# Comparative analysis of CAGE-Seq across tissues reveals transcription start sites unique to cattle

**M. Salavati[1*], R. Clark[2], D. Becker[3], C. Kühn[3,4], G. Plastow[5], G.C.M. Moreira[6], C. Charlier[6,7] and E.L. Clark[1] on behalf of the BovReg consortium**

[1] The Roslin Institute, University of Edinburgh, EH25 9RG, Edinburgh, UK; [2] Genetics Core, Edinburgh Clinical Research Facility, The University of Edinburgh, EH4 2XU, Edinburgh, UK; [3] Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), 18196, Dummerstorf, Germany; [4] Faculty of Agricultural and Environmental Sciences, University Rostock, 18059, Rostock, Germany; [5] Livestock Gentec, Department of Agricultural, Food and Nutritional Science, University of Alberta, T6G 2R3, Edmonton, Canada; [6] Unit of Animal Genomics, GIGA Institute, University of Liège, 4000, Liège, Belgium; [7] Faculty of Veterinary Medicine, University of Liège, 4000, Liège, Belgium; [*] Mazdak.Salavati@roslin.ed.ac.uk

## Abstract

Recently, expressed genomic information e.g., expression quantitative loci (eQTL), have been utilised in both dairy and beef breeding to link genotype to phenotype. The fine mapping of promoters (transcription start sites [TSS]) and enhancers (divergent amplifying segments local to TSS) in different populations of cattle across multiple tissues can help to unravel breed- and tissue-specific phenotypes. This study provides a high-resolution analysis of TSS complexity including drivers of transcription using Cap Analysis Gene Expression (CAGE) sequencing in 3 populations of cattle (Holstein, Charolais x Holstein and Kinsella beef composite [KC]). We identified 55,139 TSS and 2,543 TSS-Enhancer regions shared by 3 populations along with population-specific sets. Comparative analysis revealed a cattle-specific set of TSS. Improving our understanding of the genomic architecture of gene expression and regulation in this way can help to inform the application of genomic technologies in breeding programmes for cattle.

## Introduction

Mapping of the actively transcribed regions of the genome helps identify the drivers of gene expression, regulation and phenotypic plasticity. TSS within promoter regions provide information about how genes controlling traits of interest are expressed and regulated. These TSS can be specific to a tissue or pervasive across the breed/species. To improve TSS and enhancer annotation of the current cattle genome (ARS-UCD1.2), we used CAGE sequencing to create base-pair resolution expression profiles of TSS and their co-expressed short-range enhancers (<1kb) in the ARS-UCD1.2_Btau5.0.1Y reference genome (1000bulls run7). We utilised publicly available CAGE datasets for human, chicken, mouse, rat, macaque monkey, dog and sheep for a cross-species comparative analysis of TSS and TSS-Enhancers and their co-expressive interaction. Using comparative analysis this study aims to provide a cattle-specific set of TSS and TSS-Enhancers in multiple tissues from dairy (Belgian Holstein), and cross-bred composite beef (KC) and beef/dairy cross (Charolais x Holstein) cattle.

## Materials & Methods

### *CAGE sequencing and analysis pipeline.*

A total of 105 samples from 24 different tissues were collected from 6 animals (3 populations, different ages and 2 sexes). Tissue representation for each breed was as follows: dairy (Holstein, n=43), composite beef (KC-composite, n=31) and beef/dairy cross (Charolais x Holstein, n=31). CAGE libraries were prepared as for (Salavati et al.2020) with some modifications of

the barcodes to facilitate sequencing on an Illumina NextSeq 550 (50nt SE) and mapped against the ARS-UCD1.2_Btau5.0.1Y assembly using the nf-cage pipeline (BovReg 2021).

***Transcription start site and enhancer prediction analysis.***
The base pair (bp) resolution output files were processed by CAGEfightR v.15.1 for cattle. For comparative analysis a CAGE dataset from sheep (Salavati et al. 2020) was mapped to ARS-UI_Rambv2.0 and also processed by CAGEfightR v.15.1. The uni- and bi-directional clustering algorithm of the CAGEfightR was used to predict putative TSS and TSS-Enhancers within each tissue individually (i.e. tissue-specific) and all tissues combined (species and population). Co-expression of predicted TSS and TSS-Enhancer regions was tested using a Kendall correlation test ($p<0.05$ sig.). Finally a hierarchical clustering of the super-enhancers was performed to identify longer (>1kb) regulatory stretches of the genome.

***Re-analysis of the Fantom5 CAGE datasets.***
Already mapped CAGE datasets from human (152), rat (13), mouse (17), chicken (32), dog (13) and Macaque monkey (15) were obtained from (Fantom5 2021). After conversion to bp resolution tracks similar analysis to the cattle and sheep was performed to identify putative TSS and TSS-Enhancer regions.

**Results**
An average (± SE) of 15.5±0.53 million reads per CAGE library were produced and a 94% mapping rate was achieved for the cattle dataset. Initially more than 4.5 million putative TSS and 57,412 TSS-Enhancer clusters were identified in total (min 10 read counts/cluster – Figure 1).

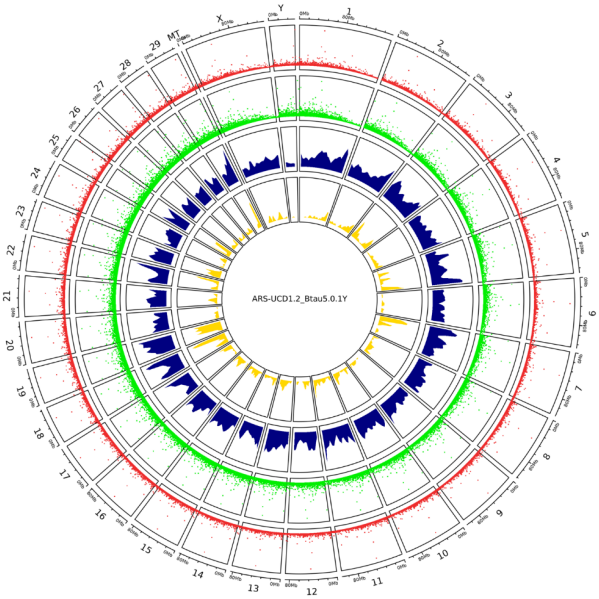


**Figure 1. Distribution of uni-directional (TSS) and bi-directional (TSS-Enhancer) CAGE clusters within the cattle genome (ARS-UCD1.2_Btau5.0.1Y). The TSS clusters (red), TSS-Enhancer (green), significant positive (blue) and negative (yellow) correlation between co-expressed Enhancer and TSS(s) are shown in genomic tracks. The height of the tracks shows scaled expression or correlation coefficients (0-1).**

We considered a putative region, real/reproducible only when present across at least 2/3$^{rd}$ of the tissues. After such filtering 55,139 TSS and 2,543 TSS-Enhancers were detected for cattle

(Table 1). Tissue-specific analysis captured, on average per tissue (± SE), 253,852 ± 24,713 TSS clusters, 41.6% of which were novel. On average per tissue 12,138 ± 889 TSS-Enhancer clusters were captured (27.6% novel). Population-specific analysis showed differences in TSS coordinates and expression levels between the populations. The highest number of population-specific TSS were in the KC-composite (3,102) followed by 1,152 in Holstein and 1,092 in Charolais x Holstein. The same pattern was observed in the TSS-Enhancer clusters (419 in KC-composite, 286 in Charolais x Holstein and 202 in Holstein). A comparison of the sheep dataset and the cattle dataset is shown in Table 1. The higher number of identified regions in cattle is partly due to inclusion of multiple tissue samples from 3 populations (compared to a single animal for the sheep data).

**Table1. Mapped and annotated CAGE uni-directional clusters (TSS) in sheep and cattle using reference assembly gene models (min 2/3rd tissue representation).**

| Genomic | ARS-UI_Ramb_v2.0 | | | | ARS-UCD1.2_Btau5.0.1Y | | | |
|---|---|---|---|---|---|---|---|---|
| region | Novel | Anno* | Total | %$ | Novel | Anno* | Total | %$ |
| promoter | 0 | 13,372 | 13,372 | 39.4 | 694 | 8,740 | 9,434 | 17.1 |
| proximal | 0 | 944 | 944 | 6 | 252 | 2,807 | 3,059 | 5.5 |
| fiveUTR | 0 | 873 | 873 | 6.7 | 141 | 2,901 | 3,042 | 5.5 |
| threeUTR | 0 | 2,197 | 2,197 | 7.9 | 38 | 2,257 | 2,295 | 4.2 |
| CDS | 0 | 4,513 | 4,513 | 16.1 | 363 | 8,977 | 9,340 | 16.9 |
| exon | 0 | 295 | 295 | 1.9 | 0 | 0 | 0 | 0 |
| intron | 0 | 2,386 | 2,386 | 10 | 485 | 8,100 | 8,585 | 15.6 |
| antisense | 1,034 | 0 | 1,034 | 4.2 | 5,991 | 17 | 6,008 | 10.9 |
| intergenic | 1,397 | 0 | 1,397 | 7.9 | 13,315 | 61 | 13,376 | 24.3 |
| Total TSS | 2,431 | 24,580 | **27,011** | 100 | 21,279 | 33,860 | **55,139** | 100 |
| Total TSS-En£ | 34 | 1,459 | **1,493** | | 575 | 1,968 | **2,543** | |
| Annotated genes/transcripts | | | 13,771 / 45,298 | | | | 13,251 / 24,756 | |

* Annotated using the reference assembly gff3 track

$ Percentage calculated based on total per genomic region category / total TSS clusters

£ TSS-Enhancers (Bi-directional clustering)

The correlation (Kendall) analysis between a bi-directional cluster and multiple uni-directional clusters (significantly co-expressed TSS and TSS-Enhancers) showed on average 3.73±0.05 (median 3) and 22.8±0.02 (median 18) links in the sheep (3,694 regions) and cattle (641,028 regions) datasets respectively. The Kendall estimates of these significant co-expressed links were 0.547 and 0.363 for the sheep and cattle tissues respectively. The analysis of the super enhancers (stretches of bi-directional CAGE clusters) encompassing multiple enhancers within each stretch in the sheep dataset resulted in 2 super enhancer predictions. These stretches were formed of 6 TSS-Enhancer clusters (total 1,493) with the longest stretch of 5,172bp [cluster of 3 enhancers]. The similar analysis of the cattle CAGE dataset from 3 populations resulted in 20 super enhancer stretches from 65 TSS-Enhancer clusters (total 2,543). The longest stretch was 25,678bp which contained 6 TSS-Enhancers. The number of discovered super enhancer regions overall was higher in the cattle dataset (3 populations) compared to the sheep (single breed). A multi species metrics for available CAGE datasets is shown in Table 2.

**Table2. Comparison of the mapped TSS and annotated genes identified in other CAGE datasets (Fantom5, OvineFAANG and BovReg). Column 'Genes' corresponds to only the genes that were annotated using the CAGE data (using the 2/3rd representation rule). The table is sorted (descending) by the number of unique TSS identified in each genome.**

| Species | Genome | TSS↓ | Genes |
|---|---|---|---|
| Human | hg38 | 209,911 | 31,184 |
| Mouse | mm10 | 164,672 | 30,501 |
| **Cow** | **ARS-UCD1.2_Btau5.0.1Y$** | **55,139** | **13,251** |
| Chicken | galGal5 | 32,015 | 7,759 |
| Rat | rn6 | 28,497 | 13,719 |
| **Sheep** | **Oar rambouillet v1.0£** | **28,148** | **13,912** |
| **Sheep** | **ARS-UI_Ramb_v2.0*** | **27,011** | **13,771** |
| Rhesus monkey | rheMac8 | 25,869 | 8,047 |
| Dog | canFam3 | 23,147 | 5,288 |

* NCBI RefSeq gff3 annotation v104
$ Ensembl gff3 annotation v103 track lifted over to 1000bulls reference genome
£ NCBI RefSeq gff3 annotation v100

**Discussion**

Recently there has been an influx of breed-specific reference assemblies in livestock species (Crysnanto et al.2021, Li et al.2019). To harness the full potential of these assemblies, identifying breed or population-specific promoter complexity is valuable within cattle genomic breeding programmes (Clark et al. 2020). To this aim, we used CAGE sequencing and publicly available datasets in a comparative analysis, to create further annotation underpinning the phenotypic diversity seen in 3 diverse cattle breeds. We identified more than 55k unique putative TSS coordinates (38% un-annotated regions of the cattle genome) compared to 27k TSS in sheep (7% unannotated regions). The promoter plasticity captured by the usage of 3 divergent populations showed population-specific dominant TSS for the same genes across the breeds e.g. a *DGAT1* Holstein-specific TSS (overlapping milk yield QTLs) without expression in other breeds or *PLAG1* Charolais-specific TSS-Enhancer region (overlapping cow weight QTLs) that was not present in the other 2 populations. We will also compare the sheep and cattle datasets with other publicly available TSS and Enhancer genomic tracks to further identify regulatory regions specific to the cattle genome. Such information could be used to understand how the genome controls traits in different species, and to identify regions that are important to conserve in breeding programmes. The CAGE data produced for this study when combined with transcriptomic datasets (mRNA, miRNA and total RNA-Seq) produced by BovReg partners will provide a new comprehensive transcriptome annotation for the cattle genome. This high resolution annotation of the cattle genome will help to inform the application of genomic technologies in breeding programmes. The BovReg project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668.

**References**

BovReg's nf-cage pipeline 2021. Available at: https://github.com/BovReg/nf-cage.git
Clark E.L., Archibald A. L., Daetwyler H.D., Groenen M. A.M., Harrison P.W. et al. (2020). Genome Biol. 21, 1–9. https://doi.org/10.1186/s13059-020-02197-8.
Crysnanto, D., Leonard, A. S., Fang, Z. H., and Pausch, H. (2021). Proc. Natl. Acad. Sci. U. S. A. 118, e2101056118. https://doi.org/10.1073/pnas.2101056118.
Fantom5 data repository 2021. Available at: https://fantom.gsc.riken.jp/5/datafiles/basic
Li R., Fu W., Su R., Tian X., Du D. et al. (2019). Front. Genet. 10, 1169. https://doi.org/10.3389/fgene.2019.01169.
Salavati M., Caulton A., Clark R., Gazova I., Smith T.P.L. et al. (2020). Front. Genet. 11, 1184. https://doi.org/10.3389/fgene.2020.580580.