Subset-based genomic prediction provides insights into the genetic architecture of free amino acid levels in dry *Arabidopsis thaliana* seeds

Kevin A. Bird[*], Sarah D. Turner[*], Timothy M. Beissinger[*, †, ‡], Ruthie Angelovici[*]

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

[†] USDA-ARS, Columbia, Missouri

[‡] Division of Plant Sciences, University of Missouri, Columbia

Correspondence may be addressed to:

Ruthie Angelovici
Division of Biological Sciences
University of Missouri, Columbia, 65211
angelovicir@missouri.edu
(573) 882-3440

Timothy Beissinger
USDA Plant Genetics Research Unit
University of Missouri, Columbia, 65211
timbeissinger@gmail.com
(573) 882-7606

Running Title: Subset-based genomic prediction of free amino acids in dry *Arabidopsis* seeds

Keywords: Genomic prediction, *Arabidopsis thaliana*, free amino acids, genetic architecture

Abstract:

Amino acids play a central role in plant growth, development, and human nutrition. A better understanding of the genetic architecture of amino acid traits will enable researchers to integrate this information for plant breeding and biological discovery. Despite a collection of successfully mapped genes, a fundamental understanding of the types of genes driving the genetic architecture of amino acid related traits in crop seeds and model systems such *Arabidopsis* has remained unresolved. To address this issue, we applied genomic prediction using distinct subsets of genes, including those belonging to the known amino acid biochemical pathways, to quantify their contributions to the genetic variation of free amino acid levels in dry seeds. First, we demonstrate that genomic prediction of free amino acid levels is moderately accurate in *Arabidopsis* seeds. Then, we explore whether specific subsets of SNPs corresponding to amino acid pathways exhibit enhanced predictability for amino acid traits. Surprisingly, for several of the traits we studied, SNPs within the amino acid pathways were no more predictive than randomly generated sets of control SNPs. This may imply a complex genetic architecture that includes other genes related to cellular processes or development. Conversely, a subset of amino acid traits did exhibit enhanced predictability based on pathway SNPs compared to control SNPs. We propose that this latter set of traits may correspond to a simpler genetic architecture. Ultimately, this study provides a potential strategy to assess the involvements of candidate genes in the genetic architecture of a traits using subset-based genomic prediction.

INTRODUCTION

Amino acids play a central role in plant growth and development, primarily by acting as building blocks for proteins. However, free amino acids are also precursors for various metabolites and cellular processes such as signaling, nitrogen assimilation, secondary metabolism, and osmotic tolerance (Wu and Messing 2014; Angelovici et al. 2010). In human and farm animal diets, plants are a major source of these necessary amino acids (Ufaz and Galili 2008). Within plant seeds, free amino acids also contribute to desiccation and seed vigor pathways, which impact seed nutritional quality traits (Galili et al. 2014; Angelovici et al. 2011). Previous work has demonstrated that the free amino acid pools in seeds can be more easily manipulated for breeding and nutritional biofortification purposes than protein-bound amino acids (Galili and Amir 2013; Zhu and Galili 2003). The importance of amino acid content for both plant development and human nutrition has motivated researchers to employ multiple techniques to dissect the genetic architecture of amino acids in plants, with the ultimate goal of providing foundational genetic knowledge that can be integrated into plant breeding and biotechnology.

One of the most powerful methods for addressing the genetic basis of metabolic traits, such as free amino acids, is linkage mapping, which has frequently been used to uncover the genetic architecture of both primary and secondary plant metabolites (Fernie and Tohge 2017). Linkage mapping involves using a population produced from a cross between two or more distinct individuals to associate genotypic information with phenotypic values, enabling identification of quantitative trait loci (QTLs; Fernie and Klee 2011; Foerster et al. 2015). Previous QTL studies have successfully identified genomic loci that contribute to primary and secondary metabolic traits (Wong et al. 2004; Wentzell et al. 2007; Chander et al. 2008; Balasubramanian et al. 2009; Vallabhaneni and Wurtzel 2009; Gutiérrez-Rojas et al. 2010; Maloney et al. 2010; Kochevenko and Fernie 2011) and, more importantly, unveiled the complexity underlying the genetics of primary metabolite traits. QTL mapping studies have identified hundreds of QTLs associated with a large number primary metabolites in maize (Wen et al. 2015) and *Arabidopsis* (Lisec et al.2008; Rowe et al. 2008; Knoch et al. 2017). Frequently,

identified QTLs were found to be unevenly distributed into QTL hotspots and often associated with expression QTLs (eQTLs), suggesting pleiotropy among regulatory genes controlling primary metabolism (Lisec et al. 2008; Rowe et al. 2008; Knoch et al. 2017). In addition to pleiotropy, epistasis has been commonly identified for primary metabolism QTLs (Rowe et al. 2008; Knoch et al. 2017). QTLs identified in these studies tend to exhibit relatively small effect sizes and explain a small proportion of total phenotypic variance. The unexplained portion of primary metabolism and amino acid variance may be due to several unidentified loci with small contributions to primary metabolite levels in *Arabidopsis*.

As generating dense genotypic information for large numbers of individuals has become increasingly affordable, so has the use of genome-wide association studies (GWAS). GWAS uses a large panel of diverse individuals with genotypic and phenotypic data to identify loci associated with traits of interest (Hirschhorn and Daly 2005; Weigel and Nordborg 2005), providing higher resolution than linkage mapping and allowing appreciably more allelic diversity to be present in the study population (Korte and Farlow 2013; Fernie and Tohge 2017). GWAS is able to identify candidate loci for amino acids and primary metabolites both independently (Riedelsheimer et al. 2012), and in conjunction with QTL studies (Angelovici et al. 2013). However, the number and the effect size of loci detected so far explain only a fraction of the observed phenotypic variation of amino acid traits, and some amino acids have proven harder to dissect than others. For example, Angelovici et al. (2013, 2016) found the strongest associations for branched-chain amino acids (BCAA) and weak signals for most other free amino acid traits. Although GWAS has limited power to reliably identify variants that are rare and/or of small effect (Korte and Farlow 2013), subsequent investigations using integrated analyses that combined GWAS, linkage mapping, and metabolic correlation networks were able to replicate previously-identified candidate loci and identify new candidate loci related to amino acid and free amino acid levels in both seeds and leaves of *Arabidopsis* (Angelovici et al. 2016; Wu et al. 2016).

The consistent finding that amino acid traits frequently have several associated loci, coupled with the difficulty of GWAS to explain a large proportion of genetic variation for these traits, suggests the strong possibility that amino acid traits may have a highly polygenic

architecture (Chen et al. 2014). While linkage mapping and GWAS are typically underpowered to map loci contributing to polygenic traits, genomic prediction methods excel at providing information when traits are polygenic. Genomic prediction allows researchers to predict an individual's breeding value, or the additive component of their genetic variation, based only on genotypic data (Meuwissen et al. 2001; Heffner et al. 2009). The efficacy of genomic prediction results from its simultaneous use of all genotyped markers, in contrast to analyzing markers one-at-a-time as is done for linkage mapping and GWAS (Heffner et al. 2009). This allows the inclusion of information from all loci to make predictions, instead of basing conclusions only on individual loci that achieve genome-wide significance, and therefore captures more of the additive genetic variance.

Recent work on complex traits in humans has led to the proposal of the omnigenic model as a rediscovery of R.A. Fisher's infinitesimal model of quantitative traits (Boyle et al. 2017, Fisher 1918). Together, these models propose that the bulk of heritability for complex traits can be found in exceedingly small-effect loci that are dispersed throughout the genome, potentially, though not necessarily, outside of trait-relevant pathways (Boyle et al. 2017). If this applies to free amino acid traits in dry *Arabidopsis* seeds, large-effect genes involved in known amino acid pathways may not be informative regarding free amino acid levels. Instead, one would expect that seemingly unrelated genes and genomic regions distributed throughout the genome may harbor the majority of genetic variability contributing to free amino acid levels. In contrast, experiments using linear mixed models and genomic prediction methods to investigate complex traits in *Drosophila melanogaster* (Edwards et al. 2016), Duroc boars (*Sus scrofa;* Sarup et al. 2016), and dairy cattle (*Bos taurus*; Edwards et al. 2015, Fang et al. 2017) have shown improved predictions by incorporating differential weightings for trait-relevant pathways that are enriched for previously-identified candidate genes. This suggests that for some traits in some species, phenotypic variance may be better-explained by genetic variation in trait-relevant pathways than those found in pathways unrelated to the trait. Similarly, when Owens et al. (2014) tested the predictive power of genome-wide SNPs, pathway-relevant SNPs from genes identified in previous linkage studies, and SNPs from eight particular genes in a carotenoid QTL study, they found that SNPs beyond the eight that were previously-implicated provided little to no

125 improvement in predictive power. This is consistent with the possibility that carotenoids in maize
126 exhibit a relatively simple genetic architecture. Together, these studies imply that, along with
127 providing predictions for breeding purposes, applying genomic prediction methods to subsets of
128 the genome can provide novel insights into the genetic architecture of quantitative traits.

129 In this study, we first demonstrate the efficacy of whole-genome prediction for free
130 amino acid levels in dry *Arabidopsis* seeds. Next, we establish that a subset of SNPs restricted to
131 genes within the amino acid pathway are not as informative for genomic prediction as SNPs
132 from all genes in the *Arabidopsis* genome. Then, we test whether amino acid-related pathway
133 SNPs are better predictors of free amino acid levels in dry *Arabidopsis* seeds than randomly
134 selected control subsets with an equal number of SNPs, which were selected from the rest of the
135 genome. This application of subset-based genomic prediction sheds light on the genetic
136 architecture of these traits. Importantly, we observed that the predictive power of the SNPs
137 within the amino acid pathway genes, as compared to their controls, varied significantly from
138 trait to trait. This implies a varying level of genetic complexity among free amino acid levels in
139 dry *Arabidopsis* seeds.

140 MATERIALS AND METHODS

141 **Phenotypes and Plant Materials:** For this study, we reanalyzed data on the absolute levels of
142 free amino acids in dry *Arabidopsis* seeds, which were previously measured for a 312 member
143 association panel (Angelovici et al. 2013, 2016). In brief, seeds were harvested from three
144 independent replicates of the 360-accession *Arabidopsis* diversity panel (Nordborg et al. 2005)
145 and absolute levels of free amino acids (nmol/mg seed) were quantified using liquid
146 chromatography–tandem mass spectrometry (see Angelovici et al. 2013 for further details).
147 Eighteen of the 20 proteogenic amino acids were measured, in addition to a composite
148 phenotype: the sum of all the free amino acids measured (total free amino acids; TFAA). For
149 each phenotype and each accession, best linear unbiased predictors (BLUPs) were generated by
150 fitting mixed models across all three replicates. These BLUPs, reported in Angelovici et al.
151 (2013) were used as the phenotypic data for this study. The BLUPs effectively removed the

152 effect of environment and reduced the number of phenotypic measurements for downstream
153 analysis from three per accession to one, but they do not take genetic information into account.

154 **Genotyping and SNP Subsetting:** Accessions were previously genotyped using a 250k SNP
155 panel (Atwell et al. 2010), version 3.06. Unless otherwise specified, all analyses were conducted
156 in R 3.4.3 (R Core Team 2015).

157 The full genome-wide SNP set was used to create three different SNP subsets: genic SNPs (GS
158 subset), SNPs related to amino acid metabolism (AAS subset), and control SNPs (CS subset), as
159 described below. These three SNP subsets were used throughout this study to assess the
160 predictive power of genes in the amino acid pathway as compared to all genes and control genes.

161 *Genic SNP (GS) subset:* Using the R package biomaRt (Durinck et al. 2005, 2009), a subset
162 comprised of all genic SNPs was generated by selecting SNPs that fell within a 5kb window of
163 the start and stop positions for all annotated genic regions of the *Arabidopsis thaliana* genome in
164 TAIR10 (Swarbreck et al. 2008). In total, this subset included 213,612 SNPs.

165 *Amino Acid SNP (AAS) subset:* The AraCyc database (Mueller et al. 2003; Zhang et al. 2005)
166 was used to identify *Arabidopsis thaliana* genes associated with amino acid biosynthesis and
167 catabolism. The start and stop positions of genes associated with amino acids were retrieved
168 from the biomaRt database (Durinck et al. 2005, 2009), comprising a total of 335 genes
169 (Smedley et al. 2015). SNPs that fell within a 5kb window of the start and stop positions of these
170 genes were selected from the full genome-wide SNP set to form an amino acid subset of 7,667
171 SNPs. SNPs up to 5kb outside of genes were permitted so that up- and downstream regulatory
172 elements would generally be included in the SNP subset.

173 *Control SNP (CS) subsets:* 1,000 random subsets of SNPs were generated to serve as controls for
174 the AAS subset. These subsets represented the same number of genes and contained the same
175 number of SNPs as the AAS subset, but the genes were randomly chosen from all non-amino
176 acid annotated genes in *Arabidopsis* without incorporating any additional pathway information.
177 Aside from the number of SNPs in the CS subsets and the total number of genes they
178 represented, there were intentionally no additional restrictions placed on which SNPs could be

179 chosen as part of the CS subsets. This was done in order to minimize potential sources of bias in
180 our analysis.

181 **Genomic Prediction using GBLUP:** For each SNP subset, including the 1,000 CS subsets,
182 genomic relationship matrices were constructed using the VanRaden algorithm (VanRaden 2008)
183 as implemented in the GAPIT software package (Lipka et al. 2012). Based on these matrices,
184 genomic best linear unbiased prediction (GBLUP; Zhang et al. 2007) was conducted in GAPIT
185 for each of the BLUP-transformed free amino acid traits using a mixed linear model without
186 compression (Lipka et al. 2012).

187 First, to establish the overall predictive accuracy of the genic SNP set for amino acid
188 traits, one iteration of GAPIT with 1,000 replications and 3-fold cross validation was run on the
189 GS subset. The correlation coefficient between observed and predicted phenotypic values was
190 recorded for each run and plotted using the R package ggplot2 (Wickham 2016). To compare the
191 predictive accuracy of the GS and AAS, 1000 replicates of 3-fold cross validation were
192 performed in GAPIT using the AAS subset, and these were compared to the 1,000 replicates
193 using the GS subset based on a Welch two sample t-test.

194 Next, for each of the 1,000 CS subsets, genomic prediction was performed 100 times
195 with 3-fold cross validation. The analysis of these 1,000 CS subsets were paired to 1,000
196 analyses of the AAS subset, each of which were also conducted with 100 replicates of 3-fold
197 cross validation. Statistical comparison of AAS and CS subsets was carried out using a binomial
198 test. For each amino acid trait, 1,000 pairwise trials were performed with success defined as the
199 AAS subset having a greater correlation coefficient than the CS subset. Then, based on the
200 number of successes, we evaluated whether or not the probability of success was significantly
201 greater than 50%, which corresponds to testing whether AAS subsets are significantly more
202 predictive than CS subsets.

203 Finally, since a significant difference doesn't necessarily imply a large difference in
204 magnitude, particularly when the number of trials is large, a second set of binomial tests were
205 conducted. For these, similar pairwise trials to those described above were conducted, with
206 success being defined as the correlation coefficient of the AAS subset being at least 5% greater

than that of the CS subset. Again, after 1,000 pairwise comparisons (one for each CS subset) we used a binomial test to determine if the observed number of success was significantly greater than 50%. This is equivalent to testing whether or not AAS subsets out-predict CS subsets by at least 5%.

Genomic Prediction using GFBLUP: In parallel to the GBLUP analysis with GAPIT, genomic feature best linear unbiased prediction (GFBLUP; Edwards et al. 2015, Edwards et al. 2016, Sarup et al. 2016, Fang et al. 2017) was used to estimate the proportion of genomic variance attributable to markers in the AAS and CS subsets. This approach follows the method used by Edwards et al. (2015) to assess the influence of genetic markers in biological pathways on health and milk production in dairy cattle. Briefly, the GFBLUP model is an extension of GBLUP that incorporates the additional effect of genetic markers that have prior evidence of an association with a given trait (e.g., gene annotation categories or QTL) (Edwards et al. 2015, Edwards et al. 2016, Sarup et al. 2016). Thus, the full set of genomic markers are still used for prediction, with the potential for marker subsets (i.e., genomic features) to be differentially weighted if they are potentially causal for the phenotype of interest.

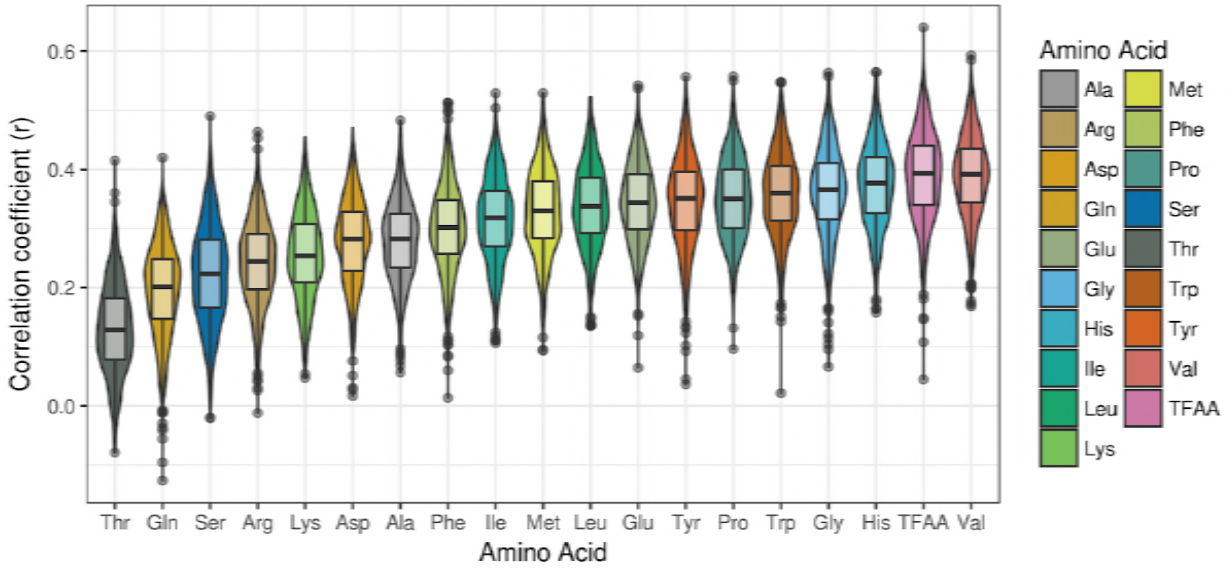
For our analysis, 1,000 models were fit, each including the AAS subset and one of the 1,000 CS subsets as genomic features. This allowed a direct comparison between the proportion of additive genomic variance (h^2) explained by the AAS subset $\left(h^2_{AAS} = \frac{\sigma^2_{AAS}}{\sigma^2_{GS} + \sigma^2_{AAS} + \sigma^2_{CS}}\right)$ and each of the 1,000 CS subsets $\left(h^2_{CS} = \frac{\sigma^2_{CS}}{\sigma^2_{GS} + \sigma^2_{AAS} + \sigma^2_{CS}}\right)$ (Edwards et al. 2015, Edwards et al. 2016, Sarup et al. 2016). The GFBLUP model was implemented via the qgg package in R (Sørensen 2017) using the REML procedure with 50 replications and 10-fold cross validation. The proportion of genomic variance explained for each genomic feature was averaged across validation sets to generate a total of 1,000 data points per amino acid trait. Similar to above, a binomial test was used to test if h^2_{AAS} was greater than h^2_{CS} for the 1,000 comparisons. As for the GBLUP results, this was followed by a second set of binomial tests to establish if the proportion of genomic variance explained by the AAS subset was at least 5% greater than that explained by the CS subset.

Computational Resources: Computation was performed using the University of Missouri Informatics Core Research Facility BioCluster (https://bioinfo.ircf.missouri.edu/). Computational nodes used to calculate genomic relationship matrices and perform genomic prediction had 64 cores and 512 GB of RAM.

Data Availability: Genotypic data for all accessions utilized is available at https://github.com/Gregor-Mendel-Institute/atpolydb/wiki. Phenotypic data are available in Supplemental File S1. All scripts used in this study are publicly available on GitHub at https://github.com/KevinABird/AA-GenomicPrediction.

RESULTS

Efficacy of genomic prediction: We first evaluated the overall accuracy of genomic prediction of free amino acid levels in dry *Arabidopsis* seeds. For this evaluation, we employed the GS subset only, as we were interested in general predictability. Our results indicate that genomic prediction is reasonably effective for absolute levels of free amino acids (nmol/mg seed) and the composite trait TFAA, with mean correlation coefficients (r) over 1,000 runs ranging from ~0.13 for threonine (Thr) to ~0.39 for TFAA (Figure 1, Table 1, Figure S1). As is expected for polygenic traits, whole genome prediction enables useful predictions of phenotype from genotype, despite the difficulty identifying large effect causal variants using GWAS.



251 **Figure 1:** Distribution of predictive accuracy of the GS subset for each of the 19 amino acid traits from 1000 3-CV
252 replicates of GAPIT.

253 **Utility of Amino Acid pathway genes for predicting free amino acid levels in dry**
254 ***Arabidopsis* seeds with GAPIT:** Our results from genomic prediction using amino acid
255 pathway-specific SNP subsets highlight the complexity of genetic architecture in free amino acid
256 traits. In addition to genomic prediction using the GS subset, we evaluated genomic prediction
257 using the AAS subset. For all traits, the mean predictability of the GS subset was higher than that
258 of the AAS subset, and, for all but Gln, this difference was significant according to a simple
259 t-test (Table 1). However, the GS subset contains several-fold more SNPs than the AAS subset,
260 so this was not unexpected. Still, it suggests the possibility that the amino acid pathway is
261 unlikely to harbour all of the variability for amino acid traits. To evaluate whether or not AAS
262 subsets contained elevated information relative to the background level of population structure
263 and frequency of causal SNPs genome-wide, we compared prediction with the AAS subset to
264 that using CS subsets. Recall that 1,000 CS subsets were specifically designed to serve as
265 random controls for the AAS subset, which included only SNPs within or near genes that are
266 annotated as part of the amino acid pathways (Mueller et al. 2003; Zhang et al. 2005). For every
267 amino acid trait tested, the GS had higher predictive accuracy than the AAS and CS, although the
268 degree to which the GS had higher predictive accuracy varied between amino acid traits (Table
269 1). Results from principal component analysis (PCA) indicated that population structure was
270 similar across the GS, AAS, and CS subsets (Figure S2).

271 **Table 1:** Mean predictive performance (r) for each of 19 traits using different AAS and GS SNP subsets. P-values
272 shown indicate whether the differences between AAS and GS prediction is significant based on a t-test.

Genomic Prediction: *r*

Trait	AAS	GS	P-value
Ala	0.245888	0.27881	< 2.2e-16
Arg	0.216031	0.242825	8.9e-16
Asp	0.191865	0.278783	< 2.2e-16

Gln	0.193691	0.195561	0.58
Glu	0.287493	0.343803	< 2.2e-16
Gly	0.294474	0.36171	< 2.2e-16
His	0.313663	0.373729	< 2.2e-16
Ile	0.298894	0.315852	3.7e-07
Leu	0.295554	0.338175	< 2.2e-16
Lys	0.213057	0.255043	< 2.2e-16
Met	0.260942	0.329325	< 2.2e-16
Phe	0.265411	0.300857	< 2.2e-16
Pro	0.327411	0.351283	7.4e-15
Ser	0.151919	0.222466	< 2.2e-16
Thr	0.123287	0.130962	0.027
Trp	0.326466	0.358462	< 2.2e-16
Tyr	0.309437	0.345061	< 2.2e-16
Val	0.328866	0.389133	< 2.2e-16
TFAA	0.285942	0.388012	< 2.2e-16

354 Using GAPIT, a binomial test showed that the AAS subset was significantly more
355 predictive than CS subsets for 12 out of 19 amino acid traits (Ala, Arg, Gln, His, Ile, Leu, Lys,
356 Phe, Pro, Thr, Trp, Tyr; Table 2, Figure 2A). However, for seven of these 19 traits (Ala, Arg, His,
357 Leu, Lys, Trp, Tyr), our analysis showed that the magnitude of AAS predictions did not
358 outperform CS predictions by at least 5%. This indicates that the higher predictive accuracy may
359 be statistically significant but low in magnitude for these seven amino acids. Angelovici et al.
360 (2013) identified BCAT2, a large-effect gene within the amino acid pathway for isoleucine (Ile)
361 content in dry seeds. In agreement with this, the AAS subset was more predictive than CS
362 subsets for Ile (Figure 2). For the remaining seven traits (Asp, Glu, Gly, Met, Ser and Val),

including the branched chain amino acid Val, predictions from the AAS subset were not significantly higher than predictions from the CS subsets (Table 1, Figure 2B).

Predictive accuracy of the AAS subset was further tested with and without BCAT2 to investigate the effect of BCAT2 on Ile prediction. Pairwise comparison of the AAS subsets with and without BCAT2 showed that the AAS subset with BCAT2 had a higher correlation coefficient in 992 out of 1000 comparisons and was significantly more predictive ($P < 2.2e-16$). Additionally, the AAS subset with BCAT2 had a correlation coefficient that was 5% higher than the AAS subset without BCAT2 in 648 out of 1000 comparisons ($P < 2.2$ e-16). However, the AAS subset without BCAT2 was more predictive than the CS subsets in in only 548 out of 1000 comparisons ($P= 0.001$), compared to 752 out of 1000 when BCAT2 is included ($P< 2.2e-16$). The AAS subset had a correlation coefficient 5% higher than the CS subset in only 396 out of 100 comparisons, compared to 624 when BCAT2 is included ($P=2.07e-15$).

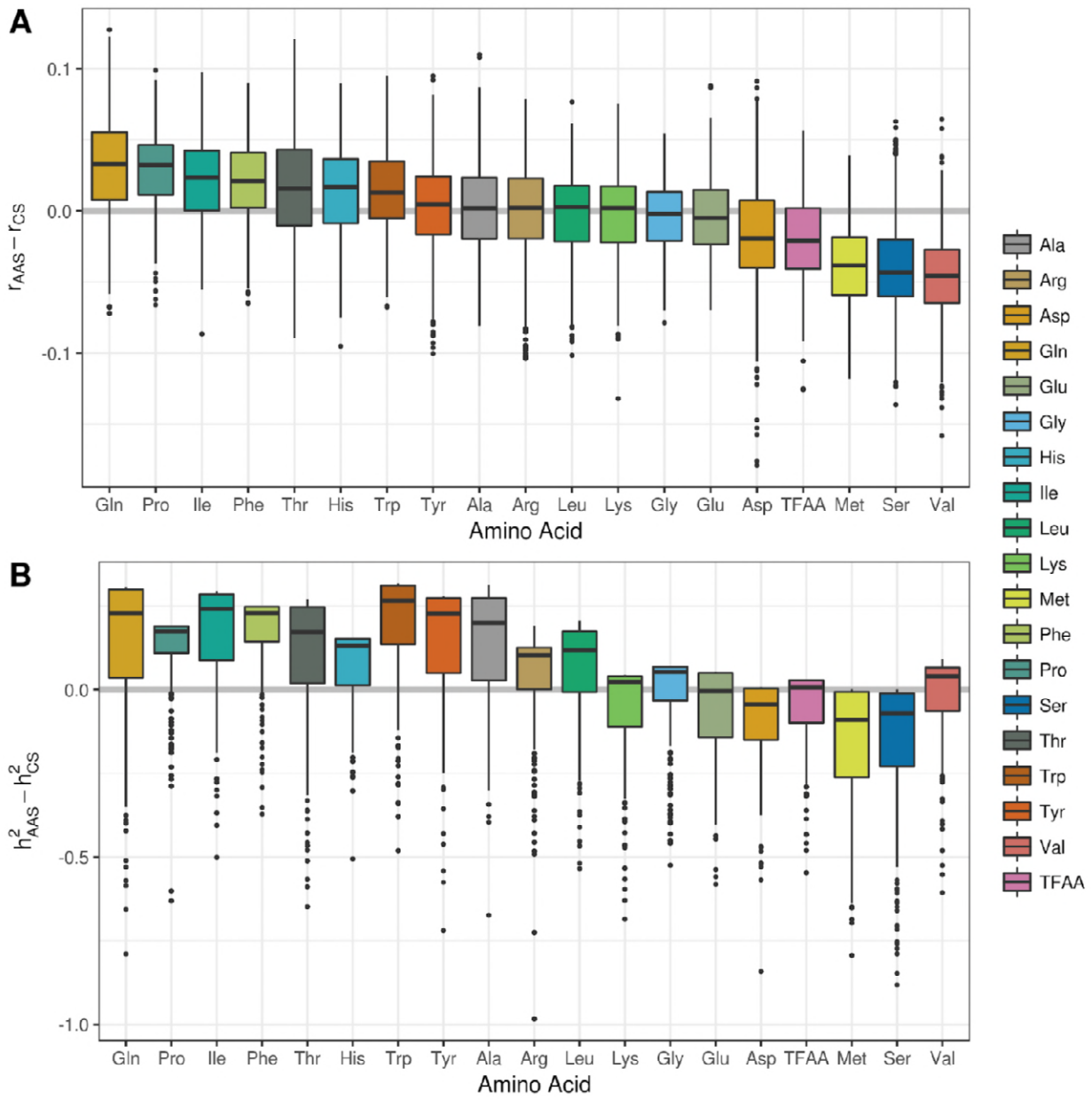


Figure 2: Comparison between the AAS subset and CS subsets. (**A**) Difference in genomic prediction accuracy between the AAS subset (r_{AAS}) and the CS subsets (r_{CS}) as estimated from GAPIT. (**B**) Difference between the proportion of additive genomic variation explained for amino acid traits by the AAS subset (h^2_{AAS}) and by the CS subset (h^2_{CS}) in the GFBLUP model. Note that the same AAS subset was compared to 1,000 independent CS subsets.

Table 2: Binomial test results comparing the AAS subset to 1,000 CS subsets. Test results are presented for relative predictive accuracy using GAPIT and for the relative proportion of additive genomic variance explained using GFBLUP.

	GAPIT				**GFBLUP**			
Amino Acid	**# AAS >CS**	**p-value**	**# AAS >CS + 5%**	**p-value**	**# AAS > CS**	**p-value**	**# AAS > CS + 5%**	**p-value**
Ala	536*	0.01	381	1	783*	< 2.2e-16	771*	< 2.2e-16
Arg	530*	0.03	390	1	750*	< 2.2e-16	738*	< 2.2e-16
Asp	314	1	240	1	304	1	304	1
Gln	801*	< 2.2e-16	737*	< 2.2e^-16	791*	< 2.2e-16	779*	< 2.2e-16
Glu	411	1	263	1	495	0.636	493	0.6824
Gly	465	0.99	236	1	695*	< 2.2e-16	692*	< 2.2e-16
His	663*	< 2.2e-16	518	0.13	786*	< 2.2e-16	779*	< 2.2e-16
Ile	752*	< 2.2e-16	624*	2.1e^15	802*	< 2.2e-16	785*	< 2.2e-16
Leu	535*	0.01	305	1	742*	< 2.2e-16	715*	< 2.2e-16
Lys	528*	0.04	362	1	570*	5.355e-06	570*	5.355e-6
Met	87	1	27	1	157	1	157	1
Phe	775*	< 2.2e-16	614*	2.8e^-13	891*	< 2.2e-16	878*	< 2.2e-16
Pro	860*	< 2.2e-16	696*	< 2.2e^-16	881*	< 2.2e-16	881*	< 2.2e-16
Ser	111	1	88	1	125	1	125	1
Thr	660*	< 2.2e-16	615*	1.8e^-13	768*	< 2.2e-16	752*	< 2.2e-16
Trp	676*	< 2.2e-16	463	0.99	845*	< 2.2e-16	839*	< 2.2e-16
Tyr	559*	0.01e-2	374	1	801*	< 2.2e-16	787*	< 2.2e-16
Val	53	1	19	1	630*	< 2.2e-16	609*	2.81e-12
Total	261	1	131	1	525	0.06061	518	0.1342

Comparing the proportion of genomic variance explained by AAS and CS subsets using GFBLUP: As an alternative test, the relative proportions of genomic variance explained by the AAS (h^2_{AAS}) and CS (h^2_{CS}) subsets were compared using GFBLUP. Results from the binomial test comparing h^2_{AAS} to h^2_{CS} generally agree with the results from GAPIT, indicating that the

AAS subset explained more genomic variance than the CS subset for Ala, Arg, Gln, His, Ile, Leu, Lys, Phe, Pro, Thr, Trp, and Tyr (Figure 2B, Table 2). In contrast to the results from GAPIT, the AAS subset was also more predictive than the CS subsets for Gly and Val. This discrepancy is especially dramatic for Val; according to our GAPIT GBLUP analysis, the AAS subset was more predictive than the CS subsets in only 53 out of 1,000 tests. However, based on GFBLUP, we found a greater proportion of variance explained by the AAS subset in 630 out of 1,000 tests (Figure 2, Table 2). Notably, the AAS subset explained equal or less genomic variation than the CS subsets for total free amino acids (TFAA), Glu, Asp, Ser, and Met, consistent with the results from GAPIT and thus indicating that the AAS subset is reliably not informative for these amino acid traits. These results were unchanged when we tested whether there was at least a 5% difference in the proportion of genomic variance explained by the AAS and CS subsets. Because a 5% difference may not be as informative for the proportion of genetic variance explained as for predictive accuracy, we also tested for a difference of up to 50% and found no changes in the overall patterns of significance (Table S1). Additionally, Figure S3 demonstrates that when the AAS subset explained significantly more genomic variance than the CS subset, the magnitude of this difference was quite large.

Interestingly, although our GFBLUP analyses demonstrated that the AAS subsets were more informative than controls for the majority of amino acid traits, we did not observe an increase in prediction accuracy for any of the 19 amino acid traits when using GFBLUP compared to GBLUP (Figure S4A). Despite this, we did observe differences among the 19 amino acid traits for the proportion of genomic variance explained by the AAS subset (Figure S4B). Amino acid traits for which the AAS subset explained a relatively high proportion of genomic variance (i.e., Trp, Gln, Ala, Ile, Thr, Phe, Tyr, Pro, Leu, and His) also showed predictive accuracies for the AAS subset that were similar to the accuracy of prediction when the entire GS subset was used (Table 1).

DISCUSSION

600 Since its development nearly two decades ago (Meuwissen et al. 2001), genomic prediction has
601 dramatically altered the speed and scale of applied genetic and breeding research (Daetwyler et
602 al. 2013). However, the use of genomic prediction has been primarily limited to agricultural
603 species (Wolc et al. 2016; Nielsen et al. 2016; Weller et al. 2017), likely because this is the realm
604 where predicted breeding values are most directly useful for breeding purposes. However, as
605 demonstrated in this manuscript, the methodology of genomic prediction can be meaningfully
606 applied for biological discovery in the context of basic science. Our study sheds light on
607 fundamental aspects that underlie the genetic architecture of free amino acid traits in *Arabidopsis*
608 seeds. As discussed below, our findings are consistent with existing literature, but also propose
609 new hypotheses.

610 **Genomic prediction of free amino acid levels in dry *Arabidopsis* seeds using a diverse panel**
611 **of individuals:** The panel of 312 *Arabidopsis* individuals we studied is diverse and represents a
612 substantial proportion of the known genetic variability present in *Arabidopsis* (Nordborg et al.
613 2005). This setting is distinct from the closed breeding populations of dairy cattle, maize, and
614 other agricultural species where genomic prediction is often utilized (Weller et al. 2017; Heffner
615 et al. 2009; Wolc et al. 2016). The moderate prediction accuracies that we observed for free
616 amino acid traits are promising, as they suggest the possibility of LD between markers and
617 causal loci, and serve as additional evidence that genomic prediction can successfully be applied
618 outside of crop species. Notably, we did not observe an increase in prediction accuracy when
619 using GFBLUP to differentially weight SNPs in amino acid pathway genes. This could result
620 from several factors, such as dilution by non-causal markers or inclusion of causal variants that
621 do not explain at least 10% of the genomic variance (Edwards et al. 2016, Sarup et al. 2016).

622 **Genetic variability for free amino acids in dry *Arabidopsis* seeds is not restricted to known**
623 **amino acid pathways:** For all but one of the amino acid traits we studied, the full GS subset was
624 more predictive than the AAS subset. More importantly, for six out of 18 amino acids studied
625 (~33%), as well as the composite trait TFAA (seven traits in all), the subset of SNPs restricted to
626 amino acid pathway genes (AAS subset) did not exhibit a greater predictive accuracy than the
627 same number of SNPs restricted to an equivalent number of random control genes (CS subsets).
628 These seven traits spanned four out of five families (Asp, Glu, BCAA and Ser amino acid

629 families). For an additional seven traits (Ala, Arg, His, Leu, Lys, Trp, Tyr), the AAS had
630 significantly higher predictive accuracy than the CS, but by a margin less than 5% of the
631 predictive accuracy for the CS, suggesting that the qualitative difference between the AAS and
632 the expected background predictive accuracy is minimal. This 5% margin was less meaningful
633 when using GFBLUP to compare the proportion of genomic variance explained, as the AAS
634 subset significantly explained at least 5% more genomic variance than the CS subset for all
635 seven of these amino acid traits. This could indicate that a 5% difference in predictive accuracy
636 is more extreme than a 5% difference in the proportion of genomic variance explained. Since the
637 effectiveness of genomic prediction often relies on linkage disequilibrium (LD) between SNPs
638 and causal alleles, this may imply that for these traits, an increased proportion causal loci fall
639 outside of known amino acid pathways. However, it is well established that genomic prediction
640 can succeed as a result of phenotypic correlations between related individuals manifesting as
641 co-segregation between markers and causal loci. This can be the case regardless of whether or
642 not LD between SNPs and causal alleles is present (de los Campos et al. 2013). Such a
643 phenomenon could contribute to the reason that the AAS subset does not outperform the CS
644 subsets for these seven traits, and produces significant, but low magnitude improvements for
645 another seven traits. Considering that the other five traits did exhibit significantly greater
646 predictive performance from the AAS subset than the CS subsets, it appears that in certain
647 situations there can be additional information contained in the amino acid pathway when
648 compared to random control regions. Interestingly, all three Shikimate family amino acids (Phe,
649 Tyr, and Trp) fell into the category for which the AAS subset was more predictive than CS
650 subsets.

651 This leads us to hypothesize that for the seven traits for which the AAS subset did not
652 out-predict the CS subsets, and to a lesser extent the additional seven traits for which the AAS
653 did not out-predict the CS subsets by at least 5%, causal alleles for these traits are more likely to
654 be small-effect and dispersed throughout the genome. These findings provide unique insights
655 into amino acid metabolism during late seed maturation and desiccation. To date it has remained
656 unclear whether the free amino acid pool in dry seeds is a direct consequence of active amino
657 acid metabolism (biosynthesis or catabolism) or the results of processes such as protein

658 degradation or synthesis, nitrogen metabolism, or secondary metabolism. For example, a
659 metabolic characterization study of *Arabidopsis* seeds during late maturation suggested that
660 toward the final desiccation stage there is an active metabolic switch that is responsible for
661 elevation of several free amino acids (Fait et al. 2006). Our findings suggest that a metabolic
662 switch with a simple underlying genetic architecture might affect only a subset of the free amino
663 acids (potentially Gln, Ile, Phe, Pro, and Thr, for which the magnitude of AAS subset-based
664 prediction was substantially and significantly greater than CS-subset based prediction), while the
665 other thirteen free amino acids levels may be determined by additional processes. Distinguishing
666 between the two architectures can help understand how and if we can manipulate these traits for
667 breeding, and how challenging this task might be. Consistent with this hypothesis, the AAS
668 subset did not significantly predict better than the CS subset for Met, an essential amino acid that
669 is deficient in most crop seeds (Galili and Amir 2013; Ufaz and Galili 2008). This suggests that
670 variation for Met is likely driven by a complex assortment of cellular processes beyond the
671 amino acid metabolic pathway. This is also in agreement with the difficulty of enhancing Met
672 without agronomical penalties (Galili et al. 2014; Galili and Amir 2013; Ufaz and Galili 2008).

673 **A collection of traits do rely on elevated genetic variability within the amino acid pathway:**
674 For five of the traits we studied, the predictive accuracy of the AAS subset was significantly
675 greater than that for the CS subset and by a margin greater than 5% of the mean predictive
676 accuracy of the CS. This provides evidence that genetic variability for these traits appears to be
677 more concentrated within the amino acid pathway. In the case of Ile, this is consistent with
678 results that have been previously published. Angelovici et al. (2013) detected a large-effect QTL
679 contributing approximately 19% of the observed variability of Ile. The causal gene found was the
680 Branched chain amino acid transferase 2 (At1g10070), which is part of the BCAA metabolic
681 pathway (Mueller et al. 2003; Binder 2010). Our re-analysis of the data with and without SNPs
682 near the causal gene included demonstrated that a significant proportion of the enhanced
683 predictive ability of the AAS subset can be attributed to this gene alone. This observation is
684 consistent with the possibility that subset-based genomic prediction may be implemented as a
685 technique to hone in on categories or features of SNPs/genes that are more likely to harbour
686 causal variants. However, even for these five traits, our results suggest that the amino acid

687 pathway does not contain all of their genetic variability. If it did, we would expect that the AAS
688 subset would be as or more predictive than the GS subset, which we only observed for Gln
689 (Table 1). These results suggests that candidate gene approaches may be limited by the genetic
690 variation outside of amino acid pathways that contribute greatly to trait variation. Future work in
691 genetic engineering and breeding of complex metabolic traits should bear in mind that important
692 genetic variation can lie outside of known amino acid pathway genes.

693 **CONCLUSIONS**

694 Based on our results, it appears that for a proportion of the free amino acid traits in dry
695 *Arabidopsis* seeds, known amino acid pathway genes contribute no more additive genetic
696 variation than do genes distributed genome-wide. Instead, we hypothesize that the genetic
697 architecture of these traits is made up of many loci of small effect from any number of different
698 biochemical and biological pathways. However, several free amino acid traits showed the
699 opposite pattern; they were better-predicted by SNPs within the pathway than by random SNPs
700 elsewhere in the genome. At least one of these counter-examples, Ile, corresponds to an amino
701 acid already demonstrated to have a simpler genetic architecture that includes natural variation
702 for a large effect allele segregating within the pathway (Angelovici et al. 2013). We propose that
703 those amino acids with high pathway-based predictability may have a simpler genetic
704 architecture than those that cannot be well-predicted from pathway-based SNPs alone. Broadly
705 speaking, a strategy similar to that implemented here may be useful for classifying the genetic
706 architecture of complex traits, especially of metabolic traits.

707 **Acknowledgements:**

708 We are grateful for helpful input from two anonymous reviewers, whose suggestions improved
709 this manuscript. This project was supported by the USDA Agricultural Research Service, the
710 University of Missouri Division of Biological Sciences, Columbia Mo, the NSF Graduate

Research Fellowship Grant No. DGE-1424871 for KAB, and the NSF Postdoctoral Research Fellowship in Biology Grant No. 1711347 for SDT .

References

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 Physiology*, January, pp.01287.2016. doi:10.1104/pp.16.01287.

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." *The New Phytologist* 189 (1): 148–59. doi:10.1111/j.1469-8137.2010.03478.x.

Angelovici, R., G. Galili, A. R. Fernie, and A. Fait, 2010. "Seed Desiccation: A Bridge between Maturation and Germination." *Trends in Plant Science* 15 (4): 211–18. doi:10.1016/j.tplants.2010.01.003.

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." *The Plant Cell Online* 25 (12): 4827–43. doi:10.1105/tpc.113.119370.

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 (7298): 627–31. doi:10.1038/nature08800.

Balasubramanian, S., C. Schwartz, A. Singh, N. Warthmann, M. C. Kim, et al., 2009. "QTL Mapping in New Arabidopsis Thaliana Advanced Intercross-Recombinant Inbred Lines." *PLOS ONE* 4 (2): e4318. doi:10.1371/journal.pone.0004318.

Binder, S., 2010. "Branched-Chain Amino Acid Metabolism in Arabidopsis Thaliana." *The Arabidopsis Book* 8: e0137. doi:10.1199/tab.0137.

Boyle, E. A., L. I. Yang,, and J.K. Pritchard, 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." *Cell* 169 (7): 1177–86. doi:10.1016/j.cell.2017.05.038.

Campos, G. de los, and D. Sorensen. 2014. "On the Genomic Analysis of Data from Structured Populations." *Journal of Animal Breeding and Genetics* 131 (3): 163–64. doi:10.1111/jbg.12091.

Campos, G. de los, A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D.Sorensen. 2013. "Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor." *PLOS Genetics* 9 (7): e1003608. doi:10.1371/journal.pgen.1003608.

Chander, S., Y. Q. Guo, X. H. Yang, J. Zhang, X. Q. Lu, et al., 2008. "Using Molecular Markers to Identify Two Major Loci Controlling Carotenoid Contents in Maize Grain." *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 116 (2): 223–33. doi:10.1007/s00122-007-0661-7.

Chen, W., Y. Gao, W. Xie, L. Gong, K. Lu, et al., 2014. "Genome-Wide Association Analyses Provide Genetic and Biochemical Insights into Natural Variation in Rice Metabolism." *Nature Genetics* 46. doi:10.1038/ng.3007.

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 (2): 347–65. doi:10.1534/genetics.112.147983.

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 (Oxford, England)* 21 (16): 3439–40. doi:10.1093/bioinformatics/bti525.

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.” *Nature Protocols* 4 (8): 1184–91. doi:10.1038/nprot.2009.97.

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

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 (4): 1871–83. doi:10.1534/genetics.116.187161.

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 Physiology* 142 (3): 839–54. doi:10.1104/pp.106.086694.

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.” *Genetics, Selection, Evolution: GSE* 49 (1): 44. doi:10.1186/s12711-017-0319-0.

Fernie, A. R., and H. J. Klee. 2011. “The Use of Natural Genetic Diversity in the Understanding of Metabolic Organization and Regulation.” *Frontiers in Plant Science* 2: 59. doi:10.3389/fpls.2011.00059.

Fernie, A. R., and T. Tohge. 2017. “The Genetics of Plant Metabolism.” *Annual Review of Genetics* 51 (1): null. doi:10.1146/annurev-genet-120116-024640.

Fisher, R. A. (1918). The correlation among relatives on the supposition of mendelian Inheritance. *Trans. R. Soc. Edinb.* 52, 399–433

Foerster, J. M., T. Beissinger, N.de Leon, and S.Kaeppler. 2015. “Large Effect QTL Explain Natural Phenotypic Variation for the Developmental Timing of Vegetative Phase Change in Maize (Zea Mays L.).” *Theoretical and Applied Genetics* 128 (3): 529–38. doi:10.1007/s00122-014-2451-3.

Galili, G., and R. Amir. 2013. “Fortifying Plants with the Essential Amino Acids Lysine and Methionine to Improve Nutritional Quality.” *Plant Biotechnology Journal* 11 (2): 211–22. doi:10.1111/pbi.12025.

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.” *Frontiers in Plant Science* 5 (September). doi:10.3389/fpls.2014.00447.

Gutiérrez-Rojas, A., J. Betrán, M. P. Scott, H. Atta, and M. Menz. 2010. “Quantitative Trait Loci for Endosperm Modification and Amino Acid Contents in Quality Protein Maize.” *Crop Science* 50 (3): 870–79. doi:10.2135/cropsci2008.10.0634.

Heffner, E. L., M. E. Sorrells, and J. Jannink. 2009. “Genomic Selection for Crop Improvement.” *Crop Science* 49 (1): 1–12. doi:10.2135/cropsci2008.08.0512.

Hirschhorn, J. N., and M. J. Daly. 2005. “Genome-Wide Association Studies for Common Diseases and Complex Traits.” *Nature Reviews Genetics* 6 (2): 95–108. doi:10.1038/nrg1521.

Knoch, D., D. Riewe, R. C. Meyer, A. Boudichevskaia, R. Schmidt, and T. Altmann. 2017. “Genetic Dissection of Metabolite Variation in Arabidopsis Seeds: Evidence for mQTL Hotspots and a Master Regulatory Locus of Seed Metabolism.” *Journal of Experimental Botany* 68 (7): 1655–67. doi:10.1093/jxb/erx049.

Kochevenko, A., and A. R. Fernie. 2011. “The Genetic Architecture of Branched-Chain Amino Acid Accumulation in Tomato Fruits.” *Journal of Experimental Botany* 62 (11): 3895–3906. doi:10.1093/jxb/err091.

Korte, A., and A. Farlow. 2013. “The Advantages and Limitations of Trait Analysis with GWAS: A Review.” *Plant Methods* 9 (July): 29. doi:10.1186/1746-4811-9-29.

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li, et al., 2012. "GAPIT: Genome Association and Prediction Integrated Tool." *Bioinformatics (Oxford, England)* 28 (18): 2397–99. doi:10.1093/bioinformatics/bts444.

Lisec, J., R. C. Meyer, M. Steinfath, H. Redestig, M. Becher, et al., 2008. "Identification of Metabolic and Biomass QTL in Arabidopsis Thaliana in a Parallel Analysis of RIL and IL Populations." *The Plant Journal: For Cell and Molecular Biology* 53 (6): 960–72. doi:10.1111/j.1365-313X.2007.03383.x.

Maloney, G. S., A.Kochevenko, D. M. Tieman, T. Tohge, U. Krieger, et al., 2010. "Characterization of the Branched-Chain Amino Acid Aminotransferase Enzyme Family in Tomato." *Plant Physiology* 153 (3): 925–36. doi:10.1104/pp.110.154922.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." *Genetics* 157 (4): 1819–29.

Mueller, L. A., P. Zhang, and S. Y. Rhee. 2003. "AraCyc: A Biochemical Pathway Database for Arabidopsis." *Plant Physiology* 132 (2): 453–60. doi:10.1104/pp.102.017236.

Nielsen, N. H., A. Jahoor, J. D. Jensen, J. Orabi, F. Cericola, et al., 2016. "Genomic Prediction of Seed Quality Traits Using Advanced Barley Breeding Lines." *PLOS ONE* 11 (10): e0164494. doi:10.1371/journal.pone.0164494.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, et al., 2005. "The Pattern of Polymorphism in Arabidopsis Thaliana." *PLOS Biology* 3 (7): e196. doi:10.1371/journal.pbio.0030196.

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 (4): 1699–1716. doi:10.1534/genetics.114.169979.

R Core Team, 2017 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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." *Proceedings of the National Academy of Sciences* 109 (23): 8872–77. doi:10.1073/pnas.1120813109.

Rowe, H. C., B. G. Hansen, B. A. Halkier, and D. J. Kliebenstein. 2008. "Biochemical Networks and Epistasis Shape the Arabidopsis Thaliana Metabolome." *The Plant Cell Online* 20 (5): 1199–1216. doi:10.1105/tpc.108.058131.

Sarup, P., J. Jensen, T. Otersen, 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

Smedley, D., S. Haider, S. Durinck, L. Pandini, P. Provero, et al., 2015. "The BioMart Community Portal: An Innovative Alternative to Large, Centralized Data Repositories." *Nucleic Acids Research* 43 (W1): W589-598. doi:10.1093/nar/gkv350.

Sørensen, P., 2017 qgg; Statistical tools for quantitative genetic analyses. R package version 1.0.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, et al. 2008. "The Arabidopsis Information Resource (TAIR): Gene Structure and Function Annotation." *Nucleic Acids Research* 36 (Database issue): D1009-1014. doi:10.1093/nar/gkm965.

Ufaz, S., and G. Galili. 2008. "Improving the Content of Essential Amino Acids in Crop Plants: Goals and Opportunities." *Plant Physiology* 147 (3): 954–61. doi:10.1104/pp.108.118091.

Vallabhaneni, R., and E. T. Wurtzel. 2009. "Timing and Biosynthetic Potential for Carotenoid Accumulation in Genetically Diverse Germplasm of Maize." *Plant Physiology* 150 (2): 562–72. doi:10.1104/pp.109.137042.

VanRaden, P. M. 2008. "Efficient Methods to Compute Genomic Predictions." *Journal of Dairy Science* 91 (11): 4414–23. doi:10.3168/jds.2007-0980.

Weigel, D., and M.Nordborg. 2005. “Natural Variation in Arabidopsis. How Do We Find the Causal Genes?” *Plant Physiology* 138 (2): 567–68. doi:10.1104/pp.104.900157.

Weller, J. I., E. Ezra, and M. Ron. 2017. “Invited Review: A Perspective on the Future of Genomic Selection in Dairy Cattle.” *Journal of Dairy Science*, August. doi:10.3168/jds.2017-12879.

Wen, W., Li, K., Alseekh, S., Omranian, N., Zhao, L., Zhou, Y., et al., (2015). Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. *The Plant Cell*, 27(7), 1839-1856.

Wentzell, A. M., He. C. Rowe, B. G. Hansen, C. Ticconi, et al., 2007. “Linking Metabolic QTLs with Network and Cis-eQTLs Controlling Biosynthetic Pathways.” *PLOS Genetics* 3 (9): e162. doi:10.1371/journal.pgen.0030162.

Wickham, H.. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wolc, Anna, A Kranis, Jesus Arango, Petek Settar, Janet Fulton, Neil O’Sullivan, A Avendano, et al. 2016. *Implementation of Genomic Selection in the Poultry Industry*. Vol. 6.

Wong, H., H. Chan, G. M. Coruzzi, and H. Lam. 2004. “Correlation of ASN2 Gene Expression with Ammonium Metabolism in Arabidopsis.” *Plant Physiology* 134 (1): 332–38. doi:10.1104/pp.103.033126.

Wu, S., S. Alseekh, Á. Cuadros-Inostroza, C. M. Fusari, M. Mutwil,et al., 2016. “Combined Use of Genome-Wide Association Data and Correlation Networks Unravels Key Regulators of Primary Metabolism in Arabidopsis Thaliana.” *PLOS Genetics* 12 (10): e1006363. doi:10.1371/journal.pgen.1006363.

Wu, Y. and J. Messing. 2014. “Proteome Balancing of the Maize Seed for Higher Nutritional Value.” *Frontiers in Plant Science* 5: 240. doi:10.3389/fpls.2014.00240.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, et al., 2010. “Common SNPs Explain a Large Proportion of the Heritability for Human Height.” *Nature Genetics* 42 (7): 565–69. doi:10.1038/ng.608.

Zhang, P., H. Foerster, C. P. Tissier, L. Mueller, S. Paley, et al., 2005. “MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research.” *Plant Physiology* 138 (1): 27–37. doi:10.1104/pp.105.060376.

Zhang, Z., R. J. Todhunter, E. S. Buckler, L. D. Van Vleck, 2007. “Technical Note: Use of Marker-Based Relationships with Multiple-Trait Derivative-Free Restricted Maximal Likelihood.” *Journal of Animal Science* 85 (4): 881–85. doi:10.2527/jas.2006-656.

Zhu, X., and G. Galili. 2003. “Increased Lysine Synthesis Coupled with a Knockout of Its Catabolism Synergistically Boosts Lysine Content and Also Transregulates the Metabolism of Other Amino Acids in Arabidopsis Seeds.” *The Plant Cell* 15 (4): 845–53. doi:10.1105/tpc.009647.

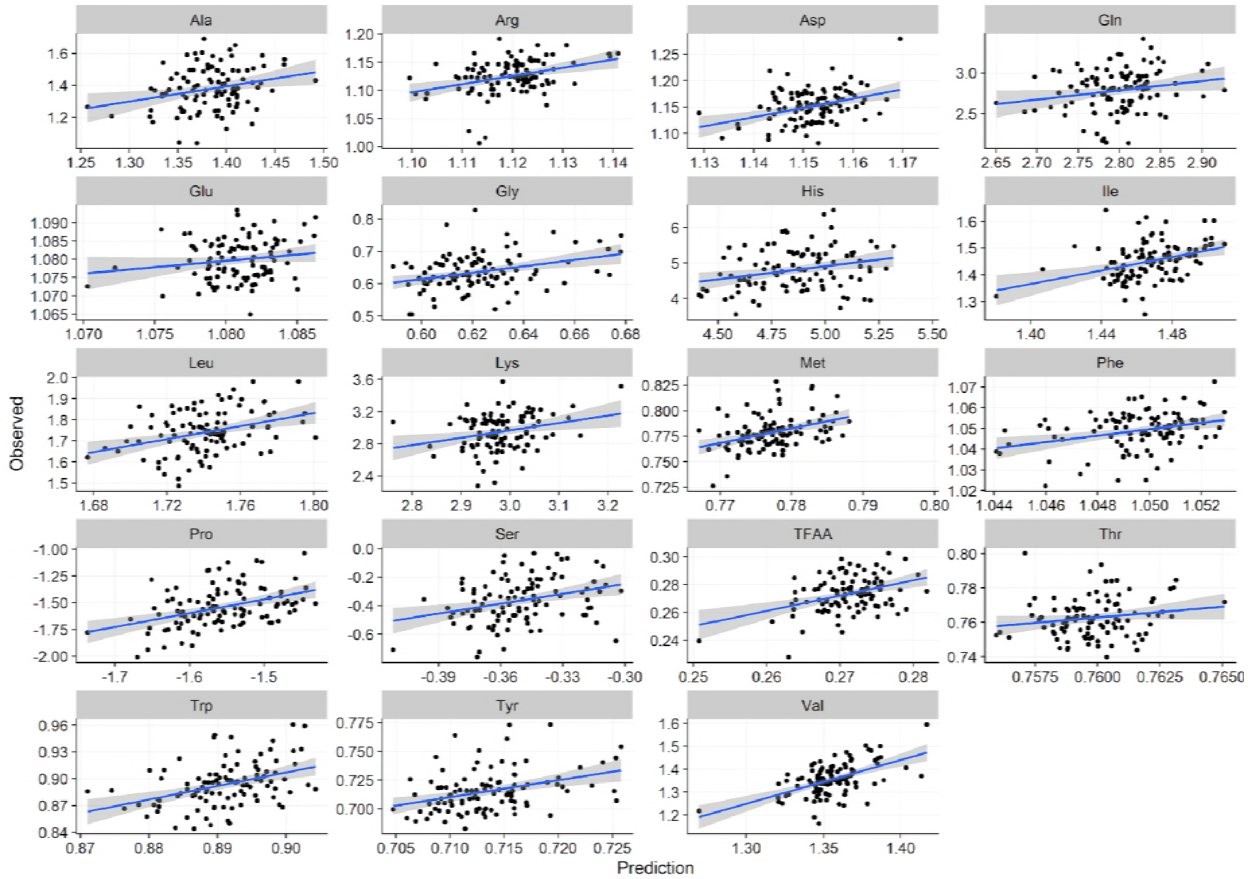


Figure S1: Correlation between predicted and observed values using the GS subset for each of the 19 amino acid traits studied. Although 1,000 replicated genomic prediction runs were conducted, this figure was generated using data from a single replicate.

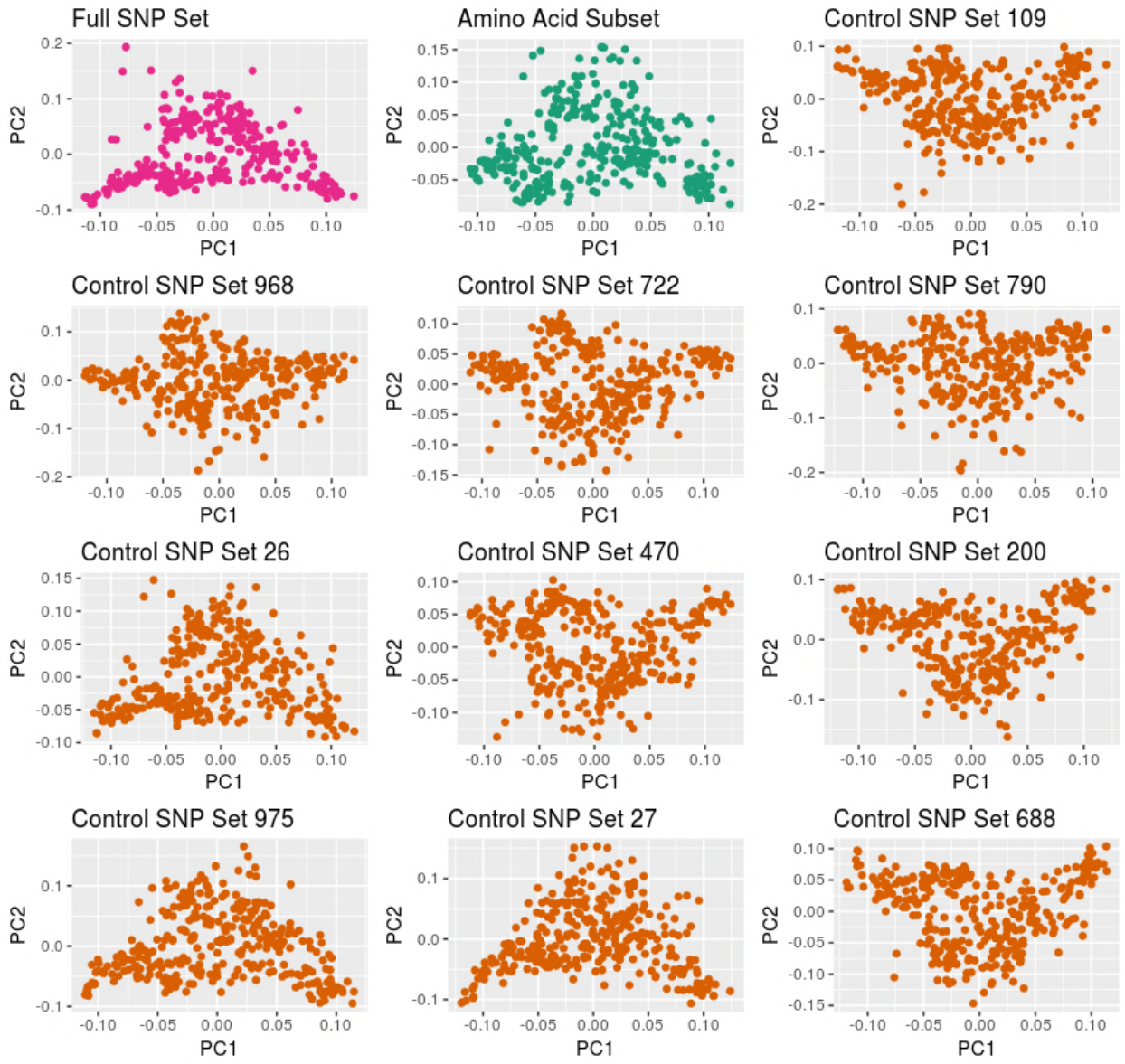


Figure S2: Principal component analysis (PCA) demonstrates similar patterns of population structure for the full genomic SNP set, amino acid SNP subset, and ten randomly selected control SNP sets. Note that although the direction of PC2 fluctuates across the control SNP sets, the pattern and magnitude of variation is consistent.

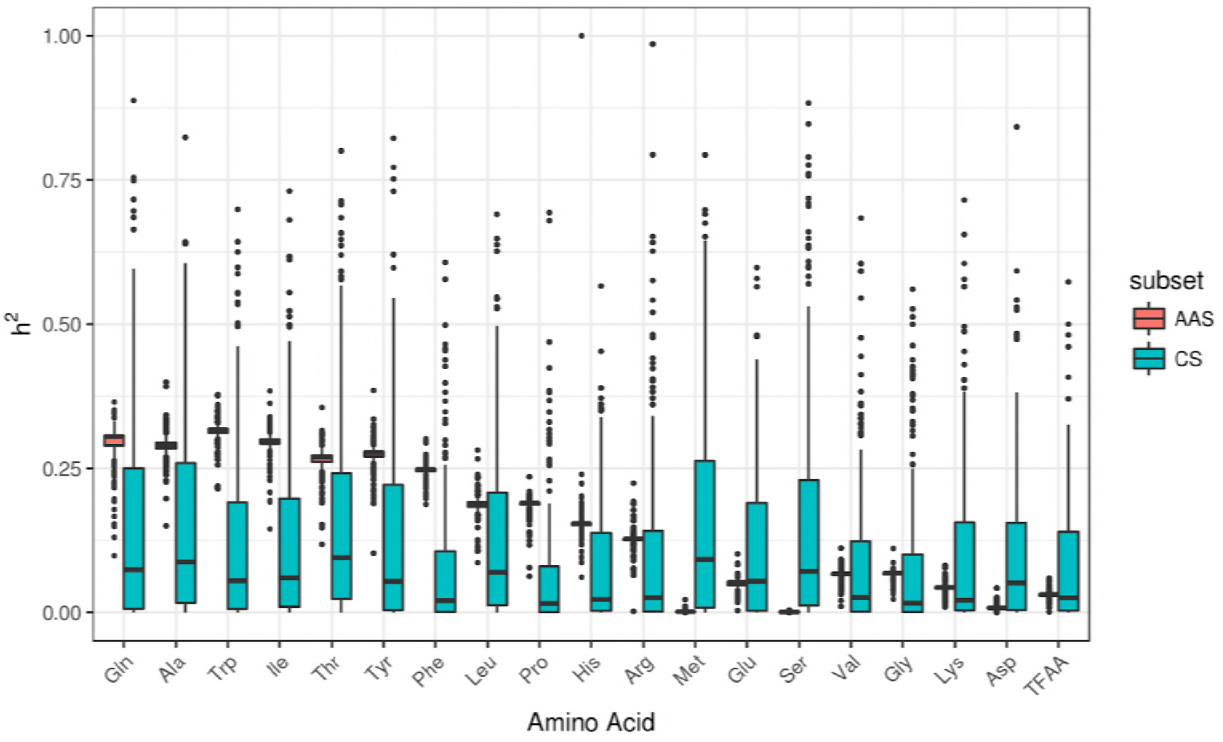


Figure S3: Comparison of the proportion of genomic variance explained (h^2) between the AAS subset and the CS subset.

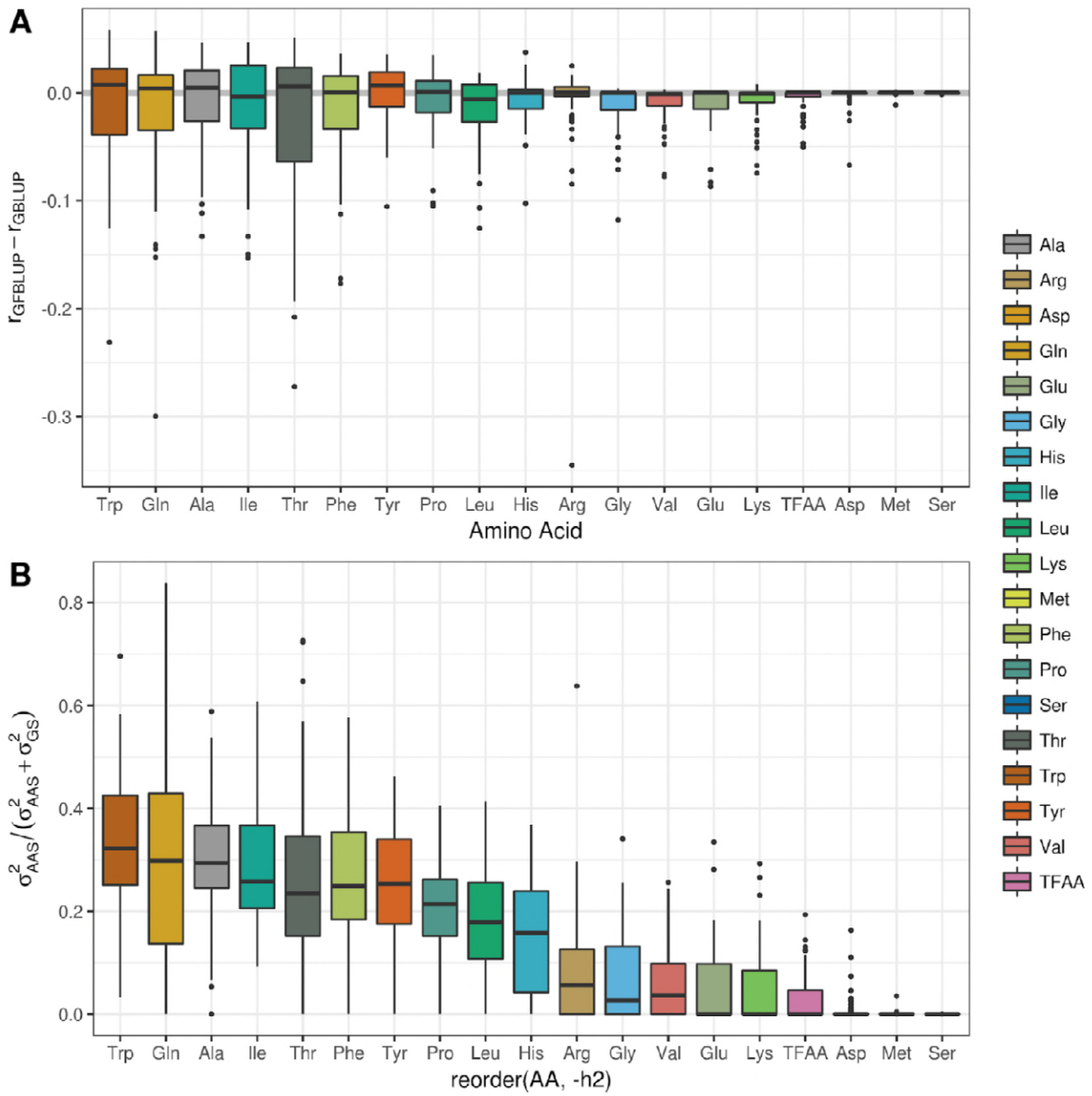


Figure S4: (**A**) Difference in prediction accuracy (*r*) between the GFBLUP model with the AAS subset weighted and the GBLUP model for 19 amino acid traits. (**B**) Proportion of genomic variance explained by the AAS subset in the GFBLUP model with the AAS subset weighted.

894 **Table S1:** Binomial test results comparing the relative proportion of additive variance explained using GFBLUP for
895 the AAS subset and the 1,000 CS subsets. The number of successes and p-values are presented for varying
896 magnitudes of difference (5%, 10%, 15%, 20%, 30%, 40%, 50%).