Proc. Assoc. Advmt. Anim. Breed. Genet. 23:11-14

INTEGRATION OF FUNCTIONAL GENOMICS AND PHENOMICS INTO GENOMIC PREDICTION RAISES ITS ACCURACY IN SHEEP AND DAIRY CATTLE

H.D. Daetwyler[1,2], R. Xiang[1,3], Z. Yuan[1,4], S. Bolormaa[1], C.J. Vander Jagt[1], B.J. Hayes[5], J.H.J. van der Werf[6], J.E. Pryce[1,2], A.J. Chamberlain[1], I.M. MacLeod[1] and M.E. Goddard[1,3]

[1]AgriBio, Centre for AgriBioscience, Agriculture Victoria, Bundoora, VIC, 3083 Australia
[2]School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia
[3]Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, VIC, 3052 Australia
[4]Lanzhou University, Lanzhou, China
[5]QUAAFI, University of Queensland, St. Lucia, QLD, 4067 Australia
[6]University of New England, Armidale, NSW, 2351 Australia

SUMMARY

In the animal breeding there is debate on whether knowledge of functional genomics is useful for genomic prediction. Black box approaches have worked well but technological change now allows for the generation of functional genomic and phenomic information at high resolution. This will allow us to come closer to actual functional variants, thereby increasing genomic prediction accuracy in animals less related to the reference population, such as across breeds and across generations. Here we demonstrate that even with current imperfect knowledge the use of functional information in genomic prediction results in immediate benefits to prediction accuracy and industry breeding decisions.

INTRODUCTION

Currently implemented industry genomic evaluations usually use single nucleotide polymorphisms (SNP) that are neutral and of medium density (e.g. 50k SNP chips in sheep and cattle). The evaluations rely on SNP being in linkage disequilibrium (LD) with causative mutations. This has been effective and has resulted in good prediction accuracy when reference populations are of sufficient size and when predictions are for animals that are relatively closely related to the reference. However, large LD blocks break down quite quickly across generations and LD is also only consistent across breeds at short distances that are not captured by medium density genotyping platforms. This reduces genomic prediction accuracy in these animal groups and imposes a shelf-life on reference populations. A solution is to find SNP that are not neutral but that are more closely linked to, or, are causative mutations. Purely statistical methods can do that with some success, but they are often limited in their ability to fine map causal variants and are susceptible to biases because it is difficult to keep association discovery and prediction reference populations independent. This is where additional independent functional information from other "omics" is helpful to prioritise SNP at finer scale. The overall idea is to reduce the millions of sequence SNP in whole genome sequence data to thousands, such that they can be routinely genotyped by industry and used in genetic evaluations without great computational challenges.

A plethora of high-resolution "omics" data can now be collected in relatively large numbers of animals providing newly defined intermediate phenotypes. Genome sequencing technologies have enabled several approaches to investigate regions of the genome that are associated with phenotypes as well as gene expression and regulation. Large global collaborative projects have created inventories of sequence variants in cattle (1000 Bull Genomes Project) and sheep (SheepGenomesDB) (Daetwyler *et al.* 2014; Daetwyler *et al.* 2017; Bouwman *et al.* 2018). The advantage of sequence

data is that the vast majority of SNP and short insertions/deletions (Indels) will be contained in the dataset, thereby enabling quicker discovery of causative mutations or variants that are very closely linked to these mutations (Hayes and Daetwyler 2019). Next-generation genome sequencing also underpins most assays that aim to interrogate gene expression and regulation, for example RNA and chromatin immunoprecipitation (ChIP) sequencing. Regulators of gene expression have been found to be important and enriched in regions that have been associated with phenotypes (Wang *et al.* 2018). Regulatory regions can be identified with expression quantitative trait loci (eQTL) mapping, where variants are associated with gene and exon expression as well as with splice variants (Chamberlain *et al.* 2018; Xiang *et al.* 2018). Similarly, SNP in highly expressed genes in relevant tissues can be identified and such information can be utilized directly in genomic prediction (MacLeod *et al.* 2019). Another functional assay that provides insight into regulatory regions is ChIP sequencing, which can provide information on histones with specific modifications that indicate regions that are likely to be enhancers, promotors or repressors of gene expression. Finally, molecular phenomics (e.g. metabolite levels) can reveal the abundance of compounds in the pathway between gene expression signals and phenotypes and can also be genetically mapped.

Our aim was to combine information from several omics-derived datasets to prioritize variants to increase the accuracy of genomic prediction. We demonstrate the advantage of using this additional information to raise the accuracy of genomic prediction with examples in sheep and dairy cattle.

MATERIALS AND METHODS

Sheep. 42 million sequence variants discovered by SheepGenomesDB Run2 (Daetwyler *et al.* 2017) were imputed into 46,000 sheep (Bolormaa *et al.* 2019). Only the 31 million sequence variants with a Minimac R2 >0.4 were used for downstream analyses. RNA sequencing was carried out on 150 wethers for muscle and liver tissues (Bolormaa *et al.* 2015). All data was aligned with the program STAR, counts were generated with the R package feature Counts, normalised for read depth. Expression QTL (eQTL) mapping was performed with gene and exon counts, as well as with splice variants at SNP 1 megabase (Mb) up and downstream of genes. A false discovery rate (FDR) threshold of 0.05 was used to determine significant SNP, which were then overlapped with significant QTL regions from a genome-wide association study on meat and carcass traits (individual animal phenotypes) also imposing a FDR of 0.05 (Bolormaa *et al.* 2016) and pruned for LD > 0.9. The same multi-breed reference population and traits as Khansefid et al (2018) were used to test two SNP sets: i) the 50k Ovine SNP chip and ii) the 50k Ovine SNP chip with the 10,000 significant eQTL sequence SNP added. Genomic prediction accuracy was validated in approximately 1000 Merino and 500 Border Leicester/Merino cross sheep for 6 meat traits (individual animal phenotypes). Validation animals were chosen to not have half-sibs in the training set to restrict relationships (Khansefid *et al.* 2018).

Dairy Cattle. 17 million sequence variants identified in the 1000 Bull Genomes Project Run6 were imputed into 44,260 animals (about 75% Holstein, 20% Jersey and 5% Australian Red breeds). Sequence variants associated with gene expression (eQTLs) and concentration of milk metabolites (mQTLs, phospholipids), and under histone modification marks (providing information on protein – DNA interactions) were discovered from multi-omics data in several tissues of over 400 cattle. Variants were also identified from 1000 Bull Genomes database (N=2,330) beef-dairy selection signatures. These analyses defined 30 variant sets and for each set we estimated the genetic variance it explained across 34 complex traits in 11,923 bulls and 32,347 cows. Only sets that explained more variance than a random set were carried forward in the analysis leaving approximately one million variants. We defined a Functional-And-Evolutionary Trait Heritability (FAETH) score indicating the proportion of the variance explained by each variant (Xiang *et al.* 2019). Further LD pruning and variant classification reduced the set to 40,000 variants that were included on a new Illumina XT SNP

chip design. Finally, we tested whether this new variant set increased genomic prediction accuracy using Bayesian genomic prediction method BayesR across milk, fat and protein yield, somatic cell count and fertility, when compared to the standard Illumina 50k SNP chip in an independent cow dataset (N range 538 (Crossbreds) to 2740 (Holstein)). Similarly to sheep, validation animals were not allowed to have sires or half-sibs in the training set.

RESULTS AND DISCUSSION

Sheep. One million eQTL were detected with significant overlap of eQTL between gene, exon expression and splice variation. Overlapping the eQTL with significant GWAS peaks resulted in 10,000 selected SNP that were added to the 50k Ovine SNP chip for genomic prediction. The increase in prediction accuracy from adding the 10,000 functional SNP was approximately 2 to 3% and varied between traits (Figure 1). In most traits Bayesian methods attained higher prediction accuracy than GBLUP as they are better at accommodating SNP with large effects (data not shown). Bias of genomic breeding values (slope of phenotypes on genomic breeding values) was unaffected compared to Ovine 50k results.

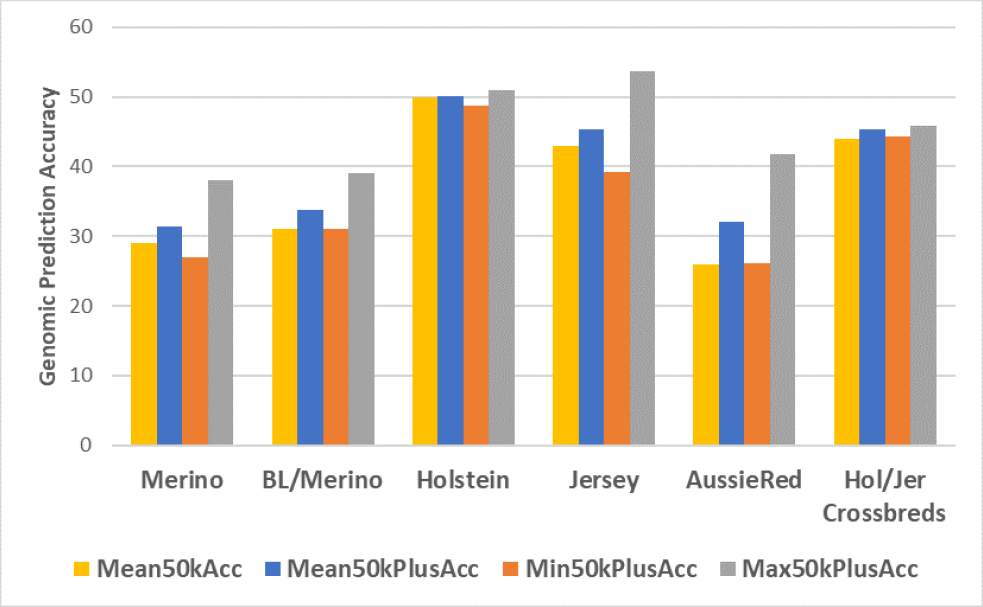

Figure 1. Genomic prediction accuracy when comparing standard 50k Ovine and Bovine SNP chips (50k) to SNP sets that include prioritised markers using functional information (50kPLus) in Merino and Border Leicester/ Merino cross sheep, as well as Holstein, Jersey, Aussie Red, and Holstein/Jersey crossbred cattle

Dairy Cattle. In the variant prioritisation work, the per-variant trait variance explained was highly consistent ($r > 0.98$) between bulls and cows across traits. Based on the per-variant heritability, the sets of mQTL, eQTL and variants associated with non-coding RNAs ranked the highest, followed by more recent mutations, those under histone modification marks, and selection signatures. A XT SNP chip with 40,000 variants from the prioritisation (as well as 8,000 markers overlapping with the Low-Density Dairy SNP chip) is currently in use for genotyping these variants directly (to avoid imputation errors). An early validation in cows not used in the prioritisation and using the imputed

high-value variants has increased prediction accuracy on average by 2.5% across all pure breed groups and traits (Figure 1). The increase in accuracy was more pronounced in crossbred, Jersey and Australian Red cattle, which is encouraging for these smaller breed groups, but could also be partly due to lower reference population sizes in those groups. Additional XT SNP chip results can be found in van den Berg *et al*. (2019).

CONCLUSIONS

A strategy to prioritize variants from whole-genome sequence using functional genomic, annotation, and phenomic information combined with target trait phenotypes has increased genomic prediction accuracy in animals that are less related to the reference population in both sheep and dairy cattle. This results in genomic breeding values that are more widely applicable across breeds (shown) and more robust across generations (not shown). The prioritized SNP sets can be utilized by industry immediately to increase prediction accuracy and genetic gain.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from DairyBio, a joint venture between Agriculture Victoria, Dairy Australia and the Gardiner Foundation. We are grateful to have access to data from the 1000 Bull Genomes Project, SheepGenomesDB, SheepCRC, SheepGenetics, and DataGene.

REFERENCES

Bolormaa S., Behrendt R., Knight M.I., … Daetwyler H.D. (2015) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **21**: 489.

Bolormaa S., Chamberlain A.J., Khansefid M., … MacLeod I.M. (2019) *Genet. Sel. Evol.* **51**: 1.

Bolormaa S., Hayes B.J., van der Werf J.H.J., … Daetwyler H.D. (2016) *BMC Genomics* **17**: 224.

Bouwman A.C., Daetwyler H.D., Chamberlain A.J., … Hayes B.J. (2018) *Nat. Genet.* **50**: 362-7.

Chamberlain A.J., Xiang R., Vander Jagt C.J., … Goddard M.E. (2018) *11th World Congress in Genetics Applied to Livestock Production*, Auckland.

Daetwyler H.D., Brauning R., Chamberlain A.J., … Kijas J.W. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **22***:* .

Daetwyler H.D., Capitan A., Pausch H., … Hayes B.J. R.F. (2014) *Nat. Genet.* **46**.

Hayes B.J. and Daetwyler H.D. (2019) *Ann. Rev. Anim. Biosci.* **7**.

Khansefid M., Bolormaa S., Swan A.A., … MacLeod I.M. (2018) *11th World Congress of Genetics Applied to Livestock Production*, Auckland, NZ.

MacLeod I.M., Bowman P.J., Chamberlain A.J., … Goddard M.E. (2019) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **23**:

van den Berg I., MacLeod I.M. & Pryce J.E. (2019) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **23**.

Wang M., Hancock T.P., Chamberlain A.J., Vander Jagt C.J., … Hayes B.J. (2018) *BMC Genomics* **19**: 395.

Xiang R., Hayes B.J., Vander Jagt C.J., MacLeod I.M., … Goddard M.E. (2018) *BMC Genomics* **19**: 521.

Xiang R., Van Den Berg I., MacLeod I.M., Hayes B.J., … Goddard M.E. (2019) *PNAS (online)*.