Transcriptome analysis of crossbred and purebred cattle using CAGE and mRNA-Seq for the BovReg Project

Mazdak Salavati[1], Richard Clark[2], Doreen Becker[3], Christa Kühn[3,4], Graham Plastow[5], Gabriel Costa Monteiro Moreira[6], Carole Charlier[6,7] and Emily L. Clark[1] on behalf of the BovReg consortium

1 The Roslin Institute, University of Edinburgh, Edinburgh, UK,;2 Genetics Core, Edinburgh Clinical Research Facility, The University of Edinburgh, Edinburgh, UK; 3 Institute of Genome Biology Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany;4 Faculty of Agricultural and Environmental Sciences, University Rostock, Rostock, Germany; 5 Livestock Gentec, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Canada; 6 Unit of Animal Genomics, GIGA Institute, University of Liège, Liège, Belgium; 7 Faculty of Veterinary Medicine, University of Liège, Liège, Belgium.

Introduction:

Mapping of transcription start sites (TSS) in multiple tissues is a key first step in understanding transcript regulation and diversity and how these might influence phenotypic plasticity in cattle. TSS mapping can provide information about complex promoter activity, pervasive transcription and tissue-specific promoter usage. To this aim we utilised tissues (24 tissue types, n=105 total samples) (Table 1) from 3 diverse/non-reference populations (beef, dairy and beef-dairy cross) of cattle (Figure 1) to maximise our understanding of TSS complexity using CAGE-sequencing. From the same cohort of animals 121 mRNA-Seq libraries were prepared in order to create a *de novo* assembled transcriptome.

A Belgian Holstein **B** Charolais X German Holstein F2 **C** Kinsella composite (Canada)

Figure 1. Three cattle populations (1 male, 1 female, n=6 per population). A: 2 Belgian Holstein calves (Dairy), B: 2 Charolais X Holstein F2 (Beef-Diary cross) and C: 2 Kinsella composites (Beef composite).

Methods:

Analysis of the mRNA-Seq datasets was performed by University of Liège using an nf-core /rnaseq pipeline: https://github.com/BovReg/rnaseq.git. Analysis of CAGE sequence data analysis was performed using the analysis pipeline described in Salavati *et. al* 2020. CAGE libraries were analysed using the CAGEfightR uni-directional (TSS) and bi-directional (TSS-Enhancer) clustering algorithms described by Thodberg et al. 2019. For this study the analysis pipeline was transferred to NextFlow/DSL2 : https://github.com/MazdaX/nf-cage.git .

Table1. Tissues and number of replicates used in the CAGE study (n=105 samples).

Tissue	Male Belgium calf Holstein	Female Belgium calf Holstein	Male Charolais x German Holstein F2	Female Charolais x German Holstein F2	Male KC composite	Female KC composite
adrenal gland cortex	Yes	Yes	Yes	Yes	Yes	Yes
cerebellum	Yes	Yes	Yes	No	No	Yes
cerebrum cortex	Yes	Yes	Yes	Yes	Yes	No
colon	Yes	Yes	Yes	Yes	Yes	Yes
duedenum	Yes	Yes	Yes	Yes	Yes	Yes
heart	Yes	Yes	Yes	No	Yes	Yes
hypothalamus	Yes	Yes	No	No	No	No
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	No	Yes	No	Yes	No	Yes
ovary	No	Yes	No	No	No	Yes
pancreas	Yes	Yes	Yes	Yes	Yes	Yes
pituitary gland	No	Yes	No	No	Yes	Yes
rumen	Yes	Yes	Yes	Yes	Yes	Yes
skeletal muscle	Yes	Yes	No	No	No	No
spleen	Yes	Yes	Yes	Yes	Yes	Yes
subcutaneous fat	Yes	Yes	No	No	No	No
Testis	Yes	No	Yes	No	No	No
thyroid gland	Yes	Yes	Yes	Yes	No	No
uterus	No	Yes	No	Yes	No	Yes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668

Disclaimer: the sole responsibility of this presentation lies with the authors. The Research Executive Agency is not responsible for any use that may be made of the information contained therein.

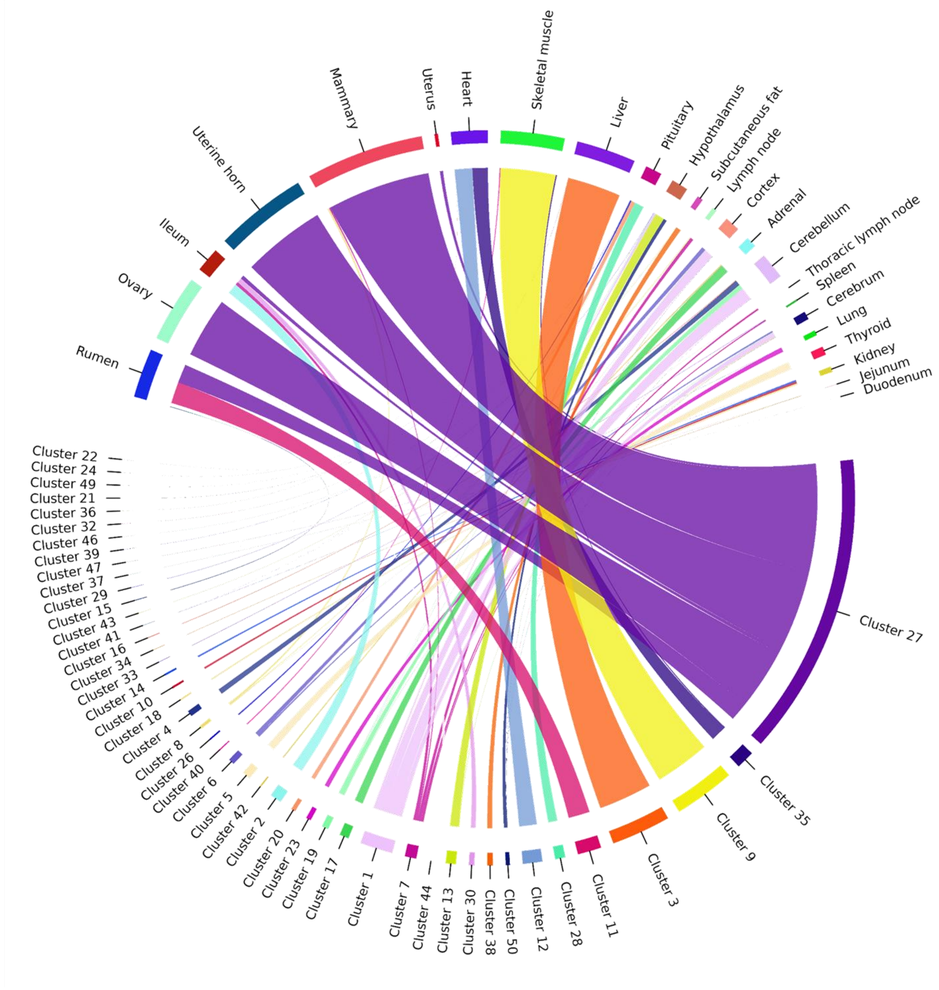


Figure 2. The Graphia network analysis of the mRNA-Seq data (23 tissue types, n=121). The input matrix was produced using the *de novo* assembled transcriptome using the nf-core/rnaseq pipeline. The transcript level counts for each tissue (Kallisto output TPM) were used for the MCL clustering (correlation threshold min 0.75 max 0.9; MCL granularity 2.0 and kNN edge reduction algorithm [multiplicity >5]). The MCL clustering resulted in a graph network of 51 components comprised of 13.8k nodes (genes) and 56k edges (correlations). The top 10 components are numbered 1-10 in bold. Each cluster is labelled according to the tissue with the highest expression level.

NB. The testis tissue was removed from this analysis.

Figure 3 - Chord diagram of the top 50 co-expressed gene clusters in different tissues types from Graphia analysis of the mRNA-Seq dataset. The expression level (TPM) of all genes within each cluster was averaged per tissue. The width of the chords represents the average TPM of all genes present within the cluster. The top half of the plot shows different tissue types while the bottom half depicts co-expressed gene clusters in the order of average TPM.
Cluster 27 showed high levels of average transcript expression in reproductive tissues. There were also tissue specific clusters e.g. for liver (3), skeletal muscle (9) and heart (35 & 12).

NB. The testis tissue was removed from this analysis.

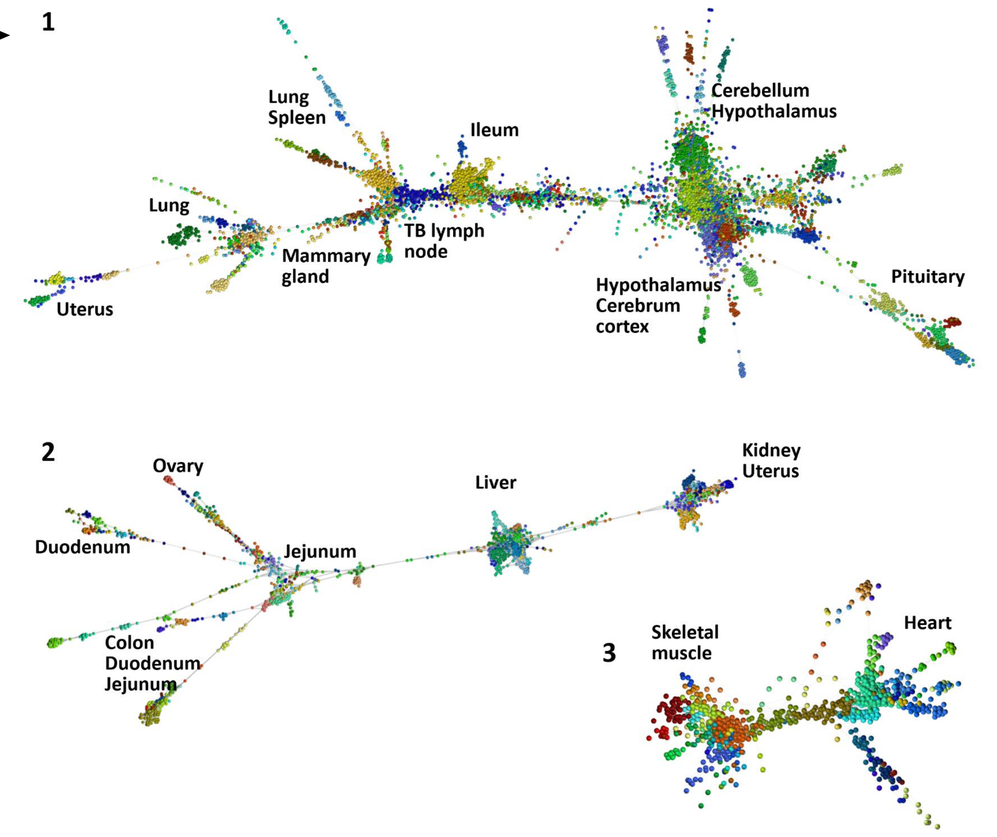


Figure 4. Dimension reduction of all putative TSS and TSS-Enhancers CAGE clusters identified in the dataset among all tissues and 3 populations. The GI and CNS organ system tissues showed a distinct profile along with tissues such as liver.

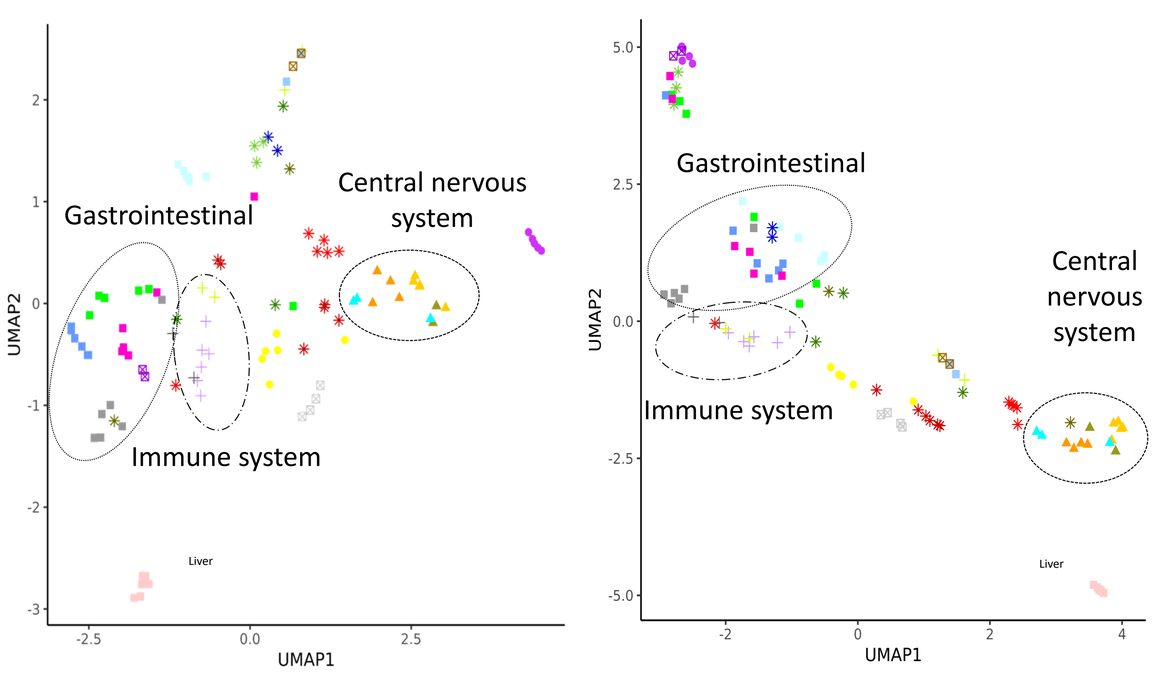
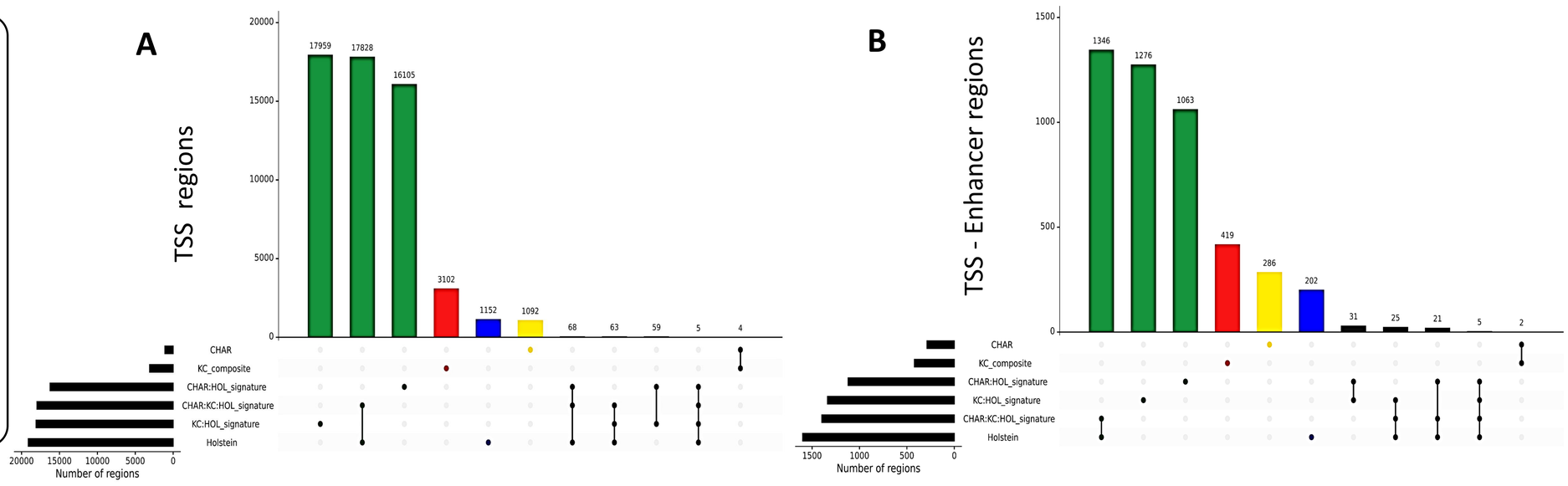


Figure 5. Intersection analysis of (A) TSS and (B) TSS-Enhancer CAGE clusters private to or shared between the 3 populations. A Holstein signature is shown in **green** shared with Charolais x Holstein and KC cattle. The population specific (i.e., not present in any tissue of another population) CAGE clusters are shown in **red** (KC), **blue** (HOL) and **yellow** (CHAR). HOL: Holstein; CHAR: Charolais; KC: Kinsella Composite.



Summary:

- The *de novo* transcriptome generated using the mRNA-Seq dataset captured 38,025 genes (12,868 novel compared to Ensembl v104 and 9,794 novel compared to NCBI v104).
- The top 50 clusters of co-expressed genes captured tissue-specific and shared expression profiles across all 23 tissue types (Figures 2 & 3).
- Clustering of tissues was largely similar between the CAGE and mRNA-Seq datasets (Figures 2 &4).
- Population specific TSS and TSS-Enhancers showed that cross-bred populations of cattle in this study have higher TSS diversity relative to pure bred populations (Figure 5).
- The Holstein CAGE cluster signature was dominant in the other two populations (> 80%) (Figure 5).
- Kinsella composite had the highest number of population specific TSS (3102) and TSS-Enhancer (419) CAGE clusters across 24 tissue types (Figure 5).
- These datasets will provide a new high resolution annotation for ARS.UCD1.2._BtauY