-specific impact on Gln levels and composition that manifested early in seed development. The elimination of GLS in seeds was associated with a large effect on seed nitrogen and sulfur homeostasis, which conceivably led to the Gln response. This finding indicates that both Gln and GLS play key roles in shaping the seed metabolic homeostasis. It also implies that select secondary metabolites might have key functions in primary seed metabolism. Finally, our study shows that an mGWAS performed on dry seeds can uncover key metabolic interactions that occur early in seed development.
in Arabidopsis seedlings grown under nitrate-deficit conditions (Scheible et al., 2004). Interestingly, extensive variation in free Gln content in dry Arabidopsis seeds has been reported across the various accessions belonging to the Arabidopsis diversity panel (Angelovici et al., 2017), but the genetic architecture regulating this trait remains poorly understood. Knowledge regarding the genes that underlie Gln levels, composition, and seed partitioning would shed light on its potential seed-specific functions, its interaction with other biological processes, and its role in downstream metabolism.

In recent years, genome-wide association studies (GWAS) as well as quantitative trait loci (QTL) mapping experiments have facilitated the identification of many loci for both primary and secondary metabolites (Wentzell et al., 2007; Chan et al., 2011; Riedelsheimer et al., 2012; Angelovici et al., 2013; Gonzalez-Jorge et al., 2013; Chen et al., 2014; Verslues et al., 2014; Angelovici et al., 2017). In-depth analyses of these QTLs have facilitated the further discovery of key structural and regulatory genes that underlie the natural variation of metabolic traits and the identification of various cellular processes involved in metabolic homeostasis. Although GWAS and QTL mapping have been conducted on FAAs in both vegetative and seed tissues across several species, no major QTLs have been identified for Gln (Riedelsheimer et al., 2012; Chen et al., 2014; Wen et al., 2014). The lack of any identifiable loci implies that Gln either has a complex genetic architecture or that these studies possibly utilized “underpowered” association panels, or both.

The use of metabolic ratios as traits in GWAS has been useful for dealing with several such calcitrant metabolites. The approach, which relies on biochemical pathways and/or represent relationships uncovered by a metabolic network correlation analysis, has yielded several significant associations even when the absolute levels of metabolites have not (Wentzell et al., 2007; Lipka et al., 2013; Angelovici et al., 2013; Gonzalez-Jorge et al., 2013; Angelovici et al., 2017). It has been postulated that metabolic ratios are less complex (since they only represent the metabolite partitioning within biochemical pathways) and, therefore, are more tractable in association mapping studies (Angelovici et al., 2017). Still, even this approach has failed to identify QTLs for Gln in dry seeds (Angelovici et al., 2017).

A different approach is clearly needed to uncover the genetic architecture of Gln. Notably, the metabolic ratios used in previous studies do not represent all the potential ratios of Gln-related traits since they were based principally on a priori pathway information, which is often incomplete.

In theory, performing a metabolic genome-wide association study (mGWAS) on all possible Gln-related metabolic ratios would potentially resolve its genetic architecture. In practice, however, such an endeavor would be challenging given the enormous number of metabolic ratios that could be derived from the relationships between Gln and all 20 proteogenic amino acids. Therefore, as a point of departure from previous studies, we derived all possible metabolic ratios of Gln only to its proteogenic amino acid family members, thus theoretically representing all potential biologically relevant partitioning/relationship of Gln within the Glu family (Fig. 1). By combining this approach with a Fixed and Random model Circulating Probability Unification (FarmCPU), which uses fixed and random effect models for powerful and efficient GWAS studies (Liu et al., 2016), we uncovered many significant QTLs for various Gln-derived traits in dry seeds. More importantly, our analysis of the candidate genes revealed a surprising enrichment for genes residing in the glucosinolate (GLS) biosynthesis pathway, suggesting a potential interplay between two metabolic pathways that are not known to be directly linked (Fig. 1). We validated this association by using an independent QTL mapping approach as well as by characterizing Gln and other FAAs in mutant plants that have a disrupted GLS composition and loading to the seeds. Our data support an association between GLS natural diversity and Gln levels and composition in seeds, and also reveal that GLS loading to the seeds has a profound effect on seed nitrogen and sulfur homeostasis as well as Gln levels and composition. Our results strongly suggest that an interaction between Gln and GLS plays a key role in seed metabolic homeostasis.

RESULTS

The Four Glu Family Members Vary in Abundance, Relative Composition, and Broad-Sense Heritability Across the Arabidopsis Diversity Panel

In a previous study, we quantified and described the natural variation of 18 out of the 20 proteogenic FAAs
(excluding Cys and Asn) measured from dry seeds of three biological repeats of a 313-accession Arabidopsis diversity panel (Angelovici et al., 2013; Angelovici et al., 2017). In the current study, we used that data to assess the natural variation among only the proteogenic FAAs in the Glu family (i.e. Glu, Pro, Gln, and Arg).

Our analysis showed that the four Glu family members vary in abundance, relative composition, and broad-sense heritability (Supplemental Table S1A). Glu was the most abundant amino acid with a relative composition mean of 0.35, whereas Gln was the least abundant with a relative composition mean of 0.015. We defined relative composition as the ratio of an individual amino acid to the sum of the 18 measured amino acids (e.g. Gln/Total, Glu/Total). Arg and Pro had a relative composition means (Arg/Total; Pro/Total) of 0.04 and 0.017, respectively. Gln demonstrated moderate heritability (0.52) along with Pro and Glu (0.48 and 0.63, respectively), whereas Arg had the highest heritability (0.74). Interestingly, Gln had the largest relative so, whereas Glu had the smallest despite its high abundance (~61% and 23% relative standard deviation, respectively).

To evaluate the relationship between Gln and the other Glu family members, we performed a correlation-based network analysis among the four FAAs and visualized the results using Cytoscape version 3.6.1 (Supplemental Fig. S1). All correlations (r) were significant at α = 0.001 and ranged from 0.12 to 0.54. Gln was moderately correlated with Arg and Glu and weakly correlated with Pro, which had a significant but weak correlation with all Glu family members.

Figure 1. Simplified metabolic pathways of the Glu amino acid family and the aliphatic GLS. Genes that are within the genomic region of GS-ELONG or GS-AOP loci are represented in red and include MAM1, MAM2, and MAM3 as well as AOP2 and AOP3. These gene products are responsible for most of the GLS natural diversity. MAM1 is responsible for the production of C4 GLS 4-methylsulfinylbutyl GLS (4msb); MAM2 (in the absence of MAM1) is required for the production of C3 GLS 3-methylsulfinylpropyl GLS (3msp); MAM3 is responsible for the production of C8 GLS 8-methylsulfinyloctyl (8msop). C3 GLS 3msp can be further converted to 3-hydroxypropyl GLS (3ohp) by AOP3 or 2-propenyl by AOP2. C4 GLS 4msb can be converted to 4-hydroxybutyl GLS (4ohb) by AOP3 or 3-butenyl and then OH-3-butenyl by AOP2.

mGWAS Identified Significant Single Nucleotide Polymorphism-Trait Associations for Six Gln-Related Traits

In our previous study, no significant associations were identified when seed Gln traits or any Gln-related traits derived from a priori knowledge of the Glu metabolic pathway or correlation-based network analysis were used for the mGWAS (Angelovici et al., 2017). Therefore, we took a slightly different approach in this study by using all possible Gln metabolic ratios that could be derived from Gln relationships with the other members of the Glu family. The various relationships were represented by calculating all the possible ratios in which Gln is the numerator and is divided by a sum of every combination of the four Glu family members, including Gln itself [i.e. Gln/(Gln+Arg+Pro+Gu) = (and/or)]. We consider this a semicombinatorial approach since it relies on both a priori knowledge of the Glu family as well as all the possible combinations of the Glu family FAAs in the denominator. The traits and their corresponding means, ranges, and broad-sense heritability scores are listed in Supplemental Table S1B. For simplicity, we used a one letter code in our trait representations. The sum of the FAA in the denominator of each trait is represented by a string of one letter codes. For example, Q/EP is Gln divided by the sum of Glu and Pro. This approach yielded 16 Gln-related traits: 14 ratio-based traits (Supplemental Table S1B), one free Gln absolute level, and the Gln relative composition (Gln/Total; Supplemental Table S1A). Of all these 16 traits, Q/QP had the highest heritability (0.53)
and Q/RP had the lowest (0.35). In general, the derived traits had low to moderate heritability.

We used the FarmCPU package in R (version 1.02; Liu et al., 2016) to perform an mGWAS on the 16 Gln-related traits. Because FarmCPU may be prone to a type I error, we chose to use the more conservative Bonferroni multiple testing correction procedure instead of the Benjamini and Hochberg (1995) false discovery rate-controlling procedure. We also considered single nucleotide polymorphism (SNP)-trait associations significant only at an \( \alpha = 0.01 \) Bonferroni correction level. At this significance threshold, we identified 21 SNP–trait associations for six traits: Q/P, Q/R, Q/QP, Q/RP, Q/RQ, and Q/RQP (Fig. 2; Supplemental Dataset S1); only 16 SNPs were identified from the 21 signals. None of the six traits included Glu in their denominator but did include either Arg or Pro, or both. The heritability of these six traits ranged from low to moderate (0.35–0.53; Supplemental Table S1B). No significant associations were observed on chromosome 1. One was observed on chromosome 2, and three on chromosome 3. The majority of significant SNPs were identified on either chromosome 4 or 5 (Fig. 2; Supplemental Dataset S1). The five SNPs with the lowest \( P \) values were located on chromosomes 4 or 5 (Table 1); three of these SNPs fell within a gene, whereas the remaining two were located in a transposable element and an intragenic region. The three genes are annotated as encoding Brassinosteroid suppressor 1 (BSU1), a MATE efflux family protein, and methylthioalkylmalate synthase 1 (MAM1).

**Genes Within Haploblocks Spanning Significant SNPs Are Enriched for Glucosinolate Biosynthetic Process**

We compiled a candidate gene list based first on genes that contain a significant SNP. We then expanded the list to include those genes that are in strong linkage disequilibrium (LD; defined as regions with nonrandom associations calculated using a 95% confidence bounds on D prime) with the significant SNPs identified by our mGWAS, since significant SNPs identified by GWAS may tag causal variants in neighboring genes that are in LD (Atwell et al., 2010). To that end, we identified haploblocks that spanned the 16 SNPs using Haploview version 4.2 (See “Materials and Methods”; Barrett et al., 2004) and considered all spanned genes as candidates. If a haploblock was not identified for a given SNP and did not fall within a gene, then the gene directly upstream or downstream was recorded. Overall, we found 27 unique genes. The entire list of genes associated with all 16 SNPs is summarized in Supplemental Table S2A.

Next, we used agriGO (http://bioinfo.cau.edu.cn/agriGO/) to perform a Gene Ontology (GO) enrichment analysis of the 27 genes. We analyzed all genes identified across the six traits since, collectively, they represent the potential genetic architecture of the Gln partition within the Glu family and its relationships to the other members. The analysis revealed a significant enrichment for the following terms: secondary metabolic process, carbohydrate metabolic process, sulfur metabolic process, S-glycoside biosynthetic process, and glucosinolates biosynthetic process (Supplemental Table S2B).

All the significant enrichment terms resulted from three genes: MAM1 (AT5G23010), AOP1 (AT4G03070), and AOP3 (AT4G03050), all of which are annotated as involved in the biosynthesis of aliphatic GLS. Notably, one of our top five significant SNPs fell within MAM1 (Q/P; Table 1). AOP1 was associated with traits Q/RQ and Q/RQP, and AOP3 was associated with trait Q/RQ (Fig. 2; Supplemental Dataset S1). Although these genes are located in three different haploblocks, AOP1 and AOP3 are in very close proximity within the

---

**Figure 2.** The chromosomal distribution of the 21 significant SNP–trait associations identified by the mGWAS. A total of 21 significant SNPs-trait associations were identified at an \( \alpha = 0.01 \) Bonferroni for six traits. SNPs are represented by short lines that are color-coded based on their \( P \) value. The histogram on top of the heatmap illustrates the number of occurrences of each SNP identified across all traits. Arrows indicate SNPs located within a gene or haploblock of interest. Asterisks designate traits that have a significant association with a SNP in a GLS gene or in high LD with GLS genes. Q, Gln; R, Arg; P, Pro.
Table 1. Top five SNP-trait associations with the lowest $P$ values

<table>
<thead>
<tr>
<th>Trait</th>
<th>SNP</th>
<th>Chromosomal Position</th>
<th>$P$ Value</th>
<th>Functional Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>S204486</td>
<td>Q/P</td>
<td>5 21544586</td>
<td>6.18E-2</td>
<td>Transposable element</td>
</tr>
<tr>
<td>S127076</td>
<td>Q/RP</td>
<td>4 1360042</td>
<td>6.17E-1</td>
<td>None</td>
</tr>
<tr>
<td>S124151</td>
<td>Q/R</td>
<td>4 152626</td>
<td>5.32E-2</td>
<td>MATE efflux family protein</td>
</tr>
<tr>
<td>S190878</td>
<td>Q/RP</td>
<td>5 15830557</td>
<td>5.22E-1</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>S175365</td>
<td>Q/P</td>
<td>5 7704845</td>
<td>3.79E-3</td>
<td>Methylthioalkylmalate synthase 1</td>
</tr>
</tbody>
</table>

The high LD with neighboring SNPs suggests that this SNP may tag a causal variation in one or both of these AOP genes (Supplemental Fig. S2A). Similarly, SNP S175365, which resides in the same haploblock as MAM1, is in strong to moderate LD with SNPs associated with MAM2 (S175355; $r^2 = 0.908$) and MAM3 (S175394; $r^2 = 0.649$; Supplemental Fig. S3).

Overall, we found six genes involved in aliphatic GLS biosynthesis that are in moderate ($>0.5$) to strong ($>0.8$) LD with three of significant SNPs in the region. It is likely that either one or an allelic combination of all six genes contributes to the natural variation of free Gln and its related traits in dry seeds.

QTL Analysis of the Bayreuth-0 and Shahdara Mapping Population Supports the GWAS Finding

The finding of an association between Gln and GLS in dry seeds was surprising. Glucosinolates are not
synthesized in seeds but rather are transported to the seed from the maternal plant (Magrath and Mithen, 1993). Therefore, to independently confirm our results from the mGWAS and to further support the association between Gln and the two GLS-related QTLs, we performed a biparental QTL mapping using the Bayreuth-0 (Bay) and Shahdara (Sha) recombinant inbred population (Loudet et al., 2002). Previous work has shown that Bay and Sha segregate at the GS-ELONG and GS-AOP loci and have an epistatic relationship (Kliebenstein et al., 2001; Kroymann et al., 2003; Textor et al., 2004; Kliebenstein et al., 2007; Wentzell et al., 2007). We hypothesized that if these GLS-related QTLs are indeed responsible for the natural variation of Gln in dry seeds, then the Bay × Sha mapping population should recapitulate the QTL for the Gln-related traits.

To test this hypothesis, we used the FAA quantifications from 158 recombinant inbred lines of the Bay × Sha population, as described previously (Angelovici...
et al., 2013; Angelovici et al., 2017), and performed a QTL analysis of our 16 Gln-related traits using Multiple QTL Mapping (MQM) in the R/quantile package in R (Arends et al., 2010). This approach yielded a total of 25 QTLs for eight traits (for the full list see Supplemental Dataset S2). Six traits had significant log of the odds (LOD) maxima on chromosome 5 at marker MSAT5.14 (position 7498509 bp): Q/RQ, Q/RQP, Q/R, Q/RP, Q/QP, and Q/P. The supporting interval overlapped with the GS-ELONG locus (Table 2). Both the highest percent of total phenotypic variation and the highest LOD were observed for Q/QP and Q/P. These two traits also had a LOD maxima on chromosome 4 at marker MSAT4.43 with supporting intervals spanning the GS-AOP locus.

Interaction between the two QTLs has been observed previously in GLS traits (Kliebenstein et al., 2001; Kliebenstein et al., 2007). Therefore, we tested whether interactions between the two loci existed for our Gln-related traits. Visual inspection of the interaction plots between markers MSAT4.43 and MSAT5.14 clearly indicated interaction between these markers that seem to heavily influence the Q/QP and Q/P trait means (Supplemental Fig. S4).

The Presence or Absence of Specific GLS Has a Significant Effect on the Levels of the Gln-Related Traits in Dry Seeds

To further validate the association between GLS natural variation and the Gln-related traits, we grew 133 accessions from the Arabidopsis diversity panel and measured both FAA and Gln levels in the dry seeds (Supplemental Dataset S3). Next, we tested whether the presence or absence of one of the four GLS, which result from the different allelic combinations at the GS-ELONG and GS-AOP loci (Fig. 1), was associated with high or low levels of our traits of interest (i.e. the 16 Gln-related traits analyzed in our mGWAS). The four GLS analyzed for presence/absence were 30hp (requiring the presence of MAM2 and AOP3), 2-propenyl (requiring the presence of MAM2 and AOP2), 40hb (requiring the presence of MAM1 and AOP3), and 3butenyl/ÔH-3-butenyl (requiring the presence of MAM1 and AOP2). To evaluate this association, we performed t tests on the levels of the Gln-related traits measured from accessions that either had a specific GLS chemotype (i.e. 30hp or 40hb) or completely lacked it (see “Materials and Methods” for more details regarding the statistical analysis).

Our results showed that Gln absolute levels were significantly less in the presence of 2-propenyl (Supplemental Table S3). However, the presence/absence of both 30hp and 40hb had the most significant effect on our traits. The presence of 30hp had a negative effect on most of the Gln-related ratios and had a positive effect on the absolute levels of Arg, Glu, and Pro. By contrast, the presence of 40hb had the opposite effect on most of the Gln-related traits in addition to the absolute levels of Glu and Pro (Supplemental Table S3). Taken collectively, these results both confirm that GLS diversity can significantly affect the Gln-related traits and further supports the association between these two pathways.

FAA Characterization of Mutants in GLS Genes Present in the GS-ELONG and GS-AOP Showed Only Small Effects on Gln-Related Traits in the Col-0 Background

We performed a transgenic approach to further confirm the association between aliphatic GLS and Gln content in dry Arabidopsis seeds. We obtained null and overexpression (OX) mutants of the six relevant genes located in the GS-ELONG or GS-AOP locus and involved in aliphatic GLS biosynthesis. All plants were grown to maturity, and their dry seeds harvested and analyzed for FAA content and composition. We also obtained and quantified the dry seed FAA content of a bsu1 null mutant, which lacks the BSU1 genes that contain the significant SNP (i.e. S127076) identified for traits Q/RP and Q/RQP (Fig. 4; Table 1). The T-DNA insertion lines were ordered from the SALK and WISc T-DNA collections and included insertions in the AT4G03070 (aop1), AT4G03050 (aop3), AT5G23020 (mam3), and AT4G03080 (bsu1) genes. The T-DNA insertion locations are summarized in Supplemental Fig. S5. Null homozygous mutants were isolated and confirmed by the absence of the full transcript in a tissue of high expression (Supplemental Figs. S5 and S6). Based on the eFP browser expression data (Schmid et al., 2005; Winter et al., 2007), AOP1 expression was evaluated in imbibed seeds, AOP3 was evaluated in young siliques, MAM1 and MAM3 were evaluated in seedlings, and BSU1 was evaluated in leaves. The reverse transcription-PCR primers used are listed in Supplemental Table S4.

Interestingly, all genes, excluding AOP2, showed some transcript expression during seed development, despite a lack of GLS synthesis at the seed level. MAM2 does not exist in the Columbia-0 (Col-0) ecotype and does not have any publicly available expression profiles.

In addition to null mutants, we also obtained mutants with altered GLS composition in the Col-0 background. These mutants included gsm1, which accumulates C3 GLS and has large reductions in 4-methyl sulfanylbutyl and 6-methylsulfinyl glucosinolates (Haughn et al., 1991; Kroymann et al., 2001). Because the Col-0 accession does not contain MAM2 and has a truncated nonfunctional AOP2 protein (Kroymann et al., 2001; Wentzell et al., 2007; Jensen et al., 2015), we also analyzed a previously characterized AOP2 overexpression mutant in the Col-0 background that accumulates alken GLS (Rohr et al., 2009; Burow et al., 2015). Collectively, these mutants represent some of the potential GLS composition alterations that can occur in the Col-0 background. The ability of any single gene mutant to capture the diversity of GLS is limited because it arises from a complex allelic combination (Kliebenstein et al., 2001).

We quantified the dry seed FAA for each of these single gene mutants and then assessed the fold change (FC), as compared with its respective wild-type control (Col-0 or Col-3), for 16 Gln-related traits (Supplemental
Table 2. QTL analysis of the (Gln)-related traits from the Bay × Sha mapping population

Only QTL that span the GS-ELONG and/or GS-AOP region are shown. For full analysis, see Supplemental Dataset S2. Genotypic and phenotypic data from 158 recombinant inbred lines were analyzed using the R/qtl2 package to identify significant QTLs. Asterisks indicate traits with significant SNP-trait associations in our GWAS.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Chr</th>
<th>Marker Name</th>
<th>Marker Position</th>
<th>Supporting Interval</th>
<th>Supporting Interval</th>
<th>Log of the Odds</th>
<th>Phenotypic Variance</th>
<th>No. GSL Genes in Supporting Interval</th>
<th>Gene Name</th>
<th>Gene Location</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Q/RQ</em></td>
<td>5</td>
<td>MSAT5.14</td>
<td>26.6</td>
<td>7.499</td>
<td>17.7–45.4</td>
<td>4.67–13.96</td>
<td>3.49</td>
<td>9.67</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AT5G23020 (MAM3)</td>
<td>7718118–7721866</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT5G26000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td><em>Q/RQP</em></td>
<td>5</td>
<td>MSAT5.14</td>
<td>26.6</td>
<td>7.499</td>
<td>17.7–41.8</td>
<td>4.67–12.11</td>
<td>7.82</td>
<td>20.38</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AT5G23020 (MAM3)</td>
<td>7718118–7721866</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT5G26000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td>Q/R</td>
<td>5</td>
<td>MSAT5.14</td>
<td>26.6</td>
<td>7.499</td>
<td>17.7–45.4</td>
<td>4.67–13.96</td>
<td>2.80</td>
<td>7.84</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td><em>Q/RP</em></td>
<td>5</td>
<td>MSAT5.14</td>
<td>26.6</td>
<td>7.499</td>
<td>17.7–41.8</td>
<td>4.67–12.11</td>
<td>5.97</td>
<td>15.97</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td>Q/QP</td>
<td>4</td>
<td>MSAT4.43</td>
<td>10.7</td>
<td>2.576</td>
<td>2–24.2</td>
<td>0.41–7.55</td>
<td>3.97</td>
<td>10.93</td>
<td>AT4G03050 (AOP3)</td>
<td>1343845–1346436</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT4G03060 (AOP2)</td>
<td>1351688–1354096</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT4G03063 (AOP1)</td>
<td>1355667–1356018</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT4G03070 (AOP1)</td>
<td>1358267–1359698</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT4G12030</td>
<td>7210807–7213376</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q/QP</td>
<td>5</td>
<td>MSAT5.14</td>
<td>26.6</td>
<td>7.499</td>
<td>17.7–41.8</td>
<td>4.67–12.11</td>
<td>15.79</td>
<td>36.89</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td><em>Q/P</em></td>
<td>5</td>
<td>MSAT5.14</td>
<td>26.6</td>
<td>7.499</td>
<td>17.7–41.8</td>
<td>4.67–12.11</td>
<td>12.10</td>
<td>29.72</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td><em>Q/P</em></td>
<td>4</td>
<td>MSAT4.43</td>
<td>10.7</td>
<td>2.576</td>
<td>2–15.8</td>
<td>0.41–5.63</td>
<td>5.88</td>
<td>15.75</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
<tr>
<td><em>Q/P</em></td>
<td>5</td>
<td>MSAT5.14</td>
<td>26.6</td>
<td>7.499</td>
<td>17.7–41.8</td>
<td>4.67–12.11</td>
<td>12.10</td>
<td>29.72</td>
<td>AT5G23010 (MAM1)</td>
<td>7703041–7706973</td>
</tr>
</tbody>
</table>
Dataset S4A). Gln absolute levels in the aop1, aop3, and AOP2-OX mutants did not change significantly. An elevated amount of Arg in the aop3 mutant led to reductions in two Gln-related traits, Q/R and Q/RQ (0.54 and 0.75 FC, respectively; Fig. 5; Supplemental Table S5A; Supplemental Dataset S4B). In addition, Glu and Pro were reduced in the AOP2-OX mutant but did not lead to any significant changes in the Gln-related ratios (Fig. 5B; Supplemental Table S5B). The bsu1 mutant had significantly high levels of Arg and Glu (a 1.62 and 1.43 FC, respectively), but the levels of Gln and related ratios were unchanged (Fig. 5; Supplemental Table S5B). The FAA quantifications of the AOP-related mutants showed that, in addition to minor alterations in the Glu family FAAs, few other FAAs changed significantly (Fig. 5A; Supplemental Table S5B). Our analysis of the MAM-related mutants showed that levels of Gln, Glu, and Pro were slightly elevated (a 1.39, 1.19, and 1.35 FC, respectively) in the gsm1 mutant, which led to slight increases in nine traits Gln related ratios (Fig. 5B; Supplemental Table S5). In sum, the single gene mutants showed only a small effect of the altered GLS composition on the Gln-related traits.

Elimination of Aliphatic GLS Triggers a Strong Seed-Specific Increase in Free Gln

To further characterize the association between aliphatic GLS and the Gln-related traits, we quantified the absolute levels of each FAA in the dry seeds of three null mutants (myb28/29, myb34/51, and gtr1/2) with altered GLS compositions and the Col-0 ecotype. The log2 of the average FC, defined as the ratios between individual amino acid levels in the mutants and their levels in their respective controls, were calculated and used to create heat maps of the FAAs (Fig. 6; Supplemental Dataset S4). The myb28/29 double knockout mutant is a null mutant of two transcription factors that regulate the aliphatic GLS in Arabidopsis: MYB28 (AT5G61420) and MYB29 (AT5G07690). This double knockout eliminates all aliphatic GLS from the entire plant, including the seed (Sonderby et al., 2007). A double knockout of GTR1 (AT3G47960) and GTR2 (AT5G62680), resulting in the gtr1/2 mutant, abolishes the transport of all GLS to the seeds (Nour-Eldin et al., 2012). Finally, a double knockout of the two transcription factors MYB51 (AT1G18570) and MYB34 (AT5G60890), resulting in the myb34/51 mutant, eliminates the indole GLS from the entire plant (Frerigmann and Gigolashvili, 2014).

The FAA analysis revealed that Gln levels were significantly higher in the myb28/29 and gtr1/2 mutants, but not in the myb34/51 mutant, as compared with Col-0 (Fig. 6; Supplemental Table S5A; Supplemental Dataset S4A). In fact, Gln showed the most pronounced FC among all FAAs measured: a 97 FC in the myb28/29 mutant and a 598 FC in the gtr1/2 mutant (Fig. 6; Supplemental Table S5). In addition to Gln, three other Glu family members increased significantly in the myb28/29 and gtr1/2 mutants: a 35.1 and 64.5 FC...
for Arg, a 3.3 and 4.7 FC for Glu, and a 1.3 and 4 FC for Pro, respectively (Supplemental Table S5, A and B). Alterations in these Glu family FAAs led to significant FC increases in all Gln-related ratios, ranging from a 1.5 to 1.9 FC in Q/\(R_Q\) and a 76.3 and 150.7 FC in \(Q/P\) in the \(myb28/29\) and \(gtr1/2\) mutants, respectively (Fig. 6B; Supplemental Table S5A). In the \(myb28/29\) and \(gtr1/2\) mutants, we also observed increases in Asn (10.40 and 9.87 FC, respectively) and His (8.78 and 47.28 FC, respectively). Glu and Asp also showed a consistent elevation (~3–5 FC) in both mutants (Fig. 6A; Supplemental Table S5B). The total sum of the FAAs (TFAA) measured also increased significantly in both \(myb28/29\) and \(gtr1/2\), by 4.73 and 12.58, respectively (Supplemental Table S5B).

Because TFAA changed in both mutants, we also calculated the percent of each FAA to the sum of the TFAA measured in all genotypes including Col-0 (Supplemental Dataset S4C; Supplemental Table S5C). In both mutants, the largest increase was in the relative composition of Gln, which increased from ~1% in Col-0 to 22.82% in the \(myb28/29\) mutant and to 53.10% in the \(gtr1/2\) mutant (Fig. 6C; Supplemental Table S5C). Arg and His were the only other FAAs that consistently increased in both the \(myb28/29\) and \(gtr1/2\) mutants: from ~1% of the total FAA in Col-0 to 8.82% and 6.10%, respectively, for Arg and to 2.44% and 4.95%, respectively, for His. The relative compositions of the remaining FAAs were consistently lower in both mutants (excluding Asn, which showed opposite trends in the two mutants; Fig. 6C; Supplemental Table S5C). The largest decreases were in the two most abundant FAAs in the Col-0 seeds, Glu and Gly, which had relative abundances of 28.81% and 18.77% in Col-0, 19.94% and 10.65% in \(myb28/29\), and 6.66% and 2.83% in \(gtr1/2\), respectively (Fig. 6C; Supplemental Table S5C).

Next, we tested whether a reduction in GLS (rather than its complete elimination) would result in significant alterations in Gln levels. We quantified the dry seed FAA levels from the \(myb28\) and \(myb29\) single mutants, which have approximately half the seed GLS as the Col-0 ecotype (Francisco et al., 2016). The \(myb28\) mutant had significant FCs only in Pro levels (a 1.23-FC increase; Supplemental Table S5, A and B). The \(myb29\) mutant, by contrast, showed minor but significant increases in both Gln absolute levels (1.55 FC) and relative composition (Gln/Total 1.26 FC) as well as FCs (1.7–1.47) in several Gln-related traits (i.e. Q/REP, Q/E, Q/P, Q/RE, Q/QE, Q/QP, Q/EP, Q/RQE, Q/QEP, Q/RQEP) in the \(myb29\) mutant (Fig. 6B; Supplemental Table S5A). Nevertheless, levels of Asp, Gly, Leu, and Phe were also elevated significantly in this mutant, with FCs of 1.23 to 1.42 (Fig. 6A; Supplemental Table S5B).

Collectively, this genetic analysis indicated to us that
Gln levels were extensively altered in response to a complete absence of aliphatic GLS either in the plant or specifically in the seed.

To evaluate if the response was seed specific, we analyzed the FAA content in the rosette leaves and stems of the myb28/29 and gtrl/2 double mutants and the respective Col-0 control. Tissues were collected ~20 d after bolting to capture the metabolic steady state of the FAA in these tissues during seed setting and filling. Neither mutant had significant fold changes in Gln levels in either its leaves or stems (Supplemental Table S6; Supplemental Dataset S5). In contrast with the seeds, we also found no elevation in TFAA (as explained above) in either mutant. The results support the genetic evidence that the elevated Gln levels in the mutant seeds are occurring at the seed level rather than resulting from specific increases in the maternal tissue.

### Gln Levels Are Elevated During Early Seed Maturation in Both the myb28/29 and the gtrl/2 Mutants

During seed maturation, FAAs (especially Gln) are incorporated into the SSPs especially during seed filling/maturation (Fait et al., 2006). Hence, we assessed whether Gln levels are elevated during the early stages of seed development. To do this, we isolated developing seeds at 12, 14, 16, and 18 d after flowering (DAF) and at the dry seed stage from the myb28/29 and gtrl/2 mutants and the Col-0 ecotype and analyzed the FC in FAA levels across these time points (Supplemental Dataset S6). Our analysis indicated that, as compared with the Col-0 control, the seeds from both mutants had substantial increases in Gln as early as 12 DAF (Fig. 7; Supplemental Table S7). At 12 DAF, there was a 24-FC increase of Gln in the myb28/29 mutant and a 37-FC increase in the gtrl/2 mutant (Supplemental Table S7). Gln levels were higher across all the developmental time points in both mutants. Although Gln levels in all genotypes showed an overall reduction trend, the FC observed in the mutants continued to increase as the seed progressed to desiccation (Fig. 7, A and B; Supplemental Table S7). Gln absolute levels at all time points exceeded the levels of any other amino acid (Supplemental Dataset S6).

Because the TFAA changed in both mutants, we also evaluated the changes in FAA relative composition as described above. The relative composition of Gln dropped from 9.5% (12 DAF) to ~1.11% (dry seed) in the Col-0 and dropped from ~54.1% (12 DAF) to 22.82% (dry seed) in the myb28/29 mutant (Supplemental Table S7B). Surprisingly, the Gln content in the gtrl/2 mutant remained between 54.53% and 61.40% throughout the entire seed maturation process, despite a drop in Gln absolute levels (Fig. 7C; Supplemental Table S7B). Hence, Gln is only a minor amino acid in Col-0 but the most abundant one in the mutants. By contrast, Glu is most abundant in the seeds, and its levels increased from 21.3% (12 DAF) to 28.8% (dry seed) in the Col-0, remained constant at ~20% in the myb28/29 mutant throughout development, and decreased from 13.9% (12 DAF) to 10.6% (dry seed) in the gtrl/2 mutant (Supplemental Table S7B). Very pronounced changes were also recorded in the composition of Gly, which had a lower relative composition as compared with the Col-0 throughout seed development (Fig. 7C; Supplemental Table S7). Notably, at all seed developmental stages, the FC never exceeded 2 for Gly or 6 for Glu (Supplemental Table S7A).

Collectively, these results show that compositional alteration to FAAs in the glucosinolate mutants occurs very early in seed maturation and persists in the dry seeds.

### Both Sulfur and Nitrogen Significantly Changed in Seeds that Lacked GLS

GLS are high in nitrogen and sulfur compounds. A lack of GLS in seeds may cause a change in their homeostasis, which is known to have a substantial impact on Gln levels (Nikiforova et al., 2005; Nikiforova et al., 2006). To test this possibility, we measured nitrogen, carbon, and sulfur in the myb28/29 and gtrl/2 mutants and in the Col-0 control (Table 3). We found that, as compared with Col-0, nitrogen was higher in both mutants (by 8% and 15% [w/w], respectively), sulfur was significantly lower (by 79% and 90% [w/w], respectively), and carbon was unaltered (Table 3). Finally, we assessed whether the elevated levels of Gln and other FAAs reflected any changes in the levels or composition of proteins. To do this, we analyzed the protein-bound amino acids (PBAA) in the dry seeds of the two mutants and in Col-0. The analysis revealed no significant or consistent alterations in PBAA levels (Supplemental Table S8; Supplemental Dataset S7).

### DISCUSSION

GWAS have successfully uncovered many genes involved in the natural variation and regulation of various metabolic traits, including FAAs in seeds (Magrath et al., 1994; Parkin et al., 1994; Chan et al., 2011; Angelovici et al., 2013; Lipka et al., 2013; Diepenbrock et al., 2017). Yet, none of these studies have identified any significant SNP associations with free Gln in dry seeds. The intractability of this trait would suggest that Gln has a highly complex genetic architecture. When faced with such complex metabolic traits, some researchers have enlisted metabolic ratios based on a priori knowledge or unbiased network analysis, an approach that has yielded additional QTLs that could not be retrieved using direct measurements of the absolute traits (Angelovici et al., 2013; Angelovici et al., 2017; Diepenbrock et al., 2017). Unfortunately for free Gln in seeds, neither absolute measurements nor specific metabolic ratios have resulted in significant associations.
In this study, we used a semicombinatorial approach to formulate metabolic ratios as traits in a mGWAS. Unlike previous studies, this approach yielded several novel SNP-trait associations. Interestingly, we identified unique SNP-trait associations across the different Gln-related traits, suggesting a slightly different genetic architecture for each metabolic ratio (Fig. 2; Supplemental Dataset 1). Because all the traits represent the Gln partition or a relationship to the other Glu family members, we treated all the SNPs as contributing to one genetic architecture of Gln metabolism. This collective analysis enabled us to compile a comprehensive candidate gene list that, upon further analysis, revealed a strong association between Gln and an unexpected metabolic pathway, the GLS biosynthesis. We argue that this approach could help elucidate the genetic basis of other complex metabolites and further reveal unexpected metabolic pathway associations.
Unexpected Association Between the Gln-Related Traits and the Aliphatic GLS Natural Diversity Is Supported by Multiple Independent Lines of Evidence

Our semicombinatorial mGWAS analysis revealed that the natural variation of the Gln-related traits measured from dry seeds is strongly associated with natural variation of aliphatic GLS. Not only did we identify an enrichment of GLS biosynthesis genes in our collective candidate gene list, but we also identified two aliphatic GLS biosynthetic genes in our top significant SNP-trait associations analysis (Table 1; Supplemental Table S2B). This association is surprising because GLS biosynthesis has three main steps (chain elongation of either Met, branched chain, or aromatic amino acids; core structure formation; secondary modifications; Kliebenstein et al., 2001), none of which involve Gln. In general, GLS are nitrogen- and sulfur-containing compounds that likely evolved from cyanogen glucosides but are largely limited to the Brassicales (Halkier and Gershenzon, 2006). Their breakdown products display a variety of biological activities explaining their defensive roles (Johnson et al., 2009). Although GLS accumulate to very high levels in seeds, they are synthesized in the vegetative tissue and transported from the maternal plant to the seed (Magrath and Mithen, 1993). Nevertheless, our study provides multiple lines of evidence confirming an association between the natural variation of Gln-related traits and the natural diversity of aliphatic GLS. First, it is important to note that the three significant SNPs associated with aliphatic GLS fell within two well-characterized QTLs: the GS-ELONG and the GS-AOP (Magrath et al., 1994). Previous studies have shown that the presence and absence of five genes within these QTLs account for much of the diversity in the aliphatic GLS profile in Arabidopsis. These genes are MAM1–3, AOP2, and AOP3 (Halkier and Gershenzon, 2006). Pairwise LD analysis of the three significant SNPs identified in these two regions revealed that these SNPs are likely tagging all five genes within these two key QTLs (Supplemental Figs. S2 and S3). Second, an independent QTL mapping of the Gln-related traits measured from the Bay/Sha mapping population (which segregates for these two key QTLs; Wentzell et al., 2007) also identified significant associations of both GS-ELONG and GS-AOP loci with several Gln-related traits (Table 2; Supplemental Dataset 2). Finally, the presence/absence of various chemotypes, arising from different allelic combinations of the MAM and AOP genes (Fig. 1), resulted in significantly different levels in the Gln-related traits (Supplemental Table S3). GLS 3ohp and 4ohb, in particular, showed strong associations with the Gln-related traits and are among the most abundant class of GLS in seeds (Petersen et al., 2002; Velasco et al., 2008). In addition, the aliphatic GLS are the most abundant GLS in Arabidopsis seeds (Kliebenstein et al., 2001). Interestingly, their precise function in this tissue is unclear. Taken together, our results show that, although unexpected, the pathway level association revealed by our mGWAS approach is strongly supported by multiple independent approaches.

The Nature of the Association Between the Gln-Related Traits and the GLS Natural Diversity Is Complex and Seed Specific

The precise nature of the association between GLS and the Gln-related traits is unclear. Our data indicate that the association is not simple. Analysis of known single gene mutants of the genes related to GLS in the GS-ELONG and GS-AOP regions in the Col-0 background (which lacks the expression of AOP2 and MAM2; Kroymann et al., 2001) showed relatively small changes in the Gln-related traits (Fig. 5; Supplemental Table S5). This finding is perhaps not surprising since GLS diversity relies on the presence of a complex epistatic interaction network of different GLS QTLs (Burow et al., 2010), and the ability of a single gene elimination in a set genotypic background to capture all the potential allelic combinations is very limited. In addition, a reduction of about half of the aliphatic GLS

### Table 3. Nitrogen, carbon, and sulfur absolute levels measured from the dry seeds of Arabidopsis myb28/29 and gtr1/2 double mutants and the Col-0 ecotype (n = 5)

The table lists the averages of the absolute levels, the percentage of each absolute level in the mutants relative to Col-0, the percentage of increase or decrease, and the significance of the difference after using a Duncan’s Multiple Range Test, with different lowercase letters indicating significant differences at the α = 0.05 level.

<table>
<thead>
<tr>
<th>Element</th>
<th>Line</th>
<th>Average Absolute levels</th>
<th>Control</th>
<th>Difference</th>
<th>Duncan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>µg mg⁻¹, %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrogen</td>
<td>Col-0</td>
<td>34.15 ± 0.32</td>
<td>100</td>
<td>0</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>myb28/29</td>
<td>36.94 ± 0.48</td>
<td>108.17</td>
<td>8.17</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>gtr1/2</td>
<td>39.18 ± 0.19</td>
<td>114.73</td>
<td>14.73</td>
<td>a</td>
</tr>
<tr>
<td>Carbon</td>
<td>Col-0</td>
<td>525.14 ± 4.16</td>
<td>100</td>
<td>0</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>myb28/29</td>
<td>519.1 ± 5.18</td>
<td>98.85</td>
<td>−1.15</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>gtr1/2</td>
<td>508.76 ± 2.94</td>
<td>96.88</td>
<td>−3.12</td>
<td>a</td>
</tr>
<tr>
<td>Sulfur</td>
<td>Col-0</td>
<td>5.37 ± 0.03</td>
<td>100</td>
<td>0</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>myb28/29</td>
<td>1.15 ± 0.01</td>
<td>21.43</td>
<td>−78.57</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>gtr1/2</td>
<td>0.54 ± 0.02</td>
<td>9.96</td>
<td>−90.04</td>
<td>c</td>
</tr>
</tbody>
</table>
through single mutations in either the *myb28* or
*myb29* mutants (Francisco et al., 2016) did not result
in any large effects on the Gln-related traits (Fig. 6;
Supplemental Table S5). However, the elimination of all
GLS transported to the seeds in the *gtr1/2* double mu-
tant or removal of the aliphatic GLS in the *myb28/29*
from the entire plant had a profound effect on the
composition of all FAAs and most prominently on Gln
(Fig. 6; Supplemental Table S5). These findings
emphasize that the association between Gln and GLS relies
on a complete elimination of specific GLS in the seed.
This observation is further supported by our statistical
analysis of the association between levels of the Gln-
related traits and the presence/absence of specific
GLS in a natural population (Supplemental Table S3).
More importantly, lack of FAA alteration in the stem
and leaf measured from the double mutant clearly
showed that the association between GLS and Gln is
seed specific and is not the cause of a pleotropic effect
that could arise from a lack of GLS in the mother plant
or a direct interaction of the MYB genes with any Gln-
related pathway genes (Supplemental Table S6). In line
with our observation, a study of the perturbation of
aliphatic GLS biosynthesis in Arabidopsis showed mild
alteration in leaf FAA, including free Gln; in fact, the
study found that Gln levels in leaves slightly decreased
(Chen et al., 2012). Interestingly, our FAA analysis per-
duced during early seed maturation further indicated
that the response of Gln to the lack of GLS, especially
aliphatic, occurs early (Fig. 7; Supplemental Table S7).
Overall, this early seed-specific interaction strongly
suggests that both GLS and Gln have key functions in
seed metabolic homeostasis that are not manifested in the
vegetative tissues. Moreover, it also demonstrates that an
mGWAS of FAA in dry seeds can reveal associations of
biological processes taking place in early development.

The Association between Gln and GLS Is Likely Indirect
and Induced by Alterations in the Seed
Metabolic Homeostasis

The molecular mechanism that underlies the inter-
action between GLS and Gln in the seeds is not clear.
The Gln response appears to depend on the presence/
absence of aliphatic GLS that is manifested in a specific
tissue and is not dosage dependent. This suggests that
the interaction is likely indirect and is potentially me-
diated through alteration of signaling/sensing path-
ways or other aspects of cell metabolism. Consistently,
previous studies in Arabidopsis leaves have shown that
perturbation of the aliphatic GLS alter several proteins
and metabolites involved in various physiological
processes, including photosynthesis, oxidative stress,
hormone metabolism, and specific amino acids (Chen
et al., 2012). It also has been shown in Arabidopsis
specifically that indole GLS activation products can
interact with the conserved TIR auxin receptor to alter
auxin sensitivity (Katz et al., 2015). Furthermore, ex-
genous application of a specific aliphatic GLS (3ohp)
causes an alteration in root meristem growth in an array
of plant lineages, even those that have never been
reported to produce GLS (Malinovsky et al., 2017).
These authors have established that this response is due
to the interaction between GLS and the TOR pathway,
which is a key primary metabolic sensor that controls
growth and development and is conserved back to the
last common eukaryotic ancestor (Henriques et al.,
2014). These findings highlight the potential interac-
tions of aliphatic GLS with primary metabolism and a
conserved sensing mechanism. Consistent with these
observations, our data show that the presence of specific
GLS compounds has a significant effect on the levels of
the Gln-related ratios: 3ohp had a negative effect on most
of the Gln-related ratios, whereas 4ohb had the opposite
effect (Supplemental Table S3). These two GLS may
possibly interact with distinct conserved metabolic reg-
ulatory pathways that affect Gln metabolism.

Our data also indicate that the strong seed-specific
association between the Gln-related traits and GLS in
the seeds lacking aliphatic GLS (i.e. *myb28/29 and gtr1/2*)
may be induced due to substantial alteration in the
overall cell metabolic homeostasis. Our analysis of the
carbon, nitrogen, and sulfur contents of the two double
mutants lacking aliphatic GLS in seeds support this
hypothesis. The results show that carbon remains rel-
atively stable, whereas both the nitrogen and sulfur
homeostasis is severely altered: total sulfur is dramati-
cally decreased, and nitrogen is increased (Table 3).
GLS are compounds rich in both nitrogen and sulfur,
which are present in high levels in seeds. It was previ-
ously suggested that GLS may function as a sulfur
storage, due to the large induction of the GLS break-
down pathway during broccoli (*Brassica oleracea* var
*italic*) seed germination (Gao et al., 2014). Gln is also
known to increase upon both high nitrogen availability
and sulfur deficiency (Nikiforova et al., 2005; Nikiforova
et al., 2006). A study of sulfur starvation in Arabidopsis
seedlings showed that plants convert the accumulated
excess nitrogen into nitrogenous compounds, including
Gln (for review, see Nikiforova et al., 2006). Hence, it is
possible that the lack of stored sulfur in the form of GLS
in seeds may lead to sulfur deficiency, in turn leading to
an elevation in FAAs, especially Gln. It is worth men-
tioning that no coherent pattern of alteration of the PBAA
composition was observed in the *myb28/29* and the *gtr1/2*
mutants as compared with the Col-0 ecotype, indicating
that the elevation in Gln is not due to a lack of incorpo-
ration of Gln into SSP (Supplemental Table S8). The latter
finding further supports the conclusions that sulfur re-
duction is due mainly to GLS reduction and that the in-
teraction between the pathways is mediated through
signaling/sensing cascades that are induced in response to
the alterations to seed metabolic homeostasis.

CONCLUSION

In this study, we demonstrated that free Gln in Ara-
bidopsis seeds is strongly affected by glucosinolate
diversity and presence in this organ. This finding clearly highlights that the presence of specific secondary metabolites can profoundly affect primary metabolism in seeds and that selected specialized metabolites may play a larger role in the metabolic homeostasis of this tissue than originally believed. Evolutionary theory predicts that the diversity and composition of plant defense compounds, such as the glucosinolates, in the different plant tissues reflect past selection pressures imposed on plants by their environment (Jones and Firn, 1991), pressures that are believed to be key driving forces of compound diversity and composition (Benderoth et al., 2006). Our study supports this claim and further suggests that the GLS effect on core metabolism may have played a role in shaping its diversity and composition; further studies are needed to reveal the extent of this phenomenon and its implication for seed fitness. Our study also aligns with previous work that has shown that defense mechanisms, such as GLS, although evolutionarily more recent and often species- and taxa-specific, have established connections with conserved regulatory/signaling pathways involved in core metabolism and other essential cellular processes. The latter was suggested to be evolutionarily advantageous in helping plants coordinate both defense metabolism and growth (Malinovsky et al., 2017). Finally, this study demonstrates that performing a semi-combinatorial ratio based mGWAS using metabolites measured in dry seeds can capture events occurring early in seed development. This finding has practical implications for future metabolic analyses since it is easier to perform an mGWAS on dry seeds than on developing seeds.

MATERIALS AND METHODS

Plant Growth and Seed Collection
All Arabidopsis (Arabidopsis thaliana) genotypes were grown at 22°C/24°C (day/night) under long-day conditions (16-h of light/8-h of dark). Growth of the Arabidopsis diversity panel (Nordborg et al., 2005; Platt et al., 2010; Horton et al., 2012) was as described (Angelovici et al., 2013).

Seed and Tissue Collection
Developing siliques were marked to track their developmental stage. Siliques were lyophilized, and the seeds were isolated and ground for the metabolic analysis.

Isolation of T-DNA Insertion Mutants and Genotypic Characterization


combinations with the T-DNA left border primer. Primers spanning the full-length transcript were used to confirm lack of transcripts for respective genes. The list of primers can be found in Supplemental Table S4.

The AOP2 overexpression line (Barlow et al., 2015), the myb28 and myb29 single mutants, the myb28/29 and myb34/51 knockout mutants (Sonderby et al., 2010; Frerichs and Gigué, 2011), and the GSK1 mutant (Haughn et al., 1991) were provided by Dr. Dan Kleißenstein from the University of California, Davis. The GLS transporter mutant gtr1/2 (Nour-Eldin et al., 2012) was provided by Dr. Hussam Hassan Nour-Eldin from Copenhagen University.

Transcript Analysis

Total RNA extracted from dry and developing seeds was isolated using a hot borate method (Birtic and Kraner, 2006) and purified using Direct-zol RNA Miniprep Plus filter columns (Zymo Research). Total RNA from leaves was extracted using the Direct-zol RNA Miniprep Plus Kit (Zymo Research). First-strand complementary DNA was synthesized from 1 μg of purified, total RNA using the first-strand cDNA Synthesis Kit (Bio-rad). RT-PCR was used to determine transcript levels using the Dream Taq Green Master Mix (Thermo Fisher Scientific) with ACTIN2 (AT3G18790) as a control. (For primers, see Supplemental Table S4.)

Data Analysis and mGWAS

mGWAS was performed using the genotypic data from the 250K SNP chip that was performed on the Arabidopsis diversity panel (Horton et al., 2012). SNPs with a minor allele frequency <0.05 were removed, leaving 214,052 SNPs for our association mapping. The 14 derived ratios and Gln absolute value and relative composition were used as trait inputs. The association tests were conducted in R 3.3.2 (R Core Team, 2014) using the R package Fixed and Random model Circulating Probability Unification (FarmCPU; Liu et al., 2016). The Bonferroni correction was used to control the experiment wise type I error rate at α = 0.01.

Statistical Analysis of the Gln-Related Trait Levels in Accessions Harboring Specific GLS

A subset of 133 accessions from the 313-accession population were grown in single replicates and analyzed for FAA and GLS contents. Accessions were grouped based on the presence or absence of four GLS chemotypes: 3-hp, 2-propenyl, 4ohb, and 3-butenyl/OH-3-butenyl. A/3 was used to test for the presence/absence of the group based on levels of Gln-related traits. Because sample sizes of some groups were small, a 1,000 permutations procedure was conducted to determine statistical significance at α = 0.05 (Churchill and Doerge, 1994).

GO Enrichment

For the candidate genes associated with the SNPs, a GO enrichment analysis was performed using agrigo (Du et al., 2010) with the following parameters: a hypergeometric test with an α = 0.05 FDR correction (γ = 3), an agrigo suggested background, Arabidopsis as the select organism, and a complete GO ontology.

Correlation Analysis

A Spearman’s rank correlation was used to calculate an r correlation matrix between all raw absolute levels of FAA in the Glu family after one round of outlier removal was performed in R. A P value was calculated to reflect the significance of each correlation. Results were filtered using a r 2 = 0.1 threshold and a P value = 0.001. Results were visualized with Cystoscope (Shannon et al., 2003) using the method previously described in Batushansky et al. (2016).

Haplotype Analysis and Pairwise LD Analysis

Because the average LD in Arabidopsis is 10 kb (Kim et al., 2007), the haploblock analysis was performed on a 10-kb window, where the 5 kb to the left and right of each significant SNP were used as inputs. The haploblock analysis was completed using Haploview version 4.2 (Barrett et al., 2004). Pairwise LD values (r 2) were calculated between the significant SNP of interest and

Downloaded on January 2, 2021. - Published by https://plantphysiol.org
Copyright (c) 2020 American Society of Plant Biologists. All rights reserved.
neighboring (upstream and downstream) SNPs in a ±4−/−5 kb window. All SNPs were filtered at a 5% minor allele frequency <0.05 were removed, leaving 214,052 SNPs for our association mapping. The 14 derived ratios and Gln absolute value. Default Gabriel block parameters were used, resulting in blocks that contained at least 95% of the SNPs in strong LD. Any genes contained, or partially contained, in the haploblock with a significant SNP were saved as putative genes of interest. If no haploblock was identified for a respective SNP, then the genes immediately upstream and downstream of the SNP were saved.

**QTL Analysis**

A QTL analysis was conducted on the 158 RILs from the Bay × Sha population (Loudet et al., 2002) in R 3.3.2 (R Core Team, 2014) using the R/qtl2 package (Broman et al., 2014). Data were previously quantified and described in Angelovic et al. (2013). Publicly available genotype markers spanned the five chromosomes for a total of 69 markers. The `genotype` function in R was used to calculate genotype probabilities using a step value and an assumed genotyping error rate of 0.001. Missing values were replaced with the most probable values using the `fill.geno` function; unsupervised cofactor selection was completed through backward elimination. Genome-wide LOD significance thresholds were determined using 1,000 permutations for each trait. For each QTL, confidence intervals were determined by a 1.5-LOD drop from peak marker. Percent variance explained was calculated using the following formula: Percent variance explained = 1 − 10−(2/LOD), where n is the sample size. Epistatic interactions were explored using the `effectplot` function.

**Metabolic Analyses**

**Amino acid analysis**

Amino acids were analyzed from four biological replicates (n = 3–4) for seed tissues and from five biological replicates (n = 5) for leaf and stem tissues. The FAA were extracted from −3 mg of dry seed by performing acid hydrolysis (200 μL 6 N HCl at 110°C for 24 h) followed by the FAA extraction method described in Yobi and Angelovic, 2018. FAA were extracted with 1 mM PBAAs were extracted from tissues and from seeds digested with a concentrated HNO3 with an internal standard. Seeds were harvested at five time points during seed maturation in the Bay × Sha population. Amino acid analysis was carried out by plotting against external standards for both molecules. Amino acids were analyzed from four biological replicates (n = 5) for leaf and stem tissues. The FAA were extracted from −3 mg of dry seed by performing acid hydrolysis (200 μL 6 N HCl at 110°C for 24 h) followed by the FAA extraction method described in Yobi and Angelovic, 2018. FAA were extracted with 1 mM PBAAs were extracted from tissues and from seeds digested with a concentrated HNO3 with an internal standard. Seeds were harvested at five time points during seed maturation in the Bay × Sha population. Amino acid analysis was carried out by plotting against external standards for both molecules.

**Glu analysis**

GLS identification and quantification were completed using HPLC with diode-array detection (HPLC/DAD), as described previously (Kliebenstein et al., 2001).

**Nitrogen and carbon analyses**

Nitrogen and carbon levels were determined using an ECS 4010 CHNS analyzer (Costech Analytical Technologies), following the instructions in the manufacturer’s manual. Briefly, ~2 mg tissue from five biological replicates (n = 5) were placed in tin capsules (Costech Analytical Technologies) and analyzed along with 2 mg of 2,5-bis-(3-tert-butylbenzoxazolyl)thiophene (Thermo Fisher Scientific) as an internal standard. Helium was used as a carrier gas, and separation was performed on a GC column maintained at 110°C. Detection was based on thermal conductivity, and quantification was carried out by plotting against external standards for both molecules.

**Sulfur analysis**

Sulfur measurements were performed using the procedure described in Ziegler et al. (2013). Briefly, ~50 mg of seed from six biological replicates (n = 6) were digested with a concentrated HNO3 with an internal standard. Seeds were soaked at room temperature overnight and then incubated at 105°C for 2 h. After cooling, the samples were diluted and analyzed with an Elan 6000 DRC (dynamic reaction cell) mass spectrometer (PerkinElmer SCIX) connected to a Perfuoroalkoxy (PFA) microflow nebulizer (Elemental Scientific) and Apex HF desolvator (Elemental Scientific) as described in (Ziegler et al., 2013).

**Accession Numbers**

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession numbers: At4g03070 (AOP1), At4g03060 (AOP2), At4g03050 (AOP3), At1g23010 (MAM1), At5g23020 (MAM3), At4g03063 (BSU1), At5g61420 (MYB28), At5g07690 (MYB29), At3g47960 (GTR1), At5g61650 (GTR2), AT1G18570 (MYB51), AT5G6890 (MYB34).

**Supplemental Data**

The following supplemental materials are available.

**Supplemental Figure S1.** Correlation analysis of the four Glu family FAAs.

**Supplemental Figure S2.** Pairwise LD estimates (r2) of S127076 with SNPs spanning all the GLS-related genes in the region.

**Supplemental Figure S3.** Pairwise LD estimates (r2) of S175365 with SNPs spanning all the GLS-related genes in the region.

**Supplemental Figure S4.** Interaction plots from QTL associated with the G5-AOP and G5-ELONG loci showing epistatic interactions between markers MSAT4.43 and MSAT5.14 for traits Q/P and Q/QP, respectively.

**Supplemental Figure S5.** AOP1, AOP3, MAM3, and BSU1 genomic structures and transcripts levels in tissues of high expression from the respective knockout mutants.

**Supplemental Figure S6.** Heatmap of expression data for AOP1, AOP2, AOP3, MAM1, MAM3, and BSU1 across various Arabidopsis tissues and seed developmental stages.

**Supplemental Table S1.** Mean, relative content, so, relative so, range and broad-sense heritability of the Glu family and the Gln-related traits.

**Supplemental Table S2.** List of all unique genes that are included in the haploblocks spanning all the identified significant SNPs.

**Supplemental Table S3.** Statistical analysis summary of the Gln-related levels from the 133 accessions grouped by GLS chemotypes.

**Supplemental Table S4.** A list of primers used to test for the expression of the full transcript of relevant genes involved in GLS biosynthesis in Arabidopsis.

**Supplemental Table S5.** Summary of the FC in the average FAA and Glu-related traits measured from the dry mutant seeds as compared to the control (Col-0).

**Supplemental Table S6.** Summary of the FC in the stem and leaf FAA average levels measured from the myb28/29 and gtr1/2 mutants as compared to the Col-0.

**Supplemental Table S7.** Summary of the FAA measured from seeds harvested at five time points during seed maturation in the myb28/29 and gtr1/2 mutants and Col-0 ecotype.

**Supplemental Table S8.** Summary of the FC in the PBAAs absolute level averages (nanomole per milligram) in myb28/29 and gtr1/2 mutant dry seeds.

**Supplemental Dataset S1.** mGWA results summarizing all significant SNP-trait associations identified at a 1% Bonferroni level from the 16 Gln-related traits.

**Supplemental Dataset S2.** The full QTL analysis of the 16 Gln-related traits from the Bay × Sha mapping population.

**Supplemental Dataset S3.** Summary of Gln-related traits and GLS measurements from the dry seeds in the 133-accession Arabidopsis panel.

**Supplemental Dataset S4.** Summary of the seed FAA quantification from mutants of relevant genes involved in the natural variation of aliphatic GLS in the Col-0 background.

**Supplemental Dataset S5.** Summary of the FAA absolute level measurements from the leaf and stem tissues of Arabidopsis myb28/29 and gtr1/2 mutants and the Col-0 ecotype.

**Supplemental Dataset S6.** Summary of the FAA measured from seeds harvested at five time points during seed maturation in the myb28/29 and gtr1/2 mutants and the Col-0 ecotype.
Supplemental Dataset S7. Summary of the protein-bound amino acid (PBA) absolute levels (nmol/mg) measured from the dry seeds of the myb28/29 and gtr1/2 mutants.

ACKNOWLEDGMENTS

The authors acknowledge Melody Kroll for assistance with editing the manuscript and Ivan Baxter and Greg Ziegler for sulfur analysis.

Received January 15, 2020; accepted March 26, 2020; published April 21, 2020.

LITERATURE CITED


