HHS Public Access

Author manuscript

J Anim Breed Genet. Author manuscript; available in PMC 2015 June 09.

Published in final edited form as:

J Anim Breed Genet. 2014 June ; 131(3): 163–164. doi:10.1111/jbg.12091.

On the genomic analysis of data from structured populations

G. de los Campos[1] and **D. Sorensen**[2]

G. de los Campos: gcampos@uab.edu

[1]University of Alabama at Birmingham, Birmingham, AL USA

[2]Aarhus University, Tjele, Denmark

The availability of dense SNP panels allows assessing similarity between distantly related individuals, including those with different genetic background. This has renewed the interest in the analysis of data from heterogeneous populations. Examples include the genomic analysis of data from multiple breeds (Hayes *et al*. 2009, Gen. Sel. Evol. **41**, 51), or from structured populations (e.g. de los Campos *et al*. 2009, Genetics **182**, 375–385; Daetwyler *et al*. 2012, J. Anim. Sci. **90**, 3375–3384).

Whole genome regression (*WGR*) *methods* (Meuwissen, Hayes, and Goddard, 2001, Genetics **157**, 1819– 1829), where allele substitution effects are assumed to be homogeneous across subjects, have been used for the analysis of data from structured populations (e.g. Hayes *et al*. 2009, Gen. Sel. Evol. **41**, 51; de los Campos *et al*. 2009, Genetics **182**, 375–385). This approach allows borrowing information across groups and, under some circumstances, can increase prediction accuracy. For example, in a combined analysis of Holstein and Jersey data, Hayes *et al*. (2009, Gen. Sel. Evol. **41**, 51) showed that prediction accuracy of estimated breeding values could be increased, relative to a within-breed analysis, in Jerseys but not in Holsteins. However, assuming that marker effects are constant across groups ignores the fact that dominance, epistasis or differences in the marker–QTL LD (linkage disequilibrium) patterns can lead to group-specific marker effects.

Principal Components (PCs) methods are commonly used in genome-wide association studies to account for population structure (e.g. Price *et al*. 2006, Nat. Gen. **38**, 34–41; Marchini *et al*. 2004, Nat. Gen. **36**, 512–517). Drawing on these ideas, some authors suggested expanding WGRs such as the G-BLUP (genomic best linear unbiased predictor) by adding marker-derived PCs as fixed effect covariates. This approach has been used to account for stratification in the estimation of variance components (e.g. Yang *et al*., 2010, Nat. Genet. **42**, 565–569) and in the prediction of breeding values (e.g. Daetwyler et al., 2012, J. Anim. Sci. **90**, 3375–3384). However, Janss *et al*. (2012, Genetics **192**, 693–704) demonstrated that adding eigenvectors as fixed effects in G-BLUP can create important inferential problems. Indeed, Gaussian processes, including the G-BLUP, are equivalent to a random regression on all marker-derived PCs (e.g. de los Campos *et al*., 2010, Genetics Research **92**, 295–308). Therefore, the PCs that are added as fixed effects in the G-BLUP, typically those with the largest eigenvalue, enter twice in the model, and this can have

adverse effects on inferences on variance components. The problem is aggravated by the fact that in G-BLUP, despite the random nature, the effects of eigenvectors with large eigenvalues are effectively estimated as fixed effects. In their article, Janss *et al*. showed how the standard G-BLUP, parameterized using PCs, can be used to draw inferences and predictions based on all or some PCs in a coherent statistical framework. This approach should be preferred over the one using PCs as fixed effects in a G-BLUP model. However, regardless of how PCs are dealt with, when only a subset of PCs is used for inferences, the connection with the original model is lost and parameters have no genetic interpretation.

Back to basics

Allele substitution effects are defined as the regressions of genetic values on allele content (e.g. Falconer and Mackay, 1996). It is well established that such effects, and functions thereof, are allele frequency dependent. Moreover, even if allele substitution effects of QTL are constant across groups, differences in the marker–QTL LD patterns can translate into group-specific marker effects. In short, there are important reasons that support the idea that in heterogeneous populations, marker effects should be allowed to vary between groups. Neither the standard G-BLUP nor PC methods address this fundamental problem. We argue that alternative approaches are needed for the analysis of data from structured populations.

Interactions can be used to model heterogeneity

Schulz-Streeck *et al*. (2012, Crop Sci **52**, 2465–2461) used this approach in G-BLUP for the analysis of multiple breeding populations, assuming, however, homogeneous genetic variance across groups. A more general approach models marker effects in subpopulation g as the sum of an effect common to all groups, b_{j0}, plus a term, b_{jg}, of group-specific deviations; that is, for marker j: $\beta_{jg} = b_{j0} + b_{jg}$. The term common to all groups allows using information from all groups, and the group-specific deviations allow for marker effects and functions thereof (genomic values and variances) to vary between groups. The hyperparameters of the distribution of b_{j0} and of b_{jg} control the trade-offs between borrowing of information between groups and allowing effects to vary between groups. These can be estimated from data using standard Bayesian methods. With this approach, the covariance between marker effects across any two groups is the variance of the common group effect, $\sigma^2_{b0} = Var\,(b_{j0})$, and each group has its own variance. Whenever the number of groups is greater than two, this treatment imposes restrictions on the variance–covariance matrix of marker effects.

A more general approach regards the vectors of effects of markers across the N_g groups $\beta_j = (\beta_{j1}, \ldots, \beta_{jN_g})'$ (for marker j) as IID draws from a multivariate distribution. For instance, β_j may be a draw from a multivariate normal density (e.g. Olson *et al.* 2012, J Dairy Sci., **95**: 5378–83.). However, in principle, the multivariate approach is not restricted to Gaussian models and could be applied with any of the priors commonly used in WGR models. Within group g, the variance–covariance matrix of marker effects is $I\sigma^2_{\beta_g}$ and between groups g and g', $I\sigma_{\beta_g\beta_{g'}}$(subscripts here are labels for groups, and the dimension of the identity matrix I is equal to the number of markers). In a G-BLUP context, identifiability of the covariance

terms requires that the genomic relationship matrix describing relationships between members of different groups is different from zero.

The interaction and the multivariate models lead to group-specific genomic variances and heritabilities. Perhaps more importantly, the approach leads to group-specific breeding values. For the *i*th subject, with genotype x_i, and regardless of the group to which it belongs, one can define customized breeding values that predict the expected performance of its progeny when bred with a randomly selected mate of the *g*th group, using $u_{ig}=x_i'\beta_g\ (g=1,\ldots,N_g)$. In this set-up, the ranking of individuals can change from group to group, opening new opportunities for mate group allocation and intergroup breeding.

The interaction and multivariate approaches are statistically similar in spirit to early work on multienvironment models by Falconer (1952, Am. Naturalist, LXXVI, **83**: 293–298) and to work on the joint analysis of pure-bred and cross-bred data (e.g. Wei and van der Werf, 1994, Anim. Prod. **59**, 401–413; Cecchinato *et al*., 2010, J. Anim. Sci. **88**, 481–490), recently extended to genomic prediction (Christensen *et al*. 2014, Gen. Sel. Evol. **46**, 23). However, the approaches discussed here are not restricted to the combined analysis of pure- and cross-bred individuals and can be used, in principle, with any genetic grouping scheme such as breeds, breeding programmes or clusters derived from genomic data.

Acknowledgments

GDLC acknowledges financial support from NIH Grants GM099992 and GM101219, and DS acknowledges financial support from the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research.