Design of a sheep information nucleus as a reference population for genomic prediction

Julius van der Werf

Background and Aim

In Australia, both the sheep and the beef industry are looking for breeding program models where they can utilize genomic selection. The idea of genomic prediction is that based on a DNA test the breeding value of selection candidates can be predicted with a reasonable accuracy at a young age such that rates of genetic improvement can be increased, especially for traits that are not well measured at the time of selection. The DNA profile of the selection candidate is compared with that of a large number of animals in a reference population that have been genotyped for dense SNP markers as well as having phenotypic information on many traits relevant to the breeding objective, including those that are hard to measure or that are measured later in an animals' life.

The purpose of this document is to bring insight in the design of a reference population for genomic selection in sheep. The two main design questions about a reference population are 1) how large should it be? and 2) which animals should be in it? The first question is most important as the size is directly affecting the accuracy of genomic predictions, and therefore the benefit gained with genomic selection. It also directly affects costs. The latter question refers to the genetic constitution: which breeds, how many from each breed, which sires, how may progeny per sire, which dams, existing stud animals or newly created animals for the purpose?

Theoretical prediction of genomic prediction accuracy

Genomic prediction can be seen as predicting the effects of many segments of the genome of differences in phenotypes. Another way of looking at it is to compare the similarity of genome segments among individuals. The phenotype (or breeding value) of young animals can be predicted by comparing their segments with those that have been both genotyped and phenotyped. Those are referred to as the reference population.

Reference populations for genomic selection need to be large, with thousands of animals measured for phenotype and genotype. The accuracy of genomic predictions of breeding value (GBV) depends on the size of the reference population, how related it is to the animals to be predicted, the effective population size of the breed, the heritability of the trait. It also depends on the genetic model underlying the trait. If phenotypic variation is due to polymorphisms in only a few major genes with large effects, the prediction accuracy can be higher. However, evidence so far points to the fact that many traits are being affected by very many genes, each with very small effects on trait variation, and that these genes are generally scattered across the complete genome.

Goddard (2009) presented formulas that could be used to predict the accuracy of genomic prediction by estimating the number of independent segments in a genome, and the accuracy of the effects of each of these segments. The number of independent segments is dependent on the size of the genome, but also on the effective population size.

The effective population size (N_e) is formally a measured via the rate of inbreeding and is a reflection how related animals with a population are to each other. It is totally unrelated to true population size, and mainly depends on the genetic diversity of sires used in breeding programs. For example, the Holstein Friesian population has a population size of millions of dairy cattle, but due to a narrow genetic base (most sires used worldwide are descending from a handful of grandsires), the effective population size is less than 100. The means that the rate of inbreeding is the same as in a population of 100 individuals with equal number of males and females.

The effective size is strongly related to genomic selection accuracy as it affects the size of the genomic segments that animals in the population share. In populations with small effective size, such as Holstein Friesian and Border Leicester (N_e is smaller than 100) these shared segments are large whereas they are small in the merino breed which is more diverse and has a much larger effective size (N_e is around 1000). In populations with small N_e there are larger segments, hence there are fewer of them (there are fewer 'independent loci), hence their effects are larger and easier to estimate.

The effective population size is not so easy to measure. Most breeding populations have seen a large reduction in effective size due to more targeted selection and more use of common sires (hence, more inbreeding). Moreover, populations are often not fully closed, there is breed admixture, and they are not always homogeneous, e.g. the merino breed really consists of subpopulations of fine, medium and strong wool types. Therefore, in the modeling of genomic prediction accuracy it is a challenge to us the correct parameter for effective size.

The other important parameter determining accuracy is heritability. With lower heritability, there is more error variance, and it is more difficult to estimate the effect of genome segments accurately. More records are need for accurate prediction. Goddard's formula allows to show the relationship between number of records used (size of the reference population) and genomic prediction accuracy. This is illustrated in Figure 1 for 2 levels of heritability and for two values for effective population size. The populations refer to merino and terminal sire breeds, respectively.

The figure is derived using the formula of Goddard (2009) and using an approximation of the number of loci as $2.N_eL$, where L is the genome length. The approximation was suggested by Hayes et al. (2009) although there is still some uncertainty about this approximation. Hence, the figure maybe somewhat conservative in predicted accuracy, and consequently, the numbers required for a reference population maybe somewhat large.

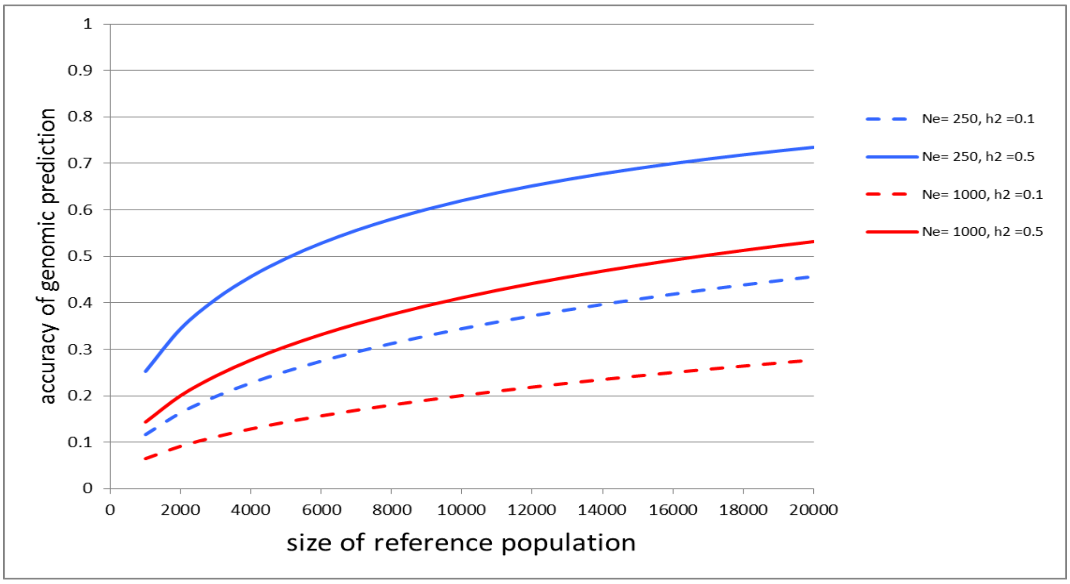

Figure 1: Accuracy of genomic prediction depending on size of reference populations for various heritability (h2) and effective population size (Ne: 1000 ~ merino; 250 ~ terminals)

Another consideration is that the accuracies shown are for 'unrelated individuals', i.e. animals share genome segments by being member of the same breeding population, but they are not closely related through pedigree. We can call this 'baseline accuracy' for genomic prediction. There will be more accuracy of genomic prediction if the reference population is related to the selection candidates. Hayes et al (2009) presented formulas for genetic prediction accuracy with relatives. They assumed genomic prediction was based on gBLUP, using the genomic relationship matrix to replace the pedigree based numerator relationships as used in traditional BLUP. Table 1 shows how the genomic predictions can gain in accuracy with relatives in the reference. The additional gains compared to current BLUP are not large, but the accuracy from using information on a few relatives is large compared to the accuracy obtained from information on 'non-relatives'. For example, information on 100 half sibs gives an accuracy of 0.44 for a low heritable trait (Table 1) whereas this would require many thousands of animals in a reference population if only unrelated individuals were used.

Table 1: Accuracy based on information on N half sibs, using either pedigree based relationships (A) or genomic relationships (G) *, Hayes et al. 2009.*

	N	accuracy	
		A	G
h2=0.5	10	0.37	0.39
	100	0.48	0.52
	1000	0.50	0.63
h2=0.1	10	0.22	0.22
	100	0.42	0.44
	1000	0.49	0.56

Genomic relationships provide more information than pedigree relationships because they give more accurate information about shared genetic information. Whereas pedigree relationships predict that full sibs have 50% and half sibs have 25% of genetic material in common (by descend) the genomic information will give a better prediction of actual gene sharing. The actual relationship among half sibs can vary from 0.2 to 0.3, with 0.25 only being the average. Therefore, genomic relationships give more accurate predictions than pedigree relationships.

Results in Table 1 show that, when using genomic relationship with relatives, the increase in accuracy relative to pedigree based prediction is relatively small unless there are extremely large family sizes. However, in genomic prediction we can also use information on less related individuals, and there can be very many of those. This information is not used in pedigree based prediction. The baseline accuracy of genomic prediction is in effect information on 'unrelated members of the same population' and if there are very many of these, accuracy can be high, as illustrated in Figure 1. In the context of a breeding program, it may be difficult to form reference populations with direct relatives of all future selection candidates. Large reference populations allow predicting also those not directly related to the reference. Furthermore, relationships become small and predictions based on direct relatedness only lasts for one to two generations. Predictions based many unrelated individuals in the reference population will wear out less quickly, and persist over more generations.

Genomic predictions are often discussed as based on linkage disequilibrium (LD) versus based on (genomic) relationships. LD refers to the marker alleles (SNPs) being in linkage disequilibrium with QTL (quantitative trait locus) alleles such that individuals with the same SNP allele (in a given region) are likely to have the same gene effects in that region. Within families such predictions are possible over large segments of chromosomes. Marker based predictions are possible within families even if marker and QTL are not close; this is termed prediction based on 'linkage'. Within breeds, there will be LD between SNPs and QTL over small distances, e.g.

<50 kb in sheep. With dense markers, we can assume that all QTL are at least in LD with one marker, making predictions based on SNP genotypes possible within breed. The amount of LD (i.e. the distance on the chromosome over which loci are in LD) is smaller for populations with large N_e. The ovine SNP chip has 50,000 SNPs, with the average distance between SNPs being around 60 kb (kilo bases). This is enough density to guarantee QTL-SNP LD within most breeds, and marginally enough for breeds with high N_e such as merino. LD across breeds exists even over shorter distances and the 50k chip is not expected to predict breeding value across breed as there is not sufficient LD. LD and relationships are to some extend highly related. One could say that there is a lot of LD within families, and also that all individuals within a breed are somewhat more related relative to relatedness across breed. It has been shown that genomic prediction accuracy declines when an animal is less related to the reference population. The accuracy of ‘unrelated’ animals falls back to the baseline accuracy, which is dependent on the number of individuals in the reference population. Prediction across breed is not possible as the ovine SNP chip density is insufficient. Whether a denser SNP chip will overcome that problem will also depend on other considerations, e.g. whether SNP effects are somewhat consistent across a wide variety of genetic backgrounds (i.e. across breeds).

Table 2: Comparison of prediction accuracy between groups differing in relatedness with the reference population for pedigree based BLUP prediction based on a shallow pedigree (BLUP-S), a deep pedigree (BLUP-D) and genomic relationship matrix (GBLUP).

Method	Close	Distant	Unrelated
Relatedness with reference	0-0.25	0-0.125	0-0.05
Method		Accuracy	
BLUP-S	0.39	0.00	0.00
BLUP-D	0.42	0.21	0.04
GBLUP	0.57	0.41	0.34

A simulation study (Clark et al. 2011) illustrates how accuracy is affected by relatedness (Table 2). They predicted a test set of animals based on a reference population that was either highly or moderately related, or unrelated to the test set. The first group had at least 20 half sibs in the reference set, those with moderate relationships had cousins and the unrelated reference set was up to 10 generations removed from the test set. The test set consisted of 750 animals, the reference set of 1750 animals, the heritability was 0.3 and the effective population size was 100.

For closely related animals, the accuracy of pedigree based BLUP predictions was high and the accuracy of GBLUP was somewhat higher. For moderately related animals, genomic prediction accuracy was lower, but pedigree based BLUP was much lower and even zero if only a one generation pedigree was used. For the ‘unrelated’ test set, pedigree based BLUP had virtually zero accuracy but GBLUP still gives a decent accuracy. As it turns out, the value of 0.38 is close to the theoretical value predicted by the Goddard equation, which gives 0.36 for ‘baseline accurcay’.

This illustrates that the accuracy predicted from the Goddard equation (Figure 1) can be considered as accuracies for individuals that are not directly related to the reference, but are member of the same breed (breeding population). Hence, they maybe not be related within 2 generations, but they are related at least within the last 10 generations. The average prediction accuracy of selection candidates will be somewhat higher because many will have more direct relatives in the reference population.

Empirical evidence of genomic prediction accuracy

Based on analysis of sheep CRC data we have estimated accuracies of genomic prediction. These accuracies were observed by comparing genomic breeding values (GBVs) of a few hundred sires in a test set, with their ASBVs, which were highly accurate and based on many progeny (Daetwyler, 2010). The ASBVs are acting here like the true breeding values, and the correlation between GBV and ASBV is a lower limit of accuracy, as the ASBV accuracy was usually lower than 100%, i.e. the ASBV is only an approximation of the true breeding value. For traits without ASBV we used cross validation, which is to sample repeatedly a test set, while treating the remainder as a reference set. The genomic breeding value accuracies obtained so far are listed in Table 3.

Table 3. Accuracy of genomic prediction based on sheep CRC data, augmented with Sheep genomics data) for a range of traits, and for main types of sheep (MER=merino; MAT = maternal (Border Leicester) and TERM = terminal (mainly White Suffolk and Poll Dorsett) *(Aug 2011)*

Trait	nr in reference	MER	MAT	TERM
BWT	4500	0.36	0.4	0.10
WWT	4500	0.47	0.38	0.16
PWWT	7180	0.35	0.36	0.10
PEMD	7166	0.40	0.21	0.40
PFAT	7163	0.40	0.12	0.25
Year.GFW	3341	0.60		
Year.FD	2831	0.55		
Year.SS	2471	0.25		
Adult GFW	3341	0.55		
Adult FD	2831	0.55		
BRWR (late)	4584	0.55		
CVFD	3057	0.46		
Staple Length	1734	0.50		

The correlations in Table 3 are obtained by using the full set of animals as a reference population, i.e. from a variety of breeds, and correlating GBV separately for the three breed types. The reference population for slaughter traits consists for ~60% of merino haplotypes, 30% of terminal haplotypes and 10% of maternal haplotypes. For wool traits it is 100% merino. As predictions across breeds generally give no additional accuracy, as illustrated in Table 4, the actual reference population size is therefore lower per breed. The obtained accuracies in Table 3 are generally much higher than expected, when based on the predicted accuracies in Figure 1. For example, effectively we used only about 5000 merinos to obtain genomic prediction accuracies for weight traits around 0.40. The heritability is between 20% and 30%, hence the expected accuracy was only 0.24 (using N_e = 1000). For wool traits the expected accuracy would be around 30%, whereas we observed values higher than 0.5.The discrepancy can be explained perhaps by the fact that merino is a heterogeneous breed with quite distinct subpopulations. Consequently, the prediction based on a N_e value of 1000 maybe too conservative, as the subpopulations are effectively smaller. Another explanation is that compared to Figure 1, which gives ‘baseline accuracies, there is some degree of relationship between the animals in the reference and those validated based on their ASBV. The terminal breed accuracies were also higher than expected, as effectively per breed we used less than 1000 animals. This would give a predicted baseline accuracy of around 20% for a trait with a heritability value of 0.3. The terminal breeds are also not completely pure and separate, and some of the admixture that cannot be fully picked up based on pedigree will contribute to genomic prediction accuracy. Hence, based on real data, the obtained accuracies are higher than the baseline accuracies predicted with the Goddard formula. This means that the baseline accuracies are probably conservative approximations of genomic prediction accuracy in sheep. The realized accuracies are higher, probably due to additional information of related individuals and due to capturing some of the between breed variation that is not always captured by pedigree.

Genomic accuracy results from dairy cattle are often more in line with baseline values. However, dairy cattle often use large reference populations, so the baseline accuracy is high. The additional value of relatives' information is lower in that case. Information on relatives is relatively much more important if the baseline accuracy is low.

Deciding on the required size of a reference populations

The amount of resources that one can afford to invest in reference populations will depend on the additional gains that are expected from genomic selection. Larger reference population will deliver more accuracy of genomic prediction, but there is a diminishing return. The additional genetic gain that can be achieved through genomic selection depends on the breeding objective and varies between species and between breeds and objectives. It is mainly driven by how well traits can be selected for with current methods.

Genetic gain in breeding programs is based on a multiple traits objective. Multiple trait selection is facilitated by an index that weigh the information according to the relative economic importance of the breeding objective traits. Some traits are easier to select for than others, and accordingly, more progress can be made via phenotypic selection. Traits that are easy to improve have high heritability, can be measured before the time of selection of breeding animals, and are not unfavourably correlated to other important breeding objective traits. Traits that are difficult to improve are those that cannot be easily measured at the time of selection, traits that are sex limited, have low heritability or are too expensive to measure. Genomic selection especially favours traits that are hard to select for otherwise, as those have relatively the largest increase in accuracy of breeding value at the time of selection. Breeding value accuracy is directly related to potential progress. The potential gain of genomic selection will depend on the 'measurability' of the main traits in the breeding objective. For example, the main traits in dairy can only be measured in females, so genomic selection has a large impact on potential gain as it allows accurate selection of young bulls, rather than waiting for their progeny test. Model studies have shown that the potential gains of genomic selection in sheep are moderate, varying from 40% in wool breeding objectives to 20% for meat objectives (van der Werf, 2009). The gain in wool is mainly based on the possibility to select earlier for adult wool traits. The gain is meat sheep are limited as measurement of weight measures and ultrasound scanning give accurate predictions of breeding value for the breeding objective. It is possible that carcass and meat quality traits will become more important, in which case genomic selection in meat sheep would become of larger value.

A certain size of the reference population will give certain genomic prediction accuracy, and this will vary depending on trait heritability (Fig. 1). More records are needed for low heritable traits to achieve the same accuracy. However, typically in breeding programs, it is not possible or efficient to obtain an equally high accuracy for all traits, but rather the accuracy is increased proportionally to the heritability of the trait. The squared value of accuracy of an EBV is a measure of how much variance of the true breeding value is captured with the EBV. The additional genetic gain is more or less proportional to the explained variance. For example, the increase in rates of genetic improvement with genomic selection was up to 40% for merino objectives, and up to 20% for terminal objectives for a scenario with accurate genomic selection where the squared accuracy of GBV was equal assumed to be equal to trait heritability (Van der Werf, 2009). In a scenario where the genomic selection accuracy was reduced (squared accuracy equal to half of heritability) the corresponding additional genetic gains were about halved; 20% for a merino objective and 10% for a terminal objective.

To achieve genomic prediction accuracy close to the square root of heritability for each trait, the size of the reference population needs to be 30,000 for merino and 10,000 for terminal breeds and 5000 for Border Leicester. These figures are derived from Fig. 1 and based on Goddard's formula assuming N_e = 1000, 250 and 100, respectively. Using the same formula for achieving accuracy close to the square root of half the heritability would require 12000, 4000 and 2000, for merino, terminal breeds and Border Leicester, respectively. Note that these targets are not fully achieved for the high heritable traits (Table 5). These are very large reference populations and as noted before, the predicted accuracies maybe to conservative when comparing them with the realised estimated from analysis of CRC data. Two critical parameters in the Goddard formula are difficult to determine. One is the effective population size, and the other is the effective number of loci (approximated currently as $2N_e.L$). Other reasons why realised accuracies could be higher are that the animals in the test set maybe somewhat related to the reference population, and that breeding populations are not fully closed and homogeneous. It is difficult to predict the additional accuracy that could arise as some of these factors are difficult to quantify. Hence, values in Table 5 maybe baseline accuracies and somewhat lower than what can be achieved in practice, but it is hard to predict by how much. Results obtained for CRC analysis (Table 3) suggest that the difference could be fairly substantial.

Table 5 Accuracy (lower limit) of genomic prediction depending on size of reference population, for various values of heritability and effective size.

Breed	merino	WS, PD	BL
Ne	1000	250	100
Size of reference pop'n	30,000	10,000	5,000
Progeny measured per year[1]	3750	1250	625
h2=0.1	0.33	0.34	0.35
h2=0.3	0.51	0.53	0.54
h2=0.5	0.60	0.62	0.63
Predicted benefit in dG	40%	20%	?
Size of reference pop'n	12,000	4,000	2,000
Progeny measured per year[1]	1500	500	250
h2=0.1	0.22	0.23	0.23
h2=0.3	0.36	0.37	0.38
h2=0.5	0.44	0.46	0.47
Predicted benefit in dG	20%	10%	8%

[1] *assuming the reference population is 'refreshed' every 8 years.*

The accuracies predicted by Goddard's formula are expected to last over several generations, as they are not based on direct relationships. Therefore, such large reference populations 'last' at least for several generations. The number required to be measured each year can be a proportion of the total such that the reference is gradually refreshed and kept up to date. For example, if the reference population was 'refreshed' every 8 years, then only 12.5% would have to be measured each year. For the 'low accuracy scenario' this would still mean that 1500, 500 and 250 progeny need to be tested for merinos, terminals and maternals, respectively. This would give a predicted increase in genetic gain of 20% for merinos and 10% for terminals. Note that these are numbers per breed; there would need to be 500 for each of the Poll Dorsett and White Suffolk breeds. Actual accuracies maybe higher as predicted due to animals having relatives in the reference population. The average relatedness (based on pedigree) of the current CRC INF sires and the 2010 drop of young rams is discussed in Appendix 2.

Prediction across breeds

Based on the genomic selection theory, prediction across breeds is not possible without a marker density that allows linkage disequilibrium across breeds. The current 50k ovine chip is barely enough for LD within merino. How much LD exists between breeds needs to be investigated with higher density marker panels. Empirical evidence based on analysis of CRC data has confirmed that prediction across breeds gives low to zero accuracy (Daetwyler et al., 2012). Even breeds that seem to be closely related cannot be estimated from related breeds. For example, Dohnes and SAMMs are closely related to merino, but in fact they are still many generations removed. More detail on across breed prediction within the CRC data is in Appendix 1.

Breeds that are not well represented in the reference set cannot be predicted with adequate accuracy. The consequence is that each breed that requires genomic predictions will need to be represented in a reference population. The reference population for a breed will depend on its effective size, as discussed before. Some small breeds may have arisen from imports of only a few rams, and consequently would have very small effective size, provided they remained a closed breeding population. For example, a population based on 5 sires would have an N_e of about 20, and a reference population of 1000 animals measured would already give a high prediction accuracy. Hence, for small breeds it is recommended to estimate the effective size as breeds with small N_e require smaller reference populations.

Design of the reference population.

Given that reference populations are needed for each breed, the main question is how to select sires, how many progeny and how would these sires be picked?

Animals to be tested in the reference populations should be selected from a diverse genetic background within the breed, but also from family lines that can be expected to contribute to the future gene pool in that breed. So there needs to be a balance between merit and diversity. A good strategy is to select progeny from young sires that have a high genetic merit, yet that are relatively unrelated to each other. The number of progeny tested per sire should be small, about 10 progeny per sire, but accuracy is not highly dependent on that number. However, smaller progeny groups allow testing more sires which is desirable from an ‘industry engagement’ point of view.

All progeny should be genotyped. Genotyping sires only give no information about segregating alleles within the sire families. A strategy where only sires are genotyped requires a lot more progeny measurement. For example, for N_e = 250, h^2=0.3 and 200 progeny measured and genotyped gives an accuracy of 0.27. The same accuracy would be achieved if 1300 sires were genotyped, each with 10 progeny, where ‘heritability of progeny mean’ is equal to 0.45. Hence, although fewer animals need to be genotyped, a lot more will require a phenotype.

For traits not usually measured, such as carcass traits, new animals needs to be generated and measured to inform selection candidates in the breeding nucleus. This is as in the current INF model. However, other traits useful for genomic selection could be measured on-farm, e.g. NLW, Adult weight, Adult wool traits. For such traits the reference population could be formed by the previous generations, i.e. ancestors of the current selection candidates. This is current practice in the dairy industry. This model has some challenges as it will be more difficult to test the genomic diversity of the breed.

Cost benefit

A more detailed cost benefit analysis is needed and requires a separate discussion paper. However, to give a ballpark figure, the following example can be used. At an industry level, the benefits of breeding programs are large, and benefits of small increases of rates of genetic gain are large as well. However, a problem is that the benefits are not all captured by breeders or even producers. Most benefits are ultimately captured by consumers.

Assume 20 million sheep are affected by improvement and current rate of genetic improvement is $2/production ewe per year. The current NPV of the genetic improvement over the next 20 years is $3.32 billion. Then a 10% increase in rate of genetic gain would give a cumulative NPV over the next 20 years of about 250 million; from 3.32 billion to 3.57 billon. Much of this benefit would end up with consumers. If only 10% would be captured by the farming sector, they could afford to invest of up to 2.2 million per year, as the NPV of such a cost per year over the next 20 years would be about $25 million.

Acknowledgements

The paper was based on analysis of sheep CRC data and discussion with the team and other colleagues. The data analysis were mainly done by Hans Daetwyler, Sam Clarke, Andrew Swan and Nasir Moghaddar, and further discussion with Ben Hayes, John Henshall, Brian Kinghorn, John Hickey, Bruce Tier, Andrew Swan, Daniel Brown, John Gibson, Hans Graser, Rob Banks, Alex Ball and Sam Gill are acknowledged.

References

Clark S, Hickey J.M. and van der Werf J.H.J. (2011) *Proc. Assoc. Adv. Anim. Breed. Genet*. **19.** 38

Daetwyler H., Kemper K., Hayes B.J. and van der Werf J.H.J. (2011) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **19.**

Daetwyler, H.D., K. Kemper, J.H.J van der Werf and B.J. Hayes. 2012 Components of the Accuracy of Genomic Prediction in a Multi-Breed Sheep Population J ANIM SCI 90:3375-3384; doi:10.2527/jas.2011-4557

Daetwyler H.D., Hickey J.M., Henshall J.M., Dominik S.,Gredler B., van der Werf J.H.J. and Hayes B.J. (2010) *Animal Production Science*, **50**: 1004-1010.

Goddard ME (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136**, 245-257.

Hayes, B.H. Visscher and Goddard (2009). Genet. Res. **91**:47-60.

Van der Werf, J.H.J. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **18**: 38.