eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

GLOBAL CHARACTERIZATION OF THE IMMUNE RESPONSE TO INOCULATION OF ALUMINIUM HYDROXIDE-BASED VACCINES BY RNA SEQUENCING

Endika Varela Martínez

Supervisor: B.M. Jugo

Euskal Herriko Unibertsitatea / Universidad del País Vasco

2020

Acknowledgments

I would like to thank my supervisor Begoña M. Jugo for her constant support and guidance through my PhD project. I am thankful to the postdoctoral researcher Naiara Abendaño and the fellow PhD student Martín Bilbao for promoting and maintaining a cordial working environment and for their contributions. This work would have not been possible without the collaboration of the research group directed by Dr. L LLuján in the Department of Animal Pathology in the University of Zaragoza and the research team directed by Dr. D de Andrés in the Institute of Agrobiotechnology (IdAB) in Mutilva, Navarra.

Despite being a sort stay, I would like to express my appreciation to Dr. Jan Gorodkin and other members of the Center for non-coding RNA in Technology and Health (RTH) from the University of Copenhagen, Denmark. Working with that team has been a great privilege and I have learned a great deal about circular RNAs (circRNAs) and their annotation, apart to be an opportunity to learn about Denmark and more specifically about Copenhagen.

Last but not least, I dedicate this thesis to my family. Most of all, to my parents and brother who have supported me through these years and specially to my grandmother for her continuous and unparalleled love and encouragement.

Funding

This thesis is the result of my PhD project carried out from November 2015 to _____ 2020 at the Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country (UPV/EHU). This project was supported by a MINECO project grant (AGL2013-49137-C3-3-R) and by UPV/EHU projects (GIU14/23 and PPGA18/11). E Varela-Martínez was a recipient of a UPV/EHU predoctoral grant (PIF15/361), a grant for mobility of research staff in stays of 30 to 150 days from the UPV/EHU and a Short-Term Scientific Mission (STSM) from COST (COST Action CA15112, Functional Annotation of Animal Genomes – European network), which is an EU-funded programme.

Abstract

Introduction

Vaccines that require the addition of substances named adjuvants for enhancement of the immune response are mainly those composed of killed organisms or highly purified antigens. By themselves, antigens are not able to induce a strong and long-lasting immune response. Thus, adjuvants that increase the speed of the response to the antigen, that reduce the quantity of antigen that need to be exposed to for a long-lasting immune response or that bias the immune response towards specific cells of the immune system (e.g., bias the response towards a T_H1 or a T_H2 immune response) are added to the vaccines. Aluminium salts, especially aluminium hydroxide and aluminium phosphate, are among the most widely used adjuvants in human and veterinary vaccines.

Despite aluminium adjuvants have been in use for a long time, with the first aluminium salt used as adjuvant in the 1930's, the mechanism of action by which they elicit an immune reaction is not fully understood. Multiple mechanisms of action have been proposed and it seems that rather than having a dominant one, aluminium acts simultaneously through multiple pathways. In addition to its partially understood mechanism of action, some concerns regarding its safety have recently been raised. It is well known that Al is a non-essential element for the human body and is thought that it lacks any essential biological function. The fact that the body is not able to excrete all injected Al in a short-term period of time through normal mechanisms such as urine, points towards a high persistence of the material in the body. Moreover, the fact that Al may be able to reach distant organs though translocation of the material by phagocytic immune cells from the monocytic cell line (with translocation of the material to lymph nodes demonstrated) have raised some concerns, since the material may be translocated to such a distant tissue as brain after accumulation of the material from multiple exposures of aluminium-based vaccines in predisposed individuals. Thus, in this work a long-term experiment in which sheep receive multiple aluminium-based vaccines was developed to study the fate of injected aluminium and to asses if the aluminium of multiple vaccinations accumulated in the body due to its high persistence.

Few works have analysed the immune response to aluminium adjuvants in a large mammal such as sheep through high throughput technology (RNA sequencing). Most studies on aluminium adjuvants have been done in mice models, which may fail to recreate some aspects of the mechanism of action. Larger animals like sheep share similarities to human regarding physiology, anatomy, metabolism, genetics and size, making them a good alternative. First and foremost, the route for vaccine administration is the same in human and farm animals, through the subcutaneous or intramuscular route. Secondly, large farm animals such as sheep have aluminium-based vaccines designed specifically for them.

Objectives

The general aim of this work is the study of genes and regulatory elements involved in the immune response induced by aluminium hydroxide after repetitive inoculation of commercial vaccines composed of said element. For that purpose, RNA sequencing (RNA-seq) libraries from multiple tissues of sheep were prepared and analysed with the protocols described in this thesis. Expression levels for mRNAs and miRNAs were quantified for each animal in peripheral blood mononuclear cells (PBMCs) and parietal lobe cortex. In addition, circular RNAs (circRNAs) were characterized in those sheep samples. The study of the mentioned elements and their

interactions may help to understand better the mechanism of action of aluminium adjuvants and to discern if the aluminium is able to reach a distant organ such as brain.

Material and methods

Samples in this study were obtained from the Rasa Aragonesa sheep breed, a breed from the Northeast of Spain raised mostly for meat production. Three months old purebred lambs were selected from a single pedigree flock and after a short period of two months to acclimatize to the new environment, the sheep were randomly distributed in three different treatment groups: one group receiving multiple inoculations of commercial vaccines composed of aluminium hydroxide, named Vac group; another group receiving equivalent doses of the adjuvant (Alhydrogel) diluted in phosphate-buffered saline, named Adj group; and a group only receiving phosphate-buffered saline, termed Control group. The complete experiment lasted 475 days, from February 2015 to June 2016. During that period of time, nine different vaccines were administered to each animal, which comprises a total of 19 inoculations throughout 16 different inoculation dates. All commercial vaccines used in this study are common vaccines given to sheep during their productive period in Spain.

Comparisons of multiple tissues from these animals would allow to discern key elements of Al mechanism of action, in addition to check if Al accumulates in a distant tissue such as brain. In this work, sequencing results from PBMCs and parietal lobe cortex will be presented. This being a coordinated project, the histopathological analyses and behavioural changes of the animals were studied by other research groups under the same project. RNA-seq was applied by our research group to characterize molecular changes in the transcriptome in the previously commented tissues. Therefore, libraries for total RNA-seq, which undergo ribosomal RNA depletion, retaining any non-coding RNA, and miRNA-seq were prepared. It was taken into account that the sheep reference genome is still in progress, with non-coding element such as lncRNAs and circRNAs being poorly annotated. Thus, paired-end libraries were prepared and were sequenced with a high enough depth for novel element characterization. In this work, the differential expression results from the previously commented tissues and novel circRNA characterization will be presented.

Results and discussion

The expression of the *NLRP3* inflammasome has been related to Al adjuvant activity and it has been reported that *IL1B* activation is dependent of the expression of the inflammasome. In PBMC samples, *NLRP3* had a constant expression when sheep that received commercial vaccines (Vac-injected sheep) were compared to their initial stage, before any vaccination, while it was found downregulated in sheep that received the aluminium hydroxide (AH) adjuvant diluted in phosphate-buffered saline (Adj-injected sheep). Thus, it seems that the inflammasome is not required for Al adjuvant activity in sheep under the conditions of this experiment, which point towards an inflammasome independent activation of the immune response. The contradictory results regarding the requirement of the inflammasome may be explained by the multifaceted activation of the immune response by Al adjuvants. Multiple mechanisms have been reported for Al activated immune response: formation of a depot; creation of a local pro inflammatory environment, which results in recruitment of different immune cells at the injection site; due to tissue damage at the injection site, endogenous danger signals that induce an inflammatory response (uric acid and host DNA in the case of Al) are released from necrotic cells; enhancement of antigen uptake and presentation; and as previously stated, activation of the NLRP3 inflammasome, which leads to production of the pro inflammatory cytokine IL-1β. Instead of having a dominant mechanism for immune stimulation, it seems that multiple pathways are

activated simultaneously, and when one of the pathways is not activated, the others may act compensatory to elicit a strong enough immune response. Differences regarding the role of the inflammasome in adjuvant activity may be attributed to differences in the studies such as the formulation used (different combinations of adjuvant and antigen, which results in differences in agglomerate size and adsorption rate), immunization protocols, animal models and route of administration.

The consequent increase in inflammatory signals led to the activation of the NF-κB signaling pathway in both Vac- and Adj-inoculated sheep when compared against their initial state before any vaccination. There were multiple genes from the NF-κB family, such as *NFKB2*, *RELA* and *RELB*, which were highly expressed in both groups simultaneously. The main differences in both groups were in the expression of genes from the cytokine-cytokine receptor interaction pathway, which were clearly downregulated in Adj-injected animals. This was consistent with the induction of an ongoing immune response against the vaccine, but suggesting a milder induction of the immune response in Adj-inoculated animals. Although the adjuvant is not antigenic per se, it seems that is able to produce a non-specific induction of proinflammatory responses when the adjuvant is injected alone without any pathogen. Something that would be concordant with multiple reports showing immune stimulation without any adsorbed antigen.

Regarding the miRNA differential expression analysis in PBMCs, there were some miRNAs that had been previously related to Al adjuvants. miR-125b, which was upregulated in Vac-inoculated sheep, has been shown to be a reactive oxygen species (ROS)- and NF-κB up-regulated miRNA by Al. Despite not being directly linked to Al, miR-99a, which was also upregulated in Vac injected sheep, has been shown to be promoted by NF-κB. There is a broad activation of the NF-κB pathway and it seems that said pathway is highly regulated by multiple miRNA expression in the immune response to Al adjuvants. When miRNA-mRNA co-expression was checked, multiple factors related to cellular response to DNA damage stimulus, RNA binding and response to stimulus were found to be negatively correlated. Among them, there were some negatively correlated miRNA-mRNA pairs related to the NF-κB pathway, pointing as previously mentioned to a highly regulated expression of said pathway by miRNAs in Al elicited immune response. *MAP3K2* (*MEKK2*), which is a predicted target of *let-7b* (upregulated in Adj-injected sheep), is a kinase that controls the persistent activation of *NF-κB* in response to stimulation with proinflammatory cytokines through the formation of the MAP3K2:IκB-β:NF-κB:IKK complex. *SNX27*, which is a predicted target of *miR-125b* (upregulated in Vac-injected sheep), has been shown to cause NF-κB hyperactivation after its silencing. Apart to pairs related to NF-κB, it was found a pair related to DNA damage response. *CHEK1*, predicted to be targeted by *miR-16b* (miRNA upregulated in Adj-injected animals), has been shown to have reduced levels of expression after exposure to aluminium chloride or aluminium chlorohydrate.

In contrasts to the high expression change seen in PBMC samples, , it was shown nearly no differential expression in the parietal lobe cortex samples of animals vaccinated with commercial vaccines and the quantity of Al detected in parietal lobe samples from the Vac-inoculated sheep was similar to those of the control group, which indicate that commercial formulations are pretty safe under the conditions of this experiment. Completely different was the case of Adj-inoculated animals, in which a tendency to higher Al content was detected when compared to control samples. It must be pointed that most of the Al accumulation measurements made in the parietal lobe were below 1 μg/g, a level considered safe. With nearly 5 times more DEGs in Adj-injected sheep than the Vac-inoculated animals, among the differentially expressed genes there were terms usually found dysregulated in neurological diseases, namely: *VCAM1*, *TRPM4*, *GDF10* and *NTN1*. Taken together, it seems that under the

terms of this experiment Al was able to reach the cortex and induce molecular changes when is free from any antigen. Thus, it raises some concerns on the safety of a large number of vaccine trials, which uses Al adjuvant-containing placebo groups.

Regarding the miRNA differential expression analysis, a pattern similar to the total RNA-seq differential expression analysis was seen, with nearly no significant change in Vac-inoculated animals. When miRNA-mRNA co-expression was checked for differentially expressed miRNAs, multiple factors related to mitochondria function, maintenance of neural polarity and DNA damage were found. Among them, there were some predicted targets that would help to broad our knowledge on Al toxicity in brain, pointing to a dysregulation of mitochondrial functions. *ACTR10*, which is a predicted target of the up-regulated *let-7b* in Adj-inoculated animals, has been shown to disrupt mitochondrial retrograde transport at its absence, leading to accumulation of mitochondria in axon terminals. In addition, *MRS2*, another predicted target of let-7b, is a mitochondrial Mg transporter that has been related to defects in the organelle and apoptosis. Taken together, it seems that Al may be causing an imbalance in metal ion levels, which would be concordant to what have been seen in other species such as rats treated with an intragastric administration of Al gluconate.

As previously stated, paired-end libraries were sequenced with a high enough depth for novel sheep RNA characterization. circRNAs are non-coding RNAs with a cyclic structure. Recently, circRNAs have caught the attention of researchers due their tissue specific expression, conservation across species and their involvement in multiple biological functions, such as neuronal differentiation, neuronal apoptosis and BBB dysfunction in brain and basic immune functions and transcription regulation in blood. Thus, this study attempted to annotate novel circRNAs in sheep and to discern if circRNAs have any role in Al adjuvancy. It must be pointed that there is not database recording sheep circRNAs and the tissues used in this study have not been used previously for circRNA annotation. Thus, this study will broad the current sheep circRNA annotation.

After circRNA characterization by two different tools, a wide expression of circRNAs was found in both tissues. A total of 2,510 and 3,403 circRNAs were detected in parietal lobe cortex and PBMCs, respectively, of which 1,379 were completely novel circRNAs (841 exclusive to PBMC samples, 421 exclusive to encephalon samples and 117 expressed in both tissues). Most of the identified circRNAs originated from annotated genes, and supposing that all exons were retained, they were generally formed by two or three distinct exons, in agreement with what has been previously reported in human and mouse data. In addition, it has been described that some circRNAs have a tissue-dependent or developmental stage-dependent expression pattern. In our samples, 1,236 circRNAs (36.32% of all detected blood circRNAs) were detected in both tissues, which is concordant with approximately 30% of the detected blood circRNAs overlapping with circRNAs expressed in the cerebellum of human and mouse data.

The circRNAs detected in this study were compared to other sheep circRNA identified in pituitary gland and in longissimus dorsi muscle, since as previously pointed there is no database recording them. Only a few circRNA backsplice junctions were detected in all tissues at the same time, while several hundreds of circRNAs were exclusive to each tissue, which shows how some circRNAs have a tissue-dependent expression. Furthermore, since multiple studies have shown that circRNAs have evolutionary conservation between human and mouse, the circRNAs detected in this work were compared to a human circRNA database. Approximately the 63% of circRNAs in both tissues had completely conserved backsplice sites when compared to human backsplice junctions.

Among the functional roles that have been proposed for circRNAs, there is binding activity between circRNAs and RNA binding proteins (RBPs), which suggests that circRNAs can

impact the same functional processes in which the corresponding linear host gene is involved. Under the assumption that the function of a circRNA may be associated with the known function of its parental gene, PBMC circRNAs were related to multiple immune functions such as B- and T-cell proliferation, neutrophil degranulation, the MAPK cascade and the NF-κB signaling, while parietal lobe cortex circRNAs were related to synapse regulation, behaviour, learning process and brain development. Furthermore, multiple circRNAs have been reported to act as miRNA sponges, which are defined as sequences with multiple miRNA binding sites that compete with target genes for miRNA binding. There was only one circRNA exclusively expressed in cortex samples, and hosted by the CDR1 gene, containing multiple binding sites for miR-7 and miR-1224, both reported to be expressed in the mammalian brain.

Regarding the differential expression analysis, it must be pointed that there is no tool designed specifically for circRNA expression data based on rRNA depleted total RNA-seq libraries. Most researchers use tools such as DESeq2 and edgeR, which are based in a negative binomial distribution, but there have not been any study showing if the negative binomial model is adequate for circRNA expression data, which only counts backsplice junction reads. At least in our samples, circRNA expression data had generally very low counts (a few highly expressed circRNAs originated most of the counts) and is zero-inflated. Due to the different structure of circRNA expression data and uncertainty of whether the methods used so far are adequate, multiple methods were applied to our data. Independent of the choice of method, we did not detect any differentially expressed circRNAs in any of the two tissues, which indicates that circRNAs may not be connected with aluminium adjuvancy.

Conclusions

Briefly, a general activation of the immune response was seen in both Vac- and Adj-inoculated animals in PBMCs, with the activation of the NF-κB signaling pathway in both treatment groups. There were similarities in the immune response seen in both treatments, but it was shown that without the presence of any antigen a milder immune response was induced in Adj-inoculated sheep. Interestingly, the expression of the *NLRP3* inflammasome was not required for aluminium adjuvant activity in the animals of this study. Regarding the expression of miRNAs and their interaction with the studied mRNAs, it was shown that multiple miRNAs such as *let-7b* (upregulated in Adj tf vs. Adj t0 comparison), *miR-125b* and *miR-99a* (both upregulated in Vac tf vs. Vac t0 comparison) may be related to the NF-κB pathway, pointing towards a highly regulated expression of the pathway by miRNAs in response to aluminium adjuvants.

In parietal lobe cortex samples, it was shown nearly no differential expression in the animals vaccinated with commercial vaccines, while in Adj-inoculated sheep a few more differentially expressed genes (nearly 5 times more) were found. This was highly concordant with the Al levels detected in the tissue, with no difference of aluminium content between control samples and Vac-inoculated sheep and with a tendency to higher aluminium levels in Adj-inoculated sheep. From the results of this study, it seems that the aluminium from commercial vaccines is not able to reach the parietal lobe cortex after a long-term exposure.

Regarding the circRNA characterization, a wide expression of said elements was found in both tissues. In addition, sheep circRNAs detected in this study were highly conserved when compared to human circRNAs. Independent of the choice of differential expression method, there were no differentially expressed circRNAs in both tissues, which may indicate an aluminium independent expression of circRNAs.

Resumen

Introducción

Las vacunas compuestas por organismos muertos o antígenos altamente purificados requieren la adición de ciertas sustancias para la mejora de la respuesta inmune. Dichas vacunas no son capaces de inducir una respuesta inmune fuerte y duradera por sí solas y necesitan la adición de substancias denominadas adyuvantes. Se entiende como adyuvante cualquier substancia que aumente la velocidad de la respuesta al antígeno, que reduzca la cantidad de antígeno al que se debe exponer para una respuesta inmune duradera o que sesgue la respuesta inmune hacia células específicas del sistema inmune (por ejemplo, hacia una respuesta inmune T_H1 o T_H2). Las sales de aluminio, especialmente el hidróxido de aluminio y el fosfato de aluminio, se encuentran entre los adyuvantes más utilizados en vacunas humanas y veterinarias.

A pesar de que los adyuvantes de aluminio han estado en uso durante un largo periodo de tiempo, siendo la primera sal de aluminio utilizada como adyuvante en la década de 1930, el mecanismo de acción por el cual provocan una reacción inmune no está completamente estudiado. Se han propuesto múltiples mecanismos de acción y en lugar de tener uno dominante, parece ser que el aluminio (Al) actúa simultáneamente a través de múltiples vías. Además de su mecanismo de acción parcialmente estudiado, recientemente se han planteado algunas preocupaciones con respecto a su seguridad. Es bien sabido que el Al es un elemento no esencial para el cuerpo humano y se cree que carece de cualquier función biológica esencial. El hecho de que el cuerpo no pueda excretar todo el Al inyectado en un período corto de tiempo a través de mecanismos como la orina, apunta hacia una alta persistencia del material en el cuerpo. Además, el hecho de que el Al pueda ser capaz de alcanzar órganos distantes a través de la translocación del material mediante células inmunes fagocíticas de la línea celular monocítica (con la translocación del material a los ganglios linfáticos ya demostrado) ha generado algunas preocupaciones, sobre todo en cuanto a la todavía por demostrar capacidad de translocación al cerebro en individuos predispuestos después de exposiciones a múltiples vacunas con base de aluminio. Por lo tanto, en este trabajo se analizó un experimento a largo plazo en el que ovejas reciben múltiples vacunas compuestas de hidróxido de aluminio con el objetivo de estudiar el destino del aluminio inyectado y evaluar si el aluminio se acumula en el cuerpo debido a su más que demostrada alta persistencia.

Pocos trabajos han analizado la respuesta inmune a los adyuvantes de aluminio en un mamífero grande como la oveja a través de tecnologías de alto rendimiento (secuenciación de ARN o ARN-seq). La mayoría de los estudios sobre adyuvantes de aluminio se han centrado en el uso de ratones como modelo animal, que pueden no recrear algunos aspectos del mecanismo de acción. Animales más grandes como las ovejas comparten muchas similitudes con los humanos en cuanto a fisiología, anatomía, metabolismo, genética y tamaño, lo que los convierte en una buena alternativa. En primer lugar, la ruta para la administración de la vacuna es la misma en humanos y animales de granja, siendo esta la ruta subcutánea o intramuscular. En segundo lugar, animales de granja como la oveja tienen vacunas a base de aluminio diseñadas específicamente para ellas después de pasar rigurosos controles.

Objetivos

El objetivo principal de este trabajo es el estudio de genes y elementos reguladores involucrados en la respuesta inmune inducida por el hidróxido de aluminio después de inoculaciones repetitivas de vacunas comerciales. Para ello, se prepararon y analizaron bibliotecas de secuenciación de ARN (ARN-seq) de múltiples tejidos ovinos con los protocolos descritos en esta

tesis. Los niveles de expresión de ARNm y miARN fueron cuantificado para cada animal en células mononucleares de sangre periférica (PBMC) y en la corteza del lóbulo parietal. Además, los ARN circulares (circARN) fueron caracterizados en esas mismas muestras ovinas. El estudio de los elementos mencionados y sus posibles interacciones puede ayudar a comprender el mecanismo de acción del hidróxido de aluminio y a discernir si puede alcanzar un órgano tan distante como el cerebro.

Materiales y métodos

Las muestras de este estudio provinieron de la raza ovina Rasa Aragonesa, una raza del noreste de España criada principalmente para la producción de carne. Se seleccionaron corderos de tres meses de pedigrí de raza pura de un solo lote. Después de un corto período de aclimatación de dos meses, las ovejas se distribuyeron aleatoriamente en tres grupos de tratamiento diferentes: un grupo que recibió múltiples inoculaciones de vacunas comerciales compuestas de hidróxido de aluminio, denominado grupo Vac; otro grupo que recibió dosis equivalentes del adyuvante (Alhydrogel) diluido en tampón fosfato salino, denominado grupo Adj; y un grupo que solo recibió tampón fosfato salino, denominado grupo control. El experimento completo duró 475 días, desde febrero de 2015 hasta junio de 2016. Durante ese período de tiempo, se administraron nueve vacunas distintas a cada animal, resultando en un total de 19 vacunaciones en 16 fechas de inoculación distintas. Todas las vacunas comerciales utilizadas en este estudio son vacunas comunes que usualmente se administran en España a las ovejas durante su período productivo.

Las comparaciones de los múltiples tejidos en estos animales permitirían discernir elementos clave del mecanismo de acción del Al, además de verificar si se acumula en un tejido distante como el cerebro. Al tratarse de un proyecto coordinado, los análisis histopatológicos y los cambios de comportamiento de los animales fueron estudiados por otros grupos de investigación bajo el mismo proyecto. En este trabajo se presentarán los resultados de la secuenciación (ARN-seq) de PBMCs y la corteza del lóbulo parietal. Se prepararon bibliotecas de secuenciación de ARN total, en el que el ARN ribosómico es removido reteniendo cualquier secuencia de ARN no codificante, y de secuenciación de miARN. Se tuvo en cuenta que el genoma de referencia ovino está incompleto, sobre todo en elementos no codificantes como lncARNs y circARNs. Por lo tanto, se prepararon bibliotecas paired-end y se secuenciaron con una profundidad suficientemente alta como para permitir la caracterización de nuevos ARNs no codificantes. En este trabajo se presentarán los resultados de la expresión diferencial de los tejidos previamente comentados y de la caracterización de circARNs.

Resultados y discusión

En otros estudios, la expresión del inflamasoma *NLRP3* se ha relacionado con la actividad adyuvante del Al y se ha descrito que la activación de IL-1β depende de la expresión del inflamasoma. En nuestras muestras de PBMCs, *NLRP3* tenía una expresión constante cuando las ovejas que recibieron vacunas comerciales (grupo Vac) se compararon con su etapa inicial, antes de recibir cualquier vacunación, mientras que se encontró una regulación negativa en las ovejas que recibieron el adyuvante diluido en tampón fosfato salino (grupo Adj). Por lo tanto, parece que bajo las condiciones de este experimento las ovejas no requieren de la expresión del inflamasoma para la actividad adyuvante del Al, apuntando hacia una activación de la respuesta inmune independiente del inflamasoma. Los resultados contradictorios con respecto al requerimiento del inflamasoma pueden explicarse por la activación multifacética de la respuesta inmune del Al. Se han descrito múltiples mecanismos para la activación de la respuesta inmune por el Al: formación de depósitos; creación de un entorno proinflamatorio local, que da como

resultado el reclutamiento de diferentes células inmunes en el sitio de inyección; liberación de señales de peligro endógenas (ácido úrico y ADN del huésped en el caso de Al) de células necróticas que inducen una respuesta inmune, todo debido al daño causado en el tejido en el sitio de inyección; mejora de la captación y presentación de antígeno; y como se indicó anteriormente, la activación del inflamasoma *NLRP3*, que conduce a la producción de la citoquina proinflamatoria IL-1β. En lugar de tener un mecanismo dominante para la estimulación inmune, parece que se activan múltiples vías simultáneamente, y cuando una de las vías no se activa, las otras pueden actuar de manera compensatoria para provocar una respuesta inmune lo suficientemente fuerte. Las diferencias con respecto al papel del inflamasoma en la actividad adyuvante también se pueden atribuir a diferencias en los estudios, como la formulación utilizada (diferentes combinaciones de adyuvante y antígeno, que dan como resultado a diferencias en el tamaño del aglomerado y tasa de adsorción), protocolos de inmunización, modelos animales y ruta de administración.

El aumento de las señales inflamatorias condujo a la activación de la vía de señalización NF-κB en ovejas de los grupos Vac y Adj en comparación con su estado inicial antes de cualquier vacunación. Se hallaron múltiples genes de la familia NF-κB, como NFKB2, RELA y RELB, altamente expresados de manera significativa en ambos grupos. Las principales diferencias en ambos grupos se hallaban en la expresión de genes de la vía de interacción de receptores de citoquinas y citoquinas, que estaban claramente con regulación negativa en animales del grupo Adj. Esto es consistente con la inducción de una respuesta inmune contra la vacuna, pero sugiere una inducción más leve en animales inoculados solo con el adyuvante. Aunque el adyuvante no es antigénico *per se*, parece que es capaz de producir una inducción no específica de respuestas proinflamatorias incluso cuando ningún antígeno se halla presente. Esto concuerda con múltiples trabajos anteriores que muestran estimulación inmune sin ningún antígeno adsorbido.

Respecto al análisis de expresión diferencial de miARNs en PBMCs, se hallaron diferencialmente expresados algunos miARNs que ya se habían relacionado previamente con el Al. Se ha demostrado que *miR-125b*, que estaba sobreexpresado en ovejas del grupo Vac, es un miARN sobreexpresado por especies de oxígeno reactivo (EOR o ROS) y por NF-κB en la respuesta al Al. A pesar de no estar directamente relacionado con Al, *miR-99a*, que también estaba sobreexpresado en ovejas del grupo Vac, ha demostrado ser promovido por NF-κB. Existe una amplia activación de la ruta NF-κB y parece que dicha ruta está altamente regulada por la expresión de miARNs en la respuesta inmune al Al. Cuando se analizó la coexpresión de miARN-ARNm, se descubrieron múltiples factores negativamente correlacionados relacionados con la respuesta celular a estímulos de daño de ADN, unión del ARN y respuesta a estímulos. Entre ellos, había algunos pares de miARN-ARNm negativamente correlacionados relacionados con la ruta NF-κB, apuntando como se ha mencionado anteriormente, a una expresión altamente regulada de dicha ruta por miARNs en la respuesta inmunitaria provocada por el Al. *MAP3K2* (*MEKK2*), que es una diana predicha de *let-7b* (sobreexpresado en el grupo Vac), es una quinasa que controla la activación persistente de NF-κB en respuesta a la estimulación con citoquinas proinflamatorias a través de la formación del complejo MAP3K2:IκB-β:NF-κB:IKK. Por otro lado, se ha visto que *SNX27*, que es una diana predicha de *miR-125b* (sobreexpresado en el grupo Vac), causa hiperactivación de NF-κB. Además de los pares relacionados con NF-κB, se encontró un par relacionado con la respuesta al daño del ADN. Se ha demostrado que *CHEK1*, diana predicha de *miR-16b* (sobreexpresado en el grupo Adj), tiene niveles reducidos de expresión después de la exposición al cloruro de aluminio o clorhidrato de aluminio.

En contraste a los cambios de expresión observados en las muestras de PBMC, casi no se observó expresión diferencial en las muestras de la corteza del lóbulo parietal en animales del grupo Vac y se pudo comprobar que la cantidad de Al detectada en dichas muestras era

similar a la del grupo de control, lo que indica que las vacunas comerciales son bastante seguras en las condiciones de este experimento. Completamente diferente es el caso de los animales del grupo Adj, en los que se detectó una tendencia a un mayor contenido de Al en comparación con las muestras de control. Debe señalarse que la mayoría de las mediciones de acumulación de Al realizadas en el lóbulo parietal estaban por debajo de 1 µg/g, un nivel considerado seguro. Con casi 5 veces más genes diferencialmente expresados, entre los genes diferencialmente expresados en el grupo Adj había términos que generalmente se encuentran desregulados en enfermedades neurológicas, a saber: *VCAM1*, *TRPM4*, *GDF10* y *NTN1*. Parece que, bajo los términos de este experimento, el Al pudo llegar a la corteza e inducir leves cambios moleculares cuando está libre de cualquier antígeno. Por lo tanto, plantea algunas dudas sobre la seguridad de una gran cantidad de ensayos de vacunas que utilizan grupos placebo con solo el adyuvante de Al.

Con respecto al análisis de expresión diferencial de miARNs, se observó un patrón similar al análisis de ARNm, con apenas algún cambio significativo en los animales del grupo Vac. Cuando se verificó la coexpresión de miARN-ARNm, se encontraron múltiples factores relacionados con la función de las mitocondrias, el mantenimiento de la polaridad neural y el daño del ADN. Entre ellos, había algunas dianas predichas que podrían ayudar a ampliar nuestro conocimiento sobre la toxicidad de Al en el cerebro, apuntando a una desregulación de las funciones mitocondriales. Se ha demostrado que *ACTR10*, que es una diana predicha del *let-7b* (sobreexpresado en el grupo Adj), interrumpe el transporte retrógrado mitocondrial en su ausencia, lo que lleva a la acumulación de mitocondrias en axones terminales. Además, *MRS2*, otro target predicho de *let-7b*, es un transportador de Mg mitocondrial que se ha relacionado con defectos en el orgánulo y apoptosis. En conjunto, parece que el Al puede estar causando un desequilibrio en los niveles de iones metálicos, lo que sería concordante con lo que se ha visto en otras especies, como en ratas tratadas con una administración intragástrica de gluconato de Al.

Como se ha indicado previamente, las librerías paired-end se secuenciaron con una profundidad suficientemente alta como para la caracterización de nuevos ARN ovinos. Los circARN son ARN no codificantes con una estructura cíclica. Recientemente, los circARN han llamado la atención de los investigadores debido a su expresión específica mostrada en múltiples tejidos, su conservación entre especies y su participación en múltiples funciones biológicas, como la diferenciación neuronal, la apoptosis neuronal y la disfunción de la barrera hematoencefálica en el cerebro y funciones inmunes básicas y la regulación de la transcripción en la sangre. Por lo tanto, en este estudio se anotaron nuevos circARNs ovinos para discernir si dichos elementos tienen algún papel en la adyuvancia del Al. Debe señalarse que no existe una base de datos que registre los circARN ovinos y que los tejidos utilizados en este estudio no se han utilizado previamente para la anotación de circARN. Por lo tanto, este estudio ampliará la anotación actual de circARN ovinos.

Después de la caracterización de circARNs mediante dos programas distintos, se encontró una amplia expresión de circARNs en ambos tejidos. Se detectaron un total de 2,510 y 3,403 circRNAs en la corteza del lóbulo parietal y PBMCs, respectivamente, de los cuales 1,379 eran completamente nuevos (841 exclusivos en PBMCs, 421 exclusivos en encéfalo y 117 expresados en ambos tejidos simultáneamente). El origen de la mayoría de los circARNs identificados eran genes anotados, y suponiendo que todos los exones son retenidos, generalmente estaban formados de dos o tres exones distintos, de acuerdo con lo que se ha descrito previamente en humanos y ratones. Además, se ha descrito que algunos circARNs tienen un patrón de expresión dependiente del tejido o del estadio de desarrollo. En nuestras muestras, se detectaron 1,236 circRNAs (36.32% de todos los circRNAs sanguíneos detectados)

expresados simultáneamente en ambos tejidos, lo cual concuerda con aproximadamente el 30% de los circRNAs sanguíneos detectados que se superponen con los circRNAs expresados en el cerebelo en datos humanos.

Los circARNs detectados en este estudio se compararon con otros circARNs ovinos identificados en la glándula pituitaria y en el músculo longísimo, ya que, como se ha señalado anteriormente, no existe una base de datos que los registre. Solo unos pocos circARNs eran expresados simultáneamente en todos los tejidos, mientras que varios cientos de circARNs eran exclusivos de cada tejido, lo que muestra cómo algunos circARNs tienen una expresión tejido-dependiente. Además, dado que múltiples estudios han demostrado que los circARNs tienen una conservación evolutiva entre humanos y ratones, los circARNs detectados en este trabajo se compararon con una base de datos de circARN humanos, CIRCpedia. Aproximadamente el 63% de los circARNs ovinos en ambos tejidos estaban conservados al compararlos con la base de datos humana.

Entre los roles funcionales que se han propuesto para los circRNA, se ha demostrado actividad entre circRNA y proteínas de unión a ARN (RBP), lo que sugiere que los circRNA pueden afectar los mismos procesos funcionales en los que está involucrado el gen correspondiente. Bajo el supuesto de que la función de un circARN puede estar asociada con la función conocida de su gen parental, los circARN estaban relacionados con múltiples funciones inmunes como la proliferación de células B y T, la desgranulación de neutrófilos, la cascada de MAPK y la vía NF-κB en PBMCs, mientras que los circARN de la corteza del lóbulo parietal estaban relacionados con la regulación de la sinapsis, el comportamiento, el proceso de aprendizaje y el desarrollo del cerebro. Además, se ha informado que los circARNs pueden actuar como esponjas de miRNAs, que se definen como secuencias con múltiples sitios de unión con el miRNA que compiten con los genes huésped por la unión con el miRNA. Solo existía un circARN expresado exclusivamente en muestras de corteza, y alojado en el gen CDR1, que contenía múltiples sitios de unión para *miR-7* y *miR 1224*.

Con respecto al análisis de expresión diferencial, debe señalarse que no existe una herramienta diseñada específicamente para los datos de expresión de circARNs. La mayoría de los investigadores usan herramientas como DESeq2 y edgeR, que se basan en una distribución binomial negativa, pero no se ha realizado ningún estudio que demuestre si el modelo binomial negativo es adecuado para los datos de expresión de circARNs, que solo cuenta las lecturas de los extremos de la circularización. Al menos en nuestras muestras, los datos de expresión de circARNs generalmente tenían recuentos muy bajos (unos pocos circRNA altamente expresados originaron la mayoría de los recuentos). Debido a la diferente estructura de los datos de expresión de circARNs y la incertidumbre de si los métodos utilizados hasta ahora son adecuados, se aplicaron diversos métodos a nuestros datos. Independientemente de la elección del método, no detectamos ningún circARN diferencialmente expresado en ninguno de los dos tejidos, lo que puede indicar que los circRNA pueden no estar conectados con la adyuvancia de aluminio.

Conclusiones

En resumen, se observó una activación general de la respuesta inmune en animales de los grupos Vac y Adj en PBMCs, con la activación de la ruta NF-κB en ambos grupos de tratamiento. Se observaron similitudes en la respuesta inmune en ambos tratamientos, pero se demostró que sin la presencia de ningún antígeno la respuesta inmune era más leve. Curiosamente, la expresión del inflamasoma *NLRP3* no era necesaria para la actividad adyuvante de aluminio en los animales de este estudio. Con respecto a la expresión de miARNs y su interacción con los ARNm estudiados, se demostró que múltiples miARNs como *let-7b* (sobreexpresado en la

comparación Adj tf vs. Adj t0), *miR-125b* y *miR-99a* (ambos sobreexpresados en la comparación Vac tf vs Vac t0) puede estar relacionado con la ruta NF-κB, apuntando hacia una expresión altamente regulada de la ruta por miARNs en respuesta a los adyuvantes de aluminio.

En las muestras de corteza del lóbulo parietal, casi no se observó expresión diferencial en los animales inoculados con vacunas comerciales, mientras que en las ovejas inoculadas con adyuvante únicamente se encontraron algunos genes expresados diferencialmente (casi 5 veces más en comparación con el primer grupo). Esto concuerda con los niveles de Al detectados en el tejido, sin diferencias en el contenido de aluminio entre las muestras de control y las ovejas del grupo Vac y con una tendencia a niveles más altos de aluminio en las ovejas del grupo Adj. A partir de los resultados de este estudio, parece que el aluminio de las vacunas comerciales no es capaz de alcanzar la corteza del lóbulo parietal después de una exposición a largo plazo.

Con respecto a la caracterización de circARNs, se encontró una amplia expresión de dichos elementos en ambos tejidos. Además, los circARNs ovinos detectados en este estudio estaban altamente conservados en comparación con los circARNs humanos. Independientemente de la elección del método de expresión diferencial, no hubo circRNA expresados diferencialmente en ambos tejidos, lo que puede indicar una expresión independiente al aluminio de los circARNs.

Publications and contributions throughout the thesis

Articles related to the thesis

(1) Related to chapter 3 - **Varela-Martínez E**, Abendaño N, Asín J, Sistiaga-Poveda M, Pérez MM, Reina R, de Andrés D, Luján L, Jugo BM. Molecular signature of Aluminum hydroxide adjuvant in ovine PBMCs by integrated mRNA and microRNA transcriptome sequencing. Front Immunol (2018) 9:2406. doi:10.3389/FIMMU.2018.02406

(2) Related to chapter 4 - **Varela-Martínez E**, Bilbao-Arribas M, Abendaño N, Asín J, Pérez MM, de Andrés D, Luján L, Jugo BM. Whole transcriptome approach to evaluate the effect of Aluminium hydroxide in ovine encephalon. [Under revision]

(3) Related to chapter 5 - **Varela-Martínez E**, Corsi GI*, Anthon C, Gorodkin J, Jugo BM. Novel circRNA discovery in sheep shows evidence of high backsplice junction conservation. [In preparation]

Collaborations

(a) Bilbao-Arribas M, Abendaño N, **Varela-Martinez E**, Reina R, de Andrés D, Jugo BM. Expression analysis of lung miRNAs responding to ovine VM virus infection by RNA-seq. *BMC Genomics* **20**, 1–13 (2019).

(b) Alonso-Hearn, Canive M, Blanco-Vazquez C, Torremocha R, Balseiro A, Amado J, **Varela-Martinez E**, Ramos R, Jugo BM, Casais R. RNA-Seq analysis of ileocecal valve and peripheral blood from Holstein cattle infected with Mycobacterium avium subsp. paratuberculosis revealed dysregulation of the CXCL8/IL8 signaling pathway. *Sci. Rep.* **9**, 14845 (2019).

(c) Bilbao-Arribas M, **Varela-Martinez E**, Abendaño N, Jugo BM. LncRNAs dysregulated by aluminium adjuvants in sheep PBMCs are related to genes of the immune response. [In preparation]

*PhD student Giulia Corsi is co-author on the publication on circ-RNAs and has contributed to the analysis of the data (more specifically RNA sponges) and to the manuscript writing.

List of abbreviations

AD	Alzheimer's Disease
Adj	Adjuvant group
Ag	Antigen
AH	Aluminium Hydroxide
Al	Aluminium
ALS	Amyotrophic Lateral Sclerosis
APC	Antigen-Presenting Cell
ASIA	Autoimmune/inflammatory Syndrome Induced by Adjuvants
BBB	Blood Brain Barrier
BP	Biological Process
BTV	Bluetongue Virus
cDNA	Complementary DNA
CDS	Coding Region
circRNAs	Circular RNAs
ciRNAs	Intronic circular RNAs
CNS	Central Nervous System
DAMPs	Damage-Associated Molecular Patterns
DC	Dendritic Cells
DE	Differential Expression
DEGs	Differentially Expressed Genes
DLN	Draining Lymph Node
dTTP	Deoxythymidine triphosphate
dUTP	Deoxyuridine triphosphate
ecRNAs	Exonic circular RNAs
EIciRNA	Exon-intron circular RNAs
FDR	False Discovery Rate
GO	Gene Ontology
GS	Gene Significance
IgE	Immunoglobulin E
IgG1	Immunoglobulin G1
INDEL	Insertion or Deletion
KEGG	Kyoto Encyclopedia of Genes and Genomes
lncRNAs	Long Non-Coding RNAs
LPS	Lipopolysaccharide
MHC	Major Histocompatibility Complex
miRISC	miRNA-induced silencing complex
miRNAs	microRNAs
MM	Module Membership
MMF	Macrophagic Myofasciitis
MS	Multiple Sclerosis
mRNAs	Messenger RNA
NC cells	Natural Killer Cells
NCL	Non-co-linear
ncRNAs	Non-coding RNAs
NGS	Next Generation Sequencing

PBMCs	Peripheral Blood Mononuclear Cells
PBS	Phosphate-Buffered Saline
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PD	Parkinson Disease
PE	Paired-End
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
PRRs	Pattern Recognition Receptors
RIN	RNA Integrity Number
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
ROS	Reactive Oxygen Species
SD	Standard Deviation
SE	Single-End
SNPs	Single Nucleotide Polymorphisms
T_H1	T Helper 1 Cell
T_H2	T Helper 2 Cell
T_H17	T Helper 17 Cell
T_{FH}	T Follicular helper Cell
T_{reg}	Regulatory T cells
TOM	Topological Overlap Measure
UTR	Untranslated Region
tRNA	Transfer RNA
Vac	Vaccine group
WGCNA	Weighted network correlation analysis

Table of contents

Chapter 1

Introduction & Literature Review

Since their discovery, adjuvants have been widely used on vaccines to elicit a strong immune response. Aluminium-based adjuvants are the most widely used in veterinary medicine. However, their mechanism of action is not totally understood. Thus, is imperative to further investigate the adjuvants and how they exercise their function.

This chapter briefly introduces a general background on sheep and the adjuvants, mainly focusing in aluminium (Al) hydroxide, the adjuvant this work focuses on. Then I provide an overview of the RNA sequencing technology and the options to analyse the data produced by this technology. Furthermore, a sort summary about microRNAs (miRNAs) is added since data from total RNA and miRNA sequencing is going to be analysed throughout this work. In addition, a brief description of circular RNAs (circRNAs) is given at the end.

1.1 Sheep (Ovis aries)

Being one of the earliest animals domesticated for agricultural purposes, sheep are raised for meat, milk and wool production. Selection for different traits has resulted in a great variety of breeds worldwide. The United Kingdom and Spain accounts for a large proportion (about %45) of the European Union's (EU) sheep, Spain being the second largest producer of sheep meat in the EU (1). In Spain, sheep production is an important field as the country holds about 16 million of sheep.

Apart from studies with agricultural applications, lately livestock species have also been useful as models for human diseases like neurological diseases and heart diseases. Despite the fact that mouse models have a lower cost, they sometimes fail to recreate some aspects of the disease at study, while larger animals like sheep share more similarities to human regarding physiology, anatomy, metabolism, genetics and size, making them a good alternative (2–4). Notwithstanding such advantages, the purchase and maintenance of larger animals is more expensive and require special facilities.

In recent years there have been an increase in the research of multiple diseases and traits of agricultural interest in small ruminants by next-generation sequencing technology, specially RNA sequencing (RNA-seq). This is due to the recent reduction in cost of such technology, allowing the inclusion of more samples per experiment, and due to the power of the tool for exploring the genetic architecture of complex traits. However, the ovine reference genome lag behind those of other domestic species such as cattle and further work is needed to improve its annotation.

Due to the importance of sheep production in Spain, sheep are highly supervised with disease prevention and control measures and they are put under highly strict vaccination schedules. All flocks are inoculated on a routine and systematic basis with aluminium-based adjuvanted vaccines against different pathogens and for antiparasitic treatment. Taking all that into account, each animal is given an average of 4 inoculations (range 2-9, depending of the

flock) with aluminium hydroxide (AH) salts as adjuvant per year during the 7-8 years in which the animal is considered economically productive (5).

Figure 1.1: Photo of a Rasa Aragonesa sheep (photo adapted from the feagas federation web page, https://feagas.com/razas/ovino/rasa-aragonesa/#!).

All the analyses carried out in this work have been done in a specific sheep breed, Rasa Aragonesa (See Figure 1.1). The Rasa Aragonesa is a sheep breed that can be found in the Northeast of Spain in most of the Ebro basin (See Figure 1.2). They are very well adapted in all geographical means of the Ebro Valley, from the most arid environments to the valleys and ports of the Pyrenees.

Rasa Aragonesa sheep are physically characterized as hornless, without wool in their head, average size and weight with balanced proportions and a sub-convex profile, reaching convex in males. This breed is mainly raised for meat production (https://feagas.com).

Figure 1.2: Distribution of the Rasa Aragonesa breed investigated in this work in Spain (photo adapted from the feagas federation web page, https://feagas.com/razas/ovino/rasa-aragonesa/#!).

1.2 Adjuvants

Adjuvants can be described as substances added to vaccines, especially to those containing killed organisms or highly purified antigens (Ag), to enhance the immune response. Adjuvants can be used to increase the speed of the response to the antigen, to reduce the quantity of antigen needed to induce a long lasting immune response, to reduce the number of doses needed or to modify the immune response to particular cells of the immune system, acting directly on T or B cells or using them to bias towards a T_H1 or T_H2 response. Independently of their mode of action, the adjuvants achieve an immunostimulatory effect. The choice of adjuvant would be determined by the antigen, the desired immune response and the route of administration (6).

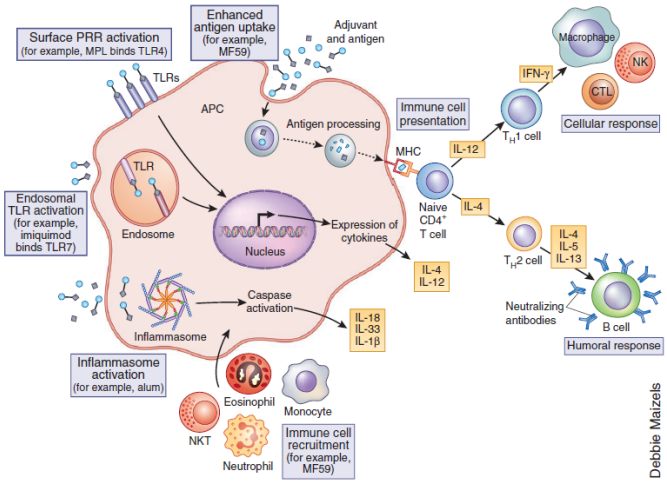


Figure 1.3: Proposed mechanism of action for adjuvants. Many adjuvants act as ligands for pattern recognition receptors (PRRs), inducing the production of cytokines and chemokines that direct the immune response towards a T_H1 or T_H2 response. Activation of the inflammasome has also been shown to induce the secretion of pro-inflammatory cytokines IL-1β and IL-18. Some adjuvants also influence antigen presentation by major histocompatibility complex (MHC) or induces cell recruitment at the injection site (figure adapted from (6)).

Table 1.1: List of some of the most common adjuvants in humans and in animals in use or in clinical trial classified by their source. The table is based on references (7,8,10–18).

Adjuvant	Type of immune response	State
Mineral salts		
Alum	T_H2, T_{FH}, IgG1/IgE	Not used anymore
Inject-alum	T_H2, T_{FH}, IgG1/IgE	Only experimental immunology
Alhydrogel [aluminium hydroxide]	T_H2, T_{FH}, IgG1/IgE	Numerous licensed products
Adju-Phos [aluminium phosphate]	T_H2, T_{FH}, IgG1/IgE	Numerous licensed products
Calcium phosphate	T_H1	
Oil emulsions and surfactant-based formulations		
MF59	T_H1, T_H2, IgG2a, IgG1	Fluad
QS-21 [purified saponin]	T_H1, T_H2, T_c	Clinical Phase 3
AS02 [MPL + QS-21 in oil-on-water emulsion]	T_H1	Development discontinued
AS03	T_H1, T_H2	Pandemrix
ISA-51 and ISA-720	T_c	ClimaVax-EGF [only available in Cuba and some south American countries]
Particulate adjuvants		
virosomes	T_H1, T_H2	Epaxal, Inflexal V
AS04 [Al salt + MPL]	T_H1	Cervarix, Fendrix
ISCOMS [structured complex of saponins and lipids]	T_H1, T_H2, T_c, IgG2a	Clinical Phase 2
PLG [polylactide co-glycolide]		Clinical Phase 1
Microbial derivatives		
MPL [monophosphoryl lipid A]	T_H1, T_{FH}, T_c	
Detox [MPL + *M. Phlei* cell wall skeleton]		
AGP	T_H1	
DC-Chol	T_H1	
CpG motifs	T_H1	Heplisav-B
Modified LT and CT	T_H1, T_H2, IgA, IgG1, IgE	
Endogenous human immunomodulators		
hGM-CSF	T_H1, T_H2	Clinical Phase 1/2
hIL-12	T_H1, IgG2a, IgG2b	
Inert vehicles		
Gold particles		

Abbreviations: T_H1, T helper 1 cells; T_H2, T helper 2 cells; T_{FH}, Follicular B helper T cells; T_c, Cytotoxic T cell; IgG1, Immunoglobulin G1; IgE, immunoglobulin E; CT, cholera toxin; LT, *Escherichia coli* heat labile enteroroxin;

The discovery of vaccine adjuvants can be traced back to 1925, where Gaston Ramon, while developing a diphtheria and tetanus vaccine, demonstrated that adding different

substances to the antigen (tapioca, lecithin, oil, saponin,…) could improve the immune response (11). Later in 1926, Alexander Thomas Glenny demonstrated an increase in effectiveness of the diphtheria toxoid by precipitating it with aluminium salts (17). In the 1930's, aluminium salts were the first adjuvants used for human vaccines and they continue to be the most commonly used adjuvants in human and veterinary vaccines. For more than 60 years, aluminium salts have been the only licensed adjuvants in commercial vaccines, but around 1990's new vaccines with other compounds as adjuvants (mainly squalene emulsion based adjuvants and virosomes) emerged for commercial use (17). Currently, there is ongoing research in an attempt to improve current adjuvant formulation and to find new compounds for triggering different immune reactions.

Adjuvants are an heterogeneous group of compounds and they can be classified by different traits: their source, the mechanism of action (See figure 1.3) or physical or chemical properties (13). In table 1.1 can be seen a list of the current most common adjuvants in use or in clinical trial classified by their source. The Food and Drug Administration in the United States and the European Union has only approved in humans the use of aluminium salts, AS04 (Cervarix vaccine, which is a vaccine to prevent cervical cancer. No longer commercialized in U.S.), AS03 (in a vaccine for prevention of H5N1 influenza or commonly known as bird flu), MF59 (Fluad vaccine, which is a vaccine to prevent influenza in adults over 65 years), AS01B (Shingrix vaccine, for the prevention of shingles) and CpG 1018 (Heplisav-B vaccine, for the prevention of infection caused by hepatitis B virus) as adjuvants. Other adjuvants like mineral oil emulsions or complete Freund's adjuvant are thought to be too reactogenic for human use, some of them being only licensed for veterinary vaccines.

Since their discovery, aluminium salts are the most widely used compounds in human and veterinary vaccines. Despite the multiple intents trying to decipher their mechanism of action, it remains elusive and only a partial picture has been achieved. In some cases, even contradictory elements have been reported as will be seen below. All the experiments carried out in this work have been done with vaccines composed of aluminium hydroxide (Alhydrogel).

1.2.1 Aluminium Hydroxide (Alhydrogel)

1.2.1.1 Physicochemical properties

Despite being called Aluminium Hydroxide (Alhydrogel), the adjuvant used in vaccine formulations is not chemically $Al(OH)_3$. Such adjuvants are usually prepared adding alkali to the solution of aluminium salts, generating a poorly crystalline aluminium oxyhydroxide [AlO(OH)] (18), also termed poorly crystalline boehmite (or Pseudo-boehmite). It is usually prepared using $AlCl_3$ or alum [$AlK(SO_4)_2$] as the aluminium solution and sodium hydroxide as alkali (19). In spite of the fact that "aluminium hydroxide-based adjuvant" does not reflect its true chemical composition, in this thesis, we will continue using this term to refer to the adjuvant, since that naming has been used and accepted for many years.

Aluminium hydroxide gels are composed of elongated nanoparticles (crystals) with an approximate size of 4 x 2 x 10 nm (19) (See figure 1.4). These nanoparticles tend to self-associate forming micro- and nanoparticles of variable size. In its native form, Shardlow et al. (2016) showed by photon correlation spectroscopy that Alhydrogel nanoparticles had a size range between 955-7456 nm with the majority of the intensity being between 2.2-3.3 µm (20). Although it has to be noted that such spectroscopy has a maximum limit, being only capable of detecting particles up to 10 µm. It has been reported that the nanoparticles can form aggregates

up to 17 µm in size (21). Moreover, AH adjuvant has a mean surface area of 509 m^2/g (with a 95% confidence interval of 30 m^2/g), being such high value an important characteristic of the adjuvant, which is one of the reasons for its high protein absorption capacity (22). Burrell et al. (2000) reported that aluminium-containing adjuvants stored at room temperature during a 15 month period become more ordered, with a consequent reduction of the surface area and a lower absorption capacity (23).

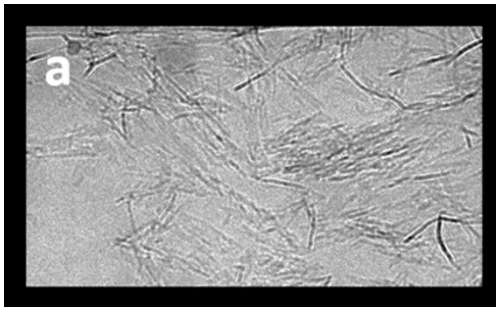


Figure 1.4: Structure of aluminium hydroxide adjuvant. (figure adapted from (19))

1.2.1.2 Mechanism of Action

Despite aluminium adjuvants being one of the most used in human and veterinary vaccines since their discovery, the mechanism of action by which they elicit an immune reaction is not fully understood. Several mechanisms have been proposed and it seems that rather than having a dominant one, aluminium elicit a strong immune response operating simultaneously through multiple pathways. Some different mechanisms have been proposed: 1) The depot effect, also known as repository effect; 2) Induction of inflammation, leading to activation and maturation of immune cells, with a direct effect in the uptake of antigens by APCs; 3) The "Danger Theory".

After inoculation of the vaccine into the organism, a depot is formed at the inoculation site. The antigen is slowly released by the adjuvant and long-lasting interactions between antigen and APCs are promoted. The release of the antigen would be influenced by properties of both antigen and adjuvant previously described, such as association strength or adjuvant particle size. Experiments by Harrison et al. (1935) showed that the antitoxin was present in the serum after re-injection of the nodules formed after inoculation of an alum precipitated toxoid from a guinea pig to a second pig (24), verifying the depot effect. However, the depot effect alone cannot explain completely the immune reaction caused by aluminium adjuvants. In addition, there have been different reports suggesting that the depot effect is not necessary for the immune reaction caused by aluminium adjuvants. Holt et al. (1950) inoculated diphtheria toxoid adsorbed with aluminium adjuvant into guinea pigs and demonstrated that even cutting off the tissue where the inoculation was done after 7 days post vaccination, the vaccine was still able to carry on its immune modulation (25). Another study showed in mice that removal of the injection site with the associated alum depot 2 hours after immunizations had no effect in the magnitude of antigen-specific T- and B-cell responses (26). Together, these studies indicate that the depot effect is not necessary for aluminium adjuvants to carry out their role as long as the concentration of Ag is high at the injection site and engulfed by APCs.

It has been shown that aluminium adjuvants create a local pro-inflammatory environment, recruiting different immune cells at the site of injection (See figure 1.5). Such cells are composed of neutrophils, eosinophils, natural killer cells (NK cells), CD11b* monocytes, macrophages and immature dendritic cells (DC) (27,28). The recruitment of those cells lead to expression of messenger RNAs (mRNAs) related to chemokines, cytokines and cell adhesion

molecules and their secretion (29). Different studies have been trying to address the signalling pathways that triggers DC and macrophages activation after Al exposure (See figure 1.6).

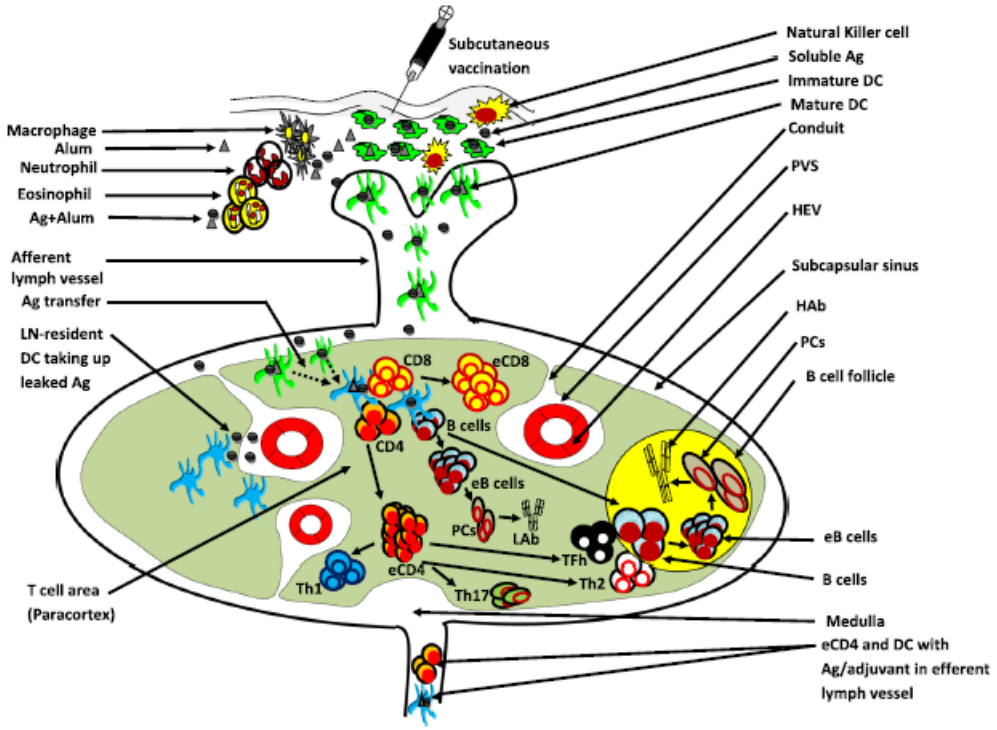


Figure 1.5: Current understanding of aluminium based adjuvants *in vivo*. After aluminium administration, different cells are recruited at the injection site. Immature DC take up the soluble Ag or particulate Ag together with the aluminium and migrate towards the draining lymph node (DLN). Then, after leaking out of conduits the Ag, resident DCs take up the Ag and present it to naïve T cells. As a result of CD8 and CD4 T cell activation, effector CD8 and CD4 T cells (eCD8 and eCD4, respectively) are produced. eCD4 cells can polarize into T helper (T_H) 1, 2, 17 or T follicular helper (T_{FH}) cells. In the case of aluminium based adjuvants, they mostly polarize into T_H2 and T_{FH} cells. Those cells can reach to the border of B cell follicle, activating B cells that produce effector B cells (eB) and then differentiate in plasma cells (PCs). As a result, high-affinity antibodies are secreted. (figure adapted from (29))

According to the danger theory, pathogens or other elements such as adjuvants cause tissue damage at the injection site. Then, endogenous danger signals are released from the damaged tissues which induces an inflammatory response, initiating an adaptative immune response (28). These endogenous danger signals, also known as damage-associated molecular patterns (DAMPs), have been claimed to be released from necrotic cells or stressed cells, due to endocytosis of the adjuvant (30). Aluminium hydroxide has been reported to produce granulomatous inflammatory reactions and promote local necrosis in vaccinated muscle tissue (31). Usually uric acid, ATP, host DNA or HMGB1 are released after tissue damage. In the case of aluminium based adjuvants, the release of uric acid (32) and host DNA (33) has been reported at the injection site.

Uric acid release activates the NLRP3 inflammasome (34), also known as NALP3 and cryopyrin. The NLRP3 inflammasome is a large multiprotein complex composed of the NLR (nucleotide-binding oligomerization domain and leucine-rich repeat-containing receptor) protein NLRP3, the adapter ASC (apoptosis-associated speck like protein containing a caspase recruitment domain) and pro-caspase-1 and it participates in the production of the pro-inflammatory cytokines IL-1β and IL-18. There is controversy about the requirement of the NLRP3 inflammasome in the alum-induced response. *In vitro*, aluminium-containing adjuvants stimulate the production of IL-1 β through the inflammasome (35,36). Despite the agreement

on the role of the inflammasome *in vitro*, there is no consensus in how this translates to *in vivo*. There have been recent studies in which NLRP3 deficient mice vaccinated with aluminium hydroxide-based adjuvant showed no significant effect on T and B cell responses (37,38). Despite not knowing if the inflammasome is needed for Al adjuvancity, there is *in vitro* and *in vitro* evidence that shows the importance of inducible HSP70 in IL-1β expression induced by Al (39). Another study showed that both IL-1α and IL-1β were essential for neutrophil infiltration at the injection site and that the NLRP3 inflammasome was dispensable for such effect, while cathepsin S (*CTSS*) was indispensable for IL-1β induction at the injection site (40). Futhermore, aluminium adjuvants have been shown to induce host DNA release. It has been proposed that host DNA signalling differentially regulates production of immunoglobulin E (IgE), through an IRF3 (interferon response factor 3)-dependent mechanism, and immunoglobulin G1 (IgG1) (33). TLR9 can sense DNA and it would lead to IL-1β production via NF-κB activation or IFNβ production via IRF3 activation (27).

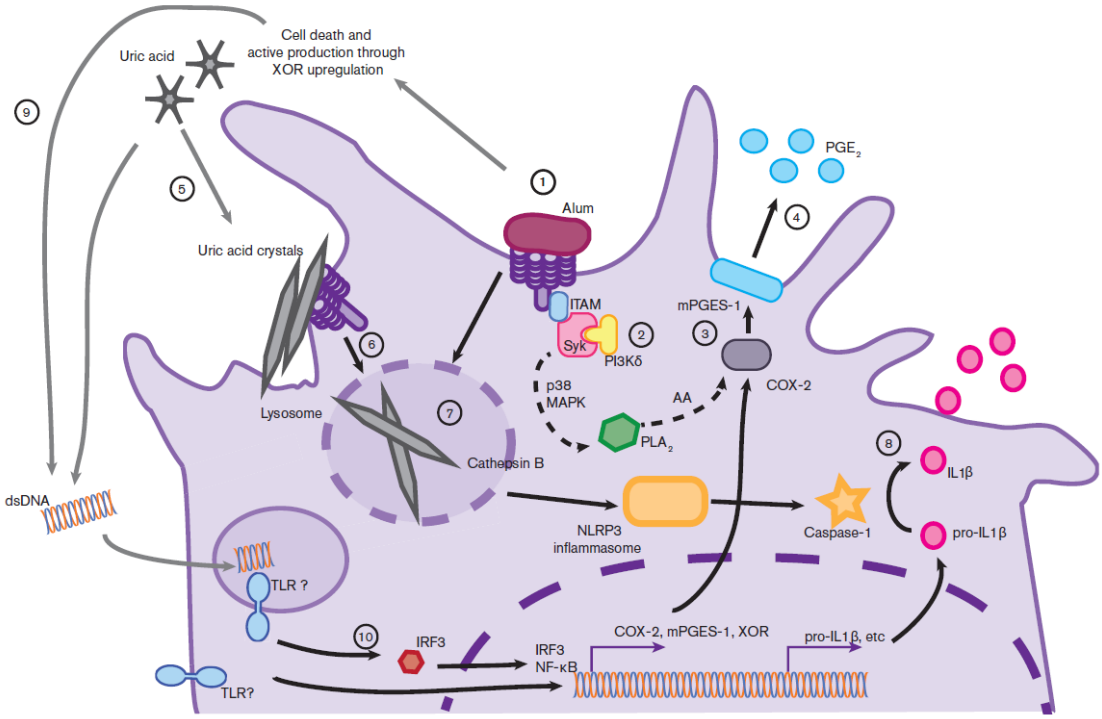


Figure 1.6: Pathways induced by aluminium in DCs and macrophages. **(1)** The binding of alum at the cell membrane causes lipid raft formation. **(2)** In the lipid raft, ITAM-containing receptors cluster and activate the Syk-PI3Kδ pathway. **(3)** Then, cytosolic phospholipase A2 (cPLA$_2$) is activated probably via the p38 MAP kinase, which in turn results in release of arachidonic acid (AA) from membrane lipids. COX2 and mPGES-1 convert AA to prostaglandin E2 (PGE$_2$). **(4)** PGE$_2$ instructs the TH$_2$ response upon release from the cell. **(5)** In addition, aluminium induces the release of uric acid. **(6)** Lysosomal damage is induced after phagocytosis of aluminium or uric acid crystals. **(7)** Activation of the NLRP3 inflammasome by the release of enzymes like Cathepsin B into the cytoplasm. **(8)** Activation of caspase-1 lead to cleavage of proIL-1β into active IL-1β. **(9)** Aluminium can also induce the release of host DNA due to necrosis at the injection site. **(10)** Host DNA activates monocytes through IRF3, being critical for migration of inflammatory monocytes. (figure adapted from (27) and based on (41))

Despite AH not being antigenic itself, there have been multiple reports on immune stimulation without any adsorbed antigen and, as consequence, without long-term release of antigen (30,42). Güven et al. (2013) showed that aluminium hydroxide adjuvant activated the three complement pathways with a major involvement of the alternative complement pathway, providing an explanation to how the adjuvant was able to stimulate an immune response by

providing a “surface” for complement activation, antigen opsonisation and stimulation of antigen removal through complement receptors (43).

As previously mentioned, several mechanisms of action have been proposed for the aluminium adjuvant, and rather than having a dominant one, it seems that multiple pathway are activated simultaneously. In some cases, contradictory results have been obtained, making it difficult to reach an agreement. Those differences may be explained by the physicochemical properties of the adjuvant used in each study, with varying particle size or difference in antigen and adjuvant under study, or by the animal model in use, as each animal model has different Toll-like receptor (TLR) expression patterns (6). Despite those differences, the fact that AH induces mainly a T_H2 response, T_{FH} cells and the antibody isotypes IgG1 and IgE is not under dispute (7). The limited capacity of Al based adjuvants to induce a T_H1 immune response has been addressed by different studies. It has been shown that Al adjuvants activates the Syk-PI3 pathway (41) and activation of the PI3 kinase inhibits secretion of IL-12p70 cytokine by DCs, which is a key cytokine that drives the polarisation of naïve T cells towards a T_H1 phenotype (44). In addition, other study showed that Al adjuvants promote IL-10 expression, which can block T_H1 responses, and IL-10-deficient mice were able to show an increase in T_H1 responses after vaccination of Al adjuvants (45).

1.2.1.3 Aluminium translocation to other tissues

Although the benefits of vaccination have never been questioned, there is disagreement on the degree of safety of aluminium-containing vaccines. There is some controversy regarding the capacity of aluminium adjuvants to reach distant organs after a long-term exposure. In a research study in which New Zealand White rabbits were injected intramuscularly with labelled AH, it was determined by accelerator mass spectrometry that Al was detectable in blood one hour after vaccination and that the body was able to partially eliminate through urine the Al absorbed from the adjuvant, but only a 6% of the AH adjuvant dose was eliminated after 28 days post vaccination (46). The retention level seen for AH is consistent with its expected low solubilization rate (47). In the same study, the following distribution profile to tissues for the AH adjuvant was seen for the short time of the experiment: kidney > spleen > liver > heart > lymph node > brain (46). Such study had a limited number of samples and there is a need of further long-term studies on a larger number of animals.

In a more recent study, after intramuscular injection of the aluminium adjuvant in mice, the material was translocated at a very slow rate in normal conditions to draining lymph nodes (DNL) and thereafter was detected associated with phagocytes in blood and spleen (48). It has been shown in mice that AH particles are transported by cells of monocytic lineage to the DLN and bloodstream, reaching distant organs such as the spleen (48,49). Furthermore, there is some controversy regarding the capacity of Al to reach such a distant organ as the brain. Multiple studies has shown a slow Al translocation to the brain after intramuscular injection (48,50), while in another study in which CD1 and C57BL/6J mice received an intramuscular injection of AH, the CD1 mice showed a lack of Al translocation to brain and ,in contrast, in the C57BL/6J mice Al was still detected one year later (51). Such differences in both strains may be due to differences in the genetic background. Interestingly, C57BL/6J strain mice produce more MCP1/CCL2 than other strains (51), a key cytokine which recruits monocytes, memory T cells and DCs to the sites of inflammation.

Taking all into account, it is clear that aluminium adjuvants are extremely biopersistent and such biopersistence can be seen at the site of injection and in distant organs such as the

DLN and spleen, although the translocation of Al to brain needs further research. The slow translocation and clearance of Al has raised some concern due to extensive vaccination campaigns that some farm animals are subject to during their productive period.

1.2.1.4 Toxicity

Aluminium is the most abundant metal element in the biosphere and humans are frequently exposed to it as Al is being widely used in different fields such as pharmaceuticals and to a lesser extent in foods (e.g. in food additives or due to contaminants) and water (due to water treatment process or from weathering rocks and soils). Al is routinely taken up by the human body through ingestion and inhalation. The neurotoxic effects of aluminium on brain are well known. It has been shown that Al nanoparticles are capable of surpassing the blood brain barrier (BBB) and elicit an immune inflammatory response (52). High-level exposure to Al could impair neuronal functions, while low-level and long-term exposure to Al has been linked to some neurological diseases like Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), dementia and Parkinson disease (PD) (53). It has been shown that Al causes oxidative stress and mitochondrial dysfunction, being involved in the production of reactive oxygen species (ROS) (54). Recent reports about translocation of Al adjuvants to the brain has gained more attention, although it remains a topic of constant debate. It has been pointed that there is prolonged retention of a fraction of Al that enters the brain, which may accumulate within repeated exposures (54). Environmental Al has been suspected to act as a co-factor for some neurological diseases and, due to some reports of Al adjuvant translocation to brain, the idea that Al adjuvants may not be totally safe over a long-term exposure in predisposed individuals has emerged.

There are multiple studies about the toxic properties of aluminium in humans exposed via inhalation, oral, or dermal exposure, supported by a large number of studies in laboratory animals. Numerous studies have reported that subjects exposed to airborne Al showed reduced respiratory functions (55) and aluminium welders and workers exposed to high levels of Al have showed a declining performance in neuropsychological tests (56). In contrast, the absorption of Al via dermal exposure is poorly understood, although there are some reports on antiperspirants which use Al as a component, mainly in the form of aluminium chloride and aluminium chlorohydrate, that has linked their use with breast cancer (57). Such link is not well supported by consistent scientific data (56,58) and needs further research. In addition, cases of skin irritations associated with cosmetic products containing aluminium have been reported in human (59). Regarding the Al intake via the oral route, multiple animal studies have identified the nervous system as the major target of Al toxicity, while others have found adverse effects such as impaired erythropoiesis in rats, erythrocyte damage and increased susceptibility to infection (60).

In a recent research study by Crépeaux et al. (2016), mice were subject to multiple AH adjuvant injections to study the dose-response (200, 400 and 800 µg Al/Kg groups) and an unusual pattern limited to the low-dose group was observed (61). Mice injected with 200 µg Al/Kg displayed a decreased locomotor activity and a higher cerebral Al level, associated with a nearly complete disappearance of the aluminium-induced granulomas. In contrast, those changes were not seen in the highest dose groups. Such difference may be due to aluminium agglomerates size, as the higher doses formed larger agglomerates, probably making it more difficult for transportation by monocyte-lineage cells.

To sum up, aluminium is a non-essential element for the human body and is thought to serve no essential biological function, but humans are constantly exposed to such element in their daily lives. Under low levels of exposition, the body is able to handle it, but upon high level exposures, long-term exposures or predisposed individuals, it seems that Al can affect the health, especially impairing the central nervous system (CNS).

1.2.1.5 Aluminium and diseases

As previously stated, Al has been linked to some neurological diseases like Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), dementia, multiple sclerosis (MS) and Parkinson disease (PD) (53). See table 1.2 for a list of diseases associated with aluminium. Accumulation of Al within the CNS seems to reach a threshold in which it is able to induce proinflammatory signalling, dysregulation of gene expression, brain cell damage and a functional decline in neurons that results in cognition, memory and behaviour deficits (62). Despite Al elicits its neurotoxic effect through a myriad of mechanism, oxidative stress (accumulation of high levels of ROS) seems to be a common mechanism in many of such diseases (63). In addition to oxidative stress, Al is able to modify hippocampal calcium signal pathways, which are crucial to neural plasticity and therefore to memory (56).

AD is a neurodegenerative disease disorder characterized by the presence of extracellular senile plaques containing the amyloid protein (Aβ) and neurofibrillary tangles composed of hyperphosphorylated tau protein, perturbed metal (copper, iron and zinc) homeostasis, oxidative stress, neuroinflammation, irreversible loss of neurons and other pathologies (64,65). Since the discovery that brain tissues of AD patients contains more Al than non-demented age-matched controls and that Al is not homogeneously distributed in the tissue (66), Al is thought to be a potential environmental risk factor for the disease. However, it is not clear whether the Al is the cause of the disease onset, or whether is a change produced due to the disease pathology. Furthermore, the contribution of cumulative doses of Al to AD is still unknown and further research is needed.

Similar to AD, exposure to aluminium has been identified as a possible contributor to MS. MS is a demyelinating neurodegenerative disease of unknown aetiology. Patients with relapsing-remitting and progressive MS were found to have a significant increase in Al concentration in urine (67). In addition, in a recent study, brain tissues from 14 donors diagnosed with MS were evaluated with transversely heated graphite furnace atomic absorption spectrometry and all patient had at least one brain tissue with a pathologically significant Al concentration (68).

PD is a neurodegenerative disease characterized with neuronal loss and presence of Lewy bodies (abnormal aggregates of protein, g.e. α-syn, that develops in nerve cells) in the surviving neurons. Mitochondrial dysfunction, oxidative/nitrative stress, microglia activation and inflammation have been suggested responsible for the neuronal death (69). Several studies have shown increased Al concentrations in the substantia nigra of PD patients compared to controls (70).

Apart from neurodegenerative diseases, aluminium has been linked to some autoimmune/inflammatory diseases such as macrophagic myofasciitis (MMF) and autoimmune/inflammatory syndrome induced by adjuvants (ASIA). Autoimmune diseases arise when the immune system recognise self-antigens as foreigners, leading to inflammation and tissue death. The possible link between some adjuvant and autoimmune diseases is quite controversial. Such diseases are complex and a combination of genetic, hormonal and/or

environmental factors may play a role, making it difficult to attribute causality (71). Despite multiple studies have pointed out a probable relationship between aluminium adjuvants and genetically predisposed individuals, there is no clear evidence of a causal association.

Table 1.2: Diseases associated to aluminium exposure. It must be pointed that in the majority of cases is not clear whether Al is the direct cause of the disease or only an environmental factor associated with it. Most of the associations are under constant discussion in the scientific community. The table is based on references (41,72–77).

Disease	Al Exposure	Symptoms
Potroom Asthma	Al dust and gases (Inhalation)	Workers on a broad range of aluminium factories affected. Shown respiratory symptoms such as cough, phlegm, dyspnea, wheezing and chest tightness. Some workers have shown a long-term impairment and an asthma-like syndrome.
Dialysis Encephalopathy	Dialysis	A complication surged from prolonged haemodialysis exposure in chronic renal failure patients, linked to Al-containing phosphate binders. Characterized by speech alterations, dyspraxia, unconsciousness and psychosis following ataxia and dementia, between others.
Alzheimer Disease	Dietary (Oral)	High concentration of Al increase amyloid aggregation and deposition, which is one of the main features of the disease. Typical sign of the disease are neurofibrillary tangles, deposition of extracellular senile plaques and loss of synapses and neurons, between others.
Parkinson's Disease	Dietary (Oral)	Al^{+3} toxicity impairs iron metabolism, which results in accumulation of iron in neurons and, consequently, oxidative damage. Characterized by selective neuronal death in substantia nigra, its main symptoms are difficulties in speaking and a decrease in motor abilities.
Multiple Sclerosis	Unknown Adjuvants associated	High Al concentrations in urine and brain samples of patients with the disease. Characterized as a demyelinating disease of unknown cause.
Gulf War Syndrome	Adjuvants associated	A spectrum of disorders among veterans of the Persian Gulf War (1990-1991). Characterized by fatigue, muscle pain, emotional disorders, stress and memory loss.
Amyotrophic Lateral Sclerosis	Adjuvants associated	A neurological disease characterized by muscle weakness, disability and eventually death by respiratory failure. It may be part of Gulf War Syndrome.
ASIA	Adjuvants associated	Condition in which repeated exposure to aluminium-based adjuvants lead to aberrant autoimmune responses in susceptible individuals. The syndrome encompasses a diverse group of disorders including siliconosis, MMF, Gulf war syndrome (GWS) and post-vaccination phenomena.
Macrophagic Myofasciitis	Adjuvants associated	Characterized by inflammatory macrophages with agglomerates of nanocrystals, which contain aluminium, in their cytoplasm. Patients show diffuse myalgia, arthralgia and fatigue.

MMF is a disease (described first in France) characterized by inflammatory macrophages with agglomerates of nanocrystals, which contain aluminium, in their cytoplasm and associated

microscopic muscle necrosis in biopsy samples of the deltoid muscle. MFF patients are characterized by diffuse myalgia, arthralgia and fatigue. Taking into account that macrophages contained Al, that patients manifesting the lesion had no particular exposure to Al other than previous vaccinations and that the lesions were localised near the usual injection site for vaccines in the deltoid muscle, it was thought at first to be related to the immunization. Now is clear that the rapid emergence of MFF in France was mainly due to three factors: (1) change of vaccination route in the early 1990s, from the subcutaneous route to the intramuscular route; (2) extension of hepatitis B vaccination to the adult population; and (3) use of the deltoid muscle for routine muscle biopsy, while the biceps brachialis and the quadriceps femoris are preferred in most other countries (77). The Global Advisory Committee on Vaccine Safety (GACVS) from the World Health Organization (WHO) has discussed the safety of aluminium-containing vaccines, reaching the conclusion that at present there is no evidence of a health risk from such vaccines and suggest that further research is necessary to determine if there is a link between MFF and aluminium-containing vaccines, as it may be only a coincidence.

Table 1.3: Suggested diagnostic criteria for ASIA (table adapted from (78)).

Diagnosis criteria for ASIA
Major Criteria:
1. Exposure to an external stimulus (infection, vaccine, silicone, adjuvant) prior to clinical manifestations. 2. The appearance of "typical" clinical manifestations: - Myalgia, Myositis or muscle weakness. - Arthralgia and/or arthritis. - Chronic fatigue, un-refreshing sleep or sleep disturbances. - Neurological manifestations (especially associated with demyelination). - Cognitive impairment, memory loss. - Pyrexia, dry mouth. 3. Removal of inciting agent induces improvement. 4. Typical biopsy of involved organs.
Minor Criteria:
1. The appearance of autoantibodies or antibodies directed at the suspected adjuvant. 2. Other clinical manifestations (i.e. irritable bowel syndrome) 3. Specific HLA (i.e. HLA DRB1, HLA DQB1) 4. Evolvement of an autoimmune disease (i.e. Multiple Sclerosis, Systemic Sclerosis).

ASIA is a recently identified new condition (for now is not recognised as an official diagnosis) in which repeated exposure to aluminium-based adjuvants lead to aberrant autoimmune responses in susceptible individuals. The syndrome encompasses a diverse group of disorders including siliconosis, MMF, Gulf war syndrome (GWS) and post-vaccination phenomena, which share clinical and pathogenic resemblances. A recent study has suggested that the interval from exposition to severe ASIA manifestation can range from 2 days to 23 years (76). In table 1.3 can be seen the suggested diagnostic criteria for ASIA. Patient must show at least 2 major criteria or 1 major and 2 minor criteria to be diagnosed. Since the proposition of the syndrome and the diagnostic criteria, it has been severely criticized by part of the scientific community. One of the major criticisms to the syndrome has been that the diagnostic criteria are exceptionally broad, it appears to include all patients with an autoimmune disorder (79). A registry for ASIA has been recently created in an attempt to increase the knowledge on the clinical and laboratory presentation (80). An animal model of ASIA in commercial sheep has been recently described (5). Sheep exposed to multiple aluminium-based vaccinations from bluetongue vaccines showed

an acute phase occurring 2-6 days after vaccination. Animals were characterized by nervous clinical signs such as lethargy, bruxism, stupor, abnormal behaviour, disorientation and low response to external stimuli. Most sheep recovered from this phase, but some of the sheep exposed to external stimuli (mainly cold weather) progressed towards a chronic phase (it could appear directly without previous manifestation of the acute phase). In the chronic phase, aluminium was detectable in blood and animals were characterized by neurological abnormalities, cachexia, anasarca, coma and ending in death (81).

Despite aluminium been linked to different neurological and autoimmune diseases, a causative role for it needs to be probed. This is a theme under constant debate through the scientific community and further research, preferably *in vivo* and in a long-term experiment, needs to be carried out.

1.3 RNA sequencing (RNA-seq)

RNA-seq methods are part from what is known as the second-generation high-throughput sequencing technology, also known as next generation sequencing (NGS). Such methods allow the determination of individual nucleotides from RNAs in a massively parallel format, obtaining millions of sequences from samples. As a result, a picture of the transcriptome is achieved from cells or tissues. A typical RNA-seq experiment returns sequences between 50 or 150 nucleotides long, so such methods are not capable of capturing completely long RNAs in a single reaction. Recently, the third-generation sequencing technology has started to be applied, in which individual molecules of DNA or RNA are used as templates, obtaining in each reaction sequences of 1500 nucleotides approximately. Despite the gain in sequence length, there is a loss in sequencing depth, as only a few millions of sequences are obtained in each run of the technology. There is a continuous increase of studies applying RNA-seq, as the technology price is decreasing, making it more affordable to all laboratories across the world.

Little by little, RNA-seq has replaced DNA microarrays for gene expression pattern analyses. Briefly, DNA microarrays consist of predetermined nucleic acid probes attached to a surface, in which labelled (with a fluorescent marker) complementary DNA (cDNA) derived from cellular RNA is put over the array and the quantity of cDNA in each probe is assessed by the use of lasers. When comparing both technologies, RNA-seq has some clear advantages. First of all, RNA-seq does not require *a priori* knowledge of the organism under study, making it an ideal technology for the discovery of uncharacterized transcripts, while microarrays only search for transcripts whose known sequence has been attached to a probe. Due to that, RNA-seq can be used well with non-model organisms such as sheep, goat and pig. Secondly, RNA-seq can detect lowly- and highly-expressed transcripts more effectively. In microarrays, lowly-expressed transcripts emit such a low fluorescence that it is hard to differentiate from background fluorescence, while for highly-expressed transcripts the signal become saturated. Moreover, each probe in a microarray can differ in their hybridization properties. Despite the advantages of RNA-seq, such methods produce enormous quantities of data that need to be analysed more meticulously, the pipelines for analysis of data produced in RNA-seq protocols are not totally in consensus and it still remains an expensive technology (although its drop of prize has made it more affordable for laboratories). For the study of model organisms and well-defined transcripts, microarrays still remain as the preferred option, since for a cheaper prize more samples can be included into the experiment. In addition, it has been shown that microarray intensity levels and RNA-seq mapped reads are highly correlated, differing when the array intensities are large and the sequence counts low (82).

1.3.1 Library preparation

Prior to library preparation, the isolated RNA must be quality checked for degradation, purity and quantity. One of the most used platforms for this are the NanoDrop Spectrophotometer (Thermo Scientific Inc, Bremen, Germany) for RNA quantity and purity and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) for RNA integrity and concentration. Nanodrop measures with a spectrometer utilizing a linear charge-coupled device array the light (produced by a pulsed xenon flash lamp) passing through a sample. 1µl samples are enough to ensure accurate and reproducible results. The platform is not able to differentiate between DNA and RNA and degraded RNA give similar readings as intact RNA, so the platform is unable to check for DNA contamination or the quality of the sample (83). Despite all of that, NanoDrops is able to measure the sample concentration in ng/µl and the 260/280 and 260/230 absorbance ratios. The 260/280 absorbance ratio is used to assess the purity of DNA or RNA, being a ratio of ~2 considered as "pure" for RNA. Lower ratios may indicate the presence of protein, phenol or other contaminants. The 260/230 absorbance ratio is a secondary measure of nucleic acid purity which commonly is in the range of 1.8-2.2. Lower ratios may indicate the presence of co-purified contaminants. In contrast, the Agilent 2100 Bioanalyzer is a microfluidics-based automated electrophoresis used for sample quality. Similar to the NanoDrop system, 1µl samples are enough for nucleic acids. The system calculates the RNA Integrity Number (RIN), which is a numerical value between 1 and 10 calculated from the entire electrophoretic trace, 1 indicating a degraded profile and 10 indicating an intact sample. Such value was introduced to substitute the 18S to 28S ribosomal subunits ratio for determining the degradation level, in an attempt to standardize the process and remove any individual-dependent interpretation.

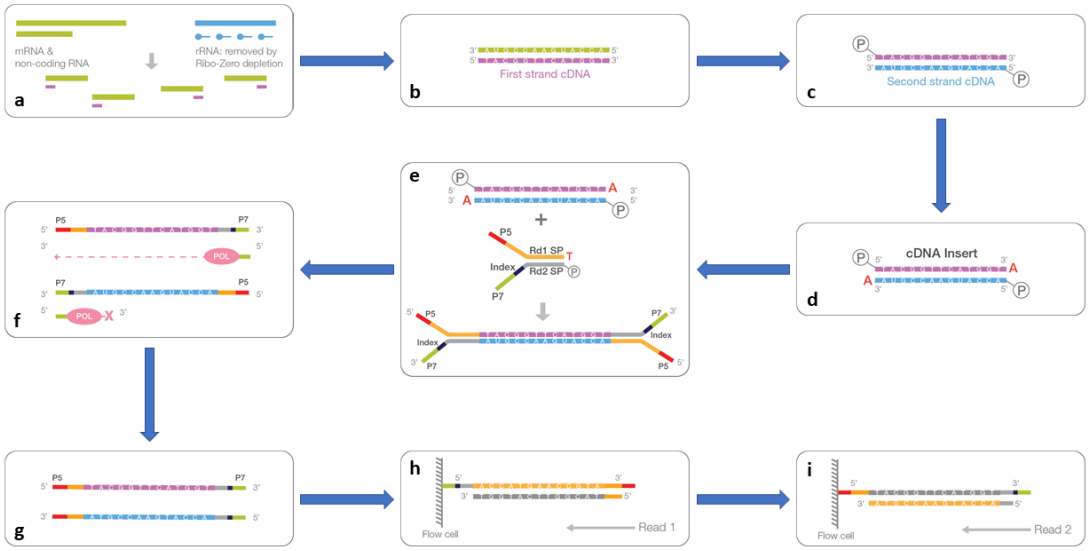


Figure 1.7: Illumina total RNA library preparation. **a)** Ribo-Zero depletion and RNA fragmentation. **b)** cDNA first strand synthesis. **c)** cDNA second strand synthesis. **d)** 3' end adenylation. **e)** Adaptor ligation (single-index adapters shown in image). **f)** DNA fragment enrichment. **g)** Final library. **h)** Cluster generation and read 1 sequencing. **i)** Paired-end turnaround and read 2 sequencing (figures adapted from the TruSeq Stranded Total RNA manual from the Illumina webpage, https://emea.illumina.com/).

Once good RNA quality samples are obtained, the samples need to be prepared for sequencing. The RNAs are converted into a cDNA library due to the improved chemical stability

of such molecules, making them more manageable for the sequencing protocols from each platform. As each sequencing platform has its own library protocols with varying steps, and the objective of this thesis is not to give a detailed description of each one (in case of interest, each protocol can be found in each company's website), only a brief description of the protocol for the sequencing of total RNA from the Illumina platform, which is the one used throughout this thesis, will be provided. In figure 1.7 a schematic workflow for the Illumina Total-RNA library preparation is shown.

Briefly, the Illumina TruSeq Stranded Total RNA library preparation includes the following major steps (the protocol is optimized for 0.1-1 µg of total RNA):

1. Ribosomal RNA (rRNA) is removed annealing total RNA to biotinylated and target-specific oligo magnetic beads. Depending of the kit used, cytoplasmic or/and mitochondrial rRNA in case of eukaryotic organism or cytoplasmic or/and chloroplast rRNA in case of plants can be removed. In case of blood samples, there is another kit that depletes goblin-encoding mRNAs in addition to the previous rRNA species.
2. Following purification, the RNA is fragmented into small pieces using divalent cations under elevated temperature. Other platforms use enzymes or heat alone for fragmentation.
3. RNA fragments are copied into first strand cDNA using reverse transcriptase and random hexamer primers.
4. Second strand cDNA is synthetized using DNA Polymerase I and RNase H. To achieve strand specificity, previous to the second strand synthesis, deoxythymidine triphosphate (dTTP) is replaced by deoxyuridine triphosphate (dUTP), which quenches the second strand during amplification.
5. A single 'A' nucleotide is added to the 3' ends of the double-stranded cDNA (ds cDNA) to prevent them from ligating to each other during the adapter ligation reaction.
6. Adapters are ligated to prepare the ds cDNA for hybridization onto a flow cell. The adapters can be indexed for each library reaction, allowing for pooling libraries later for sequencing.
7. The DNA fragments are enriched with polymerase chain reaction (PCR) and purified to achieve the final cDNA. The polymerase used in this step does not incorporate past dUTP, quenching the second strand effectively.
8. Once the library is prepared, it is sent to the sequencing facility for cluster generation and sequencing.

In addition, for small RNA sequencing, especially microRNAs (miRNAs), from the same good RNA quality samples, library preparation changes slightly. In figure 1.8 can be seen a schematic workflow of the Illumina TruSeq Small RNA library preparation. The protocol is optimized for 1 µg of total RNA. Briefly, the protocol use adapters that take advantage of a common structure of most miRNA molecules: mature miRNAs have a 5'-phosphate and a 3'-hydroxyl group. First, 3' and 5' adapters are ligated to the small RNAs. Then, reverse transcription followed by PCR is used to create cDNA. PCR is performed with two primers that anneal to the ends of the adapters. Finally, amplified cDNA is purified by a gel electrophoresis. From the gel, individual bands can be used for sequencing. The 147 nt band primarily contains mature miRNAs (~22 nt in size), while the 157 nt band contains piwi-interacting RNAs, some mature miRNAs and other regulatory small RNA molecules (~30 nt sized fragments).

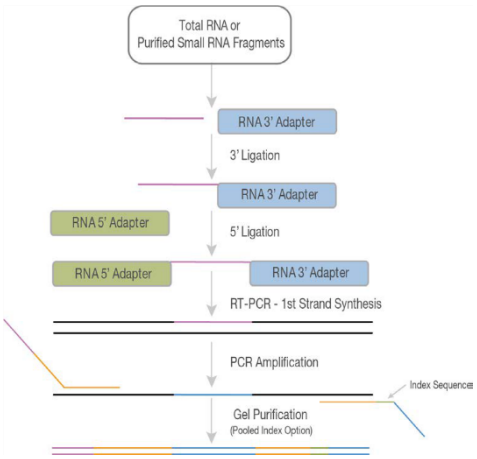


Figure 1.8: Illumina TruSeq Small RNA library preparation (figures adapted from the Illumina webpage, https://emea.illumina.com/).

1.3.2 RNA-seq platforms

There are multiple platforms with totally different approaches to produce RNA-seq sequences. Despite technologically different, such platforms rely on similar workflows for the generation and analysis of sequencing libraries. Generally, NGS technologies capture the light emitted when a correct base is incorporated into the sequencing reaction to the template being sequenced (84). Thus, the raw output from sequencing machines consist of image records of the light emitted by every single reaction. Those images are processed to extract numerical values for every base. Then, these values are used to determine the bases, ending with a list of sort sequences with base quality values. In table 1.4 can be seen some of the major RNA-seq platforms and their general properties.

Table 1.4: Major RNA-seq platforms and their general properties (table adapted from (83)).

Platform	Sequencing Chemistry	Detection Chemistry
Illumina	Sequencing by synthesis	Fluorescence
SOLiD	Sequencing by ligation	Fluorescence
Roche 454	Pyrosequencing	Luminescence
Ion Torrent	Sequencing by synthesis	Proton release
PacBio sequencing	Single-molecule, Real-Time (SMRT)	Real-time fluorescence
Oxford nanopore	Electrophoresis	Electrical current difference per nucleotide through a pore

As each sequencing platform has its own method, and the objective of this thesis is not to give a detailed description of each one, only a brief description for the sequencing from the Illumina platform, which is one of the most popular sequencing platforms and the one used throughout this thesis, will be provided. In figure 1.9 a schematic workflow for Illumina sequencing can be seen. Briefly, sequencing templates from the library are loaded and hybridized to the flow cell surface. Then, through solid-phase bridge amplification, clusters of up to 1,000 identical copies of each template are created. The reverse strands are cleaved and washed away. The generation of such clusters are necessary due to the signal emitted by the synthesis of a single deoxynucleotide triphosphate (dNTP) is not strong enough to be detected. Finally, reagents are added to the flow cell to execute sequencing by synthesis, a process in which a labelled nucleotide is imaged thanks to the fluorescence emitted in each cycle. The nucleotide label serves as a terminator for polymerization, so after each dNTP incorporation, the signal emitted is imaged and the label is cleaved to allow the incorporation of the next base. The reconstruction of the sequence of added nucleotides in a specific location on the flow cell

corresponds to nucleotide sequences of a template ds cDNA, which are recorded in a massive parallel manner. Each base has an assigned quality score. Compared to other platforms such as Roche 454 and SOLiD, Illumina sequencing is the cheapest option with the biggest output at the expense of accuracy (85).

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) was a next generation DNA sequencing platform commercialized by Life Technologies, who was later acquired by Thermo Fisher Scientific Corporation. This approach is based on ligation, it exploits the mismatch sensitivity of a DNA ligase enzyme to determine the underlying sequence of the target DNA molecule. The major shortcoming of the platform is sort read lengths, but it has one of the highest accuracies from all the platform previously listed (85).

Roche 454 was the first NGS method commercially available. Instead of detecting labelled dNTPs, this platform used pyrosequencing technology, in which chemiluminescence of pyrophosphate release during dNTP binding is detected. The main advantages of this platform is long read lengths (up to 700 bp) and the fast speed of the method, but it has a high cost due to the reagent prize, a high error rate in terms of poly-bases longer than 6 bp and low throughput (85).

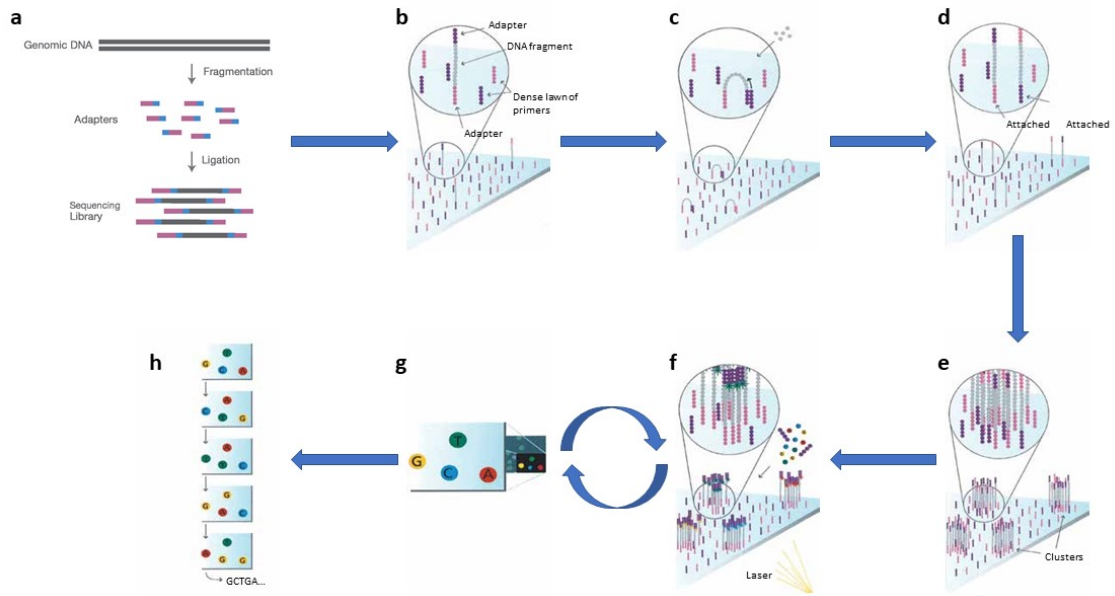


Figure 1.9: Illumina sequencing. **a)** Library preparation. **b)** The library is loaded into flow cell and hybridized to the surface. **c)** solid-phase bridge amplification by adding unlabelled nucleotides and enzyme, building double-stranded bridges on the solid-phase substrate. **d)** Double-stranded molecules are denatured, leaving single-stranded templates anchored to the substrate. **e)** The process is repeated until dense clusters of double-stranded DNA are generated. **f)** Four fluorescently labelled nucleotides, primers and DNA polymerase are added. **g)** After laser excitation, the fluorescence emitted from each cluster is captured. This last two steps are repeated until all the sequence from the template is captured. **h)** The emission wavelength and intensity are used to identify the bases in each fragment (figures adapted from the sequencing protocols from the Illumina webpage, https://emea.illumina.com/).

Ion Torrent is another next generation DNA sequencing platform from Thermo Fisher Scientific. It is based on the detection of hydrogen ions released during DNA polymerization. When a nucleotide is added to the template complementary strand, an ion-sensitive field-effector transistor (ISFET), a device used for measuring ion concentration in solution, detects the reaction due to a change in pH. The main advantages of the platform are its speed and low cost. When homopolymer repeats of the same nucleotide are present in the template, multiple nucleotides are inserted in the same reaction, resulting in a greater pH change and electronic

signal. Due to that, it is quite difficult for the platform to enumerate long repeats. It has been shown that Ion Torrent and Illumina platforms have a high degree of concordance at the gene-level read counts, detecting similar sets of differentially expressed genes (DEGs) and reaching similar conclusions at the pathway level (86).

PacBio sequencing or SMRT (Single-molecule, Real-Time) sequencing from Pacific BioSciences is a third-generation high throughput technology and it collects data from millions of wells using the DNA replication to sequence long fragments of DNA or RNA. Single-stranded DNA templates are circularized by ligation of hairpin adapters at both ends and the light pulses emitted by labelled nucleotides are detected when a polymerase bound to a hairpin adaptor adds nucleotides to the template. In comparison to any second-generation throughput technologies, PacBio offers longer reads (over 10 kb) and faster runs, but in return it has a lower throughput (0.5-1 billion bases), higher error rates (around 11%-15%) and a higher cost (87). The long reads make the technology ideal for identification and quantification of isoforms.

Oxford nanopore sequencing is another third-generation high throughput technology. When an ion current is passed through a nano-scaled hole, the device measures the change in current in real-time as the DNA molecule pass through or near the nanopore, using the information about the current change to identify the exact nucleotides in the template. In comparison to PacBio system, Oxford nanopore returns slightly longer reads and the quantity of reads per flow cell is higher, but in return the data of PacBio is of higher base quality (88).

1.3.3 RNA-seq applications

RNA-seq can be applied to answer a broad range of questions such as transcript quantification, differential gene expression profiling, study of alternative splicing events, discovery of new transcripts, annotation of non-coding molecules (such as miRNAs, lncRNAs and circRNAs), *de novo* genome assembly and identification of single nucleotide polymorphisms (SNPs). Depending of the interests of the study and the organism under study, different experimental designs would be applied. For example, if a reference genome is available for the organism of interest, the expression level can be quantified directly mapping the reads onto the genome. In contrast, if there is no reference genome available, the reads must be assembled into contigs previous to any quantification. In general, there is no consensus pipeline for each analysis type, being dozens of different algorithms to chose from for each step. There is no optimal tool for all RNA-seq data sets. In addition, RNA-seq data can be combined with other genome-wide data such as DNA sequencing, DNA methylation or ChIP-seq (Chromatin immunoprecipitation sequencing) to connect gene regulation with specific aspects of molecular physiology (89).

1.3.3.1 Differential expression

Differential expression analysis is the most common application for RNA-seq data. The expression values, which are based in the number of reads that map to the characteristic of interest (at the exon, transcript or gene level), are compared among samples in different conditions in an attempt to discover key elements in diseases, treatments or pathways. In this kind of experiment is recommended to have at least three biological samples of the same condition to be able to measure the biological variation between samples and differentiate it from the variation caused by the treatment. The more samples per condition, the more precise the mean expression levels, leading to a more accurate modelling of the data. Normally, there is no necessity for technical replicates at the sequencing level (repeated measurements of the

same biological sample used to test the variability of a protocol), as the reproducibility of RNA-seq is quite high.

In one RNA-seq study in Merino sheep, peripheral blood mononuclear cell (PBMC) expression levels at three different time points from sheep infected with the fasciola hepatica parasitic trematode were compared relative to uninfected controls (90). Over 100 million reads were obtained. Key genes involved with the immune response to the parasite were identified and the study showed that events related to the complement system, the chemokine signalling, T-cell activation and metabolic processes were altered by the presence of the trematode. This study is a clear example of how RNA-seq data can be used to study the temporal progression of a disease, treatment or condition (in this case the temporal progression of the host response to *Fasciola hepatica*).

In addition, RNA-seq data can be used to find elements that regulates different traits of interest, to select those traits in animal breeding, for example. In a study realized in Spanish Churra and Assaf breeds, two breeds with totally different milk production traits, milk somatic cells from ewes on four different time points were sequenced (91). A total of 1,116 million paired-end 75 nt reads were obtained. In the comparison between the two sheep breeds, 256 annotated differentially expressed genes were observed. Some of those genes were shown to be related to the endopeptidase and channel activity GO terms, finding some genes that may explain the higher cheese yield of the Churra breed.

1.3.3.2 *De novo* assembly

Before the introduction of NGS technology, very few well-studied organisms had their genome fully sequenced with an enriched annotation, being some of the protein-coding genes computationally predicted. RNA-seq allowed to different laboratories to build complete transcriptomes from non-model organisms and to identify new genes, being those new genes annotated by sequence similarity to other organisms with known gene functions. Transcriptomes for the 'Bouche de Bétizac' and 'Madonna' chestnuts (92), blueberry (93), four different species of common tropical crustose coralline algae (94), pieris rapae butterfly (95) and two related geese species (96) have been recently created *de novo*, to name just a few examples.

The first draft of the sheep (*Ovis aries*) reference genome was done by The International Sheep Genomics Consortium using whole-genome shotgun sequencing of a Texel ewe and Illumina paired-end sequence data from a Texel ram (97). In addition, data from seven different tissues in conjunction to existing ovine EST collections and other NGS datasets were used to predict genes and to annotate them in the de novo assembly. The final Oar v3.1 assembly had a contig N50 length of ~40 kb and a total assembled length of 2.61 Gb, with ~99% anchored to the 26 autosomes and the X chromosome (98).

1.3.3.3 Alternative splicing

Splicing patterns of a gene are not fixed, from one gene multiple mRNAs can be produced. Thus, alternative splicing is a post-transcriptional process that increases the proteome diversity, particularly in mammals. There are seven distinct alternative splicing events in total: exon skipping, intron retention, alternative 3' splice site, alternative 5' splice site, alternative first exon, alternative last exon and mutually exclusive exons (99). Alternative splicing has been studied with reverse transcription polymerase chain reaction (RT-PCR), sequencing of expressed sequence tags (ESTs) or designed microarrays, but such technologies have low-throughput, high

noise or are limited to known splicing patterns (100). RNA-seq has improved the capacity to study splicing patterns, although it is still complex to quantify each alternatively spliced transcript expression, since it is quite complex to assign short reads to an exon of a transcript when said exon is part of multiple transcripts. Such drawback of RNA-seq technology has been partially solved with third-generation sequencing technologies such as PacBio and Oxford Nanopore, as it is possible to sequence complete mRNA molecules in one reaction, but at the cost of a higher error rate.

In a study analysing muscle transcriptomes of Dorper and Small-Tailed Han sheep (99), multiple alternative splicing events were found. Approximately 26% of the reference genes underwent some splicing event, being alternative 3'/5' splice sites the most common, while intron retention was the least frequent. In addition, they were able to identify splicing event exclusive to each species. In a more recent study, liver tissues of Mongolian and Lanzhou fat-tailed sheep were analysed (101). Higher rates of alternative splicing were found in the samples (from 30.54% to 38.33%) and in contrast to the previous study, the most common splicing event was intron retention. Such differences may be explained by breed differences or may be indicative of the lack of information on alternative splicing in sheep, in any case more studies should be conducted.

1.3.3.4 Variant discovery

RNA-seq data allows for detection of transcriptome variants such as SNPs and short insertions or deletions (INDELs) at a large scale. Some of those variants can be related to different phenotypes of interest or diseases. It must be noted that RNA-seq data is though not to be an ideal source for SNP detection due to high false positive rates caused by the complexity of alignment due to RNA splicing, sequencing errors, random errors introduced during RT-PCR or RNA editing (102). It has been shown that there are large differences (up to 10%) between genotypes inferred from DNA and RNA sequencing. With RNA-seq data alone, it is complex to distinguish between a true SNP or an RNA editing process.

In a recent study from high-throughput RNA-seq data, tissue samples from longissimus dorsi muscle, perinephric fat and tail fat in three different sheep breeds (Lanzhou fat-tail sheep, small-tail Han sheep and Tibetan sheep) were compared in an attempt to find genetic variations related to the fat-tail phenotype (103). They reported 33 SNPs distributed across a chromosome 3 region previously related to fat deposition in tails of sheep, in addition to three genes (*CREB1*, *WDR92* and *ETAA1*) that may be associated with fat tail development. In other study, RNA-seq data from milk somatic cells from Churra and Assaf sheep were analysed to identify genetic variations related to milk yield (104). From 216,637 detected variants (SNPs and INDELs), 57,795 were detected in regions harbouring Quantitative Trait Loci (QTL) for milk yield, protein percentage and fat percentage. In addition, 20 mutations having great effects on principal milk proteins and lipid metabolism proteins were found.

1.3.3.5 Long non-coding RNAs (lncRNAs)

Despite lncRNAs were known before the introduction of NGS technology, it was the use of RNA-seq that allowed to see how extensively expressed were in different tissues. lncRNAs can be described as sequences longer than 200 nucleotides (transcribed from both protein-coding and non-coding DNA regions) without open reading frames and 3'-untranslated regions (3'-UTRs) (105). It has been shown that these molecules are able to interact with proteins, DNA or other

RNA molecules, having principal roles in multiple pathways such as gene regulation, cell differentiation and chromatin remodelling. RNA-seq, with an adequate library preparation, allows identification and quantification of such molecules.

In a recent study, Alpine merino sheep skin samples were sequenced on a Illumina machine to integrate mRNA and lncRNA information (106). A total of 884 lncRNAs were identified. From those lncRNA, they were able to see that lncRNAs were shorter than mRNAs in length distribution, that lncRNAs were composed of less exons and that sheep lncRNAs were longer in comparison to the ones detected in human and mouse. In a more recent study, Subo Merino sheep (a Chinese breed) skin samples were sequenced to study the involvement of lncRNAs in the development of hair follicles (107). From 10,193 (1,540 known and 8,653 novel) identified lncRNAs, 471 were differentially expressed. In agreement with the previous study, they showed that lncRNA were shorter in comparison to mRNAs, have a lower expression and are composed of less exons (the majority of two exons). In addition, among the differentially expressed mRNAs there were predicted targets of lncRNAs with functions in hair follicle development and morphogenesis, epidermis development and morphogenesis and cell differentiation and migration.

Furthermore, RNA sequencing technology can be used for the annotation of other non-coding molecules such as microRNAs (miRNAs) and circular RNAs (circRNAs). Those applications will be explained later in sections specific for them.

1.3.4 Biases in RNA-seq

There is a broad range of platforms for RNA-seq, each with its own benefits and drawbacks. The objective of this thesis is not to give a detailed list for each one, only those biases related to all platforms and the ones related exclusively to Illumina sequencing (the one used throughout this thesis) are going to be explained in more detail. Illumina sequencing has been shown to be highly replicable and with fairly little technical variation (82). Despite such advantages, it has been shown that the technology suffers from some biases introduced during library preparation and others related to the sequencing itself.

One of the most important characteristics of NGS data that must be taken into account is its high error rate. For the Illumina HiSeq2000 and MiSeq sequencing machines, an error rate of ~0.1% per base has been reported, being the most common error single nucleotide substitutions (108). As previously noted, Illumina uses labelled dNTPs to detect inserted nucleotides to the template by fluorescence. A and C nucleotides can be detected upon a light excitation produced by a red laser, while G and T nucleotides needs light excitation through a green laser. Due to similar emission spectra, C to A and G to T are the most common substitutions in Illumina sequencing (109). Other mechanisms that cause incorrect base calling in the sequencing are the ones known as post-phasing and pre-phasing. Pre-phasing can occur due to inadequate flushing of the flow cell, resulting in non-incorporated nucleotides remaining after a cycle, which can lead to incorporation of more than one nucleotide in some sequences from the cluster in the following cycles. In contrast, post-phasing can occur due to an incomplete removal of the terminator in a cycle, which results in a lag in the synthesis of some sequences in a cluster (109,110). Both processes cause an incorrect elongation of the sequence, as a result the fluorescence signal from the cluster has interferences. This can explain why at the end of the reads the quality of base-calling usually drops in Illumina sequencing, since the longer the read the weaker the signal from the cluster due to accumulation of such events (110). Such sequencing errors are particularly relevant for SNP discovery, as it is complex to differ a

sequencing error from a true SNP (84). This can be avoided increasing the sequencing depth to reach a level where a sequence is sequenced multiple times, which can help to correct for sequencing errors.

Previous to the synthesis of cDNA during library preparation, RNA molecules are fragmented into smaller pieces. Depending of the platform, the fragmentation can be done at the RNA level through hydrolysis or nebulization or at the DNA level through DNase I treatment or sonication, for example. Illumina prepares their libraries through RNA fragmentation with divalent cations under elevated temperatures. Such fragmentation causes a little bias in the transcript coverage and it has been shown that there is a lower coverage of both 3'/5' transcript ends, while in cDNA fragmentation methods there is a strong bias towards identification of the 3' end (111). Due to a lower coverage of the transcript ends, it is usually complex to clearly define the 3'/5' ends of transcripts through Illumina sequencing, something that can be partially solved with an adequate sequencing depth. cDNA synthesis is another step in which bias is introduced at the library preparation. Illumina uses random hexamer priming to generate complementary reads across the full transcript. It has been shown that such priming leads to a not totally uniform coverage of the transcript due to a bias in the nucleotide composition at the start of the first strand cDNA (112,113). This bias can be seen as a strong pattern in nucleotide frequencies in the first 13 positions at the 5' end. Despite the bias introduced by random hexamer priming, it is preferred to other methods such as oligo(dT) priming, which are highly biased towards the 3' ends (112).

Library preparation procedures usually employs PCR amplification before sequencing. This step has been shown to be the major cause of uneven read coverage in regions with enriched GC or AT content (114,115), being such regions underrepresented when sequencing. It has been found that Illumina sequencing has a dependence between GC content of the full DNA (not only the sequenced read) and read count and that such bias in not consistent between samples (116), making it hard to correct for. Moreover, since GC content is sometimes correlated with functionality, it is complex to differ between GC bias and the true signal. Library preparations with optimized PCR protocol or with a PCR-free protocol have been developed in an attempt to correct for GC bias (114). It must be taken into account that PCR-free methods require large amounts of starting RNA material, which makes them inadequate when dealing with small amounts of RNA from clinical isolates. In contrast, other studies have tried to deal with the GC content bias at the data analysis step, developing normalization strategies specific for GC bias correction (117). The RNA selection step (rRNA removal or polyA selection) has been pointed as another source of variability in the coverage of transcripts (118). It has been shown that a combination of a TRIzol RNA extraction and RiboZero RNA-seq protocol (based on rRNA removal) produce a significant increase of intronic sequences in comparison to other RNA-seq protocols, being the origin of these intronic sequences pre-mRNA or splicing by-products (119).

Other characteristics of transcripts that seem to influence expression quantification of RNA-seq data are the length and real expression levels. Due the inherent nature of RNA-seq data, which is short reads from the full transcript formed at the fragmentation step, the counts for a transcript will be proportional to the real expression level and its length, since longer transcripts originate more fragments and this results in more sequencing reads (120). Normalization methods that divide the obtained counts with transcript length have been proposed as a countermeasure for this kind of bias, but such normalization methods only mitigate it. In addition, the ability of RNA-seq data to detect rare transcripts is influenced by the sequencing depth, as RNA-seq is biased towards highly expressed genes which concentrate the vast majority of sequencing reads (121). The probability to sequence lowly expressed transcripts increase with greater sequencing depths, but more reads also means noisier data that results in

a more difficult differential expression analysis. For transcriptomes of similar size to the human, it has been estimated that approximately 10 million reads of 35 nt in length (~10x coverage) would be required for quantification of 80% of expressed genes, while approximately 700 million reads would be needed for quantification of more than 95% of the expressed transcripts (122).

Despite not being a bias directly related to library preparation or sequencing process, it has been shown that RNA degradation, which occurs in most of isolated RNA samples with a varying degree that depends on sample collection and storage conditions, impairs accurate quantification of transcripts (123). It has been pointed that RIN values used for RNA degradation assessment has several issues, since its calculation relies heavily in 18S and 28S ribosome RNA properties and as a result it is not a direct measure of mRNA integrity. In addition, it has to be taken into account that RNA decay rate is transcript-specific. Different metrics had been proposed in an attempt to improve the quantification of RNA integrity, such as transcript integrity number (TIN) (123) or DV_{200} metric proposed by Illumina (www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/evaluating-rna-quality-from-ffpe-samples-technical-note-470-2014-001.pdf). Illumina laboratories has pointed out that mean RNA fragment size is a more reliable metric for quantification of RNA quality for the TruSeq library preparation kit and they defined the DV_{200} metric, which is the percentage of RNA fragments > 200 nucleotides. In contrast, TIN is a metric that measures the percentage of the transcript that has uniform coverage and it has been demonstrated to be quite useful to adjust gene expression and improve differential expression analysis.

Until now library preparation and sequencing biases of total RNA have been addressed. The preparation of libraries for miRNA-seq has slight variations, which results in other kind of bias in the data. In those library preparations, the adapter sequence ligation at both ends of small RNAs is a critical step and it has been shown that differences in adapter ligation efficiency among protocols result in drastically altered expression patterns of individual miRNAs (124). In addition, the formation of adapter dimers lead to detection of reads without miRNA insert, reducing the number of informative reads.

1.4 Non-coding RNAs (ncRNAs)

Approximately ~21,000 protein-coding genes have been described in human, but these protein-coding regions only encompass ~1.5% of the human genome (125). With the exception of some well-known ncRNAs with specific roles in translation of protein-coding RNAs such as rRNA or transfer RNA (tRNA), the rest of the DNA sequence was thought to represent noise and was referred as junk DNA. With improvements in technology and additional research, this notion has changed markedly in recent years and a wide variety of different ncRNAs with distinct functional roles have been described, seeing that the majority of the human genome (up to 90%) is transcribed into RNAs (126). The emergence of NGS technology has allowed the discovery of thousands of ncRNAs with important roles in gene expression, such as miRNAs and lncRNAs. It has been shown that such ncRNA elements are not only rare transcripts and, despite generally having a lower expression level than protein-coding molecules, ncRNAs are expressed more broadly than initially thought.

In spite of the fact that thousands of ncRNAs are being described, further research is needed for their functional annotation. ncRNAs can be classified into two main categories with multiple subcategories inside of them: small non-coding RNAs and long non-coding RNAs. The structure and function and how they interact with other molecules is best known for small ncRNAs. In contrast, lncRNAs, which are usually defined as non-coding transcripts longer that

200 nucleotides, not have a clearly defined structure and functional role. New lncRNAs are constantly defined and new categories and functions are published annually. In table 1.5 can be seen a brief list of different ncRNA classes with some of their properties. This thesis does not intend to give a detailed description of all ncRNAs, only miRNAs and circRNAs are going to be described in more detail, since they are the two main ncRNA classes that are going to be annotated and analysed in sheep with the data presented in this thesis.

Table 1.5: Classes of non-coding RNAs and their approximate size and functions (human as reference) (table adapted from (83,126–130)).

Class	Size	Function
microRNA (miRNA)	21-23 nt	RNA that, in complex with AGO protein, binds to mRNA to induce deadenylation and decay or translational regulation.
Long non-coding RNA (lncRNA)	>200 nt	Long elements transcribed by RNA polymerase II, occasionally capped and polyadenylated. Roles in post-transcriptional regulation of mRNA and cis regulation.
PIWI-associated RNA (piRNA)	25-33 nt	Involved in the epigenetic and post-transcriptional silencing of transposons and other repeat-derived transcripts.
Ribosomal RNA (rRNA)	120, 160, 1,868, 5,025 nt	RNA compound of the small and large ribosomal subunit.
Small conjugation-specific RNA (scan RNA/scnRNA)	28 nt	Double-stranded RNAs processed by a Dicer-related RNase that recognize genomic internal eliminated sequences in the developing macronucleous of ciliates and target them for destruction.
Small nuclear RNA (snRNA)	100-300 nt	Molecules found within the splicing speckles and Cajal bodies. Participates in the splicing of pre-mRNA.
Small nucleolar RNA (snoRNA)	60-300 nt	Participate in chemical modification of other RNAs, mainly rRNA, tRNA and snRNA.
Small Cajal body-associated RNA (scaRNA)	200-300 nt	A class of snoRNAs localised in Cajal bodies. Guide chemical modifications of RNA polymerase II transcribed spliceosomal RNAs (U1, U2, U4, U5 and U12).
Small interfering RNA (siRNA)	20-25 nt	Double-stranded RNA molecules that interfere with complementary target RNA by degrading mRNA after transcription.
Tranfer RNA (tRNA)	70-90 nt	Helps in the translation of mRNA into protein connecting an mRNA codon with its corresponding amino acid.
Enhancer RNA (eRNA)	50-2000 nt	Transcribed from DNA enhancer regions, play a role in transcriptional regulation in *cis* and *trans*. Highly correlated proximal gene expression.
microRNA-offset RNA (moRs/moRNA)	~23 nt	Derived from miRNA precursor sequences. Function still unknown.
Promoter-associated small RNA (PASR)	20-200 nt	Derived from promoter regions. May be involved in chromatin modifications within promoter regions, thus modulating their host gene.
Terminus-associated small RNA (TASR)	22-200 nt	May play a role in increasing the copy number of the transcripts of their host genes.
Circular RNAs (circRNAs)	1-5 exons	Circularized transcripts usually formed from coding segments. Play roles as miRNA sponges, form RNA-protein complexes and may be translated into proteins.
Natural antisense transcript (NAT)	-	Derived from both DNA strands at the same locus but in opposite direction from the gene. Precise role unknown, suspected to play a role in gene expression regulation.

1.4.1 microRNAs (miRNAs)

MicroRNAs (miRNAs) are a class of short endogenous ~22 nt non-coding RNAs that play key roles in gene regulation through translational repression or mRNA decay. The biogenesis of miRNAs is tightly controlled and their dysregulation has been associated with different diseases. The first miRNA to be discovered was *lin-4* in C. elegans in 1993 (131). In this first study, it was shown that *lin-4* did not encode any protein and that two transcripts of different length, one of 22 nt (later known as the mature miRNA form) and other of 61 nt (later known as the pre-miRNA form), were being expressed. In addition, the conservation of this molecule in three other Caenorhabditis species and a complementary sequence element repeated seven times in the *lin-14* 3'UTR were found. It was proposed that *lin-4* acted by negatively regulating the expression levels of *lin-14* protein. After this initial report of a small non-coding molecule, it was needed some years until a new molecule resembling in structure to *lin-4* to appear. In 2000, a 21-nucleotide RNA which regulates developmental timing known as *let-7* was reported in C. elegans (132). In this study, it was shown that *let-7* has complementary sequences to elements in the 3'UTRs of genes *lin-14*, *lin-28*, *lin-41*, *lin-41* and *daf-12*, suggesting a regulatory role of *let-7* on these genes. These short RNA sequences gained more attention from the scientific community and an increasing number of miRNAs were reported shortly after. In a popular miRNA database such as miRbase (Release 22), there are 1,917 entries for human miRNAs and 39,417 in total for all annotated species. A substantial number of those annotations has a dubious origin, probably being false positives.

1.4.1.1 miRNA-seq

Long ago it was though that DNA fragments that did not encode any protein was "junk", with no known purpose. With the introduction of NGS technology, it was shown that the majority of the DNA was transcribed but not translated into proteins and that those non-coding RNAs had regulatory roles in multiple pathways. Among those non-coding RNAs, small molecules such as miRNAs, small interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs) have been studied in great detail. There are special library preparations for small non-coding RNA enrichment for subsequent sequencing, using the data to identify the sequence, structure, abundance and function of such molecules. Studies using these data have been focusing mainly on the detection and functional annotation of miRNAs, although it can be used for the analysis of other small non-coding RNAs such as small nuclear/nucleolar RNAs (snRNAs) or endogenous siRNAs. Despite miRNA-seq analyses being nowadays common, being various studies in sheep, in a miRNA database such as miRbase [Release 22.1] there is only 153 mature miRNAs annotated for sheep.

In one early miRNA-seq study in sheep, longissimus dorsi muscle of Texel and Ujumqin embryos and lambs through eight different time points were sequenced in a Solexa machine, covering all representative periods of growth and development during gestation (133). After comparing to other mammalian mature miRNAs and the sheep genome, 2,914 mature miRNAs representing 2,319 unique miRNAs were predicted. Among the main characteristics of those predicted sheep miRNAs, it was pointed that:

- Few miRNAs account for nearly all the expression data. In their case, the 21 most expressed miRNAs accounted for 85.06% of the expression data.

- Highly expressed miRNAs are more stable, while lowly expressed ones are more easily influenced by development stages.

- Almost all of the sequenced miRNAs were able to produce different isomiRs (miRNA sequences showing variations respect to the reference sequence).

1.4.1.2 miRNA biogenesis

The miRNA biogenesis process in animals usually starts with transcription by RNA polymerase II, although there are some exceptional cases transcribed by RNA polymerase III (83). Most of the transcribed primary miRNAs (pri-miRNAs), which may be up to 1kb long, are usually intragenic and processed from introns and a few exons of protein coding genes, while the remaining ones originate from intergenic regions (134). It has to been pointed out that in some cases clusters of miRNAs in close proximity to each other forms polycistronic transcriptional units (135). All the miRNAs in such units are usually co-transcribed together and are individually regulated post-transcriptionally. Typically, a pri-miRNA is formed by a stem loop of 33-35 nt, a terminal loop and single stranded RNA segments at both ends. After transcription, the pri-miRNA undergoes maturation in the cell nucleus prior to transportation into the cytoplasm.

Following pri-miRNA formation, it matures into precursor miRNA (pre-miRNA) with the help of the microprocessor complex, which consist in the RNA binding protein DiGeorge Syndrome Critical Region 8 (*DGCR8*) and the ribonuclease III enzyme *Drosha*. *DGCR8* is able to recognize some N6-methyladenilated motifs in the pri-miRNA, while *Drosha* cleaves the pri-miRNA at the base of the hairpin structure, resulting a pre-miRNA (~70 nt) with a 2 nt long 3' overhang (134). Then, the pre-miRNA is exported to the cytoplasm, where the maturation process finish. This last step is done with the help of the complex formed by exportin 5 (*EXP5*) and *RanGTP*. Once in the cytoplasm, the pre-miRNA is further processed by the RNAase III endonuclease *Dicer*. In this step, *Dicer* cleaves the terminal loop, yielding a ~21 nt miRNA duplex with protruding 2-nucleotide 3' ends (136). At first it was thought that only one strand of the duplex was selected for further processing, while the other was degraded. It has been shown that both strands from the duplex can be loaded into Argonaute proteins (*AGO1-4* in humans) to act as mature miRNAs (134), forming what is known as miRNA-induced silencing complex (miRISC).

In the case of plants, miRNA biogenesis varies slightly. The maturation process is executed completely in the nucleus and the pri-miRNA processing instead of being done by the complex formed by *Drosha* and *DGCR8* (there are no homologous in plants), it is executed by DICER-LIKE 1 (*DCL1*), *DAWDLE,* the zinc-finger protein SERRATE (*SE*) and the double stranded RNA-binding protein Hyponastic Leaves 1 (*HYL1*) (135). Them, the pre-miRNA or mature miRNA is transported to the cytoplasm by HASTY (*HST*) and loaded into Argonaute proteins (usually AGO1).

Apart from the canonical miRNA biogenesis pathway, multiple non-canonical pathways (see figure 1.10) has been described. This non-canonical biogenesis is grouped into Drosha/DGCR8-independent and Dicer-independent pathways (134,135).

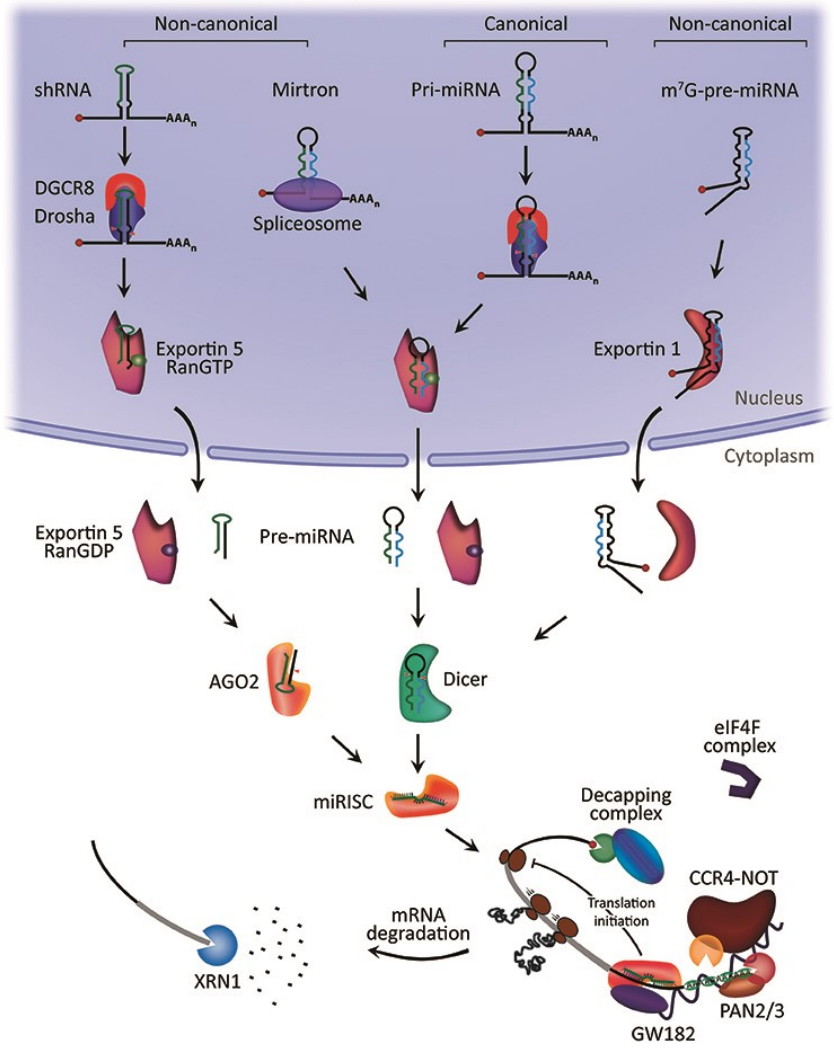


Figure 1.10: Animal miRNA biogenesis. The majority of miRNAs are transcribed from DNA sequences into primary miRNAs and then are processed into precursor miRNAs and finally into mature miRNAs. Depending the molecules assisting the process, different pathways has been described (figure adapted from (134)).

1.4.1.3 Mechanism of action

Once the miRISC complex is formed, miRNAs usually bind by sequence complementarity to 3' UTR sequence region of genes (although it has been reported cases binding to the coding regions, 5' UTR or promoter regions of a gene) and can induce mRNA translational repression or mRNA decay. Furthermore, it has been reported that in some especial conditions miRNAs or components of the miRISC complex may be able to act as translational activators (136). In plants, miRNAs binds nearly with perfect complementarity to mRNAs, while in metazoan bind with imperfect base-pairing following some rules (136) (see figure 1.11 for examples of miRNA- mRNA pairing):

- miRNA nucleotides 2 to 8, usually named the "seed" region, are essential for complementarity with the mRNA. This region usually binds with consecutive Watson-Crick base-pairing, although imperfect complementarity may be allowed if other regions bind compensating any mismatch in the seed. An A nucleotide at position 1 or an A or U

at position 9 of the mRNA improve the efficiency, despite not being directly paired to the miRNA.

- Bulges or mismatches must be present at the central region of the miRNA-mRNA duplex.

- There may be some complementarity to the miRNA 3' half (particularly nucleotides 13-16). This base binding can compensate for a seed mismatch

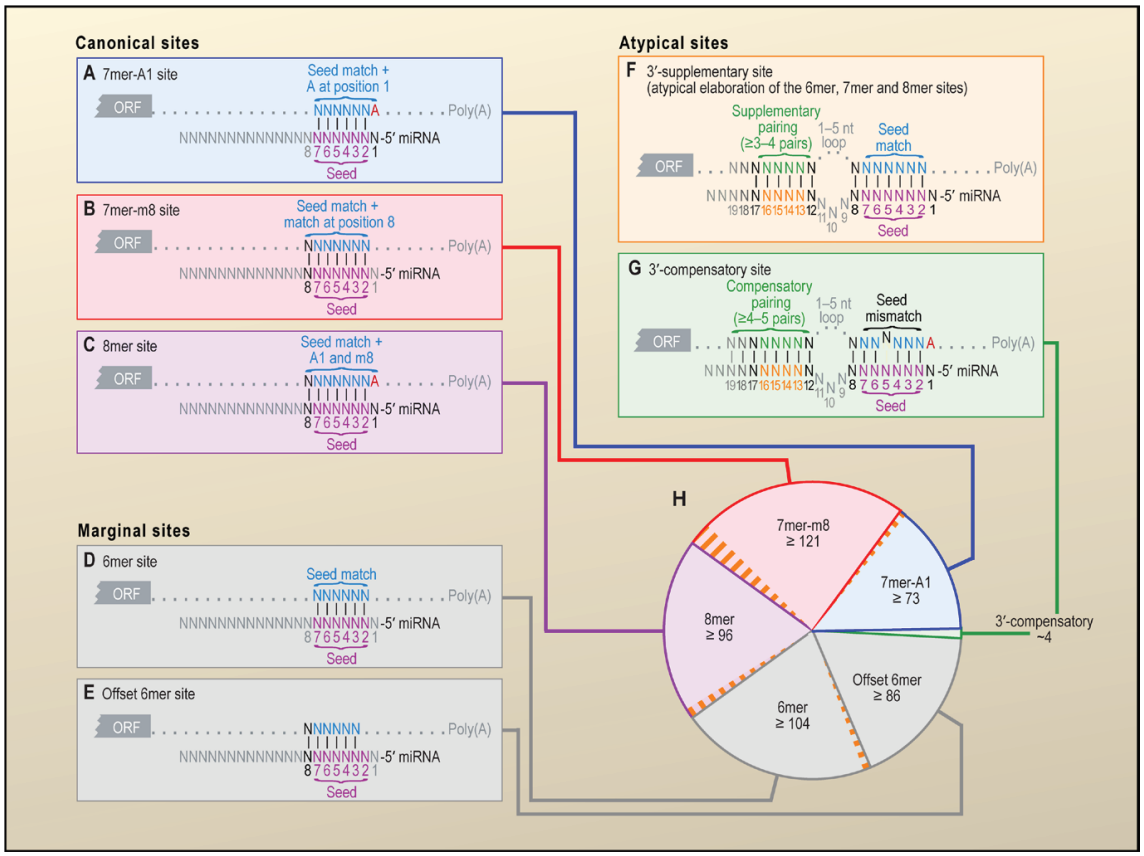


Figure 1.11: Types of miRNA-mRNA pairing. Vertical dashes indicate contiguous Watson-Crick pairing. **(A-C)** Canonical seed pairing. **(D-E)** 6 nt seed pairing. These 6mer sites have reduced efficacy and some target prediction algorithms discard them. **(F-G)** Imperfect seed pairing and a 3'-compensatory site (usually miRNA nucleotides 13-16) (figure adapted from (137)).

Multiple algorithms to predict miRNA targets has been made, each one using different characteristics. In addition to the previously defined bindings, it has been shown that multiple binding sites for the same miRNA in a 3'UTR can strengthen the regulation (138). Furthermore, it seems that many 6-8mer seed matches are evolutionary conserved, mainly 6mer (match of 6 nt from the seed), 7mer-m8 (seed binding supplemented with an additional nucleotide pairing at position 8), 7mer-A1 (seed binding and an A at position 1) and 8mer sites (seed binding and supplemented by both m8 and A1) (139).

The degree of complementarity between the miRNA and mRNA would determine whether there is an AGO2-dependent slicing of the mRNA (full complementarity) or translational inhibition and target mRNA decay (134). Most miRNA-mRNA interactions are partial, which results in the prevention of AGO2 endonuclease activity. In addition, it has to be stressed that the binding of a single miRNA is unlikely to produce a significant effect in the mRNA expression pattern (136). The formation of the silencing miRISC complex starts with recruitment of the *GW182* family of proteins, which recruits other effector proteins such as *PAN2-PAN3* and *CCR4-*

NOT (see figure 1.10). Then, after miRNA-mRNA interaction, the mRNA is poly(A)-deadenylated with the collaboration of *PAN2/3* and *CCR4-NOT*. Finally, the mRNA is decapped thanks to decapping protein 2 (*DCP2*) and degraded due to exoribonuclease 1 (*XRN1*) activity. The degradation or its final steps is thought to occur in P-bodies (cellular structures enriched in mRNA-catabolizing enzymes and translational repressors) (136).

In addition to translational repression and mRNA decay, there have been multiple reports indicating that miRNAs can up-regulate translation under special conditions. It has been shown that miR369-3 and let-7 can activate translation under growth-arrest conditions and it has been proposed that miRNAs can oscillate between repression and activation depending of the cell cycle (140). Furthermore, it seems that miR-10a is able to interact with the 5' UTR of ribosome protein-coding mRNAs to enhance ribosomal biogenesis and miR-328 indirectly up-regulates the transcription factor CEBPA by binding to PCBP2 (141). It still needs to be addressed if this miRNA-related activation of protein translation is a general phenomenon or it is just some exception of the usual mechanism of action.

1.4.1.4 Databases and nomenclature

Over the last decade thousands of miRNAs have been reported in multiple species. As previously noted, most of the annotated miRNAs in public databases such as miRbase are not robustly supported, with many false positives. At first, the nomenclature used for annotation of new miRNAs was pretty chaotic and it passed some time until a uniform system for annotation of miRNAs was presented. The minimum requirements for miRNA annotation and some of the nomenclature used in databases such as miRbase are revised in (142). For miRNA annotation the following would be required:

- An RNA stem consisting of two 20-26 nt long reads (with a median length of 22-23 nt) with 2 nt offsets between both arms need to be expressed.
- At least a section of 16 nt showing complementarity between both arms.
- 5' end homogeneity (most of the reported reads start with the same nucleotide).
- Both arms must be separated by a loop sequence of at least 8 nt and a maximum of 40 nt in species with a single *Dicer* protein, with no maximum requirement is species with multiple *Dicer* proteins.

As miRbase (http://www.mirbase.org/) is one of the most used miRNA databases, its nomenclature system will be explained briefly. One example of a miRbase miRNA would be oar-mir-374b. The first three letters represent the organism ("oar" being ovis aries and "hsa" homo sapiens, for example). Then, the following three letters would be mir or miR and it represents the miRNA gene (with the stem-loop) or the mature miRNA sequence respectively. At the end a numerical value is given. The numbering is given in sequential order, therefore, assuming that the last mouse miRNA was mir-352, the next new discovered would be mir-353. In case of a lettered suffix, it represents closely related mature sequences (for example oar-miR-374a and oar-miR-374b). In addition, if different genomic loci express identical mature sequence, each locus is represented by a second numerical value of the form hsa-miR-121-1 and hsa-miR-121-2. Finally, as each miRNA gene can generate two distinct mature sequences, the mature miRNA

identifier will have “-5p” or “-3p” to indicate the arm of origin (miR-142-5p and miR-142-3p, for example). It has to be pointed out that if the mature sequence of a new discovered miRNA is identical to a previously annotated miRNA in the database, it is suggested to name the new miRNA with the same identifier. In the database there are some exceptions to these rules, such as let-7 and lin-4, and these names have been retained for historical reasons.

1.4.2 Circular RNAs (circRNAs)

Circular RNAs (circRNAs) are a recent class of covalently closed circular single-stranded non-coding RNAs, formed when a downstream 5’ splice donor and upstream 3’ splice acceptor from a linear RNA are linked together, a process also called backsplicing (143). circRNAs does not have terminal structures such as 5’ cap or 3’ end poly(A) tail (144). Due to their circular structure, circRNAs are more stable, resistant to RNAse R and have longer half-lives than linear RNAs (145), making them good candidates for disease biomarkers. Despite being discovered long ago, they were thought to be low abundance products derived from splicing errors (146). The first circular molecules (viroids) were discovered by electron microscopy in 1976 (147). Despite being other reports of circular RNAs without functional potential, it was not until 1991 that the first endogenous circRNA originating from the *DCC* tumour suppressor was reported in humans (148). Shortly after, in 1993, it was shown that *Sry* gene was circularized in mouse testis samples, being these molecules one of the most abundant transcripts (149). In addition, the presence of long inverted repeats flanking the mouse *Sry* gene were demonstrated to be necessary for the formation of the circular transcript (150). It not was until recently, with the recent increase in total RNA-seq-based studies, that it was shown that circRNAs were more common than initially thought and that some of them had important roles in different pathways. One of the first studies that reported the circularization of hundreds of transcripts by RNA-seq data was done in 2012 (151). Since then, the circRNAs has gained the attention of the scientific community and multiple studies in different species have been reported. Despite the recent increase in circRNA studies, their functional role remains under constant debate and only a few circRNAs are well characterized.

1.4.2.1 Total RNA-seq

circRNAs are a more recently discovered class of non-coding RNAs, characterized by their closed circular form which gives them more stability and longer half-lives in comparison to other RNA classes. Their functions are not totally understood, but some circRNAs has been shown to act as miRNA sponges (e.g., the circRNAs related to CDR1-AS and SRY sequester miR-138 and miR-7, respectively) (152) , to have coding ability (e.g., circ-ZNF609) (153), although it remains to be probed actual translation into protein *in vivo*, or to have protein-binding activity (e.g., the circ-FOXO3 forms a ternary complex with p21 and CDK2) (154). Total RNA libraries with rRNA depletion allow for their detection through RNA-seq.

Being a new discovered molecule (or at least they have gained recently the attention of the scientific community due to their probed regulatory roles in gene expression), there are few studies of circRNA identification on sheep. Different studies taking pituitary gland samples (155,156) and longissimus dorsi muscle samples (157,158) from Kazakh sheep have identified multiple circRNAs. In the studies focused on the pituitary gland have been found numerous circRNAs interacting with miRNAs related to development and endocrine functions of the pituitary and other circRNAs enriched in neuromodulation pathways, such as dopaminergic

synapse and glutamatergic synapse. In contrast, in the studies focused on longissimus dorsi muscle, circRNAs involved in growth and development of muscle related signals such as positive regulation of myoblast differentiation, muscle fiber development and muscle organ development were identified. Such different functions of circRNAs in the tissues goes in agreement to what have been shown in human and mouse, that circRNA expression is tissue-dependent.

1.4.2.2 circRNA biogenesis

circRNAs are formed by circularization of linear transcripts, bringing a downstream 5' splice donor and upstream 3' splice acceptor close to each other. Depending the composition of circRNAs, they can be divided in three main categories: exonic circRNAs (ecRNAs), intronic circRNAs (ciRNAs) and exon-intron circRNAs (EIciRNAs) (144). Although it must be taken into account that in archaea, such as haloferax volcanii, splicing of some tRNA has resulted in intron circularization (159), the circularization process related to the spliceosome machinery, which is the most common one, will be explained in more detail in this thesis. Two main models have been proposed for circularization of transcripts: intron-pairing-driven circularization (see figure 1.12b) and lariat-driven circularization (for intronic circRNA see figure 1.12a and for circular product derived from exon skipping see figure 1.12c).

In the intron-pairing-driven circularization model, the splicing donor and acceptor are put under close proximity thanks to reverse complementary sequences in the flanking introns. It has been shown that short inverted repeats of 30-40 nt (e.g., the *Alu* elements in primates) are sufficient for circularization (160). Up to 90% of predicted circRNAs in human have reverse complementary sequences in their flanking introns (161). In addition, there is an alternative mechanism by which the circularized ends are put under close proximity. There have been different reports of RNA-binding proteins recruited to flanking introns that dimerize, thereby putting each end in close proximity (160,161). Specifically, the regulators of linear splicing Quaking (*QKI*), Muscleblind (*MBL*) and Fused-in-sarcoma (*FUS*) have been shown to dimerize in flanking introns.

In the lariat-driven circularization model, some exon-containing lariats formed due to an exon skipping event are further spliced. In this case, a mRNA without the spliced exons and a circRNA are expected to be generated, thus, a correlation between both transcripts is expected to some extent. This point remains under debate as many studies has corroborated the correlation, while others have not seen any trend at all or it is only limited to a few transcripts. It must be pointed out that both models of circularization are not mutually exclusive, since inverted repeats can facilitate the circularization inside an excised lariat.

It has been shown that RNA circularization depends on multiple factors. Some of the features that have been associated with circRNA formation at the gene level are (a) flanking introns longer than average, (b) single exon circRNAs are formed from unusually long exons, (c) circRNAs are typically composed of few exons (2-3 exons the most common) (d) circRNA-producing genes are longer and contain more exons than average and (e) circRNA-producing genes are usually transcribed by RNA polymerase II at a faster rate (152,161). Furthermore, it has been shown in flies carrying a variant of the large subunit of the RNA polymerase II, which decrease the elongation capacity of the polymerase, that their circRNA production capacity was reduced significantly (162). Thus, it has been proposed that circRNA synthesis is dependent of the polymerase elongation rate and other factors that compromise the linear splicing.

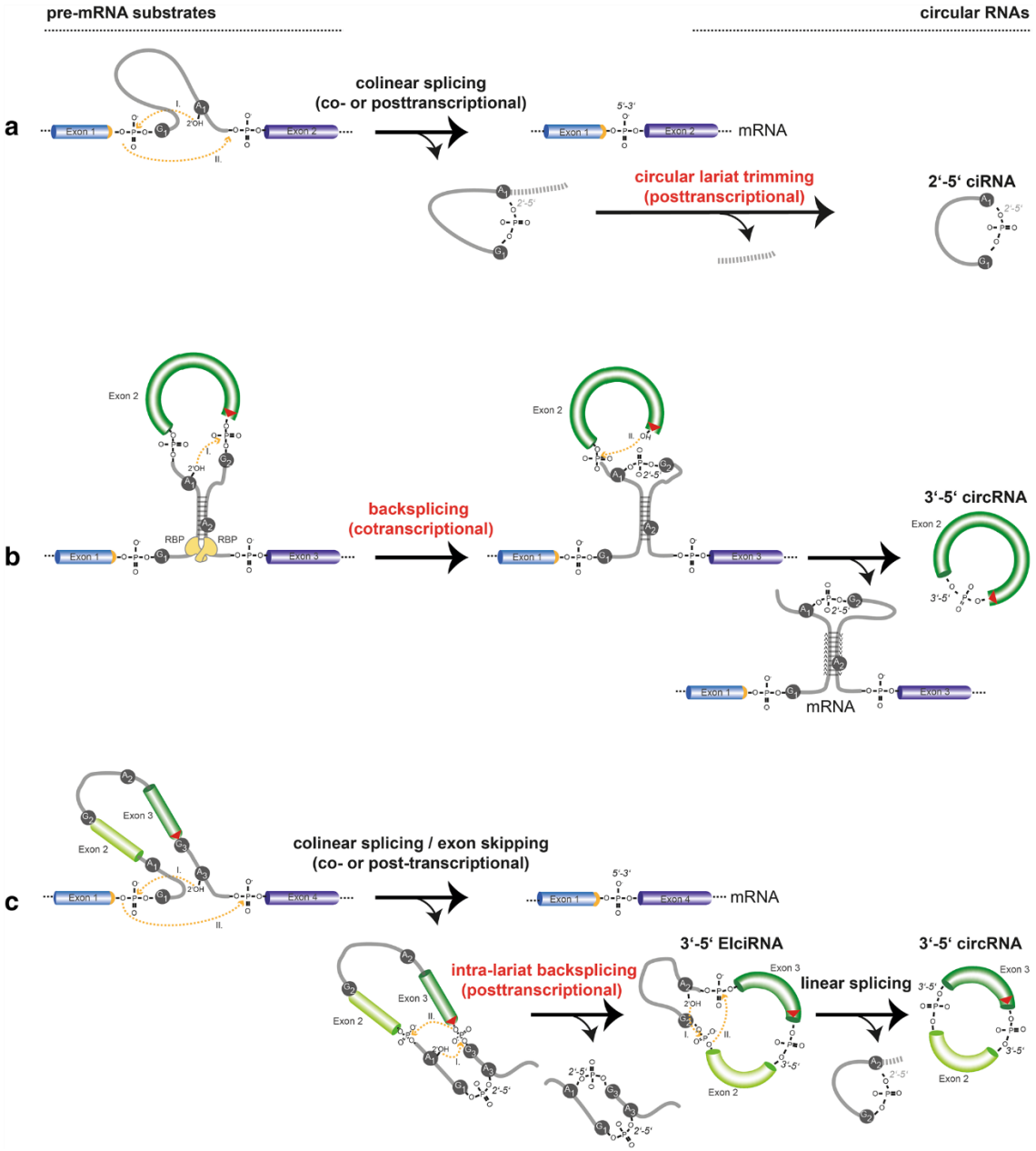


Figure 1.12: circRNA biogenesis. The starting molecule is at the left and the circRNA products on the right. **a)** From conventional linear mRNA splicing 2'→5'-linked intronic lariats are formed. Normally these lariats are degraded, but in some cases are further processed giving rise to ciRNAs. **b)** Formation of 3'→5'-linked circRNAs by co-transcriptional backsplicing. The splicing donor and acceptor are put under close proximity thanks to reverse complementary sequences in the flanking introns or due to dimerization of RNA-binding proteins that binds to flanking introns. **c)** Formation of 3'→5'-linked circRNAs by post-transcriptional backsplicing. Due to an exon-skipping event, an exon containing lariat is formed. This lariat can be further processed to produce an EIciRNA or an ecRNA (if the intronic sequence is spliced in a second reaction). The circRNA products from co- and post-transcriptional splicing are molecularly identical (figure adapted from (161)).

1.4.2.3 Mechanism of action

A 3'→5' circRNA (produced by co- or post-transcriptional splicing) has an average lifetime of 19-24h, in some cases reaching 48h, which is much longer compared to the average mRNA lifetime (4-9h) (161). These long-lived and stable molecules have been proposed to act through totally different mechanism depending their biogenesis and cellular localization (see figure 1.13). At first, the majority of circRNAs were thought to act as miRNA sponges, but with further research it has been shown that it is a rare function, since there are few circRNAs with multiple binding sites for a single miRNA or miRNA family. For now, most laboratories are focusing their attention in annotation of new circRNAs from total RNA-seq samples in multiple species and very few circRNAs have been studied in more detail, excluding bioinformatic predictions, to decipher their actual functional role. As a result, the functional role of few high-confidence circRNAs from public databases are well known.

Many exonic circRNAs with retained introns (EIciRNAs) and intronic circRNAs (ciRNAs) are found predominantly expressed in the nucleus, where they regulate transcriptional activity in a direct manner interacting with RNA polymerase II or indirectly antagonizing colinear splicing. Interaction of EIciRNAs with RNA polymerase II has been shown to be dependent of the small nuclear U1 snRNA, which forms an EIciRNA:U1 complex that stimulates the transcription of both linear mRNA and EIciRNA in the same locus in a feed-forward loop (161). In addition, ciRNAs has been shown to have little enrichment for miRNA binding sites and their knockdown has led to reduced expression of their host gene (163). As an example, the ciRNA ci-ankrd52 has been shown to associate with the polymerase II elongation machinery and acts as a positive regulator of transcription. Apart from the direct regulation through RNA polymerase II interaction, circRNAs can indirectly reduce the pool of canonically spliced transcripts, since it seems that in some cases backsplicing and linear splicing are under competition. This splicing competition model is still under debate. It has been shown that depletion of the RNA A→I editing factor *ADAR*, which is able to associate to circRNA-generating intron:intron duplexes and antagonize base pairing, increased the circRNA pool, but in contrast, the canonical linear splicing did not decreased (161). In contrast, in (146) it has been reported that for 45% of the detected ecRNAs, the corresponding colinear transcript product of the exon skipping model was detected. A second direct regulation of circRNAs to control host gene linear splicing has been shown, in which some circRNAs are able to bind to their host gene DNA forming an R-loop (RNA:DNA hybrid). The circRNA derived from exon 6 of the SEPALLATA3 (*SEP3*) gene in Arabidopsis forms an R-loop with their cognate DNA locus, which results in transcriptional pausing, favouring the recruitment of splicing factors and increasing the abundance of the exon-skipped alternatively sliced variant (164).

In addition to the nuclear regulation of alternative splicing or host canonical linear splicing of some circRNAs, it has been shown that circRNAs localized in the cytoplasm act through totally different pathways. One of the first functions discovered for circRNAs was their miRNA sponge activity. It was shown that some circRNAs has multiple binding sites for a unique miRNA or miRNA family and they act by sequestering the miRNA mature sequence, allowing the normal expression or stopping the degradation of the targeted mRNA. Two well-known miRNA sponges are the circRNAs originating from the *CDR1-AS* and *SRY* genes. The *CDR1-AS* related circRNA has been shown to have more than 60 binding sites for miR-7 (165,166), while the *SRY* related circRNA has 16 binding sites for miR-138 (167), both circRNAs with probed miRNA sponge activity. At first, it was though that the miRNA sponge activity was an extended function of circRNAs, but there are actually few circRNAs containing enough miRNA binding sites to act as sponges, while other circRNAs are only exceptions (144).

Certain circRNAs has been shown to be able to interact with proteins in the cytoplasm, sequestering them or forming different complexes. One example is the *FOXO3* related circRNA,

which has been shown to interact with senescence-related proteins *ID1* and *E2F1* and stress-related proteins *HIF1a* and *FAK* (168), sequestering them and avoiding their usual biological activities. In addition, this circRNAs has been shown to form a ternary complex with p21 and CDK2, which results in arrested function of CDK2 and in blocked cell cycle progression (154). Another examples of protein binding circRNAs are the *ANRIL* and *PABPN1* related ones. circANRIL was demonstrated to bind to members of the PeBoW complex, which functions in pre-rRNA processing during 60S ribosome maturation, and results in reduced ribosome biogenesis, while circPABPN1 is thought to compete with its parental host gene for binding to *HuR*, which is a RNA-binding protein that augments stability of other mRNAs and can bind to introns to modulate splicing of pre-mRNAs (161).

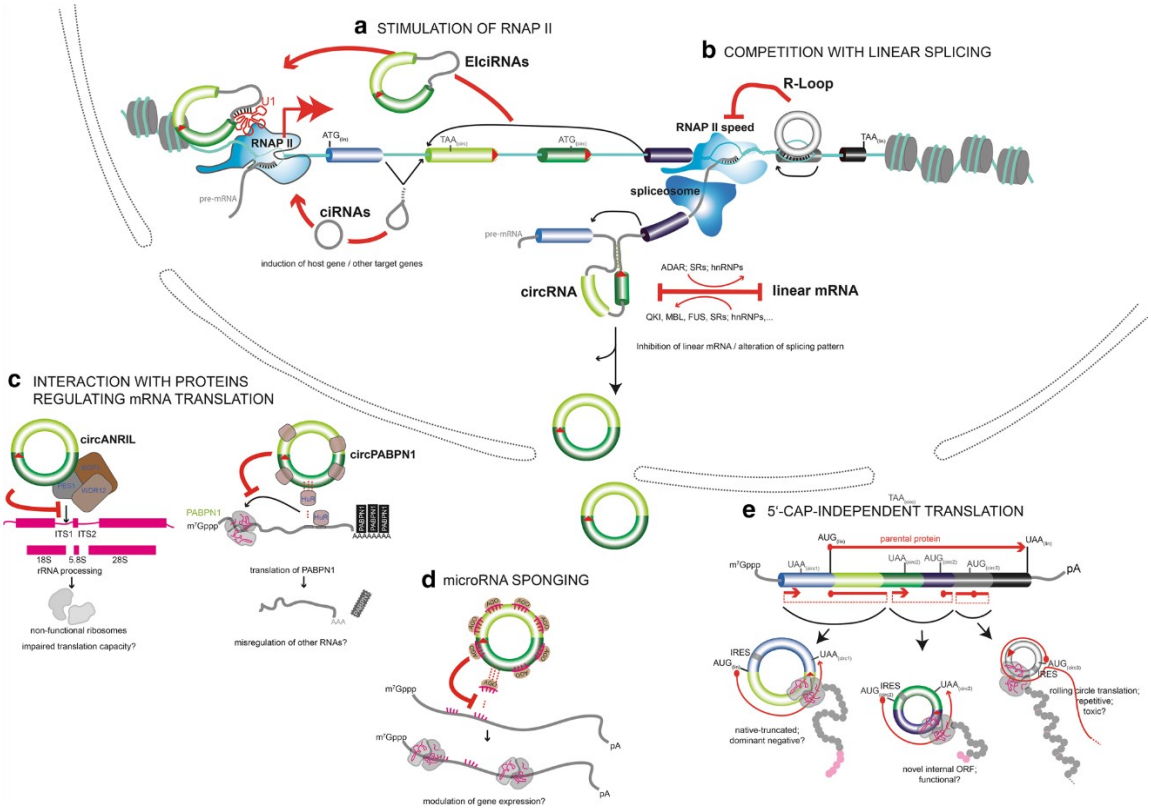


Figure 1.13: Mechanism of action and cellular localization of circRNAs. **(a-b)** Nuclear functions of circRNAs. **(c-e)** Cytoplasmatic functions of circRNAs. **a)** EIciRNAs and ciRNAs can stimulate RNA polymerase II transcriptional initiation at the transcription start site. **b)** circRNAs bind to DNA forming a DNA:RNA hybrid (R-loop) that impairs RNA polymerase II, which results in stimulation of parental exon skipping. In other cases, the circRNA production directly antagonizes the production of the colinear transcript. **c)** circRNAs can interact with proteins and inhibit their usual function. Two specific cases are shown. **d)** circRNAs can act as miRNA sponges and impede the mRNA decay or translation inhibition of miRNAs. **e)** Some circRNAs are thought to be translated into proteins, but there is not *in vivo* demonstration of such capacity (figure adapted from (161)).

After seeing that circRNAs lack both 5' cap and poly(A) tail, they were considered to not undergo any translation and, thus, they were considered non-coding elements. But most circRNAs originate from coding genes and contain different exons, and therefore, some circRNAs have been shown to contain open-reading frames (ORFs) or internal ribosome entry sequences (IRES), so it would be a possibility that some circRNAs could be translated. In a recent study, it was shown by a recent technique named TrIP-seq (Transcript Isoforms in Polysomes

sequencing), which is a technique based on polysome profiling and RNA-seq to identify different populations of RNAs associated with different numbers of ribosomes (Briefly, ribosomes are stalled during elongation with cycloheximide and then different polyribosomal complexes are separated using sucrose density gradient fractionation to prepare later sequencing libraries), that multiple circRNAs co-sediment with translating ribosomes (152). In addition, N^6-methyladenosine (m^6A) modifications have been previously related to cap-independent translation initiation and it has been shown that certain circRNAs with potential translational activity are highly methylated (160,168). Despite all that, the potential regulatory mechanism of circRNA translation is totally unknown, but some factors such as eukaryotic initiation factor 4 gamma (eIF4G2), methyltransferase-like 3 (METTL3) and methyltransferase-like 14 (METTL14) have been previously related to the circRNA translational activity (160). Further experimental validation is needed to test whether these circRNA candidates are truly translated or whether their association with ribosomes has other functionality.

Finally, different studies have pointed out that some circRNAs could be reverse transcribed in cDNA and integrated again into the genome, generating circRNA-derived pseudogenes (144). Taking into account that a circRNA-derived pseudogene would need an exon-exon junction in reverse order (non-colinear), 33 pseudogenes were predicted to originate from the RFWD2 related circRNA (144).

1.4.2.4 Databases and nomenclature

In contrast to miRNAs, circRNAs does not have a clear structure, which makes them hard to annotate and predict their functionality. The two main databases for circRNAs are circBase (http://www.circbase.org/) and CIRCpedia (http://www.picb.ac.cn/rnomics/circpedia/) and each one has their one nomenclature system. There is no consensus nomenclature for the circRNAs in common. In addition, both databases only have human, mouse, fly and worm circRNAs for now (CIRCpedia has some additional rat and zebrafish circRNAs). Furthermore, different studies have pointed out that those databases based on circRNA predictions have a high percentage of false positives. Apart from circRNA annotation databases, there exist others recording miRNA-circRNA predicted interaction such as circRNABase (http://starbase.sysu.edu.cn/starbase2/mirCircRNA.php) and other recording circRNA-disease associations such as circ2Traits (http://gyanxet-beta.com/circdb/).

1.5 Project origin

In 2008, there was an outbreak of bluetongue virus (BTV), which causes the ruminant disease bluetongue, through Europe. BTV is a non-contagious insect-borne *Orbivirus*, which mainly affects sheep and with less frequency cattle and goats. Until recently at that time, there were only live attenuated vaccines developed against such disease, but the risks associated with such vaccines were enough to discourage their use in several countries. The first new case of BTV serotype 8 (BTV-8) in Europe during 2008 was probably detected in Cantabria, Spain (169). Early in 2008, several inactivated vaccines for BTV-8 were made commercially available. At the same time, the BTV serotype 1 (BTV-1) was spreading. A rapid development of an inactivated vaccine was possible as government and industry in Spain were already collaborating in the production of an inactivated vaccine against BTV serotype 4 (169).

The re-emergence of the disease and the availability of inactivated vaccines made the European Union to decide to start a vaccination campaign. In Spain, animals received vaccines

against two serotypes of the virus (mainly BTV-1 and BTV-8) with prime inoculation and boosting, thus receiving a total of 4 vaccines in less than a month. After such vaccinations, both composed of aluminium hydroxide salts, a new syndrome with an acute nervous phase normally followed by a chronic cachectic phase was seen in some animals and it had severe consequences in the sheep industry in Spain (5). Such symptoms were thought to be caused by the repetitive inoculation of aluminium adjuvants and their accumulation in the organism, as there were previous reports of adverse effects caused by the repetitive inoculation of vaccines in other organisms, although such reports remain controversial and are under constate debate in the scientific community. The clear advantages of vaccination are well documented, whereas the risk of adverse effect is poorly documented or hypothetical in some cases.

The project started in an attempt to check if the symptoms shown after vaccination were caused by the adjuvant, trying to reproduce the syndrome in a flock of sheep exposed to multiple vaccinations. The experiment would allow to study in more detail the mechanism of action of aluminium hydroxide as adjuvant, as it is not totally understood how the adjuvant perform its function in the organism.

1.6 Aims and outline of the thesis

For approximately 90 years, different aluminium compounds have been used as adjuvants in veterinary and human vaccines. Aluminium Hydroxide, Aluminium Phosphate and Aluminium Sulfate are the main forms of aluminium used as adjuvants. Despite its widespread use, the mechanism of how aluminium-based adjuvants exert their beneficial effects is not completely known. In addition, in some cases adverse effects have been described after inoculation.

In 1998, a new inflammatory muscle disorder was described in which the presence of aluminium conglomerates in macrophages was demonstrated and was linked to the inoculation of vaccines with aluminium adjuvant, the disease is currently known as macrophagic Myofasciitis (MMF). Different studies with animal models have concluded that vaccines with aluminium hydroxide as an adjuvant can cause local tissue damage and behavioural neurological changes similar to those seen in MMF. In sheep, a form of adjuvant-induced autoimmune/ auto-inflammatory syndrome (ASIA – Autoimmune /Autoinflammatory Syndrome Induced by Adjuvants) has been described in association with repeated inoculation of aluminium-based vaccines. This syndrome was detected in sheep after a mandatory vaccination in ruminants against the bluetongue virus in 2008 in Spain.

Despite indications of negative effects after vaccinations in predisposed individuals, the need for vaccinations and the beneficial effects they have for the general population is undeniable. All this is a subject of great controversy in the scientific community. The fact of having been used for so long without fully knowing its mechanism of action does not help. More studies on vaccination are necessary to try to decipher how aluminium acts as an adjuvant, and it can be beneficial to improve future vaccine developments.

In this thesis project, the study of the effect of repetitive vaccination (by means of vaccines with aluminium hydroxide as an adjuvant) with transcriptomics and bioinformatics technologies is being carried out, in an attempt to decipher aluminium hydroxide mechanism of action. RNA sequencing (total RNA-seq and miRNA-seq) is being used to study global gene expression. This technique has several basic advantages over other technologies. It has a large dynamic range of expression and is not limited to detecting transcripts that correspond to pre-existing genomic sequences, which makes it particularly attractive for use in non-model organisms whose sequence is not completely determined, as is the case with sheep. The works carried out by RNA-seq in sheep are scarce, although they are increasing little by little thanks to the lowering of cost of the technology. In addition, there are still few works in which the new

RNA sequencing technologies have been applied to motorize the immune response to vaccination.

Thus, the general aim of this work is the identification of genes and regulatory elements involved in the immune response induced by the repetitive inoculation of vaccines with aluminium hydroxide by genomic and bioinformatic techniques. It is intended to deepen the knowledge of the effect of aluminium hydroxide as an adjuvant, especially at the genomic level, which would allow us to understand the process in a more detailed manner. The hypothesis of this doctoral thesis work is that there are differentially expressed genes and regulatory elements in animals exposed to aluminium and that such effects can reach distant organs. These genes and the routes involved will allow us a better understanding of the body's response to aluminium.

The specific objectives of this doctoral thesis work are:

i) Quantification of the level of expression at the genomic level of mRNAs and miRNAs in healthy animal tissues (prior to the repetitive vaccination experiment) and that have been exposed to repetitive vaccinations during the experiment. Samples from two different tissues will be sequenced: peripheral blood mononuclear cells (PBMCs), with samples at the start and at the end of the experiment; and parietal lobe cortex, with only samples after the vaccination schedule.
ii) Comparison of the expression, co-expression and interaction at the genomic level of mRNAs and miRNAs, which will be studied in parallel, in the groups of animals analysed.
iii) Detection and characterization of new non-coding transcripts as circular RNAs, in addition to studying possible roles of these elements in the mechanism of action of aluminium.

Chapter 2

Materials and methods

Due to affordable tariffs and its ability to characterize without prior knowledge the transcriptomes of non-model organisms, RNA-seq technology (alone or in conjunction with other NGS technologies) has become the main choice for genome wide analyses. When working with this kind of data, laboratories must deal with storage of enormous quantities of data and analysis workflows that are not fixed and require high computational resources.

This chapter briefly introduces a general background for each step of the analysis done in our data. First, a detailed description of the sampling procedure and the schedule followed during animal treatment will be given. Then, a brief description of parameters that must be taken into account for an RNA-seq experiment design will be presented, explaining the set-up of this thesis. Finally, the workflows executed in the data presented in this thesis will be given, explaining each step in detail and giving a brief description of the programs that can be used. A total of three different workflows will be explained: for both total RNA-seq and miRNA-seq differential expression analysis and another one for circRNA annotation.

2.1 Sampling workflow

The sampling procedure employed in this work is presented in figure 2.1. Briefly, twenty-one Rasa Aragonesa purebred lambs were selected from a single pedigree flock of certified good health at three months of age, with the condition of not having undergone any vaccination before the experiment. The flock analysed in this study was stablished at the experimental farm of the University of Zaragoza and was always maintained indoors, with ideal controlled conditions of housing, management and diet. Before the start of the experiment, the animals were care for two months to acclimatize to the new environment, so they were five months old when the experiment started. Then, all lambs were randomly distributed in different treatment groups, each consisting of 7 animals. One of the groups, from now on denominated vaccine group (Vac), received a subcutaneous treatment with commercial vaccines based on aluminium hydroxide adjuvant. Another group, denominated adjuvant group (Adj), received equivalent doses to the commercial vaccines of aluminium hydroxide only (Alhydrogel®, CZ Veterinaria, Spain) diluted in phosphate-buffered saline (PBS). Finally, PBS was administered to the control group (Control). Blood samples were taken at the start (before any vaccination) and at the end of the experiment, while for encephalon (parietal lobe cortex) and spleen only samples at the end were taken. A more detailed description of the samples from each tissue can be seen later in their corresponding chapter.

It must be pointed that all experimental procedures were approved and licensed by the Ethical Committee of the University of Zaragoza (ref: PI15/14). Requirements of the Spanish Policy for Animal Protection (RED53/2013) and the European Union Directive 2010/63 on protection of experimental animals were always fulfilled.

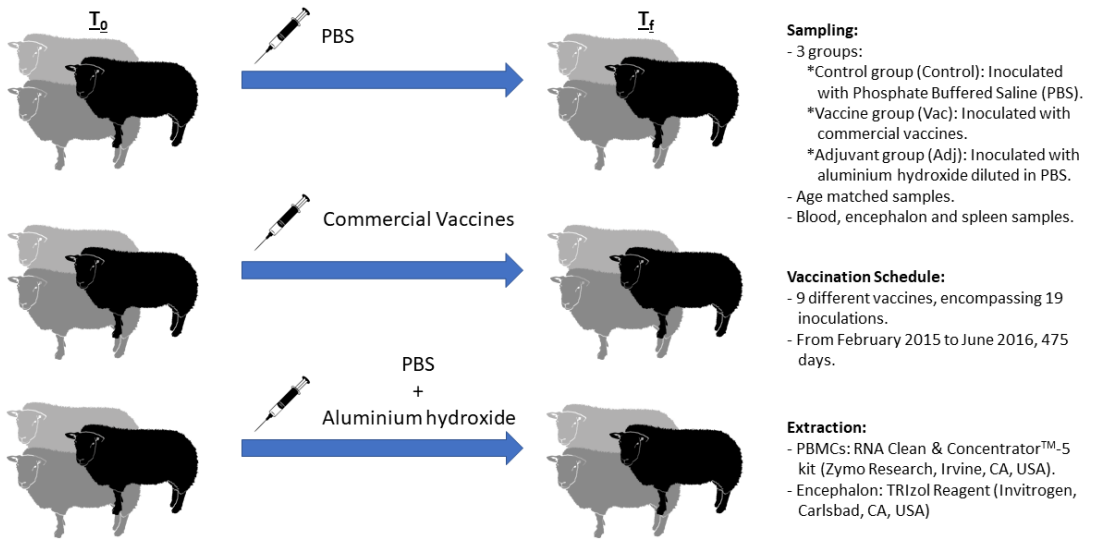


Figure 2.1: Experimental design. Samples from three different groups were obtained for sequencing (total RNA-seq and miRNA-seq).

2.2 Vaccination Schedule

The complete experiment lasted 475 days, from February 2015 to June 2016. During that period of time, nine different vaccines were administered to each animal, which comprises a total of 19 inoculations throughout 16 different inoculation dates. All vaccines, except for Heptavac P Plus which contains toxoid, are composed of inactivated bacteria or viruses, and they not contain purified antigens for a specific pathogen. As previously stated, sheep were divided in three different group, each receiving a different inoculum (see figure 2.1). The Vac group only received commercial vaccines with aluminium hydroxide as adjuvant. All commercial vaccines were administered following manufacturer recommendations and leaving between each inoculation the recommended time. For the Adj group, for each time of inoculation a solution was prepared with a concentration equivalent in aluminium in mg/g. To do so, Alhydrogel was diluted in PBS at the concentration of the commercial vaccine that was used each time. A total amount of 81.29 mg of Al per animal was given in the Vac and Adj groups throughout all the experiment. A detailed list of the commercial vaccines used in this study can be seen in table 2.1, while in figure 2.2 a detailed timeline with the followed vaccination schedule can be seen.

2.3 RNA sequencing

2.3.1 Experimental Design

2.3.1.1 Parameters that must be taken into account

When designing an RNA-seq experiment, multiple parameters must be taken into account. Depending of the organism under study or the main objectives of the experiment, different platforms or set ups could be chosen. If the main objective of the experiment is the annotation of new genes, understanding gene as any element of the DNA that is transcribed regardless of

whether it is protein coding or not, or a *de novo* transcriptome assembly of a non-model organism, a platform that returns longer fragments would be preferred. In contrast, if the objective of the experiment is a differential expression analysis of a well or partially annotated organism, a platform that returns more fragments at the cost of being shorter is preferred for an adequate quantification of lowly expressed transcripts. Apart from the platform selection, there are other parameters that must be chosen before the sequencing of libraries. Among the most important characteristics that would be explained in more detail below are: single-end or paired-end libraries, strand specificity, poly(A) or rRNA depleted libraries, sequencing depth, read length and number of replicates.

Table 2.1: Commercial vaccines used on sheep in the experiment of this thesis.

Vaccine number	Commercial name	Manufacturer	Antigen/s	Inoculation day	mg of Al per dose
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

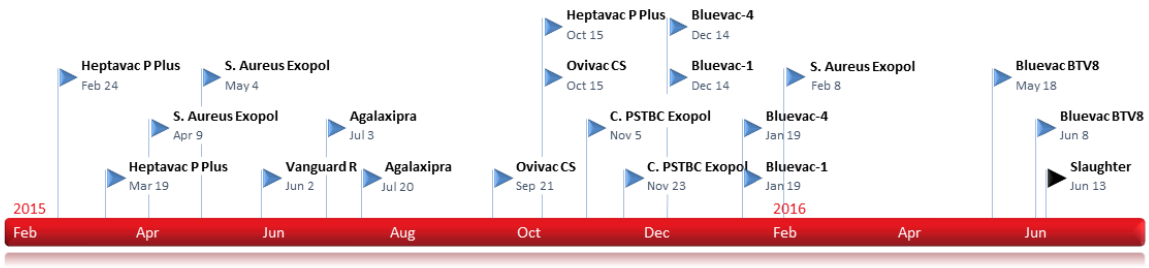


Figure 2.2: Followed timeline with the vaccination dates. Nine different vaccines were administered to each animal, which comprises a total of 19 inoculations throughout 16 different inoculation dates.

One of the most important parameters to take into account when designing an RNA-seq experiment is the sequencing depth or coverage (i.e. *number of reads* X *read length* / *target size*). Despite being thought at first that RNA-seq technology produced an unbiased

quantification of the transcriptome, now it is known that longer transcripts generate more reads than shorter ones of similar expression and that a great proportion of reads is concentrated on a few highly expressed transcripts (121). Thus, the sequencing depth would determine the ability to detect rare or lowly expressed transcripts, since higher sequencing rates would allow for a more accurate quantification of expression levels. Blencowe et al. (2009) estimated that for single-end sequencing ~700 million reads would be required to quantify >95% of transcripts, but with less than 10 million reads >80% of transcripts were accurately sequenced (170). If the main objective of the study is a differential expression analysis, it has been shown that with a moderate sequencing depth a stable detection of protein-coding genes can be achieved (121). In contrast, if the main objective is the detection of rare transcripts such as non-coding RNAs, higher depths are needed. In table 1.4 can be seen the recommended depths or the range of depth of other studies found for each analysis type. Such recommendations are not graved into stone, as such values can vary greatly depending on the organism under study or the platform used for sequencing between others. It must be pointed that there is no clear consensus in the required minimum depth in the case of long non-coding RNAs (lncRNAs) and circRNAs identification. It has been suggested that a minimum of 100 million reads may be needed for reliable lncRNA prediction (171), but studies with a mean sequencing depth greater than 40 million reads have been able to report lncRNAs (172,173). Furthermore, studies from rRNA depleted total RNA libraries with a mean sequencing depth of 30 million reads have been able to detect circRNAs (174,175), although higher sequencing depths may be recommended as only junction spanning reads are used to identify circRNAs (176).

Table 2.2: Recommended sequencing depth for each RNA-seq application.

Analysis type	Recommended depth (million reads)	References
Differential expression (DE)	10-25M	https://genohub.com/ngs/
Alternative splicing	50-100M	https://genohub.com/ngs/
De novo assembly	100-200M	https://genohub.com/ngs/
SNP discovery	50-100M	https://genohub.com/ngs/
miRNA DE	1-2M	https://genohub.com/ngs/
miRNA discovery	5-8M	https://genohub.com/ngs/
lncRNA annotation	40-100M	(171–173)
circRNA annotation	>30M	(174,175)

Other parameter that must be taken into account when designing RNA-seq libraries is whether a single-end (SE) or paired-end (PE) sequencing will be done. In SE sequencing, reads from one end of an RNA fragment are generated, while in PE sequencing, after the first read with a specified length is generated from one end, the opposite end of the fragment is sequenced. As the fragmentation step and size selection step during library construction produces RNA fragments of known length and the reads produced during sequencing are of a predetermined size, the distance between each paired read is known and this information can be used when aligning reads. Because of that, PE sequencing allows for a better detection of genomic rearrangements (insertions, deletions or inversions), repetitive regions and novel transcripts. Despite the clear advantage of PE sequencing, it must be pointed out that it is more expensive and depending of the interest of the study, a SE approach can be enough (e.g., as most small non-coding RNAs are short, some of them being shorter than the reads generated from the sequencing machine, a SE sequencing is enough).

In addition, at the library construction step, it must be chosen between a poly(A) selection or an rRNA depletion procedure. This step is necessary to remove some highly expressed transcripts without interest such as rRNAs (they can represent over 90% of the RNAs present in a cell (177)), which in case of keeping them during the sequencing, they would take the majority of reads, leaving the rest of the transcriptome underrepresented. In the poly(A) selection method, poly(dT) oligomers attached to a surface are used for selection of RNAs with poly(A) tails at the 3' end, which results of selection of mature mRNA transcripts, ignoring most non-coding elements. In contrast, in the rRNA depletion method, oligomers complementary to rRNAs are used for removal of nearly all rRNA transcripts, which results in a full transcriptome sequencing (including non-coding elements). It has been shown that rRNA depletion libraries return more intronic reads than poly(A) selection ones, most of them originating from non-spliced immature transcripts (178). If the main objective are protein-coding genes, a poly(A) selection is recommended. To achieve an exon coverage similar to a poly(A) library, a higher depth would be necessary in a rRNA depletion library. As a result of a broader fraction of the transcriptome being sequenced in rRNA depletion libraries, there is discrepancies between multiple genes in both methods, making the expression levels of both of them incompatible and hard to compare without a prior correction (179).

Another consideration that must be taken into account is whether to generate a strand-specific library, in which the orientation of the original RNA transcript is known. Without the strand information, the quantification of overlapping genes in opposite strands remains a challenge. Depending the alignment of the reads, two strand specific libraries can be distinguished (see figure 2.3 for more information).

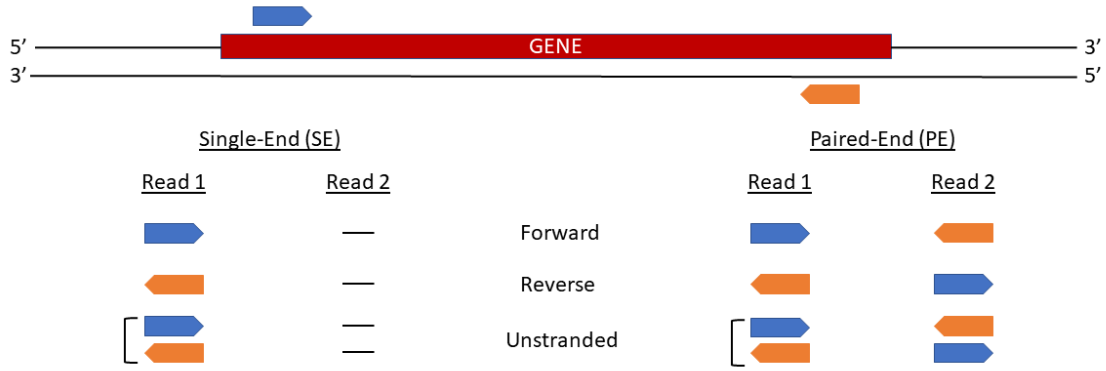


Figure 2.3: Library strandness. If read 1 (or single-end reads) comes from the forward strand (original RNA strand) and reads 2 comes from the reverse strand, the library is said to be "stranded". At the contrary, if read 1 (or single-end reads) comes from the reverse strand and read 2 comes from the forward strand, it is said to be "reverse stranded". In the case that both cases happen in a library in a similar proportion, it is said to be "unstranded". The "reverse stranded" libraries originate from library preparations based on dUTP methods (like the Illumina TruSeq stranded libraries), while "stranded" libraries surge from less common library preparations (such as standard SOLID).

Another consideration that must be taken into account when preparing libraries for sequencing is the desired read length, which would depend on the objectives of the experiment. It has been shown that reads of 50 bp are enough for differential expression analyses in well annotated organism, not being any substantial change with longer reads, while for splice junction detection, the longer the reads the better, showing 100 bp PE libraries the best performance (180). Illumina itself (https://emea.support.illumina.com/) recommends at least

SE 50-75 bp length reads for gene expression of the coding transcriptome, PE 75 bp reads for full transcriptome analysis, which would allow for *de novo* assembly or identification of novel variants and splice sites, and SE 50 bp reads for small RNA analysis.

Finally, the number of replicates that should be included in an RNA-seq experiment is a crucial factor. There are two kind of replicates, technical and biological, and each one is used to address different questions. Technical replicates refer to repeated measurements of a sample and they are used to test the variability of a protocol or assay. These types of replicates evaluate the precision and reproducibility. In the case of RNA-seq experiments, it has been shown to have high technical reproducibility and there is no need of technical replicates (181). On the other hand, biological replicates refer to parallel measurements of biologically distinct and independently generated samples. These types of replicates are used to test the biological variation. The more samples available, a better improvement in the strength for making statistical inferences. Due to the high cost of RNA-seq studies, most laboratories restrict their inferences to small sample groups and try to balance the number of replicates and sequencing depth. Generally, more replicates are preferred than a higher sequencing depth for differential expression analyses, always taking into account that if the interest is the detection of lowly expressed genes or alternative-splicing, higher depths would be needed. Independent of the biological question of interest, it has been pointed out that to make inferences on the population a minimum of three biological replicates per group are required (89).

2.3.1.2 Design Setup

As previously stated, the main objective of this study is the differential expression between the different groups (described in the sampling workflow section) in an attempt to discern the mechanism of action of aluminium-based adjuvants. It must be pointed that the sheep reference genome (Oar v3.1), like other ruminant reference genomes, are assembled and with an extensive annotation in protein-coding genes, but they are constantly being improved and some pathways are still partially described. In addition to differential expression, we are also interested in annotation of non-coding elements (mainly miRNAs, circRNAs and lncRNAs) to check if they have any role in the adjuvant mechanism. Taking into account that such non-coding elements are poorly annotated in the reference genome (there is no database recording sheep circRNAs; there are ~155 sheep miRNAs in miRbase; and most of the annotated sheep lncRNAs in NCBI or ENSEMBL are predicted), it is clear that a PE stranded library would be required with a high sequencing depth and an rRNA depletion step.

Total RNA-seq libraries were prepared according to the TruSeq Stranded Total RNA kit with Ribo-Zero (Illumina, San Diego, CA, USA) for encephalon samples and the TruSeq Stranded Total RNA kit with Ribo-Zero Globin (Illumina, San Diego, CA, USA) for PBMCs. Total RNA libraries were sequenced on a HiSeq2000 sequencer with a mean sequencing depth of 75 million and 70 million 75 base-pair (bp) paired-end reads at CNAG (Centro Nacional de Análisis Genómico, Barcelona, Spain) for encephalon samples and PBMCs, respectively. miRNA libraries were sequenced in a Hiseq2500 sequencer with a mean sequencing depth of 19 million and 17 million 50 bp single-end reads at CRG (Centro de Regulación Genónimca, Barcelona, Spain) for encephalon samples and PBMCs, respectively. A detailed description of the sequenced samples and summary statistics will be given later in the corresponding chapters for each tissue.

2.3.2 Total RNA-seq differential expression

Once the RNA libraries have been prepared and sequenced, the returned huge files need to be analysed through multiple steps, which would depend on the main objective. The process begins with separate files for each sample (in FASTQ format) in which the sequenced fragments are recorded (in the case of PE samples, two files are returned, one for each end of the fragment). Briefly, for a differential expression (DE) analysis, the main scheme would be composed of the following steps:

1. Quality control and pre-processing of reads.

2. Alignment of reads to a reference genome (In case of a non-model organism or poorly annotated organism, a *de novo* assembly).

3. Quantification of gene (or transcript or exon) expression levels.

4. Differential expression analysis.

5. Gene set enrichment analysis.

In this section, a brief description of the major steps for a DE analysis of total RNA data will be provided, with a brief summary of the different programs that can be used in each step. At the end, the exact workflow used in the data presented in this thesis will be provided.

2.3.2.1 Quality control and pre-processing

Quality problems can arise from different sources, mainly the samples themselves (e.g., RNA degradation), at the library preparation step or at sequencing. Among the most common bias sources, incorrect base calling at sequencing, non-uniform transcript coverage (e.g., consequence of the sequence GC content), untrimmed adapters, sequence contamination (e.g., other organism) or duplicate sequences (due to PCR, RNA degradation or bad rRNA depletion, among others). Prior to any analysis step, the quality of the samples must be checked, since some biases can affect future steps, such as alignment or quantification. Some biases can be corrected by trimming programs (mostly bad quality base calls, adapters or repetitive sequences), others must be accounted by normalization or batch effect removal methods after gene quantification, while others are not possible to correct for and must be taken into account when interpreting the data. In table 2.3 can be seen a brief list of different quality control programs that can be applied to NGS data. Among them, FASTQC is one of the most used by laboratories, as it returns quality metrics at a fast pace. Since most of the tools returns similar metrics, the ones from FASTQC would be described in more detail, explaining what would be expected in presence of any kind of bias.

All sequencing machines returns multiple fragments in which each base has an assigned quality measure called Phred score. Phred scores, Q, represents the estimated probability of a base being incorrect in Illumina sequencing and is determined based on base calling peaks. It would be interpreted as follow:

$$Q = -10\ log_{10}\ P$$, being P the error probability.

Table 2.3: List of tools for quality control of RNA-seq samples (some of them are able to do trimming of samples), with their main characteristics and returned statistics and plots. For a more detailed description of the returned statistics by each program, their corresponding webpage is advised. Table based on tool references and corresponding webpages.

Tool	Last Version*[1]	Main Features	Returned Statistics and Plots
FastQC (182)	v0.11.8	Developed in Java. Has a standalone version with a graphical user interface. Multithreading. Only SE reads*[2].	Per base sequence quality. Per base sequence content. Per base GC content. Per sequence GC content. Per base N content. Sequence length distribution. Duplicated Sequences. Overrepresented sequences. Overrepresented K-mers
Fastp (183)	v0.20.0	New version of AfterQC developed in C++. Multithreading. Can perform filtering and trimming of adapters and bad quality bases. Allows SE and PE reads.	Adapters presence. Duplication rate. Inserts size estimation. Per base sequence quality. Per base sequence content. Overrepresented sequences. K-mer counting.
Fastqp (184)	V0.3.4	Developed in Python. Quite similar to FASTQC. Not multithreading. Only SE reads*[2].	Quality score heatmap. Per base sequence quality. Per base GC content. Per sequence GC content. Per base sequence content. Reference mismatches by cicle. K-mer content. Read length distribution.
FASTX-Toolkit (185)	v0.0.13	Web-based and command-line versions. Can perform filtering and trimming of adapters and bad quality bases. Only SE reads*[2].	Per base sequence quality. Per base sequence content.
HTSeq (186)	v0.11.2	Developed in Python. Not multithreading. Allows SE and PE reads. More focused for post alignment quality.	Per base sequence quality. Per base sequence content for aligned and non-aligned reads. Transcript coverage.
KRAKEN (187)	v13-274	Tool with three modules designed mainly for small RNA NGS data. Its Reaper module can be used for quality control and trimming.	Per base sequence quality. Per base sequence content. Sequence length distribution.
MultiQC (188)	-	Not being directly a program that calculates quality metrics, it can aggregate quality results (pre-processing ones and post alignment ones) from multiple samples from numerous bioinformatics tools.	Generates aggregate reports from 77 different bioinformatic tools (FastQC, Cutadapt, Trimmomatic, Bowtie, STAR, HISAT2, TopHat, Salmon, StrinTie, feaureCounts, GATK, HTSeq, Picard, RSeQC, Samtools, SnpEff, etc.).

[1] Last checked on 29/10/2019.

[2] Each PE file must be executed separately.

Table 2.3: (continued).

Tool	Last Version*[1]	Main Features	Returned Statistics and Plots
NGS QC Toolkit (189)	v2.3.3	Developed in Perl. Multithreading. Can perform filtering and trimming of reads. Allows SE and PE reads.	Per base sequence quality. Per base sequence content. Per sequence GC content. Summary of quality check and filtering.
PRINSEQ (190)	v0.20.4	Developed in Perl. Web-based and command-line versions. Can perform filtering and trimming. Not multithreading. Allows SE and PE reads.	Per base sequence quality. Per base sequence content. Sequence length distribution. Per sequence GC content. Per base N content. Duplicated Sequences. Sequence complexity (DUST and Entropy). Sequence contamination (PCA).
RNA-SeQC (191)	V2.3.4	Developed in Java. It is built on the GATK and Picard API. Not multithreading. Allows SE and PE reads.	Per base sequence quality. Per base GC content. Duplicated Sequences. Sequence length distribution. rRNA reads. Transcript coverage. Other plots related to alignment results.
RSeQC (192)	v3.0.1	Developed in Python. Allows SE and PE reads. Plotting modules for alignment quality.	Per base sequence quality. Per base sequence content. Clipping profile. Deletion profile. Insertion profile. Mismatch profile. Transcript coverage. Calculates inner distance between read pairs. Junction annotation. Junction saturation. Duplicated Sequences. Per sequence GC content. Hexamer frequency. Transcript integrity number (TIN) calculation.
SolexaQA++ (193)	v3.1.7.1	Developed in C++. Support Illumina, Ion Torrent and 454 data. Can be used in Windows systems. Can perform filtering and trimming. Allows SE and PE reads.	Sequence quality heatmap. Mean quality distribution. Sequence length distribution.
Trim Galore (194)	V0.6.4	Developed in Perl. Wrapper tool around FastQC and Curadapt. For quality control and sequence trimming.	Similar features to FastQC and Cutadapt.

*[1] Last checked on 29/10/2019.

*[2] Each PE file must be executed separately.

Thus, if a Phred quality score of a base is 30, the chance of that base being incorrectly assigned is 1 in 1000 (or has a 99.9% of being accurately assigned). Using the information given by these quality scores, bad quality bases can be trimmed off from the fragment to improve alignments to the reference genome. One of the most common plots from pre-processing tools such as FastQC or PRINSEQ is a box plot representing the median Phred quality for each base position. It gives an initial representation of the sequencing quality. In this kind of plots is common to see a linear decrease in base calling quality at the end of fragments, since most sequencing machines accumulate pre-phasing and post-phasing errors with long sequences (explained in more detail in the RNA-seq biases section in the Introduction). Other plot such as the average quality distribution can point if there is a subset of sequences with bad quality that need to be completely removed from the analysis. Furthermore, it is expected during library preparation a uniform sequence sampling from any long transcript, so if the sampling was truly random, the probability to found any base at each position would be constant. It has been shown that random hexamer priming introduces a bias in the first ~12 bases. Other plots such as GC content distribution or highly duplicated sequences can point for library contaminations (e.g., due to the presence of other organisms or deficient rRNA depletion). Finally, PRINSEQ calculates dinucleotide odds ratios and use that information to detect possible sequence contaminations through a principal component analysis (PCA).

Once the quality of samples has been checked, adapter sequences, bad quality bases and low complexity sequences must be removed to improve the alignment to the reference genome or the *de novo* assembly. Apart from some tools listed in table 2.3 that can perform both quality checks and sequence filtering, in table 2.4 can be seen a brief list of some tools specifically designed for sequence filtering, base trimming or sequence correction. Generally, in RNA-seq data originating from long transcripts, it is uncommon to have adapter sequences in reads due to the limited sequenced lengt. In addition, it is always recommended to execute a length filtering step after quality trimming, since very short fragment can result from the trimming. Short reads can lead to ambiguously mapped reads. It should be noted that some biases (such as GC content) can not be corrected with these tools and must be addressed after sequence alignment. It has been shown da SolexaQA achieves the best performance (highest quality fragments while keeping the highest number of aligned reads) in low quality RNA-seq datasets (195). In addition, in the same study, it has been pointed out that the best performances are always achieved when trimming using intermediate quality thresholds (Phred quality score between 20 and 30). Sequence trimming must be done with caution, since it has been shown that aggressive trimming strategies can lead to changes in gene expression estimates across multiple data sets and to a reduction in correlation with microarray data (196).

2.3.2.2 Alignment to reference genome or transcriptome

Once low-quality bases have been removed from the sequenced fragments, these sequences must be aligned or mapped to a reference genome to estimate the loci of origin. This step allows the determination of the exact coordinates of most reads, which would allow for novel transcript or gene discovery (if a transcriptome assembly is done after the alignment to the reference genome) or expression level quantification (at exon, transcript or gene level). In absence of a reference genome, fragments can be aligned against a reference transcriptome (In absence of both, a *de novo* assembly would be recommended). The alignment can be challenging due to several reasons: 1. Generally too short reads are aligned, but reference genomes can be large and with multiple repetitive regions (low complexity regions, repeats and pseudogenes), as a

consequence the sequence align to multiple places; 2. Aligners must cope with sequencing errors and variation; 3. Millions of reads must be aligned, a computationally intensive task; 4. Reference genomes from mammals and other organisms have intronic sequences, so some reads may align in a non-continuously manner and these reads must be spliced to determine correctly exon-intron boundaries; and 5. Sample specific attributes such as SNPs and INDELs.

Table 2.4: List of tools for RNA-seq sequence filtering and trimming. All tools are compatible with paired-end data, unless otherwise stated. Table based on tool references and corresponding webpages.

Tool	Last Version*[1]	Main Features
BBDuk (197)	v38.70	Included in the BBMap package. In addition to usual quality-related trimming and filtering, it can perform contaminant filtering via kmer matching, GC filtering, entropy-filtering and quality-score recalibration. A good choice for rRNA transcript filtering with a database such as SILVA, although other databases can be used if there is any suspected source of contamination of known origin.
Cutadapt (198)	v2.6	Adapter removal, quality trimming (if average quality below a pre-defined range), minimum length, discard reads containing more than a threshold of N bases.
PRINSEQ (190)	v0.20.4	Available as web application and command line. Quality trimming, minimum length, GC filtering (separate sequences in a bi-modal distribution of the GC content), N base filtering, entropy filtering, trim ends by quality scores.
Scythe (199)	v0.994	Tool using a Naive Bayesian approach to classify contaminant substrings for 3'-end adapter removal. It recommends to be executed to any other quality trimming tool.
SEECER (200)	v0.1.3	Tool for sequencing error correction, based on a hidden Markov Model. Removes mismatch and indel errors from the data.
TagCleaner (201)	v0.16	Tool for tag sequence (e.g. WTA tags) removal. Recommended to be run before any trimming.
Trimmomatic (202)	v0.39	Adapter removal. Multiple options for sequence trimming and cropping of Illumina data (sliding windows, minimum length, leading, trailing, average quality filtering).

*[1] Last checked on 30/10/2019.

There are a great variety of tools for RNA-seq data alignment, each with its own strategy. These tools can be classified depending of their approach, mainly being two main categories: unspliced aligners and spliced aligners. Unspliced aligners align reads to the reference without any large gap and are limited to identify known exon and junctions, while spliced aligners can handle large gaps and are perfect for the detection of intron-spanning reads. Unspliced aligners can be further divided in "seed methods" and "Burrows-Wheeler transform methods", while the spliced ones in "exon first" and "seed and extend" methods (203). In "seed methods" short sub-sequences, called seeds, are aligned with perfect complementarity to the reference and, when a match is found, more sensitive methods are used to extend the seed alignment. In contrast, in "Burrows-Wheeler transform methods", the reference is transformed in a structure that is efficient for searching complementary sequences. Among spliced aligners, "exon first" methods first map fragments in a continuously manner with an unspliced aligner and then, unmapped

reads are divided into shorter segments to try to align non-continuously. Alternatively, “seed and extend” methods directly split reads in shorter segments, seeds, and try to align with perfect complementarity and, when a match is found, more sensitive methods are used to extend and detect spliced alignments. The tool will be chosen depending the organism of interest and the main interest of the study. Each tool has its own benefits and drawbacks, some of them being faster or being more sensitive for splice junction detection, while others such as “exon first” based ones have been shown to be more deficient in some regions (Some spliced reads could be misaligned continuously to a pseudogene, instead of being aligned preferentially in a non-continuously manner to the corresponding gene) (203). In addition, each aligner handles multimapping reads in a different manner, most of them discard such reads, while other allocate them randomly, based on an estimate of coverage or following a statistical model. In table 2.5 can be seen a brief list of both unspliced and spliced aligners. In addition, in figure 2.4 can be seen a timeline with different aligner publication dates up to 2017.

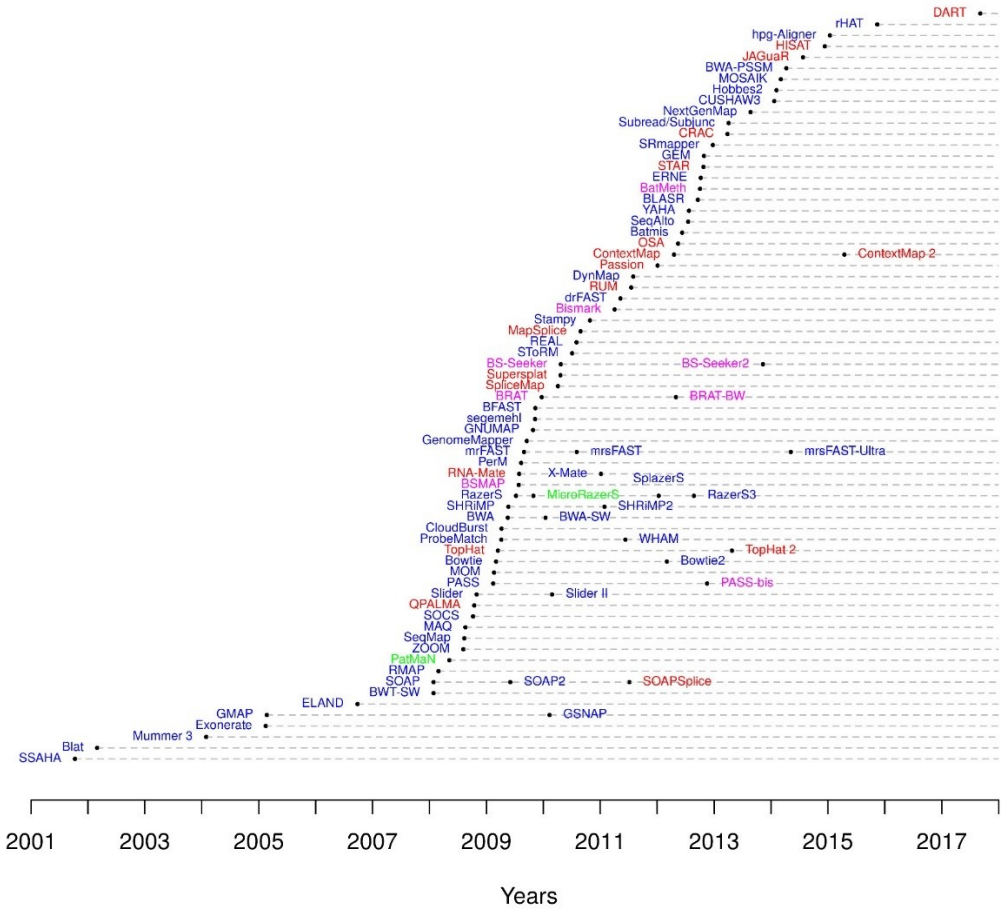


Figure 2.4: Timeline (up to 2017) with the publication date of different aligners (figure adapted from https://www.ebi.ac.uk/~nf/hts_mappers/).

Among all the listed aligners, TopHat2, Hisat2 and STAR are one of the most popular seeing how many times their reference article has been cited by other researchers. In a study comparing different alignment protocols in 76-nucleotide (nt) PE data and simulated data with different alignment tools, among them GSNAP, MapSplice, STAR, TopHat and GEM, it was shown that MapSplice was quite a conservative aligner respect to mismatch frequency, indel and exon junction calls, while aligners such as GSNAP and STAR returned many false exon junctions if the

Table 2.5: List of tools for RNA-seq data alignment to some reference. Table based on tool references and corresponding webpages.

Tool	Last Version*[1]	Strategy	Main Features
Unspliced aligners			
Bowtie (204)	v1.2.3	Burrows-Wheeler Transform (BWT)	Index the reference genome with a Burrows-Wheeler index based in the full-text minute-space (FM) index. Mainly for short read (25-50 bp) alignment.
BWA (205)	v0.7.17	Burrows-Wheeler Transform (BWT)	Need to construct first a FM index. It not considers base qualities when evaluating hits. Perform Smith-Waterman alignment for unmapped reads.
NovoAlign (206)	v3.032.04	Hashtable of k-mers	It uses Needleman-Wunsch algorithm and iteratively search the best alignment. Heuristics are used in calculation of alignment quality scores. Short read aligner that allows paired-end data.
SHRiMP (207)	v2.2.3	Smith-Waterman extension	There is not support since 2014, although the code remains available.
SOAP (208)	-	Seed and hash look-up table algorithms	Appears as temporarily unavailable.
Subread (209)	v2.0.0	Seed-and-vote	Use a large number of equi-spaced seeds (shorter than conventional ones), called subreads. All the seed in conjunction are used to obtain the optimal loci. Uses dynamic programming to complete the alignment.
Spliced aligners			
Bowtie2 (210)	v2.3.5.1	FM index (based on BWT)	Improved version of Bowtie that support gapped and paired-end alignments.
GEM3 (211)	v3.6	Custom FM-index and adaptative gapped search	Designed to obtain best results with alignment of long sequences (up to 1 Kb). Supports SE and PE modes. Also, supports both global alignment and local alignment models for different error models (i.e., hamming, edit, gap-affine).
GMAP (212,213)	2019-09-12	Segment chaining using genomic hash tables, a greedy match-and-extend algorithm using suffix arrays, hash tables, and nucleotide-level dynamic programming procedures.	Uses an oligomer chaining method that involves neither seeds nor extensions. It finds all matching 8-mers between the cDNA and genomic sequence, and then uses dynamic programming to find an optimal chain of 8-mers.
GSNAP (212,214)	2019-09-12	Segment chaining using genomic hash tables, a greedy match-and-extend algorithm using suffix arrays, hash tables, and nucleotide-level dynamic programming procedures.	From the same authors of GMAP. Can align both SE and PE reads as short as 14 nt and arbitrarily long length. Can detect splicing, including interchromosomal, using probabilistic models or known splice junction databases. Allows for bisulfite-treated DNA alignment.
HISAT2 (215)	v2.1.0	Hierarchical indexing (Burrows-Wheeler and FM index)	Same authors of TopHat. Uses an indexing scheme called Hierarchical Graph FM index (HGFM). It first does a short read alignment with bowtie2.

Table 2.5: (continued).

Tool	Last Version*[1]	Strategy	Main Features
MapSplice2 (216)	v2.2.1	Bayesian regression	First, it splits reads in shorter segments and align them with bowtie. Then, unmapped reads are aligned with gaps to infer splice junctions, with a quality score determined using a Bayesian regression. Supports PE and SE reads and can align reads of variable length.
QPALMA (217)	V0.9.3	Machine learning (uses extended Smith-Waterman algorithm)	Tool that exploits reads (including its quality information), splice site predictions, intron length and genome information. The only precondition is that there are genomic reads available that can be used to generate artificially spliced reads for training.
RazerS 3 (218)	v3.5.8	Based on counting q-grams	Searches for matches of reads with a percent identity above a given threshold, whereby it detects alignments with mismatches as well as gaps. Supports reads of arbitrary length with a large number of INDELs. An algorithm created with the sequencing of long fragments in mind, which usually have higher error rates.
RNASequel (219)	-	-	A tool that uses the spliced-read output of any aligner and *de novo* splice junction detection to perform an error-tolerant realignment. Recommended to use in combination with STAR by authors.
RUM (220)	V2.0.4	Burrows–Wheeler based algorithms, BLAT	Reads are first mapped with Bowtie against the genome and transcriptome. Then, the information is merged and unmapped reads are sent to BLAT. Finally, Bowtie ans BLAT alignments are merged. It can be used for DNA sequencing (e.g., ChIP-Seq) and microarray probe mapping.
STAR (221)	v2.7	Maximal Mappable Prefix (MMP)	The algorithm consists of two major steps: a seed searching step, in which a MMP is search for each read through an uncompressed suffix array, and a clustering, stitching and scoring step, in which alignments of the entire read are built by stitching all seed alignments trough a frugal dynamic algorithm. Recognised as fast aligner, but with high computational requirements.
TopHat2 (222)	v2.1.1	-	It is in a low maintenance, low support state as it is superseded by HISAT2, which has the same core functionality, but in a more accurate and efficient way (said by its own authors). It first does a short-read alignment with bowtie. Then, with unmapped reads, it detects potential splice sites. It uses these candidate splice sites in a subsequent step to correctly re-align multiexon-spanning reads with bowtie2. For PE data, the process is run separately for each end and, at the final stage, both reads are analysed together.

*[1] Last checked on 04/11/2019.

junctions were not filtered out by the number of supporting reads (223). In that study, TopHat2 and STAR were the fastest aligners, with STAR being the fastest one by a big difference. In other study, STAR, TopHat, GSNAP, RUM and MapSplice were compared in a simulated RNA-seq dataset and it was shown that all aligners exhibited desirable receiver operating characteristics (ROC) curves at high values of detection thresholds, while STAR exhibited the lowest false-positive rate at the lowest detection threshold of 1 read per junction (221). In addition, in the same study, those aligners were compared to a real experimental RNA-seq dataset and it was shown that STAR and GSNAP achieved a higher percentage of reads aligned to the reference and that all aligners shown similar sensitivities to annotated junctions, while STAR, RUM and TopHat2 performed similarly for unannotated junctions. In other study using both simulated data and real data comparing STAR, GSNAP, OLego, TopHat2 and HISAT , it was shown that tools using a two-pass strategy (They do a first alignment and update the junction splice site information for in a second run achieve a better alignment of reads with short anchors) such as STAR, HISAT and TopHat2 had better alignment sensitivities (224). Futhermore, HISAT and STAR aligned the greatest number of reads. Despite simulated data allow for precise calculations of false-positive and -negative rates, it must be pointed that comparisons of aligners in simulated data may not reflect real experimental errors, and such comparisons must be taken with caution.

2.3.2.3 Quantification

Once a good quality alignment has been achieved, being the only interest of the study annotated genes, gene quantification (at exon, transcript or gene level) must be performed. In the case of any interest in novel genes, then a transcriptome assembly step would be necessary prior to quantification. With the quantification of genes, new quality measures such as sequencing depth saturation, read distribution between different genomic features, principal component analysis (PCA) for biases in data and coverage uniformity would be available. The simplest way to calculate expression levels would be at the gene and exon level, where each read is counted as an expression unit of a gene/exon if the reads is concordant with the annotated coordinates of the molecule of interest. In this case, only those genes with some genomic overlap (or genes located in opposing strands in the same locus in unstranded libraries) would be hard to count for. Quantification at the transcript level is harder due to different isoforms of a gene having a high proportion of genomic overlap. There are multiple gene quantification tools, each with a different way of dealing with multimapping reads (some ignoring these reads, others dividing them equally or with a probabilistic model) or with reads that overlaps with more than one molecule or partially falls in intronic sequence. In figure 2.5 can be seen different strategies to count reads implemented in the HTSeq tool, while in table 2.6 can be seen a brief list of quantification specific tools (some transcriptome assemblers are able to perform quantification at gene and transcript level) and their main characteristics. Despite the modes of action depicted in figure 2.5 are exclusive of HTSeq, it shows typical problems that most quantification tools must deal with: reads that spans intronic and exonic sequence (probably from pre-mRNA products), reads overlapping multiple genes and multimapping reads, among the most common features. While some quantification tools returns raw reads and leave bias correction for normalization procedures or batch effect correction programs, other tools try to account for such biases with probabilistic models as can be seen in table 2.6.

In a recent paper, Sailfish, eXpress, Kallisto and RSEM were compared in simulated and real data and it was shown that all tools had similar accuracies, but Kallisto showed a higher performance on paralogs within a family and a higher run speed (225). In other study,

featureCounts and HTSeq-count were compared and it was shown that for SE data both tools returned quite identical counts, while for PE data there were more discrepancies mainly due to the way each tool handles those pairs that overlaps multiple genes (In the case that read1 aligns to gene A and gene B at the same time, but read2 only to gene B, feaureCounts assign that read to gene B, while HTSeq-count treats it as ambiguous) (226). In a more recent paper, seven different quantification tools (Cufflinks, RSEM, TIGAR2, eXpress, Sailfish, kallisto and Salmon) were compared in both experimental and simulated datasets, focusing on isoform quantification (227). It was shown that all methods, even with reasonable absolute abundances, have difficulties to quantify expression levels of isoforms whose relative abundance is low.

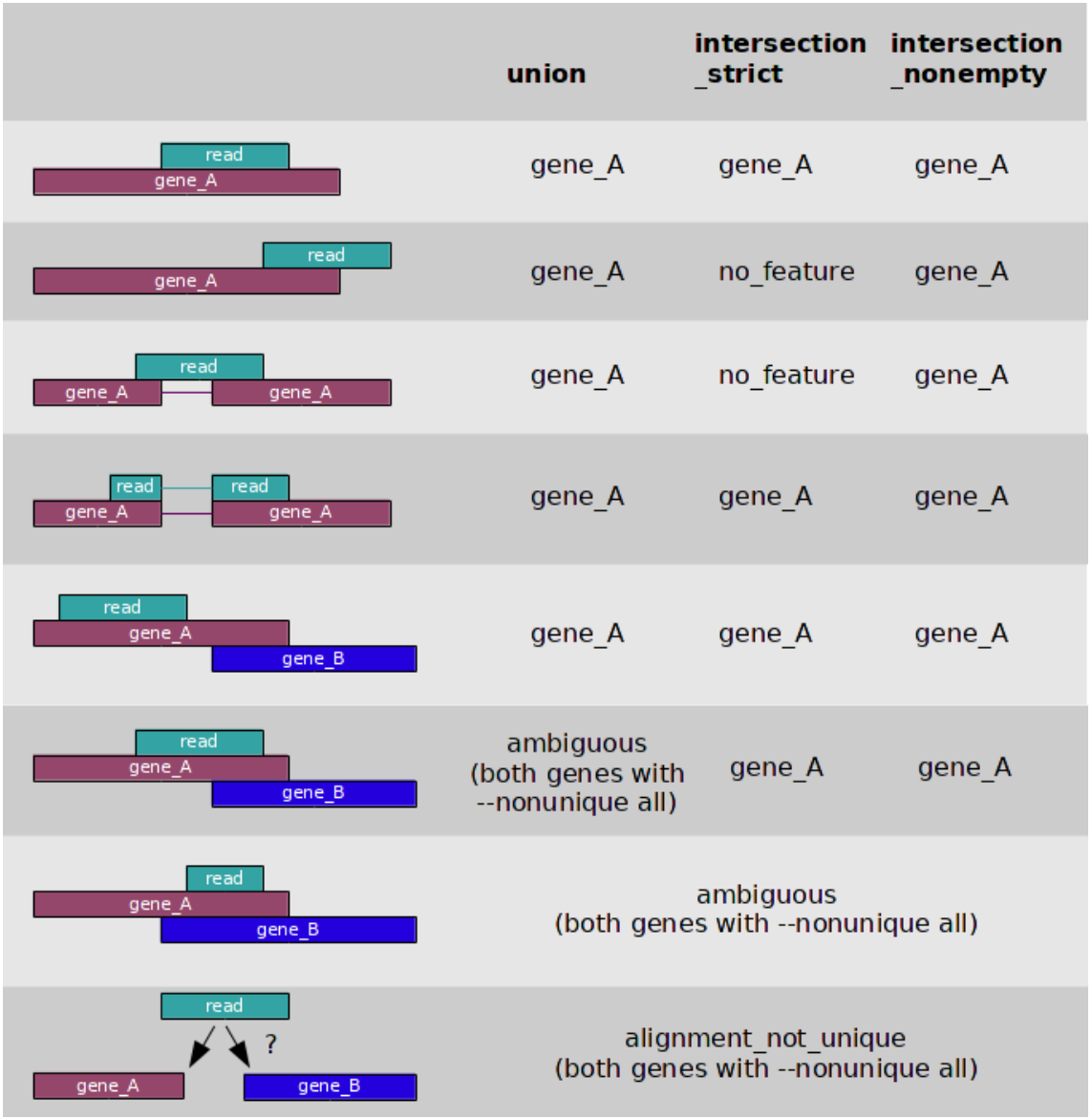


Figure 2.5: Quantification modes implemented in HTSeq. It can be seen to which feature (gene) a read (green box) is assigned to in three different modes (union, intersection_strict and intersection_nonempty) depending of the read alignment. In the union mode a read is always assigned to a gene if the full read falls upon a gene unambiguously (independent if part of the read falls in intronic sequence). In the intersection_strict mode, a read must fall completely into the sequence of a gene unambiguously (those with intronic sequence are discarded). The intersection_nonempty mode is quite similar to the previous one, but allowing for intronic sequence overlap (figure adapted from https://htseq.readthedocs.io/en/release_0.11.1/count.html).

Table 2.6: List of tools for RNA-seq data quantification. Table based on tool references and corresponding web pages.

Tool	Last Version*[1]	Main Features
Casper (228)	v2.20.0	Isoform level quantification for PE data through a Bayesian probabilistic model. Estimates probabilities of a read pair originating from an isoform by considering fragment length distribution and possible read non-uniformity coverage.
EQP-QM (229)	v2.1.0	Gene, exon or junction quantification. The contribution of a read to a feature count is based on the read weight which is the inverse of the genome alignments of the read.
eXpress (230)	v1.5.0	There is no longer development (last version 2017). Transcript-level quantification with a probabilistic assignment of ambiguously mapped sequences. Parameters for fragment length distribution, sequence bias and sequence read errors are used. Only known genes.
featureCounts (226)	v2.0.0	From the Subread packge, quantification for genomic features such as genes, exons, promoters and genomic bins (user defined). Flexible in counting multimapping reads and reads overlapping multiple features (They can be excluded, fully counted or fractionally counted). Allows for chimeric read counting. A minimum number of overlapping bases and a minimum faction of overlapping bases to assign a read to a feature can be defined.
HTSeq-count (186)	v0.11.1	Quantification to any given feature (see figure 2.5 for alignment modes).
Kallisto (225)	v0.46.1	Transcript-level quantification through read pseudoalignment. Only k-mer length and the mean of the fragment length distribution are required for quantification (estimated during run time). Only known genes.
RSEM (231)	v1.3.1	Gene and transcript quantification and credibility interval estimation. It aligns reads to features with Bowtie using parameters specifically chosen for quantification. After alignment, it computes maximum likelihood abundance estimates using the Expectation-Maximization algorithm for its statistical model. Only known genes.
Salmon (232)	v1.0.0	Transcript-level aquantification. It can do the alignment of reads to reference (quasi-mapping) or take the output of other alignments against the transcriptome (not genome). It employs a new dual-phase statistical inference procedure and sample-specific bias models that account for sequence-specific, fragment GC-content, and positional biases. Also accounts for 5'- and 3'-sequence-specific biases. Only known genes.
Seal (197)	v38.71	From BBMap package, an alignment-free quantification tool, based on which reference sequences share the most k-mers with the query. Can handle ambiguous reads.

*[1] Last checked on 06/11/2019.

2.3.2.4 Normalization

Once an expression table with all samples is obtained, data must be normalized to account for differences between libraries (different sequencing depths, transcript length, differences in GC content per sample or other biases such as non-uniform coverage). Multiple normalization methods have been developed, each one with different assumptions and objectives in mind. It must be pointed that some differential expression tools such as EdgeR and DESeq2 (in the following section described in more detail) expect raw counts and apply their own normalizations. The biases to normalize for can by classified as within-sample effects and between-sample effects. Within-sample effects refer to those features that affect comparison of read counts between different genes of a sample (e.g., length and GC content), while between-sample effects refer to those that affect comparisons of read counts of the same gene in different samples (e.g., sequencing depth). In this section some of the most used normalization methods will be described in more detail:

1. Read per kilobase million (RPKM) and Fragment per kilobase million (FPKM) (233):
   Both methods are quite similar, being RPKM for SE reads and FPKM for PE reads. In PE data, two reads may correspond to a single fragment and it should be counted once. These metrics try to normalize for differences in sequencing depth and gene length. Supposing that n_i (i=1,…,G) represents the number of reads (for SE data) or fragments (for PE data) aligned to gene *i* and that l_i represents the length of gene *i*, it is calculated as can be seen in formula (1):

   $$RPKM|FPKM_i = \frac{n_i}{l_i \times \sum_j n_j} \times 10^9 \qquad (1)$$

   It must be pointed that for comparison of a gene in different samples, as in a differential expression analysis, there is no need of a normalization that takes into account the feature length, since such characteristic would be the same in different samples.

2. Counts per million (CPM) (234):
   A normalization method similar to RPKM, but without a length normalization. It has been used in the EdgeR and limma differential expression packages. Using the same annotation, CPM is calculated as follow:

   $$CPM_i = \frac{n_i}{\sum_j n_j} \times 10^6 \qquad (2)$$

3. Transcripts per kilobase million (TPM) (235):
   It was pointed that RPKM was an inconsistent measure for between sample comparison, having the potential to cause inflated statistical significance values (235), and TPM was proposed as an alternative, which measures the relative abundance of transcripts.

   $$TPM_i = \frac{n_i}{l_i} \times \left(\frac{1}{\sum_j \frac{n_j}{l_j}} \right) \times 10^6 \qquad (3)$$

   With the development of new normalization methods, it has been noted that these “per million” methods are not optimal for differential expression analyses, since neither performs robust between-sample normalization. For example, TPM values are

dependent of the total number of different transcripts, which can vary greatly between samples, and thus makes it unsuitable for between sample comparisons.

4. Upper Quartile (UQ) (236):
   It has been shown that in RNA-seq experiments few highly expressed transcripts originate most of the sequencing reads and the rest of genes remain underrepresented. The UQ normalization was presented in an attempt to account for such effect. Using the same annotation as before and defining q_i^{75} as the 75th percentile of mapped reads of the genes in the sample after removing all genes with zero counts in all libraries and *n* as the total number of samples, UQ can be represented as:

$$UQ_i = \frac{n_i}{q_i^{75}} \times \frac{\sum_{j=1}^{n} q_j^{75}}{n} \qquad (4)$$

5. GC normalization procedures (237):
   It has been shown that GC-rich and GC-poor segments tend to be underrepresented in RNA-seq data and, in addition, such effects can be lane specific. Three different normalization procedures were proposed and implemented in the EDAseq Bioconductor R package to deal with such biases. Using the same annotation as before and defining gc_i as the GC content (proportion of G and C nucleotides) in gene i and n_i' as the normalized expression measure, and normalizing the logarithms of gene counts (since regression in the log-scale is more robust to outliers), these three normalization procedures are defined in EDAseq as:

   - Regression normalization:
   Gene counts, $log(n_i)$, are regressed on gc_i using a loess robust local regression and normalized expression values n_i' are achieved by shifting the residuals to recover scale of raw counts, i.e.,

$$n_i' = log(n_i) - \hat{n}_i + T(n_1, \dots, n_G) \qquad (5)$$

   where $\hat{n}_i$ represents fitted values and T is a summary statistic such as the median.

   - Global-scaling normalization:
   In this case genes are stratified into K equally-sized bins based on GC-content and is calculated as follow:

$$n_i' = log(n_i) - T\big(n_j : j \in k(i)\big) + T(n_1, \dots, n_G) \qquad (6)$$

   where k(i) represents the stratum to which gene i belongs.

   - Full-quantile normalization (FQ):
   Similar to global-scaling, genes are stratified based on GC content. The quantiles of read count distributions are then matched between bins, by sorting counts within bins and taking the median of quantiles across bins.

6. Trimmed mean of M values (TMM) (238):
   It has been pointed that expression levels are not only dependent on depth and length; it can be affected by the composition of the RNA population that is being sequenced. Thus, if a gene is exclusively expressed (or highly expressed) in one condition, the rest of genes remains underrepresented as a consequence and, if such effect is not

corrected, the DE can be skewed towards one condition. For such bias correction, the TMM normalization was presented and implemented in the EdgeR Bioconductor R package. In this normalization method, scaling factors across several samples are estimated taking one sample as reference and calculating TMM factors for each non-reference sample. Those factors can be used later in downstream statistical analyses. This strategy equates the overall expression levels of genes between samples under the assumption that the majority of them are not differentially expressed. Suppose that Y_{ij} represents the observed count value for gene i (i=1,…,G) and sample j (j=1,…,n), N_j the total number of reads in sample j and the sample r ∈ {1,…,n} is taken as reference. We define the gene-wise log-fold-change and absolute expression level respectively as:

$$M_i = log_2 \frac{Y_{ij}/N_j}{Y_{ij'}/N_{j'}} \qquad (7)$$

$$A_i = \frac{1}{2} log_2 (Y_{ij}/N_j \times Y_{ij'}/N_{j'}) \text{ for } Y_{i\cdot} \neq o \qquad (8)$$

Then, the TMM_i value is calculated taking a weighted mean of M values after removing the upper and lower 30% and 5% (those values can be adapted) of the M_i and A_i values, respectively. If G* represents the set of genes that are not trimmed, then:

$$log_2 \left(TMM_j^{(r)} \right) = \frac{\sum_{g \in G^*} w_{gj}^{(r)} M_{gj}^{(r)}}{\sum_{g \in G^*} w_{gj}^{(r)}}$$

$$\text{where } M_{gj}^{(r)} = \frac{log_2(Y_{gj}/N_j)}{log_2(Y_{gr}/N_r)} \text{ and } w_{gj}^{(r)} = \frac{N_j - Y_{gj}}{N_j Y_{gj}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}} \qquad (9)$$

7. Median of ratios (239):
   Other method that account for RNA composition is the one called median of ratios, which is implemented in the DESeq2 Bioconductor R package. Similar to TMM values, these values are calculated under the assumption that the majority of genes are not differentially expressed. Such size factors are calculated taking the median ratio of gene counts relative to the geometric mean per gene across all samples:

$$s_j = \underset{i}{median} \frac{Y_{ij}}{(\prod_{v=1}^{n} Y_{iv})^{1/m}} \qquad (10)$$

In different studies comparing normalization methods, it has been noted that “per million” methods such as RPKM/FPKM perform poorly and several authors recommends to avoid their use for DE analyses (240). In a study comparing UQ, median of ratios, TMM and RPKM, among others, it was shown in simulated data that only TMM and median of ratios were able to maintain a reasonable false-positive rate without loss in power in the presence of highly expressed genes (241). Although it must be taken into account that those two methods are constructed under the assumption that most genes are non-DE, and that for those DE there is a balanced proportion of over- and under-expression. In other study, despite similarly showing that the median of ratios normalization performed satisfactorily, contradictory results in comparison to the previous study were achieved regarding TMM, in which the majority of studied cases showed a poor display (242). The differences in results may be explained by the chosen metrics for performance quality or the selected datasets (some of them being simulated) for comparisons. It is clear that there is no global normalization method applicable to all datasets and they must be selected with caution, since incorrect normalization can result in inflated false positives in DE analyses.

2.3.2.5 Batch effect removal

Apart from the normalization methods described in the previous section to control for the variation caused by different artifacts (e.g., length, depth, RNA composition, …), there is sometimes the need of assessment and adjustment of technical sources of variation of unknown origin (in some cases it is exactly known and can be inserted directly into the model), also called batch effects. These technical effects can be complex to deal with if they are correlated with the biological factor under study. Thus, to avoid or soften such batches, it is recommended a good experimental design in which the batch, in case of not being able to avoid it, is evenly distributed in all groups. There are mainly two ways to correct these effects: 1. If the batch is known or estimated (e.g., through a PCA analysis), adding it to the statistical model (there are some tools that do not allow for such procedure); or 2. Estimate and remove the batch effect, creating a batch-free data, and then, perform the statistical analysis on the corrected data. The problem with this methodology is in the presence of unbalanced data (i.e., when the batch is not equally represented in all studied groups), since in that case group differences and batch effects are interdependent, and it would be needed to check if the batch removal does not introduce another batch in the data (or if the introduced batch has a lesser effect and represents an improvement over the not corrected data) (243). In table 2.7 can be seen a brief list of some batch-effect removal tools.

To better understand what some of these principal components-based tools do, a brief explanation of SVA (Surrogate Variable Analysis) will be given in more detail. The idea behind it is to model the data as a combination of known variables of interests, known batches and unknown and unmeasured batch effects (244). Supposing that Y_{ij} represents the expression for gene i (i=1,…,G) and sample j (j=1,…,n), y_j the phenotype of sample j, a_j known batch variables of sample j and u_j unknown batch variables of sample j, a simple model with only one variable in each category would be:

$$\underbrace{Y_{ij}}_{gene\ expression} = \underbrace{b_{i0}}_{baseline} + \underbrace{b_{i1}y_j}_{phenotype\ effect} + \underbrace{c_i a_j}_{known\ batch} + \underbrace{d_i u_j}_{unknown\ batch} + \underbrace{e_{ij}}_{mean\ error} \quad (11)$$

As it is now, it is hard to estimate unknown batches from the data directly with so many parameters, but knowing that it must exists a subset of genes with no DE where $b_i=c_i=0$, the formula would be reduced to:

$$\underbrace{Y_{ij}}_{gene\ expression} = \underbrace{b_{i0}}_{baseline} + \underbrace{d_i u_j}_{unknown\ batch} + \underbrace{e_{ij}}_{mean\ error}$$

It must be pointed that it is not necessary to exactly known d_i and u_j to make statistical inferences about b_{i1}, just to known their linear combination $d_i \times u_j$ is enough. If all data from non-DE genes were collected and their mean extracted to remove the baseline effect, the matrix formula would have the following form (m_u is the number of genes were $b_i=c_i=0$):

$$\underbrace{G}_{m_u \times n} = \underbrace{\vec{d}}_{m_u \times 1}\ \underbrace{\vec{u}}_{1 \times n} + \underbrace{E}_{m_u \times n}$$

To this expression can be applied matrix decompositions like single value decomposition or PCA to estimate unknown batches in this subset of data and if this subset of data is large enough, consistent estimations for the more general u_j can be made.

Table 2.7: Tools related to batch effect removal for RNA-seq data. Table based on tool references.

Tool	Last Version*[1]	Main Features
ComBat (245)	v3.34.0	Tool for directly removing known batch effects from the SVA R package. It uses an empirical Bayesian framework (by default a parametric one, though a nonparametric Bayesian adjustment is also available). Initially developed for microarray data, extended for RNA-seq. Assumes that phenomena resulting in batch effects often affect many genes in similar ways.
gPCA (246)	v1.0	An extension of a regular PCA to quantify the existence of batch effects.
Harman (247)	v1.14.0	Tool for removing unknown batch effects. It can process multiple different datasets (e.g., microarray, RNA-seq, methylation). First, the tool separates the data in its principal components and scans each component for variance arising from batches. Then, it removes the detected batches under a tolerance level defined by the user. Finally, the principal components after batch removal are recombined and transformed in the original dataset format, which can be used in downstream analyses.
removeBatchEffect (248)	v3.42.0	Function from the limma differential expression tool for known batch effect removal. The function fits a linear model, including treatment and batch variables, and remove the components due to the batch.
RUVSeq (249)	v1.20.0	Tool for removing unknown batch effects by performing factor analysis. To estimate the factors, it needs a set of genes assumed not to be influenced by the covariates of interest (e.g., housekeeping genes, spike-in controls). If such set of genes are not known *a priori*, the less significantly DE genes from a first-pass DE analysis can be used instead.
svaseq (244,250,251)	v3.34.0	Tool for removing unknown batch effects from the SVA R package. It estimates the batches by principal component or singular vectors. Then, the resulting components can be used in downstream analyses to correct the batch effect.

*[1] Last checked on 11/11/2019.

To sum it up, these methods try to remove noise caused by technical artifacts (e.g., processing samples on different days, post-processing effects, samples sequenced at different lanes) at the risk of also removing the biological signal. Extra caution must be taken with unbalanced designs (independent of sample size), since it could lead to over-confidence in the results (243). In a recent study comparing some of these normalization approaches in simulated data, methods such as SVA, RUVSeq and PCA, it was shown that SVA outperformed other methods by correctly estimating the number of batch effects (252). In other study comparing Harman and ComBat, it was shown that there was always a setting for Harman with better noise rejection and signal preservation than ComBat (247).

2.3.2.6 Differential Expression (DE) analysis

Once a normalized expression matrix is achieved (or raw counts in some tools for DE analysis, since they apply their own normalization), the next step would be to identify genes that are differently expressed in distinct group of samples (e.g., healthy vs. diseased, control vs. treatment, at different time points, different tissues), in an attempt to find relevant genes to the biological question under study. From now on we will always use the word "genes" when describing the differential expression analysis to simplify, but it should be borne in mind that such analyses can be performed on other genomic features such as transcripts and exons. In this type of analysis, thousands of genes from few samples in different groups are compared against each other. Such a problem dimensionality, where the expression of thousand genes are recorded in few samples, makes it hard to fit a statistical model that takes into account the expression of all genes at the same time (which would make sense, since the expression of one gene would be correlated or dependent of others in the same pathway or family or to other regulatory elements). At the end, DE analyses are normally performed for each gene independently, in a univariate way, and a multiple testing correction procedure is applied.

When analysing RNA-seq data, it must be pointed that integer counts are being studied, a discrete variable, which would be totally different to microarray data where continuous intensity levels are recorded. When RNA-seq appeared for the first time, a great number of laboratories performed DE analysis by transforming the data to a continuous range (e.g., log transformation, *voom* function in limma) and using tools developed for microarray data, which usually requires a normal distribution. Later, specific tools for RNA-seq using discrete distributions such as Poisson and negative binomial or using non-parametric methods were developed. The choice of method would greatly depend on the number of biological replicates at hand. Non-parametric methods do not make any assumption about the statistical distribution, but they need multiple biological replicates per group (at least 5-10 samples per group) to have enough statistical power (83). In most cases, due to the high cost of RNA-seq among others, most studies are limited to a low number of biological replicates per group (usually in the range of 3-4 samples per group), in which non-parametric methods are underpowered and DE tools such as EdgeR and DESeq2 that use a negative binomial distribution are preferred. In table 2.8 can be seen a brief list of DE tools with their main characteristics.

Despite being a great variety of tools for DE analysis, each one based on different distributions such as Poisson, negative binomial or non-parametric ones, the majority of laboratories have chosen EdgeR and DESeq2 as their main DE tools. These tools can be installed as R packages from Bioconductor. It is important to highlight that both tools are built under the assumption that most genes are not differentially expressed. Both tools follow a similar philosophy regarding the distributional choice, being their major differences: DESeq2 uses raw counts and models normalization, while EdgeR first normalize the data and then fits the model; the chosen normalization method (TMM in EdgeR and median of ratios in DESeq2); and DESeq2 performs independent filtering. Since both tools are going to be used throughout this thesis, a brief explanation of their statistical model will be given in more detail. For a more detailed description, it is recommended their principal papers (253,254) and online manuals in Bioconductor (https://bioconductor.org/). Supposing that Y_{ij} represent the expression level of gene i $(i = 1, \cdots, G)$ for sample j $(j = 1, \cdots, n)$, EdgeR uses for differential expression analysis the following generalized linear model:

Table 2.8: List of tools for differential expression analysis of RNA-seq data.

Tool	Last Version*[1]	Approach	Main Features
ABSSeq (255)	v1.41.0	Model counts differences between conditions through a Negative Binomial model.	Can analyse complex experimental designs. Allows the calculation of fold change shrinkage to facilitate gene ranking and outlier detection. The authors affirm to be a robust method against outliers compared to other methods.
Ballgown (256,257)	v2.18.0	Log transformation and linear modelling.	Allows for differential expression at the isoform level. The default parameters assume a modest sample size. Can analyse complex experimental designs such as time-course experiments. Similar to limma strategy.
baySeq (258)	v2.20.0	Negative Binomial model. Estimation of posterior probability through an empirical Bayes approach.	Can analyse complex experimental designs, but not paired sample analysis. Computationally intensive, but takes advantage of parallel processing. Shows improvements in performance for large numbers of libraries compared to EdgeR.
BitSeq (259)	v1.30.0	Bayesian inference for expression levels. Log-normal model of the estimates used to infer mean expression and ranking transcripts based on the likelihood of DE.	Differential expression at the isoform level. Computationally intensive tool. Only pairwise comparisons. Incorporates RPKM/FPKM normalization.
CuffDiff2 (260)	v2.2.1	Isoform deconvolution + count-based test	Differential expression at the gene and isoform level. Controls for both variability across replicates and uncertainty in abundance expression estimates caused by ambiguously mapped reads by using a model for fragment counts based on the beta negative binomial distribution. Only pairwise comparisons.

*[1] Last checked on 12/11/2019.

Table 2.8: (Continued)

Tool	Last Version*[1]	Approach	Main Features
DESeq2 (254)	v1.26.0	Negative Binomial generalized linear model.	Differential expression at gene level. Expects un-normalized counts, since its model internally corrects for library size. It performs independent filtering by default using the mean of normalized counts as filter statistic. Can analyse complex experimental designs.
DEXSeq (261)	v1.32.0	Negative Binomial generalized linear model.	Differential expression at exon level. When fitting a model for an exon, it sums up the counts from all the other exon and use only the total, rather than the individual counts in the model. It uses the same normalization method as DESeq2. Can analyse complex experimental designs. Approach similar to DESeq2.
EBSeq (262)	v1.26.0	Negative Binomial distribution and empirical Bayes model.	Differential expression at the gene and isoform level. Allows comparisons of multiple conditions. Authors point to be strong when outliers are present.
EdgeR (253,263)	v3.28.0	Negative Binomial generalized linear model.	Differential expression at pre-defined genomic features (preferentially non-overlapping ones). Methodology similar to DESeq2, but EdgeR uses internally TMM normalization. Can analyse complex experimental designs. Can be used for differential methylation analysis.
GPseq (264)	-	Two-parameter Generalized Poisson distribution.	Not updated since 2011. Differential expression at the gene and exon level. Can analyse complex experimental designs.
Limma (234,248)	v3.42.0	Linear models of continuous data.	Differential expression at the gene and isoform level. Originally developed for microarray data. Needs to process expression matrix to continuous values (log2(CPM) values). Then, the mean-variance relationship is modelled either with precision weights (“voom” method) or with an empirical Bayes prior trend (“limma-trend” method). Can analyse complex experimental designs.

*[1] Last checked on 12/11/2019.

Table 2.8: (Continued)

Tool	Last Version*[1]	Approach	Main Features
NBPSeq (265)	v0.3.0	Over-parameterized version of the negative binomial model.	Differential expression at the gene level. Only two-group comparisons.
NOISeqBIO (266)	v2.30.0	Non-parametric test. Implements an empirical Bayes approach to improve the handling of variability specific to each gene.	Differential expression at pre-defined genomic features. Only two-group comparisons. Allow application of external normalization procedures. It returns a DE probability that is equivalent to FDR adjusted P-values. It has implemented an optional batch effect removal tool denominated ARSyNseq. It has a great variety of plots for bias detection.
SAMseq (267)	v3.0	Non-parametric test. Uses the rank of the expression values in a Wilcoxon statistic.	Can analyse complex experimental designs. With moderate sample size, this non-parametric method gives competitive results comparable to popular parametric methods. While parametric models can suffer at the presence of outliers, certain amount of outliers barely hurt the performance of this non-parametric method.
TSPM (268)	-	Two-stage Poisson model. Relies on likelihood ratio and Pearson test statistics that have no exact distributions under a Poisson model. Therefore, it is relied on the fact that both follow asymptotic $\chi 2$ distributions.	Can analyse complex experimental designs. It should only be used when there are at least six degrees of freedom to estimate dispersion. Otherwise, the TSPM may provide over-estimates of significance. Allow application of external normalization procedures.
tweeDESeq (269)	v1.32.0	Poisson-Tweedie distribution family.	Differential expression at pre-defined genomic features. Only two-group comparisons. Takes advantage of parallel processing.

*[1] Last checked on 13/11/2019.

$$Y_{ij} \sim NB(\mu_{ij}, \alpha_i)$$
$$\mu_{ij} = N_j \pi_{ij} \qquad (12)$$
$$log\ \mu_{ij} = x_j^T \beta_i + log N_j$$

where Y_{ij} counts are modelled using a negative binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. π_{ij} represents the fraction of fragments in sample j that originates from gene i ($\sum_{i=1}^{G} \pi_{ij} = 1$), while N_j denotes the total of aligned reads to sample j. x_j is a vector of covariates that specifies the treatment conditions of sample j, while β_i is a vector of regression coefficients that gives the log2 changes for gene i for the corresponding covariate. The dispersion parameter defines the relationship between the variance and mean value:

$$Var(Y_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2 \qquad (13)$$

Then, the coefficient of variation (standard deviation divided by the mean) would be:

$$CV^2 = {}^{1}/_{\mu_{ij}} + \alpha_i$$

The dispersion parameters are estimated using expected mean values from the maximum likelihood estimate and the Cox-Reid profile-adjusted likelihood method. EdgeR has the option to specify how to calculate dispersion: a common dispersion for all genes, a trended dispersion dependent of gene expression values, or individual dispersions for each gene.

In a study comparing EdgeR, DESeq, baySeq and TSPM, it was shown that EdgeR and DESeq performed similarly to baySeq, which was the one with best performance in term of ranking genes for smaller FDR values, while TSPM performed poorly in comparison when the sample size was small (270). In other study comparing EdgeR, DESeq and NBPSeq, it was shown that the performance was dependent of the number of samples available (271). DESeq was conservative in all scenarios, EdgeR overestimated DE detection for small sample scenarios and NBPSeq overestimated detection of DE in all scenarios. In (272) eight different DE tools (Cuffdiff2, SAMseq, baySeq, EBSeq, limma, NOIseq, DESeq and EdgeR) were compared and it was shown that under small numbers of replicates per group, DESeq and limma were the most conservative approaches, EdgeR showed variable results depending of the dataset and SAMseq had low power. It must be pointed that such comparisons have been done with early versions of these methods and since then, methods such as EdgeR, DESeq and limma have been updated. In a more recent study comparing different DE tools (baySeq, DESeq2, EBSeq, EdgeR, limma+voom, NOIseq, SAMseq and sleuth), it was shown that NOIseq, limma+voom and DESeq2 were the most balanced ones regarding precision, accuracy and sensitivity and that a combination of multiple tools produce more precise and accurate results (273). There is no optimal DE tool under all circumstances, the choice would greatly depend of different characteristics of the data and the experimental design.

2.3.2.7 Gene Set Enrichment

The differential expression analysis usually ends with a large list of genes that must be interpreted by researchers to achieve biological meaningful results about the treatment or disease under study. Since the simultaneous interpretation of hundred DE genes is not affordable by the researcher itself, they are compared against databases that group genes into functional categories, in an attempt to find overrepresented biological functions that may

underlie the differences between condition in the data. Two of the most used databases are Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). GO database clusters genes in different biological functions with respect to three main domains: molecular function (molecular-level activities performed by gene products), cellular component (the location relative to cellular structures at which a gene performs its function) and biological process (large processes accomplished by multiple molecular activities). In contrast, the KEGG database is a collection of manually drawn pathway maps that links genes in a genome to gene products in complex networks of molecular interactions and reactions.

In an enrichment analysis is tested if in the DE gene list there is more genes belonging to a functional category or pathway than one would expect from a random sample of all the "universe" or background (e.g, in an RNA-seq experiment, the "universe" refers to all expressed genes in the samples). There are multiple webpages and standalone tools, such as DAVID tools (274), PANTHER (275) and g:Profiler (276), that use GO term and KEGG pathway information (apart from their own databases) for the functional enrichment. The main difference of these tools is how they test for gene enrichment. In DAVID, a modified Fisher Exact p-value (EASE score) is used to measure gene-enrichment in annotation terms. In contrast, PANTHER uses the binomial test for each functional term to determine if there is overrepresentation or underrepresentation. In addition, g:Profiler uses the cumulative hypergeometric test. All the tools correct obtained p-values by multiple testing correction techniques (Bonferroni, Benjamini and FDR).

When realizing such analysis, it must be taken into account different biases in an RNA-seq data analysis. The transcriptome of a tissue is highly specialized, so it is obvious that if all annotated genes are used as background, the enrichment will be biased towards the functional role of the tissue itself and it will not tell much about the DE list (277). To avoid this bias, it is always recommended to use all sequenced genes as background, although it stills not reflect the true "universe" due to RNA-seq inherent bias (e.g., depending of read depth, some lowly expressed transcripts could not be sequenced). Furthermore, enrichment tests based on the Fisher's exact test tend to select as significant large pathways, so it is recommended to put an upper limit of gene set size (278). In addition, non-coding molecules such as lncRNA, circRNAs and miRNAs lack any functional annotation and they must be studied separately. Finally, some enrichment analysis methods assume statistical independence among genes, something unrealistic since most genes in a pathway or term may be co-expressed. Thus, standard FDRs tend to be more or less conservative than expected, although they can still be used for exploratory analysis and hypothesis generation (278).

2.3.2.8 *Weighted correlation network analysis (WGCNA)*

As it has been mentioned before, genes belonging to a pathway are usually co-expressed. In an attempt to address the correlation patterns among genes, correlation networks can be constructed. In these networks the genes are represented by nodes and the edges that join the nodes will be calculated based on a measure that quantifies the strength of the co-expression between two nodes. There are multiple tools using different measures and methods. In this section the Weighted Correlation Network Analysis (WGCNA) R package (279) and some general terminology will be explained in detail, since it will be used later in the data presented in this thesis.

A network can be fully described by an adjacency matrix a_{ij}, a symmetric n x n matrix (if n is the total of genes sequenced) with values in [0,1] in which a_{ij} represents the connection

strength between nodes i and j. Different measures can be used to represent the connection strength. In WGCNA, the default measure is the absolute value of the correlation coefficient (Pearson correlation) between two gene expression profiles. The package gives the option to use alternative co-expression measures such as Spearman correlation and biweight midcorrelation (also called bicor). The bicor correlation is a median-based measure, which is less sensitive to the presence of outliers, and it has been shown that bicor coupled to topological overlap matrix transformation leads to more significantly enriched co-expression modules (280). Thus, if x_i represent the expression profile of a gene in each sample, a weighted network adjacency can be defined as:

$$a_{ij} = \left|bicor(x_i, x_j)\right|^{\beta}$$

with β>1. The parameter β is chosen in a way that the constructed network has a scale-free topology, i.e. a network whose degree distribution follows a power law, which translates in some nodes having a great number of connections to other nodes, while most of the nodes have only few connections. The choice of a scale-free network is based in the fact that Barabási and colleagues (281) showed that various protein-interaction networks of cells and a cellular metabolic network of 43 different organisms had a scale-free structure. And the more networks studied, the more scale-free topologies are discovered.

Once the network has been constructed, clusters of interconnected genes (modules) are detected. Different metrics can be used for gene interconnectedness, but WGCNA uses the topological overlap measure (TOM):

$$TOM_{ij} = \begin{cases} \dfrac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{min\{\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{ju}\} + 1 - a_{ij}} & if\ i \neq j \\ 1 \quad if\ i = j \end{cases}$$

Then, WGCNA identifies modules by unsupervised clustering. At this step, the user must select multiple parameters such as the minimum cluster size or a criterion to merge clusters that are quite similar. Finally, it can be checked if there is any module related to the treatment or variable of interest. This can be achieved by seeing how correlated are the variable of interest and the eigengene (which is a gene expression representing all the module and calculated from the first component of a PCA).

Other terms that need to be described for later are module membership (MM), gene significance (GS) and hub genes. Module membership is the correlation between a gene and the eigengene of the module in which it is clustered in. Gene significance refers to the absolute value of the correlation between gene and the trait of interest. Finally, hub genes are defined as those with a high correlation/connectivity in each module and they are supposed to play key roles in the module. There are multiple ways to define hub genes, depending of the objectives and interest of the study. In this thesis, hub genes are defined as those belonging to the ≥85th percentile for both MM and GS in each module (282).

2.3.2.9 Workflow

In figure 2.6 can be seen the workflow used to analyse the data in this thesis. Briefly, a quality check was performed on the raw data files with FASTQC [v0.11.5] (182) to assess the most appropriate read quality filtering and trimming. The following criteria were used with Trimmomatic [v0.36] (202): (1) remove adaptor sequences with the “palindrome” mode for paired-end data, allowing up to two mismatches (It must be pointed that few adaptor sequences

are expected in total RNA-seq data, since read length is generally shorter than fragment length and, as a consequence, it is uncommon that the sequencer reads through the adaptor sequence); (2) remove reads in which the average Phred quality score within a sliding window of five nucleotides fall below 20; and (3) remove reads with a length <36 nucleotides. The data was checked again with FASTQC to ensure that the filtering was adequate.

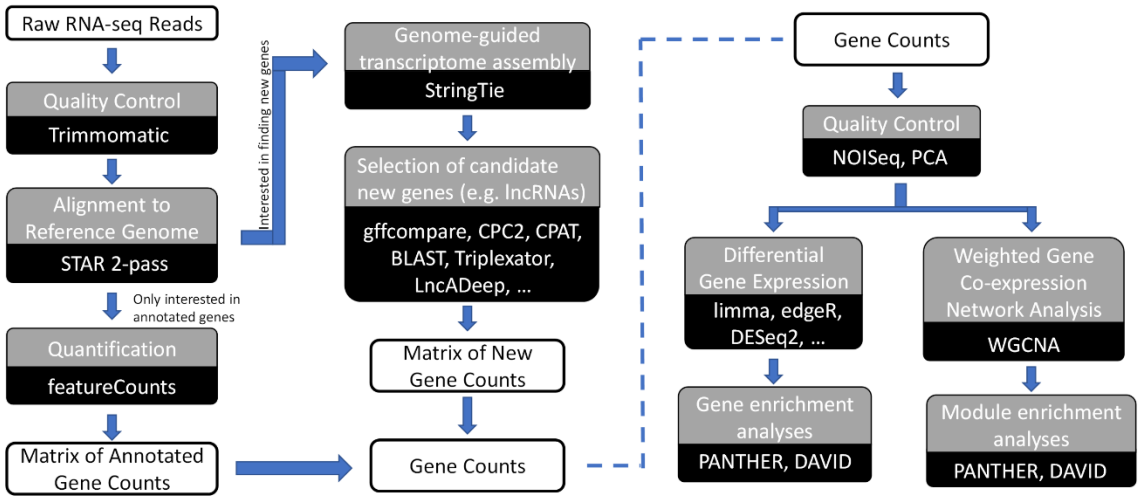


Figure 2.6: Pipeline followed for total RNA-seq data analysis.

Then, the quality checked reads were aligned with the STAR algorithm [v2.5.2b PBMCs and v2.5.4a encephalon] (221) against the *Ovis aries* genome build Oar3.1 from Ensembl [version 89.31] (97) using the 2-pass mode. In this mode, STAR does a first alignment to the reference genome and uses the new detected junction information in all the samples to update the archive of known junctions, to then make a second alignment using that information. Despite this 2-pass strategy is computationally demanding (since each samples is aligned twice with a program known for its high computational requirements), it allows for a better alignment of spliced reads, which would be necessary with an organism such as sheep whose reference genome is still not "complete" or with a quality similar to human or mouse. Once the alignment has ended, the featureCounts software from the SourceForge Subread package [v1.5.0-p1 PBMCs and v1.6.0 encephalon] (226) was applied to each library to assign uniquely aligned fragments to annotated genes in a strand-specific manner. Then, the expression levels were evaluated by a set of plots from the NOISeq package [v2.20.0] (266) and by a principal component analysis (PCA) to detect potential biases and contamination.

Although in this thesis is not going to be deepened, since it is still a work in progress, apart from annotated genes, one of the interests of our research group is to use total RNA data to find new lncRNAs and study their function in sheep. This would be done mainly for two reasons: (1) The majority of annotated lncRNAs in sheep are predicted and with unknown function, so it would improve greatly the sheep annotation; (2) Until now, there is no study that has tried to elucidate if lncRNAs have any role in aluminium adjuvancy. For that purpose, an additional step after mapping was necessary. The StringTie [v1.3.3b] (283) transcriptome assembler was used to reconstruct the transcriptome from the previous mapping. Briefly, the StringTie algorithm was run on each sample with the reference annotation from Ensembl and, in order to obtain a non-redundant set of transcripts, the –merge option was applied after the assembly to all samples. Then, StringTie was once again applied on each sample, but with the GTF transcript file obtained in the previous step in order to estimate transcript abundances.

From this assembly, only candidate lncRNAs were selected and their counts were added to the count matrix of annotated genes. The identification and functional annotation of lncRNAs is still in progress and it would be carried out and presented in another thesis of a fellow PhD student.

Prior to the differential expression, the SVA package [v3.24.0 PBMCs and v3.26.0 encephalon] (251) 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 to check how the batch effect affected the data. Once batch effect bias was dealt with, differential gene expression analysis was performed using three different R packages within Bioconductor: edgeR [v3.18.1] (263), DESeq2 [v1.16.1 PBMCs and v1.18.1 encephalon] (254) and limma+voom [v3.32.2] (234). It must be pointed that not all the named tools were applied in both tissues. In PBMCs samples, due to the higher variability in gene expression, it was decided to apply the three tools and take the intersection as true differentially expressed genes, in an attempt to remove false positives from the results. In contrast, in encephalon samples, only the DESeq2 package was applied. The DESeq2 package performs independent filtering, but for edgeR and limma packages a cut-off for filtering lowly expressed genes was set at 2 CPM. A not too strict cut-off that eliminated a large number of genes that were barely expressed.

RNA-seq counts were modelled by different generalized linear models in both tissues. In the case of PBMCs samples, since the experiment was carried as a time-series with two time points (samples at the start of the experiment before any vaccination (T0) and samples at the end after all vaccinations (Tf)), it must be kept in mind that not all samples are collected from independent subjects and it must be dealt with when fitting the model. The following variables were used in the model of PBMCs samples: *time* (T0 or Tf), *treatment* (commercial vaccine group [Vac] or adjuvant alone group [Adj]), *sample* (variable indicating the samples that comes from the same individual), and *batch* (surrogate variables calculated by svaseq from SVA package). The model included the *treatment* factor, the *batch* variable and the interactions *treatment x sample* (since there are different animals in each treatment) and *treatment x time* (to account for the treatment-specific time effects). Differential expression analyses were performed for the time points, considering the treatment group (Vac Tf vs. Vac T0 and Adj Tf vs. Adj T0), and for the treatments at the end of the experiment (Adj Tf vs. Vac Tf). In the case of encephalon samples, only samples at the end of experiment were analysed, so the differential expression analysis was performed using the following variables in the model: *treatment* (Control samples, Vac group and Adj group) and *batch* (surrogate variables calculated by svaseq from SVA package). Three different comparisons were made (Adj vs. Control, Vac vs. Control and Adj vs. Vac). In both tissues, the differentially expressed genes (DEGs) were selected as those with an adjusted p-value (using the Benjamini-Hochberg method) threshold of<0.05 and a fold change value of>1.5 or <0.667. In the specific case of PBMCs samples, only those genes that were identified as DEGs by all of the three programs were selected for further analysis.

To search for overrepresented gene functions in the lists of DEGs, gene enrichment analyses were conducted using the Gene Ontology (GO) database with PANTHER [v12.0] (275) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database with DAVID [v6.8] (274). Enriched terms were considered statistically significant with an adjusted p-value threshold of <0.05.

In addition to the DE analysis, a weighted gene co-expression network analysis was performed using the WGCNA [v1.63] (279) 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 or approximate to a scale-free topology network (R2>0.8). 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 ≥85th percentile for both MM and GS in each module (282). Those genes are likely 'key drivers' and might play important roles in the treatment.

2.3.3 miRNA-seq differential expression

The procedure to analyse the miRNAs is similar to the one employed for total RNA-seq, with a lot of programs in common. The process begins with separate files for each sample (in FASTQ format) in which the sequenced fragments are recorded. In contrast to total RNA-seq libraries, since mature miRNAs are ~22 nt long, miRNA-seq libraries are composed of SE 50 nt long reads. Briefly, for a differential expression (DE) analysis, the main scheme would be composed of the following steps:

1. Quality control and pre-processing of reads.

2. Alignment of reads to a reference genome. Since mature miRNAs are shorter than read length, they are expected to be fully sequenced in a single read. As a consequence, an unspliced aligner such as bowtie is enough for alignment to the reference.

3. Novel miRNA discovery.

4. Quantification of miRNA expression levels.

5. Differential expression analysis.

6. Target gene prediction.

7. Gene set enrichment analysis.

In this section, a brief description of the major steps for a DE analysis of miRNA data will be provided, with a brief summary of the different programs that can be used in each step. Since most of the steps (1,2,4,5 and 7) allow the use of similar methods to the ones seen in total RNA-seq (which can be reviewed in the corresponding section), only miRNA-seq specific steps will be studied in more detail. Thus, novel miRNA discovery, target gene prediction tools and miRNA-mRNA correlation analysis will be briefly explained. At the end, the exact workflow used in the data presented in this thesis will be provided.

2.3.3.1 Novel miRNA discovery

Most of the tools for novel miRNA discovery are wrapper tools that also includes: unspliced alignment to a reference with public available tools such as bowtie or with their own strategy; annotation of candidate miRNAs based on a miRNA database from different organisms (since miRNAs are highly evolutionary conserved); and annotation of novel miRNAs from unknown aligned reads, based on the prediction of the characteristic hairpin structure of the pre-miRNA around the read alignment locus. Since a brief review of unspliced alignment tools and miRNA databases have been given in their corresponding sections, we will review novel miRNA discovery. In table 2.9 can be seen a brief list of miRNA prediction tools and other wrapper tools.

Despite each tool has different strategies to predict novel miRNAs (some directly evaluating the read alignment and predicting the precursor structure through some predefined parameters or classifying them with a machine learning-based method, while others search for blocks of reads mapping close to each other or directly search miRNA duplexes), some of the parameters used are common to all tools due to the well-known pre-miRNA structure. It is known that all reads originating from a miRNA hairpin should correspond to either one of the miRNA duplex or to loop sequences; that 5’ ends (the seed sequence) are required to be homogeneous; and that a canonical pre-miRNA secondary structure should be predicted with a minimum free energy, minimum of paired based between the miRNA and “sister” miRNA sequence, the formed duplex need to have a 2 nt overhang at both 3’ ends and there must be absence of branches and bulges outside of the loop (284).

2.3.3.2 Target prediction

It is necessary to predict the candidate mRNA targets to study the functional role of all detected miRNAs or a subset (e.g., the differentially expressed ones). Although it is not fully known how miRNAs interplay with their target mRNA, some key aspects have been elucidated. Some of the most used features by target prediction programs are the complementarity of the seed sequence and the thermodynamic stability of the miRNA-mRNA complex. For a more detailed description of the miRNA mechanism of action it is recommended to review section 1.4.1.3 from Chapter 1. It must be pointed that most target prediction tools has concentrated their efforts in predicting miRNA target sites at 3’ untranslated regions (3’ UTRs), but there has been evidence that they can also target 5’ UTRs and coding regions (CDSs) (285). In addition, those tools have been reported to return a significant number of false positives, so in an attempt to reduce them, it is recommended to execute multiple target prediction tools based on different methods and take their intersection. Furthermore, it has been pointed out that a single miRNA can target multiple different mRNAs and, conversely, a mRNA can be targeted by multiple different miRNAs. In table 2.10 can be seen a brief list of target prediction algorithms and their main features.

Table 2.9: List of tools for novel miRNA identification from miRNA-seq data. Table based on (284) and corresponding tool reference.

Tool	Last Version*[1]	Main Features
BlockClust (286)	v1.1.0	Tool available as a galaxy repository. Briefly, the tool partition the reads of an expression profile in a sequence of blocks and then discretize the statistics of the read distribution in each block. Then, such blocks are clustered by an unsupervised method and classified with built-in class specific discriminative models for C/D box snoRNA, H/ACA box snoRNA, miRNA, rRNA, snRNA, tRNA and Y_RNA.
CoRAL (287)	v1.1.1	Machine-learning based tool for classification of RNA molecules, using features such as fragment length and cleavage specificity. It needs a training set for the classifier generation, which could be complex in non-model organisms.
deepBlockAlign (288)	v1.3.1	Overlapping mapped reads are merged into blocks and then closely spaced blocks are combined. Then, blocks are compared to determine similarity scores. In a second stage, block patterns are compared by a modified Sankoff algorithm. Finally, hierarchical clustering of blocks separates most miRNAs and tRNAs.
miRanalyzer (289)	-	Tool that has been replaced by sRNAbench.
miRDeep2 (290)	v2.0.1.2	It uses bowtie for sequence alignment and RNAfold for secondary structure prediction. First, the tool does a quantification of known miRNAs if precursor and mature miRNA sequence files from miRbase are given. Then, potential miRNA precursors are excised from the genome using the read mappings as guidelines. Then, reads a realigned to known and excised precursors. Finally, the secondary structure is predicted by RNAfold.
miRDeep-P2 (291)	v1.1.4	Plant specific tool for miRNA analysis. Improved version of miRDeep-P, whose model was adapted from miRdeep, but taking into account plant specific miRNA characteristics such as more variable precursor lengths and more prevalent large paralogous families.
miRDentify (292)	v1.00	First, it aligns reads using Bowtie (no mismatches allowed). Then, it assembles duplex-forming reads within a 46-80 nt distance and evaluates different parameters from each candidate. A cut-off and FDR is selected based on annotated miRNAs.

*[1] Last checked on 20/11/2019.

Table 2.9: (Continued).

Tool	Last Version*[1]	Main Features
MiRdup (293)	v1.4	Two main functions: 1) Given a miRNA and a pre-miRNA, it validates pre-miRNAs predictions from other tools. To that purpose, it uses a trained model on a particular set of species in order to maximize species-specificity. The model is trained on 100 features with adaboost on random forest (miRbase of other sequence data can be used to train the model). 2) Given a pre-miRNA and a model, it predicts a potential miRNA.
miReader (294)	v2.0	Tool to identify mature miRNAs without any dependence on reference genome or homologous references.
MIReNA (295)	v2.0	To identify pre-miRNA/miRNA pairs, it explores a multidimensional space defined by only five parameters (for more details check reference). These parameters characterize suitable pre-miRNA structures.
miRExpress (296)	v2.0	Generates miRNA expression profiles from high-throughput sequencing of RNA without the need for sequenced genomes. It align sequences directly to known miRNA databases and for reads not mapping to known miRNAs, a cross-species comparison is done.
miRPlex (297)	v0.1	Mature miRNAs are predicted through a multi-stage process, involving filtering, miRNA:miRNA* duplex generation and duplex classification using a support vector machine (one model for plants and nother for animals).
miRSeqNovel (298)	v1.3	An R based workflow. First, reads are mapped to a reference with an unspliced aligner such as Bowtie or BWA. Known miRNAs are identifies using miRbase database. For novel miRNA discovery, hairpin-like structures are searched in aligned reads, their secondary structure predicted with RNAfold and ranked according to different features.
sRNAbench (299)	v1.5	Module from the sRNAtoolbox. Uses bowtie for sequence alignment. Allows isomiR analysis.

*[1] Last checked on 20/11/2019.

Most of target prediction tools need a mature miRNA sequence file and an additional 3' UTR sequence file, whose boundaries may be not well defined in some non-model organisms. Once a list of predicted targets has been achieved, it can be further studied by functional annotation through enrichment analysis using the miRNA-target genes (assuming that miRNAs have similar functions of their target genes) or by correlation or other network methods to infer miRNA-mRNA regulatory modules, integrating sequence and expression profile data from mRNAs and miRNAs. It is though that single miRNAs, normally expressed at lower levels than their target mRNAs, are not enough to alter significantly mRNA expression levels by themselves.

Thus, it is necessary to study mRNA expression levels changes that are regulated by multiple miRNAs targeting them.

Table 2.10: List of tools for miRNA target prediction. Table adapted from (300,301) and corresponding tool reference.

Tool	Last Version*[1]	Main Features
EIMMo (302)	-	Not available anymore. It was a web-based tool with a Bayesian target prediction algorithm.
microT-CDS (303,304)	v5.0	Can identify binding sites on the 3' UTRs and CDS regions. It checks how conserved are the binding sites: 1) For CDSs: Calculates excess sequence conservation above the one required for amino acid conservation. 2) For 3'UTRs: Asses evolutionary conservation across 16 species and the conservation score is defined as the ratio of the number of species in which the binding position is conserved and the respective number using the maximal number of species having any conservation in the whole 3'-UTR region.
miRanda (305)	aug2010	Predict 3' UTRs targets based on: 1) Sequence complementarity. 2) Binding energy (through Vienna package). 3) Evolutionary conservation (PhastCons score).
mirSVR (306)	-	Machine learning method for ranking miRNA target sites. It trains a regression model on features extracted from miRanda-predicted target sites.
miRTar (307)	-	Web based system that can identify binding sites on the 3' UTRs, 5' UTRs and CDS regions. miRTar can analyze and highlight a group of miRNA-regulated genes that participate in particular KEGG pathways to elucidate the biological roles of miRNAs in biological pathways.
psRNATarget (308)	2017 Update	Plant sRNA (e.g., miRNAs and siRNAs) target prediction server. It evaluates: 1) Sequence complementarity. 2) Target site accessibility.
PicTar (309)	-	Given a 3' UTR file and miRNA mature sequences, it searches 7 base (seed) alignments that pass an optimal free energy filter and fall into overlapping positions in the alignments for all species under consideration. Each UTR in the alignment is scored with the central PicTar maximum likelihood procedure.
PITA (310)	v6	Predict 3' UTRs targets using a non-parametric model that scores microRNA-target interactions by an energy score equal to the difference between the energy gained by binding of the microRNA to the target and the energy required to make the target region accessible for microRNA binding.
RNA22 (311)	v2	Pattern-based algorithm that not relies in cross-species conservation information.

*[1] Last checked on 21/11/2019.

Table 2.10: (Continued).

Tool	Last Version*[1]	Main Features
RNAhybrid (312)	v2.1.2	The program finds the energetically most favourable hybridization sites of a small RNA in a large RNA. Statistical significance of predicted targets is assessed with an extreme value statistics of length normalized minimum free energies, a Poisson approximation of multiple binding sites, and the calculation of effective numbers of orthologous targets in comparative studies of multiple organisms.
TargetS (313)	-	Can identify binding sites on the 3' UTRs, 5' UTRs and CDS regions. It does not rely on evolutionary conservation
TargetScan (314)	v7.2	Can identify canonical binding sites on the 3' UTRs and CDS regions. Uses a model (context++) which considers site type and another 14 features (e.g., 3'-supplementary pairing, local AU content, structural accessibility, conservation, and distance from the closest 3'-UTR end) to predict the most effectively targeted mRNAs.

*[1] Last checked on 21/11/2019.

Multiple studies have compared performance of target prediction algorithms and there is a consensus regarding how all tools usually returns a great number of false positives. In (137) is remarked that tools with an stringent seed pairing criteria such as TargetScan, PicTar, EMBL and EIMMo have a high degree of overlap, with TargetScan having the most robust target ranking. In a recent study comparing TargetScan, miRanda-mirSVR, PITA and RNA22 using the validated miRNA target database from miRTarBase (only a limited number of miRNAs were selected for the analysis) (315), it was shown that miRanda had the best performance with a balanced sensitivity, specificity and precision, while TargetScan and PITA showed a better precision with the lower number of false positives. It is clear that there is not program consistently superior to all others and it has become pretty common to execute multiple target prediction algorithms to take their intersection for further research.

2.3.3.3 miRNA-mRNA correlation analysis

Target prediction algorithms based only in sequence information return a high number of false positives. In order to achieve more reliable results, and if mRNA expression levels from the same samples are available, it is possible to combine both miRNA and mRNA expression levels to construct miRNA-mRNA interaction networks. Different approaches have been developed for that purpose, e.g. methods using directly the correlation of miRNAs and their corresponding targets such as MMIA, mirConnX, MAGIA and TargetMiner (316); methods using graph mining techniques such as iSubgraph (317); and other methods using bi-clustering algorithms and Bayesian network models (318). It must be pointed out that methods that use only pairwise correlations to infer networks do not consider miRNAs interactions. As it has been mentioned before, an mRNA may be regulated by multiple miRNAs and its expression would be affected by all targeting miRNAs. In addition, such methods are not able to distinguish between changes

produced by a direct interaction with the miRNA and changes produced as a secondary consequence of miRNA targeting (e.g., a miRNA that targets a transcription factor).

Although methods that merge miRNA and mRNA expression profiles could increase our understanding on miRNA targeting, they are usually limited to the assumption that miRNAs negatively regulate their target mRNA and it may not be sufficient in some cases. Multiple miRNAs have been reported to not produce any mRNA expression change (319), but causes an expression change at the protein level. And inversely, full or partial inhibition of a mRNA does not mean that the protein expression levels would change substantially. In order to account for these cases, multiple studies have integrated miRNA, mRNA and protein expression profiles (318,320). Furthermore, adding complexity to miRNA targeting, it has been shown that some lncRNA can be targeted by miRNAs, while other lncRNAs can act as miRNA sponges (sequestering miRNAs and avoiding miRNA repression) or can produce miRNAs (321). In addition, some circRNAs have also been shown to act as miRNAs sponges (152). It is clear that miRNA interactions are quite complex, being multiple players at different levels (mRNAs, lncRNAs, circRNAs and transcription factors). In this thesis, only miRNA and mRNA interactions will be studied and some circRNA sponges will be defined for the first time in sheep.

2.3.3.4 Workflow

In figure 2.7 can be seen the workflow used to analyse the miRNA-seq data in this thesis. Briefly, a quality check was performed on the raw data files with FASTQC [v0.11.5] (182) to assess the most appropriate read quality filtering and trimming. The Trimmomatic [v0.36] (202) program was used to remove adaptor sequences (In contrast to total RNA sequencing, mature miRNA length is shorter that read length and, as a consequence, it is expected that the sequencer reads through the adaptor) and to filter reads shorter than 16 bp. The data was checked again with FASTQC.

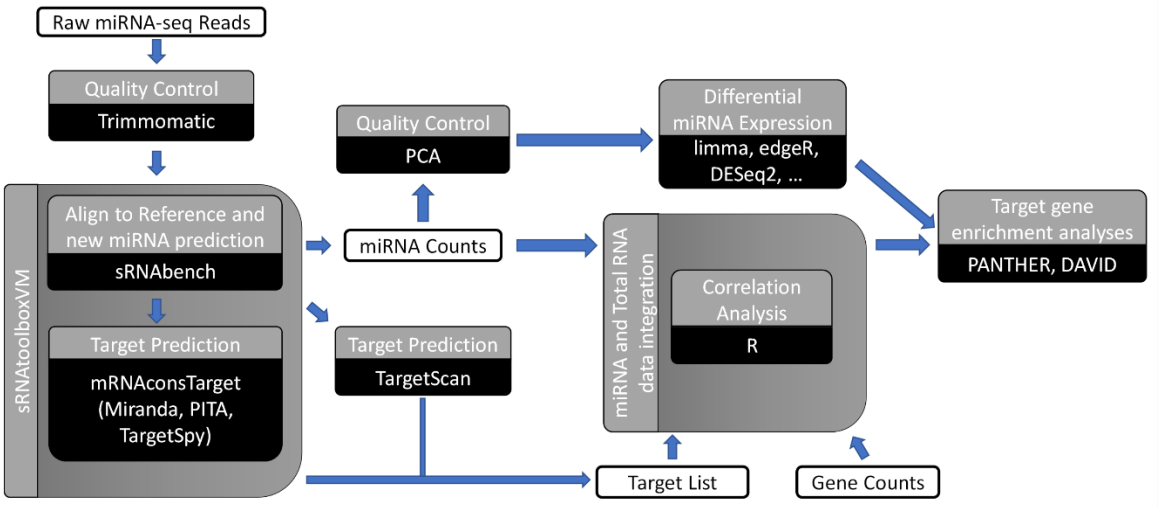


Figure 2.7: Pipeline followed for miRNA-seq data analysis.

For subsequent analyses, some of the sRNAtoolboxVM (322,323) modules were applied. sRNAtoolbox is a set of interconnected tools for small RNA sequencing data analysis. First, the sRNAbench module was used to align sequences to the *Ovis aries* reference genome Oar3.1 from Ensembl [version 89.31] (97), to profile the expression of small RNAs and to predict novel miRNAs. The program uses bowtie (204) behind the scenes to map all the sequences to the

reference genome, and it searches in the miRBase [v21] (324) database for known miRNAs in sheep. Furthermore, Rfam data was used to identify other small RNAs originating from rRNA, tRNA, snRNA, and snoRNA and exclude them from the analysis. The remaining sequences were searched against the mature miRNAs of human and other species, including cow, goat and mouse, in miRBase to identify miRNA homologs. For the discovery of new miRNAs, the remaining sequences were used to predict their folding secondary structure and, if a hairpin structure was predicted, their free energy of hybridization. Ultimately, the predicted new miRNAs were searched in the RNACentral [v6] database with blastn to ascertain if they have been previously identified.

Once a miRNA expression matrix is obtained, a differential expression analysis was performed with edgeR and DESeq2 for PBMCs and encephalon, respectively. The same model as for the total RNA-seq analysis was applied to each tissue, applying first the SVA package to remove unwanted variation. Similar to the total RNA-seq comparisons, differential expression analyses were performed for the time points, considering the treatment group (Vac Tf vs. Vac T0 and Adj Tf vs. Adj T0), and for the treatments at the end of the experiment (Adj Tf vs. Vac Tf) in PBMCs. In the case of encephalon samples, since only samples at the end of experiment were analysed, three different comparisons were made (Adj vs. Control, Vac vs. Control and Adj vs. Vac). In both tissues, the differentially expressed genes (DEGs) were selected as those with an adjusted p-value (using the Benjamini-Hochberg method) threshold of<0.05 and a fold change value of>1.5 or <0.667.

In parallel to the differential expression analysis, the mRNAconsTarget module from sRNAtoolbox was used to predict potential miRNA target genes, which uses miRanda (305) and PITA (310) for its predictions. In addition, the target prediction algorithm TargetScan (314) was applied independently, making a total of three distinct target prediction algorithms. In a attempt to reduce false positives and select trustworthy target genes, the following criteria was used: in miRanda a pairing score > 150 and an energy score < −15; in PITA, an energy score < −15; and in TargetScan, a contex++ score < −0.7. Only those genes that were common across the three programs were selected for further analysis.

Next, integrating total RNA and miRNA analyses, correlation between miRNA and target mRNA expression values were determined using the R statistical software [v3.4.1 PBMCs and v3.5.0 encephalon]. A test for association between paired samples using the Spearman's rank correlation coefficient was applied with the R *cor.test* function. The obtained p-values were used for estimation of the FDR (q-value) with the *qvalue* R package and the Benjamini-Hochberg method, using a threshold of <0.05 to indicate significant miRNA-mRNA pairs. Apart from the correlation analysis, in an attempt to discover miRNA-gene patterns, a subgraph mining tool was applied. For that purpose, the iSubgraph (317) algorithm was used, which searches for frequent cooperative regulations of genes and miRNAs happening in a minimum group of samples. Briefly, iSubgraph transforms the sequencing data from miRNA and total RNA into graphs for each sample and uses the target prediction information to create a bipartite graph for each sample. From the sample graphs constructed in the first part, using graph mining algorithms, iSubgraph searches for frequent cooperative regulations of genes and miRNAs. The parameters were set as follow: the threshold for Up and Down tags was set at 0.75; and to report a pattern, that pattern needed to be found at least in three samples.

2.3.4 circRNA analysis

Apart from the differential expression analysis, the data was used for circRNA identification and functional annotation, since total RNA libraries were prepared with a rRNA depletion strategy which retain most non-poly(A) ncRNAs (e.g., circRNAs and lncRNAs). The strategy for circRNA analysis is quite similar to the one followed for differential expression analysis. The main difference is that at the alignment step a tool that can detect topologically inconsistent split reads must be used. In addition, the counting strategy for circRNAs is different and will be explained in more detail bellow.

2.3.4.1 Alignment to the reference and circRNA identification

circRNA identification tools search for topologically inconsistent split reads, in which a downstream 3' splice site and an upstream 5' splice site is linked, formally known as backsplicing. The tools for circRNA annotation can be classified according to their dependency on genome annotation, being two different strategies (see figure 2.8 for more details): a pseudo-reference strategy, in which reads are aligned to all possible combinations of annotated exons in an attempt to detect inconsistent reads, and a fragment-based strategy, in which reads are split in segments and aligned to the reference genome. Tools based in the pseudo-reference strategy are limited to find circRNAs whose origin is an annotated exon junction, which makes them unsuitable for poorly or partially annotated organisms (160). In table 2.11 can be seen a brief list of tools for circRNA identification from total RNA-seq data.

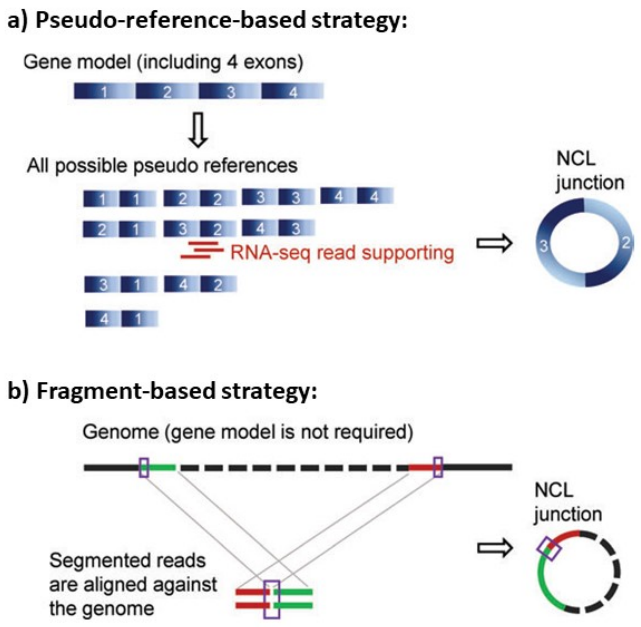


Figure 2.8: Two different strategies for circRNA detection: a) Pseudo-reference-based strategy: Using genome annotation, all pseudo-references (all possible combinations of exons, including topologically inconsistent ones) are constructed. Then, all reads aligning to junctions of topologically inconsistent pseudo-references are regarded as circRNA candidates. b) Fragment-based strategy: Reads are directly aligned to the reference genome with an aligner that can split reads and align them in a non-linear fashion. (figure adapted from (160)).

Most circRNA annotation tools apply some filters after read alignment to distinguish between real circRNA candidates and other non-co-linear (NCL) products that can produce topologically inconsistent read alignments, such as trans-spliced RNAs and genetic rearrangements, or sequencing errors and other *in vitro* artifacts. Among the most used filters are: to select a minimum read counts to treat a circRNA as expressed; to detect a circRNA in multiple samples; and if paired end data is available, to check that both ends align in the predicted circRNA or that both ends are in the correct orientation.

Table 2.11: List of tools for circRNA identification.

Tool	Last Version*[1]	Main Features
ACFS (325)	v2.1	No need of annotation. Designed for SE data, but can deal with PE data with less sensitivity. First, it aligns reads with TopHat2 and, then, unmapped reads are aligned with BWA. Uses a maximal entropy model to pinpoint circRNAs.
CIRCexplorer2 (326)	v2.3	Two modes: with annotation or annotation-free. Supports both SE and PE data. Supports multiple circRNA-aware aligners (TopHat2, STAR, MapSplice, BWA and segemehl)
Ciri2 (327)	v2.0.6	No need of annotation. Align reads with BWA-MEM. Allows multithreading. Uses an adapted maximum likelihood estimation based on multiple seed matching to identify back-spliced junction reads and to filter false positives derived from repetitive sequences and mapping errors.
DCC (328)	v0.4.7	No need of annotation. Supports both SE and PE data. Align reads with STAR with chimeric alignment detection activated. It returns two expression files: one for the detected circRNAs and another one for their corresponding linear counterparts. Only returns circRNAs from canonical GT/AG splicing sites and discards mitochondrial ones. Filters those whose origin is in a repetitive region.
find_circ (329)	Custom Script (2013)	No need of annotation. First, reads are aligned contiguously to the genome. From the unmapped reads, 20mers from both ends are extracted and aligned independently with Bowtie2 to detect circRNA spanning reads. Only returns circRNAs from canonical GT/AG splicing sites.
KNIFE (330)	v1.5	Need of annotation. Supports both SE and PE data (for PE data, each end is aligned independently as SE). Uses Bowtie2 for read alignment. Statistically based splicing detection for circular and linear isoforms.
MapSplice2 (331)	v2.2.1	No need of annotation. First, reads are aligned contiguously to the genome. In the second step, segments that do not have an exonic alignment are considered for spliced alignment using a splice junction search technique that starts from neighbouring segments already aligned.
NCLScan (332)	v1.6.5	Need of annotation. Detects NCL transcripts (fusion, trans-splicing, and circular RNA) from PE data. First, aligns reads contiguously with BWA and, then, unmapped reads are aligned again with Novoalign.

*[1] Last checked on 26/11/2019.

Table 2.11: (Continued).

Tool		
Segemehl (333)		Spliced aligner that has circRNA detection. No need of annotation. Support SE and PE data. Allows multithreading. Has high computer requirements, ~50 Gb, and high run times. For circRNA detection, it is recommended to use another spliced aligner first and pass unmapped reads to segemehl.
TopHat-Fusion (334)	v2.1.1	Included in TopHat2. It was designed for fusion gene discovery, which results from chromosome rearrangements, but can be adapted for circRNA discovery with the adequate filters. No need of annotation. Supports both SE and PE data
UROBORUS (335)	v2.0.0	Based on TopHat and Bowtie, it has a similar strategy to find_circ. Only for exonic circRNAs.

*[1] Last checked on 26/11/2019.

Interest in circRNAs has recently raised and most circRNA detection tools are quite recent. Those tools for RNA-seq data report a great number of circRNAs, a lot of them being false positives. It has been shown that around 31-76% of the NCL events reported in rRNA depleted data are not detected in both poly(A)-depleted and RNase R-selected data (which depletes all linear RNAs) (160). In a recent study comparing the performance of 11 circRNA detection tools (among them Ciri, DCC, find_circ, MapSplice, NCLScan, segemehl, UROBORUS and KNIFE) on synthetic and real datasets (336), the following was was shown: 1) Ciri and KNIFE had a balanced performance regarding sensitivity and precision; 2) NCLScan and MapSplice were quite conservative, with less sensitivity compared to the previous ones; 3) Segemehl was sensitive, but returned many false positives; 4) There was no single tool that dominated the rest in all the used metrics. In other study comparing the performance of circRNA_finder, CIRCexplorer, MapSplice, Ciri and find_circ, it was shown that CIRCexplorer and MapSplice returned the most reliable results, which can be partially explained by the fact that both tools require a mandatory annotation file (337). In addition, they showed that the false positive rate (measured by the RNase R resistance of circRNA candidates) ranged from 12% to 28%.

2.3.4.2 Quantification

In the case of circRNA detection by rRNA depleted total RNA-seq data, only the backsplicing junction is being detected. It is not possible to infer the remaining structure of the circRNA (e.g., in case of an exonic circRNA, which exons compose the circRNA or if any intron is retained), since the origin of the remaining reads can be both the circRNA or the linear counterpart. Some tools have tried to infer their structure by statistical modelling, using a similar strategy of some transcriptome assemblers and, thus, with similar disadvantages. All the tools for circRNA detection previously listed only count reads aligning to the backsplicing junction. As a consequence, some tools also return the expression of the host linear isoform as the reads aligning to the linear junctions, to be able to compare circRNA and host linear expression.

2.3.4.3 Differential expression

Since the circRNA quantification is totally different to the total RNA-seq, it is not clear if circRNA expression can be modelled in a similar fashion (e.g., using the negative binomial model from DESeq2 or edgeR). In general, circRNA expression has a totally different structure, mainly: 1) circRNAs are usually detected with a low quantity of assigned reads, lower than their linear counterparts (although there are some exceptions); 2) their expression varies greatly between samples; 3) there are a lot of samples with no expression of the corresponding circRNA. There is no study for now that has tried to address the circRNA expression modelling and most of the studies have limited themselves to tools such as DESeq2, edgeR2 and limma and to non-parametric tests. As a consequence, circRNA differential expression results must be taken with caution.

2.3.4.4 Workflow

In figure 2.9 can be seen the workflow used to identify circRNAs in total RNA-seq data in this thesis. Briefly, a read quality filtering and trimming was performed with Trimmomatic [v0.38] (202) using the following criteria: 1) adaptor removal with the "palindrome" mode for paired-end data; 2) trimming of bases from the start or end of a read if their quality dropped below a Phred value of 20; 3) trimming of reads if the average quality within a sliding window of five nucleotides falls below 20; and 4) read filtering if their length was sorter than 40.

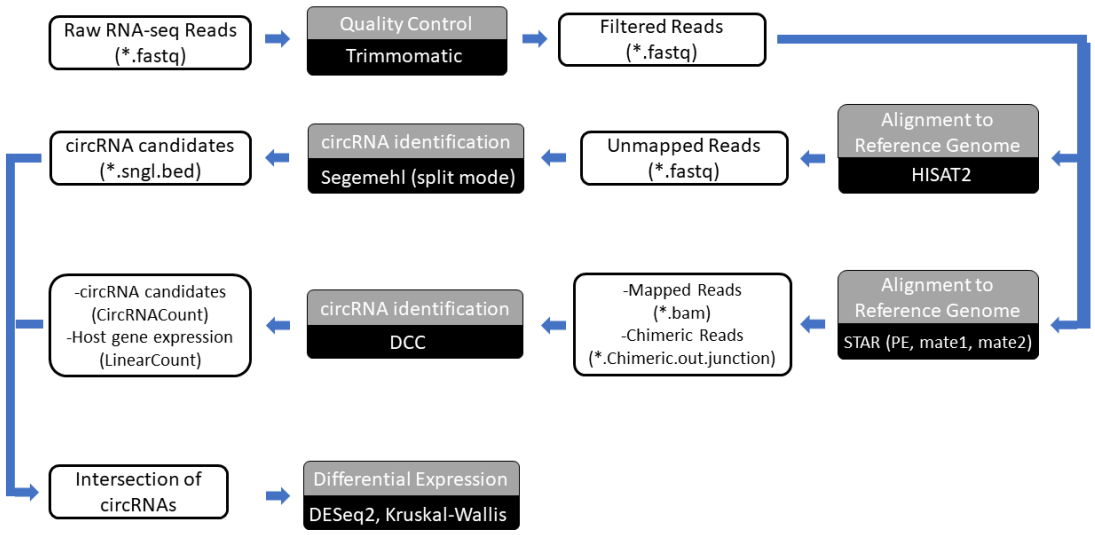


Figure 2.9: Workflow followed for circRNA detection.

As previously mentioned, circRNA identification tools search for topologically inconsistent split reads, in which a downstream 3' splice site and an upstream 5' splice site is linked, formally known as backsplicing. Since the sheep annotation is still in progress, two fragment-based circRNA identification tools were selected: segemehl [v0.3.4] (333) and DCC [v0.4.7] (328). In the case of segemehl, quality filtered reads were first aligned to the sheep reference genome (Oar_v3.1) with HISAT2 [v2.1.0] (224). Then, the set of non-aligned reads from the previous step were used to detect circRNAs in segemehl with default parameters. In contrast, for DCC, the quality filtered reads were first aligned to the reference genome with STAR [v2.6.1d] (338) following DCC author recommendations. Then, the chimeric.out.junction files

from the previous alignments and a file with repetitive regions in the sheep genome downloaded from the UCSC table browser (RepeatMasker and Simple Repeats tracks) were passed to DCC. DCC was run with default parameters, changing that a circRNA had to be expressed with one read in at least one sample to be reported. For further analysis, different filtering criteria were checked and different filters were applied to the tissues, as they had different experimental setups. In both tissues circRNAs needed a minimum of 2 read counts to be taken as expressed. In addition, in encephalon, circRNAs needed to be expressed at least in the same three samples in both tools, while in PBMCs needed to be expressed at least in the same three samples from one group in both tools. The intersection of both tools was taken to select trustworthy circRNAs. The expression counts for the detected circRNAs and host genes were taken as reference from DCC for further analysis.

Once an expression matrix was achieved, different metrics and plots were checked in R with self-scripts and plotting functions from multiple packages. Multiple authors have pointed out that some circRNAs are tissue specific and evolutionary conserved (339). The main databases of circRNA annotation are focused in human, mouse, rat, zebrafish, fly and worm, being sheep circRNA data not submitted to any database to date. A literature search of articles in which circRNAs in sheep are detected and are given at least as supplementary material was done in an attempt to compare the circRNAs annotated in this study. Four different studies in different tissues were found: two from the pituitary gland (155,156) and another two from the longissimus dorsi muscle (157,158). In addition, as circRNAs are evolutionally conserved to some extent in different species, the detected circRNAs were compared to the ones annotated in CIRCpedia (340) for human. The following steps were carried:

- The 5’ and 3’ flank coordinates of each circRNA found in sheep were converted to human coordinates with the liftOver tool from UCSC (341) with default parameters (min. ratio of remapped bases = 0.95).

- The resulting coordinates were screened for overlap with human annotated circRNAs in CIRCpedia. Splice sites detected in +/- 2 nt intervals around the putative human sites were considered homologous.

- Different categories were assigned to each circRNA: “not-aligned”, the sheep coordinates were not translated to human with liftOver; “no homologous”, No human circRNA detected near both splice sites; “5’ site utilized”, a human circRNA that only use the 5’ splice site is detected; “3’ site utilized”, a human circRNA that only use the 3’ splice site is detected; “Both sites utilized”, both splice sites are used by different circRNAs in human; and “homologous”, a human circRNA using both splice sites is detected.

Then, detected circRNAs whose origin were in an annotated gene were further analysed. Supposing that circRNAs have functions related to their host genes, gene enrichment analyses were conducted using the GO and KEGG databases in g:Profiler (276). The tool computes the p-values using a Fisher’s exact test and applies the Benjamini-Hochberg multiple test correction. The set of all expressed genes detected in the total RNA-seq libraries was set as background and related terms associated with the host genes of the circRNAs were checked for enrichment. Terms composed of more than 400 genes, due to limited interpretative value, or composed of less than 5 genes, due to the decrease in statistical power by multiple testing correction, were removed from the analysis. Those terms with an FDR less than 0.05 were selected for further

analysis. For visualization purposes, the list of enriched GO term was further analysed with Cytoscape using EnrichmentMap and Autoannotate plugins (278). Briefly, GO term information is inherently redundant, as genes often participate in multiple pathways, and collapsing redundant pathways into a single biological theme simplifies interpretation. EnrichmentMap generates a network in which pathways are visualized as nodes connected between them if the terms share many genes. Pathways with common genes often represent similar biological processes and are grouped together as sub-networks. Then, the clustering algorithm (clusterMaker2) automatically group similar pathways into major biological themes and then summarizes each cluster on the basis of word frequency within the pathway names (WordCloud app). Then, those words that were automatically selected were manually renamed to better explain groups. To note, clusters with less than 3 interconnected nodes were removed for visualization purposes.

In an attempt to check circRNAs acting as sponges, a list of predicted miRNA sponges, identified as clusters of miRNA binding sites, previously reported in the human genome (hg19) was downloaded (342). The genomic coordinates of each sponge candidate were converted to hg38 with liftOver (min. ratio of remapped bases = 0.95) and intersected with those of the circRNAs identified in this study, already lifted from the sheep reference genome to hg38, with bedtools (min. fraction overlap = 75%). Results were then filtered by excluding circRNAs targeting miRNAs for which no orthologue sequence was reported in sheep according to Ensembl (release 97). All human miRNAs hairpins were screened for similarity with the Oar3.1 genome with BLAST, requiring a minimum sequence identity of 90% on at least 95% of the hairpin. The sequences of the processed miRNAs were downloaded from miRbase (343) (Release 22.1) and the corresponding sheep orthologous were extracted from the alignment provided by Ensembl. CircRNAs were screened for miRNA binding sites with RIsearch2 (344), using the following parameters: -s 1:8/6 -e -10 -l 20 -p2. In the same way we re-evaluated the clusters of miRNA binding sites identified in human

Finally, a differential expression analysis was performed in both tissues. For the encephalon samples, the differential expression analysis was performed via two different methods. First, the analysis was done with DESeq2 and those circRNAs with an adjusted p-value<0.05 were taken as cut-off. An alternative method was also applied, given that DESeq2 is not designed specifically to work on circRNA expression data. In this case, for normalization of the circRNA expression data, not only the circRNA counts were taken into consideration to calculate library sizes, the total amounts of reads aligned to the reference annotation was considered. The data was normalized by SRPBM (Spliced Reads per Billion Mapped Reads) (146) (see formula 14). After normalization, a Kruskal-Wallis test was done to check for differences between groups and the p-values were adjusted for multiple comparisons by the Benjamini & Hochberg method. Those circRNAs with an adjusted p-value<0.05 were taken as cut-off. For the PBMC samples, a batch effect removal program, harman [v1.12.0] (247), was applied after normalizing data by SRPBM. Then, the package limma and the Kruskal-Wallis test were applied to check for differential expression. Those circRNAs with an adjusted p-value<0.05 were taken as cut-off.

$$SRPBM = \frac{circRNA\ counts}{Total\ mapped\ reads} \times \frac{10^6}{Read\ length} \qquad (14)$$

Chapter 3

Response to Aluminium in PBMCs

3.1 Introduction

Aluminium hydroxide (AH) is one of the most used compounds as adjuvant in human and veterinary vaccines. Despite its widespread use, the mechanism of action is not fully known. Although aluminium adjuvants are extremely effective at enhancing antibody responses, are well tolerated and have the strongest safety record, recently some safety concerns have been raised: 1) It seems that the body is not able to discharge all injected aluminium; 2) It seems that it can reach distant organs; 3) It has a long-lasting biopersistence; 4) Some studies has related Al long-term exposure to multiple diseases in susceptible individuals. Independent of those concerns, which some remains quite controversial in the scientific community, it is clear that further research is need, preferentially *in vivo* to be able to capture the full picture of changes in the immune system after exposure to Al adjuvants. Understanding how Al adjuvants exert their role and the factors that affect them may lead to new and more efficient adjuvant formulations with better immunogenicity and safety profiles.

An inflammatory muscle disorder was described in humans by Gherardi et al. (345) in which the presence of aggregates of aluminium-containing macrophages was detected in intramuscular inflammation, which was linked at the same time to inoculation of aluminium-containing vaccines. This disorder was later known as macrophagic myofasciitis (MFF). In recent years the ASIA syndrome (Autoimmune/Autoinflammatory Syndrome Induced by Adjuvants) has been defined (78). This syndrome encompasses four different medical conditions known by their hyperactive immune responses, namely the Gulf War syndrome, the MFF, siliconosis and post-vaccination phenomena. These four conditions have been previously related to exposure to some kind of adjuvant, suggesting that a common denominator to them is adjuvant activity. In sheep, symptoms that can be related to ASIA has been described after a compulsory vaccination campaign against the bluetongue virus of ruminants in 2008 (5). The symptoms seen in sheep appeared in two different phases: an acute phase, which was seen in 25% of the flocks and affected to less than 0.5% of animals in a herd, characterized by a low response to external stimuli and acute meningoencephalitis; and an acute phase triggered by external stimuli (mainly low temperatures), which could be seen in 50-70% of flocks in a specific area and could affect to nearly 100% of animals in a herd, and characterized by an initial excitatory phase, followed by weakness, extreme cachexia, tetraplegia and death.

There have been few in vivo sequencing studies on the effect of aluminium hydroxide adjuvant and its influence on the immune response after a long-term exposure to multiple Al-based vaccines. Furthermore, the majority of studies have focused their efforts in brain changes caused by Al neurotoxicity and in using rat or mouse as models. In a recent study, male rats were exposed to intraperitoneal injections of a complex of aluminium chloride hexahydrate and maltolate every other day for 3 months and hippocampus samples were sequenced (53). Genes related to glial cell differentiation, neuronal transmission and vesicle trafficking were found differentially expressed. In other study comparing multiple adjuvants (Al among them) for an

inactivated rabies virus vaccine, blood samples from ICR mice inoculated intraperitoneally were sequenced (346). Alum-vaccinated mice were enriched in terms related to leukocyte differentiation and activated the *antigen processing and presentation* pathway. One drawback of these studies using mice is that researchers use intraperitoneal injections instead of subcutaneous or intramuscular injections, so any outcome regarding Al translocation to other tissues such as brain trough the intraperitoneal route would be hardly extended to any large mammal, in which vaccination is usually through the other routes. In addition, there are not commercial vaccines intended for use in mice, and most of these studies use generally the adjuvant alone or reduced versions of the human vaccines. Depending of the aluminium adjuvant and inactivated pathogen combination (and the proportion of both elements), the vaccine could take different conformations (mainly agglomerate size) that could cause different reactions. Therefore, it would be recommended to study Al adjuvancy in larger mammals to achieve more meaningful results that could be extrapolated to human. The use of farm animals or some domestic animals would be preferable, since there are vaccines specifically designed for them. Sheep is an attractive animal model to study adjuvants due to: 1. their low cost (at least compared to other farm animals); 2. Il8 has been identified, not in mice; 3. Large lifespan allowing long-term studies; and 4. their previous use as animal models for specific vaccines (e.g., to asses vaccines against the human respiratory syncytial virus designed for infants and to test the zoonotic pathogen Rift Valley fever virus) (347). Although there are some drawbacks when compared to mice, such as higher costs (food and facilities), worse state of reference genome and a smaller research community.

In this study, lambs received a parallel subcutaneous treatment with either commercial vaccines containing aluminium hydroxide or an equivalent dose of only this compound with the aim of identifying the activated molecular signature. Blood samples were taken from each animal at the beginning and at the end of the experiment and PBMCs were isolated. Then, total RNA and miRNA libraries were prepared and sequenced. In an attempt to decipher AH adjuvant mechanism of action, three expression comparisons were made: vaccinated animals at the beginning and at the end of the treatment, adjuvanted animals at the same times, and animals of both treatments at the end of the experiment.

The main aim of this sequencing study was to characterize the immune response to a long-term and intensive vaccination schedule and to check previously studied pathways related to AH adjuvant in an *in vivo* experiment in sheep. The hypothesis of this work was that a prolonged exposure to vaccine adjuvants, always following manufacture's recommendations, would result in a hyperactivation of the immune system. For that purpose, two different sequencing libraries per sample were constructed, namely total RNA-seq and miRNA-seq. Thus, the objectives of this work were:

1. to identify genes and regulatory elements involved in the immune response induced by the repetitive inoculation of vaccines with AH and to check the state of some pathways previously related to AH by others in literature.
2. to predict potential targets of miRNAs that can be related to Al adjuvancy and test for correlation between both miRNA and predicted mRNA target expression data.

3.2 Material and methods

In this section only a brief description of animal samples, extraction method and validation by qPCR (quantitative polymerase chain reaction), also known as real-rime PCR, will be provided.

The analysis workflow for total RNA-seq and miRNA-seq data has been described in full detail in their corresponding section in Chapter 2 – Material and Methods.

3.2.1 Animals

All experimental procedures were approved and licensed by the Ethical Committee of the University of Zaragoza (ref: PI15/14). Requirements of the Spanish Policy for Animal Protection (RED53/2013) and the European Union Directive 2010/63 on the protection of experimental animals were always fulfilled.

As previously stated, Rasa Aragonesa pure breed lambs were selected from a single pedigree flock of certified good health at 3 months old and with the requirement of not having undergone any vaccination before the experiment. For the purpose of the present work, they were randomly distributed in different treatment groups, n = 7 each. Each group received a parallel subcutaneous treatment with either commercial vaccines containing aluminium hydroxide as adjuvant (Group Vac) or aluminium hydroxide diluted in PBS (Group Adj; Alhydrogel, CZ Veterinaria, Spain), always inoculating with the equivalent dose of aluminium applied in the vaccinated group. Aluminium content was established by inductively coupled plasma atomic emission spectrometry and calculated per total dose. Nine different vaccines were used (see table 2.1 in Chapter 2), and a total of 19 inoculations were applied in each group throughout 16 different inoculation dates (see figure 2.2 in Chapter 2), thus entailing a total amount of 81.29mg of aluminium per animal. Intervals between inoculations ranged from 17 to 100 days (mean = 31.3 ± 22.1 days). The complete study lasted 475 days, from February 2015 to June 2016. All injections were subcutaneous, in the area encompassing scapula and ribs. 16 out of 19 inoculations were performed in the right flank, and the rest, corresponding to double injection dates, were performed in the left flank.

For total RNA-seq and miRNA-seq analysis, samples at the start and at the end of the experiment from the same 6 animals (3 sheep inoculated with vaccines and 3 sheep inoculated with the adjuvant alone) were used for library preparation. The rest of the animals from those two groups (4 sheep inoculated with vaccines and 4 sheep inoculated with the adjuvant alone) were used for validation of the sequencing data by qPCR. A list of the samples used in this experiment can be seen in table 3.1.

Table 3.1: Samples used for sequencing and RT-qPCR. Vaccine refers to the group vaccinated with commercial vaccines, while adjuvant refer to the one inoculated with aluminium hydroxide diluted in PBS. In addition, T0 refers to the start of the experiment and Tf to the end.

Treatment	Animals	Time	Samples
Sequencing			
Vaccine	121, 124, 125	T0	121-A, 124-A, 125-A
		Tf	121-B, 124-B, 125-B, 125-B*
Adjuvant	111, 114, 116	T0	111-A, 114-A, 116-A
		Tf	111-B, 114-B, 116-B
RT-qPCR			
Vaccine	122, 123, 126, 127	T0	122-A, 123-A, 126-A, 127-A
		Tf	122-B, 123-B, 126-B, 127-B
Adjuvant	112, 113, 115, 117	T0	112-A, 113-A, 115-A, 117-A
		Tf	112-B, 113-B, 115-B, 117-B

*Same RNA sample obtained with a conventional TRIzol extraction method.

3.2.2 Blood collection, RNA extraction and sequencing

For the isolation of ovine peripheral blood mononuclear cells (PBMCs), blood was collected from the jugular vein of 14 Rasa Aragonesa sheep. Blood samples were taken from each animal at the beginning (day 0, T0), before any vaccination, and at the end of the treatment (day 475, Tf), which was 5 days after the last inoculation. Blood was collected into heparinized Vacutainer tubes (Becton, Dickinson and Company, Sparks, MD), transferred into 50-ml centrifuge tubes and diluted 1:2 in HBSS. Twenty-five millilitres of blood:HBSS were layered over 10ml of Ficoll-Paque (1.084 g/cm^3) (GE HealthCare Bio- Sciences, Uppsala, Sweden) in 50-ml centrifuge tubes. The cells were centrifuged at 900 × g for 30min to separate erythrocytes and polymorphonuclear cells from PBMCs. PBMCs were collected from the HBSS-Ficoll-Paque interface, washed with HBSS by centrifugation at 400 × g for 10min, lysed in 1ml of TRIzol and stored at −80°C until further use.

Total RNA was extracted from PBMCs using an RNA Clean & Concentrator™-5 kit (Zymo Research, Irvine, CA, USA) following manufacturer's instructions and stored at −80°C. RNA quantity and purity were assessed with a NanoDrop 1000 Spectrophotometer (Thermo Scientific Inc., Bremen, Germany). The RNA integrity and concentration were assessed with a 2100 Bioanalizer (Agilent Technologies, Santa Clara, CA, USA). Two numeric parameters concerning RNA integrity were estimated, the 28S/18S (ribosomal RNA) ratio and the RNA integrity number (RIN value). The RNA samples with a RIN value >7.5 and a 260/280 ratio >1.8 were used. A summary with the sample qualities can be seen in table 3.2.

Table 3.2: Quality summary of sequenced samples with their 260/280 and 260/230 absorbance ratios and RIN values. Vac, group vaccinated with commercial vaccines; Adj, group inoculated with aluminium hydroxide diluted in PBS; T0, start of the experiment (day 0); Tf, end of the experiment (day 475).

CNAG ID	Sample name	Group	260/280 Absorbance ratio	260/230 Absorbance ratio	RIN
AD1395	121-A	Vac T0	2	1.86	9,3
AD1396	124-A	Vac T0	1.94	1.27	9,3
AD1397	125-A	Vac T0	2.06	1.9	9,4
AD1398	111-A	Adj T0	2.1	2.11	7,8
AD1399	114-A	Adj T0	2.09	2.14	8,1
AD1400	116-A	Adj T0	2.06	2.2	9,3
AD1401	121-B	Vac Tf	2.05	2.05	9,6
AD1402	124-B	Vac Tf	1.98	1.64	9,2
AD1407	125-B	Vac Tf	1.94	1.79	9,2
AD1403	125-B*	Vac Tf	1.87	-	6.4
AD1404	111-B	Adj Tf	1.87	-	8,1
AD1405	114-B	Adj Tf	1.91	-	8,5
AD1406	116-B	Adj Tf	1.93	-	7,8

*Same RNA sample obtained with a conventional TRIzol extraction method. Only used for total RNA-seq.

Total RNA-seq libraries were prepared according to the TruSeq Stranded Total RNA kit with Ribo-Zero Globin (Illumina, San Diego, CA, USA) to deplete the samples of cytoplasmic rRNA and globin mRNA. The miRNA-seq libraries were prepared according to the TruSeq Small RNA library prep kit (Illumina). Total RNA and miRNA libraries were sequenced on a HiSeq2000 sequencer and HiSeq2500 sequencer, respectively. RNA-seq was conducted for a total of 13 samples, with a mean sequencing depth of 70 million 76 base pair (bp) paired-end reads at CNAG

(Centro Nacional de Análisis Genómico, Barcelona, Spain). miRNA-seq included 12 samples, with a mean sequencing depth of 17 million 50 bp single- end reads at CRG (Centro de Regulación Genómica, Barcelona, Spain).

3.2.3 qPCR validation

To validate changes that were identified by RNA-seq experiments, the relative expression levels of 9 genes (CNTLN, EGR2, GPRC5C, HGF, NRXN2, SAMD4B, SKAP2, TREM1, WDR5B) and 3 miRNAs (oar-let-7b, oar-miR-19b, oar-miR-25) that were selected based on significant changes seen in the RNA-seq and miRNA-seq analyses were verified by qPCR. For quantification of mRNA transcripts, primers were designed using PrimerQuest and OligoAnalyzer tools of Integrated DNA Technologies (IDT). GAPDH, ATPase, ACTB and TFRC were used as reference genes. For quantification of miRNAs, primers were designed using Qiagen platform. U6 snRNA, oar-miR-30d and oar-miR-191 were used as internal standards. These last two miRNAs were selected for their expression stability in our samples. Supplementary Table S3.1 and Table S3.2 shows the list of the amplified ovine genes and miRNAs and the corresponding primer sequences, respectively. The Real-time qPCR amplifications of cDNA pools was accomplished using PowerUp™ SYBR™ Green Master Mix (Applied Biosystem, Foster City, CA, USA) in a 10 µl final volume reaction, according to the manufacturer´s instructions. qPCR reactions were conducted on a QuantStudio® 3 detection system (Applied Biosystem) under the following conditions: 1 cycle of 50℃ for 2 min, 1 cycle of 95℃ for 2 min, 40 cycles of denaturation at 95℃ for 15 s, annealing at 60℃ for 60 s, and a dissociation curve to measure the specificity of the amplification. Appropriate controls (no template and no retrotranscription) were included. Primer concentrations that did not produce non-specific fragments or primer dimers and generated the lowest Ct value were selected for the final analysis.

The expression study has been based on the analysis of mRNA and 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 expression analysis 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. Ct values and real-time PCR analysis was carried out with Fluidigm Real-Time PCR Analysis Software [v3.1.3]. PCR efficiency calculation and correction, reference gene and miRNA stability analysis and normalization have been done with GenEx software of MultiD [v5.4]. Most of the genes and miRNAs showed high amplification efficiencies with a mean value of 96 % and 99 % respectively. The stability of candidate reference genes and miRNAs was analysed using both NormFinder (348) and GeNorm (349) algorithms integrated in GenEx. The two most stable genes were ACTB and GAPDH and normalization has been performed using these two reference genes. The two most stable miRNAs were oar-miR-30d and oar-miR-191 and normalization has been performed using these two miRNAs.

Changes in gene and miRNA expression (n-fold) or relative quantification (RQ) were determined by the Δ(ΔCt) method. Based on the sequencing results, three comparisons were done: Vac Tf vs Vac T0, Adj Tf vs Adj T0 and Adj Tf vs Vac Tf. The results are expressed as relative quantifications and fold changes, which was standardized by log2 transformation. Normal distribution was checked using the Shapiro-Wilk test in the IBM SPSS statistical package [v24]. Changes in expression between different groups (Vac Tf vs Vac T0, Adj Tf vs Adj T0 and Adj Tf vs Vac Tf) were compared with the Tukey HSD or Games-Howell *post-hoc* test (ANOVA) or with

non-parametric Kruskal-Wallis test of the SPSS package. In all analyses, differences were considered significant when p-values were <0.05.

3.3 Results

3.3.1 Total RNA-seq

3.3.1.1 Sequencing quality

After preparing 13 total RNA-seq libraries for sequencing, a mean value of 69.7 million 75 nt paired-end reads per library were achieved. After trimming of adaptor sequences and low-quality fragments with Trimmomatic (chosen parameters can be seen in the corresponding section of Chapter 2 – Material and Methods), a mean of 68.5 (SD=6.95) million reads (98.33%) were considered as good quality segments for subsequent analyses. The average quality of the trimmed samples can be seen in figure 3.1. The drop in quality seen in the first bases is common in Illumina sequencing data due to random hexamer priming during the cDNA generation step (112). As it can be seen, all samples have good quality after trimming.

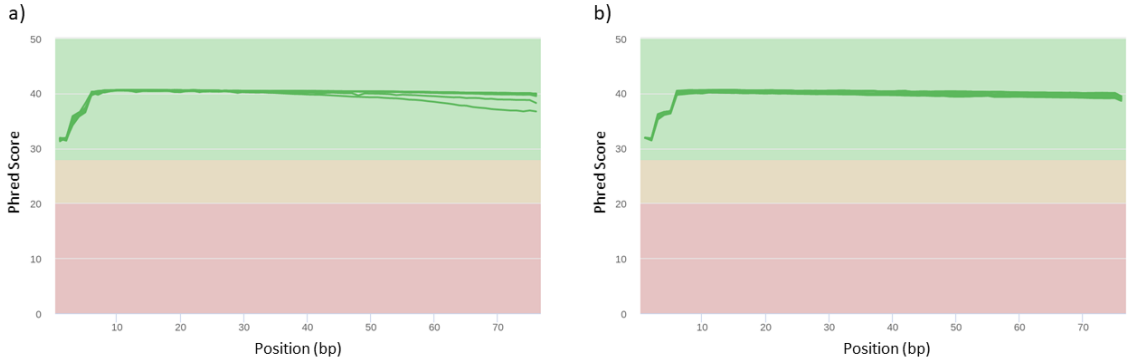


Figure 3.1: Average quality per base for all sequenced samples for paired-end data. a) pair 1 and b) pair 2. The y-axis shows the average quality in Phred scale, while the x-axis is a representation of all bases in a sequencing read of 75 nt for each pair. The background of the graph is divided by different colours in three main section, with the green section indicating good quality bases, the orange reasonable quality bases and the red one poor quality bases. The per sample quality plots were produced by FASTQC and then, they were aggregated by MultiQC.

3.3.1.2 Alignment to reference genome

Trimmed reads were aligned to the *Ovis aries* reference genome Oar_v3.1 from Ensembl with STAR. The alignment yielded a mean value of 52.7 (SD=10.4) million read pairs (76.95%) mapping to a unique locus, 11.4 (SD=5.8) million read pairs (16.70%) mapping to multiple loci and 4.3 (SD=2.6) million read pairs (6.32%) not mapping to any loci in the genome. A more detailed summary of the alignment can be seen for each sample in table 3.3. Only uniquely mapped reads were used for gene expression quantification by featureCounts in a strand-specific manner. A mean value of 36.1 million read pairs (68.42%) per sample were successfully assigned to sheep annotated genes. Once an expression matrix was achieved, different metrics were evaluated with the NOISeq R package from Bioconductor and RSeQC package. Among the checked features were a splice junction class pie chart (see figure 3.2), a “saturation plot” (see figure 3.3) and a

"sensitivity plot" (see figure 3.4). Only a junction class pie chart from sample 121-A is shown, but the rest of the samples follow a similar distribution of junctions. It can be seen that despite a great number of junctions are known (~50%), the sheep reference annotation is still in process and it suffers from a lack of annotation in some pathways since approximately 30% and 20% of the detected junction in our samples are completely novel or partially novel, respectively.

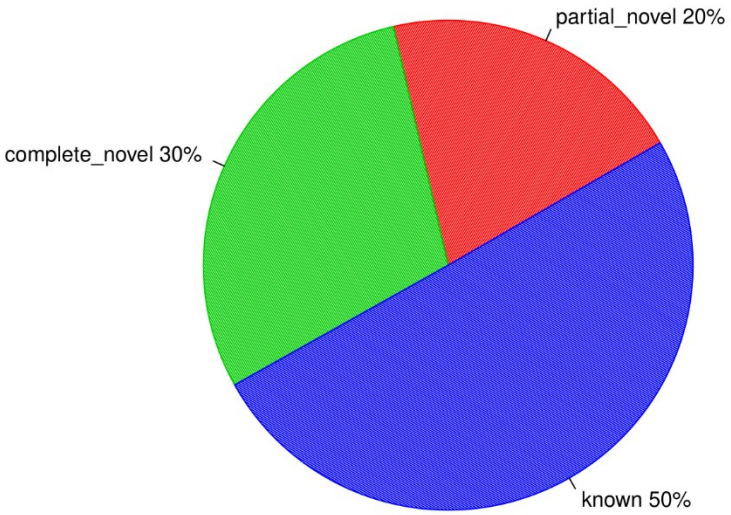


Figure 3.2: Detected splice junctions for sample 121-A. The junctions are divided as novel, partially novel (only one splice site is novel) and known (both splice sites are annotated in the reference genome).

The saturation plot shows how many features (in our case genes) are detected with more than a pre-defined count (more than 0 counts in our case) with the sequencing depth of the sample, and other sequencing depths are simulated from this total sequencing depth to see how many features would be detected with lower/higher depths. As it could be seen in figure 3.3, there would be few gains with higher sequencing depths for our samples, at least for the annotated genes. In addition, in the "sensitivity plot" (figure 3.4) can be seen depicted for each sample the percentage of features with more than 0, 1 ,2,5 and 10 counts per million (CPM). This plot can be used to select a filtering criterion for lowly expressed features, which are less reliable and can introduce noise into the differential expression analysis.

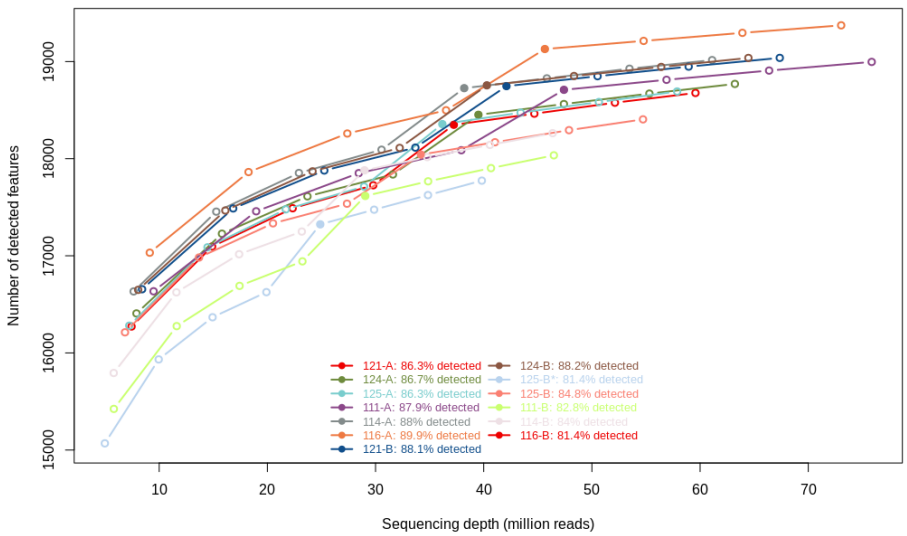


Figure 3.3: Saturation plot. The y-axis shows the number of detected features (genes) with more than 0 counts at different sequencing depths, x-axis, for each sample. Filled dots correspond to the values detected in our libraries, while the empty dots correspond to the simulated values.

Table 3.3: Summary statistics from the sequence alignment step for total RNA-seq data.

ID	Total Read-Pairs	Read-Pairs Surviving Trimming	Uniquely Mapped Read-Pairs	Read-Pairs Mapping to Multiple Loci	Unmapped Read-Pairs
121-A	64357288	63394838 (98.50%)	53552843 (84.48%)	6764150 (10.67%)	2884465 (4.55%)
124-A	68629982	67702143 (98.65%)	56800802 (83.90%)	7500326 (11.08%)	3202311 (4.73%)
125-A	61903639	60920287 (98.41%)	51751591 (84.95%)	6417402 (10.53%)	2583020 (4.24%)
111-A	85026279	83827905 (98.59%)	71104450 (84.82%)	9181708 (10.95%)	3302819 (3.94%)
114-A	68079080	67073386 (98.52%)	56167786 (83.74%)	7895668 (11.77%)	2830497 (4.22%)
116-A	80035364	78642855 (98.26%)	65954369 (83.87%)	9596562 (12.20%)	2870464 (3.65%)
121-B	73966873	72364211 (97.83%)	60629956 (83.78%)	8062333 (11.14%)	3422827 (4.73%)
124-B	70779453	69135213 (97.68%)	58449065 (84.54%)	7472133 (10.81%)	2993555 (4.33%)
125-B	60098045	59571141 (99.12%)	50176958 (84.23%)	7023235 (11.79%)	2162432 (3.63%)
125-B*	68718234	67334569 (97.99%)	38560204 (57.27%)	20816453 (30.91%)	7352935 (10.92%)
111-B	69869296	68584349 (98.16%)	42759051 (43.02%)	17835426 (34.14%)	11227258 (16.37%)
114-B	72223271	71055465 (98.38%)	43208875 (60.81%)	20973868 (29.52%)	6323936 (8.90%)
116-B	62221512	61215065 (98.38%)	36354141 (59.39%)	19189726 (31.35%)	5148187 (8.41%)

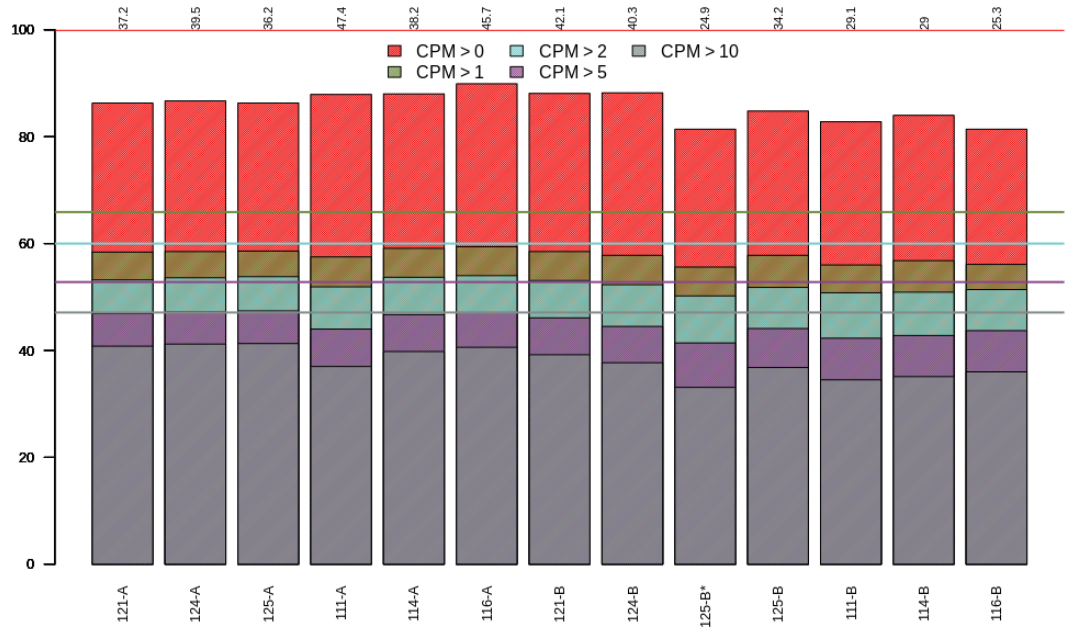


Figure 3.4: Sensitivity plot. The bars show the percentage of annotated features within each sample with more than 0, 1, 2, 5 and 10 CPM. In the upper side of the plot, the sequencing depth of each sample (in millions) is given. The horizontal lines are the corresponding percentage of features with those CPM in at least one of the samples.

It must be pointed that from 27,054 annotated genes in Ensembl, 21,274 (78.63%) were expressed with at least one sequence read count in one of the 13 RNA-seq libraries. Detected genes whose expression was lower than 2 CPM and could be found in less than 6 individual libraries were treated as lowly expressed genes and were removed from the differential expression analysis. These cut-offs were selected after checking that less stringent criteria introduced genes with high variability and expressed in only a few animals of each group. Those genes may not provide enough statistical evidence for reliable judgments and may confound the differential expression analysis if left in the data (350). After filtering lowly expressed genes, 11,395 (42.12%) remained for subsequent analyses.

3.3.1.3 Differential expression analysis

Prior to the differential expression analysis, the *svaseq* function from the SVA package was applied to remove unwanted variation and accurately measure the biological variability. The obtained surrogate variables were incorporated into the testing model of the DE analysis. A principal component analysis (PCA) was done with the corrected data (see figure 3.5). The different groups in our data are clearly depicted separately by the main components, pointing towards differences in gene expression. It must be pointed that samples from the vaccine and adjuvant groups at the initial time are not expected to be too different (apart from animal specific differences), since both groups are samples before the inoculation of any vaccine component.

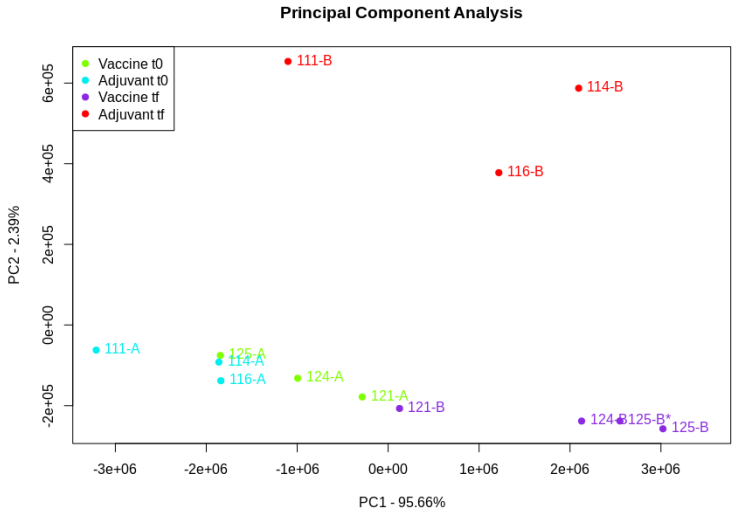


Figure 3.5: Principal Component Analysis (PCA) in total RNA-seq data from sheep PBMCs after the batch effect removal with SVA R package.

After this first exploratory analysis, three different programs (edgeR, DESeq2 and limma) were selected for the differential expression analysis. The same design model, previously described in Chapter 2, was applied in all tools and those genes with an adjusted p-value<0.05 and a fold-change>1.5 or <0.667 in all tools at the same time were designated as true DEGs. It must be pointed that, once the intersection of all the tools is done, the results from edgeR were taken as reference. As previously described in Chapter 2, edgeR and DESeq2 can estimate the dispersion of the model by three different manners: a common dispersion for all genes, a trended dispersion which depends mostly on the expression levels of each gene and a gene wise dispersion, in which a different dispersion is estimated for each gene. The estimated dispersion values by edgeR can be seen in Figure 3.6. This kind of plot can help to decide if the chosen fit model is good and if there is any suspicious data in our samples. Generally, for RNA-seq data, it is expected for the trended dispersion to decrease smoothly with abundance and to asymptotic to a constant value for highly expressed genes (350). For the DE analysis, the gene wise dispersion was used, which is the default and recommended choice in edgeR.

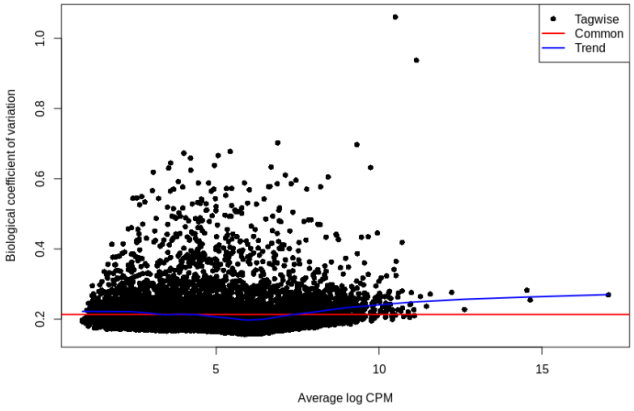


Figure 3.6: Biological coefficient of variation (square root of the negative binomial dispersion) against gene abundance (in log2 count per million) in PBMC samples. The red line is the estimated common BCV, the blue line the estimated trended BCV and the black dots the gene wise (or Tagwise) BCV.

Three different comparisons were made, mainly: Vac Tf vs. Vac T0, Adj Tf vs. Adj T0 and Adj Tf vs. Vac Tf. Thus, 2,473 DEGs were identified in the Vac Tf vs. Vac T0 comparison (Figure 3.7A), of which 1,208 and 1,265 displayed increased and decreased expression, respectively. Showing a similar pattern, 2,980 DEGs were identified in the Adj Tf vs. Adj T0 comparison (Figure 3.7B), of which 1,474 and 1,506 were upregulated and downregulated, respectively. Finally, in the Adj Tf vs. Vac Tf comparison, 429 DEGs were identified (Figure 3.7C), of which 132 were upregulated and 297 were downregulated. The exact results by gene of the DE analysis are available as a supplementary table in a previously published publication (347). In addition, the achieved results can be seen as MA plots in Figure 3.8, which serves as a way to visualize the differences between counts from two groups.

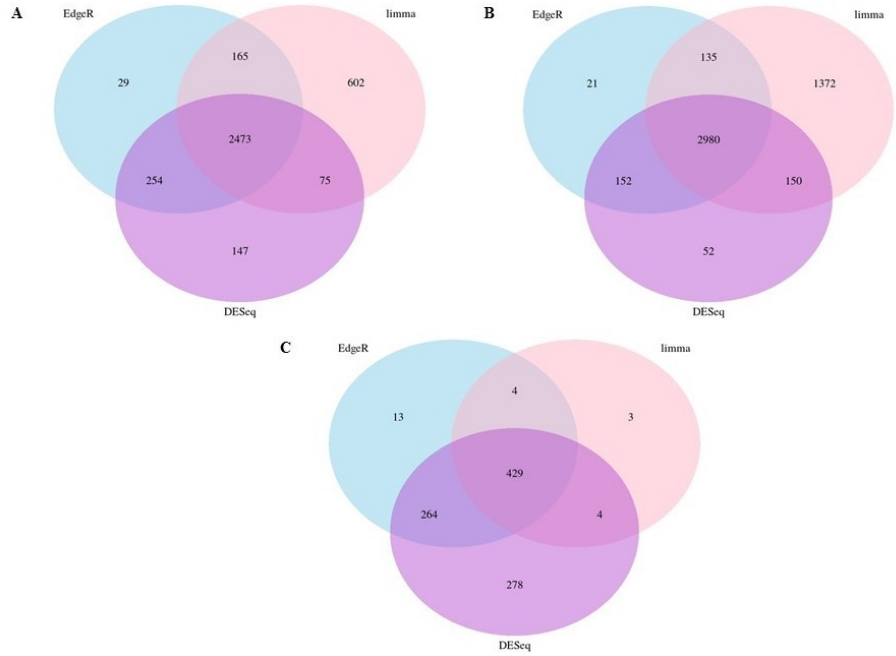


Figure 3.7: Venn diagrams depicting the differentially expressed genes detected with edgeR, DESeq2 and limma in total RNA-seq data. A) Vac Tf vs. Vac T0 comparison. B) Adj Tf vs. Adj T0 comparison. C) Adj Tf vs. Vac Tf comparison.

Among the most significant DE genes in each comparison (Figure 3.9), there are factors that are clearly related to apoptosis (*TP53BP2*, *CSRNP1*, *TEAD3*, *CDCA7*, *PPP1R15A*), immune response (*OSM*, *AMPD3*, *BTLA*, *SKAP2*, *IGSF6*, *LST1*, *FGR*, *MAPK13*), regulation of inflammatory response (*CD40*, *S100A12*, *ADGRE3*, *TREM1*, *STEAP4*, *NR4A3*), DNA replication and repair (*FEN1*, *HIST1H4L*), cell growth (*ARID5A*, *VPS37B*, *HGF*, *CSF3R*), cell adhesion and cell signalling (*NRXN2*, *CLEC12A*, *AREG*), nervous system development (*RAPGEF*, *CASZ1*, *EGR2*, *L1CAM*), and a gene involved in the pathogenesis of Alzheimer's disease (*APBB1*).

3.3.1.4 Functional enrichment analysis

In addition to check the most significant differentially expressed genes, in an attempt to decipher the functions of DEGs, a functional enrichment analysis was performed with PANTHER and DAVID tools. For the three main domains of the GO database (cellular component, molecular function and biological process), the PANTHER webtool was used for each list of DEGs (three in total). In all comparisons clearly appeared different terms related to the immune system,

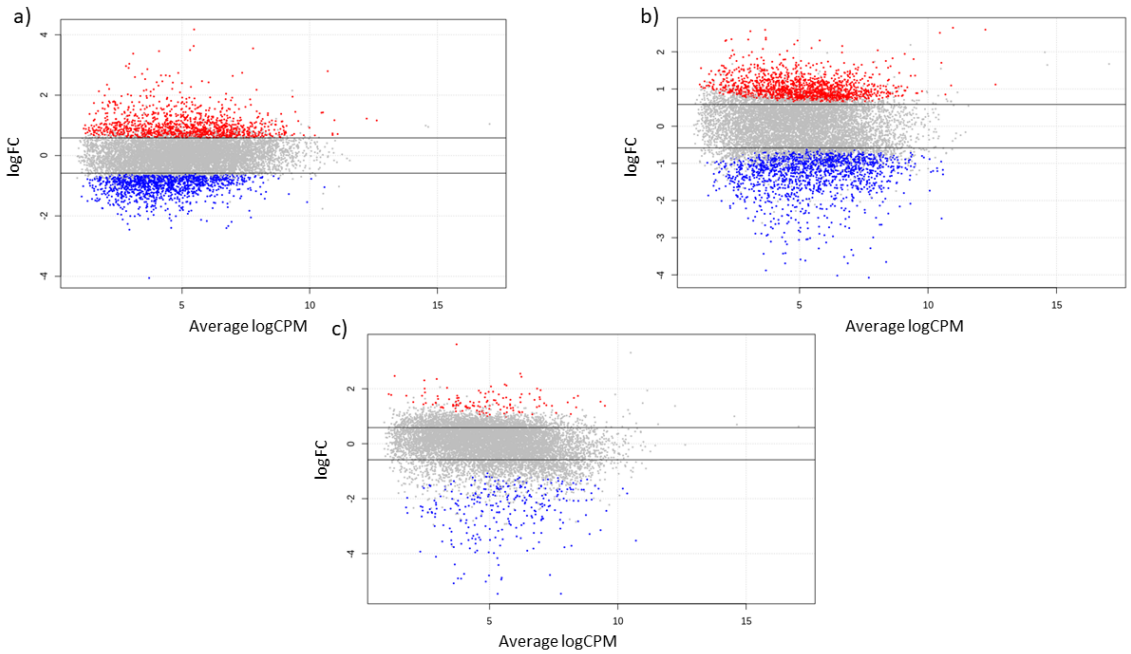


Figure 3.8: Scatter plots with the average expression (in log counts per million) of genes on the x-axis and the fold change (in log2) in the y-axis. These plots are named MA plots. Blue dots represent significant (in edgeR, DESeq2 and limma) downregulated genes in their corresponding comparison, while red dots represent significant upregulated genes. Horizontal lines correspond to FC=1.5 and FC=0.667. a) Vac Tf vs. Vac T0. b) Adj Tf vs. Adj T0. c) Adj Tf vs. Vac Tf.

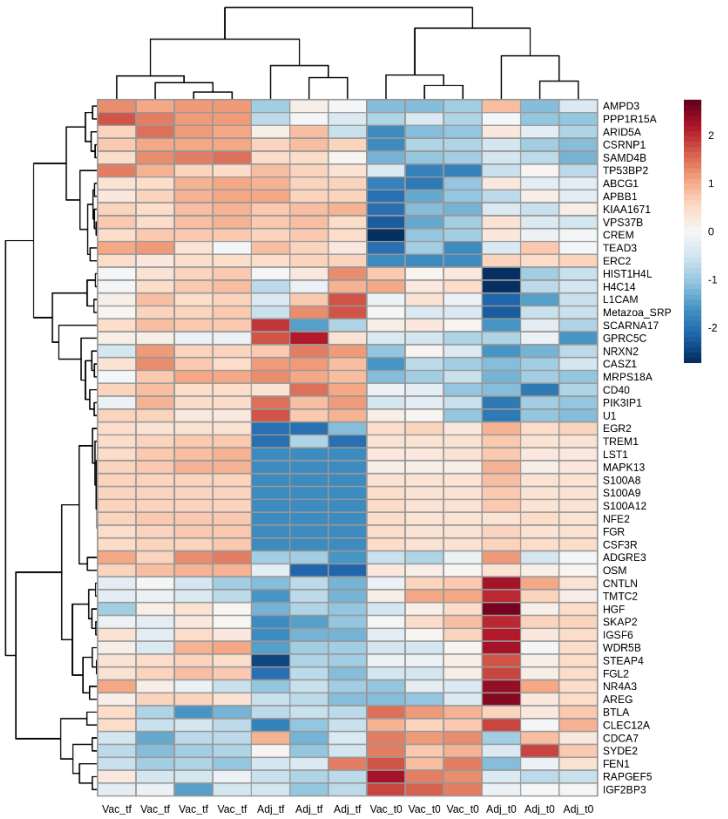


Figure 3.9: Heatmap with the log2(Fold Change) of the top 10 significant up- and down-regulated genes in the Vac Tf vs. Vac T0, Adj Tf vs. Adj T0, and Adj Tf vs. Vac Tf comparisons. The genes were selected from those found differentially expressed in 3 different programs: limma, edgeR and DESeq2.

apoptotic signalling pathway or response to DNA damage. In the Vac Tf vs. Vac T0 comparison, 46 significantly over-represented GO terms were found in total. Among the terms from the Biological Process (BP) category (Figure 3.10), terms such as *intracellular signal transduction* (GO:0035556), *cellular response to lipopolysaccharide* (GO:0071222), *regulation of cytokine production* (GO:00018117), *DNA repair* (GO:0006281) and *regulation of defense response* (GO:0031347) were identified. In addition, 72 enriched GO terms were identified in the Adj Tf vs. Adj T0 comparison. Among the terms from the BP ontology (Figure 3.11), there were *positive regulation of GTPase activity* (GO:0043547), *regulation of cellular response to stress* (GO:0080135), *cellular response to DNA damage stimulus* (GO:0006974), *positive regulation of proteolysis* (GO:0045862), *regulation of apoptotic process* (GO:0042981), *cellular response to chemical stimulus* (GO:0070887), *regulation of autophagy* (GO:0010506) and *regulation of immune system process* (GO:0002682). Both treatments have terms related to DNA damage in common. Lastly, in the Adj Tf vs. Vac Tf comparison, 23 overrepresented terms were identified (Figure 3.12), including *positive regulation of cytokine production* (GO:0001819), *positive regulation of immune system proces*s (GO:0002684), *inflammatory response* (GO:0006954), *immune response* (GO:0006955), *regulation of response to external stimulus* (GO:0032101), *cellular response to cytokine stimulus* (GO:0071345), *and neutrophil chemotaxis* (GO:0030593).

DAVID tools was used for the enrichment analysis of KEGG pathways, revealing overrepresented pathways related to the immune system, inflammatory response and some autoimmune diseases. The results were represented as networks with Cytoscape for each comparison: Vac Tf vs. Vac T0 (Figure 3.13), Adj Tf vs. Adj T0 (Figure 3.14) and Adj Tf vs. Vac Tf (Figure 3.15). In both treatment groups (adjuvant- and vaccine-treated animals), the *NF-κB signaling pathway* was enriched. Other pathways enriched exclusively in the Vac Tf vs. Vac T0 comparison were: *TNF signaling pathway*, *Toll-like receptor signaling pathway*, *p53 signaling pathway*, *DNA replication*, *purine metabolism* and *endocytosis*. Furthermore, in the Adj Tf vs. Adj T0 comparison, immune related pathways such as *T cell receptor signaling pathway* and *B cell receptor signaling pathway* were enriched. Seeing the enriched GO terms and KEGG pathways, it is clear that the treatments, including the aluminium adjuvant without the presence of any antigen, are able to stimulate the immune response. Notably, as can be seen in the enriched KEGG pathway *cytokine-cytokine receptor interaction* in the Adj Tf vs. Vac Tf comparison, nearly all cytokines and cytokine receptors are downregulated in the adjuvanted animals, except *CCR6*.

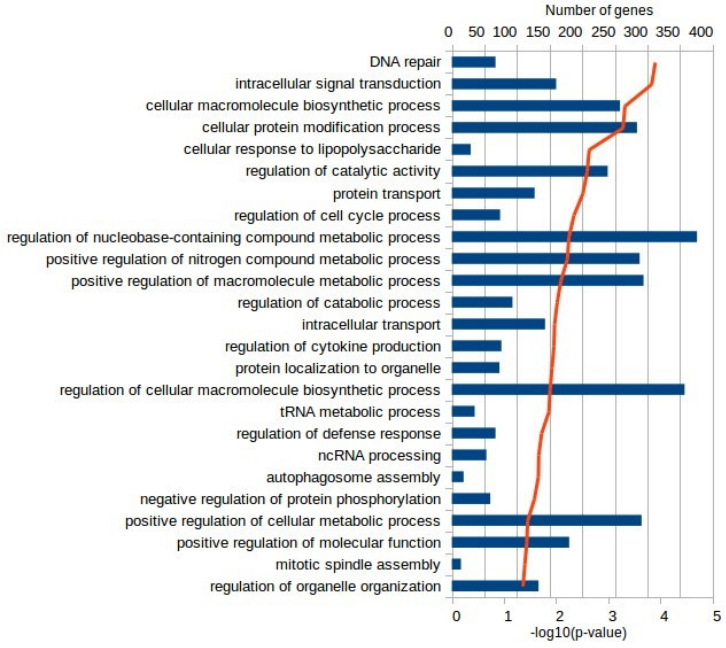
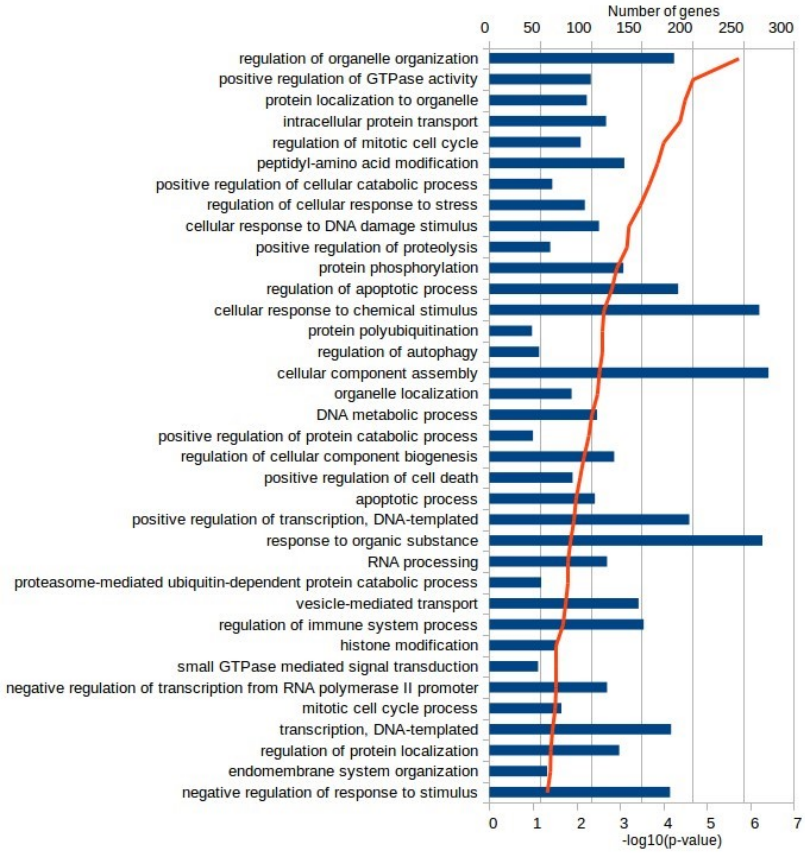


Figure 3.10: Enriched GO terms from the BP ontology in the Vac Tf vs. Vac T0 comparison in PANTHER. The significance is calculated with a Fisher's exact test and adjusted for multiple comparison correction with Benjamini-Hochberg False Discovery Rate correction. The blue bars depict number of DEGs in the corresponding term, while the red line depicts the adjusted p-value (after -log10 transformation).

Figure 3.11: Enriched GO terms from the BP ontology in the Adj Tf vs. Adj T0 comparison in PANTHER. The significance is calculated with a Fisher's exact test and adjusted for multiple comparison correction with Benjamini-Hochberg False Discovery Rate correction. The blue bars depict number of DEGs in the corresponding term, while the red line depicts the adjusted p-value (after -log10 transformation).



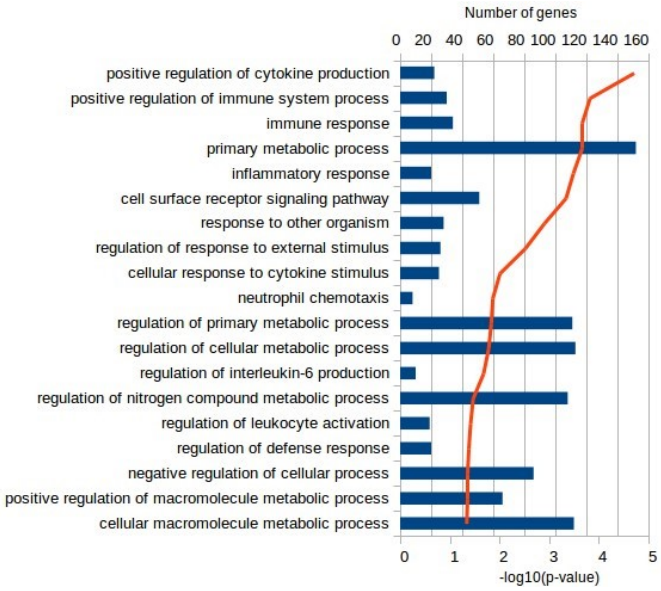


Figure 3.12: Enriched GO terms from the BP ontology in the Adj Tf vs. Adj T0 comparison in PANTHER. The significance is calculated with a Fisher's exact test and adjusted for multiple comparison correction with Benjamini-Hochberg False Discovery Rate correction. The blue bars depict number of DEGs in the corresponding term, while the red line depicts the adjusted p-value (after -log10 transformation).

In concordance with what have been seen in other studies, aluminium adjuvants are able to upregulate NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) in both groups (351,352). NF-κB is a family of structurally-related transcription factors that regulates a great variety of genes from the innate and adaptative immune response. Among the functions of NF-κB targeted genes, there are processes such as immune and inflammatory response, cellular growth, stress response, proliferation, differentiation, development and apoptosis (353). The expression of NF-κB targeted genes which are differentially expressed in our study in at least one comparison can be seen as a radar plot in figure 3.16A. Among the NF-κB target

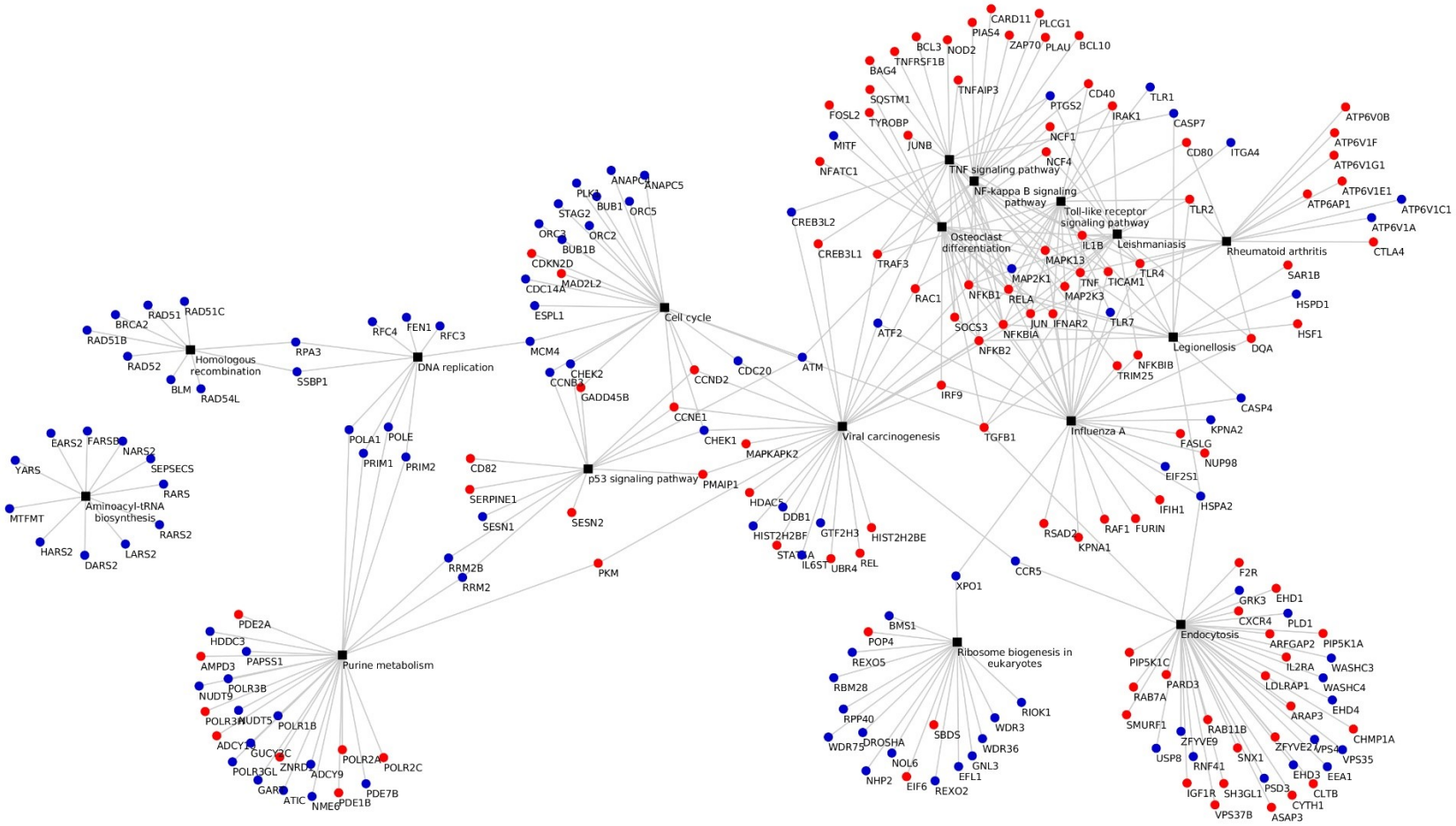


Figure 3.13: Enriched KEGG pathways in the Vac Tf vs. Vac T0 comparison in DAVID tools with EASE score (a modified Fisher's exact test) and Benjamini-Hochberg False Discovery Rate (FDR) correction. Black boxes represent enriched pathways and points differentially expressed genes in each pathway, up-regulated ones in red and down-regulated ones in blue.

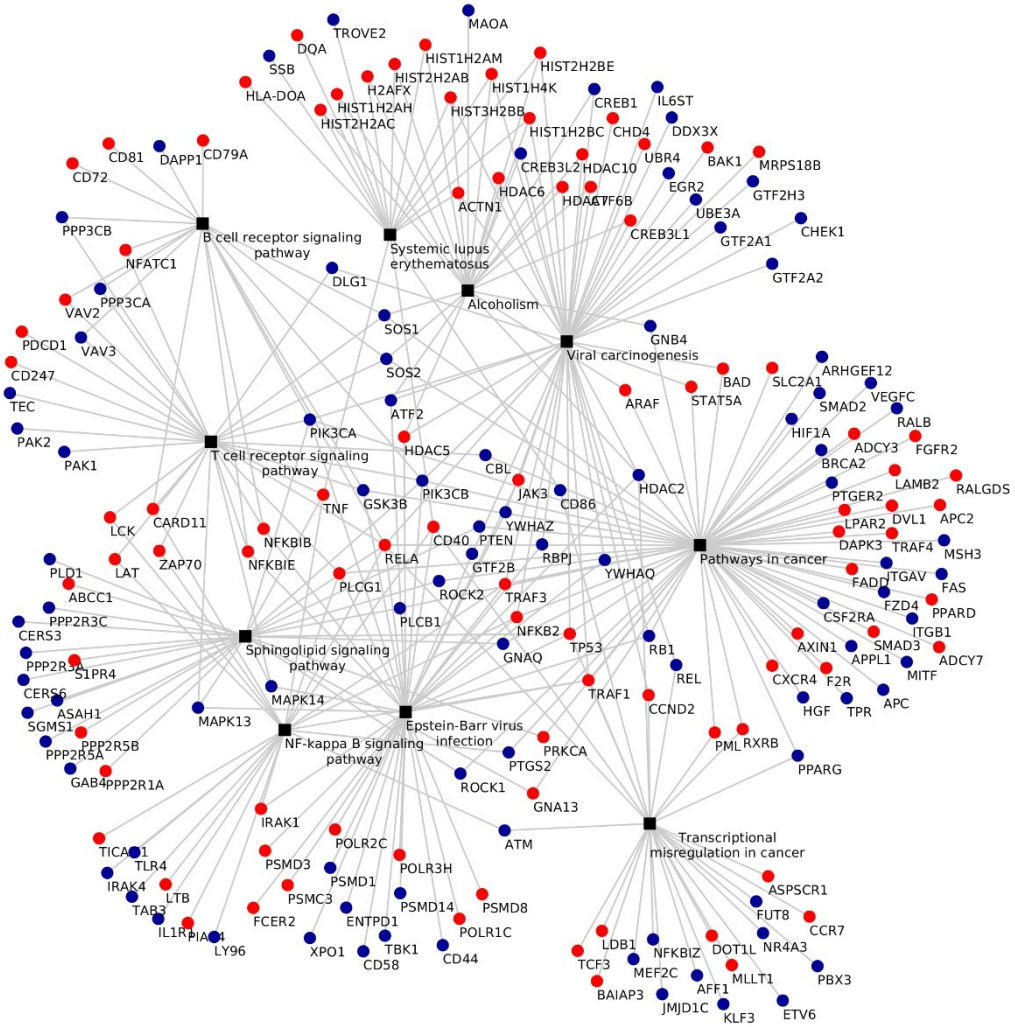


Figure 3.14: Enriched KEGG pathways in the Adj Tf vs. Adj T0 comparison in DAVID tools with EASE score (a modified Fisher's exact test) and Benjamini-Hochberg False Discovery Rate (FDR) correction. Black boxes represent enriched pathways and points differentially expressed genes in each pathway, up-regulated ones in red and down-regulated ones in blue.

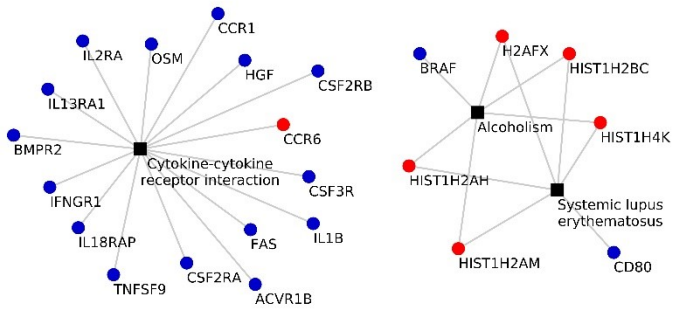


Figure 3.15: Enriched KEGG pathways in the Adj Tf vs. Adj T0 comparison in DAVID tools with EASE score (a modified Fisher's exact test) and Benjamini-Hochberg False Discovery Rate (FDR) correction. Black boxes represent enriched pathways and points differentially expressed genes in each pathway, up-regulated ones in red and down-regulated ones in blue.

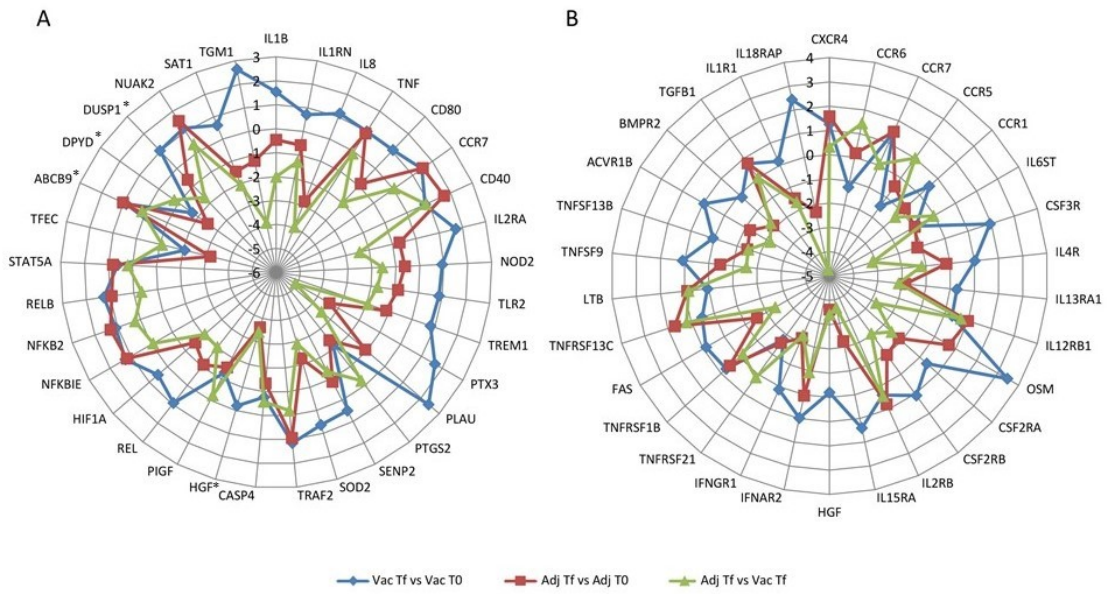


Figure 3.16: Radar plots with the log2(Fold Change) of differentially expressed genes in multiple pathways. Each colour line represents log2(Fold Change) values in a different comparison: Vac Tf vs. Vac T0 (blue), Adj Tf vs. Adj T0 (red) and Adj Tf vs. Vac Tf (green). A) Differentially expressed genes that are targeted by NF-κB (list of target genes obtained from https://www.bu.edu/nf-kb/gene-resources/target-genes/). * indicates that the gene has a κB site in the promoter, but the gene has not clearly been shown to be controlled by NF-κB. B) Differentially expressed genes related to the *cytokine-cytokine receptor interaction pathway*.

genes are cytokines/chemokines and their modulators (*IL1B*, *IL1RN*, *IL8* and *TNF*), immunoreceptors (*CD80*, *CCR7*, *CD40*, *IL2RA*, *NOD2*, *TLR2*, *TREM1*), acute phase proteins (*PTX3*, *PLAU*), stress response genes (*PTGS2*, *SENP2*, *SOD2*), regulators of apoptosis (*TRAF2*, *CASP4*), growth factors (*HGF*, *PIGF*), transcription factors and their modulators (*REL*, *HIF1A*, *NFKBIE*, *NFKB2*, *RELB*, *STAT5A*, *TFEC*), and enzymes (*ABCB9*, *DPYD*, *DUSP1*, *NUAK2*, *SAT1*, *TGM1*). Furthermore, within the DEGs related to the cytokine-cytokine receptor interaction pathway, there are chemokines (*CXCR4*, *CCR6*, *CCR7*, *CCR5*, *CCR1*, *IL8*), haematopoietins (*IL6ST*, *CSF3R*, *IL4R*, *IL13RA1*, *IL12RB1*, *OSM*) and genes belonging to the platelet-derived growth factor (PDGF) family (*CSF2RA*, *CSF2RB*, *IL2RA*, *IL2RB*, *IL15RA*, *HGF*), interferon family (*IFNAR2*, *IFNGR1*), tumour necrosis factor (TNF) family (*TNFRSF21*, *TNFRSF1B*, *FAS*, *CD40*, *TNFRSF13C*, *TNF*, *LTB*, *TNFSF9*, *TNFSF13B*), transforming growth factor beta (TGFB) family (*ACVR1B*, *BMPR2*, *TGFB1*), and interleukin-1 (IL1) family (*IL1R1*, *IL18RAP*, *IL1B*).

The vaccination induced a clear upregulation of *IL1B*, *IL2RA* and *PTX3*. *IL1B* (interleukin-1 beta) is a potent pro-inflammatory cytokine produced by activated macrophages who is able to induce neutrophil recruitment and activation (354), T/B cell activation and has a central role in the differentiation of T_H17 cells (355). In addition, *IL2RA* (a subunit of the high affinity receptor for interleukin-2) expression is constitutive in regulatory T cells (T_{reg}) (356), while *PTX3* is expressed and released by cells of the monocyte-macrophage lineage exposed to inflammatory signals (357). Taken together, the expression of these genes is consistent with the induction of an ongoing immune response against the vaccine. In contrast, in the animals inoculated with aluminium alone (without any antigen), the expression of some proinflammatory gene mRNAs were downregulated (e.g., *IL1B*, *IL8*, *TLR2*, *NOD2*, *IL2RA*), suggesting a milder induction of the immune response. Cytokine receptor interaction in vaccinated animals evidenced the induction of *IL18RAP*, involved in sensing the proinflammatory *IL18* cytokine, and *CSF3R*, which is the

receptor for granulocyte colony stimulation factor (G-*CSF*), a key cytokine that controls myeloid cell function.

3.3.1.5 Weighted gene correlation network analysis (WGCNA)

A weighted gene co-expression network analysis was performed with the WGCNA [v1.63] (279,358) R package. This kind of 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. It must be pointed that in addition to annotated gene data, expression data of detected new lncRNAs (data not shown) was used for network construction. Briefly, a similarity matrix was constructed from normalized data using the biweight midcorrelation, which was chosen for its robustness against outliers in comparison to Pearson correlation. Then, the similarity matrix was raised to a power β to calculate the adjacency matrix. The parameter β needs to be selected based on the minimum value required to get a scale-free topology network, which correspond to a scale-free topology fit index R^2>0.8. If the index fails to reach values over 0.8 for reasonable powers (less than 15 for unsigned networks and less than 30 for signed networks, by authors recommendation), it may be indicative of a subset of samples with strong differences from the rest of samples. As can be seen in figure 3.17, there is not a power β lower than 30 that achieves a R^2>0.8. Taken the differential expression results (Figure 3.7), it is clear that there are strong expression changes between conditions. The lack of a scale-free topology does not invalidate the data by itself and, if the cause is an interesting variable, the authors provide different soft-thresholding powers to achieve a conservative network based on the type of network and the number of samples in the data. For a signed network constructed from less than 20 samples, a β parameter of 18 is recommended.

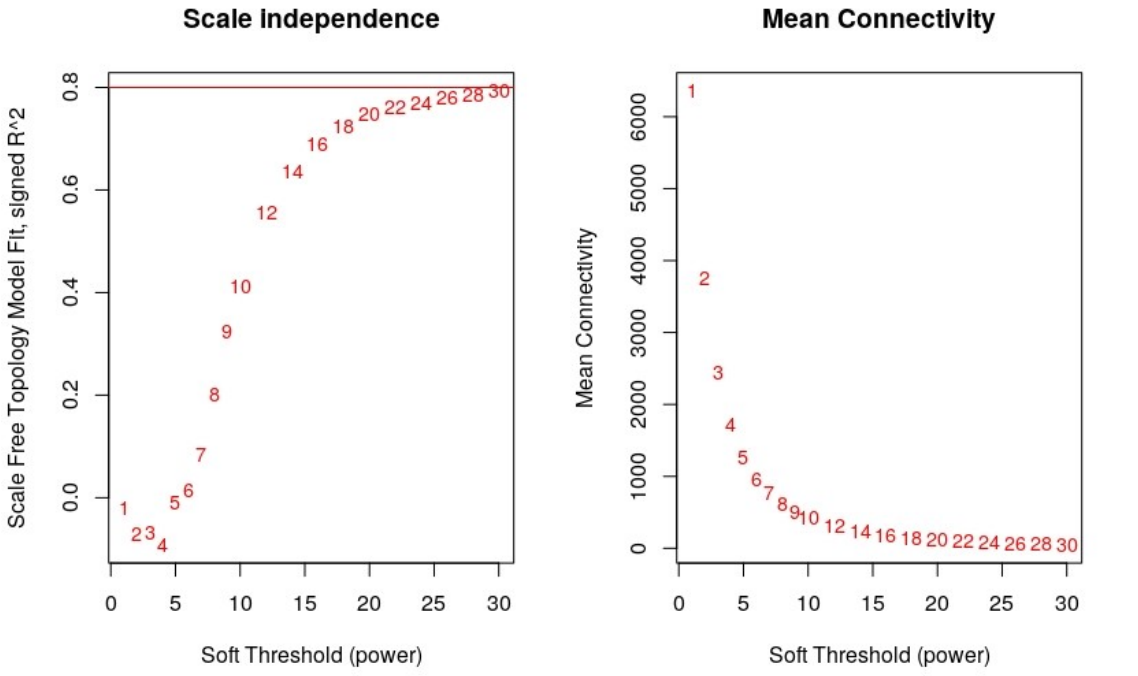


Figure 3.17: Summary network indices (y-axes) as functions of the soft thresholding power (x-axes). Numbers in the plots indicate the corresponding soft thresholding powers. For each power, the scale-free topology fit index (R^2) is calculated (left panel) and returned along the network mean connectivity (right panel).

Once an adequate β parameter was chosen and the network was constructed, different modules (clusters of densely interconnected genes) were detected by an unsupervised hierarchical clustering algorithm (*hclust* function in R). The TOM metric was used for the clustering as a dissimilarity matrix by subtracting it from 1. In addition, a minimum size of 30 genes was required for a module to be reported. A total of 106 co-expressed gene modules were detected. Then, modules with similar expression profiles were merged based on a height cut-off threshold of 0.3. In the end, a total of 32 co-expressed gene modules were detected (figure 3.18(a) and 3.18(b)), module size ranging from 39 to 1,956 genes. Each module was assigned a random colour name. Then, significant Pearson's correlations among module eigengenes (the first principal component of each module) and treatment variables (all possible dichotomized combinations, in which one group is against the other two. Treat, samples at the start of the experiment against samples at the end; TreatVac, samples from the vaccine group at the end of the experiment against the rest of samples; and TreatAdj, samples from the adjuvant group at the end of the experiment against the rest of samples) were searched. After selecting a q-value of 0.05 as cut-off, the following modules were found significant for each dichotomized treatment variable (Figure 3.18(c)): lavenderblush3 (1956 genes, r=0.86, q-value=0.004), darkgreen (1066 genes, r=0.67, q-value=0.04), plum1 (125 genes, r=0.7, q-value=0.04), coral1 (695 genes, r=-0.75, q-value=0.03), darkolivegreen2 (39 genes, r=-0.81, q-value=0.01), pink (1492 genes, r=-0.94, q-value=9e-05) and yellowgreen (1131 genes, r=-0.78, q-value=0.02) for Treat variable; plum3 (64 genes, r=-0.73, q-value=0.03), darkred (223 genes, r=-0.68, q-value=0.04), grey60 (751 genes, r=-0.71, q-value=0.03), thistle (188 genes, r=-0.67, q-value=0.04) and pink (1492 genes, r=-0.68, q-value=0.04) for TreatVac variable; and lavenderblush3 (1956 genes, r=0.69, q-value=0.04), salmon4 (280 genes, r=0.74, q-value=0.03), skyblue3 (126 genes, r=0.65- q-value=0.05), antiquewhite4 (254 genes, r=0.79, q-value=0.02) and lightpink4 (976 genes, r=-0.73, q-value=0.03) for TreatAdj. In addition, it was checked how many DEGs were inside each significant module and it was shown that all those modules had from 10 to 1111 DEGs. The most outstanding modules were lavenderblush3, coral1, pink, yellowgreen and lightpink4 with 1111 from 1956, 370 from 695, 803 from 1492, 434 from 1131 and 349 from 976 differentially expressed genes, respectively.

The obtained treatment associated modules were further studied for enrichment of GO terms and KEGG pathways. Among the modules correlated to both treatments (Treat), lavenderblush3 and darkgreen showed positive correlation, which indicates that genes belonging to those modules usually have higher expressions at the end of the experiment, while coral1, pink and yellowgreen showed negative correlation, which indicates a lower expression at the end of the experiment. Among terms from the BP ontology, lavenderblush3 module was enriched in *thymic T cell selection* (GO:0045061), *interstrand cross-link repair* (GO:0036297), *T cell receptor signaling pathway* (GO:0050852) and *intrinsic apoptotic signaling pathway by p53 class mediator* (GO:0072332); darkgreen module was enriched in *endoplasmic reticulum (ER) overload response* (GO:0006983) and *regulation of type I interferon production* (GO:0032479); coral1 module was enriched in multiple processes such as immune response (*immune response-activating signal transduction* (GO:0002757), *positive regulation of innate immune response* (GO:0045089) and *positive regulation of cytokine production* (GO:0001819)), T cell functions (*regulation of T-helper 17 cell differentiation* (GO:2000319)), inflammation (*negative regulation of acute inflammatory response* (GO:0002674)), MAPK and JNK cascades, cell proliferation and motility; the most correlated pink module (r=-0.94) was enriched in *DNA repair* (GO:0006281), *methylation* (GO:0032259) and *cellular protein modification process* (GO:0006464); and the

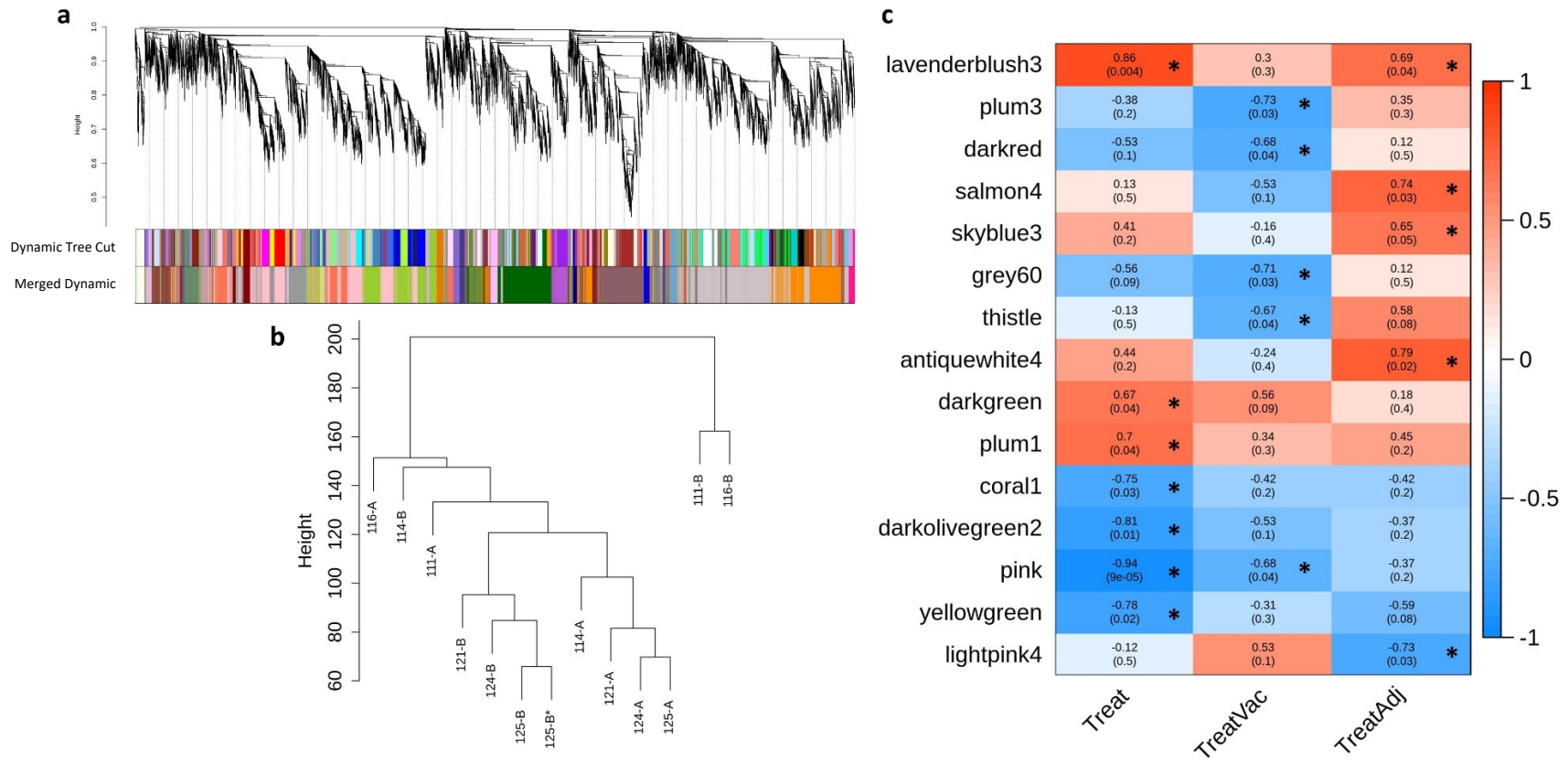


Figure 3.18: Weighted gene expression co-variance network analysis (WGCNA) summary. (a) Gene dendrogram obtained by average linkage hierarchical clustering. The colour rows underneath the dendrogram shows the module assignment before (Dynamic Tree Cut) and after (Merged Dynamic) modules with similar 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 (color-coded) and adjusted p-values. Only modules associated with at least one trait are shown (significant ones marked with and asterisk).

yellowgreen module was enriched in terms related to mRNA processing and protein regulation such as *ubiquitin-dependent ERAD pathway* (GO:0030433), *negative regulation of intracellular protein transport* (GO:0090317), *protein folding* (GO:0006457), *protein localization to organelle* (GO:0033365), *RNA splicing* (GO:0008380) and *mRNA processing* (GO:0006397). Finally, the lightpink4 module, which was negatively correlated with adjuvant only inoculated animals (TreatAdj), was enriched in terms related to response to external stimuli, cytokines (*cellular response to chemokine* (G0:1990869), *regulation of cytokine biosynthetic process* (GO:0042035), *cytokine-mediated signaling pathway* (GO:0019221) and *regulation of cytokine secretion* (GO:0050707)) and immune cell regulation (*negative regulation of leukocyte differentiation* (GO:1902106), *regulation of interferon-gamma production* (GO:0032649), *regulation of interleukon-6 production* (GO:0032675), *positive regulation of leukocyte differentiation* (GO:1902107), *regulation of inflammatory response* (GO:0050727),…). Of all the modules related to a treatment variable, it is clear that lavenderblush3 and coral1 modules are composed of genes crucial for the correct function of the immune system and that those genes are similarly expressed in both treatments, indicating that aluminium hydroxide regardless of the presence of any antigen is capable to activating an immune response. In addition, a negatively regulated module to TreatAdj such as lightpink4 is composed of genes related to cytokines that were downregulated in the Adj Tf vs. Vac Tf comparison. The rest of trait-associated modules had almost no significant enrichment, probably due to the small number of annotated genes in them.

Regarding KEGG pathway enrichment, few modules showed any significant results. The lavenderblush3 module was enriched in *viral carcinogenesis* (oas05203), the yellowdreen module in *RNA transport* (oas03013), *Parkinson's disease* (oas05012), *protein processing in endoplasmic reticulumoas* (oas04141), *ribosome biogenesis in eukaryotes* (oas03008), *RNA degradation* (oas03018), *spliceosome* (oas03040*), cell cycle* (oas04110) and *mRNA surveillance pathway* (oas03015), and the lightpink4 module in *osteoclast differentiation* (oas04380), *systemic lupus erythematosus* (oas05322), *HIF-1 signaling pathway* (oas04066) and *TNF signaling pathway* (oas04668).

3.3.2 miRNA-seq

3.3.2.1 Sequencing quality

For the same samples used in Total RNA-seq, 12 miRNA-seq libraries were prepared. After sequencing, a mean value of 17.2 million 50 nt single-end reads per library were achieved. Adaptor sequence trimming and low quality read filtering were done with Trimmomatic following criteria previously described (Chapter 2 – section 2.3.3.4), which resulted in a mean 14.2 (SD=4.4) million reads (82.38%) of good quality reads. The average quality of the trimmed samples can be seen in figure 3.19. As it can be seen, all samples have good quality after trimming.

Trimmed reads were aligned to the Ovis aries reference genome Oar_v3.1 from Ensembl with the sRNAbench module from sRNAtoolbox, which uses bowtie for read alignment. Up to 20 multiple mappings per read were allowed. The alignment yielded a mean value of 12.9 (SD=4.2) million read pairs (91.40% of the filtered reads). A more detailed summary of the alignment can be seen for each sample in table 3.4. It must be pointed that for small RNA annotation, all sequences were searched in miRbase to identify annotated miRNAs and in Rfam to identify other small RNAs originating from rRNA, tRNA, snRNA and snoRNA. As can be seen in figure 3.20, 39.43% of all successfully aligned reads were annotated miRNAs from the miRbase database,

10.87% were annotated sheep small nucleolar RNAs (snoRNAs), 3.10% to tRNAs, 22.95% to other RNAs of RNACentral and few reads were assigned to other small RNAs such as rRNAs and snRNAs. The detected miRbase miRNA expression values were taken for the expression matrix construction and the unassigned reads (23.32%) were used for novel miRNA prediction.

Figure 3.19: Average quality per base for all sequenced samples. The y-axis shows the average quality in Phred scale, while the x-axis is a representation of all bases in a sequencing read of 75 nt for each pair. The background of the graph is divided by different colours in three main section, with the green section indicating good quality bases, the orange reasonable quality bases and the red one poor quality bases. The per sample quality plots were produced by FASTQC and then, they were aggregated by MultiQC.

Table 3.4: Summary statistics from the sequence alignment step for miRNA-seq data.

ID	Total Reads	Reads Surviving Trimming	Mapped Reads
121-A	12366489	10133595 (81.94%)	9186077 (90.65%)
124-A	17380329	14569307 (83.83%)	13340726 (91.57%)
125-A	16571561	13662699 (82.45%)	12384329 (90.64)
111-A	18387274	15449821 (84.02%)	14029289 (90.80)
114-A	24859612	21630010 (87.01%)	20245260 (93.60%)
116-A	22502245	19329539 (85.90%)	18107867 (93.68%)
121-B	18715220	14364474 (76.75%)	12924511 (89.97%)
124-B	24297910	20905227 (86.04%)	18987581 (90.83%)
125-B	14680915	11702300 (79.71%)	10534334 (90.02%)
111-B	12209255	9430188 (77.24%)	8541387 (90.57%)
114-B	11878265	9068630 (76.35%)	8111086 (89.44%)
116-B	12913201	10094890 (78.17%)	9301333 (92.14%)

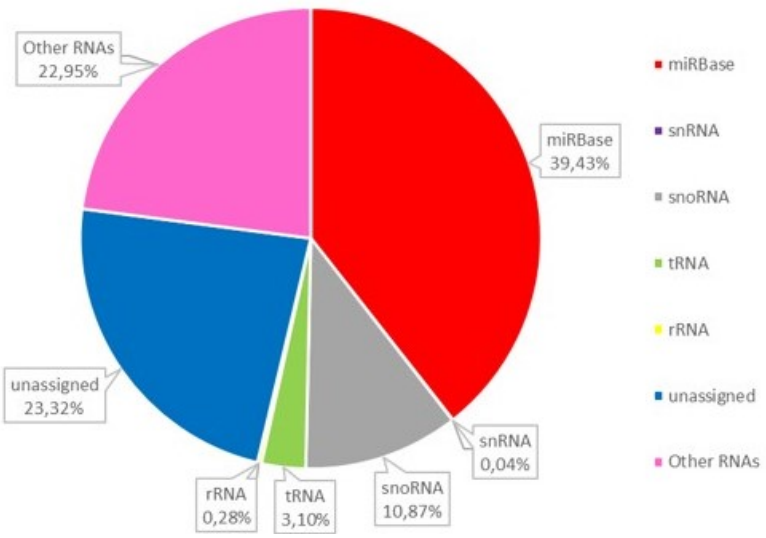


Figure 3.20: Class of molecules to which miRNA sequencing reads align.

Apart from sequence alignment to the reference, the sRNAbench module is capable of new miRNA prediction from unassigned reads. In total, 56 annotated *Ovis aries* miRNAs were expressed with at least one sequence read count in at least one of the 12 sample libraries. Furthermore, 39 new miRNAs were predicted with at least one sequence read in one sample. These miRNAs were blasted against miRbase miRNAs from other species and miRNAs described in the RNACentral database, to verify if they have been previously described or if there are homologous sequences in other species. Of these 39 new miRNAs, 11 were completely homologous to other miRNAs described in miRbase for *Bos Taurus*, another 8 were previously described in other sheep studies from RNACental, other was described in *Canis lupus* and another one in *Cervus elephus*. The rest were completely new miRNAs. More information of the new miRNAs can be seen as supplementary material in the corresponding publication (359). The length distribution of all miRNAs was checked, and it was shown that the majority of reads has a size of 21-24 nt, a range distribution common in mammalian miRNAs (360).

3.3.2.2 Differential expression analysis

Those miRNAs with an expression lower than 1 CPM and detected in <6 individual libraries were treated as lowly expressed and they were filtered out for further analysis. In total, 64 miRNAs remained for the differential expression analysis. Prior to any other analysis, and similar to the total RNA-seq data, the *svaseq* function from the SVA package was applied to remove unwanted variation and accurately measure the biological variability. The obtained surrogate variables were incorporated into the testing model of the DE analysis. A principal component analysis (PCA) was done with the corrected data (see figure 3.21). It can be seen that that samples group together according to treatment condition.

Figure 3.21: Principal Component Analysis (PCA) in miRNA-seq data from sheep PBMCs after the batch effect removal with SVA R package.

After the surrogate variables where calculated, the differential expression analysis was performed with edgeR using the same testing model as in total RNA-seq data analysis. Those miRNAs with an adjusted p-value<0.05 and a fold-change>1.5 or <0.667 were taken. Thus, a total of 3, 6 and 1 differentially expressed miRNAs were identified in the Vac Tf vs. Vac T0, Adj Tf vs.

Adj T0 and Adj Tf vs. Vac Tf comparisons, respectively (Table 3.5). Among the differentially expressed miRNAs, there were molecules previously characterized in other studies: a cattle-specific miRNA (*miR-2284ab-5p*) from the largest miRNA family in cattle, whose predicted targets have been related to the insulin signalling pathway (a pathway known to contribute to metabolic differences between ruminants and non-ruminants) (361); *miR-125b*, which has been related to primary B cell differentiation (362,363); *miR-19b*, which together with *miR-19a*, has been related to T_H2 cytokine-promoting activity (364);and *miRNA-99a*, which has been related to inflammation (361).

Table 3.5: List of differentially expressed miRNAs detected by edgeR with an adjusted p-value <0.05 and a fold-change >1.5 or <0.667.

Vaccine tf VS Vaccine t0			Adjuvant tf VS Adjuvant t0			Adjuvant tf VS Vaccine tf		
miRNA	**logFC**	**FDR**	**miRNA**	**logFC**	**FDR**	**miRNA**	**logFC**	**FDR**
new-miR-2284ab-5p	-8,613	1,063E-06	oar-miR-25	2,171	2,003E-03	new-miR-2284ab-5p	12,072	1,483E-03
oar-miR-125b	2,225	6,024E-04	oar-miR-379-5p	-4,208	2,003E-03			
oar-miR-99a	1,654	1,369E-02	oar-miR-411a-5p	-6,556	6,028E-03			
			oar-miR-16b	1,732	2,023E-02			
			oar-miR-19b	-1,799	2,635E-02			
			oar-let-7b	1,576	3,259E-02			

3.3.2.3 Target prediction and miRNA-mRNA data integration

Three different target gene prediction programs (miRanda, PITA and TargetScan) were selected and applied to the differentially expressed miRNAs. The intersection of all tools was taken and treated as potential target candidates. miRNAs usually act via translational repression and/or mRNA cleavage, although there is evidence of miRNAs upregulating translation by diverse mechanism in specific situations (140,141). However, it must be determined whether activation of protein translation is a general phenomenon or is only an exception in the mechanism of miRNA action. For that reason, only those miRNA-target pairs with negative correlation were further studied. A total of 70 significant miRNA-target pairs with negative Spearman's rank correlation coefficient (rho) value between -0.853 and -0.657 were predicted (Figure 3.22). Among the predicted interactions, *oar-let-7b* had 33 predicted targets, while *oar-miR-25* and *oar-miR-125b* has 13 and 11 predicted targets, respectively. There were factors related to cellular response to DNA damage stimulus (*STXBP4*, *RNF169*, *ZBTB4*, *NFATC2*), positive regulation of cell migration (*RDX*, *ATP8A1*) and response to stimulus (*HSPA14*, *MAP3K2*, *CHEK1*, *MKNK1*, *ANTXR2*, *NBEAL1*, *NFATC2*).

3.3.3 Validation by RT-qPCR

For RNA-seq data validation, 9 mRNAs (*CNTLN*, *EGR2*, *GPRC5C*, *HGF*, *NRXN2*, *SAMD4B*, *SKAP2*, *TREM1*, *WDR5B*) were verified using the Fluidigm Biomark HD Nanofluidic qPCR system, while for miRNA-seq data 3 miRNAs (*oar-let-7b*, *oar-miR-19b*, *oar-miR-25*) were verified. Log2 fold changes (log2FC) in gene expression between the different groups calculated by RT-qPCR are shown in figure 3.23 for RNA-seq data and figure 3.24 for miRNA-seq data. In the case of RNA seq data, despite the log2FC values for the expression of some genes measured by RNA-seq or RT qPCR were different, in terms of log2FC direction, the gene expression patterns of most genes (6 in Vac Tf vs. Vac T0, 9 in Adj Tf vs. Adj T0, and 7 in Adj Tf vs. Vac Tf) (81.5%) were reproducible by the RT-qPCR analysis. In the case of miRNA-seq data, the results confirmed the upregulated

expression of 2 miRNAs (oar-let-7b and oar-miR-25) and downregulated expression of oar miR 19b. The miRNA data and RT-qPCR showed a high degree of concordance.

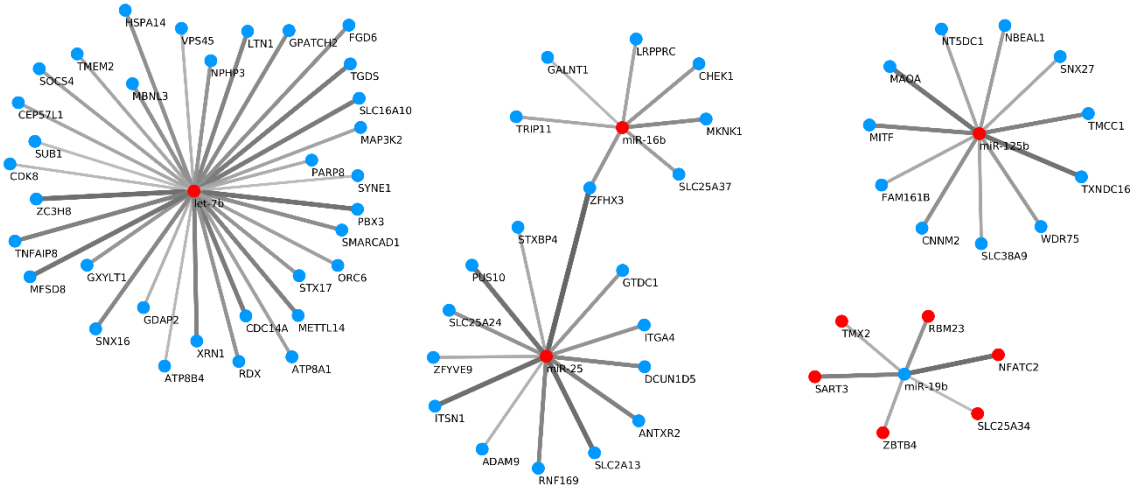


Figure 3.22: Significant negative correlation (with an adjusted p-value <0.05) between differentially expressed miRNAs and predicted targets represented as a network. Red points represent up-regulation in their corresponding comparison, while blue ones down-regulation. The grater the absolute value of the Spearman's rank correlation coefficient (rho), the broader and darker is the line joining the miRNA and predicted target.

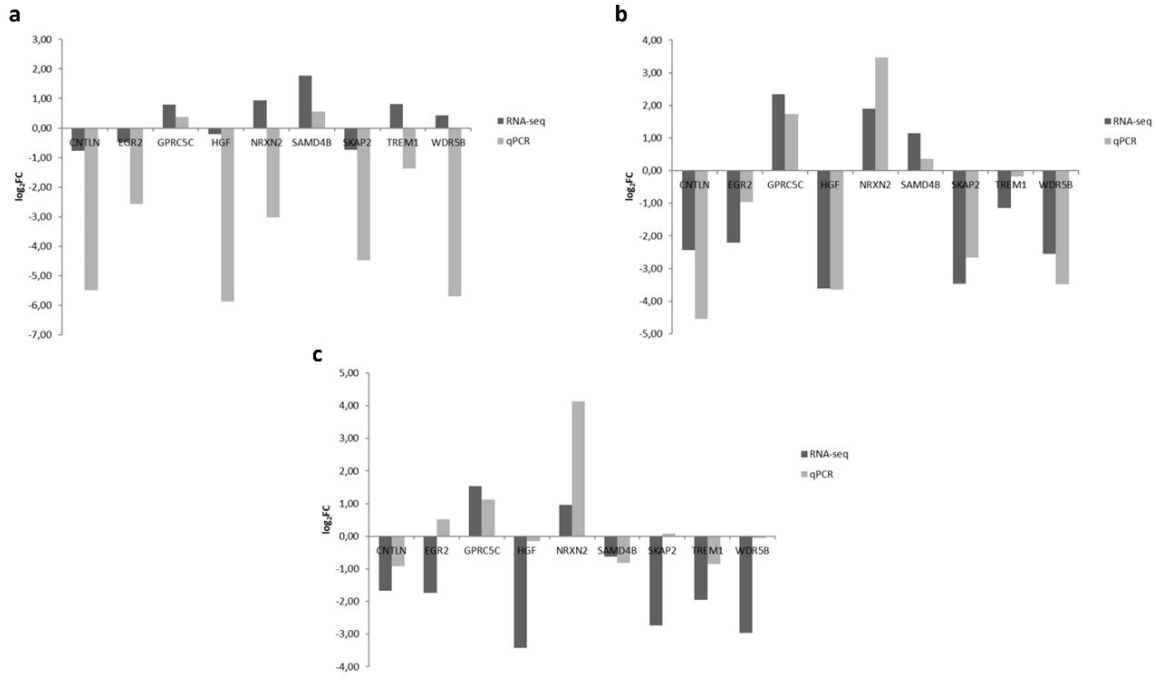


Figure 3.23: Validation of the RNA-seq data for the selected mRNAs by RT-qPCR. Log_2FC are shown for a) Vac Tf vs. Vac T0; b) Adj Tf vs. Adj T0; and c) Adj Tf vs. Vac Tf.

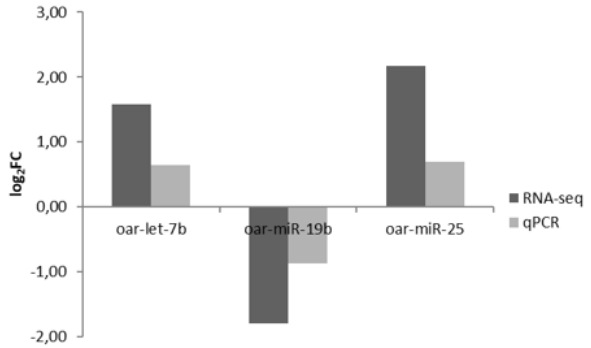


Figure 3.24: Validation of the miRNA-seq data for the selected miRNAs by RT-qPCR. Log2FC are shown in the Adj Tf vs. Adj T0 comparison.

3.4 Discussion

Aluminium based adjuvants, especially those containing aluminium hydroxide, are the most widely used compounds in human and veterinary vaccines (361,362). Despite Al mineral salts have been used over a century, their mechanism of action is only partially understood. In an attempt to decipher its mechanism of action, multiple sheep were subjected to a vaccination schedule with commercial vaccines or with an equivalent quantity of aluminium hydroxide alone through 475 days. PBMCs samples were extracted at the start (day 0, before any inoculation) and at the end (day 475) of the experiment.

The main objective of this study was to describe the RNA molecular patterns elicited upon vaccination mimicking a field practical situation, independently of the role of specific vaccine antigens/inactivated pathogens. The choice of commercial vaccines as the main treatment group was based in the idea that they are commercial products under strict evaluation, ensuring that immune stimulation has been proven. The rationale for the choice of a long-term response experiment was based on the field situation represented by the presence of unexplained adverse effects after an intensive vaccination campaign. This is the first long-term *in vivo* study dealing with the molecular genetic basis of the immune response in sheep after repetitive inoculation with aluminium containing components. As already noted, and it should be highlighted, the experiment is *in vivo*, since most studies dealing with the immune response induced by Al adjuvants have been realized *in vitro*, on immune system cells. There have been contradictory results (mainly regarding the requirement of the NLRP3 inflammasome in Al adjuvancy), which in some cases may be explained with cell behavioural patterns and interactions with their environment that cannot be fully grasped in *in vitro* designs and differences in structure of adjuvants and absorbed antigens in commercial formulations used in different studies. In addition, Al adjuvants have a low dissolution rate and there have been multiple reports showing that low doses of Al remain in the organism after long periods of time (48). Taking all into consideration, there is a need for more long-term *in vivo* studies to address the discussed elements (47), as every year farm animals are exposed to multiple aluminium-based vaccines. The selected study design is a fairly schematic approach for documenting aluminium-based adjuvants on whole animals in a real-life setting.

There are some limitations inherent to the selected study design that must be taken into account when interpreting the results. First of all, the main profiles being compared are t=0 versus t=475, and as a consequence, it is hard to determine if certain changes have been elicited by the overall administration schedule or by the last dose prior sample collection. Something that could be corrected with samples from middle time points, but due to funding limitations, only two time points were performed. This experiment was planned to study the pathology caused by administration of Al adjuvant-containing products, independently from the

identification of individual inoculations, the granuloma exact age or the role of specific vaccine antigens. We expect to see the cumulative effect of all inoculations, without ruling out that the latter has a greater effect on the response of the animals than the previous ones. In addition, animals that had been under treatment were analysed only, so that the final result of the repetitive experiment was appreciated with the initial situation in each animal. It must be pointed that the immune system is completely developed in 5-month-old lambs (365) and differences related to development should not account for the differences observed in this study. Moreover, the vaccines used in sheep are the same regardless of age, with the same dose quantity and the same administration protocols. Ideally, the process would be better seen in adult sheep, but it is quite difficult to find adult animals without any prior vaccine for conducting the experiment in an acceptable time. Priority has been given to the homogeneity of the animals analysed in the different groups, using young animals from the same herd, without any prior vaccination before the experiment and with a period of adaptation to the new environment under the best conditions of feeding and temperature. Finally, the number of animals analysed by RNA-seq is limited, but always over the minimum required for statistical inferences in this kind of data.

Among the up- and down-regulated genes, there were some previously described in other studies dealing with alterations in expression caused by aluminium, namely: *NLRP3*, *IL1B*, *IL8*, *TNF*, *NFKB2*, *RELA* and *RELB*. NLRP3 is a member of the NLR (nucleotide-binding oligomerization domain and leucine-rich repeat-containing receptor) family and is part of the molecular platform called inflammasome (366). As previously stated, there is controversy regarding the involvement of the inflammasome in aluminium-adjuvant induced immune response. It has been shown in multiple *in vitro* studies that aluminium stimulation of *IL1B* is dependent on the *NLRP3* inflammasome (35,36). In contrast, in *in vivo* experiments of aluminium adjuvant changes, multiple studies have found no involvement of the inflammasome in dendritic cell and lymphocyte activation (36–38), supporting an inflammasome-independent mechanism for aluminium adjuvant immune response. In addition, it is known that aluminium hydroxide induces a T_H2-skewed response, whereas it has been shown that hyper-activation of the inflammasome in absence of any particulate produces T_H17 and T_H1 responses (367). Other studies *in vivo* have shown that the inflammasome is not required for aluminium-induced increase of serum IgG1/IgE antibody production (368). In our study in PBMCs samples, NLRP3 was significantly downregulated in Adj-injected sheep, while maintained a constant expression in Vac-injected animals. Taken together, it seems that the inflammasome is not totally required to induce an immune response in this *in vivo* experiment.

Another set of differentially expressed genes that seems to have an important role in aluminium-induced response, especially when the antigen is present, are those related to proinflammatory cytokines. Several reports have shown that secretion of inflammatory cytokines is induced by Al (36,369–371). The consequent increase in inflammatory signals led to the activation of the NF-κB signaling pathway. In our study, there were multiple genes from the NF-κB family, such as *NFKB2*, *RELA* and *RELB*, which had a greater expression in Vac-injected and Adj-injected animals. As previously stated, activation of NF-κB regulates a great variety of genes from the innate and adaptative immune response, including the cytokines *TNF*, *IL1B* and *IL8*. It has been shown that Al hydroxide upregulates *TNF* in human monocytes (372), while IL1B mRNA has been shown to have an increased expression in bovine PBMCs treated with Al (75). Both molecules were significantly upregulated in Vac-injected animals. In contrast, other proinflammatory cytokines (mainly *TNF* and *IL16*) were found upregulated in Adj-injected animals, supporting a non-specific induction of proinflammatory responses when the adjuvant is injected alone without any pathogen.

Apart from differentially expressed genes related to the proinflammatory response, there were genes from other pathways previously linked to Al adjuvancy. There were multiple genes related to apoptosis upregulated in Vac- or Adj-injected sheep, among them: *TP53BP2*, *CSRNP1*, *TEAD*, *CDCA7* and *PPP1R15A*. This is in agreement with other studies showing apoptosis in human neuroblastoma cell lines (373) and pro-apoptotic gene expression in human brain cells (374) after aluminium stimulation. In addition, there were genes related to the immune response (*SKAP2*, *IGSF6*, *LST1*, *FGR*, *MAPK13*), inflammatory response (*S100A12*, *ADGRE3*, *TREM1*, *STEAP4*, *NR4A3*), cell growth (*HGF*, *CSF3R*) and cell-cell signaling (*AREG*) upregulated in Vac-injected sheep, but downregulated in Adj-injected sheep. In contrast, some genes related to DNA replication and repair (*FEN1*, *HIST2H4A*) and involved in RNA binding, synthesis and metabolism (*IGF2BP3*) were downregulated in Vac-injected animals, but upregulated in Adj-injected animals. In fact, there have been multiple reports of danger signals released from necrotic or damaged cells at the inoculation site. In the case of Al hydroxide, release of host DNA and uric acid has been reported, which rapidly degrades RNA and DNA (32).

Interestingly, two autoimmune processes appeared in the pathway analysis, but in different comparisons: rheumatoid arthritis in Vac-inoculated sheep and systemic lupus erythematosus in Adj-inoculated sheep. Therefore, it is possible that a previously described syndrome in sheep after inoculation of vaccines based on Al hydroxide (5) resembles these autoimmune diseases and that the autoimmune effect of the adjuvant alone differs slightly from that obtained in combination with the pathogens in vaccines.

A co-expression analysis was also performed for mRNAs and lncRNAs with WGCNA software, and 32 different modules were obtained. Interestingly, 7 modules correlated with the treatment independent of presence of pathogens, 5 exclusively related to the vaccination of commercial vaccines and 5 exclusively related to the Adjuvant group. Some of the modules were in concordance to what have been seen in the differential expression analysis: lavenderblush3 and coral1 modules were composed of genes crucial for the correct function of the immune system and that those genes were similarly co-expressed in both treatments, indicating that aluminium hydroxide regardless of the presence of any pathogen is capable of activating an immune response; lightpink4 module is composed of genes related to the cytokines that were downregulated in Adj-injected animals, indicating a milder immune response in the absence of pathogens.

These results would be in concordance with the histopathological analyses done in an independent study in the same animals (375). Briefly, all Vac-injected sheep and nearly all Adj-injected lambs presented injection site granulomas, mostly composed of activated macrophages. The presence of Al in macrophages was demonstrated by fluorescence microscopy with lumogallion staining and by electron microscopy. In addition, macrophage-driven translocation of Al to regional lymph nodes was seen. It must be pointed that higher numbers of granulomas at the injection site and higher percentage of Al translocation to lymph nodes was seen in Vac-inoculated animals. The lower persistence of granulomas in Adj-inoculated animals, which might indicate a quicker clearance of Al, might be caused by the milder immune reaction from the lack of pathogen seen in the differential expression analysis.

Regarding the miRNA differential expression analysis, there were some miRNAs previously described in other studies related to expression changes induced by Al, namely: *miR-19b* (downregulated in Adj-injected animals) and *miR-125b* (upregulated in Vac-injected animals). miR-125b has been identified as a reactive oxygen species (ROS)- and NF-κB up-regulated miRNA by aluminium-sulfate in human astroglial cells (376). In addition, this miRNA shows an expression pattern similar to those observed in Alzheimer's disease (377,378). In contrast, dysregulation of miR-19b has been related to multiple nervous diseases, including

Parkinson's disease (379,380), and its downregulation has been shown in PBMC patients with Alzheimer's disease (381). Taken together, miRNA pattern analysis links central nervous system damage pathways seen in nervous diseases with the intensive vaccination employed in this study. Moreover, it has been shown in other studies that miR-99a expression, which is upregulated in our Vac-injected animals, is promoted by NF-κB (382).

Within the mRNA-miRNA correlated pairs, there were factors related to cellular response to DNA damage stimulus (*STXBP4*, *RNF169*, *ZBTB4*, *NFATC2*), RNA binding (*WDR75*, *SART3*, *LRPPRC*, *SYNE1*, *RDX*, *XRN1*, *ZC3H8*,*SUB1*, *MBNL3*) and response to stimulus (*HSPA14*, *MAP3K2*, *CHEK1*, *MKNK1*, *ANTXR2*, *NBEAL1*, *NFATC2*). As previously stated, host DNA released from necrotic or damaged cells has been shown to act as a danger signal that modulates immunity via cytotoxic effects in Al hydroxide. It has been reported that Al adjuvants produces granulomatous inflammatory reactions and promotes local necrosis in muscle tissue (31) and in the peritoneum of mice (383), something completely concordant with the histopathological analyses of sheep from the vaccine group (375).

In addition, some of the predicted and correlated miRNA targets have been previously linked to the immune system. Activation of *NF-κB* has been shown to be a biphasic process, a transient phase regulated by IκB-α and a persistent phase regulated by IκB-β (384). *MAP3K2* (*MEKK2*), which is a predicted target of *let-7b* (upregulated in Adj-injected sheep), is a kinase that controls the persistent activation of *NF-κB* in response to stimulation with proinflammatory cytokines through the formation of the MAP3K2:IκB-β:NF-κB:IKK complex (385). This kinase has been shown to directly phosphorylate and activate IκB kinases. *SNX27*, which is a predicted target of *miR-125b* (upregulated in Vac-injected sheep), has been shown to cause NF-κB hyperactivation after its silencing in human Jurkat T cells (386). This would be in concordance to what have been seen in animals vaccinated with commercial vaccines, in which the *SNX27* mRNA is downregulated and there is a general upregulation of *NF-κB* induced genes. Another interesting target gene is *CHEK1*, predicted to be targeted by *miR-16b* (upregulated in Adj-injected animals). This gene has been shown to be involved in DNA damage response. In accordance with our study, Farasani et al. (387) showed reduced levels of this gene after exposition of aluminium chloride or aluminium chlorohydrate in MCF10A-immortalized non-transformed human breast epithelial cells. This suggest that aluminium is not only able to damage DNA, but it can also compromise DNA repair systems.

In summary, it has been shown for the first time in a sheep model that Al hydroxide is able to produce an immune response independent of the presence of pathogens, with a significant increase in expression of proinflammatory cytokines, NF-κB regulated genes and apoptotic genes. The absence of inactivated pathogens results in a milder immune response, as seen in Adj-injected animals and the general downregulation of genes associated with the *cytokine-cytokine receptor interaction pathway*. In addition, different mechanisms for the regulation of the NF-κB pathway through miRNAs (especially miR-125b and let-7b) has been proposed in the aluminium adjuvant activity. Furthermore, it has been shown that aluminium affects multiple genes related to DNA repair and cellular response to DNA damage stimulus, some of them by a probable miRNA-mediated regulation (e.g., *miR-16b* and its predicted target *CHEK1*). Moreover, at least in our samples, the inflammasome does not seem necessary for Al adjuvant to induce an immune response. Finally, *miR-25*, *miR-16b* and *let-7b* differential expression has been associated for the first time with Al adjuvancy.

3.5 Appendix

Table S3.1: List of selected genes and the corresponding primer sequences for the validation of the total RNA-seq experiment in PBMCs.

Gene	GenBank ID	Primer Code	Location*[1]	Exon Junction	Sequence (5'-3')
Target Genes					
CNTLN	XM_015093231.1	CNTLN-F	2541-2563	yes	CACTGTTCTCAATCACTCCATC
		CNTLN-R	2616-2638		TTCAGAATCACTGCTTTCACTC
EGR2	XM_004021395.3	EGR2-F	175-192	yes	CACGTCGGTGACCATTT
		EGR2-R	242-261		TGTTGATCATGCCATCTCC
GPRC5C	XM_015098975.1	GPRC5C-F	2328-2345	yes	AGTGCCAACTCCACCCT
		GPRC5C-R	2396-2414		GGGACTGAGCCTTCCTTG
HGF	XM_012176562.2	HGF-F	1052-1072	yes	TCAAGTGCAAGGACCTAAGA
		HGF-R	1124-1145		CAACTCGGATGTTTGGATCAG
NRXN2	XM_015103168.1	NRXN2-F	986-1007	yes	GCATTATCTGGTGACCATCTC
		NRXN2-R	1047-1067		GAGCCCAGCATAGTGTAATC
SAMD4B	XM_015100576.1	SAMD4B-F	7164-7183	yes	CAGCCCTCTTCTCACAGAT
		SAMD4B-R	7218-7239		TGACATTCTGAGACTCCAAGT
SKAP2	XM_015095341.1	SKAP2-F	992-1012	yes	ACCACACCACAGGAGATAAA
		SKAP2-R	1060-1082		ATGACAGTTCATCAGAAAGAGC
TREM1	XM_012100643.2	TREM1-F	55-75	yes	CTCTTCGTTCCAGCCAGAAG
		TREM1-R	111-130		CCTCTGTGATTGCCAGTGT
WDR5B	XM_004002975.3	WDR5B-F	1066-1086	no	GCTCATTCTGACCCAGTTTC
		WDR5B-R	1132-1154		AGATTCGACAGACACCATCATA
Reference Genes					
GAPDH	NM_001190390.1	GAPDH-F*[2]	-	yes	GGCGTGAACCACGAGAAGTATAA
		GAPDH-R*[2]	-		CCCTCCACGATGCCAAAGT
ATP1A1	NM_ 001009360	ATP1A1-F*[2]	-	yes	GACTTGAACCGAGGCTTAACAAC
		ATP1A1-R*[2]	-		TCTGGCTAGGATCTCAGCAGC
ACTB	NM_001009784.2	ACTB-F	453-474	yes	ATGTTTGAGACCTTCAACACC
		ACTB-R	531-548		TCCATCACGATGCCAGT
TFRC	XM_004003001.2	TFRC-F	1888-1908	yes	GAGCTGGACCTGAACTATGA
		TFRC-R	1962-1984		CAGACCCATATCCCTTATGTCT

*[1] Corresponding Start-End coordinates from NCBI gene annotation.
*[2] Primers from (388). All other primers have been newly designed.

Table S3.2: List of selected miRNAs and the corresponding primer sequences for the validation of the miRNA-seq experiment in PBMCs.

miRNA	Assay product number (Qiagen)	Sequence (5'-3')
Target miRNA		
oar-let-7b	YP02115207	UGAGGUAGUAGGUUGUGUGGU
oar-miR-19b	YP00204450	UGUGCAAAUCCAUGCAAAACUGA
oar-miR-25	YP02110541	AUUGCACUUGUCUCGGUCUGA
Reference miRNA		
oar-miR-30d	YP02110767	UGUAAACAUCCCCGACUGG
oar-miR-191	YP00205972	CAACGGAAUCCCAAAAGCAGCU
U6 snRNA	YP00203907	–

Chapter 4

Response to Aluminium in brain

4.1 Introduction

Aluminium salts, and especially aluminium oxyhydroxide based Alhydrogel, are the most predominant adjuvants in human and veterinary vaccines. Their safety record in human vaccinations, their effectiveness to enhance antibody responses and the fact that these adjuvants are well tolerated and do not cause pyrexia makes them the first choice when developing new vaccines against pathogens and T_H2 responses are desired (aluminium salts are reported to rarely induce cellular immune responses) (389). Furthermore, any new adjuvant formulation is usually compared against Al hydroxide. However, their exact mechanism of action is under-studied and its importance has been under-appreciated for a long time (6).

Despite their safety record, some minor side effects have been reported for Al hydroxide: formation of granulomas when the route of administration is subcutaneous or intradermal rather than intramuscular (389); injection site pain and tenderness, which may be a reflect of cell necrosis; post-immunization headache, arthralgia and myalgia; and increased risk of allergy and anaphylaxis due to the T_H2 biased response and increased eosinophil and immunoglobulin E (IgE) production (390). In addition to those accepted rare adverse effects, there have recently been a large number of studies linking aluminium adjuvants and autoimmune reactions or transport of the material to the brain. In a research study with New Zealand White rabbits injected intramuscularly with labelled AH, it was determined by accelerator mass spectrometry that Al was detectable in blood one hour after vaccination and that the body was able to partially excrete through urine the Al absorbed from the adjuvant, but only a 6% of the AH adjuvant dose was eliminated after 28 days post vaccination (46). Aluminium is a non-essential element for the human body and is thought to serve no essential biological function, so the fact that the body is not able to excrete all injected Al in a short period of time, and that Al by itself is an element known to have cytotoxic properties (391–393) (although actual translocation of Al adjuvants to tissues such as brain remains to be probed), has raised some concerns regarding their safety in predisposed individuals in part of the research community.

It is known that AH induces a strong innate immune response, and this results in infiltrating immune-responsive phagocytic cell types harvesting the complex of adjuvant and pathogens. AH is taken up by immature dendritic cells at the site of injection and transported via the afferent lymph vessels to the lymph node for presentation to T cells (394). The capability of APCs to endocytose significant amounts of Al without suffering toxicity has taken the interest of multiple investigators, since these cells may act as vehicles for the trafficking of Al adjuvants though the body (with transport to lymph nodes probed) (395). Multiple studies have focused their attention in the relationship between Al and some neurological and autoimmune diseases, but direct causality remains to be demonstrated. In addition, prior to any such study, Al adjuvant transport to the brain must be demonstrated in an *in vivo* experiment, discarding Al up take from other routes (e.g., drinking water or food).

The potential effect of this kind of compound on the nervous system has been tested mainly in animal models such as mouse. In CD1 mice, with a subcutaneously injected dose of 100 µg/Kg of Alhydrogel adjuvant, cognitive alterations associated with death of motor neurons and an enormous increase of reactive astrocytic cells in an inflammatory response was reported (396). Moreover, with a dose of 300 µg/Kg, microglial and astroglial reactions were detected in the spinal cord of the same mice type, in addition to altered motor and cognitive functions (397). In other immunization experiment with fluorescently labelled oxyhydroxide particles in mice, an average of 15 solid Al particles were detected in mice brain at 21 days after immunization. In vitro studies performed in parallel confirmed the toxicity of Al adjuvant to neuronal cell cultures (398).

In an attempt to check if Al adjuvant is able to translocate to brain and to induce molecular changes, a long-term in vivo experiment in a large mammal was carried out. In this study, lambs received a parallel subcutaneous treatment with commercial vaccines containing aluminium hydroxide, an equivalent dose of only this compound diluted in PBS or PBS only. Parietal lobe cortex samples were taken from each animal at the end of the experiment. Then, total RNA and miRNA libraries were prepared and sequenced. Three expression comparisons were made: vaccinated animals against control samples, adjuvanted animals against control samples, and animals of both treatments at the end of the experiment against each other.

Very few studies have analysed the Al effects in the nervous system by RNA-seq technology. In a recent work, Xu et al. (53) identified by RNA-seq in hippocampus samples of Al treated rats 96 up-regulated and 652 downregulated mRNAs, in addition to 37 dysregulated lncRNAs. Among the most significant GO terms of dysregulated genes, there were terms related to glial cell differentiation, neural transmission and vesicle trafficking. Moreover, the results of this study pointed toward glial cell related genes having a relevant effect in the mechanism associated to Al neurotoxicity.

In parallel to the RNA-seq analyses in brain samples carried out in this chapter, histopathological analyses from the same animals in lumbar spinal cord and parietal lobe samples were done (399). Briefly, those brain samples were analysed by transversely heated graphite furnace atomic absorption spectrometry and lumogallion staining for Al quantity measurement and Al localization, respectively. Sheep showed significantly higher Al content in lumbar spinal cord samples of both treatment groups (Vac- and Adj-injected samples), while in the parietal lobe samples there were no differences, only a tendency to higher Al content ($p=0.074$) in Adj-injected sheep. Al deposits were localized by lumogallion staining mainly in the gray matter of both tissues, with strikingly more abundant deposits in the lumbar spinal cord. The Al content and deposits were always more abundant in Adj-injected samples in comparison to Vac-injected samples. This could be related to the lower persistence of granulomas at the injection site in Adj-inoculated sheep (375). Al aggregates are notably larger when pathogens are adsorbed (400), and the smaller size of Al aggregates in Adj-injected sheep may explain the earlier mobilization and systemic distribution of the material to other tissues. It must be pointed that few animals of the adjuvant group showed high Al content, >2 µg/g, in the lumbar spinal cord, a value considered potentially pathologic in humans (401), while levels >3 µg/g are considered pathologically significant (402). Generally, Al content in parietal lobe was lower than 1 µg/g. Lambs in this experiment underwent an accelerated vaccination schedule to mimic, in an acceptable time frame, the Al load that these animals receive during their productive life. Thus, the outcome might differ if the same amount of Al was given in a longer period of time.

The main aim of this sequencing study was to characterize the molecular changes in brain after a repetitive vaccination schedule in an in vivo experiment in sheep, in the same group of animals as our previous work (359). The hypothesis of this work was that a prolonged

exposure to vaccine Al adjuvants, always following manufacture's recommendations, would result in transport of the material to brain in predisposed individuals. For that purpose, two different sequencing libraries per sample were constructed, namely total RNA-seq and miRNA-seq. Thus, the objectives of this work were:

1. to identify genes and regulatory elements altered by Al from a repetitive inoculation experiment and to check the state of some pathways previously related to Al neurotoxicity by others in literature.

2. to predict potential targets of miRNAs that can be related to Al neurotoxicity and test for correlation between miRNA and predicted mRNA target expression data.

4.2 Material and methods

In this section only a brief description of animal samples, extraction method and validation by qPCR will be provided. The analysis workflow for total RNA-seq and miRNA-seq data has been described in full detail in their corresponding section in Chapter 2 – Material and Methods.

4.2.1 Animals

All experimental procedures were approved and licensed by the ethical committee of the University of Zaragoza (ref: PI15/16). Methods were carried out under the following guidelines: Spanish Policy for Animal Protection (RED53/2013) and the European Union Directive 2010/63 on protection of experimental animals.

The animals studied in this work were previously analysed for a different tissue (PBMCs) (359). A detailed description of the experimental design can be found in the corresponding section in chapter 3 (3.2.1 Animals). For total RNA-seq and miRNA-seq analysis, parietal lobe cortex samples at the end of the experiment from 12 animals (4 sheep inoculated with vaccines, 4 sheep inoculated with the adjuvant alone and 4 sheep from the control group) were used for library preparation. There was one more control sample for the miRNA-seq library preparation. The remaining 9 animals, 3 of each treatment group, were used for validation of the sequencing data by qPCR. A list of the samples used in this experiment can be seen in table 4.1.

Table 4.1: Samples used for sequencing and RT-qPCR. Vaccine refers to the group vaccinated with commercial vaccines, adjuvant the one inoculated with aluminium hydroxide diluted in PBS and control the one inoculated with PBS. In addition, Tf refers to the end of the experiment.

Treatment	Time	Samples
Sequencing		
Vaccine	Tf	121-E, 122-E, 124-E, 126-E
Adjuvant	Tf	114-E, 115-E, 116-E, 117-E
Control	Tf	131-E, 133-E*[1], 135-E, 136-E, 137-E
RT-qPCR*[2]		
Vaccine	Tf	123-E, 125-E, 127-E
Adjuvant	Tf	111-E, 112-E, 113-E
Control	Tf	132-E, 133-E, 134-E

*[1]Only sequenced for miRNA-seq data.

*[2]Validation on total RNA-seq data only.

4.2.2 Tissue collection, RNA extraction and sequencing

A sample from encephalon (parietal lobe cortex) was aseptically taken from each animal and tissue sections were preserved in RNAlater solution (Ambion, Austin, TX, USA) at −80 ^{o}C. The experimental procedure to obtain RNA was similar to the one previously performed in the analysis of PBMCs (359). 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 ^{o}C. RNA quantity and purity was assessed with NanoDrop 1000 Spectrophotometer (Thermo Scientific Inc, Bremen, Germany). RNA integrity was assessed on an 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 average RIN value of 8.06 and a 260/280 ratio >1.7. A summary with the sample qualities can be seen in table 4.2.

Table 4.2: Quality summary of sequenced samples with their 260/280 and 260/230 absorbance ratios and RIN values. Vac, group vaccinated with commercial vaccines; Adj, group inoculated with aluminium hydroxide diluted in PBS; Control, group inoculated with PBS alone.

CNAG ID	Sample name	Group	260/280 Absorbance ratio	260/230 Absorbance ratio	RIN
AD1408	114-E	Adj	1.71	1.72	8.6
AD1409	115-E	Adj	1.98	1.92	8.6
AD1410	116-E	Adj	2.05	0.81	7.6
AD1411	117-E	Adj	2.07	1.82	7.5
AD1412	121-E	Vac	2.05	1.87	7.6
AD1413	122-E	Vac	2.06	2.07	6.2
AD1414	124-E	Vac	2.06	1.65	8.7
AD1415	126-E	Vac	2.04	2.13	7.7
AD1416	131-E	Control	2.11	2.2	8.3
AD1417	135-E	Control	2.1	2.2	8.3
AD1418	136-E	Control	2.04	2.01	8.3
AD1419	137-E	Control	2.06	2.07	8.0
AD1420	133-E*	Control	2.06	1.93	8.6

*Only sequenced for miRNA-seq data.

For total RNA-seq libraries, the TruSeq Stranded Total RNA kit with Ribo-Zero (Illumina, San Diego, CA, USA) was used, while the TruSeq Small RNA library prep kit (illumine) was used for miRNA-seq. 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). 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).

4.2.3 qPCR validation

To validate changes identified by RNA-seq experiments, the relative expression levels of 13 mRNAs that were selected based on significant changes seen in the RNA-seq analyses were verified by qPCR. The strategy followed was similar to the one previously done for PBMCs samples (359). Briefly, primers were designed using the PrimerQuest and OligoAnalyzer tools of Integrated DNA Technologies (IDT). Supplementary table S4.1 shows the list of amplified ovine genes and the corresponding primer sequences. Quantitative PCR amplifications were performed using PowerUp™ SYBR™ Green Master Mix (Applied Biosystem, Foster City, CA, USA) in a 10 µl final volume reaction on a QuantStudio® 3 detection system (Applied Biosystem). The conditions were as follows: 1 cycle of 50℃ for 2 min, 1 cycle of 95℃ for 2 min, 40 cycles of denaturation at 95℃ for 15 s, annealing at 60℃ for 60 s, and a dissociation curve to measure the specificity of the amplification. The stability of candidate endogenous control was analysed using GenEx software of MultiD [v5.4] (NormFinder (348) and GeNorm (349) algorithms). HPRT and ATP1A1 were the two most stable genes, so these two reference genes were used as an internal control to normalize the data. The expression level of mRNA transcripts was calculated using the $2^{-\Delta(\Delta Ct)}$ method. Statistical significance of the comparison between results obtained with RNA-seq and RT-qPCR was calculated by using t-test. In all analyses, differences were considered significant when p values were <0.05.

4.3 Results

4.3.1 Total RNA-seq

4.3.1.1 Sequencing quality

After sequencing 12 total RNA-seq libraries, an average depth of 74.1 million 75 nt paired-end reads per library were achieved. Then, after adaptor removal and quality filtering with Trimmomatic (chosen parameters can be seen in the corresponding section of Chapter 2 – Material and Methods), a mean of 68.8 (SD=6.95) million reads (92.80%) were considered as good quality segments for subsequent analyses. The average quality of the trimmed samples can be seen in figure 4.1. All samples had good quality after trimming.

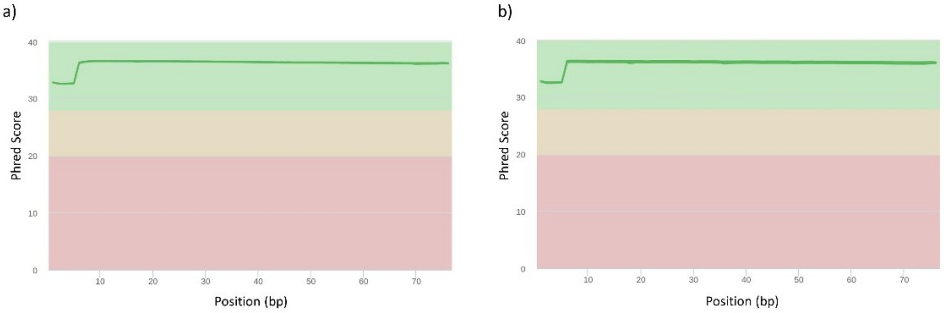


Figure 4.1: Average quality per base for all sequenced samples for paired-end data. a) pair 1 and b) pair 2. The y-axis shows the average quality in Phred scale, while the x-axis is a representation of all bases in a sequencing read of 75 nt for each pair. The background of the graph is divided by different colours in three main section, with the green section indicating good quality bases, the orange reasonable quality bases and the red one poor quality bases. The per sample quality plots were produced by FASTQC and then, they were aggregated by MultiQC.

4.3.1.2 Alignment to reference genome

Trimmed reads were aligned to the Ovis aries reference genome Oar_v3.1 from Ensembl with STAR. The following results were achieved: a mean value of 60.7 (SD=6.25) million read pairs (88.33%) mapping to a unique locus, 5.9 (SD=0.74) million read pairs (8.54%) mapping to multiple loci and 2.1 (SD=0.28) million read pairs (3.13%) not mapping to any loci in the genome. A more detailed summary of the alignment can be seen for each sample in table 4.3. Only uniquely mapped reads were used for subsequent analyses. A mean value of 32.97 million read pairs (54.28%) per sample were successfully assigned to sheep annotated genes. Similar to PBMC samples, different metrics were evaluated with the NOISeq R package from Bioconductor and RSeQC package. Among the checked features were a splice junction class pie chart (see figure 4.2), a "saturation plot" (see figure 4.3) and a "sensitivity plot" (see figure 4.4). Only a junction class pie chart from sample 115-E is shown, but the rest of the samples follow a similar distribution of junctions. A similar distribution to PBMC samples can be seen in encephalon, with ~47% of junctions known, but with a bit higher percentage of junctions (~40%) completely novel, pointing towards to a poorer brain sheep transcriptome annotation in comparison to the blood transcriptome.

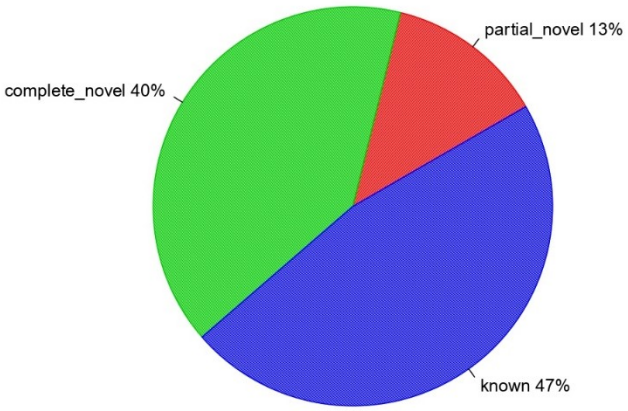


Figure 4.2: Detected splice junctions for sample 115-E. The junctions are divided as novel, partially novel (only one splice site is novel) and known (both splice sites are annotated in the reference genome).

In addition, as it could be seen in figure 4.3, there would be few gains with higher sequencing depths for our samples, at least for the annotated genes. In addition, in the "sensitivity plot" (figure 3.4) can be seen depicted for each sample the percentage of features with more than 0, 1 ,2,5 and 10 counts per million (CPM). This plot can be used to select a filtering criterion for lowly expressed features, which are less reliable and can introduce noise into the differential expression analysis.

It must be pointed that from 27,054 annotated genes in Ensembl, 21,487 (79.42%) were expressed with at least one sequence read count in one of the 12 RNA-seq libraries. Detected genes whose expression was lower than 1 CPM and could be found in less than 4 individual libraries were treated as lowly expressed genes and were removed from the differential expression analysis. These cut-offs were selected after checking that less stringent criteria introduced genes with high variability and expressed in only a few animals of each group. Those genes may not provide enough statistical evidence for reliable judgments and may confound the differential expression analysis if left in the data (350). After filtering lowly expressed genes, 14,387 (53.18%) remained for subsequent analyses.

Table 4.3: Summary statistics from the sequence alignment step for total RNA-seq data.

ID	Total Read-Pairs	Read-Pairs Surviving Trimming	Uniquely Mapped Read-Pairs	Read-Pairs Mapping to Multiple Loci	Unmapped Read-Pairs
114-E	79087585	73799900 (93.31%)	66105547 (89.57%)	5557132 (7.53%)	2140197 (2.90%)
115-E	70719509	65610966 (92.78%)	58456249 (89.10%)	5235755 (7.98%)	1915840 (2.92%)
116-E	68664277	63870962 (93.02%)	55715005 (87.23%)	5761161 (9.02%)	2395161 (3.75%)
117-E	65386555	60583932 (92.66%)	52304796 (86.33%)	6191678 (10.22%)	2084087 (3.44%)
121-E	80675987	74931076 (92.88%)	66535255 (88.80%)	6204293 (8.28%)	2187987 (2.92%)
122-E	79248637	74006266 (93.38%)	65046123 (87.89%)	6497750 (8.78%)	2457008 (3.32%)
124-E	66961909	62429123 (93.23%)	55610270 (89.08%)	5106702 (8.18%)	1716800 (2.75%)
126-E	81037553	75696789 (93.41%)	67277037 (88.88%)	5972477 (7.89%)	2452576 (3.24%)
131-E	83873049	77940602 (92.93%)	68206294 (87.51%)	7178329 (9.21%)	2556458 (3.28%)
135-E	74942848	70177373 (93.64%)	61568462 (87.73%)	6582637 (9.38%)	2028126 (2.89%)
136-E	75240320	70257641 (93.38%)	62286220 (88.65%)	5803281 (8.26%)	2156909 (3.07%)
137-E	63437982	55955752 (88.21%)	49842243 (89.07%)	4409313 (7.88%)	1706650 (3.05%)

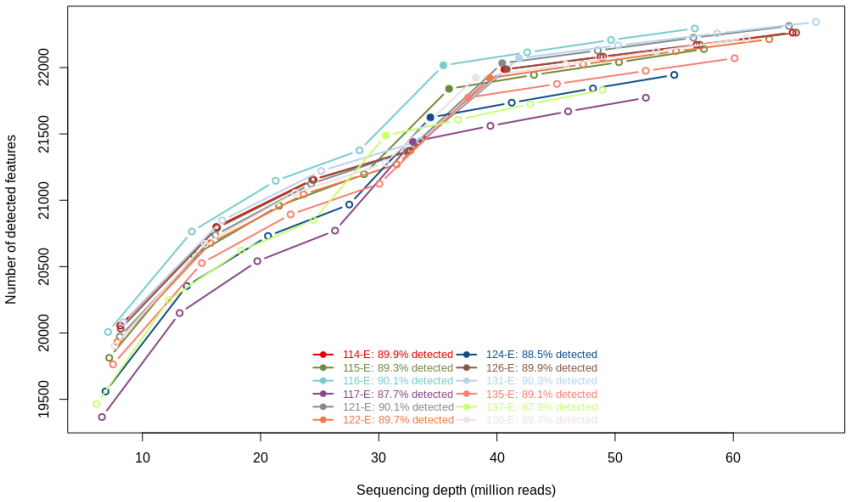


Figure 4.3: Saturation plot. The y-axis shows the number of detected features (genes) with more than 0 counts at different sequencing depths, x-axis, for each sample. Filled dots correspond to the values detected in our libraries, while the empty dots correspond to the simulated values.

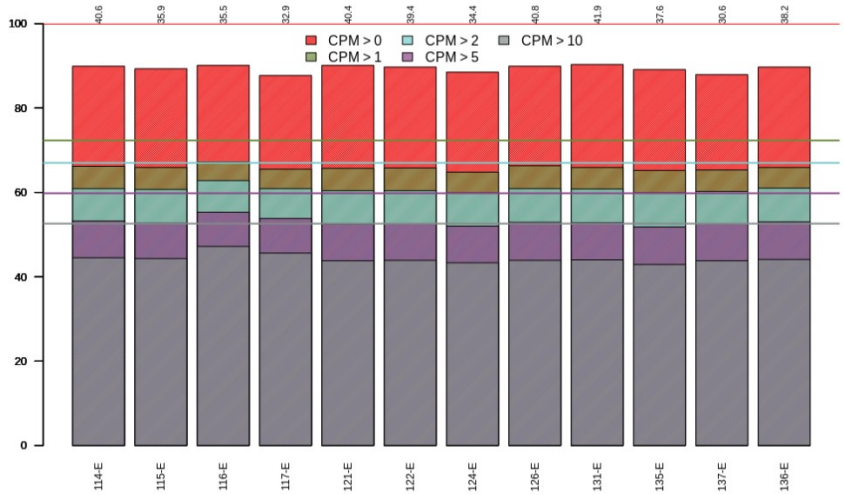


Figure 4.4: Sensitivity plot. The bars show the percentage of annotated features within each sample with more than 0, 1, 2, 5 and 10 CPM. In the upper side of the plot, the sequencing depth of each sample (in millions) is given. The horizontal lines are the corresponding percentage of features with those CPM in at least one of the samples.

4.3.1.3 Differential expression analysis

One sample from the adjuvant group (116-E) was treated as an outlier and was removed from the analysis. Despite having an adequate RIN value (7.6), it was observed a low 260/230 absorbance ratio (a secondary measure of nucleic acid purity) of 0.81 for that sample. Lower ratios of 1.8 may indicate the presence of co-purified contaminants. Prior to the differential expression analysis, the svaseq function from the SVA package was applied to remove unwanted variation and accurately measure the biological variability. The obtained surrogate variables were incorporated into the testing model of the DE analysis. A principal component analysis (PCA) was done with the corrected data (see figure 4.5).

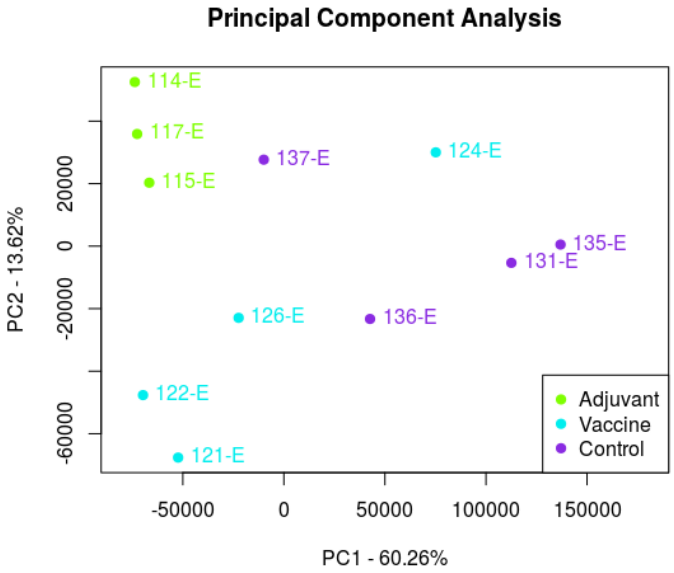


Figure 4.5: Principal Component Analysis (PCA) in total RNA-seq data from sheep encephalon after the batch effect removal with SVA R package.

DESeq2 was applied for the differential expression analysis with the design model previously described in Chapter 2. Those genes with an adjusted p-value<0.05 and a fold-change>1.5 or <0.667 were selected as significant. Three different comparisons were done, mainly: Vac vs. Control, Adj vs. Control and Adj vs. Vac. In the Adj vs. Control comparison 33 differentially expressed genes were found, of which 20 were up-regulated and 13 were down-regulated. In the Vac vs. Control comparison 6 DEGs were found, of which 2 and 4 were up-regulated and down-regulated, respectively. Finally, in the Adj vs. Control comparison, 45 DEGs were identified, including 33 and 12 with increased and decreased expression, respectively. The exact results by gene of the DE analysis would be available as a supplementary table in the corresponding article that is under review. In addition, the DEGs can be seen as a heatmap in figure 4.6.

Among the differentially expressed mRNAs, there are factor 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) (ATP13A5, HIST1H1C) and Amyotrophic Lateral Sclerosis (ALS) (ANXA2) (see figure 4.7).

4.3.1.4 Functional enrichment analysis

In addition to check the most significant differentially expressed genes, in an attempt to decipher the functions of DEGs, a functional enrichment analysis was performed with PANTHER and DAVID tools. For the three main domains of the GO database (cellular component, molecular function and biological process), the PANTHER webtool was used for each list of DEGs (three in total). 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

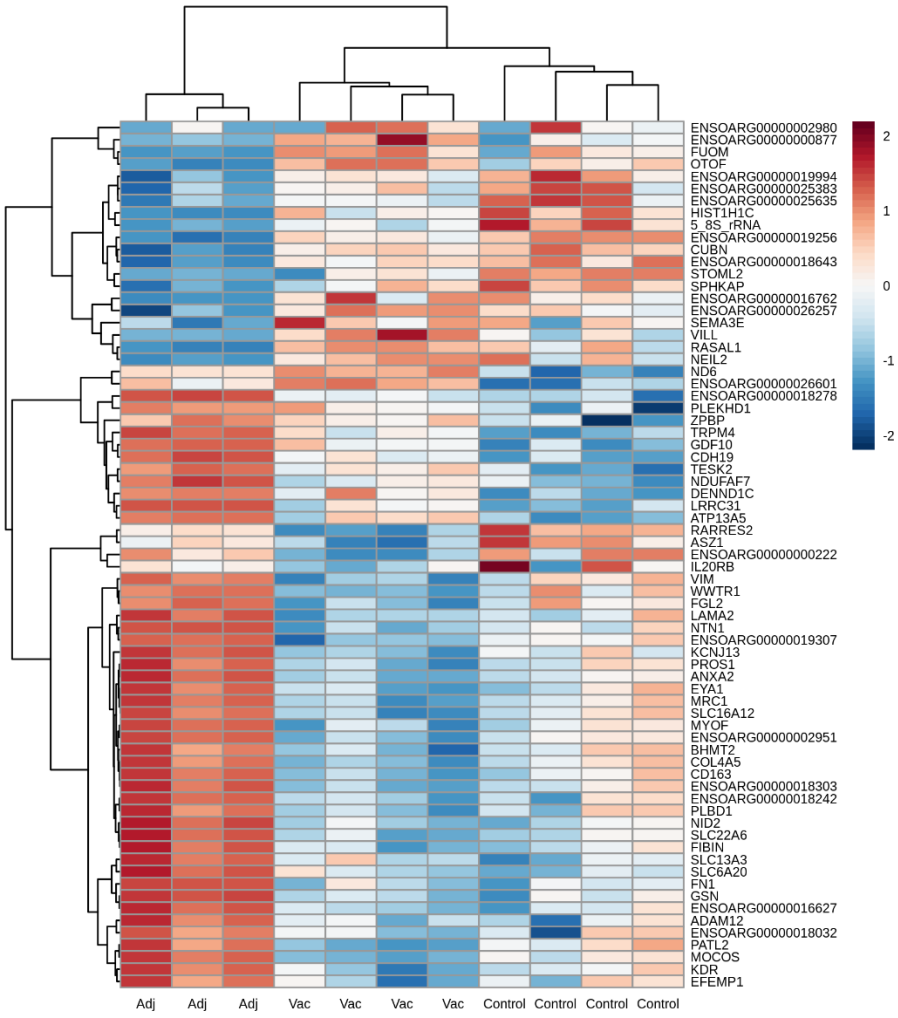


Figure 4.6: Heatmap constructed with the differentially expressed gene expression data after *SVA* package correction and DESeq2 variance stabilizing transformation (vst) normalization.

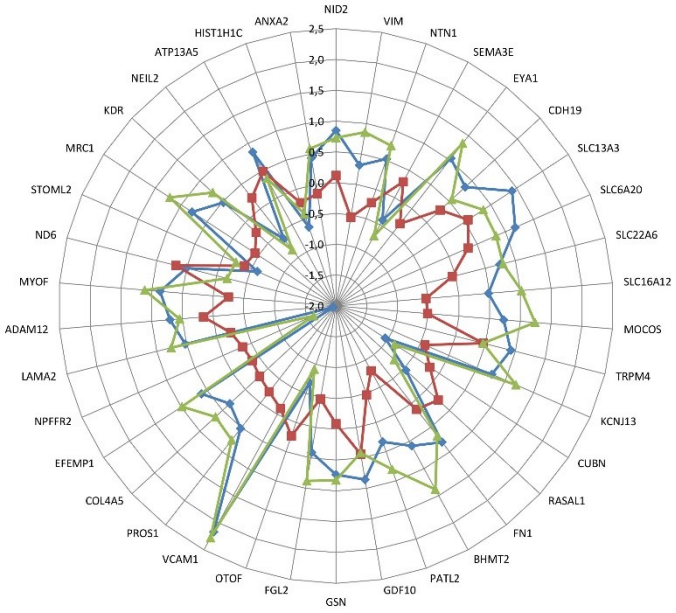


Figure 4.7: Radar plot with the log2(FC) for DEGs in each comparison.

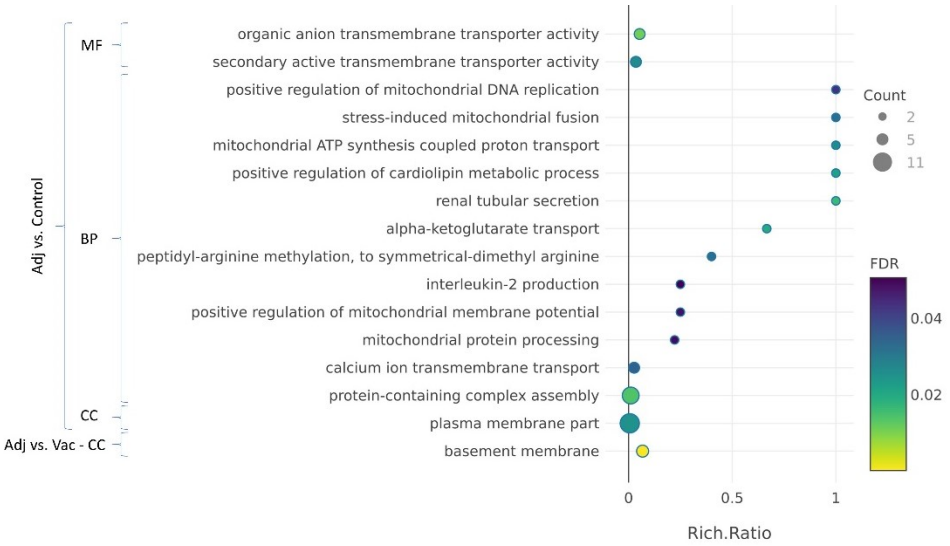


Figure 4.8: GO enrichment term analysis of differentially expressed genes in the Adj vs. Control and Adj vs. Vac comparisons. The bubble plot shows in the Y-axis the enriched GO terms, while in the X-axis the rich ratio is represented (rich ratio=amount of differentially expressed genes in the term/all 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.

arginine (GO:0019918), positive regulation of mitochondrial membrane potential (GO:0010918), mitochondrial protein processing (GO:0034982) and calcium ion transmembrane transport (GO:0070588) (Figure 4.8). Due to the few DEGs found in each comparison, there was no significant GO term or KEGG pathway in the remaining comparisons. Taken together, it is clear that terms related to mitochondria are being altered in the Adj-injected animals.

4.3.1.5 Weighted gene correlation network

A weighted gene co-expression network analysis was performed with the WGCNA [v1.63] (279,354) R package. It must be pointed that in addition to annotated gene data, expression data of detected new lncRNAs (data not shown) was used for network construction. Following a similar pipeline to the PBMC samples, the first step was to select an adequate parameter β based on the minimum value required to get a scale-free topology network, which correspond to a scale-free topology fit index R^2>0.8. As can be seen in figure 4.9, a parameter β=28 was enough to achieve a scale-free topology. After network construction, a total of 255 co-expressed gene modules were detected. Then, modules with similar expression profiles were merged and, in the end, a total of 46 co-expressed gene modules were detected (figure 4.10(a) and 4.10(b)), module size ranging from 37 to 2,424 genes. Each module was assigned a random colour name. Then, significant Pearson's correlations among module eigengenes (the first principal component of each module) and treatment variables (all possible dichotomized combinations, in which one group is against the other two. TreatControl, control samples against treated animals; TreatVac, samples from the vaccine group against the rest of samples; and TreatAdj, samples from the adjuvant group against the rest of samples) were searched. After selecting a q-value of 0.05 as cut-off, the following modules were found significant for each dichotomized treatment variable (Figure 4.10(c)): 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 (1325 genes, r=0.88, qvalue=0.01) and the

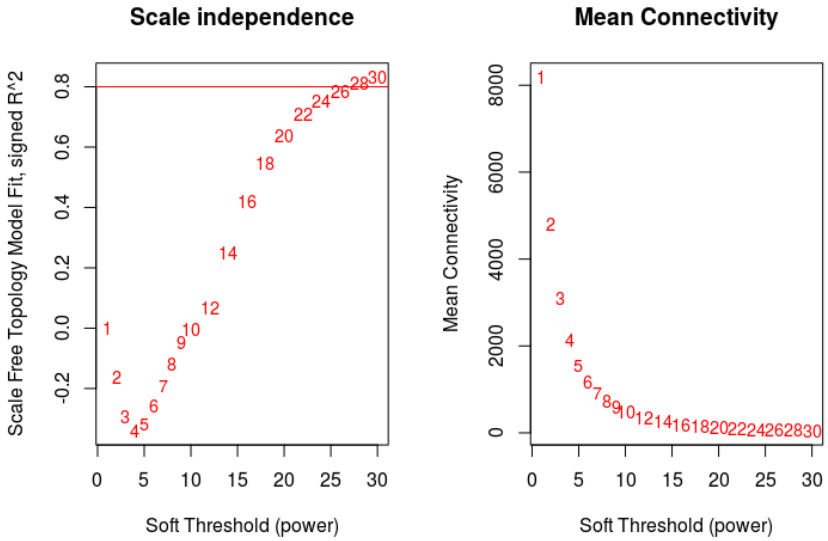


Figure 4.9: Summary network indices (y-axes) as functions of the soft thresholding power (x axes). Numbers in the plots indicate the corresponding soft thresholding powers. For each power, the scale-free topology fit index (R2) is calculated (left panel) and returned along the network mean connectivity (right panel).

burlywood1 module (228 genes, r=-0.83, qvalue=0.04) for Adj group (Figure 4(c)). There were no co-expressed modules associated with the Control group. In addition, it was checked how many DEGs were inside each significant module and it was shown that all those modules had from 3 to 36 DEGs. The most outstanding module was the maroon with 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. Among the modules correlated to Adj-inoculated sheep (TreatAdj), maroon showed positive correlation, which indicates that genes belonging to this module usually have higher expressions, while burlywood1 showed negative correlation, which indicates a lower expression. Among terms from the BP ontology, the maroon module was enriched in *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 response* (*GO:0050778*), *phagocytosis* (*GO:0006909*), *receptor-mediated endocytosis* (*GO:0006898*), *cellular response to oxygen-containing compound* (*GO:1901701*) and *positive regulation of reactive oxygen species metabolic process* (*GO:2000379*). A more detailed list of enriched terms from the Biological Process (BP) ontology for the maroon module can be seen in figure 4.11 (terms with more than 50 genes were removed for visualization purposes, being these terms way too general to be of interest). The remaining modules did not show significant enrichment for any GO ontology, except the burlywood1 module that was enriched with some terms from the Cellular Component (CC) ontology such as *histone methyltransferase complex* (*GO:0035097*) and *mitochondrion* (*GO:0005739*).

Regarding KEGG pathway enrichment, only the maroon module showed any significant results, mainly: *ECM-receptor interaction* (*oas04512*), *amoebiasis* (*oas05146*), *focal adhesion* (*oas04510*), *PI3K-Akt signaling pathway* (*oas04151*) and *protein digestion and absorption* (*oas04974*). In addition, there were two additional pathways with multiple genes that were nearly significant: *sulfur metabolism* (*oas00920*, adjusted p-value=5.05E-02) and *NF-kappa B signaling pathway* (*oas04064*, adjusted p-value=5.19E-02).

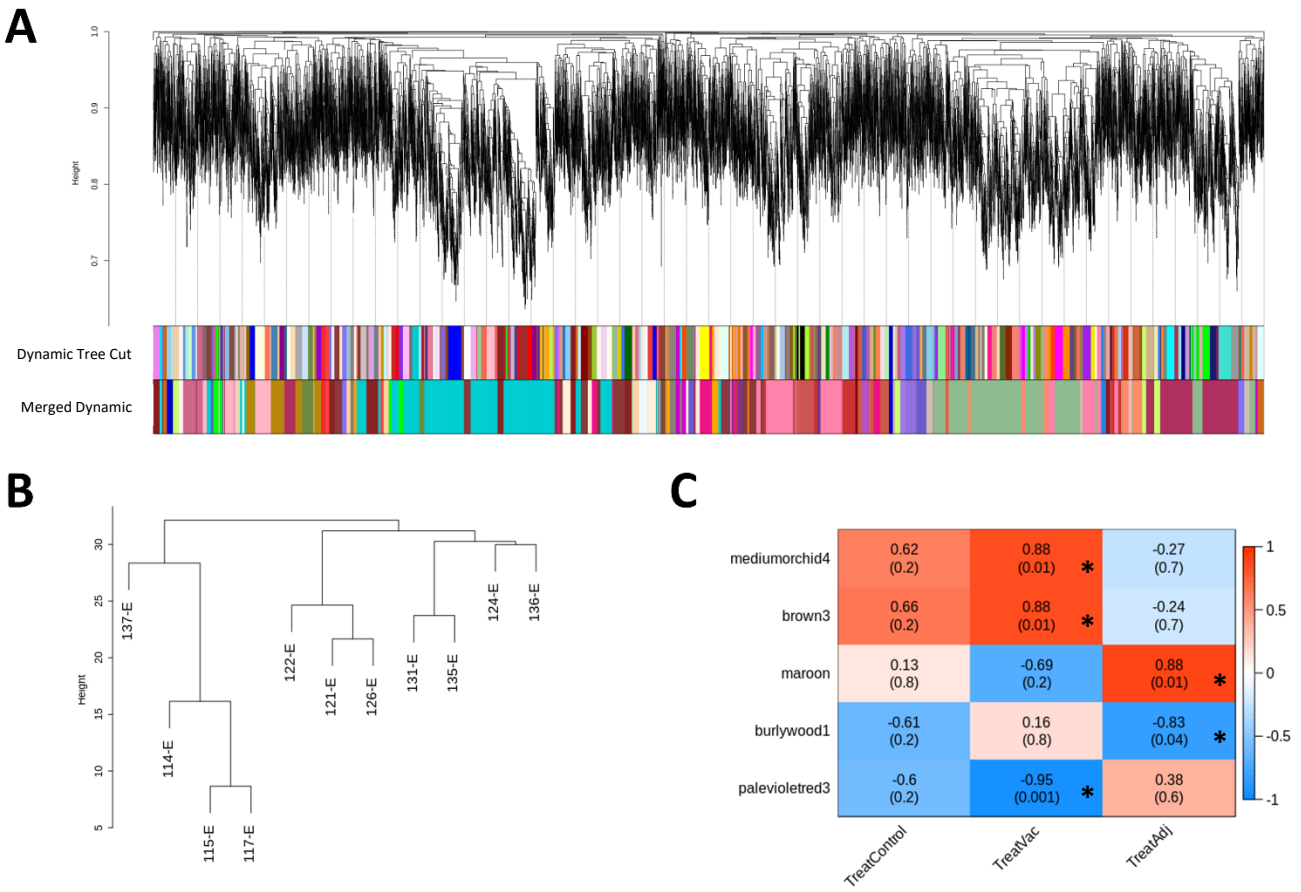


Figure 4.10: Weighted gene expression co-variance network analysis (WGCNA) summary. (a) Gene dendrogram obtained by average linkage hierarchical clustering. The colour rows underneath the dendrogram shows the module assignment before (Dynamic Tree Cut) and after (Merged Dynamic) modules with similar 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 (color-coded) and adjusted p-values. Only modules associated with at least one trait are shown (significant ones marked with and asterisk).

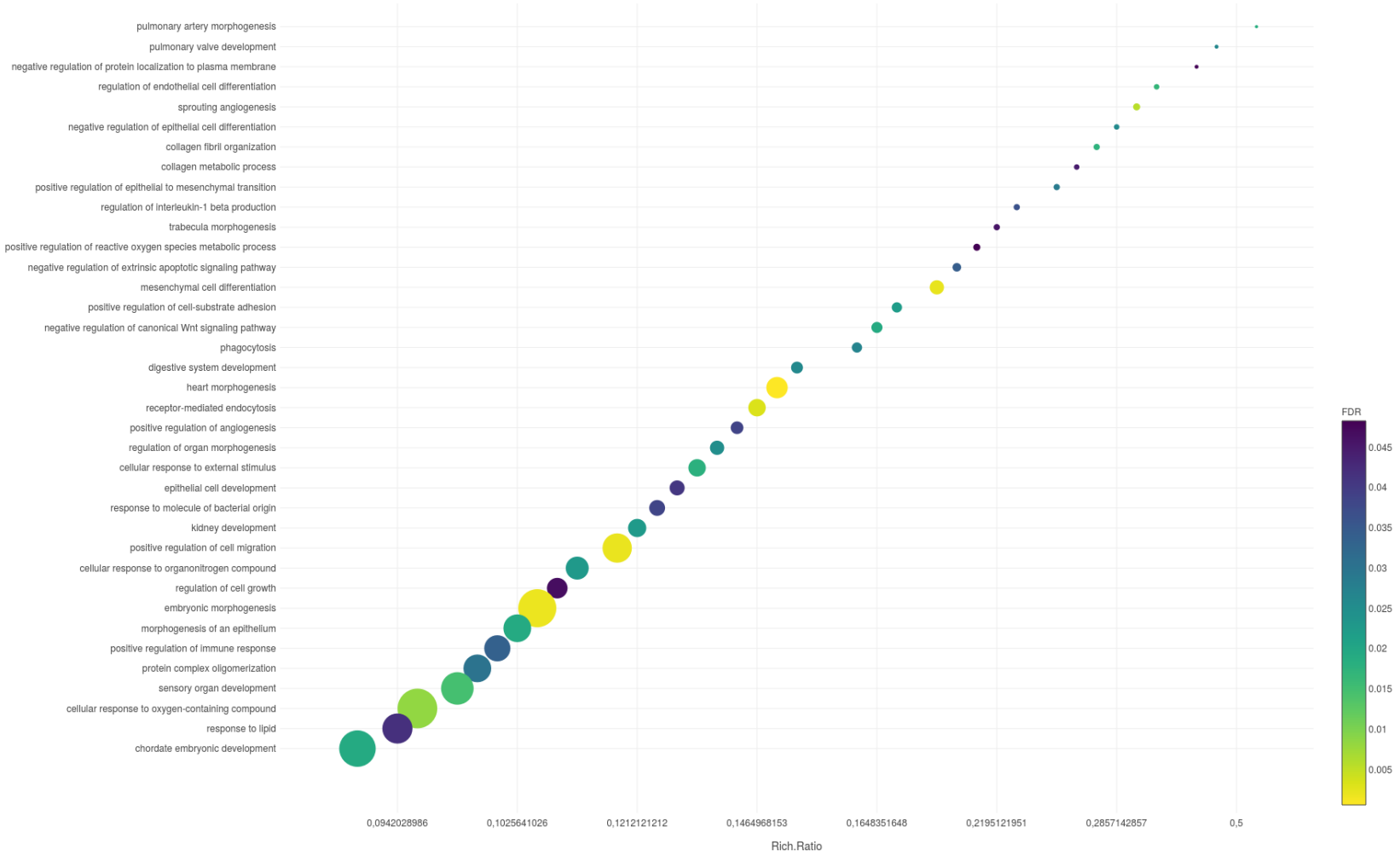


Figure 4.11: GO enrichment term analysis from the Biological Process category in the maroon module. The bubble plot shows in the Y-axis the enriched GO terms, while in the X-axis the rich ratio is represented (rich ratio=amount of differentially expressed genes in the term/all genes included in the term). Size and colour of the bubble represent the number of differentially expressed genes in the GO term and enrichment significance (FDR), respectively.

Finally, treatment-related hub genes, which are defined as those belonging to the ≥85th percentile for both MM and GS, were checked (Table 4.4). To note the maroon module, in which 17 of the hub genes (including a few lncRNAs that are not shown) are DEGs. Some of them, as previously detailed, had been related to brain injury (*GSN*, *LAMA2*, and *PROS1*), neuronal development (*NTN1* and *NID2*) and different diseases in brain (*MRC1* and *ANXA2*). In addition to DEGs, there were other hub genes related to insulin signaling (*INSR*, *IGFBP2* and *IGF2BP2*), blood brain barrier (*ADGRA2* and *NTN1*), ERK signaling (*INSR*, *ITGA9*, *OSMR*, *COL18A1*, *LAMA2*, *BCL2L11*, *ADAM17*, *COL4A3*, *COL4A4*, *COL4A6*, *COL2A1* and *BMP4*) and calcium signaling (*APOOL*, *HOMER3* and *TMBIM1*). It seems that the maroon module is composed of genes essential for the correct function of the brain.

Table 4.4: Hub genes in the modules related to any treatment.

Module	Hub Genes
mediumorchid4	SCML4, PROKR2, TTC25, HHLA2
brown3	KRBA1, MAST2, RET, RUNDC1, FLRT1, ABLIM3, SMG5, PODN, TBR1, MROH2B, GABBR2, SIRT4, THRA, TENM2, NKAP, MCTP1, SOWAHB, CCKBR, ENSOARG00000001628, ENSOARG00000013353, ENSOARG00000017276
palevioletred3	RARRES2, ATPAF2, AKR1A1, PDGFD, NQO1, RNF187, TYROBP, MTFR1L, IL6ST, NET1, TMEM186, ACAT1, GNE, NECAP2, SCARF1, GNG5, TCIRG1, CH25H, C2orf69, TTC32, HERPUD1, PLA1A, GNL2, ENSOARG00000010890
maroon	INSR, SLC22A6, ITGA9, ADGRA2, TFPI2, CAV1, SNX33, APOOL, MYOF, HPSE, NTN1, RASSF3, CSGALNACT1, ACSF2, KCNT1, TRIP10, STOX1, MOCOS, ZIC4, GSN, PLEKHH2, GHR, TIMELESS, COLEC12, DAB2, TNFRSF11B, OSMR, TTC23, HOMER3, MYOM1, SUCLG2, SPTLC3, SLC6A13, ISYNA1, MRC1, SLC4A5, COL18A1, C1orf115, NEXN, LAMA2, CROT, BCL2L11, SVIL, FIBIN, ADAM17, DSN1, PROS1, XPNPEP3, CILP, COL4A6, ALDH7A1, SNX24, COL2A1, KCNJ13, IGFBP2, TMBIM1, FSTL1, FAM43A, COL4A4, COL4A3, IGF2BP2, EHHADH, NID2, ANXA2, BMP4, L3HYPDH, ENSOARG00000002862, ENSOARG00000002951, ENSOARG00000003361, ENSOARG00000003373, ENSOARG00000008491, ENSOARG00000009919, ENSOARG00000016627, ENSOARG00000018242, ENSOARG00000018303
burlywood1	RGS17, ST6GALNAC5, DPCD, ENSOARG00000001664, ENSOARG00000002028, ENSOARG00000002172, ENSOARG00000005831, ENSOARG00000013502, ENSOARG00000014615, ENSOARG00000019256

4.3.2 miRNA-seq

4.3.2.1 Sequencing quality

The same 12 samples from Total RNA-seq, in addition to an extra sample in the control group (133-E), were prepared for miRNA sequencing. After sequencing, an average depth of 18.2 million 50 nt single-end reads per library were achieved. Adaptor sequence trimming and low quality read filtering were done with Trimmomatic following criteria previously described (Chapter 2 – section 2.3.3.4), which resulted in a mean 16.3 (SD=3.4) million reads (89.55%) of

good quality reads. The average quality of the trimmed samples can be seen in figure 4.12. As it can be seen, all samples have good quality after trimming.

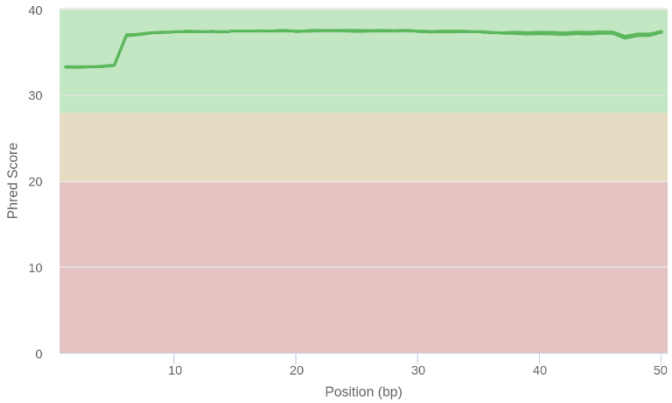


Figure 4.12: Average quality per base for all sequenced samples. The y-axis shows the average quality in Phred scale, while the x-axis is a representation of all bases in a sequencing read of 75 nt for each pair. The background of the graph is divided by different colours in three main section, with the green section indicating good quality bases, the orange reasonable quality bases and the red one poor quality bases. The per sample quality plots were produced by FASTQC and then, they were aggregated by MultiQC.

Trimmed reads were aligned to the Ovis aries reference genome Oar_v3.1 from Ensembl with the sRNAbench module from sRNAtoolbox, allowing up to 20 multiple mappings per read. An average of 13.1 (SD=3.1) million read pairs (80.21% of the filtered reads) were aligned to the reference per library. A more detailed summary of the alignment can be seen for each sample in table 4.5. It must be pointed that for small RNA annotation, all sequences were searched in miRbase to identify annotated miRNAs and in Rfam to identify other small RNAs originating from rRNA, tRNA, snRNA and snoRNA. As can be seen in figure 4.13, 42.91% of all successfully aligned reads were annotated miRNAs from the miRbase database, 3.01% were annotated sheep small nucleolar RNAs (snoRNAs), 12.96% to tRNAs, 24.71% to other RNAs from RNACentral and few reads were assigned to other small RNAs such as rRNAs and snRNAs. The detected miRbase miRNA expression values were taken for the expression matrix construction and the unassigned reads (15.60%) were used for novel miRNA prediction.

Table 4.5: Summary statistics from the sequence alignment step for miRNA-seq data.

ID	Total Reads	Reads Surviving Trimming	Mapped Reads
114-E	18593838	15731291 (84.60%)	12172383 (77.38%)
115-E	20260075	17405047 (85.91%)	13730047 (78.89%)
116-E	11367619	9594763 (84.40%)	6806110 (70.94%)
117-E	13522424	11829370 (87.48%)	8497618 (71.83%)
121-E	15423663	13756770 (89.19%)	11331974 (82.37%)
122-E	16590429	14580150 (87.88%)	12064292 (82.74%)
124-E	16239506	14851157 (91.45%)	11826273 (79.63%)
126-E	19208637	17588010 (91.56%)	14479206 (82.32%)
131-E	20848339	19413620 (93.12%)	16388368 (84.42%)
135-E	19982899	18452005 (92.34%)	15215629 (82.46%)
136-E	22059443	20165918 (91.42%)	16637673 (82.50%)
137-E	23603819	21750993 (92.15%)	17744010 (81.58%)
133-E	18996475	16838934 (88.64%)	13117106 (77.90%)

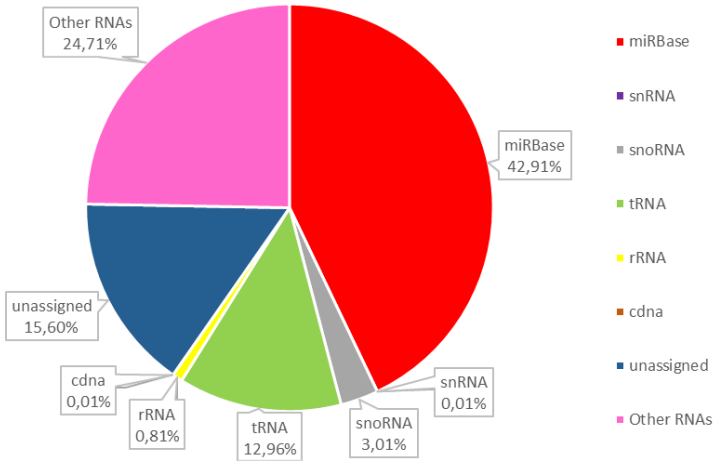


Figure 4.13: Class of molecules to which miRNA sequencing reads align.

After identification of miRNAs from miRbase and new miRNA prediction, 299 miRNAs were detected in encephalon samples with at least one sequence read in at least one of the 13 miRNA-seq libraries. From the detected 299 miRNAs, 141 were already annotated as *Ovis aries* miRNAs in miRbase, while others were previously annotated in other species (84 in *Capra hircus*, 20 *Bos taurus* and 44 in others). Ten were completely new miRNAs. A detailed list of all detected miRNAs will be available as supplementary material in the corresponding article that is under review.

4.3.2.2 Differential expression analysis

Those miRNAs with an expression lower than 1 CPM and detected in <4 individual libraries were treated as lowly expressed miRNAs and were filtered out for the differential expression analysis. In total, 259 miRNAs remained after filtering lowly expressed ones. The same sample, as in the total RNA-seq analysis, from the adjuvant group (116-E) was treated as an outlier and was removed from the analysis. Prior to any other analysis, and similar to the total RNA-seq data, the svaseq function from the SVA package was applied to remove unwanted variation and accurately measure the biological variability. The obtained surrogate variables were incorporated into the testing model of the DE analysis. A principal component analysis (PCA) was done with the corrected data (see figure 4.14).

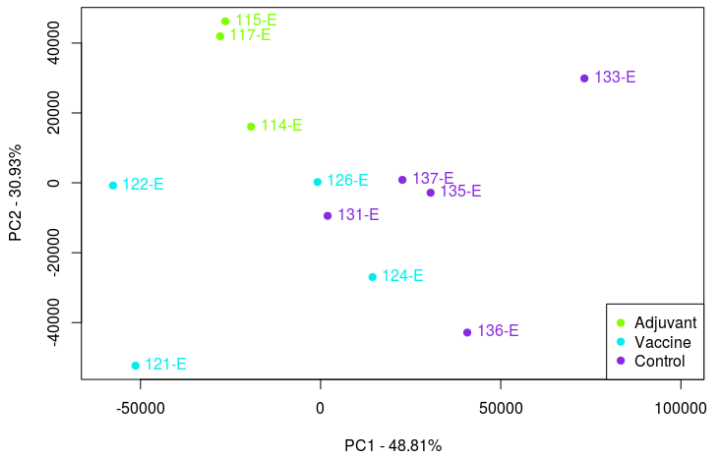


Figure 4.14: Principal Component Analysis (PCA) in miRNA-seq data from sheep PBMCs after the batch effect removal with SVA R package.

The differential expression analysis was performed with DESeq2 with the same testing model used for total RNA-seq data. Those miRNAs with an adjusted p-value<0.05 and a fold-change>1.5 or <0.667 were taken. Thus, a total of 38, 2 and 7 DE-miRNAs were identified in the Adj vs. Control, Vac vs. Control and Adj vs. Vac comparisons, respectively. The DEGs can be seen as a heatmap in figure 4.15. Within the DE-miRNAs there are factors that have been previously related to brain injury (*let-7b*, *miR-423-3p*, *miR-99b-3p*, *miR-874-3p*, *miR-29b/c*, *miR-328-3p*, *miR-99a*) and neurodegenerative diseases like AD (*miR-181c-3p*, *miR-29b/c*), PD (*miR-99b-3p*, *miR-29b/c*), ALS (*miR-181a*, *miR-30b*) and Multiple Sclerosis (MS) (*miR-369-5p*, *miR-370*, *let-7b/c*) or autoimmune diseases like lupus erythematosus (miR-410-3p).

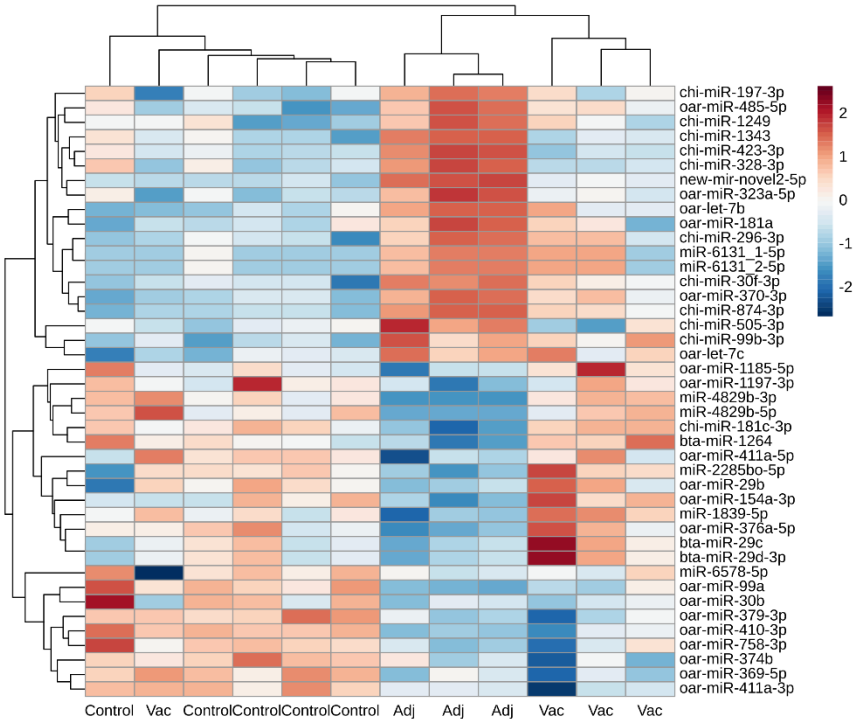


Figure 4.15: Heatmap constructed with the differentially expressed miRNA expression data after SVA package correction and DESeq2 variance stabilizing transformation (vst) normalization.

4.3.2.3 Target prediction and miRNA-mRNA data integration

Three different target gene prediction programs (miRanda, PITA and TargetScan) were selected and applied to the differentially expressed miRNAs, taking the intersection of all tools as potential target candidates. Two different approaches were applied to integrate miRNA and mRNA data: a correlation analysis and the iSubgraph algorithm. In table 4.6 can be seen the list of correlated miRNA-mRNA pairs after multiple-testing correction. The majority of the significant miRNA-mRNA pairs were positively correlated, something not expected if the miRNA acts via translational repression and/or mRNA cleavage. However, there is evidence of miRNAs enhancing translation in special conditions like cell cycle arrest (140) or mitochondrial translation (403). Positive correlation may denote the existence of feed-forward regulations mediated by transcription factors (404). Among the negatively regulated targets, there were some genes related to mitochondria (*ACTR10* and *MRS2*, both targeted by let-7 family members), to maintenance of neuronal polarity and axon growth (*RUFY3* targeted by *let-7b*) and to apoptosis (*NAA50* and *UNC5D* targeted by *miR-197-3p* and *miR-410-3p*, respectively).

Table 4.6: Significant miRNA-targets correlations after multiple testing correction. rho, Spearman's rank correlation coefficient; pvalue, significance level from the cor.test R function that test for association/correlation between paired samples; qvalue, false discovery estimation from the *qvalue* Bioconductor R package.

miRNA	Transcript	rho	pvalue	qvalue
let-7b	ACTR10	-0,84	1,33E-03	0,0484
let-7b	CEP135	0,93	3,97E+08	0,0163
let-7b	GGH	-0,86	6,12E-04	0,0377
let-7b	PALD1	0,85	8,07E-04	0,0415
let-7b	RUFY3	-0,90	1,60E-04	0,0247
let-7c	ACTR10	-0,85	8,07E-04	0,0415
let-7c	FBXL12	0,85	1,05E-03	0,0445
let-7c	MRS2	-0,84	1,33E-03	0,0484
let-7c	SLC20A1	-0,91	1,06E-04	0,0217
miR-181c-3p	ZDHHC2	0,89	2,33E-04	0,0262
miR-197-3p	GDF11	0,90	1,60E-04	0,0247
miR-197-3p	NAA50	-0,89	2,33E-04	0,0262
miR-197-3p	PM20D2	-0,89	2,33E-04	0,0262
miR-197-3p	RIN2	0,86	6,12E-04	0,0377
miR-29c	AP2B1	0,86	6,12E-04	0,0377
miR-29c	ASXL3	0,95	4,99E+08	0,0031
miR-29c	MARCH9	0,85	1,05E-03	0,0445
miR-29c	NAV3	0,85	8,07E-04	0,0415
miR-29d-3p	AP2B1	0,86	6,12E-04	0,0377
miR-29d-3p	ASXL3	0,95	4,99E+08	0,0031
miR-29d-3p	MARCH9	0,85	1,05E-03	0,0445
miR-29d-3p	NAV3	0,85	8,07E-04	0,0415
miR-30b	JOSD1	0,85	1,05E-03	0,0445
miR-30f-3p	FAM167A	0,84	1,33E-03	0,0484
miR-30f-3p	MRPS18C	-0,86	6,12E-04	0,0377
miR-30f-3p	SIGLEC1	0,84	1,33E-03	0,0484
miR-30f-3p	SIRT2	0,88	3,30E-04	0,0313
miR-323a-5p	NAT10	0,87	4,55E-04	0,0377
miR-379-3p	LAP3	-0,84	1,33E-03	0,0484
miR-379-3p	NDST1	-0,86	6,12E-04	0,0377
miR-410-3p	UNC5D	-0,88	3,30E-04	0,0313
miR-485-5p	MRPS27	-0,85	1,05E-03	0,0445
miR-485-5p	SEC61A1	0,91	1,06E-04	0,0217
miR-485-5p	SLC36A1	0,92	6,66E+09	0,0205

Finally, the iSubgraph algorithm was applied to encephalon samples, a tool designed to identify miRNA-gene patterns, even if they occur only in some groups/samples, by graph mining methods. After applying the algorithm, only a network was found in the three Adj-injected sheep (Figure 4.16). All the miRNA-gene pairs detected were positively correlated. From the network the miR-29 family and their targets stand out, some of them previously related to neurodegenerative diseases (*NAV3*, a member of the neuron navigator family, and *IREB2*, which encodes a protein that is a regulator of the cellular iron metabolism). Both genes have been previously reported to be affected at protein level by the miRNA while their mRNA level

remained stable in brain samples (405,406). The positive correlation of the miR-29c/NAV3 pair was also detected in the previous correlation analysis.

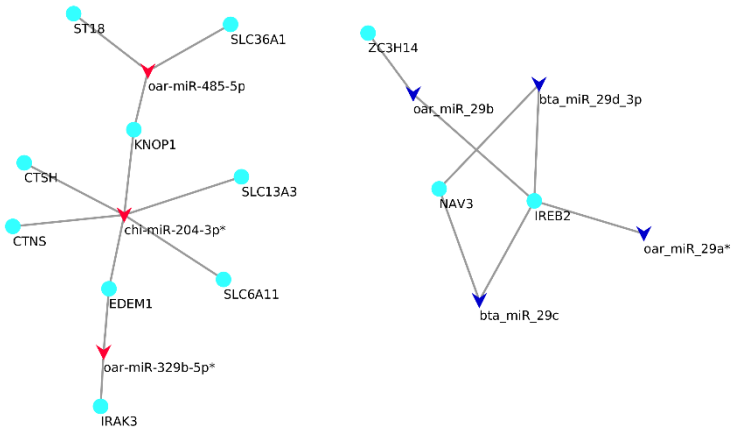


***Figure* 4.16:** Subgraph detected by iSubgraph algorithm in the adjuvant samples. In red and blue the miRNAs with higher and lower expression in the adjuvant samples in comparison to the control group, respectively. All the pairs are positively correlated. miRNAs marked with an asterisk were no DE.

4.3.3 Validation by RT-qPCR

For RNA-seq data validation, 13 mRNAs (*CUBN*, *GDF10*, *NDUFAF7*, *SLC13A3*, *SLC6A20*, *SPHKAP*, *ASZ1*, *ND6*, *RARRES2*, *EYA1*, *GNS*, *LAMA2* and *VIM*) were verified by RT-qPCR. Log2 fold changes (log2FC) in gene expression between the different groups calculated by RT-qPCR are shown in figure 4.17 for RNA-seq data. Data from RNA-seq and RT-qPCR showed a high degree of concordance (11/13, the gene expression patterns of most genes were reproducible) and there were no significant differences in fold change between RNA-seq and RT-qPCR (p-value>0.05), indicating that the sequencing results are reliable.

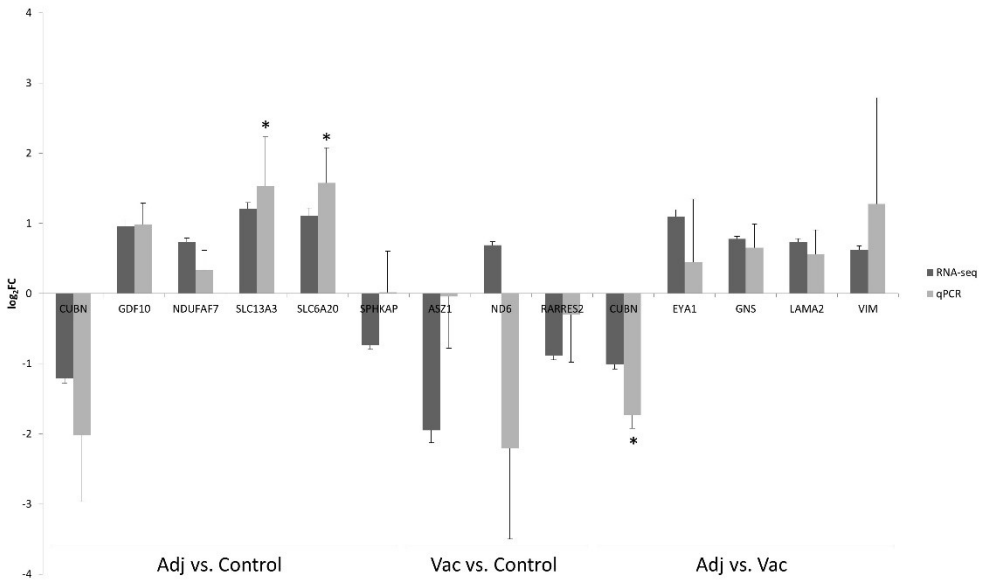


Figure 4.17: Validation of the RNA-seq data for the selected mRNAs by RT-qPCR. Log2FC are shown for each comparison in the bars, while error bars depict the standard error of the mean (SEM), which is the precision by which the mean value has been determined. An asterisk indicates statistically significant differences between both groups.

4.4 Discussion

Aluminium is a non-essential element and it has no physiological purpose in the human body (407). Nowadays, humans and animals are frequently exposed to multiple sources of Al, as it is being widely used in different fields, such as pharmaceuticals (e.g. in vaccines as adjuvants) and, to a lesser extent, in foods (e.g. in food additives or due to contaminants) and water (due to water treatment process or from weathering rocks and soils). In addition, Al is known to have cytotoxic properties and to be a biopersistent molecule, the body not being able to excrete all internalized Al in a short period of time. In a research study with New Zealand White rabbits injected intramuscularly with labelled AH, it was determined by accelerator mass spectrometry that Al was detectable in blood one hour after vaccination and that the body was able to partially eliminate through urine the Al absorbed from the adjuvant, but only a 6% of the AH adjuvant dose was eliminated after 28 days post vaccination (46). Taking all into account, in addition to multiple reports linking Al and some neurological diseases such as AD and PD (66) or some autoimmune diseases such as MMF and ASIA (5,52,77), some researchers have raised some concerns regarding the safety of Al-adjuvanted vaccines in pre-disposed individuals. The possibility of injected Al being able to reach distant organs is a topic of constant debate and further research *in vivo* is needed.

The main objective of this project was to describe the RNA molecular patterns elicited upon vaccination mimicking a field practical situation, independently of the role of specific vaccine antigens/inactivated pathogens. In addition, it was tested the ability of Al particles to reach a distant organ such as brain (399). For that, encephalon samples from already analysed sheep in a different tissue (359) were used. Briefly, sheep were inoculated multiple times with either Al-containing commercial vaccines, equivalent doses of Al diluted in PBS or PBS only during 16 months. Then, after detection and characterization of mRNAs and miRNAs by RNA-seq technology, the molecular differences withing treatment groups were tested, mainly: Adj-inoculated vs control, Vac-inoculated vs control and Adj-inoculated vs Vac-inoculated.

Few differentially expressed genes were found in each comparison, with nearly 5 times more DEGs in Adj-injected sheep than the Vaccine group when contrasted against control animals. Among the up-regulated genes, there were some previously described in other studies dealing with neurological disorders, namely: *VCAM1*, *TRPM4*, *GDF10* and *NTN1*. The first three were significantly up-regulated in Adj-inoculated animals and the latter was up-regulated in the Adj vs. Vac comparison. *VCAM1* (Vascular Cell Adhesion Molecule 1), member of the immunoglobulin superfamily, is a cellular adhesion molecule needed for the migration of immune cells across the blood brain barrier in inflammatory central nervous system diseases (408). It has been related with neuronal apoptosis after intracerebral haemorrhage, although the exact molecular mechanism for apoptosis regulation remains to be explored, and it may play a role in the development of rheumatoid arthritis (409,410). In addition, *TRPM4* (Transient Receptor Potential Cation Channel Subfamily M 4) is a calcium-activated nonselective ion channel permeable only to monovalent ions that contributes to inflammation-induced neurodegeneration and it may play a role in various neurological diseases like experimental autoimmune encephalomyelitis and MS (411). *GDF10* had been seen to be induced in peri-infarct neurons in mice, non-human primates and humans (412). *GDF10* is considered a stroke-induced signal that promotes axonal outgrowth and enhanced functional recovery after stroke. Finally, NTN1 (Netrin 1), which is included in a family of laminin-related secreted proteins, has been shown to be an important regulator of BBB and to protect the CNS against inflammatory conditions such as MS and experimental autoimmune encephalomyelitis (413). It

has been suggested that *NTN1* acts reducing serum levels of pro-inflammatory mediators and limiting the entrance of immune cells into the CNS (414).

In parallel to the analyses in brain samples carried out in this chapter, those samples were analysed by transversely heated graphite furnace atomic absorption spectrometry and lumogallion staining for Al quantity measurement and Al localization, respectively (399). Sheep showed significantly higher Al content in lumbar spinal cord samples of both treatment groups (Vac- and Adj-injected samples), while in the parietal lobe samples there were no differences, only a tendency to higher Al content (p=0.074) in Adj-injected sheep. More abundant Al deposits were found in the lumbar spinal cord. The Al content and deposits were always more abundant in Adj-injected samples in comparison to Vac-injected samples. It must be pointed that most of the Al accumulation measurements made in the parietal lobe were below 1 μg/g, a level considered safe. The limited quantity of aluminium that reached this tissue could explain the low number of DEGs, when compared to other tissues such as PBMCs. In other study, after intraperitoneal injection of $AlCl_3$ in neonatal rats, significant higher concentrations of Al were found in the hippocampus, diencephalon and cerebellum (415).

The functional enrichment analysis returned multiple GO terms related to mitochondria in the Adj-injected animals, among them: *stress-induced mitochondrial fusion*, *mitochondrial ATP synthesis coupled proton transport*, *positive regulation of mitochondrial membrane potential*, *mitochondrial protein processing* and *positive regulation of mitochondrial DNA replication*. Multiple reports have associated Al toxicity with the production of ROS, which could lead to mitochondrial bioenergetic impairment and to the generation of oxidative stress (416,417). Changes in mitochondrial functions produce oxidative stress, leading to DNA damage and cell death. In addition to mitochondria-related terms, *positive regulation of cardiolipin metabolic process* and *alpha-ketoglutarate transport* GO terms were enriched in Adj-injected sheep. Cardiolipin, a phospholipid located mainly in the inner mitochondrial membrane, is known to promote brain cell viability and to be associated with brain homeostasis, and reduced levels of this phospholipid can result in mitochondrial dysfunction (418). Alpha-ketoglutarate is a source of glutamate, a neurotransmitter that is involved in neurotoxicity (in which ROS takes place, likely due to calcium influx in the cytosol) (419) and the transport of calcium across the inner mitochondrial membrane plays an important role in neuronal physiology and pathology (420).

Then, a co-expression analysis was performed for mRNA and lncRNAs (lncRNA data will be not show) with WGCNA, and 45 modules were achieved. Interestingly, 5 modules correlated with different treatment groups, that is, 3 modules correlated with Vac-injected sheep (mediumorchid4, brown3 and palevioletred3) and 2 with Adj-injected sheep (maroon and burlywood1). Among them, the maroon module was distinguished since it contained 36 DEGs and was enriched in the following KEGG pathways: *ECM-receptor interaction*, *amoebiasis*, *focal adhesion*, *PI3K-Akt signaling pathway* and *protein digestion and absorption*. In a recent study, male rats were exposed through intraperitoneal injections to a complex of aluminium chloride hexahydrate and maltolate, and enrichment analysis revealed terms such as *ECM-receptor interaction*, *protein digestion and absorption*, *focal adhesion focal adhesion* and *PI3K-Akt signaling pathway* (53). These terms are highly concordant to the terms found enriched in the maroon module, which indicates that the few changes seen in our parietal lobe samples may be caused by Al. Among these pathways, the *PI3K-Akt signaling pathway* is expressed during central nervous system development (421) and it is well known that this pathway is particularly important for mediating neuronal survival, differentiation and metabolism (422). In addition, focal adhesion and ECM-receptor interaction signalling are known to be involved in the regulation of synaptic plasticity (423) and NF-κB pathway plays a crucial role on neurogenesis,

cellular responses to neurological injury and neuroinflammation (424,425). Currently, there are few reports regarding the role that these pathways play in the neurotoxicity caused by aluminium.

Regarding the differential expression analysis of miRNA data, the miRNAome of Adj-inoculated animals was clearly dysregulated, while nearly no significant change was detected in Vac-inoculated sheep. Among all the differentially expressed miRNAs, there were some previously related to multiple neurological diseases. One of those was the *let-7b* miRNA, which was found upregulated in Adj-inoculated animals. It has been shown that extracellular let-7b can act as activator of TLR7, which leads to neurodegeneration (426). In addition, *miR-374b* and *miR-30b* were downregulated in the adjuvant group. These miRNAs have been found with a decreased expression in serum samples of patients with sporadic ALS and have been correlated with disease progression (427,428), but their exact mechanism of action regarding ALS is unknown (*miR-30b* is involved with the ECM receptor pathway) . Apart from miRNAs related to neurodegenerative diseases, there were some previously described in studies related to brain injury. The expression levels of *miR-874-3p* and *miR-423-3p* were increased and the expression levels of *miR-99a* and *miR-29c* were decreased in the adjuvant group. *miR-874-3p* expression has been reported to increase after injury in neurons and his over-expression leads to increased stress and vulnerability, affecting inflammatory and apoptotic processes (429). In contrast, *miR-423-3p* might be compensatorily over-expressed in response to apoptosis and exert anti-apoptotic effects in chronic temporal lobe epilepsy (430). Both *miR-99a* and *miR-29c* have been involved in oxidative stress and apoptosis (431,432).

After integrating mRNA and miRNA expression profiles, there were genes related to mitochondria function, maintenance of neural polarity and DNA damage control within the negatively correlated pairs. Mitochondrial transport is crucial for the function of the nervous system due to the particular cellular morphology of neurons and the need to supply energy to remote regions (433). *ACTR10*, which is a negatively regulated and predicted target of *let-7b*, is part of the dynactin complex and absence of the protein encoded by this gene has been shown to disrupt mitochondrial retrograde transport, leading to accumulation of mitochondria in axon terminals (434). In addition, mitochondria are one of the major pools of intracellular Mg and its deficiency seems to be related to mitochondrial dysfunction. *MRS2*, which is other predicted target of *let-7b*, is a mitochondrial Mg transporter that has been related to defects in the organelle and apoptosis (435). It should be pointed out that generally miRNAs function in the cell cytoplasm, but there is evidence of miRNAs located in other locations, being the *let-7* family one of the miRNAs found in mitochondria cytoplasm (436). In adjuvant-only vaccinated animals Al might be causing an imbalance in metal ion levels, among them Mg^{2+}, something that has been seen in rats treated with an intragastric administration of Al gluconate (437). Taken together, the results point towards a miRNA regulation of apoptotic pathways, mitochondrial dysfunction and ECM related pathways upon an intensive vaccination with the adjuvant alone.

4.5 Appendix

Table S4.1: List of selected genes and the corresponding primer sequences for the validation of the total RNA-seq experiment in encephalon.

Gene	GenBank ID	Primer Code	Location*[1]	Exon Junction	Sequence (5'-3')
Target Genes					
CUBN	XM_015099599.1	CUBN-F	975-998	yes	ATCCAAATATGATGACTGTGAGG
		CUBN-R	1057-1074		CTGTACTCGGGCTCTCC
GDF10	XM_004021551.3	GDF10-F	1486-1507	yes	GGACATAGGGTGGAATGAGTG
		GDF10-R	1563-1581		GGACCATCTTGGGCATCG
NDUFAF7	XM_004006018.3	NDUFAF7-F	433-451	yes	GCAGCTTTCCAACTGGTG
		NDUFAF7-R	489-513		CCAAGTTGACTAAATACCCTCAAA
SLC13A3	XM_004014618.3	SLC13A3-F	646-668	yes	CTCAAGAGTTTCTTCCCACAGT
		SLC13A3-R	704-723		AGCAGCATGAGAGGAAAGG
SLC6A20	XM_015102462.1	SLC6A20-F	1468-1485	yes	CTGTCCCTGCTGCTCAT
		SLC6A20-R	1527-1547		GGTCGCTTTCAAATCTGCTC
SPHKAP	XM_012147712.2	SPHKAP-F	721-739	yes	CGCTTCTGTCTGCTTTGT
		SPHKAP-R	786-807		GTGAAACACTGACCAACTTCT
ASZ1	NM_001195309.2	ASZ1-F	985-1007	yes	GCCCTTAAAGAACTGGAAGTAG
		ASZ1-R	1049-1070		GGAACTCATCACCACTGATTT
ND6	DQ320083.1	ND6-F	270-292	no	AGGGACGTTTATTACTGGTTTA
		ND6-R	324-347		CAATTTCCACCTCCTTATCTTTC
RARRES2	XM_012143266.2	RARRES2-F	335-352	yes	GGGCAGTTTGTGAGGCT
		RARRES2-R	408-426		CATTGGGCTTGACCTTGC
EYA1	XM_012183841.2	EYA1-F	528-548	yes	CCAATGGCACCGAAGTTAAA
		EYA1-R	606-627		CAATGGCTGAACCTGAGAAAT
GSN	NM_001246006.1	GSN-F	2041-2059	yes	TCATGCTTCTGGACACCT
		GSN-R	2094-2117		GCTTCTGTCTTCTCTTCTTCTTG
LAMA2	XM_015097399.1	LAMA2-F	736-756	yes	ACCTTGAATGCCGATTTGAT
		LAMA2-R	804-827		CCTTGACCGAGTAGTAGTATCTT
VIM	XM_004014247.3	VIM-F	1361-1380	yes	GGAGAGGAGAGCAGGATTT
		VIM-R	1433-1452		TGTCAACCAGAGGAAGTGA
Reference Genes					
GAPDH	NM_001190390.1	GAPDH-F*[2]	-	yes	GGCGTGAACCACGAGAAGTATAA
		GAPDH-R*[2]	-		CCCTCCACGATGCCAAAGT
ATP1A1	NM_ 001009360	ATP1A1-F*[2]	-	yes	GACTTGAACCGAGGCTTAACAAC
		ATP1A1-R*[2]	-		TCTGGCTAGGATCTCAGCAGC
HPRT	NM_001034035	HPRT-F*[2]			TGGTGGAGATGATCTCTCAACTTTAA
		HPRT-R*[2]			TTCGACAATCAAGACATTCTTTCC
ACTB	NM_001009784.2	ACTB-F	453-474	yes	ATGTTTGAGACCTTCAACACC
		ACTB-R	531-548		TCCATCACGATGCCAGT
TFRC	XM_004003001.2	TFRC-F	1888-1908	yes	GAGCTGGACCTGAACTATGA
		TFRC-R	1962-1984		CAGACCCATATCCCTTATGTCT

*[1] Corresponding Start-End coordinates from NCBI gene annotation.

*[2] Primers from (388). All other primers have been newly designed.

Chapter 5

circRNA annotation

5.1 Introduction

Circular RNAs (circRNAs) are a new class of covalently closed circular non-coding RNAs, formed when a splice donor and upstream acceptor from a linear RNA are linked together, a process also called backsplicing (143). Due to their circular structure, circRNAs are more stable, resistant to RNAse R and have longer half-lives than linear RNAs (145), making them good candidates for disease biomarkers. Despite being discovered long ago, with the first circular molecules (viroids) revealed by electron microscopy in 1976 (147) and the first endogenous circRNA originating from the *DCC* tumour suppressor reported in humans in 1991 (148), for a long time circRNAs were thought to be low abundance products derived from splicing errors (146). With the recent increase in high-throughput sequencing studies, it was shown that these molecules are more common than initially thought and that some of them have important roles in different pathways (438,439). Although the biological function of most circRNAs remains unknown, some circRNAs have been shown to contain clusters of miRNA binding sites that function as miRNA sponges (e.g., the circRNAs related to CDR1 and SRY sequester miR-7 and miR-138, respectively) (152). Other circRNAs have been shown to contain sequences that can act as internal ribosome entry sites (IRESes), such as circ-ZNF609(153), thus can potentially code for proteins. However, their actual translation *in vivo* remains to be probed. Last, circRNAs can regulate a number of processes via protein-binding activity (e.g., the circ-FOXO3 forms a ternary complex with p21 and CDK2) (154).

Despite there have been a myriad of published articles for circRNA annotation in human and mouse, few have realised functional studies in candidate circRNAs. The brain has been one of the most studied tissues for circRNA annotation in human and mouse (440). In a recent study, it has been shown that the circRNA originating from the *SLC45A4* gene (circSLC45A4), which is one of the highest expressed circRNAs in the human frontal cortex, is required to keep neural cells in a progenitor state (441). In another research, circHIPK2 has been shown to be involved in the differentiation of neuronal stem cells, in which overexpression of the circRNA reduced neuronal differentiation (442). In addition, multiple circRNAs has been related to neurological diseases. circZIP-2, which originates from the *SLC39A2* gene, has been associated with aggregation of α-synuclein in a transgenic *C. elegans* model of PD and it may possibly sponge miR-60 (443). One of the most studied brain specific circRNAs is circCDR1as, also known as ciRS-7 due to its ability to acts as a sponge for *miR-7*, which at the same time is known to downregulate the activity of ubiquitin protein ligase A (*UBE2A*) that is involved in clearance of toxic amyloid peptides in AD (444). circHDAC9 has been shown to act as a sponge for *miR-138*, with a subsequent reversion of SIRT1 suppression and excessive amyloid beta production in AD patients (445). Apart from neurological diseases, other studies have found circRNAs associated to a great variety of pathways such as neuronal apoptosis and BBB dysfunction. Function assay has shown that circPTK2 regulates microglia-induced neuronal apoptosis via sponging *miR-29b*, a miRNA known to induce SOC1, block JNK2/STAT3 signaling and inhibit IL-1β production (446).

Regarding BBB, it has been shown that circDLGAP4 acts as a sponge for *miR-143* and overexpression of the circRNA results in inhibition of endothelial-mesenchymal transition (its activation contributes to BBB disruption) (447).

Despite not being so well studied as brain, multiple studies has demonstrated that circRNAs are abundantly expressed in the bloodstream, with expression levels comparable to the cerebellum (448). Furthermore, due to their stability and longer half-lives, circRNAs have been proposed as promising biomarkers for diagnosis of human diseases (449). A recent study on mouse macrophage cells treated with lipid A (the active component of lipopolysaccharide (LPS)) identified the LPS-inducible circRNA *mcircRasGEF1B*, which was shown to regulate the stability of the mature *ICAM-1* mRNA (an intracellular adhesion molecule that plays a role in immune response) (450). In other study with CD28(+)CD8(+) T cells and CD28(-)CD8(+) T cells from healthy elderly or adult control subjects, circRNA100783 was proposed to play a role in phosphoprotein-associated functions during CD28-related CD8(+) T cell ageing (451). circNR3C1 has been shown to act as *miR-382-5p* sponge in blood serum of age-related macular degeneration samples and the miRNA sequestering resulted in PTEN expression increase and inhibition of the AKT/mTOR pathway (452).

In sheep, a number of circRNAs were previously identified from RNA sequencing data. Li et al. detected 6,133 and 10,226 circRNAs in prenatal and postnatal muscle and pituitary glands of sheep, respectively (156,157). Interestingly, they observed an association of some circRNAs with economically important traits, such as the growth and development of muscle related signaling pathways in the first tissue and the regulation of hormone secretion in the second. In addition to this, the same group identified 9,231 circRNAs differentially expressed in the estrus and anestrus pituitary system of sheep (155). Last, 886 circRNAs were detected in the skeletal muscle by Cao et al., and some of them were reported to be involved in muscle cell development and signaling pathway (158). Characterizing the circRNA profiles of specific tissues and cell types is a promising way to reveal functional properties of circRNAs.

Thus, seeing how circRNAs participates in a myriad of pathways in brain samples and that blood circRNAs may participate in the fine-tuning of immune responses (448), it would be interesting to address any function that those molecules may have in Al adjuvancy. In addition, this work would improve the still in progress circRNA annotation in sheep in two tissues not previously studied (regarding circRNA annotation). Thus, the objectives of this work were:

1. to characterize circRNAs from ribosomal depleted RNA sequencing libraries in encephalon (parietal lobe cortex) and PBMCs, improving the sheep transcriptome.

2. to address the functional role of circRNAs in Al adjuvancy (if any).

5.2 Material and methods

The samples employed to annotate circRNAs have been previously used for differential expression analyses. For a detailed summary of the samples and RNA extraction methods see the corresponding section, chapter 3 for PBMC samples (3.2 Material and methods) and chapter 4 for parietal lobe cortex samples (4.2 Material and methods). In addition, a detailed description of the workflow for circRNA characterization can be seen in chapter 2 (2.3.4.4 Workflow).

5.3 Results

5.3.1 circRNA characterization

Ribosomal RNA depleted total RNA-seq datasets (parietal lobe cortex and PBMCs) from the previous studies were re-analysed to characterize circRNAs. Two bioinformatics tools, Segemehl (333) and DCC (328), were selected for circRNA identification, which resulted in 12,475 and 60,375 candidate circRNAs in encephalon and 19,611 and 63,138 candidate circRNAs in PBMC samples by segemehl and DCC, respectively, in each tissues. It must be pointed that these tools only detect the backsplice junction in their search for topologically inconsistent reads (non-colinear splicing) and they do not make any assumptions on the retained exons or introns, since the remaining reads cannot be distinguished if they belong to the linear or circular transcript. Out of all the circRNAs detected in the encephalon, 4,996 had concordant coordinates in both tools. After filtering circRNAs based on their abundance and expression patterns among samples, 2,510 circRNAs were selected for subsequent analyses. In PBMCs, 10,414 circRNAs were concordant between tools. After filtering, 3403 circRNAs were retained for further analysis. The list of circRNAs and their details will be published as supplementary material in a research article. The naming of circRNAs in each tissue list was performed by assigning sequential unique numeric identifiers. From the 2,510 and 3,403 circRNAs detected in encephalon and PBMCs, 1,236 (49.24% of the encephalon circRNAs and 36.32% of the PBMC circRNAs) were present concordantly in both tissues (see figure 5.1). In other studies, it has been shown that approximately 30% of the detected blood circRNAs overlapped with circRNAs expressed in the cerebellum (448). The counts from DCC were taken as reference abundance values.

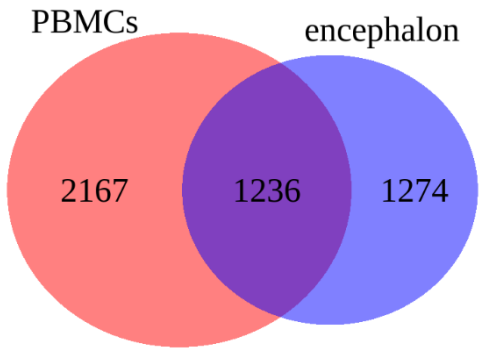


Figure 5.1: Venn diagram with the number of circRNAs detected in each tissue after filtering for a minimum expression in at least three samples.

In the available literature a number of studies have described the principal characteristics of circRNAs in human and mouse (152,161). In our sheep data, in both tissues, we observe that the longer the chromosome, the more circRNAs are detected (Figure 5.2a and 5.2b). In addition, supposing that all exons between backsplice junction coordinates are conserved, the circRNAs are most commonly formed by two or three exons, being those composed of two exons the most prevalent ones (Figure 5.2c and 5.2d). This is in accordance with what was previously described in other species (152). A representation of the location of each circRNA in the reference genome is given in figure 5.3 for encephalon and figure 5.4 for PBMCs. In those figures can be seen regions with high concentrations of circRNAs, generally regions in which a few genes host multiple circRNAs. In the case of encephalon samples (Figure 5.3), the regions chr1:188-189Mb, chr6:101-102Mb and chr25:10-11Mb contained 11, 10 and 11 circRNAs respectively. The most outstanding region is the one from chromosome 25, in which

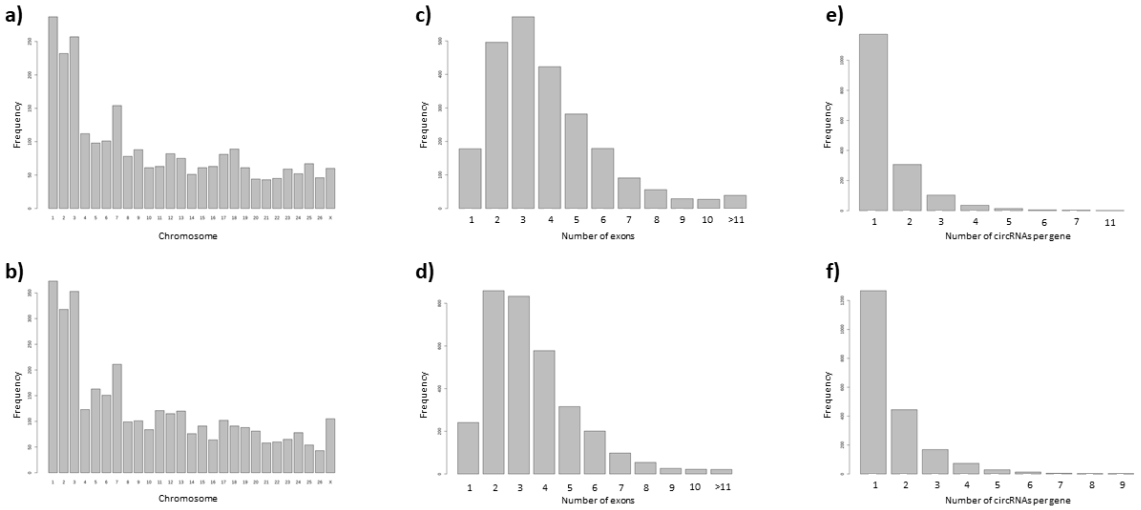


Figure 5.2: Plots depicting some characteristic circRNA properties. a) and b) Number of circRNAs identified in each chromosome in encephalon and PBMCs, respectively. c) and d) Number of exons inside each circRNA whose origin is an annotated gene in encephalon and PBMCs, respectively. On the x-axis the number of exons a circRNA has from start-end coordinates and on the y-axis the number of circRNAs that are composed of a determined number of exons. e) and f) Bar plots in which the x-axis represents how many circRNAs are from the same host gene and the y-axis shows the number of genes that host a specific number of circRNAs in encephalon and PBMCs, respectively.

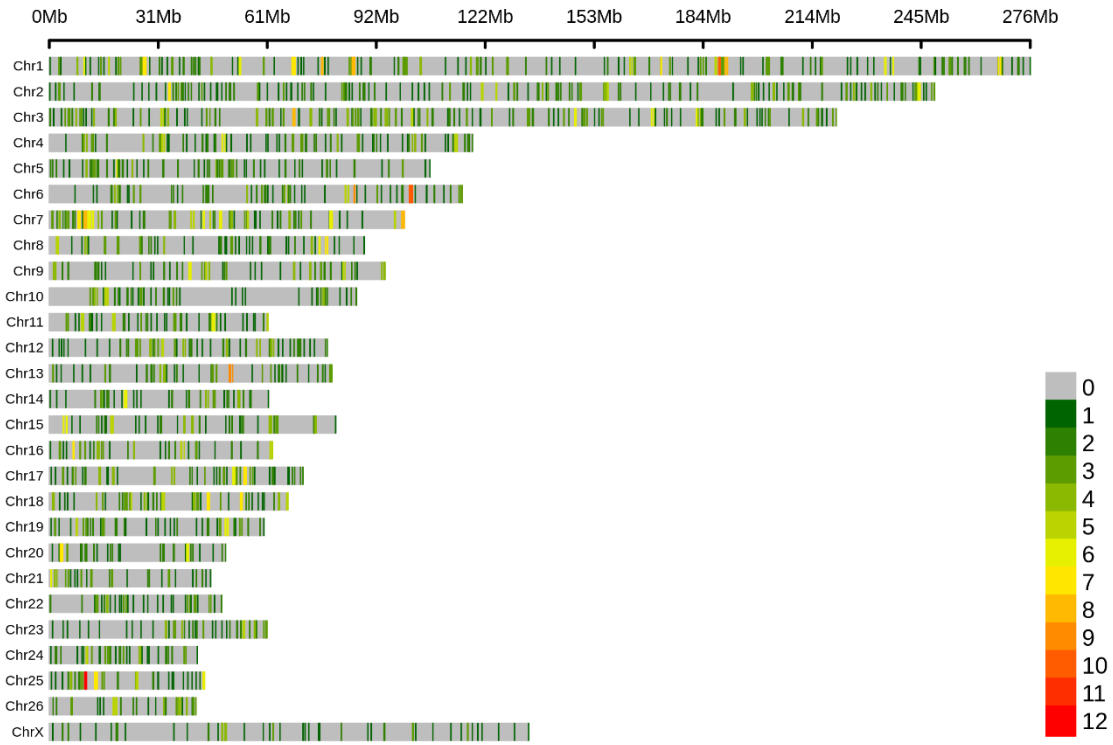


Figure 5.3: Location of detected circRNAs in encephalon. Each chromosome was divided in bins of 1Mb and the number of circRNAs was counted in each bin. The colour code represents the number of circRNAs detected in the bins.

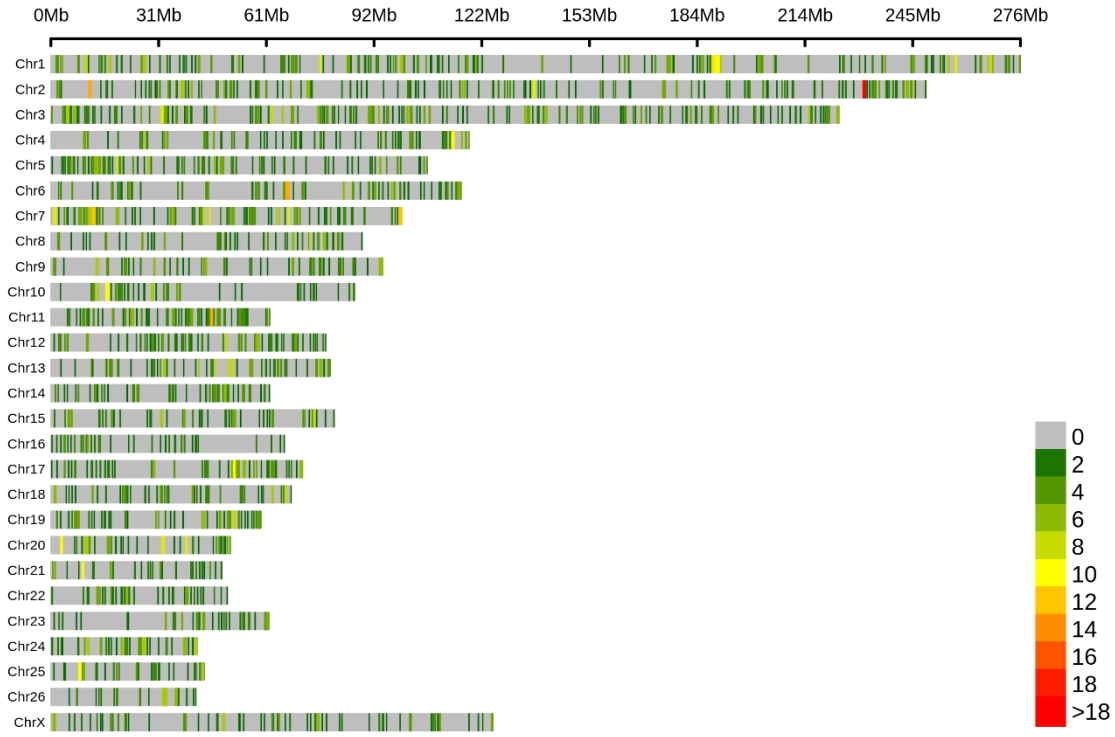


Figure 5.4: Location of detected circRNAs in PBMCs. Each chromosome was divided in bins of 1Mb and the number of circRNAs was counted in each bin. The colour code represents the number of circRNAs detected in the bins.

the 11 circRNAs surge from the same gene, ryanodine receptor 2 (*RYR2*). In a recent study, the RYR2 gene was predicted by bioinformatic tools to have 177 circRNA isoforms in human heart samples (453) , understanding as isoforms in circRNAs the multiple forms that host a single gene. The other two regions had multiple genes that hosted multiple circRNAs. The region of chromosome 1 had the genes *ZNF148* (with 1 circRNA), *SNX4* (with 7 circRNAs), *LMLN* (1) and *TFRC* (1), while the region in chromosome 6 had the genes *PTPN13* (3), *AFF1* (2) and *SPARCL1* (4). In addition, in the case of PBMC samples (Figure 5.4), the regions chr2:11-12Mb, chr2:231-232Mb, chr6:66-68Mb and chr7:99-100Mb contained 12, 18, 17 and 12 circRNAs, respectively. Similar to the regions of encephalon samples, these regions had a few circRNAs that hosted multiple circRNAs, mainly: in chr2:11-12Mb, genes *PTBP3* (4) and *SUSD1* (4); in chr2:231-232Mb, genes *SP110* (5), *ENSOARG00000020646* (7) and *SP100* (5); in chr6:66-68Mb, genes *TEC* (5) and *SLAIN2* (5); and in chr7:99-100Mb, genes *TTC7B* (2), *RP56KA5* (3), *CASC4* (5) and *CTDSPL2* (2).

Out of the 2,510 candidate circRNAs detected in encephalon, 2,372 overlap with 1,642 annotated sheep genes. Of those circRNAs that originated from an annotated gene, 1,927 were concordant with an annotated exon-intron boundary in both ends, while in the other cases, despite the overlap with an annotated gene, at least one end was not concordant with an annotated exon-intron boundary. Concerning the 3,403 circRNAs detected in PBMCs, 3,249 were found to originate from 2,006 annotated sheep genes. Of these, 2,597 were concordant with an annotated exon-intron boundary in both ends. In some cases, the cause of the discrepancy between the annotated exon-intron boundaries and the circRNA backspliced junctions could be explained by the incomplete state of the sheep gene annotation. The majority of genes host only one circRNA in both tissues (Figure 5.2e and 5.2f). There were a few cases in which a gene hosted more than 6 isoforms in both tissues. In the encephalon samples, there were 3 genes (*PANK2*, *DZIP3* and *SNX4*) that each one hosted 7 circRNAs and one gene (*RYR2*)

that hosted 11 different circRNA isoforms. In the PBMC samples, there were 5 genes (*KMT2C*, *JCHAIN*, *COP1*, *KIF3A* and *ACAP2*) that each one hosted 7 circRNAs, 2 genes (*EFCAB3* and *GPCPD1*) hosting 8 isoforms and other 2 genes (*PICALM* and *VPS13C*) with 9 isoforms. In a recent study on genes of the *MLL* rearranged acute leukaemia in normal blood cells (454), 31 circRNA isoforms of the *PICALM* gene were reported in monocytes.

Despite studies on circRNA annotation in different species are growing in number, there is a lack of functional characterization. One of the most well characterized circRNAs is the one related to the CDR1 gene, whose function as a sponge for miR-7 has been shown to be specific of neuronal tissue in mouse (455). Although CDR1 is not annotated in sheep, blasting the human sequence of this gene against the sheep reference genome results in a single hit, matching a region of circRNA4960, detected in our encephalon samples. We lifted the coordinates of the sheep backsplice junctions (sheep genome version Oar_3.1) to the human genome (version hg38) with the UCSC liftOver tool (341) and found that circRNA4960 is homologous to the human CDR1-AS. Interestingly, circRNA4960 was one of the most expressed in our cortex samples. Among the highly expressed circRNAs detected in encephalon (Table 5.1) other two were homologous to previously characterized human circRNAs, circRNA4266 and circRNA4357, which originate from HOMER1 and ZNF609 genes, respectively. Within the circRNAs with highest levels of expression detected in PBMCs (Table 5.2) there was no functionally characterized circRNA previously described in other species.

Table 5.1: Top 10 highly expressed circRNAs in encephalon for each sample. “Sample”, number of samples in which the circRNA is detected among the 10 most expressed.

Name	Chr	Start	End	Strand	Gene	Gene Name	Samples
circRNA4266	7	10226343	10245400	-	ENSOARG00000017295	HOMER1	12
circRNA1370	15	74333115	74346193	-	ENSOARG00000003032	PHF21A	12
circRNA4960	X	89513503	89514803	-		CDR1-AS	11
circRNA22	1	10099323	10102102	+	ENSOARG00000019470	ZMYM4	10
circRNA1805	18	66676757	66681832	+	ENSOARG00000005628	EIF5	9
circRNA2718	23	35401154	35419026	+	ENSOARG00000008819	ROCK1	7
circRNA598	10	28577657	28578297	-	ENSOARG00000010917	PDS5B	7
circRNA2217	2	146153913	146169101	+	ENSOARG00000005956	KCNH7	6
circRNA1395	16	6732929	6736996	+	ENSOARG00000004394	FAM169A	6
circRNA4387	7	48356079	48359956	+	ENSOARG00000020844	SLTM	6
circRNA438	1	234794720	234799994	-	ENSOARG00000004134	MED12L	5
circRNA1840	19	9567084	9578895	+	ENSOARG00000016371	ARPP21	5
circRNA913	12	42986759	43003148	+	ENSOARG00000009688	RERE*[2]	5
circRNA2780	23	58208176	58223449	+	ENSOARG00000005569	ZNF532	4
circRNA2747	23	43282769	43284768	-	ENSOARG00000001837	SPIRE1	2
circRNA969	12	68716251	68720208	+	ENSOARG00000009633	ANGEL2	2
circRNA4225	7	3056580	3056998	+	ENSOARG00000016256	KCNN2	2
circRNA3977	5	85636228	85648776	-	ENSOARG00000016099	MEF2C	2
circRNA4357	7	42641887	42642760	-	ENSOARG00000020735	ZNF609	1
circRNA317	1	171863910	171882996	+	ENSOARG00000019010	DZIP3	1
circRNA1601	17	52116136	52117521	+	ENSOARG00000006283	PITPNM2	1
circRNA3748	4	86579656	86598085	-	ENSOARG00000018580	AASS	1
circRNA4708	9	14695072	14695957	-	ENSOARG00000002124	ADGRB1	1
circRNA2092	2	82728978	82746338	-	ENSOARG00000013900	ZDHHC21	1
circRNA666	10	77178839	77187483	-	ENSOARG00000004729	NALCN	1

Table 5.2: Top 10 highly expressed circRNAs in PBMCs for each sample. “Sample”, number of samples in which the circRNA is detected among the 10 most expressed.

Name	Chr	Start	End	Strand	Gene	Gene Name	Samples
circRNA1097	10	11535094	11536868	-	ENSOARG00000006858	ELF1	13
circRNA2237	13	32987272	33001090	-		ZEB1	13
circRNA9166	7	43514992	43519072	-	ENSOARG00000020780	USP3	11
circRNA3570	18	1380967	1386808	-	ENSOARG00000009094	UBE3A	10
circRNA9195	7	48356079	48359956	+	ENSOARG00000020844	SLTM	8
circRNA4776	2	177871845	177873523	-			7
circRNA2303	13	46999890	47008442	-	ENSOARG00000017627	GPCPD1	7
circRNA7862	4	112437200	112438114	-	ENSOARG00000001140		5
circRNA1078	1	273591055	273596954	-	ENSOARG00000015604	SATB1	5
circRNA5791	23	37284690	37289374	+	ENSOARG00000009781	SMCHD1	5
circRNA3022	15	74333115	74346193	-	ENSOARG00000003032	PHF21A	5
circRNA1811	12	24282247	24297165	-	ENSOARG00000013880	AIDA	4
circRNA7465	3	207812049	207812406	+	ENSOARG00000007751	IFFO1	4
circRNA3410	17	52116136	52117521	+	ENSOARG00000006283	PITPNM2	4
circRNA6610	3	25707012	25707671	-	ENSOARG00000016596	SMC6	3
circRNA6790	3	61888465	61889612	+	ENSOARG00000000446	LIMS1	3
circRNA3496	17	62670908	62672930	-	ENSOARG00000013244	SPPL3	3
circRNA2644	14	45143572	45144245	+	ENSOARG00000004796	CD22	3
circRNA8972	7	12355216	12355633	-	ENSOARG00000018058	DENND4A	3
circRNA5601	22	21584852	21595148	-	ENSOARG00000016804	FBXW4	3
circRNA6617	3	28061734	28076540	-	ENSOARG00000017236	PUM2	2
circRNA4121	19	59224897	59225941	-	ENSOARG00000004600	EEFSEC	2
circRNA8856	6	115601583	115611675	-	ENSOARG00000015178	RNF4	2
circRNA8583	6	66537573	66540717	-		SENP6	1
circRNA9141	7	42256500	42256834	+	ENSOARG00000020714	PTGDR	1
circRNA7242	3	151073625	151077096	-	ENSOARG00000001039	RAP1B	1
circRNA4228	2	27958067	27961323	-	ENSOARG00000008048	FAM120A	1
circRNA5157	20	11010704	11011610	+	ENSOARG00000014288	FGD2	1

5.3.2 Conservation

circRNAs have been reported to be evolutionary conserved and to be tissue specific to some extent. As previously stated, it has been shown that approximately 30% of the detected blood circRNAs overlapped with circRNAs expressed in the cerebellum of human and mouse data (448). In our samples, 1236 circRNAs (36.32% of all detected blood circRNAs) were detected in both tissues. As there is no actual circRNA database recording sheep circRNAs, a search of sheep circRNA characterization articles was done, requiring that the detected circRNAs were given at least as supplementary material. A total of 4 different articles in pituitary gland and longissimus dorsi muscle were found (155–158). Those circRNAs were compared to the ones described in our samples. Notably, only 175 circRNAs were consistently detected in all tissues at the same time, including ours (Figure 5.5). Such low concordance is in agreement with other studies, which showed that the expression of circRNAs can be tissue-dependent (456). In addition, our

results showed that 421 (16.77%) and 841 (24.71%) circRNAs were exclusive to the encephalon and PBMCs data.

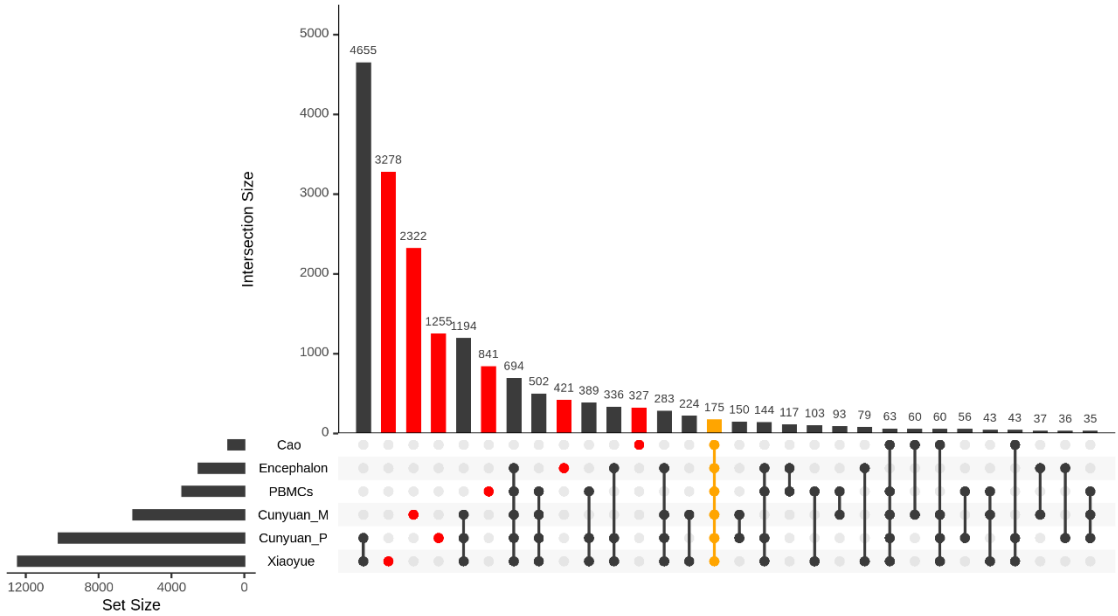


Figure 5.5: UpSet plot with the comparison of detected circRNAs in different studies. Encephalon and PBMCs refers to the circRNAs detected in this study, while Cunyuan_P (pituitary gland), Xiaoyue (pituitary gland), Cunyuan_M (longissimus dorsi muscle) and Cao (longissimus dorsi muscle) refers to the circRNAs detected in (155–158), respectively. Cells filled with a dot indicate the circRNA is in the corresponding database, while empty cells indicate that the circRNA is not present in the corresponding database. In red the circRNAs that are exclusively expressed in one database and in orange the circRNAs common to all databases. Intersections with less than 30 elements were removed for visualization purposes.

In addition to this, the detected circRNAs were compared to the human circRNAs annotated in CIRCpedia (340). First, sheep circRNA coordinates were translated to human ones with the UCSC liftOver tool (341) and classified based on their backsplice junction conservation. Out of the 2,510 detected circRNAs in encephalon, 52 splice sites coordinates could not be lifted. For the rest, nearly all had at least one reported human circRNA utilizing one of the splice sites. A total of 1,606 (63.98%) circRNAs were completely homologous to a human circRNA (Figure 5.6a). In PBMCs, out of the 3,403 detected circRNAs, 93 splice sites coordinates were not lifted to human, while 2,114 (62.12%) circRNAs were found to be completely homologous to a human circRNA (Figure 5.6b).

5.3.3 Functional enrichment analysis

A functional enrichment analysis was conducted with g:Profiler (276) on the GO (457) and KEGG (458) databases for both tissues, by considering the terms annotated for the parental genes of the detected circRNAs and after setting as background all the genes expressed in the corresponding tissue. Terms with an FDR less than 0.05 were selected as significant. For visualization of the significant GO term clustering, the complete networks are represented in figures 5.7 and 5.8 for encephalon and PBMC, respectively. A more detailed zoom of some clusters can be seen in figures 5.9 and 5.10. The 20 most enriched KEGG pathways are shown in

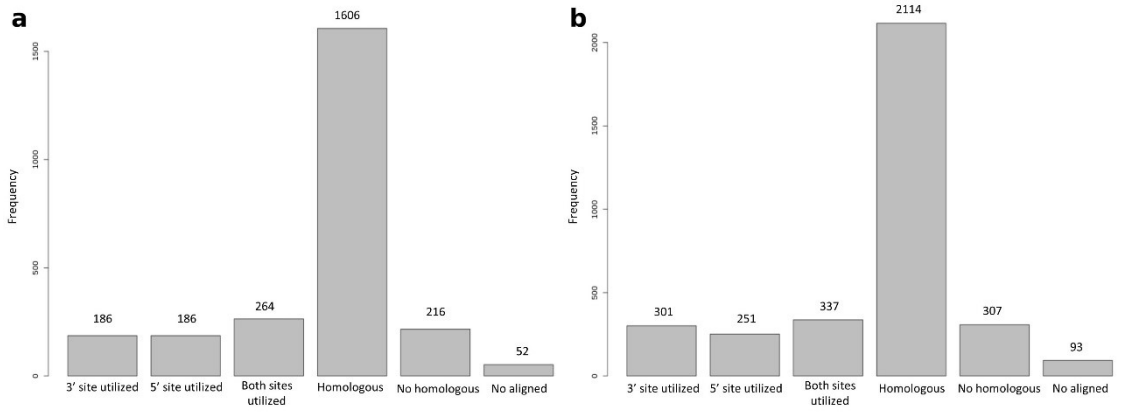


Figure 5.6: Bar plot with the result of the conservation analysis. In the x-axis the different categories described in material and methods and in the y-axis the number of circRNAs in each category. a) Encephalon; b) PBMCs.

figures 5.11a and 5.11b, for encephalon and PBMCs, respectively. Among the GO terms significantly enriched in encephalon, there are a number of terms related to synapse regulation, presynaptic endocytosis, behaviour, brain development and myelination, while among the KEGG pathways glutamatergic synapse, dopaminergic synapse and serotonergic synapse were enriched, suggesting an important role for some circRNAs in synaptic functions. This would be in concordance with brain circRNAs discovered in human and mouse, in which an enrichment of circRNAs in synapses has been shown and important roles in synaptic plasticity and neuronal function have been suggested (339,440). Instead, in PBMCs, we retrieved GO terms related to B- and T-cell proliferation, T-cell differentiation, activation and regulation of immune response and neutrophil degranulation. In addition, there were some clusters related to housekeeping functions such as DNA repair, DNA replication, cell cycle, mRNA splicing and telomere maintenance. In a recent study using datasets of human hematopoietic cells from the SRA repository, it was found that circRNA-hosting genes were enriched in housekeeping functions such as DNA repair, regulation of cell cycle and transcription (459). In both tissues, the KEGG T-cell receptor signaling pathway and B-cell receptor signaling pathway were enriched, suggesting that some circRNAs may be involved in basic immune system functions.

5.3.4 circRNA sponges

To identify circRNAs which could function as miRNA sponges, we compared all 2,510 (encephalon) and 3,403 (PBMCs) predicted circRNAs with clusters of miRNA binding sites reported by Pan *et al*. (460) in the human genome, a dataset that comprises a total of 3,673 predicted sponges for 1,250 miRNAs. Only 3 (encephalon) and 4 (PBMCs) sheep circRNAs overlapped one or more candidate sponges-miRNA pairs, and those entries for which the predicted sponged miRNA does not have a homologous pre-miRNA in sheep were filtered out. As a result, one circRNA (circRNA4960) overlapping predicted sponges for two miRNAs (miR-7 and miR-1224) was predicted in encephalon tissue, while two circRNAs, circRNA2342, which overlaps predicted sponges for miR-409, miR-383, miR-370, miR-369 and miR-212, and circRNA8181 for miR-124, were predicted for PBMC samples. Then, those sponge candidates in sheep and their overlapping human sponges were screened for miRNA binding sites with RIsearch2 (344). After removing overlapping binding sites as described in Pan *et al.* (460), 44 and 65 binding sites were respectively found on circRNA4960 for miR-7 and miR-1224. Although the

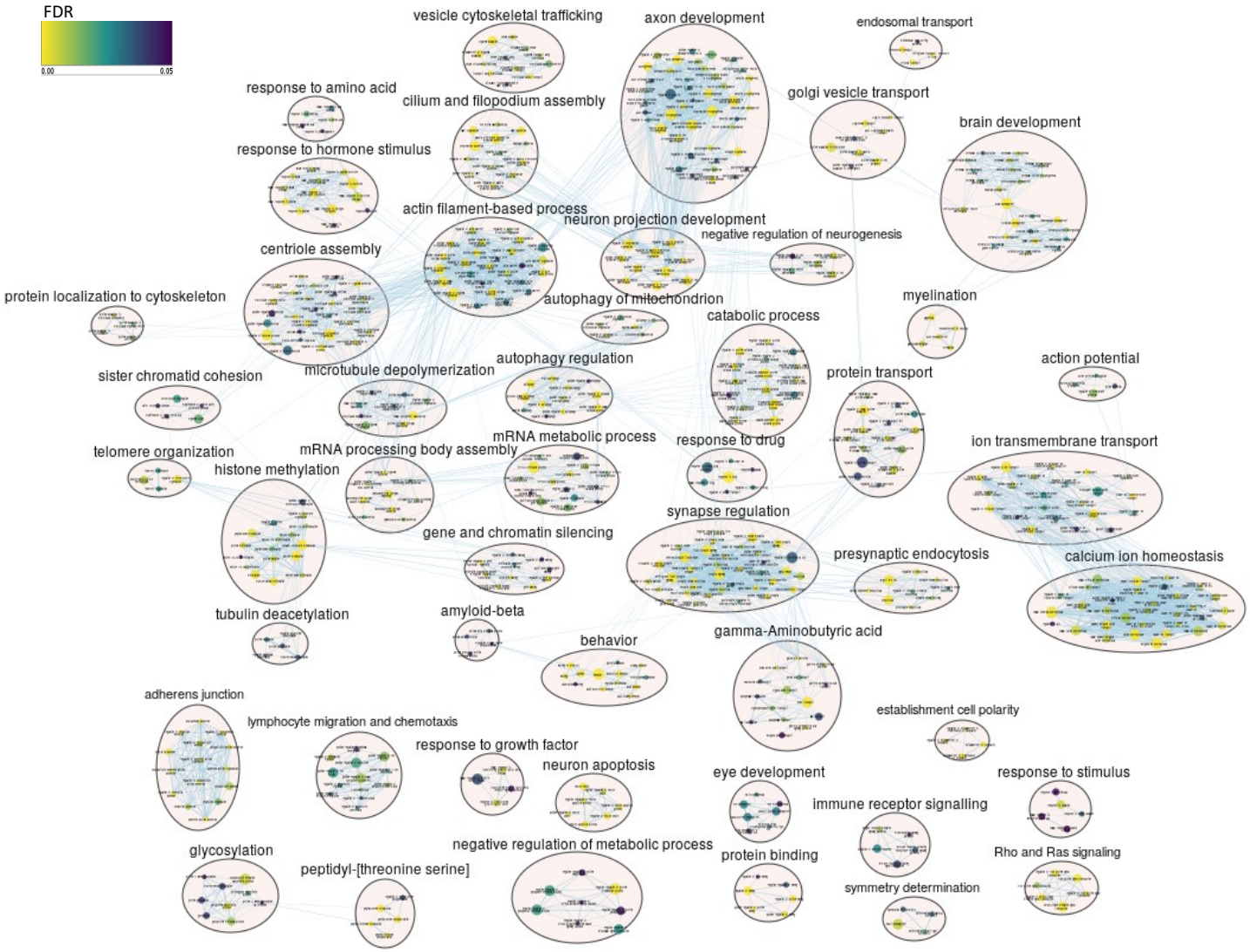


Figure 5.7: Complete network from enriched GO terms by g:Profiler in encephalon and visualized in Cytoscape after clustering with Autoannotate. Node size correspond to number of genes expressed from the term; edge size represents the number of genes that overlap between different terms; and colour represents the significance level (FDR).

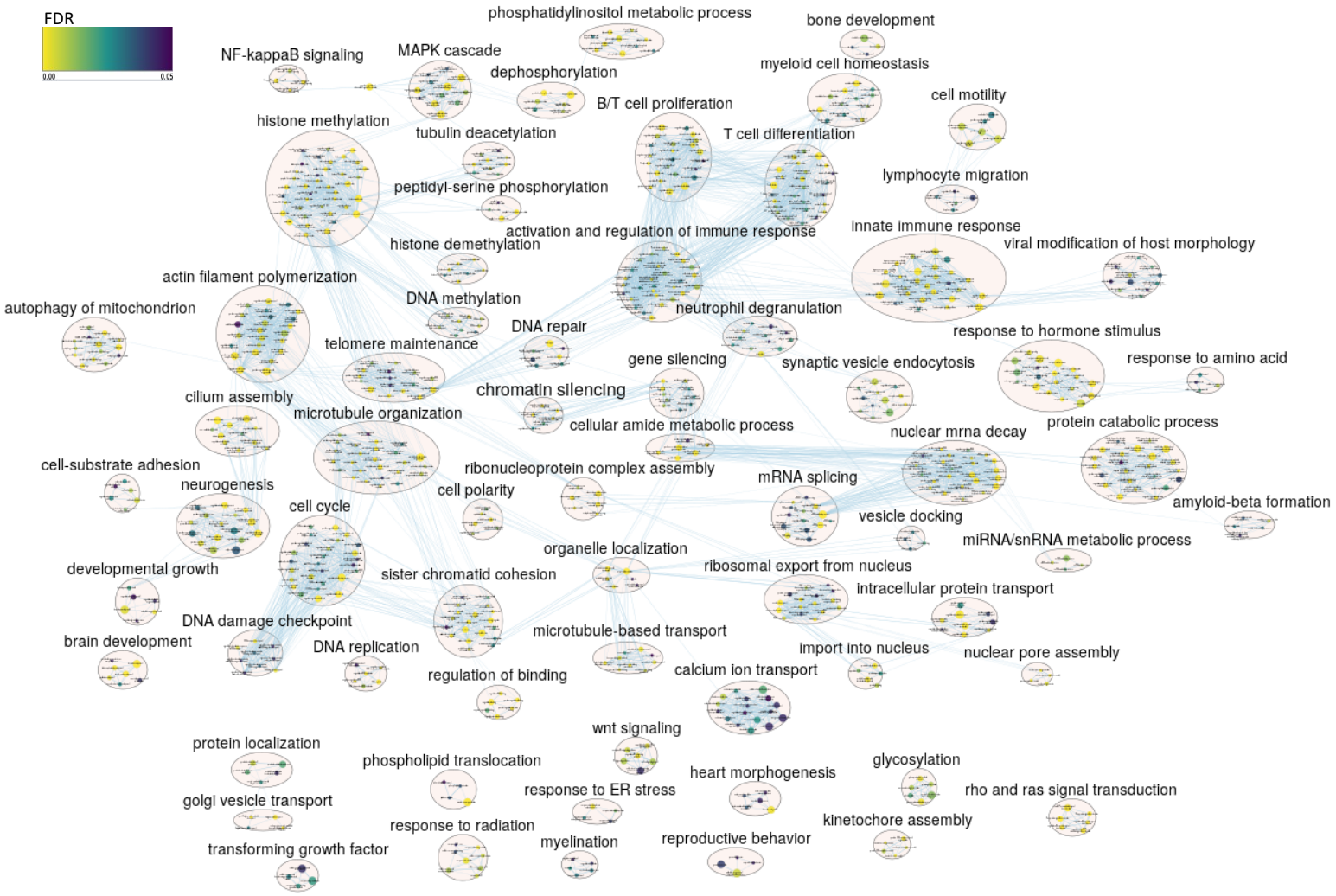


Figure 5.8: Complete network from enriched GO terms by g:Profiler in PBMCs and visualized in Cytoscape after clustering with Autoannotate. Node size correspond to number of genes expressed from the term; edge size represents the number of genes that overlap between different terms; and colour represents the significance level (FDR).

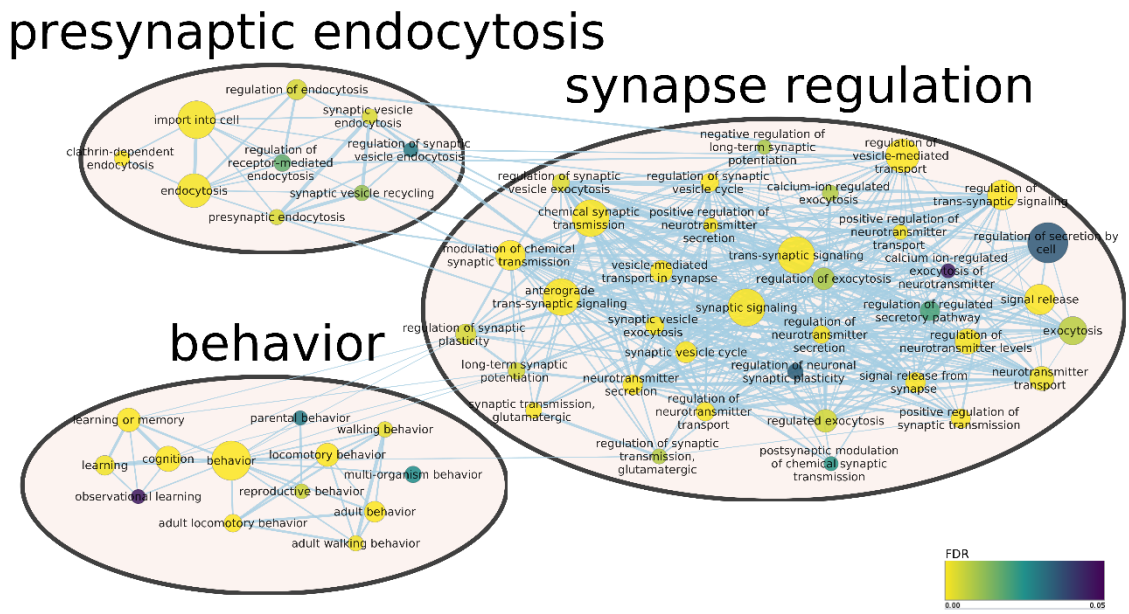


Figure 5.9: Sub-network from enriched GO terms by g:Profiler in encephalon and visualized in Cytoscape after clustering with Autoannotate. Node size correspond to number of genes expressed from the term; edge size represents the number of genes that overlap between different terms; node colour represents the significance level (FDR).

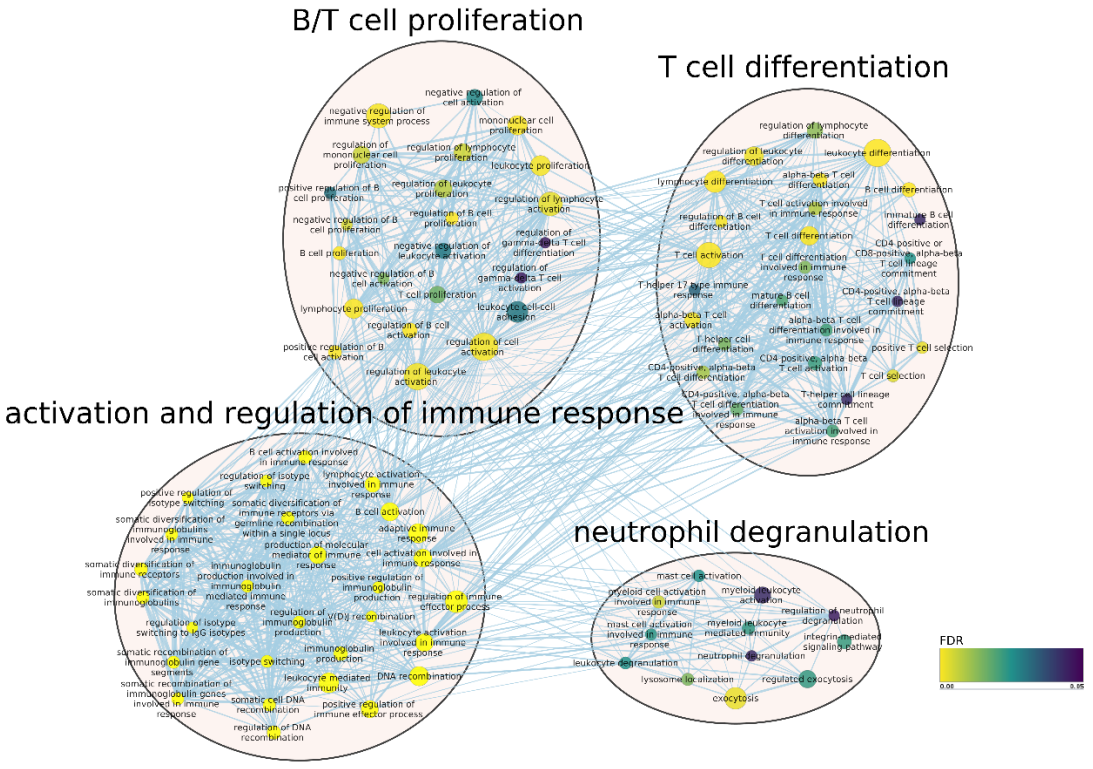


Figure 5.10: Sub-network from enriched GO terms by g:Profiler in PBMCs and visualized in Cytoscape after clustering with Autoannotate. Node size correspond to number of genes expressed from the term; edge size represents the number of genes that overlap between different terms; node colour represents the significance level (FDR).

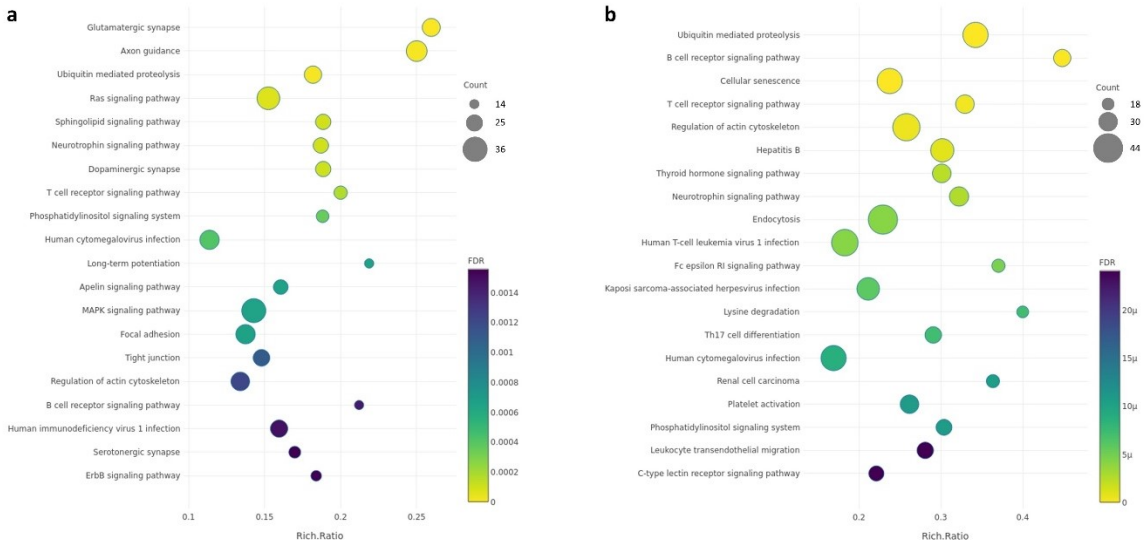


Figure 5.11: The 20 most enriched KEGG pathways by g:Profiler. The bubble plots show in the Y-axis the enriched KEGG pathways, while in the X-axis the rich ratio is represented (rich ratio=amount of differentially expressed genes in the term/all genes included in the term). Size and colour of the bubble represent the number of differentially expressed genes in the KEGG pathway and enrichment significance (FDR), respectively. a) Encephalon; b) PBMCs.

sheep circRNA4960 (CDR1-AS in human) is shorter than the corresponding cluster of miRNA binding sites detected in human for miR-7 and miR-1224, the per-base binding sites ratio is higher in sheep, further underlying a possible functional role of this molecule in the sheep brain. Recent studies have shown that miR-671 has sufficient complementarity with CDR1-AS to induce AGO2 endonucleolytic cleavage and, based on this, an alternative function for this circRNA molecule as miRNA shuttle system, releasing its miR-7 cargo upon binding with miR-671, has been proposed (461). Interestingly, the binding pattern of miR-671 in sheep is identical to the human one and includes 13 canonical base pairs in the seed region, and only 1 mismatch over the entire sequence, therefore it is sufficient to obtain cleavage by AGO2. Hence, our results support both the miRNA sponge and the miRNA shuttle functions previously proposed for CDR1-AS in brain and suggest a possible similar mechanism for miR-1224, which is reported as highly expressed in brain according to the Genotype-Tissue Expression (GTEx) Project v8.

In contrast, when the biding sites of sheep and human sponge sequences were reanalyzed in PBMCs by RIsearch2, it was shown that miRNA binding sites were scattered far away from one another over both the exonic and the much longer intronic regions of circRNA2342 and circRNA8181, with few bindings overlapping with the clusters of miRNA binding sites identified in human, hence we could not infer any sponge activity for these circRNAs. The complete list of binding sites identified for sheep circRNA-miRNA pairs in both encephalon and PBMCs candidate circRNA sponges is intended to be given as supplementary material in a future publication.

5.3.5 Differential expression analysis

A principal component analysis (PCA) was done with the circRNA expression data from encephalon and PBMCs (Figure 5.12). Similar to the differential expression analysis in Chapter 4, the sample 116E from encephalon was treated as an outlier and it was removed from the analysis. Despite having an adequate RIN value (7.6), it was observed a low 260/230 absorbance

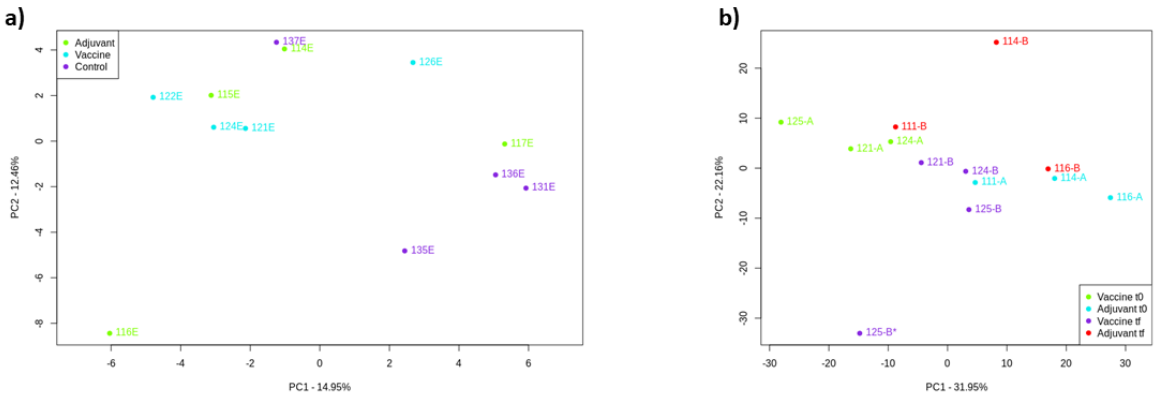


Figure 5.12: Principal component analysis (PCA) of the encephalon and PBMCs circRNA data. a) PCA plot for the encephalon samples. b) PCA plot for the PBMC samples after batch effect correction with the Harman package from R.

ratio (a secondary measure of nucleic acid purity) of 0.81 for that sample. Lower ratios of 1.8 may indicate the presence of co-purified contaminants.

Then, the differential expression analysis was performed with the R package DESeq2 (254). We did not detect any differentially expressed circRNA in any comparison after considering an adjusted p-value<0.05 as cut-off. We also performed differential expression analysis normalizing the data as spliced reads per billion mapping (SRPBM), and by applying a Kruskal-Wallis test before correcting for multiple comparison with the Benjamini & Hochberg method. Also in this case, there were no significant differences between groups when an adjusted p-value<0.05 was taken as cut-off.

For the PBMCs samples, the Harman R package (247) was applied to remove any batch effect in the data after normalizing by SRPBM. Then, both the limma package (248) and Kruskal-Wallis test were used to test for differential expression, but no circRNA was found to be differentially expressed in any comparison with an adjusted p-value<0.05.

5.4 Discussion

CircRNAs are a novel (or have at least recently been dropped from be treated as low abundance products derived from splicing errors) class of endogenous non-coding RNAs with a cyclic structure formed through a covalent bind of a linear transcript. Lately, circRNAs have gained more attention due to their abundance, their expression levels in specific tissues and their involvement in different biological functions, particularly studied in human and mouse (462–464). Among all studied tissues in human and mouse, it has been shown that brain samples and blood samples are highly enriched in circRNAs (448). In the case of brain circRNAs, it has been shown that they are particularly enriched in synapses (440) and that they can have a great variety of functions, such as neuronal differentiation (442), neuronal apoptosis (446) and BBB dysfunction (447), while others have been related to some neurological diseases (443–445). In the case of blood circRNAs, the functional roles are not so well studied, but some have been related to immune functions (450), but most of them have been related to transcription regulation (448). In addition, studies on circRNAs in non-model organism, such as sheep, are still lacking, and there is no database recording such data yet.

Taking all together, the main objectives of this study were twofold: first, to improve the annotation of circRNAs in sheep, studying two tissues still not used for circRNA annotation in sheep; secondly, to address any function that circRNAs may have in Al adjuvancy, since blood

circRNAs may participate in the fine-tuning of immune responses (448). For that, the samples of encephalon and PBMCs used in previous chapters (PMBC samples in chapter 3 and encephalon samples in chapter 4) were reanalysed to annotate novel circRNAs.

A total of 2,510 and 3,403 circRNAs were detected in parietal lobe cortex and PBMCs, respectively, via *in silico* analysis of ribo-minus total RNA sequencing data. From these circRNAs, 1,379 were completely novel circRNAs (841 exclusive to PBMC samples, 421 exclusive to encephalon samples and 117 expressed in both tissues). The annotation of sheep circRNAs was improved by adding relevant information such as conservation and potential function. Most of the identified circRNAs in both tissues are from annotated genes, generally formed by two or three distinct exons, in agreement with what has been previously reported in human and mouse data (465). In addition, we observe that circRNAs are widely expressed in both these tissues in sheep, which was somewhat expected since circRNAs are enriched in mammalian brain and human PBMCs (466).

Some circRNAs have a tissue-dependent or developmental stage-dependent expression pattern (456). 1236 circRNAs (36.32% of all detected blood circRNAs) were detected in both tissues, which is concordant with approximately 30% of the detected blood circRNAs overlapping with circRNAs expressed in the cerebellum of human and mouse data (448). In addition, the circRNAs detected in this study were compared to other sheep circRNA identified in pituitary gland (155,156) and in longissimus dorsi muscle (157,158). Only 175 circRNAs were detected in all tissues at the same time, while several hundreds of circRNAs were exclusive to each tissue, which shows how some circRNAs have a tissue-dependent expression. Furthermore, given that numerous circRNAs have exhibited evolutionary conservation between human and mouse (467), the circRNAs detected in this study were analysed for backsplice site conservation, by comparing them to the human circRNAs available in CIRCpedia. We found that 1,606 (63.98%) and 2,114 (62.12%) sheep circRNAs had completely conserved backsplice sites between human and sheep in encephalon and PBMCs, respectively. Among the most expressed circRNAs, circRNA4266 and circRNA4357, in order originating from the HOMER1 and ZNF609 genes, had been previously characterized in other species. Consistent with this, it has been shown that the circRNA related to HOMER1 has a regulatory role in cell growth in human bronchial epithelial cells, as its silencing promotes cell proliferation (468). The circRNA originated from ZNF609 has been shown to adsorb miR-150-5p and to upregulate SP1 transcription factor, promoting the proliferation of nasopharyngeal carcinoma cells (469). In addition, this circRNA has been related to myoblast proliferation and the fact that its sequence includes an open reading frame and that a fraction of this circRNA is loaded into polysomes indicates that it may encode for proteins (153).

It was previously proposed that the binding activity between circRNAs and RNA binding proteins (RBPs) can have regulatory effects (162), which suggests that circRNAs can impact the same functional processes in which the corresponding linear host gene is involved. Under the assumption that the function of a circRNA may be associated with the known function of its parental gene, GO analysis indicated that the circRNAs identified in encephalon are related to synapse regulation, behaviour, learning process and brain development, while KEGG pathway analysis also related these circRNAs to synapses and to pathways implicated in cell proliferation such MAPK/ERK pathways, the last ones being previously linked to circRNAs (465). In contrast, in the PBMCs samples, GO terms associated with the immune system such as B- and T-cell proliferation, neutrophil degranulation, the MAPK cascade and the NF-κB signaling were enriched, as well as DNA methylation and histone modification, supporting the possibility that circRNAs could be related to epigenetic alterations, as previously suggested (470). In both tissues the B- and T-cell receptor signalling pathways were enriched, in addition to Fc epsilon RI signaling pathway, Th17 cell differentiation and platelet activation in PBMCs samples, indicating a potential functional role for circRNAs in the immune system response.

In addition, it has been shown that circRNAs can act as miRNA sponges through abundant binding sites for miRNAs in their sequence and modulating the activity of miRNAs in their target genes (471). For that purpose, the circRNAs detected in encephalon and PBMCs were screened for the presence of clusters of miRNA binding sites. The circRNA CDR1-AS, which corresponds to circRNA4960 in this study, contains numerous binding sites for miR-7 and miR-1224, both reported to be expressed in the mammalian brain. In agreement with our expectations, we observed that this circRNA is highly expressed only in our encephalon samples.

After the characterization of the circRNAs, a differential expression analysis was performed, in an attempt of identifying any circRNA that may play a role in Al adjuvancy. We did not detect any differentially expressed circRNAs in any of the two tissues, which indicates that circRNAs may not be connected with aluminium adjuvant effects. Despite this, it should be noted that no differential expression analysis software has been specifically designed to handle circRNA data, in which expression levels are generally lower compared to mRNA and are subjected to greater variability.

Chapter 6

General discussion and conclusions

This chapter will provide a general discussion off all the work performed throughout this thesis, mainly: the differential expression analysis of PBMC and encephalon samples, and the annotation of circRNAs in both tissues. Some suggestions will be given regarding possible future works and the reached conclusions will be listed at the end.

6.1 General discussion

6.1.1 Experimental design

The present work aimed to characterize the response of aluminium (Al) adjuvant-based vaccinations in sheep and to identify novel genes or regulatory elements that partake in the still partially unknown mechanism of action of Al adjuvants. For that purpose, Rasa Aragonesa sheep were subjected to repetitive inoculations of commercial vaccines composed of aluminium hydroxide (AH) or to equivalent doses of aluminium containing adjuvant diluted in phosphate-buffered saline (PBS).

It is well known that Al is a non-essential element for the human body and is thought that it lacks any essential biological function. Al by itself has been shown to have cytotoxic properties: at concentrations of 100 μg/ml, Alhydrogel was shown to increase cellular mortality in T_H1 cells (391); in erythrocytes of common carp exposed to different concentrations of Al, it was shown that Al exposure produced oxidative stress, which induced DNA damage and gene modifications in cells (392); and in thymocytes and lymphocytes of young mice exposed to different concentrations of aluminium chloride, it was shown a dose- and time-dependent damage of the plasma membrane (but not causing acute cell death) (393). Most researches have studied Al adjuvancy in vaccines in short term experiments, in which a single or a few doses of the adjuvant are administered. The fact that the body is not able to execrate all injected Al in a short-term period of time through normal mechanisms such as urine (46), points towards a high persistence of the material in the body. Moreover, the fact that Al may be able to reach distant organs though translocation of the material by phagocytic immune cells from the monocytic cell line (472,473) or other unknown mechanism are among the most recent concerns. A long-term experiment would be helpful to study the fate of injected Al and to assess if the Al of multiple vaccinations accumulates in the body due to its high persistence.

Different animal models can be used for that purpose. Despite the fact that mice models have a lower cost and have been extensively used for Al adjuvant experiments, they may fail to recreate some aspects of the mechanism at study. Some of those studies give the Al adjuvant through intraperitoneal injections (37,53), a route completely distinct from human or other large mammals, so the Al distribution and mechanism from those mouse models may differ from the animals that receive the inoculations through the subcutaneous or intramuscular route. Larger animals like sheep share more similarities to human regarding physiology, anatomy, metabolism, genetics and size, making them a good alternative (2–4). Furthermore, farm

animals have commercial vaccines for multiple diseases designed specifically for them. In contrast, mouse experiments of Al adjuvant usually use vaccines intended for human use, some by dilution with saline or buffer, which is not recommended since it can change the main properties of the adjuvant in the formulation (agglomerate size, adsorbed antigen,…), while others use a fraction of the human dose (around one-fifth or one-tenth, 0.1-0.05 ml) (19).

Our experiment solves some of the problems that have been presented in other experiments. First, commercially approved AH-based vaccines that are usually given during their productive period to sheep were chosen for the experiment. Second, an intensive vaccination schedule was planned, in which sheep received in 475 days, always following manufacturer recommendations, most of vaccines that would be expected throughout their lifespan. The Al load in the tissues under a longer vaccination schedule, and more similar to the field conditions, may differ to the one seen in these sheep. However, such a long experiment is not feasible. Other option would be to use adult animals, but it would be nearly impossible to find adult animals that have grown under controlled conditions, in addition to not be able to distinguish if the detected Al is from the vaccine or other source such as food, water or other probable exposures to the element throughout their life.

In the experimental design, priority has been given to the homogeneity of the animals analysed in the different groups, using young animals from the same herd, without any prior vaccination before the experiment and with a period of adaptation to the new environment under the best conditions of feeding and temperature. Regarding vaccination in young sheep (the animals in this experiment started their vaccination at the age of 5-months old), It must be pointed that the immune system is completely developed in 5-month-old lambs (365) and differences related to development should not account for the differences observed in this study. Moreover, the vaccines used in sheep are the same regardless of age, with the same dose quantity and the same administration protocols.

The tissues studied in this work, peripheral blood mononuclear cells (PBMCs) and parietal lobe cortex, were chosen mainly to characterize the immune response elicited by the adjuvant in the case of PBMCs and to study any molecular changes in brain, in the case of detecting Al in said tissue, in cortex samples. The histopathological analyses and behavioural changes of the animals were studied by other research groups under the same project (399,474). As a novelty in comparison to other studies of Al adjuvant activity, our group applied RNA sequencing (RNA-seq) to characterize molecular changes in the transcriptome in the previously commented tissues. Therefore, libraries for total RNA-seq, which undergo ribosomal RNA depletion, retaining most non-polyadenylated non-coding RNA, and miRNA-seq were prepared.

This work has a limited number of samples, mainly due to restricted funding and the experimental design. Despite of that, it meets the minimum required, which is three biological replicates per group, to make inferences on the population (89). It must be pointed that the profiles under study are at the beginning (t=0) and at the end of the experiment (t=475), and as a consequence, it is complex to determine if some changes are elicited by the overall administration schedule or if the last dose has a greater effect. This experiment was carried out to study the effects of repetitive inoculations of Al adjuvant-containing products and their expected accumulation in the organism, independently from the identification of individual inoculations, the granuloma exact age or the role of specific vaccine antigens. It is expected that the cumulative effect of the inoculations would be seen, but it cannot be discarded that the latter has a greater effect on the response of the animals than the previous ones.

6.1.2 Bioinformatic analysis

So far, the sheep reference genome annotation (Oar_v3.1) is still a work in progress, being non-coding elements poorly annotated. One example of this would be miRNAs. In the last release of miRbase (Release 22.1), 153 mature miRNAs can be found annotated for *ovis aries*, while other species such as *bos taurus* or *homo sapiens* have over 1045 and 2693, respectively. Other elements such as lncRNAs or circRNAs have nearly no annotation and despite there are already several research groups working on their annotation, there is no database with a clear nomenclature system recording them. The study of said molecules and their interactions with coding genes may help to improve the current annotation of the sheep refence genome, in addition to broad our knowledge of the mechanism of action of Al adjuvant. For the purpose of annotating novel non-coding elements in our sheep libraries, in addition to a high sequencing depth, a paired-end library was chosen to achieve a better alignment and characterization of those novel non-coding RNAs.

The steps followed in the pipelines for the differential expression analyses of total RNA-seq and miRNA-seq datasets are well stablished, with a wide variety of tools to choose from for each step (89). The choice of tools would depend mostly of the objectives of the study, the species under study and the state of their reference genome, the type of sequencing library (paired-end or single-end, strand specificity, sequencing depth, etc.) and desired performance (variations in specificity and sensitivity). As previously mentioned, it must be taken into account that the sheep reference genome and annotation is still a work in progress and is lacking specifically in annotations of non-coding elements such as miRNAs, lncRNAs and circRNAs.

The choice of aligner, STAR, was based mainly by the performance shown by multiple aligner comparisons (221,223,224). There are variations in performance depending on the dataset used for comparison, but generally STAR was always found to be among the best. In addition, this aligner has in favour to be very fast, to have implemented in its code a two-pass strategy in which a first alignment is done to update junction splice site information and to have an extensive and detailed documentation.

In addition to the differential expression analysis, a weighted correlation network analysis was performed in data from both tissues. Taking into account that genes belonging to a pathway are usually co-expressed, genes are clustered by their correlation and represented as a network in tools such as WGCNA (279). These networks can help to broad the knowledge in the clinical trait of interest by correlation of the trait with module eigengenes (a representative gene expression of a module, calculated as the first component of a PCA). In addition, hub genes (highly connected genes) from those modules related to the clinical trait of interest may have key roles in the treatment under study.

Regarding the miRNA-seq data analysis, miRNA target prediction tools use multiple features from the transcript sequence for their purpose. Nearly all tools use the complementarity of the seed sequence, the thermodynamic stability of the miRNA-mRNA complex and the predicted secondary structure of the miRNA, while other use in addition other specific information such as sequence evolutionary conservation and target site accessibility. Regardless of the tool used, there is a general consensus regarding how all these tools usually returns a great number of false positives (475). In addition, it has been reported that most predicted targets may be functionally insensitive to the miRNA repression (475). In an attempt to reduce the number of false positives returned by these tools, three tools (Miranda, PITA and TargetScan) using different features for target prediction were used and their intersection was taken as candidate miRNA targets. It must be pointed that these are predictions based on

sequence information and are further studied checking the correlation between miRNA-target pairs. For a precise dissection of a miRNA regulatory network further experiments will be required and it would be an interesting starting point for future work in Al mediated miRNA-mRNA changes. Until recently, most researchers have done miRNA gain-of-function and loss-of-function experiments to identify differentially expressed genes upon alteration of the expression of specific miRNAs. Recently, interrogation of specific miRNA-target interaction by a CRISPR/Cas9-mediated system has been shown (476), which allows to disrupt or restore individual interactions on demand.

Completely different is the pipeline used for circRNA characterization based on total RNA-seq data, despite there are multiple tools for their characterization (325,327,331,333), it is still a work in progress. The major problems for their characterization in this kind of datasets are that circRNAs constitute a small fraction of reads in common cells lines and that most of them are lowly expressed (477). In addition, all tools for circRNA characterization developed so far rely in the detection of junction reads that align in a non-linear manner, also called backsplice junctions. Thus, most tools only return the counts detected in a fixed window (the backsplice junction) and they cannot discern if the remaining reads belong to the linear or circular isoform of a gene, not giving any information of the retained exons/introns between the backsplice junction coordinates. In an attempt to reduce false-positive calls, it was decided to use the intersection of two different circRNA characterization tools, segemehl and DCC, in addition to a filtering criterion in which a circRNA needed to be detected in multiple samples to be taken as a true circRNA candidate. Then, multiple characteristics of the detected circRNAs were retrieved, assuming that all exons along the backsplice junction coordinates were retained. Despite this limitation, this is a first step toward circRNA annotation in sheep, providing a general background on the functions that circRNAs may have in the studied tissues. For a more detailed characterization of the circRNA sequences, improvements in the experimental design have been proposed. Some researchers have started to use RNase R, which is an exonuclease that digest nearly all linear transcripts, to enrich for circRNAs before sequencing (477). However, it must be pointed that some circRNAs have been shown to be sensitive to RNase R, in addition to some linear transcripts still remaining after RNase R treatment (478).

6.1.3 Differential gene expression related to aluminium

In this work, changes previously related to Al were observed in both tissues. In the PBMC samples, there was a clear modulation of the immune response caused by Al stimulation, with thousands of genes differentially expressed in Vac-inoculated and Adj-inoculated animals, while in encephalon samples only a few genes were found differentially expressed in Adj-inoculated animals only. This dissimilarity in gene expression in both tissues may be explained by the low levels of Al detected in the cortex samples (399), but most of the few differentially expressed genes in cortex were expressed in a similar manner to other Al exposure works. After the differential expression analysis in cortex samples, it was shown nearly no differential expression in the animals vaccinated with commercial vaccines and the quantity of Al detected in parietal lobe samples from the Vac-inoculated sheep was similar to those of the control group (399), which indicate that commercial formulations are pretty safe under the conditions of this experiment. Completely different was the case of Adj-inoculated animals, in which a tendency to higher Al content was detected when compared to control samples (399). It must be pointed that most of the Al accumulation measurements made in the parietal lobe were below 1 μg/g, a level considered safe. With nearly 5 times more DEGs in Adj-injected sheep than the Vac-

inoculated animals, among the differentially expressed genes there were terms usually found dysregulated in neurological diseases, namely: *VCAM1*, *TRPM4*, *GDF10* and *NTN1* (409–413). Taken together, it seems that under the terms of this experiment Al was able to reach the brain and induce molecular changes when is free from any antigen. Thus, it raises some concerns on the safety of a large number of vaccine trials, which uses Al adjuvant-containing placebo groups (479).

Regarding the changes observed in PBMCs, a clear secretion of inflammatory cytokines, previously reported to be induced by Al (36,369–371), and activation of the NF-κB signalling pathway was observed in both adjuvant-treated groups. The main differences in both groups were in the expression of genes from the cytokine-cytokine receptor interaction pathway, which were clearly downregulated in Adj-injected animals. This may reflect what has been seen in the granulomas formed after inoculation in those animals (375), in which Adj-inoculated animals showed a lower persistency of granulomas, pointing towards a quicker clearance of Al at the injection site. Thus, explaining the milder immune response observed in those sheep. The fact that Al is cleared from the injection site at a faster pace in adjuvant only animals, in addition to the smaller size of Al particles when is free of antigen, may also explain why higher levels of Al is detected in lumbar spinal cord and a tendency to higher content in parietal lobe is observed (399). In addition, macrophages and other phagocytic immune cells may be the main players in the systemic distribution of Al seen in Adj-inoculated animals in this work. It was shown that macrophages at the injection site contained Al and that Al-containing macrophages tended to form aggregates in the lymph nodes (375).

Finally, it must be pointed that after circRNA characterization in both tissues, there were not found circRNAs with striking roles in the Al adjuvant activity. There is no tool designed specifically for circRNA expression data based on rRNA depleted total RNA-seq libraries. Most researchers use tools such as DESeq2 and edgeR, which are based in a negative binomial distribution, but no study has been done to show if the negative binomial model is suitable for circRNA expression data, which only counts backsplice junction reads. At least in our samples, circRNA expression data has generally very low counts (a few highly expressed circRNAs originated most of the counts) and is zero-inflated. Due to the different structure of circRNA expression data and uncertainty of whether the methods used so far are adequate, multiple methods were applied to our data. Independent of the choice of method, we did not detect any differentially expressed circRNAs in any of the two tissues, which indicates that circRNAs may not be connected with aluminium adjuvancy. This is the first work characterizing circRNA expression levels after exposure to Al adjuvants to the date.

6.1.4 Guidelines for future work

As a first exploratory analysis, this study has returned some interesting miRNAs and novel predicted targets, mostly those related to mitochondria in encephalon and to NF-κB pathway in PBMC samples, that would be interesting to study in functional studies to understand better the Al mechanism of action and neurotoxicity capacity, mainly: let-7b/MAP3K2, miR-125b/SNX27 and miR-16b/CHEK1 in PBMCs and let-7b/ACTR10 and let-7b/MRS2 in encephalon. One option to validate the importance of a miRNA/target pair would be a gain-of-function experiment by a miRNA mimic in the interrogated cell type and subsequent Western blot using a specific antibody of the changed protein (480). Recently, interrogation of specific miRNA-target interaction by a CRISPR/Cas9-mediated system has been shown (476), which allows to disrupt or restore individual interactions on demand, a more precise method for direct target validation.

In addition, a RT-qPCR analysis of some candidate circRNAs would be advisable for validation of the bioinformatic analyses. The main difference of RT-qPCR for circRNAs compared to linear transcripts is in the design of divergent primes which face away from each other in the linear RNA and the use of RNAse R treatment to enrich the circRNA population (481).

6.2 Conclusions

6.2.1 Experimental design and bioinformatic analysis

1.1. Total RNA-seq libraries in this work had a mean sequencing depth greater than 70 million reads in both tissues, while miRNA-seq libraries has a sequencing depth greater than 16 million reads. Thus, the high sequencing depths of the libraries in this study has allowed the discovery of 1,379 novel circRNAs in the studied tissues, in addition to 39 and 148 novel sheep miRNAs (at least not annotated in the miRbase database) in PBMCs and cortex samples, respectively.

1.2. Due to the limited number of samples, to improve the detection of differentially expressed genes, different DE tools have been applied to data analysis, mainly: edgeR, DESeq2 and limma. Their choice, especially for edgeR and DESeq2, was based mainly for the true positive identification rate and controlled FDR at lower fold changes shown by these tools when a low number of replicates are available. In addition, these tools allow for confounding factors to be added when modelling the dataset, which enable their use in complex datasets. It was completely necessary in the PBMC dataset, which was a longitudinal study with two time points (before any vaccination, t0; and after all vaccinations, tf).

1.3. A weighted correlation network analysis (WGCNA) was performed in datasets from both tissues, PBMCs and cortex, to group in clusters co-expressed genes, which are supposed to belong to the same or related pathways. Thus, the constructed networks would allow to understand better complex interactions of genes in the sheep samples and see which pathways may be related to Al adjuvant treatment. It must be pointed that the results achieved must be interpreted with caution, as with any other analysis based in correlations, it is important to have a great number of samples per treatment for the correlation to be meaningful. However, the networks built in this work are a first exploratory analysis that will allow identifying elements related to Al adjuvancy.

6.2.2 Expression changes due to aluminium in PBMCs

2.1. The increase in inflammatory signals detected by RNA-seq analysis led to the activation of the NF-κB signaling pathway in both treatment groups, pathway that was enriched in the Vac tf vs. Vac t0 and Adj tf vs. Adj t0 comparisons. There were multiple genes from the NF-κB family, such as *NFKB2*, *RELA* and *RELB*, which were highly expressed in Vac- and Adj-inoculated animals simultaneously.

2.2. Despite there were similarities in the immune response in both treatments, there were a few discrepancies. Commercial vaccines induced a clear upregulation of *IL1B*, *IL2RA*, and *PTX3*, consistent with the induction of an ongoing immune response against the vaccine. In contrast, inoculation with Al alone generally downregulated the mRNA expression of several proinflammatory genes, including *IL1B*, *IL8*, *TLR2*, *NOD2*, or *IL2RA*, suggesting a milder induction of the immune response. In concordance with a milder

immune response in Adj-inoculated sheep, it was shown that genes from the *cytokine-cytokine receptor pathway* were downregulated compared to Vac-inoculated sheep. In addition to immune related genes, there were upregulated genes related to apoptosis in both treatment groups, among them: *TP53BP2*, *CSRNP1*, *TEAD*, *CDCA7* and *PPP1R15A*. This would be concordant with release of danger signals such as uric acid or host DNA from necrotic or damaged cells.

2.3. The expression of the *NLRP3* inflammasome has been previously related to Al adjuvant activity and it has been reported that *IL1B* activation is dependent of the expression of the inflammasome. In this work, the *NLRP3* inflammasome had a constant expression when sheep that received commercial vaccines were compared to their initial stage, before any vaccination, while it was found downregulated in sheep that received the AH adjuvant diluted in phosphate-buffered saline. Thus, it seems that the inflammasome is not required for Al adjuvant activity in sheep under the conditions of this experiment, which point towards an inflammasome independent activation of the immune response.

2.4. In the gene co-expression analysis of PBMC samples, it was shown that the lavenderblush3 and coral1 modules, which were related with both Vac- and Adj-inoculated animals, were composed of genes crucial for the correct function of the immune system. Genes expressed in those pathways were similarly co-expressed in both treatments, showing that an immune response is elicited by Al adjuvants in both treatment groups, the one receiving commercial vaccines and the other one receiving equivalent doses of the adjuvant diluted in PBS.

2.5. Among the differentially expressed miRNAs, *let-7b* (upregulated in Adj tf vs. Adj t0 comparison), *miR-125b* and *miR-99a* (both upregulated in Vac tf vs. Vac t0 comparison) were found to be related to the NF-κB pathway. There is a broad activation of the NF-κB pathway in our samples and it seems that said pathway is highly regulated by multiple miRNAs in the immune response to Al adjuvants.

6.2.3 Expression changes due to aluminium in encephalon

3.1. It was shown nearly no differential expression in the animals vaccinated with commercial vaccines and the quantity of Al detected in parietal lobe samples from the Vac-inoculated sheep was similar to those of the control group, which indicate that commercial formulations are pretty safe under the conditions of this experiment.

3.2. In contrast, with nearly 5 times more DEGs in Adj-injected sheep than the Vac-inoculated animals, among the differentially expressed genes there were terms usually found dysregulated in neurological diseases, namely: *VCAM1*, *TRPM4*, *GDF10* and *NTN1*. In the case of Adj-inoculated animals, a tendency to higher Al content was found in cortex samples. It seems that Al is able to reach the brain when is free of any antigen.

3.3. Among the differentially expressed miRNAs in Adj-inoculated animals, multiple predicted targets related to mitochondria function (*ACTR10* and *MRS2*, both targeted by let-7 family members), maintenance of neural polarity and axon growth (*RUFY3,* targeted by *let-7b*) and apoptosis (*NAA50* and *UNC5D* targeted by *miR-197-3p* and *miR-410-3p*, respectively) were found. *MRS2*, which is other predicted target of *let-7b*, is a mitochondrial Mg transporter that has been related to defects in the organelle and apoptosis. In adjuvant-only vaccinated animals Al might be causing an imbalance in metal ion levels, among them Mg^{2+}.

3.4. The majority of the differentially expressed genes were found in the maroon module from the co-expression analysis, a module correlated with Adj-inoculated sheep enriched in terms such as *ECM-receptor interaction*, *amoebiasis*, *focal adhesion*, *PI3K-Akt signaling pathway* and *protein digestion and absorption*. The changes seen in our Adj-inoculated samples are quite similar to those observed in brain samples after Al exposure in other species such as rats. Among the hub genes of the maroon module, in addition to a great number of differentially expressed genes, there were terms related to blood brain barrier (*ADGRA2* and *NTN1*), ERK signaling (*INSR*, *ITGA9*, *OSMR*, *COL18A1*, *LAMA2*, *BCL2L11*, *ADAM17*, *COL4A3*, *COL4A4*, *COL4A6*, *COL2A1* and *BMP4*) and calcium signaling (*APOOL*, *HOMER3* and *TMBIM1*). This module was composed of genes essential for the correct function of the brain and a more detailed study of these genes and related pathways may help to understand better the Al mechanism of action and how it is able to reach the brain.

6.2.4 circRNAs and aluminium

4.1. Circular RNAs (circRNAs) were characterized for the first time in sheep parietal lobe cortex samples and PBMCs. A wide expression of circRNAs was found in both tissues. A total of 2,510 and 3,403 circRNAs were detected in parietal lobe cortex and PBMCs, respectively, of which 1,379 were completely novel circRNAs (841 exclusive to PBMC samples, 421 exclusive to encephalon samples and 117 expressed in both tissues). This study broads the current sheep circRNA annotation with two tissues not previously used for circRNA characterization in sheep.

4.2. 1236 circRNAs (36.32% of all detected blood circRNAs) were detected in both tissues, which is concordant with approximately 30% of the detected blood circRNAs overlapping with circRNAs expressed in the cerebellum of human and mouse data. Thus, it seems that some sheep circRNAs have a tissue-dependent or developmental stage-dependent expression pattern.

4.3. The circRNAs detected in this work were compared to a human circRNA database, CIRCpedia. It was shown that approximately the 63% of circRNAs in both tissues had completely conserved backsplice sites when compared to human backsplice junctions. Thus, it seems that most circRNAs are evolutionary conserved between human and sheep.

4.4. Under the assumption that the function of a circRNA may be associated with the known function of its parental gene, PBMC circRNAs were related to multiple immune functions such as B- and T-cell proliferation, neutrophil degranulation, the MAPK cascade and the NF-κB signaling, while parietal lobe cortex circRNAs were related to synapse regulation, behaviour, learning process and brain development. In this work, a homologous sheep circRNA to a human circRNA from the CDR1 gene locus was characterized exclusively in cortex samples and it was shown that the sheep sequence was enriched in binging sites for miR-7, supporting that the miRNA sponge activity demonstrated in humans for miR-7 was also retained in sheep.

4.5. Independent of the choice of differential expression method (edgeR, DESeq2 or non-parametric test such as Kruskal-Wallis), we did not detect any differentially expressed circRNAs in any of the two tissues, which indicates that circRNAs may not be connected with aluminium adjuvancy, or that we could not detect it with the planned experimental design.

Bibliography

1. Eurostat. *Agriculture, forestry and fishery statistics - 2018 edition*. Publications Office of the European Union (2018). doi:10.2785/340432

2. Rogers CS. Engineering large animal species to model human diseases. *Curr Protoc Hum Genet* (2016) **2016**:15.9.1-15.9.14. doi:10.1002/cphg.18

3. Camacho P, Fan H, Liu Z, He J-Q. Small mammalian animal models of heart disease. *Am J Cardiovasc Dis* (2016) **6**:70–80. doi:10.3390/JCDD3040030

4. Perentos N, Martins AQ, Watson TC, Bartsch U, Mitchell NL, Palmer DN, Jones MW, Jennifer Morton A. Translational neurophysiology in sheep: Measuring sleep and neurological dysfunction in CLN5 Batten disease affected sheep. *Brain* (2015) **138**:862–874. doi:10.1093/brain/awv026

5. Luján L, Pérez M, Salazar E, Álvarez N, Gimeno M, Pinczowski P, Irusta S, Santamaría J, Insausti N, Cortés Y, et al. Autoimmune/autoinflammatory syndrome induced by adjuvants (ASIA syndrome) in commercial sheep. *Immunol Res* (2013) **56**:317–324. doi:10.1007/s12026-013-8404-0

6. Reed SG, Orr MT, Fox CB. Key roles of adjuvants in modern vaccines. *Nat Med* (2013) **19**:1597–1608. doi:10.1038/nm.3409

7. McKee AS, Marrack P. Old and new adjuvants. *Curr Opin Immunol* (2017) **47**:44–51. doi:10.1016/j.coi.2017.06.005

8. Heffernan MJ, Zaharoff DA, Fallon JK, Schlom J, Greiner JW. In vivo efficacy of a chitosan/IL-12 adjuvant system for protein-based vaccines. *Biomaterials* (2011) **32**:926–932. doi:10.1016/j.biomaterials.2010.09.058

9. Heegaard PMH, Dedieu L, Johnson N, Le Potier MF, Mockey M, Mutinelli F, Vahlenkamp T, Vascellari M, Sørensen NS. Adjuvants and delivery systems in veterinary vaccinology: Current state and future developments. *Arch Virol* (2011) **156**:183–202. doi:10.1007/s00705-010-0863-1

10. Awate S, Babiuk LA, Mutwiri G. Mechanisms of action of adjuvants. *Front Immunol* (2013) **4**:114. doi:10.3389/fimmu.2013.00114

11. Garçon N, Di Pasquale A. From discovery to licensure, the Adjuvant System story. *Hum Vaccines Immunother* (2017) **13**:19–33. doi:10.1080/21645515.2016.1225635

12. Van Doorn E, Liu H, Huckriede A, Hak E. Safety and tolerability evaluation of the use of Montanide ISATM51 as vaccine adjuvant: A systematic review. *Hum Vaccines Immunother* (2016) **12**:159–169. doi:10.1080/21645515.2015.1071455

13. Giese M. *Introduction to molecular vaccinology*. Cham: Springer International Publishing (2016). doi:10.1007/978-3-319-25832-4

14. Rhee JH, Lee SE, Kim SY. Mucosal vaccine adjuvants update. *Clin Exp Vaccine Res* (2012) **1**:50. doi:10.7774/cevr.2012.1.1.50

15. Perales MA, Yuan J, Powel S, Gallardo HF, Rasalan TS, Gonzalez C, Manukian G, Wang J,

Zhang Y, Chapman PB, et al. Phase I/II study of GM-CSF DNA as an adjuvant for a multipeptide cancer vaccine in patients with advanced melanoma. *Mol Ther* (2008) **16**:2022–2029. doi:10.1038/mt.2008.196

16. Yoon HA, Aleyas AG, George JA, Seong OP, Young WH, John HL, Cho JG, Seong KE. Cytokine GM-CSF genetic adjuvant facilitates prophylactic DNA vaccine against pseudorabies virus through enhanced immune responses. *Microbiol Immunol* (2006) **50**:83–92. doi:10.1111/j.1348-0421.2006.tb03773.x

17. Christensen D. Vaccine adjuvants: Why and how. *Hum Vaccines Immunother* (2016) **12**:2709–2711. doi:10.1080/21645515.2016.1219003

18. He P, Zou Y, Hu Z. Advances in aluminum hydroxide-based adjuvant research and its mechanism. *Hum Vaccines Immunother* (2015) **11**:477–488. doi:10.1080/21645515.2014.1004026

19. HogenEsch H, O'Hagan DT, Fox CB. Optimizing the utilization of aluminum adjuvants in vaccines: you might just get what you want. *npj Vaccines* (2018) **3**:51. doi:10.1038/s41541-018-0089-x

20. Shardlow E, Mold M, Exley C. From Stock Bottle to Vaccine: Elucidating the Particle Size Distributions of Aluminum Adjuvants Using Dynamic Light Scattering. *Front Chem* (2017) **4**:48. doi:10.3389/fchem.2016.00048

21. HogenEsch H. Mechanism of immunopotentiation and safety of aluminum adjuvants. *Front Immunol* (2013) **3**:406. doi:10.3389/fimmu.2012.00406

22. Johnston CT, Wang SL, Hem SL. Measuring the surface area of aluminum hydroxide adjuvant. *J Pharm Sci* (2002) **91**:1702–1706. doi:10.1002/jps.10166

23. Burrell LS, White JL, Hem SL. Stability of aluminium-containing adjuvants during aging at room temperature. *Vaccine* (2000) **18**:2188–2192. doi:10.1016/S0264-410X(00)00031-1

24. Harrison WT. Some Observations on the Use of Alum Precipitated Diphtheria Toxoid. *Am J Public Health Nations Health* (1935) **25**:298–300. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18014174 [Accessed May 15, 2019]

25. Holt LB. *Developments in Diphtheria Prophylaxis.* William Heinemann-Medical Books Ltd. (1950). Available at: https://www.cabdirect.org/cabdirect/abstract/19502702965 [Accessed May 15, 2019]

26. Hutchison S, Benson RA, Gibson VB, Pollock AH, Garside P, Brewer JM. Antigen depot is not required for alum adjuvanticity. *FASEB J* (2011) **26**:1272–1279. doi:10.1096/fj.11-184556

27. Kool M, Fierens K, Lambrecht BN. Alum adjuvant: Some of the tricks of the oldest adjuvant. *J Med Microbiol* (2012) **61**:927–934. doi:10.1099/jmm.0.038943-0

28. Ghimire TR. The mechanisms of action of vaccines containing aluminum adjuvants: an in vitro vs in vivo paradigm. *Springerplus* (2015) **4**: doi:10.1186/s40064-015-0972-0

29. Ghimire TR. The mechanisms of action of vaccines containing aluminum adjuvants: an in vitro vs in vivo paradigm. *Springerplus* (2015) **4**:181. doi:10.1186/s40064-015-0972-0

30. Svensson A, Sandberg T, Siesjö P, Eriksson H. Sequestering of damage-associated molecular patterns (DAMPs): a possible mechanism affecting the immune-stimulating properties of aluminium adjuvants. *Immunol Res* (2017) **65**:1164–1175. doi:10.1007/s12026-017-8972-5

31. Goto N, Kato H, Maeyama JI, Shibano M, Saito T, Yamaguchi J, Yoshihara S. Local tissue irritating effects and adjuvant activities of calcium phosphate and aluminium hydroxide with different physical properties. *Vaccine* (1997) **15**:1364–1371. doi:10.1016/S0264-410X(97)00054-6

32. Kool M, Soullié T, van Nimwegen M, Willart MAM, Muskens F, Jung S, Hoogsteden HC, Hammad H, Lambrecht BN. Alum adjuvant boosts adaptive immunity by inducing uric acid and activating inflammatory dendritic cells. *J Exp Med* (2008) **205**:869–882. doi:10.1084/jem.20071087

33. Marichal T, Ohata K, Bedoret D, Mesnil C, Sabatel C, Kobiyama K, Lekeux P, Coban C, Akira S, Ishii KJ, et al. DNA released from dying host cells mediates aluminum adjuvant activity. *Nat Med* (2011) **17**:996–1002. doi:10.1038/nm.2403

34. Martinon F, Pétrilli V, Mayor A, Tardivel A, Tschopp J. Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* (2006) **440**:237–241. doi:10.1038/nature04516

35. Kool M, Petrilli V, De Smedt T, Rolaz A, Hammad H, van Nimwegen M, Bergen IM, Castillo R, Lambrecht BN, Tschopp J. Cutting Edge: Alum Adjuvant Stimulates Inflammatory Dendritic Cells through Activation of the NALP3 Inflammasome. *J Immunol* (2008) **181**:3755–3759. doi:10.4049/jimmunol.181.6.3755

36. Li H, Willingham SB, Ting JP-Y, Re F. Cutting Edge: Inflammasome Activation by Alum and Alum's Adjuvant Effect Are Mediated by NLRP3. *J Immunol* (2008) **181**:17–21. doi:10.4049/jimmunol.181.1.17

37. McKee AS, Munks MW, MacLeod MKL, Fleenor CJ, Van Rooijen N, Kappler JW, Marrack P. 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* (2009) **183**:4403–4414. doi:10.4049/jimmunol.0900164

38. Franchi L, Núñez G. The Nlrp3 inflammasome is critical for aluminum hydroxide-mediated IL-1beta secretion but dispensable for adjuvant activity. *Eur J Immunol* (2008) **38**:2085–2089. doi:10.1002/eji.200838549

39. Wang Y, Rahman D, Lehner T. A comparative study of stress-mediated immunological functions with the adjuvanticity of alum. *J Biol Chem* (2012) **287**:17152–17160. doi:10.1074/jbc.M112.347179

40. Oleszycka E, Moran HBT, Tynan GA, Hearnden CH, Coutts G, Campbell M, Allan SM, Scott CJ, Lavelle EC. IL-1α and inflammasome-independent IL-1β promote neutrophil infiltration following alum vaccination. *FEBS J* (2016) **283**:9–24. doi:10.1111/febs.13546

41. Flach TL, Ng G, Hari A, Desrosiers MD, Zhang P, Ward SM, Seamone ME, Vilaysane A, Mucsi AD, Fong Y, et al. Alum interaction with dendritic cell membrane lipids is essential for its adjuvanticity. *Nat Med* (2011) **17**:479–487. doi:10.1038/nm.2306

42. Wang XY, Yao X, Wan YM, Wang B, Xu JQ, Wen YM. Responses to multiple injections with alum alone compared to injections with alum adsorbed to proteins in mice. *Immunol Lett* (2013) **149**:88–92. doi:10.1016/j.imlet.2012.11.005

43. Güven E, Duus K, Laursen I, Højrup P, Houen G. Aluminum Hydroxide Adjuvant Differentially Activates the Three Complement Pathways with Major Involvement of the Alternative Pathway. *PLoS One* (2013) **8**:e74445. doi:10.1371/journal.pone.0074445

44. Oleszycka E, Lavelle EC. Immunomodulatory properties of the vaccine adjuvant alum.

Curr Opin Immunol (2014) **28**:1–5. doi:10.1016/j.coi.2013.12.007

45. Oleszycka E, McCluskey S, Sharp FA, Muñoz-Wolf N, Hams E, Gorman AL, Fallon PG, Lavelle EC. The vaccine adjuvant alum promotes IL-10 production that suppresses Th1 responses. *Eur J Immunol* (2018)1–40. doi:10.1002/eji.201747150

46. Flarend RE, Hem SL, White JL, Elmore D, Suckow MA, Rudy AC, Dandashli EA. In vivo absorption of aluminium-containing vaccine adjuvants using 26Al. *Vaccine* (1997) **15**:1314–1318. doi:10.1016/S0264-410X(97)00041-8

47. Masson JD, Crépeaux G, Authier FJ, Exley C, Gherardi RK. Critical analysis of reference studies on the toxicokinetics of aluminum-based adjuvants. *J Inorg Biochem* (2018) **181**:87–95. doi:10.1016/j.jinorgbio.2017.12.015

48. Khan Z, Combadière C, Authier FJ, Itier V, Lux F, Exley C, Mahrouf-Yorgov M, Decrouy X, Moretto P, Tillement O, et al. Slow CCL2-dependent translocation of biopersistent particles from muscle to brain. *BMC Med* (2013) **11**:99. doi:10.1186/1741-7015-11-99

49. Mold M, Shardlow E, Exley C. Insight into the cellular fate and toxicity of aluminium adjuvants used in clinically approved human vaccinations. *Sci Rep* (2016) **6**:31578. doi:10.1038/srep31578

50. Gherardi RK, Eidi H, Crépeaux G, Authier FJ, Cadusseau J. Biopersistence and brain translocation of aluminum adjuvants of vaccines. *Front Neurol* (2015) **6**:4. doi:10.3389/fneur.2015.00004

51. Crépeaux G, Eidi H, David MO, Tzavara E, Giros B, Exley C, Curmi PA, Shaw CA, Gherardi RK, Cadusseau J. Highly delayed systemic translocation of aluminum-based adjuvant in CD1 mice following intramuscular injections. *J Inorg Biochem* (2015) **152**:199–205. doi:10.1016/j.jinorgbio.2015.07.004

52. Guimarães LE, Baker B, Perricone C, Shoenfeld Y. Vaccines, adjuvants and autoimmunity. *Pharmacol Res* (2015) **100**:190–209. doi:10.1016/j.phrs.2015.08.003

53. 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* (2018) **34**:220–232. doi:10.1007/s12640-018-9879-1

54. Kumar V, Gill KD. Oxidative stress and mitochondrial dysfunction in aluminium neurotoxicity and its amelioration: A review. *Neurotoxicology* (2014) **41**:154–166. doi:10.1016/j.neuro.2014.02.004

55. Abbate C, Giorgianni C, Brecciaroli R, Tringali MA, D'Arrigo G. Spirometric function in non-smoking workers exposed to aluminum. *Am J Ind Med* (2003) **44**:400–404. doi:10.1002/ajim.10276

56. Klotz K, Weistenhöfer W, Neff F, Hartwig A, Van Thriel C, Drexler H. The health effects of aluminum exposure. *Dtsch Arztebl Int* (2017) **114**:653–659. doi:10.3238/arztebl.2017.0653

57. Darbre PD. Aluminium and the human breast. *Morphologie* (2016) **100**:65–74. doi:10.1016/j.morpho.2016.02.001

58. Mandriota SJ. A Case-control Study Adds a New Piece to the Aluminium/Breast Cancer Puzzle. *EBioMedicine* (2017) **22**:22–23. doi:10.1016/j.ebiom.2017.06.025

59. Afssaps. Risk assessment related to the use of aluminum in cosmetic products. (2011). Available at:

http://www.ansm.sante.fr/var/ansm_site/storage/original/application/424d34a9741c36907c95baa1ac838183.pdf [Accessed August 26, 2019]

60. Bezak-Mazur E, Widlak M, Ciupa T. Priority Data Needs for Aluminium. (2001). Available at: http://www.pjoes.com/pdf/10.4/263-267.pdf [Accessed August 26, 2019]

61. Crépeaux G, Eidi H, David MO, Baba-Amer Y, Tzavara E, Giros B, Authier FJ, Exley C, Shaw CA, Cadusseau J, et al. Non-linear dose-response of aluminium hydroxide adjuvant particles: Selective low dose neurotoxicity. *Toxicology* (2017) **375**:48–57. doi:10.1016/j.tox.2016.11.018

62. Lukiw WJ, Kruck TPA, Percy ME, Pogue AI, Alexandrov PN, Walsh WJ, Sharfman NM, Jaber VR, Zhao Y, Li W, et al. Aluminum in neurological disease – a 36 year multicenter study. *J Alzheimer's Dis Park* (2019) **8**: doi:10.4172/2161-0460.1000457

63. Cheignon C, Tomas M, Bonnefont-Rousselot D, Faller P, Hureau C, Collin F. Oxidative stress and the amyloid beta peptide in Alzheimer's disease. *Redox Biol* (2018) **14**:450–464. doi:10.1016/j.redox.2017.10.014

64. Simunkova M, Alwasel SH, Alhazza IM, Jomova K, Kollar V, Rusko M, Valko M. Management of oxidative stress and other pathologies in Alzheimer's disease. *Arch Toxicol* (2019) doi:10.1007/s00204-019-02538-y

65. Liang R. "Cross Talk Between Aluminum and Genetic Susceptibility and Epigenetic Modification in Alzheimer's Disease," in *Advances in Experimental Medicine and Biology*, 173–191. doi:10.1007/978-981-13-1370-7_10

66. Crapper DR, Krishnan SS, Dalton AJ. Brain aluminum distribution in Alzheimer's disease and experimental neurofibrillary degeneration. *Science (80-)* (1973) **180**:511–513. doi:10.1126/science.180.4085.511

67. Exley C, Mamutse G, Korchazhkina O, Pye E, Strekopytov S, Polwart A, Hawkins C. Elevated urinary excretion of aluminium and iron in multiple sclerosis. *Mult Scler* (2006) **12**:533–540. doi:10.1177/1352458506071323

68. Mold M, Chmielecka A, Rodriguez M, Thom F, Linhart C, King A, Exley C. Aluminium in Brain Tissue in Multiple Sclerosis. *Int J Environ Res Public Health* (2018) **15**:1777. doi:10.1016/j.jtemb.2017.11.012

69. Inan-Eroglu E, Ayaz A. Is aluminum exposure a risk factor for neurological disorders? *J Res Med Sci* (2018) **23**:51. doi:10.4103/jrms.JRMS_921_17

70. Altschuler E. Aluminium-containing antacids as a cause of idiopathic Parkinson's disease14. *Med Hypotheses* (1999) **53**:22–23. doi:10.1054/mehy.1997.0701

71. van der Laan JW, Gould S, Tanir JY. Safety of vaccine adjuvants: Focus on autoimmunity. *Vaccine* (2015) **33**:1507–1514. doi:10.1016/j.vaccine.2015.01.073

72. Hekman JP, Johnson JL, Kukekova A V. Transcriptome Analysis in Domesticated Species: Challenges and Strategies. *Bioinform Biol Insights* (2015) **9S4**:21–31. doi:10.4137/BBI.S29334

73. Hogenesch H, O'hagan DT, Fox CB. Optimizing the utilization of aluminum adjuvants in vaccines: you might just get what you want. (2018) **3**:51. doi:10.1038/s41541-018-0089-x

74. Ohlsson L, Exley C, Darabi A, Sandén E, Siesjö P, Eriksson H. Aluminium based adjuvants and their effects on mitochondria and lysosomes of phagocytosing cells. *J Inorg*

Biochem (2013) **128**:229–236. doi:10.1016/j.jinorgbio.2013.08.003

75. Harte C, Gorman AL, McCluskey S, Carty M, Bowie AG, Scott CJ, Meade KG, Lavelle EC. Alum activates the bovine NLRP3 inflammasome. *Front Immunol* (2017) **8**:1494. doi:10.3389/fimmu.2017.01494

76. Jara LJ, García-Collinot G, Medina G, Cruz-Dominguez M del P, Vera-Lastra O, Carranza-Muleiro RA, Saavedra MA. Severe manifestations of autoimmune syndrome induced by adjuvants (Shoenfeld's syndrome). *Immunol Res* (2017) **65**:8–16. doi:10.1007/s12026-016-8811-0

77. Gherardi RK, Authier FJ. Macrophagic myofasciitis: Characterization and pathophysiology. *Lupus* (2012) **21**:184–189. doi:10.1177/0961203311429557

78. Shoenfeld Y, Agmon-Levin N. "ASIA" - Autoimmune/inflammatory syndrome induced by adjuvants. *J Autoimmun* (2011) **36**:4–8. doi:10.1016/j.jaut.2010.07.003

79. Ameratunga R, Gillis D, Gold M, Linneberg A, Elwood JM. Evidence Refuting the Existence of Autoimmune/Autoinflammatory Syndrome Induced by Adjuvants (ASIA). *J Allergy Clin Immunol Pract* (2017) **5**:1551-1555.e1. doi:10.1016/j.jaip.2017.06.033

80. Watad A, Quaresma M, Bragazzi NL, Cervera R, Tervaert JWC, Amital H, Shoenfeld Y. The autoimmune/inflammatory syndrome induced by adjuvants (ASIA)/Shoenfeld's syndrome: descriptive analysis of 300 patients from the international ASIA syndrome registry. *Clin Rheumatol* (2018) **37**:483–493. doi:10.1007/s10067-017-3748-9

81. Ruiz JT, Luján L, Blank M, Shoenfeld Y. Adjuvants- and vaccines-induced autoimmunity: animal models. *Immunol Res* (2017) **65**:55–65. doi:10.1007/s12026-016-8819-5

82. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* (2008) **18**:1509–1517. doi:10.1101/gr.079558.108

83. Eija Korpelainen, Tuimala J, Somervuo P, Huss M, Wong G. *RNA-Seq Data Analysis: A practical approach*. (2015).

84. Marguerat S, Bähler J. RNA-seq: From technology to biology. *Cell Mol Life Sci* (2010) **67**:569–579. doi:10.1007/s00018-009-0180-6

85. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. *Role Bioinforma Agric* (2014) **2012**:1–25. doi:10.1201/b16568

86. Lahens NF, Ricciotti E, Smirnova O, Toorens E, Kim EJ, Baruzzo G, Hayer KE, Ganguly T, Schug J, Grant GR. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genomics* (2017) **18**:602. doi:10.1186/s12864-017-4011-0

87. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma* (2015) **13**:278–289. doi:10.1016/j.gpb.2015.08.002

88. Buck D, Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, Au KF. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* (2017) **6**:100. doi:10.12688/f1000research.10571.2

89. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* (2016) **17**:13. doi:10.1186/s13059-016-0881-8

90. Alvarez Rojas CA, Scheerlinck JP, Ansell BRE, Hall RS, Gasser RB, Jex AR. Time-course study of the transcriptome of peripheral blood mononuclear cells (PBMCs) from sheep infected with Fasciola hepatica. *PLoS One* (2016) **11**:e0159194. doi:10.1371/journal.pone.0159194

91. Suárez-Vega A, Gutiérrez-Gil B, Klopp C, Robert-Granie C, Tosser-Klopp G, Arranz JJ. Characterization and Comparative Analysis of the Milk Transcriptome in Two Dairy Sheep Breeds using RNA Sequencing. *Sci Rep* (2015) **5**: doi:10.1038/srep18399

92. Acquadro A, Torello Marinoni D, Sartor C, Dini F, Macchio M, Botta R. Transcriptome characterization and expression profiling in chestnut cultivars resistant or susceptible to the gall wasp Dryocosmus kuriphilus. *Mol Genet Genomics* (2019) doi:10.1007/s00438-019-01607-2

93. Qi X, Ogden EL, Ehlenfeldt MK, Rowland LJ. Dataset of de novo assembly and functional annotation of the transcriptome of blueberry (Vaccinium spp.). *Data Br* (2019) **25**:104390. doi:10.1016/j.dib.2019.104390

94. Page TM, McDougall C, Diaz-Pulido G. De novo transcriptome assembly for four species of crustose coralline algae and analysis of unique orthologous genes. *Sci Rep* (2019) **9**:12611. doi:10.1038/s41598-019-48283-1

95. Kumar R, Srinivasan R, Rawdzah MA, Malini P. Mapping and identification of potential target genes from short–RNA seq for the control of Pieris rapae larvae. *Genomics* (2019) doi:10.1016/j.ygeno.2019.08.017

96. Sello CT, Liu C, Sun Y, Msuthwana P, Hu J, Sui Y, Chen S, Zhou Y, Lu H, Xu C, et al. De Novo Assembly and Comparative Transcriptome Profiling of Anser anser and Anser cygnoides Geese Species' Embryonic Skin Feather Follicles. *Genes (Basel)* (2019) **10**:351. doi:10.3390/genes10050351

97. Archibald AL, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, McEwan JC, Hutton Oddy V, Raadsma HW, Wade C, et al. The sheep genome reference sequence: a work in progress. *Anim Genet* (2010) **41**:449–53. doi:10.1111/j.1365-2052.2010.02100.x

98. Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W, et al. The sheep genome illuminates biology of the rumen and lipid metabolism. *Science (80-)* (2014) **344**:1168–1173. doi:10.1126/science.1252806

99. Zhang C, Wang G, Wang J, Ji Z, Liu Z, Pi X, Chen C. Characterization and Comparative Analyses of Muscle Transcriptomes in Dorper and Small-Tailed Han Sheep Using RNA-Seq Technique. *PLoS One* (2013) **8**:e72686. doi:10.1371/journal.pone.0072686

100. Zhao S. Alternative splicing, RNA-seq and drug discovery. *Drug Discov Today* (2019) **24**:1258–1267. doi:10.1016/j.drudis.2019.03.030

101. Cheng X, Zhao SG, Yue Y, Liu Z, Li HW, Wu JP. Comparative analysis of the liver tissue transcriptomes of mongolian and lanzhou fat-tailed sheep. *Genet Mol Res* (2016) **15**: doi:10.4238/gmr.15028572

102. Guo Y, Zhao S, Sheng Q, Samuels DC, Shyr Y. The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. *BMC Genomics* (2017) **18**:690. doi:10.1186/s12864-017-4022-x

103. Ma L, Li Z, Cai Y, Xu H, Yang R, Lan X. Genetic variants in fat- and short-tailed sheep from high-throughput RNA-sequencing data. *Anim Genet* (2018) **49**:483–487. doi:10.1111/age.12699

104. Suárez-Vega A, Gutiérrez-Gil B, Klopp C, Tosser-Klopp G, Arranz JJ. Variant discovery in the sheep milk transcriptome using RNA sequencing. *BMC Genomics* (2017) **18**:170. doi:10.1186/s12864-017-3581-1

105. Dhanoa JK, Sethi RS, Verma R, Arora JS, Mukhopadhyay CS. Long non-coding RNA: Its evolutionary relics and biological implications in mammals: A review. *J Anim Sci Technol* (2018) **60**:25. doi:10.1186/s40781-018-0183-7

106. Yue Y, Guo T, Yuan C, Liu J, Guo J, Feng R, Niu C, Sun X, Yang B. 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* (2016) **11**:e0156890. doi:10.1371/journal.pone.0156890

107. Sulayman A, Tian K, Huang X, Tian Y, Xu X, Fu X, Zhao B, Wu W, Wang D, Yasin A, et al. Genome-wide identification and characterization of long non-coding RNAs expressed during sheep fetal and postnatal hair follicle development. *Sci Rep* (2019) **9**:8501. doi:10.1038/s41598-019-44600-w

108. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl* (2014) **1**: doi:10.4172/jngsa.1000106

109. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep* (2018) **8**:10950. doi:10.1038/s41598-018-29325-6

110. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* (2015) **43**:e37–e37. doi:10.1093/nar/gku1341

111. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* (2009) **10**:57–63. doi:10.1038/nrg2484

112. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* (2010) **38**:e131. doi:10.1093/nar/gkq224

113. van Gurp TP, McIntyre LM, Verhoeven KJF. Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS One* (2013) **8**:e85583. doi:10.1371/journal.pone.0085583

114. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, MacInnis B, Kwiatkowski DP, Swerdlow HP, et al. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics* (2012) **13**:1. doi:10.1186/1471-2164-13-1

115. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* (2011) **12**:R18. doi:10.1186/gb-2011-12-2-r18

116. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* (2012) **40**:e72. doi:10.1093/nar/gks001

117. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* (2011) **12**:480. doi:10.1186/1471-2105-12-480

118. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS, et al. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol* (2014)

15:R86. doi:10.1186/gb-2014-15-6-r86

119. Sultan M, Amstislavskiy V, Risch T, Schuette M, Do¨kel S, Ralser M, Balzereit D, Lehrach H, Yaspo ML. Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* (2014) **15**: doi:10.1186/1471-2164-15-675

120. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* (2009) **4**:14. doi:10.1186/1745-6150-4-14

121. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. *Genome Res* (2011) **21**:2213–2223. doi:10.1101/gr.124321.111

122. Cerdà J, Manchado M. Advances in genomics for flatfish aquaculture. *Genes Nutr* (2013) **8**:5–17. doi:10.1007/s12263-012-0312-8

123. Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, Vedell PT, Barman P, Wang L, Weinshiboum R, et al. Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* (2016) **17**:58. doi:10.1186/s12859-016-0922-z

124. Baran-Gale J, Lisa Kurtz C, Erdos MR, Sison C, Young A, Fannin EE, Chines PS, Sethupathy P. Addressing bias in small RNA library preparation for sequencing: A new protocol recovers microRNAs that evade capture by current methods. *Front Genet* (2015) **6**:352. doi:10.3389/fgene.2015.00352

125. Richard Boland C. Non-coding RNA: It's Not Junk. *Dig Dis Sci* (2017) **62**:1107–1109. doi:10.1007/s10620-017-4506-1

126. Latgé G, Poulet C, Bours V, Josse C, Jerusalem G. Natural antisense transcripts: Molecular mechanisms and implications in breast cancers. *Int J Mol Sci* (2018) **19**: doi:10.3390/ijms19010123

127. Cech TR, Steitz JA. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell* (2014) **157**:77–94. doi:10.1016/j.cell.2014.03.008

128. Yu D, Ma X, Zuo Z, Wang H, Meng Y. Classification of transcription boundary-associated RNAs (TBARs) in animals and plants. *Front Genet* (2018) **9**:168. doi:10.3389/fgene.2018.00168

129. Ma X, Han N, Shao C, Meng Y. Transcriptome-wide discovery of PASRs (promoter-associated small RNAs) and TASRs (terminus-associated small RNAs) in Arabidopsis thaliana. *PLoS One* (2017) **12**:e0169212. doi:10.1371/journal.pone.0169212

130. Perez P, Jang SI, Alevizos I. Emerging landscape of non-coding RNAs in oral health and disease. *Oral Dis* (2014) **20**:226–235. doi:10.1111/odi.12142

131. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* (1993) **75**:843–854. doi:10.1016/0092-8674(93)90529-Y

132. Reinhart BJ, Slack FJ, Basson M, Pasquienelll AE, Bettlnger JC, Rougvle AE, Horvitz HR, Ruvkun G. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* (2000) **403**:901–906. doi:10.1038/35002607

133. Zhang S, Zhao F, Wei C, Sheng X, Ren H, Xu L, Lu J, Liu J, Zhang L, Du L. Identification and Characterization of the miRNA Transcriptome of Ovis aries. *PLoS One* (2013) **8**:e58905. doi:10.1371/journal.pone.0058905

134. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front Endocrinol (Lausanne)* (2018) **9**:402. doi:10.3389/fendo.2018.00402

135. Ha M, Kim VN. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* (2014) **15**:509–524. doi:10.1038/nrm3838

136. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? *Nat Rev Genet* (2008) **9**:102–114. doi:10.1038/nrg2290

137. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* (2009) **136**:215–233. doi:10.1016/j.cell.2009.01.002

138. Fang Z, Rajewsky N. The impact of miRNA target sites in coding sequences and in 3'UTRs. *PLoS One* (2011) **6**:e18067. doi:10.1371/journal.pone.0018067

139. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* (2009) **19**:92–105. doi:10.1101/gr.082701.108

140. Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: MicroRNAs can up-regulate translation. *Science (80-)* (2007) **318**:1931–1934. doi:10.1126/science.1149460

141. Ling H, Fabbri M, Calin GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nat Rev Drug Discov* (2013) **12**:847–865. doi:10.1038/nrd4140

142. Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E, et al. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu Rev Genet* (2015) **49**:213–242. doi:10.1146/annurev-genet-120213-092023

143. Guo JU, Agarwal V, Guo H, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol* (2014) **15**:409. doi:10.1186/s13059-014-0409-z

144. Xu S, Zhou L, Ponnusamy M, Zhang L, Dong Y, Zhang Y, Wang Q, Liu J, Wang K. A comprehensive review of circRNA: from purification and identification to disease marker potential. *PeerJ* (2018) **6**:e5503. doi:10.7717/peerj.5503

145. Bonizzato A, Gaffo E, Te Kronnie G, Bortoluzzi S. CircRNAs in hematopoiesis and hematological malignancies. *Blood Cancer J* (2016) **6**:e483. doi:10.1038/bcj.2016.81

146. Burd CE, Wang K, Sharpless NE, Sorrentino JA, Jeck WR, Marzluff WF, Slevin MK, Liu J. Circular RNAs are abundant, conserved, and associated with ALU repeats. *Rna* (2012) **19**:141–157. doi:10.1261/rna.035667.112

147. Sanger HL, Klotz G, Riesner D, Gross HJ, Kleinschmidt AK. Viroids are single stranded covalently closed circular RNA molecules existing as highly base paired rod like structures. *Proc Natl Acad Sci U S A* (1976) **73**:3852–3856. doi:10.1073/pnas.73.11.3852

148. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, Kinzler KW, Vogelstein B. Scrambled exons. *Cell* (1991) **64**:607–613. doi:10.1016/0092-8674(91)90244-S

149. Capel B, Swain A, Nicolis S, Hacker A, Walter M, Koopman P, Goodfellow P, Lovell-Badge R. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* (1993) **73**:1019–1030. doi:10.1016/0092-8674(93)90279-Y

150. Dubin RA, Kazmi MA, Ostrer H. Inverted repeats are necessary for circularization of the mouse testis Sry transcript. *Gene* (1995) **167**:245–248. doi:10.1016/0378-1119(95)00639-7

151. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* (2012) **7**:e30733. doi:10.1371/journal.pone.0030733

152. Ragan C, Goodall GJ, Shirokikh NE, Preiss T. Insights into the biogenesis and potential functions of exonic circular RNA. *Sci Rep* (2019) **9**:2048. doi:10.1038/s41598-018-37037-0

153. Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, Fatica A, Santini T, Andronache A, Wade M, et al. Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Mol Cell* (2017) **66**:22-37.e9. doi:10.1016/j.molcel.2017.02.017

154. Du WW, Yang W, Liu E, Yang Z, Dhaliwal P, Yang BB. Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. *Nucleic Acids Res* (2016) **44**:2846–2858. doi:10.1093/nar/gkw027

155. Li X, Li C, Wei J, Ni W, Xu Y, Yao R, Zhang M, Li H, Liu L, Dang H, et al. Comprehensive Expression Profiling Analysis of Pituitary Indicates that circRNA Participates in the Regulation of Sheep Estrus. *Genes (Basel)* (2019) **10**:90. doi:10.3390/genes10020090

156. Li C, Li X, Ma Q, Zhang X, Cao Y, Yao Y, You S, Wang D, Quan R, Hou X, et al. Genome-wide analysis of circular RNAs in prenatal and postnatal pituitary glands of sheep. *Sci Rep* (2017) **7**:16143. doi:10.1038/s41598-017-16344-y

157. Li C, Li X, Yao Y, Ma Q, Ni W, Zhang X, Cao Y, Hazi W, Wang D, Quan R, et al. Genome-wide analysis of circular RNAs in prenatal and postnatal muscle of sheep. *Oncotarget* (2017) **8**:97165–97177. doi:10.18632/oncotarget.21835

158. Cao Y, You S, Yao Y, Liu Z-J, Hazi W, Li C-Y, Zhang X-Y, Hou X-X, Wei J-C, Li X-Y, et al. Expression profiles of circular RNAs in sheep skeletal muscle. *Asian-Australasian J Anim Sci* (2018) **31**:1550–1557. doi:10.5713/ajas.17.0563

159. Noto JJ, Schmidt CA, Matera AG. Engineering and expressing circular RNAs via tRNA splicing. *RNA Biol* (2017) **14**:978–984. doi:10.1080/15476286.2017.1317911

160. Chen I, Chen CY, Chuang TJ. Biogenesis, identification, and function of exonic circular RNAs. *Wiley Interdiscip Rev RNA* (2015) **6**:563–579. doi:10.1002/wrna.1294

161. Holdt LM, Kohlmaier A, Teupser D. Molecular roles and function of circular RNAs in eukaryotic cells. *Cell Mol Life Sci* (2018) **75**:1071–1098. doi:10.1007/s00018-017-2688-5

162. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. CircRNA Biogenesis competes with Pre-mRNA splicing. *Mol Cell* (2014) **56**:55–66. doi:10.1016/j.molcel.2014.08.019

163. Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, Zhu S, Yang L, Chen LL. Circular Intronic Long Noncoding RNAs. *Mol Cell* (2013) **51**:792–806. doi:10.1016/j.molcel.2013.08.017

164. Conn VM, Hugouvieux V, Nayak A, Conos SA, Capovilla G, Cildir G, Jourdain A, Tergaonkar V, Schmid M, Zubieta C, et al. A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. *Nat Plants* (2017) **3**:17053.

doi:10.1038/nplants.2017.53

165. Zhang L, Li Y, Liu W, Li H, Zhu Z. Analysis of the complex interaction of CDR1as-miRNA-protein and detection of its novel role in melanoma. *Oncol Lett* (2018) **16**:1219–1225. doi:10.3892/ol.2018.8700

166. Han C, Seebacher NA, Hornicek FJ, Kan Q, Duan Z. Regulation of microRNAs function by circular RNAs in human cancer. *Oncotarget* (2017) **8**:64622–64637. doi:10.18632/oncotarget.19930

167. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. Natural RNA circles function as efficient microRNA sponges. *Nature* (2013) **495**:384–388. doi:10.1038/nature11993

168. Qu S, Liu Z, Yang X, Zhou J, Yu H, Zhang R, Li H. The emerging functions and roles of circular RNAs in cancer. *Cancer Lett* (2018) **414**:301–309. doi:10.1016/j.canlet.2017.11.022

169. Wilson AJ, Mellor PS. Bluetongue in Europe: Past, present and future. *Philos Trans R Soc B Biol Sci* (2009) **364**:2669–2681. doi:10.1098/rstb.2009.0091

170. Blencowe BJ, Ahmad S, Lee LJ. Current-generation high-throughput sequencing: Deepening insights into mammalian transcriptomes. *Genes Dev* (2009) **23**:1379–1386. doi:10.1101/gad.1788009

171. Sun Z, Nair A, Chen X, Prodduturi N, Wang J, Kocher JP. UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. *Sci Rep* (2017) **7**:14196. doi:10.1038/s41598-017-14595-3

172. Feng M, Dang N, Bai Y, Wei H, Meng L, Wang K, Zhao Z, Chen Y, Gao F, Chen Z, et al. Differential expression profiles of long non-coding RNAs during the mouse pronuclear stage under normal gravity and simulated microgravity. *Mol Med Rep* (2019) **19**:155–164. doi:10.3892/mmr.2018.9675

173. Brazão TF, Johnson JS, Müller J, Heger A, Ponting CP, Tybulewicz VLJ. Long noncoding RNAs in B-cell development and activation. *Blood* (2016) **128**:e10–e19. doi:10.1182/blood-2015-11-680843

174. Nicolet BP, Engels S, Aglialoro F, Van Den Akker E, Von Lindern M, Wolkers MC. Circular RNA expression in human hematopoietic cells is widespread and cell-type specific. *Nucleic Acids Res* (2018) **46**:8168–8180. doi:10.1093/nar/gky721

175. Li L, Zheng YC, Kayani MR, Xu W, Wang GQ, Sun P, Ao N, Zhang LN, Gu ZQ, Wu LC, et al. Comprehensive analysis of circRNA expression profiles in humans by RAISE. *Int J Oncol* (2017) **51**:1625–1638. doi:10.3892/ijo.2017.4162

176. Cooper DA, Cortés-López M, Miura P. Genome-wide circRNA profiling from RNA-seq data. *Methods Mol Biol* (2018) **1724**:27–41. doi:10.1007/978-1-4939-7562-4_3

177. Fang N, Akinci-Tolun R. Depletion of ribosomal RNA Sequences unit 7.27 from single-cellRNA-sequencing library. *Curr Protoc Mol Biol* (2016) **2016**: doi:10.1002/cpmb.11

178. 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* (2018) **8**: doi:10.1038/s41598-018-23226-4

179. Bush SJ, McCulloch MEB, Summers KM, Hume DA, Clark EL. Integration of quantitated expression estimates from polyA-selected and rRNA-depleted RNA-seq libraries. *BMC*

Bioinformatics (2017) **18**: doi:10.1186/s12859-017-1714-9

180. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol* (2015) **16**: doi:10.1186/s13059-015-0697-y

181. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics* (2014) **30**:301–304. doi:10.1093/bioinformatics/btt688

182. Babraham Bioinformatics. FASTQC, a quality control tool for the high throughput sequence data. *Http://WwwBioinformaticsBbsrcAcUk/Projects/Fastq/* (2012) Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

183. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. in *Bioinformatics* (Oxford University Press), i884–i890. doi:10.1093/bioinformatics/bty560

184. Shirley M. Simple FASTQ quality assessment using Python. Available at: https://github.com/mdshw5/fastqp [Accessed October 29, 2019]

185. FASTX-Toolkit. Available at: http://hannonlab.cshl.edu/fastx_toolkit/ [Accessed October 28, 2019]

186. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* (2015) **31**:166–169. doi:10.1093/bioinformatics/btu638

187. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods* (2013) **63**:41–49. doi:10.1016/j.ymeth.2013.06.027

188. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (2016) **32**:3047–3048. doi:10.1093/bioinformatics/btw354

189. Patel RK, Jain M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One* (2012) **7**: doi:10.1371/journal.pone.0030619

190. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* (2011) **27**:863–864. doi:10.1093/bioinformatics/btr026

191. Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* (2012) **28**:1530–1532. doi:10.1093/bioinformatics/bts196

192. Wang L, Wang S, Li W. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* (2012) **28**:2184–2185. doi:10.1093/bioinformatics/bts356

193. Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* (2010) **11**: doi:10.1186/1471-2105-11-485

194. Krueger F. TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. Available at: https://github.com/FelixKrueger/TrimGalore [Accessed October 29, 2019]

195. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One* (2013) **8**:e85024.

doi:10.1371/journal.pone.0085024

196. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* (2016) **17**: doi:10.1186/s12859-016-0956-2

197. Bushnell B. BBMap. (2019) Available at: https://sourceforge.net/projects/bbmap/ [Accessed October 30, 2019]

198. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* (2011) **17**:10. doi:10.14806/ej.17.1.200

199. Vince Buffalo. scythe: A 3'-end adapter contaminant trimmer. (2014) Available at: https://github.com/vsbuffalo/scythe [Accessed October 30, 2019]

200. Le HS, Schulz MH, Mccauley BM, Hinman VF, Bar-Joseph Z. Probabilistic error correction for RNA sequencing. *Nucleic Acids Res* (2013) **41**: doi:10.1093/nar/gkt215

201. Schmieder R, Lim YW, Rohwer F, Edwards R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* (2010) **11**:341. doi:10.1186/1471-2105-11-341

202. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) **30**:2114–2120. doi:10.1093/bioinformatics/btu170

203. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* (2011) **8**:469–77. doi:10.1038/nmeth.1613

204. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* (2009) **10**: doi:10.1186/gb-2009-10-3-r25

205. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) **25**:1754–1760. doi:10.1093/bioinformatics/btp324

206. Novocraft. Available at: http://www.novocraft.com/ [Accessed October 31, 2019]

207. Rumble SM, Lacroute P, Dalca A V., Fiume M, Sidow A, Brudno M. SHRiMP: Accurate mapping of short color-space reads. *PLoS Comput Biol* (2009) **5**: doi:10.1371/journal.pcbi.1000386

208. Li R, Li Y, Kristiansen K, Wang J. SOAP: Short oligonucleotide alignment program. *Bioinformatics* (2008) **24**:713–714. doi:10.1093/bioinformatics/btn025

209. Liao Y, Smyth GK, Shi W. The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* (2013) **41**: doi:10.1093/nar/gkt214

210. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* (2012) **9**:357–359. doi:10.1038/nmeth.1923

211. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: Fast, accurate and versatile alignment by filtration. *Nat Methods* (2012) **9**:1185–1188. doi:10.1038/nmeth.2221

212. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. "GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality," in *Methods in Molecular Biology* (Humana Press Inc.), 283–334. doi:10.1007/978-1-4939-3578-9_15

213. Wu TD, Watanabe CK. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* (2005) **21**:1859–1875. doi:10.1093/bioinformatics/bti310

214. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* (2010) **26**:873–881. doi:10.1093/bioinformatics/btq057

215. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* (2019) **37**:907–915. doi:10.1038/s41587-019-0201-4

216. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* (2010) **38**: doi:10.1093/nar/gkq622

217. De Bona F, Ossowski S, Schneeberger K, Rätsch G. Optimal spliced alignments of short sequence reads. *Bioinformatics* (2008) **24**:i174–i180. doi:10.1093/bioinformatics/btn300

218. Weese D, Holtgrewe M, Reinert K. RazerS 3: Faster, fully sensitive read mapping. *Bioinformatics* (2012) **28**:2592–2599. doi:10.1093/bioinformatics/bts505

219. Wilson GW, Stein LD. RNASequel: Accurate and repeat tolerant realignment of RNA-seq reads. *Nucleic Acids Res* (2015) **43**:e122–e122. doi:10.1093/nar/gkv594

220. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* (2011) **27**:2518–2528. doi:10.1093/bioinformatics/btr427

221. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) **29**:15–21. doi:10.1093/bioinformatics/bts635

222. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* (2013) **14**:R36. doi:10.1186/gb-2013-14-4-r36

223. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* (2013) **10**:1185–91. doi:10.1038/nmeth.2722

224. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods* (2015) **12**:357–360. doi:10.1038/nmeth.3317

225. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* (2016) **34**:525–527. doi:10.1038/nbt.3519

226. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* (2014) **30**:923–930. doi:10.1093/bioinformatics/btt656

227. Zhang C, Zhang B, Lin LL, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* (2017) **18**: doi:10.1186/s12864-017-4002-1

228. Rossell D, Stephan-Otto Attolini C, Kroiss M, Stöcker A. Quantifying Alternative Splicing From Paired-End Rna-Sequencing Data. *Ann Appl Stat* (2014) **8**:309–330. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/24795787%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4005600 [Accessed November 5, 2019]

229. Novartis. EQP-QM: Unix based RNA-seq quantification module. Available at: https://github.com/novartis/EQP-QM [Accessed November 6, 2019]

230. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* (2013) **10**:71–73. doi:10.1038/nmeth.2251

231. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *Bioinforma Impact Accurate Quantif Proteomic Genet Anal Res* (2014) **12**:41–74. doi:10.1201/b16589

232. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* (2017) **14**:417–419. doi:10.1038/nmeth.4197

233. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* (2008) **5**:621–8. doi:10.1038/nmeth.1226

234. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* (2014) **15**:R29. doi:10.1186/gb-2014-15-2-r29

235. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* (2012) **131**:281–285. doi:10.1007/s12064-012-0162-3

236. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* (2010) **11**:94. doi:10.1186/1471-2105-11-94

237. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* (2011) **12**:480. doi:10.1186/1471-2105-12-480

238. Li X, Brock GN, Rouchka EC, Cooper NGF, Wu D, OToole TE, Gill RS, Eteleeb AM, O'Brien L, Rai SN. A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. *PLoS One* (2017) **12**: doi:10.1371/journal.pone.0176185

239. Anders S, Huber W. Differential expression analysis for sequence count data. (2010). doi:10.1186/gb-2010-11-10-r106

240. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform* (2018) **19**:776–792. doi:10.1093/bib/bbx008

241. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* (2013) **14**:671–83. doi:10.1093/bib/bbs046

242. Zyprych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, Siatkowski I. The Impact of Normalization Methods on RNA-Seq Data Analysis. *Biomed Res Int* (2015) **2015**: doi:10.1155/2015/621690

243. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining

group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* (2016) **17**:29–39. doi:10.1093/biostatistics/kxv027

244. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* (2014) **42**:e161-. doi:10.1093/nar/gku864

245. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* (2007) **8**:118–127. doi:10.1093/biostatistics/kxj037

246. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, de Andrade M, Kocher J-PA, Eckel-Passow JE. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* (2013) **29**:2877–83. doi:10.1093/bioinformatics/btt480

247. Oytam Y, Sobhanmanesh F, Duesing K, Bowden JC, Osmond-McLeod M, Ross J. Risk-conscious correction of batch effects: Maximising information extraction from high-throughput genomic datasets. *BMC Bioinformatics* (2016) **17**:332. doi:10.1186/s12859-016-1212-5

248. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* (2015) **43**:e47. doi:10.1093/nar/gkv007

249. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* (2014) **32**:896–902. doi:10.1038/nbt.2931

250. Leek JT, Johnson EW, Parker HS, Fertig EJ, Jaffe AE, Storey JD, Zhang Y, Torres LC. sva: Surrogate Variable Analysis. (2018) doi:10.18129/B9.bioc.sva

251. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* (2012) **28**:882–883. doi:10.1093/bioinformatics/bts034

252. Abbas-Aghababazadeh F, Li Q, Fridley BL. Comparison of normalization approaches for gene expression studies completed with highthroughput sequencing. *PLoS One* (2018) **13**:e0206312. doi:10.1371/journal.pone.0206312

253. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* (2012) **40**:4288–4297. doi:10.1093/nar/gks042

254. Love MI, Huber W, Anders S, Lönnstedt I, Speed T, Robinson M, Smyth G, McCarthy D, Chen Y, Smyth G, et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* (2014) **15**:550. doi:10.1186/s13059-014-0550-8

255. Yang W, Rosenstiel PC, Schulenburg H. ABSSeq: A new RNA-Seq analysis method based on modelling absolute expression differences. *BMC Genomics* (2016) **17**: doi:10.1186/s12864-016-2848-2

256. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol* (2015) **33**:243–246. doi:10.1038/nbt.3172

257. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* (2016) **11**:1650–1667. doi:10.1038/nprot.2016.095

258. Hardcastle TJ, Kelly KA. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* (2010) **11**:422. doi:10.1186/1471-2105-11-422

259. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* (2012) **28**:1721–1728. doi:10.1093/bioinformatics/bts260

260. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* (2013) **31**:46–53. doi:10.1038/nbt.2450

261. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* (2012) **22**:2008–2017. doi:10.1101/gr.133744.111

262. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* (2013) **29**:1035–1043. doi:10.1093/bioinformatics/btt087

263. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (2009) **26**:139–140. doi:10.1093/bioinformatics/btp616

264. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res* (2010) **38**: doi:10.1093/nar/gkq670

265. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat Appl Genet Mol Biol* (2011) **10**:1–28. doi:10.2202/1544-6115.1637

266. Tarazona S, Furió-Tarí P, Turrà D, Di Pietro A, Nueda MJ, Ferrer A, Conesa A. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* (2015) **43**:e140. doi:10.1093/nar/gkv711

267. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* (2013) **22**:519–536. doi:10.1177/0962280211428386

268. Auer PL, Doerge RW. A two-stage poisson model for testing RNA-Seq data. *Stat Appl Genet Mol Biol* (2011) **10**: doi:10.2202/1544-6115.1627

269. Esnaola M, Puig P, Gonzalez D, Castelo R, Gonzalez JR. A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. *BMC Bioinformatics* (2013) **14**:254. doi:10.1186/1471-2105-14-254

270. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot* (2012) **99**:248–56. doi:10.3732/ajb.1100340

271. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* (2012) **13**:484. doi:10.1186/1471-2164-13-484

272. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform* (2013) doi:10.1093/bib/bbt086

273. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* (2017) **12**:e0190152. doi:10.1371/journal.pone.0190152

274. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* (2009) **4**:44–57. doi:10.1038/nprot.2008.211

275. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* (2003) **13**:2129–2141. doi:10.1101/gr.772403

276. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* (2019) **47**:W191–W198. doi:10.1093/nar/gkz369

277. Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol* (2015) **16**:186. doi:10.1186/s13059-015-0761-7

278. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* (2019) **14**:482–517. doi:10.1038/s41596-018-0103-9

279. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* (2008) **9**:559. doi:10.1186/1471-2105-9-559

280. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: Mutual information, correlation, and model based indices. *BMC Bioinformatics* (2012) **13**: doi:10.1186/1471-2105-13-328

281. Barabási AL, Bonabeau E. Scale-free networks. *Sci Am* (2003) **288**:60–69. doi:10.1038/scientificamerican0503-60

282. Kost MA, Perales HR, Wijeratne S, Wijeratne AJ, Stockinger E, Mercer KL. Differentiated transcriptional signatures in the maize landraces of Chiapas, Mexico. *BMC Genomics* (2017) **18**: doi:10.1186/s12864-017-4005-y

283. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* (2015) **33**:290–295. doi:10.1038/nbt.3122

284. Bortolomeazzi M, Gaffo E, Bortoluzzi S. A survey of software tools for microRNA discovery and characterization using RNA-seq. *Brief Bioinform* (2019) **20**:918–930. doi:10.1093/bib/bbx148

285. Ni WJ, Leng XM. Dynamic miRNA-mRNA paradigms: New faces of miRNAs. *Biochem Biophys Reports* (2015) **4**:337–341. doi:10.1016/j.bbrep.2015.10.011

286. Videm P, Rose D, Costa F, Backofen R. BlockClust: Efficient clustering and classification of non-coding RNAs from short read RNA-Seq profiles. *Lect Notes Informatics (LNI), Proc - Ser Gesellschaft fur Inform* (2014) **P-235**:12–22. doi:10.1093/bioinformatics/btu270

287. Leung YY, Ryvkin P, Ungar LH, Gregory BD, Wang LS. CoRAL: Predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Res* (2013) **41**: doi:10.1093/nar/gkt426

288. Langenberger D, Pundhir S, Ekstrøm CT, Stadler PF, Hoffmann S, Gorodkin J.

DeepBlockAlign: A tool for aligning RNA-seq profiles of read block patterns. *Bioinformatics* (2012) **28**:17–24. doi:10.1093/bioinformatics/btr598

289. Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM. miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* (2009) **37**:W68–W76. doi:10.1093/nar/gkp347

290. Friedländer MR, MacKowiak SD, Li N, Chen W, Rajewsky N. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* (2012) **40**:37–52. doi:10.1093/nar/gkr688

291. Kuang Z, Wang Y, Li L, Yang X. MiRDeep-P2: Accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics* (2019) **35**:2521–2522. doi:10.1093/bioinformatics/bty972

292. Hansen TB, Venø MT, Kjems J, Damgaard CK. MiRdentify: High stringency miRNA predictor identifies several novel animal miRNAs. *Nucleic Acids Res* (2014) **42**: doi:10.1093/nar/gku598

293. Leclercq M, Diallo AB, Blanchette M. Computational prediction of the localization of microRNAs within their pre-miRNA. *Nucleic Acids Res* (2013) **41**:7200–7211. doi:10.1093/nar/gkt466

294. Jha A, Shankar R. miReader: Discovering Novel miRNAs in Species without Sequenced Genome. *PLoS One* (2013) **8**:e66857. doi:10.1371/journal.pone.0066857

295. Mathelier A, Carbone A, Hofacker I. MIReNA: Finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. in *Bioinformatics* (Oxford University Press), 2226–2234. doi:10.1093/bioinformatics/btq329

296. Wang WC, Lin FM, Chang WC, Lin KY, Huang H Da, Lin NS. MiRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* (2009) **10**:328. doi:10.1186/1471-2105-10-328

297. Mapleson D, Moxon S, Dalmay T, Moulton V. MirPlex: A Tool for Identifying miRNAs in High-Throughput sRNA Datasets Without a Genome. *J Exp Zool Part B Mol Dev Evol* (2013) **320**:47–56. doi:10.1002/jez.b.22483

298. Qian K, Auvinen E, Greco D, Auvinen P. MiRSeqNovel: An R based workflow for analyzing miRNA sequencing data. *Mol Cell Probes* (2012) **26**:208–211. doi:10.1016/j.mcp.2012.05.002

299. Aparicio-Puerta E, Lebrón R, Rueda A, Gómez-Martín C, Giannoukakos S, Jaspez D, Medina JM, Zubkovic A, Jurak I, Fromm B, et al. sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res* (2019) **47**:W530–W535. doi:10.1093/nar/gkz415

300. Liu B, Li J, Cairns MJ. Identifying miRNAs, targets and functions. *Brief Bioinform* (2014) **15**:1–19. doi:10.1093/bib/bbs075

301. Riffo-Campos ÁL, Riquelme I, Brebi-Mieville P. Tools for sequence-based miRNA target prediction: What to choose? *Int J Mol Sci* (2016) **17**: doi:10.3390/ijms17121987

302. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* (2007) **8**:69. doi:10.1186/1471-2105-8-69

303. Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. Functional microRNA

targets in protein coding sequences. *Bioinformatics* (2012) **28**:771–776. doi:10.1093/bioinformatics/bts043

304. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res* (2013) **41**: doi:10.1093/nar/gkt393

305. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. *Genome Biol* (2003) **5**:R1. doi:10.1186/gb-2003-5-1-r1

306. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. (2010). doi:10.1186/gb-2010-11-8-r90

307. Hsu JB, Chiu CM, Hsu S Da, Huang WY, Chien CH, Lee TY, Huang H Da. MiRTar: An integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics* (2011) **12**:300. doi:10.1186/1471-2105-12-300

308. Dai X, Zhuang Z, Zhao PX. PsRNATarget: A plant small RNA target analysis server (2017 release). *Nucleic Acids Res* (2018) **46**:W49–W54. doi:10.1093/nar/gky316

309. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, Da Piedade I, Gunsalus KC, Stoffel M, et al. Combinatorial microRNA target predictions. *Nat Genet* (2005) **37**:495–500. doi:10.1038/ng1536

310. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. *Nat Genet* (2007) **39**:1278–1284. doi:10.1038/ng2135

311. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I. A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell* (2006) **126**:1203–1217. doi:10.1016/j.cell.2006.07.031

312. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. *Rna* (2004) **10**:1507–1517. doi:10.1261/rna.5248604

313. TargetS: Liu Lab (UT Medical School). Available at: http://liubioinfolab.org/targetS/mirna.html [Accessed November 21, 2019]

314. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* (2015) **4**:101–112. doi:10.7554/eLife.05005

315. Oliveira AC, Bovolenta LA, Nachtigall PG, Herkenhoff ME, Lemke N, Pinhal D. Combining results from distinct microRNA target prediction tools enhances the performance of analyses. *Front Genet* (2017) **8**: doi:10.3389/fgene.2017.00059

316. Su N, Qian M, Deng M. Integrative Approaches for microRNA Target Prediction: Combining Sequence Information and the Paired mRNA and miRNA Expression Profiles. *Curr Bioinform* (2013) **8**:37–45. doi:10.2174/1574893611308010008

317. Ozdemir B, Abd-Almageed W, Roessler S, Wang XW. iSubgraph: Integrative genomics for subgroup discovery in hepatocellular carcinoma using graph mining and mixture models. *PLoS One* (2013) **8**:e78624. doi:10.1371/journal.pone.0078624

318. Seo J, Jin D, Choi CH, Lee H. Integration of MicroRNA, mRNA, and protein expression data for the identification of cancer-related MicroRNAs. *PLoS One* (2017) **12**: doi:10.1371/journal.pone.0168412

319. Cloonan N. Re-thinking miRNA-mRNA interactions: Intertwining issues confound target discovery. *BioEssays* (2015) **37**:379–388. doi:10.1002/bies.201400191

320. Tang D, Chen Y, He H, Huang J, Chen W, Peng W, Lu Q, Dai Y. Integrated analysis of mRNA, microRNA and protein in systemic lupus erythematosus-specific induced pluripotent stem cells from urine. *BMC Genomics* (2016) **17**:488. doi:10.1186/s12864-016-2809-9

321. Yoon JH, Abdelmohsen K, Gorospe M. Functional interactions among microRNAs and long noncoding RNAs. *Semin Cell Dev Biol* (2014) **34**:9–14. doi:10.1016/j.semcdb.2014.05.015

322. Gómez-Martín C, Lebrón R, Rueda A, Oliver JL, Hackenberg M. "sRNAtoolboxVM: Small RNA analysis in a virtual machine," in *Methods in Molecular Biology* (Humana Press Inc.), 149–174. doi:10.1007/978-1-4939-6866-4_12

323. Rueda A, Barturen G, Lebrón R, Gómez-Martín C, Alganza Á, Oliver JL, Hackenberg M. SRNAtoolbox: An integrated collection of small RNA research tools. *Nucleic Acids Res* (2015) **43**:W467–W473. doi:10.1093/nar/gkv555

324. Kozomara A, Griffiths-Jones S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* (2014) **42**:D68–D73. doi:10.1093/nar/gkt1181

325. You X, Conrad TO. Acfs: Accurate circRNA identification and quantification from RNA-Seq data. *Sci Rep* (2016) **6**:38820. doi:10.1038/srep38820

326. Zhang XO, Dong R, Zhang Y, Zhang JL, Luo Z, Zhang J, Chen LL, Yang L. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res* (2016) **26**:1277–1287. doi:10.1101/gr.202895.115

327. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. *Brief Bioinform* (2018) **19**:803–810. doi:10.1093/bib/bbx014

328. Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* (2016) **32**:1094–1096. doi:10.1093/bioinformatics/btv656

329. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* (2013) **495**:333–338. doi:10.1038/nature11928

330. Szabo L, Morey R, Palpant NJ, Wang PL, Afari N, Jiang C, Parast MM, Murry CE, Laurent LC, Salzman J. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol* (2015) **16**:126. doi:10.1186/s13059-015-0690-5

331. Huang Y, Zeng Z, Liu J, MacLeod JN, Singh D, Savich GL, Mieczkowski P, Chiang DY, Coleman SJ, Wang K, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* (2010) **38**:e178–e178. doi:10.1093/nar/gkq622

332. Chuang TJ, Wu CS, Chen CY, Hung LY, Chiang TW, Yang MY. NCLscan: Accurate identification of non-co-linear transcripts (fusion, trans-splicing and circular RNA) with a good balance between sensitivity and precision. *Nucleic Acids Res* (2016) **44**: doi:10.1093/nar/gkv1013

333. Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt LM, Teupser D, Hackermüller J, et al. A multi-split mapping algorithm for circular RNA,

splicing, trans-splicing and fusion detection. *Genome Biol* (2014) **15**:R34. doi:10.1186/gb-2014-15-2-r34

334. Kim D, Salzberg SL. TopHat-Fusion: An algorithm for discovery of novel fusion transcripts. *Genome Biol* (2011) **12**:R72. doi:10.1186/gb-2011-12-8-r72

335. Song X, Zhang N, Han P, Moon BS, Lai RK, Wang K, Lu W. Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res* (2016) **44**: doi:10.1093/nar/gkw075

336. Zeng X, Lin W, Guo M, Zou Q. A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput Biol* (2017) **13**: doi:10.1371/journal.pcbi.1005420

337. Hansen TB, Venø MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. *Nucleic Acids Res* (2015) **44**: doi:10.1093/nar/gkv1458

338. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) **29**:15–21. doi:10.1093/bioinformatics/bts635

339. Hanan M, Soreq H, Kadener S. CircRNAs in the brain. *RNA Biol* (2017) **14**:1028–1034. doi:10.1080/15476286.2016.1255398

340. Dong R, Ma XK, Li GW, Yang L. CIRCpedia v2: An Updated Database for Comprehensive Circular RNA Annotation and Expression Comparison. *Genomics, Proteomics Bioinforma* (2018) **16**:226–233. doi:10.1016/j.gpb.2018.08.001

341. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* (2019) **47**:D853–D858. doi:10.1093/nar/gky1095

342. Pan X, Wenzel A, Jensen LJ, Gorodkin J. Genome-wide identification of clusters of predicted microRNA binding sites as microRNA sponge candidates. *PLoS One* (2018) **13**:e0202369. doi:10.1371/journal.pone.0202369

343. Kozomara A, Birgaoanu M, Griffiths-Jones S. MiRBase: From microRNA sequences to function. *Nucleic Acids Res* (2019) **47**:D155–D162. doi:10.1093/nar/gky1141

344. Alkan F, Wenzel A, Palasca O, Kerpedjiev P, Rudebeck AF, Stadler PF, Hofacker IL, Gorodkin J. RIsearch2: Suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. *Nucleic Acids Res* (2017) **45**: doi:10.1093/nar/gkw1325

345. Gherardi RK, Coquet M, Chérin P, Authier FJ, Laforêt P, Bélec L, Figarella-Branger D, Mussini JM, Pellissier JF, Fardeau M. Macrophagic myofasciitis: An emerging entity. *Lancet* (1998) **352**:347–352. doi:10.1016/S0140-6736(98)02326-5

346. Shi W, Kou Y, Xiao J, Zhang L, Gao F, Kong W, Su W, Jiang C, Zhang Y. Comparison of immunogenicity, efficacy and transcriptome changes of inactivated rabies virus vaccine with different adjuvants. *Vaccine* (2018) **36**:5020–5029. doi:10.1016/j.vaccine.2018.07.006

347. Gerdts V, Wilson HL, Meurens F, Van den Hurk S van DL, Wilson D, Walker S, Wheler C, Townsend H, Potter AA. Large animal models for vaccine development and testing. *ILAR J* (2015) **56**:53–62. doi:10.1093/ilar/ilv009

348. Andersen CL, Jensen JL, Ørntoft TF. 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*

(2004) **64**:5245–5250. doi:10.1158/0008-5472.CAN-04-0496

349. Vandesompele J, De Preter K, Pattyn I, Poppe B, Van Roy N, De Paepe A, Speleman R. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* (2002) **3**:research0034.1–0034.11. doi:10.1186/gb-2002-3-7-research0034

350. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* (2016) **5**:1438. doi:10.12688/f1000research.8987.2

351. Morris G, Puri BK, Frye RE. The putative role of environmental aluminium in the development of chronic neuropathology in adults and children. How strong is the evidence and what could be the mechanisms involved? *Metab Brain Dis* (2017) **32**:1335–1355. doi:10.1007/s11011-017-0077-2

352. Kinoshita T, Imamura R, Kushiyama H, Suda T. NLRP3 Mediates NF-κB activation and cytokine induction in microbially induced and sterile inflammation. *PLoS One* (2015) **10**:e0119179. doi:10.1371/journal.pone.0119179

353. Oeckinghaus A, Ghosh S. The NF-kappaB family of transcription factors and its regulation. *Cold Spring Harb Perspect Biol* (2009) **1**: doi:10.1101/cshperspect.a000034

354. Rider P, Carmi Y, Guttman O, Braiman A, Cohen I, Voronov E, White MR, Dinarello CA, Apte RN. IL-1α and IL-1β Recruit Different Myeloid Cells and Promote Different Stages of Sterile Inflammation. *J Immunol* (2011) **187**:4835–4843. doi:10.4049/jimmunol.1102048

355. Lasigliè D, Traggiai E, Federici S, Alessio M, Buoncompagni A, Accogli A, Chiesa S, Penco F, Martini A, Gattorno M. Role of IL-1 beta in the development of human TH17 cells: Lesson from NLPR3 mutated patients. *PLoS One* (2011) **6**: doi:10.1371/journal.pone.0020014

356. Belot MP, Castell AL, Le Fur S, Bougnères P. Dynamic demethylation of the IL2RA promoter during in vitro CD4+ T cell activation in association with IL2RA expression. *Epigenetics* (2018) **13**:459–472. doi:10.1080/15592294.2018.1469893

357. Alles VV, Bottazzi B, Peri G, Golay J, Introna M, Mantovani A. Inducible expression of PTX3, a new member of the pentraxin family, in human mononuclear phagocytes. *Blood* (1994) **84**:3483–3493. doi:10.1182/blood.v84.10.3483.bloodjournal84103483

358. Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat Appl Genet Mol Biol* (2005) **4**:Article17. doi:10.2202/1544-6115.1128

359. Varela-Martínez E, Abendaño N, Asín J, Sistiaga-Poveda M, Pérez MM, Reina R, de Andrés D, Luján L, Jugo BM. Molecular signature of Aluminum hydroxide adjuvant in ovine PBMCs by integrated mRNA and microRNA transcriptome sequencing. *Front Immunol* (2018) **9**:2406. doi:10.3389/FIMMU.2018.02406

360. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* (2004) **116**:281–297. doi:10.1016/S0092-8674(04)00045-5

361. Bao H, Kommadath A, Sun X, Meng Y, Arantes AS, Plastow GS, Guan LL, Stothard P. Expansion of ruminant-specific microRNAs shapes target gene expression divergence between ruminant and non-ruminant species. *BMC Genomics* (2013) **14**:609. doi:10.1186/1471-2164-14-609

362. Sun YM, Lin KY, Chen YQ. Diverse functions of miR-125 family in different cell contexts. *J Hematol Oncol* (2013) **6**:6. doi:10.1186/1756-8722-6-6

363. Gururajan M, Haga CL, Das S, Leu CM, Hodson D, Josson S, Turner M, Cooper MD. MicroRNA 125b inhibition of B cell differentiation in germinal centers. *Int Immunol* (2010) **22**:583–592. doi:10.1093/intimm/dxq042

364. Simpson LJ, Patel S, Bhakta NR, Choy DF, Brightbill HD, Ren X, Wang Y, Pua HH, Baumjohann D, Montoya MM, et al. A microRNA upregulated in asthma airway T cells promotes T H 2 cytokine production. *Nat Immunol* (2014) **15**:1162–1170. doi:10.1038/ni.3026

365. Corpa JM, Pérez V, García Marín JF. Differences in the immune responses in lambs and kids vaccinated against paratuberculosis, according to the age of vaccination. *Vet Microbiol* (2000) **77**:475–485. doi:10.1016/S0378-1135(00)00332-1

366. Pétrilli V, Dostert C, Muruve DA, Tschopp J. The inflammasome: a danger sensing complex triggering innate immunity. *Curr Opin Immunol* (2007) **19**:615–622. doi:10.1016/j.coi.2007.09.002

367. Spreafico R, Ricciardi-Castagnoli P, Mortellaro A. The controversial relationship between NLRP3, alum, danger signals and the next-generation adjuvants. *Eur J Immunol* (2010) **40**:638–642. doi:10.1002/eji.200940039

368. Quandt D, Rothe K, Baerwald C, Rossol M. GPRC6A mediates Alum-induced Nlrp3 inflammasome activation but limits Th2 type antibody responses. *Sci Rep* (2015) **5**:1–12. doi:10.1038/srep16719

369. Ohlsson L, Exley C, Darabi A, Sandén E, Siesjö P, Eriksson H. Aluminium based adjuvants and their effects on mitochondria and lysosomes of phagocytosing cells. *J Inorg Biochem* (2013) **128**:229–236. doi:10.1016/j.jinorgbio.2013.08.003

370. Eisenbarth SC, Colegio OR, O'Connor W, Sutterwala FS, Flavell RA. Crucial role for the Nalp3 inflammasome in the immunostimulatory properties of aluminium adjuvants. *Nature* (2008) **453**:1122–1126. doi:10.1038/nature06939

371. Hornung V, Bauernfeind F, Halle A, Samstad EO, Kono H, Rock KL, Fitzgerald KA, Latz E. Silica crystals and aluminum salts activate the NALP3 inflammasome through phagosomal destabilization. *Nat Immunol* (2008) **9**:847–856. doi:10.1038/ni.1631

372. Kooijman S, Brummelman J, van Els CACM, Marino F, Heck AJR, Mommen GPM, Metz B, Kersten GFA, Pennings JLA, Meiring HD. Novel identified aluminum hydroxide-induced pathways prove monocyte activation and pro-inflammatory preparedness. *J Proteomics* (2018) **175**:144–155. doi:10.1016/j.jprot.2017.12.021

373. Zhu M, Li B, Ma X, Huang C, Wu R, Zhu W, Li X, Liang Z, Deng F, Zhu J, et al. Folic Acid Protected Neural Cells Against Aluminum-Maltolate-Induced Apoptosis by Preventing miR-19 Downregulation. *Neurochem Res* (2016) **41**:2110–2118. doi:10.1007/s11064-016-1926-9

374. Lukiw WJ, Percy ME, Kruck TP. Nanomolar aluminum induces pro-inflammatory and pro-apoptotic gene expression in human brain cells in primary culture. *J Inorg Biochem* (2005) **99**:1895–1898. doi:10.1016/j.jinorgbio.2005.04.021

375. Asín J, Molín J, Pérez M, Pinczowski P, Gimeno M, Navascués N, Muniesa A, de Blas I, Lacasta D, Fernández A, et al. Granulomas Following Subcutaneous Injection With Aluminum Adjuvant-Containing Products in Sheep. *Vet Pathol* (2018)030098581880914.

doi:10.1177/0300985818809142

376. Pogue AI, Percy ME, Cui JG, Li YY, Bhattacharjee S, Hill JM, Kruck TPA, Zhao Y, Lukiw WJ. Up-regulation of NF-kB-sensitive miRNA-125b and miRNA-146a in metal sulfate-stressed human astroglial (HAG) primary cell cultures. *J Inorg Biochem* (2011) **105**:1434–1437. doi:10.1016/j.jinorgbio.2011.05.012

377. Lukiw WJ, Pogue AI. Induction of specific micro RNA (miRNA) species by ROS-generating metal sulfates in primary human brain cells. *J Inorg Biochem* (2007) **101**:1265–1269. doi:10.1016/j.jinorgbio.2007.06.004

378. Bhattacharjee S, Zhao Y, Hill JM, Percy ME, Lukiw WJ. Aluminum and its potential contribution to Alzheimer's disease (AD). *Front Aging Neurosci* (2014) **6**:62. doi:10.3389/fnagi.2014.00062

379. Cao XY, Lu JM, Zhao ZQ, Li MC, Lu T, An XS, Xue LJ. MicroRNA biomarkers of Parkinson's disease in serum exosome-like microvesicles. *Neurosci Lett* (2017) **644**:94–99. doi:10.1016/j.neulet.2017.02.045

380. Gui YX, Liu H, Zhang LS, Lv W, Hu XY. Altered microRNA profiles in cerebrospinal fluid exosome in Parkinson disease and Alzheimer disease. *Oncotarget* (2015) **6**:37043–37053. doi:10.18632/oncotarget.6158

381. Martins M, Rosa A, Guedes LC, Fonseca B V., Gotovac K, Violante S, Mestre T, Coelho M, RosaMá MM, Martin ER, et al. Convergence of mirna expression profiling, α-synuclein interacton and GWAS in Parkinson's disease. *PLoS One* (2011) **6**:e25443. doi:10.1371/journal.pone.0025443

382. Bao MH, Li JM, Luo HQ, Tang L, Lv QL, Li GY, Zhou HH. NF-κB-regulated MIR-99a modulates endothelial cell inflammation. *Mediators Inflamm* (2016) **2016**:5308170. doi:10.1155/2016/5308170

383. Marichal T, Ohata K, Bedoret D, Mesnil C, Sabatel C, Kobiyama K, Lekeux P, Coban C, Akira S, Ishii KJ, et al. DNA released from dying host cells mediates aluminum adjuvant activity. *Nat Med* (2011) **17**:996–1002. doi:10.1038/nm.2403

384. Thompson JE, Phillips RJ, Erdjument-Bromage H, Tempst P, Ghosh S. IκB-β regulates the persistent response in a biphasic activation of NF-κB. *Cell* (1995) **80**:573–582. doi:10.1016/0092-8674(95)90511-1

385. Schmidt C, Peng B, Li Z, Sclabas GM, Fujioka S, Niu J, Schmidt-Supprian M, Evans DB, Abbruzzese JL, Chiao PJ. Mechanisms of proinflammatory cytokine-induced biphasic NF-κB activation. *Mol Cell* (2003) **12**:1287–1300. doi:10.1016/S1097-2765(03)00390-3

386. Tello-Lafoz M, Rodríguez-Rodríguez C, Kinna G, Loo LS, Hong W, Collins BM, Teasdale RD, Mérida I. SNX27 links DGKζ to the control of transcriptional and metabolic programs in T lymphocytes. *Sci Rep* (2017) **7**:16361. doi:10.1038/s41598-017-16370-w

387. Farasani A, Darbre PD. Effects of aluminium chloride and aluminium chlorohydrate on DNA repair in MCF10A immortalised non-transformed human breast epithelial cells. *J Inorg Biochem* (2015) **152**:186–189. doi:10.1016/j.jinorgbio.2015.08.003

388. Larruskain A, Bernales I, Luján L, de Andrés D, Amorena B, Jugo BM. Expression analysis of 13 ovine immune response candidate genes in Visna/Maedi disease progression. *Comp Immunol Microbiol Infect Dis* (2013) **36**:405–413. doi:10.1016/j.cimid.2013.02.003

389. Petrovsky N, Aguilar JC. Vaccine adjuvants: Current state and future trends. *Immunol Cell Biol* (2004) **82**:488–496. doi:10.1111/j.0818-9641.2004.01272.x

390. Petrovsky N. Comparative Safety of Vaccine Adjuvants: A Summary of Current Evidence and Future Needs. *Drug Saf* (2015) **38**:1059–1074. doi:10.1007/s40264-015-0350-4

391. Mold M, Shardlow E, Exley C. Insight into the cellular fate and toxicity of aluminium adjuvants used in clinically approved human vaccinations. *Sci Rep* (2016) **6**:31578. doi:10.1038/srep31578

392. García-Medina S, Angélica Núñez-Betancourt J, Lucero García-Medina A, Galar-Martínez M, Neri-Cruz N, Islas-Flores H, Manuel Gómez-Oliván L. The relationship of cytotoxic and genotoxic damage with blood aluminum levels and oxidative stress induced by this metal in common carp (Cyprinus carpio) erythrocytes. *Ecotoxicol Environ Saf* (2013) **96**:191–197. doi:10.1016/j.ecoenv.2013.06.010

393. Carpenter DO, Kamalov J, Birman I. Cytotoxicity of environmentally relevant concentrations of aluminum in murine thymocytes and lymphocytes. *J Toxicol* (2011) **2011**: doi:10.1155/2011/796719

394. Singh M. *Vaccine Adjuvants and Delivery Systems*. , ed. M. Singh Hoboken, NJ, USA: John Wiley and Sons (2006). doi:10.1002/9780470134931

395. Shardlow E, Mold M, Exley C. Unraveling the enigma: Elucidating the relationship between the physicochemical properties of aluminium-based adjuvants and their immunological mechanisms of action. *Allergy, Asthma Clin Immunol* (2018) **14**:1–19. doi:10.1186/s13223-018-0305-2

396. Petrik MS, Wong MC, Tabata RC, Garry RF, Shaw CA. Aluminum adjuvant linked to Gulf War illness induces motor neuron death in mice. *Neuromolecular Med* (2007) **9**:83–100. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17114826 [Accessed May 27, 2019]

397. Shaw CA, Petrik MS. Aluminum hydroxide injections lead to motor deficits and motor neuron degeneration. *J Inorg Biochem* (2009) **103**:1555–1562. doi:10.1016/j.jinorgbio.2009.05.019

398. Eidi H, David MO, Crépeaux G, Henry L, Joshi V, Berger MH, Sennour M, Cadusseau J, Gherardi RK, Curmi PA. Fluorescent nanodiamonds as a relevant tag for the assessment of alum adjuvant particle biodisposition. *BMC Med* (2015) **13**:144. doi:10.1186/s12916-015-0388-2

399. de Miguel R, Asín J, Rodríguez-Largo A, Molín J, Echeverría I, de Andrés D, Pérez M, de Blas I, Mold M, Reina R, et al. Detection of aluminum in lumbar spinal cord of sheep subcutaneously inoculated with aluminum-hydroxide containing products. *J Inorg Biochem* (2020) **204**:110871. doi:10.1016/j.jinorgbio.2019.110871

400. Shardlow E, Mold M, Exley C. From stock bottle to vaccine: Elucidating the particle size distributions of aluminum adjuvants using dynamic light scattering. *Front Chem* (2017) **5**: doi:10.3389/fchem.2016.00048

401. Mirza A, King A, Troakes C, Exley C. Aluminium in brain tissue in familial Alzheimer's disease. *J Trace Elem Med Biol* (2017) **40**:30–36. doi:10.1016/j.jtemb.2016.12.001

402. Exley C, Mold MJ. Aluminium in human brain tissue: how much is too much? *J Biol Inorg Chem* (2019) **24**:1279–1282. doi:10.1007/s00775-019-01710-0

403. Zhang X, Zuo X, Yang B, Li Z, Xue Y, Zhou Y, Huang J, Zhao X, Zhou J, Yan Y, et al.

MicroRNA directly enhances mitochondrial translation during muscle differentiation. *Cell* (2014) **158**:607–619. doi:10.1016/j.cell.2014.05.047

404. Diaz G, Zamboni F, Tice A, Farci P. Integrated ordination of miRNA and mRNA expression profiles. *BMC Genomics* (2015) **16**:767. doi:10.1186/s12864-015-1971-9

405. Ripa R, Dolfi L, Terrigno M, Pandolfini L, Savino A, Arcucci V, Groth M, Terzibasi Tozzini E, Baumgart M, Cellerino A. MicroRNA miR-29 controls a compensatory response to limit neuronal iron accumulation during adult life and aging. *BMC Biol* (2017) **15**:9. doi:10.1186/s12915-017-0354-x

406. Zong Y, Yu P, Cheng H, Wang H, Wang X, Liang C, Zhu H, Qin Y, Qin C. MiR-29c regulates NAV3 protein expression in a transgenic mouse model of Alzheimer's disease. *Brain Res* (2015) **1624**:95–102. doi:10.1016/j.brainres.2015.07.022

407. Coleman N, Castrejon A, Blaine C, Chemmachel T. *The Toxicology of Essential and Nonessential Metals*. LULU Publishing SERVICES (2017).

408. Kallmann BA, Hummel V, Toyka K V., Rieckmann P. "Soluble VCAM-1 Release Indicates Inflammatory Blood-Brain Barrier Pathology and Further Modulates Adhesion," in *Early Indicators Early Treatments Neuroprotection in Multiple Sclerosis* (Milano: Springer Milan), 115–117. doi:10.1007/978-88-470-2117-4_11

409. McMurray RW. Adhesion molecules in autoimmune disease. *Semin Arthritis Rheum* (1996) **25**:215–233. doi:10.1016/S0049-0172(96)80034-5

410. Zhang D, Yuan D, Shen J, Yan Y, Gong C, Gu J, Xue H, Qian Y, Zhang W, He X, et al. Up-regulation of VCAM1 Relates to Neuronal Apoptosis After Intracerebral Hemorrhage in Adult Rats. *Neurochem Res* (2015) **40**:1042–1052. doi:10.1007/s11064-015-1561-x

411. Schattling B, Steinbach K, Thies E, Kruse M, Menigoz A, Ufer F, Flockerzi V, Brück W, Pongs O, Vennekens R, et al. TRPM4 cation channel mediates axonal and neuronal degeneration in experimental autoimmune encephalomyelitis and multiple sclerosis. *Nat Med* (2012) **18**:1805–1811. doi:10.1038/nm.3015

412. Li S, Nie EH, Yin Y, Benowitz LI, Tung S, Vinters H V., Bahjat FR, Stenzel-Poore MP, Kawaguchi R, Coppola G, et al. GDF10 is a signal for axonal sprouting and functional recovery after stroke. *Nat Neurosci* (2015) **18**:1737–1745. doi:10.1038/nn.4146

413. Podjaski C, Alvarez JI, Bourbonniere L, Larouche S, Terouz S, Bin JM, Lécuyer MA, Saint-Laurent O, Larochelle C, Darlington PJ, et al. Netrin 1 regulates blood-brain barrier function and neuroinflammation. *Brain* (2015) **138**:1598–1612. doi:10.1093/brain/awv092

414. Mulero P, Córdova C, Hernández M, Martín R, Gutiérrez B, Muñoz JC, Redondo N, Gallardo I, Téllez N, Nieto ML. Netrin-1 and multiple sclerosis: a new biomarker for neuroinflammation? *Eur J Neurol* (2017) **24**:1108–1115. doi:10.1111/ene.13340

415. Yuan CY, Lee YJ, Hsu GSW. Aluminum overload increases oxidative stress in four functional brain areas of neonatal rats. *J Biomed Sci* (2012) **19**:51. doi:10.1186/1423-0127-19-51

416. 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* (2017) **54**:563–570. doi:10.1007/s12035-015-9650-z

417. Kumar V, Gill KD. Oxidative stress and mitochondrial dysfunction in aluminium

neurotoxicity and its amelioration: A review. *Neurotoxicology* (2014) **41**:154–166. doi:10.1016/j.neuro.2014.02.004

418. Pointer CB, Klegeris A. Cardiolipin in Central Nervous System Physiology and Pathology. *Cell Mol Neurobiol* (2017) **37**:1161–1172. doi:10.1007/s10571-016-0458-9

419. Atlante A, Calissano P, Bobba A, Giannattasio S, Marra E, Passarella S. Glutamate neurotoxicity, oxidative stress and mitochondria. *FEBS Lett* (2001) **497**:1–5. doi:10.1016/S0014-5793(01)02437-1

420. Nicholls DG. Brain mitochondrial calcium transport: Origins of the set-point concept and its application to physiology and pathology. *Neurochem Int* (2017) **109**:5–12. doi:10.1016/j.neuint.2016.12.018

421. Shu Y, Zhang H, Kang T, Zhang JJ, Yang Y, Liu H, Zhang L. PI3K/Akt signal pathway involved in the cognitive impairment caused by chronic cerebral hypoperfusion in rats. *PLoS One* (2013) **8**:e81901. doi:10.1371/journal.pone.0081901

422. 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* (2018) **19**: doi:10.3390/ijms19123725

423. Kerrisk ME, Cingolani LA, Koleske AJ. ECM receptors in neuronal structure, synaptic plasticity, and behavior. *Prog Brain Res* (2014) **214**:101–131. doi:10.1016/B978-0-444-63486-3.00005-0

424. Koo JW, Russo SJ, Ferguson D, Nestler EJ, Duman RS. Nuclear factor-κB is a critical mediator of stress-impaired neurogenesis and depressive behavior. *Proc Natl Acad Sci U S A* (2010) **107**:2669–2674. doi:10.1073/pnas.0910658107

425. Shih R-H, Wang C-Y, Yang C-M. NF-kappaB Signaling Pathways in Neurological Inflammation: A Mini Review. *Front Mol Neurosci* (2015) **8**:77. doi:10.3389/fnmol.2015.00077

426. Lehmann SM, Krüger C, Park B, Derkow K, Rosenberger K, Baumgart J, Trimbuch T, Eom G, Hinz M, Kaul D, et al. An unconventional role for miRNA: Let-7 activates Toll-like receptor 7 and causes neurodegeneration. *Nat Neurosci* (2012) **15**:827–835. doi:10.1038/nn.3113

427. Waller R, Goodall EF, Milo M, Cooper-Knock J, Da Costa M, Hobson E, Kazoka M, Wollff H, Heath PR, Shaw PJ, et al. Serum miRNAs miR-206, 143-3p and 374b-5p as potential biomarkers for amyotrophic lateral sclerosis (ALS). *Neurobiol Aging* (2017) **55**:123–131. doi:10.1016/j.neurobiolaging.2017.03.027

428. Raheja R, Regev K, Healy BC, Mazzola MA, Beynon V, Von Glehn F, Paul A, Diaz-Cruz C, Gholipour T, Glanz BI, et al. Correlating serum micrornas and clinical parameters in amyotrophic lateral sclerosis. *Muscle and Nerve* (2018) **58**:261–269. doi:10.1002/mus.26106

429. Truettner JS, Motti D, Dietrich WD. MicroRNA overexpression increases cortical neuronal vulnerability to injury. *Brain Res* (2013) **1533**:122–130. doi:10.1016/j.brainres.2013.08.011

430. Li M-M, Jiang T, Sun Z, Zhang Q, Tan C-C, Yu J-T, Tan L. Genome-wide microRNA expression profiles in hippocampus of rats with chronic temporal lobe epilepsy. *Sci Rep* (2015) **4**:4734. doi:10.1038/srep04734

431. Lin SH, Song W, Cressatti M, Zukor H, Wang E, Schipper HM. Heme oxygenase-1 modulates microRNA expression in cultured astroglia: Implications for chronic brain disorders. *Glia* (2015) **63**:1270–1284. doi:10.1002/glia.22823

432. Tao Z, Zhao H, Wang R, Liu P, Yan F, Zhang C, Ji X, Luo Y. Neuroprotective effect of microRNA-99a against focal cerebral ischemia-reperfusion injury in mice. *J Neurol Sci* (2015) **355**:113–119. doi:10.1016/j.jns.2015.05.036

433. Schwarz TL. Mitochondrial trafficking in neurons. *Cold Spring Harb Perspect Med* (2013) **3**:a011304. doi:10.1101/cshperspect.a011304

434. Drerup CM, Herbert AL, Monk KR, Nechiporuk A V. Regulation of mitochondria-dynactin interaction and mitochondrial retrograde transport in axons. *Elife* (2017) **6**: doi:10.7554/eLife.22234

435. Merolle L, Sponder G, Sargenti A, Mastrototaro L, Cappadone C, Farruggia G, Procopio A, Malucelli E, Parisse P, Gianoncelli A, et al. Overexpression of the mitochondrial Mg channel MRS2 increases total cellular Mg concentration and influences sensitivity to apoptosis. *Metallomics* (2018) **10**:917–928. doi:10.1039/C8MT00050F

436. Ni WJ, Leng XM. Dynamic miRNA-mRNA paradigms: New faces of miRNAs. *Biochem Biophys Reports* (2015) **4**:337–341. doi:10.1016/j.bbrep.2015.10.011

437. Yu L, Jiang R, Su Q, Yu H, Yang J. Hippocampal neuronal metal ion imbalance related oxidative stress in a rat model of chronic aluminum exposure and neuroprotection of meloxicam. *Behav Brain Funct* (2014) **10**:6. doi:10.1186/1744-9081-10-6

438. Wang D, Luo Y, Wang G, Yang Q. Circular RNA expression profiles and bioinformatics analysis in ovarian endometriosis. *Mol Genet Genomic Med* (2019) **7**:e00756. doi:10.1002/mgg3.756

439. Sekar S, Liang WS. Circular RNA expression and function in the brain. *Non-coding RNA Res* (2019) **4**:23–29. doi:10.1016/j.ncrna.2019.01.001

440. Ashwal-Fluss R, Rybak-Wolf A, Bartok O, Herzog M, Rajewsky N, Glažar P, Ivanov A, Hanan M, Öhman M, Pino N, et al. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol Cell* (2015) **58**:870–885. doi:10.1016/j.molcel.2015.03.027

441. Suenkel C, Cavalli D, Massalini S, Calegari F, Rajewsky N. A Highly Conserved Circular RNA Is Required to Keep Neural Cells in a Progenitor State in the Mammalian Brain. *Cell Rep* (2020) **30**:2170-2179.e5. doi:10.1016/j.celrep.2020.01.083

442. Wang G, Han B, Shen L, Wu S, Yang L, Liao J, Wu F, Li M, Leng S, Zang F, et al. Silencing of circular RNA HIPK2 in neural stem cells enhances functional recovery following ischaemic stroke. *EBioMedicine* (2020) **52**:102660. doi:10.1016/j.ebiom.2020.102660

443. Kumar L, Shamsuzzama, Jadiya P, Haque R, Shukla S, Nazir A. Functional Characterization of Novel Circular RNA Molecule, circzip-2 and Its Synthesizing Gene zip-2 in C. elegans Model of Parkinson's Disease. *Mol Neurobiol* (2018) **55**:6914–6926. doi:10.1007/s12035-018-0903-5

444. Akhter R. "Circular RNA and Alzheimer's disease," in *Advances in Experimental Medicine and Biology* (Springer New York LLC), 239–243. doi:10.1007/978-981-13-1426-1_19

445. Lu Y, Tan L, Wang X. Circular HDAC9/microRNA-138/Sirtuin-1 Pathway Mediates Synaptic and Amyloid Precursor Protein Processing Deficits in Alzheimer's Disease.

Neurosci Bull (2019) **35**:877–888. doi:10.1007/s12264-019-00361-0

446. Wang H, Li Z, Gao J, Liao Q. Circular RNA circPTK2 regulates oxygen-glucose deprivation-activated microglia-induced hippocampal neuronal apoptosis via miR-29b-SOCS-1-JAK2/STAT3-IL-1β signaling. *Int J Biol Macromol* (2019) **129**:488–496. doi:10.1016/j.ijbiomac.2019.02.041

447. Bai Y, Zhang Y, Han B, Yang L, Chen X, Huang R, Wu F, Chao J, Liu P, Hu G, et al. Circular RNA DLGAP4 ameliorates ischemic stroke outcomes by targeting miR-143 to regulate endothelial-mesenchymal transition associated with blood–brain barrier integrity. *J Neurosci* (2018) **38**:32–50. doi:10.1523/JNEUROSCI.1348-17.2017

448. Vea A, Llorente-Cortes V, de Gonzalo-Calvo D. "Circular RNAs in blood," in *Advances in Experimental Medicine and Biology* (Springer New York LLC), 119–130. doi:10.1007/978-981-13-1426-1_10

449. Chen L, Wang C, Sun H, Wang J, Liang Y, Wang Y, Wong Corresponding G. The bioinformatics toolbox for circRNA discovery and analysis. doi:10.1093/bib/bbaa001

450. Ng WL, Marinov GK, Liau ES, Lam YL, Lim YY, Ea CK. Inducible RasGEF1B circular RNA is a positive regulator of ICAM-1 in the TLR4/LPS pathway. *RNA Biol* (2016) **13**:861–871. doi:10.1080/15476286.2016.1207036

451. Wang Y hong, Yu X hui, Luo S shun, Han H. Comprehensive circular RNA profiling reveals that circular RNA100783 is involved in chronic CD28-associated CD8(+)T cell ageing. *Immun Ageing* (2015) **12**:17. doi:10.1186/s12979-015-0042-z

452. Chen X, Jiang C, Sun R, Yang D, Liu Q. Circular Noncoding RNA NR3C1 Acts as a miR-382-5p Sponge to Protect RPE Functions via Regulating PTEN/AKT/mTOR Signaling Pathway. *Mol Ther* (2020) doi:10.1016/j.ymthe.2020.01.010

453. Tan WLW, Lim BTS, Anene-Nzelu CGO, Ackers-Johnson M, Dashi A, See K, Tiang Z, Lee DP, Chua WW, Luu TDA, et al. A landscape of circular RNA expression in the human heart. *Cardiovasc Res* (2017) **113**:298–309. doi:10.1093/cvr/cvw250

454. Molin AD, Bresolin S, Gaffo E, Tretti C, Boldrin E, Meyer LH, Guglielmelli P, Vannucchi AM, Kronnie G, Bortoluzzi S. CircRNAs are here to stay: A perspective on the MLL recombinome. *Front Genet* (2019) **10**: doi:10.3389/fgene.2019.00088

455. Haddad G, Lorenzen JM. Biogenesis and function of circular RNAs in health and in disease. *Front Pharmacol* (2019) **10**:428. doi:10.3389/fphar.2019.00428

456. Ebbesen KK, Hansen TB, Kjems J. Insights into circular RNA biology. *RNA Biol* (2017) **14**:1035–1045. doi:10.1080/15476286.2016.1271524

457. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: Tool for the unification of biology. *Nat Genet* (2000) **25**:25–29. doi:10.1038/75556

458. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* (2000) **28**:27–30. doi:10.1093/nar/28.1.27

459. Nicolet BP, Engels S, Aglialoro F, Van Den Akker E, Von Lindern M, Wolkers MC. Circular RNA expression in human hematopoietic cells is widespread and cell-type specific. *Nucleic Acids Res* (2018) **46**:8168–8180. doi:10.1093/nar/gky721

460. Pan X, Wenzel A, Jensen LJ, Gorodkin J. Genome-wide identification of clusters of predicted microRNA binding sites as microRNA sponge candidates. *PLoS One* (2018)

13:e0202369. doi:10.1371/journal.pone.0202369

461. Piwecka M, Glažar P, Hernandez-Miranda LR, Memczak S, Wolf SA, Rybak-Wolf A, Filipchyk A, Klironomos F, Jara CAC, Fenske P, et al. Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science (80-)* (2017) **357**: doi:10.1126/science.aam8526

462. Shan C, Zhang Y, Hao X, Gao J, Chen X, Wang K. Biogenesis, functions and clinical significance of circRNAs in gastric cancer. *Mol Cancer* (2019) **18**:136. doi:10.1186/s12943-019-1069-0

463. Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Hanan M, Behm M, Bartok O, Ashwal-Fluss R, Herzog M, et al. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol Cell* (2014) **58**:870–885. doi:10.1016/j.molcel.2015.03.027

464. Barrett SP, Salzman J. Circular RNAs: analysis, expression and potential functions. *Development* (2016) **143**:1838–1847. doi:10.1242/dev.128074

465. Yu CY, Kuo HC. The emerging roles and functions of circular RNAs and their generation. *J Biomed Sci* (2019) **26**:29. doi:10.1186/s12929-019-0523-z

466. Qian Z, Liu H, Li M, Shi J, Li N, Zhang Y, Zhang X, Lv J, Xie X, Bai Y, et al. Potential Diagnostic Power of Blood Circular RNA Expression in Active Pulmonary Tuberculosis. *EBioMedicine* (2018) **27**:18–26. doi:10.1016/j.ebiom.2017.12.007

467. Tang B, Hao Z, Zhu Y, Zhang H, Li G. Genome-wide identification and functional analysis of circRNAs in Zea mays. *PLoS One* (2018) **13**:e0202375. doi:10.1371/journal.pone.0202375

468. Dai X, Zhang N, Cheng Y, Yang T, Chen Y, Liu Z, Wang Z, Yang C, Jiang Y. RNA-binding protein trinucleotide repeat-containing 6A regulates the formation of circular RNA circ0006916, with important functions in lung cancer cells. *Carcinogenesis* (2018) **39**:981–992. doi:10.1093/carcin/bgy061

469. Zhu L, Liu Y, Yang Y, Mao XM, Yin ZD. CircRNA ZNF609 promotes growth and metastasis of nasopharyngeal carcinoma by competing with microRNA-150-5p. *Eur Rev Med Pharmacol Sci* (2019) **23**:2817–2826. doi:10.26355/eurrev_201904_17558

470. Su M, Xiao Y, Ma J, Tang Y, Tian B, Zhang Y, Li X, Wu Z, Yang D, Zhou Y, et al. Circular RNAs in Cancer: Emerging functions in hallmarks, stemness, resistance and roles as potential biomarkers. *Mol Cancer* (2019) **18**:90. doi:10.1186/s12943-019-1002-6

471. Xiu Y, Jiang G, Zhou S, Diao J, Liu H, Su B, Li C. Identification of Potential Immune-Related circRNA-miRNA-mRNA Regulatory Network in Intestine of Paralichthys olivaceus during Edwardsiella tarda Infection. *Front Genet* (2019) **10**:731. doi:10.3389/fgene.2019.00731

472. Gherardi RK, Eidi H, Crépeaux G, Authier FJ, Cadusseau J. Biopersistence and brain translocation of aluminum adjuvants of vaccines. *Front Neurol* (2015) **6**: doi:10.3389/fneur.2015.00004

473. Shardlow E, Mold M, Exley C. The interaction of aluminium-based adjuvants with THP-1 macrophages in vitro: Implications for cellular survival and systemic translocation. *J Inorg Biochem* (2020) **203**:110915. doi:10.1016/j.jinorgbio.2019.110915

474. de Miguel R, Asín J, Largo AR, Molín J, Echeverría I, de Andrés D, Pérez M, de Blas I,

Mold M, Reina R, et al. Detection of aluminum in lumbar spinal cord of sheep subcutaneously inoculated with aluminum-hydroxide containing products. *J Inorg Biochem* (2019)110871. doi:10.1016/j.jinorgbio.2019.110871

475. Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, Seitz H. MicroRNA target prediction programs predict many false positives. *Genome Res* (2017) **27**:234–245. doi:10.1101/gr.205146.116

476. Michaels YS, Wu Q, Fulga TA. "Interrogation of functional miRNA-target interactions by CRISPR/Cas9 genome engineering," in *Methods in Molecular Biology* (Humana Press Inc.), 79–97. doi:10.1007/978-1-4939-6866-4_7

477. Szabo L, Salzman J. Detecting circular RNAs: Bioinformatic and experimental challenges. *Nat Rev Genet* (2016) **17**:679–692. doi:10.1038/nrg.2016.114

478. Barrett SP, Salzman J. Circular RNAs: Analysis, expression and potential functions. *Dev* (2016) **143**:1838–1847. doi:10.1242/dev.128074

479. Tomljenovic L, Shaw CA. "Answers to Common Misconceptions Regarding the Toxicity of Aluminum Adjuvants in Vaccines," in *Vaccines and Autoimmunity* (Wiley Blackwell), 43–56. doi:10.1002/9781118663721.ch4

480. Kuhn DE, Martin MM, Feldman DS, Terry A V., Nuovo GJ, Elton TS. Experimental validation of miRNA targets. *Methods* (2008) **44**:47–54. doi:10.1016/j.ymeth.2007.09.005

481. Panda A, Gorospe M. Detection and Analysis of Circular RNAs by RT-PCR. *Bio-Protocol* (2018) **8**: doi:10.21769/bioprotoc.2775