Roddur, Q., Qu, J., Liang, D., Carrara, E., Lourenco, D., Valente, B., Chen, C.-Y., Holl, J., Cheng, H. (2026) Integrating longitudinal metabolite profiles improves trait prediction in pigs in a trait- and timepoint-dependent manner, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23928: https://doi.org/10.31274/wcgalp.23928 ©2026 Roddur, et al. CC BY-NC-ND

Integrating longitudinal metabolite profiles improves trait prediction in pigs in a trait- and timepoint-dependent manner

Q.A.H. Roddur[1*], J. Qu[1], D. Liang[1], E.R. Carrara[2], D. Lourenco[2], B. Valente[3], C.-Y. Chen[3], J. Holl[3], and H. Cheng[1]

[1]University of California, Davis, 1 Shields Ave, Davis, CA 95616, USA; [2]University of Georgia, 100 Green St, Athens, GA 30602, USA; [3]PIC North America, 100 Bluegrass Commons Blvd, Ste 2200, Hendersonville, TN 37075, USA; *[*]Corresponding author: qroddur@ucdavis.edu*

Introduction

Accurate prediction of phenotypic performance is essential for improving selection and management decisions in livestock production, thus accelerating productive performance (Forutan *et al*., 2023; Yin & König, 2016). Approaches that leverage genotypes and phenotypes from relatives have become standard practice in animal breeding (Mrode *et al*., 2019; Wiggans *et al*., 2011) since Meuwissen *et al*. (2001) proposed genomic selection using simulated data. In pig breeding, traits of economic importance, such as feed efficiency, growth, and reproduction, are often expensive, labor-intensive, or challenging to measure directly in selection candidates (Guan *et al*., 2024; Delpuech *et al*., 2021). Therefore, there is considerable interest in identifying biologically informative intermediate phenotypes that can be reliably measured and integrated into selective breeding programs (Legarra & Christensen, 2023; Goldansaz *et al*., 2017; Fontanesi, 2016).

Metabolite-profile data offers a dynamic snapshot of an organism's physiological and biochemical state (Fiehn, 2002). Blood metabolites are influenced by both genetic and environmental factors and have shown promise as intermediate phenotypes (Jang *et al*., 2019). Integration of metabolite-profile data into genetic evaluations has been explored in barley (Guo *et al*., 2023) and in pigs for average daily gain (Guo *et al*., 2025), yielding small increases in prediction accuracy. However, metabolite sampling time is crucial (Bromage *et al*., 2016), and the effects of metabolites on phenotypes may vary across traits (Aliakbari *et al*., 2019).

Our study aimed to address the temporal dynamics in omics-based prediction by analyzing metabolite-profile data collected at 10-week and 20-week of age. We evaluated five economically important traits for pigs: average daily feed intake (DFI), feed conversion ratio (FCR), backfat thickness (BF), test daily gain (TDG), and loin depth (LD). We hypothesize that integrating individual-level metabolite-profile data collected at multiple growth stages into prediction models will improve phenotype prediction accuracy compared to using genomic data alone, and that the degree of improvement will be timepoint and trait dependent.

Materials & Methods

Dataset. Metabolite-profile data were provided by PIC (a Genus company, Hendersonville, TN, USA). Blood samples were taken from 2,101 pigs from a single purebred population at both 10-week and 20-week of age and analyzed using 1H NMR (Proton Nuclear Magnetic Resonance) spectroscopy, generating individual metabolite-profile data for 149 metabolites. Blood samples were stored at −80 °C. The 1H NMR spectra were acquired at 310 K on a Bruker AVANCE III 600 MHz IVDr certified spectrometer equipped with a BBI probe, automatically preprocessed using Bruker software. The traits analyzed were: DFI, FCR, TDG (measured from 10-week to 20-weeks span), BF, and LD (measured at 20-week). All genotypes were imputed to a medium-density panel of 37k markers using AlphaImpute (Hickey *et al*., 2012).

For quality control (QC), metabolites with more than 20% missing data were excluded. This resulted in a final dataset of 95 metabolites. QC for genotype data included filtering out SNPs with a minor allele frequency (MAF) below 0.01, resulting in 37,199 SNPs retained for analysis, and no animals exceeded the 5% genotype missingness threshold used for exclusion. Finally, the dataset was further restricted to 1,637 individuals with complete 10-week and 20-week metabolite profiles, performance trait records, and genotype data.

Statistical Methods. For each trait, we evaluated seven linear mixed models (G, M1, M2, M1+M2, G+M1, G+M2, and G+M1+M2); where G is the genotype-only model, M1 and M2 are metabolite-only models for 10-week and 20-week metabolite profiles, respectively, and the rest are their combinations. Prior distributions of regression coefficients were specified according to the BayesC method, using two separate sets of hyperparameter values for genotype and metabolite profiles (Wang & Cheng, 2021). Eq.1 is when only genotypes were included (conventional BayesC model):

$$\mathbf{y} = \mathbf{W\theta} + \mathbf{X\beta} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the $n \times 1$ phenotype vector (n = 1,637 individuals); $\boldsymbol{\theta}$ is the vector of systematic effects and $\mathbf{W}$ is its corresponding design matrix; $\mathbf{X}$ is the $n \times p$ genotype matrix (p = 37,199 SNPs); $\boldsymbol{\beta}$ is the $p \times 1$ vector of marker effects; and $\mathbf{e}$ is the $n \times 1$ vector of residuals with the distribution of $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$.

In the BayesC framework, marker-effect estimation is based on a biologically meaningful assumption that only a subset of all genomic variants has non-zero effects, while the majority do not influence the phenotype. The model, therefore, infers a null-effect probability, π, which reflects the proportion of SNPs whose effects are zero. For the remaining SNPs, their effect magnitudes are assumed to follow a normal distribution with mean 0 and variance σ_β^2. Thus, the effect of SNP j is modeled as a mixture distribution comprising a point mass at zero and a normal distribution as expressed as:

$$\beta_j = \begin{cases} \mathrm{N}\left(0, \sigma_\beta^2\right) & \text{with probability } (1 - \pi) \\ 0 & \text{with probability } (\pi) \end{cases}$$

Metabolite-profile data were included as additional random regressors. Eq.2 is when metabolites from 10-week and 20-week were analyzed together:

$$\mathbf{y} = \mathbf{W\theta} + \mathbf{X\beta} + \mathbf{Z_1 u_1} + \mathbf{Z_2 u_2} + \mathbf{e} \quad (2)$$

where variables $\mathbf{y}, \mathbf{W\theta}, \mathbf{X\beta}$, and $\mathbf{e}$ are the same as Eq.1, $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are the matrices of metabolites from weeks 10 and 20, respectively, where each matrix entry in the i^{th} row and j^{th} column represents the j^{th} metabolite feature for individual i, and $\mathbf{u_1}$ and $\mathbf{u_2}$ are their corresponding effect vectors. For DFI and TDG, litter was included as a random effect and birth year-month was included as a systematic effect, whereas for FCR, BF, and LD, only litter was included as a random effect. The choice of including random and systematic effects in the models for each trait was based on their optimal prediction performance in the baseline genotype-only (G) model. Prediction accuracy was calculated as the Pearson correlation between estimated values and adjusted phenotypes in the testing set. The full dataset was randomly split into training (80%) and testing (20%) sets with 20 replications. Model parameters were estimated from the training set and predictions were generated for the testing set. No parent-offspring relationships were present, but related individuals may have appeared in both the training and testing sets.

Results

Figure 1 shows the Monte Carlo cross-validation distribution of predictive accuracies for all models relative to G, used as the baseline. Each point represents, for one of 20 random train/test splits, the differences in prediction accuracy between the model indicated on the x-axis and the

genotype-only (G) baseline. The box plots represent the distributions of accuracy differences relative to G for all models and traits.

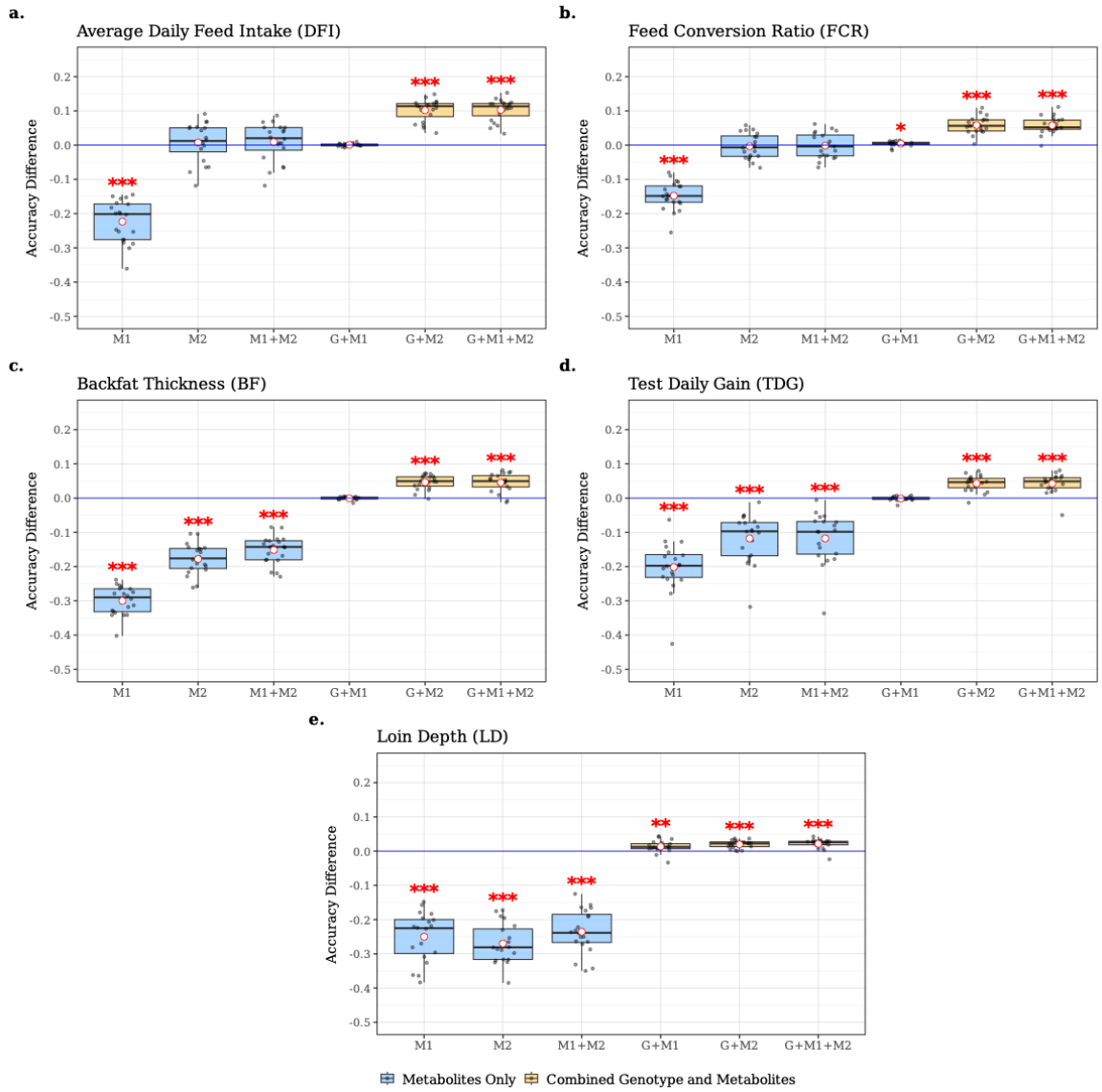

Figure 1 Box plots showing the differences in prediction accuracy across various models relative to the genotype-only baseline (Model G) for five traits: **a.** DFI, **b.** FCR, **c.** BF, **d.** TDG, and **e.** LD. Each plot represents the difference in prediction accuracy between the genotype-only model and metabolites-only models (blue boxes) or combined genotype and metabolites models (yellow boxes). Statistical significance is indicated by * (*p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001).

All metabolites-only models (M1, M2, and M1+M2) showed notably lower ($p<0.001$) prediction accuracy across traits compared to G model, except for traits DFI and FCR, where the prediction accuracy for the M2 and the M1+M2 models were similar ($p>0.05$) to the G model. Furthermore, M2 showed a clear superiority as a single timepoint across traits over M1, except for LD. Also, M1+M2 model mostly yielded higher prediction accuracies than only M1 or only M2 model, with some exceptions. Further, the accuracy for G+M2 was consistently increased ($p<0.001$) across all five traits. However, G+M1 did not improve prediction accuracy ($p>0.05$) over the genotype-only model for DFI, BF, and TDG, whereas it provided noticeable gains for FCR ($p<0.05$) and LD ($p<0.01$). Combining metabolite-profile data from both

timepoints with genotype data (G+M1+M2) achieved a significant overall improvement in accuracy ($p<0.001$) relative to the G model across all five traits.

Discussion

This study evaluates the potential of integrating individual-level metabolite-profile data from multiple timepoints for five economically important production traits in pigs. Our findings confirm that combining metabolite-profile data with genotype data led to consistent improvements in prediction accuracy across all five traits, which is aligned with Guo *et al*. (2025), who reported a small increase in prediction accuracy for average daily gain in pigs while integrating metabolite-profile data with genotype. Also, the biological relevance of incorporating metabolite-profile data in phenotypic prediction is trait dependent. For traits like BF and TDG, the inclusion of late-stage metabolite-profile data (20-week), with or without the early-stage metabolite-profile data (10-week) provides further useful signals when combined with genotype data. DFI, FCR, and LD, being influenced by both early and late biological stages, benefit from the integration of metabolite-profile data from multiple timepoints. Furthermore, for LD, the highest performance was achieved when combining both timepoint metabolite-profile data with genotypes. These suggest that late-stage metabolite profile data were more informative than early-stage metabolite profile data for predicting these traits.

Conclusion

This study demonstrates that integrating individual-level metabolite-profile data with genotype data can improve the prediction accuracy for some key production traits in pigs. While metabolite-profile data alone yields relatively low predictive power in most scenarios, their combination with genotype data consistently enhances prediction accuracy across all five studied traits. By integrating this omics data, prediction models gain access to non-redundant biological information not captured by genotypes alone.

References

Aliakbari A, Ehsani A, Vaez Torshizi R, *et al*. (2019) J Anim Sci. 97:3832-44. https://doi:10.1093/jas/skz228.

Bromage TG, Idaghdour Y, Lacruz RS, *et al*.(2016) PLoS One.11(1):e0145919. https://doi:10.1371/journal.pone.0145919.

Delpuech E, Aliakbari A, Labrune Y, *et al*. (2002) Genet Sel Evol 53:49. https://doi:10.1186/s12711-021-00642-1.

Fiehn O. (2002) Plant Mol Biol 48:155-71.

Fontanesi L. (2016) Anim Front 6(1):73-79. https://doi:10.2527/af.2016-0011.

Forutan M, Lynn A, Aliloo H, *et al*. (2023) Front Genet 14:1089490. https://doi:10.3389/fgene.2023.1089490.

Goldansaz SA, Guo AC, Sajed T, (2017) PLoS One 12(5):e0177675. https://doi:10.1371/journal.pone.0177675.

Guan R, Li Y, Hu Z, *et al*. (2024) Sci Rep 14:28222. https://doi:10.1038/s41598-024-78093-z.

Guo X, Sarup P, Bay Nord A, *et al*. (2025) Genet Sel Evol 57:24. https://doi:10.1186/s12711-025-00972-4.

Guo X, Sarup P, Jahoor A, *et al*. (2023) Genet Sel Evol. 55:61. https://doi:10.1186/s12711-023-00835-w.

Hickey JM, Kinghorn BP, Tier B, *et al*. (2012) Genet Sel Evol 44:9. https://doi:10.1186/1297-9686-44-9.

Jang C, Hui S, Zeng X, *et al*. (2019) Cell Metab 30:594-606.e3. https://doi:10.1016/j.cmet.2019.06.002.

Legarra A, and Christensen OF. (2023) JDS Commun 4(1):55-60. https://doi:10.3168/jdsc.2022-0276.

Meuwissen TH, Hayes BJ, and Goddard ME. (2001) Genetics 157:1819-29. https://doi:10.1093/genetics/157.4.1819.

Mrode R, Ojango JMK, Okeyo AM, *et al*. (2019) Front Genet 9:694. https://doi:10.3389/fgene.2018.00694

Wiggans GR, VanRaden PM, and Cooper TA. (2011) J Dairy Sci 94(6):3202-11. https://doi:10.3168/jds.2010-3866.

X.Wang Z, and Cheng H. (2021) Front Genet 12:717457. https://doi:10.3389/fgene.2021.717457.

Yin T, and König S. (2016) Animal Frontiers 6(1):65-72. https://doi:10.2527/af.2016-0010.