Genetic architecture and accuracy of genomic prediction

B. Hayes[*], J. Pryce[*], P. Bowman[*], A.C. Chamberlain[*] and M.E. Goddard[*]

Introduction

Genomic prediction of future phenotypes or genetic merit using dense SNP genotypes can be used for prediction of disease risk, for forensics, and for estimation of breeding values for use in selection of livestock, crops and forage species (Wray et al. 2007, Lee et al. 2008, VanRaden et al. 2009, Goddard and Hayes 2009). In dairy cattle, estimated breeding values predicted from genomic information are now in wide spread use (Van Raden et al. 2008, Hayes et al. 2009).

The accuracy of genomic predictions will depend on the number of phenotypes used to derive the prediction equation, the heritability of the trait, the effective population size, the size of the genome, the density of markers, and the genetic architecture of the trait, in particular number of loci affecting the trait and distribution of their effects (Daetwyler et al. 2008, Goddard 2008). In simulated data the distribution of loci effects affects the accuracy of predicting genetic values. However in real data it has been difficult to show that traits vary in this distribution. For instance, in many cases a statistical method (Best linear unbiased prediction or BLUP) designed for traits with many loci all of small effects performs as well as other methods (eg. Moser et al. 2009, Verbyla 2009). If it is true that most complex traits are controlled by very many polymorphisms of very small effect (a nearly infinitesimal model), this has important consequences for prediction of genetic merit or future phenotypes such as disease risk. Formulae for the accuracy of genomic prediction under this model (Goddard 2008) suggest that sample sizes >100,000 individuals will be needed to achieve high accuracy, except for populations with a small effective population size. Thus it is important to determine the distribution of effect sizes for a range of traits, use this information in genomic prediction and plan future experiments accordingly.

Proportion of black of the coat in Holstein Friesian cattle is one such "quantitative" trait, as it can be recorded as the proportion of the coat which is white. In this paper, we use proportion of black in the coat and two other complex traits to show that differences in the distribution of loci effects are recognisable using a new method to estimate the distribution of variance explained by each QTL. We then contrast the accuracy of genomic prediction which can be achieved for proportion of black on coat the accuracy of genomic predictions for two traits with a different distribution of effects, namely overall type, a complex trait combining scores

[*]Biosciences Research Division, Department of Primary Industries Victoria, Melbourne, Victoria , Australia.
Faculty of Land and Food Resources, University of Melbourne, Melbourne, Victoria, Australia.

for a number of aspects of cow confirmation, and fat% in milk. The results demonstrate a clear effect of trait architecture on the accuracy of genomic predictions.

Material and methods

Samples and SNPs. The data set consisted of 1200 Australian Holstein bulls. For fat% and overall type the 'phenotype' used for each bull was the mean phenotype of his daughters. To obtain this phenotype we de-regressed the Australian breeding values (ABVs) to remove the contribution from relatives other than daughters (eg VanRaden et al. 2009) while retaining the correction for non-genetic effects such as herd. All bulls with de-regressed estimated breeding values had at least 80 daughters. The traits measured in the bull's daughters were fat% in a sample of the milk on each test day, and overall type. Overall type is composite trait combining scores for a number of aspects of the cow's conformation, including frame-capacity, rump, feet and legs, fore udder, rear udder, mammary system and dairy character. For portion of black, each bull himself was scored according to the proportion of black on the entire body, from 0% to 100% black. The values ranged from 5% black to 95% black. The bulls were genotyped for the Illumina Bovine50K array, which includes 54,001 Single Nucleotide Polymorphism (SNP) markers (Matukumalli et al. 2009). The following criteria and checks were applied to the bull's genotypes. Mendelian consistency checks revealed a small number of either sons who were discordant with their sires at many (>1000) SNPs or sires with many discordant sons. These animals (17) were removed from the data set. We omitted bulls if they had more than 20% of missing genotypes. 1181 bulls passed these criteria. Criteria for selecting SNPs were; less than 5% pedigree discordants (eg. cases where a sire was homozygous for one allele and progeny were homozygous for the other allele), 90% call rate, MAF>2%, Hardy Weinberg P<0.00001. 40077 SNPs met all of these criteria. A small number of these could not be mapped and were omitted from the final data set, as were SNPs on the X chromosome. Parentage checking was then performed again, and any genotypes incompatible with pedigree were set to missing. To impute missing genotypes, the SNPs were ordered by chromosome position. All SNPs which could not be mapped or were on the X chromosome were excluded from the final data set, leaving 39,048 SNPs. To impute missing genotypes, the genotype calls and missing genotype information was submitted to fastPHASE (Scheet and Stephens 2006) chromosome by chromosome. The genotypes were taken as those filled in by fastPHASE.

The discovery dataset consisted of bulls progeny tested before 2004 (n=756). For proportion of black portion 327 bulls in the reference set had phenotypes. The bulls in the validation dataset were progeny tested during or after 2004 (n=400).

Estimating the distribution of proportion of variance explained by chromosome segments

A chromosome segment was defined as consisting of 50 adjacent SNP loci. This size segment was chosen as a compromise between having too little SNP information to accurately estimate a genomic relationship matrix, and having sufficiently small segments to enable interpretation regarding the distribution of effects on the trait. For each 50SNP segment of chromosome, we estimated the proportion of variance explained by building a genomic relationship matrix (as described above) based on the 50SNPs only (**G_1**), and a

second genomic relationship matrix ($\mathbf{G_2}$) using all SNPs except those in the current 50 SNP segment to account for popultation structure and the effects of all the other QTL. We the fitted the model $\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Zg}_1 + \mathbf{Zg_2} + \mathbf{e}$, where **y** is a vector of phenotypes, μ is the mean, $\mathbf{1_n}$ is a vector of 1s**,** **Z** is a design matrix allocating records to animals, $\mathbf{g_1}$ is a vector of genetic effects for a 50 SNP segment, assumed to be normally distributed with mean 0 and co(variance) $\mathbf{G_1}\sigma^2_{g1}$, $\mathbf{g_2}$ is a vector of breeding values based on all the other segments, assumed to be normally distributed with mean 0 and co(variance) $\mathbf{G_2}\sigma^2_{g2}$ and **e** is a vector of random normal deviates ~ $N(0, \mathbf{I}\sigma^2_e)$. Variance components were estimated with ASREML (Gilmour et al. 2002), and the proportion of variance explained by each segment was calculated as $\sigma^2_{g1} /(\sigma^2_{g1} + \sigma^2_{g2} + \sigma^2_e)$.

The estimate of the proportion of variance explained by a chromosome segment i (y_i^2) is naturally subject to some sampling error. y_i^2 is analogous to the squared correlation between the effect of the segment and the phenotype so y_i is analogous to the correlation. We modelled y_i as $y_i = t_i + e_i$ where t_i is the true correlation between segment i and phenotype and e_i is a sampling error While it is not possible to estimate the sampling error for a specific segment, we can estimate the distribution of sampling errors. To do this the phenotypes were permuted across the genotypes 1000 times and the proportion of variance explained by each segment re-calculated. Under the null hypothesis that there is no real correlation between segments and phenotypes, the distribution of the estimated proportion variance explained should be a mixture of zero and a chi-square with 1 degree of freedom. (half the time the correlation would be estimated to be negative but ML always reports an estimate within the parameter space and so half he reported estimates of variance are zero). Therefore the square roots of these estimates were assumed to be near-zero(half the time) and the positive half of a normal distribution the other half. The standard deviation of e , σ, was then taken as the square root of the average proportion of variance explained multiplied by 2 (the multiplication by two was to account for the fact that negative estimates of the proportion of variances explained are reported as zero).

We then used maximum likelihood to estimate the distribution of true chromosome segment variances (t_i^2) given that we had a sample of estimated chromosome segment variances (y_i^2) and $y_i = t_i + e_i$ with $e_i \sim N(0,\sigma)$.

We estimate the distribution of y and then convert that to a distribution of y^2. We did not wish to assume any parametric form for the distribution of y so we approximate it by a discrete distribution in which the proportion explained can only take values j=0.00, 0.005 and so on to 1 (eg 100 classes between 0 and 1, but including 0) . We then estimate the frequency of these discrete values. The probability of observing y_i given j and σ was taken

as $\phi(y_i, j, \sigma)$ if $y_i > 0$ and $\varphi(0, j, \sigma)$ if $y_i = 0$ where $\phi(p_i, x_j, \sigma)$ is the density function of the normal distribution and $\varphi(0, j, \sigma)$ is the cumulative function of the normal distribution. (If t+e is negative for a segment then y^2 would be reported as zero since negative variances are not allowed).

Then an expectation maximisation (EM) algorithm was used to estimate the proportion of chromosome segments in each class f_j. The EM algorithm had three steps

1. Initialise each f_j. to 0.01.
2. Calculate the probability of the j given the y_i was $P(j \mid y_i) = \frac{P(j \mid y_i) f_j}{\sum_{j=1}^{100} P(y_i \mid j) f_j}$
3. Update the proportion of chromosome segments in each class as

$$f_j = \frac{\sum P(j \mid y_j)}{n}$$

Steps 2 and 3 were repeated until the f_j values did not change between iterations. The results are presented as a distribution of t^2 where the frequencies all values of t between $\sqrt{0.01}$ and $\sqrt{0.03}$ are summed and presented as the frequency of $0.01<t^2<0.03$ etc.

Genomic prediction. If there are many QTLs whose effects are normally distributed with constant variance, then genomic selection is equivalent to replacing the expected relationship matrix with the realised or genomic relationship matrix (**G**) estimated from DNA markers in the BLUP equations. (e.g. Nejati-Javaremi et al. 1997; Goddard 2008). The GBLUP model was $\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Zg} + \mathbf{e}$ where **y** is a vector of phenotypes, μ is the mean, $\mathbf{1_n}$ is a vector of 1s**,** **Z** is a design matrix allocating records to estimated breeding values, **g** is a vector of breeding values and **e** is a vector of random normal deviates ~ $N(0, \sigma_e^2)$. The breeding value **g** can be modelled by the combined effects of all the SNPs **g**=**Wu** where u_j is the effect of the j^{th} SNP, and $V(\mathbf{g}) = \mathbf{WW'}\sigma_u^2$. Elements of matrix **W** are w_{ij} for the i^{th} animal and j^{th} SNP, where $w_{ij} = 0 - 2p_j$ if the animal is homozygous 11 at the j^{th} SNP, $1-2p_j$ if the animal is heterozygous and $2 - 2p_j$ if the animal is homozygous 22 at the j^{th} SNP. The diagonal elements of **WW'** will be $\sum_{j=1}^{m} 2p_j(1-p_j)$ where m is the number of SNPs. If **WW'** is scaled such that $\mathbf{G} = \frac{n\mathbf{WW'}}{\sum_{i=1}^{n} w_{ii}}$ then $V(\mathbf{g}) = \mathbf{G}\sigma_g^2$. Estimated breeding values for both phenotyped and non-phenotyped individuals can be predicted by:

$\left[\hat{\mathbf{g}}\right] = \left[\mathbf{Z'Z} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2}\right]^{-1}\left[\mathbf{Z'y}\right]$ Where **G** is the realised relationship matrix calculated as above, and σ_g^2 is a genetic variance. Variance components were estimated with ASREML (Gilmour et al. 2002).

The realised accuracy of GEBV was calculated as r(**GEBV,y**$_{val}$)/h where y_{val} was the phenotype (either deregressed estimated breeding values for overall type and fat%, or the bull's own proportion of black), for bulls in the validation set, and h is the correlation between the phenotype and the true breeding value estimated, (the square root of the heritability of the records was used).

BayesA. A Bayesian approach to simultaneously predicting the effect of all SNPs to derive the prediction equation was used, namely BayesA described by Meuwissen et al. (2001). BayesA has a prior assumption that SNP effects are t-distributed. The model fitted was:

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{Xu} + \mathbf{Zv} + \mathbf{e}$$

Where **y** is a vector of n phenotypes, **X** is (n x m) a design matrix allocating records to the marker effects with element X_{ij} =0, 1 or 2 if the genotype of animal i at SNP j is 11, 12 or 22 respectively. **u** is a (m x 1) vector of SNP effects assumed normally distributed $u_i \sim N(0, \sigma_{ui}^2)$, **e** is a vector of random deviates where σ_e^2 is the error variance, v_i is the polygenic breeding value of the i^{th} animal, with variance $\mathbf{A}\sigma_a^2$, where **A** is the numerator relationship matrix derived from pedigrees. In BayesA the prior for σ_{ui}^2 was an inverse chi square distribution with 4.012 degrees of freedom. This describes a moderately leptokurtotic distribution (eg. Meuwissen et al. 2001). Using the predicted SNP effects from each method, we predicted the GEBVs in the validation sets as $\mathbf{GEBV} = \hat{\mathbf{v}} + \mathbf{X}\hat{\mathbf{u}}$. The realised accuracy of GEBV was derived as described for BLUP above.

Results and discussion

We have attempted to describe the distribution of QTL effects by estimating the distribution of the variance explained by chromosome segments across the genome, Figure 1. For all three traits most segments explain <0.1% of the variance and for proportion black 96% of segments fall into this category. Collectively these segments appear to explain half the variance in these three traits. However, there are tens of segments that explain 0.1-4.7% of the variance for each trait. For overall type there are no segments with bigger effects but for proportion black there are segments explaining 4.7% to 18.8% and for fat% there are segments explaining 4.7- 37.5%. The total variance explained appears to be greater then 100% probably because segments next to the segment containing DGAT1, for instance, explain a significant amount of variance so that the variance explained by DGAT1 is counted more than once. For proportion of black the segments explaining the largest proportion of the variance included the genes KIT, MITF and PAX5. Mutations in KIT and MITF have

been demonstrated to cause white spotting in horses and dogs respectively (Karlsson et al. 2007, Haase et al. 2008) and PAX5 is a regulator of MITF.

Next we investigated the effect of the distributions of loci effects on the accuracy of genomic estimated breeding values using the BayesA approach. The accuracies of genomic estimated breeding value were 0.38, 0.59 and 0.73 for overall type, proportion of black and fat% respectively from the BayesA analysis, Table 1. The accuracy of these GEBVs was compared to that obtained using a statistical analysis (BLUP) that assumed all SNP effects are sampled from a normal distribution and therefore no large effects exist. Accuracies of the GEBVs using the Bayes A method were higher than those using the BLUP method for fat% and proportion of black, but not for overall type, Table 1.

Our results demonstrate that large differences exist in the architecture of different complex traits. For both proportion of black and fat% there are segregating mutations of moderate effect so that the distribution of effects is leptokurtotic. This in contrast to overall type which has only loci of small effect, and the distribution of these effects could be assumed to be normal.

It interesting to speculate on why large effects are segregating for fat% and proportion of black, but not overall type. For fat%, the fact that DGAT1 continues to segregate in the population may reflect the change in breeding goal for dairy cattle over time (as described by Grisart et al 2002). The mutant allele decreases milk fat yield but increases milk volume so artificial selection is likely to have favoured it at times but not consistently. This swept the allele to moderate frequencies in the population. Mutations causing white spotting must have been selected by breeders of black and white cattle since it is their defining feature. Thus in both cases, mutations which would have been unfavourable before domestication, were selected and still segregate at intermediate frequencies. In contrast, mutations which have a large effect on conformation and hence overall type, may be deleterious even in domesticated cattle and so have not risen in frequency to a point where they explain a substantial part of the variance. Research on genomic selection finds little evidence for genes of large effect for most complex traits (eg. Maher 2008). Thus most complex traits are like overall type in architecture. Fat% and proportion black may be examples of transient situations where a change in selection pressure has driven a mutation to intermediate frequency. Recently Eyre-Walker (2009) argued that genes of large effect should explain much of the variation in complex traits. The experimental evidence does not seem to support this prediction. It may be that, although mutations of moderate effect occur (as demonstrated here for fat% and proportion black), they are very rare compared to mutations of small effect.

Information on the degree of leptokurtosis of the distribution of effects can be used to guide the design of experiments that will subsequently enable genomic predictions. Goddard (2008) developed a deterministic method to predict the accuracy of genomic estimated breeding values. The parameters of this formula were the number of phenotypic records in the reference population (N), the heritability of the trait (h2), the length of the genome (L), and the distribution of QTL effects. The distribution of effects could be either normal or leptokurtotic. When a normal distribution of effects is assumed, the accuracy of genomic breeding values can be predicted as where $a= 1+2 \lambda/N$, and $\lambda= qk/h2$, with $k = 1/\log(2Ne)$,

where Ne is the effective population size. The parameter q = number of independent chromosome segments in the population. The value of q used here was 2NeL, where L is the length of the genome in Morgans (Hayes et al. 2009). Using the same number of phenotypic records as were used in our experiment, and the same heritabilities of the traits, the deterministic prediction of accuracies is given Table 1. For leptokurtotic distributions, there is no closed form equation for the accuracy of breeding values, but these accuracies can be derived by numerical integration of the accuracy of predicting the effects given the distribution of effects over an assumed distribution of the frequency of effects (Goddard 2008). A t distribution with 4.012 degrees of freedom was used to model the distribution of effects, and a U shaped distribution of allele frequencies as expected under the neutral model was used. As expected, the leptokurtotic distribution of effects gave higher predicted accuracies of genomic breeding value than a normal distribution of effects. The observed accuracy of GEBVs for overall type in our experiment, 0.35, matches closely the prediction for accuracy of GEBV for a quantitative trait with the same heritability and a normal distribution of effects. Conversely, both fat% and proportion of black better match the predictions when a leptokurtotic distribution of effects was used.

Table 1: Accuracy of genomic breeding value (GEBV) from BayesA analysis and GBLUP, and determinstic predictions.

	Trait		
	Overall type	Proportion of black	Fat %
Number of records in reference set	756	327	756
Heritability of records	0.63	0.74	0.83
Realised accuracy of GEBV			
BLUP	0.42	0.46	0.63
BAYESA	0.38	0.59	0.73
Deterministic prediction of accuracy of GEBV			
Normal distribution of effects	0.35	0.26	0.39
Leptokurtotic distribution of effects	0.75	0.66	0.93

Conclusion

In this paper we have demonstrated that for some traits, such as overall type, a very large number of SNPs will be required to predict the trait with any accuracy. An approach where all SNPs are fitted simultaneously to derive a prediction equation, ignoring significance levels, should lead to higher accuracies of prediction, than an approach which uses only associations detected in GWAS with stringent thresholds. The accuracies achievable with this approach can be predicted deterministically provided we have some knowledge of whether the distribution of QTL effects is normal or leptokurtotic.

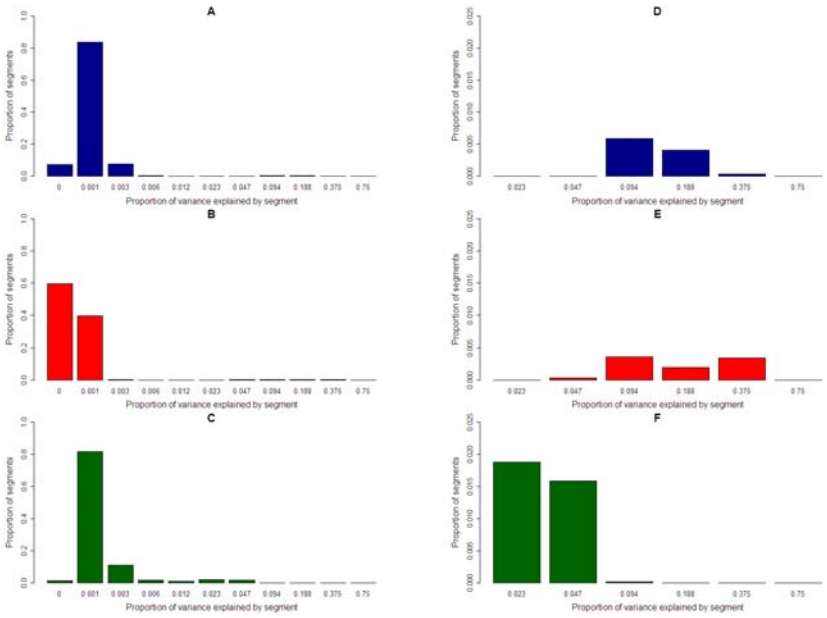

Figure 1: Distribution of proportion of variance explained by 50 SNP chromosome segments A. For proportion of black B. For fat % and C. For overall type. The x axis is on a logarithmic scale. Graphs D to F are the same results with the x axis from 0.023 to 1.0 proportion of variance explained.

References

Eyre-Walker, A. and Keightley, P.D. (2009). *Mol Biol Evol*. 26:2097-2108.

Daetwyler, H.D., Villanueva, B., Woolliams, J.A. (2008). *PLoS One.* 3: e3395.

Gilmour, A.R., Gogel, B.J., Cullis B.R. *et al*. (2002). ASReml User Guide Release 1.0. VSN International Ltd., Hemel Hempstead, UK.

Goddard, M.E., Hayes, B.J. (2009). *Nat. Rev. Genet.* 10: 381-391.

Goddard, M.E. (2008). *Genetica* Epub PMID: 18704696.

Grisart, B., Coppieters, W., Farnir., F, et al. (2002). *Genome Res.* 12: 222-231.

Haase, B., Brooks, S.A., Tozaki, T., et al. (2009). *Anim Genet*. 40: 623-629

Hayes, B.J, Bowman, PJ., Chamberlain, AC, Goddard, ME. (2009). *J. Dairy Sci.* 92:1313.

Karlsson, E.K., Baranowska, I., Wade, C.M., et al. (2007). *Nat Genet. 3*9: 1321-1328.

Lee, S.H., van der Werf, J.H., Hayes, B.J., et al. (2008). *PLoS Genet.* 4:e1000231.

Maher (2008). Nature. 2008 Nov 6;456(7218):18-21.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D. et al. (2009). *PLoS ONE* 4: e5350.

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

Nejati-Javaremi, A, Smith, C, Gibson, J. (1997). *J. Anim. Sci*. 75: 1738–1745.

Scheet, P., Stephens, M.A. (2006). *Am. J. Hum. Genet.* 78: 629–644.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R. *et al*. (2009). *J Dairy Sci* 92: 16-24.

Wray, N.R., Goddard, M.E., Visscher, P.M. (2007). *Genome Res.* 17:1520-1528.

Moser, G., Tier, B., Crump, R.E. *et al.* (2009). *Genet Sel Evol*. 31: 41:56.

Verbyla, K.L., Hayes, B.J., Bowman, P.J. Goddard, M.E. (2009). *Genet Res*. 91: 307-11.