GWAS Analysis of Unknown Sheep Phenotype

Ben Murarotto

2025-05-25

Contents

1 Introduction **1**

2 Materials and Methods. **2**

2.1 Reading in genetypic and phenotypic data. 2

2.2 Calculating linear model coefficients. 2

3 Results. **2**

4 Discussion **2**

4.1 ARHGEF39 3

4.2 TESK1 3

4.3 CA9 3

4.4 CD72 3

5 Conclusion **4**

6 References **4**

7 Appendix **4**

1 Introduction

Genome-wide association studies (GWAS) are a useful and modern mthod of exploring the genetic basis of complex traits by looking for statistical links between genetic variants and observed phenotypes. In this project, genotypic and phenotypic data from 300 sheep were used to investigate an unknown residual phenotype as part of the GENE552 course at the University of New England. The dataset included over 48,000 SNP markers aligned to the Ovis aries genome (Oar_v3.1).

After cleaning and centering the data, I ran a series of single-SNP linear models to test for associations between each SNP and the phenotype. From there, Manhattan and Q-Q plots were generated to visualize the results and assess the overall quality of the analysis. While the majority of SNPs behaved as expected under the null hypothesis, one SNP on chromosome 16 stood out, exceeding the Bonferroni significance threshold. I then delved more closely into this region and identified several nearby genes that might aid in explaining the biological basis of the association.

2 Materials and Methods.

This study was founded on data provided by the University of New England's Genomic Analysis and Bioinformatics (GENE552) course. Genotypic data was provided for 300 sheep for 48,570 SNP markers. Residual values for an unknown phenotype were used provided along with a corresponding SNP reference map aligned to *Ovis aries* genome assembly Oar_v3.1. All anaylsis was performed using R in RStudio suite, the code of which can be found in the appendix of this paper.

2.1 Reading in genetypic and phenotypic data.

To prepare the data for association analysis, data was loaded into RStudio. The genotype data was then centred by subtracting twice the allele frequency from each value, so that each SNP had a mean of zero. Transposing the matrix before and after this step helped keep individuals as rows.

2.2 Calculating linear model coefficients.

A standard linear model was fitted for each SNP. We determined strength of effect and associated p-value using the lm() function.

3 Results.

The phenotype values used in this analysis are residuals,they represent what's left over after adjusting for fixed effects like environmental or management factors. These aren't raw trait values but instead reflect the variation that couldn't be explained by those fixed effects — in other words, the "leftover" differences between individuals that might have a genetic basis.

As shown in the density plot (Figure 1), these residual values are mostly centered around zero, with a slightly left-skewed shape. This suggests a few individuals had much lower values than expected based on the fixed effects. The values range from about −22 to +22, showing a fair amount of variation across the sample. Even though the distribution is not perfectly Gaussian, it's still unimodal and suitable to fit the linear model.

After fitting the model and the corresponding significant points a Manhattan plot was generated (Figure 2). The GWAS yielded SNPs of significant effect located on chromosomes 16, that being OAR16_57327865.1, positioned at 52.6 Mb with a corresponding p-value of 7.33e-07.

A QQ plot (Figure 3) was generated to assess how well the observed SNP p-values aligned with those expected under the null hypothesis of no association. For most of the genome, the points closely followed the expected line then the curve begins to deviate upward at the tail, showing an excess of small p-values. This suggests that some SNPs may truly be associated with our trait of interest. The GIF (λ) was calculated to be 1.408, suggesting moderate inflation of the test statistics. This level of inflation could reflects that there may be residual effects that wasn't fully corrected for by the model.

4 Discussion

Initial visual inspection of the Manhattan plot revealed no clear genomic region with a dense cluster of significant SNPs. The observed associations were scattered across the genome, with no distinct peak that would indicate a strong candidate locus. This suggests that no single variant stood out as a likely causal locus, and the trait under investigation may be influenced by multiple small effect loci or complex effects. It is also possible that the sample size was insufficient to detect variants of strong effect within genome wide significance. Nonetheless, we will still assess and further investigate our most significant SNPs.

The singular significant SNP when accounting for Bonferroni correction is OAR16_57327865.1, which lies in a region on chromosome 16, 100 kb downstream of ARHGEF39 and within 200 kb of CA9, TESK1, and CD72.

4.1 ARHGEF39

The gene AHRGEF39 codes for a protein of the same name. It is highly involved in cell growth and proliferation regulatory factors through activation of Rho GTPases which facilitate cell signalling pathways. In a study by Cooke et al., (2020), ARHGEF39, along with FARP1 and TIAM2, was identified as a driver of cell motility signaling in human lung adenocarcinoma cells. These pathways suggests that mutation near ARHGEF39 in sheep might modulate how cells migrate, proliferate, or differentiate during development. This could manifest in potential phenotypes such as body weight or tissue size.

4.2 TESK1

The gene TESK1 encodes for the TESK1 protein. TESK1 is primarily understood for its role in regulating cytoskeletal organisation and cell morphology. TESK1 functions as a kinase that phosphorylates cofilin, an actin binding protein. This regulation of actin filaments has influence over processes such as cell shape maintenance, migration, and differentiation (Johne et al., 2008). TESK1 is highly expressed in testicular tissue and has been implicated in spermatogenesis, particularly during germ cell maturation. Given its involvement in actin dynamics and cellular development, mutations near TESK1 in sheep may influence traits related to reproductive performance, muscle development, or growth, potentially manifesting in phenotypes such as fertility or tissue size (Johne et al., 2008).

4.3 CA9

The gene CA9 encodes for carbonic anhydrase 9, an enzyme involved in the regulation of cellular pH by catalyzing the reversible hydration of carbon dioxide. CA9 expression is typically induced under hypoxic conditions via activation of hypoxia-inducible factor 1 alpha (HIF-1α) and is commonly used as a marker of hypoxia and solid tumor progression.

Recent work has shown that nitrosative stress, through the S-nitrosylation of DNA methyltransferases (DNMTs), can epigenetically induce CA9 expression in normal human small airway epithelial cells (SAECs). This occurs via hypomethylation of the CA9 promoter and increased recruitment of HIF-1α (Fujimoto et al., 2022). The study also demonstrated that the chemical compound DBIC can inhibit S-nitrosylation of DNMT3B, suppressing NO-induced CA9 upregulation.

Given this regulatory mechanism, variation near CA9 in sheep may influence traits associated with oxygen homeostasis, inflammation response, or tissue pH balance which translates potentially in phenotypes such as respiratory function.

Although CA9 has evidence for epigenetic regulation, it is plausible that genetic variation may affect this regulation or interact with it, influencing downstream phenotypes and thus appearing in our results.

4.4 CD72

The gene CD72 produces a protein found on the surface of B cells, which are crucial in the immune system. It is present from early B cell development through most of their life cycle, although its levels drop as B cells mature into antibody producing plasma cells. CD72 acts as a co-receptor alongside the B cell receptor helping to regulate how B cells respond to signals. Inside the cell, CD72 has two important regions: one that helps turn signals off by recruiting a phosphatase enzyme called SHP-1, and another that helps turn signals on by connecting to a signaling adaptor protein called Grb2, which activates pathways like Ras involved in cell growth and survival (Smith et al., 2023).

CD72 is involved in several important B cell functions, including proliferation, apoptosis, and development. In autoimmune diseases, CD72 expression can change.For example, a recent study found that people with primary Sjogren's syndrome (pSS)— a had higher levels of CD72 on their B cells and state that the condition is marked by overactive B cells and inflammation. The study also looked at a floating form of CD72 in the blood, called soluble CD72, suggesting CD72 might play a role in regulating immune responses in this disease (Smith et al., 2023).

5 Conclusion

Although my GWAS didn't reveal a strong cluster of significant SNPs, the OAR16_57327865.1 variant on chromosome 16 passed the Bonferroni-corrected threshold, pointing us toward a potentially meaningful region. This SNP sits near several genes with interesting and differing function: ARHGEF39 is involved in cell proliferation and motility, TESK1 regulates cytoskeletal structure, CA9 plays a role in pH regulation and hypoxia response, and CD72 is important in B cell signaling and immune function.

The QQ plot showed a slight excess of small p-values ($\lambda = 1.408$), which might reflect some inflation, though overall the results held up reasonably well. That said, the sample size of 300 sheep likely limited our ability to detect variants of smaller effect, and the phenotype itself was unknown, which makes interpretation a bit more speculative.

Still, this analysis highlights how even a single significant SNP when paired with biological context can offer useful clues about potential genetic mechanisms. Further investigation with more samples and a defined phenotype would help confirm these findings and possibly uncover more about the roles these nearby genes might play in shaping the trait.

6 References

Cooke, M., Kreider-Letterman, G., Baker, M. J., Zhang, S., Sullivan, N. T., Eruslanov, E., Abba, M. C., Goicoechea, S. M., García-Mata, R., & Kazanietz, M. G. (2021). FARP1, ARHGEF39, and TIAM2 are essential receptor tyrosine kinase effectors for Rac1-dependent cell motility in human lung adenocarcinoma. Cell Reports, 37(6), 109905. https://doi.org/10.1016/j.celrep.2021.109905

Johne, C., Matenia, D., Li, X. Y., Timm, T., Balusamy, K., & Mandelkow, E. M. (2008). Spred1 and TESK1—Two new interaction partners of the kinase MARKK/TAO1 that link the microtubule and actin cytoskeleton. Molecular Biology of the Cell, 19(4), 1391–1403. https://doi.org/10.1091/mbc.e07-07-0730

Fujimoto, H., Watanabe, Y., Osawa, S., Hori, T., Adachi, T., & Maehara, K. (2022). Nitrosative stress epigenetically regulates CA9 expression through S-nitrosylation of DNMT3B in human airway epithelial cells. Biological & Pharmaceutical Bulletin, 45(5), 620–627. https://doi.org/10.1248/bpb.b21-00968

Smith, Y., Zhou, C., Gao, Y., Chen, Y., Zhang, X., Zhang, J., & Wu, Y. (2023). Aberrant expression of CD72 and soluble CD72 in patients with primary Sjögren's syndrome. Immunologic Research, 71(4), 361–371. https://doi.org/10.1007/s12026-023-09337-7

7 Appendix

```r
geno <- read.table("geno50k.txt", header=F)
pheno <- read.table("pheno.txt", header=T)
map <- read.table("map.txt", header=T)
```

```r
dim(geno)
head(geno[1:10,1:5])
```

```r
N=dim(geno)[1] #nr ofsamples
M=dim(geno)[2] #nr of markers
```

```r
p=colMeans(geno)/2 # calculate allele frequency (p) for every SNP, p=sum/2
geno=t(geno)-2*p
geno=t(geno) # the t() function takes a transpose, used here to make arrays fit
head(geno[1:10,1:5])
```

```
nrow(pheno)
nrow(geno)
```

```
# single SNP regression with lm ==========================================
effect=numeric(M) # define an array to store effect sizes (coefficients from linear model)
pval=numeric(M) # define an array to store p-values ( from linear model)
```

```
class(pheno)
str(pheno)
nrow(pheno)
head(pheno)
```

```
for (i in 1:M) {
 res=coef(summary(lm(pheno$phenotype~geno[,i])))[2,c(1,4)]
 effect[i]=res[1]
 pval[i]=res[2]
}
```

```
library(qqman)

gwas = (cbind (map,pval))
colnames(gwas)=c("SNP","CHR","BP","P")
```

```
bonferroni_threshold <- -log10(0.05 / nrow(gwas))

jpeg('Manhattanplot.jpg', width=1200)
manhattan(
  gwas,
  chr = "CHR",
  p = "P",
  ylim = c(0, 10),
  suggestiveline = -log10(1e-4),  # Optional: suggestive line
  genomewideline = bonferroni_threshold,
  genomewideline.col = "red",
  cex = 1.2,
  cex.axis = 1.1,
  col = c("blue", "green")
)
dev.off()


chi2 <- qchisq(1 - pval, df = 1)
lambda <- median(chi2) / qchisq(0.5, df = 1)

jpeg('QQplot.jpg', width=1100)
qq(pval, main = paste0("Q-Q Plot (Lambda = ", round(lambda, 3), ")"))
dev.off()
```

```
x <- 3
top_x_idx <- order(pval)[1:x]
top_x_snp <- gwas[top_x_idx, ]

top_x_snp
```

```
library(dplyr)
library(biomaRt)
ensembl <- useEnsembl(biomart = "genes", dataset = "oaries_gene_ensembl", host = "https://asia.ensembl.org"
lookup_genes <- function(chr, position, window = 500000) {
  genes <- getBM(
    attributes = c("external_gene_name", "ensembl_gene_id", "chromosome_name", "start_position", "end_posit
    filters = c("chromosome_name", "start", "end"),
    values = list(chr, position - window, position + window),
    mart = ensembl
  )
  return(genes)
}
```

```
listDatasets(useEnsembl("genes", host = "https://asia.ensembl.org")) %>%
  subset(dataset == "oaries_gene_ensembl")
```

```
results_list <- list()

for (i in 1:nrow(top_x_snp)) {
  chr <- as.character(top_x_snp$CHR[i])
  pos <- top_x_snp$BP[i]
  nearby_genes <- lookup_genes(chr, pos)
  results_list[[i]] <- data.frame(SNP = top_x_snp$SNP[i], nearby_genes)
}

all_nearby_genes <- do.call(rbind, results_list)
```

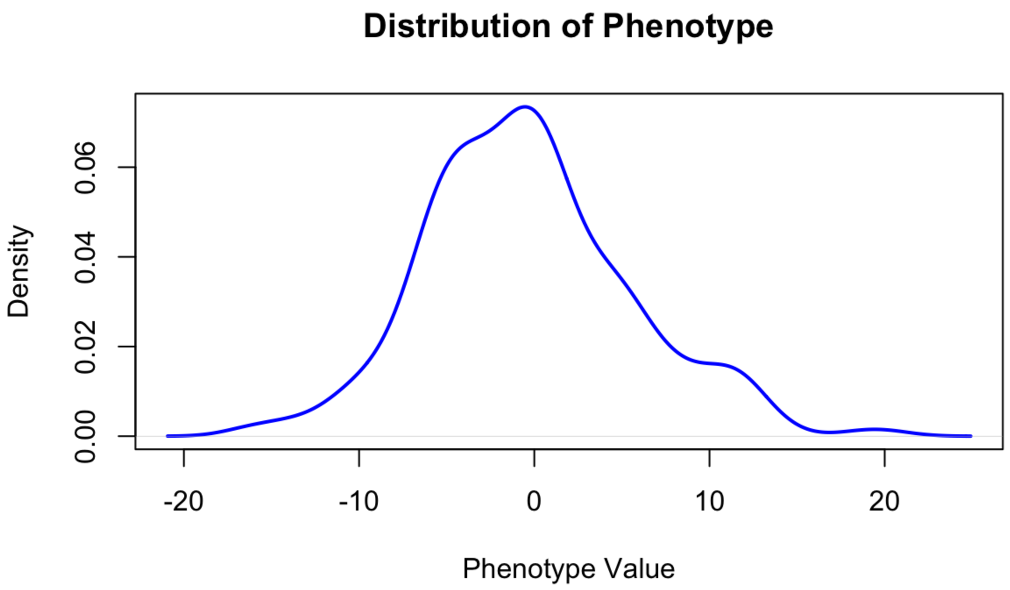

Figure 1: Phenotype Distribution

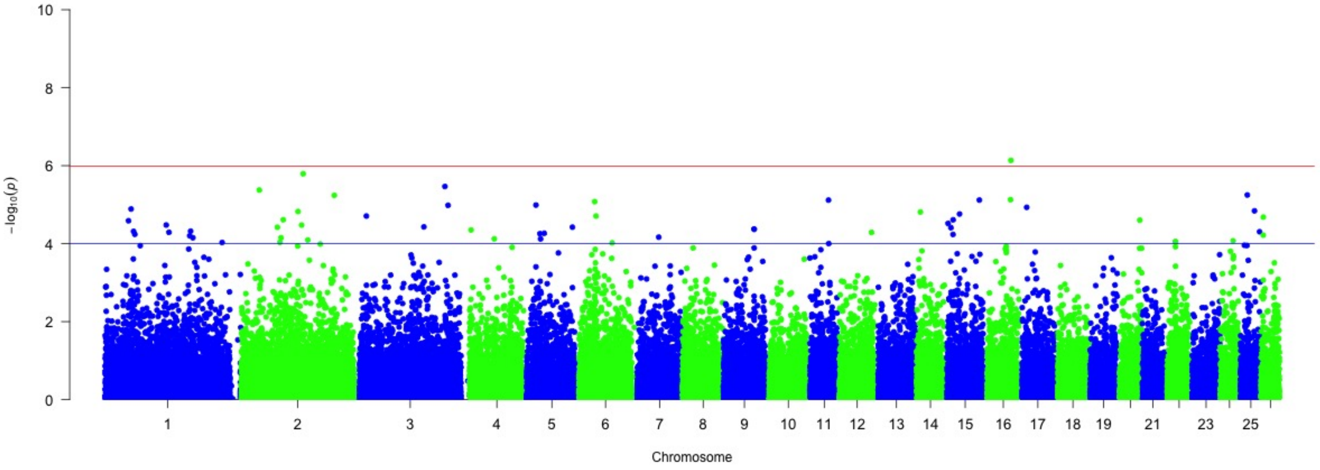

Figure 2: Manhattan Plot

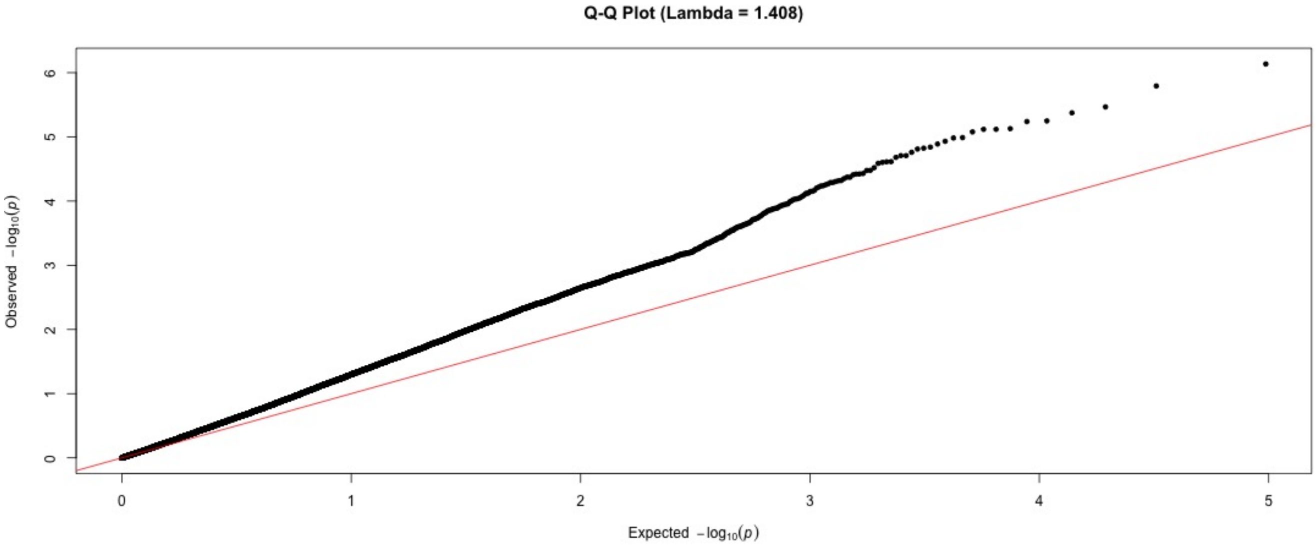

Figure 3: QQ Plot