Reciprocal Recurrent Genomic Selection for Total Genetic Merit in Crossbred Individuals

B.P. Kinghorn[*], J.M. Hickey[*], J.H.J. van der Werf[*]

Introduction

Genomic Selection (GS: Meuwissen et al, 2001) aims to provide genomic estimated breeding values (*gEBVs*) as criteria for selection by exploiting associations between genetic markers and phenotypes. This process generally assumes additive genetic merit, implicitly regressing individual phenotype on the number of specific alleles carried, separately for each locus. gEBVs are thus based on estimated allele substitution effects at QTL or at linked marker loci. This paper aims to give proof of concept that genomic selection can be used to breed for dominance effects as well as additive effects, without having to fit both these effects in statistical models. This leads to increasing heterosis in crossbred animals though genomic selection within the contributing parental lines. Should this work under real conditions, the benefits could be substantial, particularly in the pig and poultry industries.

In this paper the target is genetic merit in crossbred individuals. This has been well addressed by Ibánẽz-Escriche et al. (2009), but without consideration of non-additive QTL effects. This paper aims to exploit within-locus dominance deviations as well as additive effects while selecting *within* purebred parental lines. This can be done by estimating allele substitution effects on crossbred phenotypes, separately for each parental line. Consider the average effect of an allele substitution in gametes of line A, evaluated on its phenotypic impact in the AxB crossbred genotype. This depends on allele frequency in gametes contributed from line B, and is thus the same as the average effect of an allele substitution within line B, assuming no epistatic effects.

This leads to the contention that we should select Line A individuals based on a "genomic key" (a set of weightings to calculate *gEBVs* from genotype) calibrated in Line B, and vice versa. However, this strategy is deficient, because:

1. Some loci segregating in line A will not be segregating in line B (or have extreme frequency giving low power to estimate effect), such that the key developed in line B will not exploit these loci.
2. Marker distribution in Line B has no colinearity with pedigree in Line A, and can thus not help improve accuracy of Line A *gEBVs*. This can be a disadvantage where calibration is based on the same Line A population as those to undergo marker assisted breeding. However, it may be an advantage if the target populations to undergo selection are genetically distant from those involved in the calibration exercise.

[*] School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia

3. Phenotypes in crossbred individuals reflect the epistatic and environmental interactions that are relevant to the target commercial product, and are thus more relevant than phenotypes in the pure lines.

This paper proposes an alternative strategy, whereby phenotypes and genotypes used for genomic calibration all come from AxB crossbred animals. We want to evaluate the effects of allele substitutions in line A as they impact on crossbred phenotype. We can do this by inferring the gametotypes (genome-wide alleles carried) of the gametes contributed to each crossbred individual, as used by Ibánẽz-Escriche et al. (2009). The genomic key for a parental line is thus derived from the crossbred phenotypes and the gametotypes contributed by that parental line for each crossbred individual. This is done for each parental line in a two line cross. More complex crosses involve further considerations (see Ibánẽz-Escriche et al., 2009).

Note in relation to point 1. above that this strategy does not require segregation in the alternate line. Loci that are segregating in only one of the parental lines will be exploited in the resulting *gEBV*s for that line.

This strategy is denoted reciprocal recurrent genomic selection (RRGS). This is inspired by the concept of reciprocal recurrent selection (RRS) whereby two lines undergo repeated cycles of selection based on the performance of crosses between them.

Classic selection and RRS are based on phenotypes as selection criteria. However, at the stage of implementation, GS and RRGS are based on genotypes of candidates, such that selection and mating does not have to wait until phenotypes are recorded on candidates, or on their crossbred progeny.

RRGS is analogous to RRS if genomic key calibration is carried out in each generation of crossbreds. However, just as the move from classic selection to GS enables us to avoid recording phenotypes in each generation, a move from RRS to RRGS enables us to avoid recording crossbred phenotypes in each generation. We can select within the pure lines using appropriate *gEBV*s without ongoing retrieval of information from crosses.

This is an important advantage for RRGS, as it allows short generation intervals, without the need to wait for crossbred progeny to be recorded before selecting within the parental lines. We may choose to monitor progress in the crossbred populations, and possibly refine and update the genomic keys, but these activities do not require a delay in making selection decisions within the parental lines.

Methods

Long range phasing for inference of gametotypes

The task here is to infer phase of high density genotypes – to derive the direction of inheritance at each locus – for crossbred animals that will be phenotyped and used for genomic calibration.

Kong et al. (2008) developed a powerful approach for phasing high density genotypes. Our implementation of this approach is based on Hickey et al. (2009, 2010) and Kinghorn et al. (2009). This does not require pedigree information, but in the current setting is aided by knowledge of the line of origin of each parent, even if the parent itself is unknown.

A number of individuals from each pure line are genotyped, ideally individuals that have made greater genetic contributions to their lines and to the crossbred individuals to be phased. The phasing method infers regions where identity by descent (IBD) exists between the current crossbred proband and genotyped individuals in the parental lines. This gives power to infer phase wherever a pure line individual is homozygous at a given locus. Even if all such individuals are heterozygous at this locus, there is prospect to make a phasing call, as these individuals themselves can be phased in a similar manner.

In simulations, an average of over 98.5% of loci were correctly phased, with most of the rest being uncalled, while in real data sets typically 92% to 99% of the loci were called (John Hickey, pers. comm.). This shows at least one route to inferring gametotypes in crossbred animals, enabling application of RRGS. However, for the simple simulations in this paper, knowledge of phase was assumed.

Simulations for illustration

The purpose of these simulations was to demonstrate underlying properties of the method proposed, without testing the merit of its performance. This means that most sources of sampling error are unashamedly removed and some simplifying assumptions made.

Simulation of SNP genotypes: The first phase of SNP simulation was used to generate a library of gametotypes for each line, with appropriate linkage disequilibrium within and between lines, and appropriate genetic distance between lines. Ten thousand biallelic loci were initially generated with location sampled uniformly on a map distance of one Morgan, and with frequencies sampled using a mean of p = 0.5, and SDp = 0.1, implemented by taking the mean of n samplings from (0,1), with probability p of sampling 1 on each occasion, where n is the nearest integer to $p.(1\text{-}p)/(SDp*SDp)$. The two alleles at each locus were labelled 0 and 1. These loci were propagated in a "burn-in" simulation phase using Mendelian sampling and recombination fractions assuming Kosambi's map function, in a panmictic population of 250 individuals per generation for 1000 generations. Lines 1 and 2 were propagated as a common population until generation 800, and were split for the last 200 generations. Over all loci that were still segregating, mean frequency was 0.491 and mean frequency difference between lines was 0.278.

Loci segregating with a minor allele frequency of at least 0.1 in at least one of the two lines were accepted into the breeding program. The number of such loci was denoted n_{SNP}.

Gametotypes in the final generation of the burn-in simulation were sampled at random into the foundation individuals of the corresponding parental populations in the "breeding" simulation. SNP genotypes were propagated in the simulated line and cross populations taking account of the recombination fraction between loci, and with no effect of SNP genotypes on selection.

Simulation of QTL effects and phenotype: From the n_{SNP} loci, n_{QTL} were chosen at random as QTL. For each QTL, the effect on phenotype of 0, 1 and 2 copies of the favourable allele was set to –*a*, *d* and +*a* respectively, where both *a* and *d* are the same for each locus in this initial simple illustration. A broad-sense heritability (V_G/V_P) of 0.5 was adopted, with environmental deviations generated accordingly.

Population size and structure: Within both line 1 and line 2, n_M males were mated to n_F females each generation, with n_F / n_M females allocated per male, to produce n_P progeny per female. Selection of parents was carried out randomly from generation 0 to generation 3.

In addition, crossbred matings were made as of generation 3: n_M line 1 males were mated to n_F line 2 females, to produce n_P crossbred progeny per female. Genomic calibration was carried out in this generation, and the resulting key(s) were used to estimate *gEBVs* for selection of parents in all subsequent generations.

Genomic calibration: Crossbred progeny from generation 3 parents were used to make the first genomic calibration. Further calibrations were optionally made at specified intervals. Three genomic selection strategies were tested:

RRGS: Reciprocal Recurrent Genomic Selection. Use line gametotypes and crossbred phenotypes to make one key for each line

XBGS: Crossbred Genomic Selection. Use crossbred (XB) genotypes and crossbred phenotypes to make one key for use across both lines

WLGS: Within Line Genomic Selection. Use pure line genotypes and phenotypes to make one key for each line.

For these simulations true QTL effects were used for calibration. For the k^{th} QTL, the effect of an allele substitution for lines 1 and 2 was calculated as follows:

RRGS: $\alpha_{1,k} = a + d.(1 - 2p_2)$ and $\alpha_{2,k} = a + d.(1 - 2p_1)$

XBGS: $\alpha_{1,k} = \alpha_{2,k} = a + d.(1 - p_1 - p_2)$

WLGS: $\alpha_{1,k} = a + d.(1 - 2p_1)$ and $\alpha_{2,k} = a + d.(1 - 2p_2)$

where p_i is the observed frequency of the favourable allele in line *i*. *gEBVs* for individual *j* in pure line *i* are then estimated as:

$$gEBV_{ij} = \sum_{k=1}^{n_{QTL}} G_{i,j,k}.\alpha_{i,k} \qquad 1.$$

where $G_{\mathrm{i},j,k}$ is the number of favourable alleles at locus *k* for individual *j* within line *i*.

More realistic simulations

Methods follow "Simulations for illustration" except as follows:

Simulation of QTL effects: The number of SNP loci chosen to act as QTL was set to n_{QTL} = 2000. QTL additive and dominance effects for locus *k* were modelled as a_k, where $2a_k$ equals the difference in merit between opposing homozygotes, and d_k equals the deviation of heterozygote merit from mid homozygote value, which is set to zero.

a_k were sampled from a gamma distribution, using shape and scale parameters equal to 0.4 and 1/1.66, following Ibánẽz-Escriche et al. (2009). Recognising an association of magnitude between a_k and d_k, values for d_k were sampled contingent on the previously sampled values for a_k, using $d_k = a_k\ (x + y\mathrm{N}(0,1))$. Values of *x* and *y* were chosen to give typical values for heterosis in the initially generated crossbred progeny. The effective number of QTL of equal effect was 567. A broad-sense heritability (V_G/V_P) of 0.5 was adopted, and genetic values, summed over loci, were linearly scaled to give variance of phenotypes = 1.

The information provided to the calibration analysis includes phenotypic, genetic and pedigree information:

Phenotypic information: This is the phenotypes for all crossbred animals (RRGS and XBGS) or all purebred animals (WLGS) in the prevailing generation.

Genetic information for RRGS: This is the *gametotype* for each gamete in the *crossbred* individual, delivered as 0 or 1 copies of Allele 1 at each SNP locus.

Genetic information for XBGS: This is the *genotype* for each *crossbred* individual, delivered as 0, 1 or 2 copies of Allele 1 at each SNP locus.

Genetic information for WLGS: This is the *genotype* for each *purebred* individual, delivered as 0, 1 or 2 copies of Allele 1 at each SNP locus.

Pedigree information: This is the whole pedigree across lines and crosses, as of the foundation of the simulated breeding population.

The domain of genetic information is just (0,1) under RRGS (as in the BSAM model of Ibánẽz-Escriche et al., 2009), because the information used for calibration is haploid, rather than diploid under XBGS and WLGS. However, the domain of *G* as used in equation 1 is (0,1,2) for all strategies, including RRGS, as *gEBVs* are to be estimated for diploid pure line individuals.

Results

Simulations for illustration

Parameters we set as follows: n_{SNP} = 2249, n_{QTL} = 100, n_m = 50, n_f = 100, n_p = 4, H^2=0.5. Figure 1 shows results for a = 0 and d = 1, with genomic calibration each generation (Figure 1a), and just once in the first crossbred generation (Figure 1b).

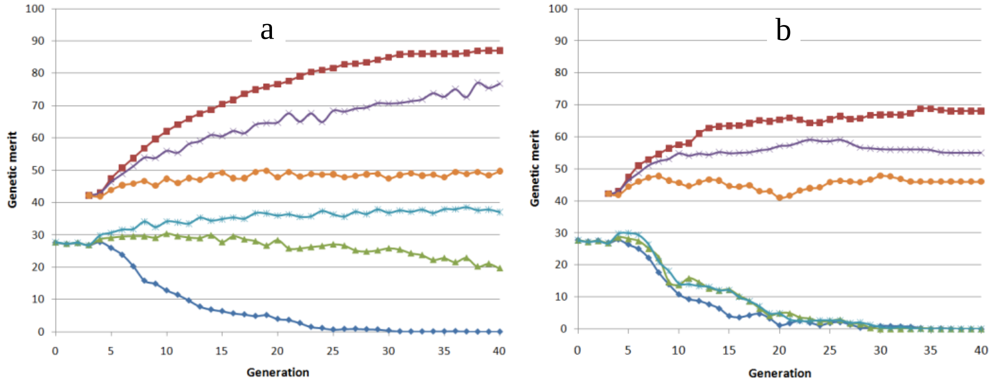

Figure 1: Responses in crossbred and purebred populations. The upper curves are the three crossbred populations (RRGS, XBGS, WLGS from top to bottom) and the others are the purebreds (WLGS, XBGS, RRGS from top to bottom on Fig1a). Fig1a: calibration each generation; Fig1b: calibration in first crossbred generation alone.

In this trivial scenario the best possible outcome is a genetic merit of 100 units (one unit at each of 100 loci) in the crossbred and 0 in both purebred lines, with complete fixation of opposing alleles in the pure lines. Figure 1a achieves 87 units for RRGS by generation 40, with full fixation at all loci within pure lines. Fixation was for the same allele in each parental line for 13 loci showing lack of ability to coordinate opposing fixation at all loci, with these population sizes and other conditions. WLGS calibrations always favour the more rare allele, driving merit towards $2pqd = 0.5$ in pure lines for loci that still segregate, and a higher value in crossbreds due to capture of dominance at loci where opposing fixation has occurred. XBGS is intermediate in its ability to use dominance in the crossbreds. The drive in this case is towards equal allele frequency *across* both lines, such that allele frequencies within the pure lines can drift to opposing extreme or fixed values, leading to increased heterosis compared to WGLS. In fact under XBGS, 33 and 32 loci became fixed in the two lines respectively.

Calibrating just once (Figure 1b) loses power because the initial calibration becomes less relevant as allele frequences change. For example, alleles that are rare in both parental lines will continue to be favoured for selection in both parental lines. With ongoing calibration under RRGS, the line with the more rare frequency at a locus will have higher selection value given to that allele. As allele frequencies deviate more between the lines, these selection values will become stronger. Figure 1b misses out on these advantages. Notice that without ongoing calibration, genetic merit in pure lines declines to zero under all treatments as the constant allele selection values drive all loci to fixation.

More realistic simulations

The numbers and effects of QTL were made more realistic, sampling from an exponential distribution and generating dominance values to give a realistic observed heterosis in the first crossbred generation (10.8%), as described in Methods.

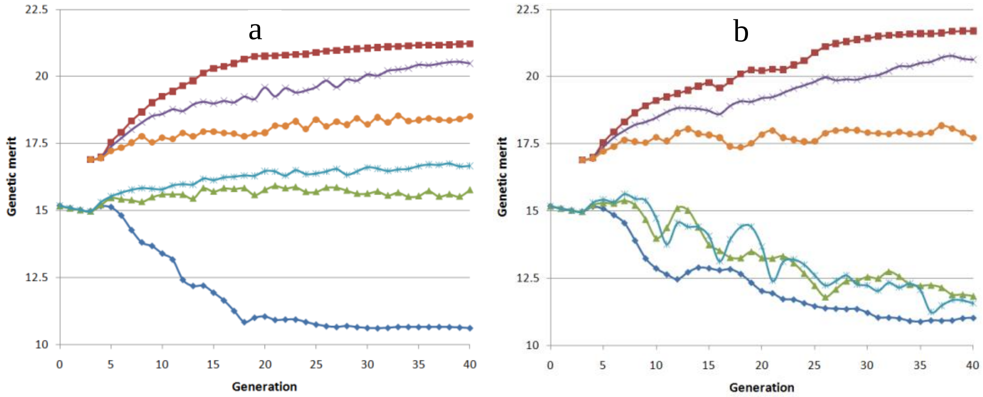

Figure 2: As for Figure 1, but with 2000 QTL sampled from an exponential distribution. However, in this case, for Fig2a calibration took place each generation, and for Fig2b calibration took place every fifth generation.

Parameters were set as follows: n_{SNP} = 2249, n_{QTL} = 2000 (Effective number =567), n_m = 50, n_f = 100, n_p = 4, H^2=0.5. Figure 2 shows results with genomic calibration each generation (Figure 2a), and just once in the first crossbred generation (Figure 2b).

This more realistic scenario gives similar results. RRGS gives substantially more response in the crossbred population, at the expense of genetic merit in the purebred lines. The effect of calibrating every five generations can be seen in oscillations in genetic merit (Fig 2b), particularly in the purelines, where alleles that have recently lost frequency are recognised every five generations for their increased value.

Discussion

Mate selection to capture heterosis in future generations has been addressed by Shepherd and Kinghorn (1998) and Li et al. (2006), based on overall line dominance effects and a single locus respecively. This paper concerns use of genome-wide marker information in Reciprocal Recurrent Genomic Selection. RRGS uses genomic selection within pure lines to give selection response in heterosis as part of the total genetic merit in crossbreds. However, these results do not demostrate likely performance in real conditions, particularly because the ability to phase genotypes and perfect estimation of QTL effects have been assumed. Proper assessment is warranted because of the big savings in generation intervals that RRGS gives over classic RRS.

Genomic calibration of gametotypes is likely to be less powerful than calibration on genotypes, as the former have a narrower domain (0,1 versus 0,1,2) and the crossbred phenotypes are of course affected by both parental gametotypes. Both the true number of QTL segregating and the number perceived in Bayesian analyses are likely to interact with the size of dominance deviations and the pattern of response under RRGS, and this warrants investigation.

There may be a case for optimal use of pure and cross gEBVs, depending on the accuracies that can be achieved, the relative contributions of pure and cross stock to harvested product, and the environments in which they perform.

If technically successful, application of this approach seems well suited to the pig and poultry industries, where fecundity is high and structured crossbreeding systems are well established. However, successful application in beef and sheep populations might be a component that would drive these industries further towards specialised sire and dam lines, and more structured crossbreeding structures.

Acknowledgement

The first author thanks Scott Newman for discussions on the use of mate selection in crossbreeding systems.

References

Hickey, J.M., Kinghorn, B.P., Tier, B., and van der Werf, J.H.J. (2009). *Proc. Assoc. Advmt. Anim. Breed. Genet.,* 18: 72-75.

Hickey, J.M., Kinghorn, B.P., and van der Werf, J.H.J. (2010). In *Proc 9th WCGALP*.

Ibánẽz-Escriche, N, Fernando, R.L., Toosi, A. and Dekkers, J.C.M. (2009). *Genet. Sel. Evol.,* 41:12-21.

Kinghorn, B.P., Hickey, J.M. and van der Werf, J.H.J. (2009). *Proc. Assoc. Advmt. Anim. Breed. Genet.,* 18: 76-79.

Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D.F., Stefansson, H., and Stefansson., K. (2008). *Nature Genetics,* 40:1068 – 1075.

Li, Yongjun, Van der Werf, J.H.J. and Kinghorn, B.P. (2006). *Genet. Sel. Evol.,* 38:147-165.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001). *Genetics*, 157:1819–1829.

Shepherd, R.K. and Kinghorn, B.P. (1998). In *Proc 6th WCGALP*, volume 25, pages 431-438.