Genomic prediction informed by biological processes expands our understanding of the genetic architecture underlying free amino acid traits in dry *Arabidopsis* seeds

Sarah D. Turner-Hissong[*], Kevin A. Bird[*], Alexander E. Lipka[†], Elizabeth G. King[*], Timothy M. Beissinger[‡,§], Ruthie Angelovici[*]

[*] Division of Biological Sciences, University of Missouri, Columbia, MO, USA

[†] Department of Crop Sciences, University of Illinois at Urbana-Champaign, IL, USA

[‡] Division of Plant Breeding Methodology, Department of Crop Science, Georg-August-Universtät, Göttingen, Germany

[§] Center for Integrated Breeding Research, Georg-August-Universtät, Göttingen, Germany

Short title: Biological pathways inform prediction of free amino acids in seeds

Keywords: amino acids, genomic prediction, MultiBLUP, complex traits, *Arabidopsis*

Corresponding author:

Ruthie Angelovici
371D Christopher S. Bond Life Sciences
University of Missouri
Columbia, MO 65201
Phone: (573) 882-3440
E-mail: angelovicir@missouri.edu

ABSTRACT

Plant growth, development, and nutritional quality depends upon amino acid homeostasis, especially in seeds. However, our understanding of the underlying genetics influencing amino acid content and composition remains limited, with only a few candidate genes and quantitative trait loci identified to date. Improved knowledge of the genetics and biological processes that determine amino acid levels will enable researchers to use this information for plant breeding and biological discovery. Towards this goal, we used genomic prediction to identify biological processes that are associated with, and therefore potentially influence, free amino acid (FAA) composition in seeds of the model plant *Arabidopsis thaliana*. Markers were split into categories based on metabolic pathway annotations and fit using a genomic partitioning model to evaluate the influence of each pathway on heritability explained, model fit, and predictive ability. Selected pathways included processes known to influence FAA composition, albeit to an unknown degree, and spanned four categories: amino acid, core, specialized, and protein metabolism. Using this approach, we identified associations for pathways containing known variants for FAA traits, in addition to finding new trait-pathway associations. Markers related to amino acid metabolism, which are directly involved in the FAA regulation, improved predictive ability for branched chain amino acids and histidine. The use of genomic partitioning also revealed patterns across biochemical families, in which serine-derived FAAs were associated with protein related annotations and aromatic FAAs were associated with specialized metabolic pathways. Taken together, these findings provide evidence that genomic partitioning is a viable strategy to uncover the relative contributions of biological processes to FAA traits in seeds, offering a promising framework to guide hypothesis testing and narrow the search space for candidate genes.

45 **INTRODUCTION**

46 Amino acids play a central role in plant growth and development, as well as human and
47 animal nutrition. In addition to serving as the building blocks for proteins, amino acids are involved
48 in essential biological processes that include nitrogen assimilation, specialized metabolism,
49 osmotic adjustment, alternative energy, and signaling (Rai 2002; Araújo *et al.* 2010; Angelovici *et*
50 *al.* 2010, 2011; Wu and Messing 2014). Therefore it is no surprise that the homeostasis for absolute
51 levels and relative composition of the free amino acid (FAA) pool is complex and depends, at least
52 in part, on various factors such as allosteric regulation, feedback loops of key metabolic enzymes
53 in amino acid synthesis pathways, and the rate of amino acid degradation (Less and Galili 2008;
54 Jander and Joshi 2010; Hildebrandt *et al.* 2015; Huang and Jander 2017; Amir *et al.* 2018). Studies
55 have also demonstrated that core metabolism has a significant impact on FAA homeostasis. For
56 example, alteration of the interconversion of pyruvate and malate in tomato fruits caused a
57 reduction in FAAs related to aspartate (Osorio *et al.* 2013). In addition, processes related to protein
58 and specialized metabolism also influence FAA homeostasis, especially in vegetative tissues
59 (Hildebrandt *et al.* 2015; Barros *et al.* 2017; Huang and Jander 2017; Hirota *et al.* 2018;
60 Hildebrandt 2018). Together, these lines of evidence highlight that FAA homeostasis is likely
61 determined by orchestration of multiple processes. However, the relative contribution of each
62 process to FAA composition remains unclear.

63 The composition of the FAA pool is especially critical in dry seeds, since it ensures proper
64 desiccation, longevity, germination, and seed vigor (Angelovici *et al.* 2011; Galili *et al.* 2014).
65 Despite this, very little is known about the function, genetic architecture and regulation of FAAs.
66 The FAA pool comprises 1-10% of total seed amino acid content in maize (Muehlbauer *et al.*
67 1994; Amir *et al.* 2018) and ~7% in *Arabidopsis thaliana* (Cohen *et al.* 2014; Amir *et al.* 2018).

68 Although the relative size of the FAA pool is small, manipulation of FAAs in seeds can have a
69 substantial contribution to crop seed nutritional biofortification (Galili and Amir 2013).
70 Nevertheless, several studies have also shown that manipulation of specific FAAs can have
71 pleiotropic effects on growth and germination, indicating their metabolism is intertwined with
72 other key metabolic processes in the seeds (Galili and Amir 2013; Amir *et al.* 2018). Thus,
73 uncovering more about the relative influence of these metabolic processes can help tailor a more
74 effective approach to FAA manipulation and biofortification.

75 Like many other primary metabolites in dry seeds, FAAs are complex traits with extensive
76 variability and high heritability across natural populations. Multiple genome-wide association
77 (GWA) studies have identified several candidate loci for amino acid traits, both independently
78 (Riedelsheimer *et al.* 2012) and in conjunction with QTL studies (Angelovici *et al.* 2013, 2016).
79 However, the number and effect size of loci detected so far explained only a fraction of the
80 observed phenotypic variation for amino acid traits, with some traits proving harder to dissect than
81 others. For example, (Angelovici *et al.* 2013, 2016) found the strongest associations for traits
82 related to histidine and branched-chain amino acids (BCAAs), but weak signals for most other
83 FAA traits. The findings that amino acid traits frequently have several associated loci and that
84 these loci explain a small proportion of the genetic variation suggest a highly polygenic
85 architecture with many loci of small effect (Korte and Farlow 2013). Additional evidence for
86 metabolic traits indicates that, although many genetic markers may contribute to overall genetic
87 variation, many of these markers are preferentially located in genes that are connected to a
88 biological pathway(s) (Lango Allen *et al.* 2010).

89 While linkage mapping and GWA studies are typically underpowered to identify variants
90 that are rare and/or of small effect, genomic prediction methods perform well when traits are highly

complex (Meuwissen *et al.* 2001; Goddard *et al.* 2009; de Los Campos *et al.* 2013). Genomic prediction models are trained on a subset of individuals with genotypic and phenotypic data, enabling researchers to predict breeding values for genotyped individuals that have an unknown phenotype (Meuwissen *et al.* 2001; Heffner *et al.* 2009). Since its development nearly two decades ago (Meuwissen *et al.* 2001), genomic prediction has dramatically altered the speed and scale of applied genetic and breeding research (Daetwyler *et al.* 2013). The efficacy of genomic prediction results from its simultaneous use of all genotyped markers and indifference to the statistical significance of individual markers, in contrast to analyzing markers one-at-a-time for significance as is done for linkage mapping and GWA studies (Heffner *et al.* 2009). This allows the inclusion of information from all loci to make predictions, instead of basing conclusions only on loci that achieve genome-wide significance, and therefore captures more of the additive genetic variance.

One of the most widely used methods for prediction of complex traits is genomic best linear unbiased prediction (GBLUP) (Meuwissen *et al.* 2001), which assumes that all variants share a common effect size distribution. Recent extensions of the GBLUP model, such as MultiBLUP (Speed and Balding 2014), genomic feature BLUP (Edwards *et al.* 2015, 2016; Sarup *et al.* 2016; Fang *et al.* 2017), and BayesRC (MacLeod *et al.* 2016), incorporate genomic partitions as multiple random effects, allowing effect size weightings to vary across different categories of variants. These partitions can be derived from prior biological information, such as physical position, genic/nongenic regions, pathway annotations, and gene ontologies. Further, genomic partitioning is most successful when a given partition is enriched for causal variant(s) (Sarup *et al.* 2016), providing a framework for guided hypothesis testing. To this end, models that incorporate genomic partitioning have enabled researchers to determine the relative influence of genomic features (e.g. chromosome segments, exons) and/or biological pathways on the variance explained for complex

114 traits. For example, annotations for several biological pathways were used to determine which
115 pathways were associated with udder health and milk production in dairy cattle (Edwards *et al.*
116 2015). Similarly, gene ontology categories were leveraged to explore the genetic basis of different
117 phenotypes in *Drosophila melanogaster* (Edwards *et al.* 2016). In maize, applications of genomic
118 partitioning models have revealed that SNPs located in exons explain a larger proportion of
119 phenotypic variance compared to other annotation categories (Li *et al.* 2012) and that genomic
120 prediction is improved for multiple traits by incorporating information from gene annotations,
121 chromatin openness, recombination rate, and evolutionary features (Ramstein *et al.* 2020). The
122 inclusion of prior biological information from transcriptomics, GWA studies, and genes identified
123 *in silico* also improved predictions of root phenotypes in cassava (Lozano *et al.* 2017).

124 In this study, we evaluated genomic partitioning as a method to estimate the relative
125 contribution of metabolic pathway annotations to variation for FAA traits in dry seeds of
126 *Arabidopsis thaliana.* This approach enabled us to incorporate prior knowledge of FAA
127 biochemistry based on metabolic pathways and to identify annotation categories with a
128 disproportionate contribution to the genomic heritability of FAA content and composition. The
129 ultimate objective of this work was to discover metabolic pathway annotations that explained
130 significant variation and improved predictive ability, with the underlying assumption that the
131 corresponding genomic regions are important for determining seed metabolic associations and
132 constraints. These findings can then be considered in future designs to support seed amino acid
133 biofortification.

134 **MATERIALS AND METHODS**

135 **Plant materials and trait data**

136 For this study, we reanalyzed data of the absolute levels (nmol/mg seed), relative
137 compositions, and biochemical ratios for FAAs in dry *Arabidopsis thaliana* seeds (see **Table S1**
138 for a list of traits). These traits were previously measured by (Angelovici *et al.* 2013, 2016) for
139 313 accessions of the Regional Association Mapping panel (Nordborg *et al.* 2005; Platt *et al.*
140 2010). Seeds were obtained from the *Arabidopsis* Biological Resource Center (ABRC,
141 https://abrc.osu.edu/, see **Table S2** for stock numbers). The panel was grown in three independent
142 replicates, each at 18°C to 21°C (night/day) under long day conditions (16 h of light/8 h of dark).
143 Following the desiccation period, dry seeds were harvested and stored in a desiccator at room
144 temperature for at least six weeks prior to analysis to ensure full desiccation (Angelovici *et al.*
145 2013).

146 Absolute levels of FAAs (nmol/mg seed) were quantified using liquid chromatography–
147 tandem mass spectrometry multiple reaction monitoring (LC-MS/MS MRM; see Angelovici *et al.*
148 2013, 2016 for further details). Eighteen of the 20 proteinogenic amino acids were measured,
149 including composite phenotypes for the sum of all FAAs measured (total FAAs) and for each of
150 five biochemical families as determined by metabolic precursor (**Figure S1, Table S1**). This prior
151 knowledge of biochemical relationships among FAAs was also used to determine metabolic ratios,
152 which can represent, for example, the proportion of a metabolite to a related biochemical family
153 or the ratio between two metabolites that share a metabolic precursor (Sauer *et al.* 1999;
154 Weckwerth *et al.* 2004; Wentzell *et al.* 2007). The inclusion of metabolic ratios was based on
155 evidence from multiple studies, which reported novel or more significant associations when using
156 metabolic ratios as compared to absolute levels of metabolites (Wentzell *et al.* 2007; Harjes *et al.*

157 2008; Vallabhaneni and Wurtzel 2009; Wurtzel *et al.* 2012; Lipka *et al.* 2013; Gonzalez-Jorge *et*
158 *al.* 2013; Angelovici *et al.* 2013, 2016; Owens *et al.* 2014).

159 For each amino acid, relative composition was calculated as the absolute level over the
160 total. Additional ratio traits were determined based on biochemical family affiliation (Angelovici
161 *et al.* 2016). Traits and their respective abbreviations are described in **Table S1**. Overall, the 65
162 traits included 25 absolute FAA levels (individual amino acids and composite traits), 17 relative
163 levels (ratio of the absolute level for an amino acid compared to total FAA content), and 23 family-
164 derived traits (ratio of the absolute level for an amino acid to the total FAA content within a given
165 family).

166 Following the guidelines for multi-stage genomic prediction (Piepho *et al.* 2012), the best
167 linear unbiased estimates (BLUEs) for each accession were used as the phenotypic data in this
168 study and were calculated using the HAPPI-GWAS pipeline (Slaten *et al.* 2020a) in R v3.6.0 (R
169 Core Team 2016). First, outlier removal was performed by fitting a mixed effects model using the
170 'lmer' function in the 'lme4' package (v1.1-21, (Bates *et al.* 2015), with the raw trait values as the
171 response variable, replicate included as a random effect, and accession included as a fixed effect.
172 Studentized deleted residuals were then used to identify outliers (Kutner *et al.* 2004). Following
173 outlier removal, the Box-Cox transformation (Box and Cox 1964) was applied for each trait to
174 avoid violating model assumptions of normally distributed error terms and constant variance.
175 Finally, to remove phenotypic variability arising from environmental conditions, the BLUE for
176 each accession was obtained from the fitted mixed model described above, which was applied
177 across all three replicates. The BLUEs for each trait were used as the response variables in all
178 subsequent prediction models.

179 **Genetic data**

180 The accessions used in this study were previously genotyped using a 250k SNP panel
181 (v3.06, Atwell *et al.* 2010). The software PLINK (v1.9, Purcell *et al.* 2007) was used to filter for
182 minor allele frequency (MAF) > 0.05, reducing the number of SNPs from 214,051 to 199,452.
183 Principal component analysis was performed on this filtered SNP set using the 'prcomp' function
184 in R. The first two principal components explained 5.6% of the variance (**Figure S2**) and were
185 included as fixed covariates in the prediction models.

186

187 **Selection of pathway SNPs**

188 To examine specific metabolic pathways, SNPs were selected based on annotation
189 categories from the MapMan annotation software (Thimm *et al.* 2004) for the TAIR10 version of
190 *Arabidopsis* (Berardini *et al.* 2015). We focused broadly on 20 pathways, which spanned four
191 categories: amino acid metabolism (three pathways), core metabolism (three pathways),
192 specialized metabolism (five pathways), and protein metabolism (nine pathways) (**Table 1**), all of
193 which are known to be involved in FAA metabolism to some extent. The SNP positions were first
194 matched to the corresponding Ensembl gene id using the 'biomaRt' package (Durinck *et al.* 2005,
195 2009) in R. We then selected all SNPs within a 2.5 kb range of the start and stop position for each
196 gene, which is within the range of the estimated average intergenic distance in *Arabidopsis* (Zhan
197 *et al.* 2006) and includes upstream promoter regions. Relative SNP positions for each pathway are
198 provided in **Figure S3**. Pathways and MapMan annotation categories, including the number of
199 genes and SNPs represented, are described in **Table 1**. We used MapMan annotations for all genes
200 except BCAT2 (At1g10070), which was moved from the amino acid synthesis pathway to the

201 amino acid degradation pathway along with other SNPs in the same haploblock (chromosome 1,
202 3274080 to 397645 bp). This decision was based on previous work, which showed that *bcat2*
203 mutants accumulate higher levels of branched-chain amino acids in seeds, thereby demonstrating
204 that BCAT2 has catabolic activity (Angelovici *et al.* 2013).

205

Table 1. Summary of selected biological pathways.

Pathway	Number of genes[a]	Number of SNPs[a]	MapMan BINCODE
Amino Acid Metabolism			
amino acid synthesis	376	2084	13.1
amino acid degradation	160	1094	13.2
amino acid transport	144	939	34.3
Primary Metabolism			
glycolysis	148	858	4
TCA cycle	167	926	8
ATP synthesis (alternative oxidase)	10	66	9.4
Specialized Metabolism			
isoprenoids	269	1788	16.1
phenylpropanoids	161	845	16.2
nitrogen containing	39	229	16.4
sulfur containing	113	733	16.5
flavonoids	171	1062	16.8
Protein Metabolism			
amino acid activation	203	1231	29.1
protein synthesis	1383	7290	29.2
protein targeting	624	3689	29.3
protein posttranslational modification	1407	8794	29.4
protein degradation	996	6405	29.5
ubiquitin	2691	16000	29.5.11
protein folding	138	814	29.6
protein glycolysis	87	459	29.7
protein assembly	44	312	29.8

[a]Includes a 2.5 kb buffer before and after the start/stop position of each gene

206 **Prediction models**

207 The Linkage Disequilibrium Adjusted Kinship (LDAK) software (v5.0, Speed *et al.* 2012,
208 http://dougspeed.com/ldak/) was used to implement two models for genomic prediction of each
209 trait: GBLUP, which uses a single marker-based additive genetic relatedness matrix, and
210 MultiBLUP, which incorporates multiple marker-based additive genetic relatedness matrices, each
211 calculated with different subsets of genome-wide markers (Speed and Balding 2014).

212 Genomic prediction was performed for all markers (p = 199,452) using a GBLUP model,
213 in which individuals were included as a random effect and the additive genetic relatedness matrix
214 was used as part of the variance-covariance matrix among the individuals (Whittaker *et al.* 2000;
215 Meuwissen *et al.* 2001). First, the pairwise genetic similarity between individuals was estimated
216 using a genomic similarity matrix (GSM), or kinship matrix (VanRaden 2008; Astle and Balding
217 2009):

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}'/p, \tag{1}$$

218 where X is an n x p design matrix of SNP genotypes, X' is the transpose of X, n is the total number
219 of individuals, and p is the total number of markers. The GBLUP model was then fit with the
220 random effects model:

$$\boldsymbol{Y} = \mu + Zu + \varepsilon, \tag{2}$$

221 where Y is the vector of phenotypic values for n individuals, μ is the overall mean, Z is a design
222 matrix connecting observations to genotypes, u is the vector of random genetic effects distributed
223 as $u \sim N(0, K\sigma_G^2)$, and ε is the random error term distributed as $\varepsilon \sim N(0, I\sigma_e^2)$, where K is the

224 GSM representing the correlation structure of u, I is a $n \times n$ identity matrix, and σ_G^2 and σ_e^2 are
225 variances.

226 The MultiBLUP model (Speed and Balding 2014) was used to incorporate biological
227 pathway information into genomic prediction. As an extension of the GBLUP model, MultiBLUP
228 is a multi-kernel model that subdivides genetic effects into at least two random effects, where
229 different subsets of markers are used to calculate GSMs for each random effect. In this study, the
230 MultiBLUP model included a random genetic effect corresponding to sets of markers within a
231 single biological pathway (m) and a second random genetic effect corresponding to the remaining
232 markers not included in the given pathway ($\notin m$). Using notation from equation (2) and Speed and
233 Balding (2014), the MultiBLUP model within the context of this work is:

234

$$Y = \mu + Zu^m + Zu^{\notin m} + \varepsilon, \quad (3)$$

235

236 where u^m is the vector of random genetic effects distributed as $u^m \sim N(0, K^m\sigma_m^2)$, with K^m
237 representing the kinship matrix calculated using markers within a given biological pathway and
238 σ_m^2 denoting the corresponding variance component; $u^{\notin m}$ is the vector of random genetic effects
239 distributed as $u^m \sim N(0, K^{\notin m}\sigma_{\notin m}^2)$, with $K^{\notin m}$representing the kinship matrix calculated using
240 markers outside of the given biological pathway and $\sigma_{\notin m}^2$ denoting the corresponding variance
241 component; and $\boldsymbol{Y}$, Z, μ, and ε are as previously described.

242 For our purposes, kinship matrices were estimated using the LDAK software for either all
243 SNPs (GBLUP) or each SNP partition (MultiBLUP, i.e. separately for SNPs belonging to a single

244 pathway and all other remaining SNPs). For each pathway, the extent of collinearity due to LD
245 was determined by examining Spearman's rank correlation between the off-diagonal elements of
246 the kinship matrices for the pathway SNPs and remaining genomic SNPs. For estimates of genomic
247 heritability, values were constrained to be positive and less than one. Additionally, the parameter
248 α, which models the relationship between heritability and MAF, was set to α = 0 under the
249 assumption that SNPs with lower MAF contribute less to heritability than SNPs with higher MAF
250 (Speed *et al.* 2017). In relation to Eq. 2 and 3, the parameter α adjusts the expected contribution
251 of each SNP to heritability, with a value of α = -1 assuming that heritability does not depend on
252 MAF (see Speed et al. 2017 for details).

253

254 **Estimation of genomic heritability**

255 For both GBLUP and MultiBLUP, average information restricted maximum likelihood
256 (REML, see Speed and Balding 2014 for details) was used to compute variance component
257 estimates for σ_m^2, $\sigma_{\notin m}^2$, and σ_e^2. The maximum number of iterations to achieve convergence was
258 set to 500. This process was repeated for each trait and pathway combination. In the case of the
259 GBLUP model, $\hat{\sigma}_m^2$ is the estimate of variance for all SNPs. These estimates were used to calculate
260 genomic heritability as the ratio of additive genomic variance explained for a given marker set
261 (σ_m^2) over the total variance explained (the sum of σ_m^2, $\sigma_{\notin m}^2$,and the residual variance, σ_e^2):

262

$$h_m^2 = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{\notin m}^2 + \sigma_e^2}\,. \tag{4}$$

263

264 For the MultiBLUP model, the proportion of genomic heritability explained was calculated as:

265

$$\frac{h_m^2}{h_m^2 + h_{\notin m}^2}, \tag{5}$$

266

267 where h_m^2 is the genomic heritability explained by SNPs in a given genomic partition and $h_{\notin m}^2$ is
268 the genomic heritability explained by all other SNPs not included in the partition.

269

270 **Assessing model performance**

271 The performance of the prediction models was determined using ten-fold cross validation
272 with a one-fold holdout, with the same training and testing sets used for both the GBLUP and
273 MultiBLUP models. For each cross validation, the genomic estimated breeding value (GEBV) was
274 derived from marker data for the excluded individuals based on estimates of random genetic effects
275 for the individuals in the training set. This process was repeated five times for a total of 50 cross
276 validations per trait and pathway combination. Predictive ability was then calculated as $r(\hat{g}, g)$,
277 where $\hat{g}$ represents the GEBVs and g represents the BLUEs for each trait. Reliability, which is the
278 coefficient of determination (r^2) scaled by heritability, was calculated as $\frac{r^2}{h^2}$ (Rincent *et al.* 2012).
279 Bias was calculated as the simple linear regression slope estimate between the GEBVs and BLUEs
280 for each trait, with a slope estimate of one indicating no bias. Lastly, the overall root mean squared
281 error (RMSE), which measures prediction bias and variability, was calculated as the square root
282 of the mean for the squared difference between the BLUEs and GEBVs across all cross-

283 validations, $\sqrt{\frac{(g-\hat{g})_1^2+\cdots+(g-\hat{g})_k^2}{k}}$, where k is the number of cross-validations. Predictive abilities for
284 the MultiBLUP and GBLUP models were compared using a one-sided, paired Welch's t-test for
285 unequal variances.

286

287 **Generation of an empirical null distribution**

288 To test if a metabolic pathway explained more variation than expected by chance, we
289 generated an empirical null distribution for each trait and pathway combination. The null
290 hypothesis was that a given biological pathway will explain a similar amount of trait variance as
291 the same number of SNPs in randomly selected gene groups (Edwards *et al.* 2015). To establish a
292 null distribution, we first defined 1000 random gene groups for each pathway, where the target
293 number of SNPs in each random gene group was comparable to the size of the pathway. Ranges
294 for the number of genes and SNPs sampled for each pathway are provided in **Table S3**. For each
295 random subset, all SNPs within 2.5 kb of the start and stop positions were sampled for a randomly
296 selected gene. This process was repeated by randomly sampling genes one at a time until the target
297 number of SNPs for each subset was achieved. Genes within a given pathway were excluded from
298 the random sampling procedure for that pathway. As discussed in Edwards *et al.* (2015), this
299 approach does not explicitly model variation in other parameters (e.g., allele frequencies, LD), but
300 it is expected that these differences are captured to some extent by the sampling process.

301 Next, we used two metrics to test if SNPs in a given pathway explained more genomic
302 variance than expected by chance and increased model fit for each trait: (1) the proportion of
303 genomic heritability explained by a pathway compared to the random gene groups described
304 above, and (2) the likelihood ratio (LR) test statistic as a measure of pathway model fit compared

305 to the model fit of random gene groups. The proportion of heritability explained was calculated as
306 described previously in equation (5) and the LR test statistic was calculated as twice the difference
307 between the log likelihood of the MultiBLUP model and the log likelihood of the GBLUP model.
308 For each pathway and trait combination, the values for proportion of heritability explained and the
309 LR test statistic were compared to the empirical cumulative distribution function for the
310 corresponding 1000 random gene groups using the 'ecdf' function in R. To determine if the
311 observed value was greater than the random values for each metric, *P*-values were computed with
312 a one-sided test using the 't_test' function in the R package 'rstatix' (Kassambara 2020).

313

314 **Correction for multiple testing**

315 For each trait, the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) was
316 used to adjust for multiple testing across pathways (n = 20) at a 10% false discovery rate (FDR).
317 Multiple testing correction was performed with the 'p.adjust' function in R for the proportion of
318 heritability explained, the LR test statistic, and predictive ability.

319

320 **Identifying biological pathways of interest**

321 In summary, a pathway was considered of interest for a trait if the MultiBLUP model passed
322 all three of the following criteria:

323 1.) The proportion of heritability explained was significantly greater than empirical values for
324 random gene groups of the same size (FDR-adjusted *P*-value < .10),
325 2.) The LR test statistic was significantly greater than empirical values for random gene groups
326 of the same size (FDR-adjusted *P*-value < .10),
327 3.) The MultiBLUP model significantly improved predictive ability compared to the GBLUP
328 model (FDR-adjusted *P*-value < .10).

329 Together, criteria (1) and (2) established that a given pathway improved model fit better than a
330 random set of SNPs. Criteria (3) was imposed to ensure that there was a meaningful difference in
331 predictive ability when pathway information was incorporated via MultiBLUP compared to the
332 naive GBLUP model that incorporated no pathway information.

333

334 **Data availability statement**

335 Genotype data were previously published (Atwell *et al.* 2010) and were accessed from
336 github.com/Gregor-Mendel-Institute/atpolydb/wiki. The scripts and phenotypic data supporting
337 the conclusions of this article are publicly available as a Snakemake workflow (v5.4.2, Köster and
338 Rahmann 2012) on GitHub at github.com/mishaploid/aa-genomicprediction. **Table S2** details the
339 ABRC stock names and accession numbers for each individual.

340

341 **RESULTS AND DISCUSSION**

342 In this study, we applied a genomic partitioning model to evaluate the contribution of
343 metabolic pathways to FAA traits in seeds. The combination of a genomic partitioning framework
344 and the model system *Arabidopsis* allowed us both to test the feasibility of this approach and to
345 further examine the relative contribution of each pathway to the genetic basis of FAA traits in
346 seeds. Additionally, because FAA traits are part of core metabolism that is highly conserved, we
347 hypothesize that our findings can be used to develop hypotheses in crop systems, where there is
348 potential to contribute to the biofortification of essential amino acids.

349 **Genomic prediction was most effective for absolute levels of free amino acids**

350 We first established the efficacy of standard GBLUP in a diversity panel of 313
351 *Arabidopsis* individuals, which represents a substantial proportion of the known genetic variability
352 present in *Arabidopsis* (Nordborg *et al.* 2005). Because this setting is distinct from the closed
353 breeding populations of dairy cattle, maize, and other agricultural species where genomic
354 prediction is often applied (e.g. Heffner *et al.* 2009; Wolc *et al.* 2016; Weller *et al.* 2017), we were
355 interested in testing how well genomic prediction would work in this panel. We were also
356 interested in testing the utility of genomic prediction for FAA traits, which are highly conserved.

357 Using the GBLUP model, we observed low to moderate predictive ability for the amino
358 acid traits measured (**Table 2**). Of these 65 FAA traits, 30 had a predictive ability greater than 0.3
359 (**Figure 1, Table 2**). In general, prediction was effective for a greater number of absolute level
360 FAA traits, with 21 out of 25 absolute traits having a predictive ability > 0.3 (84%), compared to
361 relative levels (4 out of 17, 24%) and family-derived ratios (5 out of 23, 22%). The family ratio of
362 methionine (met_AspFam) had the highest predictive ability (r = 0.47), while the relative level of
363 serine (ser_t) had the lowest predictive ability (r = 0.08) (**Table 2**). The observation of moderate

prediction accuracies for many of these traits (**Figure 1, Table 2**) suggests that there is linkage disequilibrium (LD) between markers and causal loci, providing evidence that genomic prediction can be successfully applied in this system.

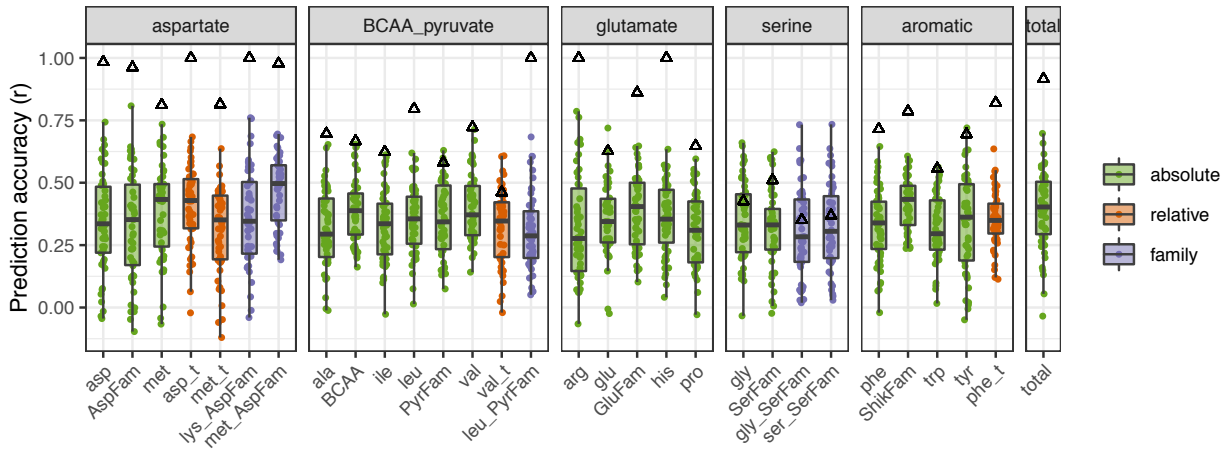


Figure 1. Genomic prediction performed well for a higher proportion of absolute traits compared to relative and family-based ratio traits.

Boxplots show free amino acid traits with predictive ability (r) > 0.3 based on genomic best linear unbiased prediction (GBLUP). Black triangles indicate the genomic heritability for each trait. Colors indicate whether the trait is an absolute level, relative level, or family-based ratio. Each point represents an individual cross-validation.

Table 2. Genomic prediction results for 65 free amino acid traits using a GBLUP model.

Trait type	Metabolic family	Trait	Predictive ability (r)		Reliability (r^2)		slope (bias)	RMSE
			mean	SE	mean	SE		
absolute	aspartate	asp	0.341	0.028	0.157	0.020	0.936	0.055
		met	0.384	0.027	0.224	0.024	1.051	0.0243
		thr	0.099	0.022	0.075	0.013	0.756	0.1348
		AspFam	0.327	0.031	0.159	0.022	1.042	0.0266
	BCAA_pyruvate	ala	0.317	0.022	0.178	0.021	1.119	0.249
		ile	0.328	0.021	0.206	0.023	1.046	0.0583
		leu	0.353	0.019	0.179	0.017	1.035	0.1261
		lys	0.270	0.024	0.122	0.017	0.921	0.4597
		val	0.397	0.019	0.242	0.022	1.027	0.1007
		BCAA	0.388	0.019	0.252	0.023	1.042	0.101
		PyrFam	0.351	0.021	0.249	0.026	1.108	0.12
	glutamate	arg	0.323	0.029	0.146	0.023	0.923	0.0552
		gln	0.178	0.025	0.114	0.018	1.032	0.5348
		glu	0.356	0.020	0.234	0.023	0.990	0.0072
		his	0.359	0.020	0.149	0.014	0.880	1.0529
		pro	0.310	0.021	0.182	0.021	1.010	0.0405
		GluFam	0.389	0.020	0.199	0.018	0.916	0.0297
	serine	gly	0.349	0.023	0.344	0.038	1.072	0.1086
		ser	0.241	0.021	0.142	0.019	1.078	0.0016
		SerFam	0.320	0.022	0.247	0.028	1.030	0.0112
	aromatic	phe	0.326	0.021	0.178	0.019	1.084	0.0202
		trp	0.317	0.018	0.209	0.021	1.019	0.0877
		tyr	0.334	0.027	0.212	0.026	1.046	0.0266
		ShikFam	0.411	0.015	0.229	0.016	1.023	0.0146
		Total	0.392	0.022	0.193	0.018	1.015	0.019
relative	aspartate	asp_t	0.405	0.022	0.189	0.017	0.933	0.0386
		met_t	0.313	0.025	0.157	0.017	1.021	0.0162
	BCAA_pyruvate	ala_t	0.261	0.026	0.154	0.022	1.239	5.1795
		ile_t	0.218	0.021	0.127	0.017	1.085	0.0197
		leu_t	0.296	0.022	0.122	0.015	1.093	0.0957
		lys_t	0.196	0.023	0.125	0.019	1.112	0.8788
		val_t	0.319	0.022	0.271	0.030	1.106	0.0138
	glutamate	arg_t	0.276	0.037	0.220	0.042	1.056	0.0145
		gln_t	0.108	0.022	0.118	0.019	1.322	5.0853
		glu_t	0.264	0.021	0.168	0.020	1.008	0.0355
		his_t	0.259	0.024	0.123	0.016	1.076	37.6486
		pro_t	0.253	0.019	0.134	0.017	1.022	0.0228
	serine	gly_t	0.268	0.026	0.567	0.082	1.127	0.0155

		ser_t	0.076	0.023	0.118	0.019	1.452	0.0185
	aromatic	phe_t	0.355	0.016	0.169	0.014	1.047	0.0181
		trp_t	0.205	0.024	0.110	0.015	0.940	0.0381
		tyr_t	0.116	0.027	0.146	0.018	1.112	0.0118
family	aspartate	asp_AspFam	0.141	0.031	0.134	0.022	1.309	0.0782
		ile_AspFam	0.165	0.029	0.139	0.021	1.197	0.0772
		lys_AspFam	0.358	0.027	0.164	0.021	0.933	0.0189
		met_AspFam	0.468	0.020	0.244	0.018	1.060	0.001
		thr_AspFam	0.118	0.024	0.090	0.013	1.049	0.0552
		AspFam_Asp	0.171	0.022	0.052	0.007	1.042	0.027
	BCAA_pyruvate	ala_PyrFam	0.216	0.019	0.092	0.013	0.905	0.0222
		ile_BCAA	0.250	0.020	0.083	0.011	0.914	0.0348
		leu_BCAA	0.251	0.024	0.091	0.012	1.076	0.075
		leu_PyrFam	0.303	0.021	0.114	0.015	0.975	0.0244
		val_BCAA	0.268	0.020	0.091	0.011	0.848	0.0224
		val_PyrFam	0.298	0.019	0.232	0.024	0.858	0.0153
	glutamate	arg_GluFam	0.205	0.032	0.193	0.029	1.243	0.0659
		gln_GluFam	0.167	0.034	0.195	0.039	1.076	0.0218
		glu_GluFam	0.139	0.022	0.153	0.022	0.992	1.1693
		GluFam_glu	0.203	0.034	0.202	0.030	0.881	0.0665
		his_GluFam	0.186	0.023	0.102	0.015	1.012	24.2851
		pro_GluFam	0.289	0.024	0.155	0.018	1.004	0.0329
	serine	gly_SerFam	0.305	0.025	0.351	0.048	1.172	0.0634
		ser_SerFam	0.325	0.024	0.364	0.047	1.179	0.0461
	aromatic	phe_ShikFam	0.218	0.028	0.149	0.021	1.097	0.0607
		trp_ShikFam	0.187	0.024	0.111	0.015	1.028	0.0658
		tyr_ShikFam	0.257	0.028	0.144	0.022	0.945	0.0331

SE, standard error; R*MSE*, root mean squared error.

373

374 **Annotations for biological pathways explained significant variation and improved predictive**
375 **ability of free amino acid traits in seeds**

376 We next applied a genomic partitioning approach, MultiBLUP, to investigate the
377 association of different metabolic annotation categories with FAA traits in dry *Arabidopsis* seeds.
378 The focus was specifically on categories which are thought to influence FAA homeostasis, but

379 where the degree of this influence is unclear, especially in dry seeds (Skirycz et al. 2010, 2011;
380 Hildebrandt et al. 2015; Hildebrandt 2018).

381 The pathway annotations listed in **Table 1** were used to subset SNPs and spanned the broad
382 categories of amino acid, core, specialized, and protein metabolism. When partitioning these
383 pathways in the MultiBLUP model, 18 trait-pathway combinations were flagged as potentially
384 related based on comparison to a null distribution (**Figure 2A, Table 3**). The observation that
385 specific pathways improved model fit based on the LR test statistic, explained a significant
386 proportion of genomic heritability, and improved predictive ability suggests that these pathway
387 annotations may have biological relevance for FAA traits.

388 For the trait-pathway combinations that passed the significance criteria, the MultiBLUP
389 model generally reduced bias and RMSE compared to the GBLUP model (**Table 3**). For six of
390 these trait-pathway combinations, the predictive ability for the MultiBLUP model was also over
391 5% higher than for the GBLUP model (**Table 3**, bold). This substantial increase in predictive
392 ability was observed in the pyruvate/BCAA family for absolute levels of leucine (leu, 5.3%) and
393 isoleucine (ile, 7%) when the model included SNPs in the amino acid synthesis pathway. The
394 highest increase in predictive ability was observed when incorporating the amino acid degradation
395 pathway for traits in the glutamate family, which included the relative level and family-based ratio
396 for histidine (his_t, 7.6%; his_GluFam, 9.7%). A similar increase in predictive ability was
397 observed when including SNPs related to phenylpropanoids for the family ratio of tyrosine
398 (tyr_ShikFam, 6.9%) and when including SNPs related to protein amino acid activation for the
399 family ratio of glycine (gly_SerFam, 7.5%).

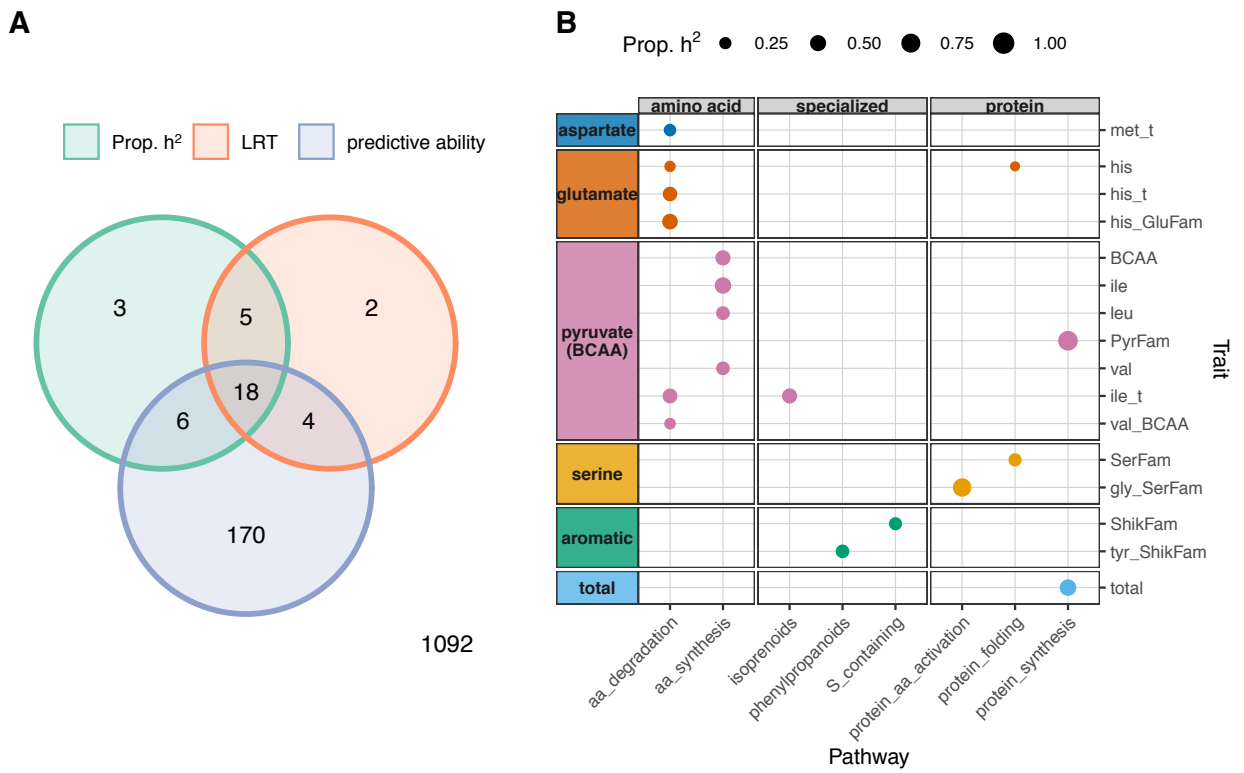


Figure 2. Biological pathways explain significant variation and improve predictive ability for free amino acid traits when incorporated into a MultiBLUP model.

(A) Venn diagram showing which trait-pathway combinations passed significance criteria (FDR adjusted *P*-value < .10) for proportion of heritability explained (Prop. h^2), likelihood ratio test statistic (LRT), and improved predictive ability for MultiBLUP compared to GBLUP. The bottom right corner indicates the number of combinations that did not pass any significance criteria. The Venn diagram was constructed using the 'limma' package in R (Smyth *et al.* 2005).

(B) Dots indicate trait-pathway combinations that passed all three significance criteria. The diameter of each dot is proportional to the amount of genomic variance explained by pathway SNPs in the MultiBLUP model. Traits are included on the y-axis and are grouped by metabolic family (aspartate, glutamate, pyruvate/BCAA, serine, aromatic). Pathways are included on the x-axis and separated into amino acid, core, specialized, and protein metabolism categories.

412 **Table 3. Free amino acid traits and pathway combinations for which the MultiBLUP model explains a significant proportion**

413 **of heritability, improves model fit relative to random gene groups of approximately the same size, and increases accuracy**

414 **compared to GBLUP.**

			Prop. h^2 explained			Likelihood ratio			Predictive ability					
Category	**Pathway**	**Trait**	**% h^2**	***P*-value**	**FDR**	**LR**	***P*-value**	**FDR**	**Δr**	***P*-value**	**FDR**	**Δreliability**	**Δslope[a]**	**ΔRMSE**
amino acid	aa_degradation	his	0.202	0.001	0.020	13.000	<0.001	<0.001	0.039	<0.001	<0.001	0.026	-0.022	-1.46E-02
	aa_degradation	**his_t**	**0.390**	**<0.001**	**<0.001**	**9.907**	**0.001**	**0.020**	**0.076**	**<0.001**	**<0.001**	**0.058**	**-0.036**	**-7.71E-01**
	aa_degradation	ile_t	0.412	0.003	0.060	5.689	0.007	0.100	0.039	0.006	0.037	0.042	-0.045	-1.60E-04
	aa_degradation	met_t	0.268	0.003	0.060	6.382	0.004	0.080	0.027	0.003	0.012	0.015	-0.079	-1.57E-04
	aa_degradation	val_BCAA	0.218	0.002	0.040	8.024	0.004	0.080	0.042	0.001	0.006	0.027	0.003	-4.67E-04
	aa_degradation	**his_GluFam**	**0.468**	**0.002**	**0.040**	**11.780**	**<0.001**	**<0.001**	**0.097**	**<0.001**	**<0.001**	**0.076**	**-0.081**	**-4.59E-01**
	aa_synthesis	**ile**	**0.529**	**0.003**	**0.060**	**11.790**	**<0.001**	**<0.001**	**0.070**	**<0.001**	**<0.001**	**0.073**	**-0.101**	**-1.73E-03**
	aa_synthesis	**leu**	**0.339**	**0.004**	**0.080**	**9.529**	**<0.001**	**<0.001**	**0.053**	**<0.001**	**<0.001**	**0.050**	**-0.040**	**-2.86E-03**
	aa_synthesis	val	0.336	0.003	0.030	4.507	0.007	0.067	0.017	0.006	0.064	0.016	-0.036	-8.14E-04
	aa_synthesis	BCAA	0.439	0.002	0.040	8.823	0.001	0.020	0.044	<0.001	<0.001	0.050	-0.064	-2.16E-03
specialized	isoprenoids	ile_t	0.429	0.008	0.080	4.575	0.010	0.100	0.044	0.001	0.005	0.032	-0.078	-2.04E-04
	phenylpropanoids	**tyr_ShikFam**	**0.332**	**<0.001**	**<0.001**	**8.811**	**<0.001**	**<0.001**	**0.069**	**<0.001**	**0.001**	**0.035**	**-0.301**	**-7.71E-04**
	S_containing	ShikFam	0.301	0.001	0.020	10.810	0.001	0.020	0.032	<0.001	0.001	0.036	-0.024	-2.39E-04
protein	**aa_activation**	**gly_SerFam**	**0.700**	**0.003**	**0.060**	**13.810**	**<0.001**	**<0.001**	**0.075**	**<0.001**	**<0.001**	**0.147**	**-0.076**	**-1.89E-03**
	protein_folding	his	0.146	0.009	0.090	11.520	0.002	0.020	0.021	0.011	0.055	0.013	-0.009	-8.36E-03
	protein_folding	SerFam	0.303	0.005	0.100	6.487	0.001	0.020	0.032	0.001	0.008	0.042	-0.033	-1.11E-04
	protein_synthesis	total	0.535	0.002	0.040	6.418	<0.001	<0.001	0.024	0.001	0.006	0.018	-0.025	-2.00E-04
	protein_synthesis	PyrFam	0.809	0.005	0.100	5.298	0.003	0.060	0.026	0.003	0.029	0.032	-0.030	-1.50E-03

SE, standard error; R*MSE*, root mean squared error.

[a] Computed as $|slope_{multiblup} - 1| - |slope_{GBLUP} - 1|$

415

416 **Amino acid synthesis and degradation pathways were significantly associated with several**
417 **FAA traits**

418 The homeostasis of FAAs is regulated by multiple allosteric enzymes and feedback loops
419 (Less and Galili 2008; Jander and Joshi 2010; Hildebrandt *et al.* 2015; Huang and Jander 2017;
420 Amir *et al.* 2018). However, the homeostasis of some FAAs, such as proline, can also be
421 determined by environmental conditions. For example, proline may serve as either an
422 osmoprotectant under stress or an energy source during development, and its elevation is mostly
423 from active synthesis (Szabados and Savouré 2010; Hayat *et al.* 2012). In addition, previous work
424 has suggested that an overarching metabolic switch occurs during late maturation to desiccation,
425 when amino acid synthesis is active (Fait *et al.* 2006). Hence, our initial hypothesis was that FAA
426 traits would be strongly associated with pathway annotations within core and amino acid
427 metabolism.

428 For amino acid metabolism, our initial hypothesis was supported by significant
429 associations between the amino acid degradation pathway and with six traits, which spanned the
430 aspartate, glutamate, and pyruvate/BCAA families (**Figure 2B**). Two BCAA traits, ile_t and
431 val_BCAA, were associated with amino acid degradation, consistent with previous work which
432 identified a large effect QTL that explained 12-19% of the variance for BCAA traits (Angelovici
433 *et al.* 2013). Based on this previous work, the causal gene was identified as the catabolic branched-
434 chain amino acid transferase 2 (BCAT2; At1g10070) (Angelovici *et al.* 2013). Our results
435 recapitulate this finding, showing that the amino acid degradation pathway, which contains the
436 BCAT2 haploblock, explained both a significant proportion of heritability (41%) and improved
437 predictive ability for BCAA traits (e.g. by 3.9% for ile_t) (**Table 3**). In contrast, the only additional
438 associations that were identified were between amino acid synthesis and BCAA traits, despite no

439 prior evidence that QTLs for BCAAs contain genes related to amino acid synthesis (Angelovici *et*
440 *al.* 2013). Surprisingly, these were also the only associations that were identified for amino acids
441 synthesis, despite evidence that levels of several FAAs and transcription of their biosynthetic genes
442 are elevated toward desiccation (Fait *et al.* 2006). This observation could arise from one of several
443 reasons: 1) the elevation of transcription for amino acid biosynthetic genes does not lead to a
444 corresponding elevation in metabolic pathway products, 2) our sample size and statistical approach
445 was unable to resolve other traits associated with amino acid synthesis, or 3) we are unable to
446 cleanly partition pathway SNPs from background genome wide markers. Nonetheless, our results
447 imply that amino acids synthesis may be more important for BCAAs than for other FAA traits at
448 this stage of development.

449 We also observed that annotations for amino acid degradation were associated with
450 histidine and methionine FAA traits (**Figure 2B, Table 3**), which, to our knowledge, has not been
451 reported in previous QTL studies for seed FAAs. Both histidine and methionine are essential amino
452 acids, which are deficient in most crop seeds, and therefore of special interest for biofortification
453 and crop improvement (Galili and Amir 2013). Notably, very little is currently known about the
454 pathway for histidine degradation in plants. Taken together, these findings suggest that the
455 MultiBLUP approach can not only recapture previous observations for FAA traits, but can also
456 generate new insights into their genetic regulation.

457 Both amino acid and core (or primary) metabolism are tightly interconnected. For example,
458 amino acids in the glutamate family are known to play a central role in core metabolism, mainly
459 by functioning as precursors for energy generation via glycolysis, amino acid metabolism, and the
460 TCA cycle. However, we found no associations for any FAA traits with the core/primary metabolic

461 pathways tested in this study, which included glycolysis, the TCA cycle, and ATP synthesis via
462 alternative oxidase (**Figure 2B, Table 3**).
463

464 **Gene annotations for specialized metabolism are associated with FAA precursors**

465 The synthesis of specialized metabolites involves many FAAs. For example, methionine
466 and aromatic amino acids (i.e. phenylalanine, tryptophan, and tyrosine) are precursors for
467 alkaloids, phenylpropanoids, and glucosinolates. Levels of these specialized metabolites are often
468 dependent on the availability of their FAA precursors (Tzin and Galili 2010; Maeda and Dudareva
469 2012). However, less is known regarding whether the extensive natural variation of these
470 specialized metabolites produces a feedback effect on FAA precursors, especially in seeds.
471 Previous work in vegetative tissues has found that perturbation of the synthesis for secondary
472 metabolites produces a pleiotropic effect on other types of metabolism, including FAAs (Chen *et*
473 *al.* 2012; Slaten *et al.* 2020b), but the nature of such interactions is not well understood.

474 Consistent with knowledge of precursors for specialized metabolites, we observed that
475 aromatic FAAs were associated with categories belonging to specialized metabolism (**Figure 2**).
476 This included associations for the combined absolute levels of FAAs in the shikimate family
477 (ShikFam) with the pathway for sulfur-containing compounds and between the family ratio of
478 tyrosine (tyr_ShikFam) with the phenylpropanoid pathway. When partitioning SNPs from the
479 phenylpropanoid pathway in the MultiBLUP model, we observed a 6.9% increase in predictive
480 ability for tyr_ShikFam (**Table 3**), suggesting SNPs in this pathway have a substantial contribution
481 to the variation for Tyr_ShikFam or are in strong LD with one or more causal variants. We also
482 found an unexpected association between isoprenoid metabolism and the relative ratio of
483 isoleucine (ile_t), which is part of the BCAA family (**Figure 2B**). The metabolic relationship is

484 less clear in this case, as isoleucine is not directly involved in phenylpropanoid metabolism, and
485 provides an avenue for further investigation.

486 A recent metabolic GWA study identified an unanticipated association between
487 glucosinolate biosynthesis and levels of free glutamine in seeds of *Arabidopsis* (Slaten *et al.*
488 2020b). This finding was further validated by evidence that elimination of seed glucosinolates
489 significantly impacted levels of glutamine during early seed development (Slaten *et al.* 2020b).
490 Notably, when partitioning SNPs for sulfur-related metabolism, the family-based ratio for
491 glutamine (GluFam_glu) passed significance criteria for proportion of heritability explained
492 (40.5%, FDR corrected *P*-value = .10) and predictive ability (3.7% increase compared to GBLUP,
493 FDR corrected *P*-value = .006), but not for the LR test statistic (5.48, FDR corrected *P*-value =
494 .12). This observation reinforces that additional studies, especially with greater statistical power,
495 may identify more connections with biological relevance.

496

497 **Annotations for protein metabolism are associated with serine family FAAs**

498 It stands to reason that FAA homeostasis will be influenced by protein metabolism since
499 FAAs serve as the building blocks for proteins. Consistent with this expectation, significant
500 increases in FAAs are observed under many abiotic stresses and suggested to result from protein
501 autophagy and turnover (Hildebrandt *et al.* 2015; Barros *et al.* 2017; Huang and Jander 2017;
502 Hirota *et al.* 2018; Hildebrandt 2018). In contrast, the *opaque2* null mutant in maize exhibits a
503 reduction in the most abundant seed storage proteins and a significant elevation of many FAAs,
504 despite an unchanged composition of protein-bound amino acids (Wang and Larkins 2001;
505 Schmidt *et al.* 2011), indicating a complex relationship between the free and bound amino acid

506 pools for protein metabolism. Hence, it is unclear to what extent protein metabolism affects FAAs,
507 particularly in seeds where protein composition is critical for nutritional quality.

508 Interestingly, we find that protein metabolism annotations are associated with five FAA
509 traits, which spanned the glutamate, pyruvate/BCAA, and serine families, and included the
510 composite trait for total FAA content (**Figure 2, Table 3**). Notably, no aromatic FAA traits were
511 associated with protein metabolic annotation categories, while the serine family FAA traits were
512 exclusively associated with this group of pathways. Further, the family-based ratio for glycine
513 (Gly_SerFam) showed an increase in predictive ability of 7.5% when partitioning SNPs related to
514 amino acid activation in the MultiBLUP model (**Table 3**). This suggests that genes related to amino
515 acid activation, such as tRNA synthetases, may contribute to the homeostasis of glycine and serine.
516 Overall, even though most protein metabolism occurs at seed maturation, we found evidence that
517 annotations for protein metabolism influence FAAs in dry seeds, suggesting that FAA levels at
518 this stage may reflect prior events occurring earlier in seed development.

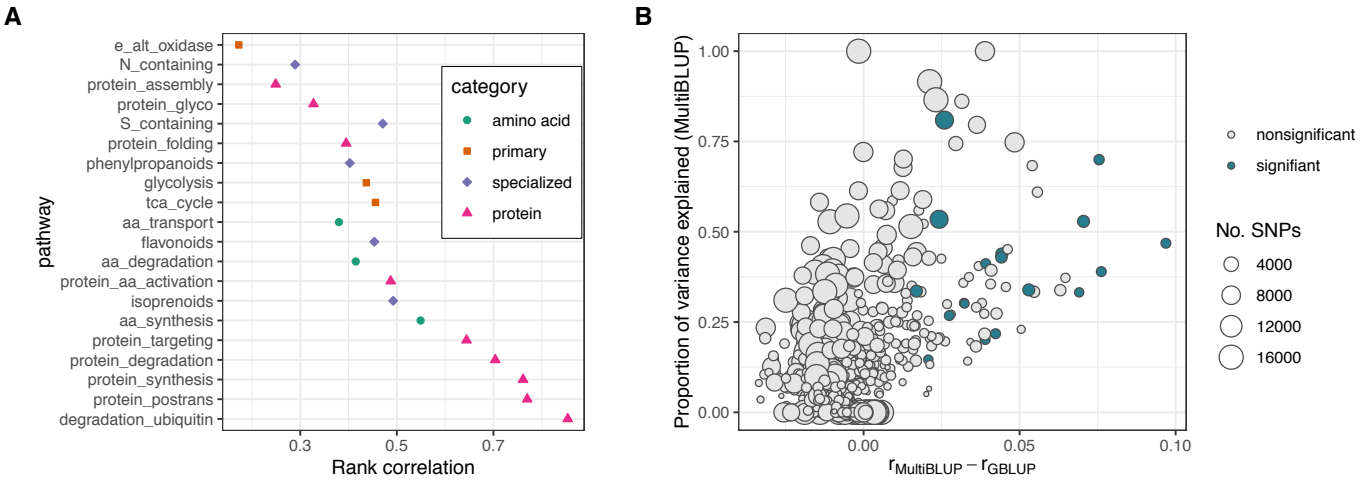
519

520 **Pathway size influences proportion of heritability explained, model fit, and predictive ability**

521 To examine the relationship between pathway size, LD, and variance partitioning, we
522 compared off-diagonal elements of the kinship matrices for pathway SNPs and remaining genomic
523 SNPs (**Figure 3A**). Spearman's correlations ranged from 0.17 for the ATP synthesis via alternative
524 oxidase category (e_alt_oxidases, 66 SNPs, 0.03% of total SNPs) to 0.85 for the protein
525 degradation by ubiquitin category (degradation_ubiquitin, 16000 SNPs, 8.02% of total SNPs)
526 (**Figure 3A**). In general, pathways containing a greater number of SNPs displayed more
527 collinearity with SNPs not contained in the pathway. Similar to observations for genomic
528 partitioning based on gene ontology terms for locomotor activity in *Drosophila* (Rohde *et al.*

529 2018), we observe that pathways which increased predictive ability also explained a large
530 proportion of genomic heritability, whereas pathways with a greater number of SNPs explained
531 less genomic heritability and did not improve predictive ability (**Figure 3B**). Further, as suggested
532 by Rohde *et al.* (2018), pathways which explained all of the genomic heritability likely represent
533 an overestimation caused by high similarity between the relationship matrices for the pathway and
534 background genomic SNPs.

535



536 **Figure 3. Pathway size influences the proportion of heritability explained and predictive ability when using a**
537 **MultiBLUP model.**

538 **(A)** Spearman's rank correlations between off-diagonal elements of the kinship matrices for each pathway and the
539 remaining genomic SNPs. Pathways are sorted from top to bottom by increasing size (number of SNPs).

540 **(B)** Difference in predictive ability between the MultiBLUP and GBLUP models compared to the proportion of
541 heritability explained by each pathway for all 1300 trait-pathway combinations (65 traits, 20 pathways). The diameter
542 of the points is proportional to the number of SNPs in the pathway and color indicates whether or not a trait-pathway
543 combination passed significance for proportion of heritability explained, likelihood ratio test statistic, and predictive
544 ability.

Conclusions

Overall, we find that predictive ability for FAA traits was improved by incorporating prior knowledge from metabolic pathway annotations for several FAA traits, adding to a growing body of literature that demonstrates the utility of genomic partitioning in the study and prediction of complex traits. This study further highlights that specific metabolic pathways are associated with natural variation of FAA traits across amino acid families. The amino acid degradation pathway was significantly associated with traits in the BCAA/pyruvate, glutamate, and aspartate families, while specialized metabolism was associated with traits in the aromatic family and protein metabolism was associated with traits in the serine, pyruvate/BCAA, and glutamate families. Thus, although the FAA metabolic network is tightly connected, the predominant genetic architecture underlying variation for specific FAA traits varies, at least for this stage of seed development. Overall, this study furthers our understanding of the contribution from specific metabolic pathway genes to amino acid trait variation and offers an additional strategy to investigate other complex metabolic traits, both in *Arabidopsis* and other species.

Competing interests

The authors declare that they have no competing interests.

Funding

This project was supported by the USDA Agricultural Research Service, the University of Missouri Division of Biological Sciences (Columbia, MO, USA), the NSF Postdoctoral Research Fellowship in Biology Grant No. 1711347 for STH, and the NSF Graduate Research Fellowship Grant No. DGE-1424871 for KAB. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

STH, KB, TB, and RA conceived the study and wrote the paper. STH and KB performed the analyses. AL, EK, and TB contributed to statistical analyses and interpretations. RA supervised the study and aided interpretation of results. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Dan Kliebenstein, Jinliang Yang, Jeffrey Ross-Ibarra, and anonymous reviewers for helpful comments and discussions that improved the manuscript. We thank Doug Speed for advice on the LDAK software and Marianne Slaten for assistance with HAPPI GWAS.

REFERENCES

Amir, R., G. Galili, and H. Cohen, 2018 The metabolic roles of free amino acids during seed development. Plant Sci. 275: 11–18.

Angelovici, R., A. Batushansky, N. Deason, S. Gonzalez-Jorge, M. A. Gore *et al.*, 2016 Network-guided GWAS improves identification of genes affecting free amino acids. Plant Physiol.

Angelovici, R., A. Fait, A. R. Fernie, and G. Galili, 2011 A seed high-lysine trait is negatively associated with the TCA cycle and slows down Arabidopsis seed germination. New Phytol. 189: 148–159.

Angelovici, R., G. Galili, A. R. Fernie, and A. Fait, 2010 Seed desiccation: a bridge between maturation and germination. Trends Plant Sci. 15: 211–218.

Angelovici, R., A. E. Lipka, N. Deason, S. Gonzalez-Jorge, H. Lin *et al.*, 2013 Genome-Wide Analysis of Branched-Chain Amino Acid Levels in Arabidopsis Seeds. Plant Cell 25: 4827–4843.

Araújo, W. L., K. Ishizaki, A. Nunes-Nesi, T. R. Larson, T. Tohge *et al.*, 2010 Identification of the 2-Hydroxyglutarate and Isovaleryl-CoA Dehydrogenases as Alternative Electron Donors Linking Lysine Catabolism to the Electron Transport Chain of Arabidopsis Mitochondria. Plant Cell 22: 1549–1563.

Astle, W., and D. J. Balding, 2009 Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat. Sci. 24: 451–471.

Atwell, S., Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton *et al.*, 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627–631.

Barros, J. A. S., J. H. F. Cavalcanti, D. B. Medeiros, A. Nunes-Nesi, T. Avin-Wittenberg *et al.*, 2017 Autophagy Deficiency Compromises Alternative Pathways of Respiration following Energy Deprivation in Arabidopsis thaliana. Plant Physiol. 175: 62–76.

Bates, D., M. Mächler, B. Bolker, and S. Walker, 2015 Fitting Linear Mixed-Effects Models Usinglme4. Journal of Statistical Software 67.:

Benjamini, Y., and Y. Hochberg, 1995 Controlling the False Discovery Rate: A Practical and Powerful

Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. 57: 289–300.

Berardini, T. Z., L. Reiser, D. Li, Y. Mezheritsky, R. Muller *et al.*, 2015 The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis 53: 474–485.

Box, G. E. P., and D. R. Cox, 1964 An analysis of transformations. J. R. Stat. Soc.

Chen, Y.-Z., Q.-Y. Pang, Y. He, N. Zhu, I. Branstrom *et al.*, 2012 Proteomics and metabolomics of Arabidopsis responses to perturbation of glucosinolate biosynthesis. Mol. Plant 5: 1138–1150.

Cohen, H., H. Israeli, I. Matityahu, and R. Amir, 2014 Seed-specific expression of a feedback-insensitive form of CYSTATHIONINE-γ-SYNTHASE in Arabidopsis stimulates metabolic and transcriptomic responses associated with desiccation stress. Plant Physiol. 166: 1575–1592.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de Los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193: 347–365.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor *et al.*, 2005 BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439–3440.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber, 2009 Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4.:

Edwards, S. M., I. F. Sørensen, P. Sarup, T. F. C. Mackay, and P. Sørensen, 2016 Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. Genetics 203: 1871–1883.

Edwards, S. M., B. Thomsen, P. Madsen, and P. Sørensen, 2015 Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. Genet. Sel. Evol. 47: 60.

Fait, A., R. Angelovici, H. Less, I. Ohad, E. Urbanczyk-Wochniak *et al.*, 2006 Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. Plant

Physiol. 142: 839–854.

Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu *et al.*, 2017 Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. Genet. Sel. Evol. 49: 44.

Galili, G., and R. Amir, 2013 Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality. Plant Biotechnology Journal 11: 211–222.

Galili, G., T. Avin-Wittenberg, R. Angelovici, and A. R. Fernie, 2014 The role of photosynthesis and amino acid metabolism in the energy status during seed development. Front. Plant Sci. 5: 447.

Goddard, M. E., N. R. Wray, K. Verbyla, and P. M. Visscher, 2009 Estimating Effects and Making Predictions from Genome-Wide Marker Data. Stat. Sci. 24: 517–529.

Gonzalez-Jorge, S., S.-H. Ha, M. Magallanes-Lundback, L. U. Gilliland, A. Zhou *et al.*, 2013 Carotenoid cleavage dioxygenase4 is a negative regulator of β-carotene content in Arabidopsis seeds. Plant Cell 25: 4812–4826.

Harjes, C. E., T. R. Rocheford, L. Bai, T. P. Brutnell, C. B. Kandianis *et al.*, 2008 Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319: 330–333.

Hayat, S., Q. Hayat, M. N. Alyemeni, A. S. Wani, J. Pichtel *et al.*, 2012 Role of proline under changing environments: a review. Plant Signal. Behav. 7: 1456–1466.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009 Genomic Selection for Crop Improvement. Crop Sci. 49: 1–12.

Hildebrandt, T. M., 2018 Synthesis versus degradation: directions of amino acid metabolism during Arabidopsis abiotic stress response. Plant Mol. Biol. 98: 121–135.

Hildebrandt, T. M., A. Nunes Nesi, W. L. Araújo, and H.-P. Braun, 2015 Amino Acid Catabolism in Plants. Mol. Plant 8: 1563–1579.

Hirota, T., M. Izumi, S. Wada, A. Makino, and H. Ishida, 2018 Vacuolar Protein Degradation via Autophagy Provides Substrates to Amino Acid Catabolic Pathways as an Adaptive Response to

Sugar Starvation in Arabidopsis thaliana. Plant Cell Physiol. 59: 1363–1376.

Huang, T., and G. Jander, 2017 Abscisic acid-regulated protein degradation causes osmotic stress-induced accumulation of branched-chain amino acids in Arabidopsis thaliana. Planta 246: 737–747.

Jander, G., and V. Joshi, 2010 Recent progress in deciphering the biosynthesis of aspartate-derived amino acids in plants. Mol. Plant 3: 54–65.

Kassambara, A., 2020 rstatix: pipe-friendly framework for basic statistical tests. R package version 0.4. 0.

Korte, A., and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 29.

Köster, J., and S. Rahmann, 2012 Snakemake--a scalable bioinformatics workflow engine. Bioinformatics 28: 2520–2522.

Kutner, M. H., C. J. Nachtsheim, and J. N. Dr., 2004 *Applied Linear Regression Models- 4th Edition with Student CD (McGraw Hill/Irwin Series: Operations and Decision Sciences)*. McGraw-Hill Education.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon *et al.*, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

Less, H., and G. Galili, 2008 Principal Transcriptional Programs Regulating Plant Amino Acid Metabolism in Response to Abiotic Stresses. Plant Physiol. 147: 316–330.

Lipka, A. E., M. A. Gore, M. Magallanes-Lundback, A. Mesberg, H. Lin *et al.*, 2013 Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. G3 3: 1287–1299.

Li, X., C. Zhu, C.-T. Yeh, W. Wu, E. M. Takacs *et al.*, 2012 Genic and nongenic contributions to natural variation of quantitative traits in maize. Genome Res. 22: 2436–2444.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345.

Lozano, R., D. P. del Carpio, T. Amuge, I. S. Kayondo, A. O. Adebo *et al.*, 2017 Leveraging

Transcriptomics Data for Genomic Prediction Models in Cassava. bioRxiv 208181.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper *et al.*, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17: 144.

Maeda, H., and N. Dudareva, 2012 The shikimate pathway and aromatic amino Acid biosynthesis in plants. Annu. Rev. Plant Biol. 63: 73–105.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Muehlbauer, G. J., B. G. Gengenbach, D. A. Somers, and C. M. Donovan, 1994 Genetic and amino-acid analysis of two maize threonine-overproducing, lysine-insensitive aspartate kinase mutants. Theor. Appl. Genet. 89: 767–774.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol. 3: e196.

Osorio, S., J. G. Vallarino, M. Szecowka, S. Ufaz, V. Tzin *et al.*, 2013 Alteration of the interconversion of pyruvate and malate in the plastid or cytosol of ripening tomato fruit invokes diverse consequences on sugar but similar effects on cellular organic acid, metabolism, and transitory starch accumulation. Plant Physiol. 161: 628–643.

Owens, B. F., A. E. Lipka, M. Magallanes-Lundback, T. Tiede, C. H. Diepenbrock *et al.*, 2014 A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. Genetics 198: 1699–1716.

Piepho, H.-P., J. Möhring, T. Schulz-Streeck, and J. O. Ogutu, 2012 A stage-wise approach for the analysis of multi-environment trials. Biom. J. 54: 844–860.

Platt, A., M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio *et al.*, 2010 The scale of population structure in Arabidopsis thaliana. PLoS Genet. 6: e1000843.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81.:

Rai, V. K., 2002 Role of Amino Acids in Plant Responses to Stresses. Biol. Plant. 45: 481–487.

Ramstein, G. P., S. J. Larsson, J. P. Cook, J. W. Edwards, E. S. Ersoz *et al.*, 2020 Dominance Effects and Functional Enrichments Improve Prediction of Agronomic Traits in Hybrid Maize. Genetics 215: 215–230.

R Core Team, 2016 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riedelsheimer, C., J. Lisec, A. Czedik-Eysenberg, R. Sulpice, A. Flis *et al.*, 2012 Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc. Natl. Acad. Sci. U. S. A. 109: 8872–8877.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel *et al.*, 2012 Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics 192: 715–728.

Rohde, P. D., S. Østergaard, T. N. Kristensen, P. Sørensen, V. Loeschcke *et al.*, 2018 Functional Validation of Candidate Genes Detected by Genomic Feature Models. G3 8: 1659–1668.

Sarup, P., J. Jensen, T. Ostersen, M. Henryon, and P. Sørensen, 2016 Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. BMC Genet. 17: 11.

Sauer, U., D. R. Lasko, J. Fiaux, M. Hochuli, R. Glaser *et al.*, 1999 Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. J. Bacteriol. 181: 6679–6688.

Schmidt, M. A., W. B. Barbazuk, M. Sandford, G. May, Z. Song *et al.*, 2011 Silencing of soybean seed storage proteins results in a rebalanced protein composition preserving seed protein content without major collateral changes in the metabolome and transcriptome. Plant Physiol. 156: 330–345.

Slaten, M. L., Y. O. Chan, V. Shrestha, A. E. Lipka, and R. Angelovici, 2020a HAPPI GWAS: Holistic Analysis with Pre and Post Integration GWAS. bioRxiv 2020.04.07.998690.

Slaten, M. L., A. Yobi, C. Bagaza, Y. O. Chan, V. Shrestha *et al.*, 2020b mGWAS Uncovers Gln-

Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis. Plant Physiol. 183: 483–500.

Smyth, G. K., M. Ritchie, N. Thorne, and J. Wettenhall, 2005 LIMMA: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health.

Speed, D., and D. J. Balding, 2014 MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 24: 1550–1557.

Speed, D., N. Cai, UCLEB Consortium, M. R. Johnson, S. Nejentsev *et al.*, 2017 Reevaluation of SNP heritability in complex human traits. Nat. Genet. 49: 986–992.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding, 2012 Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. 91: 1011–1021.

Szabados, L., and A. Savouré, 2010 Proline: a multifunctional amino acid. Trends Plant Sci. 15: 89–97.

Thimm, O., O. Bläsing, Y. Gibon, A. Nagel, S. Meyer *et al.*, 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.

Tzin, V., and G. Galili, 2010 New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. Mol. Plant 3: 956–972.

Vallabhaneni, R., and E. T. Wurtzel, 2009 Timing and biosynthetic potential for carotenoid accumulation in genetically diverse germplasm of maize. Plant Physiol. 150: 562–572.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Wang, X., and B. A. Larkins, 2001 Genetic Analysis of Amino Acid Accumulation inopaque-2 Maize Endosperm. Plant Physiol. 125: 1766–1777.

Weckwerth, W., M. E. Loureiro, K. Wenzel, and O. Fiehn, 2004 Differential metabolic networks unravel the effects of silent plant phenotypes. Proc. Natl. Acad. Sci. U. S. A. 101: 7809–7814.

Weller, J. I., E. Ezra, and M. Ron, 2017 Invited review: A perspective on the future of genomic selection in dairy cattle. J. Dairy Sci. 100: 8633–8644.

Wentzell, A. M., H. C. Rowe, B. G. Hansen, C. Ticconi, B. A. Halkier *et al.*, 2007 Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet. 3: 1687–1701.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genet. Res. 75: 249–252.

Wolc, A., A. Kranis, J. Arango, P. Settar, J. E. Fulton *et al.*, 2016 Implementation of genomic selection in the poultry industry. Animal Frontiers 6: 23–31.

Wu, Y., and J. Messing, 2014 Proteome balancing of the maize seed for higher nutritional value. Front. Plant Sci. 5: 240.

Wurtzel, E. T., A. Cuttriss, and R. Vallabhaneni, 2012 Maize provitamin a carotenoids, current resources, and future metabolic engineering challenges. Front. Plant Sci. 3: 29.

Zhan, S., J. Horrocks, and L. N. Lukens, 2006 Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. Plant J. 45: 347–357.

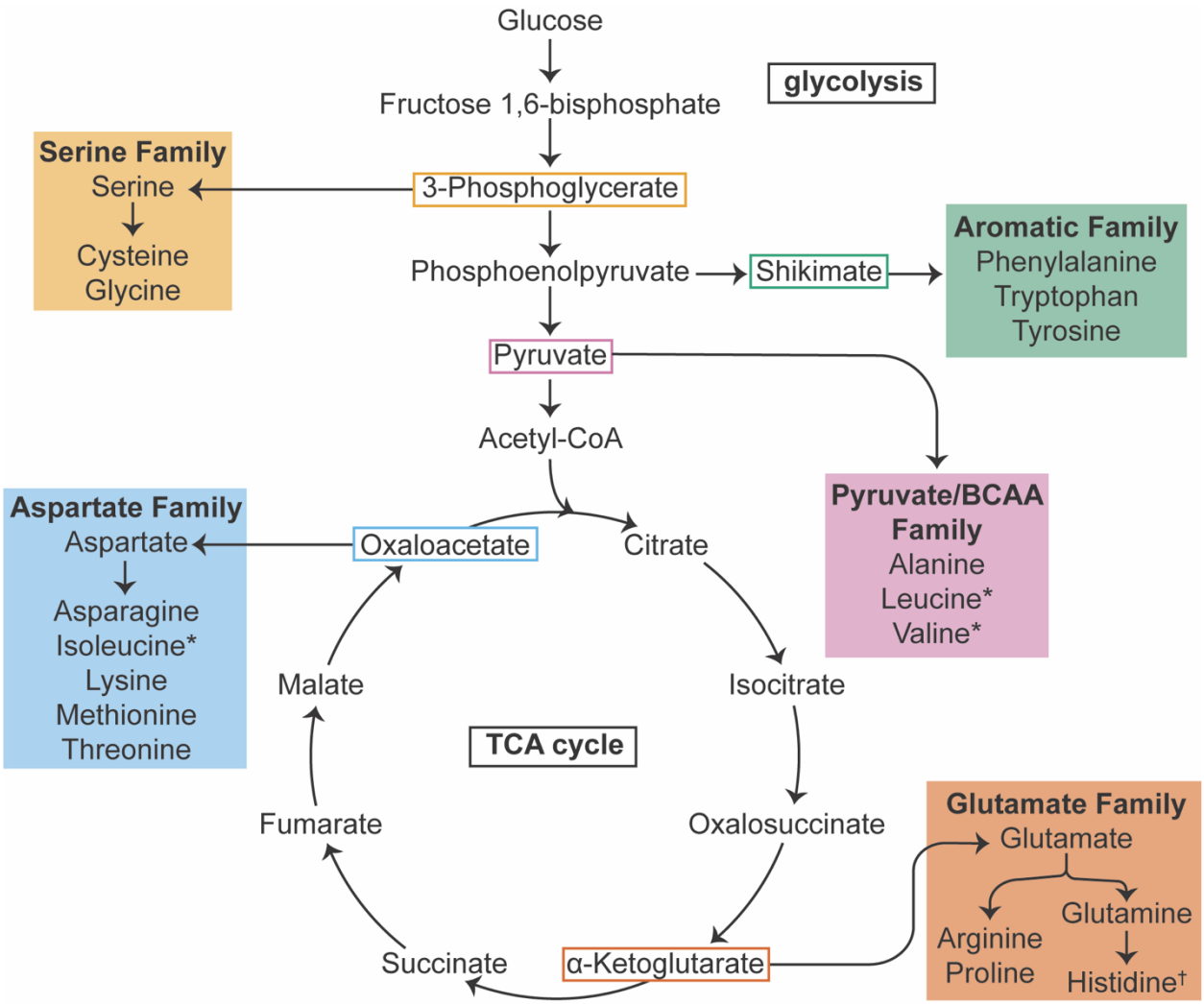


Figure S1. Biochemical relationships among amino acids.
Colors indicate different amino acid families and boxes indicate the corresponding precursor. The branched-chain amino acids include Leu, Ile, Val are split across the Aspartate and Pyruvate family and therefore denoted with asterisks (*). Note that histidine (†) does not belong explicitly to the families identified here, but often is considered as part of the glutamate family.

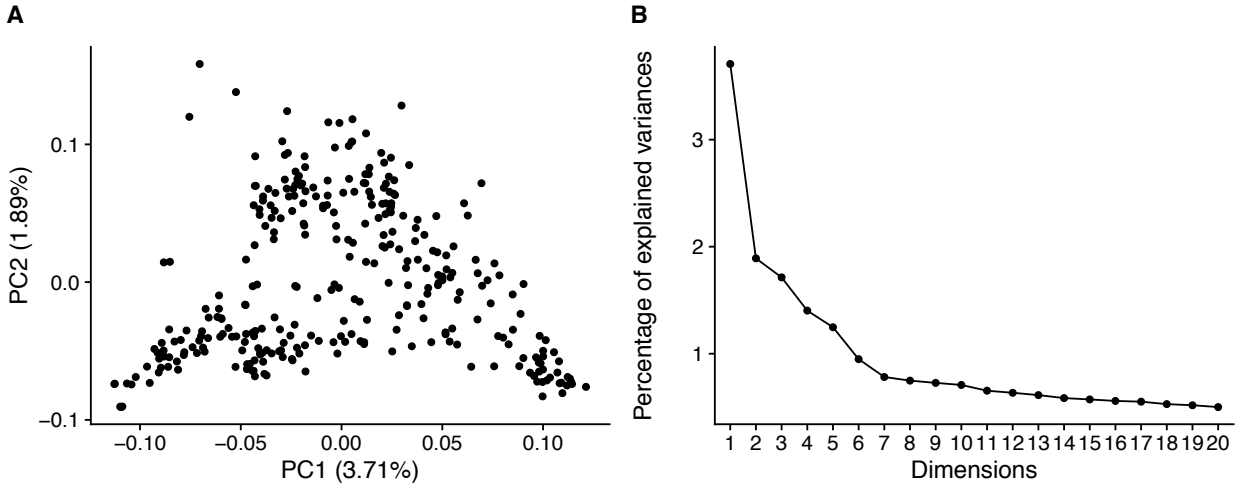


Figure S2. Principal component analysis (PCA) of genetic data for the 313 Arabidopsis accessions used in this study.

(A) PCA scatterplot and percent variation explained for the first two principal components. (B) Screeplot showing the percent variance explained by each principal component.

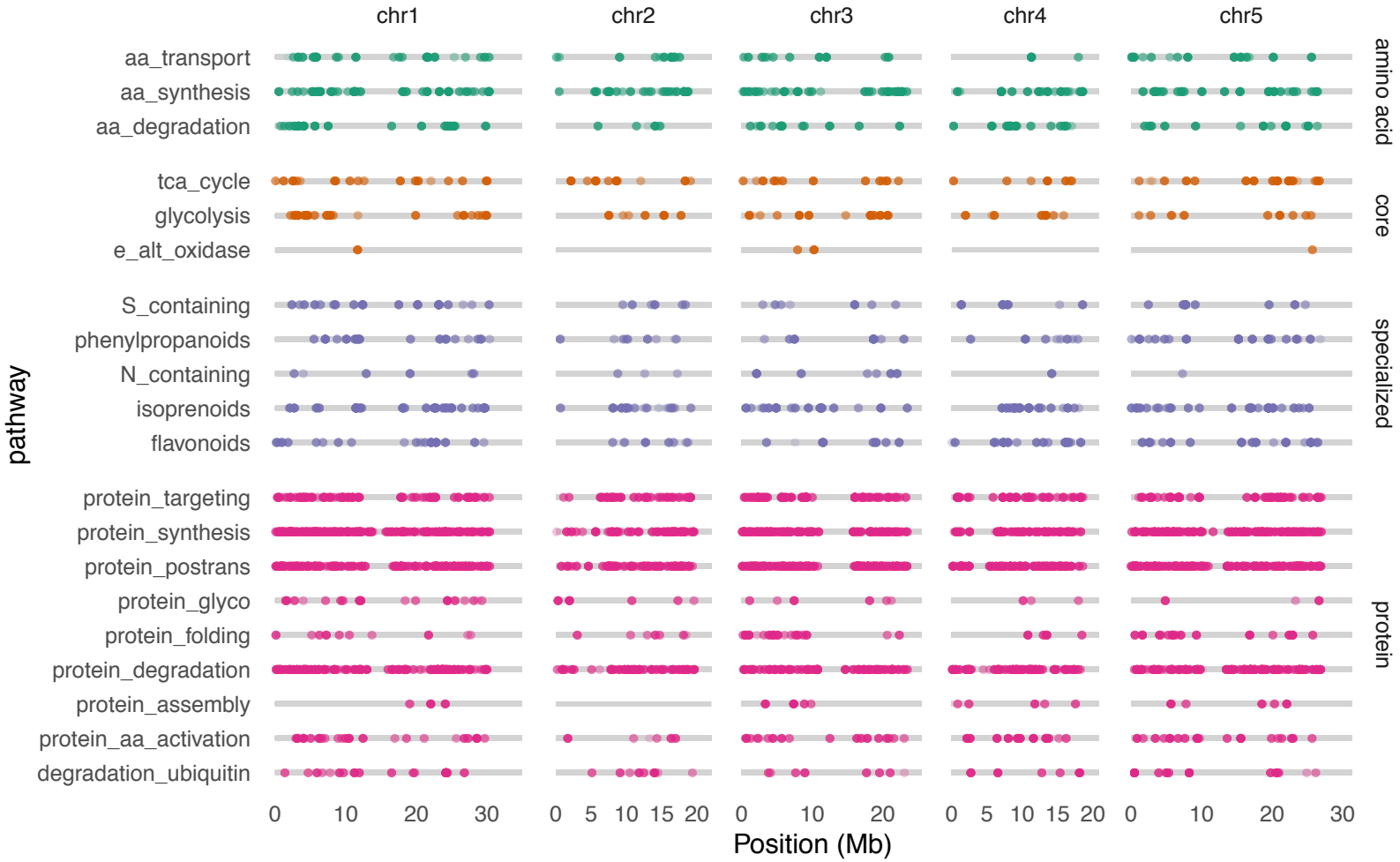


Figure S3. Genomic locations of pathway SNPs for each pathway.
The x-axis indicates genomic position on each chromosome in megabases (Mb) and the y-axis indicates each pathway tested. Pathways are grouped into broad categories of metabolism (amino acid, core, specialized, and protein) and each point represents a SNP.

Table S1. List of seed free amino acid traits calculated from the quantification of 18 FAA and biochemical family affiliations. Strings of AA letter codes represent the sum of those AAs.

Amino acids one letter code	
Ala	A
Asp	D
Glu	E
Phe	F
Gly	G
His	H
Ile	I
Lys	K
Leu	L
Met	M
Pro	P
Gln	Q
Arg	R
Ser	S
Thr	T
Val	V
Trp	W
Tyr	Y

Absolute levels
AA-Abs Total = Sum of 18 AA
A
D
E
F
G
H
I
K
L
M
P
Q
R
S
T
V
W
Y
Total
IMTDK (Asp family)
IVL (BCAA family)
LAV (Pyr family)
EHPRQ (Glu family)
WFY (Shik family)
SG (Ser family)

Relative levels to total
AA/Total
A/Total
D/Total
E/Total
F/Total
G/Total
H/Total
I/Total
K/Total
L/Total
M/Total
P/Total
Q/Total
R/Total
S/Total
T/Total*
V/Total
W/Total
Y/Total

Biochemistry based metabolic ratios, grouped by AA families' affiliation

Asp Family = Ile, Met, Thr, Asp Lys (IMTDK)
D/IMTDK
K/IMTDK
M/IMTDK
T/IMTDK
IMTDK/D

Glu Family = Glu, His, Pro, Arg, Gln (EHPRQ)
Q/EHPRQ
E/EHPRQ
H/EHPRQ
P/EHPRQ
R/EHPRQ
EHPRQ/E

BCAA Family = Ile, Val, Leu (IVL) Pyr Family=Leu, Ala, Val (LAV)
I/IVL
V/IVL
L/IVL
A/LAV
L/LAV
V/LAV
I/IMTDK

Shikimate (Aromatic) Fam = Trp, Phe,Tyr (WFY)
W/WFY
F/WFY
Y/WFY

Ser Family = Ser, Gly (Cysteine-not detected -SG)
G/SG
S/SG

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
1	ALL1-2	CS76089
2	ALL1-3	CS76090
23	CAM-16	CS76107
66	CAM-61	CS76108
81	CUR-3	CS76115
91	JEA	CS76148
94	LAC-3	CS76157
96	LAC-5	CS76158
116	LDV-25	CS76161
126	LDV-34	CS76162
149	LDV-58	CS76163
166	MIB-15	CS76181
173	MIB-22	CS76182
178	MIB-28	CS76183
223	MIB-84	CS76184
242	MOG-37	CS76189
258	PAR-3	CS76205
259	PAR-4	CS76206
267	ROM-1	CS76221
281	TOU-A1-115	CS76252
282	TOU-A1-116	CS76253
321	TOU-A1-43	CS76255
328	TOU-A1-62	CS76256
334	TOU-A1-67	CS76257
357	TOU-A1-96	CS76258
362	TOU-C-3	CS76259
366	TOU-E-11	CS76260
373	TOU-H-12	CS76261
374	TOU-H-13	CS76262
378	TOU-I-17	CS76263
379	TOU-I-2	CS76264
380	TOU-I-6	CS76265
383	TOU-J-3	CS76266
386	TOU-K-3	CS76267
390	VOU-1	CS76299
392	VOU-2	CS76300
641	LI-OF-095	CS76165
957	Belmonte-4-94	CS76095
1716	KBS-Mac-8	CS76151

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
1859	MNF-Pot-48	CS76187
1867	MNF-Pot-68	CS76188
1925	MNF-Che-2	CS76185
2057	Map-42	CS76180
2150	Paw-3	CS76208
2187	Pent-1	CS76209
2274	SLSP-30	CS76228
2290	Ste-3	CS76232
4802	UKSW06-202	CS76292
5056	UKSE06-192	CS76281
5116	UKSE06-272	CS76282
5122	UKSE06-278	CS76283
5158	UKSE06-349	CS76284
5160	UKSE06-351	CS76285
5202	UKSE06-414	CS76286
5207	UKSE06-429	CS76287
5232	UKSE06-466	CS76288
5245	UKSE06-482	CS76289
5264	UKSE06-520	CS76290
5341	UKSE06-628	CS76291
5380	UKNW06-059	CS76275
5381	UKNW06-060	CS76276
5565	UKNW06-386	CS76277
5606	UKNW06-436	CS76278
5742	UKID37	CS76272
5753	UKID48	CS76273
5785	UKID80	CS76274
5805	UKID101	CS76270
5832	App1-16	CS76092
5837	Bor-1	CS76099
5887	DraIV 1-5	CS76120
5896	DraIV 1-14	CS76119
5987	DraIV 6-16	CS76122
6005	DraIV 6-35	CS76123
6008	Duk	CS76124
6020	Fja 1-5	CS76132
6042	Lom1-1	CS76174
6074	Or-1	CS76201
6094	T1040	CS76233

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
6100	T1110	CS76236
6188	TDr-1	CS76245
6190	TDr-3	CS76248
6194	TDr-8	CS76249
6243	Tottarp-2	CS76251
6318	UduI 1-34	CS76269
6413	Ull3-4	CS76295
6448	ZdrI 2-24	CS76307
6449	ZdrI 2-25	CS76308
6709	Bg-2	CS76096
6727	CIBC-2	CS28140
6729	CIBC-4	CS28141
6730	CIBC-5	CS28142
6744	CSHL-5	CS28181
6810	KNO-11	CS28407
6847	NFC-20	CS28550
6897	Ag-0	CS76087
6898	An-1	CS76091
6899	Bay-0	CS76094
6903	Bor-4	CS76100
6904	Br-0	CS76101
6905	Bur-0	CS76105
6906	C24	CS76106
6907	CIBC-17	CS76111
6909	Col-0	CS76113
6910	Ct-1	CS76114
6911	Cvi-0	CS76116
6914	Edi-0	CS76126
6916	Est-1	CS76127
6919	Ga-0	CS76133
6921	Got-7	CS76136
6924	HR-5	CS76144
6926	Kin-0	CS76153
6928	Kno-18	CS76154
6932	Ler-1	CS76164
6933	LL-0	CS76172
6936	Lz-0	CS76179
6937	Mrk-0	CS76191
6939	Mt-0	CS76192

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
6940	Mz-0	CS76193
6942	Nd-1	CS76197
6943	NFA-10	CS76198
6944	NFA-8	CS76199
6946	Oy-0	CS76203
6951	Pu2-23	CS76215
6953	Pu2-24	CS28663
6958	Ra-0	CS76216
6959	Ren-1	CS76218
6961	Se-0	CS76226
6962	Shahdara	CS76227
6967	Sq-8	CS76230
6968	Tamm-2	CS76244
6970	Ts-1	CS76268
6973	Ull2-3	CS76293
6976	Uod-7	CS76296
6977	Van-0	CS76297
6979	Wei-0	CS76301
6980	Ws-0	CS76303
6982	Wt-5	CS76304
6983	Yo-0	CS76305
6985	Zdr-6	CS76306
6988	Alc-0	CS76088
6989	Alst-1	CS28013
6990	Amel-1	CS28014
6992	Ang-0	CS28018
6994	Ann-1	CS28049
6996	An-2	CS28017
6998	Arby-1	CS28051
7000	Aa-0	CS28007
7002	Baa-1	CS28054
7004	Bs-2	CS28097
7008	Benk-1	CS28064
7011	Be-1	CS28063
7014	Ba-1	CS28053
7026	Boot-1	CS28091
7031	Bsch-0	CS28099
7035	Blh-2	CS28090
7062	Ca-0	CS28128

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
7064	Cnt-1	CS28160
7069	Cha-0	CS28133
7071	Chat-1	CS28135
7075	Cit-0	CS28158
7078	Co-2	CS28163
7080	Co-4	CS28165
7092	Com-1	CS28193
7094	Da-0	CS28200
7098	Di-1	CS28208
7100	Db-0	CS28202
7102	Do-0	CS28210
7105	Dra-2	CS28214
7110	Ede-1	CS28217
7123	Ep-0	CS28236
7126	Es-0	CS28241
7128	Est-0	CS28243
7135	Fr-4	CS28268
7139	Fi-1	CS28252
7141	Ga-2	CS28274
7143	Gel-1	CS28279
7145	Ge-1	CS28277
7147	Gie-0	CS28280
7150	Gu-1	CS28332
7151	Go-0	CS28282
7158	Gr-5	CS28326
7163	Ha-0	CS28336
7165	Hn-0	CS28350
7166	Hey-1	CS28344
7169	Hh-0	CS28345
7178	Jm-1	CS28373
7181	Je-0	CS28364
7186	Kn-0	CS28395
7188	Kelsterbach-2	CS28382
7199	Kl-5	CS28394
7201	Kr-0	CS28419
7205	Krot-2	CS28423
7206	Kro-0	CS28420
7224	Li-3	CS28454
7227	Li-5:2	CS28457

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
7229	Li-6	CS28459
7231	Li-7	CS28461
7244	Mnz-0	CS28495
7252	Mc-0	CS28490
7255	Mh-0	CS28492
7258	Nw-0	CS28573
7260	Nw-2	CS28575
7263	Nz1	CS28578
7270	Nok-1	CS28568
7275	No-0	CS28564
7277	Ob-1	CS28580
7280	Old-1	CS28583
7282	Or-0	CS28587
7283	Ors-1	CS28848
7284	Ors-2	CS28849
7291	Pa-2	CS28595
7296	Petergof	CS76211
7300	Pla-0	CS28640
7306	Pog-0	CS28650
7307	Pn-0	CS28645
7310	Pr-0	CS28651
7316	Rhen-1	CS28685
7320	Rou-0	CS28692
7330	Sapporo-0	CS28724
7331	Sh-0	CS28734
7333	Sei-0	CS28729
7337	Si-0	CS28739
7340	Sav-0	CS28725
7343	Sp-0	CS28743
7346	Ste-0	CS28750
7351	Ty-0	CS28786
7353	Tha-1	CS28758
7355	Tiv-1	CS28760
7372	Tscha-1	CS28779
7373	Tsu-0	CS28780
7378	Uk-1	CS28787
7379	Uk-2	CS28788
7382	Utrecht	CS28795
7384	Ven-1	CS28800

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
7390	Wag-3	CS28808
7391	Wag-4	CS28809
7392	Wag-5	CS28810
7394	Wa-1	CS28804
7397	Ws	CS28823
7405	Wc-2	CS28814
7408	Wt-3	CS28833
7411	Wl-0	CS28822
7418	Zu-1	CS28847
7424	Jl-3	CS28369
7430	Nc-1	CS28527
7438	N13	CS76194
7446	N4	CS28510
7449	N7	CS28513
7460	Da(1)-12	CS28201
7472	S96	CS28720
7477	WAR	CS28812
7479	PHW-10	CS28610
7482	PHW-13	CS28613
7483	PHW-14	CS28614
7490	PHW-20	CS28620
7492	PHW-22	CS28622
7496	PHW-26	CS28626
7498	PHW-28	CS28628
7502	PHW-31	CS28631
7504	PHW-33	CS28633
7506	PHW-35	CS28635
7507	PHW-36	CS28636
7508	PHW-37	CS28637
7514	RRS-7	CS28713
7515	RRS-10	CS22689
7518	Omo2-1	CS76200
7520	Lp2-2	CS76176
7521	Lp2-6	CS76177
7522	Mr-0	CS76190
7523	Pna-17	CS76213
7525	Rmx-A180	CS76220
8213	Pro-0	CS76214
8214	Gy-0	CS76139

Table S2 ***Arabidopsis thaliana*** **accessions used in this study** (from Angelovici et al. Plant Cell 2013)

Ecotype ID	ABRC stock name	Accession number
8215	Fei-0	CS76129
8222	Lis-2	CS76170
8235	Hod	CS76141
8236	HSm	CS76146
8239	Koln	CS76155
8240	Kulturen-1	CS76156
8241	Liarum	CS76166
8256	Ba1-2	CS76093
8264	Bla-1	CS76097
8265	Blh-1	CS76098
8271	Bu-0	CS76103
8274	Can-0	CS76109
8275	Cen-0	CS76110
8296	Gd-1	CS76134
8300	Gr-1	CS76137
8304	Hi-0	CS76140
8306	Hov4-1	CS76142
8307	Hovdala-2	CS76143
8310	Hs-0	CS76145
8311	In-0	CS76147
8314	Ka-0	CS76149
8323	Lc-0	CS76159
8325	Lip-0	CS76168
8326	Lis-1	CS76169
8329	Lm-2	CS76173
8343	Na-1	CS76195
8351	Ost-0	CS76202
8354	Per-1	CS76210
8365	Rak-2	CS76217
8374	Rsch-4	CS76222
8376	Sanna-2	CS76223
8378	Sap-0	CS76224
8387	St-0	CS76231
8389	Ta-0	CS76242

Table S3 Ranges for the number of genes and SNPs in 1000 null gene groups for each pathway. Null gene groups were sampled for each pathway by excluding SNPs within the pathway.

	Number of genes		Number of SNPs	
pathway	**minimum**	**maximum**	**minimum**	**maximum**
aa_degradation	108	177	1080	1172
aa_synthesis	225	327	2068	2150
aa_transport	88	159	931	1017
degradation_ubiquitin	1924	2129	15866	16000
e_alt_oxidase	3	21	66	143
flavonoids	104	168	1054	1118
glycolysis	85	144	850	927
isoprenoids	189	269	1769	1859
N_containing	18	49	229	304
phenylpropanoids	84	139	839	909
protein_aa_activation	125	194	1223	1285
protein_assembly	25	65	310	392
protein_degradation	739	885	6363	6470
protein_folding	72	138	809	867
protein_glyco	40	81	452	529
protein_postrans	1030	1211	8725	8835
protein_synthesis	854	1023	7231	7342
protein_targeting	417	530	3670	3738
S_containing	70	119	725	800
tca_cycle	95	159	916	1003