# Multi-omics Prediction of Oat Agronomic and Seed Nutritional Traits Across Environments and in Distantly Related Populations

**Haixiao Hu** ( haixiao.work@gmail.com )
Cornell University https://orcid.org/0000-0001-5888-4839

**Malachy Campbell**
Cornell University

**Trevor Howard Yeats**
Cornell University

**Xuying Zheng**
Cornell University

**Daniel Runcie**
UC Davis: University of California Davis

**Giovanny Covarrubias-Pazaran**
CIMMYT: Centro Internacional de Mejoramiento de Maiz y Trigo

**Corey Broeckling**
Colorado State University

**Linxing Yao**
Colorado State University

**Melanie Caffe-Treml**
South Dakota State University

**Lućia Gutiérrez**
University of Wisconsin-Madison

**Kevin Smith**
University of Minnesota

**James Tanaka**
Cornell University

**Owen Hoekenga**
Cayuga Genetics Consulting Group LLC

**Mark Sorrells**
Cornell University

**Michael Gore**
Cornell University

**Jean-Luc Jannink**

Cornell University

**Research Article**

**Keywords:** multi-omics prediction, transcripts, metabolites, multi-environment trials, distantly-related populations, oat

**Posted Date:** June 14th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-581505/v1

# Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations

Haixiao Hu[1*], Malachy T. Campbell[1], Trevor H. Yeats[1], Xuying Zheng[1], Daniel E. Runcie[2], Giovanny Covarrubias-Pazaran[3], Corey Broeckling[4], Linxing Yao[4], Melanie Caffe-Treml[5], Lucía Gutiérrez[6], Kevin P. Smith[7], James Tanaka[1], Owen A. Hoekenga[8], Mark E. Sorrells[1], Michael A. Gore[1], and Jean-Luc Jannink[1,9]

[1]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

[2]Department of Plant Sciences, University of California Davis, Davis, CA 95616, USA

[3]International Maize and Wheat Improvement Center (CIMMYT), Km. 45, Carretera México-Veracruz, El Batán, Texcoco, Edo. de México, CP 56130, México

[4]Proteomics and Metabolomics Facility, Colorado State University, C130 Microbiology, 2021 Campus Delivery, Fort Collins, CO 80521, USA

[5]Department of Agronomy, Horticulture & Plant Science, South Dakota State University, Brookings, SD 57007, USA

[6]Department of Agronomy, University of Wisconsin-Madison Madison, WI 53706, USA

[7]Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA

[8]Cayuga Genetics Consulting Group LLC, Ithaca, NY 14850 USA

[9]USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853 USA

*Correspondence (email: haixiao.work@gmail.com)

# Abstract

***Key message*** **Integration of multi-omics data improved prediction accuracies of oat agronomic and seed nutritional traits in multi-environment trials and distantly-related populations in addition to the single-environment prediction.**

Multi-omics prediction has been shown to be superior to genomic prediction with genome-wide DNA-based genetic markers (G) for predicting phenotypes. However, most of the existing studies were based on historical datasets from one environment; therefore, they were unable to evaluate the efficiency of multi-omics prediction in multi-environment trials and distantly-related populations. To fill those gaps, we designed a systematic experiment to collect omics data and evaluate 17 traits in two oat breeding populations planted in single and multiple environments. In the single-environment trial, transcriptomic BLUP (T), metabolomic BLUP (M), G+T, G+M and G+T+M models showed greater prediction accuracy than GBLUP for 5, 10, 11, 17 and 17 traits, respectively, and metabolites generally performed better than transcripts when combined with SNPs. In the multi-environment trial, multi-trait models with omics data outperformed both counterpart multi-trait GBLUP models and single-environment omics models, and the highest prediction accuracy was achieved when modeling genetic covariance as an unstructured covariance model. We also demonstrated that omics data can be used to prioritize loci from one population with omics data to improve genomic prediction in a distantly-related population using a two-kernel linear model that accommodated both likely casual loci with large-effect and loci that explain little or no phenotypic variance. We propose that the two-kernel linear model is superior to most genomic prediction models that assume each variant is equally likely to affect the trait and can be used to improve prediction accuracy for any trait with prior knowledge of genetic architecture.



**Key words:** multi-omics prediction, transcripts, metabolites, multi-environment trials, distantly-related populations, oat

# Introduction

Oat (Avena sativa L.) ranks sixth in world cereal production and has increasingly been consumed as a human food (USDA, 2019). Oat has a high content of health-promoting compounds such as unsaturated fatty acids, dietary fiber, antioxidants and vitamins, which makes it an interesting target for metabolomics studies from a human health and nutrition perspective (IMARC Group, 2019). In addition, high-density genetic markers have been developed in oat (Bekele et al., 2018), a draft genome sequence has been released (PepsiCo, 2020) and a high-quality and comprehensive seed transcriptome has been characterized (Hu et al., 2020). Furthermore, recent advances in high throughput sequencing and metabolite profiling technologies enable quantification of gene expression and metabolite abundance for hundreds of samples with high precision and reasonable cost (Alseekh & Fernie, 2018; Moll et al., 2014). All these advances in technology provides an opportunity to integrate different omics data and improve predictions for phenotypes of interest.

Several multi-omics prediction studies have been reported in cereal species (Guo et al., 2016; Riedelsheimer et al., 2012; Schrag et al., 2018; Wang et al., 2019; Westhues et al., 2017; Y. Xu et al., 2017; Yang Xu et al., 2021). These studies have shed light on the merits of multi-omics prediction over traditional genomic prediction and discussed useful statistical methods for integrating omics data. For instance, Y. Xu et al. (2017) and Wang et al. (2019) suggested that best linear unbiased prediction was the most efficient method compared to other commonly used genomic prediction and non-linear machine learning methods. However, most of those studies were based on historical datasets with a limited number of metabolite features and each level of omics data was collected from different projects. Therefore, they were unable to evaluate the efficiency of multi-omics prediction in multi-environment trials and genetically distant populations. However, in plant breeding, multi-environment trials are important for assessing the performance of genotypes across environments and identifying well-adapted genotypes for a specific region (Burgueño et al., 2012; Mathew et al., 2018). In addition, prediction of breeding values of distantly-related individuals are needed in many and perhaps the most promising applications of genomic selection in both plant and animal breeding programs (Lorenz & Smith, 2015; Meuwissen, 2009; Moghaddar et al., 2019).

To fill the knowledge gaps of multi-omics prediction in plant breeding, we designed a systematic experiment to collect omics data and evaluate eight agronomic and nine fatty acid traits (Table S1) in a core set of a worldwide oat collection (termed Diversity panel) planted in one environment and advanced breeding lines adapted to the upper Midwest region in the U.S.

(termed Elite panel) planted in three environments. Our efforts included (i) comparing the accuracy of mufti-omics prediction against genomic prediction in a single-environment trial; (ii) evaluating the efficiency of multi-omics prediction in multi-environment trials; and (iii) exploring the potential of using multi-omics data to predict distantly-related individuals.

# Materials and methods

## The plant materials and experimental designs

The Diversity and Elite panels consisted of 378 and 252 lines (Table S2), respectively. The Diversity panel originally included 500 lines described by Carlson et al. (2019) that was a core set of worldwide collection of oat germpalsm, and we further selected for lines with visible anther extrusion for the convenience of collecting developing seeds for RNA sequencing. The Diversity panel was planted at Ithaca, NY, and the Elite panel was planted at Madison, WI, Crookston, MN, and Brookings, SD, respectively. An augmented incomplete design was used for both panels. The Diversity panel included 18 blocks of 23 plots each, one common check across all blocks and six secondary checks replicated in three blocks each. The Elite panel included 12 blocks of 25 plots each, one common check across all blocks and two secondary checks replicated in six blocks each.

## Phenotype evaluation and analysis

Plant height was evaluated for five randomly selected plants in each plot after anthesis. Days to heading was defined by the days from seeding to heading in >50% of total plants. 100 randomly selected seeds from each plot were dehulled with a hand dehuller for evaluation of hundred kernel weight, hundred hull weight and groat percentage. After dehulling, 50 randomly selected seeds were delivered to the Proteomics and Metabolomics Facility at Colorado State University for metabolite analysis, and the other 50 seeds were used for measuring seed length, width and height with an electronic micrometer. Fatty acids were identified and quantified with targeted GC-MS, then normalized to concentration (mg/g of oats) against the internal standard (C17:0) (details were described in the Supplemental Methods).

## Genotype analysis

Genotypic data of the two panels were downloaded from T3/oat (https://triticeaetoolbox.org/oat/). SNPs were filtered using the following criteria (i) minor allele frequency (MAF) > 2%; (ii) site missingness < 60%; and (iii) site heterozygosity < 10%. After initial SNP filtering, lines were selected if (i) call rate > 80% and (ii) heterozygosity < 10%. A

total of 73,014 markers and 568 lines (368 for the diversity panel, 232 for the elite panel, 32 in common) met these criteria and were used for further analyses. Subsequently, missing genotypes were imputed using the linear regression method glmnet described by Chan et al. (2016). The imputed genotypic data was used for constructing a neighbor-joining tree based on Rogers' distance using the ape package (Paradis et al., 2004), and the tree was visualized with the ggtree package (Yu, 2020).

## Transcript profiling

RNAseq was based on developing seeds at 23 days after anthesis (DAA). The 23 DAA was chosen based on our pilot study (Hu et al., 2020) that showed 23 DAA had slightly higher correlation between transcript and metabolite abundance than other sampled seed developmental time points. Seed sample collection, RNA extraction, library construction procedures were described in details by Hu et al. (2020). Pooled libraries were sequenced using Illumina NextSeq500 with a 150 nt single-end run. The RNAseq reads quality trimming, transcript abundance quantification, and library size normalization followed Hu et al.(2020).

## Metabolite profiling

Metabolite analysis was based on physiologically mature seeds because they have the highest level of health-promoting compounds and those compounds are stable at room temperature until germination. GC-MS non-targeted analysis and LC-MS Phenyl-Hexyl analysis were done at the Proteomics and Metabolomics Facility at Colorado State University. Details of chemical analysis, raw mass spectrometry data processing, metabolite annotation and normalization were described in the Supplemental Methods. The normalized metabolomics data was used for network analysis with the WGCNA package (Zhang & Horvath, 2005).

## Analysis of phenotypic traits, transcriptomic and metabolic features

Phenotypic traits, transcriptomic and metabolic features were analyzed following a standard linear mixed model of an augmented design accounting for effects of check genotypes and blocks (Campbell et al. 2021a). For metabolites analysis, batch effect was also included in the model to account for batch variation. All statistical models were described in the Supplemental Methods and fitted using the sommer package (Covarrubias-Pazaran, 2016).

## Single-environment prediction

The additive genomic relationship matrix was made with the A.mat function implemented in the rrBLUP package (Endelman, 2011), and relationship matrices for transcriptomics and

metabolomics data were made following Westhues et al. (2017). GBLUP, Transcriptomic BLUP (T), metabolomic BLUP (M), G+T, G+M and G+T+M models were fitted with the BGLR package (Pérez & De Los Campos, 2014). In the Diversity panel, transcriptomics and metabolomics data were collected on the same plots as the phenotypic data and therefore non-genetic (i.e., microenvironmental) factors that affected both omics features and phenotypic traits may induce non-genetic correlations among traits. Therefore, we estimated prediction accuracy as $cor_g\left(\sqrt{\hat{h}^2_{\hat{u}}}\right)$ described by Runcie and Cheng (2019), and used a 50:50 training:testing split of the data to ensure that $cor_g$ could be estimated accurately in the testing partition. This cross-validation procedure was repeated for 50 times with different random partitions.

## Multi-environment prediction

The metabolomics data were also collected on the same plots as the phenotypic data in the Elite panel, which would bias prediction accuracy if directly using metabolites to predict target phenotypes from the same environment. Therefore, when predicting target phenotypes from one environment, we used metabolites from other two environments to make the metabolomic relationship matrix. For each trait, we fitted six multi-trait mixed models on G, M and G+M kernels with different genetic and residual covariance structures (Table S3). We applied a single environment cross validation method for genomic prediction described by Mathew et al. (2018) and extended it to multi-kernel omic prediction (illustrated in Fig. S1). To predict a phenotype in the first environment, we masked 20% of lines for cross validation and used metabolites from the other two environments to construct the metabolomic relationship matrix. We then used multi-trait models treating phenotypes from all three environments as separate traits for model training but using only the phenotypic data of the masked lines from the first environment as the testing data. We further estimated prediction accuracy of the first environment as $r(\hat{y}, y)/\sqrt{h^2}$(Riedelsheimer et al., 2012), where r(ŷ,y) is the Pearson correlation between the observed (y) and predicted (ŷ) phenotypic values and $h^2$ is the heritability of the target trait. To predict the phenotype in the second and third environments, we masked 20% of lines (the same genotypes as those in the first environment) from the second and third environments, respectively, and calculated their prediction accuracies following the same procedure as that applied to the first environment. Finally, we averaged the three prediction accuracies across environments to represent the prediction accuracy of a single run. This procedure was repeated for 50 times with different random partitions.

## Prediction of distantly related individuals

Seed fatty acid concentrations were used as target traits for predicting distantly related individuals, which included two steps: likely causal loci prioritization in the Diversity panel and multiple-kernel prediction in the Elite panel.

We first performed the WGCNA on all metabolite features in the Diversity panel, and identified twenty-six network modules. Based on the metabolites annotation, we performed the Fisher's exact test to identify a subset of network modules enriched with lipids and lipid-like molecules. We then performed hierarchical clustering and GWAS on eigenvectors of the twenty-six network modules and PC1 of fatty acids. Based on these analyses, we found that a darkred module enriched with lipids and lipid-like molecules, clustered together with PC1 of fatty acids, and its eigenvector had a QTL co-located with the major-effect QTL of fatty acids on chromosome 6A. We finally prioritized 140 markers including significant markers and the markers in LD with them based on GWAS hits of the darkred module. A LD threshold of $r^2$=0.1 was used as it is frequently recommended for SNP pruning (Kawakami et al., 2014).

The prioritized markers and all rest markers were used to construct two genomic relationship kernels in the Elite panel and perform a multiple kernel prediction. Genomic predictions with GBLUP and BayesB models were used as references to compare with the two-kernel linear model. The five-fold cross-validation was ued to estimate prediction accuracies for all models and the prediction accuracy was estimated as $r(\hat{y}, y)/\sqrt{h^2}$ (Riedelsheimer et al., 2012). This cross-validation procedure was repeated for 50 times with different random partitions.

# Results

After filtering out lines with low-quality genetic markers, the Diversity and Elite panels consisted of 368 and 232 lines (Table S2), respectively, with 32 lines in common. A reconstructed phylogenetic tree revealed that most of clusters were primarily comprised of either the Diversity or the Elite panel members, although a couple of clusters had approximately equal representation from both sets (Fig. 1). This is consistent with our prior knowledge about different origins of the two panels (Carlson et al., 2019; Campbell et al., 2021b) .

## Single-environment prediction in the Diversity panel

Using GBLUP (G) as a baseline, there were 5, 10, 11, 17 and 17 traits out of the 17 total traits with improved prediction accuracy from transcriptomic BLUP (T), metabolomic BLUP (M), G+T, G+M and G+T+M models, respectively (Fig. 2, Table S4). Percent change in prediction accuracy over GBLUP ranged from 0.1% (Days to Heading, G+T model) to 70.3% (C18:0, G+M model) with a median of 21.5%. Because GBLUP does not allow for large-effect or zero-effect genetic markers, we also compared BayesB with the multi-omics models, and found BayesB showed similar results to GBLUP (Fig. S2).

To evaluate whether transcriptomic and metabolomic features equally contribute to improved prediction accuracy or if one is more important than the other, we compared multi-omics prediction models with T and M kernels added in different orders. By adding kernels in their order along the central dogma of molecular biology, median prediction accuracy changes from G to G+T models and from G+T to G+T+M models across all traits ranged from -11.6% to 35.8% (median=3.2%) and 6.5% to 55.6% (median=16.3%), respectively (Fig. S3). In contrast, when adding the M kernel first (G+M model) then followed by the T kernel (G+T+M model), percent changes in prediction accuracy ranged from 2.5% to 67.3% (median=41.7%) and -3.3% to 3.5% (median=-0.03%), respectively (Fig. S4). These results indicated that seed metabolites generally contributed more than transcripts to improving prediction accuracy of both agronomic and seed nutritional traits when combined with SNPs.

In addition to playing important roles in improving prediction accuracy when combined with other kernels, metabolites alone from mature seeds (M model) greatly outperformed SNPs (G model) and transcripts (T model) in predicting fatty acids (except C16:1, Fig. 2). To understand why metabolites are better predictors for fatty acid traits, we used the Weighted Gene Co-expression Network Analysis (WGCNA, Zhang & Horvath, 2005) that accommodated both annotated and unannotated compounds and used metabolites annotations (Table S5) to elucidate their

biological functions. The WGCNA was designed to construct gene/metabolite co-expression networks, and a co-expression module (network module) may reflect a true biological pathway (Langfelder and Horvath 2008). We identified twenty-six network modules and eight of them were enriched with lipids and lipid-like molecules (Table S6), which included 33.0% of total identified seed metabolite compounds.

## Multi-environment prediction in the Elite panel

Beyond single-environment prediction, omics data might also have merit in predicting multi-environment trials, which has not yet been investigated to our knowledge. Here we used SNPs and metabolites for analyzing the multi-environment trials in the Elite panel, because transcript profiling from a single developmental time point showed limited value for improving prediction accuracy in addition to being very labor-intensive. We focused on prediction of lines that have been evaluated in some but not in target environments (CV2, Burgueño et al., 2012). To this aim, we applied a single environment cross validation method (Mathew et al., 2018) (Fig. S1). Briefly, to predict a phenotype in the first environment, we masked 20% of lines for cross validation and used metabolites from the other two environments to construct metabolomic relationship matrices to minimize the influence of non-genetic effects on prediction accuracy. We then used multi-trait models treating phenotypes from all three environments as separate traits for model training but using only the phenotype data of the masked lines from the first environment as the testing data. This procedure was repeated for the second and third environments and prediction accuracies were averaged across the three environments for each run.

Multi-environment predictions were performed using six multi-trait models (Table S3) on three different kernels/combinations (G, M, G+M) with various genetic and residual covariance structures (Fig. 3, Fig. S5). The diagonal heterogeneous covariance structure (D-D, the uppercase letters before and after the hyphen represent genetic and residual covariance structures and D=diagonal) corresponds to a single-environment model without borrowing information from other environments. The question that we explored was whether multi-omics models (M and G+M) could improve prediction accuracy compared to corresponding multi-trait models based on SNPs alone (G model). To answer this question, within each of the five multi-trait models (the D-D model was excluded), we compared percent change in prediction accuracy of M and G+M models relative to the G model. We found the M model outperformed the G model for all seed fatty acid traits except C16:1 and C18:3, with an increase in prediction accuracy ranging from 0.1 to 15.9%. However, the G+M model outperformed the G model for all

traits except days to heading, with an increase in prediction accuracy over the G model ranging from 0.1 to 13.9%. These results confirmed the value of using multi-omics data in the multi-environment prediction.

We then used the prediction accuracy from GBLUP in the single-environment model (D-D) as a baseline to compare the performance of different multi-trait models. We found that all multi-trait models outperformed their counterpart single-environment models (Fig. 3, Figs. S6-8). The multi-trait models generally performed better when modeling the genetic covariance as unstructured (UN) or as factor-analytic (FA) than modeling genetic covariance as a diagonal structure (D). The highest prediction accuracy was achieved by either UN-D (UN and D represent genetic and residual covariance structures, respectively) or UN-UN models, although FA-D and FA-UN models provided very similar results.

## Using multi-omics data to improve genomic prediction in distantly-related populations

Although multi-omics data showed superiority over SNPs to predict phenotypes in both single and multi-environment trials, currently transcript and metabolite profiling is more expensive than SNP genotyping, which would limit their applications in plant breeding. Here we hypothesized that omics data from well characterized populations can be used to prioritize likely causal loci and improve performance of genomic prediction models in distantly-related populations. Seed fatty acid concentrations were used as target traits to test the hypothesis because their genetic architectures have been well characterized (Carlson et al., 2019) and lipid biosynthetic pathways are known to be highly conserved in higher plants (de Abreu e Lima et al., 2018).

To explore this scientific question, we first attempted to prioritize likely causal loci from the Diversity panel based on the eight network modules enriched with lipids and lipid-like molecules (Table S6). Among the eight network modules, only one (darkred) strongly correlated with fatty acids (Fig. S9). We then performed hierarchical clustering and GWAS on eigenvectors of all the 26 network modules and PC1 of fatty acids. The eigenvector of the darkred module was clustered together with PC1 of fatty acids (Fig. S10) and had significant GWAS hits on chromosome 6A (Fig. S11), which co-located with the fatty acids major-effect QTL (*QTL-6A*, Fig. S12). However, the *QTL-6A* was not detected from other network modules. We further prioritized 140 markers including significant markers and the markers in LD with them based on the darkred module GWAS hits on chromosome 6A.

The primary use of locus prioritization is to split markers in the test population into two sets for a multi-kernel model prediction, in which the two genomic relationship kernels were constructed from the two marker sets. We termed our method multi-kernel network-based prediction (MK-Network) and found it improved prediction accuracy over GBLUP and BayesB for all fatty acid traits (Fig. 4) except C14:0 and C18:3, because they had different genetic architectures from other fatty acids and no significant markers from GWAS (Fig. S12). The percent change of mean prediction accuracy over 50 cross-validation runs ranged from 4.0% to 32.0% with a mean of 14.5%.

# Discussion

### Roles of transcripts and metabolites in the single-environment prediction

In the single-environment prediction, we found that transcripts showed limited value for improving prediction accuracy either by themselves alone or by combining with SNPs. Other researchers (Westhues et al., 2017; Y. Xu et al., 2017) also reported that prediction abilities of transcripts were lower than GBLUP. The poor predictive performance of transcripts in existing studies might be explained by the fact that they were collected from a single developmental time point and subject to dynamic changes in later unsampled developmental stages or by that transcripts and SNPs tend to capture similar genetic signals for predicted traits (Guo et al., 2016).

Although metabolites played important roles when combined with other kernels in improving prediction accuracy, we found that metabolites alone from mature seeds (M model) showed mixed results for predicting agronomic traits (Fig. 2), while they greatly outperformed SNPs in predicting fatty acids. The relatively low performance of mature seed compounds in predicting agronomic traits might be explained by the fact that development of the agronomic traits and accumulation of compounds in mature seeds occurred either at different times or in different tissues. In contrast, these compounds and fatty acids were synthesized in the same tissue, a large proportion of them directly or indirectly connected with fatty acids through biochemical pathways (Tables S4-5) and different pathways relevant to lipids were likely influenced by overlapping gene sets. Therefore, they should be able to capture more genetic co-variation (including both additive and non-additive covariation) with fatty acids than SNPs fitted in an additive model. This hypothesis was partially supported by our results that combining G model and M model (G+M model) significantly improved prediction accuracies than using the G model

alone for all the 17 traits (Fig. 2, Table S7) and by findings of Guo et al. (2016) that adding metabolites to saturated SNP densities still led to significant increases in predictive abilities.

## Application of omics data in the multi-environment prediction

In the multi-environment prediction, we observed that for predicting agronomic traits, the M model performed similarly to the G model (i.e.M~G, Fig. 3), however, the M model outperformed G model for predicting fatty acids traits (i.e. M>G). This pattern is very similar to that observed in the single-environment prediction, and therefore could be interpreted similarly. Both analyses indicated that when predicting traits very distantly connected or unconnected through biological pathways, metabolites functioned similarly to DNA-based genetic markers (i.e. we need to trace back to the DNA along the central dogma); however, when predicting relevant traits that directly/indirectly connected through biological pathways, metabolites could capture more genetic co-variation with the target traits than DNA-based genetic markers, because they shared more similarities in temporal and spatial expression.

In addition, we observed that all multi-trait models outperformed their counterpart single-environment models (Fig. 3, Figs. S6-8), and the multi-trait models generally performed better when modeling the genetic covariance as unstructured (UN) or as factor-analytic (FA) than modeling genetic covariance as a diagonal structure (D). This indicated that the genetic covariance between environments played an important role in the multi-omics prediction models. These findings agree with recent genomic prediction studies (Malosetti et al., 2016; Mathew et al., 2018; Montesinos-López et al., 2016) that UN covariance structure improved prediction accuracy compared to the models with diagonal homogeneous or heterogeneous covariances. Overall, we concluded that considering genetic and non-genetic covariances is useful to improve prediction accuracy of multi-environment models using multi-omics data.

## The genetic basis of predicting distantly-related individuals and advantages of the two-kernel linear model

In the prediction of distantly-related individuals, the universal QTL of fatty acids (*QTL-6A*, Figs. S12-13) and similar LD relationships (Fig. S14) with the surrounding loci between the Diversity and Elite panels promoted the success of our likely causal loci prioritization. The network-based prioritization strategy takes advantages of pleiotropy, in which one or a few genes influence both target traits and other metabolites from related network modules. In the darkred module, 23 of 32 metabolites showed clear peaks at the *QTL-6A*, although only five of them were significant at FDR<0.05 (Fig. S15). This indicated that *QTL-6A* was likely a causal locus and influenced both

fatty acids and the darkred module. The relationships between fatty acids and the darkred module are expected to be conserved between populations. However, we were unable to test this because there is currently no robust method to map all untargeted metabolites from one panel to another and quantify them precisely.

Most genomic prediction methods assume that each variant is equally likely to affect the trait (MacLeod et al., 2016). There are certain loci that explain more phenotypic variance and they should be placed in different kernels than loci that explain little or no variance. However, the other kernel is still needed because we may unintentionally exclude important loci based on prior biological knowledge alone, for example, a prior GWAS might not identify all possible causal loci. There are many loci that have small effects, through whatever pathway, whether it is through trans effects as hypothesized in the omnigenic model (Liu et al., 2019) or through much more indirect effects like competition for photosynthates or impact on fitness (Price et al., 2018). Li et al. (2018) found that excluding those small-effect loci could not further improve prediction accuracy compared to GBLUP with all SNPs. Therefore, a two-kernel linear model that accommodates both likely casual loci and loci with minimal to no effect should be used to improve prediction accuracy for any traits with prior knowledge of genetic architecture.

## Acknowledgements

We thank Joshua Wood and Robin Buell for helping with oat seed RNA extraction; David Benscher, Amy Tamara Fox and Nicholas Kaczmar for help with planting and harvesting field trials and sample collection; Yujie Meng for phenotype evaluation; Jing Wu and Peter Schweitzer for library preparation and RNA sequencing.

## Author contribution statement

J.J., M.A.G and M.E.S designed the research. H.H. analyzed the data. H.H., M.T.C, M.A.G and J.J. wrote the manuscript. D.E.R, G.C., O.A.H and M.E.S advised H.H. on data analysis. H.H, T.H.Y, X.Z., M.C., L.C., K.P.S. J.T. performed experiments. C.B. and L.Y. performed metabolite analysis. All co-authors were involved in editing the manuscript.

## Funding

Funding for this research was provided by USDA-NIFA-AFRI 2017-67007-26502. 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.

## Compliance with ethical standards

**Conflict of interest** The authors have no conflict of interest to declare

## Figure legends

**Fig. 1.** Neighbor-joining tree of 568 oat lines in the Diversity and Elite panels. Different panels are shown in different colors (darkblue, Diversity panel; red, Elite panel, light blue, lines in common).

**Fig. 2** Distribution of prediction accuracy of the 17 phenotypic traits in the Diversity panel across 50 re-sampling runs. For each trait, boxplots with different colors represent prediction models. Medians of percent change in prediction accuracy of models relative to GBLUP are indicated below each box in blue if positive and in red if negative. G = genomic BLUP, T = transcriptomic BLUP, M = metabolomic BLUP.

**Fig. 3** Distribution of prediction accuracy of the 15 phenotypic traits in the Elite panel across 50 re-sampling runs estimated by multi-trait models. For each trait, boxplots with different colors represent models. Medians of percent change in prediction accuracy of M and G+M models relative to the G model are indicated below each box in blue if positive and in red if negative. For each model, the uppercase letters before and after the hyphen represent genetic and residual covariance structures: D=diagonal, UN=unstructured, FA=factor-analytic.

**Fig. 4** Prediction accuracy of the 10 fatty acid traits in the Elite panel estimated by GBLUP, BayesB and two-kernel BLUP models across 50 re-sampling runs. For each trait, barplots with different colors represent models. Means of percent change in prediction accuracy of all other models relative to GBLUP are indicated above each bar (in blue if positive, in red if negative, and in black if zero). MK-Network=network-based multiple-kernel prediction.

# References


Alseekh S, Fernie AR (2018) Metabolomics 20 years on: what have we learned and what hurdles remain? Plant J 94:933–942. https://doi.org/10.1111/tpj.13950

Bekele WA, Wight CP, Chao S, et al (2018) Haplotype-based genotyping-by-sequencing in oat genome research. Plant Biotechnol J 16:1452–1463. https://doi.org/10.1111/pbi.12888

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52:707–719. https://doi.org/10.2135/cropsci2011.06.0299

Campbell MT, Hu H, Yeats TH, et al (2021a) Translating insights from the seed metabolome into improved prediction for lipid-composition traits in oat (Avena sativa L.). Genetics 217:. https://doi.org/10.1093/genetics/iyaa043

Campbell MT, Hu H, Yeats TH, et al (2021b) Improving Genomic Prediction for Seed Quality Traits in Oat (Avena sativa L.) Using Trait-Specific Relationship Matrices. Front Genet 12:1–12. https://doi.org/10.3389/fgene.2021.643733

Carlson MO, Montilla-Bascon G, Hoekenga OA, et al (2019) Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (Avena sativa L.). G3 Genes, Genomes, Genet 9:2963–2975. https://doi.org/10.1534/g3.119.400228

Chan AW, Hamblin MT, Jannink JL (2016) Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data. PLoS One 11:1–17. https://doi.org/10.1371/journal.pone.0160733

Covarrubias-Pazaran G (2016) Genome-Assisted prediction of quantitative traits using the r package sommer. PLoS One 11:1–15. https://doi.org/10.1371/journal.pone.0156744

de Abreu e Lima F, Li K, Wen W, et al (2018) Unraveling lipid metabolism in maize with time-resolved multi-omics data. Plant J 93:1102–1115. https://doi.org/10.1111/tpj.13833

Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4:250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Guo Z, Magwire MM, Basten CJ, et al (2016) Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theor Appl Genet 129:2413–2427. https://doi.org/10.1007/s00122-016-2780-5

Hu H, Gutierrez-Gonzalez JJ, Liu X, et al (2020) Heritable temporal gene expression patterns correlate with metabolomic seed content in developing hexaploid oat seed. Plant Biotechnol J 18:1211–1222. https://doi.org/10.1111/pbi.13286

IMARC Group (2019) Oats Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2019-2024 http://www.reportlinker.com/p04715198-summary/view-report.html

Kawakami T, Backström N, Burri R, et al (2014) Estimation of linkage disequilibrium and interspecific gene flow in Ficedula flycatchers by a newly developed 50k single-nucleotide polymorphism array. Mol Ecol Resour 14:1248–1260. https://doi.org/10.1111/1755-0998.12270

Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9:. https://doi.org/10.1186/1471-2105-9-559

Li B, Zhang N, Wang YG, et al (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 9:1–20. https://doi.org/10.3389/fgene.2018.00237

Liu X, Li YI, Pritchard JK (2019) Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177:1022-1034.e6. https://doi.org/10.1016/j.cell.2019.04.014

Lorenz, A. J., & Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in Barley. *Crop Science*, *55*(6), 2657–2667. https://doi.org/10.2135/cropsci2014.12.0827

MacLeod IM, Bowman PJ, Vander Jagt CJ, et al (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17:1–21. https://doi.org/10.1186/s12864-016-2443-6

Malosetti M, Bustos-Korts D, Boer MP, Van Eeuwijk FA (2016) Predicting responses in multiple environments: Issues in relation to genotype × Environment interactions. Crop Sci 56:2210–2222. https://doi.org/10.2135/cropsci2015.05.0311

Mathew B, Léon J, Sillanpää MJ (2018) Impact of residual covariance structures on genomic prediction ability in multienvironment trials. PLoS One 13:1–11. https://doi.org/10.1371/journal.pone.0201181

Meuwissen, T. H. (2009). Accuracy of breeding values of “unrelated” individuals predicted by dense SNP genotyping. *Genetics Selection Evolution*, *41*(1), 1–9. https://doi.org/10.1186/1297-9686-41-35

Moghaddar, N., Khansefid, M., Van Der Werf, J. H. J., Bolormaa, S., Duijvesteijn, N., Clark, S. A., Swan, A. A., Daetwyler, H. D., & MacLeod, I. M. (2019). Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. *Genetics Selection Evolution*, *51*(1), 1–14. https://doi.org/10.1186/s12711-019-0514-2

Moll P, Ante M, Seitz A, Reda T (2014) QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods 11:i–iii. https://doi.org/10.1038/nmeth.f.376

Montesinos-López OA, Montesinos-López A, Crossa J, et al (2016) A genomic bayesian multi-trait and multi-environment model. G3 Genes, Genomes, Genet 6:2725–2774. https://doi.org/10.1534/g3.116.032359

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290. https://doi.org/10.1093/bioinformatics/btg412

PepsiCo (2020) Avena sativa – OT3098 v1. https://wheat.pw.usda.gov/GG3/graingenes_downloads/oat-ot3098-pepsico)

Pérez P, De Los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. https://doi.org/10.1534/genetics.114.164442

Price N, Moyers BT, Lopez L, et al (2018) Combining population genomics and fitness QTLs to identify the genetics of local adaptation in Arabidopsis thaliana. Proc Natl Acad Sci U S A 115:5028–5033. https://doi.org/10.1073/pnas.1719998115

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, et al (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217–220. https://doi.org/10.1038/ng.1033

Runcie D, Cheng H (2019) Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. G3 Genes, Genomes, Genet 9:3727–3741. https://doi.org/10.1534/g3.119.400598

Schrag TA, Westhues M, Schipprack W, et al (2018) Beyond genomic prediction: Combining different types of omics data can improve prediction of hybrid performance in maize. Genetics 208:1373–1385. https://doi.org/10.1534/genetics.117.300374

USDA. (2019) Grain : World Markets and Trade Competitive Pricing Suggests Rebound in EU Wheat Exports

Wang S, Wei J, Li R, et al (2019) Identification of optimal prediction models using multi-omic data for selecting hybrid rice. Heredity (Edinb) 123:395–406. https://doi.org/10.1038/s41437-019-0210-6

503 Westhues M, Schrag TA, Heuer C, et al (2017) Omics-based hybrid prediction in maize. Theor
504 Appl Genet 130:1927–1939. https://doi.org/10.1007/s00122-017-2934-0

505 Xu Y, Xu C, Xu S (2017) Prediction and association mapping of agronomic traits in maize using
506 multiple omic data. Heredity (Edinb) 119:174–184. https://doi.org/10.1038/hdy.2017.27

507 Xu Y, Zhao Y, Wang X, et al (2021) Incorporation of parental phenotypic data into multi-omic
508 models improves prediction of yield-related traits in hybrid rice. Plant Biotechnol J 19:261–
509 272. https://doi.org/10.1111/pbi.13458

510 Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in*
511 *Bioinformatics*, *69*(1), 1–18. https://doi.org/10.1002/cpbi.96

512 Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network
513 analysis. Statistical Applications in Genetics and Molecular Biology, 4(1).
514 https://doi.org/10.2202/1544-6115.1128

# Figures

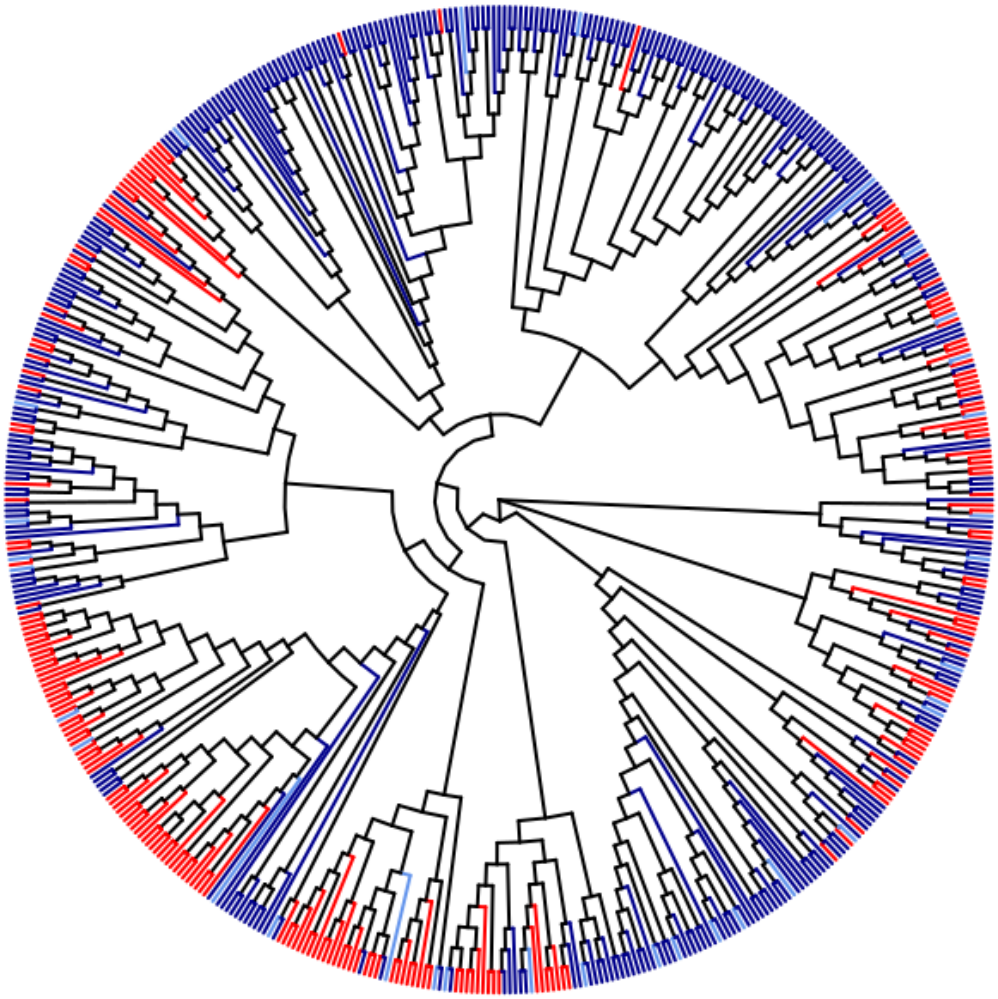
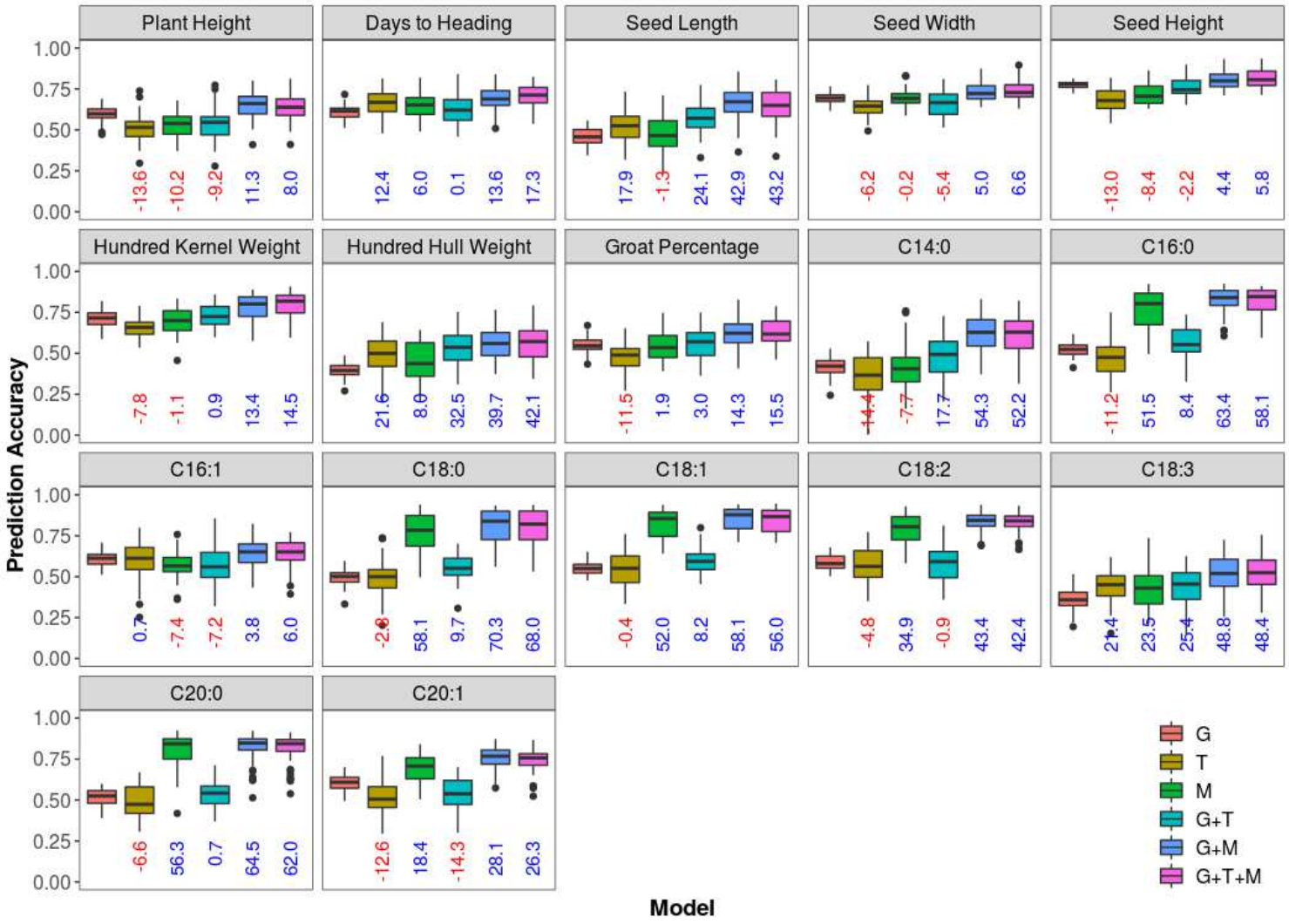


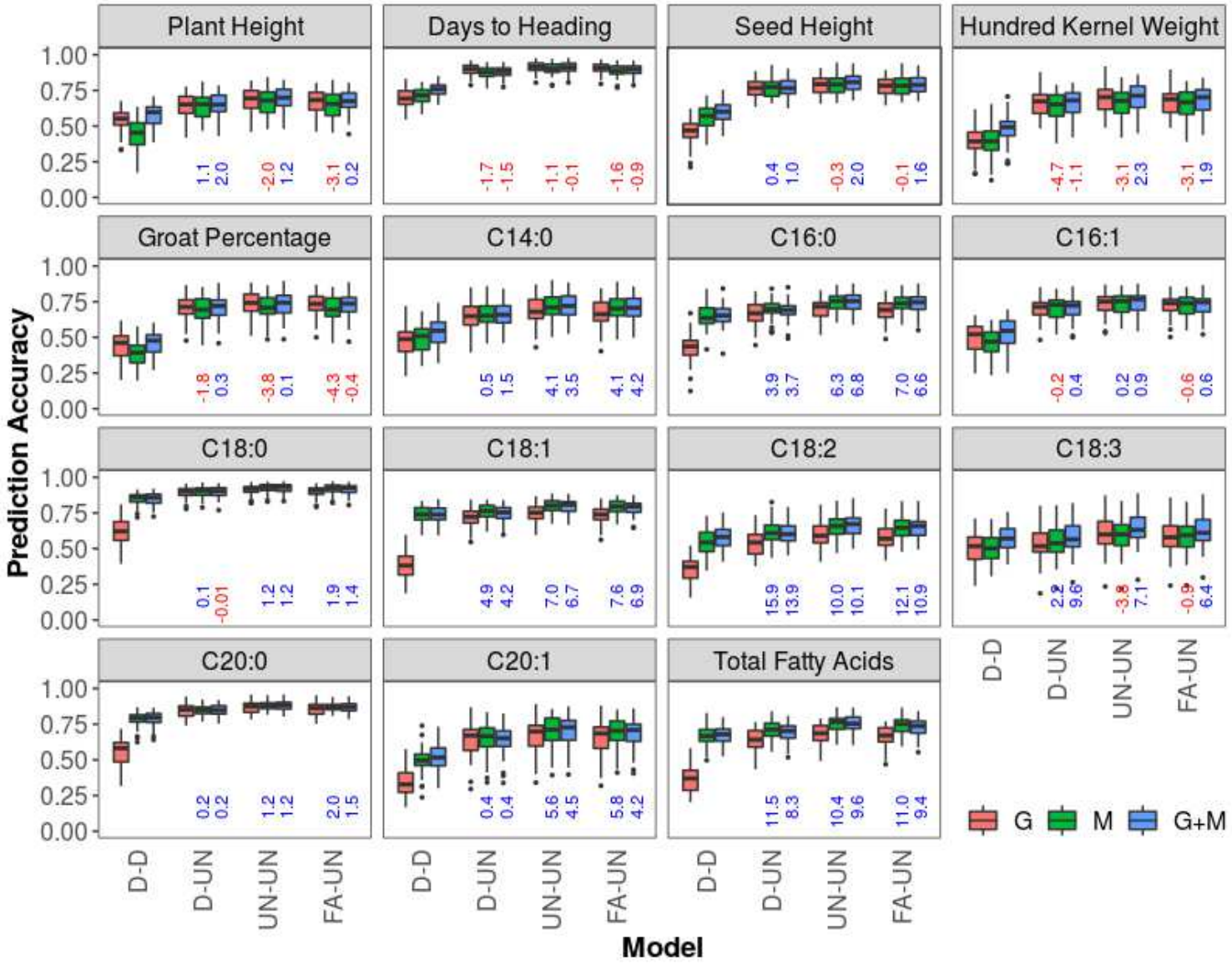
**Figure 1**

Neighbor-joining tree of 568 oat lines in the Diversity and Elite panels. Different panels are shown in different colors (darkblue, Diversity panel; red, Elite panel, light blue, lines in common)




## Figure 2

Distribution of prediction accuracy of the 17 phenotypic traits in the Diversity panel across 50 re-sampling runs. For each trait, boxplots with different colors represent prediction models, which are G, T, M, G+T, G+M and G+T+M from left to right. Medians of percent change in prediction accuracy of models relative to GBLUP are indicated below each box in blue if positive and in red if negative. G = genomic BLUP, T = transcriptomic BLUP, M = metabolomic BLUP.




## Figure 3

Distribution of prediction accuracy of the 15 phenotypic traits in the Elite panel across 50 re-sampling runs estimated by multi-trait models for multi-environment prediction. The 15 phenotypic traits in the Elite panel were evaluated at three environments. For each trait, boxplots with different colors represent models. Medians of percent change in prediction accuracy of M and G+M models relative to the G model are indicated below each box in blue if positive and in red if negative. For each model, the uppercase letters before and after the hyphen represent genetic and residual covariance structures: D=diagonal, UN=unstructured, FA=factor-analytic.

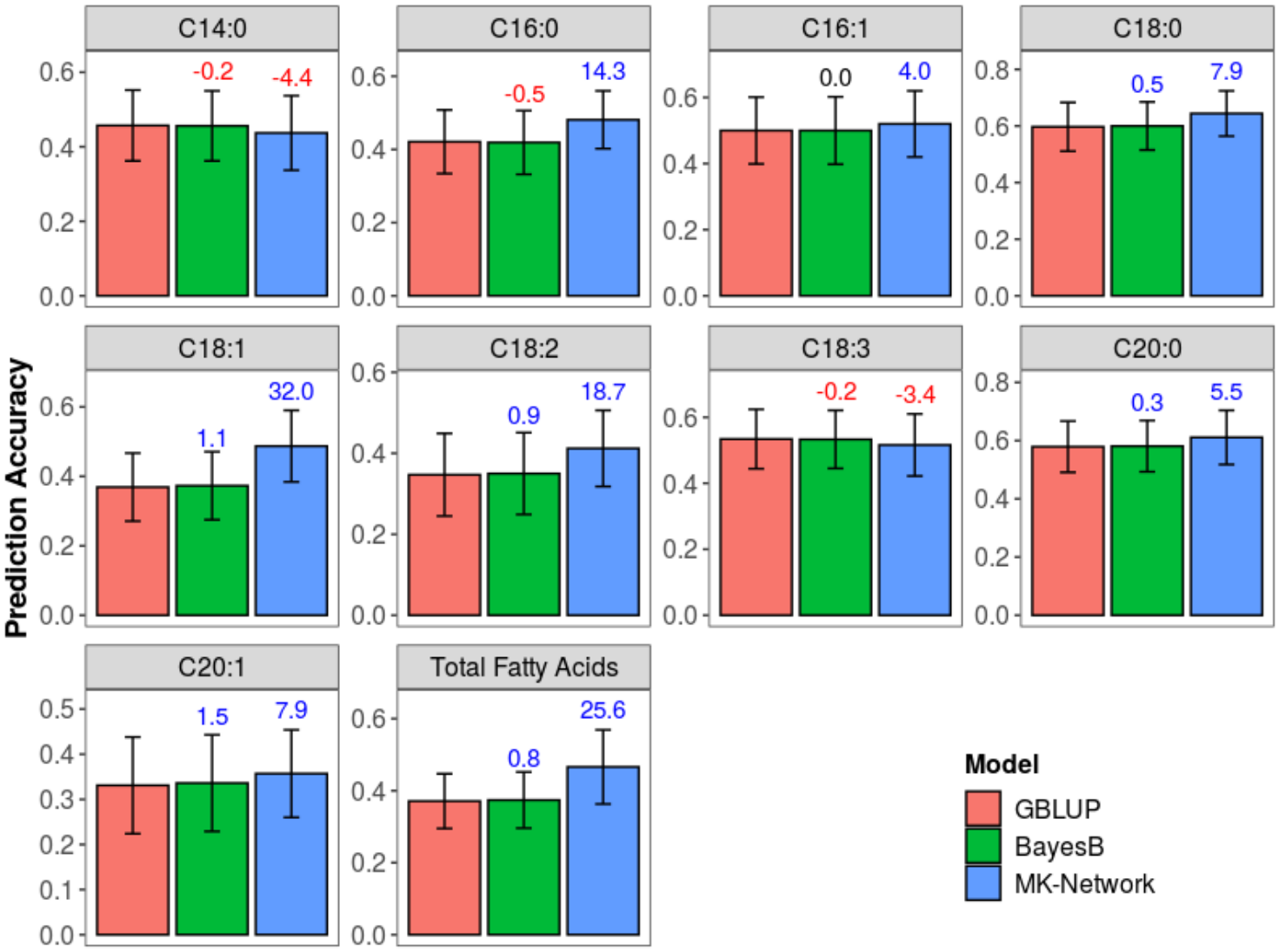



**Figure 4**

Prediction accuracy of the 10 fatty acid traits in the Elite panel estimated by GBLUP, BayesB and two-kernel BLUP models across 50 re-sampling runs. For each trait, barplots with different colors represent models. Means of percent change in prediction accuracy of all other models relative to GBLUP are indicated above each bar (in blue if positive, in red if negative, and in black if zero). MK-Network=network-based multiple-kernel prediction.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 03OatMultiomicsPredSI20210531.pdf