

## GENOME WIDE ASSOCIATION STUDIES IN DAIRY CATTLE USING HIGH DENSITY SNP SCANS

H.W. Raadsma<sup>1</sup>, M.S. Khatkar<sup>1</sup>, G. Moser<sup>2</sup>, M. Hobbs<sup>1</sup>, R. Crump<sup>3</sup>, J.A.L. Cavanagh<sup>1</sup> and B.Tier<sup>3</sup>

The CRC for Innovative Dairy Products, <sup>1</sup> ReproGen, Faculty of Veterinary Science, The University of Sydney, Camden, Australia, <sup>2</sup> Bellbowrie, QLD, 4070, Australia, <sup>3</sup> Animal Breeding and Genetics Unit, UNE, Armidale, NSW, 2351, Australia

### SUMMARY

Use of high density Single Nucleotide Polymorphic (SNP) marker information allows for prediction of genetic merit via genome wide selection and for localization of markers in gene regions of biological interest through Genome Wide Association Studies (GWAS). We report on a replicated GWAS in dairy cattle using 1,945 progeny tested bulls genotyped with three high density SNP panels representing 63,678 informative SNP. Single SNP genotypes were analysed against deregressed EBV for protein percent and fat percent using a mixed linear model accounting for SNP and animal polygenic effects. The 127,356 analyses (63,678 informative SNP by two traits) across the two data sets identified 143 and 87 significant ( $P < 0.05$ , corrected for False Discovery Rate) associations for protein % in data set 1 and 2 respectively, whilst for fat % 102 and 61 significant associations were identified in the two data sets respectively. Outputs from selected SNP analyses are discussed for significance and pleiotropic effects and compared against integrated QTL meta-assembly from public domain studies

### INTRODUCTION

Development of high-density large-scale single nucleotide polymorphism (SNP) genotyping platforms (Hardenbol *et al.* 2005) has opened the possibility of large scale genomic investigations. Typically in livestock high density SNP typing has seen the transition from linkage and QTL mapping to the prediction of total genetic merit using genome wide selection or genomic selection approaches (Schaeffer 2006; Raadsma *et al.* 2008) where marker location is not required, or focus on localised SNP in genes to dissect the genetic architecture underlying complex traits through Genome Wide Association Studies (GWAS). Whilst in human over 220 GWA studies have reported on complex disease/performance traits (National Human Genome Research Institute catalogue of published genome-wide association studies at <http://www.genome.gov/26525384>) relatively few studies describing applications of GWAS in cattle have been reported to date (Daetwyler *et al.* 2008; Kolbehdari *et al.* 2008). In this study we report on a GWA study using two data sets with overlapping SNP data against two commonly recorded milk performance traits.

### MATERIALS AND METHODS

**Bulls and Genotyping.** DNA was extracted from 1,945 Australian progeny-tested Holstein Friesian dairy bulls. All bulls were sourced from Genetics Australia, and represented a cohort of sires used for ongoing commercial use (proven) or rejected (non selected) following progeny testing. The sires were born between 1955 and 2001, with >96% of sires born after 1980. All sires had Australian Breeding Value (ABV) data calculated by the Australian Dairy Herd Improvement Scheme (ADHIS) for traits associated with lactation performance, conformation, reproductive fitness and disease resistance. The bull samples were split into two data sets, the first consisting of 1,309 bulls genotyped with 15,036 SNPs (Khatkar *et al.* 2007) and the second data set of 634 bulls genotyped with 25K (Affymetrix ) and/or a 50K SNPs platform (Illumina). Across the two data

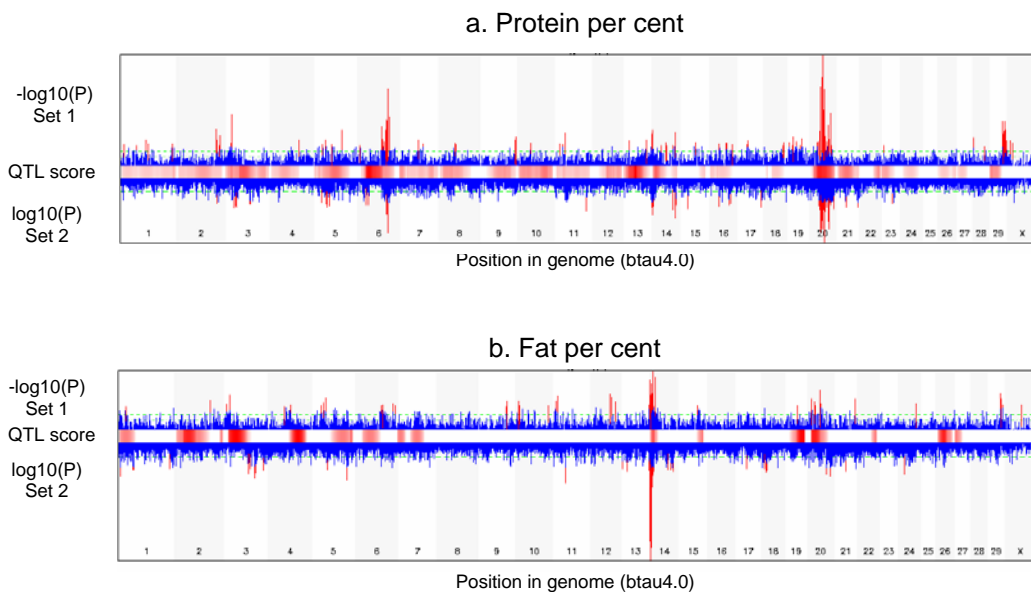
## Dairy Cattle

sets 8072 SNPs were in common. Locations of the SNPs were determined on the bovine sequence assembly Btau 4.0 (<ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20060815-freeze/>).

**Traits and Analyses.** Deregressed ABV data on protein percent (prot%) and fat percent (fat%) were selected for the GWAS. The deregressed Estimated Breeding Values (EBVs,  $\mathbf{y}$ ) were analysed against each SNP by fitting mixed linear models  $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$ , with SNP as fixed effect ( $\mathbf{Xb}$ ) and animal polygenic effect ( $\mathbf{Za}$ ) as random using ASReml (Gilmour et al., 2000) providing a nominal significance value and effect size for each SNP. All SNP effects were standardized by dividing the estimated effects by the standard deviation (SD) computed from the test sample for each trait. The effects and probabilities for each SNP were aligned in a genome browser against a spline of a QTL meta-analysis for all published QTLs based on an average score system adapted from Khatkar (2006), allowing for direct comparison with public domain studies. The browser allows selection of subsets of trait/data set combinations for internal comparison. The p-values were scaled to  $-\log_{10}(\text{p-value})$ . The false discovery rate (FDR) for the test of each marker were computed as q-value implemented in R package qvalue (Storey and Tibshirani 2003).

## RESULTS

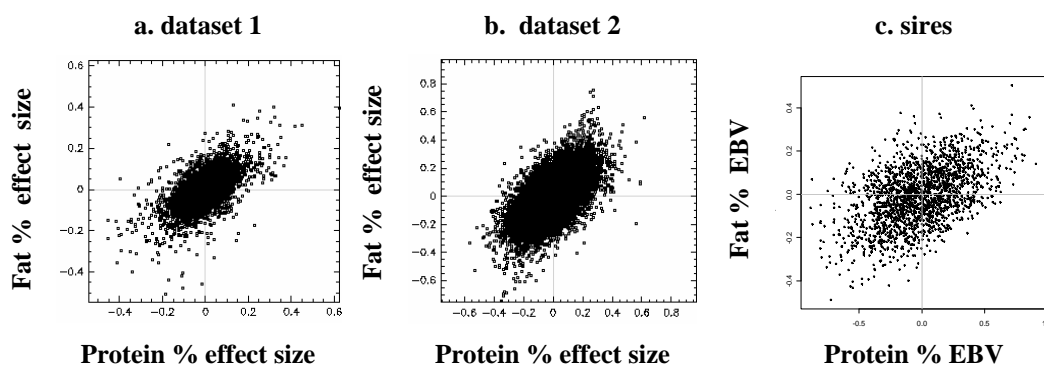
The 127,356 analyses (63,678 informative SNP by two traits) across the two data sets identified 143 and 87 significant  $P < 0.05$  FDR associations for prot% in data set 1 and 2 respectively, whilst for fat % 102 and 61 significant associations were identified in the two data sets respectively. The highly significant associations occurred across the genome in clusters of SNPs (Figure 1 a and b) some of which were strongly aligned with known QTL locations, whilst others identified novel regions for which no QTL have previously been reported.



**Figure 1. Distribution of SNP association  $-\log_{10}(P)$  along the genome for 2 data sets of bulls for (a) prot% and (b) fat% as well as alignment against QTL meta score from public domain analyses. SNP association in red exceed FDR significance threshold, green broken line  $P=0.01$ , and QTL meta score intensity increases as QTL score increases.**

The distribution of SNP effects showed a similar trend for both traits and both data sets and was strongly skewed. Only 10 and 56 SNPs showed an effect of greater than 0.5SD for prot% and fat% respectively. Similarly 10 % and 9.2 % of SNP showed a small (<0.01SD) or non existent effect for prot% and fat% respectively. The relationship of significance values between common SNPs measured in both data sets for the same trait was  $r = 0.04$  and  $r = 0.16$  based on Spearman rank correlations, for prot% and fat % respectively showing a high degree of SNP variability in detecting a significant relationships across the two data sets. However, the relationship of effect size between common SNPs measured in both data sets for the same trait was  $r = 0.28$  and  $r = 0.20$  for prot% and fat % respectively showing some degree of SNP repeatability in detecting a similar effect relationship across the two data sets.

For both data sets a strong correlation was observed between SNP showing an association for both prot% and fat%, ( $r = 0.25$  for P-values and 0.60 for effect size). Although the majority of SNP had a positive or negative pleiotrophic effect on prot% and fat%, some exceptions were evident with an antagonistic sign of effect (Figure 2a,b) Similarly a strong correlation ( $r = 0.50$ ) was observed between deregressed EBV for the 1945 sires for prot% with fat % (Figure 2c).



**Figure 2. Association between SNP effects for prot% and fat%, data set 1 (a), data set 2 (b) and (c) deregressed EBV solutions for prot% and fat% for the sires used in the association analysis.**

## **DISCUSSION AND CONCLUSIONS**

From the large number of single point associations conducted across the two data sets, only a relatively low number of SNP showed a highly significant association which exceeded the threshold for FDR (outlined in red in Figure 1). Furthermore most of the significant SNPs occurred in clusters most likely as a result of LD between closely spaced SNP and the underlying QTL or QTN. Using a replication of SNP association shows a negligible repeatability on significance probability across two data sets for individual SNPs suggesting that some of the significant SNP may still be spurious. Repeatability for effect size attributed to each of the SNP was higher possibly as a scale effect given the broader range for effect compared with P values and the extend of LD over larger segments containing several SNP. In line with expectations of likely gene effects, relatively few SNP were associated with very large effects (>0.5SD) as has been shown previously for QTL effects (Hayes and Goddard, 2001). Although it is attractive to select such SNP associations for further investigation, some caution is warranted since these effects may be over estimated and occur by chance alone. Some account of replication is essential.

The most obvious feature of a GWAS is the strong alignment of SNP to QTL which have been independently verified in previous studies, including some cases where an underlying QTN has

## Dairy Cattle

been identified. Such clusters were evident on BTA 20 for prot% and BTA 14 for fat%. ([http://www.vetsci.usyd.edu.au/reprogen/QTL\\_Map/](http://www.vetsci.usyd.edu.au/reprogen/QTL_Map/)). The cluster of SNPs on BTA 6 for prot% shows some departure in location from the widely known QTL cluster on BTA 6 for prot%. The GWAS also showed failure to detect significant SNP in known QTL regions on BTA3 and BTA10 for prot% and BTA 2,3,4,5 and 25 for fat% (Figure 1), possible reasons may be that no QTL are segregating at those locations in this population, or SNP density is too low to provide LD to underlying QTN although both reasons are unlikely and warrants further examination.

A key problem in applying of GWAS in livestock is the confounding of the estimates of SNP effects with pedigree, and the SNP may partly act as marker for detecting the relationship between individuals without having a large true effect. The BLUP procedure is likely to account for some of the unrelated polygenic effects vs. specific SNP effects, but the confounding is unlikely to be fully compensated for. However, using EBVs from progeny tested sires has the advantage that all polygenic differences between animals are captured with a high degree of accuracy. Finally although it is attractive to search for all large SNP effects independently and use these as marker panel for MAS, it is unlikely to be as efficient as genome wide selection/genomic selection procedures where the effects of all SNP are considered in predicting genetic merit. Such approaches have now shown utility in independent dairy sets (Raadsma *et al.* 2008); (Hayes *et al.* 2009). The down side from such prediction functions is a requirement that all SNPs be genotyped, unless stringent SNP selection procedures are applied which allow for subsets to be used without loss of predictive power. Alignment of SNP identified under GWAS and GS has not yet been reported.

## ACKNOWLEDGMENTS

The genotype data and phenotype was provided by the Co-operative Research Centre for Innovative Dairy Products.

## REFERENCES

- Gilmour, A.R., Thompson, R., Cullis, B.R., and Wellham, S.J. (2000) ASReml Reference Manual. Daetwyler, H.D., Schenkel, F.S., Sargolzaei, M. and Robinson, J.A. (2008) *J. Dairy Sci.* **91**:3225.
- Hardenbol, P., F. Yu, J. Belmont, J. Mackenzie, C. Bruckner, T. Brundage, A. Boudreau, S. Chow, J. Eberle, A. Erbilgin, M. Falkowski, R. Fitzgerald, S. Ghose, O. Iartchouk, M. Jain, G. Karlin-Neumann, X. Lu, X. Miao, B. Moore, M. Moorhead, E. Namsaraev, S. Pasternak, E. Prakash, K. Tran, Z. Wang, H. B. Jones, R. W. Davis, T. D. Willis and R. A. Gibbs, (2005) *Genome Res* **15**: 269.
- Hayes, B. and Goddard, M.E. (2001). *Genet. Sel. Evol.* **33**:209.
- Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. (2009) *J. Dairy Sci.* **92**:433.
- Khatkar, M.S. (2006) PhD Thesis, University of Sydney.
- Khatkar, M. S., Zenger, K.R, Hobbs, M., Hawken, R.J., Cavanagh, J.A.L., Barris, W., McClintock, A.E., McClintock, S., Thomson, P.C., Tier, B., F. Nicholas, F.W., and H. W. Raadsma, H.W. (2007). *Genetics* **176**:763.
- Kolbehdari, D., Z. Wang, J. R. Grant, B. Murdoch, A. Prasad, Z. Xiu, E. Marques, P. Stothard and S. S. Moore, (2008) *J Dairy Sci* **91**: 2844.
- Raadsma, H.W., Moser, G., Crump, R.E., Khatkar, M.S, Zenger, K., Cavanagh, J., and Tier, B. (2008) *Dev. Biologicals. (Basel)* **132**:219.
- Schaeffer, L.R. (2006) *J. Anim. Breed. Genet.* **123**:218.
- Storey, J.D. and Tibshirani, R. (2003) *Proc. Natl. Acad. Sci. U S A* **100**:9440.