Proc. Assoc. Advmt. Breed. Genet. 21: 477-480

ALLELE SPECIFIC EXPRESSION IS PERVASIVE IN CATTLE

C.J. Vander Jagt[1,4*], A.J. Chamberlain[1,4*], B.J. Hayes[1,2,4], L.C. Marett[1], M.E. Goddard[1,3,4]

[1]Department of Economic Development, Jobs, Transport and Resources, Victoria, Australia
[2]Biosciences Research Centre, La Trobe University, Victoria, Australia
[3]Faculty of Veterinary & Agricultural Science, University of Melbourne, Victoria, Australia
[4]Dairy Futures CRC, Victoria, Australia
[*]These authors contributed equally to this work

SUMMARY

Gene expression can be regarded as a complex trait phenotype, affected by a number of mechanisms, including *cis*-regulatory genetic variation. Allele specific expression (ASE) analysis can be used to determine the importance of *cis*-regulatory variation. In this study, using RNAseq data mapped to parental reference genomes, we analyse the ASE patterns of 17 tissue types and white blood cells (WBC) taken from a single lactating dairy cow. We found that 76% of all heterozygous single nucleotide polymorphisms (SNPs) tested (total 25,251) had significant ($p<0.01$) ASE in at least one tissue type and of all tested genes containing more than 1 tested SNP (7,985), 74% contained greater than 1 ASE SNP. However, there is a large variation between tissues in which genes contain SNP displaying ASE. We conclude that ASE is pervasive in cattle. Identification of these ASE SNP will aid in the detection of *cis*-regulatory variants responsible for phenotypic variation in bovine production traits, which in turn, may lead to improved selection of animals.

INTRODUCTION

Detection of ASE depends on the ability to differentiate the gene product of one parental chromosome from that of the other, and then to quantitate the relative amounts of each gene product. Using RNAseq data, this can be achieved by examining the imbalance of parental alleles expressed at heterozygous SNP (Pastinen 2010). When only one parental allele is expressed at a known heterozygous SNP, it may be indicative of gene imprinting. ASE complements the more traditional expression quantitative trait loci (eQTL) data, narrowing genomic regions of interest and has been successful in helping pin-point causative variants (Ge *et al.* 2009; Montgomery *et al.* 2010; Pickrell *et al.* 2010). The variants used to measure ASE from RNAseq data are within transcribed regions, nevertheless, identification of those ASE SNP in mRNA may serve as markers for the existence of causal regulatory variants close by. It is the identification of these causal regulatory variants affecting quantitative traits in livestock species that are of most interest, as a subset of these mutations could affect traits in the breeding goals for these species.

In this paper we present the results of an allele specific expression analysis of 17 tissues and WBC taken from a lactating Australian Holstein cow at a single point in time. This cow and her sire were sequenced as part of the 1000 bull genomes project and thus phased genotypes of all her heterozygous variants were available to create parental genomes. Alignment to parental genomes is considered the most accurate mapping method and least likely to result in mapping bias (Degner *et al.* 2009). Results of this study indicate pervasive ASE in bovine and large variation between tissues in which genes display ASE.

METHODS

100 base paired end RNA-seq reads were generated on an Illumina HiSeq2000 from 17 different tissues and WBC (in triplicate - see Table 1, column 1 for tissue types) taken from a single lactating Australian Holstein cow (25 months old, 65 days into first lactation). Reads per

tissue ranged from 40 to 100 million. Maternal and paternal reference genomes were created by editing UMD3.1 bovine genome assembly at all heterozygous variant sites from this cow using phased genotypes from 1000 bull genomes run 3 (Daetwyler *et al.* 2014). Paired RNA reads for each tissue replicate were aligned twice, once to each parental reference genome, using TopHat2 (Kim *et al.* 2013) and Ensembl release 75 genome annotation, allowing for two mismatches. Alignment files for each tissue replicate were merged, sorted and indexed using SAMtools (Li *et al.* 2009). Maternal and paternal allele counts for known heterozygous SNP for this cow (Daetwyler *et al.* 2014) were extracted using SAMtools mpileup (version 0.1.14). SNP were then filtered to only consider those falling within gene exon boundaries, with a minimum read depth of 10 in both parental reference alignments and the most abundant allele in both the maternal and paternal alignments had to agree (removing SNP falling in regions with obvious mapping bias). SNP were considered as having significant ($p<0.01$) ASE using the following Chi-squared (χ^2) test:

$$\chi^2 = \frac{\left(\frac{\left(r_m a_p - a_m r_p\right)^2 N}{ramp}\right)}{2}$$

where r was the count of reference alleles aligned to both parental genomes, a was the count of alternate alleles aligned to both parental genomes, m was the count of reference and alternate alleles aligned to the maternal genome, p was the count of reference and alternate alleles aligned to the paternal genome, r_m was the count of reference alleles aligned to the maternal genome, r_p was the count of reference alleles aligned to the paternal genome, a_m was the count of alternate alleles aligned to the maternal genome, a_p was the count of alternate alleles aligned to the paternal genome and N was the total number of alleles aligned to both parental genomes. Chi-squared values were divided by 2 to account for the value of N being derived from the counts of both parental haplotypes.

RESULTS AND DISCUSSION

Figure 1 demonstrates that there is little bias toward the reference allele for all SNP tested, with reference allele frequency normally distributed and centred at 0.5, indicating that our strategy of mapping reads separately to parental genomes was largely successful. As will be discussed later, lung has a large number of ASE SNP. Figure 1 also reveals a large number of SNP have extreme ASE (peaks at 0 and 1), however there is some bias in the SNP that display a reference allele frequency of 1. We believe this is due to errors in the whole genome sequencing of this cow, resulting in SNP called heterozygous when in fact the cow is homozygous at that position.

In total 25,251 SNP were tested for ASE in at least one tissue, and these SNP fell within 7,985 annotated genes. 89% of genes tested had significant ASE in at least one tissue (Table 1). Wang et al. (2014) state genes that have multiple SNP supporting ASE have a higher rate of successful verification. Therefore, we also tested the proportion of genes with >1 SNP with significant ASE where the gene had >1 SNP tested, this was 74% (Table 1). These results suggest that between 74-89% of genes show ASE in at least one tissue. This estimate is higher than the majority of published mouse and human literature of 4-53% (Yan *et al.* 2002; Bray *et al.* 2003; Pant *et al.* 2006; Serre *et al.* 2008; Vidal *et al.* 2011; Gao *et al.* 2012; MacEachern *et al.* 2012), though it must be acknowledged that these estimates are for single or few tissues or cell lines.

For individual tissues, the proportion of genes showing significant ASE varied from as low as 8-16% of those tested in thymus, to as high as 71-82% tested in lung. Pant et al. (2006) previously reported that 53% of genes tested showed significant ASE in a study limited to testing only 1,389 genes in WBC, our estimate of 21-33% of genes tested in WBC was lower, however we tested

more genes (4,680). Gao et al (Gao *et al.* 2012) reported 30% of the 8,779 genes tested in human mammary epithelial cells lines showing significant ASE, this also corresponds well to 16-31% of the 3,566 genes tested in our study showing significant ASE in mammary gland. The result of 71-82% of genes tested showing significant ASE in lung seems high, however to our knowledge this is the first time ASE has been tested in lung. Our estimate of 14-25% and 14-26% of the 5,462 and 5,946 genes tested in brain caudal lobe and brain cerebellum respectively are much lower than the estimate of 89% by Crowley et al (Crowley *et al.* 2015) in whole mouse brain, however they had an extremely powerful design testing greater than 12,000 genes in 96 mice from all possible pairwise crosses between the three divergent inbred lines. The power of our study comes from testing many tissues. Interestingly, a recent study undertaken by the GTEx Consortium (2015), compared between-sample and between-tissue sharing of ASE in humans with overall similarity of gene expression. They found that gene expression levels were determined by tissue, and that individuals clustered by tissue. However, allelic ratios have a higher correlation among tissues from the same individuals than among individuals for the same tissue. This suggests that ASE is primarily determined by the individual's genome. Therefore we have likely underestimated the total number of genes displaying ASE in the cattle population, and that further testing in more individuals (currently underway) will uncover more genes that undergo cis-regulation.

This study demonstrates that ASE is pervasive in cattle, supporting the argument by Pai et al. (2015) that much of the variation seen in quantitative traits is likely due to these changes in expression, and that these genes are under *cis*-regulatory control. Attention must now turn to the identification of these cis-regulatory variants. The identification of causative regulatory variants could then be used in livestock genomic selection programs leading to more accurate genomic breeding values and increases in the rate of genetic gain for economically important traits.

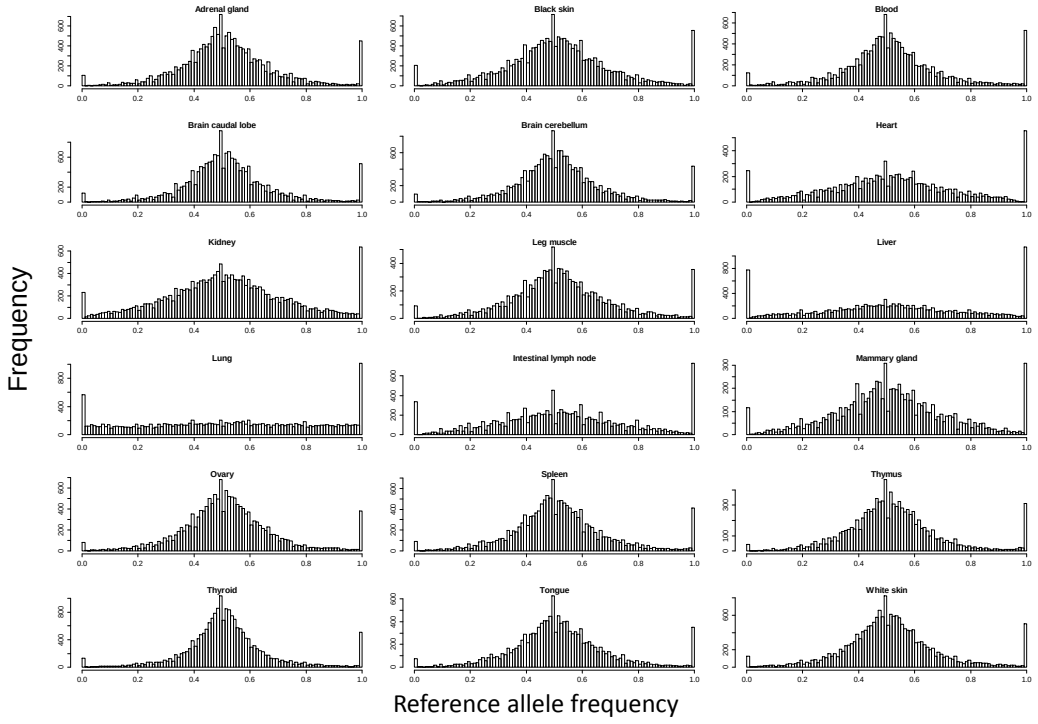

Figure 1. Reference allele frequency distributions for each tissue and WBC.

Table 1. Allele specific expression analysis results

Tissue	# SNP tested	# ASE SNP (% tested)	# Genes tested	# Genes w/ >1 SNP tested	# Genes w/ ASE SNP (% tested)	# Genes w/ >1 ASE SNP (% tested)
Adrenal	14,698	2,636 (18%)	5,462	3134	1,635 (30%)	536 (17%)
Brain caudal lobe	16,594	2,419 (15%)	5,946	3483	1,478 (25%)	494 (14%)
Brain cerebellum	15,460	2,324 (15%)	5,650	3269	1,470 (26%)	466 (14%)
Heart	9,545	2,919 (31%)	3,999	2118	1,869 (47%)	618 (29%)
Intestinal lymph	11,719	3,554 (30%)	4,684	2542	2,391 (51%)	782 (31%)
Kidney	16,616	7,442 (45%)	5,925	3457	3,958 (67%)	1,751 (51%)
Leg Muscle	11,401	2,006 (18%)	4,455	2467	1,394 (31%)	402 (16%)
Liver	12,507	6,773 (54%)	4,887	2715	3,574 (73%)	1,612 (59%)
Lung	14,238	9,216 (65%)	5,419	3032	4,448 (82%)	2,157 (71%)
Mammary	8,161	1,543 (19%)	3,566	1838	1,100 (31%)	302 (16%)
Ovary	15,108	2,043 (14%)	5,588	3229	1,407 (25%)	399 (12%)
Skin black	16,255	4,507 (28%)	5,870	3386	2,776 (47%)	999 (30%)
Skin white	17,087	3,533 (21%)	6,004	3531	2,156 (36%)	766 (22%)
Spleen	14,495	2,066 (14%)	5,317	3071	1,448 (27%)	382 (12%)
Thymus	9,781	986 (10%)	3,981	2159	634 (16%)	182 (8%)
Thyroid	18,181	3,279 (18%)	6,196	3703	2,013 (32%)	688 (19%)
Tongue	12,744	1,671 (13%)	4,850	2718	1,177 (24%)	327 (12%)
White blood cells	12,768	2,662 (21%)	4,680	2690	1,543 (33%)	552 (21%)
Total	25,251	19,082 (76%)	7,985	4856	7,067 (89%)	3,570 (74%)

REFERENCES

Bray N.J., Buckland P.R., Owen M.J. & O'Donovan M.C. (2003) *Human Genetics* **113**: 149-53.
Crowley J.J., Zhabotynsky V., Sun W, *et al.* (2015) *Nature Genetics* Advanced online article.
Daetwyler H.D., Capitan A., Pausch H., *et al.* (2014) *Nat Genet* **46**: 858-65.
Degner J.F., Marioni J.C., Pai A.A., *et al.* (2009) *Bioinformatics* **25**: 3207-12.
Gao C., Devarajan K., Zhou Y., *et al.* (2012) *BMC Genomics* **13**: 570.
Ge B., Pokholok D.K., Kwan T., *et al.* (2009) *Nature Genetics* **41**: 1216-22.
GTEx Consortium (2015) *Science* **348**: 648-60.
Kim D., Pertea G., Trapnell C., *et al.* (2013) *Genome Biology* **14**: R36.
Li H., Handsaker B., Wysoker A., *et al.* (2009) *Bioinformatics* **25**: 2078.
MacEachern S., Muir W.M., Crosby S.D. & Cheng H.H. (2012) *Frontiers in Genetics* 2.
Montgomery S.B., Sammeth M., Gutierrez-Arcelus M., *et al.* (2010) *Nature* **464**: 773-U151.
Pai A.A., Pritchard J.K. & Gilad Y. (2015) *PLoS Genetics* **11**: e1004857.
Pant P.V.K., Tao H., Beilharz E.J., *et al.* (2006) *Genome Research* **16**: 331-9.
Pastinen T. (2010) *Nature Reviews Genetics* **11**: 533–8
Pickrell J.K., Marioni J.C., Pai A.A., *et al.* (2010) *Nature* **464**: 768.
Serre D., Gurd S., Ge B., *et al.* (2008) *Plos Genetics* **4**.
Vidal D.O., De Souza J.E.S., Pires L.C., *et al.* (2011) *Genome* **54**: 120.
Wang X. & Clark A.G. (2014) *Heredity* **113**: 156.
Yan H., Yuan W., Velculescu V.E., Vogelstein B. & Kinzler K.W. (2002) *Science* **297**: 114.