Capitalizing on complex annotations in Bayesian genomic prediction for a backcross population of growing pigs

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, Andrea Rau

Capitalizing on complex annotations in Bayesian genomic prediction for a backcross population of growing pigs

F. Mollandin[1*], H. Gilbert[2], P. Croiseau[1] and A. Rau[1,3]

[1] INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas 78350, France;
[2] GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan 31320, France;
[3] BioEcoAgro Joint Research Unit, INRAE, Université de Liège, Université de Lille, Université de Picardie Jules Verne, Estrées-Mons 80203, France ; [*]fanny.mollandin@inrae.fr

Abstract

Prior biological information has the potential to guide and inform genomic prediction models, but the BayesRC approach is currently limited to the use of disjoint categorizations of genetic markers. We propose two novel Bayesian approaches to model cumulative (BayesRC+) or preferential (BayesRCπ) contributions of multiple biological categories for multi-annotated SNPs. We illustrate the performance of these approaches on data from a backcross population of growing pigs in conjunction with several different sets of annotations related to multiple production traits constructed using the PigQTLdb. On the two traits predicted, ADG and BFT, we observed improved prediction quality on ADG (up to 1.7-gain point) with both BayesRCpi and BayesRC+, and suitable annotation set.

Introduction

In plant and animal breeding, genomic prediction models have been widely developed and deployed in recent years to predict polygenic traits using genetic variants, typically single nucleotide polymorphisms (SNP). An interesting and potentially useful approach to improve upon existing genomic prediction models is to combine the use of genotype and phenotype data with prior biological information to better guide models. Most routinely used genomic prediction models are based on linear models, including notably genomic best linear unbiased prediction. Another family of models, known as the Bayesian alphabet (Habier *et al.*, 2011), uses a flexible set of assumptions about how individual SNPs contribute to the overall genomic variance. Among these, BayesR (Erbe *et al.*, 2012) assumes SNP effects arise from one of four groups (null, small, medium, or large variance) and has been shown to perform well for both prediction and quantitative trait loci (QTL) mapping (Moser *et al.*, 2015; Mollandin *et al.*, 2021). BayesRC extends BayesR to further incorporate prior biological information in the form of disjoint annotation categories (MacLeod *et al.*, 2016), but SNPs can only be assigned to a single annotation category. There thus remains a need for genomic prediction models able to capitalize on annotations of greater complexity, in particular those for which SNPs may potentially be assigned to multiple categories.

In this work, we present two novel extensions to BayesRC to deal with such complex, overlapping annotations, and we illustrate their utility on data from an experimental backcross population in growing pigs. This project is part of EuroFAANG (https://eurofaang.eu), a synergy of five Horizon 2020 projects that share the common goal to discover links between genotype to phenotype in farmed animals and meet global Functional Annotation of ANimal Genomes (FAANG) objectives.

Materials & Methods

Bayesian genomic prediction with complex, overlapping annotations. The general statistical model for genomic prediction can be defined as

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{X\beta} + \boldsymbol{e}, \quad e_i \sim N(0, \sigma_e^2) \tag{1}$$

where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{\mu}$ an intercept, $\boldsymbol{\beta}$ the vector of SNP effects, $\boldsymbol{X}$ the centered and scaled marker matrix, and σ_e^2 the variance of the residuals $\boldsymbol{e}$. We further assume that $\boldsymbol{C} = (C_{i,j})$ denotes annotation categories, such that $C_{i,j} = 1$ if SNP i is included in category j and 0 otherwise.

Using the four-component mixture of BayesR as a base, we propose two alternative models to account for overlapping annotations (where $\sum_j C_{i,j} > 1$ for some i). The first, BayesRCπ, defines a mixture-of-mixtures prior for SNP effects to assign multi-annotated markers to the single annotation category that maximizes its conditional likelihood:

$$\beta_i \sim \sum_{j \in C_{i,j}=1} p_{i,j}\left(\pi_{1,j}\delta(0) + \pi_{2,j}N(0, 10^{-4}\sigma_g^2) + \pi_{3,j}N(0, 10^{-3}\sigma_g^2) + \pi_{4,j}N(0, 10^{-2}\sigma_g^2)\right), \tag{2}$$

such that $\delta(0)$ represents the dirac function at 0, $\sum_k \pi_{k,j} = 1$ for all annotations j, σ_g^2 the total additive genetic variance, and $p_{i,j}$ the annotation mixing parameter with $\sum_j p_{i,j} = 1$ for all i. The second, BayesRC+, instead defines a cumulative mixture prior across categories for SNP effects:

$$\beta_i \sim \sum_{j \in C_{i,j}=1}\left(\pi_{1,j}\delta(0) + \pi_{2,j}N(0, 10^{-4}\sigma_g^2) + \pi_{3,j}N(0, 10^{-3}\sigma_g^2) + \pi_{4,j}N(0, 10^{-2}\sigma_g^2)\right). \tag{3}$$

In both models, all mixing proportions are assumed to follow flat Dirichlet priors and σ_g^2 an inverse χ^2 prior. A Gibbs sampler is used for inference as posterior distributions are not tractable. Both BayesRCπ and BayesRC+ have been implemented in the BayesRCO software in Fortran; additional details can be found in the User's Guide (https://github.com/fmollandin/BayesRCO).

Genotype and phenotype data from a backcross pig population. A backcross (BC) population between Large White (LW; 3/4) and Creole (CR;1/4) pigs was established as previously described (Gourdine *et al.*, 2019). BC (n = 1,297 from 130 LW sows) growing pigs raised in two environments (tropical and temperate) were related via genetically related sows sired with the same 10 F1 × CR LW boars. A common trait recording protocol was used in the two environments for phenotypic data. Phenotypes were pre-corrected for age, sex, and farm; we focus here on measures at 23 weeks for backfat thickness (BFT) and average daily weight gain (ADG). Animals were genotyped using the Illumina Porcine 60k BeadChip array; markers with minor allele frequencies greater than 0.01 were retained for the analysis (corresponding to 46,908 and 46,881 markers for ADG and BFT, respectively). To establish the potential impact of our models on prediction accuracy, we used a sibling-structured 10-fold cross validation procedure. For the descendants

from each sire in turn, we calculated the correlation between their observed corrected phenotypes and those predicted from models constructed on the descendants of the remaining 9 sires; validation correlations were averaged across the ten folds.

PigQTLdb annotations. Animal QTLdb (https://www.animalgenome.org/QTLdb) groups together curated results from genotype-phenotype association studies in several livestock species (Hu *et al.*, 2021). Cross-experiment QTL data from PigQTLdb (Release 45; SS11.1) for traits relevant to pig production were downloaded for eleven trait sub-hierarchy categories (anatomy, behavioral, blood parameters, conformation, fatness, fatty acid content, feed conversion, fowth, immune capacity, litter traits, reproductive organs). An additional "other" category was created for markers not included in PigQTLdb. Genotyped markers in our data were subsequently assigned to one or more annotation categories using three different strategies: (1) using the position of known PigQTLdb markers; (2) using the extended position of known PigQTLdb markers, including the nearest up- and downstream neighbors ("extended PigQTLdb"); and (3) using the extended position of known PigQTLdb markers as before, where neighboring markers were allowed ambiguous assignment to both trait-specific and "other" categories ("fuzzy extended PigQTLdb"). In the three annotation construction strategies, 1.3%, 4.9% and 17.7% of markers were respectively assigned to two or more categories.

Results

We compared the prediction accuracy of BayesRCπ and BayesRC+ to that of BayesR (ignoring annotation categories) and BayesRC (where a single category is allowed per marker). For the latter, multi-annotated SNPs were randomly assigned to a single category. We notably observed different trends for the two traits (Table 1). For ADG, we remark a loss in prediction accuracy compared to BayesR for all annotation-based models with straightforward pigQTLdb annotations; however, extending these annotations to include neighboring markers (extended and fuzzy extended pigQTLdb) led to improvements in prediction quality, with a 1.7-point gain in correlation for BayesRCπ with extended annotations. On the other hand, for BFT the use of annotations, regardless of how they are constructed, did not appear to lead to a marked improvement in prediction. This suggests that categorizations constructed from PigQTLdb contribute little pertinent information for the genomic prediction of BFT in our data, or that environment-dependent categories should be added to the model to account for the significant GxE affecting this trait (Gourdine *et al.*, 2019).

Discussion

In this work we have proposed two new approaches, BayesRCπ and BayesRC+, to fully capitalize on complex, overlapping annotations in genomic prediction. Both methods showed promise for incorporating partially overlapping categories from pigQTLdb in genomic prediction for a growing pig population, although a gain in predictive accuracy was observed for only one (ADG) of the two traits considered here. We also compared three strategies for constructing prior biological

categories by extending pigQTLdb annotations in various ways to include neighboring markers, which has the potential to better exploit linkage disequilibrium around relevant markers. Taken together, these results suggest that the incorporation of complex annotations can lead to modest gains in prediction performance in some cases, even for moderate marker density SNP chips, but such gains depend strongly on the choice and construction of annotations and are unlikely to be universal across traits.

Table 1. Validation correlation for two traits in pig data for BayesRCπ and BayesRC+ with different annotation strategies, as compared to BayesR and BayesRC.

Method	Annotations	ADG Mean (SD)	BFT Mean (SD)
BayesR	—	0.213 (± 0.081)	0.265 (± 0.161)
BayesRC	PigQTLdb (random)	0.200 (± 0.105)	0.265 (± 0.159)
	Extended pigQTLdb (random)	0.225 (± 0.098)	0.258 (± 0.157)
BayesRCπ	PigQTLdb	0.200 (± 0.100)	0.266 (± 0.157)
	Extended pigQTLdb	0.229 (± 0.095)	0.254 (± 0.162)
	Fuzzy extended pigQTLdb	0.226 (± 0.096)	0.262 (± 0.159)
BayesRC+	PigQTLdb	0.207 (± 0.097)	0.273 (± 0.163)
	Extended pigQTLdb	0.227 (± 0.095)	0.271 (± 0.158)

This work is part of the GENE-SWitCH project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement n° 817998. The financial support of the French National Agency of Research (ANR PigHeaT, ANR-12-ADAP-0015) is also gratefully acknowledged.

References

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., *et al*. (2012) Journal of Dairy Science 95(7): 4114–29. https://doi.org/10.3168/jds.2011-5019

Gourdine, J.-L., Riquet, J., Rosé, R. Poullet, N., Giorgi, M. *et al.* (2019) Journal of Animal Science 97(9): 3699–3713. https://doi.org/10.1093/jas/skz245

Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. (2011). BMC Bioinformatics 12(1): 186. https://doi.org/10.1186/1471-2105-12-186

Hu, Z.-L., Park, C.A., and Reecy, J.M. (2021) Nucleic Acids Research, gkab1116. https://doi.org/10.1093/nar/gkab1116

MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., *et al.* (2016) BMC Genomics. 17(1): 144. https://doi.org/10.1186/s12864-016-2443-6

Mollandin, F, Rau, A., and Croiseau, P. (2021) G3 Genes|Genomes|Genetics 11(11): jkab225. https://doi.org/10.1093/g3journal/jkab225

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R. *et al*. (2015) PLOS Genetics 11(4): e1004969. https://doi.org/10.1371/journal.pgen.1004969