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.

INTRODUCTION

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

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

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

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

While linkage mapping and GWA studies are typically underpowered to identify variants 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

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

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

MATERIALS AND METHODS

Plant materials and trait data

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

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

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

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

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

178 **Genetic data:**

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

185

186 **Selection of pathway SNPs**

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

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

204

Table 1. Summary of selected biological pathways.

Pathway	Number of genes	Number of SNPs	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
Core 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

Pathways include genes and SNPs within a 2.5 kb buffer before the start and after the stop position of each gene.

205 **Prediction models**

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

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

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}'/p, \quad (1)$$

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

$$\boldsymbol{Y} = \mu + \boldsymbol{Z}u + \varepsilon, \quad (2)$$

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

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

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

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

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

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

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

Estimation of genomic heritability

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

$$h_m^2 = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{\notin m}^2 + \sigma_e^2}. \quad (4)$$

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

$$\frac{h_m^2}{h_m^2 + h_{\notin m}^2}, \quad (5)$$

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

Assessing model performance

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

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 the MultiBLUP and GBLUP models were compared using a one-sided, paired Welch's t-test for unequal variances.

Generation of an empirical null distribution

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

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

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

Correction for multiple testing

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

Identifying biological pathways of interest

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

1.) The proportion of heritability explained was significantly greater than empirical values for random gene groups of the same size (FDR-adjusted *P*-value < .10),

2.) The LR test statistic was significantly greater than empirical values for random gene groups of the same size (FDR-adjusted *P*-value < .10),

3.) The MultiBLUP model significantly improved predictive ability compared to the GBLUP model (FDR-adjusted *P*-value < .10).

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

Data availability statement

Genotype data were previously published (Atwell *et al.* 2010) and were accessed from github.com/Gregor-Mendel-Institute/atpolydb/wiki. The scripts and phenotypic data supporting the conclusions of this article are publicly available as a Snakemake workflow (v5.4.2, Köster and Rahmann 2012) on GitHub at github.com/mishaploid/aa-genomicprediction (archived at https://doi.org/10.5281/zenodo.4048850). Free amino acid traits and details on ratio calculations are provided in **Table S1**. A list of ABRC stock names and accession numbers for each individual is in **Table S2**.

RESULTS AND DISCUSSION

In this study, we applied a genomic partitioning model to evaluate the contribution of metabolic pathways to FAA traits in seeds. The combination of a genomic partitioning framework

and the model system *Arabidopsis* allowed us both to test the feasibility of this approach and to further examine the relative contribution of each pathway to the genetic basis of FAA traits in seeds. Additionally, because FAA traits are part of core metabolism that is highly conserved, we hypothesize that our findings can be used to develop hypotheses in crop systems, where there is potential to contribute to the biofortification of essential amino acids.

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

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

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

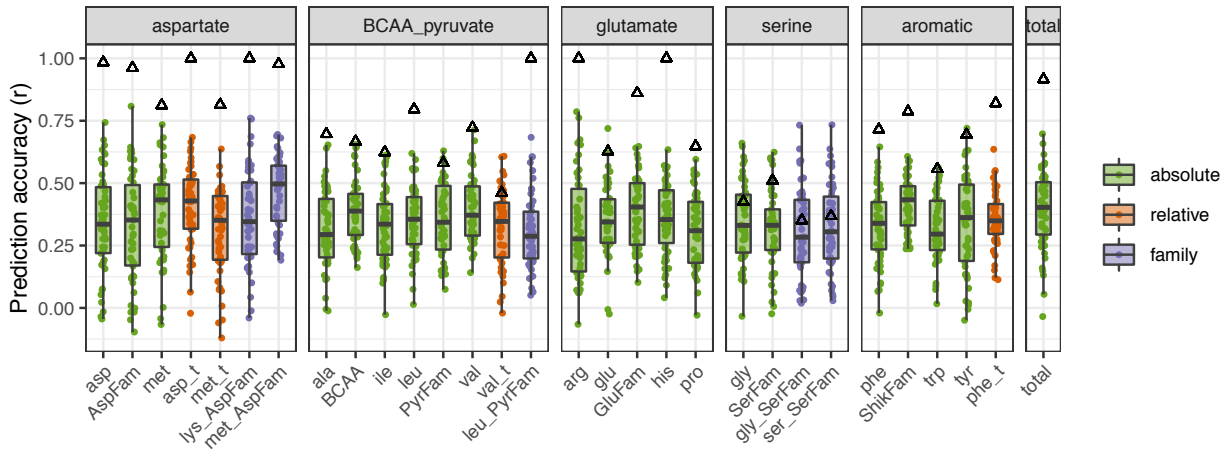

Figure 1. Genomic prediction performed well for a higher proportion of absolute traits compared to relative and family-based ratio traits.

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 (FAA) traits using a GBLUP model.

			Predictive ability (r)		Reliability (r^2)			
Trait type	**Metabolic family**	**Trait**	**mean**	**SE**	**mean**	**SE**	**slope (bias)**	**RMSE**
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

Traits are grouped by the type of trait (absolute level, relative to total FAA content, and family ratio) and metabolic family based on shared precursor. *SE*, standard error; R*MSE*, root mean squared error.

375

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

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

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

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

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

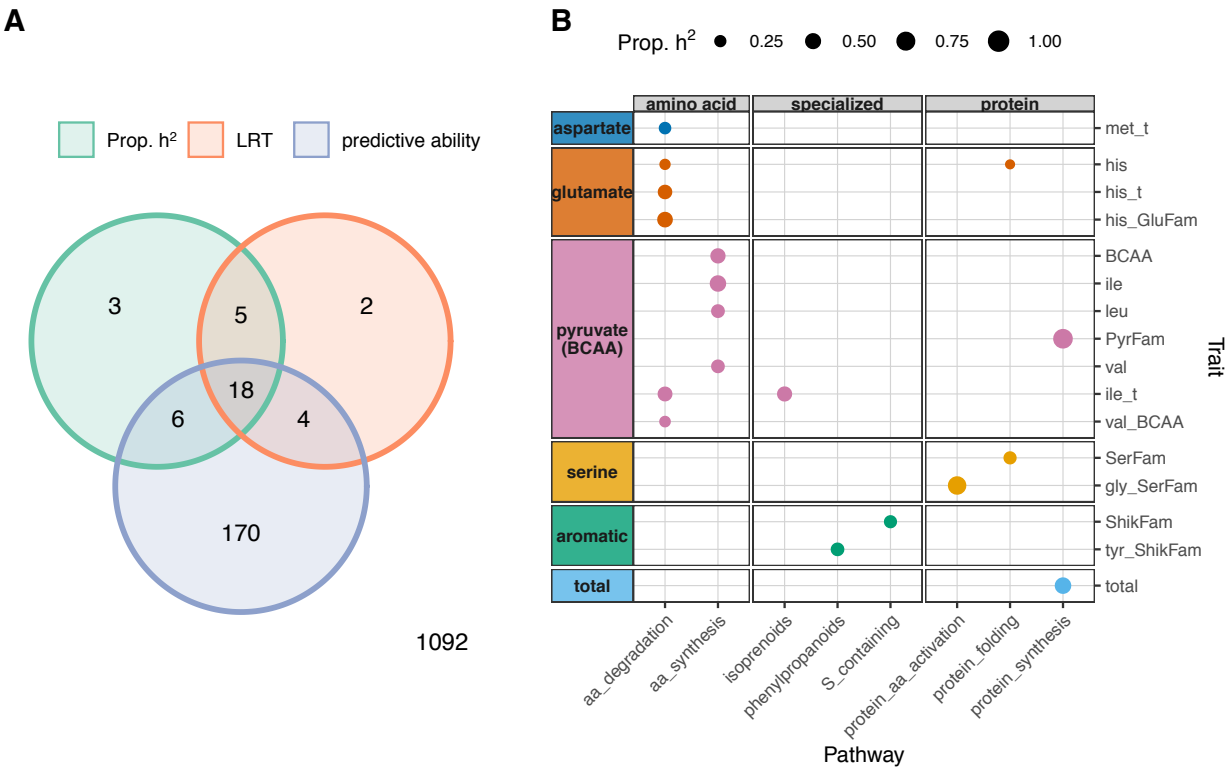

Figure 2. Biological pathways explain significant variation and improve predictive ability for free amino acid traits when incorporated into a MultiBLUP model.

(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) Points indicate trait-pathway combinations that passed all three significance criteria. The diameter of each point 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.

Table 3. Free amino acid traits and pathway combinations for which the MultiBLUP model explains a significant proportion of heritability, improves model fit relative to random gene groups of approximately the same size, and increases accuracy compared to GBLUP.

			Prop. h^2 explained			Likelihood ratio			Predictive ability					
Category	**Pathway**	**Trait**	**Prop. 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

Bolded rows indicate trait and pathway combinations that increased predictive ability by more than 5% compared to a GBLUP model. The difference in slope between the MultiBLUP and GBLUP models was computed as $|slope_{MultiBLUP} - 1| - |slope_{GBLUP} - 1|$. *RMSE*, root mean squared error.

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

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

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

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

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

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

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

Gene annotations for specialized metabolism are associated with FAA precursors

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

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

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

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

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

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

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

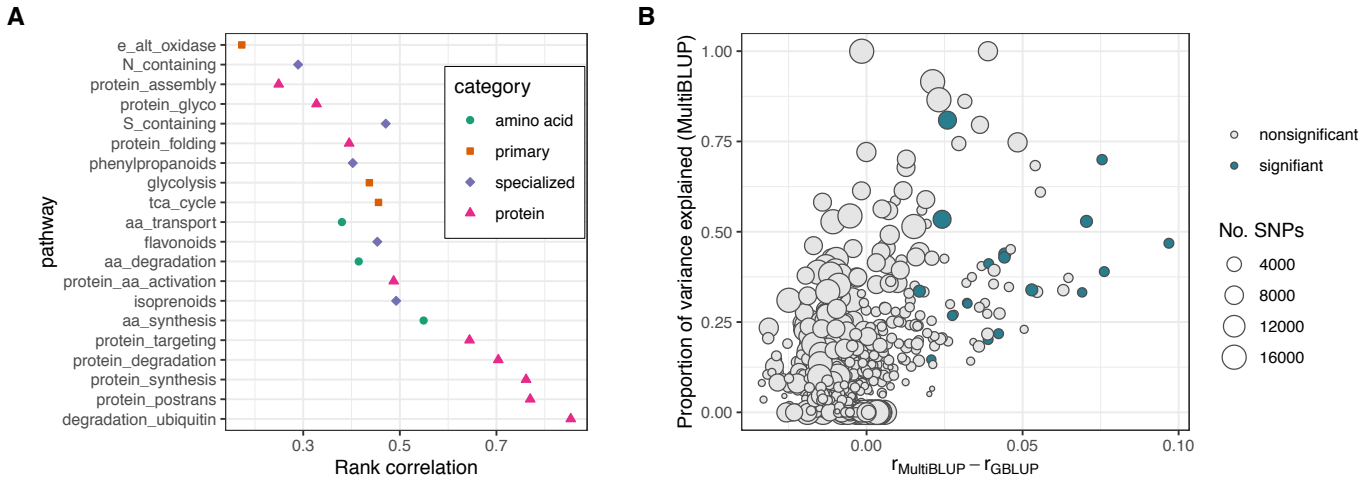

Figure 3. Pathway size influences the proportion of heritability explained and predictive ability when using a MultiBLUP model.

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

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