eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Non-coding RNAs in ovine immunity: Identification of unannotated genes and functional analyses of high throughput genomic data

Candidate
Martin Bilbao Arribas

Director
Begoña M. Jugo Orrantia

2022

Abstract

Background The mammal genome is pervasively transcribed, that is, a bigger fraction of the genome is being transcribed at some point of development than what can be attributed to protein-coding RNAs. It is known that several classes of non-protein coding RNA (ncRNA) genes are encoded in the mammal genome. Some are essential well-defined housekeeping RNAs, others small conserved regulatory RNAs that act as post-transcriptional regulators and others large non-coding transcription products with limited evidence of function. MicroRNAs (miRNAs) are the main class of noncoding small RNAs. They are molecules of approximately 22 nucleotides that regulate gene expression post-transcriptionally by binding to mRNAs. In animals, many miRNAs are highly conserved, up to 200 miRNA genes can be traced to the vertebrate ancestor of mammals and bony fish. Nevertheless, the number of confidently identified but poorly conserved genes is growing as more samples are subjected to high-throughput sequencing. Long non-coding RNAs (lncRNAs) are a heterogeneous class of transcripts longer than 200 nucleotides lacking protein-coding potential that are present across a variety of eukaryotic species. The biogenesis of lncRNAs seems to be very similar to mRNAs: they are transcribed by RNA polymerase II, are post-transcriptionally modified like mRNAs and expressed lncRNA promoters are enriched for the same histone modifications. However, they show an exceptional cell type, tissue, developmental stage and disease state specific expression, are expressed at lower abundances than mRNAs, and most of them are poorly conserved at sequence level. Several lncRNAs have been attributed a biological function, usually related to a structural or regulatory role, but most of them remain uncharacterized.

Domestic animals are of great importance as sources of high-quality products for human consumption and as disease models in biomedical research, and animal health is an essential component of the "one health" concept for the prevention of zoonotic diseases. The human and mouse genomes have been deeply annotated for non-coding genes, but the rest of the mammal genomes, including livestock species, are lacking in terms of ncRNA gene quantity and quality. These ncRNAs are important because a big fraction of the thousands of genomic regions that have been associated with complex phenotypic traits and diseases in farmed animals lie within non-coding genomic regions. miRNAs

are much more conserved than lncRNAs and have predictable structures, so there have been more transcriptomic studies in sheep. Still, the number of annotated miRNAs in the reference databases remain small, with just 103 miRNA genes in miRBase. Considering lncRNAs, they are also underrepresented in the sheep reference annotations, mainly because of the difficulty of correctly producing transcript models using in silico analyses. The vast majority of published functional genomics analyses that have profiled lncRNAs in sheep are related to phenotypic traits important for production and development of commercially interesting tissues, but few analyse the immune system.

In non-model organisms, since the non-coding gene sets are still in need for improvement in terms of quantity and quality of annotated genes, genome-wide transcriptomic studies that link ncRNAs to immune functions have been less common. The main objective of this work is the identification of ovine non-coding genes, concretely miRNA and lncRNA genes, that are involved in the innate and adaptive immune responses induced by vaccines, vaccine components and pathogen infections. For this purpose, sequencing datasets produced in the lab and datasets publicly available were analysed with bioinformatic tools and workflows in order to identify unannotated non-coding genes, profile their expression in different tissues and perform evolutionary conservation analyses. Statistical approaches for analysing their expression profiles during different immune responses include differential gene expression analysis and co-expression network analyses.

Methods Transcriptomics is defined as the study of the transcriptome, the complete set of RNA molecules, as it is the study of expressed RNA in a given cell or tissue type. For the characterisation of the ovine non-coding transcriptome, high-throughput RNA sequencing assays were analysed, mainly small RNA-seq (sRNA-seq) and ribosome-depleted RNA-seq. miRNAs were predicted from sRNA-seq data using widely used tools – such as miRDeep2 or sRNAbench – that make use of the known structural features of miRNAs such as length or folding energy of the precursor molecule. Sequencing reads are mapped against the reference genome, bona fide miRNA sequences are selected and reads are quantified. Unannotated miRNAs were named after their closest orthologue by sequence similarity. Target prediction tools were used to know which genes are regulated by miRNAs. Ribosome-depleted RNA-seq and poly-A selected RNA-seq were analysed with a tailor-made bionformatic workflow for the identification of unannotated lncRNAs. Reads are mapped to the reference genome with a splice-aware tool, the transcriptome is assembled, the transcriptome is compared with the reference annotation, novel transcripts are filtered and gene expression levels are quantified. The filtering of unannotated transcripts for selecting lncRNAs consists of length filters, coding potential assessment and protein domain searches.

Expression profiling of lncRNAs and miRNAs along protein coding genes (PCGs) was used to get clues about their involvement in the immune response to infection and vaccination. Differential gene expression analysis was used to identify genes with statistically significant differences in expression levels between animals subjected to two experimental conditions (e.g. uninfected animals and infected animals). The selected tools were DESeq2 and edgeR, both based on negative binomial regression models. Besides, gene co-expression network analysis, an unsupervised clustering method that uses correlation values as a measure of similarity, was performed with the lncRNA and PCG expression values. We determine clusters of genes with the same expression patterns, which following the guilt-by-association principle are expected to be involved in similar biological pathways.

Results and discussion The Small Ruminant Lentiviruses group includes the Visna Maedi Virus (VMV) and Caprine Arthritis Encephalitis (CAEV) viruses, which cause a disease in sheep and goats characterized by pneumonia, mastitis, arthritis and encephalitis. In Chapter 3, we performed the first study reporting miRNA profiling in sheep in response to VMV infection during different clinical stages of infection. A total of 212 miRNAs were identified, of which 46 were conserved sequences in other species but found for the first time in sheep, and 12 were completely novel. Differential expression analysis comparing the uninfected and seropositive groups showed changes in several miRNAs; however, no significant differences were detected between seropositive asymptomatic and diseased sheep. Thus, the infection could be detected before the appearances of symptoms by changes in miRNA expression. Oar-miR-21, oar-miR-148a and oar-let-7f seem to have potential implications for the host-virus interactions because of their strong upregulation by both treatments. The robust increase in the expression level of oar-miR-21 is consistent with its increased expression in other viral diseases and during lung fibrosis, a common symptom of VMV disease. Furthermore, the target prediction of the dysregulated miRNAs revealed that they control genes involved in proliferation-related signalling pathways, such as the PI3K-Akt, AMPK and ErbB pathways, also common during fibrosis. The known functions of oar-miR-21 as a regulator of inflammation and proliferation appear to be a possible cause of the lesions caused in the sheep's lungs. This miRNA could be an indicator for the severity of the lung lesions, or a putative target for therapeutic intervention.

In sheep, few miRNAs have been described in comparison with other livestock species or model organisms. In Chapter 4, we uniformly analysed 172 public ovine small RNA sequencing datasets from 21 different tissues in order to predict conserved and novel miRNA precursors and profile their expression patterns. In addition to the 106 annotated sheep miRNAs, 1047 previously unannotated miRNA precursor sequences were detected and 41% of them were assigned an orthologue from other close species. Considering expression levels, a set of miRNAs with high sequence conservation were detected in all tissues, while 733 mature miRNAs were robustly expressed in at least one tissue. 270 miRNAs showed high tissue specificity index values. Brain, male reproductive tissues and PBMCs showed the most distinct expression patterns. Strikingly, over one hundred precursors from the ruminant specific family of mir-2284/mir-2285 miRNAs were found, which were enriched in immune related tissues. This work supports the known high conservation of many miRNAs, but also highlights the potential of clade-specific innovations in ruminant evolution.

Aluminium hydroxide adjuvants are crucial for livestock and human vaccines. Few studies have analysed their effect on the central nervous system in vivo. In Chapter 5, we assessed lncRNA expression in a long-term vaccination experiment for the study of vaccine adjuvant safety. Lambs received three different treatments of parallel subcutaneous inoculations during 16 months with aluminium-containing commercial vaccines, an equivalent dose of aluminium hydroxide or mock injections. Brain samples were sequenced by RNA-seq for the expression analysis of mRNAs and long non-coding RNAs and three expression comparisons were made. Commercial vaccines with aluminium adjuvant did not show almost any effect on lncRNA expression levels in brain tissue, while the inoculation of the adjuvant alone produced the dysregulation of 30 lncRNAs. Although few differentially expressed genes were identified, some dysregulated genes were linked to neurological functions, the lncRNA TUNA among them, or were enriched in mitochondrial energy metabolism related functions. In brief, in this study aluminium hydroxide alone altered the transcriptome of the encephalon to a higher degree than commercial vaccines that present a milder effect. The expression changes in the animals inoculated with aluminium hydroxide suggest mitochondrial disfunction. Further research is needed to elucidate to which extent these changes could have pathological consequences.

In the context of the same experiment, in Chapter 6 we reanalysed sequencing data from PBMCs in order to find dysregulated lncRNAs related to the innate immune response to aluminium adjuvants. We built a transcriptome from sheep PBMCs RNA-seq data in order to identify unannotated lncRNAs and analysed their expression patterns along protein coding genes. We found 2284 novel lncRNAs and assessed their conservation in terms of sequence and synteny. In this case, we found out that inoculation with commercial vaccines or aluminium hydroxide alone caused changes in expression of 159 and 170 lncRNAs. The co-expression analysis revealed lncRNAs related to the immune response to vaccines and adjuvants. A group of co-expressed genes enriched in cytokine signalling and production highlighted the differences between different treatments. A number of differentially expressed lncRNAs were correlated with a divergently located protein-coding gene, such as the OSM cytokine. Other lncRNAs were predicted to act as sponges of miRNAs involved in immune response regulation. This work puts an accent on their involvement in the immune response to repetitive vaccination and the understanding of the mechanism of action of aluminium adjuvants.

LncRNAs are involved in several biological processes, including the immune system response to pathogens and vaccines. In Chapter 7, we take advantage of the increasing number of high-throughput functional experiments deposited in public databases in order to uniformly analyse and profile unannotated lncRNAs from 422 available ovine

RNA-seq samples of blood cells, lymphoid organs and other immune cells. We identify the lncRNA gene expression signature of a broad immune response, that is, lncRNAs that are dysregulated upon immune activation by a variety of pathogens and vaccines. We identified 12302 unannotated lncRNA genes with support from independent assays and 873 expressed annotated lncRNAs. Unannotated lncRNAs showed low expression levels and sequence conservation, with differences depending on lncRNA classification. Differential expression analyses between unstimulated samples and samples with different stimulations such as pathogen infection or vaccination resulted in hundreds of lncRNAs with changed expression. In blood cell samples there were 75 differentially expressed lncRNAs and in lymph node samples there were 46. Gene co-expression analyses revealed immune gene-enriched clusters associated with immune system activation and related to interferon signalling, antiviral response or endoplasmic reticulum stress. Besides, differential co-expression networks (DCNs) were constructed in order to find condition-specific relationships between coding genes and lncRNAs. The DCNs also revealed PCGs that, despite not being differentially expressed, could be implicated in the immune response to infection and vaccination. Examples of this would be the metabolic anzyme IDO1 in the innate response and the transcription factor CREB3 in the adaptive response.

Conclusions Multiple processes are involved in the immune response to infection and vaccination and ncRNAs play different roles in these processes. The main goals of this work were, first, to detect unannotated ovine miRNAs and lncRNAs from functional genomics experiments produced in the research group and publicly available RNA sequencing datasets from immune tissues and, second, to profile the ncRNA gene expression across a variety of immune stimulations such as pathogen infection or vaccination. The ovine miRNAs described in this dissertation show diverging levels of sequence conservation, from a set of deeply conserved miRNA families to species-specific families, passing through clade-specific miRNA families that might be important for ruminant evolution. The ovine lncRNAs show the characteristics of other published livestock lncRNAs and the known features of human lncRNAs: poor sequence conservation, low expression levels, few exon number and primarily intergenic location. The functional analyses performed with immune-stimulated samples revealed hundreds of known and novel ncRNAs with specific expression patterns during an infection or vaccination. These genes make up a prioritized set of potential candidates for deeper experimental analyses. Taken together, these results should help completing the sheep non-coding gene catalogue, and most importantly, they give evidence of immune state-specific ncRNA expression patterns in a livestock species.

Laburpena

Sarrera Ugaztunen genoma erabat transkribatzen da, hau da, genomaren zati handiago bat ari da transkribatzen garapen-uneren batean gene proteina kodetzaileei egotzi ahal zaiena baino. Jakina da zenbait proteina-kodetzaile ez diren RNA (ncRNA) gene mota ugaztunen genoman kodetuta daudela. Batzuk oinarrizko funtzioak betetzen dituzten eta ongi definituta dauden funtsezko RNAk dira, beste batzuk kontserbatutako RNA erregulatzaile txikiak dira, transkripzio osteko erregulatzaile gisa jarduten dutenak, eta beste batzuk RNA ez-kodetzaile produktu luzeak dira, funtzio-ebidentzia mugatuarekin. MikroRNAk (miRNAk) RNA ez-kodetzaile txikien artean nagusienak dira. 22 nukleotido inguruko molekulak dira, eta genearen transkripzio-osteko adierazpena erregulatzen dute, mRNA molekulekin lotuz. Animalietan, miRNA asko oso kontserbatuta daude, izan ere, 200 miRNA gene aurki daitezke ugaztunen eta hezur-arrainen arbaso ornoduneraino. Hala eta guztiz ere, konfiantzaz identifikatutako baina gaizki kontserbatutako miRNA geneen kopurua gero eta handiagoa da lagin gehiago ekoizpen-handiko metodoekin sekuentziatzen diren heinean. RNA ez-kodetzaile luzeak (lncRNAk) 200 nukleotido baino gehiagoko transkripto mota heterogeneoa da, proteinak kodetzeko ahalmenik ez dutenak, eta espezie eukariotoetan zehar agertzen direnak. LncRNAen biogenesia mRNAren oso antzekoa da: II RNA polimerasak transkribatzen ditu, transkripzio-osteko aldaketak dituzte, eta adierazten diren lncRNAen promotoreak aberastuta daude histona-eraldaketa berberetan. Hala ere, zelula-mota, ehun, garapen-etapa eta gaixotasun-egoeraren arabera adierazpen espezifiko aparta erakusten dute, mRNAk baino ugaritasun txikiagoarekin adierazten dira, eta gehienak sekuentzia mailan ez daude oso kontserbatuta. LncRNA batzuei funtzio biologiko bat egotzi zaie, normalean funtzio estruktural edo erregulatzaile batekin erlazionatua, baina gehienak karakterizatu gabe jarraitzen dute.

Etxeko animaliek garrantzi handia dute giza kontsumorako kalitate handiko produktuen iturri gisa eta ikerketa biomedikoan gaixotasun-eredu gisa. Gainera, animalien osasuna "one health" kontzeptuaren atal ezinbestekoa da gaixotasun zoonotikoen prebentziorako. Gizakiaren eta saguaren genomak sakonki anotatu egin dira gene ez-kodetzaileei dagokienez, baina gainerako ugaztunen genomak, etxe-abereenak barne, ncRNA geneen kantitate eta kalitate aldetik gabezi handiak dituzte. Gene ez-kodetzaileak garrantzitsuak dira, ohikoa baita ezaugarri fenotipiko konplexuekin eta etxe-abereen gaixotasunekin lotzen diren milaka eskualde genomikoen zati handi bat eskualde ez-kodetzaileetan kokatzea. miRNAk lncRNak baino kontserbatuagoak daude eta aurresan daitezkeen egiturak dituzte, hortaz ardia bezalako animalia ez-eredu batean gehiago ikertu dira metodo transkriptomikoak erabiliz. Hala ere, erreferentziazko datu-baseetan jasotako miRNA kopurua txikia da, soilik 103 miRNA gene baitaude miRBase-n. LncRNAk kontuan hartuta, ardien erreferentzia-anotazioetan ere urriak dira, batez ere in silico analisiak erabiliz transkripto-eredu zuzenak eraikitzeko dagoen zailtasunagatik. Ardian argitaratu diren genomika funtzionaleko analisi gehienek ekoizpenerako eta garapenerako ezaugarri fenotipiko komertzialak aztertu dituzte, baina gutxik aztertu dituzte lncRNAk immunitate-sistemaren baitan.

Animalia-eredu ez diren organismoetan, gene ez-kodetzaileen bilduma oraindik hobetzeke dagoenez anotatuta dauden geneen kantitateari eta kalitateari dagokienez, ez dira hain ohikoak izan ncRNAk funtzio immunologikoekin lotzen dituzten azterketa transkriptomikoak. Lan honen helburu nagusia gene ez-kodetzaileak identifikatzea da, zehazki miRNA eta lncRNA geneak, txertoek, txertoen osagaiek eta patogenoen infekzioek eragindako erantzun immunean parte hartzen dutenak. Horretarako, laborategian sortutako sekuentziazio datuak eta publikoki eskuragarri zeuden datu-multzoak lan-fluxu bioinformatikoekin aztertu ziren, anotatu gabeko gene ez-kodetzaileak identifikatzeko, ehun desberdinetan euren adierazpena aztertzeko eta kontserbazio-analisi ebolutiboak egiteko. Adierazpen profilak erantzun immunologiko desberdinetan aztertzeko erabili diren metodo estatistikoak geneen adierazpen diferentzialaren analisia eta ko-adierazpen sareen analisia izan dira, besteak beste.

Metodoak Transkriptomika transkriptoma ikertzen duen arloa da, zelula edo ehun mota jakin batean adierazitako RNA multzoa. Ardiaren transkriptoma ez-kodetzailearen karakterizaziorako, etekin altuko RNA sekuentziazio datuak aztertu ziren, batez ere small RNA-seq (sRNA-seq) deritzona eta erribosoma-generik gabeko RNA-seq. miRNAk sRNA-seq datuetatik abiatuta aztertu ziren, miRNA molekulen egiturazko ezaugarri ezagunak erabiltzen dituzten tresna ezagunak (miRDeep2 eta sRNAbench) erabiliz, hala nola sekuentziaren luzera edo prekurtsoreak tolesteko behar duen energia. Sekuentziazio-irakurketak erreferentziazko genomaren kontra mapatzen dira, miRNA sekuentzia fidagarriak hautatzen dira eta irakurketak kuantifikatzen dira. Anotatu gabeko miRNAei ortologo hurbilenaren izena eman zitzaien, lerrokaketak erabiliz. Jakiteko zeintzuk diren miRNAek erregulatzen dituzten geneak gene-ituak aurresateko tresnak erabili ziren. Anotatu gabeko lncRNAk identifikatzeko erribosomarik gabeko RNA-seq eta poly-A RNA-seq datuak aztertu ziren, neurrira egindako lan-fluxu bionformatiko batekin. Irakurketak erreferentziazko genomaren aurka mapatzen dira, moztitsasketa kontuan hartzen duen tresna batekin, transkriptoma eraikitzen da, transkriptoma erreferentziazko anotazioarekin alderatzen da, transkripto berriak iragazten dira eta geneen adierazpen-maila kuantifikatzen da. Anotatu gabeko transkriptoen iragaztea lncRNAk hautatzeko egiten da eta, besteak beste, luzera-iragazkiak, proteina kodetzeko ahalmenaren ebaluazioa eta proteinen domeinu-bilaketak egiten dira.

lncRNA eta miRNA geneen adierazpena gene proteina kodetzaileekin (PCG) batera aztertu zen, infekzioaren eta txertaketaren erantzun immunologikoan parte-hartzen duten ikusteko. Bi baldintza esperimentalen (adibidez, infektatu gabeko animaliak eta infektatutako animaliak) arteko desberdintasun estatistikoki esangarriak testatzeko geneen adierazpen diferentzialaren analisia erabili zen. Aukeratutako tresnak DESeq2 eta edgeR izan ziren, biak erregresio eredu binomial negatiboak oinarri dituztenak. Gainera, geneen ko-adierazpen sarearen analisia, korrelazio balioak antzekotasun-neurri gisa erabiltzen dituen gainbegiratzerik gabeko taldekatze-metodoa, lncRNA eta PCG adierazpen-balioekin egin zen. Honekin, adierazpen-eredu berbera duten gene-multzoak zehazten ditugu eta guilt-by-association printzipioari jarraituz antzeko bidezidor biologikoetan parte hartzea espero da.

Emaitzak eta eztabaida Visna Maedi birusa (VMV) eta Caprine Arthritis Encephalitis birusa (CAEV) hausnarkari txikien lentivirusen taldearen parte dira. Horiek ardietan eta ahuntzetan gaixotasun bat eragiten dute, pneumonia, mastitisa, artritisa eta entzefalitisa bezalako sintomekin. 3. Kapituluan, lehen aldiz berri ematen da miRNAren profilari buruz ardietan VMVren infekzio batean, infekzioaren fase kliniko ezberdinetan. Guztira 212 miRNA identifikatu ziren, eta horietatik 46 sekuentzia aldetik kontserbatuak zeuden beste espezie batzuetan, baina lehen aldiz ardietan aurkitu ziren beste 12 sekuentzia erabat berri. Kutsatu gabeko animaliak eta seropositiboak alderatzen dituen adierazpen diferentzialaren analisiak aldaketak aurkitu zituen hainbat miRNAtan; hala ere, ez zen aldaketa esanguratsurik antzeman sintomarik gabeko ardi seropositiboen eta gaixotutako ardien artean. Horrela, infekzioa sintomen agerpenaren aurretik detektatu ahal izango litzateke, miRNAen adierazpenaren aldaketen ondorioz. Oar-miR-21, oar-miR-148a eta oar-let-7f ostalariaren eta birusaren arteko elkarrekintzetan garrantzitsuak izan litezke, beren adierazpenaren igoera oso nabarmena baita bi tratamenduetan. Oar-miR-21 miRNAren adierazpen-mailaren hazkunde indartsua bat dator beste gaixotasun biriko batzuekin eta biriketako fibrosian duen adierazpen aldaketekin, VMV gaixotasunaren sintoma arrunta. Gainera, diferentzialki adierazitako miRNAen geneituen predikzioek agerian utzi zuten zelulen ugaritzearekin lotutako bidezidorretan, hala nola PI3K-Akt, AMPK eta ErbB bideetan, geneak kontrolatzen dituztela, fibrosian ere ohikoa dena. Jakina denez Oar-miR-21 miRNAk funtzio ezagunak ditu hanturaren eta zelulen ugaritzearen erregulazioan eta ardien biriketan eragindako lesioen kausa izan liteke. miRNA hau biriketako lesioen larritasunaren adierazlea izan daiteke, edo interbentzio terapeutikorako itu bat.

Ardietan miRNA gutxi deskribatu dira beste etxe-abere edo animalia-eredu batzuekin alderatuta. 4. Kapituluan, era uniformean aztertu ditugu ardiaren 172 sRNA-seq datu-multzo publiko, 21 ehun desberdin bilduz, kontserbatutako eta anotatu gabeko miRNAk aurresateko eta euren adierazpen mailak aztertzeko. Anotatuta dauden 106 miRNAez gain, aurretik anotatu gabeko 1047 miRNA gene aurkitu ziren. Horietatik, %41a hurbileko beste espezie baten ortologoa zela zehaztu zen. Adierazpen-maila kontuan hartuta, kontserbazio-maila handiko miRNA multzo bat ehun guztietan hauteman zen, 733 miRNAek gutxienez ehun batean adierazpen sendoa zuten bitartean. 270 miRNAek ehun-espezifikotasun indize balio handiak aurkeztu zituzten. Garunak, arren ugaltze-ehunek eta odol zelulek adierazpen-profil desberdinenak erakutsi zituzten. Deigarria da mir-2284/mir-2285 miRNA familiako, zeina hausnarkarien espezifikoa baita, 100 miRNA baino gehiago aurkitu zirela, eta hauek sistema immunearekin lotutako ehunetan gehiago adierazten ziren. Lan honek miRNA askoren kontserbazio-maila handia babesten du, baina hausnarkarien eboluzioan familia espezifikoek dakartzaten berrikuntza ebolutiboak ere nabarmentzen ditu.

Aluminio hidroxidozko txerto-laguntzaileak ezinbestekoak dira etxe-abere eta giza txertoetan. Ikerketa gutxik aztertu dute nerbio-sistema zentralean in vivo duten eragina. 5. Kapituluan, txertoaren segurtasuna aztertzeko epe luzeko txertaketa-esperimentu batean lncRNAk nola adierazten diren aztertu dugu. Arkumeek hiru tratamendu desberdin jaso zituzten larruazalpeko inokulazio paraleloetan 16 hilabetez: aluminioa duten txerto komertzialak, aluminio hidroxidoa soilik edo kontrol-inokulazioak. Garuneko laginak RNA-seq bidez sekuentziatu ziren PCGen eta lncRNAen adierazpenaren analisirako, eta hiru konparaketa egin ziren. Aluminiozko laguntzailea zuten txerto komertzialek ez zuten ia inolako eraginik izan lncRNA genen adierazpen-mailan garun-ehunean; aldiz, aluminio hidroxidoaren inokulazioak 30 lncRNAren adierazpen diferentziala baino ez zuen eragin. Nahiz eta diferentzialki adierazitako gene gutxi identifikatu, zenbait gene funtzio neurologikoekin lotuta zeuden, horien artean TUNA lncRNA, eta gene kodetzaileak metabolismo energetiko mitokondrialarekin lotutako funtzioekin erlazionatuta zeuden. Laburtuz, azterketa honetan aluminio-hidroxidoak entzefaloaren transkriptoman eragin arinagoa izan zuen txerto komertzialek baino, zeinek eragin oso txikia izan zuten. Aluminio hidroxidoz inokulatutako animalien adierazpen aldaketek disfuntzio mitokondriala iradokitzen dute. Ikerketa berriak beharrezkoak dira aldaketa horiek ondorio patologikoak zenbateraino izan ditzaketen argitzeko.

Esperimentu beraren testuinguruan, 6. Kapituluan, odol-zelula mononuklearren (PBMC) sekuentziazio-datuak aztertu ziren, aluminiozko txerto-laguntzaileen aurreko erantzun immunearekin lotutako lncRNA berriak aurkitzeko. Ardien PBMC RNA-seq datuetatik transkriptoma bat eraiki zen, anotatu gabeko lncRNAak identifikatzeko eta haien adierazpen-profila gene proteina-kodetzileekin batera aztertzeko. 2284 lncRNAs berri aurkitu ziren, eta horien kontserbazioa sekuentziaren eta sinteniaren (gene ordena) arabera ebaluatu zen. Kasu honetan, txerto komertzialekin edo soilik aluminio hidroxidoarekin egindako inokulazioek aldaketak eragin zituzten 159 eta 170 lncRNAren adierazpenean, hurrenez hurren. Ko-adierazpen analisiaren bidez txertoen eta txerto-laguntzaileen erantzun immunearekin lotutako lncRNAk identifikatu ziren. Zitokinen seinalizazioan eta produkzioan aberastutako gene-talde batek tratamenduen arteko desberdintasunak nabarmendu zituen. Diferentzialki adierazitako lncRNA batzuek korrelazioak erakutsi zituzten modu dibergentean kokatutako beste PCG batzuekin, hala nola OSM zitokinarekin. Predikzio bioinformatikoek proposatzen dute beste lncRNA batzuek belaki gisa jarduteko gai liratekeela, erantzun immunearekin lotutako miRNA garrantzitsuak doituz. Lan honek lncRNAk txertaketa errepikakorraren immunitate-erantzunean eta aluminiozko laguntzaileen ekintza-mekanismoaren ulermenean duten garrantzia azpimarratzen du.

LncRNAk zenbait prozesu biologikoekin erlazioa dute, sistema immuneak patogenoei eta txertoei ematen dien erantzunarekin barne. 7. Kapituluan, datu-base publikoetan pilatutako ekoizpen handiko RNA sekuentziazioan oinarritzen diren esperimentu funtzional kopuru handia probestu zen. Anotatu gabeko lncRNAk modu uniforme batean aztertu ziren odol-zeluletako, organo linfoideetako eta beste zelula immune batzuetako 422 lagin erabilita. Erantzun immune zabal baten lncRNA gene-sinadura identifikatu zen, hau da, patogeno eta txerto ugariren aurkako erantzun immunean adierazpenean aldaketak erakusten dituzten lncRNAak. 12302 lncRNA gene identifikatu ziren eta erreferentziazko anotazioan dauden 873 lncRNA detektatu ziren. Anotatu gabeko lncRNAek sekuentzien adierazpen- eta kontserbazio-maila apalak erakutsi zituzten, desberdintasunak zeudelarik lncRNA bakoitzaren sailkapenaren arabera. Estimulatu gabeko laginen eta estimulu desberdineko laginen arteko adierazpen diferentzialaren analisian, hala nola patogenoen infekzioa edo txertaketa, adierazpen diferentziala erakusten zuten ehunka lncRNA agertu ziren. Odol-zelulen laginetan 75 lncRNA zeuden diferentzialki adierazita, eta gongoil linfatikoen laginetan 46 zeuden diferentzialki adierazita. Gene ko-adierazpenaren analisiek immunitate sistemarekin erlazionatutako gene-multzoak erakutsi zituzten, sistema immunearen aktibatzearekin lotuta zeudenak eta interferonaren seinaleztatzearekin, birusen aurkako erantzunarekin edo erretikulu endoplasmatikoaren estresarekin zerikusia zutenak. Gainera, ko-adierazpen sare diferentzialak eraiki ziren gene kodetzaileen eta lncRNAen arteko korrelazio aldaketak aurkitzeko. Sare hauek gene kodetzailei buruzko informazio interesgarria ere eman zuten, eta, nahiz eta diferentzialki adierazi ez, gene batzuk infekzioaren eta txertaketaren aurkako erantzun immunearekin erlazionatuta egon litezke. Horren adibide dira IDO1 entzima metabolikoa sortzetiko erantzunean eta CREB3 transkripzio-faktorea erantzun adaptatiboan.

Ondorioak Hainbat prozesuk parte hartzen dute infekzioaren eta txertaketaren aurkako erantzun immunitarioan, eta gene ez-kodetzaileek prozesu horietan hainbat zeregin dituzte. Lan honen helburu nagusiak izan dira, lehenik eta behin, anotatu gabeko miRNAak eta lncRNAak detektatzea, gure ikerketa-taldean egindako genomika funtzionaleko esperimentuak baliatuz eta datu-baseetan dauden ardien RNA-seq esperimentuak baliatuz. Bigarrenik, ncRNA geneen adierazpena aztertu zen sistema immunearen hainbat estimulazioen aurrean, txertaketaren edo patogenoen infekzioaren aurrean esaterako. Lan honetan deskribatu diren ardi miRNAek sekuentzien kontserbazio-maila desberdinak dituzte: batetik, eboluzioan zehar mantendu den miRNA talde bat dago, eta, bestetik, espezie-espezifikoak diren miRNAk, tarteko kontserbazioa erakusten duten miRNA sekuentziak ere daudelarik, hausnarkarien eboluziorako garrantzitsuak izan daitezkeenak. Detektatu diren lncRNA transkriptoek beste etxe-abere batzuetan eta gizakian argitaratutako lanen ezaugarriak dituzte: sekuentziaren kontserbazio txarra, adierazpen-maila apalak, exon kopuru txikia eta kokapen intergenikoaren nagusitasuna. Immunologikoki estimulatutako laginekin egindako analisi funtzionalek agerian utzi zuten ehunka ncRNA ezagun eta berri, infekzio edo txertaketa batean adierazpen-eredu espezifikoak dituztela. Gene horiek analisi esperimental sakonagoak egiteko lehentasunezko hautagai multzo bat osatzen dute. Emaitza hauek, oro har, ardiaren gene ez-kodetzaileen katalogoa osatzen lagundu beharko lukete, eta, garrantzitsuena dena, etxe-abere espezie batean gene ez-kodetzaileen adierazpenean aldaketak daudela erakusten dute sistema immunearen aktibazioan.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AD - Alzheimer's disease

Al - Aluminium

Al(OH$_3$)$_3$ - Aluminium hydroxide

ALS - Amyotrophic lateral sclerosis

BP - Biological process

BTV - Bluetongue virus

CAEV - Caprine Arthritis-Encephalitis

CAGE-seq - Cap Analysis of Gene Expression

CaptureSeq - Targeted RNA sequencing

CC - Cellular component

cDNA - Complementary DNA

ceRNA - Competing endogenous RNA

ChIP-seq - Chromatin Immunoprecipitation Sequencing

circRNA - Circular RNA

CLR - Context likelihood of relatedness transformation

CPAT - Coding Potential Assesment Tool

CPC2 - Coding Potential Calculator 2

CPM - Counts per million

DCN - Differential co-expression network

DE - Differentially expressed

DEG - Differentially expressed gene

DGE - Differential gene expression

EBV - Epstein-Barr Virus

ELISA - Enzyme-linked immunosorbent assay

ENA - European Nucleotide Archive

ER - Endoplasmic reticulum

eRNA - Enhancer RNA

EST - Expressed sequence tag

FAANG - Functional Annotation of Animal Genomes

FC - Fold change

FDR - False discovery rate

FPR - False positive rate

GEO - Gene Expression Omnibus

GO - Gene Ontology

GRO-seq - Global run-on sequencing

GS - Gene significancy

GWAS - Genome-wide association study

HCV - Hepatitis C virus

HGCN - HUGO Gene Nomenclature Comitee

ICH - Intracerebral haemorrhage

iFMDV - Inactivated foot-and-mouth disease virus

ISG - Interferon stimulated gene

lincRNA - Long intergenic non-coding RNA

lncRNA - Long non-coding RNA

log2FC - log2 transformed fold change

MF - Molecular function

miRNA - MicroRNA

miRNA* - Star miRNA

ML - Machine learning

MM - Module membership

mRNA - Messenger RNA

MS - Multiple sclerosis

NAT - Natural antisense transcript

ncRNA - Non-coding RNA

ONT - Oxford Nanopore Technology sequencing

ORF - Open reading frame

PABP - Poly(A)-binding protein

PBMCs - Peripheral blood mononuclear cells

PCA - Principal component analysis

PCG - Protein coding gene

PD - Parkinson disease

piRNA - Piwi-interacting RNA

PPRV - Peste des petits ruminants virus

pre-miRNA - Precursor miRNA transcript

pri-miRNA - Primary miRNA transcript

PROMPT - Promoter upstream transcript

PRRSV - Porcine reproductive and respiratory syndrome virus

RAMPAGE-seq - RNA Annotation and Mapping of Promoters for Analysis of Gene Expression

RefSeq - Reference Sequence

RIN - RNA integrity number

RISC - RNA-induced silencing complex

RNAi - RNA interference

RNA-seq - RNA sequencing

ROS - Reactive oxygen species

rRNA - Ribosomal RNA

SGS - Second generation sequencing

shRNA - Short hairpin RNA

siRNA - Small-interfering RNA

sisRNA - Stable intronic sequence RNA

SPPV - Sheep pox virus

SMRT - Single-molecule real-time sequencing

snoRNA - Small nucleolar RNA

snRNA - Small nuclear RNA

SRA - Sequence Read Archive

SRLV - Small Ruminant Lentiviruses

sRNA-seq - Small RNA-seq

TAD - Topologically associated domain

TF - Transcription factor

TGS - Third generation sequencing

TLR - Toll-like receptor

TPM - Transcripts per million

tRNA - Transfer RNA

TSI - Tissue specificity index

t-SNE - t-distributed stochastic neighbor embedding

TSS - Transcription start site

TTS - Transcription termination site

VM - Visna-Maedi

VMV - Visna-Maedi virus

VST - Variance stabilizing transformation

ΔG - Free energy

Chapter 1

Introduction

1.1

The non-coding genome in mammals

Since the historic achievement that meant the sequencing of the entire human genome for the first time, several other mammals have had their genome sequenced in the last two decades, including many livestock species. After the culmination of the human genome, it became apparent that there were much less genes than previously predicted [1], at least in terms of protein-coding genes [2]. This small number of genes covered a tiny fraction of the genome, while the rest of the genome was termed as "junk DNA" because it was supposedly composed of repetitive regions, transposons, pseudogenes and other unknown functionless sequences.

Eukaryotic genomes are characterised by their large size but low protein-coding content, with just 1.1% of the human genome being protein-coding. Thus, this raised the question about the usefulness of the rest of the genome. Non-coding parts of the genome already sustained interest before the sequencing of the first genomes and early genomicists hypothesised functions for these regions that might not be far from the current dogmas: chromosomal pairing, genome integrity, gene regulation, messenger RNA (mRNA) processing or serving as a reservoir for evolutionary innovation [3]. Part of that obscure genome has been shown to include functionally important DNA elements that control gene expression like promoters, enhancers, insulators and silencers, and most recent exhaustive efforts estimate that they cover around 8% of the human genome [4].

With the advance in whole-genome technologies, it was shown that the mammal genome was pervasively transcribed, that is, a bigger fraction of the genome was being transcribed at some point of development than what could be attributed to protein-coding RNAs [5]. It was shown that several classes of constitutively expressed non-protein coding RNAs (ncRNAs) were needed in the transcription and translation processes [6]. These well-defined housekeeping RNAs are essential for normal function of the cell: Small nuclear RNAs (snRNAs), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and small nucleolar RNAs (snoRNAs). Beyond these well-known ncRNAs, a small number of conserved regulatory RNAs that act as post-transcriptional regulators had been characterized since the 90s. To date, the following small ncRNAs types account for hundreds of members in human: microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), small interfering RNAs (siRNAs) and others. Finally, the evidence of thousands of large non-coding transcription products located in intergenic regions or associated with known genes paved the way for a new, heterogeneous and very numerous class of RNAs named under the umbrella term of long non-coding RNAs (lncRNAs).

Nowadays, the annotations with highest quality and completeness agree on a gene count of about 20,000 coding genes and more than 18,000 non-coding genes in human, even though those numbers are still being debated [7]. The number of annotated coding genes in other mammals follow a similar trend, but their non-coding genome is lacking in terms of gene quantity and quality. For instance, the latest Ensembl human annotation (release 105) has five times more annotated miRNA genes and ten times more annotated lncRNA genes than the sheep annotation (Figure 1). The present dissertation focuses on these two non-coding RNA classes: miRNAs and lncRNAs.

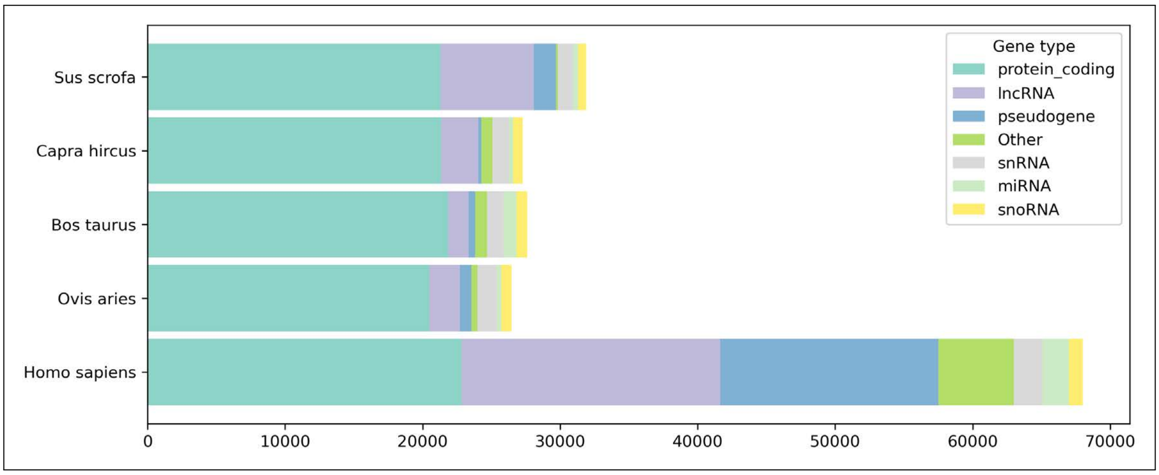


Figure 1: Types of annotated genes in human and livestock species. The figure illustrates the absolute number of genes from each relevant type according to the Ensembl release 105 annotations.

1.2

MicroRNAs (miRNAs)

MicroRNAs (miRNAs) are a class of non-coding endogenous small RNAs that play important regulatory roles in plants and animals by post-transcriptional targeting of mRNA molecules and translation repression [8]. They arise from stem-loop regions of longer precursor RNA transcripts and the mature products are around 22 nucleotides long. The first miRNA was discovered in the 90s, after it was shown that a short RNA transcript was able to block the mRNA of another gene, and was named lin-4 [9]. Later, let-7 miRNA was identified also in C. elegans [10]. Because of their relatively recent discovery knowledge on miRNAs is still in progress, but in the last decade more and more miRNAs have been characterized, with up to 2000 described miRNAs in humans. miRNAs are one of the most abundant regulators of gene expression in multicellular organisms and are said to play part in the expression of a big fraction of protein-coding genes (PCGs) [11].

The miRNA pathway in animals is derived from the more basic RNA interference (RNAi) pathway, and has independently arisen more than once in evolution in animals, plants or algae [12]. In animals, many miRNAs are highly conserved, up to 200 miRNA genes can be traced to the vertebrate ancestor of mammals and bony fish and some of them even predate the emergence of bilaterian animals. Nevertheless, the number of confidently identified but poorly conserved genes is growing as more samples are subjected to high-throughput sequencing.

1.2.1. Biogenesis of miRNAs

The genes coding for miRNAs are located across the whole genome, many are in the introns of PCGs or are part of non-coding RNA genes. Often, miRNAs are located in the same locus, forming a polycistronic cluster that is usually co-transcribed and harbours copies of the same miRNA. The canonical biosynthesis pathway of miRNAs consists in the transcription by RNA polymerase II of the miRNA genes or the gene that contains the miRNA in order to form the primary transcript (pri-miRNA), which has around one thousand nucleotides. Transcription factors and other epigenetic marks can regulate the expression of these genes [13].

In the nucleus, pri-miRNA molecules undergo the first processing of the canonical two-step maturation pathway that leads to mature miRNA molecules (Figure 2) [14]. The long flanking tails of the stem-loop are cleaved from the primary transcript to create the precursor miRNA (pre-miRNA) transcript, a hairpin of around 60 nucleotides. The Microprocessor complex, which contains the Drosha RNase III, carries out the cleavage. The second maturation step happens in the cytosol, to which the pre-miRNA is exported via the Exportin5 pathway. In the cytosol, Dicer endonuclease cleaves the loop of the hairpin creating a duplex of approximately 21 nucleotides. Dicer, like Drosha, is a class III RNase and associates with Protein Kinase, PACT and TRBP [15]. Each of the duplex strands are known as mature miRNAs. Usually, one of the molecules is functional while the other, also called passenger strand or star miRNA (miRNA*) is degraded. Some miRNA genes produce pri-miRNAs by alternative non-canonical pathways.

Eventually, miRNA duplexes are loaded into an Argonaute protein. The protein complex created by Argonaute, the miRNA and other cofactors is called the RNA-induced silencing complex (RISC). The orientation by which the duplex binds Argonaute drives the choice of mature miRNA strand to be retained, and it depends on the capacity of each strand to bind to the pocket within Argonaute. At this point, the chosen mature miRNA would be ready for targeting mRNA molecules by pairing with them.

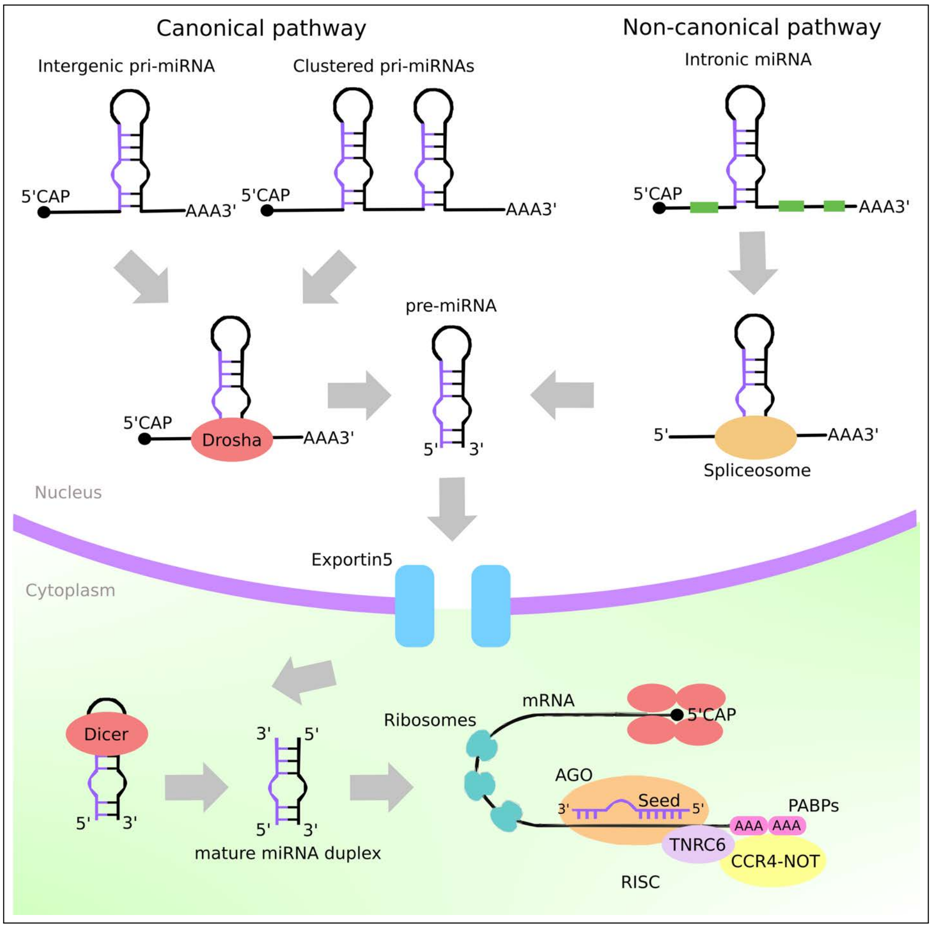


Figure 2: Biogenesis and mechanism of action of miRNAs, inspired by Saliminejad *et al.* [14]. Canonically produced pri-miRNA transcripts are cleaved in the nucleus to form the pre-miRNA transcript, which is transported to the cytoplasm. There, Dicer cleaves the loop of the hairpin producing a pair of mature miRNAs. Mature miRNAs are loaded into RISC in order to silence their target genes.

1.2.2. Mechanism of action of miRNAs

In animals, the recognition and repression of target RNA transcripts can have two mechanisms of action. If the miRNA and the target site pair with perfect or almost identity the miRNA directs slicing of the target transcript in a similar way as siRNAs work. This process is much more common for plant miRNAs than animal miRNAs. In mammals, however, the dominant repression mechanism does not cleave the target transcript and does not need an extensive pairing: Translation of the mRNA is inhibited or the mRNA is cleaved in a non-sequence-specific manner.

These two kinds of silencing without extensive pairing that take place in animals need different proteins complexes (Figure 2) [16]. On the one hand, the indirect degradation of mRNAs mediated by miRNAs requires the reduction of their stability. For that, several proteins are recruited by the adaptor protein TNRC6, such as deadenilases (CCR4-NOT and PAN2-PAN3), poly(A) binding protein (PABP) or 5′ cap removal proteins. The shortening of the poly(A) tail will make the mRNA enter the 5′-3′ degradation pathway. On the other hand, translation of mRNAs can also be inhibited without an actual change in mRNA level. The precise molecular mechanism for this remains unclear, but it seems that miRNAs inhibit translation initiation by meddling with the eukaryotic initiation factor 4F (eIF4F) complex [16].

As previously mentioned, miRNAs repress target gene mRNA by base pairing between both transcripts. Target recognition is primarily mediated by the seed region of the miRNA, defined as nucleotides 2-7 or 2-8, and the 3′ UTR regions of mRNAs [17]. Perfect matches in the 7 nucleotide seed region, sometimes extended to an 8th nucleotide, perform the bulk of the repression, as they are the most effective. Nevertheless, 6 nucleotide matches can also perform repression, albeit with less strength. Pairing to the 3′ region of the miRNA around nucleotides 13-16, known as 3′ compensatory sites, can supplement pairing to the seed region and increase repression efficacy and affinity [17]. The rest of the miRNA nucleotides can also help pairing, but the seed region remains the main driver. Because of the small size of the seed, target sites arise in the 3′ UTRs very easily, and thus, each miRNA family (miRNAs sharing the same seed) can regulate several mRNAs, with an average of 300 targets per family under selective pressure, and much more if considering the low efficiency 6 nucleotide target sites [11]. Altogether, miRNAs add a layer of post-transcriptional gene expression regulation that complements and finely tunes other regulation mechanisms.

1.3

Long non-coding RNAs

1.3.1. Definition and general features

Long non-coding RNAs (lncRNAs) are a heterogeneous class of transcripts longer than 200 nucleotides lacking protein-coding potential that are present across a variety of eukaryotic species [18]. The 200-nucleotide threshold is an arbitrary limit that serves for separating these transcripts from other well characterised smaller non-coding RNAs, such as transfer RNA (tRNA), microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs). The biogenesis of lncRNAs seems to be very similar to mRNAs: they are transcribed by RNA polymerase II and are post-transcriptionally modified like mRNAs with 5′ capping, polyadenylation and splicing [19]. Those are not universal properties of all lncRNAs due to the high heterogeneity and the lack of understanding of this class of genes. Nevertheless, lncRNAs show significant differences with mRNAs.

Regarding the expression patterns of lncRNAs, they show an exceptional cell type, tissue, developmental stage and disease state specific expression, and are expressed at lower abundances than mRNAs. These features were observed by large scale multi-tissue analyses of human gene expression [20, 21], and have also been confirmed in sheep and goat [22]. Brain and testis express the highest amount of lncRNAs, which are predominantly tissue specific or tissue enriched [20, 21], supporting the hypothesis that such transcripts are important for the acquisition of specific phenotypic traits. Besides, lncRNAs show higher expression variability across cell lines and tissues than PCGs [21], with a higher natural expression variation than PCGs between healthy human individuals [23].

LncRNAs usually have less exons than PCGs, with a striking tendency to have only two exons [21]. Their exons are slightly longer and their introns are longer than those of PCGs, and because they have less exons, lncRNAs are usually shorter [21]. Another typical feature of eukaryotic mRNAs is RNA splicing. LncRNAs can be spliced similarly to mRNAs, with the same splicing motifs and splicing machinery, but their splicing efficiency is lower due to weaker internal splicing signals [24]. LncRNAs that have been functionally characterised usually show more efficient and consistent splicing [24]. The epigenetic modifications regulating lncRNA expression profiles have also differences with mRNAs. They seemingly follow the same rules as PCGs, that is, expressed lncRNA promoters are enriched for the histone modifications H3K4me3, H3K9ac and H3K27ac [21]. Because of this, these modifications have been used as a proxy for the identification of lncRNAs. In spite of that, recent studies show that active promoters of some lncRNAs are surprisingly more enriched in the H3K9me3 histone modification, associated with transcriptional repression, than promoters of active mRNAs, and are slightly depleted in H3K4me3, H3K4me1 or H3K9ac [24]. Another epigenetic modification, DNA methylation, also differs in lncRNAs, with a higher methylation level around the TSSs than mRNAs [25, 26].

1.3.2. Classes of lncRNAs

We say that lncRNAs are part of a heterogeneous class of transcripts since not all of them have all the features explained above. The basic biology behind this class of transcripts is still under investigation, and it seems that there are specific sub-classes behind the catch-all term "lncRNA", even if we are not yet capable of clearly distinguishing them. One of the most common way of classifying lncRNAs is by their genomic relation with other genomic elements such as PCGs or regulatory elements, as those relations could give a hint about their biological function or biogenesis (Figure 3) [27].

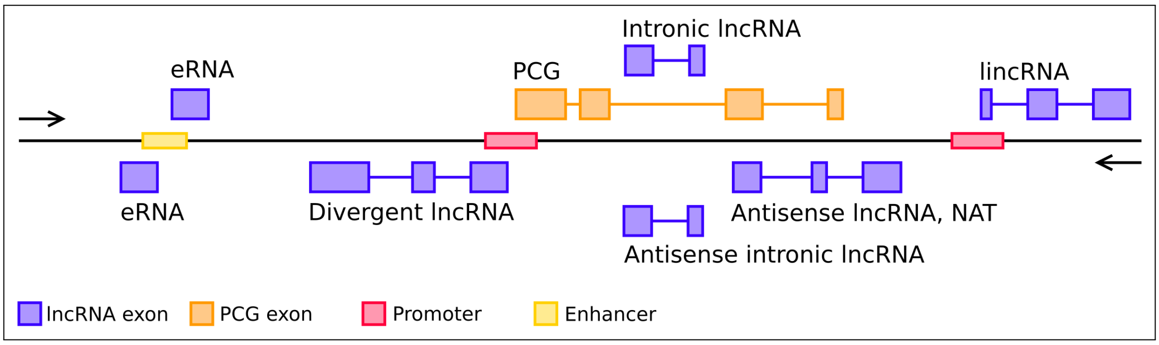


Figure 3: Classification of lncRNAs. The figure illustrates the different classes of lncRNAs based on their location in relation to PCGs and other functional elements.

The most typical class of lncRNAs are long intergenic non-coding RNAs (lincRNAs), non-coding transcripts located between genes that are not associated with nearby genes [18]. Many lncRNA studies only focus on this class since expression profiles, sequence conservation and experimental characterisation are easier to interpret than those of transcripts that overlap or are very close to other genes. Famous and well-characterised examples of this class include NEAT1 and MALAT1, which are involved in organisation of nuclear structure [28] or NORAD, which is an abundant, unspliced, polyadenylated and conserved mammalian lncRNA functioning as a decoy for RNA-binding proteins involved in genomic stability [29].

These intergenic RNAs are usually treated differently if they occur very close to another annotated gene, and sometimes are classified into a specific sub-group. If localised in a sense orientation or convergent orientation in relation to another gene they are still called lincRNAs, but because of that closeness, those genes are often first surveyed in search of potential functions. If the lncRNA is divergently located in relation to another gene, they can be called divergent lncRNAs or promoter-associated long RNAs [27]. Divergent lncRNAs are transcribed on the opposite strand of another mRNA or lncRNA gene, their transcription start sites (TTSs) are closely located and share the same bidirectional promoter [30]. The development of new sequencing techniques such as global run-on sequencing (GRO-seq) showed that divergent transcription is very common at vertebrate promoters [31] and nowadays promoters are thought to be inherently bidirectional. Nevertheless, the transcripts originated in the antisense direction of a PCG can have different properties: Some of them may be short and unstable by-products of active promoters called promoter-upstream transcripts (PROMPTs) [32], while others may be bona fide functional and stable divergent lncRNAs [30]. This class is one of the most abundant, they are coordinated with the expression of the adjacent genes and are enriched in essential developmental regulatory genes [30, 33]. Compared to standalone intergenic lncRNAs, divergent lncRNAs have stronger promoters and are less tissue specific, and this broader expression profile is related to a higher density of TF motifs in their promoters [34].

Similarly to how promoters work, active enhancers are also pervasively transcribed in a bidirectional manner, creating enhancer-derived lncRNAs or enhancer RNAs (eRNAs) [35]. These RNA products are less stable, shorter, unspliced and are not polyadenylated [36]. Due to the common features of enhancers and promoters, such as the chromatin states, eRNAs could be the equivalent RNA by-products to promoter PROMPTs [37]. Enhancer RNAs are more tissue specific and have less strong promoters than intergenic lncRNAs and divergent lncRNAs [34]. Due to these features, the biological function of eRNAs is currently debated, apart from being by-products of enhancer activity.

LncRNAs that partially or completely overlap another gene in its antisense strand are called antisense lncRNAs or natural antisense transcripts (NATs) [27, 38]. The production of non-coding transcripts from the antisense strand of PCGs is very common in eukaryotes, and it is thought that in many cases these antisense RNAs have a regulatory relationship with the gene in the sense strand [39]. They can be similar in structure to other lncRNAs, showing splicing and polyadenylation. The expression of the antisense lncRNA can have a positive or negative effect on the neighbour gene expression, forming self-regulatory loops, but antisense lncRNAs can also be independently regulated and have another function [40]. For instance, the lncRNA ANRIL (officially CDKN2B-AS1) is located in the antisense strand of the CDKN2B gene and negatively regulates it and other close genes via chromatin modifying complexes [41].

Intronic RNAs have often been regarded as junk sequences, as those sequences do not seem to have any function and spliced-out introns are rapidly degraded, but stable intronic non-coding RNAs have been described. The category of intronic lncRNA, stable intronic sequence RNA (sisRNA) or totally intronic noncoding RNA (TIN) can be given to any non-coding transcript that overlaps completely or in part an intron of another gene [27]. An early survey of mRNA and EST public sequences revealed that 74% of all annotated genes transcribed intronic RNAs from their introns and analysed their expression using a newly designed oligoarray platform [42]. Intronic RNAs can be dependent on the splicing machinery and thus be produced by the debranching of the intron lariat created by splicing, or they can be independent of the splicing machinery and be transcribed from the harbouring gene via independent promoters [43]. They are predicted to regulate host gene expression through feedback loops or disturb splicing by acting as protein decoys, but splicing-independent intronic RNAs may also function independently of the host gene [43]. Related to these RNA products, other molecules that have been classified as lncRNAs are circular RNAs, formed by several introns and exons by a mechanism called back-splicing [44]. Sense lncRNAs that overlap coding mRNAs on the same strand without encoding any proteins also exist, but this overlap makes the computational and experimental study of these transcripts quite difficult. GENCODE groups such spliced lncRNAs under their "sense overlapping" biotype [45]. In addition to the mentioned lncRNA classes, other more unclassifiable transcript groups exist, which are classified under the "processed transcript" biotype by GENCODE due to the complexity in their structure [45].

The existing classifications of lncRNAs, as well as the exact definition of what they are, are still based on very descriptive features and there is not a general agreement on them. These classifications, albeit practical, have several shortcomings caused by the yet understudied field of lncRNA biology. For instance, the lncRNA transcripts in human gene annotations differ from each other, and in non-model species the catalogue is much smaller. There can be an overlap between different classes, for example one transcript can be intronic and antisense at the same time. Finally, the functional dissection of lncRNAs will allow for a more logic definition of lncRNA classes.

1.3.3. Evolutionary conservation of lncRNA sequence and expression

The study of conservation patterns has greatly helped our understanding of gene evolution between species and has enabled the transfer of functional information from deeply studied model species like human or mouse to other less studied species. This knowledge is mainly based on the sequence evolution of PCGs and other non-coding RNA genes than lncRNAs, in which sequence conservation means that natural selection forces act in favour of functionally relevant sequences. Identification of conserved structures in lncRNA sequences should also help to detect those lncRNAs from the genome and link them to important biological functions. Nevertheless, in clear contrast to PCGs and other well-known non-coding RNA classes, in general lncRNAs show very little sequence conservation between species [18]. Most mammalian lncRNAs lack any ortholog outside of vertebrates and compared to PCG sequences, lncRNA sequences evolve very rapidly [18].

Comparative genomic approaches have been used over the last decade in order to identify and compare lncRNAs between species with different phylogenetic distances [46–50]. Overall, all studies agree on the fast evolution of lncRNA sequences compared with PCGs, with fewer orthologous genes being found as evolutionary distances grow. The selective constraint on lncRNA sequences is usually weak, but significantly above the genomic background [51]. In mice, for instance, nearly half of intergenic lncRNA loci have been gained or lost since the last common ancestor of mouse and rat, compared to 10% of PCGs [52]. In humans, these comparative genomic studies have specially highlighted the high numbers of specific lncRNAs at human, hominid and primate taxa levels. They identify 11000 primate specific lncRNAs [46] and find that around 20% of human lincRNAs are not expressed beyond chimpanzee and are undetectable even in rhesus [50]. Another common conclusion is the existence of a smaller set of highly conserved lncRNAs across taxa, with few of them spanning beyond mammals [48–50, 53]. These conserved transcripts show strong purifying selection in their genomic loci, exons and promoters, and were predicted to have diverse roles in processes from stem cell pluripotency to proliferation [53].

The same comparative genomic studies usually performed comparative transcriptomics analyses at the same time in order to assess the conservation of lncRNA expression across species, tissues and developmental stages [46–48, 50]. Expression patterns of PCGs tend to be highly conserved among mammals, and lncRNAs also show this feature, albeit with lower conservation [51]. In the case of lncRNAs, temporal changes in expression variation are more evolutionarily conserved in tissues during development [47]. Besides, the lncRNAs expressed during embryonic development show more conservation at promoter and exon sequence level [51]. Considering the lncRNAs that are expressed in mammals, they show remarkably strong conservation of tissue specificity, even if their splice-site and sequence turnover suggest that splice-sites and exact sequences are not critical [46, 48, 50].

Based on these observations, it became evident that the approaches for understanding lncRNA evolution should be different to the traditional comparative sequence analysis used to study PCGs and other noncoding RNAs. Alternative approaches have been proposed with this goal [54, 55]. First, the alignable sequences of known lncRNAs with homologues in other species are much shorter than in PCGs [48]. This calls for a focus on the conservation of shorter patches of the lncRNAs. These patches could be the functional part of the transcript, while the rest of the sequence may be dispensable and may tolerate major changes in gene architecture (e.g. splice-site turnover).

Secondly, instead of primary sequence conservation, the structure of the lncRNA could be conserved. LncRNAs as other RNA classes fold into secondary structures that could be maintained even if there are mutations due to the base-pairing properties of RNA [55]. Conservation of lncRNA structures could be widespread as the human genome has more than 4 million evolutionarily constrained RNA structures within mammals and most of them are outside of sequence-constrained regions [56]. Third, analyses could focus on the conservation of transcription status to identify homologues with low or almost inexistent sequence identity. In some cases, the promoters and tissue-specificity of lncRNAs are evolutionarily conserved while losing sequence identity in the gene body [46, 51]. In fact, it has been proposed that the action of transcription itself is only necessary in some cases for observing a regulatory effect [57].

Finally, following with the idea of the biological function being independent of RNA product sequence, we can expect that the position of the region that is transcribed would be conserved. In this way, if two lncRNAs are located between the same orthologous genes in the same relative orientation in two different species we say that they show syntenic or positional conservation and the fact that it is conserved could be a signal of functionality [58]. A recent work found 665 lncRNA promoters in mouse and human that are preserved in genomic position relative to orthologous coding genes and found that were related to developmental transcription factors [59]. As an example, there is a syntenicaly conserved lncRNA (IFNG-AS1) near the IFNG gene in human and mouse with limited sequence conservation that has been shown to regulate the expression of IFNG independently of its RNA product [60].

Taken together, the diverging conservation levels of lncRNAs and the previously discussed classification into biotypes based on their genomic environment suggest a division of lncRNAs into two broad groups [49]. On the one side, a group of lncRNAs that shows sequence conservation, even if in small patches, or structural conservation. On the other, lncRNAs whose RNA product sequence is not conserved but show conservation of promoter sequence, tissue expression or synteny. This kind of lncRNAs share properties with eRNAs and some divergent RNAs. Most conserved lncRNAs should be functional, while the biological function of the second group is debated. Importantly, thousands of reported transcribed sequences do not met any conservation criteria, which has led to their classification as noise RNA transcripts produced due to the pervasive transcription of the mammalian genomes, although biological function for all of those transcripts cannot be rolled out.

1.3.4. Biological function of lncRNAs

Even though lncRNAs do not code for proteins, they can be functional molecules. Since the first studies that described and then found a role of XIST in X chromosome inactivation [61], several other lncRNAs have been attributed a biological function. This process has been slow and is currently undergoing due to the difficulties of studying these molecules compared with PCGs. Thanks to their capability of forming different structures and interact with proteins, DNA and other RNAs, they are very heterogeneous in their mechanisms of action [62]. So far, it remains a matter of debate to which extent lncRNAs are functional: Do all lncRNAs have a function? Is the RNA product necessary for function? Two recent systematic functional profiling studies of more than 100 lncRNAs each could give a hint about the magnitude of lncRNA function. In human fibroblasts, over 25% of lncRNAs were found to affect cell growth, morphology or migration [63]; and, in fission yeast, 60% of lncRNAs deleted with CRISPR-Cas9 genome editing and 90% of overexpressed lncRNAs showed a phenotype under certain conditions [64].

LncRNAs that have been functionally characterised until now essentially carry out their functions as structural or regulatory RNAs. These transcripts have been broadly classified according to their cellular function into those that regulate local chromatin structure and/or gene expression in cis versus those that perform cellular functions outside their site of transcription in trans [62, 65]. Regardless of a cis or trans effect, lncRNAs exert their functions at different levels of the cellular machinery. They have been proved to regulate chromosomal structure, regulate chromatin accessibility, regulate polymerase II transcription, alter pre-mRNA splicing, regulate other RNA molecules post-transcriptionally, modulate mRNA stability, regulate translation or modulate post-translational modifications [66]. In terms of the exact molecular mechanism of action, lncRNA molecules can act as signals, decoys, guides or scaffolds, depending on how they interact with other molecules [67].

Cis-acting lncRNAs constitute a substantial fraction of lncRNAs with an attributed function and have been demonstrated to modulate the expression of target nearby genes by altering chromatin structure, chromatin modifications or transcription control [68]. These mechanisms seem to be related to the fact that lncRNAs tend to overlap in their TSSs with enhancers more than PCGs and that many enhancers produce stable lncRNAs. The local effect of cis-acting lncRNAs is relative, for instance, XIST is able to inactivate a whole chromosome [61], while others act within their topologically associating domain (TAD), defined as regions with high density of chromatin interactions, or just affect a single gene. For example, the human lncRNA UMLILO RNA product is required for the induction of several chemokine genes located within its TAD [69]. When stimulated with TNF, UMLILO recruits and binds to a protein complex that deposits histone modifications in the promoters of those chemokine genes, promoting their expression. Another example would be the lncRNA Morrbid, which controls apoptosis in many short-lived immune cells by regulating the neighbouring pro-apoptotic Bcl2l11 gene [70]. It does so by interacting with PRC2 and thus introducing the repressive H3K27me3 chromatin mark. Interestingly, in CD8 T cells during infection, this lncRNA has the opposite effect, it promotes Bcl2l11 gene expression, highlighting the potential of cell type-specific functions of this class of genes [71].

It is possible that a big proportion of lncRNAs actually represent RNAs that are transcribed from enhancers or promoters, do not perform sequence specific functions and have a local effect in cis [72]. This idea is backed by their predominant localisation to the nucleus, low expression level and their low sequence conservation [5, 21]. For instance, transcription of the lncRNA ThymoD, expressed from an enhancer region 700kb away from the Bcl11b gene in mouse in T cell progenitors, promotes the demethylation of CTCF motifs and supports Bcl11b expression by maintaining chromatin contacts between the enhancer and promoter [73]. Another lncRNA, PVT1, is located downstream of MYC transcription factor gene locus in vertebrates and negatively regulates the expression of MYC by competing for the binding of the same enhancers. Thus, its DNA acts as a decoy, while the RNA product has other functions [74]. The previously mentioned IFNG-AS1 gene, which controls the expression of IFNG, is another example of function independent of RNA product [60].

Various studies describing a depletion of lncRNA expression without perturbing its original gene locus demonstrate that lncRNAs have an active biological function in distal parts of the cell, from other nuclear domains to the cytoplasm [62]. For these roles, the actual RNA product should be necessary, unlike many cis-acting lncRNAs. Some lncRNAs can interact with chromatin complexes and regulate the expression of distant genes: FENDRR, transcribed bidirectionally with FOXF1, binds to PRC2 and/or TrxG/MLL complexes to regulate the expression of important transcription factors [75]. NKILA is a lncRNA that binds to the NF-κB complex to block its phosphorylation and inactivate the complex and is essential to prevent over-activation of NF-κB pathway during inflammation [76]. LncRNAs can also bind to RNA and DNA binding proteins, especially transcription factors. For instance, RMST binds to SOX2 to coregulate a large pool of downstream genes implicated in neurogenesis [77]. Finally, OIP5-AS1 (Cyrano) is a lncRNA that is part of a regulatory network where it represses miR-7 microRNA via target-directed microRNA degradation, which in turn enables the accumulation of Cdr1as circRNA in the mouse brain [78]. In this case, we say that the lncRNA is acting as a miRNA sponge or competing endogenous RNA (ceRNA).

There are many other functions associated with lncRNAs, which have been extensively reviewed elsewhere [62, 65, 66, 68], but the few functional lncRNAs mentioned in this section reflect their highly heterogeneous nature and biological implications.

1.4

The genomics of non-coding RNAs

1.4.1. Methods for transcriptome sequencing

Transcriptomics is defined as the study of the transcriptome, the complete set of RNA molecules, also known as expression profiling, as it is the study of RNA expression levels in a given cell or tissue type. It allows the study of the genome fraction that is being transcribed and the dynamics of that expression. Most transcriptome sequencing methods, from the early ones to the latest technologies, usually start with the conversion of RNA to its complementary DNA (cDNA) with a reverse transcriptase, which is more stable than the RNA molecule and can be amplified by polymerase chain reaction (PCR). Before the emergence of high-throughput sequencing technologies, the method largely used was Sanger sequencing, developed in the 70s and based on selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication [79]. This is a very arduous method, yields less than 1000 bp per read, requires a known sequence to prime to and only allows the sequencing of a single read. Nevertheless, because of its high per-base accuracy (>99.999%) and the big initial efforts using this technology to annotate the human and mouse genomes, nowadays the gene collections from NCBI RefSeq and GENCODE are still largely constituted of models obtained with Sanger sequencing. It is also still routinely used for specific applications in clinical genetics or molecular biology, for instance.

Second-generation sequencing Second-generation sequencing (SGS) methods revolutionised RNA sequencing in the mid-2000s by enabling high-throughput parallel transcriptome sequencing, obtaining millions of short sequences from a single extracted RNA pool. At the beginning various implementations were commercialised, but the technology from Solexa/Illumina sequencing became the most popular and the industry standard. The technology behind SGS RNA sequencing (RNA-seq) is based on sequencing by synthesis, tracking the addition of labelled nucleotides as the DNA chain is copied. Prior to the actual sequencing, RNA must be fragmented into shorter molecules, it is converted into cDNA and adapter sequences are ligated to the double-stranded cDNA. The modified DNA is loaded onto a flow cell where amplification and sequencing take place. The DNA molecules anchor to oligonucleotides attached on the nanowells thanks to the adapter sequences and after an amplification step, rounds of synthesis begin with modified nucleotides. These nucleotides have a reversible fluorescence blocker so the DNA polymerase can only add a single nucleotide in each round. After each round, a camera determines which was the last added nucleotide and another round begins.

The hundreds of millions of sequencing reads typically obtained from a SGS run are in the range of 30 to 150 bases, they do not provide the full cDNA sequence, and have an accuracy of 99.9% [80]. Short reads do not suppose a problem if the objective is to quantify already known genes, as we can count the number of reads that align to those genomic loci with specific software. On the contrary, recovering full-length transcripts from short reads, in order to identify its exact sequence and exon structure, asks for algorithms capable of transcript-assembly. These computational methods are able to identify new genes and isoforms without a gene annotation, but due to the transcript overlap and alternative splicing, they do not produce fully accurate transcript models [81]. Illumina sequencing has been shown to be highly replicable, with little technical variation, nevertheless, it suffers from some biases related to library preparation or the synthesis cycles in the flow cells [82, 83].

Third generation sequencing In the last years, different third generation sequencing (TGS) technologies have emerged, which are based on long-reads. These technologies enable, for the first time, the possibility to sequence RNA transcripts from 5′ to 3′ end, without molecule fragmentation and thus, without having to computationally reconstruct the transcripts. This is obtained at the expense of per-base accuracy, sequencing depth and cost. There are two main long-read sequencing platforms commercially available, using deeply distinct principles: PacBio Single-Molecule Real-Time (SMRT) and from Oxford Nanopore Technologies (ONT) sequencing. On one hand, PacBio's implementation consists in a sequencing by synthesis method using nucleotides attached to fluorescent dyes and ligating hairpin adapter sequences to the cDNA molecule. In this way, each nucleotide added is tracked and the DNA polymerase can cycle over the circular structure in order to sequence the same read many times. Because the high error rate of SMRT sequencing, around 13%, this cycling enables the creation of a consensus sequence with less than 1% error rate, still higher than the SGS Illumina implementation [80]. On the other hand, ONT sequencing utilises the electrical signal produced when DNA molecules pass through pore proteins, as each nucleotide produces a specific signal. This method, despite being cheaper than others and being able to produce very long reads, has an error rate similar to the single-pass PacBio reads [80], so, on its own, it is still not the best option for transcriptome sequencing [84]. For other uses such as monitoring of pathogens or genome sequencing it is a very useful approach.

1.4.2. Bioinformatic analysis of miRNAs

1.4.2.1. Small RNA sequencing for annotation and expression profiling

The most common assay for the annotation and expression profiling of miRNAs is high throughput RNA sequencing. It is possible to search for miRNAs *de novo* from the genome, but because not all miRNAs are conserved and they are very short in length, evidence of transcription is often needed. SGS RNA-seq is a flexible method that allows for many modifications. For miRNA profiling, small RNA enrichment library preparations are used, which are commonly known as small RNA sequencing (sRNA-seq) or miRNA-seq, even if other short ncRNAs are also enriched.

MicroRNAs are characterised by unique structural features that separate them from other RNA families. MiRNA prediction programs for bona fide miRNA annotation make use these features [85]. These requirements include: (1) 20-26 nucleotide (nt) long reads, (2) a hairpin precursor of about 59 nt, (3) 2 nt offsets between the 5p and 3p arms (consequence of Drosha and Dicer processing), (4) at least 16 nt complementarity between both arms, (5) 5′ end homogeneity of expression, (6) genome encoding, (7) loop sequence between 8 and 40 nucleotides, (8) consistent expression of both arms and (9) phylogenetic conservation (not all miRNA families).

Most of the novel miRNA annotation tools are often wrappers of other tools that add their own algorithm for *de novo* discovery. They usually include unspliced alignment to a reference genome, alignment to known miRNA sequences from other species and *de novo* prediction from unknown aligned reads based on the thermodynamics and structure of the predicted hairpin structures. The most used tool is miRDeep2 [86], the field reference tool for miRNA quantification and discovery, but there are several others such as sRNAbench, which is included in sRNAtoolbox [87]. In the last years, new tools have been developed in order to profile other small RNAs present in sRNA-seq reads, even if miRNAs are by far the most numerous biotype in this datasets. Manatee [88] and DANSR [89] are two examples of these new tools.

1.4.2.2. In silico prediction of miRNA targets

Knowing which genes are regulated by miRNAs is necessary for the biological understanding of their functions. MicroRNAs produce remarkable changes in several physiological and pathological processes, thus, identification of miRNA-mRNA target interactions is fundamental for disentangling the miRNA-governed regulatory networks [90]. Because there is a high number of potential targets for each miRNA molecule, a computational approach is often needed to prioritise a number of targets that will be experimentally validated. Many tools that are capable of computationally predict potential miRNA-mRNA interactions have been developed and even if each program uses its own strategy, most of the features are shared among them. These tools can be divided into two groups depending on their strategy: (1) Tools based on empirical sequence characteristics and (2) tools based on statistical inference with machine learning.

The first kind of approaches employ several analyses to make the predictions. Seed match is an important factor that most tools rely on, because the sequences involved tend to be the most evolutionarily conserved regions, looking for Watson-Crick matches between those 6-8 nucleotides in miRNAs and 3′ UTRs is a common approach. Thermodynamic stability is another highly important feature of target prediction. If the hybridisation of a putative target mRNA with a miRNA is strong, it is more probable that that match is real. Usually the change in free energy (ΔG) in the hybridisation reaction can be used as binding strength indicator. Another important feature is evolutionary conservation. If the matching is preserved across species it is a good indicator that the target is real. Normally, the conservation is higher in the regions matching the seed but there is also conservation in the sequences that pair with the rest of the miRNA. Other commonly used features in target prediction are accessibility of target site, target site abundance, local AU content or site position distribution. Widely used tools based on sequence features include miRanda [91], TargetScan [92] and RIsearch2 [93].

Machine learning (ML) approaches use experimentally proven miRNA-mRNA interactions as references to make new predictions from unknown data, instead of using predefined sequence features. ML can be biased by the experimental approaches to infer confident interactions but enable the prediction of functional non-canonical interactions. In a recent update, TargetScan has added a neural network based model to the algorithm [94].

1.4.2.3. Nomenclature and databases

With the number of described miRNAs in the rise, the establishment of a consensus naming system became necessary, as well as guidelines with must-have features to confidently characterize newly described miRNAs [85, 95]. With some exceptions, all miRNA names bear the "miR-" prefix followed by a sequentially given identification number, and a species-specific three letter prefix is added before it if needed. Many miRNAs show high evolutionary conservation, so it is helpful to give orthologues the same identifier in different species. The identification number should be the same in two species if the sequence is the same and the ancestor sequence is the same for both. In the case of paralogs, which are very common in many miRNA families, the identifier is kept but a suffix is given, a number or a letter, depending on if the mature sequence is the same or not. Besides, when naming mature miRNA products the arm the miRNA is coming from is marked with a "-3p" or "-5p" suffix. It should be mentioned that there have been some proposals to modify these criteria to make it more intuitive [85], and that the HUGO Gene Nomenclature Committee (HGNC) uses a different naming system that keeps the miRBase sequential numbers.

The classic miRNA database that has been giving names is miRBase [96], but it is not consistently updated, it has not added many livestock miRNAs in the last years and it seems that many of the miRNAs are not very robustly backed. For instance, of the 1881 human miRNAs in miRBase (v.21), only 523 genes met the standards for miRNA annotation [85] and there were just 106 sheep miRNA genes. A recent alternative to miRBase would be MirGeneDB [97], which in its latest update (release 2.1) gathers 16670 manually curated miRNAs from 75 metazoan species across several phyla. Its development has been focused on phylogenetically relevant species and the sole livestock species supported is cattle. RumimiR database [98] is a comprehensive repository that stores ruminant miRNA sequences from the literature as they were published and is a useful resource to find out which sequences have already been detected, but it unfortunately lacks any naming curation. Regarding the major gene annotation sources, NCBI RefSeq only includes miRBase miRNAs and Ensembl extends the miRBase miRNA gene catalogue with the annotation of other evolutionarily conserved miRNAs. Because of this, sheep miRNAs discovered in several experiments have not been properly characterized and it is difficult to find them in sequence repositories.

1.4.3. Annotation of lncRNAs

Gene annotation is the process of describing gene boundaries and structures within a given genome and is the key mechanism through which information is leveraged from sequence to function. The major targets of annotation pipelines are transcripts and the process of obtaining functional information about those transcripts, from definition of gene type to biological function dissection, is called functional annotation [99]. Eukaryotic genomes contain not only protein coding genes, but also other non-coding genes such as lncRNAs or miRNAs, and particular approaches are needed for the annotation of these gene classes.

Annotation strategies can be classified into *de novo*, using solely the genome sequence to predict transcribed sequences, or evidence-based, which adds experimental evidence of transcripts to the *de novo* predictions. Predictions from the genome in eukaryotes are difficult because of the complex gene structure with short interspersed exons and alternative splicing events. Nevertheless, existing algorithms, used by the main gene annotation consortia, work reasonably well for protein coding gene exons, as open reading frames (ORFs) can be identified and they are highly conserved [100]. For non-coding RNAs such as miRNAs, tRNAs, snoRNAs or rRNAs the conservation in terms of sequence and the prediction of thermodynamically stable secondary structures can help due to the absence of ORFs, but most lncRNAs lack those features.

Because of this, experimental evidence of the lncRNA transcription and structure is needed to properly annotate those genes [101]. This evidence mostly comes from mapping sequencing reads to the genome of interest by sequence similarity with algorithms that take splicing into account. The human and mouse annotations produced in the 2000s used the low-throughput Sanger sequencing and the gene models were manually curated by numerous research groups. It was possible thanks to big consortia with huge human and economic resources but it is nowadays unfeasible to put that effort into all other species. The hundreds of species annotated in the public databases after those model species have used SGS short read RNA-seq to help in the transcript definition because of its high yield and low cost [99]. Identifying novel transcripts using the SGS RNA-seq is a challenging tasks, because short reads rarely span across several splice junctions, making it difficult to directly infer all full-length transcripts and transcription start and end sites are not always present in the reads [102]. However, with paired-end sequencing, higher coverage and replicates it is possible to obtain a decent transcriptome for those species in which it has not been put a big effort before. Considering lncRNAs, for now, the only way to properly annotate them is through experimental evidence and the recent development of TGS methods is helping in this task, even if it has only been used in few model species. For instance, long-read sequencing, coupled with targeted RNA capture (CaptureSeq), has been used by the GENCODE consortium to enlarge the human lncRNA annotation with full-length transcripts [103]. Recently, PacBio long-read sequencing has been applied in cattle, showing that there are still many non-coding genes to be described [104].

SGS RNA-seq is a flexible method that allows for more applications than just full transcriptome sequencing by modifications in library preparation. These applications can give additional evidence for the lncRNA transcription and help in the transcript model definition [99]. The usual selection of transcripts with poly(A) tails hinders the discovery non-polyadenilated lncRNAs, so ribosome RNA depleted total RNA-seq is preferred for lncRNA profiling (Figure 4). Among other applications, Cap Analysis of Gene Expression (CAGE-seq) produces reads from the 5′ capped end of the transcripts, which can serve to identify the exact TSSs [105]. RNA Annotation and Mapping of Promoters for Analysis of Gene Expression (RAMPAGE-seq) is an inproved version of CAGE-seq that allows for longer reads [106]. PolyA-seq captures RNA sequence immediately upstream of the polyA tail, in order to identify accurate transcription termination sites (TTSs) [107]. CaptureSeq is used to pull down specific RNA transcripts and deeply sequence them, which is useful for lowly expressed lncRNAs [108]. Global run-on sequencing (GRO-seq) is an assay that measures nascent RNA of transcriptionally engaged RNA polymerases [31]. It is particularly suitable for detecting lowly expressed unstable transcripts such as eRNAs and study divergent transcription [33].

Other genome-wide "omics" assays can also provide structural or functional evidence for lncRNA annotation. ChIP sequencing (ChIP-Seq) is a powerful epigenomic approach for identifying genome-wide DNA binding sites for transcription factors and

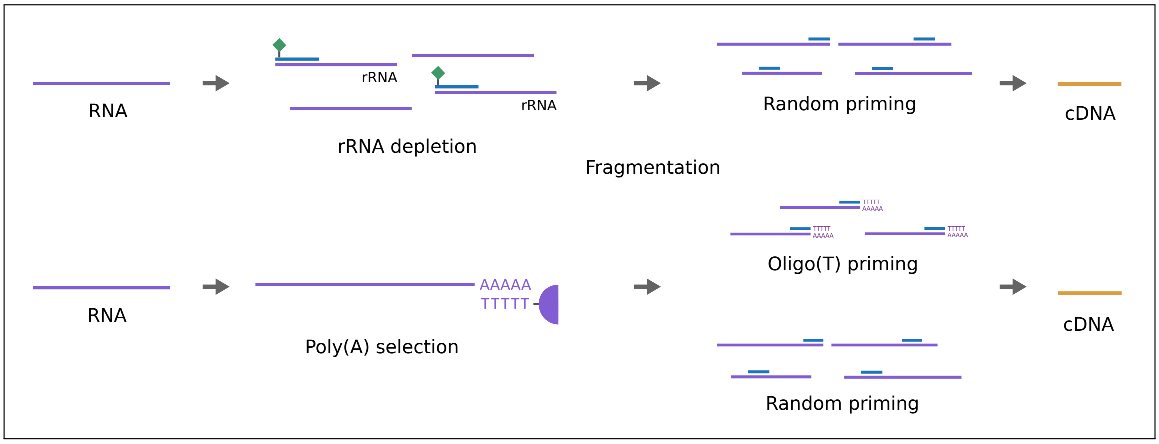


Figure 4: Schematic overview of the main library preparation protocols for SGS RNA-seq. After RNA extraction with an appropriate method, the cDNA library can be constructed by rRNA depletion or poly(A) selection.

histone modifications. Genomic overlaps between potential novel lncRNA transcripts and marks of histone modifications that are associated with promoters or enhancers can be used to provide a higher level of confidence to those loci. This strategy was first used to identify more than 1500 novel lncRNAs in mouse and humans [53].

All in all, most studies in non-model species have been based on short-read SGS methods because of their high-throughput, low cost and high accuracy, even though, with this methods, precise transcript structures are not always correctly predicted. A typical novel lncRNA identification pipeline would start with mapping the sequencing reads to the genome with a splice-aware tool followed by a transcriptome assembly [109]. Then the assembled transcriptome is filtered to search for novel transcripts. Some of the filters applied are structural, for instance, transcripts shorter than 200 nucleotides are removed and transcripts that overlap coding genes in the same strand are removed. Other filters involve the assessment of the evaluated transcripts' coding potential. There are several tools that employ machine learning algorithms and sequence features, like ORF length or K-mer frequencies, to infer the probability of a sequence to be protein coding [110]. It is also possible to scan the evaluated transcripts for known protein domains included in public databases. At the end, we obtain a set of novel non-coding transcripts to perform different analyses: expression profiling, comparative genomics, disease association or experimental functional validation, among others.

1.4.4. Searching for lncRNA function

Demonstrating the function of a non-coding RNA molecule is an arduous process, as it is the contrary, to prove that it is a product of transcriptional noise. Sequence conservation is one of the clearest evidences of function and allows the assignment of a function to those transcripts found in non-model species that have an orthologue with proven functionality in other model species. For the rest of the lncRNAs, including those with a conserved but functionally uncharacterised orthologue in humans or mice, the starting point to investigate their biological functions passes through in silico genomic data mining. This involves the utilisation of biological datasets to extract meaningful knowledge regarding a specific biological question, identifying patterns and relationships within the data.

Regarding lncRNA research, their well-known properties are their tissue and developmental specificity and their involvement in the regulation of other genes, thus expression profiling of lncRNAs along PCGs and other RNA genes can be used to get clues about their biological function. In fact, expression levels of lncRNAs should be more related to function than those of PCGs, as it is known that mRNA abundance is not always coupled with protein abundance. To disentangle the potential biological functions, often, samples from animals or cells under different biological conditions are used. The datasets needed for these kind of functional genomics analyses can come from two sources: 1) We can design an experiment specifically tailored to our biological question. This involves producing new data by extracting the genetic material from experimental cells or animals and sequencing it. 2) We can take advantage of the increasing number publicly available high-throughput datasets to ask a novel biological question, different from what the data was initially produced for [111]. The two most popular applications for RNA-seq expression data are differential gene expression (DGE) analysis and co-expression network analysis.

DGE analysis is used to identify genes with statistically significant differences in expression levels between two conditions. It is one of the most common applications of RNA-seq data, as a change in expression levels might be an indication of involvement in the biological processes activated between two conditions. In the same fashion as PCGs, if a lncRNA is found to be differentially expressed, it can give a hint about its involvement in a disease, for instance. However, even if a significant change in expression can be due to secondary processes, this method allows us to prioritize candidate genes for further functional validation. With respect to the statistical methods needed, the analysis of short-read transcriptome sequencing is not a trivial matter. Expression values are not normally distributed, sample outliers are not uncommon and the analysis consists in testing several thousands of genes from relatively few observations, so common parametric methods do not usually work. Non-parametric methods do not assume any distribution, but they are underpowered in most cases, because experiments tend to be composed of few samples due to the high cost of RNA sequencing. This has led to the development of tools based on a negative binomial distribution that fits better to expression data. Examples of these tools are edgeR [112], which uses an overdispersed Poisson model and DESeq2 [113], which uses a generalized linear model in order to account for covariates. Nevertheless, for very high number of samples, the classic non-parametric methods such as Mann-Whitney U test may perform better than these tools [114].

Gene co-expression network analysis is an unsupervised clustering method that associates genes between them using correlation strengths as a measure of similarity in order to find common patterns of expression. It allows to determine clusters of genes with the same expression pattern, and following the guilt-by-association principle, correlated genes are expected to be involved in the same biological processes and pathways [115]. By leveraging the known biological functions of the genes in one cluster, one can get a hint about the potential functions of uncharacterized genes such as lncRNAs. Besides, with data from functional experiments, it can also be tested if a particular gene cluster has a different expression pattern between different conditions. Another application of co-expression networks are differential co-expression network analyses, which have the advantage of detecting condition-dependent interactions between genes [116]. In this way, the correlations of individual gene pairs are tested to assess if there are differences between conditions. For instance, the gain or loss of correlation between a lncRNA and other PCGs can be due to regulatory relationships between them.

1.4.5. Databases

The most common resources from which lncRNAs can be obtained are the genome annotations produced by various consortia and institutions. Ensembl and NCBI Reference Sequence Database (RefSeq) are the most widely used databases of transcript structures and they contain gene annotations for hundreds of species, including the main livestock species. As for Ensembl, there are big differences between the annotations of main model organisms and other vertebrates. Human and mouse annotations have been built by the analysis of experimental data through bioinformatic pipelines and are systematically manually curated by the GENCODE consortium [117], including lncRNA genes. The GENCODE annotation comprises 18811 human lncRNA genes (version 39), very close to the amount of PCGs, and 13186 mouse lncRNAs (version M28). Remarkably, in the last releases this consortium has been using targeted long-read sequencing to get full-length accurate transcripts. In contrast, the rest of species do not have such a vast number of annotated lncRNAs and those are only predicted automatically with a bionformatic pipeline. As a consequence, livestock lncRNAs are also much more less accurate, especially regarding transcription start sites, and most of them are intergenic transcripts. Ensembl (v.105) contains 2229 sheep lncRNAs, 2705 goat lncRNAs, 1480 cattle lncRNAs, 6790 pig lncRNAs and 7241 horse lncRNAs, among other species.

Other resources containing lncRNA annotations include RefSeq, NONCODE [118] and ALDB [119]. In RefSeq, the other widely used resource, lncRNA gene annotation is done in a similar way as in Ensembl: the human gene annotation is highly manually curated while most other species are automatically annotated using their own gene annotation pipeline. NONCODE is a lncRNA-specialised database that collects data from literature and performs some additional analyses. It contains a limited number of species, excluding sheep or goat for instance, but it gathers more than 20 thousand cow lncRNA loci. ALDB is a livestock-specific lncRNA database but it is outdated and covers just three species.

1.5

Non-coding RNAs in farm animals

Domestic animals are of great importance as sources of high-quality products for human consumption, as disease models in biomedical research and for the prevention of zoonotic infections. Phenotypic variations among domestic animals can be related with mRNA expression and, therefore, with miRNAs, which are their key regulators, and with other non-coding genes. During the last decade, the non-coding genome has gained considerable attention in the pursuit of genotype to phenotype annotation. The human and mouse genomes have been extensively characterised, which has led to the discovery of a diverse and numerous set of novel noncoding genes, outnumbering PCGs. Other no-model species are beginning to have their genomes annotated with non-coding genes in the last years, but the qualitative and numeric differences with the human annotation remain vast in farm animals [109, 120].

The rich human and mouse gene and functional element annotations produced by big consortia such as ENCODE have limited translational utility in livestock species. Although mammalian PCGs are generally highly conserved, important genes involved in speciation, like those with immune or reproduction functions, are not so widely conserved [121] and those traits are under positive selection in ruminants [122, 123]. Besides, many non-coding RNA genes are not conserved between distant mammal species or even between close species. At the regulatory sequence level, the Mouse ENCODE Consortium found that there is a large degree of divergence of sequences involved in transcriptional regulation, chromatin state and higher order chromatin organization [124].

Thanks to genome-wide association studies (GWAS) based on high-resolution genotyping and sequencing, thousands of genomic regions have been associated with complex phenotypic traits and diseases in farmed animals. These associations can be browsed in the comprehensive AnimalQTLdb database [125]. However, most trait-associated variants in sheep, as in most livestock species, lie within non-coding genome regions and in regions in close proximity to protein coding genes. Therefore, it is imperative to characterise these regions of the genome in order to unravel the molecular and genetic causes, functional SNPs, causative genes and pathways underlying the variability of complex production and health traits in livestock species [126].

1.5.1. Sheep miRNAs

Because miRNAs are much more conserved than lncRNAs and have predictable structures, they began to be studied earlier in livestock species [127]. The first sheep miRNAs were identified through sequence homology search from Callipyge sheep, which display a muscle hypertrophy phenotype [128]. They identified several miRNA genes in the imprinted region that causes the hypertrophic phenotype. Another known example of the phenotypic interest of miRNAs is the case of the muscular hypertrophy of Texel sheep. In this breed, there is a G to A mutation in the 3′ UTR of the MSTN gene that creates a target site for miR-1 and miR-206, highly expressed in muscle tissue. As a consequence, the expression of MSTN is reduced and a muscle hypertrophy is developed [129].

The bioinformatic analysis of sRNA-seq data is relatively easier and more established than the analysis of lncRNAs, and direct cause-effect relationships with mRNA genes can be determined. This has led to a substantial growth in livestock and sheep miRNA studies, much more numerous than sheep lncRNA analyses. However, there are just 106 annotated sheep miRNAs in miRBase (v.22), so novel miRNA prediction is necessary to analyse the full miRNAome. The consequence of this is that each work has described a set of predicted transcripts that differ in stringency criteria, reference genomes used, naming and orthology determination of unannotated genes. In addition, similar to the lncRNA analyses, poor compliance with genomic data sharing principles prevent comparisons between studies.

Livestock miRNA functional genomic studies have explored their involvement in diseases, productivity and animal welfare, with more than 175 miRNA-related ovine publications in PubMed until 2020 [130]. Small ruminant studies have focused on muscle development, reproductive traits, mammary gland development, milk composition and hair-related phenotypes, but there is also extensive research on diseases, infection and immunity (reviewed in [130] and [131]). Because of the implication of these molecules in the pathogenesis of several diseases, they have been proposed as biomarkers for the management of livestock diseases. Nevertheless, this kind of application is hindered by the difficulty to find markers with enough specificity, accuracy and sensitivity ([130]).

1.5.2. Livestock lncRNAs

In line with these specific needs in the animal genetics field, an international consortium was presented, Functional Annotation of Animal Genomes (FAANG), whose aim is to produce comprehensive maps of functional elements in the genomes of domesticated animal species. In the last years, project collaborators have been producing genome-wide datasets on RNA expression, DNA methylation and chromatin modifications among others, as well as data analyses [132]. LncRNAs have been a particular focus of the FAANG initiative, with various groups reporting lncRNA transcript sets. Computational methods for lncRNA identification are in continuous evolution and, currently, there is not a consensus workflow for this task, with each group using their own lncRNA identification pipeline [132]. In addition, the RNA-seq datasets used different library preparation protocols, the main differences being between ribosome RNA depletion and poly(A) selection protocols. Library preparation has been seen to affect lncRNA identification in horse [133]. Considering experimental designs, these lncRNA catalogues have been produced with the aim of identifying lncRNAs expressed across many tissues and usually take samples from few animals, 2 to 6 from the same breed. In sheep, taking advantage of the sheep gene expression atlas dataset based on SGS RNA-seq assays [134], they identified lncRNAs across several tissues from six animals [22]. They highlighted how, because the exons of lowly abundant transcripts are subject to stochastic sampling, gene models are detected inconsistently between samples. LncRNAs have also been identified from multi-tissue RNA-seq datasets in other livestock species like cattle [135–137], horse [133], chicken [135, 137, 138], pig [135, 137] and goat [22, 135].

Apart from the purely descriptive datasets, functional genomics studies have profiled lncRNAs across different developmental stages, diseases or phenotypic traits, in order to connect lncRNAs with those specific biological conditions. LncRNA genes need to be defined before any functional analysis, and for a comprehensive analysis, it is usually done *de novo* due to the small number of annotated lncRNAs in livestock. Because the lack of consensus computational methods, different stringency criteria, different reference genomes used and the poor compliance with genomic data sharing principles, those lncRNA transcript sets are not easily comparable, if they are made publicly available at all.

In sheep, the vast majority of published functional analyses are related to phenotypic traits important for production and development of commercially interesting tissues (Table 1). The most studied field has been reproduction and fecundity, with works on female sex organs and hormone secreting organs. Works on the development of sex organs, hormone secreting organs and muscle tissue are also numerous. Another two tissues with commercial value that have generated interest are adipose tissue and hair follicles. There are also some lncRNA works on heat stress, milk properties, nutrition, photoperiod and hypoxia adaptation. Regarding disease-related functional genomics experiments, before the current work lncRNAs have only been profiled in a bacterial infection and a helminth infection.

Table 1: Published functional lncRNA transcriptomic studies in sheep.

Trait	Tissues	References
Reproduction and fertility	Female sex organs	[139–144]
	Hormone secreting organs	[145–152]
Development	Sex organs	[153–155]
	Hormone secreting organs	[156, 157]
	Muscle tissue	[158–162]
Fat	Adipose tissue	[163–167]
Wool	Hair follicle and skin	[168–174]
Heat stress	Liver	[175]
Milk production	Mammary gland	[176, 177]
Nutrition	Liver	[178, 179]
Photoperiod	Pituitary gland	[180]
Hypoxia	Lung	[181]
Infectious diseases	Spleen, lymph node	[182, 183]

Chapter 2

Thesis scope and objectives

The main objective of this work is the identification of ovine non-coding genes, concretely miRNA and lncRNA genes, that are involved in the innate and adaptive immune responses induced by vaccines, vaccine components and pathogen infections. For this purpose, sequencing datasets produced in the lab and datasets publicly available are analysed with bioinformatic tools in order to identify and profile non-coding gene expression. The working hypothesis is that, considering the known examples in humans and mice, several ovine miRNAs and lncRNAs should also be associated with immune responses. It will provide a foundation for future analyses on non-coding RNA function in non-model organisms such as livestock species. The specific objectives of this thesis are:

1. To characterise the miRNA transcriptome in sheep by identifying unannotated miRNAs, evaluating orthology relationships and profiling their expression levels.
2. To associate known and novel miRNAs with the immune response to a viral infection, specifically Visna-Maedi virus.
3. To develop reproducible bioinformatic pipelines to analyse RNA sequencing data in order to describe unannotated lncRNAs in a non-model organism.
4. To characterise the lncRNA transcriptome in sheep by identifying unannotated lncRNAs, evaluating their sequence conservation and profiling their expression levels.
5. To associate known and novel lncRNA genes with the immune response to vaccines, vaccine components and pathogen infections.

These objectives are fulfilled in different chapters. As reviewed in the introduction of this thesis (**chapter 1**), miRNAs and lncRNAs are non-coding genes that have attracted considerable attention in the last years. Nevertheless, livestock genomes remain under-annotated in terms of these genes and there is a lack of functional annotation regarding their involvement in important biological functions such as the immune response.

Objective number 1 is fulfilled in **chapter 3**, where we analyse the miRNA transcriptome in sheep lungs, and, with more depth, in **chapter 4**, where we analyse more than 20 different tissues. In the latter, we take advantage of hundreds of miRNA-seq samples deposited in public databases to uniformly analyse them, describing novel ovine miRNAs, analysing tissue-specific expression and describing ruminant-specific miRNA families.

Objective number 2 is fulfilled in **chapter 3**. Here we analyse the miRNA transcriptome in sheep lungs during the infection with a lentivirus (Visna-Maedi virus) causing chronic asymptomatic and clinical infections to identify dysregulated miRNAs.

Objectives number 3, 4 and 5 are fulfilled transversally in **chapter 5**, **chapter 6** and **chapter 7**. In **chapter 5** and **chapter 6**, we analysed RNA sequencing data from a collaborative study carried out in our lab, in which it was characterised the effect of Al hydroxide adjuvant on the immune response to vaccination during a long term experiment. Brain tissue and PBMCs were the tissues analysed. A custom lncRNA identification pipeline was developed for this works. In **chapter 7**, using an improved pipeline, we integrate hundreds of publicly available ovine RNA-seq samples of blood cells, lymphoid organs and other immune cells in order to identify unannotated lncRNAs. Integrated bioinformatic analyses identify hundreds of lncRNAs induced during infection with various pathogens and vaccination.

These findings are summarised, discussed and put in the context of current research in **chapter 8**. The bibliography cited across all chapters of the present dissertation are included in **chapter 9**.

This thesis chapters are organised by publications, each results chapter corresponding to an article published in a scientific journal or under revision.

Chapter 3

Expression analysis of lung miRNAs responding to ovine VM virus infection by RNA-seq

This chapter is based on the following publication:

Bilbao-Arribas, M., Abendaño, N., Varela-Martínez, E., Reina, R., de Andrés, D., and Jugo, BM. Expression analysis of lung miRNAs responding to ovine VM virus infection by RNA-seq. *BMC Genomics* **20**, 62 (2019).

3.1

Background

The Small Ruminant Lentiviruses (SRLVs) are in a group of RNA viruses in the lentivirus genus that infect cells of the monocyte/macrophage lineage from sheep and goats. This infection causes progressive inflammatory lesions in the lungs, brain, mammary glands and joints that are characterized by lymphoid hyperplasia, interstitial infiltration of mononuclear cells and interstitial pneumonia. Visna/Maedi disease (VM) has a great economic importance derived from decreased animal production and increased replacement rates [184]. Infection is present in most countries that raise sheep but the impact on production and animal welfare is affected by breed [185] and flock management [186].

Not every infected animal shows the disease due to the importance of the host genetic background [187]. In genetic association studies several molecules have been shown to be related to VMV infection: Toll like receptors (TLRs), antiviral proteins (APOBEC family, TRIM5alpha, tetherin), and cytokines (among others) [188, 189]. To our knowledge, microRNAs (miRNAs) have not been analyzed in relation to this viral disease.

miRNAs are a class of noncoding endogenous RNAs of approximately 22 nucleotides that regulate gene expression posttranscriptionally. By binding to mRNA molecules and with the help of the RNA-induced silencing complex (RISC), they can silence or cleave mRNA molecules [8]. They are one of the most abundant gene expression regulators and have an effect on phenotypic variations in domestic animals [127]. Several studies have identified miRNAs in various sheep breeds, although miRBase 21 includes only 106 miRNA precursors and 153 mature sequences (January 2018). Regarding tissue types that have been previously studied, most of the work has focused on muscle quantity, wool quality, fertility and fat deposition [190–193] with little attention to animal health and welfare.

Viruses exploit host gene pathways to accomplish their basic biological processes, from transcription to protein synthesis, thus, ensuring their own survival. MicroRNA levels can be altered due to the host's own immune response modulation [194]; however, viruses can also modulate the expression of host genes to avoid detection by the immune system or to modify cell survival pathways [195]. Furthermore, it has been proposed that host miRNAs can directly target RNA viruses either cleaving them or stabilizing them [196]. Another way that miRNA expression may change involves virally encoded miRNAs [197].

The aim of this study was to uncover the host mechanisms that are associated with VM disease in sheep. To this end, the cellular miRNAs differentially expressed at different stages of infection were identified, and information about involved genes, the mechanisms, and relevant pathways was inferred via bioinformatics analyses. These predictions could also contribute to uncover the roles of miRNAs in host-virus interactions.

3.2

Methods

3.2.1. Animals

Thirty Rasa Aragonesa adult (3 to 6 years) ewes were included in this study, in different stages of a natural infection of VMV. The samples were obtained from different commercial flocks in the routine of the Veterinary Faculty (University of Zaragoza) in the framework of the national research project ref. AGL2010–22341-C04–01. The complete experimental procedure was approved and licensed by the Ethical Committee of the University of Zaragoza (ref: PI09/10). Animals were euthanized by an intravenous injection of a barbiturate overdose (Dolethal®, Vetoquinol, Spain) and exsanguinated.

Animals were classified attending to their VMV infection status (seronegative or seropositive) using an Enzyme-Linked ImmunoSorbent Assay (ELISA) (ELITEST, Hyphen), and the clinical outcome (asymptomatic and diseased). For RNA-seq analysis, a total of 15 animals were included: Five animals were seronegative for VMV (seronegative group), five of the animals tested seropositive for VMV but did not show clinical symptoms (seropositive asymptomatic group) and, the remaining five animals were seropositive and had lung lesions (lesions group). For validation of the sequencing data 15 different animals were included (5 seronegative, 5 seropositive asymptomatic and 5 with pulmonary lesions) (Table 2).

3.2.2. Tissue collection, RNA extraction and small RNA sequencing

A sample from lung was aseptically taken from each animal and preserved in RNAlater solution (Ambion, Austin, TX, USA) at -80°C until used. Total RNA was isolated from lung tissue using Trizol (Invitrogen, Carlsbad, CA, USA) extraction. 60–70mg tissue samples were homogenized in 1ml of Trizol using Precellys®24 homogenizer (Bertin Technologies, Montigny le Bretonneux, France) combined with 1.4 and 2.8mm ceramic beads mix lysing tubes (Bertin Technologies). After adding chloroform, RNA was precipitated from the upper aqueous phase with isopropanol, washed with ethanol, suspended in RNase free water and stored at -80°C. RNA quantity and purity was assessed with NanoDrop 1000 Spectrophotometer (Thermo Scientific Inc., Bremen, Germany). RNA integrity and concentration was assessed with the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

The small RNA libraries were generated with Illumina's TruSeq small RNA library preparation kit following manufacturer's instructions. Sequencing was performed in CNAG-CRG core facility (Barcelona, Spain), using an Illumina HiSeq 2500 instrument. Single-end sequencing with 50bp read length was used for miRNAs.

3.2.3. Prediction of miRNAs

The quality control was performed with fastQC and the following computational pipeline was followed (Figure 5). Raw reads were analyzed with the sRNAbench web tool, which is included in the sRNAtoolbox collection of tools [87]. This program performed the preprocessing, mapping, expression profiling and novel miRNA prediction. Parameters were set to minimum read count of four, allowing one mismatch, with full read alignment and three species were selected to search for homologs: goat, cattle and mouse. After that, the prediction results of novel miRNAs were manually curated to remove repeated entries that just differed in one nucleotide and to give more updated miRNA names. Only miRNAs marked with high confidence by the program were selected for further analysis. Since the program only uses miRNAs present in miRBase, new predicted miRNAs that could had been previously described elsewhere were locally

Table 2: Samples used in RNA-seq and RT-qPCR study.

Status	RNA-seq	RT-qPCR
Pulmonary lesions	1P, 2P, 7P, 9P, 10P	P21, P22, P24, P25,P26
Seropositive asymptomatic	8P, 11P, 12P, P19, 4	1, 2, 3, 5, 6
Seronegative	7,10,11, 13,14	12, P-13, P-14, P-15, P-16

blasted against the whole RNAcentral database (http://rnacentral.org/) looking for perfect identity.

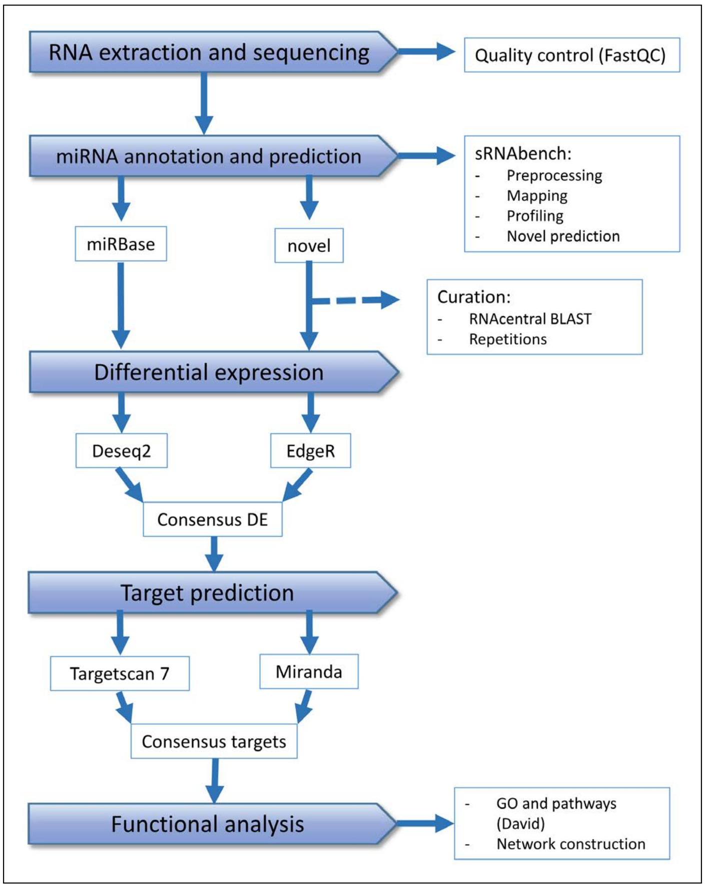


Figure 5: Computational pipeline of miRNA data analysis. The figure illustrates the four steps of the data analysis starting from the RNA extraction and sequencing: miRNA detection and prediction, differential expression, target prediction and functional analysis.

3.2.4. Differential expression

Before the differential expression analysis, the matrix of novel miRNAs was built excluding repeated miRNAs that mapped in different places, miRNAs that appeared in less than half of the samples and with counts lower than ten. This was done following common criteria in the field to perform a conservative analysis. In addition, it was performed a principal component analysis (PCA) to check the grouping of the samples with the DESeq2 Bioconductor R package (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) [113]. Three out of the 15 samples were excluded from further analysis - these outliers highly increased variability - leaving three groups with four samples each. DESeq2 results were plotted out as a heatmap with the Pheatmap function for R (https://cran.r-project.org/package=pheatmap). Differential expression analysis of both, known and novel miRNAs was performed with the sRNAde web tool included in the sRNAtoolbox collection [87]. DESeq2 and EdgeR were the methods used by the program. Three different comparisons were performed: Asymptomatic vs Seronegative, Lesions vs Seronegative and Lesions vs Asymptomatic. For a miRNA to be considered differentially expressed (DE), the adjusted p value was set to 0.05 and the absolute log2 expression fold change (FC) to one.

3.2.5. Target prediction, gene ontology and pathway analysis

Target genes for each differentially expressed miRNA were predicted using TargetScan 7 [92] and miRanda – via the miRNAconstarget tool included in sRNAtoolbox [87] – algorithms. The 3′ UTR mRNA sequences of sheep for both programs were obtained from the multi-species alignment generated from human 3′ UTRs given by the authors of TargetScan. The threshold for this program was set to absolute context++ score > 1 and the thresholds for miRanda were set to a score higher than 155 and a free energy lower than -20 kcal/mol. The consensus targets predicted by both programs were selected.

Viral-targeting miRNAs in the ovine genome were also inferred by using 11 VMV (Visna Maedi Virus) and 5 Caprine Arthritis Encephalitis Virus (CAEV) complete sequences deposited in GenBank database. The program used was standalone miRanda [198].

In order to obtain biological information from the target genes of differentially expressed miRNAs, an enrichment analysis was performed. We built three sets of genes that interacted in our predictions with any of the DE miRNAs in each comparison. Pathway and gene ontology (GO) analysis were carried out with David (https://david.ncifcrf.gov/) web tool. For pathways, KEGG pathway terms were tested and Benjamini multiple test correction value of 0.05 was applied as a threshold. We used Cytoscape version 3.5.1 [199] to build functional networks merging interactions among miRNAs, target genes and enriched pathways. This way, we were able to visualise genes in the selected pathways that are being targeted by dysregulated miRNAs.

3.2.6. RT-qPCR validation

To validate changes identified by RNA-seq experiment, the relative expression levels of 7 miRNAs (oar-miR-125b, oar-let-7b, oar-miR-181a, oar-miR-148a, oar-miR-21, oar-miR-30c, oar-miR-379-5p) selected based on significant changes seen in Lesions vs Seronegative comparison in the RNA-seq analysis, were verified by qPCR. The U6 snRNA, oar-miR-30d and oar-miR-191 were tested as internal standard controls and the last two were selected for their expression stability in our samples. Additional file 2 shows the list of the amplified miRNAs and the corresponding primer sequences. The expression study has been based on the analysis of miRNA expression with Fludigm's BioMark HD Nanofluidic qPCR System technology combined with GE 48.48 Dynamic Arrays IFC. qPCR was performed on a BioMark HD System using Master Mix SsoFastTM EvaGreen® Supermix with Low ROX (Bio-Rad Laboratories, Hercules, CA, USA). The analysis of expression with the Fluidigm Biomark HD Nanofluidic qPCR system was performed at the Gene Expression Unit of the Genomics Facility, in the General Research Services (SGIKER) of the UPV/EHU.

The software for the real-time PCR analysis and obtaining of the Ct values was Fluidigm Real-Time PCR Analysis Software [v3.1.3]. PCR efficiency calculation and correction, reference miRNA stability analysis and normalization was done with GenEx software of MultiD [v5.4]. Most miRNAs showed high amplification efficiencies (94.43–99.65%). The stability of candidate reference miRNAs was analyzed using both NormFinder [200][21] and GeNorm [201] algorithms integrated in GenEx. The two most stable miRNAs were oar-miR-30d and oar-miR-191 so normalization was performed using these two reference miRNAs. Normal distribution was checked using the Shapiro-Wilk test in the IBM SPSS statistical package [v24]. Comparison and correlation between the RNA-seq and qPCR results was performed using T-test and Pearson's correlation, respectively. In all analyses, differences were considered significant when p values were < 0.05.

3.3

Results

3.3.1. Small RNA sequencing and miRNA prediction

In the present study, the small RNAs from lung tissue of sheep with and without VMV infection were sequenced. The raw reads were high quality – only approximately 2% had Q scores below 30 – and the numbers of reads ranged from 22 to 8 million, with an average of 15 million reads. The raw reads were analyzed by sRNAbench for miRNA prediction, trimmed the adapters in around the 95% of the reads in all the samples, and 85% of the preprocessed reads were successfully mapped to the sheep genome. The read-length distribution showed a clear peak between 21 and 23 nucleotides in all of the samples, where most of the reads were located.

Out of the mapping, the program could annotate 86 known sheep miRNAs from miRBase. All of the other reads that mapped to the genome, but that did not coincide with a miRBase miRNA were subjected to novel discovery tests, from which several new miRNAs arose. Some of these new miRNAs were apparently completely novel molecules, and others were found to be conserved in other species. After cleaning the output sequences and aligning them with RNAcentral, it was found that some were already annotated in sheep and that others had homologs in other species. In total, 86 known miRNAs from miRBase, 68 known sheep miRNAs from other databases, and 58 miRNAs shown for the first time in sheep were found (Figure 6b). Twelve miRNAs out of these 58 could not be considered ovine homologs of previously described miRNAs and were considered novel. The novel miRNAs were named sequentially, but they were given the name of a homolog if one existed. Regarding the expression levels, some miRNAs were much more abundant than others (Figure 6a): the 13% most abundant miRNAs were above 10,000 counts, while the 29% least abundant miRNAs had fewer than five average counts. Furthermore, the miRNAs classified as novel or conserved had particularly low abundance, with only few of them having more than 1000 counts.

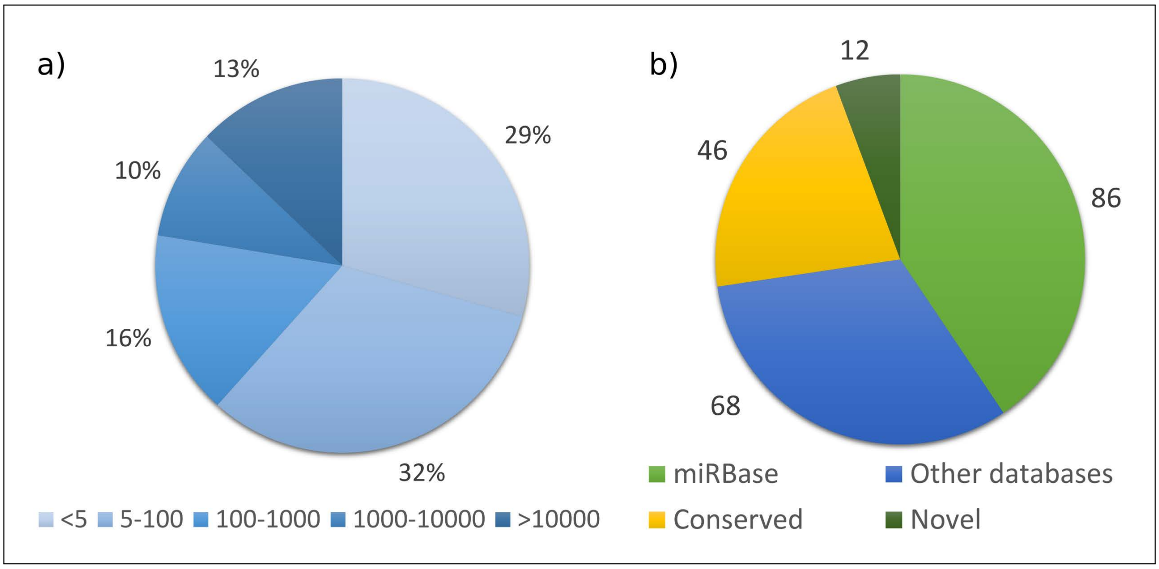


Figure 6: Statistics of RNA-seq and prediction data. a) Average counts distribution of all the miRNAs detected and predicted. b) Distribution of miRNAs according to previous knowledge about them

3.3.2. Differentially expressed miRNAs

We made pairwise comparisons among the three sample groups. Overall, the differential expression levels, as well as the PCA, pointed out that the biggest differences were between seronegative sheep and the other two seropositive groups (asymptomatic animals and animals with Lesions). Clustering of differentially expressed (DE) miRNAs detected by either of the two programs clearly grouped the seronegative samples, but failed to distinguish the other two groups, similar to the outcome of the PCA. Seropositive asymptomatic animals and animals with developed clinical symptoms seemed quite similar in terms of miRNA expression (Figure 7). By merging the results of the EdgeR and DESeq2 analyses, 34 DE miRNAs were identified between clinically affected and seronegative sheep, of which 23 were upregulated and 11 downregulated. There were also 9 upregulated and one downregulated miRNAs when comparing samples from seropositive asymptomatic animals with samples from seronegative animals, and only three miRNAs were differentially expressed between animals with clinical symptoms and seropositive asymptomatic animals. Some novel ovine miRNAs with homologs in other mammals, namely, chi-miR-30f-5p, chi-miR-449a-5p, mmu-let-7e-3p, mmu-miR-144-3p, bta-miR-142-5p, chi-mir-92a-3p, ssc-mir-7134-3p, ssc-mir-7134-5p and mmu-miR-98-5p, from goat (chi), mouse (mmu), pig (ssc) and cattle (bta), showed differences in VMV infected animals. Completely novel miRNAs did not differ significantly in their expression likely due to their low expression levels, which were sometimes even below the applied count threshold.

Among the most abundantly expressed DE miRNAs, some showed relevant increases or reductions in expression (Figure 8): oar-miR-21 was, by far, the most abundant DE miRNA, since its expression was elevated 4.3 times in seropositive asymptomatic animals and 12 times in diseased animals, with average total counts of around two million. Other highly expressed DE miRNAs, such as oar-miR-148a and oar-let-7f showed significant increases, with absolute fold changes of 3 and 2.2, respectively, in infected animals compared with seronegative animals. Furthermore, miRNAs such as oar-let-7b, oar-miR-99a and oar-miR-125b, showed reduced expression in infected sheep (Figure 8).

3.3.3. Validation of differential miRNA expression

To validate the miRNA-seq data, seven miRNAs (oar-miR-125b, oar-let-7b, oar-miR-181a, oar-miR-148a, oar-miR-21, oar-miR-30c, and oar-miR-379-5p) were verified using the Fluidigm Biomark HD Nanofluidic qPCR system. The log2FC in the miRNA expression levels calculated by qPCR in the Lesions group relative to the Seronegative group are shown in Fig. 5. The validation results confirmed the upregulated expression of 3 miRNAs (oar-miR-148a, oar-miR-21, oar-miR-379-5p) and the downregulated expression of 4 miRNAs (oar-miR-125b, oar-let-7b, oar-miR-181a, and oar-miR-30c), although only two were sta-

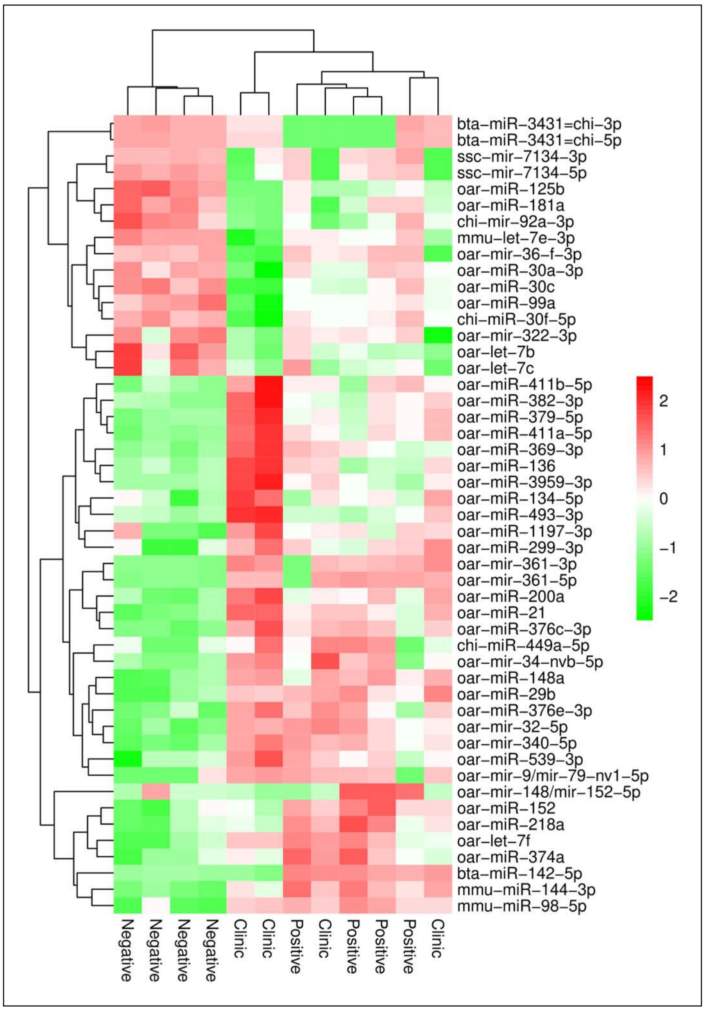


Figure 7: Hierarchical clustering heatmap. Clustering of all the DE miRNAs detected by any of both programs (DESeq2 or EdgeR) and samples. Colours and intensities depend on expression level. Green indicates gene down-regulation and red up-regulation.

tistically significant: oar-miR-21 (p=0.003) and oar-miR-30c (p=0.004). There were no significant differences in the FC data obtained from the RNA-seq and the Fluidigm Biomark HD Nanofluidic qPCR system (p=0.656) showing a high degree of concordance, with a correlation coefficient of 0.982 (p=0.000).

3.3.4. Functional analysis of dysregulated miRNAs

In this study, the targets of the DE miRNAs were predicted using the TargetScan and Miranda algorithms. TargetScan predicted a total of 1.9 million interactions for all of the identified miRNAs, and this number was reduced to 124,614 after applying the cut-off value. Miranda predicted 911,069 target sites for the same set of miRNAs and application of the threshold settings reduced this number to 41,871 targets. Next, we performed

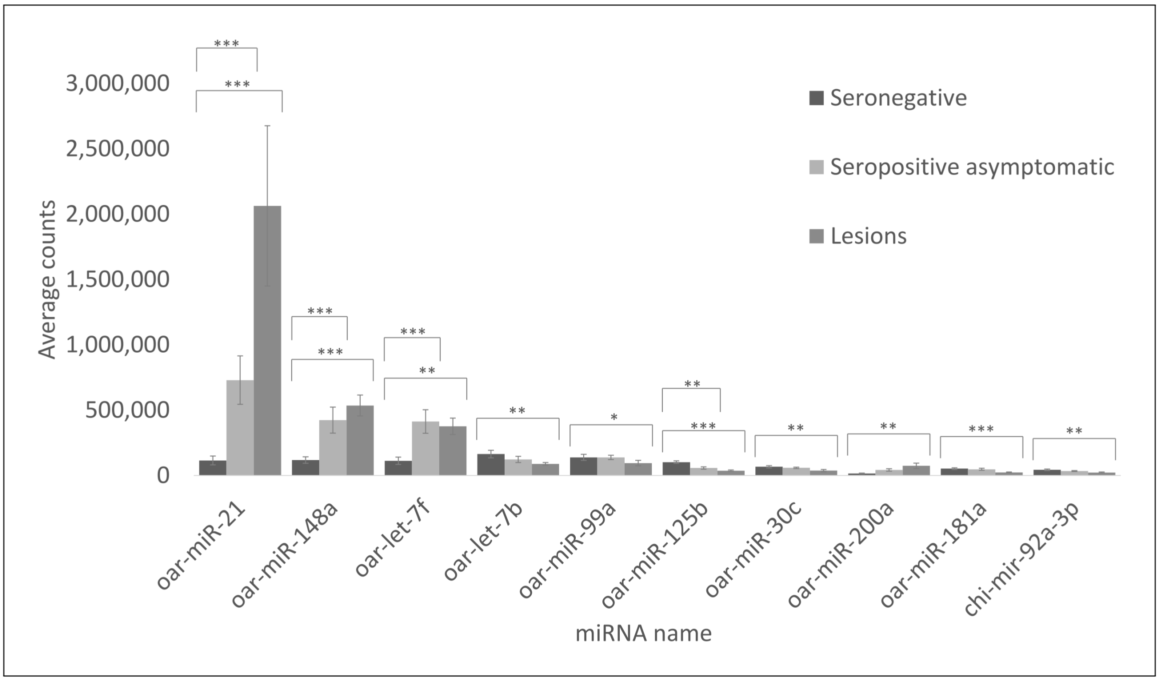


Figure 8: Expression of most abundant miRNAs. Average counts of the most expressed DE miRNAs in the three phases of disease progression. Asterisks indicate significance level between two groups (*P<0.05, **P<0.01, ***P<0.001)

an intersection analysis to enhance the confidence of the predictions, and this process reduced the number of interactions to 12,280, with 6426 unique genes. An average of 35 interactions was observed for each of the 349 mature miRNAs analyzed. Out of the collection of the predicted targets, we retrieved three sets of genes (one for each comparison) with 1736, 1135 and 190 genes each. These gene sets were then used in enrichment analyses.

The GO enrichment analysis did not identify any significantly enriched terms using the multiple testing correction, whereas some pathways were actually overrepresented, such as, signalling pathways (e.g. PI3K-Akt, AMPK and ErbB), or other terms such as ECM-receptor interaction and pathways in cancer (Table 3). The PI3K-Akt signalling pathway had the most genes involved in both comparisons – 51 and 40, respectively – and it was the most statistically significant term (corrected P values of 2.51E-04 and 0.004). The comparisons between the seropositive and seronegative sheep were the only ones yielding results, while there were no enriched terms in the comparison between the seropositive groups, based on the corrected p values.

Interaction maps incorporating the miRNAs and their targets and the pathways information were produced in an attempt to unveil how the differences in miRNA expression could affect these pathways in seropositive asymptomatic compared to seronegative animals (Figure 9) and in diseased animals compared to seronegative animals (Figure 10). Key regulators in the PI3K-Akt pathway, such as PTEN, and related transcription factors

Table 3: Enrichment analysis of pathways between both seropositive groups and the seronegative group. Significant entries with Benjamini score equal or smaller than 0.05 are shown.

Pathway	Seropositive asymptomatic-Seronegative		Lesions-Seronegative	
	Fold enrichment	**FDR**	**Fold enrichment**	**FDR**
oas04151:PI3K-Akt signaling pathway	2.327	2,51E-04	1.868	0.004
oas04152:AMPK signaling pathway	2.831	0.024	2.411	0.022
oas05202:Transcriptional misregulation in cancer	–	–	2.111	0.024
oas05161:Hepatitis B	2.542	0.027	2.134	0.047
oas04012:ErbB signaling pathway	3.296	0.034	2.519	0.048
oas05200:Pathways in cancer	–	–	1.595	0.049
oas04512:ECM-receptor interaction	3.186	0.024	2.435	0.050
oas04510:Focal adhesion	2.232	0.029	–	–
oas05215:Prostate cancer	3.296	0.034	–	–
oas04360:Axon guidance	2.643	0.034	–	–
oas04014:Ras signaling pathway	2.112	0.038	–	–
oas05206:MicroRNAs in cancer	2.190	0.050	–	–

such as FOXO3 and CREB1, appear to be targeted by dysregulated miRNAs identified between the seropositive groups and the seronegative group. Most of the miRNAs target no more than three genes in these pathways, except for oar-miR-143 and oar-mir-361-3p, which target several genes based on our predictions.

3.3.5. Virus-miRNA interactions

Regarding the highly expressed DE miRNAs, two significantly strong interactions were found between the miRNAs and the SRLV genome. The upregulated miRNA oar-miR-200a was predicted to target nine out the eleven tested sequences at nucleotides 1671 to 1689 with respect to the VMV reference genome sequence (GenBank accession number L06906.1), with a score of 155 and a folding energy of -16.1 kcal/mol. The downregulated miRNA oar-miR-99a was predicted to target nine sequences around nucleotides 5383 to 5402 with a score of 150 and a folding energy of -25.54 kcal/mol. These predicted interactions are in the "gag" and "vif" genes, respectively. These targeted sequences are all from the genotype A of SRLV. On the other hand, oar-miR-99a may also target CAEV at nucleotides 2194 to 2212 – in the "pol" gene – with respect to the CAEV reference genome (GenBank accession number M33677.1) with a score of 160 and a folding energy of -23.83 kcal/mol.

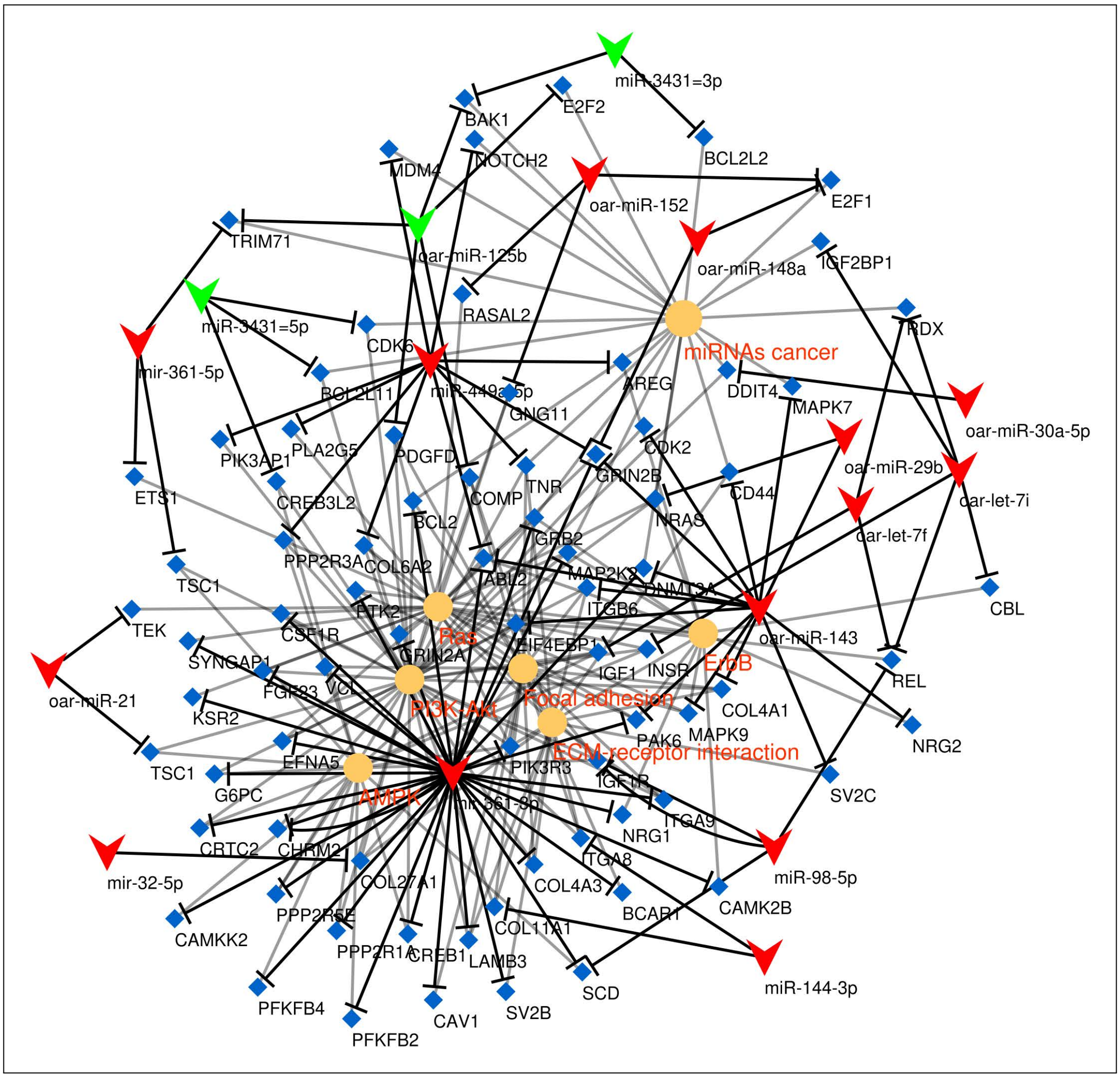


Figure 9: Functional network of the comparison between seropositive asymptomatic and seronegative sheep. It illustrates the predicted interactions of DE miRNAs with their targets and the pathways those target genes are part of. Upregulated miRNAs are coloured in red and downregulated ones in green, pathway names in orange and genes in blue.

3.4

Discussion

In this work, we used NGS techniques to analyze the expression pattern of miRNAs in seronegative sheep and in SRLV seropositive but asymptomatic animals and in diseased

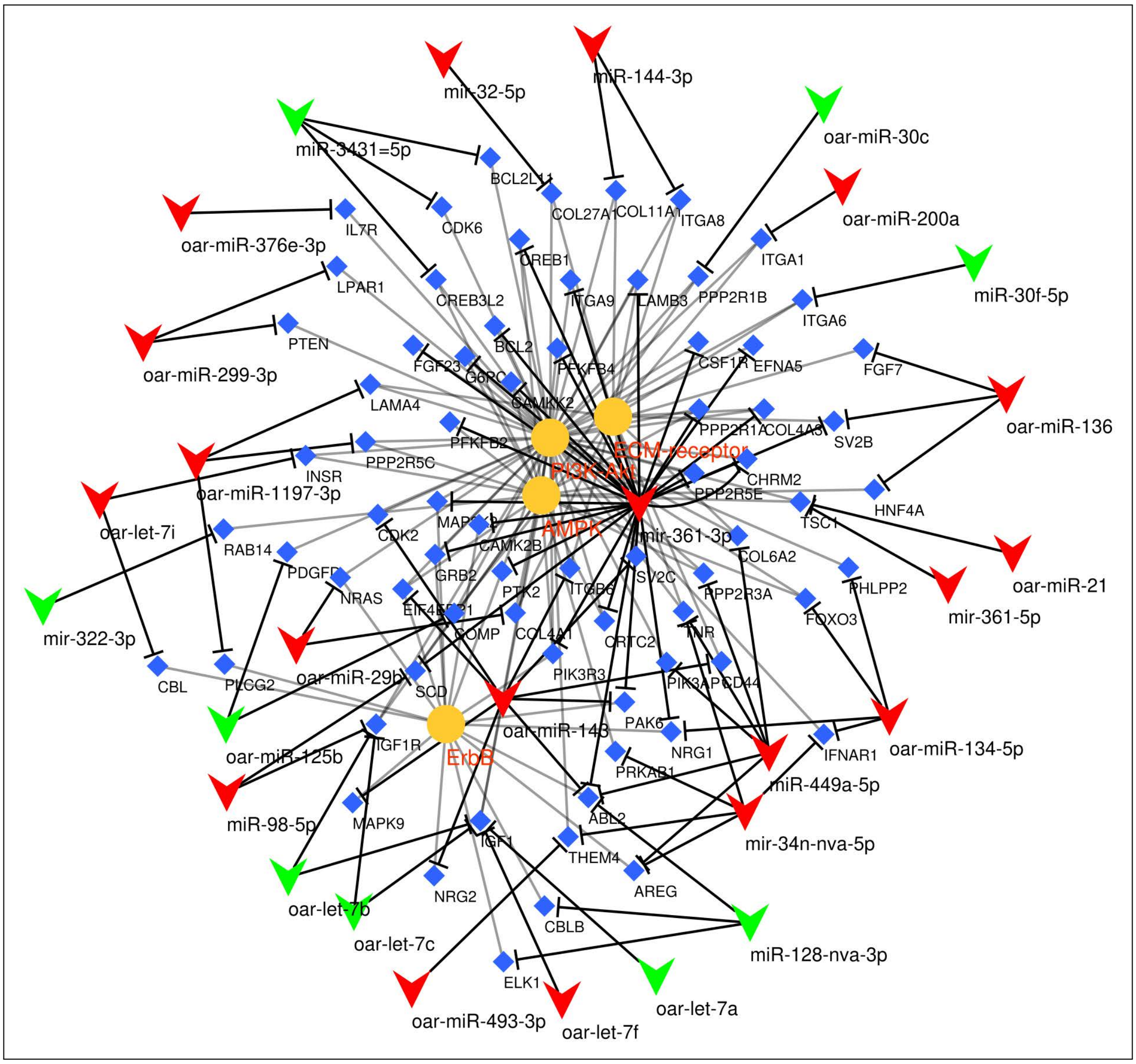


Figure 10: Functional network of the comparison between diseased and seronegative sheep. It illustrates the predicted interactions of DE miRNAs with their targets and the pathways those target genes are part of. Upregulated miRNAs are coloured in red and downregulated ones in green, pathway names in orange and genes in blue.

animals. We then made predictions of the possible regulatory functions of the miRNAs. Since we used tissue samples from naturally infected animals for the experiments, the data reflect the actual miRNA transcriptome in the lung tissue of SRLV-infected animals. Host-virus interactions modify several biological processes as a consequence of the ability of the viruses to employ the host machinery to complete their replication cycle, and of the host's attempts to deal with the infection. These changes can be observed at the miRNA expression level since miRNAs can control different pathways; therefore, understanding changes in miRNA expression could be crucial for understanding the disease.

The enriched pathways identified in this study suggest an increase in cell proliferation-related signaling. The PI3K-Akt pathway is a key pathway involved in growth and proliferation, and it has been extensively studied in the context of proliferative diseases such as cancer; furthermore, it seems to be influenced by a miRNA regulatory network as an added layer of modulation [202]. Furthermore, viruses can hijack this pathway for enhanced replication, as has been reported in several cases [203]. For instance, Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) modulates PI3K-Akt signalling via FoxO1 and Bad [204] and influenza A codes for the NS1 protein which directly interacts with the PI3K regulatory subunit p85 [205]. DE miRNAs were predicted to target very important factors in this pathway including PTEN, PI3K, FOXO3, the BCL2 family, CREB, GRB2, growth factors (FGF23) and cytokine receptors (IFNAR1). Other enriched pathways in our set of target genes were the AMPK signalling pathway, which is a regulator of cellular homeostasis and is linked to PI3K-Akt pathway, and the ErbB pathway, which is related to signal transduction involving growth factors.

Although miRNAs are fine tuners of gene expression that can act at low concentrations, the appearance of highly expressed miRNAs may be very relevant and could indicate strong modulation. Normally, a few miRNAs comprise the majority of the miRNAome, and many others are present at low concentrations. In our experiments, oar-miR-21 expression showed an interesting behaviour, as its expression is remarkably high in both seropositive groups, with its highest expression level in diseased animals. miR-21 is a fairly well-studied miRNA, and was one of the first miRNAs identified as an oncogene; it has been seen to be upregulated in several conditions including tumours [206] and viral infections. In the case of RNA viral diseases, miR-21 is upregulated by hepatitis C virus (HCV), which leads to a decreased IFN response in human cell lines [207], during dengue virus infection in human cancer cells, which promotes viral replication [208] and in HIV and in HIV-related pulmonary arterial hypertension in human plasma [209]. Furthermore, Epstein-Barr virus (EBV) induces miR-21 expression in B cells, which promotes tumorigenesis by activating the PI3K-Akt pathway, causing FOXO3a to stop repressing miR-21 [210, 211], findings that are in agreement with our current results.

The respiratory form of SRLV infection exhibits some typical histopathological lesions characterized by lymphocytic infiltration and inflammation, M2-polarized macrophages, interstitial pneumonia, lung fibrosis and decreased gas exchange [212, 213]. However, the mechanisms of this pathogenesis, which are likely immunomediated [214], are not fully characterized. There were no major differences between the infected asymptomatic animals and the sheep that did show lesions, indicating that the miRNA levels mostly change after infection, rather than when symptoms appear. It seems that most of the transcriptional changes occur in the early stages of infection and that the differences between the asymptomatic-seronegative and the lesions-seronegative comparisons could be due to disease progression and appearance of clinical symptoms.

Interestingly, these kinds of lesions could be related to some of the DE miRNAs and with the pathways regulated by them. In an artificially induced lung fibrosis in mice, miR-21 mediates the activation of pulmonary fibroblasts [215]. Furthermore, miR-21 has been recently proposed as an indicator of disease progression and potential treatment target in another mouse model [216]. MiR-21 could control pathways such as the TGF-β1 signaling pathway by targeting SMAD7 and SPRY1 or by inhibiting PTEN, which is a known negative regulator of lung fibrosis [217]. The remodelling of lung tissues caused by fibrosis related hypoxia has also been linked with miR-21 [218]. Importantly, PTEN has a crucial role in controlling the PI3K-Akt pathway, and its interaction with miR-21 has been experimentally validated several times in human and in mice [219]. The upregulated miR-148a also targets PTEN, as well as GADD45A and BCL2L11, and it accelerates the development of autoimmunity [220].

Another miRNA, miR-99a, which was downregulated in the diseased sheep, appears to target AKT1 [221] (which has an important role in the PI3K-Akt pathway) and inhibits cancer cell proliferation by targeting mTOR [222]. Thus, its downregulation in the animals with lesions should increase AKT1 and mTOR expression, stimulating proliferative signal. In our analysis, inflammation-related interleukin 13 (IL-13) was predicted as a target of miR-98-5p and let-7 family miRNAs, and it is noteworthy that previous experimental observations have shown that let-7 miRNAs can modulate inflammation through inhibition of IL-13 [223]. During bluetongue virus infection in sheep testicular cells, while IL-13 and let-7f were downregulated, let-7d was upregulated and PI3K-Akt pathway was overrepresented in the enrichment test of the DE genes [224].

The relationship between the dysregulation of some miRNAs and VM disease could be a direct consequence of virus modulation or a side effect of the host defense mechanisms. In the case of miR-21, it has been proposed as a key switch in the inflammatory response [219]. Clinical lesions observed could be a consequence of excessive cell survival signalling after the initial pro-inflammatory immune response. On the other hand, the virus itself may modulate miRNA expression, as it does in EBV and HCV infections [207, 225], during which the viruses induce miR-21 expression to promote their replication by enhancing the growth and survival of the infected cells, thus modulating the response in favour of the virus. Furthermore, PRRSV downregulated miR-125b to negatively regulate NF-κB signaling as a survival strategy [226].

Direct targeting of viruses remains controversial not only because of viral genome structure and rapid evolution but also because the normal concentrations of miRNAs are too low for efficient silencing [227]. Only some highly expressed DE miRNAs have been analyzed to determine if they could potentially silence some viral RNA. Interestingly, there were some predicted miRNA target sites in the SRLV genome, including one for oar-miR-200a. oar-miR-200a was upregulated in the lesions-seronegative comparison and could actively target the viral gag gene in the A genotype. Functional experiments are necessary to uncover the antiviral functions of these candidate miRNAs.

In this work, we performed for the first time a miRNA profiling in sheep responding to SRLV infection. Twelve completely novel miRNA molecules and more than 40 others were found for the first time in sheep. MiRNAs differentially regulated between seronegative and infected sheep, such as oar-miR-21, oar-miR-148a or oar-let-7f may have potential implications for the host-virus interaction. The miRNAs were predicted to target important genes involved in apoptosis, proliferation and growth, e.g., the PI3K-Akt and AMPK pathways. The role of oar-miR-21 as a regulator of inflammation and proliferation appeared as a possible cause for the lesions caused in sheep lungs, and this miRNA could be an indicator of the severity of the lung lesions or may be useful as a putative target for therapeutic intervention.

Chapter 4

The sheep miRNAome: characterization and distribution of miRNAs in 21 tissues

This chapter is based on the following publication:

Bilbao-Arribas, M.*, Guisasola-Serrano, A.*, Varela-Martínez, E., and Jugo, BM. The sheep miRNAome: characterization and distribution of miRNAs in 21 tissues. (under revision)

* Co-authorship

4.1

Background

MicroRNAs (miRNAs) are evolutionarily conserved small non-coding RNAs that regulate gene expression by targeting mRNAs and provoking destabilization or translational repression [12]. Moreover, they have been associated with many diseases and are used as markers for molecular diagnosis in humans [14]. In livestock species, miRNAs also show great potential as biomarkers for animal health and product quality, or as biomarkers for the selection and improvement of phenotypes of commercial interest in breeding programs [131, 228].

Livestock genomes remain under-annotated in terms of miRNAs compared with other model organisms such as human or mouse but the information in sheep is scarce. In livestock species such as goat, cattle, horse or pig hundreds of miRNAs have been described, but in sheep there are only 106 miRNA genes in miRBase database [96], there are 355 miRNA genes annotated in the latest Ensembl (v.104) annotation and sheep is not among the supported species in the latest MirGeneDB (2.1) update [97]. The RumimiR database [98] is the most comprehensive repository. It stores ruminant miRNA sequences from the literature as they were published and is a useful resource to find out which sequences have already been detected. Considering the important regulatory roles miRNAs have, their proper characterization in sheep is of prime importance.

Integrated miRNA expression profiling across tissues has been performed to identify tissue specific miRNAs in other livestock species such as horse or cattle [229, 230]. Different sheep tissues, such as ovaries, heart, lungs or intestines have been analyzed by small RNA sequencing in several functional experiments [192, 231, 232]. Recently, our group has also analyzed the miRNA expression in experiments related to the immune response in infection diseases and vaccination experiments in different tissues such as lungs [233], cerebral cortex [234], peripheral blood mononuclear cells (PBMCs) [235] and spleen (Varela-Martínez *et al.*, under revision).

Secondary analysis of genomic data deposited in public databases represents an opportunity for scientific questions that could not be possible with individual datasets [111]. In this work, we collected raw miRNA-seq samples deposited in public databases by multiple projects, comprising a wide range of sheep tissues, and analyzed them in a uniform way. Thus, the main objective of this work was to characterize the sheep miRNA microRNAome by predicting unannotated miRNAs and analyzing their expression profile across different tissues in an integrated manner. We focused on the identification of tissue-specific miRNAs and the dissection of miRNA expression distribution, as dominant miRNAs can constitute a significant fraction of the expressed miRNAome, in clear contrast to protein-coding genes [236].

4.2

Methods

4.2.1. Sample selection and data preprocessing

Firstly, searches in NCBI PubMed and SRA databases were performed using “mirna” and “sheep” keywords. 226 articles were recovered from PubMed (search performed on 20/07/2021), and all the projects including the two keywords in the SRA database were reviewed. Two conditions were established for dataset selection: high-throughput small RNA sequencing should have been performed in an Illumina platform and the raw sequencing files should have been uploaded to a public repository with clear metadata. Three conditions were set up for sample selection: samples without any experimental treatment (e.g. an infection), with at least two biological replicates in the same study and from animals older than 6 months were selected.

All samples were downloaded with the SRA toolkit version 2.10.8. Quality control was performed using Fastqc version 0.11.5. Adapters, low-quality sequences and small sequences (> 16 bp) were removed using Trimmomatic version 0.39 [237].

Some tissues were grouped into more general groups: ovary and corpus luteum samples were grouped as “Ovary”; cerebral cortex and hypothalamus samples as “Brain”; colon and intestine samples as “Intestines” and omasum and rumen samples as “Stomach”.

4.2.2. Characterization and quantification of miRNAs

Preprocessed miRNA reads were mapped against the Ovis aries reference genome Oar_rambouillet_v1.0 with the mapper.pl script from miRDeep2 v.0.1.2 [238], which internally uses bowtie [239]. During mapping, we allowed one mismatch in the seed sequence, defined as the first 18 nucleotides, and removed reads mapped to more than five locations. Then, the miRDeep2 core program was used in order to predict bona fide miRNAs. We used miRNAs from sheep and other species from miRBase release 22.1 [96] to guide the search for unannotated miRNAs. Predicted sequences with 5 or higher miRDeep2 scores were kept. All the resulting sequences were blasted against the sheep RNA sequences in RNAcentral database with standalone BLAST+ v.2.9.0 [240] in order to filter out potentially hairpin-forming RNA classes such as tRNAs and other small RNAs. A single miRNA was retained when two precursor sequences overlapped more than 16 base pairs in their location. Known miRBase miRNAs mapping to a different location were treated as unannotated copies.

Unannotated miRNAs were named based on sequence similarity with other species. Precursor sequences were blasted against individual datasets of miRBase precursor sequences of goat, cattle, horse, pig and human miRNAs with standalone BLAST+ v.2.9.0 [240]. Alignments with Q value < 0.01 and query coverage > 80 were kept, and the one with highest identity was selected in each species. When a sequence was present in more than one species, we named the unannotated miRNA with the name of the evolutionarily closest species. All miRNA loci were intersected with the Ensembl v.104 miRNA genes with bedtools v.2.26.0 [241]. All miRNA loci were grouped into clusters with a minimum genomic distance of 10kb with bedtools v.2.26.0 [241].

4.2.3. Analysis of miRNA expression

Expression quantification of known and predicted miRNAs was performed with the quantifier.pl script from miRDeep2. We conserved the expression of a single representative miRNA in the case of identical miRNAs located in different genomic loci. For that, we clustered the mature miRNA sequences into groups with complete identity using cd-hit-est from the CD-HIT suite [242]. miRNAs that were not expressed in at least one tissue with 10 reads on average were removed.

Read counts were normalized to counts per million (CPM), using the cpm function of EdgeR v.3.26.8 R package [112]. The expression matrix was log10-transformed after adding 0.1 to all the values and the normalized expression matrix was used for visualization with the t-distributed stochastic neighbor embedding (t-SNE) method of dimension reduction with Rtsne v.0.15 R package. Correlation between samples was analyzed using the Pearson correlation coefficient and was visualized as a heatmap using the pheatmap v.1.0.12 R package. The miRNA expression distribution was analyzed by averaging the CPM values in each tissue.

4.2.4. Analysis of tissue specificity

Two procedures were applied to analyze tissue specific miRNAs. Firstly, it was verified in which tissue combinations were expressed the miRNAs using a minimum of 5 CPM to consider them expressed, and it was visualized with UpSetR v1.4.0 R package [243]. Secondly, a previously described tissue specificity index (TSI) called tau (τ) was used [244]. The range of the TSI values for a miRNA is between 0 and 1, where 1 represents a miRNA expressed in a single tissue and 0 represents a miRNA expressed in all tissues. miRNAs with 0.9 or greater TSI values were considered tissue-specific and those who had a 0.25 or smaller TSI value were considered housekeeping miRNAs.

4.2.5. Target prediction

The target genes of the novel miRNAs were predicted using the standalone version of TargetScan v.7.0 [92] and the UTR sequences defined in the ovine Ensembl gene annotation (release 107). Predicted miRNA-gene interactions were filtered by a context ++ score percentile > 95 in order to retain the most confident targets. Gene set enrichment analysis of targets gene sets was done with gprofiler2 v.0.2.1 R package [245]. Benjamini-Hochberg FDR correction was applied to the p-values and the threshold was set to 0.01.

4.3

Results

4.3.1. Data retrieval

We selected 20 ovine small RNA sequencing datasets available through NCBI SRA. Among all the datasets, four had been produced in our lab and another one was produced by The Ovine FAANG Project. In total, the selected dataset comprises 172 samples and 21 tissues, with ovary being the tissue with the largest number of samples (Table 4). Information of all samples selected for the study is provided as supplementary data (Supplementary Table S1). From the 21 tissues, 3 corresponded to the female sheep reproductive system (endometrium, ovary and oviduct), 2 to the male sheep reproductive system (testis and epididymis), 4 to the digestive system (stomach, intestine, liver and gallbladder), 3 to the immune system (spleen, PBMC and lymph node) and 2 to the renal system (kidney and ureter). Four of the samples were obtained from unpublished spleen miRNA-seq experiments that were produced by our group.

Table 4: Summary of samples and publications retrieved for this work. [a]Permission from the author. [b]The Ovine FAANG Project. [c]In review

BioProject accession	Tissues (number of samples)	References
PRJNA451237	Ovary (6), endometrium (2)	[246]
PRJNA354833	Adipose tissue (2)	[247]
PRJEB22101	Ovary (10)	[192]
PRJEB32852	Corpus luteum (10)	[248]
PRJEB32852	Endometrium (10)	[249]
PRJNA392421	Intestine (3)	[231]
PRJEB20781	Lymph node (6)	Unpublished[a]
PRJNA505702	Ovary (6)	[250]
PRJNA474913	Lung (5)	[233]
PRJNA532808	Hypothalamus (12)	[145]
PRJNA511987	Heart (6), Muscle (6), Lung (6), Kidney (4), Liver (5), Spleen (6)	[232]
PRJNA414087	Gallbladder (2), Heart (2), Skin (2), Muscle (2), Lymph node (2), Colon (2), Omasum (2), Rumen (2), Oviduct (2), Ureter (2)	[251][b]
PRJNA454385	PBMC (6)	[235]
PRJNA528259	Cerebral cortex (5)	[234]
PRJNA748757	Spleen (4)	Varela-Martínez *et al.*[c]
PRJNA638028	Ovary (3)	[155]
PRJNA613135	Testis (8)	[252]
PRJNA608075	Mammary gland (6)	[253]
PRJNA694531	Epididymis (9)	[254]
PRJNA607580	Mammary gland (6)	[255]

4.3.2. Characterization of known and unannotated miRNAs

Before data preprocessing, samples had an average of 15.37 ± 0.53 million reads and 14.07 ± 0.5 million reads remained after quality filtering and adapter trimming. Samples with bad quality, very low sequencing depth or read length distribution not centered around 20-22 base pairs were discarded from the analysis. An average of 82.85 ± 0.8% of the filtered reads were unambiguously mapped against the sheep genome.

1047 sequences were selected as bona fide miRNAs, corresponding to the sequences with a probability of being a true miRNA of 0.83 ± 0.01 according to miRDeep2, and after removing other RNA classes and overlapping predictions. Despite some miRNAs being already annotated in miRBase, eleven of them did not reach the minimum established miRDeep2 score to be considered a true miRNA. Comparing with the Ensembl miRNA gene set, we detected 284 out of the 355 miRNA precursors (80%), of which 97 were annotated in miRBase and 187 were not. Sequences and coordinates of all detected miRNA loci are provided as supplementary data (Supplementary Table S2).

4.3.3. Sequence conservation and miRNA clusters

Regarding the 1047 unannotated miRNAs, they showed differing levels of conservation. 455 (43%) were found at least in another species using a stringent approach. Due to some miRNAs being located at more than a single genomic loci, they were homologs of 428 miRNAs in other species and most unannotated miRNAs were named based on a goat or cow homolog (Figure 11). 161 miRNAs (35% of the conserved miRNAs) were found in all five species and 432 miRNAs (95%) had homologs in cattle. The reason for finding so many cattle homologs is the higher number of annotated cattle miRNAs, comparable to that of humans, and especially the high number of ruminant specific miRNAs. Strikingly, we found 146 precursors of the ruminant specific family of mir-2284 and mir-2285 miRNAs, which have 206 annotated precursors in cattle and 5 precursors in goat, but none in sheep.

We identified 95 miRNA clusters, miRNA groups closely located in the genome, five of them with at least 5 miRNAs. The biggest one was the known mammal miR-379/miR-656 cluster located in chromosome 18. It harbors 48 miRNAs included in this study, mainly from miRBase, but three novel conserved loci were also found in this location. Interestingly, we found two novel clusters located on chromosome X, with 8 members each, exclusively made up of unannotated conserved miRNAs. One of them contains miRNAs homologs to the cattle bta-mir-6526 and horse eca-mir-8908 families, while the other contains miRNAs homologs to the cattle and goat miRNA families mir-424, mir-450, mir-503 and mir-542. Many of the small clusters (< 5 miRNAs) were comprised of multicopy miRNAs, very similar precursors that produce the same mature sequence.

4.3.4. Exploratory analysis of miRNA expression

In total, 1014 miRNAs were quantified, 98 of which were included in miRBase and 916 were previously unannotated miRNAs. These miRNAs represent 1985 unique mature miRNAs as different pre-miRNAs can produce the same mature miRNA product. Re-

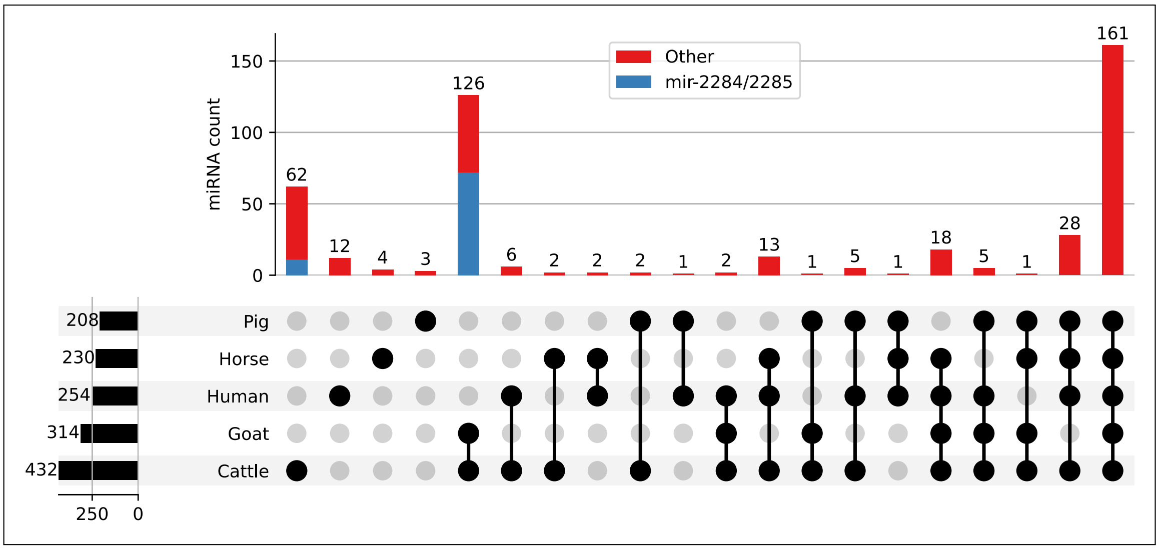


Figure 11: Comparison between species of miRNA precursor sequences. All the novel precursors found conserved between sheep and at least another species were visualized as an upset plot. Horizontal bars represent the number of sequences with a match in each species. Vertical bars represent the number of sequences common to each intersection. The family of mir-2284/mir-2285 miRNAs was given a different color in the intersection bars to highlight the ruminant specificity.

moving lowly and inconsistently expressed miRNAs, a dataset of 1082 mature miRNAs for the analysis of expression was obtained (Supplementary Table S3). Fourteen % (147) of the expressed mature miRNAs were miRBase miRNAs, 53% (574) were miRNAs with homologs in other species, and 33% (361) were unannotated novel miRNAs. MiRBase miRNAs consistently showed higher expression levels than unannotated miRNAs, regardless of homology (Figure 12A). The miRNAs included in miRBase had a mean expression of 8369.21 CPM, while conserved miRNAs and novel miRNAs had a mean expression of 525.66 CPM and 6.36 CPM, respectively. Each tissue expressed nearly 500 miRNAs on average (468.7 ± 25.6 miRNAs), with the highest number in the brain (695) and the lowest number in adipose tissue (173).

The first exploratory analysis showed that tissues were generally well grouped according with their tissue type (Figure 12B). Samples of the same tissue but from different works often did not group together, which highlights how other variables can affect miRNA expression. Brain samples were clustered into two closely related groups, based on their cerebral cortex and hypothalamus origin. We can observe a similar pattern in the sample correlation matrix (Figure 12C). Brain, male reproductive tissues and PBMCs showed the most distinct expression patterns.

The distribution of miRNA expression was skewed towards a handful of highly expressed molecules (Figure 13A). In some tissues, a single miRNA took more than half of the total expression (adipose tissue, intestine, stomach and ureter). In all tissues, the

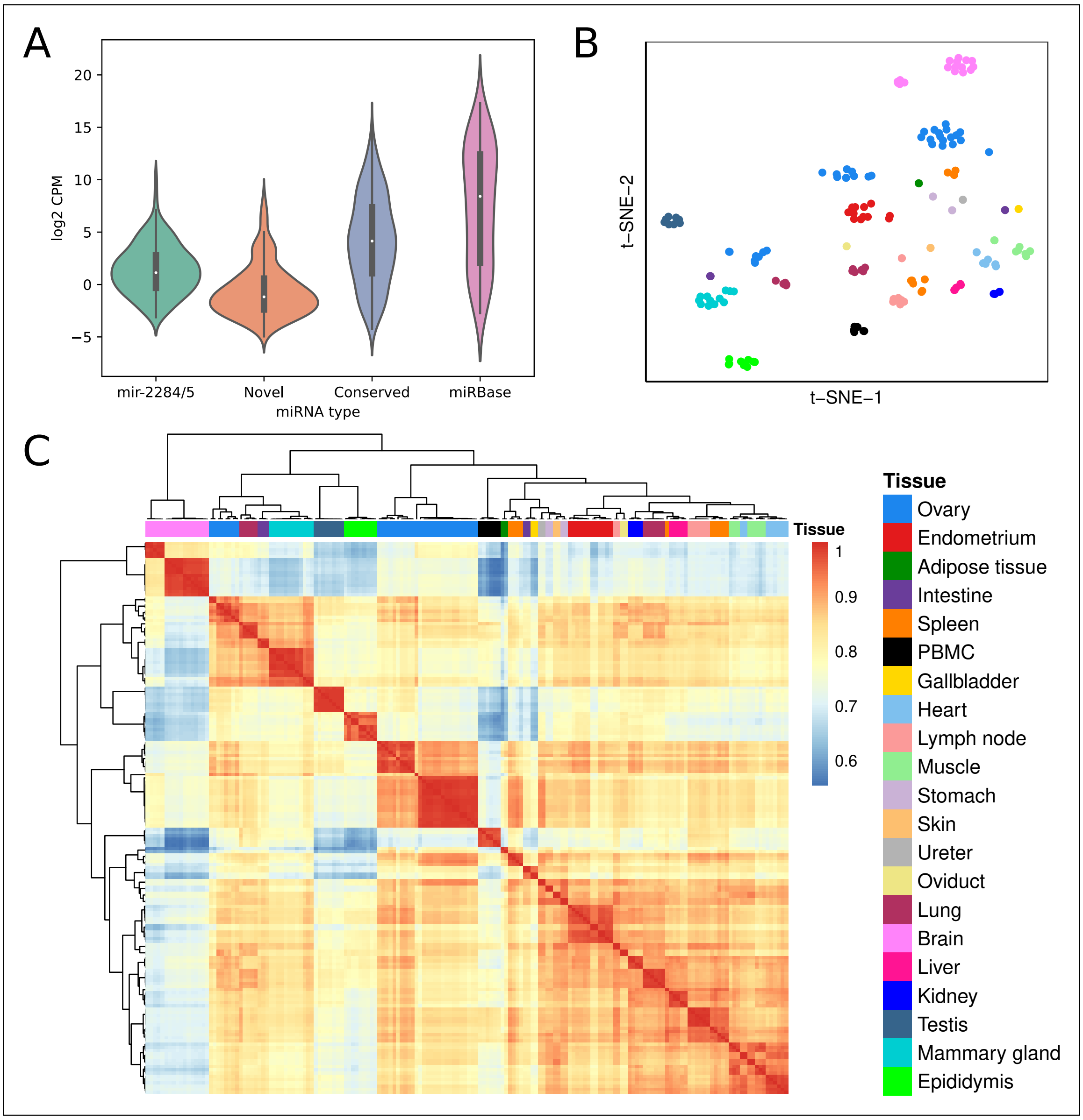


Figure 12: Exploratory analysis of miRNA expression. (A) Expression levels of all the miRNAs separated in categories based on sequence conservation. Mir-2284/5 family has been represented separately to see its specific pattern. (B) t-SNE plot of all the samples colored by tissue. (C) Correlation heat-map and clustering of all the samples using Pearson correlation. Color legend shared by (B) and (C) subfigures.

expression levels decline sharply from the eighth more expressed miRNA. The miRNAs oar-mir-143, oar-mir-26a, oar-mir-10a and oar-mir-10b were among the most expressed in most of the tissues, and in many tissues (intestine, gallbladder, lymph node, stomach, oviduct, ureter, lung, spleen and epididymis), miR-143 was the predominant miRNA. Most of the predominantly expressed miRNAs were annotated in miRBase.

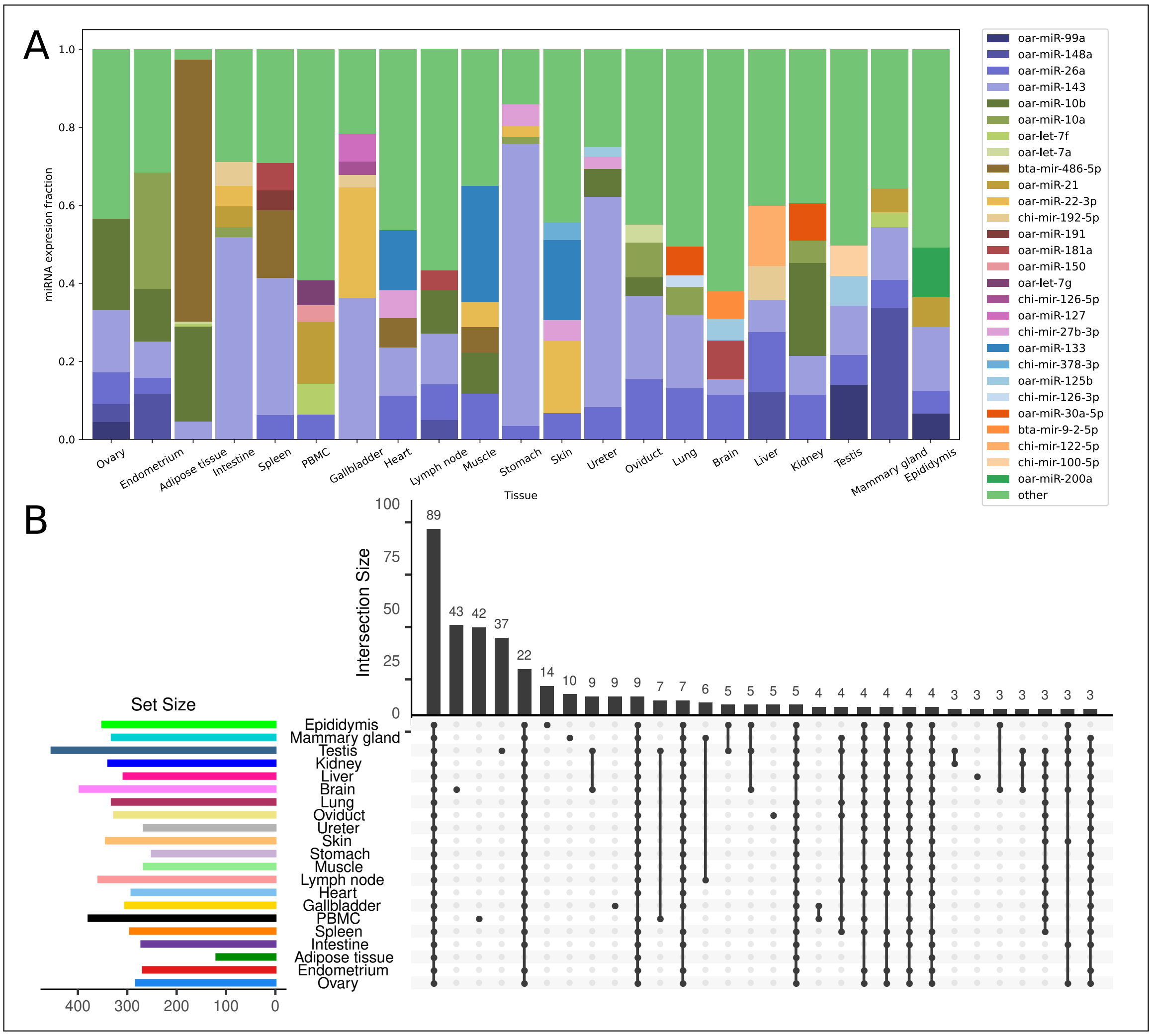


Figure 13: Distribution of miRNA expression across tissues. (A) Proportion of the five most expressed miRNAs in each tissue as a fraction of total expression, using the mean of all samples in each tissue. (B) Upset plot with the intersection of all the expressed miRNAs above 5 CPM in each tissue. Horizontal bars represent the number of expressed miRNAs in each tissue above the threshold. Intersections with at least three miRNAs are visualized in the vertical bars.

4.3.5. Tissue specificity analysis

To get a picture of the miRNAs expressed in each tissue, we considered all miRNAs expressed above a threshold of 5 CPM as strongly expressed in that tissue. 733 miRNAs were expressed above this threshold in half of the samples of at least one tissue, (Figure 13B). A set of miRNAs, containing 89 miRNAs, was expressed in all tissues. However, there were several miRNA strongly expressed exclusively in one tissue. In this analysis, 43 cerebral specific miRNAs, 42 PBMCs specific miRNAs and 37 testis specific miRNAs were detected. Other tissues with many strongly expressed exclusive miRNAs were epididymis (14), mammary gland (10) and gallbladder (9). The aforementioned tissues correspond to those that showed a lower correlation with other tissues.

To identify tissue specific and housekeeping miRNAs in a quantitative manner, a tissue specificity index (TSI) was calculated (Supplementary Table S4). 270 miRNAs could be considered tissue specific with a TSI value higher than 0.9 (Figure 14A). Of these, 18 were known miRBase sheep miRNAs and 92 miRNAs were conserved in humans, pigs, goats, cows or horses (Figure 14B). 25 miRNAs were exclusively expressed by a single tissue, most of them novel miRNAs.

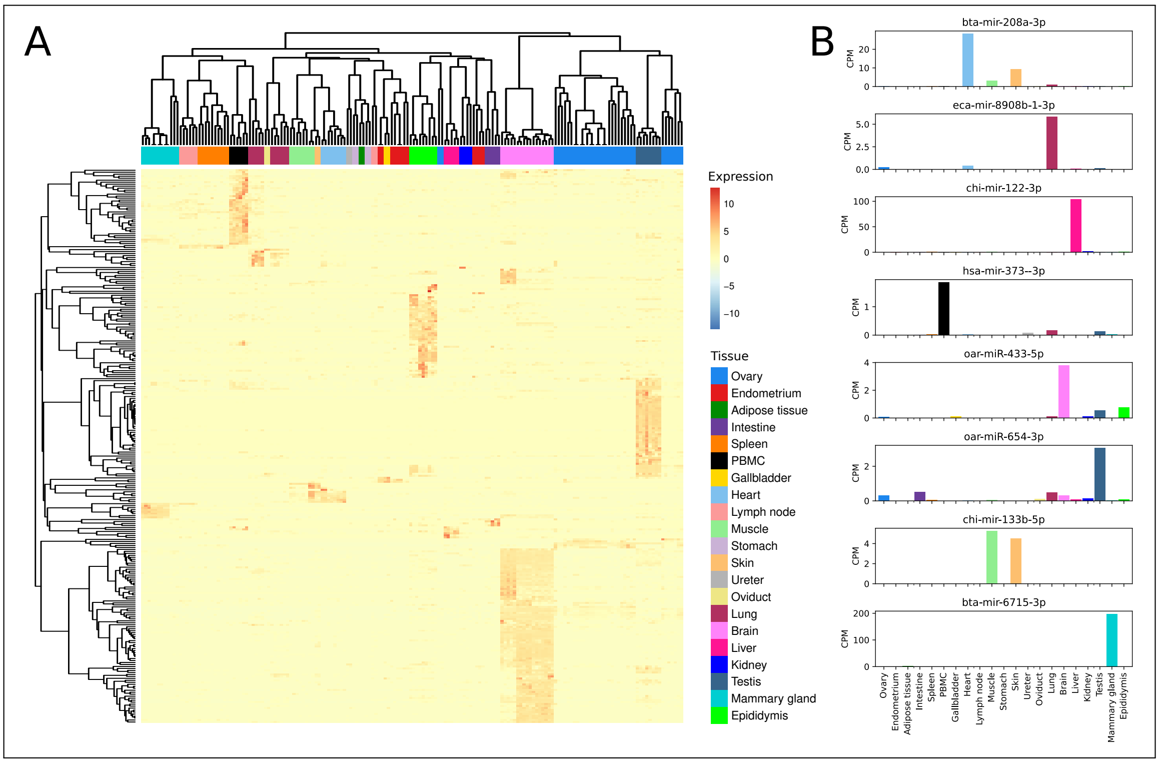


Figure 14: Tissue-specific miRNA expression. (A) Expression heatmap and sample clustering using all miRNAs with a TSI > 0.9. (B) Expression levels of selected tissue-specific miRNAs in different tissues identified in this study, using the mean of all samples in each tissue.

Nearly all tissue specific miRNAs showed their highest expression in one of the following tissues: brain (94), testis (54), epididymis (44) and PBMCs (37). It should be noted that there is a set of 46 miRNAs with their highest expression in the brain that are also highly expressed in testis or epididymis, or vice versa. These miRNAs include miRBase miRNAs such as oar-mir-433, oar-mir-1193, oar-mir-758, an orthologue of human hsa-mir-2113 or an ortholog of goat chi-mir-873. Samples were better grouped into tissues by the tissue specific miRNA heatmap (Figure 14A). Immune system-related tissues (PBMCs, spleen, and lymph nodes) were grouped together, even if there were almost no miRNAs specific to spleen or lymph nodes. This indicates that there is a common miRNA expression profile in relation to immunity.

Housekeeping miRNAs, defined here as miRNAs with a TSI < 0.25, were generally highly expressed and there was not any novel molecule among them. The 47 housekeeping miRNAs include oar-mir-143, oar-mir-26a and oar-mir-10b, present among the top 5 most expressed miRNAs in many tissues, and many members of the let-7 family. One of the only exceptions among the predominant miRNAs was chi-mir-122, which was specifically expressed in liver. Its most expressed mature arm, chi-mir-122-5p, had a TSI of 0.83, and the other arm, chi-mir-122-3p had a TSI of 0.97.

4.3.6. Ruminant specific mir-2284/mir-2285 family

In this work, we detected 146 miRNA loci expressing precursors belonging to the family of mir-2284/mir-2285 miRNAs. They were identified mainly based on sequence similarity with the over 200 cattle miRBase miRNAs from this family. Due to the high similarity between these miRNAs at precursor and mature miRNA level, exact one-to-one homologies were not given to the novel miRNAs. Instead, they were sequentially named (Supplementary Table S2). Regarding precursor sequences, there were 137 unique mir-2284/2285 family miRNAs, and regarding mature sequences, defined as the most expressed mature product from the same hairpin, there were 108 unique sequences belonging to this family. Thus, several copies of identical or very similar miRNAs are located through the sheep genome.

There were 156 mature miRNAs from the mir-2284/mir-2285 family expressed above the expression threshold. In general, their expression was lower than annotated miRNAs and other conserved miRNAs, but higher than novel miRNAs. In some tissues, their expression was significantly higher than in all tissues (Figure 15A). This was true for immune related tissues PBMCs (Mann-Whitney U test, P = 4x10-19) and lymph nodes (Mann-Whitney U test, P = 7x10-7), but not for spleen. The difference was also significant in testis (Mann-Whitney U test, P = 6x10-8). Those tissues with high mir-2284/mir-2285 expression coincide with some of the most transcriptionally distinct tissues, but, interestingly, in the brain, the most different tissue and a tissue with the most tissue specific miRNAs, the expression of mir-2284/mir-2285 family miRNAs was significantly lower (Mann-Whitney U test, P = 1x10-12). Overall, immune-related tissues and testis expressed the highest number of these miRNAs (Figure 15B). There were 16 tissue-specific miRNAs from this miRNA family, representing the 10% of expressed mir-2284/mir-2285 miRNAs, but this is lower than the fraction of tissue specific miRNAs from the whole dataset (25%) (Figure 15C). Those miRNAs were mainly specific of male reproductive tissues or PBMCs.

Because lowly expressed miRNAs with similar seeds can have an additive effect on target gene repression, the 98 genes that were predicted to be targeted by more than 10 mature miRNAs of the mir-2284/mir-2285 family were selected as their putative targets. AP3S1 was predicted to be targeted by 34 different miRNAs from this family, much more than any other gene. The GO overrepresentation test revealed that the set of 98 target genes was enriched in processes related to hormonal regulation, regulation of female sex organs and response to external stimuli (Figure 15D). The most significant term was *regulation of hormone levels* (GO:0010817, FDR=3.9x10-4), with genes such as AFP, FSHB, VAMP7 or ESR1. Other significant GO terms include *ovulation cycle process* (GO:0022602, FDR=2.4x10-3), with genes such as AFP, LHCGR or ESR1; and *regulation of response to external stimulus* (GO: 0032101, FDR=9.7x10-3), with genes such as IFNG, CXCL8 or CD200R1.

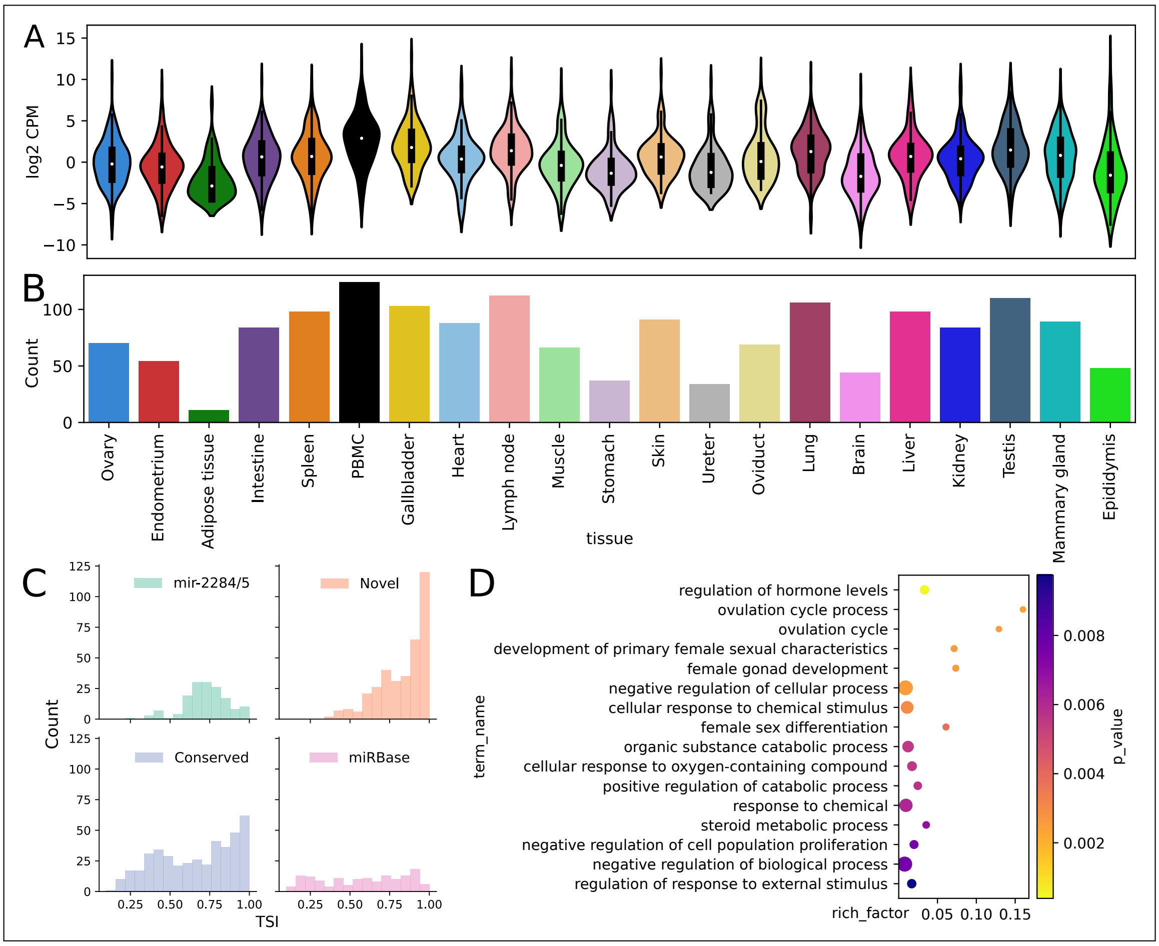


Figure 15: Expression analysis of miRNAs from the mir-2284/mir-2285 family. (A) Expression levels of all miRNAs from the mir-2284/mir-2285 family in each tissue. (B) Number of expressed miRNAs from the mir-2284/mir-2285 family with mean expression > 1 CPM. Color legend shared by (A) and (B) subfigures. (C) Distribution of TSI values in four miRNA categories based on sequence conservation. Mir-2284/5 family has been represented separately to see its specific pattern. (D) Gene Ontology enrichment results for the set of 98 target genes of the mir-2284/mir-2285 family. P values were corrected with Benjamini-Hochberg FDR.

4.4

Discussion

In recent years, thanks to the advances in sequencing technologies, many species have been extensively annotated for miRNAs. Among livestock species, nearly 1000 miRNAs have been annotated in cattle, whereas around 300 have been characterized in pig and goat. However, the current state of sheep miRNA annotation in the reference sources lies behind human and other livestock species. While there are 106 precursors in miRBase v.22 database [96] and 355 in the Ensembl v.104 annotation, we identify 1047 unannotated miRNA gene loci significantly expressed in any of the 172 samples analyzed. Moreover, most of the annotated miRNAs from both sources were detected in this study. 455 of these miRNAs were found to be conserved in another livestock species or in humans, while the remaining were classified as novel. Some of them could be orthologues of human miRNAs that were not defined as such due to the sequence divergence between species.

One of the advantages of this work is that it harmonizes the naming of all reanalyzed studies for an easier comparison between tissues. Other useful resources like the RumimiR database [98] contain an exhaustive in-depth description of the miRNAs from the literature, but we go further by reanalyzing all the raw data in an uniform way using the latest sheep genome (Oar_rambouillet_v1.0) in Ensembl (release 104). It should be noted that, due to the data being produced by different projects, there is some unavoidable variability, which could be caused by the experimental procedure for sample and RNA extraction, breed, diet, sex or age of the animals [111, 232, 256, 257]. Besides, the data for the tissues with many samples is more reliable than the tissues with few samples, more affected by variability, but, in general, the data seems representative for most tissues. We used the CPM normalization method, which compared to other methods appears to yield better reproducibility between individuals while keeping the distinction between cell types [258].

miRNAs are frequently clustered in the genome, 25% of human miRNAs are located in clusters and multicopy miRNAs tend to be in the same cluster [259]. Some interesting clusters were detected in this work. One of them, the miR-379/miR-656 cluster has already been reported in sheep, is conserved across placental mammals and is located in an imprinted region [249]. Interestingly, two novel big clusters were detected in this study on chromosome X, and one of them is potentially specific to the ruminant or ungulate clades, since its members have not been found elsewhere. This redundancy created by polycistronic loci and paralogous loci grants functional robustness to the mammalian miRNAome [236]. In addition, predominant expression of a miRNA in a tissue, observed here and in the FANTOM5 miRNA atlas [258], also plays a significant role in this functional robustness [236]. Most tissue-dominant miRNAs such as mir-143, mir-10b, mir-10a or mir-21 have been related to basic functions of cell homeostasis and division, and consequently, to cancer [260]. Still, there also are predominantly expressed miRNAs with roles related to the specific tissue or cell they are expressed in. The clearest example of this is mir-122, specifically expressed and predominant in mammal liver [261].

In this work, we have identified tissue-specific expression of miRNAs across 21 tissues. Other studies have previously generated miRNA atlases in various mammal species and have identified tissue-specific miRNAs with differing numbers of tissues. For instance, there are works in humans [258, 262], cattle [230], giant panda [263], horse [229], mice [258, 264], rat [265] and dog [266]. When comparing our dataset to those studies, we found many matching tissue-specific miRNAs, reinforcing the idea that miRNA gene expression of evolutionarily conserved miRNAs is also conserved [267].

The brain was the tissue that expressed most tissue-specific miRNAs and about half of the conserved sheep brain-specific miRNAs were also found to be specific in other works. Seven of them appeared in, at least, three other mentioned works: mir-129, mir-480, mir-551b, mir-137, mir-383, mir-380 and mir-487b [229, 230, 258, 262–266]. It has been proposed that brain miRNAs are closely related to the mental and behavioral variation during vertebrate evolution, by regulating the complex brain networks [268]. A reproductive tissue such as testis is known to be very transcriptionally complex, with a high number of expressed genes and specific genes, probably due to a more permissive chromatin [269]. In this dataset, it is the tissue with the highest amount of expressed miRNAs and 54 tissue-specific miRNAs show the highest expression in testis. Out of the 12 known or conserved tissue-specific miRNAs with highest expression in testis, 10 were also supported by other works, including mir-202 and mir-449a [262, 264–266]. miR202 mediates the proliferation, apoptosis, and synthesis function of human Sertoli cells [270] and the mir-449 cluster is essential for spermatogenesis [271]. Other tissue-specific miRNAs with extensively studied functions and supported by most of the tissue atlases include, for instance, mir-122, mir-133b and mir-208a/mir-208b. mir-122, specifically expressed in liver, is known to be involved in lipid and glucose metabolism [272]. mir-133b, expressed specifically in muscle, has an important role in the differentiation and proliferation of myoblasts [273]. The family of mir-208 miRNAs are exclusively expressed in cardiac muscle and are encoded in two myosin genes, being responsible for the control myosin content [273].

There is a set of highly conserved miRNAs [12], as well as a great correlation between the miRNA expression profiles of different mammal species [262]. The list of conserved miRNAs has been mostly completed, but there are also clade-specific miRNA families, usually lowly expressed, that could contribute to the phenotypic differences between livestock species [12, 274]. One of those families is the extensive mir-2284/mir-2285 miRNA family, which was thought to be specific to cattle. This family seems to have evolved by seed shifting and point mutation, has expanded very rapidly and might be related to insulin resistance in ruminants [275]. Homologs of the mir-2284/mir-2285 family have been previously described by other sheep studies [233, 276], but here we report a vast number of members in sheep, comparable to the amount in cattle, thus confirming that the expansion of this miRNA family is, at least, ruminant specific. Functionally, this expansion could have had two outcomes: a progressive subfunctionalization depending on the tissue [277], or an additive dosage effect on a restricted number of target genes [275]. The data from this study suggest the latter, since they are less tissue-specific than other families and many predicted target genes are shared among the paralogues.

As for the biological role of the mir-2284/mir-2285 family miRNAs, in cattle, they are expressed in immune-relevant tissues [278], and show their highest expression in lymph nodes [277]. Our sheep dataset also follows this trend, as they are highly expressed in PBMCs and lymph nodes. Nevertheless, besides the predicted targeting of immune response and inflammation associated genes, their predicted target genes were also related to hormone regulation and female sex cycle. The main target of this miRNA family was predicted to be AP3S1, which encodes a subunit of the AP3 adaptor complex, involved in intracellular vesicle trafficking. Not only AP3 is involved in the inflammatory response [279–281], but also, mice lacking AP3 show dysregulated insulin and other hormone secretion [282]. The specific subunit encoded by AP3S1 seems to play a role in the insulin receptor signalling [283] and variants within AP3S1 have been associated with type 2 diabetes in a Chinese population [284]. Considering these results, because vesicle trafficking is important for both, immune cell function and hormone regulation and evolutionary innovations have been important for the development of the unique ruminant digestive system and metabolism [123, 285, 286], mir-2284/mir-2285 family miRNAs could have evolved as an adaptation to regulate these processes. The exact phylogenetic origin and functional roles of this family remain to be studied.

In this work, we have created an expression atlas of sheep miRNAs by the integration of several small RNA sequencing experiments, including hundreds of previously unannotated and uncharacterized miRNAs. Our analyses support the high conservation of many miRNAs, but also highlight the potential of clade-specific innovations for ruminant evolution, such as the ruminant-specific family of mir-2284/mir-2285. The dataset itself and the analyses regarding expression distribution and specificity of miRNAs should be useful for the field of sheep genomic and veterinary research, as it provides sheep-specific information about the expression of any miRNA in 21 tissues.

Chapter 5

Analysis of lncRNAs to evaluate the effect of aluminium hydroxide in ovine encephalon

This chapter is based on the following publication:

Varela-Martínez, E.*, **Bilbao-Arribas, M.***, Abendaño, N., Asín, J., Pérez, M., de Andrés, D., Luján, L. and Jugo, BM. Whole transcriptome approach to evaluate the effect of aluminium hydroxide in ovine encephalon. *Sci Rep* **10**, 15240 (2020)

* Co-authorship

5.1

Background

Since the 1920's, when aluminium (Al) was discovered to enhance immune response providing more effective protection [287] vaccines have been complemented with adjuvants. Because of the effectiveness of aluminium adjuvants at enhancing humoral responses, their good tolerance without causing fever and with the longest safety record among used adjuvants [288] aluminium salts are preferably used in both animal and human vaccines. Nevertheless, the mechanism of enhancement of immune response by adjuvants has not been thoroughly analyzed and its importance has been underestimated for a long time [289].

The aluminium oxyhydroxide based Alhydrogel is one of the most common aluminium-based adjuvants used in clinically authorized vaccines. The potential effect of this kind of compounds on the nervous system has been tested mainly in animal models such as mouse. In CD1 mice, with a dose of 100 µg Al/kg, subcutaneously inoculated Alhydrogel adjuvant induced cognitive alterations associated with death of motor neurons and an enormous increase (350%) of reactive astrocytic cells in an inflammatory process[290]. Moreover, with a dose of 300 µg Al/kg, microglial and astroglial reactions were detected in the spinal cord of the same mice type, and altered motor and cognitive functions were observed [291]. In an immunization experiment in mice, after the inoculation of oxyhydroxide particles fluorescently labelled, an average of 15 solid aluminium particles were detected in the mice brain at 21 days postimmunization. In vitro studies performed in parallel confirmed the toxicity of aluminium adjuvant to neuronal cell cultures [292].

Very few studies have analysed the Al effect in animal nervous system by RNA-seq technology. In a recent work, Xu *et al.* [293] identified by means of RNA-seq 96 upregulated and 652 downregulated mRNAs, and 37 dysregulated long non-coding RNAs (lncRNAs) in the hippocampus of Al treated rats. The main functions of dysregulated genes, revealed by Gene Ontology analysis, were related with glial cell differentiation, neural transmission and vesicle trafficking. Moreover, the results of this study suggested that glial cell-related genes had relevant effects in the mechanisms associated with Al neurotoxicity and that aberrant mRNAs and lncRNAs were involved in the response to Al in the analysed tissue.

Our group has characterized the effect of Al hydroxide adjuvant and its influence on the immune response to vaccination in a long term experimental design, using sheep as a model, based in total RNA and microRNAs sequencing in peripheral blood mononuclear cells (PBMCs) [235]. With the main objective of deciphering the molecular signature activated, two different treatments were applied to lambs: commercial vaccines including Al hydroxide or Alhydrogel (aluminium hydroxide gel suspension) only in an equivalent dose. In animals of both treatments the NF-kB signalling pathway was enriched, and at the end of the experiment a downregulation of cytokines and cytokine receptors was detected in the adjuvant inoculated animals in relation to the vaccinated animals. In the adjuvanted group, differential expression of six miRNAs was also detected. Thus, aluminium could induce endogenous danger signals with an effect in the stimulation of the immune system.

Long non-coding RNAs are non-coding RNAs longer than 200 nucleotides and often transcribed. They usually do not code for proteins but their spatiotemporal-specific expression patterns indicate their diversity in functions and complexity in mechanisms [294]. They are implicated in neural function and maintenance, and many neurodegenerative diseases such as Alzheimer's disease (AD) have been linked with aberrant lncRNAs [295]. They have been also associated with chemical carcinogenicity and metal toxicity, and the relationship of some lncRNA and cadmium for example, has been reported [296].

Thus, the main objective of this study was to identify the molecular signatures activated by vaccines and adjuvants in the form of Al hydroxide in sheep encephalon, in the same group of animals as PCGs and miRNAs were analysed, by combining the molecular information provided by RNA sequencing of mRNAs and lncRNAs.

5.2

Methods

5.2.1. Animals

The animals studied in this work were previously analysed for a different tissue (PBMCs) [235]. Briefly, twenty-one Rasa Aragonesa purebred lambs were selected from a single pedigree flock of certified good health at three months of age and did not undergo any vaccination before the experiment. The flock analysed in this study was established at the experimental farm of the University of Zaragoza, with ideal controlled conditions of housing, management and diet. The experiment started after an acclimatization phase of two months, when the animals were five months old. For the purpose of the present work, they were randomly distributed in different treatment groups, n=7 each. Each treatment group was kept isolated from the others in three adjacent identical home pens with the same conditions of housing, diet and management across all the study. Each group received a parallel subcutaneous treatment with either commercial vaccines containing Al hydroxide ($Al(OH_3)_3$) as adjuvant (Group Vac), Al hydroxide only (Group Adj; Alhydrogel®, CZ Veterinaria, Spain) or PBS (Group Control). Nine different vaccines were used and a total of 19 inoculations were applied to each animal throughout 16 different inoculation dates, thus entailing a total amount of 81.29 mg of Al per animal in Vac and Adj groups (Table S1). The complete study lasted 475 days, from February 2015 to June 2016. Twelve animals were included for the RNA-seq analysis, 4 of each treatment group at the end of the experimetn (Figure S1). For the validation of the sequencing data 9 different animals were included, 3 of each treatment group (Table 5).

5.2.2. Tissue collection and RNA extraction

Tissues for pathologic studies were collected at necropsy. Samples of 1 g of parietal lobe from each sheep, with constant proportions of gray and white matter, were taken for RNA extraction and preserved in RNAlater solution (Ambion, Austin, TX, USA) at -80 °C. The experimental procedure to obtain RNA was similar to the one previously performed in the analysis of PBMCs [235]. Total RNA was isolated from encephalon tissue using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) and PureLink RNA Mini Kit (Invitrogen). 60 mg tissue samples were homogenized in 1 ml of TRIzol Reagent using Precellys®24 homogenizer (Bertin Technologies, Montigny-le-Bretonneux, France) combined with 1.4 and 2.8 mm ceramic beads mix lysing tubes (Bertin Technologies). RNA isolation was performed following manufacturer instructions and RNA was suspended in RNase free water and stored at -80 °C. RNA quantity and purity was assessed with NanoDrop 1000 Spectrophotometer (Thermo Scientific Inc, Bremen, Germany). RNA integrity was assessed on a Agilent 2100 Bioanalyzer with Agilent RNA 6000 Nano chips (Agilent Technologies, Santa Clara, CA, USA), which estimates the 28S/18S (ribosomic RNAs) ratio and the RNA integrity number (RIN value). The samples presented an av-

Table 5: Samples used in RNA-seq and RT-qPCR study.

Treatment	RNA-seq	RT-qPCR
Aluminum	114-E, 115-E, 116-E, 117-E	111-E, 112-E, 113-E
Vaccine	121-E, 122-E, 124-E, 126-E	123-E, 125-E, 127-E
Control	131-E, 135-E, 136-E, 137-E	132-E, 133-E, 134-E

erage RIN value of 8.06 and a 260/280 ratio > 1.7.

5.2.3. RNA sequencing

The TruSeq Stranded Total RNA kit with Ribo-Zero (Illumina, San Diego, CA, USA) and the TruSeq Small RNA library prep kit (Illumina) were used for Total RNA-seq and miRNA-seq, respectively. Total RNA libraries were sequenced on a HiSeq2000 with a mean sequencing depth of 75 million reads (75 bp paired-end reads) at CNAG (Centro Nacional de Análisis Genómico, Barcelona, Spain), while miRNA libraries were sequenced on a HiSeq2500 with a mean sequencing depth of 19 million reads (50 bp single-end reads) at CRG (Centro de Regulación Genómica, Barcelona, Spain). The samples used for sequencing and qPCR can be seen in Table 5.

5.2.4. Total RNA expression analysis

The bioinformatics procedure to obtain the expression matrix was similar to the one previously described in the analysis of PBMCs [235]. Briefly, after quality filtering and trimming, the reads were aligned with the STAR algorithm [v2.5.4a] [297] to the Ovis aries genome build Oar3.1 [298]. For each library, the uniquely aligned fragments were assigned to annotated genes in a strand specific manner with featureCounts [v1.6.0] [299]. Apart from annotated genes, one of the interests of this work is to find new lncRNAs and study their function in sheep brain. For that purpose, an additional step after mapping was necessary. The StringTie [v1.3.3b] [81] transcriptome assembler was used to reconstruct the transcriptome from the previous mapping. From this assembly, only candidate lncRNAs were selected (the selection process and analysis is explained below) and their counts were added to the count matrix of annotated genes.

The same sample (116-E) was treated as outlier and was filtered out from the analysis. Prior to the differential expression, the SVA package [v3.26.0] [300] was applied to remove unwanted variation and the obtained surrogate variables were incorporated into the testing model. A PCA was obtained with the corrected data (see Supplementary Fig. S1A in published article [234]). In this PCA the samples grouped according to treatment condition. The differential expression analysis was performed using DESeq2 [v1.18.1] [113] with the following variables in the model: treatment (Control, complete vaccine [Vac] or adjuvant only [Adj]) and SVA covariates (surrogate variables calculated by sva). Three different comparisons were made (Adj vs. Control, Vac vs. Control and Adj vs. Vac) in which differentially expressed genes (DEGs) were selected as those with an adjusted p-value (with the Benjamini Hochberg method) threshold of < 0.05 and a fold change > 1.5 or < 0.667. Then, gene enrichment analyses were conducted using the GO database in PANTHER [v12.0] [301] and the KEGG database in DAVID [v6.8] [302], considering enriched terms as those with an adjusted p-value threshold of < 0.05.

5.2.5. Weighted gene co-expression network analysis

A weighted gene co-expression network analysis was performed using the WGCNA [v1.63] [303] R package. Briefly, the similarity matrix was constructed from the normalized data using absolute values of the biweight midcorrelation, chosen for being more robust against outliers. Then, the adjacency matrix was defined by raising the similarity matrix to a power β. The parameter β was selected based on the minimum value required to get a scale-free topology network (R2>0.8), in our data being β=28. Once the network was constructed, module (clusters of densely interconnected genes) detection was the next step, setting a minimum module size of 30 genes. Finally, modules with similar expression profiles were merged based on a height cut-off threshold of 0.3.

Next, we sought modules with strong correlations with the treatment groups. For that purpose, the treatment variable was dichotomized in all possible combinations (one group against the other two). For each of the identified modules, eigengene values (the first principal component of each module) were generated and were used as representation of the weighted average of the gene expression profile in the modules. Pearson correlations and their associated p-values were generated for all pairwise comparisons of the module eigengene expression values and the treatment parameters. All the p-values were used for estimation of the FDR (q-value) with the qvalue R package, selecting those modules with a q-value threshold < 0.05.

Modules exhibiting high correlation with the treatment were further studied for enrichment of GO terms and KEGG pathways, considering statistically significant those with an adjusted p-value threshold of < 0.05. Apart from enrichment analysis, the hub genes of each module were obtained. For that purpose, the module membership (MM) and gene significance (GS) values were calculated. GS values are the Pearson correlations between the single expression value of each gene and the treatment parameter, whilst MM values are the Pearson correlations between the single expression value of each gene and module eigengene values. We defined hub genes as those belonging to the $\geq$ 85th percentile for both MM and GS in each module. Those genes are likely 'key drivers' and might play important roles in the treatment.

5.2.6. Analysis of lncRNAs

gffcompare software was used to classify all sequenced transcripts based on their location relative to the annotation and extract unknown intergenic transcripts (lincRNAs), intronic lncRNAs and antisense lncRNAs. Multiexonic transcripts of less than 200 nucleotides and single-exon transcripts of less than 2,000 nucleotides were filtered out. The coding potential of the remaining transcripts was assessed with three approaches. Coding Potential Calculator 2 (CPC2) is a machine learning based program with a species-neutral model able to classify coding and non-coding sequences [304]. Coding-Potential Assessment Tool (CPAT) is another machine learning based program that we trained and selected the classification threshold following authors' instructions using available bovine coding and non-coding sequences [305]. HMMER 3.1b2 [306] was used to detect Pfam protein domains in our potential lncRNAs, which were translated into the three possible frames. Transcripts classified as non-coding by CPC2 and CPAT and without protein domains detected were selected and treated as lncRNAs for their functional analysis. Besides, genes already annotated in sheep (Oar_v3.1) with "lincRNA" biotype were also added. To evaluate the sequence conservation and to look for known homologues we performed a Blast search with each lncRNA transcript to the entire RNAcentral database, which has an up-to-date collection of non-coding RNA sequences [307].

For trans acting lncRNAs potential protein-interacting lncRNAs were predicted with LncADeep tool [308] and sequences of proteins with at least evidence at transcript level or from homology were downloaded from UniProt. For more confident results, interactions were only predicted for proteins from genes in the same co-expression modules and a probability of 0.9 was set as threshold.

5.3

Results

5.3.1. Statistics for RNA-seq data

The sequenced 12 RNA-seq libraries had an average depth of 74.1 million paired-end reads. After adaptor and quality filtering, a mean of 68.8 million reads (92.80%) remained for subsequent analyses. Those reads were aligned against the Ovis aries reference genome (Oar3.1), achieving the following results in average: 60.7 million read pairs (88.33%) mapped uniquely to the reference, 5.9 million read pairs (8.54%) mapped to multiple loci and 2.1 million read pairs (3.13%) not mapped to any loci. Only uniquely mapped reads were used for subsequent analyses.

5.3.2. Identification and classification of lncRNAs

Filtering steps to improve the reliability of unknown intergenic, intronic and antisense transcripts as lncRNAs reduced the list of potential lncRNAs to 3,004. Despite their different approaches, the three methods for detecting coding sequences performed in concordance, with CPAT and CPC2 giving more similar results (Figure 16a). They are evenly distributed across all the chromosomes except for the X chromosome that harbours less transcripts than expected for its length (Figure 16b). More than half of the transcripts are longer than 5,000 nucleotides, many of the single-exon transcripts are between 2,000 and 4,999 nucleotides long and there are few transcripts with more than 3 exons (Figure 16c). We classified all the transcripts into different categories based on their relative location to their closest genes. Transcripts overlapping and in the same strand as known coding genes were not considered. Most lncRNAs are located in intergenic regions and those less than 5 kb apart from their neighbours are classified in their own category due to potential regulatory relations (Figure 16d). Intronic lncRNAs showed better correlations, in average, with their closest genes than other categories and the genes that harboured these transcripts were enriched in several functions and pathways related to neuron activity (Figure 17), while other lncRNA types did not show any overrepresented ontology or pathway terms.

In relation to the conservation of detected lncRNAs in sheep, through Blast searches against RNAcentral database, we found out that 144 unannotated transcripts (5%) had significant matches with lncRNAs already annotated in other species. Among them, the lncRNA TUNA was detected, which was differentially expressed between the adjuvant group and the other two groups. This lncRNA has been found conserved in many vertebrates like cattle (URS00008E3A0F) or human (URS000075CAB8). We also identified, albeit with incomplete alignments, similar transcripts to other human lncRNAs such as NORAD, HCG11 or COPG2IT1.

5.3.3. Analysis of differential expression of mRNAs and lncRNAs

First, lowly expressed genes, defined as those with an expression lower than 1 CPM and found in less than four individual libraries, were filtered out from the differential expression analysis. Thus, 16,369 genes remained for subsequent analysis, of which 14,387 were annotated genes in Ensembl and 1,982 were candidate lncRNAs. One sample from the adjuvant group was treated as an outlier and was extracted from the analysis.

In the Adj vs. Control comparison 63 DEGs were identified, including 33 genes, of which 20 were up-regulated and 13 were down-regulated, and 30 new lncRNAs consisting of 3 that were up-regulated and 27 down-regulated. In the Vac vs. Control compar-

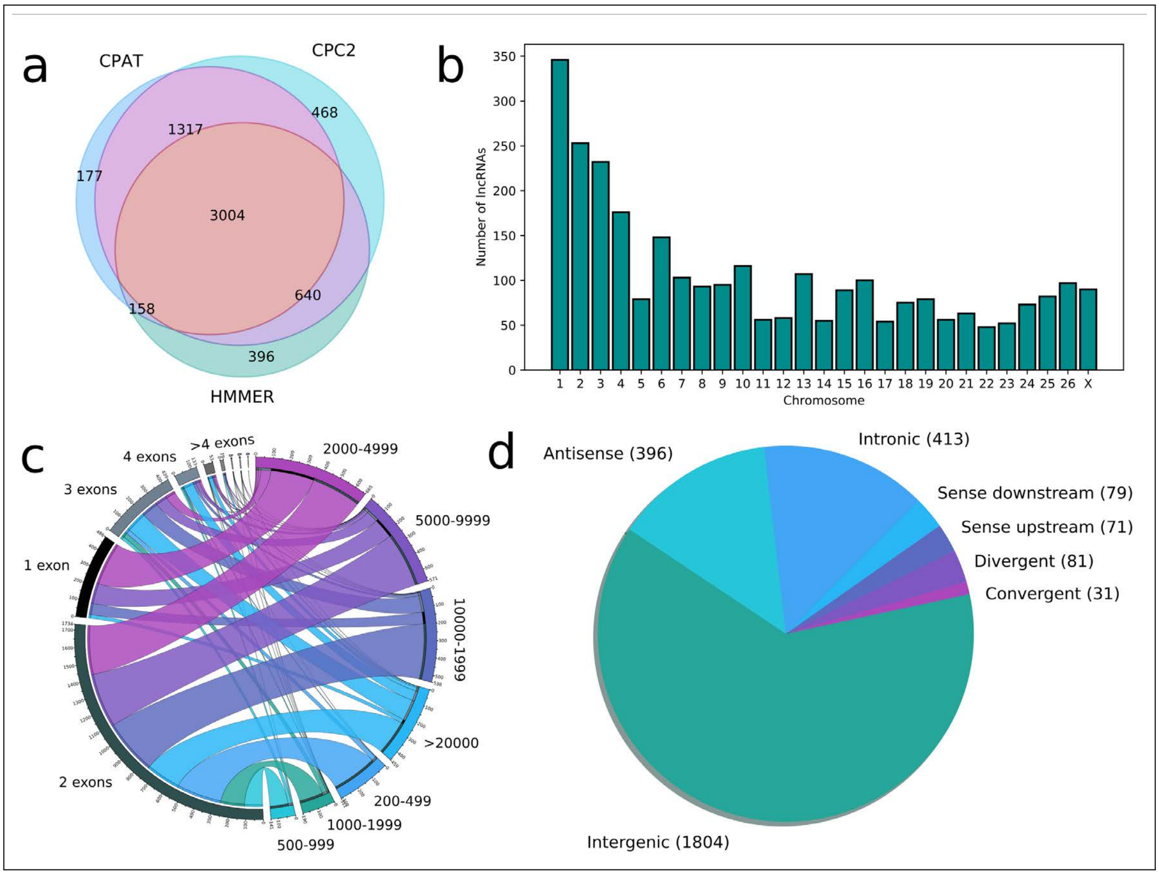


Figure 16: Summary statistics of the lncRNAs. (a) Venn diagram with the coding-potential assessment results obtained with CPAT, CPC2 and HMMER. (b) Distribution of lncRNA transcripts through chromosomes. (c) Relationship between length and exon number in the detected lncRNAs. (d) Classification of detected candidate lncRNAs by relative location to the closest annotated gene.

ison 13 DEGs were identified, including 6 genes, of which 2 were up-regulated and 4 were down-regulated, and 7 new lncRNAs consisting of 5 that were up-regulated and 2 down-regulated. Furthermore, in the Adj vs. Vac comparison 76 DEGs were identified, including 45 genes, of which 33 were up-regulated and 12 were down-regulated, and 31 new lncRNAs consisting of 4 that were up-regulated and 27 down-regulated. A detailed list of the DEGs can be seen as a heatmap (Figure 18a). In Supplementary Datasets S1 and S2 it can be seen a detailed summary of the differential expression analysis for all genes and lncRNAs that passed the filtering criteria (see published article [234]).

Within the DE-mRNAs are factors that are clearly related to neuronal development (NID2, VIM, NTN1, SEMA3, EYA1, CDH19), brain transport and neurotransmission (SLC13A3, SLC6A20, SLC6A12, MOCOS, TRPM4, KCNJ13, CUBN, MRASAL1), brain injury (FN1, BHMT2, PATL2, GDF10, GSN, FGL2, OTOF, VCAM1, PROS1, COL4A5, EFEMP1, NPFFR2, LAMA2, ADAM12, MYOF) and neurodegenerative diseases associated with Al like AD (ND6, STOML2, MRC1, KDR, NEIL2), Parkinson Disease (PD)

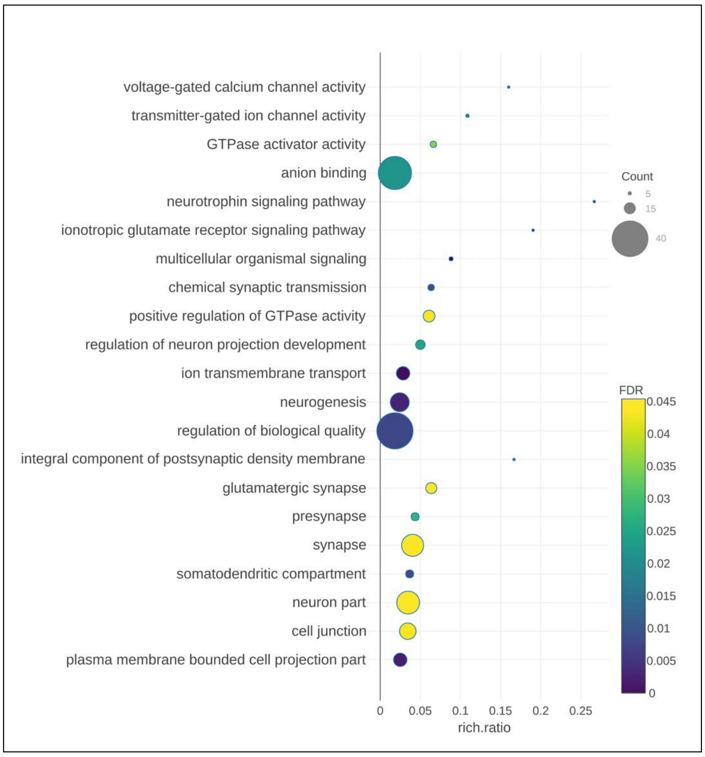


Figure 17: Pathway analysis of genes that harboured intronic lncRNAs. The bubble plot shows in the Y-axis the enriched pathways, while in the X-axis the rich ratio is represented (rich ratio = amount of genes in the term/total amount of genes in the enriched term). Size and colour of the bubble represents the number of genes in the GO term and enrichment significance (FDR), respectively.

(ATP13A5, HIST1H1C) and Amyotrophic Lateral Sclerosis (ALS) (ANXA2) (Figure 18b).

5.3.4. Functional annotation and classification for RNA-seq data

Functional characterization of the DE-mRNAs was performed with PANTHER to identify enriched GO terms in the three domains: Cellular Component (CC), Molecular Function (MF) and Biological Process (BP). In the Adj vs. Control comparison, 27 significantly overrepresented GO terms (with an adjusted p-value < 0.05) were identified in total. Among the top ranked Biological Processes were positive regulation of mitochondrial DNA replication (GO:0090297), stress-induced mitochondrial fusion (GO:1990046), mitochondrial ATP synthesis coupled proton transport (GO:0042776), positive regulation of cardiolipin metabolic process (GO:1900210), alpha-ketoglutarate transport (GO:0015742), peptidyl-arginine methylation to symmetrical-dimethyl arginine (GO:0019918), positive regulation of mitochondrial membrane potential (GO:0010918), mitochondrial protein

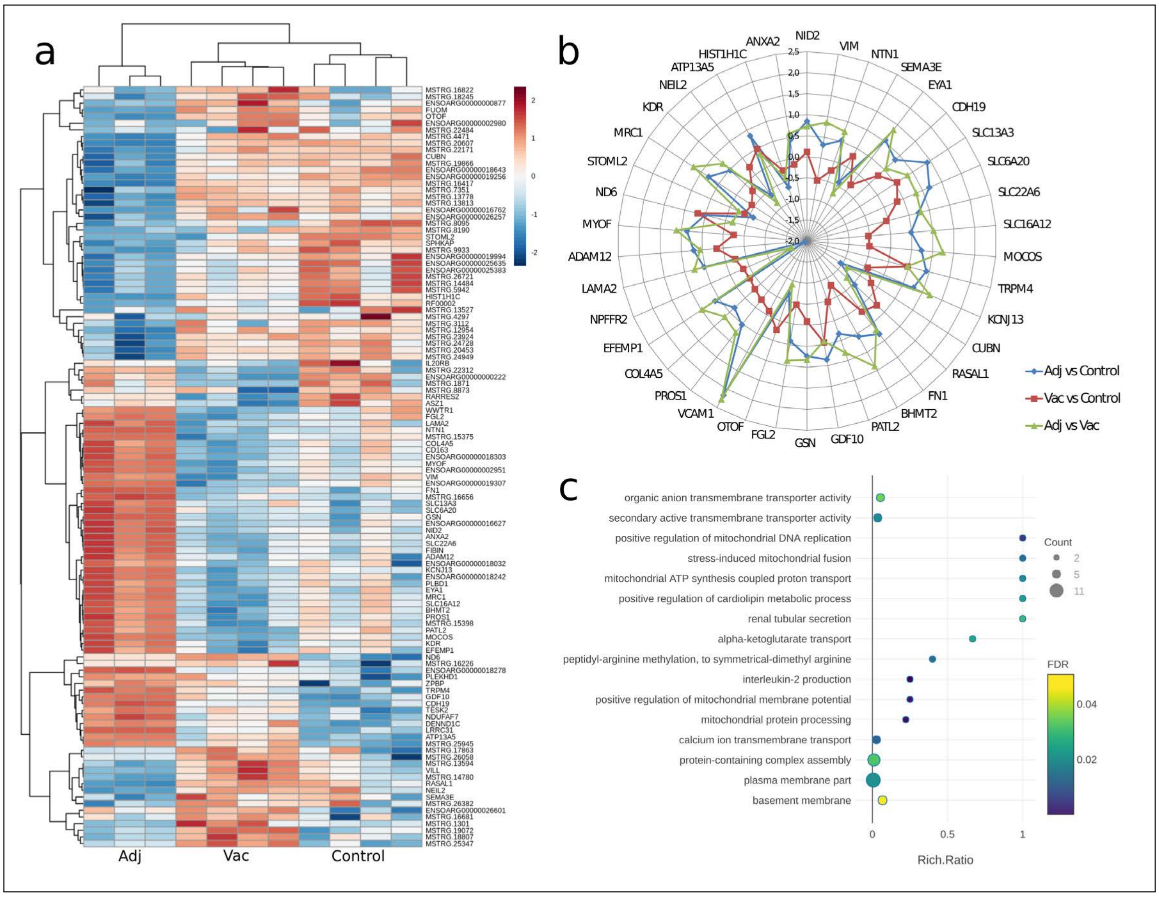


Figure 18: Differential expression of coding and lncRNA genes. (a) Heatmap depicting all the differentially expressed genes in Adj vs. Control, Vac vs. Control and Adj vs.Vac comparisons. (b) Radar plot with the log2FC of overrepresented genes related to neuronal development, neurotransmission and neurodegenerative diseases in Adj vs. Control (blue), Vac vs. Control (red) and Adj vs.Vac (green) comparisons. (c) GO enrichment term analysis of differentially expressed genes in the Adj vs. Control and Adj vs. Vac comparisons. The bubble plot shows in the Y-axis the enriched GO terms, while in the X-axis the rich ratio is represented (rich ratio = amount of differentially expressed genes in the term/all genes included in the term). Size and colour of the bubble represent the number of differentially expressed genes in the GO term and enrichment significance (FDR), respectively.

processing (GO:0034982) and calcium ion transmembrane transport (GO:0070588) (Figure 18c).

5.3.5. Results from the weighted gene co-expression network analysis

Next, a gene co-expression network was constructed with WGCNA. Such networks provide a way to account for the coordinated expression among genes and discern possible differences between individuals that may relate to differences in treatment group. A total of 45 co-expressed gene modules were detected (Figure 19a, Figure 19b), module size ranging from 37 to 2,724 genes. Each module was assigned a 'colour name'. We searched for significant correlations among module eigengenes and treatment parameters. There were no co-expressed modules associated with the Control group. In contrast, three modules showed strong correlations with Vac group and two with Adj group: the mediumorchid4 module (189 genes, r=0.88, qvalue=0.01), the brown3 module (377 genes, r=0.88, qvalue=0.01) and the palevioletred3 (275 genes, r=-0.95, qvalue=0.001) for Vac group and the maroon module (1,325 genes, r=0.88, qvalue=0.01) and the burlywood1 module (228 genes, r=-0.83, qvalue=0.04) for Adj group (Figure 19c). Interestingly, the maroon module included 36 DEGs, the remaining modules having an insignificant number of DEGs in comparison.

The obtained treatment associated modules were further studied for enrichment of GO terms and KEGG pathways. Only the modules maroon and burlywood1 had significant enrichments, while the others, probably due to the small number of annotated genes, did not have significant enrichments. The maroon module, positively correlated with the adjuvant samples, was enriched for some GO terms, among them regulation of interleukin-1 beta production (GO:0032651), negative regulation of extrinsic apoptotic signaling pathway (GO:2001237), negative regulation of canonical Wnt signaling pathway (GO:0090090), positive regulation of immune system process (GO:0002684), and inflammasome complex (GO:0061702). A more detailed list of the enriched GO terms from the Biological Process category for the maroon module can be seen as supplementary figure S3 (see published article [234]). In addition, only the maroon module was enriched in KEGG pathways, mainly: ECM-receptor interaction (oas04512), amoebiasis (oas05146), focal adhesion (oas04510), PI3K-Akt signaling pathway (oas04151), protein digestion and absorption (oas04974) and NF-kappa B signaling pathway (oas04064).

Since hub genes are likely 'key drivers' of the co-expression modules, we checked the treatment related modules. In Supplementary Table S1 there is a detailed list of the hub genes in these modules (see published article [234]). To note the maroon module, in which 17 of the hub genes are DEGs. Some of them, as previously detailed, had been related with brain injury (GSN, LAMA2 and PROS1), neuronal development (NTN1 and NID2) and different diseases in brain (MRC1 and ANXA2). Apart from the differentially expressed genes, there are other genes related to other functions such as insulin signalling (INSR, IGFBP2 and IGF2BP2), blood brain barrier (ADGRA2 and NTN1), ERK signalling (INSR, ITGA9, OSMR, COL18A1, LAMA2, BCL2L11, ADAM17, COL4A3,

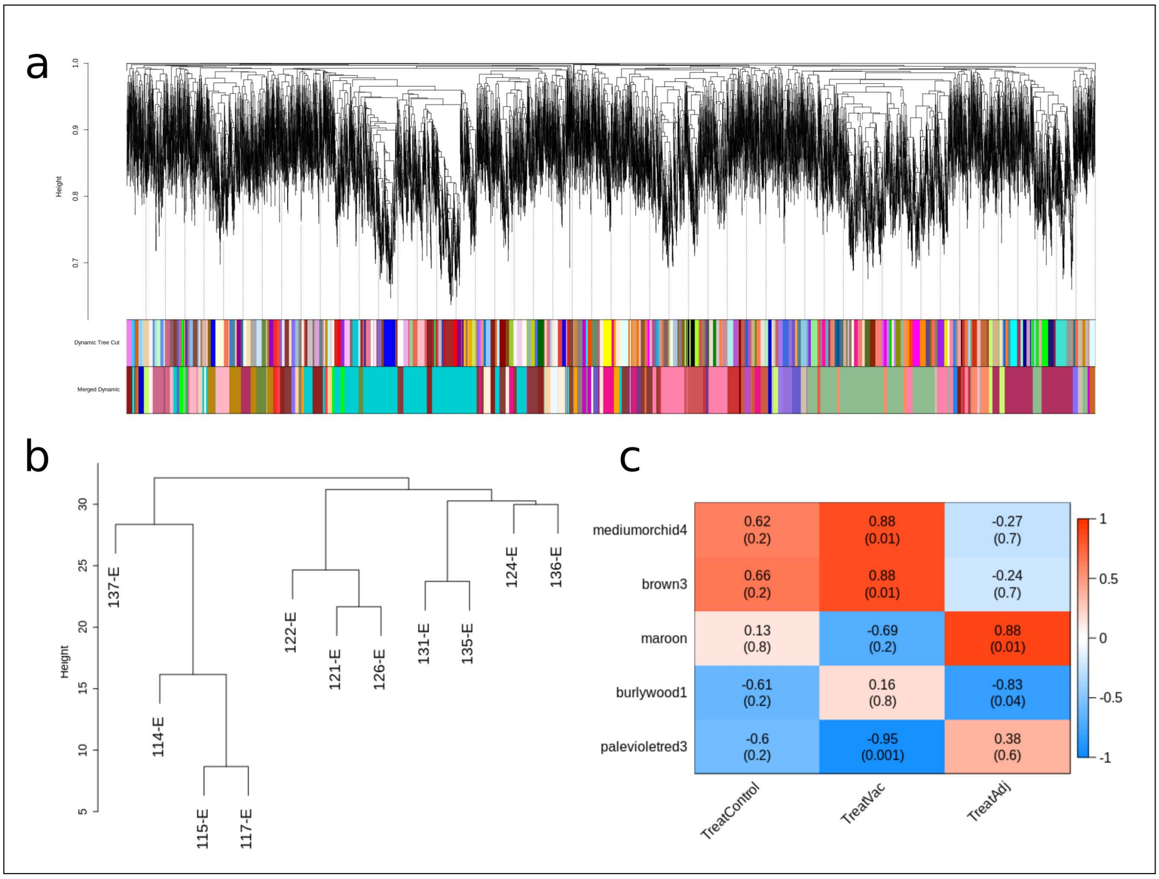


Figure 19: Weighted gene expression co-variance network analysis (WGCNA) summary. (a) Gene dendrogram obtained by average linkage hierarchical clustering. The colour rows underneath the dendrogram shows the module assignment before (Dynamic Tree Cut) and after (Merged Dynamic) modules with similar expression profiles were merged. (b) Hierarchical clustering of samples used in the analysis. (c) Module-trait associations. Each row corresponds to a module eigengene, while the columns to a trait. Each cell contains the corresponding correlations and adjusted p-values. The table is color-coded based on the correlation between the eigengene and corresponding trait. Only modules associated with at least one trait are shown.

COL4A4, COL4A6, COL2A1 and BMP4) and calcium signalling (APOOL, HOMER3 and TMBIM1). It seems that the maroon module is composed of genes essential for the correct function of the brain.

We performed predictions based on proposed mechanisms of action for the DE lncRNAs in the co-expression modules. Trans-acting lncRNAs could act in many ways to epigenetically regulate expression of distant genes, for instance, by recruiting or acting as scaffolds of proteins. 20,011 lncRNA-protein interactions were predicted in total, with an average of 235 interactions per lncRNA transcript. Top scoring interactions were used to build a network of lncRNA-protein interactions with proteins whose mRNA transcripts are correlated with DE lncRNAs (Figure 20). Among these interactions appeared all four RNA-binding proteins of the ELAV/Hu family, mainly expressed in differentiated neurons.

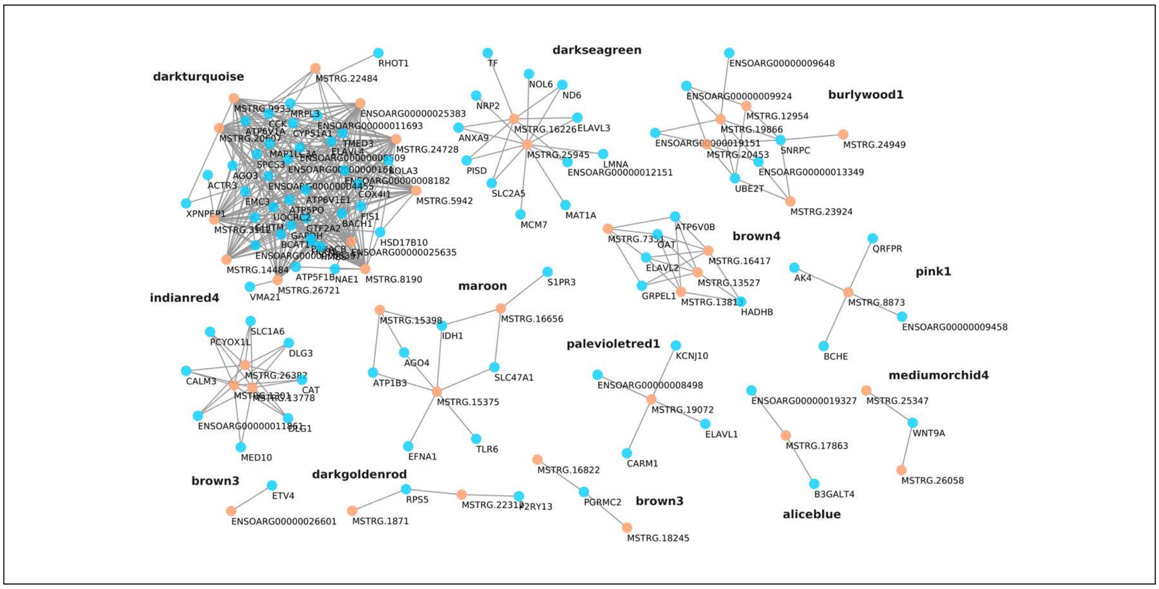


Figure 20: Interaction prediction of DE lncRNAs with correlated proteins whose mRNA genes were in the same co-expression module as the lncRNA. Interaction probability of more than 0.9 was chosen as threshold.

5.4

Discussion

In this work, the molecular signature activated in the encephalon of experimentally treated sheep has been analysed for the first time. After being inoculated with either Al hydroxide containing vaccines or an equivalent amount of Al hydroxide during 16 months, the differentially expressed mRNAs and lncRNAs were detected and functionally characterized. Previously, the transcriptome of PBMCs had been analysed at the beginning and at the end of the experiment [235]. In this study, the same group of animals was used and their transcriptomes compared with those of control animals, which only received PBS as inoculum, at the end of the experiment. Three comparisons were made with the transcriptomes: Adjuvant inoculated vs. controls, vaccinated vs. controls and adjuvant inoculated vs. vaccinated animals.

Analysis of differential gene expression from RNA-seq data identified nearly 5 times more differentially expressed genomic elements in the Adj vs control comparison than in the Vac vs. Control comparison. A very similar number of genes and lncRNAs differentially expressed was obtained in each comparison. The expression alteration of four genes that were previously described in other studies related to several neurological disorders were detected in this study, namely VCAM1, TRPM4, GDF10 and NTN1. The first three were detected as significantly upregulated in the Adj-injected sheep, while the latter was found to be upregulated in Adj vs Control and Adj vs. Vac comparisons. VCAM1 is a cellular adhesion molecule involved in the migration of immune cells across blood–brain barrier in inflammatory central nervous system diseases [309]. VCAM1 is also implicated in neuronal apoptosis and may play a role in the development of rheumatoid arthritis [310] and in the pathology of intracerebral haemorrhage (ICH) [311]. TRPM4 mediates neuronal degeneration and has been related to various neurological disorders like experimental autoimmune encephalomyelitis and MS [312]. Moreover, Li *et al.* [313] found that GDF10 was induced in peri-infarct neurons in mice, non-human primates and humans. GDF10 is considered a stroke-induced signal that promotes axonal outgrowth and enhanced functional recovery after stroke. Finally, another gene involved in blood–brain barrier integrity, NTN1, was found to be upregulated in 2 comparisons. NTN1 protects the central nervous system against inflammation.

In a recent study on aluminium accumulation in different tissues of sheep in the same experiment by means of transversely heated graphite furnace atomic absorption spectroscopy, most of the accumulation values were below 1 µg/g of aluminium in encephalon. Moreover, Al content tended to be higher in the animals of the adjuvant group compared with the control group, although without reaching statistical significance [314]. The deposits of aluminium, analysed by lumogallion technique, were cell associated and sometimes closely related to vessels. In any case, the Al deposits observed in the encephalon were lower in contrast with other tissues such as lumbar spinal cord. The limited quantity of aluminium that reached this tissue could explain the low number or differentially expressed genes, comparing with other tissues such as PBMCs.

Functional characterization of the DE-mRNAs showed that there were no overrepresented GO terms in Vac vs. Control comparison. In contrast, 27 significantly overrepresented GO terms were identified in the Adj vs. Control comparison, most of them related with the mitochondrial energy metabolism. As Aluminium is involved in the production of reactive oxygen species (ROS), it may impair mitochondrial functions [315, 316]. Changes in mitochondrial functions produce oxidative stress, leading to DNA damage and cell death. In addition, positive regulation of cardiolipin metabolic process (GO:1900210) and alpha-ketoglutarate transport (GO:0015742) GO terms were enriched in the Adj vs. Control comparison. Interestingly, cardiolipin, a phospholipid located mainly in the inner mitochondrial membrane, is associated with brain cell viability and brain homeostasis [317]. Alpha-ketoglutarate is a source of glutamate, a neurotransmitter that is involved in neurotoxicity[318] and the transport of calcium across the inner mitochondrial membrane plays an important role in neuronal physiology and pathology [319].

As far as lncRNAs expression is concerned, brain lncRNA expression is highly diverse, many lncRNA are brain-specific and some are associated with neural functions and diseases [320]. More than 3,000 candidate lncRNAs were identified in this work. Most of them presented characteristics previously described in sheep and other livestock species — poor sequence conservation, fewer exons than coding genes, diverse lengths and a majority of intergenic transcripts — even if they may vary depending on the classification methods [22, 137]. Among the few identified conserved lncRNAs, the DE TUNA, downregulated in the adjuvant group, seems an interesting element. TUNA is required for pluripotency and neural differentiation through interactions with RNA-binding proteins in its conserved sequence [321]. It regulates NANOG and SOX2 transcription factors, and FGF4 growth factor, all of them necessary for neural differentiation.

Among the candidate lncRNAs, intronic lncRNAs showed higher correlations with their closest gene and the genes that harboured intronic lncRNAs were enriched in synaptic processes. Lately, some intronic RNAs, named stable intronic sequence RNAs (sisRNAs) have been proposed as a new layer of gene regulation. They could regulate host gene expression or act as molecular sponges for miRNAs [322]. Based on GO and KEGG analysis, our data suggest that a number of intronic lncRNAs expressed in the brain may be regulating genes that act in synapses and other signalling processes, similarly to what has been proposed for brain circRNAs [323], which are also enriched in synaptic genes.

As previously described, a similar amount of DE coding genes and DE lncRNAs were detected. This feature is a sign of the importance of non-coding RNA classes in brain development, function and disease[324, 325]. Al adjuvant treatment altered the expression of several lncRNAs, which, in turn, may alter the regulation of certain genes. Since lncRNAs have been implicated in neuronal functions in diverse ways [326], we can predict potential mechanisms of action of lncRNAs. We used in silico predictors of lncRNA-protein interactions for the trans interactions. The four members of RNA-binding proteins ELAV/Hu that are mainly expressed in differentiated neurons are in the top predictions. ELAVL4, for instance, interacts with many mRNAs altering translation efficiency and stability, and is related to neuronal differentiation, self-renewal and plasticity [327]. Their activity could be altered by competing RNAs (ceRNAs) like other mRNAs or lncRNAs [328]. In fact, recent studies show that ELAVL1 interacts with several lncRNAs in mice and could have a role in neural stem cell differentiation [329].

A co-expression analysis was also performed for mRNAs and lncRNAs with WGCNA software, and 45 different modules were obtained. Interestingly, 5 of them correlated with different treatments, that is, 3 modules correlated with Vaccinated group (mediumorchid4, brown3 and palevioletred3) and 2 with Adjuvant group (maroon and burlywood1). Among them, the maroon module contained 36 DEGs and showed significant enrichments in specific KEGG pathways. Xu *et al.* [293] also found that ECM-receptor interaction, protein digestion and absorption, focal adhesion and PI3K-Akt signaling pathway were significantly enriched in the hippocampus of Al-treated rats. Among these pathways, the PI3K-Akt signaling pathway is expressed during central nervous system development [330] and it is well known that this pathway is particularly important for mediating neuronal survival, differentiation and metabolism [331]. In addition, focal adhesion and ECM-receptor interaction signalling are known to be involved in the regulation of synaptic plasticity [293, 332] and NF-κB pathway plays a crucial role on neurogenesis, cellular responses to neurological injury and neuroinflammation [333, 334]. Currently, there are few reports regarding the role that these pathways play in the neurotoxicity caused by aluminium.

Al hydroxide alone altered the expression of different mRNAs and lncRNAs important for neuronal cell survival, mitochondrial energy metabolism, metal ion balance and others associated with neurological disorders. This work is based on a long term experiment using sheep as a model. Although a considerable amount of aluminium was inoculated in a relative short period of time, the fact that certain Al salts are able to impair gene expression in a way that suggests neurotoxicity in this model should be taken into account for the production of safer vaccines.

Chapter 6

Identification of sheep lncRNAs related to the immune response to vaccines and aluminium adjuvants in PBMCs

This chapter is based on the following publication:

Bilbao-Arribas, M., Varela-Martínez, E., Abendaño, N., de Andrés, D., Luján, L. and Jugo, BM. Identification of sheep lncRNAs related to the immune response to vaccines and aluminium adjuvants. *BMC Genomics* **22**, 770 (2021)

6.1

Background

Aluminium-containing adjuvants have been used for nearly a century now both in livestock and in humans since their discovery in the early 20th century [335]. Aluminium salts such as aluminium hydroxide or aluminium phosphate are the most common compounds used as adjuvants to increase the immunogenicity of vaccines. Despite their good safety record, the mechanism of action of these adjuvants has not been fully characterised [336]. Current hypotheses include the activation of the NLRP3 inflammasome, release of DNA and uric acid danger signals, activation of the Syk-PI3K pathway and others [337], but aluminium adjuvants will most likely exert their function by multiple of these and more factors. An analysis of gene expression and proteome of Al(OH)3 treated monocytes revealed two new pathways activated by the adjuvant – IFNβ signalling and HLA class I antigen processing and presentation – and signatures of both Th1 and Th2 immune response [338].

Systems vaccinology approaches, thus application of systems biology during the development of vaccines, can be used to study the mechanism of action of adjuvants, the immune responses induced by them or, more practically, to improve the quality of vaccines [339]. Transcriptional profiles of tissues in vivo provide valuable information on the behaviour of genes after exposure to vaccines or adjuvants, including the study of non-coding transcripts, which are becoming more relevant in immunology. Recent studies have shown that lncRNAs in blood cells participate in the immune response to vaccines since the expression of several long non coding RNAs (lncRNAs) change after vaccination and correlate to antibody production [340]. In the context of sheep research, studies profiling the transcriptomic response to vaccines are scarce [341, 342], with almost none of them focusing on lncRNAs or vaccine adjuvants [234]. In human, transcriptomic studies have been used for the dissection of adjuvant mechanism of action [343, 344], and only one murine study analysed the lncRNAs induced by aluminium salts [345].

Long non-coding RNAs, defined as transcripts longer than 200 nucleotides that lack protein-coding capability and are consistently transcribed, show spatiotemporal-specific expression patterns that highlight the diverse processes in which they are involved [346]. In immune cells lncRNAs are expressed in a very cell-specific and dynamic way, even within lineages of the same cell types [347–349] and this cell-type specificity seems to be conserved among species [50]. Because of this, it is becoming apparent that lncRNAs are involved in immune system cell gene expression regulation, which should be finely regulated for the generation of a correct immunity and to avoid autoimmune responses.

Thousands of lncRNAs that may have important roles in immune processes are being described every year, but most of them remain functionally uncharacterised, especially in particular in non-human species. Many of them might simply be transcriptional noise, but several other seem to be functional [350]. In a recent collaborative project, more than the 25% of studied lncRNAs were found to affect the molecular phenotype of human fibroblasts [63]. LncRNAs do not have a single molecular mechanism. Many of the described lncRNAs function by acting as scaffolds via interactions with DNA, RNA and proteins [62]. Sometimes the act of transcription itself has a local functional output [57], which could explain the low sequence conservation of some lncRNAs. The functions of lncRNAs are generally classified as cis or trans, depending if the effect happens in a local or distant genomic region [65].

In this work, we analysed RNA sequencing data from a previous study carried out in our lab, in which it was characterised the effect of Al hydroxide adjuvant on the immune response to vaccination was characterised in a long-term experiment using sheep as a model [235] for the profiling of novel lncRNAs. We identified novel lncRNAs in sheep peripheral blood mononuclear cells (PBMCs), a subset of blood cells consisting of multiple immune cells including lymphocytes, monocytes and dendritic cells that is broadly used in infectious disease and vaccine research to get a global view of molecular and cellular events during the development of an immune response [351]. We assessed their expression kinetics along with protein coding genes (PCGs) and miRNAs by differential expression analysis and detection of co-expressed gene modules.

6.2

Methods

6.2.1. Experiment design and sequencing data

Raw data from a previous RNA-seq experiment performed by our group was analysed [235] for the detection of novel lncRNAs. All the animals used in this study were neutered male lambs of the same age without any vaccination before the experiment. The information regarding experimental design was included in [235]. In short, 14 Rasa Aragonesa lambs were divided in two treatment groups, one receiving commercial vaccines (Vac group) and the other only Alhydrogel aluminium hydroxide (Adj group), and were kept under controlled conditions for 475 days. During that time animals followed an inoculation schedule with commercial vaccines or Alhydrogel® only (Table S1).

RNA was extracted from peripheral blood mononuclear cells (PBMCs) of three animals of each group at the beginning (t0) and at the end (tf) of the treatment (Figure S1). Ribosomal RNA-depleted total RNA was sequenced in a HiSeq2000 platform with a mean sequencing depth of 70 million and 2×75 nucleotide paired-end reads at CNAG (Centro Nacional de Análisis Genómico, Barcelona, Spain). Alignment, mapping and transcriptome assembly

Quality filtering, alignment and count estimates of annotated genes was made as previously [235] and using the same parameters. In short, adaptor sequence removal and quality filtering was performed with Trimmomatic v0.36 [237], reads were mapped to the sheep genome assembly Oar_v3.1 with STAR v2.5.2b [297] and quantification of the reference transcriptome was performed with featureCounts v1.5.0-p1 [299]. For the detection of non-annotated transcripts, like most lncRNAs, it is necessary to reconstruct the transcriptome. StringTie [81] assembler was run on each sample with the reference annotation from Ensembl 95 (Oar_v3.1) and, in order to obtain a non-redundant set of transcripts, the –merge option was applied afterwards. Then, StringTie was once again applied on each sample, but with the new GTF transcript file obtained in the previous step in order to estimate transcript abundances.

6.2.2. Identification of candidate lncRNAs

GffCompare [352] software was used to classify all transcripts based on their location relative to the reference annotation. Potential lncRNAs were selected among those transcripts classified as unknown intergenic (u), fully contained within a reference intron (i) and in the opposite strand of a reference gene (x), since there is not enough evidence for other overlapping transcripts, which could arise due to errors or background noise. Potential lncRNAs were filtered by length and coding potential. First, multiexonic transcripts of less than 200 nucleotides and single-exon transcripts of less than 2000 nucleotides were filtered out. Secondly, three approaches were followed to assess the capability of the transcripts to code for proteins: Coding Potential Calculator 2 (CPC2) is a machine learning based program with a species-neutral model able to classify coding and non-coding sequences [304]. Coding-Potential Assessment Tool (CPAT) is another machine learning based program that we trained and selected the classification threshold following authors' instructions using available bovine coding and non-coding sequences [305]. HMMER 3.1b2 [306] was used to detect Pfam protein domains in our potential lncRNAs, which were translated into the three possible frames. Transcripts classified as non-coding by CPC2 and CPAT and without protein domains detected by HMMER in any frame were selected as lncRNAs.

Each of the novel lncRNAs was classified based on its position relative to its closest gene. For parsing and classification we used custom Python scripts, including the BEDTools python implementation to get the closest genes (https://github.com/daler/pybedtools). Transcription start sites (TSSs) were defined as the start or stop nucleotides, depending on strandness. Seven categories or classes were defined: (1) antisense, for those transcripts overlapping a gene in the opposite strand; (2) intronic, for transcripts fully contained within an intron; (3) intergenic, for lncRNAs at least 5 kb away from any known gene; (4) divergent, with TSSs within 5 kb and in the opposite strand; (5) convergent, with transcription stops within 5 kb and in the opposite strand; (6) sense upstream, located less than 5 kb upstream of a gene and in the same strand; and (7) sense downstream, located less than 5 kb downstream of a gene and in the same strand.

6.2.3. Sequence and synteny conservation

In order to find sequence level conservation of candidate lncRNAs, standalone Blast searches against the lncRNAs annotated in Ensembl Release 101 of four species: goat, cattle, pig and human. We libraries with lncRNA cDNA sequences for each species. We also downloaded cattle transcript sequences from NONCODE. Accounting for the low sequence conservation expected in lncRNAs, the threshold for identity was set to 50, the minimum length of the query sequence to half of the target's length, E-value of 1×10-3 and query coverage of 50%.

Synteny conservation, that is, the preservation of co-localisation of genes between different species, has been proposed as a way to deal with the low sequence conservation in lncRNAs. We downloaded from Ensembl BioMart (release 101) a custom dataset of all sheep (Oar v3.1) PCGs and their Ensembl-defined orthologues for goat (ARS1), cattle (ARS-UCD1.2), pig (Sscrofa11.1) and human (GRCh38). LncRNA annotations and cDNA sequences were also downloaded from Ensembl. Then, using a custom python script, we got the two upstream and downstream flanking orthologues for each lncRNA in the three species, which had to be located no more than 500 kb apart from it. Each sheep lncRNA was compared with all other lncRNAs. The minimum number of shared orthologues was set to two, these being the first flanking genes, and each pair of lncRNAs was scored as in the Ensembl Gene Order Conservation score. If the lncRNA was conserved in terms of synteny, an alignment was done between the novel sheep lncRNA transcript and the longest transcript of the other species' gene with the Needleman-Wunsch global pairwise alignment from EMBOSS and the longest stretch of consecutive identical nucleotides in the alignment was calculated. It is thought that even if complete sequence conservation is not the most common in lncRNAs, small functional sequences could be conserved. The analysis was also performed with the set of cattle lncRNAs in NONCODE.

6.2.4. Differential expression

The gene level expression matrix was built by keeping only the raw counts of novel lncRNAs obtained from StringTie and the count estimates of annotated genes. Before differential expression, SVA package [v3.26.0] [300] was applied to account for a known batch effect observed in the PCA analysis. After normalisation and removing of lowly expressed genes, three packages were used for differential expression: DESeq2 [113], limma [353] and edgeR [112]. Testing design included treatment, time, animal and SVA covariates, and differences were tested for the interaction of time and treatment. Thus, comparisons were made between the time points in both treatments (Vac tf vs. Vac t0 and Adj Tf vs. Adj t0) and between the treatments at the end of the experiment (Adj Tf vs. Vac Tf). The differentially expressed genes (DEGs) were selected from the intersection of the three tools of those genes with an adjusted p-value (using the Benjamini-Hochberg method) of < 0.05 and a log2 fold change (log2FC) value of > 1.

6.2.5. Gene co-expression analysis

A weighted gene co-expression network analysis was performed using the WGCNA [v1.63] R package [303]. The similarity matrix was constructed from normalised expression data using the biweight midcorrelation, a correlation more robust against outliers. Next, the adjacency matrix was defined by raising the similarity matrix to a power β=18, the minimum value required to get a scale-free topology network in our data. Modules, clusters of interconnected genes, were defined by performing a hierarchical clustering on the topological overlap measure. The minimum module size was set to 30 and modules with similar expression profiles were merged.

Once modules were defined, we looked for correlations with the treatment groups by dichotomising the groups in different combinations: samples at the beginning against samples at the end of the experiment (Treat variable), vaccine samples at the end against all other samples (TreatVac) and adjuvant samples at the end against all other samples (TreatAdj). For that purpose, Pearson correlations were generated for all pairwise comparisons of the module eigengene expression values and the treatment parameter. The eigengene is used to summarise each module with its first principal component. p-values were corrected by FDR (q-value) estimates and modules related to a variable were selected as those with a q-value < 0.05.

Every module that exhibited high correlation with a treatment or harboured many candidate lncRNAs was tested for enrichment of GO terms and KEGG pathways with gProfiler [245]. The list of all expressed genes was used as the statistical domain scope for the test and the significance threshold was set to 0.05 Benjamini-Hochberg FDR. Gene ontology term networks were created with the EnrichmentMap plugin workflow [354] for Cytoscape v3.7.1 [199] using enrichment results from gProfiler, and clusters of terms were formed by semantic similarity. Apart from enrichment analysis, the hub genes of each module were obtained by calculating the module membership (MM) and gene significance (GS) values according to WGCNA. We defined hub genes as those belonging to the $\geq$ 85th percentile for both MM and GS in each module. Those genes, including lncRNAs, are likely key drivers of expression and can give an idea about the functions or pathways of candidate lncRNAs in those modules.

6.2.6. Correlations of nearby lncRNA-PCG pairs

Candidate lncRNA-PCG pairs for cis-regulation were obtained from expression correlations between closely located pairs. Candidate lncRNAs whose TSSs were located less than 100 kb apart from the TSS of another annotated gene were selected, and the Spearman correlation was calculated between the expression profiles of both genes. Pairs with an absolute correlation R higher than 0.8 and a FDR-corrected p-value lower than 0.05 were kept.

6.2.7. Identification of potential miRNA sponges

MicroRNA expression data from the same experiment was downloaded from GEO (series GSE113897). RIsearch2.1 [93], a large-scale RNA–RNA interaction prediction tool suitable for full genome or transcriptome screening, was used to predict miRNA target sites in all the expressed transcripts. The minimum seed size was set to 6, the seed had to be within the first 8 bases of the miRNA and G-U wobbles were allowed, as proposed by the authors. Hybridization threshold was set to -15 kcal/mol. For a transcript to be classified as a potential miRNA sponge we set the minimum of 20 target sites of a single miRNA and the quantity of target sites in each transcript was averaged for visualisation at gene level. PCG, lncRNA and miRNA expression levels were normalised by TPM and Pearson correlations were performed between miRNAs and their putative sponge genes. Significant negative correlations were visualized with Cytoscape v3.7.1 [199].

6.2.8. RT-qPCR experiments

The relative quantification of 10 lncRNAs and 10 PCGs was performed by RT-qPCR using 16 different animals, 4 from each treatment group. We chose a heterogeneous set of lncRNA-PCG pairs regarding DE status and relative position of the lncRNA. They were required to be correlated at gene expression level and less than 5 kb apart. Primers were designed using PrimerQuest and OligoAnalyzer tools of Integrated DNA Technologies (IDT) (Additional file 5). GAPDH, ATPase, ACTB and G6PD were used as putative reference genes. RT-qPCR experiment was carried out using BioMark HD Nanofluidic qPCR

System technology (Fluidigm) combined with a GE 48.48 Dynamic Array integrated fluidic circuit (IFC) and the Master Mix SsoFast EvaGreen Supermix with Low ROX (Bio-Rad). RT-qPCR experiment was performed at the Gene Expression Unit of the Genomics Facility, in the General Research Services (SGIKER) of the UPV/EHU.

Analysis of amplification data was carried out using the Fludigm Real-Time PCR Analysis Software [4.1.3]. Amplification curves and melting curves were analysed to discard low quality amplifications and Ct values were corrected for efficiency differences with GenEx software of MultiD [5.4]. The stability of candidate reference genes was analysed with NormFinder and GeNorm, implemented in GenEx. G6PD and ACTB were the most stable reference genes. Relative quantification for the correlations between lncRNAs and PCGs were determined by the ΔCt method and log2 fold changes for the validation of differential expression of lncRNAs were calculated with the ΔΔCt method. Normal distribution was checked using the Shapiro-Wilk test, and because the null hypothesis was rejected, Spearman's rank correlation coefficient was used to assess the presence of significant correlation and non-parametric tests for pairwise comparisons.

6.3

Results

6.3.1. Identification and classification of lncRNAs

Unknown intergenic, intronic and antisense transcripts were filtered by length and exon count, reducing the list of potential lncRNAs from 10,340 to 4899. Transcripts were further assessed for protein coding potential, reducing the list to 2284 transcripts. These 2284 lncRNA transcripts were defined as the novel set of lncRNAs. Despite their different approaches, CPAT, CPC2 and HMMER filtered the transcripts with high overlap, with 72%, 56% and 68% of the predictions, respectively, included in the final set. Candidate lncRNAs were evenly distributed across chromosomes, with larger ones containing more transcripts (Figure 21a). Due to the 2000 nucleotide length threshold for monoexonic transcripts, 2-exon transcripts were the most numerous (Figure 21c) and showed a wider range of lengths than annotated genes (Figure 21d). Single-exon transcripts were mostly shorter than 5000 nucleotides while transcripts with more than 2 exons had diverse lengths. As for the classification of lncRNAs based on their relative location to their closest genes, the intergenic class was the most numerous (38%), followed by antisense (20%) and intronic (18%) transcripts (Figure 21b). Among those intergenic transcripts very close to an annotated gene (distance < 5 kb), we found 112 (5%) divergent lncRNAs, which are interesting because they could share the promoter with its flanking gene. PCGs were more highly expressed than lncRNAs, and mean expression levels of novel lncRNAs and annotated lncRNAs were similar (Figure 21e). These results are in concordance with some previous studies, even if due to a lack of a standardised workflow different results are obtained depending on the analyses done and applied thresholds.

We compared our shortlisted lncRNAs in PBMCs with other works in sheep that also identify novel lncRNAs by searching for transcripts that share a TSS, defined as the first transcribed nucleotide, and that are transcribed in the same direction. In brain tissue of animals from the same experiment [234] 315 transcripts (14%) shared a TSS. However, examining other works with available annotation of new lncRNA, small numbers of transcripts present in other tissues were found. Just 33 transcripts (1.44%) shared a TSS with a lncRNA from a multi-tissue catalogue [22] and 56 (2.45%) with lncRNAs from pituitary gland [156].

6.3.2. Conservation in terms of sequence and synteny

Evolutionary conservation of lncRNAs can be an indicator of function. In this way, having orthologues strengthens the evidence on sequenced transcripts, even more if the lncRNA has already been characterised in other species. As expected because of the nature of lncRNAs, few sequences had matches with other species (Figure 21f). The highest number of conserved sequences were in goat (6.67%), then cattle (4.28%), human (2.09%) and pig (1.07%). The human conserved lncRNAs included several functionally characterised lncRNAs such as CHASERR, CYTOR, CCDC26 or FTX. Just eight transcripts (0.35%) had confident matches with cattle NONCODE sequences. Note that 185 annotated sheep lncRNAs (9.96% of all annotated lncRNAs) were also detected above the minimum expression threshold in PBMCs.

In terms of gene order, more transcripts appeared to be located in conserved regions (Figure 21g), some even showing short alignments with annotated lncRNAs in the same region. We could perform the synteny analysis with roughly half of the novel lncRNAs, those surrounded with PCGs no more than 500 kb away. The 2.55% of novel sheep lncRNAs shared the same syntenic location with an annotated cattle lncRNA, and 2.19% with goat lncRNAs, a number that was higher in human (11.36%). Both sequence and conservation analyses are biased due to the vast quantity of lncRNAs annotated in the human genome (17,959) comparing with other livestock species, whose lncRNA repertoire is not fully annotated and also diverge in the quantity of lncRNA genes (1858 in sheep, 2705

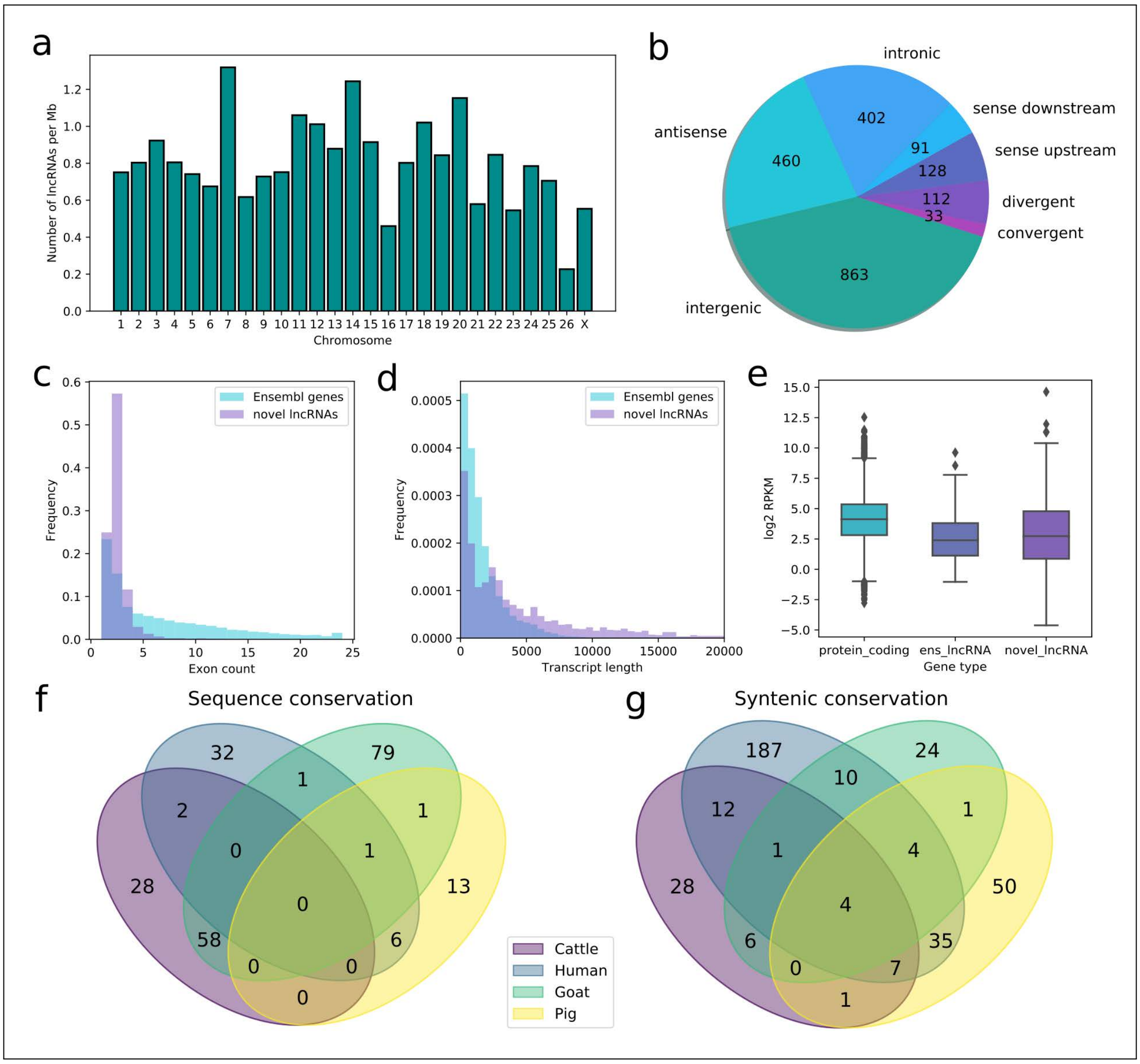


Figure 21: General characteristics of the novel lncRNAs. (a) LncRNA density per chromosome. (b) Classification of detected candidate lncRNAs by relative location to the closest annotated gene. (c) Exon number distribution in novel lncRNAs and annotated genes. (d) Transcript length distribution in novel and annotated genes. (e) Mean expression of protein coding genes, annotated lncRNAs and novel lncRNAs. (f) Novel lncRNAs conserved at sequence level comparing with selected Ensembl annotations. (g) Novel lncRNAs with conserved synteny in selected Ensembl annotations

in goat, 1480 in cattle and 6790 in pig). Because of this, when performing the same analysis with the 22,227 cattle NONCODE lncRNAs 9.93% of novel lncRNAs show syntenic conservation. Few of these lncRNAs with shared syntenic location showed short highly conserved alignments.

6.3.3. Expression analysis

In order to profile the expression of lncRNAs in the presence of aluminium adjuvants, differential expression was tested between treatment groups. The analysis was made with all annotated genes plus the newly identified candidate lncRNAs. In the same fashion as annotated genes [23], there were less DE lncRNAs in the comparison between both treatments at the end of the experiment than between each treatment at the start and end of the experiment (Figure 22). 170 lncRNAs were differentially expressed in the Adj-t0 vs. Adj-tf comparison (19 annotated and 151 candidate lncRNAs). 159 lncRNAs were differentially expressed in the Vac-t0 vs. Vac-tf comparison (11 annotated and 148 candidate lncRNAs). 65 lncRNAs were differentially expressed in the Adj-tf vs. Vac-tf comparison (4 annotated and 61 candidate lncRNAs). The expression divergence is clear when comparing time-points, while treatment-wise changes are more subtle. We found that five of the DE novel lncRNAs are conserved between sheep and human. The divergent MSTRG.24,028 lncRNA is downregulated in the Adj-t0 vs. Adj-tf comparison and is homologous to the human OTUD6B-AS1 lncRNA, which has been recently linked to regulation of apoptosis [355].

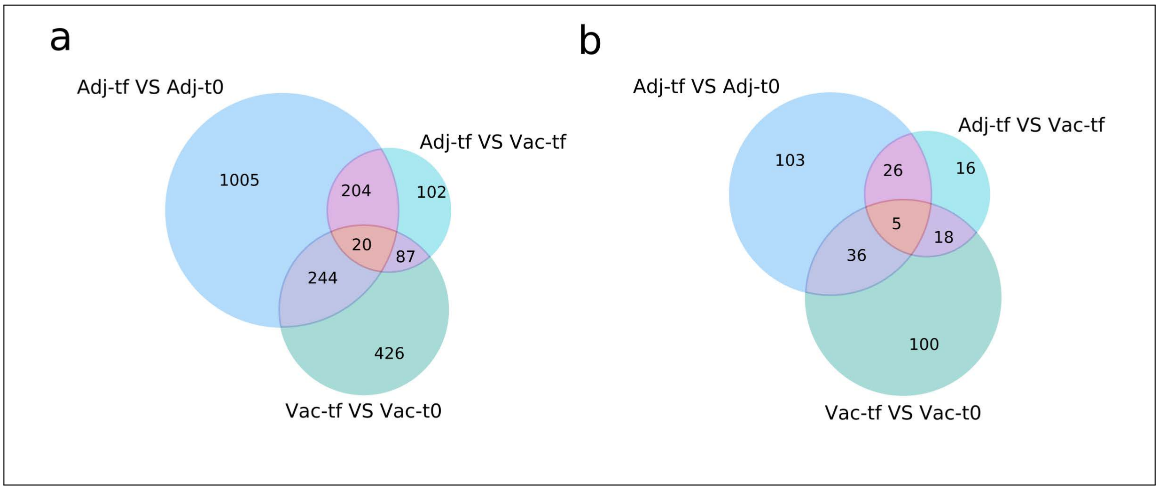


Figure 22: Venn diagrams of differential expression of coding and lncRNA genes. (a) Total differentially expressed genes. (b) Differentially expressed novel lncRNA genes. Comparisons were made between time points in vaccinated animals (Vac-tf vs. Vac-t0), between time points in adjuvant-only animals (Adj-tf vs. Adj-t0) and between the treatments at the end of the experiment (Adj-tf vs. Vac-tf)

A gene co-expression network was constructed with the same genes used for differential expression. This analysis provides valuable information about along which genes are the candidate lncRNAs expressed, and in this way, predicting their putative functions by guilt-by-association. Genes with similar expression patterns were clustered in 32 modules ranging from 39 to 1956 genes (Figure 23a). We searched for significant correlations among module eigengenes, the principal component of the genes in the module that depicts its dominant trend, and treatment parameters. 15 modules were correlated with at least one treatment: 5 modules with the adjuvant treatment, 5 modules with the vaccine treatment and 7 modules with both treatments taken together as a single group (Figure 23b).

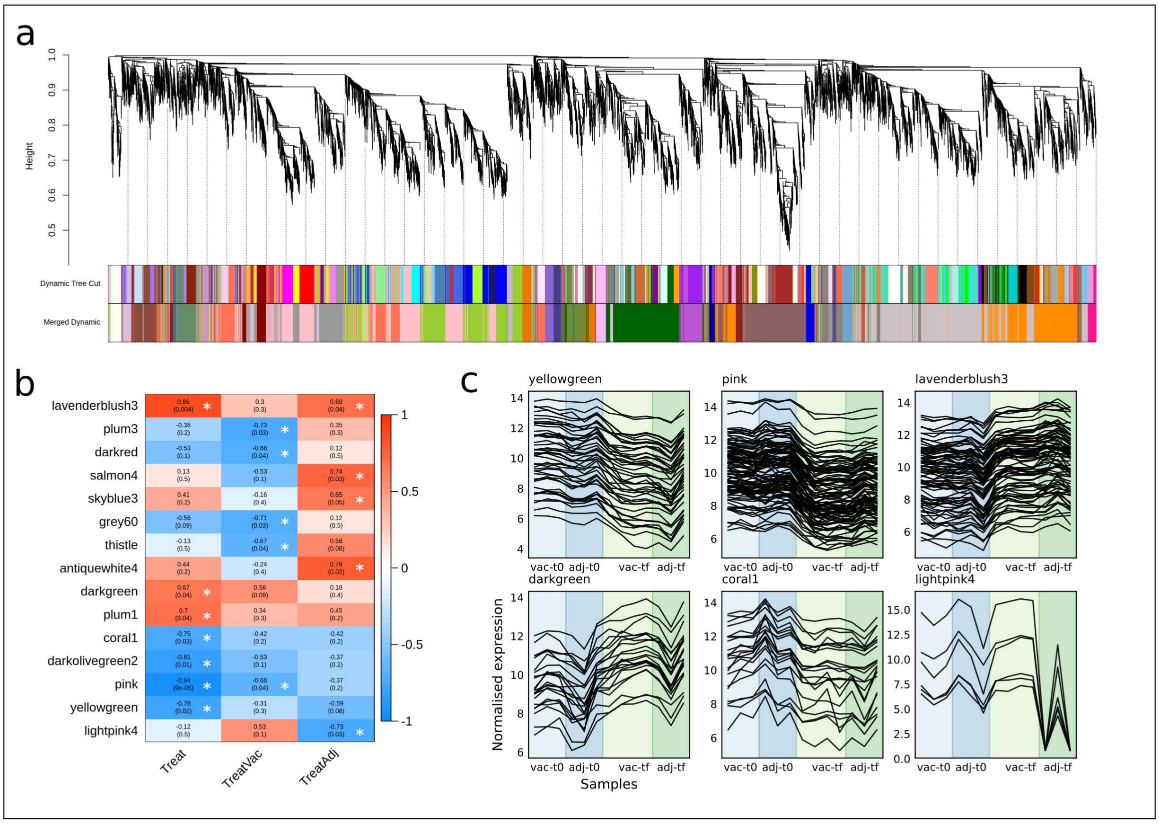


Figure 23: WGCNA co-expression analysis results. (a) Gene dendrogram obtained by average linkage hierarchical clustering. The colour bars show the module assignment before and after modules with similar expression profiles were merged. (b) Module-trait associations. Each row corresponds to a module eigengene, while columns correspond to a trait (both treatments together, vaccine and adjuvant-only). Only modules associated with at least one trait are shown. (c) Expression profiles of hub genes of modules correlated with at least one trait and that are enriched in some GO terms.

As for the module membership of candidate lncRNAs, most modules were made of both PCGs and lncRNAs, although in differing proportions. The five modules with more than 1000 genes had many co-expressed lncRNAs, while some small modules were only composed of PCGs. Integrating DE and co-expression analysis, 17 modules had DE genes within them, most of them belonging to the comparisons between time points.

Modules were characterized by gene enrichment analysis and showed involvement in distinct biological processes (Additional file 4). Some modules were not enriched in any term, mainly the smaller ones, and others were enriched in cell cycle functions or general metabolic functions. Two modules (coral1 and lightpink4) were clearly linked to the immune response with functions related to cytokines, immune cell differentiation and response to stress and external stimuli.

6.3.4. Treatment-correlated co-expression modules

Modules with significant correlations with a treatment variable were selected for further analysis, since lncRNAs in those modules are probably responding to the vaccine or adjuvants and many of them are differentially expressed. Modules whose eigengene is correlated with the treatment variable should reveal information about the general effect of aluminium on the immune response and modules whose eigengene is correlated with one of the treatments should highlight the differences between them. The expression profiles of the hub genes within each significantly correlated module show the trend of those modules across treatment groups (Figure 23c).

Among the modules correlated with both treatments at the same time, the pink module had the strongest correlation (9e-0.5 p value) and was enriched in DNA repair, methylation and general metabolic processes. Coral1 module was enriched in diverse processes such as immune response, T-helper cell functions (Th17 specifically), inflammation, cell motility or proliferation; all of these in concordance with a general response of the immune system. The yellowgreen module included genes related to the respiratory chain and cell cycle. Lavenderblush3 is highly correlated with the treatment variable, independent of its composition, and it is enriched in immune response activation, lymphocyte activation, cell cycle and metabolic processes (Figure 24).

The most prominent module correlated with a specific treatment variable was lightpink4, negatively correlated with the adjuvant treatment, suggesting a tendency for lower expression in the adjuvant group (Figure 23c). It is enriched in responses to external stimuli, cytokines and differentiation of various immune cells (Figure 24); and its expression seems to be driven by many DE genes in the Adjuvant tf vs. Vaccine tf comparison. Besides, this module includes marker genes of classical monocytes (CD14, S100A12, S100A8) and non-classical monocytes (FCGR3A) [356], possibly indicating a reduction in the monocyte lineage fraction of PBMCs in the Adjuvant tf group. S100A12 and S100A8 are known to be highly expressed in bone marrow-derived macrophages of sheep and other mammals [357]. Other abundant genes in this module are those involved in cytokine production and reception (e.g. IL6R, IL1R1, IL1R2, IRF1, PTGER4, MYD88, IL17RC, OSM, IL15RA, IL4R, CXCR1, CSF2RB, CSF3R). The genes CXCR1, CSF2RB and CSF3R are hub genes of this module.

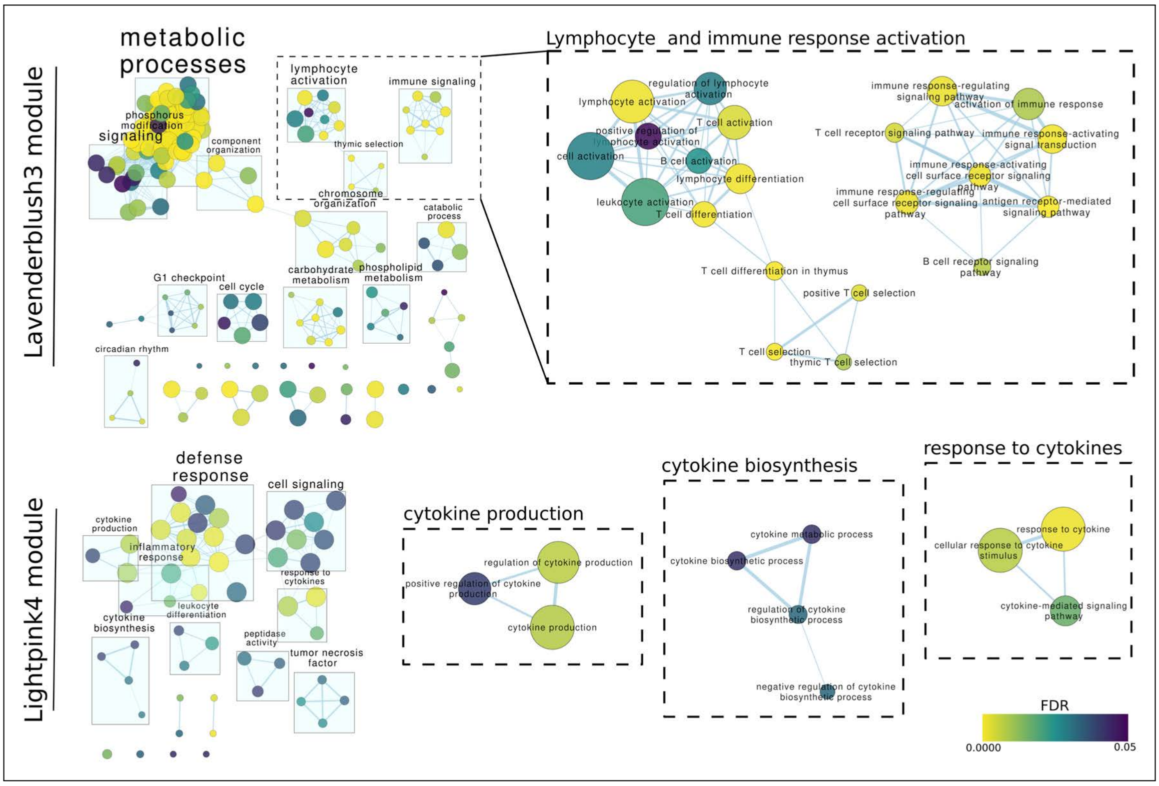

Lavenderblush3 module
metabolic processes
phosphorus modification
signaling
component organization
lymphocyte activation
immune signaling
thymic selection
chromosome organization
catabolic process
carbohydrate metabolism
phospholipid metabolism
G1 checkpoint
cell cycle
circadian rhythm
Lymphocyte and immune response activation
regulation of lymphocyte activation
lymphocyte activation
T cell activation
positive regulation of lymphocyte activation
cell activation
B cell activation
lymphocyte differentiation
leukocyte activation
T cell differentiation
T cell differentiation in thymus
positive T cell selection
T cell selection
thymic T cell selection
immune response-regulating signaling pathway
activation of immune response
T cell receptor signaling pathway
immune response-activating signal transduction
immune response-activating cell surface receptor signaling pathway
immune response-regulating cell surface receptor signaling pathway
antigen receptor-mediated signaling pathway
B cell receptor signaling pathway
Lightpink4 module
defense response
cell signaling
cytokine production
inflammatory response
response to cytokines
leukocyte differentiation
cytokine biosynthesis
peptidase activity
tumor necrosis factor
cytokine production
regulation of cytokine production
positive regulation of cytokine production
cytokine production
cytokine biosynthesis
cytokine metabolic process
cytokine biosynthetic process
regulation of cytokine biosynthetic process
negative regulation of cytokine biosynthetic process
response to cytokines
response to cytokine
cellular response to cytokine stimulus
cytokine-mediated signaling pathway
FDR
0.0000
0.05

Figure 24: Networks of enriched GO biological process functions in two trait-correlated modules: Lavenderblush3 and Lightpink4. Nodes represent GO biological process terms. Nodes are coloured by false discovery rate (FDR) and their size represents the number of genes in the module belonging to the term. Edge width represents the number of shared genes between two terms.

6.3.5. Expression of nearby PCGs and lncRNAs

Correlated lncRNA-PCG pairs were identified as a way of inferring potential cis regulation. In the RNA-seq dataset, 348 lncRNAs-PCG pairs showed correlations above the applied threshold. Most of the involved lncRNAs were sense intronic, sense upstream or sense downstream of their correlated gene, but there were 24 antisense lncRNAs, 9 divergent lncRNAs and 34 intergenic lncRNAs.

Relative expression levels of 10 pairs of correlated lncRNAs and PCGs were measured by RT-qPCR in order to validate their coordinated expression. Six differentially expressed lncRNAs and 4 non-differentially expressed lncRNAs were selected. Half of the selected lncRNAs were classified as divergent (MSTRG.9006, ENSOARG00000025373, MSTRG.17,627, MSTRG.23,098, ENSOARG00000025919), and there were two sense (ENSOARG00000026290, MSTRG.16,981), two intergenic (ENSOARG00000025821, ENASOARG00000026567) and one antisense (ENSOARG00000026120) lncRNAs. All except for the antisense one were amplified, including those that are unannotated in Ensembl and are predicted in this study. 7 out of 9 amplified lncRNAs (78%) showed significant correlations with their corresponding PCG (Figure 25).

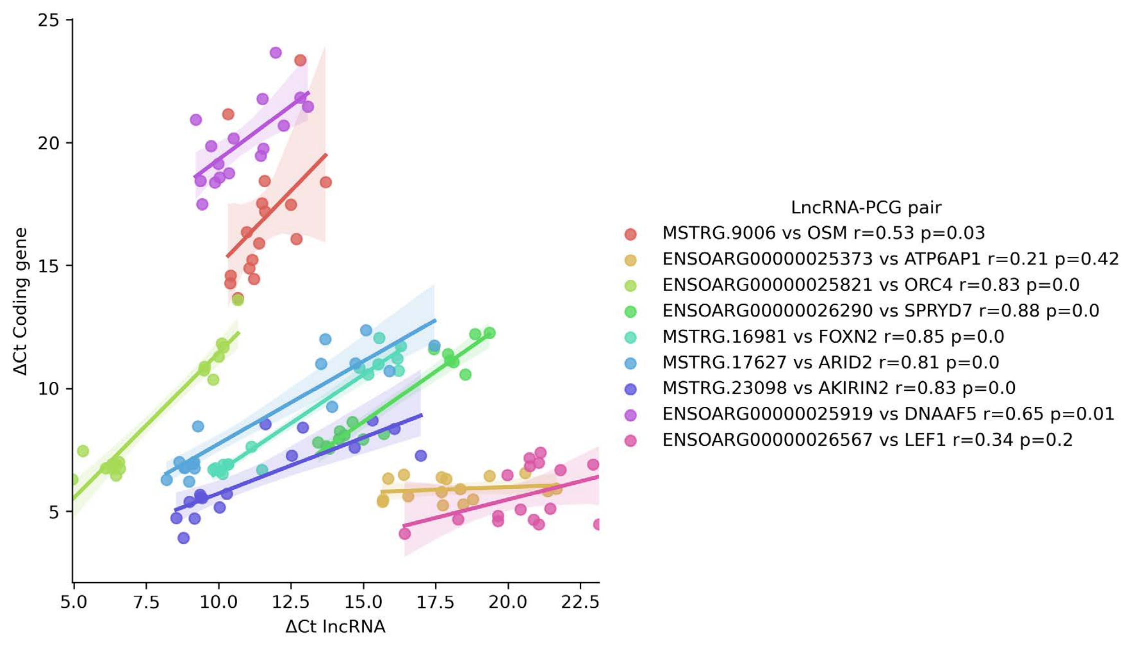


Figure 25: Expression correlations between selected lncRNA and protein coding gene (PCG) pairs assessed by RT-qPCR. Gene expression correlations were performed with efficiency corrected ΔCt values and Spearman's rank correlation

Among the studied pairs, some are interesting due to their relationship with the immune system: The gene ENSOARG00000006353, an orthologue of human and murine OSM gene, encodes for a cytokine secreted by monocytes/macrophages and T-lymphocytes, and is involved in haematopoiesis and inflammation [358]. It is divergently located to the novel monoexonic MSTRG.9006 lncRNA and both of them are differentially expressed in the vaccinated group. Another immune related gene, the transcription factor FOXN2, is correlated with the lncRNA MSTRG.16,981 located sense upstream of it and is differentially expressed in the adjuvant group. Besides, three novel lncRNAs, which were not differentially expressed in the RNA-seq dataset, showed robust correlations with coding genes ARID2, AKIRIN2 and DNAAF5 in a divergent position.

6.3.6. Novel lncRNAs as miRNA sponges

Some lncRNAs could be acting as miRNA sponges due to their high quantity of predicted miRNA binding sites. One hundred lncRNAs, 2 annotated lncRNAs and 69 PCGs had more than 20 predicted target sites for at least one expressed miRNA. 22 miRNAs were involved in those interactions. Assuming that miRNAs downregulate the expression of their targets, we calculated the expression correlations between them. 16 novel lncRNAs and 26 PCGs showed significant negative correlations with a miRNA (Figure 26). The miRNAs that target most lncRNAs are oar-let-7b and oar-miR-150. The highly expressed let-7b was upregulated in the Adj-t0 vs. Adj-tf comparison [235]. The other miRNA, oar-miR-150, was also one of the most expressed in the miRNA dataset of the same experiment [235].

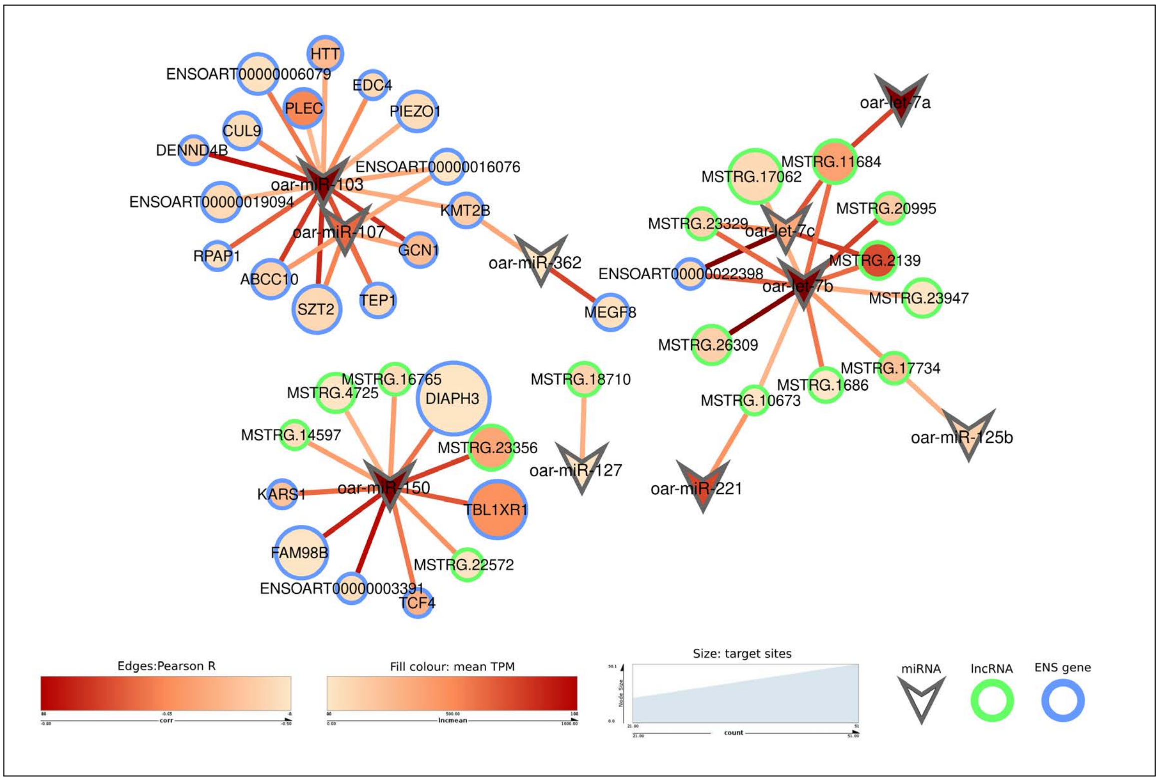


Figure 26: Network of miRNA sponge candidates. Significant negative Pearson correlations between miRNAs and target genes are depicted as edges. Size of target genes reflects the amount of target sites for a miRNA. Inner colours represent TPM expression and edge colours Pearson correlation strength (r).

6.4

Discussion

Mining lncRNAs from RNA-seq data allows the detection of large amounts of transcripts that could be classified as candidate lncRNAs. Although there was an overlap between a priori transcriptionally different tissues such as brain [234] and PBMCs of the same experimental animals, the identified lncRNAs were mostly tissue-specific, as few of them were present in other studies in sheep The newly identified lncRNAs shared similar features with those previously found in other mammal studies: lower expression than PCGs, fewer exons, limited sequence conservation and a majority of intergenic transcripts. For instance, using a multi-tissue expression dataset, 12,296 and 2657 lncRNAs with intergenic location mainly were identified in sheep and goat [22]. In a developmental tissue dataset from seven species, mostly species-specific lncRNAs were found [47]. Other sheep works analysed lncRNAs within a specific functional RNA-seq dataset and identify lncRNAs with similar characteristics [156, 157, 168, 173].

Apart from a set of highly conserved and functionally characterised lncRNAs [359], lncRNAs show low sequence conservation. Hence, some may be functionless, function by the act of transcription itself [57, 360, 361], like the bidirectionally transcribed class of eRNAs [362], or have short functional elements that escape common conservation analyses. Some of the highly conserved lncRNAs identified in this work have been experimentally tested in humans. For instance, Chaserr (LINC01578), that negatively regulates its adjacent gene CHD2, to tune its expression [363], and lnc-sox5, that promotes the expression of IDO1, which modulates T-cell behaviour [364].

A large fraction of annotated lncRNAs are divergent lncRNAs, originated upstream of an specific gene and regulated by a bidirectional promoter so they often show expression correlations with their adjacent gene, which can imply a regulatory relationship [30, 33]. Based on this statement, the function of unknown lncRNAs may be inferred from their relationship with adjacent genes. We found 112 lncRNAs which could be classified as divergent in the RNA-seq dataset. Five divergent lncRNA-PCG pairs with significant correlations were tested also by RT-qPCR. Among those pairs, the gene coding for the OSM cytokine was correlated with a 3 kb long monoexonic lncRNA not annotated in sheep. Both genes were upregulated in the vaccinated group of animals. Although pending of functional studies, this could be an example of a bidirectional promoter, known to be stronger than regular promoters [34], that increases transcription of a PCG.

To predict functions of lncRNAs, prioritise candidates and discern their transcriptional regulatory programmes a coexpression analysis network was performed, assuming that lncRNAs related to known genes are involved in the same processes or pathways. Thus, we hypothesise that differentially expressed lncRNAs co-expressed with known immune genes are more likely to be involved in immune response functions,. The gene set enrichments of co-expression modules responding to both treatments pointed to aluminium-induced inflammation, while the modules responding only to vaccines or aluminium adjuvants alone highlighted the effect of adding antigens to the adjuvant preparation, as illustrated by an immune gene-rich module with several genes involved in cytokine production and reception, and monocyte markers. This module included many novel lncRNAs, including the one divergently located to the OSM cytokine gene.

Lastly, the data sets were analyzed to investigate the interaction between two regulatory elements, lncRNAs and miRNAs. The miRNAs that target most lncRNAs were oar-let-7b and oar-miR-150. The highly expressed let-7b, being a regulator of innate immune response genes and inflammation activation [365, 366] was upregulated in the adjuvant inoculated animals [235].The second miRNA, oar-miR-150, was also one of the most expressed in the dataset [235]. It is thought to be important in the adaptive immune response due to its high expression in lymphocytes and its upregulation after vaccination [367, 368]. Thus, these lncRNAs could act as sponges by sequestrating miRNAs involved both in the innate and adaptive immune responses.

Future work should focus on annotating non-coding genes in specific immune cell types combining with functional experiments.

The lncRNA transcriptome of sheep PBMCs after multiple vaccination or adjuvant-only inoculations was analysed. More than 2000 novel lncRNAs were found, a small proportion of them being conserved across close species. Some of those lncRNAs could be involved in the immune response to vaccination and could regulate nearby immune genes although experimental work should be performed to confirm their potential regulatory functions. Moreover, both treatments induced lncRNA-containing co-expression modules, highlighting their immune response signature. At last, some lncRNAs seem to act as sponges for 2 miRNAs involved in innate and adaptive immune responses. In this case, advances in systems vaccinology can shed light on the mechanism of action of aluminium salt adjuvants, and help to understand the overall immune response to vaccines.

Chapter 7

Comprehensive analysis of ovine transcriptomic data reveals novel long non-coding RNAs related to the immune response

This chapter is based on the following manuscript:

Bilbao-Arribas, M., and Jugo, BM. Comprehensive analysis of ovine transcriptomic data reveals novel long non-coding RNAs related to the immune response.

7.1 Background

Long non-coding RNAs (lncRNAs) are a heterogeneous class of genes that transcribe transcripts longer than 200 nucleotides lacking protein-coding potential [18]. They are consistently transcribed, show lower expression, have less exons, are more enriched in the nucleus and vary in their epigenetic marks and splicing efficiency compared to protein coding genes (PCGs) [19, 21, 24]. They show spatiotemporal-specific expression and epigenetic regulation, which highlights the diverse processes in which they are involved [346, 369]. The expression of most lncRNAs varies greatly between individuals [21, 23]. In the immune system lncRNAs are expressed in a very cell-specific and dynamic way, even within lineages of the same cell types [347–349] and this cell-type specificity seems to be conserved among species [50]. Thus, lncRNAs emerge as potential regulators of immune system cell function and gene expression regulation, which should be finely coordinated for the generation of a correct immune response to external stimuli such as pathogens or vaccines.

Next-generation sequencing has expanded the mammal transcriptome attributing to thousands of poorly understood non-protein-coding transcripts the largest share of genes. There are many lncRNAs that may be involved in immune processes, but most of them remain functionally uncharacterised, especially in non-model species. Some lncRNAs might simply be transcriptional noise, but several others appear to be functional [63, 350]. LncRNAs do not have a single molecular mechanism. They can regulate gene expression through interactions with proteins, RNA or DNA and their functions can often be directed by their location, sequence or secondary structure [62]. Sometimes the act of transcription itself has a local functional output, regardless of sequence, which could explain their low sequence conservation [57, 62]. For instance, IFNG gene expression is regulated by the gene locus of an antisense lncRNA, but not by its non-coding product [60].

The lncRNA catalogues of livestock species remain under-annotated compared to the mouse or human annotations [109, 120]. Publicly available gene annotations contain more than ten thousand mouse and human lncRNA genes, while the sheep annotation contains 2229 lncRNA genes in Ensembl v.105 and 4442 lncRNA genes in NCBI Release 104. There is limited genomic overlap between both sources, most likely reflecting the highly specific expression of lncRNAs and the incompleteness of the current annotations [126]. The annotation and functional characterisation of livestock lncRNAs is essential, since most trait-associated variants in livestock lie within non-coding genome regions [126]. In sheep, lncRNAs have been profiled across a multi-tissue dataset [22], but there are few functional studies investigating their involvement in the immune response and those are difficult to compare due to differences in naming and data availability [182, 183, 370].

The exponential increase in RNA sequencing datasets in the last years offers a valuable opportunity for posing novel scientific questions or improving the statistical significance of the analyses in a cost-efficient manner [111]. This is specially suitable for the profiling of lncRNAs, due to their highly specific expression [109] and for the profiling of the gene expression signatures of immune responses [371]. There is a great interest in gene expression meta-analysis methods [372, 373], which have been successfully applied to profile the transcriptional signatures across respiratory viruses [374] or vaccines [375] in human blood samples.

In this study, we take advantage of the increasing number of high-throughput functional experiments deposited in public databases in order to uniformly analyse, profile unannotated lncRNAs and integrate 422 publicly available ovine RNA-seq samples, histone modification CHIP-seq samples and CAGE-seq samples of blood cells, lymphoid organs and other immune cells. We expand the lncRNA catalogue in sheep and identify the common expression signature of protein coding genes and lncRNAs during the immune response, evidencing the potential role of hundreds of lncRNA genes in immune functions.

7.2

Methods

7.2.1. Data collection

We selected 929 RNA-seq sequencing runs belonging to 15 BioProjects from NCBI Sequence Read Archive (SRA), which were merged into 422 samples, by the following criteria: Samples from an immune system tissue (blood, immune cells or lymphoid organs), at least five samples from a single BioProject, pair-end sequenced using an Illumina platform and genome mapping rate above 60%. Sample metadata such as tissue type, age, breed, sex, library type or experimental treatment was collected from NCBI databases and published articles. Due to metadata ambiguity, the strandness of the samples was assessed with Kallisto [376] prior to pipeline execution.

Most samples originated from functional experiments that studied the immune response to vaccines or vaccine components (PRJEB26387, PRJNA454435, PRJNA559411), helminth infections (PRJNA291172, PRJNA433706, PRJNA268183, PRJEB33476, PRJEB45790, PRJEB44063), bacterial infection (PRJEB15872) and pro-inflammatory gene upregulation (PRJNA631066). Other transcriptomic studies were not related to the immune response but were used to improve the novel lncRNA identification and as unstimulated controls (PRJNA528905, PRJNA485657, PRJNA362606). Besides, we added samples from the sheep expression atlas (PRJEB19199), including samples from bone marrow derived macrophages stimulated with LPS. All samples were dichotomized into two groups: samples from immune-stimulated animals and unstimulated or control samples.

7.2.2. Transcriptome assembly and quantification

We downloaded and analysed the 422 RNA-seq samples with a uniform workflow using custom Snakemake v.6.15.1 [377] pipelines (Figure 27). 375 reverse stranded samples were used for transcriptome construction and novel lncRNA identification, while all samples were used for quantification based on the new transcriptome. Sequencing runs were downloaded from NCBI SRA with the SRA Toolkit and were merged into samples by their experiment ID. Adapter trimming and quality filtering was performed with cutadapt v.3.5 [378]. Reads were aligned to the sheep reference genome (Oar_rambouillet_v1.0) with STAR v.2.7.3a [297] guided by the Ensembl (v102) annotation. StringTie2 v.2.0 [379] transcriptome assembler was used to reconstruct the transcriptome of each individual sample guided by the Ensembl (v102) annotation and with the –rf option. Then StringTie2 was applied again with the –merge option using all the transcriptomes in order to obtain a non-redundant transcriptome that is comparable between samples.

Quantification of gene expression was performed at transcript level with Kallisto v.0.48 [376] pseudoaligning the trimmed reads of all samples to the newly generated transcriptome, generated with GffRead v.0.11.7 [352]. The –rf-stranded option was used with the 375 stranded samples.

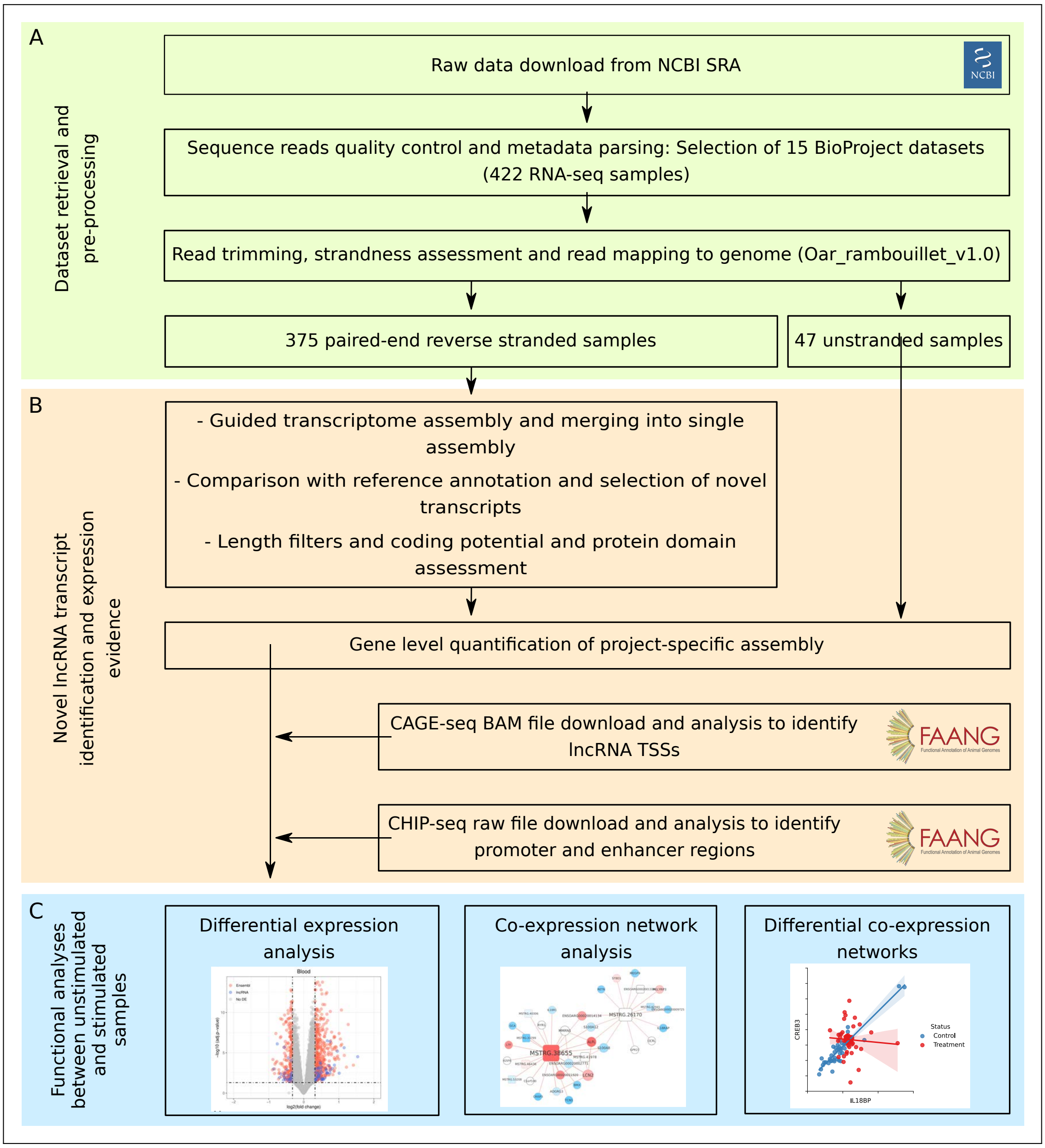

A
Dataset retrieval and pre-processing
Raw data download from NCBI SRA
NCBI
Sequence reads quality control and metadata parsing: Selection of 15 BioProject datasets (422 RNA-seq samples)
Read trimming, strandness assessment and read mapping to genome (Oar_rambouillet_v1.0)
375 paired-end reverse stranded samples
47 unstranded samples
B
Novel lncRNA transcript identification and expression evidence
- Guided transcriptome assembly and merging into single assembly
- Comparison with reference annotation and selection of novel transcripts
- Length filters and coding potential and protein domain assessment
Gene level quantification of project-specific assembly
CAGE-seq BAM file download and analysis to identify lncRNA TSSs
FAANG
Functional Annotation of Animal Genomes
CHIP-seq raw file download and analysis to identify promoter and enhancer regions
FAANG
Functional Annotation of Animal Genomes
C
Functional analyses between unstimulated and stimulated samples
Differential expression analysis
Blood
log2(fold change)
Co-expression network analysis
MSTRG.26170
MSTRG.38655
Differential co-expression networks
CREB3
IL18BP
Status
Control
Treatment

Figure 27: Bioinformatic workflow of the study. The workflow followed in this study can be divided into three sections. (A) First, sequencing data retrieval, preprocessing and mapping to the sheep genome. (B) Second, identification of unannotated lncRNA transcripts and evidence of expression. (C) Third, functional analyses between unstimulated samples and samples with an immune stimulation.

7.2.3. LncRNA identification

Potential novel lncRNAs were defined as unannotated transcripts that were located either in an intergenic region, in an intron of a known gene or in the antisense strand of a known gene. GffCompare v.0.11.2 [352] was used to compare the newly assembled transcriptome with the reference annotation and extract these transcripts. Single-exon transcripts longer than 500 nucleotides and shorter than 10kb, and multiexonic transcripts longer than 200 nucleotides and shorter than 50kb were kept. The assessment of the coding potential of the candidate transcripts was done with three different tools. The coding potential prediction module of FEELnc [380], based on a Random Forest classifier, was trained with sequences of bovine coding genes and lncRNAs from NONCODE database [118]. Coding-Potential Assessment Tool 3.0.2 (CPAT) [305] is a logistic regression-based tool that we trained and selected the classification threshold following authors' instructions using the same bovine coding and non-coding sequences. HMMER 3.3.2 [306] was used to detect Pfam protein domains in our potential lncRNAs, which were translated into the three possible frames. Transcripts classified as non-coding by FEELnc and CPAT and without protein domains detected by HMMER were kept. Transcripts classified by CuffCompare as a novel isoform of a known gene were also kept, as transcripts that had passed the coding potential tests could be legit non-coding isoforms. The selected transcripts were defined as the final set of novel lncRNAs.

Novel lncRNA transcripts were classified with a custom Python script (see Data Availability section) based on their position relative to their closest gene. Transcription start sites (TSSs) were defined as the start or stop nucleotides, depending on strandness. Seven classes were defined: 1) antisense, for those transcripts overlapping a gene in the opposite strand; 2) sense intronic or antisense intronic, for transcripts fully contained within an intron; 3) intergenic, for lncRNAs at least 5kb away from any known gene; 4) divergent, with TSSs within 5kb and in the opposite strand; 5) convergent, with transcription stops within 5kb and in the opposite strand; 6) sense upstream, located less than 5kb upstream of a gene and in the same strand; and 7) sense downstream, located less than 5kb downstream of a gene and in the same strand.

To compare the novel lncRNAs with the recently upgraded ovine NCBI RefSeq annotation (release 104), which is based on the ARS-UI_Ramb_v2.0 new reference genome [381], transcript coordinates were remapped with the NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap). They were compared with the NCBI lncRNAs using GffCompare [352]. Transcripts models with codes "=", "j", "c", "k", "o","m" or "n" were considered as overlapping, transcripts with codes "c" or "k" were considered compatible isoforms and transcripts with code "=" were considered exact matches.

7.2.4. CAGE-seq and CHIP-seq data analysis

We downloaded the mapped BAM files of CAGE-seq experiments of five immune tissues from a multi-tissue project of sheep TSSs (tonsil, alveolar macrophages, spleen, mesenteric lymph node and prescapular lymph node) [251] and analysed them using the same pipeline as the authors, with some modifications. In short, downloaded BAM files were converted to bigwig format with bedtools v.2.30.0 [241] and BedGraphToBigWig from UCSC tools [382]. The R package CAGEfightR v.1.12.0 [383] was used for normalization and clustering of CAGE tags. CAGE tags <10 read counts were removed and all the tags from any of the tissues were kept, to include tissue-specific TSSs. CAGEfightR was also used to identify bidirectional clusters. In order to get the genes supported by CAGE-predicted TSSs we used the BedTools python implementation pybedtools v.0.8.1 [384] to search for TSSs from the assembled transcriptome within 0.5 kb from them, accounting for strandness.

Sheep ChIP-seq sequencing files from alveolar macrophages [385] were downloaded from the NCBI Sequence Read Archive (SRA) and were analysed in an uniform way. Reads were aligned to the sheep genome (Oar_rambouillet_v1.0) with Bowtie2 v.2.3.5.1 [386]. SAM files were converted to BAM format with samtools v.1.7 [387], and were sorted, filtered for quality and removed duplicate reads with sambamba v.0.6.6 [388]. MACS2 v.2.2.6 [389] was used to call narrow peaks for histone modifications with a FDR cut-off of 0.05 and consensus peaks from the pairs of animals were obtained with bedtools v.2.30.0 [241]. In order to get the genes supported by CHIP-seq peaks we used pybedtools v.0.8.1 to search for TSSs from the assembled transcriptome within 0.5 kb from them.

7.2.5. Conservation in terms of sequence

Sequence level conservation was performed with standalone BLASTn (BLAST v.2.9.0) [240] by aligning the sheep lncRNA transcripts against the lncRNAs annotated in Ensembl Release 106 from five species: goat, cattle, pig, mouse and human. Because of the known low sequence conservation expected in lncRNAs, results were filtered by identity > 50, query coverage > 50, E-value > 1e-05 and it was required that the length differences between each pair of sequences was less than 50%. Visualization of the genomic context of conserved lncRNAs was performed with pyGenomeTracks 3.7 [390]. The tracks for CAGE-seq data were constructed by merging all BAM alignment files with samtools [387] into a single file and then was converted to bigwig format as previously. The tracks for histone modification CHIP-seq data were the consensus peaks obtained from MACS2.

7.2.6. Analysis of gene expression

Kallisto abundance estimates were imported to R and summarized to gene level with IsoformSwitchAnalyzeR [391] in order to set confident gene identifiers for ambiguous transcripts. Counts of annotated genes and novel lncRNA genes were kept for further analysis, discarding potential novel unannotated coding genes. For gene expression data exploration, we normalized the estimated gene counts with the variance stabilizing transformation from DESeq2 [113] and filtered out genes with less than 0.5 TPM. The first two components of the principal component analysis (PCA) and the two first dimensions of the t-Distributed Stochastic Neighbor Embedding (t-SNE) were used for visualization. LncRNAs were tagged as expressed if they could be detected above 0.1 TPM or 1 TPM in at least 20% of the samples in a tissue group. Two-sided Mann-Whitney U tests were performed to compare expression means between classes.

Differential gene expression was performed with DESeq2 [113] using the estimated counts of annotated genes and lncRNA genes expressed in at least half of each sample groups and exported from IsoformSwitchAnalyzeR [391]. Differential expression was tested separately in blood and cell samples combined on one side, and samples from lymph nodes on the other, because there were not stimulated samples from other lymphoid organs and that would unbalance the dataset. The Wald test was applied between unstimulated samples and stimulated samples using the effect of the interaction of tissue type and BioProject IDs as covariates for the lineal regression model, as those were the main drivers of the groupings seen in the exploratory analysis. Log2 fold change (log2FC) values from lowly expressed and highly variable genes were shrunken using the apeglm method [392]. Genes with an FDR-adjusted p-value lower than 0.05 and an absolute log2FC higher than 0.32, which corresponds to a 20% expression change, were kept. The relatively low log2FC filter was chosen because the large number of samples and the heterogeneity of the dataset produced differentially expressed genes with modest effect sizes and robust p-values.

Gene set enrichment analysis of differentially expressed genes was done with gProfiler R package [245]. The statistical domain scope used was the list of all expressed genes for each tissue, in order to reduce the tissue type specific expression bias. Benjamini-Hochberg FDR correction was applied to the p-values and the threshold was set to 0.05.

7.2.7. Co-expression analyses

Co-expression analyses were performed in both tissue groups separately. Genes expressed in less than half of the samples were removed and strong outlier samples were removed in order to get a better fit to a scale-free topology. We tested two network construction pipelines: 1) The pipeline proposed by the authors of GWENA [393], which consists of applying the variance stabilizing transformation (VST) from DESeq2 [113] and using spearman correlations, and 2) counts adjusted with TMM factors followed by asinh transformation, Pearson correlations and network transformation by context likelihood of relatedness (CLR) [394]. Before creating the correlation matrices, normalised gene expression was corrected for covariates with limma's removeBatchEffect function [353] to account for the effect of the interaction of tissue type and BioProject ID, as those were the main drivers of the groupings seen in the exploratory analysis. The 30% less variable genes were removed for network construction. Co-expression networks were constructed with GWENA [393] R package, which implements the WGCNA [303] R package.

Modules of co-expressed genes were detected with the threshold power and clustering threshold calculated by GWENA and a minimum module size of 30. Modules were merged if their eigengene, the first principal component of the module, correlation was higher than 0.9. Modules were associated with overall immune stimulation or specific stimulation types by correlating their eigengene to those variables. To calculate the correlation p-value threshold, we generated 1000 random gene modules ranging from 30 to 1000 genes, correlated their eigengenes with the treatment variable and calculated the false positive rate (FPR). The p-value threshold with the FPR lower than 0.05 was 1e-02, but 1e-03 was chosen for more robustness. The genes in each module were tested for Gene Ontology (GO) term enrichment with gProfiler [245] R implementation, setting the statistical domain scope to all the genes in the co-expression network and a FDR-adjusted P value threshold of 0.05.

The differential co-expression analysis was carried out by calculating the spearman correlations between all genes used in the co-expression network analysis separately in the unstimulated and the stimulated samples. The z-score method implemented in the dcanr v.1.12.0 R package [395] was used for testing the statistical differences between z-transformed correlation coefficients in both conditions. P values were adjusted for multiple hypothesis testing in order to select differentially correlated gene pairs. Differential co-expression networks (DCN) were visualized in Cytoscape v.3.8.2 [199] by integrating the differential co-expression results, co-expression modules and differential expression results. For visualization, genes without gene names in the Ensembl annotation were named after their human orthologue according to Ensembl Compara.

7.3

Results

7.3.1. Dataset description

We collected and analysed 422 publicly available RNA-seq samples of tissues related to the immune system (Table 6, Supplementary data 1) using a uniform pipeline (Figure 27). In terms of immune response induction, 49.1% of the samples had been stimulated in some way. Blood samples, as whole blood or PBMCs, represented the 64.5% of the dataset, organs and lymph nodes the 30.3% and immune cell subsets the 5.2%. The mean age of the animals was of 1.32 years and 60.1% of the samples came from male sheep. There are 12 different breeds in the dataset, with three of them being crossbreed. Library selection is an important factor for lncRNA profiling because there are transcripts that are not polyadenylated. Around half of the samples were polyA-selected and half of the samples sequenced total RNA. Besides, samples had an average of 45 million reads, summing around 19 billion reads in total (Supplementary Figure 1). Unique genome mapping rate with STAR was of 84.6% on average and pseudoalignment rate to the new transcriptome with Kallisto was of 84.7% on average (Supplementary Figure 1). The assembled and merged transcriptome annotation contained 308750 transcripts, of which 41638 were from annotated transcripts and 36067 were from novel lncRNA transcripts identified by our pipeline. These transcripts correspond to 63364 genes, including 25472 annotated genes and 21223 novel genes with at least one lncRNA isoform.

All samples were clustered based on gene expression to assess the coherence of the data. Both clustering methods used clustered together the samples based on tissue, although intra-tissue groupings were influenced by the source project (Figure 28A-B). This could be expected as each study was performed in different conditions, with different breeds, ages, sex and protocols. Immune stimulation status did not affect much the clustering probably for the same reasons and because of the strong influence of tissue type.

Table 6: Summary of samples and publications retrieved for this work.

BioProject	Samples	Tissue	Reference
PRJEB26387	72	PBMCs	[396]
PRJNA291172	36	PBMCs	[397]
PRJNA454435	13	PBMCs	[235]
PRJNA631066	6	PBMCs	[398]
PRJNA433706	5	PBMCs	[399]
PRJEB45790	48	PBMCs	[400]
PRJNA559411	76	Blood	[342]
PRJNA528905	10	Blood	[401]
PRJEB15872	16	Ileo-caecal valve lymph node	[402]
PRJNA485657	23	Spleen, tonsil, lymph node	[403]
PRJNA268183	20	Abomasal lymph node	[404]
PRJEB33476	12	Abomasal lymph node	[405]
PRJNA362606	6	Spleen	[406]
PRJEB44063	19	Hepatic lymph node	[407]
PRJEB19199	60	Several tissues	[134]

7.3.2. Novel lncRNA identification

We identified 21223 novel lncRNA genes from the sheep immune system samples that were assembled, and another 1724 annotated genes had novel non-coding isoforms classified as lncRNAs by our pipeline. Most of the novel genes with transcripts fulfilling the requisites to be classified as novel lncRNAs had all of their isoforms classified as such (17605). Some of the newly assembled gene models were coding genes missing from the Ensembl annotation that had non-coding isoforms, because they had novel transcripts with coding potential as well as lncRNA transcripts. Those unannotated genes and the 1724 annotated genes with novel non-coding isoforms were not considered as lncRNA genes for the gene-level expression analyses, even if individual transcripts could not be discarded as bona fide non-coding isoforms. We applied the same coding potential assessment methods used for novel transcripts to the annotated lncRNAs and discovered that there were transcripts potentially coding for a protein. This bias should be taken into account when comparing between the features of annotated and unannotated lncRNAs.

Regarding the characteristics of the novel lncRNAs, novel transcripts were shorter than the 2229 lncRNAs annotated in Ensembl (Figure 28C) and a great proportion of them had 2 exons, in contrast to the Ensembl lncRNAs, which are monoexonic or have more than 5 exons (Figure 28D). We classified the novel genes based on position relative to known genes (Fig. 2F, Supplementary Data 2). Intergenic lncRNAs (lincRNAs) were the most prevalent with 37% of the transcripts, followed by intronic antisense (22%) and antisense (18%) transcripts. Among the transcripts adjacent to annotated genes, the class of divergent lncRNAs was predominant (10% of all novel genes). The TSS of this kind of lncRNAs are very close to another gene's TSS, which indicates that they probably arise from a single bidirectional promoter and may have implications in terms of gene expression regulation.

We explored the sequence-level evolutionary conservation of lncRNAs with other mammal species. Most lncRNAs are known to be poorly conserved in terms of sequence, but by detecting mammalian orthologues we provide further strength to the methods by which all unannotated lncRNAs have been identified. This analysis found a small number of conserved lncRNAs (Supplementary Figure 2, Supplementary Data 3). The biggest fractions of lncRNAs with conserved orthologues were found when comparing with goat and cattle lncRNA catalogues, with 11.9% and 7.3% of transcripts with significant hits, respectively. Comparing with the human and mouse catalogues, we found much less conserved lncRNAs. Interestingly, around 3% of novel lncRNAs, corresponding to 746 unique transcripts, matched with 392 unique human lncRNAs. Among these conserved lncRNAs, widely characterized lncRNAs such as MALAT1, NEAT1, XIST, PACERR or FIRRE were successfully detected in sheep (Supplementary Figure 3). Other conserved lncRNAs were those located in Hox gene loci, such as HOTAIR, HOXA10-AS, HOXA-AS2 or HAGLR. Divergent lncRNAs were also among the conserved ones, like FMNL1-

DT, TOB1-AS1, EMSY-DT, RIPK2-DT, ATP8A1-DT or MAPK6-DT. Despite not showing enough sequence similarity, we found some sheep transcripts located in the same divergent promoter as their human counterparts, for instance the putative orthologues of HEATR6-DT or NIPBL-DT.

Because of the recent improvement of the ovine NCBI reference genome and annotation [381], the NCBI RefSeq lncRNA annotation was compared with the novel lncRNAs. After remapping to the new genome, out of the 4442 NCBI lncRNA genes, 1961 (44%) overlapped with an unannotated lncRNA. Exact matches of intron chains occurred in 571 transcripts, 238 transcripts were intron-compatible but differed in exon number and 3679 where multi-exonic transcripts with at least one intron match. The overlap between Ensembl and NCBI lncRNAs was virtually inexistent. Thus, we detected around half of the annotated NCBI lncRNA genes using only immune-related tissues, even if most of the transcript models diverged in terms of splice-junctions.

7.3.3. Expression patterns of lncRNAs

Expression levels of the novel lncRNAs detected in this study were lower than both protein coding genes and other annotated lncRNAs in the two main tissue categories (Figure 28E). In fact, after applying a minimum expression threshold in each tissue, expressed in at least 20% of the samples with 1 TPM, we were left with 2267 expressed novel lncRNAs. Besides, we also detected 482 annotated lncRNAs above the expression threshold. Interestingly, 70% of the lncRNAs annotated by Ensembl were expressed in all three main tissue categories, while only 15% of novel lncRNAs were expressed in the three tissues (Supplementary Figure 4). Setting a less stringent mean expression threshold of 0.1 TPM results in 10045 expressed novel lncRNAs, 28% of them in all three tissues. Most of the novel lncRNAs (87%) and annotated lncRNAs (93%) could be detected in the set of lymphoid organs. Overall the overlap was greater between the blood samples and "immune cell" samples for both lncRNA genes and protein coding genes, as blood contains most of those cells (Supplementary Figure 4).

The amount of detected lncRNAs in each sample significantly correlated with sequencing depth for both unannotated lncRNAs (Pearson r = 0.75) and annotated lncRNAs (Pearson r = 0.85). Expression of PCGs was also correlated (Pearson r = 0.58) with sequencing depth but the saturation curve showed a flatter slope, meaning that it saturated earlier than lncRNAs (Supplementary Figure 5). The amount of lncRNAs expressed above 0.1 or 1 TPM got saturated above around 50 million reads, while the overall amount of expressed lncRNAs at any level did not saturate even at the highest sequencing depths in the dataset (above 100 million reads) (Figure 28G).

Divergent lncRNAs showed greater expression levels than other lncRNAs classes such as intergenic lncRNAs (Mann-Whitney U test P-value 2.9e-10) or antisense lncRNAs (Mann-Whitney U test P-value 1.3e-03), and only showed significantly lower levels

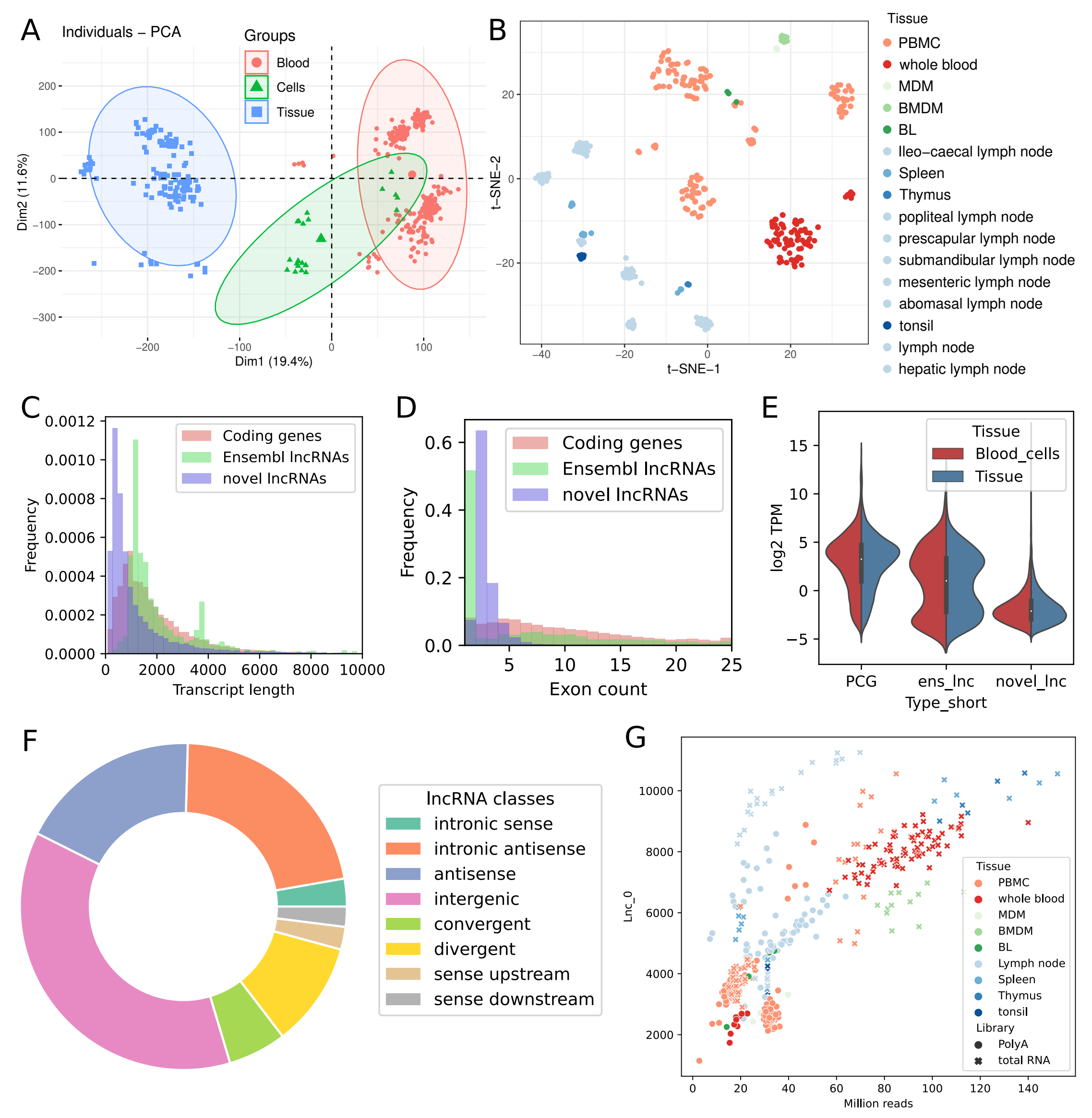


Figure 28: Characteristics of the dataset and the identification of lncRNAs. Exploratory analysis of all the samples included in the study using dimensionality reduction methods: (A) Principal Component Analysis (PCA) grouped by main tissue, (B) t-SNE plot with samples colored by tissue. (C) Transcript length distribution of PCGs and lncRNAs. (D) Exon length distribution of PCGs and lncRNAs. (E) Expression levels of PCGs and lncRNAs in blood cell samples and tissue samples. (F) Classification of lncRNAs into classes by genomic location. (G) Number of detected unannotated lncRNAs against sequencing depth.

than convergent lncRNAs (Mann-Whitney U test P-value 2.4e-03) (Supplementary Figure 6). Intronic antisense lncRNAs showed consistently lower expression than the rest of novel lncRNAs classes, in contrast with convergent lncRNAs, which were significantly more expressed than all other classes.

7.3.4. Evidence of transcription by CAGE assays and histone modifications

Independent datasets of cap analysis of gene expression sequencing (CAGE-seq) and chromatin immunoprecipitation sequencing (CHIP-seq) of histone modifications were used in order to provide evidence of lncRNA transcription at RNA and DNA level. The CAGE-seq dataset contained samples from various lymphoid organs and alveolar macrophages, so it was used to provide support of expression in two sample subsets, blood and other immune cells, and lymphoid tissues. We obtained over 2 million significant CAGE peaks and around 30 thousand bidirectional CAGE peaks present in any of the five tissues.

In both sample subsets PCGs were more strongly associated with CAGE peaks than lncRNA genes, but reducing the analysis to the genes expressed above 1 TPM instead of 0.1 TPM increased the support in all gene types (Figure 29). This increase in support specially happened in lncRNAs. 64% and 50% of the TSSs of novel lncRNAs expressed above 1 TPM in the blood subset and the lymphoid subset, respectively, were located within500 bp of a CAGE peak. LncRNAs annotated by Ensembl reached a support level comparable to that of PCGs, with more than 90% of supported TSSs at 1 TPM. Bidirectional CAGE tag clusters are usually used to identify active enhancers because it is known that bidirectional transcription of short transcripts, known as enhancer RNAs (eRNAs), is a hallmark of enhancer activation. Considering the genes expressed above 1 TPM or 0.1 TPM, novel lncRNAs were slightly less enriched in bidirectional clusters than PCGs. Around 11% and 8% of novel lncRNAs in the blood and tissue datasets, respectively, were transcribed from bidirectional sites (Figure 29). Some of them could be enhancer associated non-coding transcripts while others are divergent lncRNAs.

As for the CHIP-seq data, we analysed two histone modifications that are relevant for lncRNA transcription from a published dataset: H3K4me3, associated with promoters, and H3K27ac, associated with active enhancers and promoters. The trend of H3K4me3 peaks from alveolar macrophages and CAGE peaks were similar regarding the genes expressed in blood and other immune cells, but the overlap between histone CHIP-seq data with TSSs randomly located in the genome was much lower (Figure 29). PCGs had the highest proportion of these promoter-associated marks followed by annotated lncRNAs and novel lncRNAs. Nevertheless, regarding the H3K27ac modification, the difference between lncRNAs and PCGs was smaller, which reflects the origin of many lncRNAs from enhancer-like regions. The support from this modification was similar in novel lncRNAs and annotated lncRNAs. 20% of the TSSs of novel lncRNAs expressed above 0.1 TPM in the blood subset were associated with H3K27ac. The apparent higher support for annotated non-coding models is probably linked with their misannotation.

Providing additional evidence of the transcription of novel transcripts assembled from short-read RNA-seq reads ensures that the detected genes are reproducible. We selected 12302 assembled gene models as bona fide lncRNA genes, those which were supported by at least one of the following: CAGE tags, histone modification CHIP-seq peaks or expressed above 0.1 TPM in at least 20% of the samples in a tissue group (Supplementary Figure 10A). In this set, 47% of the lncRNAs had at least support from CAGE peaks or histone modifications. Around 1000 lncRNAs were supported by all assays, including both histone modifications. The annotation files with all unannotated lncRNA transcripts, the set of high confidence transcripts and expression values can be found in a public repository (see Data availability).

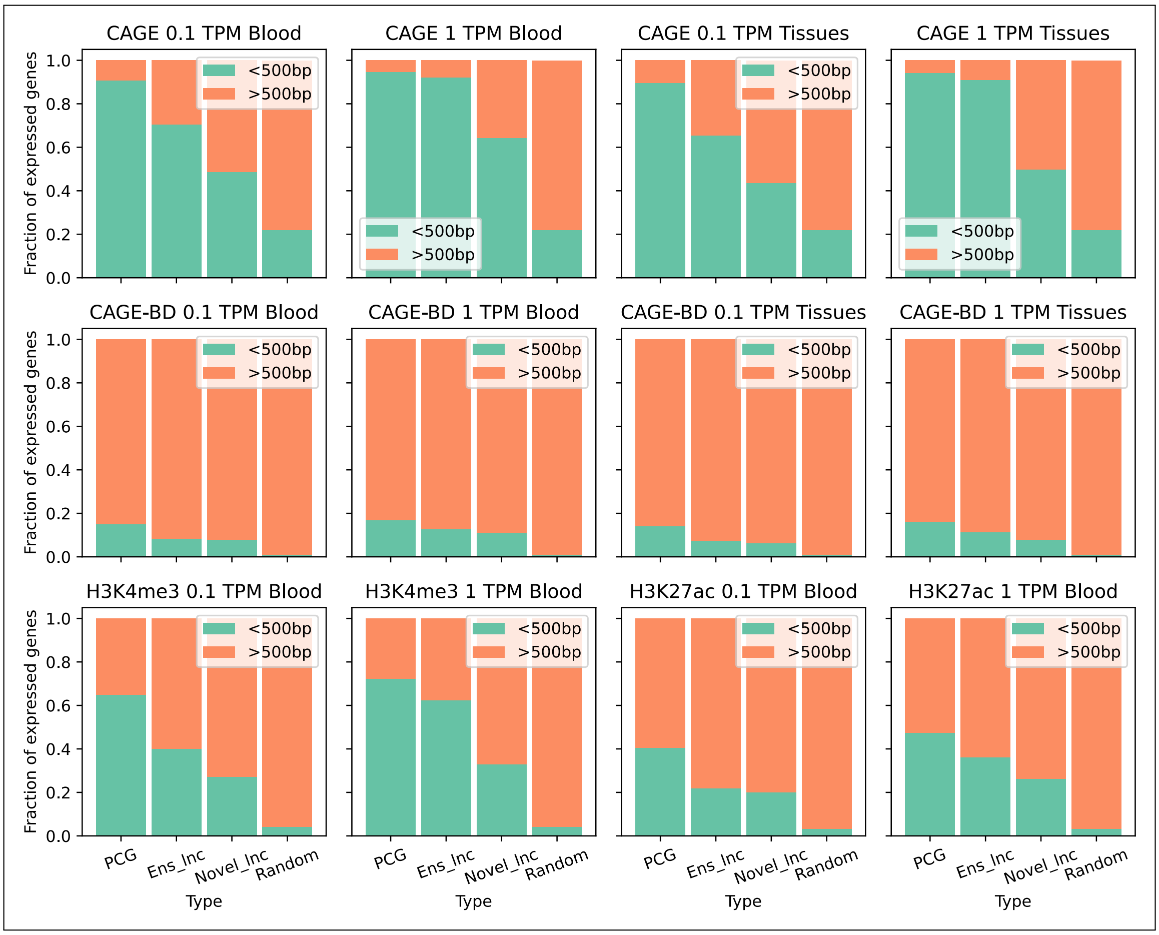


Figure 29: Support for transcription of annotated genes and novel lncRNAs as fractions of expressed genes with detected TSSs or active gene histone modifications. TSSs were obtained from CAGE-seq peaks from five immune tissues and histone modifications were obtained from CHIP-seq peaks (H3K4me3 and H3K27ac) from alveolar macrophages. PCG: Protein coding gene, Ens_lnc: Ensembl lncRNA, Novel_lnc: Novel lncRNA.

7.3.5. Differentially expressed lncRNAs and PCGs

We performed differential expression analysis between unstimulated or control samples and samples stimulated with either vaccines or a pathogen in order to identify common lncRNAs induced during an immune response. In blood samples there were 716 differentially expressed genes, including 75 novel lncRNAs and 22 annotated lncRNAs (Figure 30A, Supplementary Data 4). The large number of samples used in the blood sample dataset (222) and the heterogeneity of the data produced many differentially expressed genes with modest effect sizes but robust p-values (Supplementary Figure 7). The most significant enriched terms among the known genes were biological processes related to the immune response to external pathogens such as response to external stimulus (GO:0009605, FDR=2.86e-09), response to virus (GO:0009615, FDR=6.75e-07) or defense response (GO:0006952, FDR=1.19e-07). In lymph samples, there were 365 differentially expressed genes, including 46 novel lncRNAs and 13 annotated lncRNAs (Figure 30B, Supplementary Data 4). In this case, among the most significant enriched terms with the highest quantity of genes were general terms such as response to stress (GO:0006950, FDR=2.51e-04) and response to stimulus (GO:0050896, FDR=1.21e-03). More specifically, the terms related to T cell activation, like T cell activation (GO:0042110, FDR=3.07e-03) and regulation of T cell activation (GO:0050863, FDR=3.37e-03), reflect the critical roles of lymph nodes in adaptive immunity. Besides, there also were highly significant but smaller in size enriched terms related to response to endoplasmic reticulum (ER) stress.

There were 22 differentially expressed genes common to both datasets, among them an annotated lncRNA and an unannotated lncRNA. Some of the common PCGs are directly related with immunity, like IL21, which encodes a well known cytokine with immunoregulatory activity that induces proliferation and differentiation in several immune cell types. Other genes are related to apoptosis and inflammation (MT2, IKBIP, AEN, OSGIN1) and ER regulation (WFS1, SELENOS). Despite relatively similar number of DE genes in both comparisons, there is a big set of highly significant genes with effect sizes smaller than the threshold in the blood samples and many statistically significant but lowly expressed novel lncRNAs did not pass the fold change threshold because they were shrunken (Supplementary file 1: Fig. S7). These results give support for potential involvement of a fraction of the detected novel and annotated lncRNAs in both the innate and adaptive immune responses, following the guilt-by-association principle.

7.3.6. Co-expression network analyses detect immune-enriched gene signatures

Gene co-expression networks were constructed providing valuable information about the expression relationships of lncRNAs with PCGs and allowing the inference of their

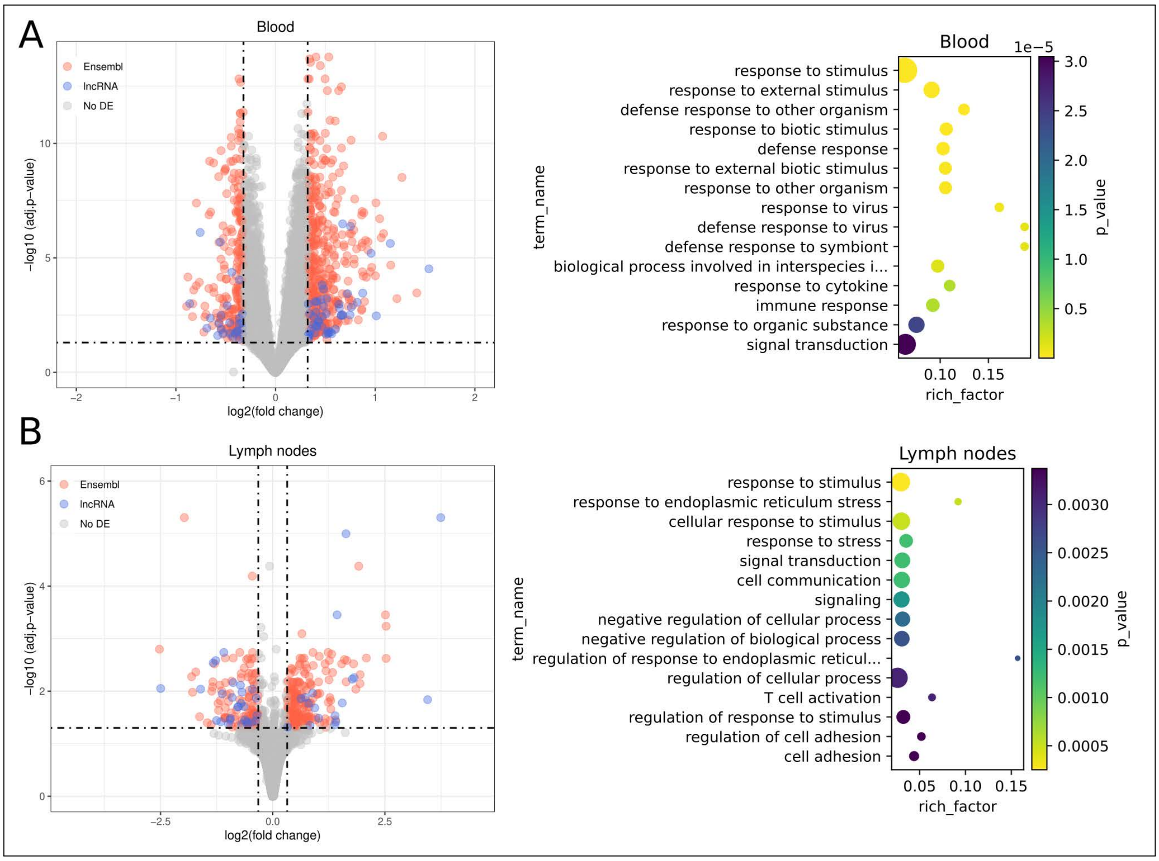


Figure 30: Differential expression results between stimulated samples and unstimulated samples in blood cell samples (A) and lymph node samples (B). For each comparison, a volcano plot using shrunken fold changes and a dot plot the results of gene ontology enrichment analysis (GO biological processes) are shown.

putative functions by guilt-by-association. We tested two different network construction pipelines and selected the one proposed by the authors of GWENA [31], as it produced networks with better fit to a scale-free topology and most of the genes could be associated to an expression module. Covariate correction for tissue type and source project enabled the construction of unbiased networks (Supplementary Figures 8-9). Filtering of lowly expressed genes, genes with low variability and outlier samples that reduced the fit to a scale-free topology resulted in co-expression networks of 12898 and 13428 genes in blood samples and lymph nodes, respectively. In the blood dataset, genes with similar expression patterns were clustered in 33 modules ranging from 54 to 1832 genes (Figure 31A, Supplementary Data 5), and in the lymph node dataset genes were clustered in 30 modules ranging from 44 to 1909 genes (Figure 32A, Supplementary Data 5). Most modules included novel lncRNAs and annotated lncRNAs, and some of them were even hub genes of their module.

We searched for significant correlations among module eigengenes, the principal component of the genes in the module that depicts its dominant trend, and treatment variables. In the blood sample dataset, 15 modules were correlated (p-val < 1e-03) with the general treatment variable, which accounts for any kind of sample stimulation (Figure 31A). Considering correlations to specific immune stimulations, helminth infection shared many correlated modules with the general treatment variable, which meant that it was one of the main drivers of variability in the dataset. Stimulation with LPS and with FMD inactivated virus (iFMDV) were correlated with specific gene modules different to those correlated to helminth infection. Other stimulations were also correlated to some modules but because of their small sample size they were not further taken into account.

Gene expression modules were characterised by GO term enrichment (Supplementary Data 5). Two of the stimulation-correlated modules (ME16 and ME19) were highly enriched in biological processes related to the immune response but they were not correlated with helminth infection. ME16 was associated with the sum of all treatments and was specially strongly correlated with LPS stimulation. The most significant enriched biological process GO terms were related to the general immune response, like immune system process (GO:0002376, FDR=2.27e-07) or immune response (GO:0006955, FDR=1.15e-05) and to cell migration and locomotion, including the terms positive regulation of locomotion (GO:0040017, FDR=1.55e-06) and leukocyte migration (GO:0050900, FDR=1.16e-05). ME19 was also associated with the sum of all treatments and was correlated to iFMDV treatment. It contained a high amount of immune response genes, for instance, from the 155 genes with GO annotations, 35 were related to response to virus (GO:0009615, FDR=9.59e-25) and 56 to immune system process (GO:0002376, FDR=1.56e-10). Besides, terms related to type I interferon response and signalling were also abundant.

In the lymph node network, the eigengenes of 11 modules were correlated (p-val < 1e-03) with the combined treatment variable (Figure 32A). The two available immune stimulation conditions, helminth infection and paratuberculosis, were correlated with a few modules, but several other significant modules emerged from the combined treatment variable correlation. The characterisation of gene expression modules by GO term enrichment revealed up to 5 immune-enriched modules: ME15, ME19, ME24, ME27 and ME28. Among them, the positively correlated modules showed functions involved in the innate immune response and general immune terms. For instance, in module ME15 the terms immune response (GO:0006955, FDR=2.10e-11) or innate immune response (GO:0045087, FDR=5.77e-07) are highly significant. In contrast, the negatively correlated modules are enriched in adaptive immune response terms. ME19 is enriched in GO terms related with T cell activation and lymphocyte proliferation while ME27 is enriched in terms related to B cell activation and proliferation. The lncRNAs present in the immune-enriched modules from both co-expression networks were classified as immune response-related lncRNAs.

In addition to the immune-enriched gene modules, another big module stood up (ME3), as it was correlated with both helminth infection and paratuberculosis. Most

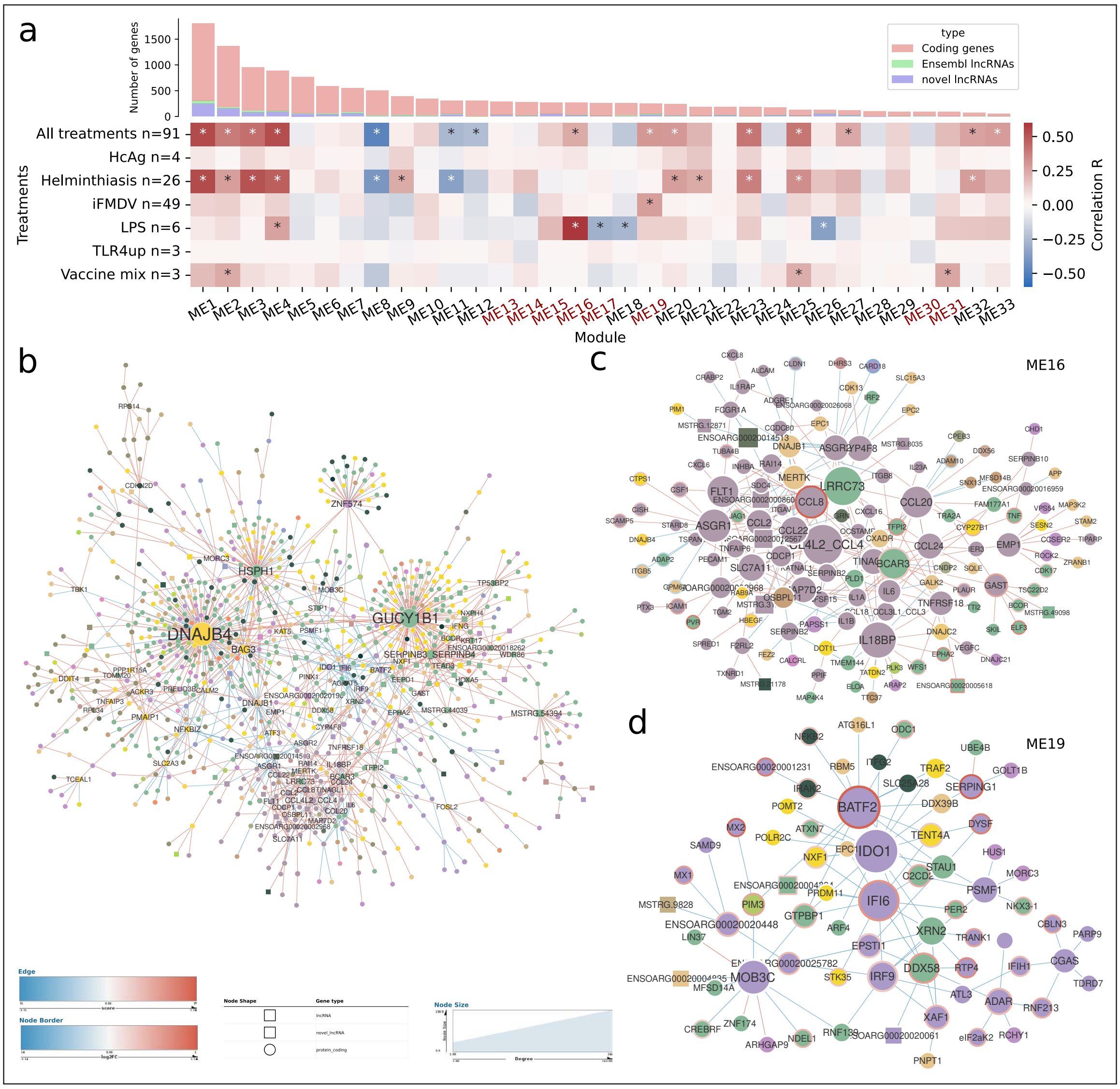


Figure 31: Co-expression analysis and differential co-expression network results in blood cell samples. (A) Correlations of gene co-expression modules with all stimulations and with each individual stimulation. Modules enriched in immune genes are highlighted in red. Number of genes in each module is depicted as a bar plot. (B) The full differential co-expression network. Node size is proportional to connectivity and differential associations are coloured by gain or loss of correlation strength. The edges of differentially expressed genes are coloured by fold change. (C) Sub-network with the differentially associated genes in module ME16. (D) Sub-network with the differentially associated genes in module ME19.

of the enriched GO terms were related to endoplasmic reticulum (ER) stress and protein post-translational processing, with terms like response to endoplasmic reticulum stress (GO:0034976, FDR=2.60e-13), Golgi vesicle transport (GO:0048193, FDR=2.42e-07) or response to unfolded protein (GO:0006986, FDR=1.72e-06).

7.3.7. Differential co-expression networks to identify regulatory relationships

Gene level differential co-expression, the gain or loss of correlation between two genes in different biological situations, indicates changes in regulatory relationships between those genes, which are often not evident from DGE results. All gene-pairs used in the co-expression network construction were tested for significant changes in correlation between control and stimulated samples and differential co-expression networks (DCN) were constructed with statistically significant gene-pairs (Supplementary Data 5). The DCN from the blood sample dataset contained 1589 differential associations (FDR < 0.05) among 1348 genes (Figure 31B) and the DCN from lymph nodes contained 2137 differential associations (FDR < 1e-03) among 1784 genes (Fig. 6B). Both networks included around 60 lncRNAs each. In terms of network topology, networks showed a small amount of nodes (genes) with many edges (differential associations), while the rest of the nodes were more loosely connected to the network. Just around 5% of the nodes had 10 or more edges. Some very interconnected nodes formed clusters according to the gene co-expression modules from the previous analysis, but most of the topology was driven by a few high-degree nodes.

Specific differential associations were observed by individually inspecting each DCN. The blood sample network was centred on two high-degree genes that had more than 100 differential associations each but did not have obvious biological relationship with the immune response: DNAJB4 and GUCY1B1. Among the rest of the 42 high-degree genes, defined as those with more than 5 differential associations, there were some lncRNAs and several immune genes such as BATF2, IDO1, IFI6, IL18BP, NFKBIZ and various CC chemokines. Focusing on the immune-enriched co-expression modules, many genes from the module ME16, most of them immune-related, formed a very interconnected subnetwork (Figure 31C). Even though the genes from this subnetwork were already correlated, they predominantly showed positive z-scores, which means that the correlations were stronger in the control samples than in stimulated samples. On the contrary, genes from the module ME19 did not form a separate cluster, but they showed negative z-scores, which means that their expression was correlated in the stimulated samples. Interestingly, many genes were up-regulated in the differential expression analysis. In the subnetwork composed by selecting the genes from this module and their differentially co-expressed pairs, there were transcription factor coding genes related to the immune response: BATF2, IRF9 and NFKB2 (Figure 31D). For instance, BATF2, upregulated in stimulated samples, is a transcription factor that controls the differentiation of lineage-specific cells in the immune system and immune-regulatory networks. There were several interferon-stimulated genes such as IFI6, MX1, MX2, ADAR, EIF2AK2, IRF9 or IFIH1, all related to antiviral functions and upregulated in the stimulated samples.

The DCN obtained from lymph node samples did not contain many immune-related

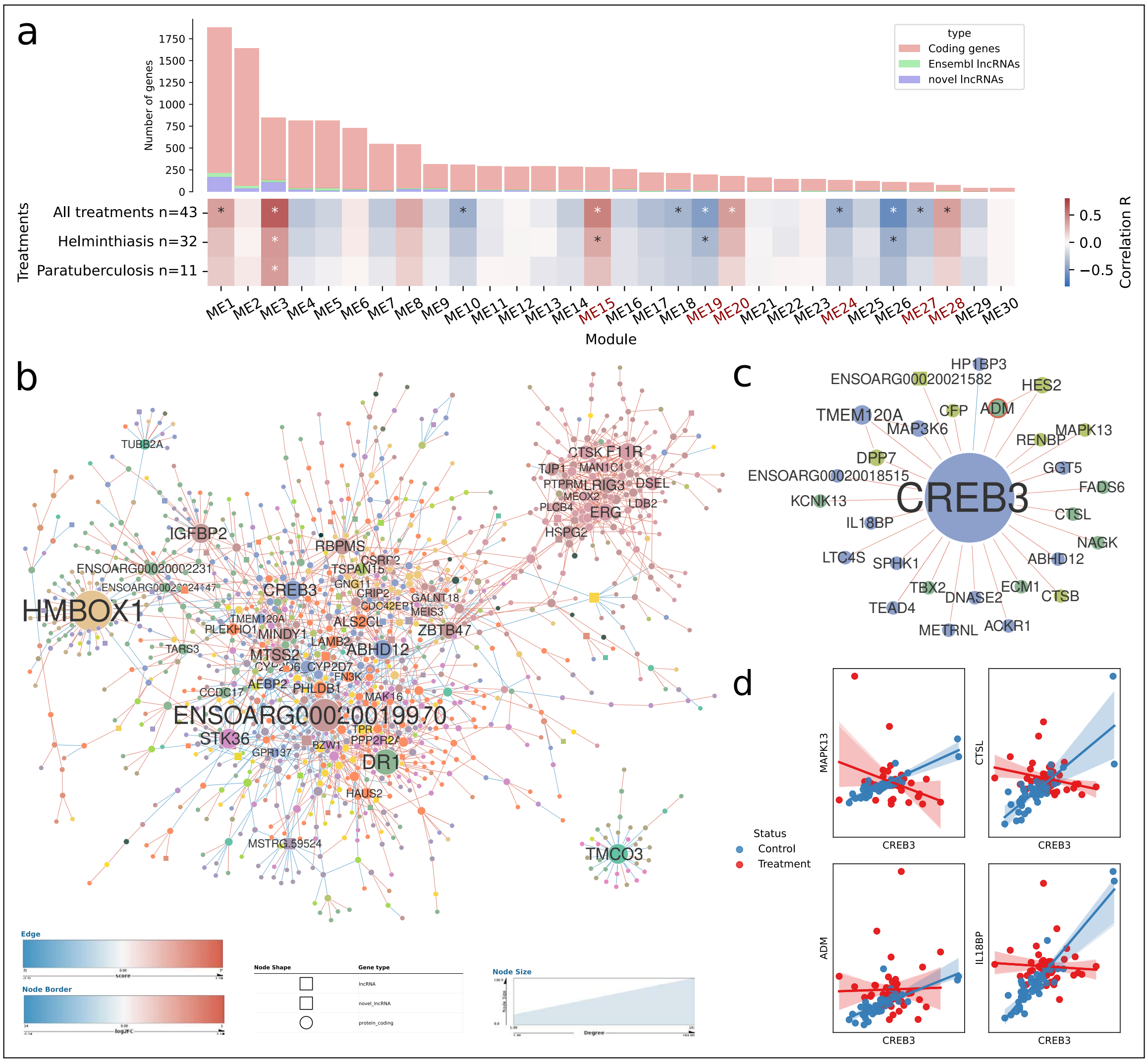


Figure 32: Co-expression analysis and differential co-expression network results in lymph node tissue samples. (A) Correlations of gene co-expression modules with all stimulations and with each individual stimulation. Modules enriched in immune genes are highlighted in red. Number of genes in each module is depicted as a bar plot. (B) The full differential co-expression network. Node size is proportional to connectivity and differential associations are coloured by gain or loss of correlation strength. The edges of differentially expressed genes are coloured by fold change. (C) The genes differentially co-expressed with CREB3 transcription factor. (D) Individual examples of statistically significant differential associations between CREB3 and four genes.

genes (Figure 32B). There were 187 high-degree genes, including 18 transcription factors coding genes that were potential drivers of the differential co-expressions, like HMBOX1, CREB3, NFATC4, NFIB or EBF4. NFATC4, for instance, is involved in T-cell activation, stimulating the transcription of IL2 and IL4 cytokine genes. CREB3, among many other functions, plays a role in the response to ER stress by promoting cell survival, a process that was previously found enriched in a co-expression module of which CREB3 was not part of. CREB3 was a high-degree node, differentially associated with 27 other genes (Figure 32C), and it showed mostly positive z-scores, thus, its expression was correlated in the unstimulated samples but those correlations were lost upon stimulation by helminth infection and paratuberculosis. Examples of differentially associated genes include IL18BP, CTSL, MAPK13 and ADM (Figure 32D). ADM, which is a known lymphangiogenic factor, was upregulated in stimulated samples and its expression decoupled from that of CREB3 in those samples. This DCN also contained several lncRNAs and 12 of them were high-degree nodes (7 known lncRNAs and 5 novel lncRNAs).

7.3.8. Integration of evidence for lncRNA expression and function

We used differential gene expression analysis, co-expression analysis and differential co-expression network analysis for the functional association of lncRNA genes with the activation of the immune response. Those three approaches resulted in 320 lncRNAs associated in at least one analysis (Supplementary Figure 10). The differential expression between stimulated and unstimulated samples showed the highest number of immune response-associated lncRNAs. Interestingly, the histone modification support in differentially expressed novel lncRNAs was much higher than in the whole set of novel lncRNA genes, 49% against 19%, and the trend was similar in the case of CAGE support. A summary of all transcription evidence and associations in an analysis for each lncRNA is available as a supplementary file (Supplementary Data 6).

7.4

Discussion

Using 422 RNA-seq samples from ovine immune tissues, we assembled a project-specific transcriptome and retrieved 17605 unannotated lncRNA loci. Around 70% of those novel genes were expressed in a sufficient number of samples and/or were supported by histone modifications or TSSs from independent experiments. LncRNAs are usually annotated with evidence-based methods, because they lack sequence features like conservation or complete ORFs [101], and this evidence mostly comes from mapping sequencing reads to the genome of interest. Model organisms have been annotated via manual curation of a variety of assays, but in the absence of this kind of data in livestock species, lncRNA annotations usually rely on automated short-read transcriptome assemblies. SGS short-read RNA-seq is widely used because of its high yield and low cost [99] and has been used in many lncRNA annotations [101], but using this kind of data is challenging, because the nature of short reads makes it difficult to completely characterize the structure of non-coding transcripts [102].

For higher confidence on the assembled transcripts, only paired-end samples were used and additional support was included from expression levels, CAGE-seq tags and histone modification CHIP-seq assays. Thus, the confidence in the existence and location of the more than 12 thousand confident lncRNA loci is high, even though not all gene boundaries and splice sites might be correct. In fact, the reproducibility of exact lncRNA short-read transcript models between samples was shown to be low in another sheep study [22]. Related to this, the amount of detected lncRNAs did not reach saturation at any sequencing depth. It has been proposed that, because of stochastic sampling, much higher sequencing depth is needed to reconstruct the vast number of lowly expressed lncRNA transcript models [22]. It should be mentioned that it is expected that a higher number of the assembled transcripts have independent evidence of expression. On one side, the signal of CAGE-seq scales with expression, similar to RNA-seq so, lowly expressed transcripts are also more weakly represented. On the other, the CHIP-seq dataset used only comprises a single cell type, while the RNA-seq dataset includes several tissue-types.

As observed in other livestock studies [22, 137], the expression levels of lncRNAs were lower than those of PCGs. The lncRNAs already present in the Ensembl annotation were more abundant, were expressed in more samples and were better supported by TSSs and histone modifications, reflecting their misannotation as lncRNA genes when many of them show coding potential. The low expression in bulk RNA-seq samples might be due to their known exceptional cell type, tissue, developmental stage and disease state specific expression [20, 21] and even to lowered transcriptional burst frequencies in single-cells [408]. In human and murine T cells and B cells, lncRNAs are expressed in a very cell-specific and dynamic way during differentiation within lineages of the same cell types [347–349]. In this manner, cell or tissue type specific lncRNAs could be involved in immunological pathways in response to infection and vaccination [340, 345], even if the perceived bulk expression was low.

The biological function of most lncRNAs remains unknown, particularly in non-model organisms. With notable exceptions, few genes can be assigned a putative function by homology with human or mouse lncRNAs. Considering sequence similarity, around 700 novel sheep lncRNA transcripts had orthologues in human, including some functionally characterised lncRNAs, and more than 3000 in goat or cattle, which are mostly uncharacterised. Because of this, we linked the sheep lncRNAs with potential broad biological functions and pathways by using classical analyses like, differential gene expression and co-expression analysis, and alternative methods like differential co-expression network analysis. In the case of the co-expression analyses, following the principle of guilt-by-association [115], association with the immune response was assigned via correlation to a group of co-expressed genes. This approach has been widely used for the functional profiling of lncRNAs by several studies [409, 410]. In addition, one of the datasets included in this study has already been analysed in this way to specifically search for candidate lncRNAs during an helminth infection [183].

Regarding the results from the blood cell dataset, with samples from whole blood, PBMCs and other cells like macrophages, all analyses resulted in the identification of genes linked to the innate immune response. Many genes were part of the interferon (IFN)-mediated immune response, which provides a first line of defence against pathogens, from viruses to parasites [411]. Upon pathogen detection and IFN stimulation, the transcription of several genes termed as IFN-stimulated genes (ISGs) is activated, which control pathogen infection by targeting pathways necessary for pathogen life cycles. Up to 21 of the most important antiviral ISGs were upregulated in the differential expression analysis, including ADAR, APOBEC3Z1, BST2, RSAD2, MX1, MX2, IFI6, IRF9 or orthologues of the OAS gene family. These genes were part of the iFMDV-associated co-expression module and many were also part of the DCN. The IFN response was mostly driven by the inactivated vaccine [342, 396] and the LPS stimulation datasets [357], while the helminth infection [397, 400] and other smaller datasets [235, 398, 399] produced a different expression profile, as seen in the stimulation-correlated co-expression modules. In the same manner as known ISGs, lncRNAs can also be induced by IFN and have important roles in controlling pathogen infection and resolution of the immune response, or they can regulate the IFN mediated host defence [412, 413]. For instance, in human, NRIR is a negative regulator of IFN antiviral response [414] and IFNG-AS1, located near the IFNG locus, regulates its expression [16]. Considering that differentially expressed lncRNAs have been proposed to function as negative or positive regulators in various critical steps of antiviral response [415], some of the ovine transcripts detected in this study could also be related to those processes.

The fact that the lymph node dataset was dominated by helminth infection experiments [404, 405, 407], except for a single bacterial infection experiment [402], greatly marked the type of genes involved in the general analysis. The different analyses revealed important immune-related genes and biological pathways, but there were many other processes involved. Parasite infections produce a different response than viral or bacterial infections and are usually associated with a non-inflammatory Th2-biased response in both parasites present in the datasets: Teladorsagia circumcincta and Fasciola hepatica [416–418]. In a human gene expression meta-analysis with different helminth species, they found upregulated immune regulatory genes while down-regulated genes were mainly involved in metabolic processes, and showed that the response was similar between species and tissues [419]. To date, there are very few studies linking lncRNAs to helminth infection in mammals. In sheep, one of the datasets included in this study [405] has been analysed for this purpose to specifically search for candidate lncRNAs during T. circumcincta infection [183]. It remains greatly important to identify novel gene candidates for this disease, as it is a source of economic loss and animal welfare deterioration [417].

Integration of several RNA-seq datasets and different bioinformatic analyses allows us to better characterise patterns that could have been overlooked in individual experiments. One of the processes consistently appeared associated with all the analyses in lymph nodes was the response to ER stress, which is an endogenous source of cellular stress that arises in the ER of cells following the accumulation of misfolded proteins during protein synthesis [420, 421]. In the immune system, this response is particularly important for resolving secretory stress and survival of highly secretory cells such as immunoglobulin producing plasma cells [420], cytokine producing Th2 cells [422] and other immune cells [421]. Among the ER stress response-related dysregulated genes, the two most important members of the IRE1a-XBP1 pathway (ERN1 and XBP1) were upregulated in the lymph node samples [421] and a co-expression module enriched in ER stress response genes was correlated with both helminth infection and paratuberculosis. This process was enriched in the sets of DEGs in the original analyses of the paratuberculosis dataset [402] and one helminth infection dataset [397], but were not further discussed in their respective publications. Furthermore, while belonging to a non-associated co-expression module and not being differentially expressed, the ER localized transcription factor CREB3 was differentially co-expressed with several other genes. CREB3 has been implicated in the ER and Golgi stress response and regulation of genes in secretory pathways [423]. LncRNAs have also been linked to proliferation and apoptosis during ER stress [424].

The DCN analyses revealed the involvement of other PCGs and lncRNAs in the ovine immune system activation. Compared to the widely employed co-expression methods, differential co-expression have the advantage of detecting condition-dependent interactions between genes [116]. For instance, the gain or loss of co-expression between a TF and its targets can be due to expression changes or post-translational modifications of the TF [395]. Apart from the mentioned CREB3 transcription factor in lymph nodes, in blood cell samples IDO1 seemed to be differentially regulated. IDO1 is a rate-limiting metabolic enzyme that converts tryptophan into downstream kynurenines, which have immunosuppressive roles, and is known to be interferon-inducible [425]. Similarly to the general differential expression analysis in this study, the original analysis of LPS effect on macrophages did not find an induction of IDO1 expression, even if it was expected [357]. Interestingly, we found that IDO1 was part of a co-expression module associated with immune stimulation and enriched in ISGs, and was differentially correlated with several genes. In stimulated samples its expression was independent from other genes, but upon immune stimulation it gained correlations with genes like the ISG DDX58. All in all, the constructed DCNs revealed several lncRNAs with stimulation-dependent associations that could have immune regulatory roles, and this approach could be useful to find novel gene candidates in each pathogen infection or vaccine component.

Multiple processes are involved in the immune response to infection and vaccination and lncRNAs might play different roles in these processes. The goals of this work were (1) to detect unannotated ovine lncRNAs from publicly available RNA sequencing datasets from immune tissues and then (2) define a lncRNA gene expression signature of the general immune activation. Poor sequence conservation and low expression, general features found in other mammal studies, were also features of ovine lncRNAs. Adding support from CAGE sequencing and histone modifications, we obtained a shortlist of more than 12 thousand unannotated high-confidence ovine lncRNAs. The functional analyses performed with immune-stimulated samples revealed hundreds of known and novel lncRNAs with specific expression patterns during an infection or vaccination. These genes make up a prioritized set of potential candidates for deeper experimental analyses. Taken together, these results should help completing the sheep non-coding RNA gene catalogue, and most importantly, they give evidence of immune state-specific lncRNA expression patterns in a livestock species.

Chapter 8

General Discussion and Conclusions

8.1

General discussion

This thesis dissertation brings together a wide range of high-throughput sequencing assays that have been used for the exploration of the currently under-annotated sheep non-coding transcriptome in the context of the immune response to pathogens and vaccines. While these projects have focused on different non-coding RNA classes (miRNAs and lncRNAs) and different aspects of the immune response, they are united by the common goal of characterising unannotated non-coding genes and their expression patterns along protein coding genes. Since results have been specifically discussed in each chapter, in this discussion we give some general remarks common to all studies and discuss the implication of non-coding genes in the immune system.

8.1.1. Methodological aspects of non-coding RNA detection in non-model species

An important aspect for lncRNA detection is the selection method used during library preparation, as effective ribosomal RNA removal is required for all RNA-seq libraries. The samples sequenced by our lab were prepared with a total RNA approach, using ribosomal RNA (rRNA) depletion, whereas in the multi-tissue immune dataset there were samples prepared by rRNA depletion and samples prepared with poly-A selection. It has been proposed that, because not every lncRNA is polyadenylated [426, 427], total RNA should be the preferred selection method for a complete transcriptome dissection. In this way, studies have shown that rRNA depletion does allow for the detection of more lncRNA transcripts, but it needs more sequencing depth than poly-A selection [428]. In fact, at the same depth, the number of lncRNA genes obtained is usually greater with poly-A selection [429], but it still lacks non-polyadenylated transcripts. This happens because the effective number of exonic reads in total RNA data is lower, since a big proportion of them come from unspliced introns, remaining rRNA contamination, small RNAs or other sources of intergenic noise [428]. In the multi-tissue immune dataset, even if there was not any paired sample with both selection methods, we could detect a higher amount of lncRNAs in samples with rRNA depletion, but differences were small when considering strongly expressed genes only. Part of these non-polyadenylated transcripts could be enhancer RNAs, a class of non-coding RNAs known to be lowly-expressed, unstable and nuclear with emerging functions in gene regulation [430].

Novel miRNA prediction is a compulsory step for the analysis of the full miRNA transcriptome in non-model organisms. For this, we applied the widely used workflows miRDeep2 [238] or sRNAbench [431] to annotate the missing ovine miRNAs from high-throughput sequencing data. These tools have been updated through the years and are now standards in the field [432].

On the contrary, the methods and pipelines used for the discovery of lncRNA genes are in constant evolution. In this work, we used genome-wide short-read RNA sequencing assays in order to reconstruct unannotated lncRNA transcripts with a specifically designed pipeline. In the first steps of the workflow sequencing reads were mapped to the sheep genome and the resulting alignments were used for transcriptome assembly using established tools [433]. As already mentioned, short reads seldom span across several splice junctions, making it challenging to infer full-length transcripts and determining transcription start and end sites [102]. Nevertheless, non-model and human studies have extensively and successfully used this approach, due to the high cost of long read sequencing and overall good performance. Some recent examples include an RNA atlas from 300 human tissues and cell lines [427], a study on human intergenic transcription [434] and an analysis of thousands of sequencing datasets in different yeast species [435]. Works in livestock species have also identified novel lncRNAs by assembling transcripts from short reads in two multi-species datasets [135, 137], as well as the novel lncRNA set obtained from The Sheep Expression Atlas [22]. Importantly, these tools are much more accurate at reconstructing lowly expressed transcripts such as lncRNAs if provided with enough sequencing depth [436]. The sequencing datasets from ovine brain and PBMCs had around 70 million paired-end reads, and the samples collected from multiple immune tissues had 45 million reads on average, with some samples having up to 150 million reads. This depth makes the datasets used in this study optimal for lncRNA detection.

As far as the bioinformatic analysis is concerned, for the actual identification of lncRNA transcripts among the great amount of RNA products that were reconstructed, we produced a workflow with Python scripts and selected tools. Existing tools for global lncRNA discovery often include several modules for extracting and filtering candidate transcripts, computing candidate transcripts' coding potential, and classifying lncRNAs based on their genomic localization [48, 380, 437, 438]. These tools, albeit useful for rapid transcriptome characterisation, can have some drawbacks: they might be only available to study specific model species, they might require extensive effort for installation and use in a high-performance computing server, or most importantly, they might lack in flexibility with the desired parameters and do not allow customisation. For instance, the widely used tool FEELnc [380] outputs a higher quantity of transcripts, but some overlap PCGs in the same strand, and it does not allow for different transcript length filters between monoexonic and polyexonic transcripts. Thus, the pipeline used through this thesis for lncRNA characterisation was developed from scratch for a tailor-made analysis of the sheep transcriptome, while using the latest tools and concepts in the lncRNA field. For the sake of reproducibility, the pipeline can be run as a Snakemake [377] pipeline with fixed program version numbers and has been posted to a public repository (https://github.com/bilbaom/immune-lncrnas-sheep). As a limitation of this work, it should be noted that the workflow has not been optimized for execution time and has not been wrapped into a single standalone tool with documentation for an easier use because, as mentioned, there are other tools that can do a similar job.

8.1.2. Characteristics of ovine lncRNA genes

Newly found ovine lncRNAs shared similar characteristics with those previously found in other livestock [22, 133, 135–137] and human [21, 427] studies: they were lowly expressed, they had few exons, they were primarily intergenic and they showed limited sequence conservation. For an additional level of confidence on the existence of the unannotated transcripts, CAGE-seq and histone modification CHIP-seq data was intersected with the set of lncRNAs in the multi-tissue immune study. This approach, which is followed by the main gene annotation pipelines [439], was used in a recent human study and showed similar levels of support as in our analysis [427].

Classifying novel transcripts based on their genomic location can give an insight about their potential biological roles, as seems that there might be different gene subtypes within the lncRNA umbrella term [27]. In each study, we classified all transcripts with a custom-made script. In general, there were more divergent lncRNAs - transcripts sharing a bidirectional promoter with a PCG - than the other PCG-associated classes and tended to be correlated with the adjacent gene. In human and murine embryonic and pluripotent stem cells it emerged as the most abundant class and they might regulate the expression of adjacent genes [30, 33]. As an example, in this thesis, the expression of a novel lncRNA in PBMCs was correlated with the adjacent OSM cytokine gene, and both genes were induced in groups treated with commercial vaccines or vaccine adjuvants (chapter 6). We also showed that they were expressed at higher levels than other classes like antisense or intergenic lncRNAs (chapter 7).

8.1.3. Functional implications in the immune system

The differentiation and activation of the innate and adaptive immune cells in response to diverse external stimuli are tightly coordinated events that cause activated immune cells to undergo rapid and dynamic changes in gene expression [440–442]. There is growing evidence of the involvement of miRNAs and lncRNAs in immune functions [443, 444], with most of the experimental work belonging to human or murine studies.

miRNAs have been shown to be essential for the host antiviral defences and viral pathogenesis [445]. In this way, the miRNAs that are upregulated or downregulated in livestock species have been proposed as potential biomarkers for a variety of infections, ranging from micobacterial infections or mastitis in cattle, to Peste des petits ruminants virus (PPRV), sheep pox virus (SPPV) or bluetongue virus (BTV) infection in small ruminants [130, 131]. For instance, miR-21 was highly upregulated in VMV infected sheep, but it has also been found differentially expressed during bovine mastitis [446, 447], influenza A virus infection in pigs [448], avian influenza infection in chicken [449] or PPRV infection in sheep [450]. miR-21 is a critically important miRNA in livestock health, development and disease, but due to its lack of specificity, it does not seem a useful biomarker [451]. Nevertheless, as previously discussed, it could serve as an unspecific marker of inflammatory lesions in VM disease and other infections, or it could be targeted for therapeutic intervention in these diseases. In fact, a clinical trial is testing a miR-21 mimic in human patients with diabetes (NCT02581098).

Unlike miRNAs, which interact with their target RNAs through base complementarity, lncRNAs control immune processes through a variety of mechanisms [444]. Despite the limited number of lncRNAs with known function, hundreds of human or murine lncRNAs have been shown to be specifically expressed in certain immune cell types and immune cell development stages [347, 349]. Besides, many others are known to respond during the immune system activation by external pathogens, vaccines or pro-inflammatory mediators [340, 452].

In non-model organisms, since the non-coding gene sets are still in need for improvement, genome-wide transcriptomic studies that link ncRNAs to immune functions have been less common. In the present thesis, we have analysed their expression profiles during different immune responses. A widely used heuristic approach, termed guilt-by-association, was followed in order to associate transcripts with unknown functions to biological processes. A known example of successful outcome of this approach was the discovery that TP53COR1 as part of a co-expression cluster of genes related to the p53 transcriptional pathway [53], which afterwards was confirmed to mediate p53-dependent apoptosis [453]. In this manner, a putative immune-related function can be assigned to the ovine lncRNAs that were part of co-expression gene modules enriched in immune biological processes and pathways. The hundreds of lncRNAs that were dysregulated during an immune stimulation with commercial vaccines or various pathogen infections should also be implicated in those responses.

8.1.4. Evolution of non-coding genes

The evolutionary dynamics of non-coding genes is different from PCGs because they are not restricted by the genetic code. Unlike lncRNAs, many miRNA genes show high evolutionary conservation, but the emergence and turnover of new miRNAs is also significant [12]. Similarly to the results obtained in this project, it is known that in mammals broadly conserved miRNAs show higher expression levels and tend to be among the annotated genes [454]. Nevertheless, the ample use of high-throughput sequencing techniques has led to an increase in the number of miRNAs with limited sequence conservation. Evolutionary young miRNAs are considered those which are specific to a clade, like primates or ruminants. These genes are usually confidently identified by miRNA prediction algorithms as *bona fide* miRNAs but show lower expression levels or very tissue, cell or development specific expression. Evolutionary studies in domestic mammals suggest that young miRNAs are expressed in a single or a few tissues when they first appear, and become more broadly expressed over time [274]. We could observe this in the multi-tissue analysis, where miRNAs with no known orthologue showed lower expression levels and many of them were highly tissue specific.

Despite the fact that in both, lung samples and brain samples, more than half of expressed miRNAs were not previously annotated, we could identify orthologues in other species by sequence similarity for the majority of those miRNAs. These results are common when profiling sheep miRNAs by sRNA-seq. For instance, recent works on female reproductive organs have also identified hundreds of unannotated sheep miRNAs and many of them are conserved in another species [192, 249]. In order to produce a comprehensive list of sheep miRNAs for future studies, we took advantage of the increasing number of small RNA sequencing experiments uploaded to public databases to detect hundreds of unannotated miRNA genes across several tissues (chapter 4). Among the over 1000 previously unannotated miRNA precursors, 41% were assigned an orthologue in a close species or human by sequence similarity.

MicroRNA evolution is characterised by punctual instances of elevated rates of miRNA innovation, such as the increase occurred in the lineage leading to human, after it split from mouse [85]. New miRNAs can be rapidly gained and lost during metazoan evolution, implying that many poorly conserved miRNAs in extant species have not yet acquired a fitness-enhancing function [12]. In the VMV dataset analysed in this work, we described the ruminant-specific novel miRNA family mir-2284/2285 in sheep, which had already been described by other studies [276], but is currently annotated only in cattle and goat. In the multi-tissue analysis, over 100 members of that family were identified in the sheep genome for the first time, thus confirming that the expansion of this family is also present in sheep. Unlike other lowly expressed and poorly conserved miRNAs, the miRNA family mir-2284/2285 has had an impressive expansion in ruminants, both in term of quantity and divergence, thus, it may represent a ruminant innovation with functional importance [275]. We found out that they are significantly more abundant in immune-related tissues and, in cattle, their predicted target genes have been linked to insulin resistance [275]. The target-gene predictions in sheep also pointed towards metabolism and immune related functions. Nevertheless, more research is needed to elucidate if this family has gained a functional role in ruminants.

Evolutionary conservation of lncRNA loci is a subject of intense research and debate. Unlike miRNAs, in most cases their function is not linked to a strict sequence feature, which makes the identification of orthologues between distant species very challenging [455]. There seems to be two sets of conserved lncRNAs: one that shows signs of purifying selection at the sequence level, and one that shows selection for transcription, small functional elements or secondary structure only. [49, 455]. Even if they are highly conserved, most members of the first set were not present in the ovine reference annotations. Comparing with other species, among the unannotated novel genes we found conserved lncRNAs such as TUNA, a nervous system-specific transcript necessary for neural development [321], and other lncRNAs widely characterised in human or mice. These included MALAT1, NEAT1, XIST, PACERR or FIRRE, for example. There were much more conserved lncRNAs between sheep and other close livestock species, showing that, even if there is a high evolutionary turnover, some could be lineage specific, similarly to what happens with miRNAs.

The meaning of the non-coding transcripts that lack evolutionary conservation is the most contested issue. It has been hypothesised that the occurrence of these non-conserved and lowly expressed transcripts might be necessary for novel gene evolution. The unstable transcripts found in the inherently bidirectional promoters and enhancers may not even have a function as mature transcripts, but they can serve as a fertile ground for more complex non-coding transcript evolution [456, 457]. A range of factors may drive the generation of longer, more stable ncRNAs from these elements and, occasionally, their functionalization [458]. The first is that these transcripts will be spatiotemporally controlled from the beginning in conjunction with one or more nearby PCGs, providing an opportunity for the creation of a negative or positive feedback loop. Second, transposable elements that are inserted close to enhancers or promoters may boost the extension and stability of developing lncRNAs. What is more, it has been recently proposed that the production of functionless no-coding transcripts could also provide a base for protein-coding gene *de novo* evolution [459].

8.1.5. Future prospects

The future of non-coding RNA research in farm animal species, important for global food production and in the emergence of zoonotic diseases, is linked to the profound functional annotation of their genomes. New emerging technologies such as long-read high-throughput sequencing and single-cell RNA sequencing will provide more accurate gene models and will disentangle the high tissue and cell type specificity of non-coding transcript expression. The non-coding genes related with the immune response to pathogens and vaccines could be useful to directly associate molecular phenotypes like gene expression to variants associated with complex traits (e.g. resistance to pathogen infection or good response to vaccination).

8.2

Conclusions

The main conclusions of the present work are the following:

1. By analysing high-throughput sequencing data from 21 tissues produced in our research group or available in public repositories, more than 1000 unannotated ovine miRNA genes were predicted. A big proportion (40%) of those sequences had a conserved orthologue in another species, but we found several clade or species-specific miRNAs that were characterised by higher tissue specificity. Among those, we detected 146 loci expressing precursors of the miR-2284/2285 miRNA family confirming its ruminant-specific expansion.

2. As for miRNAs relating to infection, comparing seronegative, asymptomatic seropositive and diseased animals, Visna-Maedi virus infection causes changes in expression in several miRNAs in lungs, which are already evident in asymptomatic animals. Oar-miR-21, oar-miR-148a and oar-let-7f seem to have potential implications for the host-virus interactions because of their strong upregulation in both asymptomatic and animals with lesions. Oar-miR-21 could serve as a marker of the lesions produced by the virus.

3. In this work, a tailor-made bioinformatic pipeline was created for the identification and classification of unannotated lncRNAs from reference annotation-guided transcriptome assemblies, which was applied through the present thesis project. This provided a useful way of selecting lncRNA transcript candidates using specific parameters not available in other tools.

4. More than 12000 unannotated ovine lncRNA genes were identified by analysing high-throughput sequencing data produced in our research group and available in public repositories. These transcripts showed the usual characteristics of these non-coding elements such as lower expression levels, higher tissue specificity and poor sequence level conservation.

5. Long non-coding RNA genes were found to be dysregulated during the immune responses to different stimulations like pathogens or vaccines, which may implicate them in those biological processes.

   - In a long-term vaccination experiment with animals treated with commercial vaccines and aluminium hydroxide adjuvant alone, lncRNA expression was altered by both aluminium hydroxide and complete vaccines. Two tissues were analysed from this experiment: the changes were strong in the innate immune response profiled in PBMCs, with a total of 304 lncRNAs with altered expression by any of both treatments. In the brain transcriptome there were also changes, but to a lesser extent, with 30 differentially expressed lncRNAs in the adjuvant-only group and 7 lncRNAs in the vaccine group.

- Using a transcriptome meta-analysis approach based on data-base data from 422 samples, a lncRNA gene expression signature of the general innate and adaptive immune system response to pathogens and vaccines was obtained. Differential co-expression networks find immune state-specific relationships between coding genes and lncRNAs. Using different analyses, we associate 320 known and unannotated lncRNAs with putative immune response functions because of their expression patterns, including response to viruses, immune cell activation, interferon response or endoplasmic reticulum stress.

Chapter 9

Bibliography

1. Vogel, F. A Preliminary Estimate of the Number of Human Genes. *Nature* **201,** 847–847 (1964).

2. Pertea, M. & Salzberg, S. L. Between a Chicken and a Grape: Estimating the Number of Human Genes. *Genome Biol.* **11** (2010).

3. Kung, J. T., Colognori, D. & Lee, J. T. Long Noncoding RNAs: Past, Present, and Future. *Genetics* **193,** 651–669 (2013).

4. Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., *et al.* Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. *Nature* **583,** 699–710 (2020).

5. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., *et al.* Landscape of Transcription in Human Cells. *Nature* **489,** 101–108 (2012).

6. Hombach, S. & Kretz, M. in *Non-Coding RNAs in Colorectal Cancer* (eds Slaby, O. & Calin, G. A.) 3–17 (Springer International Publishing, Cham, 2016).

7. Willyard, C. Expanded Human Gene Tally Reignites Debate. *Nature* **558,** 354–355 (2018).

8. Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116,** 281–297 (2004).

9. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14. *Cell* **75,** 843–854 (1993).

10. Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., *et al.* The 21-Nucleotide Let-7 RNA Regulates Developmental Timing in Caenorhabditis Elegans. *Nature* **403,** 901–906 (2000).

11. Friedman, R. C., Farh, K. K. H., Burge, C. B. & Bartel, D. P. Most Mammalian mRNAs Are Conserved Targets of microRNAs. *Genome Res.* **19,** 92–105 (2009).

12. Bartel, D. P. Metazoan MicroRNAs. *Cell* **173,** 20–51 (2018).

13. Ha, M. & Kim, V. N. Regulation of microRNA Biogenesis. *Nat. Rev. Mol. Cell Biol.* **15,** 509–524 (2014).

14. Saliminejad, K., Khorram Khorshid, H. R., Soleymani Fard, S. & Ghaffari, S. H. An Overview of microRNAs: Biology, Functions, Therapeutics, and Analysis Methods. *J. Cell. Physiol.* **234,** 5451–5465 (2019).

15. Connerty, P., Ahadi, A. & Hutvagner, G. RNA Binding Proteins in the miRNA Pathway. *Int. J. Mol. Sci.* **17,** 31 (2016).

16. Jonas, S. & Izaurralde, E. Towards a Molecular Understanding of microRNA-mediated Gene Silencing. *Nat. Rev. Genet.* **16,** 421–433 (2015).

17. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136,** 215–233 (2009).

18. Ulitsky, I. & Bartel, D. P. LincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154,** 26 (2013).

19. Quinn, J. J. & Chang, H. Y. Unique Features of Long Non-Coding RNA Biogenesis and Function. *Nat. Rev. Genet.* **17,** 47–62 (2016).

20. Cabili, M., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., *et al.* Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses. *Genes Dev.* **25,** 1915–1927 (2011).

21. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., *et al.* The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression. *Genome Res.* **22,** 1775–1789 (2012).

22. Bush, S. J., Muriuki, C., McCulloch, M. E., Farquhar, I. L., Clark, E. L. & Hume, D. A. Cross-Species Inference of Long Non-Coding RNAs Greatly Expands the Ruminant Transcriptome. *Genet. Sel. Evol.* **50,** 20 (2018).

23. Kornienko, A. E., Dotter, C. P., Guenzl, P. M., Gisslinger, H., Gisslinger, B., Cleary, C., *et al.* Long Non-Coding RNAs Display Higher Natural Expression Variation than Protein-Coding Genes in Healthy Humans. *Genome Biol* **17,** 14 (2016).

24. Melé, M., Mattioli, K., Mallard, W., Shechner, D. M., Gerhardinger, C. & Rinn, J. L. Chromatin Environment, Transcriptional Regulation, and Splicing Distinguish lincRNAs and mRNAs. *Genome Res.* **27,** 27–37 (2017).

25. Zhou, Z. Y., Li, A., Wang, L. G., Irwin, D. M., Liu, Y. H., Xu, D., *et al.* DNA Methylation Signatures of Long Intergenic Noncoding RNAs in Porcine Adipose and Muscle Tissues. *Sci. Rep.* **5,** 1–8 (2015).

26. Sati, S., Ghosh, S., Jain, V., Scaria, V. & Sengupta, S. Genome-Wide Analysis Reveals Distinct Patterns of Epigenetic Features in Long Non-Coding RNA Loci. *Nucleic Acids Res.* **40,** 10018–10031 (2012).

27. St.Laurent, G., Wahlestedt, C. & Kapranov, P. The Landscape of Long Noncoding RNA Classification. *Trends Genet.* **31,** 239–251 (2015).

28. West, J. A., Davis, C. P., Sunwoo, H., Simon, M. D., Sadreyev, R. I., Wang, P. I., *et al.* The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites. *Mol. Cell* **55,** 791–802 (2014).

29. Lee, S., Kopp, F., Chang, T. C., Sataluri, A., Chen, B., Sivakumar, S., *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164,** 69–80 (2016).

30. Luo, S., Lu, J. Y., Liu, L., Yin, Y., Chen, C., Han, X., *et al.* Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell* **18,** 637–652 (2016).

31. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

32. Ntini, E., Järvelin, A. I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., *et al.* Polyadenylation Site-Induced Decay of Upstream Transcripts Enforces Promoter Directionality. *Nat. Struct. Mol. Biol.* **20,** 923–928 (2013).

33. Sigova, A. A., Mullen, A. C., Molinie, B., Gupta, S., Orlando, D. A., Guenther, M. G., *et al.* Divergent Transcription of Long Noncoding RNA/mRNA Gene Pairs in Embryonic Stem Cells. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2876–2881 (2013).

34. Mattioli, K., Volders, P.-J., Gerhardinger, C., Lee, J. C., Maass, P. G., Melé, M., *et al.* High-Throughput Functional Analysis of lncRNA Core Promoters Elucidates Rules Governing Tissue Specificity. *Genome Res.* **29,** 344–355 (2019).

35. Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., *et al.* Widespread Transcription at Neuronal Activity-Regulated Enhancers. *Nature* **465,** 182–187 (2010).

36. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., *et al.* An Atlas of Active Enhancers across Human Cell Types and Tissues. *Nature* **507,** 455–461 (2014).

37. Andersson, R. Promoter or Enhancer, What's the Difference? Deconstruction of Established Distinctions and Presentation of a Unifying Model. *BioEssays* **37,** 314–323 (2015).

38. Khorkova, O., Myers, A. J., Hsiao, J. & Wahlestedt, C. Natural Antisense Transcripts. *Hum. Mol. Genet.* **23** (2014).

39. Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., *et al.* Molecular Biology: Antisense Transcription in the Mammalian Transcriptome. *Science* **309,** 1564–1566 (2005).

40. Pelechano, V. & Steinmetz, L. M. Gene Regulation by Antisense Transcription. *Nat. Rev. Genet.* **14,** 880–893 (2013).

41. Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., *et al.* Long Non-Coding RNA ANRIL Is Required for the PRC2 Recruitment to and Silencing of P15 INK4B Tumor Suppressor Gene. *Oncogene* **30,** 1956–1962 (2011).

42. Nakaya, H. I., Amaral, P. P., Louro, R., Lopes, A., Fachel, A. A., Moreira, Y. B., *et al.* Genome Mapping and Expression Analyses of Human Intronic Noncoding RNAs Reveal Tissue-Specific Patterns and Enrichment in Genes Related to Regulation of Transcription. *Genome Biol.* **8** (2007).

43. Chan, S. N. & Pek, J. W. Stable Intronic Sequence RNAs (sisRNAs): An Expanding Universe. *Trends Biochem. Sci.* **44,** 258–272 (2019).

44. Kristensen, L. S., Andersen, M. S., Stagsted, L. V., Ebbesen, K. K., Hansen, T. B. & Kjems, J. The Biogenesis, Biology and Characterization of Circular RNAs. *Nat. Rev. Genet.* **20,** 675–691 (2019).

45. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., *et al.* GENCODE: The Reference Human Genome Annotation for the ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

46. Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., *et al.* The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods. *Nature* **505,** 635–640 (2014).

47. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental Dynamics of lncRNAs across Mammalian Organs and Species. *Nature* **571,** 510–514 (2019).

48. Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P. & Ulitsky, I. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep.* **11,** 1110–1122 (2015).

49. Chen, J., Shishkin, A. A., Zhu, X., Kadri, S., Maza, I., Guttman, M., *et al.* Evolutionary Analysis across Mammals Reveals Distinct Classes of Long Non-Coding RNAs. *Genome Biol.* **17,** 1–17 (2016).

50. Washietl, S., Kellis, M. & Garber, M. Evolutionary Dynamics and Tissue Specificity of Human Long Noncoding RNAs in Six Mammals. *Genome Res.* **24,** 616–628 (2014).

51. Darbellay, F. & Necsulea, A. Comparative Transcriptomics Analyses across Species, Organs, and Developmental Stages Reveal Functionally Constrained lncRNAs. *Mol Biol Evol* **37,** 240–259 (2020).

52. Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., *et al.* Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. *PLoS Genet.* **8** (2012).

53. Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., *et al.* Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-Coding RNAs in Mammals. *Nature* **458,** 223–227 (2009).

54. Ulitsky, I. Evolution to the Rescue: Using Comparative Genomics to Understand Long Non-Coding RNAs. *Nat. Rev. Genet.* **17,** 601–614 (2016).

55. Diederichs, S. The Four Dimensions of Noncoding RNA Conservation. *Trends Genet.* **30,** 121–123 (2014).

56. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread Purifying Selection on RNA Structure in Mammals. *Nucleic Acids Res.* **41,** 8220–8236 (2013).

57. Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., *et al.* Local Regulation of Gene Expression by lncRNA Promoters, Transcription and Splicing. *Nature* **539,** 452–455 (2016).

58. Herrera-Úbeda, C., Barba, M. M., Pérez, E. N., Gravemeyer, J., Albuixech-Crespo, B., Wheeler, G. N., *et al.* Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates despite Absence of Sequence Conservation. *Biology* **8** (2019).

59. Amaral, P. P., Leonardi, T., Han, N., Gascoigne, D. K., Zhang, A., Pluchino, S., *et al.* Genomic Positional Conservation Identifies Topological Anchor Point (Tap) RNAs. *Genome Biol.*, 1–21 (2018).

60. Petermann, F., Pękowska, A., Johnson, C. A., Jankovic, D., Shih, H. Y., Jiang, K., *et al.* The Magnitude of IFN-γ Responses Is Fine-Tuned by DNA Architecture and the Noncoding Transcript of Ifng-as1. *Mol. Cell* **75,** 1229–1242.e5 (2019).

61. Loda, A. & Heard, E. Xist RNA in Action: Past, Present, and Future. *PLoS Genet.* **15,** 1–17 (2019).

62. Marchese, F. P., Raimondi, I. & Huarte, M. The Multidimensional Mechanisms of Long Noncoding RNA Function. *Genome Biol* **18,** 206 (2017).

63. Ramilowski, J. A., Yip, C. W., Agrawal, S., Chang, J.-C., Ciani, Y., Kulakovskiy, I. V., *et al.* Functional Annotation of Human Long Noncoding RNAs via Molecular Phenotyping. *Genome Res.* **30,** 1060–1072 (2020).

64. Rodriguez-Lopez, M., Anver, S., Cotobal, C., Kamrad, S., Malecki, M., Correia-Melo, C., *et al.* Functional Profiling of Long Intergenic Non-Coding RNAs in Fission Yeast. *eLife* **11,** 1–27 (2022).

65. Kopp, F. & Mendell, J. T. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **172,** 393–407 (2018).

66. Yao, R. W., Wang, Y. & Chen, L. L. Cellular Functions of Long Noncoding RNAs. *Nat. Cell Biol.* **21,** 542–551 (2019).

67. Wang, K. C. & Chang, H. Y. Molecular Mechanisms of Long Noncoding RNAs. *Mol. Cell* **43,** 904–914 (2011).

68. Gil, N. & Ulitsky, I. Regulation of Gene Expression by Cis-Acting Long Non-Coding RNAs. *Nat. Rev. Genet.* **21,** 102–117 (2020).

69. Fanucchi, S., Fok, E. T., Dalla, E., Shibayama, Y., Börner, K., Chang, E. Y., *et al.* Immune Genes Are Primed for Robust Transcription by Proximal Long Noncoding RNAs Located in Nuclear Compartments. *Nat. Genet.* **51,** 138–150 (2019).

70. Kotzin, J. J., Spencer, S. P., McCright, S. J., Kumar, D. B., Collet, M. A., Mowel, W. K., *et al.* The Long Non-Coding RNA Morrbid Regulates Bim and Short-Lived Myeloid Cell Lifespan. *Nature* **537,** 239–243 (2016).

71. Kotzin, J. J., Iseka, F., Wright, J., Basavappa, M. G., Clark, M. L., Ali, M. A., *et al.* The Long Noncoding RNA Morrbid Regulates CD8 T Cells in Response to Viral Infection. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 11916–11925 (2019).

72. Ali, T. & Grote, P. Beyond the RNA-dependent Function of LncRNA Genes. *eLife* **9,** 1–14 (2020).

73. Isoda, T., Moore, A. J., He, Z., Chandra, V., Aida, M., Denholtz, M., *et al.* Non-Coding Transcription Instructs Chromatin Folding and Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. *Cell* **171,** 103–119.e18 (2017).

74. Cho, S. W., Xu, J., Sun, R., Mumbach, M. R., Carter, A. C., Chen, Y. G., *et al.* Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell* **173,** 1398–1412.e22 (2018).

75. Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., *et al.* The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse. *Dev. Cell* **24,** 206–214 (2013).

76. Liu, B., Sun, L., Liu, Q., Gong, C., Yao, Y., Lv, X., *et al.* A Cytoplasmic NF-κB Interacting Long Noncoding RNA Blocks IκB Phosphorylation and Suppresses Breast Cancer Metastasis. *Cancer Cell* **27,** 370–381 (2015).

77. Ng, S. Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The Long Noncoding RNA RMST Interacts with SOX2 to Regulate Neurogenesis. *Mol. Cell* **51,** 349–359 (2013).

78. Kleaveland, B., Shi, C. Y., Stefano, J. & Bartel, D. P. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* **174,** 350–362.e17 (2018).

79. Sanger, F., Nicklen, S. & Coulson, A. R. DNA Sequencing with Chain-Terminating Inhibitors. *Proc Natl Acad Sci U S A* **74,** 5463–5467 (1977).

80. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of Age: Ten Years of next-Generation Sequencing Technologies. *Nat Rev Genet* **17,** 333–351 (2016).

81. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T. & Salzberg, S. L. StringTie Enables Improved Reconstruction of a Transcriptome from RNA-seq Reads. *Nat. Biotechnol.* **33,** 290–295 (2015).

82. Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. & Quince, C. Insight into Biases and Sequencing Errors for Amplicon Sequencing with the Illumina MiSeq Platform. *Nucleic Acids Res.* **43** (2015).

83. Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., *et al.* Systematic Evaluation of Error Rates and Causes in Short Samples in Next-Generation Sequencing. *Sci. Rep.* **8,** 1–14 (2018).

84. Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., *et al.* Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis. *F1000Res* **6,** 100 (2017).

85. Fromm, B., Billipp, T., Peck, L. E., Johansen, M., Tarver, J. E., King, B. L., *et al.* A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu. Rev. Genet.* **49,** 213–242 (2015).

86. Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., *et al.* Discovering microRNAs from Deep Sequencing Data Using miRDeep. *Nat. Biotechnol.* **26,** 407–415 (2008).

87. Rueda, A., Barturen, G., Lebrón, R., Gómez-Martín, C., Alganza, Á., Oliver, J. L., *et al.* SRNAtoolbox: An Integrated Collection of Small RNA Research Tools. *Nucleic Acids Res.* **43,** W467–W473 (2015).

88. Handzlik, J. E., Tastsoglou, S., Vlachos, I. S. & Hatzigeorgiou, A. G. Manatee: Detection and Quantification of Small Non-Coding RNAs from next-Generation Sequencing Data. *Sci. Rep.* **10,** 1–10 (2020).

89. Zhang, J., Eteleeb, A. M., Rozycki, E. B., Inkman, M. J., Ly, A., Scharf, R. E., *et al.* DANSR: A Tool for the Detection of Annotated and Novel Small RNAs. *Non-coding RNA* **8** (2022).

90. Riolo, G., Cantara, S., Marzocchi, C. & Ricci, C. miRNA Targets: From Prediction Tools to Experimental Validation. *Methods Protoc.* **4,** 1–20 (2021).

91. John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. & Marks, D. S. Human microRNA Targets. *PLoS Biol.* **2** (2004).

92. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting Effective microRNA Target Sites in Mammalian mRNAs. *eLife* **4** (ed Izaurralde, E.) e05005 (2015).

93. Alkan, F., Wenzel, A., Palasca, O., Kerpedjiev, P., Rudebeck, A. F., Stadler, P. F., *et al.* RIsearch2: Suffix Array-Based Large-Scale Prediction of RNA-RNA Interactions and siRNA off-Targets. *Nucleic Acids Res* **45,** e60 (2017).

94. McGeary, S. E., Lin, K. S., Shi, C. Y., Pham, T. M., Bisaria, N., Kelley, G. M., *et al.* The Biochemical Basis of microRNA Targeting Efficacy. *Science* **366** (2019).

95. Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., *et al.* A Uniform System for microRNA Annotation. *RNA* **9,** 277–279 (2003).

96. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. MiRBase: From microRNA Sequences to Function. *Nucleic Acids Res.* **47,** D155–D162 (2019).

97. Fromm, B., Høye, E., Domanska, D., Zhong, X., Aparicio-Puerta, E., Ovchinnikov, V., *et al.* MirGeneDB 2.1: Toward a Complete Sampling of All Major Animal Phyla. *Nucleic Acids Res.* **50,** D204–D210 (2022).

98. Bourdon, C., Bardou, P., Aujean, E., Le Guillou, S., Tosser-Klopp, G. & Le Provost, F. RumimiR: A Detailed microRNA Database Focused on Ruminant Species. *Database (Oxford)* **2019,** baz099 (2019).

99. Mudge, J. M. & Harrow, J. The State of Play in Higher Eukaryote Gene Annotation. *Nat. Rev. Genet.* **17,** 758–772 (2016).

100. Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O. & Thompson, J. D. A Benchmark Study of Ab Initio Gene Prediction Methods in Diverse Eukaryotic Organisms. *BMC Genomics* **21,** 1–20 (2020).

101. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a Complete Map of the Human Long Non-Coding RNA Transcriptome. *Nat. Rev. Genet.* **19,** 535–548 (2018).

102. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., *et al.* A Survey of Best Practices for RNA-seq Data Analysis. *Genome Biol* **17,** 13 (2016).

103. Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Pérez-Lluch, S., Abad, A., Davis, C., *et al.* High-Throughput Annotation of Full-Length Long Noncoding RNAs with Capture Long-Read Sequencing. *Nat. Genet.* **49,** 1731–1740 (2017).

104. Chang, T., An, B., Liang, M., Duan, X., Du, L., Cai, W., *et al.* PacBio Single-Molecule Long-Read Sequencing Provides New Light on the Complexity of Full-Length Transcripts in Cattle. *Front. Genet.* **12,** 1–11 (2021).

105. Takahashi, H., Kato, S., Murata, M. & Carninci, P. in *Methods in Molecular Biology* (eds Deplancke, B. & Gheldof, N.) 181–200 (Humana Press, Totowa, NJ, 2012).

106. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-Fidelity Promoter Profiling Reveals Widespread Alternative Promoter Usage and Transposon-Driven Developmental Gene Expression. *Genome Res.* **23,** 169–180 (2013).

107. Derti, A., Garrett-Engele, P., MacIsaac, K. D., Stevens, R. C., Sriram, S., Chen, R., *et al.* A Quantitative Atlas of Polyadenylation in Five Mammals. *Genome Res.* **22,** 1173–1183 (2012).

108. Mercer, T. R., Clark, M. B., Crawford, J., Brunck, M. E., Gerhardt, D. J., Taft, R. J., *et al.* Targeted Sequencing for Gene Discovery and Quantification Using RNA CaptureSeq. *Nat. Protoc.* **9,** 989–1009 (2014).

109. Lagarrigue, S., Lorthiois, M., Degalez, F., Gilot, D. & Derrien, T. LncRNAs in Domesticated Animals: From Dog to Livestock Species. *Mamm Genome* **33,** 248–270 (2021).

110. Klapproth, C., Sen, R., Stadler, P. F., Findeiß, S. & Fallmann, J. Common Features in lncRNA Annotation and Classification: A Survey. *Non-Coding RNA* **7,** 77 (2021).

111. Sielemann, K., Hafner, A. & Pucker, B. The Reuse of Public Datasets in the Life Sciences: Potential Risks and Rewards. *PeerJ* **8,** e9954 (2020).

112. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **26,** 139–140 (2010).

113. Love, M. I., Huber, W. & Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2. *Genome Biol.* **15,** 550 (2014).

114. Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. Exaggerated False Positives by Popular Differential Expression Methods When Analyzing Human Population Samples. *Genome Biol.* **23,** 1–13 (2022).

115. Wolfe, C. J., Kohane, I. S. & Butte, A. J. Systematic Survey Reveals General Applicability of "Guilt-by-Association" within Gene Coexpression Networks. *BMC Bioinformatics* **6,** 227 (2005).

116. Savino, A., Provero, P. & Poli, V. Differential Co-Expression Analyses Allow the Identification of Critical Signalling Pathways Altered during Tumour Transformation and Progression. *Int J Mol Sci* **21,** E9461 (2020).

117. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., *et al.* Gencode 2021. *Nucleic Acids Res.* **49,** D916–D923 (2021).

118. Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., *et al.* NONCODEV6: An Updated Database Dedicated to Long Non-Coding RNA Annotation in Both Animals and Plants. *Nucleic Acids Res.* **49,** D165–D171 (2021).

119. Li, A., Zhang, J., Zhou, Z., Wang, L., Liu, Y. & Liu, Y. ALDB: A Domestic-Animal Long Noncoding RNA Database. *PLoS One* **10,** e0124003 (2015).

120. Kosinska-Selbi, B., Mielczarek, M. & Szyda, J. Review: Long Non-Coding RNA in Livestock. *Animal* **14,** 2003–2013 (2020).

121. Bovine Genome Sequencing and Analysis Consortium, Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., *et al.* The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* **324,** 522–528 (2009).

122. Li, X., Yang, J., Shen, M., Xie, X. L., Liu, G. J., Xu, Y. X., *et al.* Whole-Genome Resequencing of Wild and Domestic Sheep Identifies Genes Associated with Morphological and Agronomic Traits. *Nat. Commun.* **11,** 1–16 (2020).

123. Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., *et al.* Large-Scale Ruminant Genome Sequencing Provides Insights into Their Evolution and Distinct Traits. *Science* **364** (2019).

124. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., *et al.* A Comparative Encyclopedia of DNA Elements in the Mouse Genome. *Nature* **515,** 355–364 (2014).

125. Hu, Z. L., Park, C. A. & Reecy, J. M. Bringing the Animal QTLdb and CorrDB into the Future: Meeting New Challenges and Providing Updated Services. *Nucleic Acids Res.* **50,** D956–D961 (2022).

126. Weikard, R., Demasius, W. & Kuehn, C. Mining Long Noncoding RNA in Livestock. *Anim. Genet.* **48,** 3–18 (2017).

127. Wang, X., Gu, Z. & Jiang, H. MicroRNAs in Farm Animals. *Animal* **7,** 1567–1575 (2013).

128. Davis, E., Caiment, F., Tordoir, X., Cavaillé, J., Ferguson-Smith, A., Cockett, N., *et al.* RNAi-mediated Allelic Trans-Interaction at the Imprinted Rtl1/Peg11 Locus. *Curr. Biol.* **15,** 743–749 (2005).

129. Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibé, B., *et al.* A Mutation Creating a Potential Illegitimate microRNA Target Site in the Myostatin Gene Affects Muscularity in Sheep. *Nat. Genet.* **38,** 813–818 (2006).

130. Do, D. N., Dudemaine, P.-L., Mathur, M., Suravajhala, P., Zhao, X. & Ibeagha-Awemu, E. M. miRNA Regulatory Functions in Farm Animal Diseases, and Biomarker Potentials for Effective Therapies. *Int J Mol Sci* **22,** 3080 (2021).

131. Miretti, S., Lecchi, C., Ceciliani, F. & Baratta, M. MicroRNAs as Biomarkers for Animal Health and Welfare in Livestock. *Front Vet Sci* **7,** 578193 (2020).

132. Giuffra, E. & Tuggle, C. K. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu. Rev. Anim. Biosci.* **7,** 65–88 (2019).

133. Scott, E. Y., Mansour, T., Bellone, R. R., Brown, C. T., Mienaltowski, M. J., Penedo, M. C., *et al.* Identification of Long Non-Coding RNA in the Horse Transcriptome. *BMC Genomics* **18,** 511 (2017).

134. Clark, E. L., Bush, S. J., McCulloch, M. E. B., Farquhar, I. L., Young, R., Lefevre, L., *et al.* A High Resolution Atlas of Gene Expression in the Domestic Sheep (Ovis Aries). *PLOS Genetics* **13,** e1006997 (2017).

135. Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., *et al.* Multi-Species Annotation of Transcriptome and Chromatin Structure in Domesticated Animals. *BMC Biol* **17,** 108 (2019).

136. Koufariotis, L. T., Chen, Y.-P. P., Chamberlain, A., Vander Jagt, C. & Hayes, B. J. A Catalogue of Novel Bovine Long Noncoding RNA across 18 Tissues. *PLoS One* **10,** e0141225 (2015).

137. Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H., *et al.* Genome-Wide Identification of Tissue-Specific Long Non-Coding RNA in Three Farm Animal Species. *BMC Genomics* **19,** 684 (2018).

138. Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L. & Burt, D. W. Normalized Long Read RNA Sequencing in Chicken Reveals Transcriptome Complexity Similar to Human. *BMC Genomics* **18,** 323 (2017).

139. Miao, X., Luo, Q., Zhao, H. & Qin, X. Ovarian Transcriptomic Study Reveals the Differential Regulation of miRNAs and lncRNAs Related to Fecundity in Different Sheep. *Sci Rep* **6,** 35299 (2016).

140. Feng, X., Li, F., Wang, F., Zhang, G., Pang, J., Ren, C., *et al.* Genome-Wide Differential Expression Profiling of mRNAs and lncRNAs Associated with Prolificacy in Hu Sheep. *Biosci. Rep.* **38,** 1–14 (2018).

141. La, Y., Tang, J., He, X., Di, R., Wang, X., Liu, Q., *et al.* Identification and Characterization of mRNAs and lncRNAs in the Uterus of Polytocous and Monotocous Small Tail Han Sheep (Ovis Aries). *PeerJ* **2019,** 1–21 (2019).

142. La, Y., He, X., Zhang, L., Di, R., Wang, X., Gan, S., *et al.* Comprehensive Analysis of Differentially Expressed Profiles of mRNA, lncRNA, and circRNA in the Uterus of Seasonal Reproduction Sheep. *Genes* **11,** 301 (2020).

143. Liu, A., Liu, M., Li, Y., Chen, X., Zhang, L. & Tian, S. Differential Expression and Prediction of Function of lncRNAs in the Ovaries of Low and High Fecundity Hanper Sheep. *Reprod. Domest. Anim.* **56,** 604–620 (2021).

144. Matsuno, Y., Kusama, K. & Imakawa, K. Characterization of lncRNA Functioning in Ovine Conceptuses and Endometria during the Peri-Implantation Period. *Biochem. Biophys. Res. Commun.* **594,** 22–30 (2022).

145. Zhang, Z., Tang, J., Di, R., Liu, Q., Wang, X., Gan, S., *et al.* Comparative Transcriptomics Reveal Key Sheep (Ovis Aries) Hypothalamus LncRNAs That Affect Reproduction. *Animals* **9,** 152 (2019).

146. Zheng, J., Wang, Z., Yang, H., Yao, X., Yang, P., Ren, C. F., *et al.* Pituitary Transcriptomic Study Reveals the Differential Regulation of lncRNAs and mRNAs Related to Prolificacy in Different FecB Genotyping Sheep. *Genes* **10,** 1–17 (2019).

147. Li, X., Li, C., Xu, Y., Yao, R., Li, H., Ni, W., *et al.* Analysis of Pituitary Transcriptomics Indicates That lncRNAs Are Involved in the Regulation of Sheep Estrus. *Funct. Integr. Genomics* **20,** 563–573 (2020).

148. Xia, Q., Li, Q., Gan, S., Guo, X., Zhang, X., Zhang, J., *et al.* Exploring the Roles of Fecundity-Related Long Non-Coding RNAs and mRNAs in the Adrenal Glands of Small-Tailed Han Sheep. *BMC Genet.* **21,** 1–11 (2020).

149. Li, C., He, X., Zhang, Z., Ren, C. & Chu, M. Pineal Gland Transcriptomic Profiling Reveals the Differential Regulation of lncRNA and mRNA Related to Prolificacy in STH Sheep with Two FecB Genotypes. *BMC Genomic Data* **22,** 1–17 (2021).

150. He, X., Tao, L., Zhong, Y., Di, R., Xia, Q., Wang, X., *et al.* Photoperiod Induced the Pituitary Differential Regulation of lncRNAs and mRNAs Related to Reproduction in Sheep. *PeerJ* **9,** 1–17 (2021).

151. Chen, S., Guo, X., He, X., Di, R., Zhang, X., Zhang, J., *et al.* Transcriptome Analysis Reveals Differentially Expressed Genes and Long Non-coding RNAs Associated With Fecundity in Sheep Hypothalamus With Different FecB Genotypes. *Front Cell Dev Biol* **9,** 633747 (2021).

152. Chen, S., Guo, X., He, X., Di, R., Zhang, X., Zhang, J., *et al.* Insight Into Pituitary lncRNA and mRNA at Two Estrous Stages in Small Tail Han Sheep With Different FecB Genotypes. *Front. Endocrinol.* **12,** 1–15 (2022).

153. Zhang, Y., Yang, H., Han, L., Li, F., Zhang, T., Pang, J., *et al.* Long Noncoding RNA Expression Profile Changes Associated with Dietary Energy in the Sheep Testis during Sexual Maturation. *Sci. Rep.* **7,** 1–13 (2017).

154. Yang, H., Wang, F., Li, F., Ren, C., Pang, J., Wan, Y., *et al.* Comprehensive Analysis of Long Noncoding RNA and mRNA Expression Patterns in Sheep Testicular Maturation. *Biol. Reprod.* **99,** 650–661 (2018).

155. Shabbir, S., Boruah, P., Xie, L., Kulyar, M. F.-E.-A., Nawaz, M., Yousuf, S., *et al.* Genome-Wide Transcriptome Profiling Uncovers Differential miRNAs and lncRNAs in Ovaries of Hu Sheep at Different Developmental Stages. *Sci Rep* **11,** 5865 (2021).

156. Li, X., Li, C., Wureli, H., Ni, W., Zhang, M., Li, H., *et al.* Screening and Evaluating of Long Non-Coding RNAs in Prenatal and Postnatal Pituitary Gland of Sheep. *Genomics* **112,** 934–942 (2020).

157. Yang, H., Ma, J., Wang, Z., Yao, X., Zhao, J., Zhao, X., *et al.* Genome-Wide Analysis and Function Prediction of Long Noncoding RNAs in Sheep Pituitary Gland Associated with Sexual Maturation. *Genes (Basel)* **11,** E320 (2020).

158. Ren, C., Deng, M., Fan, Y., Yang, H., Zhang, G., Feng, X., *et al.* Genome-Wide Analysis Reveals Extensive Changes in LncRNAs during Skeletal Muscle Development in Hu Sheep. *Genes* **8,** 191 (2017).

159. Li, C. Y., Li, X., Liu, Z., Ni, W., Zhang, X., Hazi, W., *et al.* Identification and Characterization of Long Non-Coding RNA in Prenatal and Postnatal Skeletal Muscle of Sheep. *Genomics* **111,** 133–141 (2019).

160. Chao, T., Ji, Z., Hou, L., Wang, J., Zhang, C., Wang, G., *et al.* Sheep Skeletal Muscle Transcriptome Analysis Reveals Muscle Growth Regulatory lncRNAs. *PeerJ* **2018,** e4619 (2018).

161. Li, Q., Liu, R., Zhao, H., Di, R., Lu, Z., Liu, E., *et al.* Identification and Characterization of Long Noncoding RNAs in Ovine Skeletal Muscle. *Animals* **8,** 1–16 (2018).

162. Yuan, C., Zhang, K., Yue, Y., Guo, T., Liu, J., Niu, C., *et al.* Analysis of Dynamic and Widespread lncRNA and miRNA Expression in Fetal Sheep Skeletal Muscle. *PeerJ* **8,** 1–21 (2020).

163. Bakhtiarizadeh, M. R. & Salami, S. A. Identification and Expression Analysis of Long Noncoding RNAs in Fat-Tail of Sheep Breeds. *G3 Genes Genomes Genet.* **9,** 1263–1276 (2019).

164. Ma, L., Zhang, M., Jin, Y., Erdenee, S., Hu, L., Chen, H., *et al.* Comparative Transcriptome Profiling of mRNA and lncRNA Related to Tail Adipose Tissues of Sheep. *Front Genet* **9,** 365 (2018).

165. Han, F., Li, J., Zhao, R., Liu, L., Li, L., Li, Q., *et al.* Identification and Co-Expression Analysis of Long Noncoding RNAs and mRNAs Involved in the Deposition of Intramuscular Fat in Aohan Fine-Wool Sheep. *BMC Genomics* **22,** 1–14 (2021).

166. Xiao, C., Wei, T., Liu, L. X., Liu, J. Q., Wang, C. X., Yuan, Z. Y., *et al.* Whole-Transcriptome Analysis of Preadipocyte and Adipocyte and Construction of Regulatory Networks to Investigate Lipid Metabolism in Sheep. *Front Genet* **12,** 662143 (2021).

167. He, X., Wu, R., Yun, Y., Qin, X., Chen, L., Han, Y., *et al.* Transcriptome Analysis of Messenger RNA and Long Noncoding RNA Related to Different Developmental Stages of Tail Adipose Tissues of Sunite Sheep. *Food Sci. Nutr.* **9,** 5722–5734 (2021).

168. Nie, Y., Li, S., Zheng, X., Chen, W., Li, X., Liu, Z., *et al.* Transcriptome Reveals Long Non-coding RNAs and mRNAs Involved in Primary Wool Follicle Induction in Carpet Sheep Fetal Skin. *Front Physiol* **9,** 446 (2018).

169. Yue, Y., Guo, T., Yuan, C., Liu, J., Guo, J., Feng, R., *et al.* Integrated Analysis of the Roles of Long Noncoding RNA and Coding RNA Expression in Sheep (Ovis Aries) Skin during Initiation of Secondary Hair Follicle. *PLoS ONE* **11,** 1–20 (2016).

170. Lv, X., Gao, W., Jin, C., Wang, Y., Chen, W., Wang, L., *et al.* Divergently Expressed RNA Identification and Interaction Prediction of Long Non-Coding RNA and mRNA Involved in Hu Sheep Hair Follicle. *Sci. Rep.* **9,** 1–12 (2019).

171. Sulayman, A., Tian, K., Huang, X., Tian, Y., Xu, X., Fu, X., *et al.* Genome-Wide Identification and Characterization of Long Non-Coding RNAs Expressed during Sheep Fetal and Postnatal Hair Follicle Development. *Sci. Rep.* **9,** 1–14 (2019).

172. Lv, X., Chen, W., Sun, W., Hussain, Z., Chen, L., Wang, S., *et al.* Expression Profile Analysis to Identify Circular RNA Expression Signatures in Hair Follicle of Hu Sheep Lambskin. *Genomics* **112,** 4454–4462 (2020).

173. Zhao, R., Li, J., Liu, N., Li, H., Liu, L., Yang, F., *et al.* Transcriptomic Analysis Reveals the Involvement of lncRNA–miRNA–mRNA Networks in Hair Follicle Induction in Aohan Fine Wool Sheep Skin. *Front Genet* **11,** 590 (2020).

174. Zhao, B., Luo, H., He, J., Huang, X., Chen, S., Fu, X., *et al.* Comprehensive Transcriptome and Methylome Analysis Delineates the Biological Basis of Hair Follicle Development and Wool-Related Traits in Merino Sheep. *BMC Biol.* **19,** 1–18 (2021).

175. Li, Y., Kong, L., Deng, M., Lian, Z., Han, Y., Sun, B., *et al.* Heat Stress-Responsive Transcriptome Analysis in the Liver Tissue of Hu Sheep. *Genes* **10** (2019).

176. Hao, Z., Luo, Y., Wang, J., Hu, J., Liu, X., Li, S., *et al.* Rna-Seq Reveals the Expression Profiles of Long Non-Coding Rnas in Lactating Mammary Gland from Two Sheep Breeds with Divergent Milk Phenotype. *Animals* **10,** 1–12 (2020).

177. Chen, W., Lv, X., Wang, Y., Zhang, X., Wang, S., Hussain, Z., *et al.* Transcriptional Profiles of Long Non-coding RNA and mRNA in Sheep Mammary Gland During Lactation Period. *Front. Genet.* **11,** 1–15 (2020).

178. Zhang, D. Y., Zhang, X. X., Li, G. Z., Li, X. L., Zhang, Y. K., Zhao, Y., *et al.* Transcriptome Analysis of Long Noncoding RNAs Ribonucleic Acids from the Livers of Hu Sheep

with Different Residual Feed Intake. *Animal* **15** (2021).

179. Guo, C., Xue, Y., Sun, D., Yin, Y., Hu, F. & Mao, S. Transcriptome Profiling of Hepatic and Renal mRNAs and lncRNAs under a Nutritional Restriction during Pregnancy in a Sheep Model. *Genomics* **113,** 2769–2779 (2021).

180. Xia, Q., Chu, M., He, X., Zhang, X., Zhang, J., Guo, X., *et al.* Identification of Photoperiod-Induced LncRNAs and mRNAs in Pituitary Pars Tuberalis of Sheep. *Front. Vet. Sci.* **8,** 1–14 (2021).

181. Lu, Z., Yuan, C., Li, J., Guo, T., Yue, Y., Niu, C., *et al.* Comprehensive Analysis of Long Non-coding RNA and mRNA Transcriptomes Related to Hypoxia Adaptation in Tibetan Sheep. *Front. Vet. Sci.* **8,** 1–11 (2022).

182. Jin, C., Bao, J., Wang, Y., Chen, W., Wu, T., Wang, L., *et al.* Changes in Long Non-Coding RNA Expression Profiles Related to the Antagonistic Effects of Escherichia Coli F17 on Lamb Spleens. *Sci. Rep.* **8** (2018).

183. Chitneedi, P. K., Weikard, R., Arranz, J. J., Martínez-Valladares, M., Kuehn, C. & Gutiérrez-Gil, B. Identification of Regulatory Functions of LncRNAs Associated With T. Circumcincta Infection in Adult Sheep. *Front Genet* **12,** 685341 (2021).

184. Minguijón, E., Reina, R., Pérez, M., Polledo, L., Villoria, M., Ramírez, H., *et al.* Small Ruminant Lentivirus Infections and Diseases. *Vet. Microbiol.* **181,** 75–89 (2015).

185. Christodoulopoulos, G. Maedi–Visna: Clinical Review and Short Reference on the Disease Status in Mediterranean Countries. *Small Ruminant Research. Keynote Lectures of the 6th International Sheep Veterinary Congress* **62,** 47–53 (2006).

186. Pérez, M., Muñoz, J. A., Biescas, E., Salazar, E., Bolea, R., de Andrés, D., *et al.* Successful Visna/Maedi Control in a Highly Infected Ovine Dairy Flock Using Serologic Segregation and Management Strategies. *Prev. Vet. Med.* **112,** 423–427 (2013).

187. Larruskain, A. & Jugo, B. M. Retroviral Infections in Sheep and Goats: Small Ruminant Lentiviruses and Host Interaction. *Viruses* **5,** 2043–2061 (2013).

188. Larruskain, A., Bernales, I., Luján, L., de Andrés, D., Amorena, B. & Jugo, B. M. Expression Analysis of 13 Ovine Immune Response Candidate Genes in Visna/Maedi Disease Progression. *Comp. Immunol. Microbiol. Infect. Dis.* **36,** 405–413 (2013).

189. Stonos, N., Wootton, S. K. & Karrow, N. Immunogenetics of Small Ruminant Lentiviral Infections. *Viruses* **6,** 3311–3333 (2014).

190. Caiment, F., Charlier, C., Hadfield, T., Cockett, N., Georges, M. & Baurain, D. Assessing the Effect of the CLPG Mutation on the microRNA Catalogue of Skeletal Muscle Using High Throughput Sequencing. *Genome Res.* **20,** 1651–1662 (2010).

191. Gao, W., Sun, W., Yin, J., Lv, X., Bao, J., Yu, J., *et al.* Screening Candidate microRNAs (miRNAs) in Different Lambskin Hair Follicles in Hu Sheep. *PLoS One* **12,** e0176532 (2017).

192. Pokharel, K., Peippo, J., Honkatukia, M., Seppälä, A., Rautiainen, J., Ghanem, N., *et al.* Integrated Ovarian mRNA and miRNA Transcriptome Profiling Characterizes the Genetic Basis of Prolificacy Traits in Sheep (Ovis Aries). *BMC Genomics* **19,** 104 (2018).

193. Miao, X., Luo, Q., Qin, X. & Guo, Y. Genome-Wide Analysis of microRNAs Identifies the Lipid Metabolism Pathway to Be a Defining Factor in Adipose Tissue from Different Sheep. *Sci. Rep.* **5,** 18470 (2015).

194. Cohen, T. S. Role of MicroRNA in the Lung's Innate Immune Response. *J. Innate Immun.* **9,** 243–249 (2017).

195. Guo, Y. E. & Steitz, J. A. Virus Meets Host MicroRNA: The Destroyer, the Booster, the Hijacker. *Mol. Cell. Biol.* **34,** 3780–3787 (2014).

196. Trobaugh, D. W. & Klimstra, W. B. MicroRNA Regulation of RNA Virus Replication and Pathogenesis. *Trends Mol. Med.* **23,** 80–93 (2017).

197. Swaminathan, G., Martin-Garcia, J. & Navas-Martin, S. RNA Viruses and microRNAs: Challenging Discoveries for the 21st Century. *Physiol. Genomics* **45,** 1035–1048 (2013).

198. Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D. S. MicroRNA Targets in Drosophila. *Genome Biol* **5,** R1 (2003).

199. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13,** 2498–504 (2003).

200. Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets. *Cancer Res* **64,** 5245–5250 (2004).

201. Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., *et al.* Accurate Normalization of Real-Time Quantitative RT-PCR Data by Geometric Averaging of Multiple Internal Control Genes. *Genome Biol* **3,** RESEARCH0034 (2002).

202. Xu, M. & Mo, Y. Y. The Akt-associated microRNAs. *Cell. Mol. Life Sci.* **69,** 3601–3612 (2012).

203. Diehl, N. & Schaal, H. Make Yourself at Home: Viral Hijacking of the PI3K/Akt Signaling Pathway. *Viruses* **5,** 3192–3212 (2013).

204. Zhu, L., Yang, S., Tong, W., Zhu, J., Yu, H., Zhou, Y., *et al.* Control of the PI3K/Akt Pathway by Porcine Reproductive and Respiratory Syndrome Virus. *Arch. Virol.* **158,** 1227–1234 (2013).

205. Ehrhardt, C., Wolff, T., Pleschka, S., Planz, O., Beermann, W., Bode, J. G., *et al.* Influenza A Virus NS1 Protein Activates the PI3K/Akt Pathway To Mediate Antiapoptotic Signaling Responses. *J. Virol.* **81,** 3058–3067 (2007).

206. Pfeffer, S. R., Yang, C. H. & Pfeffer, L. M. The Role of MIR-21 in Cancer. *Drug Dev. Res.* **76,** 270–277 (2015).

207. Chen, Y., Chen, J., Wang, H., Shi, J., Wu, K., Liu, S., *et al.* HCV-induced miR-21 Contributes to Evasion of Host Immune System by Targeting MyD88 and IRAK1. *PLoS Pathog* **9,** e1003248 (2013).

208. Kanokudom, S., Vilaivan, T., Wikan, N., Thepparit, C., Smith, D. R. & Assavalapsakul, W. miR-21 Promotes Dengue Virus Serotype 2 Replication in HepG2 Cells. *Antiviral Res.* **142,** 169–177 (2017).

209. Parikh, V. N., Park, J., Nikolic, I., Channick, R., Yu, P. B., Marco, T. D., *et al.* Brief Report: Coordinated Modulation of Circulating miR-21 in HIV, HIV-associated Pulmonary Arterial Hypertension, and HIV/Hepatitis C Virus Coinfection. *J Acquir Immune Defic Syndr* **70,** 236–241 (2015).

210. Yang, G.-D., Huang, T.-J., Peng, L.-X., Yang, C.-F., Liu, R.-Y., Huang, H.-B., *et al.* Epstein-Barr Virus_Encoded LMP1 Upregulates microRNA-21 to Promote the Resistance of Nasopharyngeal Carcinoma Cells to Cisplatin-Induced Apoptosis by Suppressing PDCD4 and Fas-L. *PLoS One* **8,** e78355 (2013).

211. Anastasiadou, E., Garg, N., Bigi, R., Yadav, S., Campese, A. F., Lapenta, C., *et al.* Epstein-Barr Virus Infection Induces miR-21 in Terminally Differentiated Malignant B Cells. *Int J Cancer* **137,** 1491–1497 (2015).

212. Pépin, M., Vitu, C., Russo, P., Mornex, J. F. & Peterhans, E. Maedi-Visna Virus Infection in Sheep: A Review. *Vet. Res.* **29,** 341–67 (1998).

213. Gayo, E., Polledo, L., Balseiro, A., Martínez, C. P., García Iglesias, M. J., Preziuso, S., *et al.* Inflammatory Lesion Patterns in Target Organs of Visna/Maedi in Sheep and Their Significance in the Pathogenesis and Diagnosis of the Infection. *J. Comp. Pathol.* **159,** 49–56 (2018).

214. Blacklaws, B. A. Small Ruminant Lentiviruses: Immunopathogenesis of Visna-Maedi and Caprine Arthritis and Encephalitis Virus. *Comp. Immunol. Microbiol. Infect. Dis.* **35,** 259–269 (2012).

215. Liu, G., Friggeri, A., Yang, Y., Milosevic, J., Ding, Q., Thannickal, V. J., *et al.* miR-21 Mediates Fibrogenic Activation of Pulmonary Fibroblasts and Lung Fibrosis. *J. Exp. Med.* **207,** 1589–1597 (2010).

216. He, S., Li, L., Sun, S., Zeng, Z., Lu, J. & Xie, L. A Novel Murine Chronic Obstructive Pulmonary Disease Model and the Pathogenic Role of MicroRNA-21. *Front Physiol* **9,** 503 (2018).

217. Kral, J. B., Kuttke, M., Schrottmaier, W. C., Birnecker, B., Warszawska, J., Wernig, C., *et al.* Sustained PI3K Activation Exacerbates BLM-induced Lung Fibrosis via Activation of pro-Inflammatory and pro-Fibrotic Pathways. *Sci. Rep.* **6,** 23034 (2016).

218. Yang, S., Banerjee, S., d. Freitas, A., Cui, H., Xie, N., Abraham, E., *et al.* miR-21 Regulates Chronic Hypoxia-Induced Pulmonary Vascular Remodeling. *AJP Lung Cell. Mol. Physiol.* **302,** L521–L529 (2012).

219. Sheedy, F. J. Turning 21: Induction of miR-21 as a Key Switch in the Inflammatory Response. *Front Immunol* **6,** 19 (2015).

220. Gonzalez-Martin, A., Adams, B. D., Lai, M., Shepherd, J., Salvador-Bernaldez, M., Salvador, J. M., *et al.* The microRNA miR-148a Functions as a Critical Regulator of B Cell Tolerance and Autoimmunity. *Nat. Immunol.* **17,** 433–440 (2016).

221. Yu, S.-h., Zhang, C.-l., Dong, F.-s. & Zhang, Y.-m. miR-99a Suppresses the Metastasis of Human Non-Small Cell Lung Cancer Cells by Targeting AKT1 Signaling Pathway. *J. Cell. Biochem.* **116,** 268–276 (2015).

222. Huang, H. G., Luo, X., Wu, S. & Jian, B. MiR-99a Inhibits Cell Proliferation and Tumorigenesis through Targeting mTOR in Human Anaplastic Thyroid Cancer. *Asian Pac. J. Cancer Prev.* **16,** 4937–4944 (2015).

223. Kumar, M., Ahmad, T., Sharma, A., Mabalirajan, U., Kulshreshtha, A., Agrawal, A., *et al.* Let-7 microRNA-mediated Regulation of IL-13 and Allergic Airway Inflammation. *J Allergy Clin Immunol* **128,** 1077-1085.e1–10 (2011).

224. Du, J., Gao, S., Tian, Z., Xing, S., Huang, D., Zhang, G., *et al.* MicroRNA Expression Profiling of Primary Sheep Testicular Cells in Response to Bluetongue Virus Infection. *Infect. Genet. Evol.* **49,** 256–267 (2017).

225. Cameron, J. E., Fewell, C., Yin, Q., McBride, J., Wang, X., Lin, Z., *et al.* Epstein-Barr Virus Growth/Latency III Program Alters Cellular microRNA Expression. *Virology* **382,** 257–266 (2008).

226. Wang, D., Cao, L., Xu, Z., Fang, L., Zhong, Y., Chen, Q., *et al.* MiR-125b Reduces Porcine Reproductive and Respiratory Syndrome Virus Replication by Negatively Regulating the NF-κB Pathway. *PLoS One* **8,** e55838 (2013).

227. Tenoever, B. R. RNA Viruses and the Host microRNA Machinery. *Nat. Rev. Microbiol.* **11,** 169–180 (2013).

228. Liu, H. C., Hicks, J. A., Trakooljul, N. & Zhao, S. H. Current Knowledge of microRNA Characterization in Agricultural Animals. *Anim. Genet.* **41,** 225–231 (2010).

229. Pacholewska, A., Mach, N., Mata, X., Vaiman, A., Schibler, L., Barrey, E., *et al.* Novel Equine Tissue miRNAs and Breed-Related miRNA Expressed in Serum. *BMC Genomics* **17,** 831 (2016).

230. Sun, H.-Z., Chen, Y. & Guan, L. L. MicroRNA Expression Profiles across Blood and Different Tissues in Cattle. *Sci Data* **6,** 190013 (2019).

231. Hou, L., Ji, Z., Wang, G., Wang, J., Chao, T. & Wang, J. Identification and Characterization of microRNAs in the Intestinal Tissues of Sheep (Ovis Aries). *PLoS One* **13,** e0193371 (2018).

232. Long, K., Feng, S., Ma, J., Zhang, J., Jin, L., Tang, Q., *et al.* Small Non-Coding RNA Transcriptome of Four High-Altitude Vertebrates and Their Low-Altitude Relatives. *Sci Data* **6,** 192 (2019).

233. Bilbao-Arribas, M., Abendaño, N., Varela-Martínez, E., Reina, R., de Andrés, D. & Jugo, B. M. Expression Analysis of Lung miRNAs Responding to Ovine VM Virus Infection by RNA-seq. *BMC Genomics* **20,** 62 (2019).

234. Varela-Martínez, E., Bilbao-Arribas, M., Abendaño, N., Asín, J., Pérez, M., de Andrés, D., *et al.* Whole Transcriptome Approach to Evaluate the Effect of Aluminium Hydroxide in Ovine Encephalon. *Sci Rep* **10,** 15240 (2020).

235. Varela-Martínez, E., Abendaño, N., Asín, J., Sistiaga-Poveda, M., Pérez, M. M., Reina, R., *et al.* Molecular Signature of Aluminum Hydroxide Adjuvant in Ovine PBMCs by Integrated mRNA and microRNA Transcriptome Sequencing. *Front. Immunol.* **9,** 2406 (2018).

236. Olive, V., Minella, A. C. & He, L. Outside the Coding Genome, Mammalian microRNAs Confer Structural and Functional Complexity. *Sci. Signal.* **8,** re2 (2015).

237. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **30,** 2114–2120 (2014).

238. Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W. & Rajewsky, N. MiRDeep2 Accurately Identifies Known and Hundreds of Novel microRNA Genes in Seven Animal Clades. *Nucleic Acids Res.* **40,** 37–52 (2012).

239. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biol.* **10,** R25 (2009).

240. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., *et al.* BLAST+: Architecture and Applications. *BMC Bioinformatics* **10,** 421 (2009).

241. Quinlan, A. R. & Hall, I. M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **26,** 841–842 (2010).

242. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **28,** 3150–3152 (2012).

243. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties. *Bioinformatics* **33,** 2938–2940 (2017).

244. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., *et al.* Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification. *Bioinformatics* **21,** 650–659 (2005).

245. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., *et al.* G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update). *Nucleic Acids Res.* **47,** W191–W198 (2019).

246. Yang, J., Li, X., Cao, Y. H., Pokharel, K., Hu, X. J., Chen, Z. H., *et al.* Comparative mRNA and miRNA Expression in European Mouflon (Ovis Musimon) and Sheep (Ovis Aries) Provides Novel Insights into the Genetic Mechanisms for Female Reproductive Success. *Heredity* **122,** 172–186 (2019).

247. Zhou, G., Wang, X., Yuan, C., Kang, D., Xu, X., Zhou, J., *et al.* Integrating miRNA and mRNA Expression Profiling Uncovers miRNAs Underlying Fat Deposition in Sheep. *Biomed Res Int* **2017,** 1857580 (2017).

248. Pokharel, K., Peippo, J., Weldenegodguad, M., Honkatukia, M., Li, M. H. & Kantanen, J. Gene Expression Profiling of Corpus Luteum Reveals Important Insights about Early Pregnancy in Domestic Sheep. *Genes* **11** (2020).

249. Pokharel, K., Peippo, J., Li, M. H. & Kantanen, J. Identification and Characterization of miRNAs during Early Pregnancy in Domestic Sheep. *Anim. Genet.* **51,** 833–836 (2020).

250. Gu, B., Liu, H., Han, Y., Chen, Y. & Jiang, H. Integrated Analysis of miRNA and mRNA Expression Profiles in 2-, 6-, and 12-Month-Old Small Tail Han Sheep Ovaries Reveals That Oar-miR-432 Downregulates RPS6KA1 Expression. *Gene* **710,** 76–90 (2019).

251. Salavati, M., Caulton, A., Clark, R., Gazova, I., Smith, T. P. L., Worley, K. C., *et al.* Global Analysis of Transcription Start Sites in the New Ovine Reference Genome (Oar Rambouillet v1.0). *Front Genet* **11,** 580580 (2020).

252. Li, T., Luo, R., Wang, X., Wang, H., Zhao, X., Guo, Y., *et al.* Unraveling Stage-Dependent Expression Patterns of Circular RNAs and Their Related ceRNA Modulation in Ovine Postnatal Testis Development. *Front Cell Dev Biol* **9,** 627439 (2021).

253. Hao, Z. Y., Wang, J. Q., Luo, Y. L., Liu, X., Li, S. B., Zhao, M. L., *et al.* Deep Small RNA-Seq Reveals microRNAs Expression Profiles in Lactating Mammary Gland of 2 Sheep Breeds

with Different Milk Performance. *Domest. Anim. Endocrinol.* **74** (2021).

254. Wu, C., Wang, C., Zhai, B., Zhao, Y., Zhao, Z., Yuan, Z., *et al.* Study of microRNA Expression Profile in Different Regions of Ram Epididymis. *Reprod. Domest. Anim.* **56,** 1209–1219 (2021).

255. Wang, J., Hao, Z., Hu, J., Liu, X., Li, S., Wang, J., *et al.* Small RNA Deep Sequencing Reveals the Expressions of microRNAs in Ovine Mammary Gland Development at Peak-Lactation and during the Non-Lactating Period. *Genomics* **113,** 637–646 (2021).

256. Benítez, R., Trakooljul, N., Núñez, Y., Isabel, B., Murani, E., De Mercado, E., *et al.* Breed, Diet, and Interaction Effects on Adipose Tissue Transcriptome in Iberian and Duroc Pigs Fed Different Energy Sources. *Genes (Basel)* **10,** E589 (2019).

257. Meder, B., Backes, C., Haas, J., Leidinger, P., Stähler, C., Großmann, T., *et al.* Influence of the Confounding Factors Age and Sex on microRNA Profiles from Peripheral Blood. *Clin. Chem.* **60,** 1200–1208 (2014).

258. De Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., *et al.* An Integrated Expression Atlas of miRNAs and Their Promoters in Human and Mouse. *Nat. Biotechnol.* **35,** 872–878 (2017).

259. Guo, L., Zhao, Y., Zhang, H., Yang, S. & Chen, F. Integrated Evolutionary Analysis of Human miRNA Gene Clusters and Families Implicates Evolutionary Relationships. *Gene* **534,** 24–32 (2014).

260. Peng, Y. & Croce, C. M. The Role of MicroRNAs in Human Cancer. *Signal Transduct Target Ther* **1,** 15004 (2016).

261. Bandiera, S., Pfeffer, S., Baumert, T. F. & Zeisel, M. B. MiR-122 - A Key Factor and Therapeutic Target in Liver Disease. *J. Hepatol.* **62,** 448–457 (2015).

262. Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., *et al.* Distribution of miRNA Expression across Human Tissues. *Nucleic Acids Res.* **44,** 3865–3877 (2016).

263. Wang, C., Li, F., Deng, L., Li, M., Wei, M., Zeng, B., *et al.* Identification and Characterization of miRNA Expression Profiles across Five Tissues in Giant Panda. *Gene* **769,** 145206 (2021).

264. Isakova, A., Fehlmann, T., Keller, A. & Quake, S. R. A Mouse Tissue Atlas of Small Noncoding RNA. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 25634–25645 (2020).

265. Smith, A., Calley, J., Mathur, S., Qian, H.-R., Wu, H., Farmen, M., *et al.* The Rat microRNA Body Atlas; Evaluation of the microRNA Content of Rat Organs through Deep Sequencing and Characterization of Pancreas Enriched miRNAs as Biomarkers of Pancreatic Toxicity in the Rat and Dog. *BMC Genomics* **17,** 694 (2016).

266. Koenig, E. M., Fisher, C., Bernard, H., Wolenski, F. S., Gerrein, J., Carsillo, M., *et al.* The Beagle Dog MicroRNA Tissue Atlas: Identifying Translatable Biomarkers of Organ Toxicity. *BMC Genomics* **17,** 649 (2016).

267. Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., *et al.* A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* **129,** 1401–1414 (2007).

268. Chen, W. & Qin, C. General Hallmarks of microRNAs in Brain Evolution and Development. *RNA Biol.* **12,** 701–708 (2015).

269. Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., *et al.* Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Rep.* **3,** 2179–2190 (2013).

270. Yang, C., Yao, C., Tian, R., Zhu, Z., Zhao, L., Li, P., *et al.* miR-202-3p Regulates Sertoli Cell Proliferation, Synthesis Function, and Apoptosis by Targeting LRP6 and Cyclin D1 of Wnt/β-Catenin Signaling. *Mol. Ther. - Nucleic Acids* **14,** 1–19 (2019).

271. Wu, J., Bao, J., Kim, M., Yuan, S., Tang, C., Zheng, H., *et al.* Two miRNA Clusters, miR-34b/c and miR-449, Are Essential for Normal Brain Development, Motile Ciliogenesis, and Spermatogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **111** (2014).

272. Lynn, F. C. Meta-Regulation: microRNA Regulation of Glucose and Lipid Metabolism. *Trends Endocrinol. Metab.* **20,** 452–459 (2009).

273. Horak, M., Novak, J. & Bienertova-Vasku, J. Muscle-Specific microRNAs in Skeletal Muscle Development. *Dev. Biol.* **410,** 1–13 (2016).

274. Penso-Dolfin, L., Moxon, S., Haerty, W. & Di Palma, F. The Evolutionary Dynamics of microRNAs in Domestic Mammals. *Sci Rep* **8,** 17050 (2018).

275. Bao, H., Kommadath, A., Sun, X., Meng, Y., Arantes, A. S., Plastow, G. S., *et al.* Expansion of Ruminant-Specific microRNAs Shapes Target Gene Expression Divergence between Ruminant and Non-Ruminant Species. *BMC Genomics* **14,** 609 (2013).

276. Laganà, A., Veneziano, D., Spata, T., Tang, R., Zhu, H., Mohler, P. J., *et al.* Identification of General and Heart-Specific miRNAs in Sheep (Ovis Aries). *PLoS One* **10,** e0143313 (2015).

277. Bell, J., Larson, M., Kutzler, M., Bionaz, M., Löhr, C. V. & Hendrix, D. miRWoods: Enhanced Precursor Detection and Stacked Random Forests for the Sensitive Detection of microRNAs. *PLoS Comput Biol* **15,** e1007309 (2019).

278. Lawless, N., Vegh, P., O'Farrelly, C. & Lynn, D. J. The Role of microRNAs in Bovine Infection and Immunity. *Front Immunol* **5,** 611 (2014).

279. Mantegazza, A. R., Guttentag, S. H., El-Benna, J., Sasai, M., Iwasaki, A., Shen, H., *et al.* Adaptor Protein-3 in Dendritic Cells Facilitates Phagosomal Toll-like Receptor Signaling and Antigen Presentation to CD4(+) T Cells. *Immunity* **36,** 782–794 (2012).

280. Mantegazza, A. R., Wynosky-Dolfi, M. A., Casson, C. N., Lefkovith, A. J., Shin, S., Brodsky, I. E., *et al.* Increased Autophagic Sequestration in Adaptor Protein-3 Deficient Dendritic Cells Limits Inflammasome Activity and Impairs Antibacterial Immunity. *PLoS Pathog* **13,** e1006785 (2017).

281. Petnicki-Ocwieja, T., Kern, A., Killpack, T. L., Bunnell, S. C. & Hu, L. T. Adaptor Protein-3-Mediated Trafficking of TLR2 Ligands Controls Specificity of Inflammatory Responses but Not Adaptor Complex Assembly. *J Immunol* **195,** 4331–4340 (2015).

282. Sirkis, D. W., Edwards, R. H. & Asensio, C. S. Widespread Dysregulation of Peptide Hormone Release in Mice Lacking Adaptor Protein AP-3. *PLoS Genet* **9,** e1003812 (2013).

283. VanRenterghem, B., Morin, M., Czech, M. P. & Heller-Harrison, R. A. Interaction of Insulin Receptor Substrate-1 with the sigma3A Subunit of the Adaptor Protein Complex-3 in Cultured Adipocytes. *J Biol Chem* **273,** 29942–29949 (1998).

284. Zhou, J.-B., Yang, J.-K., Zhao, L. & Xin, Z. Variants in KCNQ1, AP3S1, MAN2A1, and ALDH7A1 and the Risk of Type 2 Diabetes in the Chinese Northern Han Population: A Case-Control Study and Meta-Analysis. *Med Sci Monit* **16,** BR179–183 (2010).

285. Clauss, M., Hume, I. D. & Hummel, J. Evolutionary Adaptations of Ruminants and Their Potential Relevance for Modern Production Systems. *Animal* **4,** 979–992 (2010).

286. Sasaki, S.-i. Mechanism of Insulin Action on Glucose Metabolism in Ruminants. *Anim. Sci. J.* **73,** 423–433 (2002).

287. McKee, A. S., Munks, M. W., MacLeod, M. K. L., Fleenor, C. J., Van Rooijen, N., Kappler, J. W., *et al.* Alum Induces Innate Immune Responses through Macrophage and Mast Cell Sensors, But These Sensors Are Not Required for Alum to Act As an Adjuvant for Specific Immunity. *J. Immunol.* **183,** 4403–4414 (2009).

288. Petrovsky, N. & Aguilar, J. C. Vaccine Adjuvants: Current State and Future Trends. *Immunol Cell Biol* **82,** 488–496 (2004).

289. Reed, S. G., Orr, M. T. & Fox, C. B. Key Roles of Adjuvants in Modern Vaccines. *Nat. Med.* **19,** 1597–1608 (2013).

290. Petrik, M. S., Wong, M. C., Tabata, R. C., Garry, R. F. & Shaw, C. A. Aluminum Adjuvant Linked to Gulf War Illness Induces Motor Neuron Death in Mice. *Neuromolecular Med* **9,** 83–100 (2007).

291. Shaw, C. A. & Petrik, M. S. Aluminum Hydroxide Injections Lead to Motor Deficits and Motor Neuron Degeneration. *J. Inorg. Biochem.* **103,** 1555–62 (2009).

292. Eidi, H., David, M.-O., Crépeaux, G., Henry, L., Joshi, V., Berger, M.-H., *et al.* Fluorescent Nanodiamonds as a Relevant Tag for the Assessment of Alum Adjuvant Particle Biodisposition. *BMC Med* **13,** 144 (2015).

293. Xu, Y., Zhang, H., Pan, B., Zhang, S., Wang, S. & Niu, Q. Transcriptome-Wide Identification of Differentially Expressed Genes and Long Non-coding RNAs in Aluminum-Treated Rat Hippocampus. *Neurotox. Res.* **34,** 220–232 (2018).

294. Cao, H., Wahlestedt, C. & Kapranov, P. Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. *Trends Genet.* **34,** 704–721 (2018).

295. Qureshi, I. A., Mattick, J. S. & Mehler, M. F. Long Non-Coding RNAs in Nervous System Function and Disease. *Brain Res* **1338,** 20–35 (2010).

296. Zhou, Z., Liu, H., Wang, C., Lu, Q., Huang, Q., Zheng, C., *et al.* Long Non-Coding RNAs as Novel Expression Signatures Modulate DNA Damage and Repair in Cadmium Toxicology. *Sci Rep* **5,** 15293 (2015).

297. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., *et al.* STAR: Ultrafast Universal RNA-seq Aligner. *Bioinforma. Oxf. Engl.* **29,** 15–21 (2013).

298. Archibald, A. L., Cockett, N. E., Dalrymple, B. P., Faraut, T., Kijas, J. W., Maddox, J. F., *et al.* The Sheep Genome Reference Sequence: A Work in Progress. *Anim. Genet.* **41,** 449–453 (2010).

299. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. *Bioinformatics* **30,** 923–930 (2014).

300. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments. *Bioinformatics* **28,** 882–883 (2012).

301. Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., *et al.* PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res.* **13,** 2129–2141 (2003).

302. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* **4,** 44–57 (2009).

303. Langfelder, P. & Horvath, S. WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* **9,** 559 (2008).

304. Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., *et al.* CPC2: A Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features. *Nucleic Acids Res* **45,** W12–W16 (2017).

305. Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P. & Li, W. CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model. *Nucleic Acids Res* **41,** e74 (2013).

306. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).

307. The RNAcentral Consortium. RNAcentral: A Hub of Information for Non-Coding RNA Sequences. *Nucleic Acids Research* **47,** D221–D229 (2019).

308. Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., *et al.* LncADeep: An Ab Initio lncRNA Identification and Functional Annotation Tool Based on Deep Learning. *Bioinformatics* **34** (ed Birol, I.) 3825–3834 (2018).

309. Kallmann, B. A., Hummel, V., Toyka, K. V. & Rieckmann, P. in *Early Indicators Early Treatments Neuroprotection in Multiple Sclerosis* (eds Hommes, O. R. & Comi, G.) 115–117 (Springer Milan, Milano, 2004).

310. McMurray, R. W. Adhesion Molecules in Autoimmune Disease. *Semin Arthritis Rheum* **25,** 215–233 (1996).

311. Zhang, D., Yuan, D., Shen, J., Yan, Y., Gong, C., Gu, J., *et al.* Up-Regulation of VCAM1 Relates to Neuronal Apoptosis After Intracerebral Hemorrhage in Adult Rats. *Neurochem Res* **40,** 1042–1052 (2015).

312. Schattling, B., Steinbach, K., Thies, E., Kruse, M., Menigoz, A., Ufer, F., *et al.* TRPM4 Cation Channel Mediates Axonal and Neuronal Degeneration in Experimental Autoimmune Encephalomyelitis and Multiple Sclerosis. *Nat Med* **18,** 1805–1811 (2012).

313. Li, S., Nie, E. H., Yin, Y., Benowitz, L. I., Tung, S., Vinters, H. V., *et al.* GDF10 Is a Signal for Axonal Sprouting and Functional Recovery after Stroke. *Nat Neurosci* **18,** 1737–1745 (2015).

314. de Miguel, R., Asín, J., Rodríguez-Largo, A., Molín, J., Echeverría, I., de Andrés, D., *et al.* Detection of Aluminum in Lumbar Spinal Cord of Sheep Subcutaneously Inoculated with Aluminum-Hydroxide Containing Products. *J Inorg Biochem* **204,** 110871 (2020).

315. Kumar, V. & Gill, K. D. Oxidative Stress and Mitochondrial Dysfunction in Aluminium Neurotoxicity and Its Amelioration: A Review. *Neurotoxicology* **41,** 154–166 (2014).

316. Iglesias-González, J., Sánchez-Iglesias, S., Beiras-Iglesias, A., Méndez-Álvarez, E. & Soto-Otero, R. Effects of Aluminium on Rat Brain Mitochondria Bioenergetics: An In Vitro and In Vivo Study. *Mol Neurobiol* **54,** 563–570 (2017).

317. Pointer, C. B. & Klegeris, A. Cardiolipin in Central Nervous System Physiology and Pathology. *Cell Mol Neurobiol* **37,** 1161–1172 (2017).

318. Atlante, A., Calissano, P., Bobba, A., Giannattasio, S., Marra, E. & Passarella, S. Glutamate Neurotoxicity, Oxidative Stress and Mitochondria. *FEBS Lett* **497,** 1–5 (2001).

319. Nicholls, D. G. Brain Mitochondrial Calcium Transport: Origins of the Set-Point Concept and Its Application to Physiology and Pathology. *Neurochem Int* **109,** 5–12 (2017).

320. Andersen, R. E. & Lim, D. A. Forging Our Understanding of lncRNAs in the Brain. *Cell Tissue Res* **371,** 55–71 (2018).

321. Lin, N., Chang, K. Y., Li, Z., Gates, K., Rana, Z. A., Dang, J., *et al.* An Evolutionarily Conserved Long Noncoding RNA TUNA Controls Pluripotency and Neural Lineage Commitment. *Mol. Cell* **53,** 1005–1019 (2014).

322. Pek, J. W. Stable Intronic Sequence RNAs Engage in Feedback Loops. *Trends Genet* **34,** 330–332 (2018).

323. You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., *et al.* Neural Circular RNAs Are Derived from Synaptic Genes and Regulated by Development and Plasticity. *Nat Neurosci* **18,** 603–610 (2015).

324. Shi, C., Zhang, L. & Qin, C. Long Non-Coding RNAs in Brain Development, Synaptic Biology, and Alzheimer's Disease. *Brain Res. Bull.* **132,** 160–169 (2017).

325. Wei, C.-W., Luo, T., Zou, S.-S. & Wu, A.-S. The Role of Long Noncoding RNAs in Central Nervous System and Neurodegenerative Diseases. *Front. Behav. Neurosci.* **12,** 175 (2018).

326. Wang, A., Wang, J., Liu, Y. & Zhou, Y. Mechanisms of Long Non-Coding RNAs in the Assembly and Plasticity of Neural Circuitry. *Front. Neural Circuits* **11,** 76 (2017).

327. Bronicki, L. M. & Jasmin, B. J. Emerging Complexity of the HuD/ELAVl4 Gene; Implications for Neuronal Development, Function, and Dysfunction. *Rna* **19,** 1019–1037 (2013).

328. Gardiner, A. S., Twiss, J. L. & Perrone-Bizzozero, N. I. Competing Interactions of RNA-Binding Proteins, MicroRNAs, and Their Targets Control Neuronal Development and Function. *Biomolecules* **5,** 2903–2918 (2015).

329. Carelli, S., Giallongo, T., Rey, F., Latorre, E., Bordoni, M., Mazzucchelli, S., *et al.* HuR Interacts with lincBRN1a and lincBRN1b during Neuronal Stem Cells Differentiation. *RNA Biol* **16,** 1471–1485 (2019).

330. Shu, Y., Zhang, H., Kang, T., Zhang, J.-j., Yang, Y., Liu, H., *et al.* PI3K/Akt Signal Pathway Involved in the Cognitive Impairment Caused by Chronic Cerebral Hypoperfusion in Rats. *PLoS One* **8,** e81901 (2013).

331. Sánchez-Alegría, K., Flores-León, M., Avila-Muñoz, E., Rodríguez-Corona, N. & Arias, C. PI3K Signaling in Neurons: A Central Node for the Control of Multiple Functions. *Int J Mol Sci* **19,** E3725 (2018).

332. Kerrisk, M. E., Cingolani, L. A. & Koleske, A. J. ECM Receptors in Neuronal Structure, Synaptic Plasticity, and Behavior. *Prog Brain Res* **214,** 101–131 (2014).

333. Koo, J. W., Russo, S. J., Ferguson, D., Nestler, E. J. & Duman, R. S. Nuclear Factor-kappaB Is a Critical Mediator of Stress-Impaired Neurogenesis and Depressive Behavior. *Proc Natl Acad Sci U S A* **107,** 2669–2674 (2010).

334. Shih, R.-H., Wang, C.-Y. & Yang, C.-M. NF-kappaB Signaling Pathways in Neurological Inflammation: A Mini Review. *Front Mol Neurosci* **8,** 77 (2015).

335. Glenny, A. T., Pope, C. G., Waddington, H. & Wallace, U. Immunological Notes. XVII-XXIV. *J. Pathol. Bacteriol.* **29,** 31–40 (1926).

336. Ghimire, T. R. The Mechanisms of Action of Vaccines Containing Aluminum Adjuvants: An in Vitro vs in Vivo Paradigm. *SpringerPlus* **4,** 181 (2015).

337. Pellegrino, P., Clementi, E. & Radice, S. On Vaccine's Adjuvants and Autoimmunity: Current Evidence and Future Perspectives. *Autoimmun. Rev.* **14,** 880–888 (2015).

338. Kooijman, S., Brummelman, J., van Els, C. A., Marino, F., Heck, A. J., Mommen, G. P., *et al.* Novel Identified Aluminum Hydroxide-Induced Pathways Prove Monocyte Activation and pro-Inflammatory Preparedness. *J. Proteomics* **175,** 144–155 (2018).

339. Raeven, R. H., van Riet, E., Meiring, H. D., Metz, B. & Kersten, G. F. Systems Vaccinology and Big Data in the Vaccine Development Chain. *Immunology* **156,** 33–46 (2019).

340. de Lima, D. S., Cardozo, L. E., Maracaja-Coutinho, V., Suhrbier, A., Mane, K., Jeffries, D., *et al.* Long Noncoding RNAs Are Involved in Multiple Immunological Pathways in Response to Vaccination. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 17121–17126 (2019).

341. Manjunath, S., Kumar, G. R., Mishra, B. P., Mishra, B., Sahoo, A. P., Joshi, C. G., *et al.* Genomic Analysis of Host - Peste Des Petits Ruminants Vaccine Viral Transcriptome Uncovers Transcription Factors Modulating Immune Regulatory Pathways. *Vet Res* **46,** 15 (2015).

342. Jouneau, L., Lefebvre, D. J., Costa, F., Romey, A., Blaise-Boisseau, S., Relmy, A., *et al.* The Antibody Response Induced FMDV Vaccines in Sheep Correlates with Early Transcriptomic Responses in Blood. *NPJ Vaccines* **5,** 1 (2020).

343. Santoro, F., Pettini, E., Kazmin, D., Ciabattini, A., Fiorino, F., Gilfillan, G. D., *et al.* Transcriptomics of the Vaccine Immune Response: Priming With Adjuvant Modulates Recall Innate Responses After Boosting. *Front Immunol* **9,** 1248 (2018).

344. Harandi, A. M. Systems Analysis of Human Vaccine Adjuvants. *Semin Immunol* **39,** 30–34 (2018).

345. Du, J., Chen, X., Ye, Y. & Sun, H. A Comparative Study on the Mechanisms of Innate Immune Responses in Mice Induced by Alum and Actinidia Eriantha Polysaccharide. *Int. J. Biol. Macromol.* **156,** 1202–1216 (2020).

346. Ransohoff, J. D., Wei, Y. & Khavari, P. A. The Functions and Unique Features of Long Intergenic Non-Coding RNA. *Nat. Rev. Mol. Cell Biol.* **19,** 143–157 (2018).

347. Agirre, X., Meydan, C., Jiang, Y., Garate, L., Doane, A. S., Li, Z., *et al.* Long Non-Coding RNAs Discriminate the Stages and Gene Regulatory States of Human Humoral Immune Response. *Nat Commun* **10,** 821 (2019).

348. Ranzani, V., Rossetti, G., Panzeri, I., Arrigoni, A., Bonnal, R. J., Curti, S., *et al.* The Long Intergenic Noncoding RNA Landscape of Human Lymphocytes Highlights the Regulation of T Cell Differentiation by Linc-MAF-4. *Nat. Immunol.* **16,** 318–325 (2015).

349. Hu, G., Tang, Q., Sharma, S., Yu, F., Escobar, T. M., Muljo, S. A., *et al.* Expression and Regulation of Intergenic Long Noncoding RNAs during T Cell Development and Differentiation. *Nat. Immunol.* **14,** 1190–1198 (2013).

350. Ma, L., Cao, J., Liu, L., Du, Q., Li, Z., Zou, D., *et al.* Lncbook: A Curated Knowledgebase of Human Long Non-Coding Rnas. *Nucleic Acids Res.* **47,** D128–D134 (2019).

351. Touzot, M., Dahirel, A., Cappuccio, A., Segura, E., Hupé, P. & Soumelis, V. Using Transcriptional Signatures to Assess Immune Cell Function: From Basic Mechanisms to Immune-Related Disease. *J. Mol. Biol.* **427,** 3356–3367 (2015).

352. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9,** ISCB Comm J–304 (2020).

353. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., *et al.* Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. *Nucleic Acids Res* **43,** e47 (2015).

354. Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., *et al.* Pathway Enrichment Analysis and Visualization of Omics Data Using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14,** 482–517 (2019).

355. Takata, M., Pachera, E., Frank-Bertoncelj, M., Kozlova, A., Jüngel, A., Whitfield, M. L., *et al.* OTUD6B-AS1 Might Be a Novel Regulator of Apoptosis in Systemic Sclerosis. *Front Immunol* **10,** 1100 (2019).

356. Schmiedel, B. J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A. G., White, B. M., Zapardiel-Gonzalo, J., *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175,** 1701–1715.e16 (2018).

357. Bush, S. J., McCulloch, M. E. B., Lisowski, Z. M., Muriuki, C., Clark, E. L., Young, R., *et al.* Species-Specificity of Transcriptional Regulation and the Response to Lipopolysaccharide in Mammalian Macrophages. *Front Cell Dev Biol* **8,** 661 (2020).

358. Hermanns, H. M. Oncostatin M and Interleukin-31: Cytokines, Receptors, Signal Transduction and Physiology. *Cytokine Growth Factor Rev.* **26,** 545–558 (2015).

359. Johnsson, P., Lipovich, L., Grandér, D. & Morris, K. V. Evolutionary Conservation of Long Non-Coding RNAs; Sequence, Structure, Function. *Biochim. Biophys. Acta - Gen. Subj.* **1840,** 1063–1071 (2014).

360. Kaikkonen, M. U. & Adelman, K. Emerging Roles of Non-Coding RNA Transcription. *Trends Biochem. Sci.* **43,** 654–667 (2018).

361. Morf, J., Basu, S. & Amaral, P. P. RNA, Genome Output and Input. *Front Genet* **11,** 589413 (2020).

362. Gil, N. & Ulitsky, I. Production of Spliced Long Noncoding RNAs Specifies Regions with Increased Enhancer Activity. *Cell Syst.* **7,** 537–547.e3 (2018).

363. Rom, A., Melamed, L., Gil, N., Goldrich, M. J., Kadir, R., Golan, M., *et al.* Regulation of CHD2 Expression by the Chaserr Long Noncoding RNA Gene Is Essential for Viability. *Nat Commun* **10,** 5092 (2019).

364. Wu, K., Zhao, Z., Liu, K., Zhang, J., Li, G. & Wang, L. Long Noncoding RNA Lnc-Sox5 Modulates CRC Tumorigenesis by Unbalancing Tumor Microenvironment. *Cell Cycle* **16,** 1295–1301 (2017).

365. Jiang, S. Recent Findings Regarding Let-7 in Immunity. *Cancer Lett.* **434,** 130–131 (2018).

366. Nejad, C., Stunden, H. J. & Gantier, M. P. A Guide to miRNAs in Inflammation and Innate Immune Responses. *FEBS J.* **285,** 3695–3716 (2018).

367. Huang, X. L., Zhang, L., Li, J. P., Wang, Y. J., Duan, Y. & Wang, J. MicroRNA-150: A Potential Regulator in Pathogens Infection and Autoimmune Diseases. *Autoimmunity* **48,** 503–510 (2015).

368. de Candia, P., Torri, A., Pagani, M. & Abrignani, S. Serum microRNAs as Biomarkers of Human Lymphocyte Activation in Health and Disease. *Front Immunol* **5,** 43 (2014).

369. Amin, V., Harris, R. A., Onuchic, V., Jackson, A. R., Charnecki, T., Paithankar, S., *et al.* Epigenomic Footprints across 111 Reference Epigenomes Reveal Tissue-Specific Epigenetic Regulation of lincRNAs. *Nat Commun* **6,** 6370 (2015).

370. Bilbao-Arribas, M., Varela-Martínez, E., Abendaño, N., de Andrés, D., Luján, L. & Jugo, B. M. Identification of Sheep lncRNAs Related to the Immune Response to Vaccines and Aluminium Adjuvants. *BMC Genomics* **22,** 770 (2021).

371. Sparks, R., Lau, W. W. & Tsang, J. S. Expanding the Immunology Toolbox: Embracing Public-Data Reuse and Crowdsourcing. *Immunity* **45,** 1191–1204 (2016).

372. Toro-Domínguez, D., Villatoro-Garciá, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M. E. & Carmona-Saéz, P. A Survey of Gene Expression Meta-Analysis: Methods and Applications. *Brief. Bioinform.* **22,** 1694–1705 (2021).

373. Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P. & Khatri, P. Methods to Increase Reproducibility in Differential Gene Expression via Meta-Analysis. *Nucleic Acids Res* **45,** e1 (2017).

374. Andres-Terre, M., McGuire, H. M., Pouliot, Y., Bongen, E., Sweeney, T. E., Tato, C. M., *et al.* Integrated, Multi-cohort Analysis Identifies Conserved Transcriptional Signatures across Multiple Respiratory Viruses. *Immunity* **43,** 1199–1211 (2015).

375. Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., *et al.* Molecular Signatures of Antibody Responses Derived from a Systems Biology Study of Five Human Vaccines. *Nat. Immunol.* **15,** 195–204 (2014).

376. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-Optimal Probabilistic RNA-seq Quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

377. Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., *et al.* Sustainable Data Analysis with Snakemake. *F1000Res* **10,** 33 (2021).

378. Martin, M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.journal* **17,** 10 (2011).

379. Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L. & Pertea, M. Transcriptome Assembly from Long-Read RNA-seq Alignments with StringTie2. *Genome Biol* **20,** 278 (2019).

380. Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., *et al.* FEELnc: A Tool for Long Non-Coding RNA Annotation and Its Application to the Dog Transcriptome. *Nucleic Acids Res* **45,** e57 (2017).

381. Davenport, K. M., Bickhart, D. M., Worley, K., Murali, S. C., Salavati, M., Clark, E. L., *et al.* An Improved Ovine Reference Genome Assembly to Facilitate In-Depth Functional Annotation of the Sheep Genome. *Gigascience* **11,** giab096 (2022).

382. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: Enabling Browsing of Large Distributed Datasets. *Bioinformatics* **26,** 2204–2207 (2010).

383. Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R. & Sandelin, A. CAGEfightR: Analysis of 5′-End Data Using R/Bioconductor. *BMC Bioinformatics* **20,** 487 (2019).

384. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: A Flexible Python Library for Manipulating Genomic Datasets and Annotations. *Bioinformatics* **27,** 3423–3424 (2011).

385. Massa, A. T., Mousel, M. R., Herndon, M. K., Herndon, D. R., Murdoch, B. M. & White, S. N. Genome-Wide Histone Modifications and CTCF Enrichment Predict Gene Expression in Sheep Macrophages. *Front Genet* **11,** 612031 (2020).

386. Langmead, B. & Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

387. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., *et al.* Twelve Years of SAMtools and BCFtools. *Gigascience* **10,** giab008 (2021).

388. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast Processing of NGS Alignment Formats. *Bioinformatics* **31,** 2032–2034 (2015).

389. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., *et al.* Model-Based Analysis of ChIP-Seq (MACS). *Genome Biol* **9,** R137 (2008).

390. Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Grüning, B., *et al.* pyGenomeTracks: Reproducible Plots for Multivariate Genomic Datasets. *Bioinformatics* **37,** 422–423 (2021).

391. Vitting-Seerup, K., Sandelin, A. & Berger, B. IsoformSwitchAnalyzeR: Analysis of Changes in Genome-Wide Patterns of Alternative Splicing and Its Functional Consequences. *Bioinformatics* **35,** 4469–4471 (2019).

392. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-Tailed Prior Distributions for Sequence Count Data: Removing the Noise and Preserving Large Differences. *Bioinformatics* **35,** 2084–2092 (2019).

393. Lemoine, G. G., Scott-Boyer, M.-P., Ambroise, B., Périn, O. & Droit, A. GWENA: Gene Co-Expression Networks Analysis and Extended Modules Characterization in a Single Bioconductor Package. *BMC Bioinformatics* **22,** 267 (2021).

394. Johnson, K. A. & Krishnan, A. Robust Normalization and Transformation Techniques for Constructing Gene Coexpression Networks from RNA-seq Data. *Genome Biol* **23,** 1 (2022).

395. Bhuva, D. D., Cursons, J., Smyth, G. K. & Davis, M. J. Differential Co-Expression-Based Detection of Conditional Relationships in Transcriptional Data: Comparative Analysis and Application to Breast Cancer. *Genome Biol* **20,** 236 (2019).

396. Braun, R. O., Brunner, L., Wyler, K., Auray, G., García-Nicolás, O., Python, S., *et al.* System Immunology-Based Identification of Blood Transcriptional Modules Correlating to Antibody Responses in Sheep. *NPJ Vaccines* **3,** 41 (2018).

397. Fu, Y., Chryssafidis, A. L., Browne, J. A., O'Sullivan, J., McGettigan, P. A. & Mulcahy, G. Transcriptomic Study on Ovine Immune Responses to Fasciola Hepatica Infection. *PLoS Negl Trop Dis* **10,** e0005015 (2016).

398. Guo, X., Zhang, J., Li, Y., Yang, J., Li, Y., Dong, C., *et al.* Evaluating the Effect of TLR4-overexpressing on the Transcriptome Profile in Ovine Peripheral Blood Mononuclear Cells. *J Biol Res (Thessalon)* **27,** 13 (2020).

399. Wang, S., Hu, D., Wang, C., Tang, X., Du, M., Gu, X., *et al.* Transcriptional Profiling of Innate Immune Responses in Sheep PBMCs Induced by Haemonchus Contortus Soluble Extracts. *Parasit Vectors* **12,** 182 (2019).

400. Niedziela, D. A., Naranjo-Lucena, A., Molina-Hernández, V., Browne, J. A., Martínez-Moreno, Á., Pérez, J., *et al.* Timing of Transcriptomic Peripheral Blood Mononuclear Cell Responses of Sheep to Fasciola Hepatica Infection Differs From Those of Cattle, Reflecting Different Disease Phenotypes. *Front Immunol* **12,** 729217 (2021).

401. Iannaccone, M., Ianni, A., Contaldi, F., Esposito, S., Martino, C., Bennato, F., *et al.* Whole Blood Transcriptome Analysis in Ewes Fed with Hemp Seed Supplemented Diet. *Sci Rep* **9,** 16192 (2019).

402. Gossner, A., Watkins, C., Chianini, F. & Hopkins, J. Pathways and Genes Associated with Immune Dysfunction in Sheep Paratuberculosis. *Sci Rep* **7,** 46695 (2017).

403. Pan, X., Cai, Y., Li, Z., Chen, X., Heller, R., Wang, N., *et al.* Modes of Genetic Adaptations Underlying Functional Innovations in the Rumen. *Sci China Life Sci* **64,** 1–21 (2021).

404. McRae, K. M., Good, B., Hanrahan, J. P., McCabe, M. S., Cormican, P., Sweeney, T., *et al.* Transcriptional Profiling of the Ovine Abomasal Lymph Node Reveals a Role for Timing of the Immune Response in Gastrointestinal Nematode Resistance. *Vet. Parasitol.* **224,** 96–108 (2016).

405. Chitneedi, P. K., Suárez-Vega, A., Martínez-Valladares, M., Arranz, J. J. & Gutiérrez-Gil, B. Exploring the Mechanisms of Resistance to Teladorsagia Circumcincta Infection in Sheep through Transcriptome Analysis of Abomasal Mucosa and Abomasal Lymph Nodes. *Vet Res* **49,** 39 (2018).

406. Tang, Q., Gu, Y., Zhou, X., Jin, L., Guan, J., Liu, R., *et al.* Comparative Transcriptomics of 5 High-Altitude Vertebrates and Their Low-Altitude Relatives. *Gigascience* **6,** 1–9 (2017).

407. Naranjo-Lucena, A., Correia, C. N., Molina-Hernández, V., Martínez-Moreno, Á., Browne, J. A., Pérez, J., *et al.* Transcriptomic Analysis of Ovine Hepatic Lymph Node Following Fasciola Hepatica Infection - Inhibition of NK Cell and IgE-Mediated Signaling. *Front Immunol* **12,** 687579 (2021).

408. Johnsson, P., Ziegenhain, C., Hartmanis, L., Hendriks, G. J., Hagemann-Jensen, M., Reinius, B., *et al.* Transcriptional Kinetics and Molecular Functions of Long Noncoding RNAs. *Nat. Genet.* **54,** 306–317 (2022).

409. de Goede, O. M., Nachun, D. C., Ferraro, N. M., Gloudemans, M. J., Rao, A. S., Smail, C., *et al.* Population-Scale Tissue Transcriptomics Maps Long Non-Coding RNAs to Complex Disease. *Cell* **184,** 2633–2648.e19 (2021).

410. Walters, K., Sarsenov, R., Too, W. S., Hare, R. K., Paterson, I. C., Lambert, D. W., *et al.* Comprehensive Functional Profiling of Long Non-Coding RNAs through a Novel Pan-Cancer Integration Approach and Modular Analysis of Their Protein-Coding Gene Association Networks. *BMC Genomics* **20,** 454 (2019).

411. Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-Stimulated Genes: A Complex Web of Host Defenses. *Annu. Rev. Immunol.* **32,** 513–545 (2014).

412. Qiu, L., Wang, T., Tang, Q., Li, G., Wu, P. & Chen, K. Long Non-coding RNAs: Regulators of Viral Infection and the Interferon Antiviral Response. *Front Microbiol* **9,** 1621 (2018).

413. Meng, X.-Y., Luo, Y., Anwar, M. N., Sun, Y., Gao, Y., Zhang, H., *et al.* Long Non-Coding RNAs: Emerging and Versatile Regulators in Host-Virus Interactions. *Front Immunol* **8,** 1663 (2017).

414. Kambara, H., Niazi, F., Kostadinova, L., Moonka, D. K., Siegel, C. T., Post, A. B., *et al.* Negative Regulation of the Interferon Response by an Interferon-Induced Long Non-Coding RNA. *Nucleic Acids Res.* **42,** 10668–10681 (2014).

415. Ouyang, J., Hu, J. & Chen, J. L. lncRNAs Regulate the Innate Immune Response to Viral Infection. *Wiley Interdiscip. Rev. RNA* **7,** 129–143 (2016).

416. O'Neill, S. M., Brady, M. T., Callanan, J. J., Mulcahy, G., Joyce, P., Mills, K. H., *et al.* Fasciola Hepatica Infection Downregulates Th1 Responses in Mice. *Parasite Immunol.* **22,** 147–155 (2000).

417. Karrow, N. A., Goliboski, K., Stonos, N., Schenkel, F. & Peregrine, A. Review: Genetics of Helminth Resistance in Sheep. *Can. J. Anim. Sci.* **94,** 1–9 (2014).

418. Venturina, V. M., Gossner, A. G. & Hopkins, J. The Immunology and Genetics of Resistance of Sheep to Teladorsagia Circumcincta. *Vet. Res. Commun.* **37,** 171–181 (2013).

419. Zhou, G., Stevenson, M. M., Geary, T. G. & Xia, J. Comprehensive Transcriptome Meta-analysis to Characterize Host Immune Responses in Helminth Infections. *PLoS Negl Trop Dis* **10,** e0004624 (2016).

420. Todd, D. J., Lee, A. H. & Glimcher, L. H. The Endoplasmic Reticulum Stress Response in Immunity and Autoimmunity. *Nat. Rev. Immunol.* **8,** 663–674 (2008).

421. Hetz, C. & Papa, F. R. The Unfolded Protein Response and Cell Fate Control. *Mol. Cell* **69,** 169–181 (2018).

422. Pramanik, J., Chen, X., Kar, G., Henriksson, J., Gomes, T., Park, J. E., *et al.* Genome-Wide Analyses Reveal the IRE1a-XBP1 Pathway Promotes T Helper Cell Differentiation by Resolving Secretory Stress and Accelerating Proliferation. *Genome Med.* **10,** 76 (2018).

423. Sampieri, L., Di Giusto, P. & Alvarez, C. CREB3 Transcription Factors: ER-Golgi Stress Transducers as Hubs for Cellular Homeostasis. *Front Cell Dev Biol* **7,** 123 (2019).

424. Zhao, T., Du, J. & Zeng, H. Interplay between Endoplasmic Reticulum Stress and Non-Coding RNAs in Cancer. *J Hematol Oncol* **13,** 163 (2020).

425. Zhai, L., Ladomersky, E., Lenzen, A., Nguyen, B., Patel, R., Lauing, K. L., *et al.* IDO1 in Cancer: A Gemini of Immune Checkpoints. *Cell. Mol. Immunol.* **15,** 447–457 (2018).

426. Szcześniak, M. W., Wanowska, E., Mukherjee, N., Ohler, U. & Makałowska, I. Towards a Deeper Annotation of Human lncRNAs. *Biochim Biophys Acta Gene Regul Mech* **1863,** 194385 (2020).

427. Lorenzi, L., Chiu, H. S., Avila Cobos, F., Gross, S., Volders, P. J., Cannoodt, R., *et al.* The RNA Atlas Expands the Catalog of Human Non-Coding RNAs. *Nat. Biotechnol.* **39,** 1453–1465 (2021).

428. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & von Schack, D. Evaluation of Two Main RNA-seq Approaches for Gene Quantification in Clinical RNA Sequencing: polyA+ Selection versus rRNA Depletion. *Sci Rep* **8,** 4781 (2018).

429. Dahlgren, A. R., Scott, E. Y., Mansour, T., Hales, E. N., Ross, P. J., Kalbfleisch, T. S., *et al.* Comparison of Poly-A+ Selection and rRNA Depletion in Detection of lncRNA in Two Equine Tissues Using RNA-seq. *Non-Coding RNA* **6,** 32 (2020).

430. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs Are an Important Regulatory Layer of the Epigenome. *Nat. Struct. Mol. Biol.* **27,** 521–528 (2020).

431. Barturen, G., Rueda, A., Hamberg, M., Alganza, A., Lebron, R., Kotsyfakis, M., *et al.* sRNAbench: Profiling of Small RNAs and Its Sequence Variants in Single or Multi-Species High-Throughput Experiments. *Methods Gener. Seq.* **1,** 21–31 (2014).

432. Chen, L., Heikkinen, L., Wang, C., Yang, Y., Sun, H. & Wong, G. Trends in the Development of miRNA Bioinformatics Tools. *Brief. Bioinform.* **20,** 1836–1852 (2019).

433. Babarinde, I. A., Li, Y. & Hutchins, A. P. Computational Methods for Mapping, Assembly and Quantification for Coding and Noncoding Transcripts. *Comput. Struct. Biotechnol. J.* **17,** 628–637 (2019).

434. Agostini, F., Zagalak, J., Attig, J., Ule, J. & Luscombe, N. M. Intergenic RNA Mainly Derives from Nascent Transcripts of Known Genes. *Genome Biol* **22,** 136 (2021).

435. Hovhannisyan, H. & Gabaldón, T. The Long Non-Coding RNA Landscape of Candida Yeast Pathogens. *Nat Commun* **12,** 7317 (2021).

436. Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., *et al.* Assessment of Transcript Reconstruction Methods for RNA-seq. *Nat. Methods* **10,** 1177–1184 (2013).

437. Sun, Z., Nair, A., Chen, X., Prodduturi, N., Wang, J. & Kocher, J.-P. UClncR: Ultrafast and Comprehensive Long Non-Coding RNA Detection from RNA-seq. *Sci Rep* **7,** 14196 (2017).

438. Bryzghalov, O., Makałowska, I. & Szcześniak, M. W. lncEvo: Automated Identification and Conservation Study of Long Noncoding RNAs. *BMC Bioinformatics* **22,** 59 (2021).

439. Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., *et al.* Ensembl 2022. *Nucleic Acids Res* **50,** D988–D995 (2022).

440. Medzhitov, R. & Horng, T. Transcriptional Control of the Inflammatory Response. *Nat. Rev. Immunol.* **9,** 692–703 (2009).

441. Conley, J. M., Gallagher, M. P. & Berg, L. J. T Cells and Gene Regulation: The Switching On and Turning Up of Genes after T Cell Receptor Stimulation in CD8 T Cells. *Front Immunol* **7,** 76 (2016).

442. Laidlaw, B. J. & Cyster, J. G. Transcriptional Regulation of Memory B Cell Differentiation. *Nat. Rev. Immunol.* **21,** 209–220 (2021).

443. Mehta, A. & Baltimore, D. MicroRNAs as Regulatory Elements in Immune System Logic. *Nat. Rev. Immunol.* **16,** 279–294 (2016).

444. Atianand, M. K., Caffrey, D. R. & Fitzgerald, K. A. Immunobiology of Long Noncoding RNAs. *Annu. Rev. Immunol.* **35,** 177–198 (2017).

445. Bernier, A. & Sagan, S. M. The Diverse Roles of microRNAs at the Host¯Virus Interface. *Viruses* **10,** E440 (2018).

446. Lai, Y.-C., Fujikawa, T., Maemura, T., Ando, T., Kitahara, G., Endo, Y., *et al.* Inflammation-Related microRNA Expression Level in the Bovine Milk Is Affected by Mastitis. *PLoS One* **12,** e0177182 (2017).

447. Jin, W., Ibeagha-Awemu, E. M., Liang, G., Beaudoin, F., Zhao, X. & Guan, L. L. Transcriptome microRNA Profiling of Bovine Mammary Epithelial Cells Challenged with Escherichia Coli or Staphylococcus Aureus Bacteria Reveals Pathogen Directed microRNA Expression Profiles. *BMC Genomics* **15,** 181 (2014).

448. Brogaard, L., Larsen, L. E., Heegaard, P. M. H., Anthon, C., Gorodkin, J., Dürrwald, R., *et al.* IFN-λ and microRNAs Are Important Modulators of the Pulmonary Innate Immune Response against Influenza A (H1N2)

Infection in Pigs. *PLoS One* **13,** e0194765 (2018).

449. Wang, Y., Brahmakshatriya, V., Zhu, H., Lupiani, B., Reddy, S. M., Yoon, B.-J., *et al.* Identification of Differentially Expressed miRNAs in Chicken Lung and Trachea with Avian Influenza Virus Infection by a Deep Sequencing Approach. *BMC Genomics* **10,** 512 (2009).

450. Khanduri, A., Sahu, A. R., Wani, S. A., Khan, R. I. N., Pandey, A., Saxena, S., *et al.* Dysregulated miRNAome and Proteome of PPRV Infected Goat PBMCs Reveal a Coordinated Immune Response. *Front Immunol* **9,** 2631 (2018).

451. Jenike, A. E. & Halushka, M. K. miR-21: A Non-Specific Biomarker of All Maladies. *Biomark Res* **9,** 18 (2021).

452. Roux, B. T., Heward, J. A., Donnelly, L. E., Jones, S. W. & Lindsay, M. A. Catalog of Differentially Expressed Long Non-Coding RNA Following Activation of Human and Mouse Innate Immune Response. *Front Immunol* **8,** 1038 (2017).

453. Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., *et al.* A Large Intergenic Noncoding RNA Induced by P53 Mediates Global Gene Repression in the P53 Response. *Cell* **142,** 409–419 (2010).

454. Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., *et al.* Birth and Expression Evolution of Mammalian microRNA Genes. *Genome Res.* **23,** 34–45 (2013).

455. Camilleri-Robles, C., Amador, R., Klein, C. C., Guigó, R., Corominas, M. & Ruiz-Romero, M. Genomic and Functional Conservation of lncRNAs: Lessons from Flies. *Mamm Genome* **33,** 328–342 (2022).

456. Wu, X. & Sharp, P. A. Divergent Transcription: A Driving Force for New Gene Origination? *Cell* **155,** 990–996 (2013).

457. Jin, Y., Eser, U., Struhl, K. & Churchman, L. S. The Ground State and Evolution of Promoter Region Directionality. *Cell* **170,** 889–898.e10 (2017).

458. Kapusta, A. & Feschotte, C. Volatile Evolution of Long Noncoding RNA Repertoires: Mechanisms and Biological Implications. *Trends in Genetics* **30,** 439–452 (2014).

459. Kesner, J. S., Chen, Z., Aparicio, A. A. & Wu, X. *A Unified Model for the Surveillance of Translation in Diverse Noncoding Sequences* 2022.

460. Bilbao-Arribas, M. & Jugo, B. *Data from: Comprehensive Analysis of Ovine Transcriptomic Data Reveals Novel Long Non-Coding RNAs Related to the Immune Response* 2022.

Data and code availability

The datasets generated and analysed in **chapter 3** are available in the NCBI's Gene Expression Omnibus (GEO) repository and are accessible through Series accession number GSE115415. Supplementary files are available in the online version of the published article [233].

The datasets supporting the conclusions of **chapter 4** have been made publicly available by the original authors and are available in the NCBI SRA repository under the following BioProject accessions: PRJNA451237, PRJNA354833, PRJEB22101, PRJEB32852, PRJEB32852, PRJNA392421, PRJEB20781, PRJNA505702, PRJNA474913, PRJNA532808, PRJNA511987, PRJNA414087, PRJNA454385, PRJNA528259, PRJNA638028, PRJNA613135, PRJNA608075, PRJNA694531 and PRJNA607580. Sequences, coordinates and expression of all miRNAs are available as Supplementary tables (these files are available upon request and will be published together with the journal version). The code used in this chapter is available at https://github.com/bilbaom/sheep-miRNAome.

The data discussed in **chapter 5** have been deposited in NCBI's GEO and are accessible through Series accession number GSE128597. Supplementary material is available in the online version of the published article [234].

The datasets analysed in **chapter 6** are available in NCBI's GEO repository with Series accession number GSE113899. Custom python scripts used for the analysis of lncRNAs are available at (https://github.com/bilbaom/vaccine-lncrnas-sheep). Supplementary files are available in the online version of the published article [370].

All data generated or analysed in **chapter 7** are included as supplementary files and in public repositories. The RNA-seq datasets analysed in this work were obtained from the NCBI SRA repository under the following project accessions: PRJEB26387, PRJNA454435, PRJNA559411, PRJNA291172, PRJNA433706, PRJNA268183, PRJEB33476, PRJEB45790, PRJEB44063, PRJEB15872, PRJNA631066, PRJNA528905, PRJNA485657, PRJNA362606 and PRJEB19199. FAANG CAGE-seq and CHIP-seq datasets were obtained from European Nucleotide Archive (ENA) project accessions PRJEB34864 and PRJEB40528. Following the Fort Lauderdale Agreement, all used datasets have been previously published and are cited in the main text. Supplementary data files, annotations and expression quantification of novel lncRNAs can be found at https://doi.org/10.5281/zenodo.6802782 [460]. The code used in this chapter is available at https://github.com/bilbaom/immune-lncrnas-sheep.

Acknowledgements - Eskerrak

Asko eskertzen diot Begoña M. Jugo-ri lan hau aurrera eramateko aukera eman izana, eta etengabeko laguntza eta orientazioa eman izana zientziaren munduan barneratu nauen bide honetan. Gure ikerketa taldean kide izan ditudan Naiara Abendaño, Aitor Guisasola eta Endika Varela ere eskertu nahi nituzke proiektuak burutzeko izan ditugun hartu-eman guztiengatik eta tesi honetan izan duten parte-hartzeagatik, bereziki Endika, estatistikarekin laguntzeko prest beti egon delako. Genetikako laborategiko beste kide guztiak ezin ditut ahaztu, gauza desberdinetan lan egiten badugu ere, edozein lekutan lan egiteko hain garrantzitsua den giro ona sortu izan dutelako.

I am also thankful to Dr. Daniel Fischer and Professor Johanna Vilkki for having me as a guest during a short research stay at the Natural Resources Institute Finland (LUKE).

Azkenik, eskerrak eman nahi dizkiet urte hauetan nirekin izan diren lagunei eta familiari.

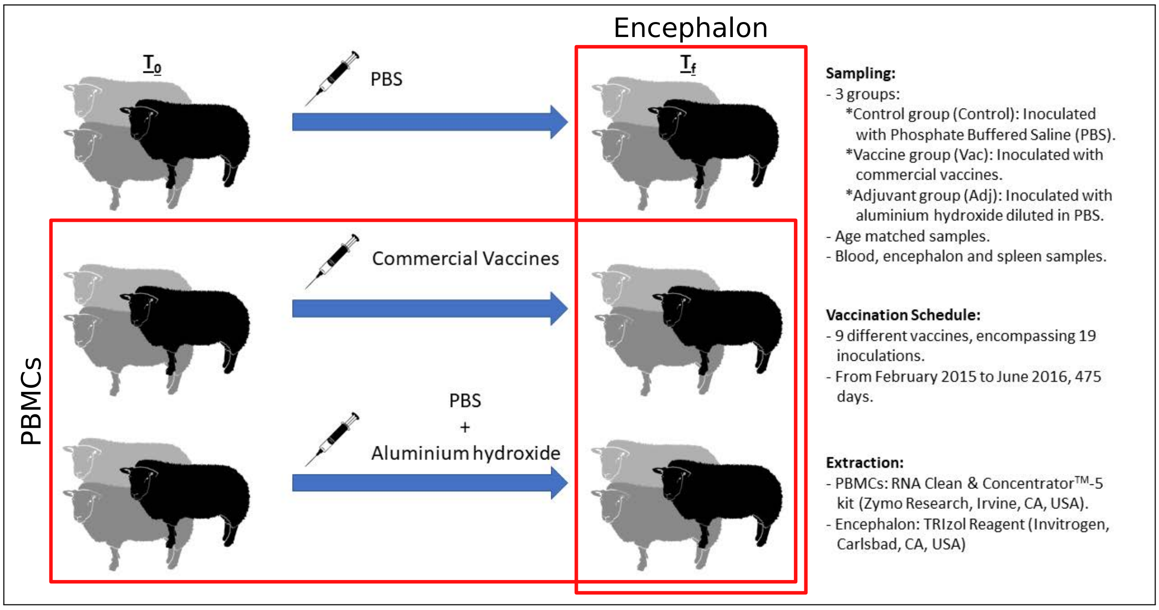
This thesis would not have been possible without the PhD scholarship (PIF17/306) granted by the University of the Basque Country (UPV/EHU). This work received funding from the Spanish Ministry of Economy and Industry project number AGL2013-49137-C3 and UPV/EHU research group grants (PPGA18/11, GIU20/071). The short time scientific mission (STSM) at LUKE (Finland) was funded by the FAANG-Europe European Union COST action (CA15112).

Appendix

Supplementary information for chapter 5 and chapter 6

Table S1: Commercial vaccines used on sheep in the experiment of this thesis.

Vaccine number	**Commercial name**	**Manufacturer**	**Antigen/s**	**Inoculation day**	**Al per dose (mg)**
1	Heptavac P Plus	MSD Animal Health S.L.	Pasteurella multocida, Mannheimia haemolytica, Clostridium spp.	0, 23, 233	7.5
2	Autogenous vaccine	Exopol	Staphylococcus aureus spp. Anaerobius	44, 69, 349	1.64
3	Vanguard R	Zoetis	Rabies virus	98	1.03
4	Agalaxipra	Hipra	Mycoplasma agalactiae	129, 146	6.76
5	Ovivac CS	Hipra	Chlamydophila abortus, Salmonella abortus ovis	209, 233	5.60
6	Autogenous vaccine	Exopol	Corynebacterium pseudotuberculosis	254, 272	1.32
7	Bluevac-1	CZ Veterinaria S.A.	Bluetongue virus serotype 1	293, 329	4.18
8	Bluevac-4	CZ Veterinaria S.A.	Bluetongue virus serotype 4	293, 329	4.16
9	Bluevac BTV 8	CZ Veterinaria S.A.	Bluetongue virus serotype 8	449, 470	4.40



Sampling:
- 3 groups:
  *Control group (Control): Inoculated with Phosphate Buffered Saline (PBS).
  *Vaccine group (Vac): Inoculated with commercial vaccines.
  *Adjuvant group (Adj): Inoculated with aluminium hydroxide diluted in PBS.
- Age matched samples.
- Blood, encephalon and spleen samples.

Vaccination Schedule:
- 9 different vaccines, encompassing 19 inoculations.
- From February 2015 to June 2016, 475 days.

Extraction:
- PBMCs: RNA Clean & Concentrator^{TM}-5 kit (Zymo Research, Irvine, CA, USA).
- Encephalon: TRIzol Reagent (Invitrogen, Carlsbad, CA, USA)

Figure S1: Experimental design of the long-term vaccination study.

Supplementary information for chapter 7

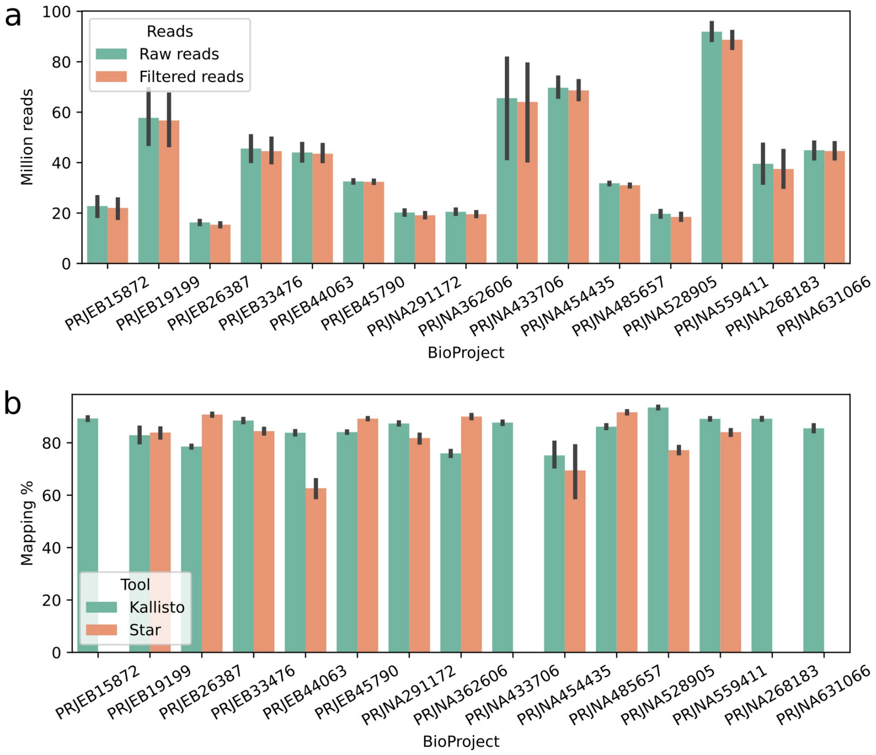
Supplementary information

Comprehensive analysis of ovine transcriptomic data reveals novel long non-coding RNAs related to the immune response

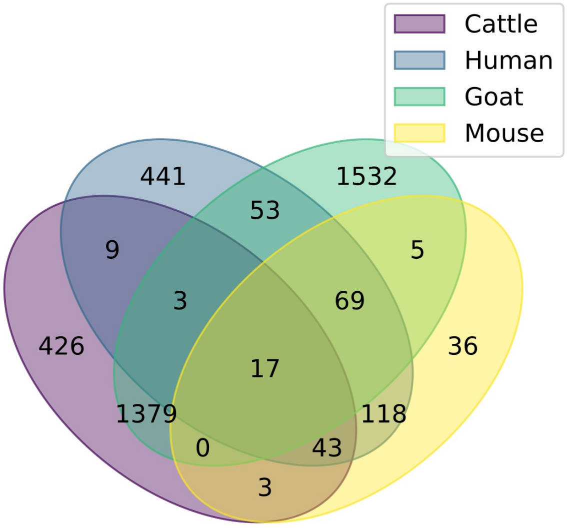
Martin Bilbao-Arribas*, Begoña M. Jugo

Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940, Leioa, Spain

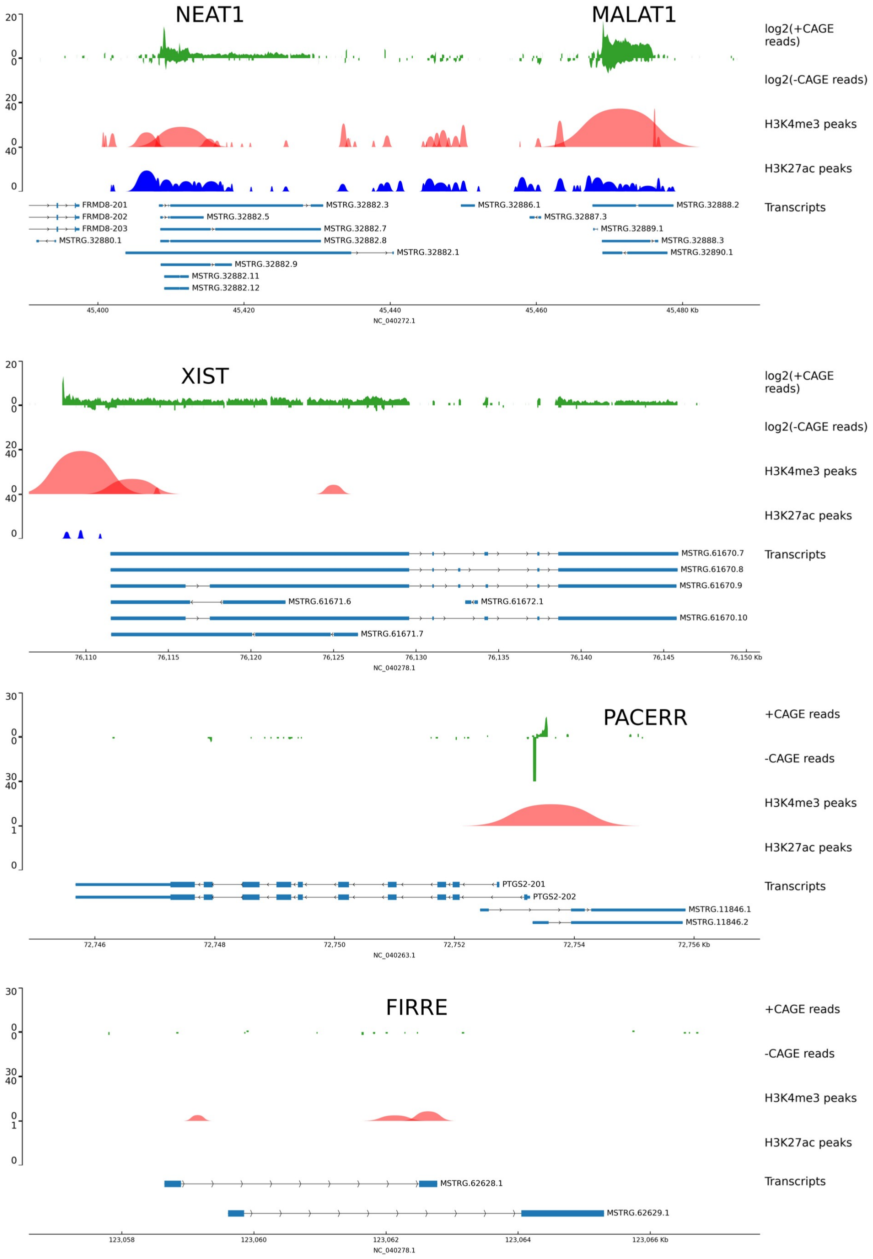
*Corresponding author



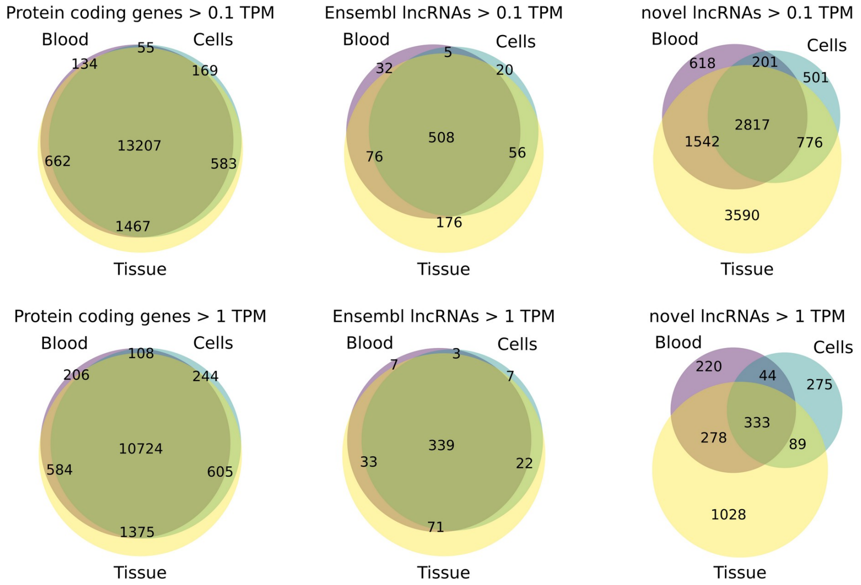
Supplementary Fig. 1: Summary statistics of the samples included in the study. (A) Average number of reads in each dataset before and after quality filtering and read adapter removal. (B) Average mapping rate to the genome (STAR) and average pseudo-alignment rate (Kallisto) to the full unfiltered transcriptome assembled with Stringtie. Unstranded samples were not used in the genome mapping for lncRNA identification.



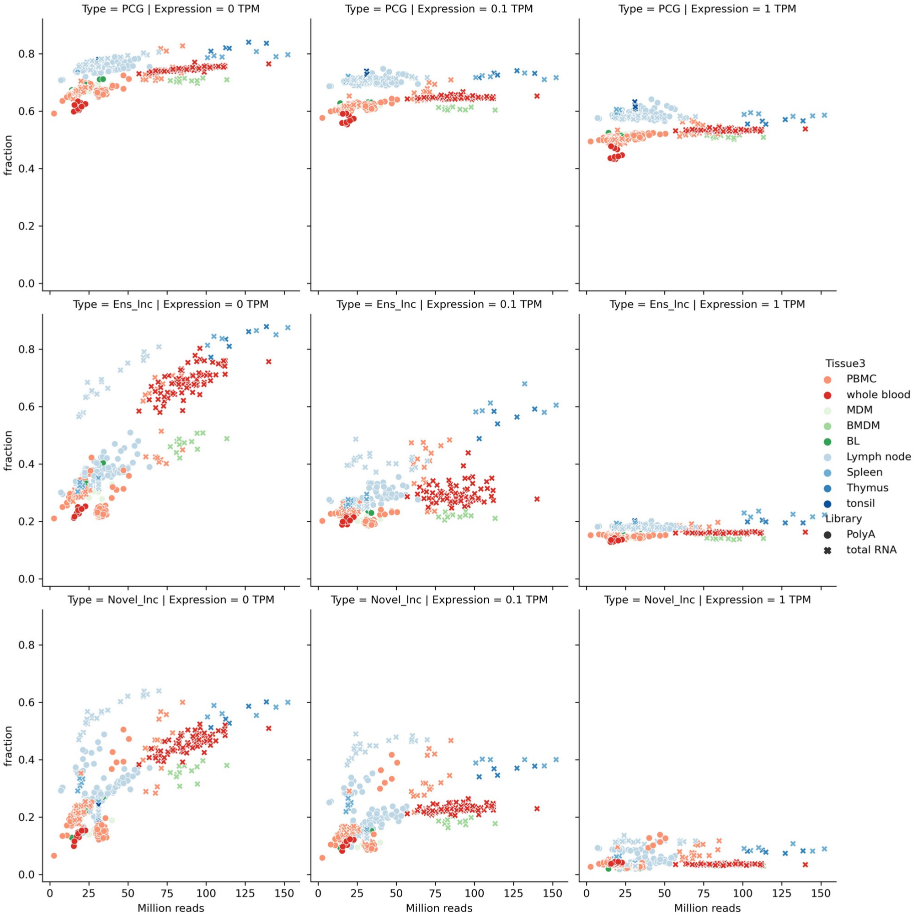
Supplementary Fig. 2: Summary of sequence conservation analysis. Number of sheep lncRNA transcripts with significant sequence similarity with annotated lncRNAs in other mammal species.



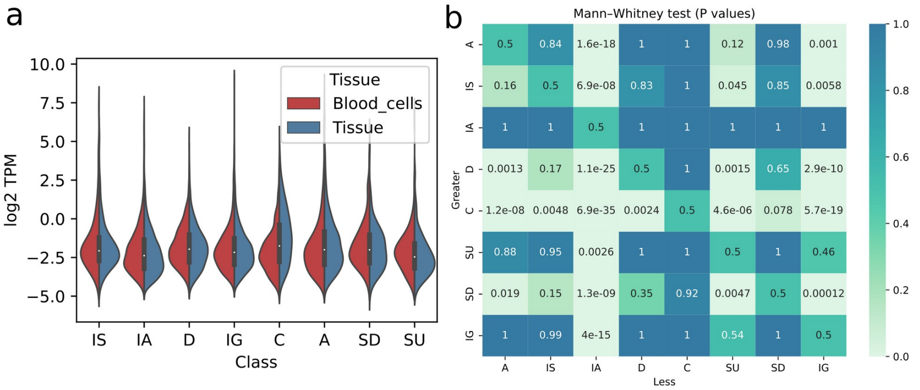
Supplementary Fig. 3: Examples of conserved lncRNAs. Genomic context of selected conserved lncRNAs between sheep and human is depicted, including CAGE-seq read mapping, predicted CHIP-seq peaks and transcripts models.



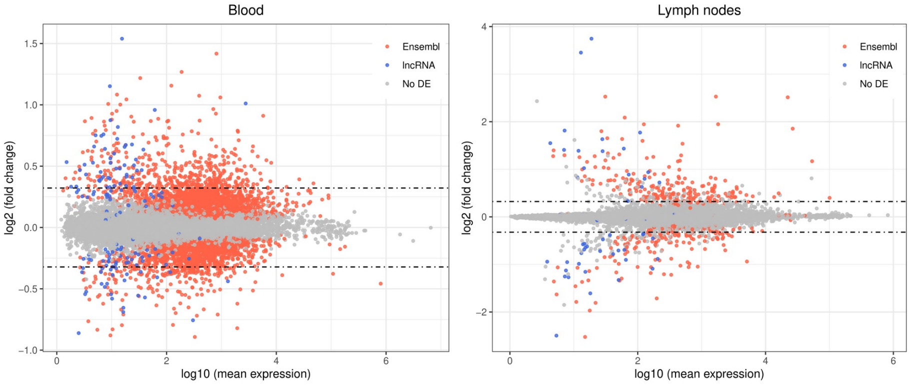
Supplementary Fig. 4: Expression Venn diagrams. Venn diagrams comparing the expression of PCGs, annotated lncRNAs and novel lncRNAs in each tissue group.



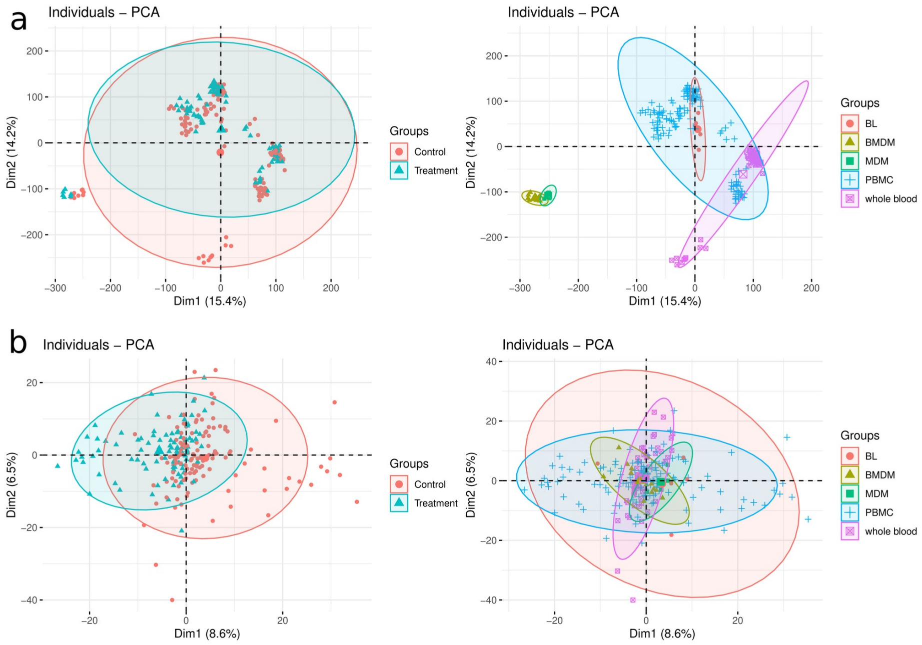
Supplementary Fig. 5: Saturation of gene detection. Number of expressed PCGs and lncRNAs in each sample as a fraction of all genes annotated from each type compared to sequencing depth of the samples.



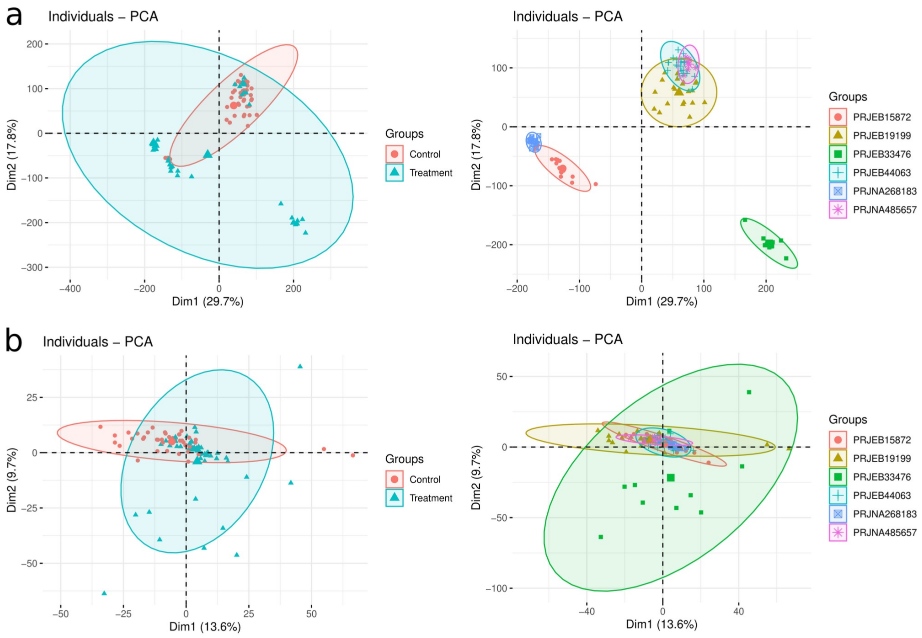
Supplementary Fig. 6: Expression of lncRNA classes. (A) Expression of each class of unannotated lncRNAs in blood and cell samples and in lymphoid tissues. LncRNAs expressed above 0.1 TPM in at least one fifth of the samples in each tissue group were used. (B) One-sided Mann-Whitney U tests between the expression levels of each class of unannotated lncRNAs. LncRNAs expressed above 0.1 TPM in at least one tenth of the samples were used.



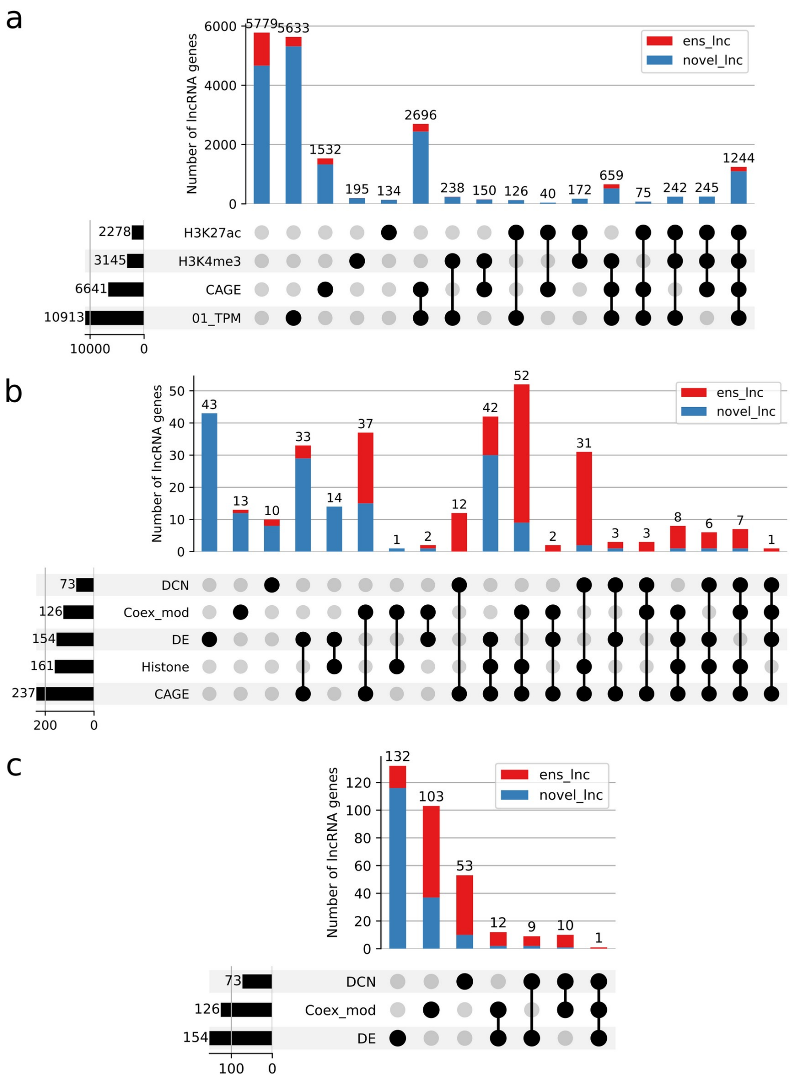
Supplementary Fig. 7: MA plots of the differential-expression results in blood cells and lymph nodes.



Supplementary Fig. 8: PCA plots of the blood cell samples used for coexpression analysis, coloured by treatment group and tissue. (A) Samples before any correction. (B) Samples after correction for covariates.



Supplementary Fig. 9: PCA plots of the lymph node samples used for coexpression analysis, coloured by treatment group and project accession. (A) Samples before any correction. (B) Samples after correction for covariates.



Supplementary Fig. 10: Upset plots of the integration of lncRNA features and the functional analyses. (A) Intersections of support from CHIP-seq histone modifications and CAGE-seq peaks in all annotated and novel lncRNAs. (B) Intersections of statistically significant genes from differential expression analysis (DE), immune-enriched modules (Coex_mod) and differential co-expression analysis (DCN) with CAGE peaks and histone modifications. Annotated and novel lncRNAs statistically significant for at least one functional analysis are depicted. (C) Intersections of statistically significant genes from differential expression analysis (DE), immune-enriched modules (Coex_mod) and differential co-expression analysis (DCN) in annotated and novel lncRNAs.