# USING SEQUENCE DATA TO IMPROVE ACCURACY OF GENOMIC PREDICTION AND QTL DISCOVERY FOR DAIRY COW FERTILITY

**I.M. MacLeod[1,2,3], B.J. Hayes[2,3,4], M. Haile-Mariam[2,3], P. Bowman[2,3], A.J.Chamberlain[2,3], C.J Vander Jagt[2,3], C. Schrooten[5] and M.E. Goddard[1,2,3]**

[1] Faculty of Veterinary & Agricultural Science, University of Melbourne, Victoria, Australia.
[2] Dairy Futures Cooperative Research Centre, AgriBio, Bundoora Victoria, Australia.
[3] AgriBio, Dept. Economic Dev., Jobs, Transport & Resources. Victoria, Australia.
[4] Biosciences Research Centre, La Trobe University, Victoria, Australia.
[5] CRV, 6800 AL, Arnhem, Netherlands

## SUMMARY

Using a Bayesian genomic prediction method, BayesR, we demonstrate improved accuracy of genomic prediction for cow fertility using high density SNP markers combined with imputed sequence variants in and close to gene coding regions. We also used the same analysis to identify candidate genes and potential causal mutations with a broad range of effects on fertility.

## INTRODUCTION

Dairy cow fertility has caused much concern over the past three decades in many countries because it had been in decline, partly due to an unfavourable genetic correlation between fertility and milk traits. Accurate genomic predictions of fertility would be of great benefit to the dairy industry because it is measured only in mature females and has low heritability. It is also important to identify genes affecting fertility to better understand genetic factors that underpin the trait.

BayesR, a Bayesian genomic prediction method, can achieve higher accuracy of genomic prediction compared to genomic best linear unbiased prediction (GBLUP), particularly for traits affected by many small effect genes as well as some of much larger effect (Erbe *et al.* 2012, Kemper *et al*. 2015). This occurs because BayesR models the single nucleotide polymorphism (SNP) effects as a mixture of four normal distributions, including a null distribution and one distribution with moderate to large variance.

BayesR should also be a more precise method for QTL discovery than GWAS (genome-wide association analysis) or GBLUP. GWAS fits SNP individually which often results in one QTL being predicted by a large number of SNP in LD. GBLUP fits all SNP simultaneously but effects are distributed as a single normal distribution so are smeared across many adjacent SNP with strong shrinkage of larger effects. Furthermore, BayesR provides a well calibrated test of the likelihood that a SNP predicts a real QTL effect (posterior probability).

Using BayesR we compare accuracy of genomic prediction for dairy cow fertility using high density SNP markers and imputed sequence variants in and close to genes. We also identify candidate genes and potential causal mutations associated with fertility.

## MATERIALS AND METHODS

We obtained dairy bull progeny test phenotypes of female fertility for 6804 bulls, including 5285 black and white Holsteins, 620 Red Holsteins, 803 Jerseys and 96 Australian Reds. Most bulls had MACE international breeding values and these were converted to de-regressed proofs (DRP) on the Australian scale (ie. corrected phenotypes: details in Haile-Mariam *et al.* (2015) – we used a subset of their data). The remainder (252) had daughter trait deviations (DTD) from the Australian Dairy Herd Improvement Scheme (ADHIS). Both the DTD and DRP were converted to the same scale using linear regression. The 620 Red Holsteins were our validation set and the remaining bulls (6184) made up the reference set.

All bulls were either genotyped or imputed for the Bovine HD SNP Illumina array ("800K") as described in Haile-Mariam *et al.* (2015). Bulls were then imputed for sequence variants in coding regions and in "regulatory regions" (defined as 5000bp either side of genes) using Beagle3 (Browning and Browning 2009). Run 3 of the 1000 Bull Genomes Project was used to discover these variants and provided the reference Holstein and Jersey bulls for imputation (Daetwyler *et al.* 2014). The 800K and imputed sequence genotypes were combined to give a third genotype set ("SEQ"). Very rare variants were pruned from SEQ (minor allele frequency: MAF < 0.002) as well one of any pair of SNP in perfect LD, preferentially keeping non-synonymous coding variants and then variants in regulatory regions. A total of 907,643 SEQ variants remained, most of which were SNP and a small number of indels.

BayesR was implemented as detailed in Kemper *et al*. (2015). Briefly we fitted the model:

$\mathbf{y} = \mathbf{Xb} + \mathbf{Wv} + \mathbf{Za} + \mathbf{e}$, where **y**=vector of phenotypes, **X**=fixed effects design matrix and **b**=vector of fixed effects (mean, breed, data type - DRP/DTD nested within breed). Here **W**=design matrix of variant genotypes centred and standardized (Yang *et al.* 2010) and **v**=vector of variant effects, distributed as a mixture of four distributions: $N(0, 0.0\sigma^2_g)$, $N(0, 0.0001\sigma^2_g)$, $N(0, 0.001\sigma^2_g)$, $N(0, 0.01\sigma^2_g)$ where $\sigma^2_g$=total genetic variance. **Z**=polygenic effects design matrix, **a**=vector of random polygenic effects ~ $N(0, \mathbf{A}\sigma^2_a)$, where **A**=pedigree relationship matrix and $\sigma^2_a$=additive genetic variance not explained by the genotypes. **e**=vector of residual errors ~ $N(0, \mathbf{E}\sigma^2_e)$ where **E**=diagonal matrix with a weighting coefficient based on effective number of daughter records per bull (Garrick *et al.* 2009). Analyses were performed with 50K, 800K or SEQ genotypes, each running for 40,000 iterations (20,000 discarded as burn-in) with 5 replicates per analysis. Results were derived from the mean of 20,000 iterations, averaged over all replicates. The posterior probability of a given SNP being included in the model was calculated as the proportion of iterations (post burn-in) that each variant fell in a non-zero distribution (averaged across replicates). Accuracy of genomic prediction was estimated as the correlation between predicted breeding value and corrected phenotypes. The mean squared error of the prediction for each analysis was calculated as the average of the squared difference between predicted breeding value and phenotype.

## RESULTS AND DISCUSSION

There was a clear trend for the accuracy of genomic prediction to increase with the density of genotypes (Table 1). The highest accuracy was achieved using SEQ genotypes, possibly because some causal mutations were included in the SEQ data. Also the imputed sequence included many more rare variants than in the 50K or 800K data (Table 1). Therefore, SEQ is more likely to include variants in strong LD with other rare/recent mutations, including rare causal mutations, than common SNP on the 50K or 800K arrays.

**Table 1. Genomic prediction using different densities of genotypes**

| Genotype Sets (Total number variants) | **50K** (37,236) | **800K** (600,640) | **SEQ** (907,643) |
|---|---|---|---|
| Proportion of variants with MAF[1] < 0.05 | 11% | 7% | 22% |
| Accuracy of Genomic Prediction (s.e.m.[2]) | **0.386** (0.0008) | **0.418** (0.0004) | **0.440** (0.0012) |
| Mean Squared Error of prediction | 166 | 148 | 144 |

[1] Minor Allele Frequency

[2] Calculated as: SD of accuracy from 5 replicates divided by √5

To better understand the contribution of sequence variants in and near coding regions, we tested the accuracy of prediction in the validation set using only the top 5000 variants (based on posterior probabilities) from either:

A. Non-synonymous coding variants and variants within 5000bp of genes, or
B. Intergenic variants excluding regions close to genes (±5000bp).

Prediction accuracy for group A = 0.35 while B = 0.27. When sets A and B were combined, prediction accuracy = 0.39. The high LD in the cattle genome means that there is considerable overlap in the predictions from set A and B. However, the results do suggest that variants in and close to coding regions explain a large proportion of the trait variance but that intergenic regions are also important in regulating trait expression, in keeping with evidence from the human ENCODE project (Skipper *et al.* 2012). The fertility trait appeared to be highly polygenic, with an average of 3305 SNP effects fitted in the model. This is expected because the fertility trait was largely based on calving interval: a complex trait influenced by many factors such as cow energy balance, oocyte health and embryo development.

In Figure 1 we present some examples of QTL discovery from among the 50 most significant QTL regions occurring within 5000bp of a gene: first to demonstrate the advantages of our approach and second to illustrate the range of genetic factors that underpin the complex female fertility trait.

Several regions only showed strong evidence for QTL in the SEQ analysis, demonstrating the improved power of SEQ genotypes. One example (Fig. 1a) is a rare variant in the 3' UTR of the SCARA5 gene showing a strong probability of being either the causal variant or one in strong LD with a QTL. This SNP is likely a relatively recent mutation because it segregated only in the black and white Holsteins (MAF=0.08) and was not in strong LD with any other SNP. SCARA5 expression is upregulated in human endometrium tissue when an early embryo is present (Duncan *et al.* 2011), and was also found to be more highly expressed in bovine ovary tissue compared to 17 other tissues (Chamberlain *et al*. 2014). A second example (Fig 1b) is two SEQ variants in high LD (Holstein only, MAF=0.025). The highest probability variant lies between SMEK1 and CCDC88C gene, while the other is a missense mutation in SMEK1. Potentially either gene could be considered to be a good candidate. CCDC88C is a negative regulator in the Wnt signalling pathway that regulates embryo germ cell development (Enomoto *et al.* 2006). SMEK1 has been demonstrated to regulate hepatic gluconeogenesis in mice (Yoon *et al.* 2010) and also appears to regulate the differentiation of embryonic stem cells (Lyu *et al.* 2011). In an analysis of the same data for milk traits (results not shown), these same mutations have a strong association with milk yield and the allele that increased milk yield reduced fertility.

A further region that showed a strong association for fertility and milk yield is between the GC and NPFFR2 genes (Fig. 1c). There is strong LD across this region and the association was spread across several variants. Again both genes are potentially good candidates: GC encodes Vitamin D transporter and disruption of the Vitamin D pathway affects oestrogen biosynthesis, while NPFFR2 interacts with kisspeptin which plays a key role in neuroendocrine regulation of reproduction (Matzuk and Lamb 2008). A number of regions on the X chromosome showed several strong QTL signals including SNP very close to KAL1 (Fig 1d) and UBE2A. In humans several mutations in KAL1 are responsible for "Kallmann syndrome", affecting the embryonic migratory pathway of neurons that synthesize gonadotropin-releasing hormone (Hardelin *et al.* 1992). This results in impaired gonad development in males and females. Mutations in UBE2A have been shown to be associated with maternal effects on early embryo survival (Matzuk and Lamb 2008).

The validity of our results are dependent on the accuracy of imputation and reference genome annotation, neither of which is perfect. However, this study demonstrates that imputed sequence genotypes with Bayesian analysis improved the accuracy of genomic prediction and the QTL discovery highlighted a broad range of genetic factors potentially affecting dairy cow fertility.

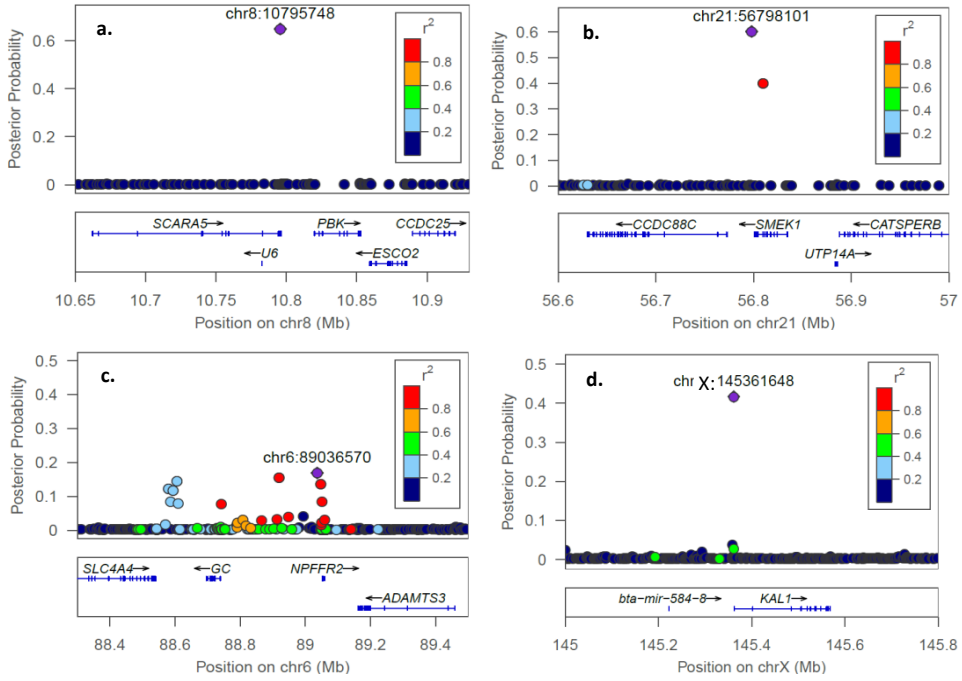


**Figure 1.** QTL discovery: The variant with the highest BayesR posterior probability is annotated and shown as purple diamond, and the LD ($r^2$) between this and other variants is colour coded. Genes are shown in blue with exons delineated by thicker bars.

## REFERENCES


Browning B.L. and Browning S.R. (2009) *American J Human Genetics* **84**: 210.

Chamberlain A.E., Vander Jagt C.J., Hayes B.J. and M. Goddard M.E. (2014) *Proc 10th World Cong. Genet. Appl. Livestock Prod*. Paper 180.

Daetwyler H.D., Capitan A., Pausch H., Stothard P., Van Binsbergen, R., *et al*. (2014) *Nat Genet.***46**: 858.

Duncan W.C., Shaw J.L.V., Burgess S., McDonald S.E., Critchley H.O.D. and Horne A.W. (2011) *PLoS ONE* **6**: e23595.

Enomoto A., Ping J. and Takahashi M. (2006) *Ann. N Y Acad. Sci.* **1086**: 169.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., *et al*. (2012) *J. Dairy Sci.* **95:** 4114.

Garrick D.J., Taylor J.F. and Fernando R.L. (2009) *Genet. Sel. Evol.* **41:** 44.

Haile-Mariam M., Pryce J.E., Schrooten C. and Hayes B.J. (2015) *J. Dairy Sci*. **98**: 3443

Hardelin J.P., Levilliers J., del Castillo I., Cohen-Salmon M., Legouis R., *et al*. (1992) *Proc. Nat. Acad. Sci*. **89:** 8190.

Kemper K., Reich C., Bowman P., Vander Jagt C., Chamberlain A., *et al*. (2015) *Gen. Sel. Evol.* **47**: 29.

Lyu J., Jho E.-h. and Lu W. (2011) *Cell research* **21**: 911.

Matzuk M.M. and Lamb D.J. (2008) *Nat. Med.* **14:** 1197.

Skipper M., Dhand R. and Campbell P. (2012) *Nature* **489**:45.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., *et al*. (2010) *Nat. Gen.* **42**: 565.

Yoon Y.-S., Lee M.-W., Ryu D., Kim J.H., Ma H., *et al*. (2010) *Proc. Nat. Acad. Sci.* **107:** 17704.