OXFORD GENETICS

GENETICS, 2021, iyaa043
DOI: 10.1093/genetics/iyaa043
Advance Access Publication Date: 3 February 2021
Genomic Prediction

Translating insights from the seed metabolome into improved prediction for lipid-composition traits in oat (*Avena sativa* L.)

Malachy T. Campbell [1,*], Haixiao Hu [1], Trevor H. Yeats [1], Melanie Caffe-Treml,[2] Lucía Gutiérrez,[3] Kevin P. Smith [4], Mark E. Sorrells,[1] Michael A. Gore [1], and Jean-Luc Jannink[1,5]

[1]Plant Breeding & Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA
[2]Department of Agronomy, Horticulture & Plant Science, South Dakota State University, Brookings, SD 57007, USA
[3]Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA
[4]Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA
[5]R.W. Holley Center for Agriculture & Health US Department of Agriculture, Agricultural Research Service, Ithaca, NY 14853, USA

*Corresponding author: Plant Breeding & Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA. campbell.malachy@gmail.com

Abstract

Oat (*Avena sativa* L.) seed is a rich resource of beneficial lipids, soluble fiber, protein, and antioxidants, and is considered a healthful food for humans. Little is known regarding the genetic controllers of variation for these compounds in oat seed. We characterized natural variation in the mature seed metabolome using untargeted metabolomics on 367 diverse lines and leveraged this information to improve prediction for seed quality traits. We used a latent factor approach to define unobserved variables that may drive covariance among metabolites. One hundred latent factors were identified, of which 21% were enriched for compounds associated with lipid metabolism. Through a combination of whole-genome regression and association mapping, we show that latent factors that generate covariance for many metabolites tend to have a complex genetic architecture. Nonetheless, we recovered significant associations for 23% of the latent factors. These associations were used to inform a multi-kernel genomic prediction model, which was used to predict seed lipid and protein traits in two independent studies. Predictions for 8 of the 12 traits were significantly improved compared to genomic best linear unbiased prediction when this prediction model was informed using associations from lipid-enriched factors. This study provides new insights into variation in the oat seed metabolome and provides genomic resources for breeders to improve selection for health-promoting seed quality traits. More broadly, we outline an approach to distill high-dimensional "omics" data to a set of biologically meaningful variables and translate inferences on these data into improved breeding decisions.

Keywords: genomic prediction; factor analysis; GWAS; metabolomics; GenPred; shared data resource

Introduction

Oat (*Avena sativa* L.) is cultivated throughout the temperate regions of the world for both human and animal consumption (Berzonsky and Ohm, 2000; Zhou *et al.* 2019). The oat seed contains a diverse array of compounds that are beneficial for human health and nutrition (Gulvady *et al.* 2013). It is widely considered a healthy food due to its high-soluble fiber content, which is unique among major cereals and has been shown to improve cardiovascular health, as well as help manage blood glucose levels (Gulvady *et al.* 2013; Kale *et al.* 2013). Oat is also a good source of protein (12.4–24.5% of seed weight), oil (3–11%), and a rich source of vitamins and minerals (Frey and Holland 1999; Gulvady *et al.* 2013). The oils found in the oat seed are primarily triglycerides, with palmitic, oleic, and linoleic acids being the primary fatty acids (Youngs 1978). Due to many of these qualities, oat has been used more recently to produce nondairy milk and yogurt products. In addition to the benefits from direct consumption, colloidal oatmeal and oat extracts have been used extensively as a topical medicine to treat skin dermatitis and reduce inflammation (Kurtz and Wallo 2007; Cerio *et al.* 2010). These benefits have been attributed to avenanthramides, flavonoids, tocopherols, polysaccharides, and lipids. Thus, the oat seed is a rich source of diverse compounds that have multifaceted effects on human health. To improve specific biochemical properties of oat, breeders must be provided with a suite of tools that allow these compounds to be quantified accurately at low cost and genomic resources that improve selection for specific seed qualities.

Advances in biochemistry have provided the research community with a breadth of tools to query the metabolome and quantify known and unknown compounds (Dunn and Ellis 2005). Untargeted metabolomics can quantify 100–1000s of metabolites in a sample, thus health-promoting and quality-related metabolites, and their intermediate or related compounds can be assessed with relative ease (Dunn *et al.* 2013; Christ *et al.* 2018).

Received: October 29, 2020. **Accepted:** December 08, 2020

These high-dimensional data can be leveraged to address basic biological questions regarding biochemical pathways that are represented in the data, as well as assess natural variation for these pathways. The effectiveness of these methods to characterize natural variation in the metabolome has been highlighted by several studies (Chan *et al.* 2010; Caspi *et al.* 2014; Matsuda *et al.* 2015; Slenter *et al.* 2018; Wu *et al.* 2018). Moreover, these data have been used as predictors, often alongside genomic data, to improve prediction for complex traits (Riedelsheimer *et al.* 2012; Guo *et al.* 2016; Xu *et al.* 2016).

Parsing these data to understand the biology of the seed metabolome can be challenging. Numerous databases are available that describe primary and secondary metabolic pathways, and are curated using information both across and within species (Kanehisa, 2002; Wishart *et al.* 2020). Metabolites can be mapped to these pathways to determine which pathways and their products are enriched in a given set of samples. While these approaches provide greater confidence over unsupervised, data-driven approaches, in many cases only a fraction of the compounds quantified via untargeted metabolomics can be mapped to these pathways (Schrimpe-Rutledge *et al.* 2016; Cui *et al.* 2018). This is especially problematic for under-characterized species or pathways. Unsupervised, data-driven approaches provide an attractive alternative that utilizes the data more completely. These approaches include coexpression-based analyses and factor analytic models. While coexpression-based analyses have been used extensively to characterize high-dimensional "omics" data, these approaches often require users to select several parameters that influence outcomes and may limit reproducibility (Langfelder and Horvath, 2008; DiLeo *et al.* 2011). Factor analytic models use a linear model to identify groups of strongly correlated metabolites. The underlying rationale for these approaches is that covariance among metabolites is driven by some unobserved (*i.e.*, latent) underlying variable(s). With this approach, the matrix of metabolites is decomposed into a lower-dimensional linear combination of factor loadings, which describe how each latent factor contributes to each compound, and a set of factor scores that ascribe a phenotypic value for all individuals for a given latent factor. Thus, these frameworks have advantages from both biological and statistical perspectives. While in some respects factor analytic models achieve the same goal as others, such as principal component analysis (PCA)—providing a reduced rank representation of the data—the defining feature of factor analytic models is that latent factors are constructed to preserve correlation among groups of related metabolites. In PCA, new constructs are defined that preserve variance in the observed variables. Constructs from factor analytic models can provide insight into biological processes driving covariation between phenotypes. Moreover, the lower-dimensional set of factor-scores can be treated as any other phenotype and will reduce the multiple testing burden often associated with high-dimensional "omics" datasets.

Improving health promoting or quality-related compounds requires decomposing phenotypic variation within the metabolome into genetic and nongenetic components, and utilizing these outcomes to inform selection decisions for quality-related phenotypes. Conventional linkage analysis or association mapping approaches have proven to be powerful approaches to identify genetic variants associated with variation in the metabolome (Rowe *et al.* 2008; Chan *et al.* 2010; Eckert *et al.* 2012; Wen *et al.* 2014; Matsuda *et al.* 2015; Xu *et al.* 2017). However, a much greater challenge is to translate genetic signal for health-promoting compounds and related metabolites to improve prediction and selection of new crop germplasm.

A number of studies have extended the conventional frameworks used for genomic prediction to accommodate prior biological information regarding genetic marker effects (Speed and Balding 2014; Edwards *et al.* 2016; MacLeod *et al.* 2016; Turner-Hissong *et al.* 2019). Although these approaches differ in how these data are treated, the motivation is similar for all—specifically, effects for variants that are more likely to be causative should be drawn from a different distribution than those lacking evidence for causality. Thus, prediction should be improved when effect sizes differ between genetic marker classes. For instance, the approaches described by Speed and Balding (2014) and Edwards *et al.* (2016) are essentially an extension of the genomic best linear unbiased prediction (gBLUP) framework, in which genomic markers are partitioned and are used to construct separate genomic relationship matrices for each random genetic effect. The framework described by MacLeod *et al.* (2016) extends the Bayesian prediction framework, BayesR, and uses biological information to partition markers into classes (Erbe *et al.* 2012). Marker effects, rather than genomic values, are sampled from each distribution. In the context of the current study, if we know what metabolites are related to quality traits and have identified variants associated with these metabolites, genomic markers can be partitioned to define biologically informed marker-sets that should be enriched for causal loci and improve prediction of genomic values.

We characterized the seed metabolomes of 375 diverse oat lines and sought to identify loci that potentially influence (co)-variation among many metabolites. Specifically, we sought to answer: (1) What pathways or metabolite classes are enriched in the seed metabolome? (2) What are the genetic controllers of the metabolome? and (3) Can these data be leveraged to improve genomic prediction for seed quality traits? To this end, we assayed the seed metabolome using untargeted LC-MS and GC-MS and used the empirical factor analysis approach described by Wang and Stephens (2018) to identify latent factors that generate covariance among many metabolites. We performed GWAS using this reduced set of latent phenotypes, and used these outcomes to inform a multi-kernel genomic prediction model for prediction of seed quality traits in two independent studies. We extract meaningful basic biological insights from "omics" data with limited annotations, and translate these outcomes to improve prediction for agriculturally important traits. This study provides a necessary foundation to characterize the oat seed metabolome and develop novel genomic resources for oat breeders to improve seed qualities.

Materials and methods

Plant materials and growth conditions

The oat diversity panel consists of 375 accessions derived from breeding programs in North America and Europe. In 2018, the diversity panel was grown in an augmented block field design in Ithaca, NY. The design consisted of 368 unreplicated entries allocated randomly to 18 blocks with 21–23 plots per block. One primary check, "Corral," was included in each of the blocks, while one of six secondary checks were randomly allocated to each block. These secondary checks were replicated four times, while the primary check was replicated 19 times (one block had two "Corral" plots). Three-hundred thirty five lines with genotypic data were used for downstream genetic analyses on latent factors.

Latent factor analysis

A description of sample preparation, metabolite quantification, and data preprocessing is provided in the Supplementary Materials. Latent factor analysis was performed using deregressed best linear unbiased predictions for 367 entries (one entry was deemed an outlier and was removed from subsequent analyses) and 1668 metabolites. Latent factor analysis seeks to identify a set of k latent factors that give rise to the observed covariance among a set p of empirical variables. This relationship is given by:

$$\mathbf{Y} = \mathbf{F\Gamma} + \mathbf{s}, \tag{1}$$

where $\mathbf{Y}$ is a centered and standardized $n \times p$ matrix of observations for p metabolites and n individuals; $\mathbf{F}$ is an $n \times k$ matrix of factor scores; $\mathbf{\Gamma}$ is a $k \times p$ matrix of loadings; and $\mathbf{s}$ is an $n \times p$ matrix of specific effects. The (co)variance matrix $\mathbf{V}$ of observations $\mathbf{Y}$ is decomposed into common covariance and specific covariance:

$$\mathbf{V} = \mathbf{\Gamma}'\mathbf{\Gamma} + \mathbf{\Psi}. \tag{2}$$

All matrices are defined as above, and $\mathbf{\Psi}$ is a $p \times p$ diagonal matrix of specific variances.

A recent framework described by Wang and Stephens (2018) uses an empirical Bayes approach to learn appropriate priors from the data given a family of densities. This approach, Empirical Bayes Matrix Factorization (EBMF), can tailor the sparsity for factor loadings and scores based on what best fits the data and was implemented using the flashr package in R (https://github.com/stephenslab/flashr/tree/master/R). Three classes of models were fit that differed in families of densities used to fit the data: Laplace, point-normal, and adaptive-shrinkage. A combination of the "Greedy" search algorithm and backfitting was used to define the model.

We evaluated the classes of models for goodness-of-fit using percent variance explained (PVE) by the common factors, as well as predictive ability using threefold orthogonal cross validation (3-OCV) (Owen and Wang 2016). PVE was defined as:

$$PVE = \frac{tr(\mathbf{\Gamma}'\mathbf{\Gamma})}{tr(\mathbf{\Gamma}'\mathbf{\Gamma} + \mathbf{\Psi}) \times 100} \tag{3}$$

with tr indicating trace of the given matrix and all other matrices defined as above. 3-OCV is similar to classical CV, but ensures that no rows and columns of the testing data ($\mathbf{Y}_{test}$) have all missing data. The model above was fitted for the training set data and predicted values for the testing set were calculated via $\hat{\mathbf{Y}}_{test} = \mathbf{F}_{test}\mathbf{\Gamma}_{test}$. The accuracy of each model was evaluated using the root mean square error (RMSE) and the correlation between predicted and observed values for observations in the testing set for each fold. Ten independent resamplings were performed. The metrics were averaged over folds, and the "best" model was selected based on the results across the 10 repeats.

Enrichment analysis for latent factors

We used the ClassyFire taxonomic hierarchies for 562 metabolites to test for functional enrichment for each factor (Feunang et al. 2016). ClassyFire uses a hierarchy of five levels to describe chemical compounds. At each level we calculated the percentage of variance explained (PVE_{kc}) for factor k by functional class c. This is given below:

$$PVE_{kc} = \frac{tr(\lambda_{kc}\lambda_{kc}')}{tr(\lambda_k\lambda_k')}, \tag{4}$$

where λ_k is a vector of loadings for a given factor k, and λ_{kc} is a vector of loadings of factor k for compounds in class c. Our null hypothesis is that the variance captured by compounds in a given class will be equivalent to that explained by a random set of compounds of equal size to that class. To test this, we generated an empirical null distribution for each functional class and factor. For each class and factor, we picked a random set of compounds with a size equivalent to the class by sampling the loadings of 1668 metabolites without replacement and computed PVE. This process was repeated 1000 times for each combination of functional class and factor. For each class-factor combination, we compared observed PVE with the empirical null distribution for that given combination and calculated P-values. q-values were calculated across all factors and classes following Storey (2002). Functional classes with fewer than five compounds were excluded from analyses to ensure that results were not biased to small classes with one or two compounds with very high loadings.

Assessing the genetic architecture of latent factors

Genome-wide association study

To identify loci associated with latent factors, the following linear mixed model was fit to factor scores for each latent factor (k):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{w_i}a_i + \mathbf{Zu} + \mathbf{e}, \tag{5}$$

where $\mathbf{y}$ is a vector of factor scores; $\mathbf{X}$ is a matrix of the first two PCs and $\mathbf{b}$ is the corresponding vector of effects; $\mathbf{w_i}$ is a vector of allele dosages for marker i and a_i is the corresponding marker effect; and $\mathbf{u}$ is a vector of polygenic effects. The first two PCs explained about 13% of the genomic relatedness among lines. We assume $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where $\mathbf{G}$ is a genomic relationship matrix calculated following the second definition provided by VanRaden (2008). These models were fitted using the rrBLUP package in R (Endelman, 2011). GWAS was performed using 62,049 SNP markers with a minor allele frequency >0.05 and 335 individuals with marker data and factor scores. Marker physical positions are based on the *Avena sativa*—OT3098 v1, PepsiCo genome assembly (https://wheat.pw.usda.gov/GG3/graingenes_downloads/oat-ot3098-pepsico).

We used the approach described by Li and Ji (2005) to account for multiple tests performed both within and across factors. We computed the number of effective tests (M_{eff}) by performing eigenvalue decomposition on the correlation matrix for 62,049 markers. This provides an estimate of the number of tests performed within each factor. Next, we multiplied this value by the total number of factors. The test criteria was then adjusted using M_{eff} with the Sidak correction below (Šidák, 1967).

$$\alpha_p = 1 - (1 - \alpha_e)^{1/(M_{eff} \times 100)} \tag{6}$$

This provided a genome-wide significance (α_p) value of 2.57×10^{-7} at $\alpha_e = 0.1$ with $M_{eff} = 4,097$. The proportion of variance explained by all significant GWAS hits for each factor (R^2_{GWAS}) was obtained by comparing two models: the full model included all significant markers and the first 10 principal components of the genomic relationship matrix, and the reduced model included only the first 10 principal components. All terms were considered

fixed. R^2_{GWAS} was calculated as the difference in the residual sum of squares of the reduced and full models.

Estimating polygenicity with Bayes Cπ

To estimate polygenicity of each factor, we used Bayes Cπ (Habier *et al.* 2011). Bayes Cπ is a Bayesian whole-genome regression approach that can be used to estimate the proportion of markers with a nonzero effect on the phenotype. Bayes Cπ assumes that marker effects are drawn from a mixture distribution. Effects drawn from a distribution with a point mass at 0 with a probability π and a Gaussian distribution with probability $(1-\pi)$. The linear model is:

$$\begin{aligned} y &= \mu + \sum_{t=1}^{T} w_t a_t + e \\ a_t|\pi,\sigma_t^2 &= \begin{cases} 0 \text{ with prob. } \pi \\ \sim N(0,\sigma_t^2) \text{ with prob. } (1-\pi) \end{cases} \end{aligned} \tag{7}$$

w_t is a vector of marker genotypes for marker t and a_t is the corresponding effect. The above model was fitted using the JWAS package in Julia using factor scores and 62,049 markers (Cheng *et al.* 2018). We used 200,000 iterations and discarded the first 100,000 iterations. Posterior means of $1-\pi$ were used as estimates of polygenicity.

Genomic prediction of seed quality traits

Two studies were used to determine whether associations from factor score-based GWAS could improve genomic prediction accuracies. The first consisted of fatty acid measurements for 500 lines, of which 338 had corresponding genotypic data consisting of 61,900 markers. These lines were evaluated at two locations in New York in 2014 (Carlson *et al.* 2019). The second consisted of six trials that evaluated protein and lipid content using near-infrared spectroscopy for 210 lines, of which 12 overlapped with the lines used for factor analysis. For this study, 58,293 markers were used for prediction. Supplementary Table S2 lists the trials used for genomic prediction and links to access these data.

A multi-kernel BLUP model was used to predict seed phenotypes across trials. Additive genetic effects were predicted using two kernels. The first is computed using markers that were identified through factor score-based GWAS and is referred to as the biologically informed kernel, while the second was computed using all other markers. This model is given by:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z_u u_{in}} + \mathbf{Z_u u_{out}} + \mathbf{Z_e s} + \mathbf{e}, \tag{8}$$

where $\mathbf{y}$ is a vector of phenotypes; $\mathbf{Z_u}$ is an $n \times q$ incidence matrix that assigns the q genomic values to n observations; $\mathbf{u_{in}}$ and $\mathbf{u_{out}}$ are genomic values predicted from biologically informed or noninformed kernels, respectively; $\mathbf{Z_e}$ is an $n \times e$ incidence matrix that assigns observations to trials and $\mathbf{s}$ are the corresponding effects. Moreover, we assume $\mathbf{u_{in}} \sim N(0, \sigma^2_{u_{in}}\mathbf{K_{in}})$, $\mathbf{u_{out}} \sim N(0, \sigma^2_{u_{out}}\mathbf{K_{out}})$, and $\mathbf{s} \sim N(0, \sigma^2_s \mathbf{Z'_e Z_e})$. Where $\mathbf{K_{in}}$ and $\mathbf{K_{out}}$ are biologically informed and noninformed kernels genomic relationship matrices, respectively, and are computed according to VanRaden (2008). We considered two marker sets to compute these matrices: markers associated with any latent factor, and markers that were associated with latent factors showing enrichment for lipid and lipid-like molecules at the superclass level ($q < 0.05$). Markers that were in weak linkage disequilibrium (LD) ($r^2 > 0.25$) with GWAS hits were included in the biologically informed kernel. LD was computed separately for each study.

The multi-kernel approaches were compared to Genomic BLUP (gBLUP) and BayesB (Meuwissen *et al.* 2001). The gBLUP model is similar to the multi-kernel model; however, the relationship matrix was constructed using all available markers for each study. All models were fit using the BGLR package in R with 20,000 iterations, of which the first 5000 were discarded (Perez and de los Campos, 2014). The model for BayesB is given by:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z_u}\sum_{t=1}^{T} w_t a_t + \mathbf{Z_e s} + \mathbf{e} \quad a_t|\pi,\sigma^2 = \begin{cases} 0 \text{ with prob. } \pi \\ \sim N(0,\sigma^2) \text{ with prob. } (1-\pi) \end{cases}. \tag{9}$$

All matrices and vectors are described above. BayesB assumes marker effects are drawn from a scaled-t mixture distribution with a probability of $(1-\pi)$ and a point mass at 0 with a probability π. BGLR places a beta prior on π.

Prediction accuracy was assessed using fivefold cross validation with 50 resampling runs, and was computed using Pearson's correlation between observed phenotypes and predicted genomic values for accessions in the testing set. Genomic values for the multi-kernel approach were computed as the sum of breeding values from each random genetic effect. Correlation coefficients were averaged across folds.

Data availability

Metabolomic data are provided via Cyverse and can be accessed using the following url https://de.cyverse.org/dl/d/614C0960-3D37-4C35-9E48-DA5CE9AB473C/DPmet.zip. All R code used for analyses is provided as Rmarkdown files and can be accessed via https://github.com/malachycampbell/OatLatentFactor.

Supplemental materials are provided at https://doi.org/10.25386/genetics.13409696. Supplemental File S1 provides deregressed BLUPs, metabolite annotations, and factor loadings for the 1668 metabolites. Supplementary File S2 lists the metabolites showing significant differences between subpopulations. Supplementary File S3 provides summary statistics for GWAS and a listing of markers used to create the biologically informed kernels. The supplemental PDF document contains all supplemental figures and methods not described above.

Results

Metabolite differences across subpopulations are primarily generated by drift

To characterize the metabolome of mature oat seed, we generated untargeted metabolite data using two mass spectroscopy (MS) pipelines (gas chromotography MS, GC-MS and liquid chromotography MS, LC-MS) for 367 diverse accessions (Supplemental File S1). The diversity panel consisted of 367 accessions that could be partitioned into six distinct genetic clusters using a k-means clustering approach (Figure 1, A and B; Supplementary Figure S1). The degree of stratification within the population was minor. For instance, the first and second principal axes explained only 7.3% and 5.9% of the variance in genetic relationships, respectively (Figure 1, A and B). In total, we quantified 1668 metabolites (601 for GC-MS and 1067 for LC-MS) across the 367 accessions. PCA of the whole metabolome dataset did not reveal any apparent clustering among accessions, and evidence of stratification between genetically defined clusters was not visually apparent (Figure 1, C and D).

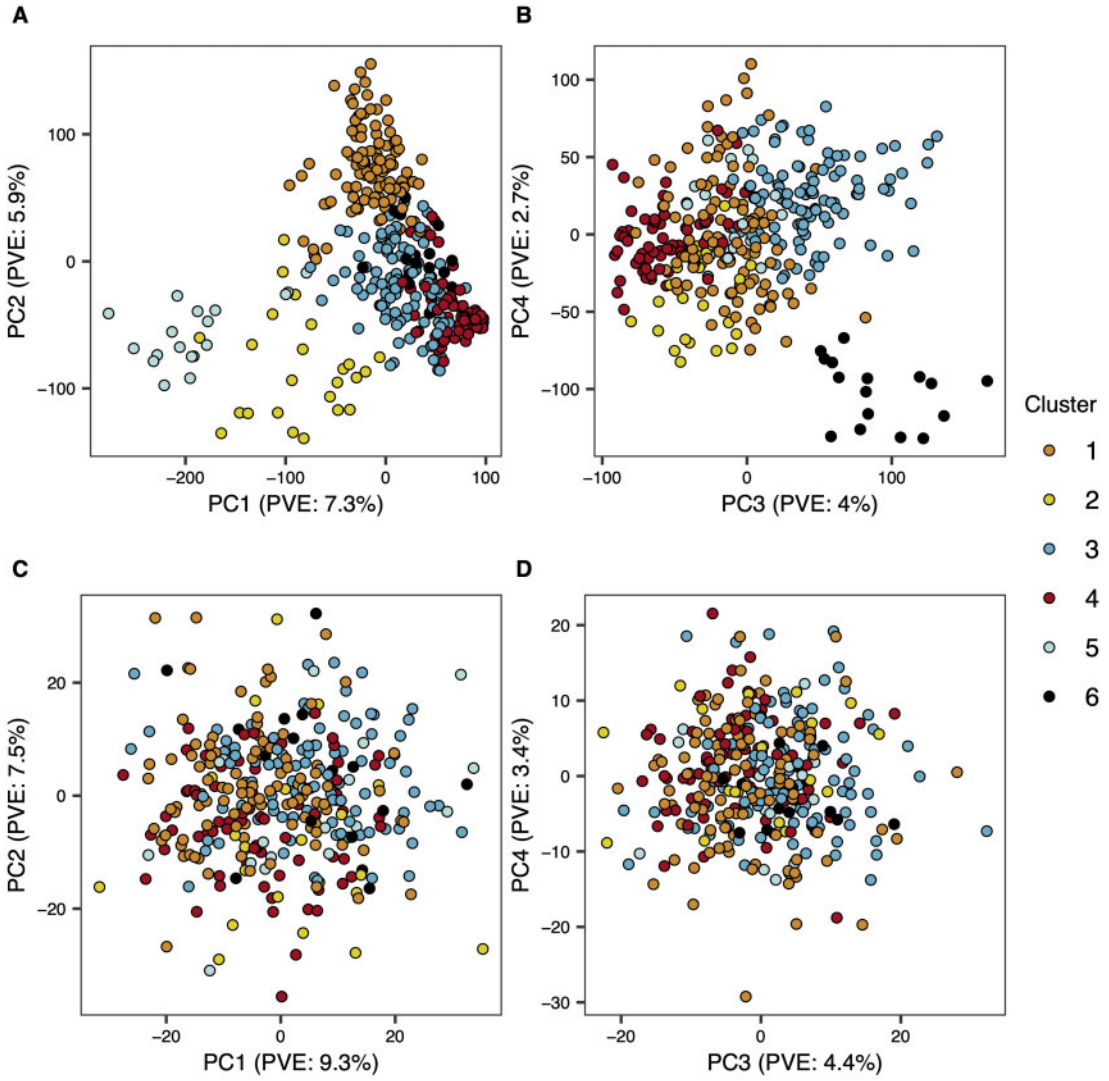

Figure 1 Principal component (PC) analysis of genotypic and metabolomic data. The first four PCs of gentoypic data are shown in panels (A and B), while the first four PCs of the metabolomic data are shown in panels (C and D). Subpopulations that were defined based on *k*-means clustering of SNP marker data are indicated by different colored points. PVE, percent variance explained.

To determine whether individual metabolites differed among clusters, we performed a one-way ANOVA for each of the 1668 metabolites (Supplemental File S2). Despite no strong differentiation of the metabolome between the six clusters, nearly 41% of the 1668 metabolites showed significant differences between one or more of the subpopulations (Benjamini–Hochberg adjusted *P*-value; $p_{BH} < 0.01$). We next elucidated whether these differences were due to selection or drift by examining P_{ST}, a measure of phenotypic divergence between clusters, and compared these values to the distribution of genetic divergence (*i.e.*, F_{st}) for all loci (Storz, 2002; Leinonen *et al.* 2013). This analysis revealed only 12 compounds with P_{st} values that were greater than 80% of the F_{ST} values, indicating that the majority of compounds differing between clusters diverged due to drift or weak selection. Only four of these compounds have annotations and were described as a putative steroidal glycosides, terpene glycoside, triterpenoid, and 1-benzopyran. These results suggest that the divergent metabolites are largely due to drift rather than selection.

Latent factor model selection

Given that only a fraction of the metabolites quantified in our population were annotated (562 compounds), we leveraged the correlation between annotated and unannotated metabolites to infer biological processes in the oat seed with the rationale that compounds participating in a related biological process will be correlated. We used an unsupervised learning approach, EBMF, that distills the covariance among the 1668 metabolites into a lower dimensional set of unobserved constructs that may cause this covariance (Wang and Stephens 2018). The observed phenotypes are approximated using a linear combination of factor loading and factor scores, and EBMF estimates priors for these terms from the data (Wang and Stephens 2018).

Three latent factor models that differed in the family of prior distributions (Laplace, point normal, and adaptive shrinkage) for factor loading and scores were evaluated, and the best model was selected based on the goodness-of-fit and predictive ability (Owen and Wang 2016) (Table 1; Supplementary Figure S2). The Laplace family of densities exhibited the lowest RMSE (0.970) and highest correlation between predicted and observed data ($r = 0.520$). This model revealed that the common covariance in the oat seed metabolome could be captured using 100 latent factors that collectively explained 58.8% of the total variance in the metabolite data. The PVE by individual factors ranged from 0.036% to 7.801%.

Table 1 Empirical Bayes matrix factorization model selection

EBNM Appr.	No. Fact.	LL	PVE	R^2_{adj}	$r_{(Y_{tst}, \hat{Y}_{tst})}$	RMSE
Ad. Shr.	102	−581716.3	59.41	0.438	0.322	1.451
Point Nor.	106	−583809.9	59.36	0.429	0.514	0.978
Laplace	100	−584317.2	58.82	0.434	0.520	0.970

Each model was fit using degressed BLUPs for 1668 metabolites. Ad. Shr.: adaptive shrinkage family of densities described by Stephens (2016). Cross-validation (CV) was based on a threefold orthogonal CV described by Wang and Stephens (2018) and Owen and Wang (2016) with 10 independent resamplings. Point Nor.: point-normal family of densities which are a normal distribution with a point mass at zero; LL indicates log-likelihood; PVE: percent variance explained; R^2_{adj}: adjusted R^2; $r_{(Y_{tst}, \hat{Y}_{tst})}$ is the Pearson's correlation between predicted and observed values for observations in the testing set; RMSE: root mean square error.

Factor analysis identifies sets of compounds coordinated by biological processes

The covariance among metabolites may be due to underlying biochemical/biological process or could be caused by an unaccounted confounding factor (*i.e.*, population structure). We sought to partition latent factors into those consistent with the former possibility (a biological process) and those consistent with the latter (a confounding effect) (Bello *et al.* 2018). Since we showed that most population structuring of metabolites was caused by drift, we expect their coordination to be largely random, and therefore unrelated to their functional class. We assessed enrichment for functional classes within each factor, as well as the relationship between factors and population structure.

To assess biological enrichment, we determined whether the variance explained by a given metabolite functional class within a factor was significantly greater than might be expected by chance. We used ClassyFire hierarchical ontologies to calculate the percentage of variance explained by each functional class for each factor, and compared these values to an empirical null distribution to calculate *P*-values. The hierarchy consists of five levels: kingdom, superclass, subclass, and parent (Feunang *et al.* 2016). Of the 100 factors identified with the EBMF approach, 37 showed significant enrichment in one or more categories at the super-class level, while 40 and 36 factors showed significant enrichment at the class level and subclass levels, respectively ($q < 0.05$). Functional classes associated with lipids were most frequently enriched in our dataset (Figure 2, A and B), indicating that many factors may be capturing components of lipid metabolism. In addition to lipids, four factors showed significant enrichment for carbohydrates and carbohydrate conjugates, as well as amino acids. These results suggest that many latent factors are capturing meaningful biological processes that shape the seed metabolome, and can help shed light on the meaning of unannotated metabolites.

To address the possibility that latent factors were due to population structure, we examined the percent of variation explained by subpopulation. A linear model was fitted to each latent factor that included subpopulation assignment as a fixed effect. The PVE by subpopulation ranged from 0.03% to 29.8%, and subpopulation explained more than 20% of the variation for factors 7 and 12. Factor 7 did not show functional class enrichment but factor 12 was enriched across all hierarchies for lipid and lipid-like molecules—specifically steroidal glycosides ($q < 0.05$). Interestingly, the P_{st} for this factor (0.27) was higher than the top 80th percentile of F_{st} (0.23), suggesting that the differences between subpopulations for this factor may be due to selection rather than drift. The high frequency of enrichment for functional classes of metabolites, as well as the relatively small amount of variation that was attributed to subpopulations suggests that these constructs can provide biochemically meaningful insights into the seed metabolome.

Elucidating the origin of latent factors

We used a Bayesian whole genome regression approach, Bayes Cπ, to estimate variance components, as well as estimate the degree of polygenicity of each factor (Habier *et al.* 2011). Bayes Cπ assumes markers have a zero effect with probability π and a nonzero effect with probability $(1 - \pi)$. π is treated as an unknown and is estimated from the data. Thus, the magnitude of $(1 - \pi)$ can provide a metric to assess the polygenicity of the trait. Narrow-sense heritability estimates (h^2) ranged from 0.01 to 0.80, indicating that variation for many of the latent factors could be attributed to additive genetic effects (Supplementary Figure S4). Moreover, as indicated by the range of $(1 - \pi)$, this genetic variance is manifested in a wide range of architectures (Supplementary Figure S4).

The distribution of loading values for each latent factor varied—some factors showed dense loadings (*i.e.*, they generate covariance for many metabolites), while others showed sparse loadings (*i.e.*, generate covariance among few compounds). These loadings were sampled from a scale mixture distribution where nonzero loadings are sampled from a Laplace distribution with a probability of $(1 - \upsilon)$ and a point-mass at zero with a probability of υ. Given that latent factors with dense loadings will generate covariance for many metabolites, we hypothesized that these factors will likely have a complex genetic architecture. To test this, we performed a partial Spearman's correlation between polygenicity and the density of factor loadings while accounting for the heritability (h^2) of each factor. To further support this relationship, we performed GWAS on factor scores for each factor and estimated the partial correlation between the proportion of variance explained by significant GWAS associations ($p < 2.57 \times 10^{-7}$) and the density of factor loadings. These analyses revealed a significant positive correlation between $(1 - \upsilon)$ and $(1 - \pi)$ ($\rho = 0.35$; $p = 4.5 \times 10^{-4}$) and a negative relationship between proportion of variance explained by significant GWAS associations and $(1 - \upsilon)$ ($\rho = -0.42$; $p = 1.72 \times 10^{-5}$), indicating that factors that capture (co)variance among many metabolites tend to be controlled by many loci with small effects (Figure 3; Supplementary Figure S3). However, several exceptions to this relationship were observed. For instance, factors 4, 13, 17, and 25 exhibited low polygenicity and dense loading patterns (Table 2), indicating that these factors may be driven by loci with pleiotropic effects on the metabolome.

Biologically informed prediction of seed quality traits

Ultimately, the aim of this study was to translate insights from the metabolome into genetic resources that can be used by breeders to make broad changes to oat seed composition. We assume that loci with large effects on multiple metabolites will be a more valuable resource to oat breeders than loci that affect one or a few metabolites. The GWAS on factor scores identified 666 markers associated with 23 factors ($p < 2.57 \times 10^{-7}$; Supplementary File S3, Supplementary Figures S6–S28). A comparison of these results with associations from GWAS on individual metabolites is provided in Supplemental File S3. We assessed whether these associations could be leveraged to improve genomic prediction for seed quality traits in two independent studies. The first study quantified 10 fatty acids (FA) in

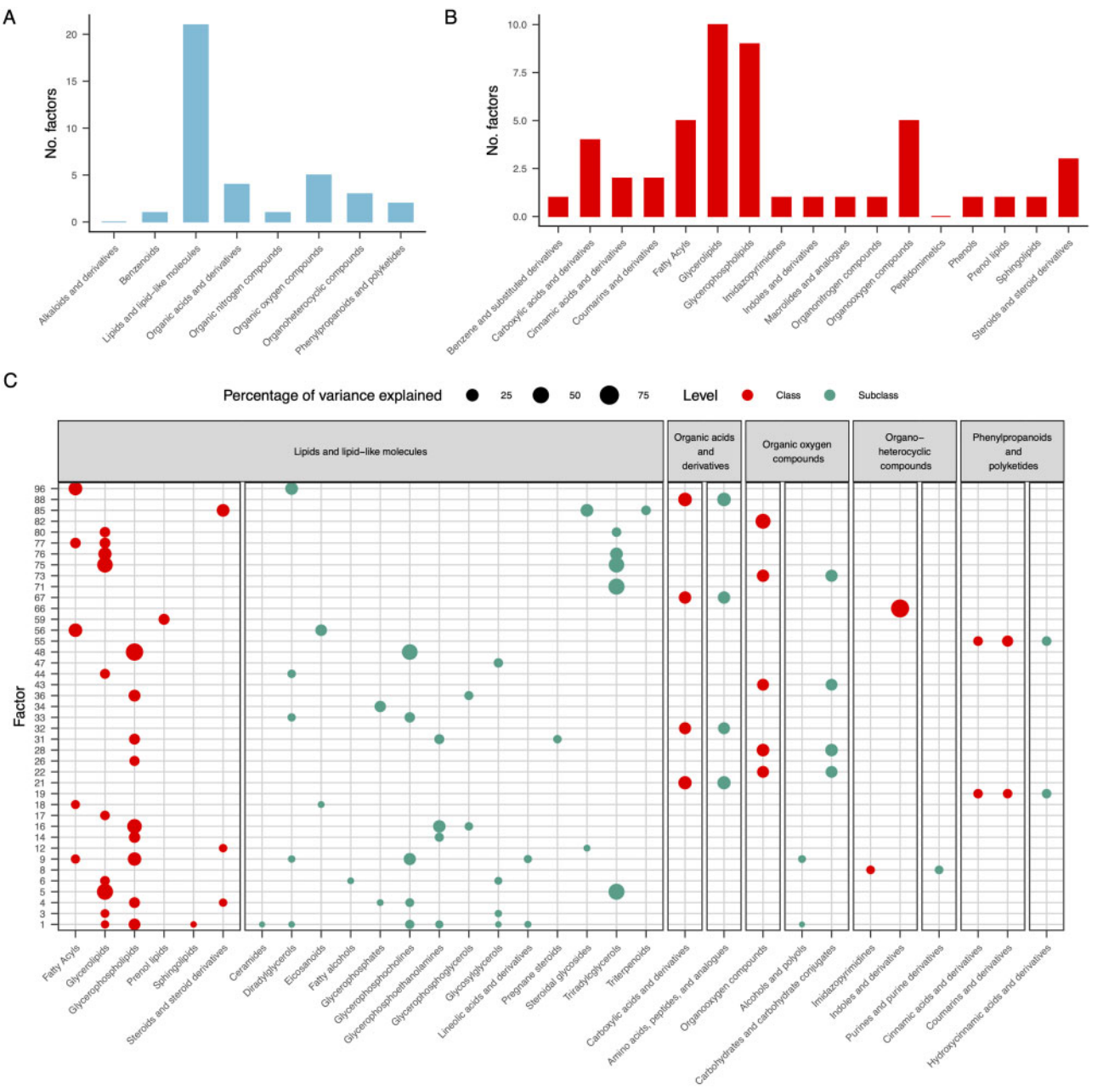

Figure 2 Functional enrichment among latent factors. Number of latent factors enriched (FDR < 0.05) for functional categories at the super-class level (A) and class level (B). Percentage of variance explained for each factor by a given functional category (C). Each point represents a functional class that was significantly enriched for one or more factors with the size of the point being proportional to the percentage of variance explained by that class for a given factor. Only factors and classes that showed significant enrichment ($q < 0.05$) at the super-class level are pictured. Colors differentiate between the class and subclass levels of the taxonomic hierarchy.

mature seed for 338 oat lines grown in two locations using targeted GC-MS. Of the 338 accessions evaluated, 330 overlapped with the panel used for factor analysis. The second study assayed seed lipid and protein content using near-infrared spectroscopy (NIRS) for 210 accessions from six trials with less than 6% of the lines (12 lines) overlapping with the panel used here for factor analysis. We compared three prediction frameworks to predict seed-quality phenotypes across trials: genomic BLUP (gBLUP), multi-kernel BLUP (MK-BLUP), and a Bayesian regression approach (BayesB) that better accommodates large effect QTL (Meuwissen *et al.* 2001).

The MK-BLUP framework uses two kernels to capture additive genetic effects, one of which is constructed from markers associated with latent factors and markers in LD with these. We refer to this as the "biologically informed" kernel. The second kernel is constructed from all other markers. We evaluated two biologically informed kernels: one that used markers associated with any latent factor (MK-all) and one that only used markers associated with factors enriched for "Lipid and lipid-like molecules" (factors 4, 17, and 34; MK-lip). Prediction accuracy was assessed using fivefold cross validation with 50 resampling runs, and the MK-BLUP models were deemed to significantly improve prediction if prediction accuracies for MK-BLUP were higher than gBLUP or BayesB in 90% of resampling runs.

In general, the MK-lip approach showed the highest prediction accuracies for most traits. MK-lip showed a significant improvement over gBLUP for seven of the 10 FA traits (Figure 4). The percent

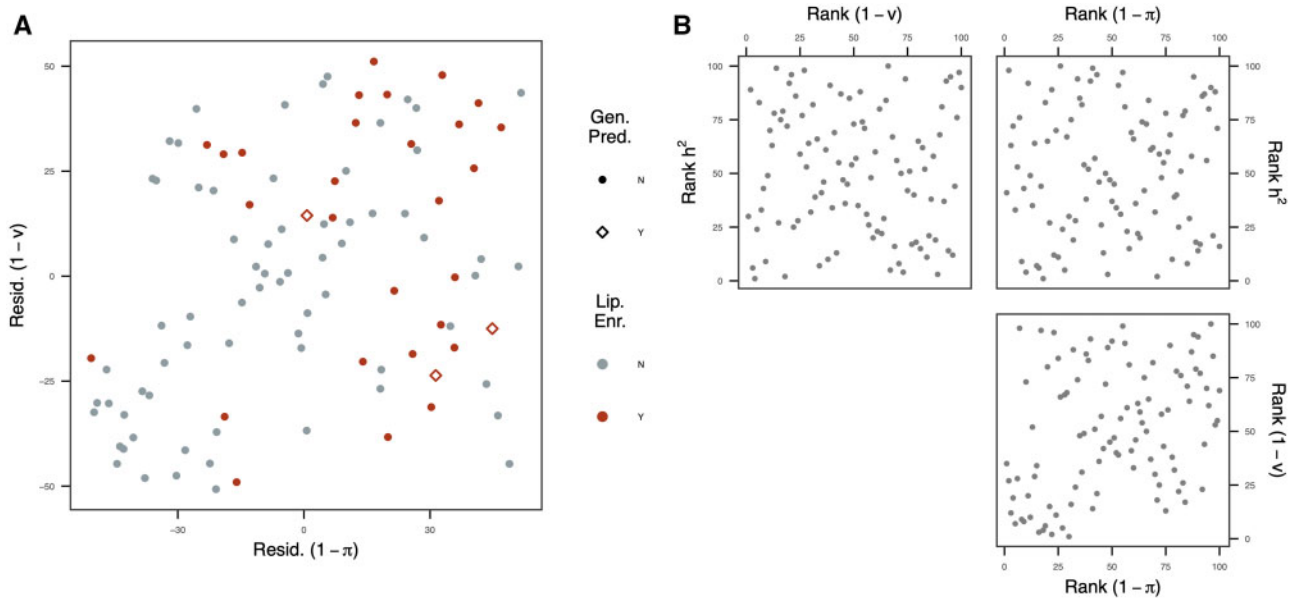

Figure 3 Relationships between polygenicity, density, and heritability. (A) Association between polygenicity ($1-\pi$) and density ranks ($1-\upsilon$) after accounting for heritability (h^2). Each variable was ranked from smallest to largest and the ranks for ($1-\pi$) and ($1-\upsilon$) were each regressed on ranks for h^2. The scatter plot depicts the relationship between the residuals (Resid.) for each of these models. Colored points indicate factors that were enriched for lipids (Lip. Enr.), and different shapes indicate whether the factor was used to inform the lipid-enriched kernel for genomic prediction (Gen. Pred.). (B) Pairwise relationships between the ranks for each variable.

Table 2 Factors capturing covariance between many metabolites with simple genetic architectures

Factor	$1-\pi$	$1-\upsilon$	R^2_{GWAS}
4	4.69×10^{-3}	0.621	0.08
13	7.80×10^{-4}	0.369	0.29
17	4.70×10^{-4}	0.413	0.19
25	5.54×10^{-4}	0.247	0.06

Polygenicity estimates were based on the posterior means of $1-\pi$ and the proportion of variance for captured by significant GWAS associations for each factor (R^2_{GWAS}), and the density of factor loadings are provided as $1-\upsilon$.

change in prediction accuracy for FA traits ranged from −0.57% to 23.10% (Figure 4B). Carlson *et al.* (2019) reported several large-effect QTL for FA traits. Whereas gBLUP provides equal shrinkage across markers, the MK-approach may shrink markers in each kernel differently. Thus, the improved prediction accuracy observed for MK-lip over gBLUP may be due to the reduced shrinkage of large-effect loci encoded by markers in the biologically informed kernel (Supplementary Figures S5, S7, and S10). To test this, we compared both multi-kernel approaches to BayesB. MK-lip significantly out-performed BayesB for two of the ten fatty acid traits [14:0 and 18:1([9); Supplementary Figure S28]. MK-all showed a significant improvement over gBLUP for 20:1, but did not show a significant improvement over BayesB for any trait. MK-lip out-performed both gBLUP and BayesB for total lipid content measured via NIRs (Figure 5; Supplementary Figure S29). On average MK-lip showed a 9.9% increase in prediction accuracy over gBLUP and a 9.5% increase over BayesB for total lipid content (Supplementary Figure S29). These results indicate that the genetic signal captured by latent factors can be leveraged to improve selection for seed compositional traits in oat.

Discussion

The oat seed harbors a rich array of biochemical compounds that are important for human health, and considerable variation for these compounds exist in oat germplasm (Peterson and Wood, 1997; Frey and Holland, 1999; Gulvady *et al.* 2013; Zhou *et al.* 2019). Accessing this variation is necessary to rapidly deliver oat varieties with beneficial nutritional profiles to the consumer. Advances in metabolic profiling over the past 20 years has provided a suite of tools to comprehensively assess these compounds, along with many others, in large populations and to elucidate their regulation (Keurentjes *et al.* 2006; Tohge and Fernie 2010). Despite these advances, significant challenges remain. Structural elucidation and metabolite identification remain a significant bottleneck in characterizing the metabolome using untargeted metabolomics (Dunn *et al.* 2013). Many of the publicly available databases do not adequately capture the rich diversity of metabolites that are produced in plant species (De Vos *et al.* 2007; Tohge and Fernie 2010). Therefore, approaches that uncover the relationships between metabolites, both known and unknown, may help shed light on the function of these compounds.

Despite being able to reliably detect the abundance of 1668 compounds in the current study, less than a third of these compounds were annotated. We used a latent factor approach that leverages the correlation between metabolites to help elucidate their function. Our rationale is that metabolites that participate in the same pathway should be correlated. Thus, by extracting the major correlation patterns in the observed variables we can begin to elucidate the biochemical pathways that shape the seed metabolome. Moreover, by studying the relationships among annotated metabolites, we can generate new hypotheses to understand the function of unannotated compounds.

Characterizing the metabolome using latent factors

Our enrichment approach helped shed light on the biochemical processes these latent factors might affect. For instance, we found significant enrichment for a range of processes associated with primary metabolism (amino acids, phospholipid

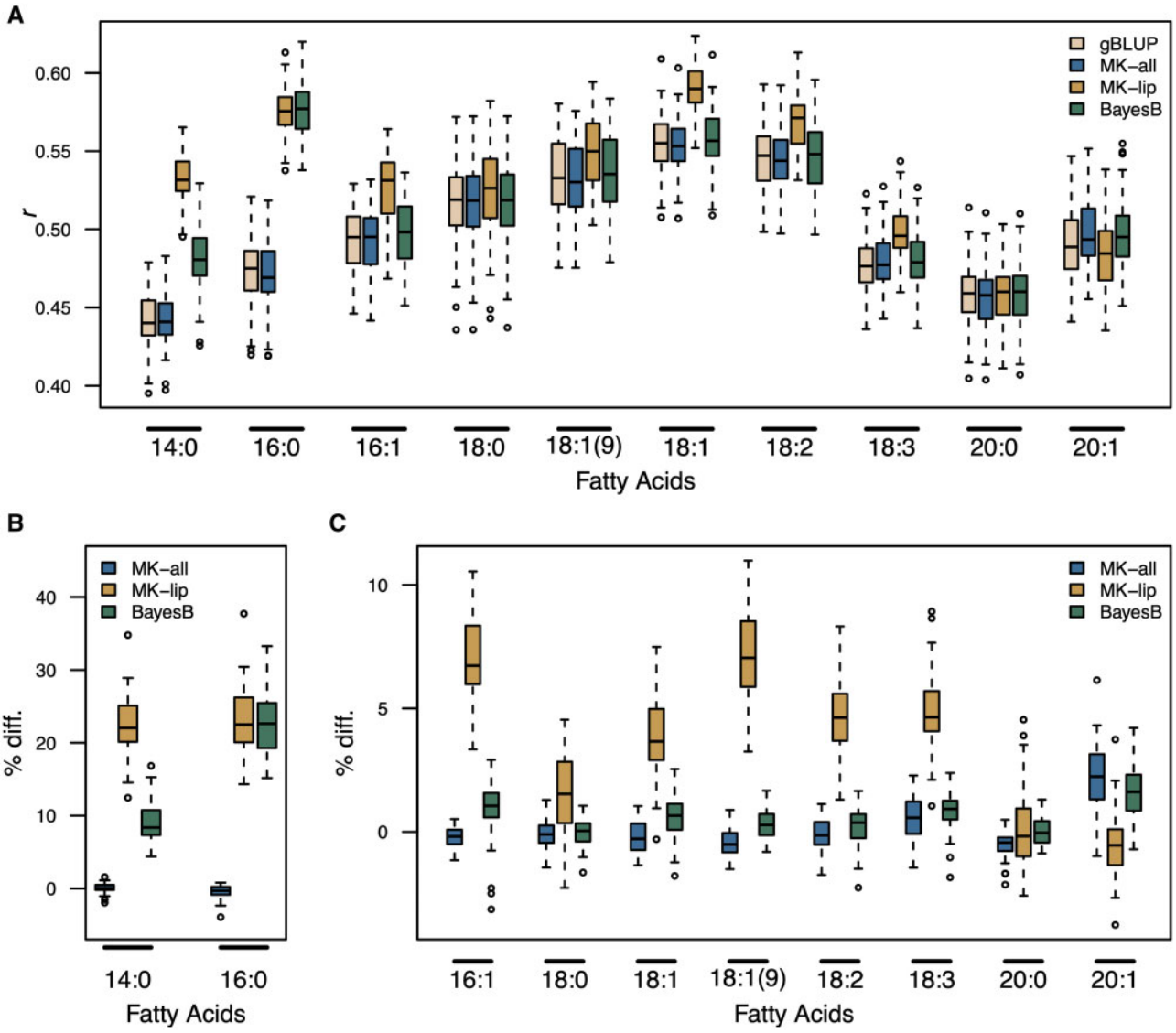

Figure 4 Genomic prediction for fatty acid compounds. Prediction accuracy was assessed using fivefold cross validation with 50 resampling runs. (A) The distribution of Pearson's correlation (*r*) coefficients between observed phenotypes and genetic values for each fatty acid compound. Panels (B and C) show the percent difference (% diff.) in prediction accuracy for the multi-kernel (MK) approach relative to genomic BLUP (gBLUP). The suffixes "-all" and "-lip" indicate models where the biologically informed kernel was constructed from markers associated with any latent factor or lipid-enriched factors, respectively. Three-hundred thirty lines used in this study were also used for factor analysis of metabolomic data.

metabolism) and secondary metabolism (coumarin and terpenoid metabolism). Since roughly 30% of the metabolites assayed had functional annotations, this enrichment approach may shed light on the function of unannotated metabolites. For instance, factor 4 showed significant enrichment for "lipid and lipid-like molecules." Although only 45 of the top 100 compounds with high loadings were annotated, the high correlation between these unknown compounds and lipid-like compounds suggests putative role in lipid metabolism. Although further analyses are necessary to elucidate the structure of these unknown metabolites, we show that enrichment provides a data-driven approach to generate hypotheses for these unannotated metabolites.

One overarching pattern observed across latent factors is the enrichment for compounds related to lipid metabolism. At the super-class level 21% of factors were significantly enriched for "lipid and lipid-like molecules," and these patterns were consistent at more specialized levels of lipid metabolism. Oat is unique among cereals in both the abundance and distribution of lipids within the seed (Price and Parsons 1975; Frey and Holland 1999; Gulvady *et al.* 2013). And with approximately 57% of the annotated metabolites in our data classified as lipid-like compounds, it is not surprising that categories associated with lipid metabolism were most frequently enriched. It is possible that other processes are prevalent in the metabolome and are reflected in the latent constructs, but remain undetected due to the annotations that were used for functional enrichment.

The ontologies used for functional enrichment are based on structural similarities between compounds, rather than pathway-based relationships. We expect compounds involved in the same pathway to be correlated, and since latent factors are defined by these correlations, they should in some sense be an abstraction of these pathways. Biochemical reactions often involve compounds with dissimilar structures, thus enrichment based on structural similarities may bias enrichment toward pathways composed of structurally similar metabolites (*e.g.*, lipid metabolism). While this enrichment approach may be imperfect,

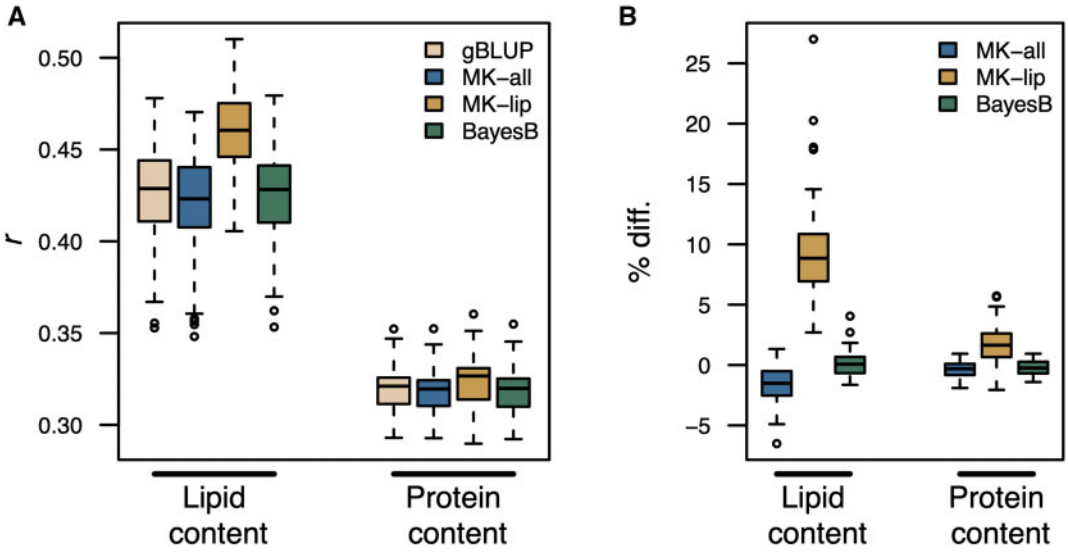

Figure 5 Genomic prediction for lipid and protein content measured via NIRS. Prediction accuracy was assessed using fivefold cross validation with 50 resampling runs. (A) The distribution of Pearson's correlation (r) coefficients between observed phenotypes and genetic values for each fatty acid compound. (B) The percent difference (% diff.) in prediction accuracy for the multi-kernel (MK) and BayesB approaches relative to genomic BLUP (gBLUP). The suffixes "-all" and "-lip" indicate models where the biologically informed kernel was constructed from markers associated with any latent factor or lipid-enriched factors, respectively. Three-hundred thirty lines used in this study were also used for factor analysis of metabolomic data.

other studies have used similar approaches to test for functional enrichment and have proven to be useful in other species (Barupal and Fiehn 2017; Fan *et al.* 2018; Marco-Ramell *et al.* 2018; Showalter *et al.* 2019). For instance, the ChemRich approach developed by Barupal and Fiehn (2017) uses the ClassyFire ontology to classify compounds into functional classes and tests for enrichment using a Kolmogorov–Smirnov test. Annotations that map metabolites to a pathway can provide additional evidence that these latent factors are indeed due to an underlying biochemical process; however, current resources that are available do not provide the breadth and resolution necessary to perform such analyses.

Understanding the origin of latent factors

Although it may seem reasonable to suggest that the observed covariance among metabolites is due to a biological cause that is manifested in the metabolome, making causal inferences from observational data is nontrivial due to the presence of confounding factors (Spirtes *et al.* 2000; Rosa and Valente, 2013; Bello *et al.* 2018). Given these data were collected on a structured population, it is expected that some of this covariance can be attributed to population structure, which can influence the construction of latent variables if not taken into account (Phillips *et al.* 2001). There are many ways to account for structure in the definition of latent factors, either by including the genomic relationship matrix, or some component(s) of it, in the factor analytic model or by regressing-out these effects prior to factor analysis. However, it is important to consider whether these steps are necessary. While such measures will control for confounding due to structure, they will also remove possibly meaningful biochemical relationships that are associated with structure. For instance, if a set of compounds participating in a common pathway happen to differ between subpopulations, correcting for structure may remove the latent factor that describes this process. We identified two latent factors, factors 7 and 12, that were associated with population structure. Enrichment analysis, as well as $P_{st} - F_{st}$, suggested that factor 12 may indeed describe a biological process (steroidal glycoside metabolism) that was affected by selection. This factor would likely be removed if structure were accounted for prior to factor analysis.

If subsequent genetic analysis are planned for latent factors, regressing-out structure may also remove meaningful genetic signal. Given the minor structure observed among accessions in the diversity panel and the importance of preserving genetic signal in the factor scores, we thought that measures to account for structure could be harmful to the study as a whole. Moreover, our downstream association mapping approaches accounted for population structure by using the first two PCs, as well as a kinship matrix based on allele dosages. In the event that some latent factors were defined based on kinship, we do not expect to recover any signal from association mapping with scores for these latent factors.

Due to the observational nature of these data, we should not place too much emphasis on causality in a purely biological sense when interpreting these latent factors. Rather it is important to consider the limitations of the study, interpret latent factors with caution, and view them as a means to generate testable hypotheses. The aims of our study were to (1) elucidate the major biochemical processes in the oat seed metabolome, and (2) to leverage these insights to improve selection for seed quality. Thus, hypotheses are generated in the former and are tested in the latter. If latent factors do not represent a causal effect then we should not see any improvement in predictions when inferences on these constructs are extended to new studies and/or populations.

Translating "omics" insights to crop improvement

Two independent studies were used to determine whether biological signal in the latent factors could be generalized to other populations and/or traits. The fatty acid dataset can be viewed as a resource to test whether the information learned by latent factors is reproducible, while the NIRS dataset provides a means to test whether this information is transmissible to related traits in new populations. We distinguish between these two because: (1) the majority of accessions included in the fatty acid dataset are

accessions that were used for the factor analysis metabolome study, while less than 6% of accessions are common between the factor analysis and the NIRS studies; (2) the fatty acid data was generated using targeted metabolomics, meaning there should be a high correspondence between the metabolites measured in the fatty acids study and those that were assayed for the factor analysis metabolome study Carlson et al. (2019).

We observed the greatest improvements in prediction accuracy among all traits for the biologically informed prediction model over gBLUP for these compounds when the kernel was constructed using associations for lipid-enriched factors. This improvement can be attributed to the ability of the multi-kernel approach to accommodate the large-effect QTL identified by Carlson et al. (2019), and/or the genetic signal associated with lipid metabolism encoded in the biologically informed kernel. MK-lip out-performed BayesB for two of the 10 fatty acid traits and generally showed higher prediction accuracies than BayesB for most FA traits. These results suggest that MK-lip both better models the genetic architecture of the FA traits, and captures relevant components of lipid metabolism. A comparison of the GWAS hits in Carlson et al. (2019) and those in our study showed little overlap, with two common associations identified for factor 13 and the tenth PC of fatty acid phenotypes in Carlson et al. (2019), and factor 17 and 14:0 in Carlson et al. (2019). Of these two factors, only factor 17 showed enrichment for "lipid and lipid like molecules" at only the super-class level. Enrichment for 1-acyl-sn-glycero-3-phosphocholines was the top-ranked category at the parental class ($q = 0.058$). Hydrolyzation of these compounds by phospholipase A1 yields a fatty acid. Although additional studies are necessary to elucidate the biochemical pathways associated with factor 17, these results provide an interesting link between 1-acyl-sn-glycero-3-phosphocholines catabolism and fatty acid abundances and the possibility of modifying 1-acyl-sn-glycero-3-phosphocholine metabolism to fine-tune fatty acid profiles in oat. Although it is difficult to connect loci associated with latent factors with changes in specific metabolites, our polygenicity analysis offers a more general explanation—specifically, that these loci may affect many metabolites. This is supported by GWAS on individual metabolite levels. Markers that were identified in both factor score-based GWAS and GWAS on metabolite tended to be associated with more than one metabolite, while marker associations that were identified only with GWAS on metabolites were predominately associated with one metabolite (Supplementary Figure S30).

The second study with NIRS-derived composition measurements provides several realistic challenges and should be a reasonable estimate of how the biologically informed model would perform in a breeding program. The population that was evaluated for NIRS phenotypes is largely independent from the population that was used for factor analysis, with only about 6% of the accessions with NIRS phenotypes also having factor scores. Moreover, the NIRS phenotypes are only approximations of total lipid or protein content. Thus, there is lower correspondence between the metabolites that were used for factor analysis and the phenotypes used for prediction. The advantage of using NIRS to estimate seed metabolites is that it is a relatively low-cost phenotyping approach compared to metabolomics and is high throughput, making it a tractable solution for many breeding programs interested in improving health-promoting compounds (Diepenbrock and Gore, 2015). Despite these challenges the multi-kernel prediction approach—when informed using markers associated with lipid-enriched factors—significantly improved prediction for lipid content compared to gBLUP and BayesB.

On the relationship between factor density and polygenicity

The positive relationship observed between the magnitude of polygenicity and loading densities, indicates that latent factors that influence many metabolites are more likely to have a complex genetic architecture. These observations are somewhat expected. If these dense latent factors are representative of some central component of the metabolomes, perturbations on these processes would likely result in large-scale changes in the metabolome and may affect fitness. Therefore, it is important that these processes are robust to mutations and are maintained at, or near some optima. This is the basis of canalization: important physiological processes will evolve to reach robust optima (Waddington 1942; Gibson 2009). And suggests that much of the oat seed metabolome is under optimizing or stabilizing selection (Slatkin 1970).

Perhaps what is more interesting are the factors that deviate from this relationship, specifically factors 4 and 17. Both exhibited dense loading patterns, oligogenic architectures (ranked 8th and 17th for density, respectively, and 50th and 73rd for polygenicity), and were enriched for lipids. The large-effect loci associated with these latent factors may have pleiotropic effects, or may consist of a set of tightly linked genes that influence the abundance of lipid-like compounds. This may explain the deviance from the density-polygenicity relationship observed for other factors. The presence of these loci raises a larger question, specifically *Why are these loci segregating in the population?* The theoretical and simulation studies by Orr (1998, 1999), as well as empirical evidence in maize and other species may help explain these observations (Doebley et al. 1997; Van Laere et al. 2003; Colosimo et al. 2004; Wang et al. 2005; Carlborg et al. 2006; Boyko et al. 2010; Brown et al. 2011). For "older" traits—*i.e.*, those associated with adaptation in natural environments—such large effect alleles at these loci would likely be removed through negative selection as these alleles may shift phenotypes far from the optimal values (Orr 1998, 1999). This was proposed by Brown et al. (2011) to explain the small effect sizes for flowering and leaf traits in maize. This is not necessarily the case for traits that are relatively "new" in evolutionary history or are not associated with adaptation. For instance, plant architecture and inflorescence traits have relatively simple genetic architectures in maize and are recent targets for artificial selection (Doebley et al. 1995, 1997; Wang et al. 2005; Brown et al. 2011; Wallace et al. 2014). This is also the case for traits under recent artificial selection in other species (Van Laere et al. 2003; Colosimo et al. 2004; Carlborg et al. 2006; Boyko et al. 2010). While it is unknown whether seed lipid content has any adaptive significance in oat, lipid content and traits that are genetically correlated with lipid content (*i.e.*, β-glucans) are important targets for many breeding programs (Welch and Lloyd 1989; Kibite and Edney 1998; Cervantes-Martinez et al. 2002). Thus, the oligogenic architectures for factors enriched for lipids may be a reflection of this relatively recent selection by breeders for lipids or traits that are genetically correlated with lipids.

Conclusions

This study shows that, we can translate biological knowledge obtained from the characterization of high dimensional "omics" data to improve prediction and selection for agriculturally important traits. The matrix factorization approach used here provides an effective means to reduce the dimensionality of the data,

while still preserving important biological features that generate correlation in the observed phenotypes. This can help reduce the multiple testing burden often experienced with GWAS on "omics" data and allow the recovery of meaningful genetic signal. We have shown that this signal can be leveraged to improve prediction in independent populations, as well as for low-cost phenotypes that provide an approximation of biochemical attributes. In a broader context, we outline an approach that can be used to manage the allocation of phenotyping resources and improve breeding decisions. For instance, breeders can phenotype a single replicate of a "discovery" population with a costly, high-resolution "omics" technology and these data can be used to inform predictions for low-cost, lower-resolution phenotypes in new populations or trials. These approaches can be easily extended to other crops, tissues and "omics" technologies to improve predictions for complex traits.

Acknowledgments

Mention of a trademark or proprietary product does not constitute a guarantee or warranty of the product by the USDA and does not imply its approval to the exclusion of other products that may also be suitable. The USDA is an equal opportunity provider and employer.

Metabolomic data were generated by H.H. and T.H.Y.; analyses were performed by M.T.C. under the guidance of M.A.G. and J.L.J.; K.P.S. and T.H.Y. generated data used for validation; M.T.C. wrote the manuscript with guidance from J.L.J. and M.A.G.; comments were provided by H.H., L.G., M.E.S., M.A.G., and J.L.J.; this study was supported by grants secured by K.P.S., L.G., M.C.T., M.E.S., M.A.G., and J.L.J.; all authors read and approved the manuscript.

Funding

Funding for this research was provided by United States Department of Agriculture - National Institute of Food and Agriculture - Agriculture and Food Research Initiative (USDA-NIFA-AFRI) grant (2017-67007-26502). The USDA is an equal opportunity provider and employer.

Conflicts of interest

None declared.

Literature cited

Barupal DK, Fiehn O. 2017. Chemical similarity enrichment analysis (chemrich) as alternative to biochemical pathway mapping for metabolomic datasets. Sci Rep. 7:11.

Bello NM, Ferreira VC, Gianola D, Rosa GJ. 2018. Conceptual framework for investigating causal effects from observational data in livestock. J Anim Sci. 96:4045–4062.

Berzonsky WA, Ohm HW. 2000. Breeding cereal small grains for value-added uses. Design Crops Added Value 40:103–145. DOI: 10.2134/agronmonogr40.

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, *et al.* 2010. A simple genetic architecture underlies morphological variation in dogs. PLos Biol. 8:e1000451.

Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, *et al.* 2011. Distinct genetic architectures for male and female inflorescence traits of maize. PLoS Genet. 7:e1002383.

Carlborg Ö, Jacobsson L, Åhgren P, Siegel P, Andersson L. 2006. Epistasis and the release of genetic variation during long-term selection. Nat Genet. 38:418–420.

Carlson MO, Montilla-Bascon G, Hoekenga OA, Tinker NA, Poland J, *et al.* 2019. Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (*Avena sativa* L.). G3 (Bethesda). 9:2963–2975.

Caspi R, Altman T, Billington R, Dreher K, Foerster H, *et al.* 2014. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucl Acids Res. 42:D459–D471.

Cerio R, Dohil M, Jeanine D, Magina S, Mahe E, *et al.* 2010. Mechanism of action and clinical benefits of colloidal oatmeal for dermatologic practice. J Drugs Dermatol. 9:1116–1120.

Cervantes-Martinez C, Frey K, White PJ, Wesenberg D, Holland J. 2002. Correlated responses to selection for greater β-glucan content in two oat populations. Crop Sci. 42:730–738.

Chan EK, Rowe HC, Hansen BG, Kliebenstein DJ. 2010. The complex genetic architecture of the metabolome. PLoS Genet. 6:e1001198.

Cheng H, Fernando R, Garrick D. 2018. JWAS: Julia implementation of whole-genome analysis software. In: Proceedings of the world congress on genetics applied to livestock production, Auckland, New Zealand, Vol. 11. p.859.

Christ B, Pluskal T, Aubry S, Weng J-K. 2018. Contribution of untargeted metabolomics for future assessment of biotech crops. Trends Plant Sci. 23:1047–1056.

Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, *et al.* 2004. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol. 2:e109.

Cui L, Lu H, Lee YH. 2018. Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases. Mass Spec Rev. 37:772–792.

De Vos RC, Moco S, Lommen A, Keurentjes JJ, Bino RJ, *et al.* 2007. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nat Protoc. 2:778–791.

Diepenbrock CH, Gore MA. 2015. Closing the divide between human nutrition and plant breeding. Crop Sci. 55:1437–1448.

DiLeo MV, Strahan GD, den Bakker M, Hoekenga OA. 2011. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PLoS One 6:e26683.

Doebley J, Stec A, Gustus C. 1995. *Teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics. 141:333–346.

Doebley J, Stec A, Hubbard L. 1997. The evolution of apical dominance in maize. Nature. 386:485–488.

Dunn WB, Ellis DI. 2005. Metabolomics: current analytical platforms and methodologies. Trends Anal Chem. 24:285–294. DOI: 10.1016/j.trac.2004.11.021

Dunn WB, Erban A, Weber RJ, Creek DJ, Brown M, *et al.* 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics. 9:44–66.

Eckert AJ, Wegrzyn JL, Cumbie WP, Goldfarb B, Huber DA, *et al.* 2012. Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. New Phytol. 193:890–902.

Edwards SM, Sørensen IF, Sarup P, Mackay TF, Sørensen P. 2016. Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. Genetics 203:1871–1883.

Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 4:250–255.

Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, *et al.* 2012. Improving accuracy of genomic predictions within and between

dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 95:4114–4129.
Fan S, Yeon A, Shahid M, Anger JT, Eilber KS, *et al.* 2018. Sex-associated differences in baseline urinary metabolites of healthy adults. Sci Rep. 8:11.
Feunang YD, Eisner R, Knox C, Chepelev L, Hastings J, *et al.* 2016. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform. 8:61.
Frey K, Holland J. 1999. Nine cycles of recurrent selection for increased groat-oil content in oat. Crop Sci. 39:1636–1641.
Gibson G. 2009. Decanalization and the origin of complex disease. Nat Rev Genet. 10:134–140.
Gulvady AA, Brown RC, Bell JA. 2013. Nutritional comparison of oats and other commonly consumed whole grains. In: ChuY, editor. Oats Nutrition and Technology. Chichester, UK: John Wiley & Sons Ltd. p. 71–93.
Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D. 2016. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theor Appl Genet. 129:2413–2427.
Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 12:186.
Kale M, Hamaker B, Bordenave N. 2013. Oat β-glucans: physicochemistry and nutritional properties. In: ChuY, editor. Oats Nutrition and Technology. Chichester, UK: John Wiley & Sons Ltd. p. 123–169.
Kanehisa M. 2002. The KEGG database. In Novartis Foundation Symposium. Wiley Online Library. Chichester, UK: John Wiley & Sons Ltd. p. 91–100.
Keurentjes JJ, Fu J, De Vos CR, Lommen A, Hall RD, *et al.* 2006. The genetics of plant metabolism. Nat Genet. 38:842–849.
Kibite S, Edney M. 1998. The inheritance of β-glucan concentration in three oat (*Avena sativa* L.) crosses. Can J Plant Sci. 78:245–250.
Kurtz ES, Wallo W. 2007. Colloidal oatmeal: history, chemistry and clinical properties. J Drugs Dermatol. 6:167–170.
Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9:559.
Leinonen T, McCairns RJS, O'Hara RB, Merilä J. 2013. Q ST–F ST comparisons: evolutionary and ecological insights from genomic heterogeneity. Nat Rev Genet. 14:179–190.
Li J, Ji L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 95:221–227.
MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, *et al.* 2016. Exploiting biological priors and sequence variants enhances qtl discovery and genomic prediction of complex traits. BMC Genomics. 17:144.
Marco-Ramell A, Tulipani S, Palau-Rodriguez M, Gonzalez-Dominguez R, Minarro A, *et al.* 2018. Untargeted profiling of concordant/discordant phenotypes of high insulin resistance and obesity to predict the risk of developing diabetes. J Proteome Res. 17:2307–2317.
Matsuda F, Nakabayashi R, Yang Z, Okazaki Y, Yonemaru J-I, *et al.* 2015. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. Plant J. 81:13–23.
Meuwissen T, Hayes B, Goddard M. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157: 1819–1829.
Orr HA. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution. 52:935–949.
Orr HA. 1999. The evolutionary genetics of adaptation: a simulation study. Genet Res. 74:207–214.
Owen AB, Wang J. 2016. Bi-cross-validation for factor analysis. Statis Sci. 31:119–139.
Perez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 198:483–495.
Peterson DM, Wood DF. 1997. Composition and structure of high-oil oat. J Cereal Sci. 26:121–128.
Phillips PC, Whitlock MC, Fowler K. 2001. Inbreeding changes the shape of the genetic covariance matrix in *Drosophila melanogaster*. Genetics. 158:1137–1145.
Price PB, Parsons JG. 1975. Lipids of seven cereal grains. J Am Oil Chem Soc. 52:490–493.
Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, *et al.* 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet. 44:217–220.
Rosa G, Valente B. 2013. Breeding and genetics symposium: inferring causal effects from observational data in livestock. J Anim Sci. 91: 553–564.
Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ. 2008. Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. Plant Cell 20:1199–1216.
Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. 2016. Untargeted metabolomics strategies–challenges and emerging directions. J Am Soc Mass Spectrom. 27:1897–1905.
Showalter MR, Wancewicz B, Fiehn O, Archard JA, Clayton S, *et al.* 2019. Primed mesenchymal stem cells package exosomes with metabolites associated with immunomodulation. Biochem Biophys Res Commun. 512:729–735.
Šidák Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc. 62:626–633.
Slatkin M. 1970. Selection and polygenic characters. Proc Natl Acad Sci USA. 66:87–93.
Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, *et al.* 2018. Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res . 46: D661–D667.
Speed D, Balding DJ. 2014. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 24:1550–1557.
Spirtes P, Glymour CN, Scheines R, Heckerman D. 2000. Causation, Prediction, and Search. Cambridge, MA, USA: MIT Press.
Stephens M. 2016. False discovery rates: a new deal. Biostatistics. 18: 275–294.
Storey JD. 2002. A direct approach to false discovery rates. J Royal Stat Soc. 64:479–498.
Storz JF. 2002. Contrasting patterns of divergence in quantitative traits and neutral DNA markers: analysis of clinal variation. Mol Ecol. 11:2537–2551.
Tohge T, Fernie AR. 2010. Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat Protoc. 5:1210–1227.
Turner-Hissong SD, Bird KA, Lipka AE, King EG, Beissinger TM, *et al.* 2019. Genomic prediction informed by biological processes expands our understanding of the genetic architecture underlying free amino acid traits in dry *Arabidopsis* seeds. bioRxiv. 272047.
Van Laere A-S, Nguyen M, Braunschweig M, Nezer C, Collette C, *et al.* 2003. A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. Nature. 425:832–836.
VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414–4423.
Waddington CH. 1942. Canalization of development and the inheritance of acquired characters. Nature. 150:563–565.
Wallace J, Larsson S, Buckler E. 2014. Entering the second century of maize quantitative genetics. Heredity. 112:30–38.

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, *et al.* 2005. The origin of the naked grains of maize. Nature. 436:714–719.

Wang W, Stephens M. 2018. Empirical bayes matrix factorization. arXiv preprint. arXiv:1802.06931.

Welch RW, Lloyd JD. 1989. Kernel (1 3)(1 4)-β-d-glucan content of oat genotypes. J Cereal Sci. 9:35–40.

Wen W, Li D, Li X, Gao Y, Li W, *et al.* 2014. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun. 5:10.

Wishart DS, Li C, Marcu A, Badran H, Pon A, *et al.* 2020. PathBank: a comprehensive pathway database for model organisms. Nucleic Acids Res. 48:D470–D478.

Wu S, Tohge T, Cuadros-Inostroza Á, Tong H, Tenenboim H, *et al.* 2018. Mapping the Arabidopsis metabolic landscape by untargeted metabolomics at different environmental conditions. Mol Plant. 11:118–134.

Xu S, Xu Y, Gong L, Zhang Q. 2016. Metabolomic prediction of yield in hybrid rice. Plant J 88:219–227.

Xu Y, Xu C, Xu S. 2017. Prediction and association mapping of agronomic traits in maize using multiple omic data. Heredity. 119: 174–184.

Youngs V. 1978. Oat lipids. Cereal Chem. 55:591–597.

Zhou S, Tong L, Liu L. 2019. Oats. In: Wang J, Sun B, Cao R, editors. Bioactive Factors and Processing Technology for Cereal Foods. Singapore: Springer. p. 185–206.

Communicating editor: H. Daetwyler