What Is The Best Phenotype For Genome-Wide Association Studies In Data With Defined Pedigrees?

C.C. Ekine[*], S.J. Rowe[†], S.C. Bishop[†] and *D.J. de Koning*[†]

Introduction

With the availability of genome-wide SNP chips for the main livestock species, genome-wide association studies (GWAS) are rapidly superseding traditional QTL mapping experiments in livestock. However, many GWAS experiments utilize family data that is already in place from previous QTL mapping experiments. Furthermore, the uptake of genomic selection, in particular by the dairy cattle industry, has resulted in the availability of large volumes of data that are also amenable to GWAS. In both cases, and arguably any GWAS involving livestock, we have to account for known family relationships in order to control the false positive rate. One way to account for pedigree relationship is to model a polygenic effect alongside the SNP effects in a so-called measured genotype (MG) approach. Because of the computational load of estimating often-complex random effects alongside every single SNP effect, a two-step approach was proposed whereby the residuals from a polygenic model are used to estimate SNP effects (Aulchenko *et al.* 2007). This was coined the GRAMMAR approach and it was shown to have a conservative type I error while maintaining favourable statistical power compared to competing methods (Aulchenko *et al.* 2007). However, it underestimates the size of the QTL effects (Crooks *et al.* 2009; Lam *et al.* 2009). The original paper only studies a small range of heritabilities and QTL effects and does not address the utility of estimated breeding values (EBV) for GWAS. It could be argued that for high heritabilities the family contribution to the EBV becomes less important and the relative contribution from the individual's own observation increases. In this study, we compare the performance of differing trait definitions across a range of heritabilities and QTL effects for different GWAS analyses. We aim to quantify the type I error, the statistical power and the bias in estimated QTL effects for the different approaches.

Material and methods

Simulations. Three pedigree structures were considered across five heritabilities and four SNP effects representing over seventy study scenarios. A human-type pedigree and livestock (pig) pedigree were compared along with a real pig pedigree with either two or five generation of individuals. The latter was to test the impact of the depth of pedigree information on performance of the EBV approach. For the human pedigree we simulated 337 nuclear families of 3 full-sibs with unrelated parents. For the pig pedigree we simulated ten sires, each mated to ten dams that had 10 or 11 offspring, resulting 1010 measured individuals for analysis. For the real pig pedigree we randomly sampled 1010 offspring from a total pedigree of ~5000 commercial pigs and included either two or five generations of

[*] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT United Kingdom
[†] The Roslin Institute and R(D)SVS, University of Edinburgh, Roslin, EH25 9PS, United Kingdom

prior pedigree information. The MORGAN genedrop program (George *et al.* 2005) was used to simulate genotypes at marker loci, trait genotypes and polygenic values contributing to the quantitative traits of the simulated pedigrees under all models considered. Quantitative traits were defined as the sum of the SNP effect, the polygenic effects and a random environmental error. Two SNPs were analysed for association, one SNP was not linked with the trait of interest and used for studying the type I error. For studying power, a causal SNP with an additive effect of 4.0 and a minor allele frequency of 0.3 was simulated explaining 0.5, 1, 2 or 3% of the total variation in the trait. Because (Aulchenko *et al.* 2007) showed no effect of the allele frequency of the SNP, we did not test the impact of different allele frequencies in this study. The simulated traits had a total heritability of 0.30, 0.40, 0.50, 0.60, and 0.80. For each scenario one thousand replicates were simulated and analyzed.

Statistical analyses. The simulated data was analysed using four different approaches:
1) *Measured genotype* (**MG**): The SNP to be tested for association was fitted as a covariate in a polygenic model (1) which accounted for familial relatedness of individuals in the pedigree. The SNP effect and heritability were estimated together using this model:

$$y = \mu + Wa + Zu + e \quad (1)$$

where y is the vector of trait values, μ is the overall mean, a, u, and e are vectors of marker effects, additive polygenic effects (random), and random residuals, respectively. W and Z are incidence matrices related to marker and polygenic effects, respectively.
2) *GRAMMAR*: The GRAMMAR approach consists of two steps: the first step accounts for the familial dependence among family members and the second step tests the single SNP effect on the remaining variation by analysis of variance (ANOVA).
Step 1: For the expression values of each probe set we fitted the following mixed model (with the same variable definitions as (1)) without the marker effect:

$$y = \mu + Zu + e \quad (2)$$

Step 2: Using the residuals from Step 1 as the new quantitative trait, the marker genotype effect of each SNP on each trait was tested by linear regression
3) *Ignoring family structure* (**IF**): The IF analysis is comparable to the second step of the GRAMMAR analysis. It uses a direct regression of the phenotypic observation on the SNP data and does not take account of family relationships.
4) *Estimated breeding value* (**EBV**): Similar to GRAMMAR but in this analysis the EBV from the polygenic model is used as a trait score for the association study.

All analyses were performed in ASReml (Gilmour *et al.* 1998). The type I error of each approach was estimated using the unlinked SNPs and a threshold of F > 3.85 ($P < 0.05$). The statistical power to detect the causal SNP was estimated using either the tabulated F-threshold of 3.85 or an empirical threshold based on a 5% error rate for the unlinked SNP.

Results and discussion

The false positive rate (FPR) for the four methods is summarised for the pig population in Table 1. The GRAMMAR approach is very conservative, while the measured genotype performs very close to the tabulated threshold. Either using the EBV or ignoring family relationships resulted in much higher levels of false positives (Table 1). The FPR increases for IF with increasing heritability while it decreases for EBV. However, in all scenarios considered the FPR for either of these methods was high. In all cases, the power of

GRAMMAR was similar to MG, but the EBV approach had markedly reduced power. For the simulated human pedigrees the FPR for using EBV was > 0.10 (between 0.12 and 0.16) across all scenarios while using IF had almost acceptable FPR for low heritabilities (0.05-0.08) but higher FPR (0.06 to 0.12) for heritabilities > 0.50). This shows that the biases from using EBVs are less extreme in the human pedigrees, as relatives contribute less information to the estimate.

Table 1: Type I error and empirical power of SNP analysis in a simulated pig pedigree

h^2	SNP effect	Type I error or False positive rate				Empirical power[5]			
		MG[1]	GRA[2]	EBV[3]	IF[4]	MG	GRA	EBV	IF
30%	0.50%	0.05	0.010	0.66	0.29	0.49	0.47	0.10	0.21
	1%	0.05	0.018	0.61	0.25	0.78	0.77	0.21	0.47
	3%	0.06	0.018	0.61	0.26	0.99	1.00	0.44	0.89
50%	0.50%	0.05	0.011	0.57	0.35	0.52	0.48	0.09	0.16
	1%	0.04	0.009	0.59	0.36	0.78	0.76	0.18	0.30
	3%	0.04	0.006	0.58	0.35	1.00	0.99	0.37	0.98
80%	0.50%	0.04	0.01	0.51	0.43	0.60	0.60	0.11	0.14
	1%	0.05	0.004	0.52	0.44	0.82	0.80	0.20	0.27
	3%	0.05	0.01	0.53	0.44	1.00	1.00	0.42	0.57

[1]MG: Measured Genotype, [2]GRA: GRAMMAR, [3]EB: Estimated Breeding Value, [4]IF: Ignoring Family, [5]Empirical threshold corresponding to 5% type I error in unlinked SNP.

The SNP estimates from different methods are summarized in Table 2 for the simulated pig pedigree. MG and IF both give unbiased estimates of the SNP effects while GRAMMAR and EBV underestimate the effect. This confirms earlier observations that GRAMMAR underestimates the gene effect (Aulchenko *et al.* 2007; Crooks *et al.* 2009).

Table 2: Estimates of SNP effects (S.E.) in a simulated pig pedigree.

h^2	SNP effect	Model			
		MG[1]	GRA[2]	EBV[3]	IF[4]
30%	0.50%	4.09 (2.01)	2.13 (1.13)	1.95 (2.70)	4.07 (3.24)
	1%	4.04 (1.53)	2.11 (0.82)	1.98 (1.96)	4.09 (2.32)
	3%	3.97 (0.86)	2.09 (0.53)	1.88 (1.13)	3.97 (1.26)
50%	0.50%	3.94 (2.05)	1.52 (0.85)	2.47 (3.61)	3.98 (3.89)
	1%	3.97 (1.49)	1.54 (0.64)	2.34 (2.58)	3.88 (2.78)
	3%	3.98 (0.83)	1.53 (0.42)	2.39 (1.58)	3.92 (1.64)
80%	0.50%	4.09 (1.91)	0.73 (0.48)	3.38 (4.63)	4.11 (4.73)
	1%	4.02 (1.42)	0.73 (0.39)	3.28 (3.28)	4.01 (3.36)
	3%	3.99 (0.78)	0.74 (0.35)	3.19 (1.89)	3.97 (1.87)

[1]MG: Measured Genotype, [2]GRA: GRAMMAR, [3]EB: Estimated Breeding Value, [4]IF: Ignoring Family; the simulated SNP effect is 4.0

A clear trend is apparent for the effect of the heritability on the estimates. With increasing heritability, the GRAMMAR estimates are further biased downwards while those from the EBV approach become less biased (Table 2). When looking at individual replicates it can be shown that the sum of the GRAMMAR and EBV estimates provide an unbiased estimate of the SNP effect. This could provide a quick way to re-estimate the true effect of significant SNPs following GRAMMAR analyses, rather than re-estimating the effect in a full mixed model as suggested previously (Aulchenko *et al.* 2007). While the type I error was less extreme for the EBV approach in the human pedigree compared to the pig pedigree, the SNP estimates were biased to the same degree for both pedigree types. The analysis of the real pig pedigree (with simulated SNP data) showed that extending the pedigree increases the FPR when using the EBV as a phenotype score (Figure 1).

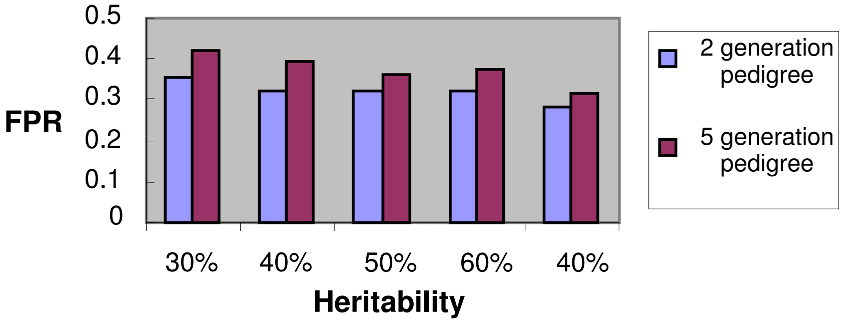

Figure 1: The false positive rate (FPR) using the EBV as a trait score when using 2 or 5 generations of pedigree information from a commercial pig population.

Conclusion

These simulations show that extreme care must be taken when defining a trait score for GWAS in pedigreed populations. In the scenarios simulated, using EBV can give unacceptable high rates of false positive results. This is due to the fact that when an individual only has a single observation; information from relatives constitutes a large part of the EBV. This may be different with progeny testing scenarios where the bulk of the information contributing to the EBV comes from offspring. We have shown that GRAMMAR is a good alternative to detect associations and that combined with EBV can give unbiased estimates of the SNP effect.

References

Aulchenko, Y.S., de Koning, D.J., and Haley, C. (2007) *Genetics* 177:577-585.

Crooks, L., Sahana, G., de Koning, D.J. *et al.* (2009) *BMC Proceedings* 3:S2

George, A.W., Wijsman, E.M., and Thompson, E.A. (2005) *Hum. Hered.* 59:98-108.

Gilmour, A., Cullis, B., Welham, S. *et al.* (1998)

Lam, A., Powell, J., Wei, W.H. *et al.* (2009) *BMC Proceedings* 3:S6