Causal Meaning of Genomic Predictors: Implication on Genome-Enabled Selection Modeling

B.D. Valente [*,†], G. Morota[†], G.J.M. Rosa[†,‡], D. Gianola[*,†,‡], K. Weigel[*].
[*]Dep. of Dairy Science,[†]Animal Sciences,[‡]Biostatistics and Medical Informatics,University of Wisconsin-Madison, WI, USA

ABSTRACT: The term "effect" in additive genetic effect suggests a causal meaning. However, inferences on such quantities for selection purposes are normally conducted as a prediction task. Predictive ability is currently the most used criterion for comparing models and evaluating new methodologies, but it is insufficient to evaluate if predictors identify causal effects. Therefore, the usual approach to infer genetic effects seems to contradict the label of the quantity inferred. Here we investigate if genomic predictors for selection should be treated as standard predictors from regression models, or if they must reflect a causal effect, asking for causal inference approaches. We demonstrate that selection requires learning causal genetic effects. However, genomic predictors may reflect non-causal signal, providing good predictions but poorly representing true genetic effects. Genomic selection models should be constructed aiming primarily for identifiability of causal genetic effects, not for predictive ability.
Keywords: causal inference; genomic selection; model comparison; prediction

Introduction

Inference of additive genetic effects by fitting predictors based on genomic (or pedigree) information is pivotal for selection and animal breeding. The meaning of these additive effects is expressed in quantitative genetics mostly using causal terms, e.g., "influence", "causes of variability" and so forth, as in Falconer (1989) or Lynch and Walsh (1998). The term "effect" by itself suggests a causal connotation. This indicates that inferring additive genetic effects from field data pertains to the realm of causal inference from observational (i.e., non-experimental) data. In animal breeding, however, this inference is commonly conducted as part of a prediction problem. Hence, predictive ability is treated as the main criterion to compare models and to evaluate novel methods and technologies. This learning approach is insufficient for drawing causal conclusions, which seems to contradict the original meaning of the quantity inferred. Furthermore, discussion on challenges involved in identifying causal effects and the concepts necessary for this discussion (Pearl (2000)) are virtually absent in genomic selection literature.

This inconsistency casts doubt whether inferring additive genetic effects should be conducted customarily as a prediction task or as inference of causal information like suggested by the label. Solving this issue is important because the approach needed to tackle one type of problem is not suitable to tackle the other type. They involve different concepts, learning methods, and assumptions. Ignoring this issue may lead to poor modeling choices and wrong conclusions.

Here we investigate if genomic predictors should be treated as standard predictors in some regression analysis or if they should reflect a causal effect to be useful for selection. Solving this issue is necessary to decide if genomic selection models should be constructed aiming primarily for maximum predictive ability or for identifiability of causal genetic effects.

Materials and Methods

Graph-theoretic terminology. Directed Acyclic Graphs (DAGs) are commonly used to qualitatively express causal relationships and assumptions among a set of variables. In such graphs (e.g., Figure 1a), nodes represent variables and arrows represent direct causal connections. One crucial aspect is how to translate the causal information in DAGs into conditional dependencies and independencies (i.e., into statistical information) which could be exploited for prediction. Paths in the graph may allow transmission of dependence between its extremities (e.g. $d \leftarrow a \rightarrow c$, $a \rightarrow d \rightarrow e$), unless they include one or more colliders (nodes with both arrows pointing towards them, such as c in $d \leftarrow a \rightarrow c \leftarrow b$). While colliders block the flow of dependence and non-colliders allow it, conditioning reverses this feature: conditioning on non-colliders blocks the flow of dependence and conditioning on colliders allows it.

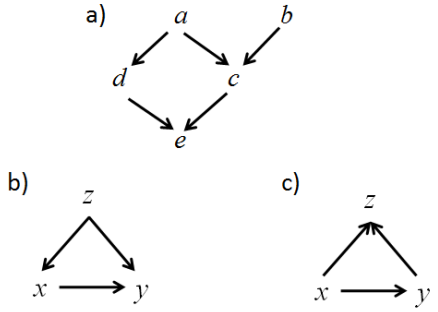

Figure 1 – Directed acyclic graphs.

Prediction and causal inference. The tasks of predicting a variable y from another variable x and inferring the effect of x on y are different. The former involves a function that allows predicting a value for y if a specific value for x was OBSERVED. A joint distribution involving both variables is sufficient to derive such function (e.g. $E(y \mid x)$).

Alternatively, inferring a causal effect involves learning about how y is expected to change if x is SET to some value through external interventions (Pearl (2000)). Unlike for prediction, joint distributions are not sufficient to derive causal information. For example, any joint distribution where x and y present some association are compatible with the following hypotheses: a) x affects y

($x \rightarrow y$), b) x and y are affected by a third (possibly unmeasured) variable ($x \leftarrow z \rightarrow y$), and c) a combination of a) and b).

As joint distributions are sufficient to obtain predictors but not causal effects, identifying the latter from evidence requires extra (causal) assumptions that cannot be deduced from this distribution. These assumptions can be expressed by a causal DAG involving x and y. The DAG is then used to deduce which identifiable function of data (or of the joint distribution) represents the target effect. For example, omitting from the DAG disturbance terms that independently affect each variable, the effect of x on y could be obtained from fitting $y_i = \mu + x_i\beta + e_i$ simply by assuming $x \rightarrow y$. According to the assumed DAG, the target effect is the only source of association between the pair of variables. On the other hand, if the causal relationship assumed involved a common influence from a third variable z (Figure 1b), then a second source contributes to the marginal association. As this is the association explored by the model above, the target effect cannot be identified from it. However, the effect could be identified from $\hat{\beta}$ after fitting $y_i = \mu + x_i\beta + z_i\alpha + e_i$ as it captures the association between x and y conditionally on z, which blocks the confounding due to $x \leftarrow z \rightarrow y$. Alternatively, if z is assumed to be affected by both x and y (Figure 1c), then fitting $y_i = \mu + x_i\beta + z_i\alpha + e_i$ would not identify the effect because conditioning on z "activates" $x \rightarrow z \leftarrow y$, which would also contribute to $\hat{\beta}$, confounding it. However, the target effect could be identified from $y_i = \mu + x_i\beta + e_i$, as $x \rightarrow z \leftarrow y$ is marginally blocked.

Notice that inferring effects is different from obtaining predictions. The criterion used to choose models for causal inference disregards predictive ability. Actually, models deemed as unsuitable for causal learning in the examples above could provide the best predictions. Clear examples are scenarios where associations due to confounding paths are more significant than those from the causal paths. Notice also that inferences can only be called causal effects if coupled with causal assumptions that support their identifiability from data or joint distribution.

Simulation. Results of this study were illustrated with simulations. Whole-genome genotypes were simulated for 2000 individuals, consisting of 4 chromosomes with 1 Morgan each, 15 QTL per chromosome and 5 markers between consecutive pairs of QTL. Simulation scenarios involved different effects from genotypes on one or two phenotypic traits, and also different causal relationships between them. For each scenario, two alternative GBLUP models were fitted using BLR (de los Campos (2013)) for one of the traits. One of them (model C) included as a covariate either a second simulated phenotypic trait or an environmental variable, while the other (model IC) ignored such covariates. Model C emulated genomic selection models that "correct for" traits (e.g., an analysis for somatic cell count that corrects for milk production, or one of age at first calving that corrects for body weight), or systematic environmental effects (e.g. farm, season). 10-fold cross validation tests were applied to compare models' predictive ability as well as the ability to identify the true genetic effect.

Results and Discussion

Selection. The basic structure involved in selection as generally described in quantitative genetics can be depicted as in Figure 2, where G is a genotype associated with a phenotype y. Also, consider G' as an individual genotype of the next generation, which is causally affected by the genotype of the parent G. G' and the phenotype y' are related as G and y, but the nature of such relationship is left unspecified at this point (Figure 2a).

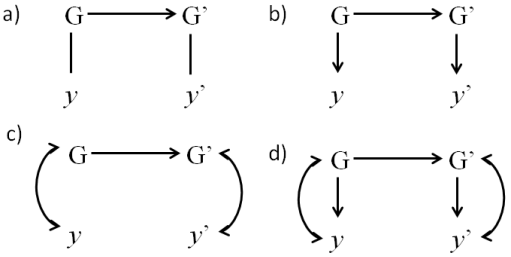

Figure 2 – Causal structures representing the selection context. The nodes G and G' represent genotypes, and the latter is assigned to a descendant of the former; y and y' are phenotypes assigned to each individual; arrows represent causal effects, bidirected arrows represent confounding paths and undirected edges represent unresolved causal relationships.

Selection involves actions to modify G' under the expectation that y' can improve from that. This implies that selection relies on a causal effect from G' to y' (e.g. Figure 2b). Even if the relationship between G (G') and y (y') was more complex, with extra sources of associations (e.g., a confounding source of covariance between genotype and environmental causal factors, as in Figure 2c), the response to selection depends only on the causal effect from G' to y'. On the other hand, any signals between G and y could be explored for genomic prediction, even if devoid of genetic effects (e.g. Figure 2d), or negatively associated with the causal signal. Since response depends on the effect of G' on y', and G' receives alleles from G, the pivotal task for selection is not to predict y from G or to identify individual G's associated with the best y's, but to study the effect of G on y and identify individuals whose genotypes produce the best effects on y.

Simulation. Simulations confirmed that good genomic predictions can exploit non-causal signals. The best predictors may be poor indicators of the true genetic effects, while genomic predictors with less predictive ability can represent the true effect more accurately. Considering that the relevant information for selection is the causal genetic effect, results illustrate that the task required for selection is not essentially a prediction problem. Therefore, cross validation and similar criteria are not sufficient to evaluate and compare models for selection. Maybe the clearest simulation scenario that demonstrates this distinction involves a non-heritable trait y_2 that affects a heritable trait y_1 (i.e. $y_2 \rightarrow y_1 \leftarrow G$). Assuming that the goal is analyzing y_2, the model that includes y_1 as a

covariate (model C) provided reasonable genomic predictive ability and, therefore, suggested variability of "genetic effects", although y_2 is not even heritable. On the other hand, model IC provided extremely poor predictions. However, model IC is the one that provides the relevant information for selection: genetics does not affect y_2, which would not respond to selection. On the other hand, results from model C suggest it would respond to selection. The reason for this "artifact" is that including y_1 as a covariate makes the genomic predictors capture the association between *G* and y_2 conditionally on y_1, which is actually affected by both *G* and y_2 (i.e., y_1 is a collider). This creates a non-causal association that could be explored for prediction, but that is not relevant for selection. We also explored scenarios where both traits are affected by the *G*, where y_1 is the target trait instead of y_2, and where the phenotypic trait is affected by an environmental effect which is associated with *G*. For all of these, predictive ability was not a good criterion to choose models that better identified true genetic effects. More details on this study are given by Valente et al. (2013).

These simulations were used as *exempla contraria*, so it is not implied that cross-validation tests always points to the wrong model. It shows, however, that ability to predict is not sufficient to claim that a model or predictor is good for selection purposes.

Discussion. As selection requires learning causal effects, inferring breeding values should be done in the framework of causal inference. As for any task of this kind, one should make causal assumptions regarding the involved variables, and verify if the relevant effects are identifiable from fitting a candidate model, according to these assumptions. This extra requirement may seem a disadvantage if compared with predictive methodologies. But if causal information is required, such assumptions are necessary anyway. Ignoring this and choosing models based on predictive ability or goodness-of-fit is theoretically inadequate and may lead to poor modelling choices.

Additionally, causal assumptions for model constructions are not necessarily difficult to accept. In the example described, as the goal is to infer the total effect of *G* on y_2, then simply assuming y_1 as heritable is sufficient for choosing to remove it from the model. Under this condition, including y_1 could either create a non-causal signal between *G* and y_2 or block part of the target effect depending on how the variables are related. A formal criterion to choose covariates for causal inference is the back-door criterion (Pearl (2000)). Given reasonable assumptions, this criterion would indicate which models identify better the true genetic effects in the simulations (Valente et al. (2013)).

Notice that here the main goal is not to present an alternative way of constructing and comparing models, or a new interpretation of genomic predictors. Results came from using causal DAGs to investigate classical quantitative genetics concepts involved in selection. They suggest that the relevance of genomic predictors for selection depends primarily on how well they represent the causal effect of *G* on *y*, and not on its predictive ability. First and foremost, a signal between genotype and phenotype should be declared as causal, which requires causal assumptions. Only then, alternatives for modeling this signal can be compared via predictive ability, provided that none of them contradict the causal assumptions.

Causal assumptions are necessary even to support standard interpretations of model parameters. That is the case, for example, if one wants to preserve the standard interpretation of heritability estimates, or to use genetic covariances to assess indirect responses to selection. Detached from the causal assumptions, what we call heritability is nothing more than a regularization parameter.

Conclusion

Fitting predictors from mixed models and genomic selection models for selection purposes is a causal inference task from non-experimental data. Such use of the predictors, as well as standard interpretations of the parameters of a model (genetic variance, heritability, genetic covariance and so forth) depend on causal interpretation of the results, as well as on modeling practices from causal inference, with explicit causal assumptions and model constructed aiming for causal identifiability of genetic effects.

Literature Cited

de los Campos, G., Perez, P., Vasquez, A. I. et al (2013). Methods. Mol. Biol. 1019: 299-320

Falconer, D. S. (1989) Introduction to quantitative genetics. Longman, New York.

Lynch, M., Walsh, B. (1998) Genetics and analysis of quantitative traits. Sinauer, Sunderland, Mass.

Pearl, J. (2000) Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge, UK.

Valente, B. D., Morota, G., Rosa, G. J. M. et al. (2013), arXiv:1401.1165 [q-bio.QM], available in http://arxiv.org/ftp/arxiv/papers/1401/1401.1165.pdf .