Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.115

The use of multi-breed reference populations and multi-omic data to maximize accuracy of genomic prediction

<1 line>

M. E. Goddard[1,2], I.M. MacLeod[2], K.E. Kemper[3], R. Xiang[1], I. Van den Berg[1], M. Khansefid[2], H. D. Daetwyler[2] & B.J. Hayes[3]<Times 12, italic, no title(s)>

<1 line>

[1] *Faculty of Veterinary & Agricultural Science, University of Melbourne, Parkville, VIC 3010, Australia. mike.goddard@dpi.vic.gov.au(Corresponding Author)*

[2]*Agriculture Victoria, Centre for AgriBioscience, Bundoora, VIC 3083, Australia.*

[3] *Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, University of Queensland, St. Lucia, QLD 4067, Australia.*

[4] *School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia.*<Times 12, italics>

<2 lines>

Summary<Times 14, bold>

<1 line>

The accuracy of genomic prediction can be increased by increasing the size of the training population by using a training population consisting of multiple breeds. This is most beneficial for numerically small breeds and for traits which are difficult to measure. The benefit is greatest when using markers selected from genomic sequence using a Bayesian statistical method. The selection of useful markers may be increased by using additional traits such as gene expression, and by using prior information about the polymorphic sites derived from functional annotation or evidence of selection. The final analysis to generate EBVs could utilise only the previously selected markers, for instance, in a one-step GBLUP analysis. <1 line>

Keywords: genomic prediction, gene expression, multiple breeds <Times 12, italic, maximum … keywords, no capitals, no & or and>

<1 line>

Introduction <Times 14, bold>

<1 line>

Genomic selection (Meuwissen et al 2001) has been widely applied in livestock breeding. For traits where a large reference or training population can be assembled within a breed (e.g. milk yield in Holsteins) the accuracy of genomic EBVs is already high. However, for traits that are not routinely recorded and for numerically small breeds it is difficult to assemble a sufficiently large training population. This is an important limitation because traits that are poorly recorded often make up a substantial part of the breeding objective (e.g. food conversion efficiency, fertility). Further, genetic diversity among commercially useful breeds will be eroded if numerically large breeds make faster genetic gains than small ones. Accuracy of genomic EBVs (GEBVs) is critically dependent on the size of the training population and one way to increase this size is to combine the training populations from multiple breeds. Another way to increase accuracy of GEBVs may be to use information on the function of sites in the genome (e.g. in gene regulation) which may help us to select polymorphisms which are useful in the prediction of breeding value. In this review we will consider how to maximize the accuracy of GEBVs using a multi-breed training population and functional data.

<1 line>

Genetic architecture of economic traits<Times 14, bold>

<1 line>

Although some single genes with a large effect on economic traits occur, most of the genetic variation in profitability is due to complex traits controlled by many genes. In fact, recent evidence is that most quantitative traits are controlled by 1000s of polymorphisms (Wood et al. 2014, Boyle et al 2017, Moser et al. 2015, (Park et al. 2011; Stahl et al. 2012)). This highly polygenic nature of economic traits affects the choice of statistical methods for calculating GEBVs as discussed below.

Two questions concerning genetic architecture that affect the use of a multi-breed reference population are: Do polymorphisms affecting a quantitative trait (QTL) segregate in multiple breeds and do they have the same effect on the trait in all breeds? The evidence to answer these questions is not conclusive. In some cases identified QTL segregate in many breeds (e.g. DGAT1, (Grisart et al. 2002);Spelman et al. 2002) and in many instances, QTL map to the same region in different breeds (e.g. Karim et al. 2011). Kemper et al. (2015a) examined 12 milk production QTL in Holsteins and found 6 of them segregated in Jerseys, although it is possible that the other 6 segregated in Jerseys at too low a minor allele frequency (MAF) to be detected. Kemper et al. (2015a) found that most QTL were older than the separation of breeds and therefore it is more likely that Jersey have lost 6 QTL through selection and drift than that Holstein have gained 6 QTL through mutation. This implies that, although QTL do not segregate in all breeds, they are usually not restricted to a single breed. Pausch et al (2017) found many QTL for fat and protein concentration in milk segregate in more than one breed.

Differences between breeds in the effect of a QTL (i.e. a QTL by breed interaction) imply non-additive effects of the QTL. Estimates of dominance and epistatic variance are usually modest to low (e.g. Jiang et al. 2017, Bolormaa et al. 2015, Aliloo et al. 2016, Aliloo et al. 2015) and this is consistent with high genetic correlations between purebred and crossbred effects of sires. Also, known QTL such as DGAT1 have at least qualitatively similar effects in different breeds (Grisart et al 2002). Therefore, our best estimate, based on inadequate data, is that the correlation between QTL effects in different breeds will typically be >0.8. Contrary to this, Khansefid et al. (2014) found that between breed effects of SNPs explained only half their variance but this reflects differences between breeds in linkage disequilibrium (LD) as well as QTL by breed interactions.

Given this background, we now consider methods to increase the accuracy of GEBVs using multi-breed reference populations. There are two problems that apply to all cases – the effects of individual QTL are very small and extensive LD exists within a breed but varies between breeds.

Increase the size of training populations

In some cases we need to genotype more animals but in many cases the limitation is the number of animals measured for hard to measure traits such as meat tenderness or FCE. In some cases this requires experimental recording of traits, and/or derivation of phenotypes from multiple data sets. In the case of food intake, international collaboration is increasing the training population (de Haas et al. 2015). To produce GEBVs for heat tolerance, Nguyen et al. (2016) combined daily milk recording data with weather station records from near the farm on which the cows were living.

Statistical method

Most of the methods currently used can be described as Bayesian methods that differ in the prior distribution assumed for the SNP effects. The accuracy of GEBVs is highest when the assumed distribution matches the true distribution of SNP effects. GBLUP assumes that all SNP effects are drawn from the same normal distribution which implies that they all have small effects, and that every SNP has an effect. A number of Bayesian methods assume a prior distribution that includes some SNPs with zero effects and some with relatively large effects (Meuwissen et al. 2001, Habier et al. 2011, Erbe et al. 2012). The accuracy of GBLUP is unaffected by the genetic architecture of the trait i.e. the number of QTL and the distribution of their effects (Daetwyler et al 2010, Clark et al. 2011). The accuracy of GBLUP can be predicted by assuming that it estimates the effects of segments of chromosome. The greater the linkage disequilibrium the smaller the number of effective chromosome segments (Me), the larger their effects, the easier they are to estimate and the higher the accuracy of GEBVs (Hayes et al. 2009). Bayesian methods give higher accuracy than GBLUP when the effects of chromosome segments are not normally distributed. This can occur because some segments have a large effect (due to containing a QTL of large effect) or because the number of QTL is less than the number of segments so that some segments have zero effect. Within Holstein the number of effective segments is approximately 4000 based on historical Ne (MacLeod et al. 2014))or the observed accuracy of genomic prediction (van den Berg unpublished) This is greater than that estimated from the genomic relationship matrix (Wientjes et al 2013) due to close pedigree relationships among the animals in the GRM. However, the number of QTL is probably > 4000 for most traits but some QTL of large effect exist for some traits, so Bayesian methods give similar or slightly higher accuracy than GBLUP within breed. On the other hand, the phase of LD is only consistent across breeds if the polymorphisms are closely linked. This implies the effect of many small chromosome segments must be estimated to obtain a prediction equation that can be used across multiple breeds i.e. the effective number of chromosome segments is high, perhaps 60,000. This is probably more than the number of QTL implying Bayesian methods will give higher accuracy than GBLUP when the training population and the target population are of different breeds or when the training population contains multiple breeds, and this expectation is born out in practice, at least for some traits ((Erbe et al. 2012), Rolf et al. 2015, (Hamidi Hay and Roberts 2017), Lu et al. 2016, Kemper et al. 2015b, Bolormaa et al. 2013).

Marker density

The number of markers, i.e. SNPs, used should be large enough so that all QTL are in LD with one or more of the SNPs. If SNPs have the same spectrum of MAF as QTL, then the proportion of the genetic variance explained by the SNPs is M/(M + Me) where M is the number of SNPs. Within a breed such as Holstein where M_e = 4000, 50,000 SNPs should explain >90% of the genetic variance. However, between breeds, M_e is much larger and perhaps 600,000 SNPs are needed for the same coverage (i.e. 600K/(600K+60K) ~ 0.9). However, if QTL have a lower MAF on average than SNPs, the SNPs may not detect as much genetic variance as expected. A possible solution is to use genome sequence instead of a SNP panel. This has the advantage that the QTL are included in the genotype data which should in theory explain all the genetic variance (though this does rely on all QTL being detected in the pipelines that call variants from sequence data).

Even if all QTL are included in the data, the majority of sequence variants are in imperfect LD with the QTL, limiting the prediction accuracy when GBLUP is used (de los

Campos et al., 2013, Perez-Enciso et al., 2014), especially for prediction across breeds (van den Berg et al., 2016a). Bayesian methods benefit more from increasing marker density than GBLUP (MacLeod et al. 2014, Macleod et al. 2016). However, in practice the increase in accuracy achieved from sequence data to date using various approaches have been very small to small, e.g. 1 -2.7%, and in Australian and North American dairy cattle (MacLeod et al. 2016, Van Raden et al. 2017), 0% in Dutch Dairy Cattle (Calus et al., 2016, Veerkamp et al. 2016), and 2-8% in Nordic and French dairy cattle (Brøndum et al. 2015, van den Berg et al., 2016b). One reason for this is that genomic prediction is based on imputed sequence, which contains errors, not actual sequence. Imputation errors that reduce the accuracy of the imputed sequence by 5% also decrease the accuracy of GEBVs by about 5% which could eliminate any gains made from the sequence (van den Berg, unpublished results). The accuracy of imputation will be improved by a larger reference population of cattle with genome sequence and perhaps by better methods of imputation. However, the ideal method is to genotype selection candidates directly for markers that have an effect in the prediction equation.

Multi-trait analysis

The Bayesian methods can be described as consisting of two parts – selecting which SNPs to include in the model and estimating the effect of those SNPs, but GBLUP includes all SNPs so only has to estimate their effects. Analysing multiple traits together could help with one or both of these tasks. If the traits are genetically correlated there must be SNPs with effects on both traits. However, as in conventional BLUP, if all animals are measured for all traits and those traits have similar heritabilities and genetic correlations similar to phenotypic correlations, then there is little gain in accuracy by multi-trait analysis. A gain in accuracy does occur when some traits are measured on some animals only, so effectively the multi-trait analysis increases the size of the training population (Jia & Jannink 2012; Maier et al. 2015).

Even if traits have little genetic correlation, they may share some QTL. In this case the multi-trait analysis could improve SNP selection but not help to estimate the effects. For instance, a mutation in SLC37A1 (a phosphorous anti-porter) has a large effect on milk phosphorous concentration and a small effect on milk yield (Kemper et al. 2016). In this case the phosphorous concentration data helps to identify the causal variant for the milk QTL (or a SNP in high LD with it). A noticeable improvement in the accuracy of GEBV using such an approach would be reliant on improved SNP selection for many such small-effect QTL.

Gene expression

One class of traits with large effects is gene expression. In particular, cis eQTL often explain a large fraction of the variance in the expression of a gene. For instance, the same polymorphism that is most significant for milk phosphorous concentration also affects the amount of mRNA from the gene SLC37A1 (Kemper et al. 2016). It appears, from results in humans, that many QTL are not coding mutants and so presumably have their effect on phenotype via an effect on gene regulation (Schaub et al. 2012), and there is some evidence that this may also be the case in cattle (Koufariotis et al. 2014). Therefore, gene expression may be widely useful in selecting polymorphisms to include in the statistical model for economic traits. However, estimating the effect on the economic trait will still rely on the data for that economic trait.

A common way to measure gene expression is by sequencing RNA (RNAseq). Short read RNA sequence can be used to estimate the expression of a gene simply by counting the number of reads from each gene. Polymorphisms affecting gene expression levels, (gene eQTL) are very common in humans and sometimes affect human complex traits (GTEx Consortium 2015). A single gene can give rise to multiple mRNA transcripts depending on which exons are sliced out. The splicing varies between individuals so the proportion of each transcript varies between individuals. Polymorphisms that affect the proportion of each transcript from a gene are known as splice eQTL and affect complex traits in humans (Li et al. 2016a, Li et al. 2016b). Polymorphisms that affect splicing also affect the level of expression of some exons within the gene and are identified as exon eQTL and overlap with splicing eQTL (Guan et al 2014). In our own research on cattle and in humans (GTex consortium 2015) exon eQTL are much more common than whole gene eQTL .

Gene expression varies between tissues, environments and physiological states (GTEx consortium 2015, Chamberlain et al. 2015). Fortunately, many eQTL affect gene expression in many conditions (Flutre et al. 2013) and combining eQTL information from multiple conditions and/or tissues increases the power of detection of causal mutations even when only about 100 animals have RNAseq data. However, it is still possible that a particular QTL that we wish to find does not affect expression in any of the available tissues and states.

Using RNAseq data we can also count the RNA copies from the two alleles of a gene separately and thus find cis eQTL that are heterozygous in an animal. This uses data (allele specific expression, ASE) that is independent of that used in conventional eQTL analysis which compares expression between animals and thus increases power to detect eQTL. While eQTL and ASE overlap greatly, there are cases where they are not the same. For example, imprinting can give rise to ASE.

eQTL are very common and this gives rise to another problem: although a trait QTL and an eQTL might map to the same region they may not be the same mutation. Thus methods are needed to show that an eQTL is indeed the same as the trait QTL (e.g. Hormozdiane et al 2016). If it can be established that they are the same, an advantage of gene expression over other traits is that it indicates which gene a QTL acts through and thus contributes directly to knowledge of the biology of the economic trait.

The use of multi-trait analysis, for example gene expression data with complex trait data, is to make the effect of a “small” QTL “larger” so that it is easier to identify and estimate. However, this does not overcome the other difficulty associated with selecting the right variant to use which is the LD between markers. That is, there may be a number of SNP all in high LD and as a result the association tests (with gene expression or other phenotypes) cannot differentiate them. The next source of information (functional data) does potentially overcome this problem.

Functional information

We use this term to mean information about the function of a site in the genome which does not depend on variation in that site. For instance, annotation of the genome tells us which sites are coding for proteins and the effect that a mutation might have e.g. synonymous or non-synonymous. However, most QTL are probably non-coding sites (e.g. Schaub et al. 2012), so a great effort has been made to annotate non-coding DNA in humans (the ENCODE project) and, for instance by the FAANG project, in animals (The FAANG consortium 2015). Many different assays have been used but several aim to detect chromatin that is ‘open’ and so can be bound by transcription factors and other molecules needed for transcription. Parts of the genome marked by these assays are enriched for complex trait QTL in humans (Li et al 2016). In cattle, Wang et al. (2017) recently found that sites identified by histone methylation and acetylation in the bovine liver were enriched for milk production QTL. However, a large number of sites will have regulatory annotations so it will still be difficult to tell which one might be responsible for a specific QTL.

Knowledge about genes rather than individual nucleotides might also be useful if we could predict which genes were likely to affect a given trait. For instance, Bolormaa et al. (2016) found many genes associated with fat metabolism affect the composition of fat in sheep. Moore et al (2016) found that genes that were differentially expressed in the corpus luteum and endometrium were enriched for SNPs affecting cow fertility. However, in general our ability to predict which genes affect a trait is weak at best (Boyle et al 2017).

Evidence of selection

If a site in the genome has been subject to selection it must have some effect on phenotype. If a site is conserved across mammals it must be deleterious if it is mutated. How much of the genetic variation in a trait is explained by polymorphism in such sites is unknown. Genomic sites might also be identified as having been under selection (selection signatures) and therefore having an effect on phenotype ((Hayes et al. 2009). However, selection signatures for complex traits are not easily detected ((Kemper et al. 2014) and it may still be difficult to distinguish which site among those in high LD is the causal polymorphism.

Utilising prior information in calculation of EBVs

When good prior biological information is available about QTL sites (as outlined above), this can be utilised in genomic prediction of breeding values by defining classes of sites and estimating the probability that each class affects the trait and/or the variance of the effects within each class (Macleod et al 2016). Using a multi-breed reference population, Macleod et al. 2016 showed that, this approach has the potential to improve the accuracy of genomic prediction.

Computational methods

Ideally we want to analyse data from many animals each with many markers including perhaps full sequence data. This is not well suited to routine genetic evaluation which must run quickly without experimentation. One way to overcome this problem is to conduct two analyses – one which finds the best markers to include in the model and perhaps estimates their variances and one that uses these markers to calculate EBVs. The first, 'research' analysis could use a Bayesian method and the second 'production' analysis could use GBLUP. Ideally the sequence variants included in the production analysis would be genotyped directly because, if they are imputed, there is a loss of accuracy due to imputation errors (van den Berg et al., 2017).

Even within a research analysis, use of full sequence data on a large number of animals is computationally demanding. We have improved the efficiency of the computation by using an EM algorithm (Wang et al. 2017), by parallelising the analysis, by progressively dropping SNPs from the model (Wang et al. 2017, van den Berg et al., 2017). However, all these methods tend to sacrifice some accuracy. For example, chromosome-wise selection of sequence variants gave promising results in a simulation study, while in real data, accuracies were at most similar to those obtained with HD SNPs (van den Berg et al., 2017). Selecting the best sequence variants from a dataset and then using them for genomic prediction with the same training dataset can lead to bias which decreases accuracy of EBVs.

Even if the number of SNPs is not too large, the number of animals included in the 'production' analysis can be very large especially if all genotyped and non-genotyped animals are included in a one-step analysis. Misztal and Legarra (2017) review computational methods for such an analysis. Most one-step methods are equivalent to imputing genotypes in the ungenotyped animals by a linear regression. Meuwissen et a (2015) argue that higher accuracy can be obtained by imputing the genotypes of the ancestors of genotyped animals using a segregation analysis.

In dairy cattle, Interbull have combined information from many countries so that dairy farmers can select the best bulls regardless of origin. This service has been of great value but the value is reducing as we move to selection of young bulls based on DNA genotypes. A new approach would be to combine information on SNP solutions (for the most predictive SNP derived from sequence data) from different countries rather than combining EBVs on bulls. This could allow the benefits of a large reference population without countries having to share raw data.

Conclusion

A multi-breed reference population can increase the accuracy of genomic prediction when the within-breed reference population is small. To gain benefit from multiple breeds we need dense markers, ideally genome sequence, and a statistical method that selects from the huge number of sequence variants those that are useful. This selection can be improved by the use of prior information about the sequence variants based, for instance, on assays for open chromatin. Many studies find that QTL for complex traits are enriched in some class defined by functional assays but unfortunately the enrichment is usually not great enough to clearly identify all the QTL for a complex trait and therefore the increase in accuracy of EBVs is limited. However, additional information is accumulating rapidly and we hope in the near future that the combination of multi-breed reference populations, genome sequence and functional information will lead to substantial increases in accuracy.

List of references

Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ. Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows. BMC Genet. 2015 Jul 22;16:89.

Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet Sel Evol. 2016 Feb 1;48:8.

Bolormaa S, Pryce JE, Kemper K, Savin K, Hayes BJ, Barendse W, Zhang Y, Reich CM, Mason BA, Bunch RJ, Harrison BE, Reverter A, Herd RM, Tier B, Graser H-U and **Goddard ME**. (2013) Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle. J. Anim. Sci. 91: 3088-3104.

Bolormaa S, Pryce JE, Zhang Y, Reverter A, Barendse W, Hayes BJ, Goddard ME. Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. Genet Sel Evol. 2015 Apr 2;47:26.

Bolormaa S, Hayes BJ, van der Werf JH, Pethick D, Goddard ME, Daetwyler HD. Detailed phenotyping identifies genes with pleiotropic effects on body composition. BMC Genomics. 2016 Mar 12;17:224.
Boyle EA, Li YI and Pritchard JK (2017) An expanded view of complex traits: From polygenic to omnigenic. Cell 169: 1177-1186.
Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, Lund MS. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 2015 Jun;98(6):4107-16.
Calus MP, Bouwman AC, Schrooten C, Veerkamp RF. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genet Sel Evol. 2016 Jun 29;48:49.
Clark SA, Hickey JM, van der Werf JH. Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol. 2011 May 17;43:18.
Chamberlain AJ, Vander Jagt CJ, Hayes BJ, Khansefid M, Marett LC, Millen CA, Nguyen TT, Goddard ME. Extensive variation between tissues in allele specific expression in an outbred mammal. BMC Genomics. 2015 Nov 23;16:993.
de Haas Y, Pryce JE, Calus MP, Wall E, Berry DP, Løvendahl P, Krattenmacher N, Miglior F, Weigel K, Spurlock D, Macdonald KA, Hulsegge B, Veerkamp RF. Genomic prediction of dry matter intake in dairy cattle from an international data set consisting of research herds in Europe, North America, and Australasia. J Dairy Sci. 2015 Sep;98(9):6522-34.
de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS genetics. 2013 Jul 11;9(7):e1003608.
Daetwyler, H. D., et al. (2010). The impact of genetic architecture on genome-wide evaluation methods. Genetics **185**: 1021-1031.

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 95:4114-4129

Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. 2013;9(5):e1003486.
Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12:222-231

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348:648–60.
Guan L, Yang Q, Gu M, Chen L, Zhang X. Exon expression QTL (eeQTL) analysis highlights distant genomic variations associated with splicing regulation. Quantitative Biology. 2014;2(2):71-9.
Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011 May 23;12:186.
Hamidi Hay E, Roberts A. Genomic prediction and genome-wide association analysis of female longevity in a composite beef cattle breed. J Anim Sci. 2017 Apr;95(4):1467-1471.
Hayes BJ, Chamberlain AJ, Maceachern S, Savin K, McPartlan H, MacLeod I, Sethuraman L,

Goddard ME (2009) A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. Animal Genetics 40:176-184

Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb). 2009 Feb;91(1):47-60.

Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo WJ, Bilow M Sul JH, Sankararaman S, Pasaniuc B & Eskin E. 2016. Amer. J. Hum. Genet. 99: 1245-1260.

Jia Y, Jannink JL. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 2012 Dec; 192(4):1513-22.

Jiang J, Shen B, O'Connell JR, VanRaden PM, Cole JB, Ma L. Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. BMC Genomics. 2017 May 30;18(1):425.

Karim, L, Takeda, H., Lin, L., Druet, T., Arias, J.A., Baurain, D., Cambisano, N., Davis, S.R., Farnir, F., Grisart, B., Harris, B.L., Keehan, M.D., Littlejohn, M.D., Spelman, R.J., Georges, M. & Coppieters, W. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nature Genetics 43, 405-13 (2011).

Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME (2014) Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics 15:246

Kemper KE, Hayes BJ, Daetwyler HD, Goddard ME. How old are quantitative trait loci and how widely do they segregate? J Anim Breed Genet. 2015a Apr;132(2):121-34.

Kemper KE, Reich CM, Bowman PJ, Vander Jagt CJ, Chamberlain AJ, Mason BA, Hayes BJ, Goddard ME. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 2015b Apr 17;47:29.

Kemper KE, Littlejohn MD, Lopdell T, Hayes BJ, Bennett LE, Williams RP, Xu XQ, Visscher PM, Carrick MJ, Goddard ME. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. BMC Genomics. 2016 Nov 3;17(1):858.

Khansefid M., Pryce J. E., Bolormaa S., Miller S. P., Wang Z., Li C., Goddard M. E. Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. Journal of animal science. 2014;92(8):3270-3283.

Koufariotis L, Chen YP, Bolormaa S, Hayes BJ. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics. 2014 Jun 6;15:436.

Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016a;352(6285):600-4.

Li YI, Knowles DA, Pritchard JK. LeafCutter: Annotation-free quantification of RNA splicing. bioRxiv. 2016b:044107.

Lu D, Akanno EC, Crowley JJ, Schenkel F, Li H, De Pauw M, Moore SS, Wang Z, Li C, Stothard P, Plastow G, Miller SP, Basarab JA. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. J Anim Sci. 2016 Apr;94(4):1342-53.

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, Schrooten C, Hayes BJ, Goddard ME. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016 Feb 27;17:144.

MacLeod IM, Hayes BJ, Goddard ME (2014) The Effects of Demography and Long-Term

Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics 198:1671-1684

Maier R, Moser G, Chen GB, Ripke S, Cross-Disorder Working Group of the Psychiatric Genomics Consortium, Coryell W, Potash JB, Schefner WA, Shi J, Weissman MM, Hultman CM, Landen M, Levinson DF, Kendler KS, Smoller JW, Wray NR, Lee SH. Joint analysis of psychiatric disorders increases accuracy of risk ofprediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet. 2015 Feb 5; 96(2):283-94.
Meuwissen THE, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome wide dense marker maps. *Genetics* **157**: 1819-1829.
Meuwissen THE, Svendsen M, Solberg T & Odegard J (2015) Genomic predictions based on animal models using genotype imputation on a national scale in Norwegian Red cattle. Genet. Sel. Evol. 47:79.
Misztal I and Legarra A (2017). Invited review: efficient computation strategies in genomic selection. Animal 11: 731-736.
Moore SG, Pryce JE, Hayes BJ, Chamberlain AJ, Kemper KE, Berry DP, McCabe M, Cormican P, Lonergan P, Trudee F & Butler ST (2016). Differentially expressed genes in endometrium and corpus luteum of Holstein cows selected for high and low fertility are enriched for sequence variants associated with fertility. Biol. Reprod. 94: 1-11.
Moser G , Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLoS Genetics DOI: 10.1371/journal.pgen.1004969.
Nguyen TTT, Bowman PJ, Haile-Mariam M, Pryce JE, Hayes BJ. Genomic selection for tolerance to heat stress in Australian dairy cattle. J Dairy Sci. 2016 Apr;99(4):2849-2862.
Pausch H , Emmerling R, Gredler-Grandl B, Fries R, Daetwyler HD & Goddard ME (2017). Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentage in milk at nucleotide resolution. bioRxiv doi:https://doi.org/10.1101/143404.
Park J-H, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Chatterjee N (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proceedings of the National Academy of Sciences 108:18026-18031

Pérez-Enciso M, Rincón JC, Legarra A. Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genetics Selection Evolution. 2015 May 9;47(1):43.
Rolf MM, Garrick DJ, Fountain T, Ramey HR, Weaber RL, Decker JE, Pollak EJ, Schnabel RD, Taylor JF. Comparison of Bayesian models to estimate direct genomic values in multi-breed commercial beef cattle. Genet Sel Evol. 2015 Apr 1;47:23.
Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012 Sep;22(9):1748-59.
Spelman RJ, Ford CA, McElhinney P, Gregory GC, Snell RG. Characterization of the DGAT1 gene in the New Zealand dairy population. J Dairy Sci. 2002 Dec;85(12):3514-7.
Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FAS, Kathiresan S, Wijmenga C, Gregersen PK, Alfredsson L, Siminovitch KA, Worthington J, de Bakker PIW, Raychaudhuri S, Plenge RM (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet 44:483-489

The FANNG consortium (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology* **16**:57

van den Berg I, Boichard D, Guldbrandtsen B, Lund MS. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study. G3: Genes, Genomes, Genetics. 2016a Aug 1;6(8):2553-61.

van den Berg I, Boichard D, Lund MS. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol. 2016b Nov 4;48(1):83.

van den Berg I, Bowman PJ, Macleod IM, Hayes BJ, Wang T, Goddard ME. Multi-breed genomic prediction using Bayes R with sequence data and dropping variants with a small effect. Genet. Sel Evol. 2017 VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 2017 Mar 7;49(1):32.

VanRaden PM, Tooker ME, O'Connell JR, Cole JB & Bickhart DM (2017). Selecting sequence variants to improve genomic predictions for dairy cattle.

Genet. Sel. Evol. 49:32Veerkamp RF, Bouwman AC, Schrooten C, Calus MP. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genet Sel Evol. 2016 Dec 1;48(1):95.

Wang M, Hancock TP, MacLeod IM, Pryce JE, Cocks BG, Hayes BJ. Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. Genet Sel Evol. 2017 Jul 6;49(1):56.

Wang T, Chen YP, MacLeod IM, Pryce JE, Goddard ME, Hayes BJ. Application of a Bayesian non-linear model hybrid scheme to sequence data for genomic prediction and QTL mapping. BMC Genomics. 2017 Aug 15;18(1):618.

Wientjes Y. C. J., et al. (2013). The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics **193**(2): 621-631.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics 46, 1173-86 (2014).

Yang YI, van de Geijin B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y and Pritchard JE (2017)