Addis Ababa University

Institute of Biotechnology

Deciphering the Conserved *Cis*-Regulatory Elements of Major Milk Protein Genes by Computational Analysis

M.Sc. Thesis

By: Addis Tekaw

April, 2024

Addis Ababa, Ethiopia

Deciphering the Conserved *Cis*-Regulatory Elements of Major Milk Protein Genes by Computational Analysis

By: Addis Tekaw

A thesis submitted to the Institute of Biotechnology, Addis Ababa University in partial fulfillment of the requirement for a Master of Science in Bioinformatics

Advisor: Abiy Zegeye (Ph.D.)

April, 2024

Addis Ababa, Ethiopia

Declaration

I, the undersigned, declare that the Master thesis entitled "Deciphering the Conserved *cis*-Regulatory Elements of Major Milk Protein Genes by Computational Analysis" is my thesis work submitted to Addis Ababa University, Institute of Biotechnology. It has not been presented to any other university for the award and all the sources of materials used have been duly acknowledged.

Name: Addis Tekaw

Signature ---------- Date----------

Approval of Thesis by Supervisor for Submission

I certify that ----------‘s M.Sc. thesis entitled “Deciphering the Conserved *Cis*-Regulatory Elements of Major Milk Protein Genes by Computational Analysis” has been carried out under my direct supervision and recommend the thesis to be accepted as fulfilling the requirement for the degree of Master of Science in Bioinformatics.

Supervisor ---------- Signature ---------- Date ----------

Director ---------- Signature---------- Date----------

Acknowledgement

I express my sincere gratitude to Dr. Abiy Zegeye for his exceptional guidance and mentorship during the extensive research and development of my skills in the field of bioinformatics. I am truly grateful for his unwavering support and invaluable insights that have greatly contributed to my growth and success in this area.

I would like to express my gratitude to the Institute of Biotechnology at Addis Ababa University for pioneering the launch of the Master of Science program in Bioinformatics, a groundbreaking initiative that had not been introduced before. Their efforts and support were instrumental in enabling me to complete the course.

Furthermore, following divine guidance, I am profoundly grateful to my dear family, supportive friends, and committed colleagues who have consistently provided me with encouragement and assistance on this path. Their support, understanding, and collaborative efforts have been instrumental in molding my professional growth and accomplishments.

Table of Content

Contents

List of Tables

List of Figures

List of Acronyms and Abbreviations

AP-1	Activator protein 1
BLG	Beta-lactoglobulin
BP	Biological processes
C/EBP	CCAAT enhancer binding protein
CC	Cell component
CCD	Coiled-coil domain
CN	Casein
CSN1S1	Alpha s1 casein
CSN1S2	Alpha s2 casein
CSN2	Beta casein
CSN3	Kappa casein
DBD	DNA-binding domain
EGF	Epidermal growth factor
ER	Epithelial rest
EDR	False discovery rate
G/L	Gram per litter
GAS	Gamma interferon activation site
GCs	Glucocorticoids
GO	Gene ontology
GOmo	Gene ontology for candidate motifs
GPCR	G-protein coupled receptor
HTML	Hyper Text Markup Language
INS	Insulin
KB	Kilo base pair
KEGG	Kyoto Encyclopedia of Genes and Genomes
LALBA	Alpha-lactalbumin
LD	Linker domain
MECs	Mammary epithelial cells
MF	Molecular function
MG	Mammary gland

NJ	Neighbor-joining
NTD	N-terminal domain
OCT1	Octamer-binding protein 1
OOPS	One occurrence per sequence
OXTR	Oxytocin receptor
PRL	Prolactin
PRLR	Prolactin receptor
RPKM	Reads per kilobase of transcript per million mapped reads
SCPP	Secretory calcium-binding phosphoprotein
SH2	Src homology 2
SPARCL1	SPARC-like 1
STAT	Signal transducer and activator of transcription
TAD	Transcription activation domain
TFBSs	Transcription factor binding sites
TFs	Transcription factors
TSS	Transcription start sites
WAP	Whey acidic protein
YY1	Yin Yang 1

ABSTRACT

Milk genes are exclusively expressed in MECs during lactation, regulated by lactogenic hormones like prolactin that act through specific TFs. The recruitment of TF is determined by cis-regulatory motifs in their gene promoters; but there is limited evidence of shared motifs among the major milk gene across species. In this study, Jalview, Meme-Suite, TomTom, and GOMo software were utilized to construct phylogenetic trees, discover motifs, identify TFs, and determine GO terms, respectively, in the 2kb upstream putative promoter region of CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG milk genes among twenty-three mammalian species. The analysis revealed three common TFBSs for STAT1, STAT5a, and STAT5b TFs in all milk genes across the species studied, except for BLG, within the region -390 to -80bp from the canonical TSS, with a few shared TFBSs located in the distal promoter region upstream of -600bp. STAT3 was also detected in CSN1S1, CSN2, CSN3, and LALBA genes, sharing binding sites with STAT1, STAT5a, and STAT5b. Furthermore, TFBSs for Sox9 and Sox6 TFs were found to be shared between the LALBA and BLG putative promoters within -400 to -100bp. Moreover, GO analysis linked GPCR with regulatory motifs in all six milk genes across 23 species, essential for enhancing STAT5 phosphorylation through Gαq activation in the JAK-STAT pathway. The totality of the result suggests that the relative abundance of STAT5 and STAT3 proteins may play a role in regulating the expression level of caseins and LALBA commensurate with the stage of lactation. Together, these results offer valuable insights into the potential of the milk genes' promoter for constructing eukaryotic expression vectors and provide essential information for transgenic studies.

Keywords: MEME-Suite, Promoter, *Cis*-regulatory Element, Motifs, Mammals, Mammary gland, Milk genes.

1. INTRODUCTION

Mammals are defined by lactogenesis, which is the synthesis of milk for the sustenance of the newborn. Mammary epithelial cells (MECs) are secretory cells in mammals and form the basis of lactation in the mammary gland (MG). The number and activity of MECs are strongly linked to the lactation phase of the MG and play a vital role in MG development. The ability of MGs to produce milk relies on both the quantities of milk-secretary cells and their level of activity. The primary protein constitute of milk are the four caseins (alpha S1 casein (CSN1S1), alpha S2 casein (CSN1S2), beta-casein (CSN2) and kappa-casein (CSN3) along with two whey protein (alpha-lactalbumin (LALBA) and beta-lactoglobulin (BLG)), which are synthesized by MECs (Fox *et al*., 2015; Le *et al*., 2017; Maity *et al*., 2020; Sebastiani *et al*., 2020).

The MG is a dynamic exocrine organ that goes through a cycle of growth, functional differentiation, and regression, all of which are connected to reproductive functions. Mammary development initiates in early fetal stages and progresses minimally during estrous cycles. The full development of the MG occurs primarily during pregnancy, reaching full functionality after parturition to provide nourishment to the newborn. Following weaning, the mammary tissue undergoes regression during involution but can undergo re-specialization if a subsequent pregnancy occurs. The different stages of MG development are precisely coordinated spatio-temporally, influenced by both systemic hormones and local factors (Truchet and Honvo-Houéto, 2017).

Research on the regulation of milk protein gene expression has shown that key hormone complex, specifically the lactogenic hormone prolactin (PRL), glucocorticoids (GCs), and insulin (INS), activate milk protein gene expression synergistically (Qian and Zhao, 2014b). Of the several factors involved in milk gene expression, PRL plays a crucial role as a fundamental hormone in all mammals. It governs lactogenesis by influencing MG cell proliferation, reducing apoptosis, and promoting milk production and secretion. Originating from the anterior pituitary gland, this hormone binds to PRL receptor (PRLR) in MECs. This binding regulates milk protein gene expression through regulatory elements in the putative promoter region of milk protein genes, triggering a range of cellular effects from growth stimulation to the initiation of milk protein synthesis (Wang *et al*., 2022).

Promoter studies has revealed that gene regulation can be influenced by transcription factors (TFs), and investigating transcription factor binding site (TFBSs) helps to fill the knowledge gap concerning milk gene expression (Najafi *et al*., 2014). TFs and their binding sites within promoters are recognized as fundamental functional elements in any genome. Perturbations in these protein-DNA interactions may play a role in the onset of diverse disorders. These interactions govern numerous crucial processes, including key developmental stages and responses to environmental cues (Parveen *et al*., 2023).

Study of promoter *cis*-elements and regulation of expression of CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG genes have been studied in goats and sheep (Ramunno *et al*., 2004; Najafi *et al*., 2014; Zhang *et al*., 2015; Morammazi *et al*., 2016), in horses (Lenasi *et al*., 2005), in camels (Pauciullo *et al*., 2013; Pauciullo *et al*., 2014; Parveen *et al*., 2023, Pauciullo *et al*., 2024), in rat, mice, rabbits, humans and cattle (Malewski, 1998; Gerencsér *et al*., 2002; Debeljak *et al*., 2005), and in water buffalo (Feng *et al*., 2021). However, there is limited direct evidence regarding universally present *cis*-regulatory elements among all major milk protein gene promoters across various species (Patel *et al*., 2014; Amandykova *et al*., 2022). Therefore, this study was conducted to deciphering the shared regulatory elements and their characterizing among the CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG genes' 2kb upstream putative promoter regions from the canonical TSS across twenty-two eutherian mammals and one marsupial, namely, human, horse, goat, cow, mouse, bonobo, lion, chimpanzee, Norway rat, sheep, Arabian camel, tiger, olive baboon, narwhale, macaque, megabat, domestic cat, dog, giant panda, rabbit, pig, kangaroo rat, and opossum based on available sequences in public databases through computational analysis. This will elucidate the promoters' potential use for constructing a eukaryotic expression vector and provide crucial details for transgenic study (Shepelev *et al*., 2018).

1.1. Statement of the Problem

The putative promoter regions of CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG milk genes are commonly utilized to express recombinant proteins, providing an opportunity to boost protein yield in milk among various mammalian species. The *cis*-elements and expression induction from these promoters of these milk genes have been reported in goats and sheep (Ramunno *et al*., 2004; Najafi *et al*., 2014; Zhang *et al*., 2015; Morammazi *et al*., 2016), in horses (Lenasi *et al*., 2005), in camels (Pauciullo *et al*., 2013; Pauciullo *et al*., 2014; Parveen *et al*., 2023a; Pauciullo *et al*., 2024), in rat, mice, rabbits, human and cattle (Malewski, 1998; Gerencsér *et al*., 2002; Debeljak *et al*., 2005), and in water buffalo (Feng *et al*., 2021). However, there is limited evidence regarding a universally present *cis*-regulatory element(s) that enable mammals to spatio-temporally regulate gene expression among all major milk protein genes across various species (Patel *et al*., 2014; Amandykova *et al*., 2022). This underscores the necessity of identifying universally shared *cis*-regulatory element(s) among CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG milk genes' 2kb putative promoter regions across twenty-three mammalian species based on publicly available sequence data.

1.2. Research Questions

What shared *cis*-regulatory element(s) govern the precise spatio-temporal expression of CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG milk genes of twenty-two true eutherians and one marsupial species? Answers to this question will elucidate the shared *cis*-regulatory element(s) that govern the spatio-temporal expression of milk genes across eutherians and marsupial, providing an important entry point for customizing milk composition and enhancing milk yield through genetic improvement.

1.3. Objectives

1.3.1. General Objective

This study aims to computationally identify universally conserved *cis*-regulatory element(s) within the 2 kb putative promoter region immediately upstream of the canonical TSS or first exon for six major milk protein genes in 23 different species, namely: human, horse, goat, cow, mouse, bonobo, lion, chimpanzee, Norway rat, sheep, Arabian camel, tiger, olive baboon, narwhale, macaque, megabat, domestic cat, dog, giant panda, rabbit, pig, kangaroo rat, and opossum.

1.3.2. Specific Objective

- Identify and retrieve putative 2kb promoter sequences immediately upstream of the TSS or first exon.
- Sort major milk protein putative promoter regions into distinct evolutionary diverged groups, or clades based on phylogenetic analysis
- Discover candidate regulatory motifs and their associated TFs.
- Characterize the discovered candidate regulatory motifs and TFs via gene ontology analysis

1.4. Significant of the Study

This study aims to enhance our understanding of the shared *cis*-regulatory elements that control the precise spatio-temporal expression of key milk protein genes, which are crucial for establishing and maintaining lactation and determining milk composition in twenty eutherian species and one marsupial species. Understanding *cis*-regulatory elements is essential for understanding how TFs regulate the expression of milk protein genes, and this knowledge has potential applications for expressing transgenic proteins in the milk of various species.

2. LITERATURE REVIEW

2.1. Major Milk Proteins

The intricate process of milk production and secretion commences with MG development and is intricately regulated by systematic hormones and local factors. These components collaborate to ensure MECs function in unison to provide newborns with the appropriate composition and volume of milk (Truchet and Honvo-Houéto, 2017). Milk proteins are categorized into whey (serum) and casein families (McMeekin and Polis, 1949) and are an important part of a neonate's diet. Scientific research has demonstrated that milk proteins offer a range of health benefits, both nutritionally and biologically. Milk proteins, whether intact or as derivatives, have been associated with anticarcinogenic effects, antihypertensive properties, immune system regulation, and other metabolic traits (Davoodi *et al*., 2016). This distinctive dietary matrix comprises on average 87% water, 4% lipids, and 9% solid-non-fat compounds including protein, lactose, and various minerals and vitamins, facilitating fractionation and isolation processes (Burke *et al*., 2018).

Milk supplies neonates' energy and food requirements for optimal growth and development. It contains considerable amounts of protein, minerals, and fats and can offer sustenance and immunological protection to newborns. One of the primary functions of milk is to provide necessary amino acids and minerals. These amino acids and minerals are necessary for neonatal growth, as well as optimal muscle and tissue function (Gigli, 2016). Species, breed, age, lactation quantity, environmental factors (such as nutrition, temperature, and milking method), and pathological conditions all influence the composition, features, and qualities of milk proteins, which account for changes in milk contents. Nutritional and genetic factors influence milk protein concentration; however, the latter is thought to be the more essential due to intrinsic regulatory mechanism (Asim *et al*., 2022).

2.1.1. Casein Milk Protein Genes

Caseins, originating from the Latin term "caseus" meaning cheese, make up a group of phosphoproteins that serve as the primary constituents of milk protein. They contain bioactive peptide precursors and immunomodulators, playing essential roles in various biological functions (Olumee-Shabon *et al*., 2013; Scumaci *et al*., 2015; Izquierdo-González *et al*., 2019). The major proteins in casein, namely CSN1S1, CSN1S2, CSN2, and CSN3, represent around 38%, 10%, 36%, and 13% of the casein fraction, respectively (Auestad and Layman, 2021). Together, these proteins establish an optimal balance of essential amino acids. Notably, they are characterized by a high proline content (16% of amino acids) and low cysteine content, resulting in the absence of disulfide bonds and alpha-helix structures commonly found in most proteins. Additionally, these phosphoproteins exhibit a strong affinity for calcium binding and are hydrophobic, rendering them insoluble in water. In their natural state within milk, caseins form a micellar structure to remain suspended in water. However, under acidic conditions (pH < 4.6), casein proteins coagulate, leading to curd formation, a pivotal step in cheese production (Bhat *et al*., 2016; Han *et al*., 2021). Studies have shown that casein can improve calcium absorption and bone mineral density, prevent muscle protein breakdown, and contribute to a moderate yet sustained muscle protein synthesis (Jeong *et al*., 2022).

Caseins have evolved in various mammalian species to perform specialized activities in milk, including delivering nutrients and minerals, particularly calcium, to offspring while maintaining fluidity in MGs (Runthala *et al*., 2023). Furthermore, caseins are the key milk proteins that provide amino acids and immunity to newborns (Roy *et al*., 2020). According to Dalgleish *et al*. (2004), these milk proteins are phosphoproteins generated by MECs under multi-hormonal regulation that resemble raspberries in electron micrographs. They are found in all animals' milk, but their overall concentrations and relative fractions differ greatly between species (Martin *et al*., 2013). The three primary caseins, CSN1S1, CSN1S2, and CSN2, are known as calcium-sensitive caseins because they precipitate quantitatively with calcium chloride. CSN3, on the other hand, does not precipitate with calcium chloride and instead stabilizes calcium-sensitive caseins (Qian and Zhao *et al*., 2014a).

Previously, it was discovered that casein genes are only found in mammalian genomes and that all casein genes belong to the secretory calcium-binding phosphoprotein (SCPP) gene family (Kawasaki *et al*., 2011). The SCPP gene family arose from the duplication of the SPARCL1 (SPARC-like 1) gene in an early vertebrate (Kawasaki *et al*., 2007), and numerous SCPPs are involved in bone and tooth mineralization in current vertebrates (Kawasaki and Weiss, 2008). SCPP genes have a well-conserved and identifiable exon-intron structure, which supports the hypothesis that all casein genes originated by gene duplication (Kawasaki *et al*., 2005).

Caseins are an important molecular model for evolutionary research, providing insights into the genetic makeup of understudied species (Kawasaki *et al*., 2011). Exploring genomic architecture and evolutionary events is critical for understanding the regulatory mechanisms of the casein gene family in mammals. While caseins are found in all mammalian milk, the overall content and relative fractions differ between species. Different species investigated extensively have varying milk amounts, with casein concentration ranging from 2.4 to 52.6 g/L, with human milk having one of the lowest protein levels (2.4 - 4.2 g/L) and sheep milk having one of the highest (41.8 - 52.6 g). Aside from the wide range in milk casein concentration, there are considerable variances in the relative distribution of caseins among species. Notably, CSN1S2 is absent from human milk (Roy *et al*., 2020; Runthala *et al*., 2023).

Table 1. Casein concentration and composition of milk in some selected mammalian species.

Animal	Total casein concentration (g/L milk)	Relative casein composition (%)				
		CSN1S1	CSN1S2	CSN2	CSN3	CSN1S1+CSN2
African elephant	Unknown	-	-	89	11	89
Buffalo	32-40	40	9	35	12	75
Cow	24.6-28	38	10	40	12	78
Camel	22.1-26	22	9	65.5	3.5	87
Goat	23.3-46.3	20	16	41	17	61
Horse	9.4-13.6	17.7	1.5	79	1.8	96
Human	2.4-4.2	3	-	70	27	73
Sheep	41.8-52.6	50	-	40	10	90
Range	-	3-55	9-28	26-89	3.5-27	61-96

(Source: Runthala *et al*., (2023), except the African elephant which is from Madende *et al*., (2015)).

2.1.2. Whey Milk Protein Genes

Whey is the aqueous portion remaining after cheese production and is rich in high-quality proteins known for their essential amino acid content, bioavailability, and bioactivities. Whey

proteins are notably abundant in lysine, methionine, leucine, and tryptophan, essential amino acids that are often limited in other dietary sources. Among the key proteins found in whey are BLG and LALBA. LALBA, a significant calcium-binding whey protein specific to MGs, plays a crucial role in lactose synthesis. It assists in lactose production by regulating the galactosyltransferase part of the enzyme system (Pauciullo *et al*., 2024).

Lactose synthesis is vital for milk synthesis because it generates osmotic pressure, which draws water into the MG and increases overall milk volume (Layman *et al*., 2018). LALBA comprises approximately 22% of total protein and 36% of whey protein in human milk. This soluble protein with 129 amino acids is remarkable for its high concentrations of tryptophan, lysine, cysteine, and branched-chain amino acids. Aside from its high protein quality, which promotes growth and development, LALBA has other bioactive qualities associated with sleep, mood modulation, gastrointestinal function, mineral absorption, and immunological response (Layman *et al*., 2018). Due to its unique amino acid composition and significant presence in milk, LALBA has been extensively studied for potential use in infant formulas. Many dairy-based infant formulas are enriched with whey proteins, often containing preparations with added LALBA. However, LALBA has not received substantial attention for use in adult nutrition (Layman *et al*., 2018).

BLG, a lipocalin, belongs to a diverse protein family that binds small hydrophobic molecules and functions as specific carriers, akin to serum retinol-binding protein (Berry *et al*., 2010). It is the most common whey protein in ruminant milk and can be found in the milk of a variety of mammals including dogs, cats, pigs, horses, kangaroos, wallabies, and dolphins, but it is not found in rodents or human milk (Wodas *et al*., 2020). In bovine milk, BLG accounts for roughly 10% of total milk protein and 58% of whey protein. This relatively small peptide chain comprises 162 amino acids and forms a dimer with a molecular weight of 36kDa in bovine milk. BLG exhibits high solubility and clarity across a wide pH range (pH 3 to 7) and displays exceptional gelling and foaming properties, making it valuable for various food applications (Chatterton *et al*., 2006). Nonetheless, concerns regarding BLG allergies have been raised due to its absence in human milk (Wei *et al*., 2018). The functional roles of BLG in ruminant offspring and its potential applications in human nutrition have been investigated but remain speculative (Chatterton *et al*., 2006; Wei *et al*., 2018).

2.2. Hormonal and Transcriptional Control of Milk Protein Genes

The regulation of major milk protein-encoding genes is complicated, involving steroid and polypeptide hormones, local growth factors, and cell-environment interactions. Throughout lactation, MECs create a considerable amount of milk proteins, with more than 90% originating from the transcription of tissue-specific genes. These genes are regulated by a complicated hormonal network that includes both transcriptional and post-transcriptional processes. Furthermore, for the MG to function as a bioreactor, it must have an adequate supply of amino acids, as well as efficient translation and transportation pathways during lactation (Vilotte *et al*., 2013).

Lactogenic hormones such as insulin, prolactin, and glucocorticoids, as well as local growth factors and cell-cell and cell-substratum interactions, cause specific transcription of essential milk protein genes and modify chromatin structure. These chromatin structure alterations include nucleosome remodelling and post-translational histone modification, which occur both at the nucleosome level and within broader chromatin domains (Xu *et al*., 2007). In contrast, hormones such as progesterone inhibit this stimulation in early pregnancy, favoring cell expansion over cell specialization. In many circumstances, it is difficult to distinguish between direct milk protein transcription induction and indirect differentiation-related activities. Major milk protein gene activation does not occur concurrently during pregnancy, probably due to varied hormonal settings and differential responses of individual milk protein genes to these environments (Kabotyanski *et al*., 2006 and 2009).

The use of bioinformatics methods in genome-wide research has improved our fundamental understanding of the distinct genomic structure of regulatory elements that drive cell-type-specific gene expression (Ong *et al*., 2011 and 2012). Conserved DNA motifs thought to regulate gene transcription were discovered by comparing the sequences of these distinct genomic structures within a given gene across different species or among diverse milk protein genes (Kolb, 2002). With over fifty years of research on regulating milk protein gene expression in the MG, the insights gained from these investigations have not only provided essential knowledge for genetic and nutritional improvements in milk composition and production, but have also shed light on the molecular mechanisms behind temporal and tissue-specific gene expression (Forsyth *et al*., 2009).

2.3. Gene Expression Dynamics of Major Milk Genes

Milk contains two primary milk proteins: caseins (CN) and whey proteins. Genes CSN1S1, CSN1S2, CSN2, and CSN3 encode caseins and are found on chromosome 6 in bovine. These genes produce the milk proteins alpha S1-CN, alpha S2-CN, beta-CN, and kappa-CN, respectively (Caroli *et al*., 2009). The relative expression of these milk protein genes (CSN1S1, CSN1S2, CSN3, CSN2) showed the greatest fold changes at day 250 (late lactation), but reduced dramatically as lactation progressed (Figure 1).

The BLG and LALBA genes encode the major whey proteins, beta-lactoglobulin and alpha-lactalbumin, respectively. These two genes likewise show the most significant fold changes at day 250 (late lactation), and, like CN genes, show a considerable reduction as lactation progresses (Figure 1). Notably, the proportions of total milk proteins, whey proteins, and casein are rather consistent throughout lactation. However, mRNA gene expression patterns indicate a higher transcription rate for casein and whey proteins during transition lactation. One possible explanation for this discrepancy is that these abundant caseins and whey proteins are enzymatically cleaved into bioactive peptides, leading to their concentrations not being accurately reflected in the analysis of major milk component proteins. In cows, it has been observed that bioactive peptides resulting from the breakdown of caseins and whey proteins are more prevalent at the onset of lactation (Wickramasinghe *et al*., 2012).

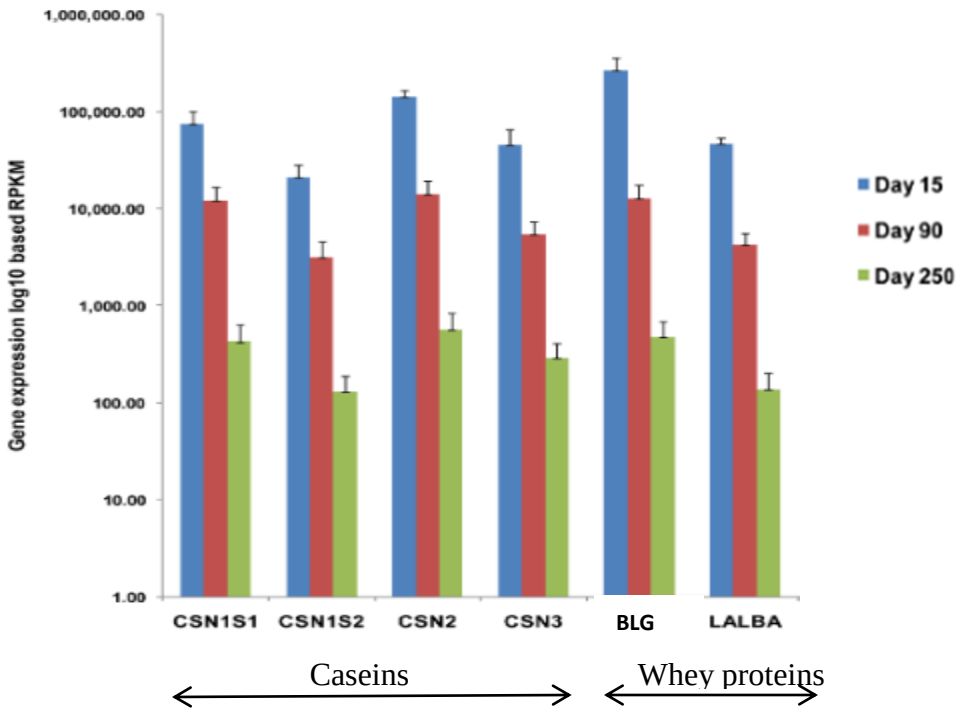


Figure 1. Variations in caseins and whey genes expression during bovine lactation stages. The x-axis represents gene symbols, while the y-axis displays log2 RPKM (reads per kilobase of transcript per million mapped reads) expression values at 15, 90, and 250 days. Notably, all genes reached their highest expression levels on day 15 and then declined significantly as lactation progressed. * Adapted from Wickramasinghe *et al*. (2012).

2.4. Transcriptional Regulation: Insights into Promoter Structure and Regulatory Elements

The primary and most critical stage in protein gene expression initiation is the dynamic collaboration of the promoter and RNA polymerase II (Thomas and Chiang, 2006). The promoter, a nucleotide region spanning tens to hundreds of base pairs either upstream or downstream of the TSS (Figure 2), regulates TSS initiation and transcription level (Haberle and Lenhard, 2016). Recognizing promoters is critical for understanding gene structure and transcriptional control. Traditional methods for identifying promoters often necessitate complex biological studies, such as DNA microarray (Brown and Botstein, 1999), DNAase footprinting (Galas and Schmitz, 1978), and chromatin immunoprecipitation (Javed *et al*., 2004).

Although traditional experimental procedures can provide precise and consistent results, identifying most promoters across the entire genome is a time-consuming and labor-intensive task. Large-scale high-throughput sequencing initiatives and the accumulation of genetic data

have given rise to promoter prediction methods based on bioinformatics and computational biology. This method may easily extract useful information from large amounts of genomic data and swiftly identify promoters (Liu *et al*., 2023).

Identifying TSS is an important first step in computational analysis, which leads to the prediction of regulatory elements. This task is particularly simple in lower eukaryotes due to their high gene density compared to genome size. For example, the yeast Saccharomyces cerevisiae dedicate around 70% of its genome to protein encoding, with short intergenic intervals of about 440bp (Goffeau *et al*., 1996). Conversely, the human genome has a reduced gene density, with only around 3% of the genome producing proteins (Venter *et al*., 2001), making it difficult to find both promoters and regulatory elements. Despite the complexities of gene expression regulation in higher eukaryotes, recent advances in investigating promoter-related features have resulted in the effectiveness of computational prediction methods based on machine learning for promoter prediction (Zhang *et al*., 2019; Lai *et al*., 2019; Bhandari *et al*., 2021). Furthermore, deep learning algorithms have been used in promoter prediction studies (Umarov *et al*., 2019; Amin *et al*., 2020; Tayara *et al*., 2020; Zhu *et al*., 2021; Zhang *et al*., 2022), which provide a comprehensive review of cutting-edge computational methods for predicting prokaryotic and eukaryotic promoter sequences.

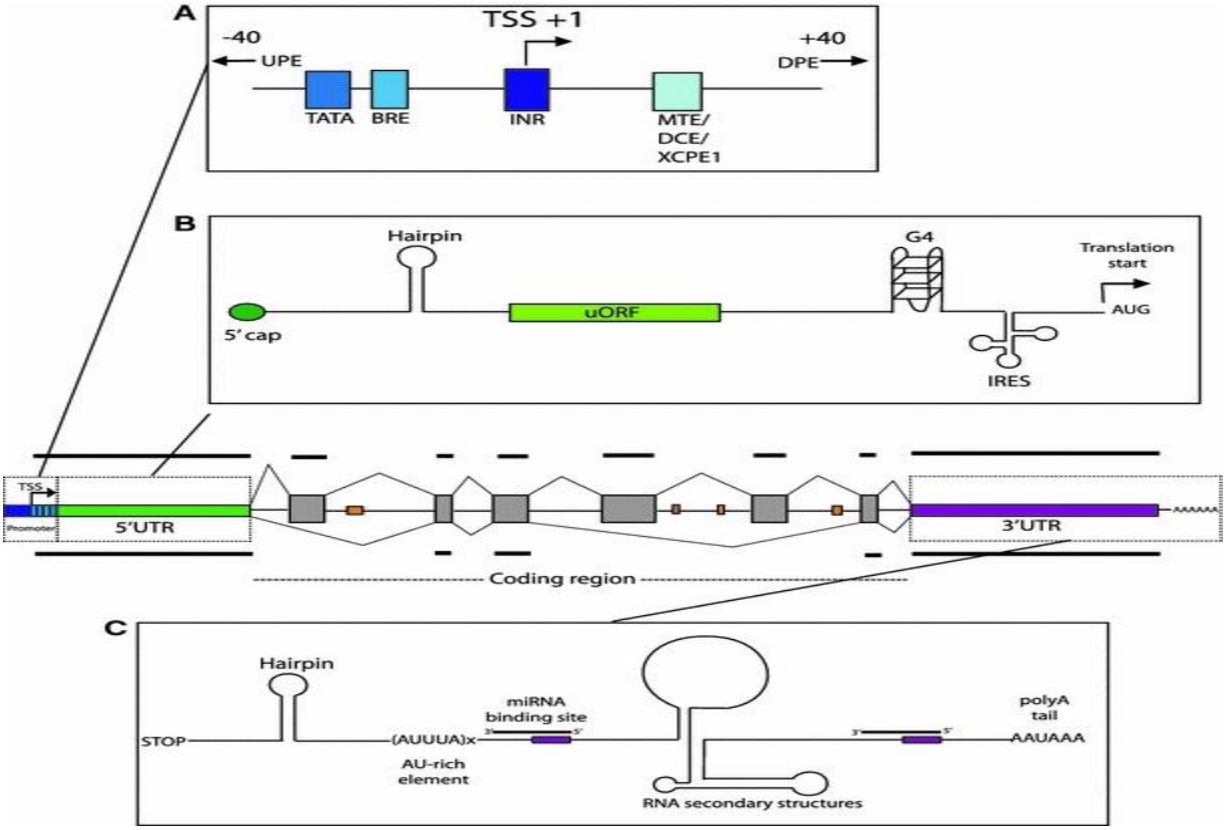


Figure 2. Regulatory elements within the noncoding regions.
The central graphic is a typical gene, with exons in grey. The orange rectangles indicate intronic enhancer elements. A). Regulatory elements in the promoter region. The arrows show upstream and downstream promoter elements situated outside of the core promoter region. B) Regulatory elements in the 5' UTR. C) Regulatory features of the 3' UTR (Barrett *et al*., 2012).

The promoter is an important *cis* component that controls the strength and timing of transcription. It controls gene expression and sheds light on transcriptional regulatory mechanisms in physiological and immunological processes. TFBSs are particular DNA sequences within the promoter to which TFs bind. TFs play an important role in regulating the expression of cell and tissue specific genes. Exploring regulatory elements such as TFBS modules in the promoter region can uncover co-regulated genes, allowing for the building of regulatory models and a thorough analysis across many species (Gorji *et al*., 2019).

Gene expression control at the transcriptional level is strongly dependent on how TFs interact with specific *cis*-DNA elements located in the gene's promoter region. The *cis*-regulatory region, which is typically a stretch of DNA spanning 100-1000 base pairs, contains clusters of TFBSs organized into modular architectures. These structured areas contain both positive and negative

regulatory elements, which coordinate signal transduction pathways via protein-DNA and protein-protein interactions. They establish diverse temporal and spatial gene expression patterns and frequently bring together TFs from different families at the same promoter region (Doppler *et al*., 2002). The amount of each protein found in milk correlates to the extent of transcription of the associated gene (Rosen *et al*., 1999).

2.5. Transcriptional Regulation of milk Genes

The transcription of genes encoding milk proteins is regulated by intricate interactions between TFs and composite response elements in the 5′-upstream region, which are controlled by lactogenic hormones. The presence of distinctive *cis*-regulatory motifs in the promoter and enhancer regions is required to recruit a distinct collection of TFs (Rosen *et al*., 1999; Kabotyanski *et al*., 2006; Patel *et al*., 2014). Signal transducer and activator of transcription 5 (STAT5), glucocorticoid receptor (GR), activator protein 1 (AP-1), and CCAAT enhancer binding protein (C/EBP) are key transcription factors that regulate milk protein gene expression (Rosen *et al*., 1999; Song *et al*., 2022). Furthermore, putative repressor elements such as Yin Yang 1 (YY1), CIS3, SOCS-1, and SOCS-3 contribute to milk gene expression (Feng *et al*., 2021).

The investigation of CSN1S1 gene promoters from various species has provided useful insights into the promoter's structure and conservation across species. The functional importance of the many *cis*-regulatory regions in milk protein genes has been primarily elucidated through studies of the beta-casein gene's regulatory region. These studies show how lactogenic hormones including prolactin, progesterone, and glucocorticoids regulate casein gene expression, as well as transcription factors like STAT5, Octamer-binding protein 1 (Oct1), C/EBP, and YY1. These factors regulate casein gene expression by interacting with particular binding motifs in the proximal promoter region. Along with the casein gene promoter, it was discovered that the whey protein gene promoters had binding sites for NFI (nuclear factor I), the glucocorticoid receptor, and STAT5 (Waston *et al*., 1991).

2.6. DNA Motifs in Gene Expression Regulation

DNA motifs are short, recurring patterns of nucleotides with biological significance found in regulatory regions of eukaryotic genes. These motifs, typically 6-12 base pairs in size, are

conserved and repeated, helping in the identification of TFBSs essential for understanding gene expression regulation mechanisms (Zhang *et al*., 2013; Bailey *et al*., 2015). Recognizing the regulatory motifs bound by TFs offers valuable insights into transcriptional regulation mechanisms (Liu *et al*., 2000). Transcription regulation is key in establishing and maintaining temporal and spatial gene expression patterns. The interaction between TFs and specific DNA sequences in promoter regions is crucial for precise control over when and where genes are expressed. TFs can activate or suppress milk gene expression by binding to their specific DNA sequence elements (Mariño-Ramírez *et al*., 2009).

2.7. Importance of Mammary Tissue-Specific Promoter Regulation

Studying the specific promoter region of mammary tissue provides valuable insights into its tissue-specific and developmental regulation, offering the potential for driving high-level expression of bioactive proteins in transgenic animal milk (Patel *et al*., 2014; Wodas *et al*., 2020; Du *et al*., 2020; Song *et al*., 2022). The promoter is a critical element that determines tissue specificity in transgene expression. Various promoters from genes encoding milk proteins have been effectively used to produce recombinant proteins in the MG (Shepelev *et al*., 2018). A thorough understanding of the mechanisms governing the regulation of mammary-specific milk protein genes is essential for improving milk production yield, quality, and efficiency. Moreover, it aids in identifying key signaling factors and pathways involved in mammary development, differentiation, lactation, and disease (Qian and Zhao, 2014a).

3. MATERIALS AND METHODS

3.1. TSS determination

Ensembl (Cunningham *et al*., 2022) and UCSC Genome Browser (Navarro Gonzalez *et al*., 2021) were used to identify canonical TSS accurately without the need for direct experimental investigation. The TSSs of the targeted gene were identified by accessing the gene summary page in a well-annotated Ensembl Genome browser. Subsequently, the promoter sequence was located by examining the region immediately upstream of the genes' first exon. Similarly, in the UCSC Genome Browser, the TSSs were found by navigating to the genes and gene prediction section.

3.2. Acquisition of Promoter Sequences

The 2 kb putative promoter sequences immediately upstream of the canonical TSS (or first exon) of CSN1S1, CSN1S2, CSN2, CSN3, LALBA, and LGB of human, horse, goat, cow, mouse, bonobo, lion, chimpanzee, Norway rat, sheep, Arabian camel, tiger, olive baboon, narwhale, macaque, megabat, domestic cat, dog, giant panda, rabbit, pig, kangaroo rat, and opossum were obtained from Ensembl and the UCSC Genome Browser. These 2 kb upstream regions for milk genes of these species are expected to contain putative TFBSs. The sequences were acquired in the Fasta format for this purpose.

3.3. Phylogenetic Tree Construction

The retrieved 2 kb putative promoter sequences immediately upstream of the TSS were imported into Jalview (Waterhouse *et al*., 2009). The sequences for each gene were aligned using Clustal Omega (Sievers and Higgins, 2014) within Jalview, using default parameters. The aligned sequences of the major milk genes of twenty-three mammalian species were utilized to construct a neighbor-joining (NJ) phylogenetic tree to sort into evolutionarily divergent groups, or clades.

3.4. Analysis of the Presence of Candidate Motifs and TFs

The upstream sequences for each clade were analyzed using the expectation maximization algorithm MEME-Suite (http://meme.nbcr.net) version 5.5.3, updated on June 15, 2023, by the National Biomedical Computation Resource (Bailey *et al*., 2015) to look for motifs and TFs that regulate the expression of CSN1S1, CSN1S2, CSN2, CSN3, LALBA, and BLG genes across the twenty-three mammalian species. This technique was used to determine the occurrence of motifs

that function as binding sites for the TFs expected to regulate the expression of milk protein genes.

The optional inputs in MEME included Classic mode for motif discovery, DNA for sequence alphabet, and one occurrence per sequence (oops) for site distribution, as these settings are generally sufficient for most motif findings (Peng *et al*., 2018). The number of motifs that MEME-Suite was instructed to find was set at five. Additionally, the motifs were set to be 6-25bp in length, with an E-value less than 0.05, which is the default parameter for MEME-Suite.

Each of the original putative promoter 2 kb sequences underwent random shuffling, maintaining the composition and number of nucleotides as in the primary sequence. This was done to compare and check if the discovered motifs could appear at random in sequences of the same length and composition when subjected to MEME analysis, using the same set of parameters used in the original sequences. This iterative technique allowed for the assessment of E-values associated with motifs identified in "random" sequence datasets that resemble the primary (original) dataset. Furthermore, this iterative approach assisted in defining a reasonable E-value threshold for motifs discovered in the unshuffled primary/original sequence. The results generated by the MEME-Suite algorithm were visualized in MEME HTML (Hyper Text Markup Language).

The MEME HTML format provided a structured framework for the *de novo* discovery of motifs and their respective distributions/locations along the sequence, illustrated through block diagrams. Following the acquisition of the MEME-Suite results, the discovered motifs underwent analysis using TomTom (https://meme-suite.org/meme/tools/tomtom), a component of the Meme-Suite (Gupta *et al*., 2007). This analysis utilized the JASPAR 2022 CORE vertebrate database with default parameters. TomTom revealed a notable similarity between the query motif and the binding motif (TFBS) present in the 2 kb upstream promoter regions of major milk protein genes. The top five TFs for each motif were collected using TomTom with a p-value significance cut-off of 0.01. Subsequently, UniProt IDs of the candidate TFs were obtained and used to assess their possible biological function.

3.5. Gene Ontology Analysis of the Discovered Candidate Motifs

The functional roles of the discovered motifs were determined using the Gene Ontology for candidate Motifs (GOMo) web-server version 5.5.3, hosted by the National Biomedical Computation Resource at (https://meme-suite.org/meme/tools/gomo) (Buske *et al*., 2010). GOMo scans the discovered motifs to determine if any motif is significantly associated with genes linked to one or more Gene Ontology (GO) term(s). For GOMo's optional settings, a significance threshold q-value of less than 0.05 was employed, with 1000 rounds of score shuffling, and multiple species for each clade were included in the analysis. Q-value denotes the lowest acceptable false discovery rate (FDR) at which a particular GO term is considered significant. Meanwhile, the number of score shuffling rounds was used by GOMo to calculate the empirical p-values. This involves rearranging the association between sequences and their scores through shuffling, which ultimately determines the likelihood of the observed outcome being a chance occurrence. The identification of significant GO terms provided insights into the potential biological roles of the motifs.

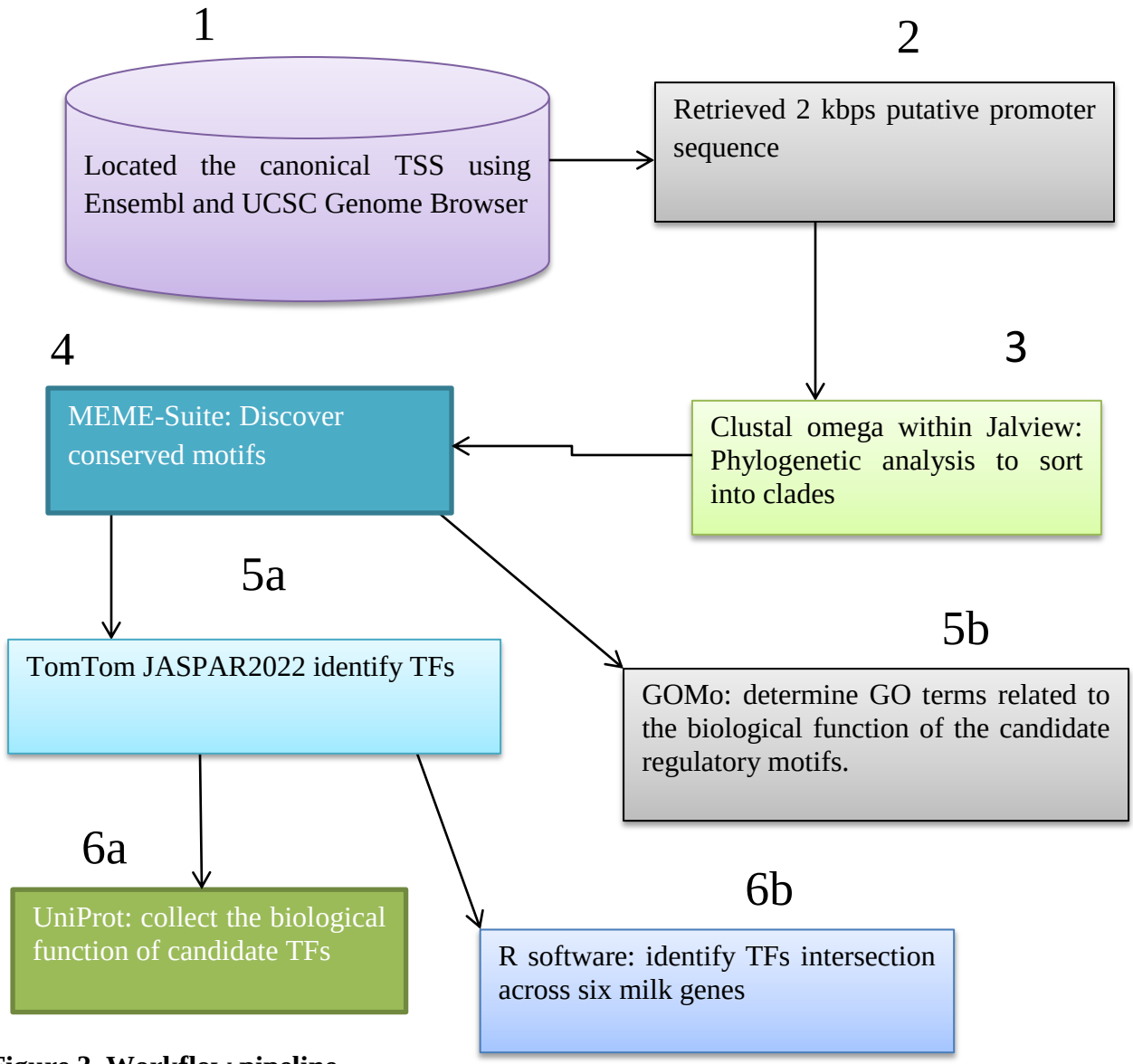


Figure 3. Workflow pipeline.
(1) Ensembl and UCSC Genome Browser were used to locate the canonical TSS using annotation information. (2) Retrieve 2 kb upstream putative promoter sequences of CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG for 23 species. (3) Phylogenetic analysis using Clustal omega within jalview was performed to place species into their respective clades. (4) All upstream sequences for each clade were run through MEME-Suite. (5a) The resulting identified motifs were searched through the TomTom JASPAR2022 CORE vertebrate database to find matches to known TFs. (5b) the biological GO term(s) for each motif were obtained using GOMo. (6a) UniProt IDs were collected for each TF to determine the biological function of candidate TFs. (6b) R software was used in identifying common TFs intersection across six milk genes.

4. RESULTS

4.1. Identification of the TSS

A comprehensive investigation of the TSSs of six (6) milk protein genes (CSN1S1, CSN1S2, CSN2, CSN3, LALBA, and BLG) across 23 mammalian species was undertaken. Using the Ensembl and UCSC genome browsers, the locations of the canonical TSSs were precisely determined. Furthermore, a 2 kb sequence upstream of each TSS, which encompasses the critical putative promoter region, was retrieved from the accession number shown in Table 2.

4.2 CSN1S1 Putative Promoter

4.2.1 Phylogenetics of CSN1S1 Promoter Region

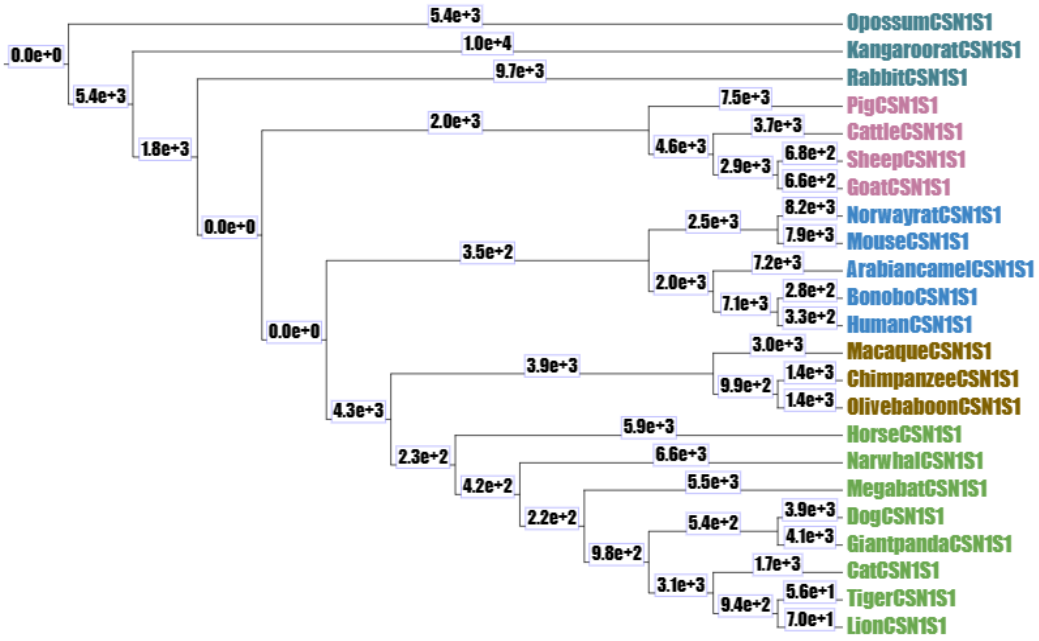


Figure 4. Phylogenetic tree of CSN1S1 gene 2 kb putative promoter with branch lengths. Different colors indicate different clades.

The phylogenetic tree in Figure 4, generated using the Clustal-Omega package in Jalview software, illustrates the evolutionary relationships of the 2 kb putative promoter region of CSN1S1 across 23 taxa. Based on close evolutionary distance, the tree shows five distinct clades from their common ancestor. This shows relatively a closer evolutionary distance among the sequences of the first clade that contains lions, tigers, cats, giant pandas, and dogs. Similarly, the

second clade groups olive baboons, macaques, and chimpanzees, in congruence with the species' close evolutionary relationship. Humans, Arabian camels, bonobos, mice, and Norway rat also form a distinct clade, showing the shared evolutionary history of the putative promoter region. Similarly, the tree also reveals a close evolutionary distance among goats, sheep, cattle, and pigs, further supporting their shared lineage within the ungulate group. However, opossum and kangaroo rat are placed in Clade five as an out-group in the tree. Rabbits, which belong to the order *lagomorpha,* are also placed in this out-group. This shows the evolution of the CSN1S1 putative promoter region doesn't closely reflect the species-level phylogenetic relationship among the 23 species considered. Humans and their very close relatives, the bonobos, are placed in the same clade as the rodent, mouse Norway rat, and Arabian camels, Rabbit, a *lagomorpha* which is closer to order *rodentia,* is placed in the clade for the out-group. The same is true for the kangaroo rat, another rodent, which is placed along with the out-group clade. This demonstrated that since they last shared a common ancestor, the 2kb putative promoter region of CSN1S1 has shown a rate of evolution that is not in concordance with the species level evolution.

Table 2. Accession numbers: Major milk genes inclusive of 2kb putative promoter region.

Species	Gene Symbol	Respective accession number
Human	CSN1S1, CSN2, CSN3 and LALBA	ENST00000246891.9, ENST00000353151.4, ENST00000304954.4, ENST00000301046.6
Mouse	CSN1S1, CSN2, CSN3 and LALBA	ENSMUST00000094641.9, ENSMUST00000197422.5, ENSMUST00000113271.3, ENSMUST00000023726.5,
Cattle	CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG	ENSBTAG00000007695, NM_174528, ENSBTAT00000003409.7, ENSBTAT00000028685.6, ENSBTAT00000007701.3, ENSBTAT00000019538.7
Goat	CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG	XM_018049133.1, XM_013964678.2, XM_013964699.2, NM_001285587.1, ENSCHIT00000032014.1, ENSCHIT00000020316.1
Sheep	CSN1S1,CSN1S2, CSN2, CSN3, LALBA and BLG	NM_001009795, NM_001009363, NM_001009373, ENSOART00020001919.2, ENSOART00000020933.1, ENSOART00020007400.2
Horse	CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG	ENSECAT00000015064.4, NM_001170767, ENSECAT00000064283.3, ENSECAT00000001652.3, XM_001915789.4, ENSECAT00000013425.2
Norway rat	CSN1S1, CSN2, CSN3 and LALBA	ENSRNOT00000088491.2, ENSRNOT00055028349.1, ENSRNOT00055028077.1, ENSRNOT00060011011.1
Tiger	CSN1S1, CSN2, CSN3 and LALBA	ENSPTIT00000010778.1 , ENSPTIT00000016113.1, ENSPTIT00000019918.1 and ENSPTIT00000022769.1
Lion	CSN1S1, CSN2, CSN3 and LALBA	ENSPLOT00000017378.1, ENSPLOT00000017548.1, ENSPLOT00000030439.1, ENSPLOT00000013724.1
Cat	CSN1S1, CSN2, CSN3, LALBA and BLG	ENSFCAT00000054514.1, ENSFCAT00000000640.6, ENSFCAT00000014279.5, ENSFCAT00000086163.1, ENSFCAG00000005742
Dog	CSN1S1, CSN2, CSN3 and LALBA	ENSCAFT00845018004.1, ENSCAFT00845018124.1, ENSCAFT00845018457.1, ENSCAFT00845044950.1
Bonobo	CSN1S1, CSN2, CSN3 and LALBA	ENSPPAT00000045088.1, ENSPPAT00000000084.1, ENSPPAT00000048405.1, ENSPPAT00000037413.1

Species	Gene symbol	Respective accession number
Chimpanzee	CSN1S1, CSN2, CSN3 and LALBA	ENSPTRT00000030034.5, ENSPTRT00000030036.4, ENSPTRT00000030043.3, ENSPTRT00000009035.5
Narwhal	CSN1S1, CSN2, CSN3 and LALBA	ENSMMNT00015015381.1, ENSMMNT00015031619.1, ENSMMNT00015031540.1, ENSMMNT00015031882.1
Pig	CSN1S1, CSN1S2, CSN2, CSN3,LALBA and BLG	NM_001004029, NM_001004030, ENSSSCT00000010141.5, ENSSSCT00000010148.5, ENSSSCT00000025257.3, ENSSSCT00030031191.1
Rabbit	CSN1S1, CSN2, CSN3 and LALBA	NM_001082391, ENSOCUT00000009490.2, ENSOCUT00000045471.1, NM_001082052.1
Macaque	CSN1S1, CSN2, CSN3 and LALBA	ENSMMUT00000064058.2, ENSMMUT00000042085.3, ENSPVAT00000009961.1, ENSMMUT00000000415.4
Olive baboon	CSN1S1, CSN2, CSN3 and LALBA	ENSPANT00000048822.2, ENSPANT00000023656.3, ENSPANT00000005673.3, ENSPANT00000073781.1
Giant panda	CSN1S1, CSN2, CSN3 and LALBA	ENSAMET00000044764.1, ENSAMET00000029205.1, ENSAMET00000006379.2, ENSAMET00000008076.2
Arabian camel	CSN1S1, CSN2, CSN3 and LALBA	ENSCDRT00005012639.1, NM_001303563.1, ENSCDRT00005030424.1, ENSCDRT00005015260.1
Mega bat	CSN1S1, CSN2, CSN3 and LALBA	ENSPVAT00000012919.1, ENSPVAT00000012922.1, ENSPVAT00000009961.1, ENSPVAT00000017943.1
Kangaroo rat	CSN1S1, CSN2, CSN3 and LALBA	XM_042678378.1, XM_013020075.1, XM_042678381.1, ENSDORT00000009838.2
Opossum	CSN1S1, CSN2, CSN3 and LALBA	NC_077232.1, XM_044681038.1, NC_077232.1, XM_007502942.2

(Source: Ensembl and UCSC genome browser)
Accession numbers starting with the prefix ENS were retrieved from Ensembl and all others are from the UCSC Genome Browser.

4.2.2. Identification of Candidate Motifs and Associated TFs of CSN1S1

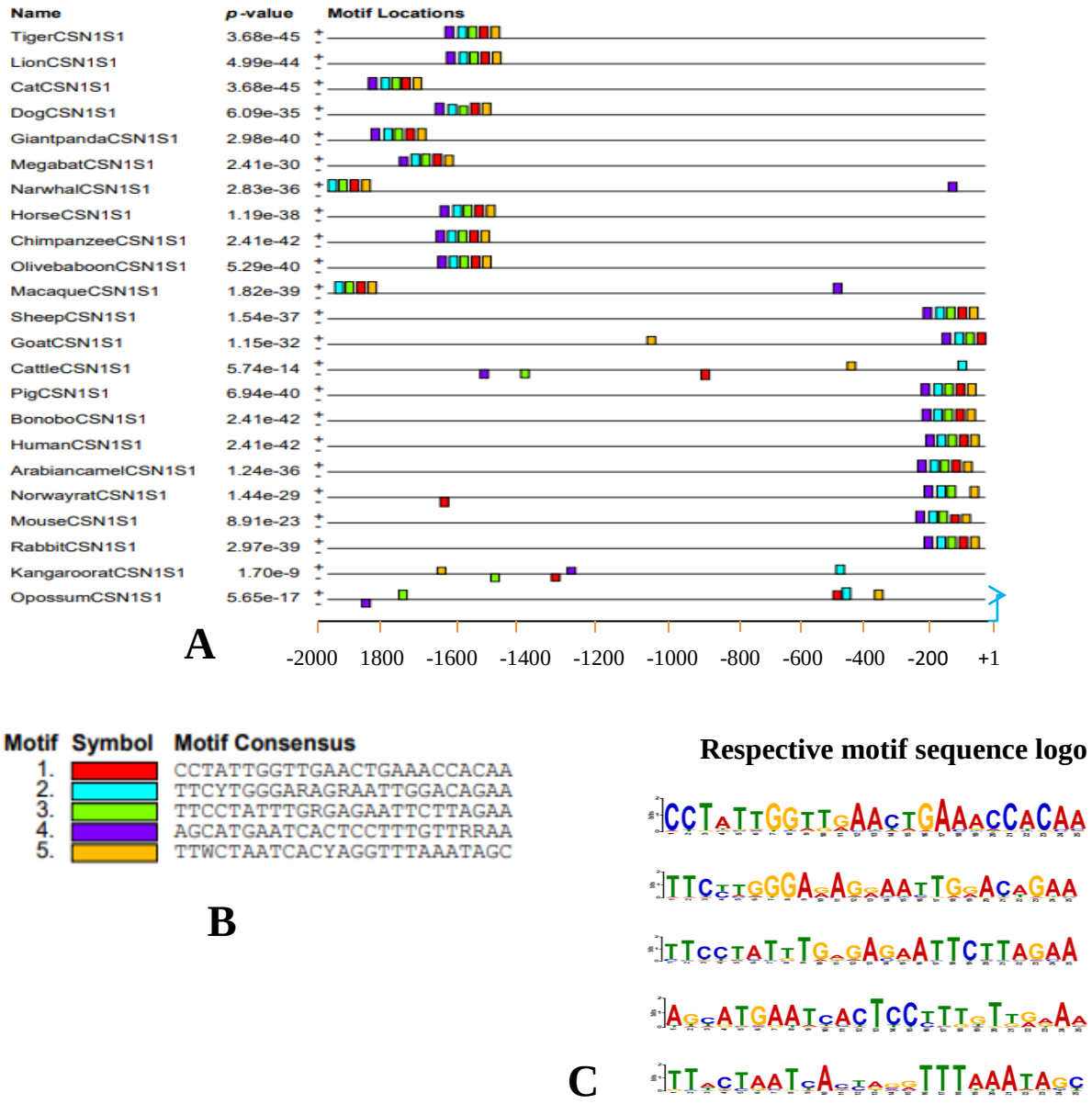


Figure 5. CSN1S1 predicted putative promoter profile.
A) Promoter motif distribution relative to TSSs. B) Promoter motifs' symbols and consensus sequence, and (C) predicted promoter motif sequence logos (The height of a letter indicates its relative frequency at the given position in the motif).

Figure 5 illustrates the results of the analysis of the 2kb putative promoter region of CSN1S1 using the MEME-Suite algorithm on 23 mammalian species. The analysis sought to identify the top five conserved motifs with an e-value less than 0.05 within the CSN1S1 putative promoter region, with motif widths ranging from 6 to 25 nucleotides. The top five discovered motifs

appear in a cluster on the positive strand in the same relative order and distance among motifs in all eutherian species, except narwhale, macaque, goat and cattle. The motif profiles in kangaroo rats and opossums are markedly different from the others and their p-values, though significant, are the highest among the species considered. Interestingly, the cattle putative promoter p-value of $5.7e^{-14}$, though expected to resemble the motif profile of the other ungulates (sheep and goat) does not appear to reflect the same conserved regularity and order of the identified five motifs. Further, the cluster of motifs in the caprine (goat) putative promoter region is the closest to the canonical TSS and does not include motif 5. This may be due to incomplete data acquisition and the motif may reside in the 5' UTR. Also in macaque and Narwhale, only four motifs are discovered with motif four not showing in the examined 2kb sequences. It is possible that motif 4 is further 5' of the 2kb sequence and as such not reflected in the cluster of motifs. The motifs found in the opossum and kangaroo rat, are distinctly different from the pattern observed in the other species. Notably, 94.8% of the identified motifs are positioned on the positive strand of the DNA.

Based on the cluster of motifs' relative distance from the canonical TSSs, they broadly fall into two groups (not including cattle, kangaroo rats and opossums). In those species listed from tiger to macaque (45%), the motifs reside within the distal promoter region, located over -1.5kb upstream of canonical TSS; while the remaining (55%), not including cattle, localize in the proximal putative promoter region to within less than -240bp upstream from the canonical TSSs. The motifs in cattle and kangaroo rats and opossums appear scattered across the examined putative promoter region and do not exhibit the same relative order, strand orientation, and distance from the TSS. The obtained p-values for cattle, kangaroo rats, and opossums are the relatively highest as compared to the rest of the sample species, thus ascribing to them a relatively lower level of significance.

After subjecting the conserved motifs in the CSN1S1 putative promoter region to the TOMTOM software, the top 5 most significant TFs for each motif with p-value less than 1.0e-02 were selected, which resulted in the identification of a total of 22 TFs (Table3).

Table 3. CSN1S1 TFs match the query motif in the JASPAR2022 CORE vertebrates.

Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
RUNX2	Runt-domain factors	Runt-related factors	MA0511.2	Q13950	$2.68e^{-04}$
ZBED2	C_2H_2 zinc finger factors	BED zinc finger factors	MA1971.1	Q9BTP6	$2.44e^{-03}$
FOXH1	Fork head/winged helix factors	FOX	MA0479.1	O75593	$2.45e^{-03}$
SIX2	Homeo domain factors	HD-SINE	MA1119.1	Q9NPC8	$2.59e^{-03}$
RUNX1	Runt-domain factors	Runt-related factors	MA0002.2	Q01196	$3.30e^{-03}$
Matched TFs with motif 2	Class	Family	Matrix ID	UniProt ID	P-value
NFIC	SMAD/NF-IDNA-binding domain factors	Nuclear factor 1	MA0161.2	P08651	$1.41e^{-03}$
GABPA	Tryptophan cluster factors	Ets-related	MA0062.3	Q06546	$5.30e^{-03}$
STAT1	STAT domain factors	STAT factors	MA0137.3	P42224	$4.40e^{-03}$
STAT5a::STAT5b	STAT domain factors	STAT factors	MA0519.1	P42230::P42232	$4.40e^{-03}$
STAT3	STAT domain factors	STAT factors	MA0144.2	P40763	$1.86e^{-03}$
Matched TFs with motif 3	Class	Family	Matrix ID	UniProt ID	P-value
STAT5a	STAT domain factors	STAT factors	MA1624.1	P042230	$3.32e^{-03}$
BCL6B	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA0731.1	A8KA13	$4.87e^{-03}$
Matched TFs with motif 4	Class	Family	Matrix ID	UniProt ID	P-value
KLF13	C_2H_2 zinc finger factors	Three-zinc finger Kruppel related	MA0657.1	Q9Y2Y9	1.84^{e-03}
Elf5	Tryptophan cluster factors	Ets-related	MA0136.3	Q8VDK3	$3.10e^{-03}$
TBX1	T-Box factors	TBX1-related factors	MA0805.1	O43435	$3.12e^{-03}$
FOS1L::JUND	Basic leucine zipper factors (bZIP)	Fos-related::Jun related	MA1142.1	P15407::P17535	$3.15e^{-03}$
FOSL2	Basic leucine zipper factors (bZIP)	Fos-related	MA0478.1	P15408	$3.32e^{-03}$
Matched TFs with motif 5	Class	Family	Matrix id	UniProt ID	P-value
SATB1	Homeo domain factors	HD-CUT	MA1963.1	Q01826	$6.09e^{-04}$
MEF2B	MADS-box factors	Regulator of differentiation	MA0660.1	Q02080	$6.38e^{-04}$
ARGFX	Homeo domain factors	Paired-related HD factors	MA1463.1	A6NJG6	$9.04e^{-04}$
HOXD3	Homeo domain factors	HOX	MA0912.2	P31249	$2.03e^{-03}$
LBX1	Homeo domain factors	NK	MA0618.1	P52954	$2.78e^{-03}$

4.2 CSN1S2 Putative Promoter

4.2.1 Phylogenetics of CSN1S2 Promoter Region

In Figure 6, the phylogenetic tree of the putative promoter regions of CSN1S2 is depicted across five species for which sequence data were available publically using Clustal-Omega within the Jalview software. The complete 2 kb upstream genomic sequences of human, mouse, bonobo, lion, chimpanzee, Norway rat, Arabian camel, tiger, olive baboon, narwhale, macaque, megabat, domestic cat, dog, giant panda, rabbit, kangaroo rat, and opossum were not publicly available or are CSN1S2 is not expressed. The tree shows that ruminants (goats, sheep, and cattle) formed a clade in the evolutionary relationships with a relatively higher degree of sequence similarity in their CSN1S2 putative promoter region than cattle do with either goats or sheep. This is evident from the shorter branch length between goats and sheep compared to the branch length between cattle and either goats or sheep. The pig was found in a separate clade in the tree with a branch length of $4.8e^3$. From their last shared common ancestor, pigs appear to have evolved slower than the other four species while, horses seem to have evolved the most (with a branch length of $13.9e^3$ since the point of differentiation from pigs.

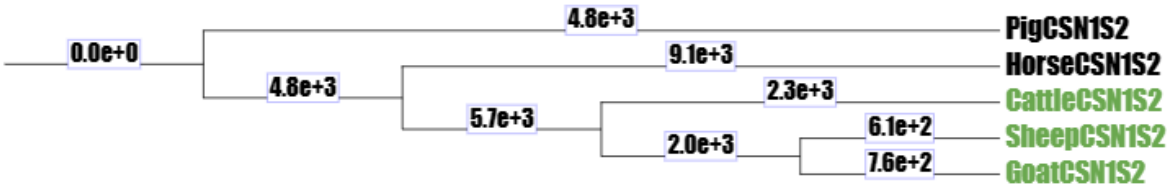


Figure 6. Phylogenetic tree of CSN1S2 gene 2 kb putative promoter with branch lengths.

4.2.2 Identification of Candidate Motifs and Associated TFs of CSN1S2

Figure 7A showcases the conserved motifs obtained by applying Meme-Suite algorithms to analyze the CSN1S2 putative promoter region across the 5 species. The analysis focused on identifying motifs with a minimum and maximum width of 6 and 25bp, respectively, selecting those with an e-value of less than 0.05. Figure 5A illustrates that motifs 1, 2, 3, and 5 appear in a cluster within -280bp upstream of the TSS in the same relative order and relative distance from each other. Motif 4, however, is the most distal from the TSS ranging from -810 to -500bp

upstream of the TSS. Figure 5C shows that the sequences of motifs 1, 2, 3, and 5 exhibit near-perfect conservation across all five species, while motif 4 also demonstrates strong conservation. It is noteworthy that all of these motifs are present on the positive DNA strand.

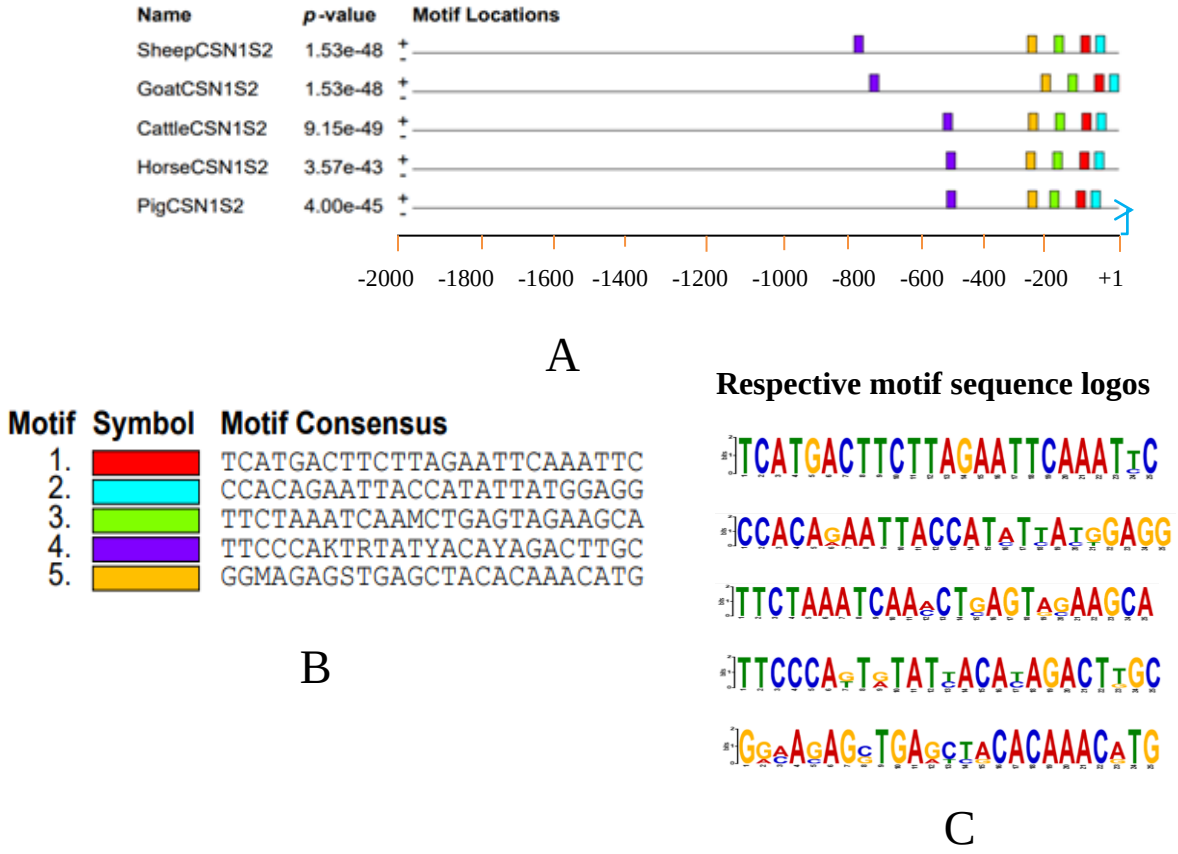


Figure 7. CSN1S2 predicted putative promoter profile.
A) Promoter motif distribution relative to TSSs. B) Promoter motif symbols and consensus sequence, and (C) predicted promoter motif sequence logos (The height of a letter shows its relative frequency at the given position in the motif).

The conserved motifs obtained from the Meme-Suite in the putative promoter sequence of CSN1S2 were subjected to the Tomtom algorithm by querying the JASPAR Core vertebrate database of known TFs; the top 5 TFs for each motif with p-values less than 1.0e-02 were selected. A total of 18 TFs were identified (Table 4). Motif 2 only contained TFs of the Pou domain factors. While motif 4, whose relative position from the TSS was the most distal and whose location range is broader than the other motifs, contained only one matching TF Sox2.

Table 4 CSN1S2 TFs match the query motif in the JASPAR2022 CORE vertebrates.

Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
LIN54	CRC domain	-	MA0619.1	FINCE0	$1.83e^{-04}$
BCL6B	C_2H_2 zinc finger factors	More than 3 adjacent zinc finger factors	MA0731.1	A8KA13	$2.00e^{-04}$
PROX1	Homeo domain factors	HD-PROS factors	MA0794.1	Q92786	$1.05e^{-03}$
STAT5a	STAT domain factors	STAT factors	MA1621.1	P42230	$2.05e^{-03}$
STA5b	STAT domain factors	STAT factors	MA1625.1	P42232	$2.68e^{-03}$
Matched TFs with motif 2	Class	Family	Matrix ID	UniProt ID	P –value
POU2F1	Homeo domain factors	POU domain factors	MA0785.1	P14859	$3.61e^{-04}$
POU3F2	Homeo domain factors	POU domain factors	MA0788.1	P20264	$1.78e^{-03}$
POU1F1	Homeo domain factors	POU domain factors	MA0784.2	P28069	$3.25e^{-03}$
POU3F4	Homeo domain factors	POU domain factors	MA0789.1	P49335	$3.86e^{-03}$
POU3F2	Homeo domain factors	POU domain factors	MA0787.1	P20265	$4.34e^{-03}$
Matched TFs with motif 3	Class	Family	Matrix ID	UniProt ID	P –value
RGFX	Homeo domain factors	Paired-related HD factors	MA1463.1	A6NJG6	$6.32e^{-03}$
STAT1::STAT2	STAT domain factors	STAT factors	MA0517.1	P42224::P52630	$7.20e^{-03}$
Matched TFs with motif 4	Class	Family	Matrix ID	UniProt ID	P-value
SOX2	High-mobility group (HMG) domain factors	SOX-related factors	MA0143.4	P48431	$4.29e^{-03}$
Matched TFs with motif 5	Class	Family	Matrix ID	UniProt ID	P-value
Ptf1A	Basic helix-loop-helix factors (bHLH)	Tal-related	MA1619.1	Q9QX98	$1.72e^{-03}$
FOXD2	Fork head/winged helix factors	FOX	MA0847.3	O60548	$1.80e^{-03}$
Bhlha15	Basic helix-loop-helix factors (bHLH)	Tal-related	MA1472.2	Q9QYC3	$2.90e^{-03}$
MYF5	Basic helix-loop-helix factors (bHLH)	MyoD/ASC-related factors	MA1641.1	P13349	$4.03e^{-03}$
ZNF667	C_2H_2 zinc finger factors	More than 3adjacent zinc finger factors	MA1984.1	Q5HYK9	$4.64e^{-03}$

4.3 CSN2 Putative Promoter

4.3.1 Phylogenetics of CSN2 Promoter Region

The CSN2 putative promoter among 23 distinct species was aligned using the Clustal-Omega package within the Jalview software for phylogenetic analysis, and it reveals six distinct clades (Figure 8). The tree indicates a closer evolutionary relationship among the sequences of putative promoters in the first clade namely; humans, chimpanzees, dogs, horses, and rabbits. The second clade consists of sheep, goats, cattle, pigs, and Arabian camels. Similarly, mice and Norway rat form the third clade, and then the fourth clade follows with species including lion, tiger, cat, giant panda, megabat, and narwhal. Olive baboons and macaques form the fifth clade. Lastly, the sixth clade represents the bonobo, kangaroo rat and opossum, which form the out-group whose sequence does not closely related to any of the other species.

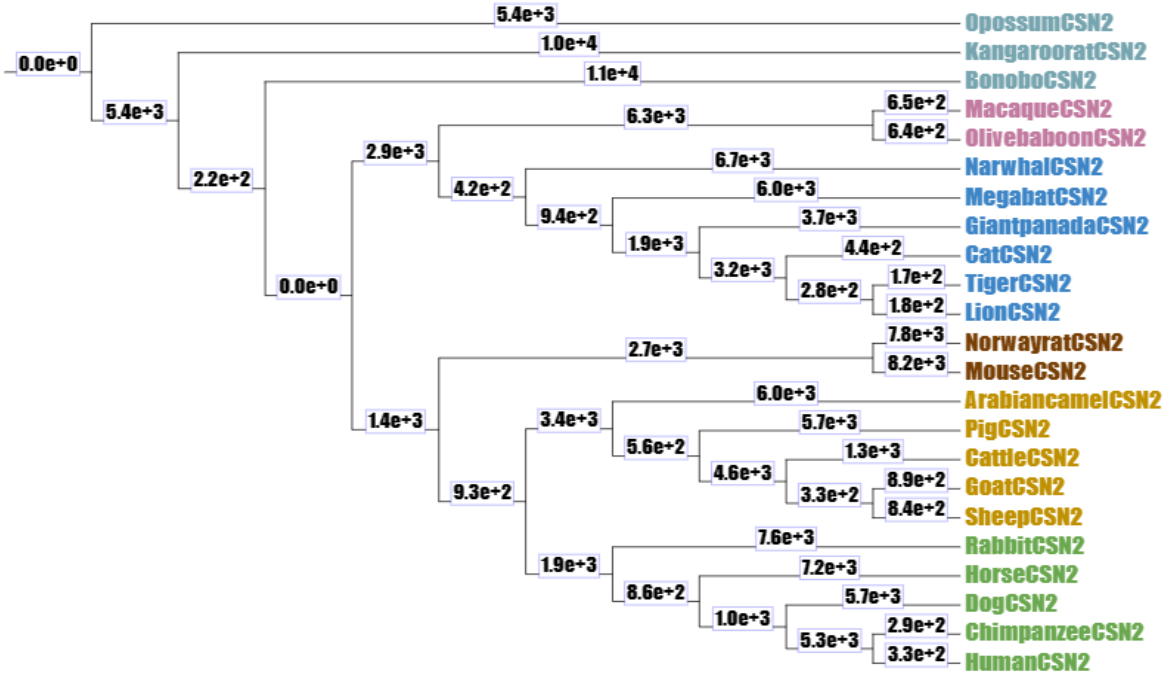


Figure 8. Phylogenetic tree of CSN2 gene 2 kb putative promoter with branch lengths.

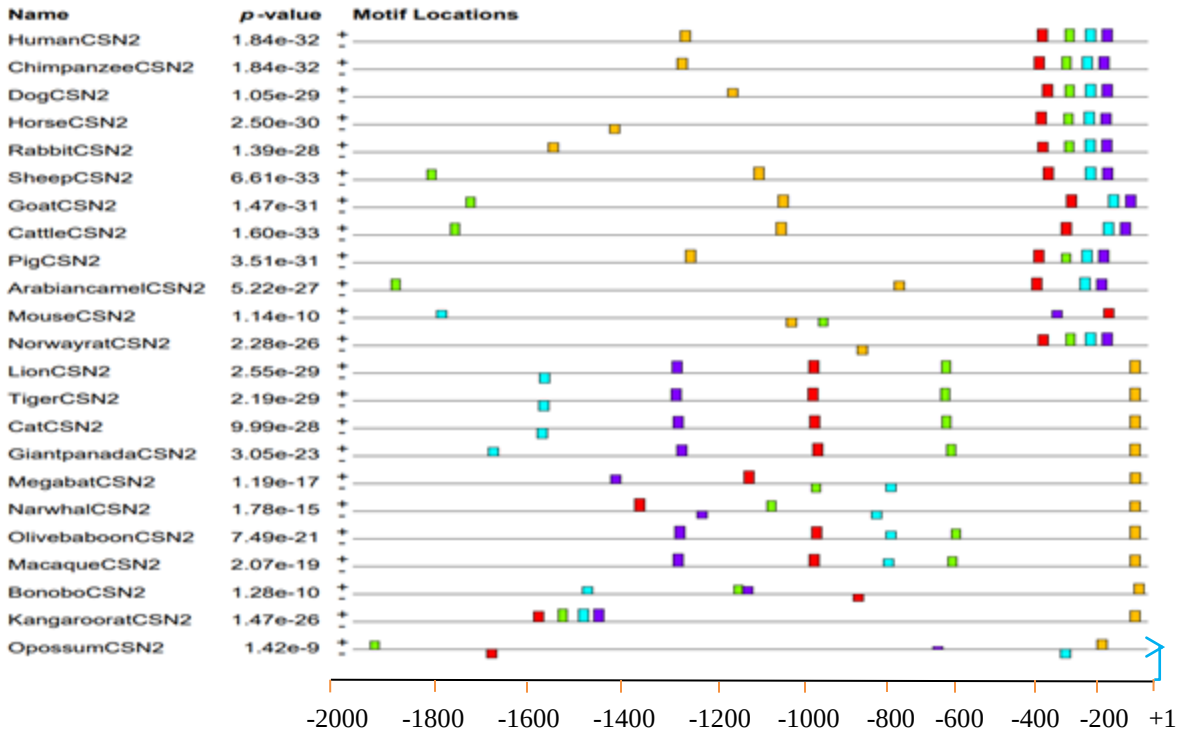
As would be expected, the ruminants and cats fall within the same clade, however, primates fall within three clades, the last of which is within the outlier group that includes opossums.

4.3.2 Identification of Candidate Motifs and Associated TFs of CSN2

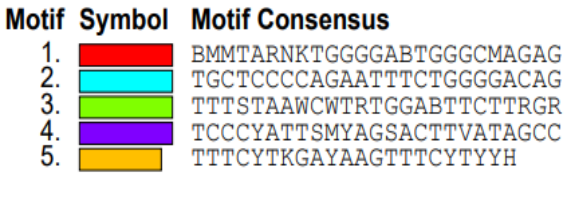
The top five motifs in the putative promoter sequence of the CSN2 gene upstream of the TSS were identified using the MEME-Suite algorithm with an e-value threshold of 0.05 and motif widths set to between 6 and 25 nucleotides. Based on the motifs relative order and distance it appears that those listed from human to Norway rat (p-value ranging from $1.84e^{-32}$ to $2.28e^{-26}$) follow a similar pattern, except motifs discovered in mice (p-value of $1.14e^{-10}$). Particularly motifs 1, 2, 3 and 4 form a cluster residing in the region within -390bp in the species listed from human to compartmented stomach from TSS. The motif presentation in the remaining species shows similarity among them but differs distinctly from those mentioned earlier. Motif 5 is located near the TSS within -40bp within the core promoter in the species listed from lion to kangaroo rat (Figure 7A).

Similarity can also be observed in the relative order and distance of motifs from the canonical TSS in those listed from lion to giant panda (p-values ranging from $2.55e^{-29}$ to $1.42e^{-9}$). Though similar motifs are also found in the remaining species, they do not appear to follow the same strict ordering and relative distance from the canonical TSS (Figure 7A). It is also worthwhile to note that the motif 5 consensus sequence is shorter than the rest with only 22 residues (Figure 9B). The motifs discovered in opossum, as would be expected, are markedly different from the other 22 species. While kangaroo rat exhibits the same cluster of motifs 1, 2, 3, and 4 presented in the same relative order, however in contrast to the group containing human to Norway rats, the cluster is located upstream of -1420bp relative to the canonical TSS, with exception of motif 5 located very close to TSS.

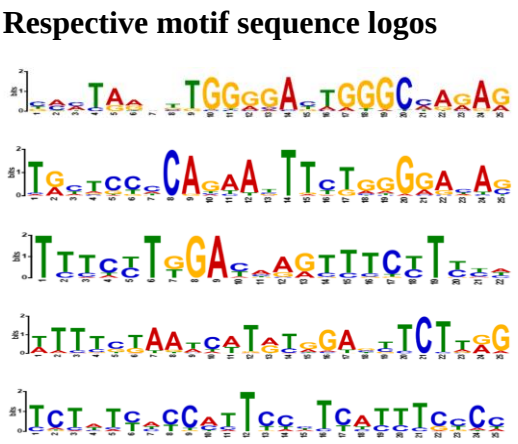
The motifs identified in the CSN2 gene promoter using the MEME-Suite algorithm were further analyzed using the TOMTOM algorithm to determine the top five TFs associated with each motif with p-values less than $1.0e^{-02}$. This analysis yielded 23 distinct TFs (Table 5). Motifs 2 and 3 are largely composed of the STAT factors family of TFs. The next most prevalent family of TFs is the zinc finger which appears in all but motifs 3 and 4.



A



B



C

Figure 9. CSN2 predicted putative promoter profile.
A) Promoter motif distribution relative to TSSs. B) Promoter motifs symbols and consensus sequence, and (C) predicted promoter motif sequence logos (The height of a letter shows its relative frequency at the given position in the motif)

Table 5. CSN2 TFs match the query motif in the JASPAR2022 CORE vertebrates.

Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
KLF15	C_2H_2 zinc finger factors	Three zinc factor Kruppel-related	MA1513.1	U9UIH9	$3.19e^{-04}$
ZNF682	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1599.1	O95780	$5.09e^{-04}$
MZF1	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA0056.2	P28698	$6.33e^{-04}$
FOXH1	C_2H_2 fork head/winged helix factors	FOX	MA0479.1	O75593	$1.08e^{-03}$
GSC2	C_2H_2 homeo domain factors	Paired-related HD factors	MA0891.1	O15499	$1.16e^{-03}$
Matched TFs with motif 2	Class	Family	Matrix ID	UniProt ID	P-value
ZNF416	C_2H_2 zinc finger factors	More than 3adjacent zinc fingers	MA1979.1	Q8NA42	$1.92e^{-03}$
STAT5a::STAT5b	STAT domain factors	STAT factors	MA0519.1	P42230::P42232	$2.13e^{-03}$
STAT5b	STAT domain factors	STAT factors	MA1625.1	P42232	$3.31e^{-03}$
STAT5a	STAT domain factors	STAT factors	MA1624.1	P42230	$3.90e^{-03}$
STAT1	STAT domain factors	STAT factors	MA0137.3	P42224	$8.02e^{-03}$
Matched TFs with motif 3	Class	Family	Matrix ID	UniProt ID	P-value
STAT5a	STAT domain factors	STAT factors	MA1624.1	P42230	$5.08e^{-04}$
STAT1	STAT domain factors	STAT factors	MA0137.3	P42224	$5.96e^{-04}$
STAT3	STAT domain factors	STAT factors	MA0144.2	P40763	$3.39e^{-03}$
Six4	Homeo domain factors	HD-SINE	MA2001.1	Q61321	$5.87e^{-03}$
STAT5a::STAT5b	STAT domain factors	STAT factors	MA0519.1	P42230::P42232	$6.27e^{-03}$
Matched TFs with motif 4	Class	Family	Matrix ID	UniProt ID	P-value
GATA6	Other C_4 zinc finger-type factors	C_4-GATA –related	MA1104'2	Q92908	$4.29.e^{-03}$
TRPS1	Other C_4 zinc finger type-actors	C_4-GATA-related	MA1970.1	Q9UHF7	$5.10e^{-03}$
ONECUT1	Homeo domain factors	HD-CUT	MA0679.2	Q9UBC0	$5.94e^{-03}$
GATA1	Other C_4 zinc finger-type factors	C_4-GATA-related	MA0035.4	P15976	$6.24e^{-03}$
Matched TFs with motif 5	Class	Family	Matrix ID	UniProt ID	P-value
ZNF24	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1124.1	P17028	$1.14e^{-04}$
EWSR1::FLI1	Tryptophan cluster factors	Ets-related	MA0149.1	F1JVV7::F1JVV8	$2.94e^{-04}$
SPi1	Tryptophan cluster factors	Ets-related	MA0080.6	P17433	$3.22e^{-04}$
ZKSCAN5	C_2H_2 zinc finger factors	Factors with multiple dispersed zinc fingers	MA1652.1	Q9Y2L8	$4.86e^{-04}$
SPIB	Tryptophan cluster factors	Ets-related	MA0081.2	Q01892	$4.99e^{-04}$

4.4 CSN3 Putative Promoter

4.4.1 Phylogenetics of CSN3 Promoter Region

The phylogenetic tree of CSN3 putative promoter sequences presented in Figure 10 was constructed using the Clustal Omega multiple sequence alignment package within Jalview software and it elucidated eight distinct clades which is the highest among all milk gene studied. The clade constructions were based on evolutionary distances, depicting the phylogenetic relationships among the 23 species. Clustering in clade one includes goats, bonobos, chimpanzee goats humans, and olive baboons, while clade two comprises Arabian camels, horses, and pigs. Clade three encompasses sheep, goats, and cattle. Similarly, clade four consists of dogs and cats followed by clade five containing mice and Norway rats. Species such as lion, tiger, and giant panda form clade six, and mega bats, macaques, and narwhals form the seventh clade in the phylogenetic tree. However, the last clade namely: kangaroo rat, rabbit, and opossum were placed as an out-group in the tree.

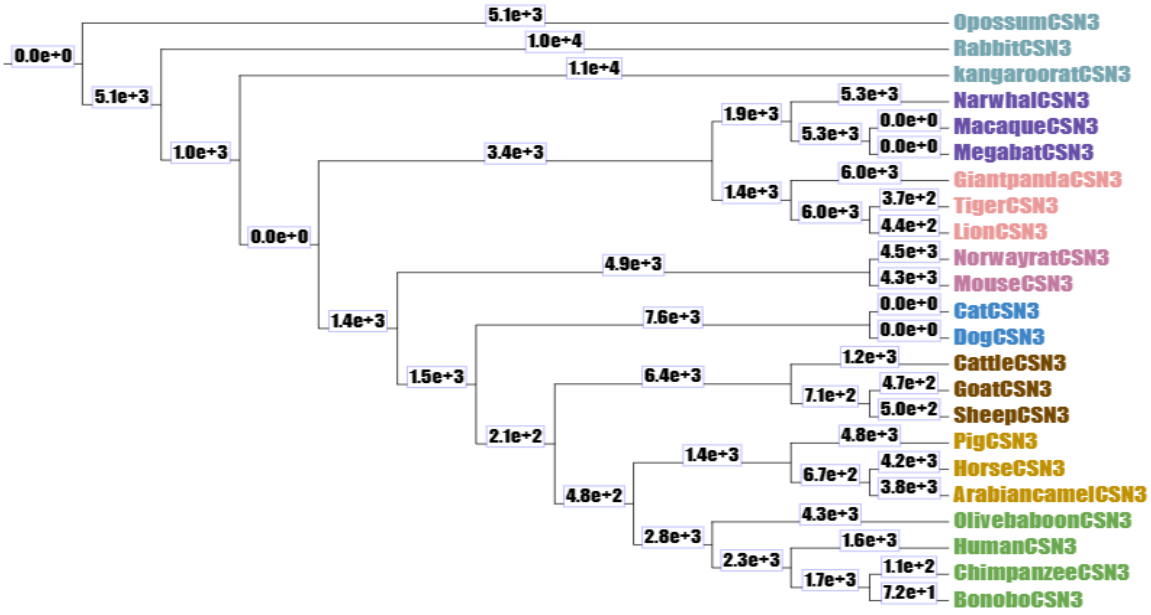


Figure 10. Phylogenetic tree of CSN3 gene 2 kb putative promoter with branch lengths.

4.4.2 Identification of Candidate Motifs and Associated TFs of CSN3

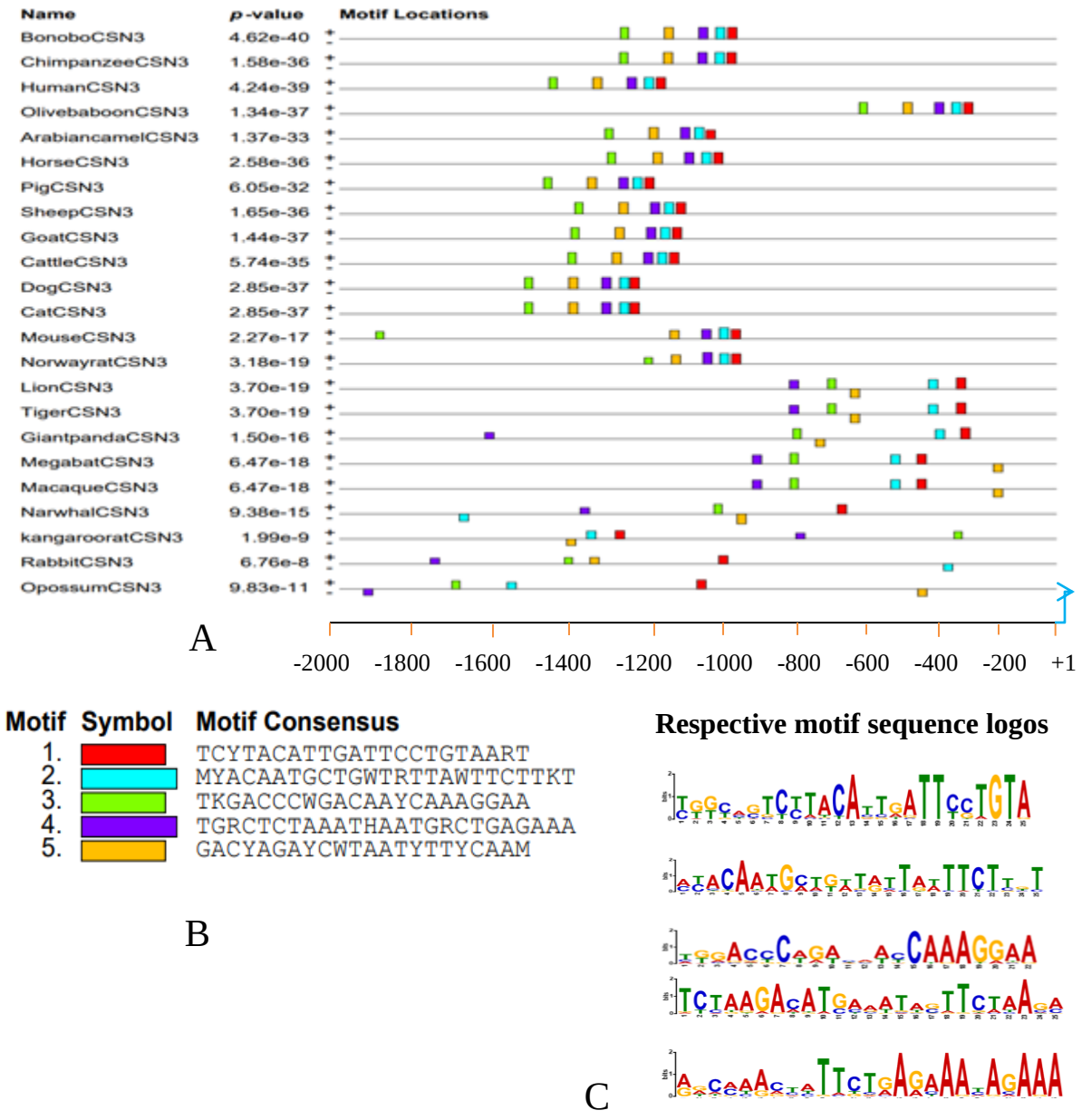


Figure 11. CSN3 predicted putative promoter profile.
A) Promoter motif distribution relative to TSSs. B) Promoter motif symbols and consensus sequence, and (C) predicted promoter motif sequence logos (The height of a letter shows its relative frequency at the given position in the motif).

A comprehensive analysis of 2 kb putative promoter sequences of CSN3 from 23 distinct species was conducted using the MEME-Suite algorithm. To capture various motifs, the search parameters were set to a minimum motif width of 6 and a maximum motif width of 25 nucleotides. To ensure statistical significance, only motifs with an E-value of less than 0.05 were

considered. This rigorous approach led to the identification of the top 5 most significant motifs within the CSN3 putative promoter regions.

A noticeable similarity is observed in the relative distance among and order of a cluster of five motifs in five clades from Norway rat to bonobo. The exception is motif 3 in mice (whose p-value of 2.27e-17 is much higher than other species in this group), which is located further upstream at about -1840bp. Also, the distance of the cluster of the five motifs from their respective canonical TSSs ranges from -1570 to -960bp except for the cluster of motifs in Olive baboon, which is situated at -620 to -210bp from its TSS.

For most taxa considered, there is a distinct pattern regularity of the five identified motifs. Based on the observed relative distance from the canonical TSS, two distinct groups are present. The first grouping includes those listed from bonobo to Norway rat, except motifs discovered in olive baboon where cluster of the motifs, are situated -1570 to -1000bp upstream of the canonical TSS; while the second grouping, includes those listed from lion to macaque which appears to display two sets of the cluster with two motifs each. One cluster centered around -400bp contains the motifs 1 and 2; while the second cluster is centered near -800bp and contains motifs 3 and 4. Motif 5 does not appear strictly limited to a specific location or relative order *vis-a-vis* the other motifs and in all cases is present in the negative strand of the DNA. The remaining samples which comprise narwhal, kangaroo rat, rabbit, and opossum do not exhibit properties that are similar to the other two groups, nor among each other (Figure 9A). It is also worthwhile to note that except for motifs 2 and 4, the other 3 motifs consensus are 22 residues long (Figure 11B).

The conserved candidate motifs of CSN3 were subjected to the Tomtom algorithm to identify the top 5 conserved TFs associated with each motif with a p-value of less than 1.0e-02, resulting in 21 statistically significant TFs (Table 6). The STAT5a: STAT5b, located in motif 5, had the highest significance level with a p-value $9.89e^{-5.}$

Table 6. CSN3 TFs match the query motif in the JASPAR2022 CORE vertebrates.

Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
ZNF24	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1124.1	P17028	$3.28e^{-03}$
ETV2::DRGX	Tryptophan cluster factors:: Homeo domain factors	Ets-related::paired related HD factors	MA1940.1	A6NNA5::Q3KNT2	$6.85e^{-03}$
ETV5::DRGX	Tryptophan cluster factors:: Homeo domain factors	Ets-related::paired related HD factors	MA1944.1	A6NNA5::P41161	$7.24e^{-03}$
ERF:HOXB13	Tryptophan cluster factors:: Homeo domain factors	Ets-related:: HOX related factors	MA1937.1	P50548::Q92826	$7.36e^{-03}$
PBX1	Homeo domain factors	TALE-type homeo domain factors	MA0070.1	P40424	$7.38e^{-03}$
Matched TFs with motif 2	Class	Family	Matrix ID	UniProt ID	p-value
Sox1	High mobility group (HMG) domain factors	Sox-related factors	MA0870.1	P53783	$1.48e^{-03}$
CDX2	Homeo domain factors	HOX	MA0465.2	Q99626	$2.15e^{-03}$
Foxq1	Fork head/winged helix factors	FOX	MA0040.1	Q63244	$2.24e^{-03}$
ONECUT3	Homeo domain factors	HD-CUT	MA0757.1	O60422	$3.60e^{-03}$
ZNF85	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1720.1	Q03923	$4.05e^{-03}$
Matched TFs with motif 3	Class	Family	Matrix ID	UniProt ID	
TCF7L2	High mobility group (HMG) domain factors	TCF-7-related factors	MA0523.1	Q9NQP0	$4.83e^{-04}$
Lef1	High mobility group (HMG) domain factors	TCF-7-related factors	MA0768.2	P27782	$1.30e^{-03}$
Sox2	High mobility group (HMG) domain factors	Sox-related factors	MA0143.4	P48431	$1.31e^{-03}$
Bcl11B	C_2H_2 zinc finger factors	Factors with multiple dispersed zinc finger	MA1989.1	Q99PV8	$1.38e^{-03}$
Sox3	High mobility group (HMG) domain factors	Sox-related factors	MA0514.2	P53784	$3.70e^{-03}$
Matched TFs with motif 4	Class	Family	Matrix ID	UniProt ID	P-value
RUNX3	Runt domain factors	Runt –related factors	MA0684.2	Q13761	$7.93e^{-03}$
Matched TFs with motif 5	Class	Family	Matrix ID	UniProt ID	P-value
STAT5a::STAT5b	STAT domain factors	STAT factors	MA0519.1	P42230::P42232	$9.89e^{-05}$
STAT2	STAT domain factors	STAT factors	MA1623.1	Q9WVL2	$1.08e^{-04}$
STAT4	STAT domain factors	STAT factors	MA0518.1	P42228	$1.62e^{-03}$
STAT3	STAT domain factors	STAT factors	MA0144.2	P40763	$2.87e^{-03}$
STAT1	STAT domain factors	STAT factors	MA0137.3	P42224	$3.55e^{-03}$

4.5 LALBA Putative Promoter

4.5.1 Phylogenetics of LALBA Promoter Region

Phylogenetic analysis of the LALBA putative promoter sequences from 23 distinct species was performed using the Clustal-Omega program within the Jalview software environment. The analysis revealed four separate clades based on their respective evolutionary distance values. The species in these clades include bonobo, chimpanzee, human, macaque, olive baboon, Norway mouse, and dog in the first clade; tiger, lion, cat, giant panda, megabat, Arabian camel, and narwhal in the second clade; and sheep, goat, and cattle in the third clade. Pigs, kangaroos, rabbits, horses, and opossums were placed as an out-group in the tree (Figure 12).

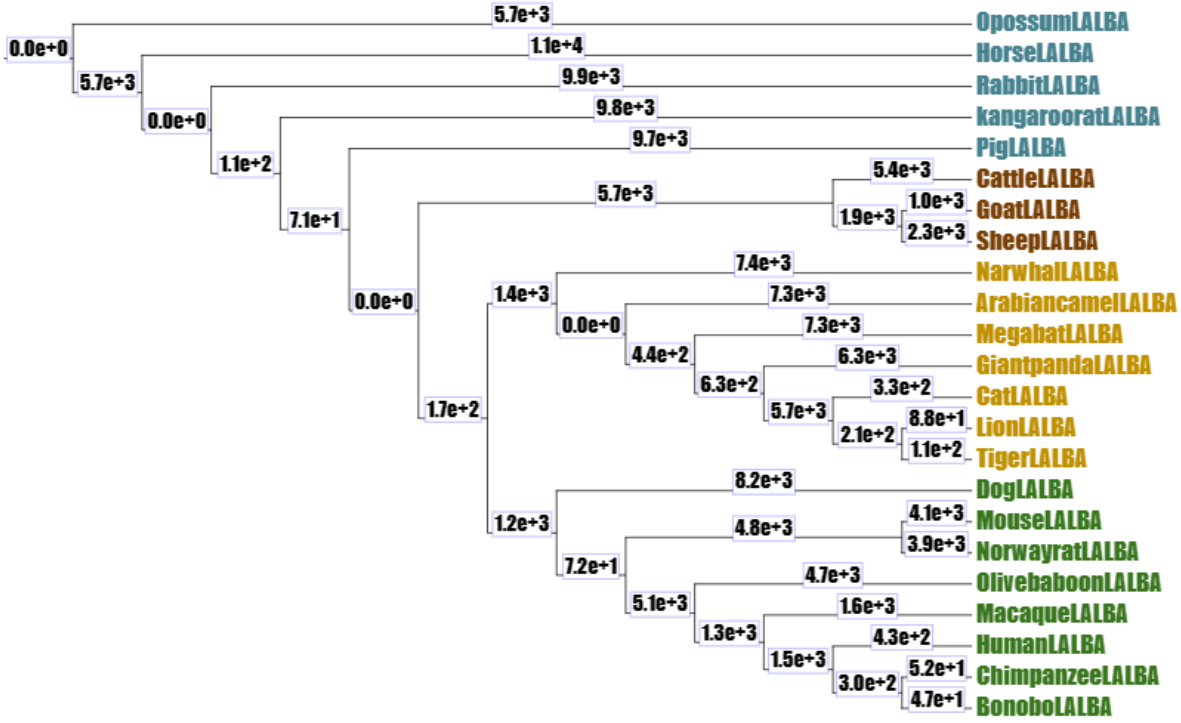
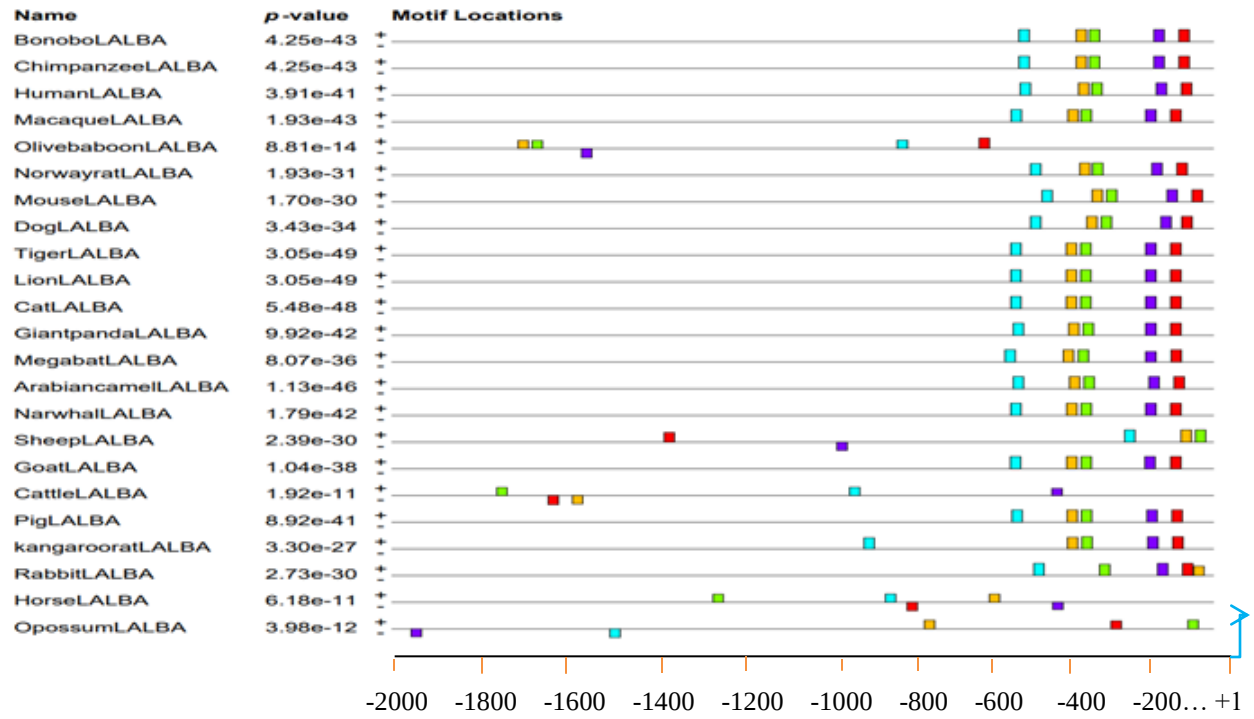


Figure 12. Phylogenetic tree of LALBA gene 2 kb putative promoter with branch lengths.

The ruminant species were placed in the same clade while, the felines were closely placed in clade 2. Similarly the five primates were all placed in clade 1, but along with the two rodents and dog. Interestingly, rabbit, horse and pig were placed in the same clade as the out-group opossum indicating possible divergent and slower evolution in these putative promoter regions as compared to their species-level relatives.

4.4.2 Identification of Candidate Motifs and Associated TFs of LALBA

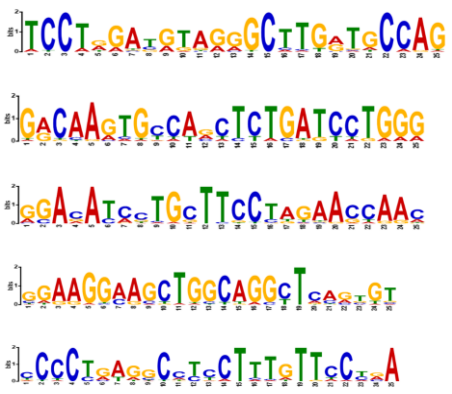


A

Motif	Symbol	Motif Consensus
1.		TCCTRGATGTAGGGCTTGRTGCCAG
2.		GRCAAGTGCCARCTCTGATCCTGGG
3.		GGACATCCTGCTTCCTAGAACCAAC
4.		GGAAGGAAGCTGGCAGGCTCAGTGT
5.		CCCCTGAGGCCTCCTTTGTTCCYRA

B

Respective motif sequence logos



C

Figure 13. LALBA predicated putative promoter profile.
A) Promoter motif distribution relative to TSSs. B) Promoter motif symbols and consensus sequence, and (C) predicted promoter motif sequence logos (The height of a letter shows its relative frequency at the given position in the motif).

The MEME Suite algorithm identified the candidate conserved motifs within the putative promoter sequences of LALBA across 23 diverse species. The motif width parameters were set

between 6 and 25 nucleotides, and an E-value threshold of less than 0.05 was applied to ensure statistical significance. The motifs' relative order and distance (Figure 13 A) reveal a similar pattern from bonobo to rabbit (except for motifs found in olive baboons, sheep, and cattle) all located within -580bp from the canonical TSS. The incongruity of the motif profile of the Olive, baboon (p-value of $8.81e^{-14}$) despite the sequence of the putative promoter showing strong relatedness to the other primate species, is surprising. Similarity is also observed in the relative order and distance of motifs from the canonical TSS in kangaroo rats, which, in previous putative promoters considered typically, behaves as an outlier. However, motifs in horse and opossum species do not exhibit the same strict ordering and relative distance from the canonical TSS. In sheep, motifs 2, 3 and 5 are observed exhibiting similar patterning are much closer to the TSS, and their predicted motifs 1 and 4 are much further upstream and in the case of motif 4, it is position in the negative strand. This could be due to incomplete data acquisition and the missing motifs may reside 5' of the canonical TSS.

The top five conserved motifs were subjected to the TomTom algorithm to identify the five most statistically significant TFs based on p-values of less than $1.0e^{-02}$. The process resulted in the identification of 25 TFs that are highly likely to play a crucial role in LALBA gene regulation (Table 7). Motif 3 is highly composed of the class of STAT TF domain factors, while motif 5 is Sox-related domain factors with p- value of less than $1e^{-4}$ and motif 2 is solely composed of bHLH factors with p- value of $1e^{-4}$.

.

Table 7. LALBA TFs match the query motif in the JASPAR2022 CORE vertebrates.

Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
NFIC::TLX1	SMAD/NF-1 DNA binding::Homeo domain factors	Nuclear factor1::NK	MA0119.1	P08651	$4.69e^{-04}$
SPDEF	Tryptophan cluster factors	Ets-related	MA0686.1	P31314::O95238	$3.72e^{-03}$
NFIX	SMAD/NF-1 DNA binding domain factors	Nuclear factor1	MA1528.1	Q14938	$4.08e^{-03}$
ZNF449	C_2H_2 zinc finger factors	More than 3 adjacent zinc finger factors	MA1656.1	Q6P9G9	$4.75e^{-03}$
Plag11	C_2H_2 zinc finger factors	More than 3 adjacent zinc finger factors	MA1615.1	O35745	$5.63e^{-03}$
Matched TFs with motif 2	Class	Family	Matrix ID	UniProt ID	P-value
HEY2	Basic helix-loop helix factors (bHLH)	Hairy-related factors	MA0649.1	Q9UBP5	$9.50e^{-05}$
HEY1	Basic helix-loop helix factors (bHLH)	Hairy-related factors	MA0823.1	Q9Y5J3	$1.05e^{-04}$
HES1	Basic helix-loop helix factors (bHLH)	Hairy-related factors	MA0616.2	Q9Y543	$1.41e^{-04}$
HES1	Basic helix-loop helix factors (bHLH)	Hairy-related factors	MA0821.2	Q5TA89	$2.49e^{-04}$
HES1	Basic helix-loop helix factors (bHLH)	Hairy-related factors	MA1493.1	Q96HZ4	$6.45e^{-04}$
Matched TFs with motif 3	Class	Family	Matrix ID		P-value
Rfx6	Fork head/winged helix factors	RFX-related factors	MA1724.1	Q8C7R7	$8.06e^{-04}$
STAT4	STAT domain factors	STAT factors	MA0518.1	P42228	$9.59e^{-04}$
STAT1	STAT domain factors	STAT factors	MA0137.3	P42224	$1.19e^{-03}$
STAT5a:: STAT5b	STAT domain factors	STAT factors	MA0519.1	P42230::P42232	$1.32e^{-03}$
STAT3	STAT domain factors	STAT factors	MA0144.2	P40763	$1.63e^{-03}$
Matched TFs with motif 4	Class	Family	Matrix ID	UniProt ID	P-value
ZNF530	C_2H_2 zinc finger factors	More than 3 adjacent zinc finger factors	MA1981.1	Q6P9A1	$3.65e^{-04}$
EWSR1-FL11	Tryptophan cluster factors	Ets-related	MA0149.1	F1JVV7::F1JVV8	$6.15e^{-04}$
ASCL1	Basic helix-loop helix factors (bHLH)	MyoD/ASC-related factors	MA1100.2	P50553	$1.21e^{-03}$
Neurod2	Basic helix-loop helix factors (bHLH)	Tal-related	MA1993.1	Q62414	$2.74e^{-03}$
Neurod1	Basic helix-loop helix factors (bHLH)	Tal-related	MA1109.1	Q13562	$3.20e^{-03}$
Matched TFs with motif 4	Class	Family	Matrix ID	UniProt ID	P-value
Sox11	High mobility group (HMG) domain factors	Sox-related factors	MA0869.2	Q7M6Y2	$6.27e^{-06}$
Sox4	High mobility group (HMG) domain factors	Sox-related factors	MA0867.2	Q06945	$8.64e^{-06}$
Sox10	High mobility group (HMG) domain factors	Sox-related factors	MA0442.2	P56693	$1.35e^{-04}$
Sox9	High mobility group (HMG) domain factors	Sox-related factors	MA0077.1	P48436	$5.92e^{-04}$
Sox6	High mobility group (HMG) domain factors	Sox-related factors	MA0515.1	P40645	$6.59e^{-04}$

4.6 BLG Putative Promoter

4.6.1 Phylogenetics of BLG Promoter Region

Figure 14 illustrates the alignment of BLG promoter sequences using the Clustal-Omega program within the Jalview software package for phylogenetic analysis. Only 6 mammalian species were considered, the complete upstream sequences of BLG of the other species were not available and some species under consideration do not express BLG. The neighbour-joining tree shows that goats, sheep, and cattle are closely related and cluster together in a clade. In contrast, pigs, horses, and cats are positioned as the out-group. The genetic distances represented in the figure provide a quantitative measure of the genetic differentiation and evolutionary divergence among the selected mammalian species

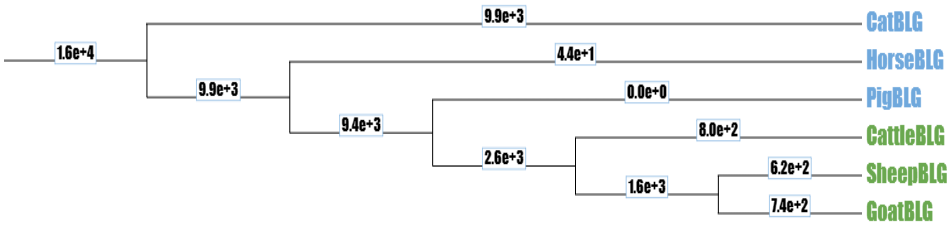
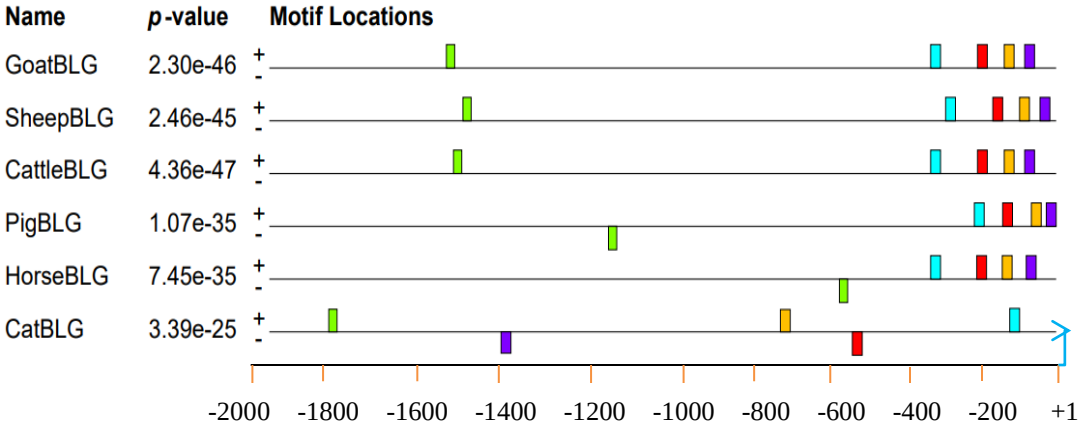
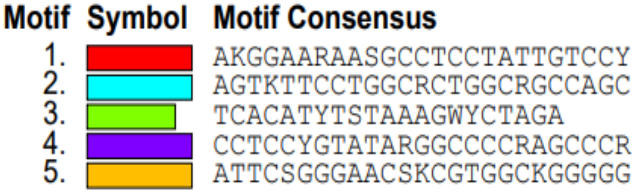


Figure 14. Phylogenetic tree of BLG gene 2 kb putative promoter with branch lengths.

4.6.2 Identification of Candidate Motifs and Associated TFs of BLG

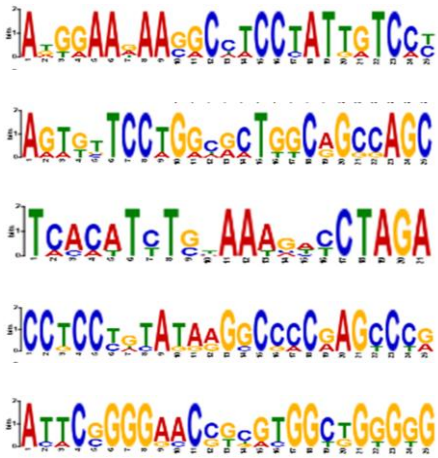


A



B

Respective motif sequence logos



C

Figure 15. BLG predicated putative promoter profile.
A) Promoter motif distribution relative to TSSs. B) Promoter motif symbols and consensus sequence, and (C) predicted promoter motif sequence logos (The height of a letter shows its relative frequency at the given position in the motif).

Figure 15A presents the distribution of the top 5 conserved motifs identified from BLG sequences of six species using the MEME suite algorithm. The minimum and maximum motif widths were restricted to 6 and 25, respectively, and only motifs with e-values of 0.05 were selected for further analysis. The relative order and distance of motifs from goat to pig follow a consistent pattern. Motifs 1, 2, 4 and 5 in goats, sheep, cattle, pigs and horse are typically situated within -490bp from the canonical TSS, but motif 3, which is shorter than the other

motifs (21bp) (Figure 15B), is further upstream with that of pig and horse being on the negative strand. Although similar motifs are found in cats, they do not display the same strict ordering and relative distance from the canonical TSS. Further, 12 and 11 nucleotides in the consensus of motifs 1 and 5 are perfectly conserved, respectively (Figure 15C).

As shown in Table 8, the motifs identified using MEME suite analysis were further subjected to Tomtom software to decipher the top 5 most significant TFs in each motif with p-values less than $1.0e^{-02}$. This analysis revealed 25 TFs potentially involved in the regulation of the BLG gene. The C_2H_2 TFs are found in four motifs, with motif 5 being composed exclusively of this TF class. Four of the five TFs in motif 1 are of the Sox-related factors family.

Table 8. BLG TFs match the query motif in the JASPAR2022 CORE vertebrates.

Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
Zic3	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA0697.2	Q62521	$6.43e^{-04}$
Sox8	High mobility group (HMG) domain factors	Sox-related factors	MA0868.2	P57073	$1.43e^{-03}$
Sox5	High mobility group (HMG) domain factors	Sox-related factors	MA0087.2	P35710	$1.65e^{-03}$
Sox6	High mobility group (HMG) domain factors	Sox-related factors	MA0515.1	P40645	$2.09e^{-03}$
Sox9	High mobility group (HMG) domain factors	Sox-related factors	MA0077.1	P48436	$2.96e^{-03}$
Matched TFs with motif 2	Class	Family	Matrix ID	UniProt ID	P-value
ZNF582	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1983.1	Q96NG8	$1.46e^{-03}$
YY2	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA0748.2	O15391	$2.78e^{-03}$
ZBTB12	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1649.1	Q9Y330	$2.92e^{-03}$
RFX3	Fork head/winged helix factors	RFX-related factors	MA0798.3	P48380	$3.81e^{-03}$
RFX1	Fork head/winged helix factors	RFX-related factors	MA0509.3	P48743	$3.96e^{-03}$
Matched TFs with motif 3	Class	Family	Matrix ID	UniProt ID	P-value
TWIST1	Basic helix-loop-helix factors (bHLH)	Tal-related factors	MA1123.2	Q15672	$1.21e^{-03}$
TBX1	T-box factors	TBX1-related factors	MA0805.1	O43435	$2.57e^{-03}$
TGIF1	Homeo domain factors	TALE-type homeo domain factors	MA0796.1	Q15583	$2.70e^{-03}$
MGA	T-box factors	TBX6-homeo domain factors	MA0801.1	P48380	$2.83e^{-03}$
TBX15	T-box factors	TBX1-related factors	MA0803.1	P48743	$2.83e^{-03}$
Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
MYC	Basic helix-loop-helix factors (bHLH)	bHLH-ZIP	MA0143.3	P01106	$1.06e^{-04}$
KLF15	C_2H_2 zinc finger factors	Three-zinc finger Kruppel-related	MA1513.1	Q9UIH9	$2.47e^{-04}$
ZNF320	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1976.1	A2RRD8	$2.54e^{-04}$
MYCN	Basic helix-loop-helix factors (bHLH)	bHLH-ZIP	MA0104.4	P04198	$1.28e^{-03}$
KLF16	C_2H_2 zinc finger factors	Three-zinc finger Kruppel-related	MA0741.1	Q9BXK1	$1.38e^{-03}$
Matched TFs with motif 1	Class	Family	Matrix ID	UniProt ID	P-value
ZNF610	C_2H_2 zinc finger factors	C_2H_2 zinc finger factors	MA1713.1	Q8N9Z0	$5.88e^{-04}$
RREB1	C_2H_2 zinc finger factors	Factors with multiple dispersed zinc fingers	MA0073.1	Q92766	$5.94e^{-04}$
ZNF189	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1725.1	O75820	$6.06e^{-04}$
KLF9	C_2H_2 zinc finger factors	Three-zinc finger Kruppel-related	MA1107.2	Q13886	$1.06e^{-03}$
ZNF528	C_2H_2 zinc finger factors	More than 3 adjacent zinc fingers	MA1597.1	Q3MIS6	$1.14e^{-03}$

Table 9. Common TFs among all six milk genes and distance from TSSs.

Gene name	Common TFs	Location of TFBSs relative to TSS	Consensus DNA-recognition motif
CSN1S1 and CSN1S2	STAT1 STAT5a STAT5b	-200 to -80bp and -1580bp upstream Exception: *CSN1S1, binding motifs are scattered in cattle, kangaroo rats and opossum*.	TTC[A/T]N[G/A]GAA
CS1N1, CSN1S2, CSN2 and CSN3	STAT1 STAT5a STAT5b	-300 to-80bp and further upstream from -600bp	TTC[A/T]N[G/A]GAA
CS1N1, CSN1S2, CSN2, CSN3 and LALBA	STAT1 STAT5a STAT5b	-390 to -80bp and further upstream from -600bp	TTC[A/T]N[G/A]GAA
CS1N1, CSN2, CSN3 and LALBA	STAT3	-390 to -80bp and further upstream from -600bp	TTC[A/T]N[G/A]GAA
LALBA and BLG	Sox6 Sox9	-400 to -100bp Exception: *LALBA – cattle and olive baboon (distal); opossum and horse (scattered). BLG: cat -500 to -570bp*.	C[A/T]TTG[A/T][A/T] C[A/T]TTG[A/T][A/T]
CS1N1, CSN1S2, CSN2, CSN3, LALBA and BLG	Null	Null	Null

The STAT transcription factors (STAT1, STAT5a, and STAT5b) share TFBSs within the -390 to -80bp range in the promoter regions of five milk genes (CSN1, CSN1S2, CSN2, CSN3, and LALBA), extending up to -600bp upstream from the TSS. However, in CSN1S1 of kangaroo rat and cattle, their shared TFBSs for STAT1, STAT5a and STAT5b are found dispersed along the breadth of promoter regions. These STAT TFs recognize a common consensus (TTC[A/T]N[G/A]GAA sequence. On the other hand, Sox6 and Sox9 transcription factors, common to LALBA and BLG, share TFBSs between -400 and -100bp which are known to bind the consensus C[A/T]TTG[A/T][A/T] sequence. In LALBA, exceptions exist where the shared TFBSs for cattle and olive baboon are in the distal promoter, while in opossum and horse, they are scattered across the promoter region. In cats, the TFBSs for these shared TFs are specifically located between -500 and 570bp (Table 9).

4.7 GO Analysis Results of Candidate Motifs

Table 10. GO analysis of candidate motifs of caseins milk genes

GO terms (GO ID)	Gene	M1*	M2*	M3*	M4*	M5*
Biological process (BP)						
G-protein coupled receptor signaling pathway (GO:0007186)	CSN1S1	✓	**✓**	**✓**	✓	✓
	CSN1S2	**✓**	✓	**✓**	✓	✓
	CSN2		**✓**	**✓**	✓	
	CSN3	✓	✓	✓	✓	**✓**
Signal transduction (GO:0007165)	CSN1S1	✓	**✓**	**✓**	✓	✓
	CSN1S2	✓	✓	✓	✓	
	CSN2		✓	✓	✓	✓
	CSN3	✓	✓	✓	✓	✓
Immune response (GO:0006955)	CSN1S2					✓
Cell differentiation (GO:0030154)	CSN2	✓				
Myeloid leukocyte activation (GO:0002274)	CSN2	✓				
Cell communication (GO:0007154)	CSN2		✓	✓	✓	
Regulation of epithelial cell proliferation (GO:0050678)	CSN2		✓	✓	✓	✓
Molecular function (MF)						
Sequences specific DNA binding (GO:0043565)	CSN2	✓				✓
Calcium ion binding (GO:0005509)	CSN2	✓				
Transcription factor activity (GO:0003700)	CSN2					✓
Cellular component (CC)						
Integral to membrane (GO:0016021)	CSN1S1	✓	**✓**	**✓**	✓	✓
	CSN1S2	**✓**	✓	**✓**	✓	
	CSN2		**✓**	**✓**	✓	
	CSN3	✓	✓	✓	✓	**✓**
Extra cellular region (GO:0005615)	CSN1S2					✓
	CSN2					✓
Proteinaceous extracellular matrix (GO:0005578)	CSN2	✓				

*M1: Motif 1; M2: Motif 2; M3: Motif 3; M4: Motif 4 and M5: Motif 5
Bold checkmarks signify the GO term linked to shared binding motifs of STAT family TFs among the four caseins and LALBA.

After analyzing the significant motifs for each of the casein milk genes using the Meme- Suite algorithm, further GO function enrichment analyses were carried out. The outcomes of the analysis were categorized into three main groups: BP, CC, and MF. Among all caseins, the most commonly observed GO terms for BP were the G-protein coupled receptor signaling pathway (GO:0007186) and Signal transduction (GO:0007165). Additionally, Integral to membrane (GO:0016021) emerged as one of the predominant CC GO terms, while the remaining GO terms were less frequently observed across the caseins. These enrichments were significant at FDR < 0.05 (Table 10).

Table 11. GO analysis result of candidate motifs of whey (LALBA and BLG) milk genes.

Biological process (BP)	Gene	M1*	M2*	M3*	M4*	M5*
G-protein coupled receptor signaling path way (GO:0007186)	LALBA			✓		☑
	BLG	✓		✓		
Positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	BLG				✓	✓
Transcription (GO:0006350)	BLG		✓		✓	✓
Negative regulation of inflammatory response (GO:0050728)	LALBA		✓			
Regulation of interleukin-5 production (GO:0032674)	LALBA			✓		
Regulation of production of small RNA for gene silencing by RNA (GO:0070920)	LALBA					
Mammary gland duct morphogenesis (GO:0060603)	LALBA				✓	
Cell communication (GO:0007154)	LALBA		✓			☑
Regulation of estrogen receptor signaling pathway (GO:0033146)	BLG					✓
Protein amino acid phosphorylation (GO:0006468)	BLG					
Signal transduction (GO:0007165)	BLG			✓		
Molecular function (MF)	Gene	M1	M2	M3	M4	M5
MF serine-type endopeptidase inhibitor activity (GO:0004867)	LALBA	✓				
Serine-type endopeptidase activity (GO:0004252)	LALBA			✓		
Transmembrane receptor protein tyrosine kinase activity (GO:0004714)	LALBA				✓	
Cytokine activity (GO:0005125)	LALBA					☑
Purinergic nucleotide receptor activity, G-protein coupled (GO:004508)	BLG	☑				
Calcium ion binding (GO:0005509)	BLG		✓			
Protein dimerization activity (GO:0046983)	BLG		✓			
Transcription factor activity (GO:0003700)	BLG				✓	✓
ATP binding (GO:0005524)	BLG				✓	✓
Zinc ion binding (GO:0008270)	BLG					
Cellular Component (CC)	Gene	M1	M2	M3	M4	M5
Integral to membrane (GO:0016021)	LALBA		✓	✓		☑
	BLG	☑		✓		
Extracellular region (GO:0005576)	LALBA	✓				
Myosin filament (GO:0032982)	LALBA				✓	
Intracellular membrane-bounded organelle (GO:0043231)	BLG		✓			

*M1: Motif 1; M2: Motif 2; M3: Motif 3; M4: Motif 4 and M5: Motif 5
Bold checkmarks signifies the GO term linked to share binding motifs of STAT family TFs among the four caseins and LALBA, while a checkbox denotes the common binding motifs for the TFs associated with whey proteins.

To gain a deep understanding of LALBA and BLG putative promoter regions, GO analyses were performed using the GOMo software. The annotations of these candidate genes were categorized into BP, CC, and MF. The analysis revealed that G-protein coupled receptor signaling pathway (GO:0007186) emerged as the most prevalent GO term for BP, while integral to membrane (GO:0016021) stood out as the predominant term for CC in both genes. Furthermore, the analysis indicated that other GO terms appeared less frequently in the examination of both genes. These enrichments were notably significant at FDR < 0.05, as detailed in Table 11.

5. DISCUSSION

The present study identified three conserved *cis*-regulatory elements that are universally present in the 2kb putative promoter sequences of four caseins (CSN1S1, CSN1S2, CSN2 and CSN3) and LALBA and one *cis*-regulatory element in a similar promoter region of CSN1S1, CSN2, CSN3, and LALBA genes. Additionally, the study shows the presence of two shared TFBSs between LALBA and BLG. Interestingly, LALBA shares common TFBSs with all caseins and BLG, positioning it as an intermediate between the caseins and the two whey proteins. However, there is no single shared TFBSs among all six milk genes in the examined region. This may be due to genetic drift that resulted in these genes' promoter regions likely undergoing loss of certain promoter variants or the fixation of others over time, potentially resulting in differences in TFBSs among genes (Vaishnav *et al*., 2022; Santpere, 2023).

It is known that, milk protein compositions in different species vary markedly. Significant variations exist in the relative proportion of the various milk proteins in different species; and, in some instances, this variations extend to the presence, or absence, of milk proteins across different species. This variation may be attributable to the presence/absence of TFBSs, their nucleotide composition and the resultant affinity to their TFs. In a study by Runthala *et al*. (2023), it was shown that bovine, camel, goat, and horse milk contain CSN1S1, CSN1S2, CSN2, and CSN3 proteins, but in varying relative ratios of 38:10:40:12, 22:9:65.5:3.5, 20:16:41:17 and 17.7:1.5:79:1.8, respectively. Human milk lacks CSN1S2, and the relative ratio of CSN1S1, CSN2, and CSN3-caseins in human milk is 3:70:27, respectively. On the other hand, BLG is typically expressed in ruminant species and is present in many, but not all, mammalian species, with its notable absence in rodents and human milk (Wodas *et al*., 2020).

In this current study, the phylogenetic trees of the putative promoter sequences of the six milk genes of the considered species namely; ruminants, rodents, carnivores, and primates indicates close evolutionary relatedness. As is to be expected, the opossum, a metatherian, emerges as an outlier species in the CSN1S1, CSN2, CSN3 and LALBA genes (Figures 4, 8, 10, and 12), aligning with findings from previous studies (Malewski, 1998; Murphy *et al*., 2001; Sims *et al*., 2009; Madende and Osthoff, 2019; Feng *et al*., 2021; Hassanin *et al*., 2022; Parveen *et al*., 2023).Within primates, bonobos exhibit variations from humans and chimpanzees in the CSN2 and CSN1S1 putative promoter regions, potentially attributable to mutation accumulation.

Furthermore, in the phylogenetic tree of CSN1S2 and BLG genes, there is a close evolutionary relationship observed among ruminant species (Figures 6 and 14).

In the current study, a comparative analysis of the 5' upstream flanking region among 23 species shows a high level of motif conservation among ruminants, rodents, primates, and carnivorous species, but significant variations were observed in other species (Figures 5A, 7A, 9A, 11A, 13A, and 15A). The study revealed that the putative promoters of the CSN1S1, CSN1S2, CSN2, CSN3, LALBA and BLG genes contain binding sites for a total of 134 TFs potentially involved in the regulation of milk gene expression (Tables 3-8). Among these, three TFs have been identified to have overlapping binding sites for CSN1S1, CSN1S2, CSN2, CSN3 and LALBA, while two TFs share binding sites for LALBA and BLG. The presence of these shared binding sites in the promoter regions of these genes suggests that they are under the influence of a similar gene expression regulatory mechanism in the MG, particularly during late pregnancy and lactation (Rosen *et al*., 1999; Kabotyanski *et al*., 2006; Buser *et al*., 2011).

The large concentration of TF binding motifs identified in the present study are situated in the proximal and core promoter regions, specifically between -400 to -80bp upstream of the canonical TSSs in concordance with prior research (Xu *et al*., 2018; Shijun *et al*., 2020; Song *et al*., 2022), indicating the critical role of the core or proximal promoter region in regulating transcriptional activity in ruminant species. The proximity of putative TF binding motifs to the transcription initiation site increases the likelihood that these factors are involved in the regulation of gene expression (Malewski, 1998; Patel *et al*., 2014; Song *et al*., 2022). Additionally, the findings from this study indicate that a few common TFBSs are also located -600bp further upstream from the TSSs in the distal promoter region as summarized in Table 9, consistent with the findings of Gerencsér *et al*. (2002), who reported that STAT5 binding sites are highly conserved across species in the distal part of the CSN3 promoter sequence centered around -800bp upstream of the TSS. To the best of my knowledge no one has previously documented, the conservation of transcription factor binding motifs in primates, marsupials, and carnivores within the -400 to -100bp regions and also extending upstream of -600 from the canonical TSS in all the analyzed milk genes. The existence of common motifs among the milk genes currently being studied potentially suggests that these genes may have similar regulatory patterns (Zemke *et al*., 2023)

In this study, a comparative analysis of TFBS organization in the putative promoters of CSN1S1, CSN1S2, CSN2, CSN3 LALBA and BLG genes reveals notable similarities in their promoter structures. The key *trans*-elements identified that may bind in the putative promoters' region situated between -390bp and -80bp, excluding BLG, are STAT1, STAT5a, and STAT5b. Additionally, STAT3 is found to be specifically within the promoter regions of CSN1S1, CSN2, CSN3, and LALBA genes. The presence of STAT family TFs in these genes' core and proximal promoters is consistent with previous findings indicating the presence of STAT5 binding sites in the core or proximal promoter region of CSN2 across humans, cows, and rodents, extending ~250bp upstream from the TSS (Rijnkels *et al*., 2003; Qian and Zhao, 2014b). In certain species, binding sites for these *cis*-elements were also observed further upstream of positions -600bp from the TSS, except for CSN1S1 in kangaroo rat and cattle, where their TFBSs are scattered along the breadth of the promoter regions (Figure 5A, 7A, 9A, 11A, 13A, 15A).

Furthermore, Qian and Zhao, (2014b) demonstrated that TFBSs for STAT5 are conserved in the core or proximal promoters of the CSN1S1 gene across cows, sheep, goats, camels, and humans. Thus, except BLG, all five milk genes (CSN1S1, CSN1S2, CSN2, CSN3 and LALBA) exhibit the highest motif pattern similarity in both the proximal and core promoter regions, with some similarities also present in the distal region across the species under consideration. Therefore, the presence of STAT1, STAT3 STAT5a and STAT5b binding sites in both the core promoter region and the proximal region leads us to speculate that these STAT TF families could significantly impact the transcriptional regulation of these milk genes.

In the current study, these elements are recognized for binding to the same small palindromic consensus sequence TTCNNNGAA (Figures 5C, 7C, 9C, 11C, 13C, 15C), which defines a gamma interferon activation site (GAS) and facilitates the binding of dimerized STAT family TFs (Decker *et al*., 1997). The current findings are consistent with previous research indicating that STAT1, STAT5a, and STAT5b are capable of binding to the consensus motif TTCNNNGAA (Decker *et al*., 1997; Liongue and Ward, 2013; Zhou and Chen, 2021). Previous evidence shows that the structure of STAT1, STAT3, STAT5a, and STAT5b TFs are highly conserved across different species, typically consisting of around 750-900 amino acids and containing six domains: the N-terminal domain (NTD), coiled-coil domain (CCD), DNA-binding domain (DBD), linker domain (LD), Src homology 2 (SH2) domains, and transcription

activation domain (TAD) (Zhu *et al*., 2023). Among these domains, the DBD (AA 332-583) is what enables STAT5a/b dimer to bind to the consensus GAS (TTCNNNGAA) motifs found in the regulatory elements of target genes (Liao *et al*., 2010).

The high-level conservation of STAT5a and STAT5b TFBSs among the gene promoters of CSN1S1, CSN1S2, CSN2, CSN3 and LALABA in the current study suggests that the milk gene promoter of one mammalian species can effect the regulated transcription of a transgene in MECs of another species. These are comparable with the previous studies, goat and cow CSN2, cow CSN1S1, human LALBA, and sheep BLG gene promoters enabling the synthesis of a desired protein at a significantly high level in milk (up to tens of grams per liter of milk) (Huang *et al*., 2007; Baldassarre *et al*., 2008; Wang *et al*., 2008; Amiri *et al*., 2013; Li *et al*., 2013; Shepelev *et al*., 2018).

STAT5a and STAT5b are the two isoforms of STAT5 which are essential TFs involved in PRL signaling (Wyszomierski and Rosen, 2001). Moreover, these two STAT5 isoforms exhibit homology (96% protein similarity) with differences primarily found in their C-terminal end (Rani and Murphy, 2016). STAT5, a member of the STAT transcription factor family, is well-known for its vital role in regulating milk gene expression in MECs during lactation (Watson and Neoh, 2008). It is prominently activated during late pregnancy and lactation, with notable levels of STAT5 observed in the nucleus of MECs while being undetectable during the involution phase (Bednorz *et al*., 2011). Recent studies have further highlighted that STAT5 plays a pivotal role in the transcription of milk protein genes and acts as a crucial regulator of MG development (Shin *et al*., 2019; Iskandar *et al*., 2021; Song *et al*., 2022; Guo *et al*., 2023).

STAT5 is a latent cytoplasmic TF that can be activated in response to various cytokines, hormones, and growth factors, leading to its translocation into the nucleus (Wingelhofer *et al*., 2018; Kim *et al*., 2022). Its activation can be triggered by prolactin (PRL) and growth hormone (GH) and act as through the JAK2/STAT signaling pathway or by the epidermal growth factor (EGF) through the Src-kinase/STAT signaling pathway (Figure 17). This activation process involves phosphorylation and dimerization, enabling STAT5 to translocate into the nucleus, bind to DNA as dimers, and stimulate the transcription of caseins and whey genes by binding to clustered STAT5 binding sites in promoters (Xie *et al*., 2002; Tian *et al*., 2020). Recent studies

have highlighted the crucial role of STAT5 activation in regulating the transcription of milk protein genes (Shin *et al*., 2019). Among the two STAT5 protein isoforms, STAT5a serves as the primary and essential mediator of MEC differentiation and milk protein gene expression, while STAT5b plays a more significant role in growth hormone (GH) signaling in the liver (Qian & Zhao, 2014b).

STAT1 is primarily involved in MG development (Watson, 2001; Khan *et al*., 2020), but there are some indications that it may also play some role in milk production traits, such as milk fat (Cobanoglu *et al*., 2006; Cobanoglu *et al*., 2016). Furthermore, it has been proven that the hormone prolactin regulates STAT1 expression. When prolactin binds to its receptor, a cascade of events occurs, resulting in the activation of STAT1, which is subsequently regulated during MG growth and development (Bole-Feysot *et al*., 1998; Cobanoglu *et al*., 2006). In addition to STAT5a/b and STAT1, studies have shown that the prolactin receptor (PRLR) also activates STAT3 (Bachelot and Binart, 2007). In this current investigation it have been demonstrates that STAT3 binds to the consensus DNA sequence TTCNNNGAA (Figures 5C, 9C, 11C, and 13C), as previously reported by Decker *et al*. (1997). Furthermore, Hughes and Watson *et al*. (2018) discovered that the STAT3 TF acts as a mediator of MG post-lactation regression (involution), which is normally induced by the phosphorylation of a particular tyrosine residue.

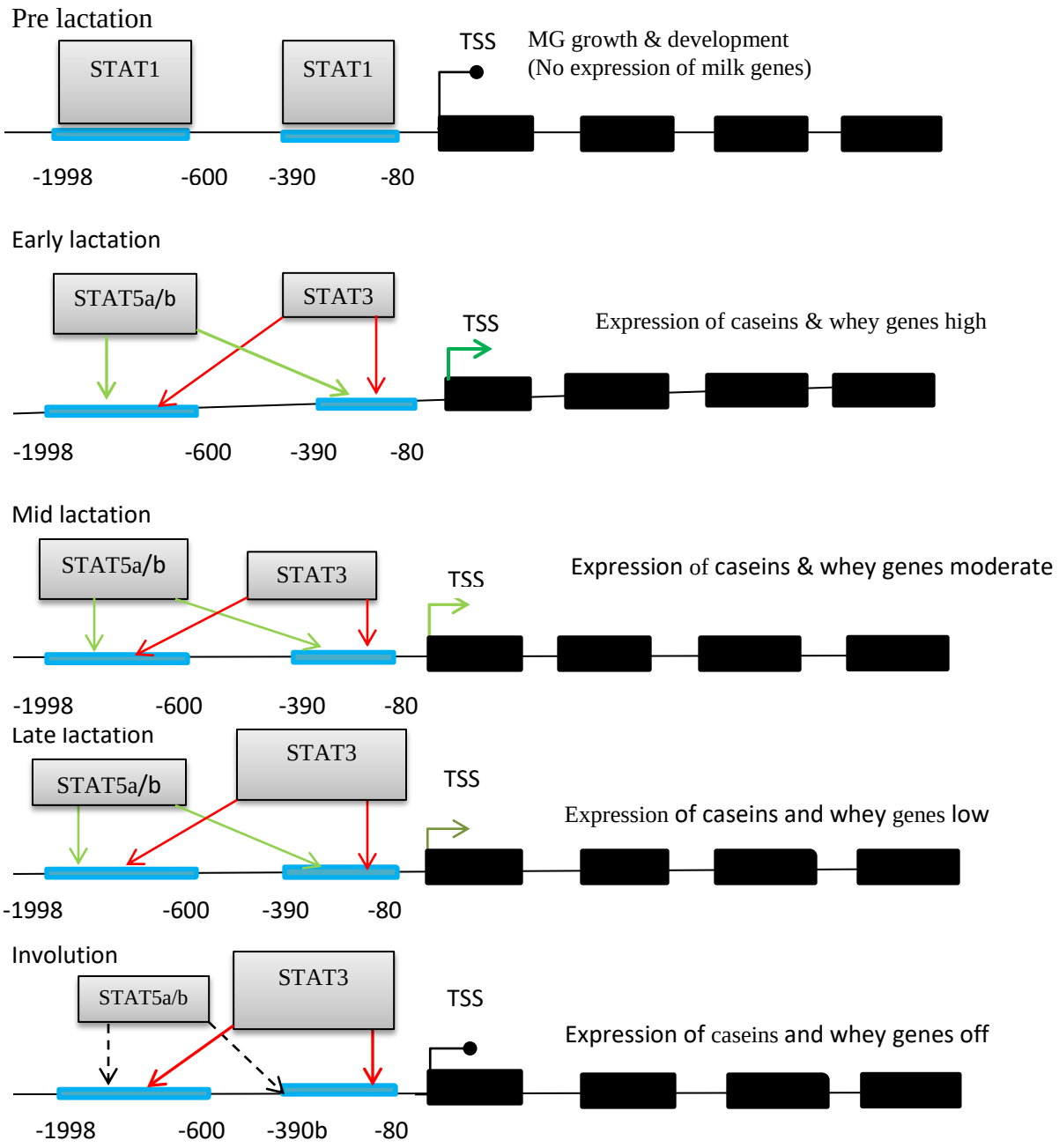


Figure 16. Model of the interplay of STAT1, STAT3 and STAT5 in regulation of milk genes.

The black boxes represent exons, interspersed with the thin lines representing introns, while the light blue thick lines represent the STAT TFs binding region.

In pre-lactation, STAT1 is predominant, while early lactation sees elevated STAT5a/b levels and lower STAT3 expression. Mid-lactation maintains significant STAT5a/b expression for milk production, potentially decreasing slightly from early lactation. Late lactation may witness a decline in STAT5a/b expression due to reduced milk demand, accompanied by increased STAT3 expression, which regulates mammary gland involution and lactation cessation.

(Sources for levels of STATs' expression are: Philp *et al*., 1996; Watson, 2001; Watson and Neoh, 2008; Walker *et al*., 2013)

Figure 16 illustrates a proposed model for the regulation of the expression of milk genes commensurate with the stages of lactation. The model proposes that expression level of caseins and LALBA as being dependent on the relative abundance of the different species of the STAT factors. Since all the STAT factors share the same TFBS GAS consensus sequence, it is likely that these factors compete for the same binding site in the milk genes' promoter regions. Consequently, it is speculated that STAT5 and STAT3 competitively bind to the same TFBSs to regulate milk gene expression with STAT5 driving expression and STAT3 depressing its expression. The oscillation of the fold-change in STAT5 and STAT3 implies that these STAT factors play antagonistic roles in the level of milk gene expression (Philip *et al*., 1996).

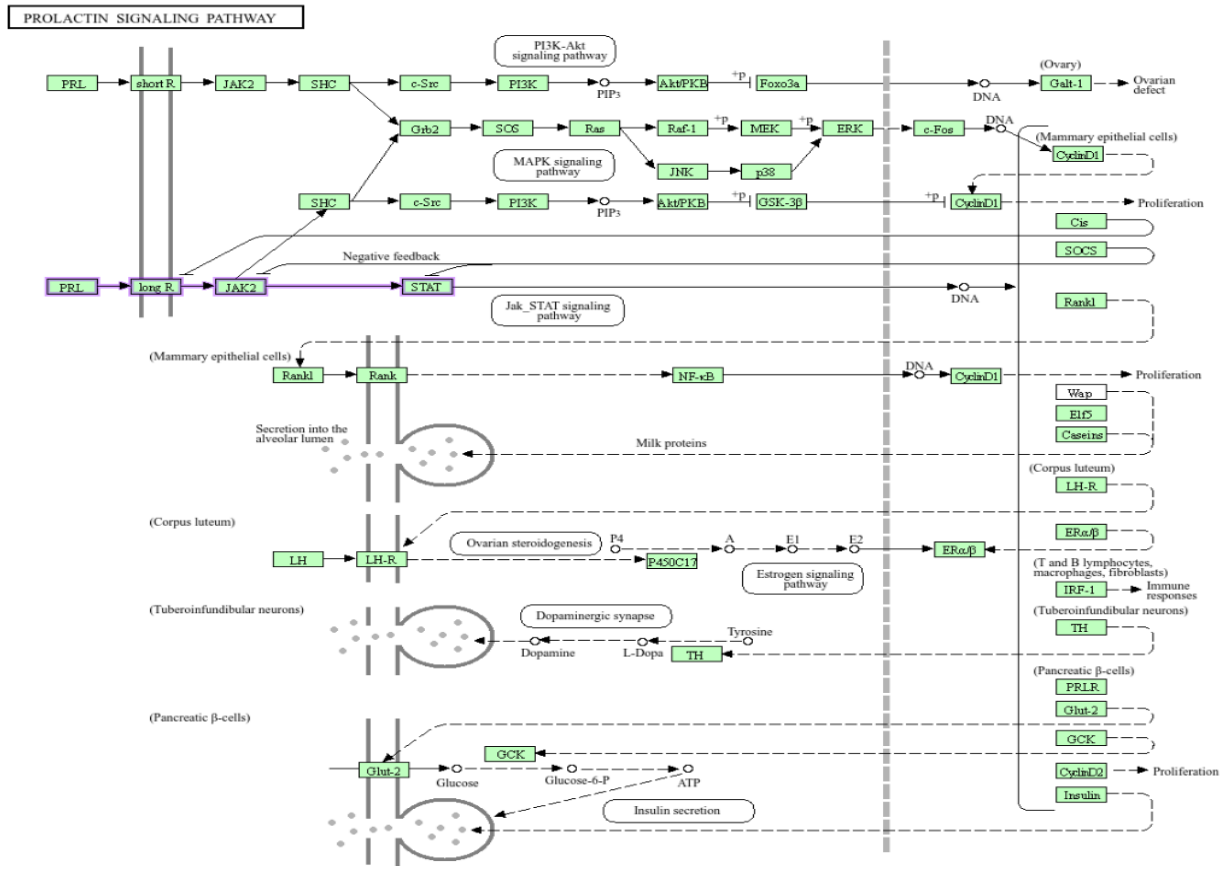


Figure 17. The prolactin signaling pathway of the universally present STAT family TFs. (Source: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis.)

Furthermore, in the current study, two TFBSs for Sox6 and Sox9, have been identified in the putative promoter region shared between two whey milk proteins (LALBA and BLG), as indicated in Tables 7 and 8. These shared TFBSs are located within the -400 to -100bp range

upstream of the TSS. In LALBA, variations are noted, with these TFBSs for opossums and horses scattered along the breadth of the promoter region, while cattle and olive baboon's TFBSs are localized in the distal promoter region. Variations are also observed in BLG of cats, where the TFBSs for these common TFs are specifically located between -500 and -570bp. Notably, these Sox TFs are capable of binding to the same potential consensus sequence (C[A/T]TTG[A/T][A/T]), as reported by previous research Liu *et al*. (2015).

Sox9 is known to regulate mammary gland (MG) development and stem/progenitor cell function, contributing to alveoli formation during pregnancy (Wang *et al*., 2017; Malhotra *et al*., 2014). Similarly, Sox6 expression in luminal cells is linked to maintaining differentiated lineages and alveologenesis (Kendrick *et al*., 2008). However, there is currently no evidence demonstrating a direct role for Sox6 and Sox9 in regulating milk gene expression**.**

In the gene ontology analysis of putative promoter motifs in all six milk genes studied, three categories - BP, MF, and CC were significantly enriched with a false discovery rate (FDR) below 0.05 (Tables 10 and 11). Among the frequently identified GO terms in this study, the G-protein coupled receptor (GPCR) protein signaling pathway (GO:0007186) is consistently represented with the common regulatory motifs found in all milk genes studied.

GPCRs, which are 7-transmembrane proteins, interact with heterotrimeric G proteins on the intracellular side of the membrane (Thal *et al*., 2018). Previous studies have identified three G protein subunits: Gα (which binds to GTP/GDP), Gβ, and Gγ (Wettschureck and Offermanns, 2005). Further studies by Syrovatkina *et al*. (2016) have also shown four parts of Gα (Gαs, Gαi, Gαq/11, and Gα 12/13), which play a role in activating distinct downstream signals of different GPCRs. Activation of GPCRs linked to the Gαq subunit has been shown to increase STAT5 phosphorylation (Tian *et al*., 2020). The oxytocin receptor (OXTR), a GPCR that binds to Gαq, when overexpressed, enhances prolactin-induced STAT5 activation, leading to premature secretory differentiation and premature milk production without pregnancy and in early pregnancy. Conversely, a decline in OXTR-induced STAT5 phosphorylation leads to reduced milk production (Li *et al*., 2018). Another GPCR associated with Gαq is GPR54, which can be activated by kisspeptins (Kps) and is highly expressed during lactation. When activated, GPR54 promotes β-casein synthesis in the MG by activating STAT5 signaling pathways (Sun *et al*.,

2017). Therefore, the identification of the GO terms GPCR in casein and whey milk genes in the present study highlights its key role in the GPCR signaling pathway on milk gene expression.

6. CONCLUSION

The computational study identified and pinpointed the GAS consensus sequence that can serve as a shared TFBSs for STAT1, STAT3, STAT5a, and STAT5b in the putative promoters of CSN1S1, CSN1S2, CSN2, CSN3, and LALBA genes within the -390 to -80bp range, with a few additional shared GAS TFBSs located further upstream of -600bp. An exception in this case is BLG, which is only expressed in ruminants and some, but not all, mammals and the power of the analysis may have been limited due to the small number of publicly available BLG putative promoter sequence data. Furthermore, TFBSs for Sox6 and Sox9 were found to be shared between the whey protein genes LALBA and BLG within the -400 to -100bp range. The totality of these results leads us to conclude that though there are shared substantial commonalities among the *cis*-regulatory elements of milk genes of various species that likely operate through a similar signaling pathway to elicit milk gene expression in lactating mammals, exceptions to these norms exist that point to parallel mechanisms of milk gene expression that have yet to be elucidated. Further, the variations observed in these *cis*-acting regulatory elements may be responsible for differing proportion of caseins and whey proteins in milk.

Noting the oscillation of the various STAT species depending on the different physiological stages of the MG, beyond the immediacy of the STAT TFs regulation of milk genes expression, it is apparent that another mechanism exists that regulates the STATs relative expression and, hence, supersedes the STATs as being the ultimate precise regulators of spatio-temporal expression regulation of milk genes.

7. RECOMMENDATIONS

Future research should prioritize the experimental validation of the shared TFs identified in all milk genes and their corresponding TFBSs to verify their regulatory function in the expression of caseins and whey milk genes. Furthermore, it is suggested to expand the study of the *cis*-regulatory elements of other milk genes including whey acidic protein (WAP) and lactoferrin in search of identifying a master switch for lactogenesis.

8. REFERENCES

Amandykova, M., Dossybayev, K., Mussayeva, A., Bekmanov, B., & Saitou, N. (2022). Comparative analysis of the polymorphism of the casein genes in camels bred in Kazakhstan. *Diversity*, *14*(4), 285.

Amin, R., Rahman, C. R., Ahmed, S., Sifat, M. H. R., Liton, M. N. K., Rahman, M. M., ... & Shatabda, S. (2020). iPromoter-BnCNN: a novel branched CNN-based predictor for identifying and classifying sigma promoters. Bioinformatics,36(19), 4869-4875.

Amiri Yekta, A., Dalman, A., Eftekhari-Yazdi, P., Sanati, M. H., Shahverdi, A. H., Fakheri, R., ... & Gourabi, H. (2013). Production of transgenic goats expressing human coagulation factor IX in the mammary glands after nuclear transfer using transfected fetal fibroblast cells. *Transgenic research*, *22*, 131-142.

Asim, M., Saif-ur-Rehman, M., & Hassan, F. U. (2022). Comparative genomic characterization and screening of regulatory regions of casein gene family in Bos taurus and Bubalus bubalis. *Journal of global innovations agricultural sciences*, *10*, 147-158.

Auestad, N., & Layman, D. K. (2021). Dairy bioactive proteins and peptides: a narrative review. *Nutrition reviews*, *79*(Supplement_2), 36-47.

Bachelot, A., & Binart, N. (2007). Reproductive role of prolactin. *Reproduction*, *133*(2), 361-369.

Bailey, T. L., Johnson, J., & Grant, C. E. (2015). Noble. 2015. The MEME Suite. *Nucleic acids research*, *43*, W39-W49.

Baldassarre, H., Hockley, D. K., Doré, M., Brochu, E., Hakier, B., Zhao, X., & Bordignon, V. (2008). Lactation performance of transgenic goats expressing recombinant human butyryl-cholinesterase in the milk. *Transgenic research*, *17*, 73-84.

Baldassarre, H., Hockley, D. K., Olaniyan, B., Brochu, E., Zhao, X., Mustafa, A., & Bordignon, V. (2008). Milk composition studies in transgenic goats expressing recombinant human butyrylcholinesterase in the mammary gland. *Transgenic research*, *17*, 863-872.

Barrett, L. W., Fletcher, S., & Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and molecular life sciences*, *69*, 3613-3634.

Bednorz, N. L., Brill, B., Klein, A., Gäbel, K., & Groner, B. (2011). Tracking the activation of Stat5 through the expression of an inducible reporter gene in a transgenic mouse line. *Endocrinology*, *152*(5), 1935-1947.

Berry, S. D., Lopez-Villalobos, N., Beattie, E. M., Davis, S. R., Adams, L. F., Thomas, N. L., … & Snell, R. G. (2010). Mapping a quantitative trait locus for the concentration of β-lactoglobulin in milk, and the effect of β-lactoglobulin genetic variants on the composition of milk from Holstein-Friesian x Jersey crossbred cows. *New Zealand veterinary journal*, *58*(1), 1-5.

Bhandari, N., Khare, S., Walambe, R., & Kotecha, K. (2021). Comparison of machine learning and deep learning techniques in promoter prediction across diverse species. *PeerJ computer science*, *7*, e365.

Bhat, M. Y., Dar, T. A., & Singh, L. R. (2016). Casein proteins: structural and functional aspects. *Milk proteins-from structure to biological properties and health aspects*, *10*, 64187.

Bole-Feysot, C., Goffin, V., Edery, M., Binart, N., & Kelly, P. A. (1998). Prolactin (PRL) and its receptor: actions signal transduction pathways and phenotypes observed in PRL receptor knockout mice. *Endocrine reviews*, *19*(3), 225-268.

Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature genetics*, *21*(1), 33-37.

Burke, N., Zacharski, K. A., Southern, M., Hogan, P., Ryan, M. P., & Adley, C. C. (2018). The dairy industry: process, monitoring, standards, and quality. *Descriptive food science*, *162*.

Buser AC, Obr AE, Kabotyanski EB, Grimm SL, Rosen JM, & Edwards DP (2011). Progesterone receptor directly inhibits beta-casein gene transcription in mammary epithelial cells through promoting promoter and enhancer repressive chromatin modifications. *Molecular endocrinology*, 25(6):955–968.

Buske, F. A., Bodén, M., Bauer, D. C., & Bailey, T. L. (2010). Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, *26*(7), 860-866.

Caroli, A. M., Chessa, S., & Erhardt, G. J. (2009). Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *Journal of dairy science*, *92*(11), 5335-5352.

Chatterton, D. E., Smithers, G., Roupas, P., & Brodkorb, A. (2006). Bioactivity of β lactoglobulin and α-lactalbumin—Technological implications for processing. *International dairy journal*, *16*(11), 1229-1240.

Cobanoglu, O., Gurcan, E. K., Cankaya, S., Kul, E., & Abaci, H. S. (2016). The detection of STAT1 gene influencing milk-related traits in Turkish Holstein and Jersey cows. *Journal of agricultural science technology*, *A*, *6*, 261-269.

Cobanoglu, O., Zaitoun, I., Chang, Y. M., Shook, G. E., & Khatib, H. (2006). Effects of the signal transducer and activator of transcription 1 (STAT1) gene on milk production traits in Holstein dairy cattle. *Journal of dairy science*, *89*(11), 4433-4437.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., ... & Flicek, P. (2022). Ensembl 2022. *Nucleic acids research*, *50*(D1), D988-D995.

Dalgleish, D. G., Spagnuolo, P. A., & Goff, H. D. (2004). A possible structure of the casein micelle based on high-resolution field-emission scanning electron microscopy. *International dairy journal*, *14*(12), 1025-1031.

Davoodi, S. H., Shahbazi, R., Esmaeili, S., Sohrabvandi, S., Mortazavian, A., Jazayeri, S., & Taslimi, A. (2016). Health-related aspects of milk proteins. *Iranian journal of pharmaceutical research: IJPR*, *15*(3), 573.

Debeljak MA, Frajman PO, Lenasi TI, Narat MO, Baldi AN, & Dovc PE (2005). Functional analysis of the bovine beta-and kappa casein gene promoters using homologous mammary gland derived cell line. *Archives animal breeding*. 48(4):334-45.

Decker, T., Kovarik, P., & Meinke, A. (1997). GAS elements: a few nucleotides with a major impact on cytokine-induced gene expression. *Journal of interferon & cytokine research*, *17*(3), 121-134.

Doppler, W., Geymayer, S., & Weirich, H. G. (2002). Synergistic and Antagonistic Interactions of Transcription Factors in the Regulation of Milk Protein Gene Expression: Mechanismsof Cross-talk Between Signaling Pathways. *Biology of the mammary gland*, 139-146.

Du, C., Deng, T. X., Zhou, Y., Ghanem, N., & Hua, G. H. (2020). Bioinformatics analysis of candidate genes for milk production traits in water buffalo (Bubalus bubalis). *Tropical animal health and production*, *52*, 63-69.

Feng, T., Wu, S., Luo, X., Lei, A., Luobu, B., Hassan, F. U., & Liu, Q. (2021). Comparative genomics, evolutionary and gene regulatory regions analysis of casein gene family in Bubalus bubalis. *Frontiers in genetics*, *12*, 662609.

Forsyth, I. A., & Neville, M. C. (2009). Introduction: hormonal regulation of mammary development and milk protein gene expression at the whole animal and molecular levels. *Journal of mammary gland biology and neoplasia*, *14*(3), 317-319.

Fox, P. F., Uniacke-Lowe, T., McSweeney, P. L. H., O'Mahony, J. A., Fox, P. F., Uniacke-Lowe, T., ... & O'Mahony, J. A. (2015). Chemistry and biochemistry of cheese. *Dairy chemistry and biochemistry*, 499-546.

Galas, D. J., & Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic acids research*, *5*(9), 3157-3170.

Gerencsér, Á., Barta, E., Boa, S., Kastanis, P., Bösze, Z., & Whitelaw, C. B. A. (2002). Comparative analysis on the structural features of the 5' flanking region of kappa casein genes from six different species. *Genetics selection evolution*, *34*(1), 117-128.

Gigli, I. (Ed.). (2016). *Milk proteins: From structure to biological properties and health aspects*. BoD–Books on Demand.

Goffeau, André, Bart G. Barrell, Howard Bussey, Ronald W. Davis, Bernard Dujon, Heinz Feldmann, Francis Galibert *et al*. (1996). "Life with 6000 genes." *Science,* 274, no. 5287, 546-567.

Gorji, A. E., Roudbari, Z., Sadeghi, B., Javadmanesh, A., & Sadkowski, T. (2019). Transcriptomic analysis on the promoter regions discovers gene networks involving mastitis in cattle. *Microbial pathogenesis*, *137*, 103801.

Guo, H., Li, J., Wang, Y., Cao, X., Lv, X., Yang, Z., & Chen, Z. (2023). Progress in research on key factors regulating lactation initiation in the mammary glands of dairy cows. *Genes*, *14*(6), 1163.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, *8*(2), 1-9.

Haberle, V., & Lenhard, B. (2016). Promoter architectures and developmental gene regulation. *Seminars in cell & developmental biology* (Vol. 57, pp. 11-23). Academic Press.

Han, R., Shi, R., Yu, Z., Ho, H., Du, Q., Sun, X., ... & Yang, Y. (2021). Distribution and variation in proteins of casein micellar fractions response to heat-treatment from five dairy species. *Food chemistry*, *365*, 130640.

Hassanin, A. A., Osman, A., Atallah, O. O., El-Saadony, M. T., Abdelnour, S. A., Taha, H. S., ... & Eldomiaty, A. S. (2022). Phylogenetic comparative analysis: Chemical and biological features of caseins (alpha-S-1, alpha-S-2, beta-and kappa-) in domestic dairy animals. *Frontiers in veterinary science*, *9*, 952319.

Huang, Y. J., Huang, Y., Baldassarre, H., Wang, B., Lazaris, A., Leduc, M., ... & Langermann, S. (2007). Recombinant human butyrylcholinesterase from milk of transgenic animals to protect against organophosphate poisoning. *Proceedings of the national academy of sciences*, *104*(34), 13603-13608.

Hughes, K., & Watson, C. J. (2018). The multifaceted role of STAT3 in mammary gland involution and breast cancer. *International journal of molecular sciences*, *19*(6), 1695.

Iskandar, I., As'ad, S., Mappaware, N., Alasiry, E., Hendarto, H., Hatta, M., ... & Syam, A. (2021). Gene prolactine receptor (PRLR) and signal transducer and activator of transcription 5 (STAT5) on milk production. *Medicina clínica práctica*, *4*, 100223.

Izquierdo-González, J. J., Amil-Ruiz, F., Zazzu, S., Sánchez-Lucas, R., Fuentes-Almagro, C. A., & Rodríguez-Ortega, M. J. (2019). Proteomic analysis of goat milk kefir: Profiling the fermentation-time dependent protein digestion and identification of potential peptides with biological activity. *Food chemistry*, *295*, 456-465.

Javed, A., Zaidi, S. K., Gutierrez, S. E., Lengner, C. J., Harrington, K. S., Hovhannisyan, H., ... & Stein, G. S. (2004). Chromatin immunoprecipitation. *Cell cycle control and dysregulation protocols: Cyclins, cyclin-dependent kinases, and other factors*, 41-44.

Jeong, E. W., Park, G. R., Kim, J., Baek, Y., Go, G. W., & Lee, H. G. (2022). Whey proteins-fortified milk with adjusted casein to whey proteins ratio improved muscle strength and endurance exercise capacity without lean mass accretion in rats. *Foods*, *11*(4), 574.

Jiang, S., Ren, Z., Xie, F., Yan, J., Huang, S., & Zeng, Y. (2012). Bovine prolactin elevates hTF expression directed by a tissue-specific goat β-casein promoter through prolactin receptor-mediated STAT5a activation. *Biotechnology letters*, *34*, 1991-1999.

Kabotyanski, E. B., Huetter, M., Xian, W., Rijnkels, M., & Rosen, J. M. (2006). Integration of prolactin and glucocorticoid signaling at the β-casein promoter and enhancer by ordered recruitment of specific transcription factors and chromatin modifiers. *Molecular endocrinology*, *20*(10), 2355-2368.

Kabotyanski, E. B., Rijnkels, M., Freeman-Zadrowski, C., Buser, A. C., Edwards, D. P., & Rosen, J. M. (2009). Lactogenic hormonal induction of long distance interactions between β-casein gene regulatory elements. *Journal of biological chemistry*, *284*(34), 22815-22824.

Kawasaki, K., & Weiss, K. M. (2008). SCPP gene evolution and the dental mineralization continuum. *Journal of dental research*, *87*(6), 520-531.

Kawasaki, K., Buchanan, A. V., & Weiss, K. M. (2007). Gene duplication and the evolution of vertebrate skeletal mineralization. *Cells tissues organs*, *186*(1), 7-24.

Kawasaki, K., Lafont, A. G., & Sire, J. Y. (2011). The evolution of milk casein genes from tooth genes before the origin of mammals. *Molecular biology and evolution*, *28*(7), 2053-2061.

Kawasaki, K., Lafont, A. G., & Sire, J. Y. (2011). The evolution of milk casein genes from tooth genes before the origin of mammals. *Molecular biology and evolution*, *28*(7), 2053-2061.

Kawasaki, K., Suzuki, T., & Weiss, K. M. (2005). Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. *Proceedings of the national academy of sciences*, *102*(50), 18063-18068.

Kendrick, H., Regan, J. L., Magnay, F. A., Grigoriadis, A., Mitsopoulos, C., Zvelebil, M., & Smalley, M. J. (2008). Transcriptome analysis of mammary epithelial subpopulations identifies novel determinants of lineage commitment and cell fate. *BMC genomics*, *9*(1), 1-28.

Khan, M. Z., Khan, A., Xiao, J., Ma, Y., Ma, J., Gao, J., & Cao, Z. (2020). Role of the JAK-STAT pathway in bovine mastitis and milk production. *Animals*, *10*(11), 2107.

Kim, U., & Shin, H. Y. (2022). Genomic mutations of the STAT5 transcription factor are associated with human cancer and immune diseases. *International journal of molecular sciences*, *23*(19), 11297.

Kolb, A. F. (2002). Structure and regulation of the murine γ-casein gene. *Biochimica et biophysica acta (BBA)-gene structure and expression*, *1579*(2-3), 101-116.

Lai, H. Y., Zhang, Z. Y., Su, Z. D., Su, W., Ding, H., Chen, W., & Lin, H. (2019). iProEP: a computational predictor for predicting promoter. *Molecular therapy-nucleic acids*, *17*, 337-346.

Layman, D. K., Lönnerdal, B., & Fernstrom, J. D. (2018). Applications for α-lactalbumin in human nutrition. *Nutrition reviews*, *76*(6), 444-460.

Le, T. T., Deeth, H. C., & Larsen, L. B. (2017). Proteomics of major bovine milk proteins: Novel insights. *International dairy journal*, *67*, 2-15.

Lenasi, T., Kokalj-Vokac, N., Narat, M., Baldi, A., & Dovc, P. (2005). Functional study of the equine β-casein and κ-casein gene promoters. *Journal of dairy research*, *72*(S1), 34-43.

Li, D., Ji, Y., Zhao, C., Yao, Y., Yang, A., Jin, H., ... & Zheng, Y. (2018). OXTR overexpression leads to abnormal mammary gland development in mice. *Journal of endocrinology*, *239*(2), 121-136.

Li, H., Liu, Q., Cui, K., Liu, J., Ren, Y., & Shi, D. (2013). Expression of biologically active human interferon alpha 2b in the milk of transgenic mice. *Transgenic research*, *22*, 169-178.

Liao, Z., Lutz, J., & Nevalainen, M. T. (2010). Transcription factor STAT5a/b as a therapeutic target protein for proSTATe cancer. *The international journal of biochemistry & cell biology*, *42*(2), 186-192.

Liongue, C., & Ward, A. C. (2013). Evolution of the JAK-STAT pathway. *Jak*-STAT, *2*(1), e22756

Liu, C. F., & Lefebvre, V. (2015). The transcription factors SOX9 and SOX5/SOX6 cooperate genome-wide through super-enhancers to drive chondrogenesis. *Nucleic acids research*, *43*(17), 8183-8203.

Liu, X., Brutlag, D. L., & Liu, J. S. (2000). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *biocomputing, 2001*, (127-138).

Liu, X., Teng, L., Luo, Y., & Xu, Y. (2023). Prediction of prokaryotic and eukaryotic promoters based on information-theoretic features. *Biosystems*, *231*, 104979.

Madende, M., & Osthoff, G. (2019). Comparative genomics of casein genes. *Journal of dairy research*, *86*(3), 323-330.

Madende, M., Osthoff, G., Patterton, H. G., Patterton, H. E., Martin, P., & Opperman, D. J. (2015). Characterization of casein and alpha lactalbumin of African elephant (Loxodonta africana) milk. *Journal of dairy science*, *98*(12), 8308-8318.

Maity, S., Bhat, A. H., Giri, K., & Ambatipudi, K. (2020). BoMiProt: A database of bovine milk proteins. *Journal of proteomics*, *215*, 103648.

Malewski, T. (1998). Computer analysis of distribution of putative *cis*-and *trans*-regulatory elements in milk protein gene promoters. *Biosystems*, *45*(1), 29-44.

Malhotra, G. K. *et al*., (2014). The role of Sox9 in mouse mammary gland development and maintenance of mammary stem and luminal progenitor cells. *BMC developmental biology*. 14, 47.

Mariño-Ramírez, L., Tharakaraman, K., Spouge, J. L., & Landsman, D. (2009). Promoter analysis: Gene regulatory motif identification with A-GLAM. *Bioinformatics for DNA sequence analysis*, 263-276.

Martin, P., Cebo, C., & Miranda, G. (2013). Interspecies comparison of milk proteins: quantitative variability and molecular diversity. *Advanced dairy chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition*, 387-429.

McMeekin, T. L., & Polis, B. D. (1949). Milk proteins. *Advances in protein chemistry*, *5*, 201-228.

Morammazi, S., Masoudi, A. A., Torshizi, R. V., & Pakdel, A. (2016). Differential expression of the alpha S1 casein and beta-lactoglobulin genes in different physiological stages of the Adani goats mammary glands. *Iranian journal of biotechnology*, *14*(4), 278.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., & O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature*, *409*(6820), 614-618.

Najafi, M., Mianji, G. R., & Pirsaraie, Z. A. (2014). Cloning and comparative analysis of gene structure in promoter site of alpha-s1 casein gene in Naeinian goat and sheep. *Meta gene*, *2*, 854-861.

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., ... & Kent, W. J. (2021). The UCSC genome browser database: 2021 update. *Nucleic acids research*, *49*(D1), D1046-D1057.

Olumee-Shabon, Z., Swain, T., Smith, E. A., Tall, E., & Boehmer, J. L. (2013). Proteomic analysis of differentially expressed proteins in caprine milk during experimentally induced endotoxin mastitis. *Journal of dairy science*, *96*(5), 2903-2912.

Ong, C. T., & Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews genetics*, *12*(4), 283-293.

Ong, C. T., & Corces, V. G. (2012). Enhancers: emerging roles in cell fate specification. *EMBO reports 13*(5), 423-430.

Parveen, S., Zhu, P., Shafique, L., Lan, H., Xu, D., Ashraf, S., ... & Liu, Q. (2023). Molecular Characterization and Phylogenetic Analysis of Casein Gene Family in Camelus ferus. *Genes*, *14*(2), 256.

Patel, A. K., Singh, M., & Suryanarayana, V. V. S. (2014). Buffalo alpha S1-casein gene 5′-flanking region and its interspecies comparison. *Journal of applied genetics*, *55*, 75-87.

Pauciullo, A., Giambra, I. J., Iannuzzi, L., & Erhardt, G. (2014). The β-casein in camels: molecular characterization of the CSN2 gene, promoter analysis and genetic variability. *Gene*, *547*(1), 159-168.

Pauciullo, A., Shuiep, E. S., Cosenza, G., Ramunno, L., & Erhardt, G. (2013). Molecular characterization and genetic variability at κ-casein gene (CSN3) in camels. *Gene*, *513*(1), 22-30.

Pauciullo, A., Versace, C., Gaspa, G., Letaief, N., Bedhiaf-Romdhani, S., Fulgione, A., & Cosenza, G. (2023). Sequencing and characterization of αs2-casein gene (CSN1S2) in the Old-World camels have proven genetic variations useful for the understanding of species diversification. *Animals*, 13(17), 2805.

Pauciullo, A., Versace, C., Miretti, S., Giambra, I. J., Gaspa, G., Letaief, N., & Cosenza, G. (2024). Genetic variability among and within domestic Old and New World camels at the α-lactalbumin gene (LALBA) reveals new alleles and polymorphisms responsible for differential expression. *Journal of dairy science*., 107(2), 1068-1084.

Peng, S., Cheng, M., Huang, K., Cui, Y., Zhang, Z., Guo, R., ... & Shi, B. (2018). Efficient computation of motif discovery on intel much-integrated core (mic) architecture. *BMC Bioinformatics*, *19*, 101-110.

Philp, J. A., Burdon, T. G., & Watson, C. J. (1996). Differential activation of STATs 3 and 5 during mammary gland development. *FEBS letters*, *396*(1), 77-80.

Philp, J. A., Burdon, T. G., & Watson, C. J. (1996). Differential activation of STATs 3 and 5 during mammary gland development. *FEBS letters*, *396*(1), 77-80.

Qian, X., & Zhao, F. Q. (2014a). Collaborative interaction of Oct-2 with Oct-1 in trans activation of lactogenic hormones-induced β-casein gene expression in mammary epithelial cells. *General and comparative endocrinology*, *204*, 185-194.

Qian, X., & Zhao, F. Q. (2014b). Current major advances in the regulation of milk protein gene expression. *Critical reviews in eukaryotic gene expression*, *24*(4).

Ramunno, L., Cosenza, G., Rando, A., Illario, R., Gallo, D., Di Berardino, D., & Masina, P. (2004). The goat αs1-casein gene: gene structure and promoter analysis. *Gene*, *334*, 105-111.

Rani, A., & Murphy, J. J. (2016). STAT5 in cancer and immunity. *Journal of interferon & cytokine research*, *36*(4), 226-237.

Rijnkels M, Elnitski L, Miller W, & Rosen JM. Multispecies comparative analysis of a mammalian-specific genomic domain encoding secretory proteins. *Genomics*, 2003; 82(4), 417–32.

Rosen, J. M., Wyszomierski, S. L., & Hadsell, D. (1999). Regulation of milk protein gene expression. *Annual review of nutrition*, *19*(1), 407-436.

Roy, D., Ye, A., Moughan, P. J., & Singh, H. (2020). Composition, structure, and digestive dynamics of milk from different species—A review. *Frontiers in nutrition*, *7*, 577759.

Runthala, A., Mbye, M., Ayyash, M., Xu, Y., & Kamal-Eldin, A. (2023). Caseins: versatility of their micellar organization in relation to the functional and nutritional properties of milk. *Molecules*, *28*(5), 2023.

Runthala, A., Mbye, M., Ayyash, M., Xu, Y., & Kamal-Eldin, A. (2023). Caseins: versatility of their micellar organization in relation to the functional and nutritional properties of milk. *Molecules*, *28*(5), 2023.

Santpere, G. (2023). Genetic Variation in transcription factor binding sites. *International journal of molecular sciences*, *24*(5), 5038.

Scumaci, D., Trimboli, F., Dell'Aquila, L., Concolino, A., Pappaianni, G., Tammè, L., ... & Britti, D. (2015). Proteomics-driven analysis of ovine whey colostrum. *PLoS One*, *10*(2), e0117433.

Sebastiani, C., Arcangeli, C., Ciullo, M., Torricelli, M., Cinti, G., Fisichella, S., & Biagetti, M. (2020). Frequencies evaluation of β-casein gene polymorphisms in dairy cows reared in Central Italy. *Animals*, *10*(2), 252.

Shepelev, M. V., Kalinichenko, S. V., Deykin, A. V., & Korobko, I. V. (2018). Production of recombinant proteins in the milk of transgenic animals: current STATe and prospects. *Acta naturae (англоязычная версия)*, *10*(3 (38)), 40-47.

Shijun, L., Khan, R., Raza, S. H. A., Jieyun, H., Chugang, M., Kaster, N., ... & Linsen, Z. (2020). Function and characterization of the promoter region of perilipin 1 (PLIN1): Roles of E2F1, PLAG1, C/EBPβ, and SMAD3 in bovine adipocytes. *Genomics*, *112*(3), 2400-2409.

Shin, H. Y., Hennighausen, L., & Yoo, K. H. (2019). STAT5-driven enhancers tightly control temporal expression of mammary-specific genes. *Journal of mammary gland biology and neoplasia*, *24*, 61-71.

Sievers, F., & Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Multiple sequence alignment methods*, 105-116.

Sims, G. E., Jun, S. R., Wu, G. A., & Kim, S. H. (2009). Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the national academy of sciences*, *106*(40), 17077-17082.

Song, N., Luo, J., Huang, L., Zang, S., He, Q., Wu, J., & Huang, J. (2022). Mutation of signal transducer and activator of transcription 5 (STAT5) binding sites decreases milk allergen αS1-Casein content in goat mammary epithelial cells. *Foods*, *11*(3), 346.

Sun, J., Liu, J., Huang, B., Kan, X., Chen, G., Wang, W., & Fu, S. (2017). Kisspeptin-10 induces β-Casein synthesis via GPR54 and its downstream signaling pathways in bovine mammary epithelial cells. *International journal of molecular sciences*, *18*(12), 2621.

Syrovatkina, V., Alegre, K. O., Dey, R., & Huang, X. Y. (2016). Regulation, signaling, and physiological functions of G-proteins. *Journal of molecular biology*, *428*(19), 3850-3868.

Tayara, H., Tahir, M., & Chong, K. T. (2020). Identification of prokaryotic promoters and their strength by integrating heterogeneous features. *Genomics*, *112*(2), 1396-1403.

Thal, D. M., Glukhova, A., Sexton, P. M., & Christopoulos, A. (2018). Structural insights into G-protein-coupled receptor allostery. *Nature*, *559*(7712), 45-53.

Thomas, M. C., & Chiang, C. M. (2006). The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology*, *41*(3), 105-178.

Tian, M., Qi, Y., Zhang, X., Wu, Z., Chen, J., Chen, F., ... & Zhang, S. (2020). Regulation of the JAK2-STAT5 pathway by signaling molecules in the mammary gland. *Frontiers in cell and developmental biology*, *8*, 604896.

Truchet, S., & Honvo-Houéto, E. (2017). Physiology of milk secretion. *Best practice & research clinical endocrinology & metabolism*, *31*(4), 367-384.

Umarov, R., Kuwahara, H., Li, Y., Gao, X., & Solovyev, V. (2019). Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics*, *35*(16), 2730-2737.

Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., ... & Regev, A. (2022). The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, 603(7901), 455-463.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Kalush, F. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304-1351.

Vilotte, J. L., Chanat, E., Le Provost, F., Whitelaw, C. B. A., Kolb, A., & Shennan, D. B. (2013). Genetics and biosynthesis of milk proteins. *Advanced dairy chemistry: Volume 1A: Proteins: Basic Aspects 4th Edition*, 431-461.

Walker, S. R., Nelson, E. A., Yeh, J. E., Pinello, L., Yuan, G. C., & Frank, D. A. (2013). STAT5 outcompetes STAT3 to regulate the expression of the oncogenic transcriptional modulator BCL6. *Molecular and cellular biology*, *33*(15), 2879-2890.

Wang, C., Christin, J. R., Oktay, M. H., & Guo, W. (2017). Lineage-biased stem cells maintain estrogen-receptor-positive and-negative mouse mammary luminal lineages. *Cell reports*, *18*(12), 2825-2835.

Wang, F., Van Baal, J., Ma, L., Gao, X., Dijkstra, J., & Bu, D. (2022). MRCKα is a novel regulator of prolactin-induced lactogenesis in bovine mammary epithelial cells. *Animal nutrition*, *10*, 319-328.

Wang, J., Yang, P., Tang, B., Sun, X., Zhang, R., Guo, C., ... & Li, N. (2008). Expression and characterization of bioactive recombinant human α-lactalbumin in the milk of transgenic cloned cows. *Journal of dairy science*, *91*(12), 4466-4476.

Waston, C. J., Gordon, K. E., Robertson, M., & Clark, A. J. (1991). Interaction of DNA-binding proteins with a milk protein gene promoter in vitro: identification of a mammary gland-specific factor. *Nucleic acids research*, *19*(23), 6603-6610.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2— a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9), 1189-1191.

Watson, C. J. (2001). Stat transcription factors in mammary gland development and tumorigenesis. *Journal of mammary gland biology and neoplasia*, *6*, 115-127.

Watson, C. J., & Neoh, K. (2008, August). The STAT family of transcription factors have diverse roles in mammary gland development. In *seminars in cell & developmental biology* (Vol. 19, No. 4, pp. 401-406). Academic Press.

Wei, J., Wagner, S., Maclean, P., Brophy, B., Cole, S., Smolenski, G., ... & Laible, G. (2018). Cattle with a precise, zygote-mediated deletion safely eliminate the major milk allergen beta-lactoglobulin. *Scientific reports*, *8*(1), 7661.

Wettschureck, N., & Offermanns, S. (2005). Mammalian G proteins and their cell type specific functions. *Physiological reviews*, *85*(4), 1159-1204.

Wickramasinghe, S., Rincon, G., Islas-Trejo, A., & Medrano, J. F. (2012). Transcriptional profiling of bovine milk using RNA sequencing. *BMC genomics*, *13*(1), 1-14.

Wingelhofer, B., Neubauer, H. A., Valent, P., Han, X., Constantinescu, S. N., Gunning, P. T., ... & Moriggl, R. (2018). Implications of STAT3 and STAT5 signaling on gene regulation and chromatin remodeling in hematopoietic cancer. *Leukemia*, *32*(8), 1713-1726.

Wodas, L., Mackowski, M., Borowska, A., Puppel, K., Kuczynska, B., & Cieslak, J. (2020). Genes encoding equine β-lactoglobulin (LGB1 and LGB2): Polymorphism, expression, and impact on milk composition. *PLoS One*, *15*(4), e0232066.

Wyszomierski, S. L., & Rosen, J. M. (2001). Cooperative effects of STAT5 (signal transducer and activator of transcription 5) and C/EBP β (CCAAT/enhancer-binding protein-β) on β-casein gene transcription are mediated by the glucocorticoid receptor. *Molecular endocrinology*, *15*(2), 228-240.

Xie, J., LeBaron, M. J., Nevalainen, M. T., & Rui, H. (2002). Role of tyrosine kinase Jak2 in prolactin-induced differentiation and growth of mammary epithelial cells. *Journal of biological chemistry*, *277*(16), 14020-14030.

Xu, H., Luo, J., Ma, G., Zhang, X., Yao, D., Li, M., & Loor, J. J. (2018). Acyl-CoA synthetase short-chain family member 2 (ACSS2) is regulated by SREBP-1 and plays a role in fatty acid synthesis in caprine mammary epithelial cells. *Journal of cellular physiology*, *233*(2), 1005-1016.

Xu, R., Spencer, V. A., & Bissell, M. J. (2007). Extracellular matrix-regulated gene expression requires cooperation of SWI/SNF and transcription factors. *Journal of biological chemistry*, *282*(20), 14992-14999.

Zemke, N. R., Armand, E. J., Wang, W., Lee, S., Zhou, J., Li, Y. E., ... & Ren, B. (2023). Conserved and divergent gene regulatory programs of the mammalian neocortex. *Nature*, *624*(7991), 390-402.

Zhang, M. Q. (2007). Computational analyses of eukaryotic promoters. *BMC bioinformatics*, *8*(Suppl 6), S3.

Zhang, M., Li, F., Marquez-Lago, T. T., Leier, A., Fan, C., Kwoh, C. K., ... & Jia, C. (2019). MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. *Bioinformatics*, *35*(17), 2957-2965.

Zhang, M., Zheng, Y., Chen, W., Zhang, Y., Guo, Z., Zhang, Y., & Liu, J. (2015). Identifying an optimal promoter sequence of goat β-lactoglobulin gene for constructing high-expression vectors in mammary epithelial cells. *Small ruminant research*, *131*, 70-77.

Zhang, P., Zhang, H., & Wu, H. (2022). iPro-WAEL: a comprehensive and robust framework for identifying promoters in multiple species. *Nucleic Acids Research*, *50*(18), 10278-10289.

Zhang, X., Zhou, X., & Wang, X. (2013). Basics for bioinformatics. *Basics of bioinformatics: lecture notes of the graduate summer school on bioinformatics of China*, 1-25.

Zhou, Y., & Chen, J. J. (2021). STAT3 plays an important role in DNA replication by turning on WDHD1. *Cell & Bioscience*, *11*(1), 1-10.

Zhu, M., Li, S., Cao, X., Rashid, K., & Liu, T. (2023). The STAT family: Key transcription factors mediating crosstalk between cancer stem cells and tumor immune microenvironment. In *Seminars in cancer biology* (Vol. 88, pp. 18-31). Academic Press.

Zhu, Y., Li, F., Xiang, D., Akutsu, T., Song, J., & Jia, C. (2021). Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. *Briefings in bioinformatics*, *22*(4), bbaa299.