**International Journal of Genetics and Genomics**
2019; 7(4): 103-109
http://www.sciencepublishinggroup.com/j/ijgg
doi: 10.11648/j.ijgg.20190704.13
ISSN: 2376-7340 (Print); ISSN: 2376-7359 (Online)

# Novel Candidate Genes for Somatic Cell Count in Frizarta Dairy Sheep

**Antonios Kominakis[1, *], Aggeliki Saridaki[2], George Antonakos[3]**

[1]Department of Animal Science, Agricultural University of Athens, Athens, Greece

[2]School of Environmental Engineering, Technical University of Crete, University Campus, Chania, Greece

[3]Agricultural and Livestock Union of Western Greece, Lepenou, Greece

**Email address:**
acom@aua.gr (A. Kominakis)
[*]Corresponding author

**To cite this article:**
Antonios Kominakis, Aggeliki Saridaki, George Antonakos. Novel Candidate Genes for Somatic Cell Count in Frizarta Dairy Sheep. *International Journal of Genetics and Genomics.* Vol. 7, No. 4, 2019, pp. 103-109. doi: 10.11648/j.ijgg.20190704.13

**Received**: July 18, 2019; **Accepted**: September 11, 2019; **Published**: October 21, 2019

**Abstract:** Aim of the present study was to identify genomic regions and candidate genes impacting on somatic cell count in the Frizarta dairy sheep. A total number of 482 Frizarta ewes genotyped with the medium density SNP array with available records on milk somatic cell count were used. Associations between genomic markers and the trait under study were detected by application of a multi-locus mixed model treating markers as fixed additive effects. Positional candidate genes identified within 1Mb flanking distances from significant markers were *in silico* prioritized based on their functional similarity to a training gene list including 1,120 genes associated with the term 'immunity'. Association analysis pinpointed 4 chromosome-wide significant SNPs dispersed on four autosomes (OAR2, OAR18, OAR19 and OAR22). A total number of 37 positional candidate genes were identified within the searched genomic distances while 13 candidate genes were highly prioritized. Seven highly prioritized genes (*NFIB, GFRA1, PSIP1, ARHGAP5, HECTD1, EMX2, STRN3*) along with genes *FREM1* and *GPR33* had evidenced involvement in immune-related processes. Current results extent previous findings by providing novel candidate genes for the somatic cell count phenotype in dairy sheep.

**Keywords:** Somatic Cell Count, Mastitis, Dairy Sheep, GWAS, Prioritization Analysis

## 1. Introduction

Somatic cell count (SCC) in dairy sheep milk is important in many aspects, including health and production. As in dairy cows, mastitis in dairy ewes is associated with increased SCC in milk [1]. Hence, milk with elevated SCC is usually considered as an indication of intra-mammary infection (IMI) and selection for decreased SCC could lead to reduced susceptibility to mastitis [2]. IMI in dairy sheep differs from the respective bovine infections in both incidence and aetiology. In dairy ewes, lower incidence of clinical mastitis (CM) versus subclinical mastitis (SCM) is observed, while the major pathogens in this species are the coagulase-negative *Staphylococci* [2]. Furthermore, in dairy ewes, SCC can reach highest counts (e.g. $4 \cdot 10^6$ cells/ml) without mastitis symptoms, with milk still of normal macroscopic appearance [3]. However, SCC levels between affected and non-affected udders seem to clearly differentiate, with sheep milk from udders free from IMI having an average of $0.185 \cdot 10^6$ cells/ml when contrasted to an average equal to $1.445 \cdot 10^6$ cells/ml of infected halves [3]. In milk from uninfected mammary glands, macrophages are the predominant cell type of the cell population (46 to 84%), followed by lymphocytes (11 to 20%), polymorphonuclear neutrophilic leukocytes (PMN) (2 to 28%) and epithelial cells (1 to 2%). In infected mammary glands, the percentage of PMN increases to 50% at a SCC of $0.20 \cdot 10^6$ cells/ml and up to 90% at a SCC over $3 \cdot 10^6$ cells/ml [3] playing a protective role in the mammary gland [4].

The advent of high-throughput genotyping platforms made genome-wide association studies (GWAS) a reality and the identification of the responsible functional genes involved in SCC in dairy sheep a promising task. Early genome scans in the Spanish Churra sheep identified a single genome-wide suggestive QTL (Quantitative Trait Locus) for SCS (Somatic Cell Score) on OAR20 [5] with peak QTL location close to marker OLADRBPS, which is located in the major histocompatibility complex (MHC). In a crossbred population of Sarda x Lacaune breeds, there were two genome-wide significant QTLs identified on OAR6 and OAR13, along with 11 genome-wide 'suggestive' QTLs associated with SCS (2). In a crossbred population of Lacaune x Manech breeds, a 'suggestive' genome-wide significant QTL was found on OAR14 for SCS [6]. In an Awassi x Merino crossbred population, a significant QTL for SCC was found on OAR14, whereas two 'suggestive' QTLs for this trait were reported on OAR17 and OAR22 [7]. Rupp et al. [8] identified a major QTL associated with SCC on ovine chromosome 3. Fine mapping of the region, provided one strong candidate SNP that mapped within the coding sequence of a highly conserved gene, suppressor of cytokine signalling 2 (*SOCS2*), mediated by the JAK/STAT signaling pathway. Using transcriptional profiling of the milk somatic cells of susceptible and resistant sheep infected by *Staphylococcus sp*., Bonnefont et al. [9] provided a list of differentially expressed genes between the resistant and susceptible animals that were associated with immune and inflammatory responses, leukocyte adhesion, cell migration and signal transduction. Most recently, Banos et al. [10] identified SNPs associated with mastitis traits on chromosomes 2, 3, 5, 16 and 19 and proposed relevant candidate genes such as *SOCS2*, *CTLA4*, *C6*, *C7*, *C9*, *PTGER4*, *DAB2*, *CARD6*, *OSMR*, *PLXNC1*, *IDH1*, *ICOS*, *FYB* and *LYFR*, implicated in innate immunity.

In the present study, we performed a genome scan of 482 Frizarta dairy ewes using a medium density genotyping SNP array to identify genomic regions associated with the SCC phenotype. We then applied *in silico* gene prioritization analysis of the positional candidate genes to identify the most plausible functional candidate genes for the trait. Current findings are expected to contribute to a better understanding of the genetic mechanisms underlying the SCC phenotype in the Frizarta dairy sheep.

## 2. Material and Methods

### *2.1. SNP Genotyping and Quality Control*

A total of 524 dairy ewes of the Frizarta breed were originally sampled. These ewes were randomly chosen from 7 herds of the Cooperative of the Agricultural and Livestock Union of Western Greece (ALUWG) in Agrinio (north-western Greece). Ewes are kept under an intense production system, with standardized conditions and feeding regime. DNA was extracted from blood samples using the NucleoSpin Blood kit (Macheray-Nagel). Sample genotyping based on the Illumina OvineSNP50 BeadChip was performed commercially at Neogen Europe, Ltd. From the 524 original samples, one sample could not be genotyped. Quality Control (QC) of the remaining 523 genotypes was performed in two stages, first on an 'individual' and second on a 'marker' basis. On the first level, samples were removed if they had: i) call rate<0.95 and ii) overall autosomal heterozygosity rate outside the 1.3 inter-quartile range (0.346-0.389). Using these criteria, the number of available animals (samples) was reduced to 503. Marker QC was based on the following criteria: (i) call rate (>0.95), (ii) minor allele frequency (MAF)≥0.05 and (iii) Fisher's Hardy-Weinberg equilibrium (HWE) $p<0.0001$. Only mapped autosomal SNPs were considered. From the originally available SNPs (n=54,013), the final number of SNPs retained for the GWAS was 42,884.

### *2.2. Phenotypic Data*

There were SCC records for 482 out of the 503 genotyped animals fulfilling the QC criteria. These records were obtained using a MilkoScan™ FT system. Prior to analysis, original SCC values (given in $10^5$ cells/ml) were logarithmically transformed to approximate normality using the following transformation function: lSCC=log(SCC.$10^{-5}$ cells/ml). lSCC ranged from 0.07 (104,972 cells/ml) to 4.42 (2,140,684 cells/ml) with average and standard deviation equal to 1.93 (381,055 cells/ml) and 0.96 (194,531 cells/ml), respectively. SCC records were obtained from ewes dispersed in 7 herds, 5 lactations, 4 production years, 3 lactation stages and 8 classes of month measurements. Detailed inspection of the number of observations per class effect showed that some of month measurement classes were not adequately represented. For this reason, grouping of observations for this particular effect was performed by grouping classes as follows: calendar month 1 (1, 2 and 3), 8 (8 and 9) and finally 11 (11 and 12). Furthermore, three lactation stages were defined as follows: early: (1-100 days), middle (101-160 days) and late (>160 days) days of lactation. The final data set included SCC records per 7 herds, 5 lactations (classes: 2, 3, 4, 5 and 6), 4 production years (classes: 2011, 2012, 2013 and 2014), 4 months of measurement (classes: 1, 8, 10 and 11) and 3 lactation stages (classes: early, middle and late) (Table 1).

***Table 1.*** *Number of observations (n) and means (M) with standard errors (SEM) for SCC at the logarithmic scale [log(SCC.$10^{-5}$ cells/ml)] per fixed effect class.*

| Herd | | Production year | | Lactation number | | Month of measurement | | Lactation stage | |
|---|---|---|---|---|---|---|---|---|---|
| Class (n) | M ± SEM | Class (n) | M ± SEM | Class (n) | M ± SEM | Class (n) | M ± SEM | Class (n) | M ± SEM |
| 1 (125) | 2.358 ± 0.051 | 2011 (45) | 2.445 ± 0.091 | 2 (153) | 1.979 ± 0.078 | 1 (20) | 2.150 ± 0.113 | 1 (258) | 1.727 ± 0.062 |
| 2 (51) | 1.998 ± 0.138 | 2012 (36) | 2.126 ± 0.202 | 3 (109) | 2.008 ± 0.087 | 9 (220) | 2.147 ± 0.054 | 2 (196) | 2.124 ± 0.061 |
| 3 (62) | 2.635 ± 0.088 | 2013 (42) | 2.370 ± 0.109 | 4 (83) | 1.946 ± 0.109 | 10 (192) | 1.617 ± 0.074 | 3 (27) | 2.548 ± 0.120 |
| 4 (30) | 2.289 ± 0.098 | 2014 (358) | 1.800 ± 0.050 | 5 (59) | 1.874 ± 0.128 | 11 (49) | 2.145 ± 0.155 | | |
| 5 (62) | 2.107 ± 0.080 | | | 6 (77) | 1.780 ± 0.109 | | | | |
| 6 (137) | 1.058 ± 0.080 | | | | | | | | |
| 7 (14) | 1.903 ± 0.177 | | | | | | | | |

Multifactor analysis of variance (ANOVA) of lSCC including all the fixed effects (herd, production year, lactation number, month of measurements and stage of lactation) revealed statistical significance of only herd effect ($p<0.001$). This factor alone explained 37.3% of the variance of the trait (results not shown). No interaction term was found to be statistically significant.

### *2.3. Marker Association Analysis*

A multi-locus mixed linear (MLMM) model [11] was used to select significant SNPs as fixed (additive) covariates. This method employs a stepwise mixed-model regression procedure with forward selection and backward elimination. From the three model selection criteria that are implemented in MLMM for multi-testing correction, the modified Bayesian information criterion (mBIC), the extended Bayesian information criterion (eBIC) and the multiple Bonferroni criterion (mBonf), we used the eBIC criterion to select the significant SNP co-factors using a $p$ threshold value of 0.10. Specifically, lSCC data were analysed using the following mixed model:

$$y = X\beta + wa + Zu + e$$

where $y$ is the vector of the lSCC, $X$ is the incidence matrix relating observations to fixed effects, $\beta$ is the vector of fixed (environmental) effects: herd (7 classes), lactation number (5 classes: 2, 3, 4, 5, 6), production year (4 classes: 2011, 2012, 2013 and 2014), month of measurement (4 classes: 1, 8 10 and 11) and stage of lactation (3 classes: early, middle and late). Note that although the additional fixed effects (apart from herd) were found not to be statistically significant during multifactor ANOVA, an expanding fixed model was used during association analysis as these effects jointly explained an additional proportion of 3% of the lSCC variance. Furthermore, w is the vector of the SNP effects with elements coded as 0, 1 or 2 for homozygote of the reference allele, heterozygote and homozygote of the other allele, $\alpha$ is the vector of the fixed effect for the reference allele of the candidate SNP to be tested for association, $Z$ is the incidence matrix relating observations to the random polygenic random effects, $u$ is the vector of random polygenic effects, and $e$ is the vector of random residuals. The random effects were assumed to be normally distributed with zero means and the following covariance structure:

$$\mathrm{Var}\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G\sigma_u^2 & 0 \\ 0 & I\sigma_e^2 \end{bmatrix}$$

where $\sigma_u^2$ and $\sigma_e^2$ are the polygenic and error variance components, $I$ is the nxn identity matrix, and $G$ is the nxn genomic relationship matrix with elements of pairwise relationship coefficient using all the 42,884 SNPs. The genomic relationship coefficient between two individuals $j$ and $k$, was estimated as follows:

$$\frac{1}{42{,}884}\sum_{i=1}^{42{,}884}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-2p_i)}$$

where $x_{ij}$ and $x_{ik}$ the numbers (0, 1 or 2) of the reference allele(s) for the $i_{th}$ SNP of the $j_{th}$ and $k_{th}$ individuals, respectively, and $p_i$ is the frequency of the reference allele. Note that inclusion of the genomic relationship matrix in the model has been shown to correct for possible population structure and stratification in the data [12]. This analysis was carried out with SNP and Variation Suite ver. 8.7.0 (Golden Helix, Inc. 2016).

### *2.4. Search for QTLs and Candidate Genes*

Since in this breed levels of linkage disequilibrium (LD) were higher than 0 between markers at genomic distances up to 1 Mb (results not shown), we searched within 1 Mb upstream and downstream each significant SNP for reported QTLs and positional candidate genes for the trait under study. The SheepQTLdb (release 36, August 22nd, 2018) and the latest sheep genome *Oar_v4.0*: http://www.ncbi.nlm.nih.gov/genome/?term=ovis+aries/ along with NCBI annotation release 102 of the sheep genome (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ovis_aries/102/) were used, respectively.

### *2.5. In Silico Gene Prioritization Analysis*

We performed *in silico* prioritization analysis (PA) of the positional candidate genes using the ToppGene portal (https://toppgene.cchmc.org/prioritization.jsp). PA was based on the functional similarity of the candidate genes to a training gene list including n=1,221 genes that have been associated with the term ‘immunity’. The InnateDB (https://www.innatedb.com/) that is a knowledgebase of genes, proteins, experimentally-verified interactions and signaling pathways involved in the innate immune response of humans, mice and bovines to microbial infection was mined to retrieve the relevant associated genes. The following semantic annotations were used during PA: GO: Molecular Function, GO: Biological Process, Human and Mouse Phenotype, Pathway, Interaction, Gene Family and Co-expression. From the initial number of training genes, a total number of n=1,120 genes were mapped and finally used for training during PA. Note that two candidate genes i.e. *FREM1* and *GPR33* were omitted from PA as they were included in the training gene list. Genes with overall $p$-values lower than 0.05 were considered as highly prioritized.

## 3. Results

### *3.1. Significant SNPs*

Figure 1 shows the Q-Q (Quantile-Quantile) plot of the expected and the observed $p$-values (on the $-\log_{10}$ scale) of the 42,884 SNPs. As Q-Q plot clearly shows, there is no evidence of any systematic bias due to population structure or analytical approach, a suggestion that was also supported by the estimated value for the genomic inflation factor ($\lambda$=1.051). The Q-Q plot along with the Manhattan plot depicted in Figure 2 also show that 4 SNPs depart from the expected probability indicating that they might be associated with the trait under study.

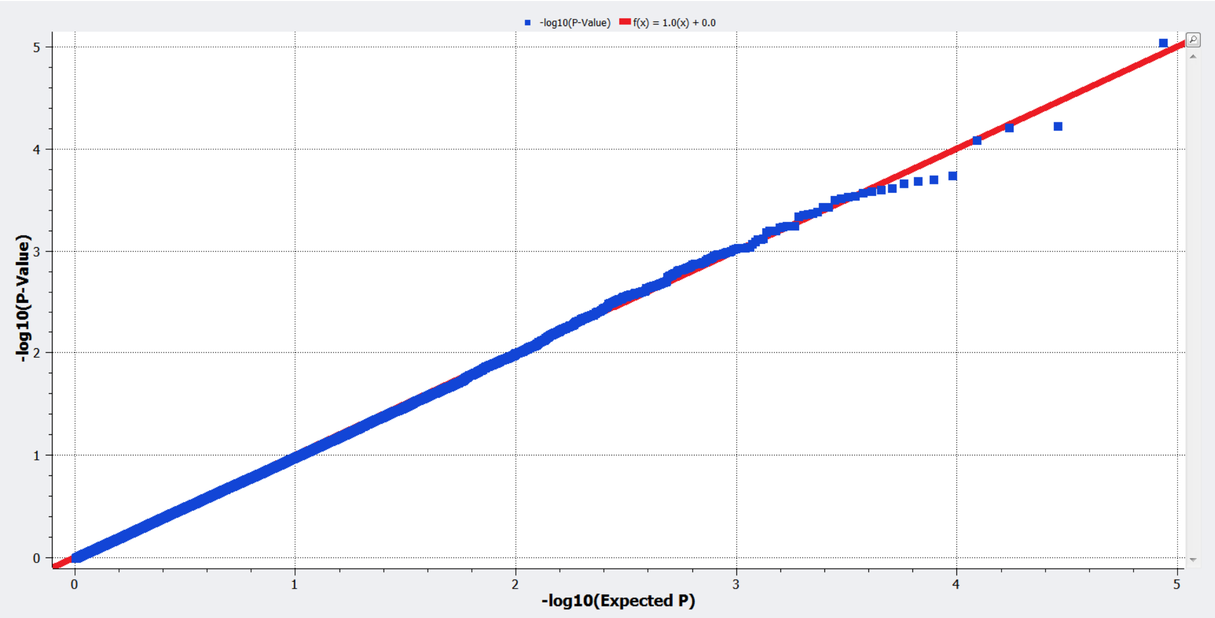


***Figure 1.*** *Quantile-Quantile (Q-Q) plot of the expected (x-axis) versus the observed (y-axis) p-values (-log$_{10}$ scale) of the SNPs.*

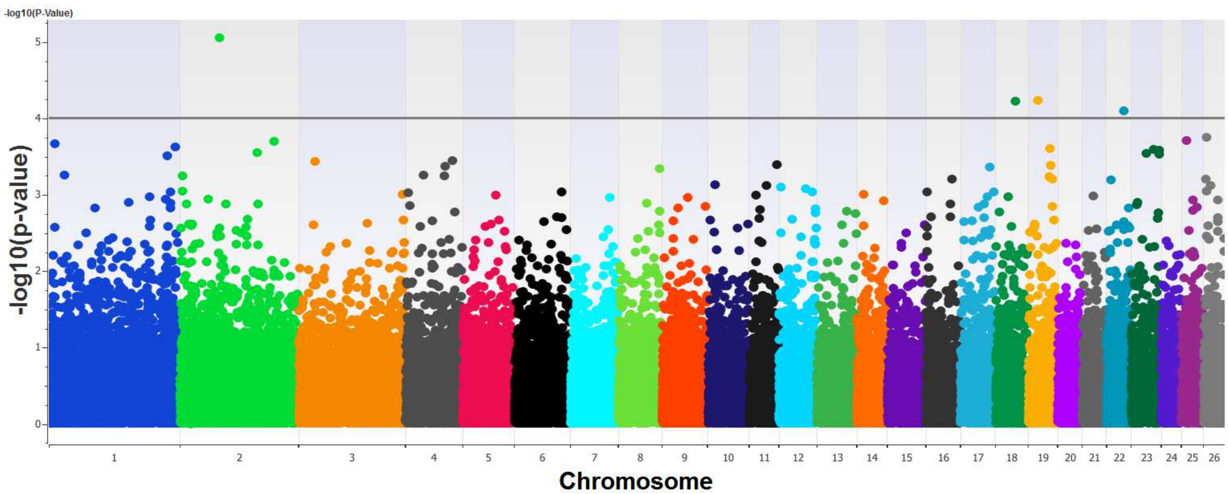


***Figure 2.*** *Manhattan plot depicting SNP associations with the SCC phenotype in the Frizarta dairy sheep. The plot shows the negative log-base-10 of the p value (y-axis) for each of the 42,884 SNPs across the 28 ovine autosomes (x-axis). Horizontal line denotes significance threshold (−log$_{10}$(p value)=4.094). The four SNPs, dispersed on OAR2, OAR18, OAR19 and OAR22 passing the chromosome-wise significance threshold are also shown.*

***Table 2.*** *Name, position on ovine chromosomes and p-values of chromosome-wide significant SNPs.*

| SNP | OAR | Position | *p*-value | -log$_{10}$ (*p*-value) |
|---|---|---|---|---|
| OAR2_87772629.1 | 2 | 82,595,501 | 8.973E-06 | 5.047 |
| OAR19_20722254.1 | 19 | 1,9801,023 | 5.924E-05 | 4.227 |
| OAR18_44175536.1 | 18 | 41,492,409 | 6.043E-05 | 4.219 |
| OAR22_41013052.1 | 22 | 36,223,181 | 8.060E-05 | 4.094 |

A detailed description of the 4 departed SNPs is provided in Table 2. Specifically, the 4 SNPs that reached chromosome-wise statistical significance i.e. −log$_{10}$(p value)=4.094 were detected on OAR2, OAR18, OAR19 and OAR22.

### 3.2. Searched QTLs and Positional Candidate Genes

A search for ‘Somatic Cell Score’ QTLs at SheepQTLdb revealed two QTLs, one on OAR2 mapped from 86.5 to 117.9 Mb and another on OAR22 mapped from 5.8 to 31.8 Mb (results not shown). In both cases, reported QTLs lie about 4 Mb away from the significant SNPs. Table 3 presents a list of positional candidate genes within 1 Mb distance from the 4 significant SNPs. A total number of 37 genes were identified within the searched regions with 8, 10, 2 and 17 candidate genes located on OAR2, OAR18, OAR19 and OAR22, respectively (Table 3).

***Table 3.*** *Positional candidate genes located within 1 Mb distances from significant SNPs.*

| Gene | Gene description | Gene location in *Oar_v4.0* | NCBI gene ID | Gene-SNP distance (bp) |
|---|---|---|---|---|
| *NFIB* | nuclear factor I/B | 2: 82,203,556..82,423,726 | 100913158 | 171,775 |
| *ZDHHC21* | zinc finger, DHHC-type containing 21 | 2: 82,728,390..82,793,425 | 101121148 | 132,889 |
| *CER1* | cerberus 1, DAN family BMP antagonist | 2: 82,810,222..82,815,069 | 101122006 | 214,721 |
| *FREM1* | FRAS1 related extracellular matrix 1 | 2: 82,833,553..83,016,760 | 101121399 | 238,052 |
| *TTC39B* | tetratricopeptide repeat domain 39B | 2: 83,148,036..83,303,125 | 101121659 | 552,535 |
| *SNAPC3* | small nuclear RNA activating complex polypeptide 3 | 2: 83,384,632..83,421,591 | 101121912 | 789,131 |
| *PSIP1* | PC4 and SFRS1 interacting protein 1 | 2: 83,426,169..83,462,898 | 100233239 | 830,668 |
| *CCDC171* | coiled-coil domain containing 171 | 2: 83,486,297..83,832,243 | 101122165 | 890,796 |
| *STRN3* | striatin 3 | 18: 40,416,405..40,520,814 | 101102517 | 971,595 |
| *AP4S1* | adaptor-related protein complex 4, sigma 1 subunit | 18: 40,520,132..40,567,008 | 101115405 | 925,401 |
| *HECTD1* | HECT domain containing E3 ubiquitin protein ligase 1 | 18: 40,568,259..40,641,561 | 100135435 | 850,848 |
| *HEATR5A* | HEAT repeat containing 5A | 18: 40,701,525..40,800,708 | 101115909 | 691,701 |
| *DTD2* | D-tyrosyl-tRNA deacylase 2 (putative) | 18: 40,826,489..40,836,811 | 101116351 | 655,598 |
| *GPR33* | G protein-coupled receptor 33 | 18: 40,852,409..40,853,736 | 101102762 | 638,673 |
| *NUBPL* | nucleotide binding protein-like | 18: 40,920,031..41,233,661 | 101116606 | 258,748 |
| *ARHGAP5* | Rho GTPase activating protein 5 | 18: 41,366,614..41,440,797 | 101116866 | 51,612 |
| *AKAP6* | A-kinase anchoring protein 6 | 18: 41,574,089..42,073,462 | 101117111 | 81,680 |
| *NPAS3* | neuronal PAS domain protein 3 | 18: 42,307,737..43,138,131 | 101103016 | 815,328 |
| *GRM7* | glutamate receptor, metabotropic 7 | 19: 18,525,586..19,471,907 | 443520 | 329,116 |
| *LOC105603423* | epidermal growth factor-like protein 6, human EGFL6 | 19: 19,846,286..19,851,084 | 105603423 | 45,263 |
| *ATRNL1* | attractin like 1 | 22: 34,467,463..35,279,716 | 101119865 | 943,465 |
| *GFRA1* | GDNF family receptor alpha 1 | 22: 35,379,304..35,612,886 | 101120125 | 610,295 |
| *CCDC172* | coiled-coil domain containing 172 | 22: 35,668,110..35,725,843 | 101103108 | 497,338 |
| *PNLIPRP3* | pancreatic lipase-related protein 3 | 22: 35,778,132..35,820,846 | 101103362 | 402,335 |
| *PNLIP* | pancreatic lipase | 22: 35,864,081..35,882,662 | 101103610 | 340,519 |
| *LOC101120372* | Inactive pancreatic lipase-related protein 1 | 22: 35,890,049..35,905,240 | 101120372 | 317,941 |
| *PNLIPRP2* | pancreatic lipase-related protein 2 | 22: 35,916,056..35,935,295 | 101103860 | 287,886 |
| *C22H10orf82* | chromosome 22 open reading frame, human C10orf82 | 22: 35,960,075..35,970,248 | 101120621 | 252,933 |
| *HSPA12A* | heat shock protein family A (Hsp70) member 12A | 22: 35,975,685..36,027,950 | 101120879 | 195,231 |
| *ENO4* | enolase family member 4 | 22: 36,149,792..36,177,110 | 101104109 | 46,071 |
| *SHTN1* | shootin 1 | 22: 36,179,953..36,293,552 | 101104358 | 0 |
| *VAX1* | ventral anterior homeobox 1 | 22: 36,384,969..36,389,063 | 101121136 | 161,788 |
| *KCNK18* | potassium channel, two pore domain subfamily K, member 18 | 22: 36,439,825..36,451,077 | 101121389 | 216,644 |
| *SLC18A2* | solute carrier family 18, member 2 | 22: 36,475,616..36,513,970 | 101104796 | 252,435 |
| *PDZD8* | PDZ domain containing 8 | 22: 36,516,506..36,599,063 | 101121648 | 293,325 |
| *EMX2* | empty spiracles homeobox 2 | 22: 36,738,845..36,744,354 | 101105042 | 515,664 |
| *RAB11FIP2* | RAB11 family interacting protein 2 (class I) | 22: 37,177,703..37,220,788 | 101105457 | 954,522 |

### *3.3. Prioritized Candidate Genes*

A total number of 13 genes (*NFIB, GFRA1, PSIP1, ARHGAP5, GRM7, HECTD1, SLC18A2, SHTN1, VAX1, CER1, STRN3, NPAS3* and *EMX2)* were highly prioritized (overall *p*-value<0.05) according to the semantic annotations applied during PA (Table 4). The aforementioned genes along with *FREM1* (OAR2) and *GPR33 (*OAR18) that were among the training genes, were considered as most plausible functional candidates for the trait under study.

***Table 4.*** *Ranked gene list according to prioritization analysis. Genes with overall p-value<0.05 are considered as highly prioritized.*

| Rank | Gene | Average score | Overall *p*-value |
|---|---|---|---|
| 1 | *NFIB* | 0.920 | 0.010 |
| 2 | *GFRA1* | 0.956 | 0.011 |
| 3 | *PSIP1* | 0.846 | 0.011 |
| 4 | *ARHGAP5* | 0.855 | 0.012 |
| 5 | *GRM7* | 0.811 | 0.014 |
| 6 | *HECTD1* | 0.767 | 0.018 |
| 7 | *SLC18A2* | 0.808 | 0.019 |
| 8 | *SHTN1* | 0.760 | 0.029 |
| 9 | *VAX1* | 0.814 | 0.034 |
| 10 | *CER1* | 0.804 | 0.036 |
| 11 | *STRN3* | 0.818 | 0.037 |
| 12 | *NPAS3* | 0.733 | 0.040 |
| 13 | *EMX2* | 0.750 | 0.043 |
| 14 | *ATRNL1* | 0.574 | 0.050 |
| 15 | *ZDHHC21* | 0.683 | 0.051 |
| 16 | *RAB11FIP2* | 0.697 | 0.051 |
| 17 | *NUBPL* | 0.675 | 0.052 |
| 18 | *EGFL6* | 0.724 | 0.053 |
| 19 | *PNLIP* | 0.640 | 0.054 |
| 20 | *DTD2* | 0.672 | 0.055 |
| 21 | *AKAP6* | 0.708 | 0.061 |
| 22 | *AP4S1* | 0.689 | 0.081 |
| 23 | *ENO4* | 0.913 | 0.085 |
| 24 | *HSPA12A* | 0.569 | 0.106 |
| 25 | *PDZD8* | 0.569 | 0.161 |
| 26 | *PNLIPRP1* | 0.527 | 0.162 |
| 27 | *TTC39B* | 0.403 | 0.177 |
| 28 | *PNLIPRP2* | 0.489 | 0.185 |
| 29 | *SNAPC3* | 0.461 | 0.203 |

| Rank | Gene | Average score | Overall *p*-value |
|---|---|---|---|
| 30 | *CCDC171* | 0.443 | 0.226 |
| 31 | *KCNK18* | 0.389 | 0.245 |
| 32 | *CCDC172* | 0.378 | 0.265 |
| 33 | *PNLIPRP3* | 0.375 | 0.310 |
| 34 | *HEATR5A* | 0.456 | 0.326 |
| 35 | *C10orf82* | 0.248 | 0.551 |

## 4. Discussion

GWAS are powerful in determining genomic regions associated with a trait. Nevertheless, these regions often contain tens or hundreds of positional candidate genes and experimentally identifying the true causal genetic variants requires considerable costs, effort and time. One of the most intriguing challenge is thus as how to narrow down the candidates list and pinpoint the most plausible genetic variants for the trait under investigation. To address this challenge, in the present study, we applied *in silico* PA of the positional candidate genes and ended up with a total of 13 highly prioritized (top) genes. A thorough search of the respective literature with regard to the biological function(s) of the top prioritized genes followed. This search showed that 7 prioritized genes i.e. *NFIB*, *GFRA1, PSIP1, ARHGAP5, HECTD1, STRN3* and *EMX2* have documented involvement in immune-related processes. Specifically, *NFIB (*Nuclear Factor I/B, ranked $1^{st}$ in PA) is a member of the nuclear factor I family of proteins; the latter are known to be involved in viral and cellular transcription and specifically CD4 transcription [13]. *GFRA1* (*GDNF* family receptor alpha 1, ranked $2^{nd}$ in PA) is connected to *GDNF* (Glial cell-derived neurotrophic factor) gene that exerts its effect on target cells by binding to *GDNF* family receptor-a. GDNFs are reported to be regulated by inflammatory cytokines [14]. *PSIP1* (PC4 and SFRS1 interacting protein 1, ranked $3^{rd}$) encodes the lens epithelium-derived growth factor p75 (LEDGF/p75), an important host co-factor that interacts with HIV-1 integrase to target integration of viral cDNA into active the outcome of HIV-1 infection genes [15]. *ARHGAP5 (*Rho GTPase activating protein 5, ranked $4^{th}$) participates in focal adhesion and specifically in leukocyte transendothelial migration, while *HECTD1* (HECT domain containing E3 ubiquitin protein ligase 1, ranked $6^{th}$) participates to focal adhesion and macrophage activation [16]. *STRN3* (striatin 3, ranked $11^{th}$) and *EMX2* (empty spiracles homeobox 2, ranked $13^{th}$) are members of the gene network of Wnt/b-catenin pathway that plays a critical role in cell differentiation, growth, proliferation, survival and immune cell function [17].

Apart from the aforementioned genes, the list with the most plausible candidate genes should also comprise genes *FREM1* and *GPR33* that they were found among the training genes, thus having evidenced involvement in immune-related processes. Specifically, *FREM1* (FRAS1 related extracellular matrix 1) is an extracellular protein with multiple annotated functional domains that interact with integrin, collagen, fibronectin, and interleukin 1 receptor (IL1R1) and influence transendothelial migration, epithelial integrity and inflammatory responses [18]. *GPR33* (G protein-coupled receptor 33) is an orphan member of the chemokine-like receptor family and is highly expressed in dendritic cells that provide a functional link between innate and acquired immunity and orchestrate the interplay between T- and B-lymphocytes [19].

We further explored the role of the current candidate genes with regard to immunity by constructing a network depicting human genes co-expressed in memory CD4 T-cells using information from the Immuno-Navigator database and the Network Analyst platform (https://www.networkanalyst.ca/). The resulting gene network is shown in Figure 3. As Figure 3 shows, this network is formed by 6 member genes including four of the candidate genes (*ARHGAP5*, *STRN3*, *HECTD1*, *PSIP1*) along with two connected genes (*STK38*, *SART3*).

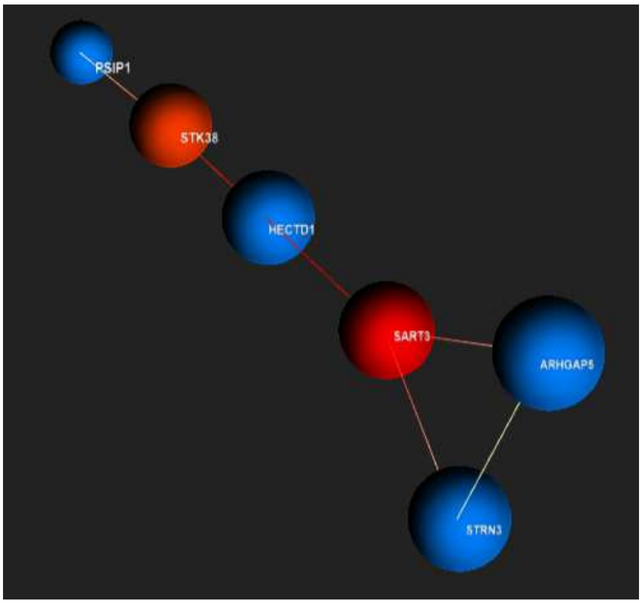


***Figure 3.*** *A minimum network showing 6 member genes co-expressed in human memory CD4 T cells. Apart from the four candidate genes (ARHGAP5, STRN3, HECTD1, PSIP1) denoted as blue coloured spheres, this network includes two more connected genes (STK38, SART3) denoted as red coloured spheres. Network was constructed using the NetworkAnalyst platform (ver. 3.0, https://www.networkanalyst.ca/) using data mined from the Immuno-Navigator database.*

Intuitively, genes that include or are in close proximity to the lead SNPs while having functional relevance with the trait under study are considered ideal functional candidates. However, proximity of a gene to the significant marker does not guarantee functional relevance and causative candidate genes may also exist among distantly located loci from the associated SNP in both qualitative [20] and quantitative traits [21]. In line with this scenario, the distance of the most plausible candidate genes from the respective significant SNPs ranged from 52 kb (*ARHGAP5*) to 972 (*STRN3*) kb, in the present study.

## 5. Conclusion

Employment of *in silico* prioritization analysis on results of genome wide associations helped identifying novel candidate genes for SCC with documented involvement in immune-related processes such as focal adhesion, transendothelial migration, macrophage activation and inflammatory responses. In light of the relatively small number of animals used here, further studies employing higher number of animals with higher density arrays are warranted to disentangle the genetic basis of the SCC phenotype in dairy sheep.

# Funding

This study was supported by the Greek Ministry of Education and Religious Affairs, Action 'Cooperation 2011' (grant number 447919/11SYN_3_1087).

# References

[1] Bergonier D, de Crémoux R, Rupp R, Lagriffoul G and Berthelot X. 2003. Mastitis of dairy small ruminants. Vet Res 34: 689-716.

[2] Barillet F, Rupp R, Mignon-Grasteu S, Astruc JM and Jacquin M. 2001. Genetic analysis for mastitis resistance and milk somatic cell score in French Lacaune dairy sheep. Genet Sel Evol 33: 397–415.

[3] Paape MJ, Poutrel B, Contreras A, Marco JC and Capuco AV. 2001. Milk somatic cells and lactation in small ruminants. J Dairy Sci 84, Supplement, E237-E244.

[4] Persson-Waller K, Colditz IG and Seow HF. 1997 Accumulation of leucocytes and cytokines in the lactating ovine udder during mastitis due to Staphylococcus aureus and Escherichia coli. Res Vet Sci 62: 63-66.

[5] Gutiérrez-Gil B, El-Zarei MF, Bayón Y, Alvarez L, de la Fuente LF, San Primitivo F and Arranz JJ 2007. Short communication: detection of quantitative trait loci influencing somatic cell score in Spanish Churra sheep. J Dairy Sci 90: 422-426.

[6] Barillet F, Arranz JJ, Carta A, Jacquiet P, Stear M and Bishop S 2006. Final consolidated report of the European Union contract of acronym "genesheepsafety" (QTLK5-CT-2000-00656) p. 145.

[7] Raadsma HW, Jonas E, McGill D, Hobbs M, Lam MK and Thomson PC 2009. Mapping quantitative trait loci (QTL) in sheep. II. Meta-assembly and identification of novel QTL for milk production traits in sheep. Genet Sel Evol 41: 45.

[8] Rupp R, Senin P, Sarry J, Allain C, Tasca C, Ligat L, Portes D, Woloszyn F, Bouchez O, Tabouret G, Lebastard M, Caubet C, Foucras G and Tosser-Klopp G 2015. A point mutation in suppressor of cytokine Signalling 2 (Socs2) increases the susceptibility to inflammation of the mammary gland while associated with higher body weight and size and higher milk production in a sheep model. PLoS Genet 11, e1005629.

[9] Bonnefont CM, Toufeer M, Caubet C, Foulon E, Tasca C, Aurel MR, Bergonier D, Boullier S, Robert-Granié C, Foucras G and Rupp R 2011. Transcriptomic analysis of milk somatic cells in mastitis resistant and susceptible sheep upon challenge with Staphylococcus epidermidis and Staphylococcus aureus. BMC Genomics 12: 208.

[10] Banos G, Bramis G, Bush SJ, Clark EL, McCulloch MEB, Smith J, Schulze G, Arsenos G, Hume DA and Psifidi A 2017. The genomic architecture of mastitis resistance in dairy sheep. BMC Genomics 18, 624.

[11] Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M and 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature Genetics 44: 825-830.

[12] Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S and Buckler ES 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38: 203-208.

[13] Sheeter D, Du P, Rought S, Richman D and Corbeil J 2003. Surface CD4 expression modulated by a cellular factor induced by HIV type 1 infection. AIDS Res Hum Retroviruses 19: 117-123.

[14] Esseghir S, Todd SK, Hunt T, Poulsom R, Plaza-Menacho I, Reis-Filho JS and Isacke CM 2007. A role for glial cell derived neurotrophic factor induced expression by inflammatory cytokines and RET/GFR alpha 1 receptor up-regulation in breast cancer. Cancer Res 67: 11732-11741.

[15] Passaes CP, Cardoso CC, Caetano DG, Teixeira SL, Guimarães ML, Campos DP, Veloso VG, Babic DZ, Stevenson M, Moraes MO and Morgado MG 2014. Association of single nucleotide polymorphisms in the lens epithelium-derived growth factor (LEDGF/p75) with HIV-1 infection outcomes in Brazilian HIV-1+ individuals. PLoS One 9, e101780.

[16] Zhou Z, Jiang R, Yang X, Guo H, Fang S, Zhang Y, Cheng Y, Wang J, Yao H and Chao J 2018. circRNA mediates silica-induced macrophage activation via HECTD1/ZC3H12A-dependent ubiquitination. Theranostics 8: 575-592.

[17] Zhang Y, Cao G, Yuan QG, Li JH and Yang WB 2017. Empty spiracles homeobox 2 (EMX2) inhibits the invasion and tumorigenesis in colorectal cancer cells. Oncol Res 25: 537-544.

[18] Luo M, Sainsbury J, Tuff J, Lacap PA, Yuan XY, Hirbod T, Kimani J, Wachihi C, Ramdahin S, Bielawny T, Embree J, Broliden K, Ball TB and Plummer FA 2012. A genetic polymorphism of FREM1 is associated with resistance against HIV Infection in the Pumwani sex worker cohort. J Virol 86: 11899–11905.

[19] Bohnekamp J, Böselt I, Saalbach A, Tönjes A, Kovacs P, Biebermann H, Manvelyan HM, Polte T, Gasperikova D, Lkhagvasuren S, Baier L, Stumvoll M, Römpler H and Schöneberg T 2010. Involvement of the chemokine-like receptor GPR33 in innate immunity. Biochem Biophys Res Commun 396: 272–277.

[20] Brodie A, Azaria JR and Ofran Y 2016 How far from the SNP may the causative genes be? Nucleic Acids Res 44: 6046-6054.

[21] Kominakis A, Hager-Theodorides AL, Saridaki A, Zoidis E, Antonakos G and Tsiamis G 2017. Combined GWAS and 'guilt by association' based prioritization analysis identified functional candidate genes for body size in sheep. Genet Sel Evol 49, 41.