Contributed paper

GENOME WIDE ANALYSIS OF BOVINE ENHANCERS AND PROMOTERS ACROSS DEVELOPMENTAL STAGES IN LIVER

M. Forutan[1], C.J Vander Jagt[2], E. Ross[1], A.J. Chamberlain[2], B. Mason[2], L. Nguyen[1], S. Moore[1], J.B. Garner[4], R. Xiang[3] and B.J. Hayes[1]

[1] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, QLD 4072, Australia
[2] Agriculture Victoria, Centre for AgriBiosciences, Bundoora, VIC 3083, Australia
[3] Faculty of Veterinary & Agricultural Science, The University of Melbourne, Parkville, VIC 3052, Australia
[4]Agriculture Victoria, Animal Production Sciences, Ellinbank Dairy Centre, Ellinbank, Victoria 3821, Australia

SUMMARY

Gene transcription is controlled by functional interactions between promoters and enhancers. Cap analysis of gene expression (CAGE) sequencing has allowed for the accurate annotation of most gene promoters (transcription start sites, TSS) and active enhancers. To date, TSSs and enhancer regions in the bovine genome are poorly characterised. To explore bovine developmental-specific patterns of enhancer-TSS usage and model TSS-enhancer interaction, CAGE-seq was applied to 6 bovine liver samples comprised of two different developmental stages (foetal and adult) obtained from 3 cows and their 3 foetuses. We identified approximately 30k and 20k TSSs and enhancer candidates, respectively, across the liver samples. About 231 significant TSS-enhancer interaction candidates were found by looking for closely spaced TSSs and enhancers that have highly correlated expression levels ($r > 0.75$; P-value < 0.05). Differential expression between development stages of TSS and enhancer candidates was performed using the Bioconductor package DESeq2 and identified 2050 (6) TSS (enhancer) candidates significantly differentially expressed across developmental stages (P-value < 0.05). The resulting catalogue of TSSs and active enhancers enables classification of developmental-specific TSSs-enhancers and modelling their interaction and provides major target regions for investigation of DNA methylation changes with aging. The information will also be useful in refining regions likely to contain causative mutations for complex traits associated with liver gene expression, such as feed efficiency.

INTRODUCTION

Identifying active regulatory regions in the genome is critical for understanding gene regulation and assessing the impact of genetic variation on phenotype. Although multiple processes are involved in gene expression regulation, the key role of promoters and enhancers has been a central focus of genome annotation for the past decade. Previous studies have confirmed that most genes have an array of close transcription start sites (TSSs) instead of the expected single TSS (FitzGerald *et al.* 2006; Hoskins *et al.* 2011; Djebali *et al.* 2012; Rojas-Duran and Gilbert 2012; Forrest *et al.* 2014), and the transcription of a gene may start from one of several TSSs, a phenomenon known as alternative transcriptional initiation (ATI, Landry *et al.* 2003; de Klerk and Hoen 2015). While promoters specify and enable the positioning of RNA polymerase machinery at TSSs, enhancers modulate the activity of promoters and play a key role in the formation of diverse cell types and respond to changing physiological conditions. Andersson *et al.* (2014) showed that enhancer activity can be detected through the presence of balanced bidirectional capped transcripts using Cap Analysis of Gene Expression (CAGE) (Takahashi *et al.* 2012). Active enhancers produce weak, but consistent, bidirectional transcription of capped enhancer RNA (eRNAs), resulting in a characteristic CAGE tag starting sites (CTSS) pattern of two diverging peaks approximately 400 bp

apart. A specific advantage of the CAGE method is that reads mapped to the genome provide accurate location of TSSs and active enhancers and quantify transcription (Kodzius *et al.* 2006; Carninci *et al.* 2007).

To date, enhancer regions in the bovine genome are poorly characterised. To explore bovine tissue-specific patterns of enhancer-TSSs usage, CAGE sequencing was applied to 6 bovine samples comprised of 2 different developmental-stages obtained from 3 cows and their 3 foetuses. To the best of our knowledge, this study is the first bovine TSS-enhancer discovery using CAGE-Seq data.

MATERIALS AND METHODS

CAGE library preparation and sequencing. Two liver samples were collected from one pregnant *Bos indicus* (Brahman cow) and the female cow's foetus (approximately 12 weeks old). Four liver samples were collected from two *Bos taurus* pregnant cows and their female foetuses (approximately 16 weeks old) at the Ellinbank research facility with approval from the DEDJTR Animal Ethics Committee (2014-23). Samples (cows and foetus) were collected from the same anatomical region. The samples were harvested after the cow was slaughtered, immediately snap-frozen in liquid nitrogen, and stored at -80ºC until processing (Forutan *et al.* 2021).

Read processing and alignment. Sequence read quality was assessed using FastQC (Andrews 2010), including calculation of GC content, and identification of over-represented sequences. The EcoP15I fingerprint was trimmed by cutting the first 9 bases (*CROP:9*) and Illumina adaptor trimmed by cutting the last 14 bases (*HEADCROP:36*) using Trimmomatic (Bolger *et al.* 2014) (version 0.35). Trimmed reads were aligned to *Bos taurus* reference genome (GenBank: ARS-UCD1.2) with Burrows-Wheeler Aligner (BWA**.** Li and Durbin 2009), version 0.7.13) using the BWA-MEM algorithms. The aligner was run using default parameters, the only exceptions were t=10, and k=10. Also, to alleviate the presence of universal G at the head of the read, which may be present in some of the reads, parameters L (clipping penalty) and B (mismatch penalty) were assigned as 4 and 5, respectively.

Quality controls and preliminary analyses. Only primary alignments with a quality of greater than 20 (>99% chance of true) were considered for TSSs and enhancers calling. Further filtering was applied by only selecting CTSS with 3 or more CAGE reads in at least one sample for TSSs calling. Considering that active enhancers produce weak but consistent bidirectional transcription of capped enhancer RNAs (eRNAs), more relaxed filtration was used for enhancer calling (selecting CTSS with 2 or more CAGE reads in at least one sample). The total number of reads before and after quality control and numbers of TSSs and active enhancer candidates across all samples is shown in Table 1.

TSSs and enhancers calling. *clusterUnidirectionally* function and the parameter *mergeDist* 20 available in *CAGEfightR* package (Thodberg *et al.* 2019) was used to call TSSs. Ensembl database release 104 for Bos taurus was used for annotation of the signals. Only TSSs overlapping promoter, proximal and 5'UTR regions were used for further analysis. Identification of active enhancer candidates was done using *clusterBidirectionally* function with a balance score > 0.95 in the *CAGEfightR* package. The enhancers not overlapping intergenic and intron regions were removed from the analysis. TSS-enhancer interaction candidates were identified using *findLinks* function from the *InteractionSet* package into an R session (version /4.0.2) by looking for closely spaced TSSs and enhancers that have highly correlated expression within 20 kb distance. Differential TSSs and enhancer usage across developmental stages was performed by using the Bioconductor package DESeq2 (Love *et al.* 2014) and keeping only TSSs expressed in all samples (10,813 TSSs) and enhancers observed to be bidirectional in all samples (21 bidirectional enhancers). The *findStretches* function from *CAGEfightR* package was used to identify groups of closely spaced enhancers, where all enhancers were within a 10 kb distance of another member.

Data availability. Bos taurus and Bos indicus raw sequence data are publicly available via European Nucleotide Archive (ENA) under study ID PRJEB43513 and PRJEB44817, respectively.

RESULTS AND DISCUSSION

Genome-wide association studies (GWAS) have discovered many variants for complex diseases and quantitative traits. However, many implicated variants are classified as non-coding and, they are thought to play a role in gene expression regulation. Functional annotations provide valuable information for prioritizing potential causal variants within complex-trait loci identified through GWAS. Like any specific tissue in the body, the biological features of tissue in foetal and adult stages may be determined mainly at the level of gene expression. So, identification of functional regions such as enhancer and TSSs and differential and quantitative analysis of developmental stage-specific TSS-enhancers expression could be useful to identify informative variants and ultimately improve genomic prediction. In total, 29,940 and 19,264 TSSs and candidate enhancers were detected across all samples, respectively (Table 1). Only 36% of TSSs (10,813) were expressed across all 6 samples. The lower number of enhancers was observed in the adult stage compared to the foetal stage (Table 1). In total, among the 19,264 active enhancer candidates expressed across samples, only a small proportion of enhancer candidates (less than 1%) were expressed across all samples. The enhancers are context-specific and respond to specific physiological, pathological, or environmental conditions which can cause the large variation in number of enhancers observed across samples. About 231 significant TSS-enhancer interaction candidates were found by looking for closely spaced TSSs and enhancers that have highly correlated expression levels ($r > 0.75$; P-value < 0.05). Examination of the differential TSS usage across developmental stages controlling for effect of sub-species revealed 2050 differentially significant TSSs (P-value <0.05). We found 6 developmental enhancers based on the differential enhancer usage analysis (P-value < 0.05), which could be the potential targets of DNA methylation in bovine liver. One of the developmental stage-specific genes in liver is Sulfotransferase isoform 1A1 (*SULT1A1*). *SULT1A1* is the most highly expressed hepatic sulfotransferase and plays the central role in detoxification. Out of five TSSs observed across samples for this gene *(*Figure 1), two of them were expressed in all samples (TSSs peaks located on positions 26,126,989 bp and 26,127,457 bp) and only the TSS on position 26,126,966 – 26,127,032 bp showed significantly differential expression in foetal stage compared to adult stage (log2FoldChange = -3.291495; adjusted P-value < 0.0006).

Table 1. Summary of the number of CAGE tags, transcription start site (TSS) and enhancer candidates expressed in bovine liver

Stage	Biological samples	Number of CAGE tags			Number of TSS		Number of enhancers	
		Total	For TSS calling	For enhancers calling	Total	In promoter, 5'UTR, proximal	Total	In intergenic and intron
Adult	*B.taurus* rep1	5,850,606	1,869,376	2,283,311	97,639		6,422	3,980
	B.taurus rep2	5,678,632	2,167,220	2,437,174				
	B.indicus rep1	5,183,691	4,647,556	4,796,404				
Foetal	*B.taurus* rep1	10,048,108	2,607,022	3,358,781		140,591	22,310	17,386
	B.taurus rep2	9,327,767	1,632,971	2,476,168				
	B.indicus rep1	7,448,786	5,552,082	5,844,322				
Total		43,537,590	18,476,227	21,196,160		162,275	24,605	19,264

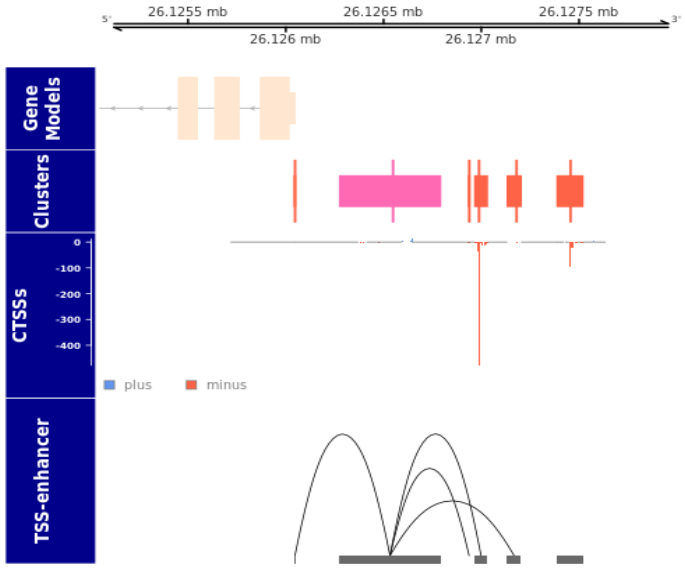

Figure 1. Plot of position of CAGE tag starting sites (CTSSs), TSSs (orange clusters), and enhancer candidate (pink cluster) of the *SULT1A1* gene in bovine liver. Gene model is plotted based on the Ensembl database (bos_taurus_core_104_12). The links between TSSs and active enhancers is plotted using arches, scaling the height of the arches according to P-values of Kendall correlation

CONCLUSIONS

Knowledge of interaction between bovine TSS and enhancer expression would be a useful starting point to predict biological function of specific genes in different developmental stages. In the current study, CAGE-seq was used for the first time to assess TSS-enhancer interactions in bovine liver. Also, we assessed differential TSSs and enhancer usages across developmental-stages in liver tissue for the first time in cattle using CAGE-seq. The results of this study will accelerate future genomic research and will assist in narrowing down candidate genes with differential TSS and enhancer usage across foetal and adult stages in liver. The information will also be useful in refining regions likely to contain causative mutations for complex traits associated with liver gene expression, such as feed efficiency. A limitation with the current study is that only one biological replicate was included for the Bos indicus cow-foetus, so analysis of additional would increase the resolution of the findings.

ACKNOWLEDGMENTS

We acknowledge financial contributions from Meat and Livestock Australia (P.PSH.0868) for the generation of the *Bos taurus indicus* CAGEseq data. We would also like to acknowledge financial contributions from DairyBio (a joint venture project between Agriculture Victoria and Dairy Australia) and Research Initiative Fund of the Faculty of Veterinary& Agriculture Sciences of The University of Melbourne for the generation of the *Bos taurus taurus* CAGEseq data. We are thankful to Dr. Brian Burns for helping source the *Bos taurus indicus* tissues, and Dr. Bronwyn Venus for collecting *Bos taurus indicus* samples. Thank you to Elise Kho for extracting some the *Bos taurus indicus* RNA.

REFERENCES

Andrews S. (2010) Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Bolger A.M., Lohse M., and Usadel, B. (2014). *Bioinformatics*. **30**: 2114.
Carninci P., Sandelin A., Lenhard B. *et al*. (2006) *Nat. Genet.* **38**:626.
de Klerk E. and Hoen P.A.C't (2015) *Trends Genet.* **31**: 128.
Djebali S., Davis C.A., Merkel A. *et al*. (2012) *Nature***. 489**:101.
FitzGerald P.C., Sturgill D., Shyakhtenko A. *et al*. (2006) *Genome Biol.* **7:** R53.
Forrest A.R., Kawaji H., Rehli M. *et al*. (2014) *Nature***. 507**:462.
Forutan, M., Ross, E., Chamberlain, A.J. *et al*. (2021) *Commun Biol.* **4**: 829.
Hoskins R.A., Landolin J.M., Brown J.B. *et al*. (2011) *Genome Res*. **21**: 182.
Kodzius R., Kojima M., Nishiyori H. *et al*. (2006) *Nat. Methods*. **3**: 211.
Landry J.R., Mager D.L. and Wilhelm B.T. (2003) *Trends Genet.* **19**: 640.
Li H. and Durbin R. (2009) *Bioinformatics*. **25**: 1754.
Love M.I., Huber W., and Anders S. (2014) *Genome Biol.* **15**: 1.
Andersson R., Gebhard C., Miguel-Escalada I. *et al*. (2014) *Nature***. 507**: 455**.**
Rojas-Duran M.F. and Gilbert W.V. (2012) *RNA.* **18**: 2299.
Takahashi H., Kato S., Murata M. *et al.* (2012) *Methods Mol Biol.* **786**: 181.
Thodberg M., Thieffry A., Vitting-Seerup K. *et al*. (2019) *BMC Bioinform.* **20**: 1.