DETECTING VARIANTS ASSOCIATED WITH COMPLEX TRAITS THROUGH CHANGING GENE EXPRESSION IN CATTLE

M. Khansefid[1,2,3], J.E. Pryce[2,3,4], S. Bolormaa[2], Y. Chen[5] and M.E. Goddard[1,2,3]

[1] Department of Agriculture and Food Systems, The University of Melbourne, VIC, Australia
[2] Department of Economic Development, Jobs, Transport and Resources, VIC, Australia
[3] Dairy Futures Cooperative Research Centre (CRC), VIC, Australia
[4] La Trobe University, VIC, Australia
[5] Department of Primary Industries, Elizabeth Macarthur Agricultural Institute, NSW, Australia

SUMMARY

Mutations can affect phenotypes by changing the amino acid sequence in a protein or by changing the expression of a gene. The gene expression in a particular tissue can be measured using mRNA sequencing and counting the number of mRNA copies of the gene. The aim of our research was to find mutations which affect expression of genes in cis. We detected the mutations that are associated with the expression of each gene in muscle (45 Angus bulls) and liver (38 Angus bulls) by correlating the mRNA count with the alleles carried at single nucleotide polymorphisms (SNPs) within 50kb of the genes that were tested. Furthermore, the SNPs and genes with at least one SNP significantly associated with one or more traits ($p < 0.001$) were found by genome-wide association studies (GWAS) in a beef cattle dataset including 6,114 genotyped animals with 20 traits recorded. We compared the results to find the SNPs significantly associated with gene expression ($p < 0.001$) and also with the variation in phenotype. The SNPs which were significantly associated with gene expression ($p < 0.001$) were more likely to be significant in GWAS for concentration of Insulin like growth factor1(IGF-1), residual feed intake (RFI) and in a multi-trait significance test.

INTRODUCTION

The mutations underlying variation in complex traits are generally identified using genotypes and phenotypes in GWAS (Goddard and Hayes 2009). However, most of the SNPs found in GWAS are not located in the coding regions of the genome and so may influence the variation of the traits by changing the expression of genes. These polymorphic sites are likely to affect the expression of the copy of the gene on the same chromosome and are called cis-expression quantitative trait loci (cis-eQTL) (Arnold *et al.* 2012).

The next generation sequencing (NGS) technology which is used for sequencing DNA is also applicable for sequencing mRNA. In genome-wide transcriptome studies using the mRNA sequencing (RNA-seq) data, the abundance of mRNA for each gene can be measured by counting the number of sequence reads aligned to the gene in the reference genome.

Although GWAS have found many associations between SNPs and traits, it has proved difficult to identify the causal mutation or the gene(s) whose structure or expression they affect. On the other hand, a cis-eQTL affects the expression of a specified gene, so if a cis-eQTL is the same as a QTL for a traditional phenotype, this defines the gene through which the QTL probably acts. In addition, the cis-eQTL may help to identify the causal mutation for both itself and the traditional QTL. However, the RNA-seq data is more expensive to obtain than most traditional phenotypes and therefore the number of animals in whole transcriptome studies is limited. Also, the results depend on the tissue which was sampled for mRNA and so, even if a QTL is identical with a cis-eQTL, this may not be discovered if the wrong tissue is sampled.

The aim of our research was to find if SNPs which were significantly associated with gene expression (cis-eQTL) were more likely to be significant in GWAS of traditional traits.

MATERIALS AND METHODS

Animals. The animals used for RNA-seq were young Angus bulls from lines of cattle divergently selected for residual feed intake (RFI) from the selection lines established in 1993 at the Agricultural Research Centre, Trangie, NSW, Australia (Arthur *et al.* 2001). The bulls with liver (L-bulls) and muscle (M-bulls) samples were from the lines after approximately 3 and 4 generations of selection, respectively.

RNA-seq. The transcriptome data was for 43 muscle and 38 liver samples. The extracted RNA from the sampled tissues (Chen *et al.* 2011) were sequenced with a HiSeq 2000 (Illumina Inc) after mRNA enrichment, by the modified protocol of Illumina sample preparation for RNA-seq. All of the raw reads were passed through quality control filters and trimming the ends of reads based on Phred quality scores (minimum quality of each nucleotide base=15, minimum average quality of the read after trimming=20, minimum read length after trimming=50 and maximum consecutive nucleotide bases with poor quality=3).

Genotypes. For 43 M-bulls we had 800K SNP chip and whole genome sequence data (WGS) with average coverage 6.7 fold from the 1000 bull genomes project (Daetwyler *et al.* 2014). The L-bulls had 800K SNP genotypes imputed from 50K (Illumina BovineSNP50K chip) using BEAGLE (Browning and Browning 2009) and then imputed to WGS using FImpute (Sargolzaei *et al.* 2014). We also used FImpute to rephase the genotypes of M and L bulls.

Alignment. The bovine genome assembly UMD3.1 was modified using the WGS data to produce a genome sequence for each individual bull. For each bull the RNA-seq data were aligned to its customised reference genome using TopHat2 (Kim *et al.* 2013).

Abundance of genes. For each gene in the reference genome, the mapped mRNA was counted using HTSeq python package (Anders *et al.* 2015). The number of sequence reads for all genes in each animal were normalised with a weighted trimmed mean of the log expression ratios using edgeR package in R (Robinson and Oshlack 2010). Finally, the normalised gene counts were log transformed to have normal distributions across animals.

eQTL mapping. The genes expressed in more than 25% of L-bulls and M-bulls were used to find eQTL. The association between the gene counts and the SNPs in WGS data within 50kb of the gene was calculated with ASReml for muscle and liver samples separately.

GWAS. A GWAS was carried out for 20 traits (including meat quality and production traits) using up to 6114 cattle and 729,068 HD SNPs. As well as individual traits, a multi-trait test was performed as described by Bolormaa et al (2014), except that only *Bos taurus* cattle were included. The SNPs were from the Illumina high density panel and were either genotyped or imputed from lower density. Only SNPs within 50 kb of a gene were used so that the same SNPs were tested in the GWAS as were tested for cis-eQTL.

The SNPs were classified as significant or not for association with gene expression ($p < 0.001$) and for association with one of the traits ($p < 0.001$) and we performed a chi-squared test of the hypothesis that SNPs affecting gene expression are more likely to be significant in GWAS. We also classified genes as either containing a significant cis-eQTL or not and as having a SNP within 50kb associated with a trait or not, and performed a chi-square test to test the hypothesis that genes containing an eQTL were also likely to be near a SNP associated with a traditional phenotype.

RESULTS AND DISCUSSION

RNAseq. In the muscle samples, there were on average about 8.5×10^6 (100%) RNA-seq raw reads per animal, 6.5×10^6 (75%) reads that passed the quality control filters, 5.9×10^6 (70%) that aligned to the reference genome and 5.5×10^6 (65%) that were mapped uniquely. In the liver samples, there were on average roughly 7.6×10^6 (100%) raw reads per animal, 5.5×10^6 (72%) reads that passed the quality control filters, 4.6×10^6 (60%) that aligned to the genome and 4.5×10^6 (52%) that were mapped uniquely. The percentage of reads mapped to the reference genome in L-

bulls is less than M-bulls probably because we used imputed genotypes to enhance the L-bulls reference genomes. In all chromosomes (and autosomes), 12,278 (11,842) in muscle and 12,233 (11,821) genes in liver were expressed (about 50% of the known genes were expressed in muscle and liver).

cis-eQTL.. In the muscle samples, each of the expressed 12,278 genes were tested for cis-eQTL using SNPs within 50kb of the gene. Among the HD SNPs, there were 240,818 SNPs that were tested for association with expression of one or more genes. 5,042 of these SNPs were significantly ($p < 0.001$) associated with expression of at least one gene in muscle. Similarly, of 227,488 SNPs tested, 2,420 were associated ($p < 0.001$) with expression of at least one gene in liver (Table 1).

Trait QTL. Table 1 presents results for two individual traits (blood concentration of IGF1 and residual feed intake) and the multi-trait test. For instance, 1,047 SNPs, out of 240,818 tested, were significantly associated with RFI (The number of SNPs tested varies slightly between M and L bulls and by trait because some SNPs had to be dropped from some analyses because they had a MAF below 1% in that dataset and only SNPs near genes expressed in that tissue were used).

Overlap between trait QTL and eQTL. There were only 3 SNP in common between the 386 SNPs that were associated with blood concentration of IGF-1 and the 5,041 associated with expression of at least one gene in muscle. This is not more than expected by chance (p=0.07) (Table 1). However, as shown in Table 1, there was more overlap between SNPs associated with traits and gene expression than expected by chance ($p < 0.05$) in 3 of the 6 tests. For instance, there were 169 SNPs associated with both the multi-trait test and with gene expression in muscle and this was far more than expected by chance (p=8.2×10^{-9}).

Where a SNP is significantly associated with a trait and with expression of a gene, the expression of this gene may also be affecting the trait. Genes identified in this way (and the number of significant SNPs associated with their expression) for the cis-eQTL in muscle and affecting IGF1 are: *PPM1H* (2), *MTHFD1* (1). For muscle cis-eQTL and RFI: *POLR2I* (33), *PTPRR* (5), *ATP5E* (4), *SSFA2* (2), *CARD6* (2), *THAP8* (1), *DNER* (1), *ATPIF1* (1) and in cis-eQTL in liver and IGF-1: *DAB1* (2), *EVC* (1), RFI: *MVK* (9), *GSK3A* (4), *LOXL3* (1), *DPYD* (1), *DNER* (1), and 2 SNPs were in an uncharacterized gene. For example, *GSK3A* (*glycogen synthase kinase 3 alpha*) was reported to regulate glycogen metabolism in liver (Ali *et al.* 2001). So it seems reasonable that this gene can affect feed efficiency and RFI. The number of SNPs found significant in eQTL mapping and in the multi-trait GWAS was 169 in muscle and 48 in liver and these SNPs were spread across 51 genes in muscle and 25 in liver.

Table 1 also contains the corresponding numbers based on testing genes instead of SNPs. 4,044 genes had a cis-eQTL, 303 genes had a SNP within 50kb associated with IGF-1 and 125 genes had both which was significantly more than expected by chance (p=0.002). Note that the SNPs associated with IGF-1 are not necessarily the same as the SNPs associated with gene expression, they are just near the same gene. In addition, the eQTL analysis used whole genome sequence whereas the QTL analysis used HD SNPs.

In all 6 tests, there were more genes that contained both a QTL and an eQTL than expected by chance (Table 1). Examples of genes that contain a QTL and an eQTL are: in muscle, *SH3-domain GRB2-like (endophilin)-interacting protein 1* (*SGIP1*) gene was previously reported to have a role in regulating food intake, fat mass, energy balance and energy homeostasis. Its roles in regulation of feeding behaviour affects any process that activates or increases the frequency, rate or extent of feeding behaviour (Trevaskis *et al.* 2005; Cummings *et al.* 2012) and therefore might affect RFI. *Bos taurus bone morphogenetic protein 2 (BMP2)* is one of the genes found to significantly associated with IGF-1 concentration and differentially expressed in muscle. BMPs, are also called growth and differentiation factors have negative regulation of the IGF receptor signalling pathway and affect any process that stops, prevents, or reduces the frequency, rate or extent of IGF receptor

signalling (Kronenberg 2003).

We conclude that traditional QTL are sometimes in fact cis-eQTL. Therefore mapping cis-eQTL will help us to identify causal variants for conventional phenotypes and the genes through which these variants act. A benefit of eQTL is that the gene whose expression they affect is known so, if the QTL is an eQTL, this identifies the gene through which the QTL acts. cis-eQTL often explain a large proportion of the variance in expression and so there is some power to identify the causal variant even in small datasets. In addition, as information builds up about regulatory regions in livestock genomes, we will have functional information to help us identify sites that might change the expression of the target gene.

Table 1. SNPs and genes association with traits variation and gene expression in muscle and liver

Tissue / Trait	**Total SNPs (genes)**	**Number of SNPs (genes) significantly associated with:**			
		QTL	**eQTL**	**QTL & eQTL**	**X^2 p-value for SNPs (genes)**
Muscle					
IGF-1	240,586 (12,278)	386 (303)	5,041 (4,044)	3 (125)	0.070 (0.002)
RFI	240,818 (12,278)	1,047 (502)	5,042 (4,044)	49 (204)	4.6×10^{-09} (1.8×10^{-04})
Multi-Trait	239,726 (11,842)	5,240 (3,102)	5,030 (3,923)	169 (1,099)	8.2×10^{-09} (0.002)
Liver					
IGF-1	227,473 (12,233)	366 (287)	2,420 (2,246)	3 (85)	0.649 (6.2×10^{-07})
RFI	227,488 (12,233)	985 (497)	2,420 (2,246)	18 (113)	0.019 (0.010)
Multi-Trait	226,564 (11,821)	4,907 (2,973)	2,413 (2,202)	48 (630)	0.550 (3.3×10^{-05})

REFERENCES

Ali A., Hoeflich K.P., and Woodgett J.R. (2001) *Chem. Rev.* **101**: 2527.

Anders S., Pyl P.T. and Huber W. (2015) *Bioinformatics* **31**:166.

Arnold M., Ellwanger D.C., Hartsperger M.L., Pfeufer A. and Stümpflen V. (2012) *PLoS One* **7**:e36694.

Arthur P.F., Archer J.A., Johnston D.J., Herd R.M., Richardson E.C., *et al.* (2001) *J. Anim. Sci.* **79**:2805.

Bolormaa S., Pryce J. E., Reverter A., Zhang Y., Barendse W., *et al.* (2014). *PLoS Genet.* **10**: e1004198.

Browning B.L. and Browning S.R. (2009) *Am. J. Hum. Genet.* **84**:210.

Chen Y., Gondro C., Quinn K., Herd R.M., Parnell P.F., *et al.* (2011) *Anim Genet* **42**:475.

Cummings N., Shields K.A., Curran J.E., Bozaoglu K., Trevaskis J., *et al.* (2012) *Int. J. Obes.* **36**:201.

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

Goddard M.E. and Hayes, B.J. (2009) *Nature Rev. Genet.* **10**: 381.

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., *et al.* (2013) *Genome Biol* **14**:R36.

Kronenberg H.M. (2003) *Nature* **423**:332

Robinson M.D. and Oshlack A. (2010) *Genome Biol* **11**:R25.

Sargolzaei M., Chesnais J.P. and Schenkel F.S. (2014) *BMC Genomics* **15**:478.

Trevaskis J., Walder K., Foletta V., Kerr-Bayles L., McMillan J., *et al.* (2005) *Endocrinology* **146**:3757.