Elimination of reference mapping bias reveals robust immune related allele-specific expression in cross-bred sheep

Citation for published version:

Digital Object Identifier (DOI):
10.1101/619122

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Unbiased ovine allele-specific expression

Elimination of reference mapping bias reveals robust immune related allele-specific expression in cross-bred sheep

Mazdak Salavati 1*, Stephen J. Bush 1, Sergio Palma-Vera 2, Mary E.B. McCulloch 1, David A. Hume 3 and Emily L. Clark 1*

1. The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, UK
2. Leibniz Institute for Farm Animal Biology (FBN), Dummerstorf, Germany
3. Mater Research Institute-University of Queensland, Translational Research Institute, Woolloongabba, Qld 4102, Australia

* Correspondence:
Dr Emily L. Clark
Emily.Clark@roslin.ed.ac.uk
Dr Mazdak Salavati
Mazdak.Salavati@roslin.ed.ac.uk

Keywords: Allele-specific expression, mapping bias, RNA-Seq, sheep, transcriptome, WASP, GeneiASE

Abstract
Pervasive allelic variation at both gene and single nucleotide level (SNV) between individuals is commonly associated with complex traits in humans and animals. Allele-specific expression (ASE) analysis, using RNA-Seq, can provide a detailed annotation of allelic imbalance and infer the existence of cis-acting transcriptional regulation. However, variant detection in RNA-Seq data is compromised by biased mapping of reads to the reference DNA sequence. In this manuscript we describe an unbiased standardised computational pipeline for allele-specific expression analysis using RNA-Seq data, which we have adapted and developed using tools available under open licence. The analysis pipeline we present is designed to minimise reference bias while providing accurate profiling of allele-specific expression across tissues and cell types. Using this methodology, we were able to profile pervasive allelic imbalance across tissues and cell types, at both the gene and SNV level, in Texel x Scottish Blackface sheep, using the sheep gene expression atlas dataset. ASE profiles were pervasive in each sheep and across all tissue types investigated. However, ASE profiles shared across tissues were limited and instead they tended to be highly tissue-specific. These tissue-specific ASE profiles may underlie the expression of economically important traits and could be utilized as weighted SNVs, for example, to improve the accuracy of genomic selection in breeding programmes for sheep. An additional benefit of the pipeline is that it does not require parental genotypes and can therefore be applied to other RNA-Seq datasets for livestock, including those available on the Functional Annotation of Animal Genomes (FAANG) data portal. This study is the first global characterisation of moderate to extreme ASE in tissues and cell types from sheep. We have applied a robust methodology for ASE profiling, to provide both a novel analysis of the multi-dimensional sheep gene expression atlas dataset, and a foundation for identifying the regulatory and expressed elements of the genome that are driving complex traits in livestock.
Unbiased ovine allele-specific expression

Introduction

Allele-specific expression (ASE) is the imbalance of allelic expression between parental (diploid) copies at the same locus (Barlow and Bartolomei, 2014). It is most commonly associated with cis-acting regulatory variation that may mediate parent-of-origin, sex- or tissue-specific transcription of one allele relative to the other (Renfree et al., 2009; Hasin-Brumstein et al., 2014). In a single individual, where there are informative sequence variants (i.e. heterozygote loci) that distinguish the products of two alleles, ASE can be detected by RNA-sequencing (Chamberlain et al., 2015; GTEx Consortium et al., 2017; Cao et al., 2019; Guilloucheau et al., 2019). The ratio of allelic read counts obtained from RNA-Seq datasets can be used as a reliable proxy for ASE (i.e. $ASE_{\text{ratio}} = \frac{\text{Counts}_{\text{Allele1}}}{(\text{Counts}_{\text{Allele1}} + \text{Counts}_{\text{Allele2}})}$) (Edsgärd et al., 2016).

Large and complex RNA-Seq datasets give rise to unique and interesting computational challenges, in particular the elimination of reference mapping bias in ASE analysis of diploid genomes. RNA-Seq data are commonly mapped against reference genomes which are typically ‘flat’, with each position represented only by the reference (most abundant) allele. As such, reads containing heterozygous loci are more likely to be erroneously mapped (Degner et al., 2009; Stevenson et al., 2013; Hodgkinson et al., 2016). This can lead to high false positive ASE locus discovery rates (Degner et al., 2009). Although development of de novo transcript assemblers (Zerbino and Birney, 2008), variant-aware aligners (Xin et al., 2013; Hach et al., 2014) and mapping-free quantification e.g. Kallisto ((Bray et al., 2016)) has removed some of the issues, reference allele mapping bias remains an unresolved challenge in ASE studies. The analysis pipeline we present in the manuscript is designed to minimise reference bias while providing accurate profiling of allele-specific expression in large and complex RNA-Seq datasets.

We have developed an ASE analysis pipeline using the combination of software available under open licence, WASP (reference mapping bias removal) (van de Geijn et al., 2015), GATK (ASEReadCounter) (McKenna et al., 2010; Van der Auwera et al., 2013) and GeneiASE (Liptak-Stouffer aggregative ASE gene model) (Edsgärd et al., 2016). The GeneiASE model is capable of testing ASE at the gene level using two approaches: i) Static ASE which measures allelic imbalance within a gene (i.e. when ASE variants are located within the boundaries of the gene) and ii) Individual condition dependent ASE (ICD) which measures inducible ASE in a gene under an environmental pressure between 2 timepoints (i.e. in stimulated or unstimulated immune cells).

In addition to ASE at the gene level, we can also measure significant ASE at the single nucleotide level (SNV). ASE has been shown to be enriched within expression quantitative trait loci (eQTL) regions (Montgomery et al., 2010), therefore identifying ASE variants can be useful for understanding the transcriptomic control of complex traits in livestock. Complex trait mapping of ASE loci has been associated with phenotypes such as resistance to Marek’s disease in chicken (Meydan et al., 2011) and pigmentation patterns in sheep (García-Gámez et al., 2011).

Understanding ASE is also important because cross-breeding now underlies most livestock production systems. Knowledge of ASE may provide insights into the molecular basis of the complex phenomenon of hybrid vigour, as emphasised by recent studies on two Chinese goat breeds and their F1 hybrids (Cao et al., 2019) and in F1 crosses of two highly inbred chicken lines (Zhuo et al., 2017). In this study we measure ASE in cross-bred sheep. Sheep are an economically important livestock species in many countries across the globe and particularly in emerging economies. The identification of prevalent ASE in populations or breeds, especially in economically relevant phenotypes and tissues, could be used to improve genomic prediction in sheep breeding programmes, such as those that have been established in...
Unbiased ovine allele-specific expression

Australia and New Zealand (Daetwyler et al., 2010).

Using the methodology we describe, for mapping bias correction and robust positive ASE discovery, we were able to profile pervasive allelic imbalance across tissues and cell types, at both the gene and SNV level, in Texel x Scottish Blackface sheep. We analysed a subset of total RNA-Seq libraries from liver, spleen, ileum, thymus and bone marrow-derived macrophages (BMDM) (+/-) lipopolysaccharide (LPS), from six individual adult cross-bred sheep to produce a detailed picture of allelic imbalance in immune-related tissues and cell types. We chose to focus this analysis on immune related tissues in part because of the depth of available sequence in those tissues, and in part because they contain abundant immune cell populations. The diversity of cell populations is reflected in the transcriptional complexity of immune tissues and cell types in the sheep gene expression atlas dataset (Clark et al., 2017; Bush et al., 2018). As such this subset of tissues gave us a transcriptionally rich dataset in which to measure ASE. We also included BMDMs stimulated and unstimulated with LPS to mimic infection with Gram-negative bacteria, in order to test whether ASE changed in response to stimulation with LPS in these cells. By measuring ASE in these tissues and cell types from sheep we were able to: i) provide insight into how pervasive ASE is across tissues at the gene and SNV level, ii) generate tissue-specific ASE profiles, iii) investigate sex-specific patterns of ASE, and iv) determine the extent to which ASE changes in response to stimulation with LPS in an immune cell type. This novel analysis of the multi-dimensional sheep gene expression atlas dataset provides a foundation for further analysis of the regulatory and expressed elements of the genome that are driving complex traits in sheep.

Methods

Sample preparation and RNA extraction

Data from three male and three female Texel x Scottish Blackface (TxBF) sheep from the sheep gene expression atlas project (Clark et al., 2017) were used in this study. The dataset including: 2 cell types (BMDMs (+/-) LPS) and 4 tissues (thymus, spleen, liver and ileum). Tissue collection, storage and RNA extraction is described in (Clark et al., 2017). BMDM’s were cultured in vitro for 7 days in the presence of macrophage colony-stimulating factor (CSF1 (10^4 U/ml) and unstimulated (0 hrs - LPS) and stimulated (7hrs +100ng/ml LPS) samples of BMDMs were obtained also as described in (Clark et al., 2017). A total of 2 samples (1 thymus and 1 spleen) did not pass the RNA quality control (RNA integrity number (Mueller et al., 2004); RIN <7) and were not included in the sheep gene expression atlas. Library preparation was performed by Edinburgh Genomics (Edinburgh Genomics, Edinburgh, UK). All total RNA Illumina TruSeq libraries (125bp paired end) were sequenced at a depth of > 100 million reads per sample.

Reference mapping bias removal

BAM files from RNA-Seq data, were previously mapped to Oar v3.1 top level DNA fasta track, using HISAT2 as described in (Clark et al., 2017). These BAM files were used to locate reads with heterozygote loci using WASP’s find_intersecting.py script (van de Geijn et al., 2015). The intersection of reads and heterozygote loci in all samples were based on the Ensembl v92 variant call format (VCF) track (Ensembl v92: ovis_aries_incl_consequences.vcf.gz). Synthetic copies of reads containing heterozygote loci, were created with the alternate allele flipped to the remaining options of A, T, C or G (up to 6 loci /read[4^6] max 64 combinations of synthetic reads) using parameters defined in WASP (van de Geijn et al., 2015). This was followed by remapping of the synthetic reads using HISAT2
(Li and Durbin, 2009; Kim et al., 2015) and eliminating the original reads (and their synthetic copies) which mapped to a different coordinate in any of its synthetic copies (WASP’s filter_reads.py) (van de Geijn et al., 2015). After merging the retained reads with non-intersecting reads, a final BAM file was produced for ASE read counting step (WASP’s remove_dup.py).

### Allelic read counts and depth filtration

Allele-specific read counting was carried out using the ASEReadCounter module of GATK v3.8 with parameters -mmq 50 -mbq 25 (McKenna et al., 2010). Multiple pre-processing steps were performed prior to GeneiASE input as instructed by (Edsgärd et al., 2016) which included, preparing per chromosome indices, merging the variant set with corresponding gene coordinates and bi-allelic expression filtering. Loci with < 10 reads mapped were excluded, as were loci with < 3 reads, or < 1% of the total reads, mapped to both the reference and alternative allele. This form of filtration will eliminate loci exhibiting mono-allelic expression (MAE) as previously described (Degner et al., 2009; Stevenson et al., 2013; Mayba et al., 2014). Similar filtration criteria have been previously used routinely in ASE studies (Mayba et al., 2014; Chen et al., 2016a; Edsgärd et al., 2016; GTEx Consortium et al., 2017; Raghupathy et al., 2018; Cao et al., 2019; Guillocheau et al., 2019; Gutierrez-Arcelus et al., 2019). The workflow of the analysis pipeline for ASE analysis is detailed in Fig.1.

### Experimental design for defining allele-specific expression

ASE was defined according to the following three categories:

i) Static ASE: which is inherent allelic imbalance (AI) in each gene calculated by

\[
ASE = \frac{Counts\ Ref\ Allele}{Counts\ Ref\ Allele + Counts\ Alt\ Allele}
\]

within the boundaries of the gene. The effect size of ASE at gene level was produced by aggregation of the ASE effect size at SNVs within the gene boundaries (the Liptak-Stouffer method, applied by the GeneiASE aggregative model). A null distribution of ASE effect size for genes in each transcriptome was produced by random sub-sampling (n = 1e+5) from a pool of genes having min 2 and max 100 loci within their boundaries. The ASE effect size of each gene (aggregated using Liptak-Stouffer) was then tested against the null distribution of same SNV number, via a modified bi-nomial test (2x1 table). Distribution of p values was examined for uniformity prior to false discovery rate correction (Supplementary file1) (Benjamini and Hochberg, 1995; Pounds and Cheng, 2006; Barton et al., 2013).

ii) Individual condition-dependent ASE (ICD-ASE): in which the same ASE effect size was calculated for each gene in the treated versus the untreated timepoints of the same sample (i.e. BMDM +/- LPS). The log2ratio (\(ASE_{treated}/ASE_{untreated}\)) was used in a beta-binomial test (2x2 table) similar to the static mode. The details of this aggregative model have been previously described in the GeneiASE publication (Edsgärd et al., 2016).

iii) Condition dependent ASE at SNV level: in which a contingency table was produced for read counts (ref and alt) for every SNV, present both in treated and untreated conditions (BMDM +/- LPS) (2x2 table) and a Fisher’s exact test performed followed by p value multiple testing correction (Benjamini and Hochberg, 1995). The p values from loci showing ASE and shared by the six adult sheep (ID and coordinate) were unified, using the Stouffer method.
(Dewey, 2016, 2018) and presented as false discovery rate (FDR) for each locus.

Static ASE was calculated in both tissues and BMDMs (each timepoint was considered separately for BMDMs). Condition dependent ASE analysis was carried out only in BMDMs ± LPS both at gene (ICD-ASE) and SNV (Fisher’s exact) level to study LPS inducible ASE.

**Statistical analysis and thresholds applied**

The extraction, transformation and loading of all datasets and subsequent statistical analysis was carried out in R v ≥ 3.4 unless stated otherwise (R Core Team, 2017). System query language join statements (Wickham and Francois, 2015) were used to compare lists of ASE genes or SNVs between samples. Raw p values resulting from all three types of ASE analysis were corrected for multiple testing via Benjamini-Hochberg false discovery rate calculations (Benjamini and Hochberg, 1995). The passing threshold of significance in all analyses was considered to be FDR < 0.1 (10%) except for the Fisher’s exact test association study. Genes showing ASE in multiple tissues were considered those for which 4 or more of the 6 sheep had significant ASE.

**Results**

**Estimation of heterozygous sites across all individuals**

To determine the level of heterozygosity present in the RNA-Seq data we first assessed the number of bi-allelic heterozygote sites per individual for each of the six sheep (range=5,673,703-6,438,497) detailed in Fig.2. Individual variation was observed in the SNVs per gene in each sheep (Fig.2A&C). However, there was no significant difference in the total number of bi-allelic SNVs captured in the RNA-Seq data across all six individuals or between the male and female sheep included in the study (Fig.2B). The bi-allelic SNVs captured in the RNA-Seq data set were annotated using the Ensembl v.92 (Zerbino et al., 2018) reference VCF track. The distribution of SNVs per gene in the Ensembl track is tail-inflated in comparison to the RNA-Seq data Fig.2A. This issue could be due to erroneous assignment of SNVs in hypervariable and repetitive regions, multi-allelic SNVs or simply that there are variants in the Ensembl track that are not expressed (transcribed). The distribution of SNVs for each individual is shown in Fig.2C.

**Reference mapping bias elimination and quality control**

We used the WASP ref bias removal script to successfully minimise ref allele mapping bias in the RNA-Seq samples. The mapping bias was assessed by global distribution of the allelic ratio i.e. \( \frac{\text{ref counts}}{\text{alt counts}} + \frac{\text{ref counts}}{\text{in each RNA-Seq sample}} \) as shown in Fig.3. The ASE discovery rate at the SNV level on average constituted 5.8% of the heterozygote loci that passed the minimum filtration criteria in each individual (0.1% of the total expressed). This portion of the transcriptomic variants belonged to an average of 103 genes in each tissue transcriptome (approx.1%) or 300 in each individual (Supplementary file4 – S4 Fig.1 and Fig.2). As shown in S4 Fig 2. expression level varies across tissues but this does not affect the distribution of ASE genes.

**Genes exhibiting tissue-specific and pervasive ASE signatures**

We used the static mode of GeneiASE to investigate pervasive and tissue-specific ASE
profiles across all of the available samples. The number of genes showing significant static ASE in immune-related tissues across the six sheep are summarised in Table 1. On average approximately 0.5% of the genes in each tissue-specific transcriptome showed significant ASE (approx. 1% of the filtered set of genes). Pervasive ASE genes in all 4 tissues for each of the 6 sheep was investigated by applying the minimum 67% shared rule (i.e. min 4 out of 6 sheep shared significant ASE for the gene). Genes exhibiting significant ASE shared by all 4 tissues were then merged and the ASE effect size averaged (Fig.4A). Six genes exhibited pervasive ASE across tissues (i.e. they were shared across all 4 tissues). In the order of allelic imbalance effect size they were NAA50 (N(alpha)-acetyltransferase 50, Histone acetyl transferase) with highest ASE effect size in spleen, UBB (Ubiquitin B, protein degradation) in thymus, HBP1 (HMG-Box transcription repressor; cell cycle regulation) and ENSOARG00000016510 both in spleen, C1orf105 (Chromosome 1 Open Reading Frame 105) in ileum and MTIF2 (mitochondrial translational initiation factor 2) in thymus.

Sets of genes with tissue-specific ASE profiles were also captured (Fig.4B-E). Thymus had the highest number of tissue-specific ASE genes (n=15) followed by liver (n=12), spleen (n=5) and ileum (n=4) (Fig.4B-E). Amongst the thymus gene set was CD244 which included 30 heterozygote loci with allelic imbalance, one of which was rs406633825. This missense allele (Chr1:110308273 C>A; pVal123Phe MAF=0.7, SIFT score=0 deleterious) has previously been reported in the Texel population characterised by the International Sheep Genome Consortium (ISGC) (Kijas et al., 2012). The liver-specific ASE profile included genes involved in amino acid metabolism, cytochrome oxidase pathways and fibrinogen: FGA (fibrinogen alpha chain), ENSOARG00000003175 (taurochenodeoxycholic 6 alpha-hydroxylase-like, cytochrome P450 family 4 member [CYP4]), ENSOARG0000001568 (novel gene, complement C4-A-like), CYP3A24 (cytochrome P450 family 3 member) and CA3 (carbonic anhydrase 3). Allelic imbalance in spleen was present in CACYBP (calcyclin binding protein), DAPK2 (death associated protein kinase [serine/threonine]) and a novel gene GIMAP8-like (ENSOARG0000001131). The ASE in GIMAP8 has been previously reported in cattle with a strong paternal parent-of-origin expression pattern (Chamberlain et al., 2015). The GIMAP/IAN family of proteins, which are involved in survival, selection and homeostasis of lymphocytes (Nitta and Takahama, 2007).

Two genes of functional interest showed evidence of strong tissue-specific ASE in the spleen: SNAP23 and MYLK. SNAP23 is a key molecule in vesicle transport machinery of the cell and has been reported to be expressed in sheep spleen. SNAP23 or Synaptosome Associated Protein 23 is part of protein complex involved in Class 1 MHC mediated antigen processing and presentation and in neutrophil degranulation (Fabregat et al., 2018). SNAP23 is also vital to lymphocyte development (both B and T) in vitro (Wong et al., 1997; Kaul et al., 2015). The myosin light chain kinase (MYLK) expression in the splenic trabeculae’s smooth muscle has been demonstrated previously (Jiang et al., 2014; Clark et al., 2017). Overall 199 heterozygote bi-allelic loci were present within the MYLK gene. The variant rs40678033 (Chr1:186347056G>A; pAla1014Val), a missense SNV in exon 17 of 33 exons in MYLK, showed consistent allelic imbalance in all spleen samples.

In summary, analysis of ASE across immune related tissues revealed there were a small number of genes that were shared across tissues. ASE signatures instead tended to be tissue-specific, within the sub-set of tissues investigated in this study.

**Individual-specific ASE signatures**

In order to investigate whether ASE profiles were either shared across all six sheep or private to individual sheep we used intersectionality (Fig.5). Each tissue was investigated separately. A number of private (to each individual) ASE genes were detected for each tissue,
Unbiased ovine allele-specific expression

ranging from: 31-123 in ileum (Fig.5A), 24-80 in liver (Fig.5B), 21-83 in spleen (Fig.5C) and 31-66 in thymus (Fig.5D). Some of the shared sets of ASE genes in these tissues were specific to either male or female sheep, these sex-specific ASE signatures are described in Fig.5. In ileum no sex-specific set was observed (Fig5A). In contrast to ileum, the ASE profile for liver included a single gene with female only membership, ENSOARG00000017409 (novel gene; 93% orthology with bovine dicarboxyl and L-xylulose reductase [DCXR]) Fig.5B. In spleen all female sheep shared significant ASE in PMS1 (DNA mismatch repair system component during post meiotic segregation), ANKRD10 (Ankyrin Repeat Domain 10) and ENSOARG00000006103 which was not present in any of the spleen profiles of male sheep (Fig.5C). In thymus there were 2 sex-specific sets: 16 genes showing ASE only in females and 5 genes only in males. The female-specific thymus gene set included: ARPP21 (cAMP regulated phosphoprotein 21), CDKL3 (cyclin dependent kinase like 3), CEP19, ENSOARG00000007110 (novel gene), ENSOARG00000001163 (novel gene), ENSOARG00000007110, ENSOARG0000001163, ENSOARG0000008981 (novel gene; T-lymphocyte surface antigen Ly-9-like), ENSOARG0000006215, ENA000000008981, ENSOARG0000009129, ENSOARG00000011375 (blood vessel epicardial substance [BVES]), ENSOARG00000015755, ENSOARG00000020354 (novel gene; 53% orthology with bovine monoacylglycerol acyltransferase [MOGAT1]), ENSOARG00000025005, ENSOARG00000026030 (novel gene), GPM6A (glycoprotein M6A), RAG1 (recombination Activating 1), STX8 (syntaxin 8). The male-specific thymus set was comprised of ENSOARG0000007267 (novel gene; T-cell surface glycoprotein CD1a-like), ENSOARG00000016841 (novel gene; 98% orthology with bovine ATP synthase membrane subunit G [ATP5MG]), ENSOARG0000007603, SNX25 (sorting nexin 25) and LDHA (lactate dehydrogenase A) Fig.5D.

In summary very few ASE genes were shared across all sheep and the majority of ASE profiles were private to each sheep. Sex-specific ASE signatures were also detected but due to the small sample size (n=3) in both cases these should be interpreted with caution.

ASE in stimulated and unstimulated BMDMs (0hrs vs 7hrs +LPS)

We examined inducible ASE after 7 hrs of exposure to LPS in BMDMs using the ICD mode of GeneiASE (Edsgärd et al., 2016). A comparison of LPS inducible ICD-ASE genes and the genes with background static ASE at 0 and 7 hr time points, was also performed. We first assessed whether differences between 0 and 7 hours could be observed using analysis of static ASE. Individual-specific ASE profiles and a limited number of shared ASE genes were also observed in BMDMs. The total number of genes with static ASE in the BMDMs is shown in Table 2.

Shared static ASE across both time points and independent of LPS induction was only observed in 5 genes. These genes have a macrophage associated function and include ITGB2 (Yee and Hamerman, 2013), SAA3 (ENSOARG0000009963) (Larson et al., 2003; Deguchi et al., 2013), CD200R1 (ENSOARG0000019357) (Ocaña-Guzman et al., 2018), DCTN5 (ENSOARG0000017281) (Habermann et al., 2001) and MTIF2 (also seen in the tissue analysis above) (Overman et al., 2003).

The ICD-ASE in BMDMs +/-LPS captured fewer ASE genes with significant LPS inducible ASE between the 2 timepoints than the static analysis of ASE. Moreover, there were large differences in the number of LPS inducible ASE in each individual sheep, indicating significant individual-specific variation in response to LPS stimulation. BMDM cultured from female 2 showed no LPS inducible response in comparison to male 3 which was a hyper responder with significant inducible ICD-ASE in 28 genes (including 634 informative SNVs total). A detailed breakdown of SNVs, aggregated within each gene, with significant ICD ASE
Unbiased ovine allele-specific expression has been summarised in Fig.6.

In summary the ICD-ASE mode of GeneiASE’s model was not capable of capturing a complete picture of differential ASE in the BMDM experiment. Static ASE was present in both time points, however, there were no inducible ASE genes that were shared across all 6 sheep. The highly diverse ASE profile of BMDMs was very individual-specific, similar to patterns observed in tissues (Supplementary file4. S4 Table 1). These individual-specific differences could be due to individual variation in the innate immune response or experimental variation introduced during primary cell culture or stimulation with LPS.

**Condition dependent ASE at the SNV level in BMDMs**

To further investigate allelic imbalance at the SNV level without the aggregative gene model of GeneiASE, Fisher’s exact test was used. The filtered read counts for bi-allelic SNVs shared by all 6 sheep BMDMs were selected (n=646 sites). Allelic read counts of each SNV were tested using Fisher’s exact test between 0hrs and 7hrs (2x2 table). Overall the 6 sheep shared 646 SNVs with identical allelic genotypes in both time-points of the BMDM RNA-Seq dataset. These SNVs were tested for association with LPS treatment and only 4 SNVs showed a strong association (FDR < 1.0e-08) and 12 SNVs had an FDR between 1.0e-02 and 1.0e-08 (Fig.7). The highest F-statistic was at rs430667535Chr17 Pos:50485358T>C, a synonymous variant in ubiquitin C (UBC), a polyubiquitin precursor, and also an intronic variant T>C or A>G in pro-apoptotic BRI3 binding protein (BRI3BP). This variant was shown to have a minor allele frequency (MAF) of 0.25 in Texel sheep based on the ISCG annotation [ISGC – Ensembl v.92](Kijas et al., 2012). The next highest peak was observed on Chr21 under SNVs within SAA3 gene boundaries (ENSOARG00000009963) at the following coordinates: Highest FDR peak was observed at SNV rs412192652 (Pos:25826978A>G missense variant [Asn145Asp], FDR=2.3e-15) surrounded by rs403064928 (Pos:25826884C>A missense variant [Asp113Glu] in exon 4 of SAA3 FDR=3.6e-07), rs426609498 (Pos:25826845A>G synonymous, FDR=5.5e-07) and rs405439099 (Pos:25826990G>C 3’UTR variant, MAF 0.6 in Texel sheep, FDR=4.1e-05). This region contains a strong LD block previously reported by the ISGC COMPOSITE population (Ensembl v.92) (Supplementary file3) e.g. rs412192652 and rs405439099 pairwise D’ statistics = 1. Two further peaks on Chr3 were observed for rs159926581 (Pos:214731375C synonymous, FDR = 3.3e-09) in ribosomal protein L3 (ENSOARG00000016495) and rs159822214 (Pos:112164732G>A 3’UTR variant, FDR=9.5e-03) in oxysterol binding protein like 8 (OSBPL8). The last inducible ASE associated signal was on Chr16 rs420037698 (Pos:6887423G >A, FDR= 6.9E-09) in ENSOARG0000004700, a known synonymous variant in the Texel population (MAF = 0.55). The SNVs and corresponding genes from Fisher’s exact test are summarized in Table 3.

The LPS inducible ASE analysis, of SNVs, using Fisher’s exact test revealed a different picture not captured by the gene level analysis with the GeneiASE model that aggregates SNVs within each gene (Fig.6 vs Fig7). The aggregative gene model did not capture any shared ASE genes in the ICD-ASE mode. Although the 6 sheep shared 646 SNVs and showed highly significant association with stimulation with the LPS (Fisher’s exact method), the aggregation of ASE effect size from SNV to gene level (ICD-ASE mode) only detected individual-specific sets of ASE genes in each sheep. This contradicted the results from Fisher’s exact test which detected 4 highly significant LPS inducible shared regions (FDR < 0.01, 16 SNVs) on chromosomes 3, 16, 17 and 21. For example, the allelic imbalance in the SAA3 genomic coordinates on chromosome 21 was not detectable in the ICD-ASE model but it was captured by the Fisher’s exact test in all individuals (Fig.7B chromosome 21).

Fisher’s exact analysis at SNV level revealed ASE in response to LPS in variants within CLDN1, ANXA3, BRI3BP, SAA3 and MSR1 (Table3). The anti-inflammatory macrophage
Unbiased ovine allele-specific expression

marker CLDN1 (Van den Bossche et al., 2012) and acute-phase inflammation resolution marker ANXA3 (Yamanegi et al., 2018) have been previously reported with distinct macrophage functions. The macrophage scavenger receptor 1 (MSR1) has also been shown to be involved in lipid uptake and migration ability of macrophages (Shi et al., 2019). Three noncoding RNAs (RF00221 [snoRD43], RF00593 [snoU83B] and RF01151 [snoU82P]) were amongst genes corresponding to ASE inducible SNVs. These 3 snoRNAs all overlap with the genomic coordinates of ribosomal protein L3 (ENSOARG00000016495 / RPL3). The RF00377 [snoU6-53] was also among the ASE positive targets which overlaps the protein-coding gene CDS2 (CDP-diacylglycerol synthase 2). Using bulk RNA-Seq makes it difficult to differentiate the origin of the ASE signal to specific RNA populations.

In summary ASE profiles in BMDMs were highly individual-specific at both gene and SNV level. Moreover, Fisher’s exact SNV level analysis discovered shared ASE SNVs where the aggregative gene model of ICD-ASE mode did not, indicating for condition dependant ASE analysis Fisher’s exact test is more accurate and robust at SNV level.

Discussion

This study is the first to investigate global allele-specific expression across tissues from sheep using RNA-Seq data. We focused our analysis on immune-related tissues and cell types from six adult cross-bred sheep (TxBF) from the sheep gene expression atlas. ASE profiles were highly individual-specific in the six sheep analysed in this study. We were able to identify tissue-specific sets of ASE genes, as well as LPS inducible sets in the BMDM experiment. Tissue-specific signatures of ASE have been previously reported in similar studies in mouse (Castel et al., 2015, 2016), goat (Cao et al., 2019) and cattle (Chamberlain et al., 2015).

Several steps were taken in the cattle study (Chamberlain et al., 2015) to mitigate the ref allele bias, assign parental origin using whole genome sequences and include MAE variants. The SNV filtration was based on the (Hayes and Daetwyler, 2019) 1000 bull genomes project to confirm the heterozygote sites. In our pipeline the Ensembl VCF track was used for that purpose. In the cattle study the 1 and 0 inflated ref allelic ratio was still present and the ASE analysis was done via introduction of 0.9 allelic ratio cut-off. In our pipeline no cut-off is introduced and because it is difficult to distinguish between sequencing error and MAE, a minimum filtration criteria was applied to exclude MAE. Chamberlain et al. tested 5317 genes (14,495 SNVs) in spleen and detected 382 ASE genes (with min > 1 SNV per gene, similar to this study). Although direct comparison would not be appropriate (because we have excluded MAE variants in our analysis) our analysis of sheep spleen revealed ASE in 86 genes (average in 5 sheep) from 8272 filtered genes (average in 5 sheep). Similarly, in thymus the cattle study showed 182 ASE genes from 986 informative genes (9781 SNVs) while from 7961 filtered genes in sheep thymus 134 ASE genes were captured. The differences in the numbers of genes exhibiting ASE between the two studies are likely to be a consequence of the filtration criteria applied, the exclusion of MAE, and species-specific differences between sheep and cow. Results from a more recent study in goat (Cao et al., 2019) more closely reflect our findings. They apply similar filtration criteria to our workflow and discovered 144 ASE genes in liver in comparison to 123 in our sheep liver sets (averaged across 6 sheep). Other recent studies including those focusing on production relevant tissues, such as muscle (Guillocheau et al., 2019), have also applied similar stringency in filtration criteria. The filtering criteria we have used for this analysis is stringent and focused on detecting variants of moderate to extreme effects. Further analysis of the dataset reducing these criteria might discover additional variants exhibiting ASE across individuals and tissues, but it would also increase the potential risk of false positive discovery.

For this analysis we have adapted an ASE analysis workflow with a primary focus on
Unbiased ovine allele-specific expression

mapping bias removal prior to allele-specific analysis of the transcriptome. The collection of scripts for WASP, used for this analysis, or modified versions of them have been utilised by others for mapping bias removal in reference guided genomic datasets e.g. RNA-Seq (Mozaffari et al., 2018; Zhou et al., 2018), Chip-Seq (Pelikan et al., 2018) and for methylomic and epigenetic analysis (Richard Albert et al., 2018).

The ASE analysis pipeline we have adapted for sheep for this study is also adaptable to other species and tissue types with available RNA-Seq datasets. It could be applied, for example, to profile allele-specific expression in the RNA-Seq datasets from livestock species listed on the FAANG data portal (Andersson et al., 2015; Harrison et al., 2018). We used the Ensembl VCF track to capture information at heterozygote loci however the individual VCF file from each sheep could also be used in ASE analysis. The latter strategy might enable the capture of rare variants not included in the publicly available VCF tracks but would also raise the issue of normalisation/standardisation between VCF call sets. The usage of either of these methods will be limited to the number of loci shared by coordinate and bi-allelic genotype (i.e. pervasive ASE discovery). Other studies have compared variants at the RNA and DNA level from the same individual and then removed the DNA variants not present in the RNA-Seq data from the analysis (Guillocheau et al., 2019). We believe that a strength of the pipeline we present is that it does not require parental genotypes and can therefore be applied to other RNA-Seq datasets for livestock where this information is not available.

In our analysis we have not considered either parent-of-origin or breed-of-origin-specific effects in this analysis. For parent-of-origin or breed-of-origin assignment of these ASE profiles DNA level genotypes from the parents of the six sheep from the gene expression atlas (i.e. Texel sire and Scottish Blackface dam) would be required, and these are unfortunately not available. In this study ASE expression profiles also might be affected by the direction of the cross (i.e. Texel sire x Scottish Blackface dam). To fully characterise parent-of-origin or breed-of-origin reciprocal cross experiments would be required. Reciprocal cross studies in mouse (Huang et al., 2017), chicken (Zhuo et al., 2017) and cross-bred cattle (Chen et al., 2016b) have shed light on the complexity of such pervasive ASE markers and parent-of-origin effects. Though potentially very interesting these experiments are lengthy and costly to perform in sheep. Particularly in this case because the reciprocal cross (Scottish Blackface dam x Texel sire) is rarely used in the UK sheep industry and as a consequence has limited relevance to production.

Our approach also excludes mono-allelic expression. The minimum filtration criteria utilised in our workflow along with the reference mapping bias removal step ensures an unbiased ASE discovery in the transcriptome by excluding the ambiguity surrounding MAE variants. This form of analysis is based on the principle that absence of evidence (reads) for either allele of a heterozygote site does not directly amount to evidence of their absence i.e. MAE--The pattern of ASE (ratio of Alt/Ref+Alt) is dependent on the bi-allelic expression of loci within the genomic coordinates of the gene or genomic element of interest. For an ASE effect to be captured by the GeneiASE model, the following criteria must be met: i) Biallelic expression of the locus; ii) Min depth criteria for each allele (min 3 reads , total 10 reads at that site and > 1% of total reads containing that allele); iii) The allelic imbalance or departure from bi-allelic balanced expression being inducible by an environmental trigger (i.e. LPS in ICD ASE experiment with BMDM data). These stringent criteria secure robust transcriptome-wide ASE discovery while maximising the usage of read counts from short read RNA-Seq datasets without considering mono-allelic sites. MAE patterns are impossible to differentiate from sequencing error or random nonsense mediated decay in bulk RNA-Seq, unless arbitrary cutoffs are introduced such as ratio of allelic read counts > 0.9 (Chamberlain et al., 2015) or > 0.7 (Cao et al., 2019). We decided to exclude MAE from this study for the aforementioned reasons.

The tissues utilised for ASE analysis in this study (thymus, spleen, liver and ileum) are
Unbiased ovine allele-specific expression

highly influential on performance of the immune system. ASE profiles shared across tissues and cell types were limited and instead they tended to be highly specific. We identified tissue-specific ASE in several genes in Thymus, for example, that are involved in the T-cell mediated immune response, including CD47 and CD244. These tissue-specific and cell-type-specific ASE profiles may underlie the expression of economically important traits such as disease resistance. Assessment of the connection between economically relevant phenotypes and tissue-specific ASE profiles could be useful for the improvement of genomics enabled sheep breeding programmes, particularly those using specialised sire and dam lines (Georges et al., 2018). Loci exhibiting ASE have been associated with production traits including milk-fat percentage (Hayes et al., 2010; Suárez-Vega et al., 2017), trypanotolerance in small ruminants (Kadarmideen et al., 2011; Álvarez et al., 2016), mastitis in goat (Ilie et al., 2018), Johne’s disease in cattle (Mallikarjunappa et al., 2018) and Marek’s disease in chicken (Maceachern et al., 2011; Meydan et al., 2011; Cheng et al., 2015). Although there is no general consensus currently on the correlation of allelic expression haplotypes and phenotypes under selection in sheep, this form of ASE analysis could pave the way for functional validation at population level (e.g. breed or haplotype-specific aseQTL studies in a larger population of sheep).

Examples of population level aseQTL, eQTL and sQTL (QTLs associated with RNA splicing) already exist for cattle (Wang et al., 2018; Xiang et al., 2018). Knowledge of favourable ASE in critical genes involved in traits of interest could be used as a performance indicator or included as weighted SNVs in genomic prediction algorithms to enhance livestock breeding programmes (Georges et al., 2018). Currently the UK sheep industry is on the cusp of applying genomic prediction, but suitable genomics enabled breeding programmes for sheep already exist in New Zealand and Australia (Daetwyler et al., 2010).

Conclusions

In this study we characterise extreme to moderate allele-specific expression, at the gene and SNP level, in immune-related tissues and cells from six adult sheep (TxBF) from the sheep gene expression atlas dataset. Reference mapping bias removal was an integral component of the analysis pipeline applied in this study. The correction of reference bias prior to obtaining the allelic read counts is a critical step toward true ASE discovery. The workflow developed as part of this manuscript provides an RNA-Seq only dependent tool, without the need for individual DNA sequences. We note that the stringent filtering process applied would remove loci where the allelic imbalance was less extreme but might still be of biological significance.

This study is a novel analysis of an existing large-scale complex RNA-Seq dataset from sheep. Using the pipeline we have adapted for this analysis, we were able to identify ASE profiles that were pervasive in each sheep and specific to the tissues and cell types investigated. These tissue and cell-type-specific ASE profiles may underlie the expression of economically important traits and could be used to identify variants that could be weight in genomic prediction algorithms for the improvement of sheep breeding programmes. In summary we have adapted a robust methodology for ASE profiling, using the sheep gene expression atlas dataset, and provided a foundation for identifying the regulatory and expressed elements of the genome that are driving complex traits in livestock.

Data Availability

The RNA-Seq sequence data is available via the European Nucleotide Archive (ENA) under PRJEB19199 (https://www.ebi.ac.uk/ena/data/view/PRJEB19199). The BAM files were already produced as part of the sheep gene expression atlas following the publicly available protocol (FAANG, 2018) and as described in (Clark et al. 2017). The BAM files have been
Unbiased ovine allele-specific expression

uploaded to ENA under accession numbers ERZ827944, ERZ827949, ERZ827951, ERZ827955, ERZ827972, ERZ827988, ERZ827995, ERZ827997, ERZ828001, ERZ828016, ERZ828019, ERZ828036, ERZ828044, ERZ828046, ERZ828050, ERZ828070, ERZ828073, ERZ828160, ERZ828167, ERZ828168, ERZ828172, ERZ828188, ERZ828192, ERZ828209, ERZ828215, ERZ828217, ERZ828221, ERZ828240, ERZ828244, ERZ828261, ERZ828268, ERZ828270, ERZ828274, ERZ828293 and ERZ828297. The Oar v3.1 reference FASTA and VCF file from Ensembl v92 were used throughout the pipeline. The ASE analysis pipeline (https://msalavat@bitbucket.org/msalavat/asewrap_public.git) was wrapped using bash scripting on Edinburgh Compute and Data Facility computing resource Eddie Mark 3 (Edinburgh, 2018).

Author’s contributions

MS and ELC coordinated and designed the analysis component of the study with assistance from SJB and SPV. MS and SPV designed, optimised and tested the ASE pipeline. DAH acquired the funding for the sheep gene expression atlas project. MEBM, ELC and DAH designed the LPS experiment and generated the data. MEBM performed the LPS stimulation of bone marrow derived macrophages and RNA extraction. SJB performed all bioinformatic analyses prior to analysis with the ASE pipeline. MS performed ASE analysis, visualisation of the results and wrote the manuscript. ELC contributed to manuscript editing and drafting. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank University of Edinburgh Computing and Data Support staff (viz. Dr Andy Law and Dr Steve Thorn) for their input towards optimisation of the pipeline and efficient parallel computing. Rachel Young and Lucas Lefevre isolated the bone marrow derived macrophages for the sheep gene expression atlas project and Iseabail Farquhar helped to collect and archive the tissue samples. The authors would also like to thank Professor Kim Summers and Professor Mick Watson for their comments, which helped to improve the manuscript.

Conflict of Interest

The authors have no competing interest regarding the findings presented in this publication.

Funding

The work was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) (http://www.bbsrc.ac.uk) Grant BB/L001209/1 ('Functional Annotation of the Sheep Genome'). Also, BBSRC Institute Strategic Program Grants: 'Farm Animal Genomics' (BBS/E/D/20211550), 'Transcriptomes, Networks and Systems' (BBS/E/D/20211552) and 'Blue Prints for Healthy Animals' (BB/P013732/1). Edinburgh Genomics is partly supported through core grants from BBSRC (BB/J004243/1), NERC (http://www.nerc.ac.uk) (R8/H10/56) and the Medical Research Council (MRC) (https://www.mrc.ac.uk) (MR/K001744/1). SJB was supported by the Roslin Foundation. ELC is supported by the University of Edinburgh Chancellor’s Fellowship programme. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Unbiased ovine allele-specific expression

References


Unbiased ovine allele-specific expression


Harrison, P. W., Fan, J., Richardson, D., Clarke, L., Zerbino, D., Cochrane, G., et al. (2018). FAANG, establishing metadata standards, validation and best practices for the farmed...
Unbiased ovine allele-specific expression
Unbiased ovine allele-specific expression


R Core Team (2017). R: A language and environment for statistical computing. Available at: https://www.r-project.org/.


Unbiased ovine allele-specific expression


Tables

Table 1. Total number of genes with significant static ASE in proportion to genes containing informative SNVs (filtered). Tissue break-down has been presented as count (%ASE/filtered).

<table>
<thead>
<tr>
<th>Sheep</th>
<th>Thymus</th>
<th>Spleen</th>
<th>Liver</th>
<th>Ileum</th>
<th>Total filtered</th>
<th>Total expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female 1</td>
<td>-</td>
<td>136</td>
<td>153</td>
<td>92</td>
<td>10,379</td>
<td>21,150</td>
</tr>
<tr>
<td>Female 2</td>
<td>136 (1.28%)</td>
<td>70 (0.66%)</td>
<td>116 (1.09%)</td>
<td>70 (0.66%)</td>
<td>10,572</td>
<td>21,420</td>
</tr>
<tr>
<td>Female 3</td>
<td>151 (1.22%)</td>
<td>75 (0.60%)</td>
<td>105 (0.85%)</td>
<td>140 (1.13%)</td>
<td>12,326</td>
<td>21,689</td>
</tr>
<tr>
<td>Avg.</td>
<td>143 (1.25%)</td>
<td>94 (0.85%)</td>
<td>125 (1.13%)</td>
<td>101 (0.89%)</td>
<td>11,092</td>
<td>21,419</td>
</tr>
<tr>
<td>Male 1</td>
<td>95 (0.92%)</td>
<td>-</td>
<td>157</td>
<td>86</td>
<td>10,282</td>
<td>21,361</td>
</tr>
<tr>
<td>Male 2</td>
<td>125 (0.99%)</td>
<td>54 (0.43%)</td>
<td>86 (0.68%)</td>
<td>80 (0.63%)</td>
<td>12,514</td>
<td>21,822</td>
</tr>
<tr>
<td>Male 3</td>
<td>106 (1.01%)</td>
<td>71 (0.67%)</td>
<td>110 (1.04%)</td>
<td>82 (0.78%)</td>
<td>10,480</td>
<td>21,545</td>
</tr>
<tr>
<td>Avg.</td>
<td>109 (0.97%)</td>
<td>62 (0.55%)</td>
<td>118 (1.08%)</td>
<td>83 (0.74%)</td>
<td>11,092</td>
<td>21,576</td>
</tr>
<tr>
<td>Total avg.</td>
<td>126 (1.13%)</td>
<td>78 (0.70%)</td>
<td>121 (1.09%)</td>
<td>92 (0.82%)</td>
<td>11,092</td>
<td>21,497</td>
</tr>
</tbody>
</table>

Table 2. Total number of genes with significant static ASE in BMDMs +/- LPS.

<table>
<thead>
<tr>
<th>Sheep</th>
<th>BMDM 0hrs -LPS</th>
<th>BMDM 7hrs +LPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female 1</td>
<td>237</td>
<td>252</td>
</tr>
<tr>
<td>Female 2</td>
<td>192</td>
<td>233</td>
</tr>
<tr>
<td>Female 3</td>
<td>227</td>
<td>263</td>
</tr>
<tr>
<td>Female avg.</td>
<td>219</td>
<td>249</td>
</tr>
<tr>
<td>Male 1</td>
<td>205</td>
<td>193</td>
</tr>
<tr>
<td>Male 2</td>
<td>300</td>
<td>260</td>
</tr>
<tr>
<td>Male 3</td>
<td>260</td>
<td>261</td>
</tr>
<tr>
<td>Male avg.</td>
<td>255</td>
<td>238</td>
</tr>
</tbody>
</table>

Table 3. The variant IDs of LPS inducible ASE SNVs (Fisher’s exact) and their respective genes. Data was obtained using Ensembl BioMart query builder. Highly significant SNVs are highlighted in bold (FDR < 1.0e-08)

<table>
<thead>
<tr>
<th>ID</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Gene ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs418350332</td>
<td>1</td>
<td>195297663</td>
<td>195297663</td>
<td>ENSOARG00000020472</td>
<td>CLDN1</td>
</tr>
<tr>
<td>rs159822214</td>
<td>3</td>
<td>112164732</td>
<td>112164732</td>
<td>ENSOARG00000014876</td>
<td>OSBPL8</td>
</tr>
<tr>
<td>rs159926581</td>
<td>3</td>
<td>214731375</td>
<td>214731375</td>
<td>ENSOARG00000016495</td>
<td>-</td>
</tr>
</tbody>
</table>
Unbiased ovine allele-specific expression

rs159926581    3    214731375 214731375  ENSOARG000000022372 RF00221
rs159926581    3    214731375 214731375  ENSOARG000000025124 RF01151
rs159926581    3    214731375 214731375  ENSOARG000000025150 RF00593
rs193634916    4    53430934 53430934  ENSOARG00000001754 MDFIC
rs159926581    3    214731375 214731375  ENSOARG000000025124 RF01151
rs418697916    13    46654781 46654781  ENSOARG000000017292 CDS2
rs162298949    9    76759747 76759747  ENSOARG00000001261 SPAG1
rs162298949    6    92979793 92979793  ENSOARG000000018710 ANXA3
rs420037698    16    6887423 6887423  ENSOARG00000004700 -
rs430667535    17    50485358 50485358  ENSOARG00000017707 BRI3BP
rs430667535    17    50485358 50485358  ENSOARG00000018177 UBC
rs403064928    21    25826884 25826884  ENSOARG00000009963 SAA3
rs405439099    21    25826990 25826990  ENSOARG00000009963 SAA3
rs412192652    21    25826978 25826978  ENSOARG00000009963 SAA3
rs161531178    26    19663270 19663270  ENSOARG00000009626 MSR1

Figure captions

Figure 1. A flowchart of the allele-specific expression analysis pipeline applied to the sheep gene expression atlas dataset and optimised for WASP and GeneiASE programmes. The remapping was carried out using HISAT2 (Kim et al., 2015) in combination with SAMtools (Li et al., 2009). The Genome Analysis Toolkit v 3.8 was used for the ASE read counting section.

Figure 2. Distribution of biallelic SNVs expressed per gene in each of the 6 TxBF sheep. The total number of SNVs was averaged across thymus, liver, ileum and spleen for every animal. A) Histogram of SNVs per gene counts in the reference track (Ensembl in grey) and 6 sheep in red (females) and blue (males) overlaid. B) The overall numbers of genes and SNVs detected in each animal (averaged over 4 tissues). C) Individual histograms from section A with females in red and males in blue.

Figure 3. The histogram of a global reference allelic ratio at every locus in the tissues. The distribution of ref allelic ratio showed a balanced profile without any 0 or 1 inflation which is observed in the presence of reference mapping bias. The allelic ratio above 0.51 is shown in blue and below 0.49 in red while balanced bi-allelic expression (0.49-0.51) is coloured in grey. Ref.dp = read counts for reference allele. Alt.dp = read counts for alternate allele. The y axis is square root scaled. As discussed in the text SNP that display MAE are not present in any of the samples analysed, indicating there was no inflation in either 0 or 1 allelic ratio.

Figure 4. Genes exhibiting static ASE shared across tissues from all six sheep. A) genes shared by four tissues with significant (FDR < 0.1) static ASE. B) ASE genes private to Ileum. C) Private to liver. D) Private to spleen. E) Private to thymus.
Unbiased ovine allele-specific expression

**Figure 5.** Intersectionality analysis of genes expressing significant ASE across all six sheep. In each tissue from left to right the set count of genes (dots connected by lines) illustrates the number of sheep sharing the gene. The private sets of genes are located at the far right of each graph (single dots with no line). The intersections are coloured in to illustrate the size of the set of shared genes (red [common to all 6 sheep], green [shared by 5 or 4 sheep], yellow [only in females] and purple [only in males]). Detailed lists of genes with ASE shared by at least 4 sheep are presented above each graph for A) Ileum, B) Liver, C) Spleen and D) Thymus. Two sex-specific sets of genes are highlighted: 16 genes showing ASE only in females (in yellow) and 5 genes only in males (in purple).

**Figure 6.** Intersection analysis of SNVs under genes with significant ICD-ASE in the BMDMs ± LPS. From left to right the set number of genes (dots connected by lines) have been illustrated in order according to the number of sheep sharing the SNV. The private sets of SNVs are located at the far right of each graph (single dots with no line).

**Figure 7.** Scatter plot of the adjusted p values from Fisher’s exact test (unified using Stouffer unification) in BMDMs comparing expression from different alleles at 0 vs 7hrs at SNV level (LPS inducible ASE). A) The graph shows 646 loci exhibiting LPS inducible allelic imbalance shared across all 6 sheep. B) Four loci on chromosomes 3, 16, 17 and 21 with FDR < 1.0e-08.

FDR < 1.0e-02 red line (n=16 SNVs) and FDR < 1.0e-08 blue line (n = 4 SNVs).
Unbiased ovine allele-specific expression

**Figure 2.**

A. Distribution of SNVs per gene in 6 sheep and 1 reference track.
   - i. Over 5e+07 SNVs in the Ensembl v.82 VCF track shown in grey.
   - ii. SNVs expressed in females shown in red and males in blue.
   - iii. Number of SNVs per gene (average of 4 tissue RNA-Seqs in each animal).

B. Table showing SNVs per gene:
   - Female1: 21150, 6151417
   - Female2: 21420, 5837076
   - Female3: 21689, 5961473
   - Male1: 21361, 5982189
   - Male2: 21822, 6438497
   - Male3: 21545, 5673703
   - Ref: 23974, 50752529

C. Distribution of total annotated SNVs (bi-allelic) per gene.
Unbiased ovine allele-specific expression
Figure 4.

Unbiased ovine allele-specific expression
Unbiased ovine allele-specific expression

Figure 5.

A

B

C

D

The figure shows bar charts for different sets of alleles across genders. Each chart represents a different set size with bars indicating the expression levels. The x-axis represents the set size, while the y-axis shows the expression levels. The charts are color-coded to distinguish between different sets and genders.
Unbiased ovine allele-specific expression

Figure 6.

![Graph showing unbiased ovine allele-specific expression](image)

Figure 7.

![Graphs showing allele-specific expression on chromosomes](image)
Supplementary material

**A bimodal p value distribution was observed in GeneiASE output as shown in S1 Fig.1:**

![GeneiASE original script](image1)

**S1 Figure 1.** Distribution of p values resulted from the beta-binomial test (Static ASE example for spleen in female 1). The uniformity of the p value distribution showed a bimodal (high frequency close to 1 followed by an extensive lower tail close to 0). R machine characteristics for printing float numbers were limited to 6 significant numbers after the decimal point by default which resulted in a high number of absolute zero (0+e0.0) output-values.

![qnorm modified for R version >= 3.4](image2)

![pvalue ==1 removal and FDR recalculation](image3)

**S1 Figure 2.** The figure shows the distribution of p values produced by GeneiASE after cut off introduction (top) and after removal of p value == 1 values (middle). The distribution of FDR calculated by GeneiASE and after correction are shown at the bottom (right and left respectively). To correct for multiple testing by recalculation of FDR and eliminating the non-uniformity of p value distribution around 1 the following steps were performed as shown in S1 Fig.2: i) An arbitrary cut off (1e-7) was placed over all p values smaller than the said limit including the absolute 0e+0.0; ii) The genes showing p values equal to 1 were removed from each data frame; iii) The false discovery portion and rate were recalculated in each sample. iv) Gene showing FDR < 0.1 selected for further analysis.
Unbiased ovine allele-specific expression

S1 Figure3. The QQ plot of observed vs expected p values in corrected (left) and GeneiASE raw (right) calculated outputs.

Supplementary_file2.zip
The zip file includes all the significant ASE values for genes and SNVs in the study.

Supplementary_file3.pdf
**Figure 1.** The LD block (A) and LD Manhattan plot (B) of SAA3 transcript (ENSOART00000010843). The LD block produced as part of International Sheep Genome Consortium is publicly available on Ensembl using the COMPOSITE population (n = 98).

---

Supplementary_file4.pdf
Unbiased ovine allele-specific expression

**S4 Figure 1 supp to Figure 2.** The proportional distribution of SNVs being expressed, passing minimum filtration criteria and showing significant ASE (FDR 10%). A) Count of genes in 6 sheep expressed in grey, informative/filtered in blue and ASE positive in red. B) Break down of genes and SNVs that are expressed, pass minimum filtration criteria and are ASE positive.

**S4 Figure 2 supp to Figure 3.** Scatter plot of allelic depth (read count) in all 4 tissues of the 6 TxBF sheep. The SNVs passing minimum filtration criteria are coloured in grey and significant ASE (FDR10%) are marked by red and blue. If reference read count > alternate read count the point is shown in blue and vice versa in red. Expression level varies across tissues but does not affect the ASE profile.
Unbiased ovine allele-specific expression

**S4 Table 1 supp to Figure 6.** Mean allelic imbalance (log2(ASE\_LPS/ASE \_\_LPS)) in all 6 sheep.

Only 1 gene (VCAN including 7 SNVs) exhibited inducible ASE which was shared between 3 sheep (in bold), while two (GCNT1 and SYNE3) showed inducible ASE in two sheep. Diverse and highly individual ASE response to LPS was observed in each sheep.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Female1</th>
<th>Female2</th>
<th>Female3</th>
<th>Male1</th>
<th>Male2</th>
<th>Male3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOD1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ASH2L</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.9</td>
<td>-</td>
</tr>
<tr>
<td>CASP8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.8</td>
</tr>
<tr>
<td>CCDC80</td>
<td>-</td>
<td>-</td>
<td>9.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CLDN1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10.3</td>
</tr>
<tr>
<td>CPOX</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.5</td>
</tr>
<tr>
<td>EDNRB</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11.4</td>
</tr>
<tr>
<td>EMP1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23.9</td>
</tr>
<tr>
<td>ENSOARG00000001174</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.0</td>
<td>-</td>
</tr>
<tr>
<td>ENSOARG00000001910</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7.8</td>
</tr>
<tr>
<td>ENSOARG00000002108</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8.0</td>
</tr>
<tr>
<td>ENSOARG00000002898</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.3</td>
</tr>
<tr>
<td>ENSOARG00000004106</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7.5</td>
</tr>
<tr>
<td>ENSOARG00000006889</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>31.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ENSOARG00000007547</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8.2</td>
<td>-</td>
</tr>
<tr>
<td>ENSOARG00000007603</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ENSOARG00000014875</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.6</td>
</tr>
<tr>
<td>ENSOARG00000016081</td>
<td>-</td>
<td>6.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ENSOARG00000018177</td>
<td>-</td>
<td>-</td>
<td>11.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ENSOARG00000021562</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.0</td>
</tr>
<tr>
<td>FDX1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>17.3</td>
<td>-</td>
</tr>
<tr>
<td>GCNT1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>16.9</td>
<td>-</td>
<td>30.4</td>
</tr>
<tr>
<td>GSDME</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HNMT</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>15.1</td>
</tr>
<tr>
<td>HUNK</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.0</td>
</tr>
<tr>
<td>IL1RN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.7</td>
</tr>
<tr>
<td>KANSL1L</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.7</td>
</tr>
<tr>
<td>KMO</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.7</td>
</tr>
<tr>
<td>MB21D2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MYO10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7.8</td>
</tr>
<tr>
<td>NRP1</td>
<td>-</td>
<td>-</td>
<td>9.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OGFRL1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>16.5</td>
</tr>
<tr>
<td>PDE4D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11.7</td>
</tr>
<tr>
<td>PDK4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PGM2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.5</td>
</tr>
<tr>
<td>PLK2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.7</td>
</tr>
<tr>
<td>PP1F</td>
<td>-</td>
<td>6.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PRNP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8.8</td>
</tr>
<tr>
<td>RASSF3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Unbiased ovine allele-specific expression

<table>
<thead>
<tr>
<th>Gene</th>
<th>Allele 1</th>
<th>Allele 2</th>
<th>Allele 3</th>
<th>Allele 4</th>
<th>Allele 5</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGS1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8.4</td>
</tr>
<tr>
<td>SAFB2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.3</td>
</tr>
<tr>
<td>SIGLEC15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>SLC35F2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.5</td>
</tr>
<tr>
<td>SYDE1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.1</td>
</tr>
<tr>
<td>SYNE3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7.2</td>
<td>-</td>
<td>10.7</td>
</tr>
<tr>
<td>TBC1D1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.3</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>TFEC</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10.3</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>TRAF1</td>
<td>12.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>VCAN</td>
<td>23.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14.0</td>
<td>13.9</td>
</tr>
<tr>
<td>VLDLR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.7</td>
<td></td>
</tr>
</tbody>
</table>

1089
1090 f