This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Transcriptomic analysis of sheep macrophages and their response to lipopolysaccharide

Mary Elizabeth Blanche McCulloch

Thesis submitted for the degree of Doctor of Philosophy,
College of Medicine and Veterinary Medicine
University of Edinburgh
July 2018
Declaration

I declare that this thesis presented for the degree of Doctor of Philosophy at the College of Medicine and Veterinary Medicine, University of Edinburgh, has been composed by myself and help from other people is clearly acknowledged. I confirm that the work has not been submitted for any other academic degree or professional qualification at this University or any other institution.

Mary McCulloch

July 2018
Abstract

Sheep are ruminant animals, highly adapted to exploit pastures of low biological value and an economically important livestock species. They also provide a useful animal model for multiple areas of human medicine. Their productivity is compromised by many viral, bacterial, fungal, protozoan and helminthic pathogens. The innate immune system evolves under stringent selective pressure from pathogens. In this project, I aimed to identify genes involved in innate immunity in sheep by dissecting the transcriptome of sheep macrophages. The analysis focused on identification of genes that may be associated with either disease susceptibility or resistance traits. The project formed part of the broader transcriptomic atlas for sheep.

A method was established and validated for the production of sheep bone marrow derived macrophages (BMDMs) grown in macrophage colony-stimulating factor (CSF1). These cells responded to bacterial Lipopolysaccharide (LPS), an archetypal agonist of the pattern recognition receptor, TLR4, with induction of inflammatory cytokines, but unlike rodent macrophages, sheep produced no nitric oxide. Bone marrow-derived macrophages were produced from male (n=3) and female (n=3) Scottish Blackface X Texel animals used in The Sheep Atlas project, and mRNA was isolated from the cells at 0, 2, 4, 7 and 24 hours following stimulation. Two different protocols of mRNA preparation were used. For 0 and 7 hour samples, samples were depleted of rRNA, and RNA-Seq was carried out at a depth of 100 million reads. For all samples in the time course (including 0 and 7), mRNA was prepared by polyA selection, and RNA-Seq was performed at lower depth (25 million reads) and a detailed analysis of the different outcomes is presented.

Two pipelines, Kallisto and StringTie were used to identify and quantify transcripts in the LPS time course transcriptomic data, along with other subsets of innate immune cells from the wider atlas. The former pipeline provides quantitative known transcript expression estimates, the latter generates novel transcript and gene models. Analysing the transcriptional signatures of these samples provided insight into the metabolic, regulatory and innate response pathways that sheep share with other animals, with both Myeloid differentiation primary response 88 protein (MYD88) dependent and independent pathways being activated following LPS stimulation, with hundreds of the same downstream cytokines being highly expressed in response, such as TNF and many interleukins as seen in other species innate responses. It has also highlighted aspects of the
response that separate sheep from other animals, such as their metabolism and biosynthesis of steroid and tryptophan as well as demonstrating differences in specific expression of certain genes. Two highly regulated and expressed genes noted during the peak response at 7hrs (ENSOARG00000005159 and ENSOARG00000006889), were both assigned a functional annotation as being the protease inhibitor Serpin family B member 2 gene (SERPINB2). This appears to be a gene duplication. This sheep expression profile of SERPINB2 is shared with mice but distinct from humans and pigs. Similar to other animals, individual animals were found to vary markedly in their transcriptional response to LPS, demonstrating hyper, early and delayed responses between the individuals.

The current reference transcriptome OarV3.1 contains 28757 transcripts, of which only 18488 are functionally annotated. More than 85% of all reference transcripts were detected in sheep macrophages and over 300 candidate annotations for genes identified only by Ensembl sheep (Ovis aries) gene ID (ENSOARG) numbers were manually assigned a functional annotation by clustering co-expression estimates using the network analysis tool, Miru and inferring function by ‘guilt by association’ of these unknown genes which demonstrated similar expression profiles to those with known function. The nearest orthologues and synteny with other species were used to validate these suspected annotations. A select few examples include PYD and CARD domain containing gene (PYCARD), all the early growth response factors and many of the C-X-C Motif chemokines.

In summary, this project identified multiple sheep-specific aspects of the innate immune response, whilst assigning hundreds of genes a functional annotation and detecting new transcript models for multiple known genes. Many regulated, highly expressed, multi exoned, novel gene models have also been identified which are worthy of future investigation.
Lay Summary

Macrophages are cells of the immune system that are found in every tissue in the body. Their main function is to recognise and destroy pathogens that cause infectious disease. Every species has evolved to deal with specific infectious challenges and as a consequence, the genes that are needed for effective immunity vary between species and between individuals. Variation between individuals explains differences in disease resistance.

Sheep are ruminant animals and an important livestock species. This project aimed to identify the sets of genes that are required for effective immune defence in the sheep, and to dissect the ways in which sheep differ from other animals. The outcomes will support effective breeding of sheep with greater resistance to infectious disease.
Acknowledgements

Professor David Hume, Professor Kim Summers and Dr James Prendergast. I would never have got through this without you all and this thesis owes everything to your brilliance. I will always be in awe of your knowledge and the patience you have all shown me in your teaching and the help given in editing this thesis. Thank you for being so generous with the time you gave, the remarkable kindness and the support that you have shown me throughout my time undertaking this PhD, which without doubt has been the toughest challenge of my life.

Professor David Hume, it is hard to articulate the lasting impression and impact you have had on me. Your phenomenal brain and scientific knowledge have no limits, yet you have always been so understanding, patient and encouraging of the crumbling, sleep deprived and hormonally challenged brain that has been mine for much of this PhD. Thank you so much for all your teaching, for sharing your love for macrophages and innate immunity, your excitement and enthusiasm for all genes and species, and for all the editing and amazing, generous help you have given in producing and bringing together this thesis. I am incredibly grateful to you for allowing me to join team Sheep.

Professor Kim Summers, thank you for all the help, support, kindness and understanding you have given me in supervision of my project and with the production of this thesis, as well as on a personal level. Your clarity of explanation, teaching and insight as to where and when help is needed, your open door and your generosity with time is all greatly appreciated. Thank you for sharing your incredible eye for detail and way with words, and for all the encouragement you have given.

Dr James Prendergast, thank you for introducing me to the world of computers, your patience, kindness and understanding with teaching me from scratch and for sharing a fraction of your amazing bioinformatic skills and knowledge. I am so grateful for all your explanations and help when things went wrong. Your remarkable ability to solve problems and find solutions to ‘bugs’ when they arose will be forever remembered and much appreciated.

Thank you to Professor David Argyle of R(D)SVS for being instrumental in my decision to undertake this PhD, your initial encouragement and for organising the funding, which in partnership with Professor David Hume and The Roslin Institute made this PhD possible.
Thank you to my thesis committee for their encouragement and support which includes Dr Gerry McLachlan and Dr Fiona Houston.

Dr Emily Clark, I am forever grateful for all your help, wise words and calmness throughout my time at Roslin, and your incredible organisational skills, gentle understanding and your kindness and generosity with offering help and teaching, all of which have made a lasting impression. I will always be incredibly grateful for everything you did to help and support me in the lab upon my return. Dr Steve Bush, thank you for sharing your enthusiasm for bioinformatics, the many tricks and some of the magic that you know (highlighting, the right click, the double click and the need for organisation, notepad and screens all made a massive difference – thank you!). Dr Clare Pridans, who knows everything about keeping macrophages happy in culture, Dr Kristin Sauter and Dr Rachel Young, thank you all for introducing me (as a rough and not so ready, aged vet from practice) to scientific laboratory work, code, conduct and equipment. Thanks to your teaching, knowledge and expertise, the data for this project was created and everything and one is still in one piece – thank you.

Thank you also to the many members of the Hume lab and Roslin institute who I would like to mention as they have helped, supported and encouraged me in this project and my time at Roslin; Iseabail Farquhar, Lindsey Waddell, Lucas Lefevre, Dr Anna Raper, Dr Tim Regan, Dr Rocio Rojo, Dr Sara Clohisey, Charity Muriuki, Dr Zofia Lisowski, Dr Mark Barnett, Ailsa Carlisle, Dr Kay Boulton, Dr Carys Pugh and Dr Stefan Hoj-Edwards. Thank you to Tim King, Peter Tennant and Adrian Ritchie from Dryden farm for your initial help on farm.

A special thanks to Dr Fiona Allan and Dr Iveta Gazova for your understanding, kindness and encouragement and all the time given to me and my family with writing up this thesis. Thank you Norman Russell for generating the beautiful photos of the Atlas Sheep, including my favourite which I hope makes the perfect finish to this thesis.

There are many people that I owe a huge debt of gratitude to for their continued personal support, friendship and encouragement: Professor Simone Meddle, thank you for all your listening, advice and the many dog walks which have literally been my salvation. You are such an inspiration, always kind, understanding and with boundless enthusiasm. Dr Susan Kempson, thank you for your encouragement and support of my return from vet practice to academia and Dr Cali Ingham for your continued support and honesty with how tough balancing life would be/ has been and all the kindness and support you have given me. Dr
Thalia Blacking, Dr Guraa Bergkvist, Dr Erika Abbondati and Dr Jorge Del-Pozo you have all been so supportive, kind and most brilliant friends throughout, thank you all so much.

So many local friends and neighbours have shown remarkable kindness, support and helped when possible; Emma and Robin Hutt, you are the greatest neighbours anyone could wish for. Cathi Gibson, Belle Dieckfoss and Claire Witney, thank you for helping look after Kitty when you did and for your kindness, support and encouragement for me to ‘keep going’ and just ‘get it finished’.

My long suffering, wonderful family; without your loyalty, support, love, encouragement and kindness this PhD would never have been completed. My mother Susan Davidson, father Roderick Davidson and dear brother Euan Davidson, and my parents-in-law, John and Cicely McCulloch, you all know just how much you have helped us as a family get through this. Thank you for always being there when we needed you and for all the help, support and kindness you have given. Last but by no means least: my beloved, loving, patient, supportive and kind, dear husband Robert, and my greatest achievements and most treasured and loved little people, Roddy and Kitty. This thesis is dedicated to all of you.
Publications


List of Figures

Figure 1.1 Summary of some of the diverse plasma membrane receptors of macrophages (taken from (Gordon & Pluddemann, 2017))…………………….6

Figure 1.2 Summary of the differentiation of monocytes from bone marrow haematopoietic stem cells (HSC) and their migration and recruitment into different tissues (adapted from (Mosser & Edwards, 2008))………………... 12

Figure 1.3 LPS signalling through TLR4………………………………………………………………….. 21

Figure 2.1 Morphological differentiation of sheep bone marrow cells in response to CSF1……………………………………………………………………………………………………….35

Figure 2.2 The two pipelines used to quantify transcript and gene expression for samples utilised in this project……………………………………………………………….39

Figure 2.3 An overview of the experimental design, the biological sample sources, the data generated and the bioinformatics utilised for downstream interpretation in this thesis…………………………………………………………………….46

Figure 3.1 The distribution of maximum transcript expression estimates for StringTie (FPKM) and Kallisto (TPM) pipelines……………………………………………………….54

Figure 3.2 Correlation of Kallisto and StringTie gene expression estimates of resting BMDMs (at 0 hr LPS)........................................................................................................56

Figure 3.3 Correlation of Kallisto and StringTie gene expression estimates of peak response BMDMs (at 7 hr LPS)........................................................................................................57

Figure 3.4 PCA to demonstrate library type differences.................................................63

Figure 3.5 Scatter plots to demonstrate correlation of gene expression estimates at 0hr LPS, detected by different library types...............................................................65

Figure 3.6 Scatter plots to demonstrate correlation of gene expression estimates at 7hr LPS, detected by different library types........................................................................66

Figure 3.7 Venn diagram of all OarV3.1 transcripts detected using Kallisto, for each of the library types used for 0hr and 7hr LPS samples. .................................67

Figure 3.8 Retrovirus genome structure............................................................................69

Figure 4.1 Principal components of all Sheep Atlas immune cell samples.................79

Figure 4.2 Sample-based networks of transcript expression........................................81

Figure 4.3 A network layout based on average transcript expression across Sheep Atlas immune cells ........................................................................................................83

Figure 4.4 Blood Leukocyte (BL) and peripheral blood mononuclear cell (PBMC) clusters extracted from the main network analysis.................................84

Figure 4.5 Macrophage enriched clusters extracted from the main network analysis.87
Figure 4.6  Alveolar macrophage enriched clusters extracted from the main network analysis..................................................................................................................................................................................................................................................................................................................93
Figure 5.1  PCA plot of all polyA selected BMDM LPS response samples........................................105
Figure 5.2  Identification of LPS responsive BMDM enriched clusters .............................................109
Figure 5.3  Sample-based networks comparing Stringtie and Kallisto expression estimates from individual sheep BMDM timecourse samples..........................115
Figure 5.4  BMDM response gene-based network expression graphs..............................................116
Figure 5.5  Kallisto expression estimate network graph highlighting each key cluster of interest.........................................................................................................................117-118
Figure 5.6  Identification of clusters with decreased expression over time in response to LPS..................................................................................................................................................................................................................................................120
Figure 5.7  Visualisation of the clusters demonstrating a regulated peak in response to LPS at 2hrs..................................................................................................................................................................................................................................................................................122
Figure 5.8  Expressed transcript varaints of NFKBIZ across the timecourse ................................123-124
Figure 5.9  Visualisation of the clusters demonstrating a regulated peak in response to LPS at 4hrs..................................................................................................................................................................................................................................................................................125
Figure 5.10 Visualisation of the clusters demonstrating a regulated peak in response to LPS at 7hrs..................................................................................................................................................................................................................................................................................127
Figure 5.11 Visualisation of the clusters demonstrating a regulated peak in response to LPS at 24hrs..................................................................................................................................................................................................................................................................................128
Figure 5.12 Mammalian arginine metabolism pathway in macrophages (taken from manuscript) ..................................................................................................................................................................................................................................................................................136
Figure 5.13 Sashimi plots for ADGRE1 across multiple species (taken from submitted manuscript)..................................................................................................................................................................................................................................................................................143

List of Tables

Table 3.1  Comparative analysis of alignment rates for different analysis pipelines..52
Table 3.2  Numbers and proportions (%) of transcript types detected in each of the library types..................................................................................................................................................................................................................................................................................68
Table 5.1: Transcript types utilised over the sheep BMDM response to LPS.................78
Table 5.2  Protein coding and total number of transcripts (OarV3.1.81) detected at each BMDM time point following LPS stimulation.................................................103
Table 5.3  Genes identified in the BMDM LPS response clusters of The Sheep Atlas which have been found to vary in expression pattern across other species. 107

Table 5.4  Sheep specific LPS responsive genes 110-112

Table 5.5  Differential expression patterns of enzymes associated with arginine metabolism and the production of NO during the macrophage response to LPS. 137
List of Abbreviations

AM  alveolar macrophage
bp  base pair
BL  blood leucocyte
CD  cluster of differentiation
CSFs colony stimulating factors
CLRs C-type lectin receptors
OarV3.1 current sheep genome assembly
DAMPs damage-associated molecular patterns
DC  dendritic cell
DNA deoxyribonucleic acid
ERV endogenous retrovirus
ER endoplasmic reticulum
FDR false discovery rate
FACS fluorescence-activated cell sorting
FPKM fragments per kilobase of exon per million reads mapped
FFAR free fatty acids receptor
FANTOM Functional annotation of mammalian genome
FAANG Functional Annotation of Animal Genomes
GO gene ontology
GTF gene transfer format
GC glucocorticoid
IRFs interferon regulatory factors
ISGS International Sheep Genomics Consortium
KO knockout
LRRs leucine-rich repeats
LPS lipopolysaccharide
lincRNA long intergenic non-coding RNA
IncRNA long non-coding RNA
MHC major histocompatibility complex
mRNA messenger RNA
Contents

DECLARATION ................................................................................................................... I

ABSTRACT ......................................................................................................................... II

LAY SUMMARY ................................................................................................................... IV

ACKNOWLEDGEMENTS ..................................................................................................... V

PUBLICATIONS .................................................................................................................. VIII

LIST OF FIGURES ............................................................................................................... IX

LIST OF TABLES ................................................................................................................ X

LIST OF ABBREVIATIONS ................................................................................................. XII

CONTENTS ........................................................................................................................ I

CHAPTER 1  INTRODUCTION .......................................................................................... 1

1.1 Innate immunity: function and overview ................................................................. 1

1.2 Major macrophage functions .................................................................................... 4
    1.2.1 Pattern recognition receptors and surface receptors of macrophages .............. 5
    1.2.2 Toll like receptor (TLR) signalling ................................................................. 7
    1.2.3 Intracellular pathogen receptor signalling .................................................... 10

1.3 Macrophage differentiation and colony stimulating factors .................................... 11
    1.3.1 Macrophage cell types and origins ................................................................. 11
    1.3.2 Macrophage colony stimulating factors ....................................................... 13
    1.3.3 Macrophage colony-stimulating factor (M-CSF/CSF1) ................................ 15
    1.3.4 Colony Stimulating factor 1 receptor (CSF1R) ............................................. 16

1.4 Regulators of macrophage function and their receptors .......................................... 16

1.5 Lipopolysaccharide (LPS) recognition and TLR4 activation .................................. 18

1.6 Variation in innate response ..................................................................................... 23
1.6.1 Evidence for differential gene expression across species ...................................................... 23
1.6.2 Variation in response between individuals ............................................................................ 24

1.7 Immunometabolism .............................................................................................................. 25

1.8 The relevance of sheep: Man’s most ancient provider ........................................................... 26
1.8.1 Worldwide importance of sheep ........................................................................................... 26
1.8.2 Sheep Genome and Transcriptome ....................................................................................... 27

1.9 RNA-Seq technology and transcriptomics .............................................................................. 28

AIM OF THE PROJECT ..................................................................................................... 30

CHAPTER 2 GENERAL METHODS AND APPROACHES USED TO QUANTIFYING
TRANSCRIPT AND GENE EXPRESSION ............................................................................ 31

2.1 Ethics statement ................................................................................................................... 31

2.2 Animal Selection ................................................................................................................... 31

2.3 Killing, tissue collection, cell isolation and cryopreservation ................................................. 32
2.3.1 Bone Marrow Processing ....................................................................................................... 32
2.3.2 Processing of blood cells ........................................................................................................ 33
2.3.3 Alveolar Macrophage Isolation .............................................................................................. 33

2.4 Cell culture of Bone Marrow Derived Macrophages and LPS treatment ......................... 34

2.5 LPS Stimulation, cell photography and Trizol reagent treatment ......................................... 35

2.6 RNA extraction, quality and quantification checks prior to submission for RNA-Sequencing at
Edinburgh Genomics .................................................................................................................... 36

2.7 RNA sample submission ........................................................................................................ 36
2.7.1 Illumina TruSeq protocols for Ribodepleted (total) stranded RNA Library Preparation and
PolyA-selected (mRNA) stranded Library Preparation used by Edinburgh Genomics ............... 37

2.8 Quality control checks on raw reads (Fastqc tables) ............................................................... 38

2.9 Pipelines used to quantify transcript and gene expression (StringTie and Kallisto2) and
BAMstat ...................................................................................................................................... 39
2.9.1 Kallisto Pipeline for processing raw reads .......................................................................... 39
2.9.2 StringTie Pipeline for processing raw reads ................................................................. 40

2.10 Visualisation tools (Muru, R, IGV) .............................................................. 42

2.11 Methods and tools used to functionally annotate key genes of interest. .......... 43
   2.11.1 Annotating reference transcripts with no functional annotation ............... 44

2.12 Summary of Chapter 2 ....................................................................................... 45

CHAPTER 3 EVALUATION OF TRANSCRIPTOME ANALYSIS METHODS ............ 47

3.1 Introduction ........................................................................................................... 47

3.2 Comparison of the outputs of Kallisto2 and StringTie Pipelines ...................... 50

3.3 Identification of differentially-expressed transcripts ........................................ 60

3.4 Ribodepleted Total RNA compared to PolyA-selected mRNA libraries ........... 62

3.5 Differential detection of transcripts with different library preparation .......... 64

3.6 Examination of repeat regions within genome .................................................. 68

3.7 Discussion .......................................................................................................... 73

CHAPTER 4 IDENTIFICATION AND ANNOTATION OF SHEEP MACROPHAGE
TRANSCRIPTS USING IMMUNE CELL DATA FROM THE WIDER SHEEP ATLAS ... 75

4.1 Introduction .......................................................................................................... 75

4.2 Methods .............................................................................................................. 77

4.3 Results and Discussion ....................................................................................... 78
   4.3.1 Examination of a principal components analysis of Sheep Atlas immune cells ... 78
   4.3.2 Network analysis ............................................................................................ 79
   4.3.3 Identification and annotation of transcripts enriched in the sheep macrophage lineage. ............................................................................................................... 86
   4.3.4 The sheep alveolar macrophage signature ...................................................... 92

4.4 Conclusion .......................................................................................................... 97

CHAPTER 5 DETAILED ANALYSIS OF THE SHEEP BMDM RESPONSE TO LPS 99
5.1 Introduction .......................................................................................................................... 99

5.2 Methods ............................................................................................................................. 100

5.2.1 Establishing a sample network expression graph using Kallisto expression estimates for individual sheep BMDM responses ............................................................................................... 102

5.2.2 Expression estimates for individual sheep BMDM responses using StringTie ..................... 102

5.3 Results ................................................................................................................................ 103

5.3.1 Overview of the sheep temporal LPS response ................................................................... 103

5.3.2 BMDM response clusters from The Sheep Atlas: Clusters 45 and 52 ............................. 106

5.3.3 The BMDM response clusters from the Immune cell averaged expression network expression graph ........................................................................................................................................ 108

5.3.4 Network analysis of the BMDM response to LPS in individual sheep............................. 114

5.3.5 Down-regulated transcripts in the response to LPS ............................................................. 119

5.3.6 Inter-individual variation in the early response phase ....................................................... 121

5.3.7 Inter-individual variation at the mid early response phase ............................................. 124

5.3.8 Inter-individual variation at the peak inflammatory response phase ............................ 126

5.3.9 Inter-individual variation at the resolution of inflammation phase ................................ 127

5.4 The metabolic shifts of sheep BMDMs in response to LPS ............................................... 128

5.4.1 Metabolism in immune cells ............................................................................................. 128

5.4.2 Analysis of genes involved in intermediary metabolism ................................................. 130

5.5 Discussion ........................................................................................................................... 138

CHAPTER 6 SUMMARY OF ACHIEVEMENTS AND FUTURE DIRECTIONS .................. 144

6.1 General discussion .............................................................................................................. 144

6.2 Genome annotation and the new genome assembly .......................................................... 145

6.3 The diversity of macrophage responses .............................................................................. 146

6.4 Individual variation in the response to LPS ....................................................................... 147

6.5 Conclusion .......................................................................................................................... 148

APPENDICES ...................................................................................................................... 149

REFERENCES ....................................................................................................................... 150
Chapter 1 Introduction

The immune response involves the temporal induction and repression of thousands of genes, and has evolved to protect all living organisms from both internal and external threats to their survival, under intense selective pressures from pathogens (Kimbell & Beutler, 2001), whilst tolerating the independent co-evolution of beneficial host-microbe relationships (Bishop, 2015). Sheep are man’s most ancient provider, being the first domesticated livestock species more than 12,000 years ago (Larson et al., 2014), and as ruminant animals their immune response has evolved alongside a symbiotic relationship with their fermentative rumen microbiome, which is thought to have evolved 35-40 million years ago (Hackmann & Spain, 2010). The products of fermentation are vital in meeting the sheep’s metabolic needs (Shabat et al., 2016).

Mice have dominated immunological research to date, being the most commonly used animal model for human medicine (Perlman, 2016), however there are significant differences between mice and humans in how these species respond to the same immunological stimuli (Schroder et al., 2012). There is divergence in the response between different mouse strains and between individual mice (Raza et al., 2014) and human individuals vary in their immune response (Brodin & Davis, 2017). Different breeds of pig and individual pigs can vary in their response but pigs have been found to resemble humans more closely than mice (Kapetanovic et al., 2012, Kapetanovic et al., 2013). As sheep depend upon their microbiome to a greater extent than any of these monogastric mammals, it is expected that their immune response will vary between individuals and be different to that of monogastric animals.

1.1 Innate immunity: function and overview

The mammalian immune response is dependent upon regulation of a careful interplay between two broad, complementary arms of defence, innate and adaptive immunity (Paul, 2011). The innate immune system is evolutionarily ancient, rapid, independent of antigen and provides the first line of defence for all single cell organisms and multicellular classes of plants and animals (Tauben, 2003, Beutler, 2004). The adaptive immune system evolved more recently, being restricted to vertebrates with jaws (gnathostomes) (Cooper & Alder,
and is highly specific, adaptable and demonstrates classical immunological memory (Ehrlich, 1900, Litman et al., 2010).

Adaptive immunity is mediated by B and T lymphocytes of the lymphoid system (Alberts et al., 2002). Innate immunity is mediated by cells of the mononuclear phagocyte system (MPS), which encompasses bone marrow pluripotent cells, blood monocytes, dendritic cells (DCs) and tissue macrophages. Cells of the MPS share similar morphology (stellate with endocytic activity), are capable of specific enzyme production for example non-specific esterase, lysosomal hydrolases and ectoenzymes, and have the ability to phagocytose ‘foreign’ particles such as latex, colloidal carbon and display endocytic receptors for Fc and complement (Hume et al., 2002, Hume, 2006). Cells of the innate immune system interact with the lymphoid system at many levels, and the lymph nodes themselves provide clearance of dead macrophages, which has been extensively reviewed (Zhang et al., 2016, Gordon & Pluddemann, 2017).

Function 1 of the immune system is to distinguish between self and non-self, whilst recognising self-cells that have become abnormal and therefore present a threat to survival. For example cancerous cells, damaged cells or cells containing intracellular pathogens, have to be distinguished from normal self-cells, and non-self-cells which are desirable or vital for the host such as gut microbiota, have to be distinguished from pathogens (Brestoff & Artis, 2013). This function is mediated through the process of tolerance which establishes the key hallmarks of self-cells during development and by interaction between the innate and adaptive immune systems (Mowat, 2003). Professional phagocytic cells play a major role, as they are able to recognise, capture, process, and present potential antigens to cells of the adaptive immune system, driving the activation, expansion and differentiation of B and T lymphocytes which improve clearance of the pathogen (Batista & Harwood, 2009). Thus, the innate immune system not only functions as the first line of defence, initiating the inflammatory response, but also provides protection whilst adaptive responses develop and influences the nature of adaptive responses (Janeway, 1989, Palucka & Banchereau, 1999, Pasare & Medzhitov, 2005, Paul, 2011). Indeed, vaccine design is based upon a detailed understanding of this innate immune signalling (Sundling & Sandgren, 2016).

Function 2 of the immune system is to maintain immunological memory so that when a pathogen is encountered again the system can respond rapidly and efficiently. Whilst this is
primarily mediated through the adaptive immune system, the innate immune system has also been found to develop “trained immunity”, through transcriptional and epigenetic reprogramming (although shorter lived and less specific than that of the adaptive immune system), when a pathogen is reencountered (Netea et al., 2011, Netea et al., 2016). Epigenomics encompasses the complexities of chromatin architecture and the precise physiological targets of regulatory proteins involved in recognising, depositing and removing chemical modifications from chromatin, as well as other factors which can alter gene expression (van der Heijden et al., 2018). Initial evidence for this arose from studies which found murine macrophages developed tolerance to the bacterial endotoxin, lipopolysaccharide (LPS) at the molecular level, where gene-specific chromatin modifications were responsible for the transient silencing of certain groups of genes whilst priming other groups (Foster et al., 2007). More recently the mouse microbiome has been found to be responsible for long-term functional reprogramming of bone-marrow derived DCs, which overall shows that communication from the microbiome to the bone marrow can have significant immunomodulatory effects (Burgess et al., 2014). Trained innate memory has also been suggested as a contributing factor to chronic immune-mediated chronic diseases such as atherosclerosis (van der Heijden et al., 2018).

The organised system of the innate response is evolutionarily ancient and vital in ensuring the survival of all multicellular organisms, hence the system is highly optimized by natural selection and the Red Queen co-evolution between hosts and parasites. Two Red Queen processes are known to provide the extremes of the spectrum of co-evolution; the Red Queen arms race which is driven by positive selection and reduces genetic variation, and Red Queen dynamics which preserves genetic variation for the long-term (Lighten et al., 2017). Genes available to a given species have often been lost or changed function through evolution (Viljakainen, 2015). The molecular mechanisms for this evolution may involve base substitutions, gene duplications, deletions, alternative splicing, gene recombination, domain shuffling, retrotransposition, and gene conversion (Buchmann, 2014). The patterns of these changes increase understanding of the evolutionary dynamics between host and pathogen. This thesis presents a study of sheep innate immune genes to add to this understanding.

Herbivorous ruminant animals such as sheep are particularly dependent upon their diverse gut microbiota (bacteria, fungi and protozoa) to convert complex carbohydrates such as
cellulose into volatile fatty acids (VFAs) for ruminant metabolism (Doreau & Ferlay, 1994). Cellulose represents the most abundant biopolymer on earth, a highly specific and architecturally diverse macromolecule made of up of repeating glucose units (Klemm et al., 2005). The sheep immune system must be able to distinguish subtle differences between desirable and undesirable components, so that only organisms and compounds that pose a true risk to the animal are attacked rapidly and efficiently. Therefore, the ruminant immune system must be highly specialised, and important evolutionary strategies could be discovered by comparing different ruminants with each other and non-ruminant animals.

1.2 Major macrophage functions

Macrophages are professional phagocytes that were first recognised by Elie Metchnikoff (the father of innate immunity) in 1882, who revealed how important phagocytosis is in development, homeostasis and disease (Gordon, 2016). They represent the tissue compartment of the MPS. They are situated throughout the body, in every mammalian tissue, and strategically are most concentrated at areas of greatest risk of invasion of pathogens, lining all epithelia and microvasculature, which ensures they are the first line of defence (Hume et al., 2002, Hume, 2006).

Their exquisite sensitivity to their microenvironment is achieved through a diverse repertoire of surface receptors which are expressed in infinite combinations, with interaction between ligand and receptor leading to a cascade of gene transcription, changing the gene expression pattern to result in “activation” (Mosser & Edwards, 2008, Jenkins & Hume, 2014, Gordon & Pluddemann, 2017). Any gene expression program can be fixed in certain tissue macrophages and induced in others, depending on what is required and the surrounding microenvironment, which means that the resident tissue macrophage becomes increasingly adapted in a dynamic and continuous way to its microenvironment (Wynn et al., 2013, Hume, 2015).

As well as surveillance and phagocytosis of microbes and dead/damaged/dying cells, activation of macrophages involves the acquisition of microbicidal effector functions, the secretion of proinflammatory cytokines and the downstream recruitment of other immune cells (Adams, 1989, Schroder et al., 2004). Macrophages process and recycle compounds such haemoglobin and iron (Knutson & Wessling-Resnick, 2003). Cytokines released by macrophages in response to stimuli have local and systemic effects, influencing cellular proliferation, differentiation, acquired immunity, embryonic development and wound
repair, which means that as a cell type, macrophages are akin to a disperse homeostatic organ (Gordon & Pluddemann, 2017) which largely dictates the outcome of any disease (Murray & Wynn, 2011, Gordon et al., 2014, Hume, 2015, Gordon & Pluddemann, 2017).

Their broad function, anatomical diversity and plasticity is owed to an equally diverse, complex and highly regulated transcriptome. The cascade of activation and downstream transcription involves thousands of genes being tightly regulated in organised sequential waves of expression, either being induced and increasing in expression or being repressed and decreasing, involving multiple signalling pathways and overall creating a complex regulatory expression network. Feedback control genes are induced in conjunction with proinflammatory genes, acting at every possible level of the cascade (Hume, 2008, Biswas & Mantovani, 2010, Gautier et al., 2012, Wynn et al., 2013, Gordon et al., 2014, Jenkins & Hume, 2014, 2015, Hume, 2015).

Activation of macrophages and the downstream inflammatory response does result in a transient decline in tissue function. This results in some undesirable aspects of the stress response, including heat shock, hypoxia, high levels of reactive oxygen species and glucose and amino acid deprivation. These aspects affect cellular function and the capacity to switch metabolic homeostatic set points (Medzhitov, 2010). By comparing species and individuals in their macrophage transcriptional response to stimuli, causative genetic variations between species, breeds and individuals may help explain differences between species and individuals in susceptibility and resilience to disease.

1.2.1 Pattern recognition receptors and surface receptors of macrophages
Classic macrophage cell surface receptors are shown in Figure 1 (taken from (Gordon & Pluddemann, 2017)) and include F4/80(ADGRE1), CD68, CSF1R, MERTK and CD64 (which can all also serve as surface markers for immunocytochemistry or fluorescence-activated cell sorting (FACS) analysis). Numerous receptors/markers mediate phagocytosis of a wide range of particles including apoptotic cells (such as CD36, SRA and TIM4), circulating ligands (such as CCR2 which detects the chemokine MCP-1), growth factors (colony stimulating factors (CSFs)), sialic acid-binding immunoglobulin-like lectin 1 (SIGLEC1 / CD169), which along with CD36, play important roles in adhesion, and scavenger receptors such as SR-A which internalises polyanionic ligands (Gordon & Pluddemann, 2017). The large repertoire of receptors summarised in Figure 1.1 (taken from (Gordon & Pluddemann, 2017)) reflects the multiple complex functions of macrophages.
Figure 1.1 Summary of some of the diverse plasma membrane receptors of macrophages (taken from (Gordon & Pluddemann, 2017)).
Receptors involve a range of structural domains, described in the key.

Surveillance and recognition of particles is achieved through the expression of conserved germline encoded pattern recognition receptors (PRRs), first identified over 30 years ago (Janeway, 1989). PRRs recognise endogenous and exogenous conserved pathogen-associated molecular patterns (PAMPs), such as lipopolysaccharide (LPS), flagellin and lipoproteins, and damage-associated molecular patterns (DAMPs), associated with host danger signals, which are released in response to stress, tissue damage and necrotic cell death (Akira et al., 2001, Gordon, 2002, Janeway & Medzhitov, 2002, Ravasi et al., 2002, Hoffmann, 2003, Taylor et al., 2005, Akira et al., 2006, Hume, 2008). Following phagocytosis or endocytosis, pathogens may enter cells and become a target for a wide array of intracellular recognition receptors (Schroder & Tschopp, 2010, Patel & García-Sastre, 2014).

PRRs include toll like receptors (TLRs), RIG-I like receptors (RLRs), nod-like receptors (NLRs), C-type lectin receptors (CLRs) and the intracellular DNA sensors DAI and AIM2 which have been extensively reviewed (Akira et al., 2006, Schroder & Tschopp, 2010). CLRs such as macrophage mannose receptor (MMR/MRC1) and attractin (CD93/C1QR1), which are expressed on the cell surface as transmembrane proteins, can be secreted as molecules,
which are able to recognise carbohydrates, proteins, lipids and even inorganic components such as ice (Brown et al., 2018).

PRRs can act alone or in concert, with the sequential detection of pathogens by different receptors in different cellular compartments resulting in a complex interplay of diverging downstream signalling cascades, that tailor the response specifically to the particular class of microbe (Franchi et al., 2008, Shaw et al., 2008). Many of these signalling pathways are highly conserved, such as the activation of transcription factors nuclear factor-[kappa]B (NFKB) and Activator protein 1 (AP-1) which lead to inflammatory cytokine expression. Interferon regulatory factors (IRFs) lead to type I interferon production and prime and activate adaptive responses. Together these instruct the host response to infection or injury (Janeway & Medzhitov, 2002, Akira et al., 2006, Kawai & Akira, 2010, Takeuchi & Akira, 2010).

1.2.2 Toll like receptor (TLR) signalling
Vertebrate toll like receptors (TLRs) are evolutionarily conserved both in coding sequence and signalling pathways and function (Kim & Ausubel, 2005, Roach et al., 2005). They are type I transmembrane receptors, present on the cell surface and intracellularly on the endoplasmic reticulum (ER), endosome, lysosome or endolysosome, and their importance, signalling and function in both innate immunity and in regulating activation of adaptive immunity has been extensively reviewed (Akira et al., 2001, Schnare et al., 2001, Gordon, 2002, Akira et al., 2006, Chang, 2010, Kawai & Akira, 2011, Broz & Monack, 2013, Takeda & Akira, 2015). TLRs recognise a wide range of lipids, lipoproteins, proteins and nucleic acid ligands. They were first identified with their involvement in dorsoventral patterning in the development of *Drosophila* flies, but it was later discovered that they were important in the response to fungal infection of *Drosophila* (Lemaitre et al., 1996) and then the human homologue of the *Drosophila* toll protein was shown to induce the expression of inflammatory cytokines via the NF-KappaB (NFKB) pathway as well as the expression of the co-stimulatory molecule B7.1, required for activating the adaptive immune response (Medzhitov & Janeway, 1997).

TLRs are known to be involved in many vertebrate diseases (Arbour et al., 2000, Nischalke et al., 2011, Piñero et al., 2017). They are central to intestinal development and protection,
as commensals are recognised under homeostatic conditions and TLRs regulate the critical intestinal epithelial barrier (Rakoff-Nahoum et al., 2004). TLR polymorphisms directly affect signalling, with reduced signalling predisposing to many diseases and health concerns including septic shock and gram negative bacterial infections (Lorenz et al., 2002), inflammatory bowel disease and ulcerative colitis (Torok et al., 2004), meningococcal infections (Smirnova et al., 2003), premature birth (Lorenz et al., 2002), leprosy and mycobacterial infections (Bochud et al., 2003) to name but a few, yet a decreased risk of atherogenesis, possibly due to a diminished inflammatory response and so lower concentrations of inflammatory cytokines (Kiechl et al., 2002, Miller et al., 2005).

TLRs all consist of an ectodomain containing varying numbers of leucine-rich repeats (LRRs), which is responsible for recognition of the PAMP/DAMP. The transmembrane domain has a role in receptor oligomerization, important for assembly and function of TLRs (Godfroy et al., 2012). The conserved cytoplasmic tail with a Toll/interleukin-1 (IL1) receptor (TIR) domain in the cytoplasm recruits required signalling molecules (Akira et al., 2006), which in turn activate transcription factors including NFkB, IRFs, and MAP kinases which dictate the downstream inflammatory response and signalling, involving cytokines, chemokines and type I IFNs (Akira & Takeda, 2004, Kawasaki & Kawai, 2014).

Adaptor proteins define the particular biological response of any TLR (Yamamoto et al., 2004). There are five important adaptor proteins involved in TLR signalling which have been extensively reviewed (O’Neill & Bowie, 2007, Belinda et al., 2008). All are recruited for signalling via the Toll/interleukin-1 (IL-1) receptor (TIR) domain (Xu et al., 2000).

1) Myeloid differentiation primary-response protein 88 (MyD88), which is utilised by all TLRs except TLR3, and is also found to be involved in IFN gamma, IL1 and IL18 signalling (Lord et al., 1990, Hultmark, 1994, Medzhitov et al., 1998).

2) TIR-domain containing adaptor protein (TIRAP/MAL) is a sorting adaptor that recruits MyD88. It is required by TLRs 1, 2, 4 and 6, but also endosomal TLR9, which demonstrates how this adaptor can bind to different lipids and so initiate specific transduction pathways (Kagan & Medzhitov, 2006). It can also be degraded by suppressor of cytokine signalling 1 (SOCS1) (Horng et al., 2002, Yamamoto et al., 2002).
3) TIR-domain containing adaptor protein inducing interferon β (TRIF), which is utilised by TLRs 3 and 4 and which activates IRF3 and NFKB and mediates apoptosis (Yamamoto et al., 2002, Yamamoto et al., 2003).

4) TRIF-related adaptor molecule (TRAM), which is required by TLR4 to recruit TRIF and regulated by protein kinase C (Fitzgerald et al., 2003, Yamamoto et al., 2003).

5) Sterile α- and armadillo-motif-containing protein (SARM), which has been shown to negatively regulate TRIF and so signalling of TLR3 and TLR4 (Mink et al., 2001, Couillault et al., 2004, Carty et al., 2006).

There are 6 major families of vertebrate TLRs, 10 TLR genes that have been recognised in humans and sheep and 12 TLR genes in mice, 9 receptors encoded in the Drosophila genome, and 20 putative TLR variants in zebrafish (Roach et al., 2005, Li et al., 2015). This repertoire is very distinct from invertebrates, with 253 TLR genes that have been detected in the purple sea urchin (Strongylocentrotus purpuratus) genome (Buckley & Rast, 2012, Kawasaki & Kawai, 2014). Strong selective pressures have resulted in these different families of TLR, which vary in length in their LRR domains and the general class of PAMP recognised. Vertebrate TLRs produce a “star phylogeny”, implying that they are all evolving at about the same slow rate. Selection for maintaining a functional endogenous signalling network (such as conserving binding with accessory molecules like CD36 or cytokines such as IL1B) appears to be the strongest evolutionary pressure, which is unusual for multigene families, where often some members derive new functions (Roach et al., 2005).

TLRs 3, 7, 8 and 9 are all expressed on endosomal membranes inside cells, so their ligands require internalisation. Other TLRs are expressed on the cell surface (TLRs 1, 2, 4, 5, 6, 10, 11, 12). This distribution throughout the cell and the trafficking from the ER to the Golgi to the specific membrane for expression, is critical for PAMP recognition and determining self from non-self which otherwise leads to autoimmunity (Tabeta et al., 2006, Lee et al., 2013).

TLRs 1, 2, 4 and 6 recognise bacterial lipids, TLRs 3, 7, and 8 recognise viral RNA, TLR9 recognises bacterial DNA, and TLR5 and 10 recognise bacterial and parasite proteins. TLRs 11 and 12 expressed in mice recognise flagellin and profilin (Koblansky et al., 2013).

Enhanced antigen uptake by TLRs is important for presentation (Hemmi & Akira, 2005,
Kawai & Akira, 2010), and each TLR induces distinct patterns of gene expression, involving the same signalling as IL1R (Akira & Takeda, 2004).

Several transmembrane molecules help regulate TLR signalling including a glycoposphatidylinositol-anchored protein CD14, which is a co-receptor with TLR4 and also required for TLR7 and 9 induction of proinflammatory cytokines (Baumann et al., 2010) and MD2 which is needed for LPS recognition and endocytosis (Zanoni et al., 2011).

1.2.3 Intracellular pathogen receptor signalling

Intracellular PRRs that act as cytosolic sensors include retinoid acid-inducible gene-I (RIG-I)-like helicases (RLHs) such as RIG-I and MDA5, DNA sensors such as DAI and AIM2 (Patel & García-Sastre, 2014) and nucleotide oligomerization domain (NOD)-like receptors (NLRs) which act as molecular scaffolds within the cytoplasm and can activate NFKB, MAPKs and control activation of inflammatory caspases which in turn are responsible for the maturation and release of potent cytokines such as IL1B and IL18 (Franchi et al., 2009, Schroder & Tschopp, 2010).

Nod-like receptors (NLRs) represent the largest family of PRRs. There are 23 human genes and at least 34 mouse genes that encode NLRs. These receptors are expressed in the cytosol, primarily in macrophage and neutrophil cells (Franchi et al., 2009). They have multiple N-terminal domains responsible for downstream signalling. These include a caspase recruitment domain (CARD), pyrin domain (PYD), acidic domain, and a NACHT domain which is the only domain shared by all NLRs. The NACHT domain is centrally located, determining 3 families of NLRs; NODs (NOD1-2, NOD3/NLRC3, NOD4/NLRC5, NOD5/NLRCX1), the NLRPs (NLRP1-14, also called NALPs) and the IPAF subfamily, consisting of IPAF (NLRC4) and NAIP (Schroder & Tschopp, 2010).

The key signalling platform of the innate immune system has been called the inflammasome. Some NLRs are known to provide a scaffold for canonical inflammasome activation, which are then connected to caspase 1 via an adaptor protein ASC (encoded by PYCARD). NLRP1, NLRP3 and IPAF are able to activate caspase 1, a proinflammatory caspase which mediates IL1B processing, release and therefore signalling (Martinon et al., 2002) and IL18 (Latz et al., 2013).

RIG-I like receptors (RLRs) are expressed in the cytosol of cells. They are able to sense ‘foreign’ 5’-triphasphate (5’-ppp) double stranded RNA molecules produced during
infection (as host mRNAs synthesised by RNA polymerase II produce a 5'-ppp end which is then processed to have an added 7-methyl guanosine cap) and following ligation, trigger efficient downstream signalling to mount an antimicrobial response against many viruses and intracellular bacteria (Yoneyama et al., 2004, Patel & García-Sastre, 2014, Bordon, 2015). RIG-1 is a protein belonging to the DExH-box family, and has been shown to have roles in proinflammatory pathways, including the response to bacterial LPS, as it regulates expression of cyclooxygenase-2 (COX-2) (Imaizumi et al., 2002), and the response to IFN-γ, which induces expression of RIG-1 (Cui et al., 2004).

1.3 Macrophage differentiation and colony stimulating factors

1.3.1 Macrophage cell types and origins

During early mammalian embryonic development, the first macrophages appear in the yolk sac (Cuadros et al., 1992, Bertrand et al., 2005). Once circulation is established, haematopoietic stem cells (HSC) from the yolk sac start to colonise the liver, and both yolk sac and liver macrophages seed populations to the brain (microglia), skin (Langerhans cells) and lung (alveolar macrophages), which can self-renew (Kikuchi & Kondo, 2006, Lux et al., 2008, Swirski et al., 2014, Hume et al., 2016). The replenishment of tissue macrophages as to whether they are derived through self-renewal or from circulating monocytes varies between adult tissues, inflammation and malignancy states, promoting the concept of a layered myeloid system (Hashimoto et al., 2013, Bain et al., 2014, Amit et al., 2016, Ginhoux & Guilliams, 2016, Perdiguero & Geissmann, 2016).

In the adult animal, myelopoiesis occurs in the bone marrow, and haematopoietic multipotential progenitor cells commit to the myeloid lineage, differentiating in response to specific haematopoietic growth factors, primarily CSF1 (Sordet et al., 2002) and IL34 (Lin et al., 2008) which has been extensively reviewed (Hume et al., 2002, Hume, 2006, Orkin & Zon, 2008, Wynn et al., 2013) and is summarised in Figure 1.2. Both CSF1 and IL34 share a highly conserved receptor encoded by the c-fms proto-oncogene, Macrophage colony-stimulating factor 1 receptor (CSF1R) (Pixley & Stanley, 2004, Chitu & Stanley, 2006, Garceau et al., 2010).
Figure 1.2 Summary of the differentiation of monocytes from bone marrow haematopoietic stem cells (HSC) and their migration and recruitment into different tissues (adapted from (Mosser & Edwards, 2008)).

All monocytes have the ability to differentiate to become macrophages if provided with the correct stimuli (Yona et al., 2013). In humans, so-called classical monocytes possess high levels of the co-receptor for LPS (CD14) and low levels of the Fc receptor CD16. In mice, these cells are defined by high levels of the marker Ly6c. Classical monocytes possess high levels of the CC-chemokine receptor 2 and respond to the ligand CCL2 (monocyte chemotactic peptide 1; MCP1). In response to inflammatory mediators, they are released from the bone marrow and into the circulation for recruitment and differentiation into mature tissue macrophages, which over time then respond to their microenvironment and become ever more specialised and differentiated (Yona et al., 2013, Ginhoux & Guilliams, 2016). Non-classical monocytes (in humans defined by low CD14 and high CD16, and in mice by low Ly6C) express an alternative chemokine receptor, CX3CR1. They are believed to differentiate from the classical cells, are more long-lived, and function mainly as patrolling cells in the blood (Geissmann et al., 2010)
Dendritic cells (DCs) are a family of professional antigen-presenting cells able to prime naïve T cells and many are phagocytic and overlap in function with monocytes and macrophages. DCs are now classified as a separate lineage of the mononuclear phagocyte system (MPS) that have been shown to differ in their growth factor dependence compared to macrophages, depending upon FLT3L (Guilliams et al., 2014, Jenkins & Hume, 2014, Hume, 2015). There are two separate lineages: DCs derived from monocytes and those that arise from a separate adult HSC derived, DC precursor (Guilliams et al., 2014). There are three resident groups of DCs; plasmacytoid DCs (pDCs), migratory DCs and monocyte derived DCs and there are multiple further subsets for each of these groups, differing in the surface markers they express (CD11c, CD8, CD103, F4/80 and CD11b) (Liu et al., 2009, Jenkins & Hume, 2014). Inflammation and the presence of growth factors change the different DC populations and they can be mobilised, activate naïve and memory T cells and secrete inflammatory cytokines; thus they are a very important cell type during inflammatory and autoimmune responses (Wicks & Roberts, 2016). All DCs require Fms-like tyrosine kinase 3 ligand (FLT3L) growth factor for development (Ginhoux et al., 2009), and further differentiation is controlled by transcription factors such as IR8, BAFT3, NFIL3 and IRF4 (Hume et al., 2002, Ginhoux et al., 2009, Guilliams et al., 2014).

1.3.2 Macrophage colony stimulating factors

Haematopoietic growth factors are hormone-like proteins that promote blood cell growth and bone marrow proliferation. Included in this group of proteins are the colony stimulating factors (CSFs), which engage with multipotential hematopoietic stem cell progenitors to induce them to become committed to the myeloid cell line, helping govern myeloid cell survival, differentiation, maturation and proliferation (Ichikawa et al., 1966, Metcalf, 2008, Metcalf, 2016). Each factor also contributes to the host response to injury and infection and may provide future potential therapeutic targets for inflammatory and/or autoimmune diseases and cancer (Hamilton, 2008, Hamilton et al., 2016).

There are several distinct glycoprotein growth factors that can influence the differentiation of macrophages, each with a level of specificity in respect to their target cell;

1) Stem cell factor, (KIT ligand)(Broudy, 1997).

2) CSF1 (macrophage colony-stimulating factor (M-CSF)), the major regulator of macrophage formation as discussed in detail below (Stanley & Heard, 1977).
3) IL34 (Interleukin 34) which shares a receptor with CSF1 (Lin et al., 2008).

4) FLT3L (Fms related tyrosine kinase 3 ligand). Primarily associated with dendritic cells, but the receptor for this ligand is expressed on the earliest myeloid progenitors and a shared macrophage-DC progenitor (Geissmann et al., 2010).

5) CSF2 (granulocyte-macrophage colony-stimulating factor (GM-CSF)), which stimulates bone marrow committed progenitors to become monocytes and macrophages, and functions in the differentiation of antigen presenting dendritic cells (Nimer & Uchida, 1995, Hamilton, 2008), and has many other functions involved in innate and adaptive immunity (Hamilton, 2008, Wicks & Roberts, 2016), for example it is vital in alveolar macrophage processing of surfactant (Metcalf, 2016).

6) CSF3 (granulocyte colony-stimulating factor (G-CSF)), which regulates granulocyte formation, release and survival from the bone marrow (Eyles et al., 2006, Hamilton et al., 2016).

7) IL3 (Interleukin 3: multi-CSF), a powerful CSF in vitro, acting in the early stages of differentiation of multiple lineages by promoting survival and proliferation of HSCs and committed progenitors in the bone marrow (Nitsche et al., 2003). IL3 has been found to be associated with immunity to parasites and allergic inflammation (Lantz et al., 1998) and recently it has been found to downregulate matrix metalloproteinases (MMP3 and MMP13), caused by inflammatory cytokines IL1B and TNF in a mouse model of inflammatory arthritis (Kour et al., 2016).

The first act through ligand-dependent tyrosine kinase receptors. Conversely, CSF2, CSF3 and IL3 exert their effects through interacting with a corresponding cell surface receptor, which activates Janus Kinase (JAK)-dependent or Signal Transducer and Activator of Transcription (STAT)-dependent pathways and downstream signalling (Thomas et al., 2015). These factors can work individually or in synergy with one another. For example CSF2 and CSF1 combined can enhance granulocyte-macrophage colony formation (Metcalf,
Other haematopoietic growth factors include stem cell factor (SCF/c-kit/CD117), FLT3 (CD135) and IL34 that also act in a highly specific and efficient manner on their target cells (macrophages and dendritic cells) through the selective expression of corresponding receptors (Maraskovsky et al., 1996, Metcalf, 2008, Jenkins & Hume, 2014). The suppressor of cytokine signalling (SOCS) family of cytokines can modulate and terminate the proliferative responses mediated through the JAK-STAT pathway (Krebs & Hilton, 2001, Metcalf, 2016).

1.3.3 Macrophage colony-stimulating factor (M-CSF/CSF1)

CSF1 is an essential regulator of macrophage formation, actively committing cells to enter the macrophage lineage and is the universal macrophage obligate growth factor which is found in all tissues and serum. It is a basal regulator of macrophages through a negative feedback loop involving CSF1R mediated endocytosis and intracellular degradation. It is the only pleiotropic cytokine to be expressed at high levels in the blood and ubiquitously expressed by most mesenchymal cells; the other CSFs are only produced in the stimulated state (Bartocci et al., 1987, Hume et al., 2002, Gordon et al., 2014). Mutation of the CSF1 gene in rats and mice produces a severe loss of macrophages, and many pleiotropic effects on other cell types. CSF1 continues to have a critical function in adults, and when this receptor is blocked, there is rapid depletion of most tissue macrophages (Sauter et al., 2014). There are three forms of CSF1: a cell surface glycoprotein, a secreted glycoprotein and a secreted proteoglycan. The different forms have different effects on macrophages, and this cytokine has very important roles in homeostasis, recruitment and proliferation of macrophages during inflammation (Chitu & Stanley, 2006, Jang et al., 2006, Jenkins & Hume, 2014). Levels of CSF1 can be increased by cells either dying locally, emigrating or mechanisms which block degradation (e.g., TNF converting enzyme (TACE), produced by activated macrophages cleaves CSF1R, which is the sole uptake/removal mechanism) (Bartocci et al., 1987, Sester et al., 1999, Jenkins & Hume, 2014). Increased CSF1 production acts to replenish tissue macrophages, increasing macrophage tissue densities by increasing the production of macrophages via recruitment of monocytes and CCR2 chemokine receptor ligands (Tagliani et al., 2011) and also by increasing proliferation of tissue resident macrophages. Resident tissue macrophages can also proliferate, above the level induced by CSF1, in response to the cytokine, IL4 (Jenkins et al., 2013) which also directs a specific activation pathway (see below).
1.3.4 Colony Stimulating factor 1 receptor (CSF1R)

The receptor for CSF1 (and IL34), is expressed specifically by macrophage lineage cells (Rojo et al., 2017). Macrophage-specific CSF1R expression involves a conserved regulatory element in the second intron, Fms intronic regulatory element (FIRE), which is crucial in directing macrophage specific transcription (Sauter et al., 2013, Hume et al., 2017, Rojo et al., 2017). CSF1R signalling is required for monocyte and macrophage survival, proliferation and differentiation in every tissue of the body, including microglia, myeloid dendritic cells, osteoclasts, monocytes and all tissue macrophages where it governs their development and function (Sasmono et al., 2003, Wei et al., 2010, Nandi et al., 2012, Hume et al., 2016).

Ligation of CSF1R, with either CSF1 or IL34 (both of which are dimers and interact with two receptor molecules) produces a conformational change, receptor dimerization, autophosphorylation and phosphorylation of downstream molecules which increase gene expression of proteins that will remodel the cytoskeleton, enabling macrophage survival, proliferation, motility and differentiation into the mature macrophage that is able to reach full transcriptional potential including the production of a wide range of cytokines and enzymes (Sweet et al., 2002, Pixley & Stanley, 2004, Stanley & Chitu, 2014). Removal of the CSF1R signal changes gene expression in itself, as macrophages are no longer able to respond to CSF1, so genes such as TLR9 that are normally repressed by CSF1, increase in expression (Sester et al., 2005).

The two ligands, CSF1 and IL34, have no discernible sequence similarity. However they do both show similar helical cytokine folds in their biologically active regions and when IL34 is bound to the N-terminal immunoglobulin domains of CSF1R, it shows some similarity in architecture to that of the CSF1/CSF1R assembly (Ma et al., 2012). CSF1R is the only receptor for CSF1. This is different from IL34 which has been shown to act via CSF1R and via a receptor-type protein-tyrosine phosphatase-ζ (PTP-ζ), a cell surface chondroitin sulphate proteoglycan (Nandi et al., 2013).

1.4 Regulators of macrophage function and their receptors

The plasticity of macrophages means that activation and differentiation are inextricably linked to one another and regulation of macrophage differentiation and transcriptional signalling has been extensively reviewed (Ravasi et al., 2007, Ramsey et al., 2008, Bhatt et al., 2012, Amit et al., 2016, Hume et al., 2016).
Helper T cell-mediated immune responses have been classified into two classes, Th1 and Th2, which are mediated by the production of two critical cytokines, interferon gamma and interleukin 4 (IL4). Each of these cytokines acts on macrophages to produce a polarised gene expression profile, which by analogy has been termed M1 and M2 respectively (Mosser & Zhang, 2008, Wynn et al., 2013, Murray et al., 2014).

Interferons are cytokines with important properties including antiviral (Isaacs & Lindenmann, 1987), antiproliferative and immunomodulatory effects. They signal through transcriptional regulatory pathways such as the JAK-STAT pathway, the mitogen-activated protein kinase p38 cascade pathway and the phosphatidylinositol 3-kinase cascade pathway (Platanias, 2005). Their signalling is integrated with that of multiple other cytokines during the immune response (Schroder et al., 2004).

Interferon gamma (IFN gamma) was first identified in 1965 as an antiviral substance that was produced by stimulated leukocytes (Wheelock, 1965). Originally called macrophage activating factor, it can also be secreted by NK cells, B cells, NKT cells and DCs (Munder et al., 1998). IFN gamma regulates multiple transcription factor genes, IRF 1, 2 and 9, along with STAT1 and NFkB, which are involved in co-ordinating the next sequential wave of transcription. In the macrophage the main function of IFN gamma is activation of antimicrobial, antiviral and tumoricidal activity through the induction of reactive oxygen and nitrogen species, as well as increased expression of TLRs and transcription factors for TLR responsive genes, increased antigen processing functions and MHC class I and II molecule presentation (Schroder et al., 2004). IFN gamma is also involved in defence against intracellular bacteria (most notably mycobacteria), fungi and parasites, as well as in anti-viral immunity like both type I and II interferons (Zhang et al., 2008). So called M1, or classically-activated macrophages can be induced by IFN gamma alone, but the response is amplified by co-stimulation with bacterial lipopolysaccharide (discussed below) or other agonists (Martinez et al., 2006). The phenotype includes high expression of inflammatory cytokines such as IL1B, IL6, IL12, IL23 and TNF and other effectors such as CD86, CD64 and CD16, and low expression of anti-inflammatory cytokines such as IL10 (Biswas & Mantovani, 2010).

The biology of M2-like, or alternatively-activated macrophages has been reviewed extensively (Biswas & Mantovani, 2010, Gordon & Martinez, 2010). Broadly-speaking, they represent an anti-inflammatory phenotype observed during allergic inflammation and
helminth infection (Mosser & Edwards, 2008, Mantovani et al., 2013, Van Dyken & Locksley, 2013). In mice, at least, alternative activation generates a shift from nitric oxide production to the metabolism of arginine through arginase (Bronte & Zanovello, 2005). The relevance of this pathway to ruminants, and indeed all large animals, will be discussed further in this thesis.

The simplification of macrophage activation states within the M1 (classically activated macrophages, by IFN gamma or Lipopolysaccharide (LPS)) (Schroder et al., 2004) or M2 (alternatively activated macrophages, by IL4) (Lawrence & Natoli, 2011) paradigm is not universally accepted (Hume, 2015). Macrophages respond to many agents other than T cell cytokines to alter their function for specific purposes (Raza et al., 2014) and transcriptomic data has demonstrated that macrophage cells vary widely in function depending upon the individual cell, the dose and combination of agonists and the time following exposure (Hume, 2015).

One of the most clinically important agonists that regulates macrophage function is glucocorticoid (GC). The glucocorticoid receptor (GR) is highly expressed by macrophages, thus they are sensitive to the effects of GC which are anti-inflammatory and immune-suppressive, and oppose the action of CSF1 (Hume et al., 1984). By interacting with glucocorticoid response elements (GRE), GC influence many transcription factors including FOS, JUN, CREB and thus, they fine tune the induction or repression of thousands of genes (Russo-Marie, 1992, Kugelberg, 2014). Mice lacking GR have greater mortality and cytokine release in response to LPS and dexamethasone (a synthetic GC) was identified to act by inhibiting p38 MAPK and therefore downstream targets (Bhattacharyya et al., 2007). The response to GC varies greatly between species, due to gain and loss of GRE in target genes (Jubb et al., 2016).

1.5 Lipopolysaccharide (LPS) recognition and TLR4 activation
Bacteria can be divided into two major groups depending upon the different staining of their cells walls; gram positive and gram negative. The outer membrane of the cell envelope of gram negative bacteria contain the potent endotoxin lipopolysaccharide (LPS), which is the major mediator of pathology in gram negative sepsis. LPS is made of a poly- or oligosaccharide anchored by lipid A, the portion/PAMP recognised by TLR4 which is highly immune stimulatory even at very low concentrations (Miller et al., 2005). Lipid A PAMPs vary in their fatty acid side chains, terminal phosphate residues and associated
modifications, which alter the charge and molecular conformation (Schromm et al., 1998) and have been shown to affect the host response (Schromm et al., 2000).

TLR4 was first identified as the likely human orthologue of the Drosophila PRR Toll (Medzhitov et al., 1997). The first genetic evidence for TLR4 being the main receptor for LPS was obtained by positional cloning of the gene involved in natural LPS resistance in certain inbred mouse strains (Poltorak et al., 1998, Qureshi et al., 1999). These seminal discoveries led to the Nobel prize in Physiology and Medicine in 2011, being awarded to Bruce A. Beutler and Jules A. Hoffmann for their discoveries concerning the activation of innate immunity and to Ralph M. Steinman for his discovery of the dendritic cell and its role in adaptive immunity.

TLR4 is found in the membrane as a complex with myeloid differentiation protein 2 (MD2). Optimal binding with LPS to the TLR4-MD2 complex also requires the glycosyl phosphatidylinositol (GPI)-anchored, macrophage membrane protein CD14, and the LPS binding protein (LBP) which is an accessory protein that circulates freely in the blood (Jiang et al., 2000).

Mice are unresponsive to LPS if any of TLR4 (Hoshino et al., 1999), MD2 (Nagai et al., 2002, Hamann et al., 2004) or, CD14 (Haziot et al., 1995) is deleted in the genome by homologous recombination. TLR4 has a large extra-cellular domain containing multiple leucine rich repeats (LRRs). Not surprisingly, evolution across species has led to variation in the extracellular domain of TLR4 and so species vary in their response to specific forms of LPS produced by different bacteria (Vaure & Liu, 2014).

There are two main divergent TLR signalling pathways with each TLR differentially recruiting the adaptor proteins: the MyD88 dependent signalling pathway and the TRIF/TRAM MyD88 independent pathway; only TLR4 utilises both these pathways (see Figure 1.3).

The MyD88 dependent pathway involves TIRAP which recruits MyD88 to the cell surface for TLRs 2 and 4, and to the endosomal membrane surface for TLR9 (Kagan & Medzhitov, 2006). Unlike TLR2 and TLR4, TLR9 activation does not depend upon TIRAP, whose actions reflect the phosphoinositide-binding domain to which lipids are targeted (Bonham et al., 2014). Following ligation, MyD88 then binds with IRAK2, IRAK4 and then IRAK 1 producing a Myddosome, which is dependent upon molecular complementarity and surface electrostatics in order to be phosphorylated and activated (Kollewe et al., 2004, Lin et al.,
IRAK1 associates with TRAF6, UBC13 and UEV1A and together with TAK1, TAB1, TAB2 and TAB3 drives TAK1 activation which activates two pathways: the IKK complex-NFKB pathway and the MAPK pathway (Chen, 2012). NFKB induces proinflammatory gene expression, whilst MAPK signalling involves ERK1, ERK2, p38 and JNK which activate the AP1 family of transcription factors which bind FOS and JUN families of bZIP proteins which provide regulatory control.

The MyD88 independent/ TRIF dependent pathway involves TRAM recruiting TRIF. This complex then interacts with TRAF6 and TRAF3. TRAF 6 interacts with RIP1 which activates TAK1. Downstream NFKB and MAPks, and TRAF3 recruit TBK1, IKKi and NEMO for IRF3 phosphorylation which can then induce type I IFN genes (Akira et al., 2006, Kawai & Akira, 2010, Kawasaki & Kawai, 2014).
Figure 1.3 LPS signalling through TLR4
Activation of macrophages by LPS involves the sequential induction and repression of thousands of genes, demanding tight regulation and including numerous feed-back control genes (induced early on in any response), which act at every level of the cascade (Adams, 1989, Nilsson et al., 2006, Mosser & Edwards, 2008, Ghisletti et al., 2010, Ostuni & Natoli, 2011, Beyer et al., 2012, Raza et al., 2014, Hume, 2015). This means that many overlapping pathways are utilised, requiring regulation over many regulatory elements such as transcription factors, as well as the expression of responder proteins involved in the cascade of activation (Natoli, 2010, Zanoni et al., 2011).

Certain sets of genes share regulation through specific transcription factors to ensure the temporal response is appropriate for the common function that a specific set of genes is required to achieve, for example responding to interferon or genes required in early inflammation or endocytosis (Hume, 2015). Macrophages from different mouse strains have been found to vary in their transcriptional response to activation due to lineage-determining transcription factors influencing the epigenetic and transcriptional landscapes (Heinz et al., 2013).

Transcription factors function by binding to regulatory regions sharing specific motifs, either proximal promoters or distal enhancers, and initiating transcription (Schmidl et al., 2014). Macrophage enriched genes include transcription factors which are linked to likely binding sites in phagocyte-restricted promoters (Hume, 2015). Amongst the most-studied is the macrophage-specific transcription factor PU.1, which binds to purine-rich motifs repeated in the proximal promoters of most macrophage-expressed genes including CSF1R (Rojo et al., 2017). Targetted mutation of the Spi-1 protooncogene gene, PU.1 gene (SPI1) in mice led to almost complete loss of myeloid differentiation (Scott et al., 1994, McKercher et al., 1996, R. McKercher et al., 1999, Schulz et al., 2012). Members of many other transcription factor families, notably AP1, CEBP, KLF, IRF, MAF and MITF families, interact with promoter and enhancer motifs in macrophage expressed genes (Rojo et al., 2017). The most studied transcriptional activation event in the response to LPS is the translocation of the NFkB complex to the nucleus, but this is only part of a complex cascade of inducible transcription factors which includes AP1, IRF and other transcription factor families (Lawrence & Natoli, 2011).

Transcription factors mediate their actions in part by chromatin remodelling, as chromatin barriers can block transcription factor binding and therefore enable different programs of
expression, allowing separation in the temporal response between primary response genes (PRGs) and secondary response genes (SRGs) (Saccani et al., 2001, Hargreaves et al., 2009). Chromatin remodelling can be influenced by the CpG content (CpG islands) of the promotor region, which in turn affects the formation of a stable nucleosome (Ramirez-Carrozzi et al., 2009). Other epigenetic influences over transcription include histone acetylation, trimethylation at lysine 4 of histone H3 (H3K4me3), DNA hypomethylation and the degree of sensitivity to DNase and RNA polymerase II (RNAPII) (Foster et al., 2007, Hargreaves et al., 2009, Ramirez-Carrozzi et al., 2009).

1.6 Variation in innate response

1.6.1 Evidence for differential gene expression across species

Although the basic biology of macrophage signalling is conserved between species and individuals, fundamental differences in response to stimuli such as LPS have been well reported (Schneemann & Schoeden, 2007). Further evidence that species vary in their response to pathogens is reflected in the many infectious bacterial diseases which show species specificity in susceptibility. Ruminants are particularly sensitive to Clostridial disease, including *Clostridium chauvoei* (blackleg), *Clostridium septicum* (braxy), *Clostridium perfringens B or C* (dysentery/enterotoxemia) as well as other bacterial infections such as *Bacteroides melaninogenicus* (foot rot), *Chlamydophila abortus /Chlamydiosis* (EAE abortion), *Campylobacter* (vibriosis), *Mycobacterium avium* subspecies *paratuberculosis*, and parasite infections such as *Toxoplasma gondii* (Toxoplasmosis), *Haemonchus contortus* (barber’s pole worm) and *Eimeria* (coccidia).

Many previous studies in multiple species, have examined macrophage differentiation and activation in a genome wide study of gene expression (Ravasi et al., 2007, Ramsey et al., 2008, Fairbairn et al., 2011, Kapetanovic et al., 2012, Baillie et al., 2017). The macrophage response to LPS has been found to be fundamentally different between mice and humans, with only 30% of orthologous genes being induced in both species (Schneemann & Schoeden, 2007, Schroder et al., 2012). This will be discussed in more detail in later chapters of this thesis in the context of comparative analysis of the response of sheep macrophages. PRR themselves are differentially regulation between species; TLRs 2, 3 and 4 have functional promotor differences between mice and humans (Heinz et al., 2003, Heinz et al., 2013). TLR9 is expressed by mouse macrophages, but not by human.
1.6.2 Variation in response between individuals

Individuals vary within any population (Darwin, 1859) and it has long been established that individuals within any population vary in their response to pathogens. Microorganisms exert strong selective pressure on genes of the innate immune system and antagonistic co-evolution between pathogen and host is a driving force of species molecular evolution (Paterson et al., 2010). Macrophage activation involves the temporal induction of numerous feedback control genes which suppress inflammation and are known to be highly polymorphic within species and produce hyper-inflammatory phenotypes if individuals possess null mutations (Gilchrist et al., 2006, Gilchrist et al., 2008, Baillie et al., 2017).

The complexities of inflammation in relation to the checkpoints and actions of specific genes and gene products underlying individual variation have been extensively reviewed (Nathan, 2002, Wells et al., 2003, Wells et al., 2005, Medzhitov, 2008). A recent large study of human monocyte responses to LPS (Fairfax et al., 2014) demonstrated that as many as 80% of expressed genes show heritable variation in the level of expression between individuals (so-called eQTL).

The inflammatory response has to balance survival during infection and injury with the impact it has on tissue function, causing pathology (Medzhitov, 2010). The huge variation in susceptibility to pathogens demonstrates the functional genetic diversity of the immune response. Single gene (Mendelian) disorders are rare but can inform on the underlying mechanisms and protection against specific infections, an example includes heterozygosity for the sickle cell haemoglobin (HbS and HbC) which result from separate mutations in the β-globin gene, which provide protection against *Plasmodium falciparum* malaria because the environment of the sickle cell red blood cell is unfavourable to the parasite (Allison, 1954).

Another example of a single gene providing protection against pathogens is *natural resistance-associated macrophage protein 1* (*NRAMP1* or *SLC11A1*). NRAMP1 is a divalent metal transporter for iron, manganese and cobalt which is expressed in the late endosomal and phagolysosomal membranes, specifically in phagocytes (Gruenheid et al., 1997, Forbes & Gros, 2003), and promotes efficient haemoglobin iron recycling (Biggs et al., 2001). A coding variation in *NRAMP1* underlies the Bcq locus in mouse, which controls susceptibility to several intracellular pathogens such as *Salmonella, Leishmania* and *Mycobacterium* (Vidal et al., 1993, Vidal et al., 1995). By restricting the phagolysosome of iron, which is
required by the pathogens for growth (Forbes & Gros, 2001), the antimicrobial agent nitric oxide is synthesised (Fritsche et al., 2007).

Polymorphisms in multiple genes surrounding TLR signalling are known to have implications for infectious diseases in humans (Miller et al., 2005). Children with inherited IRAK4 deficiency were shown to not activate NFkB and MAPK in response to ligands of TIR bearing receptors, resulting in a failure to induce downstream cytokines and all developed pyogenic bacterial infections (Picard et al., 2003).

Amino acid polymorphisms are uncommon in human TLR4, but have been found to cluster in the cytoplasmic domain so as to alter the sensitivity of detection and magnitude of the response (Smirnova et al., 2000, Smirnova et al., 2001, Miller et al., 2005).

Mutations in the extracellular domain of TLR2 have been shown to impair mycobacterial signalling in mice, leaving them more susceptible to lepromatous leprosy and Mycobacterium tuberculosis (Bochud et al., 2003).

Recognition of many of the genetic determinants of susceptibility has not yet moved into practical clinical treatment options, but as research, understanding and reports increase then realising the potential for developing novel vaccines and medical interventions become the most important goal for future studies (Burgner et al., 2006).

1.7 Immunometabolism

Arachidonic acid metabolites such as prostacyclin (PGI2), thromboxane B2, prostaglandin E2, prostaglandin F2α, leukotriene C4 (slow reacting substance of anaphylaxis), and leukotriene B4, a potent mediator of chemotaxis have long been recognised as important in homeostasis and in the macrophage response with their release an important aspect of function (Scott et al., 1982, Adams & Hamilton, 1984).

Regardless of the stimulus, macrophage activation and effector functions place demands on energy generation and cellular metabolism. Major metabolic changes have to take place for initiation of any immune response and many of the cytokines dictate endocrine regulation over whole body metabolism. The challenge is for immune cells to find adequate energy sources in a difficult environment. Immunometabolism relates to the intracellular metabolic changes and reprogramming that take place in response to immune stimuli, which is increasingly recognised as altering the function of immune cells and has been reviewed extensively by others (O'Neill et al., 2016, Gaber et al., 2017). Nutrients and
metabolites moderate inflammatory pathways as well as immune cell differentiation and when the ordered interactions between the immune response and metabolic response are disrupted a range of pathologies develops (Hotamisligil, 2017).

Given that the sheep has such different circulating metabolites to that of humans and mice, with distinctive hepatic gluconeogenic metabolism, it is anticipated that that ruminant macrophages will have distinct regulation of intermediary metabolism in immune cells.

This thesis will focus particularly on the transcriptome surrounding immunometabolism in the sheep and will be considered in more detail in chapter 5.4, where I will examine the expression of metabolism-associated genes in sheep macrophages to seek evidence for adaptation to such nutrient availability.

1.8 The relevance of sheep: Man’s most ancient provider

1.8.1 Worldwide importance of sheep

Sheep were amongst the first domesticated livestock species, over 11,000 years ago (Larson et al., 2014). This domestication has had a major impact on the modern sheep in comparison to their wild ancestor and like all domesticated animals, many of the “domestication syndrome” traits, first recognised by Darwin, can be identified, such as docility, endocrine changes, altered reproduction pattern and output, altered coat colour, facial neotony and a reduction in size.

Today sheep represent a major livestock species throughout the world, providing meat (mutton and lamb), milk, wool and skins. They are able to exploit pastures that are unsuitable for other agricultural purposes and can be integrated into other agricultural systems for efficient grassland management or crop rotation.

Cellulose is a complex and architecturally diverse macromolecule made of up of repeating glucose units (Klemm et al., 2005). Mammals are unable to produce the enzymes required to access this energy source. The rumen evolved in some herbivorous mammals to support specialised fermentative microbes (bacteria, fungi and protozoa), to convert the lignocellulose-rich plant materials, into VFAs, methane and carbon dioxide. The ruminant animal co-evolved specialised energy and lipid metabolism to utilise the fermentative by-products as mentioned briefly above; vfas (acetic, propionic and butyric acid), the microbial protein (90% of amino acids reaching the small intestine) and vitamins provided by this specialised fermentative microbiome (Doreau & Ferlay, 1994).
Both ruminant (including their innate immune system gut barrier) and microbiome have evolved together over millions of years (Russell & Rychlik, 2001) and are inextricably linked, as they depend upon one another for survival—a classic example of mutualism (Flint et al., 2008).

Within the UK there is a stratified system of sheep production that is unique and ideally suited to traditional sheep breeding management. Hillbreeds are adapted to life in inhospitable environments and are responsible for producing the replacement hill lambs, lightweight store or finishing lambs and cast ewes. They include breeds such as the Scottish Blackface, having desirable traits of increased growth rate, litter size and mothering ability. Upland areas then utilise the cross between the regular aged or cast hill ewe and the wool producing long wool rams such as the border Leicester. The resulting hybrid ewes are then crossed with terminal sire rams such as the Texel to produce the majority of UK finished lambs for meat. In the case of the Texel breed they possess a desirable mutation, which prevents translation of their Myostatin gene which is responsible for producing growth and differentiation factor 8 (GDF8). Skeletal muscle development is not then actively inhibited, resulting in double muscling and less carcass fat (Miar et al., 2014). Down breeds are kept on low ground farms and produce finished lambs and replacement terminal sire rams. Variations of this system can also be used to produce slaughter lambs, such as mating draft hill ewes directly with terminal sires.

Heterosis generated by this management system provides significant improvement to the profitability of sheep farming with crossbred lambs shown to have improved survivability and to grow faster, crossbred ewes known to be more fertile and wean more pounds of lamb and crossbred males shown to be more fertile. It is this first cross animal (Scottish Blackface (BF) crossed with Texel (T)) that has been studied in this project.

1.8.2 Sheep Genome and Transcriptome

The International Sheep Genomics Consortium (ISGC) and United States Department of Agriculture National Institute of Food and Agriculture funded and produced the latest sheep genome assembly (OarV3.1) using male and female Texel animals and also provided transcriptome information for 40 different tissues (Bahcall, 2014, Jiang et al., 2014). This work highlighted sheep specific genetic changes surrounding lipid metabolism that result in changed skin barrier lipids and their involvement with wool synthesis, as well as noting the
increased role of volatile fatty acids in ruminants compared with nonruminant animals (Bahcall, 2014).

The BBSRC funded Sheep Atlas project at the Roslin Institute, to which my project contributed macrophage-related data, was the largest gene expression dataset of its kind, examining gene expression (RNA-Seq) data from tissues at every stage of development (blastocyst to adult) and across 120 tissues of both male and female adult sheep (Clark et al., 2017). Knowing when and where a gene is expressed during the development of an animal, provides information regarding the specific biological role as well as improved understanding of functional sequence (Freeman et al., 2012, Mabbott et al., 2013, Hume & Freeman, 2014). This project provides functional information for many of the unannotated transcripts in the wider atlas through the ‘guilt by association’ principle (Oliver, 2000, Krupp et al., 2012). The animals chosen were commercial cross-breeds between Texel, the breed chosen for the sheep reference genome (Bahcall, 2014) and the disparate Scottish Blackface, to maximise heterozygosity. All the animals were subjected to whole genome resequencing. Cross-bred animals were used to enable the identification of expressed single nucleotide variants (SNVs) that provide the basis for assessment of imbalanced expression of the two parent alleles. The use of the same atlas animals in the current project enables a comparison between the data obtained on macrophages in vitro with transcriptomic data from a wide range of primary cells and tissues. Furthermore, the data on stimulated macrophages generated herein makes a significant input to the wider atlas project. Using F1 crossbred animals also maximises the heterozygosity, so that the expressed genetic variation and allele specific expression (ASE) can be explored.

The Functional Annotation of Animal Genomes (FAANG) is ‘a coordinated international action to accelerate genome to phenome’ and recognises the need for improving functional annotation (Andersson et al., 2015). The data from the Sheep Atlas, including the BMDM response to LPS samples generated in this project is freely available to support this community, as well as being deposited in the European Nucleotide Archive.

1.9 RNA-Seq technology and transcriptomics

Gene expression microarrays have provided much insight into the changes in transcriptional regulation that take place in macrophages in response to environmental stimuli (Martinez et al., 2006). Next generation sequencing technology (first introduced in 2005) (Margulies et al., 2005), has revolutionised characterisation of eukaryotic
transcriptomes, enabling novel gene clusters to be identified owing to increased transcript quality and quantity, with novel splice variants recognised and a larger dynamic range possible for detection (Wang et al., 2009, Ozsolak & Milos, 2011).

There are two broad categories of RNAs, those which encode amino acids sequences and are translated for forming proteins and those which do not. Messenger RNAs (mRNAs) are translated into proteins and possess two structures which regulate their stability, polyadenylation (long runs of adenosine residues (polyA)) at the 3’ end and the further esterified triphosphate residue at the 5′ end that forms a structure called a cap.

The second category contains the non coding RNAs (ncRNAs) are involved in many nucleocytoplasmic functions. These include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs), which both play essential roles in the maturation of mRNA (splicing) and rRNA (RNA modifications) within the nucleus, and microRNAs (miRNAs) which have been found to regulate transcription (Mattick, 2001, Memczak et al., 2013, Zhang et al., 2013, Wu et al., 2014, Fang & Fullwood, 2016). A novel class discovered first with the cloning and sequencing of full length cDNAs is the mRNA-like, polyadenylated long non-coding RNAs (lncRNA) (Mattick, 2001). The functional annotation of the mammalian genome (FANTOM) project has been central to the discovery and identification of these transcripts (Carninci et al., 2005). Most lncRNA are expressed at low levels, a few molecules per cell, and with a few exceptions, the functions of lncRNAs are unknown. Because of their low expression they are difficult to assemble reliably from short read RNA-Sequencing as shown in a very recent paper from The Sheep Atlas project which included sequence data from my project (Bush et al., 2018).
**Aim of the project**

Sheep have evolved as a species under specific environmental pressures, which has resulted in particularly complex innate immune system biology surrounding the relationship with their complex fermentative microbiome. Sheep vary in their response to various pathogens in comparison to other mammals, including closely related ruminant species, and between individuals of the same flock. It is expected that sheep have a unique innate immune relationship with their microbiome and circulating products of fermentation and their macrophage transcriptional response to TLR4 activation.

The aim of this project was to define the transcriptome of sheep macrophages and their response to LPS using RNA-Seq technology. Two outcomes were anticipated (1) a major contribution to the annotation of the sheep transcriptome and genome since studies of other species have demonstrated that macrophages are amongst the richest and most diverse sources of novel transcripts (Wells *et al.*, 2003). (2) Comparative analysis of the sheep macrophage transcriptome with other animal species, enabling identification of ruminant-specific, or sheep-specific innate immune regulation.
Chapter 2 General Methods and approaches used to quantify transcript and gene expression

This chapter addresses the methods that were used to obtain differential transcript expression estimates for each sample used in my project. I was involved in all sample collections at Dryden Farm, assisting Tim King in the schedule 1 killing and then dissection and tissue harvesting. The immediate processing of the samples post euthanasia, back in the wet lab and the preparation of these samples for cryopreservation was performed by Dr Emily Clark, Dr Rachel Young, Ms Iseabail Farquhar, Dr Lucas Lefevre, Dr Clare Pridans, Dr Kristin Sauter, Dr Anna Raper and Ms Lyndsey Waddell. Dr Rachel Young and Dr Lucas Lefevre were responsible for the culture and RNA extraction of the monocyte derived macrophages. All the bone marrow samples were cultured, treated and had RNA extracted by myself and all the bioinformatics described in this project was done by myself, with guidance from my supervisor Dr James Prendergast. A brief outline of how RNA-Seq data was generated by Edinburgh Genomics is provided, before a detailed overview of the RNA-Seq processing, programmes and tools that were used to obtain transcript expression level estimates that are examined in subsequent chapters of this thesis.

2.1 Ethics statement

All samples for this project were collected post mortem. Euthanasia was performed by personnel licenced to perform schedule 1 procedures.

2.2 Animal Selection

The animals used in this project were six adult Scottish Blackface cross Texel animals, two disparate breeds in order to maximise heterozygosity for study, with the Texel breed having been utilised in producing the reference genome OarV3.1. Three females and three males were all approximately two years of age and in good health, bodily condition and with similar husbandry histories. Having been purchased for the study they were all kept on Dryden farm for at least 2 weeks prior to culling. Identical feeding and husbandry routines were implemented for each individual, ensuring that any differential environmental influences were minimal (Rosselot et al., 2016). These animals formed part of an extensive transcriptional sheep atlas project in the Hume laboratory funded by the BBSRC and provided a wide range of other cells and tissues for transcriptional profiling. The first phase of this project was published in 2017 (Clark et al., 2017).
My project has focussed specifically on the identification, characterisation and annotation of transcripts expressed and regulated in cells of the innate immune system, specifically on the activation of macrophages generated from bone marrow progenitors \textit{in vitro}. Other immune cell types that were collected and utilised to a lesser extent in order to compare transcriptomic profiles, include macrophages derived by lavage of the lung (alveolar macrophages (AMs)), blood leucocytes (BLs), peripheral blood mononuclear cells (PBMCs which include circulating blood monocytes) and monocyte derived macrophages (MDMs).

All immune cell populations were prepared as for the previous smaller atlas project for the domestic pig (Freeman et al., 2012) and the preparation of these cell types is discussed below. Other individuals and groups who were involved in the collection and processing are referred to and acknowledged accordingly.

2.3 Killing, tissue collection, cell isolation and cryopreservation.

All animals were killed by the schedule 1 cull method of electrocution followed by exsanguination. This method avoided the use of barbiturates for euthanasia which are known to cause significant tissue damage, irrespective of species, sex and agent used (Grieves et al., 2008). All animals were killed at the same time of day (9.30am) which minimises the confounding variable of circadian rhythmicity between individuals (Keller et al., 2009, Man et al., 2016, Rosselot et al., 2016).

I collected blood from each animal during exsanguination, into a jug containing Acidic Citrate Dextrose (ACD) to prevent clotting (10ml ACD/100ml blood) and immediately transferred it to 50ml falcon tubes to provide pure blood mononuclear cells, blood leukocytes and monocyte-derived macrophages from each animal (discussed in more detail below). All tissues were removed by a team of veterinarians including myself, within 1hr 30mins of culling, and cells immediately harvested and processed as quickly as possible, so as to minimise post mortem effects on RNA levels.

2.3.1 Bone Marrow Processing

The protocol originally developed for pigs (Kapetanovic et al., 2012) was used to prepare bone marrow cells from five posterior ribs extracted from both sides of each animal, ensuring both ends of each rib remained closed so as to avoid contamination. They were immediately placed in phosphate buffered saline (PBS) within a sterile plastic bag and transported on ice back to the laboratory for processing inside the microbiological safety
cabinet. The bones were stripped of all intercostal muscle and cleaned using ethanol, cut into Sections and flushed from both ends with RPMI-1640 (Sigma UK), containing 5 mM EDTA to prevent clotting. This marrow cell suspension was then filtered through a cell strainer and centrifuged at 400g for 5 mins. The supernatant was removed and the cell pellet resuspended in 5ml erythrocyte cell lysis buffer (10mM KHCO₃, 150mM NH₄Cl, 0.1mM EDTA pH 8) for 5 mins, after which RPMI-1640 media was added to dilute the buffer and the cells were spun again at 400g for 5 mins. The supernatant was removed and the cell pellet resuspended using sheep medium (400ml RPMI1640 (Sigma UK),100ml heat inactivated sheep serum (Sigma UK), 5ml Glutamax (Invitrogen UK) and 0.5ml Penicillin/Streptomycin (25U/25µg/ml, Gibco UK)). The cells were counted and then aliquoted in cryovials at approximately 5x10⁷ cells/ml freezing medium (80% sheep serum and 20% DMSO), adding the freezing medium slowly to avoid shocking the cells with DMSO. The labelled cryovials were then placed in an isopropanol freezing unit and placed in -80°C freezer for 24hrs and then transferred to long term storage at -155°C.

2.3.2 Processing of blood cells
Total blood leukocytes were isolated by the Hopkins laboratory using published methods (Montgomery & Sise, 1990). Peripheral blood mononuclear cells (PBMCs) were isolated as previously described for the pig (Fairbairn et al., 2013) except that 2% serum was used throughout. To isolate monocytes by adherence, PBMCs were then cultured overnight in complete RPMI + recombinant human CSF1 (rhCSF1; a gift from Chiron, Emeryville, CA) in a 100cm² Sterilin plate at 10⁶ cells/ml. The following day the supernatant was removed and TRIzol ® Reagent (Invitrogen, Darmstadt, Germany) was added to the cells. The cell lysate was then aliquoted and frozen in cryovials at -80°C. The remaining PBMCs were frozen in cryovials at a concentration of 1 x 10⁸/ml in 90%FCS/10%DMSO or cultured for 7 days in complete medium in the presence of rhCSF1 to produce the monocyte derived macrophage samples (MDMs) (discussed further in Section 2.3.5).

2.3.3 Alveolar Macrophage Isolation
Briefly, the chest was opened and trachea double clamped and the entire pluck was removed. Once exteriorized, the heart and lungs were separated and with the tracheal clamp still in place, the lungs were transported on ice back to a microbiological safety cabinet for processing. The lungs were lavaged twice with 1 litre of PBS, and cells recovered by centrifugation.
If necessary, red blood cells were lysed by resuspending the pellet in red blood cell lysis buffer (NH₄Cl, NaHCO₃, EDTA). Cells were plated out at 2 x 10⁶ cells/well in a tissue culture plastic 6 well plate overnight in 2ml sheep media with rhCSF1 (see below). The following day, supernatant was removed and replaced with 1ml TRIzol®, then contents transferred to a 1ml cryovial for storage at -80 in preparation for RNA extraction.

2.4 Cell culture of Bone Marrow Derived Macrophages and LPS treatment

For the generation of bone marrow-derived macrophages (BMDMs), cells were initially recovered from long term storage at -155°C by thawing each vial rapidly in the water bath (37°C) for 2 mins. The cryovial contents were transferred to a 15ml falcon tube and very slowly (so as to avoid the shock of sudden dilution) 9ml of sheep medium was added. The tubes were gently inverted a few times and then spun for 5mins at 400g. The supernatant was removed and the cells were counted and plated out onto 100mm² bacteriological plastic plates at 2.5x10⁷ cells/plate in a total of 20ml sheep media + 200µl of recombinant human CSF1 (rhCSF1) to provide a final concentration of 10⁴ U/ml (100ng/ml).

Plates were incubated at 37°C, inside an incubator (37 °C, 5% CO₂ in air). After 4 days, 10mls of fresh sheep medium + 100µl rhCSF1 were added. The cells were checked every 2-3 days and showed increased size, granularity, vacuolar appearance and adherence to the bacteriological plastic (Wang et al., 2013). Figure 2.1 demonstrates the phenotypic changes that were visible in all cultured BMDM samples used in this project.

At the end of day 6 all cells, including the predominating confluent adherent monolayer, were removed by gentle scraping and flushing the plate with medium using a blunt needle and syringe. The cell suspension was centrifuged at 400g for 5 mins. Cells were counted and replated in a 6 well plate at 2 x 10⁶ cells/well and a total of 2mls of sheep medium + rhCSF1 (100ng/ml). On day 7 the LPS time course commenced.
Figure 2.1: Morphological differentiation of sheep bone marrow cells in response to CSF1. Sheep bone marrow cells were cultivated in sheep medium with rhCSF1 and the plates were photographed (20 X magnification) on days 0,3,4,5,6 and 7 (images from left to right). The images are typical of the time course in each of the 6 sheep studied.

<table>
<thead>
<tr>
<th>Day 0</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
<th>Day 7</th>
</tr>
</thead>
</table>

2.5 LPS Stimulation, cell photography and Trizol reagent treatment

The cells were left overnight in the 37°C incubator to adhere to the tissue culture plastic. The five time points chosen for analysis, (0, 2, 4, 7, 24 hrs) were based upon the known temporal profile of the response to LPS in other species, including mice, humans, pigs and horses. (Kapetanovic et al., 2012, Schroder et al., 2012, Raza et al., 2014, Karagianni et al., 2016, Baillie et al., 2017). LPS was added sequentially for each time starting with the longest incubation (24hrs) so that all the time points could be harvested and processed together. To initiate the LPS response, medium was replaced with 2ml sheep medium containing LPS from Salmonella enterica serotype Minnesota Re 595 to ensure a final concentration of 100ng/million cells, as used in pigs (Kapetanovic et al., 2012). For every 2ml well, a further aliquot of rhCSF1 (100ng/ml) was added to ensure that the cells were saturated with the growth factor and that most BMDMs were in the cell cycle (Raza et al., 2014).

At the end of the incubation, cells at the five time points were examined using light microscopy to determine whether LPS had induced the anticipated spreading on the substratum (Wenzel et al., 2011). Every well was examined under the light microscope and photographed, before the medium was removed and replaced with 1ml TRizol ®/well. Each
plate containing Trizol was gently agitated and the individual well/LPS time point sample contents transferred to a cryovial and stored at -80°C.

**2.6 RNA extraction, quality and quantification checks prior to submission for RNA-Sequencing at Edinburgh Genomics**

In preparation for RNA extractions, all surfaces and devices to be utilized were thoroughly cleansed using RNase-Zap® (Ambion™, ThermoFisher Scientific UK) to remove all traces of ambient RNases and nucleic acid contaminants. Only RNase free reagents and consumables were used. The sample was thawed to room temperature for 5 minutes and 200µl BCP (bromochloropropane) (Sigma Aldrich) was added. The sample was left to stand for 3 minutes and during this time the homogenate separated into a clear upper aqueous layer (containing RNA), and interphase and red lower organic layers (containing the DNA and proteins). The sample was then centrifuged for 15 minutes at 12,000 x g, at 4°C. After centrifugation the upper aqueous phase was removed and RNeasy® Mini Kits (Qiagen UK) were used to extract total RNA from each sample. This method is widely used and reported to extract high yields and good quality RNA, avoiding the need to precipitate the RNA (Grabmuller et al., 2015). The exact protocol specified in the manufacturer’s instructions was followed for each extraction and also involved an RNase-Free DNase treatment set (Qiagen UK), again following the manufacturer’s instructions.

The quantity and quality of RNA extracted for each sample was estimated using a NanoDrop spectrophotometer (NanoDrop Products, Wilmington DE USA) and the RNA integrity estimated on an Agilent 2200 TapeStation (Agilent Technologies, Santa Clara CA USA).

**2.7 RNA sample submission**

RNA samples with the best RNA Integrity number equivalent (RINe) values were selected (all with RINe >9), from each individual at each respective time point and were submitted to Edinburgh Genomics as a single submission (see Appendix 2.1) and run on the Illumina HiSeq 2500 sequencing platform.

Samples from all individuals at all time points were sent for 125 bp, paired end sequencing at 25 million read depth using Poly A selected library preparation. In addition, the same 0 and 7 hr time point samples were sent for sequencing at 100 million read depth using ribodepleted total RNA library preparation.
2.7.1 Illumina TruSeq protocols for Ribodepleted (total) stranded RNA Library Preparation and PolyA-selected (mRNA) stranded Library Preparation used by Edinburgh Genomics.

RNA samples were subjected to either mRNA-Seq (>25M read depth, as done for the majority of Sheep Atlas samples, where the priority is to examine protein coding transcripts) or total RNA-Seq (>100M read depth, as done for chosen Sheep Atlas samples of particular interest) in order to balance costs, coverage of samples collected and read depth to enable exploration of the rarer transcript. The two protocols are described briefly here, since they influence the interpretation and analysis of the RNA-Seq data. All procedures were performed by Edinburgh Genomics.

The total RNA-Seq libraries were prepared using the Illumina Truseq total stranded RNA library preparation kit with (human/mouse/rat) Ribo-zero for the depletion of rRNA in the samples (Ribodepleted_totalRNA_ILLUMINA, 2013). The mRNA-Sequencing libraries were prepared using the Illumina Truseq stranded mRNA library preparation kit (PolyA_mRNA_ILLUMINA, 2013).

The two library preparation kits differ in the pre-treatment of the RNA prior to cDNA preparation. For the ribodepleted total RNA library, the first step involved removal of both cytoplasmic and mitochondrial ribosomal RNA (rRNA), using biotinylated, target-specific oligos combined with Ribo-zero rRNA removal beads.

The first step for the polyA-selected mRNA library used oligo-dT-coated magnetic beads to purify the polyA containing mRNA molecules.

The RNA from each library prep was then fragmented using divalent cations at high temperature. The cleaved fragments were copied into first strand cDNA using reverse transcriptase and random primers. The first strand synthesis mix also included Actinomycin D which prevents spurious DNA dependent (but not RNA dependent) DNA synthesis (Ruprecht et al., 1973). This prevents artefact second-strand cDNA which would produce false antisense transcription (Perocchi et al., 2007).

The second strand was synthesised using DNA Polymerase I and RNase H (which degrades the RNA strand of the heteroduplex), with dUTP replacing dTTP in this second strand marking mix. This means the strands can be identified from one another, as the polymerase used in the assay (a Uracil-DNA-Glycosylase(UDG)) selectively degrades the
strand containing the dUTP, leaving the remaining strand to be amplified from which to generate the cDNA library for sequencing (Borodina et al., 2011). The cDNA fragments had a single ‘A’ base added and the adapter sequence was ligated. The enzymes and oligonucleotides required are all included in the Illumina Genomic DNA Sample Prep Kit that Edinburgh Genomics used.

After this, the products were purified using a Qiagen PCR Purification Kit and enriched with PCR. Both the TruSeq stranded Total RNA and mRNA HT Sample Prep Kits used by Edinburgh Genomics utilise a 96 well plate format with 96 unique adapter combinations, to index the libraries either manually (as was used for all the PolyA mRNA library samples) or using an automated robot (as was used for all the ribodepleted total RNA library samples).

2.8 Quality control checks on raw reads (Fastqc tables)

Raw read files of high throughput sequence data were returned by Edinburgh Genomics. All subsequent processing and informatics analysis was performed by myself. The fastq files were quality checked using the programme FastQC (Andrews, 2010) in Linux. This provided a modular set of analyses to ascertain whether there were any problems or biases within the data that needed to be addressed before further analysis. The per base sequence quality was excellent for all samples and no systematic issues were identified. Results including the per sequence quality scores which report if a subset of sequences have universally low quality scores, the per base N content which indicates if the sequencer has been unable to make a base call with sufficient confidence and the sequence length distribution which indicates whether there is varying lengths of sequence fragments, were all within acceptable limits and therefore no concern was felt about the quality of the results. Occasional failures in the per base sequence content, expected since the RNA-Seq libraries were being produced by priming using random hexamers, do not adversely affect the downstream analysis and so were ignored (Andrews, 2010). The per sequence GC content was similar for each sample. Failures in the sequence duplication levels, can be explained due to the nature of this RNA-Seq experiment in which some genes have exceptionally high abundance levels.

No more than 20% of the reads showed any adaptor content after 107bp. The reads were all 125bp in length. Accordingly, no adapter trimming was undertaken, as adapter sequences will only occur at the 3’ end of the read and the local alignment soft clipping tool used in Bowtie2 will resolve this problem for those 20% of reads (see Fig 2.2).
sequencing of small RNAs which would require adaptor trimming was not the priority for this project and the current coverage of the reference genome means that there is still a chance that although adaptors are supposedly synthetic and not to occur in the genomic sequence, there is a chance that they may appear in the sheep but have not yet been recognised.

2.9 Pipelines used to quantify transcript and gene expression (StringTie and Kallisto2) and BAMstat

The RNA-Seq data generated for each sample are required to generate transcript expression estimates: both for transcripts already identified in OarV3.1 (Aken et al., 2016) and for those which have not yet been recognised (novel transcripts and gene models). This necessitated two different pipelines, Kallisto2 (Kallisto) and StringTie to maximise the strengths and ability of each. Kallisto2 (Bray et al., 2016) utilised the reference transcriptome OarV3.1.81. StringTie (Pertea et al., 2015) utilised the reference genome OarV3.1.81 (http://www.ensembl.org). The relative merits and outcomes of these pipelines are discussed in detail in chapter 3.

Figure 2.2: The two pipelines used to quantify transcript and gene expression for samples utilised in this project.
2.9.1 Kallisto Pipeline for processing raw reads

Kallisto2 (Pimentel et al., 2017) utilises a method of pseudoalignment to match reads to a reference transcriptome (containing 28,757 transcript sequences for OarV3.1.81) of known coding and non-coding transcript sequences. The reference transcriptome was produced from the Ensembl 81 FTP download site using the coding and non-coding cDNA reference FASTA files for OarV3.1.81 (Bray et al., 2016). The algorithm divides each read sequence into kmers of 31bp in length and allocates each kmer to a best match transcript. Because the read alignment does not include genomic mapping, time and computing power are saved and it is robust to errors in the reads (Bray et al., 2016). The limitation is that the potential number of transcripts identified is constrained by the completeness of the reference transcriptome.

The programme Sleuth (Pimentel et al., 2017) was used to estimate the confidence of differential expression from the gene expression estimates derived from Kallisto2. Sleuth uses statistical algorithms for differential analysis that leverage the bootstrap estimates of Kallisto2 (Pimentel et al., 2017). The bootstrap provides a measure of the accuracy of the quantification of expression of a transcript by random resampling with replacement to generate estimates for the technical variance. For all samples the number of bootstraps was specified to be 1000.

The P values and q values obtained from Sleuth (discussed in more detail in Section 3.1.4) are a result of comparing the expression estimates of samples using a Wald test which tests the null hypothesis that the covariate for a given transcript is 0.

2.9.2 StringTie Pipeline for processing raw reads

After ensuring the quality of the RNA-Seq reads using FastQC, the raw reads were aligned to the reference genome (OarV3.1.81) using the programme HISAT2 (Pertea et al., 2016). The known-splicesite-infile mode was utilised which takes a list of known splice sites which HISAT2 used to align reads using 'small anchors' (the python script that the manual provided was used to make this splice site index from OarV3.1.81). When HISAT2 was unable to make a valid alignment, the set of reads was passed to an UNMAPPED file, which would then be parsed. Parsing was carried out using a perl script provided by Dr James Prendergast which fixed a bug within HISAT2 that produces double copies of .1 (first strand) reads in the unmapped output file.
To recover and analyse the unmapped HISAT2 reads, Bowtie2 (Langmead & Salzberg, 2012) was utilised in the local alignment mode. Local alignment meant that the ends of the read could be omitted or “soft trimmed”. This meant that all the multi-mappers and reads that may have had some adapter contamination present were ‘rescued’, as well as allowing reads to be mapped where there is a problem in the reference, with this second chance to be mapped by Bowtie2. The Bowtie2 default is to find a valid alignment and continue to look for alignments that are better until the limit placed on the search effort is reached (the default for this project was a maximum of 20 dynamic processing problems failing in a row). When equally good choices are found, the program uses a pseudo-random number to choose the location. This pseudo-random number generator is re-initialized for every read using a “seed” which is a function of the read name, nucleotide string and quality string, so that the mapping results are reproducible. This project selected the “--very-sensitive” option in the command line. The wide area covering the speed/sensitivity/accuracy trade off space meant that the aligner ran more slowly in this default option but was more accurate and sensitive in deciding upon the optimal alignment. This default set the maximum number of “re-seeds” when attempting to align a read with repetitive seeds as 3 and the number of mismatches permitted per seed to be 0. The trade-off between speed and sensitivity/accuracy can be adjusted by changing the seed length, the interval between extracted seeds and the number of mismatches permitted per seed. Values used here were all default for prioritising accuracy over speed.

The two alignment files from each of these tools (HISAT2 mapped.bam and Bowtie2 mapped.bam) were then merged together to produce a single merged sequence alignment map (sam) file for each sample using SAMtools 1.2 and then sorted by reference position (Li et al., 2009). SAMtools 1.2 was also used to generate an index file for each of these merged, sorted bam files.

The programme StringTie assembled these RNA-Seq alignments for each sample into potential transcripts, using a network flow algorithm as well as de novo assembly, to assemble and quantify full length transcripts representing multiple splice variants for each gene locus (Pertea et al., 2015, Pertea et al., 2016). The reference annotation file from OarV3.1.81 was also included to see whether the reference transcripts were found in the RNA-Seq data, computing the coverage of any that were recognised.
At the start of this project, the most up to date version of StringTie was version 1.0.4 which required that the programme Cufflinks was utilised as it includes a script called cuffmerge that was able to take all the assemblies for each animal and merge them together to generate a non-redundant set of transcripts (Trapnell et al., 2010, Roberts et al., 2011). This merged transcript gene transfer format (GTF) file was then used to repeat StringTie to estimate the transcript abundances for each individual and enable a comparison between individuals and samples. Subsequently, it was recognised that this pipeline has a bug. For a small minority of transcripts that significantly overlapped each other, Cufflinks assigned a unique ID and when StringTie was then performed for the second time, these transcripts were not recognised by StringTie and were automatically assigned an expression value of 0. This issue was recognised because a key known macrophage-expressed gene, CSF1R (discussed in Introduction and Chapter 4), was assigned a unique Cufflinks ID and accordingly appeared to be absent from the StringTie outcomes. The issue arose as CSF1R (ENSOART00000006925) has an upstream promoter within the 3’ UTR of the neighbouring PDGFRB gene (Visvader & Verma, 1989), so expression from the two loci was merged and allocated a unique ID.

The more recent version of StringTie was 1.3.3, for which the need to rely on Cufflinks had been resolved and Pertea et al (2017) had added their own ‘–merge’ command to achieve what previously was not possible. In repeating the pipeline in its entirety, using StringTie 1.3.3, CSF1R was found to be expressed at high levels at 0hr LPS as expected and down-regulated by LPS. The Ensembl transcript ID for CSF1R was still assigned to have an expression of 0 FPKM, as StringTie included this upstream promotor and allocated the expression to this novel transcript model (MSTRG.26740), which has an extra exon (22 exons rather than 21) compared to the reference.

From these read coverage tables that have been generated, the Ballgown software (Frazee et al., 2015, Fu J, 2016) was used to identify the differentially expressed genes and transcripts between the different samples using linear-model based analyses.

2.10 Visualisation tools (Miru, R, IGV)

Various programmes and tools have been utilised to visualise the dataset and compare gene and transcript levels between individuals, such as R and R studio, which is a language and environment for statistical computing in order to create principal component and scatter plots, as well as the powerful visualisation tools Miru (http://kajeka.com) that
enable expression pattern recognition to create meaningful clusters. The Integrative Genomics Viewer (IGV, http://software.broadinstitute.org/software/igv/) was also used.

Miru (Kajeka, UK) calculates similarity matrices based upon pairwise correlation coefficients from expression estimates for each transcript or gene across all the samples and then renders these as networks in a 3D interactive environment. Many other large gene expression analyses have used the predecessor program BioLayout Express³D (Theocharidis et al., 2009, Mabbott et al., 2011, Freeman et al., 2012, Hume et al., 2013). Clustering of co-expressed genes in this programme uses the Markov Cluster Algorithm (MCL) (Enright et al., 2002). MCL works by examining the paths (multiple pairwise comparisons between the nodes) by taking random walks and determining the probability of ‘walking’ from one cluster/node to another using two operators, Expansion and Inflation (Theocharidis et al., 2009). Clusters of genes that share transcriptional profiles tend also to participate in common pathways(Freeman et al., 2012). This means that possible function of an unknown transcript can be inferred by examining the other members of the same cluster that do have a functional annotation.

The Integrative Genomics Viewer (IGV) enables visual, interactive exploration of large, integrated genomic datasets (Robinson et al., 2011, Thorvaldsdottir et al., 2013) and has been used to observe the coverage of reads after mapping with HISAT2 and Bowtie2 to the reference Oar V3.1.81 and highlight differences between samples, individuals and the reference models. IGV was used to create Sashimi plots for selected genes and genomic regions. This enabled quantitative visualisation of the splice junctions and comparison of differentially spliced exons and alternative transcript models across multiple samples, helping identify those that are not currently recognised in the reference OarV3.1.81 (Katz et al., 2015).

2.11 Methods and tools used to functionally annotate key genes of interest.
Biomart (http://www.ensembl.org/biomart/martview/) was used to add known information for every expressed reference transcript, such as the associated gene stable ID, gene name, chromosome/scaffold name, strand, coordinates, gene type and % GC content (Smedley et al., 2015).

Many of the biologically significant transcripts of interest identified had no functional annotation. These include OarV3.1.81 reference transcripts that lack any functional
annotation and are just identified by the Ensembl ID and those transcripts and genes that have been identified by the StringTie pipeline and are totally novel.

A variety of resources was utilised to assist in functional annotation of those transcripts/genes of interest which were unknown and lacked functional annotation, prioritising those that were found to be macrophage specific (within the sheep atlas project clustering analysis) or which demonstrated regulation and differential expression in the BMDM LPS response time course.

2.11.1 Annotating reference transcripts with no functional annotation.

All coding and non-coding reference transcripts were used for the Kallisto2 index. For those reference transcripts with no functional annotation, initially Ensembl was used to visualise the transcript alignment to the genome and the gene model (Aken et al., 2016). For each protein coding transcript Ensembl provides a list of the nearest orthologues in other species. These are generated by Ensembl utilising phylogenetic trees generated by TreeBeST (Vilella et al., 2009) which represent the evolutionary history of gene families. The degree of homology at the nucleotide level can be calculated and forms the basis for the description of orthologue relationship, be it one to one, one to many, many to many or between species paralogues (Jensen, 2001). Initially any 1:1 orthologues were examined. The initial priority was to note any that appeared within the same phylogenetic clade as the sheep, the Laurasiatheria, paying attention to the best target % id and query % id. Many of these genomes are lacking in functional annotation, similar to the sheep, so other species sets were examined, with particular attention to humans and mice in the knowledge that these are the most complete and annotated reference genomes.

To reinforce predicted orthology, the synteny of the genomic region was compared within Ensembl using the species comparative viewer. The molecular clock rates of evolution vary among organisms and are known to depend upon other biological characteristics that are also undergoing evolution, for example substitution rates are influenced by metabolism (Rand, 1994). Related organisms often show similar blocks of genes in the same relative positions within the genome, and shared synteny is critical for determining the orthology of genomic regions for different species. Each reference genome is functionally annotated to a different degree. By comparing other farm animals such as pig and cattle as well as genomes with better annotation such as human and mouse it was possible to add
confidence to proposed orthology relationships and to assign proposed annotation across species.

When genes had no nearest orthologues, Ensembl provided information about the transcript such as the domains and features recognised and any variations. The BLAST tool was used to determine or infer functional and evolutionary relationships that may be shared with other genes or gene families within the sheep genome or other species and calculate the statistical significance of any matches made (McGinnis & Madden, 2004, Ye et al., 2006).

The reference transcript position and model was validated against relevant merged.BAM files (aligned using HISAT2 and Bowtie2) using IGV (Thorvaldsdottir et al., 2013), to validate reference transcript models as well as highlight novel ones of potential future interest.

2.12 Summary of Chapter 2
This thesis uses cell culture to examine the transcriptional response of sheep macrophages to stimulation with lipopolysaccharide. High-throughput sequencing and advanced bioinformatics tools were then applied to analyse the resulting data. This chapter describes those methods that were applied in all experimental procedures, summarised in Figure 2.3. The subsequent chapters present the results and interpretation of the resulting data.
Figure 2.3 An overview of the experimental design, the biological sample sources, the data generated and the bioinformatics utilised for downstream interpretation in this thesis.

- **Sheep Atlas adult (x6)**
  - F1 cross
  - 3 x male
  - 3 x female

- **Euthanasia and tissue Harvesting** for Sheep Atlas, including blood (to generate BL, MDM, PBMC samples), lungs (to generate AM samples) and bone marrow (to generate bone marrow samples)

- **Processing for cryopreservation**

- **Day 1: Bone marrow** from given individual defrosted, counted, and cultured in rhCSF1 for 1 week

- **Day 4: Cells examined and fed with fresh sheep media and rhCSF1**

- **Day 6: Cells examined, counted and replated onto tissue culture dishes**

- **Day 7: LPS time course commenced**
  - 4pm LPS administered to '24hr' plate

- **Day 8: LPS time course completed**
  - 9am LPS administered to '7hr' plate
  - 12pm LPS administered to '4hr' plate
  - 2pm LPS administered to '2hr' plate
  - 4pm all cells photographed, including untreated '0hr' plate, Trizol admin, samples labelled and frozen

- **RNA extraction**, Rneasy Mini Kits

- **RNA quality and quantity checked**

- **RNA sent for RNA-Seq at Edinburgh Genomics**

- **StringTie** Novel transcript models and reference transcript alignment (OarV3.1), gene expression estimates (FPKM)

- **Bowtie2** mapping of HISAT2 unmapped reads using soft clipping local alignment tool

- **HISAT2** mapping to OarV3.1

- **StringTie pipeline**

- **Kallisto2** Reference (OarV3.1.81) transcript expression estimates (TPM)

- **Kallisto pipeline**

- **Differential expression analysis**

- **Ballgown** for analysis of StringTie (FPKM)

- **Sleuth** for analysis of Kallisto (TPM)

- **Ensembl**
  - Functional annotation for key transcripts

- **IGV**
  - Observe coverage of reads after mapping

- **R**
  - Scatter plots
  - PCA plots

- **MIRU MCL clustering**
  - Determine network expression graphs
  - Identify key clusters

- **RNA-Seq processing** (including RNA-Seq from Sheep Atlas immune cells)

- **FastQC**

- **0hr, 2hr, 4hr, 7hr, 24hr post LPS**
  - PolyA selected RNA
  - >25 M depth
  - paired end reads

- **0hr and 7hr post LPS**
  - Ribosomal depleted total RNA
  - >100 M depth
  - paired end reads
Chapter 3 Evaluation of Transcriptome Analysis Methods

3.1 Introduction

Once a draft genome is completed, the next focus is to identify the set of transcripts (the transcriptome) that are encoded and the ways in which genes and transcripts are utilised in any given cell type/tissue at a given developmental stage and under certain physiological conditions (Ekblom & Wolf, 2014, Conesa et al., 2016). Although there are multiple steps that determine the functional output of the genome; including RNA processing, translation and post translational processing and modifications, the first level of control is transcription. Functional annotation of the genome starts with identification of the sequence and relative abundance of transcripts expressed in the widest possible range of cell types and tissues.

Transcriptomic technologies to analyse differential gene expression have evolved rapidly over the past 30 years. Differential display (DD) and differential hybridisation (Liang & Pardee, 1992), subtraction of cDNA libraries (Hedrick et al., 1984), suppression subtractive hybridisation (SSH) (Diatchenko et al., 1996), and serial analysis of gene expression (SAGE) (Velculescu et al., 1995), all managed to identify differential gene expression. However, these methods had low reproducibility and low efficiency for investigating the complexity of transcriptomes (Frolov et al., 2003).

Microarray technology, when used for gene expression analysis, is based upon the hybridisation of a probe (cDNA or synthetic oligonucleotide). This probe is complementary to a known gene or gene fragment (immobilised on a solid substrate), to a sample of labelled cDNA which is copied from the RNA of a sample of interest (Lipshutz et al., 1999, Baldi & Hatfield, 2002, Miller & Tang, 2009). They have been used extensively to determine the gene expression profiles of cell and tissue samples, including, relevant to this project, BMDMs from the pig (Kapetanovic et al., 2012), and mouse (Schroder et al., 2012), human MDMs (Martinez et al., 2006) and bovine MDMs (MacHugh et al., 2012). Microarray data has revealed many key signatures associated with the differentiation and activation of macrophages. The consistency of formats, notably the Affymetrix platform, has enabled meta-analysis and aggregation of data from multiple laboratories (Mabbott et al., 2010, Kapetanovic et al., 2013, Mabbott et al., 2013, Regan et al., 2018). Much of the documented evidence of individual and species variation in differential expression
estimates which this study utilises, to enable comparison of species, was derived using microarrays.

Microarrays do have some inherent technical problems, such as cross-hybridisation, non-specific hybridisation and a limit to the range of detection of any given probe/target, as they cannot provide the absolute detection of very highly expressed genes or the sensitivity to detect rarely expressed genes (Draghici et al., 2006). RNA-Seq, does not rely upon any pre-determined sequence and there is no saturation limit, which means that genes with the greatest differential expression can be detected. With sufficient depth of sequencing, RNA-Seq can be superior in sensitivity for detecting low abundance genes/transcripts, and because it is not restrained by the number of target DNA sequences, it is able to detect a much greater variety of transcripts and splice variants.

RNA-Seq technology has largely superseded microarray analysis as it directly sequences RNA using high-throughput sequencing technology (Lister et al., 2008, Mortazavi et al., 2008, Nagalakshmi et al., 2008). Unlike microarrays, RNA-Seq does not depend upon a fully annotated genome for probe selection and so there are no related biases and a whole-genome transcriptome map can be created by aligning the reads to a reference genome, which enables detection of splice variants. This is important as 92-94% human genes undergo alternative splicing, with many of the variants specific to tissues and individuals (Wang et al., 2008). RNA-Seq does however pose challenges for data analysis and storage (Mutz et al., 2013).

The generation of mature mRNA in the nucleus requires pre-mRNA processing. This includes a 5’ cap addition, splicing and polyadenylation (polyA tail) to the 3’ end of the majority of eukaryotic mRNAs. The polyA tail provides stability and has important roles in nucleocytoplasmic export and translation (Moore & Proudfoot, 2009). RNA can be categorised as polyadenylated (polyA+/ polyA) or non-polyadenylated (non-polyA/ polyA) transcripts, depending upon the addition of the polyA tail at the 3’ end. mRNA-Seq involves the enrichment of polyA RNAs by oligo(dT) selection. mRNA-Seq, like total ribodeleted RNA-Seq also provides splicing information (Pan et al., 2008).

The selection for polyA transcripts excludes the detection of transcripts that are not polyadenylated, including functional long transcripts (>200 bp) such as long non-coding RNAs (lncRNAs) (Wilusz et al., 2008), ribosomal RNAs (rRNAs) and replication dependent histone mRNAs (Marzluff et al., 2008, Yang et al., 2011). There are also transcripts which
are bimorphic, which cannot be categorised as either polyA+ or polyA- owing to processing which either reduces or removes the polyA tail, and examples of excised introns that have been identified as a new class of stable non-polyadenylated IncRNAs (Yang et al., 2011). Several IncRNAs have been attributed functions in immunity (Geng & Tan, 2016) and biological processes such as alternative splicing (Yin et al., 2012), transcriptional regulation (Zhang et al., 2013), and microRNA regulation (Hansen et al., 2013, Memczak et al., 2013). The functions may include acting as decoys, scaffolds or enhancer RNAs that contribute to chromatin rearrangement (Fang & Fullwood, 2016). Despite recent advances, most IncRNA functions remain to be determined (http://fantom.gsc.riken.jp/), owing to the multiple mechanisms and complex dynamics involved in IncRNA expression involving chromatin state, core transcription factors and the post transcriptional regulation involving microRNAs (Wu et al., 2014). Many non-polyadenylated IncRNAs are lowly expressed yet highly tissue- and species- specific, with poor evolutionary conservation. The processing of non-polyadenylated IncRNAs is closely linked to splicing of the gene from which they were derived (Zhang et al., 2014). IncRNAs which are conserved across species can provide important information regarding their functional roles (Li & Yang, 2017, Bush et al., 2018). Therefore, to improve understanding of the regulatory network governing expression of the sheep genome, this project analyses both the polyA and non-polyadenylated transcriptional landscape of the sheep.

RNA-Seq does have certain limitations in assembling IncRNA transcripts (Kukurba & Montgomery, 2015), as the assembled transcriptome includes partial transcripts and artefacts due to fragment bias (due to non-uniformity of cDNA fragments within the transcripts they represent) (Roberts et al., 2011). The short sequence fragments generated in RNA-Seq (Trapnell et al., 2009) mean that for those genes that are expressed at a very low level, there may be gaps in their assembly which can lead to mistaken annotation as IncRNAs (Kozarewa et al., 2009, Sun et al., 2012).

This project is utilising both polyA-selected library preparation protocols and ribodepleted protocols to enable the capture of all types of transcripts. The reads are 125bp in length which provides a balance between accuracy in sequencing, splicing detection and cost (Chhangawala et al., 2015). All reads were paired-end which enables both ends of the fragment to be aligned, improving the quality of detection of genomic rearrangements, repetitive sequence elements, splice variants and novel transcripts.
As described in Chapter 2 and shown in Figure 2.2, two pipelines were used to map transcripts and assign transcript and gene expression values. Kallisto2 (Kallisto) rapidly and accurately pseudo-aligns reads to a reference transcriptome, providing accurate expression estimates for each reference transcript in OarV3.1. The StringTie pipeline utilises more computing power, time and two methods of precise read alignment (HISAT2 and BOWTIE2) to map reads to the reference genome, then StringTie assembles these alignments enabling gene expression estimates to be determined, as well as revealing novel transcript models. With novelty comes potential errors in those transcripts which are very lowly expressed, so both pipelines have been utilised to maximise accurate analysis of the reference transcriptome as well as provide information on novel transcript expression. Every sample submitted was processed using both pipelines and the resulting expression estimates are examined in Chapters 4 and 5. In addition, two library preparation methods were used to capture either polyadenylated RNA (25M read depth), or total RNA depleted for rRNA (100M read depth) for both LPS stimulated and unstimulated BMDM samples from each individual.

This chapter compares the two preparation methods and the two analysis pipelines and finishes with a very preliminary examination of the repeat regions (reads not utilised by Kallisto) which contribute to the ‘dark matter’ of genomic sequence that currently has unknown function within the sheep genome. How much of this ‘dark matter’ is junk or has unknown function is under scientific debate. The evolutionary significance for ‘dark matter’ is that it partly accounts for the deleterious mutational load in a genome, with the functional fraction within the human genome unable to exceed 25% (Graur, 2017). The Encyclopedia of DNA Elements Project (ENCODE) (Consortium, 2012) has revealed that biochemically active regions cover a much larger fraction of the genome than evolutionarily conserved regions, with chromatin profiling (Ernst et al., 2011, Gerstein et al., 2012, Trynka et al., 2013) and many GWAS studies showing human disease susceptibility loci to lie outside protein coding regions (Kleinjan & van Heyningen, 2005, Maurano et al., 2012, Kellis et al., 2014).

3.2 Comparison of the outputs of Kallisto2 and StringTie Pipelines

Kallisto is a fast, efficient and accurate program (Bray et al., 2016), used for quantifying abundances of transcripts from RNA-Seq data of known target sequences from a reference transcriptome (see Chapter 2.8.1). The expression level estimate for each reference
transcript from OarV3.1.81 is provided as transcripts per million (TPM), a measurement of the proportion of transcripts in the RNA sample pool.

For the purpose of improved annotation of the reference sheep genome, to identify novel transcripts, splice variants and intron/exon boundary information, a second pipeline StringTie (Pertea et al., 2016) utilised a two pass approach (see Chapter 2.8.2) to alignment, making use of HISAT2 as the priority aligner, with the relatively stringent, rapid and accurate alignment to the reference (mapping 73-86% of raw reads, see Table 3.1). The unmapped reads from HISAT2 were passed to BOWTIE2, another aligner which has a local alignment tool to ‘soft clip’ the ends of the read that maybe inaccurate. Based upon this tool, previously unmapped HISAT2 reads were ‘rescued’ and utilised in the analysis. With the combined approach, >99% of the reads were used to inform further analysis, as demonstrated in Table 3.1. Mapping is primed from the 5’ end so any adaptor contamination at the 3’ end can be trimmed by the alignment step. StringTie creates a list of all gene transcripts identified and the expression level estimates are provided as Fragments per kilobase of exon per million reads mapped (FPKM). There are multiple aligners available and it is acknowledged that this is a fast-developing field.

Table 3.1 is a simplified version of Appendix 3.1 and shows that 55-68% of all raw reads were used in estimating transcript abundance (including coding and non-coding reference transcripts) using Kallisto, compared to over 99% of all raw reads being mapped and utilised by the StringTie pipeline. The remaining 32-45% of reads that were not utilised by Kallisto are further explored and explained in Section 3.3.
Table 3.1 Comparative analysis of alignment rates for different analysis pipelines
This table compares pseudo-alignment rates from Kallisto2 (Kallisto) with the average StringTie alignment rates after merging HISAT2 and Bowtie2 for each sample

<table>
<thead>
<tr>
<th></th>
<th>AVG % Pseudoaligned with Kallisto2</th>
<th>AVG % Aligned with HISAT2 alone</th>
<th>AVG % Aligned after HISAT2 and BOWTIE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0hr- PolyA/mRNA</td>
<td>67.4</td>
<td>84.4</td>
<td>99.3</td>
</tr>
<tr>
<td>2hr- PolyA/mRNA</td>
<td>55.7</td>
<td>78.9</td>
<td>99.1</td>
</tr>
<tr>
<td>4hr- PolyA/mRNA</td>
<td>64.7</td>
<td>79.0</td>
<td>99.4</td>
</tr>
<tr>
<td>7hr- PolyA/mRNA</td>
<td>66.2</td>
<td>85.7</td>
<td>99.4</td>
</tr>
<tr>
<td>24hr- PolyA/mRNA</td>
<td>65.6</td>
<td>80.6</td>
<td>99.4</td>
</tr>
<tr>
<td>0hr- ribodepleted/ TOTAL RNA</td>
<td>60.8</td>
<td>75.4</td>
<td>99.4</td>
</tr>
<tr>
<td>7hr- ribodepleted/ TOTAL RNA</td>
<td>59.1</td>
<td>73.6</td>
<td>99.2</td>
</tr>
<tr>
<td>PBMC-PolyA/mRNA</td>
<td>54.9</td>
<td>76.6</td>
<td>99.2</td>
</tr>
<tr>
<td>AM- PolyA/mRNA</td>
<td>66.1</td>
<td>80.0</td>
<td>99.4</td>
</tr>
<tr>
<td>MDM- PolyA/mRNA</td>
<td>65.7</td>
<td>77.9</td>
<td>99.2</td>
</tr>
<tr>
<td>BL- PolyA/mRNA</td>
<td>57.0</td>
<td>86.3</td>
<td>99.5</td>
</tr>
</tbody>
</table>

From the earliest gene discovery efforts by the FANTOM consortium, mouse macrophages have been shown to have a very complex transcriptome with many novel mRNAs (Wells et al., 2003, Baillie et al., 2017). Analysis of the sheep macrophage transcriptome revealed the same complexity. The reference transcriptome in the sheep is primarily focussed on the protein-coding transcriptome, and the vast majority of genes are represented by a single transcript. The Kallisto pipeline based upon polyA+ mRNA identified 23,133 transcripts from amongst 28,757 reference transcripts (80.4%) having expression estimates > 0 TPM in any individual in the sheep macrophage transcriptome. Among the total RNA
libraries (0hr and 7hr) with the greater sequencing depth, 23828 transcripts (82.6%) were found to have expression estimates > 0 TPM in at least one individual.

Given the diversity of the protein-coding transcriptome, there was an expectation that macrophages would also be a rich source of non-coding RNAs. One class of IncRNAs is derived from the transcription of enhancers. The stimulation of mouse macrophages with LPS was associated with transcription of more than 3,000 enhancer RNAs (Kaikkonen et al., 2013). Some, but not all, of these transcripts are likely to be polyadenylated but most are unspliced. StringTie is a discovery pipeline and discovery/annotation was the rationale for sequencing the total RNA libraries using the key samples (resting 0hr and maximal response 7hr LPS samples) at a greater depth of 100 million reads/sample. This method does not positively select for polyadenylation so captures non-polyadenylated protein coding and noncoding transcripts. The greater depth of sequencing detects the rarer and lower expressed transcripts with more coverage and confidence from a wide range of RNAs.

The total number of transcript models identified by StringTie for the PolyA immune cell samples alone was 91,535, more than three times the number of transcripts currently recognised in the reference transcriptome OarV3.1.81 (Li et al., 2015) that Kallisto utilised, primarily due to the multiple splice variants that StringTie is able to detect. Fig 3.1 summarises the outcomes of the two independent pipelines. Each pipeline produces a final expression level estimate (TPM for Kallisto and FPKM for StringTie), for every transcript within the overall index; Kallisto utilising purely the reference transcripts from OarV3.1.81; whereas StringTie created an index of non-redundant transcript models across all samples and time points, merging all models created using HISAT2 and BOWTIE2, as well as the reference transcriptome from OarV3.1.81. Examining the maximum transcript expression across all individuals and BMDM response samples, for every transcript within each respective index enables a comparison of the two pipelines and it is possible to visualise how many more transcript models are robustly detected through using the StringTie pipeline.

Mammalian transcriptomes display a power-law distribution in transcript abundance; in simple terms there are few very highly expressed and many lowly expressed genes and a log transformation produces a straight line (Ueda et al., 2004). Figure 3.1 shows the distribution of expression estimates for the two pipelines separated into classes of abundance. The analysis shows that at expression level cut offs (FPKM >5 for StringTie and
TPM >1 for Kallisto), the power law relationship fails. Below this level of expression, there is greater noise because the level of coverage from RNA-Seq is insufficient to overcome the stochastic sampling effects.

Figure 3.1 The distribution of maximum transcript expression estimates for StringTie (FPKM) and Kallisto (TPM) pipelines
The numbers of transcripts with different maximum expression levels across all time points in the polyA-selected time course (0hr, 2hr, 4hr, 7hr and 24hr) over all individuals were compared to highlight the differences between the results from Kallisto and StringTie. The X-axis shows the different expression level (TPM/FPKM) class boundaries; (0,0-1, 1-5, 5-20, 20-100, 100-500, > 500). The Y axis shows the number of transcripts expressing a maximum expression estimate (the maximum expression detected in any individual or sample) within that expression boundary.

For every merged HISAT2 and Bowtie2 BAM file, SAMtools flagstat was used to provide some simple statistics including information on how many reads were included in each file, the pairing of these reads and how these pairs were mapped (see Appendix 3.2). 77-90% (Average 85%) of all paired reads were properly paired across all the samples.

To cross validate these two pipelines which estimate gene expression, the sum of mean transcript expression estimates for each reference gene using Kallisto (TPM) was compared to the mean reference gene expression estimates obtained using StringTie version 1.3.3 (FPKM), using the GffCompare utility which is part of the Cufflinks/Tuxedo suite (Pertea et
The expression estimates from StringTie for each reference transcript, had to merge the expression assigned to novel transcript models that were found for many of the reference OarV3.1 genes (such as CSF1R, see Ch.2.8.2). All expression estimates had 0.00000001 added prior to the Log10 calculation, to avoid any divisions of 0, and both the Pearson’s product-moment correlation (r) and Spearman’s rank-order correlation (Rho) were calculated, removing all expression estimates of 0 prior to calculating the Log 10 correlations r and Rho.

As noted above in Figure 3.1, and below in the scatterplots, consistent with the known power law relationships of transcript abundance observed in many other systems (Ueda et al., 2004, Lu & King, 2009), a relatively small number of transcripts contributed the majority of reads in a sample. There was minimal differential expression detected in some of the most abundant protein coding transcripts which appear as outliers in both basal (0hr LPS) and peak (7hr LPS) activation expression profiles. These outliers include SPP1, GSTCD and protein coding genes lacking any functional annotation such as ENSOARG00000014339 (subsequently annotated as Ferritin heavy chain FTH1).

Figures 3.2 and 3.3 compare the different pipeline expression estimates for 0hr and 7hr LPS treated BMDMs respectively. All plots show a positive correlation between the expression estimates derived from StringTie (FPKM) and Kallisto2 (TPM), with most reference genes from OarV3.1 demonstrating similar expression profiles, irrespective of the pipeline used to obtain them.

It is interesting to note that despite the similar broad appearance of the correlation profiles between the plots, Pearson’s product-moment correlation (r) differs in value quite dramatically between the different plots, and is more stringent when compared to Rho. This is because of the numbers of transcripts depicted in each plot, the majority of which are not visible and Pearson correlation evaluating the linear relationship between the two pipeline expression estimates, treating the estimates as continuous, which is in contrast to the Spearman correlation which is based on ranked expression values for each pipeline rather than the transcript expression levels themselves.

The disparity in completeness of the reference transcriptome between the protein coding transcriptome (selected by PolyA libraries) and the total ribodepleted transcriptome has also clearly had an effect when trying to compare the two pipelines.
Figure 3.2 Correlation of Kallisto and StringTie gene expression estimates of resting BMDMs (at 0 hr LPS)
Using A: PolyA selected libraries and B: Ribodepleted total libraries, i) and ii) show correlation plots between Kallisto (tpm) and StringTie (FPKM) for i) the average (mean) and ii) Log10 (mean + 0.00000001) expression of Ensembl ID genes across all six individuals for 0hr LPS samples for both library preparation methods. The genes are coloured according to gene type (assigned using Biomart) according to the key C. The Pearson’s product-moment correlation is provided as r, and Spearman’s rank-order correlation is provided as Rho, to provide numerical strength of the association and link respectively (for Log10 plots, all expression estimates of 0 were removed).
Figure 3.3 Correlation of Kallisto and StringTie gene expression estimates of peak response BMDMs (at 7 hr LPS)
Using A: PolyA selected libraries and B: Ribodepleted total libraries. i) and ii) shows correlation plots between Kallisto (tpm) and StringTie (FPKM) for i) the average (mean) and ii) Log10 (mean + 0.00000001) expression of Ensembl ID’s across all six individuals for 7hr LPS samples for both library preparation methods. The genes are coloured according to gene type (assigned using Biomart) according to the key C. The Pearson’s product-moment correlation is provided as r, and Spearman’s rank-order correlation is provided as Rho, to provide numerical strength of the association and link respectively (for Log10 plots, all expression estimates of 0 were removed).
In the log_{10} transformed plots of Figures 3.2 and 3.3, irrespective of library type or activation state StringTie was generally better able to assign expression of protein-coding genes and pseudogenes (represented by purple and red dots producing a ‘hump’ to the left of line of best fit), whilst Kallisto appears to assign greater expression estimates for lincRNAs and miRNAs (represented by navy blue and yellow spots producing a ‘hump’ to the right of line of best fit). Differences in the majority of protein-coding genes are likely to be attributable to StringTie’s ability to detect and assign expression to multiple novel transcript models for those given genes, as Kallisto is constrained to assigning expression only to the reference transcript models, which may not be the most highly expressed model, as demonstrated by Interleukin 1 beta (IL1B) (see Chapter 5.4.7). With lincRNAs and miRNAs generally representing shorter length transcripts, the accuracy of Kallisto in identifying these small transcript lengths explains why Kallisto is better able to assign expression estimates as opposed to StringTie, which associates expression of these transcripts with other transcript models.

Both pipelines identified the C-X-C motif chemokine ligand 8 (CXCL8), highlighted in Figure 3.3.A.i and B.i as IL-8, to be highly expressed in the peak activation samples. IL-8 is a granulocyte chemoattractant and is not present in rodent genomes but is one of the most highly LPS-induced transcripts in human, pig and horse macrophages (Nieto et al., 2009, Kapetanovic et al., 2012, Baillie et al., 2017). StringTie identified 16 different transcript models. The most similar reference transcript model was not the most highly-expressed model (MSTRG.27919.16). Kallisto oversimplifies the impression of gene level expression since most genes have several transcript models. Nevertheless, the pseudoalignment has actually provided a relatively accurate level of gene expression, for which functional annotation is generally sufficient. StringTie has provided gene level expression estimates and highlighted possible alternative splice isoforms. These novel transcript models can therefore improve upon the current reference OarV3.1.

There are significant differences between the pipelines. StringTie has assigned high expression to Glutathione S-transferase C-terminal domain containing (GSTCD) in all samples, assigning a mean expression of 4962 FPKM, whereas Kallisto assigned a mean expression of only 1.6 TPM across all samples. On closer inspection this is because Gff Cuffmerge has merged both MSTRG.27343, which has the same transcript model as ENSOART00000010840 (GSTCD) containing 11 exons, and MSTRG.27345 which is a novel
transcript model containing only one exon. Therefore, in some cases Kallisto is providing the more accurate expression of reference gene expression. Neither humans, mice or pig macrophages express GSTCD to meaningful levels (Schroder et al., 2012, Kapetanovic et al., 2013, Baillie et al., 2017). It is therefore worth curating the specific cases where the two pipelines differ.

In other protein-coding genes, Kallisto has assigned significant expression where StringTie does not. One example includes integrin subunit alpha M (ITGAM), highlighted in Fig 3.3.A.ii, which was assigned a mean expression of 204 TPM by Kallisto whereas StringTie only assigned a mean of 4 FPKM. This gene encodes for the CD11b integrin which is known to combine with ITGB2 to form macrophage receptor 1 (Mac-1), important in macrophage adhesive interactions (Fagerholm et al., 2006), and therefore robust transcript expression was expected in sheep BMDMs. ITGAM only has one, complex transcript model in OarV3.1, with 31 exons, unlike the 1:1 orthologue with humans (with highly conserved regions) where 8 transcript variants have been identified, varying in length and numbers of exons (Aken et al., 2016). ITGAM is situated within a particularly complex region of chromosome 24, with many genes with equally complex transcript models such as ITGAX (identified as having 16 transcript models with up to 58 exons, assigned a mean expression of 160 FPKM), and ITGAD (3 transcript models with 31 or 14 exons, assigned an expression of 11 FPKM). The StringTie algorithm was not efficient in accurately discriminating between these transcript and gene models.

ENSOARG00000019602/ ENSOART00000021345, annotated as heat shock factor-binding protein 1 (HSBP1) is highlighted in Fig 3.2.A.ii, and represents another example of disparate assignment of expression between the two pipelines. Kallisto assigned a mean expression of 524 TPM across all PolyA-selected LPS response samples, where StringTie did not detect any expression of HSBP1. On closer examination of the region in IGV, there is no coverage over the specified location (Chr 2: 125,821,048-125,821,259). OarV3.1 shows only one transcript model, which is 204 base pairs in length, containing only one exon, however there is virtually no synteny when the region is compared with mice or humans using Ensembl. Both mice and humans have multiple transcript variants for HSBP1 with 2, 3 or 4 exons (Li et al., 2015). This highlights the need for two independent pipelines to examine gene expression in these samples, as the reference transcriptome is not always correct. It is
likely that the new assembly of sheep, and related high resolution goat assembly will provide greater clarity.

As expected, the greatest proportion of ‘novel’ transcript models came from the ribodepleted total RNA samples which were sequenced at the greatest depth. The novel models detected at very low expression levels, with low coverage, are less likely to be reproducible, which explains the apparent divergence between expression estimates derived from each pipeline in the lowly expressed genes.

3.3 Identification of differentially-expressed transcripts

Due to the biological variation that exists between individuals, the random sampling intrinsic to RNA-Seq, possible technical variation between RNA-Seq sequencing library preps (Oshlack et al., 2010), and the multiple sampling that is intrinsic to genome-scale analysis, determining what constitutes a “significant” change in transcript abundance is not straight-forward.

The perfect RNA-Seq experiment would identify the counts of reads to be proportional to the abundance of transcripts that they originate from and follow an approximate Poisson distribution with an error rate that decreases as the depth of sequencing is increased (Bullard et al., 2010). Pseudogenes and gene families have very similar sequences which creates a problem for Kallisto as allocating ‘counts of reads’ to a specific transcript or gene can be unreliable. If genomic features have been considered, as with StringTie, reads are no longer ambiguous, technical replicates are unnecessary and the focus is entirely on modelling the biological variation between samples (Trapnell et al., 2010). The process of estimating transcript level counts with methods that do utilise precise mapping can still introduce some technical variation (Pimentel et al., 2017).

Sleuth is a program that was created to bridge the gap between count-based methods and quantification algorithms by fully exploiting the advantages of both (Pimentel et al., 2017). Sleuth utilises the bootstraps of Kallisto2 which serve as proxies for teasing apart the technical variance from the biological variance, enabling a “true” estimate of the biological variance. In simulated data Sleuth appeared to outperform other ‘count based’ statistical packages such as DEseq2, edgeR and voom (Pimentel et al., 2017).

Ballgown is a program able to take output files from the StringTie pipeline and implement a basic linear modelling strategy to rigorously apply statistical tests to determine differential
expression at the exon, gene or transcript level. FPKM values are skewed so the variance is stabilized by the programme applying a log transformation and then fitting a standard linear model. At the time of writing this thesis there was no better statistical software that allows the same level of flexibility for modelling transcript-level data (Frazee et al., 2015). Transcripts expressed at lower levels are more difficult to assemble and spurious models can produce misleading apparently significant differences (Bush et al., 2018). There has not been any pre-filtering of transcripts prior to running Ballgown or Sleuth, so the p and q values must be qualified by looking at the estimated expression levels of that given transcript or gene model. Segregation of counts amongst different transcript models reduces the statistical power; so it is intrinsically easier to demonstrate differential expression at the gene, rather than the transcript level (Pertea et al., 2016).

Both Sleuth and Ballgown provide a p value and q value indicating the accuracy of the estimates of differential expression, and can account for confounding variables, which for the purpose of this project included sex when focusing on LPS time point. The null hypothesis is that transcripts do not change in expression between time points. The p value is the probability of finding a given transcript to be differentially expressed when it is in fact not, it takes into account the mean expression and sample size and relies upon establishing the size of the difference in expression in comparison to the variance to estimate significance. A biologically meaningful value for this significance cut off is still the subject of debate (Fatovich & Phillips, 2017). With a p value of 5%, multiple tests on the same sample as with RNA-Seq can result in a large number of false positives, therefore the p value needs to be adjusted. The q value takes into account the false discovery rate (FDR), and Ballgown and Sleuth both utilise a false discovery rate (FDR) to account for multiple testing (q value). The q value controls the estimated proportion of false positives among the significant results. It is therefore more stringent than the uncorrected p value but directly correlated with it. When ranking the significance of a given set of differentially expressed transcripts, the order remains the same regardless of whether p or q values are used.

The two approaches to identify significant differences are difficult to compare as Sleuth examines Kallisto results at the transcript level of expression, whilst Ballgown examines StringTie gene level expression. Miru does not assign a p or q value to an individual node (transcript or gene), or account for confounding variables, however the clustering is entirely based on valid statistical methods, algorithms and matrices (discussed in 2.9) and provides
a more useful visual approach to extracting biologically meaningful transcripts and genes. For completeness, the respective p and q values derived using either Sleuth for the Kallisto transcript expression networks or Ballgown for the StringTie gene expression networks have been assigned to the final ‘peaking analyses’ workbooks in Chapter 5.

3.4 Ribodepleted Total RNA compared to PolyA-selected mRNA libraries

For the LPS time course, there were two methods of library preparation utilised by Edinburgh Genomics (see chapter 2.7.1). For all individuals, all five LPS timepoints were used to generate PolyA-selected libraries and paired end reads at 25M read depth were used to analyse mRNA expression. The 0hr and 7hr samples from each individual were also subjected to the alternative ribodepleted (total) RNA protocol and sequenced with paired end reads at 100M read depth. The same approach was used to analyse a subset of libraries in the wider sheep transcriptional atlas (Clark et al., 2017)

Total RNA libraries removed the abundant rRNAs through hybridization capture of rRNA followed by binding to magnetic beads. The better the efficiency in removing rRNA, the greater the numbers of reads that will map to transcripts of interest, as rRNAs make up >80-90% of the total RNA of an un-enriched sample (O'Neil et al., 2013). As shown in a detailed comparative analysis of HEK293 cells, the ribominus protocol enriches for detection of non-poly-adenylated transcripts, but also captures introns from unprocessed nuclear transcripts (Sultan et al., 2014). These can potentially be merged for expression estimates, but there is an additional benefit for the future study of allele-specific expression, and specific regulatory variants (single nucleotide variants; SNVs), since intronic transcripts are greatly enriched for SNVs (Sale et al., 2007, Kumar et al., 2017) and it is known that a SNV may profoundly alter transcription factor (TF) binding sites and significantly contribute to regulatory variation within the genome (Kumar et al., 2017).

The differential detection of transcripts in the two library methods clearly has an impact on gene expression estimates. Figure 3.4 shows a principal component analysis (PCA) plot for all of the libraries from the LPS time course. The scale of the response to LPS is evident from the clear impact of time as a variable (detected by PC2), but library type also separates samples from the same time points (0 and 7 hours) (PC1).
Figure 3.4 PCA to demonstrate library type differences
PCA plot using transcript expression data derived from Kallisto2 for all LPS samples from both polyA-selected libraries and ribodepleted total libraries, depicted as circles and squares respectively. The points are coloured according to LPS time point with 0hr being turquoise, 2 hr yellow, 4 hr orange, 7hr red and 24 hr purple. The percentage of variance is shown in brackets for each principal component.

In Chapters 4 and 5, the results from the two library types were separated for analysis and utilised independently of one another. The PolyA-selected mRNA time course was used to examine the global transcriptional changes that take place when sheep BMDM respond to LPS with the focus on polyadenylated mRNAs which will be destined for translation. The ribodepleted total RNA time points from 0hr and 7hr at the greater sequencing depth remain to be explored more thoroughly. Due to the increased costs involved in sequencing ribodepleted total RNA samples at the greater sequencing depth, it was not possible to do this sequencing for all the time points.
3.5 Differential detection of transcripts with different library preparation

Figures 3.5 and 3.6 show scatterplots comparing the expression detected by the polyA plus and ribodepleted libraries, analysed using both the Kallisto (TPM) and StringTie (FPKM) pipelines. As demonstrated in Figures 3.5 and 3.6, despite the difference in sequencing depth, the vast majority of genes demonstrate a positive correlation in expression estimate, irrespective of library method utilised. However, neither selection protocol was 100% efficient at removing undesired sequences and retaining desired sequences. Known abundant non-poly-adenylated transcripts were detectable in the polyA-selected libraries despite the lower depth of sequencing.
Figure 3.5 Scatter plots to demonstrate correlation of gene expression estimates at 0hr LPS, detected by different library types
A: Using StringTie and B: Using Kallisto, i) and ii) shows the correlation plots between polyA selected and ribodepleted total library types for i) the average (mean) and ii) Log10 (mean + 0.0000001) expression of Ensembl ID’s across all six individuals for 0hr LPS samples for both pipelines. The genes are coloured according to gene type (assigned using Biomart) according to the key C. The Pearson’s product-moment correlation is provided as r, and Spearman’s rank-order correlation is provided as Rho, to provide numerical strength of the association and link respectively (for Log10 plots, all expression estimates of 0 were removed).
**Figure 3.6** Scatter plots to demonstrate correlation of gene expression estimates at 7hr LPS, detected by different library types

A: Using StringTie and B: Using Kallisto, i) and ii) shows the correlation plots between polyA selected and ribodepleted total library types for i) the average (mean) and ii) Log10 (mean + 0.00000001) expression of Ensembl ID’s across all six individuals for 7hr LPS samples for both pipelines. The genes are coloured according to gene type (assigned using Biomart) according to the key C. The Pearson’s product-moment correlation is provided as \( r \), and Spearman’s rank-order correlation is provided as Rho, to provide numerical strength of the association and link respectively (for Log10 plots, all expression estimates of 0 were removed).
Figure 3.7 utilises data derived from Kallisto and shows a 92% overlap in the reference transcripts (OarV3.1.81) that were detected with expression > 0 TPM that are detected for each of the library types. Table 3.2 expands upon this information to show the breakdown of detection of different types of reference transcript in each library type, with the greatest difference to be in the detection of protein-coding transcripts for the polyA-selected libraries, and for the ribodepleted total libraries in the detection of miRNA, snoRNA, snRNA and to a lesser extent rRNA.

The majority of transcripts detected that were selectively enriched by the ribo-minus protocol, were either snoRNAs or snRNAs. Most of the protein-coding genes/transcripts within this group required functional annotation and were found to be histone mRNAs which are probably not polyadenylated (Dávila López & Samuelsson, 2008, Marzluff et al., 2008). All other protein coding genes over-represented in in the ribodepleted total libraries, with reliable levels of detection (some of which had associated annotation), were not significantly LPS responsive or macrophage specific and therefore not relevant to the analysis of transcriptomic changes associated with macrophage biology, the subject of this project.

**Figure 3.7 Venn diagram of all OarV3.1 transcripts detected using Kallisto, for each of the library types used for 0hr and 7hr LPS samples.**

Transcripts with a Kallisto expression estimate (TPM >0), detected in any BMDM sample (either 0hr or 7hr) in the polyA-selected and ribodepleted total library types. The percentages indicate the proportion of transcripts within the reference transcriptome OarV3.1.81.
Table 3.2 Numbers and proportions (%) of the different types of transcript detected in each of the library types.
The table shows the numbers and proportions of different types of transcript that were detected in each of the library types (polyA-selected or ribodepleted total), which showed a Kallisto expression > 0 TPM in any individual for either 0hr or 7hr LPS sample.

<table>
<thead>
<tr>
<th>Transcript Type</th>
<th>In PolyA, not total</th>
<th>In total, not PolyA</th>
<th>In polyA, not total (%)</th>
<th>In total, not polyA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lincRNA</td>
<td>33</td>
<td>58</td>
<td>5.3%</td>
<td>4.4%</td>
</tr>
<tr>
<td>miRNA</td>
<td>26</td>
<td>196</td>
<td>4.1%</td>
<td>14.8%</td>
</tr>
<tr>
<td>processed_pseudogene</td>
<td>1</td>
<td>0</td>
<td>0.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>protein_coding</td>
<td>506</td>
<td>754</td>
<td>80.6%</td>
<td>57.0%</td>
</tr>
<tr>
<td>Pseudogene</td>
<td>8</td>
<td>22</td>
<td>1.3%</td>
<td>1.7%</td>
</tr>
<tr>
<td>rRNA</td>
<td>12</td>
<td>44</td>
<td>1.9%</td>
<td>3.3%</td>
</tr>
<tr>
<td>snoRNA</td>
<td>15</td>
<td>70</td>
<td>2.4%</td>
<td>5.3%</td>
</tr>
<tr>
<td>snRNA</td>
<td>27</td>
<td>179</td>
<td>4.3%</td>
<td>13.5%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>628</td>
<td>1323</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.6 Examination of repeat regions within genome

More than two thirds of the human genome consists of repetitive elements or repeat regions (de Koning et al., 2011), with debate as to how much of this genomic sequence has function. Transposable elements (TEs), colloquially known as “jumping genes” are a major category of repeat regions that can move about both within and more rarely between genomes. Retrotransposons (Class I transposable elements) represent an important and biologically fascinating type of these mobile genetic elements and are known to generate insertion mutations, cause genomic instability, alter gene expression and contribute to genetic innovation (Cordaux & Batzer, 2009).

In humans and mice, retrotransposons and other repeat elements form a significant component of the transcriptome (Faulkner et al., 2009). Despite their potentially deleterious effects, by increasing genetic diversity within a species they can increase adaptability under new selective pressures, as well as benefit the individual (Upton et al., 2011). It is an exciting, complex and largely unknown involvement that they have had in the genomic changes that have shaped the sheep species and variety of breeds that we recognise today. The domestication of sheep since 11000 BP represents a crucial step in human civilisation and studies utilising endogenous retroviruses as genetic markers have been able to distinguish primitive breeds from modern ones which has helped preserve
Retroviruses are single stranded RNA viruses that encode the enzyme reverse transcriptase (RT) which converts RNA into DNA during the life cycle of the virus, which then becomes integrated into the host’s genome (Temin & Mizutani, 1970), and they can be transformed into a retrotransposon if the domains that enable mobilisation are inactivated or deleted. Transcription and reverse transcription is shared by other viruses and with many moveable genetic elements such as retrotransposons, with over 10% of the human genome being composed of sequences resulting from reverse transcription (retrosequences) (Temin, 1989, Temin, 1992). The retroviridae family have a wide range of hosts, with extensive molecular and genomic diversity. Their genome structure is shown in Fig 3.8 (Gifford & Tristem, 2003).

**Figure 3.8 Retrovirus genome structure**
Simplified from (Gifford & Tristem, 2003) which shows the four essential components/sequence domains: 1) Gag, encoding the viral capsid, matrix and nucleocapsid proteins; 2) pro, encoding the viral protease; 3) pol, which encodes the enzymes reverse transcriptase and integrase; 4) env which encodes the surface and transmembrane glycoproteins. At either end of the DNA form of the retroviral genome are long terminal repeat (LTR) sequences. Accessory genes determine the complexity of the viral genome.

Retroviral infection involving somatic cells means that the virus can be transmitted to uninfected hosts as “exogenous” retroviruses. When germ cells are infected the viral genome can become part of the host genome as an “endogenous” retrovirus (ERV). If movement between cells is not permitted (not infectious), the insertion of new copies into the genome is restricted to that same cell, thus relying upon germ cell infection to enable vertical transmission by subsequent generations (Naville et al., 2016).
Retrotransposons can be divided into three groups:

1) Long terminal repeats (LTRs): If they have LTRs and encode reverse transcriptase (like retroviruses). This group employ a ‘copy and paste’ mechanism to transpose themselves.

2) Long interspersed nuclear elements (LINEs, L1s): If they encode reverse transcriptase but lack LTRs and are transcribed by RNA polymerase II.

3) Short interspersed nuclear elements (SINEs): If they do not encode reverse transcriptase and are transcribed by RNA polymerase III.

The bovine genome, like that of mice and humans has been found to contain many of the same eutherian mammal repeats, in addition to long interspersed nuclear element RTE (BovB) elements that have been suggested to derive horizontally from squamata and some areas of the genome, identified by repeat regions have defined both ancestral vs ruminant-specific genomic regions (Adelson et al., 2009). In ruminant animals the most common SINE is Bov-A2, and it has been suggested that genomic instability in bovine genomes can produce extensive unequal crossing over of Bov-A2 and that repeated outbreeding may contribute to levels of polymorphism (Onami et al., 2007). There is large variation of Bov-A2 sequences, partly due to site specific micro-recombination followed by gene conversion, and short cDNAs copied by the RT have been suspected of being involved in the somatic hypermutation (SHM) process in the hypervariable regions of immunoglobulin and major histocompatibility complex (MHC) genes. Bov-A2 has also been found in the non-coding regions of several genes responsive to environmental stress, and it has been suggested that this sequence may play a role in the post-transcriptional regulation of gene expression as the “core” sequence has been identified at high levels in lymphocytes only after their activation (Damiani et al., 2008).

Genomes have evolved defence mechanisms from these mobile “retro” elements and retrotransposons. Cytosine methylation, small RNAs and apolipoprotein B mRNA editing enzyme, catalytic polypeptide family (APOBECs) may all play a role in protection against infectious retroviruses (Fanning, 2011). Most ERVs have accumulated mutations which leave them unable to produce viral particles but some have managed to keep their open reading frames (ORFs) and are utilised by the host to fulfil important biological functions, participating in gene regulation, cell growth and the immune response (Stoye, 2012). John
Bittner (Bittner, 1936) and Ludwik Gross (Gross, 1951) used mice to first show that retroviruses can cause neoplastic disease, and they are involved in many diseases such as cancers (leukaemias, mammary carcinomas), immunodeficiencies and arthritis (Vogt, 1997). The interesting exception is the Spumavirus class, which can persistently infect wild non-human primates, felines, bovines, equines and small ruminants, with no apparent pathological consequences (Murray et al., 2008).

A preliminary investigation has focussed on exploring the LTR group, as this group is suspected of accounting for the significant number of reads that Kallisto2 was unable to pseudoalign to the reference transcriptome. Both retroviruses and long terminal repeat (LTR) retrotransposons are flanked by LTR regions in direct orientation that are necessary for transcription and integration into the genome, with the main difference between them being the presence of the envelope gene.

Repeat regions in the sheep genome were identified using the UCSC genome browser for OarV3.1. Within the toolset of UCSC, the tablebrowser function can select the group concerning all variation and repeats and then specifically the repeat masker track within the reference sheep genome, requesting the specific fields: genomic sequence name (genoName), the start in genomic sequence (genoStart), the end in genomic sequence (genoEnd), the relative orientation of the strand (+ or - strand), the name of repeat (repName), the class of repeat (repClass) and the family of repeat (repFamily). From this extensive list, 396,061 locations with the LTR class of repeat (Group 1 above) were selected and the co-ordinates extracted, producing regions that have been examined in more detail using the ribodepleted 0hr and 7hr data from each sheep. The locations mapping to scaffolds were ignored and the file formatted to produce a bed file (the start location for each region begins at 0 so each start location had -1 deducted) so BEDtools could be used to count the reads overlapping by >1bp to these regions of interest from the selected BAM files of interest (specifically the 0hr and 7hr total RNA samples from each individual). After collating all the counts for each set of co-ordinates for all 0hr and 7hr ribodepleted total RNA alignment files for each individual, the TPM for each co-ordinate were determined utilising the SAMtools flagstat statistic on how many reads were in that particular alignment file. Using the pivot table function in Excel, it was then possible to examine the various names of repeat (246) and families (6) of repeat, both across all individuals as a group and then separately.
As an LTR group, differences in proportions of raw reads mapping to regions of known ERVs as a whole between 0hr and 7hr was negligible, with a Welch two sample t-test identifying a P value of 0.451, and a paired t-test P value of 0.3302. Examining the LTR group across all individuals, there was individual variation in the proportions of reads mapping at 0hr LPS compared to 7hr LPS. 5 out of the 6 animals showed decreased mapping in response to LPS but Male 1 had a greater percentage of reads mapping at 0hr compared to 7hr LPS.

There was variation between the individual Classes of ERVs in response to LPS, with 5 out of the 6 Classes recognised as having decreased mapping in response to LPS, and only one class showing an increase in the 7hr compared to 0hr LPS. This sub class of ERV1 called “ERV1?” (so-named by NCBI) had significantly (Welch two sample t-test P value = 0.02579) more reads mapping either in response or as a consequence of the LPS response, despite being the smallest class, containing just 638 regions, compared to ERVL which contains 100,223 regions.

Examining the individual names of repeat region, the most regulated (down in response to LPS), with the greatest proportion of reads mapping (more than twice the 2nd most abundant region, MLT1D) was BTLTR1. BTLTR1 is a known ruminant specific LTR-containing repeat region (Elsik et al., 2009) and it has been debated as to whether this short, shared sequence is functionally significant and resulted in the transfer from BovB which was transferred horizontally among reptiles, ticks and ruminants (Walsh et al., 2013), to BTLTR1, as a region of sequence similarity has been previously identified between BovB and BTLTR1 (Adelson, 2008, Adelson et al., 2009). BovB is known to have been incorporated into the ruminant ancestor genome <50 MYA which is thought to be recent compared to others inactive for >100 MYA such as LINE L2 and SINE MIR (Adelson et al., 2009), and the horizontal transfer of BovB has been shown to have transformed many vertebrate genomes (Walsh et al., 2013). The identification of considerable BTLTR1 mapping in sheep in this study provides compelling evidence that BTLTR1 is particularly relevant to genome evolution for the lineage leading to domestic sheep, and suggests future study of the LINEs (which includes BovB), may be productive.

The reads defined as mapping to each reference ERV region are not strand specific to the ERV region and there may be significantly differentially expressed genes that lie either within or in very close proximity to these regions, so it is possible that these are confusing the overall picture as reads have been included when they only have to overlap by <1bp.
3.7 Discussion

RNA-sequencing, using both polyA-selected and ribodepleted library preparation methods for sheep BMDMs across an LPS time course has provided genome-wide transcriptional information regarding the dynamic changes of gene expression that take place following TLR4 activation. Using two independent pipelines, Kallisto and StringTie, raw sequence data has been processed which has provided evidence that sheep BMDMs share a very complex transcriptome with macrophages of other species. Sheep macrophages have been estimated to utilise over 82% of the reference transcriptome OarV3.1. StringTie identified more than three times the number of reference transcript models than are currently recognised, and with all sequence available to Ensembl, the coverage of the current reference genome OarV3.1 has been greatly improved and will contribute to improving future versions.

The advantage of RNA-Seq, lacking a saturation limit, has proved a frustrating double-edged sword. The phenomenal abundance of certain transcripts, which are not regulated, cell specific or of immediate interest (such as SPP1, identified in Figures 3.2.A.ii and 3.3.A.ii), clearly occupy the majority of precious and expensive read coverage at the cost of detection of the rarer transcripts, which may be of great biological importance (see Figure 3.1). The computational, algorithmic and logistical challenges of processing, storing and analysing the volume of data created has also been fully appreciated.

Two pipelines, Kallisto and StringTie were used to obtain expression data. Both pipelines have been compared and generally demonstrate a positive correlation, with outliers highlighting either problematic, highly complex areas of the genome (where Kallisto determined greater expression than StringTie) or genes with many more transcript models than are currently recognised in the reference OarV3.1.81 or where the reference transcript model is not the most highly expressed model identified (where StringTie determined greater expression than Kallisto). Both pipelines have complementary strengths regarding accuracy of expression at reference transcript level (Kallisto) and gene level (StringTie), with StringTie being more labour intensive and computationally demanding, yet providing detailed information regarding intron and exon boundaries and novel splice variants. These strengths and weaknesses complement each other in appreciating the global transcriptional response across the whole genome, as well as providing validation of expression estimates.
The two library type preparation methods for RNA-Seqencing (polyA-selected and ribodepleted total) have been compared. There is a positive correlation between expression estimates derived from each library type. As expected, the greater variety of transcript type was detected in the ribodepleted libraries (which were also sequenced at the greater depth >100M). This highlighted many of the non-polyadenylated histones that are present in these samples, that were not detected by the polyA-selected libraries.

Given the abundance of individual transcripts taking up such a large proportion of the reads, and the many other types of transcript detected in the ribodepleted total samples, the polyA-selected library types detected the same proportion of the protein-coding transcriptome as the ribodepleted libraries. This was despite the polyA-selected library samples being sequenced at a quarter of the depth as the ribodepleted samples. In principle, if the additional transcripts detected in the ribo-minus libraries form a constant proportion of the “excess” detection, the effect of library type can be removed with a simple correction; removing counts attributed to all of those transcripts from the estimate of total counts used as a baseline for expression estimates/read number. This correction has now been successfully applied across the entire sheep atlas (Bush et al., 2017). For the purpose of this project, such a correction was unnecessary, since both methods were applied to the same samples (0 and 7 hours). Indeed, comparison of these two samples was used to develop and validate the correction approach.

The preliminary investigation into LTR retrotransposons identified the ruminant-specific repeat region BTLTR1 to be the repeat region which has the greatest proportion of reads to map across all individuals. As a class of repeat region, BTLTR1 was found to be down-regulated in response to LPS. Taken together this suggests that of all the LTR retrotransposons, BTLTR1 has had the greatest effect on sheep genome evolution. However, the similarity in sequence to the LINE BovB, highlights the need for continued analysis involving all classes of retrotransposon. It is also recognised that a complete reference genome is essential as there are still many scaffolds that contribute to OarV3.1, which necessitate the need for extra long read technology. When the next, most complete version of the sheep genome is released, future plans involve re-mapping all the samples used in this project and repeating this analysis, as the quality, volume and RNA-Sequence from each individual will always be available from the unprecedented resource that is The
Sheep Atlas (the genomes of each individual were also sequenced and the data is available (Clark et al., 2017)).

Chapter 4 Identification and annotation of sheep macrophage transcripts using immune cell data from the wider Sheep Atlas

4.1 Introduction

The Hume laboratory in which this study was conducted, has a long term interest in transcriptional regulation in macrophages and identification of genes required for the survival, proliferation and differentiation of the macrophage (Hume et al., 2016), as well as comparative transcriptomics of macrophages from multiple species (Kapetanovic et al., 2012, Fairbairn et al., 2013, Karagianni et al., 2016). Mice and humans are known to differ in several aspects of innate and acquired immunity (Mestas & Hughes, 2004, Marr et al., 2006). There has been ongoing debate as to the value of mouse models of macrophage biology and chronic inflammatory disease (Schroder et al., 2012, Seok et al., 2013, Takao & Miyakawa, 2015). There is an increasing interest in large animal models, where both the immune system and physiology more closely resemble that of humans (Kapetanovic et al., 2012)(see Chapter 1).

Miru (now Graphia Pro) is a computational tool which enables visualisation of large correlation networks and was discussed in detail in chapter 2.10. It has been used to identify clusters of co-expressed genes in complex datasets from multiple cancers, tissues and isolated cell populations (Freeman et al., 2012, Doig et al., 2013, Hume et al., 2013, Mabbott et al., 2013, Clark et al., 2017), and to dissect the time course of macrophage responses to agonists (Kapetanovic et al., 2012, Raza et al., 2014, Baillie et al., 2017). For the prioritised clusters, transcripts were explored individually to assign functional annotation (see Chapter 2.11.1) to as many as possible.

The web based tool GATHER (a Gene Annotation Tool to Help Explain Relationships) (Chang & Nevins, 2006) improved understanding of the pathways, processes and relationships that are shared between the transcripts of a given cluster of interest. GATHER (Chang & Nevins, 2006) provided an initial overview of the biology identified in each cluster, in conjunction with information gathered for each transcript using GeneCards (Stelzer et al.,
GATHER enabled each list of genes within a cluster to be analysed against a series of data sources including annotations from evolutionary homologs, ‘interactome’ or protein-protein interactions and literature resources which use gene-to-gene co-citation networks (Jenssen et al., 2001, Hoffmann & Valencia, 2004, Rual et al., 2005). This programme is based on human functional genomic data, so some gene members of the sheep clusters were excluded from analysis. Hence, it was regarded as a complementary tool and additional cited literature information from GeneCards for individual genes (pasted into Microsoft Excel) provided additional metadata to be scrutinised to provide improved understanding and an overview in terms of transcriptional regulation, protein complexes and relationships to known pathways for each cluster of interest.

Previous studies in the laboratory have analysed the gene expression profiles of pig macrophages, including bone marrow derived macrophages, monocyte derived macrophages and alveolar macrophages using microarrays (Freeman et al., 2012, Kapetanovic et al., 2013). Equine alveolar and peritoneal macrophages have also been compared with each other when exploring the importance of the local microenvironment (Karagianni et al., 2016). This history provides a detailed framework on which to base methods (see chapter 2) and enable the sheep, a ruminant animal, to be compared to other species.

Resident tissue macrophages are situated within every mammalian tissue, forming up to 15% of the cellular content of any given tissue. As discussed in Chapter 1, macrophages in each tissue adapt to the local microenvironment, to perform specific functions (Chapter 1.3). Lineage-trace studies in mice indicate that the majority of tissue macrophages are seeded during embryonic development, and thereafter are capable of self-renewal (Guilliams et al., 2013, Yona et al., 2013, Ginhoux & Jung, 2014, Ginhoux & Guilliams, 2016). The major exception is the large macrophage population of the intestinal wall. Macrophages in the lamina propria of the gut are continually replaced by circulating monocytes and must selectively down-regulate pattern recognition receptors in order to maintain homeostasis and prevent chronic inflammation. Blood monocytes in turn, are derived from bone marrow and myeloid stem cells, with CSF1R encoding the receptor that controls macrophage differentiation (with two ligands: CSF1 and IL34) (Section 1.3). The macrophages of the gut wall are rapidly depleted following treatment of mice with anti-CSF1R antibody (MacDonald et al., 2010, Guilliams et al., 2013). Taking these findings
together, Baillie et al (Baillie et al., 2017) have argued that monocyte differentiation in response to CSF1 provides a model for the differentiation of the macrophages of the lamina propria of the gut wall, where it is crucial that they down-modulate their response to the intestinal microflora to avoid chronic inflammation.

Macrophages of the lung, on the other hand, must provide a first line of defence against inhaled microorganisms. Comparative analysis of pig alveolar macrophages with monocyte derived and bone marrow derived macrophages grown in CSF1, revealed that the lung macrophages selectively up-regulated phagocytic receptors, including multiple C-type lectins, and pattern recognition receptors (Freeman et al., 2012, Kapetanovic et al., 2013).

In this Chapter, the aim is to systematically compare the expression profiles of BMDM with other macrophage populations and immune cell types within the wider atlas. This will identify transcripts that are either a) enriched in macrophages, b) induced in cells of the macrophage lineage or c) associated with tissue-specific adaptation and preferentially expressed in certain types of macrophage. This will highlight transcriptional signatures that sheep share with other species, but the main focus for discussion is on aspects that separate sheep from other species.

### 4.2 Methods

All available immune cell sample RNA seq from the Sheep Atlas were utilised, summarised in table 4.1. The raw RNA-Seq data was taken and processed in exactly the same way as that of the BMDM response samples (see chapter 2). The same pipelines, both Kallisto and StringTie, generated transcript and gene co-expression profiles respectively for all these immune cell samples in Table 4.1 (see Appendix 4.0). These expression estimates (transcript expression as TPM, using Kallisto, and gene expression as FPKM, using StringTie), were then examined using Miru. Clusters of transcripts that share an expression pattern common to all individuals, were extracted and examined more thoroughly.
Table 4.1. Table showing the samples from the Sheep Atlas which have been utilised in chapter 4. Where samples are missing from individual animals, there was not enough high quality RNA to be sent for sequencing, thus RNA seq data was not available. All samples utilised were sequenced as described in Chapter 2, using polyA-selection at >25M read depth.

\( \text{\textbullet} \) indicates polyA-selected libraries, >25M read depth, paired end RNA-Seq data.

<table>
<thead>
<tr>
<th>Sheep Atlas animal</th>
<th>Female 1</th>
<th>Female 2</th>
<th>Female 3</th>
<th>Male 1</th>
<th>Male 2</th>
<th>Male 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMDM 0hr LPS</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td>BMDM 2hr LPS</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td>BMDM 4hr LPS</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td>BMDM 7hr LPS</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td>BMDM 24hr LPS</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td>Alveolar macrophages (AMs)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>![image]</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Monocyte derived macrophages (MDMs)</td>
<td>![image]</td>
<td>n/a</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td></td>
</tr>
<tr>
<td>Blood Leukocytes (BLs)</td>
<td>n/a</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
<td></td>
</tr>
<tr>
<td>Peripheral blood mononuclear cells (PBMCs)</td>
<td>![image]</td>
<td>![image]</td>
<td>n/a</td>
<td>![image]</td>
<td>n/a</td>
<td>![image]</td>
</tr>
</tbody>
</table>

4.3 Results and Discussion

4.3.1 Examination of a principal components analysis of Sheep Atlas immune cells

A principal component analysis (PCA) biplot of the reference transcript expression (Kallisto, TPM) profiles for each individual immune sample is shown in Figure 4.1. As expected, the macrophages cluster together, whereas the peripheral blood leukocytes and peripheral blood mononuclear cells are clearly distinct from both the macrophages and each other. As observed in pigs, the BMDM and MDM have overlapping profiles, but the alveolar macrophages are much more similar to BMDM/MDM than they were in pigs (Kapetanovic
et al., 2013), perhaps reflecting that this analysis included both BLs and PBMCs as well as the macrophage samples.

**Figure 4.1 Principal components of all Sheep Atlas immune cell samples**
Each immune cell sample RNA-Seq data (polyA-selected 25M paired end reads) were processed through the Kallisto pipeline (See Chapter 2.8.1) and transcript expression estimates (TPM) for every reference transcript in Oar V3.1.81 were analysed across all samples. The resulting PCA biplot is coloured by cell type. The percentage of variance is in brackets for each principal component.

### 4.3.2 Network analysis
To identify genes expressed specifically in macrophages, or macrophage subpopulations, relative to other leukocytes, that provide the differential profiles evident in the PCA (Figure 4.1), I performed a network analysis using Miru. For the purpose of this analysis, only transcripts expressed above 5 TPM (deemed robust expression) in at least one sample were included, which reduced the dataset to a manageable size (13,812 transcripts) and reflects the cell-type specific transcriptome better that the more relaxed criterion. A transcript-to-transcript analysis of these data without filtering was dominated by clusters with low expression and/or individual-specific expression (not shown). A sample to sample analysis based on this set of transcripts across all individual immune cell samples (Figure 4.2A), was performed to analyse the relationships between samples. This created a network graph containing 44 nodes (a node for every immune cell sample) with 218 edges with a Pearson correlation coefficient of \((R) = 0.92\), to include all 44 nodes. An MCL inflation value of 2.2
(Miru’s optimal default setting) created clusters of samples with highly correlated expression patterns. This network graph is broadly consistent with the PCA, in that the blood leukocyte and PBMC were clearly distinct, and the AM formed a separate cluster. The inclusion of the LPS-stimulated macrophages revealed the global change in profile in response to this stimulus, and also the distinctive response of one individual, Male 3 (M3 samples).

A stringent analysis of cell type relationships was therefore performed on aggregated data. The dataset was filtered further to remove transcripts that were not expressed at >30 TPM in at least one sample. The expression values for each remaining transcript (expressed >30tpm in at least one sample), were then averaged across all individuals for each immune cell type. The clustered sample-to-sample network graph contained 9 nodes (a node for each immune cell type) with 15 edges with a Pearson correlation coefficient of \( R = 0.82 \) and MCL inflation value of 2.2 to create Figure 4.2B. This analysis reduces the divergence between the macrophage samples, which form a single cluster distinct from the other leukocytes.
Figure 4.2 Sample-based networks of transcript expression
A From individual Sheep Atlas immune cell samples.
All samples which reached > 5 TPM in at least one sample were included. The Pearson correlation coefficient threshold was 0.92 and the MCL inflation value for clustering was 2.2. Each node (sphere) represents a sample and the lines between them represent correlations of at least 0.92. Nodes are coloured according to the sample type.
B From averaged transcript expression All transcripts which reached >30 TPM in at least one sample were averaged for each immune cell type, using all immune cell data from the Sheep Atlas Project. The Pearson correlation coefficient threshold was (R) =0.82, MCL inflation value was 2.2. Each node represents one cell type or treatment and the edges between them show correlations of > 0.82.
Using the same filtered list of transcripts, a transcript-transcript co-expression network graph, using a Pearson correlation coefficient of coefficient (R) = 0.9, containing 6,654 nodes with 1,210,700 edges, clustered with an MCL inflation value of 2.2 can be seen in Figure 4.3, which also highlights the position of key clusters which have been selected for further analysis within the graph.

This transcript-transcript co-expression analysis revealed clusters of transcripts with enriched expression in specific populations of cells. All clusters were examined by looking at the mean associated expression pattern histogram and known genes within each cluster. The key clusters showing increased expression of transcripts within a given cell types of interest were extracted and grouped together according to which immune cell type they were enriched in, along with clusters of interest which contain genes of known function such as CSF1R. The transcripts within these key groups and clusters can be seen in the Excel workbook in Appendix 4.1. Ensembl Biomart (Aken et al., 2016) was used to add metadata information for each ensemble transcript stable ID. The content of each of the clusters is examined in detail below.
Figure 4.3 A network layout based on average transcript expression across Sheep Atlas immune cells

A. The co-expression network layout, based on average expression across Sheep atlas immune cell types from transcripts expressed >30 TPM in any sample. The network contains 6654 nodes and 121,069 edges and was derived based on Pearson correlation coefficient >0.9. Nodes (spheres) represent transcripts and edges (connecting grey lines) represent the correlation between expression patterns of transcripts at a Pearson correlation coefficient of 0.9 or greater. Nodes of the same colour form a cluster.

B. An overview, highlighting positioning of key clusters, which share similar expression patterns that demonstrate specific increased expression of given immune cell lineages or stage of activation (LPS responsive). Nodes represent transcripts (averaged values). Edges have been removed for ease of visualisation but can be seen in 4.4A. Minimum cluster size was 4 nodes and only those nodes that fell within a cluster are shown.

C. Selected clusters for analysis, containing genes with high expression in different sheep immune cells. These clusters of interest were extracted from the main network graph. Different views are shown, rotated clockwise 60° from view 1.

Although the BM and PBMC clusters form a continuous network, the clusters are separated from both the general macrophage and the alveolar macrophage enriched clusters, as demonstrated in Figure 4.4.B which extracts the clusters of interest from this main network graph. The largest cluster 1 (pale green) and cluster 3 (pale brown), contain transcripts
which showed high expression in BLs and PBMCs (Figure 4.4), a total of 2215 transcripts. The two clusters are distinguished by relative enrichment in BLs. Cluster 3 is more mononuclear cell-specific but taken together these clusters were likely to contain genes enriched specifically in populations of blood lymphocytes, granulocytes and NK cells as opposed to macrophages. They may also contain monocyte-specific transcripts that are down-regulated with macrophage differentiation. Clearly, a much more refined analysis could be obtained by separating the populations of blood cells as in studies by the mouse ImmGen consortium (Heng & Painter, 2008) and in humans; examining peripheral blood leukocyte populations (B-cells, CD4+ T-cells, CD 8+ T-cells, lymphocytes and granulocytes) (Palmer et al., 2006), and by the FANTOM5 consortium (Lizio et al., 2015, Noguchi et al., 2017). However, as a first pass, these clusters appear to contain transcripts that are enriched specifically in cells of the immune system. In the wider sheep transcriptional atlas thymus, spleen and multiple lymph node sample data is available (Clark et al., 2017) so future studies could easily refine sheep B cell and T cell-specific clusters by including this data which has not been examined for this thesis.

**Figure 4.4 Blood Leukocyte (BL) and peripheral blood mononuclear cell (PBMC) clusters extracted from the main network analysis**, based on average expression across Sheep atlas immune cell types from transcripts expressed >30 TPM in any sample. Two particular clusters were enriched for transcripts expressed in PBMC and BLs: Cluster 1 (1686 Nodes/transcripts) and Cluster 3 (528 transcripts). The corresponding signal histograms showing the average expression of all the transcripts contained in each cluster, for each cell type can be seen next to the extracted cluster that is pictured alongside.
This analysis reveals transcripts that are not expressed by macrophages. Consistent with that proposition, these clusters contain the classic B lymphocyte cell related surface antigen genes found in mice and humans such as CD19 which is indicative of mature naïve B cells (Otero et al., 2003) and the gene encoding a B-cell restricted molecule, CD22 (Moyron-Quiroz et al., 2002). Other well documented B cell genes include CD69 which is associated with B cell activation and CD27, indicating there are memory B cells and activation within these samples (Wei et al., 2011, Kaminski et al., 2012), and many of the B cell-associated genes identified in humans such as Major histocompatibility complex, class II, DM alpha (HLA-DMA), B-cell CLL-lymphoma 11A zinc finger protein (BCL11A) and transcription factor genes such as IRF4, TCF4 and TFEB (Palmer et al., 2006). The cluster also contains genes associated with T cells and T cell subsets (CD3 and CD8), and many known lymphocyte-associated transcriptional regulators, such as LEF1. The cytokine receptor gene IL2RA appears in this cluster and is known to be an early determinant of T lymphocyte fate specification (Chang et al., 2014) and interestingly cluster 1 also contains CNR2, a gene known to encode a molecule with immunosuppressive effects in macrophages (Sacerdote et al., 2000) and it has been suggested that this gene may play a role in maintaining homeostasis in naïve and memory T cells (Best et al., 2013).

**T cell:** In common with ImmGen clustering examining T cell activation (Best et al., 2013), there are also genes of the S1P receptor family and the prosurvival chemokine receptor CX3CR1, as well as the same members of E protein (class I) family of helix-loop-helix transcription factors (TCF3 and TCF4) that have been identified which were identified as potential repressors of genes expressed in naïve and late effector memory T cell populations (Best et al., 2013). The early NK cell activation markers CD27 and the C-type lectin receptor CD69 appear in cluster 3 but stem cell antigen 1 (SCA1) and killer cell lectin like receptor G1 (KLRG1), which have both been found to be dramatically increased with NK activation (Fogel et al., 2013) are absent from these clusters.

**Granulocyte:** Multiple granulocyte-enriched genes identified in humans are found in these clusters such as colony stimulating factor 3 receptor (CSF3R), nuclear factor (erythroid-derived 2) (NFE2), fibrinogen-like 2 (FGL2), and the ets domain transcription factor E74-like factor 4 (ELF4) (Palmer et al., 2006). The gene regulator of G-protein signalling 2 (RGS2), is also found in cluster 1, which has also been identified to be expressed in unstimulated circulating neutrophils (Bertrand et al., 2004, Palmer et al., 2006). None of the recognised
intracellular pathogen destruction genes, enriched in human granulocytes were present such as cathepsins B/ C/ S.

Somewhat less obvious or expected as they should have been depleted before the library preparation, was the finding that the most highly expressed, annotated transcripts within these clusters are ribosomal proteins (RPs) involved in assembly of small 40s ribosomal subunits: RPS8,11,19 and 21 and large 60s ribosomal subunits: RPLP0,1 and RPL8,29,32,34 and 37. These ribosomal phosphoproteins are probably involved in plasma cells which are protein secretion factories.

4.3.3 Identification and annotation of transcripts enriched in the sheep macrophage lineage.

There are several clusters in which the average expression shows enrichment across all the macrophage samples, with low or absent expression within the BL or PBMC samples (Figure 4.5). The five clusters containing the greatest number of transcripts (2, 4, 23,25, 26) have been extracted and appear in Appendix 4.1, within the worksheet named BMDMs, AMs and MDMs.

In total, the five macrophage enriched clusters contain 1577 transcripts (with each transcript being from the reference transcriptome, having been assigned an expression estimate >30tpm by Kallisto2 in at least one immune cell sample). Based upon the averaged expression estimate for each cell type, the pattern of transcript expression across the different cell types was scrutinised using Miru. Each transcript has a potential function in innate immunity, and accordingly, each transcript that was identified only by an ENSEMBL ID, was subjected to manual annotation using ENSEMBL to examine the nearest known orthologues and validate by comparing regions for synteny. Genecards (Stelzer et al., 2016) was used to check the validity of each gene that I had assigned functional annotation and determine the relevance and sense of its identification within that given cluster.

The clusters contain genes encoding numerous known transcriptional regulators implicated in lineage commitment including lymphocyte cytosolic protein 1 (LCP1), a plastin expressed only in haematopoietic lineages and malignant human cells, cellular repressor of E1A stimulated genes 1 (CREG1), which is known to regulate transcription of core macrophage-associated genes in the mouse (Gautier et al., 2012), as well as MAFBZIP Transcription
factor B (MAFB), which I assigned to ENSOART00000001452/ENSOARG00000001355, a bZIP transcription factor known to have an important role in the regulation of lineage-specific haematopoiesis, and that recently has been shown to play a central role in in promoting cholesterol efflux and anti-inflammatory macrophage polarization (Kim, 2017).

**Figure 4.5** Macrophage enriched clusters extracted from the main network analysis, based on average expression across Sheep atlas immune cell types from transcripts expressed >30 TPM in any sample. Five particular clusters were enriched for transcripts expressed across the BMDMs, AMs and MDMs. The corresponding signal histograms showing the average expression of all the transcripts contained in each cluster, for each cell type can be seen next to the extracted cluster that is pictured alongside. The columns representing immune cell types, appear in the following order from left to right: BMDM 0hr, BMDM LPS 2hr, BMDM LPS 4hr, BMDM LPS 7hr, BMDM LPS 24hr, AM, BL, MDM and PBMC.

Within these 5 macrophage clusters, 319 transcripts were identifiable only by ENSEMBL ID. Many of these were situated on scaffolds, with simple transcript models of only 1 or 2 exons and minimal information other than gene type. In total, for these 5 clusters, 138 ENSOARGs have been provided with a candidate orthology to a known gene, including several genes not previously associated with macrophage function; such as COX7A2, WDR61, PGRMC1 and SELENOP (also known as SEPP1 or Selenoprotein P).

**SELENOP**, found in cluster 4, is one of the most highly expressed transcripts, selectively within the sheep BMDMs and MDMs. **SELENOP** is widely-expressed in mice, albeit largely
absent from lymphocytes (www.biogps.org), whereas in humans, the transcript is enriched in liver, absent in monocytes and induced in MDM (Kawaij et al., 2017). The relative abundance of this transcript seen here, most highly expressed in resting sheep macrophages, is consistent with the documented relative importance of selenium for innate immune function in sheep (Kiremidjian-Schumacher and Stotzky 1987, Dhur, Galan et al. 1990, Hall, Sendek et al. 2011, Hall, Vorachek et al. 2013) and this relative high expression might underpin the species differences seen in effects of selenium and vitamin E on the immune response of domestic animals (Finch and Turner 1996).

Clusters 2 and 4 are more specifically enriched in the BMDM and MDM samples. These cells share their exposure to CSF1 and extended in vitro culture, and co-segregate in the PCA (Figure 4.1), mirroring the sample to sample network analysis (Figure 4.2). The program and tool GATHER (Chang & Nevins, 2006) was used to help with understanding the known function of a given group of transcripts as it provides the most significant GO terms (Ashburner et al., 2000). For these BMDM and MDM clusters, the most significant GO terms included G-protein coupled receptor protein signalling pathway, sensory perception, detection of external stimulus and cellular physiological process, reflecting the physiological phenotype of these cells. The larger cluster 2, contains 1247 macrophage lineage transcripts and many of the classic mature macrophage markers such as CD163 (a member of the scavenger receptor cysteine-rich superfamily), CD164 (a transmembrane sialomucin), CD14 (the coreceptor for LPS, see Chapter 1.4, along with many other genes known to be enriched in macrophages, such as phospholipase D3 (PLD3), tripeptidyl peptidase 1 (TPP1), lysosomal associated membrane protein 2 (LAMP2), transcobalamin 2 (TCN2), and ATP binding cassette subfamily A member 1 (ABCA1), present in these macrophage enriched clusters (Gautier et al., 2012).

In other species, CSF1 promotes cellular proliferation and also induces a set of specific genes associated with tissue repair and lipid metabolism (Irvine et al., 2009, Fairbairn et al., 2011). Consistent with the shared cell proliferation, these macrophage enriched clusters contain numerous cyclin dependent kinase inhibitor genes such as CDKN2C and CDKN3 as well as genes encoding components of the mitotic apparatus such as BUB1B and CDC20. Many transcripts seen here encode genes known to be involved in various cancers, including CKS2 which is known to be related to pathways in cancer and small cell lung cancer (Valles et al., 2012) and the important oncogene BIRC5, a multitasking protein that
has dual roles in promoting cell proliferation and preventing apoptosis, also found to enhance aerobic glycolysis and drug resistance (Hagenbuchner et al., 2016).

Human and mouse macrophages grown in CSF1 differ in their gene expression profiles. In human MDMs, CSF1 promotes the expression of multiple genes involved in cholesterol metabolism and lipid accumulation, including HMGCR, MVD, IDI1, FDP5, S青LE, CYP51A1, EBP, NSDHL, DHCR7, and DHCR24 (Irvine et al., 2009). Of these genes, only DHCR24 appears in this Cluster 2. However, also appearing in Cluster 2 is the ATP-binding cassette (ABC) transporter gene ABCA1, a cholesterol efflux pump in the cellular lipid removal pathway, INSIG2 and ELOVL1. All these genes have been identified in mice macrophages and to be responsive to diet, being substantially upregulated in atherogenic diets compared to restrictive diets (Renaud et al., 2014).

The long chain fatty acid translocase, encoded by CD36, was amongst the novel genes most highly expressed and annotated in the macrophage-enriched gene set in cluster 2 (assigned to ENSOART00000018636/ENSOARG00000017123). This protein gene, along with FABP7 (cluster 4) and FABP5 (cluster 2), both of which are also highly expressed, are known to have a high specificity for long-chain fatty acids and other hydrophobic ligands (Vogel Hertzel & Bernlohr, 2000), and are involved in lipoprotein metabolism in humans and mice. Polymorphisms in FABP5 are already known to be associated with type 2 diabetes (Bu et al., 2011). This is of specific interest in humans because the upregulation of scavenger receptors capable of binding modified forms of LDL (such as CD36), are hallmarks of an imbalance in lipid metabolism and the formation of so-called macrophage foam cells (macrophages laden with cholesterol), which are present in atherosclerosis, and a result of apolipoprotein B in the subendothelial space causing local inflammation, the release of chemokines and chronic inflammation (Choudhury et al., 2005, Moore et al., 2013).

As mentioned in Chapter 2.4, Figure 2.1, the sheep macrophages grown in CSF1 do have a vacuolar appearance, which might be associated with formation of lipid droplets. Other very highly expressed transcripts found in cluster 4, which are involved in cholesterol metabolism are LIPA, the lysosomal acid lipase (also known as cholesteryl ester hydrolase), which separates cholesteryl esters into cholesterol and fatty acids (Brown & Goldstein, 1986) and NPC2, a gene encoding a protein known to bind unesterified cholesterol that has been released from LDLs, to regulate the transport of cholesterol through the late endosomal/lysosomal system (Frolov et al., 2003). PLD3, the gene encoding phospholipase
D which is known to catalyse the hydrolysis of membrane phospholipids (Fazzari et al., 2017), is also highly expressed and found in cluster 2.

Taken together, these results suggest that resting sheep macrophages grown in CSF1 are specifically adapted to deal with the accumulation of lipoproteins and cholesterol within the cell, resulting from the increased availability and uptake of long chain fatty acids in the normal healthy ruminant. Clearly regulation over some aspects of this gene expression profile are shared between sheep, mice and humans. Two of the ENSEMBL IDs, present in cluster 2 (ENSOART00000001452/ENSOARG00000001355 and ENSOART00000003211/ENSOARG00000002962) have been assigned the functional annotation of the highly conserved transcription factors MafB and XBP1 respectively, which are also important and well recognised in mouse and human macrophages (Martinon et al., 2010, Kim, 2017), validating these manual annotations in sheep.

MAFB is a negative regulator of type I IFN and a pro-inflammatory cytokine, playing a pivotal role in regulating macrophage polarisation, and mediating anti-inflammatory and anti-atherogenic signalling in macrophages (Kim, 2017). MAFB regulates macrophage cholesterol efflux in mice by increasing expression of ABCA1 (also in cluster 2), and is a downstream target of IL4/STAT6 (a STAT6 transcriptional binding site is conserved between mice and humans on the MAFB promotor), being strongly inhibited by proinflammatory signalling, remaining repressed in macrophage foam cells, thus it shares this same expression profile in sheep, being repressed in response to LPS (Kim, 2017).

The transcription factor XBP1 (assigned to ENSOART00000003211/ENSOARG00000002962, also seen in cluster 2) in other species is involved in B cell differentiation and the production of immunoglobulins (Reimold et al., 2001). XBP1 also promotes DC survival in a TLR ligand independent way and may be a therapeutic target for regulating the life span of DCs during an immune response (Leavy, 2007). There are multiple complex interactions between XBP1 and other transcription factors (Ozcan et al., 2004, Ueki & Kadowaki, 2011, Chen et al., 2014, Franca et al., 2014, Kishino et al., 2017). XBP1 activation through reactive oxygen species (ROS) production from TLR4 activation has also been shown to mediate the progression of steatosis (abnormal retention of lipids within a cell) to inflammation (Ye et al., 2012) and this transcription factor has also provided the link between ER stress and intestinal inflammation (Kaser et al., 2008). Given the importance and complexity of this transcription factor, future research is required to investigate the expression and regulation
of XBP1 in sheep compared to mice and humans. Improved understanding of the transcriptional regulation of these genes could provide novel management strategies for preventing atherosclerotic plaque development and promoting regression of existing atherosclerotic plaques in humans.

Macrophages are not restricted to their involvement in cholesterol homeostasis, but macrophages in the red pulp of the spleen, bone marrow and liver are involved in the removal of senescent red blood cells, the degradation of haem (a protoporphyrin IX ring surrounding an iron atom (Hamza & Dailey, 2012)), and recycling of iron for erythropoiesis (Korolnek & Hamza, 2015, Soares & Hamza, 2016, Theurl et al., 2016). Free iron exchanges electrons with oxygen, catalysing the production reactive oxygen species (ROS) which are cytotoxic. However, iron is also essential for core biological functions, such as the production of ATP by the mitochondrial electron transport chain (Baranano et al., 2000), and many haemoproteins (Kleingardner & Bren, 2015) play a central role in regulating macrophage function, therefore macrophage iron metabolism is tightly regulated.

Amongst the most highly expressed transcripts annotated in cluster 2, many are clearly associated with iron homeostasis. Ferritin heavy chain 1 (FTH1), which was assigned to ENSOART00000015613/ENSOARG00000014339, FTH1-like which was assigned to ENSOART00000019236/ENSOARG00000017679 and ferritin light chain (FTL), which was assigned to ENSOART00000013538/ENSOARG00000012450) are all enriched in the sheep macrophages. The gene zinc-iron permease SLC39A1, encoding a major endogenous uptake transporter, also appears in cluster 2 alongside the very highly expressed glutathione peroxidase 4 (GPX4 which was assigned to ENSOART00000012030/ENSOARG00000011060 in cluster 2), and encodes an enzyme that protects cells against iron-dependent accumulation of lipid reactive oxygen species (Dixon et al., 2012, Dixon & Stockwell, 2013, Cao & Dixon, 2016, Yang & Stockwell, 2016). Interestingly this isoenzyme is also a selenoprotein, underlining the importance of selenium in innate immunity in sheep (see Section 4.3.3). The high expression in these sheep macrophage enriched clusters suggests that sheep macrophages may be particularly susceptible to ferroptosis (non-apoptotic form of cell death due to iron dependent accumulation of lipid ROS and depletion of plasma membrane polyunsaturated fatty acids (Cao & Dixon, 2016)), and could provide a useful model to answering the many questions that still surround this cell death process that is linked to multiple human diseases (Feng & Stockwell, 2018). Multiple other
protective peroxidase genes are present in these macrophage clusters such as PRDX1, Peroxiredoxin 1 a thiol-specific peroxidase, and the relative high expression of these genes may suggest that redox reactions involving iron and lipid oxidation are mechanisms that sheep macrophages rely upon.

4.3.4 The sheep alveolar macrophage signature

Alveolar macrophages are positioned within the epithelium of the airways and alveoli, providing first line of defence against inhaled pathogens and surfactant homeostasis. Gibbings et al have discovered that the environment dictates the AM development, programming and function (Gibbings et al., 2015), and as a result alveolar macrophages are highly adapted with unique characteristics to reflect their tissue location and function. Alveolar macrophages must not respond in a pro-inflammatory way to non-pathogenic antigens and are known to express a unique range of surface receptors and restrain from inflammation to ensure efficient gas exchange and close communication with alveolar epithelial cells (Hussell & Bell, 2014, Kopf et al., 2014). Several studies in mice have documented the specific adaptation of tissue macrophages to specific environments (Gautier et al., 2012, Mass et al., 2016). Pig alveolar macrophages have been profiled as part of the pig atlas (Freeman et al., 2012, Kapetanovic et al., 2013) and stimulated with LPS/IFN-γ and IL4 to induce polarisation, where it has been shown to affect virus (PRRSV) replication (Wang et al., 2017).

Alveolar macrophages have long been known to self-renew under normal steady-state conditions in the adult (Tarling et al., 1987), having been seeded during waves of embryonic haematopoiesis, with monocytes only replenishing the resident alveolar macrophage population to a certain extent (Kopf et al., 2014). Alveolar macrophages depend upon GM-CSF rather than CSF1 for their survival, proliferation and terminal differentiation (Shibata et al., 2001). Horse (Karagianni et al., 2016), mouse (Heng & Painter, 2008), pig (Kapetanovic et al., 2013) and human (Xue et al., 2014) alveolar macrophages have been profiled.

Several vital ruminant-specific processes that surround foregut fermentation are likely to impact upon the AM environment (Khadom et al., 1985). Rumination requires regurgitation of rumen boluses of semi fermented complex carbohydrates, which are saturated with fermentative microbial flora (Braun et al., 2013). The remastication and re-ensalivation of the bolus are vital processes that increase the volume of saliva (containing buffering phosphate to maintain rumen pH) produced (Kay, 1966). Lastly there is the
ruminant specific process of eructation, to rid the rumen of the huge volume of gas produced through fermentation (90% methane, 10% carbon dioxide) (Dougherty, 1968). It is likely that the pharynx and glottis fail in preventing any contamination of these products from entering the lungs, so it is inevitable that the ruminant alveolar macrophage will be frequently exposed to a range of substances that simple stomached animals are not.

The PCA and network analysis above indicated that sheep AM were not greatly different to BMDM and MDM. There were, nevertheless, several clusters of transcripts that were enriched in AM (Figure 4.6).

Figure 4.6 Alveolar macrophage enriched clusters extracted from the main network analysis, based on average expression across Sheep atlas immune cell types from transcripts expressed >30 TPM in any sample. Three particular clusters were enriched for transcripts expressed in AMs. The corresponding signal histograms showing the average expression of all the transcripts contained in each cluster, for each cell type can be seen next to the extracted cluster that is pictured alongside. The columns representing immune cell types, appear in the following order from left to right: BMDM 0hr, BMDM 2hr, BMDM 4hr, BMDM 7hr, BMDM 24hr, AM, BL, MDM and PBMC.

Cluster 5 is the largest, most AM-enriched; Cluster 20 is enriched in alveolar macrophages that also appear to be downregulated in response to LPS; Cluster 80 is enriched in alveolar macrophages and transcripts that also appear to be increased in expression in response to LPS.

Using GATHER to examine the 230 transcripts within cluster 5, the top GO terms are cell communication, signal transduction, response to biotic stimulus, organismal physiological process, immune response, defense response, sensory perception of chemical stimulus and
response to pest, pathogen or parasite, all of which underline the expected functional phenotype of these cells.

The list of transcripts in cluster 5 that are specifically enriched in sheep AM appears to lack many of the transcripts that discriminate AM from other macrophages that have been recognised in mice or pigs (Shibata et al., 2001, Gautier et al., 2012, Kapetanovic et al., 2013, Misharin et al., 2013, Xue et al., 2014). No genes encoding the classic specific AM markers such as SIGLEC7, CD163, or the GM-CSF receptor (CSF2RA) appear to be selectively expressed by sheep AM, using these MCL clustering parameters. All these genes were robustly expressed in the sheep AMs, as are both transcript variants for the transcription factor, peroxisome proliferator-activated receptor gamma (PPARG), encoding a nuclear receptor that regulates peroxisomal beta-oxidation of fatty acids and determine the differentiation and identity of AMs (Schneider et al., 2014).

The transcript encoding a fundamentally important transcription factor, Hypoxia-inducible factor (HIF1A) appears in cluster 5. This transcription factor enables organisms to adapt to changes in oxygen tension (Semenza et al., 1994). There is extensive literature on HIF1A in macrophages. Recent evidence connects HIF1A to LPS inducible changes in metabolism (Palsson-McDermott et al., 2015). It may be that constitutive expression in sheep AMs supports a response to changes in oxygen tension in the lung. Cluster 5 also contains the transcript encoding transforming growth factor (TGF) beta receptor 2 (TGFBR2). TGFBR2 binds the cytokine TGF-beta, which is also known to promote the development and homeostasis of AMs (Yu et al., 2017). Mannose-6-phosphate receptor tyrosine kinase insulin like growth factor 2 receptor gene (IGF2R), which encodes a receptor known to take up lysosomal enzymes (Kornfeld, 1992, Kang et al., 1998) also appears in this AM enriched cluster 5.

Sheep AMs also appear to express the gene encoding paired box 2 (PAX2), seen in cluster 5. PAX2 encodes a multidomain transcription factor involved in kidney differentiation (Barua et al., 2014) and the development of the urogenital tract, eyes and CNS. Within the wider atlas, PAX2 appears most highly expressed in these tissues in addition to the alveolar macrophages, suggesting a sheep AM specific novel requirement of this transcription factor.

Both reference transcripts for the C type lectin domain containing 7A (CLEC7A), which encodes dectin-1, are seen in cluster 5. Dectin-1 is the myeloid pattern recognition
receptor for beta-glucan polysaccharides that are derived from fungal cell walls and is required for effective immune response to fungi, even in the presence of opsonins (Taylor et al., 2007, Vautier et al., 2012). This highlights the danger posed from ruminal fungal populations burdening the ruminant lung and risk of fungal dissemination. Dectin-1 has been found to have an important role in non-pathogen mediated inflammation, with ligation in macrophages driving immune suppressive cellular differentiation (Daley et al., 2017), which supports the requirement for ruminant AMs to reduce any pro-inflammatory responsiveness to stimuli, and induce immune suppression in response to stimuli.

The highest expressed transcripts enriched in alveolar macrophages were from genes encoding members of the S100/calgranulin protein family. These calcium binding cytosolic proteins are involved in many intracellular and extracellular functions (Marenholz et al., 2004, Donato et al., 2013, Gross et al., 2014, Xia et al., 2017). The genes S100A12 and S100A8 appear in cluster 5, and S100A9 appears in cluster 80. S100A8 and S100A9 are known to form a complex and all three of these myeloid related proteins are released in response to stress and inflammation and to mediate responses through activating the cell surface receptors RAGE, TLR4, CD147 or GPCR which influence cell differentiation, inflammatory cell migration, cell apoptosis, cell proliferation, inflammatory cytokine production and tissue repair in macrophages (Foell et al., 2004, Xia et al., 2017). They can be used as biomarkers associated with certain diseases, with S100A12, S100A8, S100A9 (and S100B, which is also in cluster 5 but not as highly expressed as the others), being linked to auto-inflammatory diseases, stroke and trauma (Basta et al., 2006, Nakashima et al., 2010, Zhao et al., 2013, Yayar et al., 2015). The high expression of these transcripts in AM is shared with pigs (Kapetanovic et al., 2013).

Extracellular S100A8, S100A9 and S100A12 all mediate vascular inflammation through the RAGE receptor which is a multi-ligand receptor for advanced glycation end products which results in a cascade of transcription involving NFkB. This is particularly relevant humans, during diabetes and hyperglycaemia (Yao & Brownlee, 2010). It is thought that these proteins therefore mediate a further increase the generation of proinflammatory, prothrombotic and ROS molecules, exaggerating cellular damage even more (Goyette et al., 2009, Yan et al., 2010). S100A12 was the most highly expressed transcript specifically in the sheep AMs and S100A12 is known to be granulocyte-specific and able to act independently of S100A8 and S100A9 during calcium-dependent signalling.
Mice have no homolog for the gene encoding S100A12 but, when transgenic mice had smooth muscle cell targeted expression of S100A12, they demonstrated increased coronary and aortic calcification and were more vulnerable to plaques (Hofmann Bowman et al., 2011). S100A8 KO mice are embryonic lethal, however S100A9 KO mice are protected against mortality induced by endotoxic shock which suggested a detrimental role of these proteins during systemic infection and inflammation (Vogl et al., 2007). The high expression seen in sheep AMs is contradictory to this, highlighting differences between sheep and mice and the dangers of relying upon mice to provide a model for human medicine. Using Ensembl to compare the regions between humans and sheep, sheep have identical synteny with humans, with the regions containing all the S100 genes in close proximity to one another and appearing identical on Chromosome 1 (Aken et al., 2016). Taken together this suggests that sheep are more likely to resemble humans in the transcriptional regulation and control of expression of these genes and would provide a much better model for improving understanding and extrapolating their functional importance in innate immunity in future.

With respect to the lung, S100A12, S100A8/A9 are all significantly increased in human lung disease (Lorenz et al., 2008). Increased levels of S100A12 relative to the heterodimer S100A8/A9, are characteristic of respiratory diseases with neutrophilic inflammation, with neutrophils known to be the first cells to migrate into the pulmonary compartment (Nauseef & Borregaard, 2014, Cantin et al., 2015). The S100A8/A9 heterodimer has potent chemoattractant capabilities, mediating neutrophil influx following LPS stimulation in mice (Vandal et al., 2003). The differential expression of high S100A12 relative to S100A8/A9 can distinguish acute respiratory distress syndrome from chronic disease such as cystic fibrosis and chronic obstructive lung disease, with the ratio increasing during times of acute exacerbations in chronic disease patients (Lorenz et al., 2008).

Many transcripts in sheep macrophage clusters discussed previously highlight an importance for metal ions and amongst the most highly expressed transcripts in cluster 5 are those encoding genes for metal ion transporters. **TFRC (Transferrin receptor)** which is known encode a receptor responsible for cellular uptake of iron, is highly expressed as is **SLC31A1**, encoding a high affinity copper transporter and the paralog SLC31A2, encoding a copper transporter, all of which appear in cluster 5. The known macrophage specific gene SLC11A1 encoding a protein involved in iron and manganese metabolism (NRAMP1), which
is associated with resistance to many intracellular pathogens (such as tuberculosis and leprosy) is also present in this cluster (Gruenheid et al., 1997, Biggs et al., 2001, Forbes & Gros, 2001, Forbes & Gros, 2003, Fritsche et al., 2007). This suggests that sheep AMs are utilising the sequestration of metal ions to protect themselves from their own generation of ROS as well as effectively denying cations to engulfed pathogens.

Also amongst the top 10 most highly expressed transcripts of cluster 5 is ENSOART00000022207/ENSOARG00000020393 which has been annotated as Lysozyme-like. Examination of ENSEMBL indicates a genuine duplication of LYZ that is shared with cows and goats. Mice also have a duplication of lysozyme; one form LYZM is expressed by macrophages and the other, LYZP by Paneth cells in the gut. Lysozyme is known to be bacteriolytic and enhance the activity of immunoagents (Ragland & Criss, 2017), suggesting that ruminant macrophages have evolved to utilise this effective means of intracellular killing without triggering inflammation. It has been shown that the respiratory tracts of healthy sheep have dramatically different bacterial communities, dependent upon the depth within the lung, with fewer microbes the further distal from the glottis (Glendinning et al., 2016). Taken together, this suggests that sheep AMs are highly effective in providing surveillance and clearance of pathogens and have harnessed useful mechanisms to kill whilst holding back the pro-inflammatory response.

In summary, it seems that AM in sheep have distinct transcriptional profiles from those of pigs and mice, utilising different transcription factors, and their responsiveness to microbial stimuli may be somewhat attenuated to avoid excessive response and inflammation to inhaled micro-organisms. The expression of the S100/calgranulin protein genes suggests that they (especially S100A12) may be conferring supportive mediation of innate immunity, indicative of innate immune response in the face of hyperlipidaemia and increased metal ion sequestration.

4.4 Conclusion

A great evolutionary biologist Dobzhansky wrote an essay “Nothing in Biology makes sense except in the light of evolution” (Dobzhansky, 1973). Clearly the evolution of the rumen and increased circulating long chain fatty acids and lipids and associated metabolic adaptations of sheep, have had an impact on the genes that are expressed by their immune cells.
Many highly conserved cellular markers (such as CD163, CD14, and CD36 which was assigned to ENSOARG00000017123) and transcription factors (HIF1A, MAFB (assigned to ENSOARG00000001355) and XBP1 (assigned to ENSOARG00000002962)), have been identified. For example, the very high expression of S100A12 relative to S100A8 and S100A9 in the AMs has supported the view that S100A12 is protective/ supportive of macrophage surveillance and stimulation without inducing proinflammatory cytokine production as opposed to the suggestion that all these calgranulins perpetuate proinflammation in the face of atherosclerosis in humans. Improved understanding of the gene expression signature in healthy sheep macrophages has not only improved functional annotation of the reference genome but has revealed mechanisms by which the innate immune system of sheep harnesses the abundance of lipid and restrains these environmental conditions from leading to systemic inflammation, as is the case in monogastric animals.
Chapter 5  Detailed analysis of the sheep BMDM response to LPS

5.1 Introduction

Chapter 4 of this thesis focussed on the identification of transcripts enriched in resting sheep macrophages both in culture (BMDM and MDM) and in situ (AM). This chapter analyses the transcriptomics of the response of BMDM to the well-defined TLR4 agonist, LPS. The interaction between macrophages and LPS elicits a temporal cascade of gene expression changes, which has been analysed in multiple species including mice (Schroder et al., 2012, Raza et al., 2014), humans (Baillie et al., 2017) and pigs (Kapetanovic et al., 2012). On the basis of those studies, and to enable comparison to the published data in other species, time points chosen were 0hr (unstimulated), 2hr (direct early response), 4hr (mid early response), 7hr (peak inflammatory response) and 24 hr (resolution phase of response) (Sweet & Hume, 1996, Hume et al., 2001, Nilsson et al., 2006, Schroder et al., 2012, Baillie et al., 2017). Other members of the laboratory have used the same time points for studies in goat, cattle, water buffalo, rat, horse and chicken macrophages, enabling a global analysis of species diversity in the response (unpublished). Genes associated with immunity are under strong evolutionary selection. The underlying hypothesis is that novel mechanisms and patterns of gene expression have evolved that distinguish sheep from other species. Some of these unique features might be shared with other ruminants which also share exposure to rumen microorganisms and dietary inputs.

The signalling pathway triggered by TLR activation is described in detail in the introduction (Section 1.2.2). The optimal interaction between LPS and TLR4 involves three proteins, LPS binding protein (LBP), CD14 and MD-2 (see Section 1.5). The recruitment of distinct adaptors activates two separable effector pathways; 1) TIRAP/MyD88-dependent pathway, responsible for pro-inflammatory cytokine expression and 2) the TRAM/TRIF MyD88-independent pathway signalling, responsible for expression of Type I interferons and IFN inducible genes (Lu et al., 2008, Hume, 2015) (Section 1.5). Inducible feedback regulation of this signalling involves dissociation of the adaptor complexes, and expression of inhibitory regulators that impact at every level of the signalling cascade (Kondo et al., 2012, Anwar et al., 2013, Baillie et al., 2017). The response to LPS shows considerable evidence of variation between individual humans (Fairfax et al., 2014) and inbred mouse strains (Raza et al., 2014). Accordingly, this project compared six individual animals, looking for
both common pathways shared by all animals and differences between individuals. The animals were a first cross between two disparate breeds of sheep, the Scottish blackface and Texel, both of which will have been subjected to divergent selection for different traits (see Section 1.8). This intent was to maximise heterozygosity and therefore the numbers of expressed biallelic SNV variants to enable detection of differential expression of the parental alleles in RNA-Seq data and it is accepted that this strategy reduces the power of expression-based analyses such as RNA-Seq. Examination of allele specific expression is beyond the scope of this thesis but future plans involving these investigations will be discussed further in Chapter 6.

Activation of mouse macrophages elicits changes in glucose, glutamine and fatty acid metabolism that meet the energy requirements of the cell, but also lead to accumulation of metabolic intermediates with signalling functions (Biswas & Mantovani, 2012, El Kasmi & Stenmark, 2015, Kelly & O’Neill, 2015, Baseler et al., 2016, Mills & O’Neill, 2016, Langston et al., 2017, Mazzone et al., 2017).

It is unclear whether these metabolic responses are species-specific. Many LPS-inducible gene expression changes are not conserved (Schroder et al., 2012). The Hume lab has identified conserved and divergent features of the LPS response of macrophages of mice, humans and pigs under comparable ex vivo conditions (Kapetanovic et al., 2012, Schroder et al., 2012, Kapetanovic et al., 2013, Baillie et al., 2017). In this chapter the temporal response of sheep BMDM has been examined in detail to improve understanding of innate immunity in sheep as a species and to identify features that may distinguish sheep from monogastric species. The gene expression patterns from individual sheep have been examined to determine regulated genes that vary most in their levels of expression between individuals at each time point.

5.2 Methods

The methods for macrophage cell culture, LPS stimulation, mRNA isolation and RNA-Seq were described in Chapter 2. The raw RNA-Seq data was processed in two pipelines: Kallisto (described in Chapter 2.10) used the available reference OarV3.1.81 transcriptome. The StringTie pipeline provides detailed exon level co-expression data for every known reference transcript and identifies completely novel transcript models, that are absent from OarV3.1, which was important as the reference OarV3.1 is not a complete or perfect transcript library.
Expression estimates from each pipeline were analysed using the network-based cluster analysis tool Miru (described in detail in Chapter 2.9). Clusters with informative average profiles (e.g. induced with a peak at a defined time) were prioritised for further annotation. Within those prioritised clusters, index genes of known function (e.g. inflammatory cytokines, transcription factors) were associated with genes that currently lacked functional annotation.

Within priority clusters, associated q values from statistical packages were assigned to each transcript. Transcripts that were expressed above a minimal threshold and significantly regulated at any time point were curated manually. Where possible they were assigned a functional name based on their nearest orthologues and conserved synteny with known genes in other species. To improve understanding of the pathways, processes and relationships that were shared between the transcripts of each cluster, the web-based tool GATHER (Chang & Nevins, 2006) (see Chapter 4.1), was utilised, in conjunction with information gathered for each gene using GeneCards (Stelzer et al., 2016).

The sheep BMDM response data was examined within four different network expression graphs generated using Miru:

1) The Sheep Atlas (Clark et al., 2017) network expression graph, shows the BMDM response-specific clusters 45 and 52. These clusters extract transcripts specifically expressed in response to LPS and enriched in stimulated macrophages relative to all other tissues. (See Appendix 5.1)

2) The Immune cell sample network expression graph (See Section 4.3.2 for methods), extracts the BMDM response-specific clusters (See Section 5.3.3 for results and Appendix 4.1 sheet 4) and identifies the set of transcripts that is macrophage and LPS response-specific relative to other immune cells.

3) The BMDM response for individual sheep sample network expression graph using Kallisto expression estimates based on the OarV3.1.81 reference transcriptome (See Section 5.2.1 for methods and 5.3.4 for results, Appendix 5.2. P and q values were obtained using Sleuth), highlights the temporal response that is shared by all individuals, as well as clusters demonstrating individuals to vary in their response.

4) The BMDM response for individual sheep sample network expression graph using StringTie expression estimates (See Section 5.2.2 for methods and 5.3.4 for results,
Appendix 5.3. P and q values were obtained using Ballgown, highlights gene expression patterns of the temporal response, including novel transcript models and genes, that are shared by all individuals, as well as clusters demonstrating individuals to vary in their response.

5.2.1 Establishing a sample network expression graph using Kallisto expression estimates for individual sheep BMDM responses

The expression estimates across all OarV3.1.81 transcripts, derived from polyA-selected libraries from BMDM at each time point, were collated together into a single table (Appendix 5.4). The maximum and minimum expression estimates along with the sample from which that maximum or minimum value was derived are shown and the ratio ((Maximum +1)/(Minimum +1)-1) was also calculated for every transcript. The table was filtered to exclude any transcript with a maximum expression estimate < 1 TPM and a ratio < 1.5. This filtered dataset (Appendix 5.6, the input file for Miru) contained 15,212 transcripts that were then clustered using Miru with a Pearson correlation r=0.95, MCL = 2.2, with 10,304 nodes and 360,840 edges in 284 clusters containing more than 3 nodes (Appendix 5.7). 3,898 nodes, where the expression profile did not resemble that of other transcripts sufficiently to create a cluster, were grouped together by the program as ‘no class’.

5.2.2 Expression estimates for individual sheep BMDM responses using StringTie

All results from StringTie can be found in Appendix 5.6. The StringTie pipeline identified 76,411 transcripts expressed (FPKM>0) at some point over the LPS response. A total of 20,196 of these are reference transcript models, 42,154 novel transcript models to known genes and 14,061 novel transcript models associated with novel genes. Transcript variants of the same gene were grouped together resulting in 32,615 expressed genes (FPKM>0 in any LPS response sample), which includes all reference transcripts and novel gene transcript models derived by StringTie. The maximum and minimum expression estimates (FPKM) and expression ratio ((Maximum +1)/(Minimum +1)-1) was calculated for every gene. The table was then filtered to exclude any gene with a maximum expression estimate < 5 FPKM or an expression ratio < 1.5. This filtered dataset (Appendix 5.8) containing 10,552 transcripts were then clustered using Miru, with a Pearson correlation r=0.85, MCL = 2.2, resulting in 9,674 nodes (genes) and 938,068 edges in 281 clusters, of at least 3 nodes (Appendix 5.9).
5.3 Results

5.3.1 Overview of the sheep temporal LPS response

The expression estimates obtained for each reference transcript using Kallisto are contained in Appendix 5.4 and for StringTie in Appendix 5.6.

Table 5.1 summarises the proportions of different transcript types that were detected over the BMDM time course and the transcriptional diversity of the sheep BMDM response. Nearly 90% of all nuclear-encoded OarV3.1.81 protein coding transcripts were detected in at least one BMDM sample. Even taking an arbitrary expression threshold of TPM >1, 85% of the reference transcriptome was detectable.

Table 5.1: Transcript types utilised over the sheep BMDM response to LPS

Biomart was used to determine the transcript type and Kallisto was used to estimate expression level. All transcripts demonstrating a maximum TPM > 0 across all polyA-selected LPS response samples have been included. The ‘Mt RNA’ transcript type includes both Mt rRNA and Mt tRNA. ‘Pseudogenes’ transcript type includes processed-pseudogenes and other pseudogenes.

<table>
<thead>
<tr>
<th>Transcript Type</th>
<th>Transcripts expressed over LPS response</th>
<th>Reference Transcriptome</th>
<th>Percentage (%) OarV3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>lincRNA</td>
<td>1,609</td>
<td>2,020</td>
<td>80</td>
</tr>
<tr>
<td>miRNA</td>
<td>817</td>
<td>1,305</td>
<td>63</td>
</tr>
<tr>
<td>Mt_RNA</td>
<td>20</td>
<td>24</td>
<td>83</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>229</td>
<td>290</td>
<td>79</td>
</tr>
<tr>
<td>protein_coding</td>
<td>20,503</td>
<td>22,823</td>
<td>90</td>
</tr>
<tr>
<td>rRNA</td>
<td>141</td>
<td>305</td>
<td>46</td>
</tr>
<tr>
<td>snoRNA</td>
<td>542</td>
<td>756</td>
<td>72</td>
</tr>
<tr>
<td>snRNA</td>
<td>594</td>
<td>1,234</td>
<td>48</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>24,455</strong></td>
<td><strong>Total 28,757</strong></td>
<td><strong>85</strong></td>
</tr>
</tbody>
</table>

Table 5.2 Protein coding and total number of transcripts (OarV3.1.81) detected at each BMDM time point following LPS stimulation

Examining Kallisto expression estimates from the six adult sheep for each time point and extracting the number of OarV3.1 protein coding and total number of transcripts demonstrating a TPM > 0 in at least one sample.

<table>
<thead>
<tr>
<th>LPS response timepoint</th>
<th>0hr</th>
<th>2hr</th>
<th>4hr</th>
<th>7hr</th>
<th>24hr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein coding transcripts detected across all individuals</td>
<td>18,270</td>
<td>18,170</td>
<td>18,053</td>
<td>18,321</td>
<td>18,260</td>
</tr>
<tr>
<td>Total number of OarV3.1.81 Transcripts detected across all individuals</td>
<td>21,032</td>
<td>20,928</td>
<td>20,778</td>
<td>21,130</td>
<td>21,027</td>
</tr>
</tbody>
</table>
The Kallisto reference index included a total of 29,757 transcripts with a unique Ensembl transcript ID and associated gene ID. A total of 10,269 of these transcripts currently have no meaningful functional annotation or gene name. All RNA-Seq data from this project has been made available to Ensembl.

A principal components analysis (PCA) of the polyA-selected time-course samples was created using R (utilising Kallisto expression estimates). The PCA plot coloured by time points is shown in Fig 5.1A. In Fig 5.1B, each animal is coloured separately. The time-course is cyclical, with the 0hr and 24hr time points being similar to each other, and 4hr and 7hr samples demonstrating the extremes of each individual’s response. This circular expression profile over the time course mirrors that which has been seen in the much denser time course of LPS stimulation of human macrophages based upon CAGE tag sequencing (Baillie et al., 2017).

In Figure 5.1B, the three females have all been coloured in shades of pink and red, and the males all in blue. The PCA suggests a possible association of response with sex, however statistically the individual variation is too great and numbers of male and female animals too small to draw any significant conclusions regarding sex influencing the innate response.
Figure 5.1 PCA plot of all polyA-selected BMDM LPS response samples

The biplots were created using Kallisto expression estimates (TPM) for every reference transcript (OarV3.1) across all BMDM response samples from every individual. The % of Variance is shown in brackets for each principal component.

A. PCA plot of BMDM response samples - coloured by time point
0hr as blue, 2hr as green, 4hr as yellow, 7hr as red and 24hr as purple.

B. PCA plot of BMDM response samples - Coloured by animal
Female 1 as red, Female 2 as bright pink, Female 3 as pink, Male 1 as dark blue, Male 2 as light blue and Male 3 as turquoise. Given the number of individuals (3 male and 3 female) and the degree of variation in response, there was not enough power to examine the effect of biological sex on the response.
Figure 5.1B shows the significant divergence in the response between individual animals. In every animal, the maximal peak in response to LPS, (the greatest divergence from the 0 hr expression pattern), occurred between 4 and 7 hr. The individual profiles were most similar to each other at zero time, and most divergent at 24 hr. There was a general trend for the 24 hr samples to return towards the unstimulated state. The exception was male 3, which showed no evidence of decay of the response. This phenotype, an apparent lack of feedback control, has been well documented in the control of IL1 production in mice (Brandwein et al., 1987). The response of Female 1, on the other hand, peaked early and rapidly returned to the baseline, suggesting that this sheep is an ‘early responder’, again a previously documented phenotype in mice (McDermott et al., 2012).

5.3.2 BMDM response clusters from The Sheep Atlas: Clusters 45 and 52

The objective of this section was to extract, analyse and functionally annotate as far as possible the transcripts selectively enriched in LPS-stimulated BMDM from the main Sheep Atlas (Clark et al., 2017). The sheep atlas provides a unique dataset of RNA-Seq data from multiple organ systems across many stages of development ranging from blastocyst to adult sheep from which to extract co-expression information for any given transcript. The clusters enriched in LPS-stimulated BMDM relative to the entire atlas included cluster 45 (containing 44 genes) and cluster 52 (containing 31 genes) (Clark et al., 2017). Where possible, genes possessing only an Ensembl ID (15 genes/75 total genes) have been manually assigned a functional annotation (Appendix 5.1). Every gene identified in these two sheep clusters is also included in later network analyses derived using StringTie or Kallisto (See 5.3.4) or the averaged immune cell analysis (See 5.3.3). The few that do not appear in Section 5.3.4 such as ENSOARG000000014841 (which has been assigned the functional annotation CCL1), SOCS3, SLC1A5, TNFSF15 and PLAUR, did not show reproducible patterns of regulation in the six individuals and therefore failed to cluster.

GATHER produced identical GO terms for both clusters of genes, including the genes assigned a manual annotation: immune, inflammatory and defence responses; responses to biotic stimulus, wounding, pest pathogen or parasite; external biotic stimulus; stress; communication; taxis; morphogenesis; signal transduction; apoptosis and cell death. These GO terms are consistent with the known biology of the innate immune response.

Many genes included in these clusters are shared with macrophage response clusters in other species, including IL1A and IL1B, IL1RN, SOD2, multiple TNF superfamily members and
C-C and C-X-C motif chemokines (Freeman et al., 2012, Schroder et al., 2012, Baillie et al., 2017).

Table 5.3 summarises the small number of genes within clusters 45 and 52 that have a distinctive expression profile to those reported in comparable published human, mouse and pig macrophage LPS time course datasets in which cells were cultured with the same sources of CSF1 and LPS (Freeman et al., 2012, Schroder et al., 2012, Baillie et al., 2017).

Table 5.3 Genes identified in the BMDM LPS response clusters (45 and 52) of The Sheep Atlas which have been found to vary in expression pattern across other species.

<table>
<thead>
<tr>
<th>Ensembl Gene ID and gene name</th>
<th>Function</th>
<th>Expression in other species</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSOARG00000006889 SERPINB2 (Assigned)</td>
<td>regulation of protease functions</td>
<td>LPS responsive in all species</td>
</tr>
<tr>
<td>ENSOARG00000005159 SERPINB2-like (Assigned)</td>
<td>regulation of protease functions</td>
<td>Duplication shared with cattle (LPS inducible)</td>
</tr>
<tr>
<td>ENSOARG00000000835 PNP (Assigned)</td>
<td>catalyses the phosphorolysis of purine nucleosides</td>
<td>Similar response in mice, mild in human, down in pig</td>
</tr>
<tr>
<td>ENSOARG00000002754 PTX3 (Assigned)</td>
<td>regulation of innate resistance</td>
<td>Similar in human, minimal in mice, not in pig</td>
</tr>
<tr>
<td>ENSOARG00000015857 METRNL</td>
<td>Hormone that promotes energy expenditure</td>
<td>Similar in mice, marginal response in human, down in pigs</td>
</tr>
<tr>
<td>ENSOARG00000009176 PVR (Assigned)</td>
<td>Transmembrane glycoprotein</td>
<td>No response in human</td>
</tr>
<tr>
<td>ENSOARG00000021170 ARG2 (Assigned)</td>
<td>catalyses hydrolysis of arginine</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>ENSOARG00000010784 SLC1A5</td>
<td>sodium-dependent neutral amino acid transporter</td>
<td>No response in mice, down in humans and pig</td>
</tr>
</tbody>
</table>

Manual annotation of the most differentially expressed genes in cluster 52 revealed a gene duplication of SerpinB2/PAI-2, which is shared with cattle (Cattle SERPINB2: ENSTAG0000023198/ ENSTAT0000031553 and Cattle SERPINB2-like: ENSTAG0000023026/ ENSTAT0000031287). In sheep, SERPINB2 (ENSOARG0000006889) had a maximum expression of 9,003 TPM and an expression ratio
The second SERPINB2-like gene (ENSOARG00000005159), also featuring in cluster 52, has two reference transcript models (ENSOART00000005620 with 7 exons and ENSOART00000005619 with 8 exons), both of which were increased in response to LPS in all six sheep, with maximum expression estimates of 29 TPM and 2,859 TPM respectively.

Further evidence that this gene duplication was not a genome assembly error was obtained by downloading cDNA sequence data from Ensembl and producing cladogram phylogenetic trees using Clustal Omega (Sievers et al., 2011, McWilliam et al., 2013, Li et al., 2015) and Figtree (http://tree.bio.ed.ac.uk/software/figtree/), to examine all SERPINB2 genes in a range of different species and all SERPINB genes in both cattle and sheep (not shown).

The greatest homology for the duplicated gene was with SERPINB2 for both sheep and cattle. SERPINB2 is profoundly inducible in mouse macrophages (Costelloe et al., 1999) and appears to protect macrophages against apoptosis (Kumar & Baglioni, 1991, Park et al., 2005) and restrict LPS-induced inflammation (de las Casas-Engel et al., 2013, Zhao et al., 2013). Deficiency in SERPINB2 in mice is linked to impaired CCL2-mediated macrophage influx into the small intestine during infection (Shea-Donohue et al., 2014), and the transcript was upregulated in enteric pathogen infections in both humans and mice (Varro et al., 2004, Schroder et al., 2010, Zhao et al., 2013). The Hume lab has generated RNA-Seq data from cattle BMDM, with and without LPS stimulation (Young et al., 2018). The same expression picture for each SERPINB2 gene was evident in the cattle data (not shown) suggesting a ruminant specific LPS response adaptation and requirement for a second inhibitor urokinase-type plasminogen activator.

5.3.3 The BMDM response clusters from the immune cell averaged expression network expression graph

To gain an overview of changes in expression across the time course, an initial analysis was performed using values averaged across individuals at each time point (see Section 4.3.2 for methods). A network graph of immune-related cells and tissues including the LPS time course data produced eight macrophage-related and/or LPS-induced clusters as shown in Figure 5.2. These contained a total of 483 reference transcripts. A total of 91 transcripts lacked any functional annotation and were assigned a functional annotation where possible. All details of transcripts contained within each cluster can be found in Appendix 4.1.
Figure 5.2 Identification of LPS responsive BMDM enriched clusters

The clusters containing BMDM response transcripts that increase in expression in response to LPS, were extracted from the network analysis used in chapter 4. The co-expression network layout was based upon average expression across Sheep Atlas immune cell types, including transcripts expressed >30 TPM in any sample. The network contains 6,654 nodes, 1,210,699 edges and was based on a Pearson correlation co-efficient >0.9. Edges (connecting grey lines) represent the correlation between expression patterns of transcripts at a Pearson correlation coefficient of 0.9 or greater. Nodes of the same colour form a cluster. The image on the left shows the whole region; the central image shows the BMDM response clusters in this network. The corresponding signal histograms showing the average expression of transcripts within each cluster are displayed next to the extracted cluster. From left to right, the bars are labelled along the X axis as: BMDM 0hr, 2hr, 4hr, 7hr, 24hr post LPS, AM, BL, MDM, PBMC. All transcripts identified can be examined in Appendix 4.1, LPS responsive worksheet.

Table 5.4 summarises transcripts identified which appear to show a unique expression pattern in sheep, their inferred function and comparison to the published data on regulation in other species (Freeman et al., 2012, Schroder et al., 2012, Baillie et al., 2017). All references to comparative expression in the text below refer to the same datasets unless otherwise indicated.
Table 5.4 Sheep-specific LPS responsive genes

Transcripts and their corresponding gene ID and names were identified from the averaged immune cell analysis, and found to differ absolutely in expression pattern in comparison to other species (Freeman et al., 2012, Schroder et al., 2012, Baillie et al., 2017). The function assigned to each gene is based upon information from www.genecards.org.

<table>
<thead>
<tr>
<th>Ensembl IDs and gene name</th>
<th>Function of encoded protein</th>
<th>Expression in other species</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSOART00000008204 ENSOARG00000007533 ADCYAP1R1</td>
<td>Type I adenylate cyclase activating polypeptide receptor, mediated by G proteins. Activating adenylate cyclase involved in transmembrane signalling and shares homology with members of the glucagon/secretin receptor family.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>ENSOART00000017157 ENSOARG00000015761 HTR7</td>
<td>5-Hydroxytryptamine Receptor 7. Receptor for 5-hydroxytryptamine (serotonin), a biogenic hormone that functions as a neurotransmitter, a hormone, and a mitogen. Mediated by G proteins that stimulate adenylate cyclase.</td>
<td>Response shared with pig, less so in mouse, no response to LPS in human</td>
</tr>
<tr>
<td>ENSOART00000023097 ENSOART00000023098 ENSOARG00000021196 TTC9 (assigned)</td>
<td>Tetramericopeptide Repeat Domain 9</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>ENSOART00000016356 ENSOARG00000015024 PRKAR1A</td>
<td>Protein Kinase CAMP-Dependent Type I Regulatory Subunit Alpha. Activates the cAMP-dependent protein kinase, which transduces the signal through phosphorylation of different target proteins.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>ENSOART00000004064 ENSOART00000004065 ENSOARG00000003740 SUCNR1</td>
<td>Succinate Receptor 1. G-protein-coupled receptor for succinate, an intermediate molecule of the citric acid cycle.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>ENSOART00000009357 ENSOARG00000008597 SGMS1</td>
<td>Sphingomyelin Synthase 1. Five-pass transmembrane protein. Synthesise the sphingolipid, sphingomyelin, through transfer of the phosphatidyl head group, phosphatidylcholine, on to the primary hydroxyl of ceramide. The reaction is bidirectional depending on the respective levels of the sphingolipid and ceramide.</td>
<td>Responsive in sheep, pig and mouse, not in human</td>
</tr>
<tr>
<td>ENSOART00000009357 ENSOARG00000008597 SGMS2</td>
<td>Sphingomyelin Synthase 2. Enzyme that catalyses production of Sphingomyelin (which affects insulin sensitivity), a sphingolipid and major component of cell and Golgi membranes, required for cell growth. Regulates cell surface levels of ceramide, mediating signal transduction and apoptosis.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>Ensembl IDs and gene name</td>
<td>Function of encoded protein</td>
<td>Expression in other species</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td><strong>ENSOART00000017366</strong></td>
<td>Glutaredoxin 5, Mitochondrial protein, involved in the biogenesis of iron-sulphur clusters, required for iron homeostasis. Involved in protein lipoylation. Required for normal regulation of haemoglobin synthesis by the iron-sulphur protein ACO1.</td>
<td>Decreased in response to LPS in mouse, human and pig</td>
</tr>
<tr>
<td><strong>ENSOART00000012596</strong></td>
<td>Cytochrome P450 Family 26 Subfamily B Member 1. Monoxygenase which catalyse many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. Regulator of all-trans retinoic acid levels by the specific inactivation of all-trans retinoic acid to hydroxylated forms.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td><strong>ENSOARG00000000937</strong></td>
<td>Cytidine Deaminase. Enzyme which scavenges exogenous and endogenous cytidine and 2-deoxycytidine for UMP synthesis, may also participate in RNA editing.</td>
<td>Decreased in response to LPS in human</td>
</tr>
<tr>
<td><strong>ENSOARG00000000875</strong></td>
<td>Adenylsuccinate Synthase. Enzyme which catalyses conversion of inosine monophosphate to adenosine monophosphate.</td>
<td>Decreased in response to LPS in mouse, human and pig</td>
</tr>
<tr>
<td><strong>ENSOART00000006066</strong></td>
<td>Cellular Retinoic Acid Binding Protein 2. Member of the retinoic acid binding protein family and lipocalin/cytosolic fatty-acid binding protein family.</td>
<td>Marginal response in pig, absent in mouse. Decreased in response to LPS in human</td>
</tr>
<tr>
<td><strong>ENSOARG00000005565</strong></td>
<td>Wolframin ER Transmembrane Glycoprotein. Partly regulates cellular Ca(^{2+}) homeostasis, by modulating the filling state of the endoplasmic reticulum Ca(^{2+}) store.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td><strong>ENSOART00000006465</strong></td>
<td>Semaphorin 6A. Expressed in developing neural tissue and required for proper development of the thalamocortical projection, promoting reorganization of the actin cytoskeleton and important axon guidance cue.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td><strong>ENSOARG00000005941</strong></td>
<td>Tenascin C. Extracellular matrix protein, containing multiple EGF-like and fibronectin type-III domains.</td>
<td>Induced by LPS in human, unresponsive in pig and mouse</td>
</tr>
<tr>
<td>Ensembl IDs and gene name</td>
<td>Function of encoded protein</td>
<td>Expression in other species</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>ENSOART00000019500</td>
<td>Shootin 1. Important role in regenerative neurite outgrowth and the accumulation of phosphatidylinositol 3,4,5-trisphosphate (PIP3).</td>
<td>Downregulated in mouse and pig LPS response, no response in humans</td>
</tr>
<tr>
<td>ENSOARG000000017919</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHTN1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSOART00000018427</td>
<td>Coagulation factor II thrombin receptor like 2. G-protein-coupled receptor that encodes protease-activated receptor-3.</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>ENSOARG00000016918 F2RL2(PAR3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xanthine Dehydrogenase. Molybdenum-containing hydroxylase involved in purine degradation, regulating adipogenesis and peroxisome proliferator-activated receptor-γ (PPAR-γ)</td>
<td>Sheep specific</td>
<td></td>
</tr>
<tr>
<td>ENSOART00000011926</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSOARG00000010953 XDH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XDH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSOART0000000285</td>
<td>CXADR like membrane protein. Type I transmembrane protein and involved in cell-cell adhesion</td>
<td>Sheep specific</td>
</tr>
<tr>
<td>ENSOARG00000000271 CLMP (ACAM)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All time course samples were PolyA-selected for the RNA-Seq library prep method (see chapter 2.7) which should remove most non-coding transcripts. Of the 15 transcripts that were not protein-coding, yet apparently highly expressed and regulated in response to LPS, 8 are lincRNAs, 7 are miRNAs, 4 are pseudogenes, 2 are rRNAs, 1 is a snoRNA and 4 were snRNAs, suggesting that these transcripts belonging to other classes of RNA are functionally important to the sheep in responding to LPS.

From these selected BMDM LPS responsive enriched clusters from the immune cell network expression graph, 73 reference transcripts lacked any functional name. All have been examined in Ensembl, their nearest orthologues and syntenic regions compared and they have been assigned a ‘Preliminary annotation’ (see Appendix 4.1, LPS responsive worksheet).

Clusters 13 and 17 contain transcripts that were induced transiently, peaking after 2 hours. The basal level of expression is broadly shared by all other immune cells in cluster 17. However, in cluster 13, the transcripts are also enriched in the PBMCs (Figure 5.2). Transcripts encoding inflammatory cytokines such as TNF, with associated signal transduction molecule TRAF1, along with genes for surface receptors such as CD83, early transcription factors such as FOSL1 and feed forward activators such as TRAF1, TICAM1, RCAN1, IRAK2 all featured in this first ‘wave’ of increased transcription. The pattern of
expression for these transcripts is largely shared with mouse and human (Nilsson et al., 2006, Baillie et al., 2017). Amongst the highest expressed transcripts was ENSOART00000012146 annotated as CEBPD, a gene encoding a transcriptional activator, and ENSOARG00000009537 annotated as DUSP5. Transcripts with the highest expression ratio in cluster 13 include a novel miRNA (ENSOARG00000024480) and NR4A3, which encodes a transcriptional activator and member of the steroid-thyroid hormone-retinoid receptor superfamily.

Clusters 7, 9 and 24 contain transcripts that also increased transiently, peaking at 4 hr in response to LPS. Cluster 9 is specific to the BMDM response. Cluster 7 shows some enrichment in the MDMs, possibly reflecting the similar culture methods as MDM and BMDM are cultured in the presence of rhCSF1. Cluster 24 shows some enrichment in BLs and PBMCs as well as increasing in response to LPS in the BMDM samples.

Cluster 9 contains multiple chemokines and cytokines including IL6. Some were previously unannotated: ENSOARG00000025179 annotated as CXCL1, ENSOARG00000014841 annotated as CXCL2. The cluster also contains genes for multiple signalling molecules, including the tyrosine kinase JAK2 and downstream kinase MAP2K3, transcription factors such as FOSL2 and nuclear hormone receptor Vitamin D receptor (VDR) and the receptors TLR2 and TLR4. GPR84 (encoding a G protein-coupled receptor 84, detecting C9-C14 medium chain free fatty acids) is also found in cluster 9, sharing the same dramatic increase in expression seen in humans, mice and pigs.

Cluster 24 contains genes for the key enzyme in prostaglandin biosynthesis, cyclooxygenase 2 (PTGS2) (Kawahara et al., 2015, Tang et al., 2017) and NIPAL1, a transmembrane transporter of divalent cations (Goytain et al., 2008) alongside known acute response genes that are associated with TRAF signalling like BIRC3, TRAF2 and TANK (Hacker et al., 2011).

Clusters 6, 18 and 26, containing transcripts that peak at 7hrs.

The largest, cluster 6 includes further inflammatory cytokine genes such as the colony stimulating factors CSF1, CSF2 and CSF3, Interleukins IL1A and IL1B and many C-C and C-X-C motif chemokines that are also amongst the most highly-induced transcripts in all other species examined. Inflammatory resolving transcripts include both transcript models for the cytokine-inducible negative regulator of cytokine signalling, CISH. Genes encoding less
well-known mediators that are nevertheless shared with other species include *EREG* and *INHBA* and transcription factors *ETS2* and *AEBP2* (assigned to *ENSOARG00000020470*).

Many of the 30 transcripts contained in cluster 18 centre around transcriptional control, growth, adhesion and cytoskeleton organisation and endocytic trafficking, such as *ATF6*, *PPARD*, *SHTN1* and the highly inducible novel miRNA *ENSOART00000025421/ENSOARG00000021041*.

Cluster 26 contains *ENSOARG00000005159*, one of the most differentially expressed transcripts which has been assigned the functional annotation of *SERPINB2* (see Section 5.3.2). *CCL2* (encoding a monocyte-attracting chemokine) is the most highly expressed transcript in this cluster, a chemokine that mediates monocyte recruitment, enhancing LPS production of IL-10, with high expression indicative of macrophages under the influence of CSF1 (Sierra-Filard et al., 2014). Another very highly expressed transcript, *ENSOART00000010843/ENSOARG00000009963*, was annotated as a serum amyloid gene, *SAA3*, encoding a well-known acute phase protein (Zhou et al., 2017) and sharing an expression pattern with other species. *ADGRE1*, which encodes the widely used macrophage marker, F4/80 in mice (Austyn & Gordon, 1981, Hume et al., 1984) was also highly expressed and further LPS inducible in sheep.

Analysis of the resolution phase of the response at 24 hr necessitates separate examination of the individual response signatures (Section 5.3.9), because of the divergent nature of the response between individuals (Section 5.3.1). There were no shared clusters found to peak in expression in the resolution phase (24 hrs).

### 5.3.4 Network analysis of the BMDM response to LPS in individual sheep.

The initial analyses (Section 5.3.3) used averaged values. However, it was clear from the PCA (Section 5.3.1) that some averages would be highly skewed by the expression level in a single individual. Therefore, the macrophage LPS response was assayed in isolation capturing all regulated transcripts regardless of their restriction to macrophages. The Kallisto sample-sample network in Fig 5.3 (Appendix 5.7 and 5.9) clearly mirrors the PCA plots in Fig 5.1, in comparison to StringTie, which highlights the extreme individual response of Male 3.
Figure 5.3 Sample-based networks comparing StringTie and Kallisto expression estimates from individual sheep BMDC time course samples.
For StringTie, all genes which demonstrated maximum expression > 5 FPKM in any sample across the time course were included. For Kallisto, all transcripts which demonstrated a maximum expression >1 TPM in any sample across the time course were included. All genes and transcripts for both pipelines had an expression ratio \(((Maximum+1)/(Minimum+1)-1) > 1.5\). The Pearson correlation coefficient threshold was 0.92. Each node represents a sample and the lines between them represent correlations of at least 0.92. Nodes are coloured according to the sample time point.

Figure 5.4 shows the gene-based networks for both methods, using the values for individual sheep. The network expression graph using StringTie expression estimates (Figure 5.4 A), is more tightly-clustered than that based on Kallisto expression estimates (Figure 5.4 B) where the circular temporal response is clearly evident (Figure 5.4 C). The smaller discrete clusters, highlighted by clusters 75, 99 and 194, contain miRNAs and snoRNAs specific to certain individuals and were not found to contain highly expressed protein coding genes.
Figure 5.4 BMDM response gene-based network expression graphs

A. Network expression graph using StringTie expression estimates for every BMDM response sample from each individual sheep (see Section 5.2.2 for methods)

B. Network expression graph using Kallisto expression estimates for every BMDM response sample from each individual sheep (See Section 5.2.1 for methods)

C. The Kallisto-based network highlighting clusters of genes that appear regulated during the response to LPS.

Figures 5.4 C and 5.5 emphasise the circular nature of this temporal response, in common with human macrophages (Baillie et al., 2017), and Figure 5.5 highlights exactly where each cluster of interest is to be found within the main network expression graph and the associated GO terms (Chang & Nevins, 2006).
Figure 5.5 Kallisto expression estimate network graph highlighting each key cluster of interest.

Nodes in the relevant clusters for that ‘peaking time point’ are enlarged in each panel for clarity, to show their position within the overall network graph. Transcripts contained within each cluster can be viewed along with the associated signal histogram for each cluster in the relevant ‘peaking’ spreadsheet within Kallisto Peaking Analysis Excel workbook in Appendix 5.2. The GO terms across all clusters peaking at a given time point are derived using GATHER (Chang & Nevins, 2006).
Figure 5.5 Continued.

**Downregulated clusters:** 1, 2, 3, 6, 7

**Shared: down and increased response**
- cell communication
- cell proliferation
- cell surface receptor linked signal transduction
- cellular morphogenesis
- cellular physiological process
- detection of external stimulus
- development
- embryonic development
- G-protein coupled receptor protein signalling
- intracellular signalling cascade
- morphogenesis
- neurophysiological process
- organ development
- organostructural physiological process
- organogenesis
- perception of smell
- phosphate metabolism
- phosphorus metabolism
- phosphorylation
- protein amino acid phosphorylation
- regulation of biological process
- regulation of cell cycle
- regulation of cellular physiological processes
- regulation of cellular process
- regulation of physiological process
- response to external stimulus
- response to stress
- sensory perception
- sensory perception of chemical stimulus
- signal transduction

**Shared: peaking 2hrs, 4hrs, 7hrs and 24hrs**
- cell communication
- cell-cell signaling
- chemotaxis
- defense response
- development
- humoral defense mechanism (sensu Vertebrates)
- immune response
- inflammatory response
- morphogenesis
- organ development
- organostructural physiological process
- organogenesis
- regulation of cellular process
- response to biotic stimulus
- response to chemical substance
- response to external biotic stimulus
- response to external stimulus
- response to pest, pathogen or parasite
- response to stimulus
- response to stress
- response to wounding
- signal transduction
- taxis
5.3.5 Down-regulated transcripts in the response to LPS

In common with the other similar projects in human, mouse and pig, when individual values were analysed, the largest clusters (1-3) contain transcripts that decrease during the response to LPS (as shown in Figure 5.6), reaching a nadir around 7 hours. Many of the genes within clusters 1 and 2 were similarly repressed in response to LPS in mouse, human and pig macrophages. In Female 1 (the early responder) expression returned to pre-stimulation levels at 24 hours, whereas the same transcripts remained down-regulated in Male 3.

Clusters 1 and 2 were enriched for GO terms; sensory perception of chemical stimulus; cell proliferation; cell cycle; detection of external stimulus; G protein coupled receptor protein signalling; cellular physiological process; neurophysiological process; protein amino acid phosphorylation; response to stress; primary metabolism; protein modification.

This pattern reflects the known ability of LPS to block cell proliferation in macrophages, in part by down-regulating the CSF1 receptor (Sester et al., 1999, Sester et al., 2005). The most highly downregulated transcript was adenosine A3 receptor (ADORA3), a gene that is not expressed robustly in either mouse or human macrophages and in the pig appears to increase in response to LPS. This gene encodes a G-protein coupled receptor, that has been implicated in mediating both cell proliferation and death (Salvatore et al., 2000).

Two transcripts encoding uncharacterized proteins (ENSOART00000020222/ENSOARG00000018581 and ENSOART00000020267/ENSOARG00000018624) that are both members (the only members) of the same protein family PTHR24365_SF250, with Toll/interleukin-1 receptor homology domains, were also highly-expressed in unstimulated cells and rapidly down-regulated. Their function is unknown.

In cluster 2 one of the most down-regulated transcripts was ENSOART00000005567/ENSOARG00000005113 assigned to FAM72A, a gene implicated in the regulation of ROS metabolism and cell growth (Schieber & Chandel, 2014, Zhao et al., 2017). In mouse BMDM, FAM72A is also highly-expressed, transiently induced but then repressed by LPS (BioGPS.org). In humans, it does not appear to be expressed in myeloid cells.
Figure 5.6 Identification of clusters with decreased expression over time in response to LPS

A Location of the key downregulated clusters (enlarged nodes) in the wider network expression graph.

B Mean expression of transcripts contained in each cluster across individuals: Female 1 (F1), Female 2 (F2), Female 3 (F3), Male 1 (M1), Male 2 (M2) and Male 3 (M3) and across time points 0hr, 2hr, 4hr, 7hr and 24 hr (left to right for each individual). Each line represents the cluster of the same colour.

C The collection of downregulated clusters. Successive images rotated by 45 degrees.
5.3.6 Inter-individual variation in the early response phase

Clusters which peak in expression at 2hrs after LPS stimulation in most of the individuals have been extracted from the network expression graphs (Figure 5.7) created using values for individual sheep with both Kallisto (Section 5.2.1) and StringTie (Section 5.2.2) and can be viewed in Appendix 5.2 and Appendix 5.3 respectively. Many genes have been assigned a functional annotation, including genes encoding immediate early transcription factors that share similar expression profiles with humans, mice and pigs, EGR1 (assigned to ENSOART000000017938/ ENSOARG000000016470 in cluster 51) and EGR4 (assigned to ENSOART000000012890/ ENSOARG000000011851 in cluster 153). Other inducible transcripts shared with humans and mice include genes encoding AP1 transcription factor subunits, FOSB, FOSL1 and FOSL2 and MAP3K2, encoding a signalling kinase, in cluster 26.

Fig 5.7 A illustrates the positioning of these clusters within the wider network analysis described in Section 5.2.1, showing the close proximity to one another, with cluster 196 (where female 2 shows the greatest expression at 2hrs) more disparate from others such as cluster 97 (Fig 5.2C). Cluster 97 strongly demonstrates inter-individual variation, with female 1 showing particularly high expression for all the transcripts within the cluster. HSPA5 and IL23A are highly expressed and appear to be co-regulated in this cluster. Future work could investigate possible shared transcription factor binding sites for these genes.

Cluster 97 is mirrored in cluster 13 of the StringTie analysis, which contains many of the same genes, including IL23, HSPA5, MCM3Ap, TCEAL1, NRD1 and SLC25A30 (assigned to ENSOARG00000007620/ ENSOART00000008296). A total of 126 genes are contained in StringTie cluster 13, which include 25 novel gene models that are not associated with an Ensembl ID. Many of these are likely spurious models such as MSTRG.4791 which is only 273 base pairs in length and only contains one exon. Conversely, some novel models are more robust, such as MSTRG.3436 which has 3 exons and is 2988 base pairs in length.
Figure 5.7 Visualisation of the clusters demonstrating a regulated peak in response to LPS at 2hrs

A. Key clusters found to peak in expression at 2hrs (enlarged nodes) in the wider network expression graph using Kallisto in 5.2.1.

B. Mean expression of transcripts contained in each cluster of 5.2.1 across individuals: Female 1 (F1), Female 2 (F2), Female 3 (F3), Male 1 (M1), Male 2 (M2) and Male 3 (M3) and across time points 0hr, 2hr, 4hr, 7hr and 24 hr (left to right).

C. Clockwise rotation of the collection of clusters peaking at 2hrs.

D. Signal histogram associated with each cluster peaking at 2hrs, using Kallisto (5.2.1) expression estimates

E. Shows the signal histogram associated for each cluster found peaking at 2hrs, using StringTie (5.2.2) expression estimates, with labels for the x axis, as listed in B.
Clusters 22 and 479 of the Kallisto network are particularly enriched with transcripts peaking at 2 hr that were very highly expressed in Male 3. Included in these clusters are transcripts associated with acute phase inflammation such as TNF, NLRP3 (both in cluster 22) and TRIM, TNFSF9, MAPKAPK2 alongside genes encoding important transcription factors such as TCF7, KLF2 (assigned to ENSOARG00000019122/ ENSOART000000020822) and FOSL1.

The equivalent StringTie cluster, containing transcripts enriched for immediate response genes with the greatest expression in Male 3 above all other individuals is cluster 162. NFKBIZ is identified in this cluster, which is contrary to the Kallisto network analysis that identifies the only reference transcript for this gene (with 11 exons) within cluster 8 to peak at 4hrs. On closer inspection, StringTie has identified 5 additional transcript models for this highly-expressed and regulated transcriptional regulator (not shown). In humans there are 11 splice variants for NFKBIZ, illustrating how this RNA-Seq data having been made available to Ensembl will help improve annotation, coverage and recognition of more transcript variants in the future updated version of the sheep genome OarV5.

**Figure 5.8** Expressed NFKBIZ transcripts across the time course
StringTie expression estimates (FPKM) of the various transcript variants of NFKBIZ across the time course for each individual animal.
5.3.7 Inter-individual variation at the mid early response phase

Clusters which peaked in expression at 4 hr after LPS stimulation in most of the individuals were extracted from the network expression graphs (Figure 5.9) created using both Kallisto (5.2.1) and StringTie (5.2.2) with individual values and can be viewed in Appendix 5.2 and Appendix 5.3 respectively.

For both pipelines, out of all the clusters found to peak in expression at 4 hr, the most highly-expressed and regulated gene was the acute phase inflammatory cytokine *IL1B* (Ren & Torres, 2009). Kallisto assigned expression to a single transcript model: *ENSOART00000022733* (6 exons), whereas StringTie assigned expression estimates to two transcript models: the reference model (6 exons, Chr3: 60023514-60034494) and *MSTRG.21999.1* (8 exons,Chr3: 60022671-60048960). Based upon the StringTie analysis, only males 1 and 2 demonstrate robust, regulated expression of the reference transcript model, peaking at 7 hr (not shown). This suggests a differentially regulated splice variant of *IL1B* exists in some individuals.

Figure 5.9 shows the clusters peaking at 4hrs within the main Kallisto network analysis in greater detail. Clusters 33, 87 and 259 contain the most highly regulated and expressed transcripts encoding many inflammatory cytokines such as *IL1B, IL1A, IL11* and colony stimulating factors *CSF1, CSF2* and *CSF3* and multiple transcripts encoding C-X-C motif chemokines. In each case, male 3 showed the highest expression level.

Male 3 also had the highest expression of the majority of transcripts in cluster 6 from the StringTie network. In addition to those identified in the Kallisto-based network, cluster 6 also contains *IL6* and downstream interferon response and signalling genes such as *IRF3* and *IRF7, IFIT2* and *IFIT5, CCL4, CCL5, CCL8* and *CCL20, RIPK1* and *STAT2*. There are 336 genes in this interferon response dominated cluster and 57 are novel genes not associated
with any Ensembl ID. Two of these novel transcripts *MSTRG.9285* and *MSTRG.5463* were robustly expressed (maximum expression estimates 54 and 34), significantly regulated (Ballgown q values 0.000645 and 0.000000000000177 respectively), relatively long (558 bp and 6117 bp) and encoded by multiple exons.

**Figure 5.9 Visualisation of the clusters demonstrating a regulated peak in response to LPS at 4hrs**

A. Key clusters found to peak in expression at 4hr in the wider network expression graph.

B. Mean expression of transcripts contained in each cluster across individuals: Female 1 (F1), Female 2 (F2), Female 3 (F3), Male 1 (M1), Male 2 (M2) and Male 3 (M3) and across time points 0hr, 2hr, 4hr, 7hr and 24 hr (left to right).

C. Clockwise rotation by 45 degrees, of the clusters peaking at 4hr.
5.3.8 Inter-individual variation at the peak inflammatory response phase

Clusters which peak in expression at 7 hr after LPS stimulation in most of the individuals were extracted from the network expression graphs (Figure 5.10) created using both Kallisto (5.2.1) and StringTie (5.2.2) with individual values and can be viewed in Appendix 5.2 and Appendix 5.3 respectively.

There was significant overlap between the gene lists of both pipelines, with the disparity due in part to the increasingly divergent responses between individuals. The largest, cluster 5, in which Male 3 showed the greatest level of expression (peaking at 4 hr), contains more genes associated with interferon signalling such as STAT2, IRF7 and IRF9 and IFIT1, IFIT2, IFIT3, IFIT5 and IFIT7. This cluster contains a candidate novel IFN regulatory factor. *ENSOART00000001952* has both IFN regulatory factor-3 domain and SMAD/FHA domains and is lowly expressed. Based upon the nearest orthologues and comparison of syntenic regions using Ensembl (Aken *et al.*, 2016), it was found to share a 1:1 orthologue with no functional annotation with 9 other placental mammals, with shared synteny across the region and was listed as having 1:1 orthologues with various fish and aquatics, where it is annotated as *IRF10* (*IRF10* does not exist in humans or mice and was first found in birds and is present in other tetrapods). This gene is also listed as a 1:1 orthologue with *IRF9* in various birds including the chicken (shared synteny). *IRF9* is already annotated in the sheep on a different chromosome (Chr 7), so this gene (which lies on Chr 13) has been assigned the annotation of *IRF9-like*. 
Figure 5.10 Visualisation of the clusters demonstrating a regulated peak in response to LPS at 7hrs
A. Key clusters found to peak in expression at 7hrs in the wider network expression graph.
B. Mean expression of transcripts contained in each cluster across individuals: Female 1 (F1), Female 2 (F2), Female 3 (F3), Male 1 (M1), Male 2 (M2) and Male 3 (M3) and across time points 0hr, 2hr, 4hr, 7hr and 24 hr (left to right).
C. Clockwise rotation by 45 degrees of the clusters peaking at 7hrs.

5.3.9 Inter-individual variation at the resolution of inflammation phase
Clusters which peak in expression at 24 hr after LPS stimulation in most of the individuals have been extracted from the network expression graphs (Figure 5.11) created using both Kallisto (Section 5.2.1) and StringTie (Section 5.2.2) with the individual values and can be viewed in Appendix 5.2 and 5.3 respectively.

There are relatively few resolution phase response clusters that increase in all individuals in the 24hr time point after LPS stimulation, reflecting the different positions of each animal at this time point in the PCA (Figure 5.1). The majority of transcripts centre around metabolism, encoding multiple solute carriers (See Section 5.5), S100 calcium binding proteins and for enzymes such as lactate dehydrogenase (LDHA), fatty acid desaturase 3 (FADS3) and matrix metallopeptidase 25 (MMP25), highlighting the metabolic progression
associated with the inflammatory response. The metabolic changes throughout the response are discussed in greater detail in Section 5.4.

**Figure 5.11 Visualisation of the clusters demonstrating a regulated peak in response to LPS at 24hrs**

A. Key clusters found to peak in expression at 24hrs in the wider network expression graph.

B. Mean expression of transcripts contained in each cluster across individuals: Female 1 (F1), Female 2 (F2), Female 3 (F3), Male 1 (M1), Male 2 (M2) and Male 3 (M3) and across time points 0hr, 2hr, 4hr, 7hr and 24hr (left to right).

C. Clockwise rotation by 45 degrees of clusters peaking at 24hr.

---

**5.4 The metabolic shifts of sheep BMDMs in response to LPS**

**5.4.1 Metabolism in immune cells**

Regulated intermediary metabolism is widely accepted as a key step in the differentiation and functional adaptation of cells of both innate and acquired immune systems (Murray *et al.*, 2015, O’Neill *et al.*, 2016, Gaber *et al.*, 2017). Circulating metabolic substrate concentrations are very different between ruminants and monogastric species. For that reason, there was a particular interest in this study on genes encoding membrane transporters and metabolic enzymes. Glucose, glutamine and fatty acids are all utilised as energy sources for mammalian macrophages. Glucose is utilised to generate ATP (by glycolysis and mitochondrial oxidative phosphorylation), glycerol-3-phosphate (for making
phospholipids and triacylglycerols), NADPH (to produce ROS) and ribose (for RNA synthesis) (Curi et al., 2017).

Mouse and rat macrophages have been shown to metabolise glutamine at a rapid rate (Curi et al., 2017). Inhibition of glutamine synthetase (GSS), which produces glutamine from glutamate, was shown to alter the polarisation state of macrophages (Palmieri et al., 2017). The circulating glutamine concentration in ruminants is 3-5 fold lower than in monogastric species, due to a low glutamine synthetase capacity, and glutamine is not the predominant respiratory fuel for the intestine (Meijer et al., 1993). When there is increased availability of other energy substrates, there is decreased glutamine utilisation in the intestinal wall (Reynolds & Huntington, 1988). In conditions of metabolic stress (such as high milk production in dairy cows), plasma and tissue glutamine pools are both similar to those of most essential amino acids such as methionine and lysine (Meijer et al., 1993).

Fatty acids are also used by mouse macrophage mitochondria, for energy and in lipid synthesis. The phenotypic plasticity of macrophages is known to be influenced by metabolites including glucose, lipoproteins, fatty acids and many nutrients derived from the microbiome (O’Neill et al., 2016, Stienstra et al., 2017). The sheep, as a ruminant, has high circulating levels of fermentative by-products, primarily volatile fatty acids (propionate, acetate and butyrate), which are utilised within the liver for gluconeogenesis (Danfaer et al., 1995). Aside from acting as fuels, free fatty acids may be recognised by specific G protein coupled receptors (FFAR1-4, GPR 84,35,91 and 109A) in cells of the immune system as well as tissues involved in metabolic function such as the pancreas and intestinal epithelium (Alvarez-Curto & Milligan, 2016).

Increased glucose metabolism in activated macrophages and disruption of the TCA cycle is proposed to be essential for the secretion of pro-inflammtory cytokines and production of ROS in mouse macrophages (Freemerman et al., 2014, Jha et al., 2015, Ryan & O’Neill, 2017). Aside from the generation of ATP and metabolites for macromolecule synthesis, regulated glycolytic enzymes such as GAPDH (Nagy & Rigby, 1995, Millet et al., 2016), pyruvate kinase, lactate dehydrogenase, enolase and aldolase, are known to bind to mRNA and regulate protein synthesis (Castello et al., 2012, Donnelly & Finlay, 2015). Hexokinase 1 (HK1) has also been shown to regulate NLRP3 inflammasome activation (Moon et al., 2015).

Fatty acid synthesis in the cytosol, which also depends upon glucose metabolism, is also closely linked to macrophage effector functions (Feingold et al., 2012, Tannahill et al.,
For example, CSF1 up regulates genes required for fatty acid synthesis whilst driving the differentiation of human monocytes, with differentiation associated with a change in lipid synthesis from cholesterol (monocytes) to phosphatidylcholine (macrophages) (Cader et al., 2016). In this respect, human macrophages differ from mouse (Irvine et al., 2009). Conversely, the beta oxidation of fatty acids has been associated with anti-inflammatory macrophages. Lipolysis may fuel mitochondrial oxidative phosphorylation leading to reduced ER stress, ROS damage and inflammation (Malandrino et al., 2015).

In overview, circulating metabolites including products of the microbiome can influence the innate immune response (Yurkovetskiy et al., 2015, Rooks & Garrett, 2016). The original hypothesis for this project (Section 1.7), suggested that ruminants may have distinct regulation of intermediary metabolism in their immune cells. This is likely due to the ruminant symbiotic fermentative microbiome and ruminant distinctive hepatic gluconeogenic metabolism which will alter homeostatic levels of circulating fatty acids and nutrients compared to monogastric species. Based upon that hypothesis, this Section is focussed specifically on the regulation of metabolism-associated genes in sheep macrophages.

5.4.2 Analysis of genes involved in intermediary metabolism

Because of the large number of reference transcripts expressed by sheep macrophages, the thresholds required for visualisation, and the variation in expression between individuals, many metabolic genes of interest were excluded from cluster analysis in previous Sections. To enable a comprehensive analysis, Kallisto expression estimates for the BMDM response (PolyA-selected) samples were averaged across all individuals. This average expression value was then extracted for all reference transcripts involved in intermediary metabolism. A list of genes with the primary annotation of Metabolic Process was downloaded from the MGI database (URL http://www.informatics.jax.org/vocab/gene_ontology/GO:0008152). A total of 7000 genes had secondary annotation indicating the likely process/pathway. That list was curated to contain only genes with a relevant secondary annotation related to intermediary metabolism. The curated list was supplemented with known genes of interest that have more recently been annotated, such as aconitate dehydratase (ACOD1) and genes involved in the uptake of nutrients, their transportation and metabolite conversion. All annotated solute carriers and cytochrome P450 (CYP) family members were included. Genes with more than one transcript variant in the reference appear twice. The ratio of the
maximum averaged expression and mean averaged expression of each transcript was used to identify genes which were most regulated during the LPS response. The final table can be viewed in Appendix 5.10.

This table showed that the majority of genes involved in intermediary metabolism were constitutively-expressed in macrophages and showed no evidence of immune cell enrichment compared to the wider sheep atlas (Clark et al., 2017). The subset of genes constitutively highly-expressed by macrophages and macrophage-enriched compared to other tissues may enable rapid metabolic responses to stimulation. They include genes that encode the glucose transmembrane transporters, namely SLC2A1 and SLC2A6, also known as GLUT-6 (Fukuzumi et al., 1996, Freemerman et al., 2014), and a novel leukocyte hexokinase 3 gene (HK3). HK1 may have another function as a PRR (Wolf et al., 2016) as well as having a critical role in the NLRP3 inflammasome (Moon et al., 2015). It is not known whether HK3 has a similar function.

Aside from gene transcripts encoding the glucose transporter(s), gene transcripts encoding most of the glycolytic enzymes (e.g. PKM, PFKL, PFKP, PGK1, PGM2/3, LDHA) including those of the pentose pathway (PGD, TADO1) were found to be highly-expressed by sheep BMDM (>>100TPM). The transcript encoding the rate-limiting enzyme of the pentose pathway, which enables the production of NADPH for oxidative killing by the phagocyte oxidase, glucose-6-phosphate dehydrogenase (G6PD gene, annotated as ENSOARG00000004237), was also highly-expressed by sheep BMDM, alongside genes encoding the components of the phagocyte oxidase (NCF1, CYBA, CYBB) (Iles & Forman, 2002, Kim et al., 2017). Constitutive high expression of genes involved in glucose metabolism in sheep BMDM may partially reflect the culture methods used in this project (Tavakoli et al., 2017).

A second subset of the genes in Appendix 5.10 with a primary annotation of metabolic process were highly regulated by LPS. Genes encoding glucose transporters, SLC2A1 (encoding GLUT1) and SLC2A6 (encoding GLUT6) were further induced by LPS, the latter >10-fold within 2 hours. The same pattern is observed in mice and humans (Schroder et al., 2012) and in all these species, SLC2A6 expression is largely restricted to stimulated macrophages. The glycolytic activator 6-phosphofructose-2-kinase and fructose-2,6-bisphosphatase encoded by PFKFB3 is proposed to upregulate glycolysis and link glucose metabolism to cell proliferation and survival in mouse macrophages (Jiang et al., 2016).
PFKFB3 was highly-expressed and induced further by LPS. The gene transcript encoding Pyruvate kinase, muscle (PKM aka PKM2) and both transcripts encoding lactate dehydrogenase A (LDHA), as well as SLC16A3 (which encodes the monocarboxyate carrier MCT4) and SLC16A6 were further elevated later in the response, after 24 hr. The TLR-inducible expression of SLC16A3 is shared with mice, and in that species mediates the export of lactate from glycolysis as part of a positive feedback mechanism (Tan et al., 2015). SLC16A3 may also mediate uptake of the ketone body, acetoacetate, in exchange for lactate, but has a relatively low affinity for beta-hydroxybutyrate and short chain fatty acids (Dimmer et al., 2000).

Although much of the flux of glucose through glycolysis gives rise to lactate, there was no evidence for the absolute loss of TCA cycle or mitochondrial gene expression in response to LPS. The set of transcripts that were partly down-regulated by LPS (see also Section 5.4.5 above), includes some encoding mitochondria-associated proteins, for example NDUF (NADH ubiquinone oxidoreductase) subunits FA10, FA2, FB4 and FB5, along with SDHC, but most genes associated with oxidative phosphorylation were unaffected.

Intermediates of the TCA cycle are known to provide precursors for other biosynthetic pathways. For example, TCA cycle intermediate metabolites function as checkpoints for multiple key LPS response genes, including IL1B (Jha et al., 2015, Papathanassiu et al., 2017) and IL1B expression has already been highlighted as particularly interesting in the sheep (see Section 5.3.7). The levels of citrate carrier (encoded by SLC25A1), which promotes the export of citrate from the mitochondria to the cytoplasm, can influence NO, ROS and PGE2 production in human macrophages (Infantino et al., 2014). SLC25A1 was induced in mouse macrophages in response to LPS, but in sheep BMDM, in common with pig, SLC25A1 was repressed.

In the TCA cycle, citrate is initially converted to cis-aconitate by mitochondrial aconitase 2 (encoded by the gene transcript ACO2). In stimulated mouse macrophages, the TCA cycle is diverted through the induction of a novel enzyme, cis-aconitate decarboxylase 1, encoded by immune responsive gene 1 (IRG1 also known as ACOD1). ACOD1 catalyses the conversion of cis-aconitate to cis-itaconate (Michelucci et al., 2013, Diskin & Pålsson-McDermott, 2018). Cis-itaconate is a novel anti-inflammatory metabolite that exerts feedback inhibition of the inflammatory response (Jha et al., 2015, Mills et al., 2018). ACO2 was robustly expressed across the sheep BMDM time course, but not specifically regulated
in response to LPS, similar to humans, mice and pigs. ACOD1 was rapidly (by 4 hr) induced in response to LPS in all individual sheep but the induction and level of expression was low compared to mice; only reaching similar levels to aconitase 2 (ACO2). In mice, the induction of ACOD1 leads indirectly to accumulation of downstream TCA cycle intermediates, succinate, fumarate and malate. Succinate may be an important metabolite in innate immune signalling which enhances IL1B production (Tannahill et al., 2013).

Interestingly, the G-protein-coupled receptor for succinate from the TCA cycle (encoded by SUCNRI also known as GPR91)(He et al., 2004) was highly upregulated in LPS stimulated sheep macrophages, whereas it is undetectable in mouse BMDM. In mice, Gpr91 is restricted to antigen-presenting dendritic cells (Rubic et al., 2008). In humans, it has been suggested that GPR91 is involved in renovascular hypertension, closely-related to atherosclerosis, diabetes and renal failure (He et al., 2004). This fundamental difference between sheep and mice/humans may relate to production of succinate by rumen microorganisms (Kennedy et al., 1991).

In other species, activated macrophages also use glutamine through anaplerosis to replenish TCA cycle intermediates (Costa Rosa et al., 1991, Wallace & Keast, 1992, de Oliveira et al., 2016, He et al., 2016). Glutamine can also act as a nitrogen donor for nucleotide (purine and pyrimidine), and protein synthesis, and can be converted to glutamate, which in turn is widely utilised in multiple metabolic pathways (Young & Ajami, 2001). By contrast to mouse and human macrophages (where expression was low, and not regulated), in sheep BMDM SLC1A5 (also called ASCT2), encoding the major glutamine transporter (Nicklin et al., 2009, Liu et al., 2015) was highly expressed and further induced by LPS. Knocking down this transporter in human PBMCs has been shown to interfere with IL1B expression (Tannahill et al., 2013). Genes encoding two other transporters that were LPS-inducible in sheep BMDM, SLC7A5 and SLC7A11, along with SLC1A5, have all been implicated in the so-called glutamine addiction of human cancer cells (Alexander et al., 2017). Only SLC7A11 appears to be similarly regulated by LPS in mice. This suggests that future studies could investigate the hypothesis that these transporters are likely to be the rate-limiting steps for glutamine metabolism within sheep BMDM. Both transcript variants for glutaminase (GLS gene), which catalyses the hydrolysis of glutamine to glutamate and ammonia as well as downstream genes in the glutamine pathway such as GOT1/2, GLUD1/2, MDH and GLUL were constitutively-expressed and not further regulated by LPS.
One notable feature of the sheep BMDM was the very highly expressed gene encoding L-asparaginase (*ASRGL1*), which was further up-regulated later in the LPS response. This gene in humans is expressed specifically in brain and testis and has been suggested as being involved in the production of L-aspartate (an excitatory neurotransmitter in some brain regions) (www.genecards.org), (Baslow, 2000). Asparaginase may also possess glutaminase activity (Chan *et al.*, 2014) and asparagine is likely taken up by *SLC1A5/ASCT2*. Weiske reported in 1879 that asparagine supports weight and a positive nitrogen balance in sheep (Ruckebusch & Thivend, 2012). Both asparagine and glutamine serve as nitrogen sources for the growth of rumen microbes, asparaginase mainly being associated with the bacteria, and glutaminase with the protozoa (Hoshino *et al.*, 1966). In humans, asparagine is a nontoxic carrier of residual ammonia to be eliminated from the body, and regulates the uptake of other amino acids, serine, arginine and histidine metabolism and thus protein and nucleotide synthesis (Krall *et al.*, 2016).

The production of nitric oxide also depends upon the uptake and intracellular availability of the NOS2 substrate, arginine. The data generated in this project contributed to a comparative study, involving multiple colleagues in the laboratory, of the regulation of the metabolism of arginine by LPS-stimulated BMDM in multiple species: sheep, goat, cattle, horse, water buffalo, human, pig and rat (Young *et al.*, 2018). The summarised data is shown in Table 5.5, with Figure 5.12. In addition to differences in NOS2 induction, macrophages from sheep and other large animals expressed distinct sets of genes encoding amino acid transporters compared to mice, in particular lacking expression of transcripts encoding the LPS-inducible cationic amino acid transporters, *SLC7A1* and *SLC7A2* (Yeramian *et al.*, 2006, Wanasen *et al.*, 2007, Thompson *et al.*, 2008, Chaturvedi *et al.*, 2010). *SLC7A2* is particularly important in the gastrointestinal tract of mice where it appears to regulate the attachment of intestinal pathogens and downstream signalling, recruitment of myeloid cells and the adaptive immune response (Singh *et al.*, 2016). Sheep express *SLC3A2/SLC7A7* constitutively, suggesting these may be the likely transporters for arginine. A protein which is also able to transport arginine, lysine and ornithine, encoded by *SLC7A3* (assigned to ENSOART00000000566 / ENSOARG00000000528), was also induced in the sheep LPS response.

As shown in Fig 5.12, arginine may also be metabolised by alternative pathways. The balance between arginase, which promotes ornithine and urea production and therefore
pro-proliferation, repair and growth of parasitic and bacterial infections, and nitric oxide synthase (NOS) enzyme, which generates nitric oxide (NO) species required for intracellular killing of pathogens, has been shown in mice to dictate the course of any response (Bronte & Zanovello, 2005). There are at least two isoforms of mammalian arginase (ARG1 and ARG2). The induction of arginase 1, encoded by ARG1 has been associated with alternative activation of mouse macrophages (Martinez et al., 2013). In mice, *Mycobacterium tuberculosis* (Mtb) evades the innate immune response by manipulating macrophage production of antimycobacterial NO (Kaufmann et al., 2005). Mtb may block the recruitment of NOS to the phagosomal membrane (Davis et al., 2007), as well as increasing expression of ARG1 which creates competition for NOS for arginine, reducing NO production (Qualls et al., 2010). In humans, ARG1 is produced in Mtb granuloma-associated macrophages and type II pneumocytes, not lymphocytes and ARG2 is not upregulated in human Mtb infected lungs (Pessanha et al., 2012). The induction of ARG2 may be upregulated in macrophages infected by intracellular pathogens such as *Trypanosoma cruzi*, *Trypanosoma brucei* and *Helicobacter pylori* (Das et al., 2010).

Table 5.5 summarises the data on expression of transcripts associated with arginine metabolism in multiple species. In common with other ruminants, sheep showed preferential inducible expression of ARG2, which encodes a mitochondrial arginase enzyme, which catalyses the hydrolysis of arginine to ornithine and urea, as part of the urea cycle, with no detectable increase in expression of ARG1 during the macrophage response to LPS, and was contrary to the horse and pig where ARG2 was downregulated in the response. These differences could explain why sheep are more susceptible to Toxoplasma gondii than cattle (Esteban-Redondo et al., 1999). Sheep macrophages also expressed high levels of *ornithine amino transferase (OAT)* transcripts, encoding the downstream enzyme which leads to the production of glutamate. Sheep and goats (to a lesser extent) were the only species that increased expression of *ODC1*, encoding ornithine decarboxylase which gives rise to polyamines. In mice, ODC1 has also been characterised in the anti-inflammatory macrophage response to parasite and fungal infections (Seabra et al., 2004, Heby et al., 2007). The unique commitment of sheep macrophages to these pathways further highlights the importance of recycling of urea to rumen microorganisms in ruminants.
In the cluster analysis of the LPS time course (Section 5.3.7, peaking at 4hrs in response to LPS), ARG2 and ODC1 were co-expressed with UPP1, the gene for uridine phosphorylase which produces uracil and ribose- or deoxyribose-1-phosphate, used for energy, carbon and to rescue pyrimidine bases for nucleotide synthesis (www.genecards). UPP1 is already known to have associations to inflammatory cytokines and the immune response (Watanabe & Uchida, 1995, De Santa et al., 2009).
Table 5.5 Differential expression patterns of enzymes associated with arginine metabolism and the production of NO during the macrophage response to LPS.

The table is simplified from expression data gathered by Young et al, with all domestic animal data generated by RNA-Seq and human data was generated by CAGE seq, described within the paper (Young et al., 2018). Where genes had a TPM <1 for both the unstimulated (0hr) and LPS stimulated (7hr) they are displayed as 0. Where the differential expression between samples was <5TPM, it has been deemed as ‘no change’ and displayed as nc. Nitric oxide (NO) production was assessed using a Griess assay, for which I contributed the sheep data. In response to LPS, Sheep and Horse macrophages do not produce NO, goat and buffalo macrophages produce small amounts of NO and both cattle and chicken (not shown) macrophages produce similarly large amounts of NO, as indicated by the approximate maximum nitrite concentration (µM) taken from the paper (Young et al., 2018). Pig, rat and human gries assay results were not available (n/a), however it is recognised that pig and human macrophages do not produce NO, and rat macrophages, like mice produce large amounts of NO in response to LPS (Kapetanovic et al., 2012).

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Sheep</th>
<th>Goat</th>
<th>Cattle</th>
<th>Buffalo</th>
<th>Horse</th>
<th>Pig</th>
<th>Rat</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG1</td>
<td>0</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
<td>0</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
</tr>
<tr>
<td>ARG2</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
</tr>
<tr>
<td>ASL</td>
<td>↓</td>
<td>nc</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>ASS1</td>
<td>↑</td>
<td>↓</td>
<td>nc</td>
<td>↑</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>GCH1</td>
<td>↑</td>
<td>nc</td>
<td>↑</td>
<td>nc</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>NO2</td>
<td>↑</td>
<td>↓</td>
<td>nc</td>
<td>↑</td>
<td>↑</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>OAT</td>
<td>↓</td>
<td>nc</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>ODC1</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>PTS</td>
<td>nc</td>
<td>nc</td>
<td>nc</td>
<td>nc</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>SLC3A2</td>
<td>↑</td>
<td>nc</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>SLC7A1</td>
<td>nc</td>
<td>↑</td>
<td>nc</td>
<td>↑</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>SLC7A2</td>
<td>0</td>
<td>↑</td>
<td>0</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>SLC7A5</td>
<td>↑</td>
<td>nc</td>
<td>↑</td>
<td>nc</td>
<td>nc</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>SLC7A7</td>
<td>↑</td>
<td>nc</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>SPR</td>
<td>↓</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>nc</td>
<td>0</td>
<td>nc</td>
</tr>
<tr>
<td>NO production</td>
<td>0.1</td>
<td>6</td>
<td>17</td>
<td>6</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

In humans and pigs, which also do not make NO, an alternative effector pathway, involving the metabolism of tryptophan by the inducible enzyme indoleamine dioxygenase (IDO1) is strongly induced by LPS (Kapetanovic et al., 2012). IDO1 has been implicated in host defence against mycobacteria, with increased IDO levels in affected tissues corresponding to progression to clinical mycobacterial disease (Plain et al., 2011); and chlamydia, where
intracellular pools of tryptophan are degraded by IDO in response to IFN-gamma and arrest growth of the organism (Entrican et al., 2009). Surprisingly, there was no detectable induction of IDO1 mRNA in the LPS-stimulated sheep BMDM, whereas the transcript was detected in other immune tissues in the atlas, and downstream genes in the tryptophan pathway, KYNU and KMO were robustly expressed and further induced by LPS. Since IDO1 is interferon-responsive in other species (Desvignes & Ernst, 2009, Kane et al., 2016), it may be that induction in sheep macrophage requires priming by IFN-gamma.

Fatty acid oxidation is highly efficient at generating ATP (O’Neill et al., 2016). Sheep macrophages do appear to be well-adapted to use free fatty acids as fuels, perhaps reflecting their availability in the ruminant circulation. Multiple genes encoding lipases (LPL, LIPA) and cytoplasmic fatty acid binding proteins (FABP4, FABP5) were expressed at exceptionally high levels (>1,000 TPM). Most enzymes required for fatty acid oxidation were constitutively expressed, and not regulated by LPS. The gene encoding the key transcriptional regulator of peroxisomal beta-oxidation pathway of fatty acids, peroxisome proliferator activated receptor delta (PPARD), was also constitutively-expressed, but further induced by LPS. PPARD is a dietary lipid sensor which has a preference for poly-unsaturated fatty acids, but depending upon the context, activation induces both anti-inflammatory and specific stimulatory molecules, including suppressing release of kynurenine (Adhikary et al., 2015). The gene encoding the receptor for medium chain free fatty acids (C9-C14), GPR84, was also highly induced in response to LPS in sheep, as it is in mice, pigs and humans. As noted with reference to arginine metabolism, ruminant and other large animal macrophages expressed distinct sets of amino acid transporters compared to mice. The increased expression of nucleoside transporter gene SLC28A3, in response to LPS, also appears to be specific to sheep, with no detectable expression in mouse macrophages and a slight decrease in expression in response to LPS in human macrophages.

5.5 Discussion

Genes associated with immunity are under strong evolutionary selection (as discussed in Section 1.1). The hypothesis underlying this project was that sheep macrophages may have evolved novel mechanisms and patterns of gene expression in response to an immune challenge such as LPS. As seen in other animals, LPS initiates a cascade of gene expression following TLR4 ligation. As expected, the transcriptome of BMDM in the sheep is
remarkably diverse with 85% of the reference transcriptome detected above a conservative threshold over the response to LPS (see Table 5.1). The transcriptome of stimulated macrophages is similarly rich in the mouse (Wells et al., 2003) and accordingly, macrophage RNA-Seq data makes a major contribution to genome annotation. Based upon their regulated expression, hundreds of genes identifiable only by Ensembl ID were manually assigned a functional annotation, including genes for transcription factors such as EGR1, EGR4, KLF2; cell surface receptors such as CD36, PTGIR; signalling molecules (PYCARD), transporters such as SLC25A30 and numerous chemokines, cytokines and enzymes.

Four different network approaches were used to explore the BMDM response to LPS over a 24 hr time course. Firstly, an analysis of the macrophage specific clusters within a network of all samples described in The Sheep Atlas (Clark et al., 2017) using averaged values, provided a comparison of the stimulated sheep macrophage signature in relation to the wide range of other cell types and tissues available within the atlas (Section 5.4.2). A small number of genes were identified that were found to vary in expression pattern compared to other species (Table 5.3).

Secondly, also using averaged values, all immune cell samples were clustered (Section 4.3.2, Appendix 4.1) and the BMDM response specific clusters were identified, in order to extract the set of transcripts that were macrophage and LPS response-specific relative to other immune cells. Consistent with the core hypothesis, a number of genes were uniquely expressed and regulated in sheep macrophages (Table 5.4).

Thirdly, the majority of the analysis focussed on expression values for individual sheep generated by the Kallisto pipeline. Evolutionary selection acts upon individual variation within any given population, and one would also anticipate that individual sheep differ from each other, especially between breeds subjected to divergent selection. In humans, the large majority of genes expressed in human macrophages in response to LPS exhibited heritable variation in the level of expression (Fairfax et al., 2014), and such variation has been associated with susceptibility to inflammatory bowel disease (Baillie et al., 2017). In pigs, variation between individuals was also observed, whereas there was no consistent impact of breed (Kapetanovic et al., 2013). This study compared the response of 6 cross-bred sheep. There was substantial variation between individuals that reinforced the identification of sets of co-regulated genes. Female 1 appeared to give a rapid early response and regain near-basal levels of expression within 24 hrs, whereas male 3 failed to
switch off an abundance of interferon responsive genes including those encoding the inducible transcription factors, \textit{IRF1}, \textit{IRF5} or \textit{IRF8}.

The Kallisto results were validated by examining a network analysis created using expression values generated by the StringTie pipeline for each individual. Numerous novel splice variants and potential novel long non-coding RNA transcripts were revealed.

Finally, the expression and regulation of metabolism-associated transcripts in sheep BMDMs was explored in detail. This analysis highlighted those that are highly regulated and/or highly expressed during the BMDM response in sheep. The high expression of \textit{SLC1A5}, appeared to be novel and specific to sheep. The high expression profile of \textit{ARG2} is shared with other ruminants, and some gene profiles are only shared with closely related species, namely the goat, such as \textit{ODC1}. The regulated expression of \textit{ARG2} in sheep (shared with other ruminants and horses) results from evolutionary divergence of promotor sequences compared to other large and small animal species, notably the insertion of a retrotransposon (BOV-A2) around -3 kb in sheep and goats (Young \textit{et al.}, 2018). Altering the promotor architecture can result in the gain or loss of enhancers (Kapetanovic \textit{et al.}, 2012, Schroder \textit{et al.}, 2012). Similarly, a BOV-A2 insertion distinguishes the NOS2 promoters of cattle and water buffalo from sheep and goat, and may explain the differential LPS inducibility, as it contains binding sites for many macrophage specific (PU.1, CEBPB) and inducible (STAT1, IRF1, NFKB) transcription factors (Young \textit{et al.}, 2018). As discussed in Section 5.4.7, sheep macrophages did not produce detectable nitric oxide in response to LPS, presumably a consequence of high \textit{ARG2} as well as low \textit{NOS2} induction. In stimulated mouse macrophages the production of nitric oxide suppresses respiration through the inhibitory effect on haem-containing components of the electron transport chain (Van den Bossche \textit{et al.}, 2016). Since sheep macrophages do not produce nitric oxide, and also have relatively low expression of \textit{ACOD2}, they are unlikely to exhibit the decreased respiration and “broken” TCA cycle reported in activated mouse macrophages (Mills & O'Neill, 2016).

Ruminant and/or sheep/goat-specific regulation could reflect the evolution of the rumen in providing optimal conditions for microbial fermentation and the wide variety of microbes, many of which cannot be cultured in a lab (Krause & Russell, 1996). Future laboratory work is needed to determine whether the retention of a fully functional TCA cycle allows sheep
macrophages to utilise fatty acids that are a major circulating product of ruminant digestion as fuels.

The seven-transmembrane/ G-protein coupled receptors are a group of receptors, involving G-protein coupling and activation of second messenger generating enzymes, and are known to have remarkably diverse biological functions, with members evolving rapidly (Pierce et al., 2002, Kwakkenbos et al., 2004, Hume, 2006), so it was anticipated that the sheep may have different functional members and possibly novel members for themselves. Free fatty acids receptor (FFARs) act as physiological sensors of food-derived FFAs and digestion products in GIT, thus they are important for nutrition and act as signalling molecules in numerous physiological processes (Hara et al., 2013).

In ruminants, a wide variety of circulating FFA is indicative of health whereas in non-ruminants the same metabolites would be pro-atherogenic (Netea et al., 2016, Groh et al., 2017), being recognised as DAMPs by the innate immune system. In mice at least, FFAR-2 has been implicated in communication between the intestinal microbiome and the microglia of the brain (Erny et al., 2015). Induction of the gene encoding a FFA receptor, GPR84, in sheep macrophages was shared with other species, but in sheep it may have access to a much higher concentration of ligand in the circulation. The response of sheep macrophages to added FFA is clearly worthy of future study.

One novel GPCR expressed in sheep macrophages, and highly-regulated by LPS, is adenylate cyclase activating polypeptide 1 (Pituitary) Receptor (ADCYAP1R1 gene). This gene was not expressed or regulated in human, mouse and pig macrophages. ADCYAP1R1 shares significant homology with members of the glucagon/secretin receptor family and the receptor is known to regulate the release of adrenocorticotropin, luteinizing hormone, growth hormone, prolactin, epinephrine, and catecholamine (Ogi et al., 1993, Yon et al., 1998, Vaudry et al., 2000). The gene is predominantly expressed in the central nervous system in humans, mice and pigs. In the sheep atlas, the LPS-stimulated level of expression is comparable to that of the sheep brain and adrenal. In other species, the ligand, ADCYAP1, appears to be brain-restricted, but the ligand gene is not currently annotated in Ensembl, so it is not possible to assess whether it might have distinct functions in sheep.

Another macrophage-specific and LPS-regulated transcript of wider interest to macrophage biologists is ADGRE1, which in mouse encodes the macrophage-specific F4/80 antigen (Hume et al., 1984). The sheep RNA-Seq data generated herein was used in a comparative
analysis of gene expression data from multiple species (Waddell et al., 2018). That analysis revealed significant divergence in both basal expression and induction by LPS amongst species (Appendix 5.11, taken from manuscript (Waddell et al., 2018)). The mouse Adgre1 gene encodes a GPCR with an extracellular region containing 7 EGF-like calcium-binding domains, which are subject to extensive alternative splicing. The sheep macrophage RNA-Seq data revealed that the sheep ADGRE1 locus contains a complete internal duplication of the extracellular domain, encoding 14 EGF-like domains, with each subject to alternative splicing/skipping. Figure 5.13 shows the Sashimi plots for the sheep RNA-Seq data, derived from the manuscript, compared with a number of other species that appear to share the duplicated extracellular domain (Waddell et al., 2018). The extensive exon skipping is most evident in the sheep profile because of the much greater depth of sequencing. The manuscript discusses the possibility that ADGRE1 encodes a pattern recognition receptor.

The analysis of macrophage-specific and LPS-inducible regulation in this chapter has not considered the long non-coding RNAs (lncRNAs). These are a significant component of the LPS-inducible transcriptome in both mouse and human macrophages; some of them associated with transcripts derived from regulated enhancers (Melamed et al., 2016, Baillie et al., 2017). Because of their low expression, lncRNA are difficult to quantitate and to assemble reliably from short-read RNA-Seq data and most fell below expression thresholds used here. The macrophage data, alongside the wide sheep atlas and a parallel study of goat and published cattle and human data was used in cross-species inference of lncRNA in ruminants (Bush et al., 2018). The consensus lncRNA set will permit future analysis of this component of the transcriptome in sheep macrophages.

It is beyond the scope of this thesis to discuss the function of every ruminant or sheep-specific transcript. Examples were summarised in Tables 5.3 and 5.4, and others, such the gene duplication of SERPINB2, were discussed briefly in context as they arose in the analysis. In overview, the core hypothesis, that the core sheep macrophage transcriptome, and the response to LPS, would exhibit species-specific features, was clearly strongly supported by the analysis of the transcriptome.
Fig 5.13 Sashimi plots for *ADGRE1* across multiple species (taken from submitted manuscript)

The large amount of splicing results in an abundant number of different isoforms, as every exon encoding an EGF-like domain can be skipped by alternative splicing in all species.
Chapter 6  Summary of achievements and future directions

6.1 General discussion

This thesis aimed to identify and annotate genes that are associated with the innate immune functions of macrophages in sheep. The analysis was based upon the development of a cell culture system to generate and characterise large numbers of macrophages from sheep bone marrow (BMDM) (Sections 2.3.1 and 2.4). The BMDM responded to the TLR4 agonist, LPS, as evidenced by release of the proinflammatory cytokine, TNF (Section 5.4.2 and 5.3.6). Using this system, bone marrow from six healthy male (n=3) and female (n=3) adult sheep (BF X T) was used to generate BMDM and to produce RNA-Seq (both polyA-selected (mRNA) and ribodepleted (Total RNA)) data describing the response to LPS across a time course (Chapter 5). The RNA-Seq data were processed using two independent pipelines (StringTie and Kallisto), to determine genome wide transcript expression estimates for each sample. The data generated in this project made a significant contribution to the sheep transcriptional atlas (Section 5.3.2 (Clark et al., 2017)). With nearly 90% of all nuclear encoded protein coding genes being detected in at least one BMDM sample at a given time point, the sheep BMDM response to LPS data provided the expected (Ravasi et al., 2007) diverse and rich source of transcripts for study, and the data created (made available to Ensembl), will greatly improve future versions and functional annotation of the reference transcriptome.

The availability of the large sheep atlas dataset enabled a comparative analysis to identify transcripts that were enriched in macrophages and/or specifically associated with activation (5.3.2 and 5.3.3). Clustering transcript expression estimates using Miru, as other projects have done (Freeman et al., 2012, Mabbott et al., 2013, Raza et al., 2014, Clark et al., 2017), segregated the massive dataset into clusters containing transcripts that share similar patterns of expression. Based upon the principle of guilt-by-association (Oliver, 2000) the products of these transcripts are likely to share functions. Co-expression in macrophages with genes of known function provided the basis for prioritisation of genes for detailed annotation. Hundreds of genes possessing only an Ensembl ID were manually assigned a functional annotation based upon nearest orthologues and shared synteny with mouse or human (Section 2.11.1 for methods, Appendix 6.1 contains the full list of Ensembl IDs that have been annotated).
Publicly available datasets were used to compare the sheep macrophage gene expression profiles with those generated for monogastric animals; human (Baillie et al., 2017), mouse (Schroder et al., 2012), pig (Kapetanovic et al., 2013) and cow (personal communications with Dr Rachel Young). Many of these differences between sheep and other species may be attributed in part to evolution of their relationship with the fermentative microbiome and increased levels of circulating fatty acids. The current view of the association of immune and metabolic genes with innate immunity is dominated by studies of mice and human cells. Evolution has provided the ruminant animal macrophage with a unique ability to function in a microenvironment that by human standards would be associated with the production of foam cells and atherosclerosis. Indeed, many of the hallmarks of foam cells were found in the resting sheep macrophage transcriptome signature. The predicted differences in intermediary metabolism between sheep and other species, based upon RNA-Seq data, now need to be confirmed using proteomic and metabolomic analysis of sheep macrophages (starting with TCA cycle intermediate investigations) using methods that have been applied previously only to the mouse (discussed in Section 5.4.2).

6.2 Genome annotation and the new genome assembly

Two pipelines were used to quantitate gene expression based upon mapping to the current reference genome OarV3 (Jiang et al., 2014). A preliminary examination of LTR repeat regions, revealed variation between individual sheep in differential expression profiles between 0hr and 7hr, with BTLTR1 (a known ruminant specific LTR-containing repeat region (Elsik et al., 2009). BTLTR1 was found to be the most regulated repeat region, decreasing in response to LPS (Section 3.6). OarV3.1 still contains many scaffolds, which will have affected the accuracy in mapping reads to repeat regions, but with the imminent release of the new genome assembly, which contains long read technology, many of these assembly issues will be resolved. Immediate future plans involve repeating the mapping of all RNA-Seq data from this project to the improved reference genome, Rambouillet v0.1 (https://www.ncbi.nlm.nih.gov/genome/83?genome_assembly_id=351950), which is a new highly contiguous PacBio assembly, incorporating many of the scaffolds that exist in OarV3.1. The accuracy of determining expression over these important areas of the genome will be greatly improved and future plans involve repeating this examination for all the classes of repeat regions, especially given the global interest surrounding Jaagsiekte sheep retrovirus (JSRV), which causes a pulmonary adenocarcinoma in sheep and goats (Rai et al., 2001, Hofacre & Fan, 2010, Martineau et al., 2011).
There is still RNA available from many of the samples that were submitted to the sheep atlas, including the BMDM time courses. Given the decreasing cost of sequencing, it may be possible to submit the samples for the PacBio Iso-Seq protocol. A recent report described a rabbit transcriptional atlas based upon long read sequencing (Chen et al., 2017). Full length RNA sequences will improve assessment of the full diversity of alternative splicing of mature transcripts, especially for complex, multi exon genes of interest expressed in macrophages such as ADGRE1, ADCYAP1R1 and ITGAM (Sections 5.5, Figure 5.13). Continued exploration of the novel and un-annotated transcripts identified in this project as being important in the sheep macrophage response will reveal many more novel transcripts to complement those identified herein. Long read RNA-Sequencing will also improve the identification of long non-coding RNAs. The sheep atlas, including the large macrophage dataset generated in this project, contributed to identification of many novel candidate IncRNA (Bush et al., 2018), but their assembly from short-read data is challenging.

6.3 The diversity of macrophage responses
This study has examined the response of sheep BMDMs to a selective TLR4 agonist, LPS from Salmonella enterica serotype Minnesota Re 595. There remain many other forms of LPS, other TLR agonists and intracellular stimuli including specific pathogen challenge, to fully appreciate the response of sheep macrophages and individual variation in that response, as well as assessment of different breeds of sheep in their response. The response to mycobacteria would be especially interesting. Recent reports have contrasted the transcriptomic response in bovine alveolar macrophages and both the transcriptomic and proteomic response to M. tuberculosis and M. bovis (Malone et al., 2018) and of monocyte-derived macrophages to M. avium subsp paratuberculosis (Marino et al., 2017). Sheep are comparatively resistant to both of these infectious agents. Responses to other stimuli also differ between species. For example, in the response to alternative activation by IL-4, very few inducible genes were shared between mice and humans (Martinez et al., 2013). Mice and humans vary in their response to glucocorticoids owing to enhancer divergence and the rapid evolution in the response between the species (Jubb et al., 2016).

Aside from mycobacteria, many other pathogens are of particular concern to sheep health, welfare and productivity. Some are zoonotic and pose significant threat to humans or other production animals. Ongoing comparative analysis will help to identify host
determinants that underlie the differential pathology. To this end, the BMDM culture system employed in this project has been applied in the Hume laboratory to the cow, goat, rat, water buffalo (personal communication, Professor David Hume, University of Edinburgh), pig (Kapetanovic et al., 2012), human (Baillie et al., 2017), and mouse (Schroder et al., 2012). Some of the differences between these species were analysed in Section 5.4.2 and were published in analysis of genes involved in arginine metabolism (Young et al., 2018). A second study in the same set of species compares expression and evolution of ADGRE1 (Section 5.5) (Waddell et al. submitted). For each of the genes identified as demonstrating a distinctive expression pattern to other animals, the obvious question is whether there is a specific function in host-specific pathogen recognition. For example, based upon comparative analysis (Section 5.4.2) nitric oxide production is unlikely to be a major component of anti-microbial defences in sheep and goats, or in horses, pigs and humans, where it is clearly important in rodents and possibly in large ruminants.

6.4 Individual variation in the response to LPS

An expression QTL analysis in humans revealed that >80% of LPS-inducible genes in monocytes exhibited heritable variation in the level of expression (Fairfax et al., 2014). The six individual sheep were found to vary markedly in the response of their BMDMs to LPS: Male 3 appeared hyper-responsive whereas Female 1 responded more rapidly but transiently. The differences appeared to focus around interferon target genes (Section 5.3.7). A previous larger comparative study in pigs showed similar variation amongst individuals but did not find any evidence of breed-specific differences (Kapetanovic et al., 2013). The cell culture system developed for this project, which utilises frozen bone marrow cells, can potentially be applied to much larger numbers of animals. It will be of great interest to determine whether the degree of variability of response observed in this study is fortuitous, or is representative of the population as a whole. Increasing the numbers of both male and female would also permit analysis of whether the sex of sheep influences the response, and which genes (if any) are sex specific in the response.

The animals used in this study were F1 crosses from two disparate breeds: the Texel and the Scottish Blackface. The high level of heterozygosity may actually have masked individual variation in the response but was intended to maximise the numbers of expressed SNPs. Future studies are required to utilise a much greater number of animals to determine if the degree of individual variation in response to LPS that has been noted in
this project, is present at the population level. One major future direction is to utilise the genomic sequence data available for each individual and examine the extent of allele specific expression across this dataset, as has been done in human (Knight, 2004, Ardlie K, 2015), mice (Crowley et al., 2015) and cows (Chamberlain et al., 2015). In outbred cattle, 89% of the genes tested had evidence of allelic imbalance in expression in at least one tissue, and as many as 28% of transcripts showed absolute monoallelic expression (Chamberlain et al., 2015). Having identified the genes which are expressed differentially between alleles, it should be possible to infer the parent of origin, as the current reference genome is based upon a Texel, the sire breed of all the animals used in this project. A future direction will be to use the same procedure to examine both pure-bred animals, F1 crosses with the male and female parents reversed to confirm any evidence of parental imprinting, and F2 generation to begin to assess heritability. The heritability of gene expression will undoubtedly be complex, involving not only Mendelian genetic factors but parental metabolic states (Palu et al., 2017).

Genes expressed in macrophages may also differ at the level of protein sequence. Current projects in Roslin and elsewhere are producing hundreds of genomic DNA sequences from sheep, goats, cattle, water buffalo and African buffalo, including those adapted to tropical environments with high disease burdens (Borriello et al., 2006, Wang et al., 2016, Banos et al., 2017, Bhuiyan et al., 2017, Wilkinson et al., 2017). The prediction a priori is that functional variation in macrophage-expressed genes between individuals, and between breeds adapted to different environments, and between species, are more likely to be causally associated with differences in disease susceptibility.

6.5 Conclusion

The work described in this thesis has made several contributions to the understanding of macrophage biology. It has resulted in a dataset now in the public domain that can continue to be mined for insights, as indicated throughout the presentation of the results, where examples of information that can be derived from the data are shown. It has also provided a comprehensive analysis of the transcriptomic response to a cellular state change, highlighting the role of metabolic adjustments as macrophages respond to pathogen challenge. Adding the sheep data to that for other ruminant and monogastric animals has allowed a comparison which highlights the adaptations of the ruminant to maintain the critical balance with the microbiota of the gut. It opens the way to
understanding of the genetic basis of individual variation in immune response, both in sheep and more generally in other animals including humans. This work therefore has the potential to impact on human and animal health as well as production and food sustainability, key goals for research in the 21st century.

Appendices

Appendices are in files contained in the CD on the inside back cover of this thesis.

Appendix 2.1 (Edinburgh Genomics sample submission)

Appendix 3.1 (Kallisto v StringTie alignment rates)

Appendix 3.2 (SAMtools flagstat analysis)

Appendix 4.0 (StringTie and Kallisto expression estimates)

Appendix 4.1 (Kallisto immune cell clusters)

Appendix 5.1 (Sheep Atlas LPS response clusters)

Appendix 5.2 (Peaking Kallisto analysis)

Appendix 5.3 (Peaking StringTie analysis)

Appendix 5.4 (Kallisto PolyA LPS timecourse)

Appendix 5.5 (Final Kallisto PolyA post filtering)

Appendix 5.6 (Corrected StringTie PolyA samples)

Appendix 5.7(Kallisto Miru files)

Appendix 5.8 (Final StringTie PolyA post filtering)

Appendix 5.9 (StringTie Miru files)
Appendix 5.10 (Metabolism genes of interest)

Appendix 5.11 (ADGRE1 and ADGRE4 species comparison, taken from manuscript)

References


Bittner JJ (1936) SOME POSSIBLE EFFECTS OF NURSING ON THE MAMMARY GLAND TUMOR INCIDENCE IN MICE. Science (New York, NY) 84: 162.


the Sheep Lung Microbiota. Applied and Environmental Microbiology 82: 3225-3238.
Godfroy JI, Roostan M, Moroz YS, Korendovych IV & Yin H (2012) Isolated Toll-like Receptor
Gordon S (2016) Elie Metchnikoff, the Man and the Myth. Journal of innate immunity 8:
223-227.
Gordon S & Martinez FO (2010) Alternative activation of macrophages: mechanism and
functions. Immunity 32: 593-604.
biology 15: 53.
Gordon S, Pluddemann A & Martinez Estrada F (2014) Macrophage heterogeneity in
 tissues: phenotypic diversity and functions. Immunological reviews 262: 36-55.
Goyette J, Yan WX, Yamen E, et al. (2009) Pleiotropic roles of S100A12 in coronary
atherosclerotic plaque formation and rupture. Journal of immunology (Baltimore, Md :
1950) 183: 593-603.
Grabmuller M, Madea B & Courts C (2015) Comparative evaluation of different extraction
and quantification methods for forensic RNA analysis. Forensic science international
biology and evolution 9: 1880-1885.
Grieves JL, Dick EJ, Jr., Schlabritz-Loutsevich NE, Butler SD, Leland MM, Price SE, Schmidt
Groh L, Keating ST, Joosten LAB, Netea MG & Riksen NP (2017) Monocyte and macrophage
immunometabolism in atherosclerosis. Seminars in immunopathology.
Gross L (1951) "Spontaneous" leukemia developing in C3H mice following inoculation in
infancy, with AK-leukemic extracts, or AK-embrovos. Proceedings of the Society for
Experimental Biology and Medicine Society for Experimental Biology and Medicine (New
York, NY) 76: 27-32.
Gross SR, Sin CG, Barracloough R & Rudland PS (2014) Joining S100 proteins and migration:
for better or for worse, in sickness and in health. Cellular and molecular life sciences : CMLS
71: 1551-1579.
Gruenheid S, Pinner E, Desjardins M & Gros P (1997) Natural resistance to infection with
intracellular pathogens: the Nramp1 protein is recruited to the membrane of the
Yona S (2014) Dendritic cells, monocytes and macrophages: a unified nomenclature based
on ontology. Nature reviews Immunology 14: 571-578.
Guilliams M, De Kleer I, Henri S, Post S, Vanhoutte L, De Prijck S, Deswarte K, Malissen B,
Hammad H & Lambrecht BN (2013) Alveolar macrophages develop from fetal monocytes
that differentiate into long-lived cells in the first week of life via GM-CSF. The Journal of
Hacker H, Tseng PH & Karin M (2011) Expanding TRAF function: TRAF3 as a tri-faced


Haziot A, Ferrero E, Lin XY, Stewart CL & Goyert SM (1995) CD14-deficient mice are exquisitely insensitive to the effects of LPS. *Progress in clinical and biological research* **392**: 349-351.


Hultmark D (1994) Macrophage differentiation marker MyD88 is a member of the Toll/IL-1 receptor family. Biochemical and biophysical research communications 199: 144-146.


protein transgene is expressed throughout the mononuclear phagocyte system of the mouse. *Blood* 101: 1155-1163.


