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

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

Sarah D. Turner-Hissong[1], Kevin A. Bird[1], Alexander E. Lipka[2], Elizabeth G. King[1], Timothy M. Beissinger[3,4], Ruthie Angelovici[1*]

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

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

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

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

*** Corresponding author:**

E-mail: angelovicir@missouri.edu (RA)

Abstract

Amino acids are a critical component of plant growth and development, as well as human and animal nutrition. A better understanding of the genetic architecture of amino acid traits, especially in seeds, will enable researchers to use this information for plant breeding and biological discovery. Despite a collection of successfully mapped genes, a fundamental understanding of the types of genes and biological processes underlying amino acid related traits in seeds remains unresolved. In this study, we used genomic prediction with SNPs partitioned by metabolic pathways to quantify the contribution of primary, specialized, and protein metabolic processes to free amino acid (FAA) homeostasis in dry *Arabidopsis* seeds. First, we demonstrate that standard genomic prediction is effective for FAA traits. Next, we show that genomic partitioning by metabolic pathway annotations explains significant genetic variation and improves prediction accuracy for many FAA traits, including many trait-pathway associations that have not been previously reported. Surprisingly, SNPs related to amino acid and primary metabolism had limited effects on prediction accuracy for most FAA traits, with the largest effects observed for branched chain amino acids (BCAAs). In contrast, SNPs related to secondary and protein metabolism had a more extensive effect on prediction accuracy. The use of a genomic partitioning approach also revealed specific patterns across biochemical families, in which protein related annotations were the only category influencing serine-derived FAAs and primary and specialized metabolic pathways were the only categories contributing to aromatic FAAs. Based on these findings, we used pathway-guided association analysis to identify novel SNP associations for traits related to methionine, threonine, histidine, arginine, glycine, phenylalanine, and BCAAs. Taken together, these findings provide evidence that genomic partitioning is a viable strategy to uncover the complexity of FAA homeostasis and to identify candidate genes for future functional validation.

Author summary

Plant growth, development, and nutritional quality depends upon the regulation of 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. As an alternative approach, we implemented multikernel genomic prediction to test whether or not genomic regions related to specific metabolic pathways contribute to free amino acid (FAA) variation in seeds of the model plant *Arabidopsis thaliana*. Importantly, this method successfully identifies pathways containing known variants for FAA traits, in addition to identifying new pathway associations. For several traits, the incorporation of prior biological knowledge provided substantial improvements in prediction accuracy. We present this approach as 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. 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 [1–5]. The homeostasis for absolute levels and relative composition of the free amino acid (FAA) pool is complex, depending on various factors such as allosteric regulation, feedback loops of key synthetic metabolic enzymes in amino acid metabolic pathways, and the rate of amino acid degradation [6–10]. In addition, FAA homeostasis can be influenced by protein metabolism. For example, the consistently observed significant increase in FAAs under many abiotic stresses is suggested to result from autophagy and protein turnover [7,8,11–13]. Specific FAAs, such as proline, may serve as either an osmoprotectant under stress or an energy source during development, with their elevation resulting mostly from active synthesis rather than protein degradation [14,15]. Studies have also demonstrated that the composition of the FAA pool is affected when either primary or specialized metabolism is altered. For example, perturbation of the glucosinolate pathway in *Arabidopsis* plants caused a significant elevation of multiple FAAs [16], while alteration of the interconversion of pyruvate and malate in tomato fruits caused reduction in aspartate family related FAAs [17]. Therefore, FAA homeostasis is most likely determined by orchestration of multiple processes, but it remains challenging to pinpoint the main processes that are associated with homeostasis at various developmental stages.

Dry seeds, despite their metabolically dormant state, maintain a tightly regulated FAA pool, which contributes to proper desiccation, longevity, germination, and seed vigor [5,18]. This pool comprises 1-10% of total seed amino acid content in maize [6,19] and ~7% in *Arabidopsis thaliana* [6,20]. Fait et al. [21] showed that in *Arabidopsis*, several FAAs are actively synthesized during late seed desiccation to provide the necessary precursors for early germination. Other studies further demonstrated that the natural variation of histidine and branched-chain amino acid (BCAA) levels in dry *Arabidopsis* seeds are associated with amino acid catabolism or transport [22,23]. Protein metabolism has also been implicated in determining the homeostasis of FAAs in dry seeds. For instance, the *opaque2* null mutant, which results in reduction of the most abundant seed storage proteins in maize, had significant elevation of many FAAs despite an unchanged composition of protein-bound amino acids [24,25]. The goal of engineering mutants like *opaque2* is to increase accumulation of essential amino acids that are deficient in crop seeds, such as lysine. However, these mutations have negative effects on key agronomic traits such as disease resistance, germination rate, and seedling vigor [26], suggesting a tight integration of AA metabolism with both primary and specialized metabolism.

Like many other primary metabolites in dry seeds, FAAs are complex traits with extensive variability and high heritability across natural populations. Several genome-wide association studies (GWAS) have been performed on FAAs, which resulted in the successful identification of candidate loci for amino acid traits, both independently [27] and in conjunction with QTL studies [22,23]. However, the number and effect size of loci detected so far explained only a fraction of the observed phenotypic variation for these traits, with some traits proving harder to dissect than others. For example, [22,23] found the strongest associations for traits related to histidine and BCAAs, but weak signals for most other FAA traits. In addition, GWAS has limited power to reliably identify variants that are rare and/or of small effect [28]. In an attempt to uncover more of the genetic basis for FAA composition, subsequent investigations used integrated analyses that combined GWAS, linkage mapping, and metabolic correlation networks to identify new candidate loci related to FAA levels in both seeds and leaves of *Arabidopsis* [23,29]. Several metabolic studies have also integrated prior information on biological relationships to specify metabolic ratios, which can uncover novel or more significant associations compared to absolute levels of metabolites [22,23,30–36].

The consistent finding that amino acid traits frequently have several associated loci, coupled with the difficulty of GWAS to explain a large proportion of the genetic variation for these traits, suggests that amino acid traits may have a highly polygenic architecture with many loci of small effect. 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 highly complex [37–39]. Genomic prediction allows researchers to predict an individual's breeding value, or the additive component of their genetic variation, based only on genotypic data [37,40]. 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 GWAS [40]. 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.

Genomic best linear unbiased prediction (GBLUP) [37], which assumes that all SNPs share a common effect size distribution, is one of the most widely used methods for prediction of complex traits. Extensions of the GBLUP model, such as MultiBLUP [41], genomic feature BLUP (GFBLUP) [42–45], and the Bayesian method BayesRC [46] 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. Models that incorporate genomic partitioning have allowed researchers to determine the influence of genomic features (e.g. chromosome segments, exons) and/or biological pathways on variance explained for complex traits in humans [47,48], cattle [42,45,49], Duroc pigs [44], fruit flies [42,50], and maize [51]. Notably, when genomic partitions are enriched for previously identified candidate genes, these models demonstrably improve prediction accuracy [42,44–46,49]. Evidence also suggests that, although many genetic markers may contribute to the overall genetic variation, many of these markers are preferentially located in genes that are connected to a biological pathway(s) [52].

In this study, we used the framework of genomic partitioning, coupled with prior knowledge and annotations of metabolic pathways, to evaluate which biological processes and regions of the genome are disproportionately influencing FAA content and composition in seeds of a diverse *Arabidopsis* panel. The primary goal was to identify the relative importance of previously implicated metabolic pathways (i.e. amino acid, primary, specialized, and protein metabolism) in relation to FAA content and composition in dry seeds. To this end, we demonstrate that specific pathways explain more variation than expected by chance for several FAA traits and improve prediction accuracy when using genomic partitioning. Findings suggest that specialized and protein metabolism are associated with many FAAs, while amino acid metabolism is associated with a very limited number. We then used these results to apply pathway-level association mapping (e.g. [34]), which uncovered additional novel loci associated with FAA levels in *Arabidopsis* seeds. By identifying genes in metabolic pathways that explain significant genetic variation and improve prediction accuracy, we can form a more comprehensive understanding of which pathways underlie FAA homeostasis in seeds. When compared to previous GWAS results, this approach adds additional information on the orchestrated regulation of FAAs in seeds, which will help expand our understanding of complex metabolic networks in plants.

Results

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

Using the GBLUP model, we observed low to moderate prediction accuracy for the amino acid traits measured (see S1 Table for trait descriptions). Of the 65 traits measured, 26 had prediction accuracy > 0.3 (Fig 1, Table 1). In general, prediction was effective for a greater number of absolute level FAA traits (68% > 0.3) compared to relative levels (29%) and family-derived ratios (17%). The aromatic family composite trait (ShikFam, combined absolute levels of phenylalanine, tryptophan, and tyrosine) had the highest prediction accuracy (r = 0.43), while the absolute level of threonine had the lowest prediction accuracy (r = 0.11) (Table 1).

Table 1. Genomic prediction results for amino acid traits using a GBLUP model.

			accuracy		reliability		bias		
Trait type	**Metabolic family**	**Trait**	**mean**	**SE**	**mean**	**SE**	**intercept**	**slope**	**MSE**
absolute	aspartate	asp	0.286	0.024	0.115	0.014	4.64E-04	0.936	3.57E-05
		met	0.321	0.024	0.214	0.025	5.75E-06	1.051	1.14E-05
		thr	0.105	0.023	0.079	0.014	-1.23E-05	0.756	5.56E-06
		AspFam	0.235	0.023	0.123	0.017	2.64E-04	1.042	2.28E-05
	BCAA_pyruvate	ala	0.299	0.024	0.177	0.023	5.33E-04	1.119	6.69E-04
		ile	0.326	0.022	0.182	0.022	2.38E-04	1.046	1.19E-04
		leu	0.342	0.022	0.185	0.019	7.55E-04	1.035	2.13E-04
		lys	0.246	0.026	0.128	0.018	-1.74E-03	0.921	1.58E-03
		val	0.409	0.020	0.187	0.017	5.06E-04	1.027	1.18E-04
		BCAA	0.392	0.020	0.199	0.019	5.40E-04	1.042	1.65E-04
		PyrFam	0.361	0.023	0.208	0.022	4.02E-04	1.108	1.87E-04
	glutamate	arg	0.228	0.023	0.145	0.020	-1.75E-04	0.923	2.85E-05
		gln	0.193	0.024	0.149	0.024	5.11E-04	1.032	2.47E-03
		glu	0.339	0.023	0.182	0.019	2.95E-05	0.990	1.10E-06
		his	0.356	0.021	0.148	0.014	6.14E-03	0.880	6.57E-03
		pro	0.372	0.021	0.194	0.019	-4.68E-04	1.010	1.02E-03
		GluFam	0.322	0.024	0.133	0.014	-5.26E-06	0.916	1.06E-05
	serine	gly	0.363	0.023	0.252	0.027	-6.57E-05	1.072	9.49E-05
		ser	0.225	0.022	0.147	0.019	6.55E-04	1.078	7.65E-04
		SerFam	0.323	0.023	0.192	0.020	9.76E-05	1.030	6.76E-04
	aromatic	phe	0.307	0.022	0.172	0.021	2.73E-05	1.084	1.84E-06
		trp	0.348	0.019	0.223	0.022	-1.73E-04	1.019	1.23E-05
		tyr	0.344	0.026	0.202	0.024	-2.49E-04	1.046	8.96E-06
		ShikFam	0.431	0.018	0.245	0.018	6.43E-05	1.023	1.09E-06
		Total	0.395	0.024	0.183	0.017	5.43E-05	1.015	5.11E-06
relative	aspartate	asp_t	0.392	0.023	0.178	0.017	1.51E-02	0.933	5.61E-02
		met_t	0.328	0.022	0.181	0.018	-8.14E-05	1.021	2.99E-05
	BCAA_pyruvate	ala_t	0.205	0.025	0.184	0.030	2.98E-05	1.239	1.97E-05
		ile_t	0.233	0.022	0.137	0.021	-1.33E-04	1.085	3.72E-05
		leu_t	0.263	0.023	0.137	0.018	9.09E-05	1.093	2.64E-05
		lys_t	0.224	0.023	0.159	0.020	-2.12E-04	1.112	3.35E-05
		val_t	0.306	0.024	0.242	0.028	-1.96E-04	1.106	9.94E-06
	glutamate	arg_t	0.193	0.025	0.154	0.023	-1.17E-04	1.056	2.07E-05
		gln_t	0.108	0.023	0.236	0.039	1.14E-04	1.322	1.77E-04
		glu_t	0.260	0.021	0.179	0.021	2.00E-02	1.008	2.78E-01
		his_t	0.262	0.026	0.160	0.019	1.22E-03	1.076	3.24E-04
		pro_t	0.342	0.020	0.172	0.016	4.54E-05	1.022	4.72E-05
	serine	gly_t	0.276	0.025	0.290	0.038	-8.77E-04	1.127	1.37E-03
		ser_t	0.156	0.023	0.143	0.024	4.44E-04	1.452	9.95E-05
	aromatic	phe_t	0.341	0.017	0.174	0.016	-3.19E-04	1.047	2.49E-04
		trp_t	0.221	0.023	0.145	0.020	-2.27E-04	0.940	2.62E-05
		tyr_t	0.151	0.027	0.169	0.023	-2.51E-04	1.112	3.68E-05

			accuracy		reliability		bias		
Trait type	**Metabolic family**	**Trait**	**mean**	**SE**	**mean**	**SE**	**intercept**	**slope**	**MSE**
family	aspartate	asp_AspFam	0.159	0.025	0.159	0.026	-2.32E-04	1.309	1.27E-04
		ile_AspFam	0.339	0.022	0.218	0.022	-6.30E-05	1.197	2.05E-05
		lys_AspFam	0.179	0.024	0.125	0.019	-9.62E-05	0.933	4.47E-06
		met_AspFam	0.296	0.023	0.219	0.025	7.92E-05	1.060	1.62E-05
		thr_AspFam	0.211	0.023	0.132	0.017	-4.73E-05	1.049	1.11E-06
		AspFam_Asp	0.235	0.023	0.123	0.017	2.64E-04	1.042	2.28E-05
	BCAA_pyruvate	ala_PyrFam	0.272	0.019	0.091	0.010	1.37E-04	0.905	5.62E-05
		ile_BCAA	0.169	0.020	0.065	0.010	7.60E-05	0.914	9.26E-06
		leu_BCAA	0.196	0.020	0.107	0.017	8.24E-05	1.076	4.84E-06
		leu_PyrFam	0.274	0.020	0.106	0.013	-3.02E-05	0.975	5.50E-06
		val_BCAA	0.172	0.019	0.066	0.010	-8.61E-05	0.848	1.85E-05
		val_PyrFam	0.227	0.019	0.070	0.009	7.23E-05	0.858	4.20E-06
	glutamate	arg_GluFam	0.134	0.027	0.174	0.032	-1.15E-04	1.243	3.93E-06
		gln_GluFam	0.135	0.022	0.197	0.030	3.23E-03	1.076	2.38E-02
		glu_GluFam	0.229	0.028	0.184	0.025	1.47E-04	0.992	3.74E-05
		GluFam_glu	0.270	0.027	0.144	0.020	8.87E-04	0.881	1.51E-04
		his_GluFam	0.195	0.023	0.127	0.019	5.48E-02	1.012	4.98E+00
		pro_GluFam	0.349	0.025	0.209	0.019	1.51E-04	1.004	1.68E-05
	serine	gly_SerFam	0.314	0.024	0.283	0.036	2.54E-05	1.172	2.47E-05
		ser_SerFam	0.313	0.024	0.286	0.037	-2.44E-05	1.179	1.32E-05
	aromatic	phe_ShikFam	0.223	0.029	0.188	0.027	2.19E-04	1.097	1.38E-05
		trp_ShikFam	0.217	0.024	0.162	0.023	-8.03E-05	1.028	5.66E-06
		tyr_ShikFam	0.168	0.024	0.114	0.019	-5.18E-05	0.945	1.52E-06

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

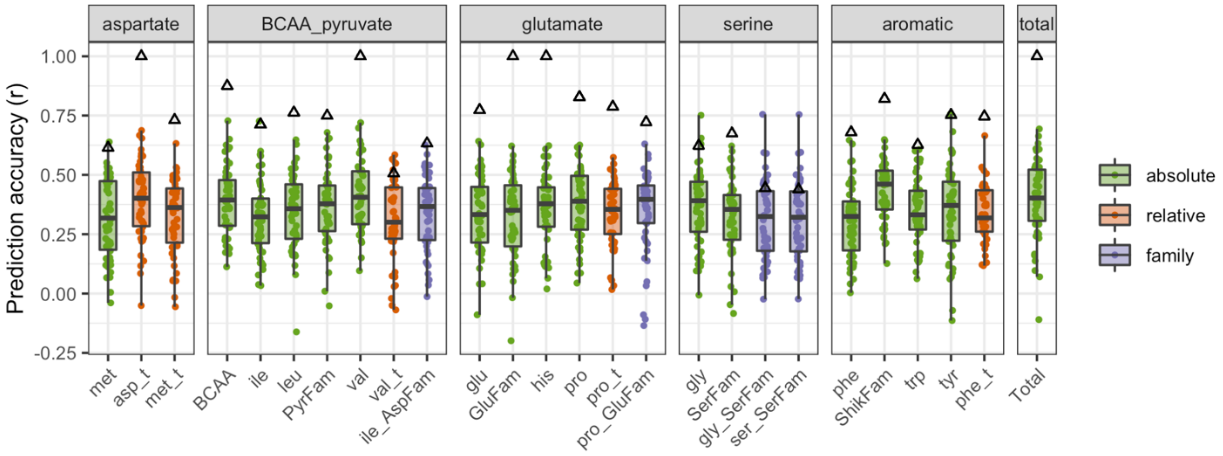


Fig 1. Genomic prediction performed well for a higher proportion of absolute traits compared to relative and family-based ratio traits.

Boxplots show free amino acid traits with prediction accuracy (r) > 0.3 based on genomic best linear unbiased prediction (GBLUP). For absolute traits, 68% had r > 0.3 compared to relative traits (29%) and family-based ratio traits (17%). Black triangles indicate the genomic heritability for each trait. Each point represents an individual cross-validation.

Annotations of biological pathways explain variation and improve prediction accuracy for free amino acid traits in seeds

The pathway annotations listed in Table 2 were used to subset SNPs and spanned amino acid, primary, specialized, and protein metabolism. When partitioning these pathways in the MultiBLUP model, 44 trait-pathway combinations were flagged as putatively related based on comparison to a null distribution (Fig 2, Table 3). Results for the null distribution of each trait, including how many random gene groups passed filtering criteria, are reported in S2 Table and S2 Fig. The observation that specific pathways improve model fit based on likelihood ratio (LR) and explain a significant proportion of genomic heritability suggests that these pathway annotations may have biological relevance for FAA traits.

Table 2. Summary of selected biological pathways.

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

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

A few patterns were noticeable when looking at absolute levels of FAAs (Fig 2). Traits in the aspartate and glutamate families showed a high proportion of genomic heritability explained for pathways related to specialized and protein metabolism. One example is the pathway for sulfur containing compounds and absolute levels of methionine, which is a precursor for aliphatic glucosinolates. The only relationship observed for absolute levels in the pyruvate/BCAA group was with amino acid synthesis. Similarly, three traits in the serine family had a significant proportion of genomic heritability explained for pathways related to protein metabolism, while three traits in the aromatic family stood out for specialized metabolism.

When looking at relative ratios of FAA traits (Fig 2), a high proportion of genomic heritability was explained for four traits in the glutamate family and pathways across amino acid metabolism, primary metabolism, specialized metabolism, and protein metabolism. A similar relationship was observed for traits in the pyruvate/BCAA family, with the exception of specialized metabolism. Traits in the serine and aromatic families again showed significant values for pathways related to protein and specialized metabolism, respectively. These relationships were similar for family-based ratios of FAA traits (Fig 2), with the exception that traits in the pyruvate/BCAA family had associations with specialized metabolism and not with primary metabolism.

For nine trait-pathway combinations, the prediction accuracy for the MultiBLUP model was over 5% higher than for the GBLUP model with limited effects on bias and MSE (Table 3, bold). This substantial increase in prediction accuracy was observed for BCAA related traits when the model included the amino acid degradation (relative levels of isoleucine, Ile_t, and the family-based ratio of valine, Val_BCAA) or isoprenoid pathway information (Val_BCAA). A similar increase in prediction accuracy was observed for relative and family-based ratios of glutamine (Gln_t and Gln_GluFam, respectively) when partitioning SNPs related to sulfur containing specialized metabolites, and for the family-based ratio of tyrosine (Tyr_ShikFam) for SNPs related to phenylpropanoids.

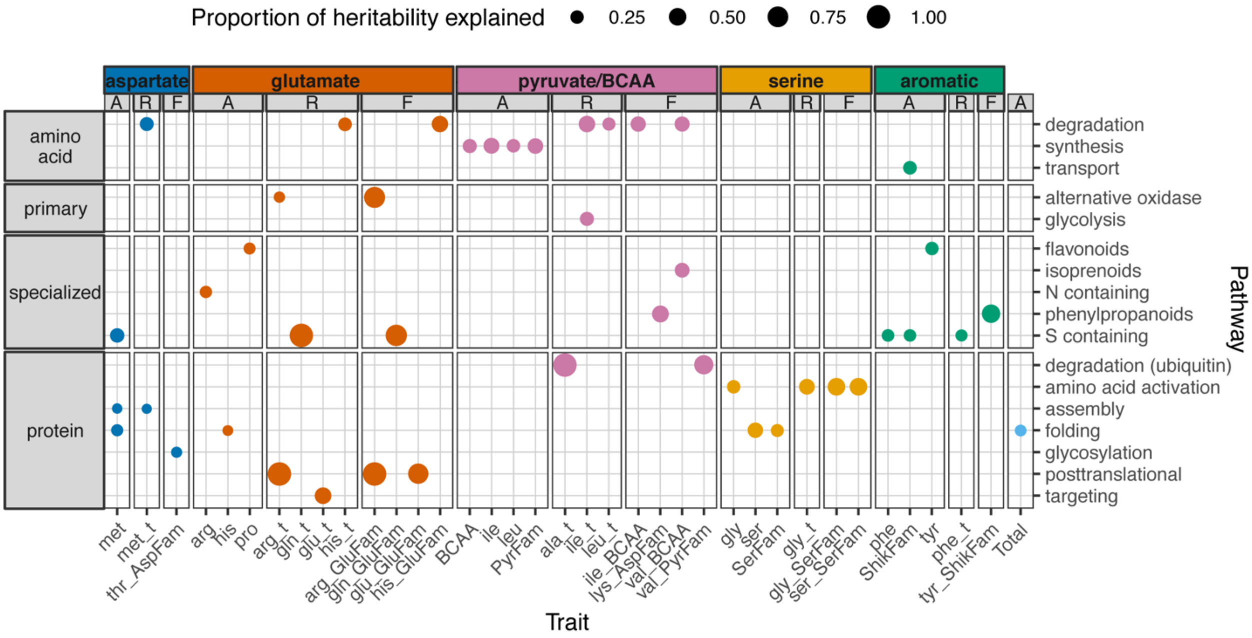


Fig 2. Biological pathways explain significant variation and improve prediction accuracy for free amino acid traits.

Dots indicate pathways that improved prediction accuracy compared to GBLUP and exceeded the 95% null thresholds for proportion of heritability explained and likelihood ratio (LR). The diameter of each dot is proportional to the amount of genomic variance explained by pathway SNPs in the MultiBLUP model. Traits are included on the x-axis and grouped by metabolic family (aspartate, glutamate, pyruvate/BCAA, serine, aromatic) and type of measurement (A = absolute, R = relative, F = family-based ratio). Pathways are included on the y-axis and separated into amino acid, primary, specialized, and protein metabolism categories.

Table 3. Free amino acid traits and pathway combinations for which MultiBLUP increases accuracy compared to GBLUP.

			Proportion h^2 explained		Likelihood ratio				Δbias		
	Pathway	Trait	95 percentile[a]	MultiBLUP	95 percentile[a]	MultiBLUP	Δr[b]	$\Delta r^2/h^2$[c]	intercept	slope	ΔMSE
amino acid	degradation	his_t	0.19	0.26	4.43	5.65	0.03	0.01	-3.2E-04	-4.5E-02	2.8E-05
	degradation	**ile_t**	**0.28**	**0.43**	**4.30**	**9.12**	**0.07**	**0.07**	**2.2E-04**	**-5.4E-02**	**4.8E-07**
	degradation	leu_t	0.22	0.24	4.05	5.17	0.03	0.03	-4.0E-05	-4.3E-02	-2.0E-06
	degradation	met_t	0.15	0.28	3.30	6.63	0.03	0.02	2.8E-05	-9.0E-03	-2.0E-06
	degradation	**his_GluFam**	**0.22**	**0.42**	**4.70**	**8.73**	**0.06**	**0.03**	**5.6E-02**	**8.3E-02**	**8.7E-01**
	degradation	ile_BCAA	0.19	0.36	4.41	5.98	0.04	0.03	-4.8E-05	2.3E-02	3.0E-07
	degradation	**val_BCAA**	**0.19**	**0.34**	**3.56**	**9.99**	**0.07**	**0.05**	**7.6E-05**	**4.2E-02**	**-1.1E-06**
	synthesis	BCAA	0.15	0.29	4.18	11.58	0.03	0.03	3.2E-04	-2.6E-02	-1.2E-05
	synthesis	ile	0.24	0.39	3.54	7.69	0.05	0.04	1.2E-04	-4.3E-02	-9.3E-06
	synthesis	leu	0.19	0.26	3.25	12.56	0.03	0.03	3.1E-04	-2.1E-02	-1.4E-05
	synthesis	PyrFam	0.16	0.37	3.43	5.17	0.03	0.03	-5.0E-05	-8.9E-02	-6.6E-07
	transport	ShikFam	0.10	0.27	3.55	6.43	0.03	0.03	-4.1E-05	-7.2E-03	2.8E-08
primary	alternative oxidase	arg_t	0.03	0.16	3.95	4.35	0.02	0.00	1.5E-05	1.5E-02	1.2E-06
	alternative oxidase	arg_GluFam	0.14	0.77	3.40	9.39	0.05	0.05	-2.6E-05	-6.7E-02	3.2E-03
	glycolysis	ile_t	0.23	0.29	3.78	5.19	0.02	0.02	-5.3E-05	-7.1E-02	2.8E-06
specialized	flavonoids	pro	0.10	0.18	2.98	3.46	0.02	0.02	-3.2E-04	-1.1E-03	-5.8E-05
	flavonoids	tyr	0.16	0.25	3.75	5.75	0.02	0.01	-6.8E-05	-8.5E-03	-1.9E-07
	isoprenoids	**val_BCAA**	**0.23**	**0.33**	**3.56**	**6.26**	**0.09**	**0.05**	**7.5E-05**	**-1.3E-01**	**8.8E-07**
	N containing	arg	0.14	0.20	3.07	3.64	0.01	0.01	-1.5E-04	4.6E-02	1.6E-06
	phenylpropanoids	lys_AspFam	0.23	0.45	3.67	3.95	0.03	0.03	-5.8E-05	4.9E-02	4.3E-08
	phenylpropanoids	**tyr_ShikFam**	**0.18**	**0.56**	**4.61**	**10.86**	**0.11**	**0.07**	**-5.0E-05**	**-5.6E-03**	**-2.7E-07**
	S containing	met	0.14	0.30	3.58	5.46	0.03	0.02	1.3E-05	-3.6E-02	-2.6E-07
	S containing	phe	0.14	0.21	3.68	4.38	0.03	0.03	-1.0E-06	-6.2E-03	7.3E-08
	S containing	ShikFam	0.10	0.21	3.55	8.18	0.02	0.02	-1.3E-05	-1.0E-02	-8.0E-11
	S containing	**gln_t**	**0.58**	**1.00**	**4.56**	**5.59**	**0.06**	**0.13**	**1.5E-04**	**-2.6E-01**	**1.1E-05**
	S containing	phe_t	0.13	0.19	3.78	4.60	0.02	0.02	2.1E-04	-1.9E-02	4.7E-06
	S containing	**gln_GluFam**	**0.33**	**0.80**	**3.44**	**6.41**	**0.06**	**0.12**	**-2.3E-03**	**-5.2E-02**	**1.5E-03**

			Proportion h^2 explained		Likelihood ratio				Δbias		
	Pathway	**Trait**	**95 percentile**[a]	**MultiBLUP**	**95 percentile**[a]	**MultiBLUP**	**Δr**[b]	**$\Delta r^2/h^2$**[c]	**intercept**	**slope**	**ΔMSE**
protein	degradation (ubiquitin)	ala_t	1.00	1.00	4.19	5.57	0.04	0.05	1.5E-05	-2.0E-01	8.9E-09
	degradation (ubiquitin)	val_PyrFam	0.47	0.63	8.76	13.53	0.05	0.02	2.3E-05	5.1E-02	-1.5E-07
	amino acid activation	gly	0.12	0.25	3.80	4.00	0.02	0.02	-3.6E-05	-2.0E-02	-6.0E-06
	amino acid activation	gly_t	0.17	0.38	3.81	5.38	0.03	0.05	-1.1E-04	-3.0E-02	-6.1E-05
	amino acid activation	**gly_SerFam**	**0.27**	**0.51**	**3.12**	**11.40**	**0.05**	**0.08**	**2.0E-05**	**-9.0E-02**	**-2.8E-06**
	amino acid activation	**ser_SerFam**	**0.28**	**0.51**	**3.16**	**11.08**	**0.05**	**0.08**	**3.6E-06**	**-9.3E-02**	**-1.5E-06**
	assembly	met	0.06	0.13	3.58	5.31	0.03	0.02	1.6E-05	6.6E-03	-6.8E-08
	assembly	met_t	0.04	0.12	3.30	4.05	0.01	0.01	-4.6E-05	-7.4E-03	-6.4E-07
	folding	his	0.09	0.14	8.86	9.56	0.02	0.01	-1.2E-03	3.2E-02	5.1E-04
	folding	met	0.14	0.19	3.58	3.67	0.01	0.01	2.3E-05	-2.8E-02	6.6E-08
	folding	ser	0.17	0.37	3.59	6.63	0.05	0.04	1.0E-03	-5.8E-02	2.3E-05
	folding	SerFam	0.13	0.23	4.16	6.75	0.04	0.03	9.3E-04	8.2E-03	2.7E-05
	folding	Total	0.09	0.18	8.48	9.06	0.02	0.01	-5.0E-05	2.3E-02	2.3E-07
	glycosylation	thr_AspFam	0.14	0.15	3.17	5.19	0.05	0.04	1.8E-05	1.5E-02	2.9E-08
	postrans	arg_t	0.93	1.00	3.95	4.45	0.02	0.02	1.4E-04	4.1E-02	-5.8E-07
	postrans	arg_GluFam	1.00	1.00	3.40	4.17	0.03	0.02	-7.4E-06	-1.4E-01	-2.4E-07
	postrans	glu_GluFam	0.63	0.71	3.11	3.28	0.02	0.00	-1.9E-05	1.1E-01	2.1E-06
	targeting	glu_t	0.30	0.44	3.36	3.98	0.02	0.02	-1.6E-02	2.7E-02	-3.4E-03

Trait and pathway combinations where the MultiBLUP model improved prediction accuracy by at least 5% are bolded. Changes in bias (zero centered) and mean squared error (MSE) were taken as the absolute value of the difference between the MultiBLUP and GBLUP model, with negative values suggesting less bias/error in the MultiBLUP model.

[a] 95 percentile based on random gene groups with the same number of markers.

[b] The difference in prediction accuracy (r) between the MultiBLUP and GBLUP models.

[c] The difference in reliability (r^2/h^2) between the MultiBLUP and GBLUP models.

Pathway-level association testing reveals novel SNP associations for FAA traits

The multiple testing correction in a typical GWAS is highly conservative, resulting in only the strongest marker-trait associations being classified as statistically significant at a genome-wide level [47,53]. To reassess previous GWAS results for FAA traits [23] at specific genomic regions, we performed pathway-level association testing for pathways which passed our significance criteria. When subsetting the GWAS *P*-values from [23] into biological pathways, we identify several novel associations that pass a false discovery rate (FDR) significance threshold of 10% (S3 Table). Similar to previous results, we found significant associations for several BCAA traits and the amino acid degradation pathway, which contained a known causal gene (BCAT2) associated with BCAA traits [22]. We also found additional associations with the amino acid degradation genes DELTA-OAT (At1g10060) and isovaleryl-CoA-dehydrogenase (IVD, At3g45300). The IVD protein was previously shown to influence all BCAAs [54], but has not been identified in GWAS or QTL mapping studies, further supporting the effectiveness of the MultiBLUP model to study genetic regulation of metabolites.

Other significant associations were found for absolute levels of methionine (Met) and SNPs in the category for protein folding (Atg01230) and for the family-based ratio of threonine (Thr_AspFam) and SNPs related to protein glycosylation (GALT31A, At1g32930; OST48, At5g66680). Single SNP associations were also identified for the family ratio of valine (Val_PyrFam) and the ubiquitin-mediated protein degradation category, for relative levels of glycine (Gly_t) and the protein amino acid activation category, and for absolute levels of phenylalanine (Phe) and the annotations for sulfur-containing specialized metabolites.

In the glutamate family, several significant associations were found for the alternative oxidase, amino acid degradation, and protein folding categories. For example, relative and family ratios of arginine (Arg_t, and Arg_GluFam, respectively) had a significant association with SNPs in the alternative oxidase 3 gene (AOX3, At1g32350), suggesting that free arginine may be related to alternative respiration. Notably, histidine related traits were associated with both the amino acid degradation category (His_GluFam) and SNPs related to protein folding (His) (Fig 3, S3 Table). Annotations for the genes that were found significant for His_GluFam (THA1, At1g08630; PED1, At2g33150; PYD4, Atg08860; LOG7, At5g06300) suggest that the metabolism of both threonine and lysine may be involved in determining the partition of histidine in dry *Arabidopsis* seeds, consistent with the observed interconnectivity within the amino acid metabolic pathway and the interdependent regulation of these amino acids [5,55].

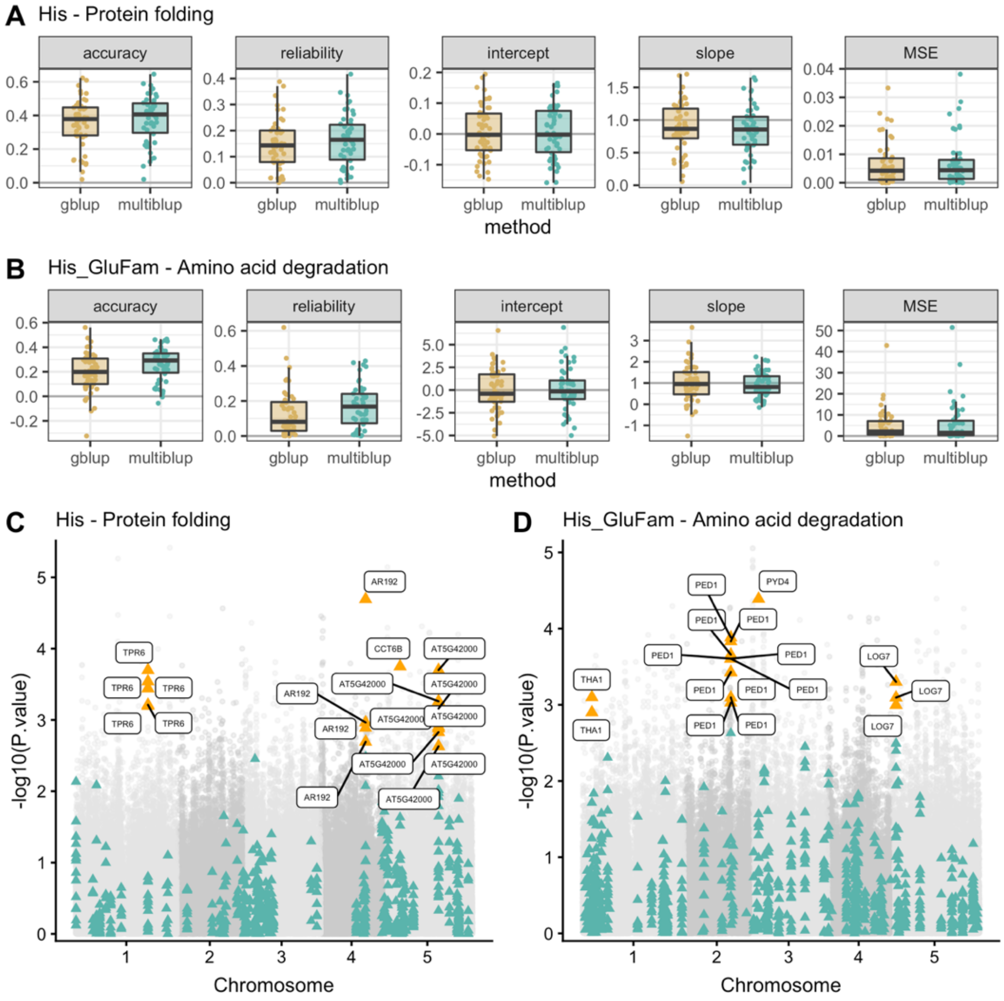


Fig 3. Increases in prediction accuracy inform pathway-guided GWAS to reveal novel SNPs related to histidine.

Comparison of MultiBLUP and GBLUP models for prediction accuracy, reliability, bias (intercept and slope), and mean squared error (MSE) for (A) absolute levels of histidine (His) and (B) the family ratio of histidine (His_GluFam). Note that the y-axis scale varies. GWAS results for (C) the protein folding category and His and for (D) the amino acid degradation category and His_GluFam. All SNPs are shown in gray, with pathway SNPs highlighted as blue triangles. SNPs with an FDR corrected p-value < 0.10 are annotated and highlighted in yellow.

Discussion

Previous studies on the genetic architecture for FAA and other metabolic traits suggest a complex genetic architecture comprised of small effect QTLs (e.g. [22,23,29]). These conclusions are recapitulated by previous biochemical and transcriptomic studies that investigated FAA homeostasis in the vegetative stage across changing environments [7,12,56,57]. Combined, these lines of evidence suggest FAA homeostasis is orchestrated by multiple pathways, including amino acid synthesis and degradation, primary metabolism, specialized metabolism, and protein metabolism (reviewed in [12]). In this study, we applied a genomic partitioning model (MultiBLUP) to investigate how FAA homeostasis is orchestrated in the model system *Arabidopsis thaliana,* allowing us to both test the feasibility of this approach and to further examine the genetic basis of FAAs in seeds. In addition to shedding light on the genetic complexity of FAA traits and the role of metabolic pathway genes in FAA homeostasis, this method can be used to develop hypotheses for biochemical and molecular studies.

Since its development nearly two decades ago [37], genomic prediction has dramatically altered the speed and scale of applied genetic and breeding research [58]. However, the use of genomic prediction has been primarily limited to agricultural species [59–61], likely because this is the realm where predicted breeding values are most directly applicable for breeding objectives. Recently, several studies have used genomic partitioning in prediction models to evaluate the relative influence of various genomic features, such as positional effects and gene annotation categories, on phenotypes of interest. Genomic partitioning is most successful when the partition is enriched for causal variant(s) [44], providing a framework for guided hypothesis testing. For example, [43] incorporated annotations for several biological pathways to determine which pathways were associated with udder health and milk production in dairy cattle. Similarly, gene ontology categories were leveraged to explore the genetic basis of different phenotypes in *Drosophila melanogaster* [42]. 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 [51]. The incorporation of prior biological information from transcriptomics, GWAS, and genes identified *in silico* also improved predictions of root phenotypes in cassava [62].

Surprisingly, genomic partitioning has not been widely applied in plants to decipher the underlying genetic contribution of biological processes to metabolic traits. We chose to use genomic partitioning to investigate amino acid traits, with the goal of advancing our understanding of metabolic systems, their complexity, and the genetic determinants that may contribute to homeostasis of FAAs in seeds. Because FAA traits are part of core metabolism that is highly conserved, we hypothesize that many of our findings can be used to develop similar hypotheses in crop systems, where there is potential to contribute to the biofortification of essential amino acids.

Genomic prediction of FAA traits in Arabidopsis seeds

We first established the efficacy of the GBLUP model in a diversity panel of 313 *Arabidopsis* individuals, which represents a substantial proportion of the known genetic variability present in Arabidopsis [63]. 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. [40,59,61]), 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. The observation of moderate prediction accuracies for many of these traits suggests that there is LD between markers and causal loci, providing evidence that genomic prediction can be successfully applied in this system. Interestingly, we observe higher prediction accuracies for a greater proportion of absolute FAA levels compared to relative levels and family-based ratios, consistent with the previous hypothesis that, compared to metabolic ratios, absolute levels of metabolites have a more complex genetic architecture, where many loci of small effect are contributing to genetic variation.

Genomic partitioning guided by metabolic processes generates new insights into the genetic basis of FAAs

We next applied a genomic partitioning approach, MultiBLUP, to investigate the association of different metabolic annotation categories with FAA traits in dry *Arabidopsis* seeds, focusing specifically on categories which are thought to influence FAA traits at this developmental stage. Our findings indicate that various FAA traits are associated with multiple biological pathways, many of which are not previously reported. On a broader scale, these results provide evidence that FAA composition in dry seeds is likely influenced by multiple metabolic processes rather than a single, predominant process. A notable caveat of this approach is that a given metabolic pathway may be in LD with an unrelated causal variant, and so the pathway itself may not be associated with the trait tested. In addition, this approach is also most effective when small SNP sets explain a large proportion of the phenotypic variance for a trait [41]. As such, some of the pathways tested in this study may have been too large to find an association.

Branched chain amino acid traits are associated with amino acid synthesis and degradation pathways. The inclusion of BCAA traits (leucine, isoleucine, valine) enabled both a proof of concept for the MultiBLUP approach and generated new insights into their genetic regulation. Previous work has demonstrated that a large effect QTL contributes to approximately 12-19% of the observed variability for BCAA traits, with the highest variance explained for relative level of isoleucine (Ile_t) [22]. The causal gene was identified as branched chain amino acid transferase 2 (BCAT2; At1g10070), which is part of the BCAA metabolic pathway [64]. Our results recapitulate this observation, showing that the amino acid degradation pathway, which contains the BCAT2 haploblock, explained both a significant proportion of heritability (43%) and improved prediction accuracy by 6.7% for Ile_t. This finding suggests that the MultiBLUP approach was effective at identifying a category of markers when a known causal variant is included.

Surprisingly, we also found that the BCAA family was the only group associated with amino acid synthesis, with a significant proportion of heritability explained for absolute levels of isoleucine, leucine, BCAA, and the pyruvate family composite trait. Previous work suggested that active amino acid synthesis is part of a metabolic switch occurring during the end of seed desiccation [21]. Under the metabolic switch scenario, we expected to see many FAA traits associated with the amino acid synthesis category. Instead, our results indicate that the effect of genes related to amino acid synthesis on FAA levels in dry seeds may be more limited. Furthermore, our findings further suggest that BCAA traits may also be influenced by genes related to glycolysis and isoprenoid metabolism, eluding to a more complex genetic architecture for these traits. Future studies will be necessary both to validate these observations and to further explore the genetic architecture for BCAA traits.

Specialized metabolism categories explain significant variation for aromatic amino acids. This study included measurements of natural variation for traits related to the aromatic amino acids (i.e. phenylalanine, tyrosine, and tryptophan). Notably, no pathway associations were identified for traits related to tryptophan. With the exception of an association of the composite aromatic family trait (ShikFam) and the amino acid transport pathway, these traits were exclusively associated with specialized metabolism pathways. Specific categories associated with aromatic FAA traits included phenylpropanoids, flavonoids, and sulfur-containing compounds (Fig 2, Table 3), consistent with the knowledge that aromatic amino acids can be converted to numerous specialized metabolites such as alkaloids, phenylpropanoids, and glucosinolates [65,66]. One notable pattern was that tyrosine-related traits were only associated with the flavonoid and phenylpropanoid categories. The finding of an association between tyrosine and flavonoids agrees with previous findings in transgenic rice seeds, which reported that flavonoids biosynthesized by exogenous enzymes may act as signaling molecules to alter amino acid biosynthesis [67]. For the family-based ratio of tyrosine (Tyr_ShikFam), we observed a 10.8% increase in prediction accuracy when SNPs from the phenylpropanoid pathway were partitioned in the MultiBLUP model, suggesting SNPs in this pathway are contributing to the variation for Tyr_ShikFam or are in strong LD with a causal variant. This is again consistent with biological expectations, as tyrosine is a known precursor for phenylpropanoid biosynthesis.

On the other hand, traits related to phenylalanine were associated with the pathway for sulfur-containing specialized metabolites (Fig 2), possibly influenced by a relationship to glucosinolates. The results from pathway-guided association mapping identified a significant SNP in the AOP1 gene (At4g03070, S3 Table), which encodes a probable 2-oxoglutarate-dependent dioxygenase involved in aliphatic glucosinolate biosynthesis. This result was surprising, as aliphatic glucosinolate biosynthesis begins with the chain elongation of methionine, suggesting that the relationship with phenylalanine in this case may be indirect. On the other hand, aromatic glucosinolates, which are produced from phenylalanine, are not considered widespread in *Arabidopsis* but are known to occur both in leaves and seeds in some ecotypes [68,69]. However, it is possible that the composition of aromatic glucosinolates in seeds and their effect on core metabolism is underestimated.

Interestingly, no association with nitrogenous specialized metabolism was detected for either phenylalanine or tyrosine, which are precursors for the nitrogen-containing compounds alkaloids. We also found no evidence of associations with protein metabolism, despite categories in this group being associated with most other amino acid families, and only one association with amino acid metabolism, suggesting that core metabolism may not play a critical role in the regulation of homeostasis for these traits.

Traits in the aspartate family, especially methionine, show relationships with amino acid degradation, specialized metabolism, and protein metabolism. Traits in the aspartate family were associated with multiple ontology categories. The most interesting of these were methionine-related traits, which were associated with amino acid degradation, specialized metabolism, and protein related metabolism. Since methionine is an essential amino acid, there have been many attempts to increase its content in seed crops via alteration of its metabolic pathway. Consistent with our observations, these attempts have also shown that alteration of methionine content in seeds affects multiple aspects of core metabolism [6,26]. We also found an association of methionine with the sulfur-containing specialized metabolism pathway. This finding is congruent with the knowledge that methionine is a precursor for aliphatic glucosinolate biosynthesis and with evidence that perturbing glucosinolates produces a significant increase in levels of free methionine in *Arabidopsis* leaves [16].

Traits in the serine family are exclusively associated with pathways related to protein metabolism. Within the serine family, traits were exclusively associated with the protein metabolism categories for amino acid activation and protein folding. Interestingly, family-based ratios for both glycine (Gly_SerFam) and serine (Ser_SerFam) showed an increase in prediction accuracy of 5% when partitioning SNPs related to amino acid activation in the MultiBLUP model. This suggests that genes related to amino acid activation, such as tRNA synthetases, may contribute to the homeostasis of glycine and serine.

Surprisingly, we did not observe a relationship of serine family traits with the amino acid synthesis category, which includes genes in the serine acetyltransferase (SAT) gene family. These enzymes catalyze the first step in the conversion of serine to cysteine (Cys), which can then be converted to methionine. In maize kernels, overexpression of SAT has been linked to increased sulfur assimilation and higher levels of methionine, without incurring detrimental effects on plant yield [70]. Notably, measurements of cysteine are not included in the present study, and thus we may be unable to fully capture the dynamics of this agronomically important relationship.

The glutamate family showed surprising associations with amino acid degradation and sulfur-containing specialized metabolism. Traits in the glutamate family were associated with amino acid degradation, primary metabolism, specialized metabolism, and protein metabolism.

Amino acids in the glutamate family are known to play a central role in core metabolism, mainly by functioning as one of the entry points for nitrogen into plants and via connections to the TCA cycle [7,71]. Hence, it was not surprising to find traits in this family associated with multiple categories, including the association of arginine traits with the pathway related to alternative oxidase activity (S3 Table). Two surprising associations were also identified: the association of His_GluFam with amino acid degradation (Fig 3) and the association of glutamine related traits with sulfur-containing specialized metabolism (Fig 2). In each case, prediction accuracy was increased substantially (>5%) (Table 3). For His_GluFam and the amino acid degradation category, pathway-guided association mapping identified SNPs in genes related to the catabolism of lysine and threonine, suggesting that these processes may be involved in the regulation of histidine composition in seeds (S3 Table). The genetic architecture for histidine is of special interest, with evidence suggesting that levels of histidine in seeds can influence important agronomic traits such as seed oil deposition [72]. However, the metabolic pathway for histidine biosynthesis and catabolism is not yet fully understood [73,74]. Previous work using network-guided GWAS has identified CAT4, a vacuolar transporter, that was associated with histidine traits [23]. Here, we present evidence that regulation of histidine may also be influenced by genes related to other aspects of amino acid degradation.

Conclusions

Our results demonstrate that genomic partitioning is a useful technique to identify genomic categories or features that are more likely to harbor causal variants. We leveraged genomic partitioning models to identify genomic regions that increase prediction accuracy. Using this approach, we are able to reduce the search space for causal variants and to identify novel candidate genes for traits related to methionine, threonine, histidine, arginine, glycine, phenylalanine, and BCAAs (S3 Table). These results can be used as a platform to further explore the biofortification of seed amino acids, to deepen our understanding of metabolic regulation, and to identify candidate regions for functional validation. Furthermore, this strategy of genomic partitioning and pathway association may be useful for classifying the genetic architecture of other complex metabolic traits in additional species.

Methods

Plant materials and trait data

For this study, we reanalyzed data of the absolute levels, relative compositions, and biochemical ratios for free amino acids in dry *Arabidopsis thaliana* seeds. These traits were previously measured in [22,23] for 313 accessions of the Regional Association Mapping panel [63,75]. In summary, seeds from two plants of each accession were harvested from three independent grow outs. Absolute levels of FAAs (nmol/mg seed) were quantified using liquid chromatography–tandem mass spectrometry multiple reaction monitoring (LC-MS/MS MRM; see [22,23] 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 (S1 Fig, S1 Table). This prior knowledge of biochemical relationships among FAAs was 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 [30,76,77]. 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 [23]. Traits and their respective abbreviations are described in S1 Table. Overall, the 65 phenotypes included 25 absolute FAA levels (individual amino acids and composite traits), 17 relative levels (ratio of absolute level for an amino acid compared to total FAA content), and 23 family-derived traits (ratio of absolute level for an amino acid to the total FAA content within a given family).

The best linear unbiased predictors (BLUPs) for each accession, reported in [22], were used as the phenotype data in this study. Briefly, BLUPs were generated by first fitting a mixed model including replicate and accessions as random effects. Outliers were then removed for 38 of the 65 traits based on Studentized deleted residuals [78]. Following outlier removal, the Box-Cox procedure [79] was applied to transform each trait to avoid violating model assumptions for normally distributed error terms and constant variance. The BLUP for each accession was then determined for all transformed traits using a mixed model fit across all three replicates. This procedure removed the effect of growing environment but did not account for genetic differences.

Genetic data

The accessions used in this study were previously genotyped using a 250k SNP panel [80], v3.06. The software PLINK v1.9 was used to filter for minor allele frequency (MAF) > 0.05 (--maf 0.05), reducing the number of SNPs from 214,051 to 199,452.

To partially account for population structure, quality filtered SNPs were first pruned for linkage disequilibrium (LD) in PLINK v1.9 using a window size of 10kb that shifted by five SNPs and a pairwise LD threshold of 0.1. The SNPs exceeding this threshold were removed, reducing the number of SNPs from 199,452 to 45,122. These LD pruned SNPs were then used as the input for principal component analysis in R v3.6.0 [86] using the ‘prcomp’ function. Phenotypes were adjusted for population structure by regressing the first six principal components, which explained 9.4% of the variance (S3 Fig), against each phenotype and returning the residuals (see similar approach in [81]). These residuals were used as the phenotypes for downstream analyses along with the full set of 199,452 quality filtered SNPs.

Selection of pathway SNPs

To examine specific metabolic pathways, SNPs were selected based on annotation categories in the MapMan software [82] for the TAIR10 annotation of *Arabidopsis* [83]. We focused broadly on four categories: amino acid metabolism (three pathways), primary metabolism (three pathways), specialized metabolism (five pathways), and protein metabolism (nine pathways) (Table 2). The SNP positions were first matched to the corresponding Ensembl gene id using the biomaRt package [84,85] in R v3.6.0 [86]. We then selected all SNPs within a 2.5 kb range of the start and stop position for each gene, which is within the range of the estimated average in *Arabidopsis* [87] and includes upstream promoter regions. Specific pathways and corresponding MapMan annotation categories, including the number of genes and SNPs represented, are described in Table 2. We followed MapMan annotations for all genes except BCAT2 (At1g10070), which was moved from the amino acid synthesis pathway to the amino acid degradation pathway along with other SNPs in the same haploblock (chromosome 1, 3274080 to 397645 bp). This decision was based on previous work in which *bcat2* mutants showed higher accumulation of branched-chain amino acids in seeds, thereby demonstrating that BCAT2 has catabolic activity [22].

Prediction models

The Linkage Disequilibrium Adjusted Kinship (LDAK) software v5.0 [88] (http://dougspeed.com/ldak/) was used to implement two models for genomic prediction of each trait: GBLUP, in which random effects are drawn from the same effect size distribution, and MultiBLUP, in which random effects can be drawn from distributions with distinct effect size variances [41]. First, the pairwise genetic similarity between individuals was estimated using a genomic similarity matrix (GSM), or kinship matrix [89,90]:

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

where X is a matrix of SNP genotypes, X' is the transpose of X, and p is the number of SNPs.

Genomic prediction was performed for all markers using a random regression BLUP (RR-BLUP) model as described in [37,91], in which phenotypes are regressed against markers that share a common effect size variance distribution. Briefly, this model equates each phenotypic value to a normally distributed random effect of each marker, and the BLUP of each random marker effect is subjected to a ridge regression penalty. The RR-BLUP model is considered equivalent to a GBLUP model, which uses a genomic relationship matrix in place of markers [37].

To model biological pathways, we used the MultiBLUP model, which extends the RR-BLUP model to incorporate multiple kinship matrices as random effects with distinct effect size variances. For this study, the MultiBLUP model included random effects for sets of markers within a biological pathway (m) and for the remaining markers not included in a given pathway ($\notin m$). Following equation (1), markers within a biological pathway have a correlation structure K^m, with the matrix form X^m, where columns refer to the set of markers in the pathway. In this case, the set of pathway markers, R_m, contains a total of p_m markers with the effect size of the j^{th} marker distributed as $\beta_j^m \sim N(0, \sigma_m^2/p_m)$. Similarly, the correlation structure for the remaining markers is $K^{\notin m}$, has the matrix form $X^{\notin m}$ for the set $R_{\notin m}$ of size $p_{\notin m}$ and the effect size of the j^{th} marker is distributed as $\beta_j^{\notin m} \sim N(0, \sigma_{\notin m}^2 / p_{\notin m})$. These terms were used in the following random regression model from [41] to perform MultiBLUP:

$$Y_i \;=\; \beta_0 + \sum_{j \in R_m} X_{ij}^m \beta_j^m + \sum_{j \notin R_m} X_{ij}^{\notin m} \beta_j^{\notin m} + \varepsilon_i, \qquad (2)$$

where Y_i is the observed phenotypic value of the i^{th} individual, β_0 is the intercept, and ε_i is the normally distributed random error term associated with the i^{th} individual.

For our purposes, kinship matrices were estimated in the LDAK software for either all SNPs (GBLUP) or each SNP partition (MultiBLUP, i.e. pathway SNPs and all other remaining SNPs) by ignoring LD adjusted SNP weightings (--ignore-weights YES) so that each marker in the model was assigned an effect. This avoids distributing a marker effect to neighboring markers that are in strong LD, which can increase noise in the prediction model, although it may bias estimates of variance. Predictors were scaled by setting the parameter α = 0 (--power 0), a commonly used value in plant and animal breeding that assumes each SNP has the same effect size distribution regardless of MAF [92].

Heritability

The GBLUP and MultiBLUP models use average information restricted maximum likelihood (REML, see [41] for details) to compute variance component estimates for $\sigma_1^2, \ldots, \sigma_M^2$ and σ_e^2. Because we were only interested in a single partition for any given pathway, we refer to variance estimates for a given partition m as $\hat{\sigma}_m^2$ and variance estimates for all other markers not included in this partition as $\hat{\sigma}_{\notin m}^2$. 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):

$$h^2 = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{\notin m}^2 + \sigma_e^2}\,. \qquad (3)$$

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

$$\frac{h_m^2}{h_m^2 + h_{\notin m}^2}, \tag{4}$$

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.

Model performance

The performance of prediction models was determined using ten-fold cross validation with a one-fold holdout, with the same training and testing sets used for the GBLUP and MultiBLUP models. For each cross validation, 10% of the data were withheld when fitting the GBLUP and MultiBLUP models. Variance estimates from REML were then used to determine the genomic estimated breeding value (GEBV) based on marker data for the excluded individuals. This process was repeated five times for a total of 50 cross validations per trait. Prediction accuracy was then calculated as $r(\hat{g}, g)$, where $\hat{g}$ represents the estimated breeding values and g represents the observed phenotype values. Reliability, which is the coefficient of determination (r^2) scaled by heritability, was calculated as $\frac{r^2}{h^2}$ [93]. Bias was calculated as the simple linear regression coefficients (i.e., the intercept and slope estimates) between the estimated breeding values and observed phenotype, with a slope estimate of one and an intercept estimate of zero indicating no bias. Lastly, mean squared error (MSE), which measures prediction bias and variability, was calculated as the mean of the squared difference between the observed phenotypes and GEBVs, $\frac{1}{n}\sum(g - \hat{g})^2$ where *n* is the number of observations.

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. 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 [43]. To establish a null distribution, we first defined 5,000 random gene groups with a target number of SNPs that ranged uniformly from 1 to 50,000 SNPs. For each random subset, all SNPs within 2.5 kb of the start and stop positions were sampled for a randomly selected gene.

This process was repeated by randomly sampling genes one at a time until the target number of SNPs for each subset was achieved. As discussed in [43], this approach does not explicitly model variation in other parameters (e.g., allele frequencies, the number of markers, and 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) as a measure of pathway model fit compared to the model fit of random SNP subsets. Each metric was evaluated by testing pathway values against the null distribution of values computed from the random gene groups described above. The proportion of heritability explained was calculated as described previously and the LR was calculated as twice the difference between the log likelihood of the MultiBLUP model and the log likelihood of the GBLUP model. Significant values for both of these metrics suggest that a given pathway annotation has biological importance [43].

To establish significance thresholds for the LR and proportion of heritability explained, we first accounted for rounding errors by setting heritability estimates that were negative to zero and greater than one to one. These negative estimates are possible because we did not constrain estimates to be non-negative in the REML solver (--constrain NO) and may occur as a consequence of small sample size and/or if the true heritability is low. Heritability estimates with negative standard deviations and/or a negative LR suggested the model did not converge and were excluded (S2 Table). Relatively few random gene groups were filtered for each trait except valine (Val), which had a high proportion (1504 observations) of random gene groups with a negative LR. Significance thresholds were then determined based on the 95th percentile of both the proportion of heritability explained and the LR using smooth quantile regression in the R package ‘quantreg’ with constraint set to ‘increasing’.

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 MultiBLUP model explained a greater proportion of the genomic heritability than the 95th percentile of the same number of randomly selected markers.
2.) The LR for the MultiBLUP model was greater than the 95th percentile of LR for the same number of randomly selected markers.
3.) The MultiBLUP model improved prediction accuracy by at least 1% compared to the GBLUP model.

Together, criteria (1) and (2) established that the pathway being tested contained significantly more information than a random set of SNPs. Criteria (3) was imposed to ensure that there was a meaningful difference in prediction accuracy when pathway information was incorporated via MultBLUP compared to the naive GBLUP model that incorporated no pathway information.

Pathway-level association analysis

If a given trait and pathway combination passed all of the above criteria, P-values for the SNPs in the pathway were selected from the GWAS results reported in [22,23]. For each trait and pathway combination, the Benjamini and Hochberg [94] procedure was conducted on the corresponding set of SNPs to control the false discovery rate (FDR) at 10%.

Data availability

Genotype data are previously published and were accessed from https://github.com/Gregor-Mendel-Institute/atpolydb/wiki [80]. The scripts and phenotypic data used for this analysis are publicly available on GitHub at https://github.com/mishaploid/aa-genomicprediction.

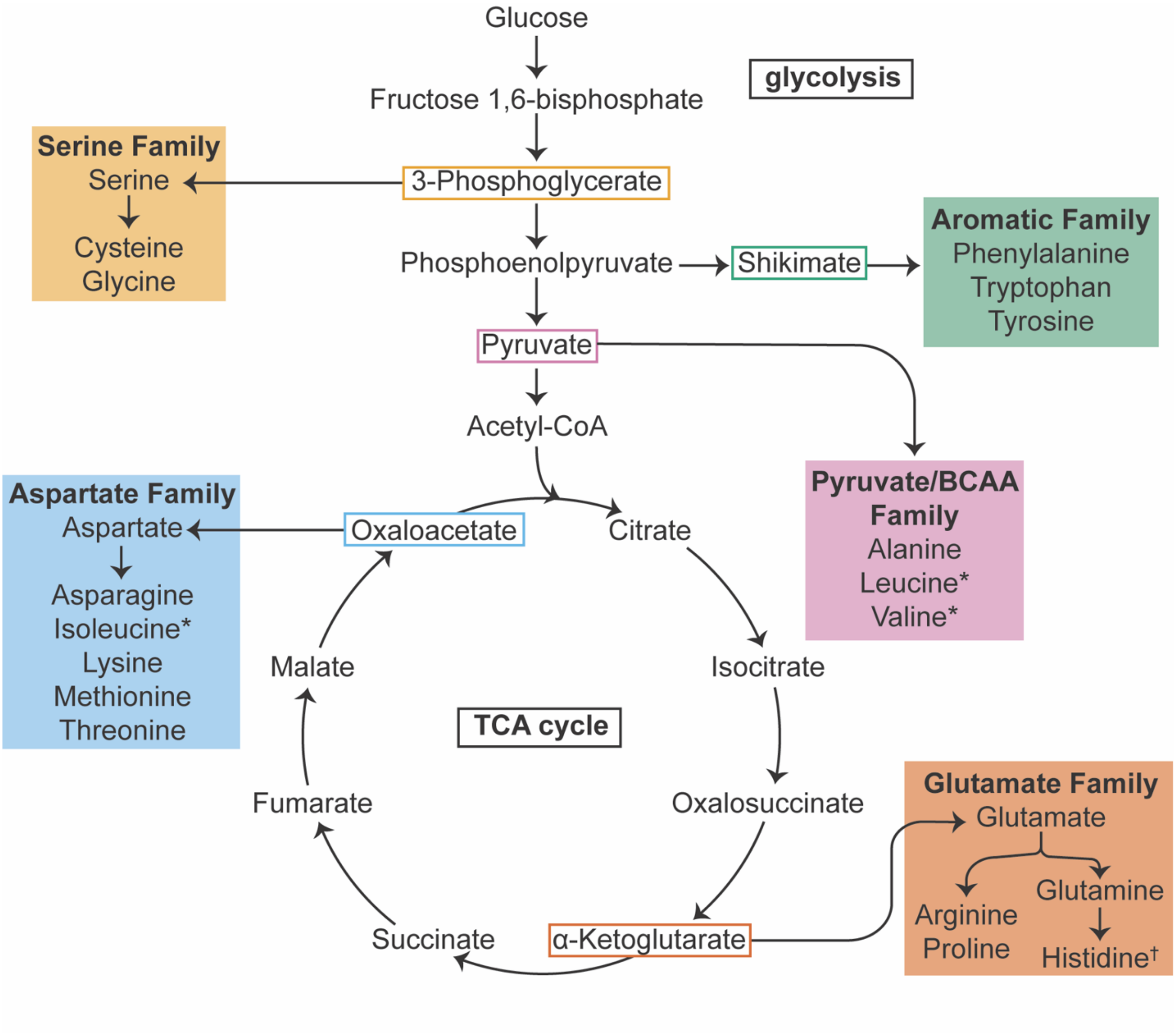
Acknowledgements

We are grateful to Dan Kliebenstein, Jinliang Yang, Jeffrey Ross-Ibarra, and two anonymous reviewers for helpful comments and discussions that improved the manuscript. We also thank Doug Speed for advice on cross-validation in LDAK.

Funding

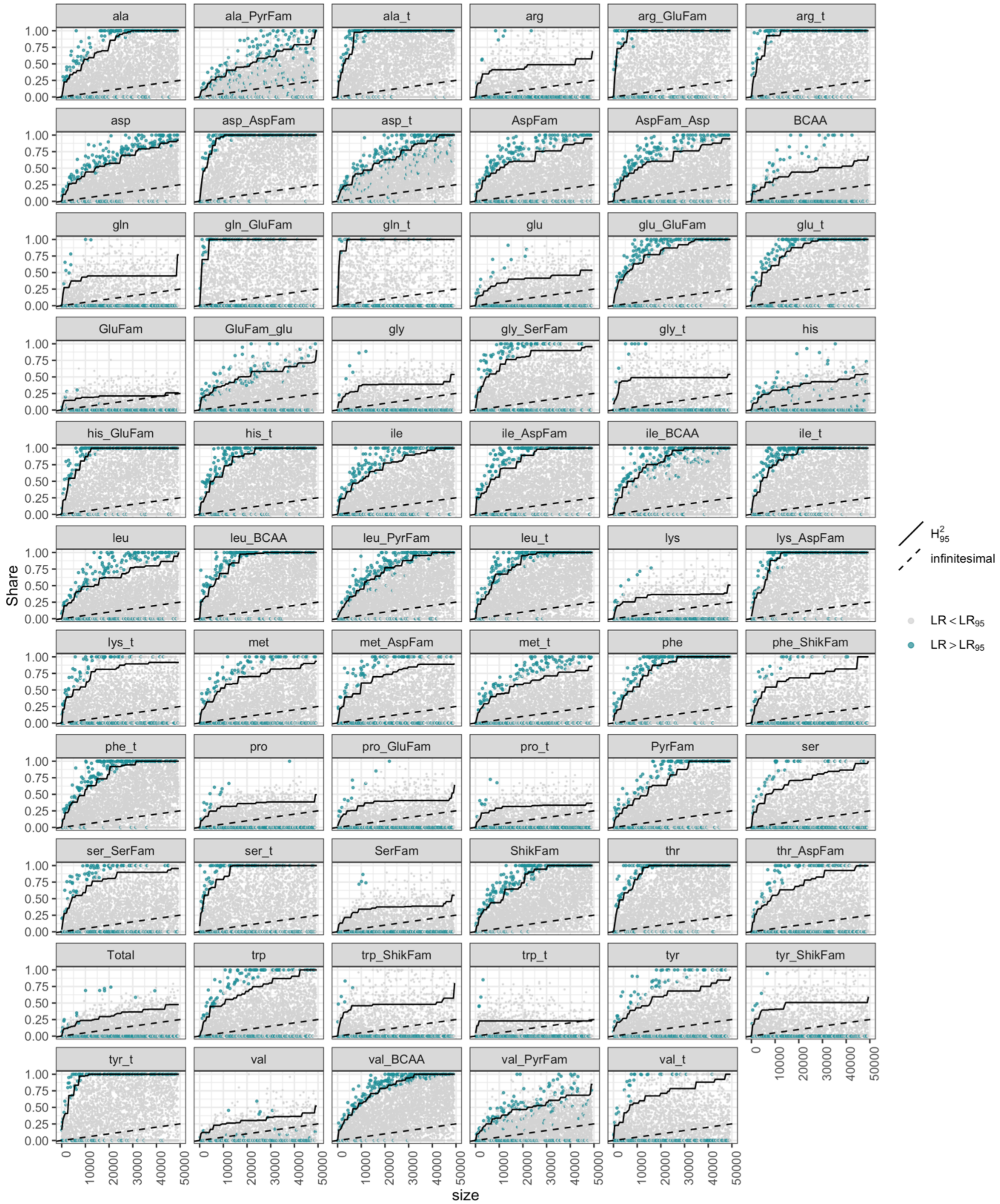
This project was supported by the USDA Agricultural Research Service, the University of Missouri Division of Biological Sciences (Columbia, MO, USA), the NSF Postdoctoral Research Fellowship in Biology Grant No. 1711347 for SDT, and the NSF Graduate Research Fellowship Grant No. DGE-1424871 for KAB. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

639 ## Supporting information

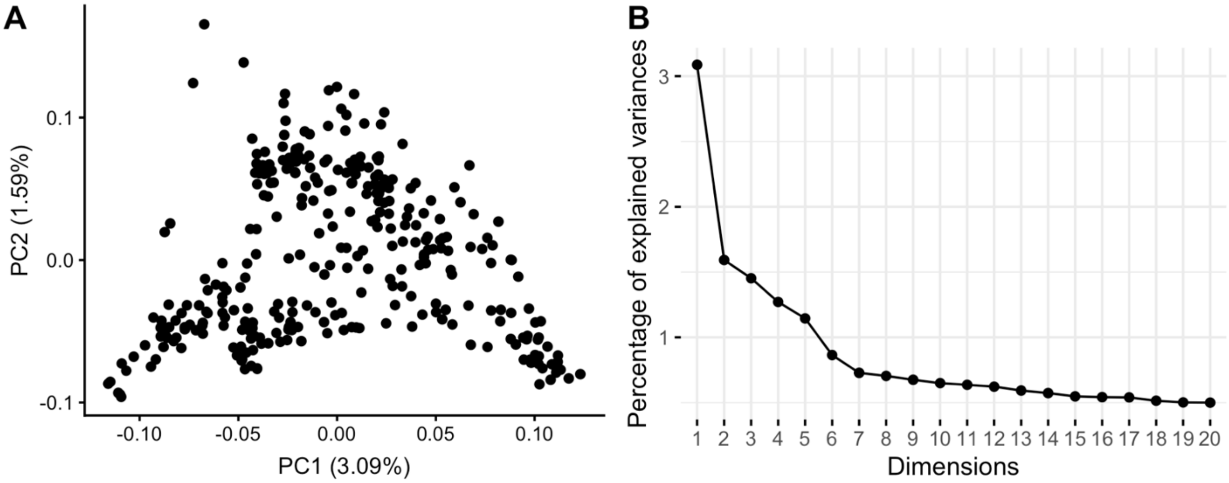


640 **S1 Fig. Biochemical relationships among amino acids.**

641 Colors indicate different amino acid families and boxes indicate the corresponding precursor. The 642 branched-chain amino acids include Leu, Ile, Val are split across the Aspartate and Pyruvate family 643 and therefore denoted with asterisks (*). Note that histidine (†) does not belong explicitly to the 644 families identified here, but often is considered as part of the glutamate family.



645 **S2 Fig. Genomic variance explained by 5,000 random SNP subsets for free amino acid traits.**
646 Each point represents a different random gene group with the number of SNPs indicated on the x-
647 axis. The solid line indicates the 95th percentile for the proportion of heritability explained and the
648 dashed line represents the expectation when all SNPs have a similar effect size. Points are colored
649 blue if the likelihood ratio for a random set exceeds 95% percentile for the LR of the same trait.



S3 Fig. Principal component analysis (PCA) of genetic data for the 313 Arabidopsis accessions used in this study.

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

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

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

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

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

* T/Total not included due to errors when generating BLUPs

Biochemistry based metabolic ratios, grouped by AA families' affiliation	
Asp Family = Ile, Met, Thr, Asp Lys (IMTDK)	**BCAA Family = Ile, Val, Leu (IVL) Pyr Family=Leu, Ala, Val (LAV)**
D/IMTDK	I/IVL
K/IMTDK	V/IVL
M/IMTDK	L/IVL
T/IMTDK	A/LAV
IMTDK/D	L/LAV
	V/LAV
	I/IMTDK
Glu Family = Glu, His, Pro, Arg, Gln (EHPRQ)	
Q/EHPRQ	**Shikimate (Aromatic) Fam = Trp, Phe,Tyr (WFY)**
E/EHPRQ	W/WFY
H/EHPRQ	F/WFY
P/EHPRQ	Y/WFY
R/EHPRQ	
EHPRQ/E	**Ser Family = Ser, Gly (Cysteine-not detected -SG)**
	G/SG
	S/SG

S2 Table. Summary of null gene groups for each free amino acid trait. Includes the number of gene groups that passed filtering criteria, failed to converge, or had a negative likelihood ratio statistic.

Trait	Number of gene groups	Failed to converge	Negative LR
ala	4998	1	1
ala_PyrFam	4958	26	17
ala_t	4998	0	2
arg	4996	0	4
arg_GluFam	4997	1	3
arg_t	4992	6	4
asp	4997	0	3
asp_AspFam	4999	1	1
asp_t	4992	5	3
AspFam	4999	0	1
AspFam_Asp	4999	0	1
BCAA	4988	0	12
gln	5000	0	0
gln_GluFam	4997	1	3
gln_t	4985	10	5
glu	4998	0	2
glu_GluFam	5000	0	0
glu_t	4994	2	4
GluFam	4954	10	36
GluFam_glu	4989	2	9
gly	4999	0	1
gly_SerFam	4996	1	3
gly_t	4994	1	5
his	4999	0	1
his_GluFam	4997	1	2
his_t	4997	0	3
ile	4999	0	1
ile_AspFam	4999	0	1
ile_BCAA	4995	3	2
ile_t	5000	0	0
leu	4998	0	2
leu_BCAA	4998	1	1
leu_PyrFam	4982	2	16
leu_t	4998	0	2
lys	5000	0	0
lys_AspFam	4998	1	1
lys_t	4997	0	3
met	4999	0	1
met_AspFam	4995	0	5
met_t	4999	1	0
phe	4997	1	2
phe_ShikFam	4998	0	2
phe_t	4997	0	3
pro	5000	0	0
pro_GluFam	4998	0	2
pro_t	4999	0	1
PyrFam	4997	1	2

Trait	Number of gene groups	Failed to converge	Negative LR
ser	5000	0	0
ser_SerFam	4996	0	4
ser_t	4997	0	3
SerFam	4997	0	3
ShikFam	4999	0	1
thr	4996	0	4
thr_AspFam	4998	0	2
Total	4999	0	1
trp	4999	0	1
trp_ShikFam	4993	1	6
trp_t	4997	1	2
tyr	4999	0	1
tyr_ShikFam	4997	0	3
tyr_t	4997	1	2
val	3491	8	1504
val_BCAA	5000	0	0
val_PyrFam	4987	7	6
val_t	4997	0	3

S3 Table. Significant results from pathway guided association testing (α = 0.10). Columns include the original GWAS p-values, the number of SNPs tested for each pathway, and the pathway-level FDR corrected p-value. (see supplementary information)

References

1. Wu Y, Messing J. Proteome balancing of the maize seed for higher nutritional value. Front Plant Sci. 2014;5: 240. doi:10.3389/fpls.2014.00240

2. Angelovici R, Galili G, Fernie AR, Fait A. Seed desiccation: a bridge between maturation and germination. Trends Plant Sci. 2010;15: 211–218. doi:10.1016/j.tplants.2010.01.003

3. Rai VK. Role of Amino Acids in Plant Responses to Stresses. Biol Plant. 2002;45: 481–487. doi:10.1023/A:1022308229759

4. Araújo WL, Ishizaki K, Nunes-Nesi A, Larson TR, Tohge T, Krahnert I, et al. 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. 2010;22: 1549–1563. doi:10.1105/tpc.110.075630

5. Angelovici R, Fait A, Fernie AR, Galili G. A seed high-lysine trait is negatively associated with the TCA cycle and slows down Arabidopsis seed germination. New Phytol. 2011;189: 148–159. doi:10.1111/j.1469-8137.2010.03478.x

6. Amir R, Galili G, Cohen H. The metabolic roles of free amino acids during seed development. Plant Sci. 2018;275: 11–18. doi:10.1016/j.plantsci.2018.06.011

7. Hildebrandt TM, Nunes Nesi A, Araújo WL, Braun H-P. Amino Acid Catabolism in Plants. Mol Plant. 2015;8: 1563–1579. doi:10.1016/j.molp.2015.09.005

8. Huang T, Jander G. Abscisic acid-regulated protein degradation causes osmotic stress-induced accumulation of branched-chain amino acids in Arabidopsis thaliana. Planta. 2017;246: 737–747. doi:10.1007/s00425-017-2727-3

9. Less H, Galili G. Principal Transcriptional Programs Regulating Plant Amino Acid Metabolism in Response to Abiotic Stresses. Plant Physiol. 2008;147: 316–330. doi:10.1104/pp.108.115733

10. Jander G, Joshi V. Recent progress in deciphering the biosynthesis of aspartate-derived

amino acids in plants. Mol Plant. 2010;3: 54–65. doi:10.1093/mp/ssp104

11. Barros JAS, Cavalcanti JHF, Medeiros DB, Nunes-Nesi A, Avin-Wittenberg T, Fernie AR, et al. Autophagy Deficiency Compromises Alternative Pathways of Respiration following Energy Deprivation in Arabidopsis thaliana. Plant Physiol. 2017;175: 62–76. doi:10.1104/pp.16.01576

12. Hildebrandt TM. Synthesis versus degradation: directions of amino acid metabolism during Arabidopsis abiotic stress response. Plant Mol Biol. 2018;98: 121–135. doi:10.1007/s11103-018-0767-0

13. Hirota T, Izumi M, Wada S, Makino A, Ishida H. 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. 2018;59: 1363–1376. doi:10.1093/pcp/pcy005

14. Hayat S, Hayat Q, Alyemeni MN, Wani AS, Pichtel J, Ahmad A. Role of proline under changing environments: a review. Plant Signal Behav. 2012;7: 1456–1466. doi:10.4161/psb.21949

15. Szabados L, Savouré A. Proline: a multifunctional amino acid. Trends Plant Sci. 2010;15: 89–97. doi:10.1016/j.tplants.2009.11.009

16. Chen Y-Z, Pang Q-Y, He Y, Zhu N, Branstrom I, Yan X-F, et al. Proteomics and metabolomics of Arabidopsis responses to perturbation of glucosinolate biosynthesis. Mol Plant. 2012;5: 1138–1150. doi:10.1093/mp/sss034

17. Osorio S, Vallarino JG, Szecowka M, Ufaz S, Tzin V, Angelovici R, et al. 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. 2013;161: 628–643. doi:10.1104/pp.112.211094

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

2014;5: 447. doi:10.3389/fpls.2014.00447

19. Muehlbauer GJ, Gengenbach BG, Somers DA, Donovan CM. Genetic and amino-acid analysis of two maize threonine-overproducing, lysine-insensitive aspartate kinase mutants. Theor Appl Genet. 1994;89: 767–774. doi:10.1007/BF00223717

20. Cohen H, Israeli H, Matityahu I, Amir R. 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. 2014;166: 1575–1592. doi:10.1104/pp.114.246058

21. Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie AR, et al. Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. Plant Physiol. 2006;142: 839–854. doi:10.1104/pp.106.086694

22. Angelovici R, Lipka AE, Deason N, Gonzalez-Jorge S, Lin H, Cepela J, et al. Genome-Wide Analysis of Branched-Chain Amino Acid Levels in Arabidopsis Seeds. Plant Cell. 2013;25: 4827–4843. doi:10.1105/tpc.113.119370

23. Angelovici R, Batushansky A, Deason N, Gonzalez-Jorge S, Gore MA, Fait A, et al. Network-guided GWAS improves identification of genes affecting free amino acids. Plant Physiol. 2016; doi:10.1104/pp.16.01287

24. Wang X, Larkins BA. Genetic Analysis of Amino Acid Accumulation inopaque-2 Maize Endosperm. Plant Physiol. 2001;125: 1766–1777. doi:10.1104/pp.125.4.1766

25. Schmidt MA, Barbazuk WB, Sandford M, May G, Song Z, Zhou W, et al. 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. 2011;156: 330–345. Available: http://www.plantphysiol.org/content/156/1/330.short

26. Galili G, Amir R. Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality [Internet]. Plant Biotechnology Journal. 2013. pp. 211–222. doi:10.1111/pbi.12025

27. Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, et al. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci U S A. 2012;109: 8872–8877. doi:10.1073/pnas.1120813109

28. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9: 29. doi:10.1186/1746-4811-9-29

29. Wu S, Alseekh S, Cuadros-Inostroza Á, Fusari CM, Mutwil M, Kooke R, et al. Combined Use of Genome-Wide Association Data and Correlation Networks Unravels Key Regulators of Primary Metabolism in Arabidopsis thaliana. PLoS Genet. 2016;12: e1006363. doi:10.1371/journal.pgen.1006363

30. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet. 2007;3: 1687–1701. doi:10.1371/journal.pgen.0030162

31. Vallabhaneni R, Wurtzel ET. Timing and biosynthetic potential for carotenoid accumulation in genetically diverse germplasm of maize. Plant Physiol. 2009;150: 562–572. doi:10.1104/pp.109.137042

32. Wurtzel ET, Cuttriss A, Vallabhaneni R. Maize provitamin a carotenoids, current resources, and future metabolic engineering challenges. Front Plant Sci. 2012;3: 29. doi:10.3389/fpls.2012.00029

33. Gonzalez-Jorge S, Ha S-H, Magallanes-Lundback M, Gilliland LU, Zhou A, Lipka AE, et al. Carotenoid cleavage dioxygenase4 is a negative regulator of β-carotene content in Arabidopsis seeds. Plant Cell. 2013;25: 4812–4826. doi:10.1105/tpc.113.119677

34. Lipka AE, Gore MA, Magallanes-Lundback M, Mesberg A, Lin H, Tiede T, et al. Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. G3 . 2013;3: 1287–1299. doi:10.1534/g3.113.006148

35. Owens BF, Lipka AE, Magallanes-Lundback M, Tiede T, Diepenbrock CH, Kandianis CB, et al. A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. Genetics. 2014;198: 1699–1716.

doi:10.1534/genetics.114.169979

36. Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, et al. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science. 2008;319: 330–333. doi:10.1126/science.1150255

37. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157: 1819–1829. Available: https://www.ncbi.nlm.nih.gov/pubmed/11290733

38. de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193: 327–345. doi:10.1534/genetics.112.143313

39. Goddard ME, Wray NR, Verbyla K, Visscher PM. Estimating Effects and Making Predictions from Genome-Wide Marker Data. Stat Sci. 2009;24: 517–529. doi:10.1214/09-STS306

40. Heffner EL, Sorrells ME, Jannink J-L. Genomic Selection for Crop Improvement. Crop Sci. 2009;49: 1–12. doi:10.2135/cropsci2008.08.0512

41. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24: 1550–1557. doi:10.1101/gr.169375.113

42. Edwards SM, Sørensen IF, Sarup P, Mackay TFC, Sørensen P. Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. Genetics. 2016;203: 1871–1883. doi:10.1534/genetics.116.187161

43. Edwards SM, Thomsen B, Madsen P, Sørensen P. Partitioning of genomic variance reveals biological pathways associated with udder health and milk production traits in dairy cattle. Genet Sel Evol. 2015;47: 60. doi:10.1186/s12711-015-0132-6

44. Sarup P, Jensen J, Ostersen T, Henryon M, Sørensen P. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in

purebred Danish Duroc pigs. BMC Genet. 2016;17: 11. doi:10.1186/s12863-015-0322-9

45. Fang L, Sahana G, Ma P, Su G, Yu Y, Zhang S, et al. 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. 2017;49: 44. doi:10.1186/s12711-017-0319-0

46. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17: 144. doi:10.1186/s12864-016-2443-6

47. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011;43: 519–525. doi:10.1038/ng.823

48. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24: 1550–1557. doi:10.1101/gr.169375.113

49. Fang L, Sahana G, Ma P, Su G, Yu Y, Zhang S, et al. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. BMC Genomics. 2017;18: 604. doi:10.1186/s12864-017-4004-z

50. Sørensen IF, Edwards SM, Rohde PD, Sørensen P. Multiple Trait Covariance Association Test Identifies Gene Ontology Categories Associated with Chill Coma Recovery Time in Drosophila melanogaster. Sci Rep. 2017;7: 2413. doi:10.1038/s41598-017-02281-3

51. Li X, Zhu C, Yeh C-T, Wu W, Takacs EM, Petsch KA, et al. Genic and nongenic contributions to natural variation of quantitative traits in maize. Genome Res. 2012;22: 2436–2444. doi:10.1101/gr.140277.112

52. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467: 832–838. doi:10.1038/nature09410

53. Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, et al. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol. 2015;24: 110–118. doi:10.1016/j.pbi.2015.02.010

54. Gu L, Jones AD, Last RL. Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. Plant J. 2010;61: 579–590. doi:10.1111/j.1365-313X.2009.04083.x

55. Toubiana D, Batushansky A, Tzfadia O, Scossa F, Khan A, Barak S, et al. Combined correlation-based network and mQTL analyses efficiently identified loci for branched-chain amino acid, serine to threonine, and proline metabolism in tomato seeds. Plant J. 2015;81: 121–133. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.12717

56. Skirycz A, Vandenbroucke K, Clauw P, Maleux K, De Meyer B, Dhondt S, et al. Survival and growth of Arabidopsis plants given limited water are not equal. Nat Biotechnol. 2011;29: 212–214. doi:10.1038/nbt.1800

57. Skirycz A, De Bodt S, Obata T, De Clercq I, Claeys H, De Rycke R, et al. Developmental stage specificity and the role of mitochondrial metabolism in the response of Arabidopsis leaves to prolonged mild osmotic stress. Plant Physiol. 2010;152: 226–244. doi:10.1104/pp.109.148965

58. Daetwyler HD, Calus MPL, Pong-Wong R, de Los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013;193: 347–365. doi:10.1534/genetics.112.147983

59. Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan NP, et al. Implementation of genomic selection in the poultry industry. Animal Frontiers. 2016;6: 23–31. Available: https://pdfs.semanticscholar.org/c7c9/899b866d8814fe32ae46c8a50dfdbe68edc7.pdf

60. Nielsen NH, Jahoor A, Jensen JD, Orabi J, Cericola F, Edriss V, et al. Genomic Prediction of Seed Quality Traits Using Advanced Barley Breeding Lines. PLoS One. 2016;11: e0164494. doi:10.1371/journal.pone.0164494

61. Weller JI, Ezra E, Ron M. Invited review: A perspective on the future of genomic selection

in dairy cattle. J Dairy Sci. 2017;100: 8633–8644. doi:10.3168/jds.2017-12879

62. Lozano R, del Carpio DP, Amuge T, Kayondo IS, Adebo AO, Ferguson M, et al. Leveraging Transcriptomics Data for Genomic Prediction Models in Cassava [Internet]. bioRxiv. 2017. p. 208181. doi:10.1101/208181

63. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, et al. The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol. 2005;3: e196. doi:10.1371/journal.pbio.0030196

64. Binder S. Branched-Chain Amino Acid Metabolism in Arabidopsis thaliana. Arabidopsis Book. 2010;8: e0137. doi:10.1199/tab.0137

65. Tzin V, Galili G. New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. Mol Plant. 2010;3: 956–972. doi:10.1093/mp/ssq048

66. Maeda H, Dudareva N. The shikimate pathway and aromatic amino Acid biosynthesis in plants. Annu Rev Plant Biol. 2012;63: 73–105. doi:10.1146/annurev-arplant-042811-105439

67. Ogo Y, Mori T, Nakabayashi R, Saito K, Takaiwa F. Transgenic rice seed expressing flavonoid biosynthetic genes accumulate glycosylated and/or acylated flavonoids in protein bodies. J Exp Bot. 2016;67: 95–106. doi:10.1093/jxb/erv429

68. Mikkelsen MD, Naur P, Halkier BA. Arabidopsis mutants in the C--S lyase of glucosinolate biosynthesis establish a critical role for indole-3-acetaldoxime in auxin homeostasis. Plant J. 2004;37: 770–777. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2004.02002.x

69. Kliebenstein DJ, D'Auria JC, Behere AS, Kim JH, Gunderson KL, Breen JN, et al. Characterization of seed-specific benzoyloxyglucosinolate mutations in Arabidopsis thaliana. Plant J. 2007;51: 1062–1076. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2007.03205.x

70. Xiang X, Wu Y, Planta J, Messing J, Leustek T. Overexpression of serine acetyltransferase

in maize leaves increases seed-specific methionine-rich zeins. Plant Biotechnol J. 2018;16: 1057–1067. doi:10.1111/pbi.12851

71. Forde BG, Lea PJ. Glutamate in plants: metabolism, regulation, and signalling. J Exp Bot. 2007;58: 2339–2358. doi:10.1093/jxb/erm121

72. Ma H, Wang S. Histidine Regulates Seed Oil Deposition through Abscisic Acid Biosynthesis and β-Oxidation. Plant Physiol. 2016;172: 848–857. doi:10.1104/pp.16.00950

73. Ingle RA. Histidine biosynthesis. Arabidopsis Book. 2011;9: e0141. doi:10.1199/tab.0141

74. Stepansky A, Leustek T. Histidine biosynthesis in plants. Amino Acids. 2006;30: 127–142. doi:10.1007/s00726-005-0247-0

75. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The scale of population structure in Arabidopsis thaliana. PLoS Genet. 2010;6: e1000843. doi:10.1371/journal.pgen.1000843

76. Sauer U, Lasko DR, Fiaux J, Hochuli M, Glaser R, Szyperski T, et al. Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. J Bacteriol. 1999;181: 6679–6688. Available: https://www.ncbi.nlm.nih.gov/pubmed/10542169

77. Weckwerth W, Loureiro ME, Wenzel K, Fiehn O. Differential metabolic networks unravel the effects of silent plant phenotypes. Proc Natl Acad Sci U S A. 2004;101: 7809–7814. doi:10.1073/pnas.0303415101

78. Kutner MH, Nachtsheim CJ, Dr. JN. Applied Linear Regression Models- 4th Edition with Student CD (McGraw Hill/Irwin Series: Operations and Decision Sciences) [Internet]. 4 edition. McGraw-Hill Education; 2004. Available: https://www.amazon.com/Applied-Linear-Regression-Models-Student/dp/0073014664

79. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc. 1964; Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1964.tb00553.x

80. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465: 627–631. doi:10.1038/nature08800

81. Lipka AE, Lu F, Cherney JH, Buckler ES, Casler MD, Costich DE. Accelerating the switchgrass (Panicum virgatum L.) breeding cycle using genomic selection approaches. PLoS One. 2014;9:e112227.

82. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, et al. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 2004;37: 914–939. Available: https://www.ncbi.nlm.nih.gov/pubmed/14996223

83. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis. 2015;53: 474–485. doi:10.1002/dvg.22877

84. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21: 3439–3440. doi:10.1093/bioinformatics/bti525

85. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4. doi:10.1038/nprot.2009.97

86. R Core Team. R: A language and environment for statistical computing [Internet]. R Foundation for Statistical Computing, Vienna, Austria; 2016. Available: http://www.R-project.org/

87. Zhan S, Horrocks J, Lukens LN. Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. Plant J. 2006;45: 347–357. doi:10.1111/j.1365-313X.2005.02619.x

88. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012;91: 1011–1021.

doi:10.1016/j.ajhg.2012.10.010

89. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91: 4414–4423. doi:10.3168/jds.2007-0980

90. Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat Sci. 2009;24: 451–471. doi:10.1214/09-STS307

91. Whittaker JC, Thompson R, Denham MC. Marker-assisted selection using ridge regression. Genet Res. 2000;75: 249–252. doi:10.1017/s0016672399004462

92. Speed D, Cai N, UCLEB Consortium, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017;49: 986–992. doi:10.1038/ng.3865

93. Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, et al. 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. 2012;192: 715–728. doi:10.1534/genetics.112.141473

94. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57: 289–300. Available: http://www.jstor.org/stable/2346101