**Identification of functionally active genomic features relevant to phenotypic diversity and plasticity in cattle**

# Deliverable 2.2

# BAM Files for CAGE Libraries

**Grant agreement no°: 815668**

Due submission date

**2020-09-30**

Actual submission date

**2021-02-27**

Responsible author(s)

**Emily Clark, UEDIN (emily.clark@roslin.ed.ac.uk)**

**Mazdak Salavati, UEDIN (mazdak.salavati@roslin.ed.ac.uk)**

**Richard Clark, UEDIN (richard.clark@ed.ac.uk)**

**Confidential No**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668. The content of this report reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

## DOCUMENT CONTROL SHEET

| | |
|---|---|
| Deliverable name | BAM files for CAGE Libraries |
| Deliverable number | 2.2 |
| Partners providing input to this Deliverable | UEDIN, ULIEGE, FBN, UALBERTA, INRAE |
| Draft final version circulated by lead party to:<br>On date | All partners in WP2<br>2021-02-23 |
| Approved by (on date) | FBN as Coordinator (2021-02-27) |
| Work package no | 2 |
| Dissemination level | Public (PU) |

## REVISION HISTORY

| Version number | Version date | Document name | Lead partner |
|---|---|---|---|
| Vs1 | 2021-02-18 | D2.2 | UEDIN |
| Vs2 | 2021-02-22 | D2.2 | ULIEGE |
| Vs3 | 2021-02-23 | D2.2-BAM_files_CAGE-18022021_v3 | UEDIN |
| Vs4 | 2021-02-24 | D2.2-BAM_files_CAGE-23022021_v3 | UEDIN |
| Vs5 | 2021-02-25 | D2.2-BAM_files_CAGE-25022021_v4. | UEDIN |

**Changes with respect to the DoA (Description of Action)**

No changes

**Dissemination and uptake**

This deliverable is for public use.

## Table of Content

## 1. Summary of Results

One of the aims of the BovReg project is to improve the genomic resources that are available for cattle. Transcription Start Sites (TSS) tell us where each gene starts in the genome. Mapping of TSS is a key first step in understanding transcript regulation and diversity. For the work in T2.1 as documented in D2.2, Cap Analysis Gene Expression (CAGE) sequencing performed for 109 tissue samples collected by BovReg partners, from two neonatal Holstein animals from Belgium (ULIEGE), two juvenile Kinsella composite animals from Canada (UALBERTA) and two Charolais x German Holstein F2 adult animals from Germany (FBN) to perform a global analysis of TSS. CAGE libraries were prepared and sequenced, and 109 BAM files generated.

The BAM files have been made available to partners within BovReg via its private data portal. All mapping was performed against the ARS-UCD2.1_Btau5.0.1Y assembly (1000 Bull Genome project reference file). Preliminary analysis revealed that a large percentage of CAGE tags mapped to intronic regions, indicating the gene models in the current annotation could be improved to better capture novel transcript and isoform complexity.

The CAGE data generated for D2.2 will be combined with RNA-Seq and small RNA-Seq from ULIEGE to generate a high-resolution transcriptome map that will significantly improve the annotation information available for cattle.

## 2. Introduction

The Deliverable D2.2 “*BAM files for CAGE libraries*”, is related to the work described in task 2.1 of the BovReg project. Mapping of transcription start sites (TSS) is a key first step in understanding transcript regulation and diversity. TSS mapping can provide information about complex promotor activity, pervasive transcription and tissue-specific promotor usage in farmed animal tissues [1,2]. Currently, the functional annotation of the bovine genome (ARS-UCD1.2) lacks precise transcription start sites and includes a low number of transcripts in comparison to human and mouse [1].

To remedy this for D2.2 Cap Analysis Gene Expression (CAGE) sequencing was used for 109 tissue samples collected from two neonatal Holstein animals from Belgium (ULIEGE), two juvenile Kinsella composite animals from Canada (UALBERTA) and two Charolais x German Holstein F2 adult animals from Germany (FBN) to perform a global analysis of TSS. CAGE measures RNA expression by 5’ cap-trapping to identify the 5’ ends of both polyadenylated and non-polyadenylated RNAs including lncRNAs and miRNAs, and has been specifically designed to allow the characterization of TSS within promoters and enhancers to single-nucleotide resolution [3].

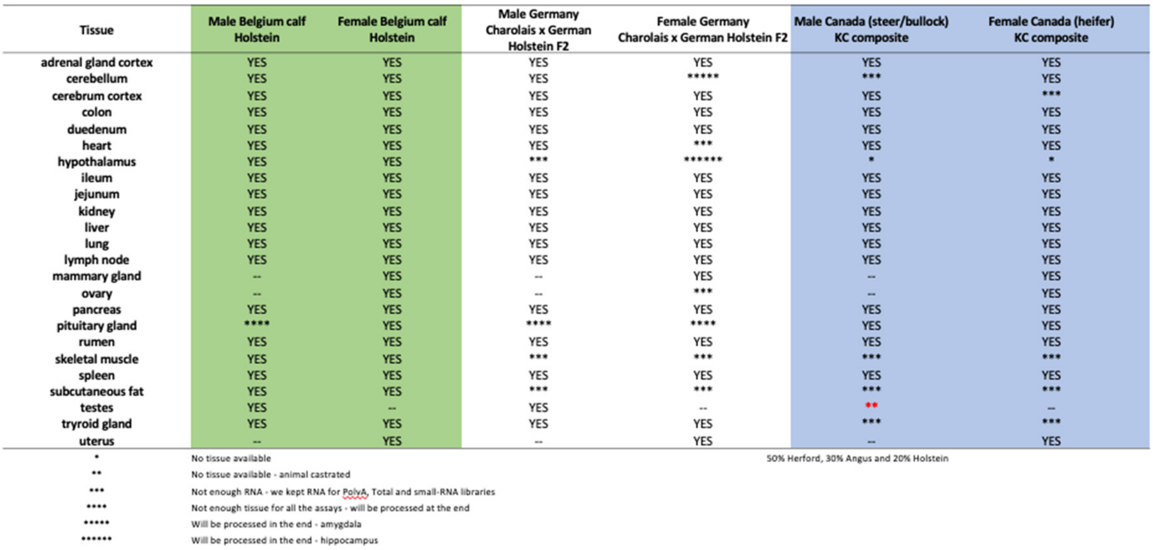
The level of resolution provided by CAGE allows investigation of the regulatory inputs driving transcript expression, and construction of transcriptional maps. The aim of Deliverable 2.2 was to generate Binary Alignment Map (BAM) files for the 109 CAGE sequencing libraries. These BAM files would then be used in D2.3 to provide a map of transcription start sites in the bovine reference genome (ARS-UCD1.2_Btau5.0.1Y).

## 3. Core Report

### Tissue Samples

CAGE-sequencing was performed on 24 tissues (Table 1) from six animals. Two animals were included (one of each gender) from a beef breed, a dairy breed and a beef/dairy (Charolais x Holstein) crossbreed. Tissues had been collected and archived for RNA extraction and library preparation by partners FBN, ULIEGE and UALBERTA prior to the start of the BovReg project. The tissues assayed were predominantly FAANG priority tissues (Tier 1 and 2 tissues), and were chosen because they were particularly relevant to BovReg traits (*e.g.* mammary gland, liver and ovary). All RNA was isolated at ULIEGE (BovReg technician) and shipped to UEDIN. Generation of 144 CAGE libraries was planned for Deliverable 2.2 but for some animals tissue samples were either not available or there was insufficient tissue or RNA (Table 1). As such 109 CAGE libraries were prepared and the associated BAM files generated. These 109 samples include tissue samples from all the major organ systems and capture the vast majority of transcriptional diversity.

**Table 1: Details of tissue samples provided for CAGE sequencing, including information relating to any missing samples.**



| Tissue | Male Belgium calf Holstein | Female Belgium calf Holstein | Male Germany Charolais x German Holstein F2 | Female Germany Charolais x German Holstein F2 | Male Canada (steer/bullock) KC composite | Female Canada (heifer) KC composite |
|---|---|---|---|---|---|---|
| adrenal gland cortex | YES | YES | YES | YES | YES | YES |
| cerebellum | YES | YES | YES | ***** | *** | YES |
| cerebrum cortex | YES | YES | YES | YES | YES | *** |
| colon | YES | YES | YES | YES | YES | YES |
| duedenum | YES | YES | YES | YES | YES | YES |
| heart | YES | YES | YES | *** | YES | YES |
| hypothalamus | YES | YES | *** | ****** | * | * |
| ileum | YES | YES | YES | YES | YES | YES |
| jejunum | YES | YES | YES | YES | YES | YES |
| kidney | YES | YES | YES | YES | YES | YES |
| liver | YES | YES | YES | YES | YES | YES |
| lung | YES | YES | YES | YES | YES | YES |
| lymph node | YES | YES | YES | YES | YES | YES |
| mammary gland | -- | YES | -- | YES | -- | YES |
| ovary | -- | YES | -- | *** | -- | YES |
| pancreas | YES | YES | YES | YES | YES | YES |
| pituitary gland | **** | YES | **** | **** | YES | YES |
| rumen | YES | YES | YES | YES | YES | YES |
| skeletal muscle | YES | YES | *** | *** | *** | *** |
| spleen | YES | YES | YES | YES | YES | YES |
| subcutaneous fat | YES | YES | *** | *** | *** | *** |
| testes | YES | -- | YES | -- | ** | -- |
| tyroid gland | YES | YES | YES | YES | *** | *** |
| uterus | -- | YES | -- | YES | -- | YES |
| * | No tissue available | | | | 50% Herford, 30% Angus and 20% Holstein | |
| ** | No tissue available - animal castrated | | | | | |
| *** | Not enough RNA - we kept RNA for PolyA, Total and small-RNA libraries | | | | | |
| **** | Not enough tissue for all the assays - will be processed at the end | | | | | |
| ***** | Will be processed in the end - amygdala | | | | | |
| ****** | Will be processed in the end - hippocampus | | | | | |

### CAGE library preparation and sequencing (performed by the Clinical Research Facility, University of Edinburgh)

### Library Preparation

CAGE libraries were prepared at the Clinical Research Facility at UEDIN from 5µg each of total RNA sample according to the protocol published by [3] with some modification of oligonucleotide sequences to allow libraries to be sequenced on Illumina paired-read flow cells such as those used on a NextSeq 500/550 instrument.

Briefly, cDNA was reverse transcribed with PrimeScript reverse transcriptase (Takara) using a random primer including the EcoP15I sequence and polyadenylated and non-polyadenylated RNA as template. Cap and 3' ends are biotinylated, and after RNase digestion of nonhybridised single-stranded RNA, 5' complete cDNAs hybridised to biotinylated capped RNAs are captured by MPG streptavidin-coated magnetic beads (Takara). The cDNA was next released from RNA and ligated to a 5' linker including a 6-base barcoded sequence and EcoP15I sequence. 5' linker ligated cDNA samples are then pooled and purified with Agencourt AMPure XP beads (Beckman Coulter) before double-strand 5' linkers are denatured to allow the biotin-modified second SOL primer to anneal to the single-stranded cDNA and prime second-strand cDNA synthesis. Subsequently, cDNA was digested with EcoP15I (NEB), which cleaves 27bp inside the 5' end of the cDNA. Next, a 3' linker containing the 3' Illumina primer sequence was ligated at the 3' end. The CAGE tags (around 106bp) were amplified (9 cycles) with the forward primer and reverse primer that are both compatible with the Illumina paired-read flow cell surface. Treatment with Exonuclease I degrades excess primers but not the double-stranded CAGE tags. Final CAGE tag pools were then purified with the MinElute PCR purification kit (QIAGEN).

**Sequencing**

Sequencing was performed using the NextSeq 500/550 High-Output v2.5 (75 cycles) Kit on the NextSeq 550 platform (Illumina Inc) over seven flow cells. Library molarity for sequencing was calculated using Qubit dsDNA quantification results and the fragment size information from Bioanalyser results. A custom sequencing primer was spiked in to the Illumina sequencing primers to allow sequencing of the CAGE tags. As the first part of each read contains identical sequence i.e. very low sequence diversity, PhiX v3 Control Library (Illumina) was spiked in to the CAGE tag pools at a concentration of 25% to increase library diversity and improve cluster resolution.

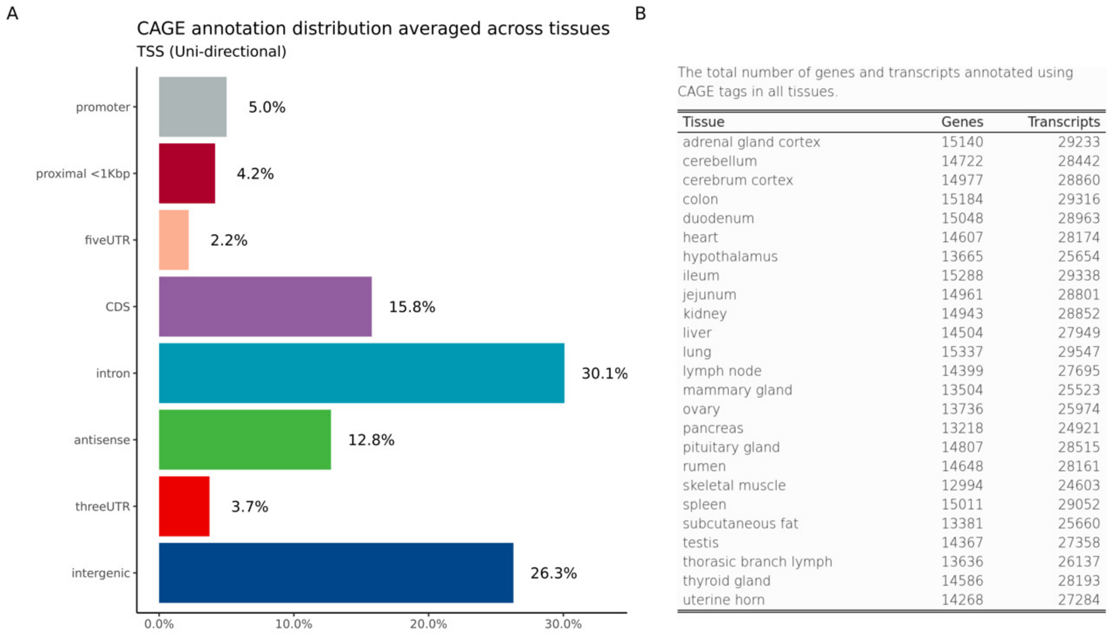
**Processing and mapping of CAGE libraries**

All sequence data were processed using in house scripting (bash and R) on the UEDIN high performance computing facility (Mazdak Salavati). The analysis protocol for CAGE is available via the FAANG Data Portal
https://data.faang.org/api/fire_api/analysis/ROSLIN_SOP_CAGE_analysis_pipeline_20191029.pdf .
To de-multiplex the data UEDIN used the FastX toolkit version 0.014 [4] for short read pre-processing. TagDust2 v.2.33 was then used [5] to extract mappable reads from the raw data and for read clean-up to remove the *EcoP1* site and barcode, according to the recommendations of the FANTOM5 consortium e.g. [6]. This process resulted in cleaned approximately 27nt reads (hereafter referred to as CAGE tags) which were mapped to the ARS-UCD1.2_Btau5.0.1Y reference genome using Bowtie2 v.2.3.5.1 in --very-sensitive mode equivalent to options *-D 20 -R 3 -N 0 -L 20 -i S,1,0.50* [7]. Multi-mapped reads were identified using Bowtie2 v.2.3.5.1 in --very-sensitive mode and excluded from the rest of the analysis. The mapped BAM files were then processed for base pair resolution strand specific read counts using bedtools v.2.29.0 [8]. On average 15 million trimmed reads were generated per tissue. The only exceptions were four pancreas samples (two from

FBN and two from UALBERTA) where the number of reads was less than 1 million per sample. Mapping rates to ARS-UCD1.2_Btau5.0.1Y were on average 94% across tissues. In order for the bedGraph files to be used in the CAGEfightR package they were converted to bigWig format using UCSCs tool BedGraphToBigWig [9].

### Normalisation and mapping of CAGE tags

For normalisation and clustering of CAGE tags (as CAGE Tags-Per-Million Mapped: CTPM) the software package CAGEfightR v.1.10.0 [10] was used. The normalisation was performed by dividing CAGE tag counts in each predicted cluster by the total mapped CAGE tags in the sample, multiplied by 1.0e6. To perform these analyses a custom BSgenome object (a container of the genomic sequence) was created for cattle from *ARS-UCD1.2_Btau5.0.1Y* using the BSgenome Bioconductor package v.1.58.0 [11]. Distribution metrics of CAGE tags across the genome were annotated and analysed using the TxDB transcript ID assignment and Genomic Features package v.1.42.1 [12]. The TxDB object was created using the Ensemblv102 gff3 gene annotation file from:
http://ftp.ensembl.org/pub/release-102/gff3/bos_taurus/Bos_taurus.ARS-UCD1.2.102.gff3.gz).
The gene annotation track was lifted over to ARS-UCD1.2_Btau5.0.1Y using liftOff v1.5.2 pipeline [13]. The distribution of CAGE tags within genomic features is shown in Figure 1A. The high proportion of CAGE tags mapping to intronic and intergenic tags relates to the lack of annotation for alternative isoforms and novel genes in the current annotation of ARS-UCD2.1. The transcriptome map BovReg will generate for D2.3, integrating the CAGE dataset with the other functional information generated for BovReg in work package 2 will remedy this. Using the CAGE data more than 15,000 genes and 29,000 transcripts have been annotated per tissue (Figure 1B). These results are very similar to those generated by [1] who used 5'-sequencing technologies to map TSS in two male and two female Holstein-Friesian cattle.



The total number of genes and transcripts annotated using CAGE tags in all tissues.

| Tissue | Genes | Transcripts |
|---|---|---|
| adrenal gland cortex | 15140 | 29233 |
| cerebellum | 14722 | 28442 |
| cerebrum cortex | 14977 | 28860 |
| colon | 15184 | 29316 |
| duodenum | 15048 | 28963 |
| heart | 14607 | 28174 |
| hypothalamus | 13665 | 25654 |
| ileum | 15288 | 29338 |
| jejunum | 14961 | 28801 |
| kidney | 14943 | 28852 |
| liver | 14504 | 27949 |
| lung | 15337 | 29547 |
| lymph node | 14399 | 27695 |
| mammary gland | 13504 | 25523 |
| ovary | 13736 | 25974 |
| pancreas | 13218 | 24921 |
| pituitary gland | 14807 | 28515 |
| rumen | 14648 | 28161 |
| skeletal muscle | 12994 | 24603 |
| spleen | 15011 | 29052 |
| subcutaneous fat | 13381 | 25660 |
| testis | 14367 | 27358 |
| thoracic branch lymph | 13636 | 26137 |
| thyroid gland | 14586 | 28193 |
| uterine horn | 14268 | 27284 |

**Figure 1. A: The average percentage of CAGE tags mapping to each genomic region; B: the total number of genes and transcripts annotated using CAGE tags for each tissue.**

### Data and Code Availability

The raw data and BAM files have been uploaded to the ENA accession number PRJEB43235 and are accessible to all partners in BovReg via the private data portal.

UEDIN published the analysis pipeline recently [2] and it is available via https://msalavati@bitbucket.org/msalavat/cagewrap_public.git. They are currently working to transfer this pipeline to Nextflow to facilitate sharing with other partners in the BovReg consortium and the other EURO-FAANG projects e.g. [14].

## 4. Conclusions

For this deliverable CAGE libraries have been generated from 109 bovine tissue samples and provided as BAM files. These BAM files are accessible to all partners in BovReg via the private data portal. UEDIN is currently performing clustering of CAGE tags to annotate TSS across tissues. The results of the clustering will then be used for D2.3 to provide a high-resolution map of the transcribed regions of the cattle genome. Based on preliminary analysis of TSS annotated in the ARS-UCD1.2 genome, the map of transcribed regions to be generated within BovReg will be a significant improvement on the current gene annotation information which is available for cattle.

## 5. References

1. Goszczynski DE, Halstead MM, Islas-Trejo AD, Zhou H, Ross PJ. (2020) Transcription initiation mapping in 31 bovine tissues reveals complex promoter activity, pervasive transcription, and tissue-specific promoter usage. bioRxiv. 2020.09.05.284547. Available from: http://biorxiv.org/content/early/2020/09/06/2020.09.05.284547.abstract
2. Salavati M, Caulton A, Clark R, Gazova I, Smith TPL, Worley KC, et al. (2020) Global Analysis of Transcription Start Sites in the New Ovine Reference Genome (Oar rambouillet v1.0) Front. Genet. p. 1184. Available from: https://www.frontiersin.org/article/10.3389/fgene.2020.580580
3. Takahashi H, Kato S, Murata M, Carninci P. (2012) CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods Mol Biol. 786:181–200. Available from: https://pubmed.ncbi.nlm.nih.gov/21938627
4. Hannon Lab. (2017) FASTX-Toolkit FASTQ/A short reads pre-processing tools.
5. Lassmann T. (2015) TagDust2: a generic method to extract reads from sequencing data. BMC Bioinformatics. 6:24. Available from: https://pubmed.ncbi.nlm.nih.gov/25627334
6. Bertin N, Mendez M, Hasegawa A, Lizio M, Abugessaisa I, Severin J, et al. (2017) Linking FANTOM5 CAGE peaks to annotations with CAGEscan. Sci Data. 4:170147. Available from: https://doi.org/10.1038/sdata.2017.147
7. Langmead B, Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–9. Available from: https://doi.org/10.1038/nmeth.1923
8. Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–2. Available from: https://pubmed.ncbi.nlm.nih.gov/20110278
9. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 26:2204–7. Available from: https://pubmed.ncbi.nlm.nih.gov/20639541
10. Thodberg M, Sandelin A. (2019) A step-by-step guide to analyzing CAGE data using R/Bioconductor. F1000Research. 8:886. Available from: https://pubmed.ncbi.nlm.nih.gov/31327999
11. Pages H. (2020) BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. R package version 1.56.0. Available from: https://rdrr.io/bioc/BSgenome/
12. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. (2013) Software for computing and annotating genomic ranges. PLoS Comput Biol. 9:e1003118–e1003118. Available from: https://pubmed.ncbi.nlm.nih.gov/23950696

13. Shumate A, Salzberg SL. (2020) Liftoff: accurate mapping of gene annotations. Valencia A, editor. Bioinformatics. btaa1016. Available from: https://doi.org/10.1093/bioinformatics/btaa1016

14. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. (2020) The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 38:276–8. Available from: https://doi.org/10.1038/s41587-020-0439-x