ABSTRACT

Title of dissertation: A WIDE SCALE INVESTIGATION INTO LNCRNA IN *BOS TAURUS*

Alexis Marceau, Doctor of Philosophy, 2023

Dissertation directed by: Professor Li Ma
University of Maryland Department of Animal and Avian Science

Although the history of genetic research has focused on genes and gene products, there is an interesting emerging subclass of genetic elements: long noncoding RNAs (lncRNAs). These are portions of the genome that are longer than 200 base pairs in length and are transcribed from DNA to RNA but do not yield a protein. The function of lncRNA is wide reaching and difficult to define; however, they are predominantly linked to the regulation of gene expression. This is done via transcriptional control, translation control, pre- and post- transcriptional and translational control, epigenetic modifications, RNA processing, as well as other methods [1].

In this dissertation, multiple *Bos taurus* tissues across various life conditions were investigated in order to identify lncRNA and to begin making predictions about the role and function of identified transcripts. First, lncRNA were identified and analyzed in *Bos taurus* rumen tissue in pre-weaning and post-weaning cattle. lncRNA were implicated in the weaning process and demonstrated enrichment in complex traits, indicating the continued impact rumen-associated lncRNA have on dairy cattle. Following this study, mammary tissues from dry and lactating cattle were used for lncRNA analysis, in relation to the lacta-

tion processes. This study revealed both the presence and impact of mammary lncRNA, and identified lncRNA associated with genes and biological processes that are strongly linked to lactation and mammary tissue function. Subsequently, immune system related tissues were analyzed for lncRNA and their roles. This investigation demonstrated lncRNA to be present in all investigated tissues, including transcripts being repeatedly present. Further analysis into identified lncRNA associated transcripts with genes and functions that are crucial to immune response. Finally, a tutorial was created to make lncRNA identification research more easily accessible to future researchers. The findings and creations of this dissertation increase the knowledge base of lncRNA and their role, allowing for further research endeavors and improvements in *Bos taurus* husbandry.

A WIDE SCALE INVESTIGATION INTO LNCRNA IN *BOS TAURUS*

by

Alexis Marceau

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Li Ma, Chair/Advisor
Professor Lindley Darden
Professor Najib El-Sayed
Dr. George Liu
Professor Mohamed Salem

Dedication

To my mom, I can't make you a coauthor so I hope this is enough.

Acknowledgments

There are more people to thank than words in this dissertation. From the smallest words of encouragement to those who read and reread my ramblings, none of this could have happened without all of you.

First, I would like to thank my advisor, Dr. Li Ma, and the defense committee; for believing in me and allowing me to work in my own strange ways. I am forever grateful for your role in this chapter of my life. This gratitude extends to the Department of Animal and Avian Science, the Graduate School, and the University of Maryland for their varied supports.

I would also like to thank all my lab mates, past and present. From commiserating to encouraging to checking in, you have all provided a support network I could not have succeeded without.

On more personal notes, there are many people who deserve to have their names associated with this work. As an homage to my time working in a library, in alphabetical order: Tera German, for holding me accountable, even when I railed against it. Jordan Graham, for contributions that are uncountable, invaluable, and everlasting. Zach Hood, for being both my moral and logical compass for the past 15 years. Tori Iqbal, for being a lighthouse in the storm of graduate school since well before my first day. Jonathan Mason, for being a new mentor when I didn't know I needed one. And Shannon Schreiner, who I truly owe everything to. I'd also like to thank my family, both human and animal, for their love and support. It was undying and unwavering even when I felt as though I was dying or wavering.

Finally, on the most personal notes, I would like to thank my partner, Nathan White, of the past decade. You have brightened my life since we were kids; without your humor, support, and love, I don't know where I'd be, but I'd sure wouldn't have laughed near as much.

Table of Contents

Chapter 1: Literature Review

1.1 Study Objectives

The overall objective of this study is to explore lncRNA across multiple *Bos taurus* tissues and developmental time points to build a deeper understanding of the genetic landscape that shapes these important production animals.

This dissertation is organized as follows: **Chapter 1** seeks to establish a knowledge base about topics related to each research endeavor. **Chapter 2** investigates lncRNA in cattle rumen tissue before and after weaning. **Chapter 3** explores lncRNA in cattle mammary tissue during the dry and lactating periods. **Chapter 4** is a multi-tissue analysis focusing on tissues associated with the immune system found in *Bos taurus*. After identification, in all studies, lncRNA are analyzed for trends and functional annotation. **Chapter 5** is a plain language tutorial for lncRNA identification from genetic sequence data. This is intended to fill the current gap in accessible lncRNA identification tools. The final chapter, **Chapter 6**, summarizes the conclusions of all lncRNA identification projects and discusses avenues for future research, as well as provides closing remarks.

To begin this dissertation, I will review pertinent literature about related topics. This will include: what are cattle; function, development, and genetic changes associated with the investigated tissues; an overview of long non-coding RNA (including their history, influential lncRNA, and the current status of identification software); reviews of recent, related studies; and the objectives of the studies conducted within this dissertation.

1.2 An Overview of Cattle

Bos taurus, or cattle, are hoofed ruminant livestock animals commonly used for their meat and dairy. Descending from wild ox, modern cattle have been traced back to a small population of aurochs dating back 10,500 years; having been selectively bred to increase meat and milk production since their domestication [4]. Today, *Bos taurus*cattle tend to be divided into two groups: those raised for meat (including breeds such as Angus and Hereford) and those raised for dairy (breeds include Holstein, Guernsey, Brown Swiss, and Jersey, to name a few), although it is possible to use a single animal for both purposes [1, 32]. They are found world-wide, being more concentrated in regions with large, open areas where herds feed on herbaceous plant material; cows utilize a ruminant digestive system, using a four compartment stomach to break down plant material into usable nutrients. [2]

Cattle tend to reproduce on a yearly basis, gestating for 9 months before birthing a single calf (although it is possible for multiples to be born, this is not common) [2]. Like most mammals, once a cow has birthed her offspring, she will begin to lactate. She will proceed to lactate for up to 12 months, where she will be milked by hand or machine, before her milk is further processed into consumer ready milk or other dairy products [31]. Depending on the nature of the farm (dairy or beef), the offspring may be kept in the herd to generate more milk and/or offspring or it may be raised to be processed for beef when it reaches the appropriate size [38].

1.3 Functions of Tissues of Interest

1.3.1 Functions of Rumen Tissue

As previously mentioned, cattle are ruminant animals. This means they utilize a multi chambered stomach to ferment feed, breaking it down from indigestible material into en-

ergy precursors [41]. The largest portion of the chambered system is the rumen; this large stomach compartment is made up of several sacs and holds 25+ gallons of material. It represent several key functions: absorption, transport, volatile fatty acid metabolism, and protection[12, 42].

When feed is ingested by cattle, it will be chewed quickly and swallowed. The swallowed material enters the rumen, liquid material passes quickly to the rest of the stomach and solid material faces one of two fates: be regurgitated and chewed further (this is called ”chewing the cud”) or form a dense mat within the rumen. The mat of solid material ferments in the rumen for up to 48 hours; during this time, a carefully balanced community of microbes goes to work breaking down the material into nutrients the animal can utilize [41]. This process is assisted by the constant movement and contractions of the rumen; a healthy animal should have one to two rumen contractions a minute. This movement functions to mix the rumen content, force rumen content into contact with undigested material, compact material into the fermenting mat, and move appropriately digested material out of the rumen [39]. This process allows the cattle to extract more nutrients from their food, as well as allows farmers to feed by-products to their livestock, keeping feed costs down while still ensuring adequate nutrition for their herd [33].

1.3.2 Mammary Tissue Function

As time has progressed, dairy cattle have been selectively bred to favor those with higher milk production as a more productive animal is a more valuable animal. The most important organ for this production is the mammary gland: a pair of apocrine glands located on each side of the anterior chest wall. This gland is made up of three distinct regions: skin, stroma, and parenchyma [23]. The skin is represented by the nipple, which functions to secrete milk, and the areola, which lubricates the nipple during nursing [3]. The stroma

is composed of fibrous and fatty components, which act together to create the framework of the breast around the parenchyma. Finally, the parenchyma is made up of 15 to 20 alveoli-containing lobules that branch into a duct network. These ducts come together to form several major (or main) ducts that each drain into a lactiferous sinus, where milk is collected. These sinuses are arranged radially around the nipple and are stimulated to release milk when a calf begins to suckle (or when a cow is milked) [23].

Given milk is primarily water, the additional nutrients must be supplied by the animal's diet. Once the food material has been adequately digested, the nutrients can be absorbed into the blood stream, where they can be shuttled to various locations, one of which is the mammary gland. Once the nutrient rich blood reaches the mammary gland, epithelial cells are able to uptake the nutrients and infuse them into the developing milk [20]. The excretion of the milk, is a two-staged process: a preparatory phase and an excretion phase. The preparatory phase is composed of cellular and enzymatic changes within the alveolar cells. The excretion phase is driven by hormonal changes, with prolactin playing a key role [40]. Overall, the milk production process is complex and multifaceted.

1.3.3 Immune System Tissues and Their Functions

The world is riddled with pathogens that, although just trying to live and reproduce, cause sickness and disease. The immune system is the line of defense against these pathogens. The immune system's function can be summarized with three main tasks: fight pathogens, neutralize harmful environmental particulates, and protect the body from rogue cells (like cancer) [21]. The immune system protects the body from outside threats (both living and environmental) with 2 distinct defense systems: the innate immune response and the adaptive immune response. The innate immune system utilizes preventative measures, such as skin and mucous membranes, and responsive cells, such as phagocytes, to act immediately,

although in a non-pathogen specific way. Phagocytes enclose and digest foreign particles, neutralizing the threat; they are also able to trigger cascades of signaling pathways to prepare the adaptive immune system for a pathogen-specific response. The adaptive immune system requires activation before engaging in a specific response, targeting specific pathogens. This system is made up of lymphocytes (both B cells and T cells) and antibodies. T cells use a cell surface protein to interact with germs in a lock-and-key function, replicating and creating more T cells once a matching pathogen has been found. There are three types of T cells: helper T cells, which use chemical messengers to activate further adaptive immune responses; cytotoxic T cells, which detect cells infected by viruses and tumor cells and destroy them; and memory T cells, helper T cells that remain after infection and are primed to initiate an immune response if the pathogen returns. B cells are activated by helper T cells: the T cell interacts with a matching B cell (the match being made between the cell surface proteins), causing the B cell to multiply and transform into antibody producing plasma cells. The transformed B cells can then release the antibodies into the blood stream, creating a flurry of pathogen-specific antibodies within the blood. The final component of the adaptive immune system is antibodies. These are protein-sugar compounds that circulate within the bloodstream and can quickly detect pathogens. Once bound to the pathogen, the pathogen is rendered useless and other immune cells are attracted to begin combating the infection. Antibodies also act in a specific lock-and-key fashion. The final function of the immune system is to protect the body from cells which have been compromised by DNA damage or viral infection. Natural killer cells scan the surface of cells, looking for abnormalities and utilize cell toxins to dispatch identified cells, in a similar fashion to cytotoxic T cells [37].

1.4 Development of Tissues of Interest

1.4.1 Development of Rumen Tissue

Although cattle are ruminants, their rumen is underdeveloped and nonfunctional at birth. Upon entering the world, the rumen takes up approximately 35% of the stomach capacity and is not utilized as the calf is exclusively feeding on milk. When suckling, the milk the calf is consuming travels down the esophageal groove, bypassing the rumen and directly entering the abomasum for digestion. Rumen maturation is two-fold, the development of rumen papillae and the development of rumen muscle mass. Fermentation is required to increase the size and muscle mass of the rumen, but this is not possible without establishing a microbial community within the rumen. Luckily, the birthing process, existing in their post-natal environment, and the introduction of dry matter all lead to the colonization of microbes in the rumen and assist in the development of this crucial microbial environment [36]. When dry feed is introduced, acetic, propionic, and butyric acids are produced, lowering the pH of the anaerobic rumen, and allowing for continued microbial growth and development; the development of the microbial ecosystem also stimulates the growth of papillae. The introduction of dry matter also physically stimulates the development of rumen mass by forcing the tissue to expand to accommodate the feed. The combination of physically increasing the rumen size and the development of necessary papillae and microbes leads to the maturation of the rumen [42].

1.4.2 Development of Mammary Tissue

Similar to rumen tissue, mammary tissue is underdeveloped at birth. Although unlike rumen tissue, mammary tissue remains underdeveloped until the cow approaches puberty. For dairy cows, puberty occurs after 13-14 months when they begin to ovulate and are

now physically ready to calve [18]. At this point, the increase in hormone levels leads to an increase in mammary tissue size. However maturation arrests after the mammary fat pad fills. The mammary tissue is larger and the duct network has begun to branch but is still not fully developed or functional [13]. The maturation of mammary tissue resumes during gestation and after birth with a surge of hormones like progesterone and estrogen. The next stage of development includes another increase in size and the finalization of the duct network branching. The duct network also becomes surrounded by connective tissue, developing into secretory alveoli. The final stage of development is triggered by an increase in prolactin and leads the ducts to fill with milk. The mammary gland is now fully developed and ready to nurse a calf. The mammary gland will be refilled with milk after the calf nurses until the calf is weaned. At that point, the gland will remain at its final size, but the alveoli will shrink and disappear until the cow is pregnant again where the cycle will continue. Once a cow reaches menopause and is no longer able to become pregnant, the lack of estrogen will cause the mammary gland to atrophy and decrease in size [23].

1.4.3 Development of the Immune System and Associated Tissues

The immune system in cattle calves begins developing from conception and does not reach full maturity until 6 months postpartum. Given the high number of pathogens the young animals are exposed to, it is crucial that their immune system develops strongly. Fetal calves are reliant on their mothers' immune system and their innate immune system, as they are yet to be exposed to any pathogens to trigger and train their adaptive immune system. It is not until late gestation that calves develop their phagocytic cells, increasing their innate immune systems' strength. Regarding their adaptive immune systems, unless a calf has been infected prior to birth, newborn calves lack antibodies completely. B and T cells are present in neonatal calves, but their numbers decrease as birth approaches.

Since newborn calves are immunologically naïve, they rely heavily on the passive immunity provided to them by their mother's colostrum (which is primarily composed of antibodies, cytokines, and cells). In healthy calves, the colostrum is absorbed by the intestinal cells, giving the baby's immune system a jump start [10]. As the immune system develops, the amount of immune system cells must increase. Most immune system cells are made in the bone marrow before traveling to other parts of the body. Lymphocytes are made in the bone marrow and those that remain in the bone marrow mature into B cells. Some lymphocytes will travel to the thymus, there they will mature into T cells. After development, many immune cells are found in strategically placed lymph nodes and within lymphatic fluid. The localization of these immune cell home bases allows for rapid immune responses when an infection looms. The spleen also holds a small army of immune system cells; when blood passes through the spleen, it can be filtered for pathogens and the local immune cells can respond to any detected threats [34].

1.5 Genetic Changes in Tissues of Interest

1.5.1 Genetic Changes of Rumen Tissue Through Development

While rumen development is reliant on feed intake, there are necessary genetic changes that occur to facilitate the development. In fact, there have been over 6,000 genes determined to be differentially expressed at various stages of tissue development and using functional annotations and protein-protein interactions, 9 genes revealed themselves to be involved directly with rumen development [45]. The uptick in butyrate as dry feed matter is introduced can lead to changes in structural integrity, epigenetic regulation, signaling pathways, and more [11]. Direct addition of butyrate via milk or top-dressed feeds led to the identification of 1,977 differentially expressed genes in rumen epithelial cells. These genes had associations with functions like regulatory and signaling pathways [27]. Changes in

chromatin structures, epigenetic interactions (for example, DNA methylation and histone modification), and extracellular interactions were also seen as a calf was weaned off milk and onto dry feed [14]. The growing microbial community within the rumen is also influential on genetic expression in early development of rumen tissue. A multifaceted study investigating the relationship between the rumen's microbiota and host transcriptomes revealed the volatile fatty acids released by the microbial colony led to changes to the mRNA and miRNA within the host's genome, with three gene modules and one miRNA associating significantly with volatile fatty acids. 13 genetic elements were also associated with a rumen bacteria cluster [30]. Findings linking rumen development, both physically and microbiologically, to genes indicate there are genetic changes happening along side the morphological changes that occur as the rumen develops.

1.5.2 Genetic Changes in Mammary Tissue As Development Progresses

Given the large changes that occur in the mammary gland as mammogenesis occurs, there must be many genes that require differential expression to generate such changes. Knockout models uncovered 2 genes that, when knocked out, disrupted the formation of mammary placodes. When investigating mammary gland aplasia, it seems a mutation in the TBX3 genes leads to the lack of mammary placodes—this gene appears to be key to mammary gland development. Ductal tree formation also appears to rely on the parathyroid hormone-releasing protein (PTHrP) gene, as disruptions in this gene lead to bud formation that ceases before branching into ductal trees [19]. When analyzing mammary tissue at various time points relating to pregnancy, and thus mammary gland development, 4,843 differentially expressed genes were identified. Included in those genes were 110 genes that showed a peak profile that was associated only with lactation, having much lower expression during pregnancy and after lactation. Also included in these identified genes were sev-

eral transcription factors, indicating there are cascading effects of differentially expressed genes as maturation progresses [9]. There are also various growth factor expression profiles that change over the course of mammary development, from prenatal mammary cell differentiation to pubescent and gestational maturation. For these profiles to vary, the expression patterns of their associated genes must also vary [29].

1.5.3 Genes Associated with the Immune System and its Activation

While there are obviously a variety of genes needed to create the various cells associated with the immune system, are there differences in gene expression in response to immune system activation? The answer is yes. Toll-like receptors (TLRs) are a key component of innate immunity and when investigating unstimulated monocytes as compared to TLR4-stimulated monocytes, 1471 expression quantitative trait loci (eQTLs) were found to be present in the stimulated monocytes while absent in the unstimulated monocytes. These eQTLs were able to be associated with genes related to TLR4 activation. These findings demonstrate there are genetic changes that are key to innate immune response [24]. Relating to the adaptive immune system, changes in gene expression are crucial. Without these changes, there is no way for adaptive immune cells to modify their cell surface proteins and become specialized. First, to transform from an early thymic progenitor cell into a T cell, B cell, natural killer cell, myeloid cell, or dendritic cell, the expression of genes like Flt3, CD24, and CCR9 must vary. The interactions these cell have with specific ligands also triggers cascades of genetic expression changes, leading cells to differentiate and follow various pathways. Activation of adaptive immune cells is also dependent on interactions leading to gene expression changes, these gene expression changes then lead to distinct subpopulations of immune cells, all ready to fight specific pathogens [8].

1.6 Long Non-Coding RNA

1.6.1 The History of lncRNA

It has been known since the 1950s that the amount of present haploid DNA showed little to no correlation with the organism's size or its complexity. If it did, the larger the genome, the more complex the creature; this simply was not the case. In the 1970s, it appeared that the genome was not all converted into genes as the amount of material present was far greater than the predicted number of genes that humans could possibly have. This led to the belief that majority of the genome was noncoding "junk," being too rife with transposons, pseudogenes, and repeating sequences to be of any use. Although the idea that the rest of the genome was just junk was short-lived, soon it was uncovered that portions of the genome were used as various forms of RNA (rRNA and tRNA being the first two major classes discovered). 1977 brought the discovery of introns and the 1980s bore snRNAs and snoRNAs. The discovery of snRNAs and snoRNAs gave way to the theory that these small portions of RNA acted in a regulatory fashion, being implicated in post-transcriptional RNA processing. It wasn't until whole genome sequencing technologies began to rise in the early 2000s that researchers were able to begin to appreciate the vastness of non-coding portions of the genome [25].

The early 1990s brought about identification of the first long noncoding RNAs (lncRNA), noncoding RNA transcripts that exceed 200 base pairs in length. These findings included H19 in 1990 and Xist in 1992, though they were hotly contested as they were not readily assigned a function, therefore casting doubt on the value of these findings. As technology advanced, it was clear lncRNA were here to stay [25]. Now, lncRNA have been discovered in all species where they have been looked for (from viruses to prokaryotes to fungi to animals), they play roles relating to cis- and trans- acting gene expression, cis- and

trans- acting chromatin remodeling, molecular scaffolding, cellular process regulation, cellular differentiation, sex determination, sequence divergence, and likely other individual lncRNA-specific functions that have yet to be discovered [22].

1.6.2 Notable lncRNA and their Functions

In January of 1990, Molecular and Cellular Biology published Brannan et al's publication "The product of the H19 gene may function as an RNA." This was the first publication to feature a lncRNA. It appeared the mouse gene H19 was a hepatic fetal-specific mRNA that was under the transcriptional control of a trans-acting locus termed "raf." The "gene" was made into mRNA, but what protein it was supposed to produce was unclear. There were many small open reading frames and multiple termination codons in the transcript. The RNA also did not appear to interact with translational machinery. All this evidence led this group to hypothesize that H19 was not a normal RNA that would be made into a functional protein, but an RNA molecule that never proceeded further in transcription-translation pathway [5].

The next big lncRNA, though not yet classified as a lncRNA, to enter the scene came a year later in January of 1991: XIST. This gene was implicated in X chromosome inactivation. In-situ hybridization demonstrated XIST repeatedly localized to the Xq13 region in inactivated X chromosomes [7]. Further research into XIST revealed the "gene product" was actually a 15kb inactive X-specific transcript that lacked a conserved open reading frame. The transcript was riddled with tandem repeats and did not associate with translational machinery; it is also almost exclusively found in the nucleus. These findings led researchers to believe XIST was a functional RNA supporting X chromosome inactivation, and was later coined as a lncRNA [6].

Perhaps one of the most recognizable lncRNA to be discovered is HOTAIR, a mem-

ber of the HOX noncoding RNA (ncRNA) family. What makes the HOX ncRNA family so memorable is the tendency to be expressed near developmental axes and differential histone modifications leading to variable RNA polymerase accessibility. HOTAIR is a specific 2.2 kilobase ncRNA located within the HOXC locus. It appears to interact with the polycomb repressive complex 2 (PCR2) and is a necessary component for both PCR2 occupancy and the histone H3 lysine-27 trimethylation of the HOXD locus. If other lncRNA act in a similar manner to HOTAIR, it would indicate lncRNA may be able to act as gene silencers from a distance, associating them with gene regulation [35].

Today, lncRNA have been implicated in a wide range of biological functions: chromatin remodeling, transcriptional activation, transcriptional interference, RNA processing, and mRNA translation. These associations connect lncRNA to growth, development, stress responses, cell differentiation, regulation, cell cycle control, and gene expression control at all levels. These non-coding transcripts are pervasive and extensive [44].

1.6.3 Gaps in lncRNA Identification Software

With the growing interest in lncRNA, it would be expected to see an increase in lncRNA identification tools, and while this is true, the available tools are sorely lacking in usability.

Developed in 2014, PLEK, or "the predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme", is a computational pipeline used to distinguish lncRNA from mRNA using a k-mer scheme. In humans, this tool was demonstrated to be robust and accurate with highly accurate results in other vertebrates. Although a useful tool, this tool functions more to distinguish between two groups than to identify lncRNA from scratch [26].

Lncident was developed in 2016 and intended to be a novel method to identify lncRNA based on the researchers' belief that rapid identification of lncRNA and mRNA were cru-

cial for functional discovery of lncRNA. With the successful development of this program, developers claim it to be high speed without the loss of accuracy. Lncident is available as an R package but this does limit its usefulness as users must be comfortable with R. A webserver was created for this tool but it was limited to human, mouse, and C. elegans models; at this time, the webserver is not available [16].

LncFinder is a tool created in 2019. It was developed to be an integrated platform to promote lncRNA research; it aimed to be used for both identification of lncRNA, as well as attribution of properties. It also aimed to improve currently available tools. The developers found the tool did outperform contemporaneous tools and was robust, providing satisfactory results. Like Lncident, this tool is available as an R package but does require users to be comfortable with R. Again, similarly to Lncident, a web server was also developed but currently is not available [15].

There are other lncRNA identification tools and tutorials available, however they all appear to have serious drawbacks. Ranging from lack of maintenance to limited functionality, it is not easy to find a useful, plain language guide to identifying lncRNA at this time.

1.7 Contemporaneous Studies

1.7.1 The Conservation and Signatures of lincRNAs in Marek's Disease of Chicken

In this study, He et al endeavored to investigate lincRNA expression as it related to the T cell lymphoma induced in chickens by the Marek's disease virus. Marek's disease virus, and the resulting Marek's disease, is highly contagious and deadly; it also is prone to increases in virulence, rendering current vaccines redundant. When investigating the chicken bursa tissue at 3 time points post-infection, 1225 transcripts in 1056 loci were revealed.

The identified transcripts were shorter, had fewer exons per transcript, showed lower expression, and less sequence conservation when compared to coding transcripts. A conserved transcript of note showed positional equivalence with the HOTTIP lincRNA, a lincRNA that has been well documented in human and mice genomes. HOTTIP is believed to act as a 5' HOXA gene activator and correlation analysis also showed this transcript to be positively correlated with HOXA expression in analyzed samples. Positional analysis was used to ascertain genes neighboring the identified lincRNA, these neighboring genes were then assessed for correlation with the transcripts, and finally correlated genes were assessed for gene ontology. This analysis revealed lincRNA to be connected to the regulation of transcription, signaling and regulation of development.

When connecting lincRNA to Marek's disease and resistance to the virus, linc-satb1 emerged as an interesting finding. This lincRNA was positively associated with defense response, inflammatory response, lymphocyte activation and response to external stimulus and negatively associated with cell cycle-related functions such as cell cycle process and DNA replication. It was also only highly expressed in birds that were infected with the virus from an MD-resistant line, implicating this lincRNA in the immune response to MDV. The nearby gene, SATB1, is a genome organizer that regulates chromatin structure and a transcription factor that controls many genes involved in T cell development and activation; these associations further connect this gene and the lincRNA to the immune response initiated by MDV. Further investigation revealed that SATB1 could promote tumor emergence and progression.

In conclusion, this study successfully identified lincRNA in chicken bursa and was able to associate identified transcripts with functions of interest. A lincRNA was also identified and assigned to play a role in the immunological response initiated by a Marek's disease virus infection [17].

1.7.2 Comprehensive Analysis of Differentially Expressed mRNA, lncRNA and circRNA and Their ceRNA Networks in the Longissimus Dorsi Muscle of Two Different Pig Breeds

In porcine, meat quality is crucial to sale price and economic value; muscle growth and fatness are complex traits that greatly influence meat quality. In this study, two pig breeds with varying muscularity traits were compared using mRNA, lncRNA, and circRNA. This investigation identified 854 mRNA, 233 lncRNA, and 66 circRNA that were differentially expressed in longissimus dorsi muscles in Chinese Huainan pigs (HN pigs) and Western commercial Duroc X (Landrace X Yorkshire) pigs (DLY pigs).

Identification efforts found 9,068 lncRNA to be present in samples and differential expression analysis isolated 4,859 lncRNA as common to both HN and DLY pigs. 233 of these lncRNA were statistically significantly differentially expressed between the two breeds. Identified differentially expressed lncRNA were then used with circRNA, miRNA, and mRNA data to construct potential regulatory networks as lncRNA have been regularly implicated as regulatory agents. Common nodes across networks included features such as the genes MYOD1 and PPARD and miRNAs miR-423-5p and miR-874. In the construction of regulatory networks, PPARD was targeted by three miRNAs, these miRNAs then targeted five lncRNA and 16 circRNA; MYOD1 showed connections between miR-296-5p; this miRNA then connected to two lncRNA and three circRNA.

Overall, identified lncRNA appear to function in the regulation of biological processes that affect muscle growth in a variety of ways. Researchers hypothesized lncRNA may change chromatin spatial conformation by binding to target genes, altering their expression level. They also theorized that lncRNA may interact with transcription factors. In the construction of regulatory networks, lncRNA were able to be associated with miRNA that led

to differential expression of genes that may contribute to the differences seen in porcine meat composition [43].

1.7.3 Genome-Wide Identification and Characterization of Long Non-Coding RNAs in Longissimus dorsi Skeletal Muscle of Shandong Black Cattle and Luxi Cattle

Although lncRNA have been repeatedly associated with the regulation of cell differentiation, fat synthesis, and embryonic development, this study sought to investigate lncRNA's role in skeletal muscle development in Shandong Black and Luxi cattle. Similarly to porcine production, meat quality in cattle is very economically important and impacted by many factors; the genetics behind improving meat quality is of utmost importance and many breeding decisions are made to maximize the development of cattle muscle. Comparisons between the two cattle breeds did reveal genetic differences.

When comparing the muscle fibers of both cattle breeds, there are demonstrable differences between the fibers: Shangdong cattle muscle fibers were significantly larger, showed less muscle fiber density, and a lower ratio of both fast and slow twitch fibers when compared to Luxi cattle. Further analysis of the genetic elements contributing to those differences identified 1,415 differentially expressed mRNA and 480 differentially expressed lncRNA. Correlation analysis between the differentially expressed mRNA and differentially expressed lncRNA yielded 43,844 correlated pairs and generated a list of 387 lncRNA to be further analyzed. Based on location, the 387 lncRNA targeted 1,164 genes including MYORG, Wnt4, PAK1, and ADCY7. Gene ontology analysis of the targeted genes revealed GO items such as cellular organization development, single multicellular organization process, and multicellular organic process. KEGG analysis of the 1,164 targeted genes showed enrichment of the calcium signaling pathway, the AMPK signaling pathway, the

cGMP-PKG signaling pathway, and the PPAR signaling pathway—these pathways related heavily to muscle development. Interaction network development also associated multiple lncRNA with several muscle types, including skeletal muscle.

The most promising lncRNA identified appeared to target bta-miR-133a. The binding between the lncRNA and this miRNA may influence the regulation of skeletal muscle development by competitively inhibiting the expression of the target gene Pax7. In conclusion, the finding of this study associated lncRNA with skeletal muscle development in cattle; the differential expression of these lncRNA between analyzed beef cattle breed indicate lncRNA may contribute to the differences seen between the two breeds and allow for continuation of improvement of beef cattle meat quality [28].

1.8 References

[1] *5 Best Beef Cattle Breeds*. URL: https://americancowboy.com/lifestyle/5-best-beef-cattle-breeds-24440/.

[2] *ADW: Bos taurus: INFORMATION*. URL: https://animaldiversity.org/accounts/Bos_taurus/.

[3] *Areola — anatomy — Britannica*. URL: https://www.britannica.com/science/areola.

[4] Ruth Bollongino et al. "Modern Taurine Cattle Descended from Small Number of Near-Eastern Founders". In: *Molecular Biology and Evolution* 29.9 (Sept. 2012), pp. 2101–2104. ISSN: 0737-4038. DOI: 10.1093/MOLBEV/MSS092. URL: https://dx.doi.org/10.1093/molbev/mss092.

[5] Camilynn I. Brannan et al. "The product of the H19 gene may function as an RNA". In: *Molecular and cellular biology* 10.1 (Jan. 1990), pp. 28–36. ISSN: 0270-7306. DOI: 10.1128/MCB.10.1.28-36.1990. URL: https://pubmed.ncbi.nlm.nih.gov/1688465/.

[6] Neil Brockdorff et al. "The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus". In: *Cell* 71.3 (Oct. 1992), pp. 515–526. ISSN: 0092-8674. DOI: 10.1016/0092-8674(92)90519-I. URL: https://pubmed.ncbi.nlm.nih.gov/1423610/.

[7] Carolyn J. Brown et al. "Localization of the X inactivation centre on the human X chromosome in Xq13". In: *Nature 1991 349:6304* 349.6304 (1991), pp. 82–84. ISSN: 1476-4687. DOI: 10.1038/349082a0. URL: https://www.nature.com/articles/349082a0.

[8] R Luz Elena Cano and H. Damaris E. Lopera. “Introduction to T and B lymphocytes”. In: (July 2013). URL: https://www.ncbi.nlm.nih.gov/books/NBK459471/.

[9] Theresa Casey et al. “Transcriptome analysis of epithelial and stromal contributions to mammogenesis in three week prepartum cows”. In: *PloS one* 6.7 (2011). ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0022541. URL: https://pubmed.ncbi.nlm.nih.gov/21829467/.

[10] Christopher C.L. Chase, David J. Hurley, and Adrian J. Reber. “Neonatal Immune Development in the Calf and Its Impact on Vaccine Response”. In: *The Veterinary Clinics of North America. Food Animal Practice* 24.1 (Mar. 2008), p. 87. ISSN: 07490720. DOI: 10.1016/J.CVFA.2007.11.001. URL: /pmc/articles/PMC7127081/%20/pmc/articles/PMC7127081/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7127081/.

[11] Qiyu Diao, Rong Zhang, and Tong Fu. “Review of strategies to promote rumen development in calves”. In: *Animals* 9.8 (Aug. 2019), p. 490. ISSN: 20762615. DOI: 10.3390/ani9080490.

[12] *Digestive Anatomy in Ruminants*. URL: http://www.vivo.colostate.edu/hbooks/pathphys/digestion/herbivores/rumen_anat.html.

[13] Laurence Finot, Eric Chanat, and Frederic Dessauge. “Bovine mammary gland development: new insights into the epithelial hierarchy”. In: *bioRxiv* (Jan. 2018), p. 251637. DOI: 10.1101/251637. URL: https://www.biorxiv.org/content/10.1101/251637v1%20https://www.biorxiv.org/content/10.1101/251637v1.abstract.

[14] Yahui Gao et al. “Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning”. In: *Genomics* 113.4 (July 2021), pp. 2045–2055. ISSN: 0888-7543. DOI: 10.1016/J.YGENO.2021.04.039.

[15] Siyu Han et al. “LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property”. In: *Briefings in Bioinformatics* 20.6 (Nov. 2019), pp. 2009–2027. ISSN: 14774054. DOI: 10.1093/BIB/BBY065. URL: https://academic.oup.com/bib/article/20/6/2009/5062950.

[16] Siyu Han et al. “Lncident: A Tool for Rapid Identification of Long Noncoding RNAs Utilizing Sequence Intrinsic Composition and Open Reading Frame Information”. In: *International Journal of Genomics* 2016 (2016). ISSN: 23144378. DOI: 10.1155/2016/9185496.

[17] Yanghua He et al. “The conservation and signatures of lincRNAs in Marek’s disease of chicken”. In: *Scientific Reports* 5.1 (Oct. 2015), pp. 1–17. ISSN: 20452322. DOI: 10.1038/srep15184. URL: www.nature.com/scientificreports.

[18] *Heifer Development: Puberty — The Cattle Site*. URL: https://www.thecattlesite.com/articles/901/heifer-development-puberty.

[19] Julie R. Hens and John J. Wysolmerski. “Molecular mechanisms involved in the formation of the embryonic mammary gland”. In: *Breast Cancer Research* 7.5 (Oct. 2005), pp. 220–224. ISSN: 14655411. DOI: 10.1186/BCR1306/TABLES/1. URL: https://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr1306.

[20] *How Do Cows Make Milk? — Organic Valley*. URL: https://www.organicvalley.coop/blog/from-grass-to-glass-how-do-cows-make-milk/.

[21] "How does the immune system work?" In: (Apr. 2020). URL: https://www.ncbi.nlm.nih.gov/books/NBK279364/.

[22] Julien Jarroux, Antonin Morillon, and Marina Pinskaya. "History, Discovery, and Classification of lncRNAs". In: *Advances in experimental medicine and biology* 1008 (2017), pp. 1–46. ISSN: 0065-2598. DOI: 10.1007/978-981-10-5203-3{_}1. URL: https://pubmed.ncbi.nlm.nih.gov/28815535/.

[23] Yusuf S. Khan and Hussain Sajjad. "Anatomy, Thorax, Mammary Gland". In: *StatPearls* (July 2021). URL: https://www.ncbi.nlm.nih.gov/books/NBK547666/.

[24] Sarah Kim et al. "Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes". In: *Nature Communications 2014 5:1* 5.1 (Oct. 2014), pp. 1–7. ISSN: 2041-1723. DOI: 10.1038/ncomms6236. URL: https://www.nature.com/articles/ncomms6236.

[25] Johnny T.Y. Kung, David Colognori, and Jeannie T. Lee. *Long noncoding RNAs: Past, present, and future*. 2013. DOI: 10.1534/genetics.112.146704. URL: /pmc/articles/PMC3583990/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583990/.

[26] Aimin Li, Junying Zhang, and Zhongyin Zhou. "PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme". In: *BMC Bioinformatics* 15.1 (Sept. 2014), pp. 1–10. ISSN: 14712105. DOI: 10.1186/1471-2105-15-311/FIGURES/3. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-311.

[27] Shudai Lin et al. "Establishment and transcriptomic analyses of a cattle rumen epithelial primary cells (REPC) culture by bulk and single-cell RNA sequencing to elucidate interactions of butyrate and rumen development". In: *Heliyon* 6.6 (June 2020), e04112. ISSN: 24058440. DOI: 10.1016/j.heliyon.2020.e04112.

[28] Ruili Liu et al. “Genome-Wide Identification and Characterization of Long Non-Coding RNAs in Longissimus dorsi Skeletal Muscle of Shandong Black Cattle and Luxi Cattle”. In: *Frontiers in Genetics* 13 (May 2022), p. 849399. ISSN: 16648021. DOI: 10.3389/FGENE.2022.849399/BIBTEX.

[29] Hector Macias and Lindsay Hinck. “Mammary Gland Development”. In: *Wiley interdisciplinary reviews. Developmental biology* 1.4 (July 2012), p. 533. ISSN: 17597684. DOI: 10.1002/WDEV.35. URL: /pmc/articles/PMC3404495/%20/pmc/articles/PMC3404495/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3404495/.

[30] Nilusha Malmuthuge, Guanxiang Liang, and Le Luo Guan. “Regulation of rumen development in neonatal ruminants through microbial metagenomes and host transcriptomes”. In: *Genome Biology* 20.1 (Aug. 2019), pp. 1–16. ISSN: 1474760X. DOI: 10.1186/S13059-019-1786-0/FIGURES/6. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1786-0.

[31] *Managing Cow Lactation Cycles — The Cattle Site*. URL: https://www.thecattlesite.com/articles/4248/managing-cow-lactation-cycles/.

[32] *Name That Cow: The 6 Great Dairy Breeds — Dairy Discovery Zone*. URL: https://www.dairydiscoveryzone.com/blog/name-cow-6-great-dairy-breeds.

[33] Adam I Orr. “How Cows Eat Grass Exploring Cow Digestion”. In: (2011). URL: http://www.fda.gov/AnimalVeterinary.

[34] *Parts of the Immune System — Children’s Hospital of Philadelphia*. URL: https://www.chop.edu/centers-programs/vaccine-education-center/human-immune-system/parts-immune-system.

[35] John L. Rinn et al. “Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs”. In: *Cell* 129.7 (June 2007), p. 1311. ISSN: 00928674. DOI: 10.1016/J.CELL.2007.05.022. URL: /pmc/articles/PMC2084369/%20/pmc/articles/PMC2084369/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2084369/.

[36] *Rumen development of calves - Beef*. URL: https://www.canr.msu.edu/news/rumen-development-of-calves.

[37] “The innate and adaptive immune systems”. In: (July 2020). URL: https://www.ncbi.nlm.nih.gov/books/NBK279396/.

[38] *The Life Cycle of Beef Cattle Production — Peterson Farm Brothers*. URL: https://petersonfarmbrothers.com/the-life-cycle-of-beef-cattle-production/.

[39] *The ruminant digestive system*. URL: https://extension.umn.edu/dairy-nutrition/ruminant-digestive-system#stomach-compartments-1000460.

[40] H Allen Tucker and East Lansing. “Physiological Control of Mammary Growth, Lactogenesis, and Lactation 1”. In: *Journal of Dairy Science* 64 (), pp. 1403–1421. DOI: 10.3168/jds.S0022-0302(81)82711-7.

[41] *Understanding the Ruminant Animal Digestive System — Mississippi State University Extension Service*. URL: http://extension.msstate.edu/publications/understanding-the-ruminant-animal-digestive-system.

[42] Ransom L Baldwin Vi and Erin E Connor. “Rumen Function and Development”. In: (2017). DOI: 10.1016/j.cvfa.2017.06.001. URL: http://dx.doi.org/10.1016/j.cvfa.2017.06.001.

[43] Jing Wang et al. “Comprehensive Analysis of Differentially Expressed mRNA, lncRNA and circRNA and Their ceRNA Networks in the Longissimus Dorsi Muscle of Two Different Pig Breeds”. In: *International Journal of Molecular Sciences* 20.5 (Mar. 2019). ISSN: 14220067. DOI: 10.3390/IJMS20051107. URL: /pmc/articles/PMC6429497/%20/pmc/articles/PMC6429497/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6429497/.

[44] Xiaopei Zhang et al. “Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels”. In: *International Journal of Molecular Sciences* 20.22 (Nov. 2019). ISSN: 14220067. DOI: 10.3390/IJMS20225573. URL: /pmc/articles/PMC6888083/%20/pmc/articles/PMC6888083/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6888083/.

[45] Yapeng Zhang et al. “Transcriptome Analysis of Bovine Rumen Tissue in Three Developmental Stages”. In: *Frontiers in Genetics* 13 (Mar. 2022), p. 821406. ISSN: 16648021. DOI: 10.3389/FGENE.2022.821406/BIBTEX.

Chapter 2: Investigation of Rumen Long Noncoding RNA Before and After Weaning in Cattle

2.1 Abstract

Background:

This study aimed to identify long non-coding RNA (lncRNA) from the rumen tissue in dairy cattle, explore their features including expression and conservation levels, and reveal potential links between lncRNA and complex traits that may indicate important functional impacts of rumen lncRNA during the transition to the weaning period.

Results:

A total of six cattle rumen samples were taken with three replicates from before and after weaning periods, respectively. Total RNAs were extracted and sequenced with lncRNA discovered based on size, coding potential, sequence homology, and known protein domains. As a result, 404 and 234 rumen lncRNAs were identified before and after weaning, respectively. However, only nine of them were shared under two conditions, with 395 lncRNAs found only in pre-weaning tissues and 225 only in post-weaning samples. Interestingly, none of the nine common lncRNAs were differentially expressed between the two weaning conditions. LncRNA averaged shorter length, lower expression, and lower conservation scores than the genome overall, which is consistent with general lncRNA characteristics. By integrating rumen lncRNA before and after weaning with large-scale GWAS results in cattle, we reported significant enrichment of both pre- and after-weaning lncRNA with

traits of economic importance including production, reproduction, health, and body conformation phenotypes.

Conclusions:

The majority of rumen lncRNAs are uniquely expressed in one of the two weaning conditions, indicating a functional role of lncRNA in rumen development and transition of weaning. Notably, both pre- and post-weaning lncRNA showed significant enrichment with a variety of complex traits in dairy cattle, suggesting the importance of rumen lncRNA for cattle performance in the adult stage. These relationships should be further investigated to better understand the specific roles that lncRNAs play in rumen development and cow performance.

Note:

This chapter was previously published by BMC Genomics. Alexis Marceau was the main contributor and is the first author. The original citation is as follows:

Marceau, A., Gao, Y., Baldwin , R.L. et al. Investigation of rumen long noncoding RNA before and after weaning in cattle. BMC Genomics 23, 531 (2022). https://doi.org/10.1186/s12864-022-08758-4

2.2 Background

For many years, genomic research has focused on the direct line of genes to gene products to phenotypic observations. Recent advances in sequencing technologies have led to more in depth exploration of the genomic regions that are transcribed into RNA but rarely into a protein product. Connections have been established between these non-coding RNAs and regulation of gene expression in many organisms. Examples include the X chromosome inactivation, allelic imprinting, pluripotency control, cancer, and many other biological processes [23]. A subset of these noncoding RNAs are the long noncoding RNA (lncRNA): these transcripts are at least 200 nucleotides in length, with some reaching up to 32,000 nucleotides, and are found across nearly all species [34]. Previous research has identified large numbers of lncRNA across many model and non-model species, ranging from over 7,000 transcripts in cattle [20] to over 270,000 lncRNA in humans [28]. Recent research in cattle has revealed more lncRNA transcripts in many tissues across cattle breeds: over 4,000 in 6 different tissues between 2 Chinese cattle breeds, nearly 10,000 in 18 different bovine tissues, over 23,000 lncRNA in bovine testes tissue as they mature, and 1,535 lncRNAs in bovine oocytes [13, 11, 22]. Additionally, almost 8,000 lncRNAs were found to be associated with metabolic efficiency [31].

A key characteristic of the cow is the four-part stomach system, where feed enters the reticulum and rumen, passes through the omasum, and reaches the abomasum where digestion occurs in a manner similar to humans and other livestock animals. Many organs comprise a small proportion of the body as the animal matures, but the rumen increases from 30 to 70% of the capacity of the gut during weaning and continues to grow throughout lactation [43, 2]. However, at birth, a calf has only fully developed the abomasum. The fermentation vat including the reticulum and rumen is sterile upon birth and takes several weeks to establish a bacterial colony suitable for ruminant digestion. Therefore, new calves

are mostly fed with milk or a milk replacer that can use an esophageal groove to bypass the under-developed digestion system, although dry feed is required to develop rumen bacterial activity. The anaerobic nature of the under-developed rumen, when dry feed is added, makes a vessel suitable for anaerobic bacterial growth, producing acetic, propionic, and butyric acids, lowering the pH, of the rumen and continuing bacterial growth [3]. As the rumen develops post-natal, the bacterial colonies grow and genetic/genomic changes likely occur to aid in the transition from a milk-based, pre-weaning diet to the feed or grass diet after weaning. It has been shown that rumen development and rumen microbiomes are affected by the weaning process across different weaning strategies [30].

It has been well documented that gene expression changes have implications in rumen pH maintenance, gastrointestinal tract cell proliferation, growth, and development [4]. LncRNA has been shown to have regulatory roles in gene expression. In humans, approximately 5% of the genome is conserved solely due to genomic regions with regulatory roles, with 80% of these regions being associated with chromatin state, adding more evidence to the hypothesis that lncRNAs are regulatory elements [12]. Given the recent increase in interest in non-coding genomic regions, it would not be surprising to find connections between these non-coding RNAs and changes in gastrointestinal behaviors, such as weaning [18].

Previous research has shown that the weaning process does lead to changes in gene expression in rumen. Butyrate, a volatile fatty acid often produced by rumen microbes, can promote rumen development through assisting with structural integrity, epigenetic regulation, signaling pathways, and more [7]. As butyrate enters the rumen via milk or top-dressed feeds, 1,977 genes were identified as differentially expressed with or without butyrate treatment in rumen epithelial tissue — these genes included many key regulators and signaling pathways [25]. Weaning also led to variability in the chromatin structures and extracellular interactions in rumen tissues. The shift from milk to feed led to epige-

netic changes including DNA methylation, histone modification, and more [10]. Lin et al. provided increasing evidence that rumen development hinges on changes in genetic interactions, both expression-wise and extracellularly [25]. The previously identified 6,679 novel lncRNA found in rumen tissue gives additional support [29]. As lncRNAs are associated with pre-transcriptional regulation, transcriptional regulation, post-transcriptional regulation, as well as acting as signaling molecules, decoys, scaffolds, and guides, it is expected to find lncRNAs functioning in the rumen tissue [24]. The previous research also lends credence to the hypothesis that the two weaning conditional rumen samples will show differential expression levels in lncRNA.

Given the regulatory roles associated with lncRNA, analysis of lncRNA and their difference before and after weaning in rumen may reveal important regulatory pathways that allow a calf to be adequately weaned and begin life with a healthy digestive system. To further explore the features and potential functions of lncRNA in cattle, here we report a genome-wide study by isolating and investigating RNA transcripts that meet coding and length criteria for lncRNA classification, using rumen epithelial tissue in calves before and after weaning. With industrial tools, samples were sequenced and mapped to a robust genome before being filtered based on size, coding status, coding potential, and similarities to known genes (Supplemental Figure 1). The resulting transcript lists were then analyzed for differential expression, sequence conservation, multi-species orthologs, and enrichment with GWAS results of 42 dairy cattle traits that involve production, reproduction, body conformation, and health components. These findings shed light on the lncRNA landscape within cattle and their rumen, allowing us to further unravel the genetic mechanisms at work for tissue development and dairy cattle performance.

2.3 Results

2.3.1 lncRNA Identification

To identify lncRNA, Illumina HiSeq 2500 (PE150) sequencing was performed on three pre-weaning rumen samples and three post-weaning samples in Holstein cattle. We generated a total of over 125,000,000 sequence reads, averaging 21,085,595 reads per sample. Double-iterative mapping via Hisat yielded approximately 87.97% alignment in each sample [21]. Stringtie and Stringtie-merge yielded a consensus sequence for pre-weaning samples with 312,432 fragments and 315,559 fragments in the post-weaning samples [41, 33].

Comparing consensus sequences to the current ARS-UCD 1.2 *Bos taurus* reference genomes annotated transcripts based on known loci and genes [37]. Results from the comparisons allowed for filtering based on overlapping with known loci and transcripts in both reference genomes that resulted in 33,975 remaining transcripts in the pre-weaning samples and 35,621 transcripts in the post-weaning samples, respectively. These transcripts were then analyzed with a coding potential calculator, a BLAST search, and comparisons to the Pfam database to remove any remaining coding transcripts [1, 35, 17]. Finally, filtered results produced a list of 404 candidate lncRNA transcripts in the pre-weaning tissue and 234 transcripts in the post-weaning tissue (Fig 2.1).

2.3.2 lncRNA Characteristics with Comparison to Protein Coding Transcripts

Based on the ARS-UCD1.2 *Bos taurus* annotation [37], size of gene transcripts ranged from 200 to over a million nucleotides, averaging 28,239 nucleotides. This is in stark contrast to the length of lncRNA transcripts we identified in rumen, which averaged 674

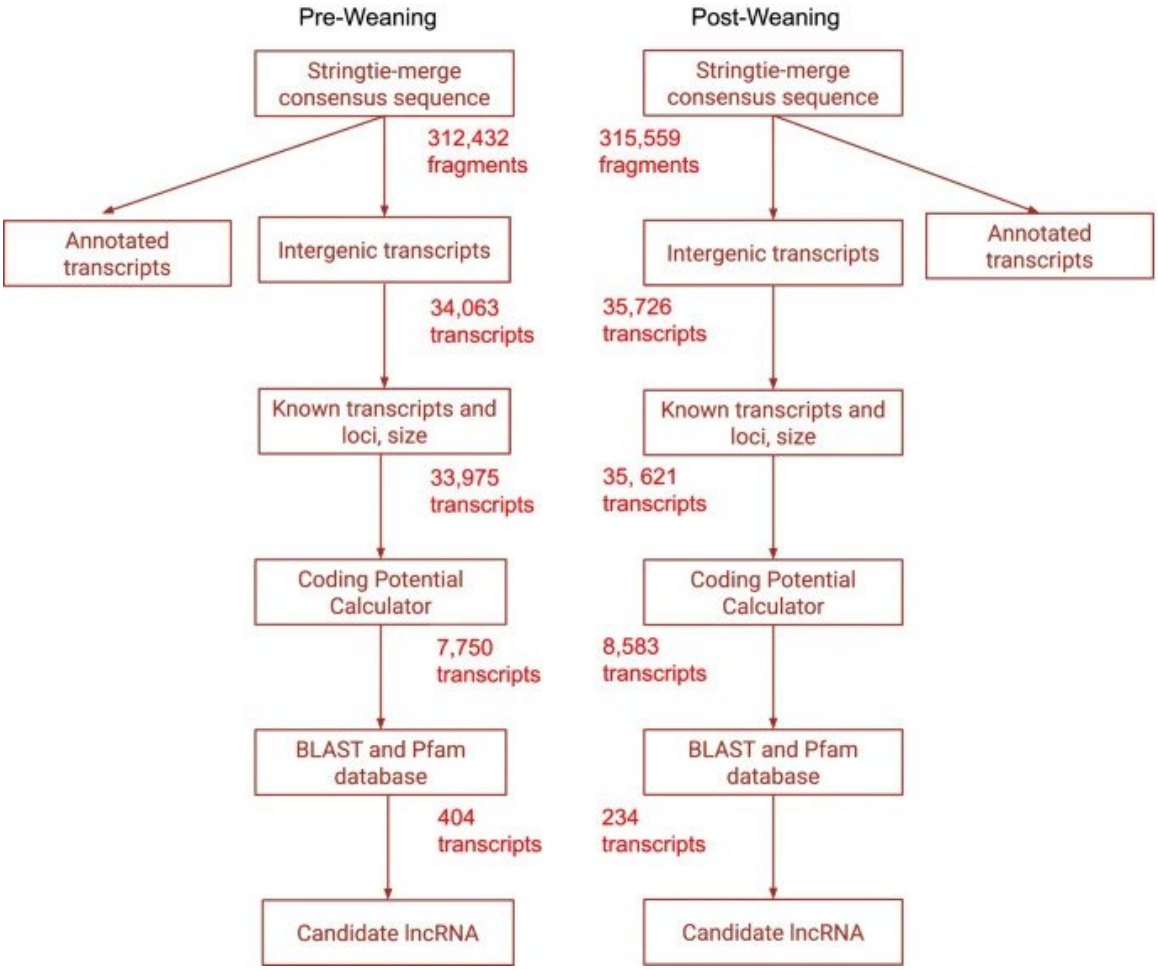

Figure 2.1: Filtering of Transcripts in Pre-weaning and Post-Weaning Rumen Tissue Samples. Once a consensus sequence was generated for each sample, a series of filtering steps were used to isolate candidate lncRNA. Steps included removing known protein coding transcripts, removing transcripts possessing coding potential, and those that demonstrated nucleotide and protein sequence homology. In the pre-weaning rumen tissue sample, 404 transcripts remained, and 234 transcripts remained in the post-weaning rumen tissue sample

nucleotides overall; pre-weaning lncRNAs averaged 466 nucleotides and post-weaning averaged 1,033 nucleotides (Fig 2.2). Although all identified lncRNAs were over 200 nucleotides in length, as required by the definition, they averaged 674 nucleotides and ranged from 200 to 18,000 base pairs in pre-weaning samples, and 200-57,000 base pairs in post-weaning samples (Fig 2.2). This is in line with previous findings of lncRNA transcripts being shorter than their coding counterparts. All identified lncRNA had, on average, 1.01 exons per transcript, whereas the NCBI gene annotation had 1-335 exons per transcript with an average of 9.2 exons per transcript. A distinctly lower number of exons per transcript in lncRNA compared to coding transcripts appears to be common across many studies [12, 6]. In pre-weaning cattle, expression level of coding genes averaged 26.12 fragments per kilobase of transcript per million mapped reads (FPKM). This is five times more than the lncRNA identified in pre-weaning samples, which measured an average of 5.24 FPKM. In post-weaning samples, lncRNA expression averaged an FPKM value of 7.89, compared to the 27.89 FPKM average value of all gene transcripts (Fig 2.3). The lower expression level of lncRNA compared to coding genes is also consistent with previous studies.

2.3.3 Differential Expression Analysis Before and After Weaning

The expression level of rumen lncRNA was similar before and after weaning with an average of 5.24 and 7.89 FPKM pre- and post-weaning, respectively. When analyzing the lncRNA profiles of pre- and post-weaning rumen tissue, the majority of lncRNAs identified were present only in one condition, with 395 found only in pre-weaning tissues and 225 only in post-weaning samples. Still, nine transcripts were present under both weaning conditions. However, none of these common transcripts showed expression levels that varied at a statistically significant level (Table 2.1).

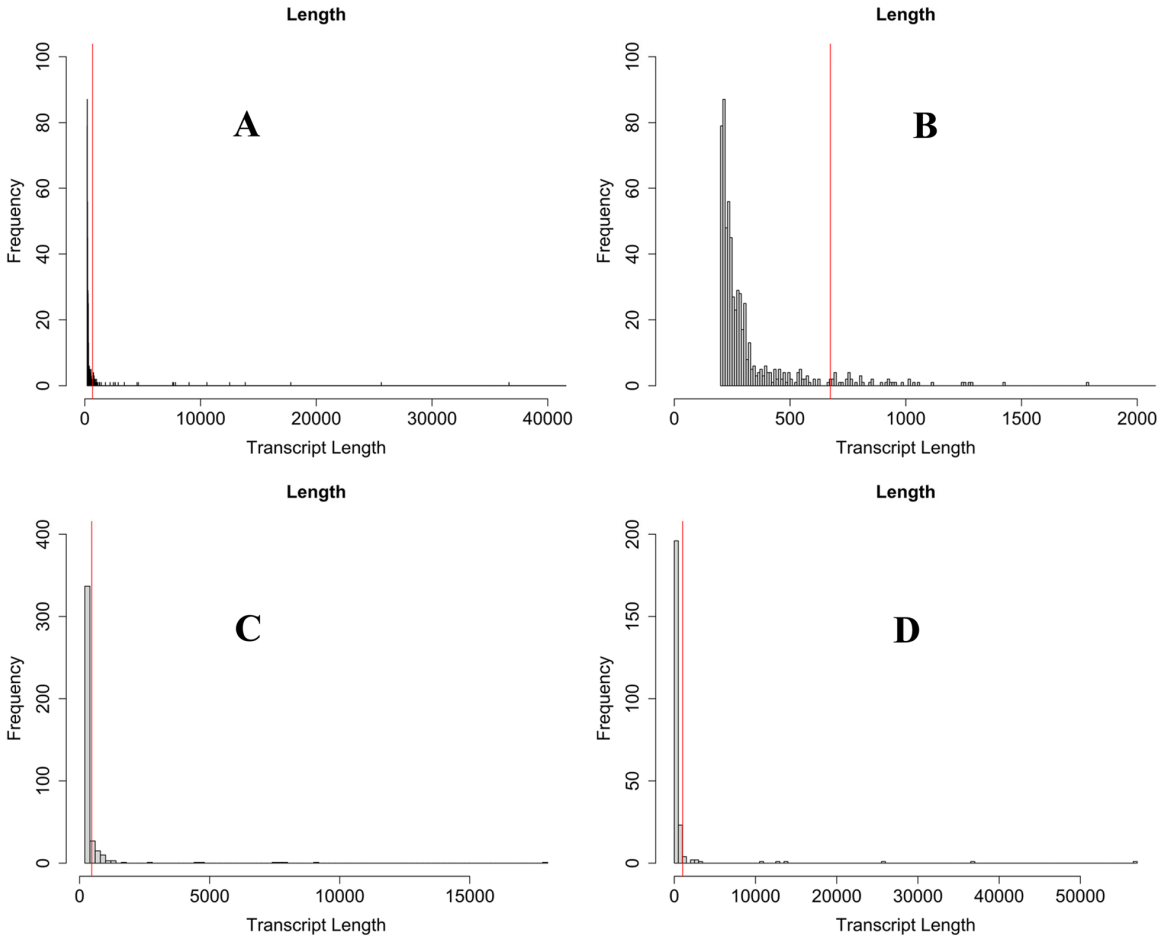

Figure 2.2: Length distribution of candidate lncRNA transcripts. **A** Length of all candidate lncRNA transcript. The average length of transcripts measured 674 base pairs, indicated by red line. **B** Zoomed in distribution of length of all candidate lncRNA transcript. Excluding those longer than 2000 base pairs for added clarity. **C** Length of pre-weaning transcripts, ranging from 200 to 17809 and averaging 466 nucleotides. **D** Length of post-weaning transcripts, ranging from 200 to 56626 and averaging 1033 nucleotides

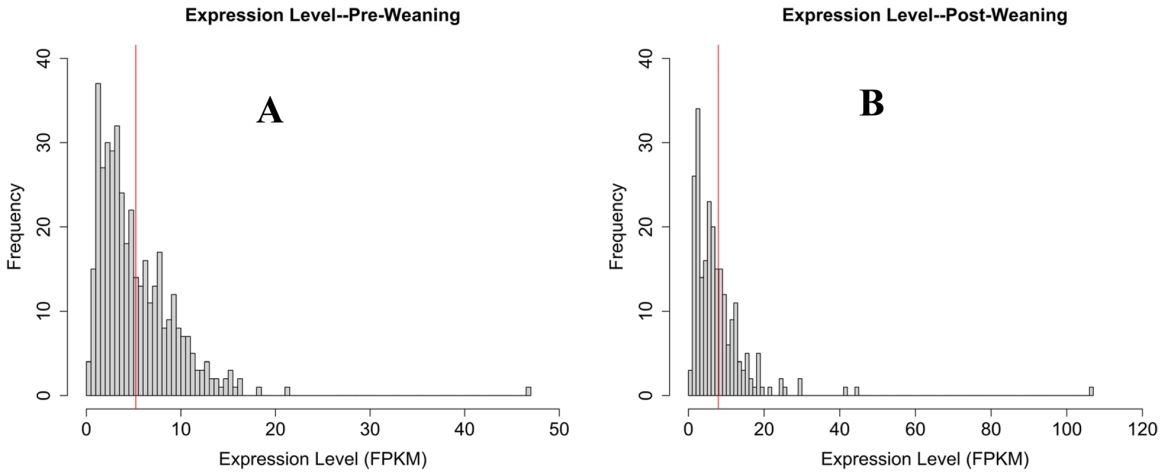

Figure 2.3: Expression of lncRNA candidate transcripts. **A** FPKM values of transcripts expressed in pre-weaning tissue. Expression levels ranged from 0.17 to 46.81 FPKM, averaging 5.24 FPKM. The average length of transcripts was indicated by red line. **B** FPKM values of transcripts expressed in post-weaning tissue. Expression levels ranged from 0.72 to 106 FPKM, averaging 7.89 FPKM. The average length of transcripts was indicated by red line

Table 2.1: T-test of expression levels between pre and post weaning rumen tissue for nine common lncRNA transcripts. Transcripts were isolated as common if they overlapped with each other, average expression was calculated for each transcript and is reported in FPKM. A paired student t-test was performed with degrees of freedom equaling 8. This yielded a critical value of 1.860, which none of the t-test scores surpassed, indicating none of the lncRNAs identified as common to both conditions were differentially expressed at a significant level

Pre-weaning		**Post-weaning**		***P*-value**
lncRNA	**Expression**	**lncRNA**	**Expression**	
Chr1:143606037–143,606,970	1.167	Chr1:143606047–143,607,059	2.963	0.743
Chr1:146868890–146,870,164	0.798	Chr1:146868985–146,869,483	2.004	0.798
Chr20:23927430–23,927,631	46.81	Chr20:23927420–23,927,884	106.4	0.584
Chr4:119469477–119,469,730	3.01	Chr4:119469677–119,470,016	1.776	0.514
Chr5:118054620–118,054,855	0.722	Chr5:118054288–118,054,860	1.607	0.654
Chr7:46217152–46,217,456	3.399	Chr7:46217122–46,217,472	5.292	0.515
Chr7:12950022–12,950,226	13.54	Chr7:12950002–12,950,242	11.11	0.464
Chr7:36273737–36,274,435	1.494	Chr7:36273748–36,274,421	1.348	0.471
Chr7:43944991–43,945,389	1.430	Chr7:43944829–43,945,374	1.757	0.516

2.3.4 Analysis of lncRNA Sequence Conservation

lncRNA tends to show lower rates of sequence conservation when compared to coding genes. To investigate this in cattle, transcript profiles for all coding transcripts, intergenic transcripts, and lncRNAs for both pre- and post-weaning conditions were converted to the human genome equivalents using the LiftOver v3.13 software and run through the PhastCons program to calculate conservation scores among transcripts of 46 vertebrates [40].

As expected, whole genome gene profiles showed higher scores overall than both intergenic and lncRNA profiles (Fig2.4). Interestingly, the median conversation score was slightly higher in lncRNA than intergenic transcripts. Although when plotted as a violin graph, the scores of lncRNA are localized near 0, intergenic regions are slightly more conserved. All transcripts averaged a PhastCons score of 0.103, intergenic regions scored an average of 0.104, and lncRNA had the lowest average score of 0.100. Complete profiles ranged in scores from 0.000111 to 0.999982, intergenic profiles ranged from 0 to 1, and lncRNA scores were as small as 0.000183 to as large as 0.998854. Between pre- and post-weaning, conservation scores were very similar (Fig 2.4). Pre-weaning samples ranged from 0.000873 to 0.879405, averaging 0.09963. Although it should be noted the median score of conserved pre-weaning transcripts was 0.0434. In post-weaning transcripts, scores ranged from 0.000183 to 0.98854 with an average score of 0.101 and a median score of 0.0485. LncRNA PhastCons scores were plotted based on score rank to illustrate clustering patterns (Fig 2.5). Most lncRNAs showed low sequence conservation scores with a small number showing much higher scores. These findings are consistent with lncRNA trends regarding conservation. The conservation score of the nine common transcripts were also calculated, which ranged from 0.008296 to 0.161162 with an average score of 0.0477 and a median of 0.0261.

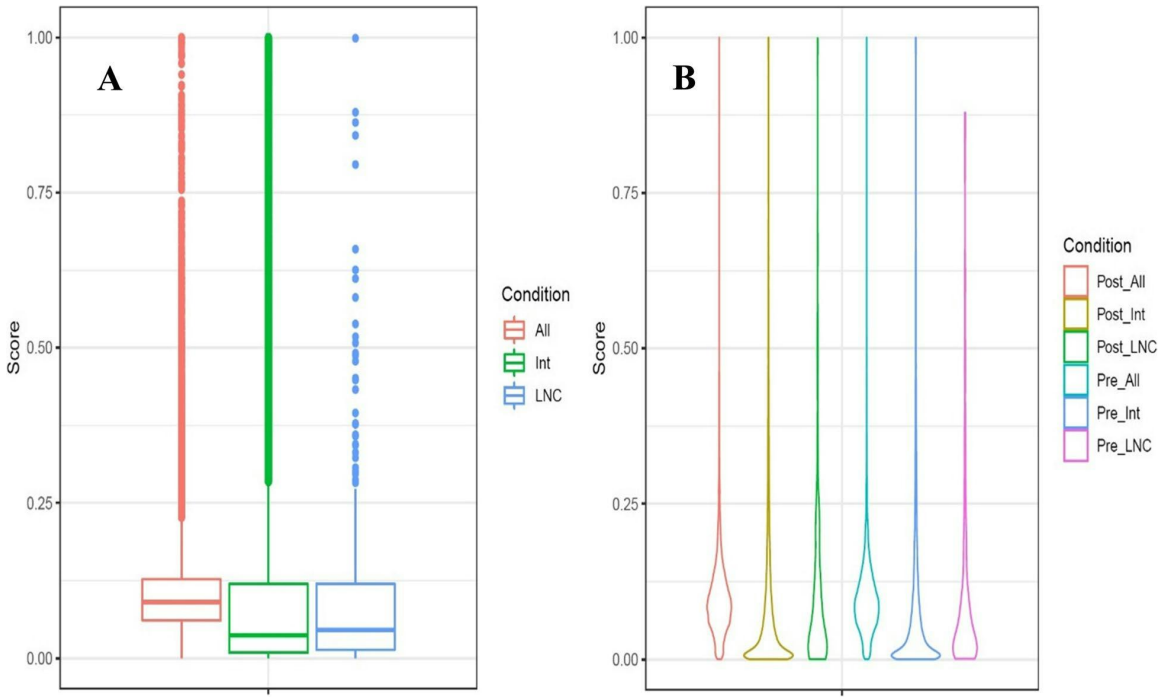

Figure 2.4: Phastcons scores of pre- and post-weaning conditions at whole genome, intergenic region, and lncRNA levels. **A** Boxplot of PhastCons scores for all transcripts, all intergenic transcripts, and all lncRNA candidate transcripts. **B** Violin plot of all six profiles: all preweaning transcripts, preweaning intergenic regions, preweaning lncRNA transcripts, all postweaning transcripts, postweaning intergenic regions, and postweaning lncRNA transcripts

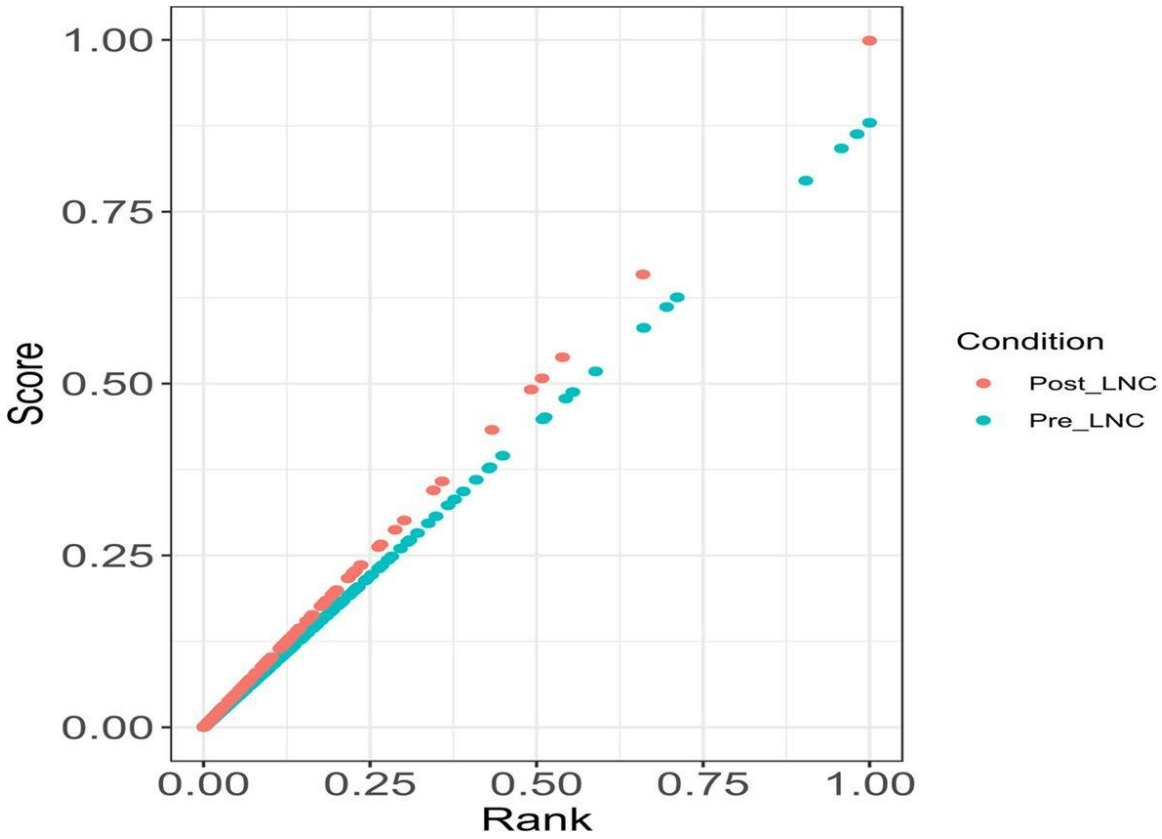

Figure 2.5: Scatter plot of lncRNA PhastCons scores. Most lncRNAs show scores well below 0.50 with a small number being well conserved across many species. Pre-weaning scores ranged from 0.000873 to 0.879405, and post-weaning scores ranged from 0.000183 to 0.658853, with an outlier of 0.98854

2.3.5 Transcriptional Annotation of Common and Conserved lncRNA

In addition to the nine common transcripts in two weaning conditions, the top 10% of lncRNAs based on conservation scores were also isolated from both weaning profiles for further analyses, in an attempt to ascertain lncRNA function. Twenty-three transcripts were kept from the pre-weaning profile, and 14 transcripts were kept for the post-weaning profile. Using the UCSC genome browser, these 46 transcripts were compared to the human and mouse genomes to identify previously annotated genes. Of the nine common lncRNA transcripts, three matched to human and mouse genes: ESYT2, SEC24A, and ZFN491/Zfp811. A fourth lncRNA matched to the Clic6 gene in Rattus norvegicus. Functions of these genes include ion channel formation, cellular lipid transport, transport vesicle formation, transcriptional activity, and more. Eleven of the 23 top conserved pre-weaning lncRNA showed matches to human and/or mouse genes. These include NLGN1, SAP130, ATP13A2/Padi3, BCL1, NTNG1, SLC2A1, APBB2, OAZ1, MEX3D, ANXA6, and DOCK11. Roles include nervous system development, cell surface protein interactions, transcriptional repressors, signal transduction, neuronal circuit formation, glucose transportation, post-transcriptional regulation, and much more. Of the post-weaning transcripts, nine of the 14 candidates matched to human/mouse genes: 4930448K20Rik, LAPTM4A, MBD5, IGSF21, F11R, EPS15, RAB21, COL25A1, UHRF1, and PLK5. Roles include nucleoside transport, heterochromatin binding, synapse inhibition, membrane traffic control, fibrillization inhibition, epigenetic regulation, cell cycle progression, and a number of other roles. Most notably, the F11R gene plays a role in epithelial tight junction formation; given the rumen is lined with stratified squamous epithelial cells, this has interesting implications in the relationship between lncRNA and rumen development [19]. Another interesting finding is the EPS15 gene, which has associations with cell growth regulation [38]. Given the physical changes taking place as the weaning process progresses, cell growth is a key com-

ponent of the rumen maturation. The large number of roles these equivalent genes play fit with previous research of the vast number of functions lncRNAs exhibit in the biological systems.

2.3.6 SNP heritability Enrichment Analysis on Cattle Traits

By integrating 404 pre- and 234 post-weaning lncRNAs with large-scale GWAS data in Holstein cattle [14, 9], we revealed an interesting relationship between rumen lncRNA and cattle complex traits of economic importance. Using a method for partitioning SNP heritability named MPH (https://github.com/jiang18/mph), we quantified SNP heritability enrichment as the ratio of per-SNP heritability near the lncRNAs to the genome-wide one. P-value was also computed by comparing the enrichment level to 1 using a Wald test. Finally, our enrichment analysis used 7,988 and 4,856 variants within the lncRNA identified for pre-weaning and post-weaning conditions, respectively.

As a result, we found a significant enrichment of per-SNP heritability near the lncRNAs under pre and post weaning conditions across cattle production, reproduction, health, and body conformation traits (Table 2.2). Overall, post-weaning lncRNAs showed slightly more significant enrichment with cattle traits than pre-weaning lncRNAs, indicating a more important function of the rumen lncRNAs after weaning for cow performance in the adult stage. Stature was highly significantly associated with lncRNAs in both pre- (11.67x; P=5.7E-6) and post-weaning (25.01x; P=1.1E-9) conditions, reflecting the important functions of rumen lncRNAs under both pre and post weaning conditions related to the overall tissue development and growth. Livability was associated with lncRNAs under both weaning conditions (11.34x; P=0.01 for pre and 15.1x; P=0.01 for post, respectively), suggesting that rumen development may have a functional role in the regulation of immunity and disease resistance. Milk's SNP heritability was only significantly enriched in post-weaning

lncRNAs (5.38x; P=0.04). And daughter pregnancy rate (DPR) was only significantly enriched with post-weaning lncRNAs (14.9; P=6.4E-4).

Table 2.2: Enrichment of rumen lncRNA in cattle GWAS results. Traits analyzed for SNP heritability enrichment included cattle production (milk), reproduction (daughter pregnancy rate, DPR), health (livability), and body conformation (stature). Enrichment was analyzed in both pre and post weaning tissue conditions

	DPR			**Livability**			**Milk**			**Stature**		
	Enr	SE	P	Enr	SE	P	Enr	SE	P	Enr	SE	P
Pre	3.49	1.76	0.08	3.76	2.55	0.14	11.67	2.43	5.7E-06	11.34	4.46	0.01
Post	5.38	2.52	0.04	14.9	4.31	6.4E-04	25.01	4.02	1.1E-09	15.07	6.23	0.01

2.4 Discussion

In this study, Illumina high throughput RNA-seq data were used to detect and analyze lncRNA in cattle rumen tissue before and after calf weaning. This was done to both identify rumen lncRNA and to find transcripts that are differentially expressed as the rumen develops from immature to mature. RNA-seq data were aligned to a robust reference genome and then progressively filtered by criteria such as coding potential, intergenic support, and size, resulting in a list of 629 lncRNA transcripts. Candidate transcripts averaged a shorter length, less exons per transcript, lower expression, and lower conservation scores when compared to whole genomic transcripts. This represents 404 transcripts in the pre-weaning profile, 234 transcripts in the post-weaning profile, and 9 transcripts common to both profiles. Interestingly, the 9 common transcripts are expressed at similar levels, indicating they likely play a basal role in rumen tissue that is independent of rumen tissue development.

This study confirms that there are lncRNA transcripts expressed in rumen tissue, and furthermore that transcripts are expressed differently as weaning progresses. Identification of hundreds of transcripts unique to before and after weaning conditions suggests the expression of these transcripts is related to the maturation of the animal, likely being tied to the weaning process. Overall, the variability in identified lncRNA transcripts indicates they likely play a role in the biological changes that occur as the calf's digestive system develops. This also supports theories that although these transcripts are not made directly into gene product, they are essential for the growth and development of organisms. Enrichment analysis demonstrated that, although isolated from rumen tissue, these lncRNA transcripts likely influence traits outside of those associated with digestion/weaning, as several of the transcripts show significant enrichment based on genome wide association studies. A subset of the unique transcripts to each tissue condition showing enrichment in complex traits could mean many things, however, it is evidence that these transcripts are involved in the

development of other characteristics as the calf is weaned. This could demonstrate that several transcripts may have broad applications in calf development and later performance overall.

When integrating the identified rumen lncRNA with existing GWAS results, we reported significant enrichments of GWAS signals in or near lncRNA regions across a wide spectrum of cattle traits, suggesting the importance of lncRNA on the genetics of complex traits. Considering the functional impact of lncRNA on gene regulation, future studies should no longer ignore lncRNA in GWAS and fine-mapping studies. In addition, genomic selection may also consider adding important SNPs near lncRNA onto newer SNP arrays and using different weights in the modelling process.

Previous studies have identified at least 7,000 lncRNA within cattle, as well as 20,000+in other systems, and this study was able to identify over 600 transcripts that met lncRNA criteria when analyzing rumen tissue from *Bos taurus* [20, 28, 16]. Compared to other studies of cattle lncRNA, the current study is limited by only focusing on the rumen tissue and two weaning conditions. With only three replicates within each condition, we may not be able to discover all the important lncRNAs in rumen due to limited detection power. Still, we emphasize the validity and importance of our results based on the evidence of consistent genomic features of the identified lncRNAs with existing studies, higher conservation across species than genome average, and enrichment with GWAS signals of important dairy cattle traits.

Further analyses of multiple tissues may also expand our findings and identify lncRNAs that are expressed exclusively in other tissues, as well as reveal those expressed in multiple tissues, including the rumen. Research has also shown lncRNA has connections with many biological processes such as chromosome X inactivation, allelic imprinting, pluripotency control, and cancer, so it is not surprising to find that lncRNA identified in *Bos taurus* rumen may be involved in development of the rumen, and likely across the organism [23].

This is supported by our findings that there are very different lncRNA profiles as a calf is weaning. Of both common and highly conserved lncRNAs, the corresponding genes in human and mouse models are indicative that these lncRNAs serve a number of roles related to rumen. Two of the identified genes are of particular interest: F11R and EPS15. F11R is related to the formation of the epithelial tight junctions, which is an interesting research endeavor given the rumen is lined with epithelial cells. The role of EPS15 in cell growth regulation may also be particularly telling about the morphological changes that occur in the rumen as the calf is weaned. Although several genes show sequence similarities to known genes in well researched organisms, further research may shed light on the exact roles of these transcripts in either individual tissue or in a systematic setting.

2.5 Conclusions

In this study, we investigated lncRNA features in cattle rumen tissue before and after weaning. As expected, cattle lncRNAs exhibited similar features as lncRNAs identified in other species. The majority of rumen lncRNAs are uniquely expressed in either the pre-weaning or post-weaning condition, indicating a functional role of lncRNAs in rumen development and transition of weaning. Overlapping with large-scale GWAS results in cattle, both pre- and post-weaning lncRNAs showed significant enrichment with all types of dairy traits of economic importance, suggesting the impact of rumen lncRNAs for cattle performance in the adult stage. The findings of this study allow for further research to be conducted to identify lncRNA in other tissues, investigate roles of candidate transcripts in traits of interest, and delve into lncRNA expression and developmental roles across an entire organism.

2.6 Materials and Methods

2.6.1 Animals and Tissue Collection

The animal handling procedures were approved by the Beltsville Area Animal Care and Use Committee Protocol Number 07–025. Six Holstein bull calves were used for the study, which were purchased from a local farm and transported by 3 days of age to the Beltsville Agricultural Research Center (BARC), Beltsville, MD, USA. All calves received colostrum at the first feeding and received standard diet afterward. We randomly selected three calves to be euthanized at 14 days of age for rumen epithelium collection to exemplify preweaning, and at 70 days of age, three calves were euthanized for the collection of rumen tissue postweaning, as previous research has demonstrated that a sample size of three is sufficient to find different gene expression related to organ development in dairy cattle [5].

The rumen tissues were collected from the ventral sac beneath the cranial pillar. Distinct visual cues present ensured the rumen tissue was predominately exposed to rumen liquor before slaughter. The collected tissues were rinsed with tap water and then rinsed in ice-cold physiological saline. According to manufacturer instructions, subsamples (.600 mg) of epithelial tissue were fixed in RNAlater RNA stabilization solution (Life Technologies, Grand Island, NY, USA) and stored at -.80° C until RNA extraction.

2.6.2 RNA sequencing, Transcriptional Mapping and Assembly

The RNA extraction procedure has been reported previously [44]. Briefly, total RNA was extracted by Trizol (Invitrogen, Carlsbad, CA, USA) followed by DNase digestion and Qiagen RNeasy column purification (Qiagen, Valencia, CA, USA). The RNA integrity was verified by an Agilent Bioanalyzer 2100 (Agilent, Palo Alto, CA, USA). After quality control procedures, total RNA was sequenced at 150-bp/paired-end sequence reads using

RNA-sequencing service supplied by Novogene Corporation, Inc. (UC Davis Sequencing Center, Davis, CA, USA).

RNA sequence data of each sample were mapped to the ARS-UCD1.2 reference genome [37] using the spliced read aligner Hisat 2.2.1. A two-iteration approach was taken to capitalize on exon junction information [21]. The first iteration mapped samples to the reference using default settings. The second iteration leveraged junction findings from the first run to re-map reads to the ARS-UCD1.2 reference. Stringtie v2.1.4 was then used to assemble accepted hits into reports for each sample [33]. The ARS-UCD1.2 *Bos taurus* annotation was provided with the '–G' option for more accurate assembly. All pre-weaning samples were merged using the 'stringtie—merge' function. The ARS-UCD1.2 annotation was provided again with the '–G' option for accuracy. The above was performed again for the post-weaning samples. This provided a consensus sequence for both conditions for further analyses [32, 37].

2.6.3 lncRNA Identification

Identification of long non-coding RNAs (lncRNAs) started by using the CuffCompare 2.2.1 program. The ARS-UCD1.2 annotation fasta files were provided to improve classifications using the '–s' option, and the ARS-UCD1.2 annotation .gtf files were provided as well via the '–r' option to annotate any possible matching reference transcripts [41]. Any transcripts marked as overlapping with coding transcripts were removed to retain only intergenic transcripts. Remaining transcripts were inputted into the coding potential calculator (CPC), a software that scores reads based on open reading frame and homology analysis. Coding potential scores range from -1 to 1, where a score below 0 indicates the transcript is likely non-coding [35]. All transcripts that scored above 0 were removed. Those transcripts that remain were submitted through the CPC again to analyze the re-

verse compliment; again, any read scoring above 0 was removed. Next, transcripts were ran through the BLAST database to compare transcripts to known coding genes, removing any transcripts that were protein-coding but unannotated [1]. Transcripts were kept if the BLAST results showed no matching transcript. In order to remove any remaining non-lncRNA transcripts, reads were converted from bases to amino acids in all three reading frames before being compared to the Pfam database [35]. This removed any transcripts that showed documented protein domains. Reads that were less than 200 bp were also excluded, as they do not meet the lncRNA criteria. This produced a final list of transcripts for pre-weaning and post-weaning tissues that (a) did not overlap a known gene based on the NCBI and ARS-UCD1.2 reference genome, (b) were longer than 200 bp in length, (c) lacked coding potential, (d) did not show sequence similarity to any known gene, and (e) did not show any amino acid pattern resembling a known protein domain.

2.6.4 Comparison to Coding Transcripts

Lengths were calculated for all transcripts in both the lncRNA and genome wide profiles under two weaning conditions. Profiles were compared to determine different patterns between subsets. Combined lncRNA and gene profiles were graphed as histograms to demonstrate contrast between profiles. The lncRNA length profiles were also graphed as a histogram with the average length calculated and plotted on the histogram using R 4.12 [36]. Additionally, a truncated graph was generated for clarity with average length denoted. Using the Stringtie v2.1.4 software, expression levels for all transcripts in all samples were estimated. Stringtie v2.1.4 does so by clustering reads, creating a splice graph for each cluster that is associated with an identified transcript, creating a flow network for each transcript, then estimating expression level with a maximum flow algorithm [32]. Expression was reported with the fragments per kilobase of transcript per million mapped reads

(FPKM) value. FPKM data for all reads were filtered into a subset containing only those reads in the final lncRNA list identified previously. Pre- and post-weaning expression levels were plotted in a histogram with the average value superimposed using R. Average expression of all transcripts was also calculated using the average command in Excel.

2.6.5 PhastCons Analysis

PhastCons analysis was performed on lncRNA transcripts, full transcript profiles, and intergenic transcript profiles in both weaning conditions. This was done by using the UCSC LiftOver web tool to convert profiles from the ARS-UCD1.2 cattle genome to the hg38 human genome. The PhastCons command was then used on the open source galaxy server to compare genomic regions across 46 vertebrate species and the placental mammal subset of the vertebrate species to determine conservation scores (species include human, chimp, gorilla, orangutan, rhesus, baboon, marmoset, tarsier, mouse lemur, bushbaby, tree shrew, mouse, rat, kangaroo rat, guinea pig, ground squirrel, rabbit, pika, alpaca, dolphin, cow, horse, cat, dog, microbat, megabat, hedgehog, shrew, elephant, rock hyrax, tenrec, armadillo, sloth, wallaby, opossum, platypus, chicken, zebra finch, green anole, western clawed frog, tetraodon, fugu, stickleback, medaka, zebrafish, and lamprey). Each set was then graphed for comparisons. The top 10% of lncRNA transcripts based on PhastCons conservation score were extracted for further analyses.

2.6.6 Transcriptional Annotation

Candidate lncRNA transcripts were selected by either being present in both tissue conditions or by representing the top 10% of conserved sequences. The top 10% cutoff was selected as there was a clear separation of Phastcons scores between the top 10% and other transcripts. These regions were then searched in the UCSC genome browser to identify

genes present in other species, primarily noting those found in humans and mice, as they represent well researched species. One transcript showed no matches to mice or humans but did match to a gene in Rattus norvegicus, and this was noted as well. Genes identified were then searched in UniProt to identify gene function.

2.6.7 Heritability Enrichment Analysis

Heritability enrichment analysis was performed on two sets of lncRNA: those only present in pre-weaning and those in post-weaning rumen samples. GWAS data of 27 K bulls and 3 M SNPs/INDELspreviously reported [15, 9] were used for the enrichment analysis, representing cattle production (milk), reproduction (daughter pregnancy rate), health (livability), and body conformation (stature) phenotypes. The GWAS was performed in 27,214 Holstein bulls that had imputed 3 million SNPs and highly reliable predicted transmitting abilities (PTAs) for production, reproduction, health and body type traits.

We used a newly developed method, MPH, for the enrichment analysis (Github).
In this analysis, MPH fits the following linear mixed model with two genetic components:

$$y = Xb + g_1 + g_2 + e$$

$$g_1 \sim N(0, G_1\sigma_g^2 1)$$

$$g_2 \sim N(0, G_2\sigma_g^2 2)$$

$$e \sim N(0, R\sigma_e^2)$$

where b is a vector of fixed effects including intercept, X is a covariate matrix corresponding to b, g_1 and g_2 are two vectors of random effects representing the genetic contributions of the variants within lncRNAs and the remaining ones on the genome, respectively, and e is

a vector of residual effects. G_1 and G_2 are two genomic relationship matrices constructed by genotypes corresponding to g_1 (lncRNA) and g_2 (remaining), respectively. We computed G using the second method as described in [42], $G = (ZZ^T)/m$, where Z represents standardized genotypes, and m is the number of variants in the GRM. We used m_1 and m_2 to represent the number of variants within lncRNAs and the number of the remaining ones on the genome, respectively. R is a diagonal matrix used to model individual reliability of deregressed PTAs. MPH uses a Monte Carlo REML algorithm similar to [26] to estimate variance components (VCs) and computes the variance-covariance matrix of VC estimates based on the Fisher information matrix.

The per-SNP genetic variance in lncRNA is equal to $\frac{\sigma^2_{g1}}{m_1}$, and its per-SNP heritability enrichment is then computed by $\rho = \frac{\frac{\sigma^2_{g1}}{m_1}}{\frac{(\sigma^2_{g1}+\sigma^2_{g2})}{(m_1+m_2)}}$. We used the delta method to compute the standard error of ρ.

A total of 404 pre- and 234 post-weaning lncRNAs with 10 Kb window extension on both sides were used to test their enrichment of per-SNP heritability for four traits. We used a 10Kb window extension to cover linked regions considering the SNP density of our GWAS data and the linkage disequilibrium level in the Holstein cattle population [8, 39, 27]. The enrichment analysis included a total of 2,803,423 autosomal variants after QC with a minimal MAF of 0.005 and a minimal HWE p-value of 1E-8. We used 7,988 and 4,856 variants within the lncRNA identified for pre-weaning and post-weaning conditions, respectively. This analysis is achieved by partitioning SNP heritability with two genomic relationship matrices (GRMs): one made using SNPs/INDELs within lncRNAs and the other made using remaining ones. SNP heritability enrichment is quantified as the ratio of per-SNP heritability near lncRNAs to the genome-wide one, whose estimate and standard error are computed by MPH. Wald tests were used to compute P-value comparing an enrichment level to 1 (i.e., the genome-wide per-SNP heritability level).

2.7 References

[1] Stephen F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 1990), pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

[2] R. L. Baldwin VI, K. R. McLeod, and A. V. Capuco. “Visceral tissue growth and proliferation during the bovine lactation cycle”. In: *Journal of dairy science* 87.9 (2004), pp. 2977–2986. ISSN: 0022-0302. DOI: 10.3168/JDS.S0022-0302(04)73429-3. URL: https://pubmed.ncbi.nlm.nih.gov/15375059/.

[3] T. Bauchop. “Rumen Anaerobic Fungi of Cattle and Sheep”. In: *Applied and Environmental Microbiology* 38.1 (1979), pp. 148–158. ISSN: 00992240. DOI: 10.1128/AEM.38.1.148-158.1979. URL: https://journals.asm.org/doi/10.1128/aem.38.1.148-158.1979.

[4] E. E. Connor et al. “Gene expression in the digestive tissues of ruminants and their relationships with feeding and digestive processes”. In: *Animal* 4.7 (2010), pp. 993–1007. ISSN: 17517311. DOI: 10.1017/S1751731109991285. URL: https://www-cambridge-org.proxy-um.researchport.umd.edu/core/journals/animal/article/gene-expression-in-the-digestive-tissues-of-ruminants-and-their-relationships-with-feeding-and-digestive-processes/44979423689DCA3D3AEE9BD878021E37.

[5] Erin E. Connor et al. “Effects of increased milking frequency on gene expression in the bovine mammary gland”. In: *BMC Genomics* 9.1 (July 2008), pp. 1–14. ISSN: 14712164. DOI: 10.1186/1471-2164-9-362.

[6] Thomas Derrien et al. “The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression”. In: *Genome Research*

22.9 (Sept. 2012), pp. 1775–1789. ISSN: 10889051. DOI: 10.1101/gr.132159.111. URL: www.genome.org.

[7] Qiyu Diao, Rong Zhang, and Tong Fu. "Review of strategies to promote rumen development in calves". In: *Animals* 9.8 (Aug. 2019), p. 490. ISSN: 20762615. DOI: 10.3390/ani9080490.

[8] Frédéric Farnir et al. "Extensive genome-wide linkage disequilibrium in cattle". In: *Genome Res* 10.2 (Feb. 2000), pp. 220–227. ISSN: 10889051. DOI: 10.1101/gr.10.2.220.

[9] Ellen Freebern et al. "GWAS and fine-mapping of livability and six disease traits in Holstein cattle". In: *BMC Genomics 2020 21:1* 21.1 (Jan. 2020), pp. 1–11. ISSN: 1471-2164. DOI: 10.1186/S12864-020-6461-Z. URL: https://link.springer.com/articles/10.1186/s12864-020-6461-z%20https://link.springer.com/article/10.1186/s12864-020-6461-z.

[10] Yahui Gao et al. "Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning". In: *Genomics* 113.4 (July 2021), pp. 2045–2055. ISSN: 0888-7543. DOI: 10.1016/J.YGENO.2021.04.039.

[11] Yuan Gao et al. "Analysis of Long Non-Coding RNA and mRNA Expression Profiling in Immature and Mature Bovine (Bos taurus) Testes". In: *Frontiers in Genetics* 0.JUL (2019), p. 646. ISSN: 1664-8021. DOI: 10.3389/FGENE.2019.00646.

[12] Yanghua He et al. "The conservation and signatures of lincRNAs in Marek's disease of chicken". In: *Scientific Reports* 5.1 (Oct. 2015), pp. 1–17. ISSN: 20452322. DOI: 10.1038/srep15184. URL: www.nature.com/scientificreports.

[13] Xianbo Jia et al. "Genome-wide identification and characterisation of long non-coding RNAs in two Chinese cattle breeds". In: *https://doi.org/10.1080/1828051X.2020.1735266*

19.1 (Dec. 2020), pp. 383–391. DOI: 10.1080/1828051X.2020.1735266. URL: https://www.tandfonline.com/doi/abs/10.1080/1828051X.2020.1735266.

[14] Jicai Jiang et al. "Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls". In: *Communications Biology 2019 2:1* 2.1 (June 2019), pp. 1–12. ISSN: 2399-3642. DOI: 10.1038/s42003-019-0454-y. URL: https://www.nature.com/articles/s42003-019-0454-y.

[15] Rui Jiang et al. "Journal Pre-proofs Transcriptome Profiling of lncRNA Related to Fat Tissues of Qinchuan Cattle". In: (2020). DOI: 10.1016/j.gene.2020.144587. URL: https://doi.org/10.1016/j.gene.2020.144587.

[16] Laurence Josset et al. "Annotation of long non-coding RNAs expressed in Collaborative Cross founder mice in response to respiratory virus infection reveals a new class of interferon-stimulated transcripts". In: *RNA Biology* 11.7 (July 2014), p. 875. DOI: 10.4161/RNA.29442. URL: /pmc/articles/PMC4179962/%20/pmc/articles/PMC4179962/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4179962/.

[17] Yu-Jian Kang et al. "CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features". In: *Nucleic Acids Research* 45.W1 (July 2017), W12–W16. ISSN: 0305-1048. DOI: 10.1093/NAR/GKX428. URL: https://academic.oup.com/nar/article/45/W1/W12/3831091.

[18] Philipp Kapranov, Aarron T. Willingham, and Thomas R. Gingeras. "Genome-wide transcription and the implications for genomic organization". In: *Nat Rev Genet* 8.6 (June 2007), pp. 413–423. ISSN: 14710056. DOI: 10.1038/nrg2083.

[19] R. N. B. Kay, W. v. Engelhardt, and R. G. White. “The digestive physiology of wild ruminants”. In: *Digestive Physiology and Metabolism in Ruminants* (1980), pp. 743–761. DOI: 10.1007/978-94-011-8067-2{_}36. URL: https://link.springer.com/chapter/10.1007/978-94-011-8067-2_36.

[20] Colin Kern et al. “Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species”. In: *BMC Genomics 2018 19:1* 19.1 (Sept. 2018), pp. 1–14. ISSN: 1471-2164. DOI: 10.1186/S12864-018-5037-7. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5037-7.

[21] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. “HISAT: a fast spliced aligner with low memory requirements”. In: *Nature Methods 2015 12:4* 12.4 (Mar. 2015), pp. 357–360. ISSN: 1548-7105. DOI: 10.1038/nmeth.3317. URL: https://www.nature.com/articles/nmeth.3317.

[22] Lambros T. Koufariotis et al. “A catalogue of novel bovine long noncoding RNA across 18 tissues”. In: *PLoS ONE* 10.10 (Oct. 2015). DOI: 10.1371/JOURNAL.PONE.0141225. URL: /pmc/articles/PMC4619662/%20/pmc/articles/PMC4619662/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4619662/.

[23] Johnny T.Y. Kung, David Colognori, and Jeannie T. Lee. *Long noncoding RNAs: Past, present, and future*. 2013. DOI: 10.1534/genetics.112.146704. URL: /pmc/articles/PMC3583990/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583990/.

[24] Zhijin Li et al. “The Role of Long Noncoding RNAs in Gene Expression Regulation”. In: *Gene Expression Profiling in Cancer*. IntechOpen, June 2019. DOI: 10.5772/intechopen.81773. URL: www.intechopen.com.

[25] Shudai Lin et al. “Establishment and transcriptomic analyses of a cattle rumen epithelial primary cells (REPC) culture by bulk and single-cell RNA sequencing to elucidate interactions of butyrate and rumen development”. In: *Heliyon* 6.6 (June 2020), e04112. ISSN: 24058440. DOI: 10.1016/j.heliyon.2020.e04112.

[26] Po Ru Loh et al. “Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis”. In: *Nat Genet* 47.12 (Dec. 2015), pp. 1385–1392. ISSN: 15461718. DOI: 10.1038/ng.3431.

[27] Duc Lu et al. “Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle”. In: *Front Genet* 3.AUG (2012), p. 152. ISSN: 16648021. DOI: 10.3389/fgene.2012.00152.

[28] Lina Ma et al. “Lncbook: A curated knowledgebase of human long non-coding rnas”. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D128–D134. ISSN: 13624962. DOI: 10.1093/nar/gky960. URL: /pmc/articles/PMC6323930/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323930/.

[29] Bizhan Mahmoudi et al. “Genome-wide identification and characterization of novel long non-coding RNA in Ruminal tissue affected with sub-acute Ruminal acidosis from Holstein cattle”. In: (). DOI: 10.1007/s11259-020-09769-w. URL: https://doi.org/10.1007/s11259-020-09769-w.

[30] Sarah J. Meale et al. “Development of ruminal and fecal microbiomes are affected by weaning but not weaning strategy in dairy calves”. In: *Frontiers in Microbiology* 7.MAY (May 2016), p. 182975. ISSN: 1664302X. DOI: 10.3389/FMICB.2016.00582/BIBTEX.

[31] Wietje Nolte et al. “Biological Network Approach for the Identification of Regulatory Long Non-Coding RNAs Associated With Metabolic Efficiency in Cattle”. In:

Frontiers in Genetics 0 (Nov. 2019), p. 1130. ISSN: 1664-8021. DOI: 10.3389/FGENE.2019.01130.

[32] Mihaela Pertea and Geo Pertea. "GFF Utilities: GffRead and GffCompare". In: *F1000Research* 9 (2020). ISSN: 1759796X. DOI: 10.12688/f1000research.23297.1.

[33] Mihaela Pertea et al. "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads". In: *Nature Biotechnology* 33.3 (2015), pp. 290–295. ISSN: 15461696. DOI: 10.1038/nbt.3122. URL: /pmc/articles/PMC4643835/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4643835/.

[34] Amaral PP et al. "lncRNAdb: a reference database for long noncoding RNAs". In: *Nucleic acids research* 39.Database issue (Jan. 2011). ISSN: 1362-4962. DOI: 10.1093/NAR/GKQ1138. URL: https://pubmed.ncbi.nlm.nih.gov/21112873/.

[35] Finn RD et al. "The Pfam protein families database". In: *Nucleic acids research* 38.Database issue (Nov. 2010). ISSN: 1362-4962. DOI: 10.1093/NAR/GKP985. URL: https://pubmed.ncbi.nlm.nih.gov/19920124/.

[36] Brian D Ripley. "The R project in statistical computing". In: *MSOR Connections The newsletter of the LTSN Maths, Stats & OR Network* 1.1 (Feb. 2001), pp. 23–25. ISSN: 1473-4869. DOI: 10.11120/msor.2001.01010023.

[37] Benjamin D Rosen et al. "De novo assembly of the cattle reference genome with single-molecule sequencing". In: *GigaScience* 9.3 (Mar. 2020), pp. 1–9. DOI: 10.1093/GIGASCIENCE/GIAA021. URL: https://academic.oup.com/gigascience/article/9/3/giaa021/5810242.

[38] Ingrid Roxrud et al. “An endosomally localized isoform of Eps15 interacts with Hrs to mediate degradation of epidermal growth factor receptor”. In: *Journal of Cell Biology* 180.6 (Mar. 2008), pp. 1205–1218. ISSN: 00219525. DOI: 10.1083/JCB.200708115.

[39] M. Sargolzaei et al. “Extent of linkage disequilibrium in Holstein cattle in North America”. In: *J Dairy Sci* 91.5 (2008), pp. 2106–2117. ISSN: 15253198. DOI: 10.3168/jds.2007-0553.

[40] Adam Siepel et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome Res* 15.8 (Aug. 2005), pp. 1034–1050. ISSN: 10889051. DOI: 10.1101/gr.3715005.

[41] Cole Trapnell et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. In: *Nature Protocols* 7.3 (Mar. 2012), p. 562. DOI: 10.1038/NPROT.2012.016. URL: /pmc/articles/PMC3334321/%20/pmc/articles/PMC3334321/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334321/.

[42] P. M. VanRaden. “Efficient methods to compute genomic predictions”. In: *J Dairy Sci* 91.11 (2008), pp. 4414–4423. ISSN: 15253198. DOI: 10.3168/jds.2007-0980.

[43] R. G. Warner, W. P. Flatt, and J. K. Loosli. “Dietary factors influencing the development of the ruminant stomach”. In: *Journal of Agricultural and Food Chemistry* 4.9 (Aug. 1956), pp. 788–792. ISSN: 15205118. DOI: 10.1021/JF60067A003/ASSET/JF60067A003.FP.PNG{_}V03. URL: https://pubs.acs.org/doi/abs/10.1021/jf60067a003.

[44] Sitao Wu et al. “Transcriptome characterization by RNA-seq unravels the mechanisms of butyrate-induced epigenomic regulation in bovine cells”. In: *PloS one* 7.5 (May 2012), e36940. ISSN: 19326203. DOI: 10.1371/journal.pone.0036940.

Chapter 3: Investigation of lncRNA in *Bos taurus* Mammary Tissue During Dry and Lactation Periods

3.1 Abstract

This study aims to collect RNA-Seq data from *Bos taurus* samples representing dry and lactating mammary tissue, identify lncRNA transcripts, and analyze findings for their features and functional annotation. This allows for connections to be drawn between lncRNA and the lactation process. RNA-Seq data from 103 samples of *Bos taurus* mammary tissue were gathered from publicly available databases (60 dry, 43 lactating). The samples were filtered to reveal 214 dry mammary lncRNA transcripts and 517 lactating mammary lncRNA transcripts. The lncRNAs met common lncRNA characteristics such as shorter length, fewer exons, lower expression levels, and less sequence conservation when compared to the genome. Interestingly, several lncRNAs showed sequence similarity to genes associated with strong hair keratin intermediate filaments. Human breast cancer research has associated strong hair keratin filaments with mammary tissue cellular resilience. The lncRNAs were also associated with several genes/proteins that linked to pregnancy using expression correlation and gene ontology. Such findings indicate that there are crucial relationships between the lncRNAs found in mammary tissue and the development of the tissue, to meet both the animal's needs and our own production needs; these relationships should be further investigated to ensure that we continue to breed the most resilient, efficient dairy cattle.

Note:

This chapter was previously published by Genes. Alexis Marceau was the main contributor and is first author. The original citation is as follows:

Marceau A, Wang J, Iqbal V, Jiang J, Liu GE, Ma L. Investigation of lncRNA in Bos taurus Mammary Tissue during Dry and Lactation Periods. Genes. 2023; 14(9):1789. https://doi.org/10.3390/genes14091789

3.2 Background

Although many research endeavors have placed emphasis on direct gene products, sequencing technology advancements have led to a shift in focus. Researchers have now taken on investigating the regions of the genome that are not made into proteins: non-coding RNA. Non-coding RNA has been established as a gene expression regulator in many roles, including (but not limited to) X chromosome inactivation, allelic imprinting, pluripotency control, and cancer [26]. An important group within the non-coding RNA family is the long noncoding RNA (lncRNA): non-coding transcripts that are longer than 200 nucleotides. Research has shown that these transcripts are far-reaching and common, with more than 270,000 lncRNAs identified in humans, approximately 21,000 lncRNAs found in mice, and over 7235 lncRNAs present in cattle [20, 22, 29]. Not only are lncRNAs found in many organisms, but they are also found in variable numbers in subspecies and breeds. In previous studies within two Chinese cattle breeds (and across six tissues), over 4000 lncRNAs were identified; an additional study utilizing 18 different bovine tissue uncovered almost 10,000 transcripts; tissue-specific studies found over 23,000 lncRNAs in bovine testes over maturation and over 1500 transcripts present in bovine oocytes [11, 17, 25, 43]. In addition to lncRNA being found in many tissues, they have been linked to biological processes: for example, nearly 8000 lncRNAs were associated with cattle metabolic

efficiency [33]. Given that lncRNAs are so wide-reaching and are under-researched, investigating the role of lncRNAs in mammary tissue may reveal more about the biological processes behind lactation, a biologically significant and economically crucial process in a dairy cow's life.

Between birth and puberty, the mammary gland (which includes the nipple, areola, stroma, and parenchyma) is and remains primarily underdeveloped. The surging levels of hormones lead the mammary tissue to increase in size during puberty, but this growth does not proceed past the filling of the mammary fat pad at which time, and the duct network has begun to branch but is not fully developed [9]. Gestation leads to the final developmental push in the mammary gland where the parenchyma increases in size and its duct system finishes branching. The duct terminals become surrounded by connective tissue and develop into secretory alveoli. At the end of pregnancy, the ducts fill with milk and the animal is ready to nurse its offspring, fulfilling the mammary gland's primary function: milk secretion. If an animal does not become pregnant and does not receive the pregnancy-based surge of estrogen and progesterone, the mammary gland development remains arrested [23].

As expected with such large physical changes, the development of the mammary gland is associated with changes in genetic expression. An example of this came from an investigation into puberty's genetic effect in cattle; researchers found 82 transcription factors in liver tissue that influence widespread tissues, demonstrating that the developmental process is reliant on gene expression changes [32]. When examining heat stress in cattle (relating to milk production), over 3000 genes were identified in the mammary gland alone, relating to anabolism, milk component synthesis, cell death, cytoskeleton degradation, and immune response, giving further support to the importance gene expression plays in mammary function [7]. Given the evidence of the genetic components of puberty and mammary gland tissues, and the vast number of morphological changes in mammary gland tissues, clearly there are many genetic elements at play in the development of mammary tissue.

Given the ever-increasing demand for dairy products, researchers and farmers alike have spent decades selectively breeding cows that produce the most and best milk possible. And, although there are well-researched genes involved in the amount and composition of milk, finding which genes are present and activated by lactation is a different beast. Research has uncovered 881 differentially expressed genes in the mammary gland tissue of lactating cows versus dry cows. The genes present in lactating cows are related to metabolic processes, protein synthesis, growth processes, immune-related processes, and more! These findings led researchers to conclude that dry cows had decreased capacity for protein synthesis, energy generation, and cell growth. However, dry cows may have a stronger immune response when compared to their lactating counterparts. Given the considerable number of coding genes identified as present between dry versus lactating cows, there is likely to be a universe of non-coding elements that are also at play [8].

Here, we report a genome-wide investigation into lncRNA isolation and identification within mammary tissue in dry versus lactating *Bos taurus* specimens. Using well-tested tools, samples were sequenced, aligned, and filtered based on lncRNA requirements. Identified lncRNA were then analyzed based on differential expression, sequence conservation, orthologous genes, GWAS enrichment, and expression correlation-based gene ontology. By uncovering more information about the lncRNA present in these tissues and their expression profiles, we both increased the knowledge about lncRNA and their connection to production traits and animal development.

3.3 Results

3.3.1 lncRNA Identification

To generate a large data set, the SRA database was used to collect a total of 103 samples, with 60 samples representing dry mammary tissue and 43 samples representing lac-

tating mammary tissue, all from *Bos taurus* [31]. Dry mammary tissue samples resulted in over 2 billion individual reads, with an average of 34,216,285 reads per sample. Lactating mammary tissue samples totaled 1.4 billion reads, averaging 31,812,950 reads per sample. Double-iterative mapping was used on all samples via Hisat2, generating an overall alignment rate of 89.61% for dry samples and 94.09% for lactating samples [24]. Using Stringtie to generate a merged, consensus sequence that was representative of each condition led to over 1 million fragments, representing 78,449 transcripts in the dry samples, and approximately 822,000 fragments, representing 100,239 transcripts in the lactating samples [42]. Given the increase in functionality of the mature mammary tissue, it is not surprising to see more transcripts represented in the lactating tissue. Once consensus sequences were generated, they were analyzed and filtered to remove fragments that did not meet lncRNA qualifications (meaning transcripts were longer than 200 basepairs in length and did not show evidence that they were made into a protein). First, consensus sequences were compared to both the ARS-UCD 1.2 Bos taurus reference genome and the UMD 3.1.1 Bos taurus reference genome to annotate known loci and transcripts, allowing their removal. This was followed by removal of transcript fragments that showed no coding potential as calculated by the coding potential calculator [21]. Noncoding transcripts were converted from sequence data to protein strings and fed through the protein family database to remove those with substantial protein domains within their sequence [37]. The final filtering step consisted of using BLAST to remove sequences that showed high sequence homology to known genes [34]. Once a final list of candidate fragments were generated, they were merged based on overlap to generate contiguous transcripts using the bedtools merge function [38]. The final list of lncRNAs for dry mammary tissue included 214 non-overlapping transcripts and 517 non-overlapping transcripts for lactating mammary tissue (Figure 3.1). Similarly to the pattern seen when observing all genome transcripts, the mature tissue has more functionality; therefore, it possesses more transcripts.

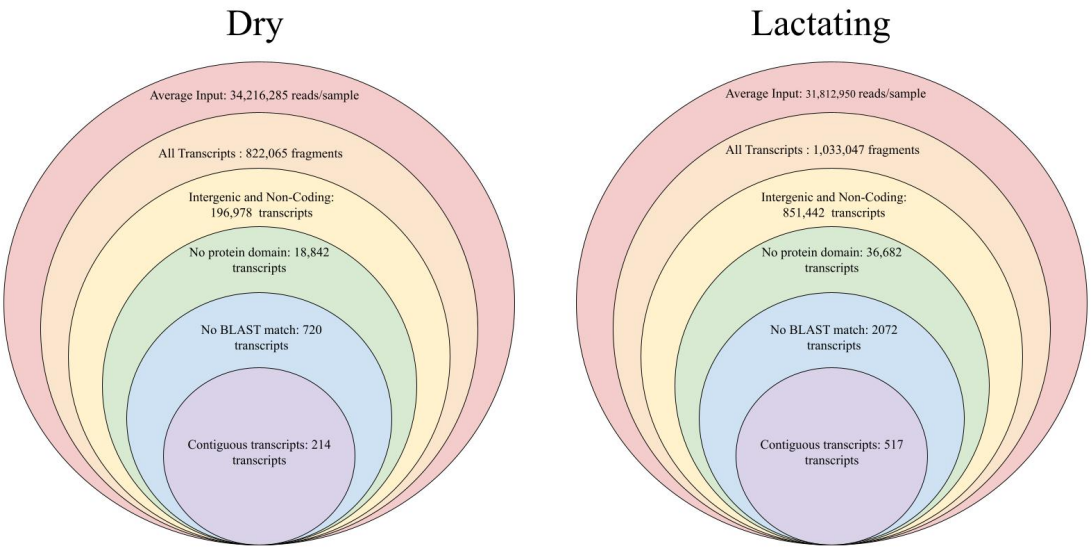

Figure 3.1: Filtering of transcripts in dry and lactating mammary tissue samples. After a consensus sequence was generated for each condition type, progressive filtering steps were taken to remove transcripts that do not meet lncRNA transcript criteria. Transcripts were aligned to the reference genome, transcripts were then kept if they were deemed intergenic based on the reference genome, noncoding as determined by the Coding Potential Calculator, lacking a definable protein domain, and lacking sequence homology. The filtering resulted in 214 contiguous transcripts in the dry tissue and 517 contiguous transcripts in the lactating tissue.

3.3.2 lncRNAs as Compared to Protein Coding Transcripts

When analyzing the entire genome, dry transcripts ranged from 53 to 1,525,771 bases, with the average length being 42,201 basepairs. Dry lncRNAs ranged from 210 to 16,503 bases, averaging 2139 bases. This fits with previous research demonstrating that even with the 200-basepairs minimum length, lncRNAs are characteristically shorter than their whole genome counterparts . These findings continued into lactating transcripts, which ranged from 13 basepairs in length to 1,838,491 basepairs, with the average length 37,873 reads in the whole genome compared to the 202 to 54,234 bases seen in lactating lncRNA, which averaged 2175 bases (Figure 3.2). Another key characteristic of lncRNA is the lower number of exons per transcript [15]. Dry whole genome transcripts averaged 7.912 exons per transcript and lactating whole genome transcripts averaged 6.5 exons per transcript. Dry lncRNAs showed 1.6 exons per transcript and lactating lncRNAs showed 2.2 exons per transcript. Both the length and exon number findings support the conclusion that the lncRNAs identified are in fact lncRNAs and not false positives.

Another indicator in favor of candidate lncRNAs being genuine is the expression level. It is common to see lncRNAs show lower expression levels as compared to their whole genome and/or protein coding counterparts [41]. Overall, dry transcripts were expressed at an average level of 15.076 FPKM and lactating transcripts showed 33.595 FPKM expression levels on average. Dry lncRNA transcripts were expressed at 2.239 FPKM on average, lactating lncRNAs were expressed at 18.794 FPKM. Although not particularly proportional, the theme of lactating transcripts being expressed at a higher level is seen in both all transcripts and lncRNAs. The levels of expression are demonstrably higher in all transcripts as compared to lncRNAs only (Figure 3.3).

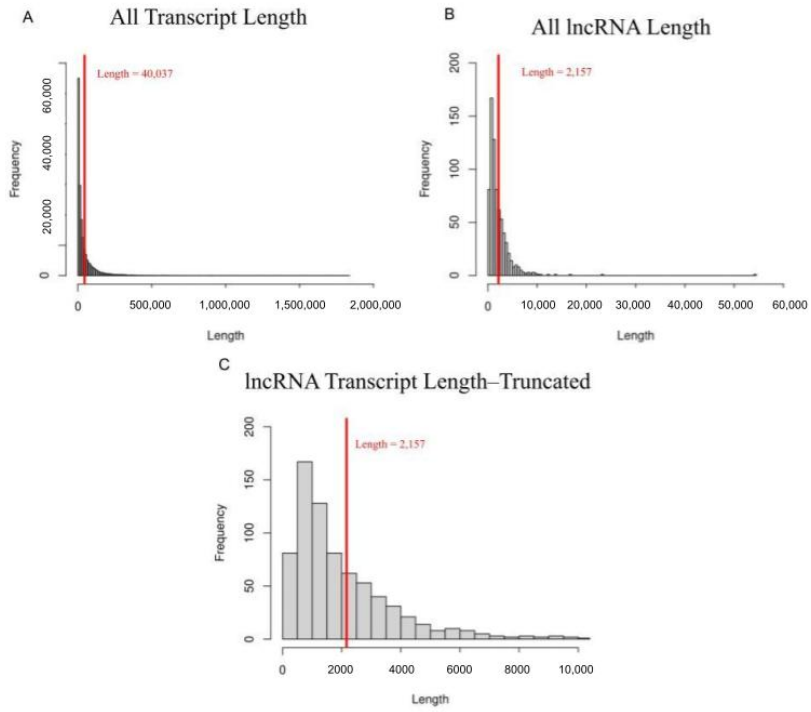

Figure 3.2: Length distribution of various transcripts. (**A**) Length of all transcripts in both dry and lactating samples. Transcripts ranged from 13 to 1,838,491 basepairs in length, the red line indicates the average transcript length of 40,037 basepairs. (**B**) Length of lncRNA transcripts in both dry and lactating samples, ranging from 202 to 16,503 basepairs. The average transcript length for candidate lncRNAs is 2157 bases and is indicated by the red line. (**C**) For added clarity, a truncated histogram representing lncRNA lengths, excluding the small number of those longer than 10,000 basepairs. Red line remains indicative of the average lncRNA length of 2157 basepairs.

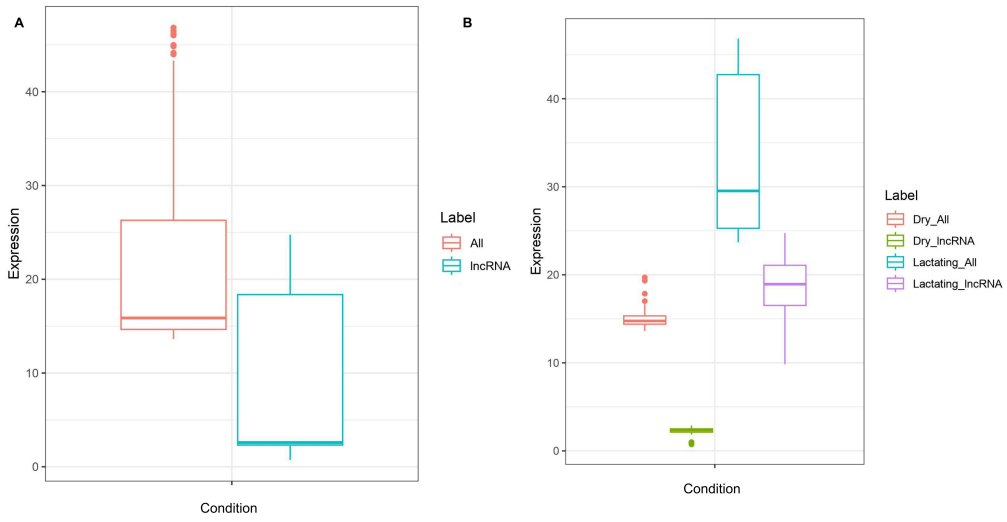

Figure 3.3: Expression of transcripts. (**A**) Using FPKM values, lncRNA are expressed at demonstrably lower levels than coding transcripts. All transcripts averaged 24.36 FPKM. lncRNA averaged 10.52 FPKM.(**B**) Expression levels additionally divided by condition type, measured via FPKM. Pattern of lower lncRNA expression as compared to all transcripts continues to hold true. Dry coding and lncRNA transcripts show lower expression than lactating transcripts, regardless of transcript type. Dry lncRNAs averaged 2.239 FPKM, all dry transcripts averaged 15.076 FPKM, lactating lncRNAs averaged 18.794 FPKM, and all lactating transcripts averaged 33.595 FPKM.

3.3.3 Differential Expression of lncRNA in Dry Versus Lactating Mammary Tissue

Of the 214 dry lncRNA transcripts and the 517 lactating lncRNA transcripts, there were 56 transcripts that covered similar regions of the genome. In total, 3 were removed due to expression not being seen in at least ½ of the samples in each condition. These lncRNA transcripts ranged from 229 basepairs to 16,503 basepairs, with dry lncRNAs averaging

2343 reads in length and lactating lncRNAs averaging 2454 basepairs. These lengths are close to the lengths seen in all lncRNAs (2139 bases in all dry lncRNAs, 2175 bases in lactating lncRNAs). The expression levels of common lncRNAs was 7.778 in dry lncRNAs and 6.747 in lactating lncRNAs. Of the 56 common lncRNAs, 44 transcripts were differentially expressed at a statistically significant level at p ¡ 0.05, when compared with a student t-test. In total, 22 of the differentially expressed transcripts were upregulated in the dry tissue and 22 were upregulated in the lactating tissue. The final 9 common transcripts were not differentially expressed at a statistically significant level (Table 4.1).

3.3.4 Sequence Conservation of lncRNA

To continue the validation of candidate transcripts, as well as to begin to attempt to assign roles to the transcripts of interest, sequence conservation was investigated. This was carried out by using the UCSC LiftOver tool to lift over all transcripts, solely intergenic transcripts, and lncRNA transcripts to their matching coordinates on the human genome. A Phastcon 100-way file was used to determine the level of sequence conservation across lncRNA-associated regions of interest [40].

Of the reads successfully lifted over the human genome, the “all transcripts” group showed higher sequence conservation on average than the other two groups. Intergenic transcripts overall were more conserved than the lncRNA group but less than the “all transcripts” group. When plotted as a violin graph, it is readily apparent that most lncRNAs are localized closer to the 0 end of the scale, more so than the “all transcripts” group and the intergenic group. All transcripts scored between 0.0003 and 1, averaging 0.15. Intergenic dry transcripts scored between 0.00004 and 1 and averaged 0.16, with intergenic lactating transcripts scoring between 0 and 1 and averaging 0.15. Dry lncRNAs scored as low as 0.0001 and as high as 0.816, with an average of 0.12, and lactating lncRNAs scored

Table 3.1: *T*-test of expression levels between lncRNAs covering comparable regions of the genome, thus deemed to be the common lncRNAs. Average expression was calculated and reported in FPKM for each transcript. Two-sample *t*-tests and Welch's *t*-tests were used to determine those transcripts expressed at levels deemed to be statistically significant based on a *p*-value of 0.05. 44 transcripts were deemed differentially expressed.

Dry		**Lactating**		***p*-Value**
lncRNA	**Expression**	**lncRNA**	**Expression**	
chr1: 88,402,354–88,403,436	0.552	chr1: 88,402,258–88,403,346	0.436	0.2258
chr10: 28,036,993–28,037,708	7.006	chr10: 28,036,875–28,037,846	4.436	$< 2.20x10^{-16}$
chr11: 102,644,124–102,648,416	0.592	chr11: 102,644,428–102,648,357	0.374	$4.48x10^{-11}$
chr11: 74,657,704–74,660,914	0.325	chr11: 74,658,354–74,660,959	0.281	0.06305
chr12: 32,163,152–32,163,433	0.953	chr12: 32,163,199–32,163,464	0.966	0.8596
chr12: 84,995,822–84,996,199	19.591	chr12: 84,995,788–84,996,023	16.090	0.03071
chr13: 29,201,242–29,205,645	0.305	chr13: 29,202,525–29,205,728	0.554	$1.80x10^{-10}$
chr13: 63,992,771–63,994,410	0.750	chr13: 63,992,252–63,994,789	0.310	$1.01x10^{-7}$
chr13: 76,610,578–76,612,313	0.617	chr13: 76,610,635–76,611,936	0.562	0.4465
chr14: 17,032–17,554	0.585	chr14: 16,830–18,164	0.733	0.001775
chr16: 42,802,614–42,812,241	33.971	chr16: 42,802,637–42,812,194	15.692	$< 2.20x10^{-16}$
chr16: 48,586,547–48,586,896	0.458	chr16: 48,584,486–48,586,868	0.612	0.0005258
chr16: 71,804,283–71,805,220	1.440	chr16: 71,804,288–71,805,212	18.671	$< 2.20x10^{-16}$
chr17: 69,703,586–69,705,950	2.227	chr17: 69,703,579–69,705,885	1.975	0.3815
chr18: 13,007,057–13,010,244	5.469	chr18: 13,007,179–13,009,811	28.722	$6.78x10^{-13}$
chr18: 1,829,126–1,830,665	0.196	chr18: 1,829,460–1,830,200	0.398	0.0001832
chr18: 38,105,619–38,111,753	6.826	chr18: 38,105,601–38,111,755	3.175	$< 2.20x10^{-16}$
chr18: 58,480,168–58,480,492	0.326	chr18: 58,479,948–58,480,580	0.448	0.0004047
chr18: 7,018,326–7,020,326	0.521	chr18: 7,018,199–7,020,326	0.283	$2.41x10^{-08}$
chr19: 41,517,108–41,519,655	0.550	chr19: 41,518,998–41,519,227	0.288	$1.28x10^{-14}$
chr2: 111,874,235–111,875,148	4.087	chr2: 111,874,066–111,877,333	2.759	$1.96x10^{-5}$
chr2: 115,539,609–115,540,162	235.756	chr2: 115,539,562–115,540,164	138.488	$7.90x10^{-10}$
chr2: 125,783,114–125,794,043	24.285	chr2: 125,777,445–125,793,948	33.604	$3.36x10^{-9}$
chr21: 58,272,796–58,274,409	1.584	chr21: 58,272,808–58,274,430	0.854	$< 2.20x10^{-16}$
chr21: 68,475,571–68,478,473	2.627	chr21: 68,475,747–68,477,894	2.432	0.6144
chr23: 11,415,564–11,420,435	19.060	chr23: 11,415,556–11,420,435	8.688	$< 2.20x10^{-16}$
chr23: 17,153,445–17,154,227	8.728	chr23: 17,153,446–17,154,226	10.971	0.0002695
chr23: 17,771,969–17,776,464	2.292	chr23: 17,771,923–17,776,476	0.578	$< 2.20x10^{-16}$
chr25: 1,035,522–1,036,687	10.186	chr25: 1,035,522–1,036,586	16.318	$5.80x10^{-15}$
chr25: 35,533,291–35,534,130	0.988	chr25: 35,533,317–35,534,096	19.412	$< 2.20x10^{-16}$
chr25: 39,512,349–39,513,122	12.600	chr25: 39,512,418–39,513,043	5.562	$< 2.20x10^{-16}$
chr25: 488,853–489,320	3.302	chr25: 488,852–489,379	2.741	0.02195
chr25: 786,947–787,788	1.176	chr25: 786,326–788,530	0.780	$13.55x10^{-8}$

Table 3.2: *Cont.*

Dry		**Lactating**		***p*-Value**
lncRNA	**Expression**	**lncRNA**	**Expression**	
chr26: 10,508,859–10,510,408	0.425	chr26: 10,508,507–10,510,415	0.526	$2.61x10^{-5}$
chr26: 21,485,237–21,487,166	0.418	chr26: 21,485,168–21,487,115	1.136	$2.49x10^{-11}$
chr26: 51,014,183–51,020,151	0.424	chr26: 51,016,793–51,020,761	0.234	$1.81x10^{-14}$
chr27: 1,092,242–1,093,287	0.894	chr27: 1,092,247–1,093,275	1.036	0.08457
chr29: 46,399,007–46,401,176	0.316	chr29: 46,399,026–46,401,485	0.302	0.745
chr29: 46,623,288–46,624,946	1.844	chr29: 46,623,270–46,625,478	3.484	$1.86x10^{-9}$
chr29: 47,279,192–47,281,419	0.440	chr29: 47,278,057–47,281,321	1.610	$< 2.20x10^{-16}$
chr29: 47,437,521–47,440,208	0.603	chr29: 47,437,552–47,440,146	1.012	$3.96x10^{-12}$
chr4: 52,093,644–52,094,267	4.437	chr4: 52,093,643–52,094,277	0.438	$1.16x10^{-14}$
chr5: 118,095,776–118,100,684	0.434	chr5: 118,095,874–118,100,834	0.303	$9.99x10^{-5}$
chr6: 3,400,389–3,401,869	0.596	chr6: 3,400,394–3,401,718	5.301	$< 2.20x10^{-16}$
chr6: 62,052,709–62,053,912	0.368	chr6: 62,052,696–62,054,059	0.511	0.002736
chr7: 10,013,140–10,014,516	1.509	chr7: 10,013,140–10,014,265	0.391	$2.82x10^{-8}$
chr7: 11,559,071–11,560,742	0.058	chr7: 11,559,861–11,561,374	0.100	0.2047
chr7: 17,181,130–17,183,216	0.703	chr7: 17,181,196–17,183,208	11.387	$< 2.20x10^{-16}$
chr7: 60,229,378–60,232,948	9.270	chr7: 60,229,471–60,232,796	4.910	$7.91x10^{-15}$
chr9: 98,186,197–98,187,338	1.152	chr9: 98,186,218–98,187,339	3.450	$< 2.20x10^{-16}$
ChrUn.004.1159: 35,363–37,180	0.171	ChrUn.004.1159: 35,361-38,038	0.344	$3.20x10^{-5}$
ChrUn.004.301: 120,666–127,066	0.410	ChrUn.004.301: 124,684-128,068	0.170	$6.17x10^{-15}$
ChrUn.004.3283: 1785-2847	0.330	ChrUn.004.3283: 1822-2792	2.147	$8.56x10^{-13}$

between 0 and 0.9, averaging 0.1 (Figure 3.4). When plotted as rank versus score, it is obvious that most lncRNAs do not show high conservation, but a small amount are highly conserved (Figure 3.5). This falls in line with lncRNA trends [19]. Common lncRNAs scored between 0.00003 and 0.816, with an average score of 0.16, demonstrating that the common lncRNAs are slightly more conserved than all transcripts, intergenic transcripts, and all lncRNA transcripts.

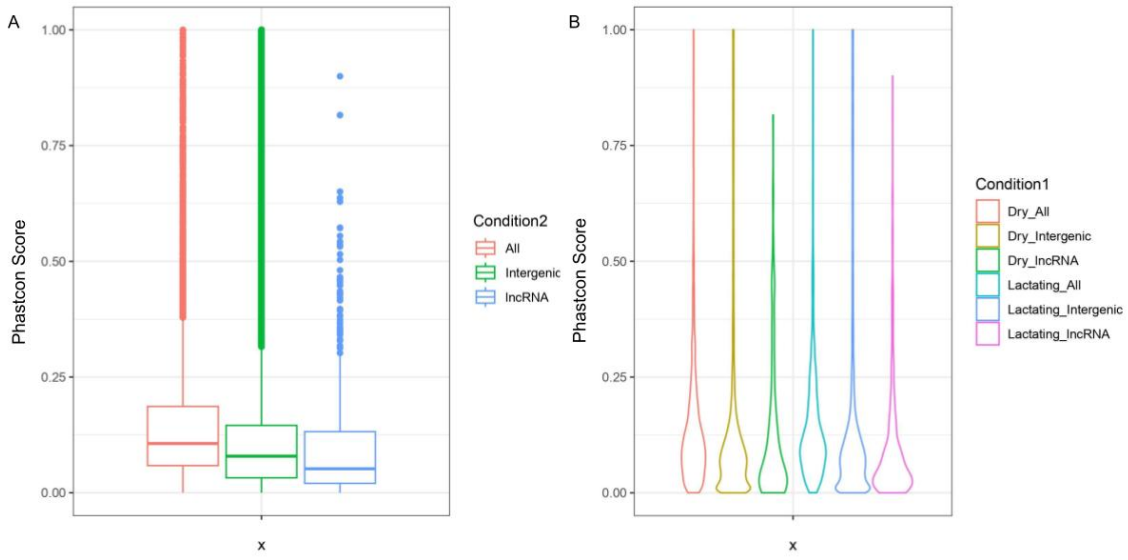

Figure 3.4: PhastCons scores of various transcript groups. (**A**) Boxplot of PhastCons scores for all transcripts, intergenic transcripts, and lncRNA transcripts. (**B**) Violin plot of all transcript profiles: all dry transcripts, dry intergenic transcripts, dry lncRNA transcripts, all lactating transcripts, lactating intergenic transcripts, and lactating lncRNA transcripts.

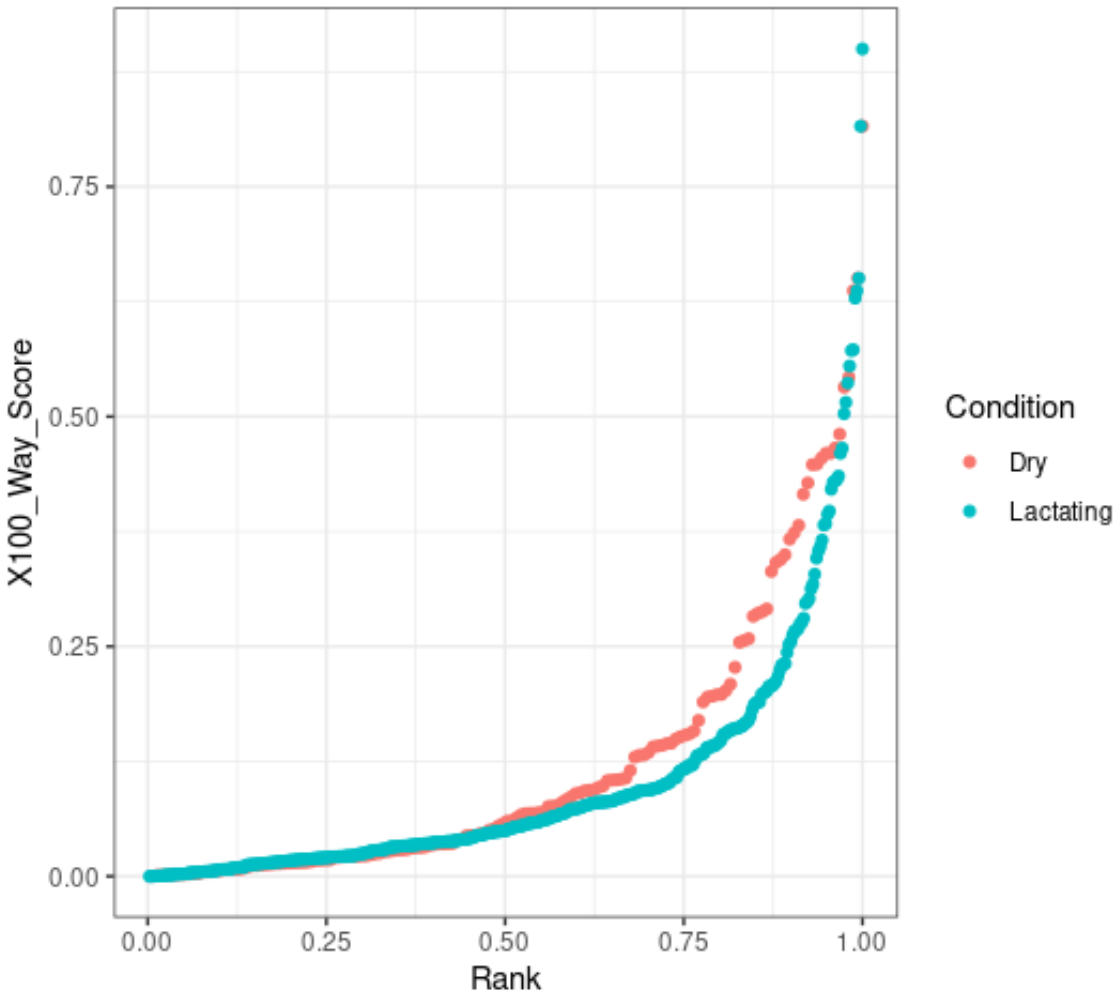

Figure 3.5: Scatter plot of lncRNA PhastCons scores. As is common, many candidate lncRNAs show very low sequence conservation with a small subset showing higher scores. Dry lncRNAs ranged from 0.001 to 0.816, with an average score of 0.12. Lactating lncRNA scores ranged from 0 to 0.9, with an average score of 0.1. These scores do not have a unit.

3.3.5 Annotation of lncRNA of Interest Based on Homology

To attempt to functionally annotate lncRNA transcripts, sequence conservation was used to predict lncRNA roles. From the whole collection of lncRNAs, those with a con-

servation score of 0.5 or higher were kept for further analysis, as well as common lncRNAs with a conservation score of 0.25. Of the 214 dry lncRNAs, 5 transcripts were kept and 4 overlapped with annotated human genes; of the 517 lactating transcripts, 11 transcripts were kept and 9 overlapped with annotated human genes. The 106 common transcripts (53 from each sample) resulted in 28 kept transcripts, 12 from lactating transcripts, and 16 from dry transcripts, with 20 transcripts overlapping with an annotated human genetic feature. Several of the overlapped genetic features represented the same feature; of the 33 identified overlapped features, there were 18 unique features. Even the unique features tended to represent common functions: transcriptional regulation/factors, translational associations, signaling, and keratin filaments were all represented several times. Other roles featured in these conserved regions included inhibition of endothelial cell proliferation, association with synapse, inositol hydrolysis, and exocytosis regulation. Additionally, regions also overlapped with a known lncRNA and a documented pseudogene. The most common overlapped elements were as follows: DLG associated protein 2, keratin-associated protein 16-1, small cysteine and glycine repeat containing protein 1 and 2, and zinc finger homeobox protein 3. The use of gene ontology has associated DLG associated protein 2 with a role in signaling. Keratin-associated protein 16-1 is associated with keratin filament in the cytosol, as seen with gene ontology. The small cysteine and glycine repeat containing proteins (protein 1 and 2) are both matrix proteins that use extensive disulfide bonds to cross-link with the cysteine residues of the hair keratins, leading to strong and resilient hair shafts. Although zinc finger homeobox protein 3 was seen more than other features, transcripts of interest also overlapped zinc finger protein 219 and zinc finger protein castor homolog 1; these features all have roles heavily associated with transcriptional regulation. Of the 9 transcripts not differentially expressed, 8 were successfully lifted over to represent human genome locations and analyzed for sequence conservation. Most of the transcripts showed low conservation; however, a small number showed conservation

with a score above 0.25. Those with the slightly higher conservation showed overlap with adhesion G protein-coupled receptor L1, nuclear receptor coactivator 1, and disks large-associated protein 2. The functions of these proteins are a potential exocytosis regulator, a transcriptional coactivator, and a synapse associated protein, respectively.

3.3.6 lncRNA-Gene Coexpression Correlation and Ontology

Investigating lncRNA from another avenue, gene co-expression analysis was used to identify genes expressed statistically significantly with a lncRNA. To avoid low expression based false correlations, any transcripts (lncRNAs or coding genes) with expression in less than 5 samples were removed and a correlation matrix was generated, yielding an R^2 value for all pairwise relationships. For each lncRNA, those coding genes whose R^2 value had a p value < 0.05 were isolated for gene ontology analysis. Although it is important to remember that correlation does not indicate causation, and showing statically significant correlation in their expression levels does not indicate that a lncRNA shares the same function as the correlated genes, the function of the gene cluster may be indicative of traits associated with the lncRNA. Like many lncRNA functional annotations, these act as reasonable hypotheses for lncRNA function. Of the 214 dry lncRNA, 126 were retained for correlation analysis. The expression levels of these lncRNAs were correlated with 29,809 coding genes, and R^2 values ranged from -0.861 to 0.956, with the average correlation coefficient being 0.016, indicating there were not a substantial number of relationships that were more positively associated or more negatively associated. When investigating the clusters of statistically significantly correlated genes, many lncRNAs clustered with coding genes that showed some sort of function. In total, 15 of the 126 lncRNAs did not associate with any *Bos taurus* biological function, and the other 111 lncRNAs demonstrated statistically significant association with 85 unique functions. Several associations were related

to regular bodily functions, such as cell differentiation, actin filament network formation, and immune responses; many associations represented regulation (in both the positive and negative direction) of cellular processes. The most highly represented association exists between lncRNAs and the regulation of telomere maintenance via telomerase, with this association representing 9 of the investigated clusters. Also associated with several lncRNAs were peptide antigen assembly with MHC class II protein complex (4 clusters), spindle assembly (3 clusters), regulation of natural killer cell mediated cytotoxicity (3 clusters), monoubiquitinated histone H2A deubiquitination (3 clusters), and the detection of chemical stimulus involved in sensory perception of smell (3 clusters). Of the 517 lactating lncRNAs, 233 were used for co-expression correlation analysis. These lncRNAs were correlated based on expression level with 24,374 coding genes. The reported R^2 values fell in similar ranges to the dry samples, ranging from -0.870 to 0.950, with the average being 0.035, indicating the relationships skew slightly more in favor of positive correlations, but not in a notable manner. Whereas 11.9% of dry lncRNA-correlated clusters investigated showed no association, 22.3% of lactating lncRNA-correlated clusters did not return an acceptable association. Lactating lncRNA-associated clusters also represented many different connections, including cell maturation, protein insertions, immune response, and many instances of regulation. Clusters represented 122 unique associations, with the most highly represented cluster being the detection of chemical stimuli involved in sensory perception of smell, representing 12 clusters. Other repeatedly represented clusters include the regulation of telomere maintenance via telomerase (7 clusters), establishment of protein localization to mitochondrial membrane (5 clusters), and positive regulation of DNA biosynthetic processes (5 clusters).

3.3.7 lncRNA-Gene Coexpression Correlation Annotation

Using the smallest *p*-value to identify the most statistically significantly correlated gene for each lncRNA, several interesting associations and patterns emerged. Similarly to gene ontology analysis, correlation is not indicative of a guaranteed relationship, but clues to lncRNA function may reveal themselves. Of the 126 investigated lncRNAs in the dry tissue condition, 107 of the lncRNAs correlated with known named genes (the other 19 correlated with novel genes) and 103 of those known genes had functions/roles assigned to them. Correlated genes represent many roles, pathways, and functions. Gene functions that were represented by several genes were cell cycle regulation and/or cell proliferation, protein transport and binding, and various membrane interactions/features. A correlated gene of note is the MTA1 gene, which has been associated with mammary adenocarcinoma cells; although the role has not been fully investigated, the presence of this protein/gene may be useful in improving cattle health. A second notable gene associated with an identified lncRNA is the BTN1A1 gene, encoding a major protein that is associated with fat droplets in milk. The regulatory nature often seen in lncRNA makes this connection interesting; could this be used to continue the improvement of milk production? In lactating tissue, 233 lncRNAs were associated with genes. Of the 233 genes, 25 associated genes were novel transcripts, and 16 have been named and not fully investigated, and the remaining 192 have been assigned a role. Trends in associated gene function included protein binding, signaling, and mRNA and tRNA processing, as well as several cancer associations, although it should be noted that many roles, functions, and pathways were represented by these correlated genes. Noteworthy genes within lactating lncRNA correlations include: FOLR1, a folate receptor (folic acid, the receptor's target, is very important in pregnancy); CAPN6, a protein highly expressed in the placenta; CSN3, a lactation-associated protein stabilizer; PTGFR, a protein implicated in uterus contractions; and OXTR, an oxytocin

receptor. Given that pregnancy is the inciting event for the final mammary tissue development, the associations with pregnancy-related proteins and processes are worth further investigation. Harkening back to highly conserved lncRNAs and their potential function, KRT17 was an associated gene that yields the type I intermediate filament chain keratin 17, a protein found in nail beds, hair follicles, sebaceous glands, and other epidermal appendages. This correlation falls in line with the keratin-associated protein 16-1 and the small cysteine and glycine repeat containing proteins identified via sequence homology.

3.3.8 SNP Heritability Enrichment Analysis on Cattle Traits

To associate lncRNAs with traits of interest, the 214 dry tissue and 517 lactating tissue lncRNA transcripts identified were integrated with Holstein GWAS data to establish single nucleotide polymorphism (SNP) heritability enrichment for 4 traits: daughter pregnancy rate (to measure reproduction and fertility), livability (to measure health), milk production (to measure milk production value), and stature (to represent body conformation). Using MINQUE for partitioning heritability (MPH) (`https://jiang18.github.io/mph/index` (accessed on June 2023)), SNP heritability enrichment was measured as a ratio of per-SNP heritability near the lncRNA of interest as compared to the entire genome. Enrichment analysis used 2,795,435 genome-wide SNPs, 4257 dry tissue SNPs, and 12,884 lactating tissue SNPs. As a result, significant enrichment of per-SNP heritability near lncRNAs was found in both dry and lactating tissues in reproduction- and production-associated traits. Enrichment was found to be significant in daughter pregnancy rate in dry tissue (8.92x; $p = 0.015$) and lactating tissue (4.96x; $p = 0.019$). Enrichment was also found to be significant in milk production, with dry tissue demonstrating an enrichment level of 7.39x ($p = 0.01$) and lactating tissue demonstrating an enrichment level of 4.73x ($p = 0.003$). Livability and stature were not statistically significantly enriched in either tissue condition

(Table 3.3).

Table 3.3: Enrichment of mammary lncRNAs in both dry and lactating tissue, based on integrated cattle GWAS data using the MPH software. Daughter pregnancy rate (DPR) represents reproduction value, livability represents health, milk represents production value, and stature represents body conformation. Data reported include estimate of per-SNP heritability enrichment, standard error of estimate of per-SNP heritability enrichment estimate, and *p*-value for per-SNP heritability enrichment estimate.

	DPR			**Livability**			**Milk**			**Stature**		
	Enr	**SE**	***p***	**Enr**	**SE**	***p***	**Enr**	**SE**	***p***	**Enr**	**SE**	***p***
Dry	8.92	3.65	0.02	6.10	5.12	0.16	7.39	2.50	0.01	2.93	2.27	0.20
Lact	4.96	1.91	0.02	−0.39	2.39	0.72	4.73	1.36	$3.00x10^{-3}$	6.10	1.53	0.10

3.4 Discussion

This study aimed to collect publicly available RNASeq data from *Bos taurus* dry and lactating mammary tissue. The purpose was to identify lncRNA present in each tissue type, identify those common to both conditions, and attempt to assign some sort of role or function to those lncRNA transcripts identified. Ideally, these findings will shed light on the role lncRNA plays in the development of the mammary gland as it develops to meet both the animal's and our own needs. RNAseq data was collected from the NCBI SRA database before being aligned to a well-researched reference genome. Transcripts were filtered based on intergenic nature, coding potential, presence of protein domains, and sequence similarities to known genes to narrow the transcripts down to a final list of 731 candidate lncRNA transcripts. When further analyzed, these transcripts showed shorter length, lower expression, fewer exons, and less sequence conservation than their whole genome counterparts; this is in line with previous lncRNA identification findings. These 731 transcripts represent 214 transcripts found in dry mammary tissue and 517 transcripts found in lactating tissue. Within these transcripts, 53 were identified to be common to both tissue conditions. The

majority of these common transcripts were found to be statistically significantly differentially expressed, with 22 transcripts showing higher expression in dry tissue, 22 transcripts showing higher expression in lactating tissue, and 9 transcripts showing similar expression levels in both conditions. The identification of these lncRNA created a framework of transcripts to further investigate and begin assessing potential functions and roles, as well as to surmise how these transcripts may fit into the development of the mammary tissue.

At the most basic level, there are lncRNA transcripts present in Bos taurus mammary tissue. The lncRNA identified also varied in both presence and levels between dry and lactating samples, demonstrating that the genetic lncRNA profile is related to the changes that occur as a result of mammary gland development. To assign roles to the identified transcripts, sequence conservation was used to determine which transcripts were highly conserved and therefore, more likely to act in a comparable way across species. Interestingly, many of the more conserved lncRNA elements tended to show high sequence conservation with documented transcriptional regulators. This supports previous research that lncRNA acts in regulatory fashions across the genome. The lncRNA identified also varied in presence and levels between dry and lactating samples, showing that the genetic lncRNA profile is related to the changes that occur due to mammary gland development. Another common theme in elements showing high sequence conservation was proteins related to keratins, this is noteworthy because research has demonstrated that keratin intermediate filaments are crucial to physical resilience of epithelial tissues, especially in mammary glands [1]. Given the function of the mammary gland and the stress put on the tissue by nursing calves and the milking process, it is of benefit for a dairy cow to have strong keratin intermediate filaments, created by the strong epithelial cells in their mammary tissue.

Using gene co-expression and correlation to establish highly correlated lncRNA-gene relationships and clusters of genes that can be linked to a single lncRNA were also tools used to begin hypothesizing about transcript function. Although the correlation should not

be mistaken for a definitive lncRNA function, a common school of thought indicates genes with a positive correlation have effects that act in the same direction [28]. When investigating clusters of genes that correlated statistically significantly with lncRNA, the functions and associations of these clusters were wide reaching and cover many distinct types of roles. The associations seen between lncRNA and general body function support the hypothesis that lncRNA are crucial to the daily functioning of the animal. It is also unsurprising to see many associations between lncRNA and regulation of biological pathways, as lncRNA are often heralded as regulatory elements. Most notably, however, is the association between gene clusters and sense of smell. lncRNA correlated gene were clusters associated with sense of smell in both dry and lactating tissues, but associated with more lncRNA in the lactating tissue. Data collected and reviewed in human pregnancy revealed that many pregnant people report an increase in sensitivity to smell as their pregnancy progressed, so it is possible this is occurring in cattle as well [4]. Given the inciting incident in lactation is pregnancy, the presence of these lncRNA correlated gene clusters may contribute to changes in olfactory abilities in cattle as well. Investigating the single most significantly correlated gene for each lncRNA also yielded interesting results. Lactating lncRNA showed several correlations with pregnancy-related genes as well. Genes relating to folic acid, oxytocin reception, uterus contractions, and lactation were all observed. Interestingly, keratin filament was again represented. When integrating GWAS results with lncRNA to determine heritability of SNPs in relation to complex traits, enrichment was deemed to be significant in daughter pregnancy rate and milk production in both dry and lactating tissue. Interestingly, livability and stature were not statistically significantly enriched. These findings demonstrate the close relationship between mammary tissue lncRNA and reproduction and production traits, as opposed to close relationships between all measured traits and these lncRNA. The enrichment results for milk are of note as the function of the mammary gland is milk production; although gene changes have been well investigated in the

milk production pathway, lncRNA are now being implicated in this pathway as well.

Although the use of public data allowed for a larger sample size than what may have been possible with live animal data collection, it does add variability into the analysis. This variability may allow for more rare variants to be collected, thus creating more in-depth findings; however, the increased variability could also lead to unintended noise within the data. The public data also means that factors such as age, number of calving rounds, breeding history, and other health traits are outside of our control. Another potential source of irregularity may come from the development of the mammary gland and its relationship to the calving history of each specimen; does the genetic landscape of the animal look different when it is dry as a heifer versus between pregnancies? This is something that could be further investigated to generate a more thorough understanding of lncRNA in this situation. The findings presented in this research project could be expanded in many different directions. The relationships between strong mammary epithelial tissue, keratin intermediate filaments, successful mammary development, and milking, and identified lncRNA could be useful to continue to develop healthy, well-producing animals. Both GWAS integrated results and co-expression correlation gene cluster ontology demonstrate that lncRNA are associated with both complex traits and a large variety of biological paths and functions, therefore lncRNA should not be ignored when investigating genetic components of biological processes.

3.5 Conclusions

In this study, dry and lactating mammary gland tissue were investigated for presence of lncRNA. lncRNA were found in both tissue conditions and demonstrated expected patterns for lncRNA, supporting findings as true lncRNA. lncRNA profiles differed between the two condition types with a small number of overlapping transcripts, contributing to the

variable expression profiles and implicating lncRNA in the morphological and biological changes that occur during lactation. Using various methods of lncRNA annotation, notable findings included strong hair keratin filaments, associations with the detection of chemical stimulus involved in sensory perception of smell, and enrichment of complex traits representing reproduction and production value. These findings imply these lncRNA are related to resilient mammary tissue, pregnancy-based changes, and complex traits of interest. The wide-reaching nature of these findings continues to add to the knowledge base and encourage the consideration of lncRNA in biologic research.

3.6 Materials and Methods

3.6.1 Data Collection

The SRA database was parsed for rumen tissue samples from *Bos taurus* animals, searching for both dry and lactating samples. In total, 60 samples were found for dry tissue and 43 samples were found for lactating tissue, and SRR IDs were recorded.

Data were downloaded locally using the SRA Toolkit, using the fastq-dump tool. The split-3 option was used to ensure all reads in the raw data were collected [14]. SRA Toolkit: `https://github.com/ncbi/sra-tools` (accessed on October 2022).

3.6.2 Sequence Assembly and Mapping

RNA-Seq data were mapped to the ARS-UCD1.2 reference genome using the spliced read aligner Hisat 2.2.1, after the ARS-UCD1.2 reference genome fasta file was used to build the necessary index file [24, 2]. Mapping was performed twice; the first approach was used to generate a junction file representing junctions between transcripts, and the second iteration used those generated junctions to increase fidelity of the mapping re-

sults. Mapped transcripts were then stitched into an assembled report for each sample using Stringtie v2.2.1. This was performed twice: once using the ARS-UCD1.2 genome and a second time with the UMD3.1.1 reference genome, for additional accuracy [42, 3]. Samples were then ready for further analysis. Hisat2: `http://daehwankimlab.github.io/hisat2/` (accessed on October 2022). Stringtie: `https://ccb.jhu.edu/software/stringtie/` (accessed on December 2022). ARS-UCD1.2 genome: `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002263795.1/` (accessed on October 2022). UMD3.1.1 genome: `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000003055.6/` (accessed on October 2022).

3.6.3 lncRNA Identification

Long non-coding RNA (lncRNA) identification started with using CuffCompare 2.2.1 software to annotate transcripts based on known genetic features [6]. Summary files representing samples aligned to the ARS-UCD1.2 reference genome were compared to the ARS-UCD1.2 reference genome, and summary files representing samples aligned to the UMD3.1.1 reference genome were compared to the UMD3.1.1 reference genome. Transcripts returning the intergenic "-u" flag were isolated, generating a list of intergenic transcripts for all summary files. For each sample, the list of intergenic transcripts from the ARS-UCD1.2 and UMD3.1.1 were combined and only unique transcripts retained, generating a summary list of transcripts deemed intergenic based on 2 reference genomes. Intergenic lists for all dry samples were merged to generate a single intergenic list for dry tissue; this was repeated for lactating samples. Those transcripts that were shorter than 200 basepairs in length were removed at this point. Intergenic transcripts were inputted into the Coding Potential Coding (CPC), where they were scored based on open reading frame and homology analysis, receiving a score between -1 and 1 [5]. Those transcripts that

scored below a 0 were deemed non-coding, and the reverse compliment of all intergenic transcripts was also analyzed with the CPC, allowing for generation of a list of intergenic transcripts deemed non-coding in both directions. Non-coding, intergenic transcripts were converted into protein sequences for each transcript and sequences were compared to the Pfam database to remove those that demonstrated known protein domains [37]. Remaining transcripts were returned to their genomic location coordinates, and the sequences associated with those coordinates were compared to known gene sequences to remove those transcripts that appeared to be known genes based on sequence homology using BLAST [34]. After all filtering steps, lists of transcripts remained for dry and lactating tissue that contained only transcripts that were: (a) intergenic, therefore not overlapping known genes; (b) longer than 200 basepairs in length; (c) did not show coding potential; (d) did not contain a known protein domain; and (e) did not have notable sequence homology with known genes. These remaining transcripts match lncRNA criteria. CuffCompare: `http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/` (accessed on December 2022). Coding Potential Calculator: `http://cpc2.gao-lab.org/` (accessed on December 2022). Pfam: `https://www.ebi.ac.uk/interpro/` (accessed on January 2023). BLAST: `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome` (accessed on January 2023).

3.6.4 Comparison to Coding Transcripts

Length was calculated for all transcripts and all lncRNA in both conditions. Average length was also calculated for each profile. This allowed for comparison of transcript length profiles. All transcript length data and lncRNA transcript length data were pooled, and these values were graphed in a histogram using R4.2.2, with the addition of an average length representative line [39]. A truncated graph was also created to add clarity.

Expression data were generated for all transcripts in all samples using the—A option of Stringtie v2.2.1 [42]. This is carried out via read clustering, leading to the creation of a splice graph for each read cluster that is associated with a transcript, creating a flow network for each transcript, allowing for estimation of expression level using a maximum flow algorithm, as measured in fragments per kilobase of transcript per million mapped reads (FPKM). Similarly to length profiles, this allowed for generation of expression profiles for both conditions, representing all transcripts and lncRNA. The average expression level was calculated. Expression data for all transcripts in both conditions and all lncRNAs in both conditions were pooled and charted in a boxplot using R, and expression profiles for all profiles were also graphed in a boxplot using R.

Stringtie v2.2.1 also reports exons per transcript, allowing for calculation of exon number profiles for all transcripts in both conditions and lncRNAs in both conditions [42]. These values were reported. R: `https://www.r-project.org/` (accessed on March 2023).

3.6.5 Differential Expression Analysis

Expression profiles for common transcripts were isolated from lncRNA expression profiles for both conditions. R was used to perform a two-sample *t*-test or a Welch's *t*-test, based on variance, where transcripts were deemed statistically significantly differentially expressed when the reported *p*-value was less than 0.05. Statistically significant results were highlighted in green for clarity.

3.6.6 Phastcon Analysis

Sequence conservation was investigated using a Phastcon-based analysis to assign a conservation score to regions of interest. Using transcript profiles for all transcripts, inter-

genic transcripts (as generated by CuffCompare), and lncRNA transcripts in both conditions, the UCSC LiftOver webtool was used to convert ARSUCD1.2 cattle genome coordinates to hg38 human genome coordinates [27]. The multiBigwigSummary tool was used in conjunction with the hg38 100-way conservation alignment table to generate a conservation score for each provided region [16, 30]. The hg38 100-way table utilizes data from 100 species ranging from simple lamprey to human. Data for both conditions for all transcripts, intergenic transcripts, and lncRNAs were pooled and graphed in a boxplot using R. Conservation scores for each profile were graphed in a violin plot using R as well. lncRNAs from dry and lactating tissues were assigned a rank based on their relative conservation score and this rank was graphed against the conservation score using R. lncRNAs whose conservation score was above a 0.5 were retained for further analysis.

[-25]LiftOver: `https://genome.ucsc.edu/cgi-bin/hgLiftOver` (accessed on January 2023). MultiBigWigSummary: `https://deeptools.readthedocs.io/en/develop/content/tools/multiBigwigSummary.html` (accessed on March 2023).

3.6.7 Transcriptional Annotation Based on Sequence Homology

All common lncRNA transcripts and those lncRNAs whose conservation score was above 0.5 were retained for transcriptional annotation, using their hg38 genome locations. Genomic regions were seared in the UCSC genome browser and overlapping genes were noted. Once all candidate regions were searched, overlapping genes were investigated with UniProt to assign a function to each gene. Patterns and common functions were notated and assessed based on findings. UCSC Genome Browser: `https://genome.ucsc.edu/` (accessed on March 2023). UniProt: `https://www.uniprot.org/` (accessed on March 2023).

3.6.8 Gene Co-Expression Correlation and Ontology

Salmon was used to quantify the expression of transcripts in all transcript profiles and lncRNA profiles for both conditions, reporting transcript per million (TPM) data for all transcripts provided within all samples. Salmon's additional function, quantmerge, was used to combine expression data to create summary files for all transcripts for both samples, with lncRNA being explicitly labeled [36, 35].

To eliminate false co-expression relationships from being generated due to low expression, any transcript lacking expression in at least 5 samples was removed from both summary files. Once these filtered expression summary files were created, a correlation matrix was generated for both tissue conditions using the corr function in R. These matrices were then filtered to only reflect correlations with lncRNAs, excluding gene-to-gene based correlations. These values were retained for reporting. The matrices were then converted to *p*-values, also using R.

Using excel for its ease of column sorting, each lncRNA was sorted based on ascending *p*-value. IDs of those genes that correlated with lncRNA with a *p*-value less than 0.05 were retained and passed to the web-based gene ontology enrichment analysis tool, parsing for biological process, limiting results to *Bos taurus* [12]. The result showing the highest hierarchical rank was noted, retaining biological process, fold enrichment value, raw *p*-value, and false discovery rate data. After all retained lncRNAs in both tissue conditions were analyzed for gene ontology, the results were able to be analyzed for pattern and significance. Salmon: `https://combine-lab.github.io/salmon/` (accessed on April 2023).

3.6.9 Transcriptional Annotation Based on Correlation

Using the correlation matrices generated for gene co-expression analysis, the gene with the most significant *p*-value for each investigated lncRNA was further investigated. This

was carried out by parsing the Ensembl gene database, noting the most significantly associated gene for each lncRNA. The function of each gene was assessed using the NCBI gene database. Most functions were defined by their role in humans as more information is available in human-based reports; however, the roles appear to be non-species specific and gene names were identical.

Once all gene functions were noted, patterns were assessed and reported.

3.6.10 Heritability Enrichment Analysis

To complete a heritability enrichment analysis, two sets of lncRNA were generated: those present in dry tissue and those present in lactating tissue. GWAS data obtained from 27,000 bulls and 3 million SNPs/INDELs were used to assess cattle production (milk), reproduction (daughter pregnancy rate), health (livability), and body conformation (stature) [18, 10]. Imputation of the SNPs allowed for reliable predicted transmitting abilities (PTAs) for the represented traits. This was carried out using MPH for enrichment analysis [13]. MPH uses the following linear mixed model with two genetic components to fit the data:

$$y = Xb + g_1 + g_2 + e$$

$$g_1 \sim N(0, G_1\sigma_g^2 1)$$

$$g_2 \sim N(0, G_2\sigma_g^2 2)$$

$$e \sim N(0, R\sigma_e^2)$$

The variables are measured as follows:

b : a vector of fixed effects including intercept

X: a covariate matrix corresponding to b, $g_!$, and g_2 g_1 : a vector of random effects representing the genetic contributions of the variants within lncRNAs

g_2: a vector of random effects representing the genetic contributions of the variants within the remaining genome

e: a vector of residual effects

G_1: a genomic relationship matrix (GRM) constructed by genotypes corresponding to lncRNAs

G_2: a genomic relationship matrix (GRM) constructed by genotypes corresponding to the remaining genome

G was computed using the following formula:

$$G = \frac{ZZ^T}{m}$$

Where Z represents the standardized genotypes, m is the number of variants in the GRM, using m_1 and m_2 to represent the number of variants within lncRNAs and the remaining genome, respectively. R is a diagonal matrix used to model individual reliability of deregressed PTAs. The method of calculation used by MPH relies on a Monte Carlo REML algorithm that estimates variance components (VCs) and uses a Fisher information matrix to compute the variance–covariance matrix of VC estimates. $\frac{\sigma_g^2 1}{m_1}$ equals per-SNP genetic variation in lncRNA, allowing for computation of per-SNP heritability enrichment using $\rho = \frac{\frac{\sigma_g^2 1}{m_1}}{\frac{\sigma_g^2 1 + \sigma_g^2 2}{m_1 + m_2}}$. A delta method was used to calculate the standard error of ρ.

For the 731 lncRNAs (214 dry lncRNAs, 517 lactating lncRNAs), each transcript was analyzed with a 10 kb extension window on both sides, testing for their enrichment of per-SNP heritability for the 4 representative traits of interest. This extension window was used to encapsulate linked regions, based on SNP density within GWAS data and linkage disequilibrium levels in the population (46–48). Enrichment analysis was performed with nearly 3 million autosomal variants, with a threshold of a minimal MAF value of 0.005 and a HWE p-value of at least $1x10^{-8}$ A total of 4257 variants and 12,884 variants were iden-

tified within dry and lactating lncRNAs, respectively. SNP heritability was partitioned into two genomic relationship matrices (GRMs), one representing SNPs and INDELs within lncRNAs, and the other representing the remaining SNPs/INDELs. Enrichment was measured as the ratio of per-SNP heritability near lncRNAs as compared to the genome-wide heritability, where MPH reported enrichment value and standard error. *p*-value was computed using a *t*-test. MPH: `https://jiang18.github.io/mph/index` (accessed on June 2023).

3.7 References

[1] A Aldahmash, D Lunny, and B Lane. "Morphogenesis of the mammary gland and the role of keratin expression". In: *Breast Cancer Research* 7.2 (2005), P3.09. ISSN: 1465-542X. DOI: 10.1186/bcr1204. URL: https://doi.org/10.1186/bcr1204.

[2] *ARS-UCD1.2 - bosTau9 - Genome - Assembly - NCBI*. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1/.

[3] *Bos_taurus_UMD_3.1.1 - Genome - Assembly - NCBI*. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF_000003055.6/.

[4] E. Leslie Cameron. *Pregnancy and olfaction: A review*. 2014. DOI: 10.3389/fpsyg.2014.00067.

[5] *CPC2 @ CBI, PKU*. URL: http://cpc2.gao-lab.org/download.php.

[6] *Cufflinks*. URL: http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/.

[7] Bethany Dado-Senn et al. "RNA-Seq reveals novel genes and pathways involved in bovine mammary involution during the dry period and under environmental heat stress". In: *Scientific Reports 2018 8:1* 8.1 (July 2018), pp. 1–11. ISSN: 2045-2322. DOI: 10.1038/s41598-018-29420-8. URL: https://www.nature.com/articles/s41598-018-29420-8.

[8] Wen ting Dai et al. "Transcriptomic profiles of the bovine mammary gland during lactation and the dry period". In: *Functional and Integrative Genomics* 18.2 (Mar. 2018), pp. 125–140. ISSN: 14387948. DOI: 10.1007/S10142-017-0580-X/FIGURES/7. URL: https://link-springer-com.proxy-um.researchport.umd.edu/article/10.1007/s10142-017-0580-x.

[9] Laurence Finot, Eric Chanat, and Frederic Dessauge. “Bovine mammary gland development: new insights into the epithelial hierarchy”. In: *bioRxiv* (Jan. 2018), p. 251637. DOI: 10.1101/251637. URL: https://www.biorxiv.org/content/10.1101/251637v1%20https://www.biorxiv.org/content/10.1101/251637v1.abstract.

[10] Ellen Freebern et al. “GWAS and fine-mapping of livability and six disease traits in Holstein cattle”. In: *BMC Genomics 2020 21:1* 21.1 (Jan. 2020), pp. 1–11. ISSN: 1471-2164. DOI: 10.1186/S12864-020-6461-Z. URL: https://link.springer.com/articles/10.1186/s12864-020-6461-z%20https://link.springer.com/article/10.1186/s12864-020-6461-z.

[11] Yahui Gao et al. “Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning”. In: *Genomics* 113.4 (July 2021), pp. 2045–2055. ISSN: 0888-7543. DOI: 10.1016/J.YGENO.2021.04.039.

[12] *Gene Ontology Resource*. URL: http://geneontology.org/.

[13] *GitHub - jiang18/mph: MINQUE for Partitioning Heritability*. URL: https://github.com/jiang18/mph.

[14] *GitHub - ncbi/sra-tools: SRA Tools*. URL: https://github.com/ncbi/sra-tools.

[15] Yanghua He et al. “The conservation and signatures of lincRNAs in Marek’s disease of chicken”. In: *Scientific Reports* 5.1 (Oct. 2015), pp. 1–17. ISSN: 20452322. DOI: 10.1038/srep15184. URL: www.nature.com/scientificreports.

[16] *Hg38 100-way conservation alignment - genomewiki*. URL: http://genomewiki.ucsc.edu/index.php/Hg38_100-way_conservation_alignment.

[17] Xianbo Jia et al. “Genome-wide identification and characterisation of long non-coding RNAs in two Chinese cattle breeds”. In: *https://doi.org/10.1080/1828051X.2020.1735266* 19.1 (Dec. 2020), pp. 383–391. DOI: 10.1080/1828051X.2020.1735266. URL: https://www.tandfonline.com/doi/abs/10.1080/1828051X.2020.1735266.

[18] Jicai Jiang et al. “Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls”. In: *Communications Biology 2019 2:1* 2.1 (June 2019), pp. 1–12. ISSN: 2399-3642. DOI: 10.1038/s42003-019-0454-y. URL: https://www.nature.com/articles/s42003-019-0454-y.

[19] Per Johnsson et al. “Evolutionary conservation of long noncoding RNAs; sequence, structure, function”. In: *Biochimica et biophysica acta* 1840.3 (2014), p. 1063. ISSN: 18728006. DOI: 10.1016/J.BBAGEN.2013.10.035. URL: /pmc/articles/PMC3909678/%20/pmc/articles/PMC3909678/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3909678/.

[20] Laurence Josset et al. “Annotation of long non-coding RNAs expressed in Collaborative Cross founder mice in response to respiratory virus infection reveals a new class of interferon-stimulated transcripts”. In: *RNA Biology* 11.7 (July 2014), p. 875. DOI: 10.4161/RNA.29442. URL: /pmc/articles/PMC4179962/%20/pmc/articles/PMC4179962/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4179962/.

[21] Yu-Jian Kang et al. “CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features”. In: *Nucleic Acids Research* 45.W1 (July 2017), W12–W16. ISSN: 0305-1048. DOI: 10.1093/NAR/GKX428. URL: https://academic.oup.com/nar/article/45/W1/W12/3831091.

[22] Colin Kern et al. “Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species 06 Biological Sciences 0604 Genetics”. In: *BMC Genomics* 19.1 (Sept. 2018), p. 684. ISSN: 14712164. DOI: 10.1186/s12864-018-5037-7. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5037-7.

[23] Yusuf S. Khan and Hussain Sajjad. “Anatomy, Thorax, Mammary Gland”. In: *StatPearls* (July 2021). URL: https://www.ncbi.nlm.nih.gov/books/NBK547666/.

[24] Daehwan Kim et al. “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 907–915. ISSN: 15461696. DOI: 10.1038/S41587-019-0201-4.

[25] Lambros T. Koufariotis et al. “A catalogue of novel bovine long noncoding RNA across 18 tissues”. In: *PLOS ONE* 10.10 (Oct. 2015). Ed. by Marinus F.W. te Pas, e0141225. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0141225. URL: https://dx.plos.org/10.1371/journal.pone.0141225.

[26] Johnny T.Y. Kung, David Colognori, and Jeannie T. Lee. *Long noncoding RNAs: Past, present, and future*. 2013. DOI: 10.1534/genetics.112.146704. URL: /pmc/articles/PMC3583990/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583990/.

[27] *Lift Genome Annotations*. URL: https://genome.ucsc.edu/cgi-bin/hgLiftOver.

[28] Samuel W. Lukowski et al. “Genetic correlations reveal the shared genetic architecture of transcription in human peripheral blood”. In: *Nature Communications* 8.1 (Dec. 2017). ISSN: 20411723. DOI: 10.1038/s41467-017-00473-z.

[29] Lina Ma et al. “Lncbook: A curated knowledgebase of human long non-coding rnas”. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D128–D134. ISSN: 13624962.

DOI: 10.1093/nar/gky960. URL: /pmc/articles/PMC6323930/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323930/.

[30] *multiBigwigSummary — deepTools 3.5.2 documentation*. URL: https://deeptools.readthedocs.io/en/develop/content/tools/multiBigwigSummary.html.

[31] *National Center for Biotechnology Information (NCBI)*. 1988.

[32] Loan T. Nguyen et al. "STAT6, PBX2, and PBRM1 emerge as predicted regulators of 452 differentially expressed genes associated with puberty in Brahman heifers". In: *Frontiers in Genetics* 9.MAR (Mar. 2018), p. 87. ISSN: 16648021. DOI: 10.3389/FGENE.2018.00087/BIBTEX.

[33] Wietje Nolte et al. "Biological Network Approach for the Identification of Regulatory Long Non-Coding RNAs Associated With Metabolic Efficiency in Cattle". In: *Frontiers in Genetics* 0 (Nov. 2019), p. 1130. ISSN: 1664-8021. DOI: 10.3389/FGENE.2019.01130.

[34] *Nucleotide BLAST: Search nucleotide databases using a nucleotide query*. URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

[35] *Overview – Salmon: Fast, accurate and bias-aware transcript quantification from RNA-seq data*. URL: https://combine-lab.github.io/salmon/.

[36] Rob Patro et al. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods 2017 14:4* 14.4 (Mar. 2017), pp. 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197. URL: https://www.nature.com/articles/nmeth.4197.

[37] *Pfam: Search Pfam*. URL: http://pfam.xfam.org/search#tabview=tab1.

[38] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033. URL: https://doi.org/10.1093/bioinformatics/btq033.

[39] *R: The R Project for Statistical Computing*. URL: https://www.r-project.org/.

[40] Adam Siepel et al. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes". In: *Genome Research* 15.8 (Aug. 2005), pp. 1034–1050. ISSN: 1088-9051. DOI: 10.1101/GR.3715005. URL: https://genome.cshlp.org/content/15/8/1034.full%20https://genome.cshlp.org/content/15/8/1034%20https://genome.cshlp.org/content/15/8/1034.abstract.

[41] Luisa Statello et al. *Gene regulation by long non-coding RNAs and its biological functions*. Feb. 2021. DOI: 10.1038/s41580-020-00315-9. URL: www.nature.com/nrm.

[42] *StringTie*. URL: http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual.

[43] Jian Wang et al. "20 Profiling circRNA and lncRNA Expression in Key Bovine Metabolic Tissues". In: *Journal of Animal Science* 99.Supplement_3 (Oct. 2021), pp. 9–9. DOI: 10.1093/JAS/SKAB235.016. URL: https://academic.oup.com/jas/article/99/Supplement_3/9/6383953.

Chapter 4: An Exploration of lncRNA in Immune System Associated Tissues in *Bos taurus*

4.1 Abstract

This research project intends to collect publicly available genetic data isolated from immune system related tissues, isolate lists of lncRNA from each tissue type, and investigate identified lncRNA transcripts for characteristics, patterns, expression, correlation, ontology, and enrichment. Analysis of lncRNA and potential biological connections expand the bank of information about lncRNA, as well as the immune system. Genetic data was gathered from the SRA database (a publicly available repository of high throughout sequencing data), representing 39 *Bos taurus*samples and 6 unique tissues (ileum-associated lymph node, jejunum-associated lymph node, lymphocyte, bone marrow, spleen, and thymus tissue). Progressive filtering steps were used to remove transcripts, leaving only transcripts deemed noncoding. After filtering, 455 lncRNA were identified across the 6 tissue types: 126 ileum-associated lymph node lncRNA transcripts, 138 jejunum-associated lymph node lncRNA transcripts, 83 lymphocyte lncRNA transcripts, 25 bone marrow lncRNA transcripts, 40 spleen lncRNA transcripts, and 43 thymus lncRNA transcripts. Patterns in characteristics of these lncRNA were reflective of historic findings. Transcripts were shorter, showed fewer exons, lower expression levels, and were conserved less than their coding counterparts. Sequence conservation analysis showed lncRNA tended to be similar to gene expression control genes, with one transcript of interest being similar to the T-cell prolifer-

ation associated gene MTCP1. Genetic correlation associated most lncRNA with general cell functions (such as binding activities and signaling), whereas correlated gene cluster ontology heavily associated lncRNA transcripts with the G protein-coupled receptor signaling pathway and the detection of chemical stimulus involved in sensory perception of smell; the functions of these processes are heavily involved in immune response, making these compelling findings. Enrichment analysis, associating lncRNA with GWAS data, revealed lncRNA-associated SNPs to be enriched in complex traits representing overall body composition and reproductive value. Overall, this study successfully identified lncRNA in immune system related tissues. These lncRNA matched lncRNA characteristics, and investigations into their functions revealed connections between immune system functions and associated genes. While connections existed between identified lncRNA and general cell functions, repeated connections to immune specific biological processes were also observed. Interestingly, the sense of smell (a key component of immunological responses to the outside world) was seen across multiple analyses and multiple identified transcripts. The identification and analysis of immune system related traits associated lncRNA with many immunological processes and opens the door to further analysis that may lead to more immunologically strong dairy cattle.

4.2 Background

The immune system is multifaceted, complex, and wide reaching. At the most simplistic level, it functions to protect its host from infectious agents of all types; when a host system is deficient in their immune system (that is, immunocompromised), they are more susceptible to infection by particulate that does not cause disease in a non-immunocompromised individual. To protect the host, the immune system must be able to discern between normal cells and intruders or damaged cells; this is done by the innate

immune system first. These cells recognize intruders and destroy them, a subset of these cells utilize a small portion of the invading cell to trigger the adaptive immune system. Once the adaptive immune system has been alerted to the incoming intruders, adaptive immune cells can begin a specialized attack on the pathogen. The immune system relies on many types of cells, most of which are made in the bone marrow first. Lymphocytes mature in the bone marrow, leading to the formation of B cells. Lymphocytes subsequently leave the bone marrow and travel to the thymus to become T cells after completing thymus –localized maturation. After creation and maturation, many immune cells are found in lymph nodes and lymphatic fluid. The location of the lymph nodes is strategic, tending to be located close to bodily entries, which would be the most likely location for a pathogen attack. Lymph nodes act as housing for immune cells, allowing for the speedy release of immune cells when an immune response is triggered. Lymph nodes then employ blood vessels and lymphatic vessels to transport the immune cells around the body, allowing for protection across the whole host body system. The largest internal organ related to the immune system is the spleen, which filters blood for pathogens and can release an immune response from its immune cell stores if a pathogen is detected. For this reason, 2 distinct lymph nodes (ileum-associated and jejunum-associated), lymphocytes, bone marrow, spleen tissue, and thymus tissue are to be investigated in this study [40].

History is riddled with proof that immunity is deeply connected to genes, with some people being naturally immune to plagues like the Black Death and tuberculosis. However, there is a subset of genetic components that have historically been neglected when researching genetic control: long-noncoding RNA (lncRNA). lncRNA are genetic elements that are transcribed from DNA into RNA but are never translated into a functional protein, the “long” portion of their name indicates they are longer than 200 base pairs in length. Research has associated these transcripts with mRNA, DNA, proteins, and miRNA;

these interactions allow lncRNA to function in gene regulation via epigenetics, transcriptional control, pre- and post-transcriptional control, translational control, and pre- and post-translational control [55]. Overall, it has been demonstrated that lncRNA are very wide reaching.

Previous research connecting lncRNA to the immune system has yielded interesting results. It appears that lncRNA are a potential immune system regulator; they may also play a role in autoimmune disorders [5]. Given economic loss due to disease is an ongoing concern for cattle farmers, there is no doubt it behooves farmers and researchers to better understand disease, immunology, and how farmers can decrease loss due to disease. To further explore the relationship between lncRNA and the immune system, 6 tissues were investigated for the presence of lncRNA. Once identified, these lncRNA were analyzed. The analyses included characteristic pattern detection, differential expression, sequence conservation, genetic correlation, gene ontology, and GWAS integrated enrichment. These findings contribute to a more complete understanding of lncRNA overall and their role in immunology. Further research based on these findings may decrease economic loss in *Bos taurus* herds due to disease.

4.3 Results

4.3.1 lncRNA Identification

To obtain data, the SRA database was parsed to collect genetic information for lymphatic system associated tissues in cattle: ileum-based lymph node tissue, jejunum-based lymph node tissue, lymphocyte tissue, bone marrow tissue, spleen tissue, and thymus tissue. These tissues represent both the primary and secondary lymphoid organs [54]. In total, 39 samples were collected, with samples possessing an average of 33,586,076 reads.

Data processing began by using a double iterative mapping approach with the alignment software Hisat2, the first iteration of alignment was performed to generate a junction file that was subsequently used in the second iteration to increase accuracy of alignment [27]. Overall, the first iteration generated a successful alignment rate of 86.92%, whereas the second iteration (the files generated by this alignment was used in subsequent steps) aligned with an average success rate of 87.41%. Following the successful alignment, the assembler tool StringTie was used to assemble aligned reads into representative transcripts [50]. This was performed twice, using 2 reference genomes to avoid loss of information: UMD3.1.1 and ARS-UCD1.3 Bos taurus genomes were employed [6, 3]. Across all samples, using both genomes, samples averaged 78,658 transcripts per sample. Given the non-coding nature of lncRNA, the first filtering step was to isolate intergenic transcripts; those deemed genic were naturally assigned to a gene, and therefore cannot be non-coding. On average, samples possessed 29,8331 intergenic transcripts. Lists of intergenic transcripts were merged for all samples within each tissue type to generate a list of all unique intergenic transcripts found in each tissue. Subsequent filtering steps were performed on a tissue basis, as opposed to a per sample basis. Intergenic transcripts were then analyzed by the Coding Potential Calculator, being assessed for features indicative of a transcript that becomes a protein, and only those deemed non-coding were retained for further filtering [10]. Across the 6 tissues, there were an average of 13,235 intergenic, non-coding transcripts identified. The remaining transcripts were then filtered by presence of protein domain and sequence homology, using the Protein Family database and NCBI's BLAST respectively [42, 37]. Protein domain homology filtering left 972 transcripts, on average, per tissue and the final filtering step of sequence homology yielded the final list of candidate lncRNA transcripts. Overlapping transcripts were combined to represent contiguous transcripts.

Beginning with over 1.24 billion reads, progressive filtering steps yielded 126 ileum-associated lymph node lncRNA transcripts, 138 jejunum-associated lymph node lncRNA

transcripts, 83 lymphocyte lncRNA transcripts, 25 bone marrow lncRNA transcripts, 40 spleen lncRNA transcripts, and 43 thymus lncRNA transcripts.

4.3.2 Comparisons Between lncRNA and Coding Transcripts

When analyzing lncRNA as compared to coding transcripts, patterns of shorter length, lower expression, and fewer exons are often observed. This data set was no different, adding validity to the findings.

When comparing transcriptional length profiles, all 6 lncRNA profiles were demonstrably shorter than all transcripts found in each tissue type. Overall, *Bos taurus* transcripts ranged from 53 base pairs to 2,246,270 base pairs in length, whereas identified lncRNA transcripts ranged from 261 base pairs to 162,258 base pairs long (it should be noted that, by definition, lncRNA transcripts must be at least 200 base pairs long). Ileum lymph node tissue transcripts were, on average, 45,400 base pairs long, ileum lymph node lncRNA transcripts ranged from 291 to 84,911 base pairs in length, averaging 6,728 bases. Jejunum lymph node transcripts averaged 44,653 base pairs long, and jejunum lymph node lncRNA transcripts ranged from 263 base pairs to 84,842 base pairs, averaging 44,653 base pairs. Lymphocyte lncRNA transcripts ranged from 249 bp to 110,644 bp, averaging 10,482 base pairs. This is in contrast to the average 40,966 bp all lymphocyte transcripts measured. Marrow lncRNA transcripts ranged from 359 base pairs to 53,293 bases, averaging 7,875 bases long, as compared to the average transcript length of 40,043 bases in all marrow transcripts. Spleen tissue transcripts averaged 21,135 base pairs long, whereas spleen lncRNA transcripts ranged from 278 bases to 36,774 bases, averaging 5,199 bases. Finally, thymus tissue lncRNA transcripts averaged 10,728, spanning 261 bases to 162,258 bases, notably shorter than the 20,769 base pairs. 4.1

Along with demonstrating shorter transcript lengths when compared to the whole

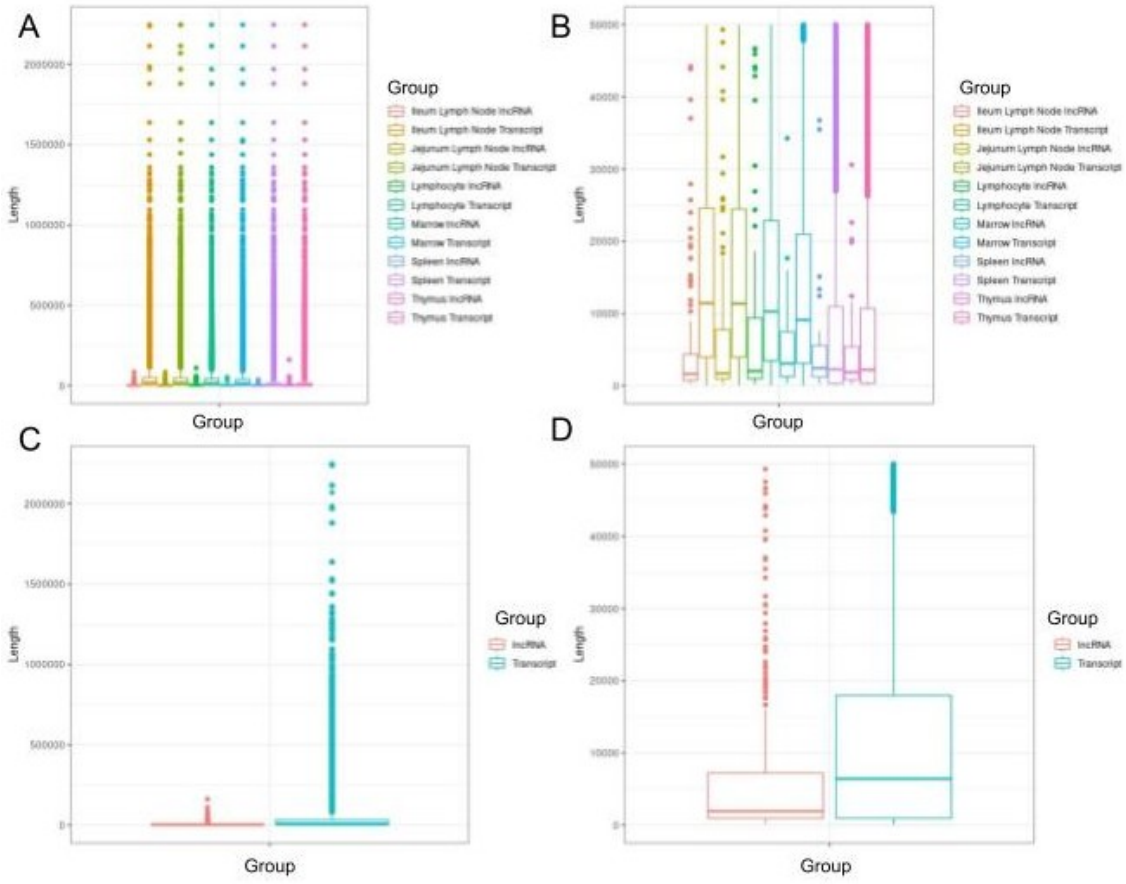

Figure 4.1: Length distribution of various transcripts. (A) Box plot representing length of transcripts in each tissue, containing both all transcripts and lncRNA transcripts only as measured in basepairs. (B) Truncated box plot representing length of transcripts in each tissue, containing both all transcripts and lncRNA transcripts only, measured in basepairs. This is the same plot as Figure 1A, with the y-axis being cut off at 50,000 basepairs. (C) Box plot representing length of all transcripts pooled together and all lncRNA transcripts pooled together. All lengths measured in basepairs. (D) A truncated box plot demonstrating both all transcripts and all lncRNA transcripts pooled together, as measured in base pairs. This is the same plot as Figure 1B, with the y-axis excluding transcripts over 50,000 basepairs.

genome, lncRNA generally demonstrate lower expression levels than their coding transcript counterparts. lncRNAs found in immune system associated tissues appear to be no different. Overall, transcripts ranged in expression from nearly 0.0 to 404,966 transcripts per million (TPM); however, across all transcripts in all samples, expression averaged 37.62 TPM.

When analyzing lncRNA expression, transcripts ranged from 0.006 to 372.108 TPM, with an average expression level of 4.33 TPM when pooling all lncRNA identified in all tissue types. More specifically, ileum lymph node lncRNA transcripts averaged 6.76 TPM, jejunum lymph node tissue averaged 7.08 TPM, lymphocyte tissue lncRNA averaged 1.95 TPM, marrow lncRNAs showed an average of 3.12 TPM, spleen tissue lncRNA averaged 1.59 TPM, and thymus lncRNA averaged 5.45 TPM. These expression levels are markedly lower than all transcripts. All ileum lymph node transcripts averaged 31.62 TPM, jejunum lymph node transcripts averaged 31.59 TPM, lymphocyte transcripts averaged 34.28 TPM, marrow transcripts averaged 36.27 TPM, spleen transcripts averaged 46.72 TPM, and thymus transcripts averaged 45.27 TPM. Overall transcript expression showed anywhere from a 4.46-fold increase in expression to a 29.38-fold increase in expression; this follows with lncRNA expression level trends overall. 4.2

A third characteristic of lncRNA, as compared to the genome on the whole, is a fewer number of exons per transcript. Immune system associated tissue lncRNA appears to follow this trend as well. Overall, across all tissue types, the genome averaged 5.1 exons per transcript; lncRNA transcripts averaged 2.79 exons per transcript. Ileum lymph node tissue averaged 5.74 exons per transcript across the genome and 2.09 exons per transcript in lncRNA. Jejunum lymph node tissue averaged 5.79 exons per transcript overall and 2.64 exons per transcript in lncRNA. Lymphocyte tissue showed 7.23 exons per transcript across the genome and 3.25 exons per transcript in lncRNA. Bone marrow demonstrated 6.34 ex-

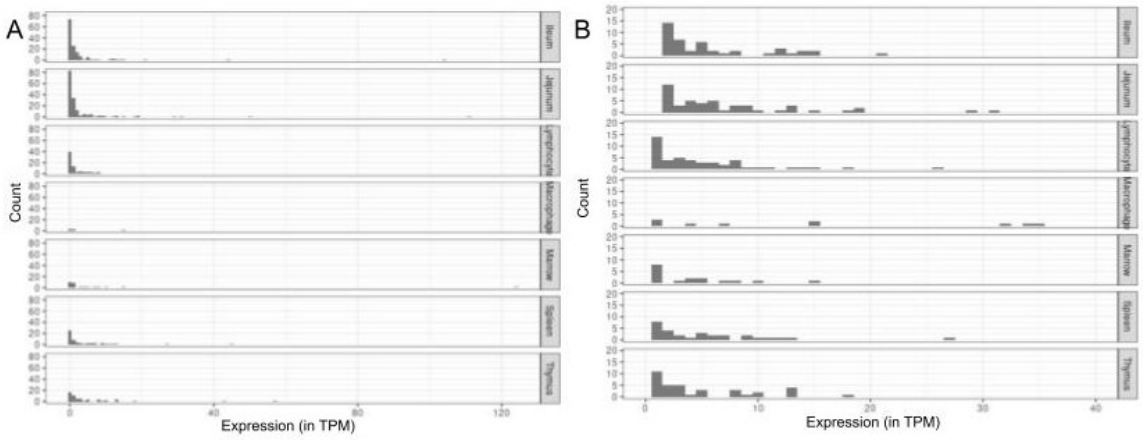

Figure 4.2: Expression of transcripts. (A) As measured in TPM, expression level of all transcripts in all investigated tissues. It should be noted, although many transcripts show lower expression, the x-axis range extends to over 120 TPM. (B) As measured in TPM, expression level of lncRNA transcripts from all investigated tissues. The x-axis of this figure ends at 40 TPM, demonstrating that lncRNA are expressed at lower levels than all transcripts.

ons per transcript overall and 3.56 exons per transcript in lncRNA transcripts. Spleen tissue had 2.81 exons per transcript and spleen lncRNA showed 2.71 exons per transcript; thymus tissue averaged 2.71 exons per transcript and 2.46 exons per transcript in lncRNA. Although the difference between exons per transcript in all transcripts and lncRNA was much smaller in thymus and spleen tissue, these tissues showed fewer exons per transcript overall and the pattern of reduction in exons per transcript was still demonstrated.

4.3.3 Common lncRNA and Differential Expression

Although a total of 455 lncRNA were detected, they represented only 364 unique regions, indicating there are several lncRNA common between various tissues. Overall, 75 lncRNA transcripts were in 2 or more tissues.

When comparing tissue on a 1-to-1 basis, lncRNA found in the 2 lymph node tissues demonstrated the most overlap, sharing 34 lncRNA transcripts. Lymphocyte tissue also

Table 4.1: Quantification of lncRNA transcripts that were common between 2 tissue types.

	Jejunum	Lymph	Marrow	Spleen	Thymus
Ileum	34	18	3	5	1
Jejunum		22	3	7	1
Lymph			1	5	1
Marrow				0	1
Spleen					8

demonstrated several overlapping lncRNA transcripts. Thymus tissue lncRNA showed the least overlap with other tissues, excluding spleen tissue, which presented a handful of overlaps. Of the 75 lncRNA transcripts found to be common to multiple tissues, 59 were found only in 2 tissues (Table 4.1).

When building more complex comparisons, several more lncRNA appeared in more than 2 tissues. 15 lncRNA transcripts were found to be present in 3 tissues. 8 lncRNA transcripts were found in ileum lymph node, jejunum lymph node, and lymphocyte tissue; 2 lncRNA transcripts were found in ileum lymph node, jejunum lymph node, and spleen tissue; 2 lncRNA transcripts were found in jejunum lymph node, lymphocyte, and spleen tissue; 1 lncRNA transcript was present in ileum lymph node, jejunum lymph node, and marrow tissue; 1 lncRNA transcript was found in ileum lymph node, lymphocyte, and marrow tissue; and finally, 1 lncRNA transcript was present in jejunum lymph node, spleen, and thymus tissue.

There was a single lncRNA that was more pervasive, being present in 4 of the 6 tissue types. This lncRNA was found in ileum lymph node, jejunum lymph node, lymphocyte, and spleen tissue.

The aforementioned findings are demonstrated in Figure 3, where most transcripts (83.52%) are found in only 1 tissue type; 12.97% of identified lncRNA are found in exactly 2 tissue types, 3.30% of lncRNA are found in exactly 3 tissue types, and a singular lncRNA

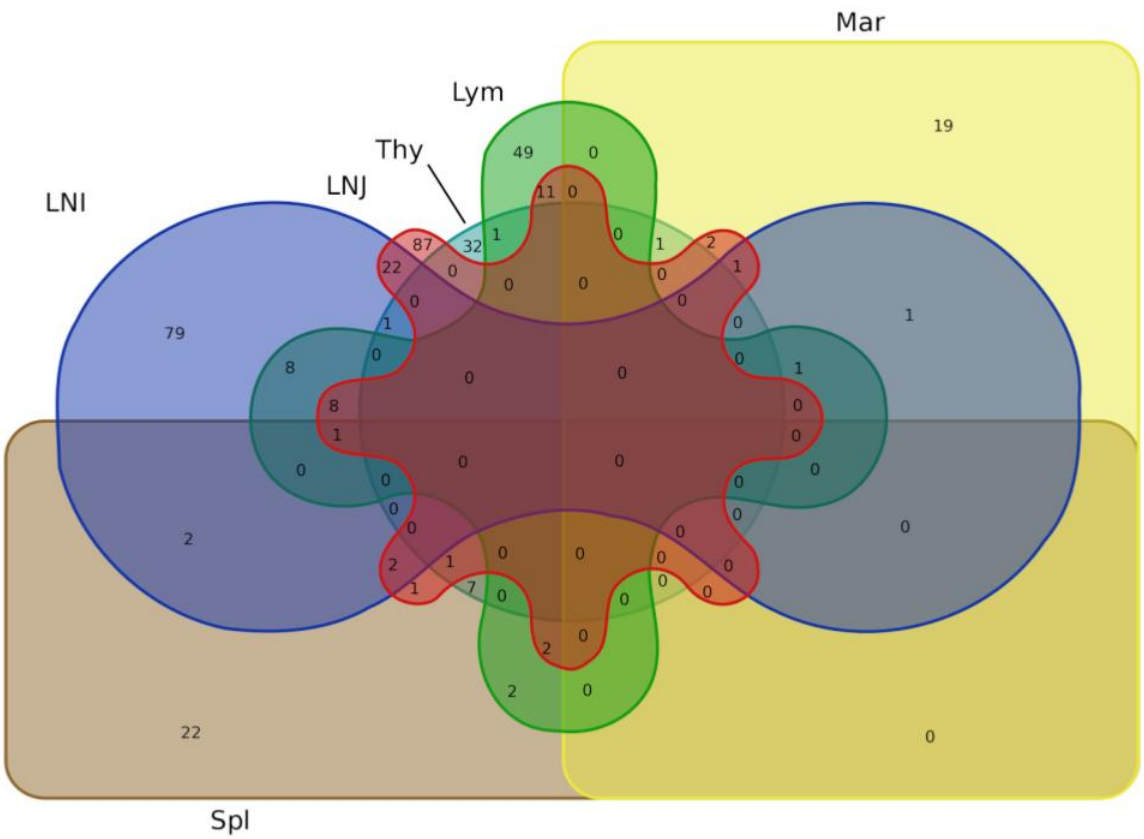

Figure 4.3: Venn diagram of overlapping lnc RNA. When comparing all lncRNA transcript profiles, there were several transcripts present in several tissues. Comparisons and overlapping transcripts are presented here.

was found in 4 tissue types (this represented 0.22% of identified lncRNA). There were no lncRNA transcripts found that were common to 5 of the 6 analyzed tissue types or all 6 tissue types (Figure 4.3)

Once common transcripts were identified, they were able to be analyzed for differential expression.

Of the 59 lncRNA found in exactly 2 tissues, 25 lncRNA transcripts were statistically significantly differentially expressed based on a heteroscedastic student t-test. Although the number of differentially expressed lncRNA found in ileum-associated lymph node tis-

sue and jejunum-associated lymph node tissue was the highest of any pairing, the ileum-associated lymph node tissue and marrow commonalities showed the highest ratio of differentially expressed lncRNA as compared to all common lncRNA with 100% of the identified common lncRNA being differentially expressed (although it should be noted there was only 1 lncRNA in common between these two tissues). This was followed by the commonalities between jejunum-associated lymph node tissue and lymphocyte tissue, where 64% of the lncRNA transcripts—7 out of the 11-transcripts were statistically differentially expressed. 50% of common lncRNA transcripts found in ileum-associated lymph node tissue and lymphocyte tissue were differentially expressed (4 out of 8). The same 50% ratio was found in jejunum-associated lymph node tissue and bone marrow common lncRNA transcripts (1 out of 2). 45.5%, or 10/22, ileum-associated lymph node and jejunum-associated lymph node common lncRNA transcripts were differentially expressed, and spleen and thymus tissue common lncRNA transcripts showed the lowest ratio of significantly differentially expression: 2 out of 7, or 28.6%.

The most significantly differentially expressed lncRNA found in 2 tissues only was the single lncRNA found in ileum-associated lymph node tissue and bone marrow tissue.

The 15 lncRNA transcripts found in 3 tissue types were next analyzed. Heteroscedastic student t-tests were used to compare all combinations of comparisons (i.e. tissue 1 to 2, tissue 1 to 3, and tissue 2 to 3). Of these 45 combinations, 26 showed statistically significant differential expression, with 12 of the lncRNA showing statistically significant differential expression in at least 1 of the comparisons. Only 1 common lncRNA showed differential expression in only 1 comparison, 8 common lncRNA showed significant differential expression in 2 of the comparisons, and 3 of the common lncRNA showed differential expression in all 3 comparisons. 2 of the common lncRNA whose expression was statistically different in all 3 tissue types were found in ileum-associated lymph node tissue, jejunum-associated lymph node tissue, and lymphocyte tissue; the other lncRNA was found

in jejunum-associated lymph node tissue, lymphocyte tissue, and spleen tissue.

The final lncRNA investigated for differential expression in multiple tissues was the lncRNA found in 4 tissues: ileum-associated lymph node tissue, jejunum-associated lymph node tissue, lymphocyte tissue, and spleen tissue. When comparing each tissues' expression profile using a heteroscedastic student t-test, 4 of the 6 comparisons showed statistically significant differential expression. Expression levels were not statistically significantly different between ileum-associated lymph node tissue/jejunum-associated lymph node tissue profiles or lymphocyte tissue/spleen tissue profiles. Ileum-associated lymph node tissue/lymphocyte tissue, ileum-associated lymph node tissue/spleen tissue, jejunum-associated lymph node tissue/lymphocyte tissue, and jejunum-associated lymph node tissue/spleen tissue profiles all showed statistically significant differential expression for the common lncRNA.

All lncRNA that demonstrated statically relevant expression profile differences in at least 1 comparison were noted and retained in a dedicated list for future analysis. Should "differentially expressed lncRNA" be mentioned further, it is these transcripts that are being referenced. The inverse will also hold true. Those data sets noted to be excluding differentially expressed lncRNA will not include those with statistically significant differential expression in at least 1 comparison.

4.3.4 Sequence Conservation of lncRNA

Beginning by converting genetic coordinates to their human genome counterparts, conservation scores for each region of interest were calculated based on sequence similarity across 100 different species.

When converting Bos taurus genomic regions into human genome coordinates, patterns emerged almost immediately. The "Coding" group for each tissue sample showed an

average successful conversion rate of 79.48%, whereas intergenic regions averaged a conversion rate of 54.62%, and lncRNA showed an average of 33.70% successful conversion; this demonstrated that there is likely higher levels of sequence conservation in coding regions as compared to noncoding regions (both the more general intergenic regions and the more specific lncRNA regions). These preliminary findings align with lncRNA patterns of lower sequence conservation.

Once profiles were converted to represent human coordinates, Phastcon analysis could be performed, measuring sequence conservation for each profile more thoroughly.

Using a sequence conservation matrix comparing human genetic regions to 99 other vertebrate genomes (ranging in complexity), successfully converted regions were scored between 0 and 1. Scores closer to 1 demonstrate more sequence conservation across all 100 species than scores closer to 0. Overall, lncRNA demonstrated the lowest score for conservation, with an average score per sample ranging from 0.05 to 0.13 and overall average score per lncRNA region being 0.09. Only showing slightly higher conservation scores, intergenic regions, on average, scored 0.10, with per sample averages ranging from 0.10 to 0.11. Unsurprisingly, coding regions were the most conserved. Regions, on average, scored 0.16, with per sample averages ranging from 0.15 to 0.16.

These patterns follow the trends seen in other lncRNA analyses, where lncRNA regions show less conservation when compared to coding regions. It should be noted, however, that even though the intergenic regions showed slightly more conservation on average, when graphed, conservation scores on intergenic regions pooled at a lower value than lncRNA regions. This demonstrates that lncRNA regions tend to score higher than a standard intergenic region, indicating the lncRNA do show some level of conservation above other noncoding regions. This trend was seen in all tissues (Figure 4.4).

When the conservation scores of investigated lncRNA regions are plotted against their

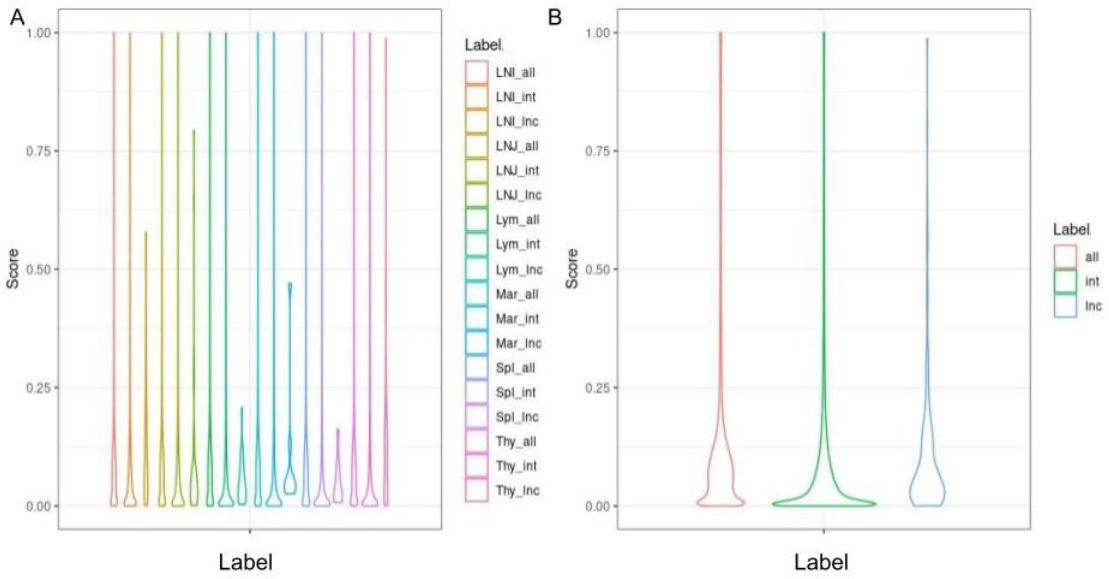

Figure 4.4: PhastCons scores across transcript profiles. (A) Violin plot of Phastcon conservation score for each profile in each tissue investigated. (B) Violin plot of Phastcon conservation scores for each profile type pooled together: all transcripts from all tissues, all intergenic transcripts from all tissues, and all lncRNA from all tissues.

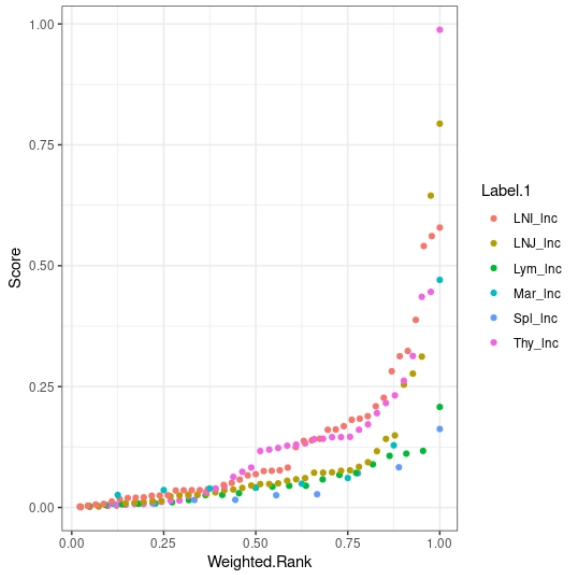

Figure 4.5: Scatter plot of lncRNA PhastCons scores from each tissue investigated, Phastcon score is plotted against the weighted rank of each lncRNA from each tissue. These scores do not have a unit.

relative ranks, it is clear to see most lncRNA regions are not well conserved, with most falling below the 0.25 score line. However, some do score higher, with a small number of regions scoring over 0.50. There does not appear to be a specific tissue type that demonstrates higher sequence conservation than another. Regions that scored above 0.25 were isolated and retained as "highly conserved regions" and further analysis was performed. This analysis will be reported in a future section (Figure 4.5)

In addition to comparing lncRNA regions to intergenic and coding regions, lncRNA regions were compared based on the number of tissues the region was present in. Many lncRNA found in multiple tissues lacked a human genome equivalent for comparison, but those that were able to be analyzed demonstrated the pattern of conservation appeared independent of the number of tissues the region is found within (Figure4.6)

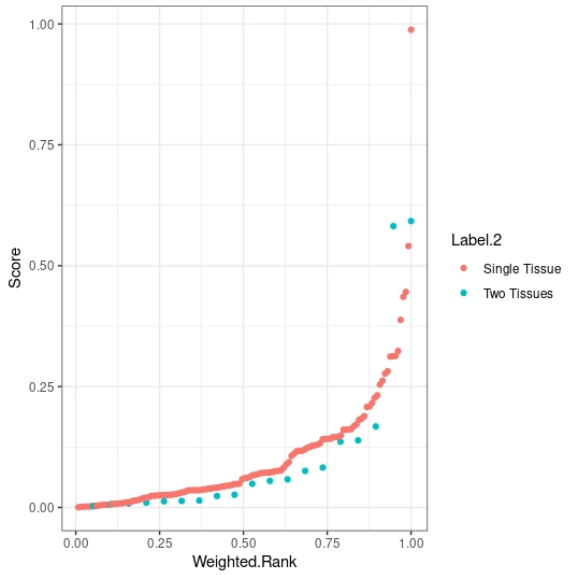

Figure 4.6: Scatter plot of lncRNA PhastCons score from 2 groups: lncRNA found only in 1 investigated tissue and lncRNA found in multiple tissues. Scores are unitless and plotted against weighted rank for each lncRNA.

4.3.5 Transcriptional Annotation of lncRNA Based on Sequence Homology

Given the nature of lncRNA, there is little information about the function of individual lncRNA transcripts; assigning a definitive function to each lncRNA transcript is also very difficult as their function is wide reaching. However, reasonable hypotheses can be made via many methods. The first method employed is based upon sequence conservation. Using the human genome coordinates from Phastcon analysis, all lncRNA found in at least 2 tissues were retained for annotation, as well as the top 25% of each tissue type lncRNA. This left 55 lncRNA regions to be analyzed for the presence of annotated genes. Of these regions, 41 regions have been assigned to various genetic features, ranging from unresearched, named loci to well researched genes.

Several of the associated genes related to transcriptional control and/or regulation—GATA6, a transcriptional activator, being the most represented gene. GATA6 appeared in regions

found in only ileum-associated lymph node tissue, only jejunum-associated lymph node tissue, and multi-tissue regions. Other found genes that are associated in some way with gene expression control include CENPA, MCM8, IKZF1, and ASCC2. Many lncRNA have demonstrated a role in gene expression regulation by acting in several fashions including interacting with RNA-binding proteins, causing mRNA degradation, stabilizing sequence-binding factors, and creating complexes with various transcripts and proteins, so these gene associations are encouraging findings for building a better understanding of lncRNA and their role within the biological system [49]. Most notably, within a thymus tissue associated lncRNA, was the MTCP1 gene; MTCP1 enhances the phosphorylation and activation of AKT1 and AKT2 and has been implicated in the t(X;14) translocations related to mature T-cell proliferation [20]. Mature T-cells are crucial to adaptive immune responses, and the thymus is the primary organ for T-cell development; this gene may have interesting implications for ensuring long-term health in high value commercial animals [51].

All common lncRNA were converted to their human coordinates, where possible, and analyzed for overlapping genes. Of these retained common lncRNA-human converts, differentially expressed lncRNA were isolated as well, and analyzed for genes within the regions. This led to the identification of 3 genes: NPLOC4, SETD2, CYTH3. NPLOC4 functions within a complex to remove misfolded proteins from the endoplasmic reticulum; SETD2 is a histone methyltransferase targeting H3K36me2 associated with many functions including DNA repair, heterochromatin maintenance, and epigenetic transcriptional activation; and CYTH3 is a PSCD family protein, this family of proteins function to regulate protein sorting and membrane trafficking [36, 45, 11].

4.3.6 Genetic Co-expression Correlation and Annotation

Another method of hypothesizing lncRNA function comes from genetic correlation. Although correlation does not indicate causation and does not definitively link lncRNA to genes, statistically significant correlation may indicate the lncRNA and gene have a functional relationship (this relationship may be positive or negative).

These relationships were investigated by generating a correlation matrix, correlating expression levels of lncRNA as compared to coding genes. To avoid false hits, transcripts were removed if they did not show expression in at least 2 samples. This left 112 ileum-associated lymph node lncRNAs, 117 jejunum-associated lymph node lncRNAs, 74 lymphocyte lncRNAs, 30 spleen lncRNAs, and 35 thymus lncRNAs. Although 19 bone marrow lncRNAs remained for analysis, the small sample size led to inconclusive results. Therefore, bone marrow was omitted from correlation-based analysis.

The first correlation-based analysis investigated the most statistically significantly correlated gene for each lncRNA. Of the complete 369 lncRNA investigated, 53 lncRNA are associated with a novel transcript. Of these novel transcripts, 27 lncRNA correlated to a novel gene at a statistically significant level of $p<0.05$, 20 lncRNA correlated to a novel gene at a statistically significant level of $0.05<p<0.10$, 6 lncRNA correlated to a novel gene at a non-statistically significant level. Of the 316 lncRNA that associated with a known Bos taurus gene, 241 lncRNA correlated with a known gene with a significance level of $p<0.05$, 51 lncRNA correlated with a known gene with a significance level of $0.05<p<0.1$, and 24 lncRNA correlated with a known gene but not at a statistically significant level, based on a p-value less than 0.1.

When analyzing the 112 ileum-associated lymph node lncRNA were correlated with genes, 96 lncRNA associated with a known gene. Of these 96 associations, 90 were deemed statistically significant at $p<0.05$. Using annotations performed in humans, common func-

tions of these lncRNA-correlated genes included various binding activities, histone modifications, protein modification/transport, signal transduction, and transcriptional activity. The most notable associated genes are CR2 (complement C3d receptor 2) and IRAK3 (interleukin 1 receptor associated kinase 3). In humans, CR2 encodes a membrane receptor protein that binds B and T lymphocytes to Epstein-Barr virus; this virus shows high conservation levels to the bovine herpesvirus 4 [11, 32]. Bovine herpesvirus 4, also known as Movar virus, is generally subclinical but can result in reproductive diseases in cattle [39, 46]. IRAK3 is integral to the Toll/IL-R immune signal transduction pathway, negatively regulating this crucial immune response pathway [24]. Toll/IL-R is responsible for the cascade leading to proinflammatory cytokine and chemokine innate immune response [35].

117 jejunum-associated lymph node lncRNAs were analyzed: 19 lncRNA associated with novel genes at various levels of statistical significance; 76 lncRNA associated with a known gene at a statistical significance level of $p<0.05$; and 22 lncRNAs associated with a known gene at a significance level of $0.05<p<0.1$. The most statistically significantly correlated genes to this lncRNA set ran a wide range of functions, with signaling/signal transduction and transcriptional connections being the most represented. Based on more extensive human research, the most notable genes correlated with the jejunum-associated lymph node lncRNA included IL17RB, C7, and SHFL. IL17RB, interleukin 17 receptor B, is a cytokine receptor that has been shown to mediate NF-kappaB activation and IL8 production; NF-kappaB regulates innate and adaptive immune function and IL8 attracts white blood cells during immune response [21, 31, 7]. C7, complement 7, is a serum glycoprotein component of the terminal complement pathway of the innate immune response [8]. SHFL, shiftless antiviral inhibitor of ribosomal frameshifting, is a gene that inhibits viral replication in response to interferon stimulation by inhibiting programmed –1 ribosomal frameshifting. This gene inhibits replication of many RNA viruses of note, such as HIV, COVID-19, Zika virus, and dengue fever [47].

Many of lymphocyte lncRNA associated genes were not statistically significant, with 40 of the 74 identified correlations being not statistically significant at $p<0.05$. Of the 34 statistically significant correlations, 4 were between lncRNA and novel genes, and the functions of the remaining 30 included cell adhesion, RNA modification, and a wide variety of biological processes. Notable gene associations uncovered included ST6GAL1 and THNSL2. ST6GAL1, ST6 beta-galactosidase alpha-2,6-sialyltransferase 1, is a membrane protein involved in the generation of differentiation antigens HB-6, CD75, and CD76; the generation of these differentiation antigens allows for differentiation of lymphocytes, altering their functions [48, 4]. THNSL2, threonine synthase like 2, has been shown to possess an alternative splice form in mice that acts as a cytokine and induces production of IL6 in osteoblasts; IL6 functions in the maturation of B cells, as well as inflammatory immune responses [22, 52].

30 spleen associated lncRNAs were analyzed, and only 1 was correlated to an unknown gene, although this correlation was not statistically significant. Of the 29 known genes correlated to lncRNA, 17 were statistically significant at $p<0.05$. There were no discernable functional patterns in the top correlated genes. The most notable gene associated with these lncRNA was IL9R. IL9R is interleukin 9 receptor, which yields a cytokine receptor protein that mediates IL9; IL9 regulates inflammatory immunity responses [17, 23].

Of the 35 thymus lncRNA investigated, 30 lncRNA/gene correlations were between a lncRNA and known gene, and 28 of those 30 were statistically significant. These genes showed patterns reflecting transcription factors/activators, zinc finger proteins, and other processes. Although interesting genes associated with human breast cancer and starvation responses were represented in this lncRNA/gene set, there were no intense immune system related genes found within thymus lncRNA/gene correlations.

Overall, genes that correlated significantly with lncRNA showed patterns of protein, DNA, and ion binding activity, apoptosis regulation, cell growth and proliferation regu-

lation, signal transduction, signaling, transcriptional factors and activators, ubiquitination, and zinc finger protein activity.

4.3.7 Gene Cluster Ontology Based on lncRNA Correlation

Utilizing the same correlation matrices as above, clusters of statistically significantly associated genes were generated for each lncRNA. These gene clusters were then analyzed for gene ontology.

Of the 112 ileum-associated lncRNAs, 103 had statistically significant correlated gene clusters. 75 of the 103 clusters associated with at least one biological process in a significant manner. The biological processes associated with each cluster were noted and patterns began to emerge. 4 unique gene clusters showed associations with both the G protein-coupled receptor signaling pathway and the detection of chemical stimulus involved in sensory perception of smell, 3 other clusters showed association with only detection of chemical stimulus involved in sensory perception of smell and 1 other cluster associated with the G protein-coupled receptor signaling pathway only. 7 gene clusters associated with T cell biological processes—costimulation, differentiation, migration, and proliferation. Interestingly, although none of the ileum-associated lncRNA correlated most highly with NF-kappaB associated genes, 1 gene cluster did associate with NIK/NF-kappaB signaling. The gene ontology profile generated for this tissue is by far the most diverse of all 5 tissue types investigated.

Although results between ileum-associated and jejunum-associated lymph node lncRNA were similar in past analysis, they began to diverge in gene cluster ontology analysis. Of the 117 jejunum-associated lymph node lncRNA, only 82 possessed statistically significant gene clusters; 48 of those clusters were not associated with a biological process. When analyzing all statistically significant biological process associations, the 34 gene clusters only

associated with 6 unique processes: the G protein-coupled receptor signaling pathway, detection of chemical stimulus involved in sensory perception of smell, the regulation of the cellular biosynthetic process, the regulation of the macromolecule biosynthetic process, the regulation of nucleobase-containing compound metabolic process, and transmembrane transport. The G protein-coupled receptor signaling pathway was present in all clusters at a significant level. The detection of chemical stimulus involved in sensory perception of smell was present in 33 of the 34 clusters. The regulation of cellular biosynthetic process, the regulation of the macromolecule biosynthetic process, the regulation of nucleobase-containing compound metabolic process, and transmembrane transport were present in 5, 4, 3, and 2 clusters respectively.

34 lymphocyte lncRNA gene clusters were generated; of the 74 lncRNA gene associations, 40 were not statistically significant. 26 of the 34 clusters associated with a known biological process. These clusters represented processes similar to jejunum-associated gene clusters. All 26 clusters associated with the G protein-coupled receptor signaling pathway at a significant level, 25 clusters showed association with the detection of chemical stimulus involved in sensory perception of smell, 11 clusters showed association with the regulation of cellular biosynthetic process, 2 clusters associated with regulation of macromolecule biosynthetic processes, and 1 cluster was associated with transmembrane transport. The similarities between lymphocyte associations and jejunum-associated lymph node associations are staggering.

17 of the 30 spleen lncRNA showed clusters of genes that were statistically significantly correlated with the lncRNA, and 10 of those clusters were associated with at least 1 biological process. Detection of chemical stimulus involved in sensory perception of smell and the G protein-coupled receptor signaling pathway were represented in 8 clusters, regulation of DNA-templated transcription was represented in 5 clusters (1 of which was a positive regulation and 1 of which was a negative regulation), the regulation of cellular biosynthetic

processes was associated with 2 clusters, the negative regulation of RNA metabolic processes and the regulation of macromolecule biosynthetic processes were each represented in 1 cluster.

Finally, thymus lncRNA gene clusters were generated and investigated. 31 of the 35 lncRNA generated statistically significant clusters and 19 of those 31 clusters associated with a biological process. As is to be expected based on previous results, the G protein-coupled receptor signaling pathway was heavily represented, being found in 13 of the 19 clusters. Regulation of DNA-templated transcription was well represented, appearing in 8 of the clusters. Regulation of cellular biosynthetic processes was present in 6 clusters. Detection of chemical stimulus involved in sensory perception of smell, regulation of gene expression, and regulation of macromolecule biosynthetic processes were all represented in 5 clusters. The regulation of nucleobase-containing compound metabolic processes was found to be associated with 1 cluster.

Across all 5 tissue types, there are a small number of biological functions that are repeatedly represented as gene cluster ontology associations. The following processes were shown to have an association with multiple clusters in all tissues: the G protein-coupled receptor signaling pathway, detection of chemical stimulus involved in sensory perception of smell, regulation of cellular biosynthetic processes, and regulation of macromolecule biosynthetic processes.

The G protein-coupled receptor signaling pathway is aptly named as it relies on the G protein-coupled receptors (GPCRs). These receptors are a large group of membrane-bound receptors that function to filter incoming energy, peptide, lipid, sugar, and protein messages; there are over 1,000 distinct GCPRs in humans and GPCRs have been shown to be highly conserved across eukaryotes overall [18]. GPCRs are emerging as crucial players in the development and activation of macrophages, as well as demonstrating a role in the spatiotemporal control of leukocytes during immune responses [29, 53]. Given the wide-

reaching nature of these receptors and the signaling pathways they trigger, their repeated presence in immune system associated tissues is not surprising.

The detection of chemical stimulus involved in sensory perception of smell was also highly represented in immune related tissues; interestingly, OR6B3, a smell receptor, was identified as the most statistically significantly correlated gene to a lymphocyte lncRNA, as well. Research has demonstrated that in response to "disgusting" odors, human participants showed a demonstrable increase in TNFalpha in their saliva, indicating a preparatory immune response. TNFalpha has been shown to play a role in pathogen defense and chemosensory perception, directly connecting smell to the immune system from a biological process perspective but given these "disgusting" odors are representative of disease cues, the increased salivary TNFalpha provides more direct evidence linking sense of smell and immune response [1].

Both the regulation of cellular biosynthetic processes and the regulation of macromolecule biosynthetic processes are very large, encompassing processes. Cellular biosynthetic processes are those that generate the formation of substance and macromolecule biosynthetic processes are those that generate a macromolecule [33, 9]. The comprehensive nature of these processes makes it near impossible to definitively link them to specific immune system functions; the inverse of that caveat is that it is almost certain these biological processes do play a key role in some immune system functions.

4.3.8 Integration of GWAS Data Using SNP Heritability Enrichment Analysis

In order to integrate lncRNA with complex traits of interest, GWAS data from 27,000 bulls was utilized to analyze the enrichment of SNPs/INDELs overlapping with the identified lncRNA. 3 million SNPs/INDELs were cross-referenced with the 455 identified lncRNA

to investigate heritability enrichment for 4 traits: daughter pregnancy rate, livability, milk production, and stature. These traits are representative of more complex traits: daughter pregnancy rate representing fertility and reproductive value, livability representing health, milk production representing production value, and stature representing overall body condition. MINQUE for partitioning heritability (MPH) link was used to measure SNP heritability enrichment; this was measured as a ratio of per-SNP heritability near identified lncRNA as compared to per-SNP heritability in the genome overall. This analysis utilized 2,795,435 genome-wide SNPs/INDELs and 4326 immune tissue associated lncRNA SNPs/INDELs. Significant enrichments of per-SNP heritability near lncRNA was found in 2 of the 4 investigated complex traits: daughter pregnancy rate (7.43x, P=0.047) and stature (8.98x, P=0.0057). Enrichment was not significant in milk production and livability. These findings are interesting; it is logical to see enrichment in daughter pregnancy rate, a sick cow cannot sustain a pregnancy; however, enrichment was not seen in livability, a measure of overall health. Enrichment was seen in stature, a trait measuring overall body condition and conformation, indicating there is some connection between immune system lncRNA and overall condition of cattle.

Table 4.2: Enrichment of immune-related tissue lncRNA based on large-scale cattle GWAS data. This analysis is completed with the MPH software. Daughter pregnancy rate (DPR) represents reproduction value, livability representes health, milk represents production value, and stature represents body conformation. This table reports enrichment of each trait based on per-SNP heritability, standard error of estimate of per-SNP heritability enrichment estimate, and p-value for per-SNP heritability enrichment estimate.

	DPR			**Livability**			**Milk**			**Stature**		
	Enr	SE	P	Enr	SE	P	Enr	SE	P	Enr	SE	P
Imm	7.43	3.83	0.047	-0.85	4.93	0.65	-1.50	1.91	0.90	8.98	3.15	0.006

4.3.9 Immune-system Related lncRNA as Compared to Other lncRNA Profiles

Given lncRNA exists both ubiquitously and in a tissue-specific manner, patterns may emerge when comparing immune tissue associated lncRNAs to other tissue lncRNA profiles. The lncRNA profiles across all 5 immune system tissues, Bos taurus rumen tissue (both pre and post weaning), and Bos taurus mammary tissue (both from dry and lactating samples) all follow the same lncRNA profile characteristics repeatedly reported: transcripts are shorter, show less expression, and less sequence conservation than coding transcripts. This both supports the findings in the 2 previous studies and this study and reinforces the lncRNA characteristics that help further the knowledge base regarding lncRNA.

When integrating gene ontology analysis for mammary tissue lncRNA profiles and immune system lncRNA profiles, the detection of chemical stimulus involved in sensory perception of smell appeared across several profiles. As previously stated, this process is found in all 5 immune tissues investigated. This ontology association was also seen in both dry and lactating mammary tissue (the process associated with more lncRNA in the lactating mammary tissue). The sense of smell is multifaceted and affects many biological processes. It is not surprising to uncover that it is potentially influenced by lncRNA in addition to coding genes.

Biosynthetic processes were also seen in both immune system lncRNA gene cluster ontology and lactating rumen tissue lncRNA gene cluster ontology; the broadness of these processes means this finding is not particularly surprising, but it is noteworthy that it was more represented in tissue profiles where new molecules need to be created. Immune system tissues need to produce things like cytokines and other immune system molecules and lactating mammary tissue needs to produce milk.

Comparison of most statistically significantly correlated genes also revealed common-

alities between the immune system lncRNA profiles and mammary lncRNA profiles. Common functions of all lncRNA investigated included cell growth, cell cycle control, cell proliferation, binding activities, signaling, and mRNA/tRNA processing and control. Given lncRNA has repeatedly been implicated in various regulatory processes, these commonalities are not surprising.

These commonalities create many opportunities for further investigation, lncRNA are very wide reaching and each connection generates the start of uncovering the role each lncRNA plays.

4.4 Discussion

The aim of this study was to collect and analyze the genetic data from immune system related tissues, in pursuit of lncRNA. The purpose is to identify lncRNA in these tissues and begin the process of assigning roles to identified transcripts. Findings from these analyses may improve genetic breeding profiles for high value production animals, as well as increase the knowledge base about lncRNA in *Bos taurus*, both general and in relation to immunology. Disease and sickness have long been devastating parts of agriculture and likely will continue to burden farmers and agriculturalists. Increasing what is known about immunology and immune responses can decrease the negative impact on production animals, saving farmers time, money, and resources. Using filtering techniques to remove transcripts that were either known genes or likely to be genes, 455 candidate lncRNA were identified from 6 unique immune system associated tissues. When looking at characteristics of these identified transcripts, lncRNA trends were followed; transcripts were shorter, expressed at lower levels, had fewer exons, and showed less sequence conservation than coding transcripts. These findings both validate the identified transcripts as true lncRNA, as well as support the patterns seen in previous studies investigating lncRNA. The 455

identified transcripts were represented by 126 ileum-associated lymph node lncRNA transcripts, 138 jejunum-associated lymph node lncRNA transcripts, 83 lymphocyte lncRNA transcripts, 25 bone marrow lncRNA transcripts, 40 spleen lncRNA transcripts, and 43 thymus lncRNA transcripts. This demonstrates lncRNA are present in immune related tissues. 16.48% (75) of identified lncRNA transcripts were common to at least 2 tissues; when analyzed for differential expression, 36 lncRNA showed differential expression deemed statistically significant between at least 2 of the analyzed tissues. The first conclusion of these results is the presence of common lncRNA between tissue types indicates lncRNA are not exclusively acting in a tissue specific manner: some lncRNA are tissue specific and some are more ubiquitous. This is in line with previous research done by Jiang et al [25]. Another conclusion drawn from these findings pertains to the variable expression patterns. Some lncRNA are expressed across multiple tissues at consistent levels, likely acting in similar ways across various tissue types. They may be playing some kind of basal support role that is unrelated to tissue type. Other lncRNA are expressed in multiple tissues but at statistically significantly different levels, being either up or down regulated depending on tissue type. Given the cascading nature of biological (and especially immune) processes and associations of lncRNA with gene expression control, these differentially expressed lncRNA could be controlling tissue specific processes. Sequence homology analysis demonstrated lncRNA transcripts showed sequence similarity predominantly to genes related to expression regulation, such as GATA6. The most notably similar gene was MTCP1, which plays a role in a pathway linked to T-cell proliferation. This lncRNA was found in the thymus, where such maturation occurs. Correlation based on expression was also done to hypothesize lncRNA function. When the most statistically significantly correlated gene for each lncRNA was analyzed, 241 correlation pairs were uncovered. Most notably, genes associated with immune cell maturation and response were seen repeatedly. Correlated genes functioned in many ways, including binding to and inhibiting viruses, playing roles

in immune response pathways, acting as cytokine receptors, and functioning in immune cell differentiation, to name a few. Taking correlation analysis a step further, gene ontology clustering was done to associate lncRNA with biological functions. This revealed several common functions that repeatedly associated with highly correlated gene clusters: the G protein-coupled receptor signaling pathway, detection of chemical stimulus involved in sensory perception of smell, regulation of cellular biosynthetic processes, and regulation of macromolecule biosynthetic processes. While regulation of cellular and macromolecule biosynthetic processes are both wide reaching processes, the association between G protein-coupled receptor signaling pathways and detection of chemical stimulus involved in sensory perception of smell are both very interesting findings given the tissues used in this study. G protein-coupled receptors have been implicated as critical for macrophage development and spatiotemporal control of leukocytes [29, 53]. Further investigation into the associated lncRNA could reveal more about the many processes involved in immune responses, allowing for better understanding of disease control. The repeated association between gene clusters and smell, as well as one lncRNA correlating more significantly with a smell related gene (OR6B3), is interesting when looked at through the lense of immune response. Smell is a first line protective defense against food-borne pathogens, many creatures (humans included) will turn their nose up to foul smelling food. It is an instinct to avoid rancid, dangerous food. Research has also shown there is an immune response upon smell of foul odors, making this connection even more interesting [1]. Could selecting animals with a more sensitive nose lead to healthier animals overall? Should animals with a more discerning sense of smell be positively selected for if the aim is disease prevention? When analyzing SNPs located near lncRNA transcripts for heritability enrichment of complex traits, both daughter pregnancy rate and stature showed statistically significant enrichment. This demonstrates the lncRNA plays some kind of role in complex traits representing crucial factors like reproductive value and overall composition. Finally, when

comparing the lncRNA profiles of immune system related tissues to previous research, the association with sense of smell appeared again, having been associated with lncRNA in gene ontology analysis done with mammary tissue associated lncRNA.

Overall, these findings begin to uncover more about the role lncRNA plays in immunity and immune response. The identification of these lncRNA and the associations they possess will allow for further research into immune response and immunity in dairy cattle. This can allow farmers to create more resilient, less disease prone animals, saving them precious time, money, and resources.

4.5 Conclusions

In this study, immune system related tissues were investigated for lncRNA transcripts. lncRNA transcripts were uncovered, and they matched with previous findings about lncRNA patterns. A portion of the lncRNA discovered was represented in multiple tissues and showed varying expression profiles. Functional annotations associated these lncRNA with functions ranging from cell cycle control to binding to the more notable immune related processes, implicating lncRNA in immune system maturation and immune responses. Integration of lncRNA with large-scale GWAS data associated lncRNA with complex traits, as well. These findings both increase the knowledge of the immune system in dairy cattle and open the door for further research to be done to connect discovered lncRNA with more specific functions. Overall, this research and subsequent studies can allow for better breeding practices in these crucial production animals.

4.6 Materials and Methods

4.6.1 Data Collection

Using the search function of the SRA database, samples from lymph nodes (ileum-associated and jejunum-associated), lymphocytes, bone marrow, spleen, and thymus tissue in *Bos taurus* were isolated and their SRR IDs recorded. This led to obtaining 39 total samples: 9 ileum-associated lymph node samples, 9 jejunum-associated lymph node samples, 7 lymphocyte tissue samples, 2 bone marrow tissue samples, 6 spleen tissue samples, 6 thymus tissue samples.

Using the fastq-dump tool from the SRA Toolkit, samples were downloaded locally. To ensure all reads from the raw data were downloaded, the split-3 option was used [16].

4.6.2 Sequence Assembly and Mapping

Using Hisat 2.2.1, genomic data was mapped. ARS-UCD1.2 was provided as a reference genome to build the necessary mapping index [28, 2].

Double-iterative mapping was used to maximize accuracy. The first round of mapping was used to establish junctions between transcripts, the second round of mapping utilized these junctions to increase mapping accuracy. Stringtie v2.2.1 was then used to stitch mapped transcripts into an assembled report; this was done once with the ARS-UCD1.2 genome and again with the UMD3.1.1 genome. This was done to ensure all known genes were annotated in samples [50, 6].

This resulted in samples being prepared for filtering.

4.6.3 lncRNA Identification

The first step in lncRNA identidcation was to annotate all transcripts assembled in the previous step. This was done using the CuffCompare 2.2.1 software, which functions to annotate transcripts based on a reference genome [12].

The ARS-UCD1.2 aligned transcripts were compared to the ARS-UCD1.2 refernce genome, the UMD3.1.1 aligned transcripts were compared to the UMD3.1.1 reference genome; this yielded a summary file for each reference genome for each sample.

The CuffCompare software annotates intergenic regions with the ”-u” designation, these regions were retained for further filtering as those transcripts that are intergenic cannot be coding genes. Intergenic regions from each sample and reference region were combined, leaving a summary file of all unique intergenic regions from each sample, based on both reference genomes. Only transcripts exceeding 200 basepairs in length were retained, as lncRNA are over 200 basepairs in length by defintion.

Next, the Coding Potential Calculator (CPC) was utilized to remove intergenic transcripts that showed coding potential. The CPC scores regions based on homology and open reading frame analysis, assigning each transcript a score between -1 and 1 and then deeming transcripts ”coding” or ”non-coding” based on that score. Intergenic regions that were deemed noncoding were retained. The reverse compliment was also analyzed to remove transcripts that showed coding potential in the opposite direction [10].

The next filtering step was to convert intergenic, non-coding into their appropriate protein sequences and assessing said sequences for protein homology using the Pfam database.

Those sequences that demonstrated a known protein family were removed [42].

The final filtering step was to compare remaining sequences (restored to their genomic coordinates) to known *Bos taurus* genes, removing those that share sequence homology with genes using the BLAST tool [37]. This left only transcripts that were intergenic, non-coding, lacking a protein domain, lacking sequence homology to known genes, and over 200 basepairs in length.. These remaining transcripts match lncRNA criteria.

4.6.4 Comparison to Coding Transcripts

The first comparison made to coding transcripts was length. Length was calculated for all transcripts in all conditions, as well as all lncRNA in all tissues. Average length was then calculated for all transcripts and all lncRNA in all tissues. Length profiles were then able to be compared. All transcripts from all tissues were also pooled and compared to all lncRNA pooled together. R4.2.2 was used to graph these comparisons in box plots. Histograms were also truncated for ease of viewing [44].

Using the Salmon tool, expression level for all transcripts and all lncRNA in all tissues were measured [41]. This allowed the calculation of average expression level for all transcripts in all tissues and all lncRNA in all tissues. All combinations of expression data were graphed using R4.2.2 using histograms. The generated profiles were also able to be compared.

Stringtie generates an exon count for all transcripts analyzed, allowing for the calculation of exons per transcript for all transcripts and all lncRNA [50]. This allowed for comparison of the exon number profiles of all tissues.

All values were reported.

4.6.5 Differential Expression Analysis

The bedtools intersect function was employed to determine which lncRNA were common across tissue types, the expression profile of these common lncRNA were then analyzed for differential expression using Excel to perform a heteroscedastic t-test on each common lncRNA [43].

The online Venn diagram generator tool created by bioinformatics gents, tool, was utilized to create a non-symmetric Venn diagram demonstrating the overlapping lncRNA regions.

4.6.6 Phastcon Analysis

Sequence conservation was analyzed for all transcripts, intergenic transcripts (as generated by CuffCompare), and lncRNA transcripts to generate profiles that can be compared. The results of this analysis also allow for further analysis to be performed. This analysis was done by comparing Phastcon scores of each region of interest. First, regions of interest were converted to human genome equivalents using the UCSC LiftOver webtool. This utilized the ARSUCD1.2 *Bos taurus* genome and returned coordinates based on the hg38 human genome [30].

Once genomic coordinates were converted, the multiBigwigSummary tools was utilized, providing the tool with the hg38 100-way conservation alignment table and .bed files with all genomic regions of interest. The hg38 100-way conservation alignment ta-

ble contains sequence conservation data from 100 distinct species, with varying levels of complexity and this tool returned a conservation score for each region based on such information [19, 34].

For all tissue types, data for all transcripts, intergenic transcripts, and lncRNA transcripts were plotted in a violin plot using R. Scores for all transcripts, intergenic transcripts, and lncRNA transcripts, regardless of tissue type, were also pooled and graphed using a violin plot, as well. lncRNA transcripts for each tissue type were also ranked and plotted a dot plot representing score over rank using R. Finally, lncRNA found in multiple immune related tissues were isolated and ranked based on score, this was plotted also as score versus rank, and compared to single tissue lncRNA, this plotting was also completed using R. All lncRNA found in multiple tissues, as well as the top 25% most conserved lncRNA in each tissue type was retained for further analysis.

4.6.7 Transcriptional Annotation Based on Sequence Homology

The retained lncRNA from PhastCon analysis, all common lncRNA and the top 25% of conserved lncRNA in each tissue, were then analyzed for potential function based on homology. The UCSC genome browser was parsed based on retained genomic regions, overlapping genes were noted. Uniprot was then used to assign a function to each overlapping gene. Patterns and commonalities were then investigated and reported.

4.6.8 Gene Co-Expression Correlation

Expression level data was generated using Salmon for previous analysis, Salmon quantmerge was used to generate expression profiles for all tissue conditions [41, 38].

To avoid false correlations due to coincidental low expression, transcripts were removed if they did not show expression in at least 2 samples. This filtered data was uploaded to R Studio and R was used to generate a co-expression correlation matrix for each tissue type. These matrices were then filtered to reflect only lncRNA correlations, removing gene-to-gene correlations. Using Excel, R^2 values were converted to Z-scores, and subsequently to p-values. The most statistically significantly correlated gene for each lncRNA were then noted. The ensembl database was used to determine the name of the most statistically significantly correlated genes for each lncRNA. Gene names were then used to parse the NCBI gene database where functional annotation has been done in humans for nearly all the associated genes. These definitions and functions were used as the knowledge base for human genes is much larger than cattle. These functions were noted for each lncRNA. Once each lncRNA had been associated with a gene, patterns and commonalities were able to be assessed and reported.

4.6.9 Gene Ontology

The correlation matrices generated for single gene correlation analysis were subsequently used for gene ontology analysis. Using Excel for the ease of column sorting, any gene that correlated with a lncRNA with a p-value of <0.05 was retained and analyzed using the web-based gene ontology tool [14]. The biological processes returned by the tool were noted for all lncRNA. Once all lncRNA had been associated with a biological process, where possible and applicable, patterns and commonalities could be reported and further researched.

4.6.10 Heritability Enrichment Analysis

lncRNA lists were generated for each tissue type and used to complete heritability enrichment analysis. Utilizing GWAS data from 27,000 bulls and 3 million SNPs/INDELs, accurate prediction of transmitting abilities (PTA) could be performed for traits representing cattle production (milk), reproduction (daughter pregnancy rate), health (livability), and body conformation (stature) [26, 13]. MPH for enrichment analysis was used to perform this analysis[15].

MPH functions with the following linear mixed model with two genetic components to fit the data:

$$y = Xb + g_1 + g_2 + e$$

$$g_1 \sim N(0, G_1\sigma_g^2 1)$$

$$g_2 \sim N(0, G_2\sigma_g^2 2$$

$$e \sim N(0, R\sigma_e^2)$$

The variables are measured as follows: b : a vector of fixed effects including intercept X: a covariate matrix corresponding to b, g1, and g2 g_1 : a vector of random effects representing the genetic contributions of the variants within lncRNAs g_2: a vector of random effects representing the genetic contributions of the variants within the remaining genome e: a vector of residual effects G_1: a genomic relationship matrix (GRM) constructed by genotypes corresponding to lncRNA G_2: a genomic relationship matrix (GRM) constructed by genotypes corresponding to the remaining genome G was computed using the following formula:

$$G = \frac{ZZ^T}{m}$$

Where Z represents the standardized genotypes, m is the number of variants in the GRM, using m_1 and m_2 to represent the number of variants within lncRNA and the remaining genome, respectively. R is a diagonal matrix used to model individual reliability of deregressed PTAs. The method of calculation used by MPH relies on a Monte Carlo REML algorithm that estimates variance components (VCs) and uses a Fisher information matrix to compute the variance-covariance matrix of VC estimates. $\frac{\sigma_g^2 1}{m_1}$ equals per-SNP genetic variation in lncRNA, allowing for computation of per-SNP heritability enrichment using $\rho = \frac{\frac{\sigma_g^2 1}{m_1}}{\frac{\sigma_g^2 1 + \sigma_g^2 2}{m_1 + m_2}}$. A delta method was used to calculate the standard error of ρ.

For the 455 lncRNA (126 ileum-associated lymph node lncRNA transcripts, 138 jejunum-associated lymph node lncRNA transcripts, 83 lymphocyte lncRNA transcripts, 25 bone marrow lncRNA transcripts, 40 spleen lncRNA transcripts, and 43 thymus lncRNA transcripts), each transcript was analyzed with a 10kb extension window on both sides. These extended windows were tested for their enrichment of per-SNP heritability for the 4 representative traits of interest. Enrichment analysis utilized nearly 3 million autosomal variants, with a threshold of a minimal MAF value of 0.005 and a HWE p-value of at least 1E-8. 4,326 SNPs/INDELs were found to be within lncRNA/the lncRNA extension window, leaving 2,795,435 SNPs/INDELs remaining. Following this, SNP heritability was partitioned into two genomic relationship matrices (GRMs), one representing SNPs and INDELs within lncRNA and the other representing the remaining SNPs/INDELs. Enrichment was measured as the ratio of per-SNP heritability near lncRNA as compared to the genome-wide heritability, where MPH reported enrichment value and standard error. P-value was computed using a t-test.

4.7 References

[1] Stephanie Anja Juran et al. “Disgusting odors trigger the oral immune system”. In: *Evolution, Medicine, and Public Health* 11.1 (Jan. 2023), pp. 8–17. ISSN: 20506201. DOI: 10.1093/EMPH/EOAC042. URL: https://dx.doi.org/10.1093/emph/eoac042.

[2] *ARS-UCD1.2 - bosTau9 - Genome - Assembly - NCBI*. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1/.

[3] *ARS-UCD1.3 - bosTau9 - Genome - Assembly - NCBI*. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.2.

[4] B. J.E.G. Bast et al. “The HB-6, CDw75, and CD76 differentiation antigens are unique cell-surface carbohydrate determinants generated by the beta-galactoside alpha 2,6-sialyltransferase”. In: *The Journal of cell biology* 116.2 (1992), pp. 423–435. ISSN: 0021-9525. DOI: 10.1083/JCB.116.2.423. URL: https://pubmed.ncbi.nlm.nih.gov/1730763/.

[5] Marco Bocchetti et al. “LncRNAs and Immunity: Coding the Immune System with Noncoding Oligonucleotides”. In: *International Journal of Molecular Sciences* 22.4 (Feb. 2021), pp. 1–30. ISSN: 14220067. DOI: 10.3390/IJMS22041741. URL: /pmc/articles/PMC7916124/%20/pmc/articles/PMC7916124/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7916124/.

[6] *Bos_taurus_UMD_3.1.1 - Genome - Assembly - NCBI*. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF_000003055.6/.

[7] Kathleen Brennan and Jialin Zheng. “Interleukin 8”. In: *xPharm: The Comprehensive Pharmacology Reference* (2007), pp. 1–4. DOI: 10.1016/B978-008055232-3.61916-6.

[8] *C7 complement C7 [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/730.

[9] *cellular biosynthetic process Gene Ontology Term (GO:0044249)*. URL: https://www.informatics.jax.org/vocab/gene_ontology/GO:0044249.

[10] *CPC2 @ CBI, PKU*. URL: http://cpc2.gao-lab.org/download.php.

[11] *CR2 complement C3d receptor 2 [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=1380.

[12] *Cufflinks*. URL: http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/.

[13] Ellen Freebern et al. "GWAS and fine-mapping of livability and six disease traits in Holstein cattle". In: *BMC Genomics 2020 21:1* 21.1 (Jan. 2020), pp. 1–11. ISSN: 1471-2164. DOI: 10.1186/S12864-020-6461-Z. URL: https://link.springer.com/articles/10.1186/s12864-020-6461-z%20https://link.springer.com/article/10.1186/s12864-020-6461-z.

[14] *Gene Ontology Resource*. URL: http://geneontology.org/.

[15] *GitHub - jiang18/mph: MINQUE for Partitioning Heritability*. URL: https://github.com/jiang18/mph.

[16] *GitHub - ncbi/sra-tools: SRA Tools*. URL: https://github.com/ncbi/sra-tools.

[17] Ritobrata Goswami and Mark H. Kaplan. "A Brief History of IL-9". In: *Journal of immunology (Baltimore, Md. : 1950)* 186.6 (Mar. 2011), p. 3283. ISSN: 0022-1767. DOI: 10.4049/JIMMUNOL.1003049. URL: /pmc/articles/PMC3074408/%20/pmc/articles/PMC3074408/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074408/.

[18] *GPCR — Learn Science at Scitable*. URL: https://www.nature.com/scitable/topicpage/gpcr-14047471/.

[19] *Hg38 100-way conservation alignment - genomewiki*. URL: http://genomewiki.ucsc.edu/index.php/Hg38_100-way_conservation_alignment.

[20] *Human Gene MTCP1 (ENST00000362018.2) from GENCODE V43*. URL: https://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=ENST00000362018.2&hgg_chrom=chrX&hgg_start=155065320&hgg_end=155147775&hgg_type=knownGene&db=hg38.

[21] *IL17RB interleukin 17 receptor B [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/55540.

[22] *IL6 Gene - GeneCards — IL6 Protein — IL6 Antibody*. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=IL6.

[23] *IL9R interleukin 9 receptor [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/3581.

[24] *IRAK3 interleukin 1 receptor associated kinase 3 [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/11213.

[25] Chunjie Jiang et al. "Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs". In: *Oncotarget* 7.6 (Feb. 2016), p. 7120. ISSN: 19492553. DOI: 10.18632/ONCOTARGET.6859. URL: /pmc/articles/PMC4872773/%20/pmc/articles/PMC4872773/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4872773/.

[26] Jicai Jiang et al. "Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls". In: *Communications Biology 2019 2:1* 2.1 (June 2019), pp. 1–12. ISSN: 2399-3642. DOI: 10.1038/

s42003-019-0454-y. URL: https://www.nature.com/articles/s42003-019-0454-y.

[27] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. "HISAT: a fast spliced aligner with low memory requirements". In: *Nature Methods 2015 12:4* 12.4 (Mar. 2015), pp. 357–360. ISSN: 1548-7105. DOI: 10.1038/nmeth.3317. URL: https://www.nature.com/articles/nmeth.3317.

[28] Daehwan Kim et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 907–915. ISSN: 15461696. DOI: 10.1038/S41587-019-0201-4.

[29] Tim Lämmermann and Wolfgang Kastenmüller. "Concepts of GPCR-controlled navigation in the immune system". In: *Immunological Reviews* 289.1 (May 2019), p. 205. ISSN: 1600065X. DOI: 10.1111/IMR.12752. URL: /pmc/articles/PMC6487968/%20/pmc/articles/PMC6487968/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6487968/.

[30] *Lift Genome Annotations*. URL: https://genome.ucsc.edu/cgi-bin/hgLiftOver.

[31] Ting Liu et al. "NF-κB signaling in inflammation". In: *Signal Transduction and Targeted Therapy 2017 2:1* 2.1 (July 2017), pp. 1–9. ISSN: 2059-3635. DOI: 10.1038/sigtrans.2017.23. URL: https://www.nature.com/articles/sigtrans201723.

[32] P. Lomonte et al. "Bovine herpesvirus 4: genomic organization and relationship with two other gammaherpesviruses, Epstein-Barr virus and herpesvirus saimiri". In: *Veterinary microbiology* 53.1-2 (Nov. 1996), pp. 79–89. ISSN: 0378-1135. DOI: 10.1016/S0378-1135(96)01236-9. URL: https://pubmed.ncbi.nlm.nih.gov/9011000/.

[33] *macromolecule biosynthetic process Gene Ontology Term (GO:0009059)*. URL: https://www.informatics.jax.org/vocab/gene_ontology/GO:0009059.

[34] *multiBigwigSummary — deepTools 3.5.2 documentation*. URL: https://deeptools.readthedocs.io/en/develop/content/tools/multiBigwigSummary.html.

[35] Kannan Badri Narayanan and Hyun Ho Park. "Toll/interleukin-1 receptor (TIR) domain-mediated cellular signaling pathways". In: *Apoptosis : an international journal on programmed cell death* 20.2 (2015), pp. 196–209. ISSN: 1573-675X. DOI: 10.1007/S10495-014-1073-1. URL: https://pubmed.ncbi.nlm.nih.gov/25563856/.

[36] *NPLOC4 - Nuclear protein localization protein 4 homolog - Homo sapiens (Human) — UniProtKB — UniProt*. URL: https://www.uniprot.org/uniprotkb/Q8TAT6/entry.

[37] *Nucleotide BLAST: Search nucleotide databases using a nucleotide query*. URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

[38] *Overview – Salmon: Fast, accurate and bias-aware transcript quantification from RNA-seq data*. URL: https://combine-lab.github.io/salmon/.

[39] Timothy J. Parkinson. "Specific Infectious Diseases Causing Infertility and Subfertility in Cattle". In: *Veterinary Reproduction & Obstetrics* (Jan. 2019), pp. 434–466. DOI: 10.1016/B978-0-7020-7233-8.00024-0.

[40] *Parts of the Immune System — Children's Hospital of Philadelphia*. URL: https://www.chop.edu/centers-programs/vaccine-education-center/human-immune-system/parts-immune-system.

[41] Rob Patro et al. “Salmon provides fast and bias-aware quantification of transcript expression”. In: *Nature Methods 2017 14:4* 14.4 (Mar. 2017), pp. 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197. URL: https://www.nature.com/articles/nmeth.4197.

[42] *Pfam: Search Pfam*. URL: http://pfam.xfam.org/search#tabview=tab1.

[43] Aaron R Quinlan and Ira M Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033. URL: https://doi.org/10.1093/bioinformatics/btq033.

[44] *R: The R Project for Statistical Computing*. URL: https://www.r-project.org/.

[45] *SETD2 - Histone-lysine N-methyltransferase SETD2 - Homo sapiens (Human) — UniProtKB — UniProt*. URL: https://www.uniprot.org/uniprotkb/Q9BYW2/entry.

[46] Iain Martin Sheldon. “The metritis complex in cattle”. In: *Veterinary Reproduction & Obstetrics* (Jan. 2018), pp. 408–433. DOI: 10.1016/B978-0-7020-7233-8.00023-9.

[47] *SHFL shiftless antiviral inhibitor of ribosomal frameshifting [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/55337.

[48] *ST6GAL1 ST6 beta-galactoside alpha-2,6-sialyltransferase 1 [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/6480.

[49] Luisa Statello et al. “Gene regulation by long non-coding RNAs and its biological functions”. In: *Nature Reviews Molecular Cell Biology 2020 22:2* 22.2 (Dec. 2020), pp. 96–118. ISSN: 1471-0080. DOI: 10.1038/s41580-020-00315-9. URL: https://www.nature.com/articles/s41580-020-00315-9.

[50] *StringTie*. URL: http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual.

[51] Puspa Thapa and Donna L. Farber. "THE ROLE OF THE THYMUS IN THE IMMUNE RESPONSE". In: *Thoracic surgery clinics* 29.2 (May 2019), p. 123. ISSN: 15585069. DOI: 10.1016/J.THORSURG.2018.12.001. URL: /pmc/articles/PMC6446584/%20/pmc/articles/PMC6446584/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6446584/.

[52] *THNSL2 threonine synthase like 2 [Homo sapiens (human)] - Gene - NCBI*. URL: https://www.ncbi.nlm.nih.gov/gene/55258.

[53] Xinming Wang et al. "Emerging roles for G-protein coupled receptors in development and activation of macrophages". In: *Frontiers in Immunology* 10.AUG (Aug. 2019), p. 457996. ISSN: 16643224. DOI: 10.3389/FIMMU.2019.02031/BIBTEX.

[54] "What are the organs of the immune system?" In: (July 2020). URL: https://www.ncbi.nlm.nih.gov/books/NBK279395/.

[55] Xiaopei Zhang et al. "Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels". In: *International Journal of Molecular Sciences* 20.22 (Nov. 2019). ISSN: 14220067. DOI: 10.3390/IJMS20225573. URL: /pmc/articles/PMC6888083/%20/pmc/articles/PMC6888083/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6888083/.

Chapter 5: A Comprehensive Tutorial for Identification of lncRNA from RNA-Seq Data

5.1 Introduction

Attempts have been made to generate lncRNA identification software and web tools, however, they tend to be limited to model organisms. The process can be reverse engineered from publications, but this requires extensive reading, experience with bioinformatics, and a heavy amount of trial and error. This guide intends to use plain language to create an accessible, easy-to-follow, useful tutorial for a bare bones approach to lncRNA identification. It includes what programs are required, how to install them, the exact steps and commands needed to generate a list of candidate lncRNA transcripts, pitfalls to avoid, and helpful tips for a novice user. Although a number of tools are required, they are all publicly available for free and all steps can be run from a command line interface.

5.2 Materials

The only requirement to start this analysis is a computer running an Ubuntu Linux operating system. The steps in this tutorial can be followed on a machine running another operating system or Linux distro, however examples provided are the Ubuntu Linux based commands.

Although not required, it is recommended to obtain and use an external hard drive to contain the data, tools, and outputs of this tutorial. Using an external hard drive allows all the storage to be allocated to the analysis, as well as providing a single space for all results to be held.

It should be noted, these files are quite large. Notes will be made within this tutorial where files can be removed to save space.

5.2.1 Introduction

In this section, we will establish a working environment and install and prepare all the tools utilized within this tutorial. Each tool section is broken into 3 sub-sections. The first, a very brief explanation of the function of the tool/package to introduce the program and its role within the tutorial. This is followed by directions to download the software and move it, if applicable, to the established working area. The third and final section details any commands needed to prepare the tool for use. All information and/or commands provided can be found on the software's site and are re-iterated within this tutorial in an attempt to be as comprehensive and useful as possible.

The provided links and instructions are intended to be as recent as possible at the time of release, however, this cannot be guaranteed. Be sure to check the website for updated links and releases.

5.2.2 Set Up

Before embarking on this endeavor, the first step is to set up a working environment free of other files to avoid confusion. If you are using an external hard drive, navigate to the drive using the following command:

```
cd [file path]
```

If you are not using an external hard drive, create a folder to contain the project:

```
mkdir [project name of choice]
cd [project name of choice]
```

Once you've successfully created your working environment, create a directory (you may have heard of this colloquially as a folder) to house your software and move into the directory.

```
mkdir software
cd software/
```

Now you're ready to begin installing the tools you'll need for this project!

5.2.3 Software Installation

Hisat2

Function: This tool functions to align reads to a reference genome.

Download:

1. Follow this link: `http://daehwankimlab.github.io/hisat2/download/#version-hisat2-221`
2. Scroll to the **Binaries** subheadings
3. Click the link following the "Linux_x86_64" label

Prepare: Use the following commands to ready the software for use:

```
mv /home/[User]/Downloads/hisat* .
unzip hisat*
```

[13]

Samtools

Function: This tool functions in multiple ways to interact with sequencing data. In this tutorial, it will function to change between file formats.

Download:

1. Follow this link: `http://www.htslib.org/download/`
2. Click the download tool for "samtools-Index*" to download the most recent version

Prepare:

1. Use the following commands to ready to software for use:

```
mv /home/[User]/Downloads/samtools* .
tar -xvf samtools*
cd samtools*
./configure
make
sudo make install
cd ..
```

[19]

ChromToUcsc

Function: This tool functions to rename chromosomes to a consistent name between reference files.

Download:

1. Use the following command to obtain the software:

```
rsync -aP \rsync://hgdownload.soe.ucsc.edu/genome/admin/exe/linux
   ↪ .x86_64/chromToUcsc ./
```

Prepare:

1. N/A

[11]

Stringtie

Function: This tool acts to assemble reads into potential transcripts.

Download:

1. Follow this link: `http://ccb.jhu.edu/software/stringtie/#install`
2. Click the "stringtie-***.Linux_x86_64.tar.gz"

Prepare:

1. Use the following commands to ready to software for use:

```
mv /home/[User]/Downloads/stringtie* .
gunzip stringtie*
tar -xvf stringtie*
```

[20]

Cufflinks

Function: This tool functions mainly for transcript assembly and differential expression analysis. In this tutorial, it will function to annotate mapped transcripts based on a reference genome.

Download:

1. Follow this link:

   `http://cole-trapnell-lab.github.io/cufflinks/install/`

2. Click the "Linux" link in the most row representing the most recent release

Prepare:

1. Use the following commands to ready to software for use:

```
mv /home/[User]/Downlods/cufflinks* .
gunzip cufflinks*
tar -xvf cufflinks*
```

[4]

lncRNA Identification Pipeline

Function: This package contains scripts written by various people and compiled in a previous project. It will function in several steps to isolate intergenic transcripts, as well as convert nucleotide sequences to protein sequences.

Download:

Use the following command to obtain the software:

```
wget https://github.com/amarceau-998/lncRNA-Identification-Pipeline/
   ↪ archive/refs/heads/main.zip
```

Prepare:

1. Use the following commands to ready to software for use:

```
unzip main.zip
```

[1]

Bedtools

Function: This package holds many functions for genomic analysis. This tutorial will only use the "getfasta" function, which will convert a bed file, a file housing genetic coordinates, into a fasta file, a file containing the sequence(s) of interest.

Download:

1. Use the following command:

```
wget https://github.com/arq5x/bedtools2/releases/download/v2
   ↪ .31.0/bedtools-2.31.0.tar.gz
```

Prepare:

1. Use the following commands to ready to software for use:

```
gunzip bedtools*
tar -xvf bedtools*
cd bedtools*
make
sudo cp bin/* /usr/local/bin/
cd ..
```

[18]

Seqtk

Function: This package mainly functions to process fasta files. The only tool used in this tutorial will retrieve the reverse complement to fasta sequences.

Download:

1. Use the following command:

```
git clone https://github.com/lh3/seqtk.git
```

Prepare:

1. Use the following commands to ready to software for use:

```
cd seqtk
make
cd ..
```

[8]

CPC2

Function: This tool assesses the coding potential of given sequences and categorizes them as either coding or non-coding.

Download:

1. Follow this link: `http://cpc2.gao-lab.org/download.php`
2. Click the “Click here” link within the ”For the version supports Python3 please click here” text
3. Click the ”source code (tar.gz)” link

Prepare:

1. Use the following commands to ready to software for use:

```
mv /home/[User]/Downloads/CPC* .
gunzip CPC*
tar -xvf CPC*
cd CPC*
cd libs/libsvm/
gunzip libsvam*
tar -xvf libsvam*
cd libsvm-*
make clean && make
cd .. -- 4 times
```

[3]

Hmmer

Function: This tool functions to compare protein sequences to a protein family database and find similarities.

Download:

1. Use the following commands:

```
wget http://eddylab.org/software/hmmer/hmmer.tar.gz
```

Prepare:

1. Use the following commands to ready to software for use:

```
tar -xvf hmmer*
cd hmmer*
./configure
make
cd ..
```

[10]

BLAST

Function: This tool functions to compare nucleotide sequences to a sequence database to find similarities.

Download:

1. Follow this link:

   `https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/`

2. Click the link ending in "-linux.tar.gz"

Prepare:

1. Use the following commands to ready to software for use:

```
mv /home/[User]/Downloads/ncbi-blast* .
gunzip ncbi*
tar -xvf ncbi2* .
```

[15]

SRA-Toolkit

Function: This tool kit allows you to interact with the SRA database. We will only be using fastq-dump to download data from the SRA database. If you are not obtaining data from the SRA database, you do not need this tool.

Download:

1. Follow this link:

   `https://github.com/ref2/sra-tools/wiki/02.-Installing-SRA-Toolkit`

2. Click the link in the "Ubuntu" row

Prepare:

1. Use the following commands to ready to software for use:

```
mv /home/[User]/Downloads/sra* .
gunzip sra*
tar -xvf sra*
```

[9]

5.2.4 Reference Files

Introduction

In this section, we will acquire the necessary reference files for our analysis. Many files are used multiple times in several steps. This tutorial will provide a source where the necessary file(s) are likely be found; however, if you are working with a more rare species, the files may not be found at the provided link and further digging may be required to obtain the files.

Similarly to the software section, there will be a brief description of the file, a link to where you are likely to find it, instructions as to how to download the file, and if needed, any pertinent information notes.

Before beginning this section, be sure you are in your project directory.

Reference Genome(s)

Description: A general transfer format (gtf) file containing lines of information for known genetic features.

Location: `https://hgdownload.soe.ucsc.edu/downloads.html`

Instructions:

1. Select your species of interest from the drop down menus
2. Click "Genome sequence files and select annotations (2bit, GTF, GC-content, etc)" under the first subheading, this will be the most recent release for the species
3. Scroll to the bottom and click "genes/"
4. Click the links for 2 of the reference genomes to begin their downloading

5. Move the files to your current directory

```
mv /home/[User]/Downloads/[Reference genome 1]* .
mv /home/[User]/Downloads/[Reference genome 2]* .
```

Notes:

- If there is only one refence genome, that is fine! Follow the same above steps for the single genome available.

[22]

Reference Fasta

Description: A file containing nucleotide sequences and sequence identifiers for the organism of interest.

Location: `https://hgdownload.soe.ucsc.edu/downloads.html`

Instructions:

1. Select your species of interest from the drop down menus
2. Click "Genome sequence files and select annotations (2bit, GTF, GC-content, etc)" under the first subheading, this will be the most recent release for the species
3. Scroll to the bottom and click the link following this format: [species name].fa.gz
4. Move the file to your current directory

```
mv /home/[User]/Downloads/[species name]* .
gunzip *.fa
```

Notes:

- There are several files ending in variations of ".fa.gz", be sure you are downloading the correct fasta file

[22]

Chromosome Alias Files

Description: A text file containing the various chromosome naming conventions allowing for simple conversions.

Location: `https://hgdownload.soe.ucsc.edu/downloads.html`

Instructions:

1. Select your species of interest from the drop down menus
2. Click "Genome sequence files and select annotations (2bit, GTF, GC-content, etc)" under the first subheading, this will be the most recent release for the species
3. Scroll to the bottom and click the link following this format: [species name].chromAlias.txt
4. Move the file to your current directory

```
mv /home/[User]/Downloads/[species name].chromAlias.txt .
```

Notes:

- There may be a similarly named file with the ".bb" file extension, be sure you are downloading that ".txt" file

- If when clicking this link, the file does not download and instead opens as a text webpage, the information can be copied into a .txt document in your project folder

[22]

Unscaffolded Chromosome File

Description: Fasta file containing sequence information for transcripts not assigned to a chromosome.

Location: `https://hgdownload.soe.ucsc.edu/downloads.html`

Instructions:

1. Select your species of interest from the drop down menus
2. Select "Sequence data by chromosome", it may be in a less recent release
3. Scroll down to the "chrUn*" link and click the link
4. Move the file to your current directory

```
mv /home/[User]/Downloads/chrUn* .
```

Notes:

- This can be very tricky to find

[22]

Pfam Database

Description: A database containing protein domains.

Location: `https://www.ebi.ac.uk/interpro/download/Pfam/`

Instructions:

1. Click either the “Pfam-A models” link or the download icon in the Pfam-A models row
2. Move the file to your current directory

```
mv /home/[User]/Downloads/Pfam* .
```

Notes:

- This is a large file so do not panic if it takes several minutes to download

[17]

BLAST Database

Description: A database containing genetic sequences of known genes.

Location: `https://useast.ensembl.org/index.html`

Instructions:

1. Under the ‘All genomes” header, click the drop down menu and select your species
2. Under the “Gene annotation” header, click the “Download fasta” link
3. Click the “cds” link
4. Click the file with the “.cds.all.fa.gz” file extension
5. Move the file to your current directory

```
mv /home/[User]/Downloads/*.cds.all.fa.gz .
```

Notes:

- Be sure this is the “cds” file as it contains only known genes, if you use “cdna” file, your final filtering step will not work!

[6]

5.3 Methods

5.3.1 Introduction

Now that we have installed and prepared all the necessary tools and collected the necessary reference files, we are ready to begin our analysis. To be as digestible as possible, this tutorial is broken into many small steps that are generally a single command or a small set of commands intended to complete a single step in the analysis process. Some early steps may be considered common practice when working with genetic data and may be skipped by more experienced users; however, as previously stated, this tutorial is intended to be as comprehensive as possible for beginner users.

Within each step, there will be a description of purpose of the step, a command with place holders, an example command, a description of the output files, and any applicable notes about the command and/or the files associated with it. The notes section is where you will find notes about files that can be removed to save space.

5.3.2 Tutorial

Obtain data

Purpose: To identify lncRNA, we first must have genetic data to analyze! If you already have genetic information in a .fasta or a .fastq format, this step can be skipped. If you do not, a database like the SRA database can be used to obtain publically available data. The following command(s) will download data from the SRA database.

Command:

```
software/sratoolkit*/bin/fastq-dump* --gzip --split-3 [SRR_ID]
```

Example:

```
software/sratoolkit*/bin/fastq-dump* --gzip --split-3 SRR24066292
software/sratoolkit*/bin/fastq-dump* --gzip --split-3 SRR24066293
software/sratoolkit*/bin/fastq-dump* --gzip --split-3 SRR18899314
```

Output: This command will yield the fastq file for an entry in the SRA databases based on the SRR ID that was provided. The file will be in a condensed format.

Notes:

- It is possible your data will need pre-analysis processing. Information on pre-processing can be found here: link, using the Phase 1 tutorial. This step should be handled on a case-by-case basis.
- It may be tempting to unzip the files, but this is not necessary and should not be done to preserve space.

Build index

Purpose: Index files are required for a future step aligning reads to the genome. This command will generate the necessary index files.

Command:

```
software/hisat*/hisat2-build [referencence fasta] Index
```

Example:

```
software/hisat*/hisat2-build canFam6.fa Index
```

Output: This command will create 8 files with the .ht2 suffix, these will be used in the next step.

Notes:

- From this point forward, commands will need to be ran on each sample
- Many commands can be ran in a loop manner, however that is not beginner friendly. If you are comfortable with writing and running loops, feel free to modify commands as needed for this function

Align data

Purpose: Using the previously generated index information, this step will align your data to the reference information.

Command:

- If your data is made up of pair-end sequences; that is, your files end in _.1.fq.gz and _.2.fq.gz:

```
software/hisat*/hisat2 -q -x [index file name] -1 [data name]_1.
   ↪ fq.gz -2 [data name]_2.fq.gz -S [data name]_hisat2-1.sam --
   ↪ novel-splicesite-outfile [data name].junctions --summary-
   ↪ file [data name]_hisat2-1_summary.txt
```

- If your data is made up of unpaired sequence data; that is, your file ends in .fq.gz only:

```
software/hisat*/hisat2 -q x [index file name] -U [data name].fq.
   ↪ gz -S [data name]_hisat2-1.sam --novel-splicesite-outfile [
```

```
    ↪ data name].junctions --summary-file [data name]_hisat2-1
    ↪ _summary.txt
```

Example:

```
software/hisat*/hisat2 -q -x Index -U SRR24066292.fastq.gz -S
    ↪ Sample1_hisat2-1.sam --novel-splicesite-outfile Sample1.
    ↪ junctions --summary-file Sample1_hisat2-1_summary.txt
software/hisat*/hisat2 -q -x Index -U SRR24066293.fastq.gz -S
    ↪ Sample2_hisat2-1.sam --novel-splicesite-outfile Sample2.
    ↪ junctions --summary-file Sample2_hisat2-1_summary.txt
software/hisat*/hisat2 -q -x Index -1 SRR18899314_1.fastq.gz -2
    ↪ SRR18899314_2.fastq.gz -S Sample3_hisat2-1.sam --novel-
    ↪ splicesite-outfile Sample3.junctions --summary-file
    ↪ Sample3_hisat2-1_summary.txt
```

Output:

- A ".sam" file that contains sequence alignment information
- A ".junctions" file that contains the junctions between aligned transcripts
- A ".txt" file containing the alignment statistics for the sample

Notes:

- This aligner is able to process data in several formats: fastq, fasta, and SRA database enteries. If your data is not in the fastq format, visit this page to determine the appropriate command: `http://daehwankimlab.github.io/hisat2/manual/`

- There are other alignment tools available, if you have a preferred alignment tool, ensure you have a ".junctions" file output before continuing.
- To save space, the ".sam" file generated by this step can be removed.

Secondary alignment

Purpose: Using the previously generated index information and the junctions found in step 3, this step will align your data to the reference information with more accuracy.

Command:

If your data is made up of pair-end sequences; that is, your files end in _.1.fq.gz and _.2.fq.gz:

```
software/hisat*/hisat2 -q -x [index file name] -1 [data name]_1.fq.gz
   ↪ -2 [data name]_2.fq.gz -S [data name]_hisat2-2.sam --novel-
   ↪ splicesite-infile [data name].junctions --summary-file [data
   ↪ name]_hisat2-2_summary.txt
```

If your data is made up of unpaired sequence data; that is, your file ends in .fq.gz only:

```
software/hisat*/hisat2 -q -x [index file name] -U [data name].fq.gz -S
   ↪  [data name]_hisat2-2.sam --novel-splicesite-infile [data name
   ↪ ].junctions --summary-file [data name]_hisat2-2_summary.txt
```

Example:

```
software/hisat*/hisat2 -q -x Index -U SRR24066292.fastq.gz -S
   ↪ Sample1_hisat2-2.sam --novel-splicesite-infile Sample1.
   ↪ junctions --summary-file Sample1_hisat2-2_summary.txt
```

```
software/hisat*/hisat2 -q -x Index -U SRR24066293.fastq.gz -S
    ↪ Sample2_hisat2-2.sam --novel-splicesite-infile Sample2.
    ↪ junctions --summary-file Sample2_hisat2-2_summary.txt
software/hisat*/hisat2 -q -x Index -1 SRR18899314_1.fastq.gz -2
    ↪ SRR18899314_2.fastq.gz -S Sample3_hisat2-2.sam --novel-
    ↪ splicesite-infile Sample3.junctions --summary-file
    ↪ Sample3_hisat2-2_summary.txt
```

Output:

- A ".sam" file that contains sequence alignment information
- A ".txt" file containing the alignment statistics for the sample

Notes:

- This command looks and operates in a very similar way to step 3; however, it is not the same command. The output name has been changed to reflect that it is the second iteration of alignment and the "–novel-splicesite-outfile" option has been changed to "–novel-splicesite-infile"
- As previously mentioned, you are free to use any preferred aligner tool if you have one. In this step, be sure the tool outputs a ".sam" file
- After this step, you can remove the ".junction" files, they are not large files but space can be precious

Convert SAM file to BAM file

Purpose: This step functions to convert the sequence alignment files to a compressed version for further analysis.

Command:

```
software/samtools*/samtools view -S -b [data name]_hisat2-2.sam > [
    ↪ data name].bam
```

Example:

```
software/samtools*/samtools view -S -b Sample1_hisat2-2.sam > Sample1.
    ↪ bam
software/samtools*/samtools view -S -b Sample2_hisat2-2.sam > Sample2.
    ↪ bam
software/samtools*/samtools view -S -b Sample3_hisat2-2.sam > Sample3.
    ↪ bam
```

Output:

- A ".bam" file containing the same information as the ".sam" file in a more condensed format

Notes:

- The ".sam" file can now be removed
- You will not be able to natively open the ".bam" file

Sort BAM file

Purpose: This step sorts the ".bam" file by genomic coordinates to avoid errors going forward.

Command:

```
software/samtools*/samtools sort [data name].bam > [data name]_sorted.
    ↪ bam
```

Example:

```
software/samtools*/samtools sort Sample1.bam > Sample1_sorted.bam
software/samtools*/samtools sort Sample2.bam > Sample2_sorted.bam
software/samtools*/samtools sort Sample3.bam > Sample3_sorted.bam
```

Output:

- A ".bam" file containing the same information as the input file, sorted by location

Notes:

- You can remove the original ".bam" file
- As previously stated, you cannot natively open the ".bam" file

Standardize reference genome naming conventions

Purpose: Different databases use different chromosomal naming conventions; this step functions to standardize those names to avoid errors. If you only have 1 reference genome, perform only one of the commands.

Command:

```
software/chromToUcsc -a [chromAlias.txt] -i [reference 1].gtf -o ref1.
    ↪ gtf
software/chromToUcsc -a [chromAlias.txt] -i [reference 2].gtf -o ref2.
    ↪ gtf
```

Example:

```
software/chromToUcsc -a canFam6.chromAlias.txt -i refGene.gtf.gz -o
    ↪ ref1.gtf
software/chromToUcsc -a canFam6.chromAlias.txt -i ncbiRefSeq.gtf.gz.
    ↪ gtf -o ref2.gtf
```

Output:

- 2 modified ".gtf" files with a consistent chromosome naming convention

Notes:

- You can remove the original reference files to avoid confusion
- From this point forward, if you only have one reference, follow the commands for one of the reference genomes and disregard the other

Transcript Assembly

Purpose: This step will take the previously generated alignment data and create a transcript assembly. This tutorial uses 2 reference genomes to improve accuracy.

Command:

```
software/stringtie*/stringtie [data name]_sorted.bam -o [data name]
    ↪ _ref1.gtf -G [standardized reference genome 1].gtf -A [data
    ↪ name]_ref1.tab
software/stringtie*/stringtie [data name]_sorted.bam -o [data name]
    ↪ _ref2.gtf -G [standardized reference genome 2].gtf -A [data
    ↪ name]_ref2.tab
```

Example:

```
software/stringtie*/stringtie Sample1_sorted.bam -o Sample1_ref1.gtf -
    ↪ G ref1.gtf -A Sample1_ref1.tab
software/stringtie*/stringtie Sample1_sorted.bam -o Sample1_ref2.gtf -
    ↪ G ref2.gtf -A Sample1_ref2.tab
software/stringtie*/stringtie Sample2_sorted.bam -o Sample2_ref1.gtf -
    ↪ G ref1.gtf -A Sample2_ref1.tab
software/stringtie*/stringtie Sample2_sorted.bam -o Sample2_ref2.gtf -
    ↪ G ref2.gtf -A Sample2_ref2.tab
software/stringtie*/stringtie Sample3_sorted.bam -o Sample3_ref1.gtf -
    ↪ G ref1.gtf -A Sample3_ref1.tab
software/stringtie*/stringtie Sample3_sorted.bam -o Sample3_ref2.gtf -
    ↪ G ref2.gtf -A Sample3_ref2.tab
```

Output:

- 2 ".gtf" files containing the genomic features of your data based on the reference genome
- 2 ".tab" files containing expression data for transcripts within your data

Notes:

- If there is a naming convention for your reference genome(s) that you prefer, feel free to use that in place of "_ref1" and/or "_ref2"
- You can remove the "_sorted.bam" files now if you choose

Split files

Purpose: In order to avoid errors and loss of information, files should be split to represent known chromosomes and unknown chromosomes.

Command:

```
grep 'chrUn*' [data name]_ref1.gtf > [data name]_ChrUn_ref1.gtf
grep -v 'chrUn*' [data name]_ref1.gtf > [data name]_ChrKnown_ref1.gtf
grep 'chrUn*' [data name]_ref2.gtf > [data name]_ChrUn_ref2.gtf
grep -v 'chrUn*' [data name]_ref2.gtf > [data name]_ChrKnown_ref2.gtf
```

Example:

```
grep 'chrUn' Sample1_ref1.gtf > Sample1_ChrUn_ref1.gtf
grep -v 'chrUn*' Sample1_ref1.gtf > Sample1_ChrKnown_ref1.gtf
grep 'chrUn*' Sample1_ref2.gtf > Sample1_ChrUn_ref2.gtf
grep -v 'chrUn*' Sample1_ref2.gtf > Sample1_ChrKnown_ref2.gtf
grep 'chrUn*' Sample2_ref1.gtf > Sample2_ChrUn_ref1.gtf
grep -v 'chrUn*' Sample2_ref1.gtf > Sample2_ChrKnown_ref1.gtf
grep 'chrUn*' Sample2_ref2.gtf > Sample2_ChrUn_ref2.gtf
grep -v 'chrUn*' Sample2_ref2.gtf > Sample2_ChrKnown_ref2.gtf
grep 'chrUn*' Sample3_ref1.gtf > Sample3_ChrUn_ref1.gtf
grep -v 'chrUn*' Sample3_ref1.gtf > Sample3_ChrKnown_ref1.gtf
grep 'chrUn*' Sample3_ref2.gtf > Sample3_ChrUn_ref2.gtf
grep -v 'chrUn*' Sample3_ref2.gtf > Sample3_ChrKnown_ref2.gtf
```

Output:

- A ".gtf" file containing transcript information for unknown chromosomes based on the ensembl reference genome
- A ".gtf" file containing transcript information for known chromosomes based on the ensembl reference genome
- A ".gtf" file containing transcript information for unknown chromosomes based on the ref2 reference genome
- A ".gtf" file containing transcript information for known chromosomes based on the ref2 reference genome

Notes:

- This set of commands will be used several times

Cuffcompare

Purpose: Cuffcompare will function to annotate identified transcripts and assign a class code that can be utilized for filtering.

Command:

```
software/cufflink*/cuffcompare [data name]_ChrUn_ref1.gtf -s [unknown
    ↪ chromosome].fa -r ref1.gtf -o [data name]
    ↪ _ChrUn_ref1_Cuffcompare
software/cufflink*/cuffcompare [data name]_ChrKnown_ref1.gtf -s [
    ↪ reference fasta file].fa -r ref1.gtf -o [data name]
    ↪ _ChrKnown_ref1_Cuffcompare
software/cufflink*/cuffcompare [data name]_ChrUn_ref2.gtf -s [unknown
    ↪ chromosome].fa -r ref2.gtf -o [data name]
    ↪ _ChrUn_ref2_Cuffcompare
```

```
software/cufflink*/cuffcompare [data name]_ChrKnown_ref2.gtf -s [
    ↪ reference fasta file].fa -r ref2.gtf -o [data name]
    ↪ _ChrKnown_ref2_Cuffcompare
```

Example:

```
software/cufflink*/cuffcompare Sample1_ChrUn_ref1.gtf -s chrUn.fa -r
    ↪ ref1.gtf -o Sample1_ChrUn_ref1_Cuffcompare
software/cufflink*/cuffcompare Sample1_ChrKnown_ref1.gtf -s canFam6.fa
    ↪  -r ref1.gtf -o Sample1_ChrKnown_ref1_Cuffcompare
software/cufflink*/cuffcompare Sample1_ChrUn_ref2.gtf -s chrUn.fa -r
    ↪ ref2.gtf -o Sample1_ChrUn_ref2_Cuffcompare
software/cufflink*/cuffcompare Sample1_ChrKnown_ref2.gtf -s canFam6.fa
    ↪  -r ref2.gtf -o Sample1_ChrKnown_ref2_Cuffcompare
software/cufflink*/cuffcompare Sample2_ChrUn_ref1.gtf -s chrUn.fa -r
    ↪ ref1.gtf -o Sample2_ChrUn_ref1_Cuffcompare
software/cufflink*/cuffcompare Sample2_ChrKnown_ref1.gtf -s canFam6.fa
    ↪  -r ref1.gtf -o Sample2_ChrKnown_ref1_Cuffcompare
software/cufflink*/cuffcompare Sample2_ChrUn_ref2.gtf -s chrUn.fa -r
    ↪ ref2.gtf -o Sample2_ChrUn_ref2_Cuffcompare
software/cufflink*/cuffcompare Sample2_ChrKnown_ref2.gtf -s canFam6.fa
    ↪  -r ref2.gtf -o Sample2_ChrKnown_ref2_Cuffcompare
software/cufflink*/cuffcompare Sample3_ChrUn_ref1.gtf -s chrUn.fa -r
    ↪ ref1.gtf -o Sample3_ChrUn_ref1_Cuffcompare
software/cufflink*/cuffcompare Sample3_ChrKnown_ref1.gtf -s canFam6.fa
    ↪  -r ref1.gtf -o Sample3_ChrKnown_ref1_Cuffcompare
```

```
software/cufflink*/cuffcompare Sample3_ChrUn_ref2.gtf -s chrUn.fa -r
   ↪ ref2.gtf -o Sample3_ChrUn_ref2_Cuffcompare
software/cufflink*/cuffcompare Sample3_ChrKnown_ref2.gtf -s canFam6.fa
   ↪  -r ref2.gtf -o Sample3_ChrKnown_ref2_Cuffcompare
```

Output:

- A ".combined.gtf" file for each reference, a gtf file containing all the transfrags
- A ".loci" file for each reference, a file containing points of interest
- A ".refmap" file for each reference, a file assigning Cufflink matches to each transcript
- A "tmap" file for each reference, a file listing closely matched Cufflink transcripts
- A ".stats" file for each reference, a file detailing the queries and matches within each comparison
- A ".tracking" file for each comparison between transcripts

Notes:

- Note that there is not a file extension on the output flag (-o), this is intentional. This command produces several output files, they will all have the same name with varying file extensions.
- By this point, the file names are getting long. This will be rectified in the next step
- All "_CuffCompare" files except the ".combined.gtf" file can be removed
- If any ".fa.fai" files were generated by this step, they can also be removed

Reconstitute files

Purpose: This step is dual purposed, par down the commands needed for analysis and create an opportunity to shorten file names back to a reasonable length.

Command:

```
cat [data name]_ChrUn_ref1_Cuffcompare.combined.gtf [data name]
    ↪ _ChrKnown_ref1_Cuffcompare.combined.gtf > [data name]
    ↪ _ref1_Cuffcompare_Out.gtf
cat [data name]_ChrUn_ref2_Cuffcompare.combined.gtf [data name]
    ↪ _ChrKnown_ref2_Cuffcompare.combined.gtf > [data name]
    ↪ _ref2_Cuffcompare_Out.gtf
```

Example:

```
cat Sample1_ChrUn_ref1_Cuffcompare.combined.gtf
    ↪ Sample1_ChrKnown_ref1_Cuffcompare.combined.gtf >
    ↪ Sample1_ref1_Cuffcompare_Out.gtf
cat Sample1_ChrUn_ref2_Cuffcompare.combined.gtf
    ↪ Sample1_ChrKnown_ref2_Cuffcompare.combined.gtf >
    ↪ Sample1_ref2_Cuffcompare_Out.gtf
cat Sample2_ChrUn_ref1_Cuffcompare.combined.gtf
    ↪ Sample2_ChrKnown_ref1_Cuffcompare.combined.gtf >
    ↪ Sample2_ref1_Cuffcompare_Out.gtf
cat Sample2_ChrUn_ref2_Cuffcompare.combined.gtf
    ↪ Sample2_ChrKnown_ref2_Cuffcompare.combined.gtf >
    ↪ Sample2_ref2_Cuffcompare_Out.gtf
```

```
cat Sample3_ChrUn_ref1_Cuffcompare.combined.gtf
    ↪ Sample3_ChrKnown_ref1_Cuffcompare.combined.gtf >
    ↪ Sample3_ref1_Cuffcompare_Out.gtf
cat Sample3_ChrUn_ref2_Cuffcompare.combined.gtf
    ↪ Sample3_ChrKnown_ref2_Cuffcompare.combined.gtf >
    ↪ Sample3_ref2_Cuffcompare_Out.gtf
```

Output:

- A ".gtf" file containing the cuffcompare annotated transcripts for both known and unknown chromosomes for each reference genome

Notes:

- The ".combined.gtf" files can be removed at this point, this will eliminate confusion
- You should also remove the "ChrUn" and "ChrKnown" files generated from step 9

Intergenic list

Purpose: Using the labels assigned to transcripts with Cuffcompare, we will now generate a list of only those transcripts marked as intergenic. If a transcript is annotated as being anything else, it can be assumed to not be a lncRNA as it associates with a known gene in some way.

Command:

```
perl software/lncRNA-Identification-Pipeline-main/
    ↪ get_intergenicLoci_list.pl -g1 [data name]_ref1_Cuffcompare_Out
    ↪ .gtf -g2 [data name]_ref2_Cuffcompare_Out.gtf -o [data name]
    ↪ _intergenic.txt
```

Example:

```
perl software/lncRNA-Identification-Pipeline-main/
   ↪ get_intergenicLoci_list.pl -g1 Sample1_ref1_Cuffcompare_Out.gtf
   ↪  -g2 Sample1_ref2_Cuffcompare_Out.gtf -o Sample1_intergenic.txt
perl software/lncRNA-Identification-Pipeline-main/
   ↪ get_intergenicLoci_list.pl -g1 Sample2_ref1_Cuffcompare_Out.gtf
   ↪  -g2 Sample2_ref2_Cuffcompare_Out.gtf -o Sample2_intergenic.txt
perl software/lncRNA-Identification-Pipeline-main/
   ↪ get_intergenicLoci_list.pl -g1 Sample3_ref1_Cuffcompare_Out.gtf
   ↪  -g2 Sample3_ref2_Cuffcompare_Out.gtf -o Sample3_intergenic.txt
```

Output:

- A ".txt" file listing only the transcripts that are annotated as intergenic

Notes:

- Both reference genomes are used in this step to increase accuracy
- If you only have 1 reference genome, supply the cuffcompare output for both -g1 and -g2

Generate intergenic GTF

Purpose: Using the list of intergenic transcripts, this step will filter the previous GTF files to reflect only intergenic transcripts for further filtering.

Command:

```
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g [data name]_ref1_Cuffcompare_Out.gtf -l [data name]
    ↪ _intergenic.txt -o [data name]_ref1_intergenic.gtf
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g [data name]_ref2_Cuffcompare_Out.gtf -l [data name]
    ↪ _intergenic.txt -o [data name]_ref2_intergenic.gtf
```

Example:

```
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g Sample1_ref1_CuffCompare_Out.gtf -l Sample1_intergenic.txt
    ↪ -o Sample1_ref1_intergenic.gtf
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g Sample1_ref2_CuffCompare_Out.gtf -l Sample1_intergenic.txt
    ↪ -o Sample1_ref2_intergenic.gtf
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g Sample2_ref1_CuffCompare_Out.gtf -l Sample2_intergenic.txt
    ↪ -o Sample2_ref1_intergenic.gtf
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g Sample2_ref2_CuffCompare_Out.gtf -l Sample2_intergenic.txt
    ↪ -o Sample2_ref2_intergenic.gtf
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g Sample3_ref1_CuffCompare_Out.gtf -l Sample3_intergenic.txt
    ↪ -o Sample3_ref1_intergenic.gtf
perl software/lncRNA-Identification-Pipeline-main/get_intergenicGTF.pl
    ↪  -g Sample3_ref2_CuffCompare_Out.gtf -l Sample3_intergenic.txt
```

```
    ↪ -o Sample3_ref2_intergenic.gtf
```

Output:

- 2“.gtf” files, one for each reference genome, containing only transcripts deemed intergenic

Notes:

- Feel free to remove the ”CuffCompare_Out” files at this point

Generate a summary file

Purpose: A GTF file contains a lot of information and can be very overwhelming. This step will generate a more streamlined file summarizing the intergenic transcripts in the sample.

Command:

```
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i [
    ↪ data name]_ref1_intergenic.gtf -o [data name]
    ↪ _ref1_intergenic_summary.txt
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i [
    ↪ data name]_ref2_intergenic.gtf -o [data name]
    ↪ _ref2_intergenic_summary.txt
```

Example:

```
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i
    ↪ Sample1_ref1_intergenic.gtf -o Sample1_ref1_intergenic_summary.
    ↪ txt
```

```
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i
    ↪ Sample1_ref2_intergenic.gtf -o Sample1_ref2_intergenic_summary.
    ↪ txt
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i
    ↪ Sample2_ref1_intergenic.gtf -o Sample2_ref1_intergenic_summary.
    ↪ txt
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i
    ↪ Sample2_ref2_intergenic.gtf -o Sample2_ref2_intergenic_summary.
    ↪ txt
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i
    ↪ Sample3_ref1_intergenic.gtf -o Sample3_ref1_intergenic_summary.
    ↪ txt
perl software/lncRNA-Identification-Pipeline-main/summary_gtf.pl -i
    ↪ Sample3_ref2_intergenic.gtf -o Sample3_ref2_intergenic_summary.
    ↪ txt
```

Output:

- 2 ".txt" files summarizing the intergenic transcripts found in the sample, one for each reference genome

Notes:

- It is recommended to retain the intergenic GTF files for your sample for future reporting purposes. The note section of this tutorial will provide examples of potential analyses to report findings.

Filter by size

Purpose: Since lncRNA by definitoin are longer than 200 basepairs in length, one filtering step needed is to remove those transcripts shorter than 200 bases. This step will perform that.

Command:

```
awk '(NR==1) || ($8 > 199) ' [data name]_ref1_intergenic_summary.txt >
   ↪  [data name]_ref1_size.txt
awk '(NR==1) || ($8 > 199) ' [data name]_ref2_intergenic_summary.txt >
   ↪  [data name]_ref2_size.txt
```

Example:

```
awk '(NR==1) || ($8 > 199) ' Sample1_ref1_intergenic_summary.txt >
   ↪ Sample1_ref1_size.txt
awk '(NR==1) || ($8 > 199) ' Sample1_ref2_intergenic_summary.txt >
   ↪ Sample1_ref2_size.txt
awk '(NR==1) || ($8 > 199) ' Sample2_ref1_intergenic_summary.txt >
   ↪ Sample2_ref1_size.txt
awk '(NR==1) || ($8 > 199) ' Sample2_ref2_intergenic_summary.txt >
   ↪ Sample2_ref2_size.txt
awk '(NR==1) || ($8 > 199) ' Sample3_ref1_intergenic_summary.txt >
   ↪ Sample3_ref1_size.txt
awk '(NR==1) || ($8 > 199) ' Sample3_ref2_intergenic_summary.txt >
   ↪ Sample3_ref2_size.txt
```

Output:

- 2 ".txt" files, one for each reference, containing intergenic transcripts that are longer than 200 basepairs in length.

Notes:

- N/A

Convert file format

Purpose: The summary text file contains lots of useful information for summarizing findings and characteristics, but there is a lot of extraneous information that will not be needed for continued analysis. This step will reduce the file to just the genomic coordinates in a ".bed" file.

Command:

```
awk '{print $3"\t"$5"\t"$6}' [data name]_ref1_size.txt > [data name]
    ↪ _ref1_size.bed
awk '{print $3"\t"$5"\t"$6}' [data name]_ref2_size.txt > [data name]
    ↪ _ref2_size.bed
```

Example:

```
awk '{print $3"\t"$5"\t"$6}' Sample1_ref1_size.txt > Sample1_ref1_size
    ↪ .bed
awk '{print $3"\t"$5"\t"$6}' Sample1_ref2_size.txt > Sample1_ref2_size
    ↪ .bed
awk '{print $3"\t"$5"\t"$6}' Sample2_ref1_size.txt > Sample2_ref1_size
    ↪ .bed
awk '{print $3"\t"$5"\t"$6}' Sample2_ref2_size.txt > Sample2_ref2_size
    ↪ .bed
```

```
awk '{print $3"\t"$5"\t"$6}' Sample3_ref1_size.txt > Sample3_ref1_size
   ↪ .bed
awk '{print $3"\t"$5"\t"$6}' Sample3_ref2_size.txt > Sample3_ref2_size
   ↪ .bed
```

Output:

- 2 ".bed" files containing the genetic coordinates of intergenic transcripts over 200 base pairs in length.

Notes

- N/A

Combine samples

Purpose: You are most likely working with several samples, maybe even many samples for different condition types! This step serves to generate a list of intergenic transcripts for each condition. The following command will need to be run for each collection of samples.

Command:

```
sort -u [data name]_ref1_size.bed ... [data name]_ref1_size.bed [data
   ↪ name]_ref2_size.bed ... [data name]_ref2_size.bed ... > [
   ↪ condition name].bed
```

Example:

```
sort -u Sample1_ref1_size.bed Sample1_ref2_size.bed Sample2_ref1_size.
   ↪ bed Sample2_ref2_size.bed Sample3_ref1_size.bed
   ↪ Sample3_ref2_size.bed > All.bed
```

Output:

- A ".bed" file for each condition, containing unique intergenic transcripts over 200 bases in length

Notes:

- Be sure you include both reference genomes in this command, this is the point where they combine
- From this point forward, all commands will need to be ran on each condition

Split bed file

Purpose: In order to avoid errors and loss of information, files should be split to represent known chromosomes and unknown chromosomes.

Command:

```
grep 'chrUn*' [condition name].bed > [condition name]_ChrUn.bed
grep -v 'chrUn*' [condition name].bed > tmp1
sort -rk2 tmp1 > tmp2
tail -n+2 tmp > [condition name]_ChrKnown.bed
rm tmp
```

Example:

```
grep 'chrUn*' All.bed > All_ChrUn.bed
grep -v 'chrUn*' All.bed > tmp1
sort -rk2 tmp1 > tmp2
tail -n+2 tmp2 > All_ChrKnown.bed
rm tmp*
```

Output:

- A ".bed" file containing coordinates for intergenic transcripts over 200 bases in length located on known chromosomes
- A ".bed" files containing coordinates for intergenic transcripts over 200 bases in length located on unknown chromosomes

Notes:

- This command is very similar to the command in step 9, the addition of "tail -n +2" is necessary to ensure the file format matches what further tools expect

Obtain fasta file

Purpose: A further filtering step requires the nucleotide sequences of interest for analysis. This step functions to use the BED files previously generated to isolate nucleotide sequences of interest.

Command:

```
bedtools getfasta -fi [unknown chromosome].fa -bed [condition name]
   ↪ _ChrUn.bed -fo [condition name]_ChrUn.fasta
bedtools getfasta -fi [reference fasta].fa -bed [condition name]
   ↪ _ChrKnown.bed -fo [condition name]_ChrKnown.fasta
```

Example:

```
bedtools getfasta -fi chrUn.fa -bed All_ChrUn.bed -fo All_ChrUn.fasta
bedtools getfasta -fi canFam6.fa -bed All_ChrKnown.bed -fo
   ↪ All_ChrKnown.fasta
```

Output:

- A ".fasta" file containing nucleotide sequences for intergenic transcripts over 200 bases in length located on known chromosomes
- A ".fasta" files containing nucleotide sequences for intergenic transcripts over 200 bases in length located on unknown chromosomes

Notes:

- There is no need to include "software/" in this command
- If any ".fai" files were generated, they can be removed

Reconstitute fasta file

Purpose: This step functions to reconstitute fasta files generated in step 19 to represent all query sequences in a single file. This avoids confusion.

Command:

```
cat [condition name]_ChrUn.fasta [condition name]_ChrKnown.fasta > [
    ↪ condition name].fasta
```

Example:

```
cat All_ChrUn.fasta All_ChrKnown.fasta > All.fasta
```

Output:

- A fasta file with nucleotide sequences for all intergenic transcripts over 200 base pairs in length on both known and unknown chromosomes

Notes:

- You can now remove the known and unknown fasta files generated in step 18 ([condition name]_ChrUn.fasta and [condition name]_ChrKnown.fasta)

Find reverse transcript

Purpose: To be as thorough as possible, the reverse compliment of each sequence of interest also needs to be analyzed for coding potential in a later step. This step serves to generate the reverse compliment sequence for all sequences of interest.

Command:

```
software/seqtk*/seqtk seq -r [condition name].fasta > [condition name]
    ↪ _RC.fasta
```

Example:

```
software/seqtk*/seqtk seq -r All.fasta > All_RC.fasta
```

Output:

- A fasta file with the reverse compliment to nucleotide sequences for all intergenic transcripts over 200 base pairs in length on both known and unknown chromosomes

Notes:

- N/A

Coding Potential Calculator

Purpose: Created to gauge the coding potential of transcripts based on sequence features like open reading frames, the coding potential calculator classifies provided transcripts as either coding or noncoding. In this step, sequences of interest and their reverse complement will be marked as either coding or noncoding. Given the nature of lncRNA, this will allow for removal of coding transcripts as they cannot be lncRNA.

Command:

```
software/CPC*/bin/CPC2.py -i [condition name].fasta -o [condition name
    ↪ ]_CPC
software/CPC*/bin/CPC2.py -i [condition name]_RC.fasta -o [condition
    ↪ name]_RC_CPC
```

Example:

```
software/CPC*/bin/CPC2.py -i All.fasta -o All_CPC
software/CPC*/bin/CPC2.py -i All_RC.fasta -o All_RC_CPC
```

Output:

- A ".txt" file containing the identifier, the coding potential score, and the classification (coding or noncoding) for each transcript of interest
- A ".txt" file containing the identifier, the coding potential score, and the classification (coding or noncoding) for the reverse complement of each transcript of interest

Notes:

- When calling the CPC2.py command, be sure to remember the /bin/ section of the path name
- You can remove the fasta files generated by step 19 and step 20
- This is a python command. If you need to troubleshoot an error, this information is critical

Coding Potential Filtration

Purpose: Now that transcripts have been labeled as coding or noncoding, you can eliminate those deemed to be coding. That is the function of this step.

Command:

```
awk '$8 == "noncoding"' [condition name]_CPC.txt > [condition name]_NC
    ↪ .txt
awk '$8 =="noncoding"' [condition name]_RC_CPC.txt > [condition name]
    ↪ _RC_NC.txt
```

Example:

```
awk '$8 == "noncoding"' All_CPC.txt > All_NC.txt
awk '$8 =="noncoding"' All_RC_CPC.txt > All_RC_NC.txt
```

Output:

- 2 ".txt" files containing entries deemed "noncoding" from the forward transcripts of interest and the reverse compliment to transcripts of interest

Notes:

- In this command, there is both a single and a double quotation mark, be careful not to miss either

Generate noncoding lists

Purpose: The Coding Potential Calculator provides lots of information, however not all of it is necessary for continuing your analysis. This step pars down the noncoding entries to only contain the information required.

Command:

```
awk '{print $1}' [condition name]_NC.txt > [condition name]_NC.list
awk '{print $1}' [condition name]_RC_NC.txt > [condition name]_RC_NC.
   ↪ list
```

Example:

```
awk '{print $1}' All_NC.txt > All_NC.list
awk '{print $1}' All_RC_NC.txt > All_RC_NC.list
```

Output:

- 2 text based files containing only the identifiers for noncoding transcripts

Notes:

- The file extension of these files is non-traditional, these are an intermediate step that will be combined into a more succinct list in the next step

Finalize noncoding list

Purpose: This step will combine the results from the forward and reverse compliment analysis into a consolidated list of intergenic, noncoding transcripts that are over 200 base pairs in length.

Command:

```
sort -u [condition name]*.list >> [condtion name]_NC.txt
```

Example:

```
sort -u All*.list >> All_NC.txt
```

Output:

- A ".txt" file containing the unique identifies of noncoding, intergenic transcripts over 200 base pairs in length

Notes:

- You can remove both intermediate files with the file extension ".list"

Convert file

Purpose: Currently the file containing transcripts being investigated holds chromosome identifiers in a "Chr:start-end" format. This information needs to be converted to a BED file format for further use. This step acts to change the format of the file to match what's needed.

Command:

```
sed "s,:,\t,g" [condition name]_NC.txt >> [condition name]_NC.bed
sed -i "s,-,\t,g" [condition name]_NC.bed
```

Example:

```
sed "s,:,\t,g" All_NC.txt >> All_NC.bed
sed -i "s,-,\t,g" All_NC.bed
```

Output:

- A ".bed" file containing the genetic coordinates of noncoding, intergenic transcripts over 200 base pairs in length

Notes:

- The output file generated by this step has the same name as the input file, note the change in file extension
- There is no output file indicated in the second step, this is not a mistake. The "-i" flag leads to changes to the file being made in place.

Split bed file

Purpose: In order to avoid errors and loss of information, files should be split to represent known chromosomes and unknown chromosomes.

Command:

```
grep 'chrUn*' [condition name]_NC.bed > [condition name]_NC_ChrUn.tmp
grep -v 'chrUn*' [condition name]_NC.bed | tail -n+2 > [condition name
    ↪ ]_NC_ChrKnown.tmp
awk {'print $1"\t"$2"\t"$3'} [condition name]_NC_ChrUn.tmp > [
    ↪ condition name]_NC_ChrUn.bed
awk {'print $1"\t"$2"\t"$3'} [condition name]_NC_ChrKnown.tmp > [
    ↪ condition name]_NC_ChrKnown.bed
```

Example:

```
grep 'chrUn*' All_NC.bed > All_NC_ChrUn.tmp
grep -v 'chrUn*' All_NC.bed | tail -n+2 > All_NC_ChrKnown.tmp
awk {'print $1"\t"$2"\t"$3'} All_NC_ChrUn.tmp > All_NC_ChrUn.bed
awk {'print $1"\t"$2"\t"$3'} All_NC_ChrKnown.tmp > All_NC_ChrKnown.bed
```

Output:

- A ".bed" file containing coordinates for intergenic transcripts over 200 bases in length located on known chromosomes
- A ".bed" files containing coordinates for intergenic transcripts over 200 bases in length located on unknown chromosomes

Notes:

- This is similar to step 19 with an additional print command to clean up the bed files.
- The ".tmp" files can be removed

Obtain fasta file

Purpose: This step functions to use the BED files previously generated to isolate nucleotide sequences of interest. These sequences will be converted to the protein sequence for a future filtering step.

Command:

```
bedtools getfasta -fi [unknown chromosome].fa -bed [condition name]
    ↪ _NC_ChrUn.bed
-fo [condition name]_NC_ChrUn.fasta
```

```
bedtools getfasta -fi [reference fasta].fa -bed [condition name]
    ↪ _NC_ChrKnown.bed -fo [condition name]_NC_ChrKnown.fasta
```

Example:

```
bedtools getfasta -fi chrUn.fa -bed All_NC_ChrUn.bed -fo All_NC_ChrUn.
    ↪ fasta
bedtools getfasta -fi canFam6.fa -bed All_NC_ChrKnown.bed -fo
    ↪ All_NC_ChrKnown.fasta
```

Output:

- A ".fasta" file containing nucleotide sequences for noncoding, intergenic transcripts over 200 bases in length located on known chromosomes
- A ".fasta" files containing nucleotide sequences for noncoding, intergenic transcripts over 200 bases in length located on unknown chromosomes

Notes:

- There is no need to include "software/" in this command
- Aside from the input and output file, this command is identical to step 19
- If any ".fai" files were generated during this step, they can be removed

Reconstitute fasta file

Purpose: This step functions to reconstitute fasta files generated in step 28 to represent all query sequences in a single file. This avoids confusion.

Command:

```
cat [condition name]_NC_ChrUn.fasta [condition name]_NC_ChrKnown.fasta
   ↪  > [condition name]_NC.fasta
```

Example:

```
cat All_NC_ChrUn.fasta All_NC_ChrKnown.fasta > All_NC.fasta
```

Output:

- A fasta file with nucleotide sequences for all noncoding, intergenic transcripts over 200 base pairs in length on both known and unknown chromosomes

Notes:

- You can now remove the known and unknown fasta files generated in step 26 ([condition name]_NC_ChrUn.fasta and [condition name]_ChrKnown.fasta)
- Aside from input and output file, this command is identical to step 20

Move codon.txt

Purpose: A future step requires a text file holding codons to convert nucleotide sequences to protein sequences. This step acts to move the necessary text file to the current working area to avoid future errors.

Command:

```
cp software/lncRNA-Identification-Pipeline-main/codon.txt .
```

Example:

```
cp software/lncRNA-Identification-Pipeline-main/codon.txt .
```

Output:

- A ".txt" file containing genetic codons

Notes:

- It may be easy to miss so note the period at the end of the command

Convert to protein sequence

Purpose: The next filtering step requires protein sequences, as opposed to nucleotide sequences, this step functions to convert the nucleotide sequences generated in step 27 to protein sequences.

Command:

```
software/lncRNA-Identification-Pipeline-main/sixFrameTranslate.pl -i [
    ↪ condition name]_NC.fasta -o [condition name].faa -l 6
```

Example:

```
software/lncRNA-Identification-Pipeline-main/sixFrameTranslate.pl -i
    ↪ All_NC.fasta -o All.faa -l 6
```

Output:

- A ".faa" file containing protein sequences for each nucleotide sequences provided in the fasta file generated in step 27

Notes:

- You can remove the fasta file generated in step 27 at this point if space is a concern

Remove excessively long transcripts

Purpose: The protein sequence parsing database function is unable to handle protein sequences that exceed 100,000 amino acids in length. A transcript of this size is very likely to contain a protein domain of interest so it would be filtered out in subsequent steps regardless, this step retains only protein sequences shorter than 100,000 amino acids.

Command:

```
awk 'BEGIN {RS = ">" ; ORS = ""} length($2) <= 100000 {print ">"$0}' [
    ↪ condition name].faa > [condition name]_Pfam_Ready.faa
```

Example:

```
awk 'BEGIN {RS = ">" ; ORS = ""} length($2) <= 100000 {print ">"$0}'
    ↪ All.faa > All_Pfam_Ready.faa
```

Output:

- A ".faa" file containing protein sequences for each nucleotide sequences provided in the fasta file generated in step 27, excluding those over 100,000 amino acids in length

Notes:

- N/A

Split protein files

Purpose: The protein sequence file can be quite large and time consuming to analyze as a single file, to speed up the process, as well as avoid having to restart the analysis if the

analysis is interrupted for whatever reason, this step will split the protein sequence file into many smaller pieces.

Command:

```
split --additional-suffix=.faa [condition name]_Pfam_Ready.faa [
   ↪ condition name]_part
```

Example:

```
split --additional-suffix=.faa All_Pfam_Ready.faa All_part
```

Output:

- Many ".faa" files holding a portion of the sequences from the ".faa" file generated in step 30

Notes:

- Don't be intimidated by the number of files! They will be analyzed and cleaned up in the next 3 steps.

Parse Pfam database

Purpose: This step will compare the protein sequences generated in step 29 to the PFam database to identify those with protein family domains within them. If a transcript has protein domain(s) within it, it can be removed in a subsequent step as the transcript is likely a functional protein.

Command:

```
for files in [condition name]_part*; do software/hmmer*/src/hmmsearch
   ↪ --tblout $files.txt [Pfam database] $files; rm $files; done
```

Example:

```
for files in All_part*; do software/hmmer*/src/hmmsearch --tblout
   ↪ $files.txt [Pfam database] $files; rm $files; done
```

Output:

- Many ".txt" files containing information about matches, or lack thereof, to known protein domains

Notes:

- This command looks different than other commands as it is a loop running over all the partial files generated in step 31
- This command removes the partial ".faa" files generated in step 31 automatically, there will still be many files present as there is one output for every partial file that went in to the command
- Do not panic over the large number of ".txt" files, they will be combined and tidied within the next steps
- This step will take a while to run, that is normal

Merge all outputs

Purpose: This step will combine the many outputs from the Pfam search performed in step 32 into a single table for further filtering and analysis.

Command:

```
cat [condition name]_part*.txt > [condition name]_Pfam.txt
```

Example:

```
cat All_part*.txt > All_Pfam.txt
```

Output:

- A ".txt" file containing all results from the Pfam database search

Notes:

- Don't worry about removing the partial results files, this will be done in a single command in the next step

Clean up

Purpose: This step will remove the partial results files from step 32.

Command:

```
rm [condition name]_part*
```

Example:

```
rm All_part*
```

Output:

- There is no output from this step

Notes:

- This should make the working space look much less overcrowded

Filter Pfam list

Purpose: This step will filter the results from the Pfam search to limit results to those that do not contain a protein family. The results at this point will only include transcripts that are intergenic, noncoding, lacking a protein family, and over 200 bases in length.

Command:

```
awk '(NR==1) || ($5 > .1)' [condition name]_Pfam.txt > [condition name
    ↪ ]_Pfam_filtered.txt
```

Example:

```
awk '(NR==1) || ($5 > .1)' All_Pfam.txt > All_Pfam_filtered.txt
```

Output:

- A ".txt" file containing only transcripts without a protein family. As mentioned in the purpose portion of this step, these transcripts are intergenic, noncoding, lacking a protein family, and over 200 bases in length.

Notes:

- If you open this text file, it's not the prettiest. It will be tidied in the next step

List tidying

Purpose: The output of the Pfam database search is not the prettiest and is not ready for the final filtering step, this step cleans the text file up and prepares it for the final filtering step.

Command:

```
awk '{print $1 }' [condition name]_Pfam_filtered.txt > tmp1
grep -v "# *" tmp1 > tmp2
sed -i 's/|/\t/g' tmp2
awk '{print $1 }' tmp2 > tmp3
sort -u tmp3 > [condition name]_Pfam.txt
rm tmp*
```

Example:

```
awk '{print $1 }' All_Pfam_filtered.txt > tmp1
grep -v "# *" tmp1 > tmp2
sed -i 's/|/\t/g' tmp2
awk '{print $1 }' tmp2 > tmp3
sort -u tmp3 > All_Pfam.txt
rm tmp*
```

Output:

- A ".txt" file containing transcripts that are intergenic, noncoding, lacking a protein family, and over 200 bases in length in an easy to read format

Notes:

- This step does a lot of small tasks in one go, be sure to not miss small details like quotation marks and open and closing brackets
- These commands are best typed to avoid formatting issues

Convert to bed file

Purpose: Similar to previous steps, the text file generated in step 36 is not in the appropriate format for further steps so this step will convert the file to be in the correct format.

Command:

```
sed -i 's/:/\t/g' [condition name]_Pfam.txt
sed 's/-/\t/g' [condition name]_Pfam.txt > [condition name]_Pfam.bed
```

Example:

```
sed -i 's/:/\t/g' All_Pfam.txt
sed 's/-/\t/g' All_Pfam.txt > All_Pfam.bed
```

Output:

- A ".bed" file containing genomic coordinates for transcripts that are intergenic, noncoding, lacking a protein family, and over 200 bases in length

Notes:

- This command is similar to other steps where our files have been converted to bed files

Split bed file

Purpose: In order to avoid errors and loss of information, files should be split to represent known chromosomes and unknown chromosomes.

Command:

```
grep 'chrUn*' [condition name]_Pfam.bed > [condition name]_Pfam_ChrUn.
    ↪ bed
grep -v 'chrUn*' [condition name]_Pfam.bed | tail -n+2 > [condition
    ↪ name]_Pfam_ChrKnown.bed
```

Example:

```
grep 'chrUn*' All_Pfam.bed > All_Pfam_ChrUn.bed
grep -v 'chrUn*' All_Pfam.bed | tail -n+2 > All_Pfam_ChrKnown.bed
```

Output:

- A ".bed" file containing coordinates for intergenic, noncoding, protein family lacking, transcripts over 200 bases in length located on known chromosomes
- A ".bed" files containing coordinates for intergenic, noncoding, protein family lacking, transcripts over 200 bases in length located on unknown chromosomes

Notes:

- Aside from input and output files, this step is identical to step 17 and 25

Obtain fasta file

Purpose: This step functions to use the BED files previously generated to isolate nucleotide sequences of interest for the final filtering step.

Command:

```
bedtools getfasta -fi [unknown chromosome].fa -bed [condition name]
    ↪ _Pfam_ChrUn.bed -fo [condition name]_Pfam_ChrUn.fasta
bedtools getfasta -fi [reference fasta].fa -bed [condition name]
    ↪ _Pfam_ChrKnown.bed -fo [condition name]_Pfam_ChrKnown.fasta
```

Example:

```
bedtools getfasta -fi chrUn.fa -bed All_Pfam_ChrUn.bed -fo
    ↪ All_Pfam_ChrUn.fasta
bedtools getfasta -fi canFam6.fa -bed All_Pfam_ChrKnown.bed -fo
    ↪ All_Pfam_ChrKnown.fasta
```

Output:

- A ".fasta" file containing nucleotide sequences for protein family lacking, noncoding, intergenic transcripts over 200 bases in length located on known chromosomes
- A ".fasta" files containing nucleotide sequences for protein family lacking, noncoding, intergenic transcripts over 200 bases in length located on unknown chromosomes

Notes:

- Aside from the input and output file, this command is identical to step 18 and 26
- If any .fai files were generated, they can be removed at this point
- You can remove the two partial bed files from step ([condition name]_Pfam_ChrUn.bed and [condition name]_Pfam_ChrKnown.bed)

Reconstitute fasta file

Purpose: This step functions to reconstitute fasta files generated in step 39 to represent all query sequences in a single file. This avoids confusion.

Command:

```
cat [condition name]_Pfam_ChrUn.fasta [condition name]_Pfam_ChrKnown.
   ↪ fasta > [condition name]_Pfam.fasta
```

Example:

```
cat All_Pfam_ChrUn.fasta All_Pfam_ChrKnown.fasta > All_Pfam.fasta
```

Output:

- A fasta file with nucleotide sequences for all protein family lacking, noncoding, intergenic transcripts over 200 base pairs in length on both known and unknown chromosomes

Notes:

- You can now remove the known and unknown fasta files generated in step 41 ([condition name]_Pfam_ChrUn.fasta and [condition name]_Pfam_ChrKnown.fasta)
- Aside from input and output file, this command is identical to step 19 and 27

Prepare BLAST database

Purpose: In order to perform the final filtering step, the known gene database needs to be prepared for parsing. This step prepares that database.

Command:

```
software/ncbi*/bin/makeblastdb -in [cds genome].fa -out BLAST_DB -
   ↪ parse_seqids -dbtype nucl
```

Example:

```
software/ncbi-blast-2.14.0+/bin/makeblastdb -in Canis_lupus_familiaris
   ↪ .ROS_Cfam_1.0.cds.all.fa -out BLAST_DB -parse_seqids -dbtype
   ↪ nucl
```

Output:

- 10 BLAST_DB files that are ready for parsing in the next step

Notes:

- Be sure you are using the cds file, if you use another file, your analysis will run but will be incorrect.
- Be sure your cds file is unzipped, so there is no file type other than ".fa" or ".fasta"

Run BLAST

Purpose: The final filtering step is to compare the remaining transcripts to known genes to remove those statistically significantly similar to said known genes. Most known genes have already been removed by prior filtering steps, this step will remove anything that previously slipped through the cracks.

Command:

```
software/ncbi*/bin/blastn -db BLAST_DB -query [condition name]_Pfam.
   ↪ fasta -outfmt 6 -max_target_seqs 1 -max_hsps 1 -out [condition
   ↪ name]_BLAST.txt
```

Example:

```
software/ncbi-blast-2.14.0+/bin/blastn -db BLAST_DB -query All_Pfam.
   ↪ fasta -outfmt 6 -max_target_seqs 1 -max_hsps 1 -out All_BLAST.
   ↪ txt
```

Output:

- A ".txt" file containing BLAST results for all query sequences generated from previous filtering steps

Notes:

- This command will likely generate a warning recommending more matches be analyzed, since the function of this step is only to remove those with a significant match, this warning can be disregarded
- You can remove the BLAST_DB files made in step 43 once this step has been completed

Filter BLAST results

Purpose: This step will limit the results from the BLAST search to only those deemed statistically significant.

Command:

```
awk '(NR==1) || ($11 < .1)' [condition name]_BLAST.txt > [condition
   ↪ name]_BLAST_Matches.txt
```

Example:

```
awk '(NR==1) || ($11 < .1)' All_BLAST.txt > All_BLAST_Matches.txt
```

Output:

- A ".txt" file containing transcripts that match to known genes at a statistically significant level

Notes:

- Unlike previous steps, this list is made up of transcripts that are known genes. This will be used in a future step to remove transcripts
- This command should look similar, it is reminiscent of step 35

Generate match list

Purpose: The BLAST output contains lots of useful information, however, it is not all necessary for the next step. This step removes extraneous information.

Command:

```
awk '{print $1}' [condition name]_BLAST_Matches.txt > [condition name]
   ↪ _BLAST_Matches.list
```

Example:

```
awk '{print $1}' All_BLAST_Matches.txt > All_BLAST_Matches.list
```

Output:

- A ".list" file containing the location of transcripts matching known genes in the "chr:start-end" format

Notes:

- This command and step is similar to step 22

Convert list file format

Purpose: Ahead of the final step, the BLAST hits file needs to be converted into a BED file format. This step does that.

Command:

```
sed -i 's/:/\t/g' [condition name]_BLAST_Matches.list
sed 's/-/\t/g' [condition name]_BLAST_Matches.list > [condition name]
    ↪ _BLAST_Matches.bed
```

Example:

```
sed -i 's/:/\t/g' All_BLAST_Matches.list
sed 's/-/\t/g' All_BLAST_Matches.list > All_BLAST_Matches.bed
```

Output:

- A ".bed" file containing genomic coordinates for transcripts that returned a statistically significant BLAST hit

Notes:

- This is the same command as step 37 with different input/output files

Generate final lncRNA list

Purpose: You've reached the end! This final step will compare the list of transcripts that went into the BLAST to the list of those deemed to be statistically significant matches and will report back only those that did not generate a BLAST hit. The transcripts returned by this command will be intergenic, noncoding, lacking a protein domain, over 200 base pairs in length, and do not match a known gene within the organism. This is good evidence that these transcripts, although transcribed, are not translated into proteins, and therefore are good candidates for lncRNA.

Command:

```
diff [condition name]_BLAST_Matches.bed [condition name]_Pfam.bed |
   ↪ sed -n -e 's/^> //p' > [condition name]_lncRNA.bed
```

Example:

```
diff All_BLAST_Matches.bed All_Pfam.bed | sed -n -e 's/^> //p' >
   ↪ All_lncRNA.bed
```

Output:

- A final ".bed" file containing coordinates of filtered transcripts that are likely to be lncRNA!!

Notes:

- Be sure to note the inputs of this command are the BED file from step 39 and the BED file from step 47

5.4 Further Analysis and Notes

5.4.1 Further Analysis

Expression

One avenue of investigation that is often pursued with lncRNA is expression analysis.
A tool that can be used for this analysis is Salmon. Salmon is ”a lightweight method for quantifying transcript abundance from RNA–seq reads” that can easily be used to assess lncRNA expression levels. This can be done by creating a set of index files based on a provided reference genome. These index files can then be used to quantify expression of transcripts within the index. Salmon provides 2 excellent tutorials for installation and usage, they can be located at the following locations: link and link.
For lncRNA quantification, the lncRNA list generated by this tutorial can be used with the bedtools function to generate a lncRNA fasta file. This is similar to steps 19, 28, and 41 in this tutorial. This fasta file can be provided to Salmon to generate the index files. The original sample files can then be assessed with this tool and index files to quantify lncRNA expression in each sample.
Salmon also possess a tool called ”quantmerge” that can combine the quantification files into one file. There is limited online documentation for this tool so using the command: ”salmon quantmerge –help” will provide the documentation for this tool’s usage.
Here is an example using the data from this tutorial, beginning in the directory where the analysis was done:

```
bedtools getfasta -fi canFam6.fa -fo lncRNA.fasta -bed All_lncRNA.bed
mkdir Salmon
cd Salmon
```

```
wget https://github.com/COMBINE-lab/salmon/releases/download/v1.5.1/
    ↪ salmon-1.5.1_linux_x86_64.tar.gz
tar -xvf salmon*
salmon*/bin/salmon index -t ../lncRNA.fasta -i Index
salmon*/bin/salmon quant -i Index -l A -1 ../SRR18899314_1.fastq.gz -2
    ↪  ../SRR18899314_2.fastq.gz -p 8 --validateMappings -o quants/
    ↪ Sample1_quant
salmon*/bin/salmon quant -i Index -l A -r ../SRR24066292.fastq.gz -p 8
    ↪  --validateMappings -o quants/Sample2_quant
salmon*/bin/salmon quant -i Index -l A -r ../SRR24066293.fastq.gz -p 8
    ↪  --validateMappings -o quants/Sample3_quant
salmon*/bin/salmon quantmerge --quants quants/Sample1_quant/ quants/
    ↪ Sample2_quant/ quants/Sample3_quant/ -o lncRNA_Expression.txt
```

The resulting files can be used as is or opened in programs such as Excel for reporting. The ”quant.sf” files for each sample contain the lncRNA name, length, effective length, expression as measured in transcripts per million (TPM), and number of reads. The output file from the quantmerge command contains the lncRNA name and expression in each merged sample. This tool can also be used with a standard reference genome and used to obtain expression data for all transcripts as well [16].

Conservation

Another investigation technique many researchers take on with lncRNA is their conservation score. This can be done with the UCSC tool bigWigAverageOverBed and a bigwig file containing conservation scores for multiple species. The caveat with the tool is

the alignment files tend to be based on human genome locations and therefore identified lncRNA must be converted to human coordinates first.

Luckily, there is an online tool that can easily do this conversion for you. Located here: LiftOver, you can set your original species, upload the lncRNA bed files generated by this tutorial, and press submit. The tool will convert your genomic locations to human equivalents, and you can click ”View Conversions” to download a new bed file with human genomic coordinates. This file can then be renamed and moved to your analysis location for the next steps [14].

After converting your lncRNA to human coordinates, your bed file will have 5 columns: human chromosome, human start position, human end position, the original transcript ID, and a number indicating if the transcript is repeated and if so, what occurrence of the transcript each new coordinate represents. There is one more modification that needs to be done to prepare your bed file, the fourth and fifth column need to be combined to act as a label for each lncRNA investigated. This can easily be done with a single command:

```
awk '{print $1"\t"$2"\t"$3"\t"$4"."$5}' [Converted lncRNA] > [Modified
    ↪ File]
```

Once your coordinates have been converted and a label has been added, you can use your file and a conservation file in the bigwig (.bw) format to assign a conservation score for each transcript. These scores will be between 0 and 1, with a higher score indicating the transcript is more conserved. Before this analysis can be performed, you’ll need said conservation file. These files can be found here: PhastCon Files. There are two folders that contain the needed files, the phastCons46way folder and the phastCons100way folder. The phastCons46way folder holds data comparing human transcripts to 45 vertebrates; there are 3 options in this folder, all 45 vertebrate comparisons, placental mammal comparisons, and primate comparisons. The phastCons100way folder holds data comparing human

transcripts to 99 vertebrate genomes; there is only one bigwig file within this folder. The file you choose to use is up to you, there are details about the species used in the comparisons within the preamble text in each folder [12]. You can click on the file name you choose and it will download, then it can be moved to the analysis directory or you can use the following command to download the file directly to your analysis folder:

```
wget [link to phastCon file]
```

Now that the lncRNA file has been converted and the conservation file obtained, the Phastcon analysis can be performed. The following documentation explains the options available with the UCSC bigWigAverageOverBed function: link.
Here is an example of obtaining and using the necessary files and tools, having already downloaded the converted lncRNA file, moved it to the analysis directory, and working in the analysis directory:

```
mkdir Phastcon
mv lncRNA_To_Human.bed Phastcon/
cd Phastcon/
awk '{print $1"\t"$2"\t"$3"\t"$4"."$5}' lncRNA_To_Human.bed >
    ↪ lncRNA_To_Human_Labeled.bed
wget https://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons46way/
    ↪ placentalMammals.phastCons46way.bw
rsync -aP \rsync://hgdownload.soe.ucsc.edu/genome/admin/exe/linux.
    ↪ x86_64/bigWigAverageOverBed ../software
../software/bigWigAverageOverBed placentalMammals.phastCons46way.bw
    ↪ lncRNA_To_Human_Labeled.bed Phastcon.tab
```

The .tab file has 2 columns you should pay attention to: the first column that is the identifier (the previous steps in this tutorial means these identifiers are the lncRNA identified originally) and the last column is the average score over the covered bases. The last column is the phastCon score for the regions investigated. You now have these scores for reporting and further analysis. This analysis can be performed again on any group of data, you may want to perform it on all transcripts and intergenic transcripts for comparison purposes [11, 2].

Annotation

Many research endeavors intend to assign a role to identified lncRNA. Although definitively assigning lncRNA a role is very difficult without choosing a single lncRNA to research thoroughly through modeling and manipulation, it is possible to associate lncRNA with potential roles and open doors for future research endeavors. This can be achieved in many ways, below is a summary of some of these ways. It may be beneficial to your research to find a contemporaneous study and examine how they annotated lncRNA. A useful tool for annotation comes from the UCSC Genome browser. You may choose to use the genome browser and examine the locations identified as lncRNA. There may be genes nearby that the lncRNA could interact with. The genome browser also allows built in comparative genomics, you can look at a region of interest in your organism and see if there is overlap in other species (predominantly human and mice genomes). A more specific way to annotate lncRNA is based on sequence homology. If a transcript has a similar sequence to other organisms, perhaps they act in similar ways. Phastcon scores can easily allow you to assess if a sequence is highly conserved. These more highly conserved regions can be retained for analysis like any other transcript, just with the knowledge that it is a more conserved region. The UCSC genome browser is full of information and tools,

it is worth playing around with and reading up on [23].

Another way lncRNA can be associated with lncRNA is by correlation. If the expression of lncRNA is correlated with a gene, perhaps they're working towards the same goal. In order to generate a correlation matrix, you would need a single file containing expression levels for all lncRNA and all transcripts in samples. This file could then use a tool such as Excel or the R function "corr" to generate a correlation matrix. Depending on how you go about generating this matrix, different data preparation may be required. A few online searches would likely help you prepare your data for this analysis.

There are also several online tools that can be used with lists of genes to associate them with functions. If you are able to identify a list of associated genes (perhaps by proximity or correlation with a lncRNA), tools like the gene ontology resource, DAVID bioinformatics resources, and the PANTHER classification system will allow you to enter said list and uncover associations. The gene ontology resource will report what biological processes are associated with the specific gene collection [7]. DAVID is able to report associated diseases, functional annotations, gene ontology, interactions, pathways, protein domains, tissue specific expression, and associated literature–this is assuming though, that these associations exist [5]. PANTHER is able to report functional classification and statistical enrichment for gene lists [21]. These tools all have their uses and will be variably useful to you depending on your data set and identified genes. Further research and investigation into each tool will allow you to make an educated decisions about what tools to use.

Other

There are many more tools that can be used to analyze lncRNA in effective manners, such as enrichment analysis and interaction network predictions. As previously recommended,

reading of literature similar to your project can be a great tool for deciding how you want to analyze your data.

5.4.2 Notes

The final section of this tutorial will provide a few notes of common problems you may encounter and how to navigate them.

Data Management

As mentioned very early in this tutorial, there are many files downloaded and generated that are quite large and can pose a challenge if you have a limited amount of space. For this reason, there are notes indicating what files can be removed when. The removal of these files is strategic, they do not intend to delete files that may be used in reporting or analysis of findings. For instance, the original files are not recommended to be removed as they can be used with tools like Salmon later in analysis; files such as the first Hisat alignment files are recommended to be removed as more accurate alignment files will be generated and the information in these files is very unlikely to be used later in the analysis. If you are truly lacking space, removing of fasta files can help conserve space without lose of information (as the fasta file can be regenerated from a file with genomic coordinates). Another data management pitfall you may find comes to the naming convention used. The commands provided by this code are intended to be descriptive and clear but you are free to name files as you see fit. Although in the moment, you may feel as if you'll remember what each file is, that may not be the case when you come back to your data after taking a break. For this reason, it is recommended to keep the naming conventions as descriptive as possible without being excessive in length.

Error Troubleshooting

As you progress through the tutorial, you may encounter an error or two. It is important to know how to navigate errors as they occur.

The first step in troubleshooting an error is to read the error itself. It may seem silly but errors tend to tell you exactly what needs to be modified. It is most likely you would receive an error that states ”No such file or directory,” this error means one (or more) arguments cannot be found. The simplest solution is to ensure you haven’t made a typo in your code, a missing space or character can cause big problems. Other errors may tell you that a file is not in the right format or data needs to be sorted in a specific manner. As stated, many errors will tell you exactly what you need to fix.

If you are unable to get a command to work after checking the syntax and reading/trying to address the error, online is full of resources and kind people. Sources like BioStars, GitHub, Galaxy, and Stack Overflow are all full of people troubleshooting errors and helping each other. A few online searches should allow you to fix many errors. Finally, if you seem to have tried every solution to getting a software to run and it just will not cooperate, there is always the option to delete and re-download a tool. Sometimes things don’t go according to plan and the best course of action is to start over.

Words of Encouragement

This tutorial is very long and involved but you can do it! It is written to in small, simple steps to make it as beginner friendly as possible so don’t let the length scare you. It can be disheartening when a command doesn’t run and it seems like a solution is unattainable; take a break, get a drink and a snack and look at it with fresh eyes. You are unlikely to be the first to have this problem so someone, somewhere has solved it before and you can too!

5.5 References

[1] *amarceau-998/lncRNA-Identification-Pipeline: Collection of scripts used to identify lncRNA*. URL: https://github.com/amarceau-998/lncRNA-Identification-Pipeline.

[2] *bigWigAverageOverBed manual with usage examples — BioQueue Encyclopedia*. URL: https://open.bioqueue.org/home/knowledge/showKnowledge/sig/ucsc-bigwigaverageoverbed.

[3] *CPC2 @ CBI, PKU*. URL: http://cpc2.gao-lab.org/download.php.

[4] *Cufflinks*. URL: http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/.

[5] *DAVID Functional Annotation Bioinformatics Microarray Analysis*. URL: https://david.ncifcrf.gov/home.jsp.

[6] *Ensembl genome browser 110*. URL: https://useast.ensembl.org/index.html.

[7] *Gene Ontology Resource*. URL: http://geneontology.org/.

[8] *GitHub - lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats*. URL: https://github.com/lh3/seqtk.

[9] *GitHub - ncbi/sra-tools: SRA Tools*. URL: https://github.com/ncbi/sra-tools.

[10] *HMMER*. URL: http://hmmer.org/.

[11] *Index of /admin/exe/linux.x86_64*. URL: https://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/.

[12] *Index of /goldenpath/hg19*. URL: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/.

[13] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. "HISAT: a fast spliced aligner with low memory requirements". In: *Nature Methods 2015 12:4* 12.4 (Mar. 2015), pp. 357–360. ISSN: 1548-7105. DOI: 10.1038/nmeth.3317. URL: https://www.nature.com/articles/nmeth.3317.

[14] *Lift Genome Annotations*. URL: https://genome.ucsc.edu/cgi-bin/hgLiftOver.

[15] *Nucleotide BLAST: Search nucleotide databases using a nucleotide query*. URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

[16] Rob Patro et al. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods 2017 14:4* 14.4 (Mar. 2017), pp. 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197. URL: https://www.nature.com/articles/nmeth.4197.

[17] *Pfam: Search Pfam*. URL: http://pfam.xfam.org/search#tabview=tab1.

[18] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033. URL: https://doi.org/10.1093/bioinformatics/btq033.

[19] *samtools Tutorial*. URL: http://quinlanlab.org/tutorials/samtools/samtools.html.

[20] *StringTie*. URL: http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual.

[21] Paul D. Thomas et al. “PANTHER: Making genome-scale phylogenetics accessible to all”. In: *Protein Science* 31.1 (Jan. 2022), pp. 8–22. ISSN: 1469896X. DOI: 10.1002/PRO.4218.

[22] *UCSC Genome Browser Downloads*. URL: https://hgdownload.soe.ucsc.edu/downloads.html.

[23] *UCSC Genome Browser Gateway*. URL: https://genome.ucsc.edu/cgi-bin/hgGateway.

Chapter 6: Conclusion

This dissertation sought to investigate lncRNA in various *Bos taurus* tissues, considering developmental time points, as well as develop a common language lncRNA identification tutorial. The findings of this research and tutorial will expand the knowledge base of cattle lncRNA, cattle developmental processes, and allow future researchers to easily identify lncRNA without prerequisite knowledge.

In **Chapter 2**, we aimed to identify lncRNA in cattle rumen tissue before and after weaning to investigate how lncRNA may influence the crucial weaning process. lncRNA were identified using several filtering steps, and identified lncRNA were analyzed for characteristics, sequence conservation, functional annotation, and multi-trait enrichment. The findings of these analysis indicated lncRNA expression pattern changed drastically as weaning progressed, there were connections to genes associated with cell growth and epithelial tight junction formation (these junctions are found all over rumen epithelial cells), and lncRNA showed enrichment with complex traits representing critical value indicators. Overall, this study demonstrated lncRNA play a role in the development of the rumen tissue as weaning progresses and further understanding of their role can assist in the weaning process.

In **Chapter 3**, mammary tissue in dry and lactating cattle was analyzed for lncRNA and their role in lactation. lncRNA was successfully identified in both tissue conditions and profiles were reminiscent of other lncRNA profiles. Annotation based on sequence homology, gene correlation, and gene cluster ontology revealed strong connections to strong

hair keratin filaments (this is associated with resilient mammary tissue) and sense of smell (which is known to become heightened during pregnancy), as well as other genes and processes of interest. The lncRNA also showed enrichment in complex traits representing reproduction and production value. Similarly to Chapter 2, these findings indicate lncRNA are likely involved in the lactation process and further research can increase the production value of milk producing cattle.

In **Chapter 4**, the final lncRNA identification study, immune system associated tissues were analyzed for lncRNA. lncRNA profiles for each tissue demonstrated lncRNA act both in a tissue specific and a more ubiquitous manner. lncRNA were also able to be associated with several immune response processes. The integration of large-scale GWAS data also associated these lncRNA with complex traits representing daughter pregnancy rate and stature. This study demonstrated lncRNA are present in immune system associated tissues and likely play some role in the immune response. Further research into identified lncRNA may allow researchers to better combat illnesses in cattle herds.

In **Chapter 5**, a tutorial was developed to fill the gap in lncRNA identification software and documentation. This tutorial was written in small, plain language steps and is intended to be as beginner friendly as possible. Given the current tools are either complex and/or not maintained, this tutorial will improve the available tools for those seeking to identify lncRNA. By making this analysis more accessible, more research can be done into lncRNA.

In conclusion, the studies in this dissertation did uncover lncRNA in various tissues and time points, and functional annotation hypotheses were developed. Future research endeavors may select a small number of lncRNA of interest for more in depth analysis. For example, the rumen tissue lncRNA associated with the F11R epithelial tight junction formation gene may be further investigated to more clearly define its role in rumen tissue epithelial junction formation. Findings from Chapter 3 and 4 indicate more research should

be done regarding sense of smell and the lncRNA associated with it as sense of smell was repeatedly associated with lncRNA in multiple tissues and in multiple analyses. Other lncRNA of notes can also lead to improved dairy cattle lactation and immune response. Finally, the tutorial in Chapter 5 requires regular maintenance and improvement as tools improve as well. Overall, the findings of this dissertation indicate there are many avenues of investigation regarding lncRNA yet to be explored.

Bibliography

5 Best Beef Cattle Breeds. (n.d.). Retrieved from https://americancowboy.com/lifestyle/5-best-beef-cattle-breeds-24440/

Abbas, W., Howard, J. T., Paz, H. A., Hales, K. E., Wells, J. E., Kuehn, L. A., … Fernando, S. C. (9 2020). Influence of host genetics in shaping the rumen bacterial community in beef cattle. Scientific Reports 2020 10:1, 10(1), 1–14. doi:10.1038/s41598-020-72011-9

Abeler-Dörner, L., Laing, A. G., Lorenc, A., Ushakov, D. S., Clare, S., Speak, A. O., … Hayday, A. C. (12 2019). High-throughput phenotyping reveals expansive genetic and structural underpinnings of immune variation. Nature Immunology 2019 21:1, 21(1), 86–100. doi:10.1038/s41590-019-0549-0

Abstract, G. (2020). A Landscape of Murine Long Non-Coding RNAs Reveals the Leading Transcriptome Alterations in Adipose Tissue during Aging. doi:10.1016/j.celrep.2020.107694

Achawanantakun, R., Chen, J., Sun, Y., Zhang, Y. (7 2015). LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. Bioinformatics (Oxford, England), 31(24), 3897–3905. doi:10.1093/BIOINFORMATICS/BTV480

ADW: Bos taurus: INFORMATION. (n.d.). Retrieved from https://animaldiversity.org/accounts/Bostaurus/

Aldahmash, A., Lunny, D., Lane, B. (2005). Morphogenesis of the mammary gland and the role of keratin expression. Breast Cancer Research, 7(2), P3.09. doi:10.1186/bcr1204

ALDB: Cow lncRNAs. (n.d.). Retrieved from http://202.200.112.245/cgi-bin/gb2/gbrowse/bostaurus/

Altenhoff, A. M., Train, C. M., Gilbert, K. J., Mediratta, I., de Farias, T. M., Moi, D., ... Dessimoz, C. (1 2021). OMA orthology in 2021: Website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Research, 49(D1), D373–D379. doi:10.1093/NAR/GKAA1007

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (10 1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., Mattick, J. S. (1 2011). lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res, 39(suppl1), D146–D151. doi:10.1093/nar/gkq1138

amarceau-998/lncRNA-Identification-Pipeline: Collection of scripts used to identify lncRNA. (n.d.). Retrieved from https://github.com/amarceau-998/lncRNA-Identification-Pipeline

Anja Juran, S., Tognetti, A., Lundström, J. N., Kumar, L., Stevenson, R. J., Lekander, M., Olsson, M. J. (1 2023). Disgusting odors trigger the oral immune system. Evolution, Medicine, and Public Health, 11(1), 8–17. doi:10.1093/EMPH/EOAC042

Areola — anatomy — Britannica. (n.d.). Retrieved from https://www.britannica.com/science/areola

ARS-UCD1.2 - bosTau9 - Genome - Assembly - NCBI. (n.d.-b). Retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCF002263795.1/

ARS-UCD1.3 - bosTau9 - Genome - Assembly - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCF002263795.2

Baldwin VI, R. L., McLeod, K. R., Capuco, A. V. (2004). Visceral tissue growth and proliferation during the bovine lactation cycle. Journal of Dairy Science, 87(9), 2977–2986. doi:10.3168/JDS.S0022-0302(04)73429-3

Ballarin, C., Povinelli, M., Granato, A., Panin, M., Corain, L., Peruffo, A., Cozzi, B. (4 2016). The Brain of the Domestic Bos taurus: Weight, Encephalization and Cerebellar Quotients, and Comparison with Other Domestic and Wild Cetartiodactyla. PLoS ONE, 11(4), 154580. doi:10.1371/JOURNAL.PONE.0154580

Bast, B. J. E. G., Zhou, L. J., Freeman, G. J., Colley, K. J., Ernst, T. J., Munro, J. M., Tedder, T. F. (1992). The HB-6, CDw75, and CD76 differentiation antigens are unique cell-surface carbohydrate determinants generated by the beta-galactoside alpha 2,6-sialyltransferase. The Journal of Cell Biology, 116(2), 423–435. doi:10.1083/JCB.116.2.423

Bauchop, T. (1979a). Rumen Anaerobic Fungi of Cattle and Sheep. Applied and Environmental Microbiology, 38(1), 148–158. doi:10.1128/AEM.38.1.148-158.1979

Bednarski, J. J., Sleckman, B. P. (2012). Lymphocyte Development: Integration of DNA Damage Response Signaling. Advances in Immunology, 116, 175. doi:10.1016/B978-0-12-394300-2.00006-5

bigWigAverageOverBed manual with usage examples — BioQueue Encyclopedia. (n.d.). Retrieved from https://open.bioqueue.org/home/knowledge/showKnowledge/sig/ucsc-bigwigaverageoverbed

Blood and the cells it contains - Blood Groups and Red Cell Antigens - NCBI Bookshelf. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK2263/

Bocchetti, M., Scrima, M., Melisi, F., Luce, A., Sperlongano, R., Caraglia, M., … Cossu, A. M. (2 2021). LncRNAs and Immunity: Coding the Immune System with Noncoding Oligonucleotides. International Journal of Molecular Sciences, 22(4), 1–30. doi:10.3390/IJMS22041741

Bollongino, R., Burger, J., Powell, A., Mashkour, M., Vigne, J. D., Thomas, M. G. (9 2012). Modern Taurine Cattle Descended from Small Number of Near-Eastern Founders. Molecular Biology and Evolution, 29(9), 2101–2104. doi:10.1093/MOLBEV/MSS092

Bos taurus Annotation Report. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/genome/annotationeuk/Bostaurus/103/

BostaurusUMD3.1.1 - Genome - Assembly - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCF000003055.6/

Boutinaud, M., Herve, L., Lollivier, V. (2015). Mammary epithelial cells isolated from milk are a valuable, non-invasive source of mammary transcripts. Frontiers in Genetics, 6(OCT). doi:10.3389/FGENE.2015.00323/ABSTRACT

Boutinaud, M., Herve, L., Lollivier, V., Saipin, N., Noophun, J., Chumyim, P., … Maza, E. (2016). Expression of HLA and Autoimmune Pathway Genes in Liver Biopsies of Young Subjects with Autoimmune Hepatitis Type 1. Journal of Pediatric Gastroenterology and Nutrition, 15(2), 1–23. doi:10.1097/MPG.0000000000003538

Bovine lymphocyte – eClinpath. (n.d.). Retrieved from https://eclinpath.com/hematology/morphologic-features/white-blood-cells/normal-leukocytes/bvlycomp/

Brannan, C. I., Dees, E. C., Ingram, R. S., Tilghman, S. M. (1 1990). The product of the H19 gene may function as an RNA. Molecular and Cellular Biology, 10(1), 28–36. doi:10.1128/MCB.10.1.28-36.1990

Brar, G. A., Weissman, J. S. (10 2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. Nature Reviews Molecular Cell Biology 2015 16:11, 16(11), 651–664. doi:10.1038/nrm4069

Brennan, K., Zheng, J. (2007). Interleukin 8. xPharm: The Comprehensive Pharmacology Reference, 1–4. doi:10.1016/B978-008055232-3.61916-6

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., … Rastan, S. (10 1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell, 71(3), 515–526. doi:10.1016/0092-8674(92)90519-I

Brown, C. J., Lafreniere, R. G., Powers, V. E., Sebastio, G., Ballabio, A., Pettigrew, A. L., … Willard, H. F. (1991). Localization of the X inactivation centre on the human X chromosome in Xq13. Nature 1991 349:6304, 349(6304), 82–84. doi:10.1038/349082a0

Burset, M., Guigó, R. (6 1996). Evaluation of gene structure prediction programs. Genomics, 34(3), 353–367. doi:10.1006/GENO.1996.0298

C7 complement C7 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene/730

Cameron, E. L. (2014). Pregnancy and olfaction: A review. Frontiers in Psychology, Vol. 5. doi:10.3389/fpsyg.2014.00067

Cano, R. L. E., Lopera, H. D. E. (7 2013). Introduction to T and B lymphocytes. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK459471/

Casey, T., Dover, H., Liesman, J., DeVries, L., Kiupel, M., VandeHaar, M., Plaut, K. (2011). Transcriptome analysis of epithelial and stromal contributions to mammogenesis in three week prepartum cows. PloS One, 6(7). doi:10.1371/JOURNAL.PONE.0022541

cells expressing KRT14 , vimentin and PROCR corresponded to basal progenitors whereas the CD49. (2018).

cellular biosynthetic process Gene Ontology Term (GO:0044249). (n.d.). Retrieved from https://www.informatics.jax.org/vocab/geneontology/GO:0044249

Chahal, J., Gebert, L. F. R., Camargo, C., MacRae, I. J., Sagan, S. M. (8 2021). miR-122-based therapies select for three distinct resistance mechanisms based on alterations in RNA structure. Proceedings of the National Academy of Sciences of the United States of America, 118(33), e2103671118. doi:10.1073/PNAS.2103671118/SUPPLFILE/PNAS.2103671118.SAPP.PDF

Chase, C. C. L., Hurley, D. J., Reber, A. J. (3 2008). Neonatal Immune Development in the Calf and Its Impact on Vaccine Response. The Veterinary Clinics of North America. Food Animal Practice, 24(1), 87. doi:10.1016/J.CVFA.2007.11.001

Chen, K., Cerutti, A. (6 2011). The function and regulation of immunoglobulin D. Current Opinion in Immunology, 23(3), 345–352. doi:10.1016/J.COI.2011.01.006

Choi, J. Y., Shin, D., Lee, H. J., Oh, J. D. (1 2019). Comparison of long noncoding RNA between muscles and adipose tissues in Hanwoo beef cattle. Animal Cells and Systems, 23(1), 50–58. doi:10.1080/19768354.2018.1512522

Comparative Genomics Fact Sheet. (n.d.). Retrieved from https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet

Connor, E. E., Li, R. W., Baldwin, R. L., Li, C. (2010a). Gene expression in the digestive tissues of ruminants and their relationships with feeding and digestive processes. Animal, 4(7), 993–1007. doi:10.1017/S1751731109991285

Connor, Erin E., Siferd, S., Elsasser, T. H., Evock-Clover, C. M., Van Tassell, C. P., Sonstegard, T. S., … Capuco, A. V. (7 2008). Effects of increased milking frequency on gene

expression in the bovine mammary gland. BMC Genomics, 9(1), 1–14. doi:10.1186/1471-2164-9-362

CPC2 @ CBI, PKU. (n.d.). Retrieved from http://cpc2.gao-lab.org/download.php

CR2 complement C3d receptor 2 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene?Db=geneCmd=DetailsSearchTerm=1380

Cufflinks. (n.d). Retrieved from http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/

Dado-Senn, B., Skibiel, A. L., Fabris, T. F., Zhang, Y., Dahl, G. E., Peñagaricano, F., Laporta, J. (7 2018). RNA-Seq reveals novel genes and pathways involved in bovine mammary involution during the dry period and under environmental heat stress. Scientific Reports 2018 8:1, 8(1), 1–11. doi:10.1038/s41598-018-29420-8

Dai, W. T., Zou, Y. X., White, R. R., Liu, J. X., Liu, H. Y. (3 2018). Transcriptomic profiles of the bovine mammary gland during lactation and the dry period. Functional and Integrative Genomics, 18(2), 125–140. doi:10.1007/S10142-017-0580-X/FIGURES/7

Daley, V. L., Dye, C., Bogers, S. H., Michael Akers, R., Rodriguez, F. C., Cant, J. P., … Hanigan, M. D. (12 2018). Bovine Mammary Gland Biopsy Techniques. Journal of Visualized Experiments: JoVE, 2018(142). doi:10.3791/58602

DAVID Functional Annotation Bioinformatics Microarray Analysis. (n.d.). Retrieved from https://david.ncifcrf.gov/home.jsp

Davidson, N. (6 2005). The challenges in translating present knowledge of the molecular biology of breast cancer into clinical use. Breast Cancer Research, 7(S2). doi:10.1186/bcr1044

Definition of red blood cell - NCI Dictionary of Cancer Terms - NCI. (n.d.). Retrieved from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/red-blood-cell

de Freslon, I., Peralta, J. M., Strappini, A. C., Monti, G. (8 2020). Understanding Allogrooming Through a Dynamic Social Network Approach: An Example in a Group of Dairy Cows. Frontiers in Veterinary Science, 7, 535. doi:10.3389/FVETS.2020.00535/BIBTEX

De Goede, O. M., Ferraro, N. M., Nachun, D. C., Rao, A. S., Aguet, F., Barbeira, A. N., ... Montgomery, S. B. (10 2019). Long non-coding RNA gene regulation and trait associations across human tissues. bioRxiv, p. 793091. doi:10.1101/793091

de Goede, O. M., Nachun, D. C., Ferraro, N. M., Gloudemans, M. J., Rao, A. S., Smail, C., ... Volpi, S. (4 2021). Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. Cell. doi:10.1016/j.cell.2021.03.050

De Roos, A. P. W., Schrooten, C., Veerkamp, R. F., van Arendonk, J. A. M. (3 2011). Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. Journal of Dairy Science, 94(3), 1559–1567. doi:10.3168/JDS.2010-3354

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., ... Guigó, R. (9 2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research, 22(9), 1775–1789. doi:10.1101/gr.132159.111

Diao, Q., Zhang, R., Fu, T. (8 2019). Review of strategies to promote rumen development in calves. Animals, 9(8), 490. doi:10.3390/ani9080490

Digestive Anatomy in Ruminants. (n.d.). Retrieved from http://www.vivo.colostate.edu/hbooks/pathphys/digestion/herbivores/rumenanat.html

Ding, C., Li, Y., Guo, F., Jiang, Y., Ying, W., Li, D., ... He, F. (10 2016). A Cell-type-resolved Liver Proteome. Molecular Cellular Proteomics: MCP, 15(10), 3190. doi:10.1074/MCP.M116.060145

DiStefano, J. K. (2018). The emerging role of long noncoding rnas in human disease. In Methods in Molecular Biology (Vol. 1706, pp. 91–110). doi:10.1007/978-1-4939-7471-96

Download R-4.1.2 for Windows. The R-project for statistical computing. (n.d.). Retrieved from https://cran.r-project.org/bin/windows/base/

Elhalal, A., Davelaar, E. J., Usher, M. (6 2014). The role of the frontal cortex in memory: an investigation of the Von Restorff effect. Frontiers in Human Neuroscience, 0(JUNE), 410. doi:10.3389/FNHUM.2014.00410

Emms, D. M., Kelly, S. (11 2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biology, 20(1), 1–14. doi:10.1186/S13059-019-1832-Y/FIGURES/5

Ensembl genome browser 110. (n.d.). Retrieved from https://useast.ensembl.org/index.html

Eusebi, P. G., Sevane, N., O'Rourke, T., Pizarro, M., Boeckx, C., Dunner, S. (12 2021). Gene expression profiles underlying aggressive behavior in the prefrontal cortex of cattle. BMC Genomics, 22(1). doi:10.1186/S12864-021-07505-5

Explaining CRISPR-CAS9 System- Step by Step – Genetic Education. (n.d.). Retrieved from https://geneticeducation.co.in/explaining-crispr-cas9-system-step-by-step/

Explore lncRNAdb data in RNAcentral. (n.d.). Retrieved from https://rnacentral.org/expert-database/lncrnadb

Farnir, Frédéric, Coppieters, W., Arranz, J. J., Berzi, P., Cambisano, N., Grisart, B., … Georges, M. (2 2000). Extensive genome-wide linkage disequilibrium in cattle. Genome Res, 10(2), 220–227. doi:10.1101/gr.10.2.220

Finot, L., Chanat, E., Dessauge, F. (1 2018). Bovine mammary gland development: new insights into the epithelial hierarchy. bioRxiv, 251637. doi:10.1101/251637

Finot, L., Chanat, E., Dessauge, F. (12 2021). Mammary gland 3D cell culture systems in farm animals. Veterinary Research, 52(1), 1–10. doi:10.1186/S13567-021-00947-5/FIGURES/2

Firat, R. B. (2 2019). Opening the "Black Box": Functions of the Frontal Lobes and Their Implications for Sociology. Frontiers in Sociology, 4. doi:10.3389/FSOC.2019.00003/FULL

Freebern, E., Santos, D. J. A., Fang, L., Jiang, J., Parker Gaddis, K. L., Liu, G. E., … Ma, L. (1 2020). GWAS and fine-mapping of livability and six disease traits in Holstein cattle. BMC Genomics 2020 21:1, 21(1), 1–11. doi:10.1186/S12864-020-6461-Z

Gao, Yahui, Fang, L., Baldwin, R. L., Connor, E. E., Cole, J. B., Van Tassell, C. P., … Liu, G. E. (7 2021). Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning. Genomics, 113(4), 2045–2055. doi:10.1016/J.YGENO.2021.04.039

Gao, Yuan, Li, S., Lai, Z., Zhou, Z., Wu, F., Huang, Y., … Dang, R. (2019). Analysis of Long Non-Coding RNA and mRNA Expression Profiling in Immature and Mature Bovine (Bos taurus) Testes. Frontiers in Genetics, 0(JUL), 646. doi:10.3389/FGENE.2019.00646

Garcia-Campos, A., Correia, C. N., Naranjo-Lucena, A., Garza-Cuartero, L., Farries, G., Browne, J. A., … Mulcahy, G. (2019). Fasciola hepatica infection in cattle: Analyzing responses of peripheral blood mononuclear cells (PBMC) using a transcriptomics approach. Frontiers in Immunology, 10(AUG), 2081. doi:10.3389/FIMMU.2019.02081/BIBTEX

Gilaberte, Y., Prieto-Torres, L., Pastushenko, I., Juarranz, Á. (8 2016). Anatomy and Function of the Skin. Nanoscience in Dermatology, 1–14. doi:10.1016/B978-0-12-802926-8.00001-X

Gene Expression Profiling in Cancer - Google Books. (n.d.). Retrieved from https://books.google.com/books?hl=enlr=id=DJn8DwAAQBAJoi=fndpg=PA33dq=The+role+of+long+nc sig=y9asIfG9K9P4VMFrcfrdJQGfbgv=onepageq=The

Gene Ontology Resource. (n.d.). Retrieved from http://geneontology.org/

GFF/GTF File Format. (n.d.). Retrieved from http://m.ensembl.org/info/website/upload/gff.html

GitHub - jiang18/mph: MINQUE for Partitioning Heritability. (n.d.). Retrieved from https://github.com/jiang18/mph

GitHub - lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats. (n.d.). Retrieved from https://github.com/lh3/seqtk

GitHub - ncbi/sra-tools: SRA Tools. (n.d.). Retrieved from https://github.com/ncbi/sra-tools

Glover, N., Dessimoz, C., Ebersberger, I., Forslund, S. K., Gabaldón, T., Huerta-Cepas, J., ... Rogers, R. (10 2019). Advances and Applications in the Quest for Orthologs. Molecular Biology and Evolution, 36(10), 2157–2164. doi:10.1093/MOLBEV/MSZ150

GO Enrichment Analysis. (n.d.). Retrieved from https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/goenrichment/tutorial.html

Goswami, R., Kaplan, M. H. (3 2011). A Brief History of IL-9. Journal of Immunology (Baltimore, Md.: 1950), 186(6), 3283. doi:10.4049/JIMMUNOL.1003049

GPCR — Learn Science at Scitable. (n.d.). Retrieved from https://www.nature.com/scitable/topicpage/gpcr-14047471/

Han, B., He, Y., Zhang, L., Ding, Y., Lian, L., Zhao, C., ... Yang, N. (12 2017). Long intergenic non-coding RNA GALMD3 in chicken Marek's disease. Scientific Reports, 7(1), 1–13. doi:10.1038/s41598-017-10900-2

Han, J., Zhang, J., Chen, L., Shen, B., Zhou, J., Hu, B., ... Zhang, W. (2014). RNA Biology Efficient in vivo deletion of a large imprinted lncRNA by CRISPR/Cas9 View supplementary material. rNa Biology 829 rNa Biology, 11(7), 829–835. doi:10.4161/rna.29624

Han, S., Liang, Y., Li, Y., Du, W. (2016). Lncident: A Tool for Rapid Identification of Long Noncoding RNAs Utilizing Sequence Intrinsic Composition and Open Reading Frame Information. International Journal of Genomics, 2016. doi:10.1155/2016/9185496

Han, S., Liang, Y., Ma, Q., Xu, Y., Zhang, Y., Du, W., ... Li, Y. (11 2019). LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. Briefings in Bioinformatics, 20(6), 2009–2027. doi:10.1093/BIB/BBY065

Han, Z., Feng, W., Hu, R., Ge, Q., Ma, W., Zhang, W., ... Zhou, X. (12 2021). RNA-seq profiling reveals PBMC RNA as a potential biomarker for hepatocellular carcinoma. Scientific Reports, 11(1). doi:10.1038/S41598-021-96952-X

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science, 92(2), 433–443. doi:10.3168/JDS.2008-1646

He, Y., Ding, Y., Zhan, F., Zhang, H., Han, B., Hu, G., ... Song, J. (10 2015). The conservation and signatures of lincRNAs in Marek's disease of chicken. Scientific Reports, 5(1), 1–17. doi:10.1038/srep15184

Heifer Development: Puberty — The Cattle Site. (n.d.). Retrieved from https://www.thecattlesite.com/articles/901/heifer-development-puberty

Heinrichs, J. (pennsylvania S. U. (2019). Rumen Development in the Dairy Calf – DAIReXNET. Retrieved from https://dairy-cattle.extension.org/rumen-development-in-the-dairy-calf/

Hens, J. R., Wysolmerski, J. J. (10 2005). Molecular mechanisms involved in the formation of the embryonic mammary gland. Breast Cancer Research, 7(5), 220–224. doi:10.1186/BCR1306/TABLES/1

Hg38 100-way conservation alignment - genomewiki. (n.d.). Retrieved from http://genomewiki.ucsc.edu/index.php/Hg38100-wayconservationalignment

Hitte, C., Le Béguec, C., Cadieu, E., Wucher, V., Primot, A., Prouteau, A., ... Derrien, T. (6 2019). Genome-wide analysis of long non-coding RNA profiles in canine oral melanomas. Genes, 10(6). doi:10.3390/genes10060477

HMMER. (n.d.). Retrieved from http://hmmer.org/

Hoekzema, E., Barba-Müller, E., Pozzobon, C., Picado, M., Lucco, F., García-García, D., ... Vilarroya, O. (12 2016). Pregnancy leads to long-lasting changes in human brain structure. Nature Neuroscience 2016 20:2, 20(2), 287–296. doi:10.1038/NN.4458

How Cows Eat Grass — FDA. (n.d.). Retrieved from https://www.fda.gov/animal-veterinary/animal-health-literacy/how-cows-eat-grass

How Do Cows Make Milk? — Organic Valley. (n.d.). Retrieved from https://www.organicvalley.coop/blog/from-grass-to-glass-how-do-cows-make-milk/

How does the immune system work? (4 2020). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK279364/

How to Design Your gRNA for CRISPR Genome Editing. (n.d.). Retrieved from https://blog.addgene.org/how-to-design-your-grna-for-crispr-genome-editing

How to detect paralogs in a list of protein pairs in a high-throughput manner. (n.d.). Retrieved from https://www.researchgate.net/post/Howtodetectparalogsinalistofproteinpairsinahigh-throughputmanner

How to Install Python 2.7 on Ubuntu LinuxMint - TecAdmin. (n.d.). Retrieved from https://tecadmin.net/install-python-2-7-on-ubuntu-and-linuxmint/

How to reduce off-target effects and increase CRISPR editing efficiency? — Molecular-Cloud. (n.d.). Retrieved from https://www.moleculARcloud.org/p/how-to-reduce-off-target-effects-and-increase-crispr-editing-efficiency

How GMOs Are Regulated in the United States — FDA. (n.d.). Retrieved from https://www.fda.gov/food/agricultural-biotechnology/how-gmos-are-regulated-united-states

Human Gene CYTH3 (ENST00000350796.8) from GENCODE V43. (n.d.). Retrieved from https://genome.ucsc.edu/cgi-bin/hgGene?hgggene=ENST00000350796.8hggchrom=chr7hggstart=6161778hggend=6272624hggtype=

Human Gene MTCP1 (ENST00000362018.2) from GENCODE V43. (n.d.). Retrieved from https://genome.ucsc.edu/cgi-bin/hgGene?hgggene=ENST00000362018.2hggchrom=chrXhggstart=155065320hggend=155147775hgg

IL17RB interleukin 17 receptor B [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene/55540

IL6 Gene - GeneCards — IL6 Protein — IL6 Antibody. (n.d.). Retrieved from https://www.genecards.org/cgi-bin/carddisp.pl?gene=IL6

IL9R interleukin 9 receptor [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene/3581

Index of /admin/exe/linux.x8664. (n.d.). Retrieved from https://hgdownload.cse.ucsc.edu/admin/exe/linux.x8664/

Index of /goldenpath/hg19. (n.d.). Retrieved from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/

IRAK3 interleukin 1 receptor associated kinase 3 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene/11213

Jarroux, J., Morillon, A., Pinskaya, M. (2017). History, Discovery, and Classification of lncRNAs. Advances in Experimental Medicine and Biology, 1008, 1–46. doi:10.1007/978-981-10-5203-31

Javier Panei, C., Takeshima, S.-N., Omori, T., Nunoya, T., Davis, W. C., Ishizaki, H., … Aida, Y. (2013). Estimation of bovine leukemia virus (BLV) proviral load harbored by lymphocyte subpopulations in BLV-infected cattle at the subclinical stage of enzootic bovine leucosis using BLV-CoCoMo-qPCR. Retrieved from http://www.biomedcentral.com/1746-6148/9/95

Jia, X., He, Y., Chen, S. Y., Wang, J., Hu, S., Lai, S. J. (12 2020). Genome-wide identification and characterisation of long non-coding RNAs in two Chinese cattle breeds. Italian Journal of Animal Science, 19(1), 383–391. doi:10.1080/1828051x.2020.1735266

Jiang, C., Li, Y., Zhao, Z., Lu, J., Chen, H., Ding, N., … Li, X. (2 2016). Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. Oncotarget, 7(6), 7120. doi:10.18632/ONCOTARGET.6859

Jiang, J., Cole, J. B., Freebern, E., Da, Y., VanRaden, P. M., Ma, L. (6 2019). Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. Communications Biology 2019 2:1, 2(1), 1–12. doi:10.1038/s42003-019-0454-y

Jiang, R., Li, H., Huang, Y., Lan, X., Lei, C., Chen, H., Gene, G. (2020). Journal Preproofs Transcriptome Profiling of lncRNA Related to Fat Tissues of Qinchuan Cattle. doi:10.1016/j.gene.2020.144587

Johnsson, P., Lipovich, L., Grandér, D., Morris, K. V. (2014). Evolutionary conservation of long noncoding RNAs; sequence, structure, function. Biochimica et Biophysica Acta, 1840(3), 1063. doi:10.1016/J.BBAGEN.2013.10.035

Josset, L., Tchitchek, N., Gralinski, L. E., Ferris, M. T., Eisfeld, A. J., Green, R. R., … Katze, M. G. (7 2014). Annotation of long non-coding RNAs expressed in collaborative cross founder mice in response to respiratory virus infection reveals a new class of interferon-stimulated transcripts. RNA Biol, 11(7), 875–890. doi:10.4161/rna.29442

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., Gao, G. (7 2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research, 45(W1), W12–W16. doi:10.1093/NAR/GKX428

Kapranov, P., Willingham, A. T., Gingeras, T. R. (6 2007). Genome-wide transcription and the implications for genomic organization. Nature Reviews Genetics, Vol. 8, pp. 413–423. doi:10.1038/nrg2083

Kay, R. N. B., v. Engelhardt, W., White, R. G. (1980). The digestive physiology of wild ruminants. Digestive Physiology and Metabolism in Ruminants, 743–761. doi:10.1007/978-94-011-8067-236

Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H., … Zhou, H. (9 2018). Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. BMC Genomics 2018 19:1, 19(1), 1–14. doi:10.1186/S12864-018-5037-7

Khan, Y. S., Sajjad, H. (7 2021). Anatomy, Thorax, Mammary Gland. StatPearls. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK547666/

Kim, D., Langmead, B., Salzberg, S. L. (3 2015). HISAT: a fast spliced aligner with low memory requirements. Nature Methods 2015 12:4, 12(4), 357–360. doi:10.1038/nmeth.3317

Kim, D., Paggi, J. M., Park, C., Bennett, C., Salzberg, S. L. (8 2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915. doi:10.1038/S41587-019-0201-4

Kim, S., Becker, J., Bechheim, M., Kaiser, V., Noursadeghi, M., Fricker, N., … Schumacher, J. (10 2014). Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. Nature Communications 2014 5:1, 5(1), 1–7. doi:10.1038/ncomms6236

Konkel, L. (11 2018). The Brain before Birth: Using fMRI to Explore the Secrets of Fetal Neurodevelopment. Environmental Health Perspectives, 126(11). doi:10.1289/EHP2268

Koufariotis, L. T., Chen, Y. P. P., Chamberlain, A., Jagt, C. V., Hayes, B. J. (10 2015). A catalogue of novel bovine long noncoding RNA across 18 tissues. PloS One, 10(10), e0141225. doi:10.1371/journal.pone.0141225

Kung, J. T. Y., Colognori, D., Lee, J. T. (2013). Long noncoding RNAs: Past, present, and future. Genetics, Vol. 193, pp. 651–669. doi:10.1534/genetics.112.146704

Lagarrigue, S., Lorthiois, M., Degalez, F., Gilot, D., Derrien, T. (11 2021). LncRNAs in domesticated animals: from dog to livestock species. Mammalian Genome 2021, 1, 1–23. doi:10.1007/S00335-021-09928-7

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., … Maza, E. (2 2018). Optimization of an RNA-seq differential gene expression analysis depending on biological replicate number and library size. Frontiers in Plant Science, 9, 108. doi:10.3389/FPLS.2018.00108/BIBTEX

Lämmermann, T., Kastenmüller, W. (5 2019). Concepts of GPCR-controlled navigation in the immune system. Immunological Reviews, 289(1), 205. doi:10.1111/IMR.12752

LaRosa, D. F., Orange, J. S. (2 2008). 1. Lymphocytes. The Journal of Allergy and Clinical Immunology, 121(2 Suppl). doi:10.1016/J.JACI.2007.06.016

Lemay, D. G., Lynn, D. J., Martin, W. F., Neville, M. C., Casey, T. M., Rincon, G., … Rijnkels, M. (4 2009). The bovine lactation genome: insights into the evolution of mammalian milk. Genome Biology, 10(4). doi:10.1186/GB-2009-10-4-R43

Li, A., Zhang, J., Zhou, Z. (9 2014). PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics, 15(1), 1–10. doi:10.1186/1471-2105-15-311/FIGURES/3

Li, A., Zhang, J., Zhou, Z., Wang, L., Liu, Y., Liu, Y. (4 2015). ALDB: A Domestic-Animal Long Noncoding RNA Database. PLOS ONE, 10(4), e0124003. doi:10.1371/journal.pone.0124003

Li, H., Huang, K., Wang, P., Feng, T., Shi, D., Cui, K., … Liu, Q. (2 2020). Comparison of Long Non-Coding RNA Expression Profiles of Cattle and Buffalo Differing in Muscle Characteristics. Frontiers in Genetics, 11, 98. doi:10.3389/fgene.2020.00098

Li, L., Stoeckert, C. J., Roos, D. S. (9 2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research, 13(9), 2178. doi:10.1101/GR.1224503

Li, Z., Zhao, W., Wang, M., Zhou, X. (6 2019). The Role of Long Noncoding RNAs in Gene Expression Regulation. In Gene Expression Profiling in Cancer. doi:10.5772/intechopen.81773

Lift Genome Annotations. (n.d.). Retrieved from https://genome.ucsc.edu/cgi-bin/hgLiftOver

Lin, S., Fang, L., Kang, X., Liu, S., Liu, M., Connor, E. E., … Li, C. J. (6 2020). Establishment and transcriptomic analyses of a cattle rumen epithelial primary cells (REPC) culture

by bulk and single-cell RNA sequencing to elucidate interactions of butyrate and rumen development. Heliyon, 6(6), e04112. doi:10.1016/j.heliyon.2020.e04112

Liu, R., Han, M., Liu, X., Yu, K., Bai, X., Dong, Y. (5 2022). Genome-Wide Identification and Characterization of Long Non-Coding RNAs in Longissimus dorsi Skeletal Muscle of Shandong Black Cattle and Luxi Cattle. Frontiers in Genetics, 13, 849399. doi:10.3389/FGENE.2022.849399/BIBTEX

Liu, T., Zhang, L., Joo, D., Sun, S. C. (7 2017). NF-B signaling in inflammation. Signal Transduction and Targeted Therapy 2017 2:1, 2(1), 1–9. doi:10.1038/sigtrans.2017.23

Liver Biopsy Procedure in Cattle — College of Veterinary Medicine at MSU. (n.d.). Retrieved from https://cvm.msu.edu/vdl/laboratory-sections/nutrition/mineral-and-vitamin-testing-sample-collection-and-handling/liver-biopsy-procedure-in-cattle

Liver Function Tests: MedlinePlus Medical Test. (n.d.). Retrieved from https://medlineplus.gov/lab-tests/liver-function-tests/

Lo, N. W., Lau, J. T. Y. (1996). Transcription of the -galactoside 2,6-sialyltransferase gene in B lymphocytes is directed by a separate and distinct promoter. Glycobiology, 6(3), 271–279. doi:10.1093/glycob/6.3.271

Loh, P. R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., … Price, A. L. (12 2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet, 47(12), 1385–1392. doi:10.1038/ng.3431

Lomonte, P., Bublot, M., Van Santen, V., Keil, G., Pastoret, P. P., Thiry, E. (11 1996). Bovine herpesvirus 4: genomic organization and relationship with two other gammaherpesviruses, Epstein-Barr virus and herpesvirus saimiri. Veterinary Microbiology, 53(1–2), 79–89. doi:10.1016/S0378-1135(96)01236-9

Lu, D., Sargolzaei, M., Kelly, M., Li, C., Voort, G. V., Wang, Z., ... Miller, S. P. (2012). Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Front Genet, 3(AUG), 152. doi:10.3389/fgene.2012.00152

Lukowski, S. W., Lloyd-Jones, L. R., Holloway, A., Kirsten, H., Hemani, G., Yang, J., ... Powell, J. E. (12 2017). Genetic correlations reveal the shared genetic architecture of transcription in human peripheral blood. Nature Communications, 8(1). doi:10.1038/s41467-017-00473-z

Lymphocyte. (n.d.). Retrieved from https://www.genome.gov/genetics-glossary/Lymphocyte

Lymphocyte Development and Structure - Immunology - Medbullets Step 1. (n.d.). Retrieved from https://step1.medbullets.com/immunology/105063/lymphocyte-development-and-structure

Ma, L., Cao, J., Liu, L., Du, Q., Li, Z., Zou, D., ... Zhang, Z. (1 2019). Lncbook: A curated knowledgebase of human long non-coding rnas. Nucleic Acids Research, 47(D1), D128–D134. doi:10.1093/nar/gky960

Macias, H., Hinck, L. (7 2012). Mammary Gland Development. Wiley Interdisciplinary Reviews. Developmental Biology, 1(4), 533. doi:10.1002/WDEV.35

macromolecule biosynthetic process Gene Ontology Term (GO:0009059). (n.d.). Retrieved from https://www.informatics.jax.org/vocab/geneontology/GO:0009059

Macrophages — British Society for Immunology. (n.d.). Retrieved from https://www.immunology.org/public-information/bitesized-immunology/cells/macrophages

Mahmoudi, B., Fayazi, J., Roshanfekr, H., Sari, M., Bakhtiarizadeh, M. R. (2 2020). Genome-wide identification and characterization of novel long non-coding RNA in Rumi-

nal tissue affected with sub-acute Ruminal acidosis from Holstein cattle. Veterinary Research Communications 2020 44:1, 44(1), 19–27. doi:10.1007/S11259-020-09769-W

Malmuthuge, N., Liang, G., Guan, L. L. (8 2019). Regulation of rumen development in neonatal ruminants through microbial metagenomes and host transcriptomes. Genome Biology, 20(1), 1–16. doi:10.1186/S13059-019-1786-0/FIGURES/6

Managing Cow Lactation Cycles — The Cattle Site. (n.d.). Retrieved from https://www.thecattlesite.com/articles/4248/managing-cow-lactation-cycles/

Marek's disease — The Poultry Site. (n.d.). Retrieved from https://www.thepoultrysite.com/disease-guide/mareks-disease

May-Sien, M., Id, T., Klepper, A., Lyden, A., Oliveira, A., Id, P., … Lammert, C. S. (2022). Transcriptomic profiling of blood from autoimmune hepatitis patients reveals potential mechanisms with implications for management. doi:10.1371/journal.pone.0264307

Meade, H. M., Echelard, Y., Ziomek, C. A., Young, M. W., Harvey, M., Cole, E. S., … Curling, J. M. (1999). EXPRESSION OF RECOMBINANT PROTEINS IN THE MILK OF TRANSGENIC ANIMALS. Gene Expression Systems, 399. doi:10.1016/B978-012253840-7/50015-8

Meale, S. J., Li, S., Azevedo, P., Derakhshani, H., Plaizier, J. C., Khafipour, E., Steele, M. A. (5 2016). Development of ruminal and fecal microbiomes are affected by weaning but not weaning strategy in dairy calves. Frontiers in Microbiology, 7(MAY), 182975. doi:10.3389/FMICB.2016.00582/BIBTEX

Mengstie, M. A., Wondimu, B. Z. (2021). Mechanism and Applications of CRISPR/Cas-9-Mediated Genome Editing. Biologics: Targets Therapy, 15, 353. doi:10.2147/BTT.S326422

Milk — The Nutrition Source — Harvard T.H. Chan School of Public Health. (n.d.). Retrieved from https://www.hsph.harvard.edu/nutritionsource/milk/

Miller, J. B., Pickett, B. D., Ridge, P. G. (2 2019). JustOrthologs: a fast, accurate and user-friendly ortholog identification algorithm. Bioinformatics, 35(4), 546–552. doi:10.1093/BIOINFORMATICS/BTY669

Mohr, S. E., Hu, Y., Ewen-Campen, B., Housden, B. E., Viswanatha, R., Perrimon, N. (9 2016). CRISPR guide RNA design for research applications. The Febs Journal, 283(17), 3232. doi:10.1111/FEBS.13777

MONOCLdb - Home. (n.d.). Retrieved from https://www.monocldb.org/

multiBigwigSummary — deepTools 3.5.2 documentation. (n.d.). Retrieved from https://deeptools.readthedocs.io/en/develop/content/tools/multiBigwigSummary.html

Name That Cow: The 6 Great Dairy Breeds — Dairy Discovery Zone. (n.d.). Retrieved from https://www.dairydiscoveryzone.com/blog/name-cow-6-great-dairy-breeds

Narayanan, K. B., Park, H. H. (2015). Toll/interleukin-1 receptor (TIR) domain-mediated cellular signaling pathways. Apoptosis: An International Journal on Programmed Cell Death, 20(2), 196–209. doi:10.1007/S10495-014-1073-1

National Center for Biotechnology Information (NCBI). (1988). National Library of Medicine (US), National Center for Biotechnology Information.

Nguyen, L. T., Reverter, A., Cánovas, A., Venus, B., Anderson, S. T., Islas-Trejo, A., … Fortes, M. R. S. (3 2018). STAT6, PBX2, and PBRM1 emerge as predicted regulators of 452 differentially expressed genes associated with puberty in Brahman heifers. Frontiers in Genetics, 9(MAR), 87. doi:10.3389/FGENE.2018.00087/BIBTEX

Nolte, W., Weikard, R., Brunner, R. M., Albrecht, E., Hammon, H. M., Reverter, A., Kühn, C. (11 2019a). Biological Network Approach for the Identification of Regulatory Long

Non-Coding RNAs Associated With Metabolic Efficiency in Cattle. Frontiers in Genetics, 0, 1130. doi:10.3389/FGENE.2019.01130

NPLOC4 - Nuclear protein localization protein 4 homolog - Homo sapiens (Human) — UniProtKB — UniProt. (n.d.). Retrieved from https://www.uniprot.org/uniprotkb/Q8TAT6/entry

Nucleotide BLAST: Search nucleotide databases using a nucleotide query. (n.d.). Retrieved from https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastnPAGETYPE=BlastSearchLINKLOC=bl

On genetically modified foods, widespread skepticism in 20 publics — Pew Research Center. (n.d.). Retrieved from https://www.pewresearch.org/fact-tank/2020/11/11/many-publics-around-world-doubt-safety-of-genetically-modified-foods/

Orr, A. I. (2011). How Cows Eat Grass Exploring Cow Digestion. Retrieved from http://www.fda.gov/AnimalVeterinary

Overview of the Immune System — NIH: National Institute of Allergy and Infectious Diseases. (n.d.). Retrieved from https://www.niaid.nih.gov/research/immune-system-overview

Overview – Salmon: Fast, accurate and bias-aware transcript quantification from RNA-seq data. (n.d.). Retrieved from https://combine-lab.github.io/salmon/

Parkinson, T. J. (1 2019). Specific Infectious Diseases Causing Infertility and Subfertility in Cattle. Veterinary Reproduction Obstetrics, 434–466. doi:10.1016/B978-0-7020-7233-8.00024-0

Parts of the Immune System — Children's Hospital of Philadelphia. (n.d.). Retrieved from https://www.chop.edu/centers-programs/vaccine-education-center/human-immune-system/parts-immune-system

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., Kingsford, C. (3 2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 2017 14:4, 14(4), 417–419. doi:10.1038/nmeth.4197

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., Salzberg, S. L. (2 2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 2015 33:3, 33(3), 290–295. doi:10.1038/nbt.3122

Pertea, M., Pertea, G. (2020). GFF Utilities: GffRead and GffCompare. F1000Research, 9. doi:10.12688/f1000research.23297.1

Peruffo, A., Giacomello, M., Montelli, S., Corain, L., Cozzi, B. (2011). Expression and localization of aromatase P450 AROM , estrogen receptor-a, and estrogen receptor-b in the developing fetal bovine frontal cortex. doi:10.1016/j.ygcen.2011.03.005

Pfam: Search Pfam. (n.d.). Retrieved from http://pfam.xfam.org/searchtabview=tab1

PHAST. (n.d.). Retrieved from http://compgen.cshl.edu/phast/index.php

Ponting, C. P., Hardison, R. C. (11 2011). What fraction of the human genome is functional? Genome Research, Vol. 21, pp. 1769–1776. doi:10.1101/gr.116814.110

Power and Sample Size Determination. (n.d.). Retrieved from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704power/bs704powerprint.html

Pp, A., Mb, C., Dk, G., Me, D., Js, M. (1 2011). lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Research, 39(Database issue). doi:10.1093/NAR/GKQ1138

Public opinion about genetically modified foods and trust in scientists — Pew Research Center. (n.d.). Retrieved from https://www.pewresearch.org/science/2016/12/01/public-opinion-about-genetically-modified-foods-and-trust-in-scientists-connected-with-these-foods/

Quinlan, A. R., Hall, I. M. (3 2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. doi:10.1093/bioinformatics/btq033

R: The R Project for Statistical Computing. (n.d.). Retrieved from https://www.r-project.org/

Rd, F., J, M., J, T., P, C., A, H., Je, P., ... A, B. (11 2010). The Pfam protein families database. Nucleic Acids Research, 38(Database issue). doi:10.1093/NAR/GKP985

Ribosome Profiling — Ribo-Seq/ART-Seq for ribosome-protected mRNA. (n.d.). Retrieved from https://www.illumina.com/techniques/sequencing/rna-sequencing/ribosome-profiling.html

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., ... Chang, H. Y. (6 2007). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs. Cell, 129(7), 1311. doi:10.1016/J.CELL.2007.05.022

Ripley, B. D. (2 2001). The R project in statistical computing. MSOR Connections The Newsletter of the LTSN Maths, Stats OR Network, 1(1), 23–25. doi:10.11120/msor.2001.01010023

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., ... Medrano, J. F. (3 2020). De novo assembly of the cattle reference genome with single-molecule sequencing. GigaScience, 9(3), 1–9. doi:10.1093/GIGASCIENCE/GIAA021

Roxrud, I., Raiborg, C., Pedersen, N. M., Stang, E., Stenmark, H. (3 2008). An endosomally localized isoform of Eps15 interacts with Hrs to mediate degradation of epidermal growth factor receptor. Journal of Cell Biology, 180(6), 1205–1218. doi:10.1083/JCB.200708115

Rumen development of calves - Beef. (n.d.). Retrieved from https://www.canr.msu.edu/news/rumen-development-of-calves

Saipin, N., Noophun, J., Chumyim, P., Rungsiwiwut, R. (6 2018). Goat milk: Non-invasive source for mammary epithelial cell isolation and in vitro culture. Anatomia, Histologia, Embryologia, 47(3), 187–194. doi:10.1111/AHE.12339

samtools Tutorial. (n.d.). Retrieved from http://quinlanlab.org/tutorials/samtools/samtools.html

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., Schaeffer, L. R. (5 2008). Extent of linkage disequilibrium in Holstein cattle in North America. Journal of Dairy Science, 91(5), 2106–2117. doi:10.3168/jds.2007-0553

Schult, P., Roth, H., Adams, R. L., Mas, C., Imbert, L., Orlik, C., ... Lohmann, V. (7 2018). microRNA-122 amplifies hepatitis C virus translation by shaping the structure of the internal ribosomal entry site. Nature Communications 2018 9:1, 9(1), 1–14. doi:10.1038/s41467-018-05053-3

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., ... Barton, G. J. (6 2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA, 22(6), 839–851. doi:10.1261/RNA.053959.115/-/DC1

Sequencing coverage and breadth of coverage. (n.d.). Retrieved from https://www.reneshbedre.com/blog/sequencing-coverage.html

SETD2 - Histone-lysine N-methyltransferase SETD2 - Homo sapiens (Human) — UniProtKB — UniProt. (n.d.). Retrieved from https://www.uniprot.org/uniprotkb/Q9BYW2/entry

Sheldon, I. M. (1 2018). The metritis complex in cattle. Veterinary Reproduction Obstetrics, 408–433. doi:10.1016/B978-0-7020-7233-8.00023-9

Shen, L., Zhang, S., Li, Q., Fu, Y., Tang, G., Jiang, Y., ... Zhu, L. (11 2018). The landscape of non-coding RNA in an adult pig model of intrauterine growth restriction. Cellular Physiology and Biochemistry, 50(5), 1764–1778. doi:10.1159/000494794

SHFL shiftless antiviral inhibitor of ribosomal frameshifting [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene/55337

Shimakami, T., Yamane, D., Jangra, R. K., Kempf, B. J., Spaniel, C., Barton, D. J., Lemon, S. M. (1 2012). Stabilization of hepatitis C virus RNA by an Ago2-miR-122 complex. Proceedings of the National Academy of Sciences of the United States of America, 109(3), 941–946. doi:10.1073/PNAS.1112263109/-/DCSUPPLEMENTAL

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (8 2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, 15(8), 1034–1050. doi:10.1101/GR.3715005

ST6GAL1 ST6 beta-galactoside alpha-2,6-sialyltransferase 1 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene/6480

Statello, L., Guo, C. J., Chen, L. L., Huarte, M. (2 2021). Gene regulation by long non-coding RNAs and its biological functions. Nature Reviews Molecular Cell Biology, Vol. 22, pp. 96–118. doi:10.1038/s41580-020-00315-9

StringTie. (n.d.). Retrieved from http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual

Strucken, E. M., Laurenson, Y. C. S. M., Brockmann, G. A. (2015). Go with the flow-biology and genetics of the lactation cycle. Frontiers in Genetics, 6(MAR). doi:10.3389/FGENE.2015.00118/ABSTRACT

Tamminga, C. A. (2004). Images in Neuroscience Frontal Cortex Function. Am J Psychiatry, 161, 12. Retrieved from http://ajp.psychiatryonline.org

Tana, M. M. S., Klepper, A., Lyden, A., Pisco, A. O., Phelps, M., McGee, B., … Lammert, C. S. (3 2022). Transcriptomic profiling of blood from autoimmune hepatitis patients reveals potential mechanisms with implications for management. PLOS ONE, 17(3), e0264307. doi:10.1371/JOURNAL.PONE.0264307

Thapa, P., Farber, D. L. (5 2019). THE ROLE OF THE THYMUS IN THE IMMUNE RESPONSE. Thoracic Surgery Clinics, 29(2), 123. doi:10.1016/J.THORSURG.2018.12.001

The innate and adaptive immune systems. (7 2020). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK279396/

The Life Cycle of Beef Cattle Production — Peterson Farm Brothers. (n.d.). Retrieved from https://petersonfarmbrothers.com/the-life-cycle-of-beef-cattle-production/

The ruminant digestive system. (n.d.). Retrieved from https://extension.umn.edu/dairy-nutrition/ruminant-digestive-systemstomach-compartments-1000460

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L. P., Mi, H. (1 2022). PANTHER: Making genome-scale phylogenetics accessible to all. Protein Science, 31(1), 8–22. doi:10.1002/PRO.4218

THNSL2 threonine synthase like 2 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/gene/55258

Trapnell, C., Pachter, L., Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105–1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (3 2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), 562. doi:10.1038/NPROT.2012.016

Tucker, H. A., Lansing, E. (n.d.). Physiological Control of Mammary Growth, Lactogenesis, and Lactation 1. Journal of Dairy Science, 64, 1403–1421. doi:10.3168/jds.S0022-0302(81)82711-7

Type II Error Calculation Tutorial. (n.d.). Retrieved from https://www.easycalculation.com/statistics/learn-beta-error.php

UCSC Genome Browser Downloads. (n.d.). Retrieved from https://hgdownload.soe.ucsc.edu/downloads.html

UCSC Genome Browser Gateway. (n.d.). Retrieved from https://genome.ucsc.edu/cgi-bin/hgGateway

Understanding the Immune System - SITC connectED. (n.d.). Retrieved from https://www.sitcancer.org/connectedold/p/patient/resources/melanoma-guide/immune-system

Understanding the Ruminant Animal Digestive System — Mississippi State University Extension Service. (n.d.). Retrieved from http://extension.msstate.edu/publications/understanding-the-ruminant-animal-digestive-system

UniProt. (n.d.). Retrieved from https://www.uniprot.org/

Vallender, E. J. (9 2009). Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. Methods (San Diego, Calif.), 49(1), 50. doi:10.1016/J.YMETH.2009.05.010

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. J Dairy Sci, 91(11), 4414–4423. doi:10.3168/jds.2007-0980

Vi, R. L. B., Connor, E. E. (2017). Rumen Function and Development. doi:10.1016/j.cvfa.2017.06.001

Wang, Jian, Koganti, P. P., Yao, J. (2 2020). Systematic identification of long intergenic non-coding RNAs expressed in bovine oocytes. Reproductive Biology and Endocrinology 2020 18:1, 18(1), 1–9. doi:10.1186/S12958-020-00573-4

Wang, Jian, Sun, H.-Z., O'Hara, E., Chen, H., Guan, L. (10 2021). 20 Profiling circRNA and lncRNA Expression in Key Bovine Metabolic Tissues. Journal of Animal Science, 99(Supplement3), 9–9. doi:10.1093/JAS/SKAB235.016

Wang, Jing, Ren, Q., Hua, L., Chen, J., Zhang, J., Bai, H., ... Bai, X. (3 2019). Comprehensive Analysis of Differentially Expressed mRNA, lncRNA and circRNA and Their ceRNA Networks in the Longissimus Dorsi Muscle of Two Different Pig Breeds. International Journal of Molecular Sciences, 20(5). doi:10.3390/IJMS20051107

Wang, X., Iyer, A., Lyons, A. B., Körner, H., Wei, W. (8 2019). Emerging roles for G-protein coupled receptors in development and activation of macrophages. Frontiers in Immunology, 10(AUG), 457996. doi:10.3389/FIMMU.2019.02031/BIBTEX

Warner, R. G., Flatt, W. P., Loosli, J. K. (8 1956a). Dietary factors influencing the development of the ruminant stomach. Journal of Agricultural and Food Chemistry, 4(9), 788–792. doi:10.1021/JF60067A003/ASSET/JF60067A003.FP.PNGV03

Warner, R. G., Flatt, W. P., Loosli, J. K. (8 1956b). Ruminant nutrition, dietary factors influencing development of ruminant stomach. Journal of Agricultural and Food Chemistry, 4(9), 788–792. doi:10.1021/jf60067a003

What are the limitations of RNA-seq for lncRNA profiling? — Arraystar. (n.d.). Retrieved from https://www.arraystar.com/reviews/what-are-the-limitations-of-rna-seq-for-lncrna-profiling/

What are the organs of the immune system? (7 2020). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK279395/

What is a good sequencing depth for bulk RNA-Seq? (n.d.). Retrieved from https://www.ecseq.com/support/ngs/what-is-a-good-sequencing-depth-for-bulk-rna-seq

Wu, S., Li, R. W., Li, W., Li, C. J. (5 2012). Transcriptome characterization by RNA-seq unravels the mechanisms of butyrate-induced epigenomic regulation in bovine cells. PloS One, 7(5), e36940. doi:10.1371/journal.pone.0036940

Yang, Y., Yu, X., Wang, Z., Sun, H., Yang, Y., Li, K., Tang, Z. (2018). Long non-coding MEG3 is a marker for skeletal muscle development and meat production traits in pigs Animal genome View project The National Key Project for Production of Transgenic Livestock, PR China View project Long non-coding MEG3 is a marker for skeletal muscle development and meat production traits in pigs. Article in Animal Genetics. doi:10.1111/age.12712

Yao, L., Wang, H., Song, Y., Sui, G. (10 2017). BioQueue: A novel pipeline framework to accelerate bioinformatics analysis. Bioinformatics, 33(20), 3286–3288. doi:10.1093/BIOINFORMATICS/BTX403

Zamrodah, Y. (2016). Expression of HLA and Autoimmune Pathway Genes in Liver Biopsies of Young Subjects with Autoimmune Hepatitis Type 1. Journal of Pediatric Gastroenterology and Nutrition, 15(2), 1–23. doi:10.1097/MPG.0000000000003538

Zare, K., Shademan, M., Ghahramani Seno, M. M., Dehghani, H. (n.d.). CRISPR/Cas9 Knockout Strategies to Ablate CCAT1 lncRNA Gene in Cancer Cells. doi:10.1186/s12575-018-0086-5

Zeng, C., Fukunaga, T., Hamada, M. (5 2018). Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data. BMC Genomics, 19(1), 414. doi:10.1186/S12864-018-4765-Z

Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., Shen, F. (11 2019). Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. International Journal of Molecular Sciences, 20(22). doi:10.3390/IJMS20225573

Zhang, Y., Cai, W., Li, Q., Wang, Y., Wang, Z., Zhang, Q., … Zhang, L. (3 2022). Transcriptome Analysis of Bovine Rumen Tissue in Three Developmental Stages. Frontiers in Genetics, 13, 821406. doi:10.3389/FGENE.2022.821406/BIBTEX

Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., … Sun, J. (3 2015). Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Molecular BioSystems, 11(3), 760–769. doi:10.1039/c4mb00511b

Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., … He, S. (1 2021). NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Research, 49(D1), D165–D171. doi:10.1093/NAR/GKAA1046

Zibitt, S. M., Hartford, C. C. R., Lal, A. (2021). Interrogating lncRNA functions via CRISPR/Cas systems. Https://Doi. Org/10. 1080/15476286. 2021. 1899500, 18(12), 2097–2106. doi:10.1080/15476286.2021.1899500

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., … Salzberg, S. L. (4 2009). A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 2009 10:4, 10(4), 1–10. doi:10.1186/GB-2009-10-4-R42