A Bayesian model for genomic prediction using metabolic networks

Akio Onogi

Department of Plant Life Science, Faculty of Agriculture, Ryukoku University, Otsu, Shiga, 520-2194, Japan

Akio Onogi

Postal address: 1-5, Yokotani, Oe-cho, Seta, Otsu, Shiga, 520-2194, Japan

Phone: +81-(0)77-599-5719

E-mail: onogiakio@gmail.com

Running head: Metabolic networks

Abstract

Motivation

Genomic prediction is now an indispensable technique in breeding and medical fields, and it is of interest how to use omics data to enhance prediction accuracy. A precedent work proposed a metabolic network-based method in biomass prediction of Arabidopsis, however, the method consists of multiple steps which possibly degrade prediction accuracy

Results

We proposed a Bayesian model that integrates all steps and jointly infers all fluxes of reactions related to biomass production. The proposed model showed higher accuracies than methods compared both in simulated and real data. The results support the excellent idea of the precedent study that the metabolic network information can be useful for prediction.

Availability and implementation

All R and stan scripts to reproduce the results of this study are available at https://github.com/Onogi/MetabolicModeling.

Contact

onogiakio@gmail.com

Supplementary information

This study provide no supplementary information

1. Introduction

Genomic prediction is a statistical approach that predicts phenotypes or genetic merits of genotypes using genome-wide DNA markers (Meuwissen *et al*., 2001). Genomic prediction is now indispensable in animal breeding (e.g., Garcia-Ruiz *et al*., 2016; Scott *et al*., 2021), and becoming important in plant breeding (e.g., Dreisigacker *et al*., 2021). In breeding, selection strategies using genomic prediction is often referred to as genomic selection. The framework of genomic prediction is also utilized in medical to predict disease risks which are known as polygenic or genetic risk scores (Wray *et al*., 2019; Lello *et al*., 2020).

One of the main purposes of studies on genomic prediction has been to increase prediction accuracy (e.g., Heslot *et al*., 2012; Onogi *et al*., 2015; Dreisigacker *et al*., 2021). This interest has motivated various statistical methods and models including Bayesian regressions (reviewed in de los Campos *et al*., 2013; Dreisigacker *et al*., 2021). It also has been explored how to utilize information other than genome to enhance prediction accuracy. The typical additional information is “omics data”. For example, transcriptome (Li *et al*., 2019; Perez *et al*., 2022) and metabolome (Riedelsheimer *et al*., 2012; Campbell *et al*., 2021) are often used together with genome, and some studies used both (Xu *et al*., 2016; Schrag *et al*., 2018). In addition, because data in biology is essentially multivariate (i.e., multiple traits can be measured for a genotype at multiple environments), it also has been of interests how to utilize multivariate information to increase prediction accuracy for target traits (e.g., Jia and Jannink, 2012; Jarquin *et al*., 2016; Atanda *et al*., 2022).

Upon such trends, Tong *et al*., (2020) proposed utilizing metabolic networks to predict biomass of *Arabidopsis thaliana* (Arabidopsis). The approach is based on the concept of flux balance: fluxes (amount) of each reaction related to biomass production are determined such that production and consumption of each metabolite are balanced. Under this constraint, the flux of biomass production is estimated with quadratic programming to match the observed biomass. When the biomass of a new genotype is predicted, first, fluxes of all reactions and biomass production are predicted using ordinal genomic prediction methods, and then, the predicted fluxes are adjusted using quadratic programming to fill flux balance.

Although their results suggest the usefulness of metabolic network information in predicting phenotypes, their approaches still have room for improvement. First, because flux is estimated for each genotype independently, dependencies (or covariances/similarities) between genotypes are not utilized. Second, because flux estimation and prediction are conducted in separated steps, uncertainty in estimation increases noise when prediction. To circumvent these disadvantages, in the present study, a Bayesian model that estimates fluxes of all genotypes simultaneously and conducts estimation and prediction of fluxes jointly is proposed. The performance of the proposed model was compared with other methods including the approach of Tong *et al*., (2020). The results showed the superiority of the proposed model and support the excellent idea of Tong *et al*., (2020) that the metabolic network information can help phenotype prediction.

2. System and methods

Let $\mathbf{M}_i$ denote the stoichiometry matrix of genotype i (Fig. 1A). $\mathbf{M}_i$ is a K by J matrix that contains coefficients of reactions where K and J denote the number of metabolites and reactions, respectively. Let the J_{th} reaction be the reaction related to the target trait, and let the other reactions (reaction j where $1 \leq j < J$) represent processes in which metabolites relevant with the target trait are produced. Here, the target trait is biomass of Arabidopsis. Let $V_{i,j}$ denote the flux of reaction j of genotype i which represents the amount that the reaction occurs in cells/tissues of the genotype, and let $\mathbf{V}$ be an N by J matrix that contains all fluxes of all genotypes where N is the number of genotypes, and $\mathbf{V}_i$ is the i_{th} rows of $\mathbf{V}$ which denotes the fluxes of genotype i.

According to the principle of flux balance analysis, production and consumption are balanced for each metabolite. This can be written as an equation such as

$$\mathbf{M}_{i,k}\mathbf{V}_i^T = 0 \text{ for } 1 \leq i \leq N \text{ and } 1 \leq k \leq K \quad (1).$$

where $\mathbf{M}_{i,k}$ denotes the k_{th} row of $\mathbf{M}_i$. Eq. (1) is also illustrated in Fig. 1A. It is assumed that the flux of the J_{th} reaction (i.e., reaction for the target trait), $\mathbf{V}_{,J}$, is proportional to the observed phenotypes of the target trait, $\mathbf{Y}$. Here $\mathbf{Y}$ is an N-length vector that contains the phenotypes of all genotypes.

In short, the approach by Tong *et al*., (2020) is to estimate $\mathbf{V}_i$ for each genotype independently such that Eq. (1) is filled and the ratio between $V_{i,J}$ and $V_{Col-0,J}$, a flux of the reference genotype Columbia-0, is matched with the ratio between Y_i and $Y_{Col\text{-}0}$. (Fig. 1B). The flux estimated with this way is hereafter referred to as **Vt**. See Section 6.2 for the explanation of prediction by Tong *et al*., (2020).

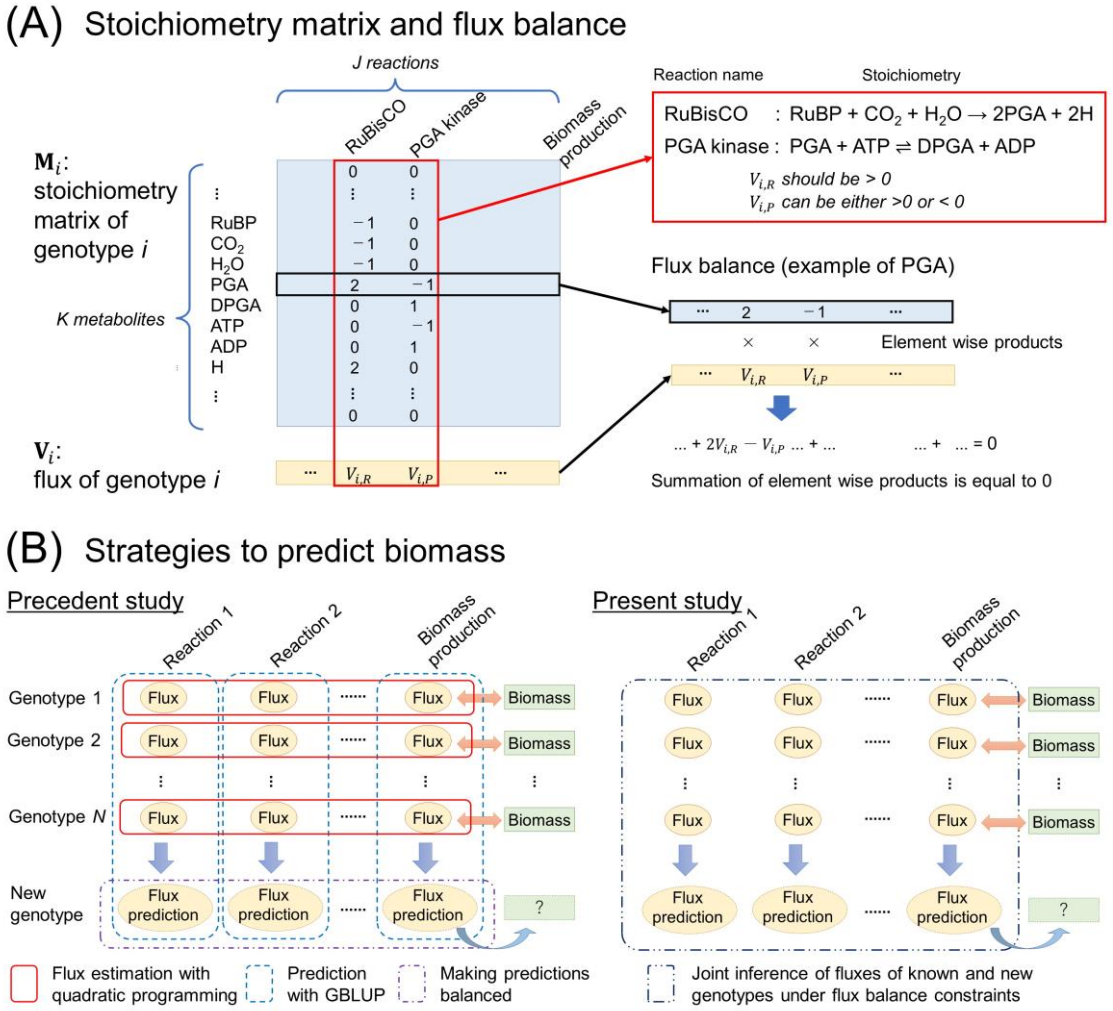

Fig. 1 Outline of the proposed model and analysis. (A) Illustration of stoichiometry matrix and flux balance. Here two reactions (RuBisCO and PGA kinase) and eight metabolites related to the reactions are shown. Flux balance analysis assumes that all metabolites in the stoichiometry matrix are balanced. Note that biomass itself is not included in the stoichiometry matrix. (B) Comparison of strategies between the preceding study (Tong *et al*., 2020) and the present study. The strategy of the precedent study largely consists of three steps: (1) Estimation of flux of each genotype under flux balance constraints using quadratic programming. (2) Prediction of fluxes of new genotypes using genomic BLUP (GBLUP). (3) Making predicted fluxes balanced using

quadratic programming and predicting biomass of new genotypes with the predicted flux of the biomass production. In contrast, the proposed model in the present study conducts these steps in one step using a Bayesian hierarchical model.

In the present study, $\mathbf{V}_{,J}$ is connected with $\mathbf{Y}$ as

$$\mathbf{Y} = \mathbf{1}a + b\mathbf{V}_{,J} + \mathbf{e} \quad (2)$$

where $\mathbf{1}$ indicates the vector of 1 and $\mathbf{e}$ is a gaussian noise, that is, $\mathbf{e} \sim N(\mathbf{0}, \quad \mathbf{I}\sigma_e^2)$ where $\mathbf{0}$ indicates the vector of 0 and $\mathbf{I}$ is an identity matrix.

The purpose of the proposed model is to estimate $\mathbf{V}_{,J}$ under the constraint of Eq. (1). To treat with the constraint in a Bayesian hierarchical model, Eq. (1) is rewritten as

$$0 = \mathbf{M}_{i,k}\mathbf{V}_i^T + \varepsilon_{ik} \text{ for } 1 \leq i \leq N \text{ and } 1 \leq k \leq K \quad (3)$$

where $\varepsilon_{ik} \sim N(0, \quad \sigma_\varepsilon^2)$. Eqs. (2) and (3) constitute the likelihood of the model. The prior distributions of σ_e^2 and σ_ε^2 are half-Cauchy distributions written as,

$$\sigma_e^2 \sim HC(0, \quad 0.1),$$

$$\sigma_\varepsilon^2 \sim HC(0, \quad S_\varepsilon),$$

$$S_\varepsilon \sim Gamma(s_{gamma}, \quad r_{gamma})$$

respectively. Herein, s_{gamma} and s_{gamma} denote the shape and rate parameters set to 0.1 and 1.0, respectively.

Positive values of flux ($\mathbf{V}$) mean that the reaction flows from left to right hand whereas negative values mean the opposite direction. Some reactions can occur in both directions whereas some only occur in either direction (Fig. 1A). Let J_n denote the set of reactions that allows both directions, and J_r denote the set of reactions with positive restriction on flux. Then the prior distributions of $\mathbf{V}$ are written as,

$$\mathbf{V}_{,j} \sim N\left(\mathbf{U}_{,j}, \quad \mathbf{I}\sigma_{V_j}^2\right) \text{ for } j \in J_n$$

$$\mathbf{V}_{,j} \sim TruncN\left(0, \quad \mathbf{U}_{,j}, \quad \mathbf{I}\sigma_{V_j}^2\right) \text{ for } j \in J_r$$

where $\mathbf{U}_{,j}$ defines the genetic component of $\mathbf{V}_{,j}$., and *TruncN* denotes truncated normal distributions. The prior distribution of $\mathbf{U}_{,j}$ is written as

$$\mathbf{U}_{,j} \sim N\left(\mathbf{1}\mu_j, \quad \mathbf{G}\sigma_{U_j}^2\right)$$

where μ_j is a known mean value described later and $\mathbf{G}$ is a genomic relationship matrix which is defined with SNPs.

The unobserved phenotype Y_i of new genotype *i* can be predicted as

$$\widehat{Y_i} = a + bV_{i,J}.$$

Herein, $V_{i,J}$ can be estimated from the constraints of Eq. (3) even when the phenotype Y_i is unavailable, and from $U_{i,J}$ which can be predicted with **G**. The outline of the proposed model is illustrated with the precedent approach by Tong *et al*., (2020) in Fig. 1B.

Some fluxes may be constrained by other fluxes based on prior physiological knowledge. For example, in the biomass prediction proposed by Tong *et al*., (2020), it was supposed that

$$0.94V_{i,oxg} \leq V_{i,car} \leq 3.81V_{i,oxg} \text{ for } 1 \leq i \leq N \quad (4) \text{ and}$$

$$0.79V_{i,suc} \leq V_{i,sta} \leq 3.37V_{i,suc} \text{ for } 1 \leq i \leq N \quad (5)$$

where $V_{i,oxg}$, $V_{i,car}$, $V_{i,suc}$, and $V_{i,sta}$ denote the fluxes of oxygenation, carboxylation, sucrose synthesis, and starch synthesis reactions, respectively. Hereafter, such constraints are referred to as “additional constraints”. These additional constraints, however, resulted in spurious predictions in the case of biomass in Arabidopsis. This issue is mentioned in the Results and Discussions

3. Algorithm and implementation

All statistical inference was conducted with R (ver. 4.1.3 on Windows machines or ver. 4.2.2 on Ubuntu 18.4 machine) (R Core Team, 2022). Posterior distributions of the proposed models were inferred with Hamiltonian Monte Carlo implemented by rstan (ver. 2.21.5 or 2.21.7). The number of iterations was 5000, the warm-up length was 4000, and thinning was 1. No-U-Turn Sampler (NUTS) was used. Because both $\mathbf{V}$ and $\mathbf{U}$ are high dimensional (N by J matrices), to facilitate exploration of posterior distribution, two modifications were made from the base model defined above when implementing the model using stan. These modifications are illustrated in subsections 3.1 and 3.2.

3.1. Scale and location

The scales ($\sigma^2_{V_j}$) and locations (μ_j) of flux are considerably different among reactions which can hamper the efficient exploration of posterior distributions. Thus, the scales and locations were adjusted to be similar among reactions using prior knowledge.

Suppose that rough estimates for $\sigma_{V_j}^2$, $\sigma_{U_j}^2$, and μ_j, $\widetilde{\sigma_{V_j}^2}$, $\widetilde{\sigma_{U_j}^2}$, and $\widetilde{\mu_j}$, can be obtained from literature and/or preceding experiments. Define δ_j as

$$\delta_j = \sqrt{\widetilde{\sigma_{U_j}^2} + \widetilde{\sigma_{V_j}^2}},$$

then divide $\mathbf{V}$ and $\mathbf{U}$ with δ_j as $\mathbf{V}_{,j}{}' = \mathbf{V}_{,j}/\delta_j$ and $\mathbf{U}_{,j}{}' = \mathbf{U}_{,j}/\delta_j$ yielding prior distributions of $\mathbf{V}_{,j}{}'$ and $\mathbf{U}_{,j}{}'$ represented as

$$\mathbf{U}_{,j}{}' \sim N\left(\mathbf{1}\frac{\mu_j}{\delta_j},\ \ \mathbf{G}\sigma_{U_j}^{2\,\prime}\right),$$

$$\mathbf{V}_{,j}{}' \sim N\left(\mathbf{U}_{,j}{}',\ \ \mathbf{I}\sigma_{V_j}^{2\,\prime}\right) \text{ for } j \in J_n, \text{ and}$$

$$\mathbf{V}_{,j}{}' \sim TruncN\left(0,\ \ \mathbf{U}_{,j}^{\prime},\ \ \mathbf{I}\sigma_{V_j}^{2\,\prime}\right) \text{ for } j \in J_r$$

where $\sigma_{U_j}^{2\,\prime}$ and $\sigma_{V_j}^{2\,\prime}$ correspond with $\sigma_{U_j}^2/\delta_j^2$ and $\sigma_{V_j}^2/\delta_j^2$, respectively. The location of $\mathbf{V}_{,j}{}'$ was then arbitrary adjusted to 10 using $\widetilde{\mu_j}$ as

$$\mathbf{V}_{,j}{}'{}' \sim N\left(\mathbf{U}_{,j}^{\prime} + \alpha_j,\ \ \mathbf{I}\sigma_{V_j}^{2\,\prime}\right) \text{ for } j \in J_n$$

$$\mathbf{V}_{,j}{}'{}' \sim TruncN\left(\alpha_j,\ \ \mathbf{U}_{,j}^{\prime} + \alpha_j,\ \ \mathbf{I}\sigma_{V_j}^{2\,\prime}\right) \text{ for } j \in J_r.$$

where $\alpha_j = 10 - \widetilde{\mu_j}/\delta_j$.

Inference was conducted using $\mathbf{V}_{,j}{}'{}'$ and $\mathbf{U}_{,j}{}'$. The constraint of Eq. (1) becomes

$$\mathbf{M}_{i,k}\mathbf{V}_i^T = \sum_{j=1}^{J} M_{i,k,j} V_{i,j} = \sum_{j=1}^{J} M_{i,k,j}\delta_j\left(V_{i,j}^{\prime\prime} - \alpha_j\right) = 0,$$

that is,

$$\sum_{j=1}^{J} M_{i,k,j}\delta_j V_{i,j}^{\prime\prime} = \sum_{j=1}^{J} M_{i,k,j}\delta_j\alpha_j.$$

Thus, Eq. (3) can be written as

$$\sum_{j=1}^{J} M_{i,k,j}\delta_j\alpha_j = \sum_{j=1}^{J} M_{i,k,j}\delta_j V_{i,j}^{\prime\prime} + \varepsilon_{ik} \text{ for } 1 \le i \le N \text{ and } 1 \le k \le K \quad (6)$$

where $\varepsilon_{ik} \sim N(0,\ \ \sigma_\varepsilon^2)$. Eq. (2) is rewritten using $\mathbf{V}_{,J}{}'{}'$ as

$$\mathbf{Y} = \mathbf{1}a' + b'\,\mathbf{V}_{,J}{}'{}' + \mathbf{e} \quad (7)$$

The prior distributions of a', b', and variances are half-Cauchy distributions written as,

$$a' \sim HC(0, \quad 10),$$

$$b' \sim HC(0, \quad 10),$$

$$\sigma_{V_j}^{2\,\prime} \sim HC(0, \quad 10), \text{ and}$$

$$\sigma_{U_j}^{2\,\prime} \sim HC(0, \quad 10000),$$

respectively.

3.2. Weight in likelihood

Eqs. (6) and (7), terms for observations and constraints, respectively, define the likelihood of the model. However, naïve combination of these two terms in the log likelihood function can result in underfitting to **Y** because of the difference in dimensions (the observations consist of N components whereas the constraints consist of $N \times K$ components). Thus, a parameter to control the relative weight between the observations and constraint was introduced to the log likelihood:

$$L \propto -\frac{N}{2}\log\sigma_e^2 - \frac{1}{2\sigma_e^2}\sum_{i=1}^{N}\left(Y_i - a' - b' V_{i,J}''\right)^2 + W\left[-\frac{NK}{2}\log\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{J} M_{i,k,j}\delta_j\alpha_j - \sum_{j=1}^{J} M_{i,k,j}\delta_j V_{i,j}''\right)^2\right] \tag{8}$$

where W is the weight parameter. The effect of W was investigated using simulated and real data.

4. Real data analysis

The real data used in this study was the same as the data used in Tong *et al*., (2020). The target trait is the biomass of *Arabidopsis thaliana* evaluated under the 12h-light optimal nitrogen condition (Sulpice *et al*., 2013). The number of genotypes (N) was 67 including Columbia-0. The stoichiometry matrix (**M**) of the reference genotype relevant with biomass was developed by Arnold and Nikoloski (2014). The matrix originally contains 407 metabolites and 549 reactions. By removing reactions with zero flux, a stoichiometry matrix with 350 metabolites and 336 reactions was obtained (i.e., $K =$

350 and $J = 336$). The matrix was sparse; the medians of the number of metabolites per reaction and the number of reactions per metabolite were 4 and 2, respectively. The J_{th} column of the stoichiometry matrix has coefficients unique to genotypes, and the coefficients for the genotypes other than the reference were estimated in Tong *et al*., (2020). Genome-wide SNPs of the 67 genotypes were genotyped by Horton *et al*., (2012), and the genomic relationship matrix **G** was calculated from exome SNPs. All real data including **G** are available at https://github.com/Hao-Tong/netGS.

Prediction ability of the proposed model was assessed with 3-fold cross validation (CV). The divisions of genotypes (N = 67) followed those used in Tong *et al*., (2020). This precedent study conducted 3-fold CV 50 times and open the genotype divisions of all CVs at the above-mentioned site. In the present study, 3-fold CV was conducted 20 times following the first 20 CVs of Tong *et al*., (2020). R package snow (ver. 0.4-4) was used to parallelize CVs.

Modifying scales and locations as described in subsection 3.1 requires rough estimates for $\sigma^2_{V_j}$, $\sigma^2_{U_j}$, and μ_j (i.e., $\widetilde{\sigma^2_{V_j}}$, $\widetilde{\sigma^2_{U_j}}$, and $\widetilde{\mu_j}$). To obtain these estimates, the following linear model was fitted to the training data:

$$\mathbf{Vt}_{,j} = \mathbf{1}\mu_j + \mathbf{U}_{,j} + \mathbf{r}_j \quad (9)$$

where $\mathbf{Vt}_{,j}$ is the fluxes of reaction j estimated by Tong *et al*., (2020), $\mathbf{U}_{,j} \sim N\left(\mathbf{0},\ \mathbf{G}\sigma^2_{U_j}\right)$, and $\mathbf{r}_j$ is the residual following $\mathbf{r}_j \sim N\left(\mathbf{0},\ \mathbf{I}\sigma^2_{V_j}\right)$. Note that $\mathbf{Vt}_{,j}$ was estimated using quadratic programming for each genotype without information of SNP genotypes. It is also notable that, in this real data analysis, $\mathbf{Vt}_{,j}$ only included the fluxes of genotypes in the training data and those in the testing data were not used in this estimation. The solutions of $\sigma^2_{U_j}$, $\sigma^2_{V_j}$, and μ_j were obtained using an R package rrBLUP ver. 4.6.1 (Endelman JB, 2011).

The unobserved phenotype Y_i of new genotype i can be predicted as

$$\widehat{Y_i} = a' + b' V'_{i,J}.$$

Prediction accuracy was evaluated using the Pearson correlation coefficient (R) between Y_i and $\widehat{Y_i}$. Following Tong *et al*., (2020), R was calculated for each fold of each CV. As the weight parameter W, six values (0.16, 0.32, 0.48, 0.64, 0.80, and 0.96) were compared. Differences of accuracy between methods were tested with the

Mann–Whitney-Wilcoxon test using the R function wilcox.test.

5. Simulation analysis

To assess the prediction ability of the proposed model, data were simulated mimicking the real data. First, using the model of Eq. (9), the location (μ_j) and variance components ($\sigma^2_{V_j}$ and $\sigma^2_{U_j}$) were estimated from all genotypes (N = 67). Then $\mathbf{U}_{,j}$ and $\mathbf{V}_{,j}$ were simulated from

$$\mathbf{U}_{,j} \sim N\left(\mathbf{1}\mu_j, \quad \mathbf{G}\sigma^2_{U_j}\right), \text{ and}$$

$$\mathbf{V}_{,j} \sim N\left(\mathbf{U}_{,j}, \quad \mathbf{I}\sigma^2_{V_j}\right),$$

respectively. If the flux of reaction j is restricted to be positive, drawing samples was continued until this restriction was filled. The stoichiometry matrix of the reference genotype (Columbia-0) was used as **M**. However, for each genotype, the last non-zero reaction of each metabolite were modified as follows to fill the constraints of Eq. (1):

$$\left(0 - \sum_{j \in J_{k,-last}} M_{i,j,k} V_{i,j}\right) \Big/ V_{i,J_{k,last}}$$

where $J_{k,-last}$ indicates the set of reactions in **M** that are non-zero for metabolite k except for the last reaction, $J_{k,last}$. For example, when metabolite k has non-zero coefficients in **M** at reactions 4, 18, and 125,

$$M_{i,125,k} = (0 - M_{i,4,k} V_{i,4} - M_{i,18,k} V_{i,18}) / V_{i,125}.$$

The observed phenotypes (**Y**) were then generated by

$$\mathbf{Y} = \mathbf{V}_{,J} + \mathbf{e}.$$

Herein **e** was drawn from the normal distribution. The variance of the normal distribution was set to 25% of that of $\mathbf{V}_{,J}$.

To facilitate model fitting in simulations, a small noise, ψ_{ik}, was added to the left hand of Eq. (6). That is,

$$\sum_{j=1}^{J} M_{i,k,j} \delta_j \alpha_j + \psi_{ik} = \sum_{j=1}^{J} M_{i,k,j} \delta_j V'_{i,j}{}' + \varepsilon_{ik} \text{ for } 1 \le i \le N \text{ and } 1 \le k \le K$$

where $\psi_{ik} \sim N(0, \quad 0.0001)$ of which the variance was determined arbitrary.

With this procedure, 10 simulated data sets were generated. Then 3-fold CV was conducted for each data set. Genotype divisions of CVs followed the first 10 CVs

of Tong *et al*., (2020). As done in the real data analysis, prediction accuracy was evaluated using R which was calculated for each fold of each CV. As the weight parameter *W*, six values (0.16, 0.32, 0.48, 0.64, 0.80, and 0.96) were compared. Differences of accuracy between methods were tested with the Mann–Whitney-Wilcoxon test.

6. Methods compared

6.1. Genomic BLUP (GBLUP)

Genomic BLUP (GBLUP), the most widely used method in genomic prediction, was conducted using the observed phenotypes ($\mathbf{Y}$) as the response variable. The model can be written as

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{e}$$

where $\mathbf{u} \sim N(\mathbf{0},\ \mathbf{G}\sigma_u^2)$ and $\mathbf{e} \sim N(\mathbf{0},\ \mathbf{I}\sigma_e^2)$, respectively. No information about metabolites and metabolic networks were used. The rrBLUP package ver. 4.6.1 was used to execute.

6.2. Quadratic programming (QP)

This method was proposed by Tong *et al*., (2020). First, fluxes of a genotype in the testing data are predicted with GBLUP using the model Eq. (9) (see also Figure 1B). The predicted fluxes, $U_{test,j}$, however, do not fill the constraints of Eq. (1). Thus, the fluxes at the steady state, $V_{test,j}$, are estimated using quadratic programming (QP) implemented in the R package CVXR (ver. 1.0-10) (Figure 1B). The objective function and constraints of this optimization are

$$\min_{V_{test,j}} \sum_{j=1}^{J} \left(\frac{V_{test,j}}{U_{test,j}} - 1 \right)^2$$

$$\text{s.t. } \mathbf{M}_{test,k}\mathbf{V}_{test}^T = 0 \text{ for } 1 \le k \le K,$$

$$V_{test,j} \ge 0 \text{ for } j \in J_r.$$

$V_{test,J}$ at the steady state was then treated as the predicted value for Y. Although the constraints of Eq. (4) and (5) were added in the original method proposed by Tong *et al*., (2020), here QP was done without these additional constraints due to the reasons described in the Results and Discussion.

6.3. MegaLMM

MegaLMM is a multi-trait mixed model that is based on factorization of genetic covariances between traits (Runcie *et al*., 2021). With this method, estimated fluxes ($\mathbf{Vt}_{,j}$) and $\mathbf{Y}$ were regarded as traits, and the following multi-trait mixed model was fitted:

$$[\mathbf{Vt}_{,1} \quad \dots \quad \mathbf{Vt}_{,J} \quad \mathbf{Y}] = \mathbf{XB} + \mathbf{F\Lambda} + \mathbf{U_R} + \mathbf{E_R}$$

where

$$\mathbf{F} = \mathbf{U_F} + \mathbf{E_F},$$

$$\mathbf{U_R} \sim \boldsymbol{N}(\mathbf{0}, \quad \mathbf{G}, \quad \mathbf{\Psi_R}),$$

$$\mathbf{E_R} \sim \boldsymbol{N}(\mathbf{0}, \quad \mathbf{I}_{\boldsymbol{N}}, \quad \mathbf{\Phi_R}),$$

$$\mathbf{U_F} \sim \boldsymbol{N}(\mathbf{0}, \quad \mathbf{G}, \quad \mathbf{\Psi_F}), \text{ and}$$

$$\mathbf{E_F} \sim \boldsymbol{N}(\mathbf{0}, \quad \mathbf{I}_{\boldsymbol{N}} \quad \mathbf{\Phi_F}).$$

Here, the second and third arguments in the normal distribution $\boldsymbol{N}$ denote the covariances between rows and columns, respectively. $\mathbf{X}$ is the design matrix which included only the intercepts for each trait, and $\mathbf{B}$ is the fixed effects. $\mathbf{F}$ is an $N \times P$ latent factor matrix, and $\mathbf{\Lambda}$ is a $P \times (J+1)$ factor loading matrix. $\mathbf{\Psi_R}$ and $\mathbf{\Phi_R}$ are diagonal matrices of dimension $J+1$ and $\mathbf{\Psi_F}$ and $\mathbf{\Phi_F}$ are diagonal matrices of dimension P. Here P is the dimension of latent variables. The prior distributions of parameters followed the vignette of the R package MegaLMM, which is available at (https://github.com/deruncie/MegaLMM/blob/master/vignettes/Running_MegaLMM.Rmd). $\mathbf{Y}$ was scalded before fitting. The total number of iterations was 10,000, thinning was 2, and the last 500 samples were used for posterior inference. P was set to 50, 80, or 160. The R package MegaLMM (ver. 0.1.0) was used to fit the model.

7. Results and Discussion

The results of CVs using simulated data are presented in Fig.2. QP and the proposed models generally showed higher accuracy than GBLUP, suggesting that the information of metabolic networks is advantageous in predicting the target trait phenotype. The accuracy of the proposed model depends on the weight parameter *W*. The accuracy showed a convex curve peaking at *W* = 0.64. The proposed models were able to predict

more accurately than QP at W = 0.48 and 0.64, suggesting that the proposed models potentially have superior prediction ability than QP as intended. MegaLMM was less accurate than GBLUP although MegaLMM also used the multivariate information of fluxes. Perhaps, factorization of genetic covariance has no meaning in these data and may make inference unstable because genetic covariances among fluxes and the target trait were not simulated,.

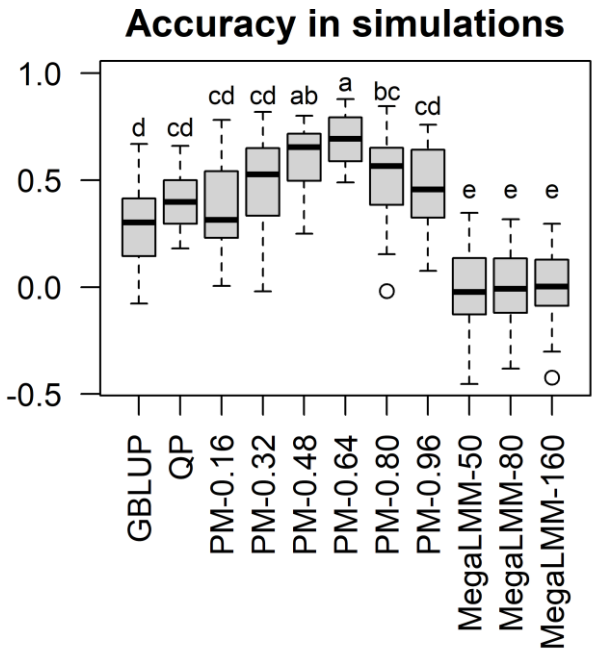

Fig. 2 Accuracy in cross-validations using simulated data. Pearson correlation coefficients between the predicted and observed phenotypes are shown. There are significant differences between different characters (P < 0.05, after Bonferroni correction). GBLUP, genomic BLUP; QP, quadratic programming; PM-*X*, proposed model with the weight parameter $W = X$; MegaLMM-*X*, MegaLMM with the number of latent factors = *X*.

The results of CVs using the real data are presented in Fig. 3. Here, QP was significantly inferior to GBLUP. The medians of accuracies were 0.012 and 0.239, respectively. The accuracy of the proposed model showed a convex curve with W as observed in the simulations, but the peak shifted to W = 0.32. At this W value, the proposed model showed significantly higher accuracy than GBLUP (the median 0.367 vs. 0.239). MegaLMM showed equivalent accuracy to GBLUP. Fig. 4 shows how strictly the constraint of Eq. (6) was filled. The Pearson correlation between

$\sum_{j=1}^{J} M_{i,k,j}\delta_j\alpha_j$ and $\sum_{j=1}^{J} M_{i,k,j}\delta_j V_{i,j}^{''}$ was almost 1.0 at each W value, suggesting the fluxes were at steady-state.

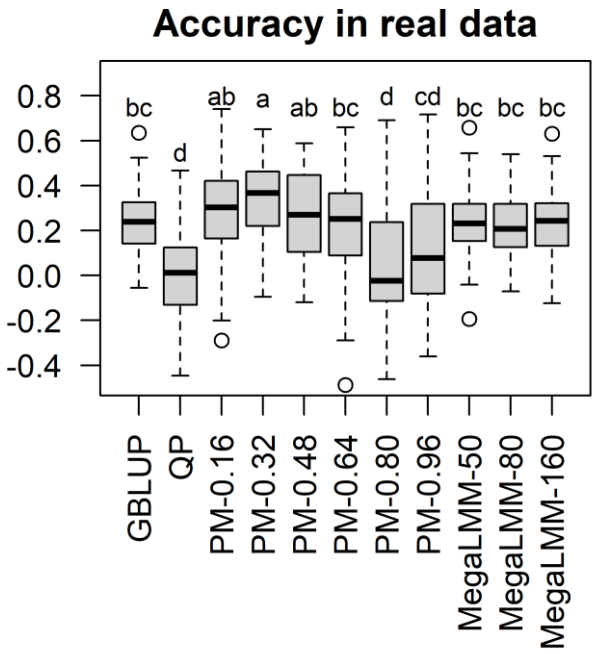

Fig. 3 Accuracy in cross-validations using real data. Pearson correlation coefficients between the predicted and observed phenotypes. There are significant differences between different characters (P < 0.05, after Bonferroni correction). GBLUP, genomic BLUP; QP, quadratic programming; PM-*X*, proposed model with the weight parameter *W* = *X*; MegaLMM-*X*, MegaLMM with the number of latent factors = *X*.

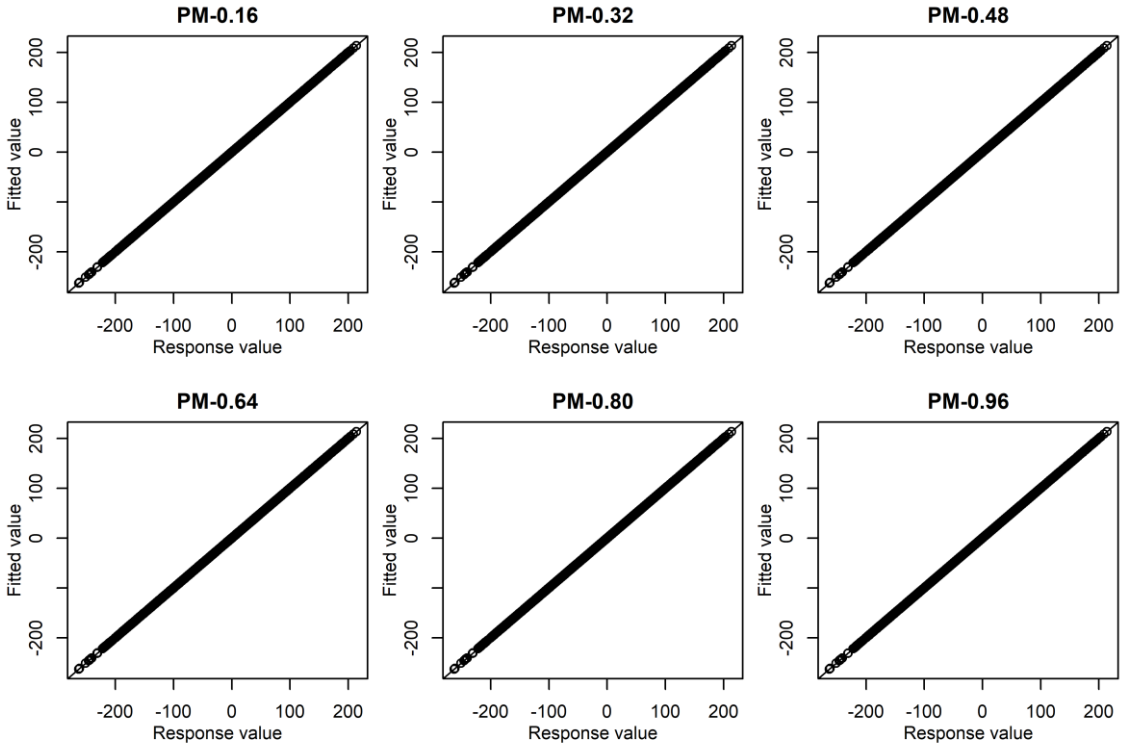

Fig. 4 Fitting status in the constraint term. The x axis is $\sum_{j=1}^{J} M_{i,k,j}\delta_j\alpha_j$ and y axis is $\sum_{j=1}^{J} M_{i,k,j}\delta_j V_{i,j}^{\prime\prime}$ of Eq. (6). PM-*X* denotes the proposed model with the weight parameter *W* = *X.* Results from all CVs are pooled. The diagonal line is the 1:1 line.

Largely, two discrepancies were observed in the real data results compared with the simulation results: (1) Less advantages in using flux balance information and (2) shift of optimal *W* values. The former might arise because of insufficient or inaccurate information on metabolic networks of biomass production of Arabidopsis. That is, all metabolites and reactions may not have been listed and/or some stoichiometry coefficients may be inaccurately determined. It will be a question for future studies what to extent such deficiencies in network information affect prediction accuracy. In terms of the second point, the relative importance between the observation (Y) and constraints (flux balance) may differ between the simulated and real data because the weight parameter *W* controls the relative contributions between these two terms to likelihood as illustrated in Eq. (8). Intuitively, if the observations are more reliable as information source to infer fluxes (**V**) than the constraints, lower *W* values may be preferable and if the observations are less reliable, higher *W* may be preferable. Supposing that the reliability as information source could be quantified with heritability,

the results may suggest that heritabilities of fluxes in real data were lower than those used in the simulations. Although the heritabilities in the simulations followed those estimated from the real data, the estimates must have been overestimated because of small data size, and overestimation for the fluxes would be more influential on inference than that for the observations (Y) because of the difference of dimensions (J vs. 1).

The unexpectedly low accuracy of QP in the real data analyses was drastically improved by introducing additional constraints illustrated in Eq (4) and (5). The median of accuracy was 0.414, which was significantly higher than GBLUP and the proposed models. However, it was found that the predicted values hardly fluctuated among CVs. The mean of the Pearson correlations among the predicted values obtained in different CVs was 0.999, suggesting that the fluctuations in prediction accuracy reported by Tong *et al*., (2020) resulted from the variation of genotype divisions in CVs. Moreover, even when the genomic relationship matrix (**G**) was permutated, prediction accuracy was still high (median 0.424), suggesting that the prediction accuracy was achieved without using genomic information. Further investigations confirmed that the constraint of Eq. (4) is sufficient to cause these phenomena (data not shown). Although it is still unclear why this constraint causes the phenomena, these results suggest that it should be careful to put constraints on fluxes when using flux balance for prediction.

Both the simulated and real data analyses showed that the proposed model is capable to improve prediction accuracy of quantitative traits. However, there are three concerns on the proposed model to be mentioned. The first is the computational cost. The Hamiltonian Monte Carlo simulations typically took 21 hours for model fitting (Windows 11 machine with Core (TM) i7-12700K 3.61 GHz), which is much slower than GBLUP (took < 1 sec.). The high computational cost of the proposed model is not surprising, because the model includes J (336) random effects of which each dimension is N (67) (**U**). The second issue is how to determine W prior to analysis. A possible solution is a grid search using CV being patient with the increased computational cost. Another solution will be to put a prior distribution on W and infer from data. The last issue is the assumption of independency among the fluxes and target trait. Because fluxes would be correlated with each other, ignoring this dependency can make inference unstable. Factorization adopted in MegaLMM will be useful to mitigate this issue. Such improvement of models will be required to make the proposed models more

applicable to various traits and species.

In summary, the proposed model was shown to be capable to improve the prediction accuracy using metabolic network information without using the additional constraints on fluxes. Although there are still issues to be addressed, this study supports the nice idea of Tong *et al*., (2020) that the metabolic network information is useful for phenotype prediction of quantitative traits.

Acknowledgments

The author thanks Ms. Yuuri Onogi for her support in conceiving and conducting this study.

Funding

This work was supported by Ryukoku University.

Data and source code availability

All real data are available at https://github.com/Hao-Tong/netGS. Stan and R scripts required to reproduce this study are available at https://github.com/Onogi/MetabolicModeling.

References

Arnold, A. and Nikoloski, Z. (2014) Bottom-up metabolic reconstruction of Arabidopsis and its application to determining the metabolic costs of enzyme production. *Plant Physiol*., **165**, 1380–1391.

Atanda, S.A. *et al*. (2022) Multi-trait genomic prediction improves selection accuracy for enhancing seed mineral concentrations in pea. *Plant Genome*, **15**, e20260.

Campbell, M.T. *et al*. (2021) Translating insights from the seed metabolome into improved prediction for lipid-composition traits in oat (*Avena sativa L*.). *Genetics*, **217**, iyaa043.

de los Campos, G. *et al*. (2013) Whole-genome regression and prediction methods

applied to plant and animal breeding. *Genetics*, **193**, 327–345.

Dreisigacker, S. *et al*. (2021) Implementation of genomic selection in the CIMMYT global wheat program, findings from the past 10 years. *Crop Breeding, Genetics and Genomics*, **3**, e210005.

Endelman, J.B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, **4**, 250–255.

Garcia-Ruiz, A. *et al*. (2016) Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc. Natl. Acad. Sci. USA*, **113**, E3995–4004.

Heslot, N. *et al*. (2012) Genomic selection in plant breeding: a comparison of models *Crop Sci.* **52**, 146–160.

Horton, MW. *et al*. (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216.

Jarquin, D. *et al*. (2016) Prospects of genomic prediction in the USDA Soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. *G3 (Bethesda)*, **6**, 2329–2341.

Jia, Y. and Jannink, J.L. (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, **192**, 1513–1522.

Lello, L. *et al*. (2020) Sibling validation of polygenic risk scores and complex trait prediction. *Sci. Rep*. **10**, 13190.

Li, Z. *et al*. (2019) Integrating gene expression data into genomic prediction. *Front. Genet.* **10**, 126.

Meuwissen, TH. *et al*. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Onogi, A. *et al*. (2015) Exploring the areas of applicability of whole-genome prediction

methods for Asian rice (*Oryza sativa L.*). *Theor. Appl. Genet.*, **128**, 41–53.

Perez, BC. *et al*. (2022) Adding gene transcripts into genomic prediction improves accuracy and reveals sampling time dependence. *G3 (Bethesda)*, **12**, jkac258.

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Riedelsheimer, C. *et al*. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.*, **44**, 217–220.

Runcie, DE. *et al*. (2021) MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biol.*, **22**, 213.

Schrag, TA. *et al*. (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics*, **208**, 1373–1385.

Scott, BA. *et al*. (2021) How genomic selection has increased rates of genetic gain and inbreeding in the Australian national herd, genomic information nucleus, and bulls. *J. Dairy Sci.*, **104**, 11832–11849.

Sulpice, R. *et al*. (2013) Impact of the carbon and nitrogen supply on relationships and connectivity between metabolism and biomass in a broad panel of Arabidopsis accessions. *Plant Physiol.*, **162**, 347–363.

Tong, H. *et al*. (2020) Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth. *Nat. Commun.*, **11**, 2410.

Wray, NR. *et al*. (2019) Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. *Genetics*, **211**, 1131–1141.

Xu, S. *et al*. (2016) Metabolomic prediction of yield in hybrid rice. *Plant J.*, **88**, 219–227.