

## Predicting Gain the Sustainable Way and its Relevance to Genomic Selection

J. A. Woolliams<sup>1</sup>, K. G. Nirea<sup>2</sup> and T. H. E. Meuwissen<sup>2</sup>

<sup>1</sup>The Roslin Institute and R(D)SVS, The University of Edinburgh, Midlothian, United Kingdom

<sup>2</sup>Dept. of Animal & Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway

**ABSTRACT:** Predicting gain for selection using optimum contributions is associated with two issues, the first concerned with inter-generational dependence of the contributions, and the second concerned with dynamic desirability. By ignoring the latter, which is valid when the accuracy of candidates approaches 1, a general formula for  $\Delta G(T, \Delta F, \alpha)$  can be obtained, where  $\Delta G(T, \Delta F, \alpha)$  is the maximum gain possible with T candidates per generation, degree of coancestry  $\alpha$ , and a rate of inbreeding  $\Delta F$ . This was shown by simulation to offer reasonable predictions, although further validation is required. The developed theory also makes testable predictions that the importance of mating designs will depend only on their impact on  $\alpha$  as accuracy approaches 1, and this prediction was also validated by simulation. The exception for mating designs affecting  $\alpha$  is because it affects both the variance of the Mendelian sampling term and the relationship between squared contributions and  $\Delta F$ .

**Keywords:** Rate of Gain, Rate of Inbreeding, Optimum Contributions, Genomic Selection, Mating Design

### Introduction

Predicting genetic gain ( $\Delta G$ ) is straightforward, the community has what is known as the 'breeders' equation  $\Delta G = \iota r \sigma_A / L$ , where  $\iota$  is the intensity,  $r$  is the accuracy,  $\sigma_A$  is the genetic standard deviation and  $L$  is the generation interval. These parameters are subject to a mix of biological, technical and operational constraints and choices. One of the constraints may be the total number of candidates  $T$ , but the value of  $\iota$  is then determined by the number of candidates selected; the fewer candidates selected, the stronger the intensity  $\iota$  and the faster  $\Delta G$ . However, whilst  $\Delta G$  increases linearly with  $\iota$ , its impact on the rate of inbreeding ( $\Delta F$ ) is quadratic with  $\iota$  so that relative small increases in  $\Delta G$ , can have much stronger increases in risks associated with the greater potential for inbreeding depression, the spread of deleterious recessives, the loss of genetic variance and the variability in the rate of genetic gain. Therefore, the use of  $T$  and the numbers selected in the 'breeders' equation disguises this dependence of  $\Delta G$  on  $\Delta F$ .

An alternative approach, which acknowledges the relationship between  $\Delta G$  and  $\Delta F$ , is to predict the rate of gain as a function of the candidates  $T$  and the  $\Delta F$ . The problem for prediction is then to determine the function  $\Delta G(T, \Delta F^*)$ , the maximum  $\Delta G$  that is feasible for an efficiently run breeding scheme with  $T$  candidates and operating with rate of inbreeding  $\Delta F^*$ . Here the asterisk is denoting a value of the constraint on  $\Delta F$ . This is the same

selection problem that is commonly known as optimum contributions selection, and to date attempts to solve this problem *a priori* have had little success.

In many respects, this problem can be characterized very well. Mathematically, *maximize*  $\Delta G$  *subject to*  $\Delta F \leq \Delta F^*$  can be reformulated using long-term genetic contributions (for random mating, but simply modified for non-random mating) as: *maximize*  $\sum r_i a_i$  *subject to*  $\sum r_i^2 \leq 4\Delta F^*$ , where  $r_i$  is the long term contribution of candidate  $i$ ,  $a_i$  is its Mendelian sampling term, and where the sums are over all  $T$  candidates per unit time. This is then repeated for each generation, which may involve several cohorts. In each generation  $\sum r_i = 1$  and with two sexes, the sums over male and female candidates each sum to  $1/2$ , and  $r_i \geq 0$ . The solution will always have  $\sum r_i^2 = 4\Delta F^*$  since, otherwise, it would be possible to increase the contribution of the individual with the most favorable Mendelian term (or its estimate) by some small value  $>0$ , so increasing the genetic gain. This would contradict the optimization of the algorithm in having found the maximum gain.

This problem has a known solution provided by Bondesson (1989), who showed using the calculus of variations that the optimum solution has a value  $a^*$  and a slope  $b$  such that the individual contributions are given by  $r_i = b(a_i - a^*)$  if  $a_i > a^*$  and  $r_i = 0$  otherwise, i.e. a linear allocation above a cut-off. The values of  $a^*$  and  $b$  are functions of the constraint  $\Delta F^*$  and the constraint that  $\sum r_i = 1$ . The difficulty with this solution is that it was provided for the case of a cash crop in forestry involving only a single generation, and does not solve the problem with recurrent selection.

Recurrent selection imposes two complications for implementing this solution. First, there is an inter-dependence between generations, in that allocation of a large long term contribution to an individual has consequences for previous generations, in that the sire and the dam will also share that contribution between them. Therefore optimal allocations cannot be decided generation by generation, but also involves the pedigree connections between generations. The second problem is that, for the most part, the EBV and, in particular, the estimates of the Mendelian sampling terms,  $\hat{a}$ , evolve over time. As a candidate, an individual's estimated  $\hat{a}$  may have poor accuracy but with selection and offspring its estimate becomes more accurate. Therefore, individuals seen as being of great merit may subsequently be viewed uninspiring and vice versa (assuming an initially poor candidate was selected). Therefore, recurrent selection on imperfect estimates brings with it the issue of dynamic desirability, adjusting contributions in line with new information on the individual.

Proposed optimum contributions algorithms (Wray and Goddard, 1994; Meuwissen, 1997; Grundy et al. 1998) seek the optimum compromise over generations, responding to the inter-generational, dynamic issues. The first two of these algorithms includes an additional down weighting of more recent generations on the assumption that their Mendelian sampling terms are being reduced because of inbreeding. However the optimizations are appropriate to the problem outlined above. What results from applying these algorithms in simulations are values of  $r_i$  with the underlying pattern described by Bondesson (1989) for the single-generation optimum of linear allocation above a cut-off, but with (i) values of  $r_i$  that are more strongly related (inevitably) to the estimate  $\hat{a}$  than  $a$ , and (ii) variance about this relationship. Howard et al. (2014) report on a broad validation of this outcome in practice, with all its associated operational constraints.

Grundy et al. (1998) predicted gain using the Bondesson (1989) solution, but the additional and intrinsic variance associated with recurrent selection means that the resulting predictions for  $\Delta G$  with optimum contributions are inevitably an overestimate. Furthermore, other approaches involving the derivation of expected long-term contributions (Woolliams et al., 1999) cannot be used. This is because such approaches assume that the evolution of  $r_i$  in future generations depends only on the selective advantage, i.e. conditional on the selective advantages, generations may be considered stochastically independent. Optimum contributions algorithms are different, in that  $r_i$  is actively managed across generations.

### Simplification and Frameworks

Dealing with both generational inter-dependence in contributions and dynamic desirability is challenging. Therefore, to obtain a meaningful answer to characterizing  $\Delta G(T, \Delta F^*)$ , the dynamic desirability will be ignored here. This is equivalent to assuming that the accuracy of evaluation is 1 in the candidate, an assumption that will be addressed later in the Discussion. This situation will occur when  $h^2 = 1$ , for example. Therefore, the remaining issue is the inter-dependence of contributions across generations which occurs because  $r_i = \frac{1}{2} \sum_j r_j$ , where the sum is over the offspring of  $i$ . In what follows below, the species is assumed diploid and monoecious, but with no selfing for simplicity.

Consider the following scenario where a scheme has been following optimum contributions selection on EBV with accuracy 1, i.e.  $\hat{a} = a$  for multiple generations, so that it has reached a (near) equilibrium in terms of population structure, and patterns of relationships. In that equilibrium,  $\mu(a)$  represents the mean value of  $r_i$  given  $a$  for any generation. Now suppose that chain of generations is broken at a point between one generation and the next. Then, instead of imposing the constraints of interdependence between the generations, each of these generations, the ‘parent’ generation and the ‘offspring’ generation, are re-optimized to minimize their squared deviations from  $\mu(a)$ . The squared contributions of the

parent generation are free of the variance imposed by accommodating the desired contributions of the offspring, but will still be constrained by accommodating the desired contribution of their parents; conversely, although the offspring generation is free of the variance imposed by accommodating the parents, they will still be constrained by accommodating their offspring. The fact that the accuracy is assumed to be 1 allows this re-distribution to have meaning, as the desired contributions are constant over time. Let  $\phi_i$  be the relaxed re-optimized values for contributions of the parent generation and  $\beta_i$  be the corresponding values for the offspring generation.

### Solutions and Predictions

Now consider what happens when the pedigree is joined back together as it was originally in the optimal contribution pedigree. Let  $y_i$  be the final solutions for contributions of the parent generation, i.e. when the pedigree is joined up, and  $z_i$  be the final solutions for the offspring generation. Joining up the generations will minimize the additional sum of squares arising, where  $S = \sum(y_i - \phi_i)^2 + \sum(z_i - \beta_i)^2$ , where the sums are over all individuals in each of the two generations and subject to the constraint imposed by the observation of parent and offspring relationships i.e.  $y_i = \frac{1}{2} \sum_j z_j$  where the sum is over the offspring of  $i$ . In vector form,  $S = (\mathbf{y} - \boldsymbol{\phi})^T (\mathbf{y} - \boldsymbol{\phi}) + (\mathbf{z} - \boldsymbol{\beta})^T (\mathbf{z} - \boldsymbol{\beta})$ . Let  $\mathbf{U}$  be the incidence matrix of parents and offspring, so  $d_{ij} = 1$  when  $j$  is an offspring of  $i$ , and 0 otherwise. In this form, the constraint is given by  $\mathbf{y} = \frac{1}{2} \mathbf{Uz}$ .  $\mathbf{U}$  summarizes the mating design and its properties will change if mating is hierarchical or factorial for example.

This constrained optimization is solved using Lagrange multipliers  $\boldsymbol{\lambda}$  and the solutions are  $\mathbf{y} = \boldsymbol{\phi} - \frac{1}{2} \boldsymbol{\lambda}$ ,  $\mathbf{z} = \boldsymbol{\beta} + \frac{1}{4} \mathbf{U}^T \boldsymbol{\lambda}$  and  $\boldsymbol{\lambda} = 2(\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T) (\boldsymbol{\phi} - \frac{1}{2} \mathbf{U} \boldsymbol{\beta})$ . Substituting for the Lagrange multiplier  $(\mathbf{y} - \boldsymbol{\phi}) = -(\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T) (\boldsymbol{\phi} - \frac{1}{2} \mathbf{U} \boldsymbol{\beta})$ ,  $(\mathbf{z} - \boldsymbol{\beta}) = \frac{1}{2} \mathbf{U}^T (\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T) (\boldsymbol{\phi} - \frac{1}{2} \mathbf{U} \boldsymbol{\beta})$ . Both are linear transformations of  $\mathbf{v} = \boldsymbol{\phi} - \frac{1}{2} \mathbf{U} \boldsymbol{\beta}$ , whose rows are defined by the parents.

The expected value of  $\mathbf{v}$  is  $E[\mathbf{v}] = E[\boldsymbol{\phi}] - \frac{1}{2} E[\mathbf{U} \boldsymbol{\beta}]$ . Let us examine this conditional on  $a_i$ , the Mendelian sampling term for each parent  $i$ . The first term is  $\mu(a_i)$ . In optimum contributions, the number of offspring is expected to be  $2T$   $\mu(a_i)$  and the expected desirable contribution of any of its offspring is  $1/T$  because the Mendelian sampling term of its offspring is independent of its parent and because of perfect accuracy, there is also no correlation in estimation errors. Finally,  $E[\mu(a_i)] = 1/T$  since all contributions must sum to 1 among the  $T$  candidates. Therefore, the expectation of  $v_i$  conditional on  $a_i$  is  $\mu(a_i) - \frac{1}{2} \cdot 2T \mu(a_i) \cdot 1/T = 0$ . Therefore, the  $(\mathbf{y} - \boldsymbol{\phi})$  and  $(\mathbf{z} - \boldsymbol{\beta})$  are deviations whose squares represent variances.

The variance of  $\mathbf{v}$  is more complicated. Both  $\boldsymbol{\phi}$  and  $\boldsymbol{\beta}$  have an unconditional expectation  $1/T$ , and both have a component of variance due to  $\mu(a)$ , say  $\sigma_\mu^2$ . The  $\boldsymbol{\phi}$  have individual variances about  $\mu(a)$  of  $\sigma_\phi^2$  but may have covariances within families from accommodating the parental contributions. Likewise,  $\boldsymbol{\beta}$  have individual

variances about  $\mu(a)$  of  $\sigma_\beta^2$  but these will be independent since they arise from ignoring the family structure. Finally, the elements of  $\boldsymbol{\phi}$  and  $\boldsymbol{\beta}$  are independent by construction. Therefore, covariances in  $\boldsymbol{\alpha}^T \mathbf{U} \boldsymbol{\beta} = 0$ . So  $\mathbf{v} \mathbf{v}^T$  is of the form  $\boldsymbol{\phi} \boldsymbol{\phi}^T + \frac{1}{4} \mathbf{U} \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{U}^T = \sigma_\mu^2 (\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T) + \sigma_\phi^2 (\mathbf{I} + \boldsymbol{\delta}) + \sigma_\beta^2 \frac{1}{4} \mathbf{U} \mathbf{U}^T$  (for some small  $\boldsymbol{\delta}$ ). In any breeding scheme generating gain efficiently,  $\sigma_\mu^2$  will contribute a large part of the total variance, and so, to the first order, the variance covariance of  $\mathbf{v}$  is  $\mathbf{v} \mathbf{v}^T \sim \sigma^2 (\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T)$  for some  $\sigma^2$ .

What then is the increase in the sum of squared contributions arising from this process of joining up generations.  $S = (\mathbf{y} - \boldsymbol{\phi})^T (\mathbf{y} - \boldsymbol{\phi}) + (\mathbf{z} - \boldsymbol{\beta})^T (\mathbf{z} - \boldsymbol{\beta}) = \mathbf{v}^T (\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T)^{-1} \mathbf{v}$  from above. Since  $\mathbf{v}$  has a variance covariance structure of the form  $\sigma_\mu^2 (\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T)$ , then  $E[\mathbf{v}^T (\mathbf{I} + \frac{1}{4} \mathbf{U} \mathbf{U}^T)^{-1} \mathbf{v}] \approx T \sigma_\mu^2$ . What is interesting is that this joining up of sum of squares is (near) independent of  $\mathbf{U}$  the matrix that embodies the design of the matings. Since the total sum of squares is  $4\Delta F^*$ , independent of the form of  $\mathbf{U}$ , and  $S$  is independent of the form of  $\mathbf{U}$ , then it implies that  $\sigma_\mu^2$  is independent of the form of  $\mathbf{U}$ , and consequently the expected  $\Delta G$  is independent of the form of  $\mathbf{U}$ . Thus, the model presented is making a clear testable prediction that when the accuracy approaches 1 benefit in  $\Delta G$  from combining a mating design with optimum contributions reduces to zero. One caveat to this prediction is that some designs, such as minimum coancestry matings, alter the variance of the Mendelian sampling term through the value of  $\alpha$ , which measures the departure from Hardy-Weinberg equilibrium caused by the mating design, i.e. heterozygosity observed is  $(1 - \alpha)$  times the heterozygosity expected given the allele frequencies in the gene pool. This is not included in the analysis above, so the prediction from the analysis is that  $\Delta G$  may be a function of  $\alpha$  but nothing else. This is a strong prediction, as one of the most potent mating designs is factorial mating, where  $\alpha$  will be close to 0 unless further actions are taken to change the coancestry of mating pairs. Factorial mating has been shown consistently to increase  $\Delta G$  for a given  $\Delta F$ , with or without the use of optimum contributions.

### Testing Predictions on Matings

Simulations were constructed to test the above predictions for mating designs when accuracy was 1, by constructing traits with heritability 1. This was done in two ways: (i) by assuming an infinitesimal model from a base generation; or (ii) by generating a finite locus base generation, where the loci were obtained by random mutations over many generations with  $N_e = 1000$ , sufficient to have attained a mutation – drift equilibrium. In the latter case, 1000 SNP loci were selected to be QTL from among those simulated. An individual's breeding value,  $A$ , was the sum of effects of the alleles it carried i.e. an additive model. Since the accuracy was to be 1, no decisions on masking QTL or not, or method of genomic evaluation was necessary.

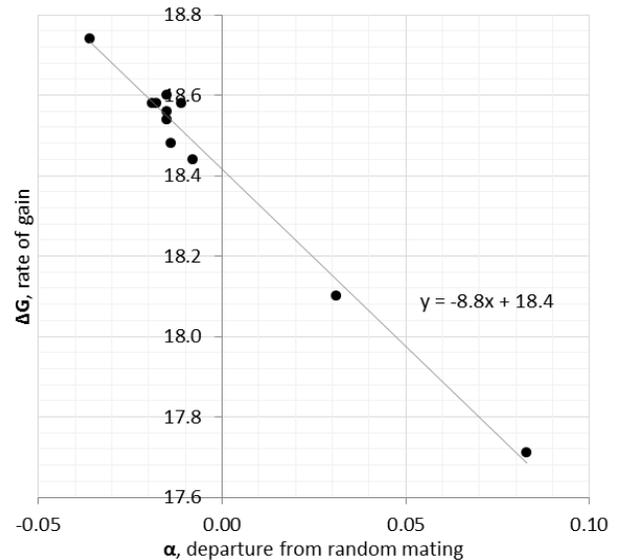
Four broad groups of mating designs were used to assign the mating pairs associated with family structure, coancestry, contributions, and breeding value of the parents. This included minimum and maximum coancestry,

minimizing and maximizing the production of full-sibs (the former is factorial mating), maximizing positive and negative correlations of parental contributions among the mates (family assortment), maximizing the positive and negative assortment of breeding values, maximizing the positive and negative assortment of Mendelian sampling terms, and positive and negative assortments based on desired contributions and deviations from desired contributions. For each of these mating designs, the following diagnostics were calculated: (i)  $\alpha$ , irrespective of whether or not  $\alpha$  was used in the optimization (often not reported in the literature); (ii) the fraction of full-sibs produced by the design.

The schemes were continued for 12 generations using optimum contributions selection for different sizes of  $T$  and  $\Delta F^*$ , and all diagnostics were calculated each generation to check for equilibrium.  $\Delta F$  was calculated using the numerator relationship matrix. For each scenario described below, 100 base populations were produced and each was replicated 4-fold.  $\Delta G$ , calculated over 5 generations, for each replicate, and the mean and standard error for each design was calculated using both variances between- and within-base populations.

The summary of the outcomes from the study are summarized very simply in Figure 1. The only discernible effect was that associated with  $\alpha$ . The clarity of the results was such that the residual variance after the regression in Figure 1 was not different from the sampling variances of the points.

**Figure 1. The simulated  $\Delta G$  for 11 different mating schemes plotted as a function of the observed departure from Hardy-Weinberg equilibrium,  $\alpha$ , which is a measure of departure from true random mating.**



The software used for these simulations has previously demonstrated expected effects of mating designs when accuracies  $< 1$ . Further confidence in the results can be obtained from noting that the impact of  $\alpha$  shown in Figure 1 can be predicted. The starting point for this is to note that  $\Delta G(T, \Delta F^*, \alpha) = (1 - \alpha)^{1/2} \Delta G(T, (1 - \alpha)^{-1} \Delta F^*, 0)$  in this case of perfect accuracy. This comes from algebra-free consideration of the observation that in its full form  $\Delta F = 1/4 (1 - \alpha) \Sigma r_i^2$ , and that  $\text{var}(a_i)$  is scaled by  $(1 - \alpha)$  from its value under random mating. Using this starting point algebraic derivation shows  $\Delta G(T, \Delta F^*, \alpha) \sim (1 - 1/2\alpha) \Delta G(T, \Delta F^*, 0)$ : the observed values was 0.48. (The primary effect of  $\alpha$  in the simulated data is through the trait's Mendelian sampling variance; but in general this regression will depend on the parameters).

In summary, the simulations validated an unexpected prediction from the preceding analysis: with perfect accuracy, mate allocation has no effect on gain when  $\Delta F$  is constrained using optimum contributions, apart from its impact on the degree of departure from random mating with respect to ancestry. It follows that the value of the mating designs such as factorial mating arises wholly from reducing the impact of dynamic desirability on the breeding scheme (Sorenson et al. 2005).

### Towards Predicting Genetic Gain

With this validation it is now possible to predict  $\Delta G(T, \Delta F^*, \alpha)$ . However random mating will continue to be assumed below. The importance of the validation is that is now plausible to ignore complications arising from  $U$ . This allows for some simplification and the results below are derived from consideration of one parent and its offspring in isolation, with only a single scalar Lagrange multiplier. The detailed algebra for this has not been described here. However the following important conclusion can be drawn from this simplification: the total variance in contributions introduced into parental contributions from accommodating the desired contributions of their offspring (and their descendants), is twice as large as the total variance in contributions introduced into offspring contributions from accommodating the desired contributions of both parents (and their ancestors).

Returning to the framework for the problem, the  $\phi$  of the parental generation have already accommodated their ancestors, and the  $\beta$  of the offspring generation have already accommodated their descendants. Therefore, it is reasonable to assume  $\sigma_\phi^2 = 1/2\sigma_\beta^2$ . It is now possible to recast  $\Sigma r_i^2 = 4\Delta F^* = T(T^{-2} + \sigma_\mu^2 + 3/2 \sigma_\phi^2)$ , the 3 terms representing the squared unconditional mean contribution, the variance of  $\mu(a)$  about this unconditional mean ( $\mu(a)$  is the expected contribution conditional on  $a$ ), and the third term results from the inter-dependence of generations. The latter term arises 1/3 from accommodating generations up to and including the parents, 2/3 from accommodating the

offspring generations and beyond. Estimating  $\sigma_\phi^2$  is difficult but feasible, and it can be characterized in terms of  $\sigma_\mu^2$ ,  $T$  and  $\Delta F^*$ . This allows  $\sigma_\mu^2$  to be estimated and, hence  $\mu(a)$  by assuming a functional form. Finally,  $\Delta G(T, \Delta F^*, 0) = E[\mu(a) a]$  (Woolliams et al., 1999). The solution is given below, but with no further derivation.

### Solutions and Fit

Assume  $\mu(a)$  takes the form  $\mu(a_i) = b(a_i - a^*)$  if  $a_i > a^*$ , and  $\mu(a_i) = 0$  otherwise; although this has two parameters,  $b$  and  $a^*$ , there is a constraint on them since all contributions sum to 1, so  $\int \mu(a) \pi(a) da = 1$  where  $\pi(a)$  is the density of the Mendelian sampling term, which will be assumed Normal. Therefore, all solutions are indexed by the single parameter  $a^*$ . From the derivation above, two further functions of  $\mu(a)$  are required:  $\sigma_\mu^2 = \int \mu(a)^2 \pi(a) da - T^{-2}$  and  $\theta = \int 1/2 T \mu(a) (1 + 1/2 T \mu(a))^{-2} \pi(a) da$ . The latter arises from the form of  $\sigma_\phi^2$  and, heuristically, is integrating the variance of error variance from matching parents to offspring. With larger families, either through  $\mu(a)$  or  $T$ , there is greater certainty in achieving desired contributions for a parent and all its offspring, and  $\theta$  approaches 0. With these definitions,  $\Delta F = 1/4 T [ T^{-2} + \sigma_\mu^2 (1 + \theta/8)(1 - \theta/4)^{-1}]$ . If  $\theta$  approaches 0, the inter-dependence of generations is small, but as  $\theta$  increases (say with small  $T$ ), more of the variance in contributions is required to match parents to offspring and less is used to deliver  $\Delta G$ .

For any constraint  $\Delta F^*$ , with random mating, the value of  $a^*$  can therefore be determined, and hence  $\mu(a)$ . Finally, the achievable gain is given by:

$$\Delta G(T, \Delta F^*, 0) = \int \mu(a) a \pi(a) da$$

and more generally

$$\Delta G(T, \Delta F^*, \alpha) = (1 - \alpha)^{1/2} \Delta G(T, (1 - \alpha)^{-1} \Delta F^*, 0)$$

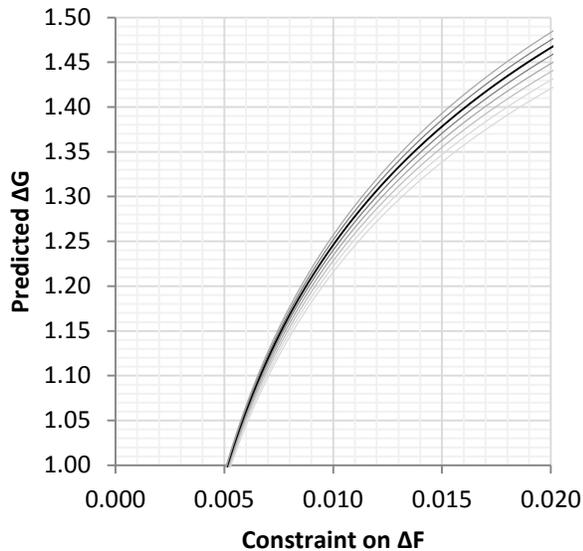
Figure 2 shows the form of the maximum gain feasible as a function of  $\Delta F^*$  for  $T=200$  and values of  $\alpha$  ranging from 0.04 to 0.10. For the range of parameters shown here, the gain shows diminishing returns with  $\Delta F^*$ , although it is not approaching its asymptote. The impact of  $\alpha$  appears to increase as  $\Delta F^*$  increases. Figure 3 gives an indication of fit to simulated data. It suggests that the fit is good but with some indication of underestimation for the highest simulated gain. More validation is required.

### Discussion

**Relevance to genomic selection.** Genomic evaluation and selection is a synergistic technology to optimum contributions selection. The use of genomic data will increase the accuracy of the estimated Mendelian sampling term, which is the principle selective advantage of an individual in a breeding scheme operating optimum contributions selection. In turn, optimum contributions algorithms maximize the gain for the loss of variance defined by the relationship matrix. Therefore, in terms of

utilizing genetic variance, genomic selection and optimum contributions selection together offer a step change in utilizing genetic variance to deliver genetic gain.

**Figure 2.** The predicted outcomes in  $\Delta G$  when  $\Delta F$  is constrained for  $T = 200$  for values of  $\alpha$  increasing from -0.04 to 0.10 in increments of 0.02. The value for  $\alpha = 0$  is the 3<sup>rd</sup> line from the top and shown in black;  $\alpha < 0$  are above this line,  $\alpha > 0$  are below this line. The intensity of the grey line diminishes with the magnitude of the deviation from 0.

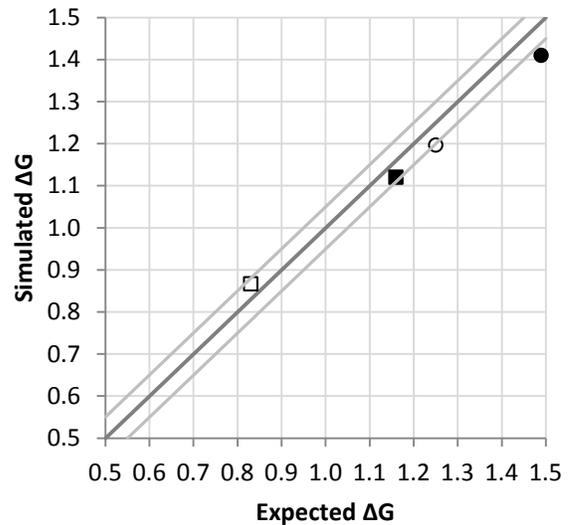


However, the relevance of genomic selection to this paper is more specific in that with genomic selection we can consider a future where dynamic desirability can be ignored. With pedigree and phenotype, where selection is on own performance and ancestors, traits with moderate heritability will always be subject to dynamic desirability. However, the advent of genomic evaluation, coupled with the prospect of big data, involving huge numbers of genotyped and phenotyped animals offers the prospect that accuracies may approach 1, as modelled here. If accuracy does not approach 1 with huge numbers, then the community needs to completely overhaul the basis of its most cherished models for genetic evaluation.

The accuracy of 1 used here is a theoretical statement, but what is necessary is that the desired long-term contribution of an individual is known at the time of selection and does not change in the process of convergence, which may be 5 to 10 generations. In the current technology of SNP chips, it may be that the chip does not capture all the variance for a trait. Following Dekkers (2007), it is plausible to decompose the trait into a component that is marker accessible and another which is independent of the markers. In this case, the theory developed will apply to the marker

component, even though the accuracy for the trait of interest is not 1.

**Figure 3.** A comparison of simulated with predicted  $\Delta G$  for two values of  $\Delta F^*$  and two values of  $T$  with  $\alpha = 0$ . Open and closed symbols denote  $\Delta F^* = 0.005$  and 0.01 respectively; squares and circles denote  $T = 200$  and 500 respectively; the central line denotes  $y = x$ , and flanking lines denote 2 standard errors of simulation values. Simulated values are means of 100 replicates.



The structure of the optimum contribution selection and managing variation at a genomic level is generalizable to genomic data. The long-term contribution of an individual is the contribution of its Mendelian sampling term, which can be a set of (correlated) terms across multiple loci. It is not a statement of fractions of its alleles that ultimately become fixed, as the long term fate of an individual's alleles will depend on the Mendelian sampling and the selection fate of its descendants. The decomposition of allele frequency changes into contributions and Mendelian terms is a unique decomposition into independent terms. The decomposition of the loss of variance as being related to the sum of squared contributions arises naturally from the decomposition of the changes in allele frequencies. It then emerges as a measure of the loss of variance promoted by an individual or its cohort.

**Methodological.** What has been presented in this paper is an analysis based upon a virtual experiment of breaking and then re-linking a converged chain of generations. The results appear to make sensible predictions but they are not particularly neat. However, the solution is neater than obtained in previous attempts (Avendano et al., 2004), but this study makes the important simplification of ignoring

dynamic desirability. The need for parameters that are known by integration or from pre-calculated tables is not unlike the case with the breeders equation (for intensity and accuracy). There may well be a simpler approach to obtaining a smarter solution. At this point, more validations need to be done and the importance of particular terms needs to be established, which may result in simpler approximations.

The magnitude of the inter-generational dependence is important. With very high  $\Delta F$ , there are a few large families, the value of  $\theta$  is small, and there is very little additional variance about desired contributions from accommodating generations. In this case, the formula of Grundy et al. (1998) is a closer approximation to  $\Delta G(T, \Delta F^*, 0)$ . For lower values of  $\Delta F$ , where the number of families is large but, as  $T$  is constant, the size of families is small, the inter-generational dependence can be as large as 50% of the variance in the desired contributions, or larger, making the Grundy formula a serious overestimate. Whilst this superficially suggests that genetic gain with optimum contributions selection is more efficient with higher  $\Delta F$ , it ignores the well-established interpretation of  $\Delta F$  as a measure of risk in a breeding scheme – the managing the impact of what is unknown. For the foreseeable future what we do not know about precision breeding will remain very considerable.

The reduced importance of mating designs with accuracies approaching 1 is a result that is strictly concerned with previously established results that show mating can influence the relationship between  $\Delta F$  and  $\Delta G$  (e.g. Sonesson and Meuwissen, 2000). The previous studies considered imperfect accuracies. However, avoidance of mating close relatives remains of value in avoiding extreme inbreeding coefficients in the offspring and the expectation that the impact of inbreeding depression will be greater in such offspring compared to their contemporaries.

### Conclusions

The prediction of gain for optimum contribution selection over recurrent generations is made difficult by two issues: the first is concerned with inter-generational dependence of the long-term contributions, since an individual's contribution will also be shared between its parents; and the second is concerned with dynamic desirability, whereby additional information on an individual's breeding value and Mendelian sampling term will change what is considered to be the ideal contribution. By ignoring the latter, which happens when the accuracy of candidates approaches 1, a general formula for the maximum gain possible with  $T$  candidates per generation, degree of coancestry  $\alpha$ , and a rate of inbreeding  $\Delta F$  can be obtained, denoted  $\Delta G(T, \Delta F, \alpha)$ . The solution involved using Lagrange multipliers to enforce the constraint that an individual's contribution is  $\frac{1}{2}$  of the sum of the contributions of its offspring, and so account for inter-generational dependence. The theory predicted that, as the accuracy approaches 1, the importance of mating designs

will depend only on their impact upon  $\alpha$ , and other benefits, e.g. in managing uncertainty, will diminish, and this was also validated by simulation. Therefore minimum coancestry mating will retain its benefits but factorial mating will not. The prediction of  $\Delta G(T, \Delta F, \alpha)$  was shown to be close to simulated values, although further validation will be required.

### Acknowledgements

JAW gratefully acknowledges funding for this work from two sources: the EC Framework 7 'Quantomics' project, and the BBSRC Institute Strategic Programme Grant BB/J004235/1.

### Literature Cited

- Avendaño, S., Woolliams, J. A., and Villanueva, B. (2004). *Genet. Res.* 83: 55-64.
- Bondesson, L. (1989). *Swed. Univ. of Ag. Sci. Report* 27.
- Dekkers, J.C.M. (2007) *J. Anim. Breed. Genet.* 124: 331-341.
- Grundy, B., Villanueva, B., and Woolliams, J. A. (1998). *Genet. Res.* 72: 159-168.
- Howard, D.M., Pong-Wong, R., Knap, P.W., Kremer, V.D. and Woolliams, J.A. (2014). *Proceedings of the 10<sup>th</sup> World Congress on Genetics Applied to Livestock Production.*
- Meuwissen, T. H. E. (1997). *J. Anim. Sci.* 75: 934-940.
- Sonesson, A. K., & Meuwissen, T. H. (2000). *Genet. Sel. Evol.* 32: 231-248.
- Sorensen, A., Berg, P., and Woolliams, J. (2005). *Genet. Sel. Evol.* 37: 57-81.
- Villanueva, B., Avendaño, S., and Woolliams, J. A. (2006). *Genet. Sel. Evol.*, 38: 127-146.
- Woolliams, J. A., Bijma, P., and Villanueva, B. (1999). *Genetics* 153: 1009-1020.
- Woolliams, J. A., and Thompson, R. (1994). *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production* 19:127-134.
- Wray, N. R., and Goddard, M. E. (1994) *Genet. Sel. Evol.* 26: 431-451.