Genomic prediction with incomplete omics data

E. Karaman*, V. Milkeviych, Z. Cai, L. Janss, G. Sahana, M.S. Lund

Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University, 8830 Tjele, Denmark; *emre@qgg.au.dk

Abstract

In animal breeding, there has been an increasing interest in investigating the added value of intermediate omics traits such as transcriptomes, metabolites and methylation patterns in genomic predictions. Such data are available only for small number of animals. The "single-step genomic prediction" machinery, which was first proposed to combine pedigree information of a large number of individuals, and genomic information of a fraction of the population, can be useful to handle incomplete omics data. Such an approach, when applied to incomplete omics data scenarios, imply a simple linear relationship from genotypes to different omics traits, which in reality may be very complex. Little is known about the accuracy of genetic evaluations when the omics traits are generated for the whole population. Here, we present two different approaches to handle incomplete omics data, and investigate their impact on genomic predictions, using simulations.

Introduction

Genomic prediction relies on the estimation of the effects for tens of thousands of genetic variants (generally single nucleotide polymorphisms-SNPs) over the whole genome (Meuwissen *et al*., 2001), and has received great attention in animal and plant breeding as well as in human genetics studies. Recently, research interests moved towards exploring the added value of intermediate omics traits, such as transcriptomes, metabolites, methylation patterns etc. in genetic evaluations.

Christensen et al. (2021) provided formulas for breeding values required for genomic evaluation when intermediate traits are included, based on a set of two models: (i) a model relating intermediate omics traits to phenotypes, and (ii) a model relating genotypes and intermediate omics traits. Such an approach has been used in transcriptome-wide association studies (TWAS) to discover gene-trait associations (Wainberg et al., 2019). Analogous to conventional genomic predictions that estimates SNP effects in a reference population, TWAS first uses a reference population of individuals with SNP genotypes and gene expression data. This population is used to estimate the effect sizes of expression quantitative trait loci (eQTL). Second, expression levels are predicted for the individuals in an independent population, for which genotypes and phenotypes are available. Third, statistical associations are tested for their significance between predicted gene expression data and the trait (Wainberg et al., 2019).

A limitation of incorporating intermediate omics traits in genomic predictions is that, they are available for only a fraction of the populations compared to the number of genotyped individuals. When omics data are available for only a subset of the population, a genomic relationship matrix between individuals with and without omics data can be used to predict (impute) the missing omics data (Christensen et al., 2021). This, however, imply a simple linear relationship from genotypes to different omics traits, which in reality may be very complex. We hypothesize here that if the structure of the gene regulatory network (GRN) responsible for the complex trait can be inferred from the available omics data, the unobserved omics traits can be predicted more accurately, and in turn accuracy in genomic evaluations can be improved. In addition, assuming the consistency of underlying true GRN across breeds (Huang et al., 2012),

we further hypothesized that GRN estimated in one breed can be used to predict expression data in other breeds, and in turn for genomic predictions.

Materials & Methods

Models. We used a two-level model as in Christensen et al. (2021) for genomic predictions.

$$\boldsymbol{y} = \mathbf{1}\mu + \mathbf{M}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \quad (1)$$

$$\mathbf{m}_{\mathrm{i}} = \mathbf{1}\mu_{o,i} + \mathbf{g}_{\mathrm{i}} + \boldsymbol{e}_{\mathrm{i}},\ i = 1, \ldots, K \quad (2)$$

where $\mu_{o,i}$ is the mean of expression level i, $\mathbf{m}_{\mathrm{i}}$ $(N \times 1)$ vector of expression levels, $\mathbf{g}_{\mathrm{i}}$ is the vector of genetic effects with $\mathbf{g}_{\mathrm{i}} \sim N\left(0, \mathbf{G}\sigma_{\mathrm{g}}^2\right)$ where $\mathbf{G}$ is a relationship matrix computed based on SNPs, and $\boldsymbol{e}_{\mathrm{i}}$ is the vector of residuals with $e_{\mathrm{i}} \sim N(0, \mathbf{I}\sigma_e^2)$, for gene i. The μ is the phenotypic mean, $\mathbf{M}$ is an $(N \times K)$ matrix of expression levels, $\boldsymbol{\alpha}$ is a $(K \times 1)$ vector of omics' effect sizes, and $\boldsymbol{\epsilon}$ is environmental effect. Equation 2 is a linear mixed model (LMM) with an underlying assumption of a linear path between SNPs to expression levels through the $\mathbf{G}$ matrix, whereas from SNPs to expression levels there can be very complex relationships.

For GRN-based description of the relationship between omics expression levels and genetic effects, we replaced Equation 2 with a structural equation model (Equation 3) to describe the expression levels on K genes for individual j (Zhou and Cai, 2022):

$$\mathbf{m}_{\mathrm{j}} = \mathbf{b} + \boldsymbol{\Lambda}_1\mathbf{m}_{\mathrm{j}} + \boldsymbol{\Lambda}_2\mathbf{t}_{\mathrm{j}}^* + \boldsymbol{\varepsilon_j} \quad (3)$$

where $\mathbf{b}$ is a $(K \times 1)$ vector of general means, $\boldsymbol{\Lambda}_1 (K \times K)$ and $\boldsymbol{\Lambda}_2\left(K \times \left(\sum_{k=1}^{K} n_k\right)\right)$ are matrices of the regulatory effects among genes behind expression levels, and among the candidate *cis*-eQTL, respectively, $\mathbf{t}_j^* = \begin{bmatrix} \mathbf{t}_{j1} \\ \vdots \\ \mathbf{t}_{jK} \end{bmatrix}$ is a $\left(\left(\sum_{k=1}^{K} n_k\right) \times 1\right)$ vector of genotypes of individual j where $\mathbf{t}_{\mathrm{jk}}$ consists of centered genotype content at n_k candidate eQTL for gene k, $\mathbf{b}$ is a $(K \times 1)$ vector of means, and $\boldsymbol{\varepsilon_j}$ is a vector of residuals with $e_{\mathrm{i}} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$. Rearranging Equation 3 we get:

$$\mathbf{m}_{\mathrm{j}} = \mathbf{b}^* + \boldsymbol{\Lambda}\mathbf{t}_{\mathrm{j}}^* + \boldsymbol{\varepsilon}_j^*$$

where $\mathbf{b}^* = (\mathbf{I} - \boldsymbol{\Lambda}_1)^{-1}\mathbf{b}, \boldsymbol{\Lambda} = (\mathbf{I} - \boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\Lambda}_2$ is a $\left(K \times \left(\sum_{k=1}^{K} n_k\right)\right)$, and $\boldsymbol{\varepsilon}^* = (\mathbf{I} - \boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\varepsilon}$.

Data simulation. Genotypic data were simulated using XSim (Cheng et al. 2015), based on an imputed (50k->HD->WGS) whole genome sequenced SNPs (WGS) dataset for 5,783 Holstein (HOL), and 1,305 Jersey (JER) cows. We restricted ourselves to a randomly selected 0.25 Morgan (1Mb = 1Morgan) from chromosome 25, for the proof of concept. In total ten generations were simulated, and population sizes were the same as initial population sizes. The numbers of males were 500 (HOL) and 100 (JER), and sex was assigned at random.

In order to simulate eQTL and gene expression data, we randomly selected 10 of the genes within the selected chromosome region, using the information in the Ensembl database (Howe et al., 2021). Then, for each of the 10 genes, we randomly selected 20 SNPs as a list of candidate eQTL, from 1kb up- and downstream of the gene's TSS site. Two out of the 20 SNPs per gene were selected at random as the cis-eQTL of the gene. Expression data of individual j were simulated with $\mathbf{m}_{\mathrm{j}} = (\mathbf{I} - \boldsymbol{\Lambda}_1)^{-1}\mathbf{b} + \boldsymbol{\Lambda}\mathbf{t}_{\mathrm{j}}^* + (\mathbf{I} - \boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\varepsilon}_j^*$ (Zhou and Cai, 2022), where $\mathbf{g}_{\mathrm{j}} = \boldsymbol{\Lambda}\mathbf{t}_{\mathrm{j}}^*$ includes the breeding values for omics traits, and each were assumed to have the same heritability, $h_m^2 = 0.6$. The variances in $\boldsymbol{\varepsilon}_j^*$ of $(\mathbf{I} - \boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\varepsilon}_j^*$, were set according to this heritability. The weight of each gene's expression data on the final trait phenotype, α_i, was simulated from the unit normal distribution. Genomic breeding value of each individual was

obtained as $\mathbf{u}_j = \mathbf{g}_j' \boldsymbol{\alpha}$, and the trait heritability was set at 0.1 or 0.5. All these procedures described above were replicated 10 times. The simulated network structure (Figure 1) was the same for both breeds, as well as the heritabilities, across all the replicates. The complexity of the simulated network, the number of genes, heritabilities of the gene expressions and the traits were varied, but only the results from the parameters described above were presented here.

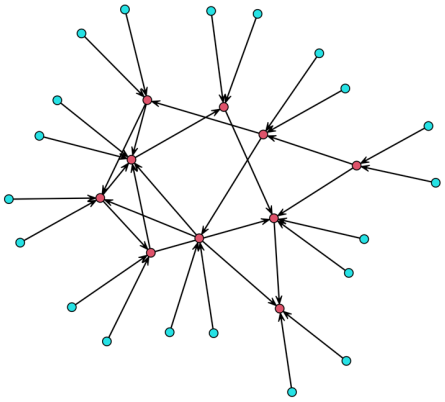

Figure 1. The structure of the simulated GRN. The red and blue circles stand for genes and eQTL.

Analyses. We used generations 9 and 10 as reference and validation populations, respectively. All individuals had genotypes, whereas only 300 and 100 individuals from generation 9 had gene expression data (data were masked), for HOL and JER respectively. Individuals in generation 9 had also phenotypes, and phenotypes of individuals in generation 10 were masked. First, we generated (imputed) the omics data to the whole population by fitting a LMM or GRN model. This was done by using HOL (or JER) individuals' available gene expression and genotype data, together with the genotypes of HOL or JER, for which the omics data will be imputed. Second, breeding values were predicted using those imputed omics data by fitting the model in Equation 1. We used JWAS package in Julia language (Cheng et al., 2018) for the analyses using Equations 1 and 2, and ssemQr package (Zhou and Cai, 2022) in R language for the analyses using Equation 3. It is worth to note that the analyses rely on 20 candidate eQTL SNPs of each gene, and two of those were used to generate the gene expression data, and thereby the phenotypes. Accuracy in genomic predictions were measured as the correlation between the true and predicted genomic breeding values, for individuals in generation 10.

Results

The accuracies from different scenarios are presented in Table 1. When the heritability of the trait was 0.5, accuracies from GRN and LMM were ranged between 0.89-0.99 and 0.82-0.95, respectively. In general, when the imputations were within-breed, genomic predictions were the highest. Using HOL as the reference to impute gene expressions of JER led to accuracies of 0.91 and 0.86, using GRN and LMM, respectively, for heritability of 0.1. Those were 0.97 and 0.93, when the heritability was 0.5.

Discussion

In this work, we have used a hierarchical modeling approach to describe the relationship between SNPs and phenotypes, which was earlier used for TWAS (Wainberg et al., 2019) and recently for integrating intermediate omics traits in genomic prediction (Christensen *et al*.,

2021). So far, the approach has not been explored extensively for genomic prediction, and little is known about the accuracy of breeding value predictions when the omics traits are extended (imputed) to the whole population. We compared two approaches, LMM and GRN, to handle incomplete intermediate omics traits for genomic prediction, and showed that the LMM-based imputation approach was inferior to GRN-based approach, even when the simulated true GRN was not highly complex (Figure 1). Our results are only preliminary and shown here primarily for the purpose of illustrating the effectiveness of accounting for complex traits' genetic architecture, in genomic predictions integrating omics data. It is worth to note that the accuracies (Table 1) are generally much higher (above 0.8) than one might observe in real data. By design, all genes leading to ultimate trait phenotype was known, and the phenotypes were linearly related to steady-state expression levels of those genes. These conditions can hardly be met for real-life complex traits. Analogous to across breed genomic predictions, we used the estimated GRN for one breed to predict gene expression levels in another breed, though GRN is influenced by many factors. When gene expression and other relevant data for, for example genotypes of the two breeds, are available, it is more appropriate to use the estimated GRN in one breed to inform the estimation of GRN in the other, but not to replace it (Zhou and Cai., 2020).

Table 1. Prediction accuracies (standard errors).

Omics data	Prediction	Model for gene expression data			
		GRN		LMM	
		$h^2 = 0.1$	$h^2 = 0.5$	$h^2 = 0.1$	$h^2 = 0.5$
HOL	HOL	0.98(0.002)	0.99(0.000)	0.94(0.003)	0.95(0.003)
JER	JER	0.95(0.002)	0.98(0.002)	0.88(0.004)	0.94(0.001)
HOL	JER	0.91(0.007)	0.97(0.002)	0.86(0.008)	0.93(0.004)
JER	HOL	0.78(0.013)	0.89(0.009)	0.74(0.010)	0.82(0.011)

References

Cheng, H., Garrick, D., Fernando, R. (2015) G3 (Bethesda, Md.) 5(7):1415-1417. https://doi.org/10.1534/g3.115.016683

Cheng H., Fernando R., Garrick D. (2018a) JWAS: Julia implementation of whole-genome analysis software. World Congr Genet Appl Livest Prod, Auckland, NZ 11:859

Christensen, O.F., Börner, V., Varona, L., Legarra, A. (2021) 219(2):iyab130, https://doi.org/10.1093/genetics/iyab130

Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J *et al*. (2021) Nucleic Acids Res. 49(1):884-891 https://doi.org/10.1093/nar/gkaa942

Huang, W., Richards, S., Carbone, M. A., Zhu, D., Anholt, R. R., *et al*. (2012) Proc. of the National Academy of Sciences of the United States of America, 109(39):15553-15559. https://doi.org/10.1073/pnas.1213423109

Meuwissen T.H.E., Hayes B.J., and Goddard M.E. (2001) Genetics 157(4):1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., *et al*. (2019) Nat Genet 51:592-599. https://doi.org/10.1038/s41588-019-0385-z

Zhou, X., and Cai, X. (2020) Bioinformatics 36(1):197–204. https://doi.org/10.1093/bioinformatics/btz529

Zhou, X., and Cai, X. (2022) Bioinformatics 38(1):149-156. https://doi.org/10.1093/bioinformatics/btab609