Contributed paper

NARROWING THE SEARCH SPACE: PUTATIVE CAUSAL VARIANTS ARE ENRICHED IN ANNOTATED FUNCTIONAL REGIONS FROM 6 BOVINE TISSUES

C.P. Prowse-Wilkins[1,2], J. Wang[2], M.E. Goddard[1,2], R. Xiang[1,2], J.B. Garner[2]and A.J. Chamberlain[2]

[1] Faculty of Veterinary & Agricultural Science, The University of Melbourne, Parkville, Victoria, Australia

[2] Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria, Australia

SUMMARY

Identifying causal variants in the bovine genome is difficult as there are millions of variants. Work in humans shows that most variants affecting complex traits lie in non-coding functional regions. However, functional regions are generally species specific and not well annotated in non-model organisms. This project annotated functional regions directly in dairy cows using a laboratory technique called ChIP-seq (Chromatin Immunoprecipitation followed by sequencing).

We generated 86 functional datasets across 6 tissues from 3 lactating Holstein dairy cows. This represents millions of putative functional regions in the bovine genome including, for the first time, in the mammary gland of lactating dairy cows. These regions were highly enriched for putative causal variants (eg milk trait QTL and eQTL). The results represent the largest database of functional regions in the bovine genome to date and can be used to narrow the search space for causal variants and improve genomic predictions.

INTRODUCTION

Genomic prediction aims to predict the phenotypes of animals based on their genotypes. It does this by finding genotypes which associate with the phenotype in a training population. However, this association could be based on linkage disequilibrium (LD) and not a direct causal relationship between the trait and the genotype. This means the accuracy of genomic predictions can break down over time as LD breaks down and is not useful in breeds which have different LD to the training population. If we could use the genetic variant which is directly affecting the phenotype (the causal variant) in our predictions, this would not occur (Hayes *et al*. 2016).

Work in other species has found that causal variants are enriched in functional regions (Schaub *et al.* 2012). Until recently, these were not well annotated in the bovine genome (Fang *et al.* 2019). Functional regions can be identified with Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) to identify functional marks which pinpoint these regions in the genome. Examples of functional marks include histone modifications and transcription factors. Histone modifications are alterations to the histone proteins which DNA is wrapped around in the cell. Four histone modifications of interest are H3K4Me3-found at promoters, H3K4Me1-found at enhancers, H3K27ac-found in active regions and H3K27Me3-found in inactive regions (Kimura 2013). Another marker of interest is the binding site for the transcription factor CTCF which is found at insulators and other regions of importance (Kim *et al.* 2015). This study annotated these functional markers in 6 tissues (mammary, liver, kidney, spleen, lung and heart) in Holstein dairy cows and tested whether these regions are enriched for causal variants.

MATERIALS AND METHODS

Chromatin Immunoprecipitation and Sequencing. Heart, kidney, liver, lung, mammary gland, and spleen were sampled from 3 Holstein dairy cows post-mortem and snap frozen in liquid nitrogen before being stored at -80°C until use. At sampling animals were at 5th, 7th, and 1st parity and 208, 173 and 65 days of lactation respectively. Ethics approval for 2 of the cows were obtained

from Department of Jobs, Precincts and Regions Ethics Committee (Application No. 2014-23). The 3rd cow was not euthanised for this study but culled as a result of injury. Frozen tissue was ground for 3 minutes in the Geno/Grinder (SPEX SamplePrep) and fixed for 10 minutes with 10% formaldehyde. Chromatin was prepared using the Magnify Chromatin Immunoprecipitation kit (ThermoFisher) as per the manufacturer's instructions. Fixed chromatin was sheared to 200-500bp using the Covaris S2 (Covaris) for three minutes, duty cycle five, % intensity four and 200 cycles per burst. Chromatin immunoprecipitation was performed using the Magnify Chromatin immunoprecipitation kit (ThermoFisher) with some modifications. Sheared chromatin was immunoprecipitated with 0.25-0.5µg of antibody for the histone modifications (H3K4Me3, H3k4Me1, H3K27ac and H3K27Me3) or 10µl of antibody for CTCF. Sequence libraries were prepared for each ChIP sample and a control for each chromatin preparation (input sample) using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) as per the manufacturer's instructions and run on the Hiseq 3000 (Illumina) in a 150 cycle paired end run. Each library was sequenced with 20-300 million reads. Raw sequence reads were trimmed of adapters and poor-quality bases at the ends (quality less than 20) using Trimmomatic (Bolger *et al.* 2014). Trimmed reads with length less than 50 were removed. Trimmed reads were mapped to UMD3.1 bovine genome using BWA mem with default settings (Li 2013). Poor-quality reads with q>15 were removed with Samtools (Li *et al.* 2009) and marked duplicate reads were also removed. MACS2 with default settings was used to call peaks from mapped ChIP reads with input reads as control (Zhang *et al.* 2008). The quality of peaks was checked with deepTools plotFingerprint (Ramirez *et al.* 2016) and SPP (Kharchenko *et al.* 2008).

Enrichment of Causal SNP in Functional Regions. Enrichment of putative causal SNP in functional regions was calculated using the formula described in (Ernst & Kellis 2010) as outlined below. A variety of SNP datasets were used as putative causal SNP (Table 1). Statistical significance of enrichment or depletion was calculated in R using a hypergeometric test.

Enrichment=(C/A)/(B/D) where:

A= number of positions under peaks

B=number of positions under peaks and also a putative causal SNP

C=number of positions that were putative causal SNP

D=number of positions in the genome

RESULTS AND DISCUSSION

In total we sequenced 86 ChIP-seq samples, with three biological replicates in 6 tissues assayed for 5 marks (four samples were excluded due to low quality). There was an average of 480,000 peaks per sample covering an average of 13% of the genome. All samples were high quality. These data represent millions of putative functional regions in the bovine genome.

Peaks were significantly enriched for putative causal variants ($P<0.001$) as expected (Table 2). The QTL for milk traits were particularly strongly enriched within peaks and particularly enriched within peaks found in the mammary gland (Figure 1). This is consistent with studies in other species which show that trait QTL are particularly enriched within histone markers specific to tissues relevant to the trait (Trynka *et al.* 2013). The 80k SNP dataset was the least enriched although these were still significantly enriched within peaks. It is possible that this is because these SNPs are contributing to multiple traits which may not be relevant to the tissues represented in this study.

CONCLUSION

This work substantially increases the number of putative functional regions found in different tissues in the bovine genome, including the mammary gland of lactating dairy cows. As seen in other species, these regions are substantially enriched for putative causal variants for important traits suggesting SNP within these regions should be prioritised for genomic selection.

Table 1. Details of putative causal SNP tested for enrichment within functional regions

Dataset	Number of SNP	Description	Reference
Allele specific eQTL	1,100,446	Allele specific expression QTL from white blood cells and milk cells in 112 holstein cows (*P*<1e-4)	(Chamberlain *et al.* 2018)
Exon eQTL	945,832	Exon expression QTL from white blood cells, milk cells, liver and muscle in 209 holstein cows (*P*<1e-4)	(Xiang *et al.* 2018, Xiang *et al.* 2019)
Gene eQTL	110,200	Gene expression QTL from white blood cells, milk cells, liver and muscle in 209 holstein cows (*P*<1e-4)	(Xiang *et al.* 2018, Xiang *et al.* 2019)
Conserved regions	378,472	SNP conserved in 100 species lifted over from human to bovine genome	(Xiang *et al* 2019)
SNP 80k	83,454	Top 80,000 sequence variants ranked for their contributions to 34 traits	(Xiang *et al.* 2021)
Splice QTL	1,112,324	Splice QTL from blood, milk cells, liver and muscle in 209 holstein cows (*P*<1e-4)	(Xiang *et al.* 2018, Xiang *et al.* 2019)
QTL Protein Yield	3,317	GWAS in 32347 cows for protein yield with *P*<1e-7	(Xiang *et al.* 2020*)*
QTL Fat yield	4,815	GWAS in 32347 cows for fat yield with *P*<1e-7	Xiang *et al.* 2020*)*
QTL Milk Yield	6,883	GWAS in 32347 cows for milk yield with *P*<1e-7	Xiang *et al.* 2020*)*
QTL Fat percentage	12,373	GWAS in 32347 cows for fat percentage with *P*<1e-7	Xiang *et al.* 2020*)*
QTL Protein percentage	17,012	GWAS in 32347 cows for protein percentage with *P*<1e-7	Xiang *et al.* 2020*)*

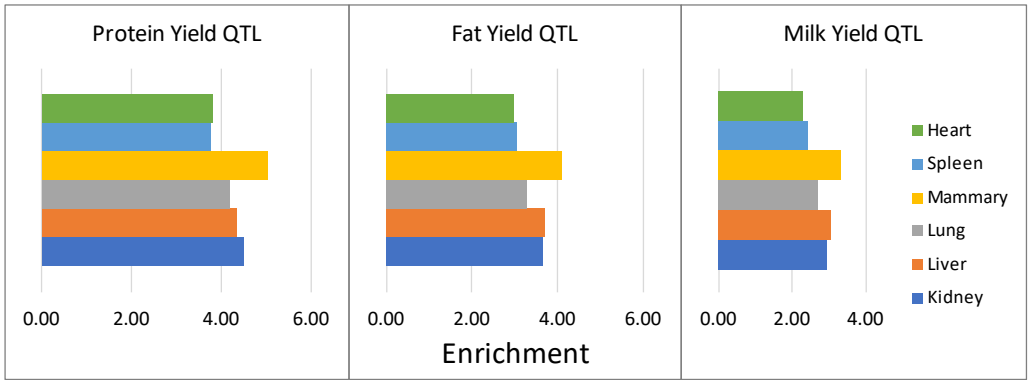

Figure 1. Enrichment of 3 sets of milk trait QTL within H3K27ac peaks. Peaks in mammary gland have the highest enrichment for these milk trait QTL

ACKNOWLEDGEMENTS

The authors would like to thank the Agriculture Victoria staff at Ellinbank and Bundoora who helped with sample collection. We also thank DataGene for access to data used in this study.

Table 2. Enrichment of causal SNP in ChIP-seq peaks. Enrichment of each SNP dataset within each histone modification or CTCF averaged across tissues

	H3K4Me3	H3K27ac	CTCF	H3K4Me1	H3K27Me3
Allele specific eQTL	1.86	1.96	1.93	1.76	1.69
Exon eQTL	1.68	2.21	1.73	1.61	1.33
Gene eQTL	2.24	2.37	2.27	1.97	1.82
Conserved regions	1.66	1.46	1.42	1.21	1.14
SNP 80k	1.20	1.16	1.18	1.16	1.15
Splice QTL	1.70	1.77	1.75	1.63	1.58
QTL Protein Yield	4.46	4.27	4.06	3.21	2.93
QTL Fat yield	3.72	3.46	3.43	2.82	2.60
QTL Milk Yield	3.09	2.79	2.85	2.35	2.24
QTL Fat percentage	2.78	2.51	2.58	2.19	2.16
QTL Protein percentage	1.85	1.91	1.80	1.58	1.40

REFERENCES

Bolger A.M., Lohse M. and Usadel B.J.B. (2014) *Bioinformatics* **30**: 2114.

Chamberlain A., Hayes B., Xiang, R., Vander Jagt C., Reich C., Macleod I., Prowse-Wilkins C., Mason B., Daetwyler H. and Goddard M. (2018) *Proc. Of. WCGALP*. **Vol Mol Gen 1**: 254

Ernst J & Kellis M. (2010) *Nature Biotech.* **28**: 817.

Fang L., Liu M., Kang X., Lin S., Li B., Connor E. E., Baldwin R.L., Tenesa A., Ma L., Liu G.E. and Li C. (2019) *BMC Biology* **17**: 68.

Hayes B., Chamberlain A., Daetwyler H., VanderJagt C. & Goddard M. (2016) *J.An. Sci.* **94**: 3.

Kharchenko P. V., Tolstorukov M. Y. and Park P. (2008) *Nature Biotech.* **26**: 1351.

Kim S., Yu N.-K. & Kaang B.-K. (2015) *Exp. & Mol. Med.* **47**: e166.

Kimura H. (2013) *J. Hum. Gen.* **58**: 439.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. and Durbin R. (2009) *Bioinformatics* **25**: 2078.

Li H. (2013) *arXiv*. 1303.3997.

Ramirez F., Ryan D.P., Gruning B., Bhardwaj V., Kilpert F., Richter A. S., Heyne S., Dundar F. and Manke T. (2016) *Nucleic Acids Res.* **44**:W160.

Schaub M.A., Boyle A.P., Kundaje A., Batzoglou S. and Snyder M. (2012) *Gen. Res.* **22**:1748.

Trynka G., Sandor C., Han B., Xu H., Stranger B.E., Liu X.S. and Raychaudhuri S. (2013) *Nature Genetics.* **45**:124.

Xiang R., Hayes B.J., Vander Jagt C.J., Macleod I.M., Khansefid M., Bowman P.J., Yuan Z., Prowse-Wilkins C.P., Reich C.M. and Mason B.A. (2018) *BMC Genomics.* **19**:521.

Xiang R., Van Den Berg I., Macleod I.M., Hayes B.J., Prowse-Wilkins C.P., Wang M., Bolormaa S., Liu Z., Rochfort S.J and Reich C.M. (2019). *PNAS.* **116**: 19398.

Xiang R., Van Den Berg I., Macleod I.M., Daetwyler H.D. and Goddard M.E. (2020). *Comm. Bio.* **3**: 1.

Xiang R., Macleod I.M., Daetwyler H.D., de Jong G., O'Connor E., Schrooten C., Chamberlain A.J. and Goddard M.E. (2021) *Nature Comms.* **12**: 860.

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown, M. and Li W. (2008) *Genome Biology.* **9**: 1.