COMPARISON OF METHODS FOR ESTIMATING THE EFFECTS OF CASEIN SNPs ON MILK TRAITS IN NORWEGIAN GOATS

ESINAM NANCY AMUZU

Norwegian University of Life Sciences
Department of ANIMAL AND AQUACULTURE SCIENCE
Master Thesis 30 credits 2011

EUROPEAN MASTER OF SCIENCE IN ANIMAL BREEDING AND GENETICS

Comparison of Methods for Estimating the Effects of Casein SNPs on Milk Traits in Norwegian Goats

ESINAM NANCY YAWO AMUZU

Main supervisor:
Associate Prof. Tormod Ådnøy
Department of Animal and Aquaculture Sciences
Norwegian University of Life Sciences
Ås, Norway

Co-supervisor:
Prof. Johann Sölkner
Department of Sustainable Agricultural Systems
Section of Livestock Sciences
University of Natural Resources and Applied Life Sciences
Vienna

DECLARATION

I hereby declare that this thesis is a bona fide record of research work done by me as a part of my Double Degree Program from the Norwegian University of Life Sciences (UMB), Ås, Norway and the University of Natural Resources and Applied life Sciences, (BOKU), Vienna, Austria.

It has not previously formed the basis for the award to me of any degree, diploma, fellowship or other similar title of any other university or society.

I hereby warrant that the thesis is based on work done by myself jointly with others; I have clearly stated exactly what was done by others and what I have contributed myself.

Ås, May 2011 E.N.Y Amuzu

……...

PREFACE

The submission of this master thesis marks the end of my 2 year MSc. program in Animal Breeding and Genetics. The study was carried out at the Department of Animal and Aquaculture Studies, Norwegian University of Life Sciences, with joint supervision from the University of Natural Resources and Applied Life Sciences, Vienna.

Single Nucleotide Polymorphisms (SNPs) in the casein genes have been the focus of many past research works all over the world in cattle, sheep and goat populations. This is due to the marked effect that they have on milk composition and sensory traits. Also, advances in genotyping methods have led to the availability of high quality data, which if utilized properly can lead to (and in some cases is already leading to) accelerated genetic improvement in livestock species. I feel statistical and data modeling methods for making logical and applicable inferences from the molecular data produced are the key to achieving this desired genetic improvement.

This study is a comparison of multivariate methods for estimating the additive effects of 38 casein SNPs in Norwegian goats. The main methods were based on principal components analysis and partial least squares regression. Another method which incorporated information on the extent of linkage disequilibrium into the mixed model equations was tested, but unfortunately excluded, from this final report because the plausibility of the results could not be assessed.

Differing from most of the other studies on the casein genes, this analysis was performed at the multi-SNP level rather than using the haplotype or single-SNP approach. The feasibility of the methods in estimating SNP effects is discussed.

E.N.Y Amuzu Ås, May 2011

ACKNOWLEDGEMENTS

To my main supervisor Associate Prof. Tormod Ådnøy: I really appreciate your willingness to accept me as your thesis student, and for the guidance and comments that lead to the successful completion of this work.

Warm thanks to my co-supervisor Prof. Johan Sölkner for his suggestions, helpful conference calls and prompt replies to my emails. I am very grateful.

Binyam S. Dagnachew, I truly appreciate all the guidance, assistance and encouragement I received right from the start of this research. You were patient and always had a smiling face to greet me whenever I called on you ☺.

To my friends at UMB: thanks for keeping my stress level down with your fun and laughter, it meant a lot to me.

To my wonderful family and Samuel, I think this is the best way to express my appreciation:
"I would thank you from the bottom of my heart but for you my heart has no bottom"
Even though you are miles away I felt your presence and support each and every day. I wouldn't have made it without you. I love you!

Thanks to the Erasmus Mundus Scholarship Program for making this MSc. study possible.

My ultimate thanks go to God; He is the reason for my existence.

ABSTRACT

The four casein proteins make up the majority of protein in goat's milk. They are encoded by 4 closely linked genes, CSN1S1, CSN2, CSN1S2 and CSN3 within a 250 kb segment on chromosome 6. Polymorphisms of these genes are of interest to animal breeders due to their effect on milk composition and quality. Due to the linkage between them, methods for analysis of this genomic region have mostly been at the haplotype level. A multi-SNP approach was used in this study. We assessed the suitability of two multivariate statistical methods, partial least squares and principal component regression, for the detection of the additive effects of casein polymorphisms on milk traits. These methods are well suited for analysis of collinear variables. Genotype information on 38 casein SNPs, and phenotypic records on milk yield, somatic cell count, fat, protein, and lactose percentages were obtained for 565 goats from 6 Norwegian farms. Three models were compared. After correcting the records for fixed and permanent environment effects, PLSR was run on single traits at a time (model 1) and then jointly for all traits but milk yield (model 2). For the third model the scores from PCA were collected and used as fixed effects in an animal model. The PLS-based methods clearly detected significant effects of SNPs in the CSN1S1 and CSN3 regions, consistent with previous findings. Three SNPs in the CSN2 gene had positive effects on fat and protein percent and negative effects on somatic cell count. A Norwegian-specific deletion in exon 12 of CSN1S1 had a significant negative effect on fat and protein percent ($p<0.05$). Estimates from Model 3 generally had higher SEs, and only identified significant effects on fat and milk yield. It was however able to detect the effect of the exon 12 deletion on fat percent. Overall, the PLS based models identified a higher number of effects as significant, fat and protein percent were better explained by the models than the other traits, and SNPs at CSN1S1 and CSN3 appear to be the most important for detecting variability in milk traits.

Key words: Casein genes, Principal component analysis, Partial least squares, Norwegian dairy goats.

TABLE OF CONTENTS

DECLARATION II

PREFACE III

ACKNOWLEDGEMENTS IV

Abstract V

TABLE OF CONTENTS VI

List of tables IX

List of figures IX

List of tables in the appendix X

Abbreviations XI

1.0 INTRODUCTION 1

1.1 Background 1

1.2 Aim of the study 4

2.0 LITERATURE REVIEW 5

2.1 Caprine casein genes and their effects on milk traits 5

2.1.1 Alpha s1 casein 5

2.1.2 Alpha s2 and Beta caseins 6

2.1.3 Kappa casein 6

2.2 Linkage disequilibrium between the casein genes 6

2.3 Multicollinearity in Regression models 8

2.4 Principal Component Analysis 9

1.5 Partial Least Squares Regression 10

3.0 MATERIALS AND METHODS 12
3.1 Materials 12
3.1.1 Genotypic data 12
3.1.2 Phenotypic records 12
3.1.3 Pedigree records 13
3.1.4 Variance components 13
3.2 Methods 14
3.2.1 Data organization 14
3.2.2 Estimation of SNP effects 14
3.2.2.1 Model 1: Partial Least Squares Regression I 14
3.2.2.2 Model 2: Partial least squares regression II 17
3.2.2.3 Model 3: PCA combined with Animal model 17
4.0 RESULTS AND DISCUSSION 20
4.1 Results 20
4.1.1 General Descriptive statistics 20
4.1.2 Estimated SNP effects 20
4.1.2.1 Model 1: Partial Least Squares Regression (PLS I) 20
4.2.2 Model 2: Partial least squares regression II (PLS II) 29
4.1.2.3 Model 3: PCA combined with Animal model 36
4.2 General Discussion 45
5.0 CONCLUSION 47

REFERENCES 48

APPENDIX I

LIST OF TABLES

Table 1. Variance components used for analysis ……13

Table 2: Descriptive statistics of traits included in the study ……20

Table 3: Summary of statistically significant SNP effects across all models……44

LIST OF FIGURES

Figure 1: LD across the casein chromosome segment……....8

Figure 2: Significance test for the additive effect of major allele of each SNP on Fat percent, estimated with Model 1……..22

Figure 3: Significance test for the additive effect of major allele of each SNP on Lactose percent, estimated with Model 1……23

Figure 4: Significance test for the additive effect of major allele of each SNP on log-transformed Somatic cell count, estimated with Model 1…….24

Figure 5: Significance test for the additive effect of major allele of each SNP on Protein percent, estimated with Model 1 ……25

Figure 6: Significance test for the additive effect of major allele of each SNP on Milk yield estimated with Model 1……26

Figures 7, 8, 9: Additive effect of major allele of each SNP on milk composition traits, estimated with Model 1……27, 28

Figure 10: Significance test for the additive effect of major allele of each SNP on log-transformed somatic cell count, estimated using Model 2……30

Figure 11: Significance test for the additive effect of major allele of each SNP on Lactose percent, estimated with Model 2……31

Figure 12: Significance test for the additive effect of major allele of each SNP on Fat percent, estimated with Model 2……32

Figure 13: Significance test for the additive effect of major allele of each SNP on Protein percent, estimated with Model 2…….33

Figures 14, 15: Additive effect of major allele of each SNP on milk composition traits, estimated with Model 2……..……34, 35

Figure 16: Significance test for the additive effect of major allele of each SNP on Fat percent, estimated with Model 3……37

Figure 17: Significance test for the additive effect of major allele of each SNP on Lactose percent, estimated with Model 3…….38

Figure 18: Significance test for the additive effect of major allele of each SNP on log-transformed Somatic cell count, estimated with Model 3…….39

Figure 19: Significance test for the additive effect of major allele of each SNP Protein percent, estimated with Model 3……40

Figure 20: Significance test for the additive effect of major allele of each SNP on Milk yield, estimated with Model 3……..41

Figures 21 and 22: Additive effect of major allele of each SNP on milk composition traits, estimated with Model 3……...42, 43

LIST OF TABLES IN THE APPENDIX

Table 1: Coding of the 38 casein SNPs……..I

Table 2a: Additive effect of major allele of each SNP on milk traits Estimated using Model 1……II

Table 2b: Additive effect of major allele of each SNP on milk traits Estimated using Model 1……...III

Table 3a: Additive effect of major allele of each SNP on milk traits
Estimated using Model 2……IV

Table 3b: Additive effect of major allele of each SNP on milk traits
Estimated using Model2……..V

Table 4a: Additive effect of major allele of each SNP on milk traits
Estimated using Model3…….VI

Table 4b: Additive effect of major allele of each SNP on milk traits
Estimated using Model3……VII

ABBREVIATIONS

CSN1S1: casein *alpha S1* locus

CSN2: casein *beta* locus

CSN1S2: casein *alpha S2* locus

CSN3: casein *kappa* locus

GAS: Gene assisted selection

HAS: Haplotype assisted selection

kb: kilo base pair

kg: kilo gram

LD: Linkage disequilibrium

LV(s):Latent variable(s)

OLS: Ordinary least squares

PCA: Principal component analysis

PC(s): Principal Components(s)

PCR: Principal Component Regression

PLS: Partial Least Squares Regression

RMSECV: Root Mean Square error of cross validation

SE: Standard error

SNP: Single Nucleotide Polymorphism

SSC: Somatic cell count

1.0 INTRODUCTION

1.1 Background

The main source of income for goat farmers in Norway is the sale of milk and milk products. This is also true for many other European countries like France, Spain, Portugal and Italy. The main use of this milk is the production of cheese, and since coagulation of caseins is the fundamental process in cheese making, there has been an immense amount of interest in the study of casein genes in goat populations across Europe over recent years.

There are four different types of casein proteins, together making up the majority of protein in goat's milk (Hayes et al., 2006). These four casein genes have been mapped on chromosome 6 in both cattle and goats (Hayes et al., 1993; Threadgill et al., 1990), with those of the caprine chromosome located within a 250 kb segment. The four genes are in the order α_{S1}-, β, α_{S2}-, and κ-casein, and are coded by the loci CSN1S1, CSN2, CSN1S2 and CSN3 respectively. Many studies have been carried out on polymorphisms in the casein genes, most of which have been done at the haplotype level due to the tight linkage between the 4 genes.

Different haplotyping methods were used in these studies: Maximum likelihood (Excoffier & Slatkin, 1995; Hawley & Kidd, 1995) and a parsimony method by Clark (1990) seem to be the most common. Stephens et al., (2001) also proposed a method using Gibbs sampling to reconstruct haplotypes from SNP data. A well-established haplotyping software, PHASE, is also commonly used for the same purpose.

The effects of the haplotypes are then estimated with varying models. Hayes et al., (2006) reported that haplotypes at CSN1S1 loci had significant effects on protein percent, fat percent and fat yield, whilst those at the CSN3 loci had significant effects on protein and fat

percentages. Similar findings on associations between casein polymorphisms and milk composition and sensory traits have been reported in Italian breeds by Pizzillo, (1996); Marletta, (2000); Meggiolaro, (2000) and in Spanish breeds by Díaz, (1993 & 1994); Angulo, (1996) and many others.

These interesting findings have raised the question of how best to include this genomic information in breeding programmes for dairy goats. Serradilla (2003) compared 3 different strategies for including CSN1S1 information to improve selection response; he found a significant increase in the rate of genetic gain when genotypic information was included in selection for protein content and concluded that further studies are necessary to optimise the use of casein gene information in breeding programmes. Hayes et al., (2006) also conclude that there is a great potential for using the casein genes in haplotype assisted selection with respect to improvement of milk quality.

Several methods have been suggested for the inclusion of genotypic information into predictors of the genetic merit of candidates for selection, one of them being multiple regression. Certain authors, e.g Clayton et al., (2004); Chapman et al., (2003) believe that genotype-based tests can outperform haplotype-based approaches. In one method, Chapman et al., (2003) used a subset of tagSNPs as regressors to test associations.

Multivariate statistical methods might be an alternative to haplotype-based methods of including genotype information in selection. They are increasingly being applied to the analysis of SNP data in recent times, the most used methods being principal component regression (PCR), partial least squares regression (PLSR) (Long et al.,2011) and other variations which are combinations, or slight alterations, of PCR and PLS. An example is seen in work by Bouveresse &Rutledge, 2009.

These methods have the advantage of eliminating multicollinearity, a problem which arises in regression models for quantitative traits when using a large number of genetic markers as predictor variables. Even in cases with relatively small numbers of markers, multicollinearity may occur because markers are intercorrelated due to linkage disequilibrium between the SNPs. Linkage disequilibrium (LD) is the non-random association between alleles at two or more loci.

Investigating the feasibility of using PCR and PLSR for the analysis of associations between casein SNPs and milk traits of economic importance is therefore interesting because these methods address both the issue of dimension reduction (if necessary), and the problem of multicollinearity among predictor variables. Also, inferring haplotypes and then carrying out a haplotype-based analysis may be more inefficient than direct SNP analysis (Morris et al., 2004). Using PCA/PLSR skips this step of haplotype inference.

It has been noted by Hoggart et al.,(2008) and Long et al., (2011) that SNPs selected by single-SNP analysis may produce more false positives than those selected by multiple-SNP analysis, because the signal at a SNP when analyzed individually is often weakened by the inclusion of other correlated SNPs. Another possible advantage of the multivariate methods is therefore that these methods will allow for the simultaneous analysis of all available genotyped SNPs, instead of performing multiple single-SNP analyses.

Genotype information on 38 SNPs from the goat casein genes have previously been analysed in the Norwegian goat population by Hayes el al (2006). They used a haplotype-based approach and found associations between about 6 haplotypes and certain milk traits.

1.2 Aim of the study

The general aim of this study is to test the feasibility of multivariate techniques as an alternative to haplotyping and single-SNP methods for the detection of associations between the casein polymorphisms and milk composition traits.

Due to the high level of linkage disequilibrium between the SNPs, the main focus will be on ways of dealing with multicollinearity between the SNPs when performing simultaneous multi-SNP analysis. This will be investigated by:

- Running partial least squares regression of SNPs on milk records (milk yield, fat percent, protein percent, somatic cell count and lactose content)
- Testing a model combining principal component analysis (PCA) and the BLUP animal model to estimate additive effects of the SNPs on milk traits.

2.0 LITERATURE REVIEW

2.1 Caprine casein genes and their effects on milk traits

2.1.1 Alpha s1 casein

The four casein genes α_{s1}-, α_{s2}-, β- and κ- casein have been described as the major milk protein genetic polymorphisms in goats (Moioli et al., 2006). Of these, α_{s1} is the most polymorphic with about 18 known co-dominant alleles (Grosclaude & Martin, 1997; Chianese et al., 1997; Bevilacqual et al., 2002; Ranummo et al., 2001 & Caroli et al., 2007). These variants are associated with different rates of protein synthesis and α_{s1}-casein content of milk (Grosclaude & Martin, 1997; Moioli et al., 2006). Effects of α_{s1} on goat flavour of cheese and fat content have also been reported (Barbier et al., 1995).

Due to its high level of polymorphism and clear distinction in levels of protein synthesized between different alleles, α_{s1} seems to be the most studied casein gene (Grosclaude et al., 1994). Studies on the effects of α_{s1} on milk yield and composition, micelle structure, renneting properties and cheese yield in French breeds are well summarized by Serradilla (2003).

Hayes et al., (2006) analysed effects of casein haplotypes on milk production traits in the Norwegian dairy goat population and came to the following conclusions: (1) CSN1S1 haplotypes had significant effects on protein percent, fat percent and fat kg, (2) CSN3 haplotypes significantly affected protein and fat percent, (3) at the individual SNP level only 2 SNPs with an effect on protein and lactose percent were significant, (4) a Norwegian-specific deletion in exon 12 of CSN1S1, found to have a very high frequency, 0.86 , (Adnoy et al., 2003), explains the effect of the CSN1S1 haplotype on fat kg, (5) there was no significant effect of the interaction of haplotypes at CSN1S1 and CSN3. The last finding is supported by reports of Caravaca et al., (2011) on a similar analysis of French breeds.

2.1.2 Alpha s2 and Beta caseins

The α_{s2} and β- casein genes are relatively less polymorphic. Independent research by Boulanger et al., (1984); Bouniol et al., (1994), Martin & Addeo, (1995); Lagonigno et al., (2001), Ramunno et al., (2001) and Erhardt et al., (2002) collectively identify 8 α_{s2} alleles reported to have effects on the synthesis levels of α_{s2} protein and ultimately on the allergenic properties of milk.

The β- casein gene has 3 variants associated with normal β- casein content and two null alleles which result in the absence of β- casein in milk (Mahe &Grosclaude, 1993; Neveu et al., 2002; Martin and Addeo, 1995).

2.1.3 Kappa casein

The κ- casein gene is also highly polymorphic, and has been studied in diverse populations in Africa, Asia, Europe and America. A large variation in the frequencies of the circa 16 polymorphisms - 13 of which are protein variants - are seen across these populations (Angiolillo et al., 2002; Yahyaoui et al., 2003; Jann et al., 2004), but the exact effect of these variants on milk production traits is not clearly stated. Hayes et al., (2006) however found suggestive effects of a cluster of SNPs in the promoter region of CSN3 on protein and fat percent in Norwegian goats. The authors believe that none of the SNPs detected so far are the causative mutation, but may be in LD with it. Caravaca et al., (2011) found a significant effect of CSN3 polymorphisms on rennet coagulation time, total casein and protein content in the Murciano-Granadina breed from Spain. They recommend that further studies on other breeds be carried out to replicate and validate their findings.

2.2 Linkage disequilibrium between the casein genes

First reported by Grosclaude et al., (1987), it is now common knowledge that the four casein genes are within a gene cluster. This region is 250 kb, located on chromosome 6 in both cattle

and goats (Hayes et al., 2006). For the Norwegian goat population, it was reported by Hayes et al., (2006) that LD was not evenly spread across the chromosome segment containing the caseins. They observed high levels of LD at either end of the segment, but low levels of LD in the middle of the segment.

The high level of LD between the four casein genes is seen as a hurdle in their analysis (Caroli et al., 2006) and is the reason that most of the association analyses so far reported have been at the haplotype level, so as to incorporate the information from all the genotyped SNPs simultaneously. These authors also feel that research focused on the haplotype level is necessary to detect important effects that could be used for the genetic improvement of goat breeds. Hayes et al., (2006) also support the idea of simultaneous analysis of all genotyped mutations.

In line with this thinking, this present study attempts to analyse all SNP information available on the entire casein cluster simultaneously through the application of multivariate methods PCA and PLS, as well as a BLUP model which incorporates information from PCA.

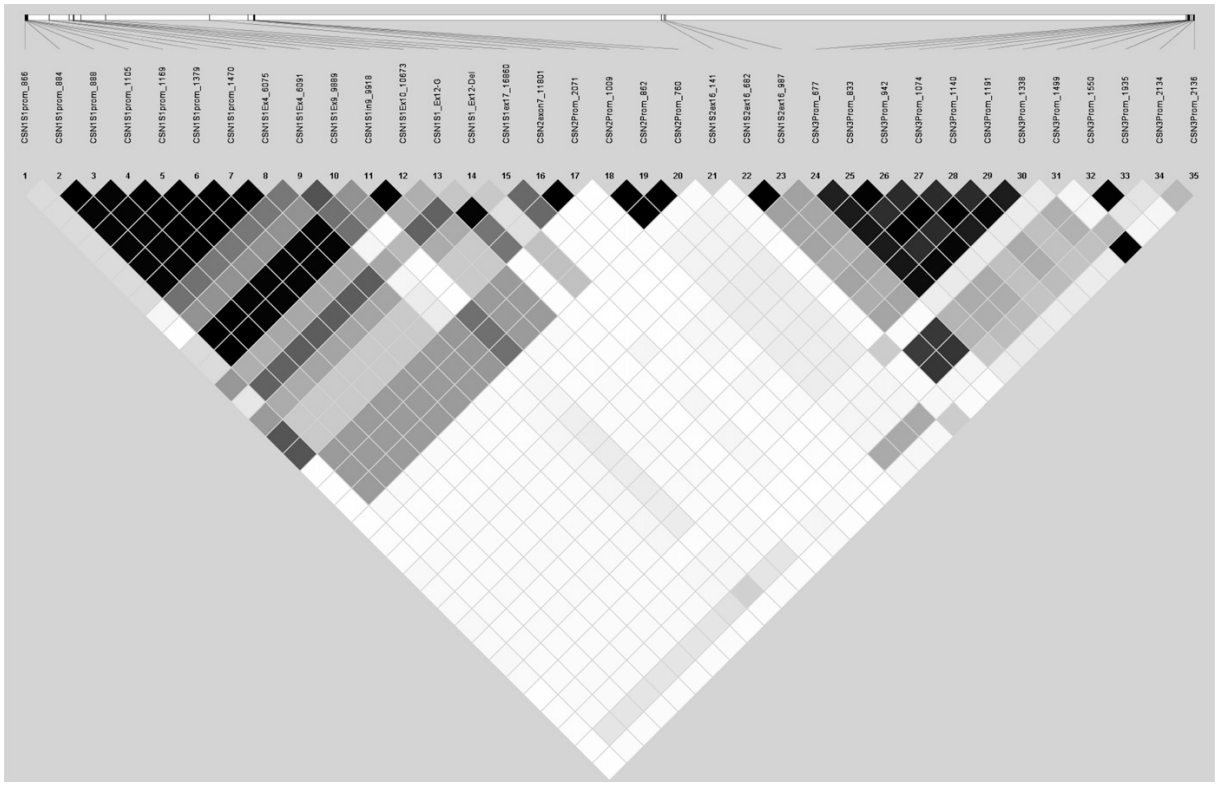


Figure 1.—LD across the chromosome segment visualized using the Haploview program (Barrett et al. 2005). Each diamond contains the level of LD measured by r^2 between the markers specified. Darker tones correspond to increasing levels of r^2. (Source: Hayes et al., 2006)

2.3 Multicollinearity in Regression models

Multicollinearity is a situation in statistical analysis where some of the predictor variables in a model are highly correlated or are perfect linear combinations of the other variables. In SNP data analysis, multicollinearity almost always exists because some of the SNPs, especially those in close proximity on a chromosome, are intercorrelated. This inter-correlation is attributed to linkage disequilibrium. As reiterated by Long et al., (2011), one consequence of multicollinearity in least-squares regression is unstable estimates, since the variance of the estimated regression coefficients will be greatly inflated. Chun and Keles (2009) state that multicollinearity is a common statistical problem that arises during regression-based modelling of modern biological data. This thus necessitates the investigation of the available methods of overcoming this, as well as perhaps the creation of new ones. Principal component

analysis and partial least squares regression have been used, albeit in different ways, by many researchers to overcome this problem.

2.4 Principal Component Analysis

Principal component analysis (PCA) is a procedure that applies mathematical algorithms to 'convert' a matrix of possibly correlated predictor variables into a set of orthogonal (uncorrelated) variables. These new variables are usually termed principal components (PCs). Each PC is a linear combination of all the initial variables. PCA decomposes a data matrix **X**, into orthogonal scores **T** and loadings **P**. (Mevik & Wehrens, 2007), giving the equation:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{t}} + \mathbf{E} \quad \text{(E is a matrix of residual errors)}$$

The PCs are ordered with respect to the amount of variance in X that they explain, with the first PC being the one with the highest proportion of explained variance. The loadings show the influence of the X variables on the scores, and are important for the interpretation of the results obtained from a PCA analysis. The scores give one of the most powerful tools that principal component-based methods can offer (Risvik, 2007), and further analysis is usually performed using scores. For example, in principal component regression, the ordinary least squares solution for:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

is given by $\widehat{\boldsymbol{\beta}} = \boldsymbol{P}(\boldsymbol{T}^{\boldsymbol{t}}\boldsymbol{T})^{-1}\boldsymbol{T}^{\boldsymbol{t}}\boldsymbol{Y}$

where the superscript '**t**' refers to the transpose of the respective matrix. One characteristic of PCA, which is considered a drawback, is that in calculation of the PCs, only the variance in X is taken into account. This may lead to suboptimal predictive power of the PCs (Mevik & Wehrens, 2007).

1.5 Partial Least Squares Regression

The methodology of partial least squares was introduced by the Swedish statistician Herman Wold. He termed it 'Projection to Latent Structures'. It is similar to PCA in that it aims at extracting from a set of *n* predictor variables, X, a set of $n<p$ orthogonal factors without losing too much of the initial variance that existed in the data (Abdi, 2010). These factors are commonly referred to as latent variables (LVs). It differs from PCA however because the LVs are chosen in such a way as to describe as much of the *covariance* between X and Y, whereas PCA concentrates on only the variance of X. The underlying models for PLS are:

$$\mathbf{X = TP^t + E}$$

$$\mathbf{Y = UQ^t + F}$$

The regression coefficients are obtained as:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{R}(\boldsymbol{T}^t\boldsymbol{T})^{-1}\,\boldsymbol{T}^t\boldsymbol{Y}$$

with $\boldsymbol{R = W(P^tW)^{-1}}$

Where **X** is a matrix of predictor variables, **Y** is a matrix of predictor variables, **T** is a matrix of X scores, **U** is a matrix of Y 'factors', **P** and **Q** are matrices of **X** and **Y** loadings respectively, **W** is a matrix of weights for **X, E** and **F** are the error terms. (Mevik & Wehrend, 2007)

PLS I is partial least squares regression on a single response variable and PLS II is the same analysis performed on multiple response variables simultaneously. Details of the PCA and PLSR algorithms and equations are not given here, as the main focus of this research is not comparing algorithms or computations of PCs and LVs, but rather on the possibilities and/or drawbacks of the use of these methods in general, for the estimation of casein SNP effects on milk traits.

In both PCA/R and PLSR one main decision is how many components to retain for further analysis. The most common criteria are the predictive ability of the model, measured by the root mean square error of prediction (RMSEP) for a test data set (or root mean square error of cross –validation (RMSECV) on the training data set), and the amount of variance in the explanatory variable that the model explains; measured by R^2. In PCR, the cumulative variance in the predictor variables that is explained by the components is usually used. There is however no hard and fast rule for this; the final decision is based on the main aim of the analysis and the researcher's opinion.

3.0 MATERIALS AND METHODS

3.1 Materials

3.1.1 Genotypic data

This study used data on 38 SNPs from the 4 caprine casein loci. The subjects were 605 does from 6 Norwegian farms. Collection of blood samples and genotyping was carried out through the combined effort of TINE SA (largest Norwegian dairy product cooperative), the Norwegian Association of Sheep and Goat Breeders (NSG), Norwegian University of Life Sciences, Norwegian Crop and Environmental Research Institute and the Centre for Integrative Genetics. Details of genotyping procedures can be found in Hayes et al., (2006). In the present report, the 38 casein polymorphisms are labelled as SNP1 – SNP40 according to the numbering used in Hayes et al., (2006); actual names are given in the appendix.

3.1.2 Phenotypic records

3127 records on milk production were available from 567 genotyped does, giving an average of 5.5 records per doe. On milk composition there were 2172 records for 565 genotyped does; an average of 3.8 records per doe. Both datasets gave a good representation of all 6 farms. The does included in the study all had kidding dates between 2004 and 2005 and production records from 2005 were used. The records taken into account were:

Milk kg: the total amount of milk produced per goat on the day of control (as a sum of morning and evening lactation)

Fat percent: the fat content of the sampled milk

Lactose percent: the lactose content of the sampled milk

Protein percent: the protein content of the sampled milk

Somatic cell count (SCC): the concentration of somatic cells per millilitre of milk

Days in milk (DIM): calculated as the number of days from the kidding date to the date of control.

Kidding information: kidding date and parity number

Herd – test day: factor which combines information on which farm and on what date records were taken

3.1.3 Pedigree records

From a database of 7323 animals, the lineage of the 605 genotyped does was traced up to 7 generations back. This information was used to compute the relationship matrix that was used in the mixed model equations.

3.1.4 Variance components

The additive, permanent environment and residual variance components for all the traits used in the study were obtained from the Norwegian Association of Sheep and Goat farmers. These variance components were calculated in 2009.

Table1. Variance components used for analysis

Variance component	TRAITS				
	Milk yield	Fat percent	Protein percent	Lactose percent	log(SCC)
Additive genetic	0.05324	0.13982	0.0149	0.01327	0.08109
Permanent environment	0.07099	0.06289	0.0073	0.00612	0.19491
Residual	0.15311	0.31173	0.01963	0.01592	0.51572

3.2 Methods

3.2.1 Data organization

General data sorting, removal of missing records, identification of lineage of the animals for the pedigree, coding, transformation of measurement scales and basic calculations (e.g. days in milk) were done using Microsoft Excel (2007).

An $n \times p$ SNP matrix was created using the following coding:

2 - homozygote for more frequent allele
1 - heterozygote
0 - homozygote for less frequent allele.

For SNP14 coding was 'deletion' = G**AAAAA**T versus
'non-deletion' G**AAGAAA**T and G**AAAAAA**T- The deletion was the more frequent allele

n is the number of records and *p* is the number of SNPs (38)

3.2.2 Estimation of SNP effects

The effect of each of the 38 SNPs on milk yield (kg), log-transformed somatic cell count, fat, lactose and protein percentages was estimated using 3 different models. All analyses were carried out using R statistical software.

3.2.2.1 Model 1: Partial Least Squares Regression I

Partial least squares regression was run using the Non-linear Iterative Partial Least Squares algorithm (NIPALS). The 38 SNPs were the predictor variables and the single traits were each responses in PLS I. The traits were pre-corrected for the fixed effects of kidding season, parity number, herd test day and days in milk (DIM). DIM was modeled as described by Jamrozik and Schaeffer (1997). To account for repeated records, the permanent environment effect of each individual was also predicted. Traits were corrected with the following model:

$$y_{ijklmn} = \mu + HTD_k + kdseason_l + KNUM_m + b_{1n}DIM + b_{2n}DIM + b_{3n}DIM + b_{4n}DIM + pe_j + e_{ijklmn}$$

where:

y_{ijklmn} is the i^{th} recorded trait for goat *j* with parity number *m* within kidding season *l*, days in milk *n* and permanent environment *j*, taken on herd test day *k*

μ is the fixed effect of the mean

HTD = fixed effect of herd test day, *k* (*k=1,2,3...,35 for milk yield and 1,2,3...,25 for milk composition)*

kdseason = fixed effect of the kidding season *l* (*l* =**1**:Dec-Feb, **2**:March to May, **3**:June to November)

KNUM= fixed effect of parity *m (m= 1, 2, 3 and 4 for unknown parity)*

DIM= effect of the stage of lactation *n where* $b_{1\text{-}4}$:

$$1 = \frac{DIM}{305}, \quad 2 = \left(\frac{DIM}{305}\right)^2, \quad 3 = ln\left(\frac{DIM}{305}\right), \quad 4 = \left[ln\left(\frac{DIM}{305}\right)\right]^2$$

pe = random effect of animal *j*'s permanent environment. *(j=1,2,…,567 for milk yield and 1,2,...,565 for milk composition).*

The model is represented in matrix notation as:

$$Y = X\beta + Zpe + \varepsilon$$

Assuming: $pe \approx N(0, I\sigma_{pe}^2)$, $\varepsilon \approx N(0, I\sigma_e^2)$, $G = \lambda_p I$ and $\lambda_p = \frac{\sigma_e^2}{\sigma_{pe}^2}$

$$E(y) = X\beta \quad and \quad V(y) = ZGZ^t + I\sigma_e^2$$

where:

X is a design matrix of all the fixed effects, *Z* is an incidence matrix relating phenotypes to individuals, *I* is an identity matrix, *pe* is a vector of the permanent environment effect of individuals, ε is the vector of residual errors associated with each observation, *G* is the covariance matrix of permanent environmental effects, σ_{pe}^2 and σ_e^2 are permanent environmental and residual variances respectively. The following equation was solved to get the estimates:

$$\begin{pmatrix} \widehat{\beta} \\ \widetilde{pe} \end{pmatrix} = \begin{bmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ + G \end{bmatrix}^{-1} \begin{bmatrix} X^ty \\ Z^ty \end{bmatrix}$$

Giving: $$\hat{Y} = X\widehat{\beta} + Z\widetilde{pe}$$

and the final 'corrected' Y values were:

$$^*Y = Y - \hat{Y}$$

A PLSR model was then used to analyze the *corrected milk traits:

$$^*Y = SNPs + residual$$

'SNPs' refers to the SNP matrix.

A different number of latent variables were used in the models for the different traits; the optimum model dimension was inferred from plots of the RMSECV.

- **Significance test**

  Student's t – distribution with degrees of freedom *n – number of fixed effects estimated* and level of significance 5% (α = 0.05) was used to test for significance of the regression coefficients.

  H_o: $\beta = 0$ versus H_a: $\beta \neq 0$ for all SNPs

  The test statistic was: $\frac{\widehat{\beta}}{\widehat{SE}_{\widehat{\beta}}}$

3.2.2.2 Model 2: Partial least squares regression II

Pre-correction of traits was done as in Model 1, and then PLSR was performed on all traits simultaneously, excluding milk yield (due to differences in the structure of the data). The test statistic was also computed as in Model 1.

3.2.2.3 Model 3: PCA combined with Animal model

The scores obtained from PCA (singular value decomposition) of the SNP matrix were used as fixed effects in an animal model. Only the scores from the first 15 PCs were used. This number was decided based upon the cumulative variance explained by the PCs, (which was 98%) as well as inference about the gene regions that had or had not been captured by the PCs. PCA decomposes a data matrix, X, such that

$\mathbf{X} = \mathbf{TP}^t$ where **T** and **P** are the scores and loadings matrices respectively.

The animal model was:

$y_{ijklmn} = \mu + scores + HTD_k + kdseason_l + KNUM_m + b_{1n}DIM + b_{2n}DIM + b_{3n}DIM + b_{4n}DIM + u_j + pe_j + e_{ijklmn}$

where:

scores are the columns of **T** corresponding to the first 15 PCs

u is the random polygenic effect (breeding value) other than casein genes of animal *j (j=1,2,…,567 for milk yield and 1,2,...,565 for milk composition).*

Other parameters are as in previous models.

Model 3 is represented in matrix notation as:

$$Y = X\beta + Zu + Zpe + \varepsilon$$

Assuming: $pe \approx N(0, I\sigma_{pe}^2)$, $u \approx N(0, A\sigma_a^2)$, $\varepsilon \approx N(0, I\sigma_e^2)$

$$G_p = \lambda_p I\,, \qquad {G_a}^{-1} = A\sigma_a^2, \qquad \lambda_p = \frac{\sigma_e^2}{\sigma_{pe}^2}, \qquad \lambda_a = \frac{\sigma_e^2}{\sigma_a^2}$$

$$E\,(y) = X\beta \quad and \quad V\,(y) = ZG_aZ^t + ZG_pZ^t + I\sigma_e^2$$

where:

X is a design matrix of all the fixed effects, *Z* is an incidence matrix relating phenotypes to individuals, *I* is an identity matrix, *u* is a vector of breeding values (polygenic effect), *pe* is a vector of the permanent environmental effect of individuals, ε is the vector of residual errors associated with each observation, G_a is the covariance matrix of the polygenic effect, G_p is the covariance matrix of permanent environmental effects, σ_a^2 , σ_{pe}^2 and σ_e^2 are additive genetic, permanent environmental and residual variances respectively. The following equation was solved to get the estimates:

$$\begin{pmatrix} \widehat{\beta} \\ \hat{u} \\ \widehat{pe} \end{pmatrix} = \begin{bmatrix} X^tX & X^tZ & X^tZ \\ Z^tX & Z^tZ + G_a{}^{-1} & Z^tZ \\ Z^tX & Z^tZ & Z^tZ + G_p{}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X^ty \\ Z^ty \\ Z^ty \end{bmatrix}$$

Only the estimated PC effects, $\hat{\beta}_{PC}$ (a sub-vector of $\hat{\beta}$) were of interest in this study. The $\hat{\beta}_{PC}$ were back- transformed into the realm of the original SNPs using the loading matrix, **P.**

$$\hat{\beta}_{SNP} = \boldsymbol{P} \times \hat{\beta}_{PC}$$

The estimated covariances of the estimated PC effects (obtained from the $\boldsymbol{C}^{-1}$ matrix) were also back-transformed into the realm of the original variables:

$$\widehat{cov}_{\beta_{SNP}} = \sqrt{\boldsymbol{P} \times \widehat{cov}_{\beta_{Pc}} \times \boldsymbol{P}^t}$$

- **Significance test**

  Student's t – distribution with degrees of freedom *n – number of fixed effects estimated,* levels of significance 5% and 10% (α = 0.05, 0.1) were used to test for significance of the estimated SNP effects.

  H_o: $\beta_{SNP} = 0$ versus H_a: $\beta_{SNP} \neq 0$ for all SNPs

  The test statistic was: $\frac{\widehat{\beta}_{SNP}}{\widehat{SE}_{\beta_{SNP}}}$

4.0 RESULTS AND DISCUSSION

4.1 Results

4.1.1 General Descriptive statistics

Table 2: Descriptive statistics of traits included in the study

Statistic	TRAITS				
	Fat %	Lactose %	log(SCC)	Protein%	Milk yield(kg)*
Mean	**4.18**	**4.38**	**2.64**	**2.98**	**2.43**
Stdev	1.00	0.24	0.57	0.28	0.80
Median	4.00	4.37	3.00	2.97	2.40
Min	1.00	3.01	1.00	2.21	0.00
Max	10.00	5.19	4.00	4.28	5.20

**per day*

4.1.2 Estimated SNP effects

4.1.2.1 Model 1: Partial Least Squares Regression (PLS I)

The number of latent variables used in the PLS I models, as well as the amount of variance in the traits that they could explain was different for each trait. Based on plots of RMSECV, the optimum model dimension was decided as the number after which there was either levelling off or an increase.

Fat percent

For Fat %, 4 LVs were used in the estimation of the SNP effects. Plots of the X loadings show that the first and second LVs mostly capture variation in the alpha s1 and kappa gene regions. The four LVs explained 87% of SNPs variation and 1.82% of the variation in fat percent.

Lactose percent

Two LVs were used in the Lactose model and they explained 33.5% of SNPs and 0.61% of trait variation. Both of these latent variables mostly captured variance in the αs1 region,

indicating that the SNPs there are relatively more important for explaining lactose percent. I therefore expected that the estimates of SNP effects in αs1 would have higher absolute values than for the other SNPs; however, analysis did not show this pattern. Also, none of the estimated effects were statistically significant. Results are in Figure 2.

Somatic cell count

Only one LV was used for the log(SCC) SNP effect estimation, and judging from the error plots, one would conclude that the SNP information was unsuitable for prediction because the null model which did not take the LVs into consideration at all had almost the same RMSECV as the model with 1 LV. Increasing the model dimension led to an increase in the error. The model with 1 LV explained 44.6% of SNP variance and 0.17% of the trait variance. Loading plots show that the first latent variable assigned positive weights to the SNPs in the κ casein region, whilst the second latent variable assigned negative weights to the very same region.

Protein percent

Protein prediction was optimum at 3 LVs. Just as would be expected, the amount of variance explained in both SNPs and trait was relatively high, with values close to those for fat percent. The final protein model explained 83.2% of the variance in the SNPs and 1.63 of that in protein.

Milk yield

The same is true for the model estimating SNP effects on milk yield, addition of more LVs increased the error, so the final model used only the first LV, which could explain 46% of SNPs and 0.1% of trait variation. Most of the variation captured by this LV was also in the κ-casein region.

SNP effects

Figure 2 shows the significance test for the estimated effects of the SNPs on fat %. Seven SNPs were found to have significant effects at 5% level of significance. 4 of these SNPs are in the αs1 region and the remaining are in the β-casein region. The deletion in exon 9 of CSN1S1 (SNP 11) had a positive effect on fat% and the deletion in exon12 of CSN1S1 (SNP 14) had a negative effect. This is consistent with previous studies on the same population, though in that study SNP effects were estimated one at a time (Dagnachew, 2009 unpublished.)

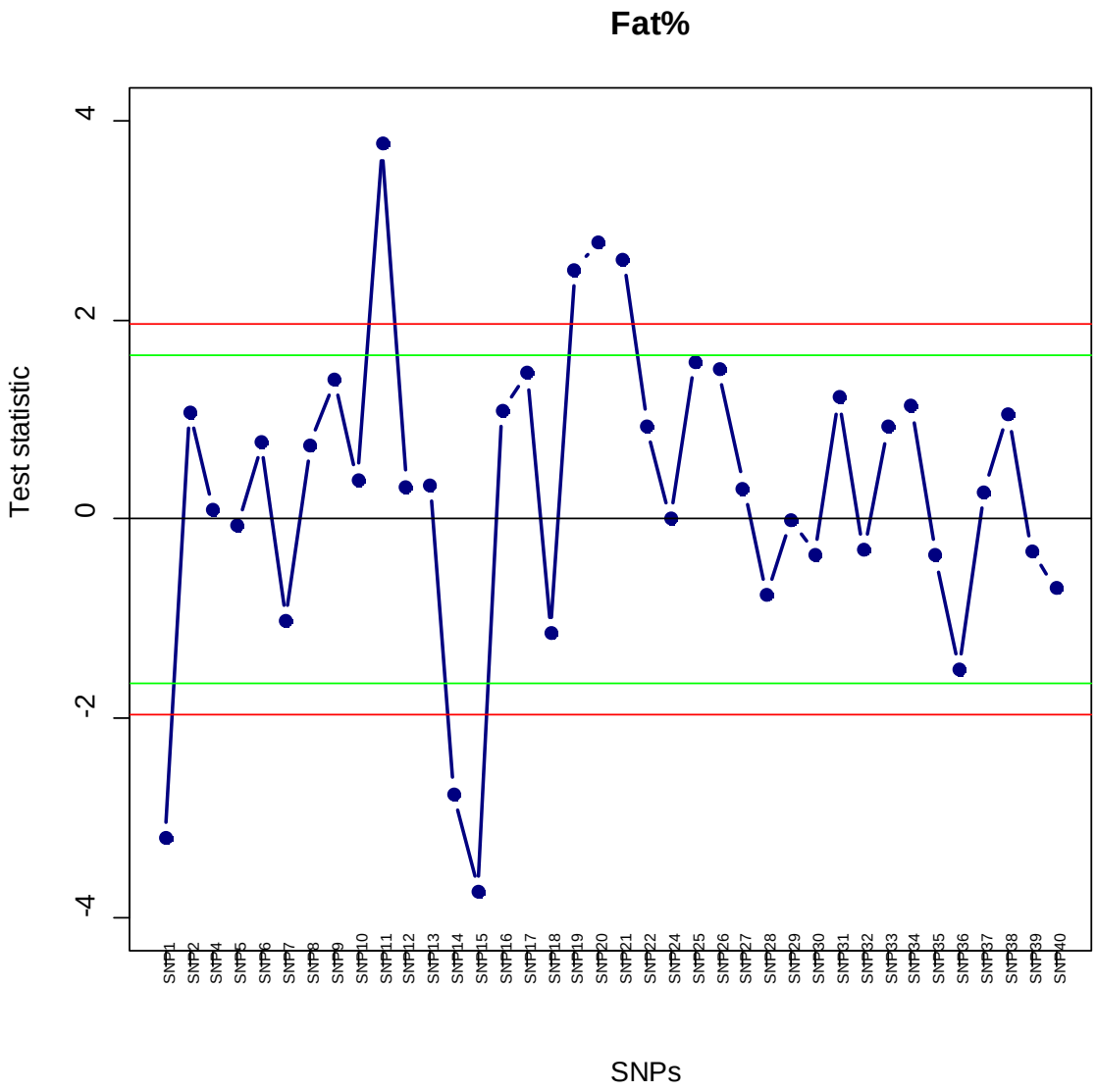


Figure 2: Significance test for the additive effect of major allele of each SNP on Fat percent, estimated with Model 1. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

PLS I did not result in the estimation of any statistically significant SNP effects on Lactose percent. This is not too surprising, because as mentioned earlier, the LVs captured very little of the variation in this trait. Figure 3 shows the significance test for the estimation of SNP effects on lactose percent.

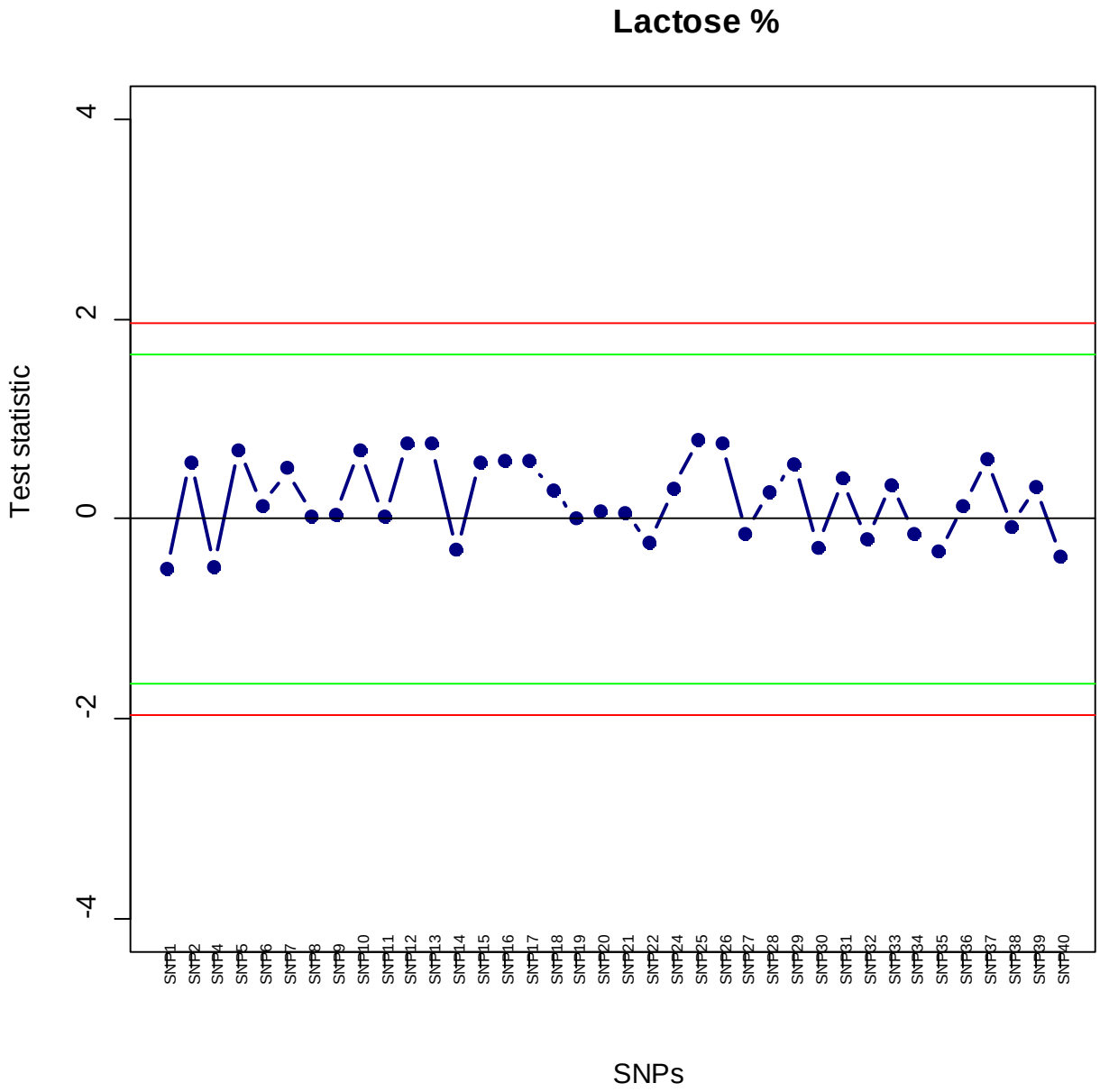


Figure 3: Significance test for the additive effect of major allele of each SNP on Lactose percent, estimated with Model 1. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

For somatic cell count, SNPs 27, 30 and 32, all in the κ-casein region, were significant at 5%. They all had a positive effect on SCC. At 10% significance level, SNP15 in the αs1 region was significant with a positive effect; SNP26 in αs2 had a negative effect and SNPs 36 and 37 both in the κ-casein region had a significant negative effect. Overall, most of the 'important' SNPs for explaining variation in SCC were in the κ-casein genes. Figure 4 gives an overview of the significance test for the additive effect of SNPs on somatic cell count.

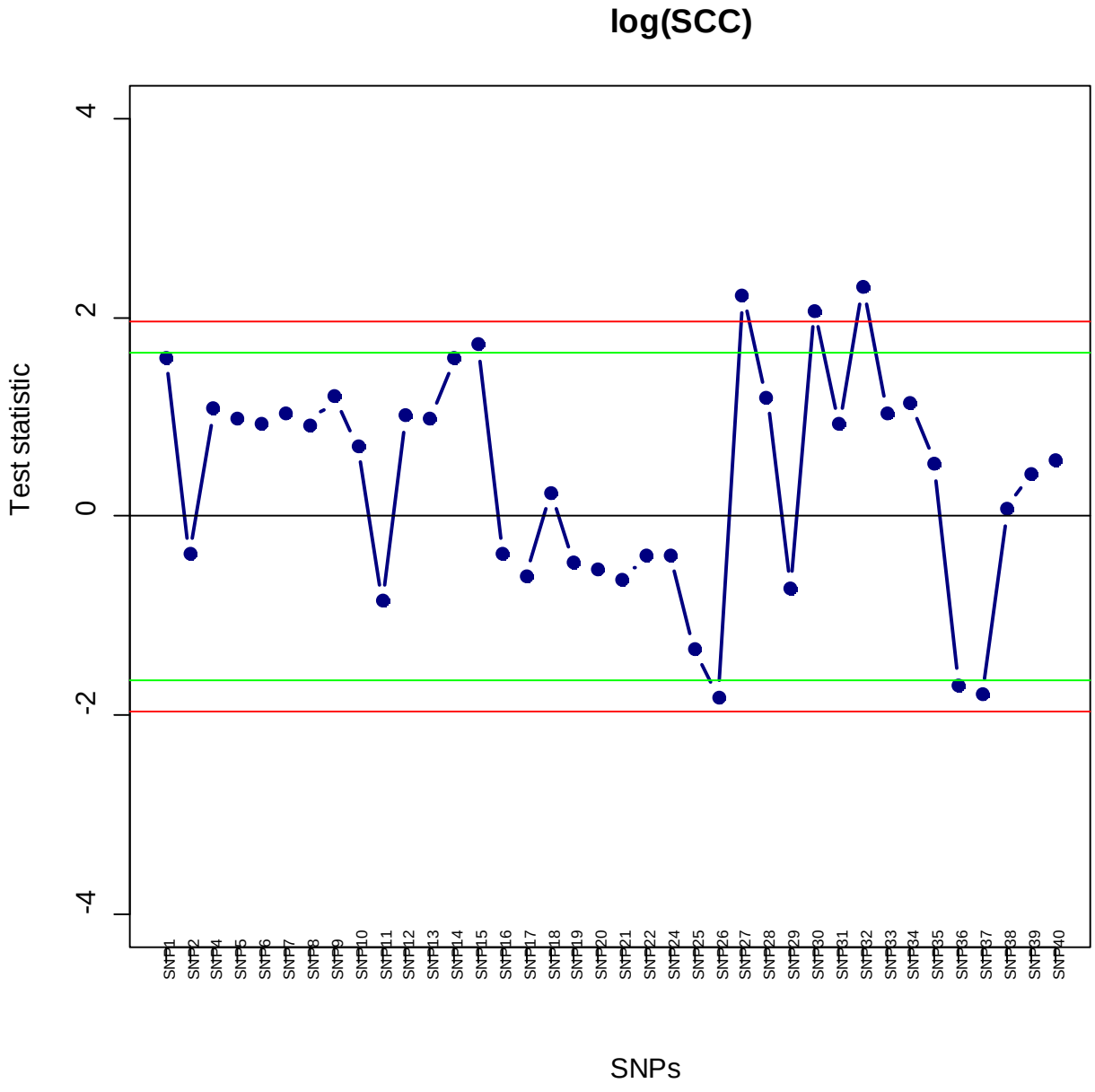


Figure 4: Significance test for the additive effect of major allele of each SNP on log-transformed Somatic cell count, estimated with Model 1. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

The significance tests for the estimated SNP effects on protein percent are shown in Figure 5. All the significant SNPs in the αs1 region had negative effects. As in the other traits, αs1 and κ –caseins genes show marked significance. In total 9 SNPs were found to be significant at 5% level. The deletion in exon 12 of CSN1S1 had a significant negative effect, as was expected.

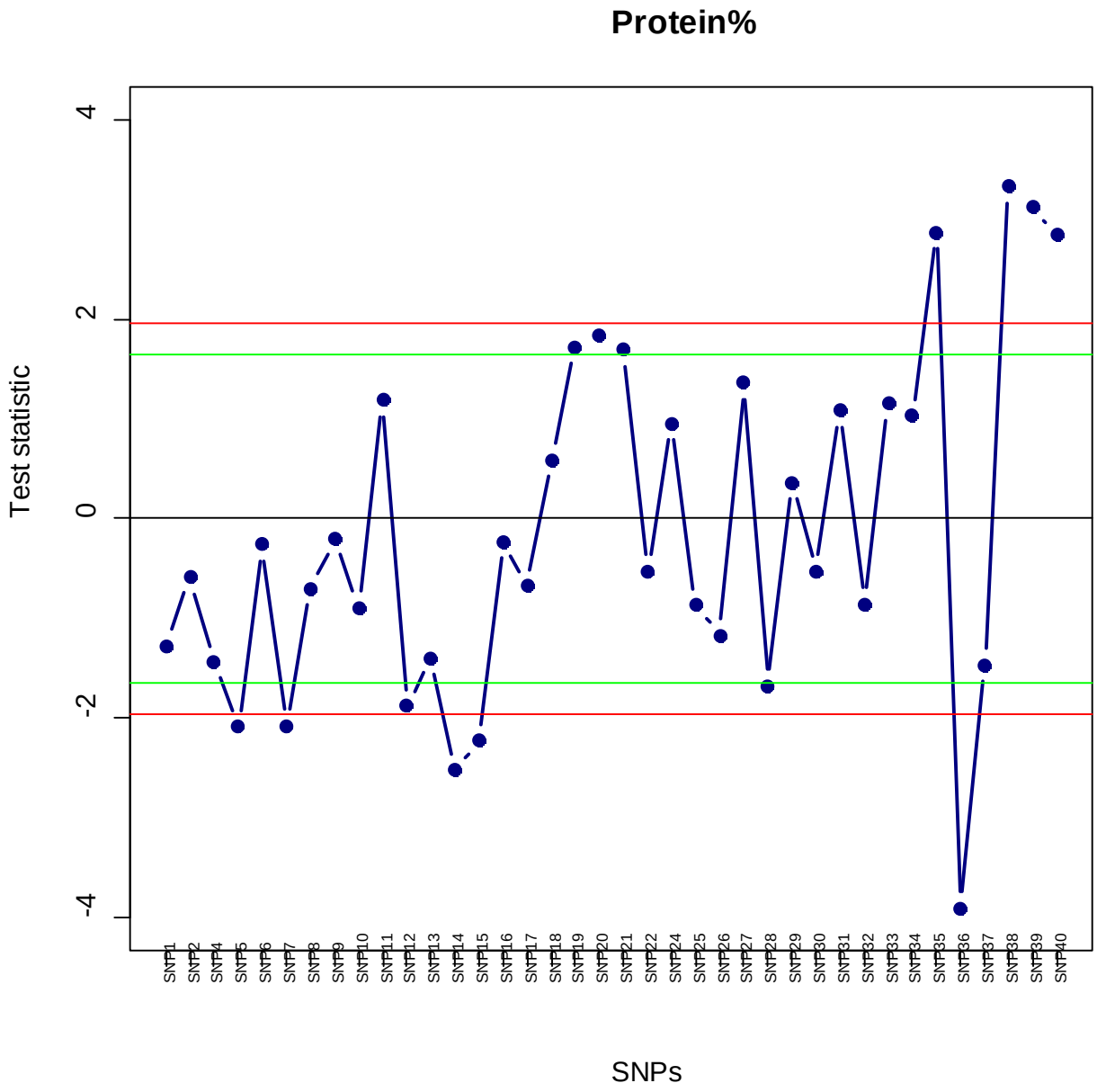


Figure 5: Significance test for the additive effect of major allele of each SNP on Protein percent, estimated with Model 1. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

For milk yield, SNPs 16 and 17 in the β-casein region had a significant negative effect at 5% level of significance. All other significant effects were in the κ-casein genes, with a cluster from SNP 27 to 34 (excluding SNP 29) all showing negative effects. Figure 6 shows the significance of estimated SNP effects on milk yield.

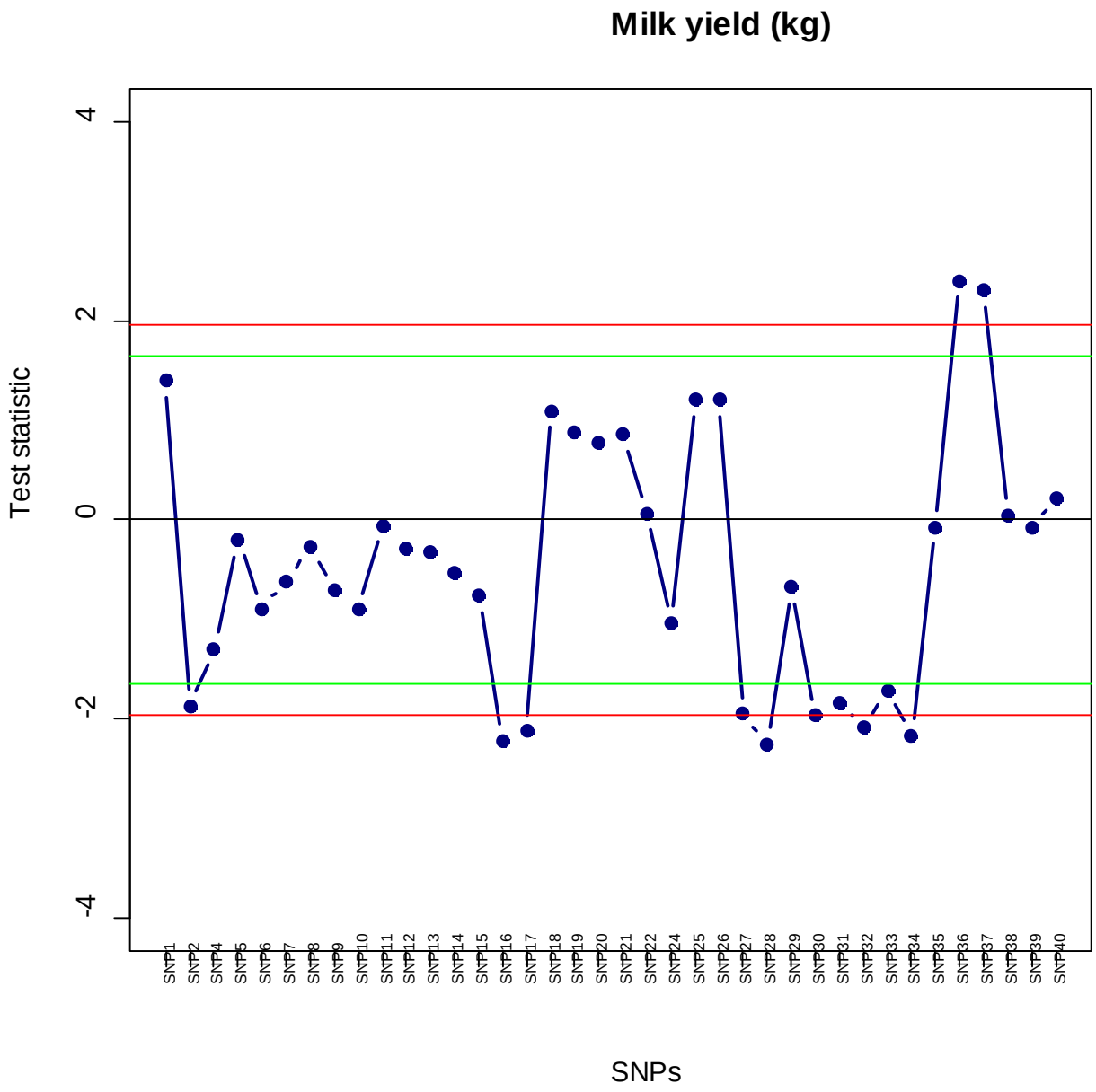


Figure 6: Significance test for the additive effect of major allele of each SNP on Milk yield (kg), estimated using Model 1. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Figures 7, 8 and 9 show the actual values of the estimated SNP effects for each of the traits. On average the highest estimates were for fat and protein, as was expected. A table including the standard errors of these estimates is in the appendix. The plots for milk yield and somatic cell count show opposite patterns, very clear especially in a cluster of SNPs in the CSN3 region. This negative correlation is consistent with findings by several authors (Zeng et al., 1995).

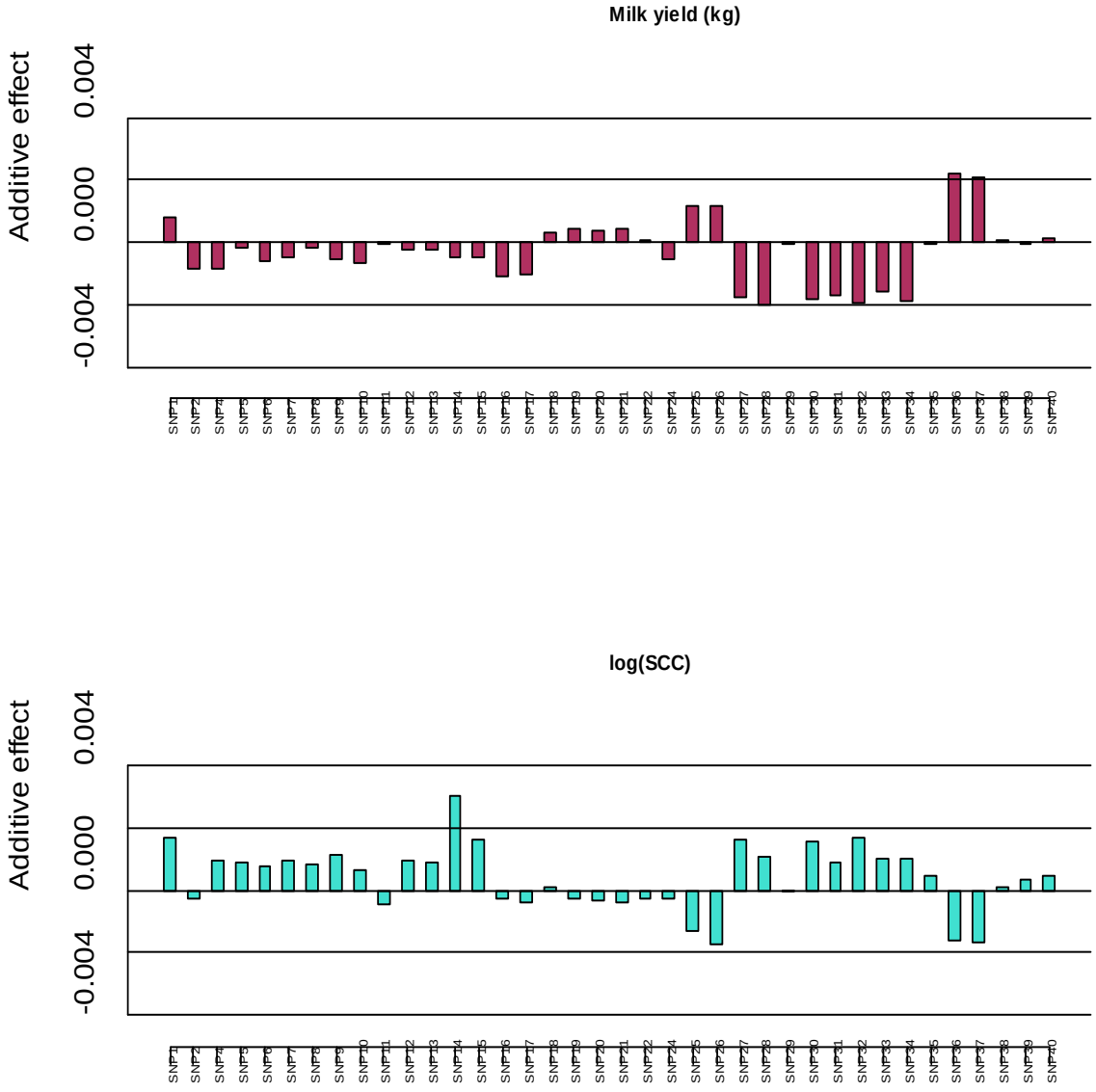
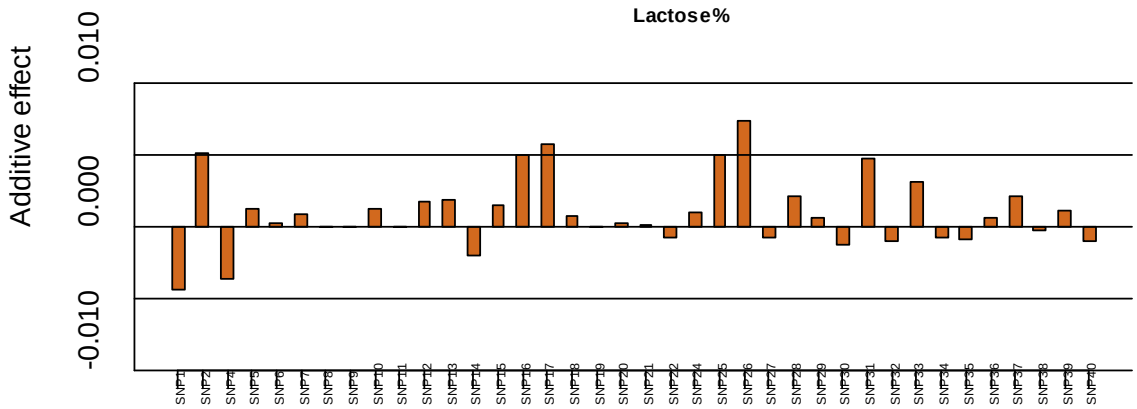


Figure 7: Additive effect of major allele of each SNP on milk composition traits (Estimated with Model)



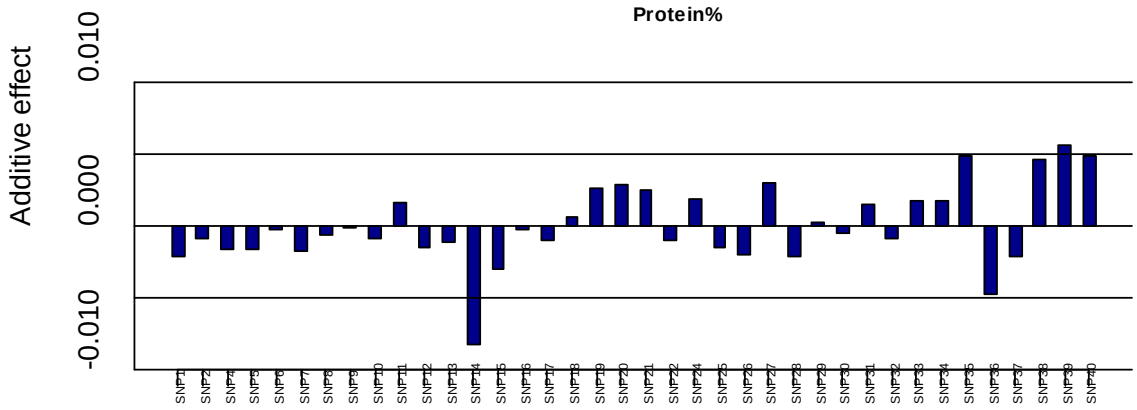


Figure 8: Additive effect of major allele of each SNP on Fat % (Estimated with Model1)

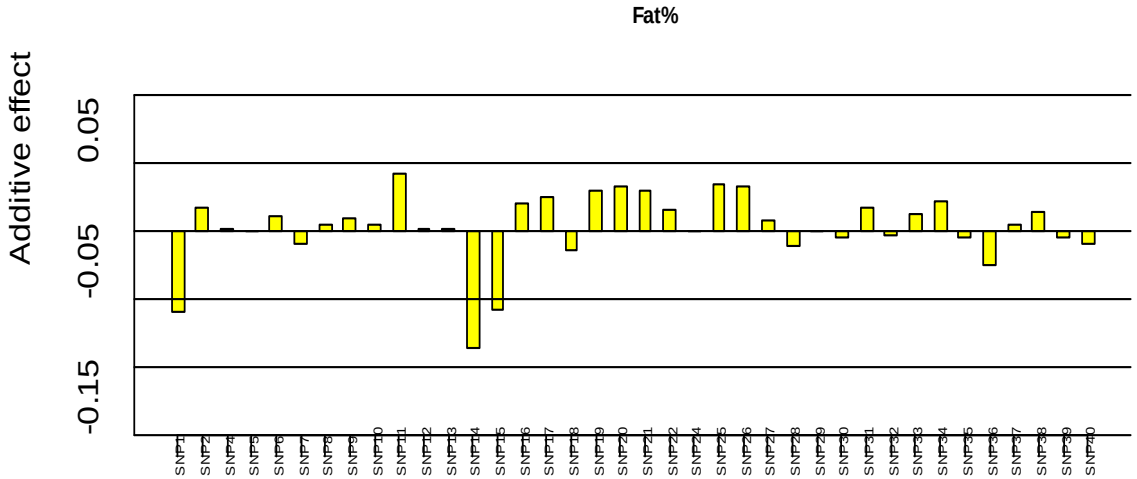


Figure 9: Additive effect of major allele of each SNP on Fat % (Estimated with Model1)

4.2.2 Model 2: Partial least squares regression II (PLS II)

PLSR II has an advantage over PLSR I. This seems to be the case with at least fat and protein percent. A joint analysis of all traits was performed to see its' effect on the estimation would increase the chance of selecting LVs that best explain the total variance/covariance between the traits. This however was not the case for this PLS II model. The optimum number of LVs was 4 and they explained 87.36% of the SNP variance, 0.81 of log (SCC), 1.81% of fat, 0.37% of lactose and 1.5% of protein.

The only increase in explained variance was for SCC, which increased 5 fold. Lactose and protein had reduced explained variances but Fat % remained the same as in the PLS I model.

The effects of this increase in explained variance of SCC can be seen in Figure 10. The effects of the αs1 SNPs were the same in terms of whether they were positive of negative, but due to the increase in explained variance, estimation ability was increased leading to smaller SEs and thus a higher number of significant SNPs. Strangely though, SNPs 31, 33 and 34 which had positive estimated effects with the PLS I model were negative with the PLS II estimation.

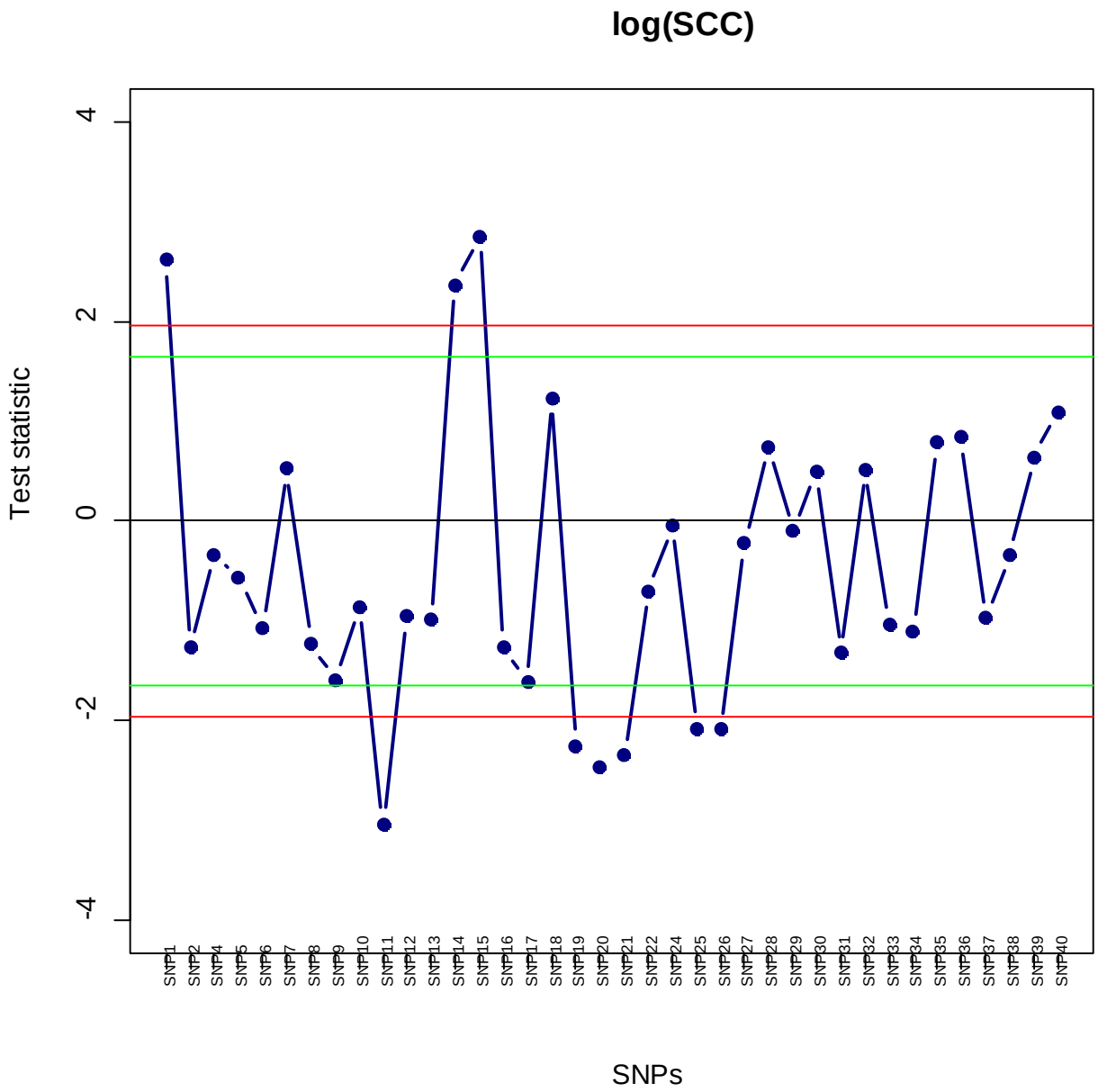


Figure 10: Significance test for the additive effect of major allele of each SNP on log-transformed somatic cell count, estimated using Model 2. *The red and green horizontal lines represent 5% and10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

For the estimation of SNP effects on lactose percent, the results are not as easy to explain. One would expect that since the amount of variance explained by the PLS II model was less than in the previous model, estimates would be less accurate and thus not significantly different from zero; this was not the case. 5 SNPs in CSN1S1 were found to have a significant positive effect at 10%, and SNP9 had a negative effect at 5% level of significance. It could be

that the additional information from the other traits are very good at explaining lactose percent. The estimated additive effects were however smaller than what was estimated with model 1.The results are presented in Figure 11.

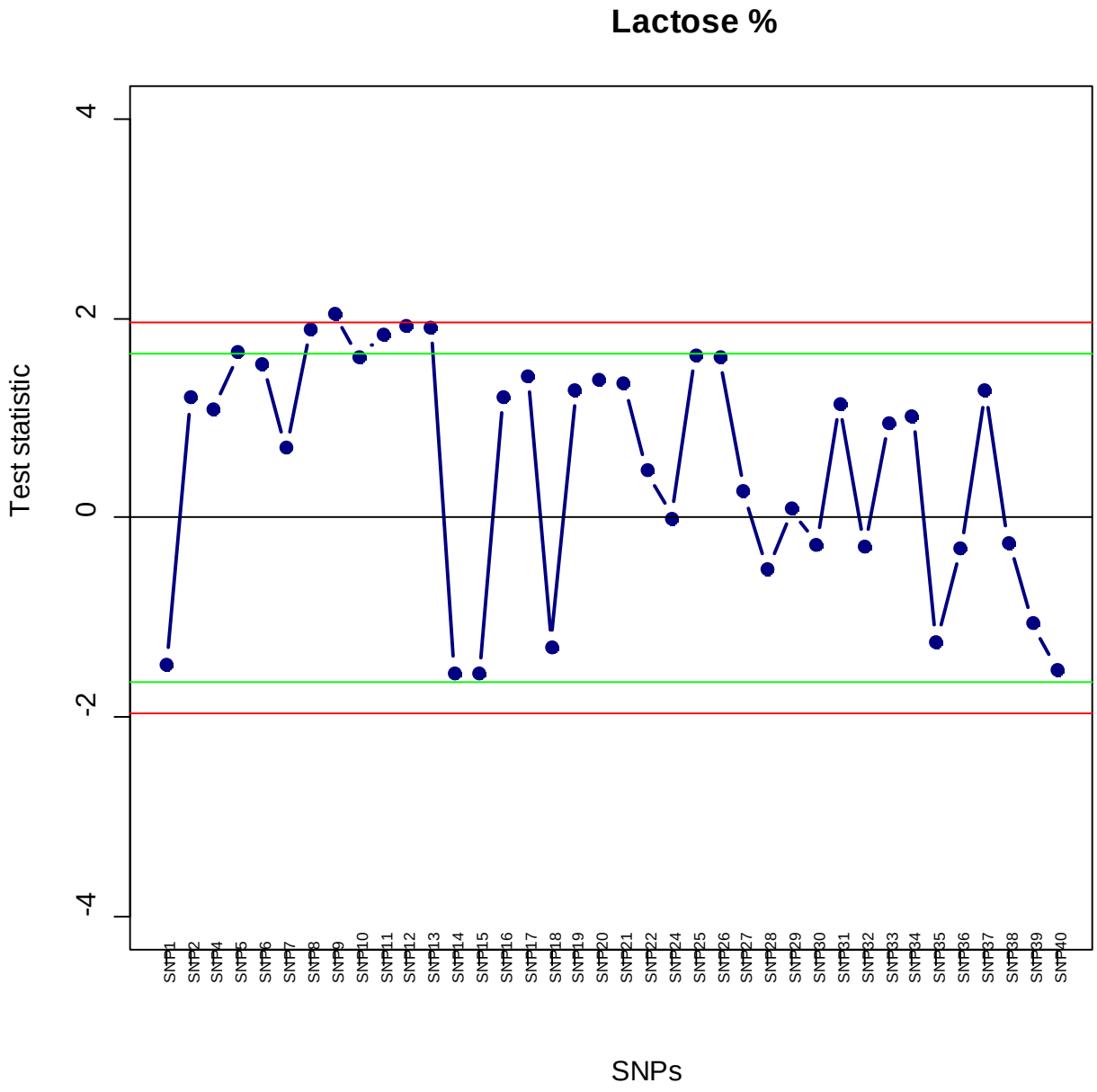


Figure 11: Significance test for the additive effect of major allele of each SNP on Lactose percent, estimated with Model 2. *The red and green horizontal lines represent 5% and10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Estimates for fat percent were completely consistent in terms of positive/negative SNP estimates, as well as SNP effects that were found to be significant at 5%. This again reflects that the casein SNPs explain fat percent relatively better than the other milk composition

traits. The deletion in exon 9 of CSN1S1 has a positive effect, than in exon 12 has a negative effect. This again shows the negative correlation between somatic cell count and milk yield. The significance test results are shown in Figure 12.

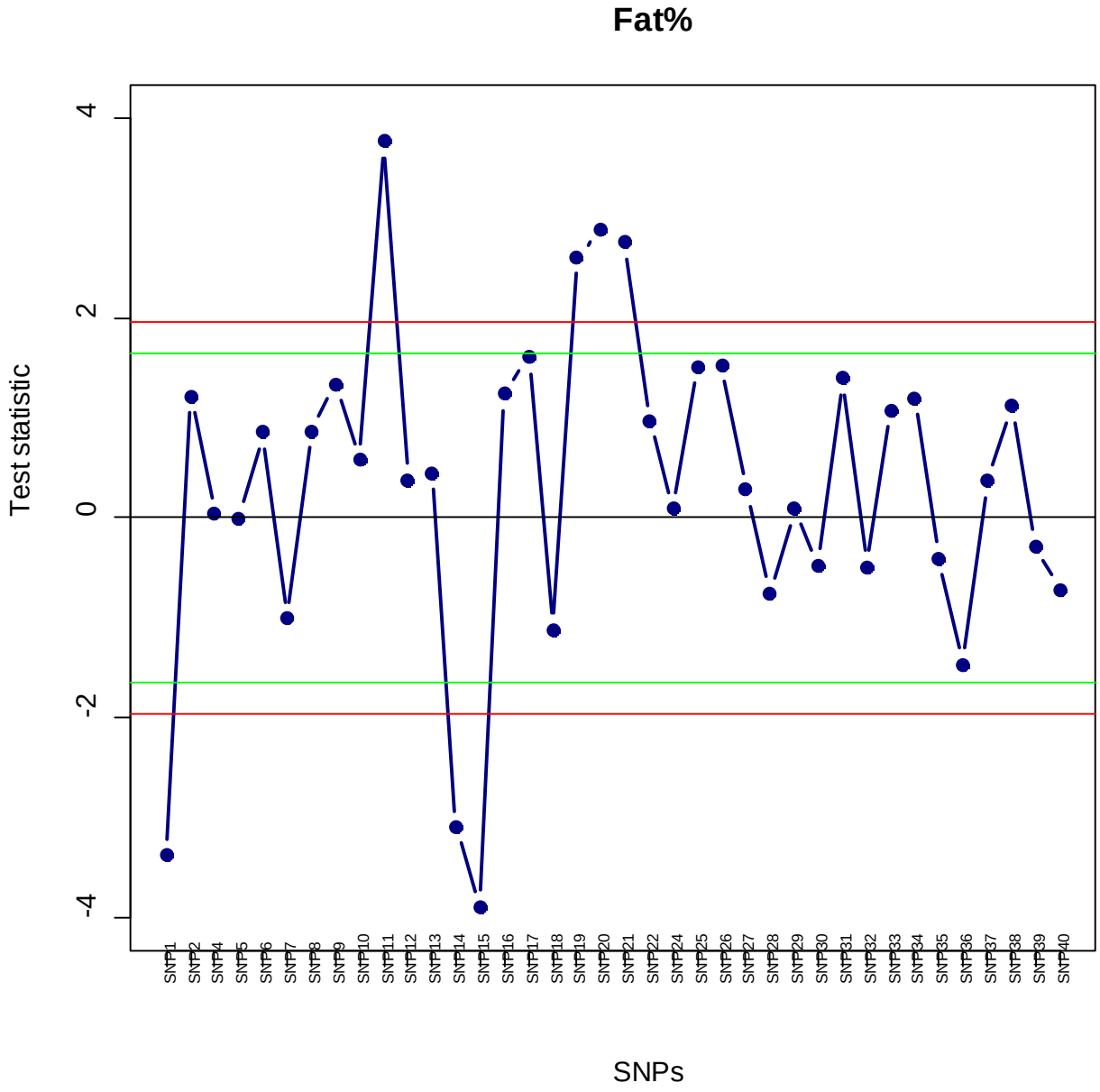


Figure 12: Significance test for the additive effect of major allele of each SNP on Fat percent, estimated with Model 2. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Protein results for model 2 were relatively consistent with model1, with respect to whether a SNP was estimated as either having a positive or negative effect. Most estimates however increased in terms of absolute value, resulting in higher significance level of 3 SNPs in the CSN2 region. On the other hand, two CSN3 SNPs: 39 and 40, which were significant with the PLS I model are now non-significant. Figure 13 shows this.

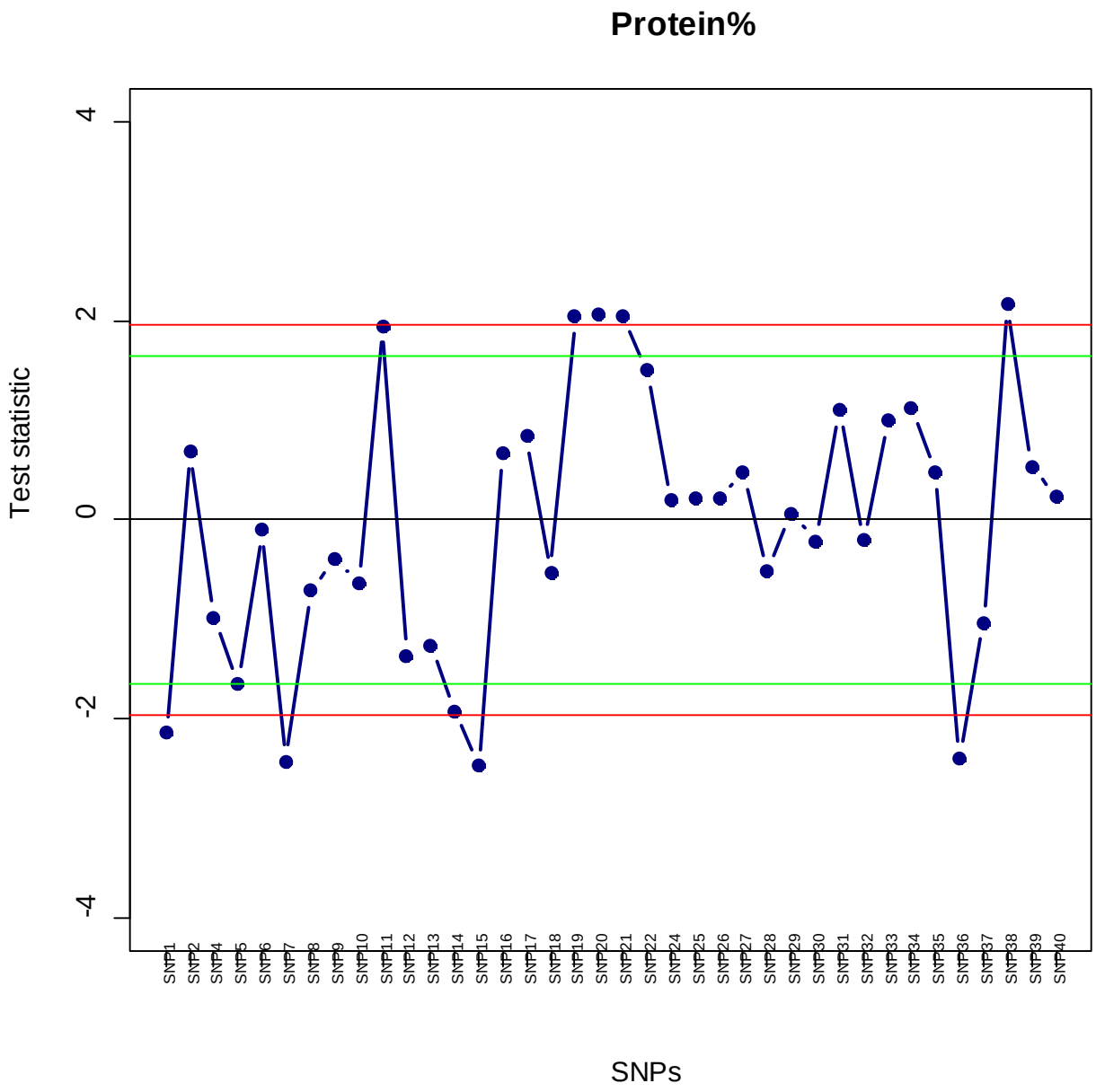


Figure 13: Significance test for the additive effect of major allele of each SNP on Protein percent, estimated with Model 2. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Figures 14 and 15 show the actual values of the estimated SNP effects. A table with the standard errors of the estimates is in the appendix.

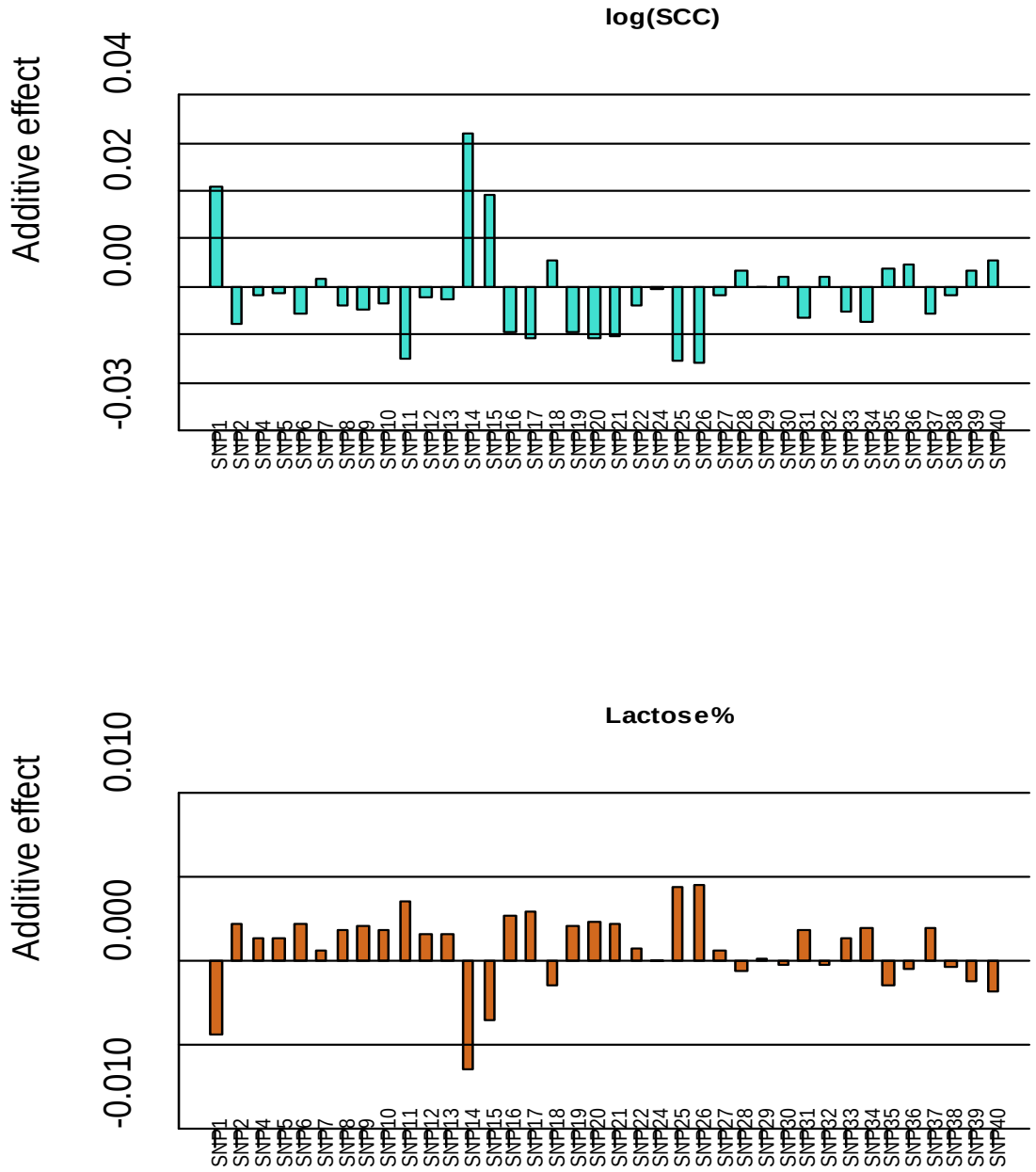


Figure 14: Additive effect of major allele of each SNP on milk composition traits (Estimated with Model2)

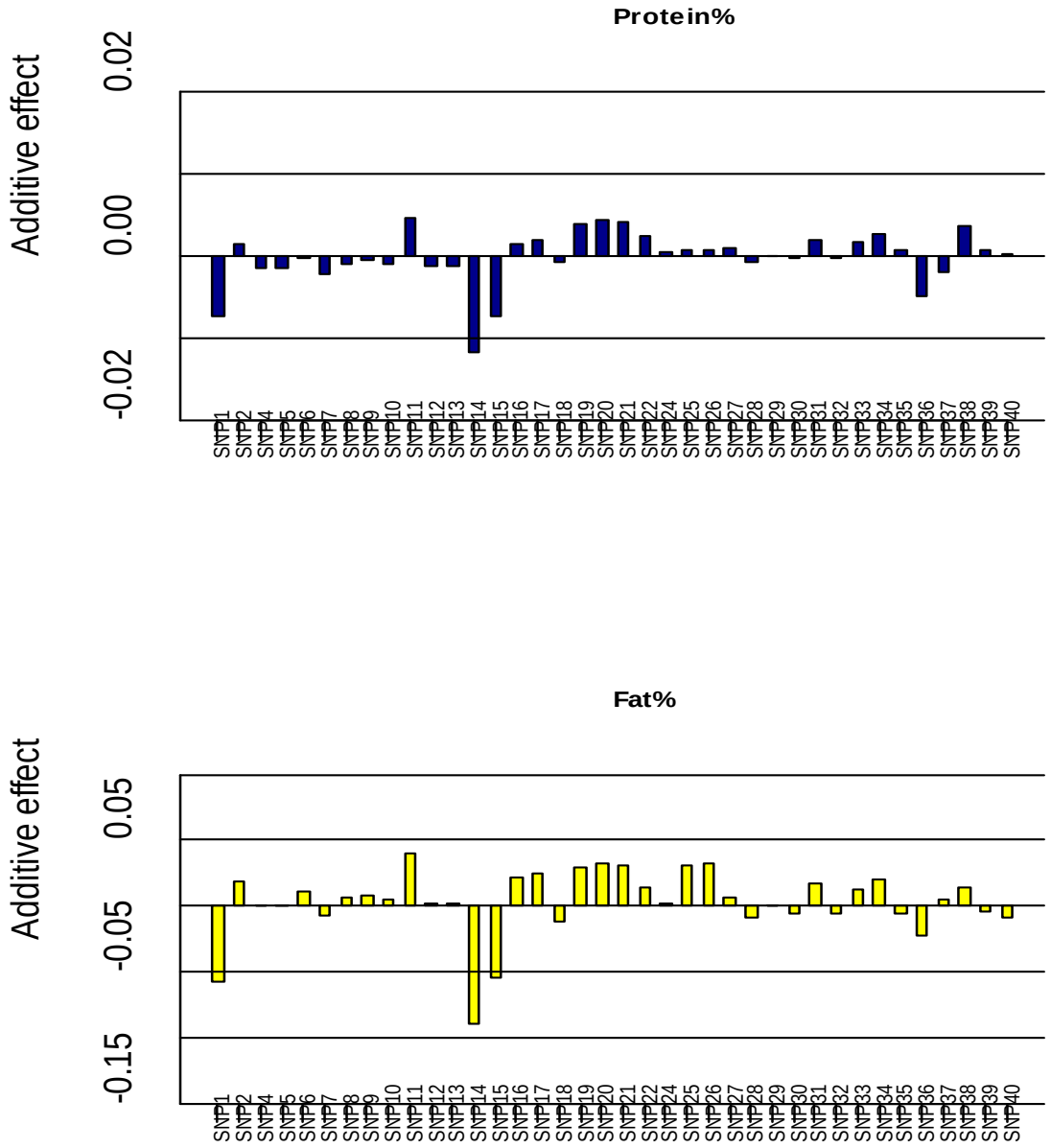


Figure 15: Additive effect of major allele of each SNP on milk composition traits (Estimated with Model2)

4.1.2.3 Model 3: PCA combined with Animal model

Apart from selecting the optimum number of PCs based on the proportion of the initial variance in the SNP information that they could explain, plots of the X loadings were studied to infer whether the variance in all portions of the 250kb casein gene complex had been captured or not.

With 10 components, the total explained variance of the SNPs was 94.3%, implying that this number would have been more than ideal if my main aim was data reduction. My main interest with this particular model however, was avoiding the problem of multicollinearity whilst ensuring (as far as possible) that variation across the entire complex was maintained; and not necessarily variable selection.

Having taken this into consideration, I used scores from the first 15 PCs in the mixed model. They explained 98% of the total variance in the original SNP data.

Fat percent

Only the deletion in exon 12 of CSN1S1 was found to be significant (α=0.1). It had a negative effect. The pattern of estimated effects in the region spanning CSN2 and CSN1S2 genes was similar in all three models but the estimates from the PLS based approaches however had higher absolute values. Results are shown in Figure 16.

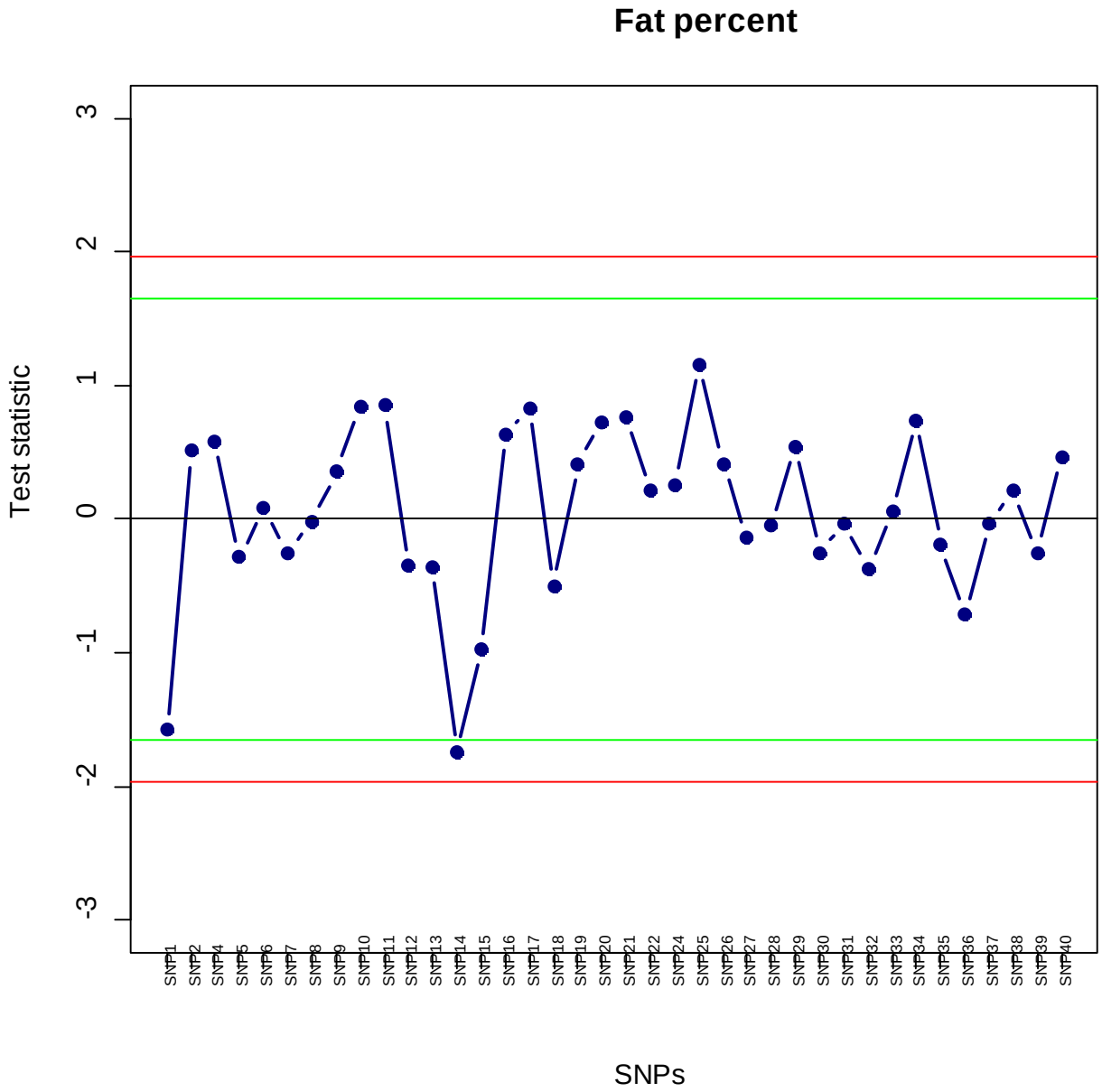


Figure 16: Significance test for the additive effect of major allele of each SNP on Fat percent, estimated with Model 3. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Lactose percent

No significant SNP effects were recorded for lactose percent. The standard errors were the highest compared to the other traits and the other models as well.

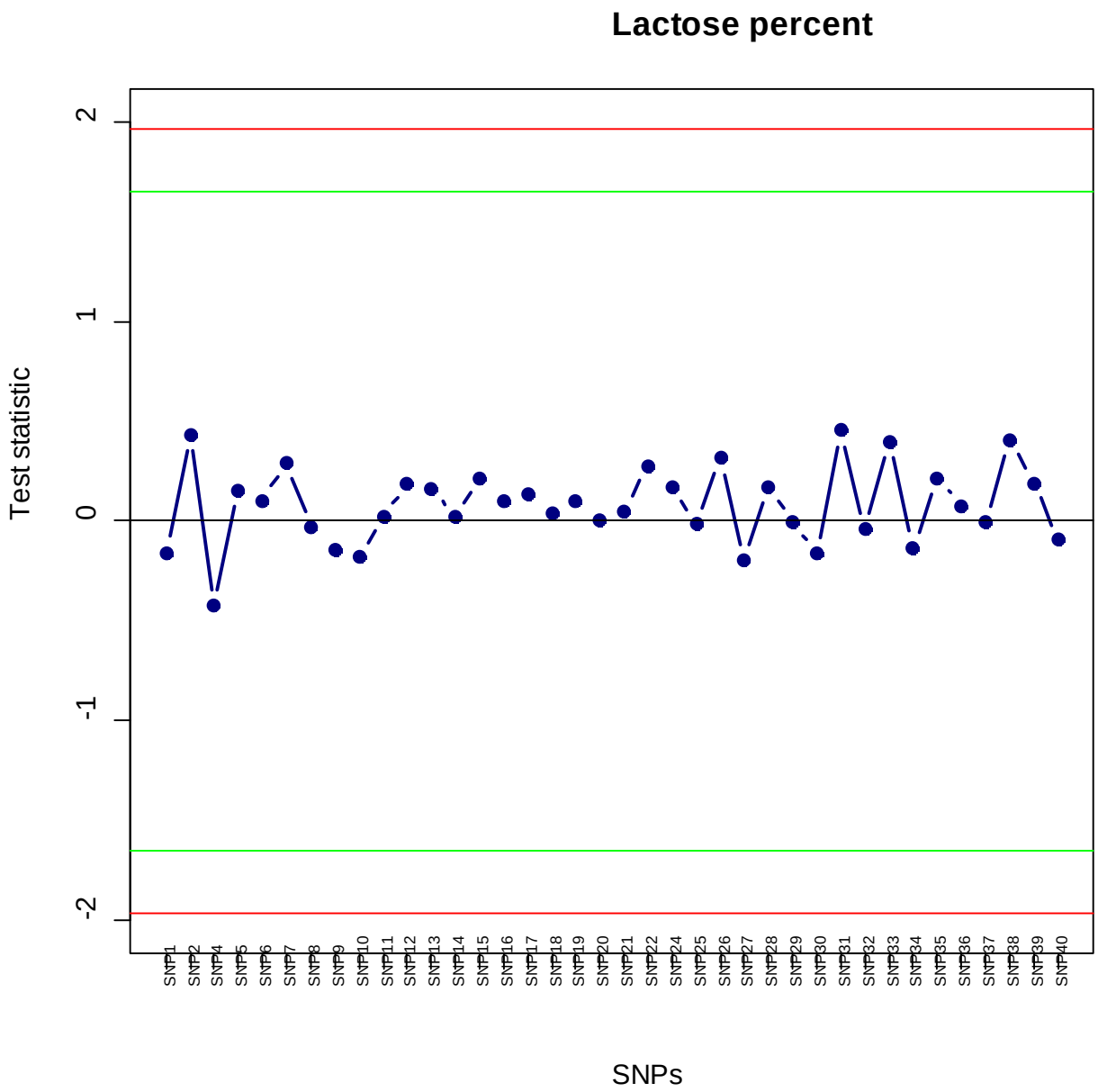


Figure 17: Significance test for the additive effect of major allele of each SNP on Lactose percent, estimated with Model 3. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Somatic cell count

Although none of the estimated SNP effects were statistically significant, the effect of exon 12 deletion clearly stands out in the plots. The results for the test of significance are in Figure 18.

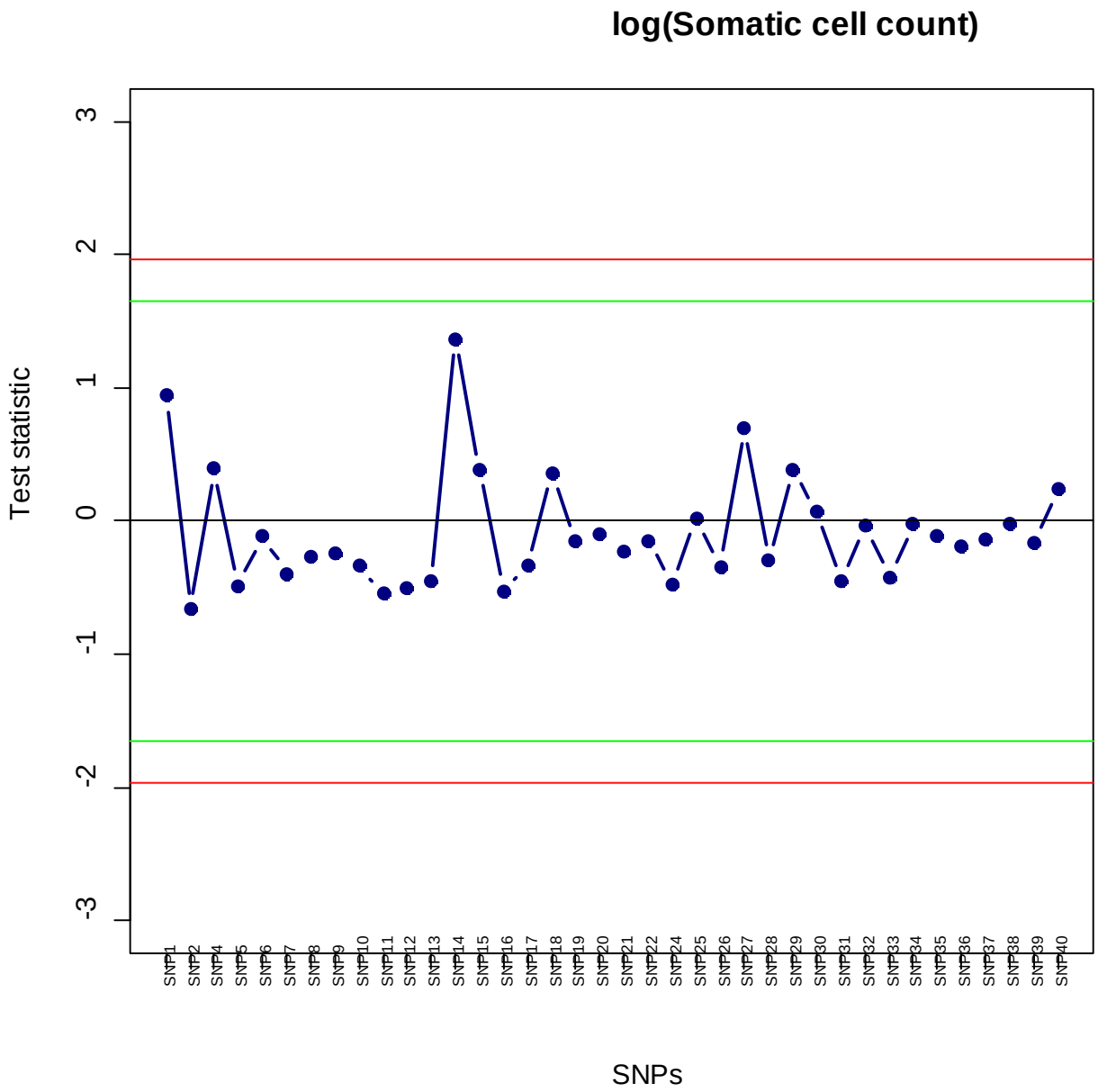


Figure 18: Significance test for the additive effect of major allele of each SNP on log-transformed Somatic cell count, estimated with Model 3. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Protein percent

The alpha s1 polymorphisms are reported to explain 48% of the additive genetic variance in protein percent (Seradilla, 2003). This might be an overestimation, but judging from results of Models 1 and 2, it is clear that CSN1S1 is important for protein percent variation. Model three failed to explain the variation in protein percent. High standard errors resulted in estimates that do not significantly differ from zero. One downside of PCA-based estimates is

this method produces biased estimates and the bias cannot be measured without knowledge of the true coefficients (Enns, 1979) The poor fit of this model indicates that the principal components were not suitable for estimation of SNP effects on protein percent.

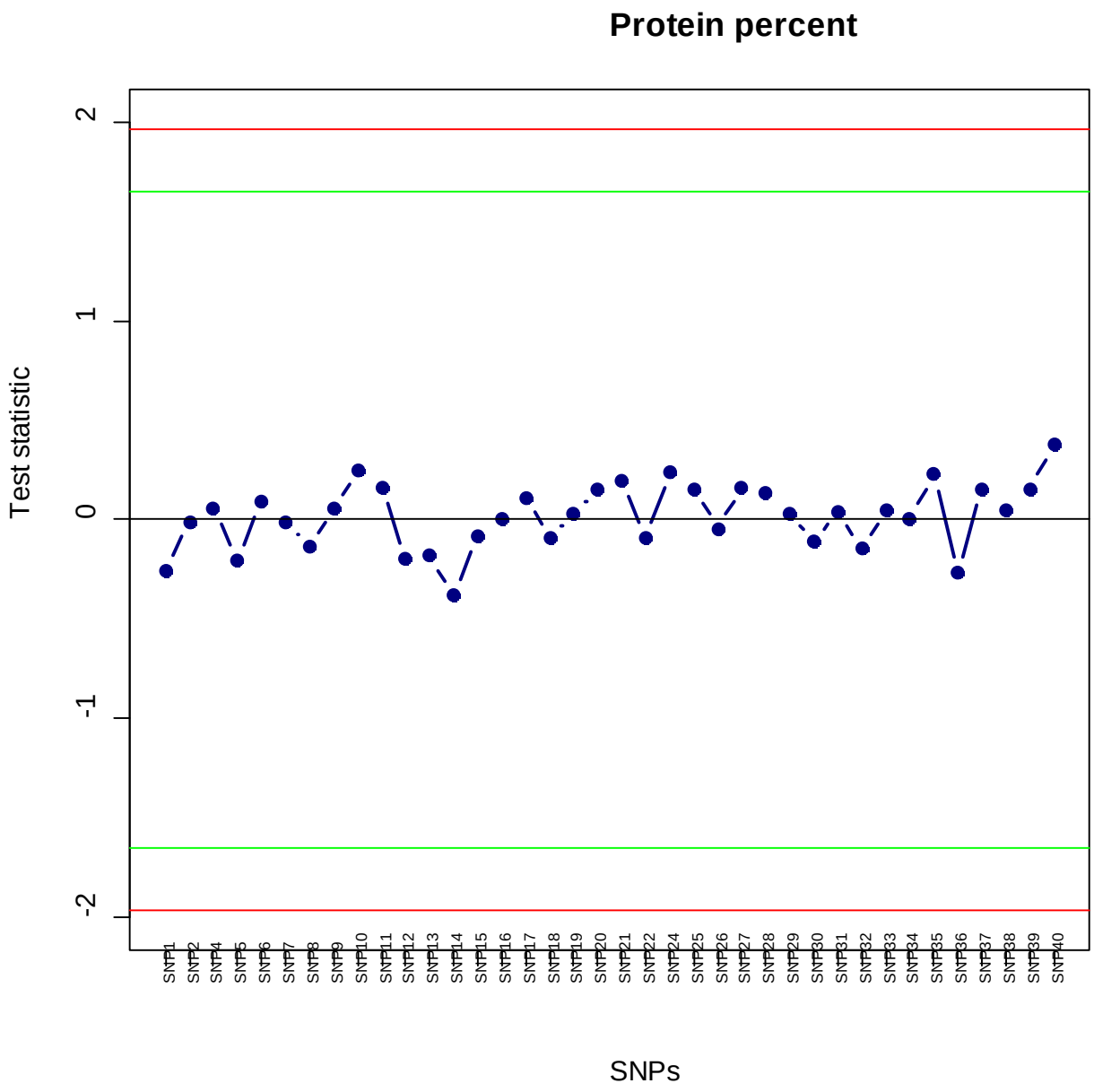


Figure 19: Significance test for the additive effect of major allele of each SNP Protein percent, estimated with Model 3. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Milk yield

Model 3 estimated significant positive effects of SNPs 5, 12 and 13. This is not consistent with the PLS I model. The values of the SNP effects also vary greatly, however both models estimated a positive effect of SNP 36 on milk yield ($p<0.05$).

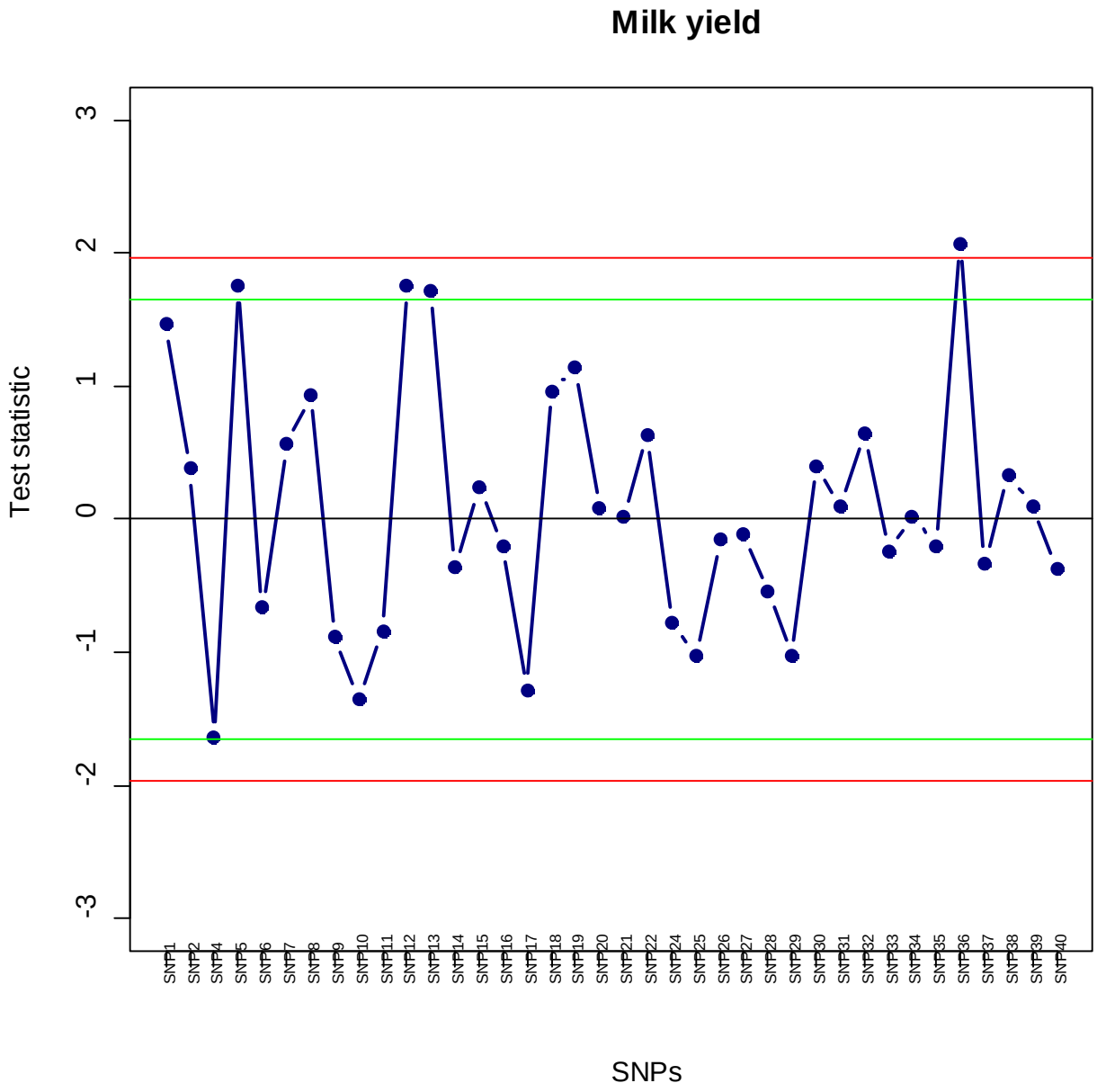


Figure 20: Significance test for the additive effect of major allele of each SNP on Milk yield (kg), estimated with Model 3. *The red and green horizontal lines represent 5% and 10% experimental-wise threshold level respectively; any SNP above the top line or below the bottom line is taken as significant.*

Figures 21 and 22 show the actual estimates of the additive effects of the SNPs on the milk composition traits, a table which shows the standard errors is in the appendix. The errors for this model were relatively large, so SNP effects that appear large may not be statistically significant.

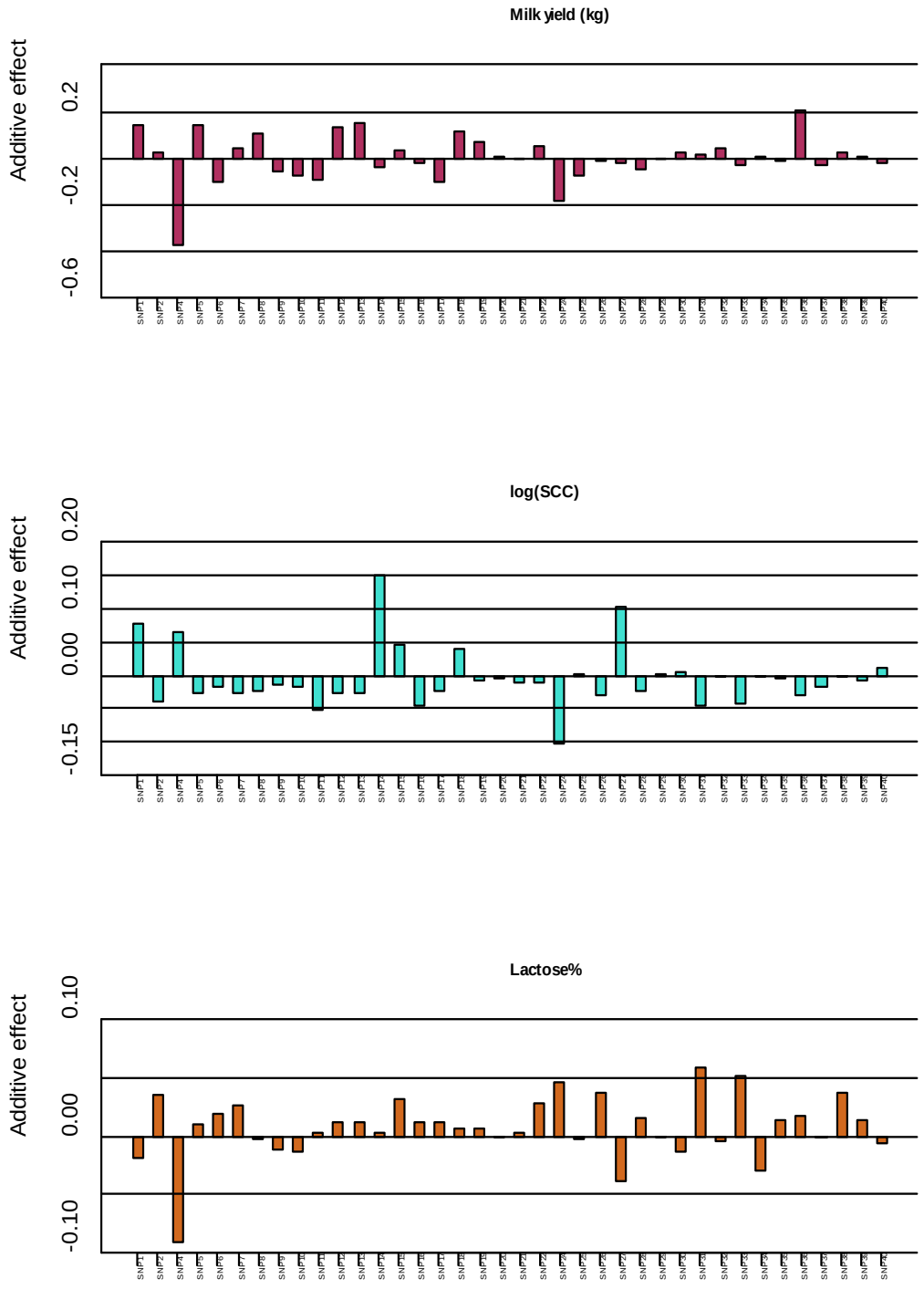


Figure 21: Additive effect of major allele of each SNP on milk composition traits (Estimated with Model 3)

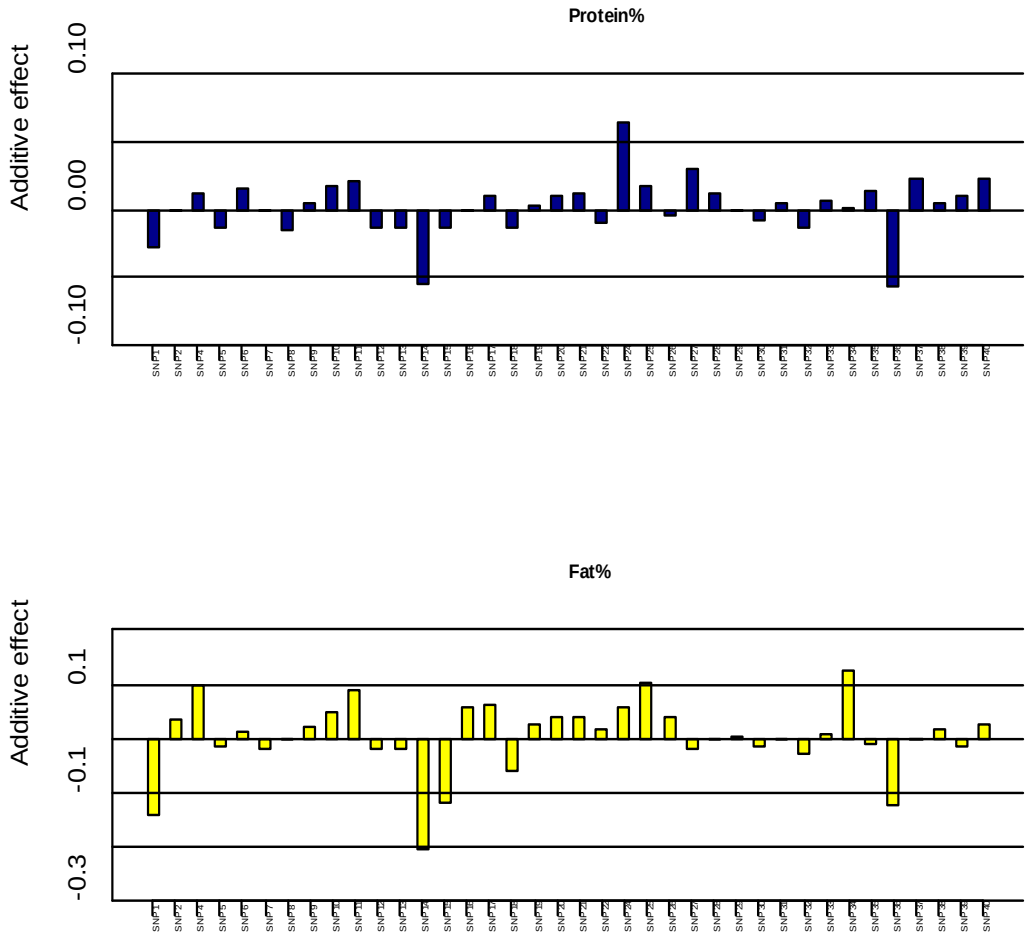


Figure 22: Additive effect of major allele of each SNP on milk composition traits (Estimated with Model 3)

Table 3: Summary of statistically significant SNP effects across all models (α=0.1)

SNPs	TRAITS													
	Fat%			Lactose%			log(SCC)			Protein%			Milk (kg)	
	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M3
SNP1	*	*						*			*			
SNP2													*	
SNP4														
SNP5										*				*
SNP6														
SNP7										*	*			
SNP8					*									
SNP9					*									
SNP10														
SNP11	*	*			*			*			*			
SNP12					*					*				*
SNP13					*									*
SNP14	*	*	*					*		*	*			
SNP15	*	*					*	*		*	*			
SNP16													*	
SNP17													*	
SNP18														
SNP19	*	*						*		*	*			
SNP20	*	*						*		*	*			
SNP21	*	*						*		*	*			
SNP22														
SNP24														
SNP25								*						
SNP26							*	*						
SNP27							*						*	
SNP28													*	
SNP29														
SNP30							*						*	
SNP31													*	
SNP32							*						*	
SNP33													*	
SNP34													*	
SNP35										*				
SNP36			*				*			*	*		*	*
SNP37							*						*	
SNP38										*	*			
SNP39										*				
SNP40										*				

**Asterisks indicate SNPs which are significant at 10% level of significance*

***M1**: PLS I model; **M2**: PLS II model; **M3**: PCA/MME model*

4.2 General Discussion

The ability to identify and test for the existence of associations between gene polymorphisms and economically important traits is an asset to animal improvement efforts. With respect to the 4 casein genes, haplotype-based approaches have been the preferred choice due to the high level of linkage disequilibrium between them. Hayes et al., (2006) identified 28 haplotypes within the casein gene region in a Norwegian goat population, 6 of which were associated milk composition traits.

Additive effects on milk yield, fat, protein and lactose percent have also been estimated by Dagnachew et al., (2010). The general similarity that can be seen between those studies and mine are that the CSN1S1 and CSN3 genes appear more relevant for especially fat and protein percent, and that SNP effects on fat percent are slightly easier to detect than in protein percent. The effects of the CSN1S1 and CSN3 polymorphisms have also been reported in goat populations in Spain, France and Italy (Angulo et al., 1994; Chiatti et al., 2005).

In the present results, the deletion in exon 12 of CSN1S1 (SNP14) and SNP36 had a negative effect on both fat and protein percentages, Dagnachew et al., (2010) reported the same, the only difference being that the PLS model used in this study estimated both effects as being statistically significant at 5% whilst in the cited work they were significant at 10%. Again, both studies indicated a cluster of genes in the kappa region that had significant effects on protein percent and milk kg, but not on lactose.

In terms of methodology, Dagnachew et al., (2011) differed completely from the present work: in their study the SNP effects were estimated one at a time, using a single trait test-day mixed model with fixed effect of single SNP's. Additive and dominance and polygenic effects were fitted. All three models in this report performed multi-SNP analysis, but neglected to

account for the dominance effects. The PLS models did not account for polygenic effects, but the PCA/MME model included a term to account for polygenic effects.

To offer an overall comparison of the models used in this study, I would say that the two PLS models showed more consistency both within this study (Table 3) and in comparison to findings by Hayes et al., (2006) and Dagnachew et al., (2010).

One concern that arose during this study was that in the pre-correction of the traits for the PLS models, some of the genetic information in the records would be lost by the fitting of the permanent environment effect, and also the fact that the PLS models took neither the relationship between the animals nor the polygenic effect into account. Notwithstanding, the results are quite comparable to previous studies.

The 'failure' of the PCA based method is difficult to explain, since almost all the variance in the SNPs was captured by the components I included in the mixed model (98%). Also, different models with varying numbers of PCs were tested and it was clear that an increase or decrease in the number of components led to very noisy estimates. Even at the optimum number of 15 PCs, the standard errors were much higher than in the two PLS models. In regular regression, PLS outperforms PCA, but I expected that the combined mixed animal model/PCA model would efficiently detect associations between the casein polymorphisms and at least fat or protein percent. This could be due to the fact that PCA-based estimates may be biased.

Apart from comparing which SNP effects Dagnachew et al., (2011) found to be significant, verifying the actual values of the SNP effects would have been interesting. They were however not included in the report that was available to me.

5.0 CONCLUSION

The PCA/MME model was not able to detect as many associations between the casein polymorphisms and milk composition traits as compared to previous studies. I must state however that for milk yield 4 SNPs, and for fat percent 2 SNPs were significant at 10%. The 2 SNPs that were identified as significant for fat were also significant in the research by Dagnachew et al, (2010). This could indicate that the model itself may not be the problem, but that it did not have enough power in the case of this data set.

Further studies are necessary before one can conclude whether a combination of principal component analysis and the mixed model equations is feasible or not.

The PLS models show a lot of promise for identifying associations between casein polymorphisms and milk composition traits. Results obtained had many similarities with those obtained both from haplotype based approaches and single SNP analysis in similar goat populations. The associations between certain polymorphisms and milk traits have somewhat been confirmed, but further analyses is necessary to verify the actual size of the effects. Dagnachew et al., (2010) reported the existence of non-additive effects of the exon 12 deletion on milk yield, fat and protein percent. I suggest that future work in this area should account for this when studying associations between the casein polymorphisms and milk traits.

REFERENCES

Abdi, H. (2010), Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdisciplinary Reviews: Computational Statistics, 2: 97–106. doi: 10.1002/wics.51

Ådnøy, T., Vegarud, G., Devold,T. G., Nordbø, R., Colbjørnsen, I., Brovold, M., Markovic, B. ,Roseth, A. and Lien, S. (2003). Effects of the 0- and F- alleles of alpha S1 casein in two farms of Northern Norway. The International Workshop on Major Genes and QTL in Sheep and Goat, Toulouse (France).

Andonov, S., Ødegard, J., Boman, I. A., Svendsen, M. Holme, I. J., Adnøy, T.,Vukovic, V. and Klemetsdal, G. (2007). "Validation of Test-Day Models for Genetic Evaluation of Dairy Goats in Norway " Journal of Dairy Science 90: 4863–4871.

Angiolillo, A., M. H. Yahyaoui, A.Sanchez, F. Pilla, and J. M. Folch. (2002). Characterization of a new genetic variant in the caprine -casein gene. J. Dairy Sci. 85:2679–2680.

Angulo, C., Diaz Carrillo, E., Munoz, A., Alonso, A., Jimenez, I. and Serradilla, J.M. (1994).Effect of electrophoretic goat's k-casein polymorphism on milk yield and main components V. World Congress on Genetis Applied to Livestock Production. Guelph: 333-336.

Barbieri, M.E., Manfredi, E., Elsen, J.M., Ricordeau, G., Bouillon, J., Grosclaude, F., Mahé, M.F. and Bibé, B. (1995). Influence du locus de la caséine αs1 sur les performances laitières et les paramètres génétiques des chèvres de race Alpine. *7* 27: 437-450

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

Barros A.S, Rutledge, D.N (2004) Principal components transform-partial least squares: A novel method to accelerate cross-validation in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 73 (2), pp. 245-255

Bevilacqua C., Ferranti P., Garro G., Veltri C., Lagonigro R., Leroux C., Pietrolà E., Addeo F., Pilla F., Chianese L., Martin P. (2002): Interallelic recombination is probably responsible for the occurrence of a new αS1-casein variant found in the goat species. Eur. J. Biochem., 269, 1293–1303

Bouniol, C., G. Brignon, M. F. Mahe´, and C. Printz. (1994). Biochemical and genetic analysis of variant C of caprine _s2-casein. Anim. Genet. 25:173–177

Caravaca, F., M. Amills, J. Jordana, A. Angiolillo, P. Aguera, C.Aranda, A. Menendez-Buxadera, A. Sanchez, J. Carrizosa, B.Urrutia, A. Sanchez, and J. M. Serradilla . (2008). Effect of αs1-casein (*CSN1S1*) genotype on milk CSN1S1 content in Malaguena and Murciano-Granadina goats. J. Dairy Res. 75:481– 484

Caroli, A. , Jann, O. , Budelli, E. , Bolla, P. , Jäger, S. and Erhardt, G. (2001), Genetic polymorphism of goat κ-casein (*CSN3*) in different breeds and characterization at DNA level. Animal Genetics, 32: 226–230. doi: 10.1046/j.1365-2052.2001.00765.x

Caroli, A., Chiatti, F., Chessa, S., Rignanese, D., Bolla, P., and Pagnacco,G. Focusing on the goat casein complex, *J. Dairy Sci.* **89** (2006), pp. 3178–3187.

Caroli, A., F. Chiatti, S. Chessa, D. Rignanese, E. Ibeagha-Awemu, and G. Erhardt. (2007). Characterization of the casein gene complex in West Africa goats and description of new S1-casein polymorphism. J. Dairy Sci. 90:2989–2996.

Chianese, L., P. Ferranti, G. Garro, R. Mauriello, and F. Addeo. (1997). Occurrence of three novel αS1-casein variants in goat milk. Pages 259–267 in Proc. IDF Milk Protein Polymorphism. Seminar II. International Dairy Federation, Brussels, Belgium.

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31

Chiatti, F., Caroli, A., Chessa, S., Bolla, P., and Pagnacco, G. (2005). Relationships between goat k-casein (CSN3) polymorphism and milk composition: The role of Biotechnology. Turin, Italy.

Chun and S. Keles (2009), Expression quantitative loci mapping with sparse partial least squares, Genetics, 182:79-90.

Clark AG. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122.

Clayton, D., Chapman, J., Cooper, J., 2004. Use of unphased multilocus genotype data in indirect association studies. Genet. Epidemiol., 27(4):415-428. [doi:10.1002/gepi. 20032]

Dagnachew, B.S. 'Casein SNPs and haplotypes: effect on milk production and quality in Norwegian goats'. MSc thesis. Norwegian University of Life Sciences, 2009.

Dagnachew, B.S,Thaller, G, Lien, S and Ådnøy,T (2010) Over-dominance in Breeding: Case of Deletion in Exon12 of *CSN1S1* in Norwegian Goats. Presented at the 9th World conference on Genetics Applied to Livestock Production. 1-6 August, 2010.

Díaz, E., Angulo-Heras, C., Jordana, J., Amills, M., Sánchez, A. and Serradilla, J.M. (1994). Casein variation in one Spanish breed of goats and its relation with production traits. In: *Book of abstracts in 44th annual meeting of EEAP*. Edinburgh (United Kingdom), 5-8 September 1994,_pp. 296.

Enns, P G. (1979). Using simulation to measure bias in principal components regression. In: *Proceedings of the 11th conference on Winter simulation - Volume 1* (WSC '79), Harold Joseph Highland (Ed.), Vol. 1. IEEE Press, Piscataway, NJ, USA, 153-157.

Excoffier L, Slatkin,M. (1995). Maximum-likelihood estimation of molecular haplotypes frequencies in a diploid population. Mol Biol Evol 12:921–927

Grosclaude, F., Ricordeau, G., Martin, P., Remeuf, F. and Vassal, L. (1994). Du gène au fromage: Le polymorphisme de la caséine αs1 caprine, ses effets, son évolution. *INRA PROD: ANIM*. 7: 3-19.

Grosclaude, F. and Martin, P. (1997). Casein polymorphisms in the goat. International Dairy Federation, New Zealand.

Hawley M, Kidd K (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Hayes B., Hagesæther N., Ådnøy T., Pellerud G., Berg P., Lien S. (2006). Effects on production traits of haplotypes among casein genes in Norwegian goats and evidence for a site of preferential recombination. Genetics, 174, 455-464.

Hayes H, Petit E, Bouniol C, Popescu P. (1993). Localisation of the alpha-S2- casein gene (CASAS2) to the homologous cattle, sheep and goat chromosome 4 by in situ hybridisation. Cytogenet. Cell Genet. 64, 282-285.

Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ, 2008 Simultaneous Analysis of All SNPs in Genome Wide and Re-Sequencing Association Studies. PLoSGenet 4(7):e10001 30 .doi:10.1371/journal.pgen.1000130

Jamrozik,J and Schaeffer L.R , Estimation of genetic parameters for a test day model with random regressions for yield traits of first lactation holsteins,*J. Dairy Sci.* 80 (1997), pp. 762–770.

Jann, O., Prinzenberg, E.M., Luikart, G., Caroli, A., Erhardt, G. (2004). High polymorphism in the k-casein (CSN3) gene from wild and domestic caprine species revealed by DNA sequencing. J. Dairy Res. 71, 188-195.

Lagonigro, R., E. Pietrola, M. 'Andrea, C. Veltri, and F. Pilla. (2001). Molecular genetic characterization of the goat _S2-casein E allele. Anim. Genet. 32:391–393.

Lamberet G., Degas C., Delacroix-Buchet A., Vassal L. (1996) Effect of characters linked to A and F caprine α_{s1}-casein alleles on goat flavours: cheesemaking with protein-fat exchange. Lait 76 , 349–61.

Mahé M.F. & Grosclaude F. (1989) α_{s1}-Cn^{D}, another allele associated with a decreased synthesis rate at the caprine α_{s1}-casein locus.Génétique, Séléction et Evolution 21 , 127–9.

Mahé, M.F., Manfredi, E., Ricordeau, G., Piacère, A. and Grosclaude, F. (1994). Effet des variants de la caséine αs1 sur les performances laitières: Analyse intradescendance de boucs de race Alpine.*Genet.Sel Evol.* 26: 151-157

Marletta, D., Bordonaro, S., Guastella, A.M., Avondo, M. and D'Urso, G. (2000). The Argentata dell'Etna goat population: Analysis of the polymorphism of alphas1- and beta- caseins. In: Proc. Of the Seminar on Genetic Markers and Milk Production: State of Art and Perspectives, Rome (Italy),15-16 February 2000.

Martin, P. and Addeo, R. (1996). Genetic polymorphism of casein in milk of goat and sheep. In: *Production and Utilization of Ewe and Goat milk. Proceeding of the IDF/ Greek committee*. Crete (Greece), pp. 45-58.

Martin P. (1993) Polymorphisme génétique des lactoprotéines caprines. Lait 73 , 511–32.

Martin P., Ollivier-Bousquet M., Grosclaude F. (1999) Genetic polymorphism of caseins: a tool to investigate casein micelle organization.International Dairy Journal 9 , 163–71.

Martin, P., Szymanowska, M., Zwierzchowski, L., and Leroux, C. (2002). "The impact of genetic polymorphisms on the protein composition of ruminant milks." Reproduction Nutrition Development 42: 433-459.

Meggiolaro, D., Crepaldi, P., Verdoglia, L., Marilli, M. and Cicogna, M. (2000). Preliminary study on alphas1-casein polymorphism in Val di Livo goats. In: Proc. of the Seminar on Genetic Markers and Milk Production: State of Art and Perspectives,Rome (Italy),15-16 February 2000. *B*26(3): 149-152.

Moioli, B., D'Andrea, M. and Pilla, F. (2007). "Candidate genes affecting sheep and goat milk quality." Small Ruminant Research 68: 179 - 192.

Morris C.A Hickey, S.M, Cullen N.G, Prosser C.G, Anderson, R.M, Tate, M.L (2005). Associations Between Beta-Casein Genotype And Milk Yield And Composition In Grazing Dairy Cows

Neveu, C., D. Molle`, J. Moreno, P. Martin, and J. Le`onid. (2002). Heterogeneity of caprine beta- casein elucidated by RP-HPLC/MS: Genetic variants and phosphorylation. J.Protein Chem. 21:557–567.

Pirisi A., Colin P., Laurent F., Scher J., Parmentier M. (1994) Comparison of milk composition, cheesemaking properties and textural characteristics of the cheese from two groups of goats with a high or low rate of α_{s1}-casein synthesis. International Dairy Journal 4 , 329–45.

Prinzenberg E.M., Krause I., Erhardt G. (1999) SSCP analysis at the bovine *CSN3* locus discriminates six alleles corresponding to known protein variants (*A, B, C, E, F, G*) and three new DNA Polymorphisms (*H, I, A_1*). Animal Biotechnology 10 , 49–62.

Pizzilo, M., Cogliandro, E., Rubino, R., and Fedele, V. 1996. Relationship between somatic cells and milk quality in different goat production systems. In: Rubino, R. (Ed.), Proceedings of the International Symposium on Somatic Cells and Milk of Small Ruminants. EAAP Publication No 77, Wageningen Pers, Bella, Italy, September 25-27, 1994. pp, 269-273

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.2.8. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramunno L., Cosenza G., Pappalardo M., Pastore N., Gallo D., Di Gregorio P., Masina P. (2000): Identification of the goat CSN1S1 F allele by means of PCR-RFLP. Anim. Genet., 31, 333–346.

Ramunno, L., Cosenza, G., Pappalardo, M., Longobardi, E., Gallo, D., Pastore, N., Di Gregorio, P., and Rando, A. (2001). "Characterization of two new alleles at the goat CSN1S2 locus." Animal Genetics , 32: 264 - 268.

Ramunno L., Cosenza G., Rando A., Pauciullo A., Illario R., Gallo D., Di Bernardino D., Masina P. (2005): Comparative analysis of gene of goat CSN1S1 F and N alleles and characterization of CSN1S1 transcript variants in mammary gland. Gene, 345, 289–299.

Rando A., Ramunno L., Masina P. (2000) Mutations in casein genes. Zootecnica e Nutrizione Animale 26 , 105–14.

Remeuf F., Verdalet-Guzman I., Lenoir J. (1995) Technological adjustment of goat milk containing low synthesis-rate α_{s1}-casein variants. International Dairy Journal 5 , 381–92.

Risvik, H. “Principal Component Analysis (PCA) & NIPALS algorithm”, http://folk.uio.no/henninri/pca_module/pca_nipals.pdf, 2007

Sacchi, P. Chessa, S., Budelli, E., Bolla, P., Ceriotti, G., Soglia, D., Rasero, R., Cauvin, E., Caroli, A. (2005). Casein haplotype structure in five Italian goat breeds. J. Dairy Sci. 88,1561–1568.

Serradilla, J.M (2003). The Goat alpha s1 casein gene: the paradigm of the use of a major gene to improve milk quality? Options Mediterreneenes. Serie 55:99-106

Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotypes reconstruction from population data. American Journal of Human Genetics 68: 978-989

Threadgill D.W., Womack J.E. (1990) Genomic analysis of the major bovine milk protein genes.Nucleic Acids Res., 18, 6935–6942.

Yahyaoui, M. H. (2003). Genetic polymorphism in goat: study of the kappa casein, beta lactoglobulin, and stearoyl coenzyme A desaturase gene. Department of Animal Sciences. Bellaterra, University of Barcelona. Phd: 169.

Zeng, S.S., Escobar, E.N.: Effect of parity and milk production on Somatic cell count, standard plate count and composition of goat milk. Small Rum. Res., 1995, 17, 269-274.

APPENDIX

Table 1: Coding of the 38 casein SNPs

Coding	Gene	Location	Alleles	Frequency of minor allele
SNP1	CSN1S1	Promoter	A(G)	0.050
SNP2	CSN1S1	Promoter	C(T)	0.049
SNP4	CSN1S1	Promoter	G(A)	0.130
SNP5	CSN1S1	Promoter	G(A)	0.145
SNP6	CSN1S1	Promoter	G(A)	0.147
SNP7	CSN1S1	Promoter	C(T)	0.146
SNP8	CSN1S1	Promoter	G(A)	0.144
SNP9	CSN1S1	Exon 4	T(C)	0.150
SNP10	CSN1S1	Exon 4	C(G)	0.160
SNP11	CSN1S1	Exon 9	C(D)*	0.037
SNP12	CSN1S1	Intron 8	A(G)	0.148
SNP13	CSN1S1	Exon 10	C(G)	0.148
SNP14	CSN1S1	Exon 12	D(N)**	0.245
SNP15	CSN1S1	Exon 17	C(T)	0.116
SNP16	CSN2	Exon 7	T(C)	0.062
SNP17	CSN2	Promoter	A(G)	0.061
SNP18	CSN2	Promoter	G(A)	0.024
SNP19	CSN2	Promoter	A(G)	0.059
SNP20	CSN2	Promoter	(A)T	0.061
SNP21	CSN2	Promoter	C(T)	0.062
SNP22	CSN1S2	2 Exon 3	G(A)	0.078
SNP24	CSN1S2	2 Exon 16	C(G)	0.047
SNP25	CSN1S2	2 Exon 16	C(T)	0.318
SNP26	CSN1S2	2 Exon 16	A(T)	0.315
SNP27	CSN3	Promoter	G(A)	0.421
SNP28	CSN3	Promoter	G(A)	0.493
SNP29	CSN3	Promoter	(A)G	0.002
SNP30	CSN3	Promoter	T(A)	0.494
SNP31	CSN3	Promoter	T(A)	0.466
SNP32	CSN3	Promoter	G(C)	0.494
SNP33	CSN3	Promoter	T(G)	0.465
SNP34	CSN3	Promoter	T(G)	0.476
SNP35	CSN3	Promoter	A(G)	0.092
SNP36	CSN3	Promoter	T(C)	0.317
SNP37	CSN3	Promoter	G(T)	0.328
SNP38	CSN3	Promoter	A(G)	0.180
SNP39	CSN3	Promoter	G(A)	0.092
SNP40	CSN3	Exon 4	C(T)	0.097

The minor alleles are in parentheses

*D represents a deletion **N represents a non-deletion

Table 2a: Additive effect of major allele of each SNP on milk traits (Estimated using Model 1: PLS I)

SNPs	TRAITS									
	Milk yield (kg)		log(SCC)		Lactose %		Fat %		Protein %	
	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
SNP1	0.0008	*0.001*	0.0017	*0.001*	-0.0044	*0.009*	**-0.0590**	*0.018*	-0.0022	*0.002*
SNP2	-0.0008	*0.000*	-0.0002	*0.001*	0.0051	*0.009*	0.0180	*0.017*	-0.0009	*0.001*
SNP4	-0.0008	*0.001*	0.0010	*0.001*	-0.0036	*0.008*	0.0012	*0.013*	-0.0016	*0.001*
SNP5	-0.0002	*0.001*	0.0009	*0.001*	0.0013	*0.002*	-0.0005	*0.007*	**-0.0017**	*0.001*
SNP6	-0.0006	*0.001*	0.0008	*0.001*	0.0003	*0.002*	0.0113	*0.015*	-0.0003	*0.001*
SNP7	-0.0004	*0.001*	0.0009	*0.001*	0.0008	*0.002*	-0.0085	*0.008*	**-0.0018**	*0.001*
SNP8	-0.0002	*0.001*	0.0008	*0.001*	0.0001	*0.002*	0.0056	*0.008*	-0.0007	*0.001*
SNP9	-0.0005	*0.001*	0.0011	*0.001*	0.0001	*0.002*	0.0100	*0.007*	-0.0002	*0.001*
SNP10	-0.0006	*0.001*	0.0006	*0.001*	0.0013	*0.002*	0.0042	*0.011*	-0.0009	*0.001*
SNP11	0.0000	*0.000*	-0.0005	*0.001*	0.0000	*0.002*	**0.0422**	*0.011*	0.0016	*0.001*
SNP12	-0.0002	*0.001*	0.0009	*0.001*	0.0018	*0.002*	0.0020	*0.006*	-0.0015	*0.001*
SNP13	-0.0002	*0.001*	0.0009	*0.001*	0.0019	*0.003*	0.0022	*0.006*	-0.0012	*0.001*
SNP14	-0.0005	*0.001*	0.0030	*0.002*	-0.0020	*0.007*	**-0.0863**	*0.031*	**-0.0084**	*0.003*
SNP15	-0.0005	*0.001*	0.0016	*0.001*	0.0015	*0.003*	**-0.0568**	*0.015*	**-0.0030**	*0.001*
SNP16	**-0.0011**	*0.000*	-0.0003	*0.001*	0.0050	*0.009*	0.0212	*0.019*	-0.0004	*0.002*
SNP17	**-0.0010**	*0.000*	-0.0004	*0.001*	0.0058	*0.010*	0.0252	*0.017*	-0.0010	*0.002*
SNP18	0.0003	*0.000*	0.0001	*0.000*	0.0007	*0.002*	-0.0143	*0.012*	0.0006	*0.001*
SNP19	0.0005	*0.001*	-0.0003	*0.001*	0.0000	*0.003*	**0.0298**	*0.012*	0.0026	*0.002*
SNP20	0.0004	*0.001*	-0.0003	*0.001*	0.0002	*0.003*	**0.0324**	*0.012*	0.0028	*0.002*
SNP21	0.0004	*0.001*	-0.0004	*0.001*	0.0002	*0.003*	**0.0306**	*0.012*	0.0025	*0.001*

Effects that are significant at 5% are in **bold** print
The standard errors (SE) for the given estimated SNP effects are in *italics*

Table 2b: Additive effect of major allele of each SNP on milk traits (Estimated using Model 1: PLS I)

SNPs	TRAITS									
	Milk yield (kg)		log(SCC)		Lactose %		Fat %		Protein %	
	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
SNP22	0.0000	*0.001*	-0.0003	*0.001*	-0.0007	*0.003*	0.0155	*0.017*	-0.0010	*0.002*
SNP24	-0.0006	*0.001*	-0.0002	*0.001*	0.0010	*0.003*	0.0001	*0.026*	0.0018	*0.002*
SNP25	0.0012	*0.001*	-0.0013	*0.001*	0.0051	*0.006*	0.0348	*0.022*	-0.0015	*0.002*
SNP26	0.0011	*0.001*	-0.0017	*0.001*	0.0074	*0.010*	0.0337	*0.022*	-0.0021	*0.002*
SNP27	-0.0018	*0.001*	**0.0016**	*0.001*	-0.0007	*0.005*	0.0074	*0.024*	0.0030	*0.002*
SNP28	**-0.0020**	*0.001*	0.0011	*0.001*	0.0021	*0.008*	-0.0103	*0.013*	-0.0021	*0.001*
SNP29	-0.0001	*0.000*	0.0000	*0.000*	0.0006	*0.001*	-0.0001	*0.003*	0.0002	*0.000*
SNP30	-0.0018	*0.001*	**0.0016**	*0.001*	-0.0013	*0.004*	-0.0043	*0.012*	-0.0005	*0.001*
SNP31	-0.0017	*0.001*	0.0009	*0.001*	0.0048	*0.012*	0.0167	*0.014*	0.0015	*0.001*
SNP32	**-0.0019**	*0.001*	**0.0017**	*0.001*	-0.0009	*0.005*	-0.0036	*0.012*	-0.0009	*0.001*
SNP33	-0.0016	*0.001*	0.0010	*0.001*	0.0032	*0.009*	0.0126	*0.014*	0.0017	*0.001*
SNP34	**-0.0019**	*0.001*	0.0010	*0.001*	-0.0007	*0.005*	0.0219	*0.019*	0.0017	*0.002*
SNP35	0.0000	*0.001*	0.0004	*0.001*	-0.0008	*0.003*	-0.0051	*0.014*	**0.0048**	*0.002*
SNP36	**0.0022**	*0.001*	-0.0016	*0.001*	0.0006	*0.004*	-0.0247	*0.016*	**-0.0048**	*0.001*
SNP37	**0.0021**	*0.001*	-0.0017	*0.001*	0.0022	*0.004*	0.0043	*0.016*	-0.0022	*0.001*
SNP38	0.0000	*0.001*	0.0001	*0.001*	-0.0002	*0.003*	0.0142	*0.014*	**0.0046**	*0.001*
SNP39	0.0000	*0.001*	0.0003	*0.001*	0.0012	*0.004*	-0.0047	*0.014*	**0.0056**	*0.002*
SNP40	0.0001	*0.001*	0.0005	*0.001*	-0.0010	*0.003*	-0.0097	*0.014*	**0.0048**	*0.002*

Effects that are significant at 5% are in **bold** print
The standard errors (SE) for the given estimated SNP effects are in *italics*

Table3a: Additive effect of major allele of each SNP on milk traits (Estimated with Model 2:PLS II)

SNPs	TRAITS							
	log(SCC)		Fat %		Lactose %		Protein %	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
SNP1	**0.021**	*0.008*	**-0.058**	*0.017*	-0.004	*0.003*	**-0.007**	*0.003*
SNP2	-0.008	*0.006*	0.018	*0.015*	0.002	*0.002*	0.001	*0.002*
SNP4	-0.002	*0.005*	0.000	*0.011*	0.001	*0.001*	-0.001	*0.001*
SNP5	-0.001	*0.003*	0.000	*0.006*	0.001	*0.001*	-0.001	*0.001*
SNP6	-0.006	*0.005*	0.011	*0.013*	0.002	*0.001*	0.000	*0.002*
SNP7	0.002	*0.003*	-0.007	*0.007*	0.001	*0.001*	**-0.002**	*0.001*
SNP8	-0.004	*0.003*	0.006	*0.007*	0.002	*0.001*	-0.001	*0.001*
SNP9	-0.005	*0.003*	0.009	*0.007*	**0.002**	*0.001*	0.000	*0.001*
SNP10	-0.004	*0.004*	0.006	*0.010*	0.002	*0.001*	-0.001	*0.001*
SNP11	**-0.015**	*0.005*	**0.041**	*0.011*	0.004	*0.002*	0.005	*0.002*
SNP12	-0.002	*0.002*	0.002	*0.006*	0.002	*0.001*	-0.001	*0.001*
SNP13	-0.003	*0.003*	0.003	*0.006*	0.002	*0.001*	-0.001	*0.001*
SNP14	**0.032**	*0.014*	**-0.089**	*0.029*	-0.007	*0.004*	-0.012	*0.006*
SNP15	**0.019**	*0.007*	**-0.054**	*0.014*	-0.004	*0.002*	**-0.007**	*0.003*
SNP16	-0.009	*0.007*	0.022	*0.018*	0.003	*0.002*	0.002	*0.002*
SNP17	-0.011	*0.007*	0.025	*0.016*	0.003	*0.002*	0.002	*0.002*
SNP18	0.006	*0.005*	-0.013	*0.011*	-0.002	*0.001*	-0.001	*0.001*
SNP19	**-0.010**	*0.004*	**0.029**	*0.011*	0.002	*0.002*	**0.004**	*0.002*
SNP20	**-0.011**	*0.004*	**0.032**	*0.011*	0.002	*0.002*	**0.004**	*0.002*
SNP21	**-0.010**	*0.004*	**0.031**	*0.011*	0.002	*0.002*	**0.004**	*0.002*

Effects that are significant at 5% are in **bold** print

The standard errors (SE) for the given estimated SNP effects are in *italics*

Table 3b: Additive effect of major allele of each SNP on milk traits (Estimated using Model 2:PLS II)

SNPs	TRAITS							
	log(SCC)		Fat %		Lactose %		Protein %	
	Estimate	*SE*	Estimate	SE	Estimate	*SE*	Estimate	*SE*
SNP22	-0.004	*0.006*	0.015	*0.015*	0.001	*0.001*	0.002	*0.002*
SNP24	0.000	*0.009*	0.002	*0.023*	0.000	*0.002*	0.000	*0.002*
SNP25	**-0.016**	*0.008*	0.031	*0.021*	0.004	*0.003*	0.001	*0.003*
SNP26	**-0.016**	*0.008*	0.032	*0.021*	0.004	*0.003*	0.001	*0.003*
SNP27	-0.002	*0.008*	0.006	*0.021*	0.001	*0.002*	0.001	*0.002*
SNP28	0.003	*0.005*	-0.009	*0.012*	-0.001	*0.001*	-0.001	*0.001*
SNP29	0.000	*0.001*	0.000	*0.003*	0.000	*0.000*	0.000	*0.000*
SNP30	0.002	*0.004*	-0.005	*0.011*	0.000	*0.001*	0.000	*0.001*
SNP31	-0.007	*0.005*	0.017	*0.012*	0.002	*0.002*	0.002	*0.002*
SNP32	0.002	*0.004*	-0.005	*0.010*	0.000	*0.001*	0.000	*0.001*
SNP33	-0.005	*0.005*	0.013	*0.012*	0.001	*0.001*	0.002	*0.002*
SNP34	-0.008	*0.007*	0.021	*0.017*	0.002	*0.002*	0.003	*0.002*
SNP35	0.004	*0.005*	-0.005	*0.013*	-0.001	*0.001*	0.001	*0.002*
SNP36	0.005	*0.006*	-0.022	*0.015*	-0.001	*0.002*	**-0.005**	*0.002*
SNP37	-0.006	*0.006*	0.005	*0.014*	0.002	*0.001*	-0.002	*0.002*
SNP38	-0.002	*0.005*	0.014	*0.012*	0.000	*0.002*	**0.004**	*0.002*
SNP39	0.003	*0.005*	-0.004	*0.013*	-0.001	*0.001*	0.001	*0.002*
SNP40	0.006	*0.005*	-0.009	*0.013*	-0.002	*0.001*	0.000	*0.002*

Effects that are significant at 5% are in **bold** print

The standard errors (SE) for the given estimated SNP effects are in *italics*

Table 4a: Additive effect of major allele of each SNP on milk traits (Estimated using Model 3:PCA/MME)

SNPs	TRAITS									
	Milk yield (kg)		los(SCC)		Lactose %		Fat %		Protein %	
	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
SNP1	0.138	*0.094*	0.077	*0.082*	-0.018	*0.113*	-0.140	*0.089*	-0.028	*0.110*
SNP2	0.027	*0.071*	-0.041	*0.061*	0.035	*0.082*	0.033	*0.065*	-0.001	*0.080*
SNP4	-0.374	*0.228*	0.065	*0.165*	-0.091	*0.212*	0.099	*0.170*	0.011	*0.207*
SNP5	**0.143**	*0.081*	-0.027	*0.056*	0.011	*0.072*	-0.016	*0.057*	-0.014	*0.070*
SNP6	-0.105	*0.158*	-0.016	*0.142*	0.019	*0.184*	0.012	*0.146*	0.015	*0.179*
SNP7	0.045	*0.079*	-0.028	*0.070*	0.027	*0.091*	-0.018	*0.073*	-0.001	*0.089*
SNP8	0.105	*0.113*	-0.023	*0.087*	-0.003	*0.113*	-0.002	*0.090*	-0.015	*0.110*
SNP9	-0.059	*0.066*	-0.015	*0.062*	-0.011	*0.081*	0.023	*0.064*	0.004	*0.078*
SNP10	-0.075	*0.055*	-0.018	*0.054*	-0.013	*0.073*	0.048	*0.057*	0.018	*0.071*
SNP11	-0.096	*0.114*	-0.052	*0.096*	0.003	*0.130*	0.088	*0.103*	0.020	*0.127*
SNP12	**0.132**	*0.075*	-0.027	*0.053*	0.013	*0.069*	-0.019	*0.055*	-0.013	*0.067*
SNP13	**0.147**	*0.085*	-0.026	*0.058*	0.012	*0.075*	-0.021	*0.060*	-0.013	*0.073*
SNP14	-0.044	*0.123*	0.152	*0.111*	0.003	*0.149*	**-0.205**	*0.118*	-0.055	*0.145*
SNP15	0.031	*0.133*	0.045	*0.117*	0.033	*0.154*	-0.119	*0.122*	-0.012	*0.150*
SNP16	-0.020	*0.099*	-0.047	*0.089*	0.012	*0.117*	0.059	*0.093*	0.000	*0.114*
SNP17	-0.102	*0.080*	-0.023	*0.069*	0.013	*0.091*	0.060	*0.073*	0.010	*0.089*
SNP18	0.113	*0.118*	0.040	*0.109*	0.006	*0.147*	-0.059	*0.117*	-0.013	*0.143*
SNP19	0.068	*0.059*	-0.009	*0.055*	0.007	*0.072*	0.024	*0.058*	0.002	*0.070*
SNP20	0.004	*0.055*	-0.005	*0.050*	0.000	*0.066*	0.038	*0.053*	0.010	*0.064*
SNP21	0.001	*0.053*	-0.011	*0.048*	0.003	*0.064*	0.039	*0.051*	0.012	*0.062*

Effects that are significant at 5% are in **bold** print
The standard errors (SE) for the given estimated SNP effects are in *italics*

Table 4b: Additive effect of major allele of each SNP on milk traits (Estimated using Model 3:PCA/MME)

SNPs	TRAITS									
	Milk yield (kg)		los(SCC)		Lactose %		Fat %		Protein %	
	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
SNP22	0.050	*0.080*	-0.011	*0.076*	0.029	*0.103*	0.017	*0.081*	-0.009	*0.100*
SNP24	-0.185	*0.236*	-0.102	*0.216*	0.047	*0.283*	0.057	*0.224*	0.065	*0.275*
SNP25	-0.080	*0.078*	0.001	*0.083*	-0.002	*0.111*	0.101	*0.088*	0.017	*0.108*
SNP26	-0.012	*0.078*	-0.031	*0.090*	0.037	*0.117*	0.038	*0.093*	-0.005	*0.113*
SNP27	-0.018	*0.162*	0.105	*0.150*	-0.039	*0.193*	-0.021	*0.154*	0.031	*0.188*
SNP28	-0.051	*0.093*	-0.022	*0.075*	0.016	*0.098*	-0.003	*0.078*	0.013	*0.096*
SNP29	-0.007	*0.007*	0.002	*0.006*	0.000	*0.007*	0.003	*0.006*	0.000	*0.007*
SNP30	0.028	*0.069*	0.004	*0.061*	-0.013	*0.080*	-0.016	*0.063*	-0.009	*0.078*
SNP31	0.010	*0.113*	-0.045	*0.100*	0.059	*0.131*	-0.003	*0.104*	0.005	*0.128*
SNP32	0.045	*0.069*	-0.002	*0.073*	-0.004	*0.095*	-0.028	*0.075*	-0.013	*0.093*
SNP33	-0.028	*0.115*	-0.041	*0.099*	0.052	*0.129*	0.005	*0.103*	0.006	*0.126*
SNP34	0.004	*0.188*	-0.003	*0.162*	-0.029	*0.215*	0.125	*0.169*	0.001	*0.209*
SNP35	-0.010	*0.051*	-0.005	*0.047*	0.014	*0.063*	-0.010	*0.050*	0.014	*0.061*
SNP36	**0.207**	*0.101*	-0.031	*0.167*	0.017	*0.216*	-0.122	*0.171*	-0.056	*0.210*
SNP37	-0.030	*0.092*	-0.017	*0.123*	-0.001	*0.160*	-0.003	*0.126*	0.023	*0.155*
SNP38	0.023	*0.070*	-0.002	*0.069*	0.038	*0.092*	0.016	*0.073*	0.004	*0.090*
SNP39	0.005	*0.049*	-0.009	*0.053*	0.013	*0.070*	-0.014	*0.056*	0.011	*0.068*
SNP40	-0.019	*0.050*	0.011	*0.047*	-0.006	*0.064*	0.023	*0.051*	0.023	*0.062*

Effects that are significant at 5% are in **bold** print
The standard errors (SE) for the given estimated SNP effects are in *italics*