Identification and characterization of developmental, tissue, and evolutionary transcription start sites in Bos taurus indicus

Mehrnush Forutan (mehrnush.forutan@gmail.com)
University of Queensland https://orcid.org/0000-0002-8145-1035
Elizabeth Ross
The university of Queensland
Amanda Chamberlain
AgriBio
Loan Nguyen
The university of Queensland
Brett Mason
Agriculture Victoria
Stephen Moore
University of Queensland
Ruidong Xiang
The University of Melbourne
Ben Hayes
University of Queensland https://orcid.org/0000-0002-5606-3970

Article

Keywords: CAGE, Developmental stage, Evolutionary divergence, Functional annotation, Heat stress genes, promoter, Tissue-specific, Transcription start site

Posted Date: September 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-50937/v1

Version of Record: A version of this preprint was published at Communications Biology on July 1st, 2021.
See the published version at https://doi.org/10.1038/s42003-021-02340-6.

Abstract

To further the understanding of the evolution of transcriptional regulation, we profiled genome-wide transcriptional start sites (**TSSs**) in two sub-species, *Bos taurus taurus* and *Bos taurus indicus*, that diverged approximately 500,000 years ago. Evolutionarily divergent TSSs were observed in more than half of the genes expressed across the sub-species, ranging from extreme cases in which a TSS was observed only in one sub-species to intermediate situations in which a corresponding TSS had been translocated by > 50 nucleotides, to situations where the number of TSS differed between the sub-species. Fetal and adult stages not only had their own regulatory profile of active and inactive genes but also their own pattern of TSSs. Given *indicus* are more adapted to heat, we also specifically investigated TSSs for heat shock proteins. More variation was observed in number of TSSs for heat shock proteins in *indicus* than *taurus*. This study confirmed that most genes are regulated in a tissue-specific manner.

Introduction

Understanding how expression of genes are regulated is one of the essential goals of genomics. Gene regulation mechanisms also contribute to evolutionary processes associated with within species divergence such as breed formation[1]. Variations in gene expression is largely due to cis mechanisms where regulatory molecules bind to elements such as promoters and enhancers close to genes to initiate transcription. Transcription Start Sites (**TSSs**) act as an integration region for a wide range of molecular signals to control transcription and expression levels[2–5]. Previous studies[6, 7], assumed that promoters have a TATA-box, which directs the positioning of the preinitiation complex, in effect initiating transcription from a single nucleotide. In contrast, more recent studies[8] have shown that the majority of human and mouse RNA Polymerase II core promoters have an array of close TSSs instead of the expected single TSS. In agreement with this finding, the FANTOM consortium project highlighted that few genes are true 'housekeeping' (with one TSS), whereas many mammalian promoters are composed of several closely separated TSSs[9]. Furthermore, a large number of genes have several strong core promoters, which force alternative splicing and ultimately, production of different protein isoforms[8, 10]. There is evidence in human genes that different isoforms are produced as a result of the usage of alternative TSSs[11, 12].

Several efficient technologies have been developed to dissect gene regulation mechanisms recently, eg. ChIP-seq[13] and ATAC-seq[14]. These technologies detect sites across the genome which can be used to infer regulatory elements including promoters and enhancers. These technologies are also used as the primary tools for consortia to annotate human (ENCODE [15]) and animal (FAANG[16]) genomes. The FANTOM consortium has focused on the mapping of TSSs using Cap Analysis of Gene Expression (**CAGE**) to identify promoters and enhancers across a large collection of primary cell types in the human and mouse genomes[9, 17].

CAGE has been developed as one of the main high-throughput assays for studying TSSs and their expression[18]. Sequencing short reads (or tags) from the 5'end of full-length cDNA allows TSSs to be mapped and their expression to be studied. A specific advantage of the CAGE method is that reads mapped to the genome provide accurate location of TSS and quantify transcription[8, 19]. As CAGE tags can be aligned to a reference genome without the need for transcript annotations, it can detect not only TSSs of known mRNAs but also mRNA from alternative TSSs that might often be tissue or developmental stage specific[20].

It has been well established that changes in transcriptional regulation underlie much of the phenotypic variation between species[21], and there is numerous evidence for gene expression divergence between even closely related lineages (e.g., [22]). Previous research has shown that transcription factor binding sites[23], centromeres[24], and TSSs[25] are affected by gain-and-loss of functional genetic elements, called "turnover". Thus far, TSS locations have mostly been explored between human and mouse (e.g., [25, 26]), which diverged approximately 96 million years ago[27]. In this study we exploit a much closer evolutionary split between *Bos taurus taurus* and *Bos taurus indicus* using CAGE data sets for the first time to assess changes in TSSs for closely related (cattle) species, with the aim of gaining new insights into the evolution of TSSs. CAGE-Seq (CAGE followed by sequencing) was performed on nine tissues at adult stages, including liver, lung, kidney, thyroid, spleen, muscle, uterus in *indicus* and liver, spleen, muscle, mammary, heart in *taurus* sub-species, and two tissues in adult and fetal stages, including liver and lung in *indicus* and liver in *taurus* sub-species. This paper highlights rapid evolutionary divergence in TSS usage and is the first bovine TSS discovery study using CAGE-Seq data.

Results And Discussion

Evolutionary divergence in TSS usage comparing *Bos taurus* and *Bos indicus*

Firstly, statistical analysis of the reproducibility of TSSs from CAGE within and across sub-species when different number of biological replicates (1 or 2), coverage of samples (total or half tags), and minimum tag count thresholds (5, 10, 15, 20 and 25) for distinct positions were investigated for TSS calling. For this purpose, fetal lung, adult liver and lung, respectively with low, medium and high sample coverage were randomly divided into two sub-samples and TSSs called in each. The TSS called in each half sample were then compared. Based on these results, using a fixed minimum tag count threshold irrespective of the number of replicates and sample coverage will result in inconsistency in number of TSSs obtained in each tissue and make the comparison between tissues difficult (Supplementary Table 1). To improve signal confidence while taking into account the effect of number of replicates and sample coverage, the minimum tag count threshold (t_i) was set using the formula $N_{max} \times \frac{TPM_i}{N_i}$, where N_{max} is the number of biological replicates for a tissue having the highest replicates among the all tissues, e.g here $N_{max} = 3, TPM_i$is total tags per million CAGE tag counts in sample *i*, N_i is the

number of biological replicates for sample *i*. Also, for samples with high coverage such as kidney and lung (t_i > 25), to avoid missing a noticeable number of true TSSs, the t_i was set to 25 tag counts. A minimum reproducibility of 76%, measured as the fraction of genes with the same number of TSSs in the total and half samples, was obtained when the coverage was reduced to less than one million CAGE tag counts for fetal lung (Supplementary Table 2). This could be a result of the reduction of t_i to three tags, and therefore it resulted in reducing the signal confidence in fetal lung sub-samples. In total, based on these results, the reproducibility of TSS is principally a consequence of the depth of sampling in individual replicates. Considering the relatively good sample coverage in the current study, the average reproducibility across samples was higher than 80%.

To study TSS evolution, we compared TSS positions and the distribution of the cluster of tags within TSSs across three tissues, including liver, muscle, and spleen in adult *Bos taurus* and *Bos indicus*. The number of TSSs in each tissue is shown in Table 1. The genes were divided into four groups based on the differences in the number of TSSs: 1) genes with single TSSs in both sub-species, 2) genes with multiple TSSs in both sub-species, 3) genes with divergent single/multiple TSSs across sub-species, *i.e.* a single TSS in *taurus* and multiple TSSs in *indicus* or vice versa, and 4) extreme cases in which a single or multiple TSSs were observed in one sub-species but no TSS were observed in the other. In total, 20% (liver), 43% (muscle), and 30% (spleen) of all genes were in the first group, while only a minority of genes (3–9%) belonged to the second group. Interestingly, about 12% (spleen and muscle) and 10% (liver) of the total genes had divergent single/multiple TSSs. The fourth group of extreme cases in which single or multiple TSSs were observed in one species but no TSS were observed in the other ranged between 77% and 48% of genes in liver and muscle, respectively.

Table 1
Summary of the number of TSSs for each sample along with the biological replicate, number of CAGE tags and number of genes with single or multiple TSSs

Tissue	Biological replicate	CAGE tags	TSSs			Genes with TSSs		
			Total	In known genes	In promoter, proximal, and 5' UTR of known genes	Total	Single TSS	Multiple TSSs
Bos indicus								
Adult stage								
Blood	1	2688190	5974	5733	5389	4352	3549	803
Kidney	1	13530755	14190	13469	12186	8449	5774	2675
Muscle	1	4873320	7234	6916	6310	4843	3770	1073
Ovary	1	6298852	9680	9377	8595	6444	4854	1590
Spleen	1	3693393	8211	7994	7622	5814	4418	1396
	2	4077550						
Thyroid	1	11418581	8503	8283	7930	6030	4541	1489
	2	4077550						
Uterus	1	7478607	7125	6872	5956	4660	3717	943
Liver	1	5897856	11928	11421	10694	7804	5654	2150
Lung	1	24258474	15151	14586	12279	8611	6084	2527
Fetal stage								
Liver	1	7981054	7498	7164	6641	5293	4204	1089
Lung	1	1647510	11562	11199	10470	7640	5549	2091
Bos taurus								
Adult stage								
Heart	1	11874030	3355	3194	2958	2509	2150	359
	2	11795755						
Liver	1	6267975	5046	4864	4502	3704	3065	639
	2	6070584						
Mammary	1	21558896	2495	2359	2199	1949	1741	208

Tissue	Biological replicate	CAGE tags	TSSs			Genes with TSSs		
			Total	In known genes	In promoter, proximal, and 5' UTR of known genes	Total	Single TSS	Multiple TSSs
	2	17327550						
	3	8235522						
Muscle	1	3967633	6784	6523	5996	4729	3754	975
	2	4181101						
Spleen	1	7876366	5555	5331	5063	4275	3618	657
	2	9644352						
Fetal stage								
Liver	1	10777732	3595	3427	3153	2704	2330	374
	2	9998986						

Further analysis of genes with single TSSs in both sub-species confirmed that TSS positions on the genome were not fixed, and TSSs mostly (~ 75–92%) lay within < ± 10 nucleotides of the other (Fig. 1). However, about ~ 50% of TSS in genes with divergent TSSs across sub-species had a TSS within < ± 20–25 nucleotides in the other sub-species. A noticeable number of these genes (~ 30–40%) had TSSs that had been translocated by > ± 25 and < ± 50 nucleotides from the TSS in the other sub-species. This is consistent with previous evolutionary research between human and mouse that suggested that the signals encoding TSSs are highly flexible and evolvable[25]. In line with evolutionary research from human and mouse[23, 26] suggesting two types of evolutionary pathways lead to TSS turnover, we observed sliding of the TSSs along the genome and gradual shifting of usage from one TSS to an alternative TSS in the other sub-species. It seems that degradation of a weak TSS or an emerging new one over evolutionary time was likely responsible for observing the variability in the number of TSSs for a specific gene across the sub-species (Fig. 2, bar charts).

A previous study in humans and mice[8] classified TSSs into four "shape" categories using four different criteria, which were applied in a specific order. 1) A TSS was assigned as a single peak (SP) shape if the distance between the 75 and 25 tag density percentile within a TSS (i.e. interquartile range (IQR)) was less than 4 bp; 2) If the ratio between the first and second tag peak was > 2 and the TSS was not classed as SP, it was classified as a broad with dominant peak (PB); 3) If the TSS was not classified as SP or PB and there was one or more consecutive tag density 5th percentile pairs with a distance exceeding 10 bp a TSS was classified as bi- or multimodal (MU); 4) If none of the above applied, the TSS was classified as broad (BR). That study[8] used a genomic interval, the ratio between the first and second tag peak, and a

distance between the tags for the classifying TSS, but in the current study to better visualize the probability distribution of the cluster of tags within a TSS we classified TSS into four shape categories based on the genomic interval covered by the cluster of tags and the probability of the tags distribution within a single TSS. In the single dominant peak class (SP), the tags were concentrated to no more than four consecutive start positions. The clusters spanning a broader region (IQR > 4 bp) were further divided into three categories based on the shapiro test for normality and Hartigans' Dip test for unimodality in R packages: 1) If the P-value for shapiro test > 0.05 a TSS was classified as normal broad distribution (NB); 2) If the P-value for the Dip test was < 0.05 a TSS was classified as a bi- or multimodal broad distribution (MB); 3) If none of the above applied, the TSS was classed as a general broad distribution (BR). Generally, in both sub-species, the first and second highest fraction of TSSs were observed in SP and BR shape groups, and *taurus* in comparison with *indicus* tended to have higher frequency of SP TSSs except for genes with divergent TSS in spleen and muscle in which *taurus* and *indicus* had single and multiple TSSs, respectively (Fig. 3). Furthermore to visualize the TSS width, we plotted the frequency of TSSs based on the genomic interval covered by the cluster of tags within the TSS in genes with divergent and multiple TSSs across sub-species separately for each tissue (see Fig. 2 for histogram charts). Based on these results, TSSs in indicus tended to be wider, spread over a larger genomic region, in genes with divergent TSSs in which *indicus* and *taurus* sub-species had single and multiple TSSs, respectively (Fig. 2 left panel).

A crossover TSS event was identified if the dominant TSS, the TSS with the highest expression, switched between the two sub-species for a specific gene. To assess crossover switching events, the expression of TSSs was compared for genes having multiple TSSs in both sub-species. Our hypothesis is that differential TSS usage in different sub-species can result in significant regulatory changes. In total 56 (0.64%), 78 (1.14%), and 97 (1.67%) genes had crossover TSS switching events between *taurus* and *indicus* in liver, spleen, and muscle, respectively (Supplementary Table 3). Significantly enriched biological process (**BP**) and gene ontology (**GO**) terms for these crossover switching genes in the three tissues included *catabolic* and *metabolic process* in muscle and liver and *positive regulation of T cell activation* and *phospholipid binding* in spleen (P-value < 0.05). Based on these results, one reason for the noticeable heterosis which is observed in *Bos taurus* × *Bos indicus* crossbreds could be that there are more functional TSSs in crossbreds compared to the pure species. This mechanism for heterosis would have to confirmed by demonstrating a dominance effect on fitness.

Characterization of TSSs architecture in heat shock protein-related genes

Bos taurus breeds are best suited to sub-tropical and temperate regions. They have thicker coats that allow them to endure cooler winters, and they do not have the notable ‘hump’ of their *Bos indicus* relatives. *Bos indicus* cattle in contrast have large ears and dewlap, which help to keep them cool and are well-suited to tropical environments. With the hypothesis that heat shock proteins may be involved in this adaptation, ten heat shock protein-related genes such as *HSP* family genes, including *HSP32*, *HSP60*,

HSP70, *HSP90*, and *HSP105* (Supplementary Table 4) and two apoptotic-related genes (BAX, and BCL2) were analysed for variation in TSS expression and number, and the distribution of the tag clusters within TSS in adult tissues between and within the sub-species. In *indicus*, except for *BAX*, *BCL2*, *HSPA14*, and *HSPH1* which had single or multiple TSSs across some of the tissues and no TSS in the others, the remaining genes were expressed in all tissues. However, in *taurus* most of the genes had no TSS across all or some of the tissues, except for *HSPA5*, *HSP90B1*, and *HSP90AA1* (Supplementary Table 5). Generally, more variation was seen in the number of TSSs in *indicus* compared to *taurus* sub-species (t-Test; P-value < 0.05). A previous study [28] profiling the genome-wide TSSs in lungs of C57BL/6 mice 24 h after intratracheal instillation of the multiwalled carbon nanotube Mitsui-7, which is associated with increased stress and inflammatory response, revealed that many of the Mitsui-7-responsive TSSs were alternative TSSs for known genes, and the number of alternative TSSs used for a given gene was positively correlated with overall Mitsui-7 response. These researchers suggested that individual TSSs may simply be additive, or may have expression output limitations that make transcription of additional TSSs favourable, also they suggested that having multiple TSSs could enable a faster response to stimuli[28]. Similarly, HSP genes have been extensively reported to respond to external stress stimuli[29], so observing more variation in the number of TSSs in *indicus* sub-species may indicate their higher ability to respond to stress. Future research exploring a time series response could shed light on the latter hypothesis.

All the expressed HSP genes had a higher total TSS expression level in *indicus* compared to *taurus* sub-species (t-Test; P-value < 0.05). The highest TSS expression was observed for gene *HSP90B1* in both sub-species (Supplementary Table 6). A heat map of the transcript expression for the most common TSS (core TSS) in genes related to heat stress across the *indicus* adult tissues (Fig. 4A) and analysis of the core TSS shape (fraction of broad or sharp TSS for each gene across tissues; Fig. 4B) demonstrated association between TSS expression and peak shape in different genes. Core TSS, the most common TSS among all tissues, in the three highest expressed genes (*HSP90B1*, *HSP90AA1*, and *HMOX1*) were observed to have sharp, well defined TSS peaks (Fig. 4).

Variation in TSS number between developmental stages within each sub-species

Like any specific tissue in the body, the biological features of tissue in fetal and adult stages might be determined mainly at the level of gene expression and transcription. So differential and quantitative analysis of TSS expression and distribution patterns could be useful for the identification of developmental-stage-specific genes and facilitate our understanding of the mechanisms of coexistence of a gene's transcription pattern in fetal and adult stages. TSSs were detected in adult and fetus of two *indicus* and one *taurus* tissues, with 7,498 (11,928) and 11,199 (15,151) detected in fetal (adult) liver and lung *indicus* tissues and 3,427(4,864) in fetal (adult) *taurus* liver, respectively (Table 1). In the *indicus* and *taurus* liver, the proportion of TSSs that matched known genes was the same (~ 96%), and about 93% (94%) and 92% (93%) of those were located in the promoter, proximal region, or 5' UTR of genes at the

fetal (adult) stage in *indicus* and *taurus* liver, respectively. In the *Bos indicus* lung, 97% (96%) matched known genes, and 93% (84%) were located in the promoter, proximal region, or 5' UTR at the fetal (adult) stage. In the pilot Encyclopaedia of DNA Elements study, it was shown that there are long transcripts that can bridge genes or even span several genes, often starting in the middle of a gene structure[30]. A small proportion of TSSs observed within introns could be the result of recapping due to post-transcriptional modifications[30]. In the current study, only TSS located in the promoter, proximal, and 5' UTR regions of known genes were used for all analyses.

Developmental-stage TSS changes in *Bos taurus* and indicus liver

In both sub-species a higher number of genes had a single TSS in the fetal stage compared to the adult (Fisher's exact test, P-value < 0.0002). In total, 79% (72%) and 86% (83%) of genes in the fetal (adult) *indicus* and *taurus* liver had a single TSS, respectively (Fig. 5A, and B). This may be due to the use of less complex expression functions in the fetal stage compared to the adult stage. Among the genes with a single TSS in both fetal and adult stages *metabolic process*, *cellular process, gene expression*, and *cellular biosynthetic process* were ranked highest based on the gene ratio (i.e. proportion of genes associated with the BP GO term divided by the number of selected genes) in both sub-species (Bonferroni-corrected P-value < 0.05, Supplementary Table 7). Also, *nuclear division, mitotic cell cycle, mitosis, and organelle fission* were the common BP GO terms for genes with single TSS expressed only in *indicus* and *taurus* fetal liver (Bonferroni-corrected P-value < 0.05, Supplementary Table 8). This is in line with the specific features and functions of the fetal tissues, *i.e* cell proliferation and differentiation.

In the current study, about 8.7% and 12.5% of the total 4,250 and 9,215 expressed genes showed divergent single/multiple TSSs usage with aging in *taurus* and *indicus* liver, respectively. These results are in line with a previous study in human and indicate that TSS switching events are common and can play a significant role in development[11]. About 50–64% of genes with a single TSS expressed in fetal liver, showed an increase in the number of TSS with aging from single to multiple. One example of a gene in this category is *ATP6V0E1*, ATPase H + transporting V0 subunit e1. It mainly has function in hydrolase activity and proton transmembrane transporter activity. It is located on chromosome 20 in the region between 4732524 and 4762537 bp. Two TSSs were found in *taurus* and *indicus* adult liver at positions 4732518 and 4732488, however only position 4732518 was observed in *taurus* and *indicus* fetal liver. Considering that TSS expression at position 4732518 doubled from fetal to adult stage, it seems that the additional TSS 30 bp apart from the main TSS in adult tissue could have increased the transcription of the main TSS. Gene ontology analysis of genes with divergent TSSs across fetal and adult stages is shown in Supplementary Table 7. In the mammalian genome, the alternative usage of multiple TSSs is an important mechanism to increase mRNA and protein diversity from a single gene[31]. When comparing the number of TSSs between developmental stages in the liver, there were a number of cases (3-4.6%, Fig. 5A and B) where the number of TSS went from multiple to single TSS with aging. Genes in this

category were involved in *cellular* and *primary metabolic process*, and *cellular process* (only significant in *indicus* liver at Bonferroni-corrected P-value < 0.05, see Supplementary Table 7).

Furthermore, 34 and 94 genes of the total 4,250 and 9,215 expressed genes had crossover TSS switching events in *taurus* and *indicus* liver with aging, respectively, which suggests variation in the dominant TSS over time (Supplementary Table 9). Since the alternative mRNA isoforms could be translated into functionally different products, a crossover switching event may suggest that one gene can play different roles at different time points in development. The most significant BP terms related to this group was *catabolic process (*P-value < 0.05).

Developmental-stage TSS changes in *Bos indicus* lung

Lung tissue plays an important role in the respiratory system of mammals after birth. Before birth, the lung is full of liquid[32–34] and does not participate in gas-exchange because of high pulmonary vascular resistance and immature respiratory function[35]. Therefore, it is necessary for the lung to be sufficiently developed at birth to perform the function of gas-exchange, which requires numerous physiological changes to occur[36]. We performed quantitative and expression analysis of CAGE-Seq data in fetal and adult lung tissues of a single *Bos indicus* cow-fetus pair. A total of 9,784 genes were expressed across both stages. Similar to liver tissues, a higher number of genes had a single TSS in the fetal stage compared to the adult (Fisher's exact test, P-value < 0.005; Fig. 5C). In total, 73% (70%) of genes at the fetal (adult) stage had a single TSS in the lung. Similar to fetal liver tissues, genes in fetal lung with a single TSS were mainly involved in *mitosis*, *nuclear division*, and *cellular process* (Bonferroni-corrected P-value < 0.05, see Supplementary Table 8). *Immune response and response to stimulus* were the common significant BP GO terms for genes with a single or multiple TSSs expressed only in adult lung, respectively (Bonferroni-corrected P-value < 0.05, see Supplementary Table 8).

Similar to liver tissue, three patterns of TSS numbers per gene were seen in lung tissue between developmental stages (Fig. 5C). The first pattern was genes with multiple TSSs in the fetal stage, but just one TSS in the adult stage (e.g., *HSP90AA1*); the 2nd pattern was relatively constant numbers at both ages either single (e.g., *SFTPC*) or multiple TSSs (e.g., *ACTB*); and the 3rd pattern (e.g., *HSP14*) had one TSS in the fetal, but multiple TSSs in adult tissue. Functional annotation of genes with divergent TSSs across fetal and adult stages is shown in Supplementary Table 7 (Bonferroni-corrected P-value < 0.05). Interestingly, the first pattern was observed in all the assessed heat shock protein related genes with exception of *HSPA9*, *HSPA13*, *HSPD1*, and *HSPH1* which reflected the 2nd pattern with aging, and for *HSP14* in which the third pattern was observed (Supplementary Table 10).

Crossover switching of TSSs occurred in 221 genes between the two stages (Supplementary Table 9) and were enriched for the BP GO terms *biological regulation, protein localization*, and *regulation of transcription* (P-value < 0.05). In line with these results, it has been reported that the usage of alternative TSSs in some genes expressed in the cerebellum could play a regulatory role, such as *temporally regulated expression*, *amplitude of expression*, *mRNA stability*, and *mRNA translational efficiency*[37]. One example of crossover TSS switching with aging in lung tissue is *BLVRB*, biliverdin reductase B. We found

two common TSSs, which were 31 bp apart for this gene. The dominant TSS shifted with aging and the highest total transcript gene expression was seen at the fetal stage. The crossover TSS switching in this gene may play a regulatory role, such as temporally upregulated expression in fetal lung.

Alternative TSSs usage across *Bos indicus* adult tissues

The number of TSSs in seven tissues from a single *Bos indicus* adult female is shown in Table 1. In total, the number of TSSs identified ranged between 5,974 (blood) and 15,151 (lung). About 95–97% of TSSs identified in different tissues were located in known gene transcripts. In general, 4% (thyroid) to 16% (lung) of total TSS peaks identified in known gene transcripts overlapped with coding sequence (**CDS**), intron, exon, 3' UTR, and antisense (i.e. genes on the opposite strand) regions. Although there is some evidence that many novel core promoters especially for novel non-coding RNA are located in intergenic regions[30], for simplicity the final dataset was restricted to the promoters-proximal, and 5' UTR regions.

In total, 53.3% of genes expressed in all tissues had a single TSS, while only a minority (6.6%) of genes had multiple TSSs. Interestingly, a noticeable proportion of the genes (40.1%) had divergent TSS numbers, i.e. they had only one TSS in one or some tissues while having multiple TSSs in other tissues. In about 68% of genes with divergent TSSs the number of TSSs per gene in single tissues having multiple TSSs was two, 24.3% had three or more TSSs, and 7.6% had two or more TSSs across tissues. Table 1 describes in detail the number of genes with single or multiple TSSs in different tissues. This highlights that about half of the genes in cattle have multiple promoters or alternative TSSs per promoter and therefore high potential for differential transcriptional regulation across tissues. Several genome wide analyses have shown that about half of human and mouse genes have multiple promoters[38, 39]. Our results were in agreement with this research and indicate tissue specific usage of TSS in mammals[4, 40].

Tissues are distinguished by gene expression patterns, indicating distinct regulatory processes. Individual genes, or even sets of genes, in each tissue cannot adequately capture the diversity of structure and function that exist among different tissues[41], and multiple regulatory elements, including transcription factors and TSSs, that work together with other genetic and environmental factors must control the transcription of genes and production of proteins[42]. Alternative TSSs can significantly alter the 5' UTR structure and therefore result in higher or lower rate of protein synthesis [43, 44]. When tissues were clustered based on their correlation between TSS numbers in different genes across tissues, tissues grouped mainly together into clusters reflecting their function, liver clustered with thyroid and spleen tissues, ovary with uterus and lung (Fig. 6). The lowest proportions of single TSS genes (72%) was observed in more complex tissues such as liver tissue (Fig. 7).

A TSS survey from the FANTOM consortium which included 573 human primary cell samples, 152 human tissues, and 250 human cancer cell lines, has demonstrated that on average four TSSs exist per gene [9]. In the current study, the maximum number of TSS peaks per gene ranged from 7 (thyroid; gene *CCND3, UBE2D3*) to 19 (blood; gene *LOC112444653*). *LOC112444653* is a 5.8S ribosomal RNA located on

chromosome 27. The minimum and maximum number of TSSs for this gene was observed in liver (single TSS), and blood (19 TSSs peaks), respectively. Gene *SEPT9* is another example of a complex gene with alternative splicing resulting in six confirmed protein isoforms. It is a member of the *septin* family involved in cytokinesis and cell cycle control and both its expression levels and isoform composition differ among cell types [45]. The lowest and highest number of TSSs for this gene was observed in ovary and muscle (single TSS), and blood (6 TSS), respectively (Fig. 8). Although the *septin* family of genes has diverse functions in cytokinesis[46], a growing body of research has confirmed their biological roles in several cellular processes, development and physiology of specific tissues and organs, and various pathophysiological states[45]. Both genes are perfect examples of genes where they are regulated in a tissue-dependent manner and consistent with previous research[9].

When we looked at the genes with different numbers of TSSs across tissues, the most significant BP GO terms were *cytoskeleton organization*, *actin filament-based process*, *actin cytoskeleton organization*, *cellular macromolecule catabolism* (Bonferroni-corrected P-value < 0.05, Fig. 9). By assessing the terms associated with the genes that had only one TSS (one promoter) in all tissues, we found that single TSS genes tended to exist among the housekeeping genes with BP functions such as *RNA processing/splicing*, *translation* and *protein catabolic process, protein localization, protein catabolic process* (Bonferroni-corrected P-value < 0.05). In agreement with these results, a previous study [47] indicated that genes with single TSS were enriched in encoding proteins with housekeeping functions, while genes with multiple TSS were enriched in regulatory pathways.

In conclusion, our results give some insight into how TSSs and frequent local insertions, deletions and duplications in the regions containing them can drive rapid evolution of species and sub-species. For example, duplication of TSSs can allow for neo-functionalization of genes, where an original gene takes on tissue specific functions following the duplication event. This is similar to the neo-functionalization following whole genome duplication model proposed by previous study[48], but on a gene scale rather than a genome scale. However the TSSs duplication process will obviously be much more frequent, as demonstrated here, than whole genome duplication events.

Conclusion

Knowledge of TSS expression and distribution would be a useful starting point to predict biological function of specific genes in different developmental stages or tissues. In the current study, we used CAGE-Seq data from *Bos taurus taurus* and *Bos taurus indicus* diverged up to 0.5 million years ago from extinct wild aurochs (*Bos primigenius*) for the first time to assess changes in TSSs for these closely related cattle species. Also, we assessed age-related changes in TSS counts and TSS switching events in lung and liver tissues for the first time in cattle. Our results confirmed that TSSs evolve rapidly between species and even sub-species. The results of this study will accelerate future genomic research and will assist in narrowing down candidate genes with differential TSS usage. Our results also constitute an

atlas of potential target sites (TSSs) for tissue specific knock out or knockdown of gene expression with CRISPR/Cas9.

Methods

The methods were performed in accordance with relevant guidelines and regulations and approved by the Queensland Department of Agriculture and Fisheries Animal Ethics Committee. See Supplementary Note, Supplementary Table 1–10 for details of methods and extended biological findings.

Declarations

Contributions

MF, ER, BH, AC, and RX conceived and designed the experiments. LN and BM prepared the RNA. MF analysed the data. SM led the project that investigated the Brahman genome which lead to this work. MF wrote the paper and all authors revised it critically for important intellectual content. All authors read and approved the submitted manuscript.

Acknowledgements

We acknowledge financial contributions from Meat and Livestock Australia (Project P.PSH.0868 - "Characterisation of the Brahman Genome") 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 the Bos taurus indicus RNA. We are also thankful to Rodger and Lorena Jefferis of Elrose Brahman Stud who donated Neomi the Brahman cow used in this study to science.

Ethics declarations

The authors declare no competing interests.

References

1. Naval-Sanchez, M. *et al.* Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. **9**, 1–13 (2018).
2. Smale, S.T. & Kadonaga, J.T.J.A.r.o.b. The RNA polymerase II core promoter. **72**, 449–479 (2003).
3. Kadonaga, J.T.J.W.I.R.D.B. Perspectives on the RNA polymerase II core promoter. **1**, 40–51 (2012).

4. Lenhard, B., Sandelin, A. & Carninci, P.J.N.R.G. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. **13**, 233–245 (2012).
5. Haberle, V. & Stark, A.J.N.R.M.C.B. Eukaryotic core promoters and the functional basis of transcription initiation. **19**, 621–637 (2018).
6. Hampsey, M.J.M.M.B.R. Molecular genetics of the RNA polymerase II general transcriptional machinery. **62**, 465–503 (1998).
7. Thomas, M.C., Chiang, C.-M.J.C.r.i.b. & biology, m. The general transcription machinery and general cofactors. **41**, 105–178 (2006).
8. Carninci, P. *et al.* Erratum: Genome-wide analysis of mammalian promoter architecture and evolution (Nature Genetics (2006) 38,(626–635)). **39**, 1174 (2007).
9. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. **507**, 462–470 (2014).
10. Kimura, K. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. **16**, 55–65 (2006).
11. Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T.H.-M.J.T.i.G. The functional consequences of alternative promoter use in mammalian genomes. **24**, 167–177 (2008).
12. Bernstein, B. *et al.* Consortium EP. An integrated encyclopedia of DNA elements in the human genome. **489**, 57–74 (2012).
13. Park, P.J.J.N.r.g. ChIP–seq: advantages and challenges of a maturing technology. **10**, 669–680 (2009).
14. Buenrostro, J.D., Wu, B., Chang, H.Y. & Greenleaf, W.J.J.C.p.i.m.b. ATAC-seq: a method for assaying chromatin accessibility genome-wide. **109**, 21.29. 21-21.29. 29 (2015).
15. nature, E.P.C.J. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. **447**, 799 (2007).
16. Giuffra, E., Tuggle, C.K. & biosciences, F.C.J.A.r.o.a. Functional annotation of animal genomes (FAANG): current achievements and roadmap. **7**, 65–88 (2019).
17. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. **507**, 455–461 (2014).
18. Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks, in *Gene regulatory networks* 181–200 (Springer, 2012).
19. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. **3**, 211–222 (2006).
20. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. **309**, 1559–1563 (2005).
21. King, M.-C. & Wilson, A.C.J.S. Evolution at two levels in humans and chimpanzees. **188**, 107–116 (1975).
22. De, S., Teichmann, S.A. & Babu, M.M.J.G.r. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. **19**, 785–794 (2009).

23. Ludwig, M.Z., Bergman, C., Patel, N.H. & Kreitman, M.J.N. Evidence for stabilizing selection in a eukaryotic enhancer element. **403**, 564–567 (2000).
24. Ventura, M., Archidiacono, N. & Rocchi, M.J.G.R. Centromere emergence in evolution. **11**, 595–599 (2001).
25. Frith, M.C. *et al.* Evolutionary turnover of mammalian transcription start sites. **16**, 713–722 (2006).
26. Young, R.S. *et al.* The frequent evolutionary birth and death of functional promoters in mouse and human. **25**, 1546–1557 (2015).
27. Nei, M., Xu, P. & Glazko, G.J.P.o.t.N.A.o.S. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. **98**, 2497–2502 (2001).
28. Bornholdt, J. *et al.* Identification of gene transcription start sites and enhancers responding to pulmonary carbon nanotube exposure in vivo. **11**, 3597–3613 (2017).
29. Rout, P., Kaushik, R., Ramachandran, N.J.C.S. & Chaperones Differential expression pattern of heat shock protein 70 gene in tissues and heat stress phenotypes in goats during peak heat stress period. **21**, 645–651 (2016).
30. Carninci, P. *Cap-Analysis Gene Expression (CAGE): The Science of Decoding Genes Transcription*, Edn. 1st (Jenny Stanford).
31. Landry, J.-R., Mager, D.L. & Wilhelm, B.T.J.T.i.G. Complex controls: the role of alternative promoters in mammalian genomes. **19**, 640–648 (2003).
32. Hooper, S.B., Polglase, G.R. & Roehr, C.C.J.P.r.r. Cardiopulmonary changes with aeration of the newborn lung. **16**, 147–150 (2015).
33. Rudolph, A.M.J.A.R.o.P. Fetal and neonatal pulmonary circulation. **41**, 383–395 (1979).
34. Hooper, S.B. *et al.* Cardiovascular transition at birth: a physiological sequence. **77**, 608–614 (2015).
35. Enders, A.J.P. Reasons for diversity of placental structure. **30**, 15–18 (2009).
36. Mortola, J.P. *Respiratory physiology of newborn mammals: a comparative perspective*. (JHU Press, 2001).
37. Zhang, P. *et al.* Relatively frequent switching of transcription start sites during cerebellar development. **18**, 461 (2017).
38. Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L. & Myers, R.M.J.G.r. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. **16**, 1–10 (2006).
39. Baek, D., Davis, C., Ewing, B., Gordon, D. & Green, P.J.G.r. Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. **17**, 145–155 (2007).
40. Li, H. *et al.* Genome-wide analysis of core promoter structures in Schizosaccharomyces pombe with DeepCAGE. **12**, 525–537 (2015).
41. Sonawane, A.R. *et al.* Understanding tissue-specific gene regulation. **21**, 1077–1088 (2017).
42. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M.J.N.R.G. A census of human transcription factors: function, expression and evolution. **10**, 252–263 (2009).

43. Pozner, A. *et al.* Transcription-coupled translation control of AML1/RUNX1 is mediated by cap-and internal ribosome entry site-dependent mechanisms. **20**, 2297–2307 (2000).
44. Courtois, V. *et al.* New Otx2 mRNA isoforms expressed in the mouse brain. **84**, 840–853 (2003).
45. Gönczi, M. *et al.* Septins, a cytoskeletal protein family, with emerging role in striated muscle. 1–15 (2020).
46. Hartwell, L.H.J.E.c.r. Genetic control of the cell division cycle in yeast: IV. Genes controlling bud emergence and cytokinesis. **69**, 265–276 (1971).
47. Wang, X., Hou, J., Quedenau, C. & Chen, W.J.M.S.B. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. **12** (2016).
48. Gout, J.-F., Lynch, M.J.M.b. & evolution Maintenance and loss of duplicated genes by dosage subfunctionalization. **32**, 2141–2148 (2015).

Figures

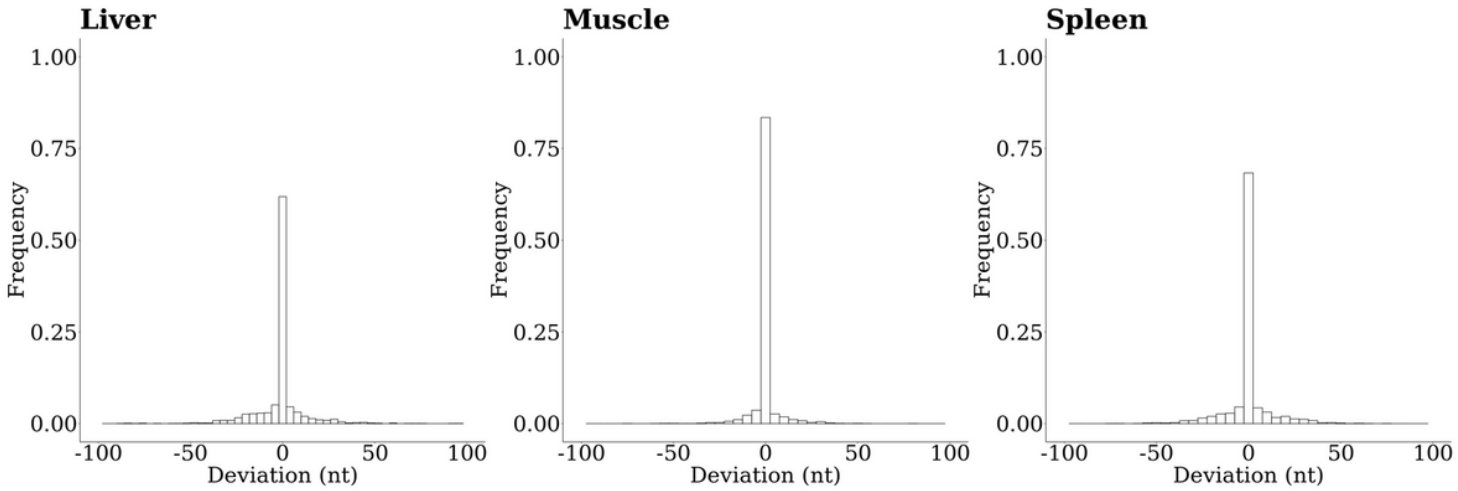

Figure 1

Histogram of distances between equivalent TSSs in genes with single TSS across Bos taurus and Bos indicus sub-species. The x-axis is calculated with respect to position in Bos taurus. Thus, positive numbers are downstream from Bos taurus and negative values are upstream from Bos taurus

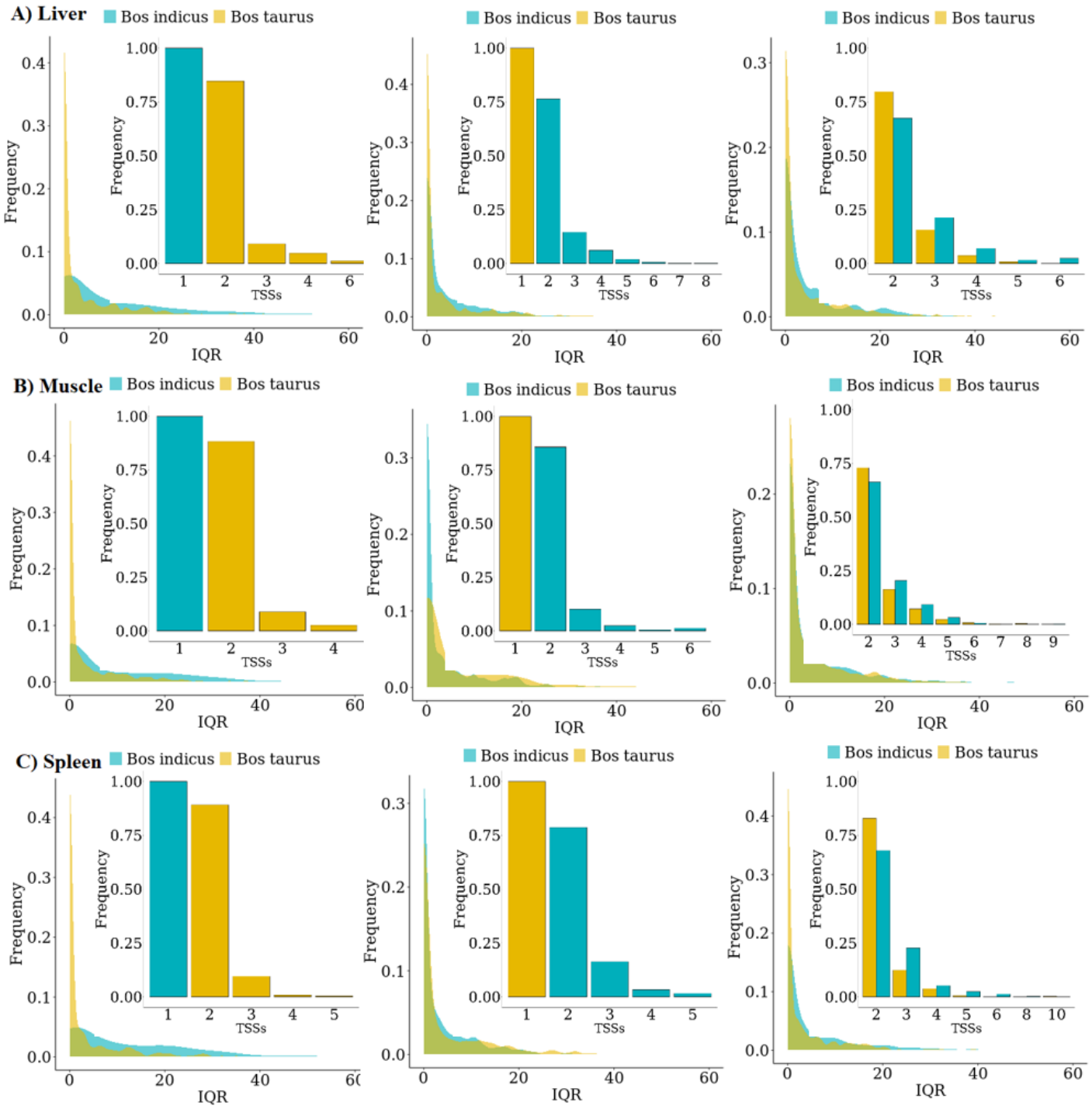

Figure 2

Frequency distribution of TSS width, and bar chart of TSS numbers in three tissues (A-liver, B-muscle, and C- spleen) for genes with single TSS in Bos indicus and multiple TSSs in Bos taurus (Left column), single TSS in Bos Taurus and multiple TSSs in Bos indicus (mid column), and multiple TSSs in both species (right column).

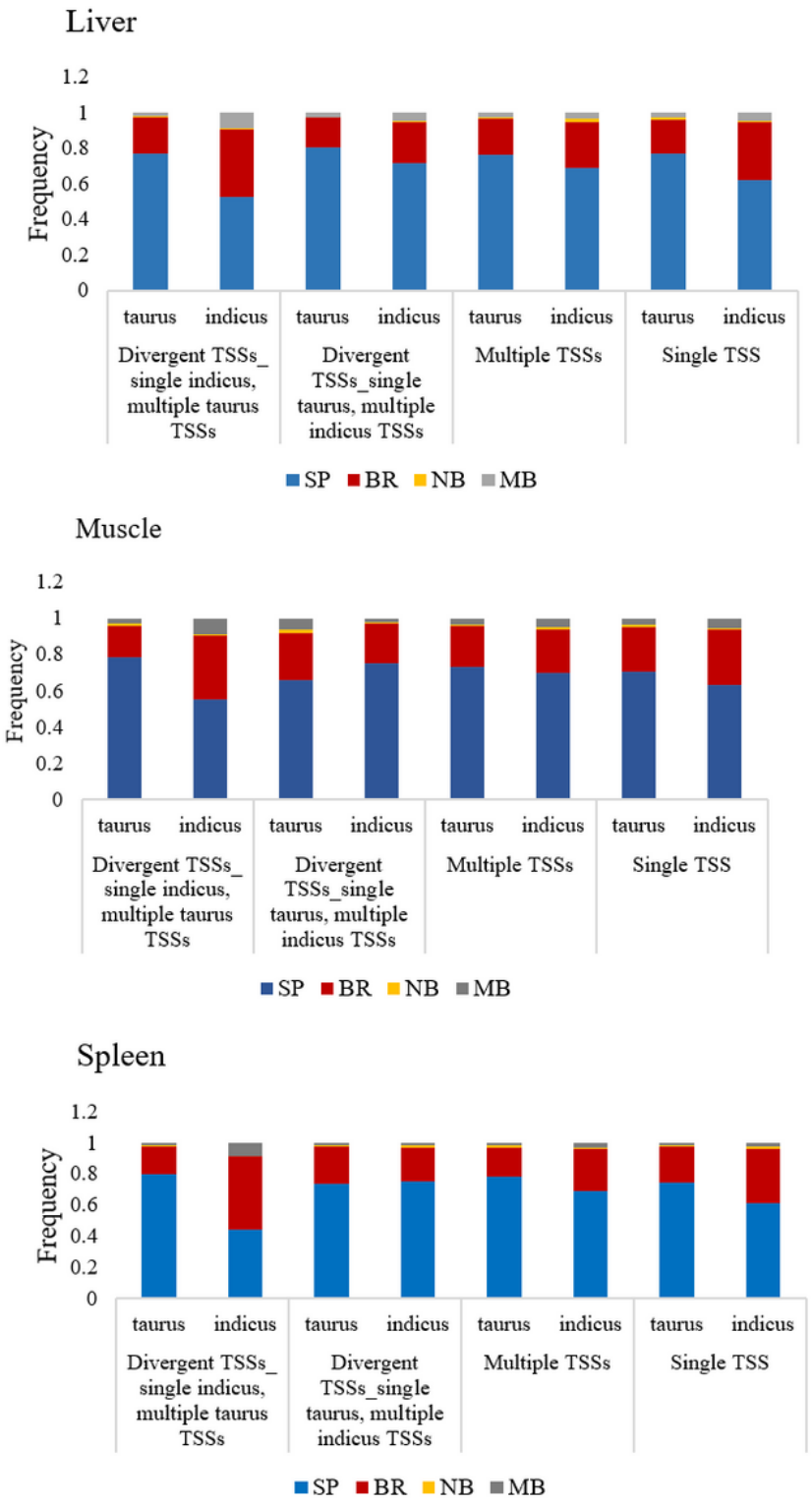

Figure 3

Comparison of TSS shape class (SP: TSS with IQR<4, NB: broad TSS (IQR>4) with normal distribution of cluster of tags, MB: broad TSS (IQR>4) with bi- or multi modal distribution of cluster of tags, and BR: broad distribution without normal or bimodal distribution) in three tissues (liver, muscle, and spleen) between Bos indicus and Bos taurus sub-species

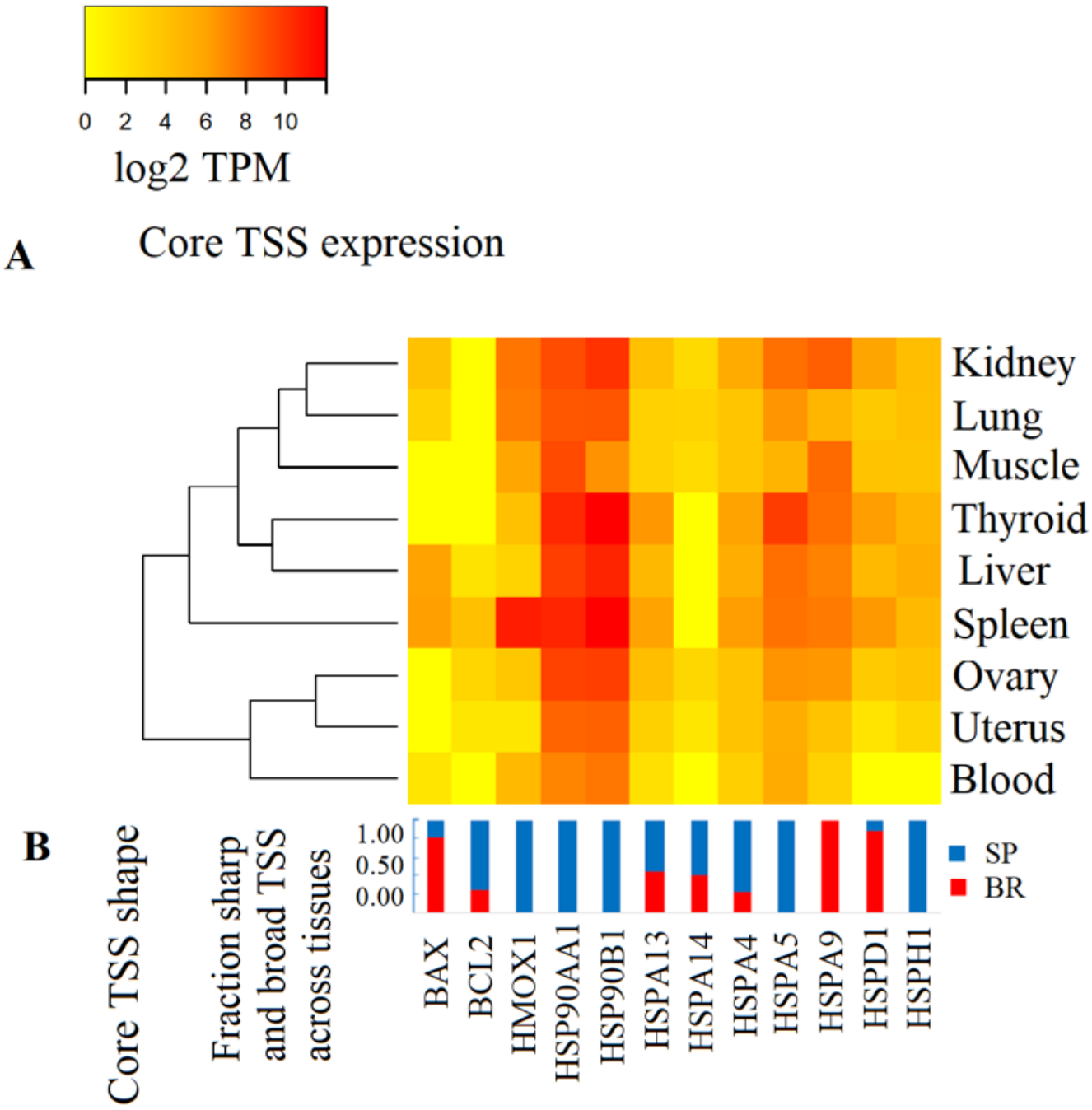

Figure 4

Heat map of core TSS expression (A) and shape propensity (B, fraction of broad or sharp TSS for each gene across tissues) in genes related to heat stress across the adult tissues (blood, spleen, thyroid, liver, lung, kidney, muscle, uterus, and ovary) in Bos indicus. Core TSS implies the most common TSS among all tissues. Sharp TSS (SP), indicative of a TSS spaned <=4 IQR and the broad (BR) shape indicates TSSs distributed over the larger distance (IQR>4). Tag per million (TPM) count was used as a measure of the expression level of RNAs in each tissue. Tissues grouped mainly together into clusters reflecting their similarities in TSS expression for genes related to heat stress.

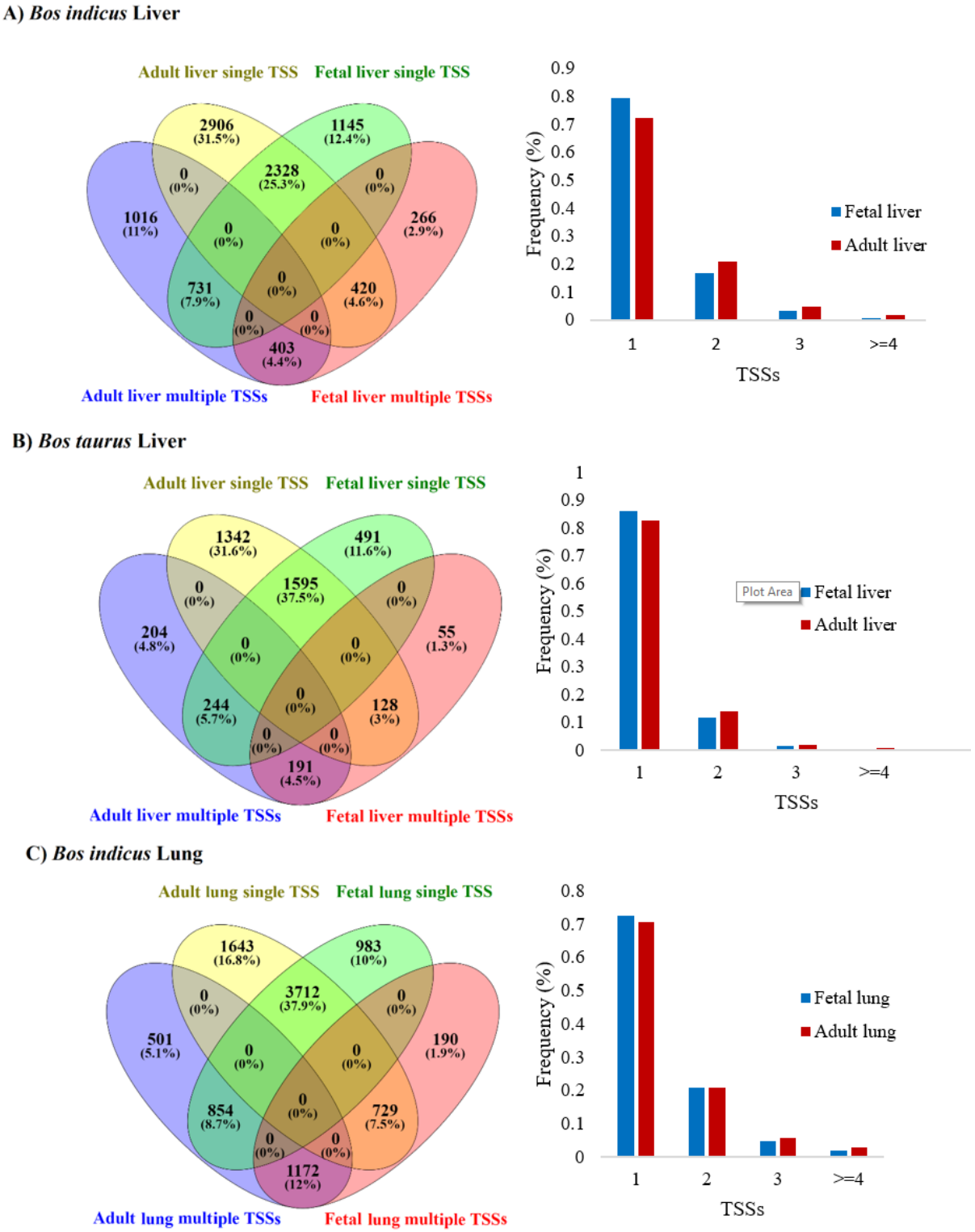

Figure 5

A. Overview of the TSS numbers (single or multiple) per gene in fetal and adult stages in Bos indicus liver (A), Bos taurus liver (B), and Bos indicus lung (C).

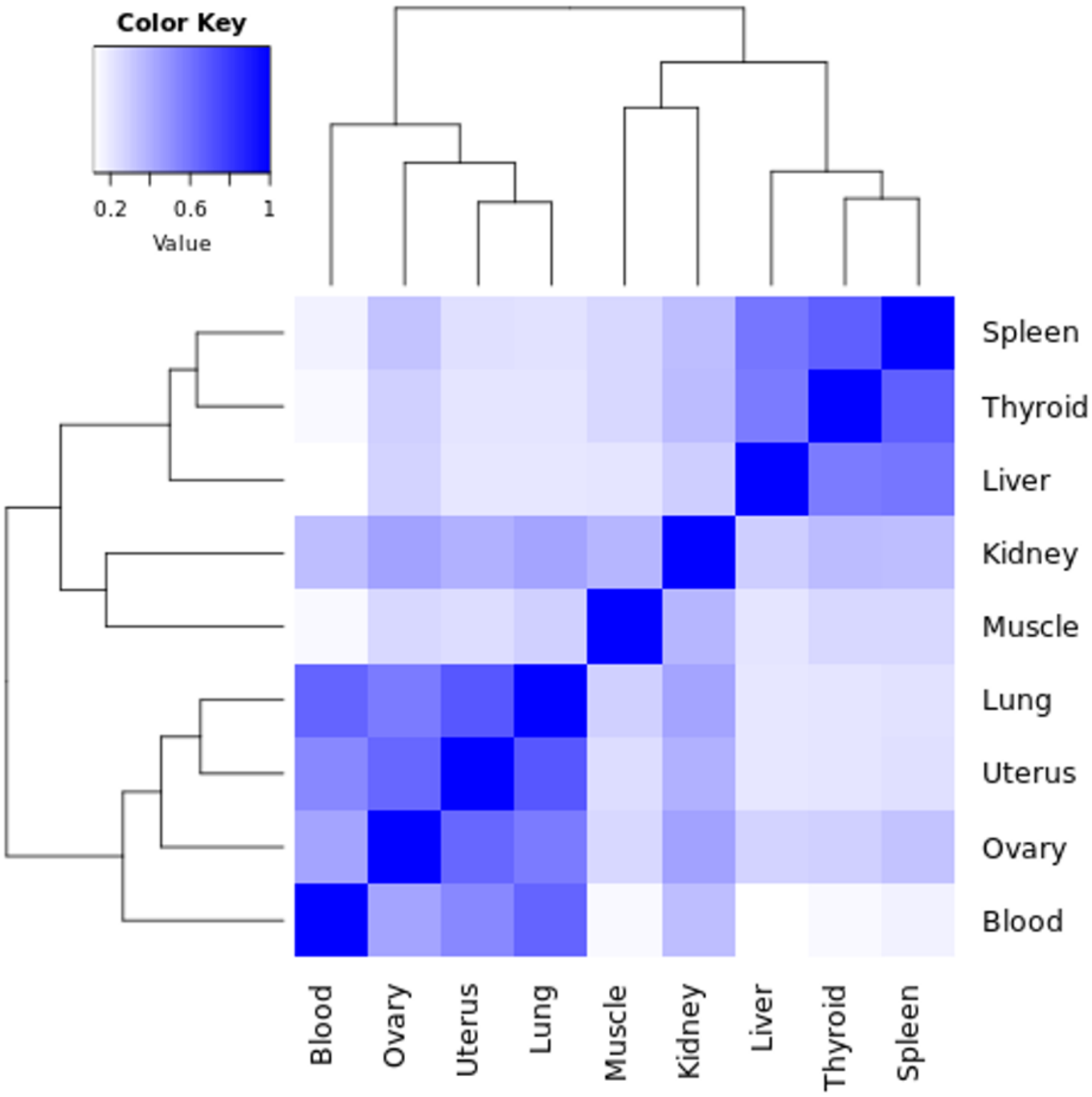

Figure 6

Heat map of the correlation between TSS numbers per gene expressed in a single adult Bos indicus across tissues blood, spleen, thyroid, liver, lung, kidney, muscle, uterus, and ovary.

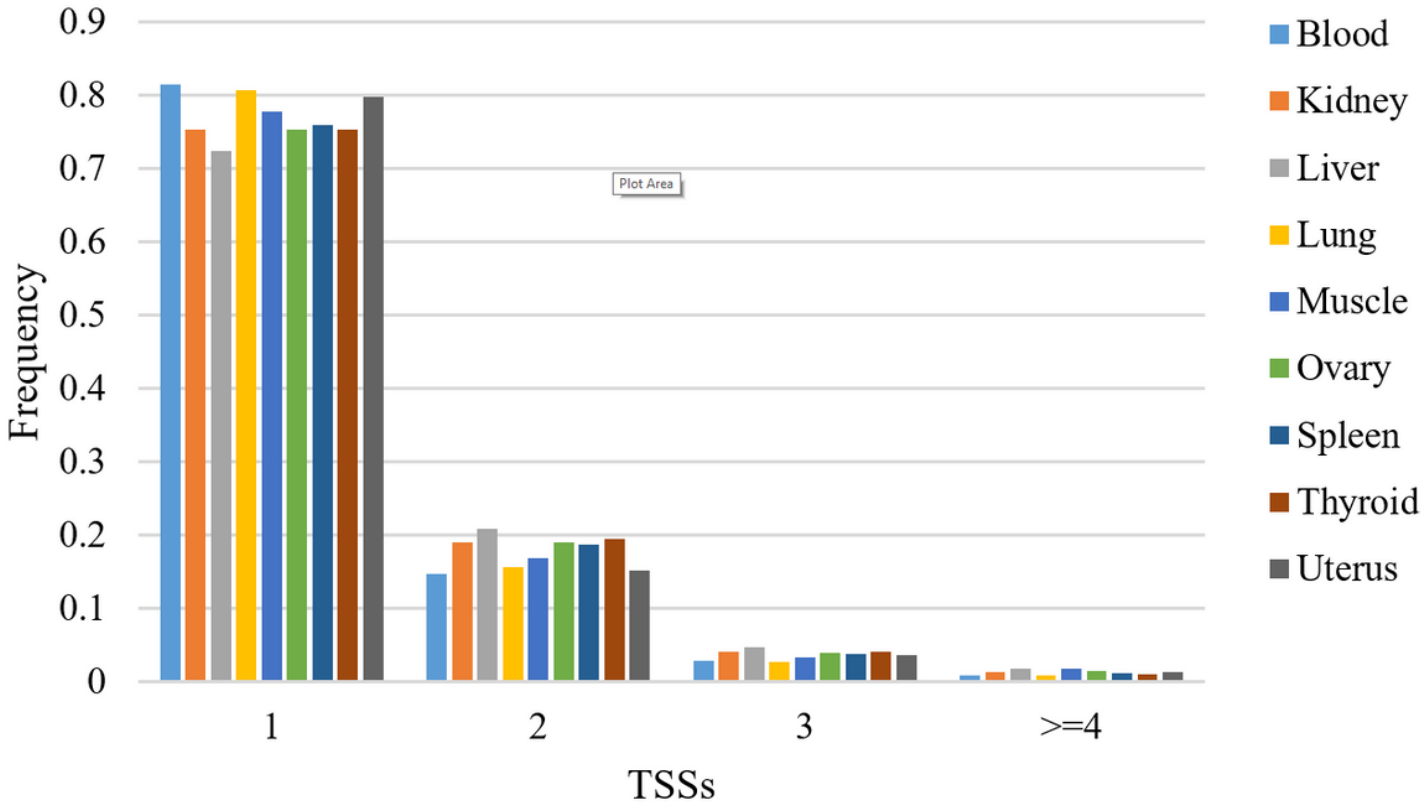

Figure 7

Frequency histogram of the number of TSS per gene for blood, kidney, liver, lung, muscle, ovary, spleen, thyroid and uterus in a single adult Bos indicus female

Figure 8

Genome browser view of TSS candidates for SEPT9 gene across Bos indicus tissues. The bottom track shows the TSS peaks mapping to each bp position in different tissues. The top track shows transcript gene models (lines/ thin blocks/ thick blocks are intronic/ UTR/ coding sequence regions, respectively).

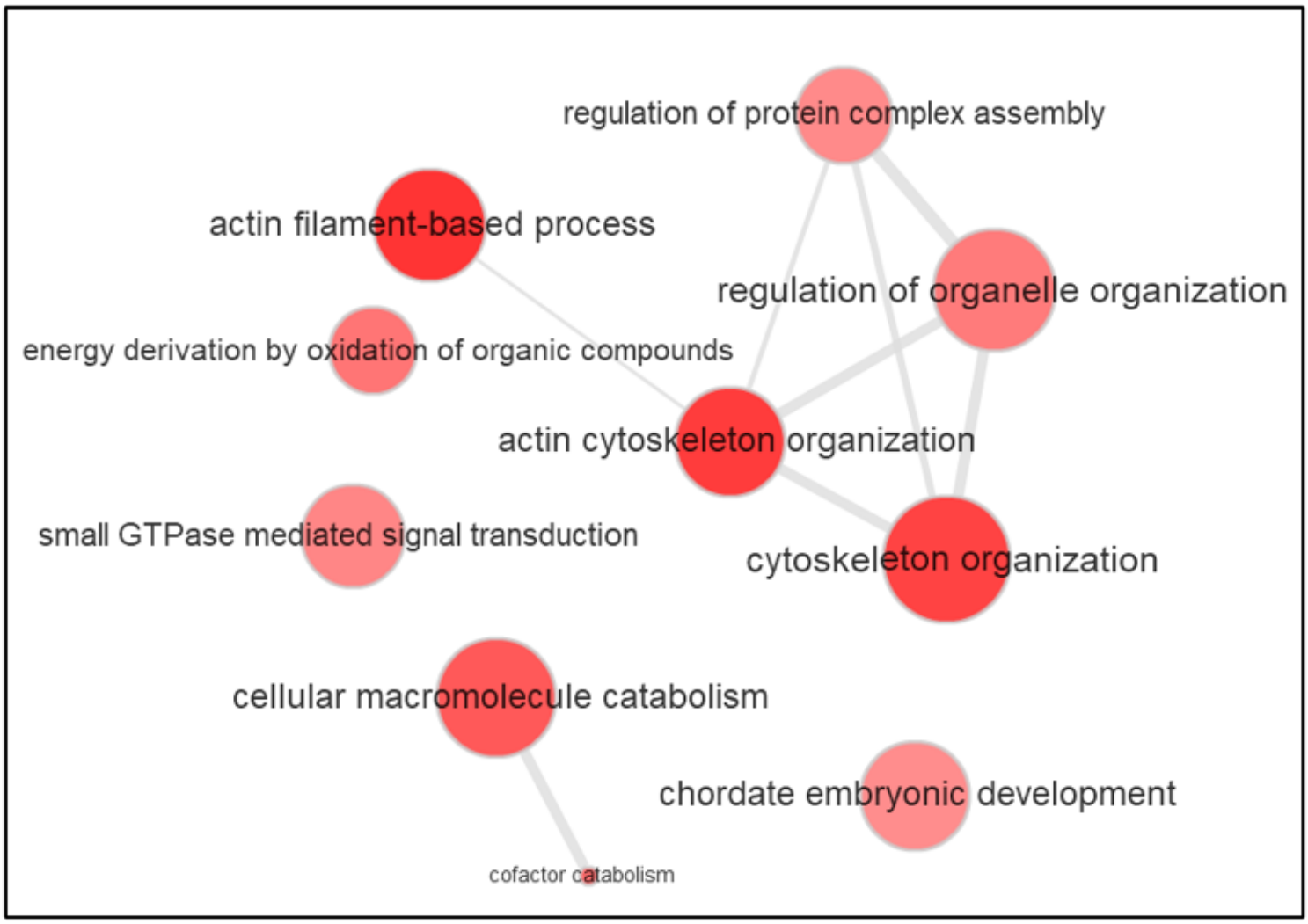

Figure 9

Interactive graph of biological process gene ontology (BP GO) terms of the genes with divergent single/multi TSSs across adult tissues (blood, spleen, thyroid, liver, lung, kidney, muscle, uterus, and ovary) in Bos indicus (Bonferroni-corrected P-value <0.05)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryNote.docx
- supplementaryTablesAC200708MF.xlsx