On the advantage of identifying causal genetic variants for genomic prediction

T.H.E. Meuwissen[1*] and M.E. Goddard[2,3]

[1]Norwegian University of Life Sciences, Box 5003, 1432 Ås, Norway; [2]Agriculture Victoria, Bundoora, Australia; [3]Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Australia; theo.meuwissen@nmbu.no

Abstract

Genome Wide Association Studies have been successful in identifying Quantitative Trait Locus regions. However, the identification of causal variants (CVs) has proven difficult in many cases. One may hypothesize that knowing the CV would not improve the (genomic) prediction of phenotypes much since otherwise phenotypes would clearly point to the CV. The effect of (not) knowing the CV on local estimated breeding values (LEBV) in 200kb regions surrounding the CV was estimated for 3 well-known CVs (in ABCG2, DGAT and GHR) in a large whole genome sequence dataset. Knowing the CV improved the accuracy of LEBV for all three CVs: for ABCG2 it was instrumental in pointing to the correct CV; the DGAT-CV was poorly imputed, and knowing the CV would result in improved genotyping; and GHR was poorly predicted by other SNPs in the region. In addition, knowing CVs is important for understanding trait-biology and for targeted gene-editing.

Introduction

Quantitative traits are generally complex in the sense that there are many genes, i.e. causal genetic variants, plus environmental effects that determine the trait phenotype. As an example there are 9,863 independent variants with significant ($p<5*10^{-8}$) effects on human height (Visscher et al., 2020). If many causal variants affect a trait then most of their effects must be relatively small, but some causal variants may have large effects. The distribution of effects sizes of causal variants can be modelled by a thick-tailed distribution such as a Gamma distribution with shape parameter <1 which includes more large effects than a Normal distribution. (Hayes and Goddard, 2001).

The identification of these causal variants with large effects is the aim of Genome Wide Association Studies (GWAS). GWAS have been successful in identifying many Quantitative Trait Loci (QTL), i.e. small chromosomal regions that explain genetic variance for a quantitative trait. However, identifying the causal variant (CV) within a QTL region is often difficult for a number of reasons:

1. The CV may not be included in the genotype data, even if Whole Genome Sequence (WGS) genotype data are used. For instance, the CV may not be included in the reference set of sequenced animals because it is too rare or hard to call as structural variants often are.
2. GWAS is based on Linkage Disequilibrium (LD), but many variants may be in high LD with the CV making it difficult to distinguish between these variants. If the CV and another non-causal variant (NCV) are in close LD, the NCV may show the higher GWAS signal due to an unfortunate sample of animals entering the GWAS. This problem may be reduced by increasing the sample size, but if the LD between the CV and NCV is 100%, a GWAS study cannot distinguish these two variants (since their genotypes are identical). If the LD is say 90%, a huge data set may be needed to ensure that the CV yields the highest GWAS signal.

3. Large WGS data sets are often obtained by genotype imputation, and genotype imputation errors of the CV genotypes may reduce the association between the CV genotypes and the trait.

Given that the identification of causal variants is problematic, the question arises whether the identification of the CV is important? From the perspective of understanding the trait biology, a sharp GWAS peak, without identification of the CV, may identify the gene that is affecting the trait, but it will not be able to answer the question what genetic change in (or around) the gene is causing the (positive or negative) trait effect. The latter also prevents studies investigating the mechanism by which the CV affects the trait. In addition, the CVs provide clear targets for gene editing (Jenko et al., 2015).

However, here, we will take the perspective of genomic prediction, i.e. we are only interested in the accuracy of genomic prediction, and whether knowing the CV increases the accuracy of genomic prediction. Instead of performing computer simulations (where the CV is always known), we will address this question using real data and consider 3 known and well-studied CVs affecting milk traits in dairy cattle: the Y581S mutation in ABCG2 on BTA 6 (Cohen-Zinder et al., 2005); the K323A mutation in DGAT1 on BTA 14 (Grisart et al., 2004); and the F2Y mutation in GHR on BTA 20 (Blott et al., 2003).

Materials & Methods

Data. The dataset is described in detail in Meuwissen et al. (2021). Briefly, it consisted of WGS genotypes and daughter yield deviations (DYD; in case of bulls) or yield deviations (YD; in case of cows) for milk yield of 35,549 Holsteins, Jerseys, and Australian Red bulls and cows. Animals were either directly genotyped with the Illumina 800K BovineHD bead chip (HD), or first genotyped with the Illumina BovineSNP50K chip or a lower density SNP chip, and subsequently imputed to HD. All individuals were imputed to WGS using a reference population of Holstein, Jersey and Australian Red bulls and cows from Run 5 of the 1000 bulls genome project. After filtering out variants with a minor allele frequency (MAF) lower than 0.002 and LD pruning ($r^2 > 0.9$), 4,809,520 variants were retained for the analysis. All three mutations (in ABCG2, DGAT1 and GHR) were included in these data.

Comparison of alternatives. All analyses were performed by the BayesGC model (see Meuwissen et al., 2021, for details), which is a Bayesian variable selection model that fits a genomic relationship matrix (**GRM**) as well as the variable selection term, which selects individual variants from the set of 4,809,520 variants using a Monte Carlo Markov Chain (MCMC) method. The GRM was based on HD SNP genotypes, which did not include the ABCG2, DGAT1, and GHR CVs. The following alternative analyses were compared:

- **AV**: all 4,809,520 variants were included in the analysis, including the 3 CV (treated the same as all other SNPs);
- **AV-CV**: the same as AV but excluding the CV (assessing the value of having the CV in the data or not);
- **AV+CV**: the same as AV except that the 3 CV were assumed known and fitted with a prior probability of 1, i.e. they were always included in the model;
- **CV**: fits only the three CVs with prior probability 1;
- **TOP10**: fits 10 SNPs with highest posterior probability in each of the three QTL regions, i.e. a total of 30 SNPs were fitted with prior probability 1.

The **TOP10** analysis considers the situation where the QTL region is detected but it is uncertain which of the SNPs is the causal variant. In order to deal with this uncertainty, the model fits 10 top SNPs (the 10 SNPs with the highest posterior probability in a 200kb region surrounding the CV). Since there are 3 QTL regions, this results in a total of 30 SNPs being fitted. It may be noted that the top-10 of the variates in the DGAT1 region did not contain the DGAT1-CV probably due to the poor imputation accuracy of this SNP (which was variable possibly because it is a 2-SNP mutation which may confuse some software), and possibly also due to other CV in the region affecting the association. In case of ABCG2, the ABCG2-CV ranked number 8 in its 200kb region, i.e. it was not the most important SNP in the region. The GHR-CV ranked number 1 in its 200 kb region, and was clearly the most important SNP.

The bayesGC model requires a variance assumed for the fitted variants, which was a fraction of 0.0005 of the total genetic variance (Meuwissen et al., 2021) for the **AV**, **AV-CV** and **AV+CV** analyses. For the **CV** analysis this factor of the genetic variance was assumed 0.01, which assumes that the CV explain a factor of 0.01 of the genetic variance each (and if the actual explained variance is bigger, this would hardly affect the accuracy of prediction). The **TOP10** analysis assumes that each SNP explains 1/10 of the variance of the locus, and in order to agree with the **CV** analysis each SNP was assumed to explain a factor of 0.001 of the genetic variance.

Criteria for comparison. Because the effect of a single CV on the total EBV is often small, and to avoid investigating very small differences in accuracy, we concentrated on the local EBV (LEBV) of each of the 200kb-regions surrounding the three CVs. As the criterion for comparison of two analyses we used the correlation between the LEBV coming from a part data set to those from the complete data set, following Legarra and Reverter (2018). This criterion estimates the ratio of the accuracies of the part data set to that of the complete data set. Here complete data means more SNPs included in the analysis or more prior information about these SNPs such as which is the CV.

Table 1: Ratio of accuracies of LEBV for the alternative analyses.

Data		Ratio of accuracies:		
Partial	Complete	ABCG2	DGAT	GHR
AV-CV	**AV**	0.993	0.999	0.726
AV	**AV+CV**	0.805	0.9996	0.998
TOP10	**AV**	0.852	0.998	0.781
TOP10	**AV+CV**	0.758	0.998	0.754
TOP10	**CV**	0.154	0.531	0.730
CV	**AV**	0.094	0.551	0.996
CV	**AV+CV**	0.662	0.547	0.999

Results

The accuracy of **AV-CV** versus **AV** shows it is only in the case of GHR important to include the CV in the BayesGC analysis (Table 1). The latter is presumably because the surrounding SNPs did not fully pick-up the GHR-CV effect and the accuracy of **AV-CV** LEBV is reduced by ~27%. The **AV** versus **AV+CV** comparison shows that in the case of GHR, the **AV** model fitted the GHR-CV with a high posterior probability anyway, so giving it a prior probability of 1 hardly affected the LEBV. For DGAT, probably most of the LEBV was due to the effect

of other SNPs and fitting the DGAT-CV with a prior probability of 1 did not change this much. However, the ABCG2-CV was somewhat important to the LEBV of its region and giving it a prior probability of 1 did improve the accuracy of the LEBV. The **TOP10** versus **AV(+CV)** comparisons show that fitting only the top-10 SNPs in a region is not sufficient in the case of ABCG2 and GHR. Other SNPs also seem to contribute significantly. In the case of DGAT, the top-10 SNPs could predict the LEBV of the entire region. The **TOP10** versus **CV** comparison shows that the top10 SNPs could generally not predict the effect of the CV. The **CV** versus **AV(+CV)** comparisons show that the CV alone often does not explain all genetic variation that could be explained in the region. For the DGAT-CV this may be expected due to its imputation inaccuracy. The GHR-CV is an exception to this result where fitting its effect alone yielded virtually the same LEBV as fitting all SNPs in the region.

Discussion

The results showed that for some CVs a higher prediction accuracy was achieved when the genotype data included the CV whereas for others the surrounding SNPs were able to pick-up its effect. Knowing the CV (giving it a prior probability of 1 in the analysis) improved prediction accuracy for some CV, especially if the analysis had insufficient power to identify the CV. Generally, including all SNPs next to the CV in the analysis seemed to increase accuracy. Fitting the top-10 SNPs in a QTL region seemed generally to result in a reduced prediction accuracy compared to fitting all SNPs.

Albeit, for different reasons, the results indicate that for all three cases prediction accuracies would be improved by knowing the CVs: for the ABCG2-CV prediction accuracy increased because the variable selection analysis did not clearly identify this SNP as the CV; for the DGAT-CV the imputation accuracy was a limiting factor and knowing the importance of this CV would ensure that it would be amongst the set of genotyped SNPs; and without prior knowledge on the GHR-CV it may have been excluded from the analyses by any of the data quality control filters, but Table 1 shows that all analyses excluding the GHR-CV yield markedly reduced prediction accuracies. Thus, it seems important to identify the CVs, from a perspective of (i) understanding the biology of phenotypes, (ii) identifying targets for genome editing; and, as shown here, (iii) to improve accuracies of genomic predictions.

Acknowledgements: Funding from Norwegian Research Council: 255297 & 309611.

References.

Blott S., Kim J-J, Moisio S., Schmidt-Kuntsel A., Cornet A. et al. (2003) Genetics 163:253-66. doi: 10.1093/genetics/163.1.253.

Cohen-Zinder M., Seroussi E., Larkin D.M., Loor J.J., Everts-van der Wind A. et al. (2005) Genome Res. 15:936-44. doi: 10.1101/gr.3806705

Grisart B. Farnir F., Karim L., Cambisano N., Kim JJ. et al. (2004) Proc Natl Acad Sci 101:2398-403. doi: 10.1073/pnas.0308518100

Hayes B.E. and Goddard M.E. (2001) Gen. Sel. Evol. 33 doi: 10.1186/1297-9686-33-3-209.

Jenko J., Gorjanc G., Cleveland M.A., Varshney R.K., Whitelaw C.B. et al. (2015) Genet. Sel. Evol. 47-55. doi: 10.1186/s12711-015-0135-3

Legarra A. and Reverter A. (2018) Gen. Sel. Evol. 50:53. doi: 10.1186/s12711-018-0426-6

Meuwissen T:H.E., VandenBerg I., Goddard M.E. (2021) Gen. Sel. Evol. 53:19. doi: 10.1186/s12711-021-00607-4

Visscher P.M. (2020) ICQG6, Brisbane, Australia, available at: icqg6.org/program/.