Genetics Selection Evolution

Research

Deregressing estimated breeding values and weighting information for genomic regression analyses

Dorian J Garrick*[1,2], Jeremy F Taylor[3] and Rohan L Fernando[1]

Addresses: [1]Department of Animal Science, Iowa State University, Ames, IA 50011, USA, [2]Institute of Veterinary, Animal & Biomedical Sciences, Massey University, Palmerston North, New Zealand and [3]Division of Animal Sciences, University of Missouri, Columbia 65201, USA

E-mail: Dorian J Garrick* - dorian@iastate.edu; Jeremy F Taylor - taylorjerr@missouri.edu; Rohan L Fernando - rohan@iastate.edu
*Corresponding author

Abstract

Background: Genomic prediction of breeding values involves a so-called training analysis that predicts the influence of small genomic regions by regression of observed information on marker genotypes for a given population of individuals. Available observations may take the form of individual phenotypes, repeated observations, records on close family members such as progeny, estimated breeding values (EBV) or their deregressed counterparts from genetic evaluations. The literature indicates that researchers are inconsistent in their approach to using EBV or deregressed data, and as to using the appropriate methods for weighting some data sources to account for heterogeneous variance.

Methods: A logical approach to using information for genomic prediction is introduced, which demonstrates the appropriate weights for analyzing observations with heterogeneous variance and explains the need for and the manner in which EBV should have parent average effects removed, be deregressed and weighted.

Results: An appropriate deregression for genomic regression analyses is EBV/r^2 where EBV excludes parent information and r^2 is the reliability of that EBV. The appropriate weights for deregressed breeding values are neither the reliability nor the prediction error variance, two alternatives that have been used in published studies, but the ratio $(1 - h^2)/[(c + (1 - r^2)/r^2)h^2]$ where $c > 0$ is the fraction of genetic variance not explained by markers.

Conclusions: Phenotypic information on some individuals and deregressed data on others can be combined in genomic analyses using appropriate weighting.

Background

Genomic prediction [1] involves the use of marker genotypes to predict the genetic merit of animals in a target population based on estimates of regression of performance on high-density marker genotypes in a training population. Training populations might involve genotyped animals with alternative types of information including single or repeated measures of individual phenotypic performance, information on progeny, estimated breeding values (EBV) from genetic evaluations, or a pooled mixture of more than one of these information sources. In pooling information of different types, it is desirable to avoid any bias introduced by pooling and to account for heterogeneous variance so that the best use is made of available information.

Uncertainty as to whether or not EBV should be used directly or deregressed or replaced by measures such as

daughter yield deviation (DYD) [2], and the manner in which information should be weighted, if at all, has been apparent for some time in literature related to discovering and fine-mapping quantitative trait loci (QTL). Typically in fixed effects models with uncorrelated residuals, observations would be weighted by the inverse of their variances. Morsci et al. [3] pointed out the counter intuitive behavior of using the reciprocal of the variance of breeding values as weights in characterization of QTL and followed the arguments of Rodriguez-Zas et al. [4] in using reliability as weights. Rodriguez-Zas et al. [4] did analyses that were limited by features of the chosen software so EBV/2 (i.e. predicted transmitting ability PTA) were multiplied by the square root of reliability and analyzed unweighted. Georges et al. [5] deregressed PTA to construct DYD and weighted these using the inverse of the variance of the DYD. Spelman et al. [6] had direct access to DYD and similarly weighted these by the inverse of their scaled variance, equivalent to using the inverse of reliability as weights. Other researchers have reported the use of PTA [7], standardized PTA [7,8] or DYD weighted by respective reliabilities [8]. The uncertainty associated with using information for QTL discovery has recently been extended to genomic prediction. An Interbull survey [9] of methods being used in various countries for genomic prediction of dairy cattle reported that some researchers used deregressed proofs weighted with corresponding reliabilities, others used DYD weighted by effective daughter contributions, while yet others used EBV without any weighting. The objective of this paper is to present a logical argument for using deregressed information, appropriately weighted for analysis. For simplicity, we consider the residual variance from the perspective of an additive model but the deregression and weighting concepts extend to analyses that include dominance and epistasis.

Methods
An ideal model
Genomic prediction involves the use of genotypes or haplotypes to predict genetic merit. Conceptually, it involves two phases, a training phase where the genotypic or haplotypic effects are estimated, typically as random effects, in a mixed model scenario, followed by an application phase where the genomic merit of selection candidates is predicted from the knowledge on their genotypes and previously estimated effects from the training phase. The ideal data for training would be true genetic merit data observed on unrelated animals in the absence of selection. In that case, the model equation would be:

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{M}\mathbf{a} + \varepsilon, \tag{1}$$

where $\mathbf{g}$ is a vector of true genetic merit (i.e. breeding value BV) with $var(\mathbf{g}) = \mathbf{T}\sigma_g^2$, the scalar σ_g^2 is the genetic variance and $\mathbf{T}$ can be constructed using the theory from combined linkage disequilibrium and linkage analyses [10], μ is an intercept, $\mathbf{M}$ is an incidence matrix whose columns are covariates for substitution, genotypic or haplotypic effects, $\mathbf{a}$ are effects to be estimated, $var(\mathbf{M}\mathbf{a}) = \mathbf{G}\sigma_M^2$, $\mathbf{G}$ is a genomic relationship matrix [11-13], ε is the lack of fit, $var(\varepsilon) = \mathbf{E}\sigma_\varepsilon^2$, hopefully small and will be $\mathbf{0}$ if BV could be perfectly estimated as a linear function of observed marker genotypes. In different settings, $\mathbf{a}$ might be defined as a vector of fixed effects [14] or a vector of random effects [1]. Even when $\mathbf{a}$ is fixed, $\mathbf{M}\mathbf{a}$ is random because $\mathbf{M}$, which contains genotypes, is random. However, in genomic analyses $\mathbf{M}$ is treated as fixed because the analysis is conditional on the observed genotypes. The philosophical issues related to the randomness of $\mathbf{M}$ and $\mathbf{a}$ are discussed in detail by Gianola [15] but for our context it is sufficient to define $var(\mathbf{M}\mathbf{a}) = \mathbf{G}\sigma_M^2$ without explicitly specifying distributional properties of $\mathbf{M}$ or $\mathbf{a}$.

Genotypes used as covariates in $\mathbf{M}\mathbf{a}$ are unlikely to capture all the variation in true genetic merit, either because they are not comprehensively covering the entire genome, or because linkage disequilibrium between markers and causal genes is not perfect. Knowledge of $\mathbf{E}$ is required in the analysis whether $\mathbf{a}$ is treated as a fixed (e.g. GLS) or random effect (e.g. BLUP). In practice with experiments that involve related animals, it is unreasonable to assume $\mathbf{E}$ has a simple form such as a diagonal matrix since that implies a zero covariance between lack of fit effects for different animals, however, it can be approximated using knowledge on the pedigree using the additive relationship matrix, $\mathbf{A}$ [16]. These lack of fit covariances can be accommodated by fitting a polygenic effect for each animal, in addition to the marker genotypes [17], or accounted for by explicitly modeling correlated residuals. For a non-inbred animal, $\sigma_g^2 = \sigma_M^2 + \sigma_\varepsilon^2$, therefore $\sigma_\varepsilon^2 = \sigma_g^2 - \sigma_M^2$ and the proportion of the genetic variance not accounted for by the markers can be defined to be $c = \dfrac{\sigma_\varepsilon^2}{\sigma_g^2} = 1 - \dfrac{\sigma_M^2}{\sigma_g^2}$. The scalar c, will be close to 0 if markers account for most of the genetic variation and close to 1 if markers perform poorly.

A model using individual phenotypic records
In practice we do not have the luxury of using true BV as data in genomic prediction. A more common circumstance might involve training based on phenotypic observations that include fixed effects on phenotype denoted $\mathbf{X}\mathbf{b}$ where $\mathbf{X}$ is an incidence matrix for fixed non-genetic effects in $\mathbf{b}$. An appropriate model equation for phenotypes is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{g} + \mathbf{e}, \tag{2}$$

where $\mathbf{e}$ is a vector of random non-genetic or residual effects. In comparison to (1), the use of $\mathbf{y}$ for training involves the addition of the vectors $\mathbf{Xb}$ and $\mathbf{e}$ to the left- and right-hand side, inflating the variance and giving

$$\mathbf{y} = (1\mu + \mathbf{Xb}) + \mathbf{Ma} + (\varepsilon + \mathbf{e}), \qquad (3)$$

with $\mathrm{var}(\varepsilon + \mathbf{e}) = c\mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$ since $\mathrm{cov}(\varepsilon, \mathbf{e}') = \mathbf{0}$. This model can be fitted by explicitly including a random polygenic effect for ε, or by accounting for the non-diagonal variance-covariance structure of the residuals defined as var $(\varepsilon + \mathbf{e})$. Including a polygenic term is not typically done in genomic prediction analyses [12,18], and when undertaken does not seem to markedly alter the accuracy of genomic predictions [Habier D. Personal communication]. Assuming var $(\varepsilon + \mathbf{e})$ is a scaled identity matrix facilitates the computing involved in fitting this model, as the relevant mixed model equations can be modified by multiplying the left- and right-hand sides by the unknown scale parameter as is typically done in single trait analyses. However, this is not an option if residuals are heterogeneous, for example, because they involve varying numbers of repeated observations.

A model using repeated records on the individual

Consider the circumstance where the training observations are a vector $\bar{\mathbf{y}}_n$ representing observations that are the mean of n observations on the individual with n potentially varying. In that case, equation (3) becomes

$$\bar{\mathbf{y}}_n = (1\mu + \mathbf{Xb}) + \mathbf{Ma} + (\varepsilon + \bar{\mathbf{e}}_n), \qquad (4)$$

With $\mathrm{var}(\bar{\mathbf{e}}_n) = \mathbf{D}$, a diagonal matrix with elements $\mathrm{var}(\bar{e}_n) = \left(\frac{[1+(n-1)t]}{n} - h^2 \right)\sigma_p^2$ with σ_p^2 being the phenotypic variance, heritability h^2, and repeatability t. Ignoring off-diagonal elements of $\mathbf{E}$, the elements of the inverse of $\mathbf{R}$ with $\mathbf{R} = \mathrm{var}(\varepsilon) + \mathbf{D}$ would for non-inbred animals be $[c\sigma_g^2 + \mathrm{var}(\bar{e}_n)]^{-1}$. In fixed effects models, this matrix can be arbitrarily scaled for convenience. In univariate random effects models, a common practice is to formulate mixed model equations using the ratios of residual variance to variances of the random effects. Here, it makes sense to factor out the residual variance of one phenotypic observation, i.e. σ_e^2, from the expression for the residual variance of the mean of n observations. In this circumstance, a scaled inverse of the residual variance being $w_n = \sigma_e^2 / [c\sigma_g^2 + \mathrm{var}(\bar{e}_n)]$ or equivalently

$$w_n = \frac{1-h^2}{ch^2 + \frac{1+(n-1)t}{n} - h^2}, \qquad (5)$$

which can be used for weighted regression analyses treating marker effects as fixed or random. When $c = 0$, the genetic effects can be perfectly explained by the model, and for $n = 1$, a single observation on the individual, the weight is 1 for any heritability. Scaling the weights is convenient because records with high information exceed 1 and the weights are trait independent which is useful when analysing multiple traits with identical heritability and information content.

Offspring averages as data

In some cases the training data may represent the mean of p individual measurements on several offspring, rather than the mean phenotype of the genotyped animal. In that circumstance, the residual variance includes a genetic component for the mate and Mendelian sampling. For half-sib progeny means with unrelated mates and no common environmental variance, $\mathrm{var}(\bar{e}_p) = \frac{(0.75\sigma_g^2 + \sigma_e^2)}{p}$. However, the half-sib progeny mean contains only half the genetic merit of the parent, therefore the genotypic covariates need to be halved, or the mean doubled, in order to analyse data that includes records on genotyped individuals and records on off-spring of genotyped individuals. The variance for twice the progeny mean is $2 \times \mathrm{var}(\bar{e}_p) = \frac{4(0.75\sigma_g^2 + \sigma_e^2)}{p}$, and adding $\mathrm{var}(\varepsilon) = c\sigma_g^2$, factoring out σ_e^2 and inverting gives

$$w_p = \frac{1-h^2}{ch^2 + \frac{(4-h^2)}{p}}. \qquad (6)$$

For full-sib progeny means the intraclass correlation of residuals will include a genetic component and perhaps a common environmental component (e.g. litter, with variance σ_l^2 and $l^2 = \frac{\sigma_l^2}{\sigma_g^2}$ giving $\mathrm{var}(\bar{e}_p) = \sigma_l^2 + \frac{(0.5\sigma_g^2 + \sigma_e^2)}{p}$ for unrelated parents. Adding variation due to $c\sigma_g^2$ factoring out σ_e^2 and inverting gives

$$w_p = \frac{1-h^2}{ch^2 + l^2 h^2 + \frac{(1-0.5h^2)}{p}}. \qquad (7)$$

This expression can be used as weights in the fixed or random regression of full-sib progeny means on parent average marker genotypes.

Estimated breeding values as training data

An estimated breeding value, typically derived using BLUP, can be recognised as the true BV plus a prediction error. That is, $\hat{g} = g + (\hat{g} - g)$. Accordingly, training on EBV might be viewed as extending the model equation in (1) by the addition of the prediction error, in the same way that (3) was derived by the addition of a residual nongenetic component. The model equation would therefore be

$$g + (\hat{g} - g) = \hat{g} = 1\mu + Ma + (\varepsilon + (\hat{g} - g)). \qquad (8)$$

There are at least two issues with this formulation of the problem, which may not be immediately apparent, and which both result from properties of BLUP. The first issue is that the addition of the prediction error term to the left- and right-hand side of (8) actually reduces rather than increases the variance, despite the fact that diagonal elements of $\text{var}(\hat{g} - g)$ must exceed 0, in contrast to the addition of non-genetic random residual effects in (3). That is $\text{var}(g_i) > \text{var}(\hat{g}_i)$, whereas $\text{var}(g_i) < \text{var}(y_i)$, due to shrinkage properties of BLUP estimators [19]. Generally, $\text{var}(\hat{g}_i - g_i) = \text{var}(g_i) + \text{var}(\hat{g}_i) - 2\,\text{cov}(\hat{g}_i, g_i)$, but for BLUP $\text{cov}(\hat{g}_i, g_i) = \text{var}(\hat{g}_i)$ so that $\text{var}(\hat{g}_i - g_i) = \text{var}(g_i) - \text{var}(\hat{g}_i)$ implying $\text{var}(g_i) \geq \text{var}(\hat{g}_i) \geq 0$. The reduction in variance of the training data comes about because prediction errors are negatively correlated with BV as can be readily shown since $\text{cov}(g_i, \hat{g}_i - g_i) = \text{cov}(\hat{g}_i, g_i) - \text{var}(g_i) = \text{var}(\hat{g}_i) - \text{var}(g_i) < 0$. This means that superior animals tend to be underevaluated (i.e. have negative prediction errors) whereas inferior animals tend to be overevaluated. This is a consequence of shrinkage estimation and prediction errors being uncorrelated with EBV, i.e. $\text{cov}(\hat{g}_i, \hat{g}_i - g_i) = \text{var}(\hat{g}_i) - \text{cov}(\hat{g}_i, g_i) = 0$. In order to account for the covariance between the prediction errors and the BV, a model that accounted for such covariance would need to be fitted. Such models are computationally more demanding compared to models whereby the fitted effects and residuals are uncorrelated. The second issue resulting from the properties of BLUP, is that it is a shrinkage estimator, that shrinks observations towards the mean, the extent of shrinkage depending upon the amount of information. This is apparent if one considers the regression of phenotype on true genotype (i.e. BV) which is 1, whereas the regression of EBV on BV is equal to $r_i^2 \leq 1$, where r_i^2 is the reliability of the EBV (for animal i) or squared correlation between BV and EBV. In the context of any marker locus, the contrast in EBV between genotypes at a particular locus is shrunk relative to the contrast that would be obtained if BV or phenotypes were used as data, with the shrinkage varying according to r_i^2. We are, however, interested in estimating the effect of a marker on phenotype, but we get a lower value for the contrast if EBV with $r_i^2 \leq 1$ are used as data, rather than

using phenotypes. A further complication is that training data based on EBV typically comprise individuals with varying r_i^2. This problem can be avoided by deregressing or unshrinking the EBV.

Deregressing estimated breeding values

The solution to the model fitting problems associated with the reduced variance of EBV and the inconsistent regression of EBV on genotype according to reliability can both be addressed by inflating the EBV. Rather than fitting (8), we will fit the linearly inflated data represented as $K\hat{g}$ for some diagonal matrix K. That is, we will fit:

$$K\hat{g} = g + (K\hat{g} - g) = 1\mu + Ma + (\varepsilon + (K\hat{g} - g)), \qquad (9)$$

for some matrix K chosen so that $\text{cov}(g_i, k_i\hat{g}_i - g_i) = 0$ and $\text{cov}(k_i\hat{g}_i, g_i)$ is a constant. Since $\text{cov}(g_i, k_i\hat{g}_i - g_i) = k_i\text{var}(\hat{g}_i) - \text{var}(g_i)$ then this expression will be 0 when $k_i = \dfrac{\text{var}(g_i)}{\text{var}(\hat{g}_i)} = \dfrac{1}{r_i^2}$.

For this value k_i, $\text{cov}(k_i\hat{g}_i, g_i) = k_i\text{var}(\hat{g}_i) = \dfrac{\text{var}(g_i)}{\text{var}(\hat{g}_i)}\text{var}(\hat{g}_i) = \text{var}(g_i)$, a constant for all animals regardless of their reliability. Accordingly, the deregression matrix is $K = diagonal\{r_i^{-2}\}$ and the deregressed observations are $\hat{g}_i / r_i^2$. Note in passing that the nature of the deregression will depend upon the EBV base. Genetic evaluations are typically adjusted to a common base before publication, by addition or subtraction of some constant. The EBV should be deregressed after removing the post-analysis base adjustment or by explicitly accounting for the base in the deregression procedure [20]. To show the dependence of the deregression to the post-analysis base, supposes that EBV are adjusted to a base, b. Then a linear contrast in deregressed EBV without removing the base effect

is $\dfrac{\hat{g}_i - b}{r_i^2} - \dfrac{\hat{g}_j - b}{r_j^2} = \left(\dfrac{\hat{g}_i}{r_i^2} - \dfrac{\hat{g}_j}{r_j^2}\right) - \left(\dfrac{b}{r_i^2} - \dfrac{b}{r_j^2}\right) \neq \left(\dfrac{\hat{g}_i}{r_i^2} - \dfrac{\hat{g}_j}{r_j^2}\right)$

unless $r_i^2 = r_j^2$. Marker effects are typically estimated as linear combinations of data, and will therefore be sensitive to the base adjustment.

A deregressed observation represents a single value that encapsulates all the information available on the individual and its relatives, as if it was a single observation with $h^2 = r^2$. This can be shown by recognising that h^2 is the regression of genotype on phenotype. Taking the deregressed observation to be the

phenotype, $h^2 = \dfrac{\text{cov}(\hat{g}_i / r_i^2, g)}{\text{var}(\hat{g}_i / r_i^2)} = \dfrac{1/r_i^2\,\text{var}(\hat{g}_i)}{1/r_i^4\,\text{var}(\hat{g}_i)} = r_i^2$.

Training on deregressed EBV is therefore like training

on phenotypes with varying h^2. Provided $r_i^2 > h^2$, training on deregressed EBV is equivalent to having a trait with higher heritability. However, as explained later, we recommend removing ancestral information from the deregressed EBV.

Weighting deregressed information

Deregressed observations have heterogeneous variance when r^2 varies among individuals. The residual variance of a particular deregressed observation is

$\text{var}(\varepsilon_i + k_i\hat{g}_i - g_i) = \text{var}(\varepsilon_i) + \text{var}(k_i\hat{g}_i - g_i) = \text{var}(\varepsilon_i) + k_i^2\,\text{var}(\hat{g}_i) + \text{var}(g_i) - 2k_i\,\text{var}(\hat{g}_i)$

but $\text{var}(\hat{g}_i) = r_i^2\,\text{var}(g_i)$ and $k_i r_i^2 = 1$ so the residual variance expression simplifies to

$\text{var}(\varepsilon_i + k_i\hat{g}_i - g_i) = \text{var}(\varepsilon_i) + \dfrac{(1-r_i^2)}{r_i^2}\text{var}(g_i)$. Ignoring

the off-diagonal elements of $\text{var}(\varepsilon)$ as before, the diagonals of the inverse of the residual variance after factoring out σ_e^2

are $\dfrac{\sigma_e^2}{[c+(1-r_i^2)/r_i^2]\sigma_g^2}$ which simplifies to give

$$w_i = \frac{1-h^2}{[c+(1-r_i^2)/r_i^2]h^2} \qquad (10)$$

an expression analogous to (5) with $n = 1$ and $h^2 = r_i^2$. Note that the weight in (10) approaches $\dfrac{1-h^2}{ch^2}$ as $r_i^2 \to 1$ in which case the weight tends to infinity as $c \to 0$. This is the same as would occur when the number of offspring $p \to \infty$, and p is used as a weight.

Removing parent average effects

Animal model evaluations by BLUP using the inverse relationship matrix shrink individual and progeny information towards parent average (PA) EBV [21]. It makes sense to remove the PA effect as part of the deregression process for two reasons. First, some animals may have EBV with no individual or progeny information. These animals cannot usefully contribute to genomic prediction. This is apparent if one imagines a number of halfsibs with individual marker genotypes and deregressed PA EBV. These animals cannot add any information beyond what would be available from the common parent's genotype and EBV. Second, if any parents are segregating a major effect, about half the offspring will inherit the favourable allele and the others will inherit the unfavourable allele. However, the EBV of both kinds of offspring will be shrunk towards the parent average. Parent average effects can be eliminated by directly storing the individual and offspring deregressed information and corresponding r^2 during the iterative solution of equations carried out for the purposes of genetic evaluation [2]. In some cases researchers do not have access to the evaluation system used to create the EBV on their training populations. In

those circumstances, it is necessary to approximate the evaluation equations and backsolve for deregressed information free of the effects of parent average. This can be done for one training animal at a time, given h^2 and knowledge of only the EBV (unadjusted for the base) and r^2 on the animal, its sire and its dam. First, compute parent average (PA) EBV and reliability for animal i with *sire* and *dam* as parents:

$\hat{g}_{PA} = \dfrac{\hat{g}_{sire} + \hat{g}_{dam}}{2}$, and $r_{PA}^2 = \dfrac{r_{sire}^2 + r_{dam}^2}{4}$. Assuming sire

and dam are unrelated and not inbred, the additive genetic covariance matrix for PA and offspring is

$\mathbf{G} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 1 \end{bmatrix}\sigma_g^2$ with inverse $\begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix}\sigma_g^{-2}$. Using

this result, recognise that the equations to be solved are:

$$\begin{bmatrix} Z'_{PA}Z_{PA} + 4\lambda & -2\lambda \\ -2\lambda & Z'_iZ_i + 2\lambda \end{bmatrix}\begin{bmatrix} \hat{g}_{PA} \\ \hat{g}_i \end{bmatrix} = \begin{bmatrix} \gamma^*_{PA} \\ \gamma^*_i \end{bmatrix}, \qquad (11)$$

where γ^*_i is information equivalent to a right-hand-side element pertaining to the individual, $Z'_{PA}Z_{PA}$ and Z'_iZ_i reflects the unknown information content of the parent average and individual (plus information from any of its offspring and/or subsequent generations), $\lambda = (1 - h^2)/h^2$ is assumed known. Define

$$\begin{bmatrix} Z'_{PA}Z_{PA} + 4\lambda & -2\lambda \\ -2\lambda & Z'_iZ_i + 2\lambda \end{bmatrix}^{-1} = \begin{bmatrix} c^{PA,PA} & c^{PA,i} \\ c^{i,PA} & c^{i,i} \end{bmatrix} = \mathbf{C}$$

then using the facts [19] that $r_i^2 = \dfrac{\text{var}(\hat{g}_i)}{\text{var}(g_i)}$ and $\text{var}(\hat{\mathbf{g}}) = \mathbf{G} - \mathbf{C}\sigma_e^2$ leads to $r_{PA}^2 = 0.5 - \lambda c^{PA,PA}$, and $r_i^2 = 1.0 - \lambda c^{i,i}$. Rearranging these equations, $c^{PA,PA} = (0.5 - r_{PA}^2)/\lambda$, and $c^{i,i} = (1.0 - r_i^2)/\lambda$. The formula to derive the inverse of a 2 × 2 matrix applied to the coefficient matrix from (11) gives $c^{PA,PA} = (Z'_iZ_i + 2\lambda)/det$, and $c^{i,i} = (Z'_{PA}Z_{PA} + 4\lambda)/det$ for $det = (Z'_{PA}Z_{PA} + 4\lambda)(Z'_iZ_i + 2\lambda) - 4\lambda^2$.

Equating these alternative expressions for $c^{PA,\,PA}$ leads to

$$(Z'_iZ_i + 2\lambda)/[(Z'_{PA}Z_{PA} + 4\lambda)(Z'_iZ_i + 2\lambda) - 4\lambda^2] = (0.5 - r_{PA}^2)/\lambda, \qquad (12)$$

and equating the expressions for $c^{i,\,i}$ leads to

$$(Z'_{PA}Z_{PA} + 4\lambda)/[(Z'_{PA}Z_{PA} + 4\lambda)(Z'_iZ_i + 2\lambda) - 4\lambda^2] = (1.0 - r_i^2)/\lambda. \qquad (13)$$

Second, solve these nonlinear equations for $Z'_{PA}Z_{PA}$ and Z'_iZ_i. Although not obvious, there is a direct solution for $Z'_{PA}Z_{PA}$ and Z'_iZ_i. It can be derived by dividing (12)

by (13), defining $\delta = (0.5 - r_{PA}^2)/(1.0 - r_i^2)$, and rearranging to get

$$Z_i'Z_i = \delta Z_{PA}'Z_{PA} + 2\lambda(2\delta - 1). \qquad (14)$$

Substituting the expression for $Z_i'Z_i$ in (14) into the denominator of (13), defining $\alpha = 1/(0.5 - r_{PA}^2)$, and rearranging leads to a quadratic expression in $Z_{PA}'Z_{PA}$, namely $0.5(Z_{PA}'Z_{PA})^2 + \lambda(4 - 0.5\alpha)(Z_{PA}'Z_{PA}) + 2\lambda^2(4 - \alpha - 1/\delta) = 0$, which has a positive root that can rearranged to

$$Z_{PA}'Z_{PA} = \lambda(0.5\alpha - 4) + 0.5\lambda\sqrt{(\alpha^2 + 16/\delta)}. \qquad (15)$$

Application of (15) provides the solution for $Z_{PA}'Z_{PA}$ that can be substituted in (14) to solve for $Z_i'Z_i$, together enabling reconstruction of the coefficient matrix of (11).

Third, the right-hand side of (11) can be formed by multiplying the now known coefficient matrix by the known vector of EBV for PA and individual. The right-hand side on the individual, free of PA effects is y_i^* The equation to obtain an estimate of EBV for animal i, free of its parent average, $\hat{g}_{i-PA}$, based only on y_i^*, is $[Z_i'Z_i + \lambda][\hat{g}_{i-PA}] = [y_i^*]$ and the corresponding r_i^{2*} for use in constructing the weights in (10) is given by $r_i^{2*} = 1.0 - \lambda/(Z_i'Z_i + \lambda)$. The deregressed information is $\frac{\hat{g}_{i-PA}}{r_i^{2*}}$, which simplifies to $\frac{y_i^*}{Z_i'Z_i}$ and is analogous to an average. An iterative procedure using mixed model equations to simultaneously deregress all the sires in a pedigree, while jointly estimating the base adjustment and accounting for group effects was given by Jairath et al [20]. However, that method requires knowledge on the numbers of offspring of each sire.

Double counting of information from descendants
Genetic evaluation of animal populations results in EBV that are a weighted function of the parent average EBV, any information on the individual, adjusted for fixed effects, and a weighted function of the EBV of offspring, adjusted for the merit of the mates [2]. The previous section has argued for the removal of parent average effects in constructing information for genomic analyses. It could be argued that information from genotyped descendants should also be removed to avoid double counting. This can be achieved during the evaluation process, and is desirable in the absence of selection. If the genotyped descendants are a selected subset, the removal of their information will lead to biased information on the individual. Simulation suggests that the double counting of descendants performance has negligible impact on genomic predictions (results not shown).

Results
Weights for different information sources
Comparative weights for individual and average of n individual observations using (5), and for progeny means of p halfsibs using (6) and deregressed EBV of varying reliability using (10) are in Table 1.

Removing parent average effects
Suppose genomic training is to be undertaken for a trait using EBV available from national evaluations that have yet to be deregressed. Widely-used bulls have been genotyped and the EBV and r^2 of those bulls are available, along with corresponding information on the sire and dam of each bull. Such a trio might have values of $\hat{g}_{sire}$ = 10, r_{sire}^2 = 0.97; $\hat{g}_{dam}$ = 2, r_{dam}^2 = 0.36; and $\hat{g}_i$ = 15, r_i^2 = 0.68. Given h^2 = 0.25, λ = 0.75/0.25 = 3, the PA information is $\hat{g}_{PA} = \frac{10+2}{2} = 6$, and $r_{PA}^2 = \frac{0.97+0.36}{4} = 0.333$. Using (15), with α = 5.97, δ = 0.523, then $\widehat{Z_{PA}'Z_{PA}}$ = 9.16 which substituted in (14) gives $\widehat{Z_i'Z_i}$ = 5.08.

Substituting these information contents into the coefficient matrix or left-hand side of (11) is
$\begin{bmatrix} 9.16+12 & -6 \\ -6 & 5.08+6 \end{bmatrix}$ with inverse $\begin{bmatrix} 0.0558 & 0.0302 \\ 0.0302 & 0.1066 \end{bmatrix}$.
These values correspond to $\hat{r}_{PA}^2$ = 0.5 - 3 × 0.0558 = 0.33

and $\hat{r}_i^2$ = 1.0 - 3 × 0.1066 = 0.68 the reported r_{PA}^2 and r_i^2 confirming the equations used to determine the information content. The right-hand side of (11) can then be reconstructed by multiplying the coefficient matrix by the vector of EBV as $\begin{bmatrix} 9.16+12 & -6 \\ -6 & 5.08+6 \end{bmatrix}\begin{bmatrix} 6 \\ 15 \end{bmatrix}$. The element of interest is the right-hand side element corresponding to the individual, obtained as y_i^* = -6 × 6 + 11.08 × 15 = 130. The deregressed information for use in subsequent analysis is obtained as $\frac{y_i^*}{Z_i'Z_i} = \frac{130}{5.08} = 25.6$ and the corresponding reliability of this information free of PA effects is r_i^{2*} = 1.0 - 3/(5.08 + 3) = 0.63. The relevant scaled weight for use with the deregressed information on this individual assuming c = 0.5 can be found using (10) as $w = \frac{0.75}{[0.5+(0.37/0.63)]0.25} = 2.76$. This implies that the deregressed information is 2.76 times more valuable than a single record on the individual.

Discussion
The relative value of alternative information sources varies according to c, the parameter that reflects the ability of the genotypic covariates to predict genetic

Table 1: Relative weights[a] for *n* phenotypic observations on the individual, *p* observations in twice the halfsib progeny mean with heritability 0.25 and repeatability 0.6, or deregressed EBV with reliability r^2 for varying values of *c*, the proportion of genetic variation for which genotypes cannot account

Information Source		c			
		0.8	**0.5**	**0.25**	**0.1**
Mean of *n* repeated records	*n*				
	1	0.79	0.86	0.92	0.97
	2	1.00	1.11	1.22	1.30
	5	1.19	1.35	1.52	1.65
	10	1.27	1.46	1.66	1.81
2 × mean of *p* half-sib offspring	*p*				
	5	0.79	0.86	0.92	0.97
	10	1.30	1.50	1.71	1.88
	20	1.94	2.40	3.00	3.53
Deregressed EBV with reliability r^2	r^2				
	0.1	0.31	0.32	0.32	0.33
	0.2	0.63	0.67	0.71	0.73
	0.3	0.96	1.06	1.16	1.23
	0.4	1.30	1.50	1.71	1.88
	0.5	1.67	2.00	2.40	2.73
	0.6	2.05	2.57	3.27	3.91
	0.7	2.44	3.23	4.42	5.68
	0.8	2.86	4.00	6.00	8.57
	0.9	3.29	4.91	8.31	14.21
	1.0	3.75	6.00	12.00	30.00

[a]Weights are diagonal elements of the inverse of the scaled residual variance-covariance matrix (with the scalar σ_e^2 factored out before inversion). Weights are relative to the information content of an individual observation with c = 0.

merit. Genomic prediction models that fit well have small values for *c* and result in greater relative emphasis of reliable information than is the case when the genomic prediction model fits poorly and the residual variation is dominated by contributions from lack-of-fit. For example, the mean of 20 halfsib progeny has about 3.6 times the value of the mean of 5 progeny when *c* is 0.1, and 2.5 times the value when *c* is 0.8. Deregressed EBV with reliability 1.0 are 11 times as valuable as reliability 0.5 when *c* is 0.1 but only 3 times as valuable when *c* is 0.5. These results indicate that collecting genotypes and phenotypes on training animals with low to moderate reliability will be of more relative value to genomic predictions that account for only 50% genetic variation (i.e. correlation 0.7 between genomic prediction and real merit) than they will for genomic predictions that account for a high proportion of variance.

The impact of the assumed *c* is to influence the relative value of individuals with reliable information, such as progeny test results, in comparison to individuals with information from less reliable sources, such as individual records. The use of too large a value of *c* will result in overemphasis of less accurate information in relation to more accurate information. The use of too small a value of *c* will result in too little emphasis on less accurate records. The correct value of *c* will not be known prior to training analyses but can be estimated from validation analyses. Training analyses could then be repeated using the estimated value of *c*. Alternatively, sensitivity to *c* could be assessed by training using a range of values. The sensitivity to *c* varies according to the heterogeneity of information content in the training data.

In practice, information sources of phenotypic data on training individuals can vary more widely than the examples derived in this paper. For example, training individuals might have their own and a mix of half-and fullsib progeny observed. In such cases, a practical approach is to first set up the mixed model equations that would be appropriate to estimate breeding values on the training individuals and use these to solve for the deregressed information [2]. This approach could also be useful in circumstances where training individuals do not all have the appropriate phenotypes. Consider a situation where some individuals have carcass measurements while others have correlated observations such as live animal ultrasound measures. A bivariate analysis of these two traits could be used to produce a single

deregressed value for the carcass trait for each animal that accounted for appropriately weighted ultrasound information.

Conclusions

The arguments put forward in this manuscript support the use of deregressed information, in agreement with practices adopted by many researchers [22]. The weighting factors proposed in this paper differ from any reported in the literature except when the parameter c = 0 in which cases the weights are effectively the same as those used by Georges et al. [5] and Spelman et al. [6]. In practice, the benefit of deregression and the subsequent weighting of alternative information sources will depend on the extent to which the number of repeat records, number of progeny and/or r^2 varies among individuals in the training population.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DJG derived the formulae following debate with JFT and RLF as to appropriate weights for training analyses with disparate data. JFT derived the direct solution for removing parent average effects. DJG drafted the manuscript and RLF and JFT helped to revise and finalize it. All authors read and approved the final manuscript.

Acknowledgements

References

1. Meuwissen THE, Hayes BJ and Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819–1829.
2. VanRaden PM and Wiggans GR: **Derivation, calculation, and use of national animal model information.** *J Dairy Sci* 1991, **74(8):** 2737–2746 http://www.hubmed.org/display.cgi?uids=1918547.
3. Morsci NMTJ and Schnabel RD: **Association analysis of adinopectin and somatostatin polymorphisms on BTA1 with growth and carcass traits in Angus Association analysis of adinopectin and somatostatin polymorphisms on BTA1 with growth and carcass traits in Angus cattle.** *Anim Genet* 2006, **37:**554–562.
4. Rodriguez-Zas SL, Southey BR, Heyen DW and Lewin HA: **Interval and composite interval mapping of somatic cell score, yield, and components of milk in dairy cattle.** *J Dairy Sci* 2002, **85 (11):**3081–3091.
5. Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR and Zhao X: **Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing.** *Genetics* 1995, **139(2):**907–920.
6. Spelman RJ, Coppieters W, Karim L, van Arendonk JA and Bovenhuis H: **Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population.** *Genetics* 1996, **144(4):**1799–1808.
7. Ashwell MS, Da Y, VanRaden PM, Rexroad CE and Miller RH: **Detection of putative loci affecting conformational type traits in an elite population of United States Holsteins using microsatellite markers.** *J Dairy Sci* 1998, **81(4):**1120–1125.
8. Van Tassell CP, Sonstegard TS and Ashwell MS: **Mapping quantitative trait loci affecting dairy conformation to chromosome 27 in two Holstein grandsire families.** *J Dairy Sci* 2004, **87(2):**450–457.
9. Loberg A and Durr JW: **Interbull survey on the use of genomic information.** *Proc Interbull Intl Workshop* 2009.
10. Meuwissen THE and Goddard ME: **Prediction of identity by descent probabilities from marker-haplotyes.** *Genet Sel Evol* 2001, **33:**605–634.
11. Nejati-Javaremi A, Smith C and Gibson JP: **Effect of total alleleic relationship on accuracy of evaluation and response to selection.** *J Anim Sci* 1997, **75:**1738–1745.
12. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91(11):**4414–4423.
13. Strandén I and Garrick DJ: **Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit.** *J Dairy Sci* 2009, **92(6):** 2971–2975 http://www.hubmed.org/display.cgi?uids=19448030.
14. Falconer DS and Mackay TFC: *Introduction to Quantitative Genetics* New York: Longman, Inc; fourth1996.
15. Gianola D, de los Campos G, Hill WG, Manfredi E and Fernando R: **Additive genetic variability and the Bayesian alphabet.** *Genetics* 2009, **183:**347–363.
16. Van Vleck LD: *Selection index and introduction to mixed model methodsBoca Raton: CRC 1993 chap. Genes identical by descent - the basis of genetic likeness; 49.*
17. Calus MPL, Meuwissen THE, de Roos APW and Veerkamp RF: **Accuracy of genomic selection using different methods to define haplotypes.** *Genetics* 2008, **178:**553–561.
18. Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, Rosa GJ and Gianola D: **Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers.** *J Dairy Sci* 2009, **92(10):**5248–5257.
19. Henderson CR: **Best linear unbiased estimation and prediction under a selection model.** *Biometrics* 1975, **31:**423–449.
20. Jairath L, Dekkers JC, Schaeffer LR, Liu Z, Burnside EB and Kolstad B: **Genetic evaluation for herd life in Canada.** *J Dairy Sci* 1998, **81(2):**550–562.
21. Mrode R: *BLUP univariate models with one random effect. In Linear Models for the Prediction of Animal Breeding Values* Cambridge: CABI; 2005.
22. Thomsen H, Reinsch N, Xu N, Looft C, Grupe S, Kuhn C, Brockmann GA, Schwerin M, Leyhe-Horn B, Hiendleder S, Erhardt G, Medjugorac I, Russ I, Forster M, Brenig B, Reinhardt F, Reents R, Blumel J, Averdunk G and Kalm E: **Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for QTL.** *J Anim Breed Genet* 2001, **118:**357–370.