# A MAP OF BOVINE LONG NON CODING RNA ACROSS 18 TISSUES

**Lambros T. Koufariotis[1,2,3], Yi-Ping Phoebe Chen[1], Amanda Chamberlian[2], Christy Vander Jagt[2], Ben J. Hayes[1,2,3]**

[1]College of Science, La Trobe University, Melbourne, VIC 3086, Australia
[2]Department of Environment and Primary Industries, AgriBio, Bundoora, VIC 3086, Australia
[3]Dairy Futures Co-operative Research Centre, 5 Ring Road, Bundoora, VIC 3086, Australia

## SUMMARY

Long non-coding RNA (lncRNA) are common elements in vertebrates and other lesser organisms that possess numerous regulatory and cellular roles. Long ncRNA are well characterized in humans and mice, however in other species, there is comparatively little information of these elements. Identifying lncRNA in bovine could aid in identifying additional sites in the genome where mutations are likely to contribute to variation in complex traits along with understanding the evolutionary importance and constraints of these transcripts. This is important in bovine, since genomic predictions are increasingly used for genetic improvement of milk and meat production. We address the main challenge in identifying lncRNA, namely distinguishing lncRNA transcripts from unannotated genes, by developing a strict lncRNA filtering pipeline. Our aim was to identify and annotate novel lncRNA transcripts in the bovine genome captured from RNA Sequencing (RNA-Seq) data across 18 tissues, sampled in triplicate. We find 9,886 transcripts passed strict filtering criteria and show moderate to high expression. Further we find many unique lncRNA transcripts are downregulated in a tissue specific manner. This study also identified a large number of novel unknown transcripts in the bovine genome, many having high protein coding potential, indicating a clear need for better annotations of protein coding genes in the bovine genome.

## INTRODUCTION

The mammalian genome is highly complex, with protein coding genes considered some of the most important elements within the genome, however these only account for only a small portion of the entire transcriptome. It has recently been revealed that about 1-2% of the human genome is transcribed to messenger RNA (mRNA) (Frith *et al*. 2005) and up to 50% of the transcribed genome does not align to known protein coding regions (Hung and Chang. 2010). It is hypothesized that these non-protein coding RNA can either be transcriptional artifacts due to RNA Polymerase II errors in elongation (Van Bakel *et al*. 2010) or non-coding RNA (Kapranov *et al*. 2010). Evidence is accumulating for the later hypothesis, with studies across a range of species, including humans (Cabili *et al*. 2011), mouse (Dinger *et al*. 2008) and bovine (Qu and Adelson. 2012, Weikard *et al*. 2013) finding many novel ncRNA across a range of tissues.

Recent advances in transcriptome sequencing has allowed for the discovery of a new class of non-coding RNA transcripts that are surprisingly long, known as long noncoding RNA (lncRNA) (Marques and Ponting. 2014). Long noncoding RNA are classified as having an arbitrarily defined length of more than 200 nucleotides with weak or no protein coding potential and generally have lower expression levels than mRNA (Marques and Ponting. 2014). Functions of lncRNA are quite diverse, but some of the better studied lncRNA have described functions in regulating and guiding epigenetic marks and gene expression. These elements are coded almost anywhere in the genome including intergenic regions (also known as long intergenic ncRNA (Qu and Adelson. 2012). One of the best studied examples is *Xist*, a gene responsible for facilitation of imprinting the X chromosome that is in fact a lncRNA (Clemson *et al*. 1996).

While there have been a few studies in bovine isolating novel lncRNA (Weikard *et al*. 2013,

Billerey *et al*. 2014) there is still comparatively little information when compared to the repertoire of lncRNA found in human and mouse genomes. In this study we describe a comprehensive catalogue of putative bovine lncRNA expressed in 18 tissues and located within intergenic regions. Given the main challenge in identifying lncRNA is distinguishing them against transcripts from unannotated genes, we used stringent filtering methods to discriminate potentially protein coding RNA from ncRNA, acknowledging that the stringent filters may discard some true lncRNAs. We also compared our putative lncRNA to catalogues from mouse and human, to gain insights into the evolution of lncRNA across species. This information is of particular value since mutations that might be found within these lncRNA elements can potentially contribute to variations in complex traits.

## MATERIALS AND METHODS

**RNA extraction, tissue sampling, sequencing and alignment.** The tissues used in this study include: adrenal gland, black skin, white blood cells, caudal lobe of brain, brain cerebellum, heart, kidney, leg muscle (semimembranosus), liver, lung, intestinal lymph node, mammary gland, ovary, spleen, thymus, thyroid, tongue and white skin.

The quality control, filtration, read alignment to the reference genome and generation of the SAM files for the 18 tissue samples were performed as described in another study (Chamberlain *et al*. 2014).

**Finding intergenic long noncoding RNA.** We used a Cufflinks/Cuffmerge/Cuffcompare pipeline to assemble transcripts for all three technical replicates in each tissue sample to the Ensemble reference gene set release 75. Entries that had a class code of either "u", (unknown intergenic transcript), or "x", (exonic overlap with the reference genome but on the opposite strand) were extracted and kept for further analysis. Similar to (Weikard *et al*. 2013) we used Cuffcompare to compare our transcripts to those in the NCBI iGenomes repository to filter out transcript with protein sequences, giving us a total of 47,117 transcripts with unknown annotations. We used the UCSC utility twoBitTofa to obtain the nucleotide sequences for the transcripts.

**Long non-coding RNA filtering pipeline.** To find transcripts most likely to be noncoding RNA transcripts, we developed a 3 stage pipeline to filter out the transcripts that had a high chance of having protein coding potential.

*Stage 1. ORF Analysis.* getorf from the EMBOSS software package was used to find all possible open reading frames (ORF) in all directions of the transcript. We performed a blastp search on all ORF sequences to determined possible protein coding domains using an E-value of 1e-06 as cut-off. If no significant sequence matches were determined then the transcript was considered to be a potential lncRNA.

*Stage 2. Blastx.* We determined if our transcripts had any significant matches with protein sequences by using the tool blastx. An E-value of 1e-06 was used as cut-off. Only transcripts that did not show any significant matches with known protein coding sequences were considered.

*Stage 3. CPC Tool.* The third stage used the tool Coding Potential Calculator (CPC) which predicts the coding and noncoding potential of a transcript. We selected for transcripts as potentially noncoding if they have a score of $< -0.5$.

**Read counts, filtering of low read counts and differential expression analysis.** Read counts were obtained using the tool HTSeq and was run with default parameters only specifying for non-stranded (--stranded=no) and union mode (--mode=union) to get the counts matrix for each unkown transcript across all tissues and replicates. The final counts matrix file was used as input for the tool EdgeR for normalization and for filtering transcripts that had very low read counts (read count $<25$ across all three replicates for each tissue).

Differential analysis was carried out by performing a t-test for each tissue with all the other

tissue samples. The standard error was calculated by subtracting the mean across all tissue with the mean for each tissue. If the t-test had a *P*-value of <0.05 and a positive standard error, then the transcript was considered to be upregulated. If the t-test had a *P*-value <0.05 and a negative standard error then that transcript was considered to be downregulated. All other transcripts were considered to have no differential expression.

**Homology analysis with ncRNA in human and mouse.** Human and mouse ncRNA were obtained from; GENCODE v7, NONCODE v4 and lncRNAdb databases. A blastn search was performed using an E-value of 1e-06 to blast the unknown transcripts with the human or mouse databases. From this we extracted the transcripts that had significant matches with a known lncRNA.

## RESULTS AND DISCUSSION

After transcript assembly and annotation of RNA-Seq reads a total of 47,117 transcripts that aligned to the genome but did not align to protein coding genes or had protein coding annotations were found. These assembled transcripts were passed through the filtering pipeline to determine coding or noncoding potential. We defined putative lncRNA only if the transcripts passed all 3 stages of the filtering pipeline (methods) and had moderate to high expression levels after filtering for low read counts with EdgeR. A total of 9,886 putative lncRNA passed all three filters and were considered for further analysis.

We find that tissues involved in similar organ functions share very similar expression of putative lncRNA. These correlations are lower that what we find in the protein coding genes from the same datasets (Chamberlain *et al*. 2014). The expression patterns of our putative lncRNA show that 37% are downregulated, while 4% are upregulated and 59% show no differential expression (Figure 1).

The vast majority of the lncRNA are found to be within intergenic regions of the genome, however we do find a total of 1,501 lncRNA (about 15% of total lncRNA) that are located either near the 5' or 3' end of protein coding genes or are located within 5 kilobases upstream or downstream of protein coding genes. Due to the lack of stranded information, it is difficult to attempt to identify independently coded transcripts that are coded in the opposite direction of the neighbouring gene. Therefore we measured the concordance of expression between the lncRNA transcript and the neighbouring protein coding gene. A Pearson's and Spearman's rank correlation analysis showed that many lncRNA had high correlations with their neighbouring genes, and therefore could be unannotated exons, however a significant minority show no correlations, these may indicate independently coded transcripts.

**Comparative analysis with human and mouse lncRNA**. To identify putative lncRNA that show sequence conservation we performed a blastn search between our lncRNA and the lncRNA in both human and mouse lncRNA databases. Of the 9,886 lncRNA, only 289 show significant sequence similarities with known human lncRNA and 119 show significant sequence similarities with known mouse lncRNA. Further, only 36 putative lncRNA show sequence similarities with both a human and mouse lncRNA. Long ncRNA were also compared to other bovine lncRNA found in similar studies using either pigmented or non-pigmented skin cells (Weikard *et al*. 2013) or bovine muscle cells (Billerey *et al*. 2014). Of the catalogue of lncRNA in the skin cells we find 848 (out of 4,948) lncRNA that overlap with our catalogue of lncRNA. Of the 584 lncRNA found in muscle cells, we find a total of 129 that overlap with our lncRNA. Due to the fact that lncRNA are tissue specific and also can be expressed in different developmental stages we acknowledge that these catalogues provide valuable information of potential lncRNA in the bovine genome. Further, studying these regions will assist in finding new classes of genes that, while lacking the ability to code for proteins, can have mutations that could potentially affect complex dairy traits of interest, such as milk volume, fat percent, protein percent and mammary system.

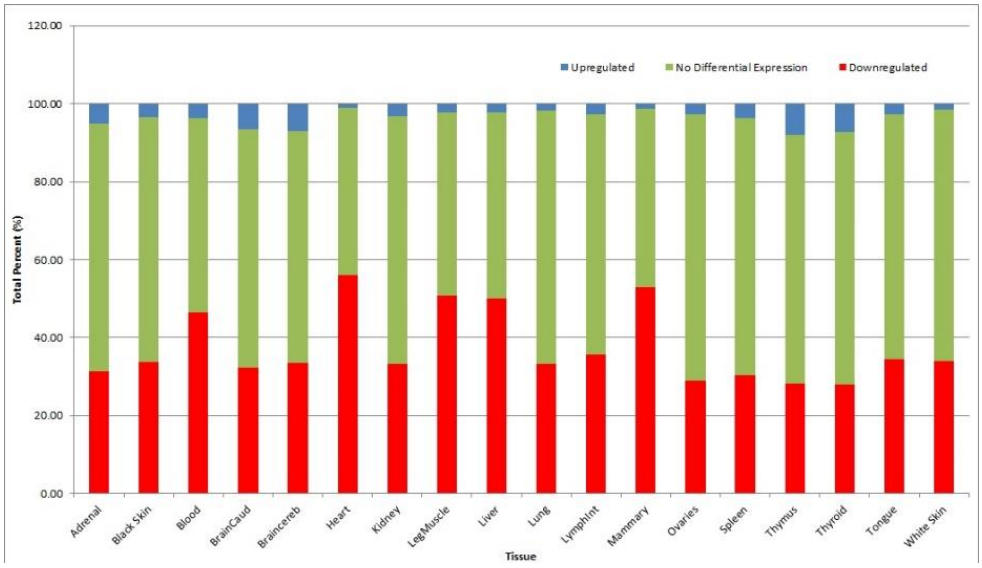


**Figure 1. Percent of lncRNA that are upregulated, downregulated or not differentially expressed**. Red bars indicate percent of downregulated lncRNA for each tissue. Blue bars indicate percent of upregulated lncRNA for each tissue. Green bars indicate no differential expression.

## REFERENCES


Billerey C., Boussaha M., Esquerre D., Rebours E., Djari A., Meersseman C., Klopp C., Gautheret D. and Rocha D. (2014) *BMC genomics* **15**:499.

Cabili M.N., Trapnell C., Goff L., Koziol M., Tazon-Vega B., Regev A. and Rinn J.L. (2011) *Genes & Development* **25(18)**:1915-1927.

Chamberlain A., Jagt C.V., Goddard M. and Hayes B. (2014). A gene expression atlas from bovine RNAseq data. *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*.

Clemson C.M., McNeil J.A., Willard H.F. and Lawrence J.B. (1996) *The Journal of cell biology* **132(3)**:259.

Dinger M.E., Amaral P.P., Mercer T.R., Pang K.C., Bruce S.J., Gardiner B.B., Askarian-Amiri M.E., Ru K., Solda G., Simons C., Sunkin S.M., Crowe M.L., Grimmond S.M., Perkins A.C. and Mattick J.S. (2008) *Genome Research* **18(9)**:1433-1445.

Frith M.C., Pheasant M. and Mattick J.S. (2005) *European journal of human genetics EJHG* **13(8)**:894-897.

Hung T. and Chang H.Y. (2010) *RNA biology* **7(5)**:582-585.

Kapranov P., St Laurent G., Raz T., Ozsolak F., Reynolds C.P., Sorensen P., Reaman G., Milos P., Arceci R., Thompson J. and Triche T. (2010) *BMC Biology* **8(1)**:149.

Marques A.C. and Ponting C.P. (2014) *Current opinion in genetics & development* **27c**:48-53.

Qu Z. and Adelson D.L. (2012) *PloS one* **7(8)**:e42638.

Van Bakel H., Nislow C., Blencowe B.J. and Hughes T.R. (2010) *PLoS biology* **8(5)**:e1000371.

Weikard R., Hadlich F. and Kuehn C. (2013) *BMC genomics* **14**:789.